Scientific Software International

GLIMs for count data using substance abuse data

Contents

1. Introduction

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.

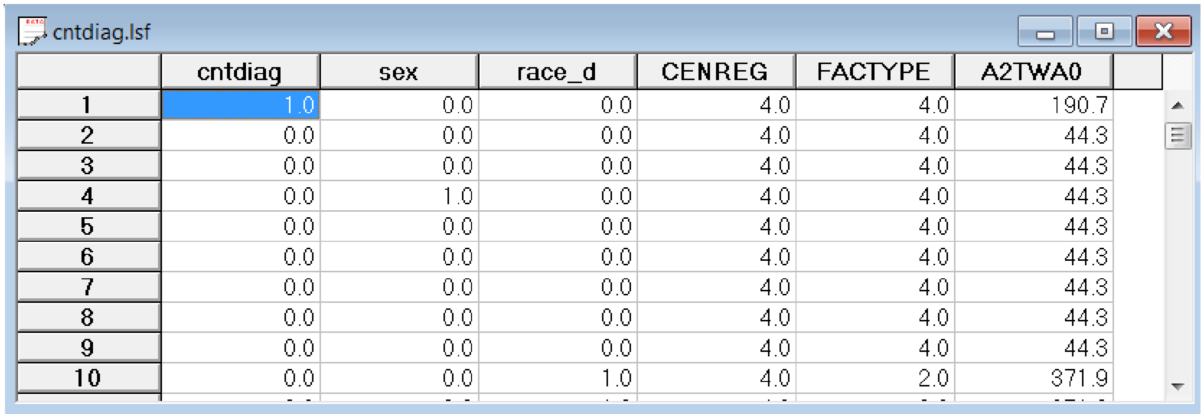
In many situations, count variables are skewed. The percentage of subjects with a score of zero or 1 is very large, those with a score of 4 or 5 or 6 considerably less common, and those with a score of 11 or 12 rare. For example, the number of delinquent acts committed by a teenager is a count variable. 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 the SurveyGLIM module of LISREL 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.

2. 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 **Generealized Linear Modeling examples** folder as the LSF **cntdiag.lsf**. The first portion of this file is shown in the following LSF window.



	cntdiag	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.0	0.0	0.0	4.0	4.0	190.7
2	0.0	0.0	0.0	4.0	4.0	44.3
3	0.0	0.0	0.0	4.0	4.0	44.3
4	0.0	1.0	0.0	4.0	4.0	44.3
5	0.0	0.0	0.0	4.0	4.0	44.3
6	0.0	0.0	0.0	4.0	4.0	44.3
7	0.0	0.0	0.0	4.0	4.0	44.3
8	0.0	0.0	0.0	4.0	4.0	44.3
9	0.0	0.0	0.0	4.0	4.0	44.3
10	0.0	0.0	1.0	4.0	2.0	371.9

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.

3. The models

The sampling distributions

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

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

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

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

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

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

The mean model

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

$$\mu_k=\exp\left(\alpha+\beta_1x_{1k}+\beta_2x_{2k}+\ldots+\beta_rx_{rk}\right)$$

where μ_k denotes the mean value of the response variable for client k, x_{jk} denotes the value of the j-th predictor ($j=1,2,\ldots,r$) for client k, and α, β_1, … 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

$$\mathrm{E}\left[\mathrm{cntdiag}_k\right]=\exp\left(\alpha+\beta_1*\mathrm{sex}_k+\beta_2*\mathrm{race\_d}_k\right)$$

where $\mathrm{E}\left[\mathrm{cntdiag}_k\right]$ denotes the mean number of diagnoses for client k, sex_k and $\mathrm{race\_d}_k$ 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 ($\mathrm{sex}_k=1$) and male ($\mathrm{sex}_k=0$) clients may be expressed as

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

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

$$\hat{\mathrm{E}}\left[\mathrm{cntdiag}_k\right] = \exp\left(\hat{\alpha} + \hat{\beta}_1 * \mathrm{sex}_k + \hat{\beta}_2 * \mathrm{race\_d}_k\right)$$

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

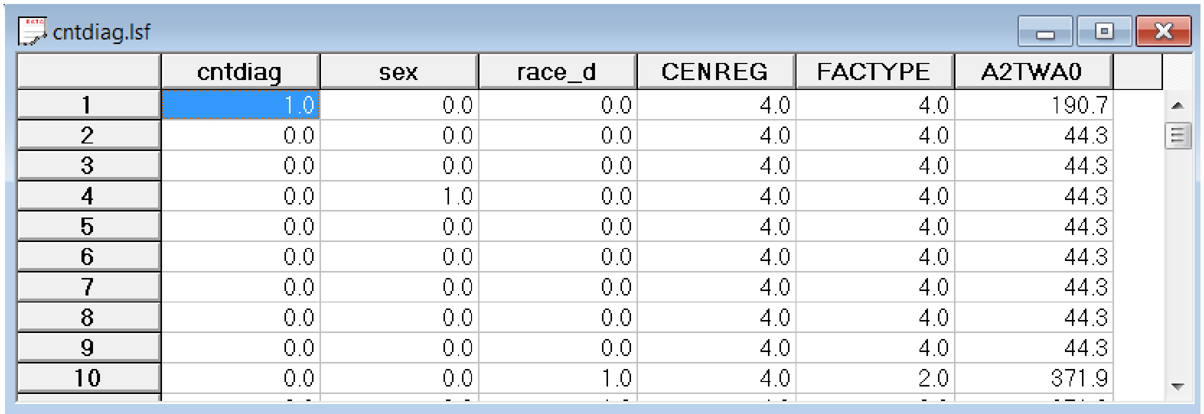
4. Analyzing counts from a complex sampling design

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

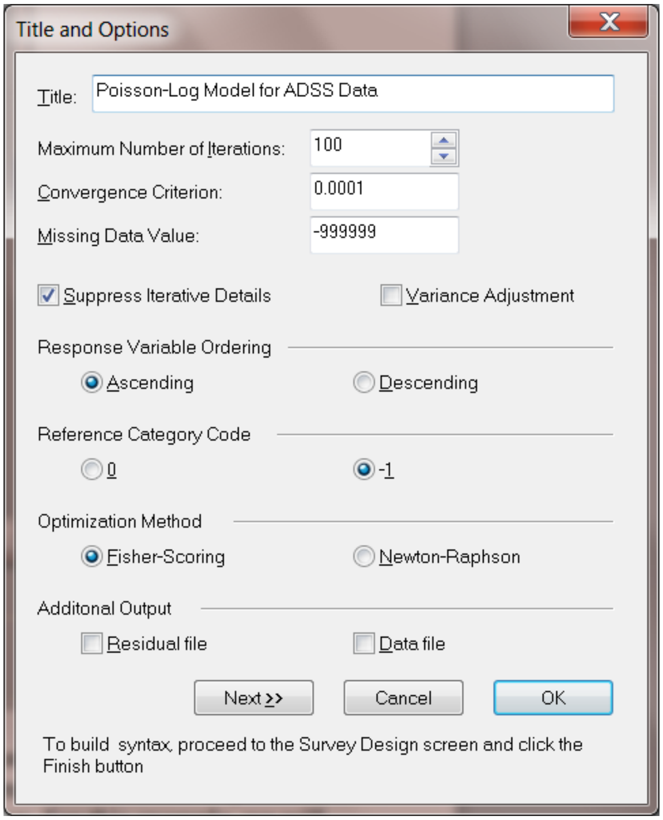
The first step is to open the LSF shown above in the LISREL LSF window. This is accomplished as follows. Use the **Open** option on the **File** menu of the root window of LISREL to load the **Open** dialog box and select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box.

Browse for and open the file **cntdiag.lsf.** Click on the **SurveyGLIM** menu to produce the following LSF window.



	cntdiag	sex	race_d	CENREG	FACTYPE	A2TWA0
1	1.0	0.0	0.0	4.0	4.0	190.7
2	0.0	0.0	0.0	4.0	4.0	44.3
3	0.0	0.0	0.0	4.0	4.0	44.3
4	0.0	1.0	0.0	4.0	4.0	44.3
5	0.0	0.0	0.0	4.0	4.0	44.3
6	0.0	0.0	0.0	4.0	4.0	44.3
7	0.0	0.0	0.0	4.0	4.0	44.3
8	0.0	0.0	0.0	4.0	4.0	44.3
9	0.0	0.0	0.0	4.0	4.0	44.3
10	0.0	0.0	1.0	4.0	2.0	371.9

The next step is complete the sequence of four dialog boxes of the SurveyGLIM GUI described. 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.



Title and Options

Title: Poisson-Log Model for ADSS Data

Maximum Number of Iterations: 100

Convergence Criterion: 0.0001

Missing Data Value: -999999

☑ Suppress Iterative Details　　☐ Variance Adjustment

Response Variable Ordering

● Ascending　　○ Descending

Reference Category Code

○ 0　　● -1

Optimization Method

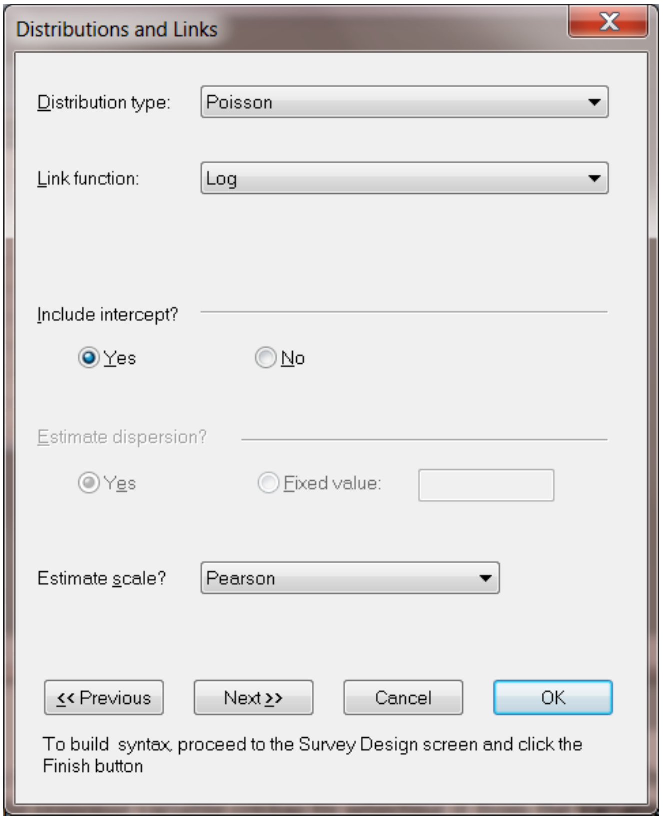
● Fisher-Scoring　　○ Newton-Raphson

Additonal Output

☐ Residual file　　☐ Data file

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

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. Select the **Pearson** option from the **Estimate scale?** drop-down list box to produce the following **Distributions and Links** dialog box.



Distributions and Links

Distribution type: Poisson

Link function: Log

Include intercept?

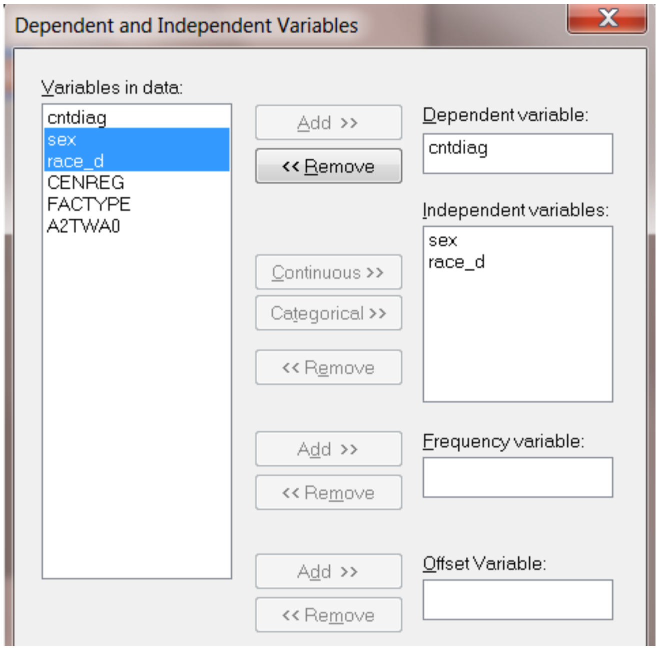
● Yes　　○ No

Estimate dispersion?

● Yes　　○ Fixed value:

Estimate scale? Pearson

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



Variables in data:
cntdiag
sex
race_d
CENREG
FACTYPE
A2TWA0

Dependent variable:
cntdiag

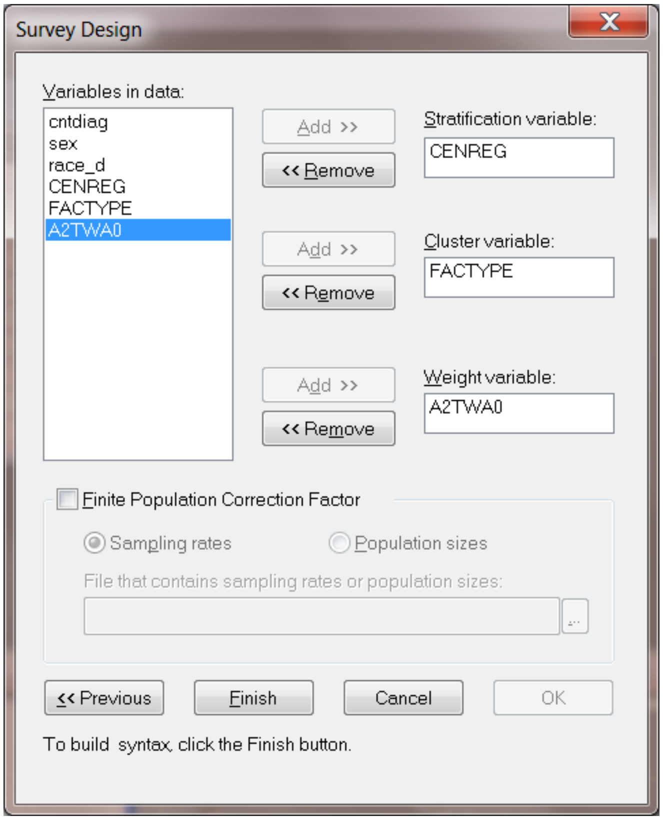
Independent variables:
sex
race_d

Frequency variable:

Offset Variable:

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.



Variables in data:
cntdiag
sex
race_d
CENREG
FACTYPE
A2TWA0

Stratification variable: CENREG

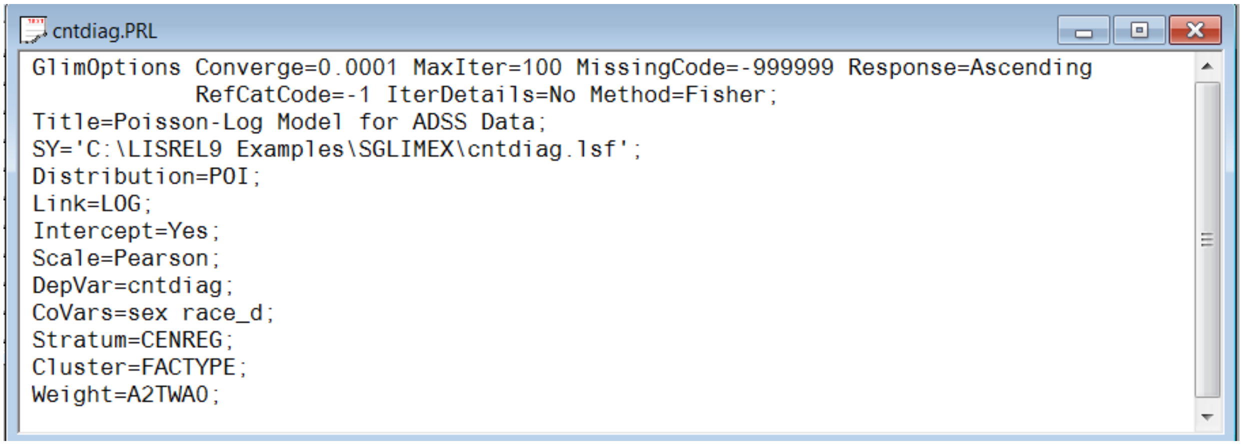
Cluster variable: FACTYPE

Weight variable: A2TWA0

□ Finite Population Correction Factor

To build syntax, click the Finish button.

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.prl**.

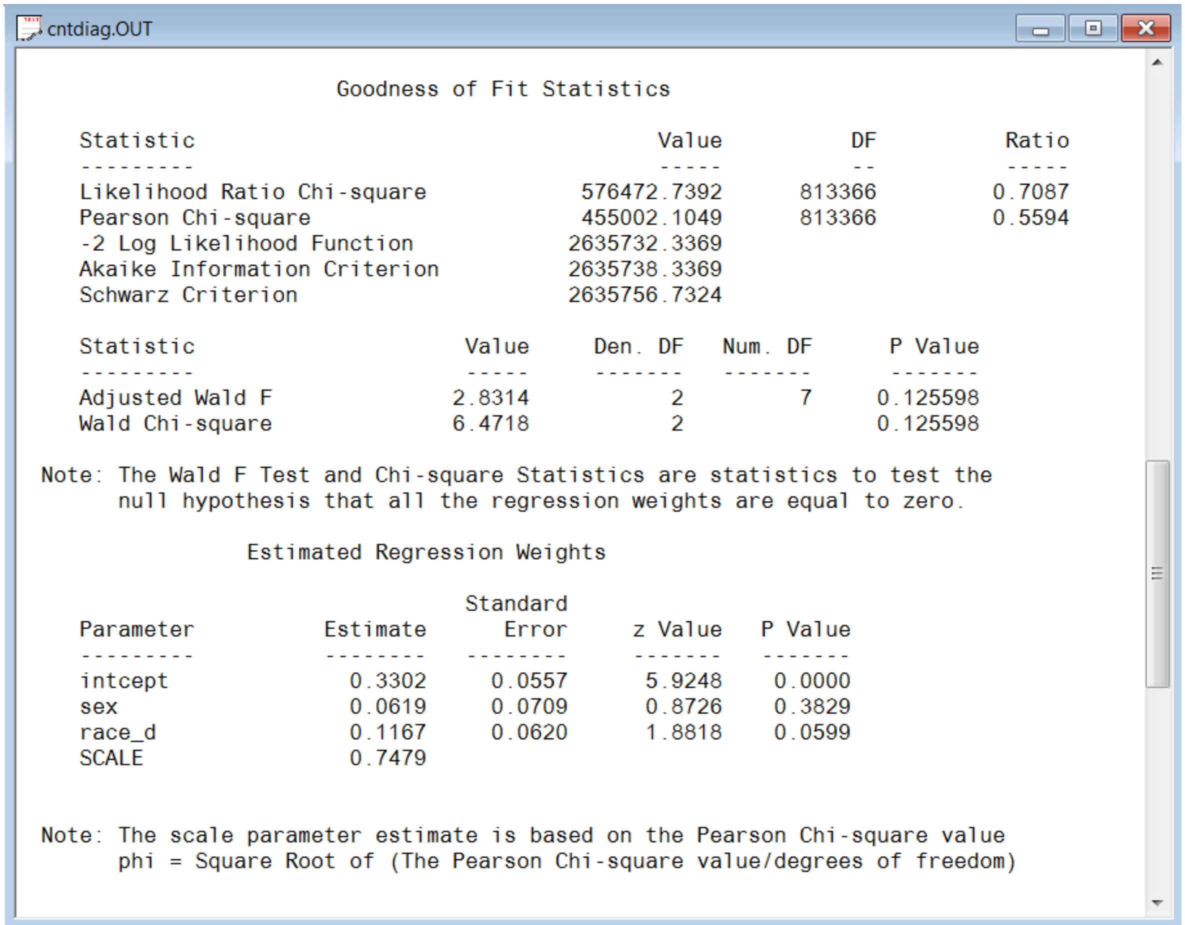


```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Poisson-Log Model for ADSS Data;
SY='C:\LISREL9 Examples\SGLIMEX\cntdiag.lsf';
Distribution=POI;
Link=LOG;
Intercept=Yes;
Scale=Pearson;
DepVar=cntdiag;
CoVars=sex race_d;
Stratum=CENREG;
Cluster=FACTYPE;
Weight=A2TWA0;
```

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.



```
                       Goodness of Fit Statistics

   Statistic                                    Value          DF           Ratio
   ---------                                    -----          --           -----
   Likelihood Ratio Chi-square            576472.7392      813366          0.7087
   Pearson Chi-square                     455002.1049      813366          0.5594
   -2 Log Likelihood Function            2635732.3369
   Akaike Information Criterion          2635738.3369
   Schwarz Criterion                     2635756.7324

   Statistic                     Value     Den. DF   Num. DF      P Value
   ---------                     -----     -------   -------      -------
   Adjusted Wald F              2.8314           2         7     0.125598
   Wald Chi-square              6.4718           2               0.125598

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

                                 Standard
   Parameter          Estimate      Error     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)

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

And

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

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, $\hat{\boldsymbol{\beta}}$ denotes the estimate of the parameter vector, $\boldsymbol{\beta}$, of regression weights and $\hat{\Upsilon}$ denotes the estimated asymptotic covariance matrix of the estimators of the elements of $\boldsymbol{\beta}$. 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

$$\hat{\mathrm{E}}\left[\mathrm{cntdiag}_k\right] = \exp\left(0.33 + 0.06 * \mathrm{sex}_k + 0.12 * \mathrm{race\_d}_k\right)$$

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 ($\mathrm{sex}_k = 1$ and $\mathrm{race}_k = 1$) is given by

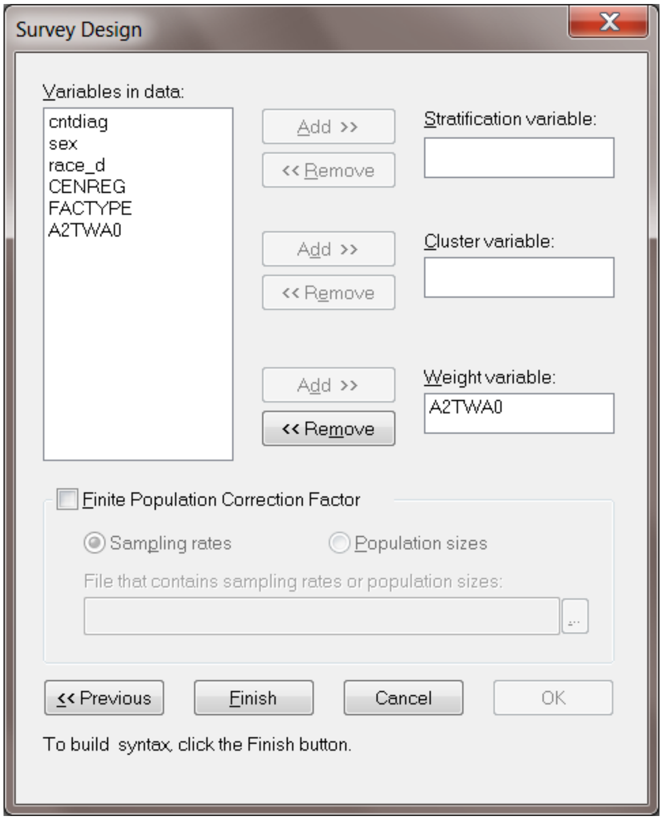
$$\exp\left(0.33 + 0.06 + 0.12\right) = \exp\left(0.51\right) = 1.67$$

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

5. 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.



Variables in data:
cntdiag
sex
race_d
CENREG
FACTYPE
A2TWA0

Stratification variable:

Cluster variable:

Weight variable: A2TWA0

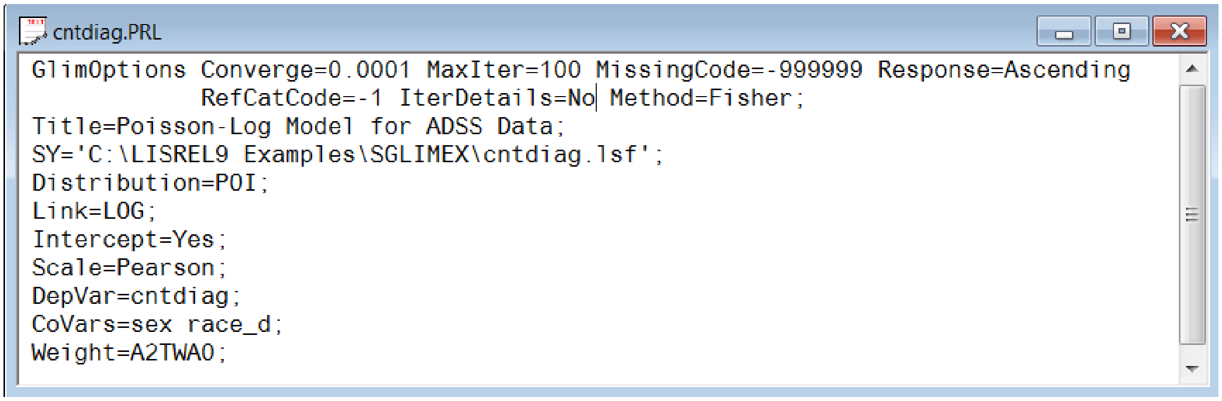
□ Finite Population Correction Factor

● Sampling rates ○ Population sizes

File that contains sampling rates or population sizes:

To build syntax, click the Finish button.

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

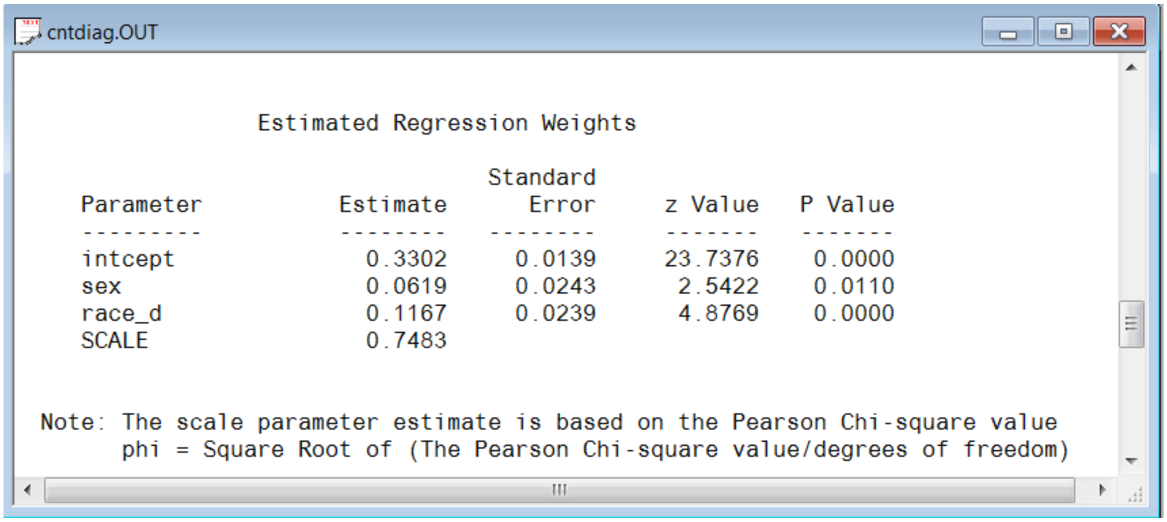


```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Poisson-Log Model for ADSS Data;
SY='C:\LISREL9 Examples\SGLIMEX\cntdiag.lsf';
Distribution=POI;
Link=LOG;
Intercept=Yes;
Scale=Pearson;
DepVar=cntdiag;
CoVars=sex race_d;
Weight=A2TWA0;
```

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

                                      Standard
     Parameter           Estimate       Error       z Value    P Value
     ---------           --------     --------      -------    -------
     intcept               0.3302       0.0139      23.7376     0.0000
     sex                   0.0619       0.0243       2.5422     0.0110
     race_d                0.1167       0.0239       4.8769     0.0000
     SCALE                 0.7483


  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)
```

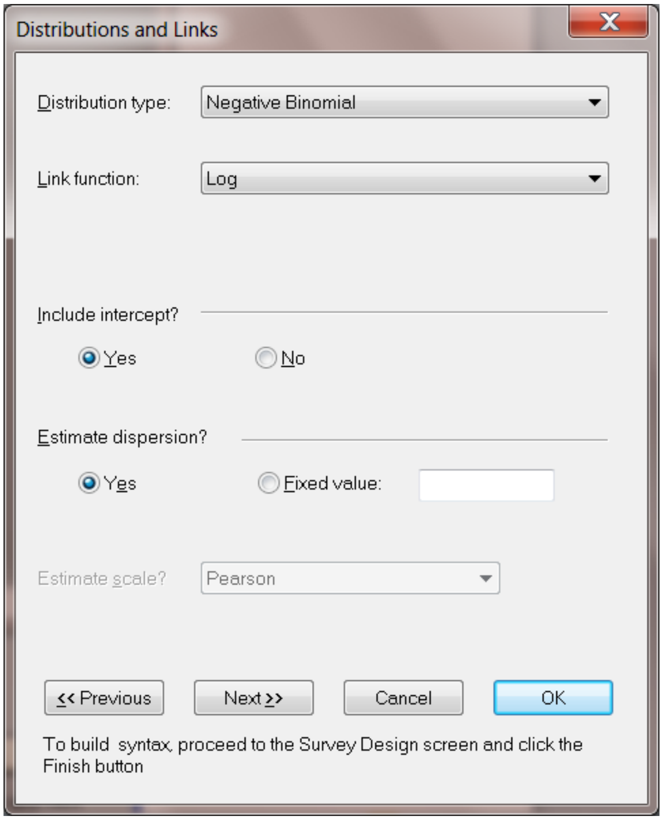
The results above indicate that although the parameter estimates are identical to those obtained when the design of the complex survey was taken into account, the standard error estimates are significantly smaller (*cf*. Brogan, 1998). As a consequence, both gender and race appear to have a statistically 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.

6. 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.lsf**, 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.



Distribution type: Negative Binomial

Link function: Log

Include intercept?

● Yes ○ No

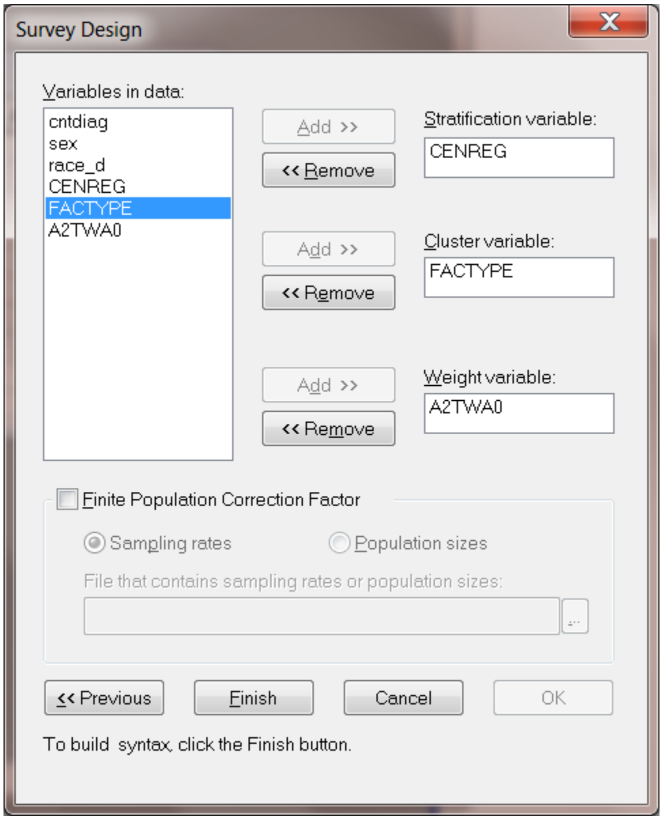
Estimate dispersion?

● Yes ○ Fixed value:

Estimate scale? Pearson

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

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.



Variables in data:

cntdiag
sex
race_d
CENREG
FACTYPE
A2TWA0

Stratification variable: CENREG

Cluster variable: FACTYPE

Weight variable: A2TWA0

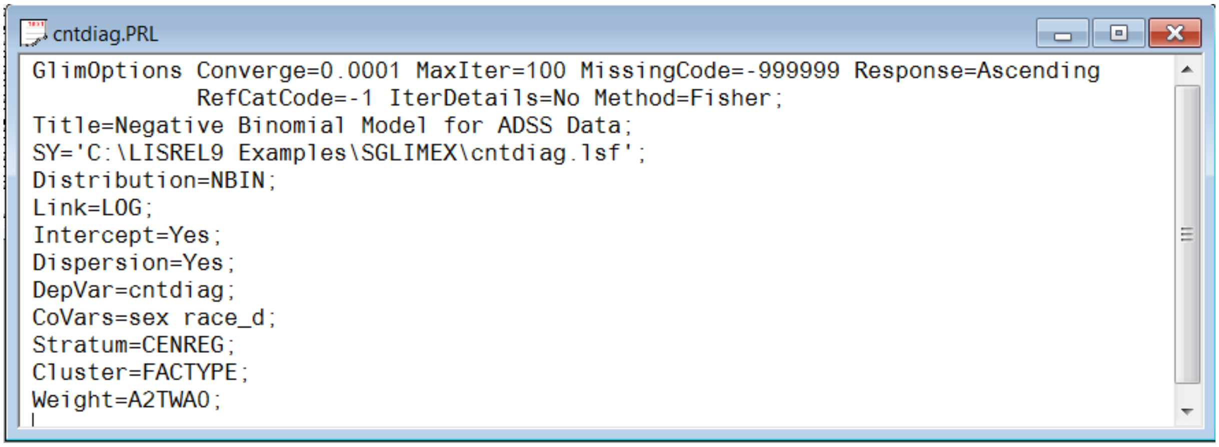
□ Finite Population Correction Factor

● Sampling rates ○ Population sizes

File that contains sampling rates or population sizes:

To build syntax, click the Finish button.

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

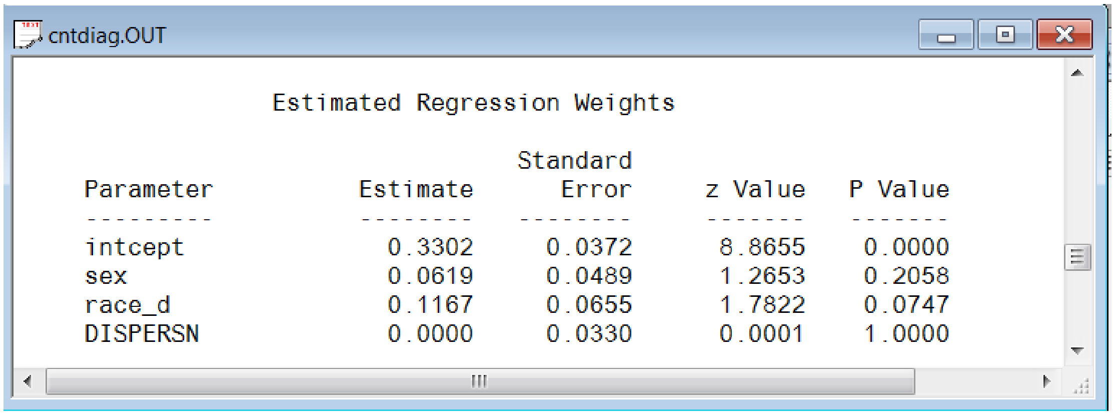


```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Negative Binomial Model for ADSS Data;
SY='C:\LISREL9 Examples\SGLIMEX\cntdiag.lsf';
Distribution=NBIN;
Link=LOG;
Intercept=Yes;
Dispersion=Yes;
DepVar=cntdiag;
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.



```
                   Estimated Regression Weights

                                      Standard
  Parameter             Estimate        Error       z Value    P Value
  ---------             --------      --------      -------    -------
  intcept                 0.3302        0.0372       8.8655     0.0000
  sex                     0.0619        0.0489       1.2653     0.2058
  race_d                  0.1167        0.0655       1.7822     0.0747
  DISPERSN                0.0000        0.0330       0.0001     1.0000
```

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.