SuperMix

MIXED EFFECTS MODELS

Don Hedeker, Robert Gibbons, Mathilda du Toit, and Yan Cheng

Please download the examples from our website and unzip them into a convenient folder location.

Preface

This primer describes the SuperMix program for mixed-effects models, also called multilevel, hierarchical, or random-effects models. These models can be used for the analysis of longitudinal data, where each individual may be measured at a different number of occasions. They can also be used for clustered data, for example, patients nested within clinics.

The primer has been written to introduce mixed-effects models to researchers new to this field. It also serves as a guide to mental health researchers who are already familiar with the basic principles of hierarchical models.

SuperMix fits mixed-effects models with continuous, count, ordinal, nominal, and survival outcome variables to nested data. Although SuperMix allows for up to three levels of nesting, the illustrative examples given in the primer are for two-level data. The only exception is an example of a three-level model with a continuous outcome variable. This example demonstrates that it is not difficult to interpret the results of three-level models once the concepts and interpretation of two-level models are mastered.

In the primer, the focus is on the "how to" part of mixed-effects modeling. Chapter 1 starts with a motivation for using mixed-effects models rather than traditional regression models. Chapter 2 describes the graphical user interface and Chapter 3 contains examples of all the outcome variable types. Since graphical displays form an integral part of conveying findings to policy makers, Chapter 4 reviews the graphical capabilities of SuperMix with illustrations based on the data sets and models used in the various sections of Chapter 2.

The authors hope that the easy to use SuperMix user's interface, together with the selection of examples contained in the primer, will make mental health researchers aware of the strengths of mixed-effects modeling, and that it will contribute to a

better understanding of the complex relationships that often exist among variables in the case of nested data.

Table of contents

List of tables and figures

1   Introduction to mixed-effects models

In mental health research, data often have a hierarchical structure. Consider, for example, the study of the effect of administering medication, such as an antidepressant, over time to a patient diagnosed with depression. For each patient, the effect of the drug over time can be modeled in terms of the time since the start of treatment, and in terms of any other information obtained at each time of measurement during the study. For example, a measure of family support at the time of measurement can be incorporated into such a model. The outcome would be described as a function of the information collected at the measurement level, and could be viewed as a measurement-level model for each individual patient. However, the gender of the patient, and other characteristics that may influence the outcome but that do not change over time, cannot easily be accommodated in the model proposed, as the model is at a measurement, rather than a patient, level. It may also be of interest to compare patients in terms of their improvement trajectories, which is easier when outcomes are described in terms of patients rather than measurements.

To allow us to study all of these areas of interest simultaneously, a model that acknowledges the data's inherent hierarchical structure (measurements nested within individual patients), and allows the study of both measurement and patient level models along with the way these models are related to each other, is needed. As patients may drop out during the study period, the model should also be suitable for the analysis of unbalanced longitudinal data where each individual may be measured at a different number of occasions, or even at different time points. In this chapter, data from a study described in Reisby *et. al.*, (1977) that focused on the longitudinal relationship between imipramine and desipramine plasma levels and clinical response in 66 depressed inpatients are used to illustrate the need for and basic characteristics of a mixed-effects regression model.

In the study, patients received 225 mg/day doses of imipramine for a period of four weeks, following a placebo period of 1 week. Subjects were rated with the Hamilton depression rating scale (HDRS), and ratings occurred twice during the baseline placebo week (at the start and end of this week), and at the end of each of the four

treatment weeks of the study. Plasma level measurements of both IMI and its metabolite DMI were made at the end of each week, and the measurements are represented in the data to follow as a single measure. Additional information includes a diagnosis of endogenous or non-endogenous depression for each patient. Of the 66 patients, 37 were classified as having endogenous depression, and 29 as having non-endogenous depression.

Although the total number of subjects in this study was 66, the number of subjects with measures at each of the weeks fluctuated, as shown in Table 1.1. The sample size, in terms of ratings and plasma level measurements, is 375.

Table 1.1: Number of subjects at measurement occasions

Time of measurement	No of patients
Week 0 (start of placebo week)	61
Week 1 (end of placebo week)	63
Week 2 (end of first drug treatment week)	65
Week 3 (end of second drug treatment week)	65
Week 4 (end of third drug treatment week)	63
Week 5 (end of fourth drug treatment week)	58

Data for the first 10 patients are shown in Table 1.2 below. The HDRS rating is given in the second column of the table, followed in each case by the week in which information was obtained. The column with header ENDOG gives the depression classification for the patient as made at the beginning of the study.

Table 1.2: Data for 10 depressed patients from Reisby data

PATIENT	HDRS	WEEK	ENDOG	PATIENT	HDRS	WEEK	ENDOG
101	26	0	0	107	21	0	1
101	22	1	0	107	21	1	1
101	18	2	0	107	16	2	1
101	7	3	0	107	19	3	1
101	4	4	0	107	6	5	1
101	3	5	0	108	21	0	1
103	33	0	0	108	22	1	1
103	24	1	0	108	11	2	1
103	15	2	0	108	9	3	1
103	24	3	0	108	9	4	1
103	15	4	0	108	7	5	1
103	13	5	0	113	21	0	0
104	29	0	1	113	23	1	0
104	22	1	1	113	19	2	0
104	18	2	1	113	23	3	0
104	13	3	1	113	23	4	0
104	19	4	1	114	17	1	0
104	0	5	1	114	11	2	0
105	22	0	0	114	13	3	0
105	12	1	0	114	7	4	0
105	16	2	0	114	7	5	0
105	16	3	0	115	16	1	1
105	13	4	0	115	16	2	1
105	9	5	0	115	16	3	1
106	21	0	1	115	16	4	1
106	25	1	1	115	11	5	1
106	23	2	1				
106	18	3	1				
106	20	4	1				

Of interest here is whether administering the medication led to an improved, *i.e.*, lower, HDRS rating over the study period. These data have the characteristic that,

while plasma level measurements and HDRS ratings were made repeatedly for each patient, the other variable of interest thought to influence the HDRS ratings is whether a patient was classified as suffering from endogenous depression or not. Thus, the influence of the type of depression on the outcome over time needs to be evaluated along with the effect of the medication over time.

While the time of measurement, plasma level measurements, and HDRS ratings were repeatedly recorded and can be viewed as time-related characteristics of the repeated measurements design described here, the depression classification is, like gender, a patient characteristic that does not change over time. As such, the data as a whole can be viewed as having a hierarchical structure, with measurement-related characteristics over time forming the lowest level of the hierarchy; all measurements for each patient are therefore *nested* within that patient. The patients, in turn, form the next level of the hierarchy, and the depression classification is a potential predictor at this level.

1.1   Fixed-effects regression ignoring data clustering

Before proceeding with a mixed-effects analysis of these data, we first look at a fixed-effects analysis that ignores the clustering of measurements within patients. Note that SuperMix can be used for this purpose, and that the analysis is essentially equivalent to performing a traditional multiple linear regression analysis using maximum likelihood, and not least squares, estimation.

Using the information for the first 10 patients, for whom 55 measurements were available, we fit a model where the HDRS rating is modeled in terms of the time at which a measurement was made. In terms of the variables shown in Table 1.2, we have

$$y_{ij} = \beta_0 + \beta_1 (\text{WEEK})_{ij} + e_{ij} \qquad (1.1)$$

where y_{ij} denotes the HDRS measurement at time j ($j = 0, 1, 2, 3, 4,$ or 5) for patient i. $(\text{WEEK})_{ij}$ indicates the associated time of measurement, and e_{ij} measurement error. The coefficients β_0 and β_1 are the fixed, but unknown, parameters to be estimated. The e_{ij} are assumed to have a normal distribution, with mean 0 and variance σ^2.

For this analysis, we obtain estimates of β_0 and β_1 of 23.746 and –2.978 respectively. The estimated HDRS rating is plotted over time in Figure 1.1.
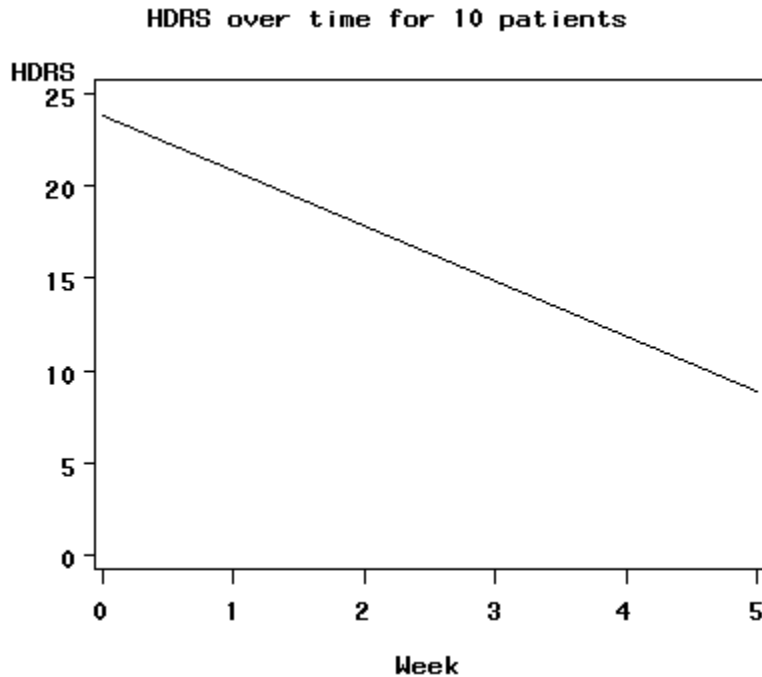
Figure 1.1: Fixed-effects regression line for 10 patients

In addition, an estimate of σ^2 of 23.3723 was obtained. The results show that the average predicted HDRS score, $\hat{\beta}_0$, at the beginning of the study is 23.746. The coefficient representing the effect of the predictor WEEK, $\hat{\beta}_1$, indicates a predicted decrease in HDRS score over the study period: a decrease of 2.978 in the HDRS score is expected between any two consecutive measurements. The coefficient β_1 is commonly referred to as a "slope" coefficient, as it indicates both the direction of the relationship between the predictor and the outcome, and the magnitude of the expected change in outcome associated with changes in the predictor. The relationship between the time of measurement and the outcome is indicated as highly significant.

1.2   Fixed-effects regression including data clustering

As noted by Hedeker, Gibbons & Flay (1994) and others, ignoring the data clustering often results in statistical tests which are too liberal, resulting in falsely rejecting the null hypothesis too often. In terms of our data, where multiple measurements "belong" to each patient, it is reasonable to expect that measurements for a given patient may be more similar to each other than to any other measurement, regardless of the patient it was obtained for. Thus, it may be reasonable to assume that the measurements for a given patient may be correlated. In addition, if it is indeed true that the depression classification of a patient impacts on the HDRS score over time, ignoring both the clustering effect and the initial classification will very likely lead to erroneous conclusions concerning the effect of treatment on the HDRS score.

To start addressing these concerns, we modify the previous model to take the clustering of measurements within patients into account. We do so by fitting a line similar to that given in Equation (1.1) for each individual patient. Table 1.3 shows the estimates of β_0 and β_1 for individual patients, and Figure 1.2 a graphical representation of the results. The estimated coefficients for the intercepts and time slopes of the individual patients ($\hat{\beta}_0$ and $\hat{\beta}_1$ respectively) in Table 1.3 show that the

predicted intercepts of patients differ considerably. Patients 114 and 115 start with a predicted initial rating of 18, which is considerably lower than the predicted initial rating of 28 for patient 104. Recall that in the previous analysis, we obtained a value of 23.746 for $\hat{\beta}_0$. While assuming an initial rating of 23.746 as a description of where patients start out may be acceptable for patients 106 or 107, it does not provide an adequate description of the initial status of patients 104, 114, or 115. A "one size fits all" policy for obtaining an estimate of the initial status of patients is clearly inadequate, and does not describe the initial status for individual patients satisfactorily.

Table 1.3: Fixed-effects regression results for 10 patients

Patient	Intercept	Week
101	26.1904	−5.1429
103	24.8000	−2.2000
104	28.1905	−4.5429
105	19.0952	−1.7714
106	23.2000	−0.9000
107	22.9324	−2.8784
108	21.0952	−3.1714
113	21.0000	0.4000
114	18.2000	−2.4000
115	18.0000	−1.0000

This conclusion is also apparent from Figure 1.2. Indeed, the patients not only start at different places, they end up in different places, too. For example, the regression line for patient 104, with a predicted initial rating of 28, is one of those closest to zero at the conclusion of the study period, with a predicted final HDRS rating of approximately 1. Not only will individual differences in initial status between patients have to be addressed, but also differences in their rates of improvement over the course of the study will have to be accommodated in the model.

HDRS over time for 10 patients
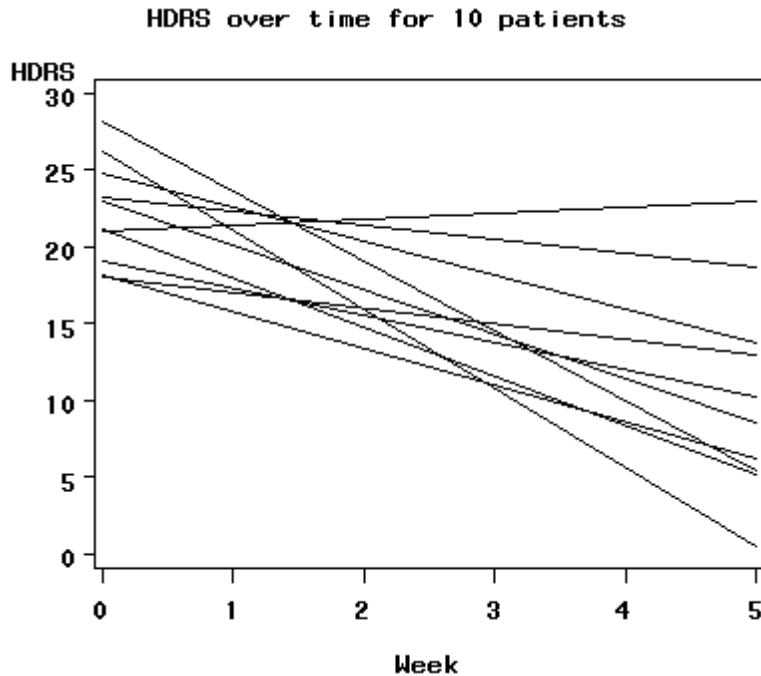
Figure 1.2: Individual fixed-effects regression lines for 10 patients

1.3 Fixed-effects regression with dummy variables

Up to this point, we have considered two approaches for the modeling of the HDRS ratings. In the first, all the data were pooled and a common regression model was fitted to the data. In the second approach, a regression line was fitted to each patient's measurements. A summary of the estimated intercepts and slopes showed substantial between-patient variation. The disadvantage of this method is that ten separate regression models are fitted. Ideally, a researcher would want to fit a single model that conveys information about between-subject variability.

One approach would be to do a regression analysis with dummy variables. Table 1.4 below shows the data for the first two and last two patients.

We use a dummy variable to represent each patient, coded as follows:

$$D_j = 1 \text{ for patient } j, \ , = 1, 2, \ldots, 10$$
$$= 0 \text{ otherwise.}$$

Table 1.4: Results of random intercept model

PATIENT	HDRS	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	WEEK
101	26	1	0	0	0	0	0	0	0	0	0	0
101	22	1	0	0	0	0	0	0	0	0	0	1
101	18	1	0	0	0	0	0	0	0	0	0	2
101	7	1	0	0	0	0	0	0	0	0	0	3
101	4	1	0	0	0	0	0	0	0	0	0	4
101	3	1	0	0	0	0	0	0	0	0	0	5
103	33	0	1	0	0	0	0	0	0	0	0	0
103	24	0	1	0	0	0	0	0	0	0	0	1
103	15	0	1	0	0	0	0	0	0	0	0	2
103	24	0	1	0	0	0	0	0	0	0	0	3
103	15	0	1	0	0	0	0	0	0	0	0	4
103	13	0	1	0	0	0	0	0	0	0	0	5
114	17	0	0	0	0	0	0	0	0	1	0	1
114	11	0	0	0	0	0	0	0	0	1	0	2
114	13	0	0	0	0	0	0	0	0	1	0	3
114	7	0	0	0	0	0	0	0	0	1	0	4
114	7	0	0	0	0	0	0	0	0	1	0	5
115	16	0	0	0	0	0	0	0	0	0	1	1
115	16	0	0	0	0	0	0	0	0	0	1	2
115	16	0	0	0	0	0	0	0	0	0	1	3
115	16	0	0	0	0	0	0	0	0	0	1	4
115	11	0	0	0	0	0	0	0	0	0	1	5

The following regression model is fitted to the data:

$$\text{HDRS}_{ij} = \beta_0 (D_1)_{ij} + \beta_1 (D_2)_{ij} + \ldots + \beta_9 (D_{10})_{ij} + \beta_{10} (\text{WEEK})_{ij} + e_{ij}.$$

This model allows for the estimation of individual intercept coefficients, but a common slope parameter β_{10}. Table 1.5 contains a summary of the results of this analysis. Although this model is a compromise between the models for pooled data and separate models for individuals' data, the number of parameters to be estimated is proportional to the number of patients and does not allow for the estimation of individual slopes. These issues have led researchers over time to develop mixed-effects models.

Table 1.5: Results of random intercept model

Variable	Parameter estimate	Standard error	t-value
D1	20.30101	1.91911	10.58
D2	27.63434	1.91911	14.40
D3	23.80101	1.91911	12.40
D4	21.63434	1.91911	11.27
D5	26.97414	1.99735	13.50
D6	22.73155	2.02303	11.24
D7	20.13434	1.91911	10.49
D8	27.37414	1.99735	13.71
D9	19.36121	2.14567	9.02
D10	23.36121	2.14567	10.89
Week	−2.78707	0.35057	−7.95

1.4 Random intercept model

From the results of the previous models, we concluded that it is not reasonable to assume that the initial status of patients, or their change in HDRS ratings over time, can be described adequately by average intercept and slope estimates while the clustering of measurements within individual patients was ignored. While the second of these analyses, where fixed-effects regression lines were fitted for each patient and thus the clustering of measurements was acknowledged, provided better information per patient, neither of these models allow us to obtain average intercept or slope coefficients while simultaneously incorporating the effect of measurements nested within individuals. To study differences in the response of patients over time, while acknowledging the clustering of measurements and allowing for patient differences in initial status, a random-effects model is needed. From the results obtained thus far, we will have to accommodate not only differences in initial status between patients, but also differences in the slopes of the improvement over time.

We start by specifying a model that takes clustering of measurements within patients into account, while allowing the initial status to vary from patient to patient. This model, a so-called random intercept model, contains both fixed and random effects, and can be expressed as

$$y_{ij} = \beta_0 + \beta_1 (\text{WEEK})_{ij} + u_{i0} + e_{ij} \qquad (1.2)$$

where y_{ij} denotes the HDRS measurement at time j ($j = 0, 1, 2, 3, 4,$ or 5) for patient i, $(\text{WEEK})_{ij}$ the associated time of measurement, and e_{ij} measurement error. The coefficients β_0 and β_1 are the fixed, but unknown, parameters to be estimated. The coefficient u_{i0}, in contrast, denotes a random parameter, and represents the amount by which the intercept of patient i differs from the average (fixed) intercept for all patients, as represented by β_0. By including u_{i0}, we allow intercepts to vary randomly over the patients. We assume that u_{i0} is normally distributed with mean 0 and variance $\phi_{(2)}$ and that the e_{ij} too, as in the first model, have a normal

distribution with mean 0 and variance σ^2 for all patients. In contrast to the model in (1.1), where all unexplained variation in HDRS ratings were captured by e_{ij}, the current model assumes that there are two potential sources of unexplained variation: variation between measurements as represented by e_{ij}, and variation between patients in terms of their intercepts, as represented by u_{i0}. Viewing the measurements as the lowest level of a nested structure in our data, with measurements nested within patients, we refer to σ^2 as the level-1 (measurement level) variance and to $\phi_{(2)}$ as the level-2 (patient level) variance.
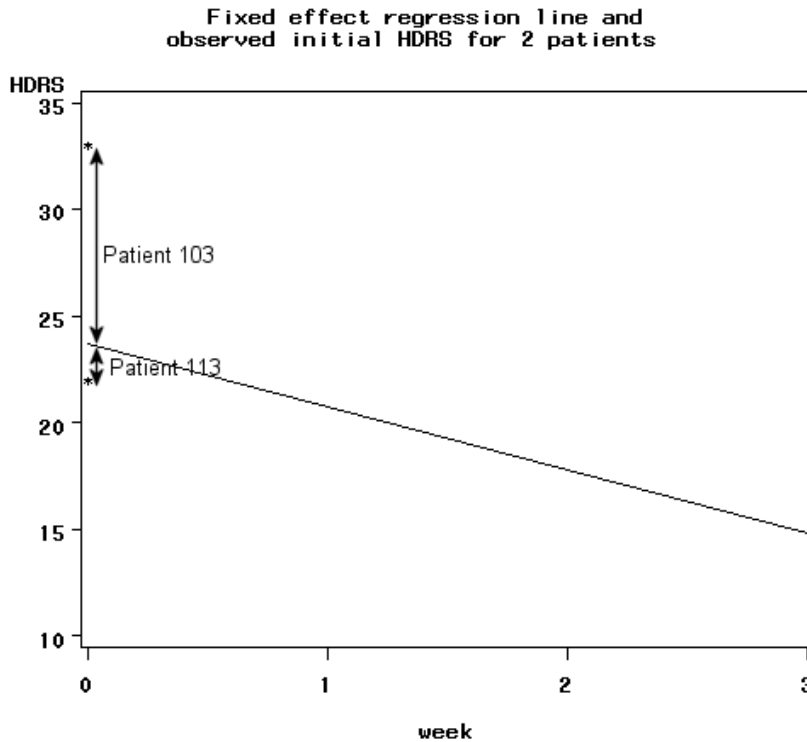
Figure 1.3: Comparison of observed initial HDRS ratings to average regression line

Combining the single regression line previously shown in Figure 1.1 with the results for the fixed-effects regression lines fitted to the data from patients 103 and 113, we can produce an approximate graph (see Figure 1.3) of what we hope the inclusion of β_{i0} in the model will add to our model. Recall from Table 1.2 that patient 103 had an observed HDRS rating of 33 at the start of the study, while patient 113 had an initial observed HDRS rating of 21. Suppose the solid line represents a fixed-effects regression model over all patients, taking clustering of measurements within patients into account. The observed initial HDRS ratings of patients 103 and 113 are indicated by star symbols, with that of patient 103 being considerably higher than that of patient 113. The distances between these and the solid line represents the differences between the individual patient intercepts and the average intercept over all patients for these two patients.

The coefficient $u_{PATIENT103,0}$, representing the unique increment to the intercept associated with patient 103 in our model, is depicted as the arrow with text "Patient 103" next to it. If our proposed model depicted here as the solid line provides a good fit to the data, we would expect $u_{PATIENT103,0}$ to be positive, indicating that the HDRS rating of patient 113 is higher than the average initial HDRS rating. In addition, the size of the estimate of $u_{PATIENT103,0}$ would provide information on the extent to which the initial status for this patient is predicted to be above the average. If the proposed model fits the data available for this patient over the course of the study well, we would expect the estimate to be reasonably close to the observed initial HDRS rating. Likewise, the coefficient $u_{PATIENT113,0}$ should, under the assumption of model fitting the data, indicate that the initial status of this patient is below average.

Estimates of the coefficients of the model given in (1.2) are given in Table 1.6. The average initial HDRS rating over all 10 patients is estimated at 23.4961, and the amount by which the HDRS rating is expected to decrease over the course of any week by –2.8625. Recall that in the first model, where the effect of clustering was ignored and no provision was made for differences between the initial status of patients, the corresponding estimates were 23.746 and –2.978 respectively. Taking clustering into account and allowing for differences in intercepts have led to a

slightly lower (in absolute value) estimates of the initial status and slope coefficients.

Table 1.6: Results of random intercept model

Coefficient	Estimate
Fixed effects:	
Intercept (β_0)	23.4961
Slope(β_1)	–2.8625
Random effects:	
Level-2 variation ($\text{var}(u_{i0})$)	5.4443
Level-1 variation ($\text{var}(e_{ij})$)	17.1203

For the first model, an estimate of σ^2 of 23.3723 was obtained. For the current model, we find that the variation in HDRS ratings between measurements is estimated at 19.0375. Allowing for random intercepts, and thus allowing for a separate coefficient to reflect the variation in initial HDRS ratings over patients, has led to some reduction in the residual, or level-1, variation. In effect, the unexplained variation in HDRS ratings is now split according to the levels of the hierarchy. Comparing the estimates of $\text{var}(u_{i0})$ and $\text{var}(e_{ij})$, we conclude that most of the remaining variation in the HDRS ratings is still at the measurement level. In addition, we conclude that it is not reasonable to assume that the initial status of patients can be described by a single, fixed coefficient β_0.

In addition to these estimates, which describe the average estimated intercept and slope over all patients, we also obtain estimates for the unique deviations from the intercept associated with each of the individual patients. The estimates of the deviations, also known as empirical Bayes residuals, are given in Table 1.7.

INTRODUCTION

Table 1.7: Estimated Bayes residuals

Patient	$\hat{u}_{i0}$
101	−1.9727
103	2.8389
104	0.32373
105	−1.0979
106	2.2277
107	−0.36753
108	−2.0821
113	2.4733
114	−2.3996
115	0.056051

Recall from the discussion of Figure 1.3 that a high initial HDRS rating of 33 was observed for patient 103. Under the assumption of good model fit and dependable data for this patient, we expected that the estimated deviation would be positive and relatively large. This is confirmed by the values shown in Table 1.7: $\hat{u}_{PATIENT103,0}$ is both positive and larger than any of the other $\hat{u}_{i0}$ estimates. On the other hand, $\hat{u}_{PATIENT113,0}$ is not negative as expected − on the contrary, the predicted initial status for this patient is positive, and substantially higher than the average predicted average status over all patients of 23.4961. This may be due to a smaller number of measurements for this patient, or data of lesser quality than for other patients.

By adding the $\hat{u}_{i0}$ estimates to the average estimated initial status $\hat{\beta}_0$, the actual predicted initial status may be obtained. In Figure 1.4, the predicted initial status of each patient is indicated by a "B," while the observed initial status is indicated by the letter "O." A reference line on the vertical axis, corresponding to the average initial status β_0 of 23.4961, allows us to distinguish between patients with either

observed or predicted initial HDRS rating higher than the average intercept and patients with ratings lower than the average estimated intercept.

Observed and predicted patient-specific initial status
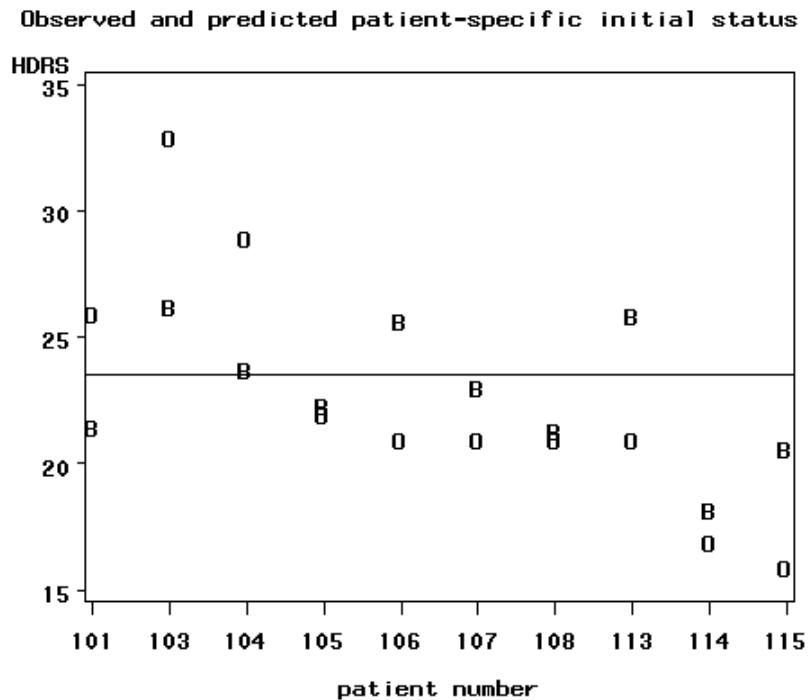
Figure 1.4: Comparison of observed and predicted initial HDRS ratings for 10 patients

While the observed and predicted initial status of patients 105, 107, 108 and 114 are very close to each other, the differences between the observed and predicted ratings for patients 103, 104, and 113 are relatively large. For patients 106 and 113, for example, observed initial status was lower than the average, but the predicted initial status is higher than the average.

Intraclass correlation

It is realistic to assume that patients in a clinical trial diagnosed with a specific type of cancer are more alike with respect to certain traits than patients who do not have

that type of cancer. This is often due to shared experiences, receiving the same type of treatment or medication, shared environment, etc. The sharing of the same context is a likely cause of dependency among observations.

The intraclass correlation is a measure of the degree of dependence of individuals. The more individuals share common experiences due to closeness in space and/or time, the more they are similar. A higher degree of dependency can, for example, be found between children born and raised in the same family. Another example of dependent observations is repeated measurements on the same person. It is realistic to assume that patients treated in the same service center are more alike with respect to certain traits than patients treated in different centers.

For data having a two-level hierarchical structure, the intraclass correlation ρ is defined as the proportion of the variance in the outcome variable that is between the second-level units:

$$\rho = \frac{\text{between group variability}}{\text{between group variability} + \text{within group variability}}$$

As pointed out by Kreft and de Leeuw (1998), if intraclass correlation is present, as is usually the case when we are dealing with clustered data, the assumption of independent observations in the traditional linear model is violated. They also pointed out that tests of significance lean heavily on the number of independent observations and that the existence of intraclass correlation makes the test of significance in traditional linear models too liberal. Barcikowski (1981) shows that in most applications of analysis of variance, the standard errors of the parameter estimates will be underestimated and that even a small intraclass correlation can inflate the alpha level substantially.

While the random intercept model has allowed us to accommodate some of our modeling concerns for an unbalanced data set such as the nesting of measurements within patients and allowing intercepts to vary over patients, other concerns remain.

From the results shown in Table 1.3, as depicted in Figure 1.2, we know that there is also potentially significant variation in the slopes of HDRS ratings over the study period. In addition, the models considered thus far made no provision for the inclusion of patient-level variables such as gender or their initial depression classification.

To address these concerns, extended models, possibly of the form

$$\text{HDRS}_{ij} = \beta_0 + \beta_1 * (\text{WEEK})_{ij} + \beta_2 * \text{ENDOG} + u_{i0} + u_{i1} (\text{WEEK})_{ij} + e_{ij},$$

are required. In the model shown here, an additional fixed effect β_2 represents the effect of the patient-level variable ENDOG, and an additional random coefficient u_{i1} allows for random variation in the WEEK slopes over patients. Examples of such models, based on the data of all 66 patients in the Reisby data instead of the 10 patients used here, are shown in detail in Section 3.2.

2 Graphical User Interface

The SuperMix graphical user interface (GUI) consists of a main window, a spreadsheet window, and a graph window. The main window is used to create or open SuperMix data files, whereas the spreadsheet window is used to display SuperMix data files and to allows access to the **Model Setup** window. The graph window is used to display SuperMix graph files. SuperMix data files have the default extension **.ss3** and are known as **ss3** or spreadsheet files, while SuperMix model files have the default extension **.mum**. SuperMix graph files have the default extension **.mug**. The main window and its menus and dialog boxes are reviewed in the next section, and the menus and dialogs of the spreadsheet and graph windows are reviewed in the sections to follow.

2.1 The main window

The SuperMix main window is accessed when you start the program. SuperMix can be opened from the **Programs** option on the Windows **Start** menu, by double-clicking on the SuperMix application or by clicking on a desktop shortcut for SuperMix. Any of these actions opens the following main window.
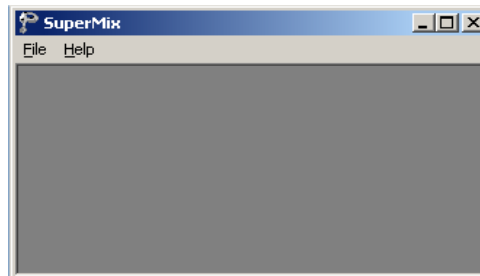
The SuperMix main window consists of a **File** menu and a **Help** menu. These menus are reviewed separately in the following two sections.

2.1.1 The File menu

The options on the **File** menu of the SuperMix main window provide access to a sequence of three dialog boxes that can be used to create or to open a SuperMix data file in a spreadsheet format.
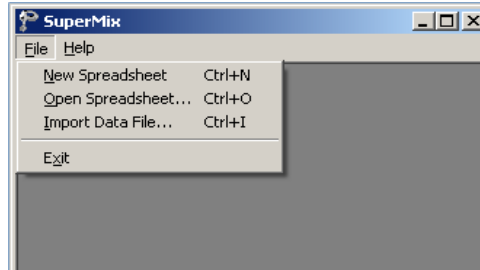
The New Spreadsheet option

Click on the **New Spreadsheet** option to open an empty SuperMix spreadsheet window.
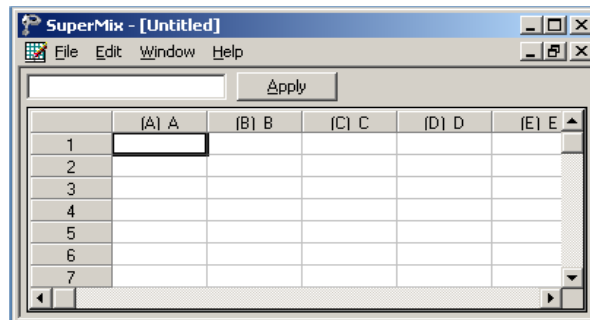
You can use the window above to enter data manually. Use the **Save As** option on the **File** menu to save the data to an **ss3** file. Alternatively, data can be imported into the empty spreadsheet via the **File**, **Import Data File** option.

The Open Spreadsheet option

The **Open** option on the **File** menu is used to open an existing SuperMix data file. Click on the **Open Spreadsheet** option to open the following **Open Spreadsheet** dialog box.
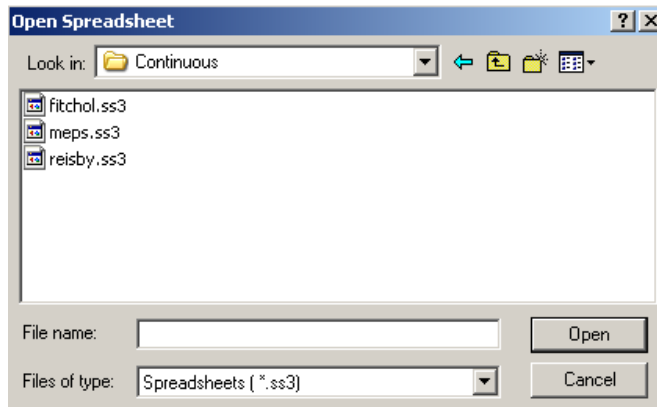
Next, browse for the **ss3** file, select it, and click on the **Open** button to open the SuperMix spreadsheet window.

The Import Data File option

Use the **Import Data File** option on the **File** menu to convert the data in a Microsoft Excel workbook (***.xls**), data files from statistical packages, or a comma delimited text file (***.csv**, ***.dat** or ***.txt**) or Bayes estimates file (***.bay2** or ***.bay3**) produced by SuperMix to a SuperMix data file. Click on the **Import Data File** option to open the following **Open a file to import** dialog box. Note that the (***.bay2** and ***.bay3** files are only created when the **Write Bayes Estimates** option is selected on the **Model Setup**, **Configuration** tab (see Section 2.4.1.)
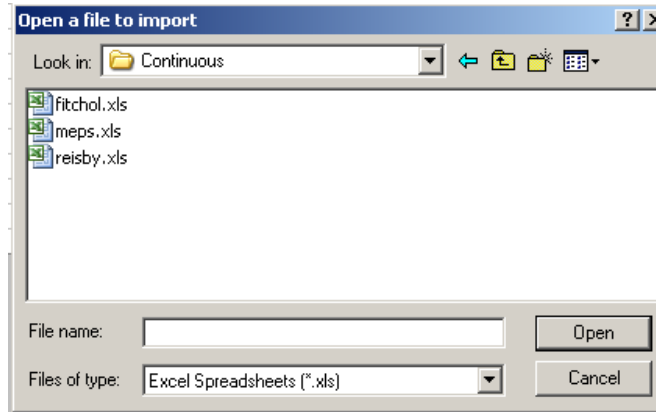
Next, browse for the Microsoft Excel workbook or the text file and select it. Click on the **Open** button to open the **Save As** dialog box. Enter a name for the **ss3** file and click on the **Save** button to open the SuperMix data file in a spreadsheet window.

The Exit option

Close the SuperMix main window by clicking on the **Exit** option on the **File** menu.

2.1.2 The Help menu

The options on the **Help** menu on the SuperMix main window provide access to the contents of the SuperMix online help file, the SuperMix user's guide, the SuperMix website, technical support and other information.

2.2 The spreadsheet window

The SuperMix spreadsheet window is used to display a new or existing SuperMix data file. The menus on the spreadsheet window can be used to manipulate the data entries in an existing SuperMix data file. It is also used to access the **Model Setup** window, which is used to specify a mixed-effects model and to edit existing

SuperMix model files. These menus can also be used to create new or open existing SuperMix graph files. In Section 2.5 some basic spreadsheet operations are illustrated. In the sections to follow, we review the four menus of the SuperMix spreadsheet window.

2.2.1 The File menu

The options on the **File** menu of the spreadsheet window are used to open a new SuperMix project, open an existing **ss3** file, create a new SuperMix model (.**mum**) file, or edit an existing model file. It is also used to create or edit a SuperMix graph file (**Exploratory…** option only). An example of the **File** menu is shown below.
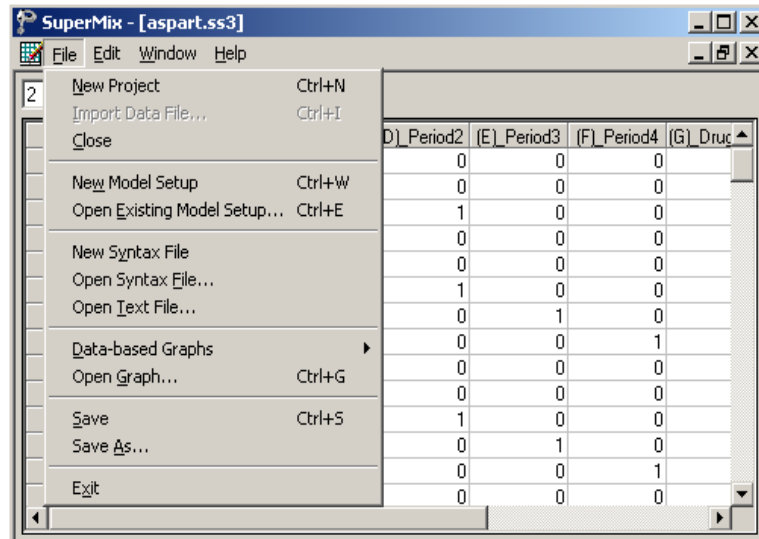
When an **ss3** file as well as a SuperMix model file are opened, the **File** menu changes as shown in the following window.
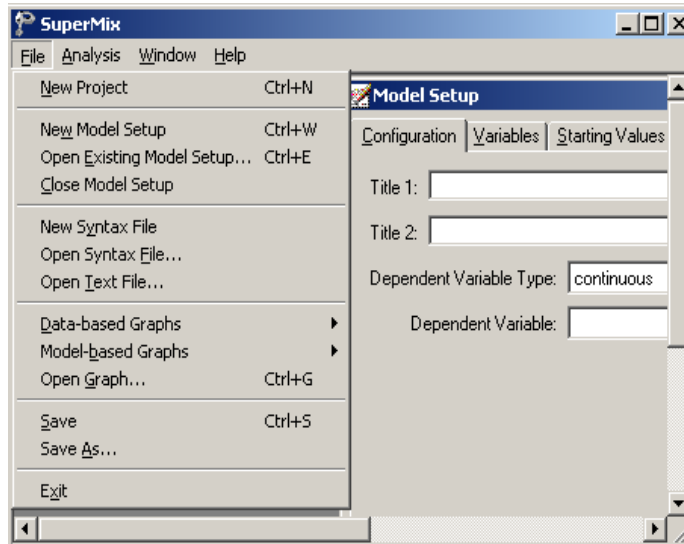
The New Project option
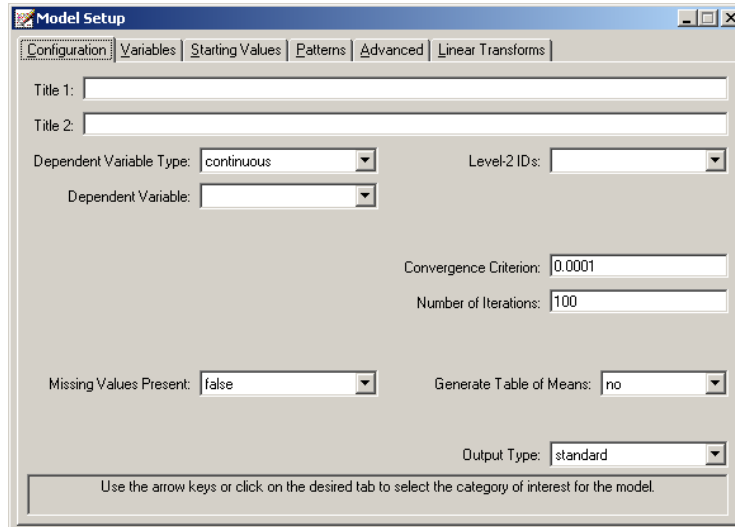
The **New Project** option is used to open an independent SuperMix main window.

The Exit option

The **Exit** option is used to close the current open SuperMix main window.

The New Model Setup option

The **New Model Setup** option of the spreadsheet window provides access to the **Configuration**, **Variables**, **Starting Values**, **Patterns**, **Advanced** and **Linear Transforms** screens of the **Model Setup** window shown below. Each screen is opened by clicking on the corresponding tab.

These screens are used to specify a mixed-effects model to be fitted to the data in the open spreadsheet window. The appearance of the screens depends on the type of outcome (dependent) variable (continuous, count, ordered, or nominal) that is selected on the **Configuration** screen shown above. A detailed description of each of these screens is given in Section 2.4. Once a model is defined, it can be saved as a **.mum** file.

The Open Existing Model Setup option

The **Open Existing Model Setup** option is used to open the **Model Setup** window of an existing SuperMix model file. This is accomplished by clicking on the **Open Existing Model Setup** option, which opens the following **Open Mixed Up Model** dialog box.

Browse for the desired SuperMix model file, select it, and click on the **Open** button to open the **Model Setup** window for the selected SuperMix model file.
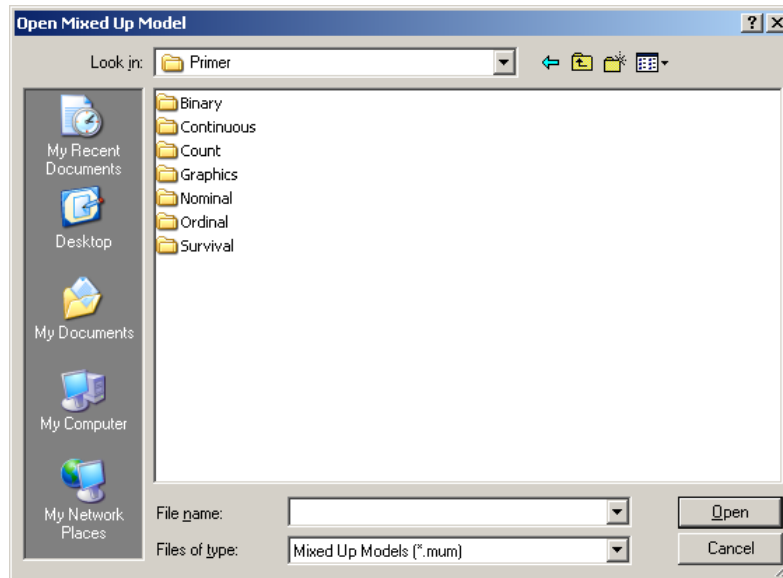
The Close Model Setup option

The **Close Model Setup** option is used to close any SuperMix **Model Setup** dialog box that is currently open.

The New Syntax File option

The **New Syntax File** option is used to open a blank syntax window.

GRAPHICAL USER INTERFACE

The Open Syntax File option

The **Open Syntax File** option is used to open an existing syntax file (**.inp**). This is accomplished by clicking on the **Open Syntax File** option, which leads to the display of the **Open Syntax File** dialog box.

Browse for the desired SuperMix syntax file, select it, and click on the **Open** button to open the syntax window for the selected SuperMix model file.
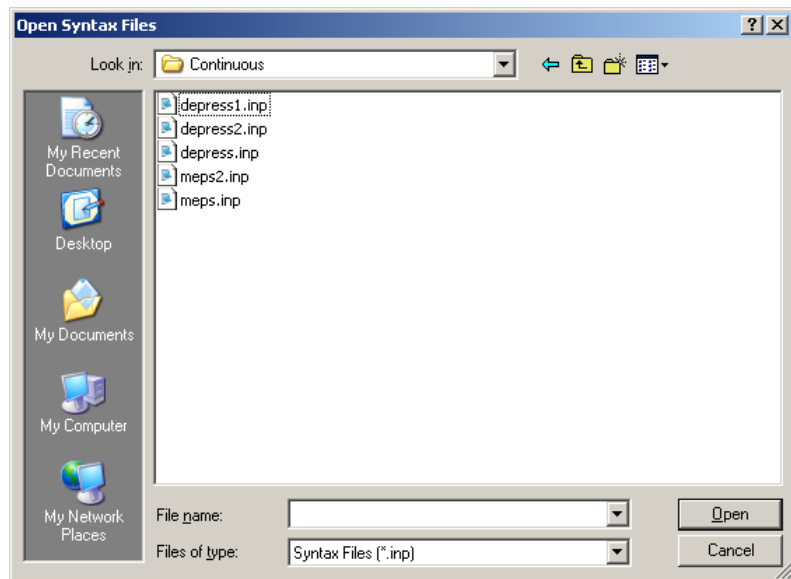
The Open Text File option

The **Open Text File** option is used to open any existing text file.

The Data-based Graphs pop-up menu

The **Data-based Graphs** pop-up menu is used to create a new SuperMix graph from the data displayed in the open **ss3** file in a SuperMix graph window. The menus and dialogs of the SuperMix graph window for new and existing SuperMix graphs are reviewed in Chapter 4.
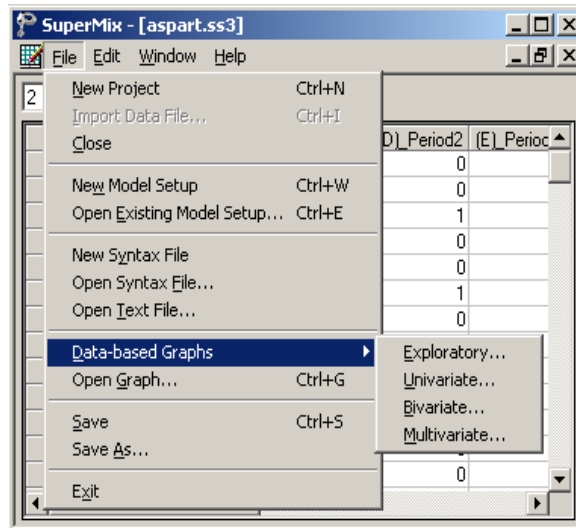
The options listed below are discussed in detail in Chapter 4, which contains examples of all the plots that SuperMix can produce.

- o The **Exploratory** option is used to produce single or overlay color-coded Y against X plots. Groups of plots are obtained by using a filter variable. A typical example is given in Section 3.2.4, and other examples can be found in Section 4.2.

- o The **Univariate** option on the **Data-based Graphs** pop-up menu is used to create a bar chart, a pie chart or a histogram for the data displayed in the spreadsheet window.

- o The **Bivariate** option on the **Data-based Graphs** pop-up menu is used to create a scatter plot, a line plot, a combination line and scatter plot, a box-and-whisker plot, or a 3-dimensional bar chart for the data in the open SuperMix data file.
- o The **Multivariate** option on the **Data-based Graphs** pop-up menu is used to make a matrix scatter plot based on the data in the open **ss3** file. This provides an organized way of simultaneously looking at a set of bivariate plots.

The Model-based Graphs pop-up menu

The options on the **Model-based Graphs** pop-up menu are used to create a new SuperMix graph from the data displayed in the open spreadsheet window. The menus and dialogs of the SuperMix graph window for new and existing SuperMix graphs are reviewed in Chapter 4.
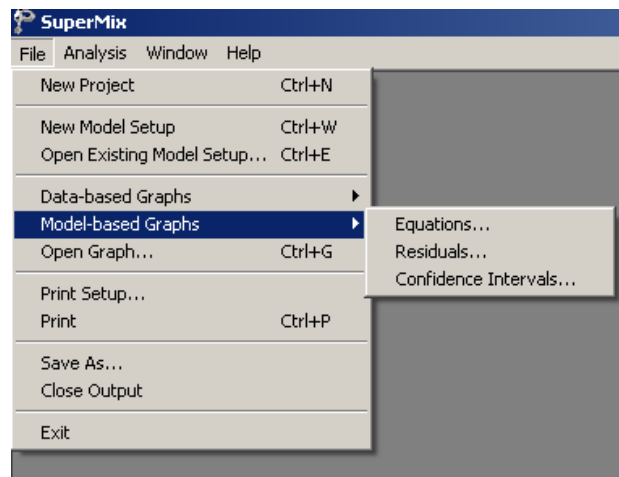
Available options are:

- o The **Equations** option on the **Model-based Graphs** pop-up menu opens the **Plot Equations for** dialog box which can be used to plot model equations of an outcome variable for given values of the predictors in the model.
- o The **Residuals** option on the **Model-based Graphs** pop-up menu provides access to the **Plot of Residuals** dialog box, which is used to create a residual plot for the residuals based on the current SuperMix analysis.
- o The **Confidence Intervals** option is used to open the **95% C.I. for Level-1 Variables** dialog box, which is used to create confidence interval plots.

The Open Graph option

The **Open Graph** option is used to open an existing SuperMix exploratory graphics file with a default extension **.mug**. Start by clicking on the **Open Graph** option to open the following **Open Mixed Up Graph** dialog box.

GRAPHICAL USER INTERFACE
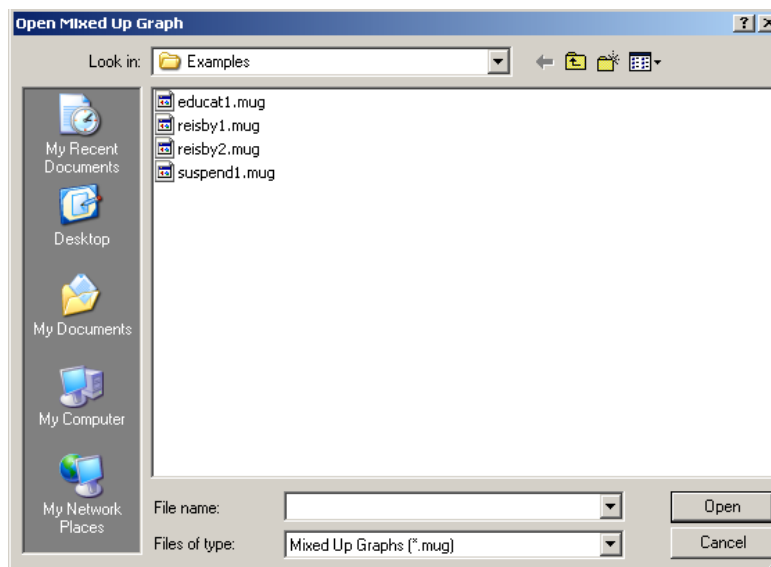
Browse for the desired SuperMix graph file, select it, and click on the **Open** button to open the graph window for the selected SuperMix graph file.

The Save option

The **Save** option on the **File** menu is used to save any changes made to the data in the open SuperMix data file. Please note that any change to the data will not be saved to file unless you use this option or the **Save As** option.

The Save As option

The **Save As** option on the **File** menu is used to save the opened **ss3** file as another SuperMix data file. Select the **Save As** option to open the following **Save As Spreadsheet Data** dialog box.
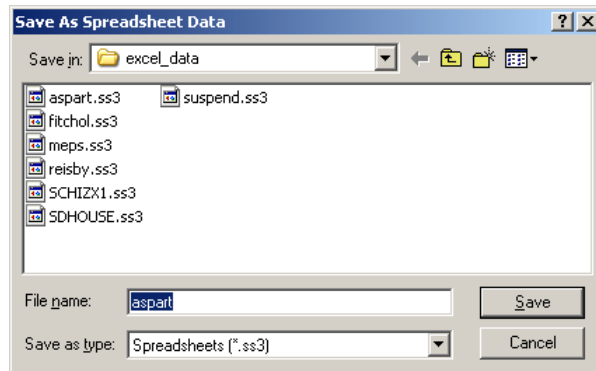
Enter the file name in the **File name** string field and click on the **Save** button to save the SuperMix data file.

2.2.2 The Edit menu

The options on the **Edit** menu of the SuperMix spreadsheet window are used to edit the data entries of the open SuperMix data file. To use these options, select the data to be edited (cell(s), row(s) or columns(s)). Then click on the **Edit** menu to produce the window shown below.

The next step is to select one of the options available, which have the following effects on the selected data.

- o The **Cut** option cuts the data selection from the spreadsheet window and places it into the Windows clipboard.
- o The **Copy** option places the data selection in the Windows clipboard.
- o The **Paste** option pastes data from the Windows clipboard into the selected area of the spreadsheet window.
- o The **Paste (value only)** option pastes only the actual values (ignoring the formats) of the data from the Windows clipboard into the selected area of the spreadsheet window.
- o The **Clear** option replaces the selected data with empty cell(s). Choosing this option activates the following drop-down menu.

- o The **Clear All** option deletes the values and the formulas of the selected data.
- o The **Clear Data** option deletes the values of the data selection, but leaves the corresponding formulas intact.
- o The **Clear Formula** option deletes the formulas of the selected data, but not the corresponding values.
- o The **Create Header from Row** creates spreadsheet headers that correspond to the labels in the selected row.
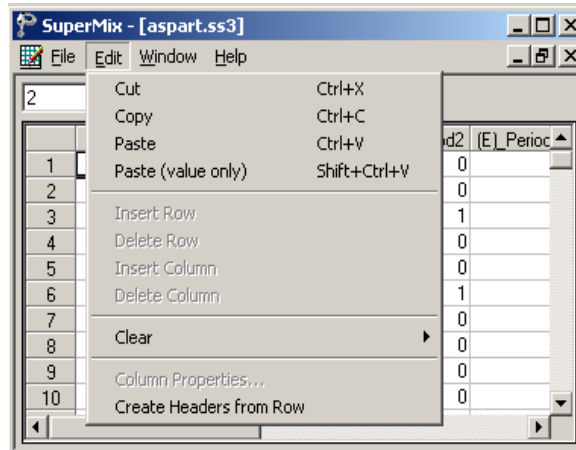
2.2.3 The Window menu

The **Window** menu is used to toggle between open SuperMix files.

2.2.4 The Help menu

The **Help** menu of the spreadsheet window is identical to that of the main window and is reviewed in Section 2.1.2.

2.3   The graph window

The SuperMix graph window is opened by creating a new SuperMix graph or by opening an existing SuperMix graph file. We accomplish this by using the **Exploratory** option on the **Data-based Graphs** pop-up menu or the **Open Graph** option on the **File** menu of the spreadsheet window reviewed in Section 2.2.1. The menus and dialogs of the SuperMix graph window are reviewed and illustrated in Chapter 4.

2.4   The Model Setup window

A SuperMix **mum** file (model setup file) is always associated with an **ss3** file (data spreadsheet file). This ensures that variable selections are maintained correctly in the **mum** file, regardless of changes to the header text and cut/paste/move operations on the columns of the **ss3** file. For this reason, the **Model Setup** window is accessed via the **File** menu of the spreadsheet window. This is done by selecting the **New Model Setup** or **Open Existing Model Setup** options. The **Model Setup** window has six tabs. By clicking on a tab, the corresponding **Configuration**, **Variables**, **Starting Values**, **Patterns**, **Advanced**, or **Linear Transforms** screen is accessed. The appearance of a screen depends on the type of outcome variable selected. The purpose of a field is displayed at the bottom of the screen when the field is clicked. Tables 2.1 to 2.15 contain summaries of these descriptions. The options that are selected in these are shown in bold typeface.

2.4.1 The Configuration screen

The **Configuration** screen is used to provide a title for the analysis, to select the type and name of the outcome (dependent) variable, and to indicate identifiers of the level-2 and level-3 units. Additionally, it contains options that control the amount of information to be saved to file and the parameters of the optimization procedure. When the **New Model Setup** or **Open Existing Model Setup** options on the **File** menu are used, the **Configuration** screen is, by default, the first screen displayed.

The same **Configuration** screen is used for continuous and count outcomes, but its contents change when the dependent variable type is ordinal or nominal. The screen is the same for ordinal and nominal outcome types. The two cases are discussed separately below.

Configuration screen for continuous and count outcomes

An example of the **Configuration** screen of the **Model Setup** window for a continuous response variable is shown below. The layout is identical when the

dependent variable type is changed from **continuous** to **count** (see Section 3.4 for an example based on a count outcome variable).

The 15 possible entries on the **Configuration** screen of the **Model Setup** window for continuous or count response variables are summarized in Table 2.1.

Table 2.1: Entries on the Configuration screen of the Model Setup window for continuous and count outcomes

Number	Caption	Purpose	Type	Action	Options
1	Title 1	To specify the first line of the title to be listed in the output file.	Text box	Enter a string of not more than 60 characters.	
2	Title 2	To specify the second line of the title to be listed in the output file.	Text box	Enter a string of not more than 60 characters.	
3	Dependent Variable Type	To specify the variable type for the response variable.	Drop-down list box	Select an option from the drop-down list box.	**continuous** (default) ordered nominal count Binary
4	Level-2 IDs	To specify the variable that defines the second level of the hierarchy in the data.	Drop-down list box	Select a variable from the drop-down list box.	
5	Dependent Variable	To specify the response variable of the model.	Drop-down list box	Select a variable from the drop-down list box.	
6	Level-3 IDs	To specify the variable that defines the 3rd level of the hierarchy in the data.	Drop-down list box	Select a variable from the drop-down list box.	
7	Write Bayes Estimates	To request a text file for the Bayes estimates.	Drop-down list box	Select an option from the drop-down list box.	**no** (default) means only means & (co)variances

Table 2.1: Entries on the Configuration screen of the Model Setup window for continuous and count outcomes (continued)

Number	Caption	Purpose	Type	Action	Options
8	Convergence Criterion	To specify the convergence criterion for the iterative algorithm.	Text box	Enter a non-zero positive real number if the default of 0.0001 is not desired.	
9	Number of Iterations	To specify the maximum number of iterations for the iterative algorithm.	Text box	Enter a positive integer if the default of 100 is not desired.	
10	Missing Values Present	To specify the missing value status of the data.	Drop-down list box	Select an option from the drop-down list box.	false (default) **true**
11	Generate Table of Means	To request the printing of a table as part of the output.	Drop-down list box	Select an option from the drop-down list box.	no (default) **yes**
12	Means Variable	To specify the variable for which the tables should be created (see 8).	Drop-down list box	Select a variable from the drop-down list box.	
13	Missing Value for the Dependent Var	To specify the missing value code for the response variable.	Text box	Enter a real number.	
14	Global Missing Value	To specify the global missing value code.	Text box	Enter a real number.	
15	Output Type	To request different type of output.	Drop-down list box	Select an option from the drop-down list box.	**standard** (default) iterative details simulation information

Configuration screen for ordered, nominal and binary outcomes

The following screen is an example of the **Configuration** screen of the **Model Setup** window in the case of an ordered response variable. An example of this screen for a nominal outcome variable is given in Section 3.5.

As shown in the image below, the 5 entries shown in bold typeface are either new or different compared with those on the **Configuration** screen of the **Model Setup** window for continuous or count outcome variables. These 5 entries are summarized in Table 2.2. Please refer to Table 2.1 for information about all the other entries.

Table 2.2: Entries of the configuration screen for ordered, nominal and binary outcomes

Number	Caption	Purpose	Type	Action	Options
3	Dependent Variable Type	To specify the variable type for the response variable.	Drop-down list box	Select an option from the drop-down list box.	continuous (default)
					ordered
					nominal
					count
5	Dependent Variable	To specify the response variable of the model.	Drop-down list box	Select a variable from the drop-down list box.	
16	Categories	To show the value of each category of the ordered dependent variable selected in 5.	Grid box		
17	Perform Crosstabulation	To specify a crosstabulation of selected variable by the outcome variable.	Drop-down list box	Select an option from the drop-down list box.	no (default)
					yes
18	Crosstab Variable	To specify the variable to be crosstabulated with the outcome variable (see 8).	Drop-down list box	Select a variable from the drop-down list box.	

2.4.2 The Variables screen

This screen has the same appearance for all outcome types except the ordered outcome. In the case of the ordinal outcome, there is no check box available to include the intercept as an explanatory variable. This screen is used to select explanatory variables and random effects. The unknown model parameters are the coefficients of the explanatory variables and the variances and covariances of the random effects. The appearance of the **Variables** screen depends on the number of levels of the model. For a two-level model, the **3** columns in the **Available** grid and the **L-3** grid will be hidden. By default, an intercept term is included in the fixed part (explanatory variables) and in the random part (random effects) of the model.

Variables screen for continuous, count, ordered, nominal and binary outcomes

The following screen is an example of the **Variables** screen of the **Model Setup** window that is used for variable selection for continuous, count, nominal, and binary response variables. The 9 possible entries of the **Variables** screen of the **Model Setup** window for continuous, count, nominal or binary response variables are summarized in Table 2.3. Note that the **Include Intercept** option for the **Explanatory Variables** is not available for the ordered outcome. The other 8 entries are the same for the ordered outcome.
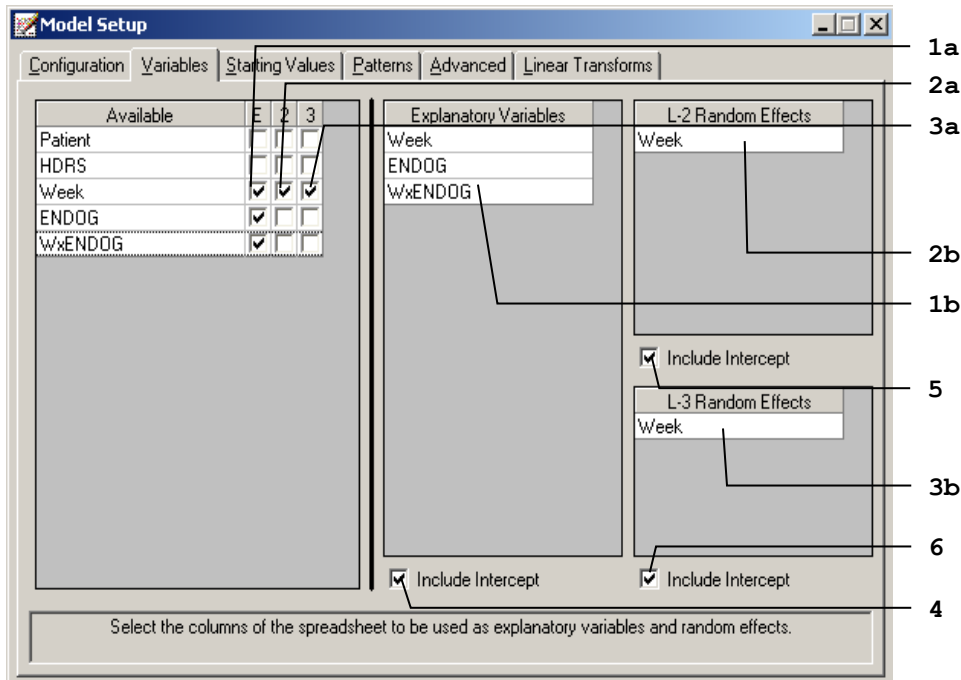
Table 2.3: Entries of the Variables screen

Number		Caption	Purpose	Type	Action	Options
1	a	E	To specify the explanatory variable(s) of the model.	Column of check box(es)	Check the E column(s) of the variable(s).	
	b	Explanatory Variables	Displays the variable(s) selected in 1a.	Grid box		
2	a	2	To specify the level-2 random effects of the model.	Column of check box(es)	Check the 2 column(s) of the variable(s).	
	b	L-2 Random Effects	Displays the variable(s) selected in 2a.	Grid box		
3	a	3	To specify the level-3 random effects of the model.	Column of check box(es)	Check the 3 column(s) of the variable(s).	
	b	L-3 Random Effects	Displays the variable(s) selected in 3a.	Grid box		
4		Include Intercept	To specify an intercept term for the fixed part of the model.	Check box	Uncheck the check box if an intercept is not desired.	**Check (default)**
						Uncheck
5		Include Intercept	To specify a random intercept at level-2 of the model.	Check box	Uncheck the check box if a level-2 random intercept is not desired.	**Check (default)**
6		Include Intercept	To specify a random intercept at level-3 of the model.	Check box	Uncheck the check box if a level-3 random intercept is not desired.	**Check (default)**
						Uncheck

2.4.3 The Starting Values screen

The unknown parameters in a mixed-effects model cannot, in general, be obtained as a closed-form expression. To estimate these parameters, use is made of an iterative procedure based on the method of maximum likelihood. For count, ordinal, and nominal outcomes, the likelihood function is approximated by numerical integration. For more than one random effect, this procedure is computationally intensive. All iterative procedures start with initial estimates of the values of the unknown parameters and, at each iteration, the algorithm attempts to improve this estimate until convergence is obtained. The closer these initial estimates (the starting values) are to the maximum likelihood solution, the fewer iterations are needed to obtain convergence and reach the final solution.

SuperMix automatically generates starting values for the model parameters and typically these values are sufficient to ensure convergence. There may, however, be cases where a model with many parameters takes a long time to run, and if small modifications are made to such a model, one can use the parameter estimates from the previous analysis as starting values for the next analysis. Alternatively, one may want to fix some of the parameter values at specific values, for example, the slope coefficient of variable X at 0.1. This can be accomplished by selecting the user-defined option and entering this value for X. Note that the value of 0.1 will remain fixed during the optimization procedure if it is specified as fixed for X on the **Patterns** screen discussed in Section 2.4.4.

Starting Values screen for continuous and count outcomes

An example of the **Starting Values** screen of the **Model Setup** window for a continuous or count response variable is shown below.
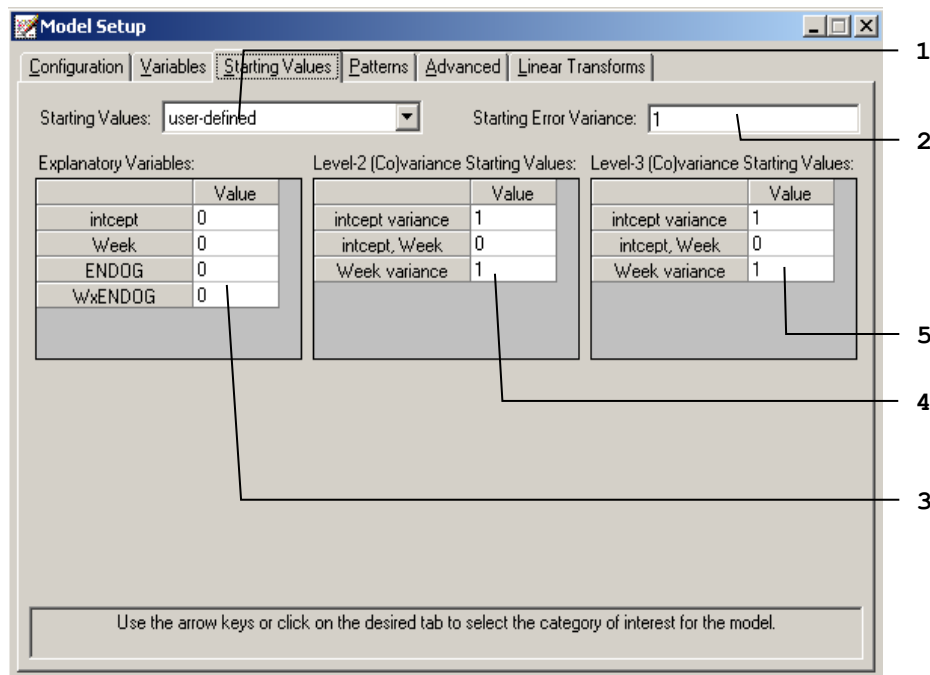
The 5 possible entries of the **Starting Values** screen of the **Model Setup** window for count or nominal response variables are summarized in Table 2.4.

Table 2.4: Entries of the Starting Values screen for continuous and count outcomes

Number	Caption	Purpose	Type	Action	Options
1	Starting Values	To specify the type of starting values to be used.	Drop-down list box	Select an option from the drop-down list box.	automatic (default)
					user-defined
2	Starting Error Variance	To specify the starting error variances.	Text box	Enter a real number (>0) if the default of 1 is not desired.	

Table 2.4: Entries of the Starting Values screen for continuous and count outcomes (continued)

Number	Caption	Purpose	Type	Action	Options
3	Explanatory Variables	To specify the starting value(s) for the coefficients of explanatory variable(s) of the fixed part of the model.	Grid box	Enter a real number in the corresponding Value box(es) of the variable(s) of interest.	
4	Level 2 (Co)variances Starting Values	To specify the starting value(s) for the variance(s) and/or covariance(s) of the level-2 random effects.	Grid box	Enter a real number (positive for variances) in the corresponding Value box(es) of the variable(s) of interest.	
5	Level 3 (Co)variances Starting Values	To specify the starting value(s) for the variance(s) or covariance(s) of the level-3 random effects.	Grid box	Enter a real number (positive for variances) in the corresponding Value box(es) of the variable(s) of interest.	

Starting Values screen for ordered outcomes

For ordinal outcomes, additional grid boxes appear to allow for user-specified starting values of thresholds and threshold-explanatory variable(s) interaction(s). The following screen is an example of the **Starting Values** screen of the **Model Setup** window.

As shown in the above image, the single entry shown in bold typeface is different from those of the **Starting Values** screen of the **Model Setup** window for continuous or count outcome variables. This entry is described in Table 2.5. Please refer to Table 2.4 for information about all the other entries.

Table 2.5: Entry of the Starting Values screen for ordered outcomes

Number	Caption	Purpose	Type	Action
6	Starting Values for Thresholds	Enter the starting values for the thresholds.	Grid box	Enter real numbers. The values must be monotonically increasing.

Starting Values screen for nominal and binary outcomes

When the nominal or binary outcome is selected, the grid boxes appear differently with a slide bar as shown below.
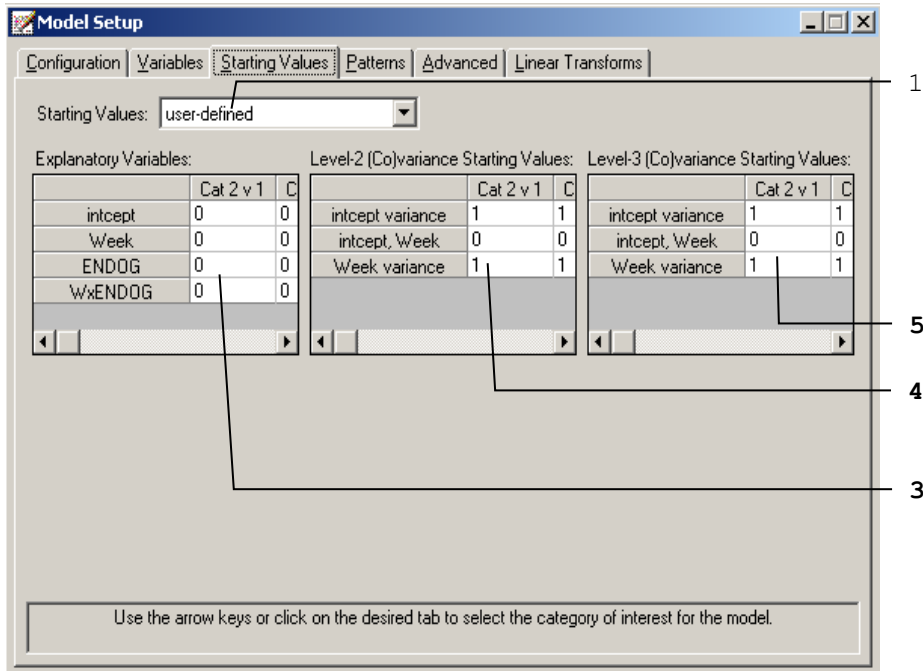
The 3 different entries shown in bold typeface are either new or different compared with those on the **Starting Values** screen of the **Model Setup** window for continuous or count outcome variables. These 3 entries are summarized in Table 2.6. Please refer to Table 2.4 for information about all the other entries.

Table 2.6: Entries of the Starting Values screen for nominal outcomes

Number	Caption	Purpose	Type	Action
3	Explanatory Variables	To specify the starting value(s) for the coefficients of explanatory variable(s) of the fixed part of the model.	Grid box with slide bar	Enter a real number in the corresponding Value box(es) of the variable(s) of interest.
4	Level 2 (Co)variances Starting Values	To specify the starting value(s) for the variance(s) and/or covariance(s) of the level-2 random effects.	Grid box with slide bar	Enter a real number (positive for variances) in the corresponding Value box(es) of the variable(s) of interest.
5	Level 3 (Co)variances Starting Values	To specify the starting value(s) for the variance(s) or covariance(s) of the level-3 random effects.	Grid box with slide bar	Enter a real number (positive for variances) in the corresponding Value box(es) of the variable(s) of interest.

2.4.4 The Patterns screen

This screen is used to specify patterns or structures for the coefficients of the explanatory variables and variances and covariances of the random effects. A typical **Patterns** screen is shown below Table 2.7. Note that the default numbers for these patterns (1, 2, 3, …) are dependent on the number of parameters listed in a grid. The default numbers indicate that all parameters are set free. On the other hand, if a number is replaced by a '0', the corresponding parameter is fixed to the default or user-specified value on the **Starting Values** screen.

Patterns screen for continuous, ordered, nominal, count and binary outcomes

The 6 different entries of the **Patterns** screen of the **Model Setup** window for continuous, ordered, nominal, count or binary response variables are summarized in Table 2.7. For an ordinal outcome variable, provision is also made for entering user-defined values for threshold parameters.

Table 2.7: Entries of the Patterns screen for continuous, count and nominal outcomes

Number		Caption	Purpose	Type	Action	Options
1	a	Explanatory Variables	To specify the pattern type for the coefficients in the fixed part of the model.	Drop-down list box	Select an option from the drop-down list box.	free (default)
						user-defined
	b		To specify the pattern for the covariance matrix of the fixed part of the model.	Text box	Enter integer values >= 0	
2	a	Level-2 (Co)variance Patterns	To specify the pattern type for the covariance matrix of the level-2 random effects.	Drop-down list box	Select an option from the drop-down list box.	correlated (default)
						independent
						unidimensional
						user-defined
	b		To specify the pattern for the covariance matrix of the level-2 random effects.	Text box	Enter integer values >= 0	
3	a	Level-3 (Co)variance Patterns	To specify the pattern type for the covariance matrix of the level-3 random effects.	Drop-down list box	Select an option from the drop-down list box.	correlated (default)
						independent
						unidimensional
						user-defined
	b		To specify the pattern for the covariance matrix of the level-3 random effects.	Text box	Enter integer values >= 0	

Examples of Patterns:

o The pattern below is used to constrain the coefficients of Treatment 1 and Treatment 2 to be equal. Likewise, the coefficients of Treatment 3 and Treatment 4 are constrained to be equal.

Explanatory Variables	Pattern
Treatment 1	1
Treatment 2	1
Treatment 3	3
Treatment 4	3

Note that a number cannot be larger than the row number on the grid. For example, the following pattern is not recognized by SuperMix:

Explanatory Variables	Pattern
Treatment 1	2
Treatment 2	2
Treatment 3	4
Treatment 4	4

o The table below shows three possible patterns for the level-2 variances and covariances of the random effects Time1, Time2, Time3, and Time4.

Level-3 co(variance)	Pattern 1	Pattern 2	Pattern 3
Variance, Time1	1	1	1
Time1, Time2	2	2	2
Variance, Time2	1	3	3
Time1, Time3	2	0	0
Time2, Time3	2	2	0
Variance, Time3	1	6	6
Time1, Time4	2	0	0
Time2, Time4	2	0	0
Time3, Time4	2	2	9
Variance, Time4	1	10	10

Pattern 1 restricts all the variances to be equal and, likewise, all the covariances to be equal. Pattern 2 specifies that all variances should be estimated freely, all covariances one time unit apart are set equal, and all covariances more than one time unit apart are fixed at the values specified on the **Starting Values** screen, the default for covariances being zero. Pattern 3 specifies that Time1 and Time2 are correlated, but uncorrelated with Time3 and Time4, which are correlated with each other.

2.4.5 The Advanced screen

The appearance of the **Advanced** screen depends on the type of outcome variable selected on the **Configuration** screen, and is used to change default settings used in SuperMix. Specific examples of the use of this screen are given in Section 3.4 to 3.7. Screens for the various outcome types are given next.

Advanced screen for continuous outcomes with normal distribution

In repeated measurement studies, the assumption of uncorrelated identically distributed level-1 error terms is often unrealistic. The options on the **Advanced** screen with normal distribution shown below allow for correlated level-1 error terms that follow a time series process. These options is currently only available for 2-level models.

The 10 different entries of the **Advanced** screen of the **Model Setup** window for continuous response variables are summarized in Table 2.8(a).

Table 2.8(a): Entries of the Advanced screen for continuous outcomes with normal distribution

Number	Caption	Purpose	Type	Action	Options
1	Unit Weighting	To select equal or differential weighting for the units of a continuous dependent variable.	Drop-down list box	Select an option from the drop-down list box.	equal (default) **differential**
2	Level-1 Weight	To specify the weight variable that defines the first level of the hierarchy in the data.	Drop-down list box	Select a variable from the drop-down list box.	
3	Level-2 Weight	To specify the weight variable that defines the second level of the hierarchy in the data.	Drop-down list box	Select a variable from the drop-down list box.	
4	Level-3 Weight	To specify the weight variable that defines level-3 of the hierarchy in the data.	Drop-down list box	Select a variable from the drop-down list box.	
5	Distribution Model	To select an appropriate distribution model.	Drop-down list box.	Select a distribution from the drop-down list box.	**normal (default)** gamma inverse Gaussian
6	'Time' Variable	To specify the time variable.	Drop-down list box	Select a variable from the drop-down list box.	fixed AC terms
7	Autocorrelation	To specify the type of autocorrelation terms.	Drop-down list box	Select an option from the drop-down list box.	no AC terms (default) fixed AC terms **estimate all**

Table 2.8(a): Entries of the Advanced screen for continuous outcomes with normal distribution (continued)

Number	Caption	Purpose	Type	Action	Options
8	Error Form	To specify a time series model for the auto-correlated errors.	Drop-down list box	Select an option from the drop-down list box.	Stationary AR1 (default)
					Non-stationary AR1
					Stationary MA1
					Stationary ARMA(1,1)
					General Auto-correlation
9	Autocorrelation Terms	To specify the number of autocorrelation terms.	Text box	Enter an integer if the default 1 is not desired.	
10	Autocorrelation Starting Values	To specify the starting value(s) for the autocorrelation(s).	Grid box	Enter a real number in the region of [-0.99, 0.99].	

Advanced screen for continuous outcomes with gamma or inverse Gaussian distribution

When the gamma or inverse Gaussian distribution is selected, the options on the **Advanced** screen are slightly different from the screen with normal distribution, as shown below.

The 3 different entries of the **Advanced** screen of the **Model Setup** window for continuous response variables with gamma or inverse Gaussian distributions are summarized in Table 2.8(b).
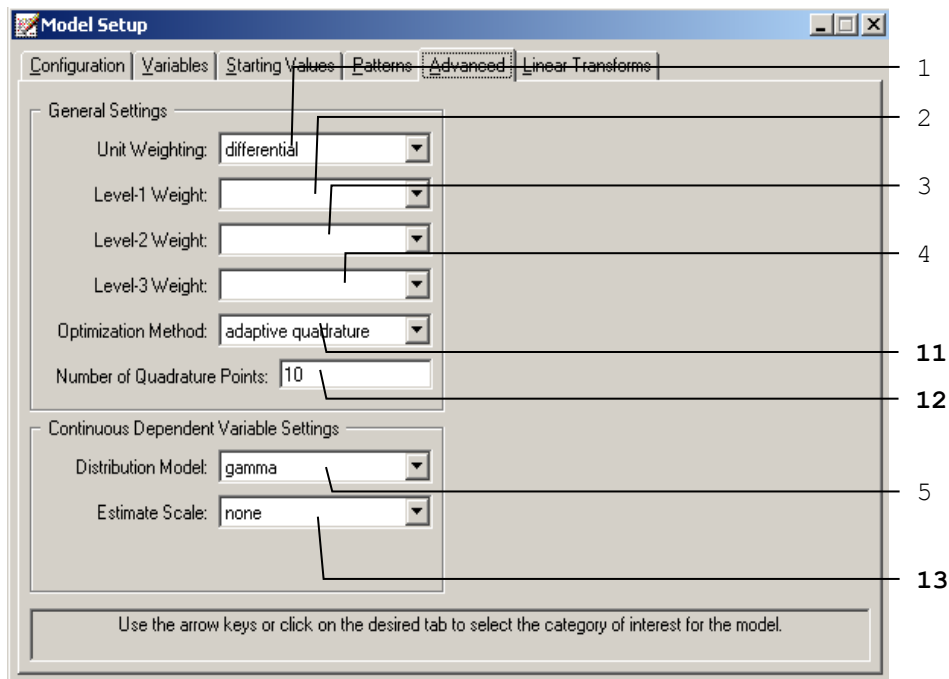
Table 2.8(b): Entries of the Advanced screen for continuous outcomes with gamma or inverse Gaussian distribution

Number	Caption	Purpose	Type	Action	Options
11	Optimization Method	To select the optimization method.	Drop-down list box	Select an optimization method from the drop-down list box.	**adaptive quadrature (default)** maximum posterior adaptive modal, non-adaptive quadrature
12	Number of Quadrature Points	To enter the number of quadrature points (per random-effect dimension).	Text box	It is usually set to 10 for 1 effect and 5 to 10 for 2 or 3 effects.	**10 (default)**
13	Estimate Scale	To select the method for estimating the scale.	Drop-down list box	Select the method from the drop-down list box.	**none (default)** deviance Pearson

GRAPHICAL USER INTERFACE

Advanced screen for ordered outcomes

An important feature of mixed-effects models with ordered outcomes is the inclusion of threshold parameters in the model. As illustrated in Section 3.6, the number of threshold parameters equals $C-2$, where C is the number of distinct categories of the outcome variable. If right-censoring (see Section 3.7) is included in the model specification, the number of thresholds becomes $C-1$ and a **Censor Variable** is selected. The mixed-effects model for ordinal outcomes additionally allows for the inclusion of **Explanatory Variable**-threshold interaction terms. If entry number 14 in the screen below is set equal to 2, for example, then interaction terms of the first two explanatory variables with each of the thresholds are included in the model. One can also select a weight variable, link function (**Function model**) and the number of quadrature points to be used for the approximation of the likelihood function by numerical integration.

Table 2.9 gives a summary of the possible 8 entries of the **Advanced** screen of the **Model Setup** window for an ordered response variable. The information for the other entries is given in Table 2.8 (a) and (b).

Table 2.9: Entries of the Advanced screen for ordered outcomes

Number		Caption	Purpose	Type	Action	Options
14	a	Include Interactions	To indicate if explanatory variable interactions should be included in the model.	Drop-down list box	Select an option from the drop-down list box.	no (default)
						yes
	b	Number of Interactions	To specify the number of interactions.	Text box	Enter an integer if the default maximum allowable value is not desired.	
15		Function Model	To specify the link function for the model.	Drop-down list box	Select a function from the drop-down list box.	**probit (default)**
						logistic
						complementary log-log
						log-log
16		Level-2 Random Thresholds	To specify if there are thresholds for the level-2 random effects	Drop-down list box	Select an option from the drop-down list box.	**no (default)**
						yes
17		Level-3 Random Thresholds	To specify if there are thresholds for the level-3 random effects	Drop-down list box	Select an option from the drop-down list box.	**no (default)**
						yes
18	a	Right-Censoring	To specify if right-censoring is included.	Drop-down list box	Select an option from the drop-down list box.	none (default)
						include
	b	Censor Variable	To specify the censor variable.	Drop-down list box	Select a variable from the drop-down list box.	

Table 2.9: Entries of the Advanced screen for ordered outcomes (continued)

Number	Caption	Purpose	Type	Action	Options
19	Model Terms	To select subtracting or adding the model terms to the threshold.	Drop-down list box	Select an option from the drop-down list box.	**subtract means** $(\gamma - \mathbf{X'\beta})$ **(default)**
					add, means $(\gamma + \mathbf{X'\beta})$

Advanced screen for nominal outcomes

This screen is similar to the one used for ordinal variables, but contains fewer entries as shown below.

Note that entries in the **Advanced** screen for the nominal outcome variable are similar to those for the continuous outcome. All the information for the first 6

entries are given in Table 2.8(a) and (b). Information on entry no. 20 is given in Table 2.10 below.

Table 2.10: Entries of the Advanced screen for nominal outcomes

Number	Caption	Purpose	Type	Action	Options
20	Reference Category	To select whether the first or last category of the outcomes should be used as the reference category.	Drop-down list box	Select a reference category from the drop-down list box.	**first (default)**
					last

Advanced screen for count outcomes with Poisson distribution

The screen below has the same functionality as the screen used for nominal outcomes, except that provision is made here for the specification of an **Offset Variable** and the choice of the **Distribution Model**.

In practice, it can occur that the coefficient of some covariate is assumed to be unity. This covariate is commonly known as an offset variable. Offsets are typically used when the response variable is a rate rather than a number or count. The available distributions available for a count variable include the Poisson and negative binomial distributions.

The 2 entries pertaining to the offset variable and the distributions on the **Advanced** screen of the **Model Setup** window for a count outcome are summarized in Table 2.11. Note that entries in the **Advanced** screen for the count outcome variable are similar to those for the continuous outcome. The information for the other entries is given in Table 2.8 (a) and (b).

Table 2.11(a): Entries of the Advanced screen for count outcomes with Poisson distribution

Number		Caption	Purpose	Type	Action	Options
21	a	Incorporate Time Offset	To select whether or not to include an offset variable.	Drop-down list box	Select an option from the drop-down list box.	no (default)
						yes
	b	Offset Variable	To specify the offset variable.	Drop-down list box	Select a variable from the drop-down list box.	
22		Distribution Model	To select an appropriate distribution model.	Drop-down list box.	Select a distribution from the drop-down list box.	**Poisson (default)**
						negative binomial

Advanced screen for count outcomes with negative binomial distribution

When the negative binomial distribution is selected, the **Advanced** screen of the count variable is slightly different, as shown below.

Note that entries 1 and 2 are different from those on the previous screen. More information on these are given in Table 2.11(b) below.

Table 2.11(b): Entries of the Advanced screen for count outcomes with negative binomial distribution

Number	Caption	Purpose	Type	Action	Options
23	Distribution Model	To select an appropriate distribution model.	Drop-down list box.	Select a distribution from the drop-down list box.	Poisson (default)
					negative binomial
24	Dispersion Parameter	To enter the dispersion parameter for the negative binomial model.	Text box	Enter any numeric value greater than 0.0. The default value is 1.0.	

Advanced screen for binary outcomes

As shown below, the **Advanced** screen used for binary outcomes includes the choices of the **Distribution Model** include Bernoulli and binomial. The new entry, **Number of Trials**, is an unique option for the binary outcomes, available when the binomial distribution is selected.

The 3 entries pertaining to the distribution model, link function and number of trials on the **Advanced** screen of the **Model Setup** window for a binary outcome are summarized in Table 2.12. Note that entries in the **Advanced** screen for the binary outcome variable are similar to those for the continuous outcome. The information for the other entries is given in Table 2.8 (a) and (b).

Table 2.12: Entries of the Advanced screen for count outcomes

Number	Caption	Purpose	Type	Action	Options
24	Distribution Model	To select an appropriate distribution model.	Drop-down list box.	Select a distribution from the drop-down list box.	Bernoulli (default)
					binomial
25	Function Model	To select from probit, logistic, complementary log-log and log-log response functions.	Drop-down list box.	Select a link function from the drop-down list box.	**probit (default)**
					logistic
					complementary log-log
					log-log
26	Number of Trials	To select the column of the spreadsheet which contains the number of trials.	Drop-down list box.	Select the desired variable.	

2.4.6 The Linear Transforms screen

Linear transforms are used to test hypotheses of the type

$$H_0 : c_1\beta_1 + c_2\beta_2 + c_3\beta_3 + ... + c_k\beta_k = 0$$

where $\beta_1, \beta_2, \beta_3, ..., \beta_k$ are model parameters and $c_1, c_2, c_3, ..., c_k$ user-specified real-valued coefficients. For each linear transform, a Z-statistic and an associated two-tailed p-value are saved to the output file. The Z-statistic is a function of the estimated parameters and the large sample covariance matrix of the estimates. The

value of the linear transform when the parameters are replaced with their estimates is also of interest. This value also appears in the output.

An example of the **Linear Transform** screen in the case of an ordinal outcome variable is given below. Also see Section 3.8 for an additional example. For continuous, count and nominal variables the **Linear Transform** screens are identical, but it differs from that for an ordinal outcome in that the screen for an ordinal outcome contains threshold parameter information.

Linear Transforms screen for continuous and count outcomes

The buttons and options on the **Linear Transforms** screen are shown below.
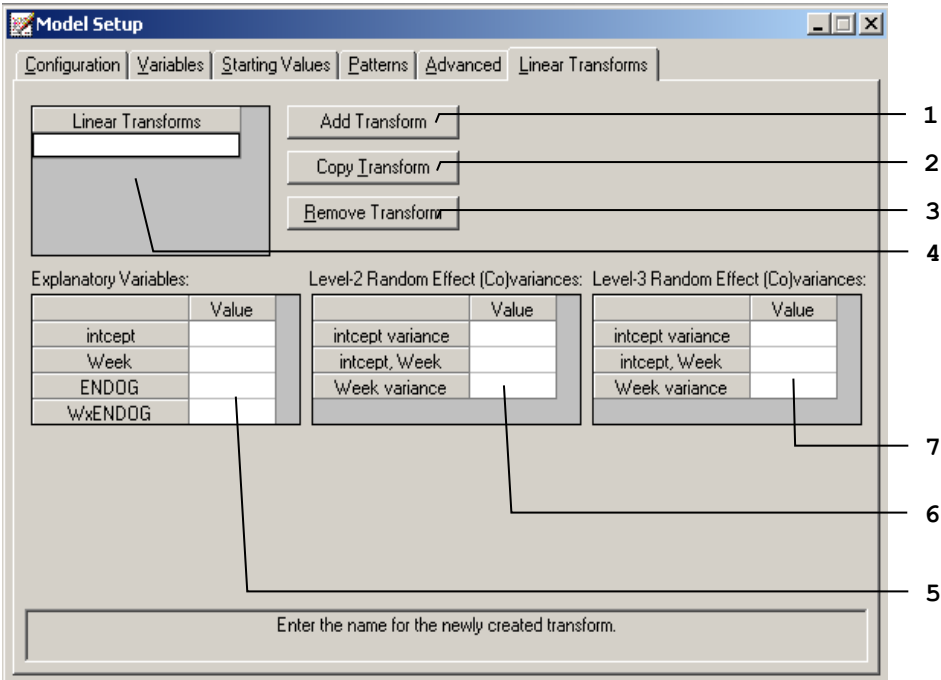
A summary of the 3 buttons and 4 different entries of the **Linear Transforms** screen of the **Model Setup** window for an ordered outcome is given in Table 2.13.

Table 2.13: Entries of the Linear Transforms screen for continuous and count outcomes

Number	Caption	Purpose	Type	Action
1	Add Transform	To create a new linear transform in 4.	Click button	Click to add a blank transform.
2	Copy Transform	To create a copy the selected transform in 4 with a different name.	Click button	Click to copy and paste the select transform.
3	Remove Transform	To delete the selected transform in 4.	Click button	Click on the button to delete the selected transform.
4	Linear Transforms	To select the linear transform and edit the components.	Grid box	Enter string(s) as names for transforms.
5	Explanatory Variables	To specify the values for the linear transformation(s) of the fixed part of the model.	Grid box	Enter real number(s).
6	Level-2 Random Effect (Co)variances	To specify the values for the linear transformation(s) of the covariance matrix of the level-2 random effects.	Grid box	Enter real number(s).
7	Level-3 Random Effect (Co)variances	To specify the values for the linear transformation(s) of the covariance matrix of the level-3 random effects.	Grid box	Enter real number(s).

For example, in the **Linear Transforms** screen shown below we wish to test the hypothesis that

$$\beta_{intercept} + \beta_{Threshold1} = 0.$$
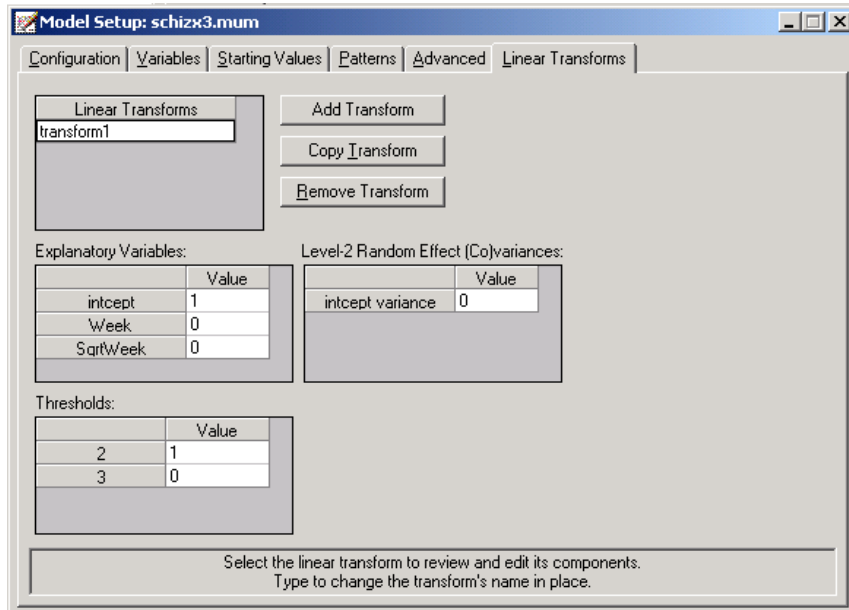
Linear Transforms screen for ordered outcomes

Additional grid boxes appear to allow for user-specified starting values of thresholds and threshold-explanatory variable(s) interaction(s) for ordinal outcomes. The following screen is an example of the **Linear Transforms** screen of the **Model Setup** window.

Only the 2 entries shown in bold typeface are either new or different compared with those on the **Linear Transforms** screen of the **Model Setup** window for continuous or count outcome variables. These 2 entries are summarized in Table 2.14. Please refer to Table 2.13 for information about all the other entries.
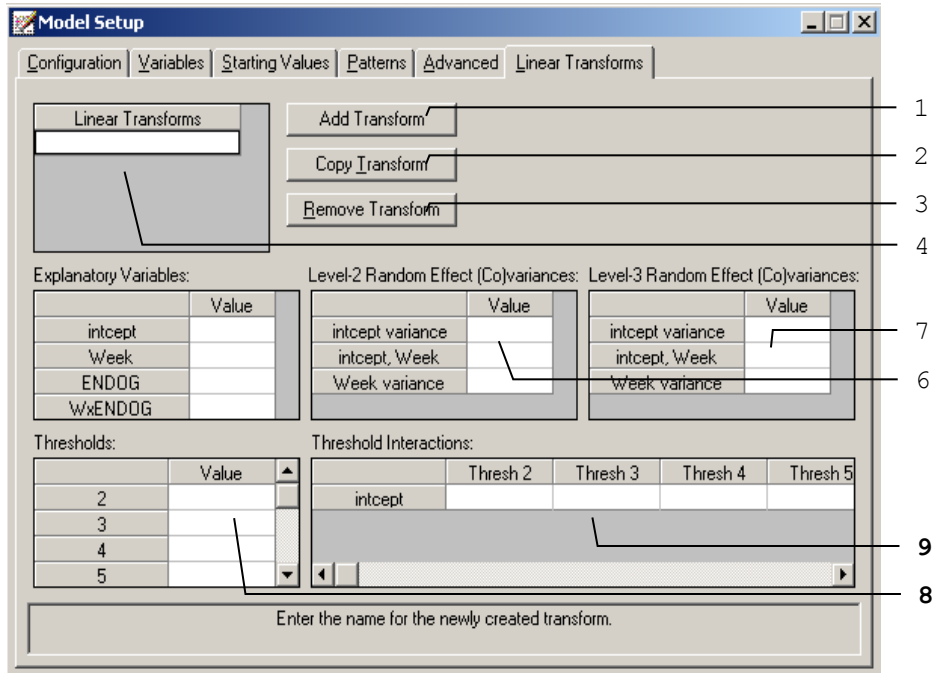
Table 2.14: Entries of the Linear Transforms screen for ordered outcomes

Number	Caption	Purpose	Type	Action
8	Thresholds	To specify the values for the linear transformation(s) of the thresholds.	Grid box	Enter real number(s).
9	Thresholds Interactions	To enter values for the linear transformations of the threshold interactions.	Grid box	Enter real number(s).

Linear Transforms screen for nominal and binary outcomes

The following screen is an example of the **Linear Transforms** screen of the **Model Setup** window. Note that additional slide bars appear for entries 5 to 7, so that values may be entered for other categories, relative to the selected reference category.
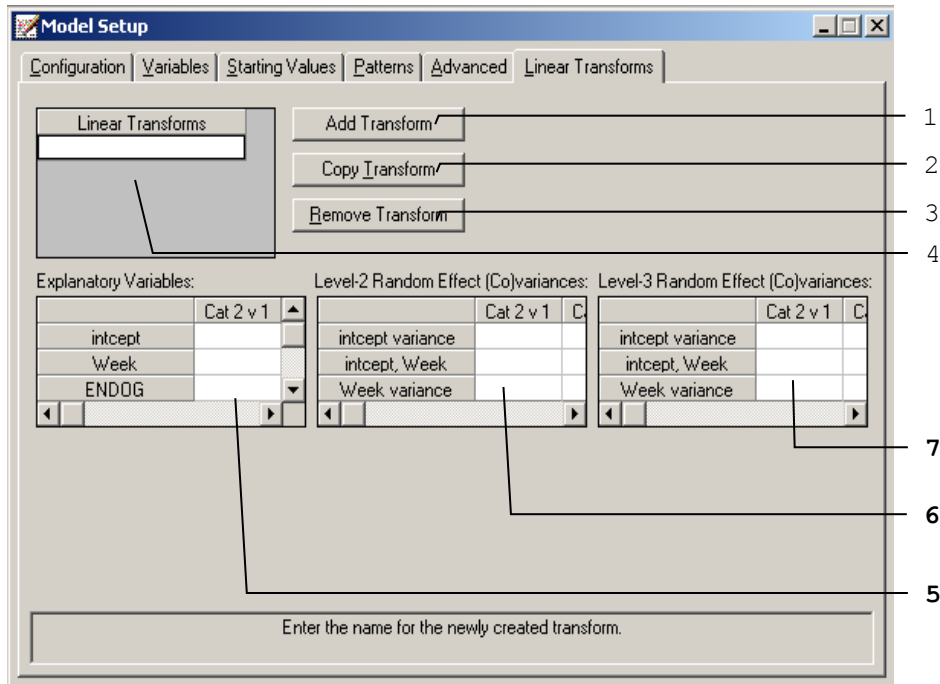
Only the 2 entries shown in bold typeface are either new or different compared with those on the **Linear Transforms** screen of the **Model Setup** window for continuous or count outcome variables. These 2 entries are summarized in Table 2.15. Please refer to Table 2.13 for information about all the other entries.

Table 2.15: Entries of the Linear Transforms screen for nominal outcomes

Number	Caption	Purpose	Type	Action
5	Explanatory Variables	To specify the values for the linear transformation(s) of the fixed part of the model.	Grid box with slide bar	Enter real number(s).
6	Level-2 Random Effect (Co)variances	To specify the values for the linear transformation(s) of the covariance matrix of the level-2 random effects.	Grid box with slide bar	Enter real number(s).
7	Level-3 Random Effect (Co)variances	To specify the values for the linear transformation(s) of the covariance matrix of the level-3 random effects.	Grid box with slide bar	Enter real number(s).

2.5   Data manipulation

The SuperMix spreadsheet can be manipulated in various ways. Rows and columns can be changed directly, and simple computations or more complex built-in functions can be used in individual cells. Some of these manipulations are discussed and illustrated in the following sections, using the data spreadsheet **Examples\Primer\Graphics\demo.ss3**. There are 15 cases (patients) in the data set.

The spreadsheet window for **demo.ss3** is opened as follows:

- o Use the **Open** option on the **File** menu of the main window to open the **Open Spreadsheet** dialog box.
- o Browse for the file **demo.ss3** in the **Examples\Primer** folder.
- o Select the file and click on the **Open** button to open the following SuperMix spreadsheet window.

The variables include:

- o Group is a variable with 3 categories, indicating the group number of the patient (5 patients in each group).
- o Age is the age of the patient.
- o WT_kg denotes the weight of the patient in kg.
- o PFat is a measure of percentage body fat.

2.5.1 Basic data manipulations

It is important to note that any change of the data file will not be saved unless you use the **Save** option on the **File** menu.

Cells

A careful examination of the data shows that the Age entry of the third observation is 330.00. This is obviously a typing error. Upon further investigation, it turns out that the correct age value is 33. To correct this error, select the cell, change the value of the formula box from 330.00 to 33.00, and then click on the **Apply** button.
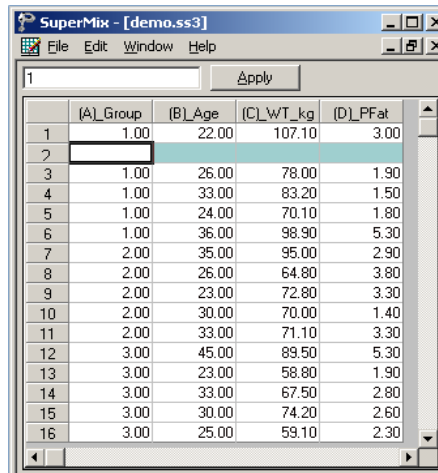
Rows

To work with a row (or rows) of the data file, click on the row tab(s) to select the complete row(s) and then right-click on the selection to display all the options from the pop-up menu.

Insert a row

For example, to insert another row (observation) between the first and the second rows, select the second row by clicking on the row 2 tab, right-click on the selected row to activate the menu and select the **Insert Row** option to create the window as shown below.
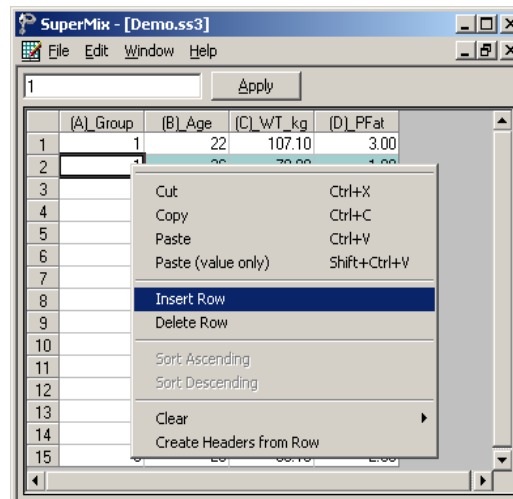
An empty row is added to the spreadsheet above the previous second row and the total sample size is changed to 16 as shown below.

Delete a row

To delete the empty row that was inserted, select the second row by clicking on the row 2 tab. Right click on the selected row and select the **Delete Row** option to delete the second row.

Columns

To work with a column (or columns) of the data file, first click on the column header(s) to select the column(s) and then right-click on one of the selected columns to see all the options listed on the pop-up menu.

Insert a column
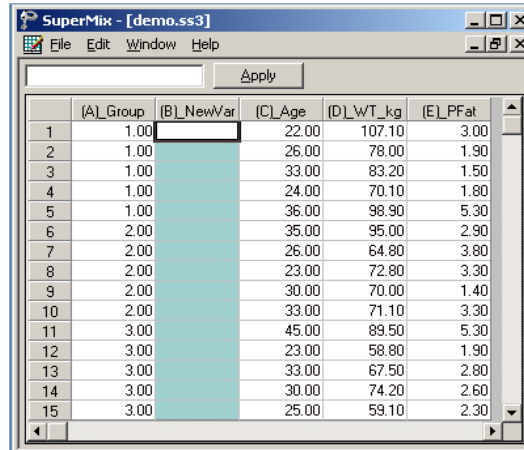
To insert another variable between (A)_Group and (B)_Age, first click on the header of (B)_Age, right-click on the column to activate the menu and select the **Insert Column** option to create the window as shown below.
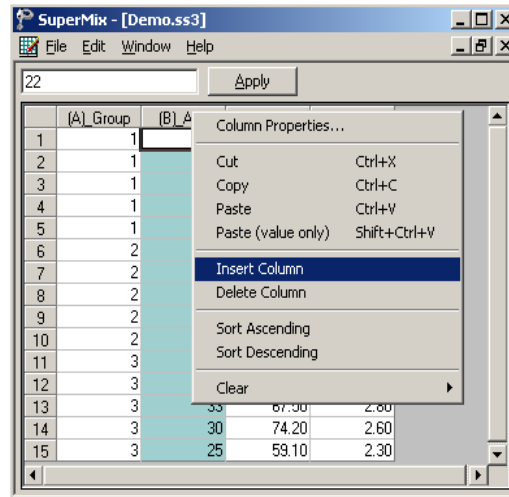
A new column (variable) is added to the spreadsheet. Change the variable name; define the variable type and decimal places by selecting the column header, right-clicking and selecting **Column Properties** to open the dialog box as shown below.
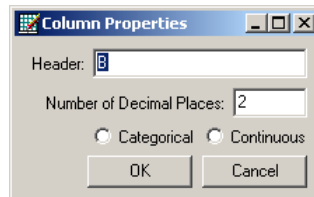
Input a variable name, such as NewVar, and then click on the **OK** button to return to the spreadsheet window as shown below.

GRAPHICAL USER INTERFACE

Delete a column

To delete the NewVar column that was inserted, select the variable NewVar by clicking on the column header. Right-click and select the **Delete Column** option to delete column B.

2.5.2 Simple computations

Assigning values to a new variable

In **demo.ss3**, the variable WT_kg is a variable denoting weight in kilograms. We would like to use the variable WT_lb, that is, the corresponding weight in pounds. To create this variable, first insert a column and change the column header to WT_lb as illustrated earlier in Section 2.5.1 to generate the following spreadsheet window.

Select the column containing the variable WT_lb, input the function **2.20462*(C1)** in the formula box and click on the **Apply** button to get the new variable WT_lb as shown below. The formula applies to each row of (D)_WT_lb provided that

- o   this column is selected (highlighted)
- o   the first cell of the variable(s) in the formula, in this case C1, is referenced.
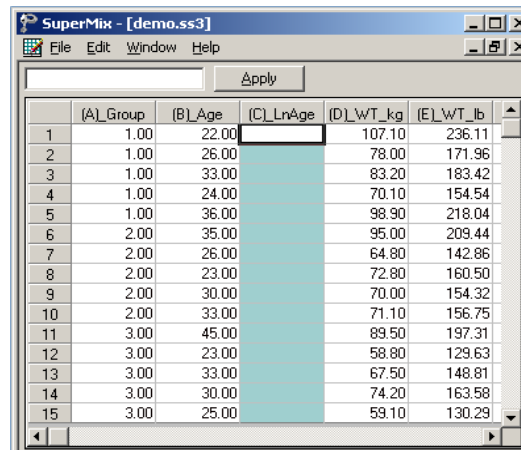
Save the changes to **demo.ss3** by clicking on the **Save** option on the **File** menu.

2.5.3 Built-in functions

LN function
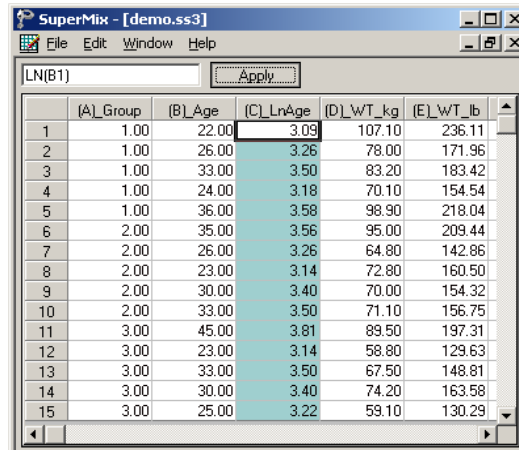
In mixed-effects models, we often consider the natural log of a variable. For example, the natural log of Age in **demo.ss3** may be a more appropriate explanatory variable than the original age given in years. The variable LnAge can easily be created by using the options available in the SuperMix spreadsheet.

First, create a new column with the header LnAge as shown below.

Select the column containing the variable LnAge, input the function **LN(B1)** in the formula box, and click on the **Apply** button. Each value of the new variable LnAge is the natural log of the corresponding values of the variable Age as shown below.
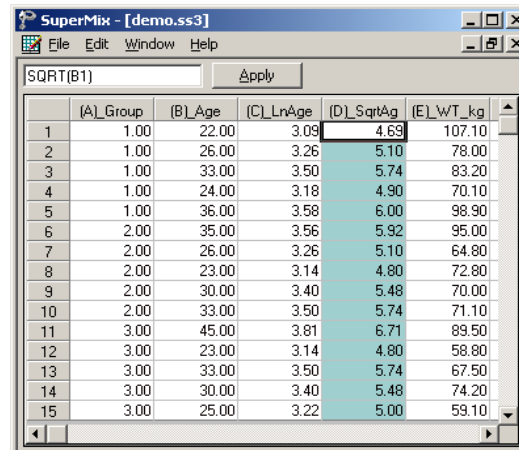
SQRT function

To add another variable, SqrtAge, which is the square root of Age, we proceed as follows. As above, first create a new column with the header SqrtAge.

Select the column containing the variable SqrtAge, input the function **SQRT(B1)** in the formula box, and click on the **Apply** button. Each value of the new variable

SqrtAge is the square root value of the corresponding value of the variable Age as shown below.

Save the changes to **demo.ss3** by clicking on the **Save** option on the **File** menu.

Table 2.16 contains a selection of the built-in functions in SuperMix. A list of values can be any of the following types:

- o (B1, B2, B3) is the selection of the first three values of the variable in column B.
- o (A1:A15) selects all the values of the variable in column A. Starting with the first and ending with the one in row 15.
- o (A1:A5, A7, A11:A15) selects the values of row 1 to row 5, row 7, and row 11 to row 15 of column A.
- o (A3:A6, B7, D12:D15) contains a list of values from more than one column. This selection includes the values of row 3 to row 6 of column A, row 7 of column B and row 12 to row 15 of column D.

Table 2.16: Selection of SUPERMIX functions

Function	Definition
ABS(value)	Absolute value
AVERAGE (list of values)	Average
EXP(value)	Exponent base e
LN(value)	Natural log
LOG(value)	Logarithm
MAX(list of values)	Maximum value
MEDIAN(list of values)	Median
MIN(list of values)	Minimum value

Table 2.16: Selection of SUPERMIX functions (continued)

Function	Definition
MODE(list of values)	Mode
SQRT(value)	Square root
SQUARE(value)	Square

2.5.4 Other useful data manipulations

The data manipulation capabilities of the SuperMix spreadsheet window can be used to create interaction terms, essentially a product of variables, for use in modeling. It can also be used to perform grand mean centering of variables.

Absolute references

If you do not want SuperMix to adjust references when you copy a formula to a different cell, use an **absolute** reference. For example, if your formula multiplies cell A5 with cell C1 (=A5*C1) and you copy the formula to another cell, SuperMix will adjust both references. You can create an absolute reference to cell C1 by placing a **dollar sign** ($) before the parts of the reference that do not change. To create an absolute reference to cell C1, for example, add dollar signs to the formula as follows: = A5*C1.

Creating an interaction term

Suppose that we want to study the possible interaction between a subject's age and weight (in pounds). This product of Age and WT_lb, is created in the SuperMix spreadsheet window as follows.

First, create a new column with header Age_WT. Then, select this column, and input the function **(B1)*(E1)** in the formula box. Click on the **Apply** button. Each value of the new variable Age_WT is equal to the product of the corresponding values of Age and WT_lb as shown below.

Grand mean centering

It is often useful to center a predictor variable around its grand mean. To illustrate, we grand mean center the variable PFat. To do so, first create a new column with the header of PFat_Mea. Then, select the PFat_Mea column, input the function **(H1)–AVERAGE(H$1:H$15)** in the formula box and click on the **Apply** button. Each value of the new variable of PFat_Mea now contains the difference between the corresponding original PFat value and the grand mean of all the PFat values. Note that the spreadsheet functions are not case sensitive.

3 Examples

3.1 Introduction

SuperMix fits mixed-effects models to nested data and allows users to specify the type of outcome (dependent) variable as continuous, count, ordinal, nominal, or survival. Each model type is based on a functional relationship between the expectation of the outcome variable and a linear combination of explanatory variables and random effects. This functional relationship is specified by a suitable link function, such as the identity link for a continuous outcome and a log link for a count outcome variable. Each of these variable types will be discussed briefly in this section. In the remainder of the chapter, a selection of examples that illustrate various types of mixed-effects models are given. Results of the analyses are discussed in detail, and are augmented with various graphical displays to demonstrate the characteristics of a hierarchical model or serve as an aid in the interpretation of results.

Continuous outcomes

In many research projects, the response variable of interest is a continuous variable. Examples of continuous response variables are inpatient expenditure of medical interns, earnings of social workers, insurance claim costs, failure times of X-ray machine parts, total cholesterol scores of heart patients, aggregate loss dollars for life insurance policies, etc. Values of continuous outcome variables have the property that they can be interpreted with respect to scale and location. For example, if the average pulse rate is 72 beats per second, a pulse rate of 79 is approximately 10% above average. A person with a pulse rate of 36 has half the pulse rate of a person with an average pulse rate.

SuperMix can fit 2-level and 3-level models with continuous response variables to hierarchical data, a feature that is illustrated in Sections 3.2 and 3.3. In Section 3.2, a level-2 model with a continuous response variable is considered. The focus is on the longitudinal relationship between imipramine (IMI) and desipramine (DMI) plasma levels and clinical response in depressed inpatients. Section 3.3 contains an example of data from a complex survey sample. Data from a longitudinal national

survey are used to examine estimates of health care use and expenditure based on the health expenditures of a sample of U.S. civilian non-institutionalized participants. Here, we assume that the strata are level-3 units and that the clusters nested within the strata are the level-2 units.

Count outcomes

Variables measured in scientific studies come in a wide assortment. When statisticians refer to a "count" variable, they mean a variable that is ordinal, typically scored 0, 1, 2, …, without fractional values such as 2.4 or 6.75. They also mean that the variable is a tally that records how often some behavior occurred, or how many incidents of a particular kind were observed in each subject of a study.

Researchers are often interested in studying response variables such as the number of traffic accidents in a year at a specific intersection, the number of children with medical problems ever born to African-American women, the number of blood cells in a blood sample, the number of infectious organisms spread on an agar plate, the number of infected children in a preschool class, the number of people in a city who contracted a pollution-related disease, the number of salamanders in a location, etc. Since these response variables involve a number, they are commonly referred to as counts.

Count variables are often analyzed in exactly the same way that continuous variables are handled, most often with a method that incorrectly assumes the count is a bell-shaped normal distribution. But counts are ordinal variables, usually skewed and with a small range. They have none of the characteristics of a continuous variable. While in many instances there are few practical problems treating them as if they were continuous variables, it is easy to find examples where an inappropriate analysis of a count variable loses important information that a better approach would convey.

Multilevel models for counts usually assume a Poisson distribution for the response variable. In Section 3.4, we illustrate how SuperMix can be used to fit Poisson regression models with a random intercept to health-related count data. We also use

this example to demonstrate how SuperMix can be used to specify an offset variable for Poisson regression models with a random intercept.

Binary outcomes

Binary response variables are often the focus of empirical studies. Examples of binary response variables are diagnosis of breast cancer (absent or present), heart disease (yes or no), damage to laser equipment (damage or no damage), and depression in substance abuse clients (yes or no), etc. In the special case of one trial for each observation, the Binomial distribution simplifies to the Bernoulli distribution, and either distribution can be used. However, if a "number of trials" variable is available, the Binomial distribution would be the appropriate choice.

In Section 3.5 we consider data from a longitudinal study on mentally ill patients randomly assigned to two treatment groups. The outcome variable, an item from the Inpatient Multidimensional Psychiatric Scale, is recoded into two distinct categories. A mixed-effects model with logit link function is fitted to the data.

Ordinal outcomes

Researchers are often involved in studying ordinal response variables such as level/degree of mental impairment (well, mild symptom formation, moderate symptom formation or impaired), patient satisfaction measured on a 5-point Likert scale, severity of lower back pain (none, mild, moderate or severe), arthritis improvement (none, some or marked), etc. It is assumed that a person who selected one category has more of a given characteristic than if s/he had chosen a lower category, but we do not know how much more. It is common practice to treat the scores 1, 2, 3, … assigned to categories as if they have metric properties, but this is wrong. Ordinal variables are not continuous variables and should not be treated as if they are. Since they do not have origins or units of measurements, the means, variances, and covariances of ordinal variables have no meaning. The only information available is counts of cases in each cell of a multi-way contingency table. Thus, to use ordinal variables in mixed-effects models requires other techniques than those traditionally employed with continuous variables.

In Section 3.6 we re-examine the data set described in Section 3.5, but recode the outcome variable into four distinct categories. We assume a probit link function and give a discussion of the results of the analysis.

Nominal outcomes

SuperMix can be used to fit models to nominal response variables. The smoking status of a patient (never smoked, former smoker or current smoker), type of special care (home for the elderly, daytime nurse, full time nurse), cancer type of female cancer patients (breast, lung, brain, leukemia, liver, colon or other) etc. are examples of nominal response variables. The difference between ordinal and nominal variables is that there is a sense of "more" or "less", "better" or "worse", associated with the categories of an ordinal variable. This type of ordering is not present in the case of nominal data. For example, it is not obvious how one would rank the categories of a variable representing marital status or home language.

In Section 3.7 we fit a model with a nominal outcome variable to examine the effect of the use of Section 8 certificates to provide independent housing to the severely mentally ill homeless.

Survival outcomes

The outcome variable of interest may be the time that has elapsed from the onset of a condition of treatment until some event has taken place. In health research, this time is often referred to as the *survival* time. An example is a measurement of the time from the initiation of radiation treatment until a patient dies (the event). Survival times are always non-negative and often the survival time is not known because the event did not occur during the observation period. If at any time t_1 the event has not occurred (for example, the patient did not die), the observation at that time point is said to be *censored*. It is possible that the event never occurs during the course of the study, which is known as *right-censoring*. On the other hand, *left-censoring* occurs if the event took place before observations begin.

In the modeling of survival rates, it is either assumed that the survival times are continuous or discrete. Discrete time survival data most commonly arise when

observations are only made at discrete time intervals, for example once a week or once a month. A further important concept in survival analysis is *hazard* or *risk.* The hazard can be defined as the probability of an event occurring at time t given that a person is still at risk (for example, still alive). Use is often made of an *event indicator* variable that assumes the value of 1 if an event has occurred and 0 otherwise.

In Section 3.8 we examine the risk of being suspended from school, using data from a longitudinal study where children are nested within therapists. We wish to determine the effect of providing financial assistance to a family on the school suspension rate of children in the study.

3.2 Two-level models for continuous outcomes

3.2.1 The data

The data set is from a study described in Reisby *et. al.*, (1977) that focused on the longitudinal relationship between imipramine (IMI) and desipramine (DMI) plasma levels and clinical response in 66 depressed inpatients (37 endogenous and 29 non-endogenous). Following a placebo period of 1 week, patients received 225 mg/day doses of imipramine for four weeks. In this study, subjects were rated with the Hamilton depression rating scale (HDRS) twice during the baseline placebo week (at the start and end of this week) as well as at the end of each of the four treatment weeks of the study. Plasma level measurements of both IMI and its metabolite DMI were made at the end of each week. The sex and age of each patient were recorded and a diagnosis of endogenous or non-endogenous depression was made for each patient.

Although the total number of subjects in this study was 66, the number of subjects with all measures at each of the weeks fluctuated: 61 at week 0 (start of placebo week), 63 at week 1 (end of placebo week), 65 at week 2 (end of first drug treatment week), 65 at week 3 (end of second drug treatment week), 63 at week 4 (end of third drug treatment week), and 58 at week 5 (end of fourth drug treatment week). The sample size is 375. Data for the first 10 observations of all the variables used in this section are shown below in the form of a SuperMix spreadsheet file, named **reisby.ss3**.

The variables of interest are:

- o Patient is the patient ID (66 patients in total).
- o HDRS is the Hamilton depression rating scale.
- o Week represents the week (0, 1, 2, 3, 4 or 5) at which a measurement was made.
- o ENDOG is dummy variable for the type of depression a patient was diagnosed with (1 for endogenous depression and 0 for non-endogenous depression).
- o WxENDOG represents the interaction between Week and ENDOG, and is the product of Week and ENDOG.

3.2.2 The models

A general two-level model for a continuous response variable y depending on a set of r predictors $x_1, x_2, \ldots, x_r$ can be expressed as

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{u}_i + e_{ij}$$

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES

where y_{ij} denotes the value of y for the j-th level-1 unit nested within the i-th level-2 unit for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n_i$. The scalar product $\mathbf{x}_{ij}'\boldsymbol{\beta}$ is the fixed part of the model, and $\mathbf{z}_{ij}'\mathbf{u}_i$ and e_{ij} denote the random part of the model at levels 2 and 1 respectively. For the fixed part of the model, $\mathbf{x}_{ij}'$ is a typical row of a design matrix $\mathbf{X}_i$ while the vector $\boldsymbol{\beta}$ contains the fixed, but unknown, parameters to be estimated. In the case of the random part of the model at level 2, $\mathbf{z}_{ij}'$ represents a typical row of a design matrix $\mathbf{Z}_i$, and $\mathbf{u}_i$ the vector of random level-2 effects to be estimated. It is assumed that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ are independently and identically distributed (i.i.d.) with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(2)}$. Similarly, the e_{ij} are assumed i.i.d., with mean vector 0 and variance σ^2. The elements of $\mathbf{z}_{ij}$ are typically a subset of those of $\mathbf{x}_{ij}$.

The random intercept and slope model

The random intercept and slope model for the response variable HDRS may be expressed as

$$\text{HDRS}_{ij} = \beta_0 + \beta_1 * (\text{Week})_{ij} + u_{i0} + u_{i1}(\text{Week})_{ij} + e_{ij}$$

where β_0 denotes the average expected depression rating scale value, β_1 denotes the coefficient of the predictor variable Week (slope) in the fixed part of the model, u_{i1} denotes the variation in the slopes over patients, and u_{i0} and e_{ij} denote the variation in the average expected HDRS value over patients and between patients respectively.

The random intercept and slope with a covariate and an interaction model

The random intercept and slope model for the response variable HDRS with the variable ENDOG as a covariate and with an interaction effect between Week and ENDOG may be expressed as

$$\text{HDRS}_{ij} = \beta_0 + \beta_1 * (\text{Week})_{ij} + \beta_2 * (\text{ENDOG})_i + \beta_3 * (\text{WxENDOG})_{ij}$$
$$+ u_{i0} + u_{i1} (\text{Week})_{ij} + e_{ij}$$

where β_0 denotes the average expected depression rating scale value, β_1 and β_2 denote the coefficients of the predictor variables Week and ENDOG in the fixed part of the model, β_3 denotes the coefficient of the interaction between Week and ENDOG in the fixed part of the model, u_{i1} denotes the variation in the Week slopes over patients, and u_{i0} and e_{ij} denote the variation in the average expected HDRS value over patients and over measurements (*i.e.*, between patients) respectively.

3.2.3 Example: Random intercept and slope model

Importing the data

The random intercept and slope model above is fitted to the data in **reisby.ss3**. The first step is to create the **ss3** file shown above from the Excel file **reisby.xls**. This is accomplished as follows.

- o Use the **Import Data File** option on the **File** menu to open the **Open** dialog box.
- o Browse for the file **reisby.xls** in the **Examples\Primer\Continuous** folder.
- o Select the file and click on the **Open** button to open the following SuperMix spreadsheet window for **reisby.ss3**.

After selecting the **File**, **Save** option, we are ready to fit the random intercept and slope model for HDRS to the data in **reisby.ss3**.

Setting up the analysis

Start by selecting the **New Model Setup** option on the **File** menu as shown below to open the **Model Setup** window. The **Model Setup** window has six tabs: **Configuration**, **Variables**, **Starting Values**, **Patterns**, **Advanced**, and **Linear Transforms**. In this example, only the **Configuration** and the **Variables** tabs are used.

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES

Starting with the **Configuration** screen, the titles Longitudinal analysis of 66 depressed inpatients and Data from Reisby et. al. (1977) are entered in the **Title 1** and **Title 2** text boxes respectively. The continuous outcome variable HDRS is selected from the **Dependent Variable** drop-down list box. The variable Patient, which defines the levels of the hierarchy, is selected as the Level-2 ID from the **Level-2 IDs** drop-down list box. We keep the default settings of all the other options in order to produce the **Configuration** screen as shown above.

Click the **Variables** tab to proceed to the **Variables** screen of the **Model Setup** window. This screen shows the list of variables available for analysis and next to it two columns, with headings **E** (for explanatory variables) and **2** (for level-2 random effects). The variable Week is specified as the covariate of the fixed part of the model by checking the **E** check box for Week in the **Available** grid. We mark the **2** check box for Week in the **Available** grid to specify the random slope at level 2 of the model. After completion, the **Variables** screen should look as shown below.

Before the analysis can be run, the model specifications have to be saved to file. To accomplish this, we select the **Save As** option on the **File** menu to open the **Save Mixed Up Model** dialog box and then enter the name **depress.mum** in the **File name** text box to produce the following dialog box.
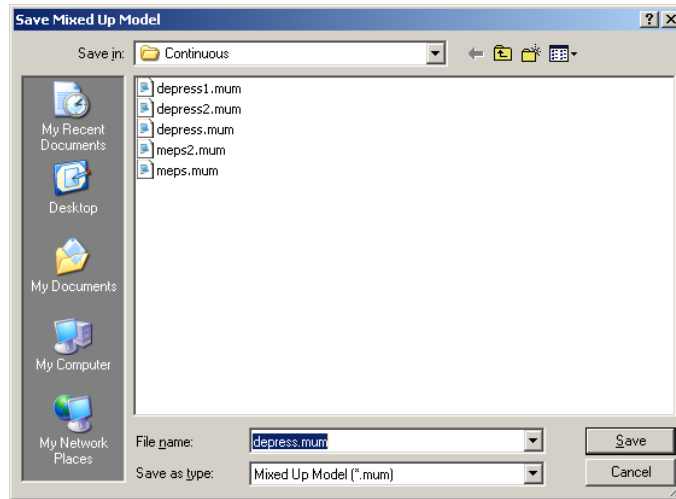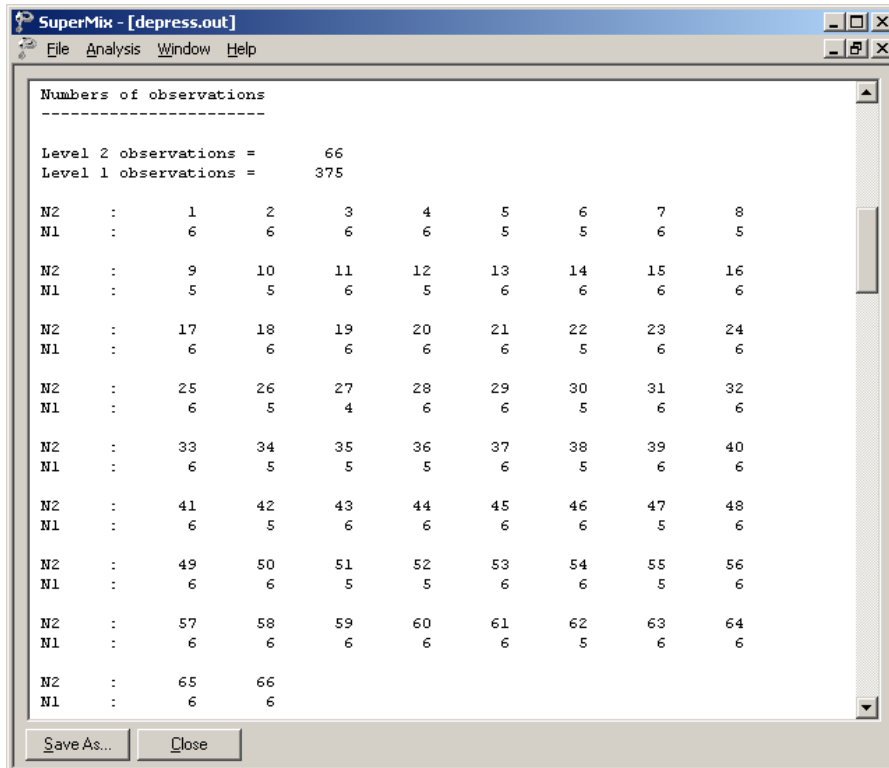
The analysis is run by selecting the **Run** option from the **Analysis** menu as shown below
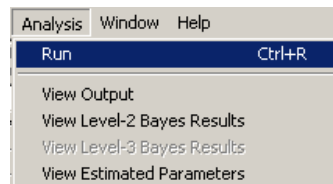
to produce the corresponding output file **depress.out**.

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES

Discussion of results

Four separate sections of the output file **depress.out** are shown below. The summary of the hierarchical structure of the data below, which is given first, shows how the 375 measurements are nested within the 66 patients. It also indicates that the number of repeated measurements per patient varies from 4 to 6 observations.

```
┌─────────────────────────────────────────────────────────────────────────┐
│  SuperMix - [depress.out]                                    _ □ ×        │
│  File  Analysis  Window  Help                                _ ฿ ×        │
├─────────────────────────────────────────────────────────────────────────┤
│                                                                    ▲      │
│  Numbers of observations                                                  │
│  -----------------------                                                  │
│                                                                           │
│  Level 2 observations =      66                                           │
│  Level 1 observations =     375                                           │
│                                                                           │
│  N2    :      1      2      3      4      5      6      7      8           │
│  N1    :      6      6      6      6      5      5      6      5           │
│                                                                           │
│  N2    :      9     10     11     12     13     14     15     16          │
│  N1    :      5      5      6      5      6      6      6      6           │
│                                                                           │
│  N2    :     17     18     19     20     21     22     23     24          │
│  N1    :      6      6      6      6      6      5      6      6           │
│                                                                           │
│  N2    :     25     26     27     28     29     30     31     32          │
│  N1    :      6      5      4      6      6      5      6      6           │
│                                                                           │
│  N2    :     33     34     35     36     37     38     39     40          │
│  N1    :      6      5      5      5      6      5      6      6           │
│                                                                           │
│  N2    :     41     42     43     44     45     46     47     48          │
│  N1    :      6      5      6      6      6      6      5      6           │
│                                                                           │
│  N2    :     49     50     51     52     53     54     55     56          │
│  N1    :      6      6      5      5      6      6      5      6           │
│                                                                           │
│  N2    :     57     58     59     60     61     62     63     64          │
│  N1    :      6      6      6      6      6      5      6      6           │
│                                                                           │
│  N2    :     65     66                                                    │
│  N1    :      6      6                                              ▼      │
├─────────────────────────────────────────────────────────────────────────┤
│   Save As...      Close                                                   │
└─────────────────────────────────────────────────────────────────────────┘
```

The following portion of the output file consists of a listing of selected descriptive statistics of the variables of the model. The descriptive results show, for example, that the mean HDRS score is 17.64 with a standard deviation of 7.19.

```
 SuperMix - [depress.out]                                             _ □ X
  File  Analysis  Window  Help                                         _ 8 X

  Descriptive statistics for all variables
  -----------------------------------------

  Variable           Minimum          Maximum            Mean      Stand. Dev.

  Dependent
  ---------
  HDRS               0.00000         39.00000        17.63733         7.19006

  Random-Effects
  --------------
  intcept  (2)       1.00000          1.00000         1.00000         0.00000
  Week     (2)       0.00000          5.00000         2.48000         1.68320
  intcept  (1)       1.00000          1.00000         1.00000         0.00000

  Fixed Regressor(s)
  ------------------
  intcept            1.00000          1.00000         1.00000         0.00000
  Week               0.00000          5.00000         2.48000         1.68320

  Save As...      Close
```

The next part of the output file contains the starting values for both fixed and random components.

```
 SuperMix - [depress.out]                                             _ □ X
  File  Analysis  Window  Help                                         _ 8 X

  Parameter starting values
  -------------------------


  Fixed regressor(s)
  ------------------

  Variable          Estimate        Std.Err.         Z-value         p-value
  -------           --------        --------         -------         -------
  intercept         23.60263         0.54761        43.10128         0.00000
  Week              -2.40536         0.18278       -13.15986         0.00000

  Log Likelihood              =      -2474.9557
  Number of free parameters   =            6


  Variance/covariance components
  ------------------------------


  Level 2                            Estimate        Std.Err.         Z-value    p-value
  -------                            --------        --------         -------    -------
  intercept      /intercept         12.73430         0.27812        45.78734    0.00000
  Week           /intercept         -1.50873         0.16119        -9.36004    0.00000
  Week           /Week               2.18703         0.18554        11.78759    0.00000

  Level 1                            Estimate        Std.Err.         Z-value    p-value
  -------                            --------        --------         -------    -------
  intercept      /intercept         12.14019         0.09064       133.93833    0.00000

  Save As...      Close
```

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES

The last part of the output shows the final estimates of the fixed and random coefficients included in the model, along with some goodness of fit measures. The results given below show that the p-values for the time effect, as represented by the variable Week, is highly significant. At the beginning of the study, when Week = 0, the average expected HDRS score is 23.57695. For each subsequent week, a decrease of –2.37707 in average HDRS score is expected. At the end of the study period, the average expected HDRS score is $23.57695 - 5(2.37707) = 11.6916$.

The p-values for the estimates of the random coefficients are also significant, with the exception of that for the covariance between the intercept and slope. From the output above we have $\hat{\text{var}}(u_{i0}) = 12.62930$, $\hat{\text{var}}(u_{i1}) = 2.07899$, $\hat{\text{cov}}(u_{i0}, u_{i1}) = -1.42093$, and that $\hat{\text{var}}(e_{ij}) = 12.21663$. It is clear that there is considerably more variation in patients' intercepts than in their slopes (12.62930 vs. 2.07899). This indicates that there are significant differences in the initial HDRS scores, but that the patients' slopes over time do not vary as much. This seems to indicate that the pattern in HDRS scores over time may be similar for patients, although they start with markedly different initial HDRS scores. Typically, one would expect most of the variation in HDRS scores at the measurement level, and thus would expect $\hat{\text{var}}(e_{ij})$ to be larger than any of the other variances/covariances. In this case, however, there is more variation in the random intercepts over patients than in the measurements nested within patients. Due to this, it may be of interest to take a closer look at the variation in HDRS scores at the two levels of the hierarchy.

In the case of a model with only a random intercept, there are two variances of interest: the variation in the random intercept over the patients, and the residual variation at level-1, over the measurements. By calculating the total variation in the HDRS score explained by such a model, obtained as $\hat{\text{var}}(e_{ij}) + \hat{\text{var}}(u_{i0})$, we can obtain an estimate of the intracluster correlation coefficient.

The intracluster coefficient is defined as

$$ICC = \frac{\overset{\wedge}{\mathrm{var}}(u_{i0})}{\overset{\wedge}{\mathrm{var}}(e_{ij}) + \overset{\wedge}{\mathrm{var}}(u_{i0})}$$

and would, for a random intercept model for this data, represent the proportion in HDRS scores between patients.

```
P depress.out                                                    _ □ ×
----------------------------------------
Convergence attained in     5 iterations
----------------------------------------


TITLE: Longitudinal analysis of 66 depressed patients
       Data from Resiby et. al (1977)

Maximum likelihood estimates
----------------------------


Fixed regressor(s)
------------------

 Variable               Estimate      Std.Err.      Z-value      p-value
 --------               --------      --------      -------      -------
 intcept                23.57695      0.54555       43.21714     0.00000
 Week                   -2.37707      0.20865       -11.39280    0.00000

Log Likelihood                  =    -1109.5188
-2 Log Likelihood (Deviance)    =     2219.0375
Akaike's Information Criterion  =     2231.0375
Schwarz's Bayesian Criterion    =     2244.1754
Number of free parameters       =            6


Variance/covariance components
------------------------------


 Level 2                       Estimate      Std.Err.     Z-value    p-value
 -------                       --------      --------     -------    -------
 intcept /intcept              12.62930      3.46653      3.64322    0.00027
 Week    /intcept              -1.42093      1.02595      -1.38500   0.16605
 Week    /Week                  2.07899      0.50417      4.12363    0.00004

 Level 1                       Estimate      Std.Err.     Z-value    p-value
 -------                       --------      --------     -------    -------
 intcept /intcept              12.21663      1.10696      11.03615   0.00000
```

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES

In the current model, which contains both a random intercept and a random slope, the situation is somewhat more complicated. This is due to the possible correlation between the level-2 random effects. When calculating an estimate of the total variation, the covariance(s) between random effects have to be taken into account in any attempt to estimate the proportion of variation in outcome at any level or for any random coefficient. In addition, the inclusion of a covariate such as ENDOG can affect the variance estimates.

The total variation in HDRS scores over patients is defined as

$$\text{Var(level 2)} = \text{var}(u_{i0}) + \text{var}(u_{i1})(\text{WEEK})_{ij}^2 + 2\left[\text{cov}(u_{i0}, u_{i1})\right](\text{Week})_{ij}$$

The total variation is a function of the value assumed by the predictor Week, which has a random slope. As such, the total variation at the beginning of the study is

$$\text{Var(level 2)} = \text{var}(u_{i0}) + \text{var}(u_{i1})(0)^2 + 2\left[\text{cov}(u_{i0}, u_{i1})\right](0)$$
$$= \text{var}(u_{i0})$$

while at the end of the study we have

$$\text{Var(level 2)} = \text{var}(u_{i0}) + \text{var}(u_{i1})(5)^2 + 2\left[\text{cov}(u_{i0}, u_{i1})\right](5)$$
$$= \text{var}(u_{i0}) + 25\,\text{var}(u_{i1}) + 10\,\text{cov}(u_{i0}, u_{i1})$$

An estimate of the total variation at this level can be obtained by using the estimates of the variances and covariance obtained under this model. By substituting $\hat{\text{var}}(u_{i0})$, $\hat{\text{var}}(u_{i1})$, and $\hat{\text{cov}}(u_{i0}, u_{i1})$ into the equations above, we obtain the estimated variation in HDRS scores over patients at different points during the study period.

At the beginning of the study, the estimated total variation in HDRS scores over patients is simply the estimated variation in the random intercept, *i.e.*, $\hat{var}(u_{i0}) = 12.62930$. At the end of the study, the total variation at level-2 is estimated as

$$\hat{var}(\text{level 2}) = \hat{var}(u_{i0}) + 25\,\hat{var}(u_{i1}) + 10\,\hat{cov}(u_{i0}, u_{i1})$$
$$= 12.62930 + 25(2.07899) + 10(-1.42093)$$
$$= 50.39475.$$

At the beginning of the study we obtain

$$\frac{\hat{var}(\text{level 2})}{\hat{var}(\text{level 2}) + \hat{var}(\text{level 1})} = \frac{12.62930}{12.62930 + 12.21663}$$
$$= 0.5083$$

and thus conclude that that 50.8% of the variation in HDRS scores at this time is over patients. At the end of the study, we find that

$$\frac{\hat{var}(\text{level 2})}{\hat{var}(\text{level 2}) + \hat{var}(\text{level 1})} = \frac{50.39475}{50.39475 + 12.21663}$$
$$= 0.8049,$$

so that only 20% of the variation in HDRS scores are estimated to be at the measurement level, with 80% at the patient level. As mentioned before, the total variation in HDRS scores is a function of the time of measurement, as represented by the variable Week. The very different estimates of variation at a patient level show how the introduction of an important predictor, in this case at the measurement

level, can have an impact on variance estimates at a different level of the hierarchy. By the end of the study period, the residual variation over measurements has been dramatically reduced, this being explained largely by the inclusion of the time effect. Most of the remaining unexplained variation is at the patient level. As a result of this finding and in the light of our original research question, whether the initial depression classification of a patient is also related to the HDRS scores over the time in which medication is administered, the model will be extended to include the covariate ENDOG. This dichotomous variable assumes a value of 1 when endogenous depression was observed, and 0 if not. In addition, we will make provision for a possible interaction between depression classification and the measurement occasion by including the interaction term WxENDOG in the model. While WxENDOG can be viewed as a cross-level interaction, as Week is a measurement-level variable and ENDOG a patient-level variable, the inclusion of the patient-level variable ENDOG may enable us to explain more of the remaining variation in the random intercepts and slopes at the patient level.

3.2.4 Example: A random intercept and slope model with covariate and interaction effect

To fit the random intercept and slope model with a covariate and an interaction effect for HDRS, we need to include the variables ENDOG and WxENDOG as covariates in the model. Significant ENDOG and WxENDOG coefficients will imply that there are significant differences in depression value ratings between the two groups of patients, or that there is significant interaction between the type of depression and the time at which a measurement was made. One way to decide whether ENDOG and WxENDOG should be included is to fit the model with and without these predictors to the data. Then the difference in the deviance (–2 log-likelihood) values can be used to obtain a χ^2-square test statistic value. The difference in the number of parameters estimated in each of these models gives the corresponding degrees of freedom.

Setting up the analysis

To create the model specifications for this model, we start by opening **reisby.ss3** in a SuperMix spreadsheet window. Then we use the **Open Existing Model Setup** option on the **File** menu to open the **Model Setup** window for **depress.mum**. We extend the string in the **Title 1** text box on the **Configuration** screen by adding the string "+ ENDOG". Since we would like to produce the Bayes estimates for the means, we select the **means & (co)variances** option from the drop-down list box next to **Write Bayes Estimates** to produce the following screen.

Next, click on the **Variables** tab to proceed to the **Variables** screen of the **Model Setup** window. The two covariates are specified by checking the **E** check boxes for ENDOG and WxENDOG in the **Available** grid respectively to produce the following **Variables** screen.

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES
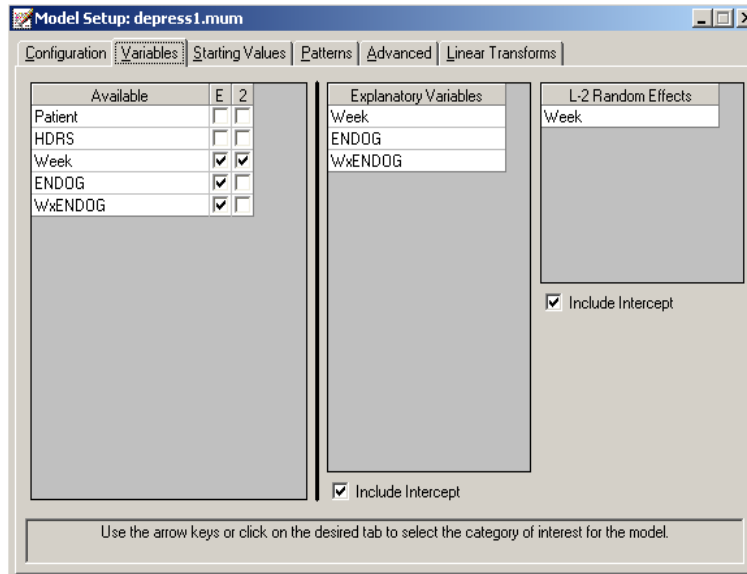
Save the changes to the file **depress1.mum** by using the **Save As** option on the **File** menu. To fit the revised model to the data, select the **Run** option on the **Analysis** menu to produce the output file **depress1.out**.

Discussion of results

A portion of the output file **depress1.out** is shown below. The interaction WxENDOG between the time variable Week and the depression classification of the patient, as represented by the variable ENDOG, is not significant. Given this, we can take a closer look at the estimated coefficients for the main effects Week and ENDOG respectively. Note, however, that the *p*-value for the ENDOG coefficient is larger than 0.05, and thus can only be considered significant at a 10% level of significance. The effect of time, on the other hand, is found to be highly significant. While the average HDRS score is predicted to decrease by –2.37 score scale units each week, patients classified as having endogenous depression (*i.e.*, ENDOG = 1) are predicted to have a HDRS score of 2 units higher at all occasions.

```
o=========================================================o
| Longitudinal analysis of 66 depressed inpatients + ENDOG |
| Data from Reisby et. al (1977)                            |
o=========================================================o

Maximum likelihood estimates
----------------------------


Fixed regressor(s)
------------------

 Variable            Estimate      Std.Err.      Z-value       p-value
 --------            --------      --------      -------       -------
 intcept             22.47626      0.79435       28.29524      0.00000
 Week                -2.36569      0.31181       -7.58693      0.00000
 ENDOG               1.98802       1.06905       1.85961       0.06294
 WxENDOG             -0.02706      0.41947       -0.06450      0.94857

 Log Likelihood                 =    -1107.4646
 -2 Log Likelihood (Deviance)   =     2214.9292
 Akaike's Information Criterion =     2230.9292
 Schwarz's Bayesian Criterion   =     2248.4465
 Number of free parameters      =            8




Variance/covariance components
------------------------------


 Level 2                    Estimate    Std.Err.    Z-value    p-value
 -------                    --------    --------    -------    -------
 intcept /intcept           11.64121    3.29646     3.53142    0.00041
 Week    /intcept           -1.40161    1.00338     -1.39689   0.16245
 Week    /Week              2.07707     0.50380     4.12283    0.00004

 Level 1                    Estimate    Std.Err.    Z-value    p-value
 -------                    --------    --------    -------    -------
 intcept /intcept           12.21847    1.10707     11.03673   0.00000
```

This is clear from a plot of the predicted HDRS scores over time for the two ENDOG groups, as shown below in Figure 3.1. For all practical purposes, the two lines are parallel to each other, again underscoring the absence of significant interaction between Week and ENDOG.

To obtain the predicted average HDRS scores as shown in these plots, the estimates obtained from the output are used:

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1}(\text{Week}) + \hat{\beta_2}(\text{ENDOG}) + \hat{\beta_3}(\text{WxENDOG})$$
$$= 22.47626 - 2.36569(\text{Week}) + 1.98802(\text{ENDOG}) - 0.02706(\text{WxENDOG})$$
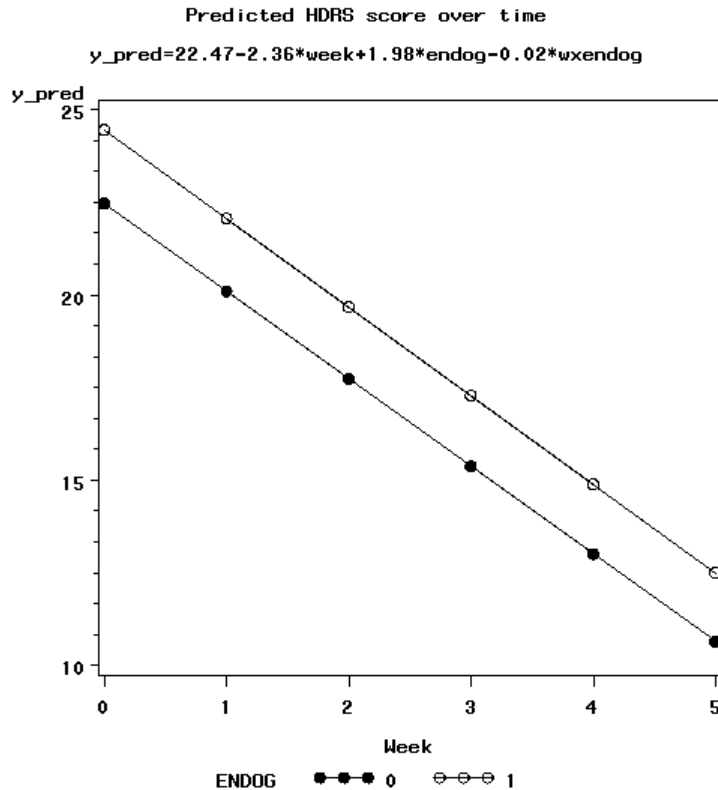
Figure 3.1: Predicted HDRS scores over time for two groups

To illustrate differences between significant and nonsignificant interaction, we examine two scenarios. In the first, suppose that a negative and significant estimate of -1.02 was obtained for the interaction between Week and ENDOG. In this case, the graph (see Figure 3.2) would have looked somewhat different, as shown below. Under this scenario, patients with endogenous depression would have improved faster than their counterparts with non-endogenous depression.
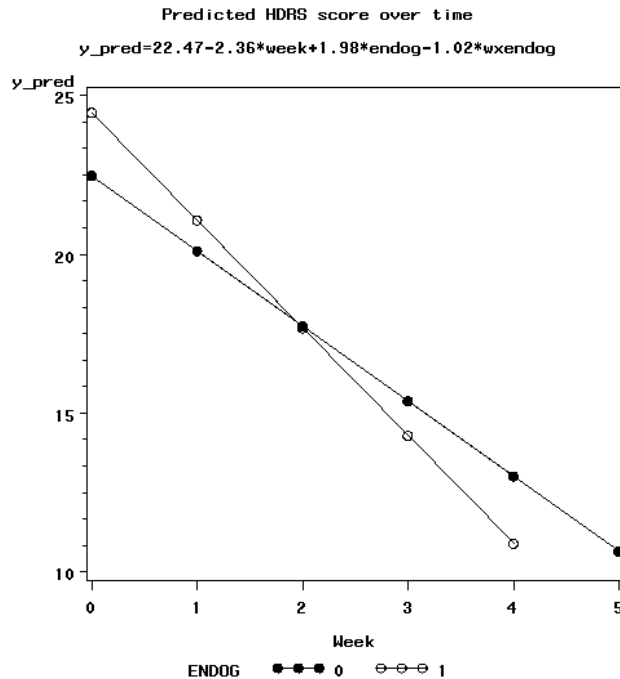
Figure 3.2: Predicted HDRS scores over time for two groups

For the second scenario, suppose that a significant and positive interaction term of 1.02 was obtained. In this case, patients with non-endogenous depression would have shown more marked improvement over time than the patients with endogenous depression. This scenario is illustrated in Figure 3.3.

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES
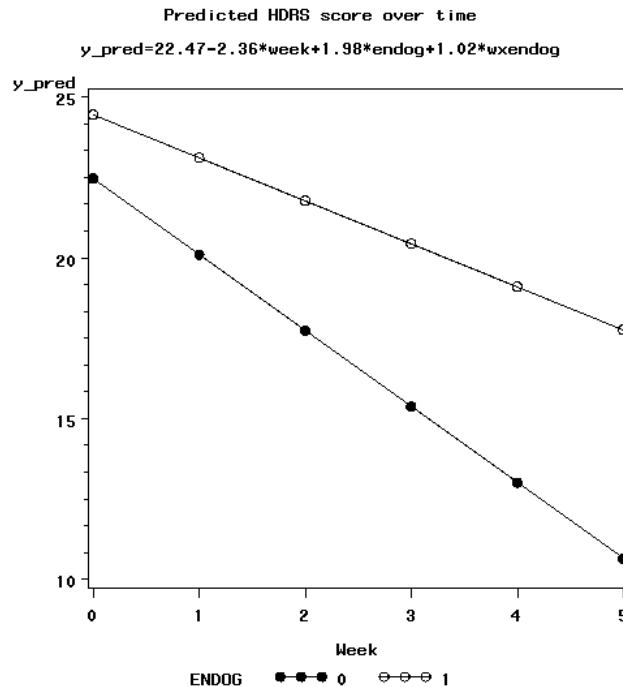
Figure 3.3: Predicted HDRS scores over time for two groups

A question that arises from inspection of the results is whether the interaction term contributes overall to the explanation of the variation in the HDRS scores. To test this, we can fit a model without the interaction term and use the deviance reported in the output to compare results for the model with interaction and the model without this term. The relevant output (**depress2.out**) from an analysis without the interaction term is shown below. We note that the deviance obtained for the simpler model is almost identical to that of the model considered in this section. Based on this, we conclude that a model without the interaction WxENDOG would fit the data as well as the one with the interaction term included.

```
SuperMix - [depress2.out]                                          _ □ ×
  File  Analysis  Window  Help                                      _ 8 ×

  Fixed regressor(s)
  ------------------

    Variable              Estimate       Std.Err.       Z-value       p-value
    --------              --------       --------       -------       -------
    intcept               22.49344       0.74839        30.05592      0.00000
    Week                  -2.38064       0.20859       -11.41317      0.00000
    ENDOG                  1.95650       0.95083         2.05769      0.03962

  Log Likelihood               =    -1107.4667
  -2 Log Likelihood (Deviance) =     2214.9334
  Akaike's Information Criterion =   2228.9334
  Schwarz's Bayesian Criterion =     2244.2610
  Number of free parameters    =             7
```
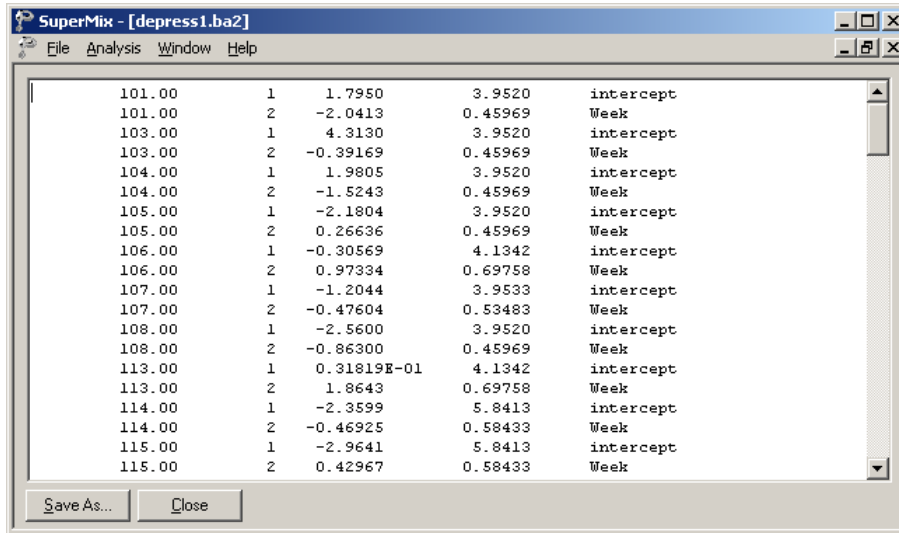
In addition, we can test the hypothesis that the model with covariate (ENDOG) fits the data better than the random intercept and slope model considered previously. To test this hypothesis, we calculate the difference between the –2 log likelihood value obtained for the previous model (**depress1.out**) and the –2 log likelihood value for the current model (**depress2.out**). It can be shown that this difference of 2219.04 – 2214.93 = 4.11 has a χ^2 distribution with associated degrees of freedom equal to the difference in the number of parameters estimated in the two examples, *i.e.*, 8 – 7 = 1 degree of freedom. Since the *p*-value for this test statistic is less than 0.05, it is concluded that the random intercept and slope model with ENDOG as a covariate does not provide a better description of the data than the original random intercept and slope model. This finding is supported by the fact that the *p*-value for ENDOG when the interaction effect between Week and ENDOG is excluded.

Residual analysis

Up to this point, we have considered results averaged over all patients. We now turn our attention to the residual file **depress1.ba2**, which offers the opportunity to take a closer look at the results by individual patient. In the image below, the contents of this file are displayed for the first 10 patients.

Two lines of information are given for each patient, containing, in order of appearance,

- o the patient ID,
- o the number of the empirical Bayes coefficient,
- o the empirical Bayes estimate,
- o the estimated variance of the Bayes coefficient, and
- o the name of the associated coefficient as used in the model.

To obtain patient-specific predicted HDRS scores the empirical Bayes estimates for each patient have to be taken into account, as these estimates indicate the extent to which the random intercept or slope for that patient deviates from the intercept and slope over all patients. Patient-specific predicted HDRS scores are calculated as

$$\hat{y}\,|\,\mathbf{u}_i = 22.47626 - 2.36569(\text{Week}) + 1.98802(\text{ENDOG})$$

$$-0.02706(\text{WxENDOG}) + \hat{u}_{i0} + \hat{u}_{i1}(\text{Week})$$

For the first patient shown in the residual file above, we have $\hat{u}_{i0}$ = 1.7950 and $\hat{u}_{i1}$ = –2.0413. From this information, we can already tell that the intercept for the patient is higher than average, but that the Week slope for this patient is lower than average. This patient was classified as having non-endogenous depression, so that ENDOG = 0. At week 1, the predicted HDRS score for this patient (Patient = 101) is found to be

$$\hat{y} = 22.47626 - 2.36569(1) + 1.7950 - 2.0413(1)$$
$$= 19.8643$$

and at the end of the study period (Week = 5) the predicted HDRS score for Patient 101 is

$$\hat{y} = 22.47626 - 2.36569(5) + 1.7950 - 2.0413(5)$$
$$= 2.3631.$$

Figures 3.4 and 3.5 are graphical representations of the average and patient-specific regression lines of six patients, three of which were classified as non-endogenous and three as endogenous. The graphs on the left show the patient-specific regression lines (solid lines) and the observed HDRS trajectories (the dotted lines) for each of the patients. The graphs on the right show the average regression line (solid) and the patient-specific regression line (dotted).

For the patients with non-endogenous depression (Figure 3.4; graphs on the right), all the average regression lines are identical. Similarly, the average regression lines for patients with endogenous depression (Figure 3.5; graphs on the right) are identical. For the latter case, the average regression lines are higher. Recall that for the predictor ENDOG an estimated coefficient of 1.98802 was obtained. A patient classified as having endogenous depression is thus expected to have a higher average HDRS score than a patient with non-endogenous depression.

Regression line and observed trajectory
Patient=101
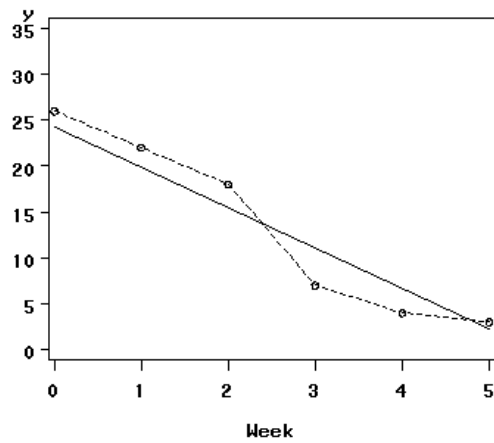
Patient and average regression lines
Patient=101
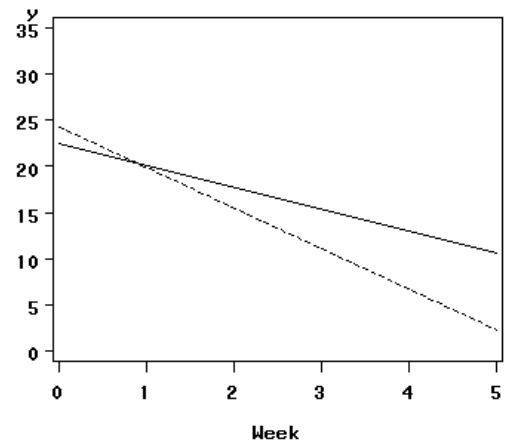
Regression line and observed trajectory
Patient=103

Patient and average regression lines
Patient=103

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES
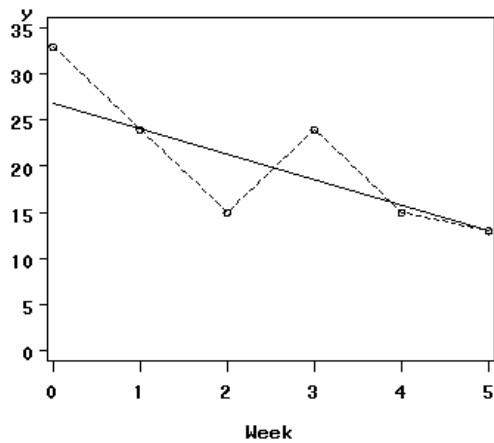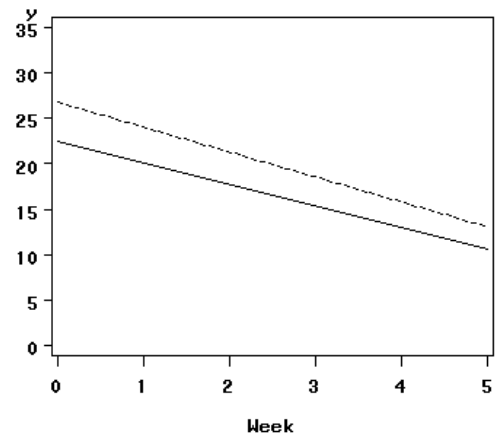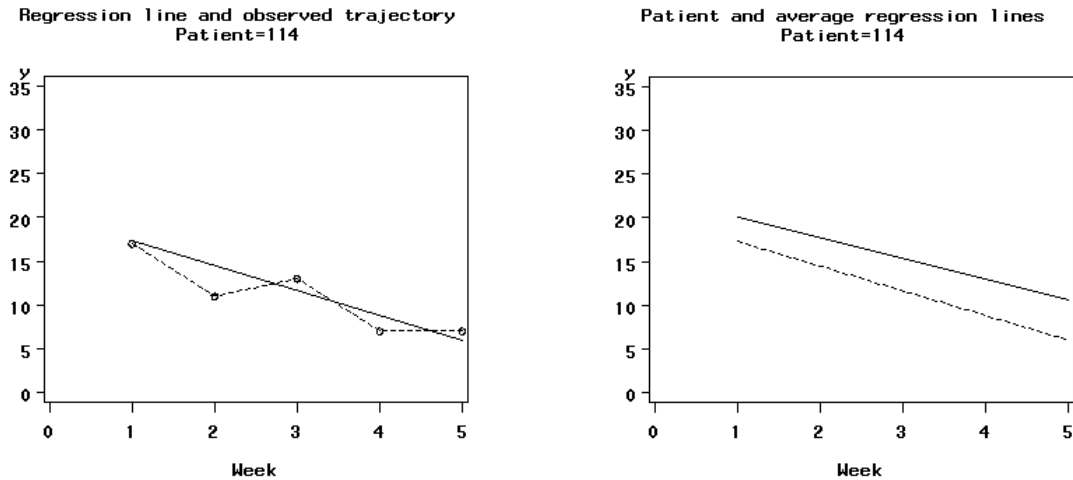
Figure 3.4: Average and unit-specific regression lines for patient with non-endogenous depression

From the top right graph in Figure 3.4 we find that patient number 101 had a higher initial HDRS score, but over time obtained a lower than average score. For patient 103, a higher than average predicted HDRS score is obtained at each time point, as illustrated in the second graph on the right. In contrast, patient 114 scored lower at each time point. Similar patterns for patients with endogenous depression are present in the graphs in Figure 3.5. Patients 104, 106 and 108 were classified as having endogenous depression. In the case of patient 108, the predicted average regression line shows a consistently higher predicted HDRS score over time when these scores are compared to the predicted average regression line for the non-endogenous patients. The observed trajectories of patients 101, 114, 106 and 108 follow their predicted patient-specific regression line more closely than is the case for patients 103 and 104.

TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES

Regression line and observed trajectory
Patient=104
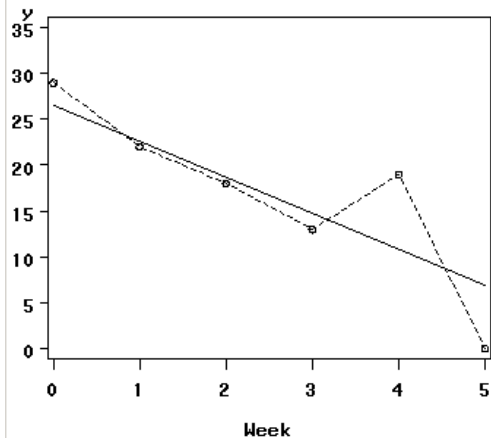
Patient and average regression lines
Patient=104
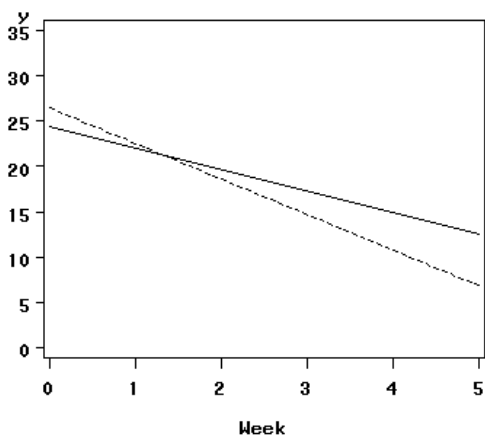
Regression line and observed trajectory
Patient=106
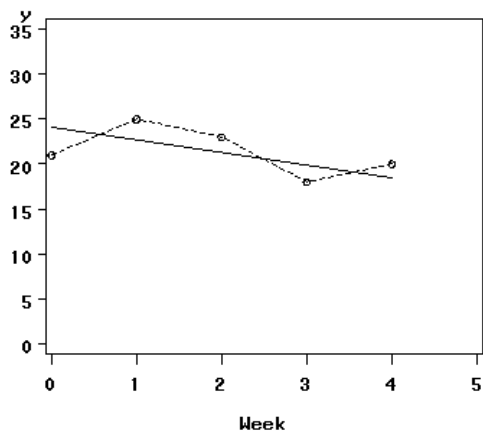
Patient and average regression lines
Patient=106
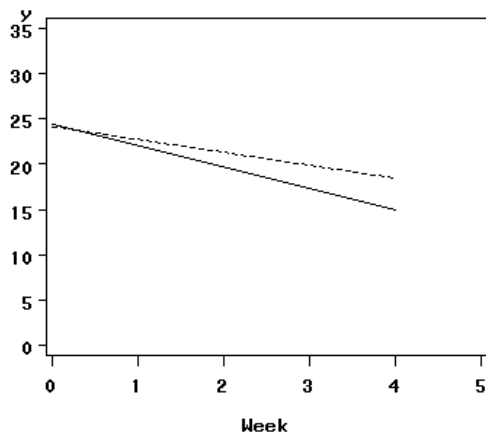
TWO-LEVEL MODELS FOR CONTINUOUS OUTCOMES   123
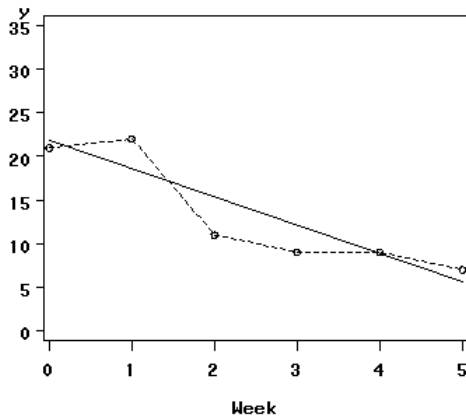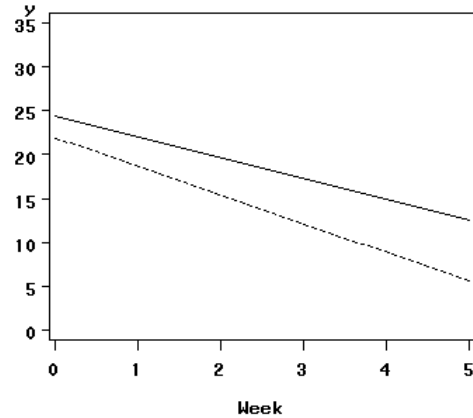
Figure 3.5: Average and unit-specific regression lines for patient with endogenous depression

Level-1 residuals can also be obtained, either for a typical or specific patient, by using the empirical Bayes estimates. The residuals for a typical patient are obtained as

$$
\begin{aligned}
\text{Average residual} &= \text{Observed HDRS score} - \hat{y} \\
&= \text{Observed HDRS score} \\
&\quad - \left[22.47626 - 2.36569(\text{Week}) + 1.98802(\text{ENDOG}) \right] - 0.02706(\text{WxENDOG})
\end{aligned}
$$

The residuals for a specific patient use the additional information given by the empirical Bayes residuals and have the form

Patient-specific residual $= $ Observed HDRS score $- \hat{y} \mid \mathbf{u}_i$

$= $ Observed HDRS score $-$

$$\left[22.47626 - 2.36569(\text{Week}) + 1.98802(\text{ENDOG}) - 0.02706(\text{WxENDOG}) + \hat{u}_{i0} + \hat{u}_{i1}(\text{Week}) \right]$$
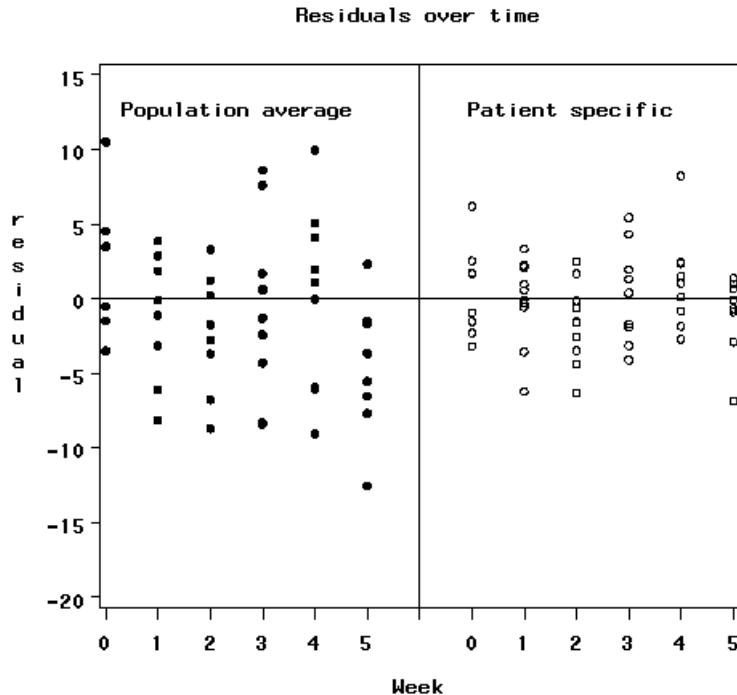
Figure 3.6: Comparison of population average and Bayes residuals for first 10 patients

Inspection of these estimates can be useful in examining the distributional assumptions for the level-1 data, in this case at the measurement level. For the current example, residuals for a typical patient have a mean of -0.022 and range between -15.3792 and 20.4995. The residuals for specific patients have a mean of 0.000 with a minimum value of -9.5942 and a maximum value of 12.6566. The range of the latter is much smaller. Empirical Bayes estimates are frequently referred to as "shrunken" estimates, as the empirical Bayes tend to pull the estimates

closer to the sample average, thus "shrinking" them. The amount of shrinkage is a function of the precision of the intercept/slope estimates. The greater the precision for an estimate, the less shrinkage. This means that more shrinkage will occur for units where there is more uncertainty concerning the accuracy of the fixed intercept and slope regression estimates. Figure 3.6 shows a comparison of the two types of residuals for our example. The Bayes residuals are much closer to the horizontal reference line at 0. Note the shrinking of the more extreme population average residuals in the left pane.

An alternative way to display the Bayes residuals is to plot 95% confidence intervals for patient intercepts and slopes, as shown in Figures 3.7 and 3.8. Information for the first 10 patients in the study were used to construct these graphs.

Confidence intervals were obtained as $mean \pm 1.96(std.dev)$. For the intercepts, the means are computed as $\hat{\beta}_0 + \hat{u}_{0j}$, $j = 1, 2, ..., 10$. Standard deviations are computed as the square roots of the variances of the empirical Bayes intercept residuals given in the **ba2** file shown earlier. For patient 1, for example,

$$\hat{u}_{01} = 1.7950, \quad var\left(\hat{u}_{01}\right) = 3.9520$$

$$\hat{u}_{11} = -2.0413, \quad var\left(\hat{u}_{11}\right) = 0.45969.$$

Therefore, the 95% confidence intervals for patient 1 are:

$$intcept : (22.47626 + 1.7950) \pm 1.96 \times (3.9520)^{1/2}$$
$$= (20.3749; 28.1677)$$

$$slope : (-2.36569 - 2.0413) \pm 1.96 \times (0.45969)^{1/2}$$
$$= (-5.7359; -3.0781).$$

In Figures 3.7 and 3.8, each mean is represented by a square. Figure 3.8 shows that the individual slopes for the first 10 patients are all negative.

95% confidence intervals for patient intercepts
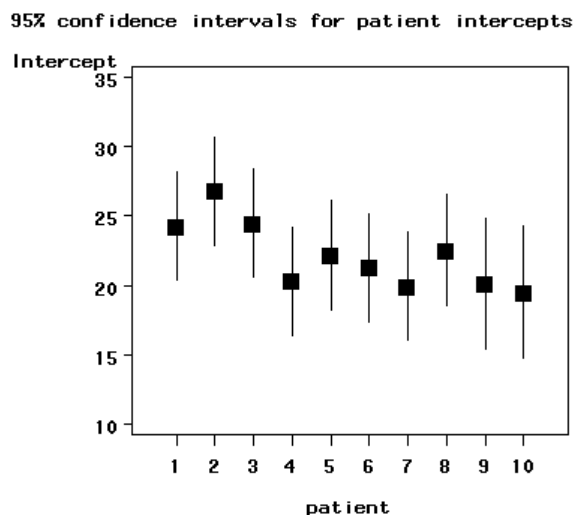
Figure 3.7: 95% confidence intervals for patient intercepts

95% confidence intervals for patient slopes
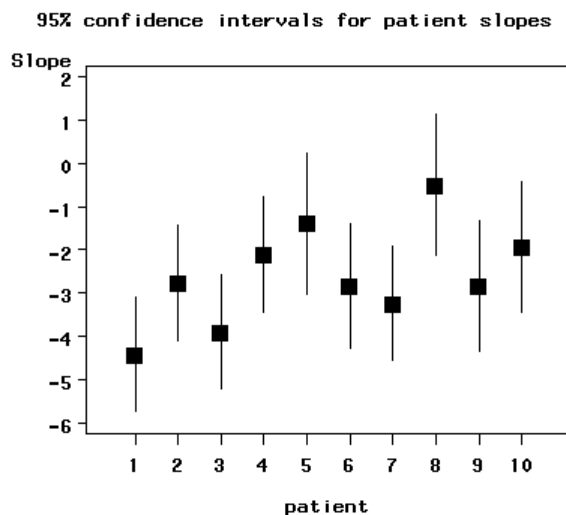
Figure 3.8: 95% confidence intervals for patient slopes

Graphical displays

It is possible to obtain a great variety of graphical displays with SuperMix. To invoke the graphics procedure, open a SuperMix data file. To illustrate this, we use **reisby.ss3** in the **Continuous** subfolder. The next step is to select the **Data-based Graphs**, **Exploratory** option on the **File** menu as shown below
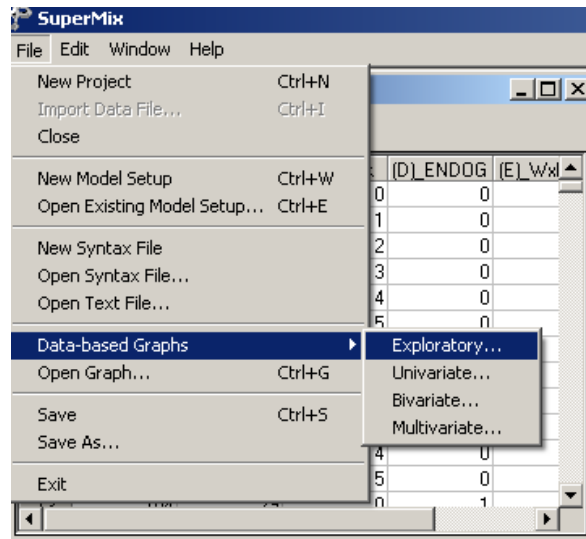
to activate the **New Graph** dialog box. Specify HDRS as the dependent (vertical axis) variable by selecting it from the **Y** drop-down list box and Week as the independent (horizontal axis) variable by selecting it from the **X** drop-down list box. A graph on the same axes-system is created for each patient by selecting the variable Patient from the **Overlay** drop-down list box. Furthermore, each graph is assigned a color by selecting ENDOG from the **Color** drop-down list box to produce the following **New Graph** dialog box.

Click on the **OK** button to produce the following graph of the reaction trajectories over time for the 66 patients.
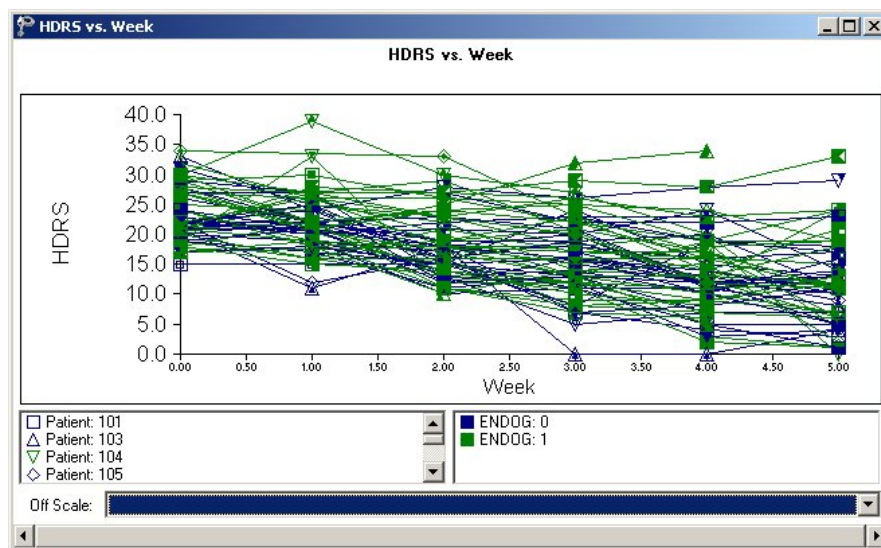
To modify the existing graphic displays, select the **Edit Graph** option from the **Settings** menu to open the **Edit Graph** dialog box. To obtain different graphs for the two categories of the covariate ENDOG, select it from the **Filter** drop-down list box to produce the following **Edit Graph** dialog box.
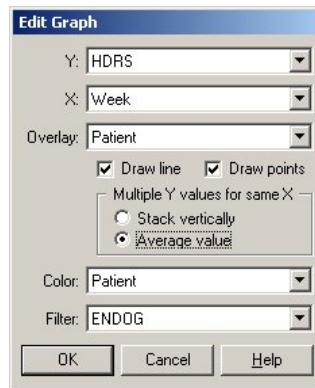
Click on the **OK** button to open the following graphics window.
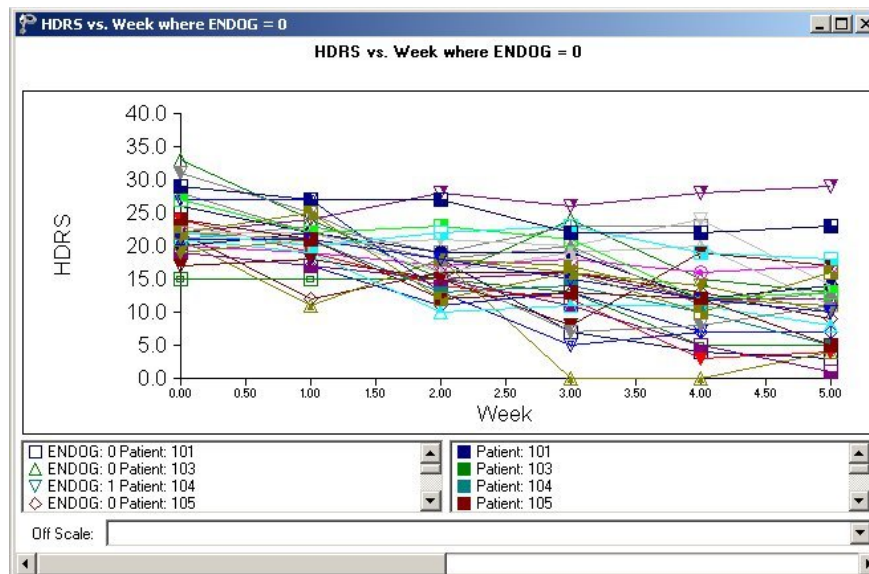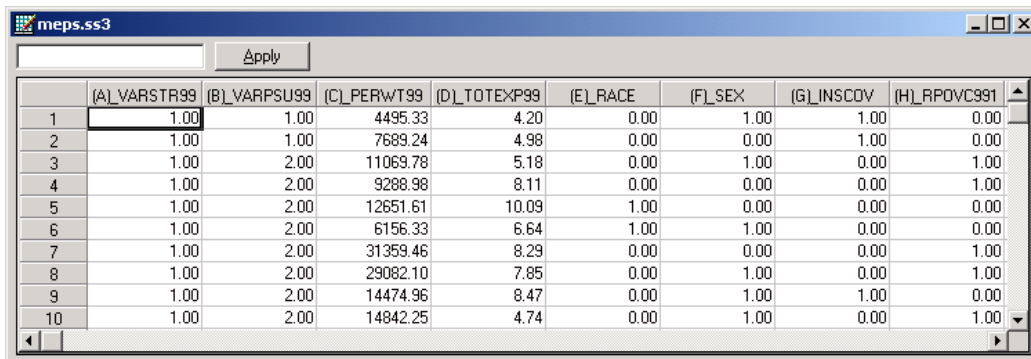
At the bottom of the graphics window is a "slider" with left and right arrows. By clicking on the right arrow, one can obtain the next graphic shown below and by clicking on the left arrow, the graphic above.
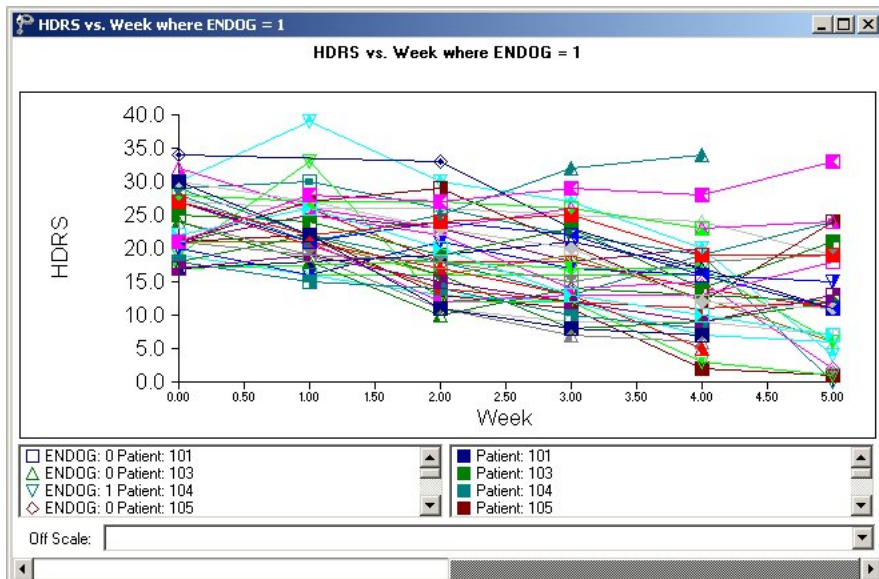
3.3 Three-level models for continuous outcomes

3.3.1 The data

The data set used here forms part of the data library of the Medical Expenditure Panel Survey (MEPS). Collected in 1999, these data from a longitudinal national survey were used to obtain regional and national estimates of health care use and expenditure based on the health expenditures of a sample of U.S. civilian non-institutionalized participants. The survey sample design utilized stratification, clustering, multiple stages of selection, and disproportionate sampling. The sample was drawn from 143 strata, divided into 460 primary sampling units (PSUs). Information on 23,565 participants included positive person-level weights and forms the data set used here, excluding the 1,053 participants in the original data with zero person-level weights. Further exclusion, of respondents with no total expenditure reported, reduced the size of the data set to 19300. Data for the first 10 participants on most of the variables used in this section are shown below in the form of a SuperMix spreadsheet file, named **meps.ss3**, which is saved in the **Examples\Primer\Continuous** folder.

	(A)_VARSTR99	(B)_VARPSU99	(C)_PERWT99	(D)_TOTEXP99	(E)_RACE	(F)_SEX	(G)_INSCOV	(H)_RPOVC991
1	1.00	1.00	4495.33	4.20	0.00	1.00	1.00	0.00
2	1.00	1.00	7689.24	4.98	0.00	0.00	1.00	0.00
3	1.00	2.00	11069.78	5.18	0.00	1.00	0.00	1.00
4	1.00	2.00	9288.98	8.11	0.00	0.00	0.00	1.00
5	1.00	2.00	12651.61	10.09	1.00	0.00	0.00	0.00
6	1.00	2.00	6156.33	6.64	1.00	1.00	0.00	0.00
7	1.00	2.00	31359.46	8.29	0.00	0.00	0.00	1.00
8	1.00	2.00	29082.10	7.85	0.00	1.00	0.00	1.00
9	1.00	2.00	14474.96	8.47	0.00	1.00	1.00	0.00
10	1.00	2.00	14842.25	4.74	0.00	1.00	0.00	1.00

The variables of interest are:

- o VARSTR99 is the stratum identification variable (143 strata in total).
- o VARPSU99 is the PSU identification variable (460 PSUs in total).

o PERWT99F represents the final sample weight, with weights ranging between 307.16 and 80061.61, correcting for both non-response and adjustments to population control totals from the Current Population Survey.

o TOTEXP99 is the natural logarithm of the total health expenditure of a respondent in 1999, ranging between 0.69 and 12.24 and representing actual expenditure of between $1.99 and $206,721.

o RACE is an ethnicity indicator, with a value of 1 indicating white respondents, and 0 denoting all other ethnic groups as well as respondents for which ethnicity is not known. This variable was recoded from the original MEPS variable RACEX.

o SEX is a gender indicator, with a value of 0 indicating a male participant and 1 a female participant; recoded from the original MEPS variable RSEX.

o INSCOV is an indicator of the level of insurance coverage, where 0 indicates private coverage any time during 1999, and 1 indicates public coverage or no insurance at all during 1999.

o RPOVC991 to RPOVC995 are five indicator variables, each associated with a category of the original MEPS variable RPOVC99 which was constructed by dividing family income by the applicable poverty line (selection of which depended on family size and composition), expressed as a percentage.

Income is a variable that is often transformed using its natural logarithm. Doing so in effect causes the impact of each additional dollar to decrease as income increases. Logarithmic transformation is also useful in lessening the influence of outliers, as the natural logarithm of a variable is much less sensitive to extreme observations than is the variable itself.

The original MEPS variable RPOVC99 assumed a value of 1 for a family with "high" income level where family income was equal to or greater than 400% of the applicable poverty line, and a value of 2 for those with a "low income" level (associated with 125% to 200% of the poverty line). Families with "middle income", "near poor" and "negative or poor" levels of income relative to poverty line income were coded 3, 4, and 5 respectively. For the "middle income" category, the ratio (as percentage) of family income to poverty line was 200% to less than 400%. In the

case of "near poor" families, the percentages ranged between 100% and 125%, and for "negative or poor", the family income was less than 100% of the relevant poverty line. Thus, a value of 1 on the indicator variable RPOVC991 indicates a family with income at the "high" level. The variables RPOVC992, RPOVC993, and RPOVC994 are associated with the categories "low income", "middle income" and "near poor" respectively. A value of 1 on the variable RPOVC995 indicates a family with "negative or poor" income level.

Note that as each of the five indicator variables for categories of RPOVC99 is coded 1 if a participant responded in that category and 0 otherwise, only four of the five indicator variables can be used in a model where an intercept is included. Here, we opted to create them prior to analysis, as illustration of that feature is not relevant to the example at hand.

3.3.2 The models

A multilevel model does not make provision for the specification of design related variables such as stratum or PSU. Instead, these design variables are used to define the hierarchical structure of the data. In this example, the stratum identification variable VARSTR99 is used as the level-3 identifier and the PSU identification variable VARPSU99 serves to identify level-2 units (*i.e.*, PSUs) nested within a given stratum. We thus use the design variables to define a three-level hierarchical structure, with participants as level-1 observations nested within PSUs, in turn nested within strata. While not explicitly acknowledging the survey design or offering a conventional design effect estimate to measure the difference in estimates obtained when implementing this design compared to estimates obtained under a simple random sample, a multilevel model offers the advantage of estimating the variation in total health care expenditure within and between PSUs.

A general three-level model for a response variable y depending on a set of r predictors $x_1, x_2, \ldots, x_r$ can be written in the form

$$y_{ijk} = \mathbf{x}'_{(f)ijk}\boldsymbol{\beta} + \mathbf{x}'_{(3)ijk}\mathbf{v}_i + \mathbf{x}'_{(2)ijk}\mathbf{u}_{ij} + \mathbf{x}'_{(1)ijk}\mathbf{e}_{ijk}$$

where $i = 1, 2, \ldots, N$ denotes the level-3 units, $j = 1, 2, \ldots, n_i$ the level-2 units, and $k = 1, 2, \ldots, n_{ij}$ the level-1 units. In this context, y_{ijk} represents the response of individual k, nested within level-2 unit j and level-3 unit i. The model shown here consists of a fixed and a random part. The fixed part of the model is represented by the vector product $\mathbf{x}'_{(f)ijk}\boldsymbol{\beta}$, where $\mathbf{x}'_{(f)ijk}$ is a typical row of the design matrix of the fixed part of the model with, as elements, a subset of the r predictors. The vector $\boldsymbol{\beta}$ contains the fixed, but unknown parameters to be estimated. The vector products $\mathbf{x}'_{(3)ijk}\mathbf{v}_i$, $\mathbf{x}'_{(2)ijk}\mathbf{u}_{ij}$, and $\mathbf{x}'_{(1)ijk}\mathbf{e}_{ijk}$ denote the random part of the model at levels 3, 2, and 1 respectively. For example, $\mathbf{x}'_{(3)ijk}$ represents a typical row of the design matrix of the random part at level 3, and $\mathbf{v}_i$ the vector of random level-3 coefficients to be estimated. The products $\mathbf{x}'_{(2)ijk}\mathbf{u}_{ij}$ and $\mathbf{x}'_{(1)ijk}\mathbf{e}_{ijk}$ serve the same purpose at levels 2 and 1 respectively. It is assumed that $\mathbf{v}_1, \mathbf{v}_2, \mathsf{K}, \mathbf{v}_N$ are independently and identically distributed (i.i.d.) with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(3)}$. Similarly, $\mathbf{u}_{i1}, \mathbf{u}_{i2}, \mathsf{K}, \mathbf{u}_{in_i}$ are assumed i.i.d., with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(2)}$, and $\mathbf{e}_{ij1}, \mathbf{e}_{ij2}, \mathsf{K}, \mathbf{e}_{ijn_{ij}}$ are assumed i.i.d., with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(1)}$.

A random intercept model with 7 predictors

Within this hierarchical framework, the first model fitted to the data uses the participant's gender, ethnicity, type of health insurance coverage, and measure of income relative to poverty level to predict the total expenditure on health care in 1999, with expenditure transformed to the natural logarithm of the actual expenses incurred.

$$\begin{aligned} \text{TOTEXP99}_{ijk} = \beta_0 &+ \beta_1 * \text{SEX}_{ijk} + \beta_2 * \text{RACE}_{ijk} + \beta_3 * \text{INSCOV}_{ijk} + \\ &\beta_4 * \text{RPOVC991}_{ijk} + \beta_5 * \text{RPOVC992}_{ijk} + \beta_6 * \text{RPOVC993}_{ijk} + \\ &\beta_7 * \text{RPOVC994}_{ijk} + v_{i0} + u_{ij0} + e_{ijk} \end{aligned}$$

where β_0 denotes the average expected total expenditure on health care in 1999, and $\beta_1, \beta_2, \text{K}, \beta_7$ indicate the estimated coefficients associated with the fixed part of the model which contains the predictor variables SEX, RACE, INSCOV and the indicator variables for categories of income relative to the poverty level. The random part of the model is represented by v_{i0}, u_{ij0} and e_{ijk}, which denote the variation in average total health related expenditure over strata, between PSUs (or, in other words, over PSUs nested within strata) and between participants at the lowest level of the hierarchy.

A random intercept model with 3 predictors

To illustrate model comparison, a simpler model was fitted to the same data. In the previous model, the estimated coefficients of only 2 of the 4 indicator variables representing a respondent's position relative to the poverty line were significant. In the second model, only the ethnicity, gender, and level of insurance coverage were included as explanatory variables. The simplified model is formulated as

$$\text{TOTEXP99}_{ijk} = \beta_0 + \beta_1 * \text{SEX}_{ijk} + \beta_2 * \text{RACE}_{ijk} + \beta_3 * \text{INSCOV}_{ijk} + v_{i0} + u_{ij0} + e_{ijk}$$

where β_0 denotes the average expected total expenditure on health care in 1999, and β_1, β_2 and β_3 indicate the estimated coefficients associated with the fixed part of the model which contains the predictor variables SEX, RACE, and INSCOV. The random part of the model is represented by v_{i0}, u_{ij0} and e_{ijk} as previously explained.

3.3.3 Example: A random intercept model with 7 predictors

Importing the data

The model is fitted to the data in **meps.ss3**. The first step is to create the **ss3** file from an Excel spreadsheet named **meps.xls**. This is accomplished as follows:

- o Use the **File**, **Import Data File** option to activate the display of an **Open** dialog box.
- o Browse for the file **meps.xls** in the **Examples\Primer\Continuous** folder.
- o Select the file and click the **Open** button to return to the main SuperMix window, where the contents of the Excel spreadsheet are displayed as the SuperMix system file with default name **meps.ss3**.

Setting up the analysis

The next step is to describe the model to be fitted. We use the SuperMix interface to provide the model specifications. From the main menu bar, select the **File**, **New Model Setup** option.

The **Model Setup** window that appears has six tabs: **Configuration**, **Variables**, **Starting Values**, **Patterns**, **Advanced**, and **Linear Transforms**. In this example, only the screens associated with the first two tabs are used.

As a first step, select the continuous outcome variable TOTEXP99 from the **Dependent Variable** drop-down list box. The stratum and cluster variables used to define the levels of the hierarchy are specified as Level-3 ID and Level-2 ID respectively by selecting them from the **Level-3 IDs** and **Level-2 IDs** drop-down list boxes. Enter a title for the analysis in the **Title 1** text box. In this example, default settings for all other options associated with the **Configuration** screen are used. Proceed to the **Variables** screen by clicking on this tab.

Model Setup: meps.mum

| Configuration | Variables | Starting Values | Patterns | Advanced | Linear Transforms |

Title 1: 3-level model for MEPS data

Title 2:

Dependent Variable Type: continuous

Dependent Variable: TOTEXP99

Level-2 IDs: VARPSU99

Level-3 IDs: VARSTR99

Write Bayes Estimates: no

Convergence Criterion: 0.0001

Number of Iterations: 100

Missing Values Present: false

Generate Table of Means: no

Output Type: standard

Use the arrow keys or click on the desired tab to select the category of interest for the model.

The **Variables** screen is used to specify the fixed and random effects to be included in the model. Start by selecting the explanatory (fixed) variables using the **E** check boxes next to the variables names in the **Available** grid at the left of the screen. Note that, as the variables are selected, the selected variables are listed in the **Explanatory Variables** grid. After selecting all the explanatory variables, the screen shown below is obtained. The **Include Intercept** check box in the **Explanatory Variables** grid is checked by default, indicating that an intercept term will automatically be included in the fixed part of the model.
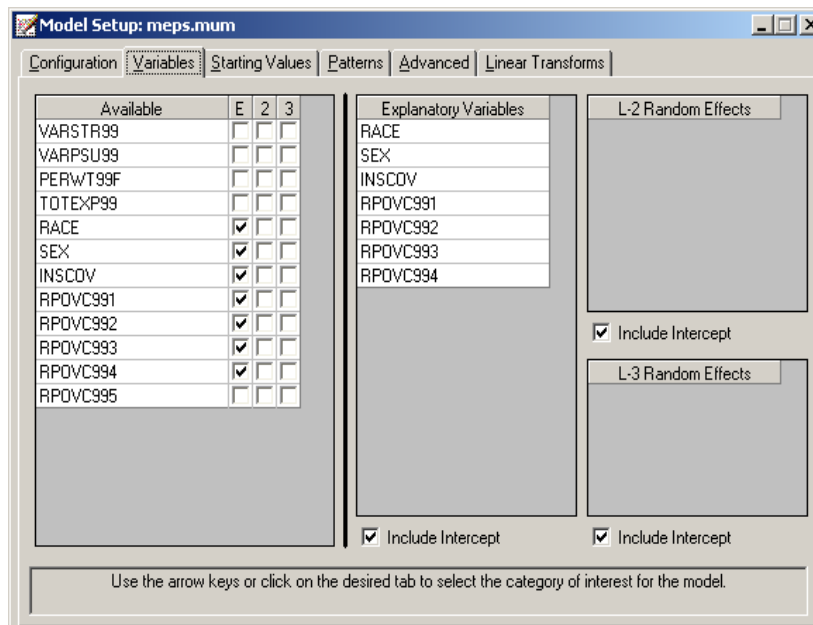
The next step is to specify the random effects at levels 2 and 3 of the hierarchy. In this example, we want to fit a model with random intercepts at levels 2 and 3. By default, the **Include Intercept** check boxes in both the **L-2 Random Effects** and **L-3 Random effects** grids are checked. If these boxes are left checked, and no additional random effects are indicated using the **2** column in the **Available** grid to the left, the model fitted will be the random intercepts only model we intend to use. No further changes on this screen are necessary.

Before running the analysis, the model specifications have to be saved. Select the **File**, **Save As** option, and provide a name (**meps.mum**) for the model specification file. Run the analysis by selecting the **Run** option from the **Analysis** menu.

Discussion of results

Portions of the output file **meps.out** are shown below.

```
 SuperMix - [meps.out]                                          _□×
  File  Analysis  Window  Help                                  _ Ð ×

 Numbers of observations
 ----------------------

 Level 3 observations =     143
 Level 2 observations =     460
 Level 1 observations =   19300

 LEVEL3 :       1       2       3       4       5       6       7       8
 N2     :       2       2       2       2       2       3       2       2
 N1     :      25      69      39      77     151      43      41      67

 LEVEL3 :       9      10      11      12      13      14      15      16
 N2     :       2       2       2      11       2       2       2       3
 N1     :      97      21      56     336      54      58     141      53

 LEVEL3 :      17      18      19      20      21      22      23      24
 N2     :       2       5       2       2       2       2       2       2
 N1     :     293     198      61      42      38      50     138     120

 LEVEL3 :      25      26      27      28      29      30      31      32
 N2     :       3       2       2       2       2       2      10       2
 N1     :     207      16      76      23      50      24     358      25

 LEVEL3 :      33      34      35      36      37      38      39      40
 N2     :       2       2       2       2      17       2       2       2
 N1     :     303      36      51      72     512      82      27     155

                                                                    OK
```

In the first section of the output file as shown above, a description of the hierarchical structure is provided. A total of 143 strata, 460 PSUs and information from 19,300 individual participants were included at levels 3, 2 and 1 of the model. This corresponds to the survey design described earlier. In addition, a summary of the number of PSUs and participants nested within each stratum is provided. For stratum number 1 (ID3: 1), data are available from only 25 participants nested within 2 primary sampling units (N2: 2). By contrast, for stratum number 12 (ID3: 12), data are available from 408 participants (N1: 336) nested within 11 primary sampling units (N2: 11).

The data summary is followed by descriptive statistics for all the variables included in the model.

```
meps.out                                                            _ □ ×

Descriptive statistics for all variables
----------------------------------------

Variable            Minimum         Maximum            Mean       Stand. Dev.

Dependent
---------
TOTEXP99            0.69315        12.23913         6.44873         1.66728

Random-Effects
--------------
intcept   (3)       1.00000         1.00000         1.00000         0.00000
intcept   (2)       1.00000         1.00000         1.00000         0.00000
intcept   (1)       1.00000         1.00000         1.00000         0.00000

Fixed Regressor(s)
------------------
intcept             1.00000         1.00000         1.00000         0.00000
RACE                0.00000         1.00000         0.81855         0.38540
SEX                 0.00000         1.00000         0.55088         0.49742
INSCOV              0.00000         1.00000         0.27974         0.44888
RPOVC991            0.00000         1.00000         0.34487         0.47534
RPOVC992            0.00000         1.00000         0.14415         0.35125
RPOVC993            0.00000         1.00000         0.31808         0.46574
RPOVC994            0.00000         1.00000         0.04933         0.21655

                                                                      OK
```
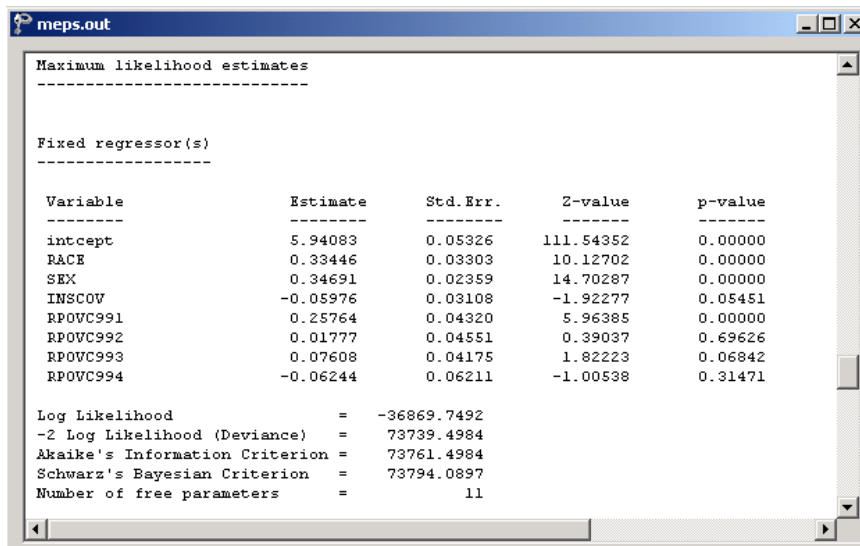
Descriptive statistics are followed by the starting values of the parameters that were used in the initial step of the iterative algorithm.

```
meps.out                                                            _ □ ×

TITLE1: 3-level model for MEPS data

Parameter starting values
-------------------------


Fixed regressor(s)
------------------

Variable             Estimate      Std.Err.      Z-value       p-value
--------             --------      --------      -------       -------
intcept               5.88952       0.04761     123.70979      0.00000
RACE                  0.29197       0.03111       9.38430       0.00000
SEX                   0.34592       0.02389      14.48028       0.00000
INSCOV               -0.09205       0.03086      -2.98255       0.00286
RPOVC991              0.33054       0.04250       7.77679       0.00000
RPOVC992              0.04650       0.04536       1.02518       0.30528
RPOVC993              0.11773       0.04144       2.84117       0.00449
RPOVC994             -0.05546       0.06203      -0.89408       0.37128

Log Likelihood          =    -43400.7859
Number of free parameters    =         11
```

The output describing the estimated **fixed effects** after convergence is shown next. The estimates are shown in the column with heading Estimate, and correspond to the coefficients $\beta_0, \beta_2, \ldots, \beta_7$ in the model specification. From the z-values and associated exceedance probabilities, we see that the coefficients associated with gender, ethnicity and insurance coverage type were all highly significant. Recall that a value of 1 for the ethnicity indicator variable RACE indicated that a participant was white, with a value of 0 assigned to participants from all other ethnic groups. The positive estimated coefficient for this variable indicates an increase of 0.3345 units in the logarithm of total health expenditure, holding all other predictors constant. Similarly, female participants (coded "1" on the gender indicator SEX), are expected to have a total health expenditure 0.3469 higher than male participants if all other variables are held constant. In contrast, participants with public coverage or no coverage have a lower expected total expenditure, as indicated by the negative estimated coefficient –0.0598.

```
meps.out                                                                  _ □ ×

Maximum likelihood estimates
----------------------------


Fixed regressor(s)
------------------

  Variable              Estimate      Std.Err.      Z-value       p-value
  --------              --------      --------      -------       -------
  intcept                5.94083       0.05326     111.54352      0.00000
  RACE                   0.33446       0.03303      10.12702      0.00000
  SEX                    0.34691       0.02359      14.70287      0.00000
  INSCOV                -0.05976       0.03108      -1.92277      0.05451
  RPOVC991               0.25764       0.04320       5.96385      0.00000
  RPOVC992               0.01777       0.04551       0.39037      0.69626
  RPOVC993               0.07608       0.04175       1.82223      0.06842
  RPOVC994              -0.06244       0.06211      -1.00538      0.31471

  Log Likelihood                   =   -36869.7492
  -2 Log Likelihood (Deviance)     =    73739.4984
  Akaike's Information Criterion   =    73761.4984
  Schwarz's Bayesian Criterion     =    73794.0897
  Number of free parameters        =           11
```

Turning to the indicator variables associated with income relative to the poverty line, it can be seen that only one of the indicator variables, RPOVC991, has an estimated coefficients that is significantly different from zero at a 5% level of

significance. In the case of families with a "high" income, the estimate of 0.2576 for RPOVC991 indicates an expected increase in expenditure, while for "near poor" families, the estimate of –0.0624 indicates an expected decrease in expenditure, holding all other variables constant.

In addition to the likelihood function value at convergence, a number of related statistical measures for model adequacy are available. The most common of these are the likelihood ratio test and Akaike's and Schwarz's criteria. Both the Akaike information criterion (AIC) and the Schwarz Bayesian criterion (SBC) are functions of the number of estimated parameters, and therefore "penalize" models with large numbers of parameters. In the SuperMix output file, all three of these are reported.

- o The deviance is defined as $-2\ln L.$ For a pair of nested models, the difference in $-2\ln L$ values has a χ^2 distribution, with degrees of freedom equal to the difference in number of parameters estimated in the models compared.
- o The AIC was originally proposed for time-series models, but is also used in regression. It is defined as $-2\ln L + 2r,$ where r denotes the number of parameters estimated in the model. The model with minimum AIC, in a set of nested models, will be the most parsimonious according to this criterion.
- o The SBC is defined as $-2\ln L + r\log n,$ where n denotes the number of units at the highest level of the hierarchy. A smaller value of this criterion would indicate the most parsimonious of the models being compared.

Estimated outcomes for different groups

To evaluate the expected effect of the measure of a family's income on the corresponding projected expenditure, suppose that the variables RACE, SEX, and INSCOV are held at zero, as would be the case for a nonwhite male participant with private insurance coverage. If such a participant originates from a family with "high" income, the logarithm of total health expenditure is expected to be

$$\hat{\beta}_0 + \hat{\beta}_4 \left(\text{RPOVC991}\right)_{ij} + \hat{\beta}_5 \left(\text{RPOVC992}\right)_{ij} + \hat{\beta}_6 \left(\text{RPOVC993}\right)_{ij} + \hat{\beta}_7 \left(\text{RPOVC994}\right)_{ij}$$

$$= \hat{\beta}_0 + \hat{\beta}_4$$
$$= 5.94083 + 0.25764$$
$$= 6.19847$$

which translates to a projected total expenditure of $e^{6.19847} = \$492$. In contrast, for a participant with similar demographic background and coverage from a "near poor" family, we obtain a projected total expenditure of

$$e^{\hat{\beta}_0 + \hat{\beta}_7}$$
$$= e^{5.94083 - 0.06244}$$
$$= \$357.23.$$

The predicted total expenditure (as natural logarithm) for similar participants from "low", "middle" or "negative or poor" families are similarly obtained by calculating $e^{\hat{\beta}_0 + \hat{\beta}_5}$, $e^{\hat{\beta}_0 + \hat{\beta}_6}$ and $e^{\hat{\beta}_0}$ respectively.

In Table 3.1, the predicted total health expenditure is given for respondents with high or near poor family income, for each of the subpopulations formed by gender, ethnicity, and insurance coverage. For purposes of the comparison, results are expressed in U.S. dollars, rather than in the natural logarithmic units of the outcome variable TOTEXP99. Respondents from families with high income consistently outspend their near poor counterparts by at least 100%, regardless of gender, ethnicity, or level of insurance coverage. In families with high income, female respondents spent more in 1999 than their male counterparts, regardless of ethnicity. This is generally also true for near poor respondents. It is also apparent that the total health expenditure in 1999 was higher for respondents with private insurance than for respondents with public or no coverage, and that white respondents spent more than respondents from other ethnic groups, regardless of gender or the level of family income. From exploratory analyses, we know that the outcome variable

TOTEXP99, when expressed in terms of dollar values instead of natural logarithmic units, is highly skewed. It has a median of $615 and a mean $2,492. When this is taken in account, we can conclude that, generally speaking, white females spent more on health in 1999 than 50% of all respondents in the sample.

Table 3.1: Predicted total health expenditure for various subgroups

Group	**Male** (SEX = 0)		**Female** (SEX = 1)	
	Insurance coverage:		**Insurance coverage:**	
	Private (INSCOV=0)	**Public/none** (INSCOV = 1)	**Private** (INSCOV=0)	**Public/none** (INSCOV = 1)
Respondents with high family income (RPOVC991 = 1)				
Nonwhite (RACE = 0)	$492	$463	$696	$656
White (RACE = 1)	$687	$648	$972	$916
Respondents with near poor income (RPOVC994 = 1)				
Nonwhite (RACE = 0)	$357	$337	$505	$476
White (RACE = 1)	$499	$470	$706	$665

The output for the **random part** of the model follows, and is shown in the image below. There is significant variation in the average estimated total health expenditure at all levels, with the most variation over the participants (level 1), and the least variation over strata (level-3).

```
┌────────────────────────────────────────────────────────────────────┐
│ ℙ  meps.out                                                 _ □ X    │
├────────────────────────────────────────────────────────────────────┤
│                                                                  ▲   │
│                                                                      │
│    Variance/covariance components                                    │
│    -----------------------------                                     │
│                                                                      │
│                                                                      │
│     Level 3                   Estimate    Std.Err.    Z-value  p-value│
│     -------                   --------    --------    -------  -------│
│     intcept /intcept          0.03494     0.01055     3.31114  0.00093│
│                                                                      │
│     Level 2                   Estimate    Std.Err.    Z-value  p-value│
│     -------                   --------    --------    -------  -------│
│     intcept /intcept          0.05032     0.00968     5.19828  0.00000│
│                                                                  ▒   │
│     Level 1                   Estimate    Std.Err.    Z-value  p-value│
│     -------                   --------    --------    -------  -------│
│     intcept /intcept          2.62733     0.02704     97.18098 0.00000│
│                                                                  ▼   │
│  ◄                                                            ►      │
├────────────────────────────────────────────────────────────────────┤
│                                                         [   OK   ]   │
└────────────────────────────────────────────────────────────────────┘
```

An estimate of the level-2 cluster effect, for example, is obtained as

$$\frac{0.05032}{0.03494 + 0.05032 + 2.62733} \times 100\% = 1.86\%$$

indicating that only 1.86% of the total variance in expenditure explained is at level 2 of the model.

3.3.4 Example: A random intercept model with 3 predictors

To illustrate model comparison, a simpler model was fitted to the same data. In the previous model, the estimated coefficients of only 2 of the 4 indicator variables representing a respondent's position relative to the poverty line were significant. In this model, only the ethnicity, gender, and level of insurance coverage were included as explanatory variables.

Setting up the analysis

The model specification file from the previous example can be used as the basis for the simplified model. As a previously saved model specification file can only be

opened once the associated **ss3** file is open, start by using the **File**, **Open** option to browse for and open **meps.ss3**. Next, open the model specification file **meps.mum** by selecting the **Open Existing Model Setup** option from the **File** menu.
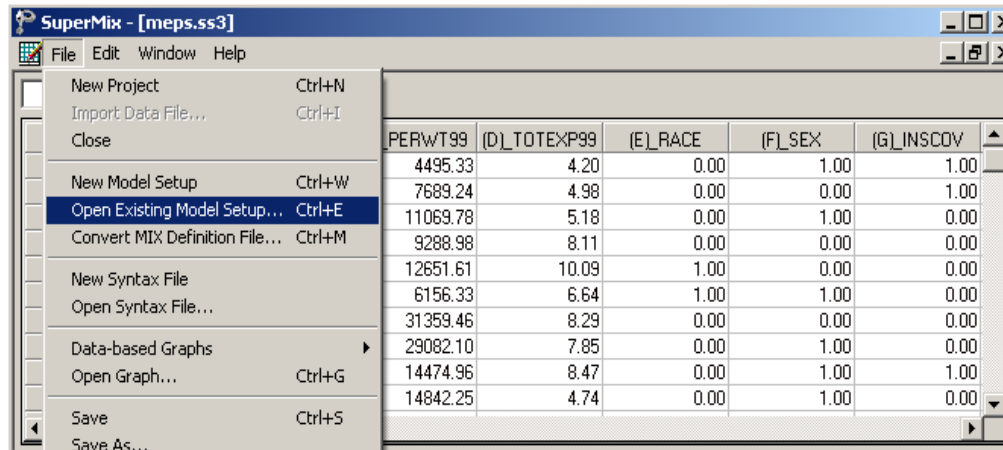
Once the **Model Setup** window is displayed, click on the **Variables** tab. Delete the first of the four indicator variables (RPOVC991) by unchecking the check box next to this variable as shown below.

Remove the other three indicator variables in the same way, and save the revised model specification file as **meps2.mum** using the **File**, **Save As** option. Finally, click the **Analysis**, **Run** option to start the analysis.
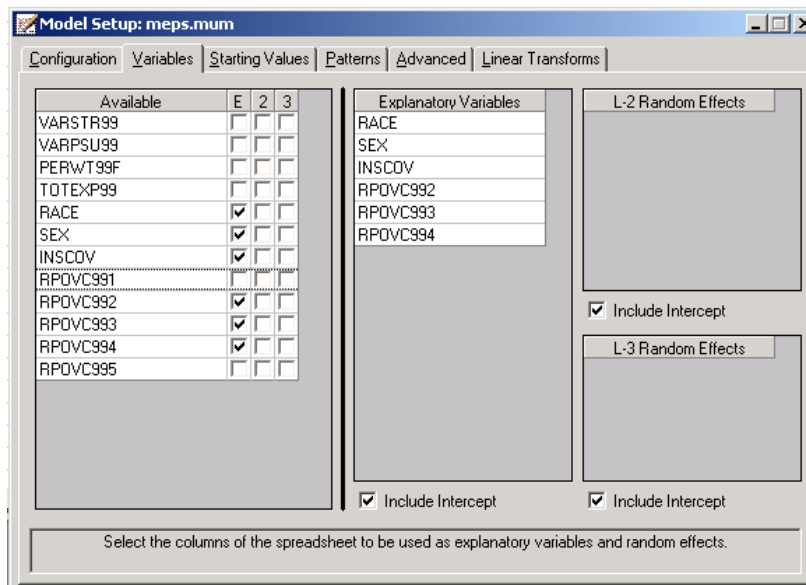
Discussion of results

A portion of the output file **meps2.out** is shown below. The estimates of the fixed effects are close to those obtained for the previous model.
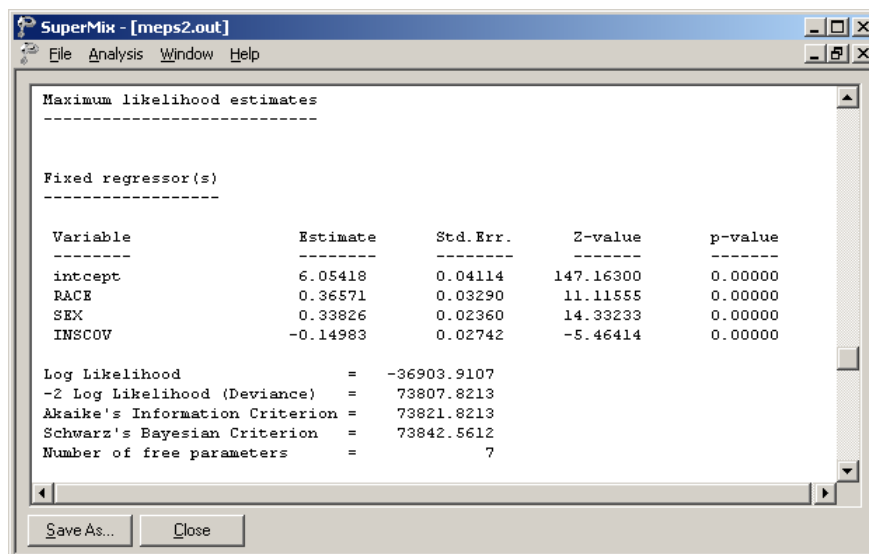
Recall that the fit measures obtained for the previous model were:

```
Log Likelihood   = -36869.7492
Deviance         =  73739.4984
AIC              =  73761.4984
SBC              =  73784.0897
Number of free parameters = 11
```
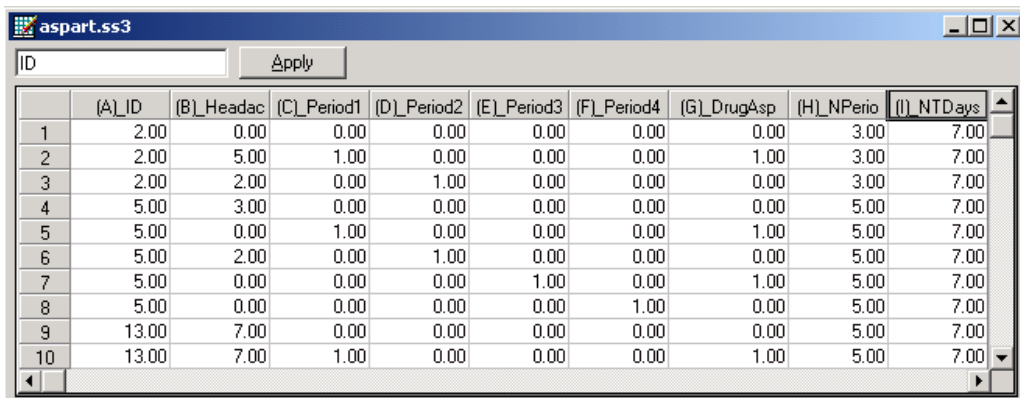
All of the reported statistical measures for model adequacy indicate that the previous model, where the variables RPOVC991 to RPOVC994 were included, offers a better fit to the data.

3.4 Two-level models for count outcomes

3.4.1 The data

The data for this example are taken from a paper by McKnight and Van Den Eeden (1993), who reported on the number of headaches in a two treatment, multiple period crossover trial. Specifically, the number of headaches per week was repeatedly measured for 27 patients. Following a seven-day placebo run-in period, subjects received either aspartame or placebo in four seven-day treatment periods according to the double-blind crossover treatment design. Each treatment period was separated by a washout day. The sample size is 122. Data for the first 10 observations of all the variables used in this section are shown below in the form of a SuperMix spreadsheet window for **aspart.ss3**.

	(A)_ID	(B)_Headac	(C)_Period1	(D)_Period2	(E)_Period3	(F)_Period4	(G)_DrugAsp	(H)_NPerio	(I)_NTDays
1	2.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	7.00
2	2.00	5.00	1.00	0.00	0.00	0.00	1.00	3.00	7.00
3	2.00	2.00	0.00	1.00	0.00	0.00	0.00	3.00	7.00
4	5.00	3.00	0.00	0.00	0.00	0.00	0.00	5.00	7.00
5	5.00	0.00	1.00	0.00	0.00	0.00	1.00	5.00	7.00
6	5.00	2.00	0.00	1.00	0.00	0.00	0.00	5.00	7.00
7	5.00	0.00	0.00	0.00	1.00	0.00	1.00	5.00	7.00
8	5.00	0.00	0.00	0.00	0.00	1.00	0.00	5.00	7.00
9	13.00	7.00	0.00	0.00	0.00	0.00	0.00	5.00	7.00
10	13.00	7.00	1.00	0.00	0.00	0.00	1.00	5.00	7.00

The variables of interest are:

- o ID is the patient ID (27 patients in total).
- o HeadAche is the number of headaches during the week (from 0 to 7).
- o Period1 is a period 1 treatment indicator (1 for the first treatment period and 0 otherwise).

- Period2 is a period 2 treatment indicator (1 for the second treatment period and 0 otherwise).
- Period3 is a period 3 treatment indicator (1 for the third treatment period and 0 otherwise).
- Period4 is a period 4 treatment indicator (1 for the fourth treatment period and 0 otherwise).
- DrugAsp indicates the type of drug being used for the treatment, (0 = placebo and 1 = aspartame). 75 observations used placebo and 47 used aspartame.
- Nperiods is the number of periods the individual was observed (from 2 to 5).
- NTDays is the number of treatment days in the period (from 1 to 7).

3.4.2 The model

A general two-level Poisson regression model for a count response variable y depending on a set of r predictors $x_1, x_1, \ldots, x_r$ may be expressed as

$$\ln\left(\mu_{ij}\right) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i$$

where y_{ij} denotes the value of y for the j-th level-1 unit nested within the i-th level-2 unit for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n_i$, the scalar product $\mathbf{x}_{ij}'\boldsymbol{\beta}$ is the fixed part of the model, and $\mathbf{z}_{ij}'\mathbf{u}_i$ denotes the random part of the model at level 2. For the fixed part of the model, $\mathbf{x}_{ij}'$ is a typical row of the design matrix while the vector $\boldsymbol{\beta}$ contains the fixed, but unknown parameters to be estimated. In the case of the random part of the model at level 2, $\mathbf{z}_{ij}'$ represents a typical row of the design matrix, and $\mathbf{u}_i$ the vector of random level-2 effects to be estimated.

TWO-LEVEL MODELS FOR COUNT OUTCOMES

The specific Poisson regression model with a random intercept for the number of headaches may be expressed as

$$\hat{\mu}_{ij} = \exp\left(\beta_0 + \beta_1 * \text{Period1}_{ij} + \beta_2 * \text{Period2}_{ij} + \beta_3 * \text{Period3}_{ij} + \beta_4 * \text{Period4}_{ij} + \beta_5 * \text{DrugAsp}_{ij} + u_{i0}\right)$$

where $\hat{\mu}_{ij}$ denotes the mean number of headaches of patient i for treatment period j; Period2$_{ij}$, Period2$_{ij}$, Period3$_{ij}$ and Period4$_{ij}$ denote the values of the dummy variables Period1, Period2, Period3 and Period4 for patient i for treatment period j respectively; DrugAsp$_{ij}$ denotes the value of the DrugAsp for patient i for treatment period j; β_0, β_1, β_2, β_3, β_4 and β_5 denote unknown parameters; and u_{i0} denotes the random intercept for patient i for $i = 1, 2, \ldots, 27$ and $j = 1, 2, 3, 4$. This model is fitted to the data in **aspart.ss3** as follows.

3.4.3 Example: Poisson regression with a random intercept

Importing the data

The first step is to create the **ss3** file, **aspart.ss3**, from the Excel workbook **aspart.xls**. This is accomplished as follows:

- o Use the **Import Data File** option on the **File** menu to open the **Open** dialog box.
- o Browse for the file **aspart.xls** in the **Examples\Primer\Count** folder.
- o Select the file and click on the **Open** button to open the following SuperMix spreadsheet window for **aspart.ss3**.

After selecting the **File**, **Save** option from the main menu bar, we are ready to fit the Poisson regression model with a random intercept for the number of headaches to the data in **aspart.ss3**.

Setting up the analysis

Start by selecting the **New Model Setup** option on the **File** menu to open the **Model Setup** window.

On the **Configuration** screen, we first enter the titles Aspartame Data – Repeated Headaches across Time and random intercept and 5 covariates for the analysis in the **Title 1** and **Title 2** text boxes respectively. The count outcome variable HeadAche is selected from the **Dependent Variable** drop-down list box. The **Dependent Variable Type** drop-down list box is used to indicate that the outcome variable is a **count**. The variable ID, which defines the levels of the hierarchy, is selected as the Level-2 ID from the **Level-2 IDs** drop-down list box.

TWO-LEVEL MODELS FOR COUNT OUTCOMES

Next, click on the **Variables** tab to proceed with variable selection. The variables Period1, Period2, Period3, Period4, and DrugAsp are specified as the fixed effects of the model by checking the **E** check boxes for Period1, Period2, Period3, Period4, and DrugAsp in the **Available** grid. These actions produce the **Variables** screen as shown above.

Finally, we enter the number of quadrature points, in this case 20, in the **Number of Quadrature Points** text box on the **Advanced** screen as shown below. Also, change the **Optimization Method** to **non-adaptive quadrature**.
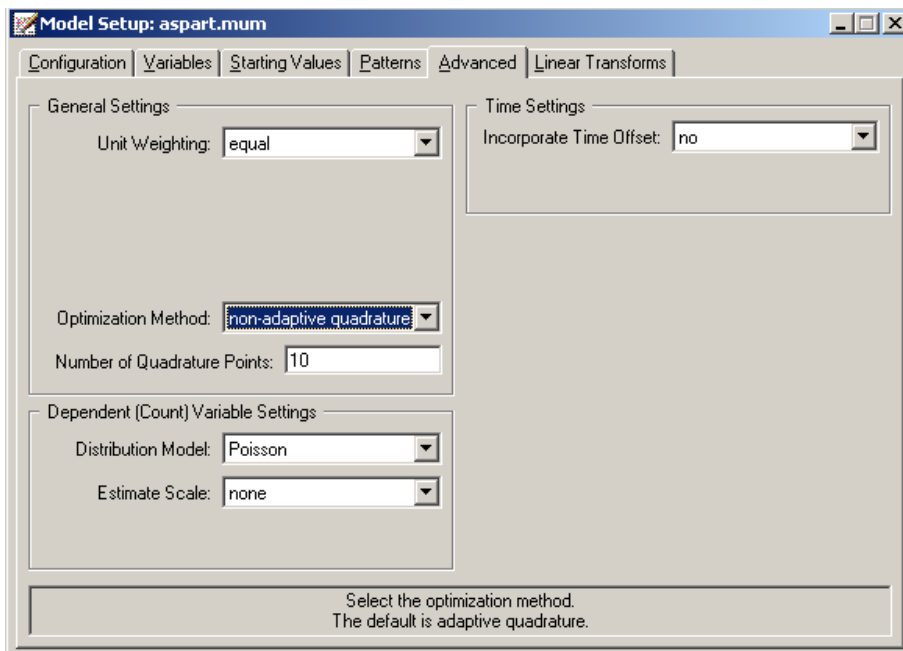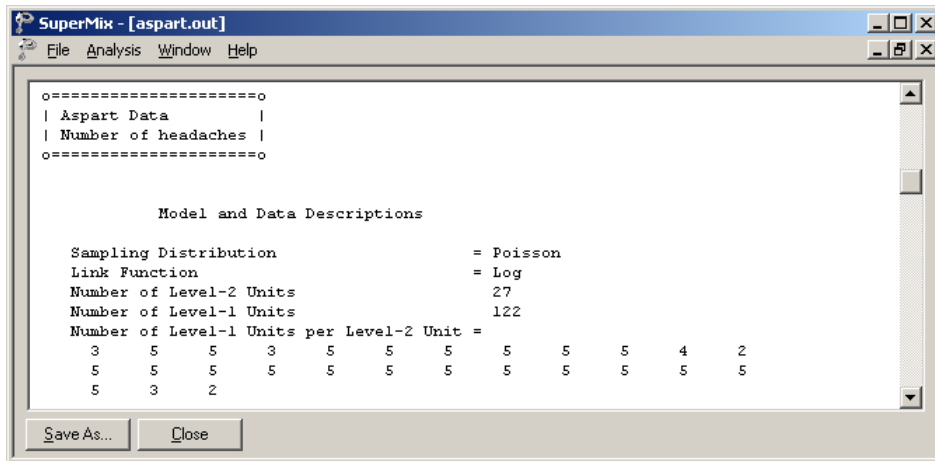
Before we can run the analysis, we have to save the model specifications to file. This is accomplished by using the **Save** option on the **File** menu to open a **Save Mixed Up Model** dialog box. First enter the name **aspart.mum** in the **File name** text box and then click on the **Save** button to save the file. The analysis is run by

selecting the **Run** option from the **Analysis** menu. This produces the corresponding output file **aspart.out**.

Discussion of results

Portions of this output file are shown below.

The above output file indicates that there are 27 subjects with 122 observations nested within them. The number of observations per subject varies between 2 and 5.

The descriptive statistics for all the variables in the model are shown next. Following the descriptive statistics, the results for the model without any random effect is given, as shown below.

```
o==========================================================o
| Descriptive statistics for all the variables in the model |
o==========================================================o
                                            Standard
    Variable    Minimum    Maximum      Mean   Deviation
    --------    -------    -------      ----   ---------
    Headache     0.0000     7.0000    1.6803      1.8863
    intcept      1.0000     1.0000    1.0000      0.0000
    Period1      0.0000     1.0000    0.2213      0.4168
    Period2      0.0000     1.0000    0.2049      0.4053
    Period3      0.0000     1.0000    0.1803      0.3860
    Period4      0.0000     1.0000    0.1721      0.3791
    DrugAsp      0.0000     1.0000    0.3852      0.4887
```

```
o==========================================================o
| Results for the model without any random effects |
o==========================================================o


              Goodness of fit statistics


    Statistic                        Value       DF       Ratio
    ---------                        -----       --       -----
    Likelihood Ratio Chi-square    243.8257      116      2.1019
    Pearson Chi-square             253.8934      116      2.1887


              Estimated regression weights


                             Standard
    Parameter     Estimate      Error    z Value    P Value
    ---------     --------    --------    -------    -------
    intcept         0.4654      0.1525     3.0516     0.0023
    Period1         0.0916      0.2265     0.4043     0.6860
    Period2         0.0131      0.2276     0.0575     0.9542
    Period3        -0.2245      0.2471    -0.9084     0.3637
    Period4        -0.1840      0.2540    -0.7242     0.4689
    DrugAsp         0.2332      0.1596     1.4612     0.1440


       Event Rate Ratio and 95% Event Rate Confidence Intervals


                                             Bounds
    Parameter     Estimate   Event Rate    Lower      Upper
    ---------     --------   ----------    -----      -----
    intcept         0.4654       1.5926    1.1811     2.1474
    Period1         0.0916       1.0959    0.7030     1.7085
    Period2         0.0131       1.0132    0.6486     1.5827
    Period3        -0.2245       0.7989    0.4923     1.2967
    Period4        -0.1840       0.8320    0.5057     1.3688
    DrugAsp         0.2332       1.2626    0.9235     1.7263
```
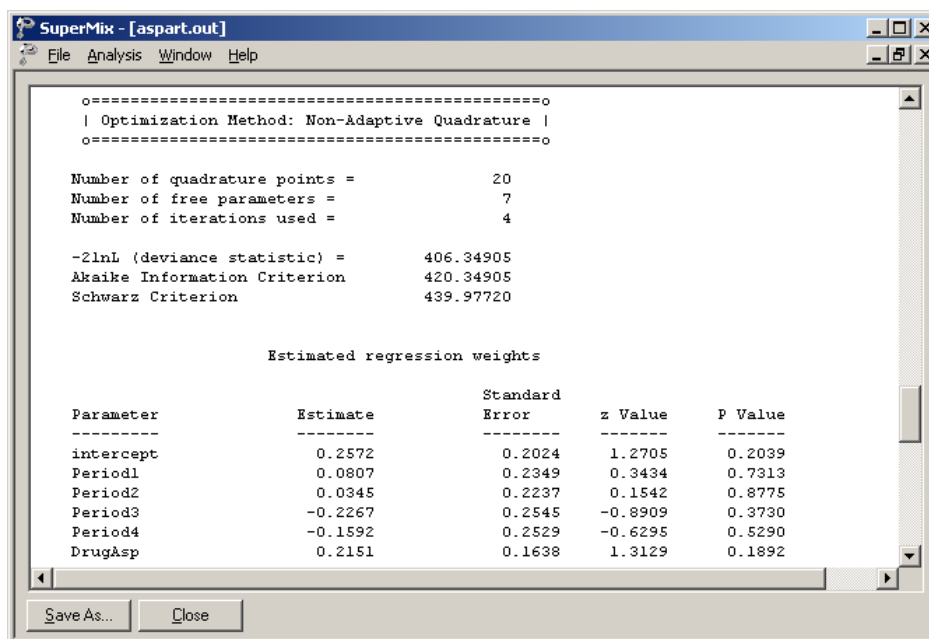
TWO-LEVEL MODELS FOR COUNT OUTCOMES

Next, the final results using non-adaptive quadrature optimization method is given. As shown below, the model converged after 4 iterations. The fixed part of the estimates is given. The *p*-values for all these estimates are not significant at 10% level.

```
SuperMix - [aspart.out]                                              _ □ ×
 File  Analysis  Window  Help                                        _ 日 ×

    o===============================================o
    | Optimization Method: Non-Adaptive Quadrature |
    o===============================================o

    Number of quadrature points =          20
    Number of free parameters =            7
    Number of iterations used =            4

    -2lnL (deviance statistic) =      406.34905
    Akaike Information Criterion      420.34905
    Schwarz Criterion                 439.97720


                   Estimated regression weights

                                       Standard
    Parameter            Estimate      Error       z Value    P Value
    ---------            --------      --------    -------    -------
    intercept             0.2572        0.2024      1.2705     0.2039
    Period1               0.0807        0.2349      0.3434     0.7313
    Period2               0.0345        0.2237      0.1542     0.8775
    Period3              -0.2267        0.2545     -0.8909     0.3730
    Period4              -0.1592        0.2529     -0.6295     0.5290
    DrugAsp               0.2151        0.1638      1.3129     0.1892

 Save As...    Close
```

```
SuperMix - [aspart.out]                                              _ □ X
 File  Analysis  Window  Help                                        _ 回 X

    Event Rate Ratio and 95% Event Rate Confidence Intervals

                                               Bounds
    Parameter           Estimate      Event Rate    Lower      Upper
    ---------           --------      ----------    -----      -----
    intercept             0.2572         1.2933     0.8697     1.9231
    Period1               0.0807         1.0840     0.6840     1.7180
    Period2               0.0345         1.0351     0.6677     1.6046
    Period3              -0.2267         0.7971     0.4841     1.3127
    Period4              -0.1592         0.8528     0.5195     1.4000
    DrugAsp               0.2151         1.2400     0.8994     1.7096


         Estimated level 2 variances and covariances

                                      Standard
    Parameter           Estimate      Error        z Value    P Value
    ---------           --------      --------      -------    -------
    intercept/intercept    0.4290        0.1715      2.5024     0.0123

 Save As...      Close
```

The random-effect variance is estimated as 0.429, with a *p*-value of 0.0123. All the regression coefficients in the fixed part of the model are non-significant.

3.4.4 Example: Mixed-effects analysis with an offset variable

The previous analysis has assumed that the counts were all observed for the same number of days. However, this was not the case since the number of treatment days in the period (the 9th field of the input data file) did vary to some degree. Most of the counts were based on the full seven days in the week; however, some observations were made only for 1 day in the given week. To take this into account, we need to specify a so-called OFFSET variable. The offset variable indicates the amount of time that each count is based on. If OFFSET = no is specified, then SuperMix assumes that all counts are based on the same amount of time.
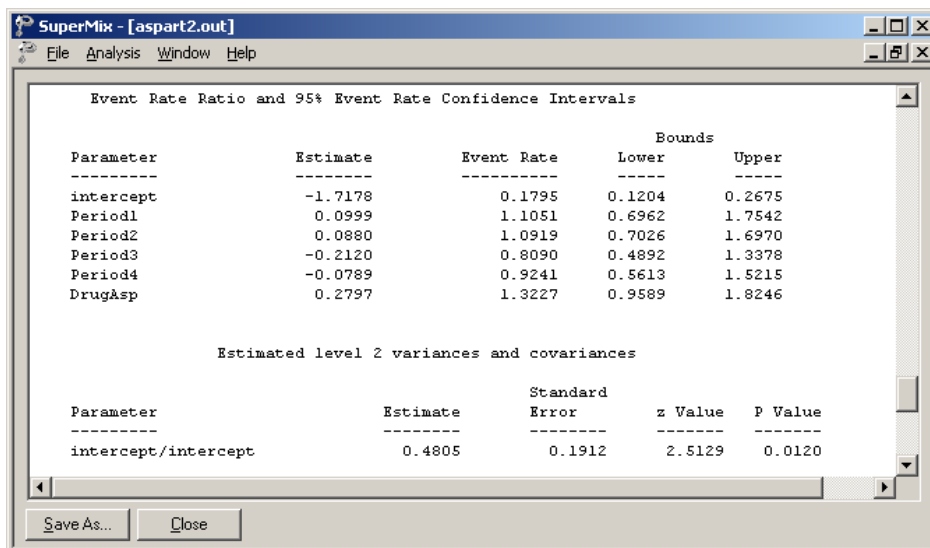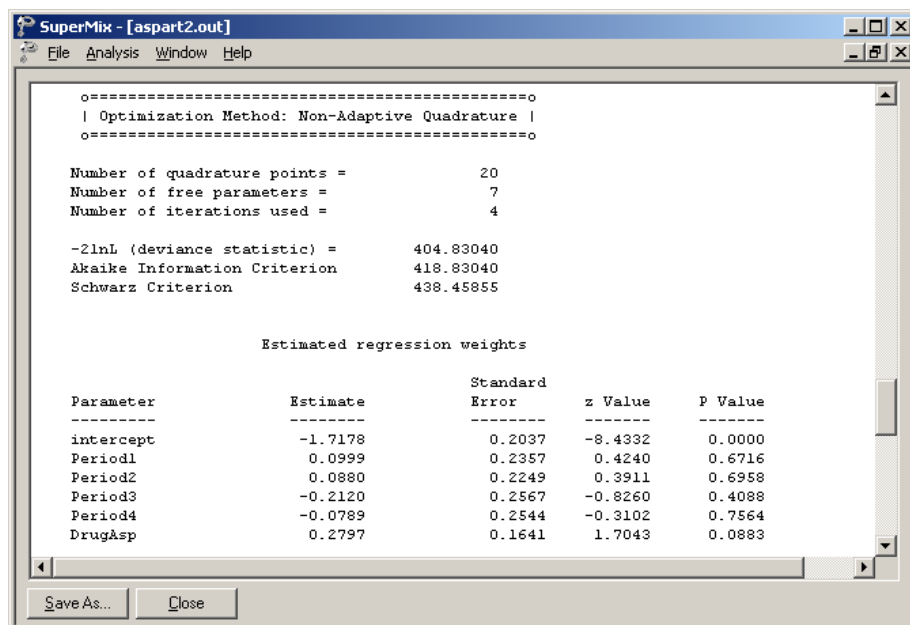
Setting up the analysis

To create the model specifications for this model, start by opening **aspart.ss3** in a SuperMix spreadsheet window. Then, use the **Open Existing Model Setup** option on the **File** menu to open the **Model Setup** window for **aspart.mum**. On the **Configuration** screen, extend the title in the **Title 1** text box by adding the string "with Offset Variable". Next, click on the **Advanced** tab of the **Model Setup** window. Select yes from the **Incorporate Time Offset** drop-down list to activate the **Offset Variable** drop-down list box. Select the variable NTDays from the drop-down list of **Offset Variable** to produce the following **Advanced** screen.

Save the changes to the file **aspart2.mum** by using the **Save As** option on the **File** menu. To fit the revised model to the data, select the **Run** option on the **Analysis** menu to produce the output file **aspart2.out**.

Discussion of results

A portion of this output file is shown below.

```
 SuperMix - [aspart2.out]                                               _ □ ×
  File  Analysis  Window  Help                                          _ 8 ×

    o===============================================o
    | Optimization Method: Non-Adaptive Quadrature |
    o===============================================o

    Number of quadrature points =           20
    Number of free parameters =              7
    Number of iterations used =              4

    -2lnL (deviance statistic) =       404.83040
    Akaike Information Criterion       418.83040
    Schwarz Criterion                  438.45855


                    Estimated regression weights

                                       Standard
    Parameter            Estimate      Error       z Value     P Value
    ---------            --------      --------     -------     -------
    intercept            -1.7178        0.2037      -8.4332      0.0000
    Period1               0.0999        0.2357       0.4240      0.6716
    Period2               0.0880        0.2249       0.3911      0.6958
    Period3              -0.2120        0.2567      -0.8260      0.4088
    Period4              -0.0789        0.2544      -0.3102      0.7564
    DrugAsp               0.2797        0.1641       1.7043      0.0883

  Save As...      Close
```

```
 SuperMix - [aspart2.out]                                               _ □ ×
  File  Analysis  Window  Help                                          _ 8 ×

    Event Rate Ratio and 95% Event Rate Confidence Intervals

                                                Bounds
    Parameter            Estimate     Event Rate   Lower      Upper
    ---------            --------     ----------   -----      -----
    intercept            -1.7178        0.1795     0.1204     0.2675
    Period1               0.0999        1.1051     0.6962     1.7542
    Period2               0.0880        1.0919     0.7026     1.6970
    Period3              -0.2120        0.8090     0.4892     1.3378
    Period4              -0.0789        0.9241     0.5613     1.5215
    DrugAsp               0.2797        1.3227     0.9589     1.8246


               Estimated level 2 variances and covariances

                                       Standard
    Parameter               Estimate   Error       z Value    P Value
    ---------               --------   --------     -------    -------
    intercept/intercept      0.4805     0.1912      2.5129      0.0120

  Save As...      Close
```

Here, we see a marginally significant positive relationship between drug treatment and number of headaches. All time effects are again non-significant.

As mentioned, the empirical Bayes estimates of the random effects are written to the file **aspart2.ba2** as shown below at the conclusion of the SuperMix run. To view the file, click the **Analysis**, **View Level-2 Bayes Results** option on the output window on as shown below.

The first few lines of this file are shown below.

The file **aspart2.ba2** contains four pieces of information per individual:

- o the individual's ID,
- o the number of the random effect,
- o the empirical Bayes estimate for that individual (which is the mean of the posterior distribution), and

      o   the associated posterior standard deviation, and

      o   the name of the relevant random coefficient.

Since they are estimates of u_{i0} for each individual, the empirical Bayes estimates are expressed on the standard normal scale. Inspection of these estimates indicates that subject 13 has a very high score. This person's estimate of 1.043 (with standard deviation .016) suggests a very high level of headaches. This agrees well with the raw data, which reveals that this person recorded 7 headaches on four occasions and 6 on the only other occasion.

Graphical displays

Figure 3.9 is a comparison (represented by a dotted line) of the predicted average number of headaches reported by each patient when taking a placebo (left axis) as opposed to the predicted average number when the treatment is aspartame (right axis). From the graphical display, it appears as if all of the lines (each representing a patient) have a positive slope. The slopes become steeper as the number of headaches increases. This suggests an increase, albeit small, in the expected average number of headaches when aspartame is used. Note that patient 13, who reported a consistently high number of headaches at all occasions, was excluded from this graph.
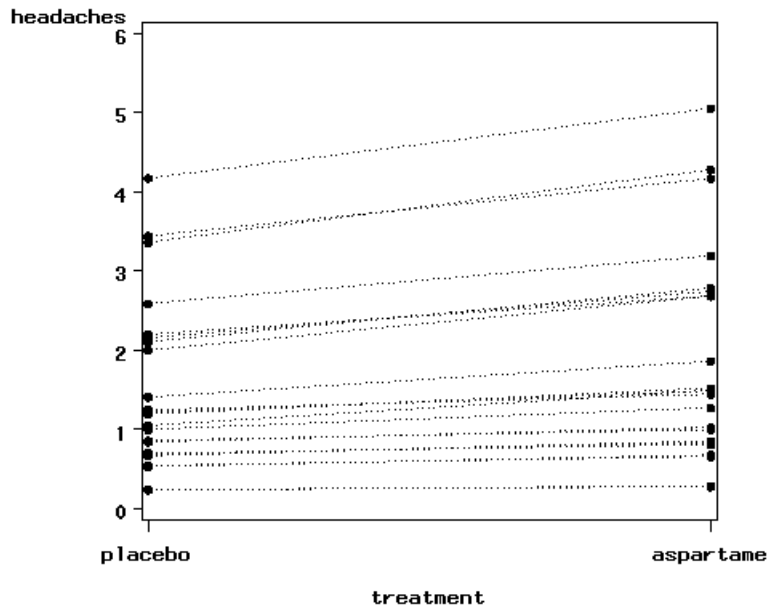
Figure 3.9: Predicted average number of headaches for placebo and aspartame

Fitted line and observed trajectory
patient=24

Fitted line and observed trajectory
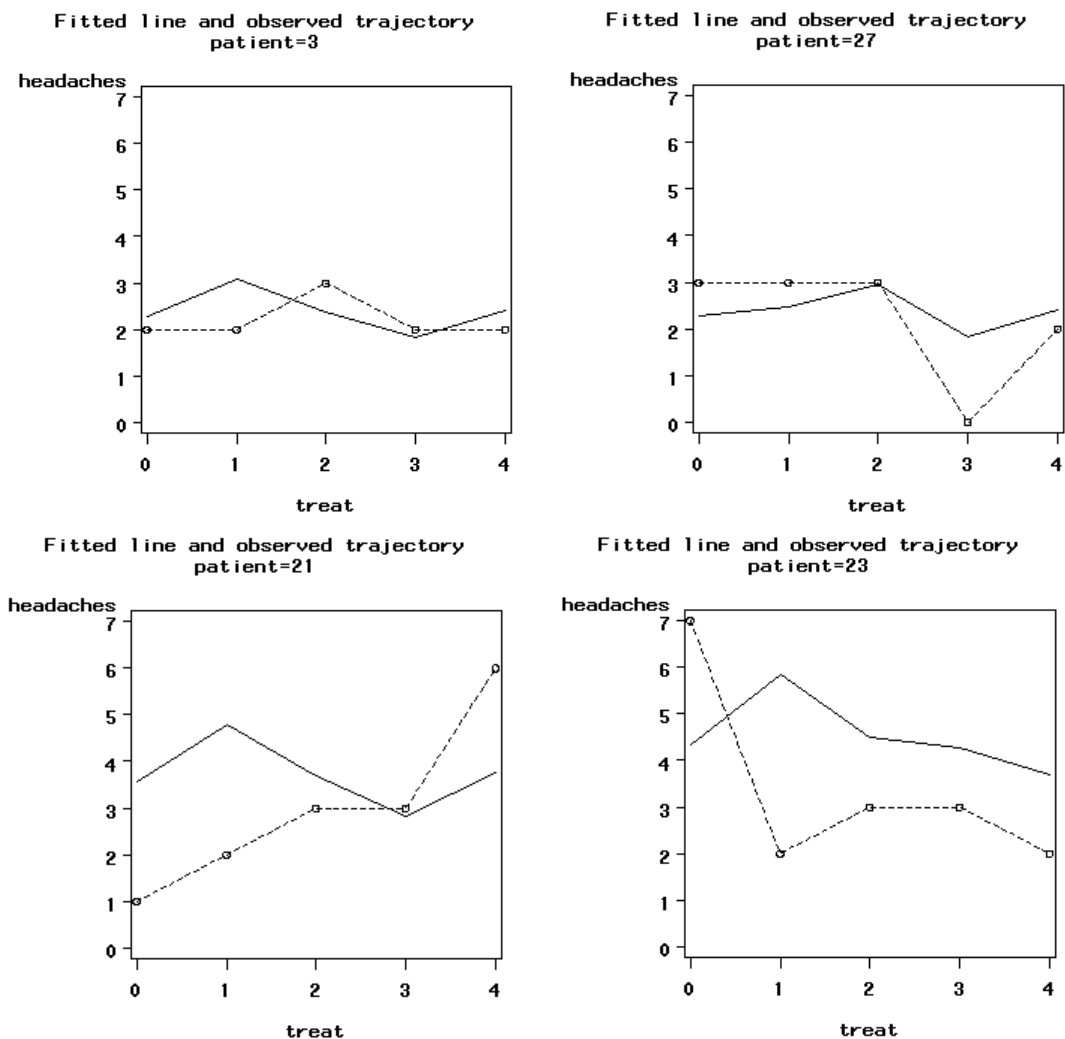patient=7

TWO-LEVEL MODELS FOR COUNT OUTCOMES 165
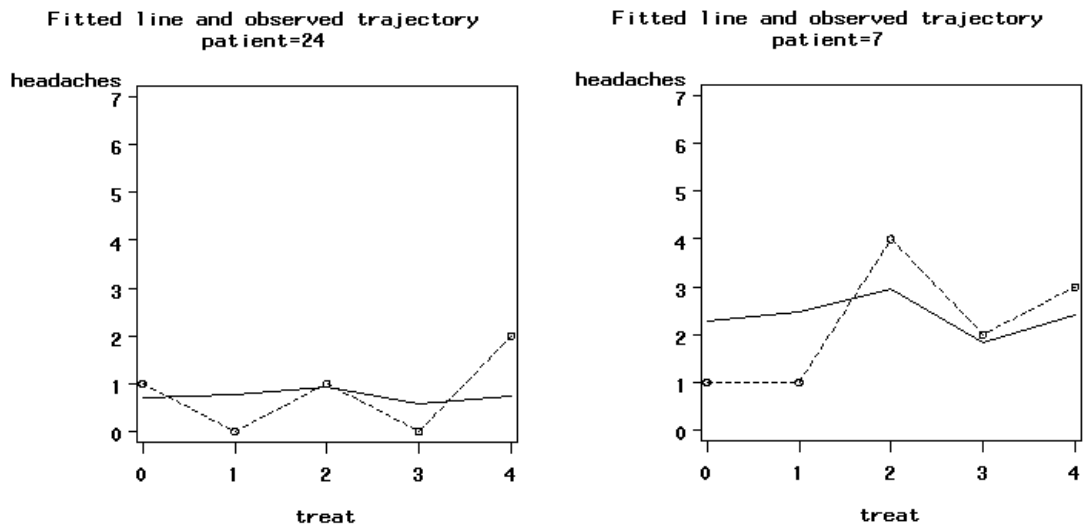
Figure 3.10: Fitted and observed trajectories

Figure 3.10 is a graphical display of the fitted trajectory (solid line) and observed trajectory (dotted line) for a sample of 6 patients. These displays are ordered from a patient who reported a relatively small number of headaches at the different

　　　　　　　　　　　　TWO-LEVEL MODELS FOR COUNT OUTCOMES

treatment occasions to one who reported a relatively high number of headaches at the treatment occasions. A study of the fitted and observed trajectories reveals that, in general, the model fit is best when the number of headaches is smaller and becomes less accurate as the number of headaches increases. For patient 13, who is not represented in the graphical display, the number of predicted headaches is almost twice the number observed.

The fitted lines were obtained as

$$\hat{y} = \exp(-1.7178 + 0.0999(\text{Period1})_{ij} + 0.0880(\text{Period2})_{ij} - 0.2120(\text{Period3})_{ij}$$
$$-0.0789(\text{Period4})_{ij} + 0.2797(\text{DrugAsp})_{ij} + \hat{u}_{i0}$$

where $\hat{u}_{i0}$ is obtained from the **aspart2.ba2** file, shown in the **discussion of results** section.

3.5   Two-level models for binary outcomes

3.5.1 The data

To illustrate the application of the mixed-effects ordinal logistic regression model to longitudinal data, data from the NIMH Schizophrenia Collaborative Study on treatment related changes in overall severity are used. Specifically, Item 79 of the Inpatient Multidimensional Psychiatric Scale (IMPS; Lorr & Klett, 1966) was used. Item 79, "Severity of Illness," (IMPS79) was scored as: 1 = normal, or not at all ill; 2 = borderline mentally ill; 3 = mildly ill; 4 = moderately ill; 5 = markedly ill; 6 = severely ill; and 7 = among the most extremely ill. An ordinal mixed-effects model, with the seven ordered categories recoded into four, is given in the section dealing with ordinal variables. In the present example, scores have been recoded to a dichotomous variable, where scores up to, but excluding 3.5 were coded 0, and scores of 3.5 or higher were coded 1. The value "0" is associated with measurements classified as normal, borderline, mildly, or moderately mentally ill, while the value "1" was assigned to measurements corresponding to "markedly ill" through "most extremely ill." In this study, patients were randomly assigned to receive one of four medications: placebo, chlorpromazine, fluphenazine, or thioridazine. Since previous analyses (Longford, 1993, and Gibbons & Hedeker, 1994) revealed similar effects for the three anti-psychotic drug groups, they were combined in the present analysis. Finally, again based on previous analysis, to linearize the relationship of the IMPS79 scores over time, a square root transformation of time was chosen.

Data for the first 10 participants on most of the variables used in this section are shown below in the form of a SuperMix spreadsheet file, named **schizx1.ss3**. The variables of interest are:

- o   Patient indicates the IDs for the 437 patients.
- o   Imps79 represents the original score on Item 79 of the Inpatient Multidimensional Psychiatric Scale.
- o   Imps79D is a recoded version of the same scale, but in binary form, with coding as discussed above.

- o Imps79O is also a recoded version of the same scale, but with the seven original categories reduced to four.
- o TxDrug indicates the treatment group, where 1 indicates membership in the treatment group, and 0 membership in the control group.
- o Week represents the time during the course of the study when a specific measurement was made, and ranges between 0 and 6.
- o SqrtWeek is the square root of Week.
- o Tx*SWeek is the product of the treatment group and the square root of Week.

	[A] Patient	[B] Imps79	[C] Imps79	[D] Imps79	[E] TxDrug	[F] Week	[G] SqrtWeek	[H] Tx*SWeek
1	1103	5.50	1	4	1	0	0.00	0.00
2	1103	3.00	0	2	1	1	1.00	1.00
3	1103	-9.00	-9	-9	1	2	1.41	1.41
4	1103	2.50	0	2	1	3	1.73	1.73
5	1103	-9.00	-9	-9	1	4	2.00	2.00
6	1103	-9.00	-9	-9	1	5	2.24	2.24
7	1103	4.00	1	2	1	6	2.45	2.45
8	1104	6.00	1	4	1	0	0.00	0.00
9	1104	3.00	0	2	1	1	1.00	1.00
10	1104	-9.00	-9	-9	1	2	1.41	1.41

In this data file, each subject's data consist of seven lines, these being the repeated measurements on seven occasions. Notice that there are missing value codes (–9) for some subjects at specific time points. The data from these time points will not be used in the analysis, but data from these subjects at other time points where there are no missing data will be used in the analysis. Thus, for inclusion into the analysis, a subject's data (both the dependent variable and all model covariates being used in a particular analysis) at a specific time point must be complete. The number of repeated observations per subject then depends on the number of time points for which there are non-missing data for that subject. The specification of missing data codes will be illustrated in the model specification section to follow.

3.5.2 The models

Continuous outcomes

A general two-level model for a continuous response variable y depending on a set of r predictors $x_1, x_2, \ldots, x_r$ can be written in the form

$$ y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{u}_i + \mathbf{e}_{ij} $$

where $i = 1, 2, \ldots, N$ denotes the level-2 units, and $j = 1, 2, \ldots, n_i$ the level-1 units. In this context, y_{ij} represents the response of individual j, nested within level-2 unit i. The model shown here consists of a fixed and a random part. The fixed part of the model is represented by the vector product $\mathbf{x}_{ij}^{'}\boldsymbol{\beta}$, where $\mathbf{x}_{ij}^{'}$ is a typical row of the design matrix of the fixed part of the model with, as elements, a subset of the r predictors. The vector $\boldsymbol{\beta}$ contains the fixed, but unknown parameters to be estimated. $\mathbf{z}_{ij}^{'}\mathbf{u}_i$ and $\mathbf{e}_{ij}$ denote the random part of the model at levels 2 and 1 respectively. For example, $\mathbf{z}_{ij}^{'}$ represents a typical row of the design matrix of the random part at level 2, and $\mathbf{u}_i$ the vector of random level-2 effects to be estimated. It is assumed that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ are independently and identically distributed (i.i.d.) with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(2)}$. Similarly, the $\mathbf{e}_{ij}$ are assumed i.i.d., with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$.

Within this hierarchical framework, the effects of TxDrug, SsqrtWeek, and Tx*SWeek can be used to predict the Imps79 score for the case where Imps79 is a continuous variable. The corresponding model may be expressed as

$$ \text{Imps79}_{ij} = \beta_0 + \beta_1 * \text{TxDrug}_{ij} + \beta_2 * \text{SqrtWeek}_{ij} + \beta_3 * \text{Tx*SWeek}_{ij} + u_{i0} + e_{ij} $$

where β_0 denotes the average expected Imps79 score, and β_1, β_2, and β_3 indicate the estimated coefficients associated with the fixed part of the model which contains the predictor variables TxDrug, SqrtWeek, and Tx*SWeek. The random part of the model is represented by u_{i0} and e_{ij}, which denote the variation in average score over patients and between measurements nested within patients at the lowest level of the hierarchy.

Binary outcomes

In the current example, the outcome variable is Imps79D, which is of a binary nature. The original scores on Item 79 of the Inpatient Multidimensional Psychiatric Scale have been recoded to a dichotomous variable, where scores up to, but excluding 3.5 were coded 0, and scores of 3.5 or higher were coded 1. In this case, the predicted value of the outcome can be viewed as the predicted probability that Imps79D is 1. Due to this, predicted values outside the interval (0,1) would not be meaningful and a model constraining predicted values to lie within this interval would be appropriate, in contrast with the model for a continuous outcome (see above) where predicted values outside this interval would be interpretable. In addition, the assumption of normality at level 1 is not realistic, as the level-1 random effect can only assume one of two values: 0 or 1. This random effect can thus not have homogeneous variance.

In order to insure that the predicted value lies within the (0,1) interval, a transformation of the level-1 predicted probability can be used. For the binary case considered here, we have

$$\text{Prob}(\text{Imps79D}_{ij} = 1 \mid \boldsymbol{\beta}, \mathbf{u}_i) = \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}$$

where η_{ij} represents the log of the odds of success, and (for the current model) can be expressed as

$$\eta_{ij} = \beta_0 + \beta_1 * \text{TxDrug}_{ij} + \beta_2 * \text{SqrtWeek}_{ij} + \beta_3 * \text{Tx*SWeek}_{ij} + u_{i0} + e_{ij}.$$

This transformation, commonly referred to as the logit link function, constrains $\text{Prob}(y_{ij} = 1 \,|\, \mathbf{u}_i)$ to lie in the interval $(0,1)$.

3.5.3 Example: Logistic regression with a random intercept

Importing the data

The model is fitted to the data in **schizx1.ss3**. The first step is to create the **ss3** file **schizx1.ss3** from an Excel spreadsheet named **schizx1.xls**. This is accomplished as follows:

- o Use the **File**, **Import Data File** option to activate the display of an **Open** dialog box.
- o Browse for the file **schizx1.xls** in the **Examples\Primer\Binary** folder.
- o Select the file and click the **Open** button to return to the main SuperMix window, where the contents of the Excel spreadsheet are displayed as the SuperMix system file with default name **schizx1.ss3**.

	(A) Patient	(B) Imps79	(C) Imps79	(D) Imps79	(E) TxDrug	(F) Week	(G) SqrtWeek	(H) Tx*SWeek
1	1103	5.50	1	4	1	0	0.00	0.00
2	1103	3.00	0	2	1	1	1.00	1.00
3	1103	-9.00	-9	-9	1	2	1.41	1.41
4	1103	2.50	0	2	1	3	1.73	1.73
5	1103	-9.00	-9	-9	1	4	2.00	2.00
6	1103	-9.00	-9	-9	1	5	2.24	2.24
7	1103	4.00	1	2	1	6	2.45	2.45
8	1104	6.00	1	4	1	0	0.00	0.00
9	1104	3.00	0	2	1	1	1.00	1.00
10	1104	-9.00	-9	-9	1	2	1.41	1.41

Setting up the analysis

The next step is to describe the model to be fitted. We use the SuperMix interface to provide the model specifications. From the main menu bar, select the **File**, **New Model Setup** option.
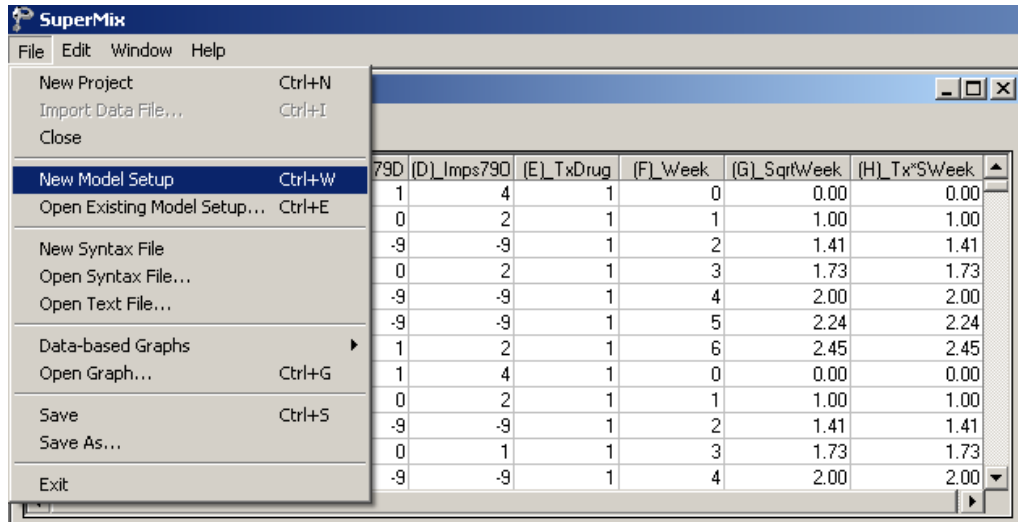
The **Model Setup** window that appears has six tabs. In this example, three of these tabs are used in model specification.

As a first step, select the dichotomous outcome variable Imps79D from the **Dependent Variable** drop-down list box on the **Configuration** screen. Specify the type of outcome as binary using the **Dependent Variable Type** drop-down list box. Once this selection is made, the **Categories** grid is displayed. As there are missing data in the **ss3** file for both outcome and potential predictors, set the **Missing Values Present** drop-down list box to true. Once this is done, the **Missing Values for Dependent Var** and **Global Missing Value** text boxes are displayed. Enter the value –9 into both these boxes.

The patient identification variable is used to define the hierarchical structure of the data, and is selected as the Level-2 ID from the **Level-2 IDs** drop-down list box. Enter a title for the analysis in the **Title 1** and **Title 2** text boxes. Request a crosstabulation of the outcome variable and the predictor SqrtWeek by setting the **Perform Crosstabulation** and **Crosstab Variable** drop-down list boxes to **yes** and SqrtWeek respectively.

The **Variables** screen is used to specify the fixed and random effects to be included in the model. Start by selecting the explanatory (fixed) variables TxDrug, SqrtWeek, and Tx*SWeek by checking the check boxes in the **E** column of the **Available** grid.
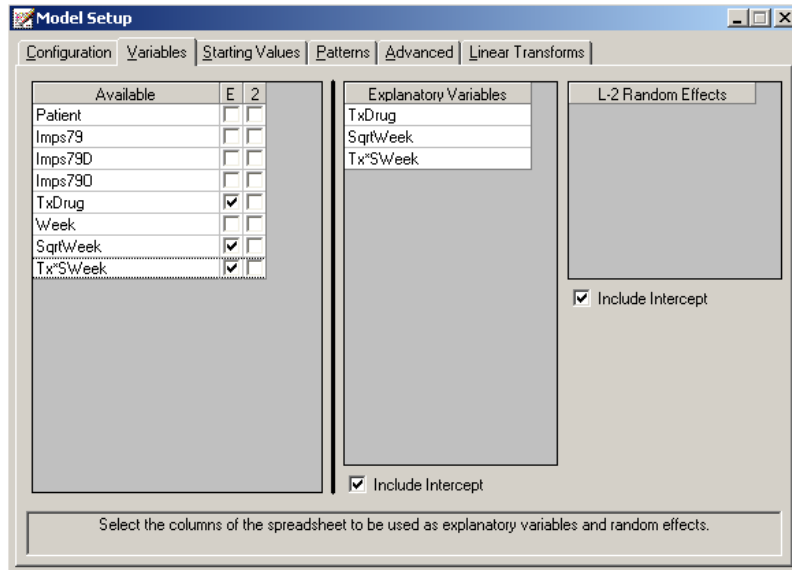
After selecting all the explanatory variables, the random effect(s) at level 2 must be selected. By default, the model will include a random intercept, as indicated by the check box for **Include Intercept** in the **L-2 Random Effects** grid. The intercept is assumed to vary randomly over higher levels of the hierarchy, while the slopes of the predictors TxDrug, SqrtWeek and the interaction between treatment and the square root of the treatment time, Tx*SWeek, are assumed to be adequately described by common, fixed coefficients that do not vary across patients.

Next, click on the **Advanced** tab. This screen is used to specify additional settings for the case where the outcome variable is binary. Request the use of 20 **non-adaptive quadrature** points for estimation by entering the number 20 in the **Number of Quadrature Points** text box. No changes are made to the **Unit Weighting** box. Use the default **Bernoulli** distribution and keep all the other default settings. The completed **Advanced** screen is shown below.
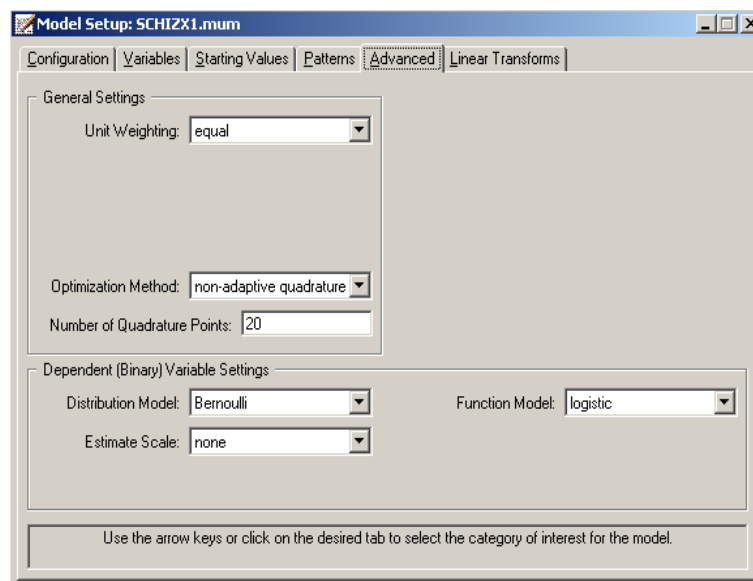
Before running the analysis, the model specifications have to be saved. Select the **File**, **Save** option, and provide a name for the model specification file. Run the analysis by selection the **Run** option from the **Analysis** menu.

Discussion of results

Portions of the output file **schizx1.out** are shown below.

In the first section of the output file, a summary of the model specifications is provided. The use of a logistic response function (logit link function), with the assumption of a Bernoulli distribution of random effects, is indicated. This is followed by a summary of the number of observations nested within each patient. The Level 2 observations entry corresponds to the number of patients for whom data were included in the analysis.
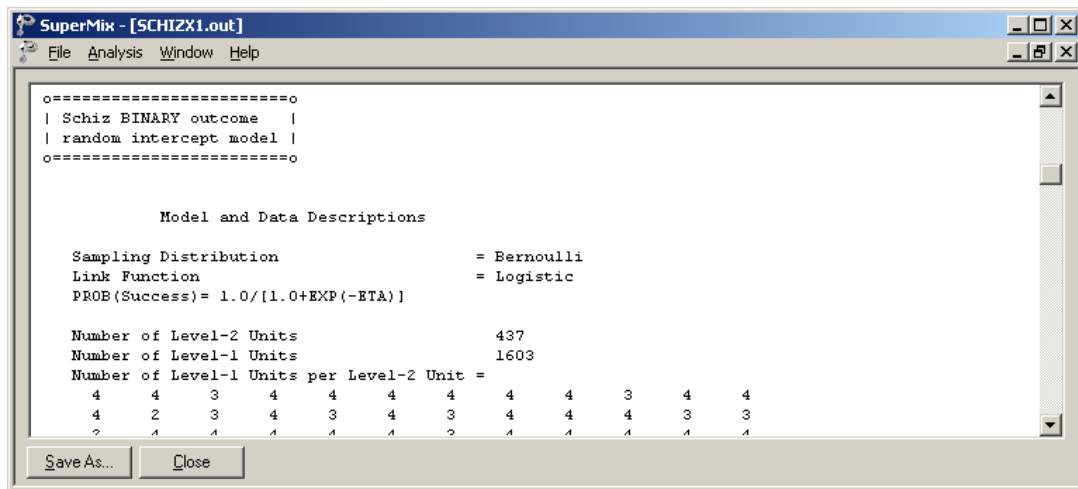
```
o========================o
| Schiz BINARY outcome   |
| random intercept model |
o========================o


            Model and Data Descriptions

   Sampling Distribution               = Bernoulli
   Link Function                       = Logistic
   PROB(Success)= 1.0/[1.0+EXP(-ETA)]

   Number of Level-2 Units             437
   Number of Level-1 Units             1603
   Number of Level-1 Units per Level-2 Unit =
       4    4    3    4    4    4    4    4    3    4    4
       4    2    3    4    3    4    3    4    4    4    3    3
       2    4    4    4    4    4    2    4    4    4    4    4
```

The data summary is followed by descriptive statistics for all the variables included in the model. We note that 23% of the patients had a value of 0 on the binary Imps79D score which indicates no or moderate signs of mental illness (Imps79D = 0).

```
o=============================================================o
| Descriptive statistics for all the variables in the model |
o=============================================================o
                                               Standard
   Variable    Minimum    Maximum      Mean    Deviation
   --------    -------    -------      ----    ---------
   Imps79D1     0.0000     1.0000     0.2321    0.4223
   Imps79D2     0.0000     1.0000     0.7679    0.4223
   TxDrug       0.0000     1.0000     0.7642    0.4246
   SqrtWeek     0.0000     2.4495     1.2204    0.8965
   Tx*SWeek     0.0000     2.4495     0.9442    0.9454
```
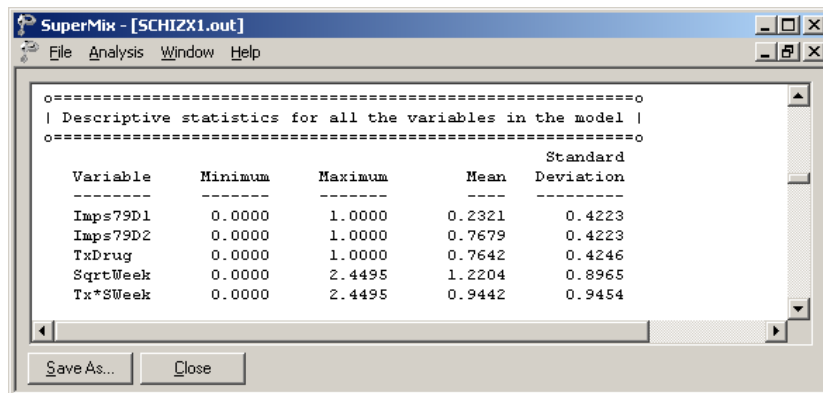
The crosstabulation of the outcome variable and the predictor SqrtWeek requested on the **Variables** screen during model specification is listed next. We see that most of the measurements (1231) are from the higher, more ill, category of Imps79D. It is also noticeable that more data were obtained at the start, the fourth time point, and

the end of the study than on the remaining occasions. The results for the model without any random effects serve as starting values for the iterative procedure.

```
SuperMix - [SCHIZX1.out]                                   _ □ X
File  Analysis  Window  Help                               _ 🗗 X

Two-way table of Response Variable by SqrtWeek
-----------------------------------------------

     Y-Category     X-Category      Frequency
     ----------     ----------      ---------
          0           0.00              6
          0           1.00             67
          0           1.41              5
          0           1.73            108
          0           2.00              7
          0           2.24              6
          0           2.45            173
          1           0.00            428
          1           1.00            359
          1           1.41              9
          1           1.73            266
          1           2.00              4
          1           2.24              3
          1           2.45            162

                     Total:          1603

Save As...      Close
```

```
SuperMix - [SCHIZX1.out]                                                       _ □ X
File  Analysis  Window  Help                                                   _ 🗗 X

   o==============================================o
   | Optimization Method: Non-Adaptive Quadrature |
   o==============================================o

   Number of quadrature points =          20
   Number of free parameters =             5
   Number of iterations used =             5

   -2lnL (deviance statistic) =      1249.73459
   Akaike Information Criterion      1259.73459
   Schwarz Criterion                 1286.63275


              Estimated regression weights

                                    Standard
   Parameter        Estimate        Error       z Value    P Value
   ---------        --------        --------     -------    -------
   intercept         5.3875          0.6307      8.5414     0.0000
   TxDrug           -0.0246          0.6535     -0.0377     0.9699
   SqrtWeek         -1.4999          0.2907     -5.1601     0.0000
   Tx*SWeek         -1.0149          0.3340     -3.0391     0.0024

Save As...      Close
```
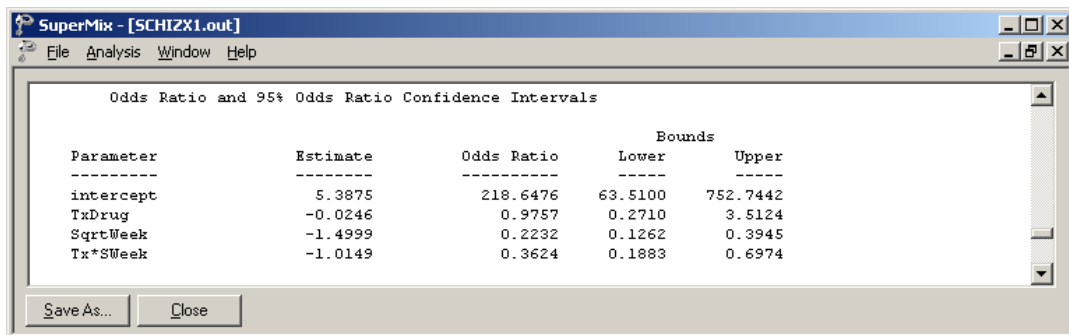
The output summarizing the estimated parameters after convergence is shown next. Five iterations were required to obtain convergence. The estimates are shown in the column with heading Estimate, and correspond to the coefficients β_0, β_1, β_2 and β_3 in the model specification. No estimate of the level-1 variance is given, as there is no single level-1 variance for this model.

The expected log-odds of having a high score at the end of the study period (Imps79D score of 1) for a patient from the control group (that is, TxDrug = Tx*SWeek = SqrtWeek = 0) is represented by the estimated intercept of 5.3905. The negative and statistically significant coefficient for SqrtWeek indicates that the probability of having a high score decreases with time and more so when the interaction between receiving treatment and time is taken into account (Tx*SWeek = −1.0158, p = 0.0024).

```
SuperMix - [SCHIZX1.out]                                              _ □ X
File  Analysis  Window  Help                                          _ 日 X

      Odds Ratio and 95% Odds Ratio Confidence Intervals

                                                    Bounds
      Parameter        Estimate     Odds Ratio    Lower      Upper
      ---------        --------     ----------    -----      -----
      intercept          5.3875      218.6476     63.5100    752.7442
      TxDrug            -0.0246        0.9757      0.2710      3.5124
      SqrtWeek          -1.4999        0.2232      0.1262      0.3945
      Tx*SWeek          -1.0149        0.3624      0.1883      0.6974

   Save As...     Close
```

As shown below, there is also evidence of significant random variation in the intercepts over patients.

```
SuperMix - [SCHIZX1.out]                                              _ □ ×
 File  Analysis  Window  Help                                          _ 8 ×

            Estimated level 2 variances and covariances

                                       Standard
      Parameter              Estimate    Error     z Value   P Value
      ---------              --------   --------   -------   -------
      intercept/intercept      4.4797    0.9473    4.7290    0.0000

 Save As...      Close
```

Estimated outcomes for groups: unit-specific probabilities

To evaluate the expected effect of TxDrug, SqrtWeek, and Tx*SWeek on the predicted probability that the Imps79D score is equal to 1, we use the expression for the predicted log odds of success given earlier

$$\hat{\eta}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 * \text{TxDrug}_{ij} + \hat{\beta}_2 * \text{SqrtWeek}_{ij} + \hat{\beta}_3 * \text{Tx*SWeek}_{ij}.$$

For a typical measurement from any patient from the control group at the beginning of the study period (TxDrug = SqrtWeek = Tx*SWeek = 0) we have

$$\hat{\eta}_{ij} = \hat{\beta}_0.$$

For a typical measurement from any patient from the treatment group at the beginning of the study period, $\hat{\eta}_{ij}$ can be expressed as

$$\hat{\eta}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 * \text{TxDrug}_{ij}$$
$$= \hat{\beta}_0 + \hat{\beta}_1.$$

Using the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ of 5.3875 and -0.0246 respectively as obtained for the current analysis, we can calculate the subject-specific probability of an Imps79D score of 1 as

$$\text{Prob}(\text{Imps79D}_{ij} = 1 \mid \text{TxDrug} = \text{SqrtWeek} = \text{Tx*SWeek} = 0) = \frac{e^{5.3875}}{1 + e^{5.3875}}$$

$$= 0.9955$$

and

$$\text{Prob}(\text{Imps79D}_{ij} = 1 \mid \text{TxDrug} = 1; \text{SqrtWeek} = \text{Tx*SWeek} = 0) = \frac{e^{5.3875 - 0.0246}}{1 + e^{5.3875 - 0.0246}}$$

$$= 0.9953$$

respectively.

On the other end of the scale in terms of time, we can consider patients from both control and treatment groups at the end of the study period. For both groups, this corresponds to a value of 2.45 on the variable SqrtWeek.

For a typical measurement from any patient from the control group at the end of the study period we have

$$\hat{\eta}_{ij} = \hat{\beta}_0 + \hat{\beta}_2 * \text{SqrtWeek}_{ij}$$
$$= \hat{\beta}_0 + 2.45 \times \hat{\beta}_2$$
$$= 1.7125.$$

For a typical measurement from any patient from the treatment group at the end of the study period, $\hat{\eta}_{ij}$ can be expressed as

$$\hat{\eta}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 * \text{TxDrug}_{ij} + \hat{\beta}_2 * \text{SqrtWeek}_{ij} + \hat{\beta}_3 * \text{Tx*SWeek}_{ij}$$
$$= \hat{\beta}_0 + \hat{\beta}_1 + 2.45\left(\hat{\beta}_2 + \hat{\beta}_3\right)$$
$$= -0.7984.$$

In Table 3.2, the predicted probabilities of a high post-treatment Imps79D score are given for patients from both experimental groups at selected time points during the study period.

Table 3.2: Predicted probability of a high post-treatment Imps79D score

Square root of week (SqrtWeek)	Control group	Treatment group
0.00	0.9954	0.9953
1.73	0.9423	0.7335
2.45	0.8472	0.3104
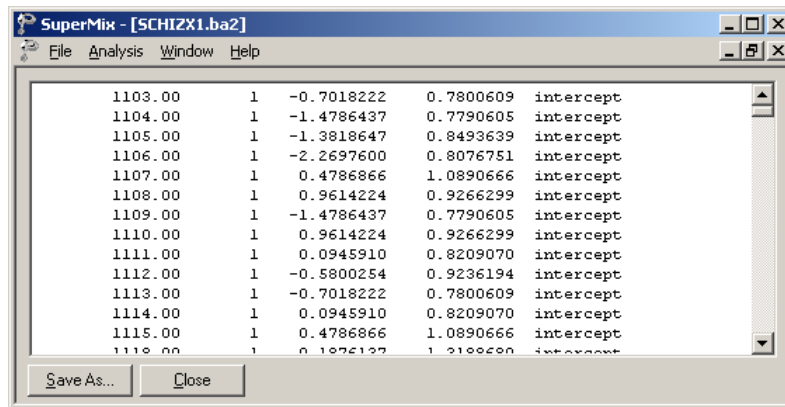
We conclude that the diagnoses for both control and treatment groups improved over the seven week study period. There is, however, a marked decrease in $P(\text{Imps79D}) = 1$ for the treatment group. Looking at the differences of probabilities at the end of the trial, we can draw the conclusion that the treatment effect is highly significant.

TWO-LEVEL MODELS FOR BINARY OUTCOMES

Estimated outcomes for different groups: population-average results

Table 3.2 contains estimated unit (patient) specific probabilities. To obtain population-average probabilities, the estimated η_{ij} - values are divided by the square root of the design effect. These adjusted η_{ij} -values are subsequently used in the computation of the probabilities. This aspect is discussed in detail in Section 3.6.

Finally, the random variation in intercept over patients is estimated at 4.4797.

In addition to the standard output file **schizx1.out**, the **Write Bayes Estimates** drop-down list box on the **Configuration** screen of the **Model Setup** window was used to request Bayes estimates for the individual random terms. This is done by selecting the **means & (co)variances** option from the **Write Bayes Estimates** drop-down list box. These estimates are written to the file **schizx1.ba2**. The first few lines of this file are shown below.


```
 SuperMix - [SCHIZX1.ba2]
 File   Analysis   Window   Help

      1103.00      1    -0.7018222    0.7800609    intercept
      1104.00      1    -1.4786437    0.7790605    intercept
      1105.00      1    -1.3818647    0.8493639    intercept
      1106.00      1    -2.2697600    0.8076751    intercept
      1107.00      1     0.4786866    1.0890666    intercept
      1108.00      1     0.9614224    0.9266299    intercept
      1109.00      1    -1.4786437    0.7790605    intercept
      1110.00      1     0.9614224    0.9266299    intercept
      1111.00      1     0.0945910    0.8209070    intercept
      1112.00      1    -0.5800254    0.9236194    intercept
      1113.00      1    -0.7018222    0.7800609    intercept
      1114.00      1     0.0945910    0.8209070    intercept
      1115.00      1     0.4786866    1.0890666    intercept
      1118.00      1     0.1876137    1.3188680    intercept

 Save As...      Close
```

Four pieces of information per patient are given:

- o   the level-3 ID (if any),
- o   the number of the random effect,

o the empirical Bayes estimate ($\hat{u}_{i0}$), and

o the estimated associated posterior variance of the patient estimate $\overset{\wedge}{\text{var}}\left(u_{i0}\right)$.

From the data, we know that the first patient, with ID = 1103, was a member of the treatment group. On the other hand, the patient with ID = 1107 was assigned to the control group. At the beginning of the study, the predictor SqrtWeek and hence the interaction term Tx*SWeek were equal to zero. In both cases, the observed Imps79D score at the beginning of the study was equal to 1. Their empirical Bayes estimates were −0.7018 and 0.4787 respectively. For measurements from each of these patients, this implies

$$\text{Prob}(\text{Imps79D}_{ij} = 1 \mid \text{TxDrug} = 1, \text{SqrtWeek=0,ID=1103}) = \frac{e^{5.3875 - 0.7018 - 0.0246}}{1 + e^{5.3875 - 0.7018 - 0.0246}}$$

$$= \frac{e^{4.6611}}{1 + e^{4.6611}} = 0.9906$$

and

$$\text{Prob}(\text{Imps79D}_{ij} = 1 \mid \text{TxDrug} = 0, \text{SqrtWeek=0,ID=1107}) = \frac{e^{5.3875 + 0.4787}}{1 + e^{5.3875 + 0.4787}}$$

$$= 0.9972$$

respectively.

Similarly, we find that at the end of the study

$$\text{Prob}(\text{Imps79D}_{ij} = 1 \mid \text{TxDrug} = 1, \text{SqrtWeek}=2.45, \text{ID}=1103)$$

$$= \frac{e^{5.3875-0.7018-0.0246-1.4999(2.45)-1.0149(2.45)}}{1+e^{5.3875-0.7018-0.0246-1.4999(2.45)-1.0149(2.45)}}$$

$$= \frac{e^{-1.5002}}{1+e^{-1.5002}}$$

$$= 0.1824$$

and

$$\text{Prob}(\text{Imps79D}_{ij} = 1 \mid \text{TxDrug} = 0, \text{SqrtWeek}=2.45, \text{ID}=1107)$$

$$= \frac{e^{5.3875+0.4787-1.4999(2.45)}}{1+e^{5.3875+0.4787-1.4999(2.45)}}$$

$$= \frac{e^{2.1914}}{1+e^{2.1914}}$$

$$= 0.8995.$$

Using the empirical Bayes estimates for each patient, we can also obtain the predicted probability of improving.

This probability, $\text{Prob}(\text{Imps79D}_{ij} = 0 \mid ID = i, \quad i = 1, 2, ..., N)$ is obtained as $1 - \left(\text{Prob}(\text{Imps79D}_{ij} = 1 \mid ID = i, \quad i = 1, 2, ..., N)\right)$. In Figures 3.11 and 3.12 below, the predicted probabilities of improving for all measurements are shown, with a linear regression line showing the mean probability for each group.
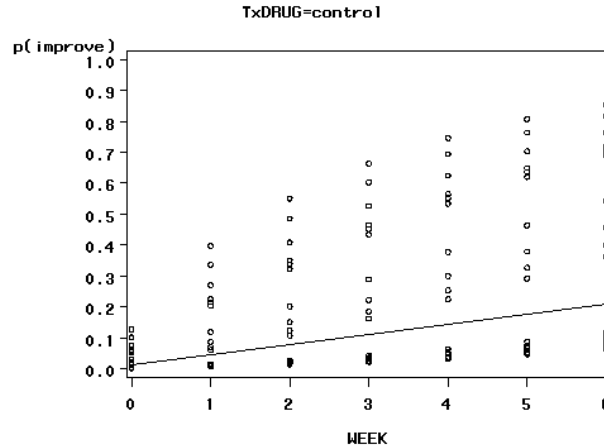
Figure 3.11: Predicted probability of improvement (control group)
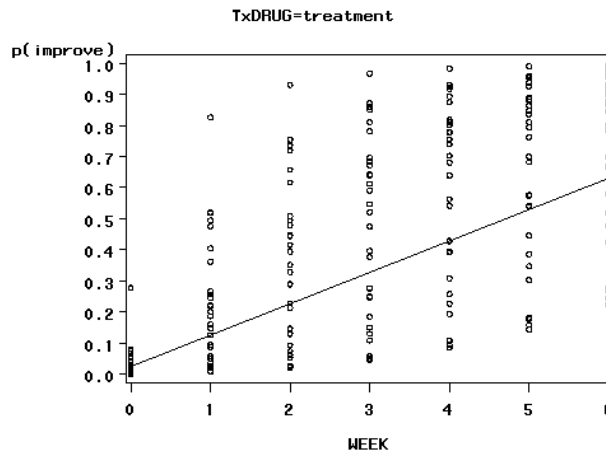
Figure 3.12: Predicted probability of improvement (treatment group)

For both groups, an increased probability of obtaining a predicted value of 0 on the outcome is observed. Recall that the value "0" is associated with measurements classified as normal, borderline, mildly, or moderately mentally ill, while the value "1" was assigned to measurements corresponding to "markedly ill" through "most extremely ill". However, the patients in the treatment group were much more likely

TWO-LEVEL MODELS FOR BINARY OUTCOMES

to show improvement, as can be seen from the slopes of the regression lines for the groups in Figure 3.13.
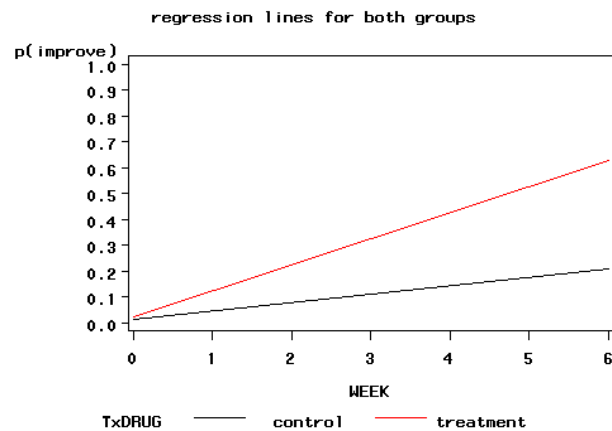
Figure 3.13: Regression lines for control and treatment groups

3.6 Two-level models for ordinal outcomes

3.6.1 The data

To illustrate the application of the mixed-effects ordinal probit regression model to longitudinal data, we examined data collected in the NIMH Schizophrenia Collaborative Study on treatment related changes in overall severity. A detailed description of these data is given in Section 3.5. For the present illustration of the ordinal mixed-effects model, we recoded the seven ordered categories of Imps79 into four: 1) normal or borderline mentally ill, 2) mildly or moderately ill, 3) markedly ill, and 4) severely or among the most extremely ill.

For both models discussed in this section and the next example, the repeated ordinal IMPS score (Imps79O) is modeled in terms of a dummy-coded drug effect (TxDrug: placebo = 0 and drug = 1), a time effect (SqrtWeek: square root of week) and a drug by time interaction (Tx*SWeek). In terms of the random effects, this example specifies a random intercepts model, while the next example will show a model that makes allowance for patients to vary in terms of both their intercept and their trend across time (random intercepts and slopes).

3.6.2 The models

To motivate the ordinal regression model, it is often assumed that there is an unobservable latent variable (y) which is related to the actual response through the "threshold concept". For the dichotomous model, one threshold value is assumed, while for the ordinal model, a series of threshold values $\gamma_1, \gamma_2, ..., \gamma_{K-1}$, where K equals the number of ordered categories, $\gamma_0 = -\infty$, and $\gamma_J = \infty$, is assumed. Here, a response occurs in category k ($Y = k$) if the latent response process y exceeds the threshold value γ_{k-1}, but does not exceed the threshold value γ_k.

Assume that there are $i = 1, \ldots, N$ level-2 units and $j = 1, \ldots, n_i$ level-1 units nested within each level-2 unit. The mixed-effects regression model for the latent response variable y_{ij} can be written as follows:

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + \mathbf{z}_{ij}' \mathbf{u}_i + e_{ij}$$

where $\mathbf{x}_{ij}$ is the $p \times 1$ covariate vector and $\mathbf{z}_{ij}$ is the design vector for the r random effects, both vectors being for the j-th level-1 unit nested within level-2 unit i. Also, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown fixed regression parameters, $\mathbf{u}_i$ is the $r \times 1$ vector of unknown random effects for the level-2 unit i, and e_{ij} is the model residual. The distribution of the random effects is assumed to be multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(2)}$, and the residuals are assumed to be independently normally distributed with mean 0 and variance σ^2.

A model with probit link function and random intercept

In the first example of this section, we consider the model

$$y_{ij} = \beta_0 + \beta_1 \mathrm{TxDrug}_{ij} + \beta_2 \mathrm{SqrtWeek}_{ij} + \beta_3 \mathrm{Tx*SWeek}_{ij} + u_{i0} + e_{ij}$$

With the above mixed-effects regression model for the underlying and unobservable variable y_{ij}, the probability, for a given level-2 unit i, that $Y_{ij} = k$ (a response occurs in category k), conditional on $\mathbf{u}_i$ and $\boldsymbol{\beta}$, is given by the following equation:

$$P(Y_{ij} = k \mid \mathbf{u}_i, \boldsymbol{\beta}) = \Phi[(\gamma_k - \eta_{ij})/\sigma] - \Phi[(\gamma_{k-1} - \eta_{ij})/\sigma]$$

where $\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i$ and $\Phi(\cdot)$ represents the cumulative standard normal density function. Without loss of generality, the origin and unit of η may be chosen arbitrarily. For convenience, let $\gamma_1 = 0$ and $\sigma = 1$, then

$$P(\text{Imps79O}_{ij} = k \mid \mathbf{u}_i, \boldsymbol{\beta}) = \Phi(\gamma_k - \eta_{ij}) - \Phi(\gamma_{k-1} - \eta_{ij}).$$

A model with probit link function with random intercept and slope

In the second example, we again consider the model

$$y_{ij} = \beta_0 + \beta_1\text{TxDrug}_{ij} + \beta_2\text{SqrtWeek}_{ij} + \beta_3\text{Tx*SWeek}_{ik} + u_{i0} + u_{i1}\text{SqrtWeek}_{ij} + e_{ij}$$

where the vector of random coefficients $\mathbf{u}_i$ in the mixed-effects regression model

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i + e_{ij}$$

is extended to include both intercept and slope (SqrtWeek) effects.

3.6.3 Example: Probit link function with random intercept

Importing the data

The model is fitted to the data in **schizx1.ss3**. Use the **File**, **Open** option to activate an **Open** dialog box, and browse for the file **schizx1.ss3** in the **Examples\Primer\Ordinal** folder.

Setting up the analysis

The next step is to describe the model to be fitted. We use the SuperMix interface to provide the model specifications. From the main menu bar, select the **File**, **New Model Setup** option.

The **Model Setup** window that appears has six tabs. In this example, the **Configuration**, **Variables**, and **Advanced** tabs are used in model specification. By default, the **Configuration** screen is displayed on top.

As a first step, enter a title for the analysis in the **Title 1** and **Title 2** text boxes. The patient identification variable (Patient) is used to define the hierarchical structure of the data, and is selected as the Level-2 ID from the **Level-2 IDs** drop-down list box. Select the ordinal outcome variable Imps79O from the **Dependent Variable** drop-down list box. Specify the type of outcome as ordered using the **Dependent Variable Type** drop-down list box. Once this selection is made, the **Categories** grid is displayed, with the distinct values of the categories shown.

The **Missing Values Present** drop-down list box is used to specify the values of missing data for both outcome and predictors. As a first step, set the value of the **Missing Values Present** drop-down list box to **True**. The appearance of the screen will change when this is done, and text boxes for the specification of the missing data codes are displayed. Start by entering the value –9 in the **Missing Value for the Dependent Var** text box. Do the same for all the predictors included in the model by entering –9 in the **Global Missing Value** text box. In terms of the missing value specification, notice that missing values are coded only for the dependent variable in the input data file and therefore the specification of a missing code in the **Global Missing Value** text box is optional. In this case, the value –9 was specified for all variables since for the dependent variable this value is the correct missing value code, while for all other model terms (Intcept, TxDrug, SqrtWeek, and Tx*SWeek) this value was never observed. Once –9 is entered in the **Missing Value for Dependent Var** text box, the categories grid will automatically be updated to display 1, 2, 3, and 4 only.

Request a crosstabulation of the variable SqrtWeek by the response variable Imps79O by selecting the **yes** option from the **Perform Crosstabulation** drop-down list box,

followed by the selection of SqrtWeek as the **Crosstab Variable**. Finally, request the writing out of Bayes estimates using the **Write Bayes Estimates** drop-down list box. The default setting for the **Number of Iterations** text box is used. Proceed to the **Variables** screen by clicking on this tab.

The **Variables** screen is used to specify the fixed and random effects to be included in the model. Select the explanatory (fixed) variables TxDrug, SqrtWeek, and Tx*SWeek using the check boxes next to the variable names in the **E** column of the **Available** grid. After selecting all the explanatory variables, the random effect(s) at level 2 must be selected. By default, a random intercept at level 2 is assumed. The intercept is assumed to vary randomly over higher levels of the hierarchy, while the slopes of the predictors TxDrug, SqrtWeek, and Tx*SWeek are assumed to be adequately described by common, fixed coefficients. The completed **Variables** screen is shown above. To change estimation settings such as the number of quadrature points used in estimation, the **Advanced** screen is used. Click **Advanced** to access this screen and complete the model specification process.

Set the **Function Model** option to **probit**. All other settings on this screen are left at their previous values, and we can save our model specification to file prior to running the analysis. Select the **File**, **Save** option, provide a name for the model specification file (such as **schizx2.mum**), and run the analysis by selecting the **Run** option from the **Analysis** menu.
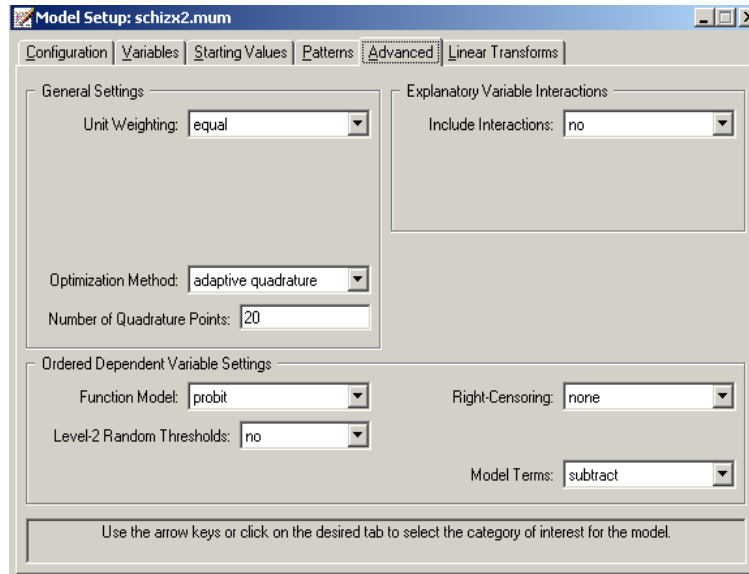
Discussion of results

Portions of the output file **schizx2.out** are shown below.
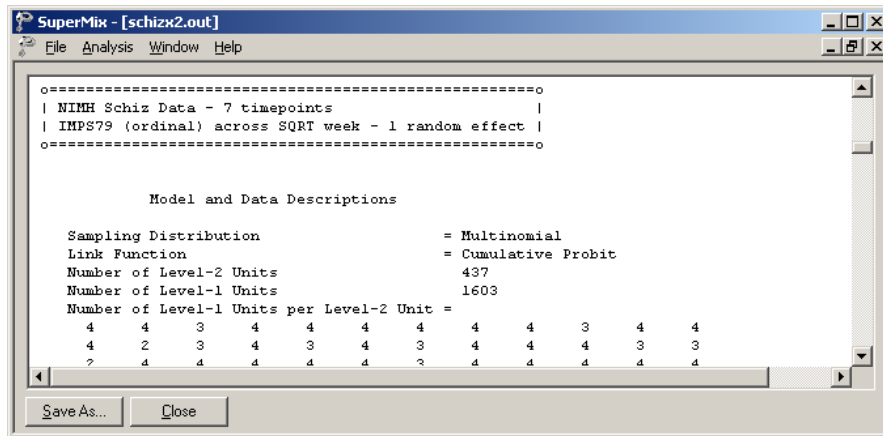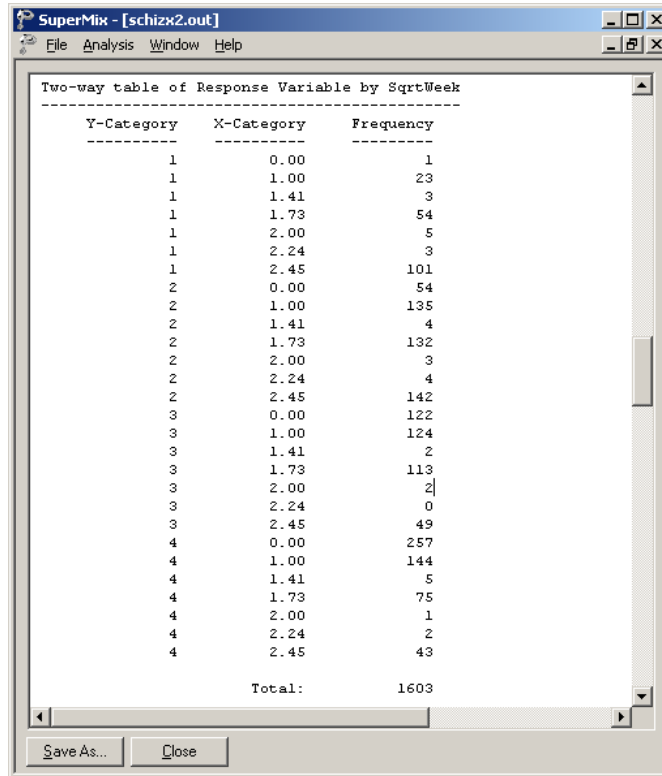
In the first section of the output file, a description of the model specifications is provided. The use of a multinomial response function with cumulative probit link function is noted. This is followed by a summary of the number of measurements nested within each subject. Note that for some of the 437 subjects, only 2 measurements are available for analysis.
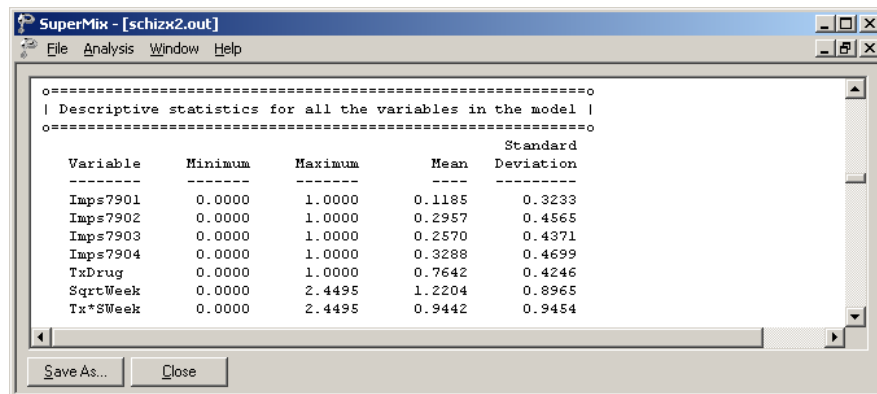
The data summary is followed by descriptive statistics for all the variables included in the model.

The following portion of the output is a crosstabulation of the seven distinct values of the variable SqrtWeek by the four categories of the outcome variable IMPS79O. We note that there are relatively few observations for the third, fifth and sixth weeks. For example, for week 5 (SqrtWeek = 2.24), measurements on only 9 of the 437 patients are available.

```
 SuperMix - [schizx2.out]                                    _ □ ×
 File  Analysis  Window  Help                                _ ₽ ×

 Two-way table of Response Variable by SqrtWeek
 -------------------------------------------------
     Y-Category    X-Category      Frequency
     ----------    ----------      ---------
              1        0.00               1
              1        1.00              23
              1        1.41               3
              1        1.73              54
              1        2.00               5
              1        2.24               3
              1        2.45             101
              2        0.00              54
              2        1.00             135
              2        1.41               4
              2        1.73             132
              2        2.00               3
              2        2.24               4
              2        2.45             142
              3        0.00             122
              3        1.00             124
              3        1.41               2
              3        1.73             113
              3        2.00               2
              3        2.24               0
              3        2.45              49
              4        0.00             257
              4        1.00             144
              4        1.41               5
              4        1.73              75
              4        2.00               1
              4        2.24               2
              4        2.45              43

                  Total:              1603

   Save As...      Close
```

The results for the model without any random effects are given next.

```
┌─────────────────────────────────────────────────────────────────────────────┐
│ ᴾ SuperMix - [schizx2.out]                                      _□×          │
│ ᴾ File  Analysis  Window  Help                                  _▌×          │
├─────────────────────────────────────────────────────────────────────────────┤
│                                                                         ▲    │
│   o==================================================o                       │
│   | Results for the model without any random effects |                       │
│   o==================================================o                       │
│                                                                              │
│                     Goodness of fit statistics                               │
│                                                                              │
│   Statistic                          Value         DF         Ratio          │
│   ---------                          -----         --         -----          │
│   Likelihood Ratio Chi-square       3745.8408      1597        2.3455         │
│   Pearson Chi-square                4442.2849      1597        2.7816         │
│                                                                              │
│                                                                              │
│                   Estimated regression weights                               │
│                                                                              │
│                                     Standard                                 │
│   Parameter          Estimate       Error       z Value     P Value          │
│   ---------          --------       --------     -------     -------          │
│   Threshold1         -2.2802        0.1109      -20.5570      0.0000          │
│   Threshold2         -1.0865        0.1017      -10.6789      0.0000          │
│   Threshold3         -0.2878        0.0993       -2.8978      0.0038          │
│   TxDrug             -0.0308        0.1136       -0.2716      0.7860          │
│   SqrtWeek           -0.3426        0.0658       -5.2071      0.0000          │
│   Tx*SWeek           -0.4202        0.0752       -5.5886      0.0000      ▼    │
│ ◄                                                                    ►        │
│   [ Save As... ]   [ Close ]                                                 │
└─────────────────────────────────────────────────────────────────────────────┘

┌─────────────────────────────────────────────────────────────────────────────┐
│ ᴾ SuperMix - [schizx2.out]                                      _□×          │
│ ᴾ File  Analysis  Window  Help                                  _▌×          │
├─────────────────────────────────────────────────────────────────────────────┤
│                                                                         ▲    │
│   Alternative Parameterization, setting Threshold1= 0                        │
│                                                                              │
│                   Estimated regression weights                               │
│                                                                              │
│                                     Standard                                 │
│   Parameter          Estimate       Error       z Value     P Value          │
│   ---------          --------       -------      -------     -------          │
│   intcept            2.2802         0.1109       20.5570      0.0000          │
│   Threshold2         1.1937         0.0483       24.6951      0.0000          │
│   Threshold3         1.9924         0.0557       35.7824      0.0000      ▼    │
│ ◄                                                                    ►        │
│   [ Save As... ]   [ Close ]                                                 │
└─────────────────────────────────────────────────────────────────────────────┘
```

The output summarizing the estimated parameters after convergence is shown next.
Four iterations were required to obtain convergence. The estimates are shown in the
column with heading Estimate, and correspond to the coefficients β_0, β_1, β_2, and β_3
in the model specification. The effect of TxDrug was the only nonsignificant effect.
The alternative parameterization obtained by setting the first threshold equal to zero
is also given in the output.

The variation in the intercept over the subjects is estimated as 1.2278, and from the associated *p*-value we conclude that there is significant variation in the (random) intercept between the patients included in this analysis. In the case of the fixed effects, a 2-tailed *p*-value is used, as the alternative hypothesis considered here is of the form $H_1 : \beta \neq 0$. As variances are constrained to be elements of the interval $[0, +\infty)$, the *p*-values used for these effects are 1-tailed. No estimate of the level-1 variance is given, as there is no single level-1 variance for this model.

```
SuperMix - [schizx2.out]                                          _ □ ×
File   Analysis   Window   Help                                   _ ₈ ×

    o=========================================o
    | Optimization Method: Adaptive Quadrature |
    o=========================================o

    Number of quadrature points =           20
    Number of free parameters =              7
    Number of iterations used =              4

    -2lnL (deviance statistic) =       3399.47482
    Akaike Information Criterion       3413.47482
    Schwarz Criterion                  3451.13224


                      Estimated regression weights

                                     Standard
    Parameter          Estimate      Error        z Value      P Value
    ---------          --------      --------     -------      -------
    Threshold1          -3.3666       0.1840      -18.2949      0.0000
    Threshold2          -1.6372       0.1642       -9.9679      0.0000
    Threshold3          -0.4267       0.1575       -2.7097      0.0067
    TxDrug              -0.0517       0.1791       -0.2887      0.7728
    SqrtWeek            -0.4591       0.0747       -6.1457      0.0000
    Tx*SWeek            -0.6723       0.0861       -7.8068      0.0000


    Alternative Parameterization, setting Threshold1= 0

                      Estimated regression weights

                                     Standard
    Parameter          Estimate      Error        z Value      P Value
    ---------          --------      --------     -------      -------
    intcept             3.3666        0.1840       18.2949      0.0000
    Threshold2          1.7294        0.0747       23.1444      0.0000
    Threshold3          2.9398        0.0955       30.7916      0.0000


              Estimated level 2 variances and covariances

                                   Standard
    Parameter             Estimate   Error      z Value   P Value
    ---------             --------   --------   -------   -------
    intercept/intercept     1.2278    0.1458     8.4200    0.0000

Save As...      Close
```

The results indicate that the treatment groups do not differ significantly at baseline (TxDrug coefficient is not significant). The placebo group seems to improve over time, as the SqrtWeek coefficient is both significant and negative. Note that the interpretation of the main effects depends on the coding of the variable, and significance of, the Tx*SWeek interaction which forms part of the model.

The connection between an ordinal outcome variable y with NCAT categories and an underlying continuous variable y^* is

$$y = j \leftrightarrow \gamma_{j-1} \leq y^* \leq \gamma_j, \ \ j = 1, 2, ..., NCAT$$

where it is assumed that $\gamma_0 = -\infty$ and $\gamma_{NCAT} = +\infty$. In addition, γ_1 is usually set to 0 to avoid identification problems. For the present example, NCAT = 4, and from the output we see that $\hat{\gamma}_2 = 1.7294$ and $\hat{\gamma}_3 = 2.9398$. These values are used in combination with the coefficients of TxDrug, SqrtWeek, and Tx*SWeek to calculate estimated outcomes for different groups of patients.

Estimated outcomes for groups: unit-specific probabilities

To evaluate the expected effect of the drug treatment TxDrug and the square of time of treatment, while allowing for the interaction between treatment and the square of time, we use the estimates for the alternative parameterization, setting $\gamma_1 = 0$:

$$\hat{\eta}_{ij} = 3.3666 - 0.0517 \ \text{TxDrug}_{ij} - 0.4591 \ \text{SqrtWeek}_{ij} - 0.6723 \ \text{Tx*SWeek}_{ij}.$$

For a typical patient from the control group (TxDrug = 0, SqrtWeek = Tx*SWeek = 0),

$$\hat{\eta}_{ij} = 3.3666.$$

The probabilities that a typical patient from the control group responded in a specific category at the start of the study are obtained by noting that $\gamma_0 = -\infty$, $\gamma_1 = 0$, $\hat{\gamma}_2 = 1.7294$, and $\hat{\gamma}_3 = 2.9398$.

Let $p = \Phi(a) = P(Z \le a)$ where Z has a standard normal distribution. To compute p for a given value of a, use is made of a probability calculator. These calculators can, for example, be found on the internet.

Since $\hat{\eta}_{ij} = 3.3666$, it follows that

$$\Phi(\gamma_0 - \eta_{ij}) = \Phi(-\infty) = 0$$
$$\Phi(\gamma_1 - \eta_{ij}) = \Phi(-3.3666) = 0.0004$$
$$\Phi(\gamma_2 - \eta_{ij}) = \Phi(-1.6372) = 0.0508$$
$$\Phi(\gamma_3 - \eta_{ij}) = \Phi(-0.4267) = 0.3348.$$

Therefore

$$P(\text{Imps79O} = 1) = \Phi(\gamma_1 - \eta_{ij}) - \Phi(\gamma_0 - \eta_{ij})$$
$$= 0.0004$$
$$P(\text{Imps79O} = 2) = \Phi(\gamma_2 - \eta_{ij}) - \Phi(\gamma_1 - \eta_{ij})$$
$$= 0.0508 - 0.0004$$
$$= 0.0504$$
$$P(\text{Imps79O} = 3) = \Phi(\gamma_3 - \eta_{ij}) - \Phi(\gamma_2 - \eta_{ij})$$
$$= 0.3348 - 0.0508$$
$$= 0.2840$$

TWO-LEVEL MODELS FOR ORDINAL OUTCOMES

and

$$P(\text{Imps79O} = 4) = 1 - P(\text{Imps79O} = 3) - P(\text{Imps79O} = 2) - P(\text{Imps79O} = 1)$$
$$= 1 - 0.2841 - 0.0504 - 0.0004$$
$$= 0.6653.$$

for the four categories of Imps79O. We can compare these predicted probabilities with the observed data for the control group as shown in Table 3.3 below. The ratio of responses in the categories of Imps79O when SqrtWeek $= 0$ are 0.000, 0.102, 0.336 and 0.561 respectively. For example, if Imps79O $= 4$, the ratio is $\dfrac{60}{107}$.

Table 3.3: Crosstabulation of Imps79O and SqrtWeek for control group

		SqrtWeek							Total
		.0000	1.0000	1.4142	1.7321	2.0000	2.2361	2.4495	
Imps79O	1	0	2	0	3	0	0	8	13
	2	11	25	0	24	0	1	24	85
	3	36	26	1	26	1	0	16	106
	4	60	52	4	34	1	1	22	174
Total		107	105	5	87	2	2	70	378

At the end of the study, the corresponding ratios for the observed data were 0.00, 0.10, 0.34, and 0.56 respectively. Here, the variable SqrtWeek $= 2.4495$ and thus

$$\hat{\eta}_{ij} = 3.3666 - 0.4591 \, \text{SqrtWeek}_{ij}$$
$$= 3.3666 - (0.4591)(2.4495)$$
$$= 2.2420.$$

The probabilities that a typical patient from the control group responded in a specific category at the conclusion of the study period are obtained as

$$\Phi(\gamma_1 - \eta_{ij}) = \Phi(-2.2420) = 0.0125$$
$$\Phi(\gamma_2 - \eta_{ij}) = \Phi(-0.5126) = 0.3041$$
$$\Phi(\gamma_3 - \eta_{ij}) = \Phi(0.6979) = 0.7574.$$

Therefore

$$P(\text{Imps79O} = 1) = \Phi(\gamma_1 - \eta_{ij}) - \Phi(\gamma_0 - \eta_{ij})$$
$$= 0.0125$$
$$P(\text{Imps79O} = 2) = \Phi(\gamma_2 - \eta_{ij}) - \Phi(\gamma_1 - \eta_{ij})$$
$$= 0.3041 - 0.0125$$
$$= 0.2916$$
$$P(\text{Imps79O} = 3) = \Phi(\gamma_3 - \eta_{ij}) - \Phi(\gamma_2 - \eta_{ij})$$
$$= 0.7574 - 0.3041$$
$$= 0.4533$$

and

$$P(\text{Imps79O} = 4) = 1 - P(\text{Imps79O} = 3) - P(\text{Imps79O} = 2) - P(\text{Imps79O} = 1)$$
$$= 1 - 0.7574$$
$$= 0.2426.$$

The estimated probabilities for a typical patient from the treatment group at the start of the study (TxDrug = 1, SqrtWeek = Tx*SWeek = 0) can be calculated in a similar fashion, using

$$\hat{\eta}_{ij} = 3.3666 - 0.0517$$
$$= 3.3149.$$

The probabilities that a typical patient from the treatment group responded in a specific category at the start of the study are obtained as

$$\Phi(\gamma_1 - \eta_{ij}) = \Phi(-3.3149) = 0.0005$$
$$\Phi(\gamma_2 - \eta_{ij}) = \Phi(-1.5855) = 0.0564$$
$$\Phi(\gamma_3 - \eta_{ij}) = \Phi(-0.3750) = 0.3538.$$

Therefore

$$P(\text{Imps79O} = 1) = \Phi(\gamma_1 - \eta_{ij}) - \Phi(\gamma_0 - \eta_{ij})$$
$$= 0.0005$$
$$P(\text{Imps79O} = 2) = \Phi(\gamma_2 - \eta_{ij}) - \Phi(\gamma_1 - \eta_{ij})$$
$$= 0.0564 - 0.0005$$
$$= 0.0559$$
$$P(\text{Imps79O} = 3) = \Phi(\gamma_3 - \eta_{ij}) - \Phi(\gamma_2 - \eta_{ij})$$
$$= 0.3538 - 0.0564$$
$$= 0.2974$$

and

$$P(\text{Imps79O} = 4) = 1 - P(\text{Imps79O} = 3) - P(\text{Imps79O} = 2) - P(\text{Imps79O} = 1)$$
$$= 1 - 0.3538$$
$$= 0.6462$$

for the four groups defined by the categories of Imps79O.

At the end of the study, the corresponding ratios for the observed data were 0.35, 0.45, 0.12, and 0.08 respectively. Here, the variable SqrtWeek = 2.4495 and thus

$$\hat{\eta}_{ij} = 3.3666 - 0.0517 \; \text{TxDrug}_{ij} - 0.4591 \; \text{SqrtWeek}_{ij} - 0.6723 \; \text{Tx*SWeek}_{ij}$$
$$= 3.3666 - 0.0517 - (0.4591)(2.4495) - (0.6723)(2.4495)$$
$$= 0.5435.$$

The probabilities that a typical patient from the treatment group responded in a specific category at the conclusion of the study period are obtained as

$$\Phi(\gamma_1 - \eta_{ij}) = \Phi(-0.5435) = 0.2934$$
$$\Phi(\gamma_2 - \eta_{ij}) = \Phi(1.1859) = 0.8822$$
$$\Phi(\gamma_3 - \eta_{ij}) = \Phi(2.3964) = 0.9917.$$

Therefore

$$P(\text{Imps79O} = 1) = \Phi(\gamma_1 - \eta_{ij}) - \Phi(\gamma_0 - \eta_{ij})$$
$$= 0.2934$$
$$P(\text{Imps79O} = 2) = \Phi(\gamma_2 - \eta_{ij}) - \Phi(\gamma_1 - \eta_{ij})$$
$$= 0.8822 - 0.2934$$
$$= 0.5888$$
$$P(\text{Imps79O} = 3) = \Phi(\gamma_3 - \eta_{ij}) - \Phi(\gamma_2 - \eta_{ij})$$
$$= 0.9917 - 0.8822$$
$$= 0.1095$$

and

$$P(\text{Imps79O} = 4) = 1 - P(\text{Imps79O} = 3) - P(\text{Imps79O} = 2) - P(\text{Imps79O} = 1)$$
$$= 1 - 0.9917$$
$$= 0.0083.$$

TWO-LEVEL MODELS FOR ORDINAL OUTCOMES

Estimated outcomes for different groups: population-average results

In this example, we defined the latent response variable model as

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i + e_{ij}, \quad j = 1, 2, ..., n_i,$$

where $\mathbf{z}_{ij}$ denotes a design vector for the random effects $\mathbf{u}_i$ and $\mathbf{x}_{ij}$ the design vector for the predictors in the fixed part of the model with corresponding vector $\boldsymbol{\beta}$ of regression parameters. Denote the covariance matrix of $\mathbf{u}_i$ by $\boldsymbol{\Phi}_{(2)}$ and the variance of e_{ij} by σ^2.

For a probit link function $\sigma^2 = 1$, and for a logistic link function $\sigma^2 = \pi^2/3$. Under the assumption that $\boldsymbol{\beta}_i$ and e_{ij} are independently distributed, it follows that

$$\sigma^2_{y_{ij}} = \mathbf{x}_{ij}'\boldsymbol{\Phi}_{(2)}\mathbf{x}_{ij} + \sigma^2.$$

Let

$$d_{ij} = \frac{\sigma^2_{y_{ij}}}{\sigma^2}.$$

The quantity d_{ij} is called the design effect. The estimated population-average probabilities (Hedeker & Gibbons, 2006) are obtained in a similar fashion as the unit-specific probabilities, but with replacing $\hat{\gamma}_k - \hat{\eta}_{ij}$ by $\left(\hat{\gamma}_k - \hat{\eta}_{ij}\right)/\sqrt{d_{ij}}$.

To illustrate, we calculate the estimated population-average probabilities for the control group associated at the end of the study (drug = 0, week = 6).

From the output, we have $\text{var}(u_{i1}) = 1.22776$, where u_{i1} denotes the random intercept coefficient. In this case, $\mathbf{x}'_{ik} = \mathbf{1}$ and hence, with $\sigma^2 = 1$ for the probit link,

$$\sigma^2_{y_{ik}} = 1 \times 1.2278 \times 1 + 1 = 2.2278.$$

Therefore

$$d_{ij} = \frac{2.2278}{1} = 2.2278.$$

Using the unit-specific values for $\hat{\gamma}_j - \hat{\eta}_{ij}$ obtained previously, it follows that

$$
\begin{aligned}
P(\text{Imps79O} = 1) &= \Phi(-2.2420 / \sqrt{2.2278}) \\
&= 0.0665 \\
P(\text{Imps79O} = 2) &= \Phi(-0.5126 / \sqrt{2.2278}) - \Phi(-1.5021) \\
&= 0.3656 - 0.0665 \\
&= 0.2991 \\
P(\text{Imps79O} = 3) &= \Phi(0.6979 / \sqrt{2.2278}) - \Phi(-0.3435) \\
&= 0.6799 - 0.3656 \\
&= 0.3143
\end{aligned}
$$

and

$$
\begin{aligned}
P(\text{Imps79O} = 4) &= 1 - P(0.0665 + 0.2991 + 0.3143) \\
&= 0.3201.
\end{aligned}
$$

The estimated population-average probabilities for the control group are summarized in Table 3.4 below.

TWO-LEVEL MODELS FOR ORDINAL OUTCOMES

Table 3.4: Estimated unit-specific probabilities for control group

		SqrtWeek						Total
		Start of study			**End of study**			
		Obs. Freq	**Ratio**	**Pred. Prob**	**Obs. Freq**	**Ratio**	**Pred. Prob**	
Imps79O	1	0	0	0.01	8	0.11	0.07	8
	2	11	0.10	0.12	24	0.34	0.30	35
	3	36	0.34	0.25	16	0.23	0.31	52
	4	60	0.56	0.61	22	0.31	0.32	82
Total		107			70			177

A comparison of these probabilities with the observed ratios given in Table 3.4 for the control group at the end of the study indicate that the population-average results are closer to the observed ratios than is the case for the unit-specific results. Results for the treatment group are obtained in a similar way and are summarized in Table 3.5 below.

Table 3.5: Estimated population-average probabilities for treatment group

		SqrtWeek						Total
		Start of study			End of study			
		Obs. Freq	Ratio	Pred. Prob	Obs. Freq	Ratio	Pred. Prob	
Imps79O	1	1	0.003	0.01	93	0.35	0.36	94
	2	43	0.13	0.13	118	0.45	0.43	161
	3	86	0.26	0.26	33	0.12	0.16	119
	4	197	0.60	0.60	21	0.08	0.05	218
Total		327			287			592

It is interesting to note that the observed frequencies at the start and end of the study are 107 and 70 for the control group, and 327 and 287 for the experimental group. The larger number of observations available in the latter case may partially explain why the predicted probabilities are closer to the observed probabilities for the experimental group when compared to those for the control group. The results clearly demonstrate the effectiveness of a seven week treatment, and the significant drug by time interaction effect: at the end of the study, the estimated population-average proportion of patients in categories 1 or 2 (normal or borderline) are 0.37 for the control group and 0.79 for the treatment group. The corresponding estimated population-average probabilities at the start of the study are 0.13 and 0.14 for the control and experimental groups respectively.

Finally, the estimated intracluster correlation (which in this case is the intra*patient* correlation) for this analysis is given, along with the variation in intercepts over patients.

```
P schizx2.out                                                   _ □ X

    Calculation of the intracluster correlation
    -------------------------------------------
    residual variance = 1 (assumed)
    cluster  variance =    1.2278

    intracluster correlation =    1.2278 / (    1.2278 + 1.000) = 0.551
```
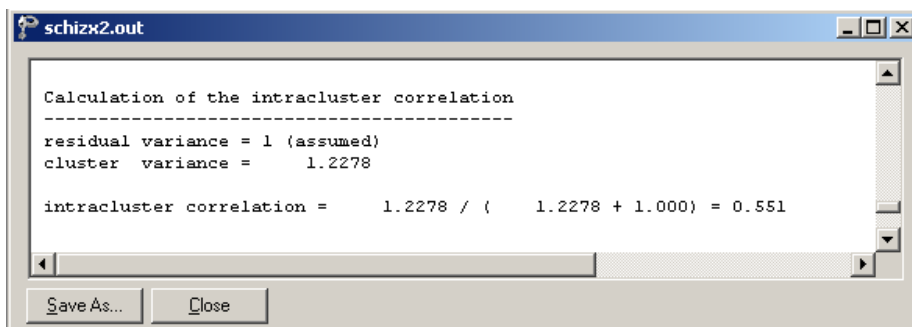
3.6.4 Example: Probit link function with random intercept and slope

To fit a model allowing patients to vary in terms of both their intercept and their trend across time (random intercepts and slopes), the model in the previous example is extended to include both intercept and slope (SqrtWeek) effects.

Setting up the analysis

The model is fitted to the data in **Examples\Primer\Ordinal\schizx1.ss3**, and the model specification file **schizx2.mum** used previously is modified to include a random slope. The first step is to open the **ss3** file, and then the model specification file **schizx2.mum**. This is accomplished by using the **File**, **Open** option to activate the display of an **Open** dialog box, and browsing for the file **schizx1.ss3** in the **Examples\Primer\Ordinal** folder. Next, use the **File**, **Open Existing Model Setup** option to activate an **Open** dialog box and browse for the file **schizx2.mum** in the same folder.

The **Model Setup** window, with entries corresponding to the contents of the model specification file **schizx2.mum** is displayed. After adjusting the title in the **Title 2** text box of the **Configuration** screen to reflect the intended changes to the model, proceed to the **Variables** screen by clicking on this tab.

Model Setup: schizx3.mum

Configuration | Variables | Starting Values | Patterns | Advanced | Linear Transforms

Title 1: NIMH Schiz Data - 7 timepoints

Title 2: IMPS79 (ordinal) across SQRT week - 2 random effects

Dependent Variable Type: ordered   Level-2 IDs: Patient

Dependent Variable: Imps79O   Level-3 IDs:

Categories:

	Value
1	1
2	2
3	3
4	4

Write Bayes Estimates: means & (co)variances

Convergence Criterion: 0.0001

Number of Iterations: 100

Missing Values Present: true   Perform Crosstabulation: yes

Missing Value for the Dependent Var: -9.0   Crosstab Variable: SqrtWeek

Global Missing Value: -9.0   Output Type: standard

Enter the main title to be displayed in the output.
The maximum length is 60 characters.

The only change needed on the **Variables** screen is to add the variable SqrtWeek to the **L-2 Random Effects** grid. This is done by checking the check box next to this variable's name under the **2** column of the **Available** grid. Doing so leads to the display of the variable name SqrtWeek in the **L-2 Random Effects** grid as shown below.
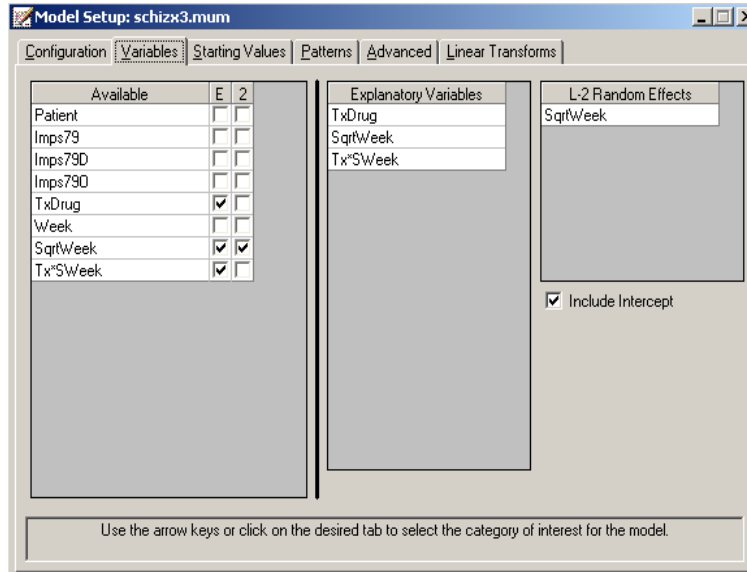
TWO-LEVEL MODELS FOR ORDINAL OUTCOMES

Before running the analysis, the model specifications have to be saved. Select the **File**, **Save** option, and provide a new name, for example **schizx3.mum**, for the model specification file. Run the analysis by selecting the **Run** option from the **Analysis** menu.

Discussion of results

Portions of the output file **schizx3.out** are shown below. The estimates are shown in the column with heading Estimate, and correspond to the coefficients β_0, β_1, β_2, and β_3 in the model specification. Results for the fixed effects in this model are similar to those obtained for the random intercepts model. There is significant random variation in the random intercepts and slopes of the patients. For this model, we have $\hat{\gamma}_2 = 2.1791$ and $\hat{\gamma}_3 = 3.6435$, both somewhat higher than those for the previous model ($\hat{\gamma}_2 = 1.7297$ and $\hat{\gamma}_3 = 2.9404$.)

```
┌─────────────────────────────────────────────────────────────────────────┐
│ ⓟ SuperMix - [schizx3.out]                                    _ □ ×       │
│ ⓟ File  Analysis  Window  Help                                 _ ₆ ×       │
│                                                                      ▲     │
│      o===========================================o                         │
│      | Optimization Method: Adaptive Quadrature |                          │
│      o===========================================o                         │
│                                                                            │
│      Number of quadrature points =          20                             │
│      Number of free parameters =             9                             │
│      Number of iterations used =             6                             │
│                                                                            │
│      -2lnL (deviance statistic) =      3327.04804                          │
│      Akaike Information Criterion      3345.04804                          │
│      Schwarz Criterion                 3393.46473                          │
│                                                                            │
│                                                                            │
│                   Estimated regression weights                             │
│                                                                            │
│                                        Standard                            │
│      Parameter            Estimate     Error      z Value     P Value      │
│      ---------            --------     --------   -------     -------      │
│      Threshold1            -4.1012      0.2543    -16.1265     0.0000       │
│      Threshold2            -1.9221      0.2127     -9.0374     0.0000       │
│      Threshold3            -0.4577      0.1964     -2.3310     0.0198       │
│      TxDrug                 0.0324      0.2190      0.1480     0.8823       │
│      SqrtWeek              -0.5059      0.1214     -4.1682     0.0000       │
│      Tx*SWeek              -0.9426      0.1400     -6.7321     0.0000   ▼    │
│   ◄                                                                  ►     │
│  ┌────────┐  ┌────────┐                                                    │
│  │Save As...│  │ Close  │                                                  │
│  └────────┘  └────────┘                                                    │
└─────────────────────────────────────────────────────────────────────────┘
```

To compare this model to the random intercepts model previously fitted, the likelihood-ratio χ^2 test can be used. Based on this test, there is clear evidence of significant variation in the linear time trends:
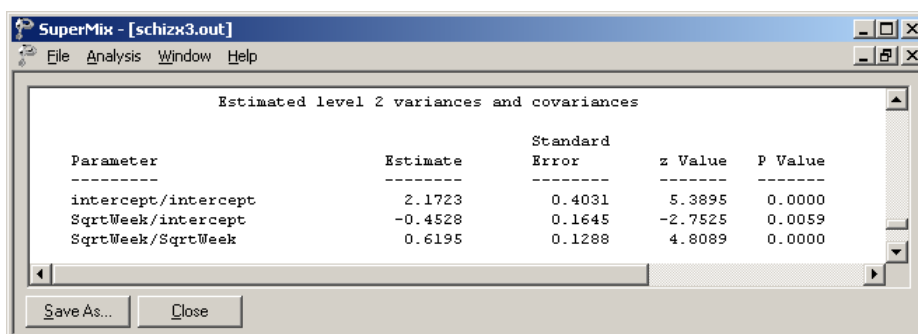
$$\chi^2 = 3399.47482 - 3327.04804 = 72.42678,$$

(2 degrees of freedom and $p < 0.001$) over and above the individual intercept variation. Significant negative association between the intercept and linear time terms is indicated, suggesting that those patients with the highest initial severity show the greatest improvement across time (*e.g.*, largest negative time trends).

Note that variances are, by definition, non-negative. It was shown by Miller (1977) and Self & Liang (1987) that, when testing the hypothesis that the variance of a random intercept or slope is zero, the alternative hypothesis is one-sided. This implies that an approximate *p*-value is obtained as half the *p*-value obtained from a

χ^2 statistic based on the difference in deviances. Snijders & Bosker (1999), compares a random intercept and slope model with a random intercept model. For their example, $\chi^2 = 8.5$ at 2 degrees of freedom with tail probability of $p < 0.02$, so halving the p-value yields a significant outcome with $p < 0.01$. Generally speaking, the likelihood-ratio test applied to random coefficients without the adjustment of p-value is conservative.

Finally, the estimated variance and covariance components, which are all significant, are shown.

```
SuperMix - [schizx3.out]
File   Analysis   Window   Help

            Estimated level 2 variances and covariances

                                        Standard
         Parameter              Estimate    Error     z Value   P Value
         ---------              --------  --------   -------   -------
         intercept/intercept      2.1723    0.4031    5.3895    0.0000
         SqrtWeek/intercept      -0.4528    0.1645   -2.7525    0.0059
         SqrtWeek/SqrtWeek        0.6195    0.1288    4.8089    0.0000

    Save As...      Close
```

3.7 Two-level models for nominal outcomes

3.7.1 The data

The McKinney Homeless Research Project study (Hough, *et. al.*, 1997; Hurlburt, *et. al.* 1996) was designed to evaluate the effectiveness of using Section 8 certificates to provide independent housing to the severely mentally ill homeless. These housing certificates, which require clients to pay 30% of their income toward rent, are meant to enable low-income subjects to choose and obtain independent housing in the community. Three hundred sixty-two clients took part in this longitudinal study employing a randomized factorial design. Clients were randomly assigned to one of two types of case management (comprehensive vs. traditional) and to one of two levels of access to independent housing (using Section 8 certificates). The project was restricted to clients diagnosed with a severe and persistent mental illness who were either homeless or at high risk of becoming homeless at the start of the study. Individuals' housing status was classified at baseline and at 6, 12, and 24 month follow-ups. Here, we focus on examining the effect of access to Section 8 certificates on housing outcomes across time. At each time point, subjects' housing status was classified as either streets/shelters, community housing, or independent housing; a partial list of these data is given below in the form of a SuperMix spreadsheet file, named **sdhouse.ss3**.

	(A)_ID	(B)_HOUSI	(C)_SECTI	(D)_TIME1	(E)_TIME2	(F)_TIME3	(G)_SECT8	(H)_SECT8	(I)_SECT8T	(J)_NOSEC	(K)_TIME	(L)_SEC8TIME
1	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1.00	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00
3	1.00	2.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	2.00	2.00
4	1.00	2.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	3.00	3.00
5	2.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	2.00	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00
7	2.00	2.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	2.00	2.00
8	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	3.00	3.00
9	3.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	3.00	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00

The variables of interest are:

- ID is the subject ID (362 subjects in total).
- HOUSING represents the housing status at the time of interview: 0 = street, 1 = community, 2 = independent, and 999 = unknown (missing).
- SECTION8 indicates the Section 8 group, with 1 representing those using Section 8 certifications, and 0 those without.
- TIME1 to TIME3 are three dummy variables for time effects, and denote whether a classification was at baseline, or at the 6, 12 or 24 month follow-up. If at the 6 months follow-up, TIME1 = 1 and TIME2 = TIME3 = 0; if at 12 months, TIME2 = 1 while TIME1 = TIME3 = 0; and at the 24 month follow-up TIME3 = 1 and TIME1 = TIME2 = 0. With this coding scheme, the baseline serves as the reference group of classification. The coding structure is shown in Table 3.6 below.
- Three Section 8 by time interaction terms follow: SECT8T1 is the product of SECTION8 and TIME1, and SECT8T2 and SECT8T3 are the products of SECTION8 and TIME2 and TIME3 respectively.
- NOSECT8 indicates the non-Section 8 group, with 0 = no, and 1 = yes.
- TIME represents the linear time contrast. At baseline, TIME = 0, at 6 months, TIME = 1, at 12 months TIME = 2, and at 24 months TIME = 3.
- SEC8TIME is the product of SECTION8 and TIME.

Table 3.6: Coding of the dummy variables TIME1, TIME2, and TIME3

	TIME1	TIME2	TIME3	TIME
baseline	0	0	0	0
6 months	1	0	0	1
12 months	0	1	0	2
24 months	0	0	1	3

Values of 999 represent missing value codes for the housing status variable. Thus, some subjects are measured at all four time points and others at fewer time points. Data from these time points with missing values are not used in the analysis; however, data are used from other time points where there are no missing data. Thus, for inclusion into the analysis, a subject's data (both the dependent variable and all explanatory variables used in a particular analysis) at a specific time point must be complete. The number of repeated observations per subject depends on the number of time points for which there are non-missing data for that subject.

The observed sample sizes and response proportions by group are given in Table 3.7 below. These observed proportions indicate a general decrease in street living and an increase in independent living across time for both groups. The increase in independent housing, however, appears to occur sooner for the section 8 group relative to the control group. Regarding community living, across time there is an increase for the control group and a decrease for the section 8 group.

Regarding missing data, further inspection of Table 3.7 indicates that there is some attrition across time; attrition rates of 19.4% and 12.7% are observed at the final time point for the control and section 8 groups, respectively. In addition, one subject provided no housing data at any of the four measurement time points. Since estimation of model parameters is based on a full-likelihood approach, missing data are assumed to be "ignorable" conditional on both the explanatory variables and observed nominal responses (Laird, 1988). In longitudinal studies, ignorable nonresponse falls under Rubin's (1976) "missing at random" (MAR) assumption, in which the missingness depends only on observed data. In what follows, since the focus is on describing use of the nominal model in SuperMix, we will make the MAR assumption. A further approach, however, that does not rely on the MAR assumption (*e.g.*, a mixed-effects pattern-mixture model as described in Hedeker & Gibbons (1997)) could be used.

Table 3.7: Observed sample sizes and response proportions by group

Time point					
Group	**Status**	**Baseline**	**6 months**	**12 months**	**24 months**
Control	Street	0.555	0.186	0.089	0.124
	Community	0.339	0.578	0.582	0.455
	Independent	0.106	0.236	0.329	0.421
	n	180	161	146	145
Section 8	Street	0.442	0.093	0.121	0.120
	Community	0.414	0.280	0.146	0.228
	Independent	0.144	0.627	0.732	0.652
	n	181	161	157	158

In preparation for the subsequent analyses, the marginal response proportions can be converted to the two logits of the nominal regression model (*i.e.*, $\log[p(C)/p(S)]$ and $\log[p(I)/p(S)]$, where S = street, C = community, and I = independent housing).[1] These logits are given in Table 3.8.

The logits clearly show the increase in community and independent housing, relative to street housing, at all follow-up time points (6, 12, and 24 months). In terms of group-related differences, these appear most pronounced at 6 months for independent housing and 12 months for community housing. While examination of these logits is instructive, the subsequent analyses will more rigorously assess the degree to which these logits vary by time and group.

---

[1]Again, street housing is treated as the reference category because its code (0) is listed as the first response category.

Table 3.8: Logits across time by group

Group	Status	Baseline	6-months	12-months	24-months
	Time point				
Control	Community vs. street	−.49	1.13	1.88	.130
	Independent vs. street	−1.66	.24	1.31	1.22
Section 8	Community vs. street	−.07	1.10	.19	.64
	Independent vs. street	−1.12	1.91	1.0	1.69
Difference	Community vs. street	.42	−.03	−1.69	−.66
	Independent vs. street	.54	1.67	.49	.47

In this example, one random subject effect (*i.e.*, a random subject intercept) is assumed and the repeated housing status classifications is modeled in terms of the dummy-coded time effects (6, 12, and 24 month follow-ups compared to baseline), a group effect (section 8 versus control), and group by time interaction terms.

3.7.2 The model

In the nominal case, we need to consider the values corresponding to the unordered multiple categories of the response variable. We thus assume that the $K+1$ response categories are coded as $0,1,2,...,K$. Let $P(y_{ij} = k \mid \mathbf{u}_{ik}, \boldsymbol{\beta})$ denote the probability that a response occurs in category k, conditional on the parameters $\boldsymbol{\beta}$ and $\mathbf{u}_{ik}$, where y_{ij} denotes the value of the nominal variable associated with level-2 unit i, $i = 1, 2,..., N$ and level-1 unit j, $j = 1, 2,..., n_i$.

　　　　　　　　　　TWO-LEVEL MODELS FOR NOMINAL OUTCOMES

Then

$$P_{ijk} = P(y_{ij} = k \mid \boldsymbol{\beta}_k, \mathbf{u}_{ik}) = \frac{\exp(\eta_{ijk})}{1 + \sum_{h=1}^{K} \exp(\eta_{ijh})} \quad \text{for} \quad k = 1, 2, ..., K$$

$$P_{ij0} = P(y_{ij} = 0 \mid \boldsymbol{\beta}_0, \mathbf{u}_{ik}) = \frac{1}{1 + \sum_{h=1}^{K} \exp(\eta_{ijh})}$$

where $\eta_{ijk} = \mathbf{x}_{ij}' \boldsymbol{\beta}_k + \mathbf{z}_{ij}' \mathbf{u}_{ik}$. Here, $\mathbf{x}_{ij}$ is the $p \times 1$ explanatory variable vector and $\mathbf{z}_{ij}$ is the design vector for the r random effects, both vectors being for the j-th level-1 unit nested within level-2 unit i. Correspondingly, $\boldsymbol{\beta}_k$ is a $p \times 1$ vector of unknown fixed regression parameters, and $\mathbf{u}_{ik}$ is a $r \times 1$ vector of unknown random effects for the level-2 unit i. The distribution of the random effects is assumed to be multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(2)}$. Notice that the regression coefficient vectors $\mathbf{u}_i$ and $\boldsymbol{\beta}$ carry the k subscript. Thus, for each of the p explanatory variables and r random effects, there will be K parameters to be estimated. Additionally, the random effect variance-covariance matrix $\boldsymbol{\Phi}_{(2)}$ is allowed to vary with k.

In the current example, the outcome variable HOUSING is coded 0, 1, and 2. Therefore

$$P\left(\text{HOUSING}_{ij} = k \mid \boldsymbol{\beta}_k, \mathbf{u}_{ik}\right) = \frac{\exp(\eta_{ijk})}{1 + \sum_{h=1}^{2} \exp(\eta_{ijh})}, \quad k = 1, 2$$

where for $k = 1$ (community housing)

$$\eta_{ij1} = \beta_{10} + \beta_{11}(\text{SECTION8})_{ij} + \beta_{12}(\text{TIME1})_{ij} + \beta_{13}(\text{TIME2})_{ij} + \beta_{14}(\text{TIME3})_{ij} +$$
$$\beta_{15}(\text{SECT8T1})_{ij} + \beta_{16}(\text{SECT8T2})_{ij} + \beta_{17}(\text{SECT8T3})_{ij} + u_{10i}.$$

For $k = 2$ (independent housing)

$$\eta_{ij2} = \beta_{20} + \beta_{21}(\text{SECTION8})_{ij} + \beta_{22}(\text{TIME1})_{ij} + \beta_{23}(\text{TIME2})_{ij} + \beta_{24}(\text{TIME3})_{ij} +$$
$$\beta_{25}(\text{SECT8T1})_{ij} + \beta_{26}(\text{SECT8T2})_{ij} + \beta_{27}(\text{SECT8T3})_{ij} + u_{20i}.$$
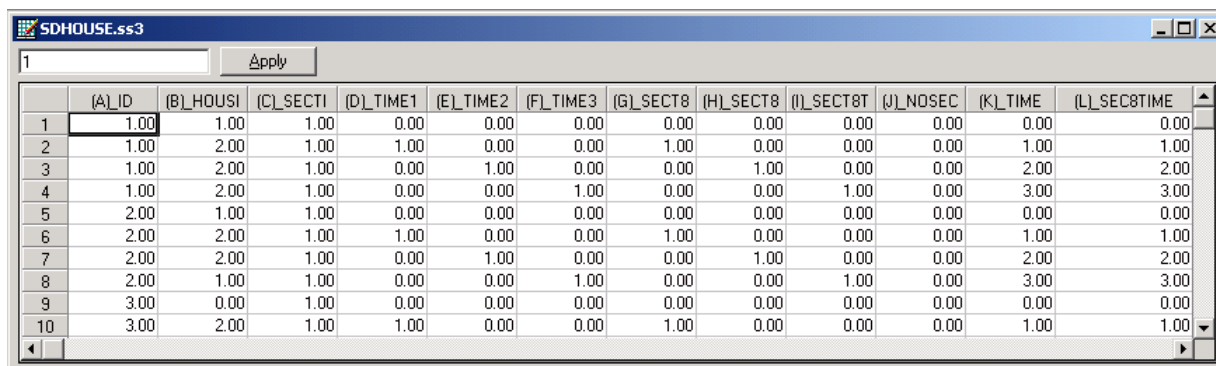
It is assumed that u_{k0i}, $i = 1, 2, ..., N$, are i.i.d. normal $(\mathbf{0}, \mathbf{\Phi}_{kk})$, $k = 1, 2$.

3.7.3 Example: Random intercept model with dummy-coded time effects

Importing the data

The model is fitted to the data in **SDHOUSE.ss3**. The first step is to create the **ss3** file shown above from an Excel spreadsheet named **SDHOUSE.xls**. This is accomplished as follows:

o Use the **File**, **Import Data File** option to activate the display of an **Open** dialog box.
o Browse for the file **SDHOUSE.xls** in the **Examples\Primer\Nominal** folder.
o Select the file and click the **Open** button to return to the main SuperMix window, where the contents of the Excel spreadsheet are displayed as the SuperMix system file with default name **SDHOUSE.ss3**.

	(A)_ID	(B)_HOUSI	(C)_SECTI	(D)_TIME1	(E)_TIME2	(F)_TIME3	(G)_SECT8	(H)_SECT8	(I)_SECT8T	(J)_NOSEC	(K)_TIME	(L)_SEC8TIME
1	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1.00	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00
3	1.00	2.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	2.00	2.00
4	1.00	2.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	3.00	3.00
5	2.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	2.00	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00
7	2.00	2.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	2.00	2.00
8	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	3.00	3.00
9	3.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	3.00	2.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00

Setting up the analysis

The next step is to describe the model to be fitted. We use the SuperMix interface to provide the model specifications. From the main menu bar, select the **File**, **New Model Setup** option.

In this example, only the **Configuration**, **Variables**, and **Advanced** tabs of the **Model Setup** window that appears will be used. By default, the **Configuration** screen is displayed first.

Start by providing a title for the analysis in the **Title 1** and **Title 2** text boxes. Next, select the outcome variable HOUSING from the **Dependent Variable** drop-down list box and indicate the type of outcome as **nominal** using the **Dependent Variable Type** drop-down list box. Once this selection is made, the **Categories** grid is displayed, with distinct values of the categories in the text boxes as shown below. The subject identification variable is used to define the hierarchical structure of the data, and is selected as the Level-2 ID from the **Level-2 IDs** drop-down list box. The bottom right portion of the screen indicates that a marginal crosstabulation table of the nominal outcome variable (*i.e.* housing status) by SECTION8 is requested. This table provides purely descriptive information and has no effect on the estimation of the model parameters.

TWO-LEVEL MODELS FOR NOMINAL OUTCOMES

Finally, we need to provide information on missing data in the **SDHOUSE.ss3** file. Some of the values of the outcome variable HOUSING are missing, and a missing value code of 999 is used to indicate this. Click on the **Missing Values Present** drop-down list, and select the **yes** option. Enter the code 999 in the **Missing Value for the Dependent Var** text box that appears. The crosstabulation table was obtained by selecting the **yes** option from the **Perform Crosstabulation** drop-down list box.

The **Variables** screen is used to specify the fixed and random effects to be included in the model. Start by selecting the explanatory (fixed) variables using the check boxes in the **Available** grid. The image below shows the completed selection of all the predictors. By default, the inclusion of both a fixed intercept coefficient and a random intercept at level-2 is assumed, as indicated by the checked boxes for **Include Intercept** in the **Explanatory Variables** and **L-2 Random effects** grids. As these selections correspond to the model we intend fitting to the data, no further changes are needed on this screen.
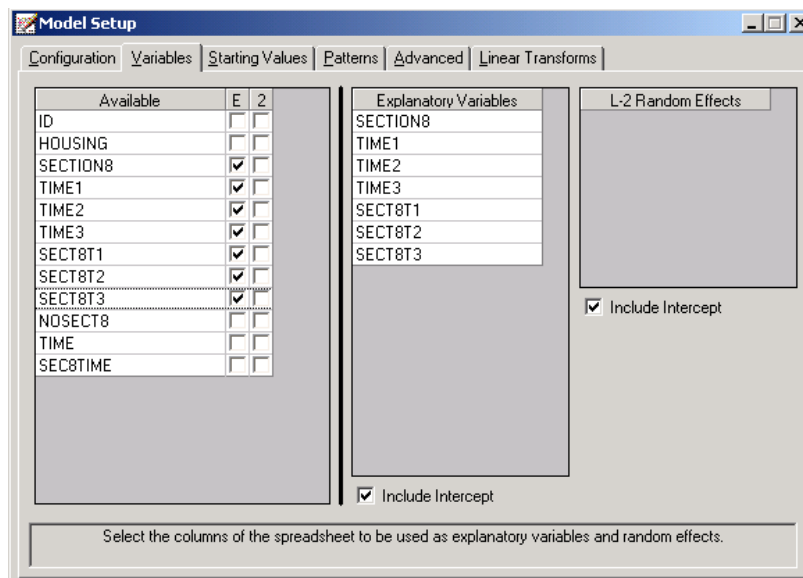
Click on the **Advanced** tab and request the use of 25 quadrature points for estimation using the **Number of Quadrature Points** text box. Also, select **non-adaptive quadrature** as **Optimization Method**. Increasing the number of points increases the accuracy of the integration, though minimal change is usually observed beyond 10 points or so. For models with only one random effect, increasing the number of points does not slow the solution down excessively. Thus, using 25 points, while perhaps not necessary, provides a safe choice.
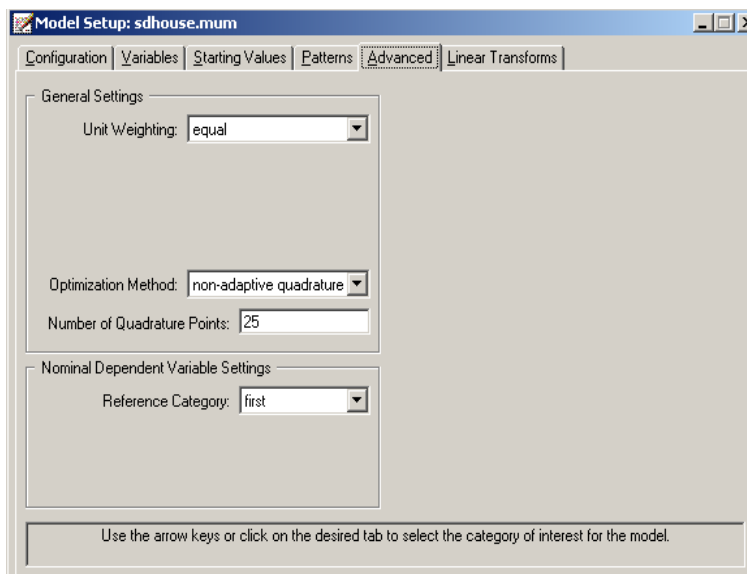
Before running the analysis, the model specifications have to be saved. Select the **File**, **Save** option, and provide a name for the model specification file, for example **sdhouse.mum**. Run the analysis by selecting the **Run** option from the **Analysis** menu.

Discussion of results

 Portions of the output file **sdhouse.out** are shown below. The first part of the output file gives a description of the model specifications. This is followed by a data
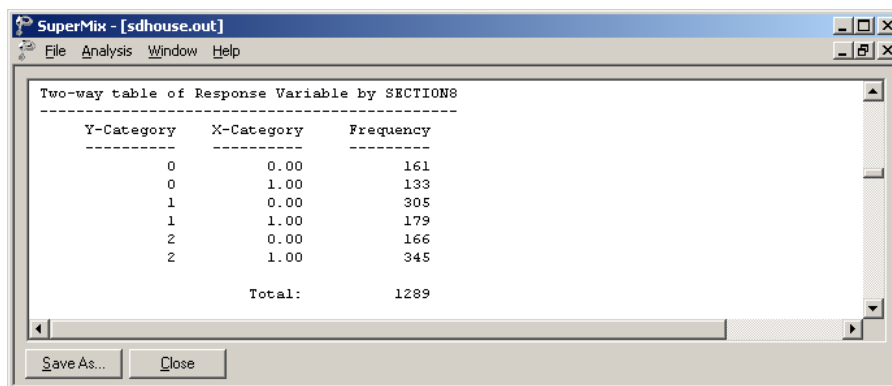
summary of the number of observations nested within each subject. The number of observations per subject (level-2 unit) ranges between 1 and 4.
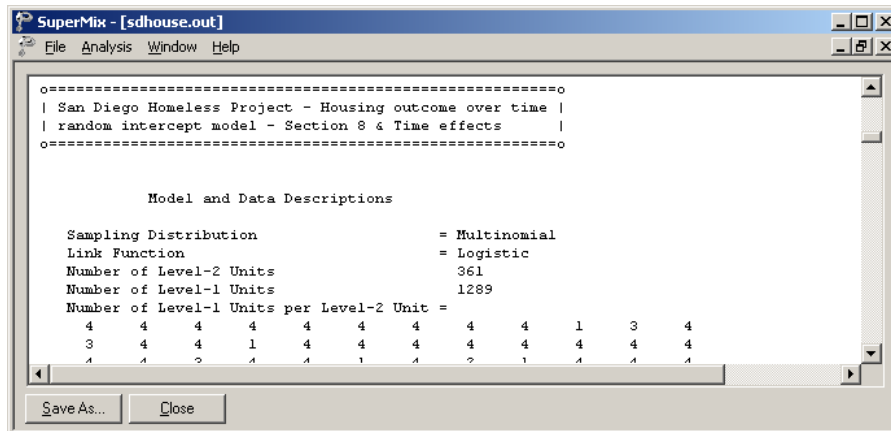
```
 SuperMix - [sdhouse.out]                                        _ □ ×
  File  Analysis  Window  Help                                  _ 🗗 ×

  o===========================================================o         ▲
  | San Diego Homeless Project - Housing outcome over time |
  | random intercept model - Section 8 & Time effects      |
  o===========================================================o


            Model and Data Descriptions

    Sampling Distribution              = Multinomial
    Link Function                      = Logistic
    Number of Level-2 Units              361
    Number of Level-1 Units              1289
    Number of Level-1 Units per Level-2 Unit =
        4    4    4    4    4    4    4    4    4    1    3    4
        3    4    4    1    4    4    4    4    4    4    4    4     ▼
    ◀                                                         ▶

  Save As...    Close
```

The data summary is followed by descriptive statistics for all the variables included in the model. We note that the most frequent response was in category 2, *i.e.* "Independent" on the nominal outcome variable HOUSING, while 23% indicated that they were living on the street (HOUSING = 0, the reference category).

```
 SuperMix - [sdhouse.out]                                        _ □ ×
  File  Analysis  Window  Help                                  _ 🗗 ×

    o===========================================================o       ▲
    | Descriptive statistics for all the variables in the model |
    o===========================================================o
                                                    Standard
    Variable        Minimum     Maximum     Mean    Deviation
    --------        -------     -------     ----    ---------
    HOUSING1        0.0000      1.0000      0.3755    0.4844
    HOUSING2        0.0000      1.0000      0.3964    0.4893
    HOUSING3        0.0000      1.0000      0.2281    0.4198
    intercept       1.0000      1.0000      1.0000    0.0000
    SECTION8        0.0000      1.0000      0.5097    0.5001
    TIME1           0.0000      1.0000      0.2498    0.4331
    TIME2           0.0000      1.0000      0.2351    0.4242
    TIME3           0.0000      1.0000      0.2351    0.4242
    SECT8T1         0.0000      1.0000      0.1249    0.3307
    SECT8T2         0.0000      1.0000      0.1218    0.3272
    SECT8T3         0.0000      1.0000      0.1226    0.3281        ▼
  ◀                                                         ▶

  Save As...    Close
```

The crosstabulation of the variable SECTION8 by the response variable HOUSING, requested on the **Variables** screen, is given next. Most of the classifications from subjects without Section 8 certificates indicated that the subject was living in community housing at the time of classification (SECTION8 = 0, HOUSING = 1). In the case of classifications from subjects with Section 8 certificates, most classifications showed the use of independent housing (SECTION8 = 1, HOUSING = 2).

```
SuperMix - [sdhouse.out]
 File  Analysis  Window  Help

Two-way table of Response Variable by SECTION8
-----------------------------------------------
    Y-Category    X-Category      Frequency
    ----------    ----------      ---------
         0           0.00            161
         0           1.00            133
         1           0.00            305
         1           1.00            179
         2           0.00            166
         2           1.00            345

              Total:               1289
```

Results for the model without any random effects are given next. The first eight values are those for the intercept, SECTION8, TIME1, …, SECT8T3 for response code 1 vs. code 0, the last eight are for the same predictors, but for response code 2 vs. response code 0.

TWO-LEVEL MODELS FOR NOMINAL OUTCOMES

```
 o==============================================o
 | Results for the model without any random effects |
 o==============================================o

                Goodness of fit statistics

 Statistic                        Value        DF       Ratio
 ---------                        -----        --       -----
 Likelihood Ratio Chi-square    2353.0456      1273     1.8484
 Pearson Chi-square             2578.0000      1273     2.0251


                Estimated regression weights

                                Standard
 Parameter          Estimate     Error     z Value   P Value
 ---------          --------    --------   -------   -------

 Response Code  1 vs Code    0
 ---------------------------
 intercept          -0.4943     0.1625    -3.0426    0.0023
 SECTION8            0.4298     0.2285     1.8805    0.0600
 TIME1               1.6257     0.2655     6.1236    0.0000
 TIME2               2.3720     0.3392     6.9922    0.0000
 TIME3               1.7936     0.3116     5.7559    0.0000
 SECT8T1            -0.4625     0.4304    -1.0748    0.2825
 SECT8T2            -2.1164     0.4869    -4.3471    0.0000
 SECT8T3            -1.0900     0.4509    -2.4171    0.0156

 Response Code  2 vs Code    0
 ---------------------------
 intercept          -1.6607     0.2503    -6.6359    0.0000
 SECTION8            0.5368     0.3370     1.5927    0.1112
 TIME1               1.8971     0.3497     5.4252    0.0000
 TIME2               2.9670     0.4005     7.4085    0.0000
 TIME3               2.8812     0.3669     7.8539    0.0000
 SECT8T1             1.1339     0.4998     2.2686    0.0233
 SECT8T2            -0.0426     0.5222    -0.0815    0.9350
 SECT8T3            -0.0670     0.4979    -0.1346    0.8929
```

The final results obtained with maximum marginal likelihood estimation are given next. Using 25 quadrature points per dimension, 5 iterations were required to obtain convergence. The log likelihood function value and deviance at convergence are included, and can be used to compare the current model with other models.

```
 o================================================o
 | Optimization Method: Non-adaptive Quadrature |
 o================================================o

 Number of quadrature points =          25
 Number of free parameters =            18
 Number of iterations used =             5

 -2lnL (deviance statistic) =      2218.72454
 Akaike Information Criterion      2254.72454
 Schwarz Criterion                 2347.63373


                    Estimated regression weights


                                    Standard
 Parameter             Estimate     Error      z Value    P Value
 ---------             --------     --------   -------    -------


 Response Code  1 vs Code   0
 ----------------------------
 intercept            -0.4519       0.1846     -2.4480     0.0144
 SECTION8              0.5210       0.2619      1.9897     0.0466
 TIME1                1.9416       0.3003      6.4658     0.0000
 TIME2                2.8199       0.3838      7.3477     0.0000
 TIME3                2.2596       0.3548      6.3678     0.0000
 SECT8T1             -0.1352       0.4617     -0.2929     0.7696
 SECT8T2             -1.9171       0.5166     -3.7113     0.0002
 SECT8T3             -0.9523       0.4808     -1.9806     0.0476

 Response Code  2 vs Code   0
 ----------------------------
 intercept            -2.6754       0.3782     -7.0742     0.0000
 SECTION8              0.7812       0.4809      1.6245     0.1043
 TIME1                2.6821       0.4360      6.1515     0.0000
 TIME2                4.0881       0.4974      8.2194     0.0000
 TIME3                4.0990       0.4698      8.7255     0.0000
 SECT8T1              2.0028       0.6192      3.2345     0.0012
 SECT8T2              0.5484       0.6507      0.8428     0.3994
 SECT8T3              0.3041       0.6196      0.4908     0.6236
```

In terms of significance of the fixed effects, the time effects are observed to be highly significant. With the inclusion of the Time by Section 8 interaction terms, the time effects reflect comparisons between time points for the control group (*i.e.*, SECTION8 coded as 0). Thus, subjects in the control group show increased use of both independent and community housing relative to street housing at all three follow-ups, as compared to baseline. Similarly, due to the inclusion of the interaction terms, the Section 8 effect is the group difference at baseline (*i.e.*, when all time effects are 0). Using a .05 cutoff, there is no statistical evidence of group differences at baseline. Turning to the interaction terms, these indicate how the two groups differ in terms of comparisons between time points. Compared to controls, the increase in community versus street housing is less pronounced for section 8

TWO-LEVEL MODELS FOR NOMINAL OUTCOMES

subjects at 12 months (the estimate for SECT8T2 equals -1.92 in terms of the logit), but not statistically different at 6 months (SECT8T1) and only marginally different at 24 months (SECT8T3). Conversely, as compared to controls, the increase in independent versus street housing (response code 2 vs. code 0) is more pronounced for section 8 subjects at 6 months (the estimate equals 2.00 in terms of the logit), but not statistically different at 12 or 24 months.

In terms of community versus street housing (*i.e.*, response code 1 versus 0), there is an increase across time for the control group relative to the Section 8 group. As the statistical test indicated, these groups differ most at 12 months. For the independent versus street housing comparison (*i.e.*, response code 2 versus 0) there is a beneficial effect of Section 8 certificates at 6 months. Thereafter, the non-significant interaction terms indicate that the control group catches up to some degree. Considering these results of the mixed-effects analysis, it is seen that both groups reduce the degree of street housing, but do so in somewhat different ways. The control group subjects are shifted more towards community housing, whereas Section 8 subjects are more quickly shifted towards independent housing.

This differential effect of Section 8 certificates over time is completely missed if one simply analyzes the outcome variable as a binary indicator of street versus non-street housing (*i.e.*, collapsing community and independent housing categories). In this case (not shown), none of the section 8 by time interaction terms are observed to be statistically significant. Thus, analysis of the three-category nominal outcome is important in uncovering the beneficial effect of Section 8 certificates.

Comparing the log-likelihood value from this analysis to one where there are no random effects clearly supports inclusion of the random subject effect (likelihood ratio $\chi_1^2 = 134.3$). Expressed as intraclass correlations, $r_1 = .19$ and $r_2 = .62$ for community versus street and independent versus street, respectively. Thus, the subject influence is much more pronounced in terms of distinguishing independent versus street living, relative to community versus street living. This is borne out by contrasting models with separate versus a common random-effect variance across the two category contrasts (not shown) which yields a highly significant likelihood

ratio $\chi_1^2 = 49.2$ favoring the model with separate variance terms. (Also see the note on the use of the likelihood-ratio test in Section 3.6.)

Estimated outcomes for groups: unit-specific probabilities

From the above output, it follows that for a typical person at 24 months from baseline (TIME1 = TIME2 = 0, TIME3 = 1) with a Section 8 certificate (SECTION8 = 1)

$$\overset{\wedge}{\eta}_{ij1} = -0.4519 + 0.5210 + 2.2596 - 0.9523$$
$$= 1.3764$$

$$\overset{\wedge}{\eta}_{ij2} = -2.6754 + 0.7812 + 4.0990 + 0.3041$$
$$= 2.5089$$

so that

$$P(\text{HOUSING} = \text{community} \mid \text{TIME3} = 1, \text{SECTION8} = 1) = \frac{e^{1.3764}}{1 + e^{1.3764} + e^{2.5089}}$$
$$= 0.2296$$

and

$$P(\text{HOUSING} = \text{independent} \mid \text{TIME3} = 1, \text{SECTION8} = 1) = \frac{12.2914}{1 + 3.9606 + 12.2914}$$
$$= 0.7125.$$

Therefore

$$P(\text{HOUSING} = \text{street} \mid \text{TIME3} = 1, \text{SECTION8} = 1) = 1 - (0.2296 + 0.7125)$$
$$= 0.0579.$$

TWO-LEVEL MODELS FOR NOMINAL OUTCOMES

The corresponding probabilities for a typical person without Section 8 certification are obtained using

$$\overset{\wedge}{\eta_{ij1}} = -0.4519 + 2.2596 = 1.8077$$

$$\overset{\wedge}{\eta_{ij2}} = -2.6754 + 4.0990 = 1.4236.$$

From these values, it follows that

$$P(\text{HOUSING} = \text{community} \mid \text{TIME3} = 1, \text{SECTION8} = 0) = \frac{6.0964}{1 + 6.0964 + 4.1520}$$
$$= 0.5420$$

$$P(\text{HOUSING} = \text{independent} \mid \text{TIME3} = 1, \text{SECTION8} = 0) = \frac{4.1520}{11.2484}$$
$$= 0.3691$$

$$P(\text{HOUSING} = \text{street} \mid \text{TIME3} = 1, \text{SECTION8} = 0) = 1 - (0.5420 + 0.3691)$$
$$= 0.0889.$$

There is therefore a higher expected proportion of subjects without Section 8 certificates who will be classified as "street/shelters" 24 months from baseline, than is the case for those with section 8 certificates.

Table 3.9 below is a summary of the predicted probabilities, calculated as described above. We conclude that the highest proportion of subjects without Section 8 certificates make use of community housing, whereas the highest proportion of subjects with certificates make use of independent housing.

Table 3.9: predicted probabilities

Section 8 certificate	Time from baseline	P(street)	P(independent)	P(community)
no	6	0.1552	0.1563	0.6885
no	12	0.0634	0.2602	0.6764
no	24	0.0889	0.3691	0.5420
yes	6	0.0420	0.6841	0.2739
yes	12	0.0522	0.8099	0.1379
yes	24	0.0579	0.7125	0.2296

The intracluster correlations for this analysis are calculated as shown below. The residual variance is assumed as $\pi^2/3$, thus the intracluster correlation for community versus street (*i.e.*, response code 1 versus 0) is

$$\text{intracluster correlation} = \frac{0.7584}{0.7584 + \pi^2/3} = 0.187,$$

and the intracluster correlation for independent versus street, respectively (*i.e.*, response code 2 versus 0) is

$$\text{intracluster correlation} = \frac{5.4468}{5.4468 + \pi^2/3} = 0.623.$$

The intracluster correlation is given at the end of the output file as shown below. For the independent versus street housing comparison the intracluster correlation is higher.

TWO-LEVEL MODELS FOR NOMINAL OUTCOMES

```
SuperMix - [sdhouse.out]
 File   Analysis   Window   Help

Calculation of intracluster correlation (O1)
--------------------------------------------
residualvariance = pi*pi / 3 (assumed)
cluster  variance =     0.7584

intracluster correlation =     0.7584 /(     0.7584 + (pi*pi/3)) = 0.187


Calculation of intracluster correlation (O2)
--------------------------------------------
residualvariance = pi*pi / 3 (assumed)
cluster  variance =     5.4468

intracluster correlation =     5.4468 /(     5.4468 + (pi*pi/3)) = 0.623

 Save As...      Close
```

3.8   Two-level survival analysis models

3.8.1   The data

The data set for this example is taken from Schoenwald & Henggeler (2005). Children in the study were assigned to therapists and followed across time. At the child level, data were collected at baseline (pre-treatment, T_0), post-treatment (T_1), 6 months post-treatment (T_2), and 12 months post-treatment (T_3). The outcome of interest is whether a child was suspended in the current school year, assessed at T_0, T_1, T_2, or T_3. Specifically, here, we will focus on the time until the first school suspension as the "survival" outcome. As indicated in more detail later, this is indicated by a combination of the variables Event and Suspend: for example, if the student was suspended, the indicator Event is given the value 1 and Suspend will indicate the time period during which this occurred. However, there are also subjects who do not experience the event (*i.e.*, were not suspended), and who drop out of the study before its end. Such subjects are considered to be right-censored in the survival analysis literature, and for these subjects the Event variable is coded 0 and the Suspend variable indicated the last time period prior to their dropout from the study. For subjects who never experience the event and who never drop out, they receive Event codes of 0 and Suspend codes equal to the final time point. In addition to these data concerning school suspension, the gender of each student was also recorded, as well as whether or not the student's family was receiving financial assistance. The first 8 cases of the data set **suspend.ss3** are shown below.

suspend.ss3						
	[A] Theraps	[B] YouthID	[C] Suspen	[D] Event	[E] SexF	[F] FinnAsst	[G] SexFin
1	18	452	1	1	0	0	0
2	18	509	1	1	0	1	0
3	18	566	1	1	0	1	0
4	18	1020	2	0	0	1	0
5	22	231	4	0	0	0	0
6	22	306	4	0	0	0	0
7	29	208	3	1	0	1	0
8	29	232	1	1	0	1	0
9	29	315	1	1	0	0	0
10	29	349	1	1	0	1	0

The variables of interest are:

- o Therapst is the patient therapist ID (443 level-2 units).
- o YouthID is the child's ID (1914 level-1 units).
- o Suspend is an ordinal outcome variable that assumes values 1, 2, 3 or 4, corresponding to the time points T_0, T_1, T_2, and T_3.
- o Event is the event indicator, where 1 indicates suspension took place and 0 that the observation was censored.
- o SexF indicates the child's gender (1 = female; 0 = male).
- o FinnAsst equals 1 if financial assistance is given to the student's family and 0 otherwise
- o SexFin equals SexF × FinnAsst and therefore assumes values of 0 and 1.

3.8.2 The model

Let y_{ij} denote an ordinal outcome variable that takes on discrete positive values $t = 1, 2, \ldots, m$. In previous examples, we assumed that y_{ij} has C categories. For example, 1 = not depressed, 2 = mildly depressed, 3 = depressed, and 4 = extremely depressed. The subscript (i, j) denotes subject j, $j = 1, 2, \ldots n_i$ nested within level-2 unit i, $i = 1, 2, \ldots, N$. In the present context the level-1 units j indicates children and the level-2 unit i indicates therapists. Note, that as another example of this type of model, one could have multiple failure times nested within individuals.

Let δ_{ij} denote the censor/event indicator, then $\delta_{ij} = 1$ if the event occurs and $\delta_{ij} = 0$ if an observation is censored. In survival analysis each ij is observed until time t_{ij} and if an event occurs $t_{ij} = t$ and $\delta_{ij} = 1$. If the observation is censored at $t_{ij} = t$ then $\delta_{ij} = 0$.

In the case of censoring it is assumed that a unit is observed at t_{ij} but not at t_{ij+1}. Hedeker, Siddiqui & Hu (2000) showed that if events occur within continuous time intervals (*i.e.*, grouped-time), for example, a student is suspended in the past year, use of the complementary log-log link for an ordinal outcome is equivalent to a proportional hazards model in continuous time. Therefore, the grouped-time proportional hazards mixed model can be written as:

$$\log\left[-\log\left(1 - P\left(t_{ij}\right)\right)\right] = \gamma_t + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i$$

where $\mathbf{x}_{ij}$ is a vector of explanatory variables and $\mathbf{z}_{ij}$ a vector of random effects. Typically, the elements of $\mathbf{z}_{ij}$ are a subset of $\mathbf{x}_{ij}$. For example, the elements of $\mathbf{z}_{ij}$ might correspond to the intercept and age, whereas $\mathbf{x}_{ij}$ would include these two terms plus any additional model covariates. It is assumed that the random effects $\mathbf{u}_i$ are from a normal distribution with mean zero and covariance matrix $\boldsymbol{\Phi}_{(2)}$.

$P\left(t_{ij}\right)$ denotes the probability that an event takes place in the interval designated at time t_{ij}. γ_t represent threshold values, and in the present context these reflect the baseline hazard (*i.e.*, the hazard when all covariates equal 0). The plus sign following γ_t means that a positive $\boldsymbol{\beta}$ indicates an increased hazard (*i.e.*, the event occurs sooner) as values of the covariate increase.

TWO-LEVEL SURVIVAL ANALYSIS MODELS

Survival data as ordinal outcomes

Assume 4 time points with no intermittent censoring and let y denote the outcome variable. Let us first consider subjects who were suspended at some point in the study. For these subjects, the variable Event will be coded as 1 and the coding of the Suspend variable will be as follows.

Suspend:

$y_{ij} = 1$: Student first suspended at T_0.

$y_{ij} = 2$: Student not suspended at T_0, but first suspended at T_1.

$y_{ij} = 3$: Student not suspended at T_0 or T_1, but first suspended at T_2.

$y_{ij} = 4$: Student not suspended at T_0, T_1 or T_2, but first suspended at T_3.

Similarly, subjects who were never censored would have the variable Event coded as 0, and the following codes for the Suspend variable.

Suspend:

$y_{ij} = 1$: Student not suspended at T_0 and no data beyond T_0.

$y_{ij} = 2$: Student not suspended at T_0 or T_1, and no data beyond T_1.

$y_{ij} = 3$: Student not suspended at T_0, T_1, or T_2, and no data beyond T_2 (*i.e.*, no data at T_3).

$y_{ij} = 4$: Student not suspended at T_0, T_1, T_2, or T_3.

Table 3.10 shows how values are assigned to y_{ij}, and the relationship between the y_{ij} outcomes and the event indicator. It should be noted that one could also fit grouped-time survival models using dichotomous indicators of event/censoring across the study time points. This approach, which is described in Singer and Willett (1993), can also be done in SuperMix, though additional data setup and manipulation is required. The advantage of representing the survival data as ordinal outcomes is that there is no need to include time indicators since the thresholds take care of this. The ordinal presentation is also more efficient in terms of data set size, especially when the number of time points is large. More information on these two different approaches can be found in Hedeker, Siddiqui & Hu (2000).

Table 3.10: Four time points with censoring

Outcome	Ordinal dep. Variable	Event indicator
Censor at T_1	1	0
Event at T_1	1	1
Censor at T_2	2	0
Event at T_2	2	1
Censor at T_3	3	0
Event at T_3	3	1
Censor at T_4	4	0
Event at T_4	4	1

3.8.3 Example: Survival analysis model

The model is fitted to the data in **suspend.ss3** as follows. The first step is to create the **ss3** file shown above from the Excel file **suspend.xls**. This is accomplished as follows:

- o Use the **Import Data File** option on the **File** menu to open the **Open** dialog box.
- o Browse for the file **suspend.xls** in the **Examples\Primer\Survival** folder.
- o Select the file and click on the **Open** button to open the following SuperMix spreadsheet window for **suspend.ss3**.

Setting up the analysis

We start by selecting the **New Model Setup** option on the **File** menu to open the **Model Setup** window.

First, enter the titles Survival Analysis Using Ordered Responses and Complementary log-log link function in the **Title 1** and **Title 2** text boxes respectively. Select the ordinal outcome variable Suspend from the **Dependent Variable** drop-down list box. The variable Therapst, which defines the levels of the hierarchy, is selected as the Level-2 ID from the **Level-2 IDs** drop-down list box. Also, set the number of iterations to 50.

Next, click on the **Variables** tab of the **Model Setup** window. SexF, FinnAsst, and SexFin are specified as the predictors (explanatory variables) of the fixed part of the model by checking the corresponding boxes in the **E** column of the **Available** grid on the **Variables** screen. These actions will produce the following screen.

Model Setup: suspend1.mum

Configuration | Variables | Starting Values | Patterns | Advanced | Linear Transforms

Available	E	2
Therapst	☐	☐
YouthID	☐	☐
Suspend	☐	☐
Event	☐	☐
SexF	☑	☐
FinnAsst	☑	☐
SexFin	☑	☐

Explanatory Variables
SexF
FinnAsst
SexFin

L-2 Random Effects

☑ Include Intercept

Use the arrow keys or click on the desired tab to select the category of interest for the model.

Model Setup: suspend1.mum

Configuration | Variables | Starting Values | Patterns | Advanced | Linear Transforms

General Settings

Unit Weighting: equal

Optimization Method: non-adaptive quadrature

Number of Quadrature Points: 25

Explanatory Variable Interactions

Include Interactions: no

Ordered Dependent Variable Settings

Function Model: complementary log-log

Level-2 Random Thresholds: no

Right-Censoring: include

Censor Variable: Event

Model Terms: add
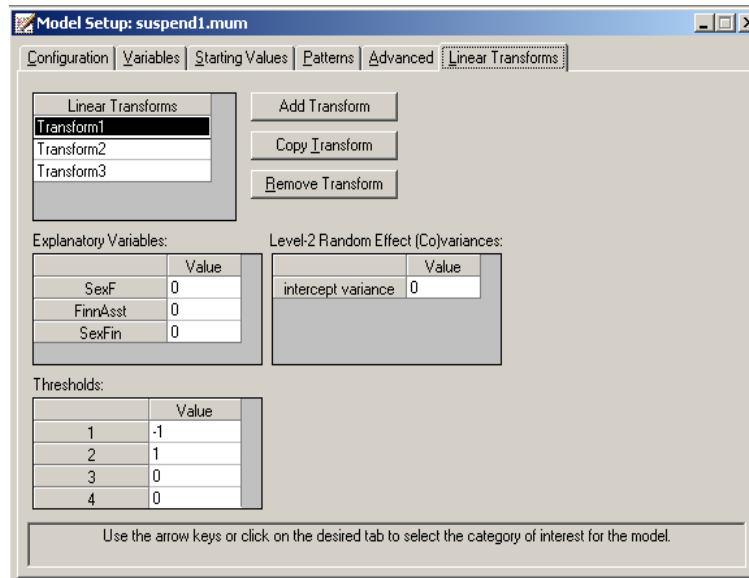
Use the arrow keys or click on the desired tab to select the category of interest for the model.

To specify the number of quadrature points and link function (**Function Model**), we proceed to the **Advanced** screen by clicking on the **Advanced** tab. Change **Model Terms** from **subtract** to **add** and select **complementary log-log** as the **Function Model**. Select **non-adaptive quadrature** as **Optimization Method**, and request 25 **quadrature points**. Finally, set the **Right-censoring** field to **include**, and select the variable Event as **Censor Variable**.

To complete the model setup, we will illustrate the use of the **Linear Transforms** option. In the current model specification, the baseline hazard at time point j, $j = 1$, 2, 3, 4 is estimated by the jth threshold parameter when all covariates equal 0. A so-called alternative parameterization of the threshold parameters is obtained by assuming that threshold 1 equals 0 and by including an intercept term in the fixed part of the model. In this case, the baseline hazard estimate at the first time point equals the estimated model intercept; the baseline hazard estimate at the second time point is the sum of the model intercept and the first estimated threshold; the baseline hazard at the third time point is the sum of the model intercept and the second estimated threshold; and the baseline hazard estimate at the final time point is a sum of the estimated intercept and the third estimated threshold. To obtain the set of alternative parameters listed in the output, use is made of linear transformations of the model estimates as specified on the **Linear Transforms** tab. The three linear transforms are specified as follows:

SexF	FinnAsst	SexFin	Random Intercpt (L-2)	Threshold 1	Threshold 2	Threshold 3	Threshold 4
0	0	0	0	−1	1	0	0
0	0	0	0	−1	0	1	0
0	0	0	0	−1	0	0	1

The screen below shows the values entered for the first transform:

For the second transform, values of –1, 0, 1, 0 are assigned for the thresholds. The third transform contains values of –1, 0, 0, 1 for the three thresholds. This step completes the model set-up. Use the **File**, **Save** option to save the model setup to a file named **suspend1.mum**. Next, use the **Analysis**, **Run** option on the main menu bar to run the analysis.

Discussion of results

The portion of the output file shown below indicates that the assumed sampling distribution is multinomial and the complementary log-log link function is used in the analysis. It indicates that there are 443 therapists. Nested within these level-2 units are 1914 subjects. A summary of the number of level-1 observations per level-2 unit (only first two lines shown) is also given.

```
 SuperMix - [suspend1.out]                                          _ □ X
  File  Analysis  Window  Help                                      _ ☐ X

 o===========================================o
 | Survival Analysis Using Ordered Responses |
 | Complementary log link function           |
 o===========================================o


              Model and Data Descriptions

    Sampling Distribution              = Multinomial
    Link Function                      = Cumulative CLL
    Number of Level-2 Units                443
    Number of Level-1 Units                1914
    Number of Level-1 Units per Level-2 Unit =
       4    2    8    1    1    1    1    1    1    1    8   11
       7    4    5    7   10   11    5   15    2    1    5    1
       2    2    6    1    2    1    4    1   26    9    3   14
       1   15    1    5    4    1   14    2    3    1    7    3
       7    8    7   14    1    6    1    6   13   17    8   13

 ◄                                                                     ►

    Save As...        Close
```

This part of the output is followed by descriptive statistics for all the variables. The variable Suspend has four categories, which are recoded into four dummy variables Suspend1, Suspend2, Suspend3 and Suspend4. We note that all the variables in this section are dichotomous.

```
 SuperMix - [suspend1.out]                                          _ □ X
  File  Analysis  Window  Help                                      _ ☐ X

 o===========================================================o
 | Descriptive statistics for all the variables in the model |
 o===========================================================o
                                                    Standard
      Variable    Minimum     Maximum      Mean     Deviation
      --------    -------     -------      ----     ---------
      Suspend1     0.0000      1.0000     0.4316     0.4954
      Suspend2     0.0000      1.0000     0.1855     0.3888
      Suspend3     0.0000      1.0000     0.1102     0.3133
      Suspend4     0.0000      1.0000     0.2727     0.4455
      SexF         0.0000      1.0000     0.3459     0.4758
      FinnAsst     0.0000      1.0000     0.3626     0.4809
      SexFin       0.0000      1.0000     0.1155     0.3197

 ◄                                                                     ►

    Save As...        Close
```

The means of the recoded dichotomous Suspend variables reflects the proportions of subjects assigned a value of 1, 2, 3 or 4 are 0.432, 0.186, 0.110 and 0.273 respectively. A crosstabulation of Suspend by Event is given in Table 3.11. It follows that, for example, 773 students out of the 1914 in the study were suspended prior to treatment (T_0). For 53 children, we only know that they were not suspended at T_0, thereafter they are missing and treated as right-censored.

Table 3.11: Crosstabulation of Suspend by Event

T_0	T_1	T_2	T_3
773	255	106	72
53	100	105	450

Parameter estimates are given in the next part of the output. We conclude that there is no gender-financial assistance interaction and that all the remaining parameter estimates are significant. The effect of SexF is negative indicating that girls have a significantly decreased hazard (*i.e.*, a longer time to the first suspension), relative to boys. The FinnAsst estimate is positive indicating an increased hazard (shorter time to first suspension) for children from families receiving financial assistance, relative to children from families not receiving this assistance.

```
┌─ SuperMix - [suspend1.out] ──────────────────────────────────── _ □ × ─┐
│  File  Analysis  Window  Help                                    _ 日 ×  │
├──────────────────────────────────────────────────────────────────────┤
│                                                                      ▲  │
│  o========================================================o            │
│  | Optimization Method: Non-Adaptive Quadrature |                       │
│  o========================================================o            │
│                                                                         │
│  Number of quadrature points =            25                            │
│  Number of free parameters =               8                            │
│  Number of iterations used =               3                            │
│                                                                         │
│  -2lnL (deviance statistic) =      4741.46570                           │
│  Akaike Information Criterion      4757.46570                           │
│  Schwarz Criterion                 4801.92131                           │
│                                                                         │
│                   Estimated regression weights                          │
│                                                                         │
│                                      Standard                           │
│  Parameter            Estimate       Error      z Value    P Value      │
│  ---------            --------       --------   -------    -------       │
│  Threshold1            -0.6566         0.0549   -11.9561    0.0000       │
│  Threshold2            -0.2239         0.0518    -4.3246    0.0000       │
│  Threshold3            -0.0328         0.0511    -0.6415    0.5212       │
│  Threshold4             0.1209         0.0512     2.3628    0.0181       │
│  SexF                  -0.3208         0.0818    -3.9203    0.0001       │
│  FinnAsst               0.2004         0.0742     2.7008    0.0069       │
│  SexFin                -0.0236         0.1345    -0.1756    0.8606       │
│                                                                         │
│                                                                         │
│  Alternative Parameterization, setting Threshold1= 0                    │
│                                                                         │
│                   Estimated regression weights                          │
│                                                                         │
│                                      Standard                           │
│  Parameter            Estimate       Error      z Value    P Value      │
│  ---------            --------       --------   -------    -------       │
│  intcept               -0.6566         0.0549   -11.9561    0.0000       │
│  Threshold2             0.4327         0.0258    16.7689    0.0000       │
│  Threshold3             0.6238         0.0306    20.3627    0.0000       │
│  Threshold4             0.7776         0.0344    22.6099    0.0000       │
│                                                                         │
│                                                                         │
│             Estimated level 2 variances and covariances                 │
│                                                                         │
│                                     Standard                            │
│  Parameter                Estimate   Error     z Value  P Value         │
│  ---------                --------   --------  -------  -------          │
│  intercept/intercept        0.0841     0.0302   2.7819   0.0054          │
│                                                                      ▼  │
│ ◄                                                                  ►     │
├──────────────────────────────────────────────────────────────────────┤
│   Save As...      Close                                                 │
└──────────────────────────────────────────────────────────────────────┘
```

The last part of the output contains an estimate of the intracluster correlation. Although this estimate indicates a modest therapist effect, the random effect variance term is highly significant. From this, we conclude that the time until suspension does vary significantly across therapists.

TWO-LEVEL SURVIVAL ANALYSIS MODELS

```
SuperMix - [suspend1.out]                                                   _ □ ×
File  Analysis  Window  Help                                                _ 8 ×

Calculation of the intracluster correlation
-----------------------------------------
residual variance = pi*pi / 6 (assumed)
cluster  variance =     0.0841

intracluster correlation =     0.0841 /(    0.0841 + (pi*pi/6)) = 0.049


             |------------------------------------|
             |          TESTING OF TRANSFORMS       |
             |    (General Linear Hypothesis Testing) |
             |------------------------------------|


  ----------------------------------------------------------------------------
   Coefficients                              Estimate    Transform No.
                                                         1        2        3
  ----------------------------------------------------------------------------
      1 Threshold1                           -0.65664  -1.0000  -1.0000  -1.0000
      2 Threshold2                           -0.22393   1.0000   0.0000   0.0000
      3 Threshold3                           -0.03280   0.0000   1.0000   0.0000
      4 Threshold4                            0.12094   0.0000   0.0000   1.0000
      5 SexF                                 -0.32076   0.0000   0.0000   0.0000
      6 FinnAsst                              0.20045   0.0000   0.0000   0.0000
      7 SexFin                               -0.02362   0.0000   0.0000   0.0000
      8 Var(intercept)                        0.08412   0.0000   0.0000   0.0000

   Transform Estimate                                   0.4327   0.6238   0.7776
   Standard Error                                       0.0258   0.0306   0.0344
   Z-Statistic                                         16.7689  20.3627  22.6099
   Exceedence Pobability                                0.0000   0.0000   0.0000

 Save As...      Close
```

A summary of the transforms is given followed by a significance test for each
transform. Setting the intercept equal to the first threshold, these provide the
estimates of the alternative parameterization.

If we remove the predictors SexF, FinnAst and SexFin from the model, the baseline
hazard estimates for the four study time points are –0.6881, –0.2607, –0.0733, and
0.0784. These can be converted to the probability scale using the inverse of the
complementary log-log function, namely,

$$P(z)=1-\exp[-\exp(z)]$$

This yields probability estimates of the baseline hazard for the first school suspension as 0.3950, 0.5372, 0.6051, and 0.6609 across these four study time points. Note that these are conditional estimates, conditional on the therapist effects. In other words, they are estimates controlling for the effect of therapist on the individual student outcomes.

4 Graphical Displays

4.1 Introduction

Graphics are often a useful data-exploring technique through which the researcher may familiarize her- or himself with the data. Relationships and trends may be conveyed in an informal and simplified visual form via graphical displays. SuperMix offers both data-based and model-based graphs, discussed in the sections to follow. In the case of data-based graphs, we distinguish between three categories: univariate, bivariate, and multivariate graphs. Univariate graphs are particularly useful to obtain an overview of the characteristics of a variable. However, they do not necessarily offer the tools needed to explore longitudinal data as completely as one would wish. For that purpose, bivariate and multivariate data-based graphs are more appropriate.
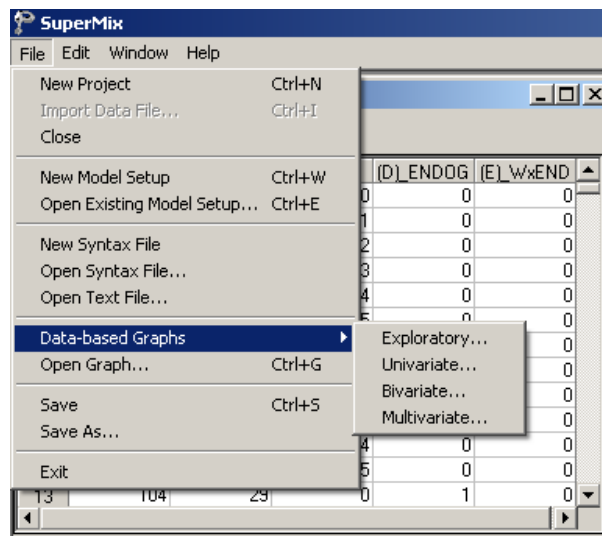
Data-based graphing options are accessed via the **File**, **Data-based Graphs** option once a **SuperMix** data file (**.ss3**) is opened, and include **Exploratory**, **Univariate**,

Bivariate and **Multivariate** graphs as shown on the pop-up menu below. Each of these options will be discussed in Sections 4.2 to 4.5.

Model-based graphing options can be accessed once a **SuperMix** data file (**.ss3**) is opened and the **New Model Setup** or **Open Existing Model Setup** option is selected from the **File** menu. Presently the available options are **Equations**, **Residuals**, and **Confidence Intervals**. These options are illustrated in Section 4.6. Note that the **Model-based Graphs** options are only displayed if the **Model Setup** screen (**.mum** file) is the active window. Furthermore, the graphical displays are only available after an analysis has been run successfully via the **Analysis**, **Run** option on the main menu bar.
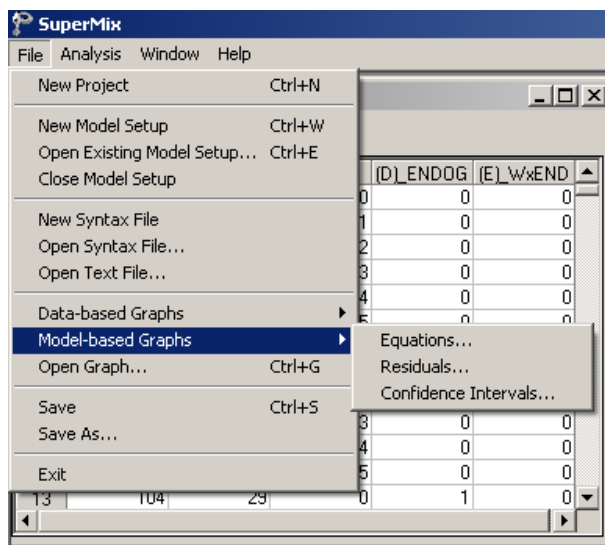
Except for the exploratory graphs discussed in Section 4.2, graph parameters that may be changed include the axes and descriptions thereof, the symbols used and the colors assigned to the symbols/text. To change any of these, simply double-click on the symbol/text to be changed to activate a dialog box in which changes can be requested. Use of such dialog boxes will be illustrated in the course of the discussion of the various graphs in the sections to follow. Section 4.7 provides an overview of the graph editing tools.

4.2 Data-based graphs: Exploratory graphics

The exploratory utility is specifically useful for the visual display of trends in longitudinal studies with continuous outcome variables.
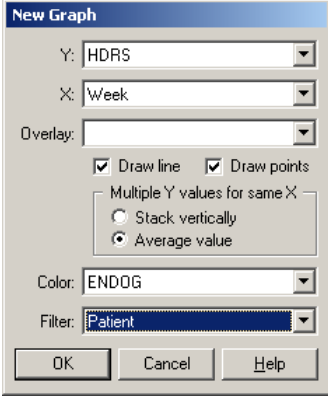
To invoke the exploratory graphics procedure, a SuperMix data file (***.ss3**) has to be opened. To illustrate we use **Examples\Primer\Graphics\reisby.ss3**. Select the **File**, **Data-based graphs**, **Exploratory** option as shown below.
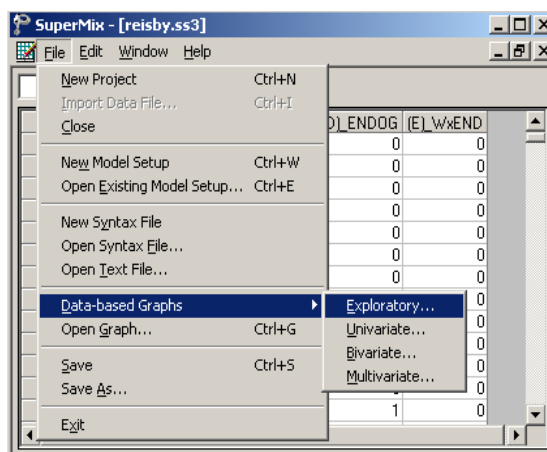
This selection activates the **New Graph** dialog box shown next. Select HDRS and Week as the dependent (vertical axis) and independent (horizontal axis) variables.

Additionally, select Patient as the so-called **Filter** variable. This specifies that a trend line will be displayed for each patient. To indicate whether a specific patient has been classified as endogenous or non-endogenous, the variable ENDOG is selected

as the **Color** attribute. The color option is very useful when the data set includes predictors such as gender (female, male) and cholesterol level (low, normal, high).

The graph below shows the trend in HDRS scores over time for patient 101, who is classified as non-endogenous.
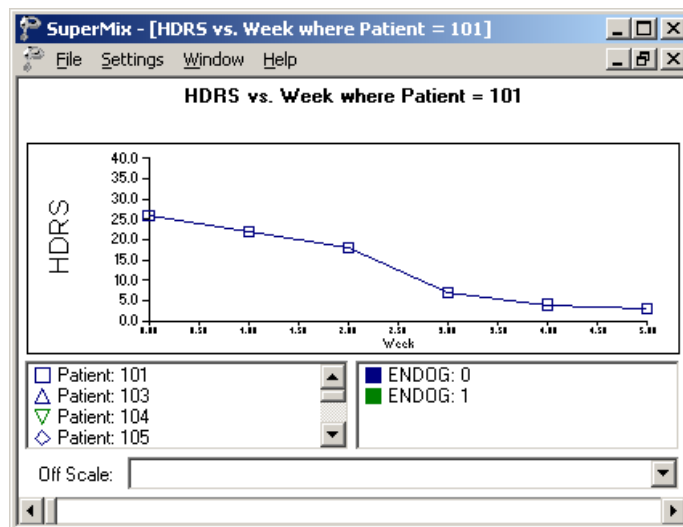
By clicking on the right-hand arrow of the slider bar at the bottom of the graphics screen, the trend in HDRS scores is shown for the next patient (patient 103). This patient is classified as endogenous.
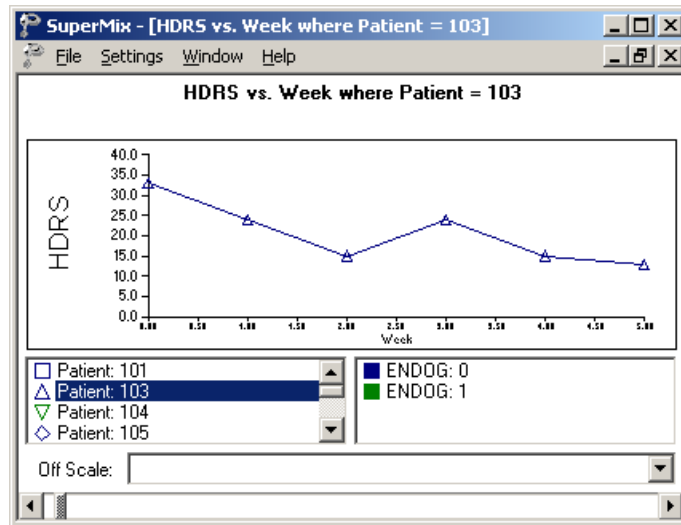
The "slider" can be used to skip over patients and the left arrow used to obtain the previous graph.

Average trends

It is often of interest to obtain a graphical display of the trend of the outcome variable over time for subgroups of the data. In the next illustration, a simultaneous graphical display of HDRS against Week is requested for non-endogenous and endogenous patients. This is accomplished by selecting ENDOG as the **Overlay** variable. The default display is black and white, but by specifying ENDOG as the **Color** attribute, the trend lines are displayed in color.
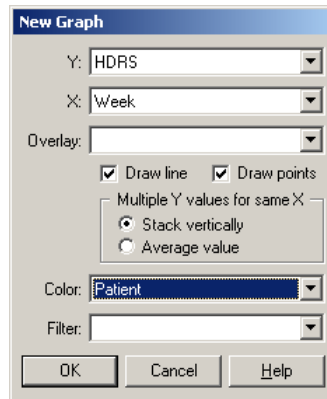
From the display below, it is evident that the average HDRS scores for non-endogenous patients (the square symbol) are consistently lower over time when compared to the endogenous patients (the triangle symbol).
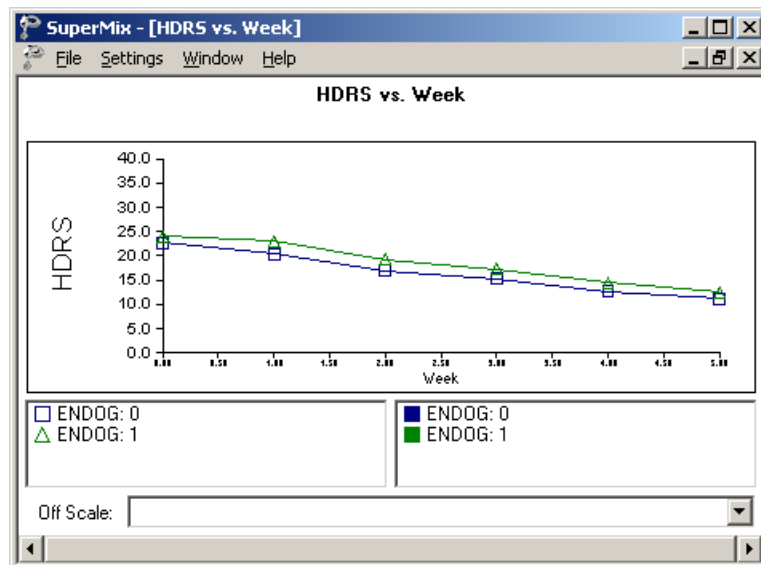
Variability in trend

It is often of interest to visually display the range of values of an outcome variable at different time points. To illustrate this type of graphic, select the **Stack vertically** option and select Patient as the **Color** attribute.

The resulting graph is shown below.
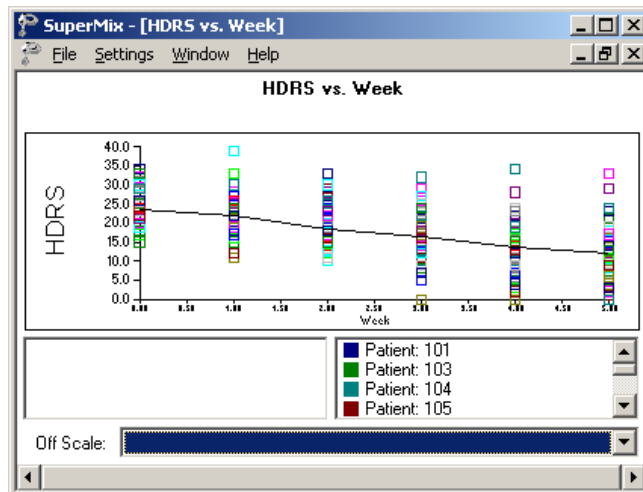
It is evident that there exists substantial variation in the HDRS scores at each time point. There also appear to be outlier values, corresponding to a few patients with much higher than average HDRS scores. At week 1 (recoded to 0), the range of HDRS values is approximately 14 to 35, whereas the range of HDRS values is approximately 0 to 35 at the end of sixth week (recoded to 5).

Editing exploratory graphs

Exploratory graphics created using the **Data-based Graphs**, **Exploratory…** option can be saved as a **.mug** file via the **File**, **Save** or **File**, **Save As** options. To illustrate, we saved the previous graph as **reisby1.mug** in the **Graphics** folder. To display or edit a saved graphics file on a later occasion, select the **File**, **Open Graph** option from the main menu bar.
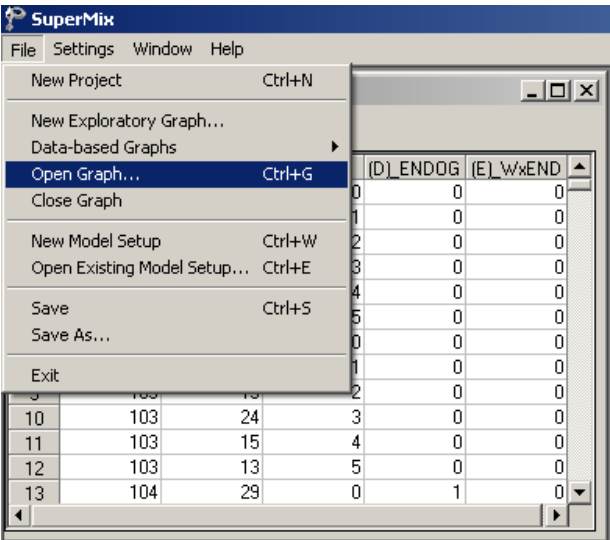
Browse for the file **reisby1.mug** and click the **Open** button to display the graph.
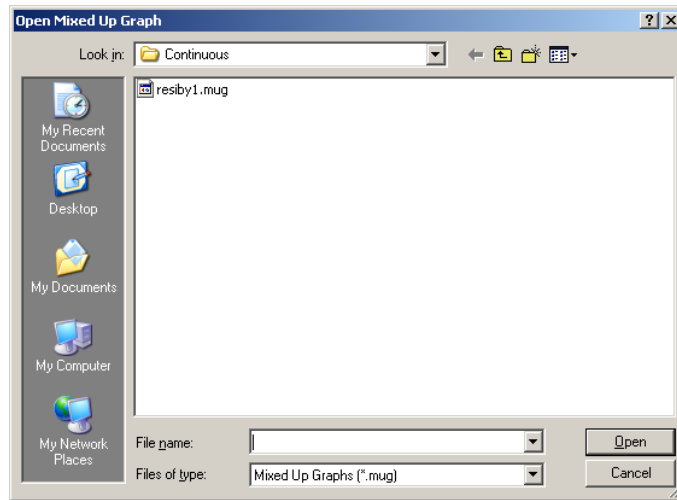
To edit the previously defined settings, select the **Settings**, **Edit Graph** option from the main menu bar.
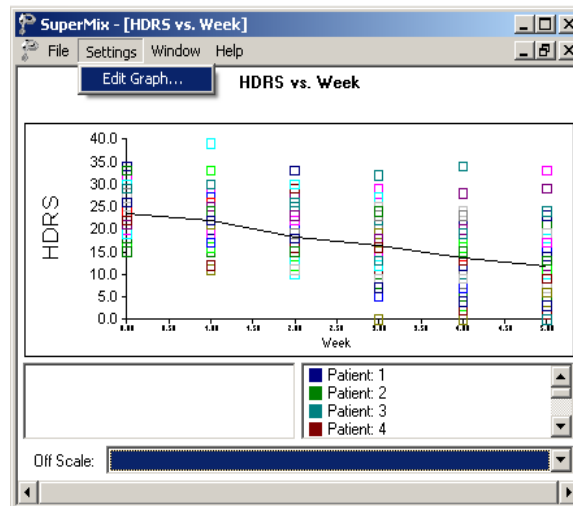
Change the previous displayed graph by choosing the settings shown on the **Edit Graph** dialog box below. Note that a trend line will be displayed for each patient.
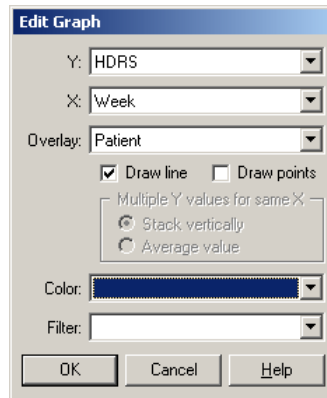
These trend lines are overlaid on a single axis system as shown below.
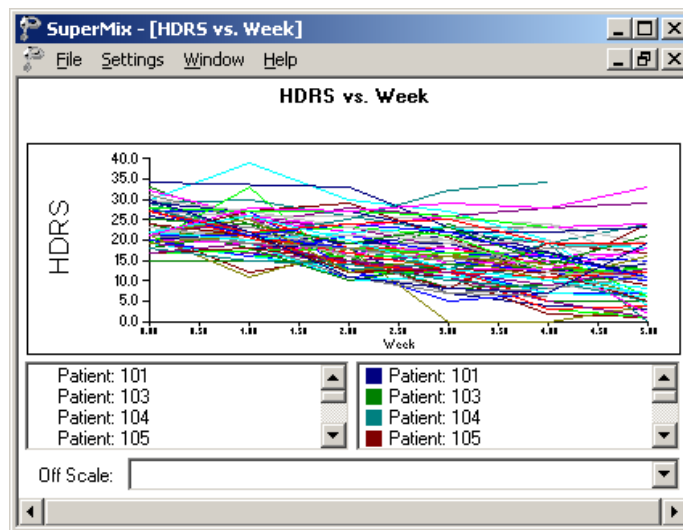
The visual display of the HDRS trends clearly demonstrates the large variation in HDRS scores over time.

4.3 Data-based graphs: Univariate graphs

4.3.1 Pie Chart

A pie chart display of the percentage distribution of a variable may be obtained if the variable does not have more than 15 distinct values. A pie chart is a graphic representation of percentages or frequencies by means of a circle that is subdivided into slices in such a way that the areas of these slices are proportional to the percentages or frequencies. Pie charts may be customized by using the graph editing dialog boxes as shown in this section.
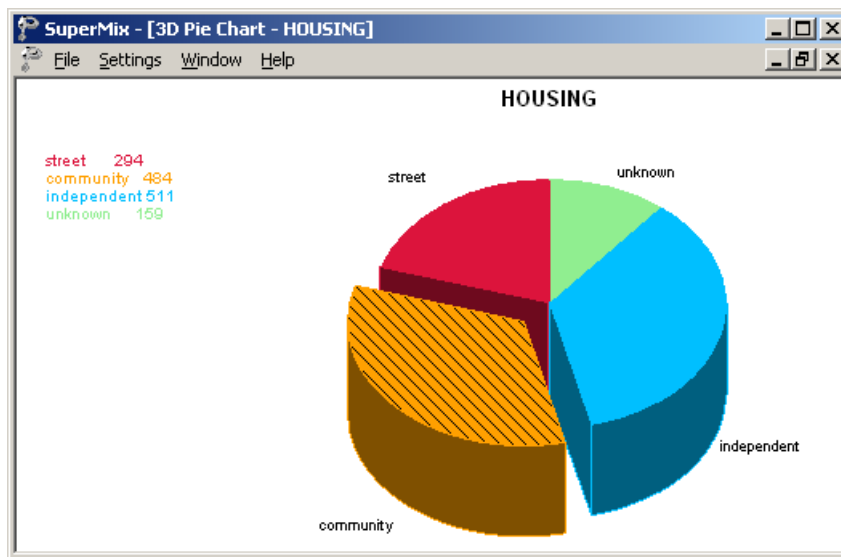
The data from the McKinney Homeless Research Project study (see Section 3.7) were used to produce a pie chart representing the variable HOUSING as given in **Examples\Primer\Graphics\sdhouse.ss3**. Recall that subjects' housing status was recorded at 4 time points. Over all time points, 511 subjects reported living in independent housing, 294 subjects were living on the street, 484 lived in community housing, and 159 cases were classified as unknown.

Creating a pie chart

To create the pie chart shown above, start by opening the spreadsheet **sdhouse.ss3**. Next, right-click on the column header of the variable HOUSING and select the **Column Properties** option from the pop-up menu.
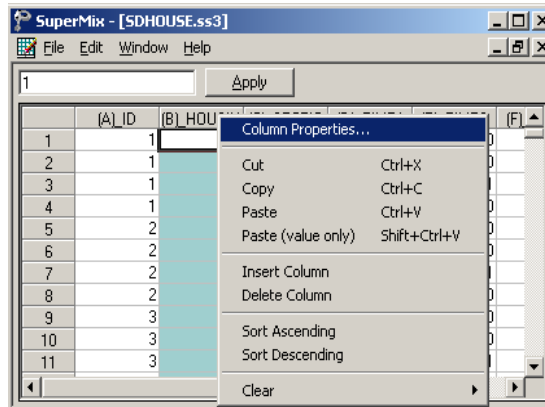
The **Column Properties** dialog box is displayed. This dialog box is used to define the type of variable (categorical or continuous) and to provide labels for the categories of nominal and ordinal variables. Indicate HOUSING as a categorical variable by clicking the appropriate radio button, and enter the labels for each category in the **Label** column as shown below. Click **OK** to return to the spreadsheet window and save the changes to the spreadsheet using the **File**, **Save** option.
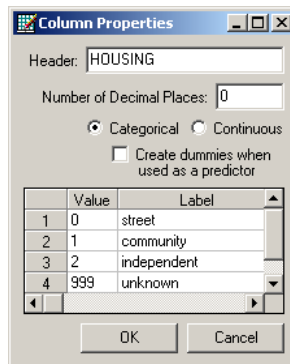
With the variable HOUSING defined as nominal and labels assigned to its categories, proceed to make the pie chart. First, select the **File**, **Data-based Graphs**, **Univariate** option from the pop-up menu as shown below.
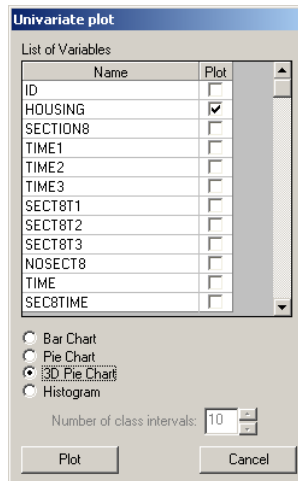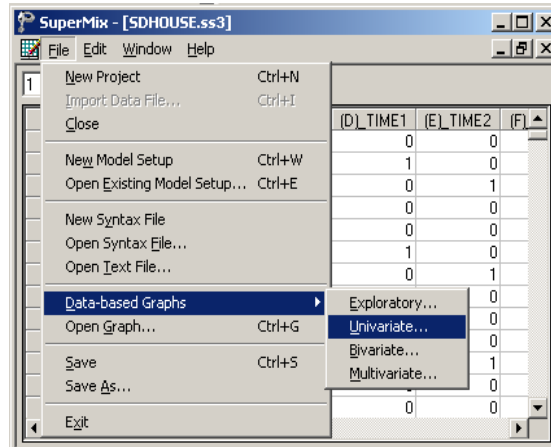
The **Univariate plot** dialog box is displayed. Select the variable HOUSING as the variable to be plotted, and request a **3D Pie Chart** by clicking the radio button next to this option. Click the **Plot** button to display the pie chart.
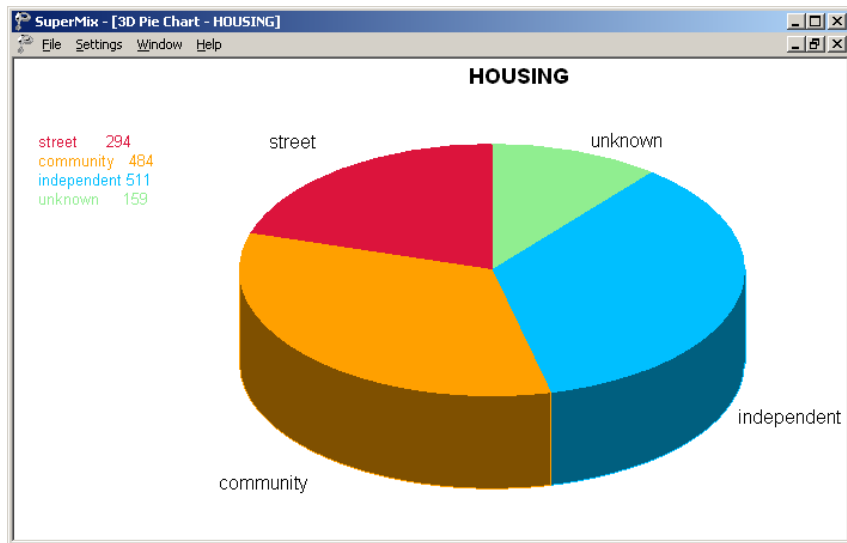
The attributes of the pie chart can be changed by clicking on areas of the graph. For example, clicking on the legend box in the top right corner of the graph will activate the **Legend Parameters** dialog box. This dialog box can be used to change the attributes of the legend box and its contents.
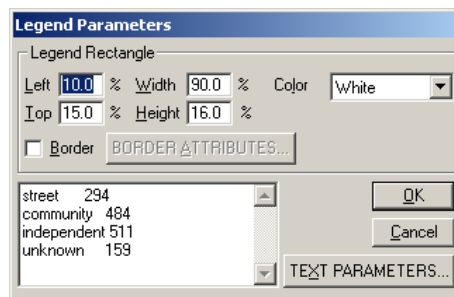
Click on the **Text Parameters** button to change the font and size of text of legend as shown below.
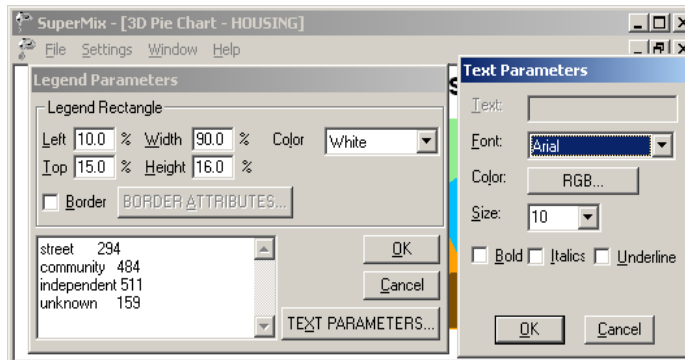
The appearance of the pie chart can be changed by clicking anywhere on the pie chart to activate the **Pie Chart Parameters** dialog box as shown below.
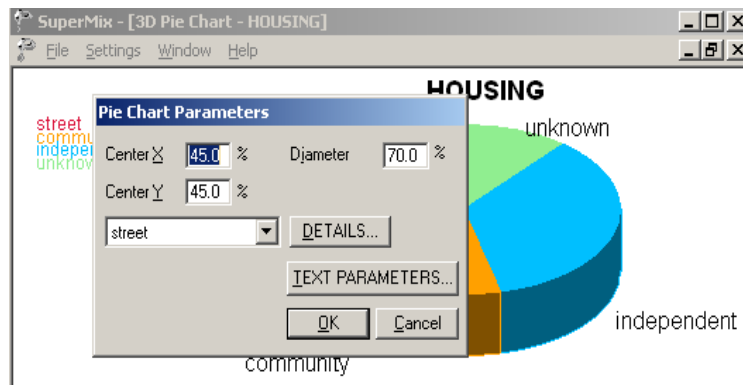
To move a slice of the pie chart, select the category of interest and click the **Details** button. Change the value in the **Explosion** field, for example to 0.2, change the hatch style and color if wanted and click **OK** on both the **Pie Slice Parameters** and **Pie Chart Parameters** dialog boxes to obtain the pie chart shown below.
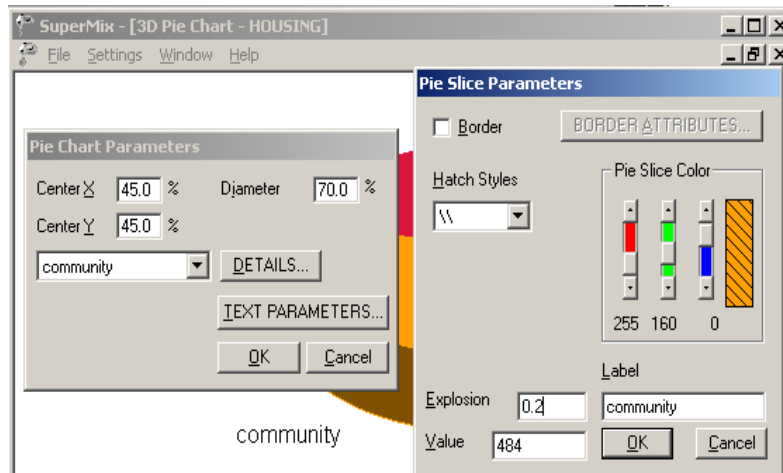
The color of each pie slice can be changed by moving each of the three color sliders shown above.

GRAPHICAL DISPLAYS
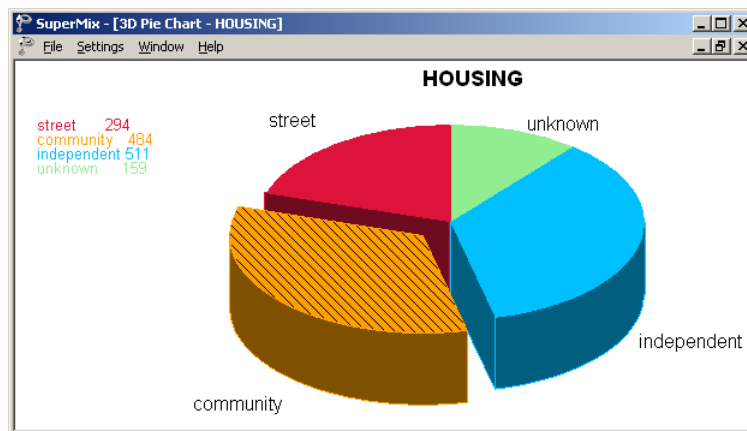
4.3.2 Bar chart

Discrete or categorical data can also be displayed graphically using a bar chart. A bar chart is a graphic representation of the frequency distribution of discrete or categorical data in which the values or categories are given on the horizontal axis and the frequencies are given on the vertical axis.

In Section 3.5 we considered longitudinal data from the NIMH Schizophrenia Collaborative Study on treatment-related changes in overall severity. An ordinal version of the Item 79 of the Inpatient Multidimensional Psychiatric Scale (IMPS) was used as the outcome variable. The variable Imps79O from the data set **schizx1.ss3** had four categories: 1 = normal or borderline mentally ill, 2 = mildly or moderately ill, 3 = markedly ill, and 4 = severely or among the most extremely ill. This variable can be represented graphically by either a pie or bar chart. In the graph below, we opted to display this variable in the form of a bar chart that offers the added advantage of retaining the ordinality of the variable in question.

Most of the total of 1603 measurements obtained over the course of the study indicated severe mental illness. The second largest group of measurements indicated mild mental illness. While this graph offers an instructive snapshot of the distribution of the ordinal variable Imps79O, it does not allow us to take the longitudinal nature of this study or the treatment group individuals were assigned to into account. To do so, we need to consider bivariate or multivariate graphics instead of univariate graphs (see Sections 4.4 and 4.5 for more information).
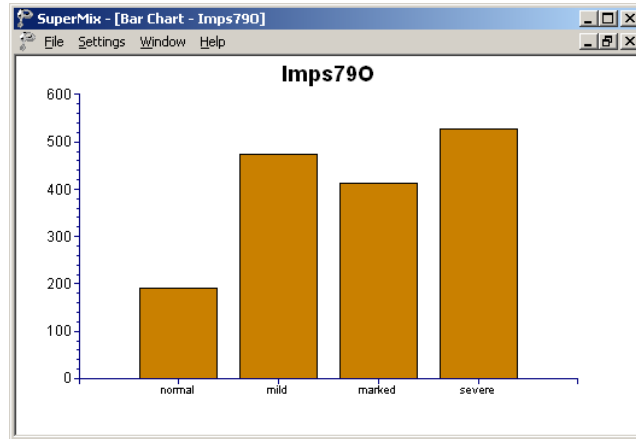
Creating a bar chart

Start by opening **Examples\Primer\Graphics\schizx1.ss3**. Right-click on the column header of the variable Imps79O and select the **Column Properties** option from the pop-up menu to open the **Column Properties** dialog box. This dialog box is used to define the type of variable (categorical or continuous) and to provide labels for the categories of nominal and ordinal variables. Indicate Imps79O as an ordinal variable by clicking the appropriate radio button and enter the labels for each category in the **Label** column as shown below. Click **OK** to return to the spreadsheet window and save the changes to the spreadsheet using the **File**, **Save** option.

Click the **Univariate** option from the pop-up menu displayed when the **File**, **Data-based Graphs** option is selected from the main menu bar.

Select the variable Imps79O by checking the box next to this variable in the **Plot** column on the **Univariate plot** dialog box and click **Plot** to display the bar chart.

Attributes of the graph are changed by activating the dialog boxes associated with sections of the graph. To change the attributes of the bars, click on a bar in the displayed area to activate the **Bar Graph Parameters** dialog box. This dialog box is usually easy to obtain if the mouse pointer is moved to the top part of the highest bar. The type of bar, color and position of the bars, and the display style can be changed by selecting options from the **Type**, **Bar Color**, **Position** and **Hatch Styles** fields. The **Bar Graph Parameters** dialog box shown below displays the default values of these fields. By clicking the **Data** button, the data plotted – in this case the frequencies associated with each category of the variable Imps79O – will be displayed in a separate window from where it can be copied to the clipboard. Note that if the data contains missing observations, the **Exclude Missing Values** on the **Settings** menu can be used to display a graph with the missing values excluded.
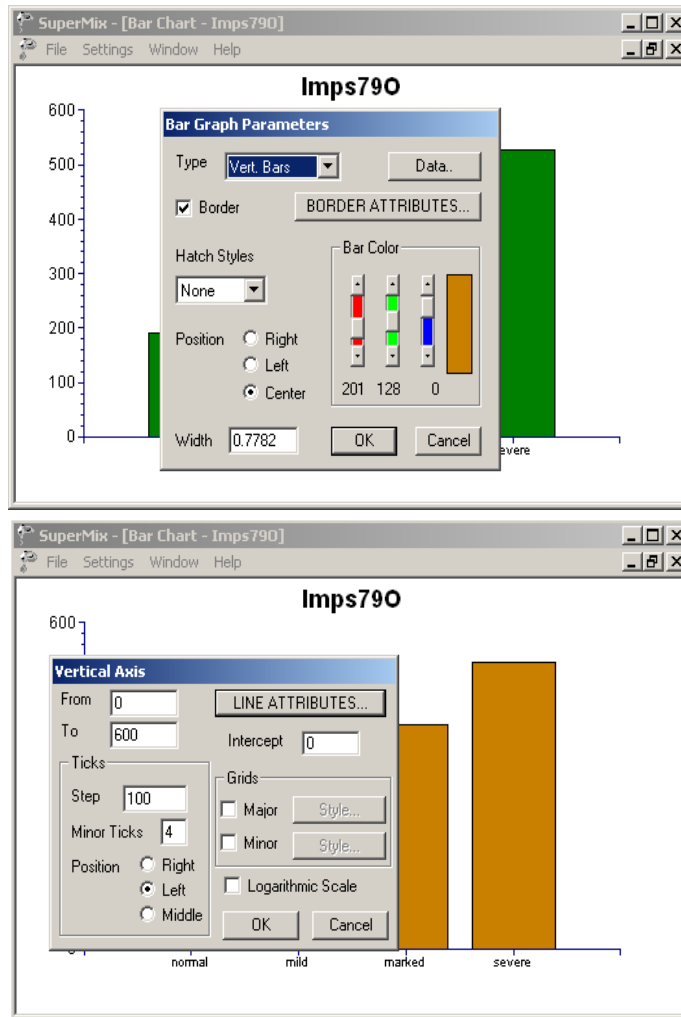
The attributes of the axes can be changed in a similar way: click anywhere on the vertical axis to display the **Vertical Axis** dialog box (shown below), or on the horizontal axis to display the **Horizontal Axis** dialog box. These dialog boxes are used to control the range, tick marks on and scale of each axis. The **Line Attributes** button provides access to the **Line Parameters** dialog box, used to change color, style and width of the axis.

4.3.3 Histogram

The distribution of a continuous variable such as the HDRS ratings in the Reisby data given in **Examples\Primer\Graphics\reisby.ss3** (see Section 3.2 for detailed analyses based on this data) can be graphically depicted by a histogram. Recall that these data are from a study described in Reisby *et. al*., (1977) that focused on the longitudinal relationship between imipramine (IMI) and desipramine (DMI) plasma levels and clinical response in depressed inpatients over a period of four weeks. The histogram offers a simple way to evaluate the distribution of the variable in question, and is a plot of the frequencies with which values occur (so-called class frequencies), against class intervals.

In the case of the HDRS ratings, we can easily verify that the mean HDRS rating is 17.637, with a standard deviation of 7.19. Ratings range between 0.0 and 39.0. The ratings can be summarized as shown in Table 4.1.

Table 4.1: Frequency distribution of HDRS ratings

Class interval	Class midpoint	Class frequency	Percentage
[0.0, 3.9)	1.95	11	2.9
[3.9, 7.8)	5.85	25	6.7
[7.8, 11.7)	9.75	39	10.4
[11.7, 15.6)	13.65	65	17.3
[15.6, 19.5)	17.55	81	21.6
[19.5, 23.4)	21.45	77	20.5
[23.4, 27.3)	25.35	49	13.1
[27.3, 31.2)	29.25	19	5.1
[31.2, 35.1)	33.15	8	2.1
[35.1, 39.0)	37.05	1	0.3

The ratings were grouped into 10 classes, each 3.9 in length as indicated in the first column. The second column shows the midpoint of each of the 10 classes, while the

third and fourth indicate the frequency and percentage of HDRS ratings that fall within a given class.

The graphical representation of the information in Table 4.1 is obtained by first selecting the **File**, **Data-based Graphs**, **Univariate** option, then selecting the variable HDRS and clicking the **Histogram** radio button, keeping the default **Number of class intervals** 10 as shown below.
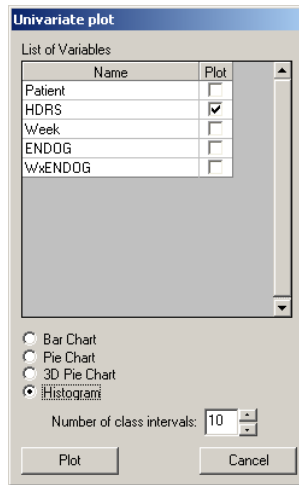
The appearance of the histogram is changed to display vertical 3D bars by clicking on the histogram to activate the **Bar Graph Parameters** dialog box, and by changing the default settings to those shown.
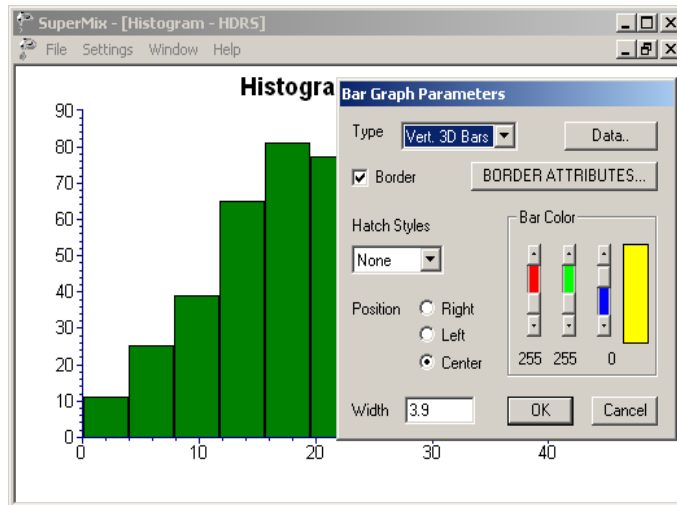
The 3D bar chart of the distribution of HDRS ratings is shown below.
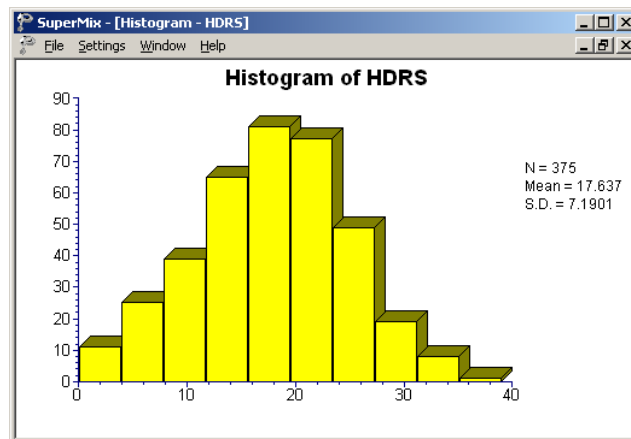
A histogram usually has the following characteristics:

o The class frequencies are plotted on the y-axis and the class intervals according to a convenient scale on the x-axis.

- o   With the class intervals as a basis, rectangles are drawn so that the area of each rectangle is proportional to the corresponding class frequency.
- o   It is common practice to leave spaces that are equal to class width on the left-hand side of the smallest class boundary and on the right-hand side of the largest class boundary.

From the histogram above, we conclude that the HDRS ratings are reasonably symmetrically distributed around the mean value of 17.637. In addition, very little data are found at the extreme left or right of the presentation. In general, the bell-shaped curve with short tails for the ratings is very similar to the familiar bell curve of a normal distribution. Treating the HDRS ratings as a normally distributed continuous variable in our previous analyses thus seems reasonable.

4.4   Data-based graphs: Bivariate graphs

4.4.1 Box-and-whisker plot for two-level data

A box-and-whisker plot is useful for depicting the locality, spread and skewness of a data set. It also offers a useful way of comparing two or more data sets with each other with regard to locality, spread and skewness. To illustrate this feature, we use **reisby.ss3** (see Section 4.3 for a brief description of these data). In the plot shown below, we requested box-and-whisker plots for the HDRS ratings that served as the outcome variable in previous analyses (see Section 3.2) at each of the measurement occasions.

The bottom line of a box represents the first quartile (q_1), the top line the third quartile (q_3), and the in-between line the median (me). The arithmetic mean is represented by a diamond. Here, a decrease in the mean HDRS rating is observed over the course of the study. In addition, the larger distances between the extremes of the boxes at the later measurement occasions indicate more variability in HDRS ratings towards the end of the study.
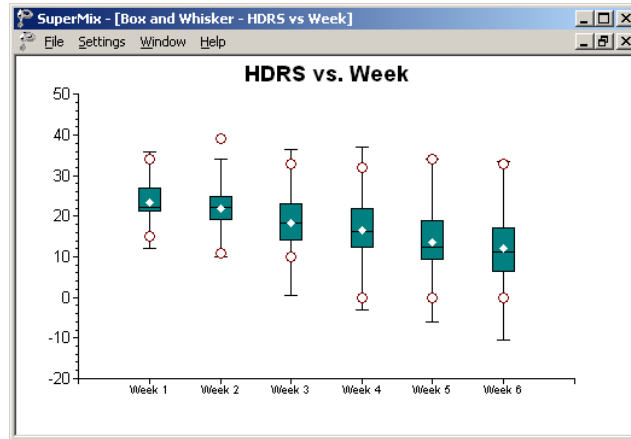
The whiskers of these boxes extend to $1.5(q_3 - q_1)$ on both sides of the median. The length of the whiskers is based on a criterion of Tukey (1977) for determining whether outliers are present. Any data point beyond the end of the whiskers is then considered an outlier. Two red circles are used to indicate the minimum and maximum values.

For symmetric distributions, the mean equals the median and the in-between line divides the box in two equal parts.

For a large sample from a normal distribution, the plotted minimum and maximum values should be close to the whiskers of the plot. This is not the case for the first box-and-whisker plot corresponding to the first measurement occasion. However, we conclude that for the remainder of the occasions the assumption of a symmetric distribution would be acceptable.

Creating a box-and-whisker chart

To create the box-and-whisker plot shown above, start by opening **Examples\Primer\Graphics\reisby.ss3**. Next, select the **Bivariate** option from the pop-up menu displayed when the **File**, **Data-based Graphs** option is selected from the main menu bar.
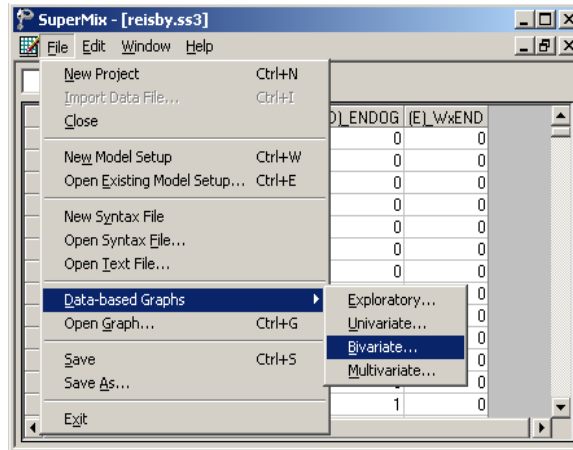
The **Bivariate Plot** dialog box is displayed. Select HDRS as the **Y** variable of interest, and Week as the **X** variable. Click the radio button next to the **Box and Whisker** option and then click **Plot** to display the box-and-whisker plot of HDRS at the measurement occasions.
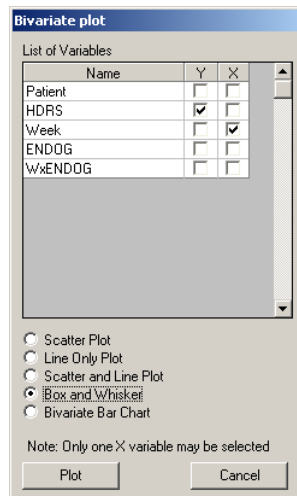
4.4.2 Box-and-whisker plot for three-level data

The data set used next (see Section 3.3 for a detailed description) forms part of the data library of the Medical Expenditure Panel Survey (MEPS). Collected in 1999, these data from a longitudinal national survey were used to obtain regional and national estimates of health care use and expenditure based on the health expenditures of a sample of US civilian non-institutionalized participants. The data is in the file **Examples\Primer\Graphics\meps.ss3**. A description of the variables TOTEXP99 and INSCOV is repeated below.

- o TOTEXP99 is the natural logarithm of the total health expenditure of a respondent in 1999, ranging between 0 and 12.24 and representing actual expenditure of between $1 and $206,721.
- o INSCOV is an indicator of the level of insurance coverage, where 0 indicates private coverage any time during 1999, and 1 indicates public coverage or no insurance at all during 1999.
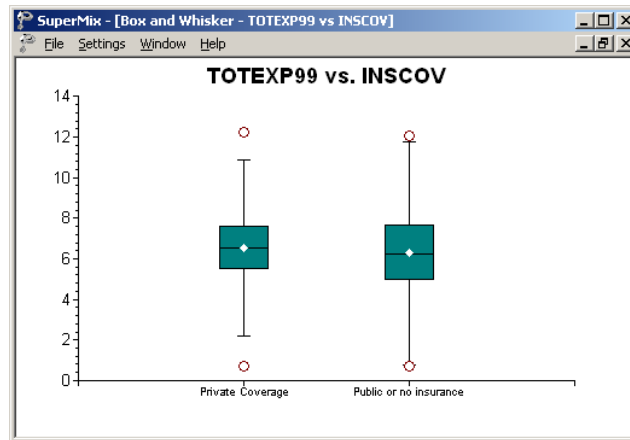
The plot shows that for both types of insurance coverage the distributions of TOTEXP99 are close to symmetrical. This follows from the fact that, for each category of INSCOV, the median and mean values are close to each other and the median line divides the box into equal parts. It is also evident that the mean (or

median) of TOTEXP99 is larger for the private coverage category. On the other hand, the variation in TOTEXP99 values for the public or no insurance categories is larger.

Creating a box-and-whisker chart

To create the box-and-whisker plot shown above, start by opening **meps.ss3**. From the **File**, **Data-based Graphs** menu, select the **Bivariate** option to activate the **Bivariate plot** dialog box. Select TOTEXP99 as the Y-variable and INSCOV as the X-variable. Click **Plot** to obtain the box-and-whisker plot shown above.
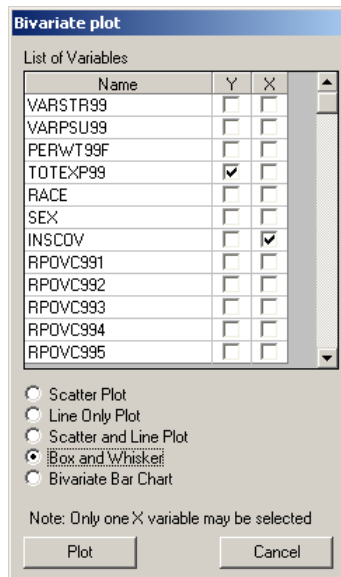
4.4.3 Scatter/line plot

A graph frequently used as part of initial exploratory analysis of data is the scatter and/or line plot. This type of plot is used to examine potential relationships between a continuous outcome variable and possible predictor variables. Scatter plots are particularly useful for the study of repeated measurements data.
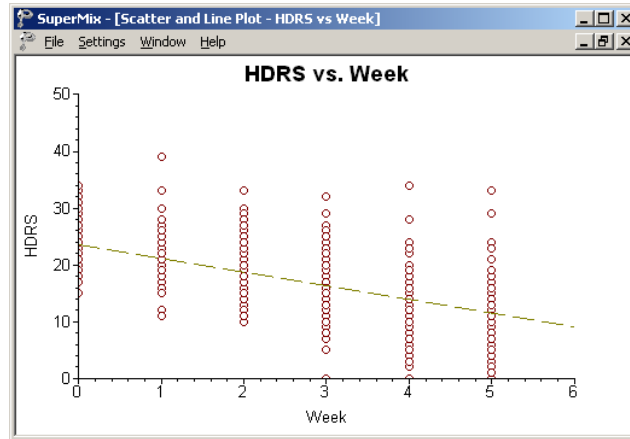
For the Reisby data, we looked at the patients' HDRS ratings at six time points. Previously, we plotted box-and-whisker plots of the HDRS ratings at the six measurement occasions. For these plots, all measurements at each of the six time points were summarized in the box-and-whisker plot for each of the occasions. A scatter/line plot allows us to also look at the trajectories of individual patients' HDRS ratings over the course of the study. The plot below shows these trajectories, with circles representing the actual measurements of the 66 patients (the scatter plot component) and a dashed line representing the fitted linear regression curve.

We note that, generally speaking, the ratings seem to decrease over the study period. Note that this graph corresponds to the discussion in Chapter 1 of this text, where data from the first 10 patients were used to illustrate the need to use a mixed-effects model for these data.

Creating a scatter/line plot

To create the scatter and line plot of HDRS ratings at the measurement occasions, open the **Examples\Primer\Graphics\reisby.ss3** spreadsheet, and select the **File**, **Data based Graphs**, **Bivariate** option from the main menu bar.
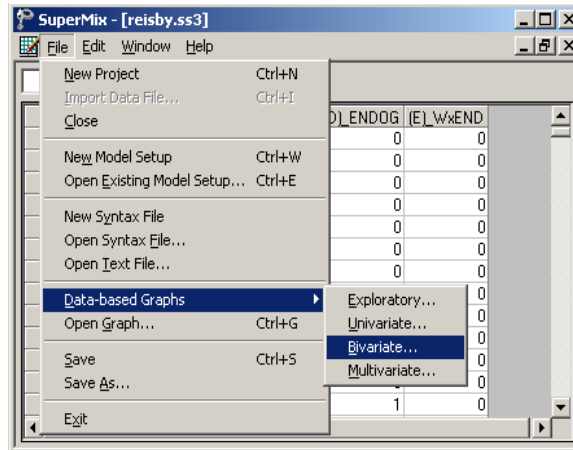
The **Bivariate Plot** dialog box is displayed. Select HDRS as the **Y** variable of interest, and Week as the **X** variable. Click the radio button next to the **Scatter and Line Plot** option and then click **Plot** to display the combined scatter and line plot of HDRS at the measurement occasions. Note that similar plots, displaying either the observed data or the average line, can be obtained by using the **Scatter Plot** or **Line Only Plot** options.
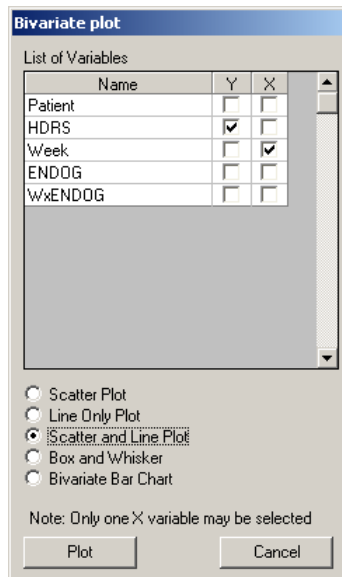
4.4.4 3D bar chart

The data from the McKinney Homeless Research Project study (see Section 3.7 and 4.3) were used for this example. Recall that the data file **sdhouse.ss3** contains subjects' housing status as recorded at 4 time points. From the pie chart, we noted, for example, that over the course of the study, subjects reported living in independent housing in 39.6% of the observations. In 22.8% of the measurements, subjects were living on the street.

A 3D bar chart allows us to graphically display a two-way frequency table. The 3D bar chart shown below is a visual display of the bivariate distribution of the variables HOUSING and TIME, where TIME represents the four measurement occasions. By including the TIME variable, we are essentially acknowledging the longitudinal nature of these data. The row of bars in the front of the graph, associated with the code "0" for HOUSING, represent the number of subjects who were living on the street at each of the four occasions. The second row of bars represent subjects living in community housing, and the third the subjects living in independent housing at the time of measurement. Over the study period, there was a marked decrease in the number of subjects living on the streets. The number of subjects living in community housing showed a decrease between the first two and the last two time points, while the subjects living in independent housing increased rapidly over time.
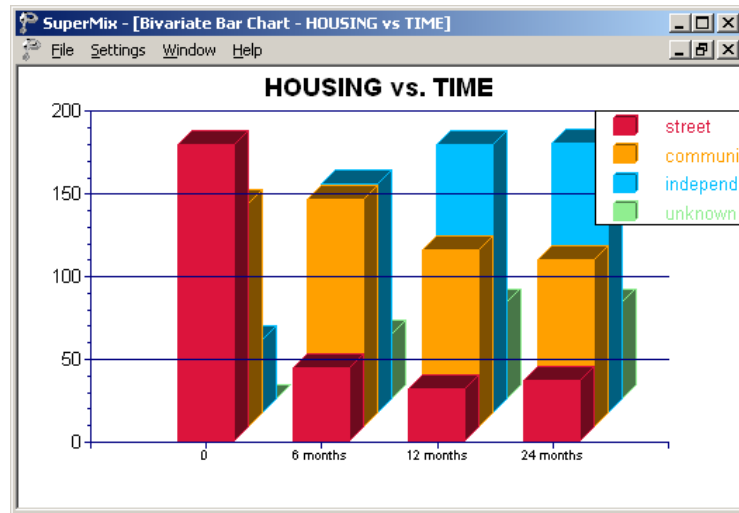
A key concern in the McKinney study was to evaluate the effectiveness of using Section 8 certificates to provide independent housing to the severely mentally ill homeless. The variable SECTION8, contained in the **sdhouse3.ss3** spreadsheet in the **Examples\Primer\Graphics** folder was used to distinguish between subjects with (SECTION8 = 1) and subjects without (SECTION8 = 0) Section 8 certificates. By portraying the variable **SECTION8** against the type of housing (HOUSING) in the form of a 3D bar chart, we can get some idea of the relationship between the type of housing reported and the use or not of a Section 8 certificate. From the 3D bar chart below, it seems as if subjects with Section 8 certificates were approximately twice as likely to report living in independent housing over the course of the study than those without, whom were more likely to report living in community housing. A slight decrease in the number living on the street is also observed in the case of reports by subjects with Section 8 certificates. Ideally, one would like to combine the information in these two 3D bar charts into one. One way to do so would be to prepare model-based graphs (see Section 4.6), based on an analysis where both the longitudinal nature of the data and the availability of Section 8 certificates are taken into account.
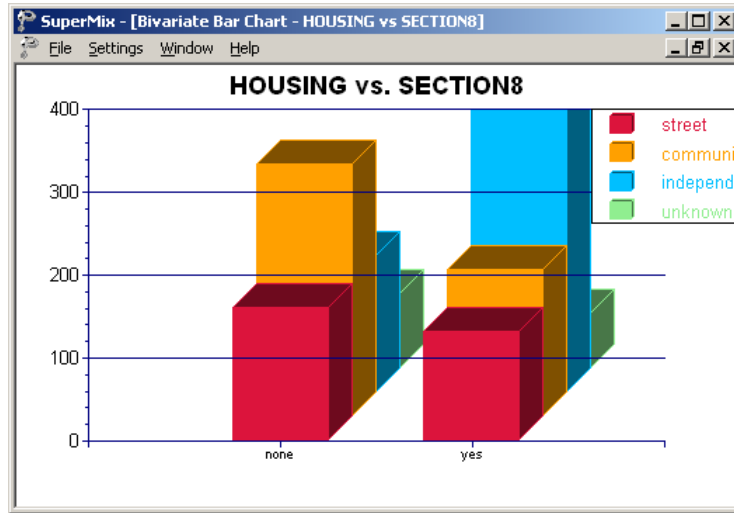
Creating a 3D bar chart

To create the 3D bar charts shown above, start by opening the data file **Examples\Primer\Graphics\sdhouse3.ss3**. Right-click on the column header of the variable TIME and select the **Column Properties** option from the pop-up menu to open the **Column Properties** dialog box. This dialog box is used to define the type of variable (nominal, ordinal or continuous) and to provide labels for the categories of nominal and ordinal variables. Indicate HOUSING as a nominal variable by clicking the appropriate radio button, and enter the labels (street, community, independent and unknown) for each category in the **Label** column. Do the same for the ordinal variable TIME as shown below. Click **OK** to return to the spreadsheet window and save the changes to the spreadsheet using the **File**, **Save** option.

Column Properties

Header: TIME

Number of Decimal Places: 0

⦿ Categorical ○ Continuous

☐ Create dummies when used as a predictor

	Value	Label
1	0	0
2	1	6 months
3	2	12 months
4	3	24 months

OK   Cancel

Select the **File**, **Data based Graphs**, **Bivariate** option from the main menu bar. The **Bivariate Plot** dialog box is now displayed. Select HOUSING as the **Y** variable of interest, and TIME as the **X** variable. To obtain a bivariate 3D bar chart of the housing status at the measurement occasions, click the radio button next to the **Bivariate Bar Chart** option and then click **Plot** to obtain the bivariate bar chart.

Bivariate plot

List of Variables

Name	Y	X
ID	☐	☐
HOUSING	☑	☐
SECTION8	☐	☐
TIME1	☐	☐
TIME2	☐	☐
TIME3	☐	☐
SECT8T1	☐	☐
SECT8T2	☐	☐
SECT8T3	☐	☐
NOSECT8	☐	☐
TIME	☐	☑
SEC8TIME	☐	☐

○ Scatter Plot
○ Line Only Plot
○ Scatter and Line Plot
○ Box and Whisker
⦿ Bivariate Bar Chart

Note: Only one X variable may be selected

Plot   Cancel

To obtain a bivariate bar chart of housing status and Section 8 certification, simply close the graphing window to return to the **Bivariate Plot** dialog box. Retain HOUSING as the **Y** variable, but uncheck TIME as the **X** variable and replace it with the variable SECTION8. Click the **Plot** button. The bivariate bar chart of HOUSING versus Section 8 certification, as shown earlier, is now displayed.

4.5   Data-based graphs: Multivariate graphs

4.5.1 Scatter Plot Matrix

The scatter plot matrix provides an organized way of simultaneously looking at bivariate plots of up to ten variables. Except for the diagonal, each frame contains a scatter plot. The diagonals contain the variable names and the minimum and maximum values of the corresponding variables. Each scatter plot is a visual summary of linearity, nonlinearity, and separated points (see, for example, Cook & Weisberg (1994)).

Two of the many factors that are known to have some influence or relevance to the condition of the human heart are physical fitness and blood cholesterol level. In a related research project, four homogeneous groups of adult males were considered. A number of plasma lipid parameters were measured on each of the 66 individuals and fitness parameters were also measured for three of the four groups. The four groups of patients are:

- o Weightlifters ($n = 17$)
- o Students (control group, $n = 20$)
- o Marathon athletes ($n = 20$)
- o Coronary patients ($n = 9$).

The triglyceride level is represented by the variable Trigl in these data, and Age represents the age of a patient in years. The cholesterol level is represented by the variable Cholest. The group identification of patients is given by the variable Group.

In the scatter plot matrix shown below, the first row of plots at the top of the matrix contains plots of Cholest against Age and Trigl respectively. The third plot in this row contains the variable name and the observed minimum and maximum values of Cholest. The first plot in the second row shows the scatter plot of Trigl against Age. The other scatter plots are essentially mirror images of the three discussed this far and are the plots for (Cholest, Trigl), (Trigl, Age) and (Cholest, Age) respectively. The diagonal elements of the matrix contain the names and the minimum and maximum values of each variable.

From the display, it is apparent that the (Age, Trigl) and (Age, Cholest) scatter plots of coronary patients, denoted by a "+" symbol, are clustered together, away from the main cluster of points formed by the other three groups. In the (Age, Age) segment the minimum and maximum values of the Age variable are given. In general, the coronary patients are older and have higher cholesterol and triglyceride levels than patients in the remaining groups.
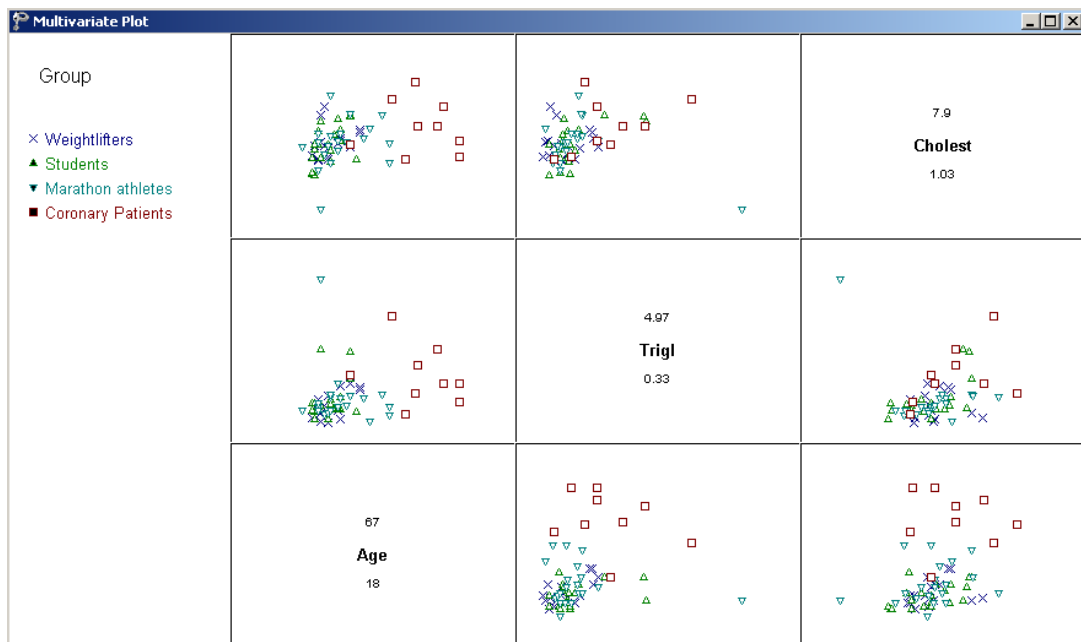
Creating a scatter plot matrix

Select **fitchol.ss3** from the **Examples\Primer\Graphics** folder. A scatter plot matrix is obtained by selecting the **Data-based Graphs**, **Multivariate** option from the **File** menu to display the **Multivariate plot** dialog box shown below.
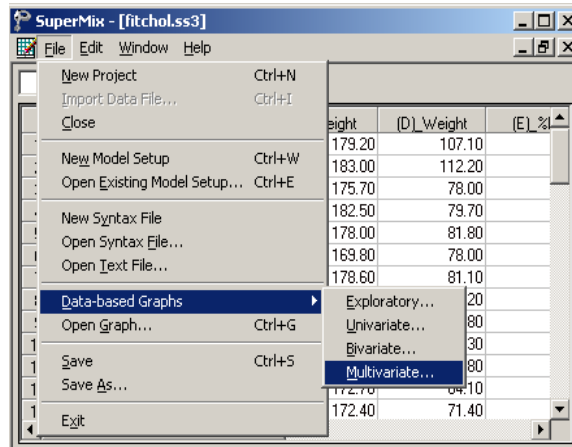
The **Multivariate plot** dialog box is displayed. Select Age, Trigl and Cholest as the **Plot** variables and select Group as the **Mark** variable as shown below. Complete the dialog box as shown and click Plot when done to obtain the graph as described earlier.
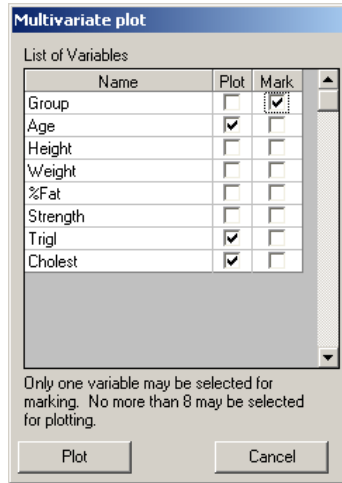
4.6 Model-based graphs

Three types of model-based graphs can be produced with **SuperMix**. Plots of predicted outcomes based on the model equations can be made, and these graphs can be displayed by group or marked by a third variable. Residual plots (normal outcomes only), again marked by an additional variable, can also be made. Finally, an option to plot model-based confidence intervals of random coefficients is available. Again, these plots can be displayed by group or by marking variable.

Model-based graphing options are accessed via the **File**, **Model-based Graphs** option. The pop-up menu activated through this selection has three options – **Equations**, **Residuals** and **Confidence Intervals** – each associated with one of the type of plots described above.
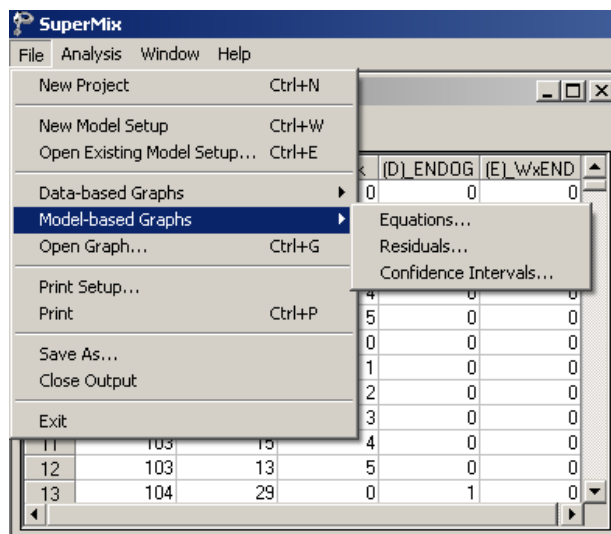
Graph parameters that may be changed include the axes and descriptions thereof, the symbols used, and the colors assigned to the symbols/text. To change any of these, simply double-click on the symbol/text to be changed to activate a dialog box in which changes can be made. Use of such dialog boxes will be illustrated in the course of the discussion of the various graphs in the sections to follow. A summary of the graph editing tools is given in Section 4.7.

4.6.1 Graphing model equations

A graph frequently used as part of initial exploratory analysis of data is the scatter and/or line plot. This type of plot is used to examine potential relationships between an outcome variable and possible predictor variables, using the observed data. After fitting a model, one can plot the relationship between the outcome variable and predictor variables, subject to the model fitted to the data.

Creating an equation based graph for a two-level model

Recall that, for the Reisby data, we looked at the patients' HDRS ratings at six time points. In Section 3.2, we fitted a model with random intercept and slope for the response variable HDRS with the fixed predictors Week, ENDOG and the interaction effect between Week and ENDOG to the data. In this model

$$\text{HDRS}_{ij} = \beta_0 + \beta_1 * (\text{Week})_{ij} + \beta_2 * (\text{ENDOG})_i + \beta_3 * (\text{WxENDOG})_{ij} + u_{i0} + u_{i1} (\text{Week})_{ij} + e_{ij}$$

β_0 denotes the average expected depression rating scale value, β_1 and β_2 denote the coefficients of the predictor variables Week and ENDOG in the fixed part of the model, β_3 denotes the coefficient of the interaction between Week and ENDOG in the fixed part of the model, u_{i1} denotes the variation in the Week slopes over patients, and u_{i0} and e_{ij} denote the variation in the average expected HDRS value over patients and over measurements (*i. e.*, between patients) respectively.

The graph above shows the predicted HDRS ratings over the study period for the two groups of patients. The average predicted HDRS rating of patient with endogenous depression is consistently higher than that of the patients with non-endogenous depression.

This trend holds for most of the patients, as illustrated in the graph below, where the predicted HDRS ratings of individual patients were plotted. In this graph, the predictors not used in the graph (i.e., ENDOG and WxENDOG) were held at the mean values, in contrast to the graph shown above where WxENDOG was held constant at a value of 0.
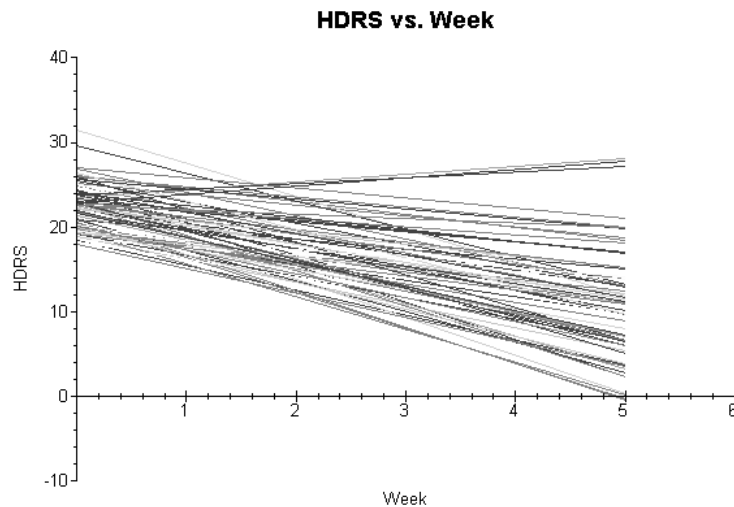
Care should be taken when making these graphs to verify that, if the default value of 0 is assumed for predictors not included, the value 0 is within the range of possible predictor values. In our case, the variables ENDOG and WxENDOG both include observed values/categories equal to zero, and the choice between holding the predictors constant at zero or the mean is a question of personal preference. However, should the variable ENDOG for example be recoded to have values 1 and 2 for the depression categories rather than the current 0,1 coding, interpretation of

the graph shown above would be problematic. The variables ENDOG and ENDOG would then both be held constant at nonexistent values in terms of their ranges of observation/coding.

Creating an equation based graph

To create the model-based graph of HDRS ratings over the study period for the two groups of patients shown previously, open both the data and model files in the **Graphics** folder, in this case **reisby.ss3** and **depress.mum**, and run the model. Next, select the **File**, **Model-based Graphs**, **Equations** option from the main menu bar. The **Plot Equations for: HDRS** dialog box is displayed. Note that it is assumed that the outcome variable used in the model, in this case HDRS, is the variable that will be displayed on the vertical axis of the graph.
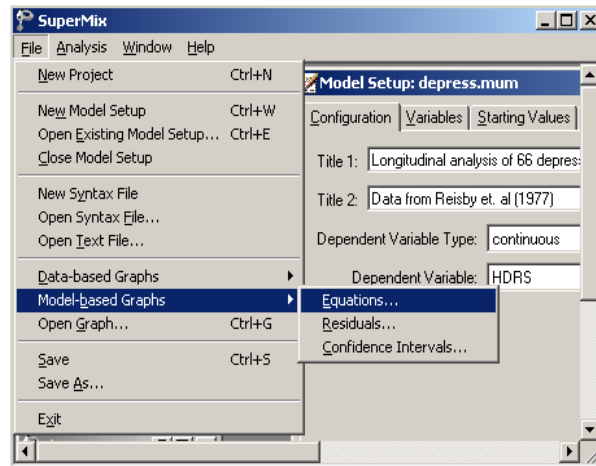
Select Week as the predictor variable that will be displayed on the X-axis of the graph by checking the box next to this variable in the **Predictor** column. To obtain separate lines for patients with and without endogenous depression, select the variable ENDOG in the **Mark** column. Note that only one grouping variable and one marking variable may be selected.

Plot Equations for Outcome Variable

List of Variables

Name	Predictor	Group	Mark
intcept	☐	☐	☐
Week	☑	☐	☐
ENDOG	☐	☐	☑
WxENDOG	☐	☐	☐
Patient		☐	☐

○ Remaining predictors fixed at 0
○ Remaining predictors fixed at their means

Note: Only one variable may be selected for grouping and only one for marking.

Plot   Cancel

By default, predictors not selected will be held constant at zero, as indicated by the default selection of the **Remaining predictors fixed at 0** option. Click **Plot** to display the graph shown below.

Note that the two lines shown in this graph correspond to the model

$$\widehat{\text{HDRS}}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 * (\text{Week})_{ij} + \hat{\beta}_2 * (\text{ENDOG})_i + \hat{\beta}_3 * (\text{WxENDOG})_{ij}$$
$$= \hat{\beta}_0 + \hat{\beta}_1 * (\text{Week})_{ij} + \hat{\beta}_2 * (\text{ENDOG})$$

as WxENDOG is held constant to zero. This variable can assume a value of 0 when a patient exhibits non-endogenous depression (ENDOG = 0), or alternatively at the beginning of the study period when WEEK = 0. Consequently, the graph reflects the predicted HDRS ratings over time for patients meeting these requirements – in effect a subset of all the patients.
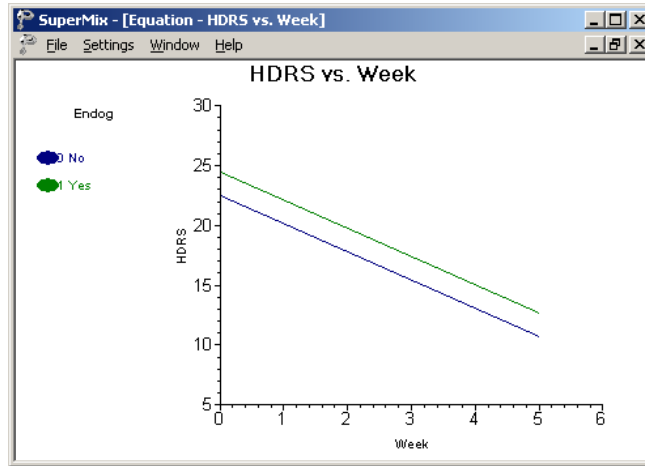
If, instead of holding the predictor WxENDOG equal to zero, this variable is held constant at its mean of 1.745, the same plot would usually look different.

For this case, the two lines shown in this graph correspond to the model

$$\widehat{\text{HDRS}}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 * \left(\text{Week}\right)_{ij} + \hat{\beta}_2 * \left(\text{ENDOG}\right)_i + \hat{\beta}_3 * \left(\text{WxENDOG}\right)_{ij}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 * \left(\text{Week}\right)_{ij} + \hat{\beta}_2 * \left(\text{ENDOG}\right) + \hat{\beta}_3 (1.745)$$

and thus corresponds to the predicted outcome assuming a mean value for the interaction term.

From Section 3.2, we have $\hat{\beta}_3 = -0.02$, and therefore $\hat{\beta}_3 (1.745) = -0.035$. The difference between this value and zero is too small to see the effect on the graphical display below, which appears to be identical to the "remaining predictors fixed at 0" selection. However, it gives support to the finding of a non-significant interaction effect.
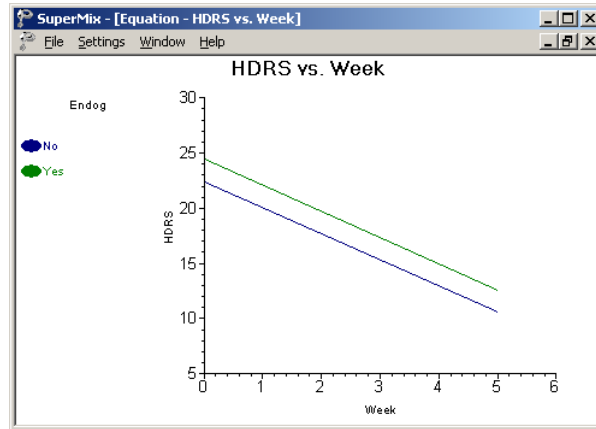
The graph shown above distinguishes between the two depression groups by using blue and green lines (green for "yes", blue for "no"). For publication purposes, the difference may be clearer if both lines are shown in black, but with different line styles. To change the graph, double-click on the lower of the two lines to activate the **Plot Parameters** dialog box. Click the **Line Attributes** button to display the **Line Parameters** dialog box and set the **Color** to black, the style to "…." and the width to 1 as shown below. Click **OK** on both dialog boxes to display the modified graph.
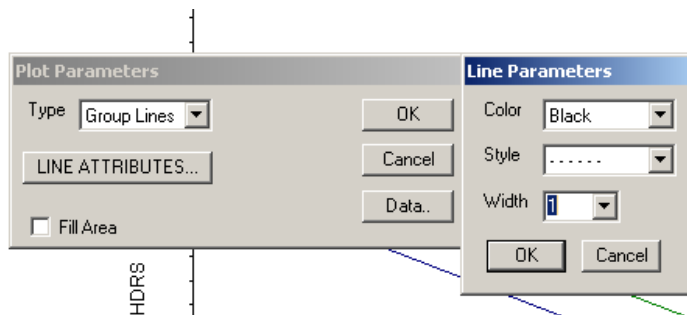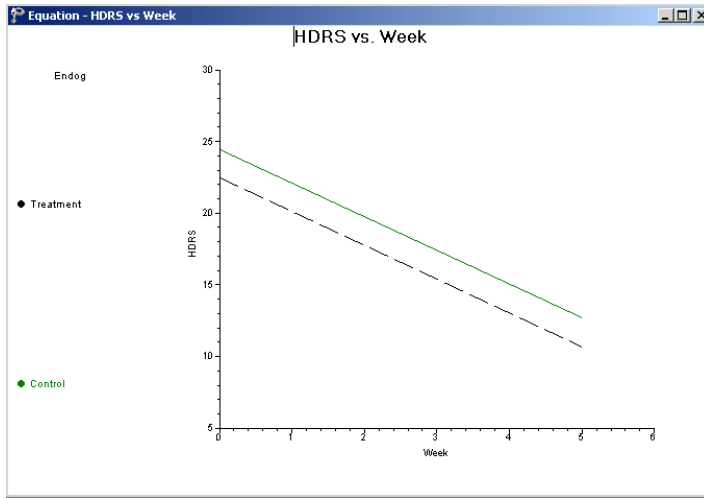
To create a graph displaying the predicted outcomes over time by individual patient, close the graphing window to return to the **Plot Equations for: HDRS** dialog box. Select the variable Patient as marking variable. Ensure that the **Remaining predictors fixed at their means** option is enabled. Click **Plot** to display this graph, as shown below.

Apart from a few patients for whom an increase in the HDRS ratings is predicted over the course of the study, the general trend is for HDRS to decrease over time.
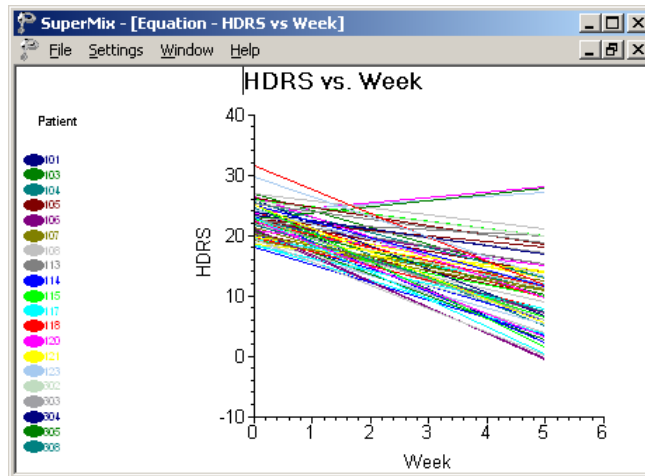
Creating an equation based graph for a three-level model

In Section 3.3 a model using the participant's gender, ethnicity, type of health insurance coverage, and measure of income relative to poverty level was fitted to predict the total expenditure on health care in 1999, with expenditure transformed to the natural logarithm of the actual expenses incurred. The model fitted was of the form

$$\text{TOTEXP99}_{ijk} = \beta_0 + \beta_1 * \text{SEX}_{ijk} + \beta_2 * \text{RACE}_{ijk} + \beta_3 * \text{INSCOV}_{ijk} +$$
$$\beta_4 * \text{RPOVC991}_{ijk} + \beta_5 * \text{RPOVC992}_{ijk} + \beta_6 * \text{RPOVC993}_{ijk} +$$
$$\beta_7 * \text{RPOVC994}_{ijk} + v_{i0} + u_{ij0} + e_{ijk}$$

For this model, the output shown below was obtained.

```
SuperMix - [meps.out]
 File  Analysis  Window  Help

Fixed regressor(s)
------------------

  Variable          Estimate     Std.Err.     Z-value     p-value
  --------          --------     --------     -------     -------
  intcept            5.94083      0.05326    111.54352     0.00000
  RACE               0.33446      0.03303     10.12702     0.00000
  SEX                0.34691      0.02359     14.70287     0.00000
  INSCOV            -0.05976      0.03108     -1.92277     0.05451
  RPOVC991           0.25764      0.04320      5.96385     0.00000
  RPOVC992           0.01777      0.04551      0.39037     0.69626
  RPOVC993           0.07608      0.04175      1.82223     0.06842
  RPOVC994          -0.06244      0.06211     -1.00538     0.31471

  Log Likelihood              =   -36869.7492
  -2 Log Likelihood (Deviance) =    73739.4984
  Akaike's Information Criterion =  73761.4984
  Schwarz's Bayesian Criterion =    73794.0897
  Number of free parameters   =           11

 Save As...      Close
```

Recall that a value of 1 for the gender indicator variable SEX indicated that a participant was female, with a value of 0 assigned to male participants. Female participants are expected to have a total health expenditure 0.93063 higher than male participants do if all other variables are held constant. In contrast, participants with public coverage or no coverage have a lower expected total expenditure, as indicated by the negative estimated coefficient –0.61785 for the predictor INSCOV.

Using the **Equations** option on the **File**, **Model-based Graphs** menu, the plot shown below was obtained. This graph shows the predicted total expenditure on health care for the two insurance coverage groups, marked by gender. As indicated by the output shown above, the total expenditure for females is appreciably higher than for males, and respondents with private insurance coverage tend to spend less than those without coverage or with public insurance coverage.
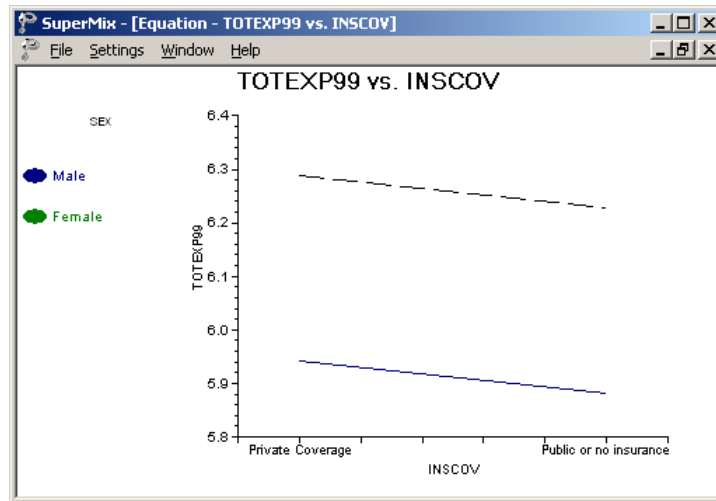
Creating an equation based graph

To create the graph shown above, start by opening both the data and model files, in this case **Examples\Primer\Graphics\meps.ss3** and **meps.mum**. After running the analysis, select the **File**, **Model-based Graphs**, **Equations** option from the main menu bar. The **Plot Equations for: TOTEXP99** dialog box is displayed. Select the predictor INSCOV for display on the horizontal axis by checking the box next to the variable name in the **Predictor** column, and request separate plots for the gender groups by selecting SEX in the **Mark** column. Click **OK**.

The graph shown below is now displayed. For publication purposes, the distinction between the two lines shown may be emphasized by changing the style of one of the lines. To change the attributes of the line for female respondents, *i.e.* the higher of the two lines, double-click on this line to activate the **Plot Parameters** dialog box.
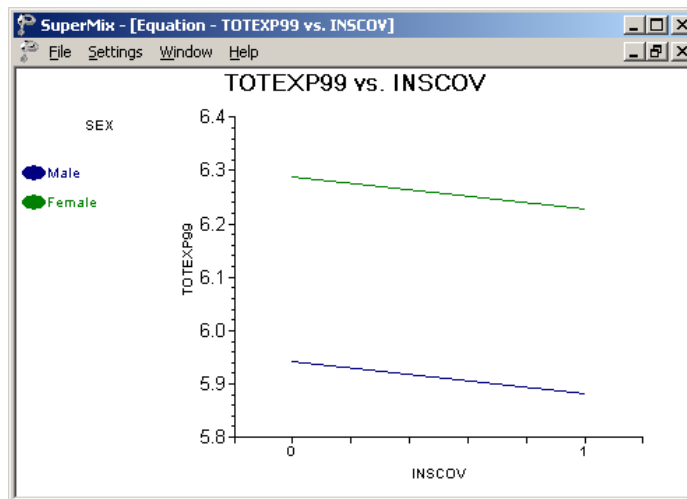
Click the **Line Attributes** button to open the **Line Parameters** dialog box. Change the **Color** of the line to black, and select a dotted line from the **Style** list box. Adjust the width of the line to 1 in the **Width** list box, and click **OK** on both the **Line Parameters** and **Plot Parameters** dialog boxes to obtain the revised graph.
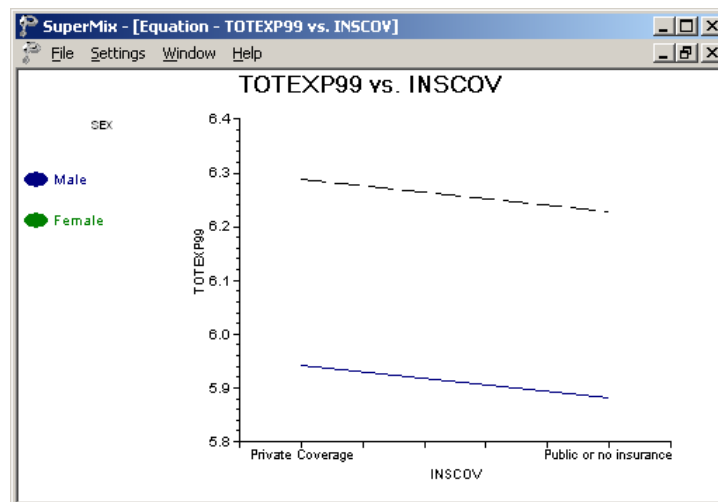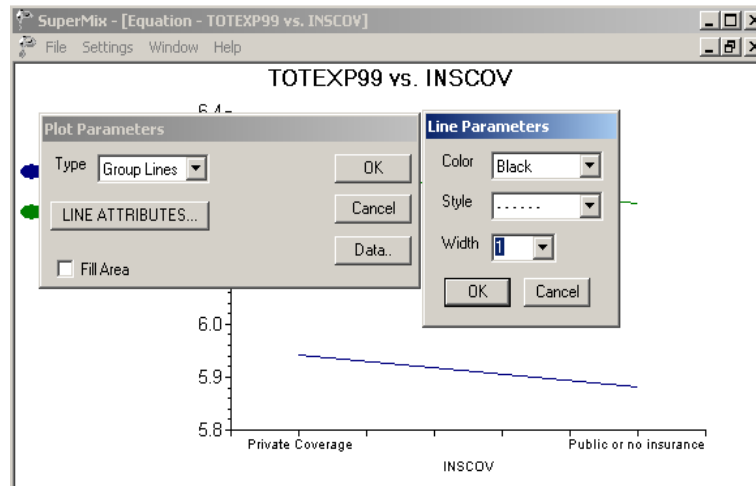
4.6.2 Residual plots

The **Residuals** option on the **File**, **Model-based Graphs** menu is used to examine the residuals obtained for a fitted model. It is useful for examining the fit of the model, and also as a check for possible distributional assumption violations. As residuals are defined as the difference between the observed and predicted outcome, trends in residuals, for example over the course of a study in a longitudinal data set, may indicate that an important predictor was not included in the model fitted to the data. To request the creation of a residual file, set the **Write Bayes Estimates** field on the **Configuration** tab to **yes**.
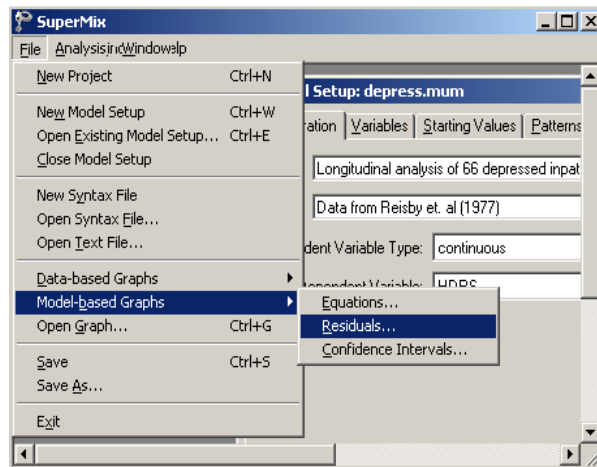
The image below shows the residuals for a model fitted to the Reisby data (see Section 3.2), in which the time of measurement, depression status, and interaction between the depression status and time of measurement, were used to predict the HDRS ratings of patients. The model fitted to the data was of the form

$$\text{HDRS}_{ij} = \beta_0 + \beta_1 * \left(\text{Week}\right)_{ij} + \beta_2 * \left(\text{ENDOG}\right)_i + \beta_3 * \left(\text{WxENDOG}\right)_{ij} + u_{i0} + u_{i1}\left(\text{Week}\right)_{ij} + e_{ij}$$

.

If a perfect fit is obtained, the residuals would all be equal to zero. In practice, this is not the case, but clustering of the residuals around 0 is usually observed. Using the 0 tick mark on the vertical axis as guideline, we observe that approximately half of the residuals are above this mark, and half below. It is interesting to note that the largest positive residuals occur for measurements associated with patients with endogenous depression (indicated by circles in the graph), and that most of the largest negative residuals are those for patients with non-endogenous depression. This indicates that, under the fitted model, some of the predicted measurements for endogenously depressed patients were smaller than the observed measurements, while the opposite is true for non-endogenously depressed patients. This occurs for a small percentage of the measurements, most of which are considerably closer to zero and tightly clustered. As such, we do not suspect systematic deviations from the fitted model over the range of the outcome variable.
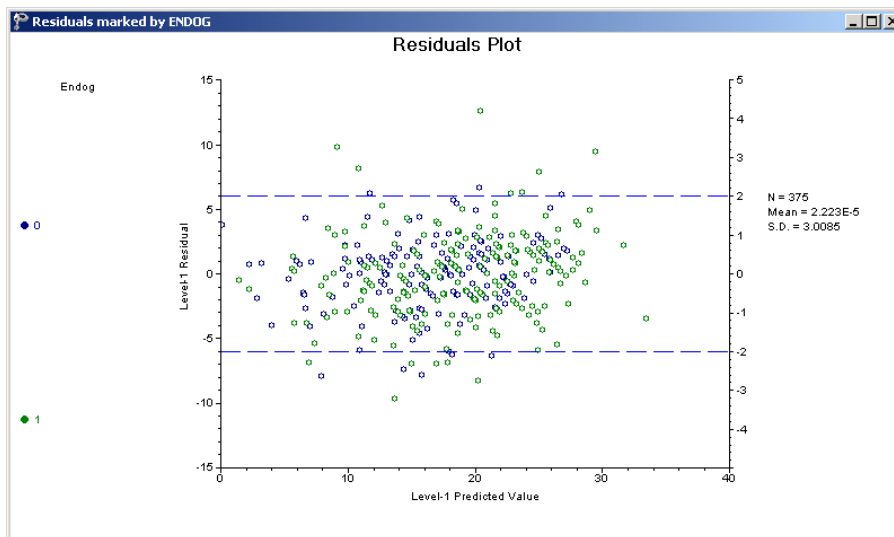
Creating a residual plot

To create the model-based graph of HDRS rating residuals shown above, start by opening the data and model files for the Reisby data **Examples\Primer\Graphics\reisby.ss3, depress.mum**) and run the analysis. Next,

select the **Residuals** option on the **File**, **Model-based Graphs** menu to open the **Plot of Residuals** dialog box.

The **Plot of Residuals** dialog box offers the option to obtain either standardized or unstandardized residual plots. By default, a standardized plot of the residuals will be displayed. All the variables included in the model specification, including the level-2 identification variable Patient, are shown in the **List of Variables**. Any single variable on this list can be used as marking variable, to request different legends for its categories. Check the box for ENDOG, the dichotomous variable indicating the depression classification of a patient and click the **Plot** button.
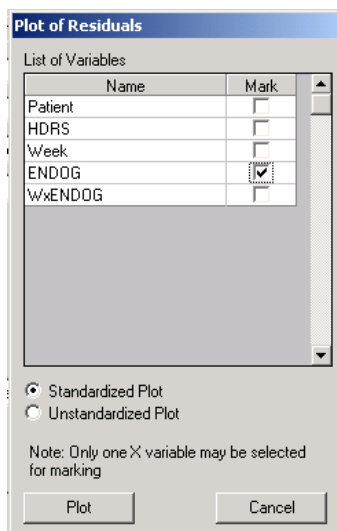
The plot obtained is the one shown above. By default, the same plotting symbol is used for both categories of the marking variable, but is shown in different colors (blue and green) for the two groups. If the plot is to be used in a report, typically produced in black and white format, it may be advisable to change the plotting symbols to make it easier to distinguish between the groups. To do so, double-click on any of the green circles towards the top of the graph.
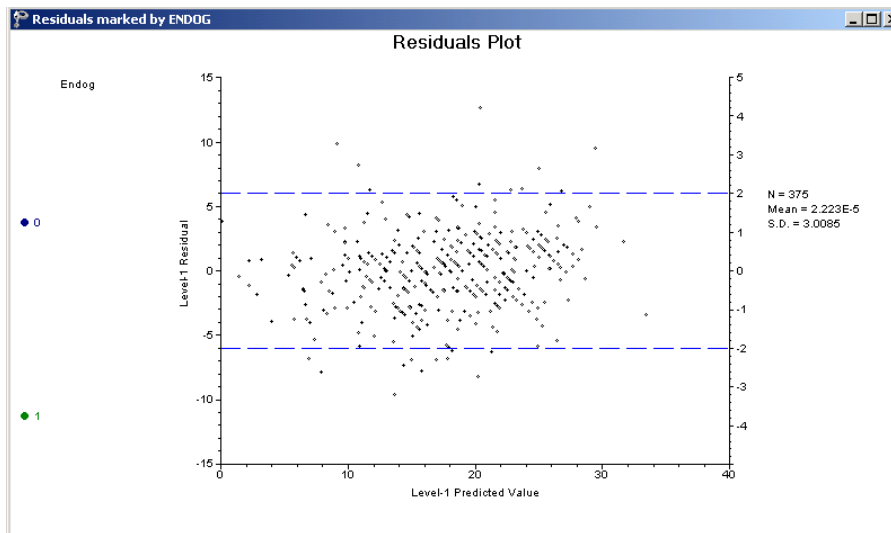
The **Plot Parameters** dialog box is activated. While retaining a circle as the **Shape** of choice for this group, change the **Color** to black, and reduce the size of the plotting symbol to 3 by selecting this option from the **Size** list box as shown below. Click **OK** to return to the graphing window.

Next, click on any of the blue plot symbols at the bottom of the graph. The **Plot Parameters** dialog box is again activated, but now shows the details of the symbols used for the second category of the marking variable ENDOG. Set the **Shape** of the symbol to a plus sign instead of a circle, and use the **Color** and **Size** list boxes to request the display of these in black and of size 3. Click **OK**. The plot shown previously is obtained.
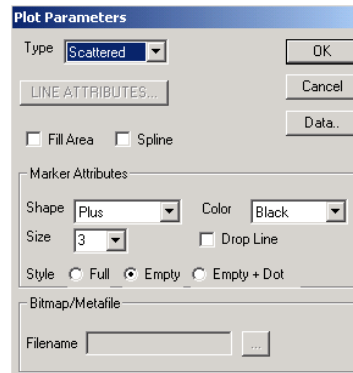
4.6.3 Confidence interval plots

The **Confidence Intervals** option on the **File**, **Model-based Graphs** menu provides the option to display confidence intervals for the empirical Bayes estimates of the random effects specified in a given model.
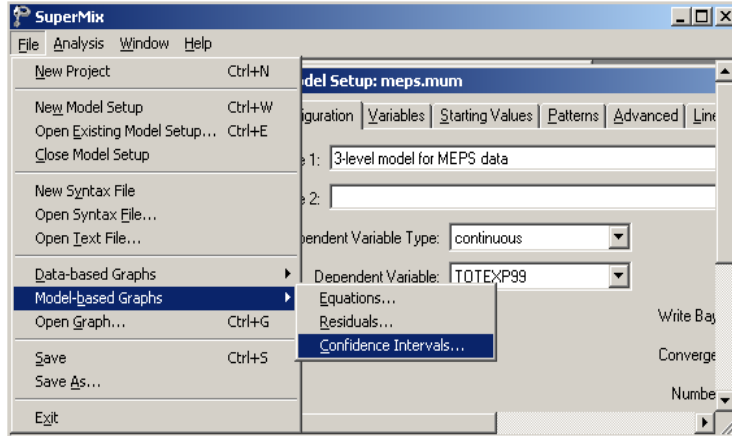
Using the three-level model discussed in Section 3.3, graphs showing the confidence intervals for the empirical Bayes estimates of the random intercepts for level-2 and level-3 units may be obtained.

Recall that the model fitted to the MEPS data, formulated as

$$
\begin{aligned}
\text{TOTEXP99}_{ijk} &= \beta_0 + \beta_1 * \text{SEX}_{ijk} + \beta_2 * \text{RACE}_{ijk} + \beta_3 * \text{INSCOV}_{ijk} + \\
&\quad \beta_4 * \text{RPOVC991}_{ijk} + \beta_5 * \text{RPOVC992}_{ijk} + \beta_6 * \text{RPOVC993}_{ijk} + \\
&\quad \beta_7 * \text{RPOVC994}_{ijk} + v_{i0} + u_{ij0} + e_{ijk},
\end{aligned}
$$

included a random intercept coefficient at both the PSU and stratum levels of the hierarchy (levels 2 and 3 respectively). Under this model, the expected total expenditure on health for a typical male respondent (SEX = 0) with no insurance coverage (INSCOV = 0) from the "negative or poor" income level (RPOVC991 = RPOVC992 = RPOVC993 = RPOVC994 = 0) can be calculated as

$$
\begin{aligned}
\hat{TOTEXP}99_{ijk} &= \hat{\beta}_0 + \hat{\beta}_1 * 0 + \hat{\beta}_2 * \text{RACE}_{ijk} + \hat{\beta}_3 * 0 + \hat{\beta}_4 * 0 + \hat{\beta}_5 * 0 + \hat{\beta}_6 * 0 + \hat{\beta}_7 * 0 \\
&= \hat{\beta}_0 + \hat{\beta}_2 * \text{RACE}_{ijk}
\end{aligned}
$$

Using the estimates of β_0 and β_2 of 4.5841 and 0.68364 respectively as obtained under this model, a white respondent (RACE = 1) was expected to have an expenditure 0.68364 logarithmic units higher than a similar respondent with different ethnic background. While this result is of interest in own right, it does not provide insight in the extent to which respondents with this background differ in terms of where they are from – that is, in terms of the PSU and stratum they belong to. These characteristics are key to the survey design, and the formulated model makes provision for intercepts to differ from stratum to stratum at the highest level, and from PSU to PSU at level 2 of the model.

To take the structure of the model into account and obtain unique estimates of the predicted total expenditure for each stratum, the extent to which the intercept of a stratum deviates from the average must be taken into account. This unique stratum contribution is represented by the random coefficient v_{i0} in the model. Estimates of these coefficients for the strata are referred to as the empirical Bayes residuals for the random level-3 intercepts. When the estimate of the v_{i0} under the model is added to the expected outcome for a typical respondent, the empirical Bayes estimate of the total expenditure specific to a stratum is obtained. Using the empirical Bayes estimates for the strata and their corresponding variances, we can plot confidence intervals for the random intercept of each stratum. A plot showing the 95% confidence intervals for the strata intercepts is shown below. Each confidence interval is obtained using

$$Empirical\ Bayes\ residuals \pm 1.96\sqrt{var\left(Empirical\ Bayes\ residuals\right)}\,.$$

95% Confidence Intervals
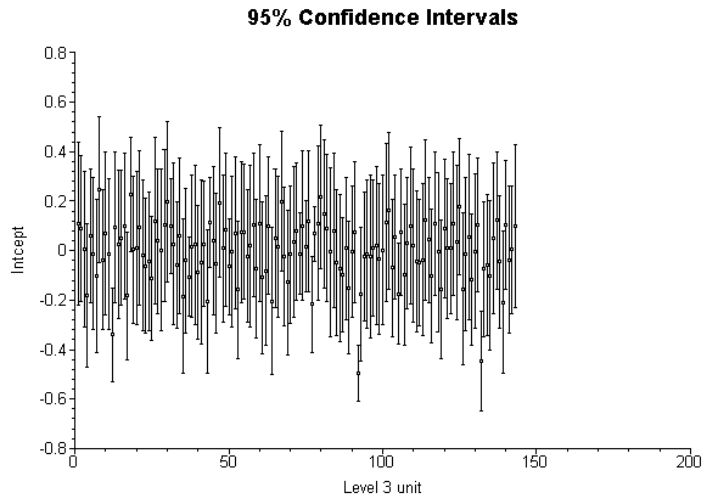
A similar graph can be obtained for the level-2 units, defined by the PSUs in this example.

95% Confidence Intervals
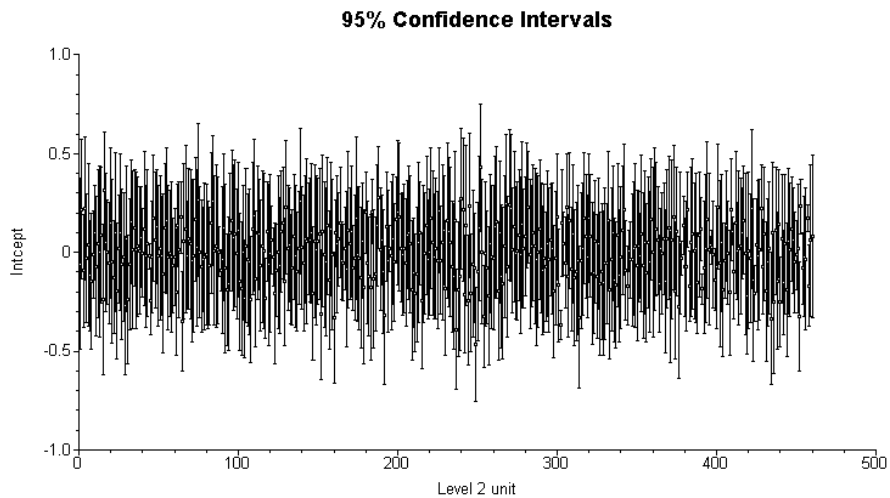
Creating confidence interval plots

Start by opening the data and model specification files for the MEPS data, named **meps.ss3** and **meps.mum** respectively. After running the analysis, select the **Confidence Intervals** option from the **File**, **Model-based Graphs** menu to open the **95% Confidence Intervals for EB estimates** dialog box. Check the box associated with **VARSTR99 Intcept** in the **Predictor** column. Note that this box also allows the selection of grouping and marking variables to be used in the graphical display. Click **Plot** to display the graph.
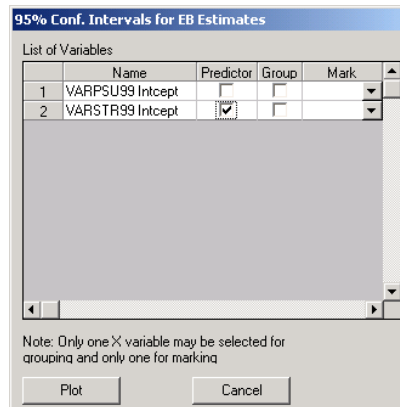
To obtain a similar graph for the empirical Bayes intercepts of level-2 units, close the graph to return to the **95% Confidence Intervals for EB estimates** dialog box. Deselect **VARSTR99 Intcept** in the **Predictor** column and select **VARPSU99 Intcept** instead as shown below. Click **Plot** to display the graph for level-2 units shown earlier. Note that one may check both the **VARSTR99** and **VARPSU99 Intcept** boxes to obtain the two graphical displays simultaneously.

95% Conf. Intervals for EB Estimates

List of Variables

	Name	Predictor	Group	Mark	
1	VARPSU99 Intcept	☑	☐	▼	
2	VARSTR99 Intcept	☐	☐	▼	

Note: Only one X variable may be selected for grouping and only one for marking

[Plot]   [Cancel]

4.7   Graph editing tools

4.7.1   Graph Parameters dialog box

The **Settings, Edit Graph** option, available for only for data-based and model-based graphs, is used to change attributes of the graph displayed and is associated with the **Graph Parameters** dialog box. This dialog box is used to change the position, size, and color of the currently selected graph and it's plotting area.
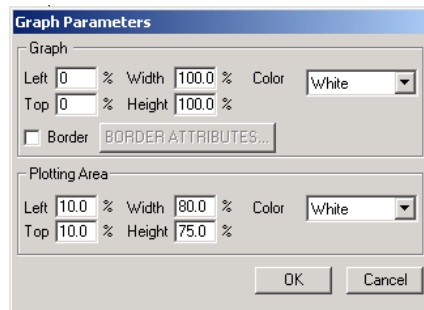
The following functions are available in this box:

- o The **Left**, **Top**, **Width**, and **Height** edit controls allow you to specify a new position and size of the graph (relative to the page window) and of the plotting area (relative to the graph window).
- o The **Color** list boxes are used to specify the graph window color and the color of the graph's plotting area.
- o If the **Border** check box is checked, the graph will have a border around it.
- o If the **Border** check box is checked, the **Border Attributes** button leads to another standard dialog box (the **Line Parameters** dialog box) that allows specification of the thickness, color, and style of the borderline.

In addition to the **Graphs Parameters** dialog box, a number of other dialog boxes may be used to change attributes of graphs. The dialog boxes accessible depend on the type of graph displayed.

The dialog boxes are:

- o **Axis Labels** dialog box
- o **Axes** dialog boxes
- o **Bar Graph Parameters** dialog box
- o **Legend Parameters** dialog box
- o **Line Parameters** dialog box
- o **Text Parameters** dialog box
- o **Plot Parameters** dialog box
- o **Pie Chart Parameters** dialog box
- o **Pie Slice Parameters** dialog box

The user may access any of these dialog boxes by double-clicking in the corresponding section of the graph. For example, double-clicking in the legend area of the graph will activate the **Legend Parameters** dialog box. Double-clicking on the

title of the graph, on the other hand, will provide access to the **Text Parameters** dialog box.

4.7.2 Axis Labels dialog box

This dialog box is used for editing axis labels and is activated by double clicking on the axis of a displayed graph.
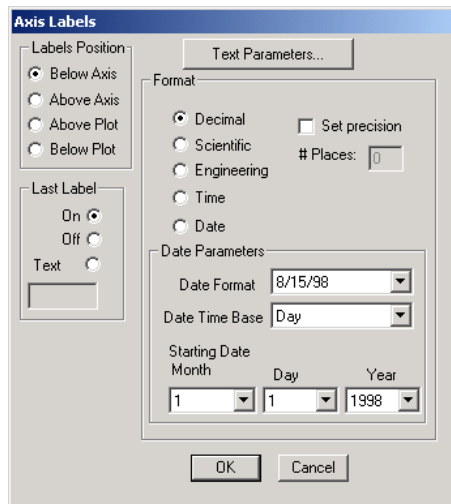
The following functions are controlled in this box:

- o The **Labels Position** group box controls the position of the labels relative to the axis or plotting area.
- o The **Last Label** group box allows manipulation of the last label drawing options. If **On** is selected, the last label is displayed like the others. If **Off** is selected, it is not displayed. If **Text** is selected, the text string entered in the edit box below will be displayed instead of the last numerical label.
- o The format of the numerical labels can be specified using the radio buttons in the **Format** group box.

- o The **Date Parameters** group box becomes active once the **Date** radio button is checked. The **Date Format** box selects the date format to use for labels, while the **Date Time Base** box selects the time base (minute, hour, day, week, month, year) for the date calculations. The **Starting Date** list boxes specify the starting date that corresponds to the axis value of 0. All dates are calculated relative to this value.
- o If the **Set Precision** check box is not checked, the labels' precision is determined automatically. If it is checked, the number entered into the **#Places** field specifies the number of digits after the decimal point.
- o The **Text Parameters** button provides access to the **Text Parameters** dialog box that controls the font, size, and color of labels.

4.7.3 Horizontal Axis dialog box

This dialog box is used for editing axis labels and is activated by double clicking on the axis of a displayed graph.
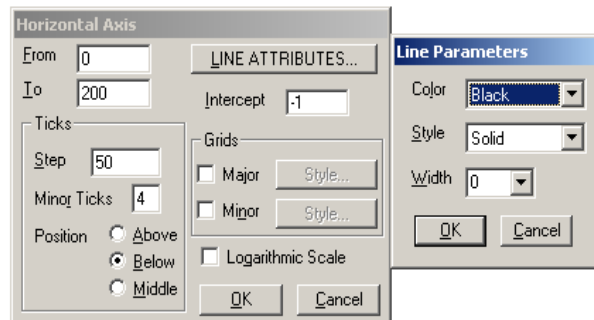
Its features include:

- o The edit controls **From** and **To** allow you to enter minimal and maximal values, in physical units, corresponding to the beginning and the end of the axis.
- o The **Intercept** edit control allows you to enter the new axis intercept value for the horizontal axis.

GRAPHICAL DISPLAYS

o The **Line Attributes** button leads to another standard dialog (see the **Line Parameters** section in this chapter) that allow you to specify the axis line thickness, color, and style.

o Controls in the group box **Ticks** allow you to enter tick mark parameters: the distance between adjacent major tick marks (**Step**), number of minor tick marks between two major ones (**Minor Ticks**), and position of the tick marks relative to the axis.

o Controls in the group box **Grids** allow you to turn major and minor grid lines on and off and modify grid line parameters.

o If the **Logarithmic Scale** box is checked, the axis and all the plots based on it will be scaled logarithmically.

4.7.4 Vertical Axis dialog box

This dialog box is used for editing axes.
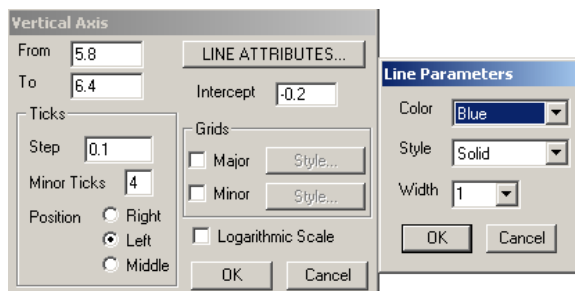
Please refer to **Horizontal Axis** section for the detailed illustration of all the features available in this dialog box.

4.7.5 Bar Graph Parameters dialog box

This dialog box is used for editing the parameters of all bars in a regular bar graph, or a selected group member of grouped bar graphs. It is displayed when a bar in the

histogram (**Histogram** option on the **Univariate** menu, bar, chart or pie-slice (see Section 4.3)) is double-clicked.
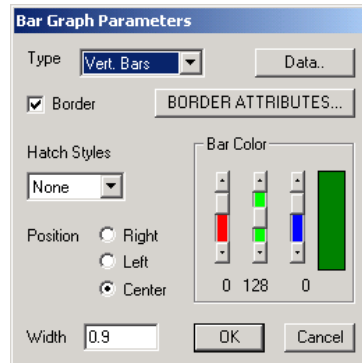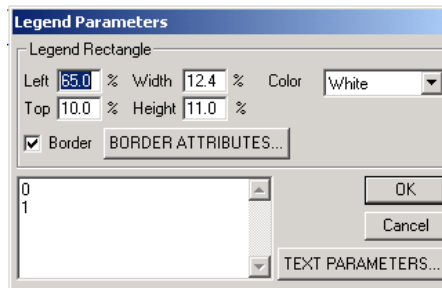
It operates as follows:

- o If the **Border** check box is checked, the bars have a border around them. In this case, the **Border Attributes** button leads to the **Line Parameters** dialog box that controls border thickness, color, and style.
- o The **Data** button leads to the spreadsheet-style window for editing plotted data points (shown below)
- o The **Hatch Style** list box allows you to choose the hatch style for bars.
- o The **Bar Color** scrolling bars control the bar color.
- o The **Position** radio buttons control the bar position relative to the independent variable values.
- o The **Width** string field allows you to enter the bar width in units of the independent variable.

#	X	Y
0	1	10
1	2	24
2	3	21
3	4	10
4	5	4
5	6	3
6	7	4
7	8	15
8	9	14
9	10	18
10	11	6
11	12	27
12	13	3
13	14	22
14	15	17
15	16	5
16	17	27
17	18	8
18	19	14
19	20	12
20	21	4

4.7.6   Legend Parameters dialog box

This dialog box allows the editing of legends. It opens when the mouse button is double-clicked while the cursor is anywhere inside the legend box, for example, when a pie chart is displayed.

This dialog box operates as follows:

- o The **Left**, **Top**, **Width**, and **Height** edit controls allow you to specify a new position and size of the legend-bounding rectangle relative to the graph window.
- o The **Color** menu specifies the legend rectangle background color.
- o If the **Border** check box is checked, the rectangle will have a border. In this case, the **Border Attributes** button leads to the **Line Parameters** dialog box that controls border thickness, color, and style of the border line.
- o The multi-line text box in the lower left corner lists and allows editing of each of the legend text strings.
- o The **Text Parameters** button leads to the **Text Parameters** dialog box discussed later.

4.7.7 Line Parameters dialog box

This dialog box is used for editing lines in the graph. It is accessed via the **Plot Parameters** dialog box, which is activated when a curve in a graph is double-clicked.

It has the following functions:

- o The **Color** list box controls the line color.
- o The **Style** list box, visible when activated, allows selection of a line style.
- o The **Width** control specifies the line width, in window pixels.

4.7.8 Plot Parameters dialog box

The **Plot Parameters** dialog box is accessed when a curve is double-clicked.
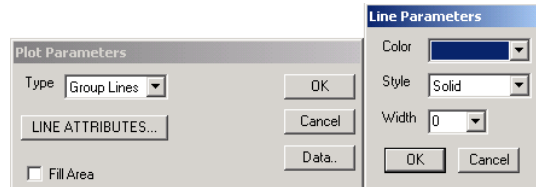
It includes the following features:

- o The type of line to be displayed may be changed using the **Type** list box.
- o To fill the area under the curve, the **Fill Area** check box may be used.
- o The type of curve fitted (spline or not) is controlled by the **Spline** check box.
- o The **Data** button provides direct access to the data used to plot the curve.
- o The **Line Attributes** button provides access to the **Line Parameters** dialog box (shown to the right of the **Plot Parameters** dialog box below). The **Line Parameters** dialog box is discussed elsewhere in this section.

4.7.9 Text Parameters dialog box

This dialog box is used for editing text strings, labels, titles, etc. It can be called from some of the other dialog boxes controlling graphic features. It may also be activated by double clicking on any text in a displayed graph.
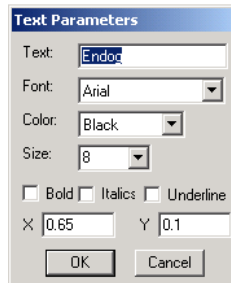
The following functions are included:

- o The **Text** edit control allows you to edit the text string.
- o The **Font** list box allows control of the typeface.
- o The text color can be selected from the **Color** menu.
- o The size of the fonts (in points) is controlled by the **Size** menu.
- o The **Bold**, **Italic** and **Underline** check boxes control the text style.

4.7.10   Pie Chart Parameters dialog box

The **Pie Chart Parameters** dialog box is accessed when a pie is double-clicked.
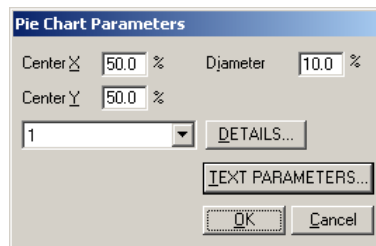
It includes the following features:

- o The **Center X**, **Center Y** and **Diameter** edit controls allow you to specify a new position and size of the pie to the graph window.
- o The drop down list contains the value or label of each category of the variable.
- o The **Details** button provides direct access to the **Pie Slice Parameter** dialog box.
- o The **Text Parameters** button provides access to the **Text Parameters** dialog box.

4.7.11 Pie Slice Parameters dialog box

The **Pie Chart Parameters** dialog box is accessed when the **Details** button on the **Pie Chart Parameters** dialog box is clicked.
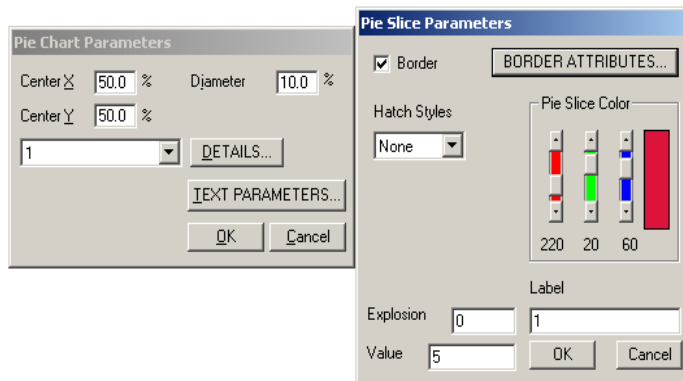
- o The **Border** check box is used if to draw the borders of the selected slice.
- o When the **Border Attributes** button is clicked, the **Line Parameters** dialog box is opened to specify the border.
- o The **Hatch Styles** drop down list enables the use to choose the hatch style.

- A real number between 0 to 1 can be entered in the **Explosion** text box to detach the selected slice from the pie.
- The **Value** text box shows the number of the observations in the selected slice.
- The **Label** text box is used to change the category label.

References

Barcikowski, R.S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, **6**(3), 267-285.

Gibbons, R.D.& Hedeker, D. (1994). Application of random-effects probit regression models, *Journal of Consulting and Clinical Psychology*, **62**, 285-296.

Gibbons, R.D. & Hedeker, D. (1994). Application of random-effects probit regression models, *Journal of Consulting and Clinical Psychology*, **62**, 285-296.

Hedeker, D. & Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Wiley: New York.

Hedeker, D. & Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis, *Biometrics*, **50**, 933-944.

Hedeker, D. & Gibbons, R.D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies, *Psychological Methods*, **2**, 64-78.

Hedeker, D., Gibbons, R.D. & Flay, B.R. (1994). Random-effects regression models for clustered data: with an example from smoking prevention research, *Journal of Consulting and Clinical Psychology*, **62**, 757-765.

Hedeker, D., Siddiqui, O., & Hu, F.B. (2000). Random-effects regression analysis of correlated grouped-time survival data, *Statistical Methods in Medical Research*, **9**, 161-179.

Hough, R.L., Harmon, S., Tarke, H., Yamashiro, S., Quinlivan, R., Landau-Cox, P., Hurlburt, M.S., Wood, P.A., Milone, R., Renker, V., Crowell, A. & Morris, E., (1997). Supported independent housing: Implementation issues and solutions in the

San Diego McKinney Homeless Demonstration Research Project. In W.R. Breakey & J.W. Thompson (eds.), *Mentally Ill and Homeless: Special Programs for Special Needs*, pp. 95-117. New York: Harwood Academic Publishers.

Hurlburt, M.S., Wood, P.A., & Hough, R.L. (1996). Providing independent housing for the homeless mentally ill: a novel approach to evaluating long-term longitudinal housing patterns, *Journal of Community Psychology*, **24**, 291-310.

Kreft, I. & De Leeuw, J. (1988). *Introducing multilevel modeling*. Sage.

Laird, N.M. (1988). Missing data in longitudinal studies, *Statistics in Medicine*, **7**, 305-315.

Longford, N.T. (1993). *Random Coefficient Models*, Oxford University Press, New York.

Lorr, M. & Klett, C.J. (1966). *Inpatient multidimensional psychiatric scale: manual*. Consulting Psychologists Press, Palo Alto, CA.

McKnight, B. & Van Den Eeden, S.K. (1993). A conditional analysis for two-treatment multiple period crossover designs with binomial or Poisson outcomes and subjects who drop out, *Statistics in Medicine*, **12**, 825-834.

Miller, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance, *The Annals of Statistics*, **5**, 746-762.

Reisby, N., Gram, L.F., Bech, P., Nagy, A., Petersen, G.O., Ortmann, J., Ibsen, J., Dencker, S.J., Jacobsen, O., Krautwald, O., Søndergaard, I., & Christiansen, J. (1977). Imipramine: clinical effects and pharmacokinetic variability, *Psychopharmacology*, **54**, 263-272.

Rubin, D.B. (1976). Inference and missing data, *Biometrika*, **63**, 581-592.

Schoenwald, S.K. & Henggeler, S.W. (2005). *Multisystemic Therapy for Adolescents with Serious Externalizing Problems*. In J. Lebow (ed.), Handbook of Clinical Family Therapy. Wiley: Jossey-Bass.

Self, G.S. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical association*, **82**, 605-610.

Singer, J.D. & Willett, J.B. (1993). It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, **18**, 155-195.

Snijders, T. & Bosker, R. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. Sage.

Subject Index

D

E

F

Advanced screen, 242

G

Global Missing Value text box
    Configuration screen, 173, 192
Grand-mean centering, 91
Graph Parameters dialog box, 268, 311
Graphics procedure
    Histogram option, 316
Graphs, 92, 128, 164, 166
    3D bar chart, 41, 272, 280, 281, 282, 283
    3D pie chart, 261
    attributes of axes, 312, 315
    bar chart, 40, 41, 265, 266, 267, 280, 281, 283, 284
    bivariate, 41, 249, 250, 273, 274, 275, 277, 278, 283
    Bivariate and Bar Chart option, 283
    Box and Whisker option, 275
    box-and-whisker plot, 41, 273, 274, 275, 277, 278
    changing attributes of, 250, 262, 268, 288, 312
    changing legends, 262, 263, 312, 317
    color attribute, 128, 252, 253, 255, 268, 294, 300, 304, 305, 312, 316, 318, 320
    confience intervals, 42
    confidence intervals, 126, 127, 250, 287, 305, 306, 307
    Confidence Intervals option, 42, 287, 305, 309
    controlling axes, 269
    data-based, 40, 41, 45, 128, 249, 250, 251, 259, 260, 261, 263, 265, 266, 267, 272, 273, 280, 281, 282, 283, 284
    displaying data for, 268, 316, 319, 321, 322
    editing, 129, 257
    editing axis labels, 312, 313, 314
    editing bar characteristics, 315
    editing bar graph parameters, 268, 271, 312, 315
    editing bar parameters, 315
    editing exploratory graphs, 256
    editing graph parameters, 268, 311
    editing legends, 317
    editing line parameters, 269, 294, 300, 312, 316, 318, 319, 321, 322
    editing lines, 318
    editing old settings, 257
    editing pie slice parameters, 263
    editing plot parameters, 294, 299, 300, 304, 305, 312, 318, 319, 320, 321
    editing text, 319
    editing text parameters, 312, 313, 314, 318, 319
    editing tools, 311
    editing vertical axis parameters, 269
    equation modeling, 250
    Equations option, 42, 287, 291, 297, 298

H

Help menu, 34
Histogram, 40, 270, 271, 272, 273, 316
Histogram option
    Graphics procedure, 316
Horizontal Axis dialog box, 269
Housing data
    analysis of, 214, 220, 223

M

N

Number of Quadrature Points text box
    Advanced screen, 156, 175, 224

P

Q

R

S

T

U

V

W