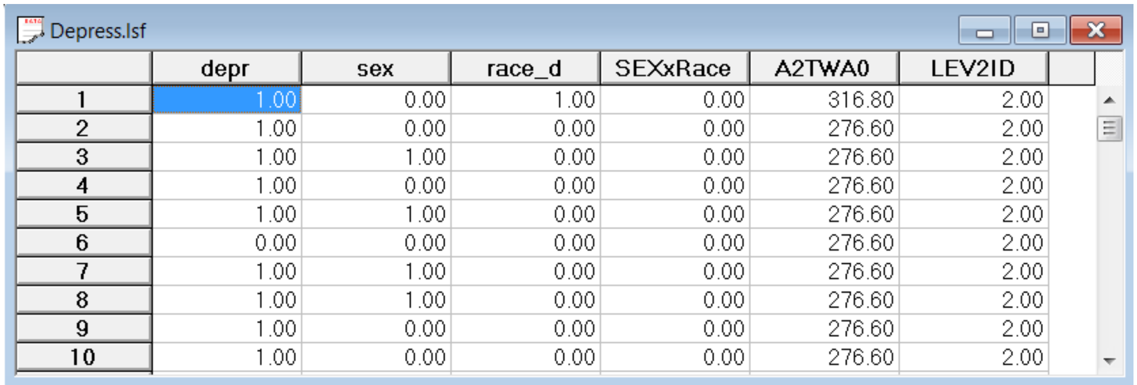
Binary models with logit link function

Contents

1. The data

The data set forms part of the data library of the Alcohol and Drug Services Study (ADSS). The ADSS is a national study of substance abuse treatment facilities and clients. Background data and data on the substance abuse of a sample of 1752 clients were obtained. The sample was stratified by census region (CENREG) and within each stratum a sample was obtained for each of three facility treatment types (FACTYPE) within each of the four census regions. More information on the ADSS and the data are available at http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/03088.xml.

The specific data set is provided in the **Multilevel Generalized Linear Model Examples** folder as the LSF file **Depress.LSF**. The first portion of this file is shown in the following LSF window.



Depress.lsf

	depr	sex	race_d	SEXxRace	A2TWA0	LEV2ID
1	1.00	0.00	1.00	0.00	316.80	2.00
2	1.00	0.00	0.00	0.00	276.60	2.00
3	1.00	1.00	0.00	0.00	276.60	2.00
4	1.00	0.00	0.00	0.00	276.60	2.00
5	1.00	1.00	0.00	0.00	276.60	2.00
6	0.00	0.00	0.00	0.00	276.60	2.00
7	1.00	1.00	0.00	0.00	276.60	2.00
8	1.00	1.00	0.00	0.00	276.60	2.00
9	1.00	0.00	0.00	0.00	276.60	2.00
10	1.00	0.00	0.00	0.00	276.60	2.00

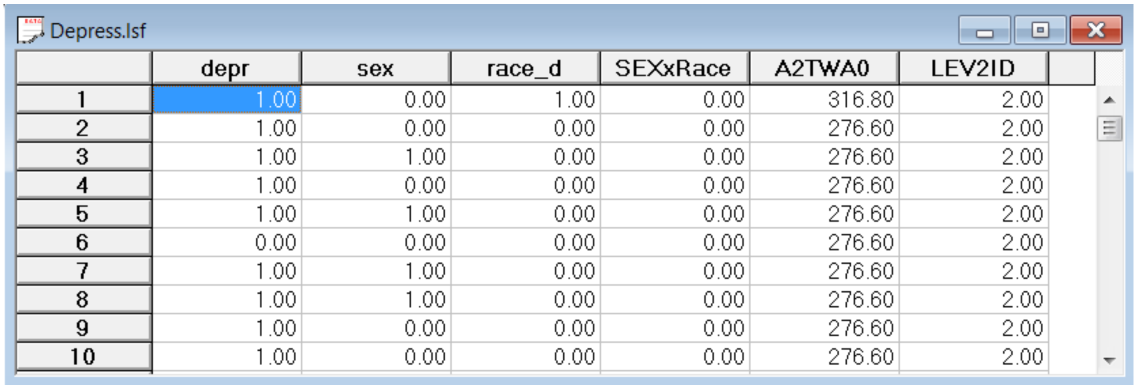
The variables of interest are:

- DEPR addresses the question of whether the patient has depression. (1 = Yes; 0 = No)
- A2TWA0 is the sampling weight
- SEX is a dummy variable indicating the gender (0 for male and 1 for female) of the client
- RACE_D is a dummy variable representing the ethnicity (0 for black and 1 for white) of the client
- SEXxRACE is the interaction term of gender and race.
- LEV2ID is the variable used to identify the level-2 ID or grouping variable.

2. Importing the data

The data set shown previously is available in the form of a spreadsheet file, named **depress.lsf**. This file contains data for the 2,214 respondents who reported some form of depression.

The first step is to create the LISREL spreadsheet file **(lsf)** from the Excel file. Use the **Import Data File** option on the **File** menu to load the **Open** dialog box. Select **Excel (*.xls)** from the **Files of type** drop-down list. Browse for the file **depress.xls**, and select the file and click on the **Open** button to open the following LISREL spreadsheet window for **depress.lsf**.



	depr	sex	race_d	SEXxRace	A2TWA0	LEV2ID
1	1.00	0.00	1.00	0.00	316.80	2.00
2	1.00	0.00	0.00	0.00	276.60	2.00
3	1.00	1.00	0.00	0.00	276.60	2.00
4	1.00	0.00	0.00	0.00	276.60	2.00
5	1.00	1.00	0.00	0.00	276.60	2.00
6	0.00	0.00	0.00	0.00	276.60	2.00
7	1.00	1.00	0.00	0.00	276.60	2.00
8	1.00	1.00	0.00	0.00	276.60	2.00
9	1.00	0.00	0.00	0.00	276.60	2.00
10	1.00	0.00	0.00	0.00	276.60	2.00

Besides EXCEL data files, LISREL is capable of importing SAS, SPSS, STATA and most of the data files in other formats. The data import processes are similar, and will not be discussed again in this document.

When the external data is imported into LISREL, the default variable type is Ordinal. Variables that have more than 15 categories are treated as continuous. To change the default settings for the variable type, click on the **Data**, **Define** variables menu and change the settings. In this example the default setting is valid, and no further changes are needed.

3. Exploring the data

Graphics are often a useful data-exploring technique through which the researcher may familiarize her- or himself with the data. Relationships and trends may be conveyed in an informal and simplified visual form via graphical displays.

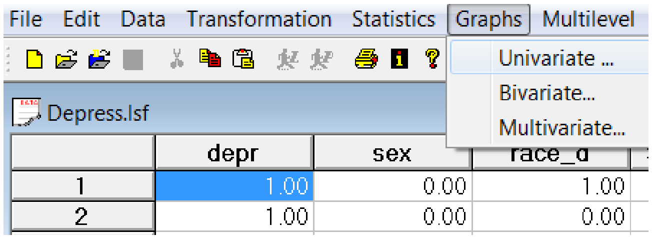
The **Graphs** option provides univariate, bivariate, and multivariate graphs. Univariate graphs are particularly useful in obtaining an overview of the characteristics of a variable.

Univariate graphs

As a first step, we take a look at the distribution of depression (DEPR), which is the potential dependent variable in this study.

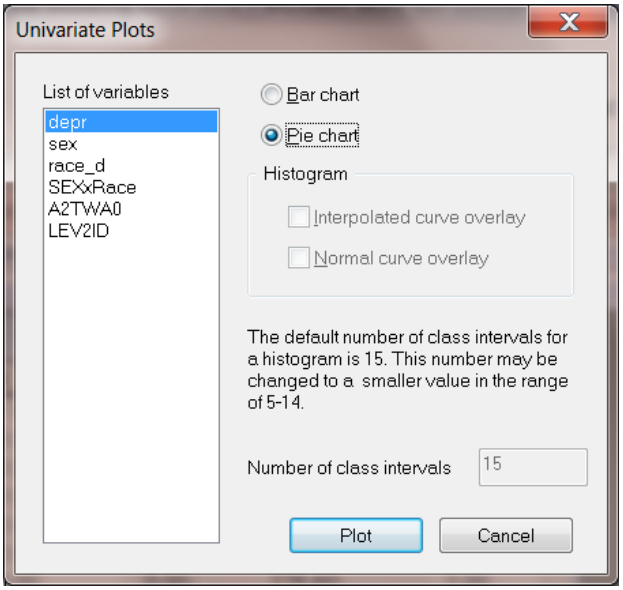
Pie chart

A pie chart gives a good picture of probability of success. To create a pie chart for DEPR, select the **Univariate** option from the **Graphs** menu as shown below.



	depr	sex	race_d
1	1.00	0.00	1.00
2	1.00	0.00	0.00

The **Univariate plot** dialog box appears. Select the variable DEPR and indicate that a **Pie chart** is to be graphed. Click the **Plot** button to display the pie chart.



List of variables

- depr
- sex
- race_d
- SEXxRace
- A2TWA0
- LEV2ID

○ Bar chart

● Pie chart

Histogram

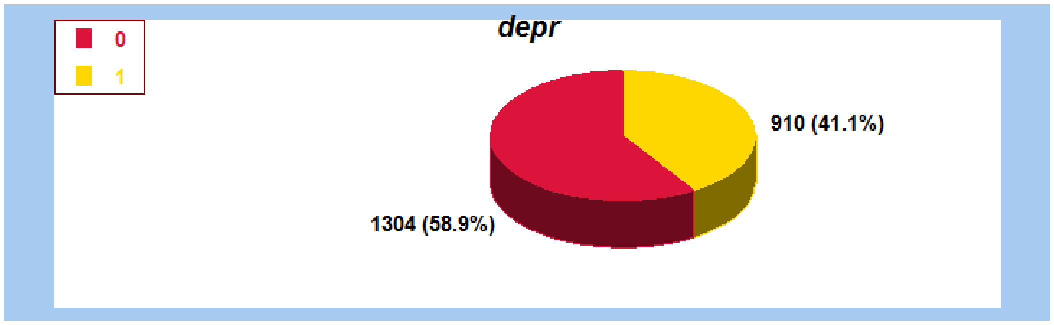
□ Interpolated curve overlay

□ Normal curve overlay

The default number of class intervals for a histogram is 15. This number may be changed to a smaller value in the range of 5-14.

Number of class intervals 15

The pie below shows that about 58.9% of the respondents have depression. Since the probability is not extremely large or small, the Bernoulli distribution should be appropriate for our study.



4. The model

The first model fitted to the depression data explores the relationship between DEPR, gender, and race, as represented by the variables SEX and RACE_D. The level-1 model is at a individual level, while the level-2 model is at a PSU level. The model can be expressed as follows.

For the binary case with logit link considered here

$$\text{Prob}(\text{DEPR}_{ij} = 1) = \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}$$

where η_{ij} represents the log of the odds of success. With the logit link function, the probability $\text{Prob}(y_{ij} = 1 | \boldsymbol{\beta})$ is transformed to lie in the interval (0,1). And (for the current model) the two-level model can be expressed as

Level-1 model:

$$\eta_{ij} = b_{0i} + b_{1i} \times (\text{SEX})_{ij} + b_{2i} \times (\text{RACE\_d})_{ij} + e_{ij}$$

Level-2 model:

$$b_{0i} = \beta_0 + u_{0i}$$
$$b_{1i} = \beta_1 + u_{1i}$$
$$b_{2i} = \beta_2$$

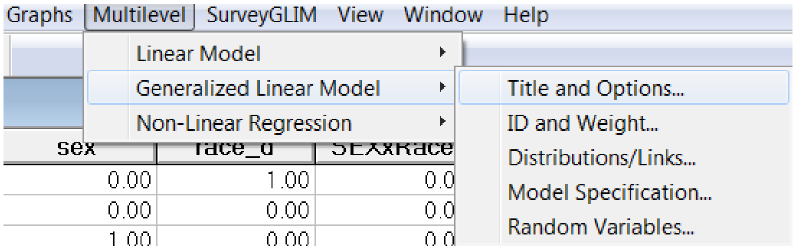
where

$$e_i \sim N(0, \sigma^2 \mathbf{I}_i)$$
$$\mathbf{u}_i \sim N(0, \mathbf{\Sigma}_i)$$

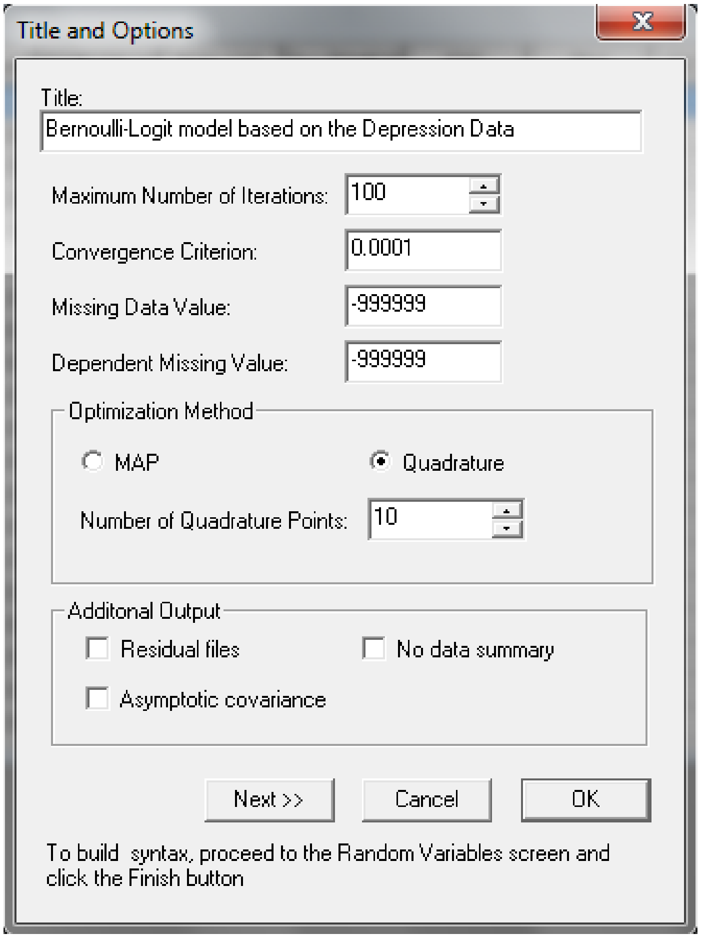
β_0 denotes the average expected η_{ij}, which can be converted to the expected probability of getting depression. β_1 denotes the coefficient of the predictor variable SEX (slope) in the fixed part of the model. The random coefficients u_{i0} and e_{ij} denote the variation in the average expected DEPR value between PSUs and between patients respectively.

5. Setting up the analysis

Open the LISREL spreadsheet **depress.lsf** used during the exploratory analysis discussed previously. The next step is to describe the model to be fitted. We use the LISREL interface to provide the model specifications. From the main menu bar, select the **Multilevel, Generalized Linear Model**, **Title and Options** option.



The multilevel generalized linear model contains five consequential dialogs boxes. The **Titles and Options** dialog box as shown below enables the user to input the title, maximum number of iteration, convergence criterion, missing values, and method and request additional output. Enter a title for the analysis in the **Title** text boxes (optional) and keep all the other settings as default.


Title and Options

Title: Bernoulli-Logit model based on the Depression Data

Maximum Number of Iterations: 100

Convergence Criterion: 0.0001

Missing Data Value: -999999

Dependent Missing Value: -999999

Optimization Method

○ MAP ● Quadrature

Number of Quadrature Points: 10

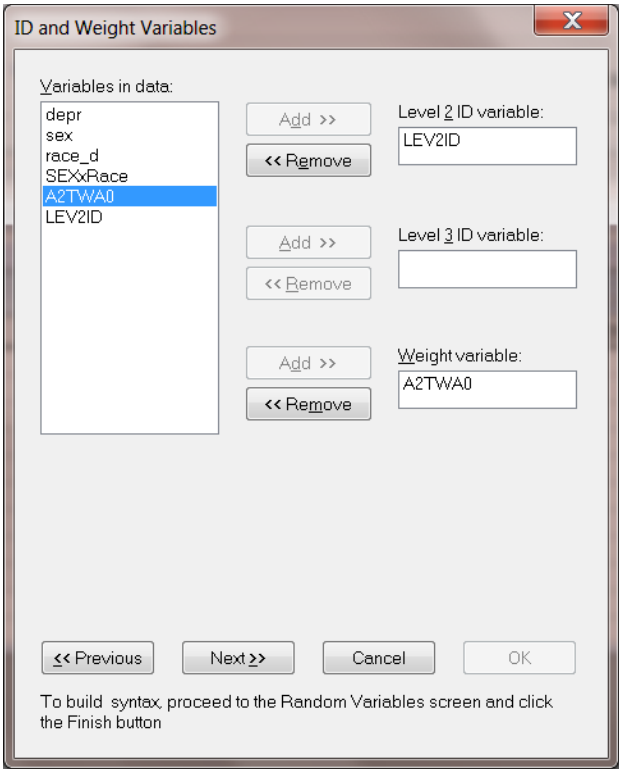
Additonal Output

□ Residual files □ No data summary

□ Asymptotic covariance

To build syntax, proceed to the Random Variables screen and click the Finish button

Proceed to the **ID and Weights** screen by clicking on the **Next** button. Highlight LEV2ID from the **Variables in data** list and click on the upper **Add** button to select is as the **Level-2 ID variable.** Similarly, highlight the variable A2TWA0 and click on the lower **Add** button to select it as the **Weight variable** and obtain the screen shown below.



Variables in data:
depr
sex
race_d
SEXxRace
A2TWA0
LEV2ID

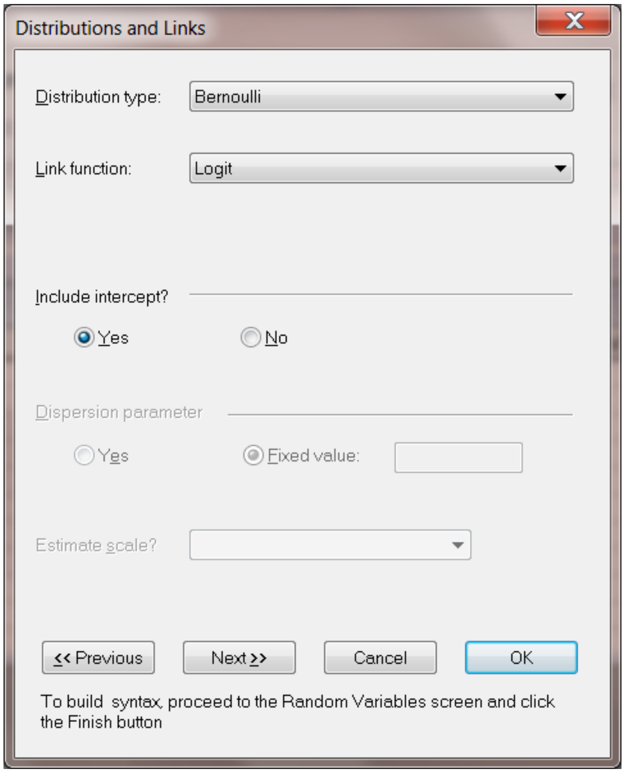
Level 2 ID variable: LEV2ID

Level 3 ID variable:

Weight variable: A2TWA0

To build syntax, proceed to the Random Variables screen and click the Finish button

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



Distribution type: Bernoulli

Link function: Logit

Include intercept?

● Yes ○ No

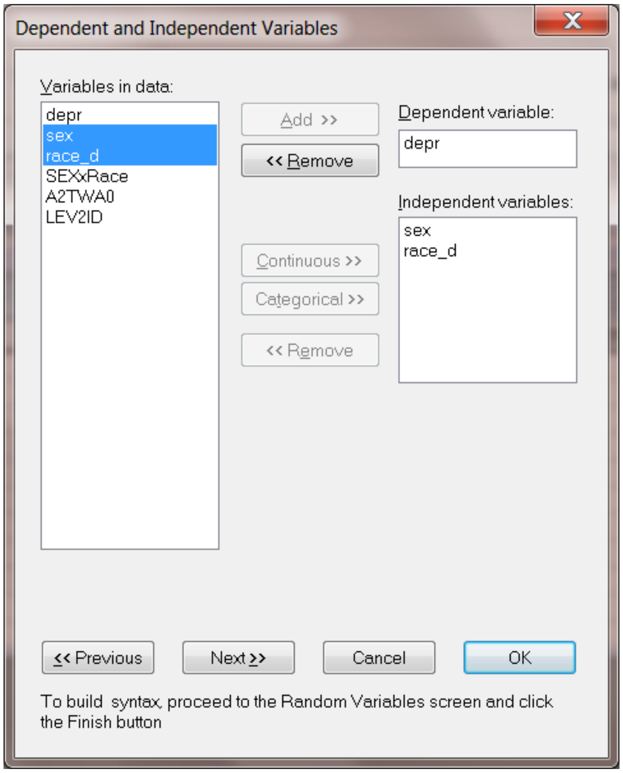
Dispersion parameter

○ Yes ● Fixed value:

Estimate scale?

To build syntax, proceed to the Random Variables screen and click the Finish button

On the **Dependent and Independent Variables** dialog box screen, first select DEPR and click on the upper **Add** button to define it as the **Dependent variable**. Then, select SEX and RACE_D and click on the **Continuous** button to add these variables in the **Independent variables** list box as shown below.



Variables in data:

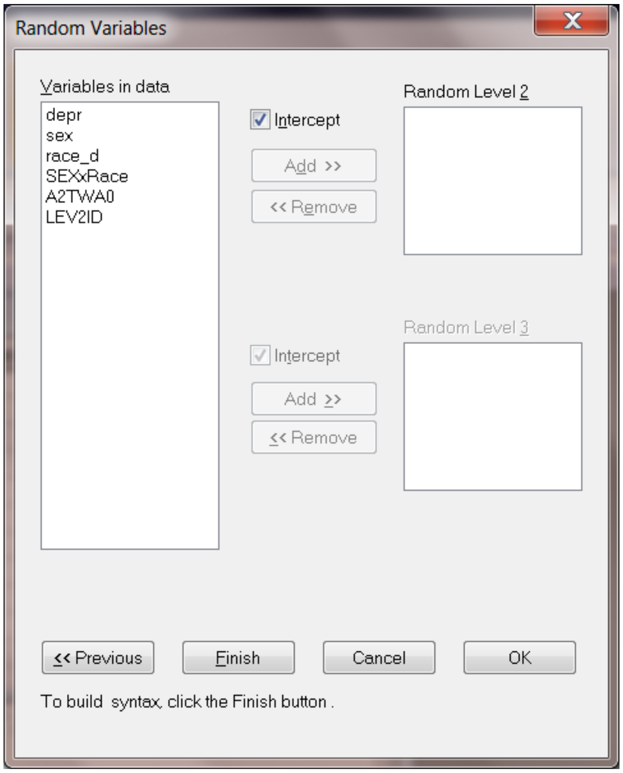
depr
sex
race_d
SEXxRace
A2TWA0
LEV2ID

Dependent variable: depr

Independent variables:

sex
race_d

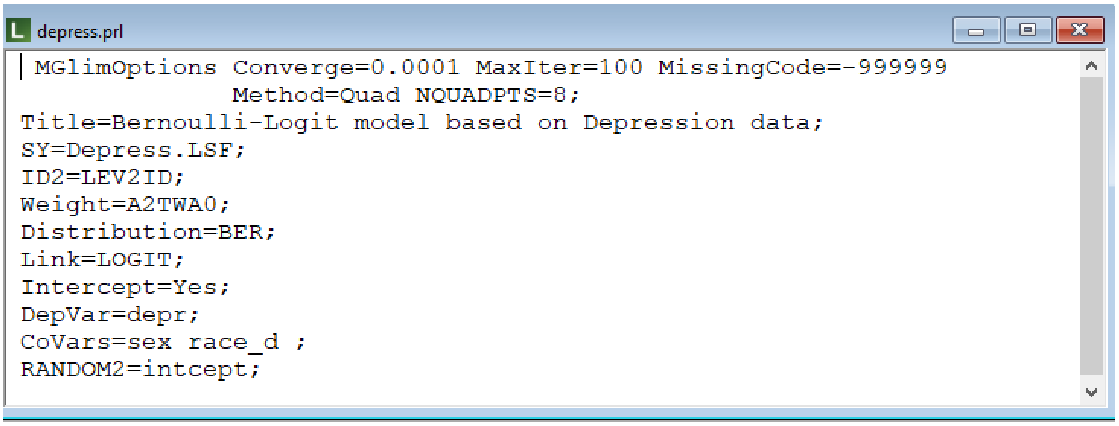
To build syntax, proceed to the Random Variables screen and click the Finish button



To build syntax, click the Finish button .

Click on the **Next** button to proceed to the **Random Variables** dialog box once these settings have been defined. Keep the **Intercept** check box checked so as to include a level-2 intercept.

Click on the **Finish** button to generate the PRELIS syntax file (**.prl**) that corresponds to the above settings. Select the **File**, **Save As** option, and provide a name (**depress1a.prl**) for the model specification file. The default folder for the syntax to be saved in is the same folder used for the data file.



```
MGlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999
               Method=Quad NQUADPTS=8;
Title=Bernoulli-Logit model based on Depression data;
SY=Depress.LSF;
ID2=LEV2ID;
Weight=A2TWA0;
Distribution=BER;
Link=LOGIT;
Intercept=Yes;
DepVar=depr;
CoVars=sex race_d ;
RANDOM2=intcept;
```

The syntax file

The syntax file contains the following information:

- The MGlimOptions keyword requests the MGLIM module to run. The first two lines, together with the Title line, correspond to the settings entered in the **Title and Options** dialog box.
- The SY line indicates the location of the **.lsf** data file.
- ID2 is the level-2 id variable, while Weight corresponds with the weight variable. These are defined in the **ID and Weights** dialog box.
- The syntax lines for Distribution, Link and level-1 Intercept are set up in the **Distribution and Links** dialog box.
- The DepVar line, which represents the dependent variable and the CoVars line, which represents the covariate variable, are defined in the **Dependent and Independent Variables** dialog box.
- Finally, the RANDOM2 syntax line corresponds to the **Random Variables** dialog box.

Understanding how the syntax works enables the user to make changes directly to the syntax file. Run the analysis by selecting the **Run PRELIS** button to generate the output file **depress.out**. 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.

6. Discussion of results

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

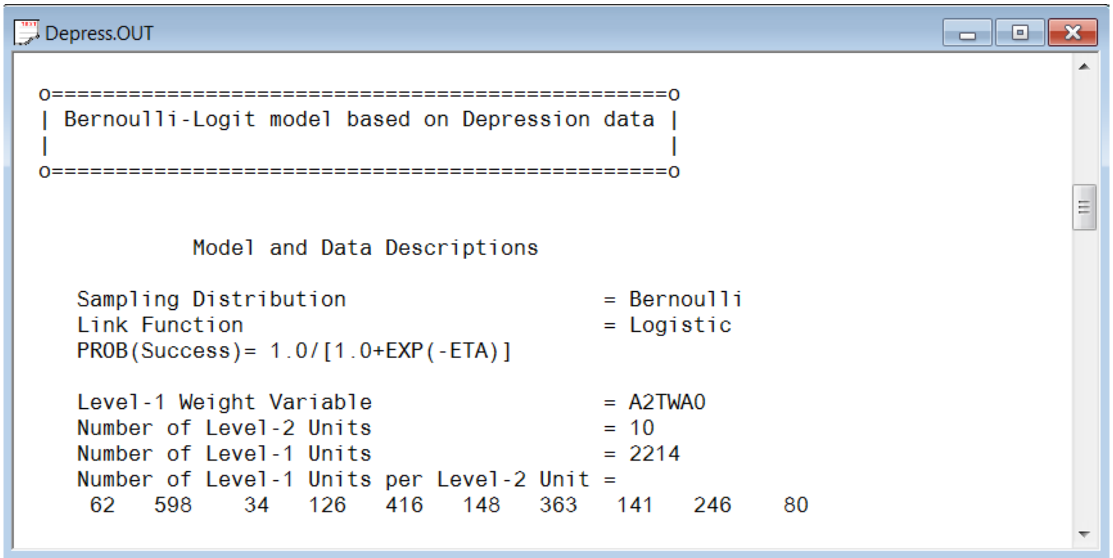
Program information and syntax

At the top of the output file, program information is given. It states the version number, corporate and technical support information, the date and time of analysis, and the locations of data file and syntax file.

Program information is followed by the Multilevel GLIM syntax. This section echoes the contents of the syntax file **depress.prl**. For more information on syntax and keywords, please see Section 2.2.3.

Model and data description

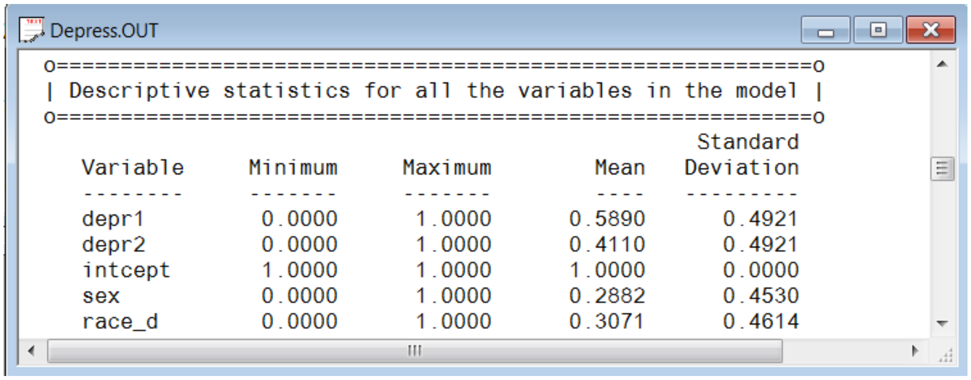
In the next section of the output file as shown above, descriptions of the distribution, the link function, the weight variable and the hierarchical structure of the data are provided. Data from a total of 10 level-2 units and 2,214 respondents were included at levels 2 and 1 of the model. In addition, a summary of the number of respondents nested within each level-2 unit is provided.



```
o===============================================o
| Bernoulli-Logit model based on Depression data |
|                                                |
o===============================================o


            Model and Data Descriptions

   Sampling Distribution                      = Bernoulli
   Link Function                              = Logistic
   PROB(Success)= 1.0/[1.0+EXP(-ETA)]

   Level-1 Weight Variable                    = A2TWA0
   Number of Level-2 Units                    = 10
   Number of Level-1 Units                    = 2214
   Number of Level-1 Units per Level-2 Unit =
    62   598    34   126   416   148   363   141   246    80
```

Descriptive statistics

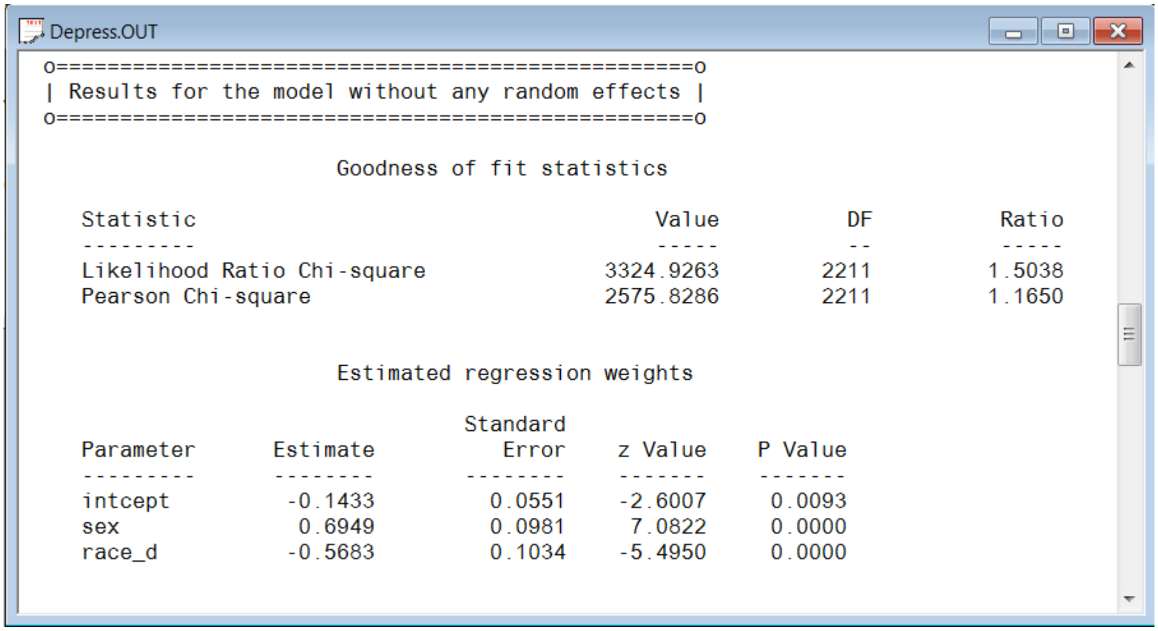
The data summary is followed by descriptive statistics for all the variables included in the model. Since DEPR is defined as a binary variable, it is presented by two dummy variables depr1 and depr2.



```
o=========================================================o
| Descriptive statistics for all the variables in the model |
o=========================================================o
                                                     Standard
   Variable      Minimum      Maximum         Mean  Deviation
   --------      -------      -------         ----  ---------
   depr1          0.0000       1.0000       0.5890     0.4921
   depr2          0.0000       1.0000       0.4110     0.4921
   intcept        1.0000       1.0000       1.0000     0.0000
   sex            0.0000       1.0000       0.2882     0.4530
   race_d         0.0000       1.0000       0.3071     0.4614
```

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 weighted least squares (WLS) regression. The goodness of WLS fit statistics are also given as shown below.

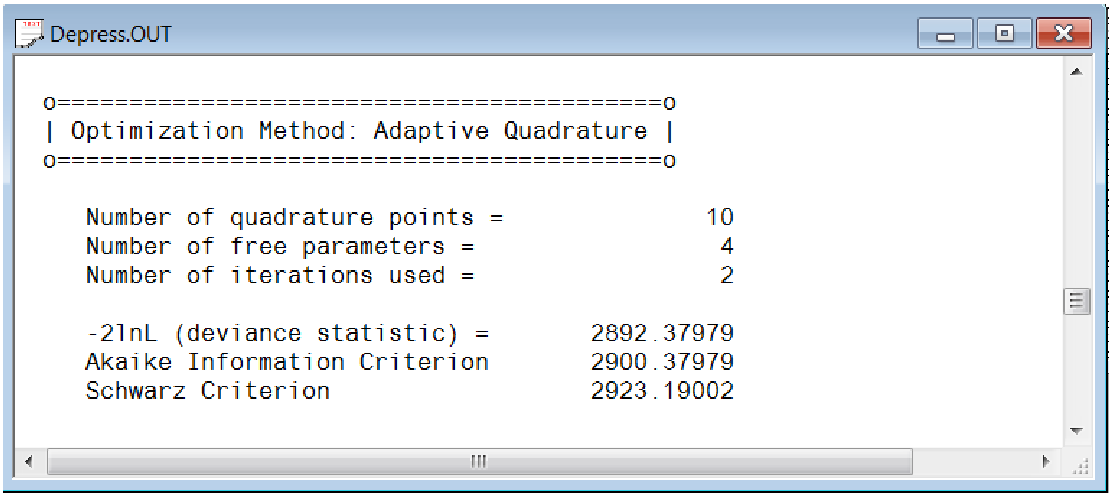


```
o==================================================o
| Results for the model without any random effects |
o==================================================o

                    Goodness of fit statistics

 Statistic                                 Value        DF        Ratio
 ---------                                 -----        --        -----
 Likelihood Ratio Chi-square           3324.9263      2211       1.5038
 Pearson Chi-square                    2575.8286      2211       1.1650


                    Estimated regression weights

                                Standard
 Parameter      Estimate           Error     z Value    P Value
 ---------      --------        --------     -------    -------
 intcept         -0.1433          0.0551     -2.6007     0.0093
 sex              0.6949          0.0981      7.0822     0.0000
 race_d          -0.5683          0.1034     -5.4950     0.0000
```

Results for the model with fixed and random effects
Number of iterations and fit statistics

The total number of (macro) iterations is reported. As shown below, there are 58 iterations to get the converged results.

In addition to the likelihood function value at convergence, a number of related statistical measures for assessing model adequacy are available. The most common of these are the likelihood ratio test, Pearson 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. A chi-square scale factor, with which a chi-square value obtained from the difference between two deviance statistics should be multiplied to yield a corrected chi-square statistic in the case of a weighted analysis, may also be found in this section.

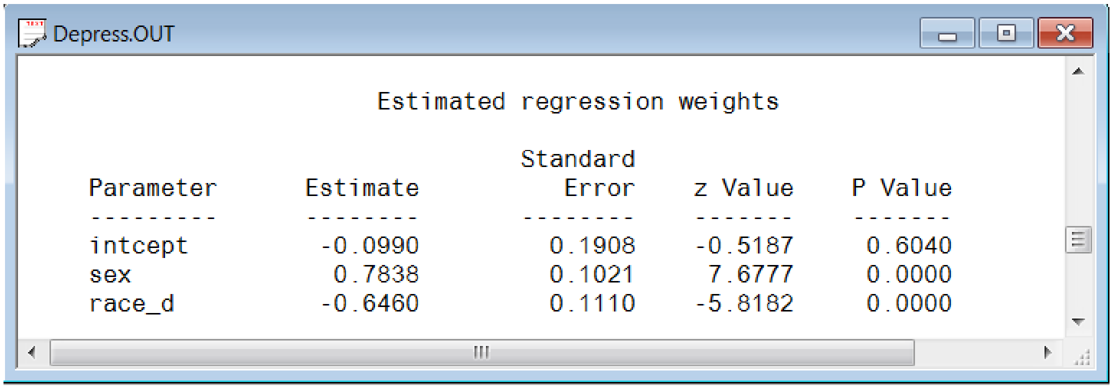


```
o==========================================o
| Optimization Method: Adaptive Quadrature |
o==========================================o

  Number of quadrature points =          10
  Number of free parameters =             4
  Number of iterations used =             2

  -2lnL (deviance statistic) =    2892.37979
  Akaike Information Criterion    2900.37979
  Schwarz Criterion               2923.19002
```

- The Pearson Chi-square is defined as $\chi_P^2 = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} \frac{w_{ijk}\left(y_{ijk} - \hat{\mu}_{ijk}\right)^2}{\hat{\sigma}^2\left(y_{ijk}\right)}$.
- The deviance is defined as $-2\ln L$. 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.
- The AIC was originally proposed for time-series models, but is also used in regression. It is defined as $-2\ln L + 2r$, where r denotes the number of parameters estimated in the model. The model with minimum AIC, in a set of nested models, will be the most parsimonious according to this criterion.
- The SBC is defined as $-2\ln L + r\log n$, where n denotes the number of units at the highest level of the hierarchy. A smaller value of this criterion would indicate the most parsimonious of the models being compared.

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 and β_2 in the model specification. From the z-values and associated exceedance probabilities, we see that all three estimates are highly significant at 10% level.



```
                    Estimated regression weights

                                   Standard
Parameter        Estimate           Error      z Value      P Value
---------        --------         --------     -------      -------
intcept           -0.0990           0.1908     -0.5187       0.6040
sex                0.7838           0.1021      7.6777       0.0000
race_d            -0.6460           0.1110     -5.8182       0.0000
```

The estimated intercept is -0.0990, which is the average logit. The estimated coefficients associated with gender (SEX) is – 0.7838, which indicates that the female respondents (SEX = 1) have a smaller $\hat{\eta}$. The estimate for the indicator of race (RACE_D) shows that white clients have higher $\hat{\eta}$ value. To describe the η's in a more accessible way to readers of reports, we need the link functions to transform them into probabilities.

Interpreting estimated regression weights by using link function

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

$$\begin{aligned}\hat{\eta}_{ij} &= \hat{b}_{0i} + \hat{b}_{1i} \times \left(\text{SEX}\right)_{ij} + \hat{b}_{2i} \times \left(\text{RACE\_d}\right)_{ij} \\ &= -0.0990 + 0.7838 \times \left(\text{SEX}\right)_{ij} - 0.6460 \times \left(\text{RACE\_d}\right)_{ij}.\end{aligned}$$

For a black male, we have SEX = 0, RACE_d = 0, thus

$$\begin{aligned}\hat{\eta}_{ij} &= -0.0990 + 0.7838 \times 0 - 0.6460 \times 0 \\ &= -0.0990\end{aligned} \quad .$$

Similarly, the calculation of $\hat{\eta}_{ij}$ for a BLACK female (SEX = 1, RACE_d = 0) is

$$\begin{aligned}\hat{\eta}_{ij} &= -0.0990 + 0.7838 \times 1 - 0.6460 \times 0 \\ &= 0.6848\end{aligned} \quad .$$

Next, we transform the $\hat{\eta}_{ij}$'s into corresponding probabilities by using the logit link function. Take the black male as the example, the probability is calculated as below.

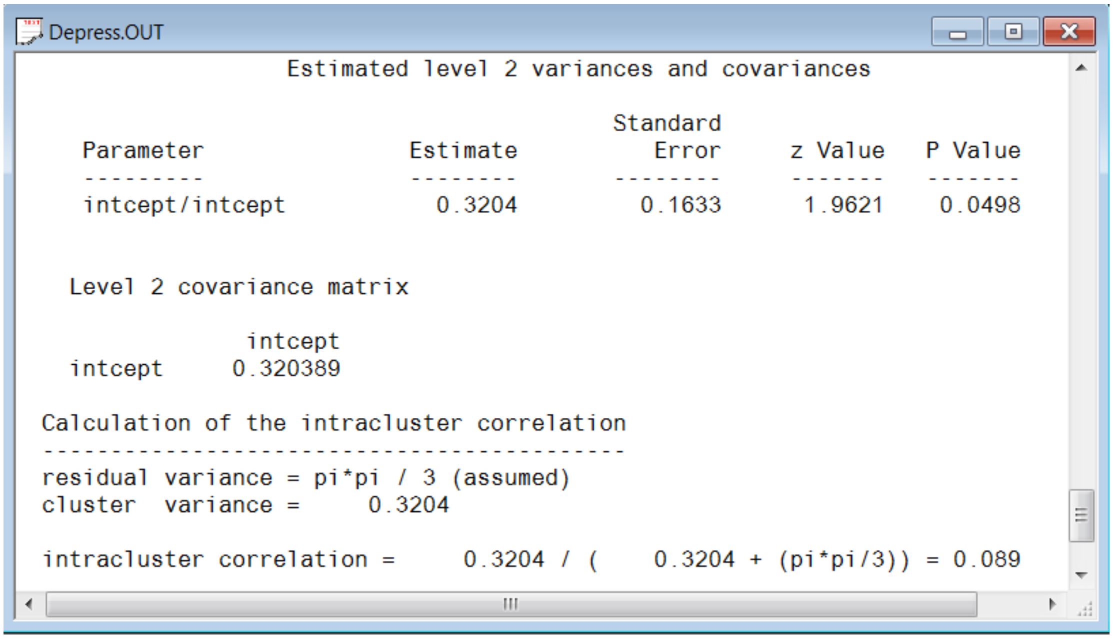
$$\text{Prob}(\text{DEPR}_{ij} = 1) = \frac{e^{\hat{\eta}_{ij}}}{1 + e^{\hat{\eta}_{ij}}} = \frac{1}{1 + e^{-\hat{\eta}_{ij}}} = \frac{1}{1 + e^{0.0990}} = 47.53\%$$

Thus we can conclude that the estimated probability for a black male who has depression is about 47.53%. Similarly, the probability of having depression for different gender and ethnicity are calculated in the following table.

Group	**Code**	$\hat{\eta}$	**Prob (DEPR = 1)**
Black, male	sex = 0, race_d = 0	0.1018	47.53%
Black, female	sex = 1, race_d = 0	0.6848	66.48%
White, male	sex = 0, race_d = 1	-0.7450	32.19%
White, female	sex = 1, race_d = 1	0.0388	50.97%

Estimated level-2 variance

The output for the estimated level-2 variance is shown in the image below. The *p* value of intercept shows the probability of getting depression differs significantly from PSU to PSU (the level-2 units).



```
                   Estimated level 2 variances and covariances

                                            Standard
   Parameter                Estimate           Error       z Value    P Value
   ---------                --------        --------       -------    -------
   intcept/intcept            0.3204          0.1633        1.9621     0.0498


  Level 2 covariance matrix

               intcept
  intcept     0.320389

 Calculation of the intracluster correlation
 -------------------------------------------
 residual variance = pi*pi / 3 (assumed)
 cluster  variance =     0.3204

 intracluster correlation =     0.3204 / (     0.3204 + (pi*pi/3)) = 0.089
```

ICCs and % variance explained

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

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

In the case of a model with only a random intercept, the variation in the random intercept at the level-2 unit, and the residual variation at level-1. The intracluster coefficient is defined as

$$ICC = \frac{\hat{\text{var}}(u_{i0})}{\hat{\text{var}}(e_{ij}) + \hat{\text{var}}(u_{i0})}$$

Level-1 variation

As mentioned earlier, for the dichotomous outcome model, it is assumed that the level-1 error variance is equal to $\pi^2/3$ for the logistic link function if the model is true (see, e.g., Hedeker & Gibbons (2006), p. 157). Thus,

$$\text{Var(level 1)} = \text{var}(e_{i0}) = \frac{\pi^2}{3} = 3.2865$$