Scientific Software International

GLIMS for nominal responses using NHIS data

Contents

1. Introduction

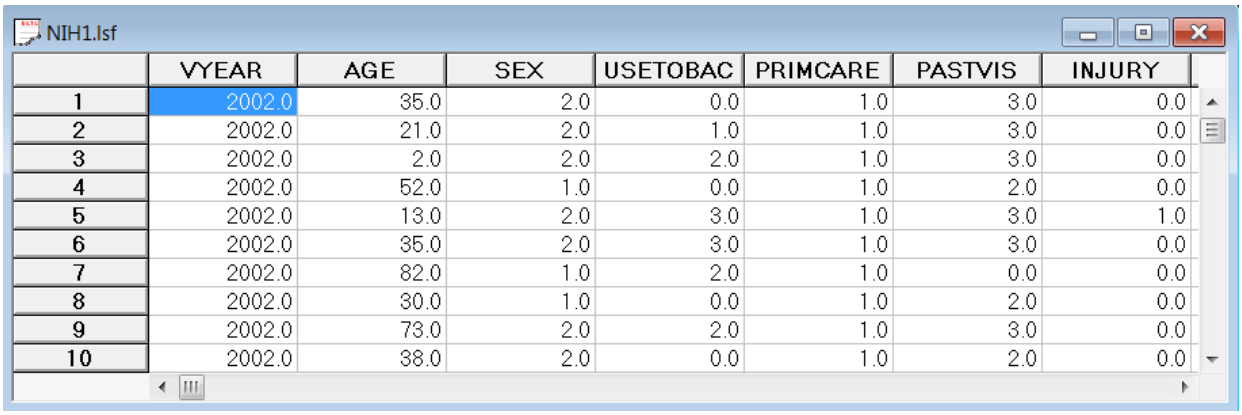
SurveyGLIM can also be used to fit models to nominal response variables. The primary food choice of alligators (fish, invertebrate, reptile, bird or other), smoking status (never smoked, former smoker or current smoker), preference for U.S. President (Democrat, Republican or Independent), cancer type of female cancer patients (breast, lung, brain, leukemia, liver, colon or other), etc. are examples of nominal response variables. In this section, we illustrate this feature by fitting a generalized logistic model to health-related data. A description of the data follows.

2. The data

The data set comes from the data library of the National Health Interview Survey (NHIS). The NHIS is a national longitudinal health survey. During 2002, background data and data on the health conditions of a sample of 28,737 participants were obtained. The 2002 sample was stratified into 64 strata and into 601 PSUs. The first portion of the data set to be used is shown in the following LSF window.

More information on the NHIS and the data are available at

http://www.cdc.gov/nchs/about/major/nhis/quest_data_related_1997_forward.htm



NIH1.lsf

	VYEAR	AGE	SEX	USETOBAC	PRIMCARE	PASTVIS	INJURY
1	2002.0	35.0	2.0	0.0	1.0	3.0	0.0
2	2002.0	21.0	2.0	1.0	1.0	3.0	0.0
3	2002.0	2.0	2.0	2.0	1.0	3.0	0.0
4	2002.0	52.0	1.0	0.0	1.0	2.0	0.0
5	2002.0	13.0	2.0	3.0	1.0	3.0	1.0
6	2002.0	35.0	2.0	3.0	1.0	3.0	0.0
7	2002.0	82.0	1.0	2.0	1.0	0.0	0.0
8	2002.0	30.0	1.0	0.0	1.0	2.0	0.0
9	2002.0	73.0	2.0	2.0	1.0	3.0	0.0
10	2002.0	38.0	2.0	0.0	1.0	2.0	0.0

The variables to be utilized in the subsequent analyses are

- CSTRATM is the stratum of the participant.
- CPSUM is the PSU of the participant.
- PATWT is the design weight of the participant.
- PASTVIS is the value of a nominal variable for the number of visits to a medical doctor during the past 12 months (1 for blank, 2 for none, 3 for 1-2 visits, 4 for 3-5 visits, 5 for 6 or more visits, 6 for unknown and 7 for not ascertained) of the participant.
- AGE is the age of the participant.
- EXERCISE is the value of a dummy variable for the exercise status (0 for do exercise and 1 for do not exercise) of the participant.

3. The models

The sampling distribution

The sampling distribution of the generalized logistic model is the Multinomial distribution whose probability density function is given by

$$f\left(\mathbf{y}_k,\boldsymbol{\pi}_k\right)=\frac{n!}{\left(\prod_{l=1}^{p-1}y_{kl}!\right)\left(n-\sum_{k=1}^{p-1}y_{ki}\right)!}\left(\prod_{l=1}^{p-1}\pi_{kl}^{y_{kl}}\right)\left(1-\sum_{k=1}^{p-1}\pi_{ki}\right)^{n-\sum_{k=1}^{p-1}y_{ki}}$$

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 client k responded with category l and $\boldsymbol{\pi}_k=[\pi_{k1},\pi_{k2},\ldots,\pi_{kp}]'$.

The probability model

The general probability model for the generalized logistic model is given by

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

where π_{kl} represents the probability that client k responded with category l, x_{jk} denotes the value of the j-th predictor ($j = 1, 2, \ldots, r$) for subject k and α_1, α_2, …, α_{p-1}, β_{11}, β_{12}, … , β_{1p-1}, …, β_{r1}, β_{r2}, …, and β_{rp-1} denote unknown parameters.

The probability model for the specific generalized logistic model is given by

$$\mathrm{P}\left(\mathrm{PASTVIS}_k = l\right) = \frac{\exp\left(\alpha_l + \beta_{1l}\mathrm{AGE}_k + \beta_{2l}\mathrm{EXERCISE}_k\right)}{1 + \sum_{l=1}^{6}\exp\left(\alpha_l + \beta_{1l}\mathrm{AGE}_k + \beta_{2l}\mathrm{EXERCISE}_k\right)} \quad l = 1, 2, \ldots, 6$$

where $\mathrm{P}\left(\mathrm{PASTVIS}_k = l\right)$ denotes the probability that client k responded with category l, and α_1, α_2, α_3, α_4, α_5, α_6, β_{11}, β_{12}, β_{13}, β_{14}, β_{15}, β_{16}, β_{21}, β_{22}, β_{23}, β_{24}, β_{25} and β_{26} denote unknown parameters.

The corresponding estimated probability model is given by

$$\hat{\mathrm{P}}\left(\mathrm{PASTVIS}_k = l\right) = \frac{\exp\left(\hat{\alpha}_l + \hat{\beta}_{1l}\mathrm{AGE}_k + \hat{\beta}_{2l}\mathrm{EXERCISE}_k\right)}{1 + \sum_{l=1}^{6}\exp\left(\hat{\alpha}_l + \hat{\beta}_{1l}\mathrm{AGE}_k + \hat{\beta}_{2l}\mathrm{EXERCISE}_k\right)} \quad l = 1, 2, \ldots, 6$$

where $\hat{\mathrm{P}}\left(\mathrm{PASTVIS}_k = l\right)$ is the estimated probability that client k responded with category l, and $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, $\hat{\alpha}_4$, $\hat{\alpha}_5$, $\hat{\alpha}_6$, $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, $\hat{\beta}_{13}$, $\hat{\beta}_{14}$, $\hat{\beta}_{15}$, $\hat{\beta}_{16}$, $\hat{\beta}_{21}$, $\hat{\beta}_{22}$, $\hat{\beta}_{23}$, $\hat{\beta}_{24}$, $\hat{\beta}_{25}$ and $\hat{\beta}_{26}$ denote the maximum likelihood estimates of α_1, α_2, α_3, α_4, α_5, α_6, β_{11}, β_{12}, β_{13}, β_{14}, β_{15}, β_{16}, β_{21}, β_{22}, β_{23}, β_{24}, β_{25} and β_{26} respectively.

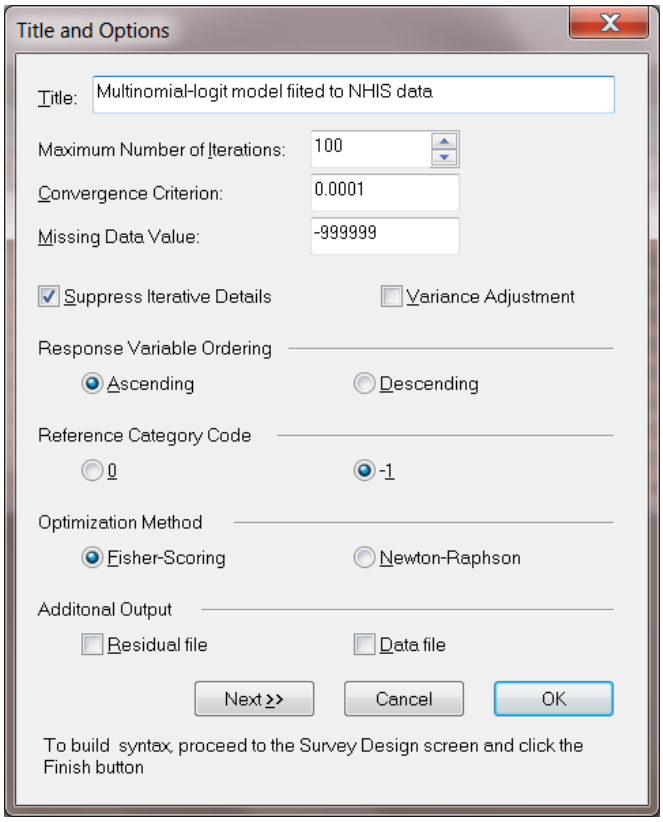
4. Analyzing nominal outcomes from complex survey designs

In this example, we wish to examine the effect of exercise and age on the number of visits (PASTVIS) to a medical doctor during the past 12 months. Since the last two categories of the outcome variable are defined as "unknown" and "not ascertained", PASTVIS is a nominal variable. A suitable GLIM model is obtained by specifying a multinomial distribution with logit link function.

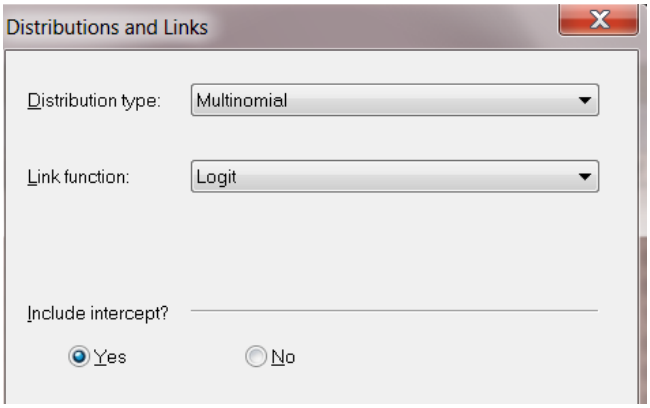
Setting up the analysis

Before the specific analysis can be specified, we need to open the file **nih1.lsf** in a LSF window. Next, click on the **SurveyGLIM** menu to produce the following LSF window.

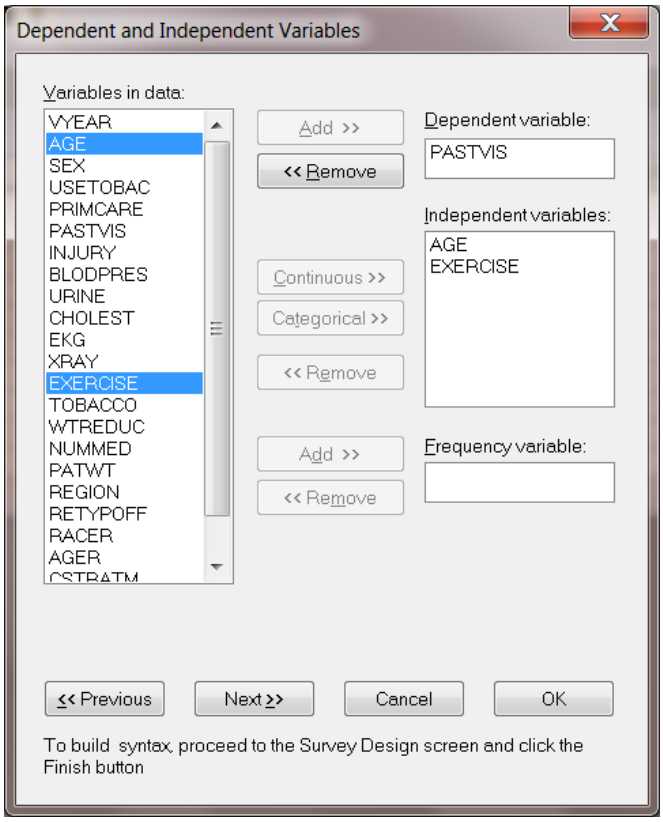
Continue by selecting the **Title and Options** option on the **SurveyGLIM** menu to access the **Title and Options** dialog box and entering the title **A Multinomial-Logit Model** into the **Title** string field to produce the following **Title and Options** dialog box.



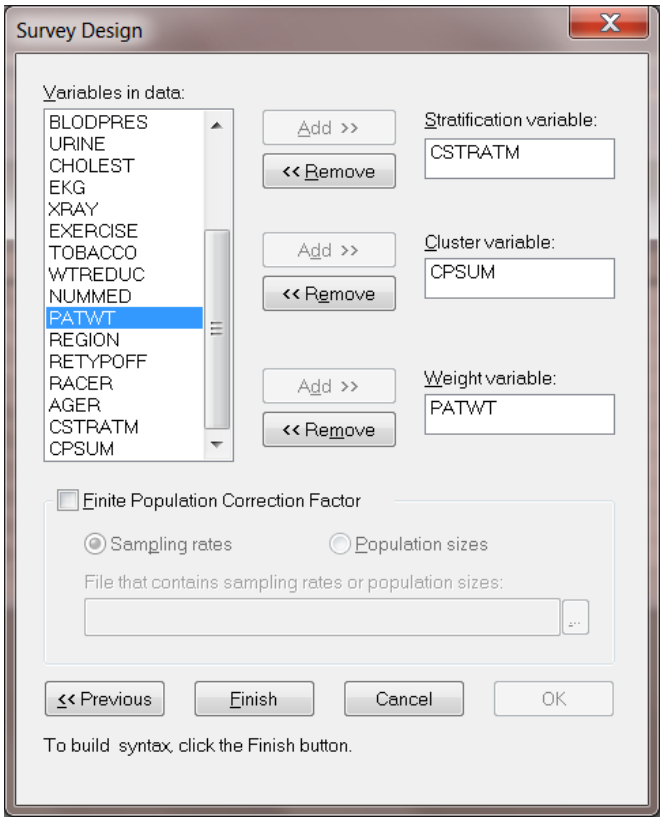
Go ahead and click on the **Next** button to access the **Distributions and Links** dialog box and select the **Multinomial** option from the **Distribution type** drop-down list box to produce the following **Distributions and Links** dialog box.



Click on the **Next** button above to go to the **Dependent and Independent Variables** dialog box. Specify the response variable, PASTVIS, 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, AGE and EXERCISE, 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 go to the **Survey Design** dialog box. Specify the stratification variable, CSTRATM, by selecting it from the **Variables in data** list box first and then clicking on the **Add** button of the **Stratification variable** section. Specify the cluster variable, CPSUM, and the weight variable, PATWT, by using the **Add** buttons of the **Cluster variable** and the **Weight variable** sections respectively to produce the following **Survey Design** dialog box.



This concludes the specifications, so click on the **Finish** button to open the following text editor window.

NIH1.PRL

```
GlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999 Response=Ascending
            RefCatCode=-1 IterDetails=No Method=Fisher;
Title=Multinomial logit model fitted to NHIS data;
SY='C:\LISREL9 Examples\SGLIMEX\NIH1.lsf';
Distribution=MUL;
Link=LOGIT;
Intercept=Yes;
DepVar=PASTVIS;
CoVars=AGE EXERCISE;
Stratum=CSTRATM;
Cluster=CPSUM;
Weight=PATWT;
```

Submit the syntax file above by clicking on the **Run Prelis** toolbar icon to obtain the corresponding output file **nih1.out**.

Discussion of results – generalized logistic model

A portion of the results in **nih1.out** is shown in the following text editor window.

NIH1.OUT

```
    Statistic                    Value     Den. DF    Num. DF      P Value
    ---------                    -----     -------    -------      -------
    Adjusted Wald F            39.1948          12        526     0.000000
    Wald Chi-square           480.1732          12                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           -1.1428     0.2505    -4.5626    0.0000
    intcept 2           -0.9696     0.1016    -9.5426    0.0000
    intcept 3            0.5666     0.0814     6.9599    0.0000
    intcept 4            0.3208     0.0907     3.5363    0.0004
    intcept 5            0.2328     0.1047     2.2246    0.0261
    intcept 6           -1.9704     0.3248    -6.0663    0.0000
    AGE 1               -0.0003     0.0048    -0.0652    0.9480
    AGE 2                0.0066     0.0018     3.6631    0.0002
    AGE 3                0.0055     0.0014     3.9140    0.0001
    AGE 4                0.0091     0.0015     5.9432    0.0000
    AGE 5                0.0091     0.0020     4.5676    0.0000
    AGE 6                0.0071     0.0056     1.2761    0.2019
    EXERCISE 1          -0.2636     0.2538    -1.0389    0.2989
    EXERCISE 2           0.3496     0.1846     1.8941    0.0582
    EXERCISE 3           0.2460     0.1217     2.0215    0.0432
    EXERCISE 4           0.2815     0.1170     2.4063    0.0161
    EXERCISE 5           0.0658     0.1658     0.3967    0.6916
    EXERCISE 6          -0.5404     0.3877    -1.3937    0.1634
```

Recall that AGE 1 represents the lowest category of the outcome variable, while AGE 6 represents the highest. At a 5% level of significance, the results above suggest that there is sufficient evidence that the age of a respondent exerts a positive influence on the probability of the number of visits to a doctor in the past 12 months by the respondent. In particular, it seems that older respondents are more likely than younger respondents to have visited a doctor more regularly in the past 12 months. The estimated coefficients for the EXERCISE variables are mostly positive, and a value of 1 on any of these indicates a patient that does not exercise. The results thus indicate that exercising exerts a significant influence on the probabilities of 1-2 and 3-5 annual visits to a doctor in the past 12 months. It appears that respondents who do not exercise are more likely than those who do exercise to have visited a doctor regularly in the past 12 months.

The estimated probability that a 60-year old respondent who does not exercise regularly does not visit the doctor (category 2) is obtained from the results above as

$$\hat{\mathrm{P}}\left(\mathrm{PASTVIS}_k = 2\right) = \frac{\exp\left(0.97 + 0.007 * 60\right)}{1 + \sum_{l=1}^{6} \exp\left(\hat{\alpha}_l + \hat{\beta}_{1l} * 60\right)} = 0.32$$

The corresponding probability that a 60-year old respondent who does exercise regularly does not visit the doctor (category 2) follows as

$$\hat{\mathrm{P}}\left(\mathrm{PASTVIS}_k = 2\right) = \frac{\exp\left(0.97 + 0.007 * 60 + 0.35\right)}{1 + \sum_{l=1}^{6} \exp\left(\hat{\alpha}_l + \hat{\beta}_{1l} * 60 + \hat{\beta}_{2l}\right)} = 0.35$$

The effect estimate for no visit to the doctor, $\hat{\beta}_{22} = 0.35$, suggests that the probability of no visit to the doctor increases for respondents who exercise regularly.