Paper 258-27

Performing Logistic Regression on Survey Data with the New SURVEYLOGISTIC Procedure

Anthony B. An, SAS Institute Inc., Cary, North Carolina, USA

Abstract

Categorical outcomes such as binary, ordinal, and nominal responses occur often in survey research. Logistic regression investigates the relationship between such categorical response variables and a set of explanatory variables. The LOGISTIC procedure can be used to perform a logistic analysis for data from a random sample. However, this approach is not valid if the data come from other sample designs, such as complex survey designs with stratification, clustering, and/or unequal weighting. In these cases, specialized techniques must be applied in order to produce the appropriate estimates and standard errors.

The SURVEYLOGISTIC procedure, experimental in SAS/STAT® , Version 9.0, brings logistic regression for survey data to the SAS® System and delivers much of the functionality of the LOGISTIC procedure. This paper describes the methodological approach and applications for this new software.

Introduction

Researchers use sample survey methodology to obtain information about a large aggregate or population by selecting and measuring a sample from the population. Categorical outcomes such as binary, ordinal, and nominal responses occur often in survey research. Logistic regression analysis is often used to investigate the relationship between these discrete responses and a set of explanatory variables. Discussions of logistic regression in sample surveys include Binder (1981, 1983), Roberts, Rao, and Kumar (1987), Skinner, Holt, and Smith (1989), Morel (1989), and Lehtonen and Pahkinen (1995).

Due to the variability of characteristics among items in the population, researchers apply scientific sample designs in the sample selection process to reduce the risk of a distorted view of the population, and they make inferences about the population based on the information from the sample survey data. In order to make statistically valid inferences for the population, they must incorporate the sample design in the data analysis. Several SAS procedures have been developed for analyzing survey data. The SURVEYSELECT procedure selects probability samples using various sample designs, including stratified sampling and sampling with probability proportional to size. The SURVEYMEANS procedure computes descriptive statistics for sample survey data, including means, totals, ratios, and domain statistics. The SURVEYREG procedure fits linear regression models and produces hypothesis tests and estimates for survey data.

The SURVEYLOGISTIC procedure, experimental in SAS/STAT, Version 9.0, brings logistic regression for survey data to the SAS System. PROC SURVEYLOGISTIC fits linear logistic regression models for discrete response survey data by the method of maximum likelihood. In the analyses, PROC SURVEYLOGISTIC incorporates complex survey sample designs, including designs with stratification, clustering, and unequal weighting.

The link functions in the regression can be the cumulative logit function (CLOGIT or PROPODD), generalized logit function (GLOGIT), probit function (PROBIT), or complementary log-log function (CLOGLOG). The maximum likelihood estimation of the regression coefficients is carried out with either the Fisher-scoring algorithm or the Newton-Raphson algorithm. Variances of the regression parameters and odds ratios are computed using a Taylor expansion approximation; see Binder (1983) and Morel (1989).

The SURVEYLOGISTIC procedure enables you to use categorical classification variables (also known as CLASS variables) as explanatory variables in an

explanatory model, using the familiar syntax for main effects and interactions employed in the GLM and LOGISTIC procedures.

The following sections discuss logistic regression in surveys and present an artificial example to illustrate how to use PROC SURVEYLOGISTIC. The syntax of PROC SURVEYLOGISTIC is given in Appendix A and the computation methods for variance estimation are described in Appendix B.

Logistic Modeling in Surveys

Logistic regression is often used to model the association of a categorical outcome with independent variables for survey data. For example,

- the National Health Interview Survey (NHIS) by the National Center for Health Statistics investigates the relationship between the health condition of the general population and demographic factors such as race, gender, and household income levels
- the Continuing Survey of Food Intakes by Individuals (CSFII) by the Human Nutrition Information Service measures the kinds and amounts of foods eaten by Americans and attitudes and knowledge about diet and health among Americans. Logistic regression can be used in modeling the levels of food sufficiency to food expenditures, participation in the Food Stamp program, and other food assistance programs

More examples of logistic regression in surveys can be found in Korn and Graubard (1999).

In logistic regression, probabilities of outcome categories are assumed to be a function of a linear combination of the explanatory variables. This function is also called a link function. Commonly used link functions are the cumulative logit function, generalized logit function, probit function, and complementary log-log function. PROC LOGISTIC in SAS/STAT is designed for such analyses.

PROC LOGISTIC computes statistics under the assumption that a sample is drawn from an infinite population by simple random sampling. However, most sample survey data are collected from a finite population with a probability-based complex sample design. In order to make statistically valid inferences for the population, the sample design should be incorporated in the data analysis. Thus, PROC SURVEYLOGISTIC is developed based on PROC LOGISTIC for logistic regression with survey data.

In PROC SURVEYLOGISTIC, the same iteration algorithms, Fisher-scoring or Newton-Raphson algorithm, are used to compute the maximum likelihood estimates of the regression coefficients for a logistic regression model, but the variance estimation is different from PROC LOGISTIC. PROC SURVEYLOGISTIC uses a Taylor expansion approximation and incorporates the sample design information, including stratification, clustering, and unequal weighting. PROC SURVEYLOGISTIC computes variances within each stratum and then pools the variance estimates together. An adjustment due to Morel (1989) is also used in the variance estimation to reduce the bias when the sample size is small. The finite population correction factor is included in the variance estimation if the sample is drawn without replacement.

The syntax of PROC SURVEYLOGISTIC is similar to PROC LOGISTIC. Additionally, you can use STRATA, CLUSTER, and WEIGHT statements in PROC SURVEYLOGISTIC to specify your sample design information. These statements are similar to other survey data analysis procedures such as PROC SURVEYMEANS and PROC SURVEYREG. See Appendix A for syntax details.

In addition to giving output similar to PROC LOGISTIC, PROC SURVEYLOGISTIC also displays the sample design information used in the analysis. For example, if you use the LIST option with the STRATA statement, the procedure will generate a table containing the summary for the strata in your data. Note that since PROC SURVEYLOGISTIC uses the Output Delivery System (ODS) to create output, the traditional OUTPUT statement of PROC LOGISTIC is not required in PROC SURVEYLOGISTIC.

Example

The following example illustrates how to use PROC SURVEYLOGISTIC to perform logistic regression for sample survey data. Note that PROC SURVEYLOGISTIC is experimental in Version 9.0; the syntax and results shown in this example are subject to change.

A market research firm conducts a survey among undergraduate students at the University of North Carolina at Chapel Hill (UNC) to evaluate three new Web designs at a commercial Web site targeting undergraduate students.

The sample design is a stratified sample where strata are students' classes. Within each class, 100 students were randomly selected using simple random sampling (without replacement). The total number of students in each class in the Fall semester of 2001 is

shown in the following table:

Class	Enrollment
Freshman	3,734
Sophomore	3,565
Junior	3,903
Senior	4,196

The total enrollment information is saved in the SAS data set **Enrollment**.

```
data Enrollment;
   input class _TOTAL_;
datalines;
1 3734
2 3565
3 3903
4 4196
;
```

Each student selected in the sample is asked to evaluate the three new Web designs and to rate them ranging from *dislike very much* to *like very much*:

1 dislike very much
2 dislike
3 neutral
4 like
5 like very much

The survey results are collected and shown in the following table, with the three different Web designs coded A, B and C.

Evaluation of New Web Designs

		Rating Counts				
Strata	Design	1	2	3	4	5
Freshman	A	10	34	25	16	15
	B	5	10	24	30	21
	C	11	14	20	34	21
Sophomore	A	19	12	26	18	25
	B	10	18	32	23	17
	C	15	22	34	9	20
Junior	A	8	21	23	26	22
	B	1	14	25	23	37
	C	16	19	30	23	12
Senior	A	11	14	24	33	18
	B	8	15	35	30	12
	C	2	34	27	18	16

These data are saved in the data set **WebSurvey**.

```
data WebSurvey;
  do class=1 to 4;
    do design=1 to 3;
      do rating=1 to 5;
        input counts @@;
        output;
      end;
    end;
  end;
datalines;
10 34 25 16 15
5 10 24 30 21
11 14 20 34 21
19 12 26 18 25
10 18 32 23 17
15 22 34 9 20
8 21 23 26 22
1 14 25 23 37
16 19 30 23 12
11 14 24 33 18
8 15 35 30 12
2 34 27 18 16
;

data WebSurvey; set WebSurvey;
  if class=1 then weight=3734/100;
  if class=2 then weight=3565/100;
  if CLASS=3 then weight=3903/100;
  if class=4 then weight=4196/100;
run;
```

The data set **WebSurvey** contains the variables **class**, **design**, **rating**, and **counts**. The variable **class**, the stratum variable, indicates four strata: freshman, sophomore, junior, and senior. The variable **design** specifies the three new Web designs: A, B, and C. The variable **rating** contains students' ratings for the new Web designs. The variable **counts** gives the frequency with which each Web design received each rating within each stratum. The variable **weight** contains the sampling weights, which are the reciprocals of selection probabilities in this example.

The following SAS statements define formats for these variables to reflect the meanings of their values.

```
proc format ;
  value CLASS 1='Freshman' 2='Sophomore'
              3='Junior' 4='Senior';
  value DESIGN 1='A' 2='B' 3='C';
  value RATING 1='dislike very much'
               2='dislike'
               3='neutral'
               4='like'
               5='like very much';
run;
```

If a sample is drawn without replacement and the sampling rate is not small enough to ignore, a finite population correction factor should be included in the analysis. For a complex sample design, sampling weights must be included in the analysis to assure an appropriate analysis.

The following SAS statements invoke PROC SURVEYLOGISTIC. The TOTAL= option specifies the data set **Enrollment** which contains the population totals in the strata. The population totals are used to calculate the finite population correction factor in the variance estimates. The response variable **rating** is ordinally scaled. A cumulative logit model is used to investigate the responses to the Web designs. In the MODEL statement, **rating** is the response variable and two indicator variables for design A and design B are explanatory variables with the design C as the reference level. Each **design** parameter compares a design to design C. Since the research firm is interested in a design that receives the most positive ratings, the DESCENDING option is specified. Because the sample design is stratified simple random sampling, the STRATA statement is used to specify the stratification variable **class**. The WEIGHT statement specifies the variable **weight** for sampling weights.

```
proc surveylogistic data=WebSurvey
  descending total=Enrollment;
  format class CLASS. design DESIGN.
         rating RATING.;
  stratum class;
  freq counts;
  class design (param=ref ref='C');
  model rating (order=data) = design;
  weight weight;
  ods select CumulativeModelTest OddsRatios;
run;
```

The output is shown in

```
                    The SAS System

            The SURVEYLOGISTIC Procedure

   Score Test for the Proportional Odds Assumption

        Chi-Square       DF     Pr > ChiSq

            8.4114        6         0.2095


              Odds Ratio Estimates

                     Point          95% Wald
 Effect           Estimate      Confidence Limits

 design A vs C       1.140       0.894       1.454
 design B vs C       1.604       1.255       2.052
```

.

The score chi-square for testing the proportional odds assumption is 8.4114, which is not significant with respect to a chi-square distribution with 6 degrees of freedom (p=0.2095). This indicates that the proportional odds model adequately fits the data.

The odds ratio for design A versus design C is 1.140 with a 95% confidence interval (0.894, 1.454). Therefore there is no significant preference between design A and design C. However, the odds ratio for design B versus design C is 1.604 with a 95% confidence interval (1.255,2.052). We can conclude that design B is significantly preferred among UNC undergraduate students over design A and design C.

Conclusion

When a complex sample design is used to draw a sample from a finite population, the sample design should be incorporated in the analysis of the survey data in order to make statistically valid inferences for the finite population. In addition to PROC SURVEYSELECT, PROC SURVEYMEANS, and PROC SURVEYREG, PROC SURVEYLOGISTIC is developed based on PROC LOGISTIC to model categorical outcomes in survey data using logistic regression. More features (for example, model selections) will be added to PROC SURVEYLOGISTIC, as well as more procedures for survey data analysis, in future releases of SAS.

Appendix A: Syntax

The following statements are available in PROC SURVEYLOGISTIC:

PROC SURVEYLOGISTIC < *options* > **;**
BY *variables* **;**
CLASS *variable* <*(v-options)*>
<*variable* <*(v-options)*>... > < / *v-options* >**;**
CLUSTER *variables* **;**
CONTRAST *'label' effect values*
$< effectvalues,...>$< / *options* >**;**
FREQ *variable* **;**
MODEL *events / trials* = < *effects* >
< / *options* >**;**
MODEL *variable* < *(variable_options)* > =
< *effects* >< / *options* >**;**
STRATA *variables* < / *options* > **;**
< *label:* > **TEST** *equation1*
< *equation2, . . .* >< / *option* >**;**
UNITS *independent1 = list1*
< *independent2 = list2 . . .* >< / *option* > **;**
WEIGHT *variable* </ *option* >**;**

The PROC SURVEYLOGISTIC and MODEL statements are required. The CLASS, CLUSTER, STRATA, and CONTRAST statements can appear multiple times. You should only use one MODEL statement and one WEIGHT statement. The CLASS statement (if used) must precede the MODEL statement, and the CONTRAST statement (if used) must follow the MODEL statement.

The PROC SURVEYLOGISTIC statement invokes the SURVEYLOGISTIC procedure and optionally identi-

fies input data sets and controls the ordering of the response levels. If your analysis includes a finite population correction factor, you can input either the sampling rate or the population total using the R= or N= option.

The CLASS statement names the classification variables to be used in the analysis.

The CLUSTER statement names variables that identify the clusters in a clustered sample design. The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata.

The CONTRAST statement provides a mechanism for obtaining customized hypothesis tests. It is similar to the CONTRAST statement in PROC LOGISTIC and PROC CATMOD, depending on the coding schemes used with any classification variables involved.

The *variable* in the FREQ statement identifies a variable that contains the frequency of occurrence of each observation. PROC SURVEYLOGISTIC treats each observation as if it appears n times, where n is the value of the FREQ variable for the observation.

The MODEL statement names the response variable and the explanatory effects, including covariates, main effects, interactions, and nested effects. If you omit the explanatory variables, the procedure fits an intercept-only model.

Two forms of the MODEL statement can be specified. The first form, referred to as *single-trial* syntax, is applicable to binary, ordinal, and nominal response data. The second form, referred to as *events/trials* syntax, is restricted to the case of binary response data. The *single-trial* syntax is used when each observation in the data set contains information on only a single trial; for instance, a single subject in an experiment. When each observation contains information on multiple binary-response trials, such as the count of the number of subjects observed and of the number responding, then the *events/trials* syntax can be used.

In the *events/trials* syntax, you specify two variables that contain count data for a binomial experiment. These two variables are separated by a slash. The value of the first variable, *events*, is the number of positive responses (or events). The value of the second variable, *trials*, is the number of trials. The values of both *events* and (*trials*$-$*events*) must be nonnegative and the value of *trials* must be positive for the response to be valid.

In the *single-trial* syntax, you specify one variable (on the left side of the equal sign) as the response variable. This variable can be character or numeric.

For both forms of the MODEL statement, explanatory *effects* follow the equal sign. You can use either continuous or classification variables. Classification variables can be character or numeric, and they must be declared in the CLASS statement. When an effect is a classification variable, the procedure enters a set of coded columns into the design matrix instead of directly entering a single column containing the values of the variable.

You use the LINK= option to specify the link function linking the response probabilities to the linear predictors. You can specify one of the following keywords. The default is LINK=LOGIT.

CLOGLOG — the complementary log-log function. PROC SURVEYLOGISTIC fits the binary complementary log-log model when there are two response categories, and fits the cumulative complementary log-log model when there are more than two response categories. Aliases: CCLOGLOG, CCLL, CUMCLOGLOG.

GLOGIT — the generalized logit function. PROC SURVEYLOGISTIC fits the generalized logit model where each nonreference category is contrasted with the reference category.

CLOGIT — the cumulative logit (or proportional-odds) function. PROC SURVEYLOGISTIC fits the binary logit model when there are two response categories, and fits the cumulative logit model when there are more than two response categories. Aliases: CUMLOGIT, PROPODD.

PROBIT — the inverse standard normal distribution function. PROC SURVEYLOGISTIC fits the binary probit model when there are two response categories, and fits the cumulative probit model when there are more than two response categories. Aliases: NORMIT, CPROBIT, CUMPROBIT.

The STRATA statement names variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample.

The STRATA *variables* are one or more variables in the input data set. These variables can be either character or numeric. The formatted values of the STRATA variables determine the levels.

The TEST statement tests linear hypotheses about the regression coefficients. The Wald test is used to jointly test the null hypotheses (H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$) specified in a single TEST statement.

Each *equation* specifies a linear hypothesis (a row of the $\mathbf{L}$ matrix and the corresponding element of the $\mathbf{c}$ vector); multiple *equations* are separated by commas. The label, which must be a valid SAS name, is used to identify the resulting output and should always be included. You can submit multiple TEST statements.

The form of an *equation* is as follows:

term < ±*term* . . . > < = ±*term* < ±*term* . . . >>

where *term* is a parameter of the model, or a constant, or a constant times a parameter. For a binary response model, the intercept parameter is named INTERCEPT; for an ordinal response model, the intercept parameters are named INTERCEPT, INTERCEPT2, INTERCEPT3, and so on. When no equal sign appears, the expression is set to 0.

The UNITS statement enables you to specify units of change for the continuous explanatory variables so that customized odds ratios can be estimated. An estimate of the corresponding odds ratio is produced for each unit of change specified for an explanatory variable. The UNITS statement is ignored for CLASS variables. If the CLODDS option is specified in the MODEL statement, the corresponding confidence limits for the odds ratios are also displayed.

The term *independent* is the name of an explanatory variable and *list* represents a list of units of change, separated by spaces, that are of interest for that variable. Each unit of change in a list has one of the following forms:

- *number*
- SD or −SD
- *number* * SD

where *number* is any nonzero number, and SD is the sample standard deviation of the corresponding independent variable. For example, $X = -2$ requests an odds ratio that represents the change in the odds when the variable X is decreased by two units. $X = 2*$SD requests an estimate of the change in the odds when X is increased by two sample standard deviations.

The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric. If you do not specify a WEIGHT statement, PROC SURVEYLOGISTIC assigns all observations a weight of 1.

Appendix B: Computation Method

Let Y be the response variable with categories $1, 2, \ldots, D, D+1$. The p covariates are denoted by a p-dimension row vector $\mathbf{x}$.

For a stratified clustered sample design, each observation is presented by a row vector,

$$(w_{hij}, \mathbf{y}'_{hij}, y_{hij(D+1)}, \mathbf{x}_{hij})$$

where

- $h = 1, 2, \ldots, H$ is the stratum number with a total of H strata
- $i = 1, 2, \ldots, n_h$ is the cluster number within stratum h, with a total of n_h clusters
- $j = 1, 2, \ldots, m_{hi}$ is the unit number within cluster i of stratum h, with a total of m_{hi} units
- w_{hij} denotes the sampling weight
- $\mathbf{y}_{hij}$ is a D-dimensional column vector whose elements are indicator variables for the first D categories for variable Y. If the response of the jth member of the ith cluster in stratum h falls in category d, the dth row of the vector is one, and the remaining elements of the vector are zero, where $d = 1, 2, \ldots, D$
- $y_{hij(D+1)}$ is the indicator variable for the $(D+1)$ category of variable Y
- $\mathbf{x}_{hij}$ denotes the k-dimensional row vector of explanatory variables for the jth member of the ith cluster in stratum h. If there is an intercept, then $x_{hij1} \equiv 1$.
- $\tilde{n} = \sum_{h=1}^{H} n_h$ is the total number of clusters in the entire sample
- $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total sample size

The following notations are also used in the following sections:

- f_h denotes the sampling rate for stratum h
- $\boldsymbol{\pi}_{hij}$ is the expected vector of the response variable

$$\begin{aligned} \boldsymbol{\pi}_{hij} &= E(\mathbf{y}_{hij}|\mathbf{x}_{hij}) \\ &= (\pi_{hij1}, \pi_{hij2}, \ldots, \pi_{hijD})' \\ \pi_{hij(D+1)} &= E(y_{hij(D+1)}|\mathbf{x}_{hij}) \end{aligned}$$

Note that $\pi_{hij(D+1)} = 1 - \mathbf{1}'\boldsymbol{\pi}_{hij}$ and $\mathbf{1}$ is a D-dimensional column vector whose elements are 1.

Let $g(\cdot)$ be a link function such that

$$\boldsymbol{\pi} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is a p-dimensional column vector for regression coefficients. The pseudo log likelihood is

$$L(\boldsymbol{\theta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\left((\log(\boldsymbol{\pi}_{hij}))'\mathbf{y}_{hij} + \log(\pi_{hij(D+1)})y_{hij(D+1)}\right)$$

Denote the maximum likelihood estimator as $\hat{\boldsymbol{\theta}}$, which is a solution to the estimating equations:

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{D}'_{hij}\left(\mathrm{diag}(\boldsymbol{\pi}_{hij}) - \boldsymbol{\pi}_{hij}\boldsymbol{\pi}'_{hij}\right)^{-1}(\mathbf{y}_{hij} - \boldsymbol{\pi}_{hij}) = \mathbf{0}$$

where $\mathbf{D}_{hij}$ is the matrix of partial derivatives of the link function f with respect to $\boldsymbol{\theta}$.

To obtain the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, the procedure uses iterations with a starting value $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}$. Let the lth step estimate be $\boldsymbol{\theta}^{(l)}$. The $(l+1)$th step estimate is

$$\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + {\mathbf{Q}^{(l)}}^{-1}\mathbf{R}^{(l)}$$

where

$$\mathbf{Q}^{(l)} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{D}^{(l)}_{hij}\left(\mathrm{diag}(\boldsymbol{\pi}^{(l)}_{hij}) - \boldsymbol{\pi}^{(l)}_{hij}{\boldsymbol{\pi}^{(l)}_{hij}}'\right)^{-1}{\mathbf{D}^{(l)}_{hij}}'$$

$$\mathbf{R}^{(l)} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{D}^{(l)}_{hij}\left(\mathrm{diag}(\boldsymbol{\pi}^{(l)}_{hij}) - \boldsymbol{\pi}^{(l)}_{hij}{\boldsymbol{\pi}^{(l)}_{hij}}'\right)^{-1}\left(\mathbf{y}_{hij} - \boldsymbol{\pi}^{(l)}_{hij}\right)$$

and $\mathbf{D}^{(l)}_{hij}$, $\boldsymbol{\pi}^{(l)}_{hij}$ are evaluated at $\boldsymbol{\theta}^{(l)}$.

The iterative scheme continues until the convergence is obtained using the convergence criterion.

By default, convergence requires that the normalized prediction function reduction is small. The iteration converges at the lth step if

$$\frac{\mathbf{g}(\boldsymbol{\theta}^{(l)})'\mathbf{H}(\boldsymbol{\theta}^{(l)})\mathbf{g}(\boldsymbol{\theta}^{(l)})}{L(\boldsymbol{\theta}^{(l)}) + 10^{-6}} < \epsilon$$

where $\mathbf{g}$ is the gradient vector and $\mathbf{H}$ is the negative (expected) Hessian matrix of the pseudo log-likelihood function. The convergence criterion ϵ is given by the GCONV= option, or by default, $\epsilon = 10^{-8}$.

Alternatively, the iteration scheme converges when the change in the log-likelihood function becomes very small at the $(l+1)$th step if

$$\frac{\left|L(\boldsymbol{\theta}^{(l+1)}) - L(\boldsymbol{\theta}^{(l)})\right|}{\left|L(\boldsymbol{\theta}^{(l)})\right| + 10^{-6}} < \epsilon$$

where the convergence criterion ϵ is set by the the FCONV= option.

Using Taylor approximation, the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{V}(\hat{\boldsymbol{\theta}}) = \widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{G}}\widehat{\mathbf{Q}}^{-1}$$

where

$$\widehat{\mathbf{Q}} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\widehat{\mathbf{D}}_{hij}\left(\mathrm{diag}(\hat{\boldsymbol{\pi}}_{hij}) - \hat{\boldsymbol{\pi}}_{hij}{\hat{\boldsymbol{\pi}}_{hij}}'\right)^{-1}\widehat{\mathbf{D}}'_{hij}$$

$$\widehat{\mathbf{G}} = \frac{n-1}{n-p}\sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(\mathbf{e}_{hi\cdot} - \bar{\mathbf{e}}_{h\cdot\cdot})'(\mathbf{e}_{hi\cdot} - \bar{\mathbf{e}}_{h\cdot\cdot})$$

$$\mathbf{e}_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij}\widehat{\mathbf{D}}_{hij}\left(\mathrm{diag}(\hat{\boldsymbol{\pi}}_{hij}) - \hat{\boldsymbol{\pi}}_{hij}{\hat{\boldsymbol{\pi}}_{hij}}'\right)^{-1}(\mathbf{y}_{hij} - \hat{\boldsymbol{\pi}}_{hij})$$

$$\bar{\mathbf{e}}_{h\cdot\cdot} = \frac{1}{n_h}\sum_{i=1}^{n_h}\mathbf{e}_{hi\cdot}$$

$\widehat{\mathbf{D}}_{hij}$ and $\hat{\boldsymbol{\pi}}_{hij}$ are evaluated at $\hat{\boldsymbol{\theta}}$.

Morel (1989) introduced an adjustment $\kappa\widehat{\mathbf{Q}}^{-1}$ to this variance estimator for the regression coefficients $\hat{\boldsymbol{\theta}}$:

$$\widehat{V}(\hat{\boldsymbol{\theta}}) = \widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{G}}\widehat{\mathbf{Q}}^{-1} + \kappa\lambda\widehat{\mathbf{Q}}^{-1}$$

where for given positive constants δ and ϕ,

$$\kappa = \max\left(\delta,\ p^{-1}(D+1)^{-1}\text{tr}\left(\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{G}}\right)\right)$$

and if $\tilde{n} - H + 1 > 3p(D+1) - 2$

$$\lambda = \frac{pD}{\tilde{n} - p(D+1) - H + 1}$$

if $\tilde{n} - H + 1 \leq 3p(D+1) - 2$, $\lambda = \phi$.

The adjustment $\kappa\lambda\widehat{\mathbf{Q}}^{-1}$ will

- reduce the small sample bias reflected in inflated Type I error rates
- guarantee a positive definite estimated covariance matrix provided that $\widehat{\mathbf{Q}}^{-1}$ exists
- be close to zero when the sample size becomes large

In this adjustment, κ is an estimate of the design effect, which has been bounded below by the positive constant δ. You can use the option DEFFBOUND=δ to specify this lower bound, and by default, the procedure uses $\delta = 1$. The factor λ converges to zero when the sample size becomes large, and λ has an upper bound ϕ. You can use the option ADJBOUND=ϕ to specify this upper bound, and by default, the procedure uses $\phi = 0.5$.

Generalized Logistic Model

Formulation of the generalized logit models for nominal response variables can be found in Agresti (1990). Without loss of generality, let the last category, $D+1$, be the reference category for the response variable Y. The link function for the generalized logistic model is defined as

$$\pi_{hijd} = \frac{e^{\mathbf{x}_{hij}\boldsymbol{\beta}_d}}{1 + \sum_{r=1}^{D} e^{\mathbf{x}_{hij}\boldsymbol{\beta}_r}}$$

where

$$\begin{aligned}\boldsymbol{\beta}_d &= (\beta_{hij1}, \beta_{hij2}, \ldots, \beta_{hijk})' \\ \boldsymbol{\theta} &= (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \ldots, \boldsymbol{\beta}_D')'\end{aligned}$$

for $d = 1, 2, \ldots, D$. The first partial derivatives matrix is

$$\mathbf{D}_{hij} = (\text{diag}(\boldsymbol{\pi}_{hij}) - \boldsymbol{\pi}_{hij}\boldsymbol{\pi}_{hij}{}') \otimes \mathbf{x}_{hij}'$$

where $\otimes$ denotes the Kronecker product. Evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$,

$$\mathbf{e}_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij}(\mathbf{y}_{hij} - \hat{\boldsymbol{\pi}}_{hij}) \otimes \mathbf{x}_{hij}'$$

Cumulative Logit Model

Details of the cumulative logit model, or proportional odds model, can be found in McCullagh and Nelder (1989). Denote the cumulative sum of the expected proportions for the first d categories of variable Y by

$$F_{hijd} = \sum_{r=1}^{d} \pi_{hijr}$$

for $d = 1, 2, \ldots, D$. Then the link function for the proportional odds model is

$$\log\left(\frac{F_{hijd}}{1 - F_{hijd}}\right) = \alpha_d + \mathbf{x}_{hij}\boldsymbol{\beta}$$

where

$$\begin{aligned}\boldsymbol{\beta} &= (\beta_1, \beta_2, \ldots, \beta_k)' \\ \boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_D)',\ \ \alpha_1 < \alpha_2 < \cdots < \alpha_D \\ \boldsymbol{\theta} &= (\boldsymbol{\alpha}', \boldsymbol{\beta}')'\end{aligned}$$

Define the D-dimensional column vector

$$\begin{aligned}\mathbf{q}_{hij} = (&F_{hij1}(1 - F_{hij1}), F_{hij2}(1 - F_{hij2}), \cdots, \\ &\cdots, F_{hijD}(1 - F_{hijD}))'\end{aligned}$$

Let U be the $D \times D$ matrix

$$U = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix}$$

The first partial derivatives matrix is

$$\mathbf{D}_{hij} = \begin{pmatrix} \text{diag}(\mathbf{q}_{hij})\mathbf{U} \\ \mathbf{q}_{hij}'\mathbf{U} \otimes \mathbf{x}_{hij}' \end{pmatrix}$$

where $\otimes$ denotes the Kronecker product. Evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$,

$$\mathbf{e}_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij} \begin{pmatrix} \text{diag}(\hat{\mathbf{q}}_{\text{hij}})\mathbf{U}\,(\text{diag}(\hat{\boldsymbol{\pi}}_{\text{hij}}))^{-1} \\ \boldsymbol{\tau}_{hij} \otimes \mathbf{x}_{hij}' \end{pmatrix} (\mathbf{y}_{hij} - \hat{\boldsymbol{\pi}}_{hij})$$

where

$$\boldsymbol{\tau}_{hij} = \hat{\mathbf{q}}_{hij}'\mathbf{U}(\text{diag}(\hat{\boldsymbol{\pi}}_{hij}))^{-1} + \hat{\pi}_{D+1}^{-1}\hat{q}_D\mathbf{1}'$$

Complementary log-log Model

For binary responses, $D = 1$. The complementary log-log is sometimes used as a link function

$$\log(-\log(1 - \pi_{hij1})) = \mathbf{x}_{hij}\boldsymbol{\beta}$$

and the parameter vector is

$$\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$$

The first partial derivatives matrix is

$$\mathbf{D}_{hij} = -(1 - \pi_{hij1}) \log(1 - \pi_{hij1}) \mathbf{x}'_{hij}$$

Evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$,

$$\mathbf{e}_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij} \frac{\log(1 - \hat{\pi}_{hij})(y_{hij1} - \hat{\pi}_{hij1})}{\hat{\pi}_{hij1}} \mathbf{x}'_{hij}$$

Probit Model

Another commonly used model for binary responses is the probit model:

$$\pi_{hij1} = \Phi(\mathbf{x}_{hij}\boldsymbol{\beta})$$

where

$$\Phi(z_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_0} e^{-\frac{1}{2}z^2} dz$$

is the cumulative distribution function of the standard normal distribution. The parameter vector is

$$\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$$

The first partial derivatives matrix is

$$\mathbf{D}_{hij} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\Phi^{-1}(\pi_{hij1})\right)^2} \mathbf{x}'_{hij}$$

Evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$,

$$\mathbf{e}_{hi\cdot} = \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{m_{hi}} w_{hij} e^{-\frac{1}{2}\left(\Phi^{-1}(\hat{\pi}_{hij1})\right)^2} \frac{y_{hij1} - \hat{\pi}_{hij1}}{\hat{\pi}_{hij1}(1 - \hat{\pi}_{hij1})} \mathbf{x}'_{hij}$$

References

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Binder, D. A. (1981), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *Survey Methodology*, 7, 157–170.

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279–292.

Korn, E. and Graubard B. (1999), *Analysis of Health Survey*, New York: John Wiley & Sons, Inc.

Lehtonen, R. and Pahkinen E. (1995), *Practical Methods for Design and Analysis of Complex Surveys*, Chichester: John Wiley & Sons, Inc.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman Hall.

Morel, G. (1989) "Logistic Regression under Complex Survey Designs," *Survey Methodology*, 15, 203–223.

Roberts, G., Rao, J. N. K., and Kumar, S. (1987), "Logistic Regression Analysis of Sample Survey Data," *Biometrika*, 74, 1–12.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, New York: John Wiley & Sons, Inc.

Acknowledgments

The author is grateful to Randy Tobias and Virginia Clark for their valuable assistance in the preparation of this paper.

Contact Information

Anthony B. An, Ph.D.
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513.
Phone (919)531-5879
FAX (919)677-4444
Email Anthony.An@sas.com