Multilevel generalized linear modeling

Contents

Multilevel generalized linear modeling 1
1. Introduction 1
2. GLIMs for counts 2
The data 2
Fitting a Binomial-logit model 3
Fitting a Poisson-log model 4
3 GLIMs for ordinal response variables 6
The data 6
Fitting a Multinomial-logit model 7
Fitting a Multinomial-cumulative logit model 8

1. Introduction

Many popular statistical methods are based on mathematical models that assume data follow a normal distribution. The most obvious among these are the analysis of variance for planned experiments and multiple regression for general analyses of independent and dependent variables. In many situations, the normality assumption is not plausible. Consequently, use of methods that assume normality may perform unsatisfactorily. In these cases, other alternatives that do not require data to have a normal distribution are attractive.

The collection of models called Generalized Linear Models (GLIMs) have become important, and practical, statistical tools. The basic idea of GLIMs is an adaption of standard regression to quite different kinds of data. The variables may be dichotomous (agree/disagree), categorical (as with a 5-point Likert scale), counts (number of arrest records), or nominal (choose among six candidates for mayor). The motivation is to tailor the regression relationship connecting the outcome to relevant independent variables so that it is appropriate to the properties of the dependent variable. The payoff is an analysis that often is more justifiable for the particular problem than a standard regression model would be.

The statistical theory and methods for fitting Generalized LInear Models (GLIMs) to simple random sample data are described in, amongst others, McCullach & Nelder (1989) and Agresti (2002). However, researchers from the social and economic sciences are often applying these methods to multilevel data. Consequently, inappropriate results are obtained if these methods are applied to multilevel data. Statistical applications such as HLM (Raudenbush & Bryk 2007) and SAS PROC NLIN (SAS Institute 2004) implement appropriate methods to fit generalized linear models to multilevel data.

LISREL (Jöreskog & Sörbom 2006) includes the statistical application MAPGLIM, which appropriately fits generalized linear models to multilevel data using the Maximum A Priori (MAP) method. Unlike other statistical software for generalized linear modeling for multilevel data such as HLM and SAS PROC NLIN, MAPGLIM allows for a wide variety of sampling distributions and link functions.

In this note, we use MAPGLIM to fit generalized linear models for counts and ordinal response variables to multilevel data.

2. GLIMs for counts

Variables measured in scientific studies come in a wide assortment. When statisticians refer to a "count" variable, they mean a variable that is ordinal, typically scored 0, 1, 2, …, without fractional values such as 2.4 or 6.75. They also mean that the variable is a tally that records how often some behavior occurred, or of how many incidents of a particular kind were observed in each subject of a study.

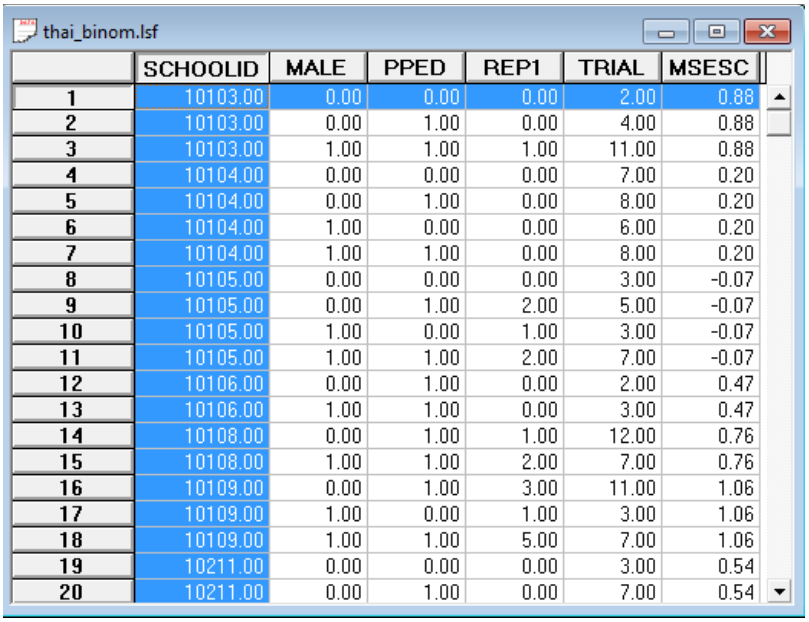
In many situations, count variables are skewed. The percentage of subjects with a score of zero or 1 is very large, those with a score of 4 or 5 or 6 considerably less common, and those with a score of 11 or 12 rare. For example, the number of delinquent acts committed by a teenager is a count variable. It is zero for the great majority. A young person who commits 1 or 2 or 3 delinquent acts is relatively rare compared to those who have no offenses. The frequencies of 1 or 2 or 3 decrease rapidly compared to those with no offenses. Juveniles who commit as many as 9 or 10 delinquent acts are very rare. As another example, the number of visits that a person makes to his or her primary care physician in a year is a count. The great majority visit the doctor not at all or once or twice in a year. Some may seek help 5, 6, or 7 times. A very few chronically ill may visit on as many as 15 occasions.

Count variables are often analyzed in exactly the same way that a continuous variable is handled, most often with a method that incorrectly assumes the count is a bell-shaped normal distribution. But counts are ordinal variables, usually skewed with a small range. They have none of the characteristics of a continuous variable. While in many instances there are few practical problems treating them as if they were continuous variables, it is easy to find examples where an inappropriate analysis of a count variable loses important information that a better approach would convey. GLIMs for counts are a special kind of model that is designed to represent the unique features of count variables in a statistically optimal way.

GLIMs for counts usually assume a Poisson, Negative Binomial or Binomial distribution for the response variable. In this section, we use MAPGLIM to fit a Poison-log and a Binomial-logit model to educational data. A description of the data follows.

The data

The data set forms part of a national survey of primary education in Thailand in 1988. The data subjects are 1097 children repeating a grade during their time at primary school and the specific data set is provided in the location **Generalized Linear modeling examples** as the PSF **thai_binom.lsf**. The first portion of this file is shown in the following LSF window.



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
11	10105.00	1.00	1.00	2.00	7.00	-0.07
12	10106.00	0.00	1.00	0.00	2.00	0.47
13	10106.00	1.00	1.00	0.00	3.00	0.47
14	10108.00	0.00	1.00	1.00	12.00	0.76
15	10108.00	1.00	1.00	2.00	7.00	0.76
16	10109.00	0.00	1.00	3.00	11.00	1.06
17	10109.00	1.00	0.00	1.00	3.00	1.06
18	10109.00	1.00	1.00	5.00	7.00	1.06
19	10211.00	0.00	0.00	0.00	3.00	0.54
20	10211.00	0.00	1.00	0.00	7.00	0.54

SCHOOLID is the school identification variable. REP1 denotes number of grade retentions for each of four subpopulations within a specific school. These 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. MSESC denotes the mother's socio-economic status score. More details about the data are provided in Raudenbush & Bhumirat (1992).

Fitting a Binomial-logit model

Select the **Open** option on the **File** menu of the main window to load the **Open** dialog box. Select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box. Browse for and open the file **thai_binom.lsf**.

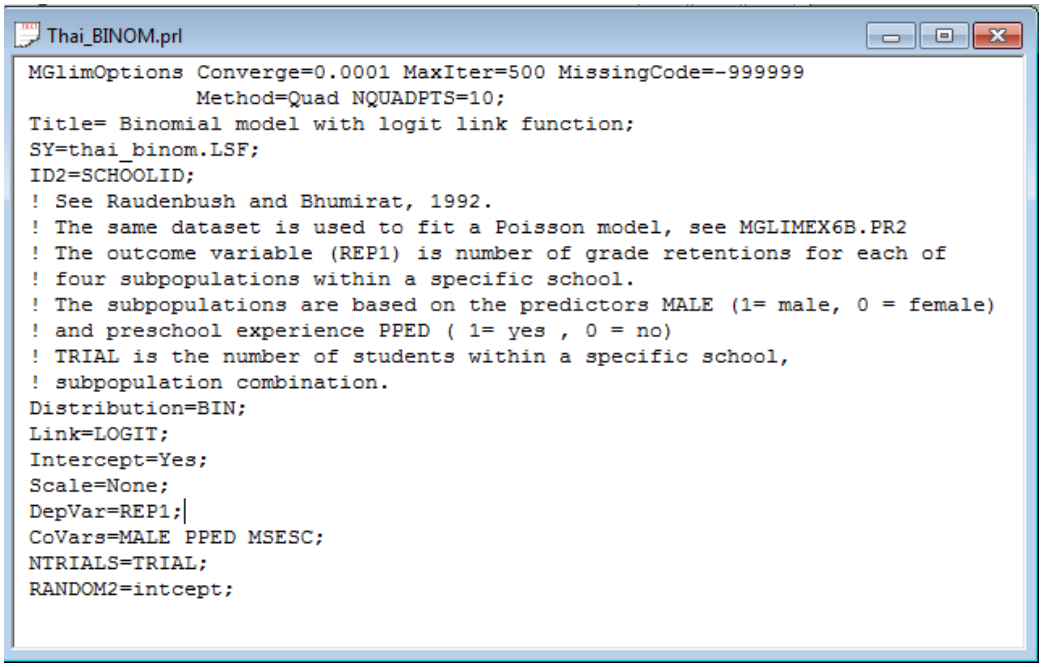
Select the **Title and Options** option on the **Generalized Linear Model** pop-up of the **Multilevel** menu to load the **Title and Options** dialog box. Enter the string Binomial-Logit Model for Thailand data into the **Title** string box. Click the **Next** button to load the **ID and Weight Variables** dialog box.

Select the variable SCHOOLID by clicking on it. Click on the **Add** button of the **Level 2 ID Variable** section.

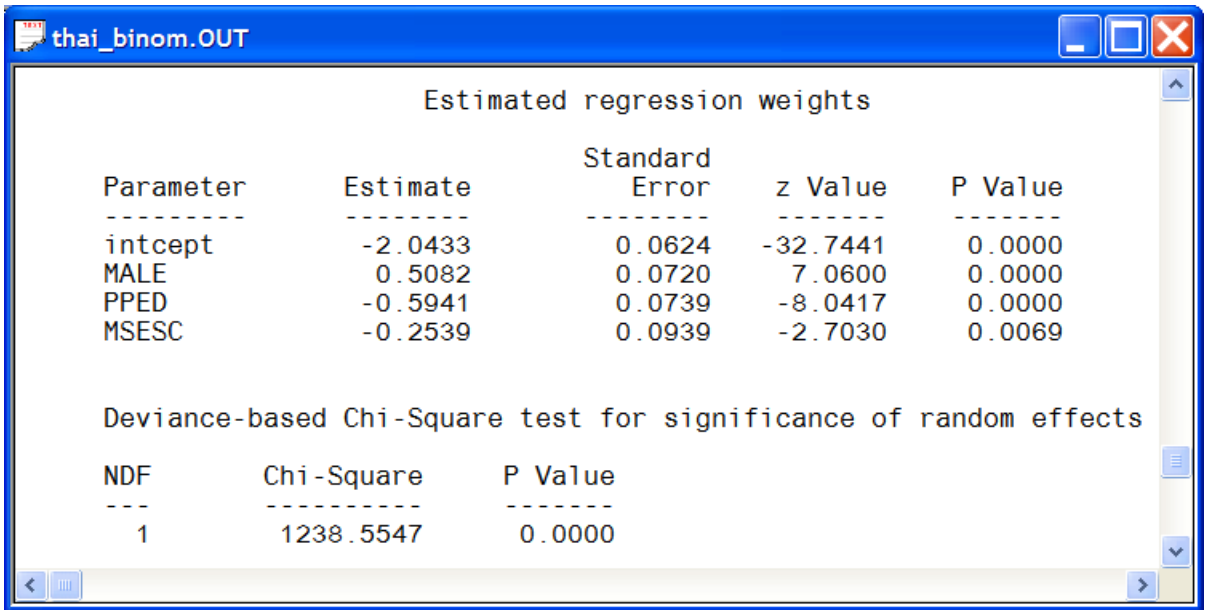
Click the **Next** button to load the **Distributions and Links** dialog box. Select the Binomial option from the **Distribution type** drop-down list box. Click the **Next** button to load the **Dependent and Independent Variables** dialog box.

Select the variable REP1 by clicking on it. Click on the **Add** button of the **Dependent variable** section. Select the variables MALE**,** PPED, and MSESC. Click on the **Continuous** button of the **Independent variables** section. Select the variable TRIAL by clicking on it. Click on the **Add** button of the **NTrials** section. Click the **Next** button to load the **Random Variables** dialog box.

Click on the **Finish** button to open the following text editor window for **thai_binom.prl**.



Click on the **Run Prelis** toolbar icon to produce the text editor window for **thai_binom.out**.



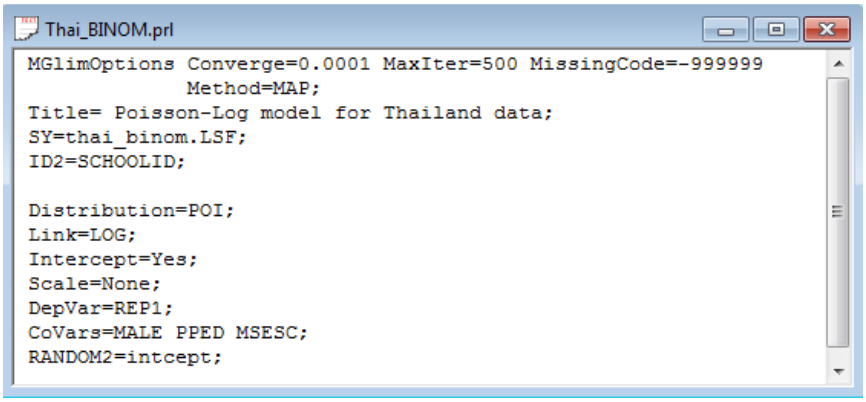
Fitting a Poisson-log model

Select the **Open** option on the **File** menu of the main window to load the **Open** dialog box. Select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box. Browse for and select the file **thai_binom.lsf** by clicking on it. Click on the **Open** button to open the file **thai_binom.lsf** in a LSF window.

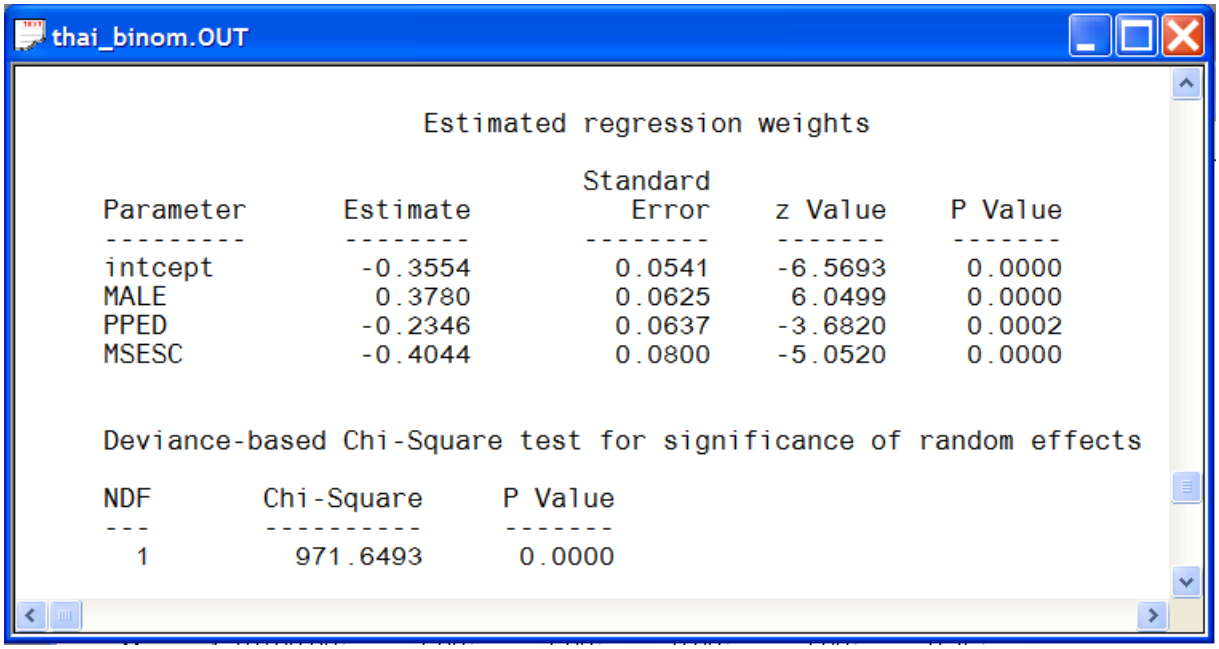
Select the **Title and Options** option on the **Generalized Linear Model** pop-up of the **Multilevel** menu to load the **Title and Options** dialog box. Enter the string Poisson-Log Model for Thailand data into the **Title** string box. Click the **Next** button to load the **ID and Weight Variables** dialog box.

Select the variable SCHOOLID by clicking on it. Click on the **Add** button of the **Level 2 ID Variable** section. Click the **Next** button to load the **Distributions and Links** dialog box. Select the Poisson option from the **Distribution type** drop-down list box. Click the **Next** button to load the **Dependent and Independent Variables** dialog box.

Select the variable REP1 by clicking on it. Click on the **Add** button of the **Dependent variable** section. Select the variables MALE, PPED, and MSESC. Click on the **Continuous** button of the **Independent variables** section. Click the **Next** button to load the **Random Variables** dialog box. Click on the **Finish** button to open the following text editor window for **thai_binom.prl**.



Click on the **Run Prelis** toolbar icon to produce the text editor window for **thai_binom.out**.



3 GLIMs for ordinal response variables

Researchers are often involved in studying ordinal response variables such as mental impairment (well, mild symptom formation, moderate symptom formation or impaired), patient satisfaction measured on a 5-point Likert scale, severity of lower back pain (none, mild, moderate or severe), arthritis improvement (none, some or marked), etc. In this section, we illustrate generalized linear modeling for ordinal response variables with MAPGLIM. A logit model and a cumulative logit model are fitted to data from the Television School and Family Smoking Prevention and Cessation Project (TVSFP).

The data

The data set forms part of the Television School and Family Smoking Prevention and Cessation Project (TVSFP). The subjects are 1600 students from 135 classrooms and 28 schools are included, where schools were randomized to one of four study conditions: a social-resistance classroom curriculum, a media (television) intervention, a social-resistance classroom curriculum combined with a mass-media intervention, and a no-treatment control group. These conditions form a 2 x 2 design of social-resistance classroom curriculum (CC [1=yes or 0=no]) by mass-media intervention (TV [1= yes or no]). THKSORD is a tobacco and health knowledge scale score. More information about this data set is provided in

http://tigger.uic.edu/~hedeker/Mixorcm.PDF

This data set is available as the LSF **tvsfpors.lsf**. The first portion of **tvsfpors.lsf** is shown in the following LSF window.

tvsfpors.lsf

	School	Class	THKSord	THKSbin	Intrcpt	PreTHKS	CC	TV	CC*TV
1	403.00	403101.00	3.00	1.00	1.00	2.00	1.00	0.00	0.00
2	403.00	403101.00	4.00	1.00	1.00	4.00	1.00	0.00	0.00
3	403.00	403101.00	3.00	1.00	1.00	4.00	1.00	0.00	0.00
4	403.00	403101.00	4.00	1.00	1.00	3.00	1.00	0.00	0.00
5	403.00	403101.00	4.00	1.00	1.00	3.00	1.00	0.00	0.00
6	403.00	403101.00	3.00	1.00	1.00	4.00	1.00	0.00	0.00
7	403.00	403101.00	2.00	0.00	1.00	2.00	1.00	0.00	0.00
8	403.00	403101.00	4.00	1.00	1.00	4.00	1.00	0.00	0.00
9	403.00	403101.00	4.00	1.00	1.00	5.00	1.00	0.00	0.00
10	403.00	403101.00	4.00	1.00	1.00	3.00	1.00	0.00	0.00
11	403.00	403101.00	3.00	1.00	1.00	3.00	1.00	0.00	0.00
12	403.00	403101.00	4.00	1.00	1.00	3.00	1.00	0.00	0.00
13	403.00	403101.00	3.00	1.00	1.00	1.00	1.00	0.00	0.00
14	403.00	403101.00	4.00	1.00	1.00	2.00	1.00	0.00	0.00
15	403.00	403101.00	2.00	0.00	1.00	2.00	1.00	0.00	0.00
16	403.00	403101.00	4.00	1.00	1.00	1.00	1.00	0.00	0.00
17	403.00	403101.00	4.00	1.00	1.00	4.00	1.00	0.00	0.00
18	403.00	403101.00	3.00	1.00	1.00	3.00	1.00	0.00	0.00
19	403.00	403101.00	3.00	1.00	1.00	0.00	1.00	0.00	0.00
	403.00	403101.00	4.00	1.00	1.00	3.00	1.00	0.00	0.00

School and Class are the identification variables for the school and the class of the student. CC*TV denotes the interaction variable between the variables CC and TV. PreTHKS is the student's knowledge scale score before the experiment.

Fitting a Multinomial-logit model

Select the **Open** option on the **File** menu of the main window to load the **Open** dialog box. Select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box. Browse for and select the file **tvsfpors.lsf** by clicking on it.

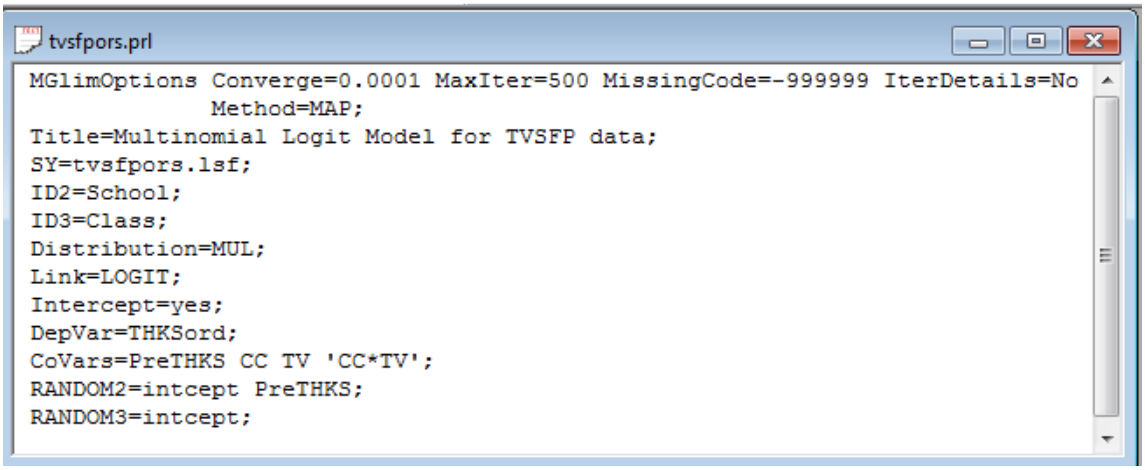
Select the **Title and Options** option on the **Generalized Linear Model** pop-up of the **Multilevel** menu to load the **Title and Options** dialog box. Enter the string Multinomial Logit Model for TVSFP data into the **Title** string box. Click the **Next** button to load the **ID and Weight Variables** dialog box.

Select the variable SCHOOL by clicking on it. Click on the **Add** button of the **Level 2 ID Variable** section. Select the variable **Class** by clicking on it. Click on the **Add** button of the **Level 3 ID Variable** section.

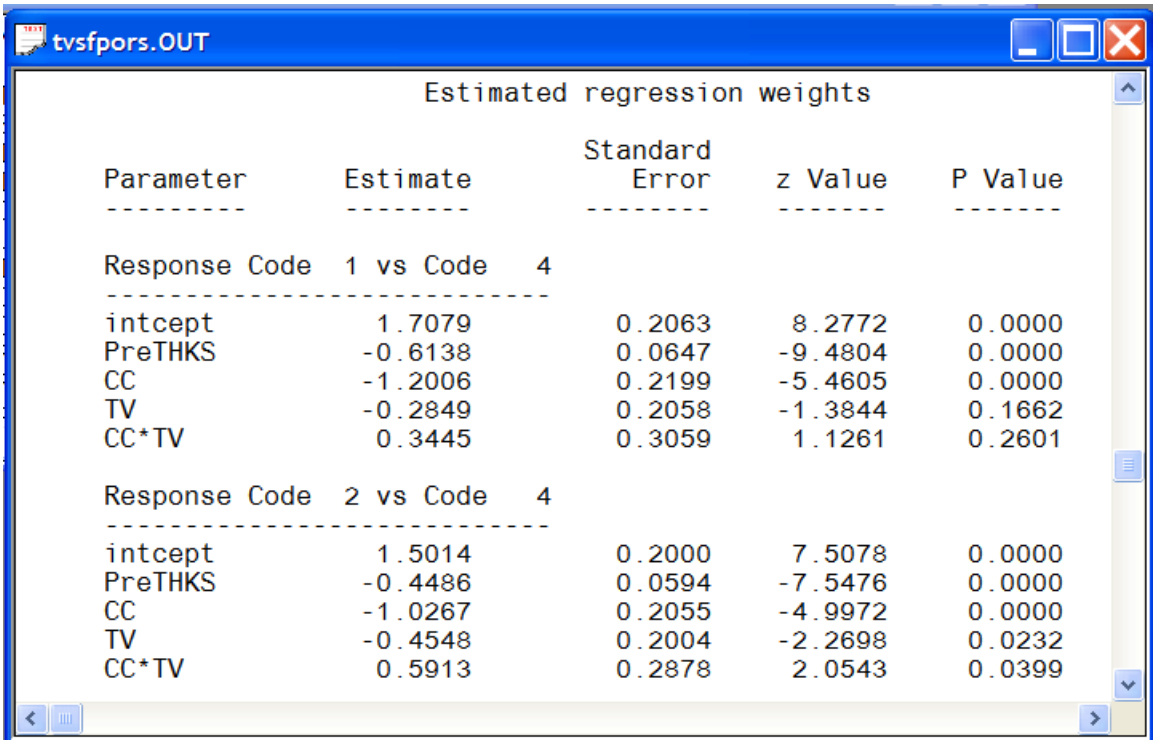
Click the **Next** button to load the **Distributions and Links** dialog box. Select the Poisson option from the **Distribution type** drop-down list box. Click the **Next** button to load the **Dependent and Independent Variables** dialog box.

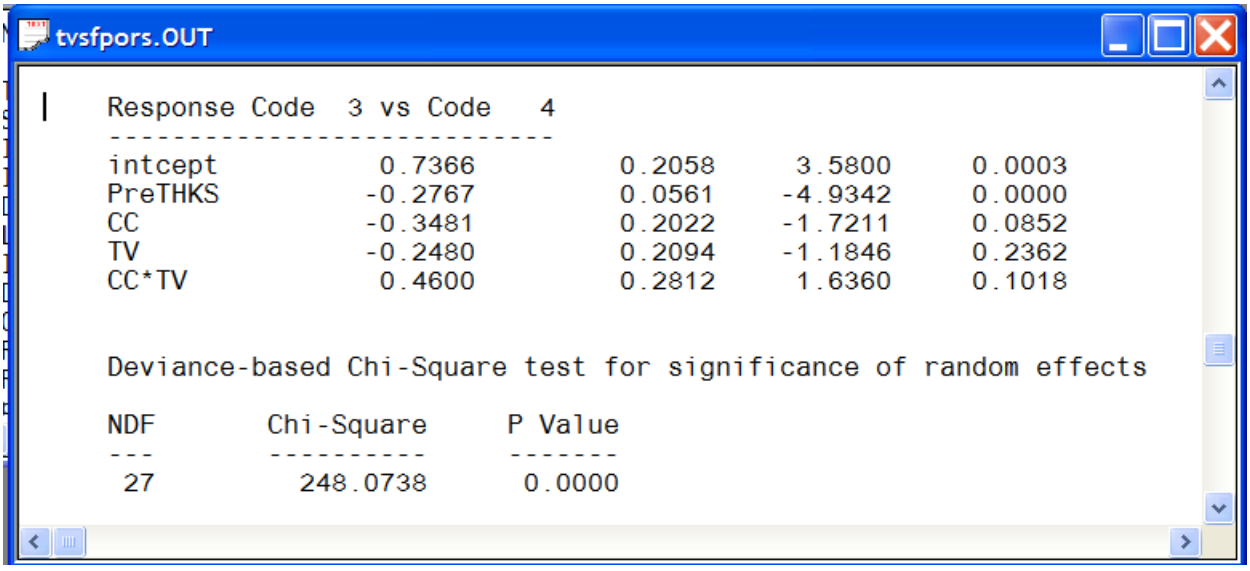
Select the variable THKSord by clicking on it. Click on the **Add** button of the **Dependent variable** section. Select the variables PreTHKS**,** CC, TV, and CC*TV. Click on the **Continuous** button of the **Independent variables** section. Click the **Next** button to load the **Random Variables** dialog box.

Select the variable PreTHKS by clicking on it. Click on the **Add** button of the **Random Level 2** section. Click on the **Finish** button to open the following text editor window for **tvsfpors.prl**.



Click on the **Run Prelis** toolbar icon to produce the results displayed in the following two text editor windows.





Fitting a Multinomial-cumulative logit model

Select the **Open** option on the **File** menu of the main window to load the **Open** dialog box. Select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box. Browse for and select the file **tvsfpors.lsf** by clicking on it.

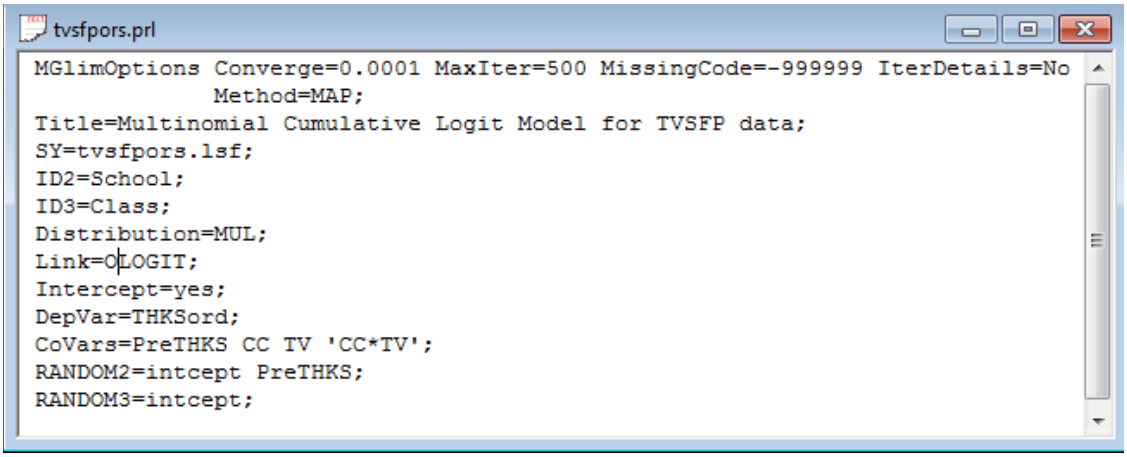
Select the **Title and Options** option on the **Generalized Linear Model** pop-up of the **Multilevel** menu to load the **Title and Options** dialog box. Enter the string Multinomial Cumulative Logit Model for TVSFP data into the **Title** string box. Click the **Next** button to load the **ID and Weight Variables** dialog box.

Select the variable SCHOOL by clicking on it. Click on the **Add** button of the **Level 2 ID Variable** section. Select the variable **Class** by clicking on it. Click on the **Add** button of the **Level 3 ID Variable** section.

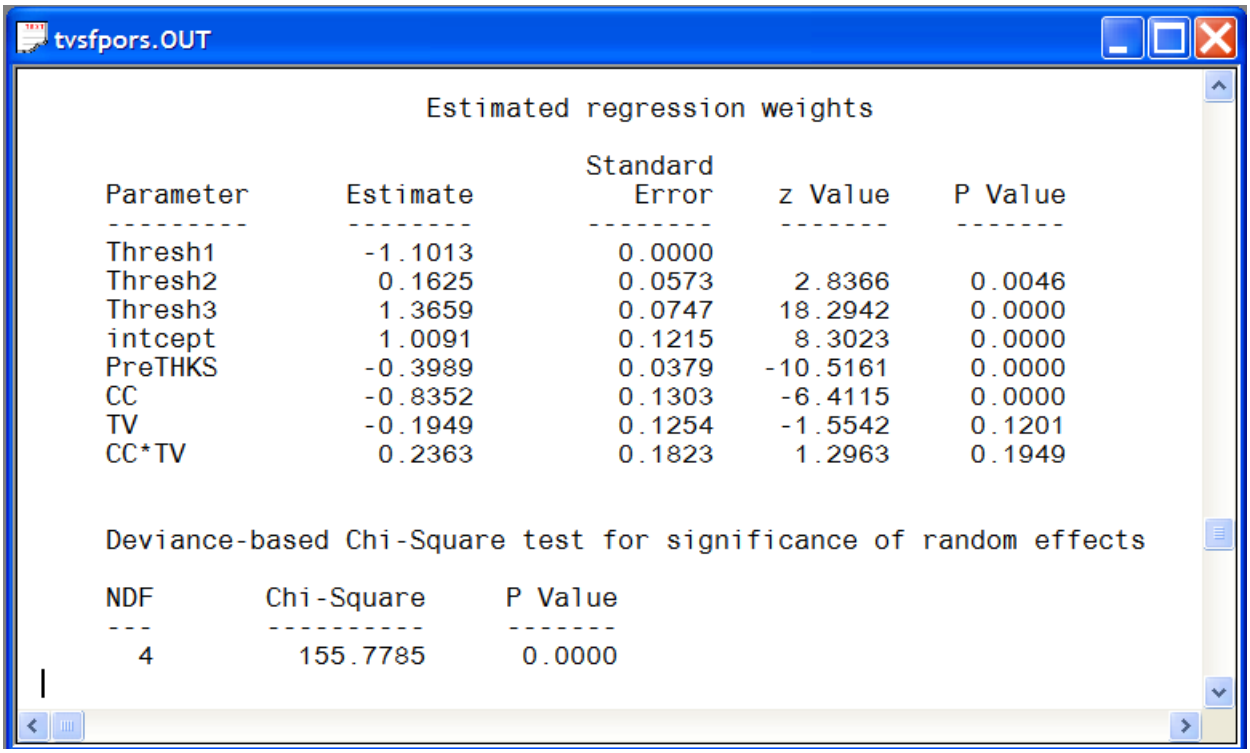
Click the **Next** button to load the **Distributions and Links** dialog box. Select the Poisson option from the **Distribution type** drop-down list box. Click the **Next** button to load the **Dependent and Independent Variables** dialog box.

Select the variable THKSord by clicking on it. Click on the **Add** button of the **Dependent variable** section. Select the variables PreTHKS**,** CC, TV, and CC*TV. Click on the **Continuous** button of the **Independent variables** section. Click the **Next** button to load the **Random Variables** dialog box.

Select the variable PreTHKS by clicking on it. Click on the **Add** button of the **Random Level 2** section. Click on the **Finish** button to open the following text editor window for **tvsfpors.prl**.



Click on the **Run Prelis** toolbar icon to produce the results displayed in the following text editor window.


tvsfpors.OUT
Estimated regression weights
Parameter Estimate Standard Error z Value P Value
Thresh1 -1.1013 0.0000
Thresh2 0.1625 0.0573 2.8366 0.0046
Thresh3 1.3659 0.0747 18.2942 0.0000
intcept 1.0091 0.1215 8.3023 0.0000
PreTHKS -0.3989 0.0379 -10.5161 0.0000
CC -0.8352 0.1303 -6.4115 0.0000
TV -0.1949 0.1254 -1.5542 0.1201
CC*TV 0.2363 0.1823 1.2963 0.1949
Deviance-based Chi-Square test for significance of random effects
NDF Chi-Square P Value
4 155.7785 0.0000