Two-level model for continuous outcomes

Contents

3.1 Models based on the TVSFP data

3.1.1 The data

The data set used here is from the Television School and Family Smoking Prevention and Cessation Project (TVSFP) (Flay *et. al.*, 1988). The study was designed to test independent and combined effects of a school-based social-resistance curriculum and a television-based program in terms of tobacco use and cessation. The data from the study included a total of 1,600 students from 135 classrooms drawn from 28 schools.

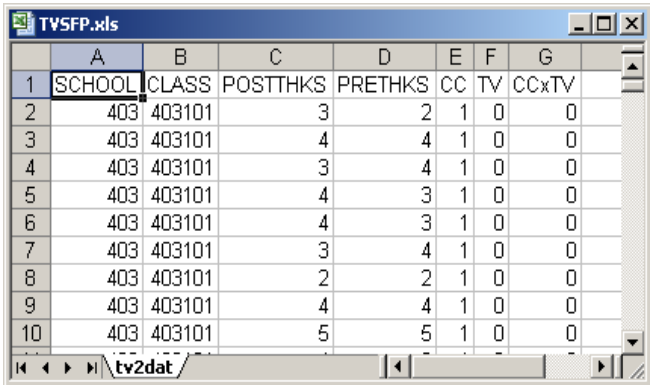
Schools were randomized to one of four study conditions:

- a social-resistance classroom curriculum
- a media (television) intervention
- a social-resistance classroom curriculum combined with a mass-media intervention, and
- a no-treatment control group

A tobacco and health knowledge scale (THKS) was used in classifying subjects as knowledgeable or not. In its original form, the student's score was defined as the number of correct answers to seven items on tobacco and health knowledge.

While the structure of this study indicates a three-level hierarchical structure, the present application uses these data to fit a two-level model, with students nested within either classes or schools, in order to present an introduction to the analysis of ordinal outcomes.

Data for the first 10 students on most of the variables used in this section are shown below in the form of an SuperMix spreadsheet file, named **TVSFP.ss3**.



TVSFP.xls

	A	B	C	D	E	F	G
1	SCHOOL	CLASS	POSTTHKS	PRETHKS	CC	TV	CCxTV
2	403	403101	3	2	1	0	0
3	403	403101	4	4	1	0	0
4	403	403101	3	4	1	0	0
5	403	403101	4	3	1	0	0
6	403	403101	4	3	1	0	0
7	403	403101	3	4	1	0	0
8	403	403101	2	2	1	0	0
9	403	403101	4	4	1	0	0
10	403	403101	5	5	1	0	0

tv2dat

The variables of interest are:

- SCHOOL indicates the school a student is from (28 schools in total).
- CLASS identifies the classroom (135 classrooms in total).
- POSTTHKS represents the post-intervention tobacco and health knowledge scale. It is treated as a continuous variable in the examples in this chapter. See Sections 4.2 and 6.2 for examples where POSTTHKS is treated as a binary or ordinal outcome.
- PRETHKS indicates the pre-intervention THKS score.
- CC is a binary variable indicating whether a social-resistance classroom curriculum was introduced, where 0 indicates "no" and 1 "yes."
- TV is an indicator variable for the use of media (television) intervention, with a "1" indicating the use of media intervention, and "0" the absence thereof.
- CCxTV was constructed by multiplying the variables TV and CC, and represents the CC by TV interaction.

3.1.1.1 Exploring the data

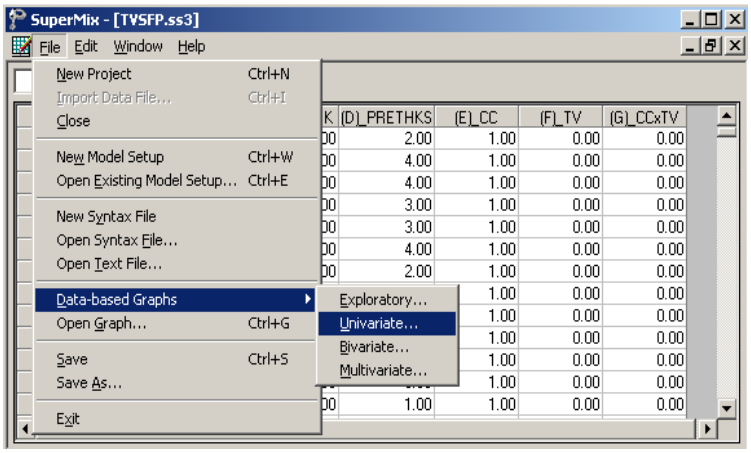
In this section, a univariate bar chart and a bivariate box-and-whisker plot are given.

Univariate graphs

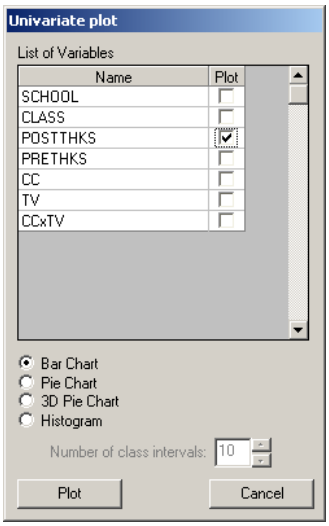
The pop-up menu below shows the data-based graphing options currently available in SuperMix. As a first step, we will take a closer look at the distribution of the total post-intervention scores (POSTTHKS), which is the potential dependent variable in this study. While scores such as these are not truly continuous variables, they are often treated as if they were.

Bar chart

To do so, select the **Univariate** option from the **Data-based Graphs** menu as shown below.



The **Univariate plot** dialog box appears. Select the variable POSTTHKS and indicate that a **Bar Chart** is to be graphed. Click the **Plot** button to display the bar chart.



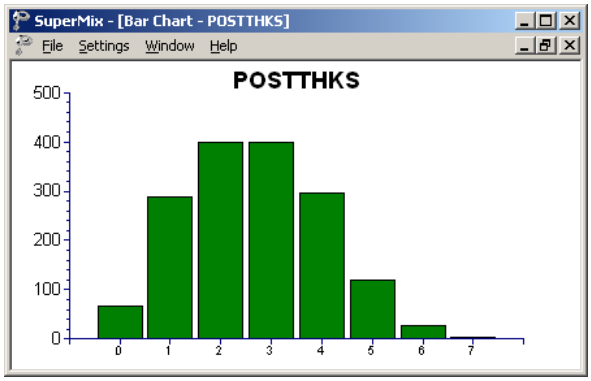


Figure 3.11: Bar chart of POSTTHKS scores

The bell-shaped bar chart above shows that the variable POSTTHKS is approximately normally distributed. Note that histograms are usually used for the depiction of the distribution of a continuous variable.

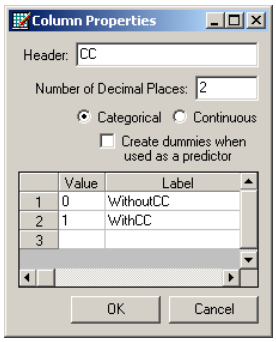
Bivariate graphs

It is hoped that the social-resistance classroom curriculum (CC), the television intervention (TV) and the CC and TV interaction combination (CCxTV) would affect the tobacco and health knowledge (POSTTHKS). Before we start with the model, we would like to show a box-and-whisker plot of POSTTHKS for each category of CC.

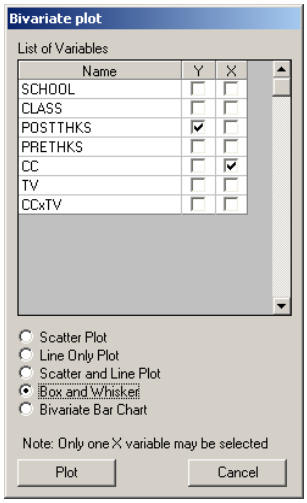
Box-and-whisker plots

A box-and-whisker plot is useful for depicting the locality, spread and skewness of variables in a data set and may be used to examine the distributions of continuous variables, such as for the different values of discrete valued predictors. This option is accessed via the **Data-based Graphs**, **Bivariate** option on the **File** menu.

To assign labels to the categories of CC, right-click on the CC column in the spreadsheet and select **Column Properties**. On the **Column Properties** dialog box, select the **Nominal** option and assign the appropriate labels and save the data file.



The **Bivariate plot** dialog box is completed as shown below: select the outcome variable POSTTHKS as the **Y**-variable of interest, and the predictor CC to be plotted on the **X**-axis. Check **the Box and Whisker** option, and click **Plot**.



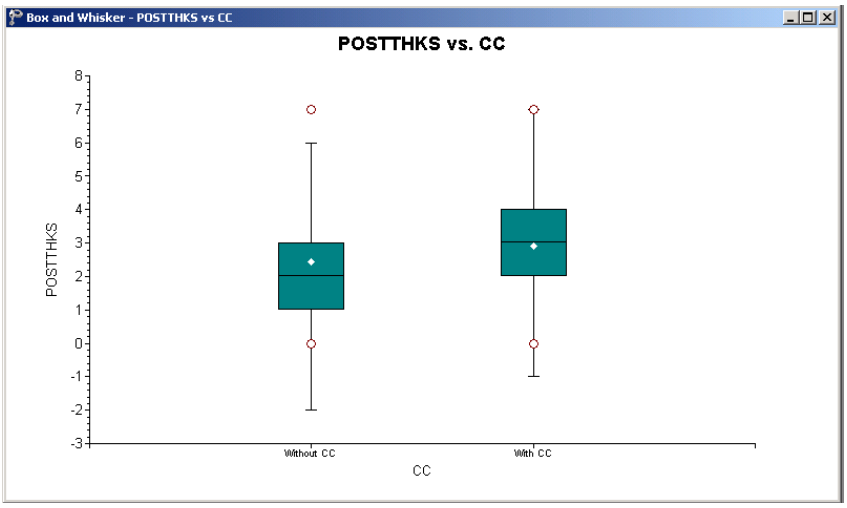


Figure 3.12: Box-and-whisker plots of POSTTHKS scores for different CC values

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, the mean of POSTTHKS is lower in the group without the social-resistance classroom curriculum (CC). The box-and-whisker plot indicates a positive relationship between CC and POSTTHKS.

3.1.2 A 2-level random intercept model using classroom as level-2 ID

3.1.2.1 The model

The first model fitted to the data explores the cluster effects of each classroom on the outcome. The mixed model can be expressed as

$$\text{POSTTHKS}_{ij} = \beta_0 + \beta_1 \text{CC}_i + \beta_2 \text{TV}_i + \beta_3 (\text{CC}_i \times \text{TV}_i) + v_{0i} + e_{ij},$$

where v_{0i} represents the classroom influence on POSTTHKS. To understand the model better, we can rewrite the model in the following way. The level-1 or within-cluster model is shown below.

Level-1 model: $(j = 1, ..., n_i)$

$$\text{POSTTHKS}_{ij} = b_{0i} + e_{ij},$$

$$e_{ij} : \ NID(0, \sigma^2)$$

The level-1 model estimates POSTTHKS as a function of the intercept b_{0i} and error term e_{ij}. Subscript i denotes the subscript for classroom, while subscript j refers to the student j. n_i is used to denote the number of students in each classroom. Because we have different numbers of students in different classrooms, n_i also varies. In this data set, $1 \le n_i \le 28$.

The level-2, or between-cluster, model describes the intercept b_{0i} as a function of cluster characteristics.

Level-2 model: $(i = 1, \text{L}\ , N)$

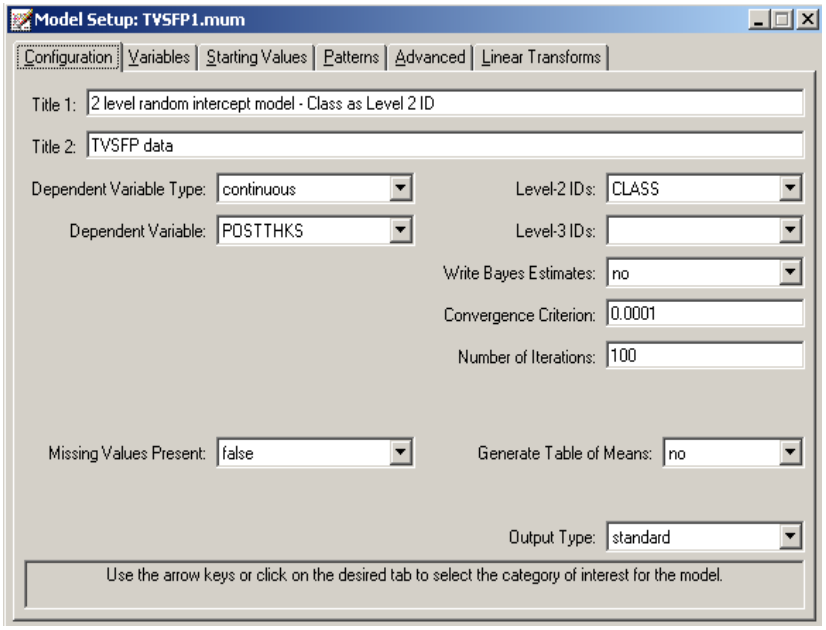
$$b_{0i} = \beta_0 + \beta_1 \text{CC}_i + \beta_2 \text{TV}_i + \beta_3 (\text{CC}_i \times \text{TV}_i) + v_{0i}$$

$$v_{0i} : \ NID(0, \sigma_v^2)$$

As shown above, the intercept b_{0i} is estimated as a function of the population average β_0, the covariates CC_i, TV_i, and $\text{CC}_i \times \text{TV}_i$, and the classroom difference v_{0i}. The coefficient v_{0i} represents the amount that unit i deviates from the average β_0, after controlling for the effects of the covariates included. The level-2 residual v_{0i} is assumed to follow $NID(0, \sigma_v^2)$ for all the is. If $v_{0i} = 0$ for all i, which implies $\sigma_v^2 = 0$, the model is the same as the ordinary regression model.

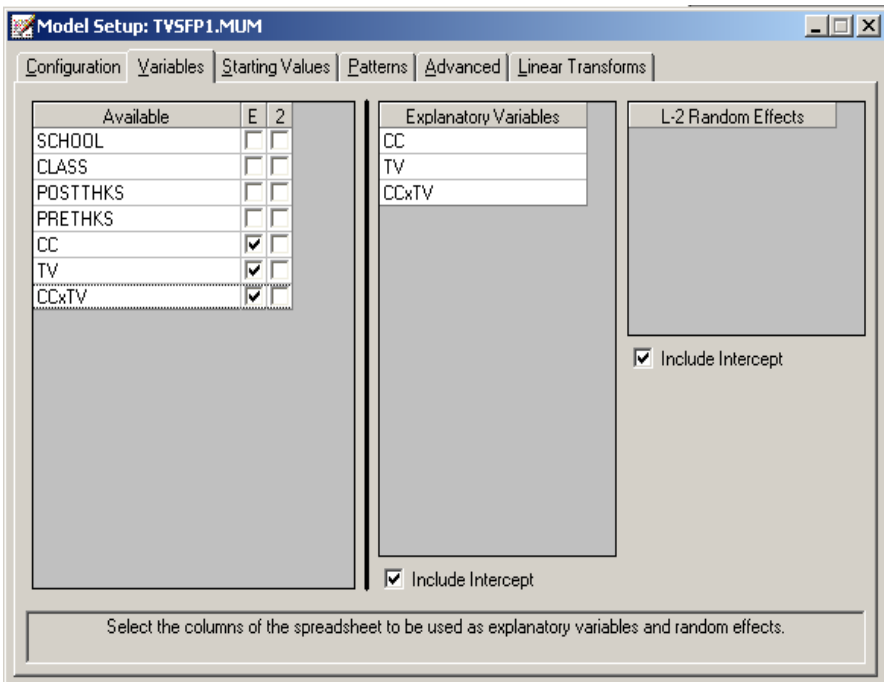
3.1.2.2 Setting up the analysis

Open the SuperMix spreadsheet **TVSFP.ss3** used during the exploratory analysis discussed previously in this chapter. 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.



Select the continuous outcome variable POSTTHKS from the **Dependent Variable** drop-down list box. Select the classroom number CLASS from the **Level-2 IDs** drop-down list box. Enter a title for the analysis in the **Title** text boxes. In this example, default settings for all other options associated with the **Configuration** screen are used.

Proceed to the **Variables** screen by clicking on that tab. The **Variables** screen is used to specify the fixed and random effects to be included in the model. Select 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.



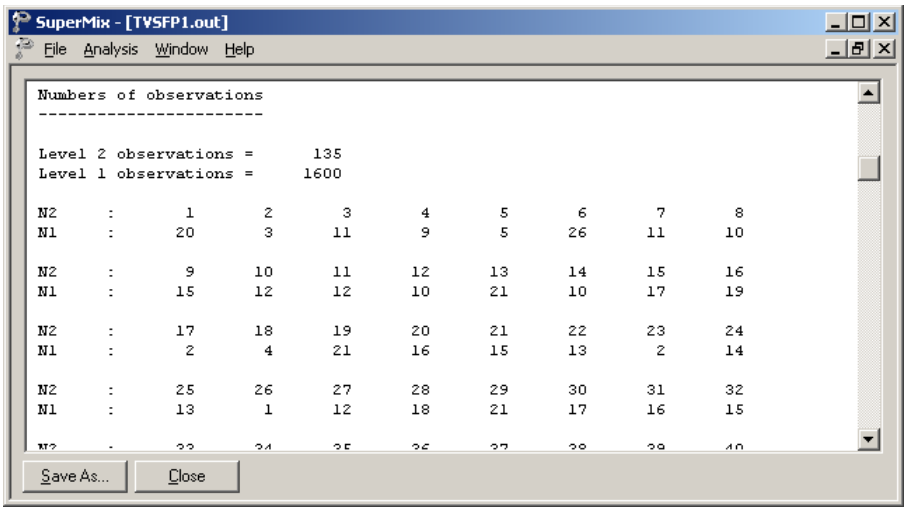
Next, specify the random effects at level 2 the hierarchy. In this example, we want to fit a model with random intercepts at level 2. By default, the **Include Intercept** check box in the **L-2 Random Effects** grid is checked. If this box is left checked, and no additional random effects are indicated 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 (**TVSFP1.mum**) for the model specification file. Run the analysis by selecting the **Run** option from the **Analysis** menu.

3.1.2.3 Discussion of results

Model and data description

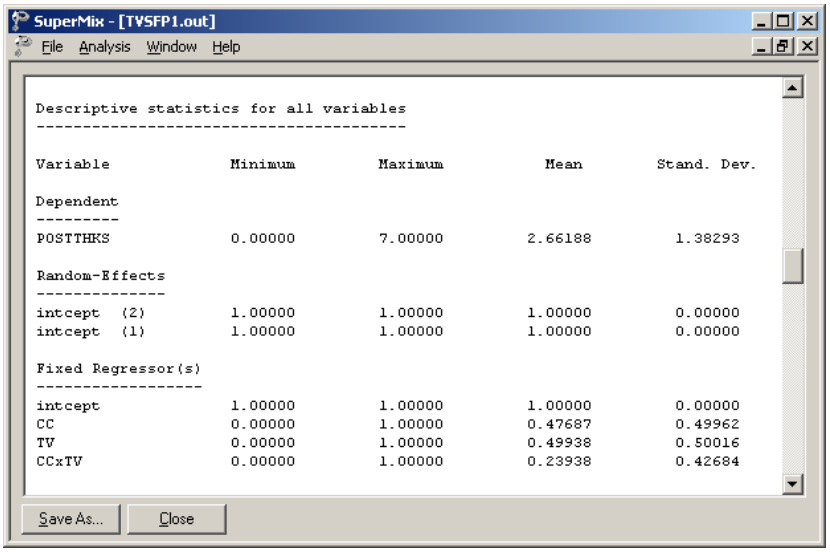
In the **numbers of observations** section, a summary of the hierarchical structure is provided.



As shown above, data from a total of 1600 students within 135 classrooms were included at levels 2 and 1 of the model. This corresponds to the study design described earlier. In addition, a summary of the number of students nested within each classroom is provided. The classroom with N2 = 6, for example, had 26 students (N1: 26). By contrast, classroom 26 had only 1 student.

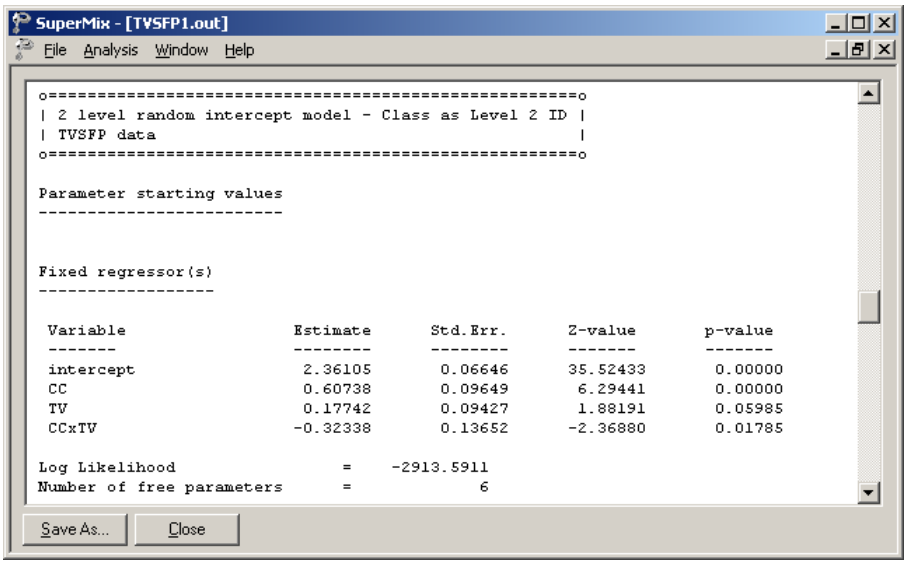
Descriptive statistics and starting values

Next, the **descriptive statistics for all variables** are given. The minimum value, maximum value, mean and standard deviation are given for all the variables included in the model. For example, the mean POSTTHKS is 2.6618 with a standard deviation of 1.38293.

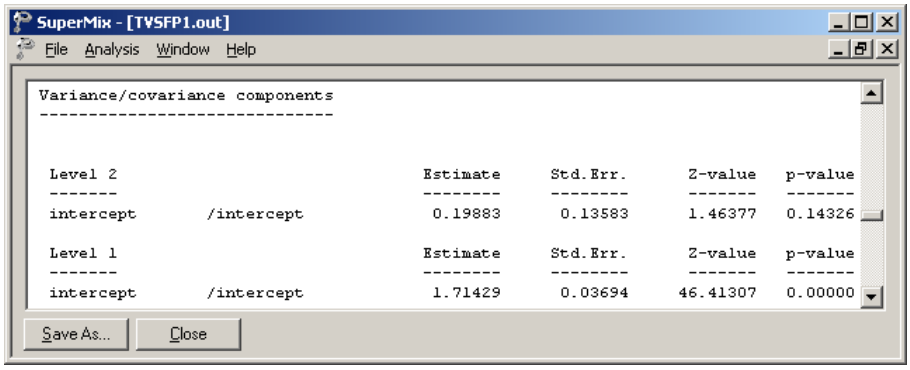


Starting values – OLS estimates

The starting values for the **fixed regressor(s)** are shown below. The **log likelihood** value and **number of free parameters** of the OLS regression are given in this part of the output.



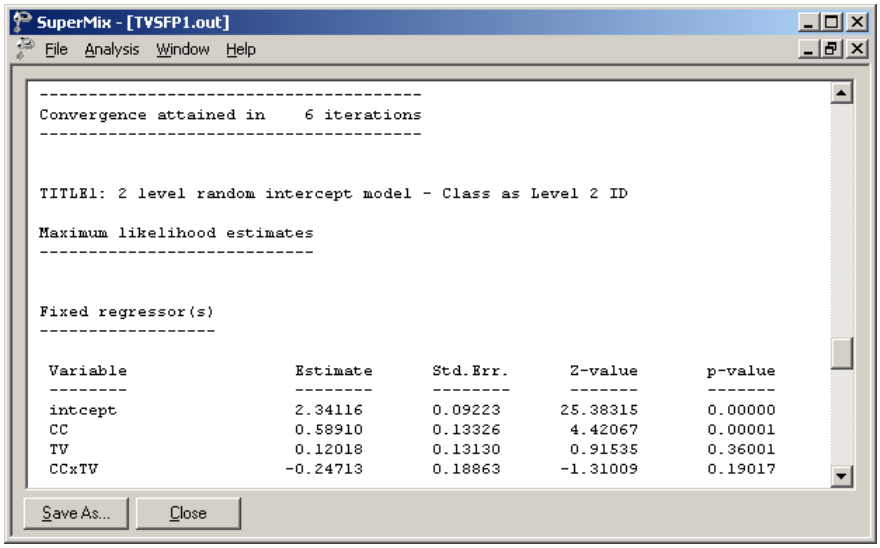
After the **number of free parameters**, the starting values of **variance/covariance components** are reported as shown.



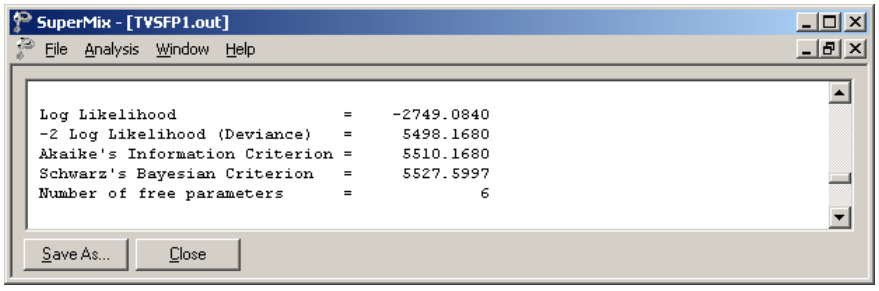
Fixed effects estimates

The number of iterations needed to obtain convergence is given after the starting values. The output describing the estimated **fixed regressor(s)** after convergence is shown next.

As shown below, the estimates for CC and TV are both positive. On average, a social-resistance classroom curriculum can improve the tobacco and health knowledge by 0.58910, and television intervention can increase the POSTTHKS score by 0.12018. However the estimate of CCxTV is negative, which implies that the students who had both CC and TV are expected to show a decrease of 0.24713 in their POSTTHKS score. The estimates associated with intercept and TV are highly significant, but estimates of the other two coefficients are not statistically significantly different from zero.

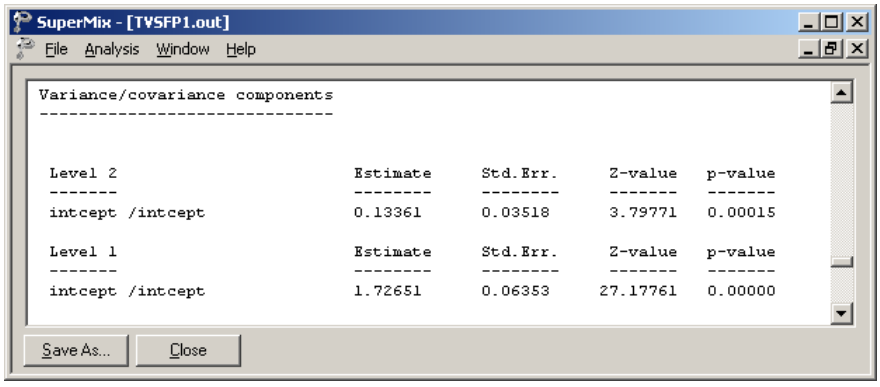


The estimates for the fixed regressors and model fit statistics are given next.



Random effect estimates

The estimates for the random part of the model are reported next. The variation in the average estimated intercept at level 2 is highly significant, which indicates that the classroom difference in intercepts does help to explain the variation in POSTTHKS scores.



The covariance and correlation matrix of level-2 and level-1 random effects are given in matrix format at the end of the output file. These values are the same as the estimates of variance/covariance components as shown above.

3.1.2.4 Interpreting the results

Estimated outcomes for different groups

For a student who participated in neither social-resistance classroom curriculum nor television intervention (CC = 0; TV = 0), the expected POSTTHKS is equal to just the intercept 2.36105. For a student who participated in both programs (CC = 1; TV = 1; CCxTV = 1), the predicted POSTTHKS is calculated as follows:

$$\begin{aligned}\hat{\text{POSTTHKS}}_{ij} &= \hat{\beta}_0 + \hat{\beta}_1 \text{CC}_i + \hat{\beta}_2 \text{TV}_i + \hat{\beta}_3(\text{CC}_i \times \text{TV}_i) \\ &= 2.34116 + 0.5891 + 0.12018\text{-}0.24713 \quad . \\ &= 2.80331\end{aligned}$$

Fit statistics and % variation explained

An estimate of the percentage of variation in the outcome at classroom level is obtained as

$$\frac{0.13361}{0.13361 + 1.72651} \times 100\% = 7.18\%$$

indicating that about 7.18% of the total variance lies between the clusters/classrooms and that 92.82% of the variance remains at the student level.

3.1.3 2-level random intercept model by using school as level-2 ID

The model in the previous section shows that only about 7% of the total variation in outcome is at the classroom level. The question that arises is whether clustering within schools may provide a better explanation of the way in which post-intervention scores vary. In this section, the model is fitted using SCHOOL, rather than classroom, as the level-2 ID.

3.1.3.1 The model

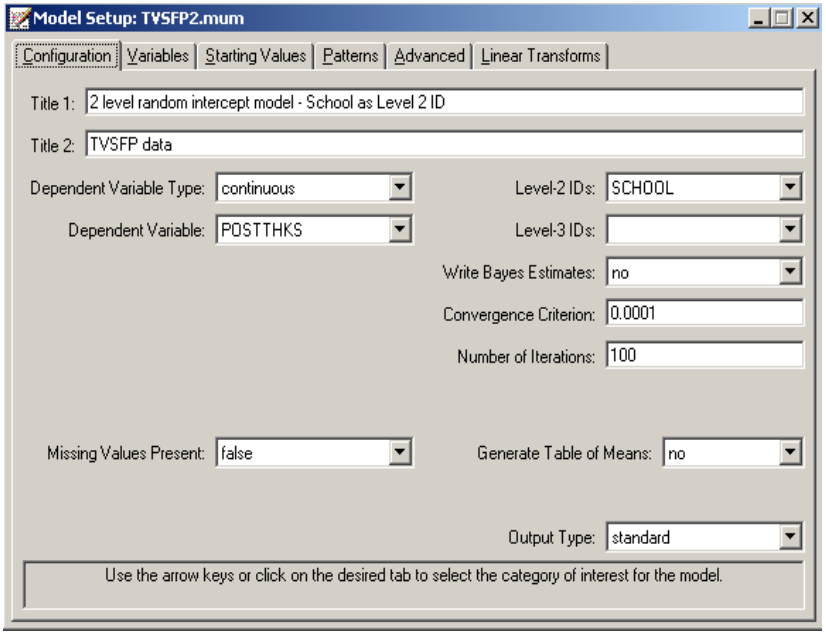
The mathematical equation of the model to be fitted is exactly the same as for the previous model.

$$\text{POSTTHKS}_{ij} = \beta_0 + \beta_1 \text{CC}_i + \beta_2 \text{TV}_i + \beta_3(\text{CC}_i \times \text{TV}_i) + v_{0i} + e_{ij},$$

The difference here is in the meaning of the subscript *i*. In the previous model, we used *i* to refer the classroom. However, the *i*s here refer to the schools.

3.1.3.2 Setting up the analysis

To create the model specifications for this model, we start by opening **TVSFP.ss3** in a SuperMix spreadsheet window.



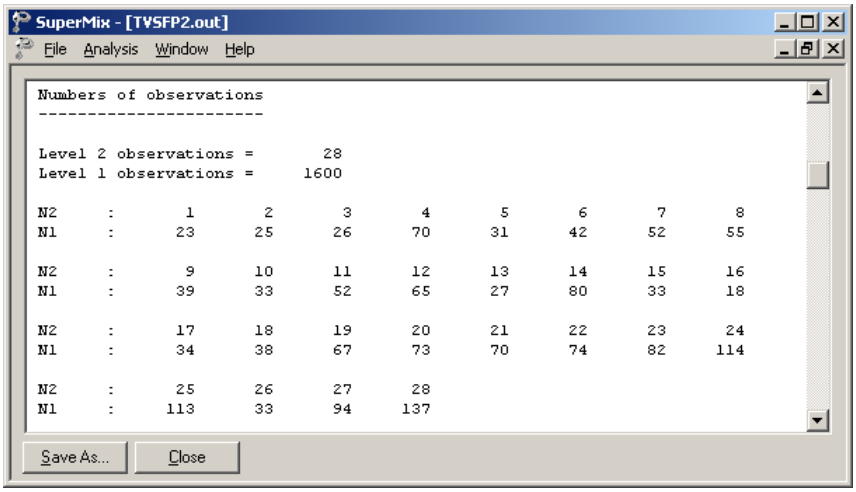
We use the **Open Existing Model Setup** option on the **File** menu to load the **Model Setup** window for **TVSFP1.mum**. Click on **File**, **Save as** to save the model setup in a new file, such as **TVSFP2.mum**. Next, change the string in the **Title 1** text box on the **Configuration** screen, and select SCHOOL as the **Level-2 ID** as shown above.

Keep all the other settings unchanged. Save the changes to the file **TVSFP2.mum** and select the **Run** option on the **Analysis** menu to produce the output file **TVSFP2.out**.

3.1.3.3 Discussion of results

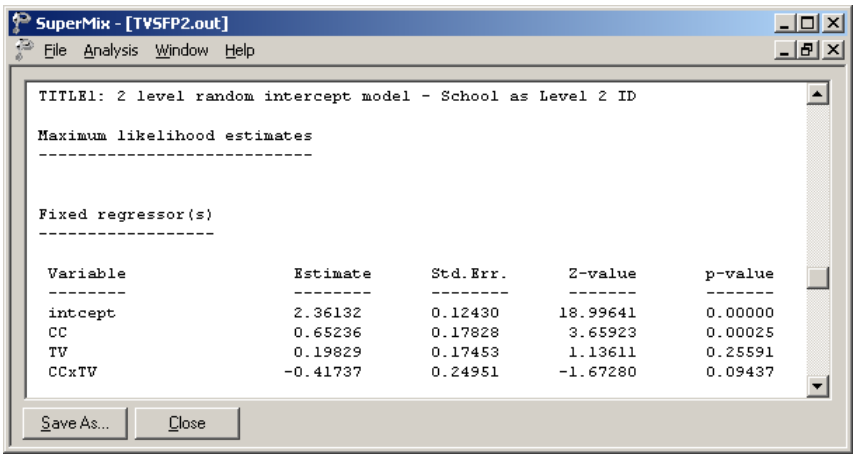
Model and data description

The **number of observations** section clearly shows that the data set contains 28 schools and each school has between 18 and 137 students as shown below.

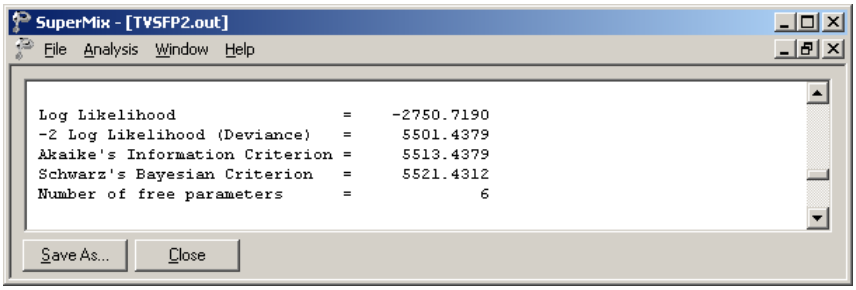


Fixed effects estimates and descriptive statistics

The estimates for the fixed estimates as shown below are close to the estimates in the previous example, but not exactly the same. For example, the estimate for CC increased by 0.06326 (0.65236 – 0.58910 = 0.06326), and the estimate for the effect of television intervention is about 0.07811 higher when using school as the level-2 ID. However, the estimate of the interaction of CC and TV is about 0.17 lower.

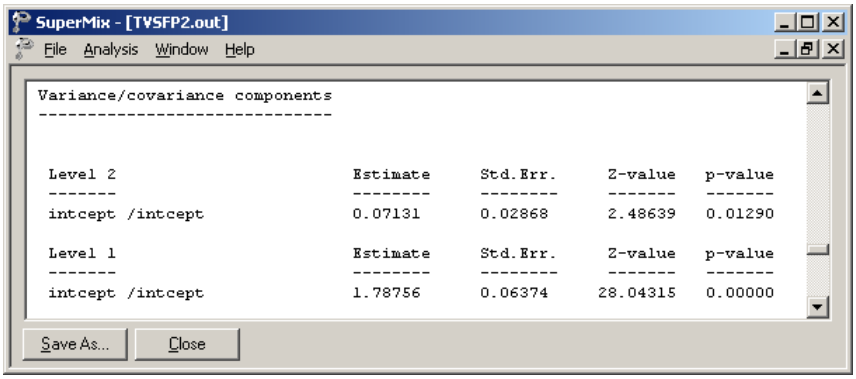


Both the deviance and Akaike information criterion (AIC) are slightly higher than the previous model. The SBC is smaller.



Random effect estimates and covariance/correlation matrices

The estimates for the random part of the model are reported next.



The variation in the average estimated intercept at level 2 is highly significant, which indicates that the difference in school intercepts also explains the variation of POSTTHKS scores. Similarly, we can calculate that about 3.84% of the total variance can be explained by the school difference:

$$\frac{0.07131}{0.07131+1.78756}\times 100\% = 3.84\%.$$