Resampling Methods in Mplus for Complex Survey Data

Tihomir Asparouhov and Bengt Muthén

May 4, 2010

1 Introduction

In this note we describe the replicate weights methodology implemented in Mplus Version 6. Replicate weights are used to compute the standard errors in analysis of complex survey data. The replicate weights can be seen as bootstrap samples that can be used to assess the variability of the estimates.

Suppose that a data set is obtained with complex sampling. To estimate correctly the standard errors for the parameters estimates a detailed description of the sampling method is usually needed. The sampling method can be quite complicated so that a secondary analyst is faced with several difficult tasks: obtaining all the sampling information from the sampling agency, interpreting the sampling method correctly, and finding appropriate methodology to use for this sampling method. Often the sampling method description is sufficient for some estimation problems but not for others. All of these problems can be resolved by the replicate weights method. The replicate weights are produced using the complete sampling information available to the agency that collected the data. The agency then distributes the data to secondary analysts without having to produce a detailed account of the data collection mechanism. This approach has the advantage that it will allows the secondary analyst to analyze the data correctly even when the analytical methods do not generally support the original survey method. Another advantage of this approach is that it prevents any possible misunderstandings about the sampling methodology that can occur between the agency, the analyst, and the software vendors that produce the analytical techniques. In addition with the replicate weights method the agency does not produce sampling information that could potentially compromise the anonymity of the data collection, such as strata and clustering information. The replicate weights should not be confused with the original sampling weights that reflect the actual sample and the actual sampling scheme that was carried out by the agency. The original sampling weight is also included in the distributed data sets along with the replicate weights.

If replicate weights are already available with the data Mplus can compute the standard errors for the estimated parameters. Mplus supports any of these four types of replicate weights: Bootstrap, Jackknife, BRR and Fay. The replicate weights are specified with the *repweight* command and the proper replicate weights method is specified with the *repse* command. Note that it is essential to specify the replicate weights method correctly, otherwise the replicate weights will produce incorrect results. More information on the replicate weights methodology can be found in Korn and Graubard (1999), Lohr (1999) and Fay (1989).

Mplus computes the standard errors for the parameters using the following method. Let θ be a model parameter. Let $\hat{\theta}$ be the estimate for that parameter using the original weight variable. Let $\hat{\theta}_r$ be the estimate for that parameter using the $r-$th replicate weight. The standard error for that parameter is computed as follows

$$\sqrt{\sum_{r=1}^{R} C_r(\hat{\theta}_r - \hat{\theta})^2} \qquad (1)$$

where the constants C_r depend on the method used for computing the replicate weights. The value for C_r is given in the following table. The values of the constants C_r are independent of r with the exception of the general Jackknife method. The Jackknife method in Mplus has three different versions: Jackknife, Jackknife1, and Jackknife2. The Jackknife refers to the general Jackknife method. For this method the constants C_r are read in by Mplus with a special file given with the *multiplier* command. This is the only case where the user is required to provide the multiplier file. The Jackknife1 method refers to the special case when the sample has only one stratum. In that case the C_r are independent of r and the value is given in the table. Jackknife2 refers to the special case when all strata have exactly two PSUs and only one of the Jackknife samples are used for computing the standard errors, i.e., for each stratum only one of the two PSUs is removed to produce a Jackknife sample. In this case $C_r = 1$. The BRR and Fay's methods are also applicable only for the special case when there are exactly two PSUs in each strata. Fay's method also requires the specification of a constant K, which also affects C_r. The default for this constant is 0.3.

2 Direct Resampling Methods without Replicate Weights

If the sampling structure is available in the data, i.e., if the strata and PSU (cluster) variable are available in the sample then Mplus can perform the Bootstrap, Jackknife, BRR and Fay variance estimation by generating the appropriate replicate weights and using the standard error computation described in the previous section. Mplus generates the replicate weights using

Table 1: Replication Methods

Method	C_r
Bootstrap	$1/(R-1)$
Jackknife	multiplier file
Jackknife1	$R/(R-1)$
Jackknife2	1
BRR	$1/R$
Fay	$1/(R(1-K)^2)$

the methods described in Stapleton (2008). Following is a brief description on how the replicate weights are constructed in Mplus.

2.1 Bootstrap

In complex sampling the bootstrap is implemented on the PSU level and the bootstrapping is done separately for each stratum. Suppose that there are H strata and that stratum h contains K_h PSUs. Each bootstrap draw is formed by selecting K_h-1 PSUs with replacement from the original K_h PSUs in stratum h. The new replicate weights are obtained using the following formula. The replicate weight for individual i in this bootstrap draw is

$$w_i f \frac{K_h}{K_h - 1}$$

where w_i is the original weight and f is the number of times the PSU that the individual belongs to was selected in the bootstrap draw. If there are strata with only one PSU then the PSU is treated as self-representing and the replicate weights in that PSU are all the same as the original sampling weight.

The number of bootstrap draws is specified in the Mplus input file using the *bootstrap* command. The number of bootstrap draws is typically between 100 and 500.

The standard errors are computed as

$$\sqrt{\sum_{r=1}^{R} \frac{1}{R-1} (\hat{\theta}_r - \hat{\theta})^2} \qquad (2)$$

2.2 Jackknife

Suppose that there are K PSUs in the sample. We form K Jackknife draws, i.e., with this method the number of replicate weights is predetermined by the data and it is not possible to change it. The $j-$th draw is formed by removing the $j-$th PSU from the sample and adjusting the weights in that stratum. Suppose that the $j-$th PSU belongs to stratum h. The replicate weights for the $j-$th Jackknife draw are computed as follows. For each individual not belonging to stratum h the replicate weights are the same as the original weights. For each individual in stratum h the sampling weight is computed as follows

$$w_i f \frac{K_h}{K_h - 1}$$

where w_i is the original sampling weight and the factor f is 0 if the individual belongs to the $j-$th PSU and 1 otherwise. The standard errors of the estimated parameters are computed as follows

$$\sqrt{\sum_{h=1}^{H} \sum_{j=1}^{K_h} \frac{K_h - 1}{K_h} (\hat{\theta}_{jh} - \hat{\theta})^2}$$

where θ_{jh} is the parameter estimate from the Jackknife draw that removes the $j-$th PSU in stratum h. Therefore the constant C_r used in formula (1) is simply $(K_h - 1)/K_h$. When using the Jackknife method with replicate weights this constant should be stored in the multiplier file. If there are strata with a single PSU they are treated as self-representing. For these strata Mplus does not create a Jackknife sample.

In the case when there are exactly two PSUs in each stratum the Jackknife2 method uses a different approach for generating the replicate weights and uses a different formula for computing the standard errors. In this method the number of replicate weights is $K/2$. For each stratum we create only one Jackknife sample by removing only the first PSU (assuming that the PSUs come in a random order), instead of the two Jackknife samples that are created with the regular Jackknife method. The replicate weights are computed as in the regular Jackknife method. The standard errors are computed as follows

$$\sqrt{\sum_{h=1}^{H} (\hat{\theta}_{1h} - \hat{\theta})^2}.$$

2.3 BRR and Fay

The BRR and Fay methods are available when each stratum contains exactly 2 PSUs. The BRR replicate weights are constructed from Hadamard matrices H. These are matrices that contain only 1 and -1 as entries and

$$HH^T = sI_s$$

where s is the size of the matrices. Such matrices exist only if s is 1, 2 or a multiple of 4. For most multiples of 4 such matrices are available but not for all. Mplus stores Hadamard matrices of size 4, 8, 12, ..., 88. For larger numbers Mplus constructs matrices that are of size $2^k n$ where $n \leq 88$ and is a multiple of 4. Using one of the available Hadamard matrices the replicate weights are constructed as follows. Suppose that there are L strata in the sample. Mplus would use the smallest Hadamard matrix of size greater than L. For example if there are 50 strata in the sample Mplus would use H of size 52. If there are 150 strata Mplus would use H of size 152. If there are 152 strata Mplus would use H of size 160. With the BRR resampling method there are as many replicate weights as the size of the Hadamard matrix. The $i-$th BRR sample is created as follows. If H_{ij} is 1 then we remove the second PSU in stratum j. If H_{ij} is -1 then we remove the first PSU in stratum j. Thus the $i-$th sample contains exactly half of the original PSUs, one from each stratum. The replicate weights are constructed as follows. In each BRR sample the weights of the PSUs that are in the sample are doubled and the weights of the PSUs that are not in the sample are set to zero. The standard errors are computed as follows

$$\sqrt{\sum_{r=1}^{R} \frac{1}{R}(\hat{\theta}_r - \hat{\theta})^2}. \tag{3}$$

Fay's method is constructed similarly using the same size Hadamard matrix as in the BRR method. The $i-$th replicate weights are constructed as follows. If H_{ij} is 1 the weights in the first PSUs in stratum j are multiplied by $2 - K$ and the weights in the second PSUs are multiplied by K. If H_{ij} is -1 the weights in the first PSUs in stratum j are multiplied by K and the weights in the second PSUs are multiplied by $2 - K$. The standard errors are then computed as follows

$$\sqrt{\sum_{r=1}^{R} \frac{1}{R(1-K)^2}(\hat{\theta}_r - \hat{\theta})^2}. \tag{4}$$

3 Example

In this section we illustrate the replication methodology using the ECLS example described in Stapleton (2008). The data has 15757 individual data points nested within 525 clusters and 89 strata. In this example we can use the Bootstrap method and the Jackknife method as well as the PML method using the linearization methodology for computing the standard errors. In addition we can also use the replicate weights provided with the data for use with the Jackknife2 method.
Following is the input file for estimating this SEM model with the bootstrap method

```
DATA: FILE IS ECLSdata2.dat;

VARIABLE: NAMES=C1RGSCAL C1RMSCAL C1RRSCAL P1NUMPLA
C1CPTW0 C1CPTSTR NEWPSU;
WEIGHT=C1CPTW0;
CLUSTER=NEWPSU;
STRAT=C1CPTSTR;

ANALYSIS: TYPE=COMPLEX;
REPSE=BOOTSTRAP;
BOOTSTRAP=500;

MODEL: ACHIEVE BY C1RGSCAL C1RMSCAL C1RRSCAL;
ACHIEVE ON P1NUMPLA;
```

Following is the input file for estimating this SEM model with the Jackknife method

```
DATA: FILE IS ECLSdata2.dat;

VARIABLE: NAMES=C1RGSCAL C1RMSCAL C1RRSCAL P1NUMPLA
C1CPTW0 C1CPTSTR NEWPSU;
WEIGHT=C1CPTW0;
CLUSTER=NEWPSU;
STRAT=C1CPTSTR;
```

```
ANALYSIS: TYPE=COMPLEX;
REPSE=JACKKNIFE;

MODEL: ACHIEVE BY C1RGSCAL C1RMSCAL C1RRSCAL;
ACHIEVE ON P1NUMPLA;
```

Following is the input file for estimating this SEM model with the linearization PML method

```
DATA: FILE IS ECLSdata2.dat;

VARIABLE: NAMES=C1RGSCAL C1RMSCAL C1RRSCAL P1NUMPLA
C1CPTW0 C1CPTSTR NEWPSU;
WEIGHT=C1CPTW0;
CLUSTER=NEWPSU;
STRAT=C1CPTSTR;

ANALYSIS: TYPE=COMPLEX;

MODEL: ACHIEVE BY C1RGSCAL C1RMSCAL C1RRSCAL;
ACHIEVE ON P1NUMPLA;
```

Following is the input file for estimating this SEM model with the 90 replicate weights provided with the data for use with the Jackknife2 method.

```
DATA: FILE IS ECLS_repw.dat;

VARIABLE: NAMES=C1RGSCAL C1RMSCAL C1RRSCAL P1NUMPLA
C1CPTW0 C1CPTSTR NEWPSU W1-W90;
USEVAR=C1RGSCAL C1RMSCAL C1RRSCAL P1NUMPLA;
WEIGHT=C1CPTW0;
REPWEIGHT=W1-W90;

ANALYSIS: TYPE=COMPLEX;
REPSE=JACKKNIFE2;
```

Table 2: Comparison of the standard errors for the SEM model using the Bootstrap, Jackknife, PML method, and the agency provided replicate weights with the Jackknife2 method.

Parameter	Bootstrap	Jackknife	PML	Jackknife2
λ_2	0.014	0.013	0.013	0.012
λ_3	0.013	0.013	0.013	0.012
β	0.053	0.051	0.051	0.049
θ_1	0.824	0.875	0.873	0.933
θ_2	0.5	0.474	0.475	0.499
θ_3	0.553	0.52	0.519	0.549
ψ	1.634	1.626	1.618	1.423

MODEL: ACHIEVE BY C1RGSCAL C1RMSCAL C1RRSCAL;
ACHIEVE ON P1NUMPLA;

The standard errors of the four methods are presented in Table 2. All four methods produced similar results.

References

[1] Fay, R.E. (1989). Theoretical application of weighting for variance calculation. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 212-217.

[2] Korn, E.L. & Graubard, B.I. (1999). Analysis of health surveys. New York: John Wiley & Sons.

[3] Lohr, S.L. (1999). Sampling: Design and analysis. Pacific Grove, CA: Brooks/Cole Publishing Company.

[4] Stapleton, L. M. (2008) Variance Estimation Using Replication Methods in Structural Equation Modeling With Complex Sample Data, Structural Equation Modeling: A Multidisciplinary Journal 15, 183 - 210.