Example 9: Variance estimates for Logistic Regression: Men and Women.

Variance estimates in SAS, SUDAAN, STATA, and WesVar for the Probability of strongly agreeing with “a young couple should not live together unless they are married” regressed on age, gender, race and Hispanic origin, and education, males and females 15-44 years of age.

Following are the programs and output for the regression of the probability of strongly agreeing that “a young couple should not live together unless they are married” on age, gender, race and Hispanic origin, and education. Regression coefficients and odds ratios 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 (for use in WesVar 4.1) 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, SET and MERGE steps create a dataset which contains the variables and recodes (‘okcohabx’, ‘black’, and ‘hieducx’) for males and females to be used in the analysis. The PROC SURVEYLOGISTIC models the relationship between a dichotomous variable (‘okcohabx’) and a set of predictors (AGER, ‘hieducx’, ‘black’, and FEMALE) specified in the MODEL statement. The WEIGHT statement identifies the weight variable (FINALWGT) to be used in estimating the model. PROC SURVEYLOGISTIC calculates standard errors appropriate to the complex sample design specified in the STRATUM and CLUSTER statements. The DEFF option, which requests calculation of design effects, is not available with PROC SURVEYLOGISTIC.

```sas
SAS 9.1 Program
data mlivtog (keep= CASEID AGER FEMALE HISPRACE HIEDUC SECU SEST FINALWGT OKCOHAB);
set NSFG.MALES;
data flivtog (keep= CASEID AGER FEMALE HISPRACE HIEDUC SECU_R SECU SEST FINALWGT OKCOHAB);
set NSFG.FEMALES;
SECU=SECU_R;
run;
proc sort data=mlivtog;   by CASEID;
proc sort data=flivtog;   by CASEID;
data NSFG.MF_LIVTOG; merge mlivtog flivtog; by CASEID;
run;
data NSFG.EX9;
set NSFG.MF_LIVTOG;
if OKCOHAB=1 then okcohabx=1; else okcohabx=5;
if HISPRACE in (1 2 4) then black=0;
if HISPRACE=3 then black=1;
if HIEDUC le 9 then hieducx=0;
else if HIEDUC gt 9 then hieducx=1;
run;
proc surveylogistic data=NSFG.EX9;
stratum SEST;
cluster SECU;
weight FINALWGT;
model okcohabx = AGER hieducx black FEMALE;
run;
```
The estimated coefficients and odds ratios are equivalent to the other software systems.

**SAS 9.1 Output**

The SURVEYLOGISTIC Procedure  
Testing Global Null Hypothesis: BETA=0  

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>433975.770</td>
<td>4</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Score</td>
<td>435582.565</td>
<td>4</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Wald</td>
<td>24.7892</td>
<td>4</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates  

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.7277</td>
<td>0.2099</td>
<td>168.9313</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AGER</td>
<td>1</td>
<td>0.00723</td>
<td>0.00703</td>
<td>1.0579</td>
<td>0.3037</td>
</tr>
<tr>
<td>hieducx</td>
<td>1</td>
<td>0.3204</td>
<td>0.1208</td>
<td>7.0301</td>
<td>0.0080</td>
</tr>
<tr>
<td>black</td>
<td>1</td>
<td>0.2710</td>
<td>0.1063</td>
<td>6.4945</td>
<td>0.0108</td>
</tr>
<tr>
<td>female</td>
<td>1</td>
<td>0.0626</td>
<td>0.1033</td>
<td>0.3675</td>
<td>0.5443</td>
</tr>
</tbody>
</table>

Odds Ratio Estimates  

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGER</td>
<td>1.007</td>
<td>0.993</td>
</tr>
<tr>
<td>hieducx</td>
<td>1.378</td>
<td>1.087</td>
</tr>
<tr>
<td>black</td>
<td>1.311</td>
<td>1.065</td>
</tr>
<tr>
<td>female</td>
<td>1.065</td>
<td>0.870</td>
</tr>
</tbody>
</table>

Association of Predicted Probabilities and Observed Responses

| Percent Concordant | Somers' D | 54.7 | 0.133 |
| Percent Discordant| Gamma     | 41.4 | 0.139 |
| Percent Tied      | Tau-a     | 3.9  | 0.021 |
| Pairs              | c          | 12285204 | 0.567 |

**SUDAAN 8.0.2**

A SAS-callable version of SUDAAN 8.0.2 was used to calculate the estimates for this example. The DATA, SET and MERGE 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 RLOGIST models the relationship between a dichotomous variable (‘okcohabx’) and a set of predictors (AGER, ‘hieducx’, ‘black’, and FEMALE) specified in the MODEL statement. The DESIGN used in this analysis is WR, with replacement. By specifying DEFT4 in the RLOGIST statement, design effects will be calculated. The NEST statement specifies the strata (SEST) and cluster (SECU) variables for calculating standard errors appropriate to the complex sample design. The WEIGHT statement identifies FINALWGT for estimation.
The estimated coefficients and odds ratios calculated by SUDAAN 8.0.2 are identical to those from SAS 9.1.

**SUDAAN 8.0.2 Output**

```
SUDAAN 8.0.2 Output

---

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 zero responses : 11503
Number of non-zero responses : 1068

Independence parameters have converged in 5 iterations

Number of observations read : 12571
Observations used in the analysis : 12571

Denominator degrees of freedom : 85

Maximum number of estimable parameters for the model is 5

File NSFG.EX8X contains 168 Clusters
168 clusters were used to fit the model

Maximum cluster size is 316 records
Minimum cluster size is 11 records

Sample and Population Counts for Response Variable OKCOHABX
0: Sample Count 11503 Population Count 111258567
1: Sample Count 1068 Population Count 11449169

R-Square for dependent variable OKCOHABX (Cox & Snell, 1989): 0.003530

\[-2 \times \text{Normalized Log-Likelihood with Intercepts Only} : 7796.96\]
\[-2 \times \text{Normalized Log-Likelihood Full Model} : 7752.50\]
Approximate Chi-Square (-2 * Log-L Ratio) : 44.46
Degrees of Freedom : 4

Note: The approximate Chi-Square is not adjusted for clustering.
Refer to hypothesis test table for adjusted test.

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable OKCOHABX: OKCOHABX

<table>
<thead>
<tr>
<th>Variables and Effects</th>
<th>Beta</th>
<th>DEFF Beta #4</th>
<th>SE Beta</th>
<th>T-Test B=0</th>
<th>P-value T-Test B=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.73</td>
<td>3.10</td>
<td>0.21</td>
<td>-13.00</td>
<td>0.0000</td>
</tr>
<tr>
<td>R's age at interview</td>
<td>0.01</td>
<td>3.71</td>
<td>0.01</td>
<td>1.03</td>
<td>0.3064</td>
</tr>
<tr>
<td>HIEDUCX</td>
<td>0.32</td>
<td>3.64</td>
<td>0.12</td>
<td>2.65</td>
<td>0.0096</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.27</td>
<td>1.53</td>
<td>0.11</td>
<td>2.55</td>
<td>0.0127</td>
</tr>
<tr>
<td>FEMALE</td>
<td>0.06</td>
<td>2.81</td>
<td>0.10</td>
<td>0.61</td>
<td>0.5480</td>
</tr>
</tbody>
</table>
```
**SUDAAN 8.0.2 Output Cont.**

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable OKCOHABX: OKCOHABX

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Degrees of Freedom</th>
<th>Wald F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVERALL MODEL</td>
<td>5</td>
<td>542.13</td>
<td>0.0000</td>
</tr>
<tr>
<td>MODEL MINUS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>4</td>
<td>6.19</td>
<td>0.0002</td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>1</td>
<td>68.97</td>
<td>0.0000</td>
</tr>
<tr>
<td>AGER</td>
<td>1</td>
<td>1.06</td>
<td>0.3064</td>
</tr>
<tr>
<td>HIEDUCX</td>
<td>1</td>
<td>7.03</td>
<td>0.0096</td>
</tr>
<tr>
<td>BLACK</td>
<td>1</td>
<td>6.48</td>
<td>0.0127</td>
</tr>
<tr>
<td>FEMALE</td>
<td>1</td>
<td>0.37</td>
<td>0.5460</td>
</tr>
</tbody>
</table>

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable OKCOHABX: OKCOHABX

<table>
<thead>
<tr>
<th>Independent Variables and Effects</th>
<th>Odds Ratio</th>
<th>Lower 95% Limit</th>
<th>Upper 95% Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.07</td>
<td>0.04</td>
<td>0.10</td>
</tr>
<tr>
<td>R’s age at interview</td>
<td>1.01</td>
<td>0.99</td>
<td>1.02</td>
</tr>
<tr>
<td>HIEDUCX</td>
<td>1.38</td>
<td>1.08</td>
<td>1.75</td>
</tr>
<tr>
<td>BLACK</td>
<td>1.31</td>
<td>1.06</td>
<td>1.62</td>
</tr>
<tr>
<td>FEMALE</td>
<td>1.06</td>
<td>0.87</td>
<td>1.31</td>
</tr>
</tbody>
</table>

**STATA 8.0**

The `use` statement specifies the dataset to be used. The `svyset` command specifies the weight (FINALWGT), strata (SEST), and cluster (SECU) 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 recodes: ‘okcohabx’, ‘hieducx’, and ‘black’. The `svylogit` the relationship between a dichotomous variable (‘okcohabx’) and a set of predictors (AGER, ‘hieducx’, ‘black’, and FEMALE) specified in the `svylogit` statement. 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 `svylogit` command.

**STATA 8.0 Program**

```stata
use "EX9.dta"
svyset [pweight=FINALWGT], strata(SEST) psu(SECU)
generate okcohabx=0
replace okcohabx=1 if OKCOHAB==1
generate hieducx=0 if HIEDUC <= 9
replace hieducx=1 if HIEDUC > 9
generate black=0
replace black=1 if HISPRACE==3
svylogit okcohabx AGER hieducx black FEMALE, 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
. svylogit okcohabx ager hieducx black female, def
Survey logistic regression
pweight: finalwgt
Strata: sest
PSU: secu

( 1.227e+08
F( 4, 81) = 5.97
Prob > F = 0.0003

okcohabx | Coef. Deff
-------------+----------------------------------------------------------------
ager | .0072349 3.57503
hieducx | .3204423 3.59665
black | .2707356 1.532045
female | .0626189 2.819198
_cons | -2.727918 3.029195
-------------+----------------------------------------------------------------
```

WesVar 4.1

Not all WesVar windows are displayed for this example. Readers may refer to Example 1 for a full set of windows. An SPSS file was imported for this analysis.

Window 1 displays the selection and categorization of variables to be used in the current analysis. After variables are selected and categorized, a new dataset is created.
Windows 2, 3 and 4 display the procedure for recoding OKCOHAB into ‘okcohax’, HIEDUC into ‘hieducx’, and HISPRACE into ‘black’. To create ‘okcohax’, ‘hieducx’, and ‘black’ select Recode under the Format menu then the New Discrete to Discrete button.
The type of regression (Logistic) and parameter settings are selected in Window 5.
Window 6 displays the statistics requested.

Additional output is selected in Window 7.
Window 8 displays the selection of the dependent (‘okcohabx’) and independent (AGER, ‘hieducx’, ‘black’, and FEMALE) variables.

The output provided by WesVar 4.1 is a list-wise statement of all the estimates requested. The estimated coefficients are identical to those calculated by the other software systems. Design effect estimates are not available for output with regression estimates.
OPTIONS: Intercept, No Standardized Coefficient, Degrees of Freedom = 7

STARTING VALUES: INTERCEPT: 0.0000
AGER: 0.0000
hieducx: 0.0000
black: 0.0000
FEMALE: 0.0000

BY: None Specified.

MISSING: 0 (UNWEIGHTED)
NONMISSING: 122707736 (WEIGHTED)

Success = records with dependent value equal to 0: 1068 (UNWEIGHTED)
11449169.250616 (WEIGHTED)

Failure = records with dependent value equal to 1: 11503 (UNWEIGHTED)
111258567.038382 (WEIGHTED)

ITERATIONS REQUIRED FOR FULL SAMPLE: 6
MAXIMUM ITERATIONS FOR REPLICATE SAMPLE: 6
-2 LOG LIKELIHOOD FOR FULL SAMPLE: 75673537.40390
-2 LOG LIKELIHOOD FOR MODEL CONTAINING INTERCEPT ONLY: 76107513.46103

Negative log likelihood: 0.006
Likelihood ratio(Cox-Snell): Maximum possible value: 0.462
Likelihood ratio(Estrella): 0.004

PARAMETER STANDARD ERROR TEST FOR H0: PARAMETER ESTIMATE OF ESTIMATE PARAMETER=0 PROB>|T| COMMENT
INTERCEPT -2.73 0.344 -7.933 0.000
AGER 0.01 0.010 0.741 0.483
hieducx 0.32 0.087 3.674 0.008
black 0.27 0.191 1.419 0.199
FEMALE 0.06 0.025 2.506 0.041

TEST F VALUE NUM. DF DENOM. DF PROB>F NOTE
OVERALL FIT 75.571 4 4 0.001
AGER 0.550 1 7 0.483
hieducx 13.498 1 7 0.008
black 2.013 1 7 0.199
FEMALE 6.278 1 7 0.041

PARAMETER ESTIMATE LOWER 95% UPPER 95% NOTE
AGER 1.000 0.984 1.016
hieducx 1.31 0.835 2.058
black 1.38 1.121 1.693
FEMALE 1.06 1.004 1.129