Replicate weights

Contents

1. Introduction

Survey data sets often include a column W_0 of design weights and additional columns $W_1, W_2, \ldots, W_{R-1}$ of replicate weights. Typically, a researcher may repeatedly fit the same model to the data by working through the sequence of weight variables $W_0, W_1, \ldots, W_{R-1}$. Means of the R sets of parameter estimates and their standard errors may subsequently be computed to obtain more accurate parameter and standard error estimates. In this section, we illustrate how to use LISREL in the case of replicate weights.

2. The data

The data set used here is **replicwts.lsf** from the **Missing data examples** folder. A few of the variables are shown below for the first 10 observations in the data set.

replicwts.lsf

	CENREG	FACTYPE	ALCEU	COCEU	MAREU	AGE	GENDER
1	3.0	2.0	1.0	-9.0	1.0	23.0	1.0
2	4.0	2.0	1.0	-9.0	1.0	26.0	2.0
3	2.0	3.0	1.0	1.0	1.0	34.0	1.0
4	3.0	4.0	1.0	-9.0	1.0	13.0	1.0
5	2.0	4.0	1.0	-9.0	-9.0	28.0	1.0
6	4.0	4.0	-9.0	-9.0	-9.0	28.0	1.0
7	4.0	4.0	1.0	-9.0	1.0	44.0	1.0
8	2.0	4.0	1.0	0.0	1.0	33.0	1.0
9	1.0	4.0	1.0	1.0	1.0	27.0	1.0
10	1.0	3.0	1.0	1.0	1.0	45.0	1.0

The contents of the LSF are obtained by selecting the **Statistics**, **Data Screening** option. A portion of the output is shown below.

replicwts.OUT

```
!PRELIS SYNTAX: Can be edited

!Contents of LSFFILE:
!--------------------
!DA NI=91 NO=5005 MI= -9.00, -999999 TR=PA
!LA
!CENREG FACTYPE ALCEU COCEU MAREU AGE GENDER RACE_D
!DEPR EDU JAILR NUMTE A2TWA0 A2TWA1 A2TWA2 A2TWA3
!A2TWA4 A2TWA5 A2TWA6 A2TWA7 A2TWA8 A2TWA9 A2TWA10 A2TWA11
!A2TWA12 A2TWA13 A2TWA14 A2TWA15 A2TWA16 A2TWA17 A2TWA18 A2TWA19
!A2TWA20 A2TWA21 A2TWA22 A2TWA23 A2TWA24 A2TWA25 A2TWA26 A2TWA27
!A2TWA28 A2TWA29 A2TWA30 A2TWA31 A2TWA32 A2TWA33 A2TWA34 A2TWA35
!A2TWA36 A2TWA37 A2TWA38 A2TWA39 A2TWA40 A2TWA41 A2TWA42 A2TWA43
!A2TWA44 A2TWA45 A2TWA46 A2TWA47 A2TWA48 A2TWA49 A2TWA50 A2TWA51
!A2TWA52 A2TWA53 A2TWA54 A2TWA55 A2TWA56 A2TWA57 A2TWA58 A2TWA59
!A2TWA60 A2TWA61 A2TWA62 A2TWA63 A2TWA64 A2TWA65 A2TWA66 A2TWA67
!A2TWA68 A2TWA69 A2TWA70 A2TWA71 A2TWA72 A2TWA73 A2TWA74 A2TWA75
!A2TWA76 A2TWA77 A2TWA78
```

The following variables included in the LSF were selected from the survey data:

- CENREG: This variable indicates the census region and has four categories, these being "Northeast," "Midwest," "South," and "West" respectively.
- FACTYPE: The facility treatment type has four categories, too, representing facilities with "residential treatment", "outpatient methadone treatment", "outpatient non-methadone treatment", and "more than one type of treatment" respectively.
- ALCEU: An indicator variable with value "1" if the respondent has ever used alcohol, and "0" otherwise.
- COCEU: An indicator variable with value "1" if the respondent has ever used cocaine, and "0" otherwise.
- MAREU: An indicator variable with value "1" if the respondent has ever used marijuana, and "0" otherwise.
- AGE: This variable denotes age at admission to a facility.

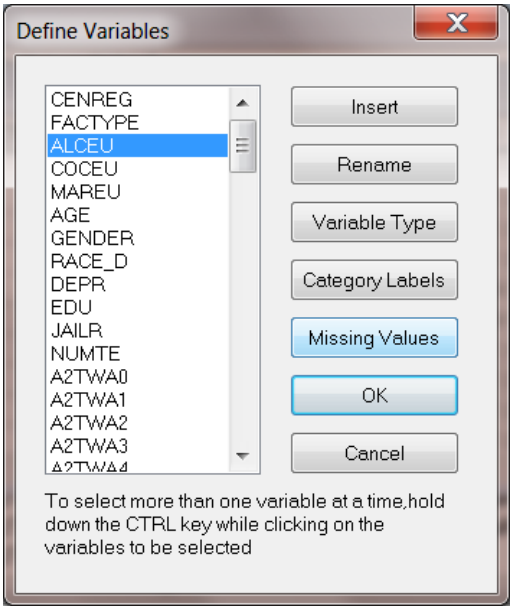
- GENDER: The respondent's gender is denoted by this indicator variable that assumes a value of "1" for female respondents.
- RACE_D: The original variable RACE recoded so that '1" denotes white and "0" other ethnic groups.
- DEPR: This indicator variable is coded "1" if the respondent is depressed, and "0" otherwise.
- EDU: A categorical variable representing the respondent's level of education at admission. It has 5 categories, these being (from 1 to 5) "less than 8 years", "8 – 11 years or less than High School graduate", "High School graduate / GED", "some college", and "college graduate / postgraduate".
- JAILR: This indicator variable indicates whether the respondent had a prison or jail record prior to admission.
- NUMTE: A count variable, indicating the total number of treatment episodes prior to admission.
- A2TWA0 – A2TWA78: These variables are abstract final full sample weights. A more complete description follows below.

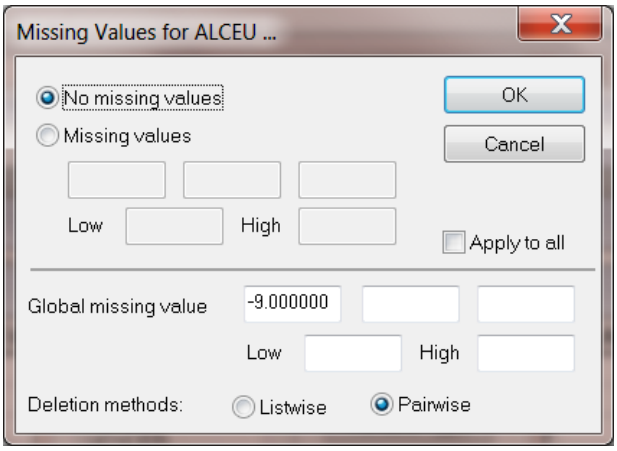
Variance estimation frequently relies on one of two techniques: Taylor series linearization or replication weights. Replicate weights are based on the same ideas as the jackknife and has recently come into use in US government surveys, where replicate weights are provided instead of information on PSUs. In such cases, replicate weights may be used to disguise and/or prevent identification of individuals within PSUs to preserve privacy.

Handling of missing data

Missing values for the variables ALCEU, COCEU, …, NUMTE in the file **replicwts.LSF** are coded -9.0. For these variables, 20.8 % of the possible values are not observed. Use of listwise deletion would result in retaining only 97 of the selected 5005 cases. Therefore, we use the full information maximum likelihood (FIML) procedure as described in Section 3.6.

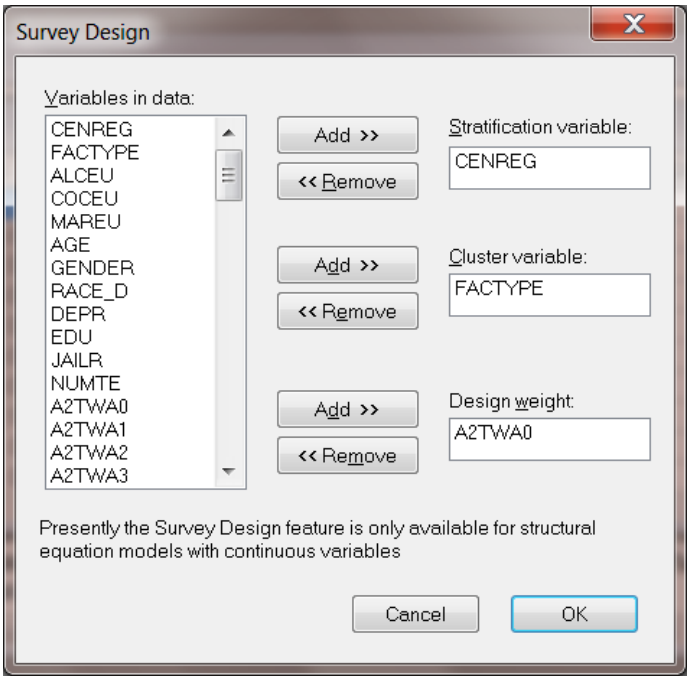
To define -9.0 as the global missing value code, select the **Data**, **Define Variables** option from the main menu. Select the variable ALCEU (or any other variable in the list) and click the **Missing Values** button to activate the **Missing Values for** dialog box. Type -9.0 in the **Global missing value** text box and select **pairwise** as the deletion method. Click **OK** when done.





Handling of zero weights

Descriptive statistics of the weight variables A2TWA0 to A2TWA78 revealed that these variables contained zero as possible values. If a zero value is encountered in any row of the data set, it was replaced by the average value of all the non-zero weights in the corresponding row.

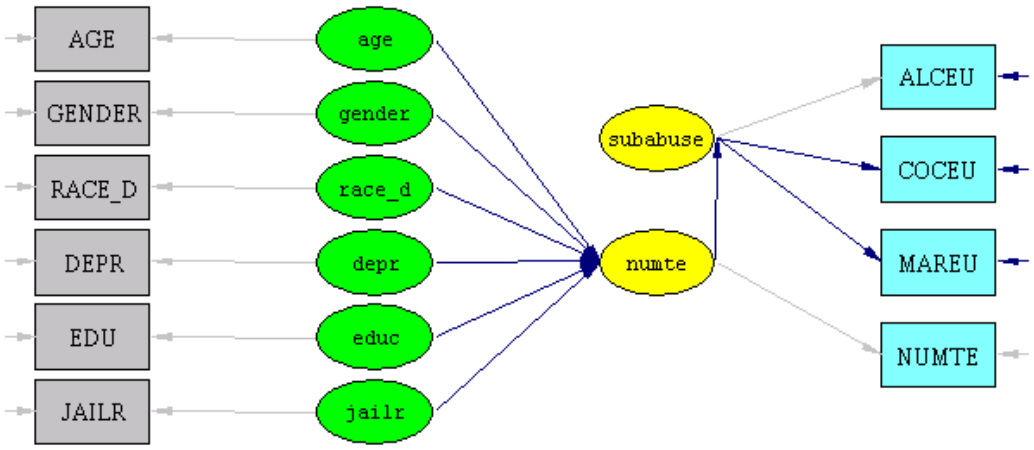


We found that multiple imputation failed in estimating weights if 0 was regarded as a missing value code. The reason for this is that all the weight variables are highly correlated and therefore the covariance matrix of the weight variables is essentially singular.

To define CENREG and FACTYPE as stratification and cluster variables, select the **Data**, **Survey Design** option from the main menu and add the variables as shown in the dialog box below. For our first analysis, select A2TWA0 as the **Design weight** variable.

3 The model

The path diagram representation of the model fitted to the data is shown below. It is assumed that each of the X-variables AGE, GENDER, RACE_D, DEPR, EDU and JAILR is a perfect indicator of the corresponding latent variable, these being age, gender, race_d, depr, educ and jailr. This implies that the error variances of the X-variables are zero, and that the path coefficients age → AGE, gender → GENDER, race_d → RACE_D, depr → DEPR, educ → EDU and jailr → JAILR are all equal to one. In the LISREL terminology, age, gender, race_d, depr, edu and jailr are exogenous (KSI) latent variables. In principle, several indicators of depression and education, if available, could be incorporated into this model.



In the Y part of the model, we include the dependent variables ALCEU, COCEU, MAREU and NUMTE. It is assumed that ALCEU, COCEU, and MAREU are indicators of the endogenous latent (ETA) variable subabuse while NUMTE is a perfect indicator of the ETA variable numte. The path subabuse → ALCEU is set equal to 1 to fix the scale of the endogenous latent variable subabuse.

Finally, we assume that numte can be predicted by age, …, jailr, and in turn, subabuse is predicted by numte. In this context, the latent variable numte is a so-called mediating variable.

The LISREL model consists of a measurement and structural part.

Measurement model

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\Lambda}_y & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_x\end{bmatrix}\begin{bmatrix}\boldsymbol{\eta}\\ \boldsymbol{\xi}\end{bmatrix}+\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\delta}\end{bmatrix}$$

where $\mathbf{x}=(AGE, GENDER, RACE\_D, DEPR, EDU, JAILR)'$, $\boldsymbol{\Lambda}_x$ is a 6×6 identity matrix and $\boldsymbol{\xi}=(age, gender, race\_d, depr, edu, jailr)'$. Also, $Cov(\boldsymbol{\delta})=\mathbf{0}$ and $Cov(\boldsymbol{\xi})=\boldsymbol{\Phi}$.

Furthermore,

$$\boldsymbol{\eta} = \begin{bmatrix} subabuse \\ numte \end{bmatrix},$$

$$\mathbf{y} = \begin{bmatrix} ALCEU \\ COCEU \\ MAREU \\ NUMTE \end{bmatrix},$$

and

$$\boldsymbol{\Lambda}_y = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \end{bmatrix}.$$

Finally, $Cov(\boldsymbol{\varepsilon})$ is a diagonal matrix with diagonal elements $\text{var}(ALCEU)$, $\text{var}(COCEU)$, $\text{var}(MAREU)$ and 0.

Structural equation model

The structural model can be written as

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{B} = \begin{bmatrix} 0 & \beta_{12} \\ 0 & 0 \end{bmatrix},$$

and

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26} \end{bmatrix},$$

that is,

$$subabuse = \beta_{12} \times numte + t_1$$
$$numte = \gamma_{21} \times age + \gamma_{22} \times gender + \gamma_{23} \times race\_d + \gamma_{24} \times depr + \gamma_{25} \times educ + \gamma_{26} \times jailr.$$

Also

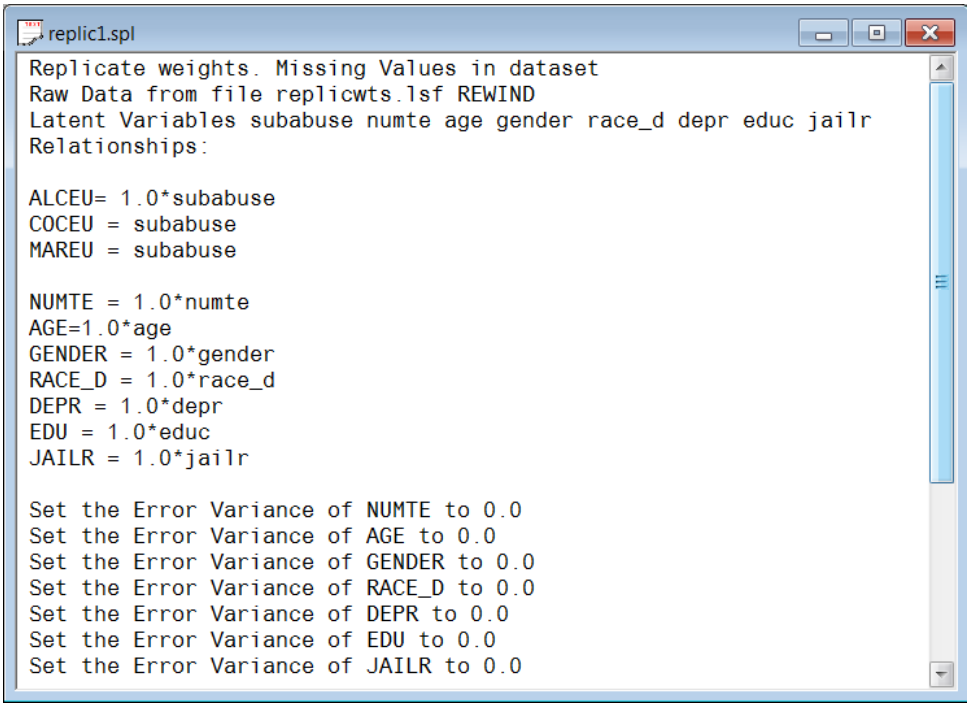
$$Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{21} \\ \psi_{21} & \psi_{22} \end{bmatrix}.$$

The unknown model parameters are therefore $\lambda_{y,21}$, $\lambda_{y,31}$, β_{12}, γ_{21}, γ_{22}, γ_{23}, γ_{24}, γ_{25}, γ_{26}, ϕ_{11}, ϕ_{21}, ϕ_{22}, …., ϕ_{66}, ψ_{11}, ψ_{21}, ψ_{22}, $\text{var}(\varepsilon_1)$, $\text{var}(\varepsilon_2)$, and $\text{var}(\varepsilon_3)$. In subsequent output, these 36 parameters will be denoted by the symbols LY21, LY31, BETA12, GAMMA21, GAMMA22, …., GAMMA26, PHI11, PHI21, …., PHI66, PSI11, PSI21, PSI22, TE11, TE22 and TE33.

3.1 Setting up the analysis using SIMPLIS syntax

It is relatively easy to specify the model described above with SIMPLIS syntax. We start by indicating that the raw data is to read from the file **replicwts.LSF**. Note that this is followed by a REWIND command which allows LISREL to repeatedly read the raw data from the same file.

The program commands AGE = 1.0*age, …, JAILR=1.0*jailr specify that each of the X-variables are assumed to be exactly equal to the corresponding latent variable. Note that the part 1.0*latent variable indicates that the path coefficient is fixed at the value of 1.0. In contrast, a command such as COCEU = (0.5)*subabuse indicates that 0.5 is a starting value (preliminary estimate) of $\lambda_{y,21}$. Since we assume that the X-variables measure the corresponding KSI latent variables without error, we set the error variances of NUMTE to JAILR to 0.



Next, we assume that subabuse is predicted by numte, and, in turn, numte is predicted by age, gender, race_d, depr, educ and jailr. This part of the syntax is shown below.

replic1.spl

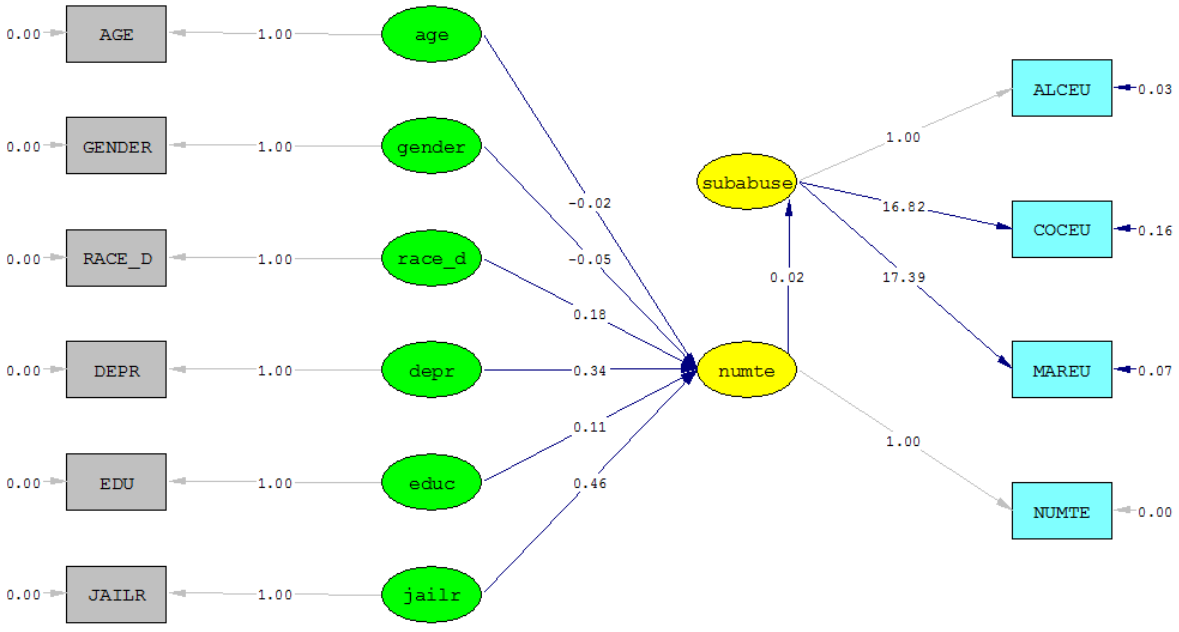
```
!--------------------------------!
! Use numte as mediating variable !
!--------------------------------!

subabuse = numte
numte = age gender race_d depr educ jailr
Set the Error Covariance of numte and subabuse Free

LISREL OUTPUT: RP=1 ND=4 AD=OFF SV=replic1.std PV=replic1.par
Path Diagram
End of Problem
```

To allow for the estimation of the covariance of the errors between the ETA latent variables numte and subabuse, we SET the error covariance between these variables free.

Finally, we use the LISREL OUTPUT: command to specify the number of repetitions (RP = 1), number of decimals (ND = 4), admissibility check off (AD = OFF), and save the standard error estimates (SV = replic1.std) and parameter estimates (PV = replic1.par) to text files **replic1.std** and **replic1.par** respectively.

3.2 Discussion of results

The estimated path coefficients for weight = A2TWA0 are shown in the path diagram given below. Although not presented here, all coefficients are statistically significant. The χ^2-statistic for goodness of fit is 113.20, degrees of freedom is 19, and the RMSEA-value is 0.031.



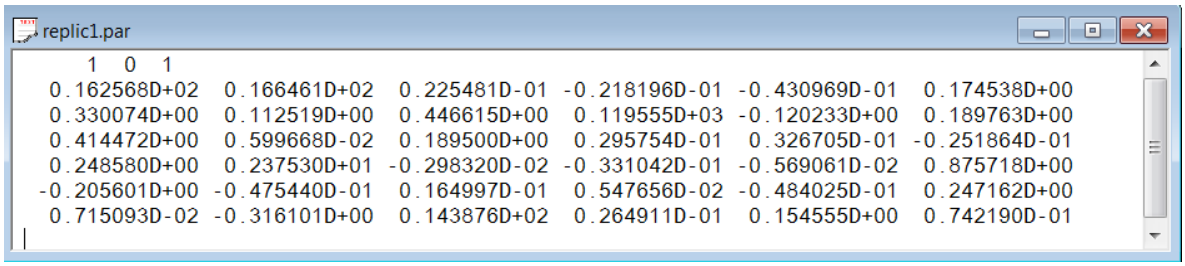
Chi-Square=113.20, df=19, P-value=0.00000, RMSEA=0.031

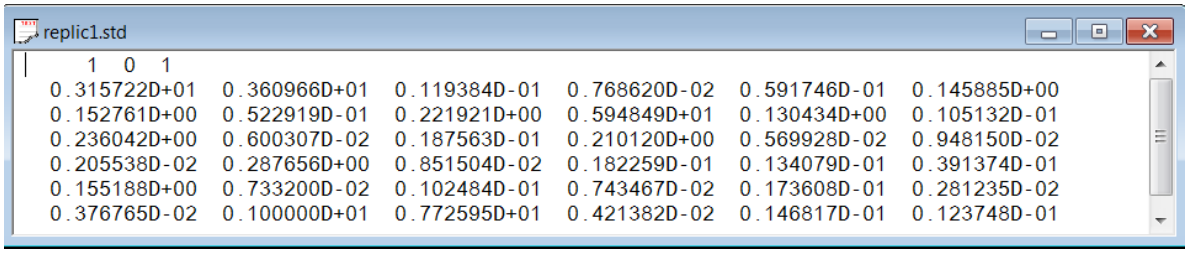
Contents of the files **replic1.par** and **replic1.std** are given below. The estimates are preceded by three numbers N1, N2, and N3, where N1 = repetition number, while N2 and N3 are zero if convergence has been attained. The parameter estimates and standard error values are given in scientific notation. For example,

$$0.165949D+02 = 0.165942\times10^{2} = 16.5942$$

$$0.234136D+0 = 0.23413\times10^{0} = 0.23413$$

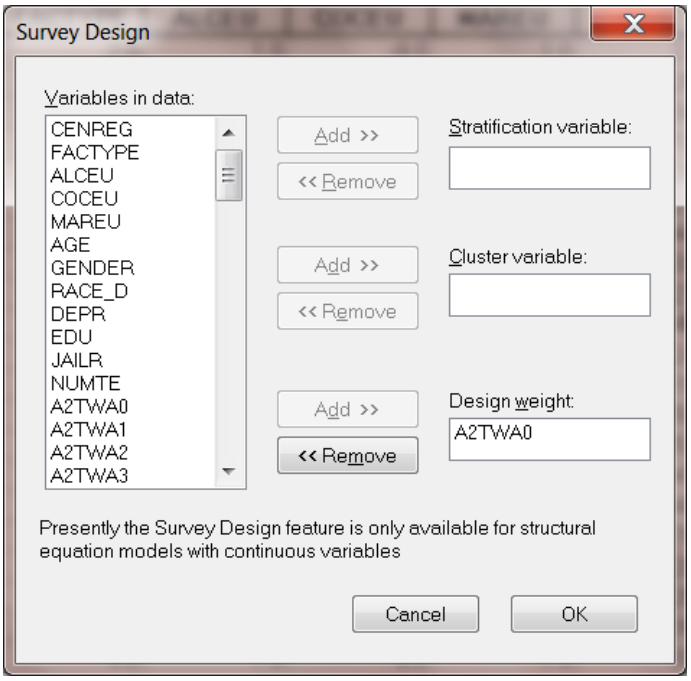
In general, $D+k$ implies that the decimal point should be moved k positions to the right, whereas $D-k$ implies the insertion of k zero values just after the decimal point. For example, $0.868283D-03 = 0.000868283$.





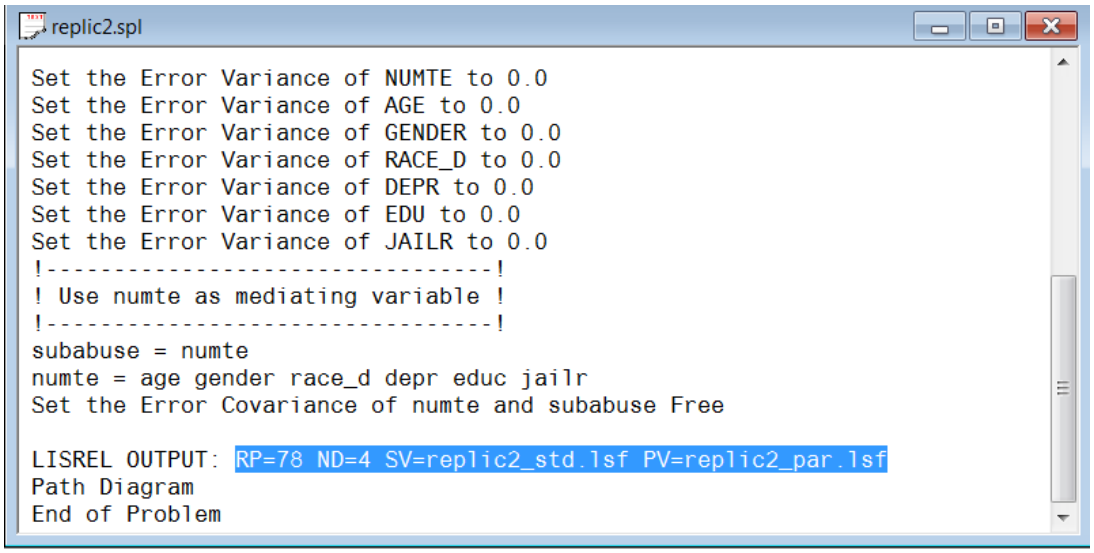
4. Use of replicate weights

We now demonstrate the use of replicate weights in LISREL and start by making the LSF window the active window in order to invoke the LSF menu bar. From the **Data** menu, select **Survey Design** and remove the stratum variable CENREG and the cluster variable FACTYPE by clicking on the corresponding **Remove** buttons. Select the **File**, **Save** option to save this change.



Next, copy **replic1.spl** to **replic2.spl** and edit this SIMPLIS syntax file by changing RP = 1 to RP = 79 (shown in bold typeface below) and use the filenames **replic2.std** and **replic2.par** for saving the standard errors and parameter estimates respectively.

It is important to note that LISREL assumes that there are a total of 78 additional weight variables in the data set, starting with the **Design weight** variable name (in the present case A2TWA0) selected in the **Survey Design** dialog box. Additionally, it is assumed that the weight columns follow one another. Note that we are saving the estimated parameters and standard errors as Lisrel system data files (***.lsf**).



4.1 Discussion of results

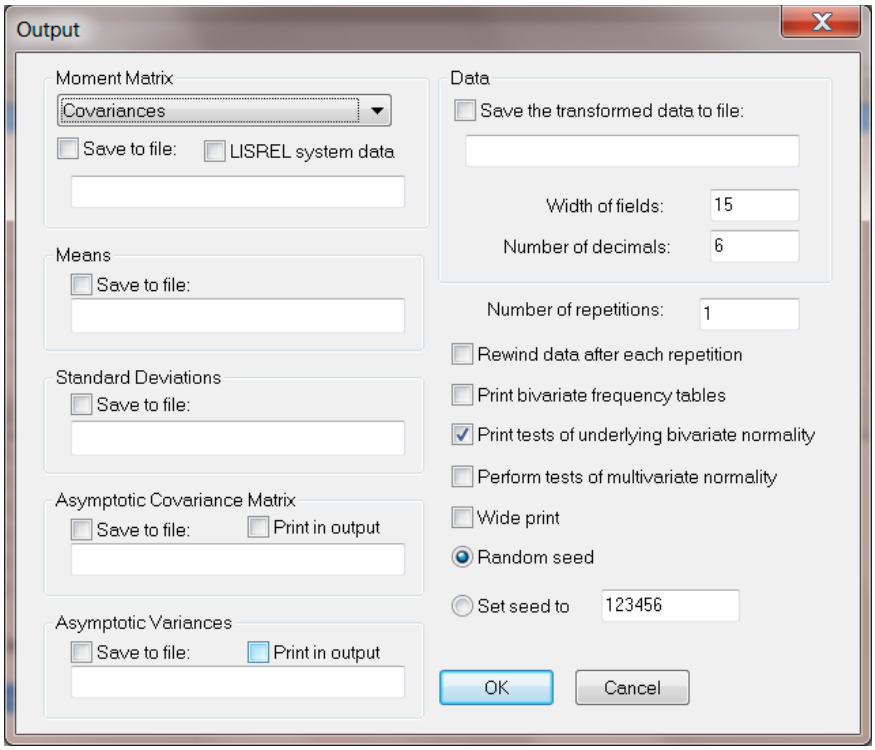
The results below show the parameter estimates and parameter standard error estimates for the model fitted with design weight A2TWA0, including stratification (CENREG) and clustering (FACTYPE) variables, and the corresponding average values using replicate weights without CENREG and FACTYPE.

Results of the parameters estimates for the first 10 repetitions are shown below.

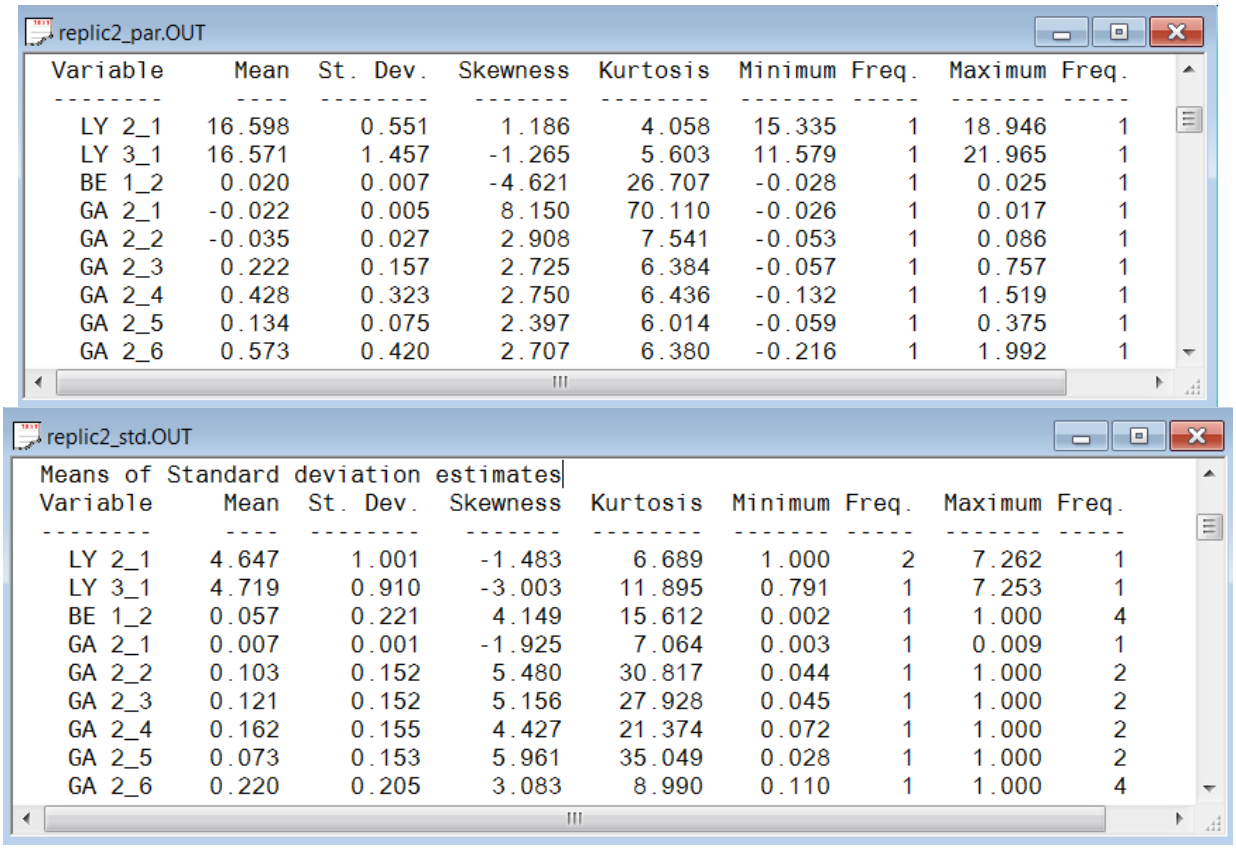
replic2_par.lsf

	LY 2_1	LY 3_1	BE 1_2	GA 2_1	GA 2_2	GA 2_3	GA 2_4	GA 2_5	GA 2_6
1	16.257	16.646	0.023	-0.022	-0.043	0.175	0.330	0.113	0.447
2	17.684	12.446	0.006	-0.021	0.054	0.757	1.519	0.375	1.983
3	17.631	12.937	0.007	-0.025	0.037	0.702	1.404	0.358	1.832
4	16.414	16.725	0.021	-0.023	-0.047	0.195	0.348	0.117	0.470
5	16.459	16.642	0.021	-0.023	-0.048	0.197	0.361	0.122	0.485
6	17.262	12.760	0.007	-0.026	0.025	0.701	1.393	0.355	1.807
7	17.459	12.577	0.006	-0.023	0.041	0.730	1.461	0.366	1.899
8	17.351	12.653	0.007	-0.024	0.033	0.720	1.432	0.361	1.858
9	16.539	16.767	0.020	-0.024	-0.046	0.202	0.366	0.124	0.497
10	16.303	16.406	0.020	-0.024	-0.046	0.205	0.385	0.129	0.516

Means of these estimates (and likewise for the standard error estimates) can be obtained by selecting **Statistics, Output Options…** from the main menu bar. This selection produces the **Output** dialog. Select **Covariances** and click **OK** to run PRELIS.



Portions of the output for the means of the parameter estimates and means of the standard error estimates are listed below.



replic2_par.OUT

Variable	Mean	St. Dev.	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
LY 2_1	16.598	0.551	1.186	4.058	15.335	1	18.946	1
LY 3_1	16.571	1.457	-1.265	5.603	11.579	1	21.965	1
BE 1_2	0.020	0.007	-4.621	26.707	-0.028	1	0.025	1
GA 2_1	-0.022	0.005	8.150	70.110	-0.026	1	0.017	1
GA 2_2	-0.035	0.027	2.908	7.541	-0.053	1	0.086	1
GA 2_3	0.222	0.157	2.725	6.384	-0.057	1	0.757	1
GA 2_4	0.428	0.323	2.750	6.436	-0.132	1	1.519	1
GA 2_5	0.134	0.075	2.397	6.014	-0.059	1	0.375	1
GA 2_6	0.573	0.420	2.707	6.380	-0.216	1	1.992	1

replic2_std.OUT

Means of Standard deviation estimates

Variable	Mean	St. Dev.	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
LY 2_1	4.647	1.001	-1.483	6.689	1.000	2	7.262	1
LY 3_1	4.719	0.910	-3.003	11.895	0.791	1	7.253	1
BE 1_2	0.057	0.221	4.149	15.612	0.002	1	1.000	4
GA 2_1	0.007	0.001	-1.925	7.064	0.003	1	0.009	1
GA 2_2	0.103	0.152	5.480	30.817	0.044	1	1.000	2
GA 2_3	0.121	0.152	5.156	27.928	0.045	1	1.000	2
GA 2_4	0.162	0.155	4.427	21.374	0.072	1	1.000	2
GA 2_5	0.073	0.153	5.961	35.049	0.028	1	1.000	2
GA 2_6	0.220	0.205	3.083	8.990	0.110	1	1.000	4

In general, parameter estimates for the two estimation methods are quite close. Standard error estimates, on the other hand, tend to be larger for the replicate method. This is clearly a topic that requires further research.