GLIMs for ordinal responses using substance abuse data

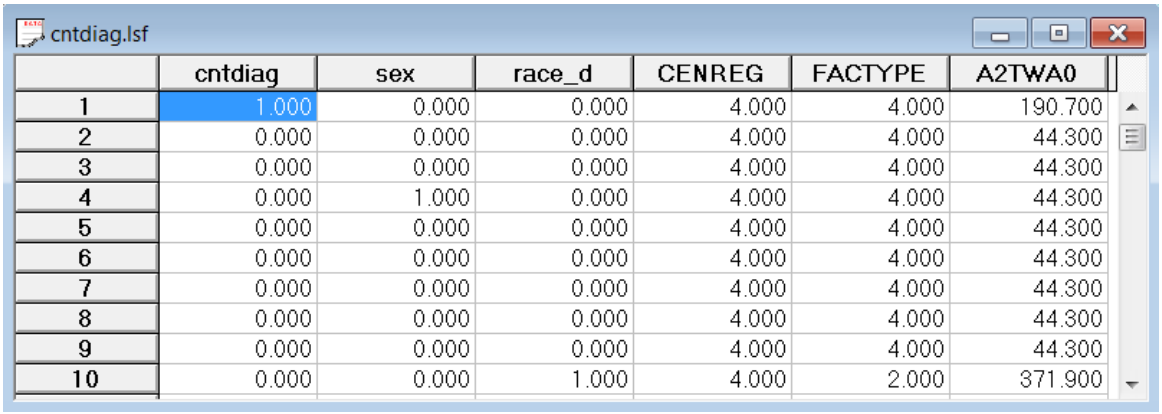
Contents

1. Introduction

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.

2. The data

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



cntdiag.lsf

	cntdiag	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	4.000	44.300
3	0.000	0.000	0.000	4.000	4.000	44.300
4	0.000	1.000	0.000	4.000	4.000	44.300
5	0.000	0.000	0.000	4.000	4.000	44.300
6	0.000	0.000	0.000	4.000	4.000	44.300
7	0.000	0.000	0.000	4.000	4.000	44.300
8	0.000	0.000	0.000	4.000	4.000	44.300
9	0.000	0.000	0.000	4.000	4.000	44.300
10	0.000	0.000	1.000	4.000	2.000	371.900

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.

3. 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 $\boldsymbol{\pi}_k = [\pi_{k1}, \pi_{k2} \ \ldots \ \pi_{kp}]'$.

The probability models

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

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

and

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

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 ($j = 1, 2, \ldots, r$) for subject k, α_1, α_2, …, α_{p-1}, β_1, …, β_{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

$$\mathrm{P}\left(\mathrm{cntdiag}_k \le l\right) = \frac{\exp\left(\alpha_l + \beta_1 * \mathrm{sex}_k + \beta_2 * \mathrm{race\_d}_k\right)}{1 + \exp\left(\alpha_l + \beta_1 * \mathrm{sex}_k + \beta_2 * \mathrm{race\_d}_k\right)} \quad l = 1, 2, 3$$

and

$$\mathrm{P}\left(\mathrm{cntdiag}_k \le l\right) = \Phi\left(\alpha_l + \beta_1 * \mathrm{sex}_k + \beta_2 * \mathrm{race\_d}_k\right) \quad l = 1, 2, 3$$

respectively where $\mathrm{P}\left(\mathrm{cntdiag}_k \le l\right)$ denotes the cumulative probability that category l was recorded for client k and α_1, α_2, α_3, β_1 and β_2 denote unknown parameters. The specific probabilities for each response category for client k for both these models may be obtained from the following expressions.

$$\mathrm{P}\left(\mathrm{cntdiag}_k = 1\right) = \mathrm{P}\left(\mathrm{cntdiag}_k \le 1\right)$$

$$\mathrm{P}\left(\mathrm{cntdiag}_k = 2\right) = \mathrm{P}\left(\mathrm{cntdiag}_k \le 2\right) - \mathrm{P}\left(\mathrm{cntdiag}_k \le 1\right)$$

$$\mathrm{P}\left(\mathrm{cntdiag}_k = 3\right) = \mathrm{P}\left(\mathrm{cntdiag}_k \le 3\right) - \mathrm{P}\left(\mathrm{cntdiag}_k \le 2\right).$$

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

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

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

$$\hat{\mathrm{P}}\left(\mathrm{cntdiag}_k \le l\right) = \frac{\exp\left(\hat{\alpha}_l + \hat{\beta}_1 * \mathrm{sex}_k + \hat{\beta}_2 * \mathrm{race\_d}_k\right)}{1 + \exp\left(\hat{\alpha}_l + \hat{\beta}_1 * \mathrm{sex}_k + \hat{\beta}_2 * \mathrm{race\_d}_k\right)} \quad l = 1, 2, 3$$

and

$$\hat{\mathrm{P}}\left(\mathrm{cntdiag}_k \le l\right) = \Phi\left(\hat{\alpha}_l + \hat{\beta}_1 * \mathrm{sex}_k + \hat{\beta}_2 * \mathrm{race\_d}_k\right) \quad l = 1, 2, 3$$

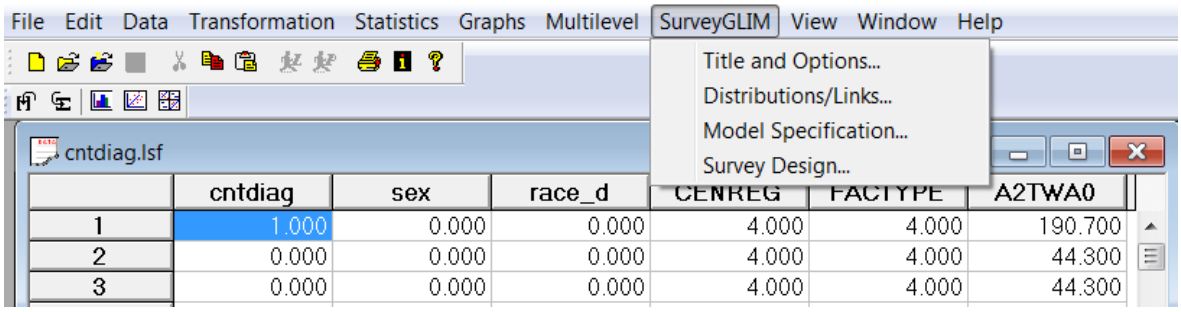
respectively where $\hat{P}(\text{cntdiag}_k \le l)$ denotes the estimated cumulative probability that at most the number of diagnoses listed in the first l categories are recorded for client k and $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the maximum likelihood estimates of α_1, α_2, α_3, β_1 and β_2 respectively.

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

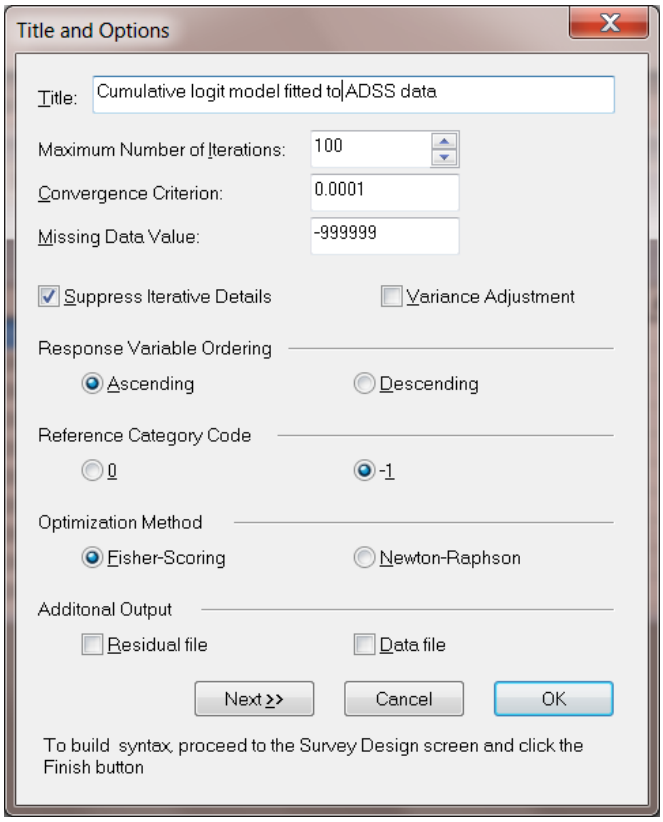
In a previous example, a GLIM with a Poisson distribution and a log link function was used to examine the possible association between ethnicity and gender effects and the number of substance abuse diagnoses (cntdiag). Since this variable assumes values between 0 and 3 in the sample data, an alternative 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

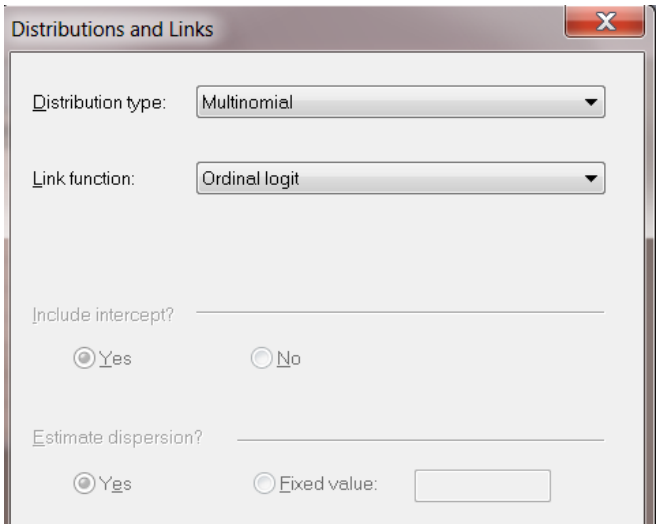
We start by opening the data file to be processed, **cntdiag.lsf**, in a LSF window as follows. Use the **Open** option on the **File** menu of the root window of LISREL to load the **Open** dialog box. Select the **LISREL Data (*.lsf)** option from the **Files of type** drop-down list box. Browse for and open the file **cntdiag.lsf**.



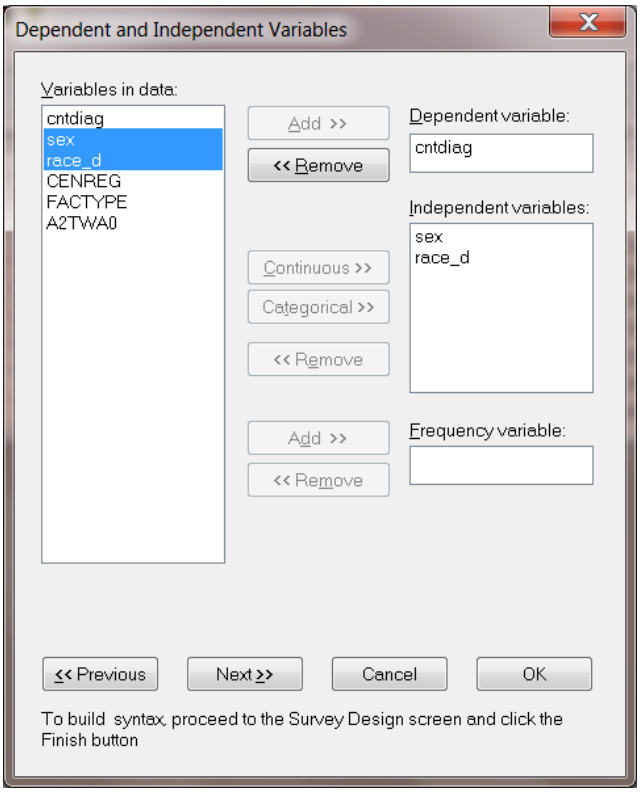
Next, we specify the analysis as follows. Select the **Title and Options** option on the **SurveyGLIM** menu to go to the **Title and Options** dialog box. Then enter the title **A cumulative logit model** into the **Title** string field to produce the following **Title and Options** dialog box.



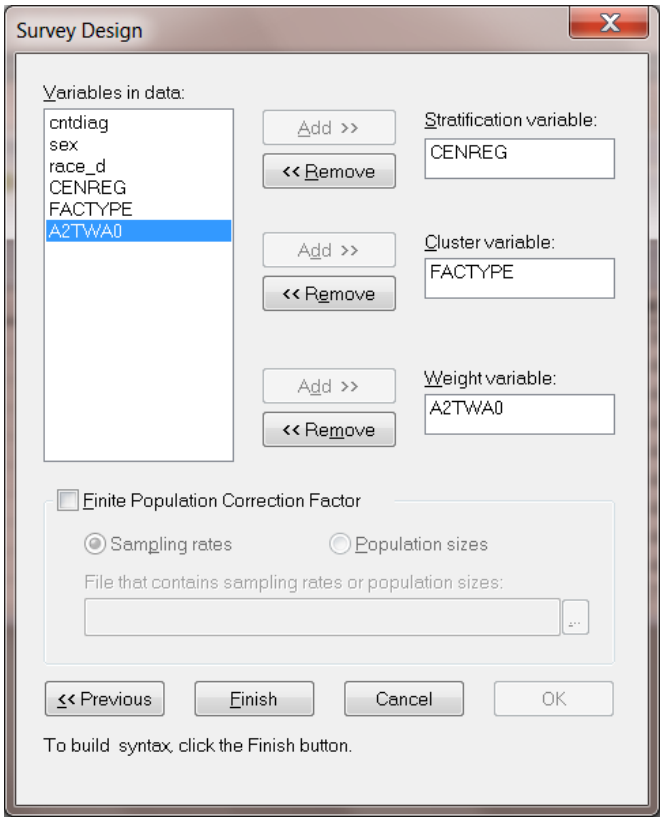
Click on the **Next** button to access the **Distributions and Links** dialog box, and select the **Multinomial** option from the **Distribution type** drop-down list box and the **Ordinal logit** option from the **Link function** drop-down list box to produce the following **Distributions and Links**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

cntdiag.PRL

```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Cumulative logit model fitted to ADSS data;
SY='C:\LISREL9 Examples\SGLIMEX\cntdiag.lsf';
Distribution=MUL;
Link=OLOGIT;
DepVar=cntdiag;
CoVars=sex race_d;
Stratum=CENREG;
Cluster=FACTYPE;
Weight=A2TWA0;
```

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	0.241973
Wald Chi-square	3.9993	2		0.241973

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	-1.6891	0.3154	-5.3554	0.0000
Alpha2	0.3493	0.1650	2.1172	0.0342
Alpha3	1.9046	0.1348	14.1297	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

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

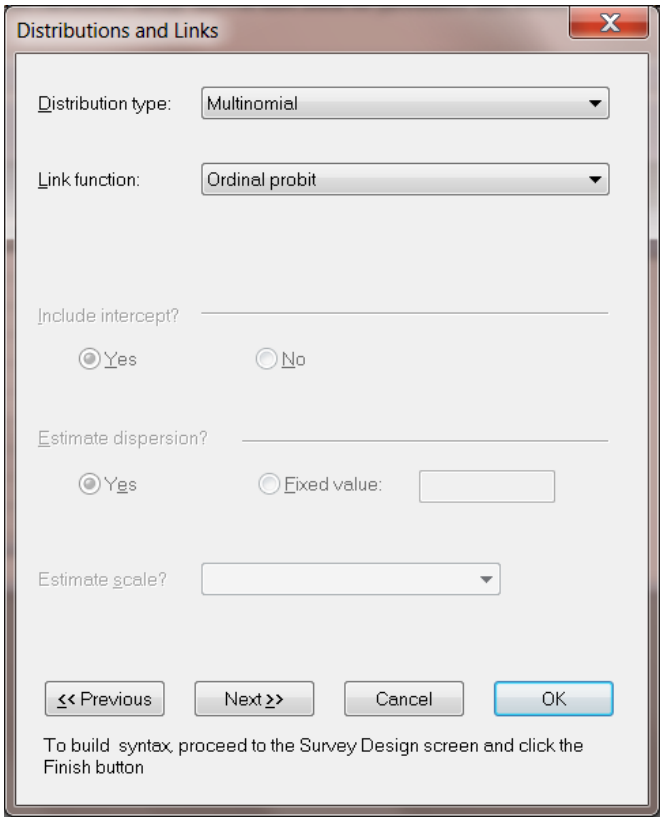
Similarly, the estimated probabilities that a white female client has at most 1 diagnosis and 2 diagnoses follow as 0.44 and 0.79 respectively. These estimated cumulative probabilities imply that the estimated probabilities that a white female client has 1 diagnosis, 2 diagnoses and 3 diagnoses are 0.44 - 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 $\exp\left(-0.20\right) = 0.82$ 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 $\exp\left(-0.39\right) = 0.68$ times the estimated probability for a nonwhite client.

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

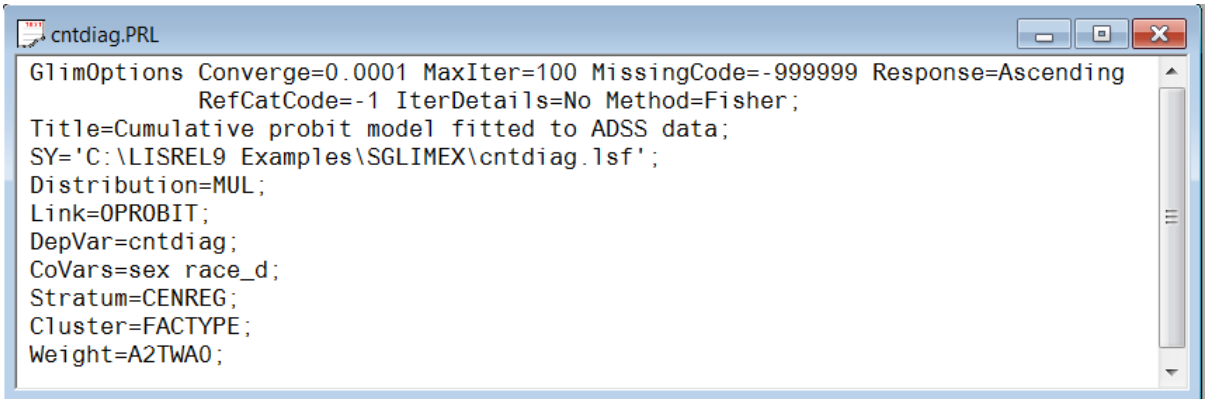
In the previous example we examined the strength of the relationship between ethnicity, gender, and the cumulative number of substance abuse diagnoses. A GLIM with a multinomial distribution and a cumulative logit link function was used to do so. To study the effect of using a different type of link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

Setting up the analysis

We fit the cumulative probit model to the data in **cntdiag.lsf** 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.



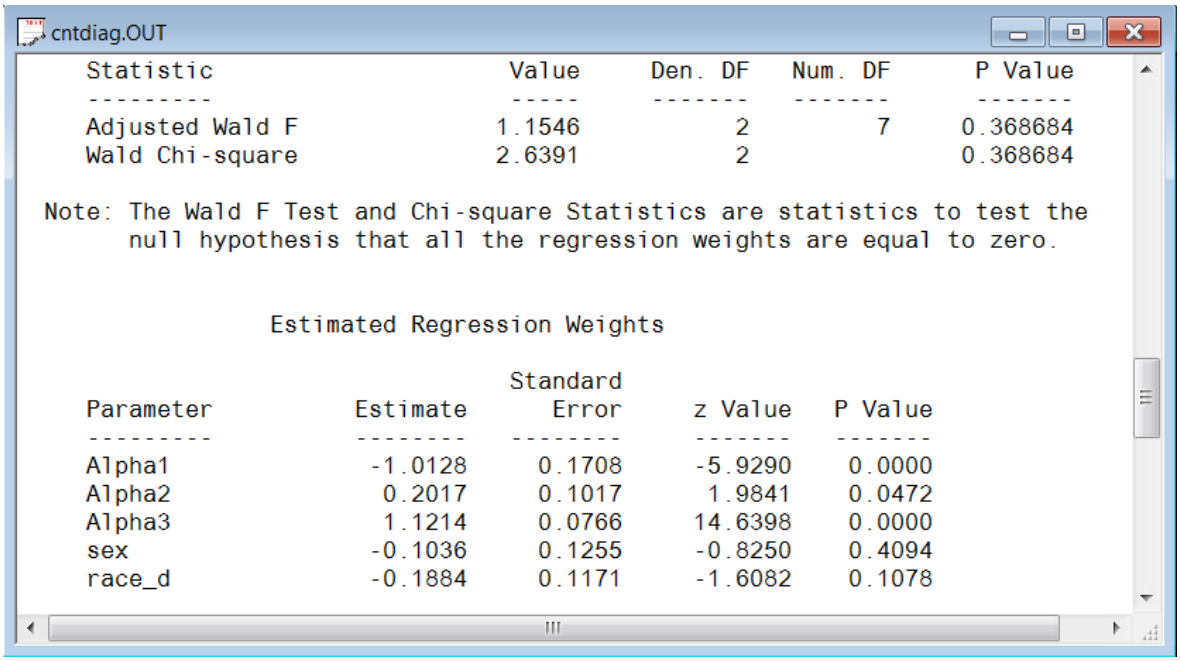
Since this concludes the modifications, click on the **Next** buttons of the **Distributions and Links** and the **Dependent and Independent Variables** dialog boxes and the **Finish** button of the **Survey Design** dialog box to open the following text editor window for **cntdiag.prl**.



Click on the **Run Prelis** toolbar icon to submit **cntdiag.prl** 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.



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 (race_d = 0, sex = 0) has no diagnoses follows from the results above as

$$\hat{\mathrm{P}}\left(\mathrm{cntdiag}_k = 1\right) = \hat{\mathrm{P}}\left(\mathrm{cntdiag}_k \le 1\right) = \Phi\left(-1.01\right) = 0.16$$

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 (race_d = 1, 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.