GLIMs for binary responses

Contents

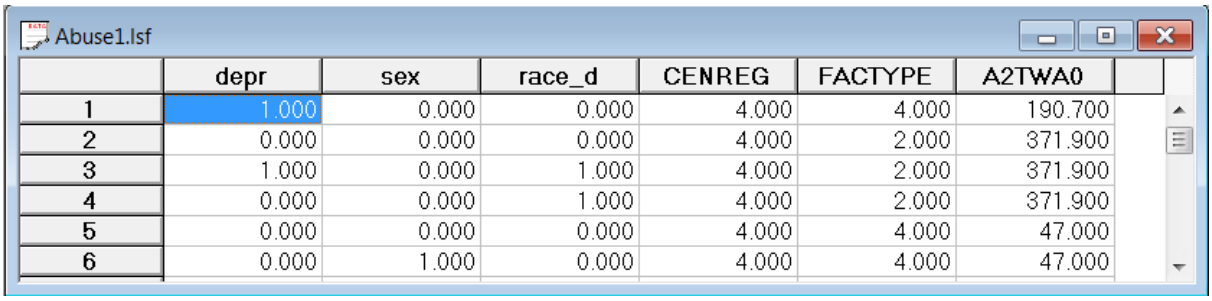
1. Introduction

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.

2. The data

The data set forms part of the data library of the Alcohol and Drug Services Study. The data set to be analyzed consists of the complete cases for a selection of variables and is provided as the LSF **abuse1.lsf** in the **Generalized Linear Modeling examples** folder. The first portion of this data set is shown in the following LSF window.



Abuse1.lsf

	depr	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.000	0.000	0.000	4.000	4.000	190.700
2	0.000	0.000	0.000	4.000	2.000	371.900
3	1.000	0.000	1.000	4.000	2.000	371.900
4	0.000	0.000	1.000	4.000	2.000	371.900
5	0.000	0.000	0.000	4.000	4.000	47.000
6	0.000	1.000	0.000	4.000	4.000	47.000

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.

3. The models

The sampling distributions

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

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

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

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

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.

The probability models

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

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

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 ($j = 1, 2, \ldots, r$) 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

$$\mathrm{P}\left(\mathrm{depr}_k = 1\right) = \frac{\exp\left(\alpha + \beta_1 * \mathrm{sex}_k + \beta_2 * \mathrm{race\_d}_k\right)}{1 + \exp\left(\alpha + \beta_1 * \mathrm{sex}_k + \beta_2 * \mathrm{race\_d}_k\right)}$$

where $\mathrm{P}\left(\mathrm{depr}_k = 1\right)$ 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 ($\mathrm{sex}_k = 1$) and a male client ($\mathrm{sex}_k = 0$) has a history of depression respectively follows as

$$\frac{\exp\left(\alpha + \beta_1 + \beta_2 * \mathrm{race\_d}\right)}{1 + \exp\left(\alpha + \beta_2 * \mathrm{race\_d}\right)} = \exp\left(\beta_1\right)$$

In a similar fashion, it follows that $\exp\left(\beta_2\right)$ 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

$$\hat{\mathrm{P}}\left(\mathrm{depr}_k = 1\right) = \frac{\exp\left(\hat{\alpha} + \hat{\beta}_1 * \mathrm{sex}_k + \hat{\beta}_2 * \mathrm{race\_d}_k\right)}{1 + \exp\left(\hat{\alpha} + \hat{\beta}_1 * \mathrm{sex}_k + \hat{\beta}_2 * \mathrm{race\_d}_k\right)}$$

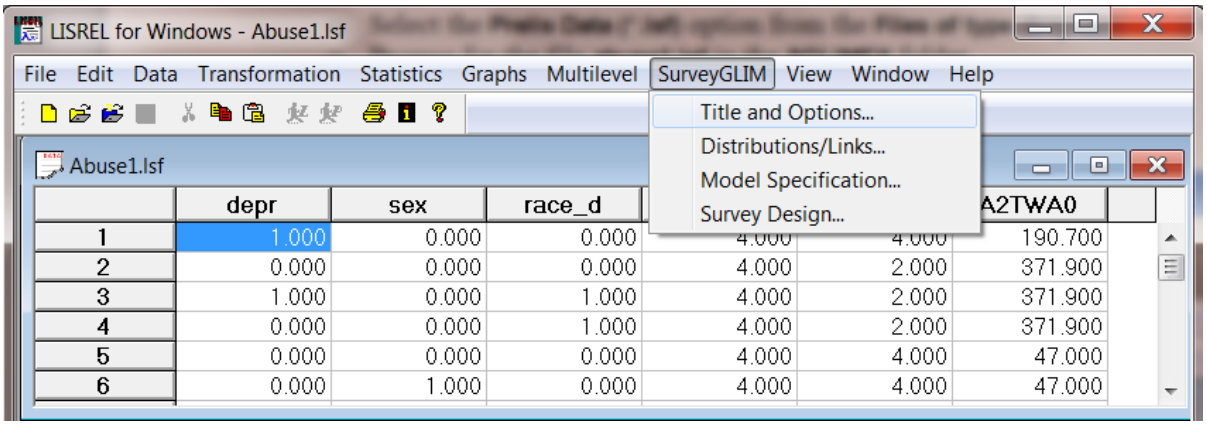
where $\hat{\mathrm{P}}\left(\mathrm{depr}_k = 1\right)$ denotes the estimated probability that client k has a history of depression and $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the maximum likelihood estimates of α, β_1 and β_2 respectively.

4. Analyzing binary outcomes from complex survey designs (method 1)

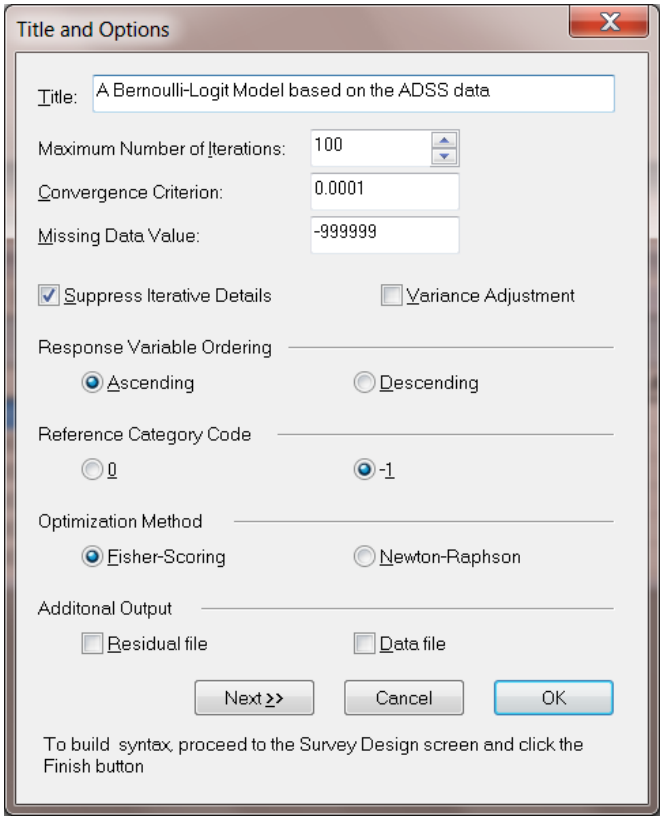
To explore a potential link between depression and a respondent's gender and ethnicity, a GLIM with Bernoulli distribution and logit link function is fitted to the data described above. The Bernoulli distribution is used since the outcome variable, depr, is dichotomous (0 for no depression history and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

Setting up the analysis

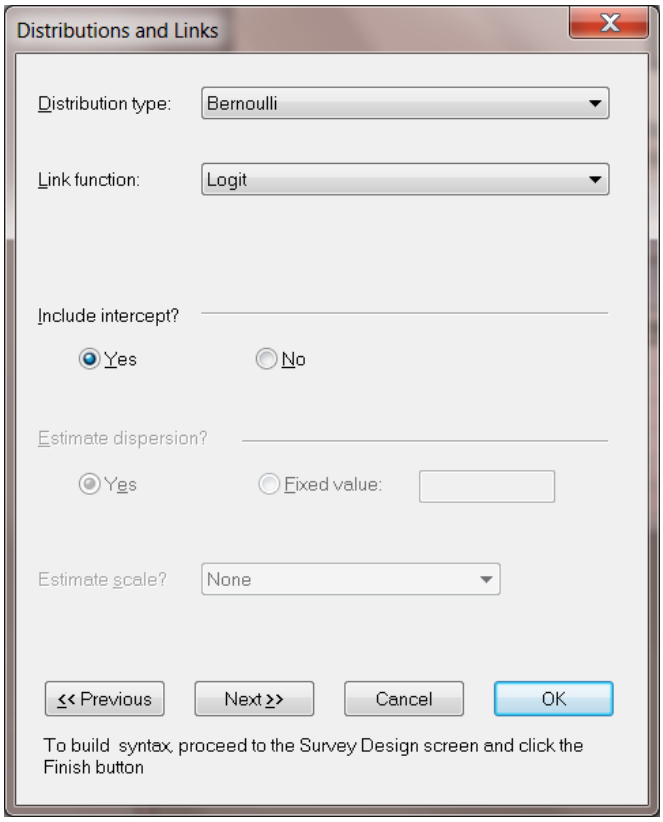
We first open the file **abuse1.lsf** in a LSF window using the following steps. Use the **Open** option on the **File** menu of the root window of LISREL to load the **Open** dialog box. Select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box. Browse for and open the file **abuse1.lsf**.



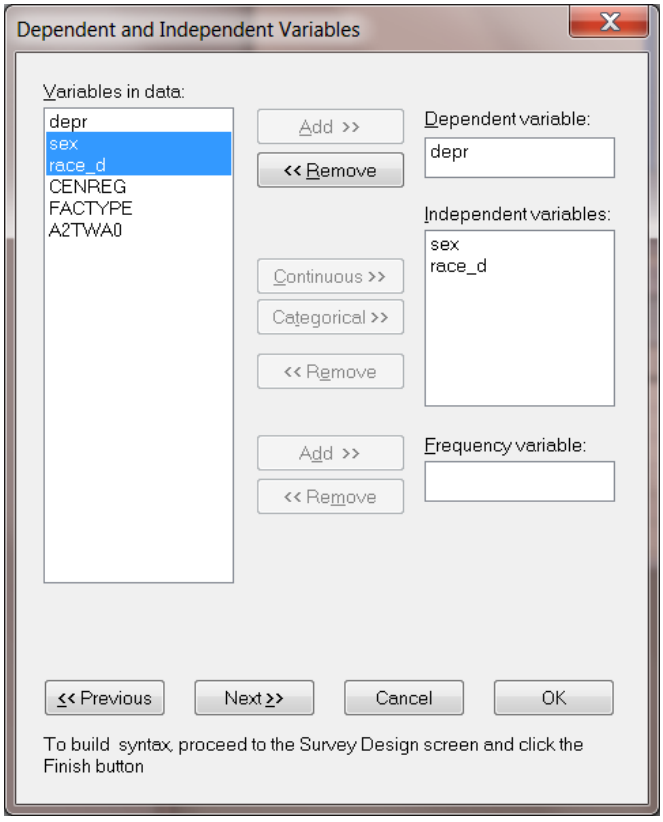
We can now use the **SurveyGLIM** menu to fit the Bernoulli-logit GLIM to the data in **abuse1.lsf**. First, select the **Title and Options** option on the **SurveyGLIM** menu to go to the **Title and Options** dialog box. Enter the title **A Bernoulli-Logit Model for ADSS Data** into the **Title** string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



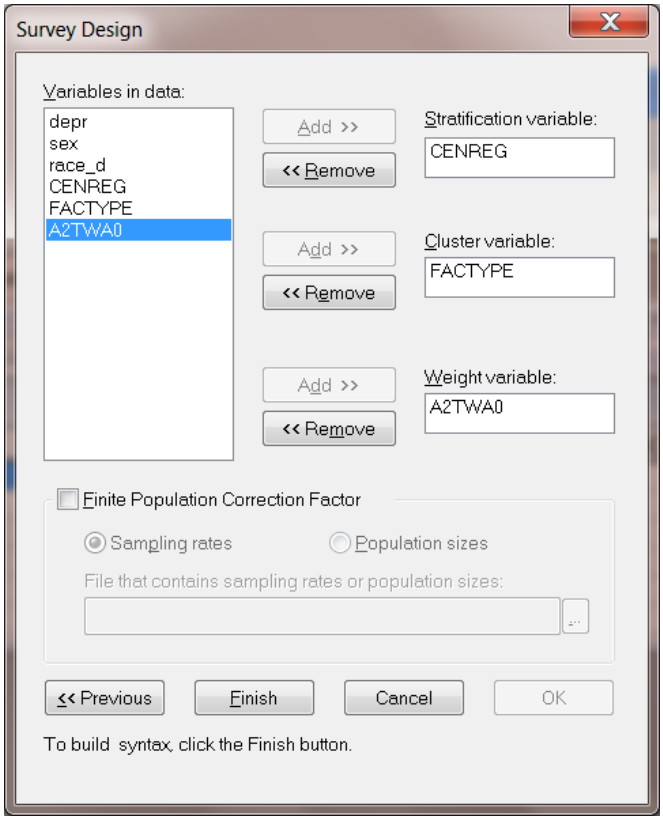
Click on the **Next** button to access the **Distributions and Links** dialog box and select the **Bernoulli** option from the **Distribution type** drop-down list box to produce the following **Distributions and Links** dialog box.



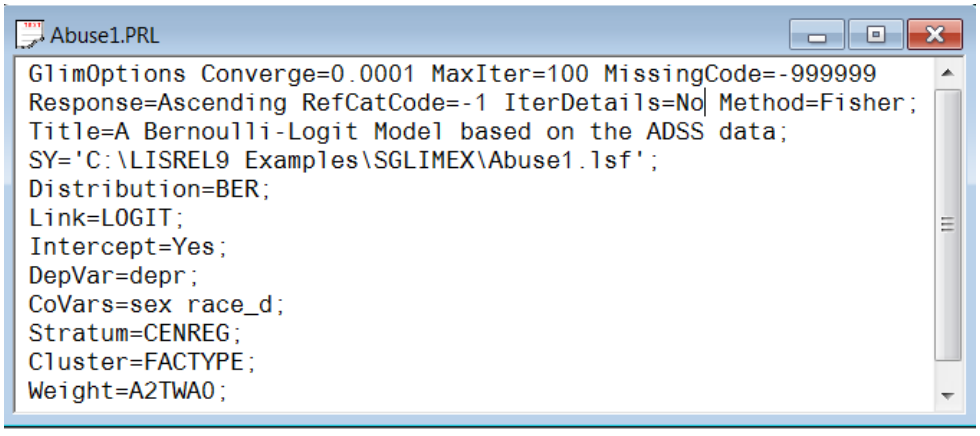
Click on the **Next** button to go to the **Dependent and Independent Variables** dialog box. Specify the response variable depr by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Dependent variable** section. Specify the covariates, sex and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



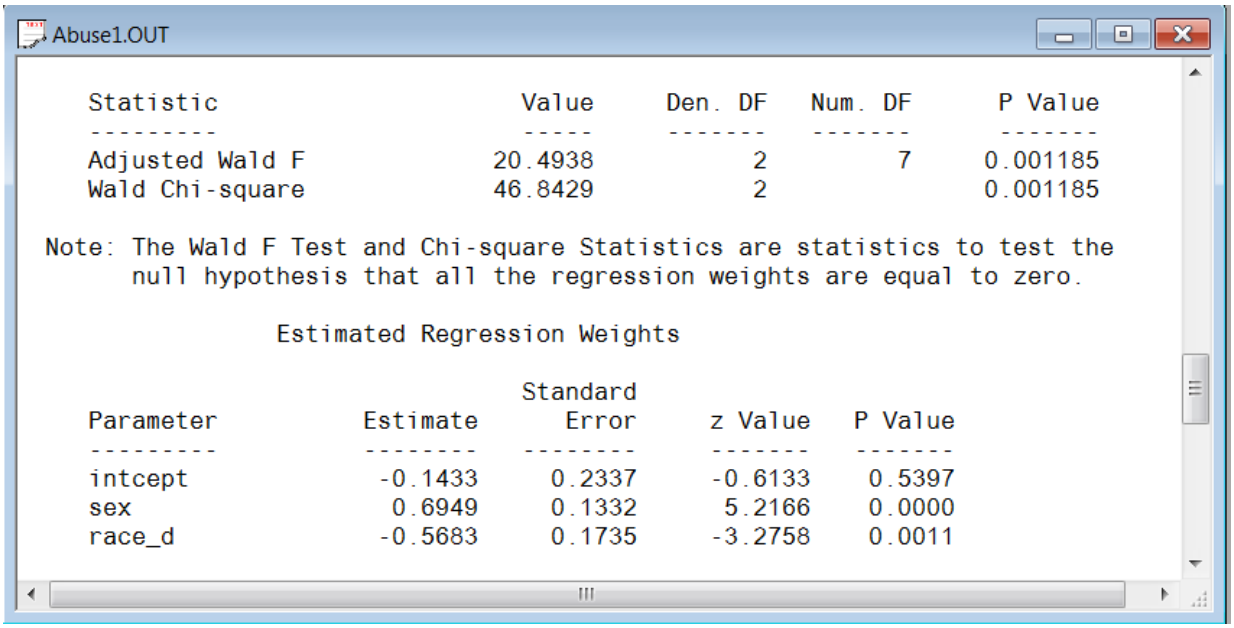
As this concludes our specifications, click on the **Finish** button to open the following text editor window for **abuse1.prl**.



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 if 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

$$\hat{\mathrm{P}}\left(\mathrm{depr}_k = 1\right) = \frac{\exp\left(-0.14 + 0.70*\mathrm{sex}_k - 0.57*\mathrm{race\_d}_k\right)}{1 + \exp\left(-0.14 + 0.70*\mathrm{sex}_k - 0.57*\mathrm{race\_d}_k\right)}.$$

The estimated probability that a nonwhite female client ($\mathrm{sex}_k = 1$ and $\mathrm{race\_d}_k = 0$) has a history of depression follows from this fitted model as

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

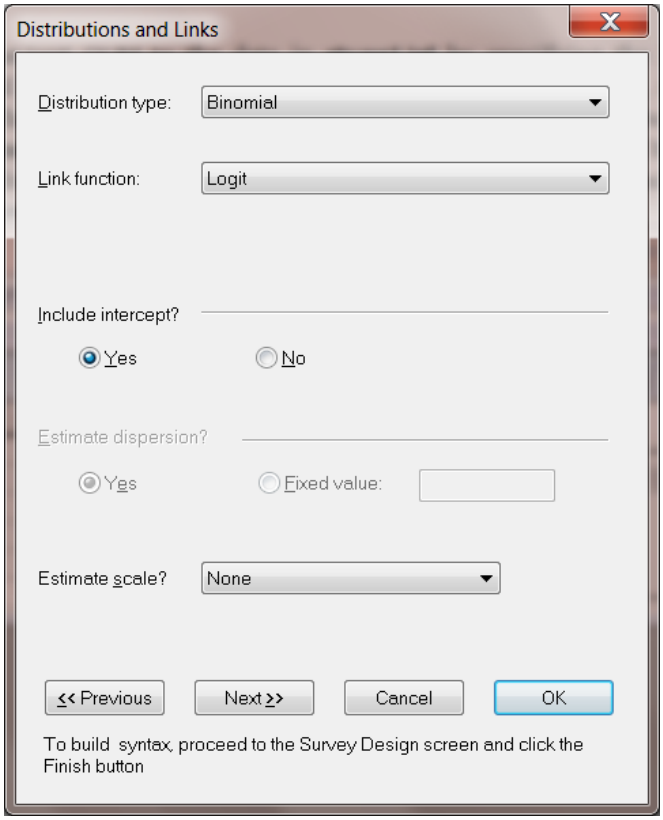
Similarly, the estimated probability that a nonwhite male client has a history of depression follows as 0.47. From the results above, it follows that $\exp\left(\hat{\beta}_1\right) = \exp\left(0.70\right) = 2.01$ which implies that female clients are twice as likely as male clients to have a history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

5. Analyzing binary outcomes from complex survey designs (method 2)

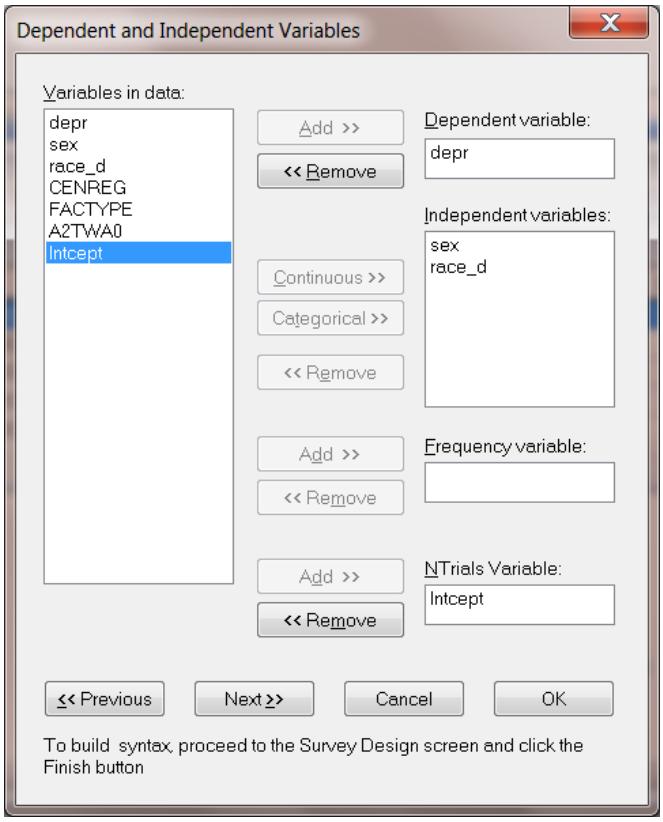
In this example, we illustrate that a GLIM with a Binomial distribution is identical to a GLIM with a Bernoulli distribution when the number of trials is one for each observation. If the NTrials command is omitted from the syntax file, the number of trials will automatically be set to unity.

Setting up the analysis

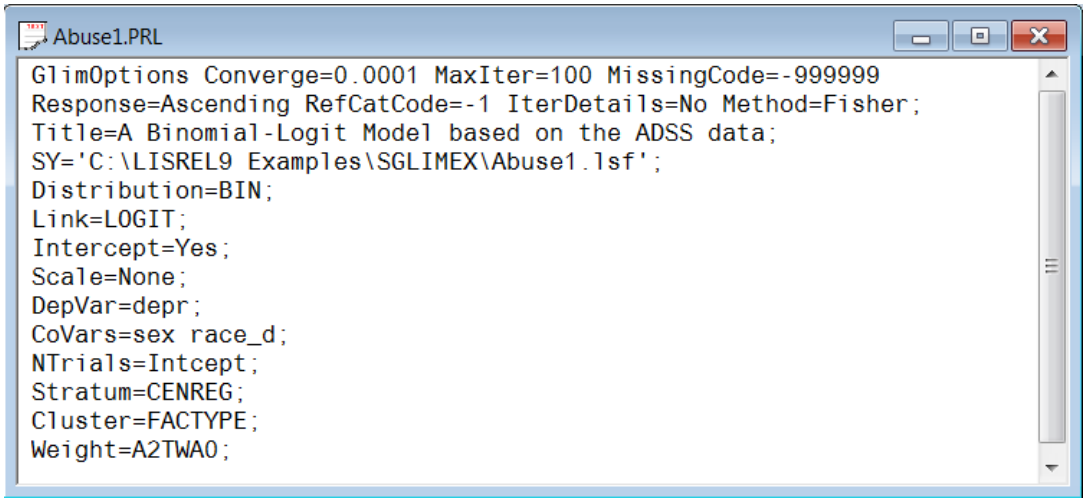
We fit the Binomial-logit GLIM to the data in **abuse1.lsf** by specifying the Binomial sampling distribution instead of the Bernoulli sampling distribution. First, however, select the **Title and Options** option on the **SurveyGLIM** menu to go to the **Title and Options** dialog box and enter the 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.



Click on the **Next** button and add the variable intcept as the **NTRIALS variable**.



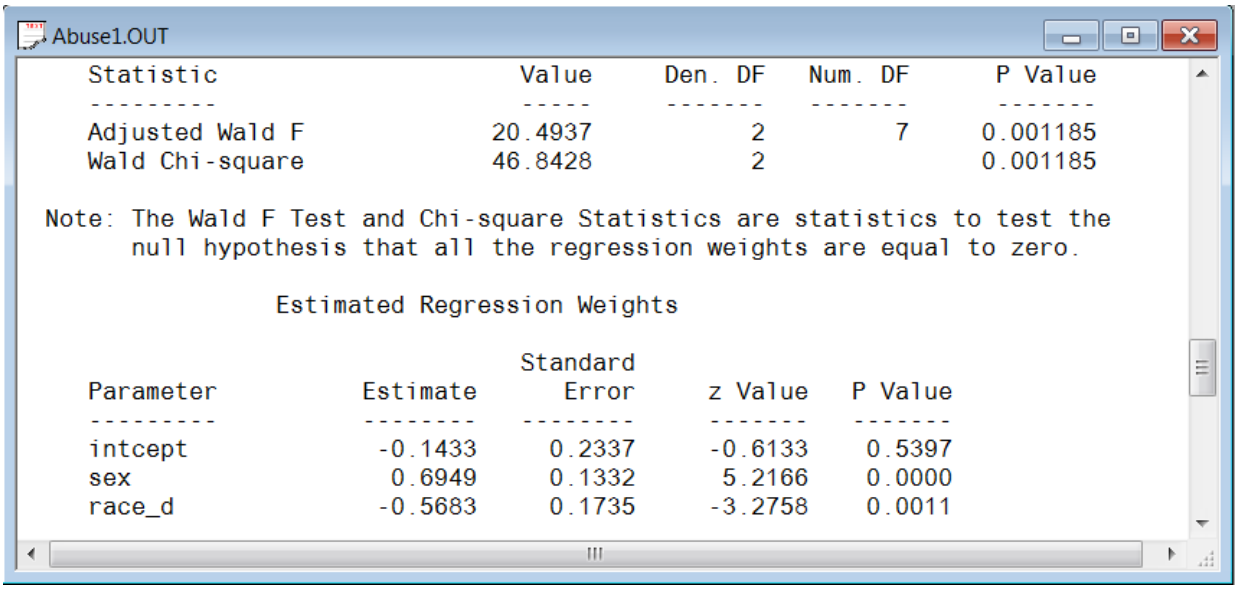
Since these are the only changes we needed to specify, click on the **Next** button of the **Dependent and Independent Variables** dialog box and the **Finish** button of the **Survey Design** dialog box to open the following text editor window for **abuse1.prl**.



Submit **abuse1.prl** 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.



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.