ESTIMATION AND ANALYSIS OF SURVEY DATA USING SAS PROCEDURES WESVAR, NASSREG, AND NASSLOG

Leyla Mohadjer, David Morganstein, Adam Chu, Mike Rhoads

1. INTRODUCTION

Standard methods of computing confidence intervals and analyzing statistical models requires the assumption that data are from simple random samples. This requirement is often not met in sample surveys since it is usually cost effective to select samples through a complex multi-stage design (e.g., involving stratification, clustering of units, and the use of several stages of selection) rather than through simple random sampling. When the standard methods of estimation and analysis are applied to data from complex sample designs the results can be misleading. For estimation and analysis of survey data from complex designs methods must be employed which reflect the sample design and any special procedures used in weighting. This paper describes an estimation and analysis computer package for use with survey data consisting of three SAS procedures WESVAR, NASSREG, and NASSLOG. All three procedures use the Balanced half-sample Repeated Replication (BRR) method to estimate the sampling errors of the survey estimates.

The package provides an effective tool for the data analyst. The user must be familiar with the sample design and the methods of computing sampling errors. Specifically, the user must determine: (a) the appropriate methods of defining half samples that, except for the sample size, simulate the original design and properly reflect all stages of sampling and estimation; (b) the number of half sample replications described for variance estimation and prepare the codes that define them; (c) the modifications necessary in the full sample estimation procedure that may be called for by samples half as large; and (d) the method of defining records and weights to be used for each of the half samples. An initial preprocessing operation is also needed to organize the data file as required for the package. At a minimum the data file must show the selection probability for each record; and it should contain sufficient information to enable assignment of records to half samples. Records comprising the half samples are defined by a matrix of replication codes to specify records that make up the half samples; the matrix of codes is supplied by the user.

The BRR method of variance estimation was chosen for its generality and its ease of use. Variances can be estimated for a wide variety of linear or non-linear statistics of interest. This paper does not discuss the BRR method in detail; the subject is treated in a number of publications. (See McCarthy, [5] and [6].)

The computations are implemented in SAS as procedures (like PROC FREQ, PROC MEANS, etc.) and they provide the user the power and flexibility of the SAS programming language as well as access to all other SAS procedures.

The WESVAR procedure computes survey estimates and their associated sampling errors using the BRR method. The procedure provides estimates and sampling errors for statistics involving user-specified transformations (for example, sums, ratios, differences, logarithms of ratios) of the survey variables and incorporates user specified ratio adjustments. PROC WESVAR is an updated version of PROC NASSVAR (Binzer, Morganstein [2]). It is more efficient than the NASSVAR procedure (includes more SAS statements and options, and requires less computer time to run), and can be used on VAX machines as well as IBM.

The NASSREG procedure applies weighted least squares to estimate the parameters of multiple regression models based on data from surveys employing complex designs. The sampling errors of model parameters are estimated by the BRR method to approximate the effects of the sample design and any special weighting adjustment used in estimation. The procedure also computes test statistics for testing the significance of a set of regression coefficients.

The NASSLOG procedure fits a logistic regression model to a binary dependent variable for data from complex surveys. Maximum-likelihood estimates of the parameters in the model are computed by the Newton-Raphson iteration model. The sampling errors of the model parameters are estimated by the BRR method. The procedure also computes test statistics for assessing the fit of the model. Brief descriptions of the three procedures are given in the following three sections.

2. THE WESVAR PROCEDURE

This section describes the WESVAR procedure. A brief summary of the functions included in the procedure is given in Section 2.1. Section 2.2 provides a list of statements that are used with the WESVAR procedure, and describes the function of each statement.

2.1 Introduction

The WESVAR procedure is designed to produce estimated totals, ratios and almost any other arithmetic function of weighted totals available in SAS (e.g., logodds ratio). It also computes associated sampling errors for these computed estimates. The procedure grammar is similar to that used in standard SAS procedures. The user must specify the variables for which estimates are to be computed, the weight variables to be used and optionally, any transformations. The full range of SAS arithmetic operators and functions may be used in specifying the computation of new variables from estimates computed by the procedure. Estimates and their associated statistics may be computed for any number of subgroups of the input file through the use of a "BY" statement. The procedure prepares the estimates as functions of weighted totals; however, the user specifies the weights to be employed when the procedure is invoked. WESVAR prepares estimates for each characteristic specified by the user within each of the half-sample replicates required for the sample design and also for the full sample. (The full sample is called "replicate zero".) The procedure computes sampling errors and variances for each given characteristic.

2.2 Specifications

The following statements are used with the WESVAR procedure:

```
PROC WESVAR options;

VAR variables;

WEIGHT variables;

BY variables;

COMPUTE newvariable=
          arithmeticexpression;

NEWLABEL newvariable='label';
```

The following options may be specified on the PROC statement:

DATA=SASdataset - names the SAS data set to be used as input.

OUTEST=SASdataset - requests that WESVAR create a new SAS data set containing estimates for the full sample and for each replicate.

OUTSTAT=SASdataset - requests that WESVAR create a new SAS data set containing full-sample estimates and sampling error statistics.

NOPRINT - suppresses the normal printed output.

PRINTREP - requests that WESVAR print the estimates for each replicate.

FPC=n - specifies a finite population correction factor to be used in calculating the sampling error statistics. The value specified for FPC must be greater than or equal to zero and less than one.

ALPHA=p - specifies the alpha level for confidence intervals. The value specified for ALPHA must be greater than 0 and less than or equal to .20. If no ALPHA level is given, .05 is used.

Statistics are calculated for each numeric variable listed on the VAR statement. If a VAR statement is not used, all numeric variables on the input data set are analyzed. Except for those listed in a BY statement or WEIGHT statement.

The WEIGHT statement supplies a list of SAS variables that contain the weights to be used for the analysis. The variable containing the full-sampling weight must be listed first, followed by the variables containing the replicate weights. The WEIGHT statement, which is required, must include at least two variables.

A BY statement may be used with PROC WESVAR to obtain separate analyses on observations defined by the BY variables.

The COMPUTE statement allows the user to obtain estimates for "computed" variables - i.e., transformed values of one or more numeric variables on the input data set. The expression may include any of the five standard arithmetic operators: addition, subtraction, multiplication, division, and raise to a power.

The NEWLABEL statement allows the user to supply variable labels for variables created in COMPUTE statements.

3. THE NASSREG PROCEDURE

This section includes a brief summary of the NASSREG procedure. Section 3.2 describes the estimation method used in PROC NASSREG. The methodology for hypothesis testing is given in Section 3.3. Section 3.4 provides the specifications for setting up the SAS statements.

3.1 Introduction

Suppose that a response (dependent) variable Y can be expressed as a linear combination of p regressor (independent) variables; X1,X2, ..., XP; i.e., for i=1,2,...,n,

$$Y_i = \beta_0 + \beta_1 X1_i + \ldots + \beta_p XP_i + \varepsilon_i$$

where n is the sample size and ε_i is a random error. In matrix notation, the model can be written as

$$Y = X\beta + \varepsilon \ , \qquad (1)$$

where Y is the nx1 column vector of observations on Y, β is the (p+1)x1 column vector of regression coefficients, X is the nx(p+1) design matrix, and ε is the nx1 column vector of random errors.

Standard methods of analyzing the model described by equation (1) such as ordinary least squares (OLS) assume that the observations are from a simple random sample. That is, the error variables have a variance-covariance structure described as an unknown constant times the identity matrix. NASSREG applies the BRR method to estimate the sampling errors of the model parameters, and provides tests of hypotheses that take account of the sample design used.

For example, suppose that Y can be predicted by a linear combination of X1 and X2. Then the model is of the form

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \varepsilon_i \ . \qquad (2)$$

To fit this model with the NASSREG procedure, specify:

```
PROC NASSREG ;
MODEL Y = X1 X2 ;
WEIGHT W0-Wk ;
```

where W0 is the full sample weight and W1-Wk represent the k replicate weights.

NASSREG computes estimates of the regression coefficients, the variance-covariance matrix of the estimated model parameters, and the square of the multiple correlation coefficient (coefficient of determination). In addition, it provides a test of the overall significance of the fitted regression model, and can be used to test the significance of a specified subset (or linear combination) of variables included in the model.

3.2 Estimation

NASSREG uses the principle of weighted least squares to produce an estimate of β for the full sample and corresponding estimates for each of the replicate samples. The replicate estimates are used to obtain the approximate sampling errors of the estimated model parameters. Essentially, the variance of the estimated model parameters is computed as a function of the sum of squares of the deviations between the full and the replicate estimates.

3.3 Testing Hypotheses

For a given regression model, NASSREG will automatically provide a test of the overall fit of the model. The user can also specify tests for the significance of a subset or linear combination of variables included in the model through the TEST statement. In general, hypotheses of the form $H_0{:}C\beta=\delta$ versus $H_1{:}C\beta\neq\delta$ can be tested in NASSREG, where C is a matrix of known values, and δ is a vector of constants (usually equal to 0). For example, to test

$$H_0{:}\beta_1=\beta_2= \ldots =\beta_p= 0$$

in the model,

$$Y_i = \beta_0 + \beta_1 X1_i + \ldots + \beta_p XP_i + \varepsilon_i,$$

we define

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

and δ to be a vector of zeros.

To test $H_0{:}a\beta_1+b\beta_2=d$ in the model $Y_i= \beta_0+ \beta_1 X1_i + \beta_2 X2_i + \varepsilon_i$, we define C = [0 a b] and δ = d. This can be specified in the TEST statement as

```
TEST A*X1 + B*X2 = D.
```

Missing Values

Observations having missing values for the full-sample weight, the dependent variable, or any of the independent variables are excluded from all estimates. All other observations are included in the full-sample estimates and in the estimates for each replicate for which the observation has a positive weight. All non-positive values for weight variables (whether zero, missing, or negative) are treated as missing.

Cautions

The t and F statistics computed by NASSREG are asymptotic approximations that depend on the degrees of freedom associated with the sample variances and covariances of the estimated regression coefficients. The maximum degrees of freedom is equal to the number of replicates used in the calculation of the variance estimates. NASSREG uses the number of replicates as the degrees of freedom associated with the various statistical tests. However, the actual degrees of freedom will probably be smaller than the number of replicates; and will vary depending on the particular survey items included in the regression model. Consequently, the t and F tests computed by the procedure are approximately valid only if the number of replicates is approximately equal to the actual degrees of freedom. The F statistic for the overall fit of a model can be computed only when the number of sample replicates is larger than the number of independent variables in the model. If this is not the case, NASSREG prints a warning message and sets the F value equal to missing.

3.4 Specifications

The following statements are used with the NASSREG procedure:

```
PROC NASSREG options;

MODEL  dependent_variable =
       independent_variables/options;

WEIGHT   fullsamp_weight_variable
         replicate_weight_variables;

TEST  linear_equation,...;
```

The PROC NASSREG statement must be accompanied by exactly one MODEL statement and exactly one WEIGHT statement. In addition, any number of TEST statements may follow the MODEL statement. The following options may be specified on the PROC statement:

DATA=SASdataset - names the SAS data set to be used.

OUTEST=SASdataset - requests that NASSREG create a new SAS data set containing regression coefficients for the full sample and for each replicate.

OUTSTAT=SASdataset - requests that NASSREG create a new SAS data set containing full-sample regression coefficients and test statistics.

COVOUT - outputs the covariance matrix for the parameter estimates to the OUTEST data set. This option is valid only if OUTEST= is also specified.

The MODEL statement specifies the variables to use in the analysis. The options below may appear in the MODEL statement after a slash (/). If no options are specified, the slash may be omitted.

NOPRINT - suppresses the normal printed output.

PRINTREP - requests that regression coefficients for each replicate be printed.

NOINT - suppresses the intercept term that is normally included in the model.

The WEIGHT statement is similar to the one used in PROC WESVAR.

The TEST statement tests hypotheses about the parameters estimated in the MODEL statement.

NASSREG computes an approximate F statistic for the joint hypothesis specified in a single TEST statement. More than one TEST statement may accompany a MODEL statement. If the hypothesis can be stated by

$$H_0 : C\beta = \delta ,$$

then the approximate F statistic is given by

$$F = \frac{k+1-c}{k.c}\,(Cb-\delta)'(C\ \widehat{Var}(b)\ C')^{-1}\,(Cb-\delta),$$

where

k = number of replicates

c = Rank(C)

b = least squares estimate of β

$\widehat{Var}(b)$ = The estimated variance-covariance matrix of b obtained by balanced repeated replication.

Examples of valid TEST statements are:

```
                     MODEL Y = X1 X2 ;
     XSUM:           TEST X1+X2=1;
     OVERALL1:       TEST X1=0 , X2=0 ;
     OVERALL2:       TEST X1 , X2 ;
```

The last two tests are equivalent; since no constant is specified after the equal sign, zero is assumed. For the first TEST statement, XSUM,

$$C = [0\ 1\ 1] \text{ and } \delta = 1 .$$

For the second and third TEST statements, OVERALL1 and OVERALL2,

$$C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \delta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Note that C and δ are determined by the row in the TEST statement and need not be explicitly specified by the user.

4. THE NASSLOG PROCEDURE

A summary of the NASSLOG procedure is given in this section. Section 4.1 provides a general definition of logistic models. Section 4.2 contains information on estimation and hypothesis testing in PROC NASSLOG. Specifications for setting up the SAS statements are given in Section 4.3.

4.1 Introduction

Suppose that a response (dependent) variable Y can take one of the two values 0 or 1. Variables of this type are called binary or dichotomous variables. For dichotomous variables such as Y, one object is to develop a method for estimating P, where P is the probability of occurrence of an event as a function of a number of independent variables. Let Y_i denote the observed value of Y for the i-th individual in the sample and, designate the corresponding vector of p regressor (independent) variables X_i, i=1,2,...,n, where

$$X_i' = (X0_i, X1_i, \ldots, XP_i).$$

The dummy regressor $X0_i = 1$ (i=1,...,n) is included to provide for the estimation of an intercept. Also define the nx(p+1) matrix of independent variables for the sample as X, where

$$X' = (X_1, X_2, \ldots, X_n).$$

Then, under the logistic regression model, the probability that Y_i is equal to 1 is assumed to be

$$P_i = 1/(1+\exp(-\beta' X_i)), \qquad (1)$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is the vector of regression coefficients in the logistic model. We can represent equation (1) in the following way

$$Y_i = 1/(1+\exp(-\beta' X_i)) + \varepsilon_i, \qquad (2)$$

where ε_i is a random error with mean zero and variance $P_i(1-P_i)$. Refer to Kleinbaum, Kupper and Morgenstern (1982), pages 421-446, for more information about logistic regression models.

Standard methods of analyzing the model described by equation (2), such as the method used in PROC LOGIST in SAS, assume that the observations are from a simple random sample. NASSLOG computes "weighted" maximum-likelihood estimates of the parameters and applies the balanced repeated replication method to estimate the sampling errors of the model parameters. NASSLOG also provides tests of hypotheses that take account of the sample design.

For example, suppose that Y can be predicted by a combination of X1 and X2. Then the probability that Y_i is equal to 1 is

$$P_i = 1/(1+\exp(-\beta_0 - \beta_1 X1_i - \beta_2 X2_i))$$

or

$$Y_i = 1/(1+\exp(-\beta_0 - \beta_1 X1_i - \beta_2 X2_i)) + \varepsilon_i.$$

To fit this model with the NASSLOG procedure, specify:

```
PROC NASSLOG;
MODEL Y=X1 X2;
WEIGHT W0-WK;
```

where W0 is the full sample weight and W1-WK represents the k replicate weights.

NASSLOG computes estimates of the regression coefficients, the variance-covariance matrix of the estimated model parameters, and an R^2 statistic which is similar to the square of the multiple correlation coefficient (coefficient of determination) in the usual regression model. It also provides a test of the overall significance of the fitted logistic model, and the significance of individual parameters included in the model.

4.2 Estimation and Hypothesis Testing

PROC NASSLOG requires an input file containing the dependent and independent variables, the full sample weights, and the precomputed half-sample weights. NASSLOG applies the Newton-Raphson iteration procedure to obtain the weighted maximum-likelihood estimates of β for the full sample and the corresponding estimates for each of the replicate samples. The replicate estimates are used to obtain the approximate sampling errors of the estimated model parameters.

For a given logistic regression model, NASSLOG will automatically provide test statistics for individual parameters in the model and the overall fit of the model. Missing values are treated the same way as PROC NASSREG. The same cautions about the t and F statistics also apply to the NASSLOG procedure.

4.3 Specifications

The following statements are used with the NASSLOG procedure:

```
PROC NASSLOG options;

MODEL dependent_variable =
       independent_variables/options;

WEIGHT  fullsamp_weight_variable
          replicate_weight_variables;

INITIAL constants;

PRINTIT replicates;
```

The PROC NASSLOG statement must be accompanied by exactly one MODEL statement and exactly one WEIGHT statement.

The INITIAL and PRINTIT statements are optional. The following options may be specified on the PROC statement:

DATA=SASdataset - names the SAS data set to be used as input.

OUTEST=SASdataset requests that NASSLOG create a new SAS data set containing regression coefficients for the full sample and for each replicate.

OUTSTAT=SASdataset - requests that NASSLOG create a new SAS data set containing full-sample logistic regression co-efficients and test statistics.

COVOUT - outputs the covariance matrix for the parameter estimates to the OUTEST data set. This option is valid only if OUTEST= is also specified.

DLIKE=value - specifies the convergence criterion, which is the difference in -2 log-likelihood between successive steps. The default value is .025.

MAXITER=n or MI=n - specifies the maximum number of iterations to perform. The default value is 25.

The MODEL statement specifies the variables to use in the analysis. After the keyword MODEL, the dependent variable is specified, followed by an equal sign and the independent variables. The options below may appear in the MODEL statement after a slash (/). If no options are specified, the slash may be omitted.

NOPRINT - suppresses the normal printed output.

PRINTREP - requests that regression coefficients for each replicate be printed.

The WEIGHT statement is similar to the one used in PROC NASSREG and PROC WESVAR.

The INITIAL statement allows the user to specify starting parameter values for the estimation process. The first constant in the list represents the initial parameter estimate for the intercept. The remaining constants specify initial parameter estimates for each of the independent variables in the MODEL statement, in the same order. When no INITIAL statement appears, NASSLOG uses starting estimates of zero.

The values specified in the INITIAL statement are used as starting estimates for the full sample only. The full sample coefficients computed by NASSLOG are always used as starting estimates for the half samples, whether or not an INITIAL statement is supplied.

The PRINTIT statement allows the user to print regression estimates and log-likelihood values at each iteration for selected replicates (or for all replicates).

REFERENCES

[1] Beaton, A. (1964), "The Use of a Special Matrix Operator in Statistical Calculus," Research Bulletin, Princeton: Educational Testing Service.

[2] Binzer, G. and Morganstein, D.R. (1983), "Automation of Estimation and Sampling Error Computations Using PROCS NASSTIM and NASSVAR," SAS User Group Conference, Feb. 1983.

[3] Goodnight, J.H., "A Tutorial on the SWEEP Operator," <u>The American Statistician</u>, August, Vol.33,No. 3.

[4] Kleinbaum, Kupper, and Morgenstern (1982), <u>Epidemiologic Research, Principles and Quantitative Methods</u>, Lifetime Learning Publications, A Division of Wadsworth, Inc., Belmont, California.

[5] McCarthy, Philip J. (1966), "Replication, An Approach to the Analysis of Data from Complex Surveys," Public Health Service Publication No. 1000,Series 2, No. 14.

[6] McCarthy, Philip J. (1969). "Pseudo-Replication: Half-Samples," <u>Review of International Statistical Institute</u>, 37, 239-264.

[7] <u>SAS Companion for OS Operating Systems and TSO</u>, 1984 Edition

[8] <u>SAS User's Guide: Basic</u>, Version 5