Scientific Software International

Poisson models for simulated data

1. Introduction

Count variable and its distributions

A count variable is used to count several discrete occurrences that take place during a time interval. For example, the occurrence of cancer cases in a hospital during a given period, 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 Poisson distribution, which was discovered by Poisson. Besides Poisson distribution, binomial and negative binomial distribution also fits the count variables.

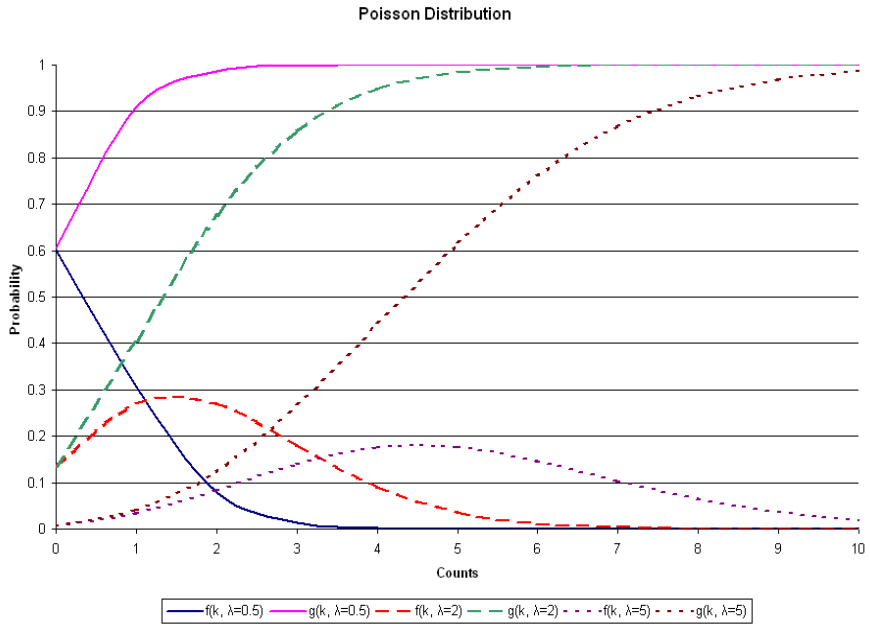
The Poisson distribution is a discrete probability distribution. It is an appropriate distribution to express the probability of several events occurring in a fixed period with a known average rate and are independent of time. The probability with k occurrences is

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

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

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

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 following chart shows Poisson probabilities $f(k)$ and cumulative probabilities $g(k)$ for λ = 0.5, 2 and 5.



As shown in the above chart, the smaller the λ is, the more skewed the probability distribution is. When the λ is large, the Poisson distribution is close to the normal distribution.

Log link function

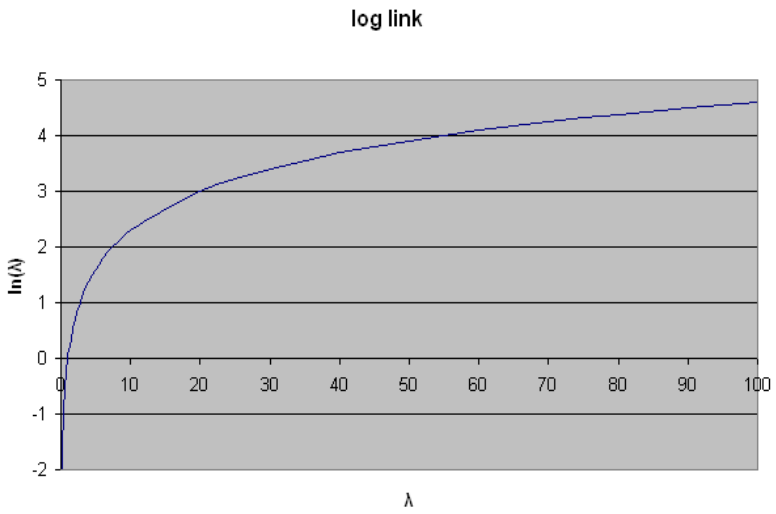
The log link function is generally used for the Poisson distribution. Assume the response measurements for a count variable $y_1,\cdots,y_n$ are independent and

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

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

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

As shown below, by using the log link function map the mean of the count variable λ with an open interval $(0,+\infty)$ to the entire real numbers $\mathbb{R}$.



2. The data

The data set is a simulated data file. The specific data set is provided in the **Multilevel GLIM Examples** folder as **simulpoisson.lsf**. The first portion of this file is shown in the following LSF spreadsheet window.

LISREL for Windows - [simulpoisson.lsf]

File Edit Data Transformation Statistics Graphs Multilevel SurveyGLIM View Window Help

	id3	id2	count	time	time_sq	X1	X2
1	1.00	1.00	6.00	0.00	0.00	0.00	0.00
2	1.00	1.00	6.00	1.00	1.00	0.00	0.00
3	1.00	1.00	6.00	2.00	4.00	0.00	0.00
4	1.00	1.00	4.00	3.00	9.00	0.00	0.00
5	1.00	1.00	6.00	4.00	16.00	0.00	0.00
6	1.00	1.00	6.00	5.00	25.00	0.00	0.00
7	1.00	2.00	7.00	0.00	0.00	0.00	0.00
8	1.00	2.00	10.00	1.00	1.00	0.00	0.00
9	1.00	2.00	8.00	2.00	4.00	0.00	0.00
10	1.00	2.00	11.00	3.00	9.00	0.00	0.00

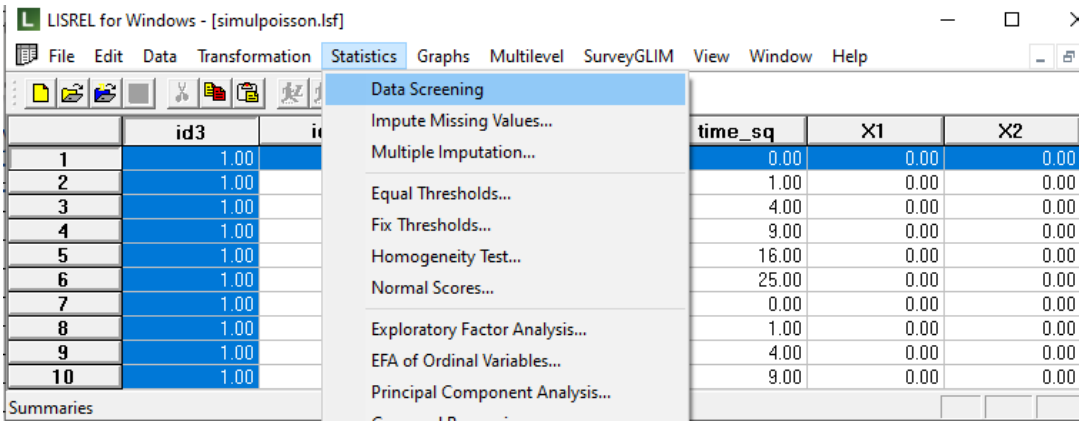
Ready

The variables are:

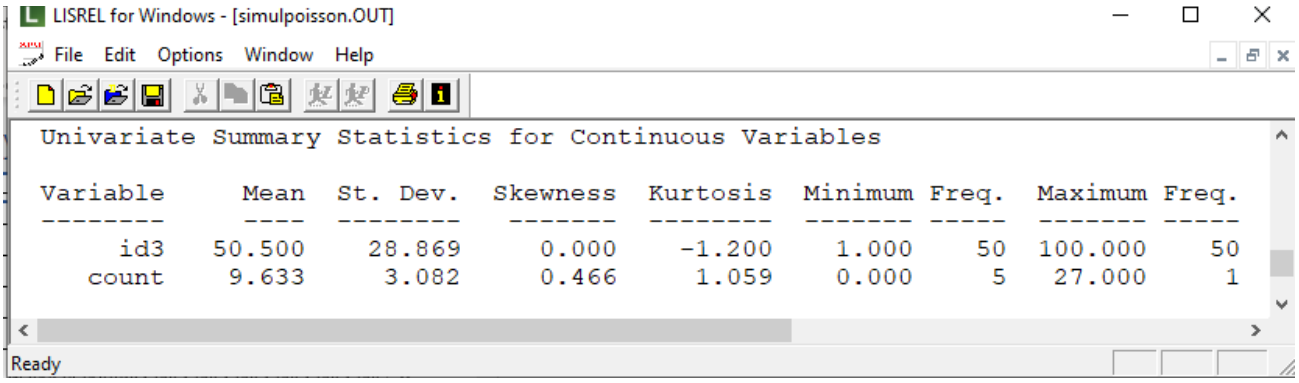
- ID3 is a simulated three level ID.
- ID2 is the level 2 ID.
- COUNT is the target outcome variable.
- TIME is a variable simulated to represent time.
- TIME_SQ is the square of the value for TIME.
- X1 and X2 are two simulated binary dependent variables.

Besides looking at the graphs, the data screening option provided by LISREL is another easy way to overview the data.

Select the **Data Screening** option from the **Statistics** menu as shown below.



We are particularly interested in COUNT. The univariate summary statistics shows that the mean for COUNT is 9.633 and the standard deviation is 3.082. The variance is $3.082^2 = 9.4987$, which is close to the mean. This satisfies the requirements for a Poisson distribution.



3 Level 3 Poisson log model with random intercept model

3.1 The model

In this section, a level 3 random intercept and random slope model is fitted by using this data set. The level 1 model is

Level 1 model $(k = 1, \cdots, n_{ij})$

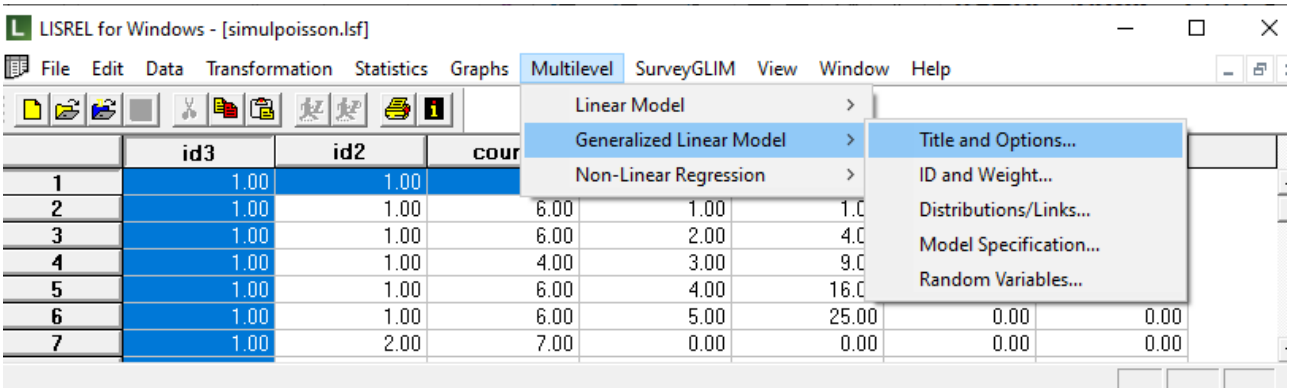
$$\log\left(\lambda_{ijk}\right) = b_{0ij} + b_{1ij} \times time_{ijk} + b_{2ij} \times time\_sq_{ijk} + b_{3ij} \times \mathrm{X1}_{ijk} + b_{4ij} \times \mathrm{X2}_{ijk} + + v_{i0} + u_{ij0} + \varepsilon_{ijk},$$

The random slopes of TIME and TIME_SQ of are considered fixed in the level 2 and the level 3 of the model. Only the intercept is assumed to vary randomly over the higher-level units.

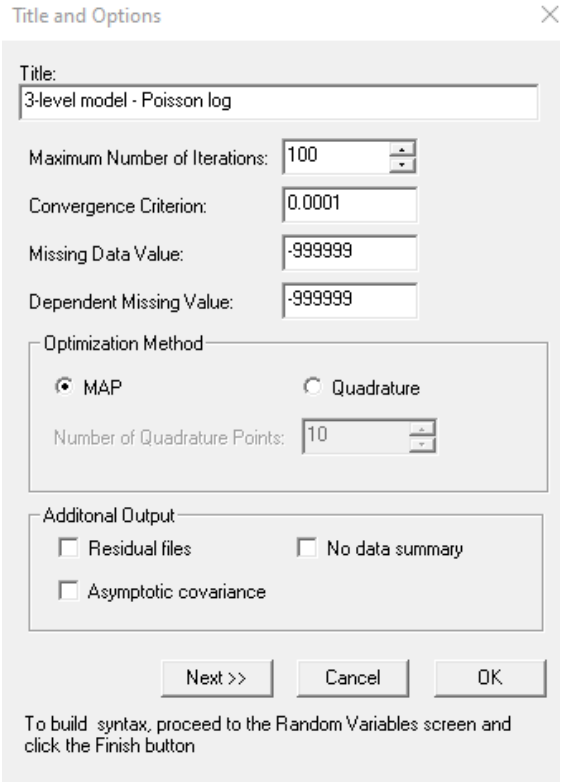
The subscript *k* refers to the *k*-th observation within the *j*-th level 2 unit and the *i*-th level 3 unit. For example, it can be the *k*-th student of the *j*-th class within the *i*-th school. Or it can be the *k*-th patient within the *j*-th hospital in the *i*-th county.

3.2 Setting up the analysis

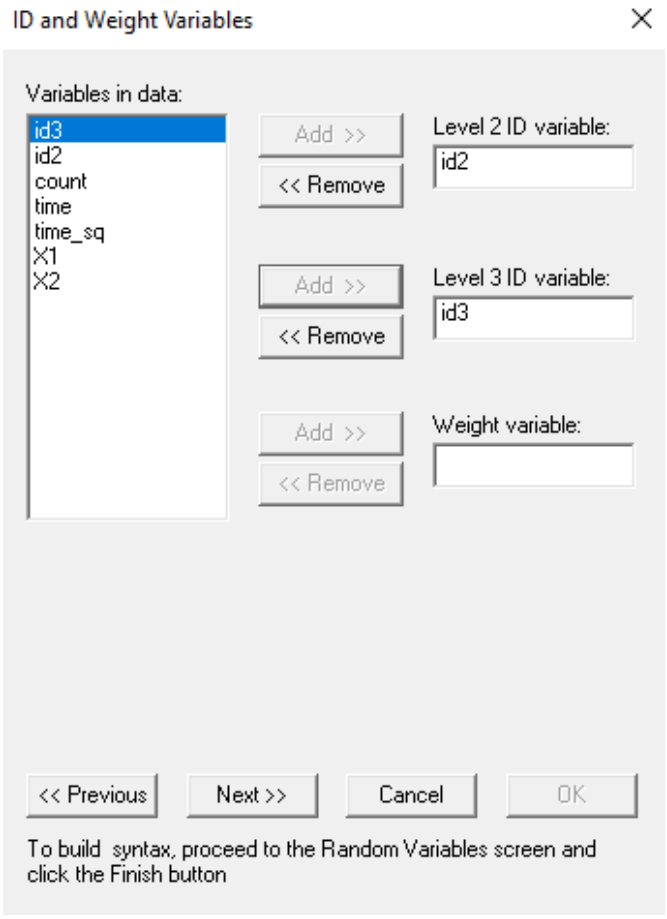
To set up the analysis, we first open the LISREL spreadsheet **simulpoisson.psf**. Select the **Multilevel, Generalized Linear Model**, **Title and Options** option as shown below.



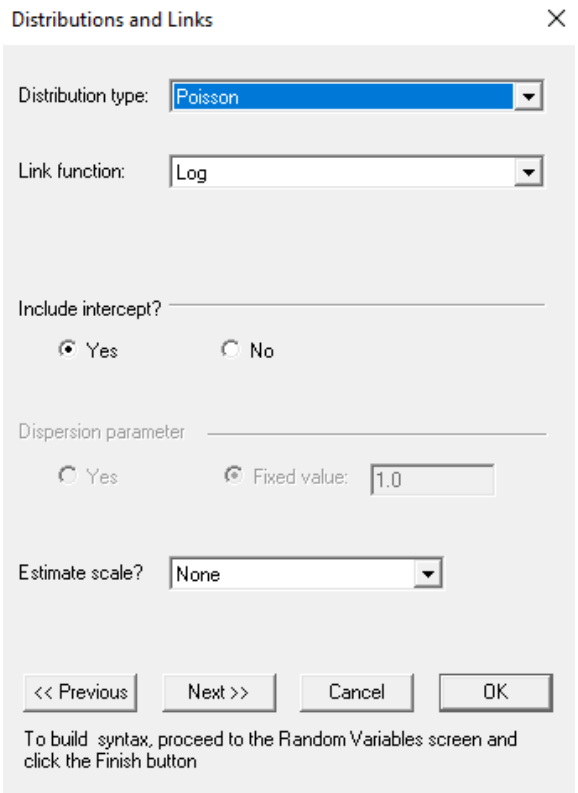
Enter a title for the analysis in the **Title** text boxes (optional) and keep all the other settings as default.



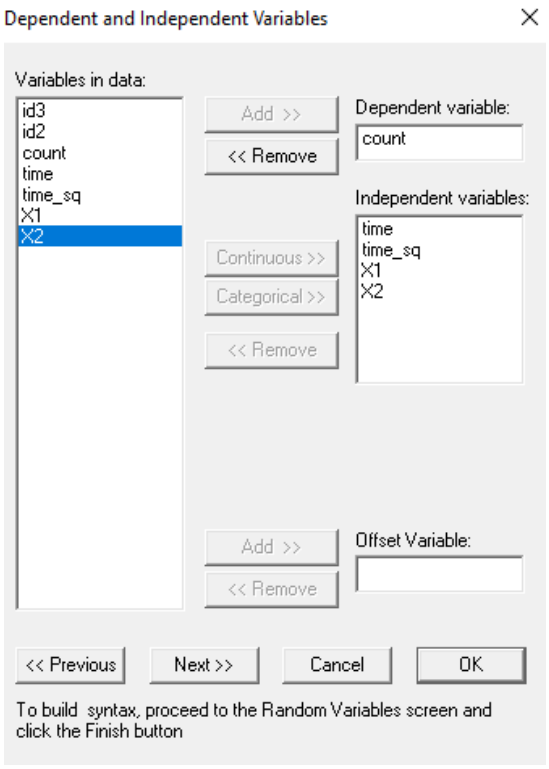
Proceed to the **ID and Weights** screen by clicking on the **Next** button. Select ID2 from the **Variables in data list** and click on the upper **Add** button to select is as the **Level 2 ID variable**. Similarly, select the ID3 and click on the middle **Add** button to select it as the **Level 3 ID variable** to obtain the screen shown below.



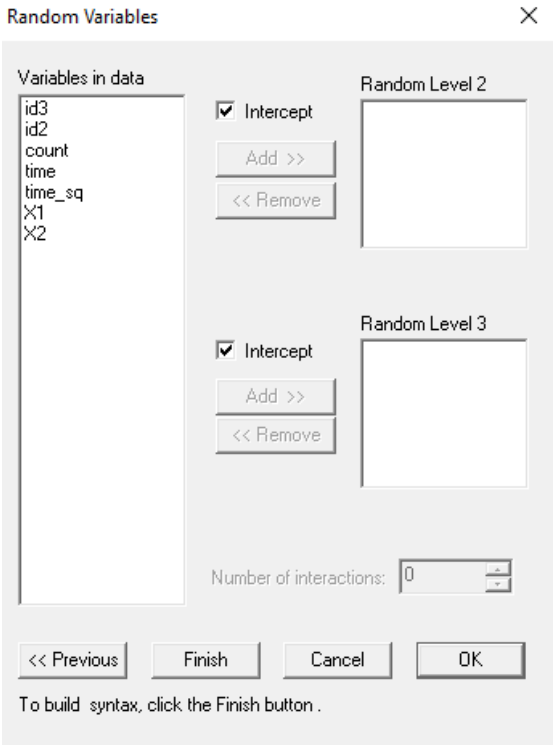
Click on the **Next** button to load the **Distribution and Links** dialog box. Select **Poisson** from the **Distribution type** dropdown list box. By default, the log link function is selected. Keep the other default settings unchanged as shown below and click on the **Next** button.



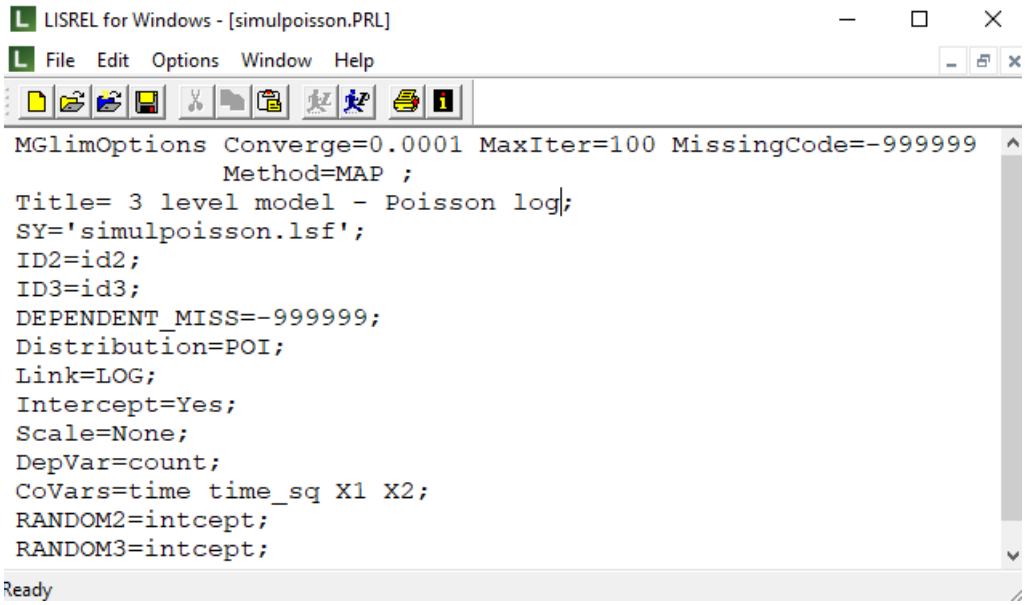
On the **Dependent and Independent Variables** dialog box screen, first select COUNT and click on the upper **Add** button to define it as the **Dependent variable.** Then, select TIME, TIME_SQ, X1 and X2 and click on the **Continuous** button to add the in the **Independent variables** list box as shown below.



Click on the **Next** button to proceed to the **Random Variables** dialog. Keep the **Intercept** check boxes checked to include level 2 and level 3 intercepts.



Click on the **Finish** button to get the PRELIS syntax file (**.prl**) generated corresponding to the settings entered as illustrated above. Select the **File**, **Save As** option, and provide a name (**simulpoisson.prl**) for the model specification file. The default folder for the syntax to be saved to is the same folder as the data file.



The syntax uses the following statements and keywords:

- MGlimOptions indicates that the MGLIM module has been selected. The first two lines together with the Title line correspond to the settings we selected in the **Title and Options** dialog box.
- The SY line indicates the location of the *.lsf* data file.
- ID2 and ID3 denote the level 2 and level 3 ID variables. These are defined in the **ID and Weights** dialog box.
- The lines starting with Distribution, Link and Intercept are set up in the **Distribution and Links** dialog box. It shows that the Poisson distribution with log link function is used in the current model.
- The lines starting with DepVar (which indicates the dependent variable) and CoVars (which indicates any covariate variable) are defined in the **Dependent and Independent Variables** dialog box.
- Finally, the RANDOM2 and RANDOM3 syntax lines correspond to the selections made in the **Random Variables dialog** box. It indicates that random intercepts are assumed for both level 2 and level 3.

Run the analysis by selecting the **Run PRELIS** button. The output file has the same file name as the syntax file with a different extension .**out**. It is saved in the same folder as the syntax file.

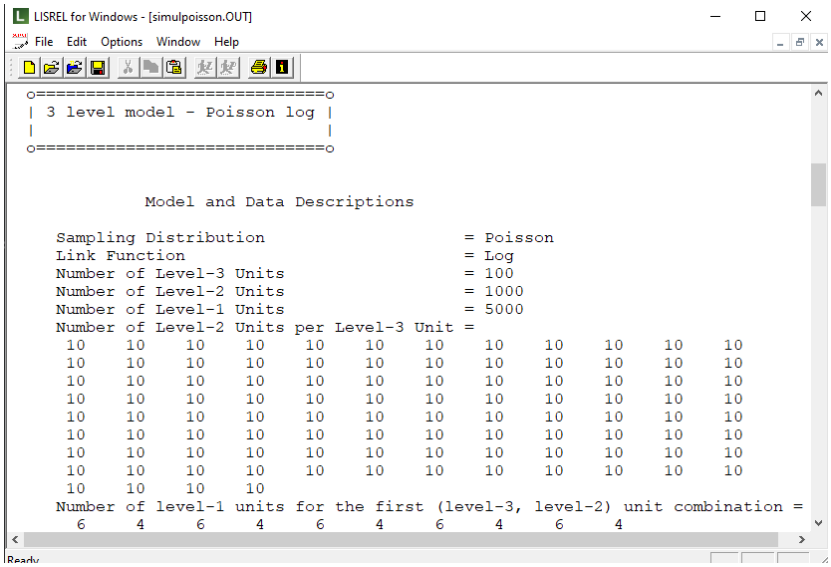
3.3 Discussion of results

Portions of the output file are shown below. Program information is followed by MAPGLIM syntax. This section echoes the contents of the syntax file.

Model and data descriptions

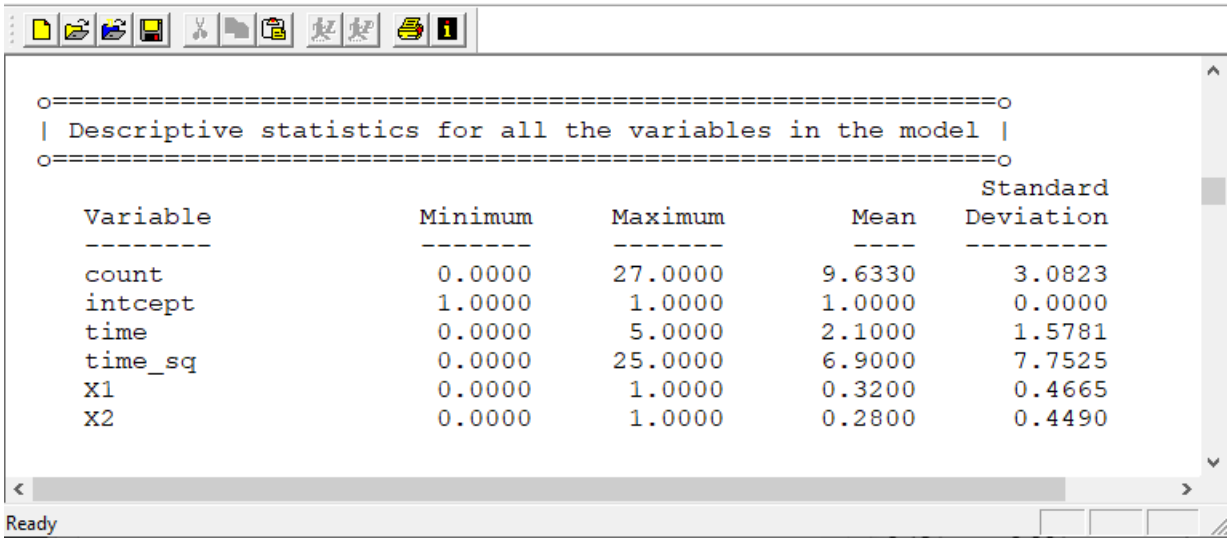
The model and data descriptions are given below the program information and the syntax section. In the next section of the output file as shown above, descriptions of the distribution, the link function. Data from a total

of 10 level 2 units and 2,214 respondents were included at levels 2 and 1 of the model. In addition, a summary of the number of respondents nested within each level 2 unit is provided.



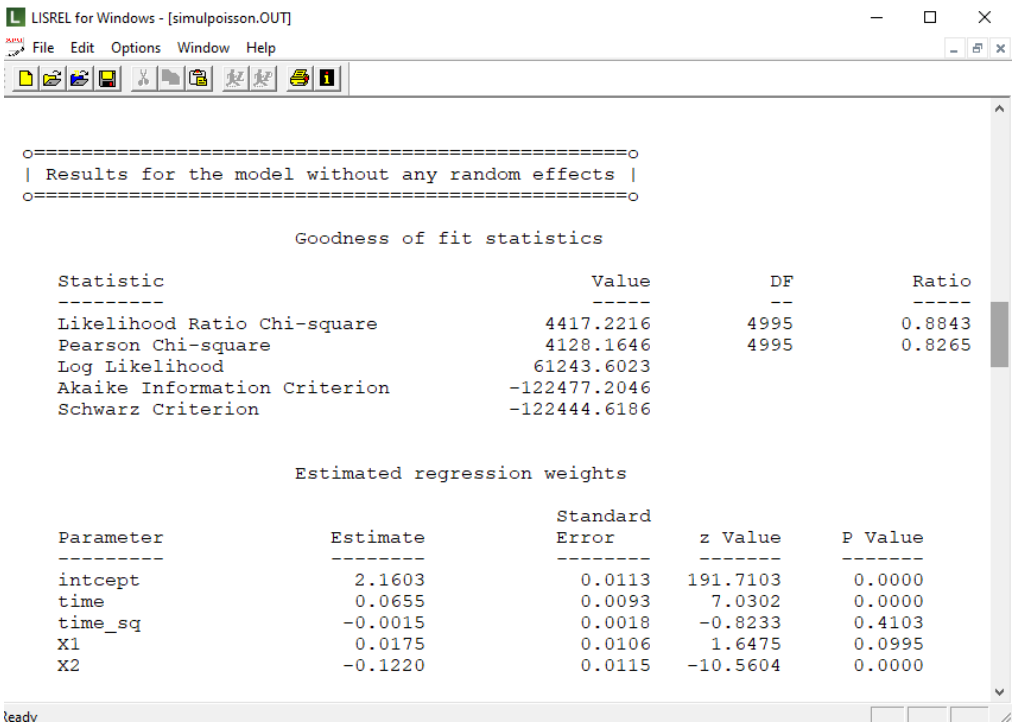
Descriptive statistics

The data summary is followed by descriptive statistics for all the variables included in the model. The mean and the standard deviation are given for each variable.



Results for the model without any random effects

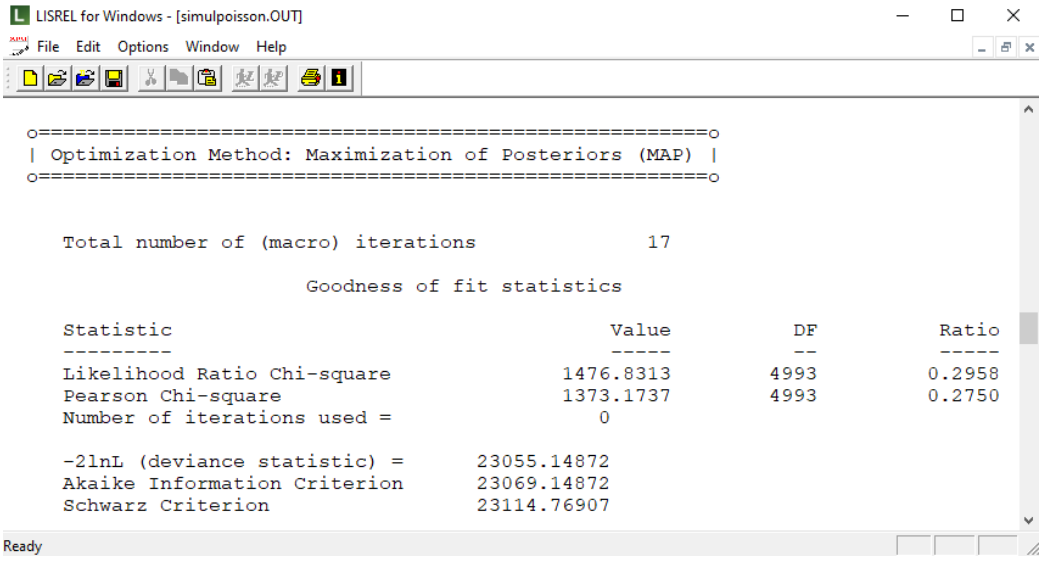
Descriptive statistics are followed by the results for the model without any random effects. These parameters are used in the initial step of the iterative algorithm. They are obtained by ordinary least squares (OLS) regression. The goodness of WLS fit statistics is also given as shown below.



Results for the model with fixed and random effects

Number of iterations and fit statistics

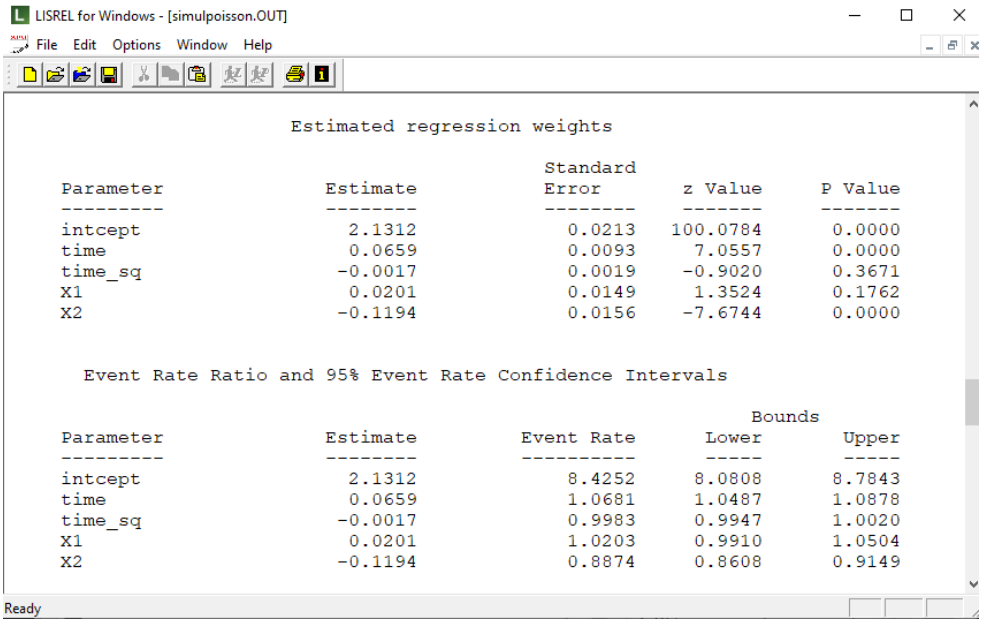
The total number of (macro) iterations is reported. In addition to the likelihood function value at convergence, related statistical measures for assessing model adequacy are available. The most common of these are the likelihood ratio test, Person chi-square, and Akaike's and Schwarz's criteria. Both 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 LISREL output file, all three of these are reported.



Estimated regression weights

The output describing the estimated regression weights after fit statistics is shown next. The estimates are shown in the column with heading Estimate, and correspond to the coefficients β_0, β_1, β_2, β_3 and β_4 in the

model specification. From the *z*-values and associated *p*-values, we see that the quadratic term of TIME is not significant at 10% level.



The estimated intercept is 2.1382, which means that the average logarithm of the estimated intercept is $\log\left(\hat{\lambda}_{ijk}\right) = 2.1382$. The estimated coefficients associated with TIME is 0.0659, which indicates that the estimated logarithm of the mean increases by 0.0659 units for each additional period. The estimate quadratic term of TIME has a negative effect on the estimated logarithm of the outcome variable. X1 has a positive estimated effect on the logarithm of the outcome variable and X2 a negative effect.

Interpret estimated regression weights by using link function

First, we substitute the regression weights and obtain the function for $\hat{\eta}_{ij}$

$$\begin{aligned}\log\left(\hat{\lambda}_{ijk}\right) &= \hat{\beta}_0 + \hat{\beta}_1 \times time_{ijk} + \hat{\beta}_2 \times time\_sq_{ijk} + \hat{\beta}_3 \times \mathrm{X1}_{ijk} + \hat{\beta}_4 \times \mathrm{X2}_{ijk} \\ &= 2.1382 + 0.0659 \times time_{ijk} - 0.0017 \times time\_sq_{ijk} \\ &\quad +0.0201 \times \mathrm{X1}_{ijk} - 0.1194 \times \mathrm{X2}_{ijk}\end{aligned} .$$

At time period 0, for an observation with X1 = 0 and X2 = 0, we have

$$\begin{aligned}\log\left(\hat{\lambda}_{ijk}\right) &= \hat{\beta}_0 + \hat{\beta}_1 \times time_{ijk} + \hat{\beta}_2 \times time\_sq_{ijk} + \hat{\beta}_3 \times \mathrm{X1}_{ijk} + \hat{\beta}_4 \times \mathrm{X2}_{ijk} \\ &= 2.1382 + 0.0659 \times 0 - 0.0017 \times 0 + 0.0201 \times 0 - 0.1194 \times 0 \\ &= 2.1382 \\ &\Rightarrow \\ \hat{\lambda}_{ijk} &= e^{2.1382} = 8.484\end{aligned}$$

Similarly, the calculation of $\hat{\lambda}_{ijk}$ for an observation with X1 = 0 and X2 = 0 at time point 5 is

$$\begin{aligned}\log\left(\hat{\lambda}_{ijk}\right) &= \hat{\beta}_0 + \hat{\beta}_1 \times time_{ijk} + \hat{\beta}_2 \times time\_sq_{ijk} + \hat{\beta}_3 \times \mathrm{X1}_{ijk} + \hat{\beta}_4 \times \mathrm{X2}_{ijk} \\ &= 2.1382 + 0.0659 \times 5 - 0.0017 \times 25 + 0.0201 \times 1 - 0.1194 \times 1 \\ &= 2.3255\end{aligned}$$

$$\Rightarrow$$

$$\hat{\lambda}_{ijk} = e^{2.3255} = 10.232$$

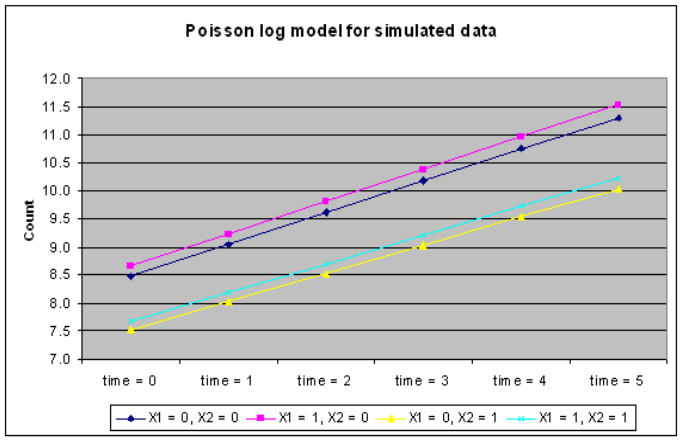
Thus we can conclude that the estimated count of the occurrence for an individual with X1 = 0 and X2 = 0 at the initial point is about 8.5 and the estimated count is about 10.2 for an individual with X1 = 1 and X2 = 1 at the end of the study (TIME =5). Similarly, the estimated count can be calculated for each group at five time points.

Table 5.4: Estimated count for Poisson log model

Code	TIME = 0	TIME = 1	TIME = 2	TIME = 3	TIME = 4	TIME = 5
X1 = 0, X2 = 0	8.484	9.047	9.614	10.182	10.747	11.304
X1 = 1, X2 = 0	8.652	9.226	9.804	10.383	10.959	11.528
X1 = 0, X2 = 1	7.530	8.029	8.533	9.037	9.538	10.033
X1 = 1, X2 = 1	7.679	8.188	8.702	9.216	9.727	10.232

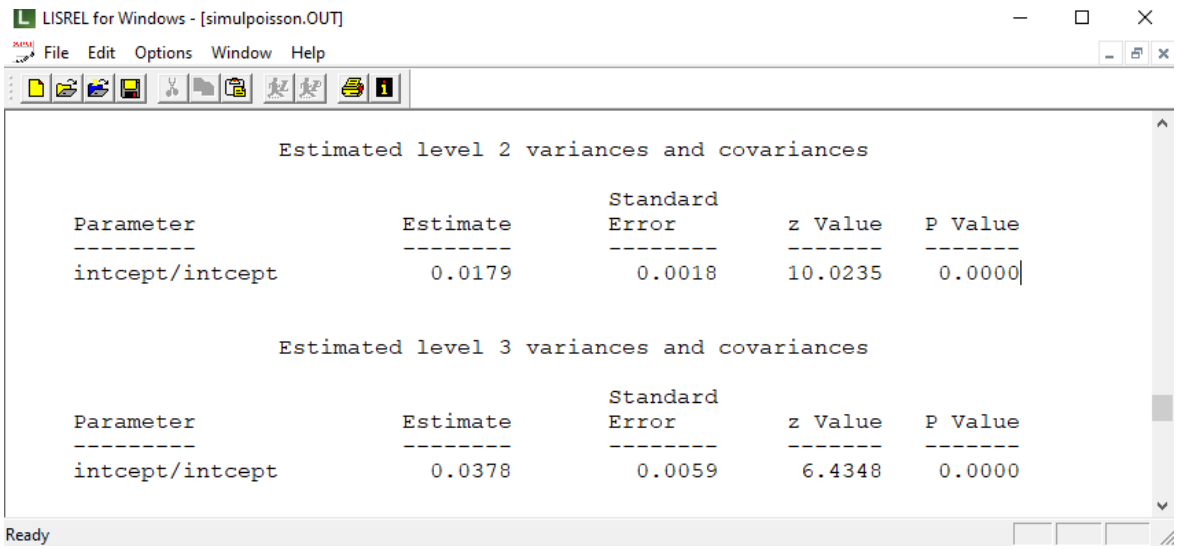
When we represent the findings in the table graphically, we can see clearly that the groups with X1 = 0 has relatively lower estimated counts. The group with X1 = 1 and with X2 = 0 has the highest count for all five time points.

Since the estimate for TIME_SQ is fairly small, the trend for the estimated outcome variable is rather linear.



Estimated level 2 and level 3 variances

The output for the estimated level 2 and level 3 variances is shown in the image below. In our model, *p*-values of intercepts show the random intercept variances are significant at both level 2 and level 3.



ICCs and % variance explained

The intraclass coefficient (ICC), or say the percentage of variance explained by level 2 unit or level 3 unit is calculated by

$$Level\ \ 2\ \ ICC = \frac{\text{level 2 variation}}{\text{level 1 variation} + \text{level 2 variation} + \text{level 3 variation}}$$

or

$$Level\ \ 3\ \ ICC = \frac{\text{level 3 variation}}{\text{level 1 variation} + \text{level 2 variation} + \text{level 3 variation}}$$

Dispersion and level 1 variation

A measure of statistical dispersion is a real number that is zero if all the data are identical and increases as the data becomes more diverse. Standard deviation, variance etc., are all dispersion statistics. When the observed variance is higher than the variance of a theoretical model, over dispersion has occurred. Conversely, under dispersion means that there was less variation in the data than predicted.

Overdispersion is often encountered when fitting very simple parametric models, such as those based on the Poisson distribution. The Poisson distribution has one free parameter and does not allow for the variance to be adjusted independently of the mean. The choice of a distribution from the Poisson family is often dictated by the nature of the empirical data. For example, Poisson regression analysis is commonly used to model count data. If over dispersion is a problem, an alternative model with additional free parameters may provide a better fit. In the case of the Poisson distribution, a Poisson mixture model like the negative binomial distribution can be used instead.

In this example, we assume no over- or under-dispersion is involved. Under such assumption, the level 1 variation is 1.

Level 2 and level 3 ICCs

Total variation is calculated as

$$\begin{aligned}\text{Total variation} &= \text{level 1 variation} + \text{level 2 variation} + \text{level 3 variation} \\ &= 1 + 0.0179 + 0.0378 = 1.056\end{aligned}$$

The level 2 ICC is

$$\text{Level 2 ICC} = \frac{\text{level 2 variation}}{\text{total variation}} = \frac{0.0179}{1.056} = 1.695\%$$

thus we conclude that about 1.4% of the variation can be explained by level 2 units. Similarly, we find that the level 3 variation is

$$\text{Level 3 ICC} = \frac{\text{level 3 variation}}{\text{total variation}} = \frac{0.0378}{1.056} = 3.580\%$$

so that about 3.6% of the variation in COUNT is estimated to be at level 3.

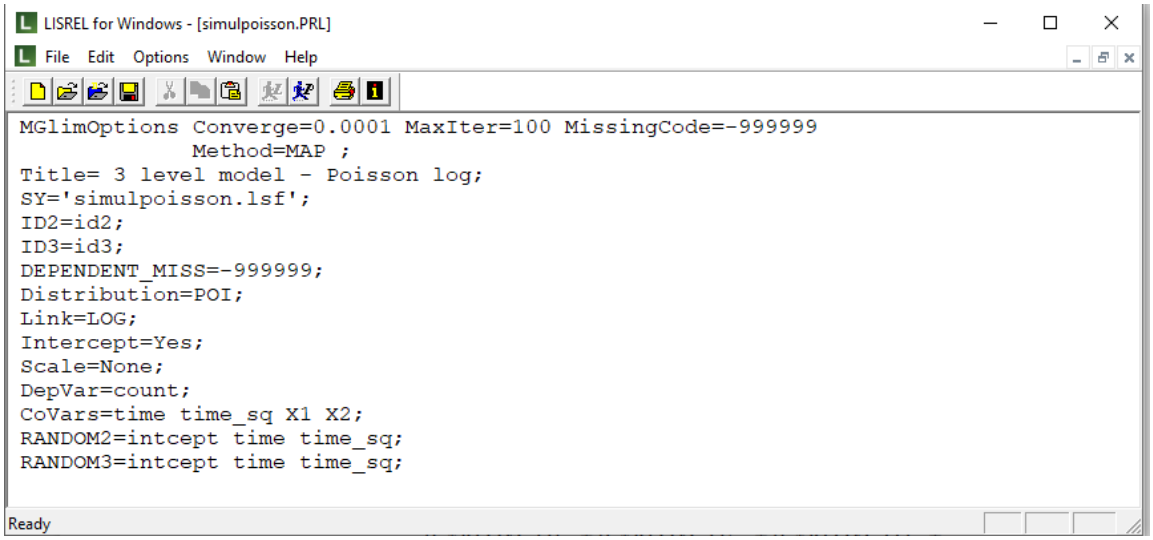
4. Level 3 Poisson log model with random intercept and random slope

4.1 The model

The first model fitted to the depression data is a three-level model that explores the relationship between COUNT and TIME and dependent variables. Retaining the same level-1 model as before, we extend the model at levels 2 and 3 to also include a random TIME and TIME _sq slope.

4.2 Setting up the analysis

Reopen the syntax file **simulpoisson.prl** and add the text TIME TIME_sq to the RANDOM2 and RANDOM3 statements to obtain the syntax shown below.



The syntax shows that only the RANDOM2 and RANDOM3 syntax lines are changed. All the rest of the model setup is the same as the previous model.

Run the analysis by selecting the **Run PRELIS** button to generate the output file.

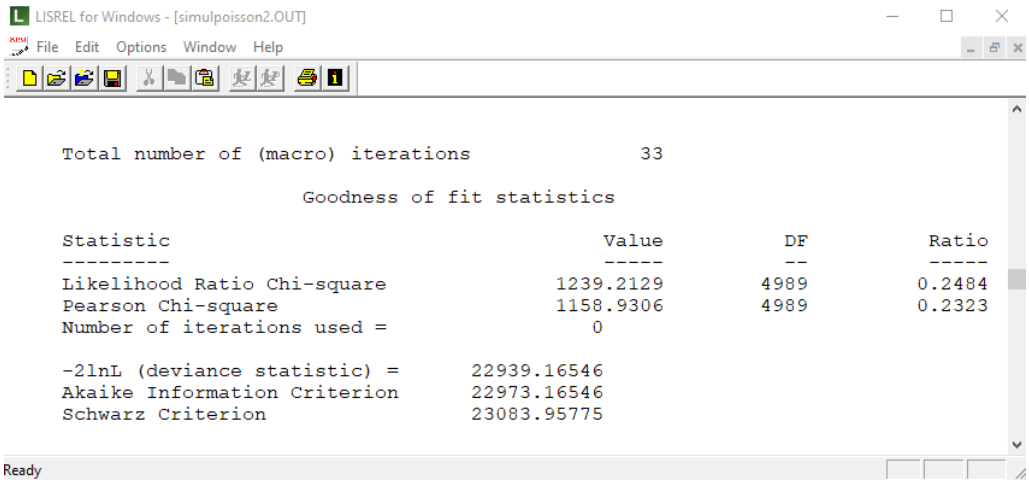
4.3 Discussion of results

Portions of the output file are shown below. Some sections of the output file, such as the model and data descriptions, the descriptive statistics for all the variables and the results without any random effects are exactly the same as the previous model, so we will not repeat.

Results for the model with fixed and random effects

Number of iterations and fit statistics

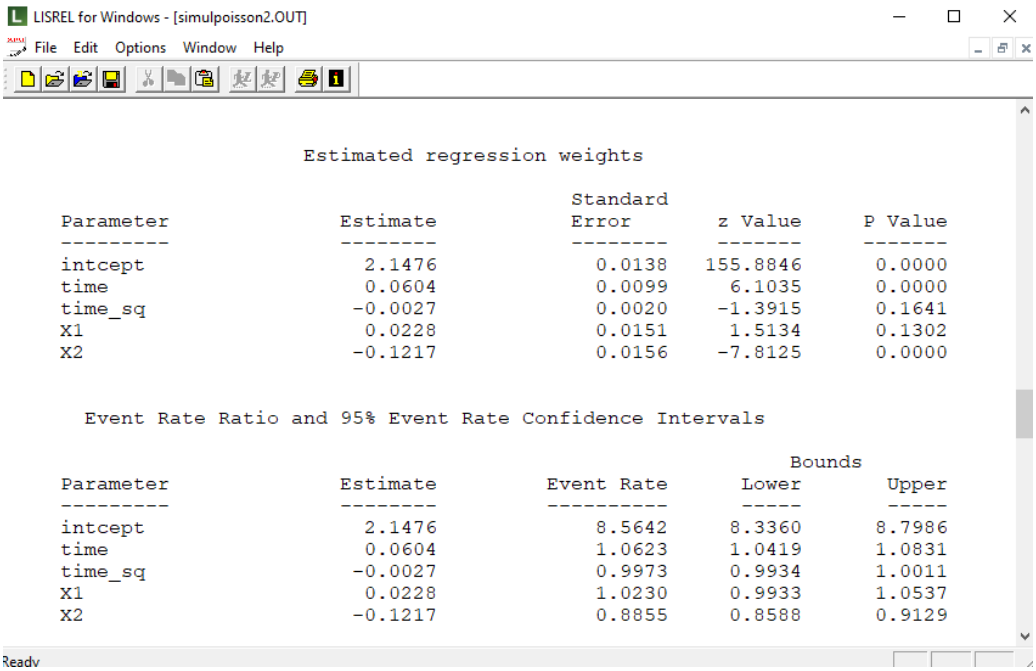
A model comparison section is given later, using the fit statistics of this model and the previous model.



Estimated regression weights

The output describing the estimated regression weights after fit statistics is shown next. The estimates are shown in the column with heading Estimate, and correspond to the coefficients β_0, β_1, β_2, β_3 and β_4 in the

model specification. From the z-values and associated p values, we see that all the quadratic terms of time are not significant at 10% level.



The estimated intercept is 2.1476. Like the previous example, the estimated coefficients associated with TIME is positive; the estimate quadratic term of TIME has a negative effect on the estimated logarithm of the outcome variable. X1 has positive estimated effect on the logarithm of the outcome variable and X2 has a negative estimated effect.

Model comparison

Nested models

Two models are nested if both contain the same terms and one has at least one additional term. Considering the two Poisson log models we had, though both of them have the same level 1 model, the level 2 and level 3 models are different.

The likelihood ratio chi-square and the Pearson's chi-square values can be used to conduct the test. The difference in the deviances follows a χ^2 distribution, where the degree of freedom is the difference of numbers of free parameters.

$$\left(chi-sq_{model1}\right)-\left(chi-sq_{model2}\right)\sim\chi^2\left(d.f.\left(\text{model 2}\right)-\left(\text{model 1}\right)\right)$$

From the output files of the two models fitted, the difference of chi-square ratios can be obtained as χ^2 =1476.83 - 1239.22 = 237.61 with $4493-4989=4$ degree of freedom, which is highly significant.