



## Two-level Poisson model

### Contents

3.1	INTRODUCTION .....	1
3.1.1	<i>Poisson distribution</i> .....	1
3.1.2	<i>Adaptive versus non-adaptive quadrature</i> .....	3
3.1.3	<i>The data</i> .....	3
3.1.4	<i>A 2-level Poisson model with 2 predictors</i> .....	5

### 3.1 Introduction

A count variable counts the number of discrete occurrences of a characteristic of interest that takes place during a time interval. Examples are the occurrence of cancer cases in a hospital during a given period of time, the number of cars that pass through a toll station per day, and the phone calls at a call center. The most common distribution for a count variable is the Poisson distribution. Besides the Poisson distribution, negative binomial distributions may also be used to describe the properties of count variables.

#### 3.1.1 Poisson distribution

The Poisson distribution is a discrete probability distribution. It is appropriate for expressing the probability of a number of events occurring in a fixed time period with a known average rate, under the assumption that the occurrences are independent of one another.

The probability of  $k$  occurrences can be expressed as

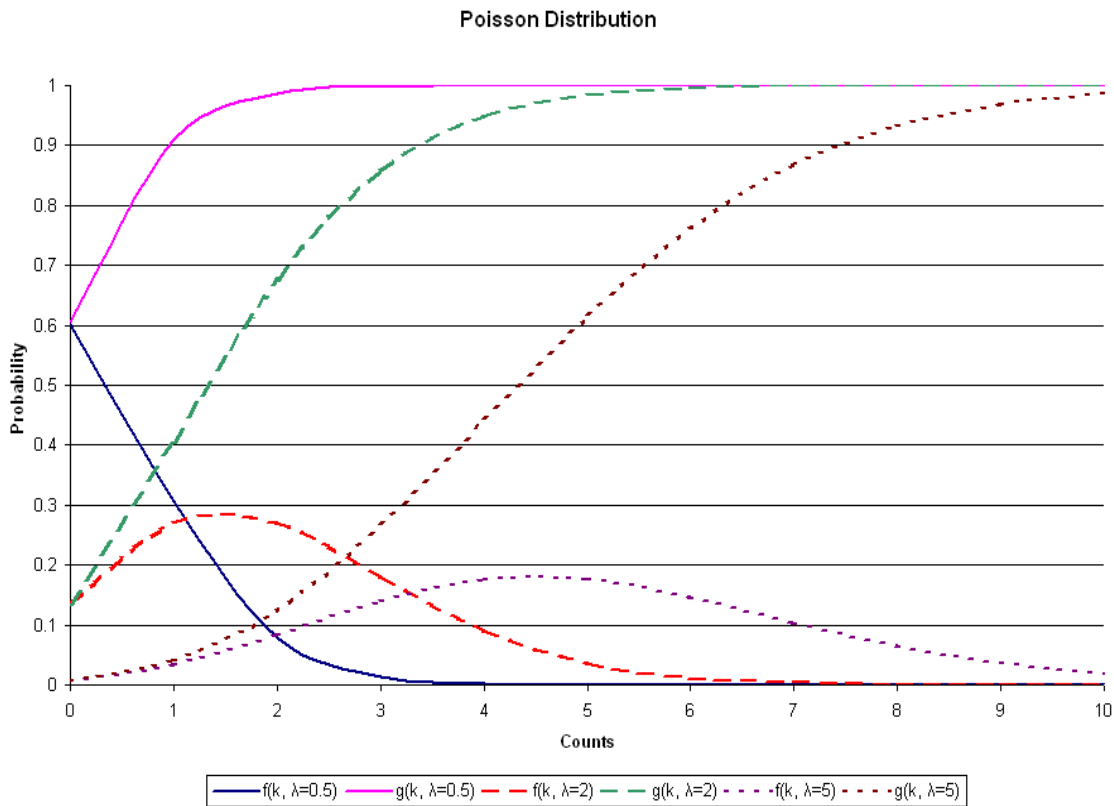
$$f(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \dots$$

where  $k$  is a non-negative integer and  $\lambda$  is a positive real number, which equals the expected number of occurrences during the given interval. The cumulative probability function is

$$\Pr(k; \lambda) = \sum_{i=0}^k \frac{e^{-\lambda} \lambda^i}{i!} \quad \text{for } k = 0, 1, 2, \dots,$$

with the single parameter  $\lambda$ . A Poisson distribution has an important property: the mean number of occurrences  $\lambda$  is equal to the variance:  $E(f) = \text{var}(f) = \lambda$ . Figure 5.1 shows Poisson probabilities  $f(k)$  and cumulative probabilities  $g(k)$  for  $\lambda = 0.5, 2$  and  $5$ .

As shown below, the smaller  $\lambda$  is, the more skewed to the right the probability distribution is. When  $\lambda$  is large, the Poisson distribution is close to the normal distribution.



**Figure 5.1: Poisson probabilities for various values of  $\lambda$**

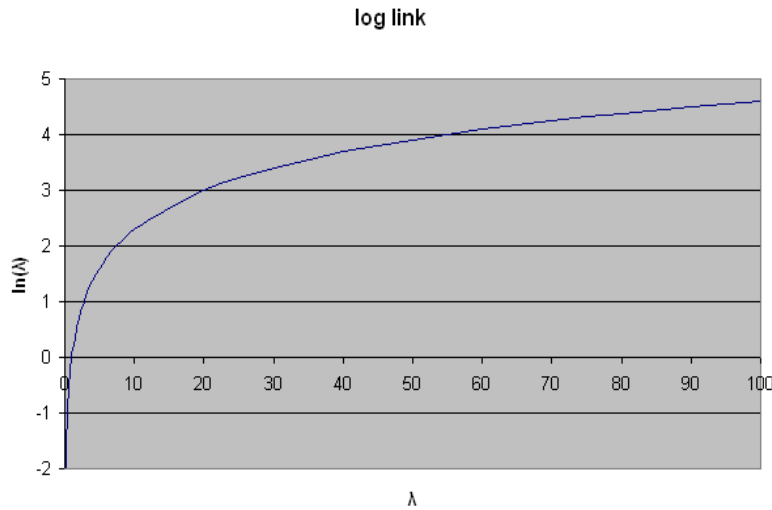
The log link function is generally used for the Poisson distribution. Assume the response measurements for a count variable  $y_1, \dots, y_n$  are independent and

$$y_i \sim \text{Poi}(\lambda_i), \quad \text{where } \lambda_i = e^{\beta_1 x_{i1} + \beta_p x_{ip}}$$

The natural logarithm of the above equation is used to define the link function:

$$\log(\lambda_i) = \beta_1 x_{i1} + \beta_p x_{ip}$$

As shown in Figure 5.2, using the log link function maps the mean of the count variable  $\lambda$  with an open interval  $(0, +\infty)$  to the set of real numbers  $(-\infty, +\infty)$ .



**Figure 5.2: Log link function**

### 3.1.2 Adaptive versus non-adaptive quadrature

Ordinary quadrature is a numeric method for evaluating multi-dimensional integrals. For mixed-effect models with count and categorical outcomes, the log-likelihood function is expressed as the sum of the logarithm of integrals, where the summation is over higher-level units, and the dimensionality of the integrals equals the number of random effects.

A problem with ordinary quadrature is that it assumes a common location and scale for each level-2 unit. This assumption often requires the use of a large number of quadrature points to calculate the log-likelihood and derivatives to an acceptable level of accuracy. To overcome this problem with ordinary quadrature, SuperMix also offers a numeric integration procedure called adaptive quadrature. The adaptive quadrature procedure uses the empirical Bayes means and covariances, updated at each iteration to essentially shift and scale the quadrature locations of each higher-level unit in order to place them under the peak of the corresponding integral. To distinguish between the two quadrature methods, SuperMix uses the terminology non-adaptive quadrature (ordinary quadrature) and adaptive quadrature. To illustrate this, this model will be fitted using the default method of adaptive quadrature.

### 3.1.3 The data

The data set is from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), which was designed to be a longitudinal survey with its first wave fielded in 2001–2002. This data file has been used in some of the examples in Chapter 3, and contains

information on the occurrences of major depression, family history of major depression and dysthymia of 2339 dysthymia respondents. After listwise deletion, the sample size is 1981.

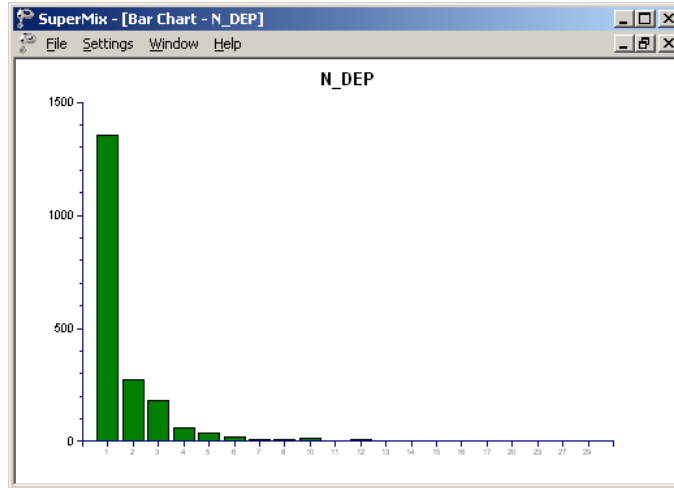
	(A) PSU	(B) FINWT	(C) CONC	(D) AGE D	(E) N DEP
1	1011	7256.15	0	51	1
2	1011	3476.67	1	48	1
3	1011	3052.10	1	59	1
4	1011	1182.03	1	36	2
5	1011	3041.05	1	17	1
6	1011	8342.94	0	16	1
7	1011	6767.06	1	29	1
8	1011	3460.29	1	43	1
9	1015	3167.29	1	55	1
10	1018	1014.56	0	37	1

The variables of interest are:

- PSU denotes the Census 2000/2001 Supplementary Survey (C2SS) primary sampling unit.
- FINWT represents the NESARC weights sample results used to form national level estimates. The final weight is the product of the NESARC base weight and other individual weighting factors.
- CONC\_DEP contains the information captured in field S4CQ3A6 of the NESARC data. It represents the response to the statement "Often had trouble concentrating/keeping mind on things," with 1 indicating "Yes," and 0 indicating "No."
- AGE\_DEP is based on field S4CQ7AR of the NESARC data. It represents the age at onset of first episode.
- N\_DEP is recoded from field S4CQ6A of the NESARC data, and gives the number of depression/dysthymia episodes. This is the count variable we would like to use as outcome variable in the examples to follow.

### 3.1.3.1 Exploring the data

Inspecting the distribution of the intended outcome variable, N\_DEP, before starting with the model is important. In the case of a count variable, this can easily be done by producing a bar chart of the observed frequencies of occurrence captured by the count variable. Select the **File, Data-based Graph, Univariate** option from the main SuperMix window and request a bar chart before clicking the **Plot** button.



**Figure 5.3: Bar chart for count variable N\_DEP**

The frequency bar chart for the count variable N\_DEP shown in Figure 5.3 is obtained. We note that the number of depression episode ranges from 1 to 29, with most respondents having a small number of reported episodes of depression.

### 3.1.4 A 2-level Poisson model with 2 predictors

#### 3.1.4.1 The model

The first model fitted to the data explores the relationship between N\_DEP and the variables indicating concentration (or lack thereof) and age, as represented by the variables CONC\_DEP and AGE\_DEP.

The level-1 model is

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 \times \text{CONC\_DEP}_{ij} + \beta_2 \times \text{AGE\_DEP}_{ij}$$

where the expected number of depression episodes is  $\lambda_{ij} = E(\text{N\_DEP}_{ij})$ .

The level-2 model is

$$\beta_0 = b_{00} + v_{i0}, \quad \beta_1 = b_{10} \quad \text{and} \quad \beta_2 = b_{20}.$$

Another way of writing the combined model is

$$\log(\lambda_{ij}) = b_{00} + b_{10} \times \text{CONC\_DEP}_{ij} + b_{20} \times \text{AGE\_DEP}_{ij} + v_{i0}.$$

In this model,  $e^{b_{00}}$  denotes the average expected count of depression episodes, and  $b_{10}$  represents the estimated coefficient for the respondent's level of concentration.

Taking exponents on both sides, we also have

$$\begin{aligned}\lambda_{ij} &= e^{b_{00}+b_{10}\times\text{CONC\_DEP}_{ij}+b_{20}\times\text{AGE\_DEP}_{ij}+v_{i0}} \\ &= e^{b_{00}} e^{b_{10}\times\text{CONC\_DEP}_{ij}} e^{b_{20}\times\text{AGE\_DEP}_{ij}} e^{v_{i0}}\end{aligned}$$

For a person who had problems concentrating (CONC\_DEP = 1), the expected average number of episodes  $e^{b_{00}}$  is multiplied by  $e^{b_1}$ , compared to an expected count of  $e^{b_{00}}$  for a person for whom CONC\_DEP = 0. Similarly, an increase of one year in age increases the estimated number of episodes by a factor of  $e^{b_{20}}$ . For example, a respondent with concentration problems who is two years older than another respondent who had no concentration problems is expected to have  $e^{b_{00}} e^{b_{10}} e^{2b_{20}}$  episodes compared to only  $e^{b_{00}}$  episodes for the younger person without concentration problems.

The random part of the model is represented by  $e^{v_{i0}}$ , which denotes the variation in average count of depression episodes over PSU and between respondents (or, in other words, over respondents nested within PSU). For a Poisson distribution, the assumption of normality at level 1 is not realistic, as the level-1 random effect can only assume a number of distinct values. Thus, this random effect cannot have homogeneous variance.

### 3.1.4.2 Setting up the analysis

Open the SuperMix spreadsheet **nesarc\_poi.ss3** used during the exploratory analysis. 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 three tabs are used: the **Configuration**, **Variables**, and **Advanced** tabs.

The **Configuration** screen is the first tab on the **Model Setup** window. It enables the user to define the outcome variable and the level-2 and level-3 IDs. Some other settings such as missing values, convergence criterion, number of iterations, etc. can be specified here. For all the available settings, please refer to Section 2.4. To obtain the model we discussed, start by selecting the outcome variable N\_DEP from the **Dependent Variable** drop-down list box. Indicate that it is a count variable by selecting the count option from the **Dependent Variable Type** drop-down list box. Next, describe the hierarchical structure of the data by selecting the level-2 ID, PSU, from the **Level-2 IDs** drop-down list box. Enter a title in the **Title** text boxes, and proceed to the **Variables** screen by clicking on this tab.

Model Setup: nesarc\_poi1.mum

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

Title 1: Level 2 Poisson log model

Title 2: NESARC data

Dependent Variable Type: count      Level-2 IDs: PSU

Dependent Variable: N\_DEP      Level-3 IDs:

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. To include the variables CONC\_DEP and AGE\_DEP as predictor variables, check the **E** check boxes next to the variables' names. Note that, as the variables are selected, the selected variables are listed in the **Explanatory Variables** grid. After selection, the screen below is obtained. Note that the **Include Intercept** check boxes in the **Explanatory Variables** grid and **L-2 Random Effects** are checked by default, indicating that an intercept term will automatically be included in the fixed and random parts of the model.

Model Setup

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

Available	E	2
PSU	<input type="checkbox"/>	<input type="checkbox"/>
WEIGHT	<input type="checkbox"/>	<input type="checkbox"/>
CONC_DEP	<input checked="" type="checkbox"/>	<input type="checkbox"/>
AGE_DEP	<input checked="" type="checkbox"/>	<input type="checkbox"/>
N_DEP	<input type="checkbox"/>	<input type="checkbox"/>

Explanatory Variables

CONC\_DEP

AGE\_DEP

L-2 Random Effects

Include Intercept

Include Intercept

Select the columns of the spreadsheet to be used as explanatory variables and random effects.

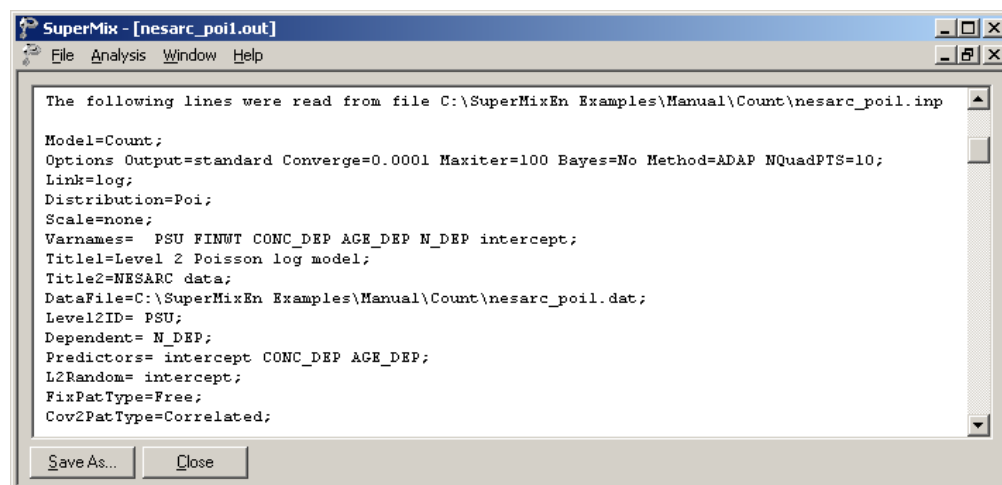
Before running the analysis, the model specifications have to be saved. Select the **File, Save As** option, and provide a name (**nesarc\_poi1.mum**) for the model specification file. Run the analysis by selecting the **Run** option from the **Analysis** menu.

### 3.1.4.3 Discussion of results

Portions of the output file **nesarc\_poi.out** are shown below.

#### Program information and syntax

As shown below, the syntax for the model setup is printed in the output file. The first line of the syntax shows the option **Model = Count**, which indicates the outcome variable is a count variable. The **Options** syntax line corresponds to the settings on the **Configuration** screen. The **Link = log** and **Distribution = Poi** options specify the use of a Poisson distribution with a log link function for the fitted model.



The screenshot shows a window titled "SuperMix - [nesarc\_poi1.out]". The window contains a text area with the following text:

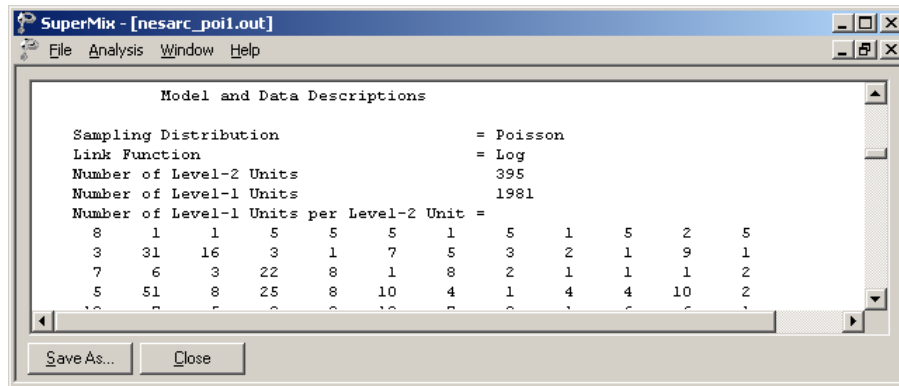
```
The following lines were read from file C:\SuperMixEn Examples\Manual\Count\nesarc_poi1.inp  
Model=Count;  
Options Output=standard Converge=0.0001 Maxiter=100 Bayes=No Method=ADAP NQuadPTS=10;  
Link=log;  
Distribution=Poi;  
Scale=none;  
Varnames= PSU FINWT CONC_DEP AGE_DEP N_DEP intercept;  
Title1=Level 2 Poisson log model;  
Title2=NESARC data;  
DataFile=C:\SuperMixEn Examples\Manual\Count\nesarc_poi1.dat;  
Level2ID= PSU;  
Dependent= N_DEP;  
Predictors= intercept CONC_DEP AGE_DEP;  
L2Random= intercept;  
FixPatType=Free;  
Cov2PatType=Correlated;
```

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

#### Model and data description

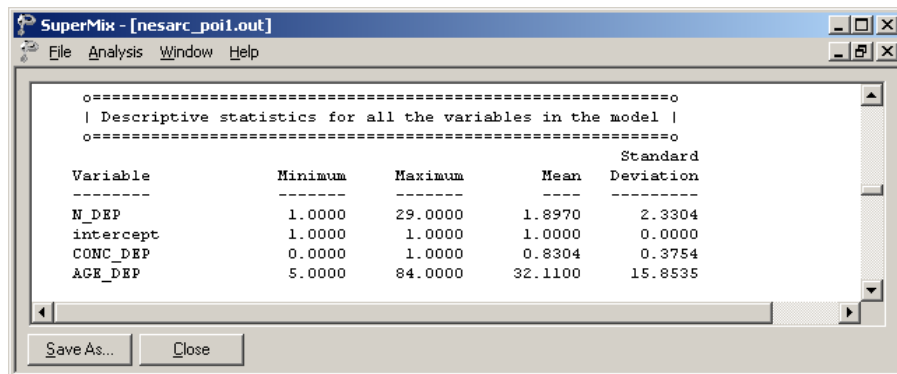
A description of the hierarchical structure follows the syntax: data from a total of 395 PSU and 1981 respondents were included at levels 2 and 1 of the model. In addition, an enumeration of the number of respondents nested within each of the 395 PSUs is provided.





### Descriptive statistics

The data summary is followed by descriptive statistics for all the variables included in the model. The mean of 1.8970 and standard deviation of 2.3304 are reported for the outcome N\_DEP indicating that, on average, 1.8970 episodes of depression were recorded.

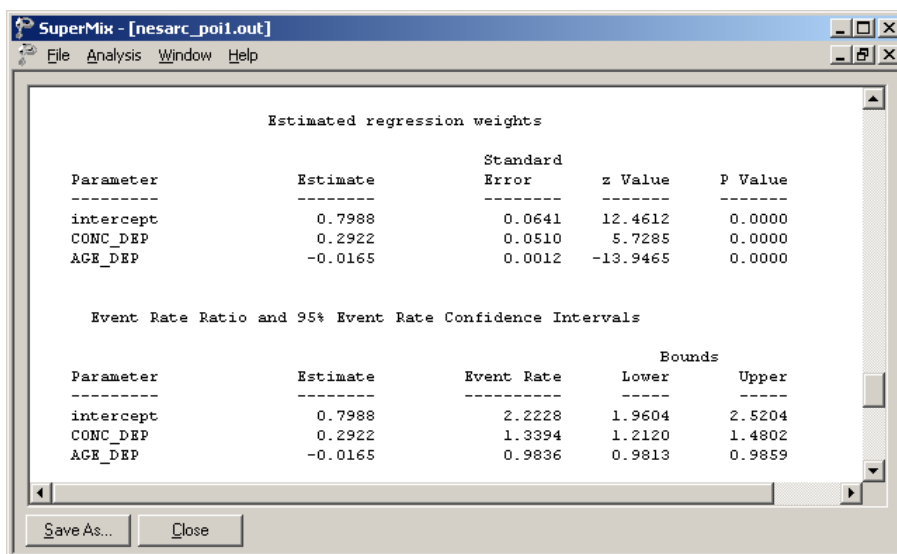
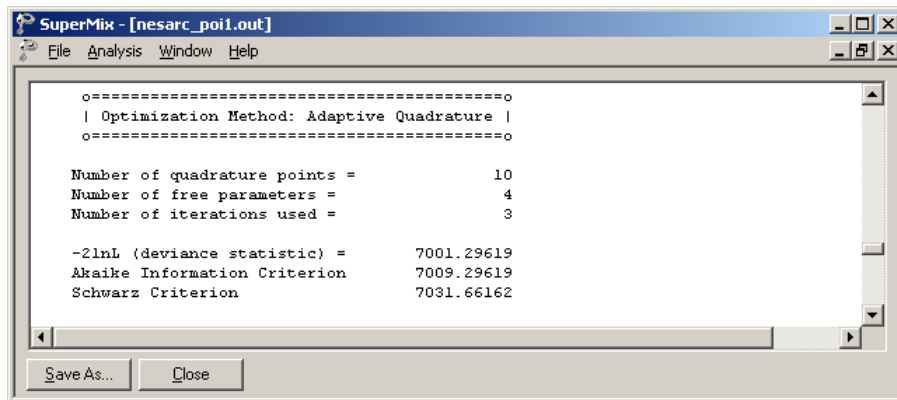


Descriptive statistics are followed by the results for a fixed-effects-only model, *i.e.* a model without random coefficients.

### Fixed effects results

At the top of the final results, the number of iterations required for convergence of the iterative procedure is given.

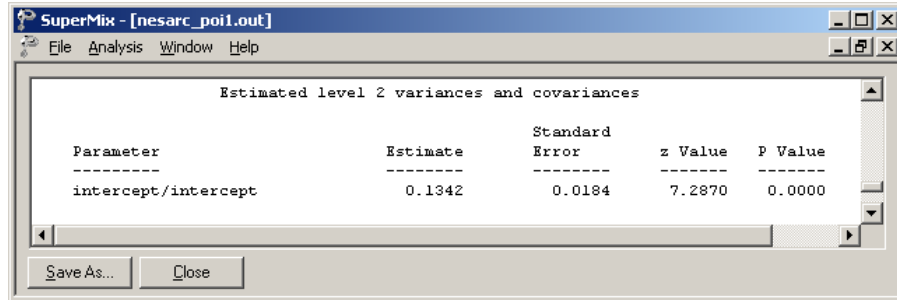
Next, the number of quadrature points per dimension is reported which, in this case, is the default number of points. The log likelihood and the deviance, which is defined as  $-2 \ln L$ , are listed next. For a pair of nested models, the difference in  $-2 \ln L$  values has a  $\chi^2$  distribution, with degrees of freedom equal to the difference in number of parameters estimated in the models compared.



The estimated intercept is 0.7982, which means that the average number of depression episodes is  $e^{0.7982}=2.2215$ , implying that on average the number of episodes is about two. The estimated coefficient for CONC\_DEP is 0.2922, which indicates that respondents who had trouble concentrating on things tended to have  $2.2215e^{0.2922}=(2.2215)(1.3394)=2.9754$  episodes at the same age as respondents without concentration problems. The estimate of the effect of age at the onset of the first episode (AGE\_DEP) shows that the onset age does not affect the number of episodes much, since  $e^{-0.0165}=0.98$ . A slight reduction in the expected number of episodes is expected with increasing age. If one compares two typical respondents with reported concentration problems, but with one respondent ten years older than the other, one would expect the older respondent to have  $(2.2215)(1.3394)e^{10(-0.0165)}=2.5229$  episodes, compared to 2.9268 expected episodes for the younger respondent. In other words, the longer it takes for the first episode to occur, the fewer episodes a respondent is expected to have. Of course, it has to be kept in mind that the younger a respondent is at the first episode, the longer that person must live with the condition and thus the more time there is for subsequent episodes to occur.

## Random effects results

The output for the level-2 random effect variance term follows next. The estimated variation in the average estimated N\_DEP at level 2 is 0.1347, which is highly significant. Respondents are different in terms of their average expected number of episodes, holding all other variables constant.



The screenshot shows a window titled "SuperMix - [nesarc\_poi1.out]" with a menu bar (File, Analysis, Window, Help). The main content area displays "Estimated level 2 variances and covariances" with the following table:

Parameter	Estimate	Standard Error	z Value	P Value
-----	-----	-----	-----	-----
intercept/intercept	0.1342	0.0184	7.2870	0.0000

At the bottom of the window are "Save As..." and "Close" buttons.

## Level-1 variation for Poisson distribution

The variance-to-mean ratio is a measure of the dispersion of a probability distribution:

$$R = \text{variance-to-mean ratio} = \frac{\sigma^2}{\mu}$$

For the Poisson distribution, where the variance equals the mean, this implies  $R = 1$ . Thus, we use a value of one as our level-1 variation. In the cases when over-dispersion ( $R > 1$ ) or under-dispersion ( $R < 1$ ) is assumed, different level-1 variation values will apply. The details of these scenarios are not discussed in this guide.

### 3.1.4.4 Interpreting the results

#### Estimated outcomes for groups: unit-specific results

First, we substitute the regression weights and obtain the following function for  $\log(\hat{N}_{\_DEP_{ij}})$ :

$$\begin{aligned}\log(\hat{N}_{\_DEP_{ij}}) &= \hat{b}_{00} + \hat{b}_{10} \times \text{CONC\_DEP}_{ij} + \hat{b}_{20} \times \text{AGE\_DEP}_{ij} \\ &= 0.7982 + 0.2922 \times \text{CONC\_DEP}_{ij} - 0.0165 \times \text{AGE\_DEP}_{ij}.\end{aligned}$$

For example, at age 40, the estimated  $\log(\hat{N}_{\_DEP_{ij}})$  for a typical respondent who does not often have trouble concentrating ( $\text{CONC\_DEP} = 0$ ), we find that

$$\begin{aligned}
\log\left(\widehat{N\_DEP}_{ij}\right) &= \hat{\beta}_0 + \hat{\beta}_1 \times \text{CONC\_DEP}_{ij} + \hat{\beta}_2 \times \text{AGE\_DEP}_{ij} \\
&= 0.7982 + 0.2922 \times \text{CONC\_DEP}_{ij} - 0.0165 \times \text{AGE\_DEP}_{ij} \\
&= 0.7982 + 0.2922 \times 0 - 0.0165 \times 40 \\
&= 0.1382.
\end{aligned}$$

Keeping in mind that we defined the relationship between  $\lambda$  and the predictors as

$$\log(\lambda_{ij}) = \beta_1 x_{i1} + \dots + \beta_p x_{ip},$$

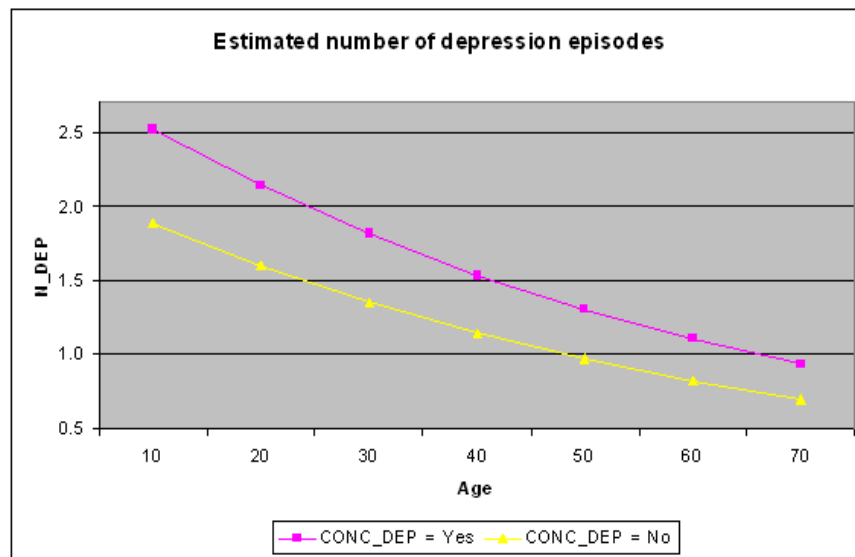
it follows that

$$\hat{\lambda}_{ij} = e^{0.1382} = 1.1482.$$

We can estimate the count of the occurrence of depression episodes for typical individuals of different ages in the same way. Results are summarized in Table 5.1. The results show a decrease in the expected number of episodes with increasing age, regardless of whether they had concentration problems or not.

**Table 5.1: Estimated number of episodes under the Poisson log model**

AGE_DEP	10	20	30	40	50	60	70
CONC_DEP = 1	2.5229	2.1391	1.8138	1.5379	1.3040	1.1056	0.9374
CONC_DEP = 0	1.8836	1.5971	1.3542	1.1482	0.9736	0.8255	0.6999



**Figure 5.4: Expected number of episodes for two groups**

The results in Table 5.1 can also be presented graphically, as shown in Figure 5.4. We clearly see that the correspondents who often had trouble concentrating (CONC\_DEP = 1) have a higher estimated number of depression episodes. It also shows that the number of episodes is expected to decrease as people get older.

### **Level 2 ICC**

The percentage of variance explained over level-2 units, or intraclass correlation coefficient (ICC), is calculated as

$$ICC = \frac{\text{level-2 variation}}{\text{level-1 variation} + \text{level-2 variation}}$$

In this example, under the assumption that the level-1 variation is fixed at a value of one, we have

$$ICC = \frac{0.1347}{1 + 0.1347} \times 100\% = 11.8\%$$

We can conclude that most of the unexplained variation in the outcome (approximately 78%) is between measurements at the lowest level of the hierarchy.