Scientific Software International

Poisson model for the NIH data

Contents

1. Introduction

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. In this guide models for the Poisson and negative binomial distributions will be discussed.

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 $E(f) = \text{var}(f) = \lambda$.

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.

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

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

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

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

2. The data

The data set comes from the data library of the National Health Interview Survey (NHIS). The NHIS is a national longitudinal health survey. During 2002, background data and data on the health conditions of a sample of 28,737 participants were obtained. The 2002 sample was stratified into 64 strata and into 601 PSUs. Using this data, we created a subset consisting of 53 PSUs (the level-2 units). A partial list of the data is given below in the form of a LISREL spreadsheet file, named **nih_subset.lsf**.

nih_subset.LSF

	PSUS	PATWT	PASTVIS	NUMMED	AGE	SEX	PRIMCARE
1	5.00	42365.00	4.00	6.00	72.00	0.00	0.00
2	5.00	62895.00	4.00	5.00	69.00	1.00	0.00
3	5.00	62895.00	3.00	0.00	11.00	0.00	0.00
4	5.00	62895.00	3.00	1.00	27.00	0.00	0.00
5	5.00	62895.00	2.00	2.00	42.00	1.00	1.00
6	5.00	62895.00	3.00	5.00	68.00	1.00	0.00
7	5.00	62895.00	3.00	1.00	35.00	0.00	0.00
8	5.00	62895.00	3.00	3.00	24.00	0.00	0.00
9	5.00	62895.00	2.00	2.00	90.00	0.00	0.00
10	5.00	62895.00	4.00	6.00	53.00	1.00	0.00

A description of the variables of interest in this data file is as follows:

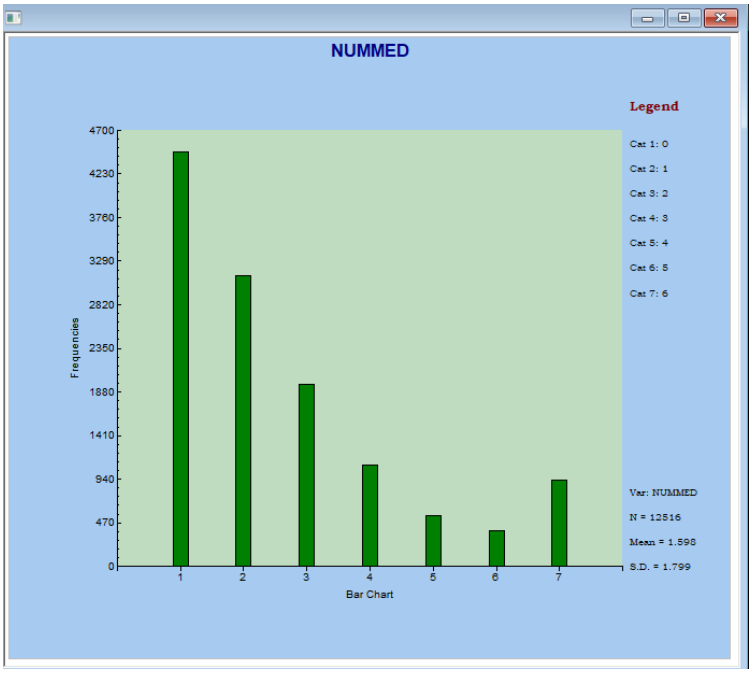
- PSU is the primary sampling unit (PSU) and is used as level-2 ID.
- PATWT is the participant design weight.
- PASTVIS is the value of the nominal variable for the number of visits to a medical doctor during the past 12 months (1 = none or unknown, 2 = 1 to 2, 3 = 3 to 5, 4 = 6 medications and more).
- NUMMED is the number of medications.

- AGE is the age of the respondent.
- GENDER, where 0 = Female and 1 = Male.
- USETOBAC indicates whether a participant smoked cigarettes or not, where 0 = no and 1 = yes.
- PRIMCARE, where 0 = none and 1 = participant has primary care.
- INJURY indicates whether a participant suffered from an injury or not (0 = no, 1 = yes).
- BLODPRES, where 0 = blood pressure not measured and 1 = blood pressure measured.
- URINE, where 0 = no urine tested, 1 = tested.
- XRAY, where 0 = no X rays taken and 1 = X ray taken.
- EXERCISE, where 0 = no exercise and 1 = participant does exercise.
- RACER indicates the ethnicity of a participant where 1 = White, 2 = Black and 3 = Other.
- AGER indicates in which age category a participant belongs. Coded as follows: 1 = Under 15, 2 = 15 to 24, 3 = 25 to 44, 4 = 45 to 64, 5 = 65 to 74, 6 = 75 and older.
- AGE1 to AGE5 are five dummy variables coded as follows:

Table: Dummy variables

Age	AGE1	AGE2	AGE3	AGE4	AGE5
Under 15	1	0	0	0	0
15 to 24	0	1	0	0	0
25 to 44	0	0	1	0	0
45 to 64	0	0	0	1	0
65 to 74	0	0	0	0	1
75 and older	0	0	0	0	0

Inspecting the distribution of the intended outcome variable, NUMMED, before starting with the model is important. The number of medications ranges from 1 to 7, with most respondents having a small number of medications.



3. The model

The first model fitted to the data explores the relationship between NUMMED and the variables indicating age, previous medical visits and results, and ethnicity.

The level-1 model is

$$\log\left(\lambda_{ij}\right) = \beta_1 \times SEX_{ij} + \beta_2 \times AGE_{ij} + \beta_3 \times PASTVIS_{ij} + \beta_4 \times PRIMCARE_{ij} + \beta_5 \times INJURY_{ij} + \beta_6 \times BLODPRES_{ij} + \beta_7 \times URINE_{ij} + \beta_8 \times CHOLEST_{ij} + \beta_9 \times RACE_{ij}$$

where the expected number of medications is $\lambda_{ij} = E\left(\text{NUMMED}_{ij}\right)$.

The level-2 model is

$$\begin{aligned}
\beta_0 &= v_{i0} \\
\beta_1 &= b_{10} \\
\beta_2 &= b_{20} + v_{i2} \\
\beta_3 &= b_{30} \\
\beta_4 &= b_{40} \\
\beta_5 &= b_{50} \\
\beta_6 &= b_{60} \\
\beta_7 &= b_{70} \\
\beta_8 &= b_{80} \\
\beta_9 &= b_{90}
\end{aligned}$$

Another way of writing the combined model is

$$\log\left(\lambda_{ij}\right) = b_{10} \times SEX_{ij} + b_{20} \times AGE_{ij} + b_{30} \times PASTVIS_{ij} + b_{40} \times PRIMCARE_{ij} + b_{50} \times INJURY_{ij} + b_{60} \times BLODPRES_{ij} + b_{70} \times URINE_{ij} + b_{80} \times CHOLEST_{ij} + b_{90} \times RACE_{ij} + v_{i0} + v_{20} \times AGE_{ij}$$

In this model, $e^{b_{00}}$ denotes the average expected number of medications, and b_{10} represents the estimated coefficient associated with the respondent's gender.

Taking exponents on both sides, we also have

$$\lambda_{ij} = e^{b_{10} \times SEX_{ij} + b_{20} \times AGE_{ij} + b_{30} \times PASTVIS_{ij} + b_{40} \times PRIMCARE_{ij} + b_{50} \times INJURY_{ij} + b_{60} \times BLODPRES_{ij} + b_{70} \times URINE_{ij} + b_{80} \times CHOLEST_{ij} + b_{90} \times RACE_{ij} + v_{i0} + v_{20} \times AGE_{ij}}$$

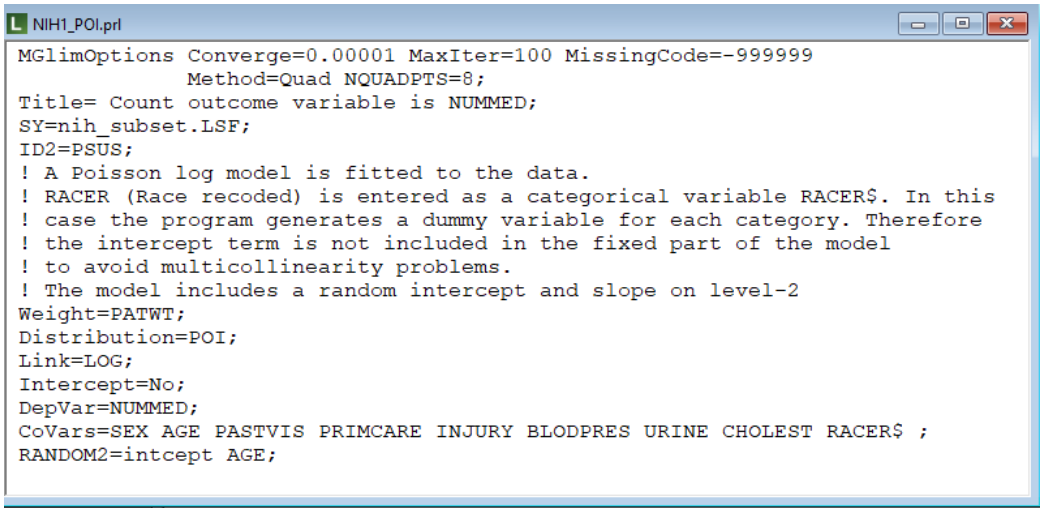
$$= e^{10 \times SEX_{ij}} e^{b_{20} \times AGE_{ij}} e^{b_{30} \times PASTVIS_{ij}} e^{b_{40} \times PRIMCARE_{ij}} e^{b_{50} \times INJURY_{ij}} e^{b_{60} \times BLODPRES_{ij}} e^{b_{70} \times URINE_{ij}} e^{b_{80} \times CHOLEST_{ij}} e^{b_{90} \times RACE_{ij}} e^{v_{i0}} e^{v_{20} \times AGE_{ij}}$$

The random part of the model is represented by $e^{v_{i0}}$ and $e^{v_{20}\times AGE_{ij}}$, which denotes the variation in average number of medications over PSU and between respondents (or, in other words, over respondents nested within PSU). The random variation over age is similarly described by $e^{v_{20}\times AGE_{ij}}$. 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. Fitting the model

Open the LISREL data spreadsheet file **nesarc_poi.lsf** and select the **Multilevel, Generalized Linear Model** option from the main menu bar. Proceed to fill in the **Title and Options** (number of quadrature points is 8); **ID and Weight** (Level-2 ID is PSUs, weight variable is PATWT); **Distributions/Links** (Poisson, log, no intercept); **Model Specification** (Dependent variable is NUMMED, predictors are the nine predictors indicated above); and the **Random Variables** dialog (Intercepts and AGE only). Note that the ethnicity variable RACE should be added as a **Categorical** predictor; all others are entered as **Continuous**.

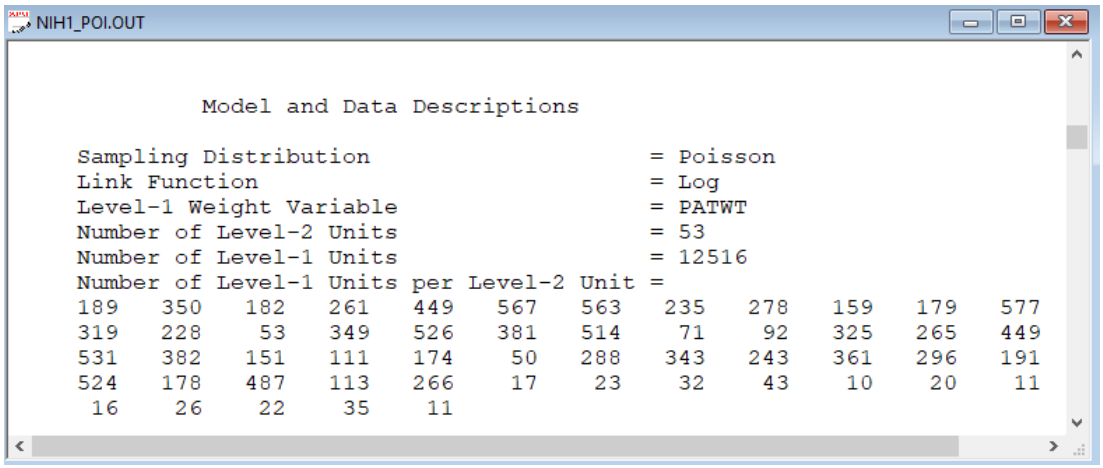
When done, click the **Finish** button to create the syntax file **nesarc_poi.prl**. Save this file as **nih1_poi.prl** using the **File, Save As** option.



Portions of the output file **nih1_poi.out** are shown below.

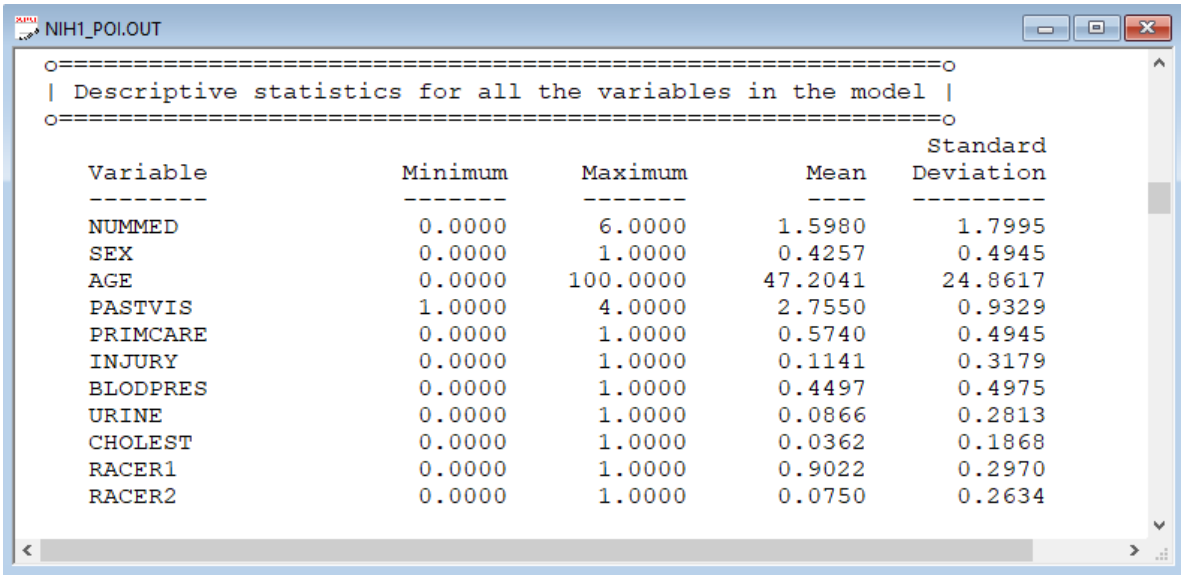
Model and data description

A description of the hierarchical structure follows the syntax: data from a total of 53 PSUs and 12516 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 PSUs is provided.



Descriptive statistics

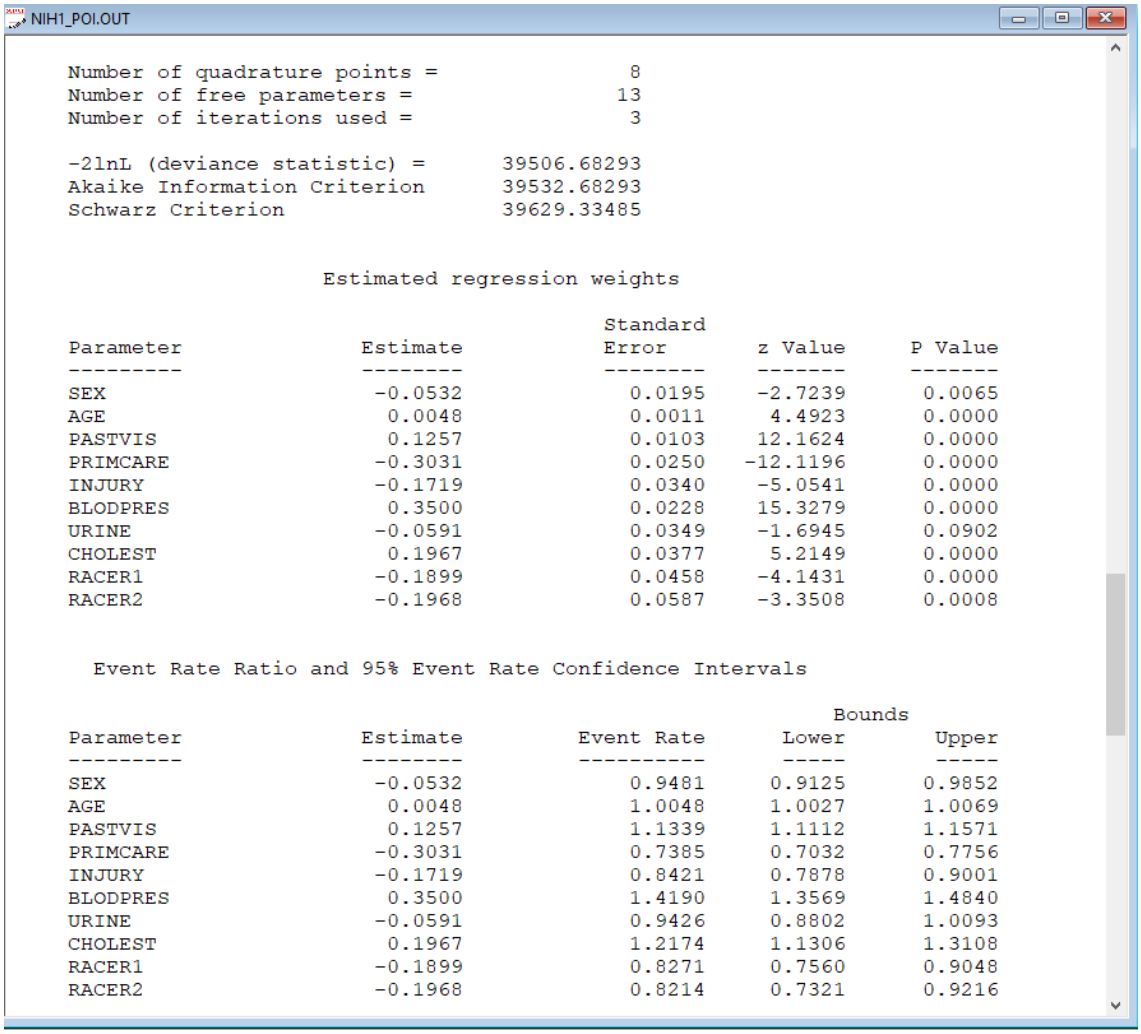
The data summary is followed by descriptive statistics for all the variables included in the model. The mean of 1.598 and standard deviation of 1.7995 are reported for the outcome NUMMED indicating that, on average, 1.6 medications are prescribed for each respondent.



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.



NIH1_POI.OUT

```
Number of quadrature points =                8
Number of free parameters =                 13
Number of iterations used =                  3

-2lnL (deviance statistic) =       39506.68293
Akaike Information Criterion       39532.68293
Schwarz Criterion                  39629.33485


                  Estimated regression weights

                                        Standard
Parameter              Estimate         Error       z Value      P Value
---------              --------         --------    -------      -------
SEX                     -0.0532           0.0195    -2.7239       0.0065
AGE                      0.0048           0.0011     4.4923       0.0000
PASTVIS                  0.1257           0.0103    12.1624       0.0000
PRIMCARE                -0.3031           0.0250   -12.1196       0.0000
INJURY                  -0.1719           0.0340    -5.0541       0.0000
BLODPRES                 0.3500           0.0228    15.3279       0.0000
URINE                   -0.0591           0.0349    -1.6945       0.0902
CHOLEST                  0.1967           0.0377     5.2149       0.0000
RACER1                  -0.1899           0.0458    -4.1431       0.0000
RACER2                  -0.1968           0.0587    -3.3508       0.0008


  Event Rate Ratio and 95% Event Rate Confidence Intervals

                                                        Bounds
Parameter              Estimate         Event Rate      Lower        Upper
---------              --------         ----------      -----        -----
SEX                     -0.0532             0.9481     0.9125       0.9852
AGE                      0.0048             1.0048     1.0027       1.0069
PASTVIS                  0.1257             1.1339     1.1112       1.1571
PRIMCARE                -0.3031             0.7385     0.7032       0.7756
INJURY                  -0.1719             0.8421     0.7878       0.9001
BLODPRES                 0.3500             1.4190     1.3569       1.4840
URINE                   -0.0591             0.9426     0.8802       1.0093
CHOLEST                  0.1967             1.2174     1.1306       1.3108
RACER1                  -0.1899             0.8271     0.7560       0.9048
RACER2                  -0.1968             0.8214     0.7321       0.9216
```

The estimated gender effect is -0.0532, which means that the average number of medications for males is $e^{-0.0532}=0.9481$. The estimated coefficient for PRIMCARE is -0.3031, which indicates that male respondents who primary care tended to have $0.9481\,e^{-0.3031}$ = (0.9481)(0.7385) = 0.7002 prescriptions, holding all other variables constant. The estimate of the effect of blood pressure shows that the high blood pressure increases the number of medications, since $e^{0.3500}=0.7047$. The estimated number of medications for a male with primary care and high bloodpressure is calculated as (0.7002)(0.7047) = 0.4934. Looking over the estimates, we note that increasing age, previous visits, and cholesterol are likely to lead to a higher number of estimated medications.

Random effects results

The output for the level-2 random effect variance term follows next. The estimated variation in the average estimated NUMMED at level 2 is 0.1069, which is highly significant. Similarly, there is evidence of significant random variation in age.

NIH1_POI.OUT

```
                Estimated level 2 variances and covariances

                                          Standard
   Parameter              Estimate        Error        z Value    P Value
   ---------              --------        --------     -------    -------
   intcept/intcept          0.1069          0.0263      4.0695     0.0000
   AGE/intcept             -0.0009          0.0004     -2.3606     0.0182
   AGE/AGE                  0.0000          0.0000      4.2505     0.0000


  Level 2 covariance matrix


              intcept        AGE

intcept      0.106917
AGE         -0.000878    0.000036
```

5. Estimating the scale parameter

The scale parameter may be obtained by using the SCALE option, as shown in the syntax below (**nih2_poi.prl**).

NIH2_POI.prl

```
MGlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999
             Method=Quad NQUADPTS=10;
Title= Poisson log model. Outcome variable is number of medications;
SY=nih_subset.LSF;
ID2=PSUS;
! RACER (Race recoded) is entered as a categorical variable RACER$. In this
! case the program generates a dummy variable for each category. Therefore
! the intercept term is not included in the fixed part of the model
! to avoid multicollinearity problems.
! The model includes a random intercept and slope on level-2
! The Scale parameter is estimated using the Pearson Chi-Square statistic
! Random intercept and slope
Weight=PATWT;
Distribution=POI;
Link=LOG;
Intercept=No;
Scale=Pearson;
DepVar=NUMMED;
CoVars=SEX PASTVIS PRIMCARE INJURY BLODPRES URINE CHOLEST RACER$ ;
RANDOM2=intcept AGE;
```

When this option is included, the following additional information is written to the output file.

```
Event Rate Ratio and 95% Event Rate Confidence Intervals

                                                          Bounds
   Parameter              Estimate        Event Rate      Lower       Upper
   ---------              --------        ----------      -----       -----
   SEX                     -0.0504            0.9509     0.9151      0.9881
   PASTVIS                  0.1339            1.1432     1.1205      1.1665
   PRIMCARE                -0.2921            0.7467     0.7110      0.7841
   INJURY                  -0.1695            0.8441     0.7896      0.9023
   BLODPRES                 0.3625            1.4370     1.3744      1.5024
   URINE                   -0.0613            0.9406     0.8783      1.0072
```

CHOLEST	0.2043	1.2267	1.1393	1.3208
RACER1	-0.1289	0.8790	0.8038	0.9613
RACER2	-0.1353	0.8735	0.7788	0.9797
SCALE	1.3513			

Note: The scale parameter estimate is based on the Pearson Chi-square value
phi = Square Root of (The Pearson Chi-square value/degrees of freedom)