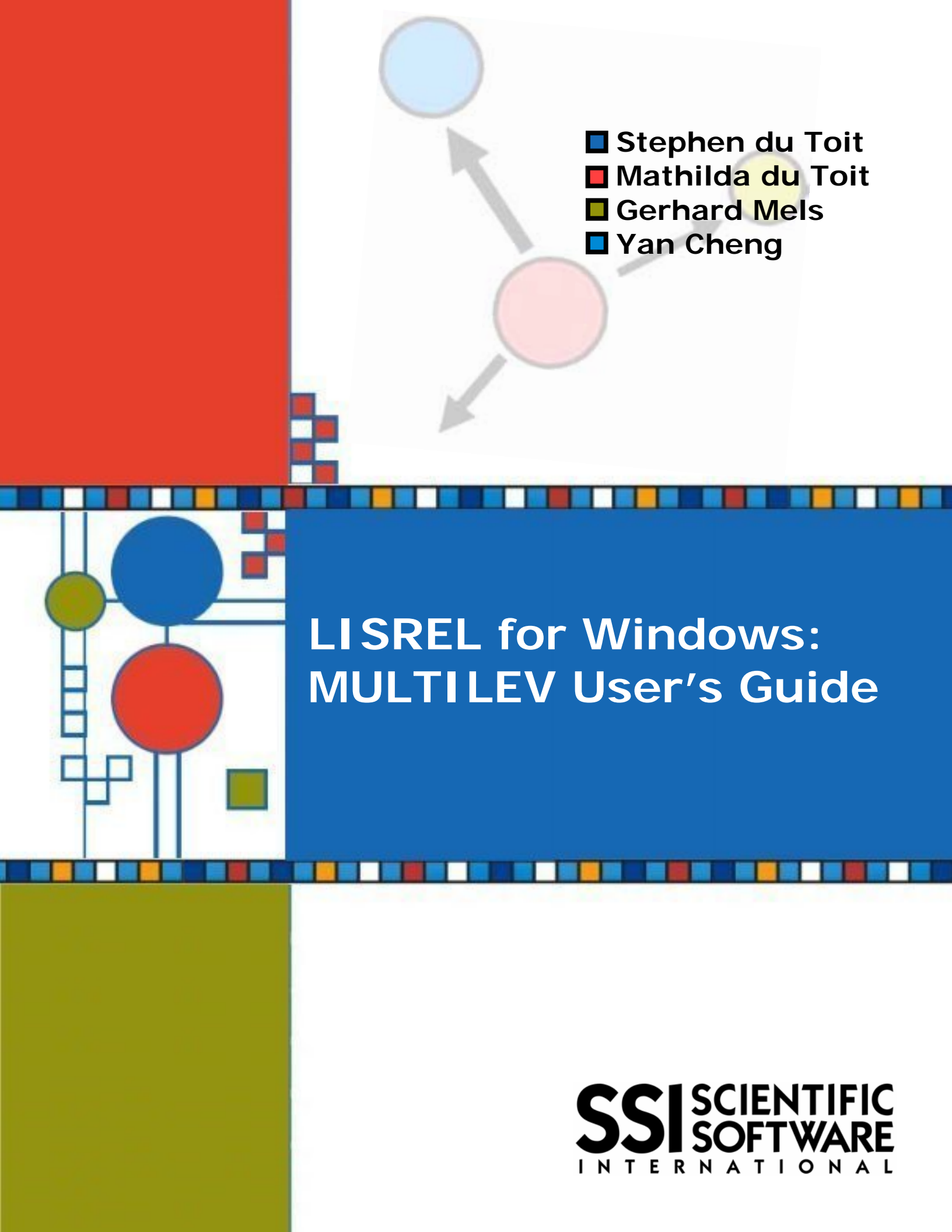


- 
- Stephen du Toit
 - Mathilda du Toit
 - Gerhard Mels
 - Yan Cheng



LISREL for Windows: SURVEYGLIM User's Guide



SSI SCIENTIFIC
SOFTWARE
INTERNATIONAL

Table of contents

INTRODUCTION	1
GRAPHICAL USER INTERFACE	2
The SURVEYGLIM menu	2
The Title and Options dialog box	2
The Distributions and Links dialog box	4
The Dependent and Independent Variables dialog box	6
The Survey Design dialog box.....	7
SURVEYGLIM SYNTAX FILES	9
The structure of the syntax file	9
CLUSTER command.....	10
COVARS command	10
DEPVAR command	11
DISPERSION command	11
DISTRIBUTION command	12
FREQ command.....	12
FPC command	13
GLIMOPTIONS command	13
INTERCEPT command.....	17
LINK command.....	17
POPULATIONSIIZES command.....	17
SAMPLINGRATES command	18
SCALE command.....	19
STRATUM command	19
SY command	20
TITLE command	20
WEIGHT command.....	21
EXAMPLES	22
GLIMs for counts	22
The data	22
The models.....	23
Analyzing counts from a complex sampling design	25
Ignoring stratification and clustering in the sample.....	31
Correcting for over-dispersion in an analysis of counts.....	33
GLIMs for continuous responses	36
The data	37
The models.....	38
Analyzing normally distributed outcomes from complex survey designs	40
Analyzing skewed outcome variables from complex survey designs (method 1)	47
Analyzing skewed outcome variables from complex survey designs (method 2)	50
GLIMs for binary responses	53
The data	53
The models.....	54
Analyzing binary outcomes from complex survey designs (method 1).....	55

Analyzing binary outcomes from complex survey designs (method 2).....	61
GLIMs for ordinal responses	63
The data	63
The models.....	64
Analyzing ordinal outcomes from complex survey designs (method 1).....	66
Analyzing ordinal outcomes from complex survey designs (method 2).....	71
GLIMs for nominal responses.....	73
The data	74
The models.....	75
Analyzing nominal outcomes from complex survey designs.....	76
STATISTICAL THEORY AND METHODS.....	83
GLIM framework.....	83
The Poisson-log model.....	85
Models for the Bernoulli sampling distribution.....	87
The logit model.....	87
The complementary log-log model.....	88
The probit model.....	88
The log model.....	89
Models for the Multinomial sampling distribution	90
The generalized logistic Model.....	90
The cumulative logit model	91
The proportional hazards model	92
The cumulative probit model.....	93
The log model.....	94
The probit model.....	95
The complementary log-log model.....	96
Models for the Binomial sampling distribution	98
The logit model.....	98
The complementary log-log model.....	98
The probit model.....	98
Models for the Gamma distribution	99
The log model.....	99
The power model	99
Models for the Inverse Gaussian distribution.....	100
The log model.....	100
The power model	100
The Negative Binomial-log model	101
The Normal-identity model	101
The estimation of scale and dispersion parameters.....	102

REFERENCES..... 104

Introduction

Many popular statistical methods are based on mathematical models that assume data follow a normal distribution, most obvious among these are the analysis of variance for planned experiments and multiple linear regressions 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 data from complex survey designs. Consequently, inappropriate results are obtained if these methods are applied to complex samples. For quite some time, these methods were extended to include the use of frequency and probability weights in an effort to deal with complex samples. Although this approach yields the appropriate estimates for complex samples, the corresponding standard error estimates are not appropriate. Using a result of Fuller (1975), Binder (1983) proposed methods to obtain the appropriate standard error estimates of the parameters of linear and nonlinear models as well as those of general estimating functions in the case of complex survey designs. These methods are implemented in, amongst others, SAS PROC SURVEYLOGISTIC (SAS Institute 2004) and AM (American Institutes for Research & Cohen 2004).

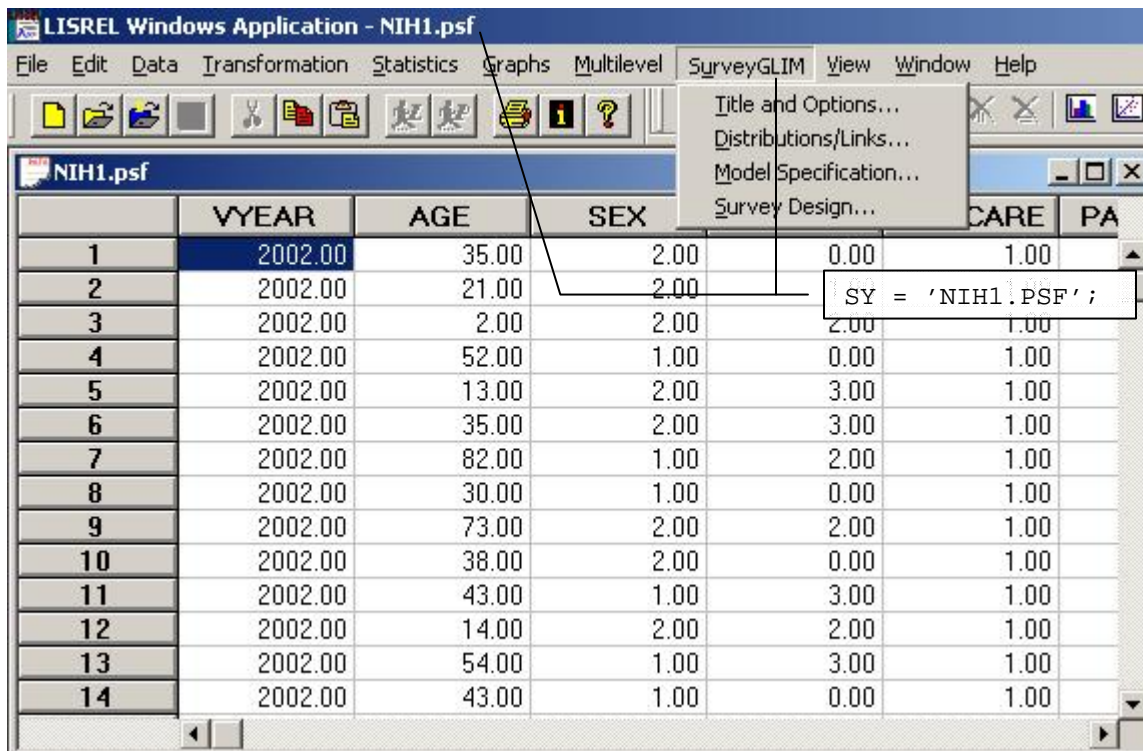
LISREL for Windows (Jöreskog & Sörbom 2005) includes the application SURVEYGLIM, which implements the methods in Agresti (2002) and Binder (1983) to fit GLIMs to complex survey data and simple random sample data. Unlike other statistical software applications for generalized linear modeling for complex survey data such as SAS PROC SURVEYLOGISTIC and AM, SURVEYGLIM allows for a wide variety of sampling distributions and link functions.

This document is an online user's guide for SURVEYGLIM. Section 2 reviews the options and dialog boxes of the SURVEYGLIM menu on the PRELIS System File (PSF) window of LISREL for Windows. SURVEYGLIM syntax files are reviewed in Section 3. Illustrative examples are provided in Section 4. In Section 5, we outline the GLIM statistical theory and methods for complex survey data and simple random sample data.

Graphical User Interface

The SURVEYGLIM menu

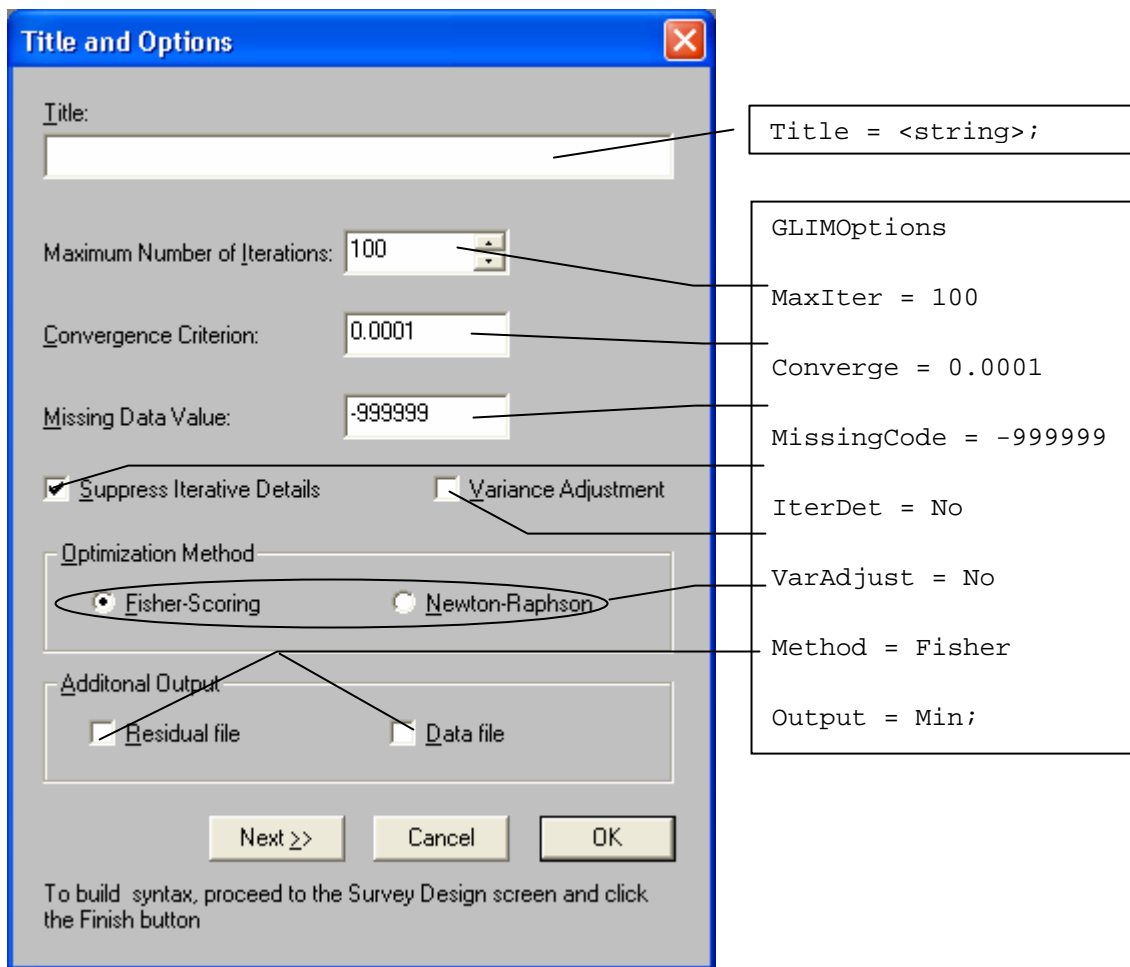
The SURVEYGLIM menu provides you access to a sequence of four dialog boxes that can be used to create a SURVEYGLIM syntax file interactively. It is located on the PSF window of LISREL for Windows which is used to display, manipulate and process raw data. In other words, you must create a PSF and open it in a PSF window before syntax can be generated interactively. To illustrate this, the PSF window for the file **NIH1.psf** in the **SGLIMEX** subfolder with the SURVEYGLIM menu expanded is shown below.



The typical next step would be to click on the **Title and Options** option to load the **Title and Options** dialog box. However, you can directly click on the **Distributions/Links**, **Model Specification** or **Survey Design** option to go to the **Distributions and Links**, the **Dependent and Independent Variables** or the **Survey Design** dialog box respectively.

The Title and Options dialog box

The **Title and Options** dialog box allows you to specify a title and the options of the GLIM analysis interactively and is accessed by selecting the **Title and Options** option on the **SURVEYGLIM** menu. This selection loads the following **Title and Options** dialog box.



Note that the **Title and Options** dialog box corresponds with the Title and GLIMOptions commands as indicated on the image above.

If desired, you can enter a descriptive title in the **Title** string field. If the raw data include missing values with a global missing value other than -999999, you need to enter the global missing value in the **Missing Data Value** number field.

Since the GLIM estimation equations do not have a closed form solution, SURVEYGLIM uses an iterative algorithm to estimate the parameters of the GLIM. In this regard, the Fisher scoring algorithm and the Newton-Raphson algorithm are available. The default algorithm is Fisher scoring; click the **Newton-Raphson** radio button to choose that algorithm instead. You can then enter the maximum number of iterations in the **Maximum Number of Iterations** field if the default of 100 is not appropriate. Enter the appropriate convergence criterion in the **Convergence Criterion** number field if the default value of 0.0001 is not to be used, and check the **Suppress Iterative Details** check box if details of the iterative algorithm should be written to the output file.

In practice, it is possible that the estimated asymptotic covariance matrix of the estimators is not positive definite, in which case the standard error estimates are unreliable. For these situations,

Morel (1989) proposed an adjustment to the estimated asymptotic covariance matrix. To request this option, you need to check the **Variance Adjustment** check box.

You can choose to export the exact raw data that SURVEYGLIM analyzed to a comma separated variable (CSV) file by checking the **Data file** check box. This file will have the same name as the PSF, except that .psf is replaced with **_RAW.CSV**. Similarly, the residuals can be exported to a CSV file by checking the **Residual file** check box. This CSV file will have the same name as the PSF, except that .psf is replaced with **_RES.CSV**.

Once you are done with the **Title and Options** dialog box, click on the **Next** button to go to the **Distributions and Links** dialog box.

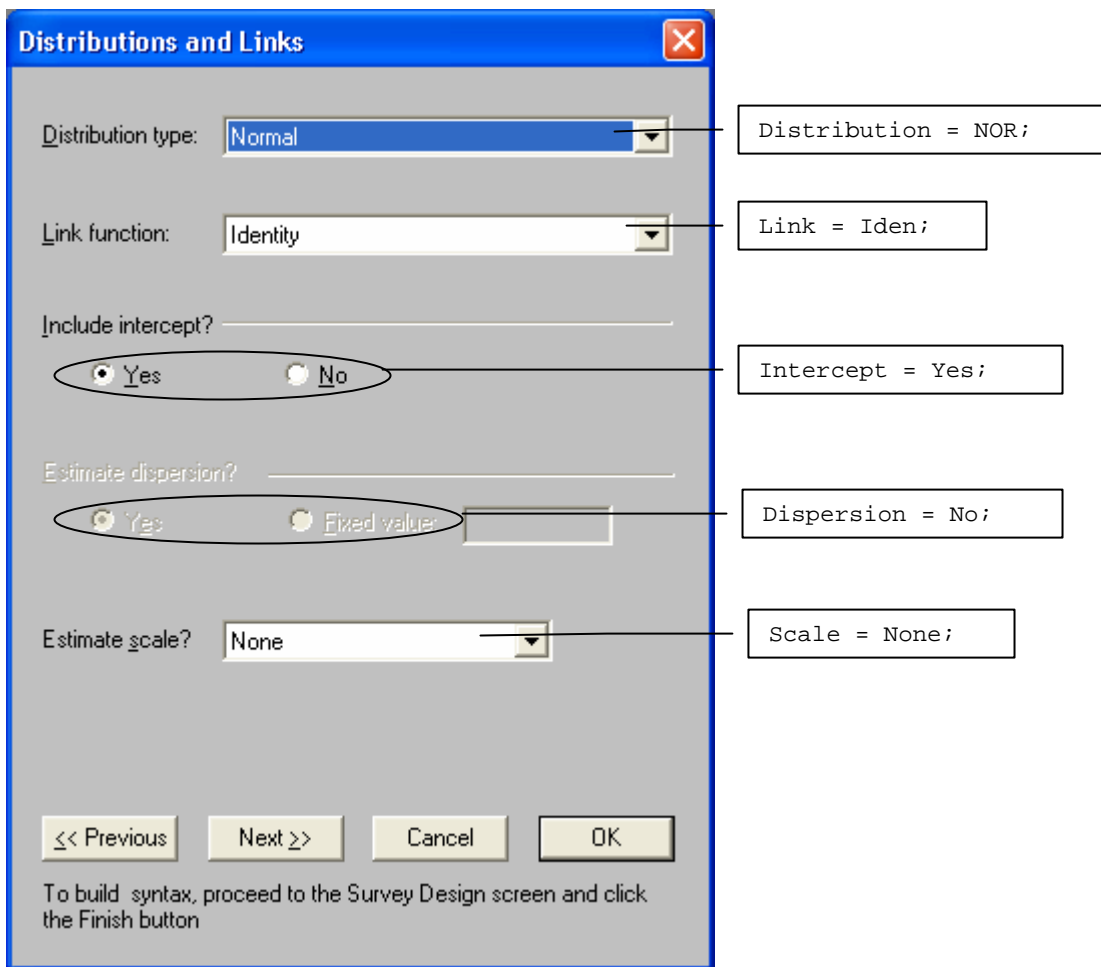
The Distributions and Links dialog box

The **Distributions and Links** dialog box allows you to specify the sampling distribution and the link function of the GLIM interactively. A summary of the combinations of sampling distributions and link functions that are available in SURVEYGLIM is listed in Table 1.

Table 1: Sampling Distribution and Link Functions

Link Distribution	CLL	Identity	Log	Logit	OCLL	OLogit	OProbit	Probit	Power
Bernoulli	x		x	x				x	
Binomial	x			x				x	
Gamma			x						x
Inverse Gaussian			x						x
Multinomial	x		x	x	x	x	x	x	
Negative binomial			x						
Normal		x							
Poisson			x						

The **Distributions and Links** dialog box is usually accessed by clicking on the **Next** button of the **Title and Options** dialog box. It can also be accessed by selecting the **Distributions / Links** option from the SURVEYGLIM menu. These actions load the following dialog box.



Note that the **Distributions and Links** dialog box corresponds with the Distribution, Link, Intercept, Dispersion and Scale commands as indicated on the image above.

Use the **Distribution type** and the **Link function** drop-down list boxes to select the distribution and link function for your GLIM. If an intercept for the mean model of the GLIM is not required, you should activate the **No** radio button.

Some GLIMs include dispersion or scale parameters. These GLIMs are listed in Table 2. If a scale parameter is desired, you can select the appropriate scale parameter from the **Estimate scale?** drop-down list box. In the case of a dispersion parameter, you can fix its value by activating the **Fixed value** radio button. Otherwise, it is estimated by means of maximum likelihood estimation.

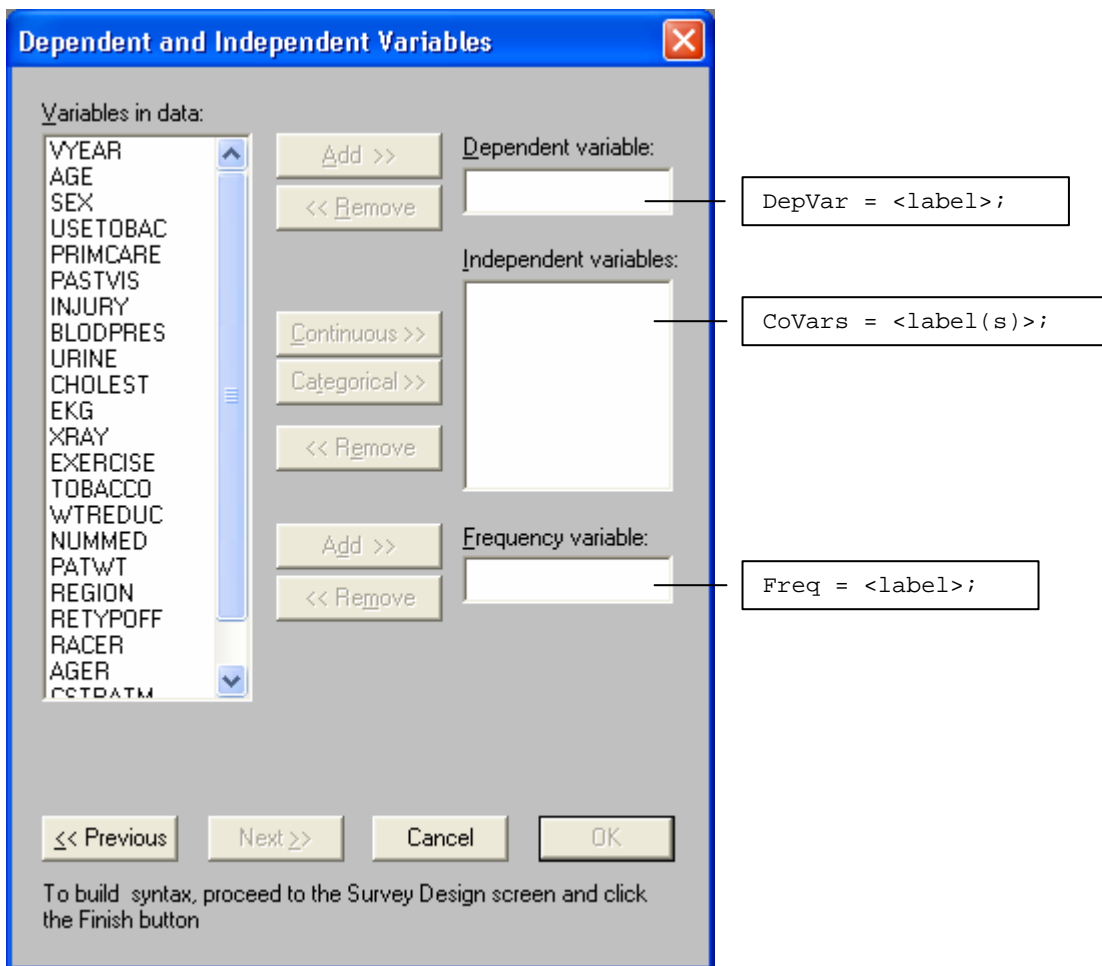
Once the **Distributions and Links** dialog box has been completed, the **Next** button is clicked to go to the **Dependent and Independent Variables** dialog box.

Table 2: Scale and Dispersion Parameters

Parameters Distribution	Scale	Dispersion	Maximum Likelihood	Pearson	Deviance
Binomial	x			x	x
Gamma	x	x	x	x	x
Inverse Gaussian	x	x	x	x	x
Negative binomial		x	x		
Normal	x	x	x	x	x
Poisson	x			x	x

The Dependent and Independent Variables dialog box

The **Dependent and Independent Variables** dialog box allows you to specify the model for the means of the outcome variable and, if applicable, a frequency variable.



Access to this dialog box is obtained by clicking on the **Next** button of the **Distributions and Links** dialog box or by selecting the Model Specification option from the SURVEYGLIM menu. An example of the **Dependent and Independent Variables** dialog box is shown above.

Note that the **Dependent and Independent Variables** dialog box corresponds with the DEPVAR, COVARS and FREQ commands as shown on the image above.

The model for the means of the outcome variable is a function of a set of covariates. You specify the outcome variable by first selecting it from the **Variables in data** list box and then by clicking on the **Add** button of the **Dependent variable** section. The covariates of the model can either be categorical or continuous variables. Dummy variables are also regarded as continuous variables. Categorical covariates are specified by first selecting the covariates from the **Variables in data** list box and then by clicking on the **Categorical** button. In a similar fashion, the **Continuous** button is used to specify the continuous covariates and dummy variables of the model.

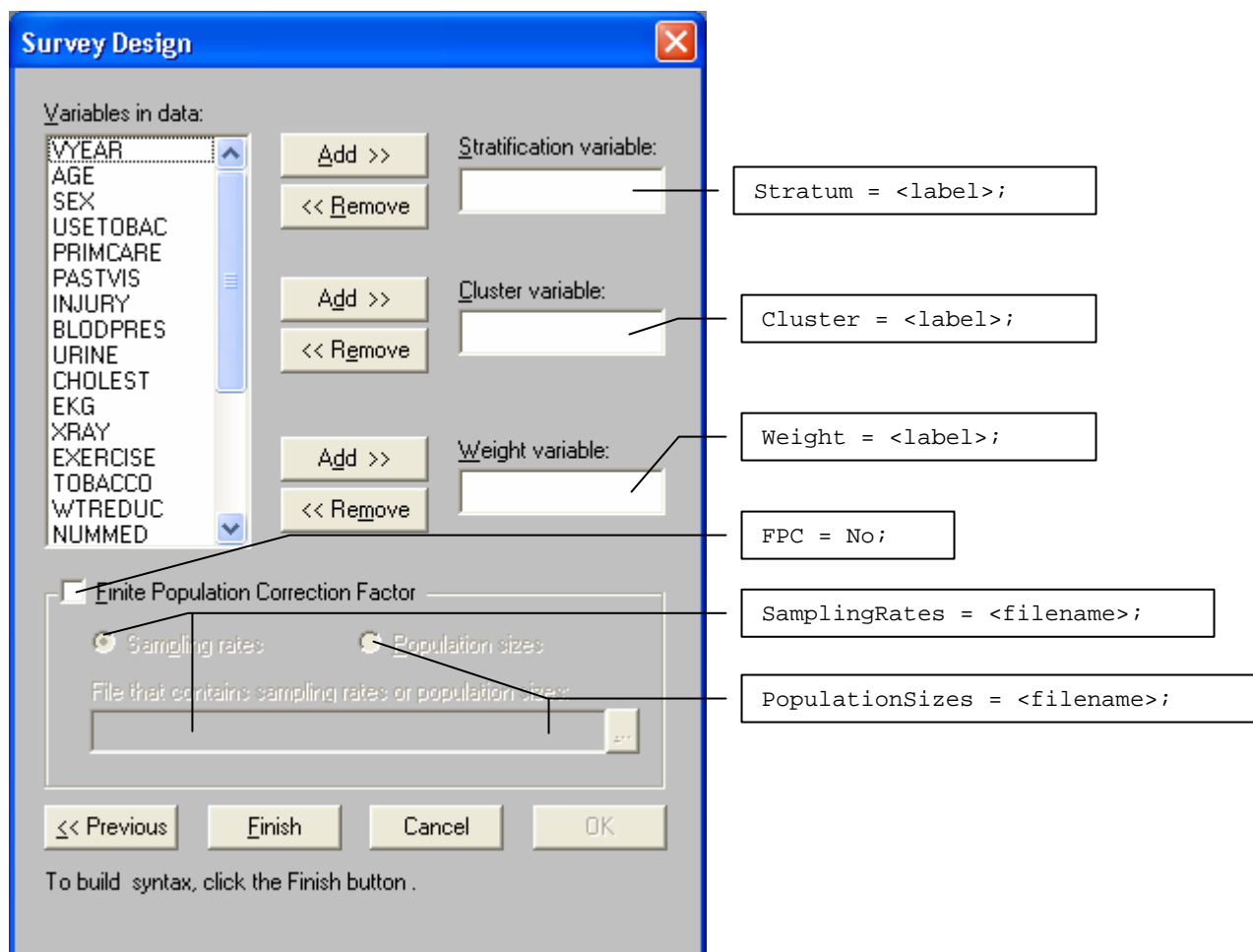
SURVEYGLIM can process raw data and frequency table data. Frequency table data are often used in the case of categorical variables, in which case the same observation often occurs more than once in the raw data. To process frequency table data, the data file must include a column that contains the observed frequencies. Specify this frequency variable by first selecting it from the **Variables in data** list box and then by clicking on the **Add** button of the **Frequency variable** section.

Once the variables have been selected, click the **Next** button to load the **Survey Design** dialog box.

The Survey Design dialog box

You can specify the design of the survey used for data collection and, if applicable, finite population information interactively by using the **Survey Design** dialog box. It is accessed by clicking on the **Next** button of the **Dependent and Independent Variables** dialog box or by selecting the Survey Design option on the SURVEYGLIM menu. An example of the **Survey Design** dialog box follows.

Note that the **Survey Design** dialog box corresponds with the STRATUM, CLUSTER, WEIGHT, FPC, SAMPLINGRATES and POPULATIONIZES commands as shown on the image below.



Complex survey designs typically stratify the target population into strata (subpopulations). These strata usually contain the primary sampling units (clusters). The ultimate sampling units are then selected from the selected clusters and design weights for the ultimate sampling units are constructed. The strata are specified by first selecting the appropriate variable from the **Variables in data** list box and then by clicking on the **Add** button of the **Stratification variable** section. Similarly, the clusters and the design weights are specified by using the **Add** buttons of the **Cluster variable** and the **Weight variable** sections respectively.

In the case of finite target populations, Fuller (1975) proposed a correction factor for the standard error estimates of the parameters. This correction is based on the sampling rates of the strata that can be computed from the actual sizes of the strata. You can prepare a text file containing either the sampling rates or the strata sizes. This file is incorporated by using the browse button of the **File that contains sampling rates or population sizes** section. If this file should contain population sizes rather than sampling rates, you need to activate the **Population sizes** radio button.

The syntax file, which was created interactively by using the four SURVEYGLIM dialog boxes, is opened in a text editor window by clicking on the **Finish** button.

SURVEYGLIM syntax files

The structure of the syntax file

The syntax file, which is generated by the interface of SURVEYGLIM, can also be prepared by using the LISREL for Windows text editor or any other text editor such as Notepad and WordPad. The structure of the syntax file follows.

```
GLIMOPTIONS <options>;
SY = '<filename>';
DEPVAR = <label>;
COVARS = <label(s)>;
DISTRIBUTION = <name>;
LINK = <function>;
INTERCEPT = <option>;
DISPERSION = <option>;
SCALE = <type>;
FREQ = <label>;
STRATUM = <label>;
CLUSTER = <label>;
WEIGHT = <label>;
FPC = <option>;
SAMPLINGRATES = <filename>;
POPULATIONSIIZES = <filename>;
TITLE = <string>;
```

where <label> denotes a case sensitive variable name used in the raw data file, <filename> denotes a complete name (including the drive and folder names) of a file, <option> is either Yes or No, <type> is one of None, Pearson, Deviance or ML (see the SCALE command section), <name> is one of BER, BIN, GAM, INVG MUL, NBIN, NOR or POI (see the DISTRIBUTION command section) and <function> is one of CLL, IDEN, LOG, LOGIT, OLOGIT, OCLL, OPROBIT, PROBIT or PWR[<n>] (see the LINK command section) where <n> denotes an integer. <options> denotes a list of options for the analysis, each of which has the following syntax:

```
<keyword> = <selection>
```

where <keyword> is one of CONVERGE, ITERDETAILS, MAXITER, METHOD, MISSINGCODE, OUTPUT or VARADJUST and <selection> denotes a number, an option or a name (see the GLIMOPTIONS command section). In many applications, optional commands and keywords can be left out if there are program default values available.

The GLIMOPTIONS, SY, DEPVAR and COVARS commands are **required** commands while the other thirteen commands are all **optional**. The GLIMOPTIONS and SY commands should be the first two

commands respectively, but the other commands can be entered in any order. Except for variable labels, the contents of the syntax file are not case-sensitive. Blank lines can be inserted in any section of the syntax file.

In the following sections, the seventeen SURVEYGLIM commands are discussed separately in alphabetical order.

CLUSTER command

The CLUSTER command is used to specify the variable for the primary sampling units of the complex survey. It is an **optional** command. For example, in the case of a simple random sample, the CLUSTER command is omitted.

Syntax

```
CLUSTER = <label>;
```

where <label> denotes the label of the cluster variable. Keep in mind that variable names are case sensitive.

Example

Suppose that the primary sampling units of the complex survey are types of facility and that the variable FACTYPE is used to indicate the facility type for each observation. Then, the corresponding CLUSTER command is

```
CLUSTER = FACTYPE;
```

COVARS command

The purpose of the COVARS command is to specify the covariates of the model for the means of the outcome variable and it is a **required** command.

Syntax

```
COVARS = <label(s)>;
```

where <label(s)> denotes the case sensitive label(s) of the covariates of the model. In the case of a categorical variable, the label should be augmented with a \$ symbol. Dummy variables are

regarded as continuous variables. Consequently, dummy variable labels are not augmented with a \$ symbol.

Example

Suppose that the covariates of the model consist of a dummy variable, *sex*, a categorical variable, *edu*, and a continuous variable, *age*. For this example, the corresponding COVARs command is given by

```
COVARs = sex edu$ age;
```

DEPVAR command

The DEPVAR command is used to specify the outcome variable of the model and it is a **required** command.

Syntax

```
DEPVAR = <label>;
```

where <label> denotes the label of the outcome variable of the model. Note that variable names are case sensitive.

Example

Suppose that the variable, *depr*, is the outcome variable to be used. In this case, the corresponding DEPVAR command would be

```
DEPVAR = depr;
```

DISPERSION command

The Negative Binomial sampling distribution, for example, has a dispersion parameter. This parameter is specified by using the DISPERSION command. Since not all sampling distributions involve a dispersion parameter, the command is **optional** with default of no dispersion to be estimated.

Syntax

DISPERSION = <option>;

where <option> is either Yes or No.

Default

DISPERSION = No;

DISTRIBUTION command

Each GLIM involves the sampling distribution of the outcome variable. The sampling distribution is specified by means of the DISTRIBUTION command, which is **optional**.

Syntax

DISTRIBUTION = <name>;

where <name> is one of BER (Bernoulli), BIN (Binomial), GAM (Gamma), INVG (Inverse Gaussian), MUL (Multinomial), NBIN (Negative Binomial), NOR (Normal) or POI (Poisson).

Default

DISTRIBUTION = NOR;

FREQ command

SURVEYGLIM can process frequency table data if a variable with the frequency is a column of the data file. This frequency variable is specified by means of the FREQ command. Since SURVEYGLIM can also analyze raw data, the FREQ command is **optional**.

Syntax

FREQ = <label>;

where <label> denotes the case sensitive label of the frequency variable.

Example

Suppose that the variable, Count, is the frequency variable. For this example, the FREQ command is given by

```
FREQ = Count;
```

FPC command

Fuller (1975) proposed a finite population correction factor for the standard error estimates of parameters if the complex survey was obtained from a finite population. The FPC command is used to request this correction.

Syntax

```
FPC = <option>;
```

where <option> is either Yes or No.

Default

```
FPC = No;
```

GLIMOPTIONS command

The purpose of the GLIMOPTIONS command is to select the iterative algorithm to be used and to specify options for the selected iterative algorithm. In addition, it is used to specify a global missing data value and the output to be generated. Finally, it allows you to specify the variance adjustment proposed by Morel (1989) if the estimated asymptotic covariance matrix of the parameter estimators is not positive definite. The GLIMOPTIONS command must always be the first command and is a **required** command.

Syntax

```
GLIMOPTIONS <options>;
```

where <options> is a list of options each of which has the following syntax:

```
<keyword> = <selection>
```

where <keyword> is one or more of CONVERGE, ITERDETAILS, MAXITER, METHOD, MISSINGCODE, OUTPUT or VARADJUST and <selection> refers to a name, a number or an option.

CONVERGE keyword

The tolerance limit of the convergence criterion of the selected iterative algorithm is specified by using the CONVERGE keyword which is an **optional** keyword.

Syntax

```
CONVERGE = <number>;
```

where <number> denotes a real number greater than zero.

Default

```
CONVERGE = 0.0001;
```

ITERDETAILS keyword

The purpose of the ITERDETAILS keyword is to suppress or request the printing of the details of the selected iterative algorithm and it is an **optional** keyword.

Syntax

```
ITERDETAILS = <option>;
```

where <option> is either Yes or No.

Default

```
ITERDETAILS = No;
```

MAXITER keyword

You can control the maximum number of iterations of the selected iterative algorithm by means of the MAXITER keyword which is an **optional** keyword.

Syntax

```
MAXITER = <number>;
```

where <number> denotes a positive integer.

Default

MAXITER = 100;

METHOD keyword

SURVEYGLIM implements the Fisher scoring and Newton-Raphson iterative algorithms to obtain the estimates and standard error estimates of the GLIM parameters.

Syntax

METHOD = <method>;

where <method> is either Fisher or Newton.

Default

METHOD = Fisher;

MISSINGCODE keyword

Raw data often include missing values. SURVEYGLIM uses list-wise deletion for handling data with missing values if you specify a global missing value by means of the MISSINGCODE option, which is **optional**.

Syntax

MISSINGCODE = <number>;

where <number> denotes a real number.

Default

MISSINGCODE = -999999;

OUTPUT keyword

SURVEYGLIM can write the raw data or residuals of the GLIM analysis to separate CSV files. The OUTPUT keyword is used to request neither, one or both of these files and is an **optional** keyword.

Syntax

OUTPUT = <amount>;

where <amount> is one of Min for the standard GLIM results, Res for adding residuals as a CSV file to the standard results, Raw for adding the data used by SURVEYGLIM as a CSV file to the standard results or All for the complete SURVEYGLIM results. The standard SURVEYGLIM results consist of the design, data and model description, the goodness of fit statistics, the estimated regression weights and standard error estimates and the estimated asymptotic covariance and correlation matrices of the parameter estimators.

Default

OUTPUT = Min;

VARADJUST keyword

Morel (1989) proposed an adjustment for the estimated asymptotic covariance matrix of the parameter estimators if it should not be positive definite. You can request this adjustment by using the VARADJUST keyword, which is **optional**.

Syntax

VARADJUST = <option>;

where <option> is either Yes or No.

Default

VARADJUST = No;

Example

Suppose that the Newton-Raphson algorithm with a maximum of 50 iterations and a convergence criterion tolerance limit of 0.0001 with printed details is required. Suppose further that the Morel (1989) variance adjustment and the complete SURVEYGLIM output are required and the global missing value for the raw data is -9. For this example, the GLIMOPTIONS command is given by

```
GLIMOPTIONS CONVERGE = 0.0001 MAXITER = 50 MISSINGCODE = -9 ITERDETAILS = Yes  
VARADJUST = Yes METHOD = Newton OUTPUT = All;
```

INTERCEPT command

Many GLIMs can either include or exclude an intercept parameter for the model for the means of the outcome variable. The purpose of the INTERCEPT command is to allow you to either include or exclude an intercept parameter and it is an **optional** command.

Syntax

```
INTERCEPT = <option>;
```

where <option> is either Yes or No.

Default

```
INTERCEPT = Yes;
```

LINK command

The link function of a GLIM describes the relationships between the means of the outcome variable and the means of the corresponding linear model. The LINK command is used to specify the link of the GLIM. It is an **optional** command.

Syntax

```
LINK = <name>;
```

where <name> is one of CLL (complementary log-log), IDEN (identity), LOG (log), LOGIT (logit), OCLL (proportional hazards), OLOGIT (cumulative logit), OPROBIT (cumulative probit), PROBIT (probit) or PWR[<n>] (power) where <n> denotes an integer.

Default

```
LINK = Iden;
```

POPULATIONSIIZES command

If the finite population correction for the standard error estimates proposed by Fuller (1975) is required, you must prepare a text file containing either the sampling rates or strata sizes. The purpose of the POPULATIONSIIZES command is to specify the file that contains the strata sizes.

Syntax

```
POPULATIONSIIZES = <filename>;
```

where <filename> denotes the complete name (including drive and folder names) of the text file containing the strata sizes. The drive and folder names may be omitted if the text file and the syntax file are in the same folder.

Example

Suppose that the text file **POPULATIONSIIZES.TXT** in the **SGLIMEX** subfolder of the C drive contains the strata sizes. In this case, the POPULATIONSIIZES command is given by

```
POPULATIONSIIZES = C:\SGLIMEX\POPULATIONSIIZES.TXT;
```

SAMPLINGRATES command

If the finite population correction for the standard error estimates proposed by Fuller (1975) is required, you must prepare a text file containing either the sampling rates or strata sizes. The purpose of the SAMPLINGRATES command is to specify the file that contains the sampling rates. It is an **optional** command.

Syntax

```
SAMPLINGRATES = <filename>;
```

where <filename> denotes the complete name (including drive and folder names) of the text file that contains the sampling rates. The drive and folder names may be omitted if the text file and the syntax file are in the same folder.

Example

If the sampling rates are contained in the text file **SampRates.txt** in the **SGLIMEX** subfolder of the C drive, the corresponding SAMPLINGRATES command is given by

```
SAMPLINGRATES = C:\SGLIMEX\SampRates.txt;
```

SCALE command

Some sampling distributions such as the Poisson, Binomial, Gamma, Inverse Gaussian and Normal distributions have an **optional** scale parameter. This parameter is specified by using the SCALE command. Since not all sampling distributions involve a scale parameter, the command is **optional**.

Syntax

```
SCALE = <type>;
```

where <type> is one of None, Pearson, Deviance or ML.

Default

```
SCALE = None;
```

STRATUM command

Complex surveys are typically obtained by stratifying the target population into subpopulations (strata). The STRATUM command allows you to specify the stratification variable. Since other types of surveys are available, the STRATUM command is an **optional** command.

Syntax

```
STRATUM = <label>;
```

where <label> denotes the case sensitive label of the stratification variable.

Example

Suppose that the target population was stratified into census regions and that the variable CENREG is the variable used to indicate the census region for each observation. In this case, the STRATUM command is given by

```
STRATUM = CENREG;
```

SY command

SURVEYGLIM can process raw data or frequency data that are available in the form of a PSF. The PSF to be processed is specified by means of the SY command. The SY command is a **required** command and must be the **second** command listed in the syntax file.

Syntax

```
SY = '<filename>';
```

where <filename> denotes the complete name (including drive and folder names) of the PSF. The drive and folder names may be omitted if the PSF and syntax file are in the same folder. Note the use of single quotes in this command.

Example

Suppose that the data to be processed are listed in the file **NIH1.psf** which is located in the **SGLIMEX** subfolder on the C drive. In this case, the SY command is given by

```
SY = 'C:\SGLIMEX\NIH1.PSF';
```

TITLE command

It is often convenient to label a specific analysis to distinguish it from other analyses. This can be accomplished by using the TITLE command which is an **optional** command.

Syntax

```
TITLE = <string> ;
```

where <string> denotes a descriptive title for the analysis.

Example

Consider an analysis in which a Bernoulli-Probit model was fitted to substance abuse data. In this case, one possible TITLE command is given by

```
TITLE = SGLIMEX1: Bernoulli Probit Model for Substance Abuse Data;
```

WEIGHT command

Design weights are constructed for the ultimate sampling units of complex surveys. The purpose of the WEIGHT command is to allow you to specify the design weight variable. Since surveys without design weights are permitted, the WEIGHT command is an **optional** command.

Syntax

```
WEIGHT = <label>;
```

where <label> denotes the case sensitive label of the design weight variable.

Example

Suppose that the variable A2TWA0 is used to capture the design weight for each observation. For this example, the WEIGHT command is given by

```
WEIGHT = A2TWA0;
```

Examples

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 distribution for the response variable. In this section, we illustrate the use of SURVEYGLIM by using some practical examples based on health-related count data. More specifically, a Poisson-log and a Negative Binomial-log model are fitted to substance abuse data. A description of the data follows.

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 and within each stratum a sample was obtained for each of three facility treatment types within each of the four census regions. The specific data set is provided in the **SGLIMEX**

subfolder of LISREL for Windows as the PSF **cntdiag.psf**. The first portion of this file is shown in the following PSF window.

	cntdiag	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.00	0.00	0.00	4.00	4.00	190.70
2	0.00	0.00	0.00	4.00	4.00	44.30
3	0.00	0.00	0.00	4.00	4.00	44.30
4	0.00	1.00	0.00	4.00	4.00	44.30
5	0.00	0.00	0.00	4.00	4.00	44.30
6	0.00	0.00	0.00	4.00	4.00	44.30
7	0.00	0.00	0.00	4.00	4.00	44.30
8	0.00	0.00	0.00	4.00	4.00	44.30
9	0.00	0.00	0.00	4.00	4.00	44.30
10	0.00	0.00	1.00	4.00	2.00	371.90
11	0.00	0.00	1.00	4.00	2.00	371.90
12	0.00	0.00	1.00	4.00	2.00	371.90
13	0.00	0.00	1.00	4.00	2.00	371.90
14	0.00	0.00	1.00	4.00	2.00	371.90
15	0.00	0.00	0.00	4.00	2.00	371.90

A brief description of the variables to be used in the subsequent GLIM analyses follows.

CENREG is the census region of the client (1 for Northeast, 2 for Midwest, 3 for South and 4 for West).

- FACTYPE is the facility treatment type of the client (1 for residential treatment, 2 for outpatient methadone treatment, 3 for outpatient non-methadone treatment and 4 for more than one type of treatment).
- A2TWA0 is the design weight of the client.
- cntdiag is the number of abuse diagnoses of the client (0, 1, 2 or 3).
- sex is the value of a dummy variable for the gender (0 for male and 1 for female) of the client.
- race_d is the value of a dummy variable for the race (0 for nonwhite and 1 for white) of the client.

More information on the ADSS and the data are available at <http://www.icpsr.umich.edu>.

The models

The sampling distributions

The sampling distribution of the Poisson-log GLIM is the Poisson distribution whose probability density function is given by

$$f(y_k, \mu_k) = \frac{e^{-\mu_k} y_k^{\mu_k}}{y_k!}$$

where y_k denotes the response variable y for respondent k and μ_k denotes the mean of y_k . The Poisson sampling distribution has the unique feature that its variance is equal to its mean. A common empirical finding in fitting a Poisson variable is that the actual variance is somewhat larger or smaller than the mean value. The data are said to have over-dispersion or under-dispersion compared to the original model. When this occurs, the variance can be freed up so that it is not exactly equal to the mean. This is handled by adding a scale parameter for the variance. When this change is implemented, the model is no longer a Poisson process. But one still can use the algorithm for generalized linear models and obtain good parameter estimates with the modified approach. Another approach for dealing with the over-dispersion problem would be to consider a more appropriate sampling distribution for the data. In this regard, the Negative Binomial distribution can be very useful. The probability density function of the Negative Binomial distribution is given by

$$f(y_k, \mu_k, \psi) = \frac{\Gamma\left(y_k + \frac{1}{\psi}\right)}{\Gamma(y_k + 1)\Gamma\left(\frac{1}{\psi}\right)} \frac{(\psi\mu_k)^{y_k}}{(1 + \psi\mu_k)^{y_k + \frac{1}{\psi}}}$$

where ψ denotes the dispersion parameter. The variance of the Negative Binomial distribution is given by

$$\sigma^2(y_k) = \mu_k + \psi\mu_k^2.$$

The mean model

The mean model for the Poisson-log and Negative Binomial-log GLIMs is given by

$$\mu_k = \exp(\alpha + \beta_1 x_{1k} + \beta_2 x_{2k} + \dots + \beta_r x_{rk})$$

where μ_k denotes the mean value of the response variable for client k , x_{jk} denotes the value of the j -th predictor ($j=1,2,\dots,r$) for client k , and α , β_1 , \dots , β_{r-1} , and β_r denote unknown parameters. In practice, it can occur that the coefficient of some covariate is assumed to be unity. This covariate is commonly known as an offset variable. Offsets are typically used when the response variable is a rate rather than a number or count. For this specific example, the mean model may be expressed as

$$E[\text{cntdiag}_k] = \exp(\alpha + \beta_1 * \text{sex}_k + \beta_2 * \text{race_d}_k)$$

where $E[\text{cntdiag}_k]$ denotes the mean number of diagnoses for client k , sex_k and race_d_k denotes the values of the variables sex and race_d respectively and α , β_1 and β_2 denote unknown parameters. From this model, it follows that the ratio of the mean numbers of diagnoses for female ($\text{sex}_k = 1$) and male ($\text{sex}_k = 0$) clients may be expressed as

$$\frac{\exp(\alpha + \beta_1 + \beta_2 * \text{race}_d)}{\exp(\alpha + \beta_2 * \text{race}_d)} = \exp(\beta_1)$$

Similarly, it follows that $\exp(\beta_2)$ is the ratio of the mean numbers of diagnoses for white and nonwhite clients. The model fitted value is a mean number of diagnoses for client k and is given by

$$\hat{E}[\text{cntdiag}_k] = \exp(\hat{\alpha} + \hat{\beta}_1 * \text{sex}_k + \hat{\beta}_2 * \text{race}_d_k)$$

where $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the maximum likelihood estimates of α , β_1 and β_2 respectively.

Analyzing counts from a complex sampling design

A question that a researcher may want to address is whether ethnicity and gender effects are associated with the number of substance abuse diagnoses. An appropriate statistical model for this type of count variable is a GLIM with a Poisson distribution and a log link function.

Setting up the analysis

The first step is to open the PSF shown above in the LISREL for Windows PSF window. This is accomplished as follows.

Use the **Open** option on the **File** menu of the root window of LISREL for Windows to load the **Open** dialog box.

Select the **Preliis Data (*.psf)** option from the **Files of type** drop-down list box.

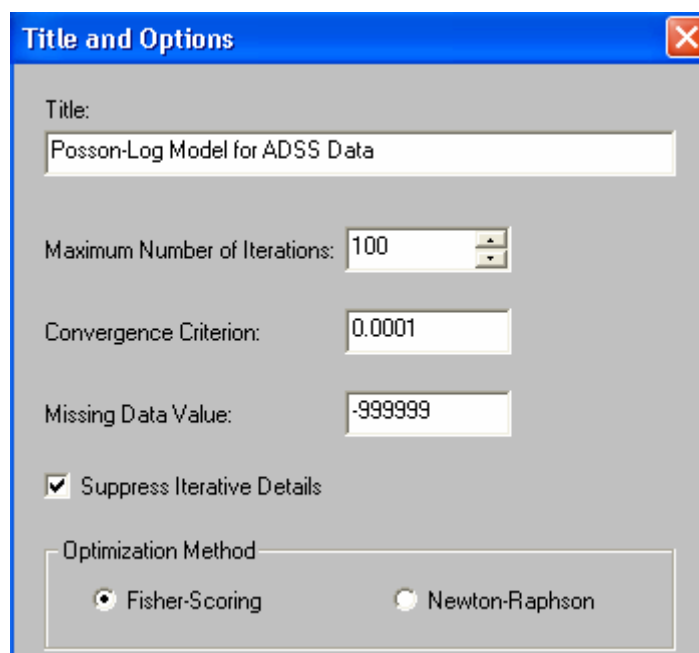
Browse for the file **cntdiag.psf** in the **SGLIMEX** subfolder.

Click on the **Open** button to open the file **cntdiag.psf** in a PSF window.

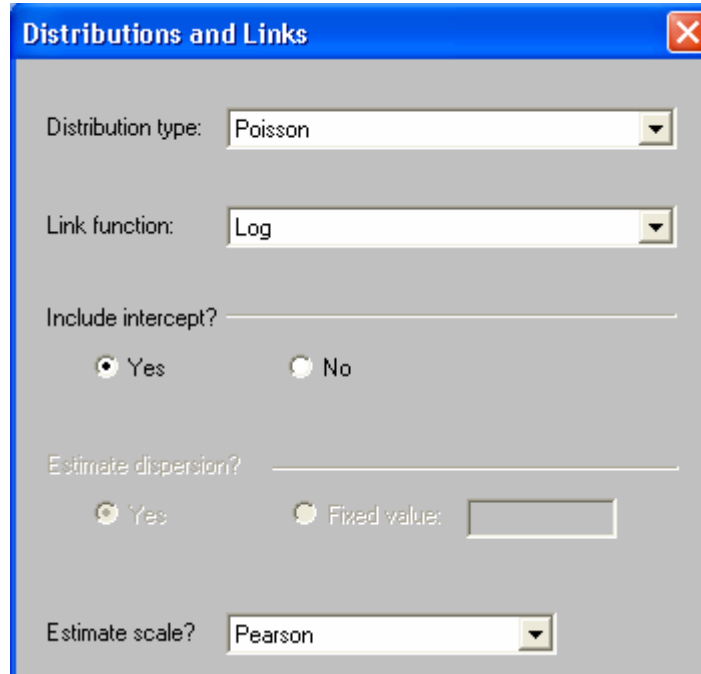
Click on the **SURVEYGLIM** menu to produce the following PSF window.

	cntdiag	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.00	0.00	0.00	4.00	4.00	190.70
2	0.00	0.00	0.00	4.00	4.00	44.30
3	0.00	0.00	0.00	4.00	4.00	44.30
4	0.00	1.00	0.00	4.00	4.00	44.30
5	0.00	0.00	0.00	4.00	4.00	44.30
6	0.00	0.00	0.00	4.00	4.00	44.30
7	0.00	0.00	0.00	4.00	4.00	44.30
8	0.00	0.00	0.00	4.00	4.00	44.30
9	0.00	0.00	0.00	4.00	4.00	44.30
10	0.00	0.00	1.00	4.00	2.00	371.90
11	0.00	0.00	1.00	4.00	2.00	371.90
12	0.00	0.00	1.00	4.00	2.00	371.90
13	0.00	0.00	1.00	4.00	2.00	371.90
14	0.00	0.00	1.00	4.00	2.00	371.90
15	0.00	0.00	0.00	4.00	2.00	371.90

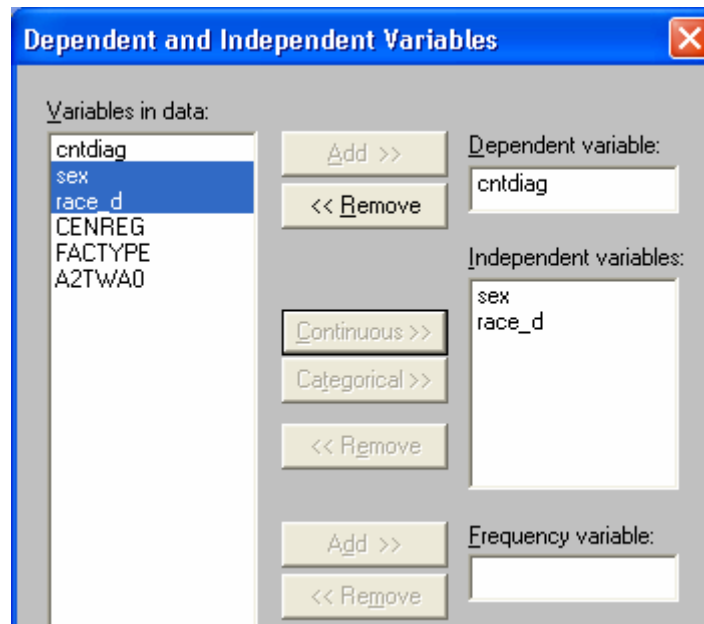
The next step is complete the sequence of four dialog boxes of the SURVEYGLIM GUI described in Section 3.2. The **Title and Options** dialog box is the first dialog box and is accessed by selecting the **Title and Options** option on the **SURVEYGLIM** menu above. In order to identify the analysis, enter the string **Poisson-Log Model for ADSS Data** into the **Title** string field to produce the following **Title and Options** dialog box.



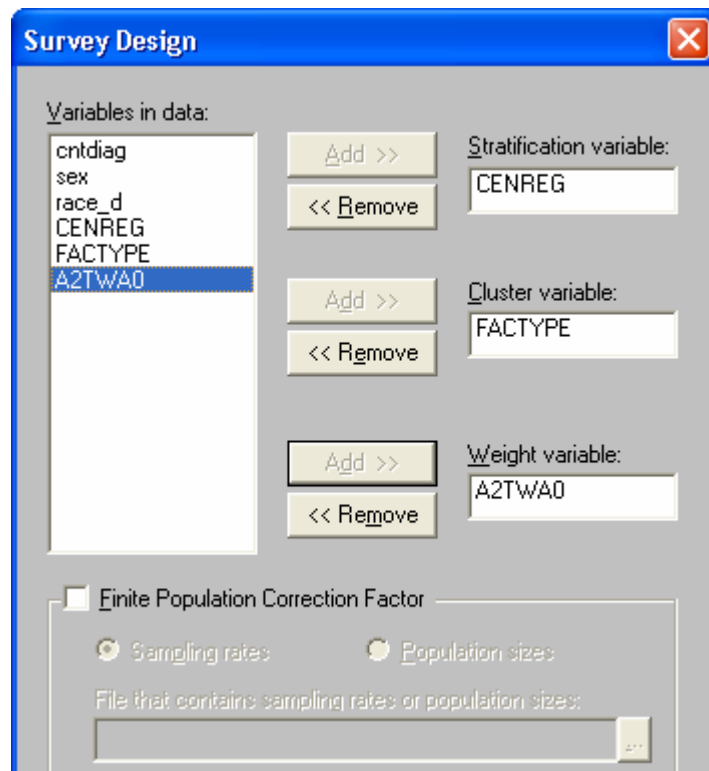
Since the default options will be used for this example, no changes are necessary. Click the **Next** button to access the **Distributions and Links** dialog box. Since we intend to fit a Poisson-log model, select the **Poisson** option from the **Distribution type** drop-down list box. For this example, we will estimate the scale parameter of the model by using the Pearson χ^2 estimate (see Section 5). Select the **Pearson** option from the **Estimate scale?** drop-down list box to produce the following **Distributions and Links** dialog box.



Move on to the **Dependent and Independent Variables** dialog box by clicking on the **Next** button. Specify the response variable `cntdiag` by selecting it from the **Variables in data** list box and clicking on the **Add** button of the **Dependent variable** section. In a similar fashion, add the covariates `sex` and `race_d` to the **Independent variables** list box to produce the following **Dependent and Independent Variables** dialog box.



Since the data are not frequency table data and no offset variable is used for this example, go to the **Survey Design** dialog box by clicking on the **Next** button. The strata are the census regions (CENREG) and are specified by selecting the variable CENREG from the **Variables in data** list box and clicking on the **Add** button of the **Stratification variable** section. Similarly, add the PSU variable FACTYPE and the design weight variable A2TWA0 to the **Cluster variable** and **Weight variable** boxes respectively to produce the following **Survey Design** dialog box.



Since no finite population information is available, we are done. The next step is to click on the **Finish** button to open the following text editor window for **cntdiag.pr2**.

```

cntdiag.PR2
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 IterDetails=No
Method=Fisher;
Title=Poisson-Log Model for ADSS Data;
SY='C:\Program Files\lisrel187\SGLIMEX\cntdiag.PSF';
Distribution=POI;
Link=LOG;
Intercept=Yes;
Scale=Pearson;
DepVar=cntdiag;
CoVars=sex race_d;
Stratum=CENREG;
Cluster=FACTYPE;
Weight=A2TWAO;

```

We are now ready to submit our GLIM analysis. This is achieved by clicking on the **Run Prelis** toolbar icon to produce the text editor window for **cntdiag.out**.

Discussion of results – Poisson-log model

A portion of the results of the Poisson-log GLIM analysis is shown in the following text editor window.

```

cntdiag.OUT

Statistic          Value      Den. DF   Num. DF   P Value
-----          -
Adjusted Wald F    2.8314      2         7         0.125599
Wald Chi-square    6.4718      2         2         0.125599

Note: The Wald F Test and Chi-square Statistics are statistics to test the
      null hypothesis that all the regression weights are equal to zero.

Estimated Regression Weights

Parameter      Estimate      Standard      z Value      P Value
-----      -
intcept        0.3302        0.0557        5.9248       0.0000
sex            0.0619        0.0709        0.8726       0.3829
race_d         0.1167        0.0620        1.8818       0.0599
SCALE          0.7479

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)

```

SURVEYGLIM reports the Adjusted Wald F and χ^2 test statistic values for testing the null hypothesis that all the regression weights are equal to zero which may be expressed as (*cf.* American Institutes for Research & Cohen, 2003)

$$F_w = \frac{\left(\sum_{h=1}^H n_h - H - r + 1 \right)}{\left(\sum_{h=1}^H n_h - H \right) * r} \hat{\beta}' \hat{\Upsilon}^{-1} \hat{\beta}$$

and

$$X_w^2 = \hat{\beta}' \hat{\Upsilon}^{-1} \hat{\beta}$$

respectively where H denotes the number of strata, $\sum_{h=1}^H n_h$ denotes the number of PSUs, r denotes the number of covariates of the model, $\hat{\beta}$ denotes the estimate of the parameter vector, β , of regression weights and $\hat{\Upsilon}$ denotes the estimated asymptotic covariance matrix of the estimators of the elements of β . If the null hypothesis is correct, F_w and X_w^2 approximately follow an F distribution with r and $\sum_{h=1}^H n_h - H - r + 1$ degrees of freedom and a χ^2 distribution with r degrees of freedom respectively.

Both the values of the Wald F and χ^2 test statistics are not statistically significant if a significance level of 5% is used. Hence, there is insufficient evidence to conclude that both gender and race influence the number of diagnoses of a client. This finding is supported by the non-significant z test statistic values for the significance of the individual parameters.

The scale parameter estimate is less than unity which indicates under-dispersion for the response variable. In other words, the sample variance of the variable cntdiag is less than its mean.

Estimated outcomes for different groups

The fitted model follows from the output file above as

$$\hat{E}[\text{cntdiag}_k] = \exp(0.33 + 0.06 * \text{sex}_k + 0.12 * \text{race}_d_k)$$

Although gender and race did not significantly affect the number of diagnoses, the following examples illustrate how the fitted model can be used to calculate the mean of number of diagnoses for various subgroups when there are statistically significant differences among them. This fitted

model implies that the mean number of diagnoses for a white female client ($\text{sex}_k = 1$ and $\text{race}_k = 1$) is given by

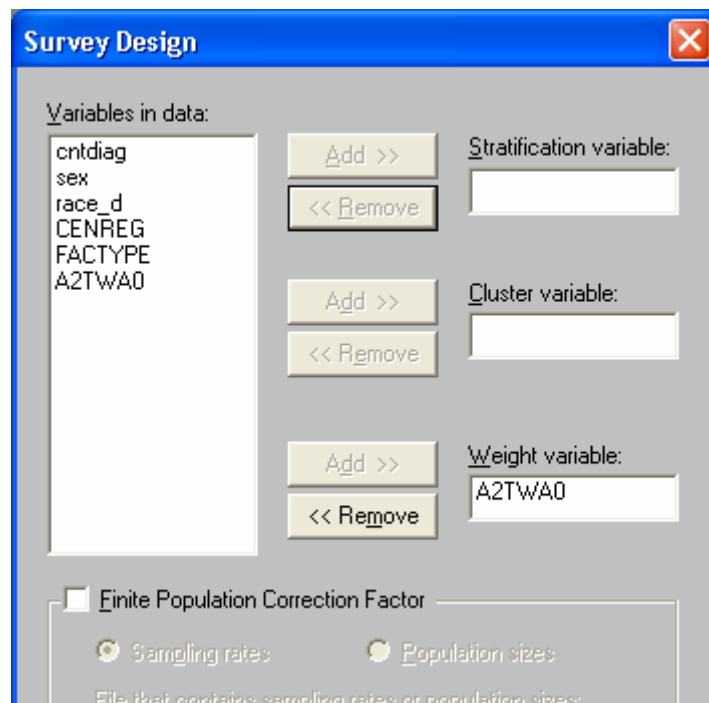
$$\exp(0.33 + 0.06 + 0.12) = \exp(0.51) = 1.67$$

Similarly, the mean number of diagnoses for a nonwhite female client ($\text{sex}_k = 1$ and $\text{race}_k = 0$) is 1.48. It also follows from the output above that $\exp(\hat{\beta}_1) = \exp(0.06) = 1.06$ is the multiplicative effect of gender on the fitted number of diagnoses for a client. This implies that, on the average, female clients have a 6% higher estimated mean number of diagnoses than male clients. Similarly, it follows that $\exp(\hat{\beta}_2) = \exp(0.12) = 1.13$ which implies that, on the average, the fitted number of diagnoses is 13% higher for white clients than for nonwhite clients.

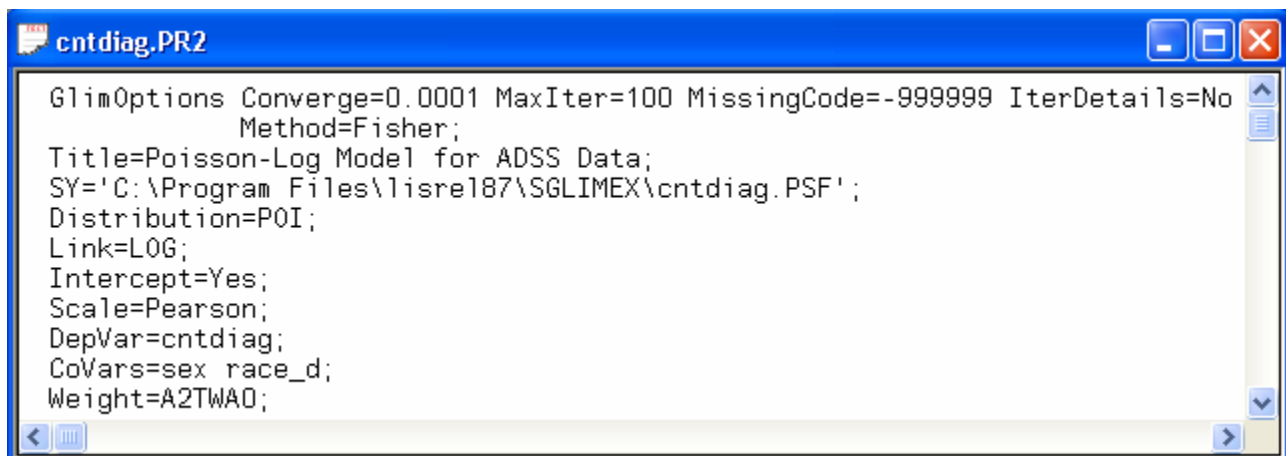
Ignoring stratification and clustering in the sample

Setting up the analysis

The stratification and clustering can be ignored by not specifying the stratification and cluster variables on the **Survey Design** dialog box. However, it is recommended to change the title of the analysis to distinguish it from the previous analysis. This is done by selecting the **Title and Options** option on the SURVEYGLIM menu to go to the **Title and Options** dialog box and then by entering the string **Fitting a Poisson-Log model with design weights only** in the **Title** string field. Since our model remains the same, click on the **Next** buttons of the **Title and Options**, the **Distributions/Links** and the **Dependent and Independent Variables** dialog boxes respectively to go to the **Survey Design** dialog box. Remove the stratification and cluster variables by clicking on the **Remove** buttons of the **Stratification variable** and **Cluster variable** sections to produce the following **Survey Design** dialog box.



As this completes our modifications, click on the **Finish** button to open the following text editor window for **cntdiag.pr2**.



As before, submit the analysis by clicking on the **Run Prelis** toolbar icon to produce the text editor window for **cntdiag.out**.

Discussion of results

A portion of the text editor window for **cntdiag.out** is shown below.

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	0.3302	0.0009	367.2546	0.0000
sex	0.0619	0.0016	39.3322	0.0000
race_d	0.1167	0.0015	75.4532	0.0000
SCALE	0.7479			

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

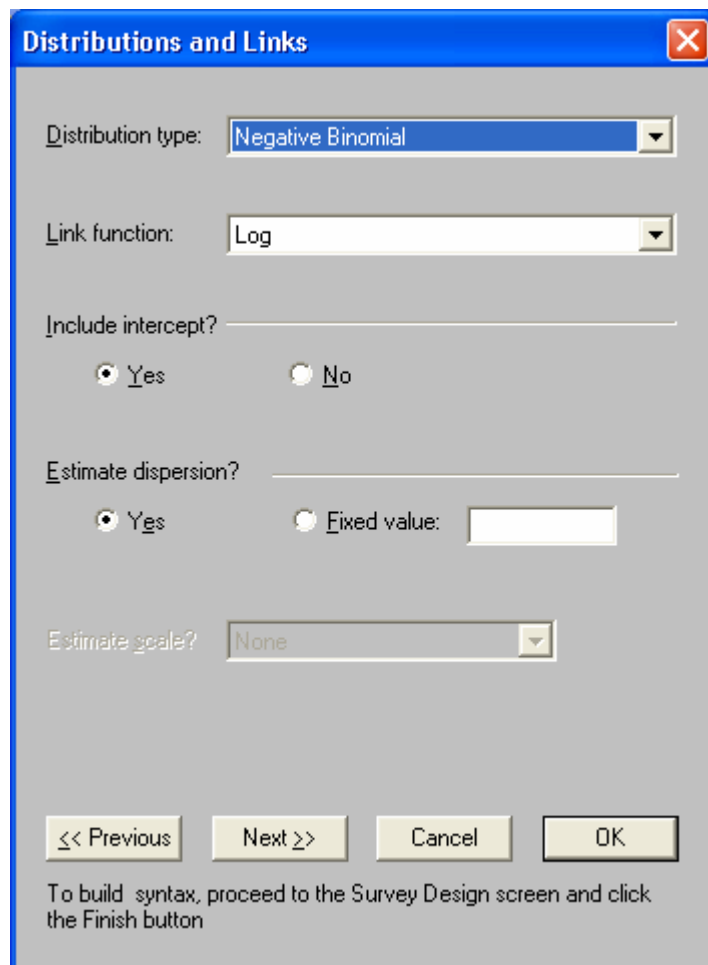
The results above indicate that although the parameter estimates are identical to those obtained when the design of the complex survey was taken into account, the standard error estimates are significantly smaller (*cf.* Brogan, 1998). As a consequence, both gender and race appear to have a statistically significant effect on the number of substance abuse diagnoses at a $p < 0.00001$ level of confidence. This is a reversal of the results obtained when the complex sampling design was taken into account. As this example indicates, inferences based on an analysis that does not correct for the reduced precision of a complex sampling design can be very misleading.

Correcting for over-dispersion in an analysis of counts

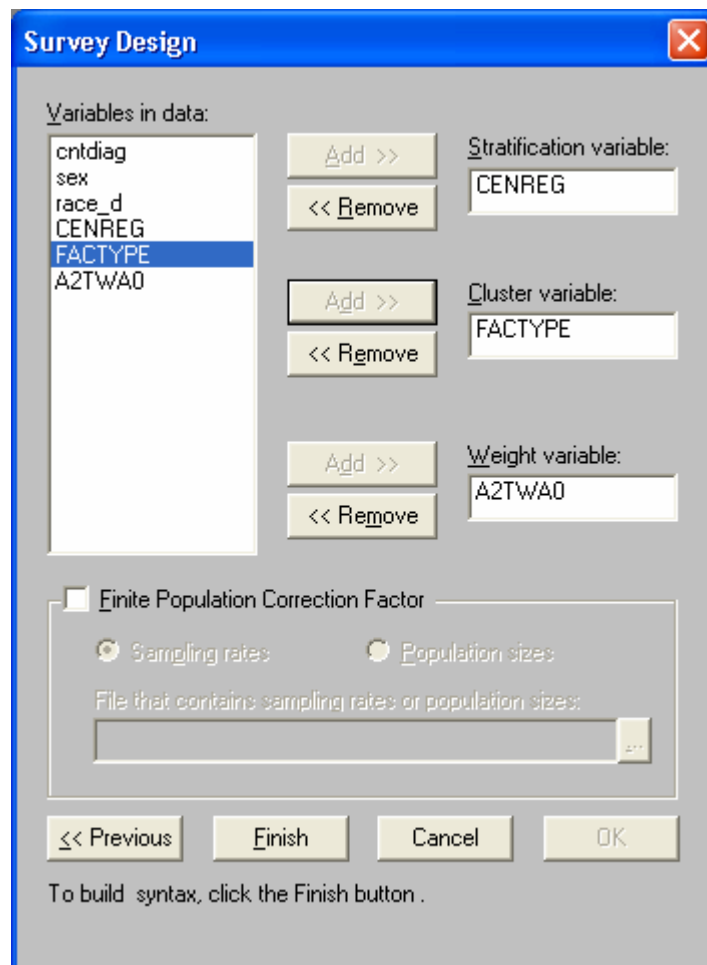
The results for the Poisson-log model indicated the presence of under-dispersion. Although the negative Binomial distribution is intended for dealing with over-dispersion, we will use it here for illustrative purposes.

Setting up the analysis

In order to fit the Negative Binomial-log model interactively to the data in **cntdiag.psf**, we only need to re-specify the sampling distribution. As in the previous analysis, start by modifying the title to **Fitting a Negative Binomial-Log model** by accessing the **Title and Options** dialog box and clicking the **Next** button to go to the **Distributions and Links** dialog box. Select the **Negative Binomial** option from the **Distribution** drop-down list box to produce the following **Distributions and Links** dialog box.



Since the rest of the model remains the same, click on the **Next** buttons of the **Distributions and Links** and the **Dependent and Independent Variables** dialog boxes respectively to go to the **Survey Design** dialog box. Specify the complex survey design again by selecting the variables CENREG and FACTYPE from the **Variables in data** list box and clicking on the **Add** buttons of the **Stratification variable** and **Cluster variable** sections respectively to produce the following **Survey Design** dialog box.



Click on the **Finish** button to open the following text editor window for **cntdiag.pr2**.

```

cntdiag.PR2
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 IterDetails=No
Method=Fisher;
Title=Poisson-Log Model for ADSS Data;
SY='C:\Program Files\lisrel187\SGLIMEX\cntdiag.PSF';
Distribution=NBIN;
Link=LOG;
Intercept=Yes;
Dispersion=Yes;
DepVar=cntdiag;
CoVars=sex race_d;
Stratum=CENREG;
Cluster=FACTYPE;
Weight=A2TWA0;
*

```

Submit the analysis by clicking on the **Run Prelis** toolbar icon to open the text editor window for the corresponding output file **cntdiag.out**.

Discussion of results – negative Binomial model

A portion of the text editor window for **cntdiag.out** is shown below.

Statistic

Statistic	Value	Den. DF	Num. DF	P Value
Adjusted Wald F	2.7650	2	7	0.130319
Wald Chi-square	6.3201	2		0.130319

Note: The Wald F Test and Chi-square Statistics are statistics to test the null hypothesis that all the regression weights are equal to zero.

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	0.3302	0.0319	10.3451	0.0000
sex	0.0619	0.0486	1.2739	0.2027
race_d	0.1167	0.0672	1.7358	0.0826
DISPERSN	0.0000	0.0000		

A comparison of these results with those obtained for the Poisson-log model shows that the estimates are the same, but that the standard error estimates are different. However, the conclusions are the same as those made based on the results for the Poisson-log model.

The zero estimate of the dispersion parameter of the Negative Binomial distribution indicates that over-dispersion seen with the Poisson distribution does not apply to this particular analysis. This finding is in agreement with the Poisson scale estimate less than unity, which indicated the presence of under-dispersion rather than over-dispersion.

GLIMs for continuous responses

In many research studies, the response variable of interest is a continuous variable. Examples of continuous response variables are inpatient expenditure of medical interns, earnings of software engineers, insurance claim costs, failure times of machine parts, total cholesterol scores of heart patients, aggregate loss dollars for life insurance policies, etc. SURVEYGLIM can also fit models with continuous response variables to complex survey or simple random sample data. This feature

is illustrated in this section by fitting a Normal-identity, a Gamma-log and an Inverse Gaussian-log model to health data. A description of the specific data set follows.

The data

The data set forms part of the data library of the Medical Expenditure Panel Survey (MEPS). The MEPS is a longitudinal national survey that is used to yield national estimates of health care expenses. During 1999, background data and data on the health expenditures of a sample of 23,565 participants were obtained. The 1999 sample was stratified into 143 strata (VARSTR99) and into 460 PSUs (VARPSU99). The first portion of the data set to be used is shown in the following PSF window.



	PERWT99F	VARSTR99	VARPSU99	racex	Rsex	Rpovc99	inscov9
1	14137.86	131.00	2.00	5.00	-1.00	3.00	2.00
2	17050.99	131.00	2.00	5.00	1.00	3.00	1.00
3	35737.55	131.00	2.00	5.00	-1.00	3.00	1.00
4	35862.67	131.00	2.00	5.00	-1.00	3.00	1.00
5	19407.11	131.00	2.00	5.00	1.00	3.00	1.00
6	18499.83	131.00	2.00	5.00	-1.00	3.00	1.00
7	18499.83	131.00	2.00	5.00	-1.00	3.00	1.00
8	22394.53	136.00	1.00	5.00	-1.00	3.00	1.00
9	27008.96	136.00	1.00	5.00	1.00	3.00	1.00
10	25108.71	136.00	1.00	5.00	-1.00	3.00	1.00
11	17569.81	136.00	1.00	5.00	-1.00	3.00	1.00
12	21478.06	136.00	1.00	5.00	-1.00	3.00	1.00
13	21415.68	136.00	1.00	5.00	1.00	3.00	1.00
14	12254.66	125.00	1.00	5.00	-1.00	5.00	2.00
15	17699.75	125.00	1.00	5.00	-1.00	5.00	1.00

The following variables are used in the subsequent analyses.

- VARSTR99 is the variance estimation stratum of the respondent.
- FACTYPE is the variance estimation PSU of the respondent.
- PERWT99F is the final design weight of the respondent.
- TOTEXP99 is the natural logarithm of the total health care expenditure of the respondent during 1999.
- racex is the value of a nominal variable for the race (1 for American Indian, 2 for Aleut or Eskimo, 3 for Asian or Pacific Islander, 4 for black and 5 for white) of the respondent.
- inscov9 is the value of a nominal variable for the type of insurance coverage (1 for private, 2 for public and 3 for uninsured) of the respondent during 1999.

More information on the MEPS and the data are available at

<http://www.meps.ahrq.gov/Puf/PufDetail.asp?ID=93>.

The models

The sampling distributions

The probability density function of the Normal sampling distribution is given by

$$f(y_k, \mu_k, \psi) = \frac{1}{\sqrt{2\pi\psi}} \exp\left(-\frac{1}{2\psi}(y_k - \mu_k)^2\right)$$

where y_k denotes the response variable y for respondent k , μ_k denotes the mean of y_k and ψ denotes the dispersion parameter. The Normal distribution is symmetric about its mean. Two examples of non-symmetric distributions are the Gamma and the Inverse Gaussian distributions. These distributions are used as statistical models for continuous variables that only take positive values. In contrast to the normal distribution, which has the same basic shape irrespective of the mean and variance, the Gamma and Inverse Gaussian can take many different shapes depending on the mean and scale parameters. Both distributions are used in situations where the variable being studied is roughly continuous, but may be strongly skewed. The corresponding probability density functions are given by

$$f(y_k, \mu_k, \psi) = \frac{1}{\Gamma\left(\frac{1}{\psi}\right) y_k} \left(\frac{y_k}{\mu_k \psi}\right)^{\frac{1}{\psi}} \exp\left(-\frac{y_k}{\mu_k \psi}\right)$$

and

$$f(y_k, \mu_k, \psi) = \frac{1}{\sqrt{2\pi y_k^3 \psi}} \exp\left(-\frac{1}{2 y_k \psi} \left(\frac{y_k - \mu_k}{\mu_k}\right)^2\right)$$

respectively.

The mean models

The mean model for the Normal-identity GLIM is given by

$$\mu_k = \alpha + \beta_1 x_{1k} + \beta_2 x_{2k} + \cdots + \beta_r x_{rk}$$

while the mean model for the Gamma-log and Inverse Gaussian-log GLIMs is given by

$$\mu_k = \exp(\alpha + \beta_1 x_{1k} + \beta_2 x_{2k} + \cdots + \beta_r x_{rk})$$

where μ_k denotes the mean value of the response variable for respondent k , x_{jk} denotes the value of the j -th predictor ($j=1,2,\dots,r$) for respondent k , and $\alpha, \beta_1, \dots, \beta_{r-1}$ and β_r denote unknown parameters. The two specific mean models are given by

$$E[\text{TOTEXP}_k] = \alpha + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k}$$

and

$$E[\text{TOTEXP}_k] = \exp(\alpha + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k})$$

where $E[\text{TOTEXP}_k]$ denotes the mean of the natural logarithm of the total medical expenditures during 1999 recorded for respondent k ; where x_{1k} (1 for Aleut or Eskimo and 0 otherwise), x_{2k} (1 for American Indian and 0 otherwise), x_{3k} (1 for Asian or Pacific Islander and 0 otherwise), x_{4k} (1 for Black and 0 otherwise) denote dummy variables for the race of respondent k . Note that $x_{1k} = x_{2k} = x_{3k} = x_{4k} = -1$ for White respondents, who serve as the reference category. Also, x_{5k} (1 for any private insurance and 0 otherwise), and x_{6k} (1 for any public insurance only and 0 otherwise) denote dummy variables for the insurance coverage category of respondent k . Here $x_{5k} = x_{6k} = -1$ represent respondents with no insurance coverage. Finally, $\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$, and β_6 denote unknown parameters. In the case of the Gamma-log and Inverse Gaussian-log GLIMs, the ratio of means of the natural logarithm of the total medical expenditures of Aleut or Eskimos may be expressed as

$$\frac{\exp(\alpha + \beta_1 + \beta_5 x_5 + \beta_6 x_6)}{\exp(\alpha + \beta_5 x_5 + \beta_6 x_6)} = \exp(\beta_1).$$

Similarly, $\exp(\beta_2)$, $\exp(\beta_3)$, $\exp(\beta_4)$ and $\exp(-\beta_1 - \beta_2 - \beta_3 - \beta_4)$ denote the ratios of the means natural logarithm of the total medical expenditures of American Indians, Asians or Pacific Islanders, Blacks and Whites and other races respectively. In addition, $\exp(\beta_5)$, $\exp(\beta_6)$ and $\exp(-\beta_5 - \beta_6)$ are ratios of the means natural logarithm of the total medical expenditures of respondents with any private insurance, public insurance only and no insurance respectively.

The estimated mean logarithmic total medical expenditures for respondent k follows as

$$\hat{E}[\text{TOTEXP}_k] = \hat{\alpha} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \hat{\beta}_3 x_{3k} + \hat{\beta}_4 x_{4k} + \hat{\beta}_5 x_{5k} + \hat{\beta}_6 x_{6k}$$

for the Normal-identity GLIM and as

$$\hat{E}[\text{TOTEXP}_k] = \exp\left(\hat{\alpha} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \hat{\beta}_3 x_{3k} + \hat{\beta}_4 x_{4k} + \hat{\beta}_5 x_{5k} + \hat{\beta}_6 x_{6k}\right)$$

for the Gamma-log and Inverse Gaussian-log GLIMs respectively where $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_6$ denote the maximum likelihood estimates of α , β_1 , β_2 , ..., β_6 respectively.

Analyzing normally distributed outcomes from complex survey designs

In this example, we are interested in exploring the linear relationship between a respondent's total health related expenditure and his/her ethnicity and gender. To make the assumption of normality more plausible, we use the natural logarithm of the total health care expenditure of the respondent during 1999 (TOTEXP99) as outcome. A normal distribution with identity link function defines the GLIM model used in this case.

Setting up the analysis

As in Section 3.4.1, the first step is to open the file **meps.psf** in a PSF window. This is done as follows.

Use the **Open** option on the **File** menu of the root window of LISREL for Windows to load the **Open** dialog box.

Select the **Prelis Data (*.psf)** option from the **Files of type** drop-down list box.

Browse for the file **meps.psf** in the **TUTORIAL** subfolder.

Click on the **Open** button to open the file **meps.psf** in a PSF window.

Click on the **SURVEYGLIM** menu to produce the following PSF window.

The screenshot shows the LISREL Windows Application interface. The title bar reads "LISREL Windows Application - [meps.PSF]". The menu bar includes "File", "Edit", "Data", "Transformation", "Statistics", "Graphs", "Multilevel", "SurveyGLIM", "View", "Window", and "Help". The "SurveyGLIM" menu is open, showing options: "Title and Options...", "Distributions/Links...", "Model Specification...", and "Survey Design...". Below the menu is a data table with 11 rows and 8 columns. The columns are labeled SEX, RACEX, POVCAT99, EXP99, and PERWT99F. The first row is highlighted in blue.

	SEX	RACEX	POVCAT99			EXP99	PERWT99F
1	1.00	5.00	4.00	2.00		7.91	14137.86
2	2.00	5.00	4.00	1.00		8.81	17050.99
3	1.00	5.00	4.00	1.00		4.09	35737.55
4	1.00	5.00	4.00	1.00		4.09	35862.67
5	2.00	5.00	4.00	1.00		6.67	19407.11
6	1.00	5.00	4.00	1.00		5.84	18499.83
7	1.00	5.00	4.00	1.00		6.52	18499.83
8	1.00	5.00	4.00	1.00		8.08	22394.53
9	2.00	5.00	4.00	1.00		7.96	27008.96
10	1.00	5.00	4.00	1.00		4.72	25108.71
11	1.00	5.00	4.00	1.00		8.06	17569.81

We are now ready to use the **SURVEYGLIM** menu to fit the Normal-identity GLIM to the data in **meps.psf**. Select the **Title and Options** option on the **SURVEYGLIM** menu. Enter the descriptive title **A Normal-Identity Model for MEPS Data** into the **Title** string field to produce the following **Title and Options** dialog box.

Title and Options

Title:
A Normal-Identity Model for MEPS Data

Maximum Number of Iterations: 100

Convergence Criterion: 0.0001

Missing Data Value: -999999

Suppress Iterative Details Variance Adjustment

Optimization Method

Fisher-Scoring Newton-Raphson

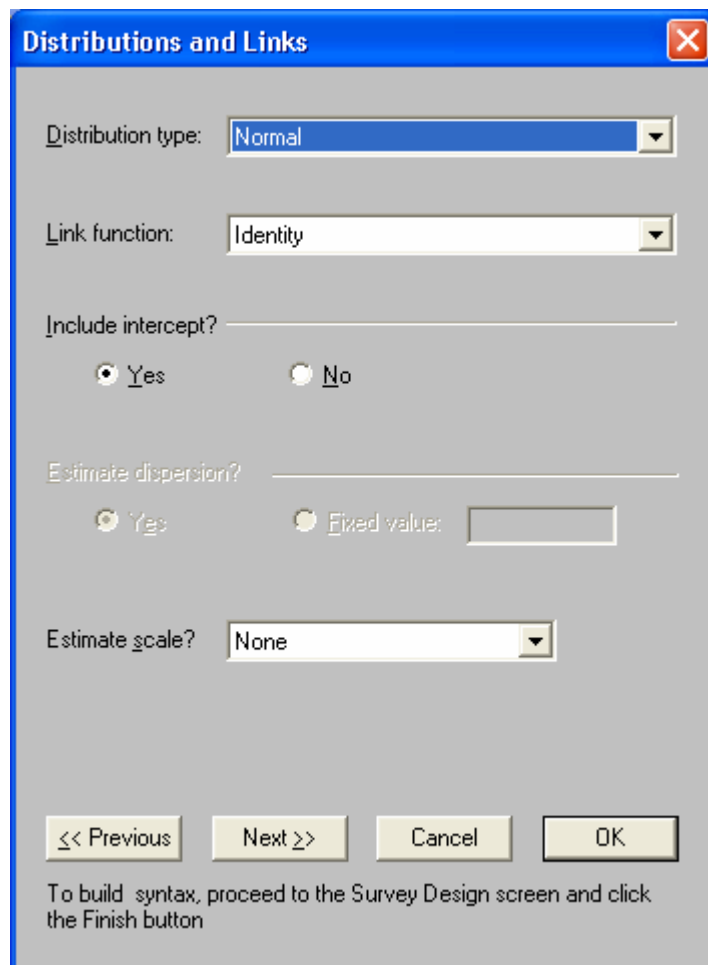
Additional Output

Residual file Data file

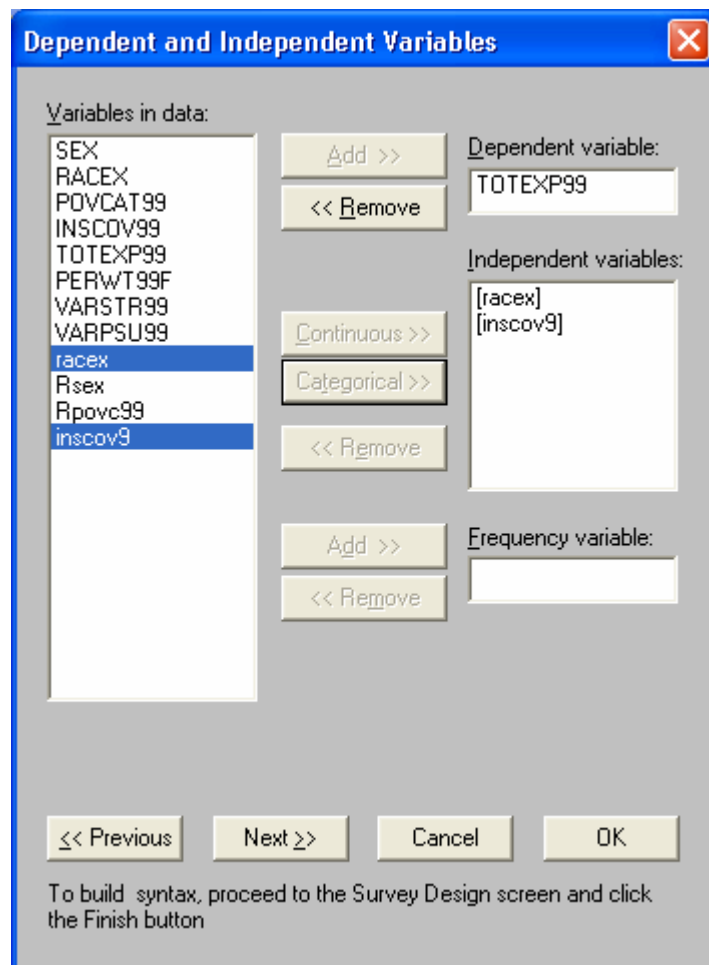
Next >> Cancel OK

To build syntax, proceed to the Survey Design screen and click the Finish button

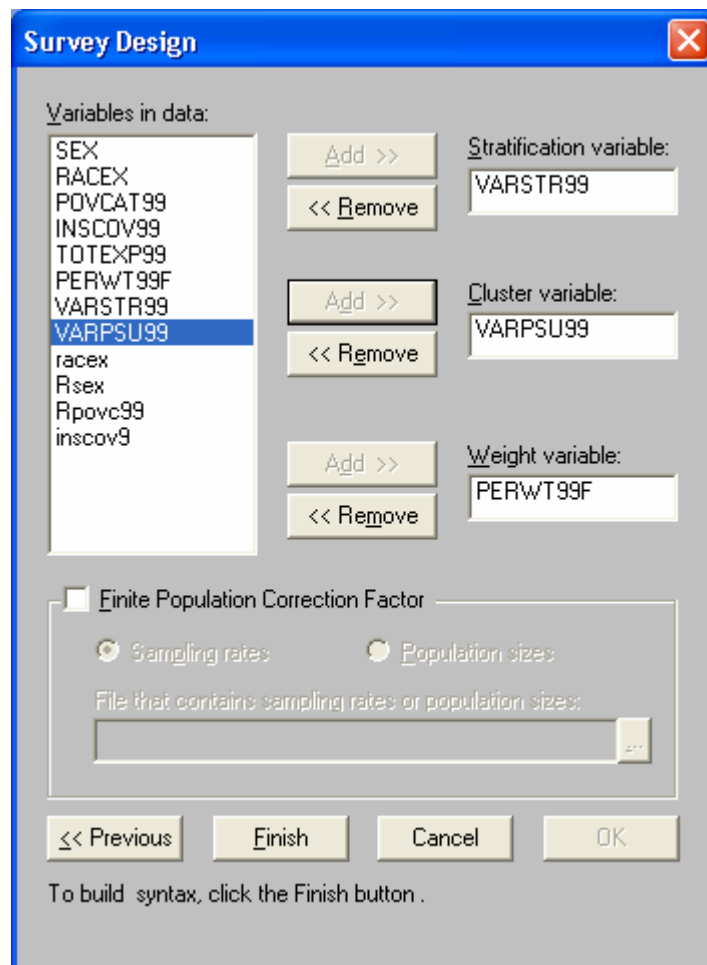
Since the default options will be used for this illustration, click on the Next button to go to the Distributions and Links dialog box.



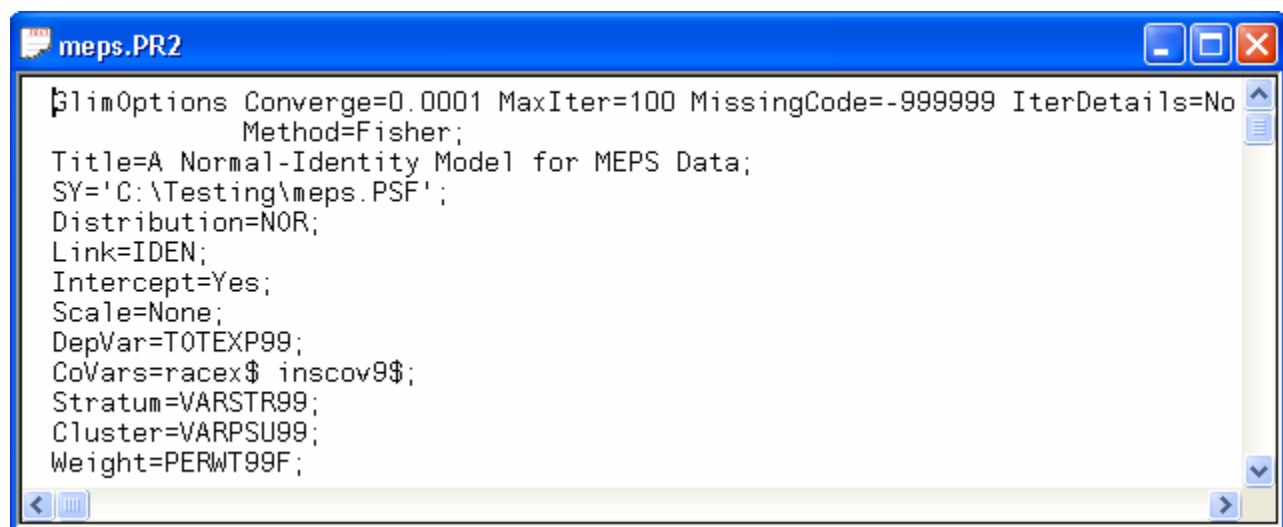
The default values are correct, so click on the **Next** button to go to the **Dependent and Independent Variables** dialog box. Specify the response variable, TOTEXP99, by selecting it from the **Variables in data** list box and then by clicking on the **Add** button of the **Dependent variable** section. Specify the two categorical covariates, racex and inscov9, by selecting them from the **Variables in data** list box and then by clicking on the **Categorical** button of the **Independent variables** section to produce the following **Dependent and Independent Variables** dialog box.



Click on the **Next** button to load the **Survey Design** dialog box. Specify the stratum variable, VARSTR99, by selecting it from the **Variables in data** list box and then by clicking on the **Add** button of the **Stratification variable** section. Similarly, use the **Add** buttons of the **Cluster variable** and the **Weight variable** sections to specify the cluster variable, VARPSU99, and the weight variable, PERWT99F, respectively to produce the following **Survey Design** dialog box.



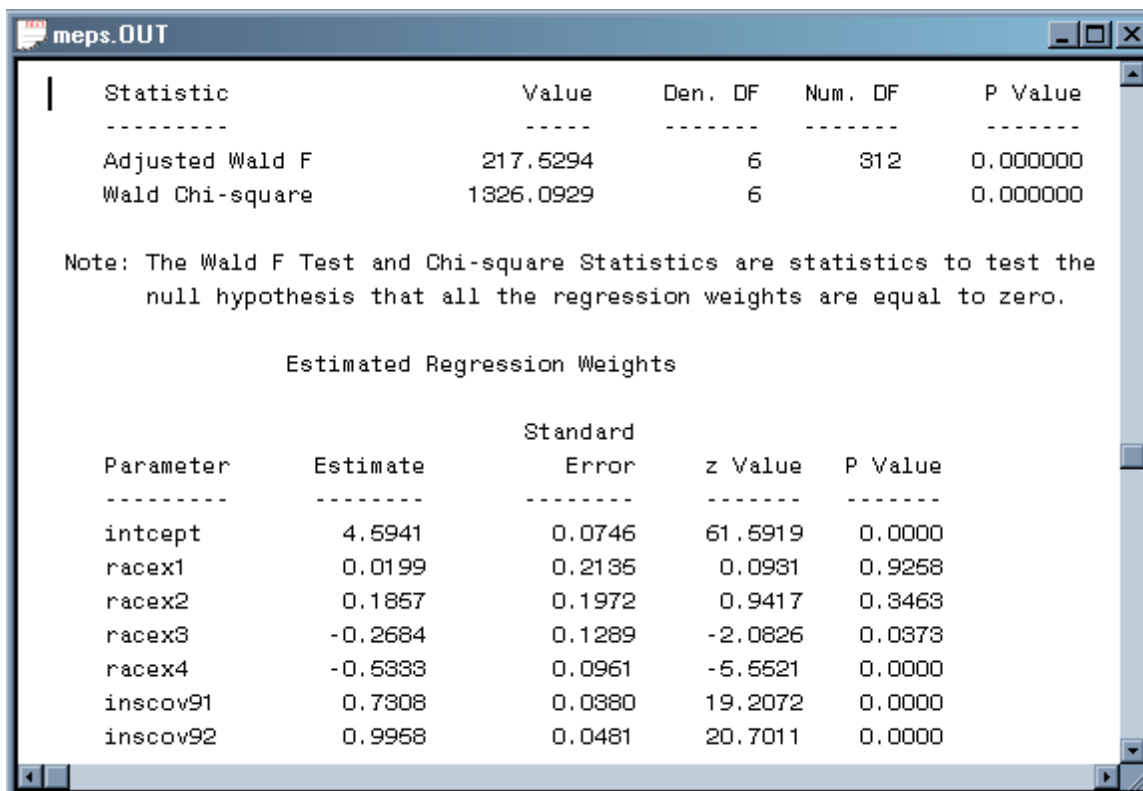
Since this completes the specification of our intended GLIM analysis, click on the **Finish** button to open the following text editor window for **meps.pr2**.



Click on the **Run Prelis** toolbar icon to submit the syntax file above and to obtain the output file **meps.out**.

Discussion of results – Normal-identity model

A portion of the output file **meps.out** is shown in the following text editor window.



The results above indicate that both the race and the insurance coverage category of a respondent exert a statistically significant influence on the respondent's total medical expenditures if a significance level of 5% is used. In particular, these results suggest that respondents with more comprehensive medical insurance coverage (inscov91 = 1 or inscov92 = 1) spend, on the average, more on medical expenses than those who have less comprehensive insurance coverage (inscov91 = inscov92 = -1). In addition, there is sufficient evidence that Whites (racex1 to racex4 = -1) spend, on the average, more on medical expenses than American Indians, Eskimos, Asians and Blacks.

Estimated outcomes for different groups

By using the results above, the estimated model may be expressed as

$$\hat{E}[\text{TOTEXP}_k] = 4.59 + 0.02x_{1k} + 0.19x_{2k} - 0.27x_{3k} - 0.53x_{4k} + 0.73x_{5k} + 1.00x_{6k}$$

The estimated model above implies that the estimated mean health care expenditure for an Asian respondent with no insurance ($x_{3k} = 1$, $x_{5k} = -1$, $x_{6k} = -1$ and $x_{1k} = x_{2k} = x_{4k} = 0$) is given by

$$\exp(4.59 - 0.27 - 0.73 - 1.00) = \exp(2.59) = \$13.33$$

Similarly, the estimated mean health care expenditures for an Asian respondent with any private insurance and public insurance only follow as \$156.39 and \$204.69 respectively. For a White respondent with any private insurance coverage ($x_{1k} = x_{2k} = x_{3k} = x_{4k} = -1$, $x_{5k} = 1$, and $x_{6k} = 0$) the mean health care expenditures is estimated as

$$\exp(4.59 - 0.02 - 0.19 + 0.27 + 0.53 + 0.73) = \exp(5.91) = \$368.70.$$

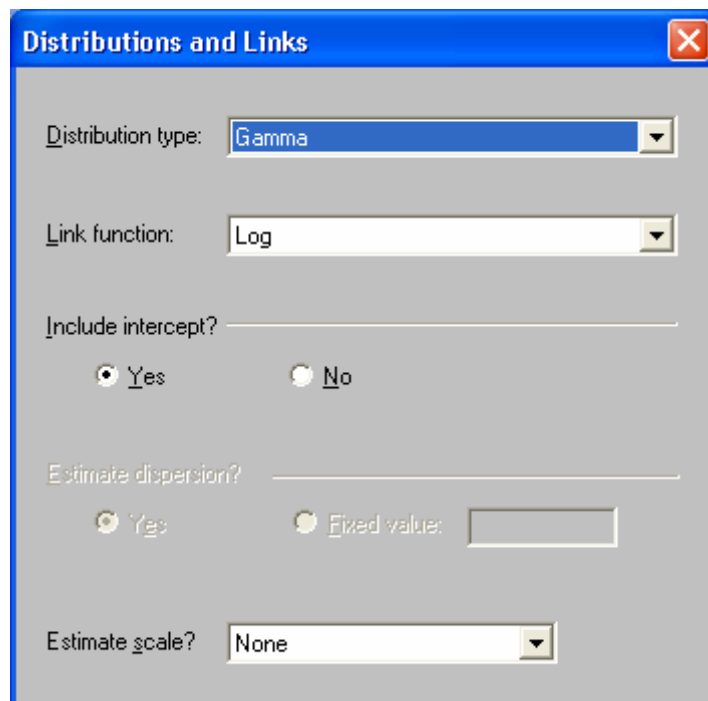
Likewise, for a White respondent with public insurance the corresponding estimate is \$482.99. This estimate of average health care expenditures will only be accurate if the outcome variable has a normal distribution. An analysis that takes the strongly skewed distribution of health care expenditures into account may produce quite different estimates, as will be seen in the next example.

Analyzing skewed outcome variables from complex survey designs (method 1)

The Normal-Identity GLIM assumes that the distribution of the response variable is symmetric about its mean. In the case of skewed response variables, which only assume values greater than zero, the Gamma and Inverse Gaussian sampling distributions will be more appropriate than the Normal distribution.

Setting up the analysis

The Gamma-log model can be fitted interactively to the data in **meps.psf** by replacing the Normal sampling distribution with the Gamma sampling distribution. Before doing so, specify a different title by selecting the **Title and Options** option on the SURVEYGLIM menu to access the **Title and Options** dialog box and then entering the title **A Gamma-Log model for MEPS Data** in the **Title** string field. Click on the **Next** button to go to the **Distributions and Links** dialog box and select the **Gamma** option from the **Distribution type** drop-down list box to produce the following **Distributions and Links** dialog box.



Since this is all we need to modify, click on the **Next** buttons of the **Distributions and Links** and the **Dependent and Independent Variables** dialog boxes and the **Finish** button of the **Survey Design** dialog box to open the following text editor window for **meps.pr2**.

```

meps.PR2
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 IterDetails=No
Method=Fisher;
Title=;
SY='C:\Program Files\lisrel870\meps2\meps.PSF';
Distribution=GAM;
Link=LOG;
Intercept=Yes;
Scale=None;
DepVar=TOTEXP99;
CoVars=racex$ inscov9$;
Stratum=VARSTR99;
Cluster=VARPSU99;
Weight=PERWT99F;
x

```

Submit the syntax file above by clicking on the **Run Prelis** toolbar icon to generate the corresponding output file **meps.out**.

Discussion of results – Gamma-log model

A portion of the resulting output file is shown in the text editor window below.

```

meps.OUT
-----
Statistic              Value      Den. DF   Num. DF   P Value
-----
Adjusted Wald F        129.7851      6         312      0.000000
Wald Chi-square        791.1897      6         0.000000

Note: The Wald F Test and Chi-square Statistics are statistics to test the
      null hypothesis that all the regression weights are equal to zero.

      Estimated Regression Weights

Parameter      Estimate      Standard      z Value      P Value
-----
intcept        1.4929        0.0169        88.3691      0.0000
racex1         0.0099        0.0393         0.2505      0.8022
racex2         0.0508        0.0465         1.0911      0.2752
racex3        -0.0554        0.0286        -1.9400      0.0524
racex4        -0.1194        0.0216        -5.5281      0.0000
inscov91       0.1742        0.0091        19.0741      0.0000
inscov92       0.2235        0.0106        21.1523      0.0000

```

At first glance, comparing the parameter estimates produced by the Normal-identity model (which assumes a normal distribution) and the Gamma-log model (which takes skewness in the outcome variable into account), it seems as if the race-related effects are radically different between the two. If, however, we order the values of the *racex* coefficients according to size, it turns out that for both the Normal-identity model and Gamma-log models the ordering is the same. This result is not unexpected since there exists a monotone relationship between any set of real numbers so that $r_1 > r_2 \rightarrow \exp(r_1) > \exp(r_2)$. Recall that for the identity link function

$$\hat{E}[\text{TOTEXP}_k] = \hat{\alpha} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \hat{\beta}_3 x_{3k} + \hat{\beta}_4 x_{4k} + \hat{\beta}_5 x_{5k} + \hat{\beta}_6 x_{6k}$$

whereas for the log-link function

$$\hat{E}[\text{TOTEXP}_k] = \exp(\hat{\alpha} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \hat{\beta}_3 x_{3k} + \hat{\beta}_4 x_{4k} + \hat{\beta}_5 x_{5k} + \hat{\beta}_6 x_{6k})$$

Substitution of the predictor values, using the appropriate parameter estimates, in any of the equations above, shows that the expected total expenditure values do not differ substantially.

Estimated outcomes for different groups

The fitted model is given by

$$\hat{E}[\text{TOTEXP}_k] = \exp(1.49 + 0.01x_{1k} + 0.05x_{2k} - 0.06x_{3k} - 0.12x_{4k} + 0.17x_{5k} + 0.22x_{6k}).$$

The estimated model above implies that the estimated mean health care expenditure for a White respondent with no insurance ($x_{1k} = x_{2k} = x_{3k} = x_{4k} = x_{5k} = x_{6k} = -1$) is given by

$$\exp(\exp(1.49 + -0.01 - 0.05 + 0.06 + 0.12 - 0.17 - 0.22)) = \exp(1.22) = \$29.58.$$

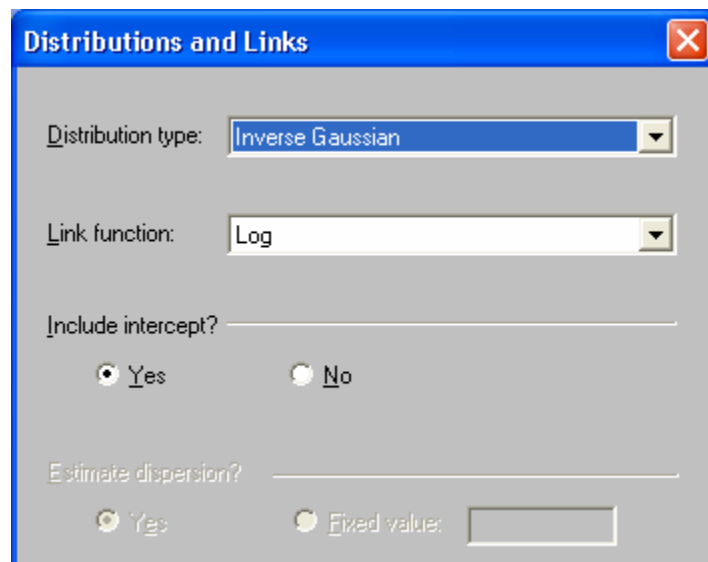
Similarly, the estimated mean health care expenditures for a White respondent with any private insurance and public insurance only follow as \$376.10 and \$509.73 respectively. The results above also indicate that $\exp(\hat{\beta}_4) = \exp(-0.12) = 0.88$ which implies that, on the average, Black respondents spent 12% less on health care in 1999 than other respondents. Similarly, it follows that $\exp(-\hat{\beta}_5 - \hat{\beta}_6) = \exp(-0.39) = 0.68$ which implies that, on the average, respondents with no insurance spent 32% less than other respondents on health care in 1999.

Analyzing skewed outcome variables from complex survey designs (method 2)

To explore the relationship between a respondent's total health related expenditure and his/her ethnicity and level of insurance coverage, we fit a GLIM model with inverse Gaussian distribution and log link function. Note that the mean model of the Inverse Gaussian-log GLIM is identical to that of the Gamma-log GLIM.

Setting up the analysis

Again, first modify the title by selecting the **Title and Options** option on the SURVEYGLIM menu and entering the title **An Inverse Gaussian-Log Model for MEPS Data** in the **Title** string field. Go to the **Distributions and Links** dialog box by clicking on the **Next** button and select the **Inverse Gaussian** option from the **Distribution type** list box to produce the following **Distributions and Links** dialog box.



This completes our modifications. Click on the **Next** buttons of the **Distributions and Links** and the **Dependent and Independent Variables** dialog boxes and the **Finish** button of the **Survey Design** dialog box to open the following text editor window for **meps.pr2**.

```

MEPS.PR2
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 IterDetails=No
          Method=Fisher;
Title=An Inverse Gaussian-Log Model for MEPS Data;
SY='MEPS.PSF';
Distribution=IMVG;
Link=LOG;
Intercept=Yes;
Scale=None;
DepVar=TOTEXP99;
CoVars= racex$ inscov9$;
Stratum=VARSTR99;
Cluster=VARPSU99;
Weight=PERWT99F;

```

The corresponding output file **meps.out** is obtained by clicking on the **Run Prelis** toolbar icon.

Discussion of results – Inverse Gamma-log model

Some selected results of the output file **meps.out** are shown in the following text editor window.

Statistic	Value	Den. DF	Num. DF	P Value
Adjusted Wald F	95.9258	6	312	0.000000
Wald Chi-square	584.7787	6		0.000000

Note: The Wald F Test and Chi-square Statistics are statistics to test the null hypothesis that all the regression weights are equal to zero.

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	1.4956	0.0206	72.6733	0.0000
racex1	0.0090	0.0401	0.2257	0.8214
racex2	0.0615	0.0620	0.9922	0.3211
racex3	-0.0577	0.0351	-1.6469	0.0996
racex4	-0.1271	0.0263	-4.8324	0.0000
inscov91	0.1729	0.0100	17.2380	0.0000
inscov92	0.2238	0.0117	19.0717	0.0000

Like the Gamma-log model, the Inverse Gaussian-log model produced results that were very different from the Normal-identity model. Since the Gamma-log model and Inverse Gaussian-log model both take the skewed distribution of the outcome variable into account, it is not surprising that they produced similar parameter estimates, standard error estimates, and estimates of statistical significance in this example.

Estimated outcomes for different groups

The estimated model follows from the results above as

$$\hat{E}[\text{TOTEXP}_k] = \exp(1.50 + 0.01x_{1k} + 0.06x_{2k} - 0.06x_{3k} - 0.13x_{4k} + 0.17x_{5k} + 0.22x_{6k})$$

The fitted model above implies that the estimated mean health care expenditure for a Black respondent with no insurance ($x_{4k} = 1$, $x_{5k} = x_{6k} = -1$, and $x_{1k} = x_{2k} = x_{3k} = 0$) is given by

$$\exp(\exp(1.50 - 0.13 - 0.17 - 0.22)) = \exp(2.69) = \$14.74$$

Similarly, the estimated mean health care expenditures for a Black respondent with any private insurance and public insurance only follow as \$106.12 and \$134.79 respectively. The results above also indicate that $\exp(\hat{\beta}_2) = \exp(0.06) = 1.06$ which implies that, on the average, American Indian

respondents spent 6% more on health care in 1999 than other respondents. Similarly, it follows that $\exp(\hat{\beta}_5) = \exp(0.17) = 1.19$ which implies that, on the average, respondents with any private insurance spent 19% more than other respondents on health care in 1999.

GLIMs for binary responses

Binary response variables are often the focus of empirical studies. Examples of binary response variables are diagnosis of breast cancer (absent or present), heart disease (yes or no), damage to solid rocket booster joints (damage or no damage), and depression in substance abuse clients (yes or no), credit risk (good or bad), etc. The analysis of GLIMs with binary response variables with SURVEYGLIM is illustrated in this section. More specifically, Bernoulli-logit and Binomial-logit models are fitted to substance abuse data.

SURVEYGLIM can also fit models with binary response variables to either simple random sample or complex sample data. This feature is illustrated in this section by fitting Bernoulli-logit and Binomial-logit models the substance abuse data. In the special case of one trial for each observation, the Binomial distribution simplifies to the Bernoulli distribution, and either distribution can be used. However, if a number of trials variable is available, the Binomial distribution would be the appropriate choice.

The data

The data set forms part of the data library of the Alcohol and Drug Services Study and is described in Section 3. The data set to be analyzed consists of the complete cases for a selection of variables and is provided as the PSF **abuse1.psf** in the **SGLIMEX** subfolder of LISREL for Windows. The first portion of this data set is shown in the following PSF window.

	depr	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.00	0.00	0.00	4.00	4.00	190.70
2	0.00	0.00	0.00	4.00	2.00	371.90
3	1.00	0.00	1.00	4.00	2.00	371.90
4	0.00	0.00	1.00	4.00	2.00	371.90
5	0.00	0.00	0.00	4.00	4.00	47.00
6	0.00	1.00	0.00	4.00	4.00	47.00
7	1.00	0.00	0.00	4.00	4.00	47.00
8	0.00	1.00	1.00	4.00	4.00	47.00
9	0.00	0.00	0.00	4.00	4.00	47.00
10	1.00	1.00	0.00	4.00	4.00	47.00
11	1.00	0.00	0.00	4.00	4.00	47.00
12	0.00	0.00	0.00	4.00	4.00	47.00
13	0.00	0.00	0.00	4.00	4.00	47.00
14	0.00	0.00	0.00	4.00	4.00	47.00
15	0.00	0.00	0.00	4.00	4.00	47.00

The variables to be used in the subsequent GLIM analyses are

- CENREG is the census region of the client (1 for Northeast, 2 for Midwest, 3 for South and 4 for West).
- FACTYPE is the facility treatment type of the client (1 for residential treatment, 2 for outpatient methadone treatment, 3 for outpatient non-methadone treatment and 4 for more than one type of treatment).
- A2TWA0 is the design weight of the client.
- depr is the value of a dummy variable for the depression status (0 for no depression history and 1 for a history of depression) of the client.
- sex is the value of a dummy variable for the gender (0 for male and 1 for female) of the client.
- race_d is the value of a dummy variable for the race (0 for nonwhite and 1 for white) of the client.

The models

The sampling distributions

The sampling distribution of the Bernoulli-logit GLIM is the Bernoulli distribution whose probability density function is given by

$$f(y_k, \pi_k) = \pi_k^{y_k} (1 - \pi_k)^{1 - y_k}$$

where y_k denotes the binary response variable y for respondent k and π_k denotes the probability that y_k assumes a unit value. Another sampling distribution for binary response variables is the Binomial distribution, which is the sampling distribution of the Binomial-logit GLIM and has the following probability density function

$$f(y_k, \pi_k) = \binom{n_k}{n_k y_k} \pi_k^{n_k y_k} (1 - \pi_k)^{n_k (1 - y_k)}$$

where n_k denotes the number of trials. In the special case of one trial for each observation, the Binomial distribution simplifies to the Bernoulli distribution. The number of trials for each observation is usually provided as a variable of the data to which the Binomial-logit GLIMs are to be fitted. Similarly to the Poisson sampling distributions, a scale parameter can be used for the Binomial distribution to address under-dispersion or over-dispersion (see Section 5).

The probability models

The general probability model for the Bernoulli-logit and Binomial-logit GLIMs may be expressed as

$$\pi_k = \frac{\exp(\alpha + \beta_1 x_{1k} + \dots + \beta_r x_{rk})}{1 + \exp(\alpha + \beta_1 x_{1k} + \dots + \beta_r x_{rk})}$$

where π_k denotes the probability that subject k has a unit value for the response variable, x_{jk} denotes the value of the j -th predictor ($j=1,2,\dots,r$) for respondent k , and α , β_1 , ..., β_{r-1} , and β_r denote unknown parameters. The probability model for the specific Bernoulli-logit and Binomial-logit GLIMs is given by

$$P(\text{depr}_k = 1) = \frac{\exp(\alpha + \beta_1 * \text{sex}_k + \beta_2 * \text{race_d}_k)}{1 + \exp(\alpha + \beta_1 * \text{sex}_k + \beta_2 * \text{race_d}_k)}$$

where $P(\text{depr}_k = 1)$ denotes the probability that client k has a history of depression and α , β_1 and β_2 denote unknown parameters. The ratio of the probabilities that a female client ($\text{sex}_k = 1$) and a male client ($\text{sex}_k = 0$) has a history of depression respectively follows as

$$\frac{\exp(\alpha + \beta_1 + \beta_2 * \text{race_d})}{1 + \exp(\alpha + \beta_2 * \text{race_d})} = \exp(\beta_1)$$

In a similar fashion, it follows that $\exp(\beta_2)$ is the ratio of the probabilities that a white client and a nonwhite client have a history of depression respectively. The corresponding estimated model follows as

$$\hat{P}(\text{depr}_k = 1) = \frac{\exp(\hat{\alpha} + \hat{\beta}_1 * \text{sex}_k + \hat{\beta}_2 * \text{race_d}_k)}{1 + \exp(\hat{\alpha} + \hat{\beta}_1 * \text{sex}_k + \hat{\beta}_2 * \text{race_d}_k)}$$

where $\hat{P}(\text{depr}_k = 1)$ denotes the estimated probability that client k has a history of depression and $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the maximum likelihood estimates of α , β_1 and β_2 respectively.

Analyzing binary outcomes from complex survey designs (method 1)

To explore a potential link between depression and a respondent's gender and ethnicity, a GLIM with Bernoulli distribution and logit link function is fitted to the data described above. The Bernoulli distribution is used since the outcome variable, *depr*, is dichotomous (0 for no depression history and 1 for a history of depression).

Setting up the analysis

We first open the file **abuse1.psf** in a PSF window using the following steps.

Use the **Open** option on the **File** menu of the root window of LISREL for Windows to load the **Open** dialog box.

Select the **Prelis Data (*.psf)** option from the **Files of type** drop-down list box.

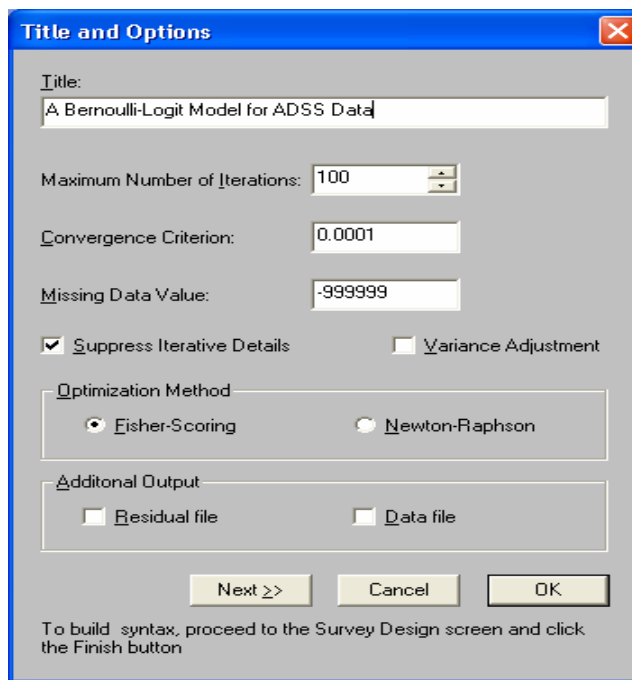
Browse for the file **abuse1.psf** in the **SGLIMEX** subfolder.

Click on the **Open** button to open the file **abuse1.psf** in a PSF window.

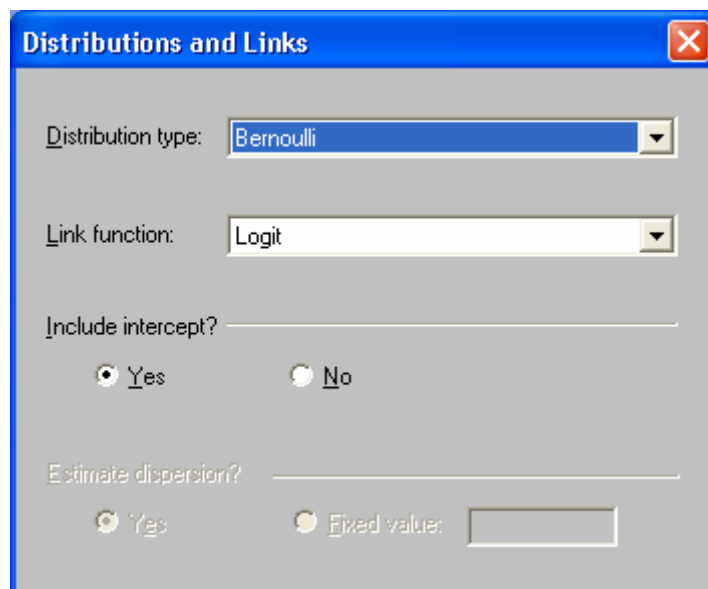
Click on the **SURVEYGLIM** menu to produce the following PSF window.

	depr	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.00	0.00	0.00	4.00	4.00	190.70
2	0.00	0.00	0.00	4.00	2.00	371.90
3	1.00	0.00	1.00	4.00	2.00	371.90
4	0.00	0.00	1.00	4.00	2.00	371.90
5	0.00	0.00	0.00	4.00	4.00	47.00
6	0.00	1.00	0.00	4.00	4.00	47.00
7	1.00	0.00	0.00	4.00	4.00	47.00
8	0.00	1.00	1.00	4.00	4.00	47.00
9	0.00	0.00	0.00	4.00	4.00	47.00
10	1.00	1.00	0.00	4.00	4.00	47.00
11	1.00	0.00	0.00	4.00	4.00	47.00
12	0.00	0.00	0.00	4.00	4.00	47.00
13	0.00	0.00	0.00	4.00	4.00	47.00
14	0.00	0.00	0.00	4.00	4.00	47.00
15	0.00	0.00	0.00	4.00	4.00	47.00

We can now use the **SURVEYGLIM** menu to fit the Bernoulli-logit GLIM to the data in **abuse1.psf**. First, select the **Title and Options** option on the **SURVEYGLIM** menu to go to the **Title and Options** dialog box. Enter the title **A Bernoulli-Logit Model for ADSS Data** into the **Title** string field to produce the following **Title and Options** dialog box.

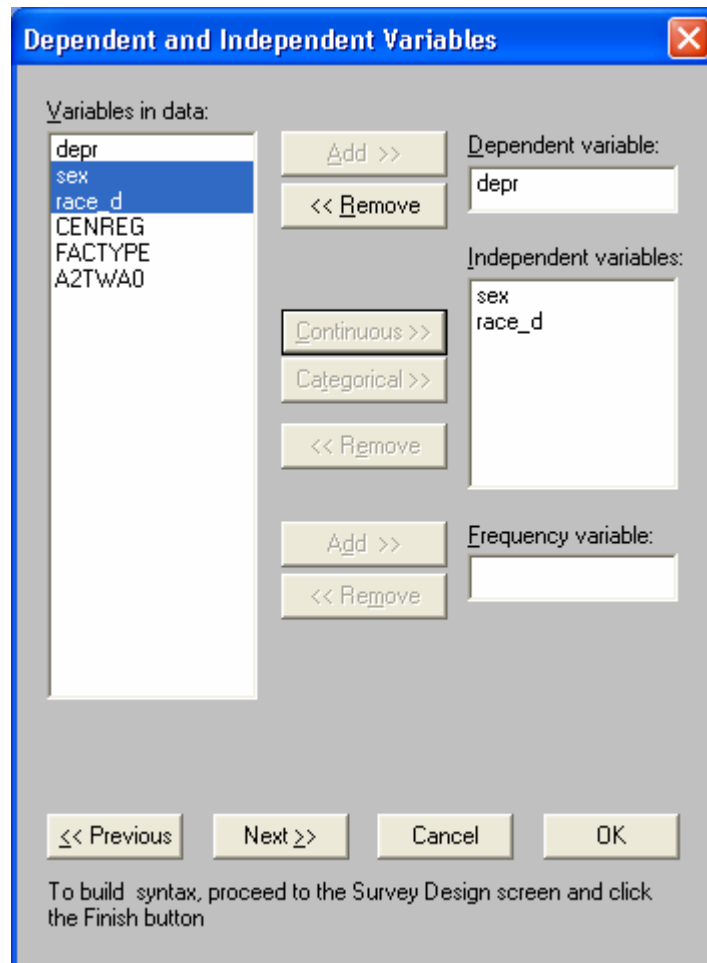


Click on the **Next** button to access the **Distributions and Links** dialog box and select the **Bernoulli** option from the **Distribution type** drop-down list box to produce the following **Distributions and Links** dialog box.

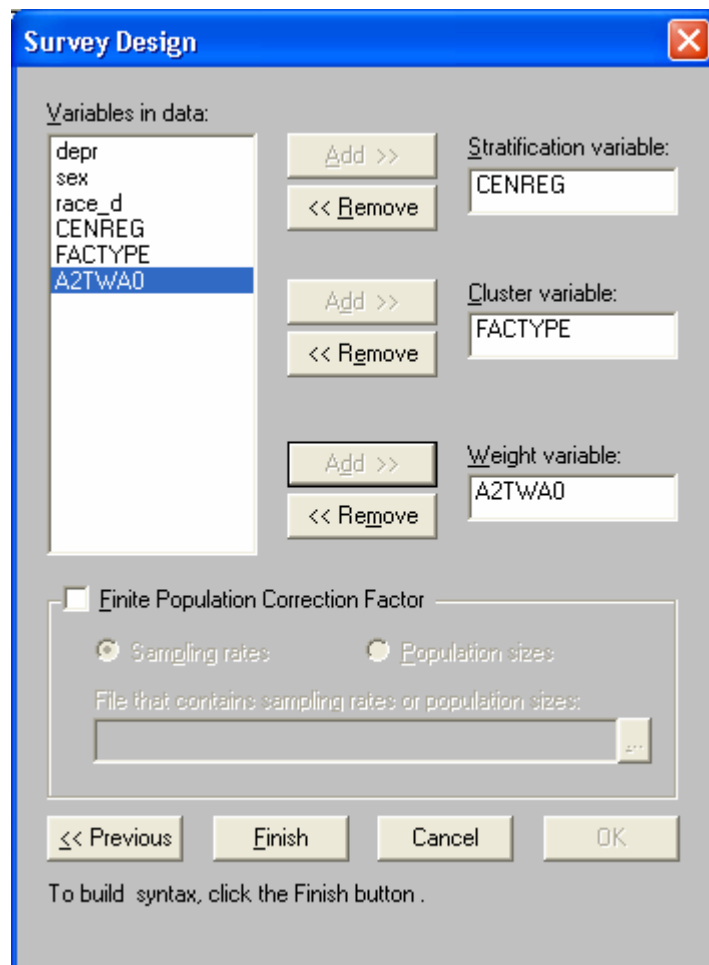


Click on the **Next** button to go to the **Dependent and Independent Variables** dialog box.

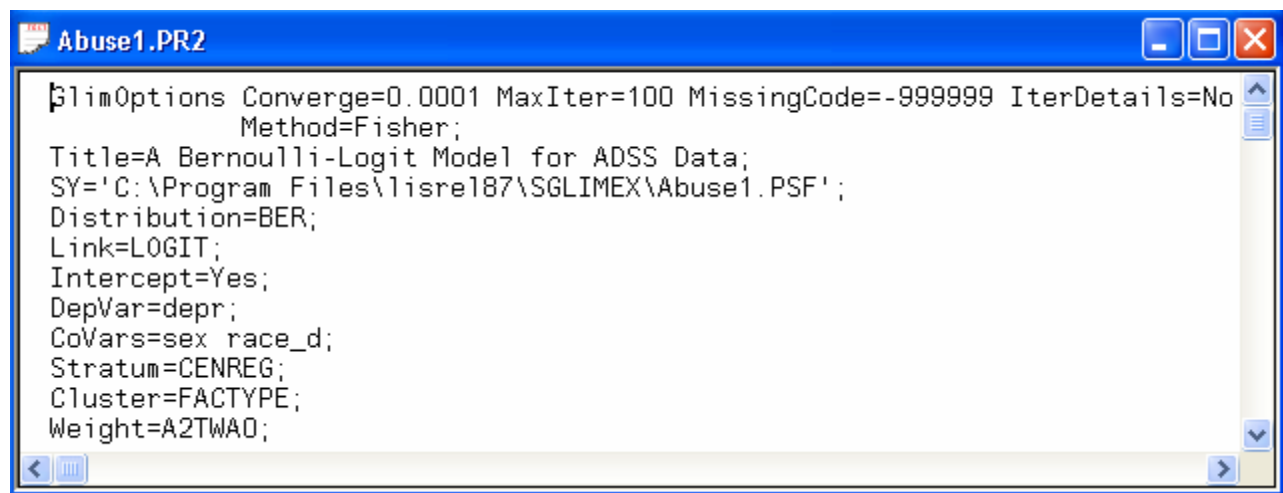
Specify the response variable depr by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Dependent variable** section. Specify the covariates, sex and race_d, by selecting them from the **Variables in data** list box and clicking on the **Continuous** button of the **Independent variables** section to produce the following **Dependent and Independent Variables** dialog box.



Click on the **Next** button to access the **Survey Design** dialog box. Specify the stratification variable, CENREG, by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Stratification variable** section. Similarly, specify the cluster variable, FACTYPE, and the weight variable, A2TWA0, by using the **Add** buttons of the **Cluster variable** and the **Weight variable** section to produce the following **Survey Design** dialog box.



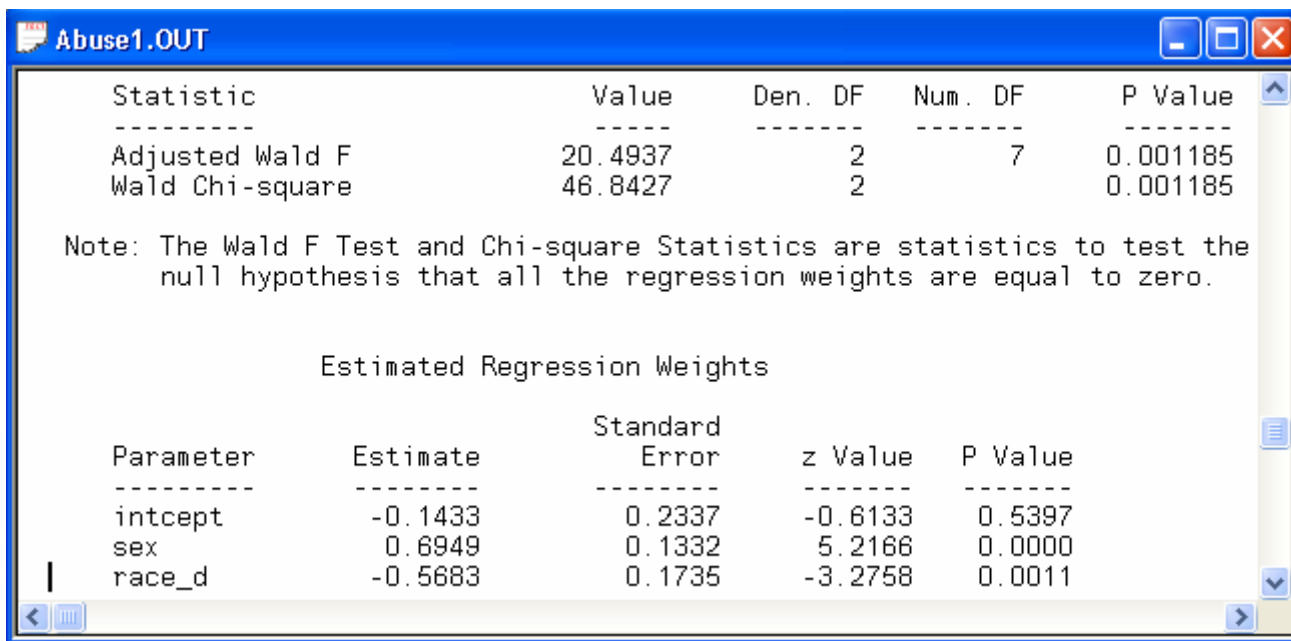
As this concludes our specifications, click on the **Finish** button to open the following text editor window for **abuse1.pr2**.



Submit the syntax file above by clicking on the **Run Prelis** toolbar icon to obtain the output file **abuse1.out**.

Discussion of results – Bernoulli-logit model

A portion of the output file **abuse1.out** is shown in the following text editor window.



The results above indicate that both the gender and the race of clients have a statistically significant influence on their depression status if a significance level of 5% is used. There is sufficient evidence to conclude that female clients (sex = 1) are more likely than male clients to have a depression history and that white clients (race_d = 1) are less likely than nonwhite clients to have a history of depression.

Estimated outcomes for different groups

The estimated model is obtained from the results above as

$$\hat{P}(\text{depr}_k = 1) = \frac{\exp(-0.14 + 0.70 \cdot \text{sex}_k - 0.57 \cdot \text{race_d}_k)}{1 + \exp(-0.14 + 0.70 \cdot \text{sex}_k - 0.57 \cdot \text{race_d}_k)}$$

The estimated probability that a nonwhite female client (sex_k = 1 and race_d_k = 0) has a history of depression follows from this fitted model as

$$\frac{\exp(-0.14 + 0.70)}{1 + \exp(-0.14 + 0.70)} = \frac{\exp(0.56)}{1 + \exp(0.56)} = 0.64$$

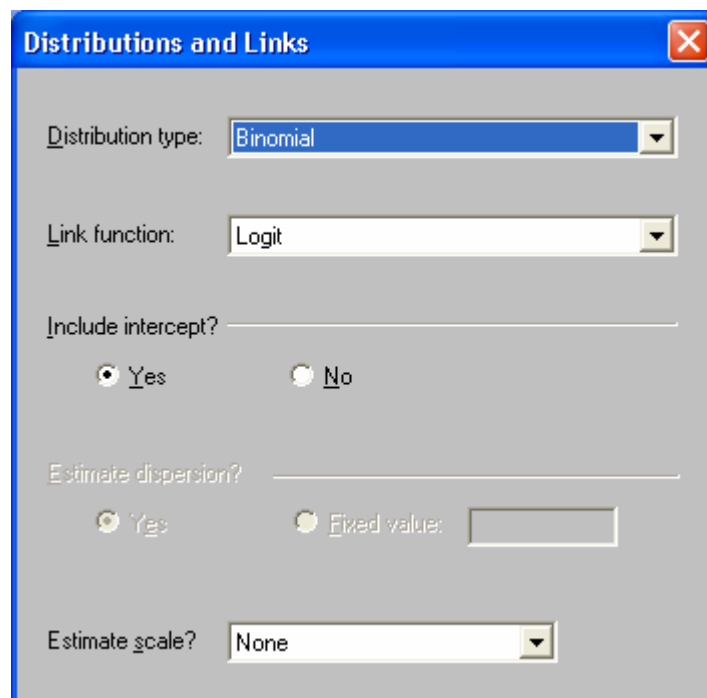
Similarly, the estimated probability that a nonwhite male client has a history of depression follows as 0.47. From the results above, it follows that $\exp(\hat{\beta}_1) = \exp(0.70) = 2.01$ which implies that female clients are twice as likely as male clients to have a history of depression. Similarly, $\exp(\hat{\beta}_1) = \exp(-0.57) = 0.57$ implies that whites are 43% less likely than nonwhites to have a history of depression.

Analyzing binary outcomes from complex survey designs (method 2)

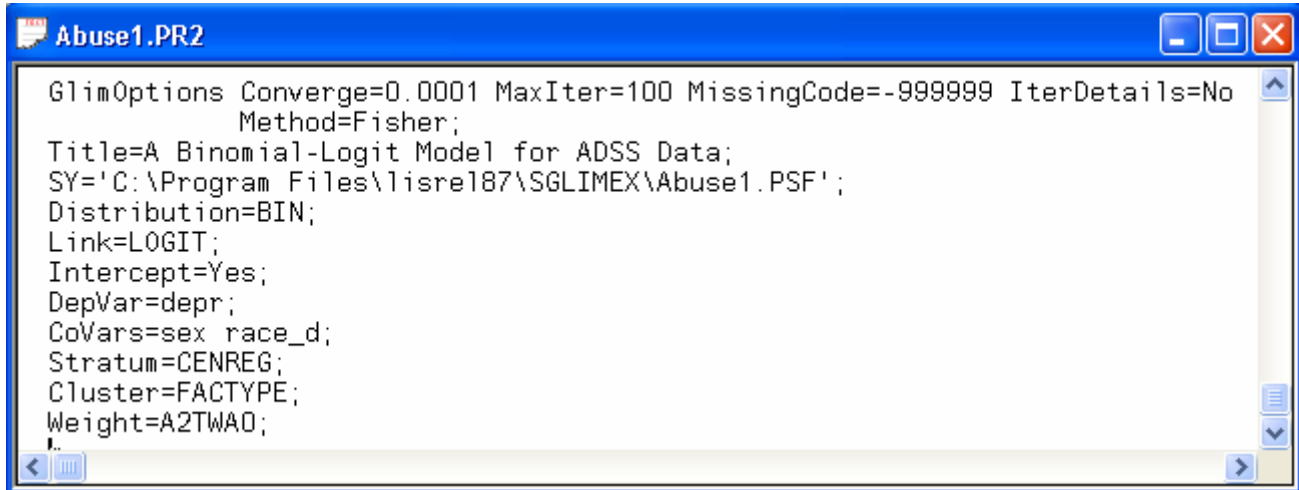
In this example, we illustrate that a GLIM with a Binomial distribution is identical to a GLIM with a Bernoulli distribution when the number of trials is one for each observation. If the `NTrials` command is omitted from the syntax file, the number of trials will automatically be set to unity.

Setting up the analysis

We fit the Binomial-logit GLIM to the data in `abuse1.psf` by specifying the Binomial sampling distribution instead of the Bernoulli sampling distribution. First, however, select the **Title and Options** option on the **SURVEYGLIM** menu to go to the **Title and Options** dialog box and enter the title **A Binomial-Logit Model for ADSS Data** into the **Title** string field. Click the **Next** button and select the **Binomial** option from the **Distribution type** drop-down list box to produce the following **Distributions and Links** dialog box.



Since these are the only changes we needed to specify, click on the **Next** buttons of the **Distributions and Links** and the **Dependent and Independent Variables** dialog boxes and the **Finish** button of the **Survey Design** dialog box to open the following text editor window for **abuse1.pr2**.



```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 IterDetails=No
          Method=Fisher;
Title=A Binomial-Logit Model for ADSS Data;
SY='C:\Program Files\lisrel87\SGLIMEX\Abuse1.PSF';
Distribution=BIN;
Link=LOGIT;
Intercept=Yes;
DepVar=depr;
CoVars=sex race_d;
Stratum=CENREG;
Cluster=FACTYPE;
Weight=A2TWA0;
```

Submit **abuse1.pr2** by clicking on the **Run Prelis** toolbar icon to generate the corresponding output file **abuse1.out**.

Discussion of results – Binomial-logit model

A selection of the results in the output file **abuse1.out** is shown in the following text editor window.

Abuse1.OUT

Statistic	Value	Den. DF	Num. DF	P Value
Adjusted Wald F	20.4937	2	7	0.001185
Wald Chi-square	46.8427	2		0.001185

Note: The Wald F Test and Chi-square Statistics are statistics to test the null hypothesis that all the regression weights are equal to zero.

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	-0.1433	0.2337	-0.6133	0.5397
sex	0.6949	0.1332	5.2166	0.0000
race_d	-0.5683	0.1735	-3.2758	0.0011

We note that the results above are identical to those obtained for the Bernoulli-logit GLIM. Hence, the conclusions based on the results above are identical to those reported for the Bernoulli-logit GLIM results. The reason for the identical results is that the number of trials was set to unity for each observation, in which case the Binomial sampling distribution simplifies to the Bernoulli sampling distribution.

GLIMs for ordinal responses

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 SURVEYGLIM. Cumulative logit and cumulative probit models are fitted to substance abuse data. Both logit and probit models usually lead to the same conclusion for the same data. Guidelines on when either of these models would be the more appropriate choice for given data are still being debated.

The data

The data set comes from part of the data library of the Alcohol and Drug Services Study and is described in Section 3. The data set to be analyzed consists of the complete cases for a selection of variables and is provided as the PSF **cntdiag.psf** in the **SGLIMEX** subfolder of LISREL for Windows. The first portion of this data set is shown in the following PSF window.

	cntdiag	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.00	0.00	0.00	4.00	4.00	190.70
2	0.00	0.00	0.00	4.00	4.00	44.30
3	0.00	0.00	0.00	4.00	4.00	44.30
4	0.00	1.00	0.00	4.00	4.00	44.30
5	0.00	0.00	0.00	4.00	4.00	44.30
6	0.00	0.00	0.00	4.00	4.00	44.30
7	0.00	0.00	0.00	4.00	4.00	44.30
8	0.00	0.00	0.00	4.00	4.00	44.30
9	0.00	0.00	0.00	4.00	4.00	44.30
10	0.00	0.00	1.00	4.00	2.00	371.90
11	0.00	0.00	1.00	4.00	2.00	371.90
12	0.00	0.00	1.00	4.00	2.00	371.90
13	0.00	0.00	1.00	4.00	2.00	371.90
14	0.00	0.00	1.00	4.00	2.00	371.90
15	0.00	0.00	0.00	4.00	2.00	371.90

A brief description of the variables to be used in the subsequent GLIM analyses follows.

- CENREG is the census region of the client (1 for Northeast, 2 for Midwest, 3 for South and 4 for West).
- FACTYPE is the facility treatment type of the client (1 for residential treatment, 2 for outpatient methadone treatment, 3 for outpatient non-methadone treatment and 4 for more than one type of treatment).
- A2TWA0 is the design weight of the client.
- cntdiag is the number of abuse diagnoses of the client (0, 1, 2 or 3).
- sex is the value of a dummy variable for the gender (0 for male and 1 for female) of the client.
- race_d is the value of a dummy variable for the race (0 for nonwhite and 1 for white) of the client.

The models

The sampling distribution

The sampling distribution of the cumulative logit and cumulative probit models is the Multinomial distribution whose probability density function is given by

$$f(\mathbf{y}_k, \boldsymbol{\pi}_k) = \frac{n!}{\left(\prod_{l=1}^{p-1} y_{kl}!\right) \left(n - \sum_{k=1}^{p-1} y_{ki}\right)!} \left(\prod_{l=1}^{p-1} \pi_{kl}^{y_{kl}}\right) \left(1 - \sum_{k=1}^{p-1} \pi_{ki}\right)^{n - \sum_{k=1}^{p-1} y_{ki}}$$

where \mathbf{y}_k denotes the vector of dummy variables for the p categories of the categorical response variable y for respondent k , π_{kl} denotes the probability that category l is recorded for client k and $\boldsymbol{\pi}_k = [\pi_{k1} \ \pi_{k2} \ \dots \ \pi_{kp}]'$.

The probability models

The general probability models for the cumulative logit and cumulative probit models are given by

$$\pi_{kl}^* = \sum_{m=1}^l \pi_{km} = \frac{\exp(\alpha_l + \beta_1 x_{1k} + \dots + \beta_r x_{rk})}{1 + \exp(\alpha_l + \beta_1 x_{1k} + \dots + \beta_r x_{rk})} \quad l = 1, 2, \dots, p-1$$

and

$$\pi_{kl}^* = \sum_{m=1}^l \pi_{km} = \Phi(\alpha_l + \beta_1 x_{1k} + \dots + \beta_r x_{rk}) \quad l = 1, 2, \dots, p-1$$

respectively where π_{km} denotes the probability that category m is recorded for subject k , x_{jk} denotes the value of the j -th predictor ($j = 1, 2, \dots, r$) for subject k , $\alpha_1, \alpha_2, \dots, \alpha_{p-1}, \beta_1, \dots, \beta_{r-1}$, and β_r denote unknown parameters and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Normal distribution. For illustrative purposes, the response variable cntdiag is treated here as ordinal rather than a count variable. The probability models for the specific cumulative logit and cumulative probit models are given by

$$P(\text{cntdiag}_k \leq l) = \frac{\exp(\alpha_l + \beta_1 * \text{sex}_k + \beta_2 * \text{race_d}_k)}{1 + \exp(\alpha_l + \beta_1 * \text{sex}_k + \beta_2 * \text{race_d}_k)} \quad l = 1, 2, 3$$

and

$$P(\text{cntdiag}_k \leq l) = \Phi(\alpha_l + \beta_1 * \text{sex}_k + \beta_2 * \text{race_d}_k) \quad l = 1, 2, 3$$

respectively where $P(\text{cntdiag}_k \leq l)$ denotes the cumulative probability that category l was recorded for client k and $\alpha_1, \alpha_2, \alpha_3, \beta_1$ and β_2 denote unknown parameters. The specific probabilities for the each response category for client k for both these models may be obtained from the following expressions.

$$P(\text{cntdiag}_k = 1) = P(\text{cntdiag}_k \leq 1)$$

$$P(\text{cntdiag}_k = 2) = P(\text{cntdiag}_k \leq 2) - P(\text{cntdiag}_k \leq 1)$$

$$P(\text{cntdiag}_k = 3) = P(\text{cntdiag}_k \leq 3) - P(\text{cntdiag}_k \leq 2).$$

In the case of the cumulative logit model, the ratio of the odds in the first l categories for a female client ($\text{sex}_k = 1$) and a male client ($\text{sex}_k = 0$) respectively follows as

$$\frac{\exp(\alpha_l + \beta_1 + \beta_2 * \text{race_d})}{\exp(\alpha_l + \beta_2 * \text{race_d})} = \exp(\beta_1)$$

Similarly, it follows that $\exp(\beta_2)$ is the ratio of the odds for a white client and a nonwhite client respectively. The corresponding estimated probability models are given by

$$\hat{P}(\text{cntdiag}_k \leq l) = \frac{\exp(\hat{\alpha}_l + \hat{\beta}_1 * \text{sex}_k + \hat{\beta}_2 * \text{race_d}_k)}{1 + \exp(\hat{\alpha}_l + \hat{\beta}_1 * \text{sex}_k + \hat{\beta}_2 * \text{race_d}_k)} \quad l = 1, 2, 3$$

and

$$\hat{P}(\text{cntdiag}_k \leq l) = \Phi(\hat{\alpha}_l + \hat{\beta}_1 * \text{sex}_k + \hat{\beta}_2 * \text{race_d}_k) \quad l = 1, 2, 3$$

respectively where $\hat{P}(\text{cntdiag}_k \leq l)$ denotes the estimated cumulative probability that at most the number of diagnoses listed in the first l categories are recorded for client k and $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}_1$ and $\hat{\beta}_2$ denote the maximum likelihood estimates of $\alpha_1, \alpha_2, \alpha_3, \beta_1$ and β_2 respectively.

Analyzing ordinal outcomes from complex survey designs (method 1)

In a previous example, a GLIM with a Poisson distribution and a log link function was used to examine the possible association between ethnicity and gender effects and the number of substance abuse diagnoses (cntdiag). Since this variable assumes values between 0 and 3 in the sample data, an alternative approach is to examine the strength of the relationship between the predictors and the cumulative number of diagnoses. A GLIM with a multinomial distribution and a cumulative logit link function may be used for this purpose.

Setting up the analysis

We start by opening the data file to be processed, **cntdiag.psf**, in a PSF window as follows.

Use the **Open** option on the **File** menu of the root window of LISREL for Windows to load the **Open** dialog box.

Select the **Prelis Data (*.psf)** option from the **Files of type** drop-down list box.

Browse for the file **cntdiag.psf** in the **SGLIMEX** subfolder.
 Click on the **Open** button to open the file **cntdiag.psf** in a PSF window.
 Click on the **SURVEYGLIM** menu to produce the following PSF window.

The screenshot shows the LISREL Windows Application window titled "cntdiag.PSF". The "SurveyGLIM" menu is open, displaying options: "Title and Options...", "Distributions/Links...", "Model Specification...", and "Survey Design...". Below the menu is a data table with 10 rows and 7 columns.

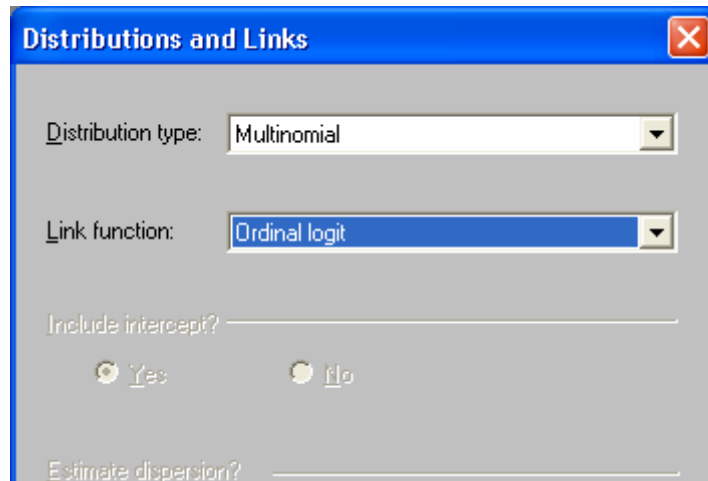
	cntdiag	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.00	0.00	0.00	4.00	4.00	190.70
2	0.00	0.00	0.00	4.00	4.00	44.30
3	0.00	0.00	0.00	4.00	4.00	44.30
4	0.00	1.00	0.00	4.00	4.00	44.30
5	0.00	0.00	0.00	4.00	4.00	44.30
6	0.00	0.00	0.00	4.00	4.00	44.30
7	0.00	0.00	0.00	4.00	4.00	44.30
8	0.00	0.00	0.00	4.00	4.00	44.30
9	0.00	0.00	0.00	4.00	4.00	44.30
10	0.00	0.00	1.00	4.00	2.00	371.90

Next, we specify the analysis as follows. Select the **Title and Options** option on the **SURVEYGLIM** menu to go to the **Title and Options** dialog box. Then enter the title **A cumulative logit model** into the **Title** string field to produce the following **Title and Options** dialog box.

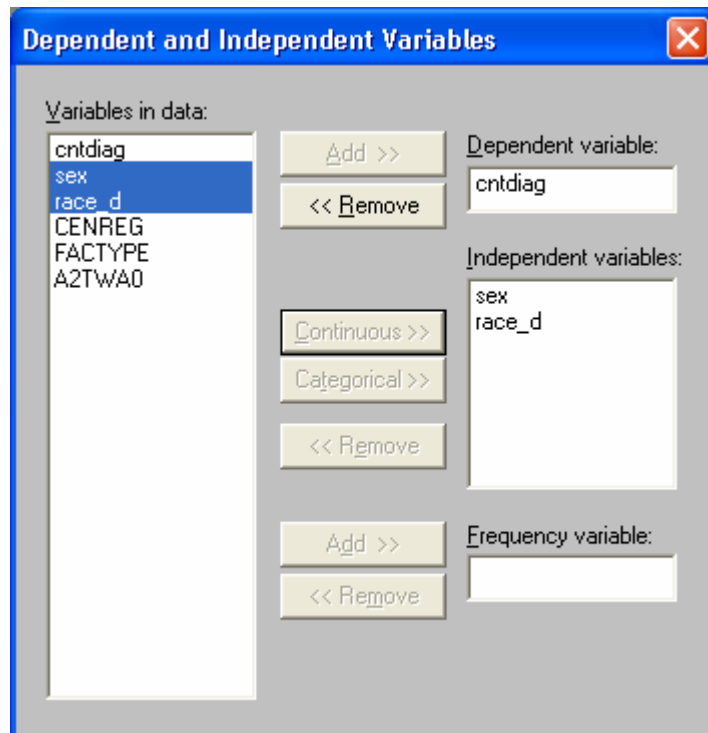
The "Title and Options" dialog box is shown with the following settings:

- Title:** A cumulative logit model
- Maximum Number of Iterations:** 100
- Convergence Criterion:** 0.0001
- Missing Data Value:** -999999

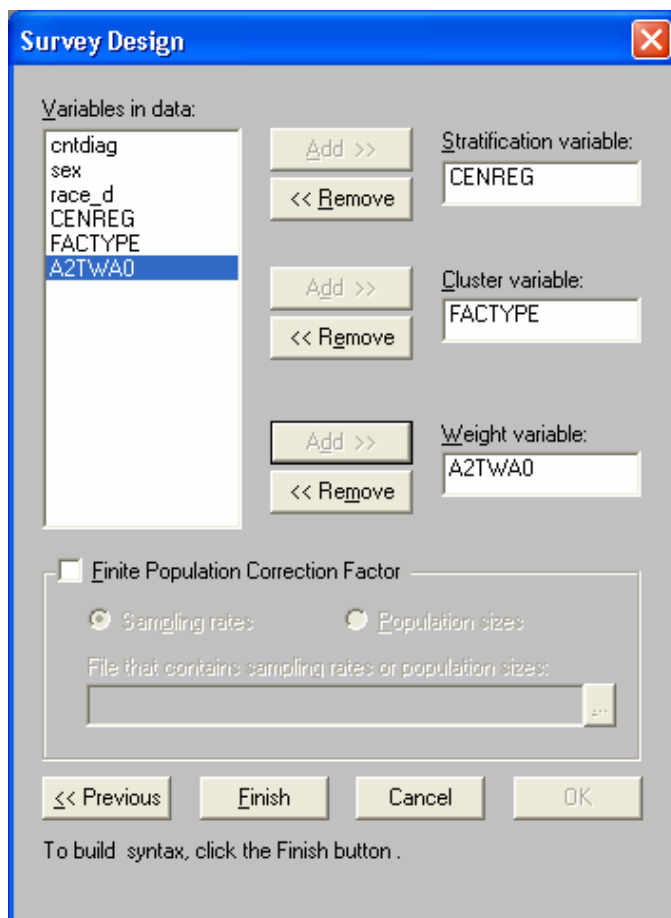
Click on the **Next** button to access the **Distributions and Links** dialog box, and select the **Multinomial** option from the **Distribution type** drop-down list box and the **Ordinal logit** option from the **Link function** drop-down list box to produce the following **Distributions and Links** dialog box.



Click on the **Next** button to go to the **Dependent and Independent Variables** dialog box. Specify the response variable cntdiag by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Dependent variable** section. Specify the covariates, sex and race_d, by selecting them from the **Variables in data** list box and clicking on the **Continuous** button of the **Independent variables** section to produce the following **Dependent and Independent Variables** dialog box.



Click on the **Next** button to access the **Survey Design** dialog box. Specify the stratification variable, CENREG, by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Stratification variable** section. Similarly, specify the cluster variable, FACTYPE, and the weight variable, A2TWA0, by using the **Add** buttons of the **Cluster variable** and the **Weight variable** sections respectively to produce the following **Survey Design** dialog box.



Since our desired analysis is now specified, click on the **Finish** button to open the following text editor window for **cntdiag.pr2**.

```

cntdiag.PR2
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 IterDetails=No
Method=Fisher;
Title=A cumulative logit model;
SY='C:\Program Files\lisrel87\SGLIMEX\cntdiag.PSF';
Distribution=MUL;
Link=OLOGIT;
DepVar=cntdiag;
CoVars=sex race_d;
Stratum=CENREG;
Cluster=FACTYPE;
Weight=A2TWAO;

```

Click on the **Run Prelis** toolbar icon to submit the syntax file above to obtain the output file **cntdiag.out**.

Discussion of results – Cumulative-logit model

A portion of the output file **cntdiag.out** is shown in the following text editor window.

```

cntdiag.OUT

Statistic              Value      Den. DF   Num. DF   P Value
-----              -
Adjusted Wald F        1.7497           2         7   0.241969
Wald Chi-square        3.9994           2         2   0.241969

Note: The Wald F Test and Chi-square Statistics are statistics to test the
      null hypothesis that all the regression weights are equal to zero.

Estimated Regression Weights

Parameter      Estimate      Standard      z Value      P Value
-----      -
Alpha1         -1.6891         0.3154        -5.3554       0.0000
Alpha2          0.3493         0.1650         2.1172       0.0342
Alpha3          1.9046         0.1348        14.1299       0.0000
sex            -0.2012         0.2157        -0.9330       0.3508
race_d         -0.3943         0.2020        -1.9518       0.0510

```

At a 5% level of significance the results above indicate that there is insufficient evidence that gender and race affect the cumulative probabilities of the number of diagnoses of clients. Although the results for race_d border on statistical significance, interpreting the test of the parameter

estimate precisely is consistent with the non-significance of the omnibus test of the model (see the Wald F-test and Wald χ^2 -statistic).

Estimated outcomes for different groups

Since $\hat{\alpha}_1 = -1.69$, the estimated probability that a white female client ($\text{race}_k = 1$ and $\text{sex}_k = 1$) has no diagnoses follows from the results above as

$$\hat{P}(\text{cntdiag}_k = 1) = \hat{P}(\text{cntdiag}_k \leq 1) = \frac{\exp(-1.69 - 0.20 - 0.39)}{1 + \exp(-1.69 - 0.20 - 0.39)} = 0.09$$

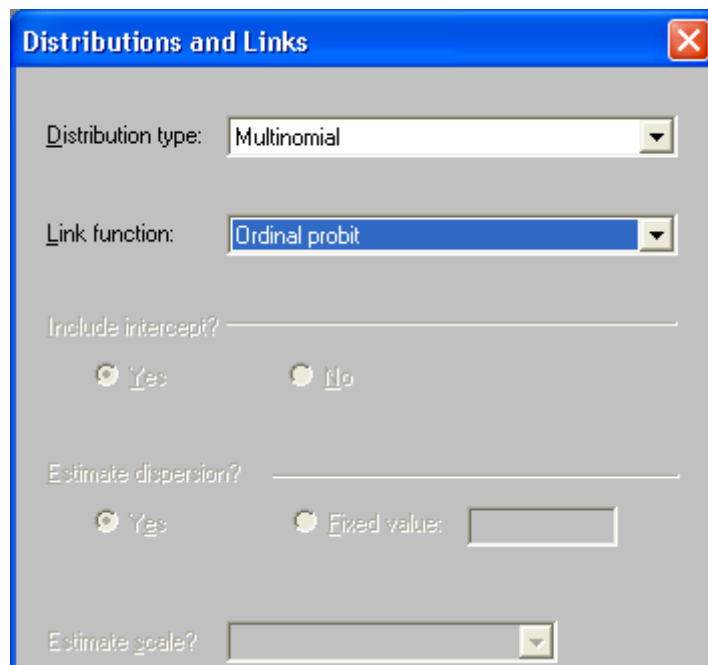
Similarly, the estimated probabilities that a white female client has at most 1 diagnosis and 2 diagnoses follow as 0.44 and 0.79 respectively. These estimated cumulative probabilities imply that the estimated probabilities that a white female client has 1 diagnosis, 2 diagnoses and 3 diagnoses are $0.44 - 0.09 = 0.35$, $0.79 - 0.44 = 0.35$ and $1 - 0.09 - 0.35 - 0.35 = 0.21$ respectively. The effect estimates, $\hat{\beta}_1 = -0.20$ and $\hat{\beta}_2 = -0.39$, suggest that the cumulative probability starting at the no diagnoses end of the scale decreases for both females and whites. Given the race of a client, the estimated probability of a number of diagnoses below any level for a female client is $\exp(-0.20) = 0.82$ times the estimated probability for a male client. Similarly, given the gender of a client, the estimated probability of a number of diagnoses below any level for a white client is $\exp(-0.39) = 0.68$ times the estimated probability for a nonwhite client.

Analyzing ordinal outcomes from complex survey designs (method 2)

In the previous example we examined the strength of the relationship between ethnicity, gender, and the cumulative number of substance abuse diagnoses. A GLIM with a multinomial distribution and a cumulative logit link function was used to do so. To study the effect of using a different type of link function, a probit link function is used here.

Setting up the analysis

We fit the cumulative probit model to the data in **cntdiag.psf** by specifying the cumulative probit link function instead of the cumulative logit link function. This is accomplished as follows. First modify the title by selecting the **Title and Options** option on the **SURVEYGLIM** menu to go to the **Title and Options** dialog box and enter the title **A cumulative probit model** into the **Title** string field. Then click on the **Next** button to access the **Distributions and Links** dialog box and select the **Ordinal probit** option from the **Link function** drop-down list box to produce the following **Distributions and Links** dialog box.



Since this concludes the modifications, click on the **Next** buttons of the **Distributions and Links** and the **Dependent and Independent Variables** dialog boxes and the **Finish** button of the **Survey Design** dialog box to open the following text editor window for **cntdiag.pr2**.

```

GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 IterDetails=No
          Method=Fisher;
Title=A cumulative probit model;
SY='C:\Program Files\lisrel87\SGLIMEX\cntdiag.PSF';
Distribution=MUL;
Link=OPROBIT;
DepVar=cntdiag;
CoVars=sex race_d;
Stratum=CENREG;
Cluster=FACTYPE;
Weight=A2TWA0;

```

Click on the **Run Prelis** toolbar icon to submit **cntdiag.pr2** to generate the corresponding output file **cntdiag.out**.

Discussion of results – Cumulative-probit model

A portion of the output file **cntdiag.out** is shown in the following text editor window.

Statistic Value Den. DF Num. DF P Value

Adjusted Wald F 1.1546 2 7 0.368680

Wald Chi-square 2.6391 2 7 0.368680

Note: The Wald F Test and Chi-square Statistics are statistics to test the null hypothesis that all the regression weights are equal to zero.

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
Alpha1	-1.0128	0.1708	-5.9290	0.0000
Alpha2	0.2017	0.1017	1.9841	0.0472
Alpha3	1.1214	0.0766	14.6400	0.0000
sex	-0.1036	0.1255	-0.8250	0.4094
race_d	-0.1884	0.1171	-1.6082	0.1078

A comparison of the results above with those obtained for the cumulative logit model indicates that although they differ, the same conclusions about the effect of gender and race on the cumulative probabilities of the number of diagnoses apply.

Since $\hat{\alpha}_1 = -1.01$, the estimated probability that a nonwhite male client ($\text{race_d} = 0, \text{sex} = 0$) has no diagnoses follows from the results above as

$$\hat{P}(\text{cntdiag}_k = 1) = \hat{P}(\text{cntdiag}_k \leq 1) = \Phi(-1.01) = 0.16$$

Similarly, the estimated probabilities that a nonwhite male client has at most 1 diagnosis and 2 diagnoses follow as 0.58 and 0.87 respectively. These estimated cumulative probabilities imply that the estimated probabilities that a white female client ($\text{race_d} = 1, \text{sex} = 1$) has 1 diagnosis, 2 diagnoses and 3 diagnoses are $0.58 - 0.16 = 0.42$, $0.87 - 0.58 = 0.29$ and $1 - 0.16 - 0.42 - 0.29 = 0.13$ respectively. The effect estimates, $\hat{\beta}_1 = -0.10$ and $\hat{\beta}_2 = -0.19$, suggest that the cumulative probability starting at the no diagnoses end of the scale decreases for both females and whites.

GLIMs for nominal responses

SURVEYGLIM can also be used to fit models to nominal response variables. The primary food choice of alligators (fish, invertebrate, reptile, bird or other), smoking status (never smoked, former smoker or current smoker), preference for U.S. President (Democrat, Republican or Independent), cancer type of female cancer patients (breast, lung, brain, leukemia, liver, colon or other), etc. are

examples of nominal response variables. In this section, we illustrate this feature by fitting a generalized logistic model to health-related data. A description of the data follows.

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. The first portion of the data set to be used is shown in the following PSF window.

More information on the NHIS and the data are available at

http://www.cdc.gov/nchs/about/major/nhis/quest_data_related_1997_forward.htm



	VYEAR	AGE	SEX	USETOBAC	PRIMCARE	PASTVIS
1	2002.00	35.00	2.00	0.00	1.00	3.00
2	2002.00	21.00	2.00	1.00	1.00	3.00
3	2002.00	2.00	2.00	2.00	1.00	3.00
4	2002.00	52.00	1.00	0.00	1.00	2.00
5	2002.00	13.00	2.00	3.00	1.00	3.00
6	2002.00	35.00	2.00	3.00	1.00	3.00
7	2002.00	82.00	1.00	2.00	1.00	0.00
8	2002.00	30.00	1.00	0.00	1.00	2.00
9	2002.00	73.00	2.00	2.00	1.00	3.00
10	2002.00	38.00	2.00	0.00	1.00	2.00
11	2002.00	43.00	1.00	3.00	1.00	3.00
12	2002.00	14.00	2.00	2.00	1.00	3.00
13	2002.00	54.00	1.00	3.00	1.00	2.00
14	2002.00	43.00	1.00	0.00	1.00	2.00
15	2002.00	39.00	1.00	0.00	1.00	3.00

The variables to be utilized in the subsequent analyses are

- CSTRATM is the stratum of the participant.
- CPSUM is the PSU of the participant.
- PATWT is the design weight of the participant.
- PASTVIS is the value of a nominal variable for the number of visits to a medical doctor during the past 12 months (1 for blank, 2 for none, 3 for 1-2 visits, 4 for 3-5 visits, 5 for 6 or more visits, 6 for unknown and 7 for not ascertained) of the participant.
- AGE is the age of the participant.

- EXERCISE is the value of a dummy variable for the exercise status (0 for do exercise and 1 for do not exercise) of the participant.

The models

The sampling distribution

The sampling distribution of the generalized logistic model is the Multinomial distribution whose probability density function is given by

$$f(\mathbf{y}_k, \boldsymbol{\pi}_k) = \frac{n!}{\left(\prod_{l=1}^{p-1} y_{kl}!\right) \left(n - \sum_{k=1}^{p-1} y_{ki}\right)!} \left(\prod_{l=1}^{p-1} \pi_{kl}^{y_{kl}}\right) \left(1 - \sum_{k=1}^{p-1} \pi_{ki}\right)^{n - \sum_{k=1}^{p-1} y_{ki}}$$

where \mathbf{y}_k denotes the vector of dummy variables for the p categories of the categorical response variable y for respondent k , π_{kl} denotes the probability that client k responded with category l and $\boldsymbol{\pi}_k = [\pi_{k1} \ \pi_{k2} \ \dots \ \pi_{kp}]'$.

The probability model

The general probability model for the generalized logistic model is given by

$$\pi_{kl} = \frac{\exp(\alpha_l + \beta_{1l}x_{1k} + \dots + \beta_{rl}x_{rk})}{1 + \sum_{l=1}^{p-1} \exp(\alpha_l + \beta_{1l}x_{1k} + \dots + \beta_{rl}x_{rk})} \quad l = 1, 2, \dots, p-1$$

where π_{kl} represents the probability that client k responded with category l , x_{jk} denotes the value of the j -th predictor ($j = 1, 2, \dots, r$) for subject k and $\alpha_1, \alpha_2, \dots, \alpha_{p-1}, \beta_{11}, \beta_{12}, \dots, \beta_{1p-1}, \dots, \beta_{r1}, \beta_{r2}, \dots, \beta_{rp-1}$ denote unknown parameters.

The probability model for the specific generalized logistic model is given by

$$P(\text{PASTVIS}_k = l) = \frac{\exp(\alpha_l + \beta_{1l}\text{AGE}_k + \beta_{2l}\text{EXERCISE}_k)}{1 + \sum_{l=1}^6 \exp(\alpha_l + \beta_{1l}\text{AGE}_k + \beta_{2l}\text{EXERCISE}_k)} \quad l = 1, 2, \dots, 6$$

where $P(\text{PASTVIS}_k = l)$ denotes the probability that client k responded with category l , and $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25}$ and β_{26} denote unknown parameters.

The corresponding estimated probability model is given by

$$\hat{P}(\text{PASTVIS}_k = l) = \frac{\exp(\hat{\alpha}_l + \hat{\beta}_{1l}\text{AGE}_k + \hat{\beta}_{2l}\text{EXERCISE}_k)}{1 + \sum_{l=1}^6 \exp(\hat{\alpha}_l + \hat{\beta}_{1l}\text{AGE}_k + \hat{\beta}_{2l}\text{EXERCISE}_k)} \quad l = 1, 2, \dots, 6$$

where $\hat{P}(\text{PASTVIS}_k = l)$ is the estimated probability that client k responded with category l , and $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\beta}_{14}, \hat{\beta}_{15}, \hat{\beta}_{16}, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\beta}_{23}, \hat{\beta}_{24}, \hat{\beta}_{25}$ and $\hat{\beta}_{26}$ denote the maximum likelihood estimates of $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25}$ and β_{26} respectively.

Analyzing nominal outcomes from complex survey designs

In this example, we wish to examine the effect of exercise and age on the number of visits (PASTVIS) to a medical doctor during the past 12 months. Since the last two categories of the outcome variable are defined as "unknown" and "not ascertained", PASTVIS is a nominal variable. A suitable GLIM model is obtained by specifying a multinomial distribution with logit link function.

Setting up the analysis

Before the specific analysis can be specified, we need to open the file **nih1.psf** in a PSF window as follows.

Use the **Open** option on the **File** menu of the root window of LISREL for Windows to load the **Open** dialog box.

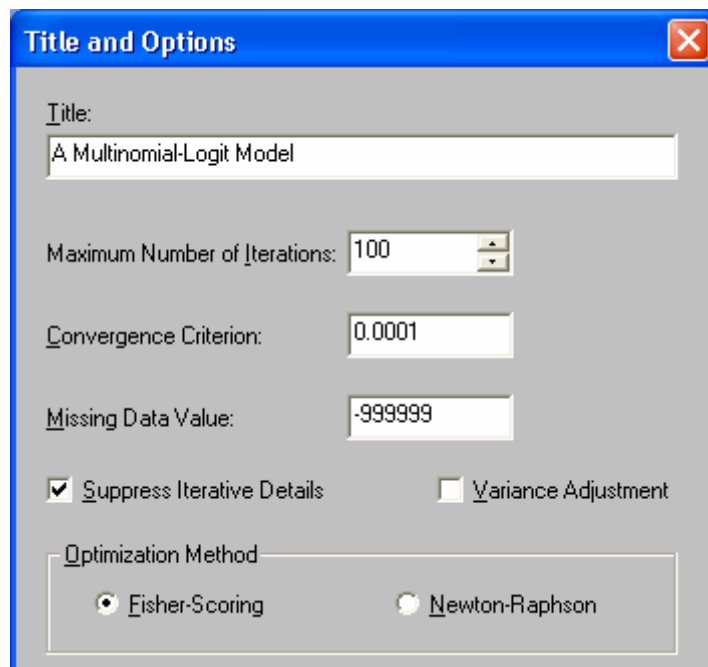
Select the **Prelis Data (*.psf)** option from the **Files of type** drop-down list box.

Browse for the file **nih1.psf** in the **SGLIMEX** subfolder.

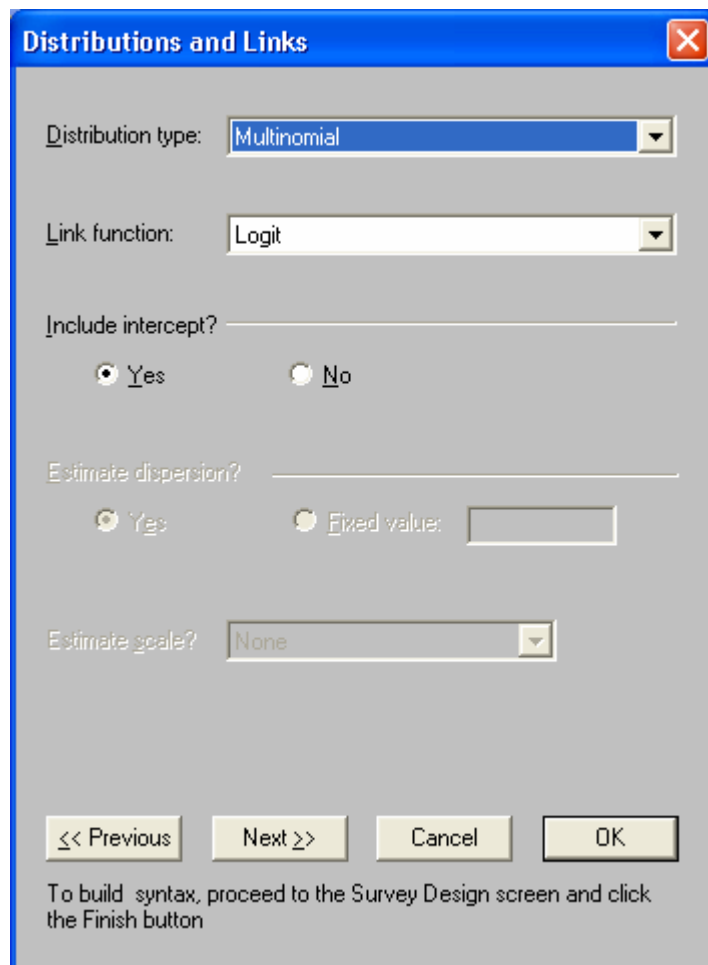
Click on the **Open** button to open the file **nih1.psf** in a PSF window.

Click on the **SURVEYGLIM** menu to produce the following PSF window.

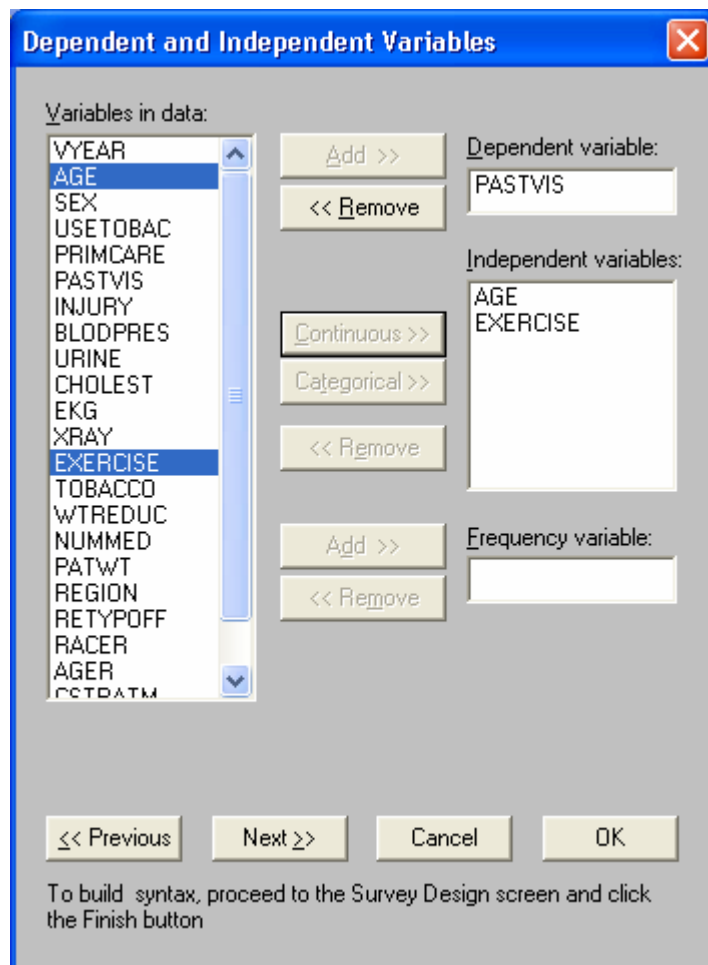
Continue by selecting the **Title and Options** option on the **SURVEYGLIM** menu to access the **Title and Options** dialog box and entering the title **A Multinomial-Logit Model** into the **Title** string field to produce the following **Title and Options** dialog box.



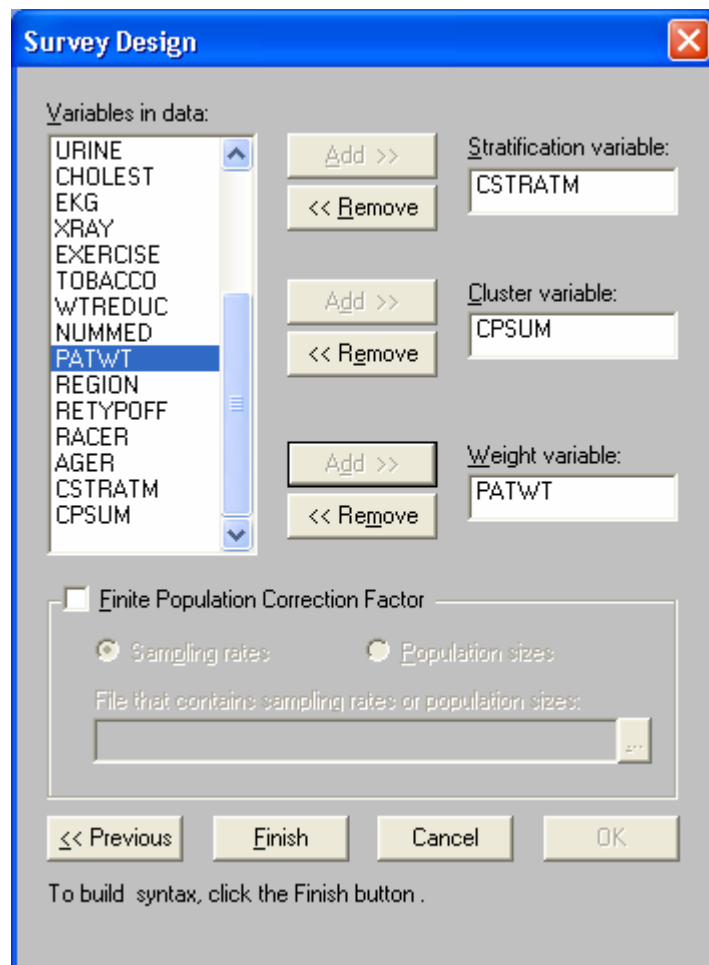
Go ahead and click on the **Next** button to access the **Distributions and Links** dialog box and select the **Multinomial** option from the **Distribution type** drop-down list box to produce the following **Distributions and Links** dialog box.



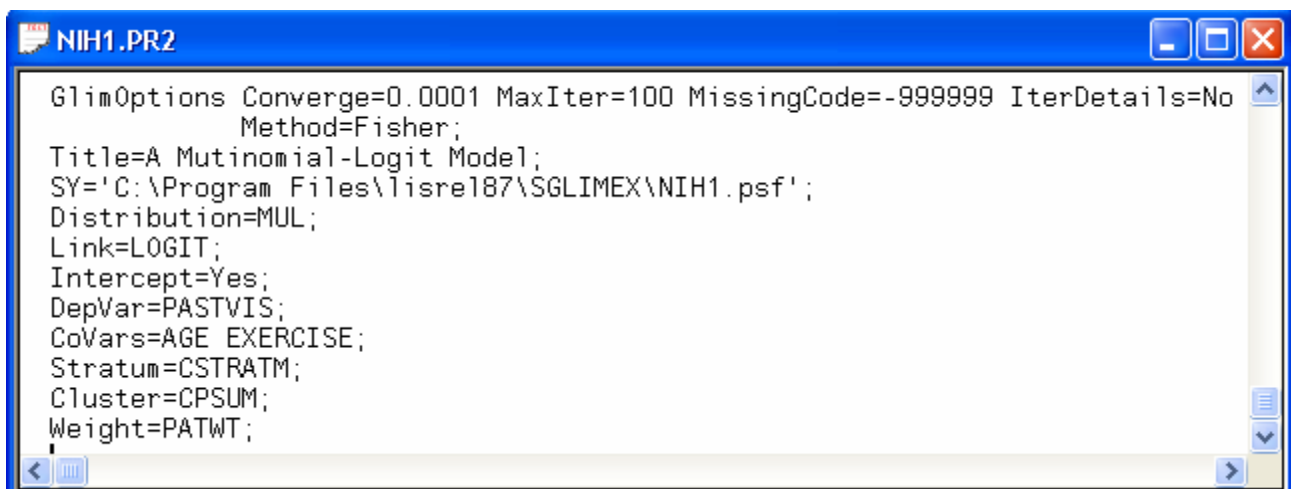
Click on the **Next** button above to go to the **Dependent and Independent Variables** dialog box. Specify the response variable, PASTVIS, by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Dependent variable** section. Specify the covariates, AGE and EXERCISE, by selecting them from the **Variables in data** list box and clicking on the **Continuous** button of the **Independent variables** section to produce the following **Dependent and Independent Variables** dialog box.



Click on the **Next** button to go to the **Survey Design** dialog box. Specify the stratification variable, CSTRATM, by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Stratification variable** section. Specify the cluster variable, CPSUM, and the weight variable, PATWT, by using the **Add** buttons of the **Cluster variable** and the **Weight variable** sections respectively to produce the following **Survey Design** dialog box.



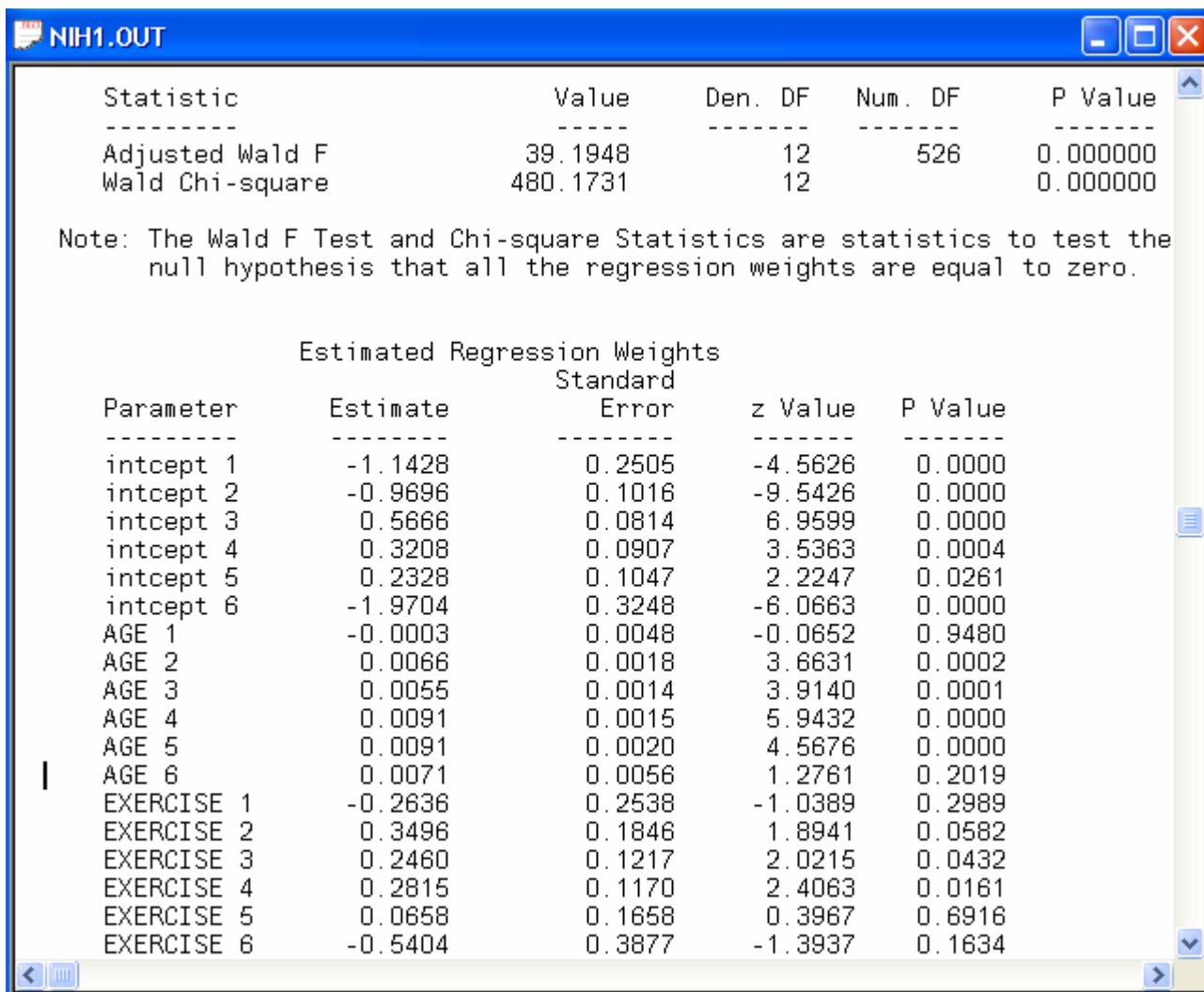
This concludes the specifications, so click on the **Finish** button to open the following text editor window.



Submit the syntax file above by clicking on the **Run Prelis** toolbar icon to obtain the corresponding output file **nih1.out**.

Discussion of results – generalized logistic model

A portion of the results in **nih1.out** is shown in the following text editor window.



Recall that AGE 1 represents the lowest category of the outcome variable, while AGE 6 represents the highest. At a 5% level of significance, the results above suggest that there is sufficient evidence that the age of a respondent exerts a positive influence on the probability of the number of visits to a doctor in the past 12 months by the respondent. In particular, it seems that older respondents are more likely than younger respondents to have visited a doctor more regularly in the past 12 months. The estimated coefficients for the EXERCISE variables are mostly positive, and a value of 1 on any of these indicates a patient that does not exercise. The results thus indicate that exercising exerts a significant influence on the probabilities of 1-2 and 3-5 annual visits to a doctor in the past

12 months. It appears that respondents who do not exercise are more likely than those who do exercise to have visited a doctor regularly in the past 12 months.

The estimated probability that a 60-year old respondent who does not exercise regularly does not visit the doctor (category 2) is obtained from the results above as

$$\hat{P}(\text{PASTVIS}_k = 2) = \frac{\exp(0.97 + 0.007 * 60)}{1 + \sum_{l=1}^6 \exp(\hat{\alpha}_l + \hat{\beta}_{1l} * 60)} = 0.32$$

The corresponding probability that a 60-year old respondent who does exercise regularly does not visit the doctor (category 2) follows as

$$\hat{P}(\text{PASTVIS}_k = 2) = \frac{\exp(0.97 + 0.007 * 60 + 0.35)}{1 + \sum_{l=1}^6 \exp(\hat{\alpha}_l + \hat{\beta}_{1l} * 60 + \hat{\beta}_{2l})} = 0.35$$

The effect estimate for no visit to the doctor, $\hat{\beta}_{22} = 0.35$, suggests that the probability of no visit to the doctor increases for respondents who exercise regularly.

Statistical theory and methods

GLIM framework

The statistical theory and methods for fitting generalized linear models (GLIMs) to complex survey data is essentially an extension of the corresponding theory and methods for simple random sample data (see McCullach & Nelder (1989) and Agresti (2002)). In this section we summarize the general GLIM theory and methods.

We assume that the target population can be stratified into H strata. Within each stratum h , n_h clusters or primary sampling units (PSUs) are drawn and within the h^{th} stratum and i^{th} cluster, n_{hi} ultimate sampling units (USUs) are drawn with design weights w_{hij} , where j denotes the j^{th} USU within the i^{th} cluster, which in turn is nested within stratum h . Furthermore, we assume that the rows of the matrix $\mathbf{Y} = [\mathbf{y}_{hij}]$ represent $n = \sum_{h=1}^H \sum_{i=1}^{n_i} n_{hi}$ observations of the p outcome variables \mathbf{y} with probability density function $f(\cdot)$ and that the rows of the matrix $\mathbf{X} = [\mathbf{x}_{hij}]$ are n observations of the r covariates \mathbf{x} . We postulate a model for the mean vector $\boldsymbol{\mu}_{hij} = E[\mathbf{y}_{hij}]$ which can be expressed as

$$\boldsymbol{\mu}_{hij} = \mathbf{m}(\mathbf{x}_{hij}, \boldsymbol{\theta}) \quad (1)$$

where $\mathbf{m}(\cdot)$ denotes a vector-valued function of \mathbf{x}_{hij} and the $q \times 1$ vector $\boldsymbol{\theta}$ of unknown parameters.

The model in (2) is transformed to a linear model by using a link function which defines the relationship between the elements of the dependent variable vector, $\boldsymbol{\eta}_{hij}$, of the linear model and the elements of the mean vector $\boldsymbol{\mu}_{hij}$. More specifically, the linear model of the GLIM is given by

$$\boldsymbol{\eta}_{hij} = \mathbf{A}_{hij} \boldsymbol{\theta} \quad (3)$$

where \mathbf{A}_{hij} denotes a known $p \times q$ design matrix and

$$\boldsymbol{\eta}_{hij} = \begin{bmatrix} \eta_{hij,1} \\ \eta_{hij,2} \\ \vdots \\ \eta_{hij,p} \end{bmatrix} = \begin{bmatrix} g(\mu_{hij,1}) \\ g(\mu_{hij,2}) \\ \vdots \\ g(\mu_{hij,p}) \end{bmatrix} = \mathbf{g}(\boldsymbol{\mu}_{hij}) \quad (4)$$

where $\mathbf{g}(\cdot) : R \rightarrow R$ denotes the link function.

The log likelihood function for the maximum likelihood estimation of the elements of $\boldsymbol{\theta}$ is given by

$$\ln L(\boldsymbol{\theta} | \mathbf{Y}) = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_j}} w_{h_{ij}} f_{h_{ij}} \ln f(\mathbf{y}_{h_{ij}}; \boldsymbol{\theta}) \quad (5)$$

where $f_{h_{ij}}$ denotes the frequency for observation h_{ij} . From (5), the maximum likelihood equations follows as

$$\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_j}} w_{h_{ij}} f_{h_{ij}} \hat{\mathbf{D}}'_{h_{ij}} \{ \hat{\boldsymbol{\Sigma}}(\mathbf{y}_{h_{ij}}) \}^{-1} [\mathbf{y}_{h_{ij}} - \hat{\boldsymbol{\mu}}_{h_{ij}}] = \mathbf{0} \quad (6)$$

where

$$\mathbf{D}_{h_{ij}} = \left\{ \begin{array}{c} \frac{\partial \boldsymbol{\mu}_{h_{ij}}}{\partial \boldsymbol{\eta}'_{h_{ij}}} \end{array} \right\} \mathbf{A}_{h_{ij}} \quad (7)$$

and $\boldsymbol{\Sigma}(\mathbf{y}_{h_{ij}})$ denotes the covariance matrix of $\mathbf{y}_{h_{ij}}$. In general, the equations in (6) do not have a closed form solution. Consequently, an iterative algorithm is required to obtain maximum likelihood estimates of the elements of $\boldsymbol{\theta}$. The Fisher scoring algorithm may be described as follows. If $\hat{\boldsymbol{\theta}}^{(t)}$ denotes the t^{th} successive approximation to $\hat{\boldsymbol{\theta}}$, then the $(t+1)^{\text{st}}$ approximation is obtained from

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + \left\{ \mathbf{I}_n(\hat{\boldsymbol{\theta}}^{(t)}) \right\}^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}}^{(t)}) \quad (8)$$

where the gradient vector $\mathbf{g}(\cdot)$ is given by

$$\mathbf{g}(\boldsymbol{\theta}) = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_j}} \mathbf{g}_{h_{ij}}(\boldsymbol{\theta}) \quad (9)$$

where

$$\mathbf{g}_{h_{ij}}(\boldsymbol{\theta}) = w_{h_{ij}} f_{h_{ij}} \mathbf{D}'_{h_{ij}} \left\{ \boldsymbol{\Sigma}(\mathbf{y}_{h_{ij}}) \right\}^{-1} [\mathbf{y}_{h_{ij}} - \boldsymbol{\mu}_{h_{ij}}] \quad (10)$$

and the Fisher information matrix is given by

$$\mathbf{I}_n(\boldsymbol{\theta}) = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_i}} w_{h_{ij}} f_{h_{ij}} \mathbf{D}'_{h_{ij}} \left\{ \boldsymbol{\Sigma}(y_{h_{ij}}) \right\}^{-1} \mathbf{D}_{h_{ij}} \quad (11)$$

By using a similar derivation to that in Chapter 2, the approximate asymptotic covariance matrix of the parameter estimators may be expressed as

$$\Upsilon = \left\{ \mathbf{I}_n(\boldsymbol{\theta}) \right\}^{-1} \boldsymbol{\Gamma} \left\{ \mathbf{I}_n(\boldsymbol{\theta}) \right\}^{-1} \quad (12)$$

where $\boldsymbol{\Gamma}$ denotes the covariance matrix of a scalar multiple of the estimated gradient vector.

The application of this general theory to the Poisson-Log model is demonstrated extensively in Section 3. Since this demonstration extends readily to the other specific GLIMs, only the necessary expressions for these GLIMs are provided in the subsequent sections.

The Poisson-log model

Suppose that the elements of the vector $\mathbf{y} = [y_{h_{ij}}]$ represent $n = \sum_{h=1}^H \sum_{i=1}^{n_h} n_{h_i}$ observations of the outcome variable y and that $y_{h_{ij}}$ follows a Poisson distribution with mean $\mu_{h_{ij}}$. In other words, the probability density function of $y_{h_{ij}}$ is given by

$$f(y_{h_{ij}}, \mu_{h_{ij}}) = \frac{e^{-\mu_{h_{ij}}} \mu_{h_{ij}}^{y_{h_{ij}}}}{y_{h_{ij}}!} \Rightarrow \ln f(y_{h_{ij}}, \mu_{h_{ij}}) = y_{h_{ij}} \ln \left\{ \mu_{h_{ij}} \right\} - \mu_{h_{ij}} - \ln \left\{ y_{h_{ij}}! \right\} \quad (13)$$

and the variance of $y_{h_{ij}}$ is given by

$$\sigma^2(y_{h_{ij}}) = \mu_{h_{ij}} \quad (14)$$

Suppose further that the following exponential model is imposed on the means of \mathbf{y}

$$\mu_{h_{ij}} = \exp \left\{ \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \right\} \quad (15)$$

where $\mathbf{x}_{h_{ij}}$ denotes observation h_{ij} of the r covariates \mathbf{x} and the elements of $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_r]'$ denote unknown parameters. The model in (15) is transformed to a linear model by using the log link function. In other words

$$\eta_{h_{ij}} = \ln \left\{ \mu_{h_{ij}} \right\} = \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (16)$$

where $\mathbf{A}_{h_{ij}} = \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$. By using (13), (15) and (16), the log likelihood function for the parameters of the Poisson-log model follows from (5) as

$$\ln L(\boldsymbol{\beta} | \mathbf{y}) = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_{ij}}} w_{h_{ij}} f_{h_{ij}} \left(y_{h_{ij}} \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} - \exp \left\{ \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \right\} - \ln \left\{ y_{h_{ij}} ! \right\} \right). \quad (17)$$

From (17), the gradient vector for the parameters of the Poisson-log model follows as

$$\begin{aligned} \mathbf{g}(\boldsymbol{\beta}) &= \frac{\partial \ln L(\boldsymbol{\beta} | \mathbf{y})}{\partial \boldsymbol{\beta}} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_{ij}}} w_{h_{ij}} f_{h_{ij}} \left[y_{h_{ij}} \frac{\partial \boldsymbol{\beta}' \mathbf{x}_{h_{ij}}}{\partial \boldsymbol{\beta}} - \exp \left\{ \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \right\} \frac{\partial \left(\boldsymbol{\beta}' \mathbf{x}_{h_{ij}} \right)}{\partial \boldsymbol{\beta}} \right] \\ &= \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_{ij}}} w_{h_{ij}} f_{h_{ij}} \left[y_{h_{ij}} - \exp \left\{ \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \right\} \right] \mathbf{x}_{h_{ij}} \\ &= \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_{ij}}} w_{h_{ij}} f_{h_{ij}} \mathbf{D}'_{h_{ij}} \left\{ \sigma^2 \left(y_{h_{ij}} \right) \right\}^{-1} \left[y_{h_{ij}} - \mu_{h_{ij}} \right] \end{aligned} \quad (18)$$

where $\mathbf{D}_{h_{ij}} = \mu_{h_{ij}} \mathbf{x}_{h_{ij}}$. The Fisher information matrix for the parameters of the Poisson-log model follows as

$$\begin{aligned} \mathbf{I}_n(\boldsymbol{\beta}) &= -E \left[\frac{\partial^2 \ln L(\boldsymbol{\beta} | \mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right] = -E \left[- \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_{ij}}} w_{h_{ij}} f_{h_{ij}} \mathbf{x}_{h_{ij}} \exp \left\{ \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \right\} \frac{\partial \left(\mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \right)}{\partial \boldsymbol{\beta}'} \right] \\ &= \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_{ij}}} w_{h_{ij}} f_{h_{ij}} \exp \left\{ \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \right\} \mathbf{x}_{h_{ij}} \mathbf{x}'_{h_{ij}} \\ &= \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_{ij}}} w_{h_{ij}} f_{h_{ij}} \mathbf{D}'_{h_{ij}} \left\{ \sigma^2 \left(y_{h_{ij}} \right) \right\}^{-1} \mathbf{D}_{h_{ij}} \end{aligned} \quad (19)$$

It is evident that expressions (18) and (19) are equivalent to the general expressions (9) and (11) respectively. Since these derivations are similar for the other GLIMs, we provide the specific expressions for each individual GLIM without derivation.

Models for the Bernoulli sampling distribution

Sampling distribution

$$f(y_{h_{ij}}) = \mu_{h_{ij}}^{y_{h_{ij}}} (1 - \mu_{h_{ij}})^{1 - y_{h_{ij}}} \quad (20)$$

Variance

$$\sigma^2(y_{h_{ij}}) = \mu_{h_{ij}} (1 - \mu_{h_{ij}}) \quad (21)$$

The logit model

Model for means

$$\mu_{h_{ij},1} = \frac{\exp\{\mathbf{x}'_{h_{ij}} \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_{h_{ij}} \boldsymbol{\beta}\}} \quad (22)$$

where the elements of $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij}} = \text{logit}(\mu_{h_{ij},1}) = \ln \left\{ \frac{\mu_{h_{ij},1}}{\mu_{h_{ij},2}} \right\} \quad (23)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (24)$$

where $\mathbf{A}_{h_{ij}} = \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \mu_{h_{ij},1} (1 - \mu_{h_{ij},1}) \mathbf{x}'_{h_{ij}} \quad (25)$$

The complementary log-log model

Model for means

$$\mu_{h_{ij}} = 1 - \exp\left\{-\exp\left\{\mathbf{x}'_{h_{ij}}\boldsymbol{\beta}\right\}\right\} \quad (26)$$

where the elements of $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij}} = \ln(-\ln(1 - \mu_{h_{ij}})) \quad (27)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}}\boldsymbol{\theta} \quad (28)$$

where $\mathbf{A}_{h_{ij}} = \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = (1 - \mu_{h_{ij}})\ln(1 - \mu_{h_{ij}})\mathbf{x}'_{h_{ij}} \quad (29)$$

The probit model

Model for means

$$\mu_{h_{ij}} = \Phi(\mathbf{x}'_{h_{ij}}\boldsymbol{\beta}) \quad (30)$$

where the elements of $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_r]'$ denote unknown parameters and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Normal distribution.

Link function

$$\eta_{h_{ij}} = \Phi^{-1}(\mu_{h_{ij}}) \quad (31)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \tag{32}$$

where $\mathbf{A}_{h_{ij}} = \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\Phi^{-1}(\mu_{h_{ij}}))^2} \mathbf{x}'_{h_{ij}} \tag{33}$$

The log model

Model for means

$$\mu_{h_{ij}} = \exp\{\mathbf{x}'_{h_{ij}} \boldsymbol{\beta}\} \tag{34}$$

where the elements of $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij}} = \ln(\mu_{h_{ij}}) \tag{35}$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \tag{36}$$

where $\mathbf{A}_{h_{ij}} = \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \mu_{h_{ij}} \mathbf{x}'_{h_{ij}} \tag{37}$$

Models for the Multinomial sampling distribution

Sampling distribution

$$f(y_{h_{ij}}) = \frac{n!}{\left(\prod_{k=1}^{p-1} y_{h_{ij},k}!\right) \left(n - \sum_{k=1}^{p-1} y_{h_{ij},k}\right)!} \left(\prod_{k=1}^{p-1} \mu_{h_{ij},k}^{y_{h_{ij},k}}\right) \mu_{h_{ij},p}^{n - \sum_{k=1}^{p-1} y_{h_{ij},k}} \quad (38)$$

Covariance matrix

$$\Sigma(\mathbf{y}_{h_{ij}}^*) = \mathbf{D}_{\mu_{h_{ij}}} - \boldsymbol{\mu}_{h_{ij}} \boldsymbol{\mu}_{h_{ij}}' \quad (39)$$

where $\mathbf{y}_{h_{ij}}^* = [y_{h_{ij},1} \ y_{h_{ij},2} \ \cdots \ y_{h_{ij},p-1}]'$ and $\mathbf{D}_{\mu_{h_{ij}}}$ denotes a $(p-1) \times (p-1)$ diagonal matrix with the elements of $\boldsymbol{\mu}_{h_{ij}} = [\mu_{h_{ij},1} \ \mu_{h_{ij},2} \ \cdots \ \mu_{h_{ij},p-1}]'$ on the diagonal.

The generalized logistic Model

Model for means

$$\mu_{h_{ij},k} = \frac{\exp\{\mathbf{x}'_{h_{ij}} \boldsymbol{\beta}_k\}}{1 + \sum_{k=1}^{p-1} \exp\{\mathbf{x}'_{h_{ij}} \boldsymbol{\beta}_k\}} \quad \forall \ k = 1, 2, \dots, p-1 \quad (40)$$

where the elements of $\boldsymbol{\beta}_k = [\beta_{k1} \ \beta_{k2} \ \dots \ \beta_{kr}]'$ $\forall \ k = 1, 2, \dots, p-1$ denote unknown parameters.

Link function

$$\eta_{h_{ij},k} = \text{logit}(\mu_{h_{ij},k}) = \ln \left\{ \frac{\mu_{h_{ij},k}}{\mu_{h_{ij},p}} \right\} \quad (41)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (42)$$

where $\boldsymbol{\eta}_{h_{ij}} = [\eta_{h_{ij},1} \ \eta_{h_{ij},2} \ \cdots \ \eta_{h_{ij},p-1}]'$, $\mathbf{A}_{h_{ij}} = \mathbf{I}_{p-1} \otimes \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = [\mathbf{D}_{\mu_{h_{ij}}} - \boldsymbol{\mu}_{h_{ij}} \boldsymbol{\mu}'_{h_{ij}}] \otimes \mathbf{x}'_{h_{ij}} \quad (43)$$

where $\mathbf{D}_{\mu_{h_{ij}}}$ denotes a $(p-1) \times (p-1)$ diagonal matrix with the elements of

$\boldsymbol{\mu}_{h_{ij}} = [\mu_{h_{ij},1} \ \mu_{h_{ij},2} \ \cdots \ \mu_{h_{ij},p-1}]'$ on the diagonal.

The cumulative logit model

Model for means

$$\mu_{h_{ij},k}^* = \sum_{l=1}^k \mu_{h_{ij},l} = \frac{\exp\{\alpha_k + \mathbf{x}'_{h_{ij}} \boldsymbol{\beta}_k\}}{1 + \exp\{\alpha_k + \mathbf{x}'_{h_{ij}} \boldsymbol{\beta}_k\}} \quad \forall \ k = 1, 2, \dots, p-1 \quad (44)$$

where the elements of $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_{p-1}]'$ and $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij},k} = \text{clogit}(\mu_{h_{ij},k}^*) = \ln \left\{ \frac{\mu_{h_{ij},k}^*}{1 - \mu_{h_{ij},k}^*} \right\} \quad (45)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (46)$$

where $\boldsymbol{\eta}_{h_{ij}} = [\eta_{h_{ij},1} \ \eta_{h_{ij},2} \ \cdots \ \eta_{h_{ij},p-1}]'$, $\mathbf{A}_{h_{ij}} = [\mathbf{I}_{p-1} \ \mathbf{1}_{p-1} \otimes \mathbf{x}'_{h_{ij}}]$ and $\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \begin{bmatrix} \mathbf{U}\mathbf{d}_{\mu_{h_{ij}}^*}^* & \mathbf{U}\mathbf{d}_{\mu_{h_{ij}}^*}^* \otimes \mathbf{x}'_{h_{ij}} \end{bmatrix} \quad (47)$$

where \mathbf{U} denotes a $(p-1) \times (p-1)$ matrix given by

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \quad (48)$$

$$\mathbf{d}_{\mu_{h_{ij}}^*}^* = \left[\mu_{h_{ij},1}^* (1 - \mu_{h_{ij},1}^*) \quad \mu_{h_{ij},2}^* (1 - \mu_{h_{ij},2}^*) \cdots \mu_{h_{ij},p-1}^* (1 - \mu_{h_{ij},p-1}^*) \right]' \quad (49)$$

The proportional hazards model

Model for means

$$\mu_{h_{ij},k}^* = \sum_{l=1}^k \mu_{h_{ij},l} = 1 - \exp\left(-\exp\left\{\alpha_k + \mathbf{x}'_{h_{ij}} \boldsymbol{\beta}_k\right\}\right) \quad \forall k = 1, 2, \dots, p-1 \quad (50)$$

where the elements of $\boldsymbol{\alpha} = [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_{p-1}]'$ and $\boldsymbol{\beta} = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij},k} = \text{cloglog}(\mu_{h_{ij},k}^*) = \ln\left(-\ln\left(1 - \mu_{h_{ij},k}^*\right)\right) \quad (51)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (52)$$

where $\boldsymbol{\eta}_{h_{ij}} = [\eta_{h_{ij},1} \ \eta_{h_{ij},2} \ \cdots \ \eta_{h_{ij},p-1}]'$ $\mathbf{A}_{h_{ij}} = [\mathbf{I}_{p-1} \ \mathbf{1}_{p-1} \otimes \mathbf{x}'_{h_{ij}}]$ and $\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \begin{bmatrix} \mathbf{U} \mathbf{d}_{\mu_{h_{ij}}^*}^* & \mathbf{U} \mathbf{d}_{\mu_{h_{ij}}^*}^* \otimes \mathbf{x}'_{h_{ij}} \end{bmatrix} \quad (53)$$

where \mathbf{U} is as defined in (48) and

$$\mathbf{d}_{\mu_{h_{ij}}^*}^* = \begin{bmatrix} -(1 - \mu_{h_{ij},1}^*) \ln(1 - \mu_{h_{ij},1}^*) \\ -(1 - \mu_{h_{ij},2}^*) \ln(1 - \mu_{h_{ij},2}^*) \\ \vdots \\ -(1 - \mu_{h_{ij},p-1}^*) \ln(1 - \mu_{h_{ij},p-1}^*) \end{bmatrix} \quad (54)$$

The cumulative probit model

Model for means

$$\mu_{h_{ij},k}^* = \sum_{l=1}^k \mu_{h_{ij},l} = \Phi(\alpha_k + \mathbf{x}'_{h_{ij}} \boldsymbol{\beta}) \quad \forall \ k = 1, 2, \dots, p-1 \quad (55)$$

where the elements of $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_{p-1}]'$ and $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_r]'$ denote unknown parameters and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Link function

$$\eta_{h_{ij},k} = \Phi^{-1}(\mu_{h_{ij},k}^*) \quad (56)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (57)$$

where $\boldsymbol{\eta}_{h_{ij}} = [\eta_{h_{ij},1} \ \eta_{h_{ij},2} \ \cdots \ \eta_{h_{ij},p-1}]'$, $\mathbf{A}_{h_{ij}} = [\mathbf{I}_{p-1} \ \mathbf{1}_{p-1} \otimes \mathbf{x}'_{h_{ij}}]$ and $\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \begin{bmatrix} \mathbf{U}\mathbf{D}_{\eta_{h_{ij}}}^* & \mathbf{U}\mathbf{d}_{\eta_{h_{ij}}}^* \otimes \mathbf{x}'_{h_{ij}} \end{bmatrix} \quad (58)$$

where $\mathbf{D}_{\eta_{h_{ij}}}^*$ denotes a $(p-1) \times (p-1)$ diagonal matrix with diagonal elements given by

$$[\mathbf{D}_{\eta_{h_{ij}}}^*]_{kk} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta_{h_{ij},k}^2} \quad \forall \ k = 1, 2, \dots, p-1, \ \mathbf{U} \text{ is as defined in (48) and}$$

$$\mathbf{d}_{\eta_{h_{ij}}}^* = \begin{bmatrix} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta_{h_{ij},1}^2} \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta_{h_{ij},2}^2} \\ \vdots \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta_{h_{ij},p-1}^2} \end{bmatrix} \quad (59)$$

The log model

Model for means

$$\mu_{h_{ij},k} = \exp(\alpha_k + \mathbf{x}'_{h_{ij}} \boldsymbol{\beta}) \quad \forall \ k = 1, 2, \dots, p-1 \quad (60)$$

where the elements of $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_{p-1}]'$ and $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij},k} = \ln(\mu_{h_{ij},k}) \quad (61)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \tag{62}$$

where $\boldsymbol{\eta}_{h_{ij}} = [\eta_{h_{ij},1} \ \eta_{h_{ij},2} \ \cdots \ \eta_{h_{ij},p-1}]'$, $\mathbf{A}_{h_{ij}} = [\mathbf{I}_{p-1} \ \mathbf{1}_{p-1} \otimes \mathbf{x}'_{h_{ij}}]$ and $\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \begin{bmatrix} \mathbf{D}_{\mu_{h_{ij}}} & \mathbf{d}_{\mu_{h_{ij}}} \otimes \mathbf{x}'_{h_{ij}} \end{bmatrix} \tag{63}$$

where $\mathbf{D}_{\mu_{h_{ij}}}$ denotes a $(p-1) \times (p-1)$ diagonal matrix with the elements of

$\mathbf{d}_{\mu_{h_{ij}}} = [\mu_{h_{ij},1} \ \mu_{h_{ij},2} \ \cdots \ \mu_{h_{ij},p-1}]'$ on the diagonal.

The probit model

Model for means

$$\mu_{h_{ij},k} = \Phi(\alpha_k + \mathbf{x}'_{h_{ij}} \boldsymbol{\beta}) \quad \forall \ k = 1, 2, \dots, p-1 \tag{64}$$

where the elements of $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_{p-1}]'$ and $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_r]'$ denote unknown parameters and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Link function

$$\eta_{h_{ij},k} = \Phi^{-1}(\mu_{h_{ij},k}) \tag{65}$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \tag{66}$$

where $\boldsymbol{\eta}_{h_{ij}} = [\eta_{h_{ij},1} \ \eta_{h_{ij},2} \ \cdots \ \eta_{h_{ij},p-1}]'$, $\mathbf{A}_{h_{ij}} = [\mathbf{I}_{p-1} \ \mathbf{1}_{p-1} \otimes \mathbf{x}'_{h_{ij}}]$ and $\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \left[\mathbf{D}_{\mu_{h_{ij}}} \ \mathbf{d}_{\mu_{h_{ij}}} \otimes \mathbf{x}'_{h_{ij}} \right] \quad (67)$$

where

$$\mathbf{d}_{\mu_{h_{ij}}} = \begin{bmatrix} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\Phi^{-1}(\mu_{h_{ij},1}))^2} \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\Phi^{-1}(\mu_{h_{ij},2}))^2} \\ \vdots \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\Phi^{-1}(\mu_{h_{ij},p-1}))^2} \end{bmatrix} \quad (68)$$

The complementary log-log model

Model for means

$$\mu_{h_{ij},k} = 1 - \exp\left\{-\exp\left\{\alpha_k + \mathbf{x}'_{h_{ij},k} \boldsymbol{\beta}\right\}\right\} \quad \forall \ k = 1, 2, \dots, p-1 \quad (69)$$

where the elements of $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_{p-1}]'$ and $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij},k} = \ln\left(-\ln\left(1 - \mu_{h_{ij},k}\right)\right) \quad (70)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (71)$$

where $\boldsymbol{\eta}_{h_{ij}} = [\eta_{h_{ij},1} \ \eta_{h_{ij},2} \ \cdots \ \eta_{h_{ij},p-1}]'$, $\mathbf{A}_{h_{ij}} = [\mathbf{I}_{p-1} \ \mathbf{1}_{p-1} \otimes \mathbf{x}'_{h_{ij}}]$ and $\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$.

The **D** matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \left[\mathbf{D}_{\mu_{h_{ij}}} \quad \mathbf{d}_{\mu_{h_{ij}}} \otimes \mathbf{x}'_{h_{ij}} \right] \quad (72)$$

where

$$\mathbf{d}_{\mu_{h_{ij}}} = \begin{bmatrix} -(1 - \mu_{h_{ij},1}) \ln(1 - \mu_{h_{ij},1}) \\ -(1 - \mu_{h_{ij},2}) \ln(1 - \mu_{h_{ij},2}) \\ \vdots \\ -(1 - \mu_{h_{ij},p-1}) \ln(1 - \mu_{h_{ij},p-1}) \end{bmatrix} \quad (73)$$

and $\mathbf{D}_{\mu_{h_{ij}}}$ denotes a $(p-1) \times (p-1)$ diagonal matrix with the elements of $\mathbf{d}_{\mu_{h_{ij}}}$ on the diagonal.

Models for the Binomial sampling distribution

Sampling distribution

$$f(y_{h_{ij}}) = \binom{n_{h_{ij}}}{y_{h_{ij}}} \mu_{h_{ij}}^{y_{h_{ij}}} (1 - \mu_{h_{ij}})^{n_{h_{ij}} - y_{h_{ij}}} \quad (74)$$

Variance

$$\sigma^2(y_{h_{ij}}) = \frac{\mu_{h_{ij}} (1 - \mu_{h_{ij}})}{n_{h_{ij}}} \quad (75)$$

The logit model

The model for means, the link function, the linear model and the **D** matrix of the Binomial-logit model are identical to those of the Bernoulli-logit model (*cf.* (22), (23), (24) and (25)).

The complementary log-log model

The model for means, the link function, the linear model and the **D** matrix of the Binomial-complementary log-log model are identical to those of the Bernoulli-complementary log-log model (*cf.* (26), (27), (28) and (29)).

The probit model

The model for means, the link function, the linear model and the **D** matrix of the Binomial-probit model are identical to those of the Bernoulli-probit model (*cf.* (30), (31), (32) and (33)).

Models for the Gamma distribution

Sampling distribution

$$f(y_{h_{ij}}) = \frac{1}{\Gamma\left(\frac{1}{\psi}\right) y_{h_{ij}}^{\frac{1}{\psi}}} \left(\frac{y_{h_{ij}}}{\mu_{h_{ij}} \psi}\right)^{\frac{1}{\psi}} \exp\left(-\frac{y_{h_{ij}}}{\mu_{h_{ij}} \psi}\right) \quad (76)$$

Variance

$$\sigma^2(y_{h_{ij}}) = \psi \mu_{h_{ij}}^2 \quad (77)$$

The log model

The model for means, the link function, the linear model and the **D** matrix of the Gamma-log model are identical to those of the Bernoulli-log model (*cf.* (34), (35), (36) and (37)).

The power model

Model for means

$$\mu_{h_{ij}} = (\mathbf{x}'_{h_{ij}} \boldsymbol{\beta})^k \quad (78)$$

where the elements of $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_r]'$ denote unknown parameters and k denotes an integer.

Link function

$$\eta_{h_{ij}} = \mu_{h_{ij}}^{1/k} \quad (79)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \quad (80)$$

where $\mathbf{A}_{h_{ij}} = \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$.

The **D** matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \frac{1}{k} \mu_{h_{ij}}^{1/k-1} \mathbf{x}'_{h_{ij}} \quad (81)$$

Models for the Inverse Gaussian distribution

Sampling distribution

$$f(y_{h_{ij}}) = \frac{1}{\sqrt{2\pi y_{h_{ij}}^3 \psi}} \exp\left(-\frac{1}{2y_{h_{ij}}} \left(\frac{y_{h_{ij}} - \mu_{h_{ij}}}{\mu_{h_{ij}}}\right)^2 / \psi\right) \quad (82)$$

Variance

$$\sigma^2(y_{h_{ij}}) = \psi \mu_{h_{ij}}^3 \quad (83)$$

The log model

The model for means, the link function, the linear model and the **D** matrix of the Inverse Gaussian-log model are identical to those of the Bernoulli-log model (*cf.* (34), (35), (36) and (37)).

The power model

The model for means, the link function, the linear model and the **D** matrix of the Inverse Gaussian-power model are identical to those of the Gamma-power model (*cf.* (78), (79), (80) and (81)).

The Negative Binomial-log model

Sampling distribution

$$f(y_{h_{ij}}) = \frac{\Gamma\left(y_{h_{ij}} + \frac{1}{\psi}\right)}{\Gamma(y_{h_{ij}} + 1)\Gamma\left(\frac{1}{\psi}\right)} \frac{(\psi\mu_{h_{ij}})^{y_{h_{ij}}}}{\left(1 + \psi\mu_{h_{ij}}\right)^{y_{h_{ij}} + \frac{1}{\psi}}} \quad (84)$$

Variance

$$\sigma^2(y_{h_{ij}}) = \mu_{h_{ij}} + \psi\mu_{h_{ij}}^2 \quad (85)$$

The model for means, the link function, the linear model and the **D** matrix of the Negative Binomial-log model are identical to those of the Bernoulli-log model (cf.(34), (35), (36) and (37)).

The Normal-identity model

Sampling distribution

$$f(y_{h_{ij}}) = \frac{1}{\sqrt{2\pi\psi}} \exp\left(-\frac{1}{2\psi}(y_{h_{ij}} - \mu_{h_{ij}})^2\right) \quad (86)$$

Variance

$$\sigma^2(y_{h_{ij}}) = \psi \quad (87)$$

Model for means

$$\mu_{h_{ij}} = \mathbf{x}'_{h_{ij}} \boldsymbol{\beta} \quad (88)$$

where the elements of $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_r]'$ denote unknown parameters.

Link function

$$\eta_{h_{ij}} = \mu_{h_{ij}} \quad (89)$$

Linear model

$$\boldsymbol{\eta}_{h_{ij}} = \mathbf{A}_{h_{ij}} \boldsymbol{\theta} \tag{90}$$

where $\mathbf{A}_{h_{ij}} = \mathbf{x}'_{h_{ij}}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$.

The D matrix for observation h_{ij}

$$\mathbf{D}_{h_{ij}} = \mu_{h_{ij}} \mathbf{x}'_{h_{ij}} \tag{91}$$

The estimation of scale and dispersion parameters

A number of sampling distributions discussed in the previous sections have a dispersion parameter and/or a scale parameter. A summary of these distributions with respect to dispersion and scale parameters and their estimates is shown in Table 3.

Table 3: Scale and dispersion parameters

Distribution	Deviance	Dispersion	Maximum Likelihood	Pearson	Scale
Binomial	x			x	x
Gamma	x	x	x	x	x
Inverse Gaussian	x	x	x	x	x
Negative binomial	x	x	x	x	
Normal	x	x	x	x	x
Poisson	x			x	x

The deviance χ^2 estimate

$$\hat{\phi}_D = \sqrt{\frac{X_D^2}{d}} \tag{92}$$

$$X_D^2 = 2 \ln L(\mathbf{y} | \mathbf{y}) - 2 \ln L(\hat{\boldsymbol{\mu}} | \mathbf{y}) \tag{93}$$

$$d = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_i}} w_{h_{ij}} f_{h_{ij}} - q \quad (94)$$

The Pearson χ^2 estimate

$$\hat{\phi}_P = \sqrt{\frac{X_P^2}{d}} \quad (95)$$

$$X_P^2 = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_i}} \frac{w_{h_{ij}} f_{h_{ij}} (y_{h_{ij}} - \hat{\mu}_{h_{ij}})^2}{\hat{\sigma}^2(y_{h_{ij}})} \quad (96)$$

The maximum likelihood estimate

The Maximum Likelihood estimate, $\hat{\psi}_{ML}$, of ψ is obtained by using a three-stage estimation procedure. In stage one, the Fisher Scoring algorithm is used to obtain Maximum Likelihood estimates of the elements of θ . These estimates are then used as fixed values for the elements of θ in a Newton-Raphson algorithm to obtain an estimate, $\hat{\psi}_{ML}$, of the dispersion parameter ψ . In this algorithm, the method of moments estimate of ψ is used as the starting values for $\hat{\psi}_{ML}$. In stage three of the procedure, the Fisher Scoring algorithm is extended to include the dispersion parameter to yield Maximum Likelihood estimates of the dispersion parameter and the elements of θ . This three-stage procedure is used in the case of the Negative Binomial, Inverse Gaussian and Gamma sampling distributions. In the case of the Gamma and Inverse Gaussian sampling distributions, the Maximum Likelihood estimate of the scale parameter, $\hat{\phi}_{ML}$, is computed from $\hat{\psi}_{ML}$ and the Delta method (Bishop, Feinberg & Holland 1988) is used to compute the corresponding standard error estimate. In the case of the Normal sampling distribution, $\hat{\psi}_{ML}$ is computed as

$$\hat{\psi}_{ML} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_i}} w_{h_{ij}} f_{h_{ij}} (y_{h_{ij}} - \hat{\mu}_{h_{ij}})^2}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{n_{h_i}} w_{h_{ij}} f_{h_{ij}}} \quad (97)$$

Corrections to standard error estimates

The standard error estimates are multiplied with the scale parameter estimate to correct them with respect to scale.

References

- Agresti, A. (2002). *Categorical data analysis, second edition*. New York: Wiley.
- American Institutes for Research & Cohen, J. (2003). *AM version 0.06.02 beta April 15, 2004 [Computer software]*. Washington, DC: National Center for Education Statistics of the U.S. Department of Education.
- Binder, D.A. (1983). On the variances of asymptotically Normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292.
- Bishop, Y.M.M., Feinberg, S.E. & Holland, P.W. (1988). *Discrete multivariate analysis: theory and practice*. Cambridge: MIT Press.
- Brogan, D.J. (1998). Pitfalls of using standard statistical software packages for sample survey data. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*. New York: John Wiley and Sons.
- Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhya, Series C*, **37**, 117-132.
- Jöreskog, K.G. & Sörbom, D. (2005). *LISREL for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International, Inc.
- McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. Chapman & Hall: London.
- Morel, G. (1989). Logistic regression under complex survey designs. *Survey methodology*, **15**, 203-223.
- SAS Institute, Inc. (2004). *SAS/STAT®: user's guide*. Cary, NC: SAS Institute, Inc.