Negative binomial model for the NIH data

Contents

1. Introduction

Previously, we fitted a Poisson model to the data described here. It was also noted that a Poisson distribution has an important property: the mean number of occurrences is equal to the variance. The negative binomial distribution can be used as an alternative to the Poisson distribution. It is especially useful for discrete data that assumes values 0, 1, 2, 3… whose sample variance exceeds the sample mean. In such cases, the observations are over-dispersed with respect to a Poisson distribution, for which the mean is equal to the variance. Since the negative binomial distribution has one more parameter than the Poisson, the second parameter can be used to adjust the variance independently of the mean. It can be shown that a model based on the negative binomial distribution with a dispersion parameter close to zero will produce results that correspond closely to those obtained for the Poisson model. In this section, we fit a negative binomial model, utilizing the same predictors to the NIH data. Again, adaptive quadrature is used as the method of optimization.

The negative binomial distribution can be expressed as

$$f\left(y_i\right)=\frac{\Gamma\left(y_i+1/\alpha\right)}{\Gamma\left(y_i+1\right)\Gamma\left(1/\alpha\right)}\times\frac{\left(\alpha\mu_i\right)^{y_i}}{\left(1+\alpha\mu_i\right)^{y_i+1/\alpha}}$$

with $\sigma^2\left(y_i\right)=\mu_i+\alpha\mu_i^2$ where α denotes an additional parameter and it can no longer be assumed that the variance is a known function of the mean. We assume α to be a fixed parameter.

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

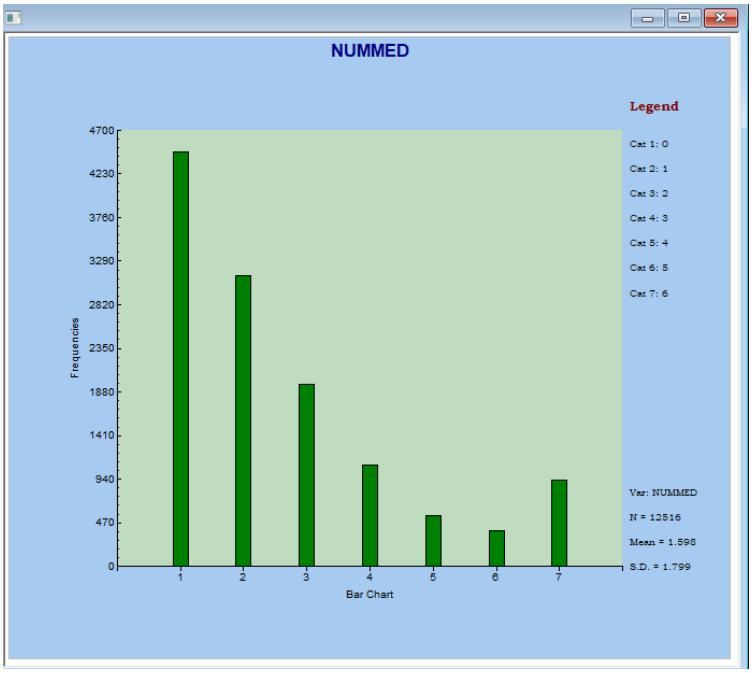
A description of the variables of interest in this data file is as follows:

- 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 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

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

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.

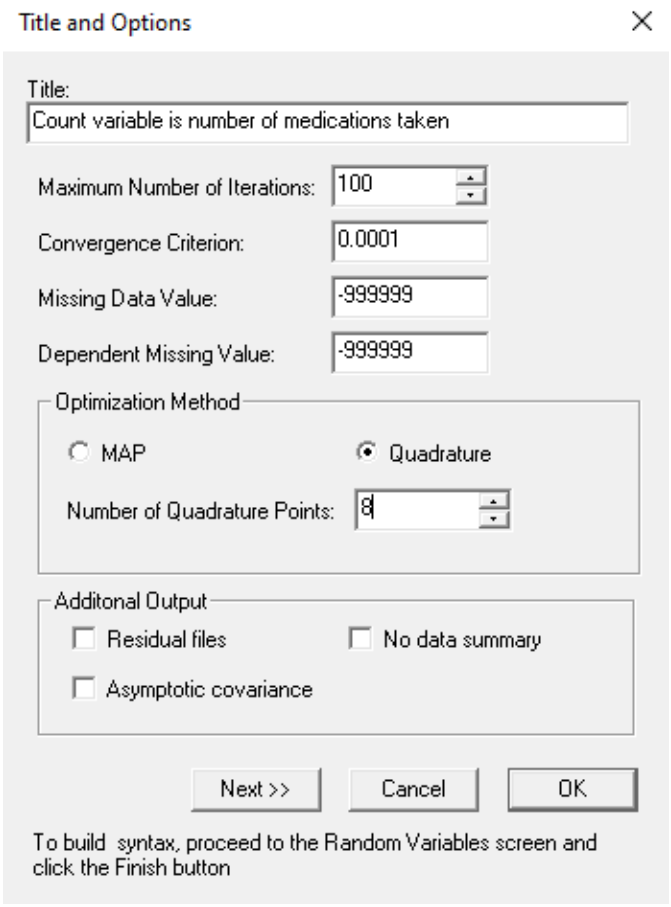
Taking exponents on both sides, we also have

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

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 $e^{v_{20} \times AGE_{ij}}$.

4. Fitting the model

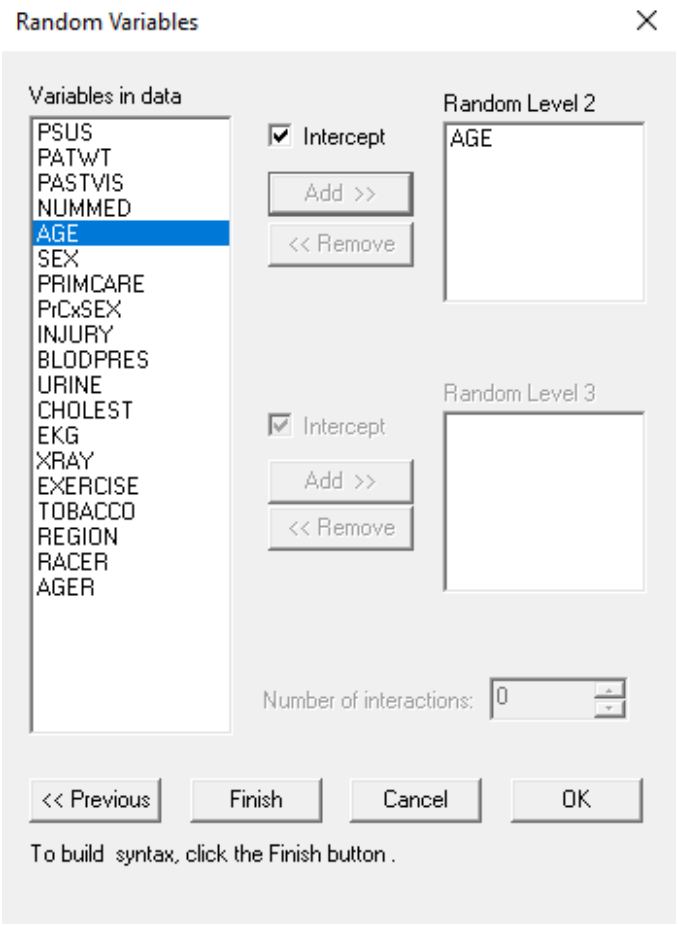
Open the LISREL data spreadsheet file **nih_subset.lsf** and select the **Multilevel, Generalized Linear Model** option from the main menu bar. Proceed to fill in the **Title and Options** (number of quadrature points is 8) dialog box and click **Next** to go to the **ID and Weight Variables** dialog box.



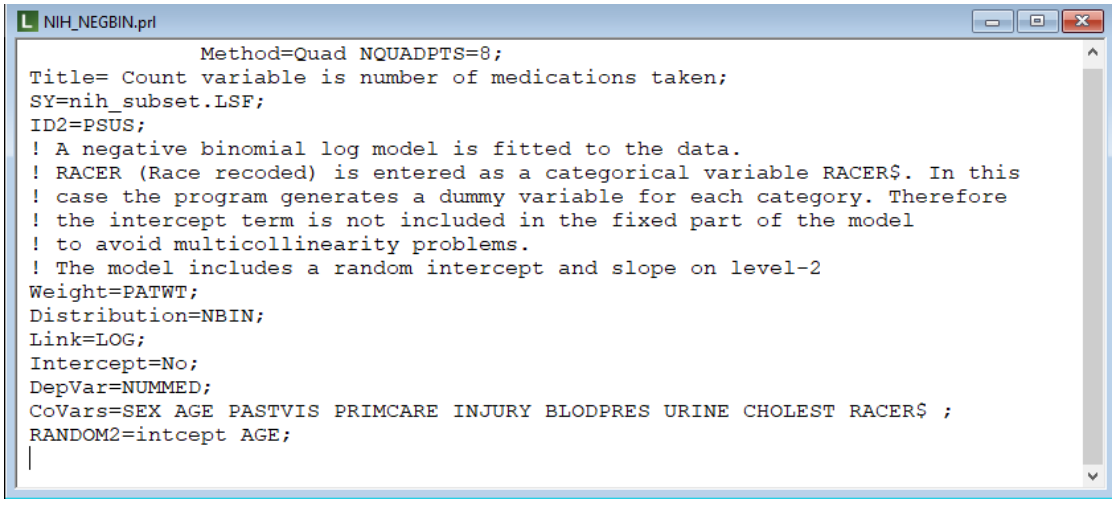
On this dialog box, indicate the level-2 ID as PSUs, and the weight variable as PATWT. The next dialog box is the **Distributions and Links** dialog box. Select the **Distribution type** and **link function** as Negative binomial and log. By default, an intercept will be included in the model. As we plan to use the variable RACER (Race recoded) in the model which will be used as categorical variable, we opt to not have an intercept included. When a categorical variable is entered, the the program generates a dummy variable for each category. Therefore, the intercept term is not included in the fixed part of the model in order to avoid multicollinearity problems.

On the **Dependent and Independent Variables** dialog box, enter NUMMED as the Dependent variable and the predictors SEX, AGE, PASTVIS, PRIMCARE, INJURY, BLODPRES, URINE, and CHOLEST as **Continuous** predictors into the **Independent variables** box. Finally, enter the variable RACER as a Categorical variable.

On the final dialog box, select AGE and add it to the **Random Level 2** field. Note that by default, it is assumed that the intercept is allowed to vary randomly over the level-2 units.

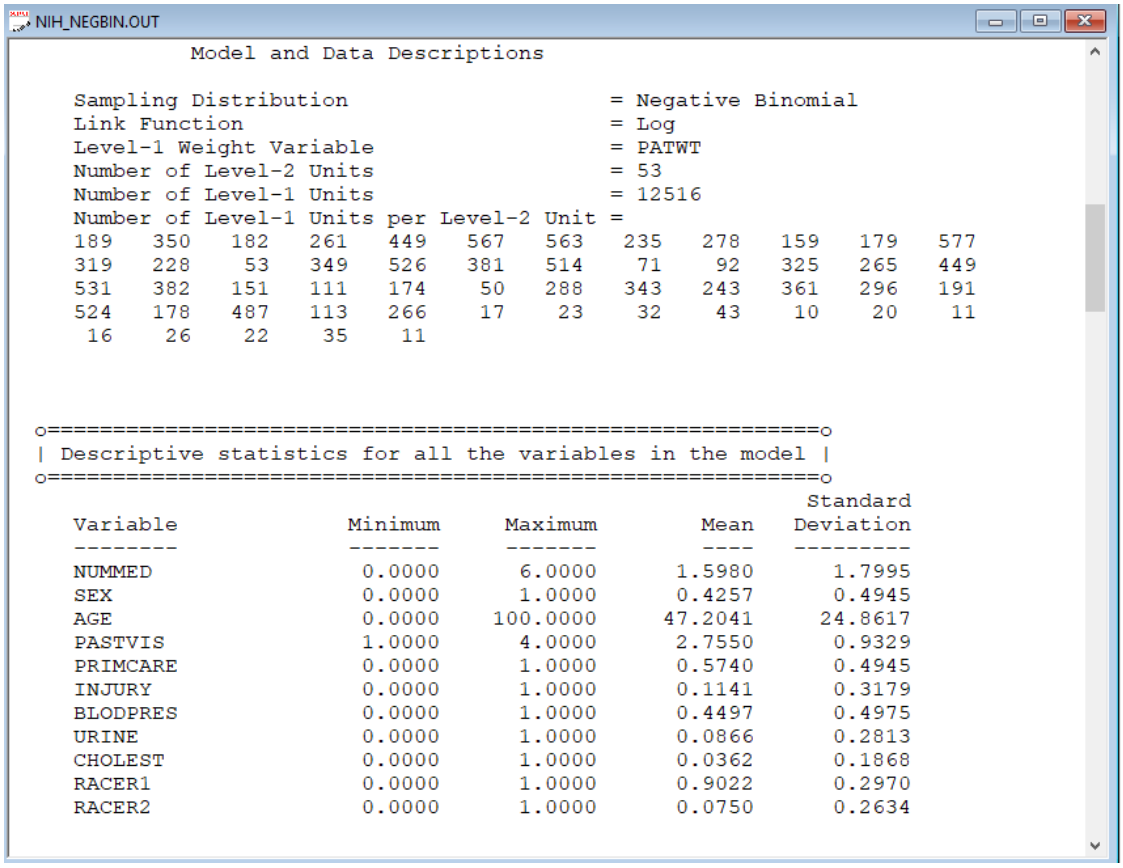


When done, click the **Finish** button to create the syntax file. Save this file as **nih_negbin.prl** using the **File, Save As** option.



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

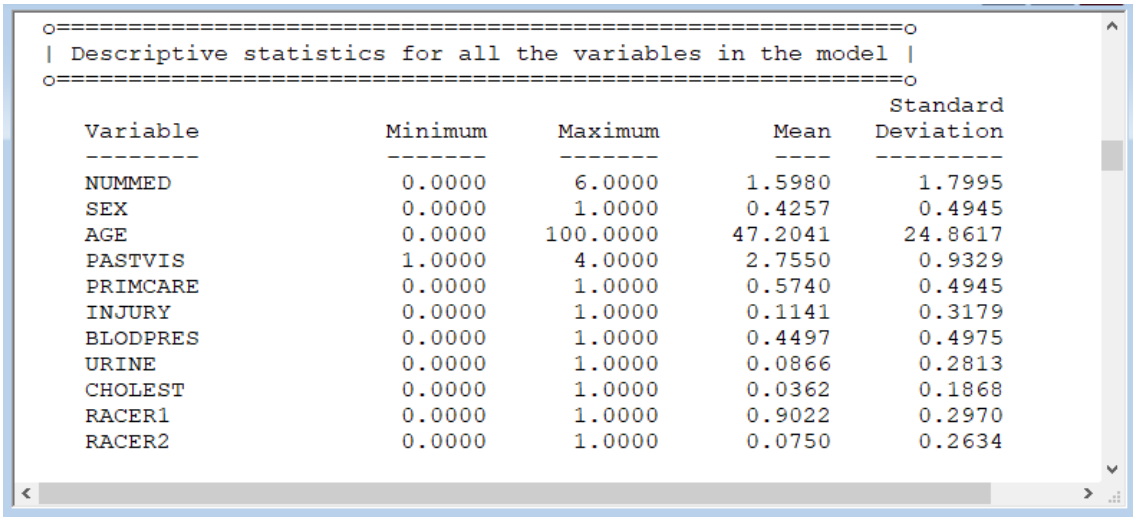
A description of the hierarchical structure follows the syntax: data from a total of 53 PSUs and 12516 respondents were included at levels 2 and 1 of the model. In addition, an enumeration of the number of respondents nested within each of the PSUs is provided.



NIH_NEGBIN.OUT

```
            Model and Data Descriptions

   Sampling Distribution                    = Negative Binomial
   Link Function                            = Log
   Level-1 Weight Variable                  = PATWT
   Number of Level-2 Units                  = 53
   Number of Level-1 Units                  = 12516
   Number of Level-1 Units per Level-2 Unit =
   189   350   182   261   449   567   563   235   278   159   179   577
   319   228    53   349   526   381   514    71    92   325   265   449
   531   382   151   111   174    50   288   343   243   361   296   191
   524   178   487   113   266    17    23    32    43    10    20    11
    16    26    22    35    11


o============================================================o
| Descriptive statistics for all the variables in the model |
o============================================================o
                                                        Standard
   Variable             Minimum     Maximum        Mean   Deviation
   --------             -------     -------        ----   ---------
   NUMMED                0.0000      6.0000      1.5980      1.7995
   SEX                   0.0000      1.0000      0.4257      0.4945
   AGE                   0.0000    100.0000     47.2041     24.8617
   PASTVIS               1.0000      4.0000      2.7550      0.9329
   PRIMCARE              0.0000      1.0000      0.5740      0.4945
   INJURY                0.0000      1.0000      0.1141      0.3179
   BLODPRES              0.0000      1.0000      0.4497      0.4975
   URINE                 0.0000      1.0000      0.0866      0.2813
   CHOLEST               0.0000      1.0000      0.0362      0.1868
   RACER1                0.0000      1.0000      0.9022      0.2970
   RACER2                0.0000      1.0000      0.0750      0.2634
```

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.



```
o============================================================o
| Descriptive statistics for all the variables in the model |
o============================================================o
                                                        Standard
   Variable             Minimum     Maximum        Mean   Deviation
   --------             -------     -------        ----   ---------
   NUMMED                0.0000      6.0000      1.5980      1.7995
   SEX                   0.0000      1.0000      0.4257      0.4945
   AGE                   0.0000    100.0000     47.2041     24.8617
   PASTVIS               1.0000      4.0000      2.7550      0.9329
   PRIMCARE              0.0000      1.0000      0.5740      0.4945
   INJURY                0.0000      1.0000      0.1141      0.3179
   BLODPRES              0.0000      1.0000      0.4497      0.4975
   URINE                 0.0000      1.0000      0.0866      0.2813
   CHOLEST               0.0000      1.0000      0.0362      0.1868
   RACER1                0.0000      1.0000      0.9022      0.2970
   RACER2                0.0000      1.0000      0.0750      0.2634
```

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.

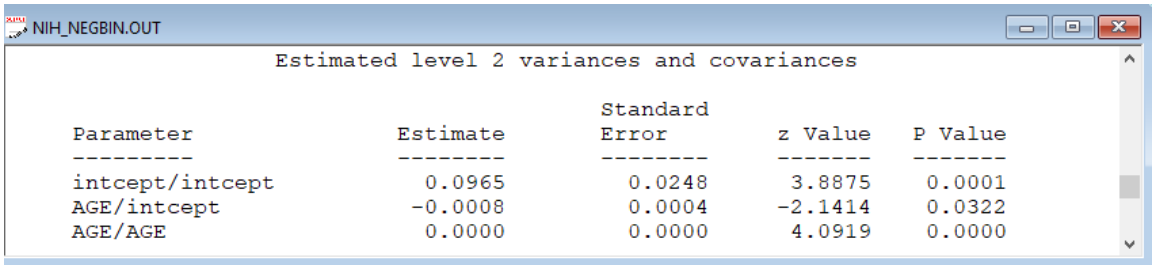
NIH_NEGBIN.OUT

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

-2lnL (deviance statistic) =       39122.89192
Akaike Information Criterion       39148.89192
Schwarz Criterion                  39245.54384


                  Estimated regression weights

                                        Standard
Parameter              Estimate         Error       z Value      P Value
---------              --------         --------    -------      -------
SEX                     -0.0417           0.0201    -2.0804       0.0375
AGE                      0.0049           0.0010     4.8421       0.0000
PASTVIS                  0.1343           0.0105    12.7667       0.0000
PRIMCARE                -0.3556           0.0254   -13.9831       0.0000
INJURY                  -0.1460           0.0336    -4.3450       0.0000
BLODPRES                 0.3533           0.0233    15.1673       0.0000
URINE                   -0.0613           0.0362    -1.6957       0.0899
CHOLEST                  0.1880           0.0414     4.5386       0.0000
RACER1                  -0.2098           0.0462    -4.5375       0.0000
RACER2                  -0.1851           0.0592    -3.1283       0.0018


  Event Rate Ratio and 95% Event Rate Confidence Intervals

                                                        Bounds
Parameter              Estimate         Event Rate      Lower        Upper
---------              --------         ----------      -----        -----
SEX                     -0.0417             0.9591     0.9222       0.9976
AGE                      0.0049             1.0050     1.0029       1.0070
PASTVIS                  0.1343             1.1437     1.1204       1.1676
PRIMCARE                -0.3556             0.7007     0.6667       0.7365
INJURY                  -0.1460             0.8642     0.8091       0.9230
BLODPRES                 0.3533             1.4238     1.3602       1.4903
URINE                   -0.0613             0.9405     0.8761       1.0096
CHOLEST                  0.1880             1.2068     1.1127       1.3089
RACER1                  -0.2098             0.8108     0.7405       0.8877
RACER2                  -0.1851             0.8310     0.7400       0.9332
```

The estimated gender effect is -0.0417, which means that the average number of medications for males is $e^{-0.0417}=0.9592$, compared to 0.9836 under the Poisson model fitted to the same data. The estimated coefficient for PRIMCARE is now -0.3556 (-0.3031 for the Poisson model), which indicates that male respondents who primary care tended to have $0.9592\ e^{-0.3556} = (0.9592)(0.7008) = 0.6722$ prescriptions, holding all other variables constant. This is a lower estimate than obtained for the Poisson model, where we had 0.7264. The estimate of the effect of blood pressure stayed essentially the same (0.3533 versus 0.3500) and still shows that high blood pressure increases the expected number of medications. 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 0.0965 compared to the 0.1069 obtained for the Poisson model, and is highly significant. Similarly, there is evidence of significant random variation in age.


NIH_NEGBIN.OUT

Estimated level 2 variances and covariances

Parameter	Estimate	Standard Error	z Value	P Value
intcept/intcept	0.0965	0.0248	3.8875	0.0001
AGE/intcept	-0.0008	0.0004	-2.1414	0.0322
AGE/AGE	0.0000	0.0000	4.0919	0.0000