9th GESIS Summer School in Survey Methodology
Cologne, August 2020

Syllabus for course 5:
"Applied Multiple Imputation"

Lecturers:	Dr. Ferdinand Geißler	Dr. Jan Paul Heisig
E-mail:	ferdinand.geissler@hu-berlin.de	jan.heisig@wzb.eu
Homepage:	https://www.sowi.hu-berlin.de/de/lehrbereiche/empisoz/a-z/geisslerferdinand	https://www.wzb.eu/en/persons/jan-paul-heisig

Date: 10-14 August 2020
Times: 09:30-12:30 + 14:30-16:30 + 16:30-17:30
Time zone: CEST, course starts on Monday at 09:30
Venue: Online via Zoom

About the Lecturers:

Dr. Ferdinand Geißler is a senior lecturer at the Chair of Social Research & Methods at the Humboldt-University Berlin, where he also earned his PhD in Sociology. From 2011 to 2013 he was a research assistant at the National Educational Panel Study (NEPS) where he worked on returns to education and the imputation of income data. His research interests include education, social inequality and quantitative methods. He regularly teaches courses on quantitative methods and has taught several specialized courses on multiple imputation.

Dr. Jan Paul Heisig is a senior researcher in the research unit "Skill Formation and Labor Markets" at WZB Berlin Social Science Center. He holds a PhD in Sociology from Freie Universität Berlin and has been a visitor at Stanford University and the University of Amsterdam. His research interests include education, labor markets, public policy, social inequality, and quantitative methods. He regularly teaches courses on multiple imputation, analysis of multilevel data, and other topics in statistics and data analysis.

Selected Publications:

- C. Aßmann, A. Würbach, S. Goßmann, S., F. Geissler, and A. Bela. 2017. Nonparametric multiple imputation for questionnaires with individual skip patterns and constraints: The case of income imputation in the National Educational Panel Study. Sociological Methods and Research, 46 (4): 864-897.
- F. Geißler. 2018. Bildung, Fähigkeiten und Arbeitsmarkterträge. Wiesbaden: VS Verlag für Sozialwissenschaften.
- W. Ludwig-Mayerhofer, U. Liebeskind, and F. Geißler. 2014. Statistik. Eine Einführung für Sozialwissenschaftler. Weinheim und Basel: Beltz Juventa.
- J.P. Heisig, M. Schaeffer, and J. Giesecke. 2017. The Costs of Simplicity: Why Multilevel Models Need to Account for Cross-cluster Differences in the Effects of Controls. American Sociological Review. 82 (4): 796-827.
- J.P. Heisig, B. Lancee, and J. Radl. 2017. Ethnic Inequality in Retirement Income: A Comparative Analysis of Immigrant-native Gaps in Western Europe. Ageing & Society. Advance Access (May 4, 2017).
- J.P. Heisig and H. Solga. 2015. Secondary Education Systems and the General Skills of Less- and Intermediate-educated Adults: A Comparison of 18 Countries. Sociology of Education. 88 (3): 202-225.

Short Course Description:

Missing data are a pervasive problem in the social sciences. Data for a given unit may be missing entirely, for example, because a sampled respondent refused to participate in a survey (survey nonresponse). Alternatively, information may be missing only for a subset of variables (item nonresponse), for example, because a respondent refused to answer some of the questions in a survey. The traditional way of dealing with item nonresponse, referred to as “complete case analysis” (CCA) or “listwise deletion”, excludes every observation with missing information from the analysis. While easy to implement, complete case analysis is wasteful and can lead to biased estimates. Multiple imputation (MI) seeks to address these issues and provides more efficient and unbiased estimates when certain conditions are met. Therefore, it is increasingly replacing CCA as the method of choice for dealing with item nonresponse in applied quantitative work in the social sciences.

The goals of the course are to introduce participants to the basic concepts and statistical foundations of missing data analysis and MI, and to enable them to use MI in their own work. The course puts heavy emphasis on the practical application of MI and on the complex decisions and challenges that researchers are facing in its course. The focus is on MI using iterated chained equations (aka “fully conditional specification”) and its implementation in the software package Stata. Participants should have a good working knowledge of Stata to follow the applied parts of the course and to successfully master the exercises. Participants who are not familiar with Stata may still benefit from the course, but will likely find the exercises quite challenging.

Keywords:

Missing Data; Item Nonresponse; Multiple Imputation; Complete-case analysis; Iterated Chained Equations

Course Prerequisites:

- Experience in the analysis of quantitative data
- Good knowledge of regression analysis
- Good working knowledge of Stata
- Basic understanding of probability theory and sampling

Target Group:

Participants will find the course useful if:

- use survey or other types of quantitative data and want to learn about MI as an alternative to CCA;
- are already using MI but want to gain a better understanding of the underlying assumptions, of current best practice recommendations, and/or of how to solve specific problems that arise in its application (e.g., imputation diagnostics, convergence problems, imputation of transformed variables such as interactions, imputation of hierarchical data).

Course and Learning Objectives:

By the end of the course participants will:

- understand basic concepts of missing data analysis such as “missing at random”;
- be familiar with different approaches of how to handle item nonresponse and with their advantages and drawbacks;
- have a solid understanding of the main assumptions and statistical theory underlying MI and of the main steps of an analysis involving MI (imputation, diagnostics, and analysis);
- know how to implement MI using chained equations in Stata;
- know how to deal with various (Stata-specific and general) practical complications that arise in the application of MI using chained equations.

Organizational Structure of the Course:

This is a five-day course with a total amount of 30 hours of virtual class time. Each day will begin with a three-hour lecture-like segment introducing the new material (9:30am-12:30pm). Exercises, most of them involving hands-on programming, will be distributed at the end of the lecture segment. Participants can start working on the exercises during the extended lunch break (12:30pm to 2:30pm). The first afternoon segment (2:30pm-4:30pm) will focus on the exercises. Participants will continue to work on the exercises, now with assistance from the lecturers, and eventually answers and solutions will be discussed with the full group. The final “flextime” segment of each day (4:30pm to 5:30pm) will serve to further discuss questions that have come up during the day and for lecturer-participant meetings that focus on individual questions and problems. Participants interested in individual consultations concerning their ongoing projects are encouraged to contact the lecturers before the course and provide a short description of the issues they would like to discuss. The individual-meetings can also be used for questions that arise during the course, however.

Software and Hardware Requirements:

The practical examples and hands-on exercises will be done in Stata. Participants should have a recent version installed on their local computer. Version 15 or later would be ideal, although most examples should work in versions 12 and later. Participants who do not own a copy of Stata will be provided with access to a full Stata license by GESIS for the duration of the course. Stata will be installed and activated prior to the course by GESIS staff through remote access on the participants' machines.

Long Course Description:

Missing data are a pervasive problem in the social sciences. Data for a given unit may be missing entirely, for example, because a sampled respondent refused to participate in a survey (survey nonresponse). Alternatively, information may be missing only for a subset of variables (item nonresponse), for example, because a respondent refused to answer some of the questions in a survey. The traditional way of dealing with item nonresponse, referred to as "complete case analysis" (CCA) or "listwise deletion", excludes every observation with missing information from the analysis. While easy to implement, complete case analysis is wasteful and can lead to biased estimates. Multiple imputation (MI) seeks to address these issues and provides more efficient and unbiased estimates when certain conditions are met. Therefore, it is increasingly replacing CCA as the method of choice for dealing with item nonresponse in applied quantitative work in the social sciences.

The goals of the course are to introduce participants to the basic concepts and statistical foundations of missing data analysis and MI, and to enable them to use MI in their own work. The course has two main parts. In the first, which will mainly take place on the first day, we provide a general introduction to the problem of missing data and introduce key concepts such as "missing data pattern" (e.g., monotone vs. non-monotone) and "missing data mechanism" (missing completely at random, missing at random, not missing at random). We then review traditional approaches of how to deal with item non-response such as listwise deletion/complete case analysis (CCA) and single imputation methods. We discuss the shortcomings of these approaches and then introduce the basic principles of MI to illustrate how it improves upon traditional methods.

The second part of the course (from day 2 onwards) provides a thorough introduction to the method of MI and its implementation in Stata. We will begin with a general introduction to the method, some simple examples, and a basic overview of Stata's tools for conducting MI and for analyzing multiply imputed data. The remainder of the course will then focus on the imputation of multivariate missing data (i.e., data with missing values on more than one variable) using the "iterated chained equations" approach, as implemented in the Stata command mi impute chained. We will focus on "best-practice" recommendations as well as issues and complications that are immediately relevant to applied work. Topics include the specification of imputation models, whether and how to impute terms such as squares or interactions that are deterministic functions of lower-order variables, or how to deal with the problem of perfect prediction. Recommendations will be linked to relevant statistical literature, including evidence from recent simulation studies, but the main emphasis is on the implications for applied work. Further topics that will be covered in greater detail include imputation diagnostics and the treatment of special data types, such as multilevel or complex survey data.

Stata examples will be given throughout the course and the lab sessions will provide plenty of opportunity for participants to improve their MI-related skills. At some points, we will highlight limitations of Stata's capabilities and briefly mention alternative software options. However, we will not actually use any statistical packages other than Stata in the course, and since we have limited experience with them, we may not be able to answer specific questions related to alternative packages. Since the basic structure of MI algorithms tends to be quite similar, participants working with other packages should nevertheless benefit from the course, including the practical elements, although they may find some of the exercises quite challenging.

Day-to-day Schedule and Literature:

Day	Topic(s)
1	**Fundamental concepts of missing data analysis** • Survey vs. item nonresponse • Missing data mechanisms • Missing data patterns **Traditional approaches to dealing with item nonresponse** • Complete-case analysis, mean imputation, hotdeck, single regression-based imputation • Why the alternatives fall short • The basic idea of MI
	Suggested reading: • Enders, C. 2010. Applied Missing Data Analysis. New York: Guilford Press. Chapters 1-2. • Van Buuren, S. 2018. Flexible Imputation of Missing Data. Boca Raton: CRC Press. Chapter 1.
2	**The Fundamentals of Multiple Imputation** • The basic idea and statistical theory behind MI • Workflow of an analysis using MI • Rubin's rules • When MI is appropriate – and when it is not • How many imputations do I need? **Multiple Imputation of missing values in Stata: First Steps** • Imputation of a single variable with missing values (univariate missingness) • Imputation of multiple variables with missing values (multivariate missingness) • Iterated Chained Equations (ICE) versus Joint Modelling (JM) • Imputation methods (linear regression, predictive mean matching, generalized linear models) and when to use them
	Suggested reading: • Enders, C. 2010. Applied Missing Data Analysis. New York: Guilford Press. Chapters 6-7. • Van Buuren, S. 2018. Flexible Imputation of Missing Data. Boca Raton: CRC Press. Chapters 2-4. • Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" International Journal of Methods in Psychiatric Research 20(1):40–49. • StataCorp. 2019. Stata Multiple-Imputation Reference Manual. Release 14. College Station: Stata Press.
3	**What is a good imputation model?** • Congeniality of imputation and analysis models • Should I impute the dependent variable? • How should I deal with terms such as squares and interactions? • Conditional imputation **Practical issues** • Convergence problems and perfect prediction
	Suggested reading: • Enders, C. 2010. Applied Missing Data Analysis. New York: Guilford Press. Chapters 8-9. • Van Buuren, S. 2018. Flexible Imputation of Missing Data. Boca Raton: CRC Press. Chapters 4-6. • Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" International Journal of Methods in Psychiatric Research 20(1):40–49. • StataCorp. 2019. Stata Multiple-Imputation Reference Manual. Release 14. College Station: Stata Press.
4	**Imputation Diagnostics** • Monitoring convergence of the ICE algorithm • How can I identify problems with the imputed data? • How can I solve such problems?

	Analysis of MI data • Rubin's rules and their application to quantities other than regression coefficients • MI degrees of freedom • Marginal effects • Testing complex hypotheses/multi-parameter tests (optional/if time permits)
	Suggested reading: • Enders, C. 2010. Applied Missing Data Analysis. New York: Guilford Press. Chapters 6-7. • Van Buuren, S. 2018. Flexible Imputation of Missing Data. Boca Raton: CRC Press. Chapters 5-6. • Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" International Journal of Methods in Psychiatric Research 20(1):40–49. • StataCorp. 2019. Stata Multiple-Imputation Reference Manual. Release 14. College Station: Stata Press.
5	**Dealing with complex data structures** • MI of hierarchical/multilevel/panel data • Complex surveys and weighting (optional/if time permits) **Open questions and feedback** • Time to discuss anything
	Suggested reading: • Van Buuren, S. 2018. Flexible Imputation of Missing Data. Boca Raton: CRC Press. Chapters 7, 11.

Preparatory Reading:

None.

Additional Recommended Literature:

Recommended books

- Enders, C. (2010). Applied Missing Data Analysis. New York: Guilford Press.
- van Buuren, S. (2018). Flexible Imputation of Missing Data. Boca Raton: CRC Press.

Papers on practical issues

- Abayomi, K., Gelman, A. and Levy, M. (2008). Diagnostics for multivariate imputations. Journal of the Royal Statistical Society: Series C (Applied Statistics), 57: 273–291
- Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" International Journal of Methods in Psychiatric Research 20(1):40–49.
- Seaman, S. R., Bartlett, J. W. and White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC medical research methodology, 12, 46.
- van Buuren, S. (2011). Multiple Imputation of Multilevel Data, in Hox, J. J., Roberts, J. K. (Eds.) Handbook of Advanced Multilevel Analysis, New York: Routledge, pp. 173–196.
- Vink, G., Frank, L. E., Pannekoek, J., van Buuren, S. (2014). Predictive Mean Matching Imputation of Semicontinuous Variables. Statistica Neerlandica, 68: 61–90.
- Vink, G. and van Buuren, S. (2013). Multiple Imputation of Squared Terms. Sociological Methods & Research, 42: 598–607
- von Hippel, P. T. (2007). Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. Sociological Methodology, 37 (1): 83–117.
- von Hippel, P. T. (2009). How To Impute Squares, Interactions, and Other Transformed Variables. Sociological Methodology, 39 (1): 265–291.
- White I. R., Royston P., Wood A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in Medicine, 30(4): 377–399.
- Young, R. and Johnson, D. R. (2015). Handling missing values in longitudinal panel data with multiple imputation. Journal of Marriage and Family, 77, 277-294.

MI in Stata

- Eddings, W. and Marchenko, Y. (2012). Diagnostics for Multiple Imputation in Stata. Stata Journal, 12: 353–367.
- StataCorp. (2019). Stata Multiple-Imputation Reference Manual. Release 14. College Station: Stata Press.