Scientific Software International

Poisson model with offset variable for the Thailand 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.

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

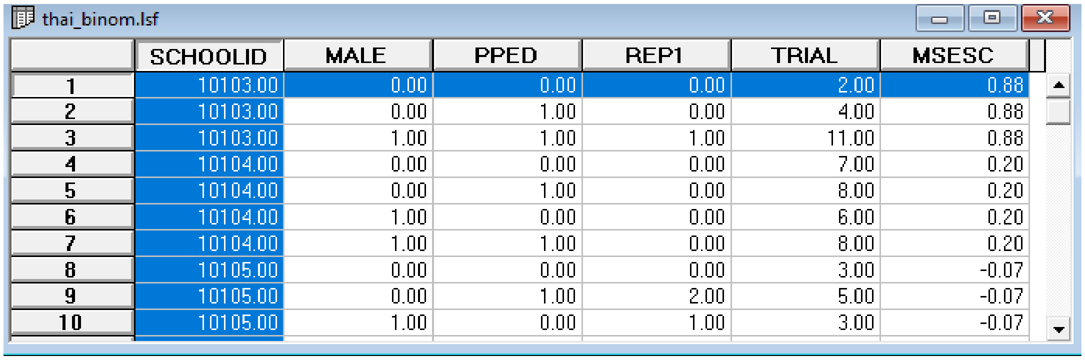
$$y_i \sim Poi(\lambda_i), \qquad where \qquad \lambda_i = e^{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}}$$

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

Data are from a national survey of primary education in Thailand (see Raudenbush & Bhumirat, 1992, for details), conducted in 1988, and yielding, for our analysis, complete data on 7516 sixth graders nested within 356 primary schools. Of interest is the probability that a child will repeat a grade during the primary years (REP1 = 1 if yes, 0 if no). It is hypothesized that the sex of the child (MALE = 1 if male, 0 of female), the child's pre-primary experience (PPED = 1 if yes, 0 if no), and the school mean SES (MSESC) will be associated with the probability of repetition. A partial list of the data is given below in the form of a LISREL spreadsheet file, named **thai_binom.lsf**.



thai_binom.lsf

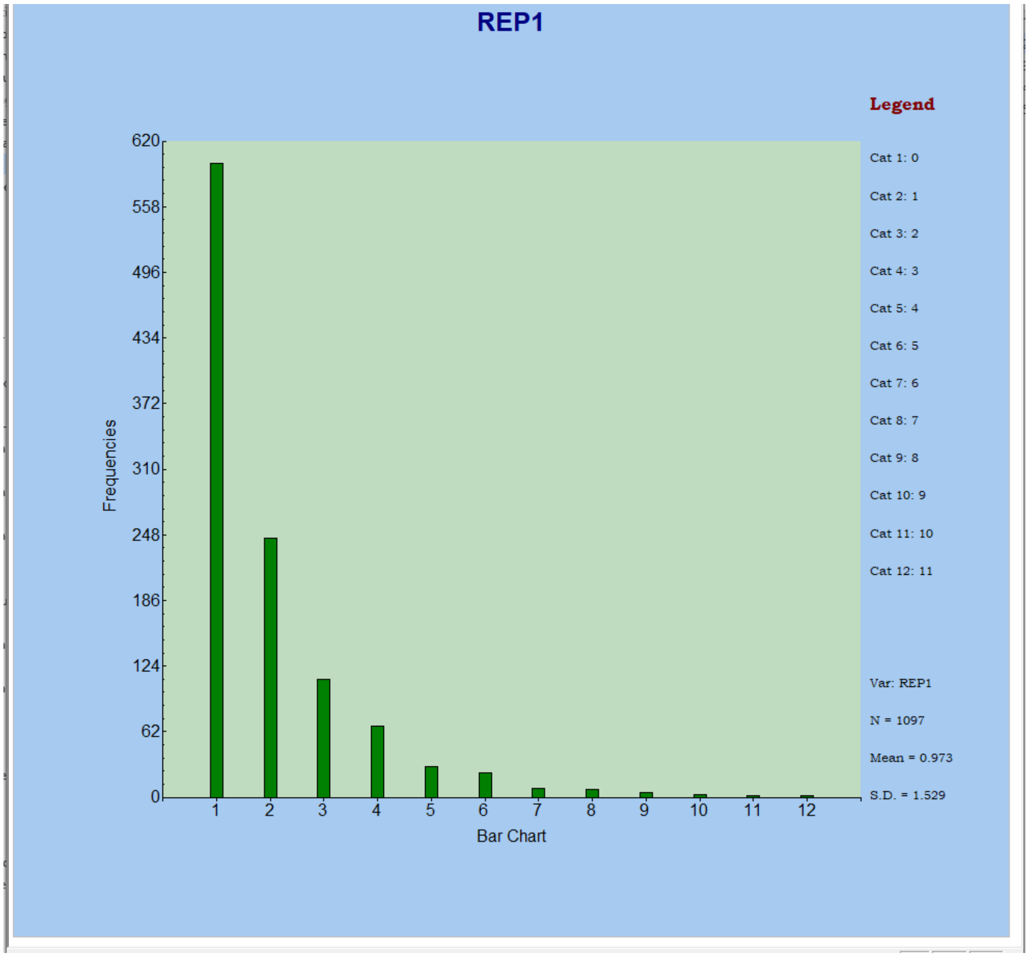
	SCHOOLID	MALE	PPED	REP1	TRIAL	MSESC
1	10103.00	0.00	0.00	0.00	2.00	0.88
2	10103.00	0.00	1.00	0.00	4.00	0.88
3	10103.00	1.00	1.00	1.00	11.00	0.88
4	10104.00	0.00	0.00	0.00	7.00	0.20
5	10104.00	0.00	1.00	0.00	8.00	0.20
6	10104.00	1.00	0.00	0.00	6.00	0.20
7	10104.00	1.00	1.00	0.00	8.00	0.20
8	10105.00	0.00	0.00	0.00	3.00	-0.07
9	10105.00	0.00	1.00	2.00	5.00	-0.07
10	10105.00	1.00	0.00	1.00	3.00	-0.07

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

- SCHOOLID is used as level-2 ID.
- MALE indicates gender, with females coded 0 and males 1.
- PPED is an indicator of pre-primary school education (1 = yes, 0 = no).
- REP1 is the number of grade retentions for each of the four subpopulations formed by MALE and PPED within each school.

- ○ TRIAL is the number of students with a specific school.
- ○ MSESC is a measure of the school's mean socio-economic status.

Inspecting the distribution of the intended outcome variable, REP1 before starting with the model is important. The number of repetitions ranges from 1 to 12, with most cells having only 1 repetition reported.



3. The model

The first model fitted to the data explores the relationship between REP1 and the variables indicating gender, pre-primary education and mean SES. In addition, we will use the TRIAL variable as the offset variable.

The level-1 model is

$$\log\left(\lambda_{ij}\right) = \beta_0 + \beta_1 \times MALE_{ij} + \beta_2 \times PPED_{ij} + \beta_3 \times MSESC_{ij}$$

where the expected number of medications is $\lambda_{ij} = E\left(\mathrm{REP1}_{ij}\right)$.

The offset variable is introduced into the Poisson model in the following way:

$$\log\left(\hat{\lambda}_{ij}\right) = \log(\text{offset variable}) + \left[\mathbf{x}_{ij}^{'}\mathbf{b}_i\right]$$

where $\mathbf{x}_{ij}$ represent the values of the covariates corresponding to level-1 unit j nested within level-2 unit i and $\mathbf{b}_i$ denotes the coefficient vector containing both fixed and random effects.

In the current situation, the variable TRIAL is the appropriate choice as the OFFSET variable. The model to be fitted to the data now changes to:

$$\log\left(\lambda_{ij}\right) = \log\left(TRIAL\right) + \beta_0 + \beta_1 \times MALE_{ij} + \beta_2 \times PPED_{ij} + \beta_3 \times MSESC_{ij}$$

The level-2 model is

$$\begin{aligned}\beta_0 &= b_{00} + v_{i0}\\ \beta_1 &= b_{10}\\ \beta_2 &= b_{20}\\ \beta_3 &= b_{30}\end{aligned}\quad.$$

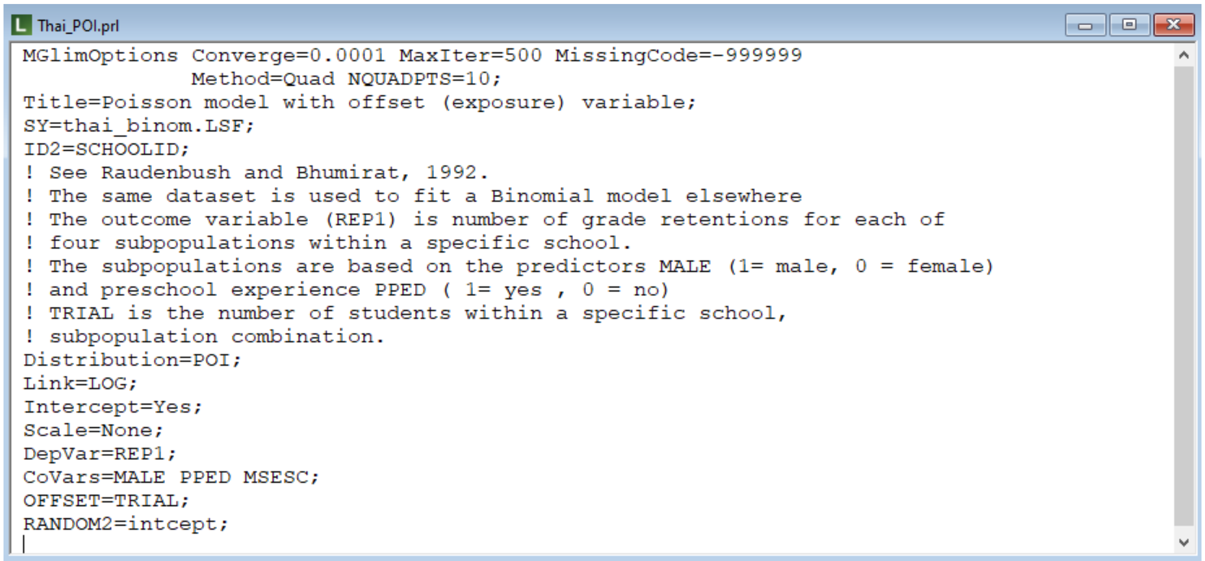
In this model, $e^{b_{00}}$ denotes the average expected number of repetitions, and b_{10} represents the estimated coefficient associated with the respondent's gender. The estimated coefficients for PPED and MSESC are b_{20} and b_{30} respectively.

The random part of the model is represented by $e^{v_{i0}}$ which denotes the variation in average number of repetitions over schools and between students (or, in other words, over students nested within the school). 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 **thai_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 10); **ID and Weight** (Level-2 ID is SCHOOLID); **Distributions/Links** (Poisson, log); **Dependent and Independent Variables** (dependent variable is REP1, predictors are the 3 predictors indicated above). All predictors are entered as **Continuous**. Select TRIAL as the **Offset variable**.

When done, click the **Finish** button to create the syntax file **thai_poi.prl.**

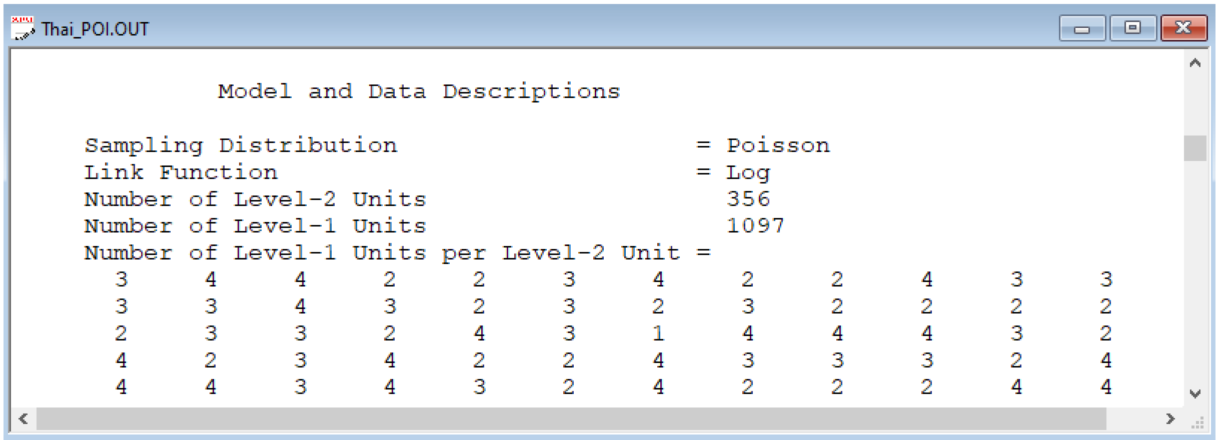


```
MGlimOptions Converge=0.0001 MaxIter=500 MissingCode=-999999
             Method=Quad NQUADPTS=10;
Title=Poisson model with offset (exposure) variable;
SY=thai_binom.LSF;
ID2=SCHOOLID;
! See Raudenbush and Bhumirat, 1992.
! The same dataset is used to fit a Binomial model elsewhere
! The outcome variable (REP1) is number of grade retentions for each of
! four subpopulations within a specific school.
! The subpopulations are based on the predictors MALE (1= male, 0 = female)
! and preschool experience PPED ( 1= yes , 0 = no)
! TRIAL is the number of students within a specific school,
! subpopulation combination.
Distribution=POI;
Link=LOG;
Intercept=Yes;
Scale=None;
DepVar=REP1;
CoVars=MALE PPED MSESC;
OFFSET=TRIAL;
RANDOM2=intcept;
```

Portions of the output file are shown below.

Model and data description

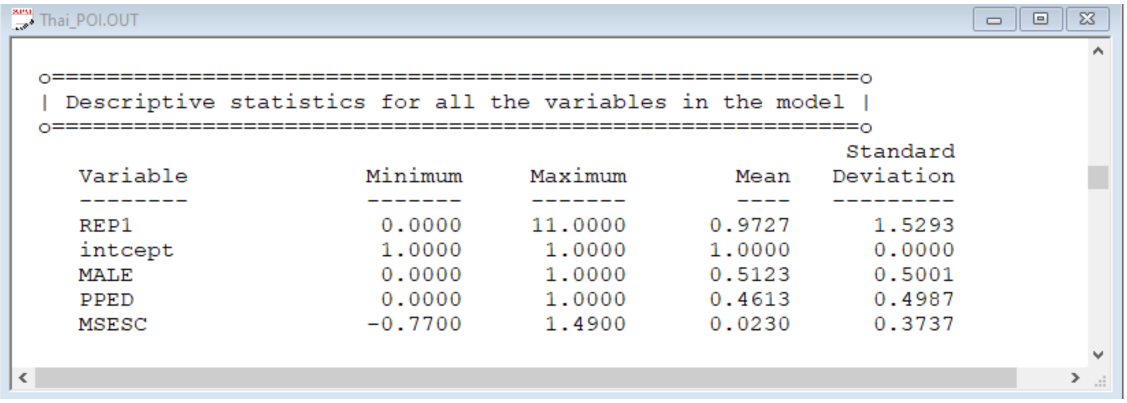
A description of the hierarchical structure follows the syntax: data from a total of 356 schools and 1097 respondents were included at levels 2 and 1 of the model. In addition, an enumeration of the number of respondents nested within each of the schools is provided.



```
              Model and Data Descriptions

     Sampling Distribution                    = Poisson
     Link Function                            = Log
     Number of Level-2 Units                    356
     Number of Level-1 Units                    1097
     Number of Level-1 Units per Level-2 Unit =
       3     4     4     2     2     3     4     2     2     4     3     3
       3     3     4     3     2     3     2     3     2     2     2     2
       2     3     3     2     4     3     1     4     4     4     3     2
       4     2     3     4     2     2     4     3     3     3     2     4
       4     4     3     4     3     2     4     2     2     2     4     4
```

Descriptive statistics

The data summary is followed by descriptive statistics for all the variables included in the model. The mean of 0.9727 and standard deviation of 1.5293 are reported for the outcome REP1 indicating that, on average, 0.9727 repetitions occur.

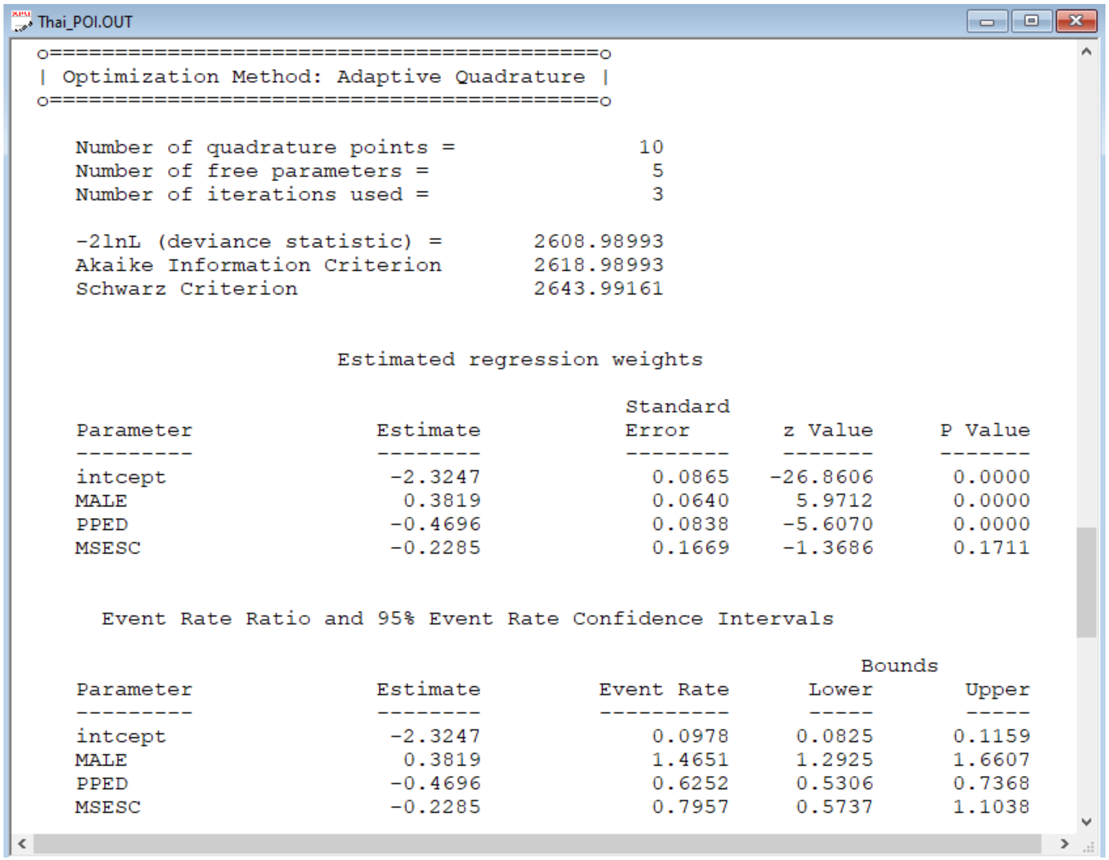


```
o===========================================================o
| Descriptive statistics for all the variables in the model |
o===========================================================o
                                                      Standard
  Variable                Minimum     Maximum      Mean   Deviation
  --------                -------     -------      ----   ---------
  REP1                     0.0000     11.0000    0.9727      1.5293
  intcept                  1.0000      1.0000    1.0000      0.0000
  MALE                     0.0000      1.0000    0.5123      0.5001
  PPED                     0.0000      1.0000    0.4613      0.4987
  MSESC                   -0.7700      1.4900    0.0230      0.3737
```

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.

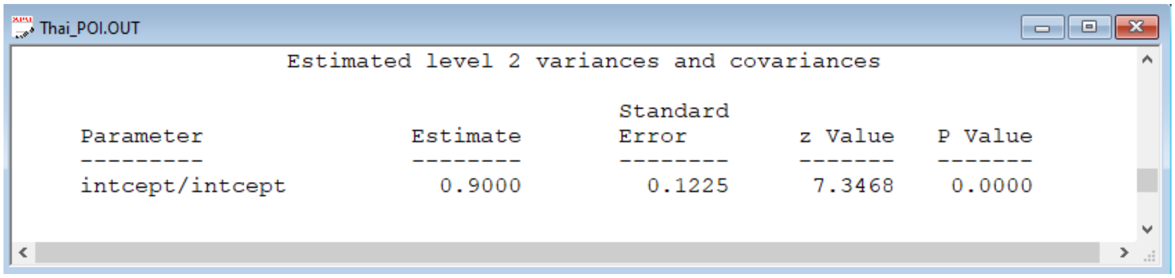


```
o==========================================o
| Optimization Method: Adaptive Quadrature |
o==========================================o

  Number of quadrature points =              10
  Number of free parameters =                 5
  Number of iterations used =                 3

  -2lnL (deviance statistic) =       2608.98993
  Akaike Information Criterion       2618.98993
  Schwarz Criterion                  2643.99161


                      Estimated regression weights

                                             Standard
  Parameter               Estimate           Error         z Value      P Value
  ---------               --------           --------      -------      -------
  intcept                  -2.3247             0.0865     -26.8606       0.0000
  MALE                      0.3819             0.0640       5.9712       0.0000
  PPED                     -0.4696             0.0838      -5.6070       0.0000
  MSESC                    -0.2285             0.1669      -1.3686       0.1711


    Event Rate Ratio and 95% Event Rate Confidence Intervals

                                                                 Bounds
  Parameter               Estimate         Event Rate       Lower        Upper
  ---------               --------         ----------       -----        -----
  intcept                  -2.3247             0.0978      0.0825       0.1159
  MALE                      0.3819             1.4651      1.2925       1.6607
  PPED                     -0.4696             0.6252      0.5306       0.7368
  MSESC                    -0.2285             0.7957      0.5737       1.1038
```

The estimated gender effect is 0.3819, which means that the average number of repetitions for males is $e^{0.03819}=1.4651$. The estimated coefficient for PPED is -0.4696, which indicates that male respondents who primary care tended to have $0.9625\,e^{-0.4696} = (1.4651)(0.6253) = 0.9161$ repetitions, holding MSESC constant. The estimate of the effect of MSESC shows that the higher socio-economic status reduces the risk of repetition, since $e^{-0.2285} = 0.7957$. For females with pre-primary education the estimated risk of repetition is $e^{-0.4696}=0.6253$. It seems reasonable to conclude that in general the risk of repetition is higher for male than for female students with pre-primary education.

Random effects results

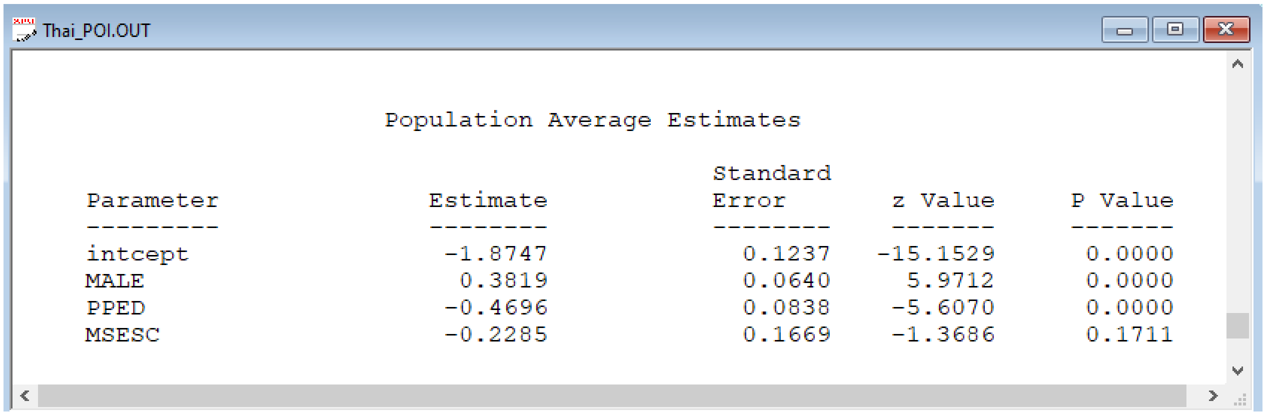
The output for the level-2 random effect variance term follows next. The estimated variation in the average estimated UREP1 at level 2 is 0.9, which is highly significant.



```
                    Estimated level 2 variances and covariances


     Parameter                  Estimate       Standard
                                               Error         z Value    P Value
     ---------                  --------       --------      -------    -------
     intcept/intcept              0.9000         0.1225       7.3468     0.0000
```

Random effects results

Finally, population-average results are reported.



```
                         Population Average Estimates


     Parameter                 Estimate           Standard
                                                  Error         z Value       P Value
     ---------                 --------           --------      -------       -------
     intcept                    -1.8747             0.1237     -15.1529        0.0000
     MALE                        0.3819             0.0640       5.9712        0.0000
     PPED                       -0.4696             0.0838      -5.6070        0.0000
     MSESC                      -0.2285             0.1669      -1.3686        0.1711
```

The regression parameters in multilevel generalized linear models have the “unit-specific” or conditional interpretation, in contrast to the “population-average” or marginal estimates that represent the unconditional covariate effects. LISREL uses numerical quadrature to obtain population-average estimates from their unit-specificcounterparts in models with multiple random effects. Standard errors for the population-average estimates are derived using the delta method. Under the model we fitted, the predicted probability for case *ij*, given u_{0j} , would be

$$E(Y_{ij} \mid u_{0j}) = \frac{1}{1+\exp\left\{-\left(\beta_0 + \beta_1 MALE_{ij} + \beta_2 PPED_{ij} + \beta_3 MSESC_{ij} + u_{0j}\right)\right\}}$$

while the population average model would be

$$E(Y_{ij}) = \frac{1}{1+\exp\left\{-\left(\beta_0 + \beta\beta_1 MALE_{ij} + \beta_2 PPED_{ij} + \beta_3 MSESC_{ij}\right)\right\}}$$

Users will need to take care in choosing unit-specific versus population-average results for their research. The choice will depend on the specific research questions that are of interest. If one were primarily interested in how a change in one of the predictors, for example PPED, can be expected to affect a particular individual school's mean, one would use the unit-specific model. If one were interested in how a change in PPED can be expected to affect the overall population mean, one would use the population-average model.