Multiple Imputation Example with Regression Analysis

Below I illustrate multiple imputation with SPSS using the Missing Values module[1] and R using the `mice` package. I used some of the variables from the study on school health behavior: the student's perceptions about how easy it is to talk to parents, `partalk`; whether parents are willing to help with homework, `hwhelp`; the number of friends, `friends`; and Hispanic ethnic identity, `hispanic`.[2] There are a variety of specific algorithms for the imputation step in MI. The methods illustrated here are called fully conditional specification (aka "chained equations" or sequential regression method or Bayesian linear regression). I am leaving out a number of descriptive analyses that are possible in both packages for examining the missing data patterns and imputation step.

SPSS

```
DATASET DECLARE i0.
MULTIPLE IMPUTATION partalk hwhelp friends hispanic
/IMPUTE MAXITER=20 NIMPUTATIONS=20 SINGULAR=1E-008
/OUTFILE IMPUTATIONS=i0.
DATASET ACTIVATE i0.
REGRESSION
/STATISTICS COEFF OUTS R ANOVA
/DEPENDENT partalk
/METHOD=ENTER hwhelp friends hispanic.
```

Imputation Specifications

Imputation Method	Automatic
Number of Imputations	20
Model for Scale Variables	Linear Regression
Interactions Included in Models	(none)
Maximum Percentage of Missing Values	100.0%
Maximum Number of Parameters in Imputation Model	100

Imputation Results

Imputation Method		Fully Conditional Specification
Fully Conditional Specification Method Iterations		10
Dependent Variables	Imputed	partalk,hwhelp,friends,hispanic
	Not Imputed(Too Many Missing Values)	
	Not Imputed(No Missing Values)	
Imputation Sequence		hispanic,friends,partalk,hwhelp

[1] An IBM SPSS module sold separately from the main SPSS package.

[2] A random subsample of the full sample was used for this data set from a WHO study of US school children. United States Department of Health and Human Services. Health Resources and Services Administration. Maternal and Child Health Bureau. Health Behavior in School-Aged Children, 2001-2002 [United States]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2008-07-24. https://doi.org/10.3886/ICPSR04372.v2

Model Summary

Imputation_ Imputation Number	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
0 Original data	1	.288 [a]	.083	.080	.84345
1	1	.274 [a]	.075	.072	.84305
2	1	.279 [a]	.078	.075	.83823
3	1	.281 [a]	.079	.076	.84516
4	1	.276 [a]	.076	.073	.83826
5	1	.279 [a]	.078	.075	.84648
6	1	.274 [a]	.075	.072	.84098
7	1	.282 [a]	.080	.077	.82729
8	1	.290 [a]	.084	.081	.83654
9	1	.298 [a]	.089	.086	.84188
10	1	.272 [a]	.074	.071	.84058
11	1	.265 [a]	.070	.067	.82981
12	1	.284 [a]	.081	.078	.82834
13	1	.278 [a]	.077	.075	.83333
14	1	.281 [a]	.079	.076	.83680
15	1	.284 [a]	.081	.078	.84002
16	1	.286 [a]	.082	.079	.83772
17	1	.302 [a]	.091	.088	.85316
18	1	.274 [a]	.075	.072	.84389
19	1	.279 [a]	.078	.075	.84750
20	1	.288 [a]	.083	.080	.84264

a. Predictors: (Constant), hispanic ethnicity, friends sum of close male and female friends, hwhelp parents willing to help with homework

Coefficients[a]

Imputation_ Imputation Number	Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Fraction Missing Info.	Relative Increase Variance	Relative Efficiency
0 Original data	1	(Constant)	1.736	.156		11.117	.000			
		hwhelp parents willing to help with homework	.227	.027	.278	8.252	.000			
		.								
		.								
		.								
		hispanic ethnicity	-.075	.067	-.035	-1.125	.261			
Pooled	1	(Constant)	1.754	.152		11.523	.000	.075	.080	.996
		hwhelp parents willing to help with homework	.221	.027		8.196	.000	.087	.095	.996
		friends sum of close male and female friends	.037	.042		.891	.373	.058	.061	.997
		hispanic ethnicity	-.081	.072		-1.118	.264	.147	.169	.993

a. Dependent Variable: partalk ave ease of talking to parents

R

```
library(mice)
#look at missing data patterns
#present data indicated by 1, missing data indicated by 0
md.pattern(d)

#based on example at https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

#mice imputation wants numeric variables rather than haven labeled
d$partalk=as.numeric(d$partalk)
d$hwhelp=as.numeric(d$hwhelp)
d$friends=as.numeric(d$friends)
d$hispanic=as.numeric(d$hispanic)


#I-step: impute 20 data sets (m=20), maxit is maximum iterations, method is type used (norm is Bayesian linear regression),
#seed sets a random number generator start for replication
impdata <- mice(d, m=20, maxit = 50, method = 'norm', seed = 500)
summary(impdata)

#build predictive model
fit <- with(data = impdata, exp = lm(partalk ~ hwhelp + friends + hispanic))

#combine results of all 5 models
combine <- pool(fit)
summary(combine)
```

```
Class: mids
Number of multiple imputations:  20
Imputation methods:
 partalk   hwhelp  friends hispanic
  "norm"   "norm"   "norm"   "norm"
PredictorMatrix:
         partalk hwhelp friends hispanic
partalk        0      1       1        1
hwhelp         1      0       1        1
friends        1      1       0        1
hispanic       1      1       1        0
```

```
> #build predictive model
> fit <- with(data = impdata, exp = lm(partalk ~ hwhelp + friends + hispanic))
>
> #combine results of all 5 models
> combine <- pool(fit)
> summary(combine)
         term    estimate  std.error  statistic       df                p.value
1 (Intercept)  1.76054523 0.15409821 11.4248261 618.0146 0.00000000000000000000
2      hwhelp  0.21915069 0.02770921  7.9089485 472.3070 0.00000000000001865175
3     friends  0.03864522 0.04246017  0.9101524 710.8288 0.36305067671685664621
4    hispanic -0.08905508 0.07367383 -1.2087750 356.3667 0.22755072014954680348
```

Resources

Analytics Vidya: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

Martijn Heymans: https://bookdown.org/mwheymans/bookmi/data-analysis-after-multiple-imputation.html

Wang, J., & Johnson, D. E. (2019). An examination of discrepancies in multiple imputation procedures between SAS® and SPSS®. *The American Statistician*, *73*(1), 80-88.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1-67. https://www.jstatsoft.org/article/view/v045i03

Stef Van Buuren's *Flexible Imputation of Missing Data, Second Edition*, free online, https://stefvanbuuren.name/fimd/ch-univariate.html