Scientific Software International

Models for count outcomes from the NESARC data

Contents

1. Count variable and its distributions

A count variable is used to count a number of discrete occurrences that take place during a time interval. For example, the occurrence of cancer cases in a hospital during a given period of time, the number cars that pass through a toll station per day and the phone calls at a call center are all count variables.

The most common distribution for a count variable is the Poisson distribution. Besides the Poisson distribution, the negative binomial distribution is also used to model count variables.

Poisson distribution

Poisson distribution is a discrete probability distribution. It is an appropriate distribution to express the probability of a number of events occurring in a fixed time period with a known average rate and are independent of time. The probability with *k* occurrences is

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

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

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

with the single parameter λ. A Poisson distribution has an important property: the mean number of occurrences λ equals the variance $E(f)=\mathrm{var}(f)=\lambda$.

The smaller the value of λ, the more skewed the probability distribution becomes. When λ is large, the Poisson distribution is close to the normal distribution.

Negative binomial distribution

The negative binomial distribution is the probability distribution of the number of failures before the *r*-th success in a Bernoulli process, with probability p of success on each trial.

Log link function

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

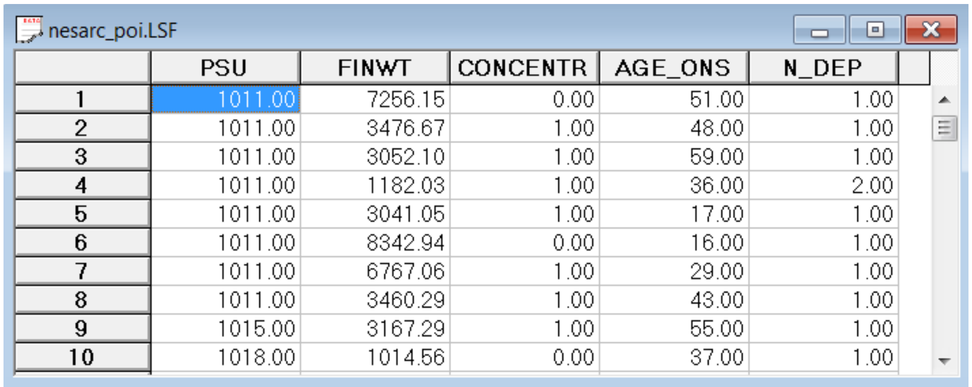
$$y_i \sim Poi(\lambda_i),\qquad where\qquad \lambda_i=e^{\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}$$

To make inference on the unknown parameters, we take the natural logarithm on the above equation.

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

2. 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 contains information on the occurrences of major depression, family history of major depression and dysthymia of 2339 dysthymia respondents. After list-wise deletion, the sample size is 1981.



	PSU	FINWT	CONCENTR	AGE_ONS	N_DEP
1	1011.00	7256.15	0.00	51.00	1.00
2	1011.00	3476.67	1.00	48.00	1.00
3	1011.00	3052.10	1.00	59.00	1.00
4	1011.00	1182.03	1.00	36.00	2.00
5	1011.00	3041.05	1.00	17.00	1.00
6	1011.00	8342.94	0.00	16.00	1.00
7	1011.00	6767.06	1.00	29.00	1.00
8	1011.00	3460.29	1.00	43.00	1.00
9	1015.00	3167.29	1.00	55.00	1.00
10	1018.00	1014.56	0.00	37.00	1.00

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.
- CONCENTR 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_ONS 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. Exploring the data

Inspecting the distribution of the intended outcome variable, N_DEP, before starting with the model is important. The number of depression episode ranges from 1 to 29, with most respondents having a small number of reported episodes of depression.

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 CONCENTR and AGE_ONS.

The level-1 model is

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

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

The level-2 model is

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

Another way of writing the combined model is

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

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

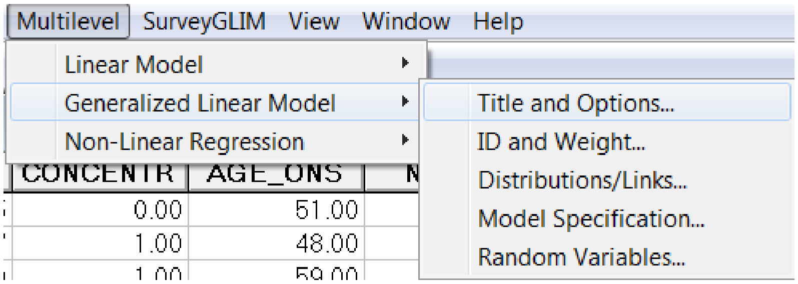
$$\begin{aligned}\lambda_{ij} &= e^{b_{00}+b_{10}\times\text{CONC\_DEP}_{ij}+b_{20}\times\text{AGE\_DEP}_{ij}+v_{i0}} \\ &= e^{b_{00}}e^{b_{10}\times\text{CONC\_DEP}_{ij}}e^{b_{20}\times\text{AGE\_DEP}_{ij}}e^{v_{i0}}\end{aligned}$$

For a person who had problems concentrating (CONCENTR = 1), the expected average number of episodes $e^{b_{00}}$ is multiplied by e^{β_1}, compared to an expected count of $e^{b_{00}}$ for a person for whom CONCENTR = 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.

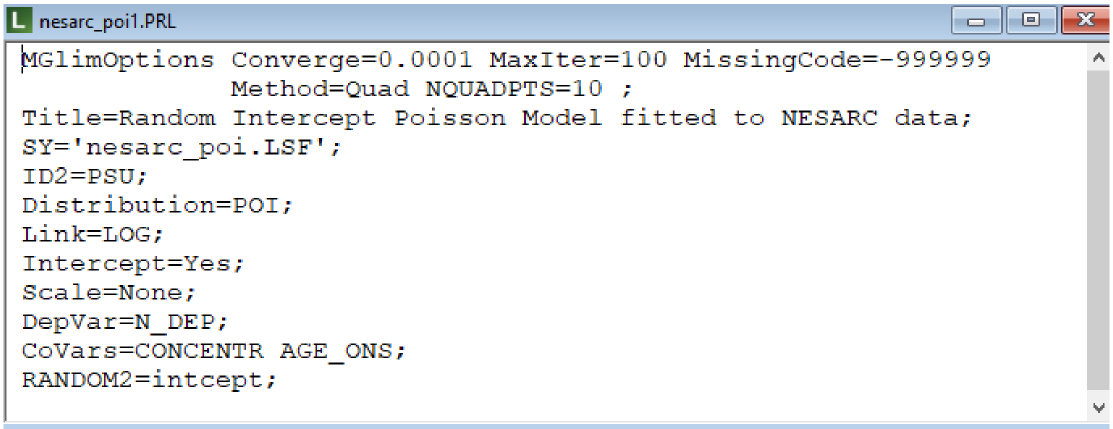
4. Setting up the analysis

Open the LISREL data spreadsheet file **nesarc_poi.lsf** and select the **Multilevel, Generalized Linear Model** option from the main menu bar as shown below.



CONCENTR	AGE_ONS
0.00	51.00
1.00	48.00
1.00	59.00

Proceed to fill in the **Title and Options** (number of quadrature points is 6); **ID and Weight** (Level-2 ID is PSU); **Distributions/Links** (Poisson, log-link); **Model Specification** (Dependent variable is N_DEP, predictors are intcept, CONCENTR and AGE_ONS); and the **Random Variables** dialog (Intercepts only). When done, click the **Finish** button to create the syntax file **nesarc_poi.prl**. Save this file as **nesarc_poi1.prl** using the **File, Save As** option.



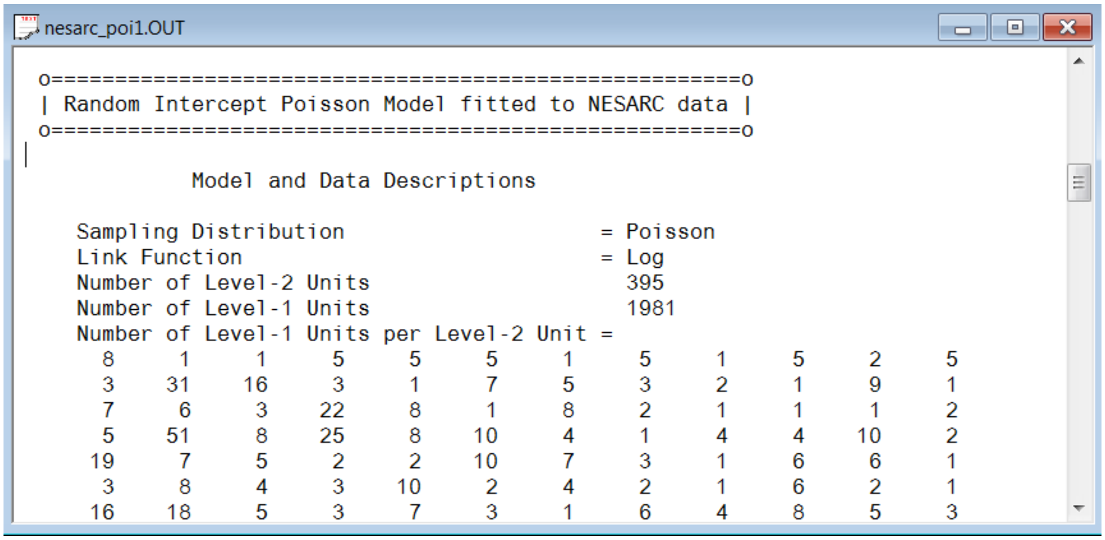
```
MGlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999
             Method=Quad NQUADPTS=10 ;
Title=Random Intercept Poisson Model fitted to NESARC data;
SY='nesarc_poi.LSF';
ID2=PSU;
Distribution=POI;
Link=LOG;
Intercept=Yes;
Scale=None;
DepVar=N_DEP;
CoVars=CONCENTR AGE_ONS;
RANDOM2=intcept;
```

5. Discussion of results

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

Model and data description

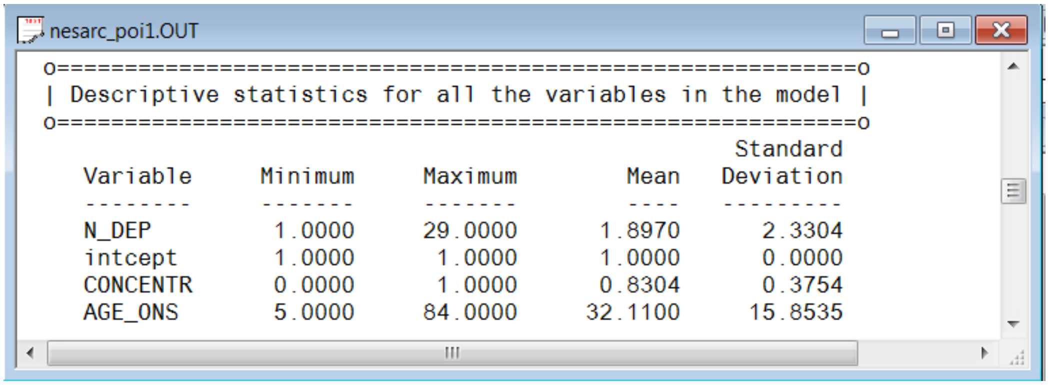
A description of the hierarchical structure follows the syntax: data from a total of 395 PSUs and 1981 respondents were included at levels 2 and 1 of the model. In addition, an enumeration of the number of respondents nested within each of the 395 PSUs is provided.



```
 o======================================================o
 | Random Intercept Poisson Model fitted to NESARC data |
 o======================================================o

               Model and Data Descriptions

   Sampling Distribution                        = Poisson
   Link Function                                = Log
   Number of Level-2 Units                        395
   Number of Level-1 Units                        1981
   Number of Level-1 Units per Level-2 Unit =
      8      1      1      5      5      5      1      5      1      5      2      5
      3     31     16      3      1      7      5      3      2      1      9      1
      7      6      3     22      8      1      8      2      1      1      1      2
      5     51      8     25      8     10      4      1      4      4     10      2
     19      7      5      2      2     10      7      3      1      6      6      1
      3      8      4      3     10      2      4      2      1      6      2      1
     16     18      5      3      7      3      1      6      4      8      5      3
```

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.



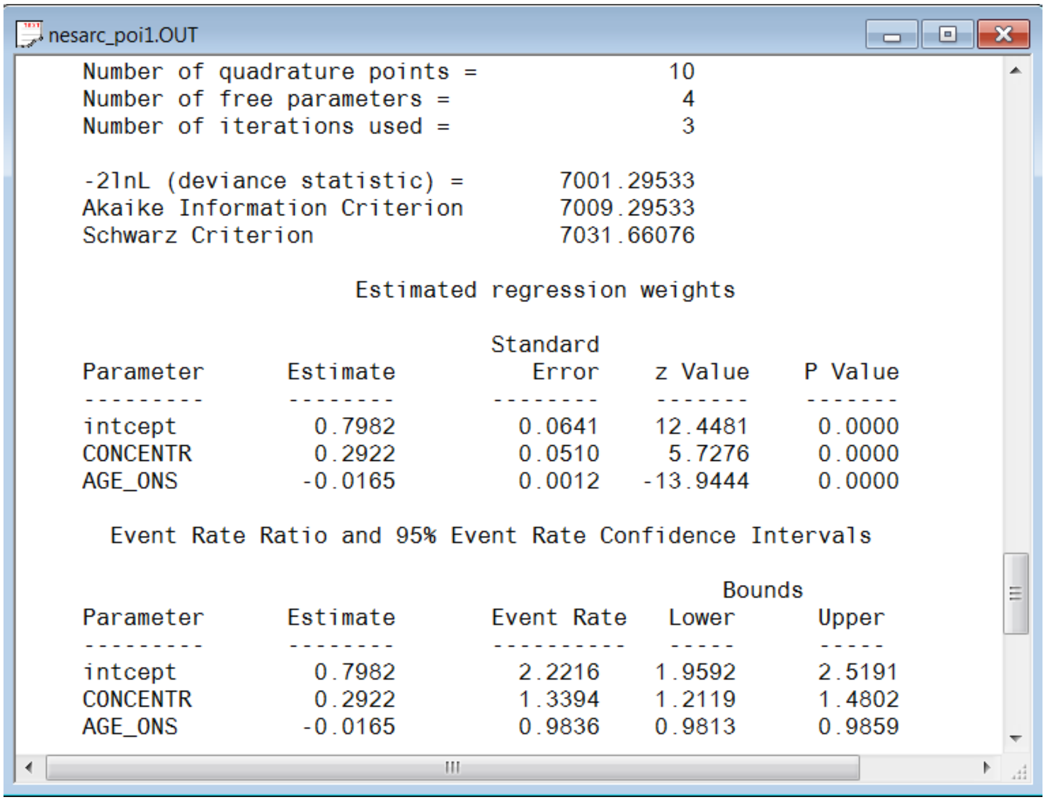
```
o=========================================================o
| Descriptive statistics for all the variables in the model |
o=========================================================o
                                                    Standard
   Variable       Minimum     Maximum        Mean   Deviation
   --------       -------     -------        ----   ---------
   N_DEP           1.0000     29.0000      1.8970      2.3304
   intcept         1.0000      1.0000      1.0000      0.0000
   CONCENTR        0.0000      1.0000      0.8304      0.3754
   AGE_ONS         5.0000     84.0000     32.1100     15.8535
```

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 χ^2 distribution, with degrees of freedom equal to the difference in number of parameters estimated in the models compared.

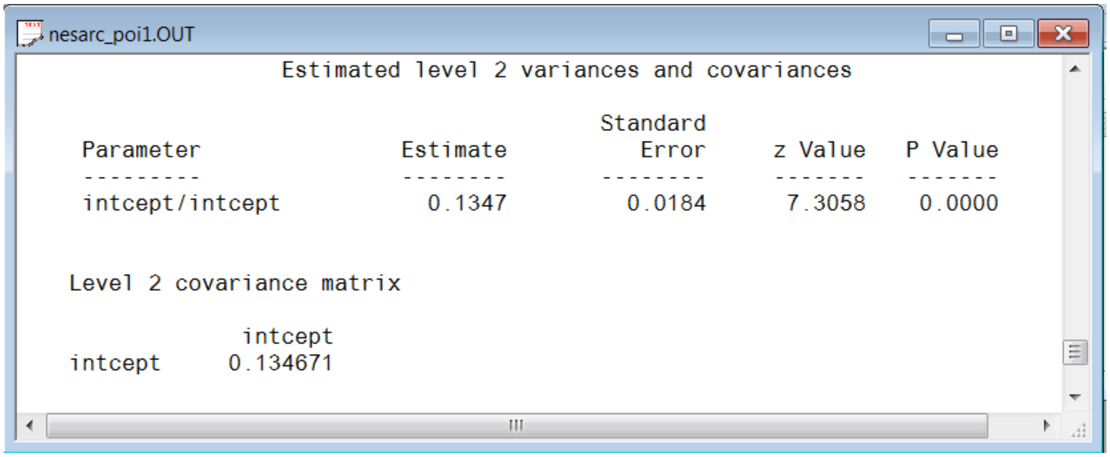


```
   Number of quadrature points =              10
   Number of free parameters =                 4
   Number of iterations used =                 3

   -2lnL (deviance statistic) =       7001.29533
   Akaike Information Criterion       7009.29533
   Schwarz Criterion                  7031.66076

                    Estimated regression weights

                                     Standard
   Parameter       Estimate             Error     z Value     P Value
   ---------       --------          --------     -------     -------
   intcept           0.7982            0.0641     12.4481      0.0000
   CONCENTR          0.2922            0.0510      5.7276      0.0000
   AGE_ONS          -0.0165            0.0012    -13.9444      0.0000

     Event Rate Ratio and 95% Event Rate Confidence Intervals

                                                      Bounds
   Parameter       Estimate          Event Rate   Lower       Upper
   ---------       --------          ----------   -----       -----
   intcept           0.7982            2.2216     1.9592      2.5191
   CONCENTR          0.2922            1.3394     1.2119      1.4802
   AGE_ONS          -0.0165            0.9836     0.9813      0.9859
```

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 CONCENTR 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_ONS) 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.



```
                  Estimated level 2 variances and covariances

                                                Standard
    Parameter                 Estimate             Error      z Value    P Value
    ---------                 --------          --------      -------    -------
    intcept/intcept             0.1347            0.0184       7.3058     0.0000


   Level 2 covariance matrix

                intcept
   intcept     0.134671
```

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.

6. Interpreting the results

Estimated outcomes for groups: unit-specific results

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

$$\log\left(N\_\hat{DEP}_{ij}\right) = \hat{b}_{00} + \hat{b}_{10} \times \text{CONC\_DEP}_{ij} + \hat{b}_{20} \times \text{AGE\_DEP}_{ij}$$

$$= 0.7982 + 0.2922 \times \text{CONC\_DEP}_{ij} - 0.0165 \times \text{AGE\_DEP}_{ij}.$$

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

$$\log\left(N\_\hat{DEP}_{ij}\right) = \hat{b}_{00} + \hat{\beta}_1 \times \text{CONC\_DEP}_{ij} + \hat{\beta}_2 \times \text{AGE\_DEP}_{ij}$$

$$= 0.7982 + 0.2922 \times \text{CONC\_DEP}_{ij} - 0.0165 \times \text{AGE\_DEP}_{ij}$$

$$= 0.7982 + 0.2922 \times 0 - 0.0165 \times 40$$

$$= 0.1382.$$

Keeping in mind that we defined the relationship between λ and the predictors as

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

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 the table below. The results show a decrease in the expected number of episodes with increasing age, regardless of whether they had concentration problems or not.

Estimated number of episodes under the Poisson log model

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

We clearly see that the correspondents who often had trouble concentrating (CONCENTR = 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

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

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.