Close

Web of Science
Page 1 (Records 1 -- 50)
[1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9]

Print

MI for big data
Non parametric
IMI

Record 1 of 444

Title: Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model

Author(s): Bartlett, JW (Bartlett, Jonathan W.); Seaman, SR (Seaman, Shaun R.); White, IR (White, Ian R.); Carpenter, JR (Carpenter, James R.)

Group Author(s): Alzheimer's Dis Neuroimaging

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 24 **Issue:** 4 **Special Issue:** SI **Pages:** 462-487 **DOI:** 10.1177/0962280214521348 **Published:** AUG 2015

Abstract: Missing covariate data commonly occur in epidemiological and clinical research, and are often dealt with using multiple imputation. Imputation of partially observed covariates is complicated if the substantive model is non-linear (e.g. Cox proportional hazards model), or contains non-linear (e.g. squared) or interaction terms, and standard software implementations of multiple imputation may impute covariates from models that are incompatible with such substantive models. We show how imputation by fully conditional specification, a popular approach for performing multiple imputation, can be modified so that covariates are imputed from models which are compatible with the substantive model. We investigate through simulation the performance of this proposal, and compare it with existing approaches. Simulation results suggest our proposal gives consistent estimates for a range of common substantive models, including models which contain non-linear covariate effects or interactions, provided data are missing at random and the assumed imputation models are correctly specified and mutually compatible. Stata software implementing the approach is freely available.

Accession Number: WOS:000358452800006

PubMed ID: 24525487

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Seaman, Shaun		0000-0003-3726-5937
Bartlett, Jonathan		0000-0001-7117-0195
Carpenter, James		0000-0003-3890-6206

ISSN: 0962-2802

eISSN: 1477-0334

Record 2 of 444

Title: missMDA: A Package for Handling Missing Values in Multivariate Data Analysis

Author(s): Josse, J (Josse, Julie); Husson, F (Husson, Francois)

Source: JOURNAL OF STATISTICAL SOFTWARE **Volume:** 70 **Issue:** 1 **Published:** APR 2016

Abstract: We present the R package missMDA which performs principal component methods on incomplete data sets, aiming to obtain scores, loadings and graphical representations despite missing values. Package methods include principal component analysis for continuous variables, multiple correspondence analysis for categorical variables, factorial analysis on mixed data for both continuous and categorical variables, and multiple factor analysis for multi-table data. Furthermore, missMDA can be used to perform single imputation to complete data involving continuous, categorical and mixed variables. A multiple imputation method is also available. In the principal component analysis framework, variability across different imputations is represented by confidence areas around the row and column positions on the graphical outputs. This allows assessment of the credibility of results obtained from incomplete data sets.

Accession Number: WOS:000373921300001

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
josse, julie		0000-0001-9547-891X

ISSN: 1548-7660

Record 3 of 444

Title: zCompositions - R Package for multivariate imputation of left-censored data under a compositional approach

Author(s): Palarea-Albaladejo, J (Palarea-Albaladejo, Javier); Martin-Fernandez, JA (Antoni Martin-Fernandez, Josep)

Source: CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS **Volume:** 143 **Pages:** 85-96 **DOI:** 10.1016/j.chemolab.2015.02.019 **Published:** APR 15 2015

Abstract: zCompositions is an R package for the imputation of left-censored data under a compositional approach. It is pertinent when the analyst assumes that the relevant information is contained on the relative variation structure of the data. For instance, in cases where the experimental data are simultaneously measured in amounts related to a same total weight or volume. The approach is used in fields like geochemistry of waters or sedimentary rocks, environmental studies related to air pollution, physicochemical analysis of glass fragments in forensic science, and among many others. In these fields, rounded zeros and nondetects are usually regarded as left-censored data that hamper any subsequent data analysis. The implemented methods consider aspects of relevance for a compositional approach such as scale invariance, subcompositional coherence or preserving the multivariate relative structure of the data. Based on solid statistical frameworks, it comprises the ability to deal with single and varying censoring thresholds, consistent treatment of closed and non-closed data, exploratory tools, multiple imputation, MCMC, robust and non-parametric alternatives, and recent proposals for count data. Key methodological aspects, new contributions, computational implementation and the practical application of the approach are discussed. (C) 2015 Elsevier B.V. All rights reserved.

Accession Number: WOS:000353730300009

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Martin-Fernandez, Josep Antoni	B-9208-2011	0000-0003-2366-1592
Palarea-Albaladejo, Javier	J-5591-2013	0000-0003-0162-669X

ISSN: 0169-7439
eISSN: 1873-3239

Record 4 of 444

Title: A comparison of two methods of estimating propensity scores after multiple imputation

Author(s): Mitra, R (Mitra, Robin); Reiter, JP (Reiter, Jerome P.)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 25 **Issue:** 1 **Pages:** 188-204 **DOI:** 10.1177/0962280212445945 **Published:** FEB 2016

Abstract: In many observational studies, analysts estimate treatment effects using propensity scores, e.g. by matching or sub-classifying on the scores. When some values of the covariates are missing, analysts can use multiple imputation to fill in the missing data, estimate propensity scores based on the m completed datasets, and use the propensity scores to estimate treatment effects. We compare two approaches to implement this process. In the first, the analyst estimates the treatment effect using propensity score matching within each completed data set, and averages the m treatment effect estimates. In the second approach, the analyst averages the m propensity scores for each record across the completed datasets, and performs propensity score matching with these averaged scores to estimate the treatment effect. We compare properties of both methods via simulation studies using artificial and real data. The simulations suggest that the second method has greater potential to produce substantial bias reductions than the first, particularly when the missing values are predictive of treatment assignment.

Accession Number: WOS:000370685000011
PubMed ID: 22687877
ISSN: 0962-2802
eISSN: 1477-0334

Record 5 of 444

Title: Imputation with the R Package VIM

Author(s): Kowarik, A (Kowarik, Alexander); Templ, M (Templ, Matthias)

Source: JOURNAL OF STATISTICAL SOFTWARE **Volume:** 74 **Issue:** 7 **DOI:** 10.18637/jss.v074.i07 **Published:** OCT 2016

Abstract: The package VIM (Templ, Alfons, Kowarik, and Prantner 2016) is developed to explore and analyze the structure of missing values in data using visualization methods, to impute these missing values with the built-in imputation methods and to verify the imputation process using visualization tools, as well as to produce high-quality graphics for publications.
This article focuses on the different imputation techniques available in the package. Four different imputation methods are currently implemented in VIM, namely hot-deck imputation, k-nearest neighbor imputation, regression imputation and iterative robust model-based imputation (Templ, Kowarik, and Filzmoser 2011). All of these methods are implemented in a flexible manner with many options for customization. Furthermore in this article practical examples are provided to highlight the use of the implemented methods on real-world applications.
In addition, the graphical user interface of VIM has been re-implemented from scratch resulting in the package VIMGUI (Schopfhauser, Templ, Alfons, Kowarik, and Prantner 2016) to enable users without extensive R skills to access these imputation and visualization methods.

Accession Number: WOS:000392513900001
Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Templ, Matthias		0000-0002-8638-5276

ISSN: 1548-7660

Record 6 of 444

Title: Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE

Author(s): Jolani, S (Jolani, Shahab); Debray, TPA (Debray, Thomas P. A.); Koffijberg, H (Koffijberg, Hendrik); van Buuren, S (van Buuren, Stef); Moons, KGM (Moons, Karel G. M.)

Source: STATISTICS IN MEDICINE **Volume:** 34 **Issue:** 11 **Pages:** 1841-1863 **DOI:** 10.1002/sim.6451 **Published:** MAY 20 2015

Abstract: Individual participant data meta-analyses (IPD-MA) are increasingly used for developing and validating multivariable (diagnostic or prognostic) risk prediction models. Unfortunately, some predictors or even outcomes may not have been measured in each study and are thus systematically missing in some individual studies of the IPD-MA. As a consequence, it is no longer possible to evaluate between-study heterogeneity and to estimate study-specific predictor effects, or to include all individual studies, which severely hampers the development and validation of prediction models.Here, we describe a novel approach for imputing systematically missing data and adopt a generalized linear mixed model to allow for between-study heterogeneity. This approach can be viewed as an extension of Resche-Rigon's method (Stat Med 2013), relaxing their assumptions regarding variance components and allowing imputation of linear and nonlinear predictors.We illustrate our approach using a case study with IPD-MA of 13 studies to develop and validate a diagnostic prediction model for the presence of deep venous thrombosis. We compare the results after applying four methods for dealing with systematically missing predictors in one or more individual studies: complete case analysis where studies with systematically missing predictors are removed, traditional multiple imputation ignoring heterogeneity across studies, stratified multiple imputation accounting for heterogeneity in predictor prevalence, and multilevel multiple imputation (MLMI) fully accounting for between-study heterogeneity.We conclude that MLMI may substantially improve the estimation of between-study heterogeneity parameters and allow for imputation of systematically missing predictors in IPD-MA aimed at the development and validation of prediction models.
Copyright (c) 2015John Wiley & Sons, Ltd.

Accession Number: WOS:000352633200004
PubMed ID: 25663182
Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Debray, Thomas	J-7413-2012	0000-0002-1790-2719
van Buuren, Stef		0000-0003-1098-2119

ISSN: 0277-6715
eISSN: 1097-0258

Record 7 of 444

Title: Exploring Diallelic Genetic Markers: The HardyWeinberg Package

Author(s): Graffelman, J (Graffelman, Jan)

Source: JOURNAL OF STATISTICAL SOFTWARE **Volume:** 64 **Issue:** 3 **Pages:** 1-23 **Published:** FEB 2015

Abstract: Testing genetic markers for Hardy-Weinberg equilibrium is an important issue in genetic association studies. The HardyWeinberg package offers the classical tests for equilibrium, functions for power computation and for the simulation of marker data under equilibrium and disequilibrium. Functions for testing equilibrium in the presence of missing data by using multiple imputation are provided. The package also supplies various graphical tools such as ternary plots with acceptance regions, log-ratio plots and Q-Q plots for exploring the equilibrium status of a large set of diallelic markers. Classical tests for equilibrium and graphical representations for diallelic marker data are reviewed. Several data sets illustrate the use of the package.

Accession Number: WOS:000352911000001

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Graffelman, Jan	L-8056-2014	0000-0003-3900-0780

ISSN: 1548-7660

Record 8 of 444

Title: imputeTS: Time Series Missing Value Imputation in R

Author(s): Moritz, S (Moritz, Steffen); Bartz-Beielstein, T (Bartz-Beielstein, Thomas)

Source: R JOURNAL **Volume:** 9 **Issue:** 1 **Pages:** 207-218 **Published:** JUN 2017

Abstract: The imputeTS package specializes on univariate time series imputation. It offers multiple state-of-the-art imputation algorithm implementations along with plotting functions for time series missing data statistics. While imputation in general is a well-known problem and widely covered by R packages, finding packages able to fill missing values in univariate time series is more complicated. The reason for this lies in the fact, that most imputation algorithms rely on inter-attribute correlations, while univariate time series imputation instead needs to employ time dependencies. This paper provides an introduction to the imputeTS package and its provided algorithms and tools. Furthermore, it gives a short overview about univariate time series imputation in R.

Accession Number: WOS:000404756200014

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Moritz, Steffen	U-8455-2018	0000-0002-0085-1804

ISSN: 2073-4859

Record 9 of 444

Title: Multiple imputation of covariates by substantive-model compatible fully conditional specification

Author(s): Bartlett, JW (Bartlett, Jonathan W.); Morris, TP (Morris, Tim P.)

Source: STATA JOURNAL **Volume:** 15 **Issue:** 2 **Pages:** 437-456 **DOI:** 10.1177/1536867X1501500206 **Published:** 2015

Abstract: Multiple imputation is a practical, principled approach to handling missing data. When used to impute missing values in covariates of regression models, imputation models may be misspecified if they are not compatible with the substantive model of interest for the outcome. In this article, we introduce the smcf cs command, which imputes covariates by substantive-model compatible fully conditional specification. This modifies the popular fully conditional specification or chained-equations approach to multiple imputation by imputing each covariate compatibly with a user-specified substantive model. We compare the smcf cs command with standard fully conditional specification imputation using mi impute chained in a simulation study and illustrative analysis of data from a study investigating time to tumor recurrence in breast cancer.

Accession Number: WOS:000357139500006

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Bartlett, Jonathan		0000-0001-7117-0195
Morris, Tim		0000-0001-5850-3610

ISSN: 1536-867X

Record 10 of 444

Title: A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables

Author(s): Wu, W (Wu, Wei); Jia, F (Jia, Fan); Enders, C (Enders, Craig)

Source: MULTIVARIATE BEHAVIORAL RESEARCH **Volume:** 50 **Issue:** 5 **Pages:** 484-503 **DOI:** 10.1080/00273171.2015.1022644 **Published:** SEP 3 2015

Abstract: This article compares a variety of imputation strategies for ordinal missing data on Likert scale variables (number of categories = 2, 3, 5, or 7) in recovering reliability coefficients, mean scale scores, and regression coefficients of predicting one scale score from another. The examined strategies include imputing using normal data models with naive rounding/without rounding, using latent variable models, and using categorical data models such as discriminant analysis and binary logistic regression (for dichotomous data only), multinomial and proportional odds logistic regression (for polytomous data only). The result suggests that both the normal model approach without rounding and the latent variable model approach perform well for either dichotomous or polytomous data regardless of sample size, missing data proportion, and asymmetry of item distributions. The discriminant analysis approach also performs well for dichotomous data. Naively rounding normal imputations or using logistic regression models to impute ordinal data are not recommended as they can potentially lead to substantial bias in all or some of the parameters.

Accession Number: WOS:000362723800002

PubMed ID: 26610248

ISSN: 0027-3171

eISSN: 1532-7906

Record 11 of 444

Title: Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates

Author(s): Quartagno, M (Quartagno, M.); Carpenter, JR (Carpenter, J. R.)

Source: STATISTICS IN MEDICINE **Volume:** 35 **Issue:** 17 **Pages:** 2938-2954 **DOI:** 10.1002/sim.6837 **Published:** JUL 30 2016

Abstract: Recently, multiple imputation has been proposed as a tool for individual patient data meta-analysis with sporadically missing observations, and it has been suggested that within-study imputation is usually preferable. However, such within study imputation cannot handle variables that are completely missing within studies. Further, if some of the contributing studies are relatively small, it may be appropriate to share information across studies when imputing. In this paper, we develop and evaluate a joint modelling approach to multiple imputation of individual patient data in meta-analysis, with an across-study probability distribution for the study specific covariance matrices. This retains the flexibility to allow for between-study heterogeneity when imputing while allowing (i) sharing information on the covariance matrix across studies when this is appropriate, and (ii) imputing variables that are wholly missing from studies. Simulation results show both equivalent performance to the within-study imputation approach where this is valid, and good results in more general, practically relevant, scenarios with studies of very different sizes, non-negligible between-study heterogeneity and wholly missing variables. We illustrate our approach using data from an individual patient data meta-analysis of hypertension trials. (c) 2015 The Authors. Statistics in Medicine Published by John Wiley & Sons Ltd.

Accession Number: WOS:000379983000009

PubMed ID: 26681666

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Carpenter, James		0000-0003-3890-6206

ISSN: 0277-6715

eISSN: 1097-0258

Record 12 of 444

Title: Multiple imputation in the presence of high-dimensional data

Author(s): Zhao, YZ (Zhao, Yize); Long, Q (Long, Qi)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 25 **Issue:** 5 **Pages:** 2021-2035 **DOI:** 10.1177/0962280213511027 **Published:** OCT 2016

Abstract: Missing data are frequently encountered in biomedical, epidemiologic and social research. It is well known that a naive analysis without adequate handling of missing data may lead to bias and/or loss of efficiency. Partly due to its ease of use, multiple imputation has become increasingly popular in practice for handling missing data. However, it is unclear what is the best strategy to conduct multiple imputation in the presence of high-dimensional data. To answer this question, we investigate several approaches of using regularized regression and Bayesian lasso regression to impute missing values in the presence of high-dimensional data. We compare the performance of these methods through numerical studies, in which we also evaluate the impact of the dimension of the data, the size of the true active set for imputation, and the strength of correlation. Our numerical studies show that in the presence of high-dimensional data the standard multiple imputation approach performs poorly and the imputation approach using Bayesian lasso regression achieves, in most cases, better performance than the other imputation methods including the standard imputation approach using the correctly specified imputation model. Our results suggest that Bayesian lasso regression and its extensions are better suited for multiple imputation in the presence of high-dimensional data than the other regression methods.

Accession Number: WOS:000385555400017

PubMed ID: 24275026

ISSN: 0962-2802

eISSN: 1477-0334

Record 13 of 444

Title: Comparison of imputation variance estimators

Author(s): Hughes, RA (Hughes, R. A.); Sterne, JAC (Sterne, J. A. C.); Tilling, K (Tilling, K.)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 25 **Issue:** 6 **Pages:** 2541-2557 **DOI:** 10.1177/0962280214526216 **Published:** DEC 2016

Abstract: Appropriate imputation inference requires both an unbiased imputation estimator and an unbiased variance estimator. The commonly used variance estimator, proposed by Rubin, can be biased when the imputation and analysis models are misspecified and/or incompatible. Robins and Wang proposed an alternative approach, which allows for such misspecification and incompatibility, but it is considerably more complex. It is unknown whether in practice Robins and Wang's multiple imputation procedure is an improvement over Rubin's multiple imputation. We conducted a critical review of these two multiple imputation approaches, a re-sampling method called full mechanism bootstrapping and our modified Rubin's multiple imputation procedure via simulations and an application to data. We explored four common scenarios of misspecification and incompatibility. In general, for a moderate sample size (n = 1000), Robins and Wang's multiple imputation produced the narrowest confidence intervals, with acceptable coverage. For a small sample size (n = 100) Rubin's multiple imputation, overall, outperformed the other methods. Full mechanism bootstrapping was inefficient relative to the other methods and required modelling of the missing data mechanism under the missing at random assumption. Our proposed modification showed an improvement over Rubin's multiple imputation in the presence of misspecification. Overall, Rubin's multiple imputation variance estimator can fail in the presence of incompatibility and/or misspecification. For unavoidable incompatibility and/or misspecification, Robins and Wang's multiple imputation could provide more robust inferences.

Accession Number: WOS:000388625700010

PubMed ID: 24682265

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
	K-7215-2019	
Hughes, Rachael		0000-0003-0766-1410
Tilling, Kate		0000-0002-1010-8926
Sterne, Jonathan		0000-0001-8496-6053

ISSN: 0962-2802
eISSN: 1477-0334

Record 14 of 444

Title: A principal component method to impute missing values for mixed data

Author(s): Audigier, V (Audigier, Vincent); Husson, F (Husson, Francois); Josse, J (Josse, Julie)

Source: ADVANCES IN DATA ANALYSIS AND CLASSIFICATION **Volume:** 10 **Issue:** 1 **Pages:** 5-26 **DOI:** 10.1007/s11634-014-0195-1 **Published:** MAR 2016

Abstract: We propose a new method to impute missing values in mixed data sets. It is based on a principal component method, the factorial analysis for mixed data, which balances the influence of all the variables that are continuous and categorical in the construction of the principal components. Because the imputation uses the principal axes and components, the prediction of the missing values is based on the similarity between individuals and on the relationships between variables. The properties of the method are illustrated via simulations and the quality of the imputation is assessed using real data sets. The method is compared to a recent method (Stekhoven and Buhlmann Bioinformatics 28:113-118, 2011) based on random forest and shows better performance especially for the imputation of categorical variables and situations with highly linear relationships between continuous variables.

Accession Number: WOS:000371235100002

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
josse, julie		0000-0001-9547-891X

ISSN: 1862-5347
eISSN: 1862-5355

Record 15 of 444

Title: Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach

Author(s): Erler, NS (Erler, Nicole S.); Rizopoulos, D (Rizopoulos, Dimitris); van Rosmalen, J (van Rosmalen, Joost); Jaddoe, VWV (Jaddoe, Vincent W. V.); Franco, OH (Franco, Oscar H.); Lesaffre, EMEH (Lesaffre, Emmanuel M. E. H.)

Source: STATISTICS IN MEDICINE **Volume:** 35 **Issue:** 17 **Pages:** 2955-2974 **DOI:** 10.1002/sim.6944 **Published:** JUL 30 2016

Abstract: Incomplete data are generally a challenge to the analysis of most large studies. The current gold standard to account for missing data is multiple imputation, and more specifically multiple imputation with chained equations (MICE). Numerous studies have been conducted to illustrate the performance of MICE for missing covariate data. The results show that the method works well in various situations. However, less is known about its performance in more complex models, specifically when the outcome is multivariate as in longitudinal studies. In current practice, the multivariate nature of the longitudinal outcome is often neglected in the imputation procedure, or only the baseline outcome is used to impute missing covariates. In this work, we evaluate the performance of MICE using different strategies to include a longitudinal outcome into the imputation models and compare it with a fully Bayesian approach that jointly imputes missing values and estimates the parameters of the longitudinal model. Results from simulation and a real data example show that MICE requires the analyst to correctly specify which components of the longitudinal process need to be included in the imputation models in order to obtain unbiased results. The full Bayesian approach, on the other hand, does not require the analyst to explicitly specify how the longitudinal outcome enters the imputation models. It performed well under different scenarios. Copyright (c) 2016 John Wiley & Sons, Ltd.

Accession Number: WOS:000379983000010
PubMed ID: 27042954

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Erler, Nicole		0000-0002-9370-6832
Franco, Oscar		0000-0002-4606-4929

ISSN: 0277-6715
eISSN: 1097-0258

Record 16 of 444

Title: Cuts in Bayesian graphical models

Author(s): Plummer, M (Plummer, Martyn)

Source: STATISTICS AND COMPUTING **Volume:** 25 **Issue:** 1 **Special Issue:** SI **Pages:** 37-43 **DOI:** 10.1007/s11222-014-9503-z **Published:** JAN 2015

Abstract: The cut function defined by the OpenBUGS software is described as a "valve" that prevents feedback in Bayesian graphical models. It is shown that the MCMC algorithm applied by OpenBUGS in the presence of a cut function does not converge to a well-defined limiting distribution. However, it may be improved by using tempered transitions. The cut algorithm is compared with multiple imputation as a gold standard in a simple example.

Accession Number: WOS:000349028500007
Conference Title: Joint IMS-ISBA Meeting (MCMSki 4)
Conference Date: JAN 06-08, 2014
Conference Location: Chamonix Mont Blanc, FRANCE

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Plummer, Martyn		0000-0001-5130-6497

ISSN: 0960-3174
eISSN: 1573-1375

Record 17 of 444

Title: A comparison of incomplete-data methods for categorical data

Author(s): van der Palm, DW (van der Palm, Daniel W.); van der Ark, LA (van der Ark, L. Andries); Vermunt, JK (Vermunt, Jeroen K.)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 25 **Issue:** 2 **Pages:** 754-774 **DOI:** 10.1177/0962280212465502 **Published:** APR 2016

Abstract: We studied four methods for handling incomplete categorical data in statistical modeling: (1) maximum likelihood estimation of the statistical model with incomplete data, (2) multiple imputation using a loglinear model, (3) multiple imputation using a latent class model, (4) and multivariate imputation by chained equations. Each method has advantages and disadvantages, and it is unknown which method should be recommended to practitioners. We reviewed the merits of each method and investigated their effect on the bias and stability of parameter estimates and bias of the standard errors. We found that multiple imputation using a latent class model with many latent classes was the most promising method for handling incomplete categorical data, especially when the number of variables used in the imputation model is large.

Accession Number: WOS:000374792800017

PubMed ID: 23166159

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
vermunt, jeroen	K-3680-2012	0000-0001-9053-9330

ISSN: 0962-2802

eISSN: 1477-0334

Record 18 of 444

Title: Using Principal Components as Auxiliary Variables in Missing Data Estimation

Author(s): Howard, WJ (Howard, Waylon J.); Rhemtulla, M (Rhemtulla, Mijke); Little, TD (Little, Todd D.)

Source: MULTIVARIATE BEHAVIORAL RESEARCH **Volume:** 50 **Issue:** 3 **Pages:** 285-299 **DOI:** 10.1080/00273171.2014.999267 **Published:** MAY 4 2015

Abstract: To deal with missing data that arise due to participant nonresponse or attrition, methodologists have recommended an inclusive strategy where a large set of auxiliary variables are used to inform the missing data process. In practice, the set of possible auxiliary variables is often too large. We propose using principal components analysis (PCA) to reduce the number of possible auxiliary variables to a manageable number. A series of Monte Carlo simulations compared the performance of the inclusive strategy with eight auxiliary variables (inclusive approach) to the PCA strategy using just one principal component derived from the eight original variables (PCA approach). We examined the influence of four independent variables: magnitude of correlations, rate of missing data, missing data mechanism, and sample size on parameter bias, root mean squared error, and confidence interval coverage. Results indicate that the PCA approach results in unbiased parameter estimates and potentially more accuracy than the inclusive approach. We conclude that using the PCA strategy to reduce the number of auxiliary variables is an effective and practical way to reap the benefits of the inclusive strategy in the presence of many possible auxiliary variables.

Accession Number: WOS:000356501600002

PubMed ID: 26610030

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Howard, Waylon	C-1176-2015	0000-0002-0355-2244

ISSN: 0027-3171

eISSN: 1532-7906

Record 19 of 444

Title: Systematic handling of missing data in complex study designs - experiences from the Health 2000 and 2011 Surveys

Author(s): Harkanen, T (Harkanen, Tommi); Karvanen, J (Karvanen, Juha); Tolonen, H (Tolonen, Hanna); Lehtonen, R (Lehtonen, Risto); Djerf, K (Djerf, Kari); Juntunen, T (Juntunen, Teppo); Koskinen, S (Koskinen, Seppo)

Source: JOURNAL OF APPLIED STATISTICS **Volume:** 43 **Issue:** 15 **Pages:** 2772-2790 **DOI:** 10.1080/02664763.2016.1144725 **Published:** DEC 2016

Abstract: We present a systematic approach to the practical and comprehensive handling of missing data motivated by our experiences of analyzing longitudinal survey data. We consider the Health 2000 and 2011 Surveys (BRIF8901) where increased non-response and non-participation from 2000 to 2011 was a major issue. The model assumptions involved in the complex sampling design, repeated measurements design, non-participation mechanisms and associations are presented graphically using methodology previously defined as a causal model with design, i.e. a functional causal model extended with the study design. This tool forces the statistician to make the study design and the missing-data mechanism explicit. Using the systematic approach, the sampling probabilities and the participation probabilities can be considered separately. This is beneficial when the performance of missing-data methods are to be compared. Using data from Health 2000 and 2011 Surveys and from national registries, it was found that multiple imputation removed almost all differences between full sample and estimated prevalences. The inverse probability weighting removed more than half and the doubly robust method 60% of the differences. These findings are encouraging since decreasing participation rates are a major problem in population surveys worldwide.

Accession Number: WOS:000384263000006

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Karvanen, Juha		0000-0001-5530-769X

ISSN: 0266-4763

eISSN: 1360-0532

Record 20 of 444

Title: Visually Exploring Missing Values in Multivariable Data Using a Graphical User Interface

Author(s): Cheng, XY (Cheng, Xiaoyue); Cook, D (Cook, Dianne); Hofmann, H (Hofmann, Heike)

Source: JOURNAL OF STATISTICAL SOFTWARE **Volume:** 68 **Issue:** 6 **Pages:** 1-23 **DOI:** 10.18637/jss.v068.i06 **Published:** DEC 2015

Abstract: Missing values are common in data, and usually require attention in order to conduct the statistical analysis. One of the first steps is to explore the structure of the missing values, and how missingness relates to the other collected variables. This article describes an R package, that provides a graphical

user interface (GUI) designed to help explore the missing data structure and to examine the results of different imputation methods. The GUI provides numerical and graphical summaries conditional on missingness, and includes imputations using fixed values, multiple imputations and nearest neighbors.

Accession Number: WOS:000384909500001

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Cook, Dianne		0000-0002-3813-7155

ISSN: 1548-7660

Record 21 of 444

Title: Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ

Author(s): Burke, DL (Burke, Danielle L.); Ensor, J (Ensor, Joie); Riley, RD (Riley, Richard D.)

Source: STATISTICS IN MEDICINE **Volume:** 36 **Issue:** 5 **Pages:** 855-875 **DOI:** 10.1002/sim.7141 **Published:** FEB 2017

Abstract: Meta-analysis using individual participant data (IPD) obtains and synthesises the raw, participant-level data from a set of relevant studies. The IPD approach is becoming an increasingly popular tool as an alternative to traditional aggregate data meta-analysis, especially as it avoids reliance on published results and provides an opportunity to investigate individual-level interactions, such as treatment-effect modifiers. There are two statistical approaches for conducting an IPD meta-analysis: one-stage and two-stage. The one-stage approach analyses the IPD from all studies simultaneously, for example, in a hierarchical regression model with random effects. The two-stage approach derives aggregate data (such as effect estimates) in each study separately and then combines these in a traditional meta-analysis model. There have been numerous comparisons of the one-stage and two-stage approaches via theoretical consideration, simulation and empirical examples, yet there remains confusion regarding when each approach should be adopted, and indeed why they may differ.
In this tutorial paper, we outline the key statistical methods for one-stage and two-stage IPD meta-analyses, and provide 10 key reasons why they may produce different summary results. We explain that most differences arise because of different modelling assumptions, rather than the choice of one-stage or two-stage itself. We illustrate the concepts with recently published IPD meta-analyses, summarise key statistical software and provide recommendations for future IPD meta-analyses. (C) 2016 The Authors. Statistics in Medicine published by John Wiley & Sons Ltd.

Accession Number: WOS:000393303200010

PubMed ID: 27747915

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Riley, Richard		0000-0001-8699-0735
Burke, Danielle		0000-0003-2803-1151
Ensor, Joie		0000-0001-7481-0282

ISSN: 0277-6715

eISSN: 1097-0258

Record 22 of 444

Title: Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis

Author(s): Siddique, J (Siddique, Juned); Reiter, JP (Reiter, Jerome P.); Brincks, A (Brincks, Ahnalee); Gibbons, RD (Gibbons, Robert D.); Crespi, CM (Crespi, Catherine M.); Brown, CH (Brown, C. Hendricks)

Source: STATISTICS IN MEDICINE **Volume:** 34 **Issue:** 26 **Pages:** 3399-3414 **DOI:** 10.1002/sim.6562 **Published:** NOV 20 2015

Abstract: There are many advantages to individual participant data meta-analysis for combining data from multiple studies. These advantages include greater power to detect effects, increased sample heterogeneity, and the ability to perform more sophisticated analyses than meta-analyses that rely on published results. However, a fundamental challenge is that it is unlikely that variables of interest are measured the same way in all of the studies to be combined. We propose that this situation can be viewed as a missing data problem in which some outcomes are entirely missing within some trials and use multiple imputation to fill in missing measurements. We apply our method to five longitudinal adolescent depression trials where four studies used one depression measure and the fifth study used a different depression measure. None of the five studies contained both depression measures. We describe a multiple imputation approach for filling in missing depression measures that makes use of external calibration studies in which both depression measures were used. We discuss some practical issues in developing the imputation model including taking into account treatment group and study. We present diagnostics for checking the fit of the imputation model and investigate whether external information is appropriately incorporated into the imputed values. Copyright (c) 2015John Wiley & Sons, Ltd.

Accession Number: WOS:000362502800002

PubMed ID: 26095855

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Siddique, Juned		0000-0002-1501-4152
Brown, C Hendricks		0000-0002-0294-2419
Crespi, Catherine		0000-0002-6150-2181

ISSN: 0277-6715

eISSN: 1097-0258

Record 23 of 444

Title: Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models

Author(s): Bondarenko, I (Bondarenko, Irina); Raghunathan, T (Raghunathan, Trivellore)

Source: STATISTICS IN MEDICINE **Volume:** 35 **Issue:** 17 **Pages:** 3007-3020 **DOI:** 10.1002/sim.6926 **Published:** JUL 30 2016

Abstract: Multiple imputation has become a popular approach for analyzing incomplete data. Many software packages are available to multiply impute the

missing values and to analyze the resulting completed data sets. However, diagnostic tools to check the validity of the imputations are limited, and the majority of the currently available methods need considerable knowledge of the imputation model. In many practical settings, however, the imputer and the analyst may be different individuals or from different organizations, and the analyst model may or may not be congenial to the model used by the imputer. This article develops and evaluates a set of graphical and numerical diagnostic tools for two practical purposes: (i) for an analyst to determine whether the imputations are reasonable under his/her model assumptions without actually knowing the imputation model assumptions; and (ii) for an imputer to fine tune the imputation model by checking the key characteristics of the observed and imputed values. The tools are based on the numerical and graphical comparisons of the distributions of the observed and imputed values conditional on the propensity of response. The methodology is illustrated using simulated data sets created under a variety of scenarios. The examples focus on continuous and binary variables, but the principles can be used to extend methods for other types of variables. Copyright (c) 2016 John Wiley & Sons, Ltd.

Accession Number: WOS:000379983000013

PubMed ID: 26952693

ISSN: 0277-6715

eISSN: 1097-0258

Record 24 of 444

Title: Convergence Properties of a Sequential Regression Multiple Imputation Algorithm

Author(s): Zhu, J (Zhu, Jian); Raghunathan, TE (Raghunathan, Trivellore E.)

Source: JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION **Volume:** 110 **Issue:** 511 **Pages:** 1112-1124 **DOI:** 10.1080/01621459.2014.948117 **Published:** SEP 2015

Abstract: A sequential regression or chained equations imputation approach uses a Gibbs sampling-type iterative algorithm that imputes the missing values using a sequence of conditional regression models. It is a flexible approach for handling different types of variables and complex data structures. Many simulation studies have shown that the multiple imputation inferences based on this procedure have desirable repeated sampling properties. However, a theoretical weakness of this approach is that the specification of a set of conditional regression models may not be compatible with a joint distribution of the variables being imputed. Hence, the convergence properties of the iterative algorithm are not well understood. This article develops conditions for convergence and assesses the properties of inferences from both compatible and incompatible sequence of regression models. The results are established for the missing data pattern where each subject may be missing a value on at most one variable. The sequence of regression models are assumed to be empirically good fit for the data chosen by the imputer based on appropriate model diagnostics. The results are used to develop criteria for the choice of regression models. Supplementary materials for this article are available online.

Accession Number: WOS:000365144600021

ISSN: 0162-1459

eISSN: 1537-274X

Record 25 of 444

Title: CONTINUOUS-TIME DISCRETE-SPACE MODELS FOR ANIMAL MOVEMENT

Author(s): Hanks, EM (Hanks, Ephraim M.); Hooten, MB (Hooten, Mevin B.); Alldredge, MW (Alldredge, Mat W.)

Source: ANNALS OF APPLIED STATISTICS **Volume:** 9 **Issue:** 1 **Pages:** 145-165 **DOI:** 10.1214/14-AOAS803 **Published:** MAR 2015

Abstract: The processes influencing animal movement and resource selection are complex and varied. Past efforts to model behavioral changes over time used Bayesian statistical models with variable parameter space, such as reversible-jump Markov chain Monte Carlo approaches, which are computationally demanding and inaccessible to many practitioners. We present a continuous-time discrete-space (CTDS) model of animal movement that can be fit using standard generalized linear modeling (GLM) methods. This CTDS approach allows for the joint modeling of location-based as well as directional drivers of movement. Changing behavior over time is modeled using a varying-coefficient framework which maintains the computational simplicity of a GLM approach, and variable selection is accomplished using a group lasso penalty. We apply our approach to a study of two mountain lions (Puma concolor) in Colorado, USA.

Accession Number: WOS:000358354400007

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Hanks, Ephraim		0000-0003-0345-7164

ISSN: 1932-6157

Record 26 of 444

Title: Multiple imputation in three or more stages

Author(s): McGinniss, J (McGinniss, J.); Harel, O (Harel, O.)

Source: JOURNAL OF STATISTICAL PLANNING AND INFERENCE **Volume:** 176 **Pages:** 33-51 **DOI:** 10.1016/j.jspi.2016.04.001 **Published:** SEP 2016

Abstract: Missing values present challenges in the analysis of data across many areas of research. Handling incomplete data incorrectly can lead to bias, over-confident intervals, and inaccurate inferences. One principled method of handling incomplete data is multiple imputation. This article considers incomplete data in which values are missing for three or more qualitatively different reasons and applies a modified multiple imputation framework in the analysis of that data. Included are a proof of the methodology used for three-stage multiple imputation with its limiting distribution, an extension to more than three types of missing values, an extension to the ignorability assumption with proof, and simulations demonstrating that the estimator is unbiased and efficient under the ignorability assumption. (C) 2016 Elsevier B.V. All rights reserved.

Accession Number: WOS:000377323600003

PubMed ID: 27647949

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Harel, Ofer	L-8906-2019	0000-0002-1054-3055

ISSN: 0378-3758

eISSN: 1873-1171

Record 27 of 444

Title: Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: The general location model

Author(s): Seaman, SR (Seaman, Shaun R.); Hughes, RA (Hughes, Rachael A.)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 27 **Issue:** 6 **Pages:** 1603-1614 **DOI:** 10.1177/0962280216665872 **Published:** JUN 2018

Abstract: Estimating the parameters of a regression model of interest is complicated by missing data on the variables in that model. Multiple imputation is commonly used to handle these missing data. Joint model multiple imputation and full-conditional specification multiple imputation are known to yield imputed data with the same asymptotic distribution when the conditional models of full-conditional specification are compatible with that joint model. We show that this asymptotic equivalence of imputation distributions does not imply that joint model multiple imputation and full-conditional specification multiple imputation will also yield asymptotically equally efficient inference about the parameters of the model of interest, nor that they will be equally robust to misspecification of the joint model. When the conditional models used by full-conditional specification multiple imputation are linear, logistic and multinomial regressions, these are compatible with a restricted general location joint model. We show that multiple imputation using the restricted general location joint model can be substantially more asymptotically efficient than full-conditional specification multiple imputation, but this typically requires very strong associations between variables. When associations are weaker, the efficiency gain is small. Moreover, full-conditional specification multiple imputation is shown to be potentially much more robust than joint model multiple imputation using the restricted general location model to mispecification of that model when there is substantial missingness in the outcome variable.

Accession Number: WOS:000432625800001

PubMed ID: 27597798

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
	K-7215-2019	
Hughes, Rachael		0000-0003-0766-1410
Seaman, Shaun		0000-0003-3726-5937

ISSN: 0962-2802

eISSN: 1477-0334

Record 28 of 444

Title: An imputation-based solution to using mismeasured covariates in propensity score analysis

Author(s): Webb-Vargas, Y (Webb-Vargas, Yenny); Rudolph, KE (Rudolph, Kara E.); Lenis, D (Lenis, David); Murakami, P (Murakami, Peter); Stuart, EA (Stuart, Elizabeth A.)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 26 **Issue:** 4 **Special Issue:** SI **Pages:** 1824-1837 **DOI:** 10.1177/0962280215588771 **Published:** AUG 2017

Abstract: Although covariate measurement error is likely the norm rather than the exception, methods for handling covariate measurement error in propensity score methods have not been widely investigated. We consider a multiple imputation-based approach that uses an external calibration sample with information on the true and mismeasured covariates, multiple imputation for external calibration, to correct for the measurement error, and investigate its performance using simulation studies. As expected, using the covariate measured with error leads to bias in the treatment effect estimate. In contrast, the multiple imputation for external calibration method can eliminate almost all the bias. We confirm that the outcome must be used in the imputation process to obtain good results, a finding related to the idea of congenial imputation and analysis in the broader multiple imputation literature. We illustrate the multiple imputation for external calibration approach using a motivating example estimating the effects of living in a disadvantaged neighborhood on mental health and substance use outcomes among adolescents. These results show that estimating the propensity score using covariates measured with error leads to biased estimates of treatment effects, but when a calibration data set is available, multiple imputation for external calibration can be used to help correct for such bias.

Accession Number: WOS:000407924800017

PubMed ID: 26037527

ISSN: 0962-2802

eISSN: 1477-0334

Record 29 of 444

Title: Simultaneous Edit-Imputation for Continuous Microdata

Author(s): Kim, HJ (Kim, Hang J.); Cox, LH (Cox, Lawrence H.); Karr, AF (Karr, Alan F.); Reiter, JP (Reiter, Jerome P.); Wang, QL (Wang, Quanli)

Source: JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION **Volume:** 110 **Issue:** 511 **Pages:** 987-999 **DOI:** 10.1080/01621459.2015.1040881 **Published:** SEP 2015

Abstract: Many statistical organizations collect data that are expected to satisfy linear constraints; as examples, component variables should sum to total variables, and ratios of pairs of variables should be bounded by expert-specified constants. When reported data violate constraints, organizations identify and replace values potentially in error in a process known as edit-imputation. To date, most approaches separate the error localization and imputation steps, typically using optimization methods to identify the variables to change followed by hot deck imputation. We present an approach. that fully integrates editing and imputation for continuous microdata under linear constraints. Our approach relies on a Bayesian hierarchical model that includes (i) a flexible joint probability model for the underlying true values of the data with support only on the set of values that satisfy all editing constraints, (ii) a model for latent indicators of the variables that are in error, and (iii) a model for the reported responses for variables in error. We illustrate the potential advantages of the Bayesian editing approach over existing approaches using simulation studies. We apply the model to edit faulty data from the 2007 U.S. Census of Manufactures. Supplementary materials for this article are available online.

Accession Number: WOS:000365144600011

ISSN: 0162-1459

eISSN: 1537-274X

Record 30 of 444

Title: Combining fractional polynomial model building with multiple imputation

Author(s): Morris, TP (Morris, Tim P.); White, IR (White, Ian R.); Carpenter, JR (Carpenter, James R.); Stanworth, SJ (Stanworth, Simon J.); Royston, P (Royston, Patrick)

Source: STATISTICS IN MEDICINE **Volume:** 34 **Issue:** 25 **Pages:** 3298-3317 **DOI:** 10.1002/sim.6553 **Published:** NOV 10 2015

Abstract: Multivariable fractional polynomial (MFP) models are commonly used in medical research. The datasets in which MFP models are applied often contain covariates with missing values. To handle the missing values, we describe methods for combining multiple imputation with MFP modelling, considering in turn three issues: first, how to impute so that the imputation model does not favour certain fractional polynomial (FP) models over others; second, how to estimate the FP exponents in multiply imputed data; and third, how to choose between models of differing complexity. Two imputation methods are outlined for different settings. For model selection, methods based on Wald-type statistics and weighted likelihood-ratio tests are proposed and evaluated in simulation studies. The Wald-based method is very slightly better at estimating FP exponents. Type I error rates are very similar for both methods, although slightly less well controlled than analysis of complete records; however, there is potential for substantial gains in power over the analysis of complete records. We illustrate the two methods in a dataset from five trauma registries for which a prognostic model has previously been published, contrasting the selected models with that obtained by analysing the complete records only. (c) 2015 The Authors. Statistics in Medicine Published by John Wiley & Sons Ltd.

Accession Number: WOS:000362426000002

PubMed ID: 26095614

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Koryakov, Dmitry E	N-4934-2015	
Zhimulev, Igor F	N-7978-2015	
Carpenter, James		0000-0003-3890-6206
Morris, Tim		0000-0001-5850-3610

ISSN: 0277-6715

eISSN: 1097-0258

Record 31 of 444

Title: Multiple imputation for continuous variables using a Bayesian principal component analysis

Author(s): Audigier, V (Audigier, Vincent); Husson, F (Husson, Francois); Josse, J (Josse, Julie)

Source: JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION **Volume:** 86 **Issue:** 11 **Pages:** 2140-2156 **DOI:** 10.1080/00949655.2015.1104683 **Published:** JUL 2016

Abstract: We propose a multiple imputation method based on principal component analysis (PCA) to deal with incomplete continuous data. To reflect the uncertainty of the parameters from one imputation to the next, we use a Bayesian treatment of the PCA model. Using a simulation study and real data sets, the method is compared to two classical approaches: multiple imputation based on joint modelling and on fully conditional modelling. Contrary to the others, the proposed method can be easily used on data sets where the number of individuals is less than the number of variables and when the variables are highly correlated. In addition, it provides unbiased point estimates of quantities of interest, such as an expectation, a regression coefficient or a correlation coefficient, with a smaller mean squared error. Furthermore, the widths of the confidence intervals built for the quantities of interest are often smaller whilst ensuring a valid coverage.

Accession Number: WOS:000375482300006

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
josse, julie		0000-0001-9547-891X

ISSN: 0094-9655

eISSN: 1563-5163

Record 32 of 444

Title: The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data

Author(s): Wood, AM (Wood, Angela M.); Royston, P (Royston, Patrick); White, IR (White, Ian R.)

Source: BIOMETRICAL JOURNAL **Volume:** 57 **Issue:** 4 **Pages:** 614-632 **DOI:** 10.1002/bimj.201400004 **Published:** JUL 2015

Abstract: Multiple imputation can be used as a tool in the process of constructing prediction models in medical and epidemiological studies with missing covariate values. Such models can be used to make predictions for model performance assessment, but the task is made more complicated by the multiple imputation structure. We summarize various predictions constructed from covariates, including multiply imputed covariates, and either the set of imputation-specific prediction model coefficients or the pooled prediction model coefficients. We further describe approaches for using the predictions to assess model performance. We distinguish between ideal model performance and pragmatic model performance, where the former refers to the model's performance in an ideal clinical setting where all individuals have fully observed predictors and the latter refers to the model's performance in a real-world clinical setting where some individuals have missing predictors. The approaches are compared through an extensive simulation study based on the UK700 trial. We determine that measures of ideal model performance can be estimated within imputed datasets and subsequently pooled to give an overall measure of model performance. Alternative methods to evaluate pragmatic model performance are required and we propose constructing predictions either from a second set of covariate imputations which make no use of observed outcomes, or from a set of partial prediction models constructed for each potential observed pattern of covariate. Pragmatic model performance is generally lower than ideal model performance. We focus on model performance within the derivation data, but describe how to extend all the methods to a validation dataset.

Accession Number: WOS:000357274300007

PubMed ID: 25630926

ISSN: 0323-3847

eISSN: 1521-4036

Record 33 of 444

Title: Penalized regression procedures for variable selection in the potential outcomes framework

Author(s): Ghosh, D (Ghosh, Debashis); Zhu, YY (Zhu, Yeying); Coffman, DL (Coffman, Donna L.)

Source: STATISTICS IN MEDICINE **Volume:** 34 **Issue:** 10 **Pages:** 1645-1658 **DOI:** 10.1002/sim.6433 **Published:** MAY 10 2015

Abstract: A recent topic of much interest in causal inference is model selection. In this article, we describe a framework in which to consider penalized regression approaches to variable selection for causal effects. The framework leads to a simple impute, then select' class of procedures that is agnostic to the type of imputation algorithm as well as penalized regression used. It also clarifies how model selection involves a multivariate regression model for causal inference problems and that these methods can be applied for identifying subgroups in which treatment effects are homogeneous. Analogies and links with the literature on machine learning methods, missing data, and imputation are drawn. A difference least absolute shrinkage and selection operator algorithm is defined, along with its multiple imputation analogs. The procedures are illustrated using a well-known right-heart catheterization dataset. Copyright (c) 2015 John Wiley & Sons, Ltd.

Accession Number: WOS:000352572200003

PubMed ID: 25628185

ISSN: 0277-6715

eISSN: 1097-0258

Record 34 of 444

Title: MIMCA: multiple imputation for categorical variables with multiple correspondence analysis

Author(s): Audigier, V (Audigier, Vincent); Husson, F (Husson, Francois); Josse, J (Josse, Julie)

Source: STATISTICS AND COMPUTING **Volume:** 27 **Issue:** 2 **Pages:** 501-518 **DOI:** 10.1007/s11222-016-9635-4 **Published:** MAR 2017

Abstract: We propose a multiple imputation method to deal with incomplete categorical data. This method imputes the missing entries using the principal component method dedicated to categorical data: multiple correspondence analysis (MCA). The uncertainty concerning the parameters of the imputation model is reflected using a non-parametric bootstrap. Multiple imputation using MCA (MIMCA) requires estimating a small number of parameters due to the dimensionality reduction property of MCA. It allows the user to impute a large range of data sets. In particular, a high number of categories per variable, a high number of variables or a small number of individuals are not an issue for MIMCA. Through a simulation study based on real data sets, the method is assessed and compared to the reference methods (multiple imputation using the loglinear model, multiple imputation by logistic regressions) as well to the latest works on the topic (multiple imputation by random forests or by the Dirichlet process mixture of products of multinomial distributions model). The proposed method provides a good point estimate of the parameters of the analysis model considered, such as the coefficients of a main effects logistic regression model, and a reliable estimate of the variability of the estimators. In addition, MIMCA has the great advantage that it is substantially less time consuming on data sets of high dimensions than the other multiple imputation methods.

Accession Number: WOS:000395004300013

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
josse, julie		0000-0001-9547-891X

ISSN: 0960-3174

eISSN: 1573-1375

Record 35 of 444

Title: Multiple imputation by chained equations for systematically and sporadically missing multilevel data

Author(s): Resche-Rigon, M (Resche-Rigon, Matthieu); White, IR (White, Ian R.)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 27 **Issue:** 6 **Pages:** 1634-1649 **DOI:** 10.1177/0962280216666564 **Published:** JUN 2018

Abstract: In multilevel settings such as individual participant data meta-analysis, a variable is systematically missing' if it is wholly missing in some clusters and sporadically missing' if it is partly missing in some clusters. Previously proposed methods to impute incomplete multilevel data handle either systematically or sporadically missing data, but frequently both patterns are observed. We describe a new multiple imputation by chained equations (MICE) algorithm for multilevel data with arbitrary patterns of systematically and sporadically missing variables. The algorithm is described for multilevel normal data but can easily be extended for other variable types. We first propose two methods for imputing a single incomplete variable: an extension of an existing method and a new two-stage method which conveniently allows for heteroscedastic data. We then discuss the difficulties of imputing missing values in several variables in multilevel data using MICE, and show that even the simplest joint multilevel model implies conditional models which involve cluster means and heteroscedasticity. However, a simulation study finds that the proposed methods can be successfully combined in a multilevel MICE procedure, even when cluster means are not included in the imputation models.

Accession Number: WOS:000432625800003

PubMed ID: 27647809

ISSN: 0962-2802

eISSN: 1477-0334

Record 36 of 444

Title: On analysis of longitudinal clinical trials with missing data using reference-based imputation

Author(s): Liu, GF (Liu, G. Frank); Pang, L (Pang, Lei)

Source: JOURNAL OF BIOPHARMACEUTICAL STATISTICS **Volume:** 26 **Issue:** 5 **Pages:** 924-936 **DOI:** 10.1080/10543406.2015.1094810 **Published:** 2016

Abstract: Reference-based imputation (RBI) methods have been proposed as sensitivity analyses for longitudinal clinical trials with missing data. The RBI methods multiply impute the missing data in treatment group based on an imputation model built using data from the reference (control) group. The RBI will yield a conservative treatment effect estimate as compared to the estimate obtained from multiple imputation (MI) under missing at random (MAR). However, the RBI analysis based on the regular MI approach can be overly conservative because it not only applies discount to treatment effect estimate but also posts penalty on the variance estimate. In this article, we investigate the statistical properties of RBI methods, and propose approaches to derive

accurate variance estimates using both frequentist and Bayesian methods for the RBI analysis. Results from simulation studies and applications to longitudinal clinical trial datasets are presented.

Accession Number: WOS:000384442400009

PubMed ID: 26418282

ISSN: 1054-3406

eISSN: 1520-5711

Record 37 of 444

Title: k-POD: A Method for k-Means Clustering of Missing Data

Author(s): Chi, JT (Chi, Jocelyn T.); Chi, EC (Chi, Eric C.); Baraniuk, RG (Baraniuk, Richard G.)

Source: AMERICAN STATISTICIAN **Volume:** 70 **Issue:** 1 **Pages:** 91-99 **DOI:** 10.1080/00031305.2015.1086685 **Published:** JAN 2 2016

Abstract: The k-means algorithm is often used in clustering applications but its usage requires a complete data matrix. Missing data, however, are common in many applications. Mainstream approaches to clustering missing data reduce the missing data problem to a complete data formulation through either deletion or imputation but these solutions may incur significant costs. Our k-POD method presents a simple extension of k-means clustering for missing data that works even when the missingness mechanism is unknown, when external information is unavailable, and when there is significant missingness in the data.

Accession Number: WOS:000373801000011

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Chi, Eric	J-2708-2019	0000-0003-4647-0895

ISSN: 0003-1305

eISSN: 1537-2731

Record 38 of 444

Title: Meta-analysis with missing study-level sample variance data

Author(s): Chowdhry, AK (Chowdhry, Amit K.); Dworkin, RH (Dworkin, Robert H.); McDermott, MP (McDermott, Michael P.)

Source: STATISTICS IN MEDICINE **Volume:** 35 **Issue:** 17 **Pages:** 3021-3032 **DOI:** 10.1002/sim.6908 **Published:** JUL 30 2016

Abstract: We consider a study-level meta-analysis with a normally distributed outcome variable and possibly unequal study-level variances, where the object of inference is the difference in means between a treatment and control group. A common complication in such an analysis is missing sample variances for some studies. A frequently used approach is to impute the weighted (by sample size) mean of the observed variances (mean imputation). Another approach is to include only those studies with variances reported (complete case analysis). Both mean imputation and complete case analysis are only valid under the missing-completely-at-random assumption, and even then the inverse variance weights produced are not necessarily optimal. We propose a multiple imputation method employing gamma meta-regression to impute the missing sample variances. Our method takes advantage of study-level covariates that may be used to provide information about the missing data. Through simulation studies, we show that multiple imputation, when the imputation model is correctly specified, is superior to competing methods in terms of confidence interval coverage probability and type I error probability when testing a specified group difference. Finally, we describe a similar approach to handling missing variances in cross-over studies. Copyright (c) 2016 John Wiley & Sons, Ltd.

Accession Number: WOS:000379983000014

PubMed ID: 26888093

ISSN: 0277-6715

eISSN: 1097-0258

Record 39 of 444

Title: Number of imputations needed to stabilize estimated treatment difference in longitudinal data analysis

Author(s): Lu, KF (Lu, Kaifeng)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 26 **Issue:** 2 **Pages:** 674-690 **DOI:** 10.1177/0962280214554439 **Published:** APR 2017

Abstract: Multiple imputation procedures replace each missing value with a set of plausible values based on the posterior predictive distribution of missing data given observed data. In many applications, as few as five imputations are adequate to achieve high efficiency relative to an infinite number of imputations. However, substantially more imputations are often needed to stabilize imputation-based inference at the analysis stage. Imputation-based inference at the analysis stage is considered stable if the conditional variability of the multiple imputation estimator, half-width of 95% confidence interval, test statistic, and estimated fraction of missing information given observed data is within specified thresholds for simulation error. For the estimation of treatment difference at study end for normally distributed responses in longitudinal trials, we calculate the multiple imputation quantities for an infinite number of imputations analytically and use simulations to assess the variability of the number of imputations needed at the analysis stage in repeated sampling.

Accession Number: WOS:000399704500009

PubMed ID: 25305196

ISSN: 0962-2802

eISSN: 1477-0334

Record 40 of 444

Title: A multiple imputation approach for MNAR mechanisms compatible with Heckman's model

Author(s): Galimard, JE (Galimard, Jacques-Emmanuel); Chevret, S (Chevret, Sylvie); Protopopescu, C (Protopopescu, Camelia); Resche-Rigon, M (Resche-Rigon, Matthieu)

Source: STATISTICS IN MEDICINE **Volume:** 35 **Issue:** 17 **Pages:** 2907-2920 **DOI:** 10.1002/sim.6902 **Published:** JUL 30 2016

Abstract: Standard implementations of multiple imputation (MI) approaches provide unbiased inferences based on an assumption of underlying missing at random (MAR) mechanisms. However, in the presence of missing data generated by missing not at random (MNAR) mechanisms, MI is not satisfactory.

Originating in an econometric statistical context, Heckman's model, also called the sample selection method, deals with selected samples using two joined linear equations, termed the selection equation and the outcome equation. It has been successfully applied to MNAR outcomes. Nevertheless, such a method only addresses missing outcomes, and this is a strong limitation in clinical epidemiology settings, where covariates are also often missing. We propose to extend the validity of MI to some MNAR mechanisms through the use of the Heckman's model as imputation model and a two-step estimation process. This approach will provide a solution that can be used in an MI by chained equation framework to impute missing (either outcomes or covariates) data resulting either from a MAR or an MNAR mechanism when the MNAR mechanism is compatible with a Heckman's model. The approach is illustrated on a real dataset from a randomised trial in patients with seasonal influenza. Copyright (c) 2016 John Wiley & Sons, Ltd.

Accession Number: WOS:000379983000007

PubMed ID: 26893215

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
resche-rigon, matthieu		0000-0003-2220-5085
GALIMARD, Jacques-Emmanuel		0000-0001-9102-4427
chevret, sylvie		0000-0001-6449-4730

ISSN: 0277-6715

eISSN: 1097-0258

Record 41 of 444

Title: MGAS: a powerful tool for multivariate gene-based genome-wide association analysis

Author(s): Van der Sluis, S (Van der Sluis, Sophie); Dolan, CV (Dolan, Conor V.); Li, J (Li, Jiang); Song, YQ (Song, Youqiang); Sham, P (Sham, Pak); Posthuma, D (Posthuma, Danielle); Li, MX (Li, Miao-Xin)

Source: BIOINFORMATICS **Volume:** 31 **Issue:** 7 **Pages:** 1007-1015 **DOI:** 10.1093/bioinformatics/btu783 **Published:** APR 1 2015

Abstract: Motivation: Standard genome-wide association studies, testing the association between one phenotype and a large number of single nucleotide polymorphisms (SNPs), are limited in two ways: (i) traits are often multivariate, and analysis of composite scores entails loss in statistical power and (ii) gene-based analyses may be preferred, e.g. to decrease the multiple testing problem.
Results: Here we present a new method, multivariate gene-based association test by extended Simes procedure (MGAS), that allows gene-based testing of multivariate phenotypes in unrelated individuals. Through extensive simulation, we show that under most trait-generating genotype-phenotype models MGAS has superior statistical power to detect associated genes compared with gene-based analyses of univariate phenotypic composite scores (i.e. GATES, multiple regression), and multivariate analysis of variance (MANOVA). Re-analysis of metabolic data revealed 32 False Discovery Rate controlled genome-wide significant genes, and 12 regions harboring multiple genes; of these 44 regions, 30 were not reported in the original analysis.
Conclusion: MGAS allows researchers to conduct their multivariate gene-based analyses efficiently, and without the loss of power that is often associated with an incorrectly specified genotype-phenotype models.

Accession Number: WOS:000352269500005

PubMed ID: 25431328

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Posthuma, Danielle		0000-0001-7582-2365

ISSN: 1367-4803

eISSN: 1460-2059

Record 42 of 444

Title: Random forest missing data algorithms

Author(s): Tang, F (Tang, Fei); Ishwaran, H (Ishwaran, Hemant)

Source: STATISTICAL ANALYSIS AND DATA MINING **Volume:** 10 **Issue:** 6 **Pages:** 363-377 **DOI:** 10.1002/sam.11348 **Published:** DEC 2017

Abstract: Random forest (RF) missing data algorithms are an attractive approach for imputing missing data. They have the desirable properties of being able to handle mixed types of missing data, they are adaptive to interactions and nonlinearity, and they have the potential to scale to big data settings. Currently there are many different RF imputation algorithms, but relatively little guidance about their efficacy. Using a large, diverse collection of data sets, imputation performance of various RF algorithms was assessed under different missing data mechanisms. Algorithms included proximity imputation, on the fly imputation, and imputation utilizing multivariate unsupervised and supervised splittingthe latter class representing a generalization of a new promising imputation algorithm called missForest. Our findings reveal RF imputation to be generally robust with performance improving with increasing correlation. Performance was good under moderate to high missingness, and even (in certain cases) when data was missing not at random.

Accession Number: WOS:000415735500001

PubMed ID: 29403567

ISSN: 1932-1864

eISSN: 1932-1872

Record 43 of 444

Title: A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies

Author(s): Kalaycioglu, O (Kalaycioglu, Oya); Copas, A (Copas, Andrew); King, M (King, Michael); Omar, RZ (Omar, Rumana Z.)

Source: JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES A-STATISTICS IN SOCIETY **Volume:** 179 **Issue:** 3 **Pages:** 683-706 **DOI:** 10.1111/rssa.12140 **Published:** JUN 2016

Abstract: Multiple-imputation (MI) methods for imputing missing data in observational health studies with repeated measurements were evaluated with particular focus on incomplete time varying explanatory variables. Standard and random-effects imputation by chained equations, multivariate normal imputation and Bayesian MI were compared regarding bias and efficiency of regression coefficient estimates by using simulation studies. Flexibility of the methods in handling different types of variables (binary, categorical, skewed and normally distributed) and correlations between the repeated

measurements of the incomplete variables were also compared. Multivariate normal imputation produced the least bias in most situations, is theoretically well justified and allows flexible correlation for the repeated measurements. It can be recommended for imputing continuous variables. Bayesian MI is efficient and may be preferable in the presence of categorical and non-normally distributed continuous variables. Imputation by chained equations approaches were sensitive to the correlation between the repeated measurements. The moving time window approach may be used for normally distributed continuous variables with auto-regressive correlation.

Accession Number: WOS:000376152200004

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Copas, Andrew		0000-0001-8968-5963

ISSN: 0964-1998

eISSN: 1467-985X

Record 44 of 444

Title: Improving Imputation Accuracy by Inferring Causal Variants in Genetic Studies

Author(s): Wu, Y (Wu, Yue); Hormozdiari, F (Hormozdiari, Farhad); Joo, JWJ (Joo, Jong Wha J.); Eskin, E (Eskin, Eleazar)

Source: JOURNAL OF COMPUTATIONAL BIOLOGY **DOI:** 10.1089/cmb.2018.0139 **Published:** OCT 1 2018

Abstract: Genotype imputation has been widely utilized for two reasons in the analysis of genome-wide association studies (GWAS). One reason is to increase the power for association studies when causal single nucleotide polymorphisms are not collected in the GWAS. The second reason is to aid the interpretation of a GWAS result by predicting the association statistics at untyped variants. In this article, we show that prediction of association statistics at untyped variants that have an influence on the trait produces is overly conservative. Current imputation methods assume that none of the variants in a region (locus consists of multiple variants) affect the trait, which is often inconsistent with the observed data. In this article, we propose a new method, CAUSAL-Imp, which can impute the association statistics at untyped variants while taking into account variants in the region that may affect the trait. Our method builds on recent methods that impute the marginal statistics for GWAS by utilizing the fact that marginal statistics follow a multivariate normal distribution. We utilize both simulated and real data sets to assess the performance of our method. We show that traditional imputation approaches underestimate the association statistics for variants involved in the trait, and our results demonstrate that our approach provides less biased estimates of these association statistics.

Accession Number: WOS:000446003300001

PubMed ID: 30272994

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Wu, Yuanqing		0000-0001-9509-2670

ISSN: 1066-5277

eISSN: 1557-8666

Record 45 of 444

Title: Combining Inverse Probability Weighting and Multiple Imputation to Improve Robustness of Estimation

Author(s): Han, PS (Han, Peisong)

Source: SCANDINAVIAN JOURNAL OF STATISTICS **Volume:** 43 **Issue:** 1 **Pages:** 246-260 **DOI:** 10.1111/sjos.12177 **Published:** MAR 2016

Abstract: Inverse probability weighting (IPW) and multiple imputation are two widely adopted approaches dealing with missing data. The former models the selection probability, and the latter models data distribution. Consistent estimation requires correct specification of corresponding models. Although the augmented IPW method provides an extra layer of protection on consistency, it is usually not sufficient in practice as the true data-generating process is unknown. This paper proposes a method combining the two approaches in the same spirit of calibration in sampling survey literature. Multiple models for both the selection probability and data distribution can be simultaneously accounted for, and the resulting estimator is consistent if any model is correctly specified. The proposed method is within the framework of estimating equations and is general enough to cover regression analysis with missing outcomes and/or missing covariates. Results on both theoretical and numerical investigation are provided.

Accession Number: WOS:000371237300017

ISSN: 0303-6898

eISSN: 1467-9469

Record 46 of 444

Title: Posterior predictive checking of multiple imputation models

Author(s): Nguyen, CD (Nguyen, Cattram D.); Lee, KJ (Lee, Katherine J.); Carlin, JB (Carlin, John B.)

Source: BIOMETRICAL JOURNAL **Volume:** 57 **Issue:** 4 **Pages:** 676-694 **DOI:** 10.1002/bimj.201400034 **Published:** JUL 2015

Abstract: Multiple imputation is gaining popularity as a strategy for handling missing data, but there is a scarcity of tools for checking imputation models, a critical step in model fitting. Posterior predictive checking (PPC) has been recommended as an imputation diagnostic. PPC involves simulating replicated data from the posterior predictive distribution of the model under scrutiny. Model fit is assessed by examining whether the analysis from the observed data appears typical of results obtained from the replicates produced by the model. A proposed diagnostic measure is the posterior predictive p-value, an extreme value of which (i.e., a value close to 0 or 1) suggests a misfit between the model and the data. The aim of this study was to evaluate the performance of the posterior predictive p-value as an imputation diagnostic. Using simulation methods, we deliberately misspecified imputation models to determine whether posterior predictive p-values were effective in identifying these problems. When estimating the regression parameter of interest, we found that more extreme p-values were associated with poorer imputation model performance, although the results highlighted that traditional thresholds for classical p-values do not apply in this context. A shortcoming of the PPC method was its reduced ability to detect misspecified models with increasing amounts of missing data. Despite the limitations of posterior predictive p-values, they appear to have a valuable place in the imputer's toolkit. In addition to automated checking using p-values, we recommend imputers perform graphical checks and examine other summaries of the test quantity distribution.

Accession Number: WOS:000357274300011

PubMed ID: 25939490

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Carlin, John	B-3492-2012	0000-0002-2694-9463
Lee, Katherine	A-2519-2016	

ISSN: 0323-3847

eISSN: 1521-4036

Record 47 of 444

Title: Efficient Quantile Regression Analysis With Missing Observations

Author(s): Chen, XR (Chen, Xuerong); Wan, ATK (Wan, Alan T. K.); Zhou, Y (Zhou, Yong)

Source: JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION **Volume:** 110 **Issue:** 510 **Pages:** 723-741 **DOI:** 10.1080/01621459.2014.928219 **Published:** JUN 2015

Abstract: This article examines the problem of estimation in a quantile regression model when observations are missing at random under independent and nonidentically distributed errors. We consider three approaches of handling this problem based on nonparametric inverse probability weighting, estimating equations projection, and a combination of both. An important distinguishing feature of our methods is their ability to handle missing response and/or partially missing covariates, whereas existing techniques can handle only one or the other, but not both. We prove that our methods yield asymptotically equivalent estimators that achieve the desirable asymptotic properties of unbiasedness, normality, and root n-consistency. Because we do not assume that the errors are identically distributed, our theoretical results are valid under heteroscedasticity, a particularly strong feature of our methods. Under the special case of identical error distributions, all of our proposed estimators achieve the semiparametric efficiency bound. To facilitate the practical implementation of these methods, we develop an iterative method based on the majorize/minimize algorithm for computing the quantile regression estimates, and a bootstrap method for computing their variances. Our simulation findings suggest that all three methods have good finite sample properties. We further illustrate these methods by a real data example. Supplementary materials for this article are available online.

Accession Number: WOS:000357437300021

ISSN: 0162-1459

eISSN: 1537-274X

Record 48 of 444

Title: Simulation-Based Study Comparing Multiple Imputation Methods for Non-Monotone Missing Ordinal Data in Longitudinal Settings

Author(s): Donneau, AF (Donneau, A. F.); Mauer, M (Mauer, M.); Lambert, P (Lambert, P.); Molenberghs, G (Molenberghs, G.); Albert, A (Albert, A.)

Source: Journal of Biopharmaceutical Statistics **Volume:** 25 **Issue:** 3 **Pages:** 570-601 **DOI:** 10.1080/10543406.2014.920864 **Published:** MAY 4 2015

Abstract: The application of multiple imputation (MI) techniques as a preliminary step to handle missing values in data analysis is well established. The MI method can be classified into two broad classes, the joint modeling and the fully conditional specification approaches. Their relative performance for the longitudinal ordinal data setting under the missing at random (MAR) assumption is not well documented. This article intends to fill this gap by conducting a large simulation study on the estimation of the parameters of a longitudinal proportional odds model. The two MI methods are also illustrated in quality of life data from a cancer clinical trial.

Accession Number: WOS:000353386300013

PubMed ID: 24905056

ISSN: 1054-3406

eISSN: 1520-5711

Record 49 of 444

Title: Propensity score analysis with partially observed covariates: How should multiple imputation be used?

Author(s): Leyrat, C (Leyrat, Clemence); Seaman, SR (Seaman, Shaun R.); White, IR (White, Ian R.); Douglas, I (Douglas, Ian); Smeeth, L (Smeeth, Liam); Kim, J (Kim, Joseph); Resche-Rigon, M (Resche-Rigon, Matthieu); Carpenter, JR (Carpenter, James R.); Williamson, EJ (Williamson, Elizabeth J.)

Source: STATISTICAL METHODS IN MEDICAL RESEARCH **Volume:** 28 **Issue:** 1 **Pages:** 3-19 **DOI:** 10.1177/0962280217713032 **Published:** JAN 2019

Abstract: Inverse probability of treatment weighting is a popular propensity score-based approach to estimate marginal treatment effects in observational studies at risk of confounding bias. A major issue when estimating the propensity score is the presence of partially observed covariates. Multiple imputation is a natural approach to handle missing data on covariates: covariates are imputed and a propensity score analysis is performed in each imputed dataset to estimate the treatment effect. The treatment effect estimates from each imputed dataset are then combined to obtain an overall estimate. We call this method MIte. However, an alternative approach has been proposed, in which the propensity scores are combined across the imputed datasets (MIps). Therefore, there are remaining uncertainties about how to implement multiple imputation for propensity score analysis: (a) should we apply Rubin's rules to the inverse probability of treatment weighting treatment effect estimates or to the propensity score estimates themselves? (b) does the outcome have to be included in the imputation model? (c) how should we estimate the variance of the inverse probability of treatment weighting estimator after multiple imputation? We studied the consistency and balancing properties of the MIte and MIps estimators and performed a simulation study to empirically assess their performance for the analysis of a binary outcome. We also compared the performance of these methods to complete case analysis and the missingness pattern approach, which uses a different propensity score model for each pattern of missingness, and a third multiple imputation approach in which the propensity score parameters are combined rather than the propensity scores themselves (MIpar). Under a missing at random mechanism, complete case and missingness pattern analyses were biased in most cases for estimating the marginal treatment effect, whereas multiple imputation approaches were approximately unbiased as long as the outcome was included in the imputation model. Only MIte was unbiased in all the studied scenarios and Rubin's rules provided good variance estimates for MIte. The propensity score estimated in the MIte approach showed good balancing properties. In conclusion, when using multiple imputation in the inverse probability of treatment weighting context, MIte with the outcome included in the imputation model is the preferred approach.

Accession Number: WOS:000454598800001

PubMed ID: 28573919

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
resche-rigon, matthieu		0000-0003-2220-5085
Leyrat, Clemence		0000-0002-4097-4577
Seaman, Shaun		0000-0003-3726-5937

ISSN: 0962-2802
eISSN: 1477-0334

Record 50 of 444

Title: Multiple imputation methods for bivariate outcomes in cluster randomised trials

Author(s): DiazOrdaz, K (DiazOrdaz, K.); Kenward, MG (Kenward, M. G.); Gomes, M (Gomes, M.); Grieve, R (Grieve, R.)

Source: STATISTICS IN MEDICINE **Volume:** 35 **Issue:** 20 **Pages:** 3482-3496 **DOI:** 10.1002/sim.6935 **Published:** SEP 10 2016

Abstract: Missing observations are common in cluster randomised trials. The problem is exacerbated when modelling bivariate outcomes jointly, as the proportion of complete cases is often considerably smaller than the proportion having either of the outcomes fully observed. Approaches taken to handling such missing data include the following: complete case analysis, single-level multiple imputation that ignores the clustering, multiple imputation with a fixed effect for each cluster and multilevel multiple imputation. We contrasted the alternative approaches to handling missing data in a cost-effectiveness analysis that uses data from a cluster randomised trial to evaluate an exercise intervention for care home residents. We then conducted a simulation study to assess the performance of these approaches on bivariate continuous outcomes, in terms of confidence interval coverage and empirical bias in the estimated treatment effects. Missing-at-random clustered data scenarios were simulated following a full-factorial design. Across all the missing data mechanisms considered, the multiple imputation methods provided estimators with negligible bias, while complete case analysis resulted in biased treatment effect estimates in scenarios where the randomised treatment arm was associated with missingness. Confidence interval coverage was generally in excess of nominal levels (up to 99.8%) following fixed-effects multiple imputation and too low following single-level multiple imputation. Multilevel multiple imputation led to coverage levels of approximately 95% throughout. (c) 2016 The Authors. Statistics in Medicine Published by John Wiley & Sons Ltd.

Accession Number: WOS:000380728800003
PubMed ID: 26990655

Author Identifiers:

Author	Web of Science ResearcherID	ORCID Number
Kenward, Michael		0000-0003-0808-4192
Grieve, Richard		0000-0001-8899-1301
DiazOrdaz, Karla		0000-0003-3155-1561

ISSN: 0277-6715
eISSN: 1097-0258

Clarivate
Accelerating innovation

© 2019 Clarivate Copyright notice Terms of use Privacy statement Cookie policy

Sign up for the Web of Science newsletter Follow us