APSTA-GE 2013: Missing Data
Spring 2021

Lecture Time:	Friday 9:30 am – 1 pm
Location:	Online
Instructor:	Dobrin Marchev
Email:	dm199@nyu.edu
Office Hours:	By appointment
TA:	TBA

Course Description and Prerequisites:
We will begin by discussing different types of mechanisms that can generate missing data. This will lay the groundwork for discussions of what types of missing data scenarios can be accommodated by each missing data method discussed subsequently. Simple missing data fixes (for example listwise deletion) will be described next as well as the problems they can create in terms of bias and loss of efficiency. We next explore some slightly more complicated fixes (for instance various types of single imputation) and the assumptions required for valid inference for each. The course will end with at least three weeks of focus on multiple imputation including discussions of the general framework, different models and algorithms and the basic theory. More detailed focus will be spent on implementation of the `mice` and `mi` packages in `R`. If time allows the course may finish with a discussion of missing data mechanisms that are not missing at random (NMAR) or more details for the Bayesian methods in multiple imputation.

The prerequisite is at least three semesters of quantitative methods beyond introductory statistics (for instance RESCH-GE.2003 and 2004 or the equivalent as approved by the instructor). It is particularly important that students should be comfortable with the following concepts before the course begins: Binomial probability model, logistic regression, transformations, and use of regression diagnostics to identify lack of model fit. Computer code will be presented in the `R` statistical software language and `R` will be required for all assignments and the project. Students are expected to have a minimum prior experience with `R`.

Course Objectives:
The goal of this course is to provide students with a basic knowledge of the implications of missing data on their data analyses as well as potential solutions. Upon completion of the course the students should be able to solve the missing data problem in their own data analysis research projects.

Assignments. Each week students will perform data analyses that correspond to that week's readings and lecture. These will be performed both on a common dataset as homework, and on students' own data (to be submitted at the end as part of the project). We will discuss the results as a class and all *students will be expected to be able to contribute to this discussion by explaining how they approached parts of the assignment*. The weekly analyses will be used towards a final project that will be turned in at the end of the semester.

Grading. Grading will be based primarily (70%) on one project comprising an amalgamation of all the weekly assignments. Short in-class quizzes will count as 15% of the semester grade. The remaining 15% are based on homework assignments.

Tasks to be completed for final project (details provided in class):

0. Find a suitable dataset with missing observations. Ideally, it should have at least 100 observations, and at least 3-4 variables, both numerical and categorical, of which at least one numerical variable is completely observed. You also need to think what you want to do with the data. That is, what model you want to run, what you want to estimate, and which variable you want to predict by the rest.

After each of the following tasks, you need to implement the analysis you have in mind and report the results/estimates.

1. Listwise deletion.
2. Mean/mode imputation.
3. Random imputation.
4. Dummy variable on predictor variables.
5. LVCF (if applicable to your data).
6. Hotdecking (nearest neighbor).
7. Regression imputation. (Note you might have to use logistic or multinomial models, depending on what type of variable you impute values for.)
8. Regression imputation with noise only on numerical and dichotomous variables.
   Multiple imputation with either `mi` or `mice` package:
9. Load your data into the package. Obtain summary, histogram and image of the data.
10. Check your data types and make changes if necessary.
11. Run the mi command and check convergence by traceplots.
12. Check r-hats.
13. Increase number of imputations if necessary.
14. Plot diagnostics.
15. Change imputation models if necessary.
16. Run pooled analysis.
17. Prepare a table with results from all imputation methods.
18. Discuss and compare.

Academic Integrity:

All students are responsible for understanding and complying with the NYU Steinhardt Statement on academic integrity. The statement is available at https://steinhardt.nyu.edu/statement-academic-integrity

Grading Scale (cutpoints):

≥ 93%:	**A**	The quality of project and homework problem sets is very high: thorough and careful data management, accurate implementations of missing data techniques, and correct interpretations of the results, clear and consistent quantitative writing in final project.
90% to 92%:	**A-**	
87% to 89%:	**B+**	The quality of project and problem sets is satisfactory: decent missing data management, mostly accurate implementations of the methods and interpretations of the results, clear quantitative writing in final project.
83% to 86%:	**B**	
80% to 82%:	**B-**	
77% to 79%:	**C+**	The quality of project and problem sets is barely satisfactory: evidence of missing data management, reasonably accurate implementations of missing data techniques and interpretations of the results, complete final project.
73% to 76%:	**C**	
70% to 72%:	**C-**	
65% to 69%:	**D+**	The quality of project and problem sets is unsatisfactory: lack of evidence of missing data management, inaccurate implementations of missing data methods and interpretations of the results, complete final project.
60% to 64%:	**D**	
< 60%:	**F**	Fail

Students with Disabilities:

Academic accommodations are available for students with disabilities. Please visit the Moses Center for Students with Disabilities (CSD) website at www.nyu.edu/csd and click on the Reasonable Accommodations and How to Register tab or call or e-mail CSD at (212-998-4980 ormosescsd@nyu.edu) for information.

Mental Health & Wellness:

If you are experiencing undue personal and/or academic stress during the semester that may be interfering with your ability to perform academically, The NYU Wellness Exchange (212 443 9999) offers a range of services to assist and support you. I am available to speak with you about stresses related to your work in my course, and I can assist you in connecting with The Wellness Exchange. The Wellness Exchange offers drop-in services on campus on a regular basis. You can find more information at https://www.nyu.edu/students/health-and-wellness/wellness-exchange.html Additionally, if you anticipate any challenges with completing the assignments, readings, exams and other work required in this course, I encourage you to register with The Moses Center (212 998 4980) in advance so that you may be granted the proper academic accommodations.

Reading materials

Required reading materials

- van Buuren, S. (2018) *Flexible Imputation of Missing Data*, 2nd ed., Chapman & Hall Available online: https://stefvanbuuren.name/fimd/

Recommended reading materials

- Venables, W. N. (2009) *An Introduction to R*
- Allison, Paul (2002) *Missing Data*, Sage University Press.
- Little, Roderick J. A., Rubin, Donald B. (2002) *Statistical Analysis with Missing Data*, Wiley-Interscience
- Enders, Craig K. (2010) *Applied Missing Data Analysis*, Guilford Press.
- Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., Verbeke, G. (2014) *Handbook of Missing Data Methodology*, Chapman & Hall/CRC

Other online resources that might be of interest:

A repository of R documentation/tutorials:
http://cran.r-project.org/

Outline of course topics and readings:

The following outline describes the topics that will be covered along with anticipated associated readings. It corresponds roughly to the course weeks though we may end up adjusting time spent on each topic as we go. Readings highlighted with an * are recommended, not required.

Topics and assigned readings:

0) Introduction to R. **Please complete the following on your own *before the first class*.**

Please complete the following tutorials:

Verzani, *simpleR*, p. 94 (installing R, external packages), pp. 1-35, pp. 41-46, 77-89, 94-100 (this document is available at cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf)

https://resources.rstudio.com/

You may also find the resources at DataCamp helpful, e.g.:
https://www.datacamp.com/courses/free-introduction-to-r

1) **Getting comfortable with R.**
Missing Data Mechanisms. How are missing data generated and why should we care?
Complete case analyses.
van Buuren, pp. 3-9, 25-35, 48, 63-65
*Allison, pp. 1-6
*Enders, Chapter 1

2) **Simple missing data fixes: listwise deletion, LVCF, mean imputation, dummy variable**
van Buuren, pp. 8-23
*Allison, pp. 6-11
*Enders, Chapter 2
*Handbook, Chapter 2

3) **More complicated missing data fixes: hotdecking, regression imputation**
van Buuren, Sections 1.3, 3.4,
*Enders, Chapter 2.7, 2.9
*Allison, pp. 11-27
Quiz 1 on 02/12 at 9:30 am

4) **Stochastic regression imputation (regression imputation with noise)**
Conceptual overview of multiple imputation
van Buuren, pp. 25-43, 53-56
*Enders, Chapter 2.8
*Allison, pp. 27-50

5) **Multiple imputation in practice**
Software in R, simple analyses, and diagnostics)
van Buuren, pp 34-51 (Remember though that we will not be using `mice` to perform imputations in `R`)
*Enders, Chapters 7, 8, 9
Quiz 2 on 02/26 at 9:30 am

6) **Multiple imputation in practice**
More complicated models and considerations, more advanced diagnostics
van Buuren, pp 53-82
Abayomi, Gelman, and Levy paper on multiple imputation diagnostics
*Enders, Chapter 7, 8, 9

7) **More advanced Bayesian imputation and other missing data methods**
Little and Rubin, Chapter 10
van Buuren, pp TBD
*Enders, Chapter 6
Quiz 3 on 03/11 at 9:30 am