



Two-level models for continuous outcomes

Contents

3.1	MODELS BASED ON A SUBSET OF THE NESARC DATA.....	1
3.1.1	<i>The data</i>	1
3.1.2	<i>2-level random intercept model with 2 predictors</i>	6
3.1.3	<i>A 2-level random intercept model with 4 predictors</i>	18

3.1 Models based on a subset of the NESARC data

3.1.1 The data

The data set is from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), a longitudinal survey with its first wave fielded in 2001–2002. The NESARC is a representative sample of the United States population, and 43,093 Americans participated in the first wave of the survey. The NESARC survey was conducted and sponsored by the National Institute on Alcohol Abuse and Alcoholism (NIAAA). Detailed information is available at <http://niaaa.census.gov/index.html>.

Section 4 of the NESARC data documentation describes data regarding major depression, family history of major depression and dysthymia. Together with the demographic information in Section 1, we produced the **nesarc_ii2.xls** data set as shown below. There are 2,339 dysthymia respondents in the survey. After listwise deletion, the sample size is 1,698.

	A	B	C	D	E	F	G	H
1	PSU	WEIGHT	WHITEOTH	BLACK	HISPANIC	M_S_DEP	ARG_DEP	AGE_DEP
2	1011	3476.6668	1	0	0	0	1	48
3	1011	3052.0965	1	0	0	1	0	59
4	1011	1182.0326	0	1	0	0	1	36
5	1011	3041.0523	1	0	0	0	1	17
6	1011	8342.9421	1	0	0	1	0	16
7	1011	6767.0638	1	0	0	0	1	29
8	1011	3460.2895	1	0	0	0	1	43
9	1011	3460.2895	1	0	0	0	0	41
10	1015	3167.2861	1	0	0	0	1	55

The variables of interest are:

- PSU is the Census 2000/2001 Supplementary Survey (C2SS) primary sampling unit (PSU).
- WEIGHT is the final weight, calculated as the product of the NESARC base weight and other individual weighting factors.
- WHITEOTH represents the white and other ethnicities, excluding African American and Hispanic. It is recoded from items S1Q1C, S1Q1D2, S1Q1D3 and S1Q1D5 in the NESARC source code (1 for white and other, 0 for African American and Hispanic).
- BLACK represents African Americans. It is recoded from items S1Q1C and S1Q1D3 in the NESARC source code (1 for African American, 0 for others).
- HISPANIC is an indicator for Hispanic. It is recoded from items S1Q1C, S1Q1D3 and S1Q1D5 (1 for Hispanic, 0 for others).
- M_S_DEP is recoded from item S4BQ10C. It is the response to the statement "Any of natural mother's full sisters ever depressed," with 1 for "Yes," and 0 for "No."
- ARG_DEP is recoded from item S4CQ43. It represents the response to the statement "Had arguments/friction with family, friends, people at work, or anyone else," with 1 for "Yes," 0 for "No."
- AGE_DEP is a renamed version of item S4CQ7AR. It represents the age at onset of first episode of dysthymia.

Inspection of the data shows that only about 2% of 43,093 respondents are of Asian and Pacific origin. Due to the skewness of the distribution of ethnicity, we recoded the variables representing ethnic origin. The resulting variable WHITEOTH represents this recoding of respondents as being either white or from other ethnic groups (blacks and Hispanics excluded).

3.1.1.1 Importing the data and defining variable types

The data set shown previously is available in the form of a spreadsheet file, named **nesarc_II2.xls**. This file contains a subset of the original NESARC data, *i.e.* data for the 1,698

respondents who reported some form of depression and for whom complete information on variables of interest was available.

The first step is to create the SuperMix spreadsheet file (*.ss3) from the Excel file:

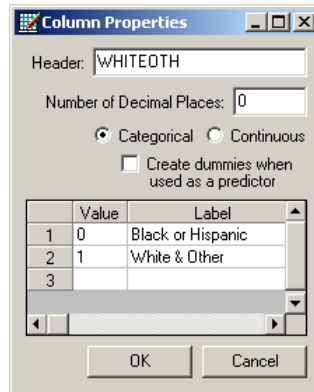
- Use the **Import Data File** option on the **File** menu to load the **Open** dialog box.
- Browse for the file **nesarc_II2.xls** in the **examples** folder of the SuperMix installation folder.
- Select the file and click on the **Open** button to open the following SuperMix spreadsheet window **nesarc_II2.ss3**.

	(A)_PSU	(B)_WEIGH	(C)_WHITE	(D)_BLACK	(E)_HISPANIC	(F)_M_S_DEP	(G)_ARG_DEP	(H)_AGE_DEP
1	1011.00	3476.67	1.00	0.00	0.00	0.00	1.00	48.00
2	1011.00	3052.10	1.00	0.00	0.00	1.00	0.00	59.00
3	1011.00	1182.03	0.00	1.00	0.00	0.00	1.00	36.00
4	1011.00	3041.05	1.00	0.00	0.00	0.00	1.00	17.00
5	1011.00	8342.94	1.00	0.00	0.00	1.00	0.00	16.00
6	1011.00	6767.06	1.00	0.00	0.00	0.00	1.00	29.00
7	1011.00	3460.29	1.00	0.00	0.00	0.00	1.00	43.00
8	1011.00	3460.29	1.00	0.00	0.00	0.00	0.00	41.00
9	1015.00	3167.29	1.00	0.00	0.00	0.00	1.00	55.00
10	1019.00	3053.85	0.00	1.00	0.00	0.00	0.00	15.00

Next, we define the variable types. Highlight WHITEOTH by clicking on the variable name, and then right click to open the following pop-up menu. Select the **Column Properties** option

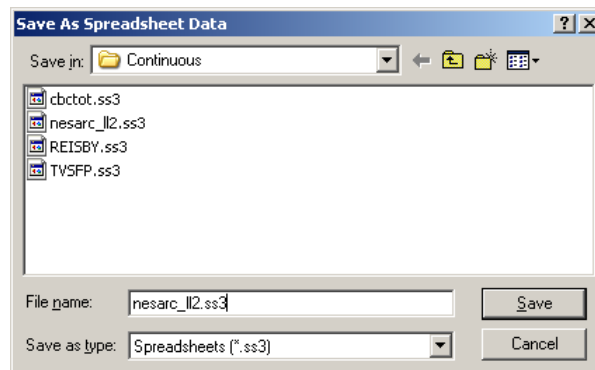
	(A)_PSU	(B)_WEIGH	(C)_WHIT	(F)_M_S_DEP	(G)_ARG_DEP	(H)_AGE_DEP
1	1011.00	3476.67		0.00	1.00	
2	1011.00	3052.10		1.00	0.00	
3	1011.00	1182.03		0.00	1.00	
4	1011.00	3041.05		0.00	1.00	
5	1011.00	8342.94		1.00	0.00	
6	1011.00	6767.06		0.00	1.00	
7	1011.00	3460.29		0.00	1.00	
8	1011.00	3460.29		0.00	0.00	
9	1015.00	3167.29		0.00	1.00	
10	1019.00	3053.85		0.00	0.00	
11	1019.00	1871.58		0.00	0.00	
12	1019.00	1819.58		1.00	1.00	

to open the **Column Properties** dialog box. Checking the **Nominal** radio button enables the user to define the labels. Input correct labels for the different categories as shown below.



Similarly define BLACK, HISPANIC, M_S_DEP and ARG_DEP as nominal variables and define AGE_DEP as continuous.

To save the **nesarc_112.ss3** spreadsheet, select the **Save As** option from the **File** menu to load the **Save As Spreadsheet Data** dialog box, and then enter the desired file name in the **File name** string field as shown below. Click on the **Save** button when done.



3.1.1.2 Exploring the data

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. 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. Model-based graphs are available after the analysis has been performed, and will be discussed later in this section.

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

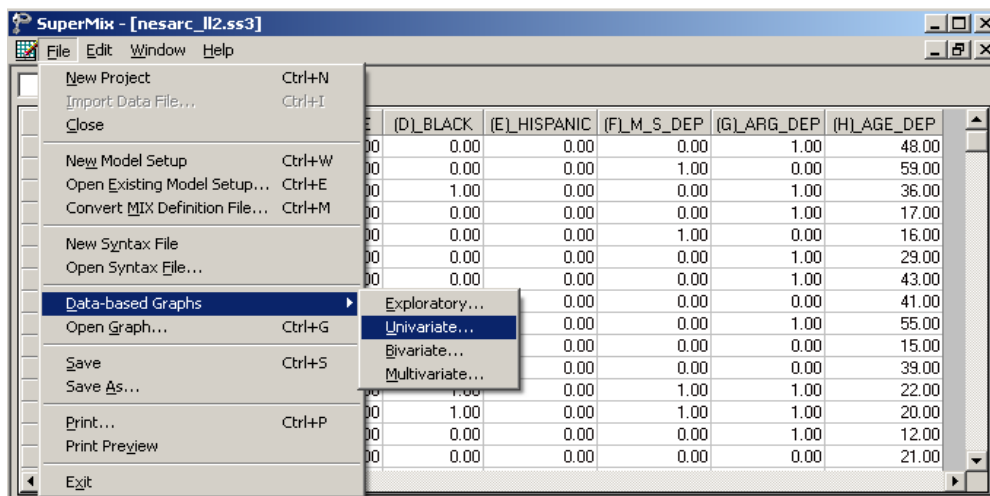
needed to explore longitudinal data as completely as one would wish. For that purpose, bivariate and multivariate data-based graphs are more appropriate.

Univariate graphs

The pop-up menu below shows the data-based graphing options currently available in SuperMix. As a first step, we take a look at the distribution of age at onset of first depression episode (AGE_DEP), which is the potential dependent variable in this study.

Histograms

A histogram represents the frequency of cases per unit interval. It gives a good picture of the distribution of a variable. To create a histogram for AGE_DEP, select the **Univariate** option from the **Data-based Graphs** menu as shown below.



The **Univariate plot** dialog box appears. Select the variable AGE_DEP and indicate that a **Histogram** is to be graphed. The desired number of intervals shown on the histogram is controlled by the **Number of class intervals** field. It is specified as 18 in this case. Click the **Plot** button to display the histogram.



The histogram, as seen below, shows that the distribution of AGE_DEP is nearly symmetrical, and should satisfy the normality assumptions implicit in a multilevel model.

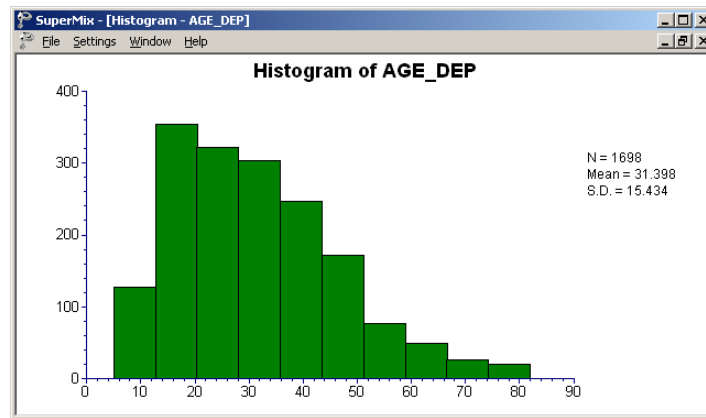


Figure 3.1: Histogram of the variable AGE_DEP

3.1.2 2-level random intercept model with 2 predictors

3.1.2.1 The model

A two-level multilevel model consists of two submodels, one at each level of the hierarchy. A general two-level model for a continuous response variable y depending on a set of p predictors x_1, x_2, \dots, x_r can be written in the form

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

where $i = 1, 2, \dots, N$ denotes the level-2 units, and $j = 1, 2, \dots, n_i$ 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 p predictors. The vector $\boldsymbol{\beta}$ contains the fixed, but unknown parameters to be estimated. $\mathbf{z}'_{ij}\mathbf{v}_i$ and 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{v}_i the vector of random level-2 effects to be estimated. It is assumed that $v_{01}, v_{02}, \dots, v_{0N}$ are independently and identically distributed (i.i.d.) with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(v)}$. Similarly, the e_{ij} are assumed i.i.d., with mean 0 and variance σ^2 .

The first model fitted to the NESARC data explores the relationship between AGE_DEP and the maternal-side depression and argument involvement, as represented by the variables M_S_DEP and ARG_DEP. The level-1 model is at a patient level, while the level-2 model is at a PSU level. The model can be expressed as

Level-1 model:

$$\text{AGE_DEP}_{ij} = b_{0i} + b_{1i} \times (\text{MS_DEP})_{ij} + b_{2i} \times (\text{ARG_DEP})_{ij} + e_{ij}$$

Level-2 model:

$$\begin{aligned} b_{0i} &= \beta_0 + v_{0i} \\ b_{1i} &= \beta_1 \\ b_{2i} &= \beta_2 \end{aligned}$$

where

$$\begin{aligned} e_i &: N(0, \sigma^2 \mathbf{I}_i) \\ \mathbf{v}_i &: N(0, \boldsymbol{\Sigma}_i) \end{aligned}$$

β_0 denotes the average expected age at onset of the first episode and β_1 denotes the coefficient of the predictor variable M_S_DEP (slope) in the fixed part of the model. Given that the variable M_S_DEP is an indicator variable, β_1 is in effect the expected change in age at onset for patients who reported maternal-side depression. Likewise, β_2 is in effect the expected change in age at onset for patients who reported arguments and stress. The random coefficients v_{i0} and e_{ij} denote the variation in the average expected AGE_DEP value between PSUs and between patients respectively.

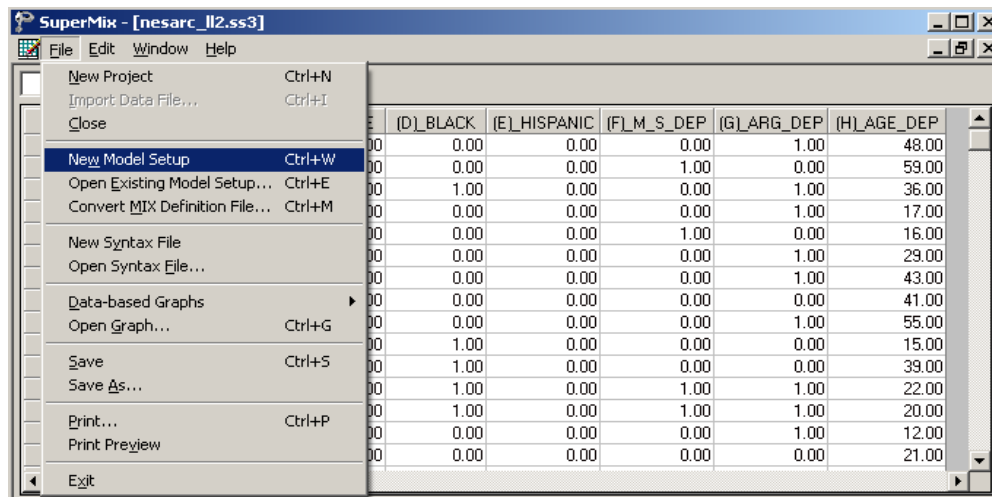
The model can also be written in so-called mixed model notation, as shown below.

$$\text{AGE_DEP}_{ij} = \beta_0 + \beta_1 * \text{M_S_DEP}_{ij} + \beta_2 * \text{ARG_DEP}_{ij} + v_{i0} + e_{ij}$$

3.1.2.2 Setting up the analysis

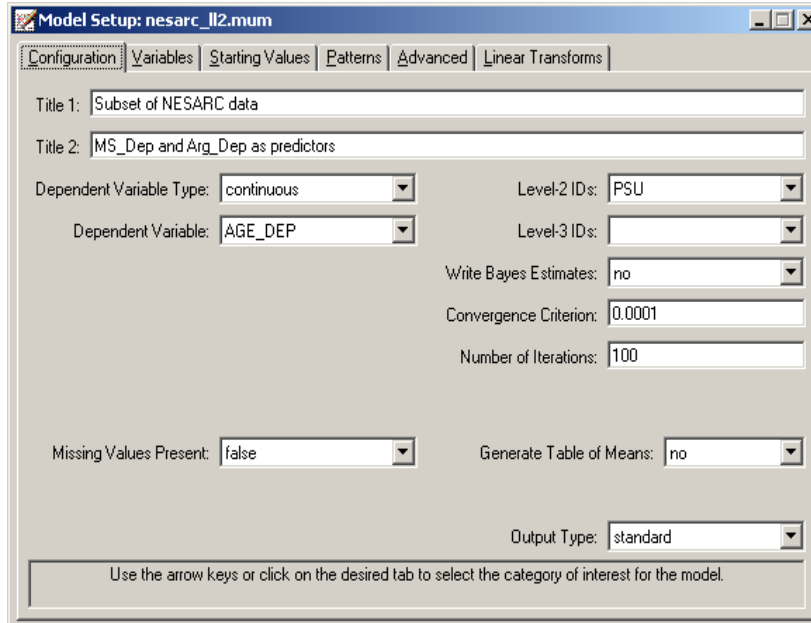
Open the SuperMix spreadsheet **nesarc_II2.ss3** used during the exploratory analysis discussed previously. The next step is to describe the model to be fitted. We use the SuperMix interface to provide the model specifications. From the main menu bar, select the **File, New Model Setup** option.

The **Model Setup** window that appears has six tabs. In this example, only the screens associated with the first two tabs are used. Information entered on these tabs are subsequently saved to a syntax file (*.mum) that can be retrieved later as needed.



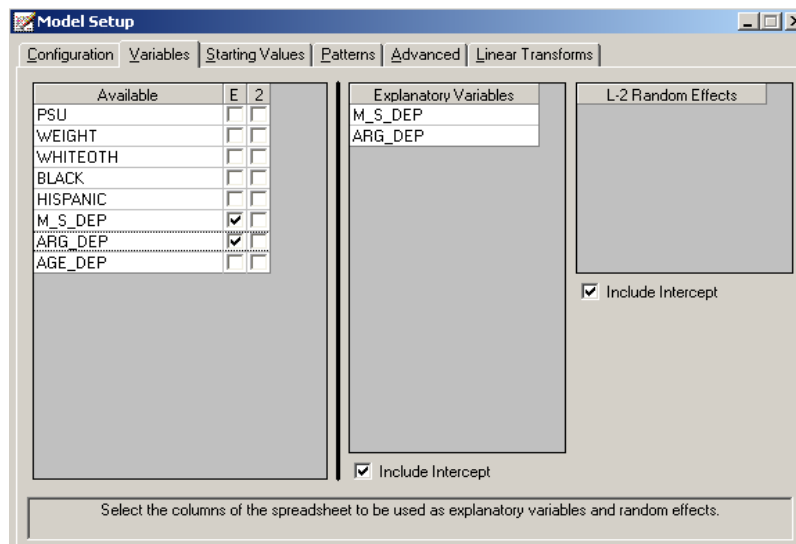
The **Configuration** screen is the first tab on the **Model Setup** window. It enables the user to define the outcome variable, level-2 and level-3 IDs. Some other settings such as missing values, the convergence criterion, the number of iterations, etc. can be specified here. To obtain the model we discussed, proceed as follows:

- Select the continuous outcome variable AGE_DEP from the **Dependent Variable** drop-down list box.
- Select PSU from **Level-2 ID** drop-down list box.
- Enter a title for the analysis in the **Title** text boxes (optional).
- Keep all the other settings on the **Configuration** screen at their default values. Proceed to the **Variables** screen by clicking on that tab.

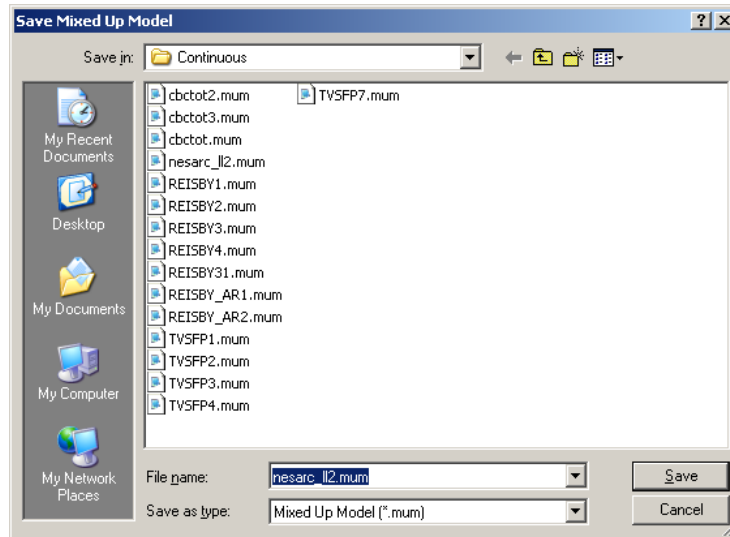


The **Variables** screen is used to specify the fixed and random effects to be included in the model. This screen shows the list of variables available for analysis and next to it two columns, with headings **E** (for explanatory variables) and **2** (for level-2 random effects). Select the explanatory (fixed) variables by checking the **E** check boxes next to the variables M_S_DEP and ARG_DEP in the **Available** grid at the left of the screen. Note that, as the variables are selected, they are listed in the **Explanatory Variables** grid. After selecting all the explanatory variables, the screen shown below is obtained.

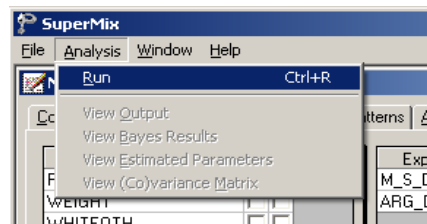
Note that the **Include Intercept** check boxes in the **Explanatory Variables** grid and **L-2 Random Effects** grid are checked by default, indicating that an intercept term will automatically be included in the fixed and random parts of the model.



Before running the analysis, the model specifications have to be saved. Select the **File, Save As** option, provide a name (**nesarc_II2.mum**) for the model specification file, and save.



Run the analysis by selecting the **Run** option from the **Analysis** menu. The standard output file opens. It can also be viewed by selecting the **View Output** option from the same menu.



3.1.2.3 Discussion of results

Portions of the output file **nesarc_II2.out** are shown below.

Program information and syntax

At the top of the output file, program information is given. It states the type, date and time of analysis, and provides contact information for technical support.

Program information is followed by model specifications. This section echoes the contents of the syntax file **nesarc_II2.mum**.

Model specifications are as follows:

```

Model=Continuous;
Options Output=standard Converge=0.0001 Maxiter=100 Bayes=No;
Link=identity;
Distribution=nor;
Varnames= PSU WEIGHT WHITEOTH BLACK HISPANIC M_S_DEP ARG_DEP AGE_DEP intercept;
Title1=Subset of NESARC data;
Title2=MS_Dep and Arg_Dep as predictors;
DataFile=C:\SuperMixEn Examples\Manual\Continuous\nesarc_112.dat;
Level2ID= PSU;
Dependent= AGE_DEP;
Predictors= intercept M_S_DEP ARG_DEP;
L1Random= intercept;
L2Random= intercept;
FixPatType=Free;
Cov2PatType=Correlated;
AutoCor=None;

```

Buttons: Save As... Close

Model and data description

Numbers of observations

```

-----
Level 2 observations =      371
Level 1 observations =     1698

```

N2	:	1	2	3	4	5	6	7	8
N1	:	8	1	5	5	4	1	4	1
N2	:	9	10	11	12	13	14	15	16
N1	:	3	2	3	2	26	15	3	2
N2	:	17	18	19	20	21	22	23	24
N1	:	6	4	2	2	8	1	8	6
N2	:	25	26	27	28	29	30	31	32

Buttons: Save As... Close

In the next section of the output file as shown above, a description of the hierarchical structure of the data is provided. Data from a total of 371 PSUs and 1,698 respondents were included at levels 2 and 1 of the model. In addition, a summary of the number of respondents nested within each PSU is provided. For example, the PSU with N2:14 had 15 respondents. Note that N2:2 had only 1 observation, which means that the estimation for this PSU might not be reliable.

Descriptive statistics and starting values

The data summary is followed by descriptive statistics for all the variables included in the model. We note that the observed average age at the onset of depression is approximately 31 years.

SuperMix - [nesarc_112.out]

File Analysis Window Help

Descriptive statistics for all variables

Variable	Minimum	Maximum	Mean	Stand. Dev.
Dependent				
AGE_DEP	5.0000	82.0000	31.3975	15.4344
Random-Effects				
intercept (2)	1.0000	1.0000	1.0000	0.0000
intercept (1)	1.0000	1.0000	1.0000	0.0000
Fixed Regressor(s)				
intercept	1.0000	1.0000	1.0000	0.0000
M_S_DEP	0.0000	1.0000	0.2644	0.4412
ARG_DEP	0.0000	1.0000	0.5960	0.4908

Save As... Close

Descriptive statistics are followed by the starting values of the parameters that were used in the initial step of the iterative algorithm. These starting values are obtained by ordinary least squares (OLS) regression, which calculates the estimates by minimizing the sum of the squares of the residuals.

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.

SuperMix - [nesarc_112.out]

File Analysis Window Help

Parameter starting values

Variable	Estimate	Std. Err.	Z-value	p-value
Fixed regressor(s)				
intercept	37.46721	0.58425	64.12858	0.00000
M_S_DEP	-4.90811	0.81540	-6.01926	0.00000
ARG_DEP	-8.00650	0.73286	-10.92493	0.00000
Log Likelihood	=	-156505.9737		
Number of free parameters	=	5		

Save As... Close

The starting values for the random effects are given next.

Variance/covariance components

Level	Estimate	Std. Err.	Z-value	p-value
Level 2				
intercept /intercept	-6.99789	0.10026	-69.79974	0.00000
Level 1				
intercept /intercept	217.71629	0.03861	5638.56252	0.00000

Fixed effects results

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 β_0 , β_1 and β_2 in the model specification. From the z-values and associated exceedance probabilities, we see that all three estimates are highly significant.

o=====o
 | Subset of NESARC data |
 | MS_Dep and Arg_Dep as predictors |
 o=====o

Maximum likelihood estimates

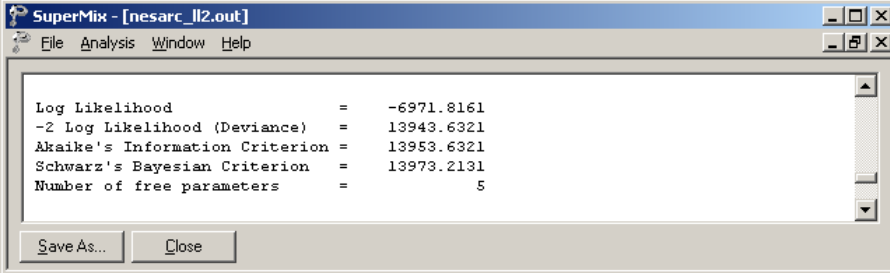
Fixed regressor(s)

Variable	Estimate	Std. Err.	Z-value	p-value
intercept	37.47246	0.59754	62.71126	0.00000
M_S_DEP	-4.89876	0.81387	-6.01913	0.00000
ARG_DEP	-7.99211	0.73247	-10.91118	0.00000

The estimated intercept is 37.472, which means that the average age of the first episode onset of the dysthymia respondents who do not have mother-side depression history and don't argue with others is around 37.4. The estimated coefficients associated with the mother-side history of depression (M_S_DEP) is - 4.898, which indicates that the respondents who have maternal-side depression history tend to get the first episode about five years earlier than those who do not (given the same response on ARG_DEP). The estimate for the indicator of argument involvement (ARG_DEP) shows that a respondent who has argument(s) with others is likely to have a first episode of depression about eight years earlier than a respondent who did not report arguing.

Fit statistics

In addition to the likelihood function value at convergence, a number of related statistical measures for assessing 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. A χ^2 scale factor, with which a χ^2 -value obtained from the difference between two deviance statistics should be multiplied to yield a corrected χ^2 statistic in the case of a weighted analysis, may also be found in this section.



The screenshot shows a window titled "SuperMix - [nesarc_112.out]" with a menu bar containing "File", "Analysis", "Window", and "Help". The main text area displays the following statistics:

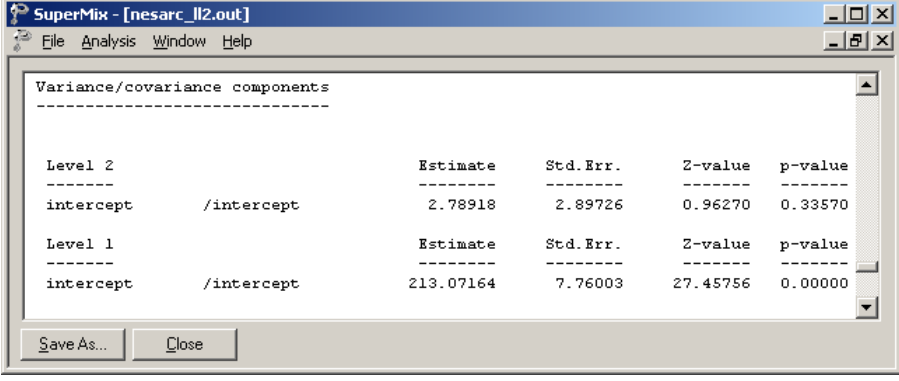
Log Likelihood	=	-6971.8161
-2 Log Likelihood (Deviance)	=	13943.6321
Akaike's Information Criterion	=	13953.6321
Schwarz's Bayesian Criterion	=	13973.2131
Number of free parameters	=	5

At the bottom of the window, there are two buttons: "Save As..." and "Close".

- 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.
- 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.
- 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.

Random effects results

The output for the random part of the model follows, and is shown in the image below. 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. There is no significant variation in the average estimated AGE_DEP at level 2 ($p = 0.33$). This indicates that the expected average age at onset of depression does not differ significantly from PSU to PSU (the level-2 units). Significant differences between the patients (the level-1 units) are reported ($p = 0.00$).



The screenshot shows a window titled "SuperMix - [nesarc_112.out]" with a menu bar (File, Analysis, Window, Help). The main content area displays "Variance/covariance components" with a table of results. The table is organized into two sections: Level 2 and Level 1. Each section has a header row with columns for Estimate, Std. Err., Z-value, and p-value. The Level 2 section shows an intercept estimate of 2.78918 with a p-value of 0.33570. The Level 1 section shows an intercept estimate of 213.07164 with a p-value of 0.00000. Buttons for "Save As..." and "Close" are visible at the bottom of the window.

Level 2		Estimate	Std. Err.	Z-value	p-value
intercept	/intercept	2.78918	2.89726	0.96270	0.33570
Level 1		Estimate	Std. Err.	Z-value	p-value
intercept	/intercept	213.07164	7.76003	27.45756	0.00000

3.1.2.4 Interpreting the results

Model-based graphs

Activate the **Model Setup** window by clicking on it. Using the **Plot Equations for: AGE_DEP** dialog box that appears when the **File, Model-based Graphs, Equations** option is selected, we can graphically depict the trend in expected age at onset of depression, taking the values of the predictors M_S_DEP and ARG_DEP into account. The dialog box below shows the selection of the predictor M_S_DEP. Marking of the plots by ARG_DEP is also requested. Two graphs will thus be displayed on the same set of axes: one for each value of the indicator variable ARG_DEP. By default, all variables present in the model, but not selected for inclusion in the graph, will be assumed to have a value of 0.

The graph below shows the result obtained when the **Plot** button is clicked after completion of the **Plot Equations for: AGE_DEP** dialog box as shown above. We note that patients who did not report arguing are expected to experience onset approximately 8 years later than patients reporting involvement in arguments.

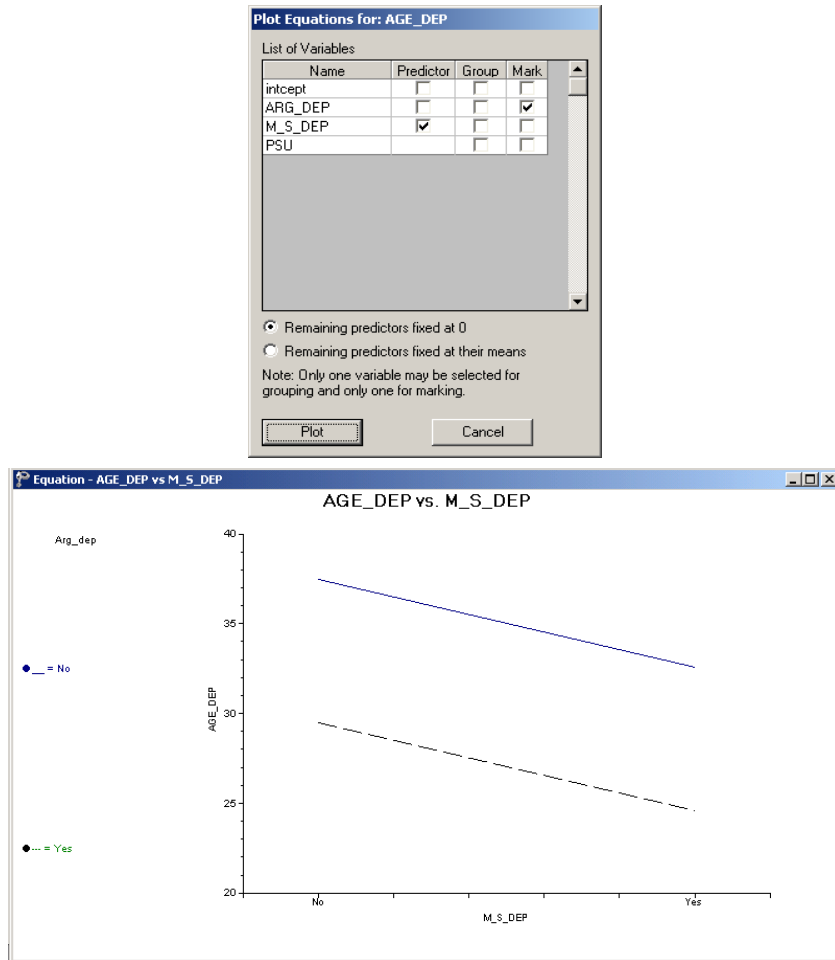


Figure 3.2: Plot of AGE_DEP versus M_S_DEP for 2 groups

A similar plot for the predictor ARG_DEP is given next. This graph was obtained by swapping the positions of the M_S_DEP and ARG_DEP variables on the **Plot Equations for: AGE_DEP** dialog box. Note that patients with maternal-side depression had their first episode approximately 5 years earlier than patients with no history of maternal-side depression. The two graphs shown represent the graphic interpretation of the fixed effect estimates shown previously.

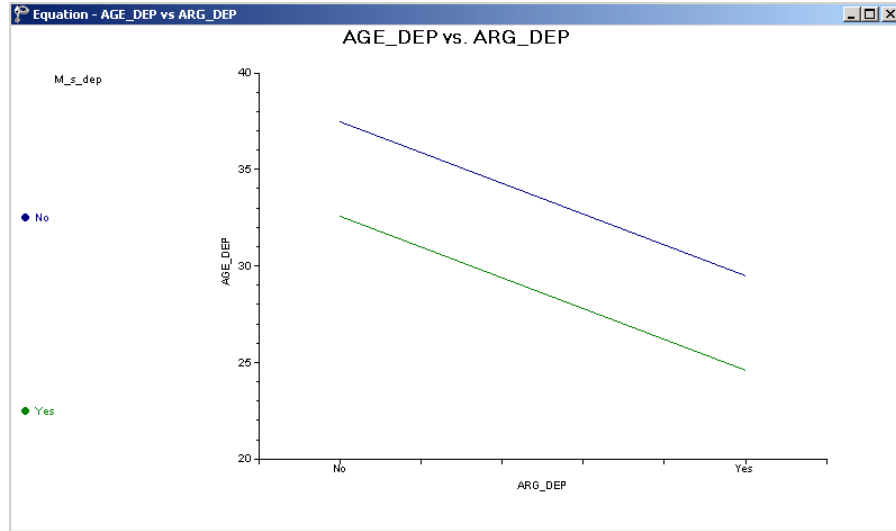


Figure 3.3: Plot of AGE_DEP versus M_S_DEP for 2 groups

ICCs and % variance explained

By calculating the total variation in the age at onset as explained by the current model, we can obtain an estimate of the intraclass correlation coefficient. We first need to calculate the total variation in the outcome variable, which for this model is defined as $\hat{\text{var}}(e_{ij}) + \hat{\text{var}}(v_{i0})$.

The intraclass coefficient is then defined as

$$ICC = \frac{\hat{\text{var}}(v_{i0})}{\hat{\text{var}}(e_{ij}) + \hat{\text{var}}(v_{i0})}$$

and represents the proportion of variation in age at onset that is between the groups (PSUs). An estimate of the percentage of variation in the outcome at a PSU level is obtained as

$$\frac{2.78918}{2.78918 + 213.07164} \times 100\% = 1.29\%$$

indicating that only 1.29% of the total variance is explained at PSU level; the rest of the variance remains at the respondent level.

3.1.3 A 2-level random intercept model with 4 predictors

3.1.3.1 The model

In the previous section, we modeled the outcome variable AGE_DEP as a function of M_S_DEP and ARG_DEP. The extended model discussed in this section takes the ethnicity of a respondent into consideration. The model fitted is expressed as follows:

$$\text{AGE_DEP}_{ij} = \beta_0 + \beta_1 * \text{BLACK}_{ij} + \beta_2 * \text{HISPANIC}_{ij} \\ + \beta_3 * \text{M_S_DEP}_{ij} + \beta_4 * \text{ARG_DEP}_{ij} + v_{i0} + e_{ij}.$$

As before, β_0 denotes the average expected age at the onset of first episode, $\beta_1, \beta_2, \dots, \beta_4$ indicate the estimated coefficients associated with the fixed part of the model, and v_{i0} and e_{ij} represent the random part of the model.

Recall from Section 3.1 that ethnicity was represented by 3 indicator variables, namely WHITEOTH, BLACK and HISPANIC. In the model formulated above, only two of these variables have been included. This was done since the inclusion of all three indicators and the intercept term in the model would cause collinearity between the fixed effects. Any of the respondents will have a value of "1" on one of the three ethnicity indicators. If the values of the indicators are added together in a column-wise fashion, a column of 1s will result. The intercept variable is represented by just such a column of 1s in the program. If a linear combination of a subset of the columns of the design matrix is a constant multiple of another column, a condition referred to as multicollinearity is present and the model cannot be estimated properly.

Consider an example where three respondents, one from each of the three ethnic groups, are considered:

Patient	WHITEOTH	BLACK	HISPANIC	Sum of Ethnicity var.	Intercept
1	1	0	0	1	1
2	0	1	0	1	1
3	0	0	1	1	1

There are two ways in which the model can be formulated to avoid running into this problem. The first is to exclude the intercept and use only the three ethnicity indicators. Such a model, as shown below,

$$\text{AGE_DEP}_{ij} = \beta_0 * \text{WHITEOTH}_{ij} + \beta_1 * \text{BLACK}_{ij} + \beta_2 * \text{HISPANIC}_{ij} \\ + \beta_3 * \text{M_S_DEP}_{ij} + \beta_4 * \text{ARG_DEP}_{ij} + v_{i0} + e_{ij}$$

would not offer an estimated coefficient of the average age at onset. Instead, the expected average age at onset for each of the three ethnic groups may be deduced from the estimated coefficients for WHITEOTH, BLACK and HISPANIC.

Alternatively, one can drop one of the ethnicity indicators from the model while retaining the intercept coefficient. This is what we have opted to do in the current example:

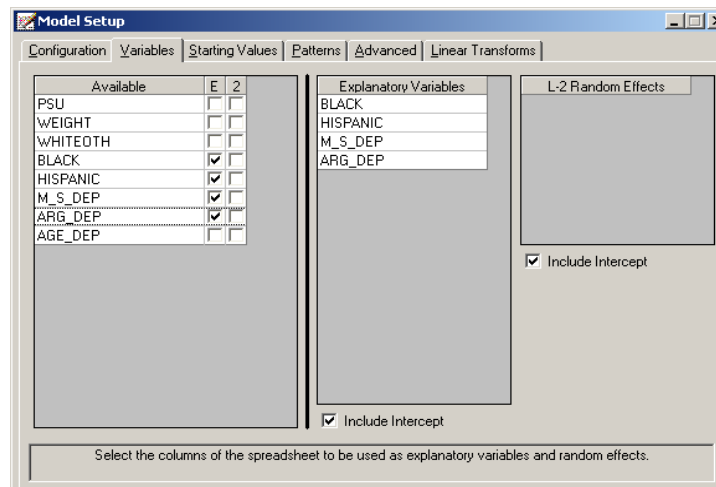
$$\text{AGE_DEP}_{ij} = \beta_0 + \beta_1 * \text{BLACK}_{ij} + \beta_2 * \text{HISPANIC}_{ij} + \beta_3 * \text{M_S_DEP}_{ij} + \beta_4 * \text{ARG_DEP}_{ij} + v_{i0} + e_{ij}$$

In the case of this formulation, the intercept coefficient represents the expected average age at onset for a patient with a value of zero on all the predictors. But if the indicators BLACK and HISPANIC assume a value of 0, it implies that the remaining ethnicity variable WHITEOTH must have a value of 1. As a result, the interpretation of the intercept coefficient would be the expected average onset age for a patient who is white or from some other ethnic origin (excluding African American and Hispanic). This ethnic group thus becomes the reference group in the current analysis. Any of the ethnic groups can be used as the reference group by simply adjusting the coding of the indicator variables; the only proviso being that the group of interest have sufficient data to serve as stable reference group.

3.1.3.2 Setting up the analysis

The SuperMix spreadsheet **nesarc_II2.ss3** and the model specification file **nesarc_II2.mum** discussed in the previous example are used a point of departure.

With the model specification file open, click on the **Variables** tab of the **Model Setup** window. Add the predictors BLACK and HISPANIC to the model by checking the boxes next to these variables in the **E** column, as shown below.

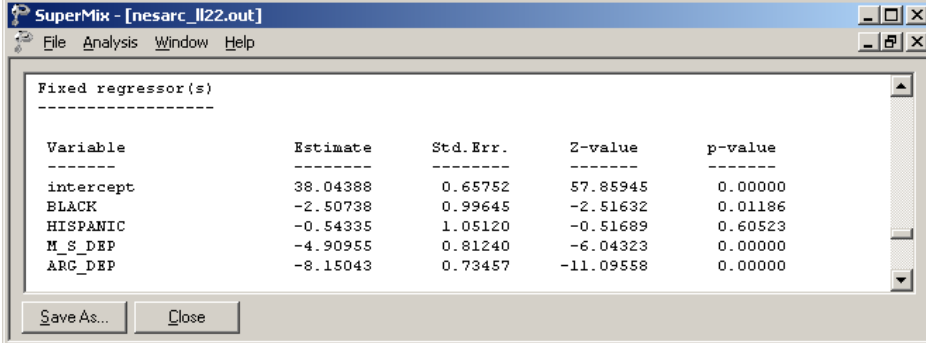


Save the modified model as **nesarc_II22.mum** specification file, and select the **Run** option from the **Analysis** menu to perform the analysis.

3.1.3.3 Discussion of results

Fixed effects results

The maximum likelihood estimates of the coefficients in the fixed part of the model are shown below. Statistically the estimate for HISPANIC is not significant ($p=0.61$). Both estimates for BLACK and HISPANIC are negative, which indicates that African American and Hispanic respondents tend to have an earlier onset of the first episode when compare with patients from white and other ethnic groups.



SuperMix - [nesarc_1122.out]

File Analysis Window Help

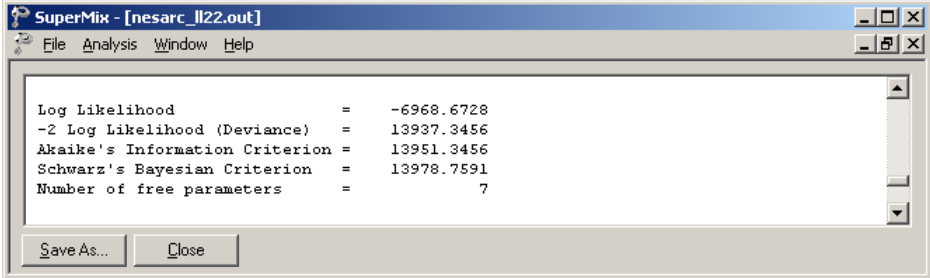
Fixed regressor(s)

Variable	Estimate	Std. Err.	Z-value	p-value
intercept	38.04388	0.65752	57.85945	0.00000
BLACK	-2.50738	0.99645	-2.51632	0.01186
HISPANIC	-0.54335	1.05120	-0.51689	0.60523
M_S_DEP	-4.90955	0.81240	-6.04323	0.00000
ARC_DEP	-8.15043	0.73457	-11.09558	0.00000

Save As... Close

Fit statistics

Fit statistics for the current model are reported as shown below.



SuperMix - [nesarc_1122.out]

File Analysis Window Help

Log Likelihood	=	-6968.6728
-2 Log Likelihood (Deviance)	=	13937.3456
Akaike's Information Criterion	=	13951.3456
Schwarz's Bayesian Criterion	=	13978.7591
Number of free parameters	=	7

Save As... Close

Random effects results

The output for the **random part** of the model is given next.

Variance/covariance components

Level 2		Estimate	Std. Err.	Z-value	p-value
intercept	/intercept	3.31353	2.97617	1.11335	0.26556
Level 1		Estimate	Std. Err.	Z-value	p-value
intercept	/intercept	211.80901	7.72711	27.41115	0.00000

The random intercept effect at level 2 is not significant. As before, most of the variation in scores is found at a respondent level, with only about 2% of the variation remaining at the PSU level.

3.1.3.4 Interpreting the results

Estimated outcomes for different groups

The estimated outcome for any patient can be obtained using the formula

$$\text{AGE_DEP}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 * \text{BLACK}_{ij} + \hat{\beta}_2 * \text{HISPANIC}_{ij} + \hat{\beta}_3 * \text{M_S_DEP}_{ij} + \hat{\beta}_4 * \text{ARG_DEP}_{ij}$$

For a white respondent, the expected AGE_DEP can be calculated as

$$\begin{aligned} \text{AGE_DEP}_{ij} &= \hat{\beta}_0 + \hat{\beta}_3 * \text{M_S_DEP}_{ij} + \hat{\beta}_4 * \text{ARG_DEP}_{ij} \\ &= 38.04388 - 4.90955 * \text{M_S_DEP}_{ij} - 8.15043 * \text{ARG_DEP}_{ij}. \end{aligned}$$

For African American respondents BLACK = 1, and thus the formula used to predict their AGE_DEP scores reduces to

$$\begin{aligned} \text{AGE_DEP}_{ij} &= \hat{\beta}_0 + \hat{\beta}_1 * \text{BLACK}_{ij} + \hat{\beta}_3 * \text{M_S_DEP}_{ij} + \hat{\beta}_4 * \text{ARG_DEP}_{ij} \\ &= 38.04388 - 2.50738 * 1 - 4.90955 * \text{M_S_DEP}_{ij} - 8.15043 * \text{ARG_DEP}_{ij}. \end{aligned}$$

The formula for a patient of Hispanic origin can be derived in a similar way. In 3.1, the same expected ages of the first episode onset for different groups are calculated based on the formulas above.

Table 3.1: Expected AGE_DEP for various groups of patients

Origin	M_S_DEP = No ARG_DEP = No	M_S_DEP = Yes ARG_DEP = No	M_S_DEP = No ARG_DEP = Yes	M_S_DEP = Yes ARG_DEP = Yes
White & Other	38.04	33.13	29.89	24.98
African American	35.54	30.63	27.39	22.48
Hispanic	37.50	32.59	29.35	24.44

The results show that the respondent who has a history of maternal-side depression or gets involved into arguments generally has an earlier onset age for the first episode. For the respondents with the same M_S_DEP and ARG_DEP values, the average first episode onset ages of African American respondents are the lowest. We also conclude that a patient involved in arguments (ARG_DEP = 1) is likely to have an earlier onset age of depression than a patient with maternal-side depression only (M_S_DEP = 1).

Fit statistics and % variation explained

Table 3.2 shows the fit indices for the previous and current models.

TABLE 3.2: Comparison of random intercept models for NESARC data

Fit indices	Model with 2 indicators	Model with 4 indicators	Difference
Log Likelihood	-6971.8161	-6968.6728	
-2 Log Likelihood (Deviance)	13943.6321	13937.3456	6.2865
Akaike's Information Criterion	13953.6321	13951.3456	2.2865
Schwarz's Bayesian Criterion	13973.2131	13978.7591	-5.5460
Number of free parameters	5	7	

The difference in deviances can be used to assess the model fit. This method is valid for nested models. A nested model may be defined as any submodel of a given model that is based on the same number of observations. Given the difference in structure between the 2-level models these models cannot, however, be compared to each other.

The difference in the deviances follows a χ^2 distribution, where the degree of freedom is the difference of numbers of free parameters.

$$(-2 \ln_{model1}) - (-2 \ln_{model2}) \sim \chi^2(d.f.(-2 \ln_{model2}) - (-2 \ln_{model1}))$$

When the deviances of the two models are compared, a χ^2 -statistic of $13943.6321 - 13937.3456 = 6.2865$ with $7 - 5 = 2$ degrees of freedom is obtained. This indicates that the

current model fits the data better. The AIC decreased from 13953.6321 to 13951.3456, and also favors the use of the 4-predictor model. The SBC, however, increased slightly, from 13973.2131 to 13978.7591, and thus favors the model previously fitted as the more parsimonious. The definitions of these indices are given in the discussion of the output of the previous model. Note, however, that the changes in all three criteria are rather small.

The estimated percentages of variation in outcome at respondent level can be calculated using the variance components reported in the random effects part of the output file:

$$\frac{211.80901}{211.80901 + 3.31353} \times 100\% = 98.46\%.$$

Once the additional level-1 predictors are taken into account, there does not seem to be significant random variation in the outcome over the intercepts of the level-2 units. The estimated average onset age of the first episode does not vary significantly from PSU to PSU.