
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 (‘okcohax’, ‘black’, and ‘hieducx’) for males and females to be used in the analysis. The PROC SURVEYLOGISTIC models the relationship between a dichotomous variable (‘okcohax’) 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 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 okcohax=1; else okcohax=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 okcohax = 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

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	433975.770	4	<.0001
Score	435582.565	4	<.0001
Wald	24.7892	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7277	0.2099	168.9313	<.0001
AGER	1	0.00723	0.00703	1.0579	0.3037
hieducx	1	0.3204	0.1208	7.0301	0.0080
black	1	0.2710	0.1063	6.4945	0.0108
female	1	0.0626	0.1033	0.3675	0.5443

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AGER	1.007	0.993	1.021
hieducx	1.378	1.087	1.746
black	1.311	1.065	1.615
female	1.065	0.870	1.304

Association of Predicted Probabilities and Observed Responses

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

SUDAAN 8.0.2 Program

```
(same recode as required in SAS9)

proc sort data=NSFG.EX9;
by SEST SECU;
proc rlogist data=NSFG.EX9 design=wr deft4;
weight FINALWGT;
nest SEST SECU;
model okcohax = AGER hieducx black FEMALE;
run;
```

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

```

                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      Weighted count:122707736
Observations used in the analysis : 12571      Weighted count:122707736
Denominator degrees of freedom   :      84

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 * Normalized Log-Likelihood with Intercepts Only : 7796.96
-2 * 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

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

Intercept          -2.73          3.10          0.21          -13.00          0.0000
R's age at interview 0.01          3.71          0.01          1.03          0.3064
HIEDUCX           0.32          3.64          0.12          2.65          0.0096
BLACK             0.27          1.53          0.11          2.55          0.0127
FEMALE           0.06          2.81          0.10          0.61          0.5460
```

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

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	5	542.13	0.0000
MODEL MINUS			
INTERCEPT	4	6.19	0.0002
INTERCEPT	1	168.97	0.0000
AGER	1	1.06	0.3064
HIEDUCX	1	7.03	0.0096
BLACK	1	6.48	0.0127
FEMALE	1	0.37	0.5460

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

Independent Variables and Effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept	0.07	0.04	0.10
R's age at interview	1.01	0.99	1.02
HIEDUCX	1.38	1.08	1.75
BLACK	1.31	1.06	1.62
FEMALE	1.06	0.87	1.31

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

```
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 okcohax ager hieducx black female, deff
Survey logistic regression

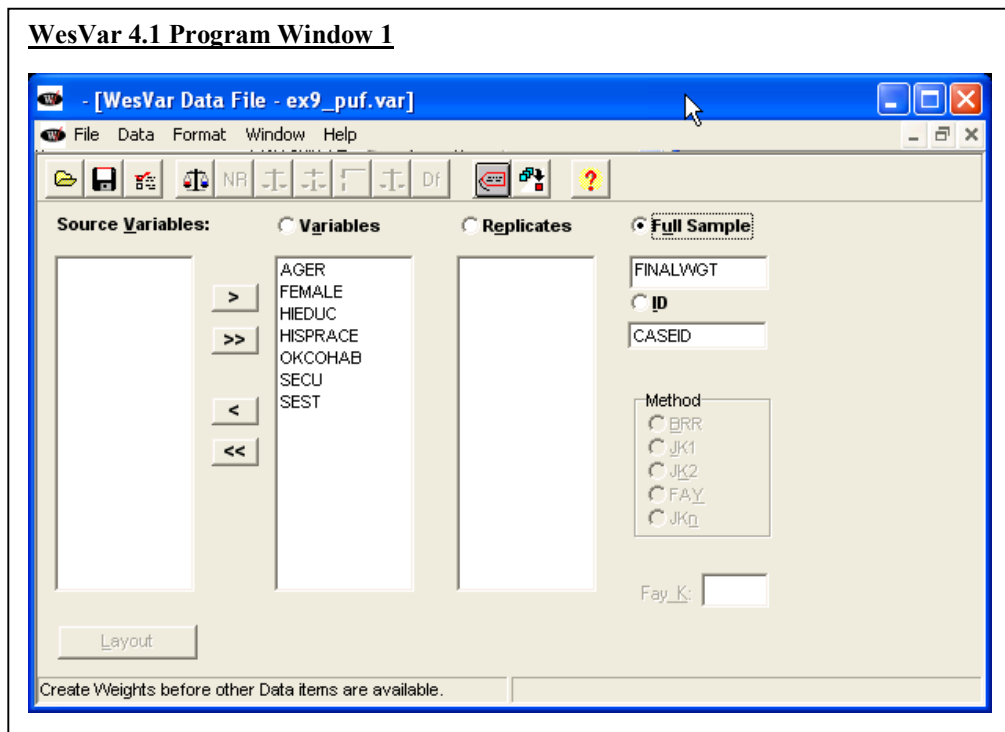
pweight: finalwgt      Number of obs   =    12571
Strata:  sest          Number of strata =     84
PSU:     secu          Number of PSUs  =    168
                          Population size = 1.227e+08
                          F( 4, 81) = 5.97
                          Prob > F = 0.0003

-----+-----
okcohax |      Coef.   Std. Err.   Deff
-----+-----
      ager |   .0072349   .0070302   3.57503
    hieducx |   .3204423   .1208391   3.59665
      black |   .2707356   .1063407   1.532045
     female |   .0626189   .1032816   2.819198
      _cons |  -2.727918   .2098602   3.029195
  
```

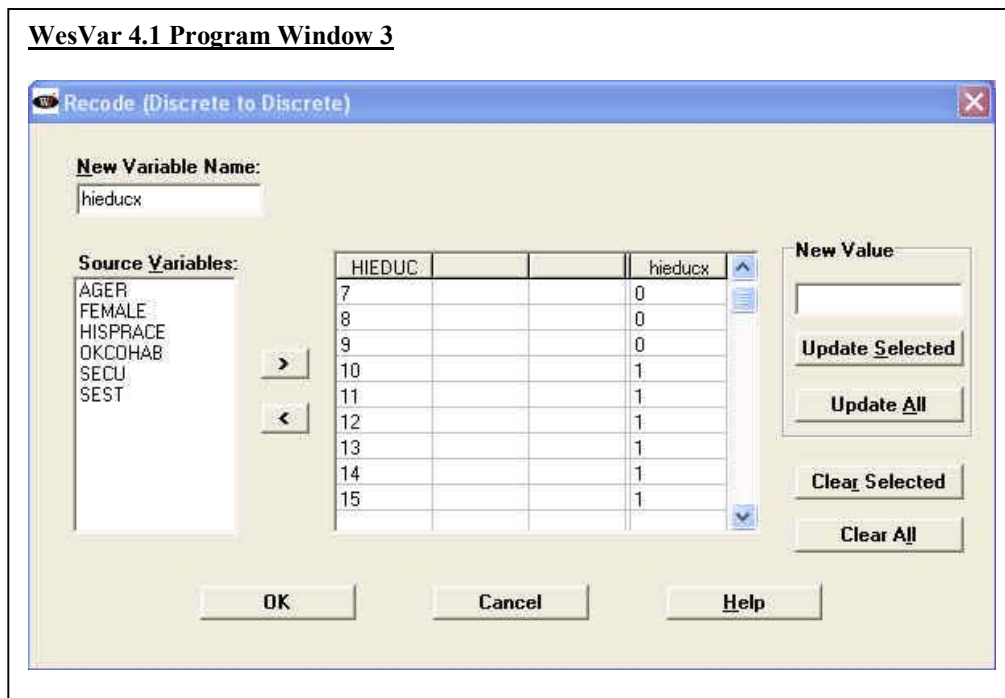
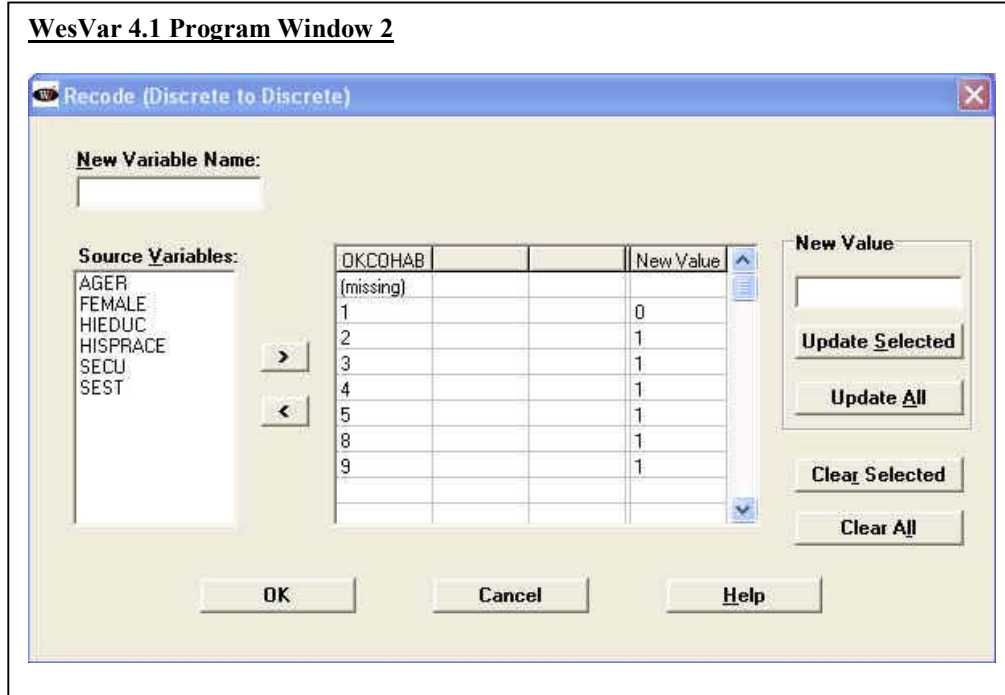
WesVar 4.1

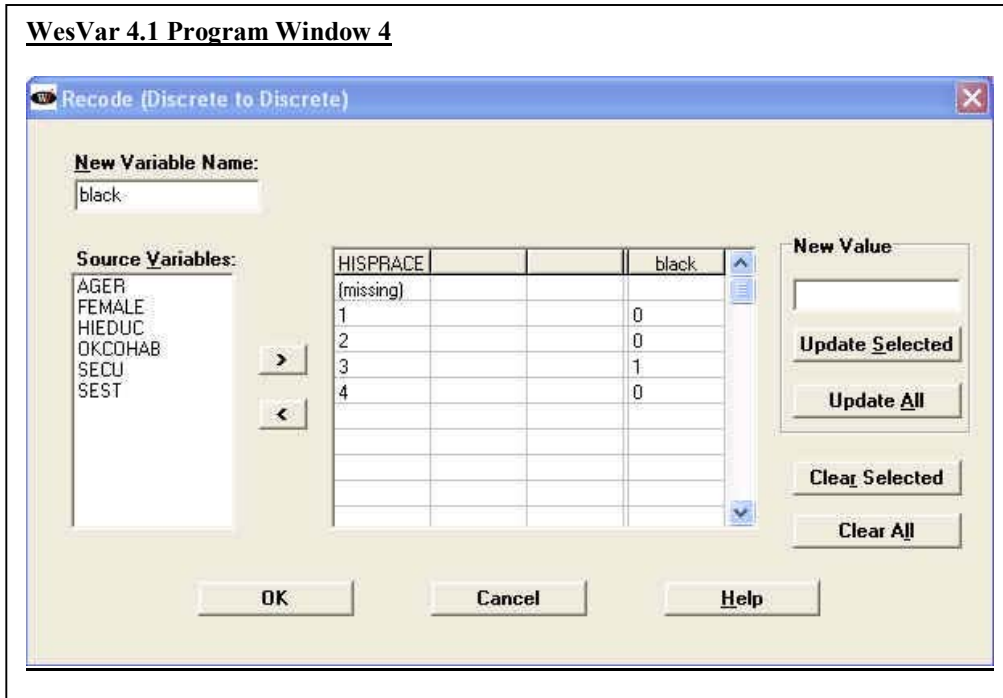
Not all WesVar windows are displayed for this example. Readers may refer to Example 1 for a full set of windows. An SPSS file was imported for this analysis.

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.

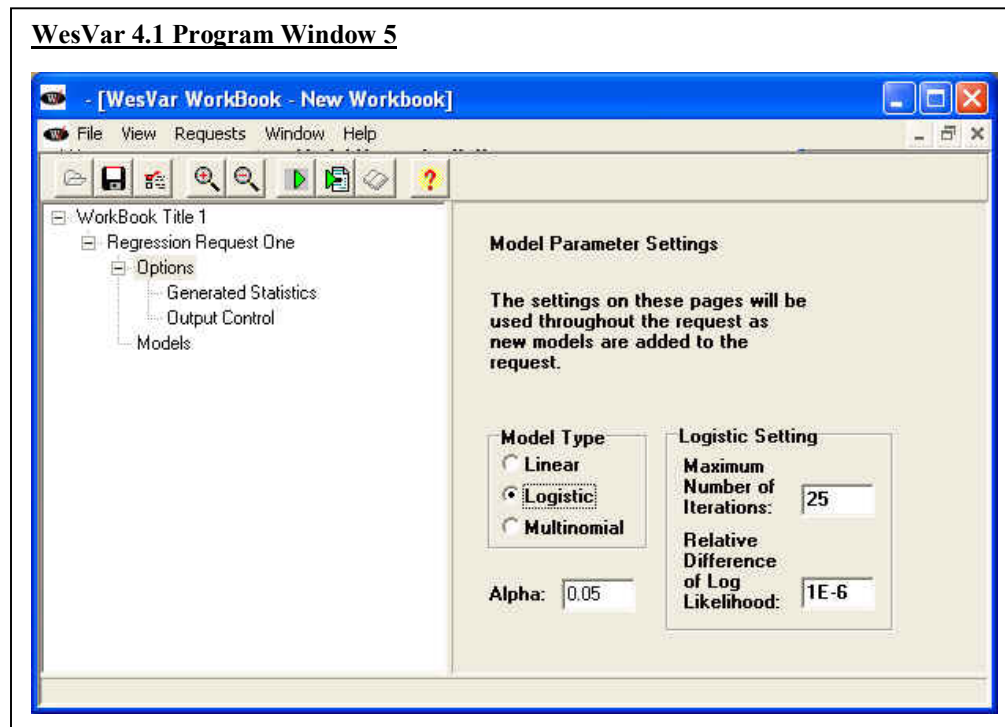


Windows 2, 3 and 4 display the procedure for recoding OKCOHAB into 'okcohabx', HIEDUC into 'hieducx', and HISPRACE into 'black'. To create 'okcohabx', 'hieducx', and 'black' select *Recode* under the *Format* menu then the *New Discrete to Discrete* button.

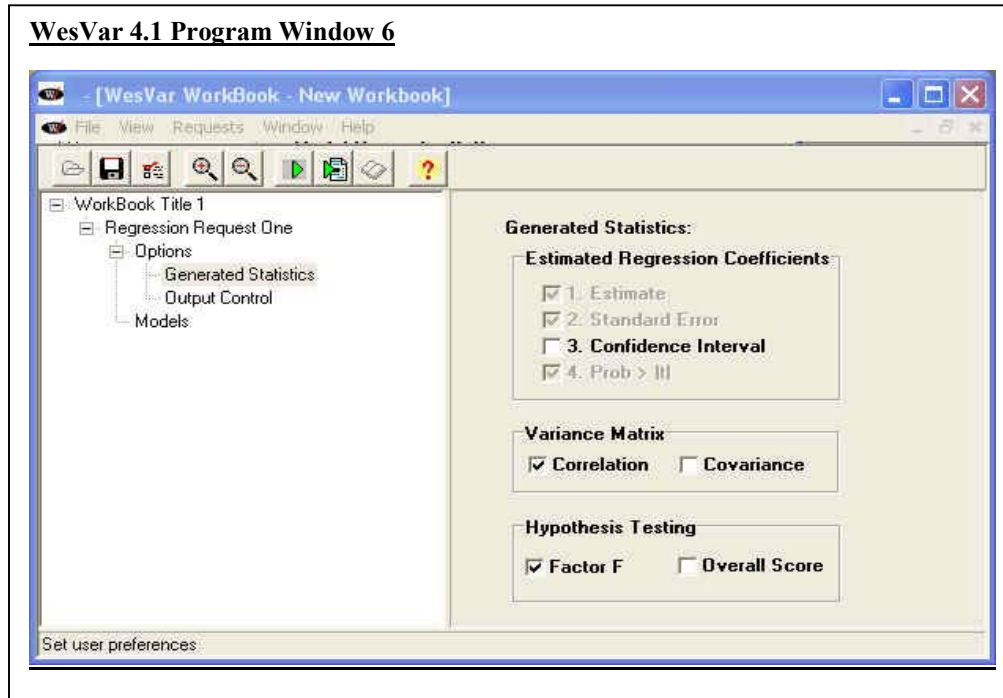




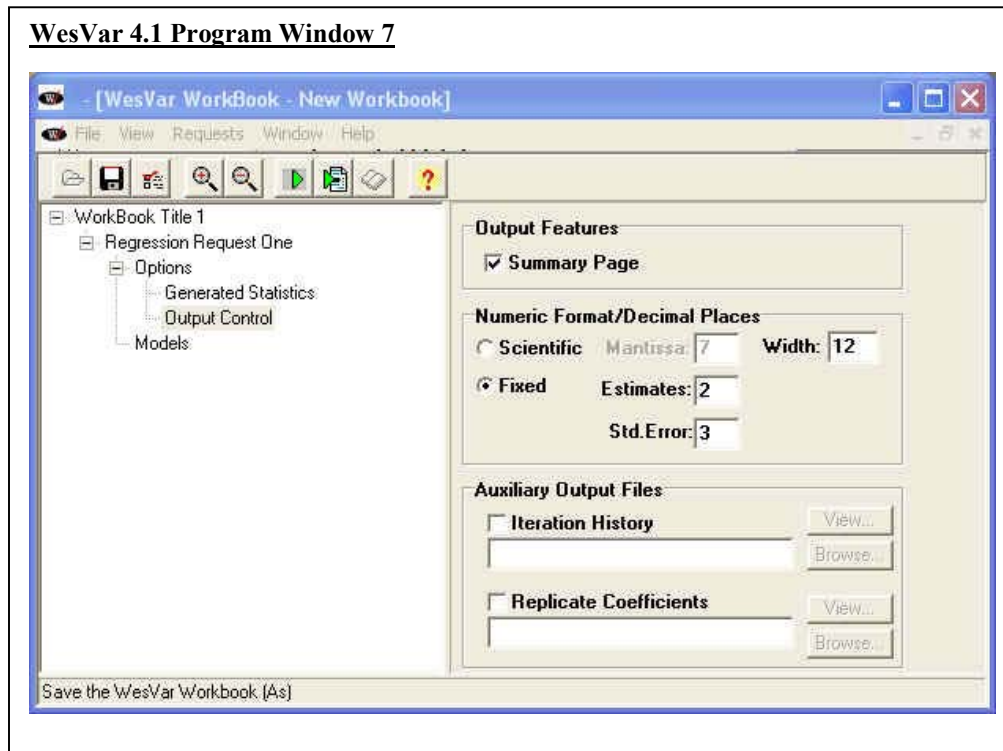
The type of regression (*Logistic*) and parameter settings are selected in Window 5.



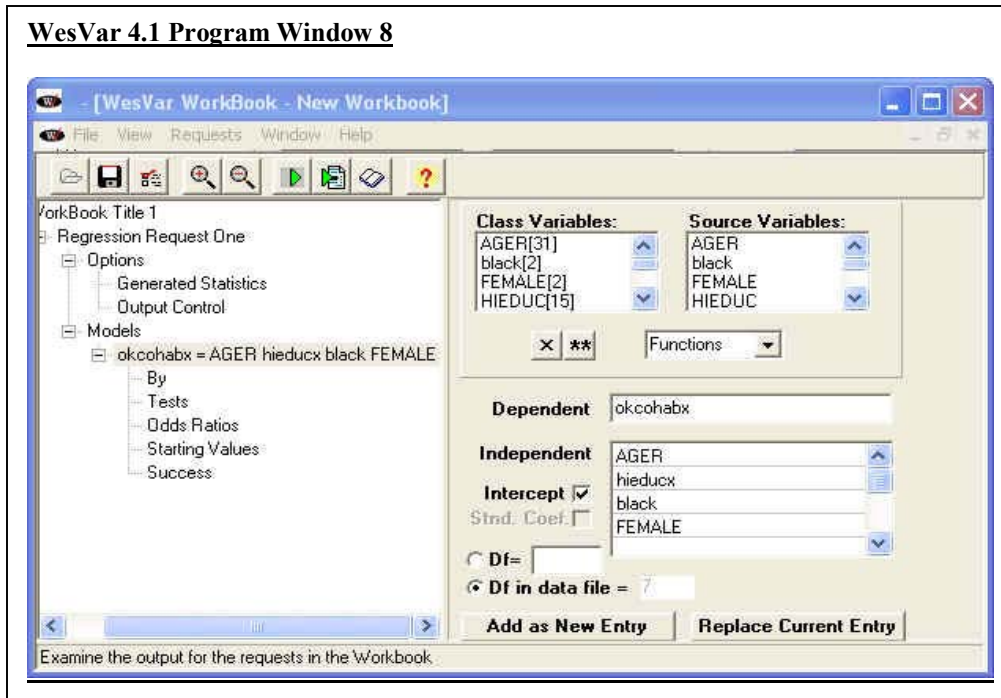
Window 6 displays the statistics requested.



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.

WesVar 4.1 Output

```

WESVAR VERSION NUMBER : v4.1
TIME THE JOB EXECUTED : 15:43:27 10/18/2004
INPUT DATASET NAME : ex9.var
TIME THE INPUT DATASET CREATED : 15:42:05 10/18/2004
FULL SAMPLE WEIGHT : FINALWGT
REPLICATE WEIGHTS : RPL1...RPL8
VARIANCE ESTIMATION METHOD : BRR

TYPE OF ANALYSIS : LOGISTIC
CONVERGENCE CRITERION : 1e-06
MAXIMUM NUMBER OF ITERATIONS : 25
VALUE OF ALPHA (CONFIDENCE LEVEL %) : 0.05000 (95.00000 %)
OPTION OUTPUT REPLICATE COEFFICIENTS : OFF
OPTION OUTPUT ITERATION HISTORY : OFF

MODEL(S): okcohabx = AGER hieducx black FEMALE

NUMBER OF REPLICATES : 8
NUMBER OF OBSERVATIONS READ : 12571
WEIGHTED NUMBER OF OBSERVATIONS READ : 122707736.289

```

WesVar 4.1 Output Cont.

OPTIONS : Intercept,
 No Standardized Coefficient,
 Degrees of Freedom = 7
 t VALUE : 2.365
 STARTING VALUES : INTERCEPT : 0.0000
 AGER : 0.0000
 hieducx : 0.0000
 black : 0.0000
 FEMALE : 0.0000
 BY : None Specified.
 MISSING : 0 (UNWEIGHTED)
 0.000000 (WEIGHTED)
 NONMISSING : 12571 (UNWEIGHTED)
 122707736,288997 (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): 0.004 Maximum possible value: 0.462
 Likelihood ratio(Estrella): 0.004

PARAMETER	PARAMETER ESTIMATE	STANDARD ERROR OF ESTIMATE	TEST FOR H0: 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	

WesVar 4.1 Output Cont.

	INTERCEPT	AGER	hieducx	black	FEMALE
INTERCEPT	1.000	-0.999	0.972	-0.913	0.285
AGER	-0.999	1.000	-0.977	0.899	-0.254
hieducx	0.972	-0.977	1.000	-0.851	0.208
black	-0.913	0.899	-0.851	1.000	-0.632
FEMALE	0.285	-0.254	0.208	-0.632	1.000

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	1.01	0.984	1.031	
hieducx	1.38	1.121	1.693	
black	1.31	0.835	2.058	
FEMALE	1.06	1.004	1.129	