Example 4: Variance estimates for Linear Regression: Women. Variance estimates in SAS, SUDAAN, STATA, and WesVar for the regression of Parity (children ever born) on age, race and Hispanic origin, and education, for women 20-44 years of age

Following are the programs and output for an analysis of the relationship between the number of children born to women 20-44 years of age interviewed in Cycle 6 of the NSFG, to race and Hispanic origin, and education. Coefficients were generated by SAS 9.1, SUDAAN 8.0.2, STATA 8.0, and WesVar 4.1. The estimates calculated are equivalent across software. However, due to specific methods used in calculations, standard errors vary slightly across packages, and design effects vary more substantially.

SAS data files were converted to STATA 8.0 and SPSS formats using DBMS/COPY 8.0. Variables in upper case are original NSFG Cycle 6 variables or recodes. Variables in lower case represent variables that were recoded as part of the variance estimation program. Library and file names are generic and it is assumed the user will apply names specific to his or her computing environment. Formatting and library options have been deleted; preferences will vary across user organizations.

SAS 9.1

The DATA and SET steps create a dataset which contains the variables for females to be used in the analysis. The PROC SURVEYREG models the relationship between a continuous variable (PARITY) and a set of predictors (AGER, 'hieducx', and 'black') specified by the MODEL statement. The WEIGHT statement identifies the weight variable (FINALWGT) to be used in estimating the means. PROC SURVEYREG calculates standard errors appropriate to the complex sample design specified in the STRATUM and CLUSTER statements. The DEFF option requests the calculation of design effects.

SAS 9.1 Program

```
data NSFG.EX4;
set NSFG.FEMALES;
if AGER lt 20 then delete;
if HISPRACE=3 then black=1;
if HISPRACE in (1 2 4) then black=0;
if HIEDUC le 9 then hieducx=0;
else if HIEDUC gt 9 then hieducx=1;
run;

proc surveyreg data=NSFG.EX4;
stratum SEST;
cluster SECU_R;
weight  FINALWGT;
model PARITY= AGER hieducx black / deff;
run;
```

The estimated regression coefficients are equivalent to the other software systems.

SAS 9.1 Output

Female Parity regressed on race and ethnicity, age, and education

The SURVEYREG Procedure

Regression Analysis for Dependent Variable PARITY

Data Summary

Number of Observations	6493
Sum of Weights	51726606
Weighted Mean of PARITY	1.50209
Weighted Sum of PARITY	77698129

Design Summary

Number of Strata	84
Number of Clusters	168

Fit Statistics

R-square	0.2282
Root MSE	1.2460
Denominator DF	84

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	3	372.67	<.0001
Intercept	1	35.60	<.0001
AGER	1	587.26	<.0001
hieducx	1	386.16	<.0001
black	1	15.17	0.0002

NOTE: The denominator degrees of freedom for the F tests is 84.

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|	Design Effect
Intercept	-0.5559592	0.09317570	-5.97	<.0001	1.58
AGER	0.0760041	0.00313634	24.23	<.0001	0.00
hieducx	-0.7451044	0.03791680	-19.65	<.0001	1.45
black	0.2244115	0.05761279	3.90	0.0002	1.63

NOTE: The denominator degrees of freedom for the t tests is 84.

SUDAAN 8.0.2

A SAS-callable version of SUDAAN 8.0.2 was used to calculate the estimates for this example. The DATA and SET statements used to create a dataset and the variables needed for this analysis are identical to those used above in the SAS 9.1 program and are omitted for this program.

The PROC REGRESS models the relationship between a continuous variable (PARITY) and a set of predictors (AGER, 'hieducx', and 'black') specified by the MODEL statement. The DESIGN used in this analysis is WR, with replacement. By specifying

DEFT4 in the REGRESS statement, design effects will be calculated. The NEST statement specifies the strata (SEST) and cluster (SECU_R) variables for calculating standard errors appropriate to the complex sample design. The WEIGHT statement identifies FINALWGT for estimated the weighted means.

SUDAAN 8.0.2 Program

```
(same recode as required in sas9)

proc sort data=NSFG.EX4;
by SEST SECU_R;
proc regress data=NSFG.EX4 design=wr deft4;
nest SEST SECU_R;
weight  FINALWGT;
model PARITY=AGER hieducx black;
run;
```

The estimated coefficients calculated by SUDAAN 8.0.2 are identical to those from SAS 9.1.

SUDAAN 8.0.2 Output

```
              Female Parity regressed on race and ethnicity, age, and education

                                  S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute      January 2003
                                Release 8.0.2

Number of observations read       :   6493    Weighted count: 51726606
Observations used in the analysis :   6493    Weighted count: 51726606
Denominator degrees of freedom    :     84

Maximum number of estimable parameters for the model is  4

File NSFG.EX7 contains  168 Clusters
 168 clusters were used to fit the model
Maximum cluster size is 155 records
Minimum cluster size is   3 records

Weighted mean response is 1.502092

Multiple R-Square for the dependent variable PARITY: 0.228243

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable PARITY: TOTAL NUMBER OF LIVE BIRTHS


Independent                                                          P-value
  Variables and        Beta         DEFF                             T-Test
  Effects              Coeff.       Beta #4       SE Beta   T-Test B=0   B=0

Intercept                  -0.56       1.58         0.09        -5.97     0.0000
AGE AT INTERVIEW            0.08       2.17         0.00        24.24     0.0000
HIEDUCX                    -0.75       1.45         0.04       -19.66     0.0000
BLACK                       0.22       1.63         0.06         3.90     0.0002
```

SUDAAN 8.0.2 Output

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable PARITY: TOTAL NUMBER OF LIVE BIRTHS


Contrast              Degrees
                      of                      P-value
                      Freedom       Wald F    Wald F

OVERALL MODEL                4      780.81     0.0000
MODEL MINUS
  INTERCEPT                  3      372.84     0.0000
INTERCEPT                    1       35.62     0.0000
AGER                         1      587.53     0.0000
HIEDUCX                      1      386.34     0.0000
BLACK                        1       15.18     0.0002
```

STATA 8.0

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (FINALWGT), strata (SEST), and cluster (SECU_R) variables to be used by STATA 8.0 in estimation. These settings are saved for the current session, but can be cleared by entering the *clear* command or running *svyset* again with different settings.

The *generate* and *replace* statements create the recodes 'hieducx' and 'black'. The *svyreg* command models the relationship between PARITY and a set of predictors (AGER, 'hieducx', and 'black'). The estimates provided are appropriate to the complex sample design identified by the *svyset* command. Design effect calculations are requested by entering *deff* after the *svyreg* command.

STATA 8.0 Program

```
use "EX4.dta"

svyset [pweight=FINALWGT], strata(SEST) psu(SECU_R)

drop if AGER <20
generate hieducx=0 if HIEDUC <=9
replace  hieducx=1 if HIEDUC >9

generate black=0
replace  black=1 if HISPRACE==3

svyreg PARITY AGER hieducx black, deff
```

The estimated coefficients as calculated by STATA 8.0 are identical to those calculated by SAS 9.1 and SUDAAN 8.0.2.

STATA 8.0 Output

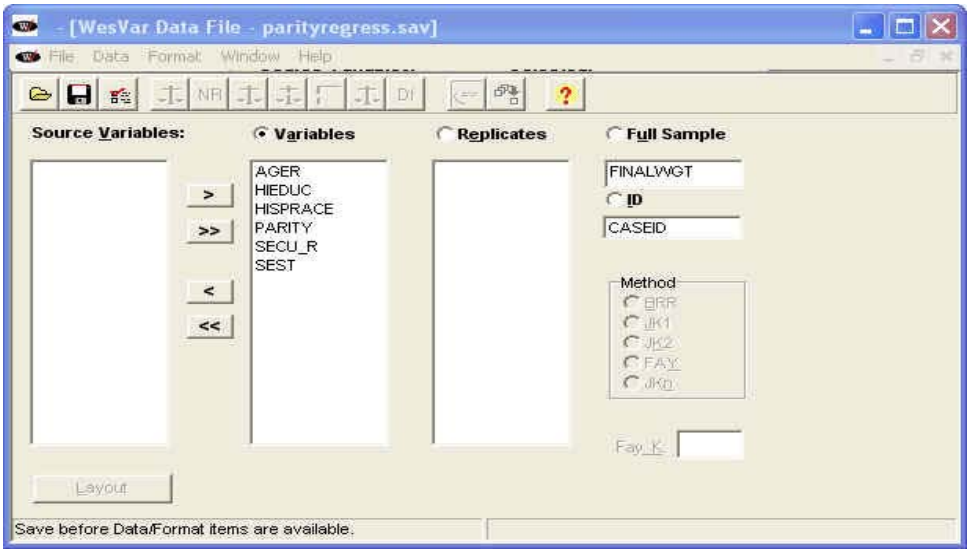
```
. svyreg parity ager hieducx black, deff

Survey linear regression

pweight:  finalwgt                              Number of obs    =      6493
Strata:   sest                                  Number of strata =        84
PSU:      secu_r                                Number of PSUs   =       168
                                                Population size  =  51726606
                                                F(   3,     82)  =    363.97
                                                Prob > F         =    0.0000
                                                R-squared        =    0.2282


      parity |      Coef.   Std. Err.      Deff
-------------+------------------------------------------------------------------
        ager |   .0760041   .0031356   2.428935
     hieducx |  -.7451044    .037908   1.322322
       black |   .2244115   .0575995   1.274843
       _cons |  -.5559592   .0931542   1.994361
```

WesVar 4.1

Not all WesVar windows are displayed for this example. Readers may refer to Example 1 for a full set of windows.

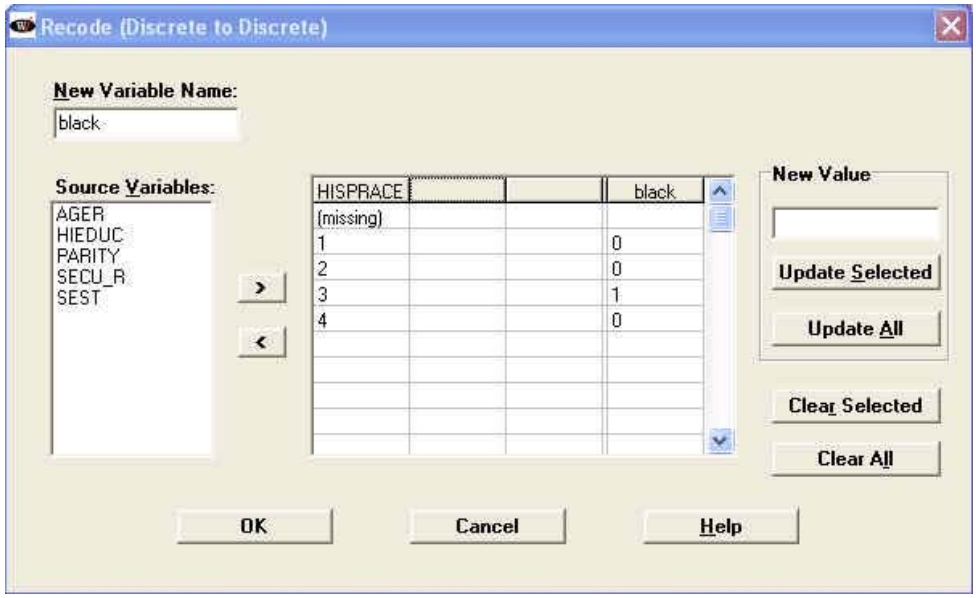
Window 1 displays the selection and categorization of variables to be used in this analysis. After variables are selected and categorized, a new dataset is created.

WesVar 4.1 Program Window 1



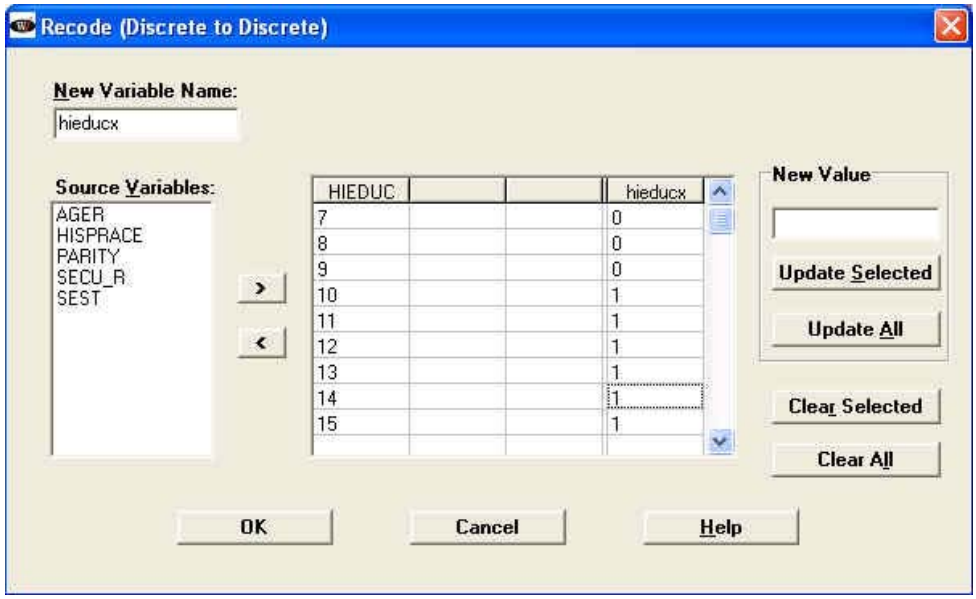
Window 2 displays the procedure for recoding HISPRACE into 'black'. Select *Recode* under the *Format* menu then the *New Discrete to Discrete* button to create 'black'.

WesVar 4.1 Program Window 2



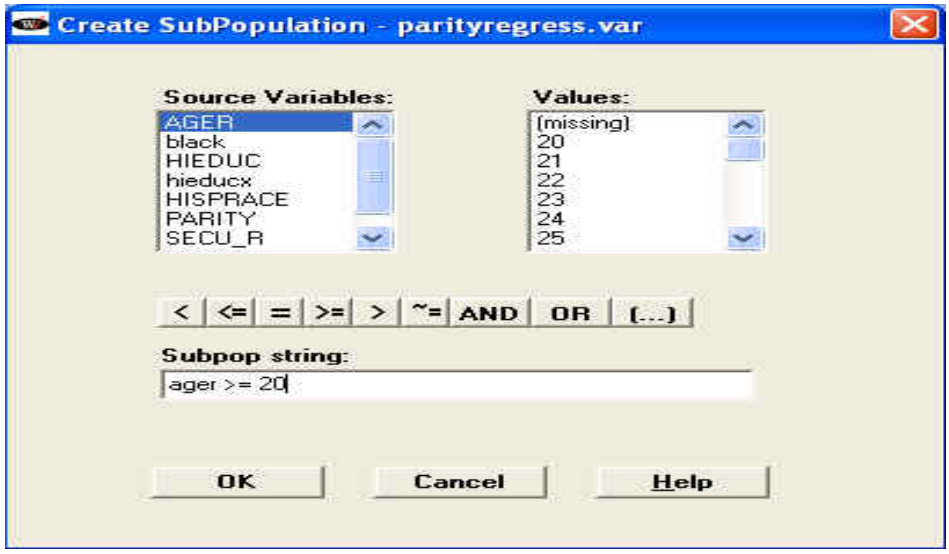
Window 3 displays the procedure for recoding HIEDUC into 'hieducx'.

WesVar 4.1 Program Window 3



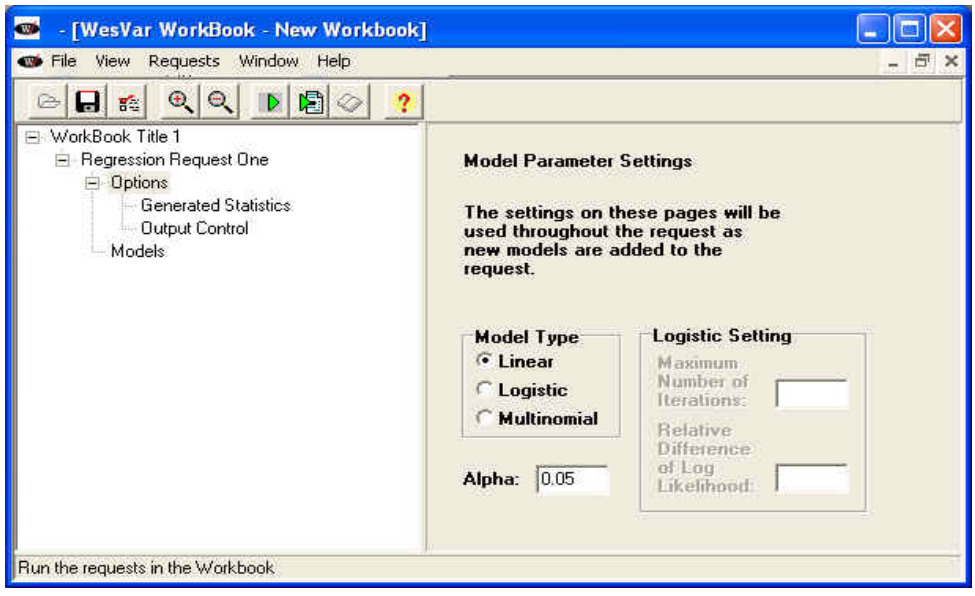
To restrict the analysis to women 20-44 years of age, create a subpopulation by selecting *Subset Population* under the *Data* menu.

WesVar 4.1 Program Window 4



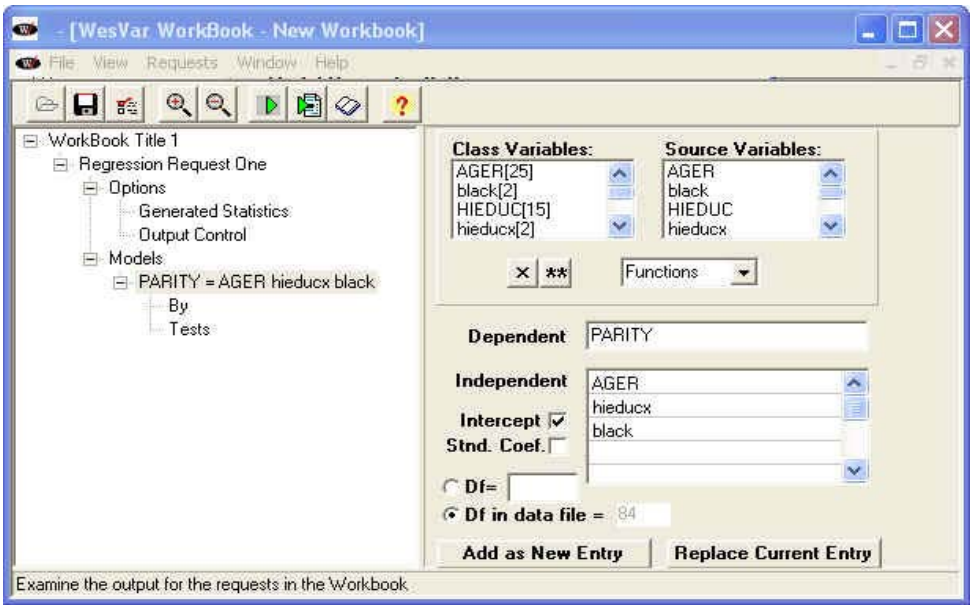
The type of regression (*Linear*) is selected in Window 5.

WesVar 4.1 Program Window 5



Window 6 displays the selection of the dependent (PARITY) and independent (AGER, 'hieducx', and 'black') variables.

WesVar 4.1 Program Window 6



WesVar 4.1 Output

```
  OPTIONS :       Intercept,
         No Standardized Coefficient,
         Degrees of Freedom = 84
         t VALUE : 1.989
  BY :    None Specified.
  MISSING :       0            (UNWEIGHTED)
         0.000000     (WEIGHTED)
  NONMISSING :    6493         (UNWEIGHTED)
         51726606.083494 (WEIGHTED)


MODEL :  23735395.718
ERROR :  80256588.944
TOTAL :     1.040e+08
R_SQUARE VALUE :        0.228


                    PARAMETER      STANDARD ERROR     TEST FOR HO:
PARAMETER             ESTIMATE       OF ESTIMATE      PARAMETER=0         PROB>|T|
INTERCEPT                -0.56             0.094           -5.895            0.000
AGER                      0.08             0.003           23.840            0.000
hieducx                  -0.75             0.039          -19.288            0.000
black                     0.22             0.058            3.878            0.000


                    INTERCEPT             AGER          hieducx            black
INTERCEPT               1.000           -0.914           -0.293            0.172
AGER                   -0.914            1.000           -0.034           -0.281
hieducx                -0.293           -0.034            1.000            0.119
black                   0.172           -0.281            0.119            1.000


TEST          F VALUE         NUM. DF         ENOM. DF         PROB>F    NOTE
OVERALL FIT 353.064                 3               82          0.000
AGER        568.342                 1               84          0.000
hieducx     372.030                 1               84          0.000
black        15.037                 1               84          0.000
```