IGNORING IT DOESN'T MAKE IT GO AWAY

Addressing Issues of Missing Data in Institutional Research

Jacob P. K. Gross
Afet Dadashova
John Moore
Mary Ziskin

IU Bloomington

AIR Annual Forum 2009 ◆ Atlanta

Overview

- What is missing data & the techniques to deal with it?
- Research questions on different missing data techniques:
  - Differences on statistical measures between multiple imputation (MI) and listwise deletion (LD) techniques?
  - Practical implications?
  - Differences among techniques with large and small sample sizes?
- Discussion of results and recommendations, focusing on use of expectation maximization (EM) algorithm imputation

Introduction

- Missing data issues are persistent and important in any form of research, whether it relies on primary or secondary data
- Missing data issues, if unaddressed, can bias results, leading to wrong conclusions or bad policy recommendations
- Social science researchers often ignore missing data issues
- Institutional researchers need to develop not only deep knowledge of why data are missing but also of how to deal with missing data issues

Background

- Types of missing data
  - Missing completely at random (MCAR)
    - The missingness of data totally unrelated to the variable itself or any other variables in the model
  - Missing at random (MAR)
    - The missingness of data unrelated to variable itself, BUT may be related to other variable(s) in the model
  - Not missing at random (NMAR)
    - The missingness of the data related to BOTH the variable itself AND to the other model variable(s)

Background (continued)

- Types of missing data (continued)
  - Some caveats
    - May be (and probably are) multiple reasons for missing data in a given data set
    - Often no true way of knowing why data are missing; thus, need to construct theoretical arguments for theories of missingness
    - Any type of missing data can threaten validity of analyses, even MCAR

Background (continued)

- Dealing with missing data
  - Strategies depend on type, pattern, and amount of missingness
  - Strategies
    - Listwise/pairwise deletion (LD)
    - Multiple imputation (MI)
    - Dummy variable adjustment
    - Single imputation
    - Expectation maximization (EM) algorithm imputation

Listwise/Pairwise Deletion (LD)

- Details
  - Eliminates cases with missing variables of interest
  - Appropriate if data are MCAR; creates a random subsample of data set
  - If data are MCAR will not alter data set parameters; reasonable estimate of parameters even with some deviations from MCAR
- Positives
  - For any type of analysis
  - Easy to use; default in most statistical packages
- Negatives
  - Reduces statistical power
  - As missingness increases, more likely that resulting data differ from original
  - With pairwise, estimates vary by statistic calculated

Multiple Imputation (MI)

- Details
  - Imputed number values randomly assigned to keep parameter estimates
  - Uses variety of techniques to create multiple data sets combined later
- Positives
  - Keeps parameter estimates while allowing variation
  - Can be used even with large amounts of missing data
- Negatives
  - One of the more challenging techniques
  - Included in only a limited number of statistical packages

Other Techniques

- Dummy variable adjustment
  - Treats “missing” as its own value
  - Assumes similar reasons for missing data
- Single imputation
  - Replaces missing values with some calculated number
    - Mean
    - Conditional mean
    - Regression outcome

Models & Methods

- Persistence model
  - Logistic regression
  - Year-to-year
- Multiple imputation
  - PROC MI in SAS v9 (six imputations)
  - PROC MI ANALYZE

Student Background (x_1)	Academic Preparation (x_2)	College Enrollment Characteristics (x_3)	Financial Aid (x_4)
•Gender •Race/ethnicity •Age •Income*	•SAT Score* •High School Rank* •High School Diploma	•Institution type •Cumulative GPA* •Residential status •Indiana Residency •Developmental Education •Credits attempted	•Grants •Loans •Private Gift-aid •Work-Study

* Denotes imputed variables.

$$Persistence = \alpha + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \varepsilon_i$$

Data Source

- Indiana Commission for Higher Education (ICHE) statewide student unit record database
  - Student information system (SIS) of all Indiana public universities, colleges, and community colleges
  - Enrollment-related transactions, e.g., demographics, course-taking, test scores, financial aid information
  - SIS data represent the universe of enrolled students at public postsecondary institutions in Indiana

Population of Interest

- 1999 cohort of first-time, first-year (i.e., freshman) students from all Indiana public postsecondary institutions (n=37,039)
- For small sample, we took about a five-percent random cut (n=1,827) of the full sample

Missingness of Data

- Financial information (income, budget, parental and student contribution)
- Academic preparation information (SAT and high school class rank)
- Cumulative college grade point average

Variable	Proportion Missing
SAT Score - Math	68%
Student Expense Budget/Summer Session	67%
SAT Score - Verbal	46%
Parent Contribution	25%
High School Rank	24%
Student Expense Budget/Academic Year	23%
Student Contribution	22%
Total Income (Independent Students) or Family's Total Income (Dependent Students)	22%
Cumulative Grade Point Average	5%

Results for Full Sample

	LD		MI		Change in our Conclusions
	OR	Sig.	OR	Sig	
Compared to Whites					
African American	1.07		0.88	**	African Americans are less likely to persist than Whites.
Not Reported	1.12		1.23	***	Students with no reported race, compared to Whites, are more likely to persist.
Age	0.98	****	1.02	****	A one year increase in age is associated with an increased likelihood of persisting.
High School Rank	1.00		1.00	***	An increase in high school rank is associated with an increased likelihood of persisting.
Compared to Regular HS Diploma					
Honors	1.35	**	1.66	****	Students who complete a more rigorous college prep curriculum more likely to perist.
Core 40	1.10		1.26	****	
Not Reported	0.70	****	1.02		
GED, Other	2.44		0.62	****	Students who completed a GED were less likely to persist
Compared to dependent					
Independent	1.02		1.32	****	Independent students are more likely to persist.
Undetermined	1.04		0.75	****	Students who did not apply for aid were less likely to persist.
N=	23,164		37,039		

Results for Small Sample

	LD		MI		Change In Our Conclusions
	OR	Sig.	OR	Sig	
Compared to Whites					
Native American/Other	-		0.02		No change, but note we now obtain an estimate.
Adjusted Gross Income ($1,000s)	1.00	*	1.00		Income not significantly related to persistence
Compared to Regular HS Diploma					
GED, Other	-		0.16	***	GED recipients less likey to persist than Regular HS diploma
Cumulative Grade Point	2.37	****	1.66	****	Magnitude of GPA is not as great in the MI model
Compared to attending a research university					
Community college	0.38	**	0.76		No significant difference between community college and reseach university.
Credits Attempted	1.02		1.02	**	An increase in credits attempted is associated with increased likelihood of persistence
Compared to dependent students					
Undetermined	0.57	*	0.73		No significant difference between dependent students and those who do not apply for aid.
N=	1,169		1,827		

****p<0.001, ***p<0.01, **p<0.05, *p<0.10

Statistical Implications

- For small sample, MI preserves 658, or 36%, of the cases (total n =1,827)
- For large sample, MI preserves 13,875, or 37%, of the cases (total n=37,039)
- Standard errors
  - Smaller in MI than in LD
  - Not as small as if original data set were completely observed
  - MI retrieves some of the lost information (Sinharay, Stern, & Russell, 2001)

Statistical Implications (continued)

- CI intervals narrower in MI; MI is more efficient
- For some variables LD and MI uncovered different relationships with the dependent variable, as demonstrated by null hypothesis testing and odds ratio
- "…MI captures the uncertainty associated with missing values and consequently leads to more valid statistical inferences in terms of both null hypothesis testing and interval estimation of the regression coefficients" (Peng, Harwell, Liou, & Ehman, 2006, p. 67)

Conclusions

- Croninger & Douglas (2005) and others point out
  - We must consider how and why data are missing
  - Our understanding of why data are missing influences methods we chose to address the issue
- Building on prior work, our work demonstrates that different approaches also have real implications for conclusions and recommendations
- No easy answers, however

Recommendations

- Know thy data
- Explore missingness
- Consider effects of missing data on results
- Address the problem as necessary
- And…if possible, test how your conclusions might vary
- Be aware of how others do or do not address the issue of missing data

Bibliography

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67*(4), 1012–1028.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

Becker, W. E., & Powers, J. R. (2001). Student performance, attrition, and class size given missing student data. *Economics of Education Review, 20,* 377–388.

Bermúdez, J., Corberán-Vallet, A., & Vercher, E. (2009, August). Forecasting time series with missing data using Holt's model. *Journal of Statistical Planning & Inference*, *139*(8), 2791–2799.

Bodner, T. (2008, October 1). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 651–675.

Chen, T., Martin, E., & Montague, G. (2009, August). Robust probabilistic PCA with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis*, *53*(10), 3706–3716.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330–351.

Bibliography (cont.)

Croninger, R. G., & Douglas, K. M. (2005). Missing data and institutional research. *New Directions for Institutional Research, 127*, 33–49.

Croy, C. D., & Douglas N. K. (2005). Methods for addressing missing data in psychiatric and developmental research. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*(12), 1230–1240.

Glas, C., & Pimentel, J. (2008, December). Modeling nonignorable missing data in speeded tests. *Educational & Psychological Measurement*, *68*(6), 907–922.

Paik, M., & Wang, C. (2009, July). Handling missing data by deleting completely observed records. *Journal of Statistical Planning & Inference*, *139*(7), 2341–2350.

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), Real data analysis (pp. 31–78). Greenwich, CT: Information Age Publishing.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525–556.

Bibliography (cont.)

Robitzsch, A., & Rupp, A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and kogistic regression analysis. *Educational and Psychological Measurement, 69*(1), 18–34.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3–15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. P*sychological Methods, 7*, 147–177.

Schafer, J. L., & Olsen, M. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545–571.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*(4), 317–329.

Yuan, K., & Lu, L. (2008). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research*, *43*(4), 621–652.

Zhang, B., & Walker, C. (2008). Impact of missing data on person-model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466–479.

Tables

Table 1
Logistic Regression Results for Full Sample

	Listwise Deletion			Multiple Imputation		
	OR	SE	Sig.	OR	SE	Sig
Men compared to women	1.13	0.04	***	1.08	0.03	***
Compared to Whites						
Native American/Other	1.14	0.28		1.07	0.20	
Asian, Pacific American	1.05	0.15		1.16	0.11	
African American	1.07	0.06		0.88	0.05	**
Hispanic	0.93	0.10		0.91	0.08	
Race Not Reported	1.12	0.16		1.23	0.07	***
Adjusted Gross Income ($1,000s)	1.00	0.00		1.00	0.00	
Age	0.98	0.01	****	1.02	0.00	****
SAT Score	1.00	0.00	****	1.00	0.00	****
High School Rank	1.00	0.00		1.00	0.00	***
Compared to Regular HS Diploma						
Honors	1.35	0.12	**	1.66	0.10	****
Core 40	1.10	0.08		1.26	0.07	****
Not Reported	0.70	0.06	****	1.02	0.03	
GED, Other	2.44	0.85		0.62	0.06	****
Cumulative Grade Point	2.40	0.02	****	1.61	0.01	****
Living on- compared to off-campus	0.83	0.06	***	0.90	0.05	**
Resident compared to non-resident	1.15	0.07	**	1.12	0.05	**
Developmental Education	1.00	0.01		1.00	0.01	
Compared to attending a research university						
Regional campus	0.82	0.08	***	0.59	0.06	****
State university	0.74	0.08	****	0.83	0.06	***
Urban institution	0.57	0.09	****	0.48	0.07	****
Community college	0.48	0.10	****	0.71	0.07	****
Credits Attempted	1.05	0.00	****	1.03	0.00	****
Compared to dependent students						
Independent	1.02	0.06		1.32	0.04	****
Undetermined	1.04	0.09		0.75	0.06	****
Total gants received	1.00	0.00		1.00	0.00	
Total loans received	1.00	0.01		1.00	0.01	
Total private gift aid received	0.98	0.03		0.96	0.02	**
Total work-study received	1.31	0.08	****	1.32	0.07	****
Received aid?	0.97	0.05		1.04	0.04	
N=	23,164			37,039		

****p<0.001, ***p<0.01, **p<0.05, *p<0.10

Table 2
Logistic Regression Results for Small Sample

	Listwise Deletion			Multiple Imputation		
	OR	SE	Sig.	OR	SE	Sig
Men compared to women	1.63	0.17	***	1.65	0.15	****
Compared to Whites						
Native American/Other	-	-		0.02	3.32	
Asian, Pacific American	1.46	0.70		0.88	0.56	
African American	0.96	0.29		0.85	0.24	
Hispanic	1.31	0.52		1.34	0.45	
Race Not Reported	0.75	0.78		1.06	0.56	
Adjusted Gross Income ($1,000s)	1.00	0.00	*	1.00	0.00	
Age	0.98	0.02		1.00	0.02	
SAT Score	1.00	0.00	*	1.00	0.00	***
High School Rank	1.01	0.00	***	1.02	0.00	****
Compared to Regular HS Diploma						
Honors	1.28	0.54		1.49	0.49	
Core 40	1.35	0.39		1.42	0.36	
Not Reported	0.69	0.27		1.08	0.23	
GED, Other	-	-		0.16	0.56	***
Cumulative Grade Point	2.37	0.10	****	1.66	0.07	****
Living on- compared to off-campus	0.89	0.27		0.88	0.25	
Resident compared to non-resident	0.79	0.34		0.85	0.27	
Developmental Education	1.01	0.04		0.98	0.03	
Compared to attending a research university						
Regional campus	0.92	0.35		0.80	0.31	
State university	0.60	0.37		0.84	0.34	
Urban institution	1.17	0.42		0.92	0.37	
Community college	0.38	0.45	**	0.76	0.37	
Credits Attempted	1.02	0.01		1.02	0.01	**
Compared to dependent students						
Independent	0.75	0.43		0.88	0.33	
Undetermined	0.57	0.30	*	0.73	0.28	
Total gants received	1.00	0.00		1.00	0.00	
Total loans received	1.02	0.03		1.01	0.03	
Total private gift aid received	0.76	0.11	**	0.83	0.09	**
Total work-study received	1.57	0.43		1.58	0.40	
Received aid?	0.93	0.26		0.83	0.22	
N=	1,169			1,827		

****p<0.001, ***p<0.01, **p<0.05, *p<0.10

Contact Us

Project on Academic Success

1900 E. Tenth Street
Eigenmann Hall, Suite 630
Bloomington, IN 47406-7512
(812) 855-0707
http://pas.indiana.edu

Jacob P. K. Gross
paugross@indiana.edu

Afet Dadashova
adadasho@indiana.edu

John V. Moore
jmooreii@indiana.edu

Mary Ziskin
mziskin@indiana.edu