Analyzing count data and correcting for over-dispersion

Contents

1. Example: Analyzing counts from a complex sampling design

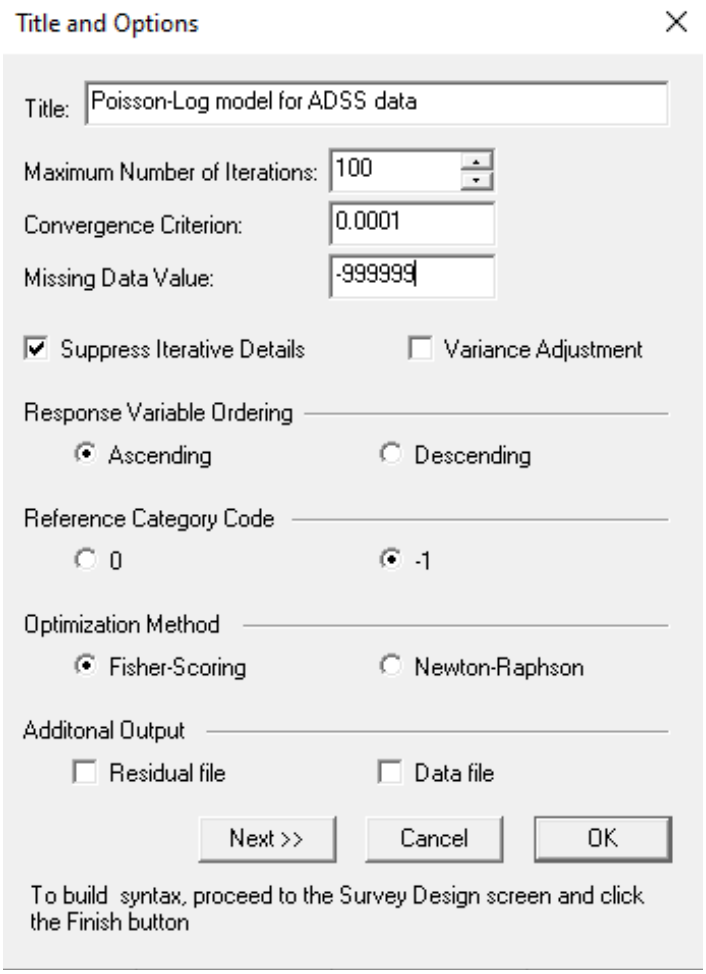
A question that a researcher may want to address is whether ethnicity and gender effects are associated with the number of substance abuse diagnoses. An appropriate statistical model for this type of count variable is a GLIM with a Poisson distribution and a log link function.

1.1 Setting up the analysis through the GUI

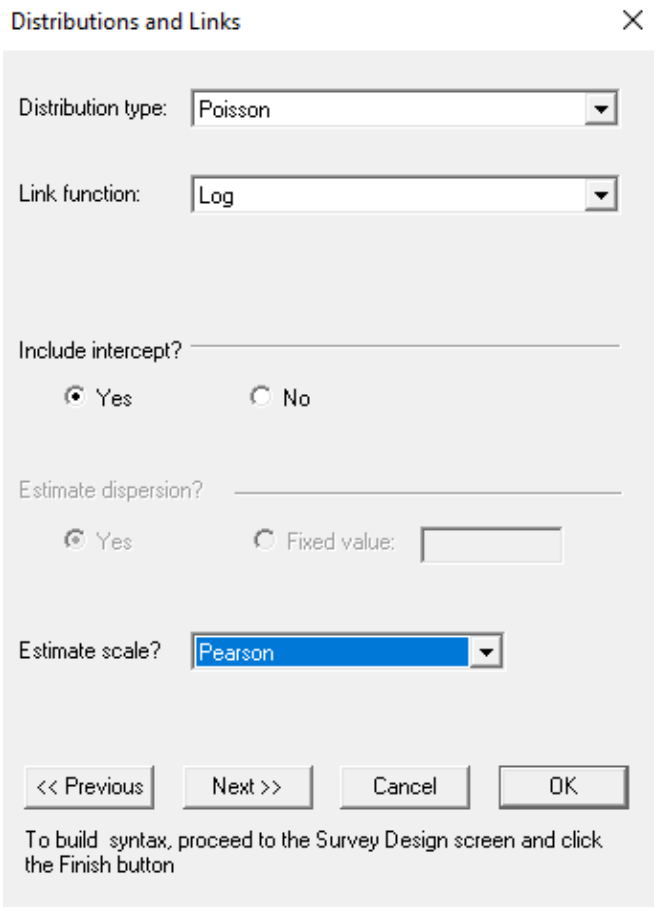
The first step is to open the LSF shown above in the LISREL LSF window. Use the **Open** option on the **File** menu of the root window of LISREL to load the **Open** dialog box. Select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box and browse for the file **cntdiag.psf** in the **Complex Survey GLIM examples** folder. Click on the **Open** button to open the file **cntdiag.lsf**.

	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

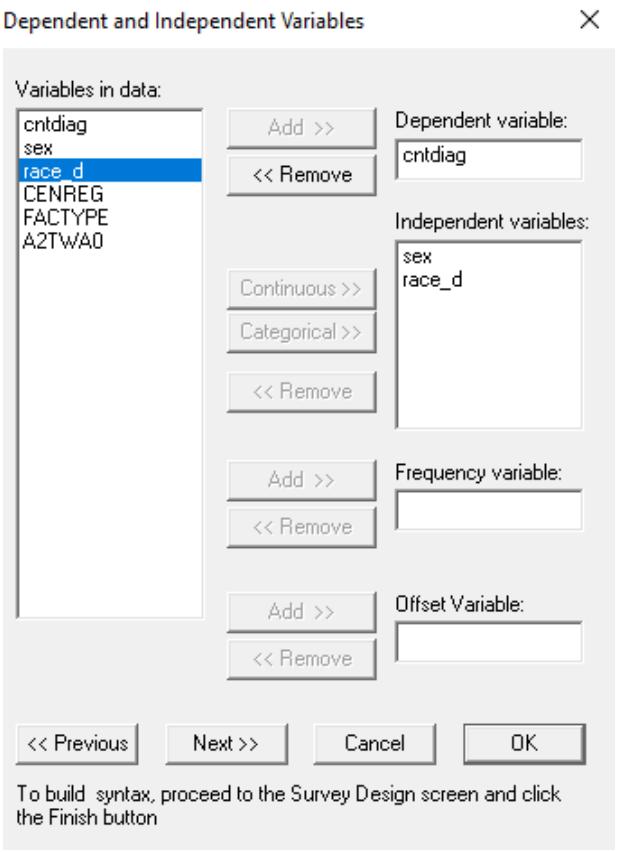
The next step is to enter the model specifications into the sequence of the four SurveyGLIM GUI dialog boxes. The **Title and Options** dialog box is the first dialog box and is accessed by selecting the **Title and Options** option on the **SurveyGLIM** menu above. In order to identify the analysis,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



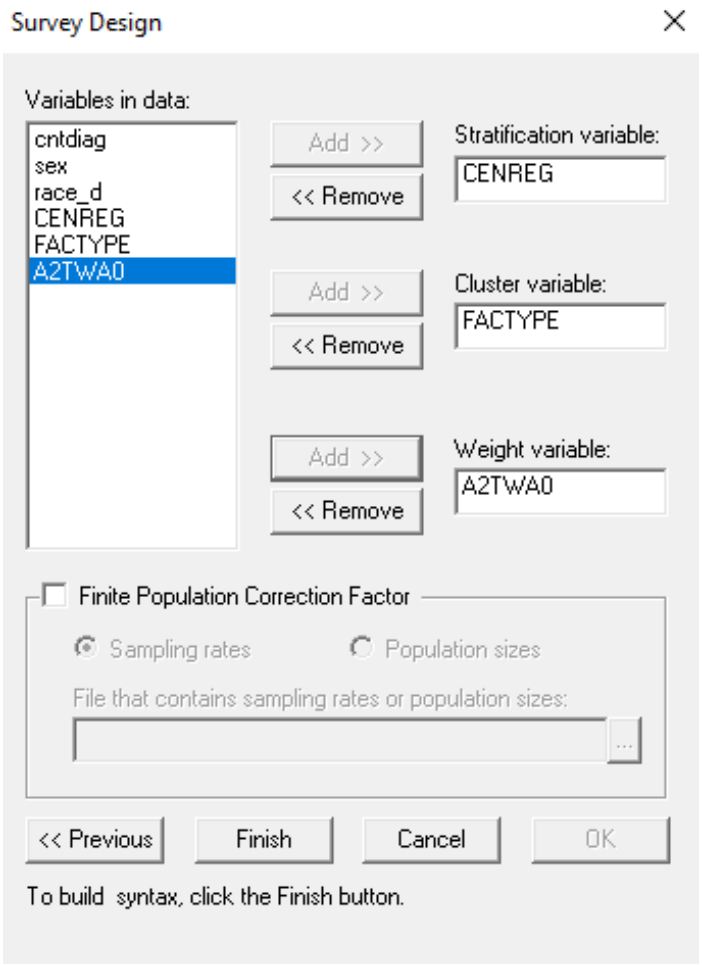
The default options will be used for this example. Click the **Next** button to access the **Distributions and Links** dialog box. Since we intend to fit a Poisson-log model, select the **Poisson** option from the **Distribution type** drop-down list box. For this example, we will estimate the scale 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.



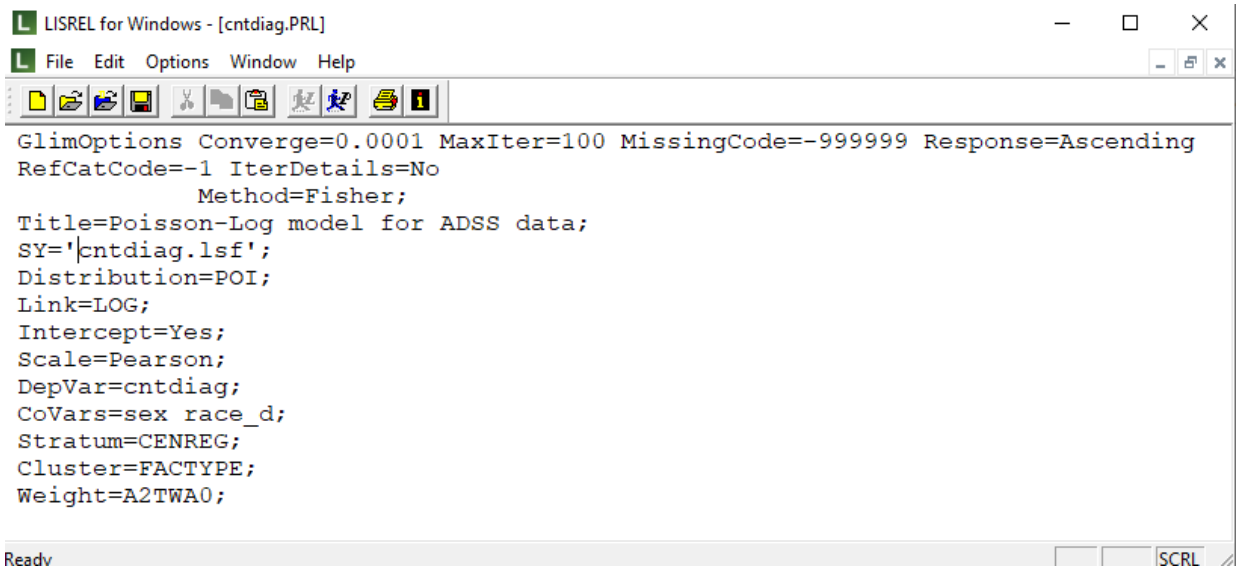
Proceed to the **Dependent and Independent Variables** dialog box by clicking on the **Next** button. Specify the response variable cntdiag by selecting it from the **Variables in data** list box and clicking on the **Add** button of the **Dependent variable** section. In a similar fashion, add the covariates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



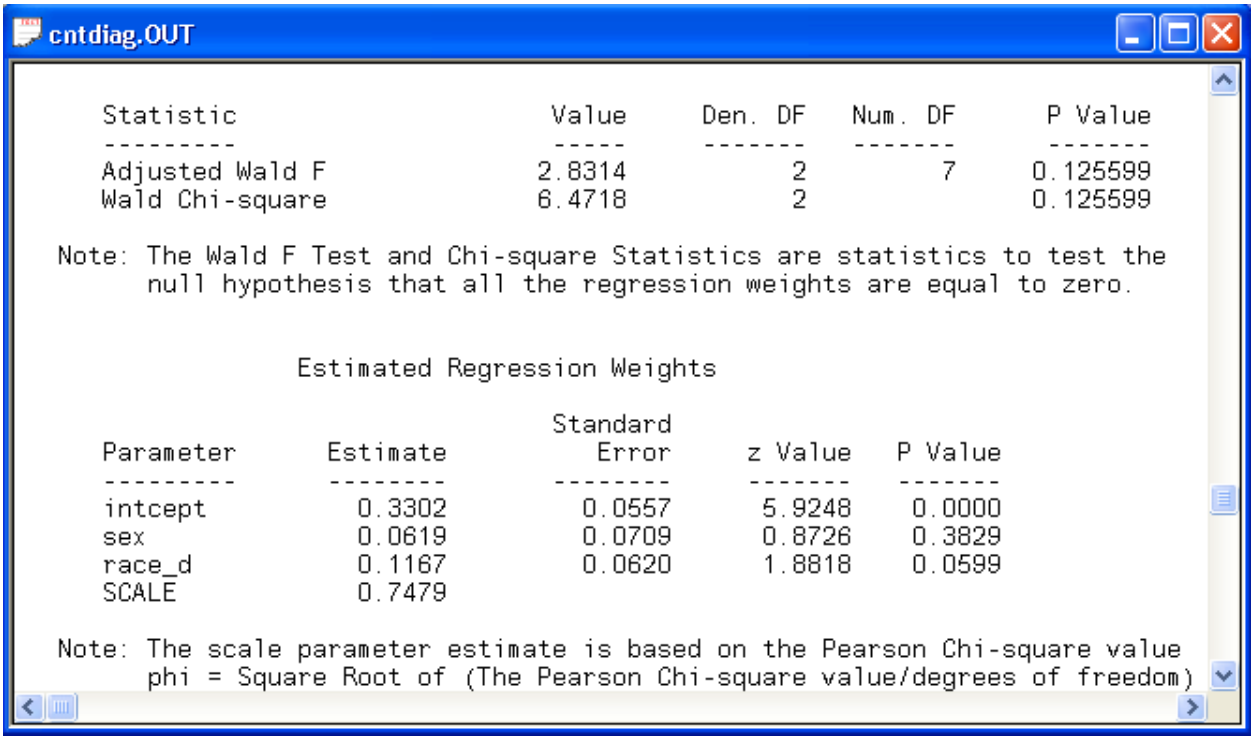
As no finite population information is available, we proceed to click on the **Finish** button to open the following text editor window for **cntdiag.prl**.



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

1.2 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.



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.

1.3 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

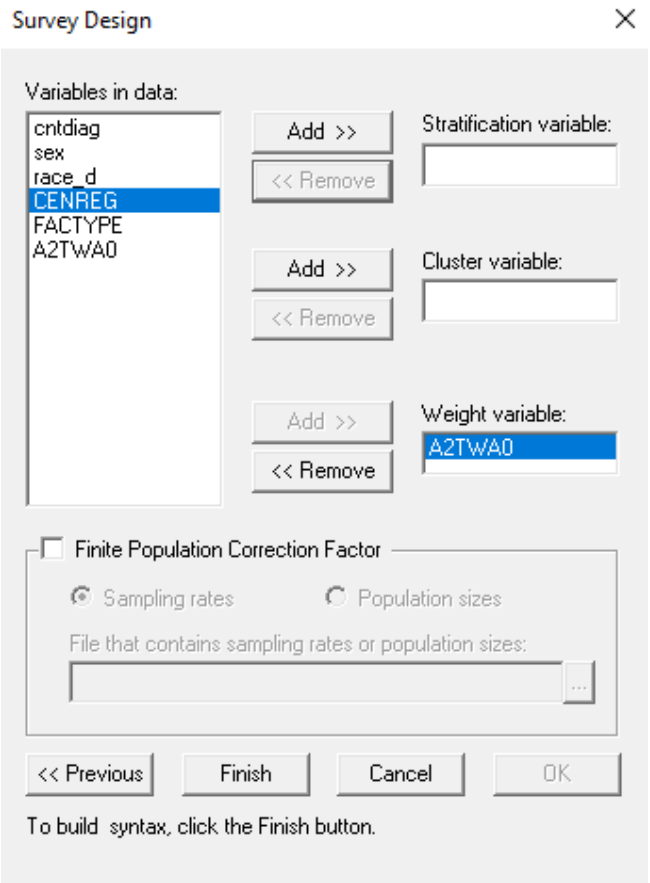
$$\exp\left(0.33 + 0.06 + 0.12\right) = \exp\left(0.51\right) = 1.67$$

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.

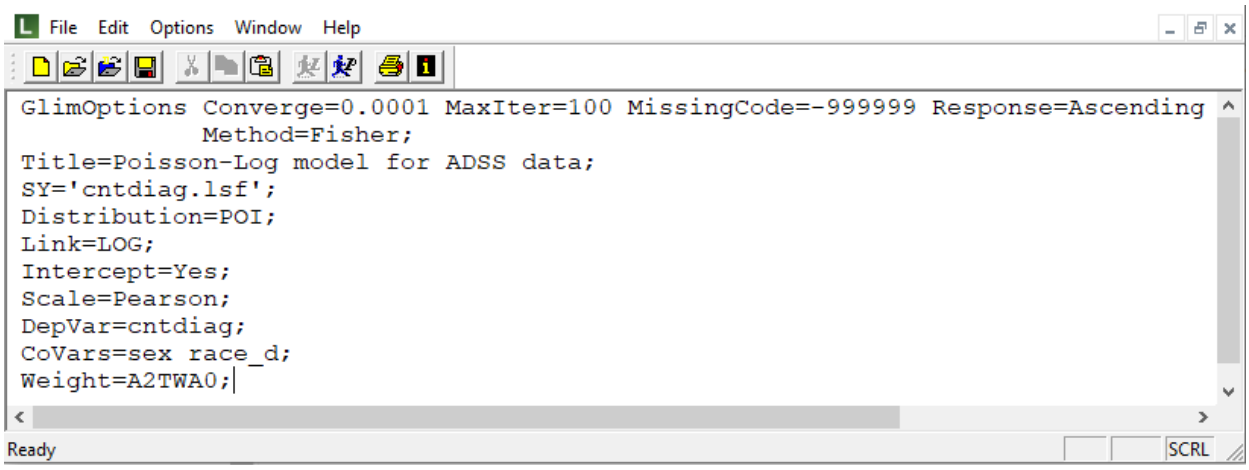
2. Ignoring stratification and clustering in the sample

2.1 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.



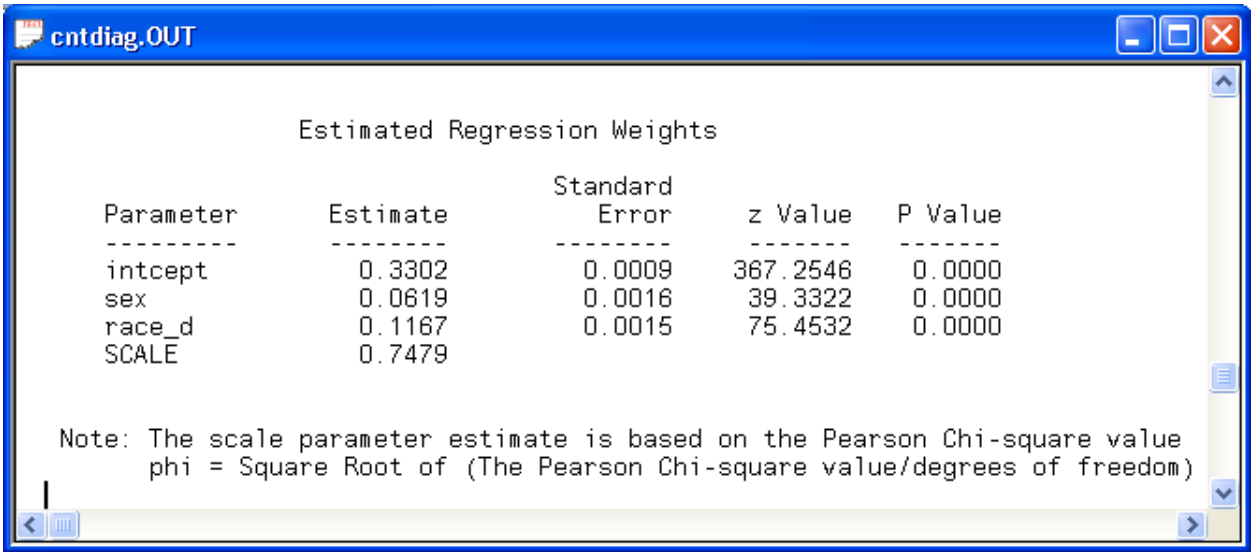
As this completes our modifications, click on the **Finish** button to open the following text editor window for **cntdiag.prl**.



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

2.2 Discussion of results

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



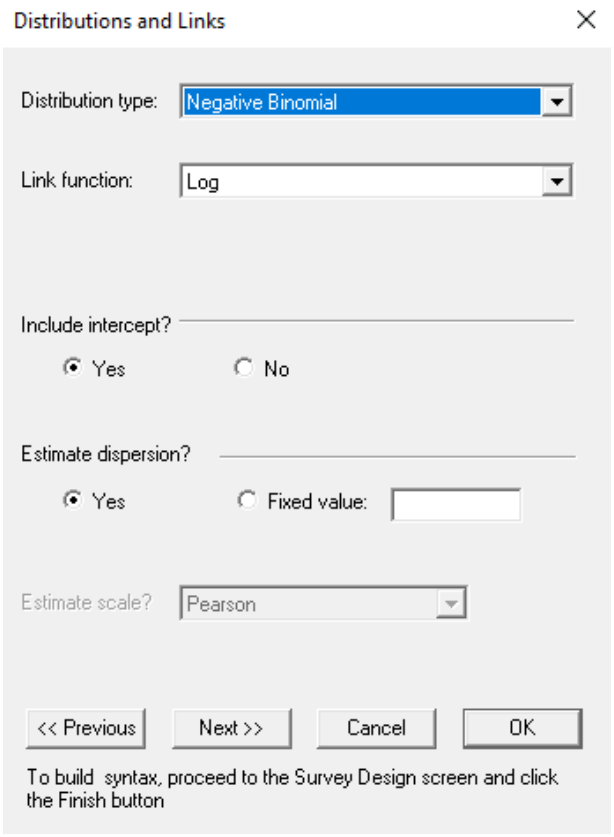
The results above indicate that although the parameter estimates are identical to those obtained when the design of the complex survey was taken into account, the standard error estimates are significantly smaller (*cf.* Brogan, 1998). As a consequence, both gender and race appear to have a statistically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

3. Example: Correcting for over-dispersion in an analysis of counts

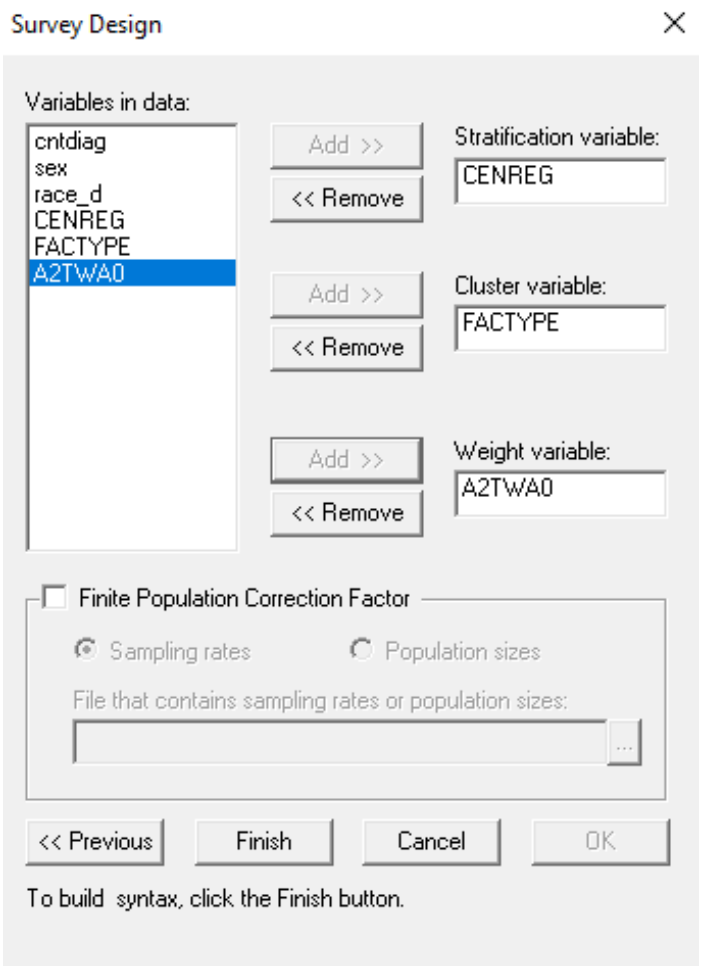
The results for the Poisson-log model indicated the presence of under-dispersion. Although the negative Binomial distribution is intended for dealing with over-dispersion, we will use it here for illustrative purposes.

3.1 Setting up the analysis

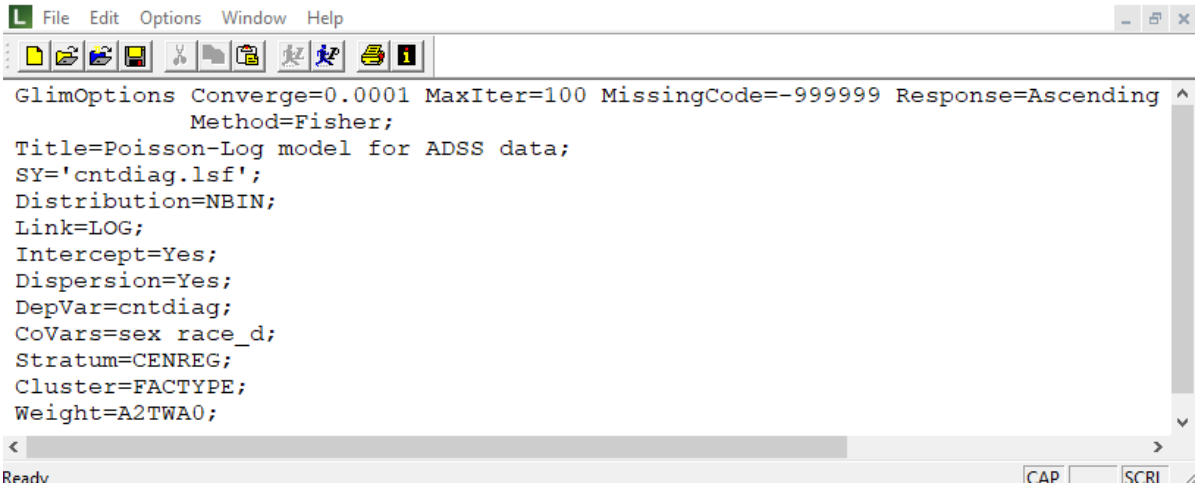
In order to fit the Negative Binomial-log model interactively to the data in **cntdiag.psf**, we only need to re-specify the sampling distribution. As in the previous analysis, start by modifying the title to **Fitting a Negative Binomial-Log model** by accessing the **Title and Options** dialog box 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.



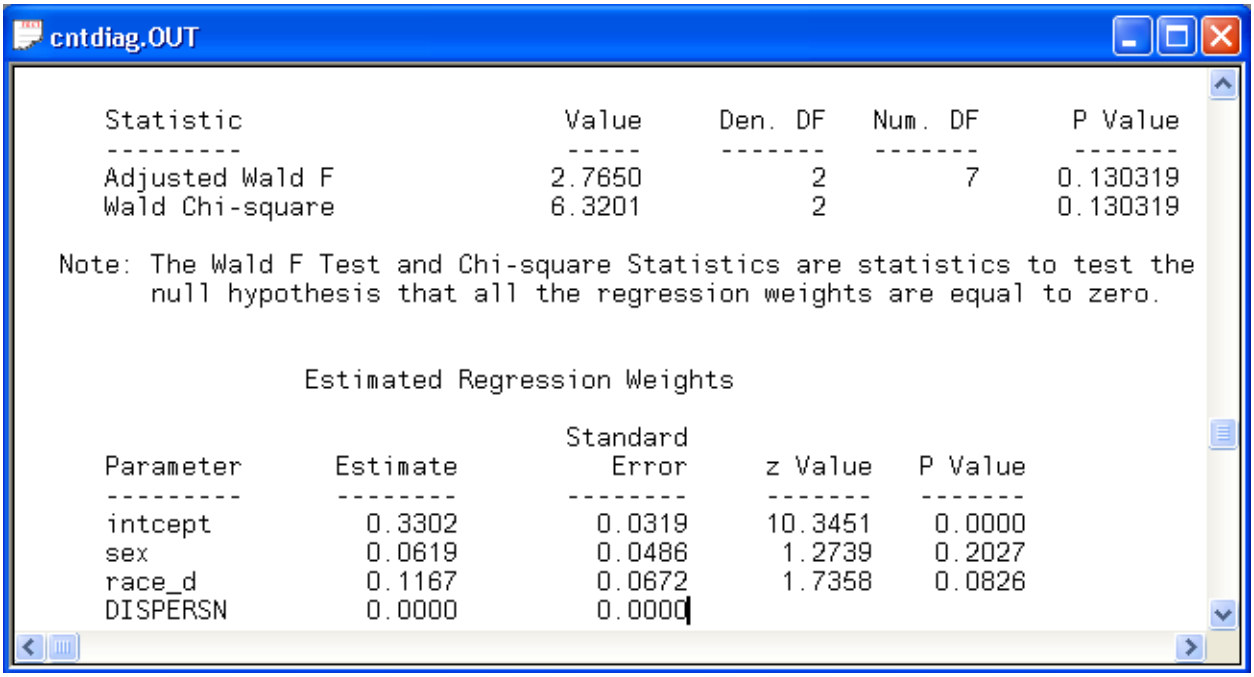
Click on the **Finish** button to open the following text editor window for **cntdiag.prl**.



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

3.2 Discussion of results – negative Binomial model

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



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.