GLIMs for continuous responses

Contents

1. Introduction

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.

2. 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 (**meps.lsf**, **Generalized Linear Modeling examples** folder) is shown in the following LSF window.

meps.lsf

	TOTEXP99	PERWT99F	VARSTR99	VARPSU99	racex	Rsex	Rpovc99
1	7.9	14137.9	131.0	2.0	5.0	-1.0	3.0
2	8.8	17051.0	131.0	2.0	5.0	1.0	3.0
3	4.1	35737.5	131.0	2.0	5.0	-1.0	3.0
4	4.1	35862.7	131.0	2.0	5.0	-1.0	3.0
5	6.7	19407.1	131.0	2.0	5.0	1.0	3.0
6	5.8	18499.8	131.0	2.0	5.0	-1.0	3.0
7	6.5	18499.8	131.0	2.0	5.0	-1.0	3.0
8	8.1	22394.5	136.0	1.0	5.0	-1.0	3.0
9	8.0	27009.0	136.0	1.0	5.0	1.0	3.0
10	4.7	25108.7	136.0	1.0	5.0	-1.0	3.0
11	8.1	17569.8	136.0	1.0	5.0	-1.0	3.0
12	5.1	21478.1	136.0	1.0	5.0	-1.0	3.0
13	7.0	21415.7	136.0	1.0	5.0	1.0	3.0
14	0.0	12254.7	125.0	1.0	5.0	-1.0	5.0
15	0.0	17699.8	125.0	1.0	5.0	-1.0	5.0

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.

3. The models

The sampling distributions

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

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

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

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

and

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

respectively.

The mean models

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

$$\mu_k=\alpha+\beta_1x_{1k}+\beta_2x_{2k}+\ldots+\beta_rx_{rk}$$

while the mean model for the Gamma-log and Inverse Gaussian-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 respondent k, 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 two specific mean models are given by

$$\mathrm{E}\left[\mathrm{TOTEXP}_k\right]=\alpha+\beta_1x_{1k}+\beta_2x_{2k}+\beta_3x_{3k}+\beta_4x_{4k}+\beta_5x_{5k}+\beta_6x_{6k}$$

and

$$\mathrm{E}\left[\mathrm{TOTEXP}_k\right]=\exp\left(\alpha+\beta_1x_{1k}+\beta_2x_{2k}+\beta_3x_{3k}+\beta_4x_{4k}+\beta_5x_{5k}+\beta_6x_{6k}\right)$$

where $\mathrm{E}\left[\mathrm{TOTEXP}_k\right]$ 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 α, β_1, β_2, β_3, β_4, β_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

$$\frac{\exp\left(\alpha+\beta_1+\beta_5x_5+\beta_6x_6\right)}{\exp\left(\alpha+\beta_5x_5+\beta_6x_6\right)}=\exp\left(\beta_1\right).$$

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

$$\hat{\mathrm{E}}\left[\mathrm{TOTEXP}_k\right]=\hat{\alpha}+\hat{\beta}_1x_{1k}+\hat{\beta}_2x_{2k}+\hat{\beta}_3x_{3k}+\hat{\beta}_4x_{4k}+\hat{\beta}_5x_{5k}+\hat{\beta}_6x_{6k}$$

for the Normal-identity GLIM and as

$$\hat{\mathrm{E}}\left[\mathrm{TOTEXP}_k\right]=\exp\left(\hat{\alpha}+\hat{\beta}_1x_{1k}+\hat{\beta}_2x_{2k}+\hat{\beta}_3x_{3k}+\hat{\beta}_4x_{4k}+\hat{\beta}_5x_{5k}+\hat{\beta}_6x_{6k}\right)$$

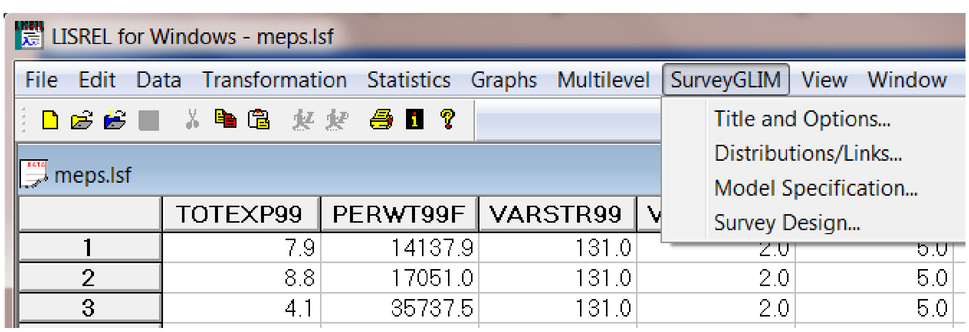
for the Gamma-log and Inverse Gaussian-log GLIMs respectively where $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_6$ denote the maximum likelihood estimates of α, β_1, β_2, …, β_6 respectively.

4. Analyzing normally distributed outcomes from complex survey designs

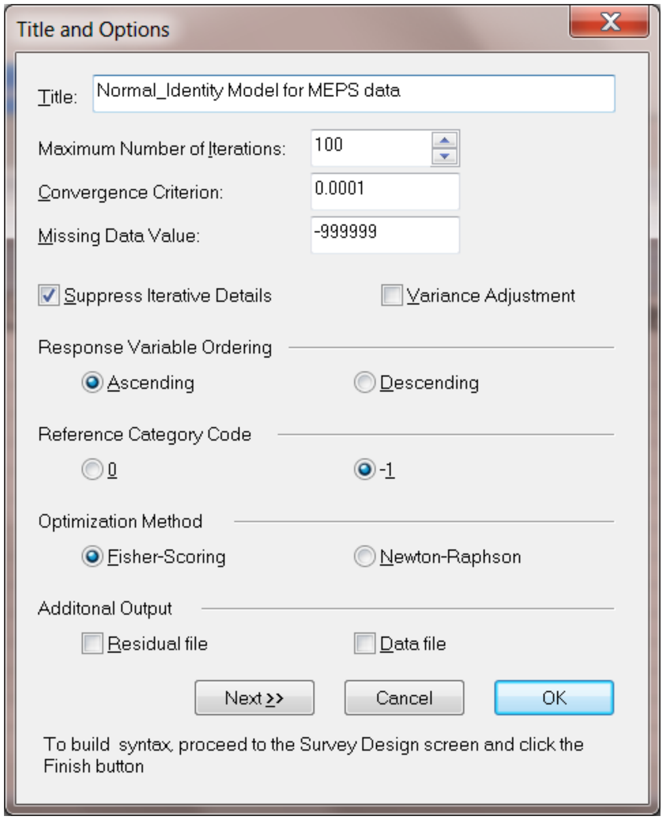
In this example, we are interested in exploring the linear relationship between a respondent's total health related expenditure and his/her ethnicity and gender. To make the assumption of normality more plausible, we use the natural logarithm of the total health care expenditure of the respondent during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

Setting up the analysis

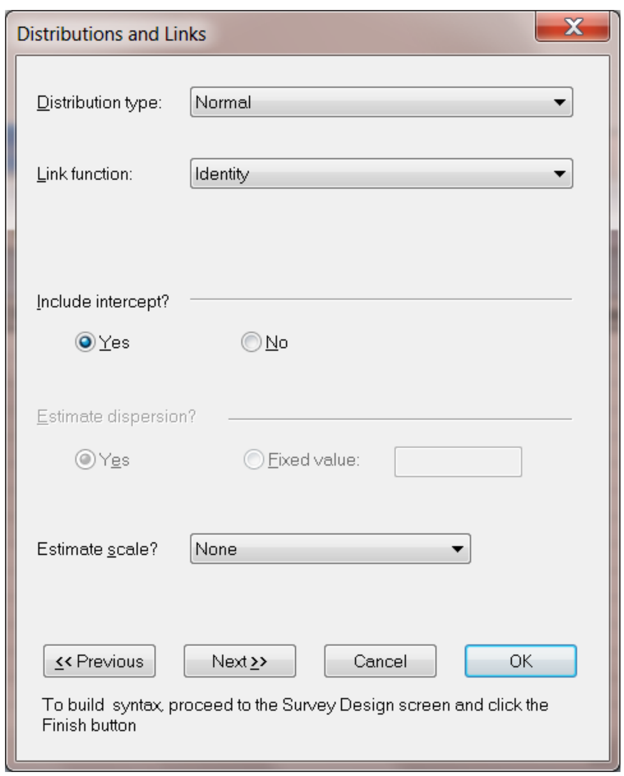
The first step is to open the file **meps.lsf** in a LSF window. This is done as follows. Use the **Open** option on the **File** menu of the root window of LISREL to load the **Open** dialog box and select the **Lisrel Data (*.lsf)** option from the **Files of type** drop-down list box. Browse for and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



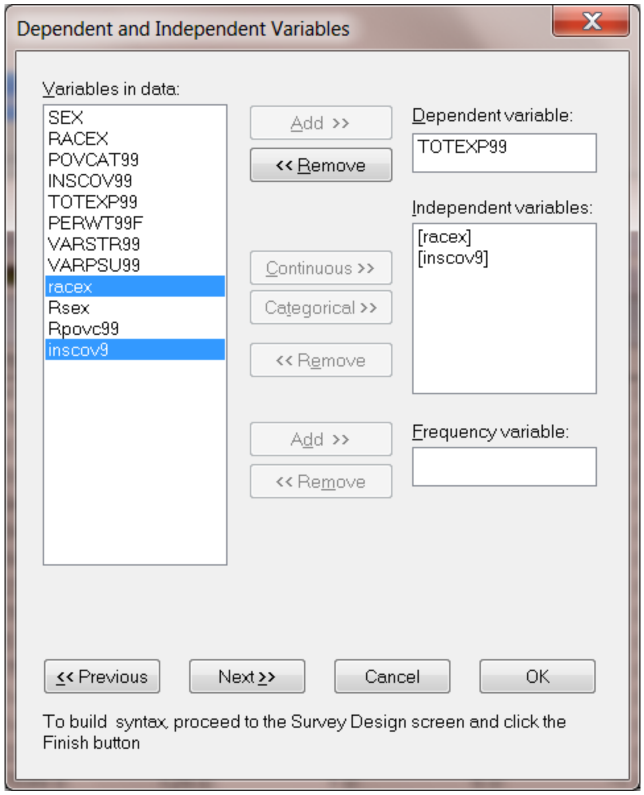
We are now ready to use the **SurveyGLIM** menu to fit the Normal-identity GLIM to the data in **meps.lsf**. Select the **Title and Options** option on the **SurveyGLIM** menu. Enter the descriptive title **A Normal-Identity Model for MEPS Data** into the **Title** string field to produce the following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



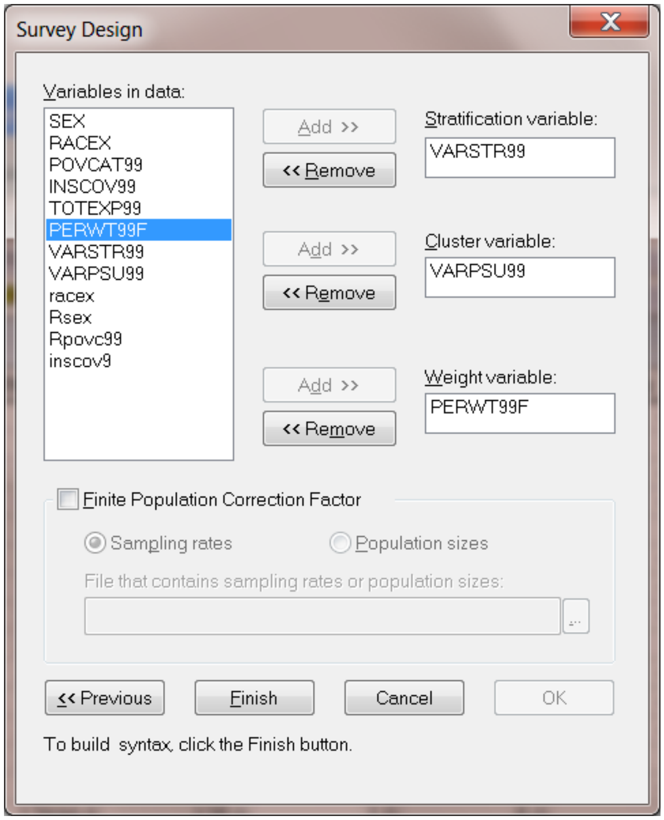
Since the default options will be used for this illustration, click on the **Next** button to go to the **Distributions and Links** dialog box.



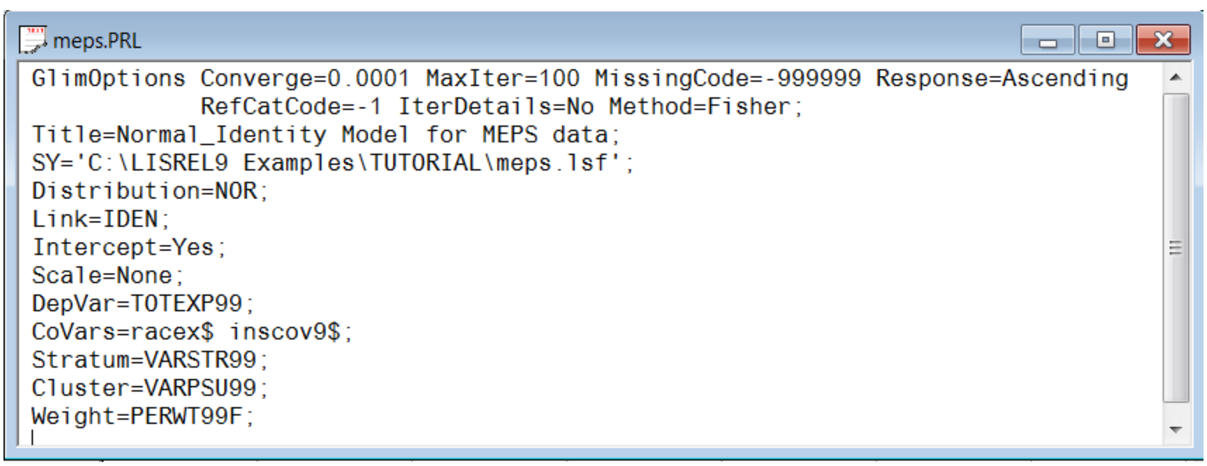
The default values are correct, so click on the **Next** button to go to the **Dependent and Independent Variables** dialog box. Specify the response variable, TOTEXP99, by selecting it from the **Variables in data** list box and then by clicking on the **Add** button of the **Dependent variable** section. 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.prl**.



```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Normal_Identity Model for MEPS data;
SY='C:\LISREL9 Examples\TUTORIAL\meps.lsf';
Distribution=NOR;
Link=IDEN;
Intercept=Yes;
Scale=None;
DepVar=TOTEXP99;
CoVars=racex$ inscov9$;
Stratum=VARSTR99;
Cluster=VARPSU99;
Weight=PERWT99F;
```

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.

meps.OUT

Statistic	Value	Den. DF	Num. DF	P Value
Adjusted Wald F	217.5296	6	312	0.000000
Wald Chi-square	1326.0937	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	4.5940	0.0746	61.5852	0.0000
racex1	0.0197	0.2135	0.0925	0.9263
racex2	0.1857	0.1972	0.9419	0.3462
racex3	-0.2684	0.1289	-2.0824	0.0373
racex4	-0.5333	0.0961	-5.5513	0.0000
inscov91	0.7308	0.0380	19.2073	0.0000
inscov92	0.9958	0.0481	20.7010	0.0000

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

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

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

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

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

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

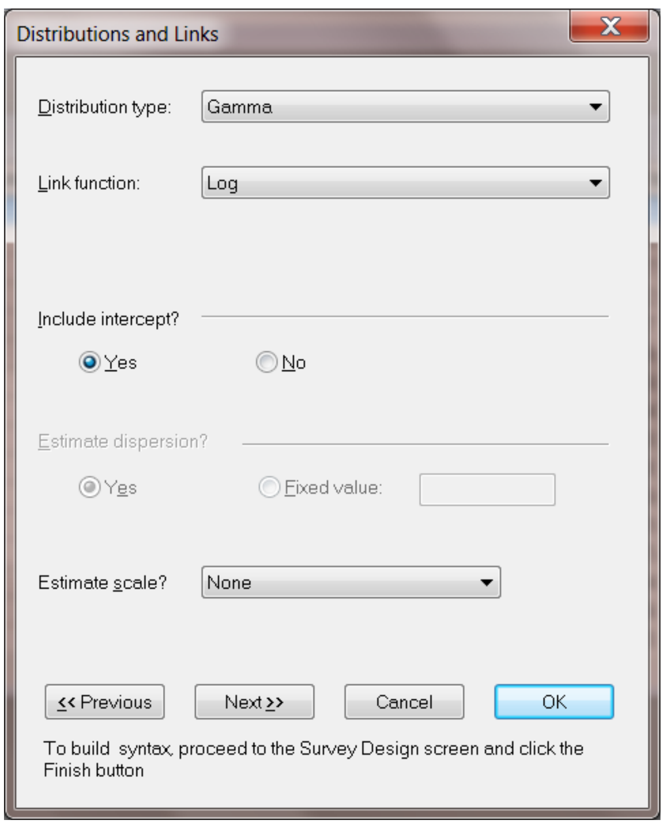
Likewise, for a White respondent with public insurance the corresponding estimate is $482.99. This estimate of average health care expenditures will only be accurate if the outcome variable has a normal distribution. An analysis that takes the strongly skewed distribution of health care expenditures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

5. 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.lsf** 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.prl**.

```
meps.PRL
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Gamma Model fitted to the MEPS data;
SY='C:\LISREL9 Examples\TUTORIAL\meps.lsf';
Distribution=GAM;
Link=LOG;
Intercept=Yes;
Scale=None;
DepVar=TOTEXP99;
CoVars=racex$ inscov9$;
Stratum=VARSTR99;
Cluster=VARPSU99;
Weight=PERWT99F;
```

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             129.7844            6        312      0.000000
     Wald Chi-square             791.1854            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

                                          Standard
     Parameter           Estimate          Error        z Value      P Value
     ---------           --------        --------       -------      -------
     intcept               1.4928          0.0169       88.3629       0.0000
     racex1                0.0098          0.0393        0.2498       0.8028
     racex2                0.0508          0.0465        1.0912       0.2752
     racex3               -0.0554          0.0286       -1.9398       0.0524
     racex4               -0.1194          0.0216       -5.5225       0.0000
     inscov91              0.1742          0.0091       19.0741       0.0000
     inscov92              0.2235          0.0106       21.1522       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

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

whereas for the log-link function

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

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

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

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

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

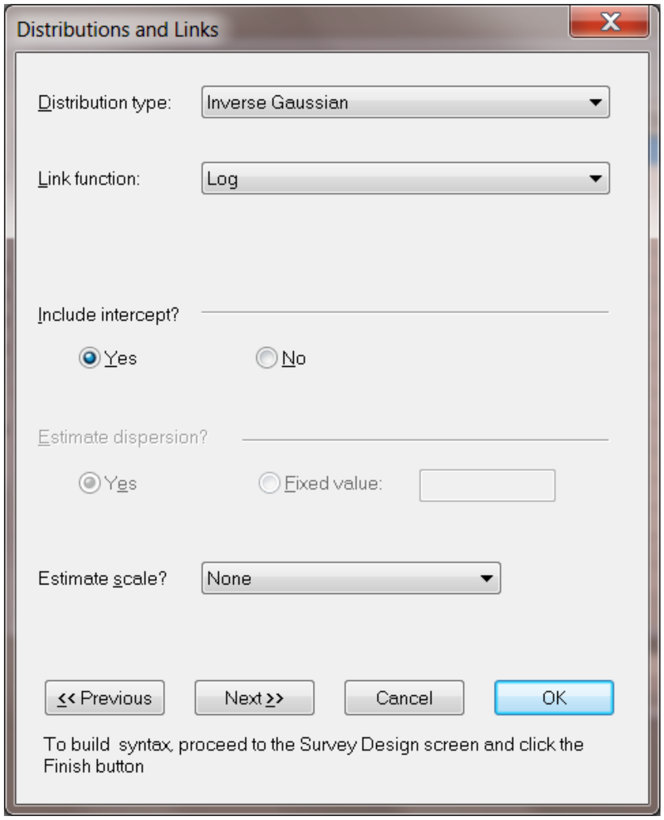
Similarly, the estimated mean health care expenditures for a White respondent with any private insurance and public insurance only follow as \$376.10 and \$509.73 respectively. The results above also indicate that $\exp\left(\hat{\beta}_4\right) = \exp\left(-0.12\right) = 0.88$ which implies that, on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

6. 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.prl**.

meps.PRL

```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Inverse Gauss Model fitted to the MEPS data;
SY='C:\LISREL9 Examples\TUTORIAL\meps.lsf';
Distribution=INVG;
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.

meps.OUT

```
    Statistic                    Value       Den. DF    Num. DF      P Value
    ---------                    -----       -------    -------      -------
    Adjusted Wald F            95.9255             6        312     0.000000
    Wald Chi-square           584.7765             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

                                      Standard
    Parameter           Estimate         Error       z Value    P Value
    ---------           --------      --------       -------    -------
    intcept               1.4956        0.0206       72.6701     0.0000
    racex1                0.0090        0.0401        0.2250     0.8220
    racex2                0.0615        0.0620        0.9924     0.3210
    racex3               -0.0577        0.0351       -1.6468     0.0996
    racex4               -0.1271        0.0263       -4.8320     0.0000
    inscov91              0.1729        0.0100       17.2380     0.0000
    inscov92              0.2238        0.0117       19.0716     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

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

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

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

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 $\exp\left(\hat{\beta}_2\right)=\exp\left(0.06\right)=1.06$ which implies that, on the average, American Indian respondents spent 6% more on health care in 1999 than other respondents. Similarly, it follows that $\exp\left(\hat{\beta}_5\right)=\exp\left(0.17\right)=1.19$ which implies that, on the average, respondents with any private insurance spent 19% more than other respondents on health care in 1999.