دانشگاه علوم پزشکی

و خدمات بهداشتی درمانی کرمان

دانشکده بهداشت

پایان نامه مقطع دکترای تخصصی (Ph.D) آمار زیستی

عنوان

مقایسه‌ی عملکرد مدل‌های افراز بازگشتی و مدل‌های جانهی پارامتری بر پایه‌ی جورسازی میانگین پیش‌بینی در جانهی چندگانه با معادلات زنجیره‌ای، درحضور اثرمتقابل

توسط

سارا جوادی

اساتید راهنما

دکتر محمدرضا بانشی

دکتر عباس بهرامپور

شماره پایان‌نامه 97/10/8/1/21 سال تحصیلی (آبان 1400)

Kerman University of Medical Sciences

Faculty of Medicine

In Partial Fulfillment of the Requirements for the Degree (PhD)

Title

Comparison of performance of recursive partitioning models and parametric imputation models based on predictive mean matching in multiple imputation by chained equations, in the presence of an interaction effect

By

Sara Javadi

Supervisor/s

Mohammad Reza Baneshi

Abbas Bahrampour

Thesis No : (97.10.8/1/21) **(Oct, 2021)**

چکیده

مقدمه و اهداف: در میان روش‌های نوین جانهی چندگانه، روش‌های جانهی چندگانه با معادلات زنجیره‌ای (MICE) بسیار پر کاربرد هستند. تمرکز اصلی در این رساله بر مقایسه عملکرد مدل‌های جانهی پارامتری بر پایه‌ی جورسازی میانگین پیش‌بینی (PMM) و روش‌های افراز بازگشتی در MICE، در حضور اثرمتقابل در مجموعه مشاهدات است. در روش PMM که پیش فرض اغلب نرم‌افزارهای آماری است، تنها اثرات اصلی وارد مدل جانهی می‌گردند و مدل جانهی PMM شامل اثرات متقابل نیست، بنابراین به کارگیری این روش در مشاهدات پیچیده که شامل اثرات متقابل هستند صحیح نمی‌باشد. در این رساله مدل‌های پارامتری با در نظر گرفتن و یا وارد نمودن اثر متقابل در مدل جانهی، با روش‌های افراز بازگشتی در MICE مقایسه گردیدند.دلیل استفاده نمودن از روش‌های افراز بازگشتی این بود که این روش‌ها اثرات غیر خطی و متقابل را به طور خودکار کشف می‌کنند و سلیقه‌ی شخصی محقق در فرآیند جانهی مقادیر گمشده دخالتی ندارد.

روش‌ها: با استفاده از شبیه‌سازی سناریوهای مختلفی با تغییر نوع پاسخ (دو حالتی و کمی)، نوع اثر متقابل (اثر متقابل بین دو متغیر دو حالتی، اثر متقابل بین دو متغیر کمی و اثر متقابل بین متغیر کمی و متغیر دو حالتی)، درصد گمشدگی در متغیر پاسخ (10% تا 50%) و سازوکار گمشدگی (گمشدگی تصادفی و گمشدگی کاملا تصادفی) طراحی گردیدند.دو روش جانهی که با در نظر گرفتن اثرمتقابل در فرآیند جانهی PMM در MICE ایجاد گردیدند تحت عناوین MICE-Interaction و MICE-Stratified مشخص شدند و دو روش افراز بازگشتی که بر پایه‌ی درخت رده‌بندی و رگرسیونی، و جنگل تصادفی ایجاد گردیدند روش‌های MICE-CART و MICE-RF بودند. چهار روش مذکور بر مشاهدات ناقص اعمال گردیدند و سپس عملکرد آن‌ها مقایسه شدند.همه‌ی مشاهدات شبیه سازی شده شامل ترکیبی از متغیرهای مستقل کمی و دو حالتی بودند. به طور کلی معیارهای ارزیابی عملکرد اریبی (خام)، اریبی نسبی، پوشش، خطای معیار مدل، خطای معیار تجربی و تغییرات منتسب به مقادیر گمشده برای$240=2\times3\times5\times2\times4$ حالت محاسبه گردیدند. داده‌ی واقعی استفاده شده در این رساله شامل مقادیر گمشده نبود. در ابتدا 300 نمونه‌ی بوت استرپ از مجموعه داده‌ی واقعی استخراج گردیدند، سپس به طور مصنوعی 10% تا 50% از مقادیر متغیر پاسخ با دو سازوکار گمشدگی کاملا تصادفی و گمشدگی تصادفی حذف گردیدند، سپس عملکرد روش‌های جانهی مقایسه شدند.

یافته‌ها:در سناریوهایی با پاسخ دو حالتی، روش MICE-Interaction موجب کاهش مقادیر قدر مطلق اریبی نسبی و افزایش پوشش نسبت به روش‌های افراز بازگشتی در MICE گردید. اگر قدرمطلق اریبی نسبی کمتر مساوی با 0.20 قابل اغماض در نظر گرفته شود، روش افراز بازگشتی MICE-CART در حداقل 70٪ از کل حالات از نظر معیارهای عملکرد مورد نظر همچون اریبی نسبی، پوشش، تغییرات منتسب به مقادیر از دست رفته

عملکرد قابل قبول و خوبی داشت. در سناریوهایی با پاسخ کمی و اثر متقابل بین دو متغیر دو حالتی، روش MICE-Interaction و روش افراز بازگشتی RF در MICE در برآورد اثر متقابل منجر به قدرمطلق اریبی نسبی کمتر از 0.05 و پوشش قابل قبول گردیدند (برای مثال در 30 % گمشدگی با سازوکار MAR اریبی نسبی و پوشش برای روش MICE-Interaction به ترتیب برابر با 0.008 - و 0.996 بودند و اریبی نسبی و پوشش برای روش MICE-RF به ترتیب برابر با 0.003- و یک بودند). در سناریوهایی با پاسخ کمی و اثر متقابل بین دو متغیر کمی در برآورد اثر متقابل، MICE-Interaction عملکرد قابل قبولی داشت (برای مثال در 30% گمشدگی در سازوکار MAR اریبی نسبی برابر با 0.016- و پوشش برابر با 0.962 بودند) و در برآورد جملات سهیم در اثر متقابل هر چند که عملکرد روش MICE-Interaction نسبت به سایر روش‌ها بهتر بود روش‌های افراز بازگشتی در MICE نیز قدر مطلق اریبی نسبی کمتر از 0.05 و پوشش کامل ایجاد نمودند. اگر مقدار عددی 0.20 برای قدرمطلق اریبی نسبی قابل اغماض باشد، هر دو روش افراز بازگشتی با در نظر گرفتن همزمان معیارهای عملکردی همچون اریبی نسبی، پوشش و تغییرات منتسب به مقادیر از دست رفته عملکرد قابل قبولی داشتند. در سناریوهایی با پاسخ کمی که اثرمتقابل بین متغیر دوحالتی و کمی وجود داشتند، در نسبت‌های گمشدگی بیشتر از 0.20، مقادیر پوشش روش‌های افراز بازگشتی قابل قبول نبودند.

بحث و نتیجه‌گیری: در مواردی که ساختار مشاهدات و اثرات متقابل مشخص باشند و محقق به طور صحیح بتواند اثر متقابل را به طور دستی وارد مدل جانهی نماید، به کارگیری روش MICE-Interaction در اغلب سناریوها منجر به عملکرد بهتر نسبت به روش‌های افراز بازگشتی در MICE گردید. در چنین شرایطی نتایج روش MICE-Interactionنتایج ایده‌آلی هستند و ممکن است در واقعیت اتفاق نیفتد زیرا تشخیص و کشف دقیق همه‌ی اثرات متقابل و روابط موجود در همه‌ی مشاهدات چالش برانگیز و دشوار هستند. در واقع اگر محقق علاقمند به برآورد اثر متقابل باشد و اطلاعات کافی از ساختار مشاهدات نداشته باشد بهتر است به سراغ روش‌های مبتنی بر درخت در MICE برود زیرا این مدل‌ها اثرمتقابل و غیرخطی را به طور خودکار و ذاتی کشف می‌نمایند و در چنین شرایطی به کارگیری روش PMM در MICE که پیش‌فرض اغلب نرم‌افزارهای آماری است توصیه نمی‌گردد.

Abstract

Background and Aim: Among the new multiple imputation methods, Multiple Imputation by Chained Equations (MICE) is a popular approach for implementing multiple imputation because of its flexibility. Our main focus in this dissertation is to compare the performance of the imputation models based on PMM and recursive partitioning methods in MICE in the presence of interaction in the data. The main focus of this dissertation is on comparing the performance of parametric (semi-parametric) models based on PMM and recursive partitioning methods in MICE, in the presence of interaction in the set of observations. In the PMM method, which is the default of most statistical software, only the main effects enter the imputation model, and the PMM imputation model does not include interactions, so it is not correct to use this method in complex observations that involve interactions. In this paper, parametric models were compared with recursive partitioning in MICE, considering the interaction in imputation model. The reason for using recursive partitioning methods is that these methods automatically detect nonlinear and interaction effects and the user's personal opinion does not interfere in the process of imputing missing values.

Method: Using simulations, different scenarios were created by changing the type of response (binary and continuous), the type of interaction (interaction between two binary variables, the interaction between two continuous variables and the interaction between a continuous and a binary variables), The percentage of missing values in the response variable (10% to 50%) and the mechanism of missing data (MAR and MCAR) were designed. Then the performance of recursive partitioning methods in MICE (MICE-CART and MICE-RF) with two parametric methods (semi-parametric) based on PMM (MICE-Interaction and MICE-Stratified) were compared. All simulation observations included a combination of continuous and binary independent variables. The performance of the methods was evaluated over 1000 simulations on the following outcome variables: bias, relative bias, coverage, model based standard error, empirical standard error and estimated proportion of the variance attributable to the missing data. The real data used in this dissertation did not include the missing values and at first 10% to 50%

of the missing values were artificially generated on the response, then the performances of the imputation methods were compared.

Results: In the scenarios with binary response variable, the proper inclusion of the interaction in the PMM model reduced the relative bias and increased the coverage compared to the recursive partitioning methods in MICE. If the absolute relative bias value is less than 0.20 negligible, the MICE-CART recursive partitioning method in at least 70% of all cases in terms of desired performance criteria such as relative bias, coverage, variations attributable to missing values had acceptable performance. In scenarios with continuous response and interaction between two binary variables, the MICE-Interaction method and the RF recursive partitioning method in MICE resulted in absolute relative bias less than 0.05 and complete coverage (for instance, at 30% missing values with MAR mechanism, the relative bias and coverage for MICE-Interaction method were -0.008 and 0.996, respectively, and the relative bias and coverage method for MICE-RF were -0.003 and one, respectively. In scenarios with continuous response and interaction between two continuous variables, MICE-Interaction had an acceptable performance in estimating the interaction (for instance, at 30% missing values in the MAR mechanism, the relative bias equal to -0.016 and coverage were 0.962) and in estimating the effects of variables that contributed to the interaction, although the performance of MICE-Interaction method was better than other methods, the recursive partitioning methods in MICE also created relative bias values less than 0.05 and complete coverage.

Discussion and Conclusion: In cases where the user can correctly enter the interaction manually into the imputation model, the application of MICE-Interaction in most scenarios led to better performance than recursive partitioning methods in MICE. In fact, if the user is interested in estimating the interaction and does not know enough about the structure of the observations, tree-based methods can be suggested to impute the missing values, but if the user can identify the appropriate model suitable for the observations, the MICE-Interaction method recommended. In general, the default PMM method in MICE did not perform well in estimating interactions in terms of performance criteria of imputation models

اظهارنامه حق انتشار

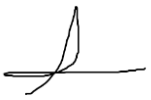
بدین‌وسیله اعلام می‌نمایم که این پایان‌نامه حاصل کار تحقیقاتی این‌جانب سارا جوادی دانشجوی مقطع دکترای آمار زیستی دانشکده بهداشت دانشگاه علوم پزشکی کرمان می‌باشد و در هیچ دانشگاه یا مرکز تحقیقاتی و آموزشی دیگری برای دریافت مدرک علمی ارائه نگردیده‌است.

اطلاعات علمی که از نتایج تحقیقات چاپ‌شده یا چاپ‌نشده دیگران اخذ شده‌است، با ذکر منبع، در متن پایان‌نامه و در فهرست منابع آورده شده‌است.

کلیه حقوق اعم از چاپ، تکثیر، نسخه‌برداری، ترجمه، اقتباس و ... از نتایج این پایان‌نامه، برای دانشگاه علوم پزشکی کرمان محفوظ است. استناد به مطالب یا نقد آن‌ها، با ذکر مأخذ بلامانع است.

نام و نام خانوادگی دانشجو: سارا جوادی

تاریخ و امضا1400/08/11



بسمه تعالی

تاریخ ۹۰/۰۴/۱۰

شماره ۱۰۰۸/۱۰۲۱

پیوست

صورتجلسه دفاع از پایان نامه

دانشگاه علوم پزشکی کرمان

مدیریت تحصیلات تکمیلی دانشگاه

جلسه دفاعیه پایان نامه تحصیلی خانم سارا جوادی دانشجوی دکتری تخصصی (Ph.D) رشته آمار زیستی دانشکده پیرا[illegible] کرمان تحت عنوان "مقایسه عملکرد مدل های افراز بازگشتی و مدل های خاتمی پارامتری در پایه خوه سازی میانگین پی[illegible] معادلات زنجیره ای" به منظور حاتمی مساعدات از دست رفته در پاسخ با ساز و کار گمشدگی تصادفی در حضور کر مدان [illegible] شنبه مورخ ۱۴۰۰/۰۸/۱۱ با حضور اعضای محترم هیات داوران به شرح ذیل:

سمت	نام و نام خانوادگی	امضا
الف-استاد(ان) راهنما	آقای دکتر محمدرضا بشتی آقای دکتر عباس بهرامپور	
ب- استاد(ان) مشاور		
ج- عضو هیات داوران (داخلی)	خانم دکتر مقدمه میرزایی	
ج- عضو هیات داوران (داخلی)	آقای دکتر یونس جهانی	
د-عضو هیات داوران (خارجی)	آقای دکتر سلیمان خیری	
د-عضو هیات داوران (خارجی)	آقای دکتر حسین فلاح زاده	
د-نماینده تحصیلات تکمیلی	سعیده حاجی مقصودی	

تشکیل گردید و ضمن ارزیابی به شرح پیوست با درجه عالی و نمره نوزده و نود و نه صدم ۱۹،۹۹

مهر و امضاء مقام آموزش

۹۰/۰۷/۰۶

فهرست منابع

1. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.

2. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. New England Journal of Medicine. 2012;367(14):1355-60.

3. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338.

4. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1994;43(1):49-73.

5. Allan F, Wishart J. A method of estimating the yield of a missing plot in field experimental work. The Journal of Agricultural Science. 1930;20(3):399-406.

6. Kang H. The prevention and handling of the missing data. Korean journal of anesthesiology. 2013;64(5):402-6.

7. Miri HH, Hassanzadeh J, Khaniki SH, Akrami R, Sirjani EB. Accuracy of Five Multiple Imputation Methods in Estimating Prevalence of Type 2 Diabetes based on STEPS Surveys. Journal of Epidemiology and Global Health. 2020;10(1):36-41.

8. Pigott TD. A review of methods for missing data. Educational research and evaluation. 2001;7(4):353-83.

9. Bartlett JW, Seaman SR, White IR, Carpenter JR, Initiative* AsDN. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical methods in medical research. 2015;24(4):462-87.

10. Schafer JL. Analysis of incomplete multivariate data: Chapman and Hall/CRC; 1997.

11. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. BMC medical research methodology. 2017;17(1):1-10.

12. Breiman L. Friedman; JH; Olshen, RA; Stone, CJ. Classification and regression trees Belmont, California: Wadsworth International Group. 1984.

13. Van der Heijden GJ, Donders ART, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. Journal of clinical epidemiology. 2006;59(10):1102-9.

14. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. The Yale journal of biology and medicine. 2013;86(3):343.

15. Zhang Z. Missing data imputation: focusing on single imputation. Annals of translational medicine. 2016;4(1).

16. Tilling K, Williamson EJ, Spratt M, Sterne JA, Carpenter JR. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. Journal of clinical epidemiology. 2016;80:107-15.

17. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006;169(3):571-84.

18. Chen S-Y, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. Journal of thoracic disease. 2017;9(6):1725.

19. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45(1):1-67.

20. Brown ML, Kros JF. Data mining and the impact of missing data. Industrial Management & Data Systems. 2003.

21. Van Buuren S. Flexible imputation of missing data: CRC press; 2018.

22. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2010:1-68.

23. van Buuren S. MICE: Multiple Imputation by Chained Equations. R Package Version 2.21. 2014.

24. Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. Statistics in Medicine. 2009;28(29):3657-69.

25. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics. 1986;4(1):87-94.

26. Landerman LR, Land KC, Pieper CF. An empirical evaluation of the predictive mean matching method for imputing missing values. Sociological Methods & Research. 1997;26(1):3-33.

27. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research. 2007;16(3):219-42.

28. Rubin DB. Multiple imputation after 18+ years. Journal of the American statistical Association. 1996;91(434):473-89.

29. Moons KG, Donders RA, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. Journal of clinical epidemiology. 2006;59(10):1092-101.

30. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.

31. Bartlett JW, Frost C, Carpenter JR. Multiple imputation models should incorporate the outcome in the model of interest. Brain. 2011;134(11):e189-e.

32. Schenker N, Taylor JM. Partially parametric techniques for multiple imputation. Computational statistics & data analysis. 1996;22(4):425-46.

33. Di Zio M, Guarnera U. Semiparametric predictive mean matching. AStA Advances in Statistical Analysis. 2009;93(2):175-86.

34. Arnold BC, Castillo E, Sarabia JM. Conditionally specified distributions: an introduction (with comments and a rejoinder by the authors). Statistical Science. 2001;16(3):249-74.

35. Moriarity C, Scheuren F. A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics. 2003;21(1):65-73.

36. Durrant GB. Imputation methods for handling item-nonresponse in the social sciences: a methodological review. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute NCRM Methods Review Papers NCRM/002. 2005.

37. Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. Statistics in medicine. 2008;27(1):83-102.

38. Royston P. Multiple imputation of missing values. The Stata Journal. 2004;4(3):227-41.

39. Spiegelhalter DJ, Thomas A, Best N, Lunn D. WinBUGS version 1.4 user manual. MRC Biostatistics Unit, Cambridge URL http://www mrc-bsu cam ac uk/bugs. 2003.

40. Heitjan DF, Little RJ. Multiple imputation for the fatal accident reporting system. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1991;40(1):13-29.

41. Little RJ. Missing-data adjustments in large surveys. Journal of Business & Economic Statistics. 1988;6(3):287-96.

42. Iacus S, Porro G. Invariant and metric free proximities for data matching: An R package. Journal of Statistical Software. 2008;25(1):1-22.

43. Harrell F. Hmisc: a library of miscellaneous S functions. 2008.

44. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing,[on line]. Vienna, Austria. 2014.

45. Slade E, Naylor MG. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. Statistics in Medicine. 2020;39(8):1156-66.

46. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. J Stat Softw. 2011;45(4):1-20.

47. Von Hippel PT. 8. How to impute interactions, squares, and other transformed variables. Sociological methodology. 2009;39(1):265-91.

48. Loh W-Y. Regression tress with unbiased variable selection and interaction detection. Statistica sinica. 2002:361-86.

49. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer series in statistics New York; 2001.

50. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC press; 1984.

51. John Lu Z. The elements of statistical learning: data mining, inference, and prediction. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2010;173(3):693-4.

52. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. American journal of epidemiology. 2010;172(9):1070-6.

53. Breiman L. Classification and regression trees: Routledge; 2017.

54. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.

55. Strobl C, Boulesteix A-L, Augustin T. Unbiased split selection for classification trees based on the Gini index. Computational Statistics & Data Analysis. 2007;52(1):483-501.

56. Loh W-Y. Classification and Regression Trees2011. 14-23 p.

57. Doove LL, Van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis. 2014;72:92-104.

58. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

59. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention science. 2007;8(3):206-13.

60. Nonyane BA, Foulkes AS. Multiple imputation and random forests (MIRF) for unobservable, high-dimensional data. The international journal of biostatistics. 2007;3(1).

61. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC medical research methodology. 2012;12(1):1-13.

62. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology. 2014;179(6):764-74.

63. Garrusi B, Baneshi MR. Eating disorders and their associated risk factors among Iranian population–a community based study. Global journal of health science. 2013;5(1):193.

64. Javadi S, Bahrampour A, Saber MM, Garrusi B, Baneshi MR. Evaluation of Four Multiple Imputation Methods for Handling Missing Binary Outcome Data in the Presence of an Interaction between a Dummy and a Continuous Variable. Journal of Probability and Statistics. 2021;2021.

65. Lewis RJ, editor An Introduction to Classification and Regression Tree (CART) Analysis," presented at Annual Meeting of the Society for Academic Emergency Medicine. Annual Meeting of the Society of Academic Emergency Medicine in; 2000: Citeseer.

66. Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. R package. 2018.

67. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in medicine. 2010;29(28):2920-31.

68. Demirtas H, Freels SA, Yucel RM. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. Journal of Statistical Computation and Simulation. 2008;78(1):69-84.

69. Demirtas H. Simulation driven inferences for multiply imputed longitudinal datasets. Statistica Neerlandica. 2004;58(4):466-82.

70. Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation. 2018;88(15):2909-30.

71. Team RC. R: A language and environment for statistical computing. 2013.

72. Goldstein H, Steele F, Rasbash J, Charlton C. REALCOM: methodology for realistically complex multilevel modelling. Bristol: Centre for Multilevel Modelling, Graduate School of Education, University of Bristol. 2008.

73. Rasbash J, Browne W, Goldstein H, Yang M, Plewis I, Healy M, et al. A user's guide to MLwiN. London: Institute of Education. 2000;286.

74. Levesque R. SPSS programming and data management. A guide for SPSS and SAS Users. 2007.

75. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician. 2007;61(1):79-90.

76. Kenward MG, Carpenter J. Multiple imputation: current perspectives. Statistical methods in medical research. 2007;16(3):199-218.

77. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377-99.

78. Dusseldorp E, Conversano C, Van Os BJ. Combining an additive and tree-based regression model simultaneously: STIMA. Journal of Computational and Graphical Statistics. 2010;19(3):514-30.

79. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of clinical epidemiology. 2019;110:63-73.

80. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine. 1999;18(6):681-94.

81. David M, Little RJ, Samuhel ME, Triest RK. Alternative methods for CPS income imputation. Journal of the American Statistical Association. 1986;81(393):29-41.

82. Van Buuren S, Oudshoorn K. Flexible mutlivariate imputation by MICE: Leiden: TNO; 1999.

83. Koller-Meinfelder F. Analysis of incomplete survey data-multiple imputation via bayesian bootstrap predictive mean matching. 2009.

84. Hsu CH, Taylor JM, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. Statistics in medicine. 2006;25(20):3503-17.

85. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical science. 1992;7(4):457-72.

فهرست مندرجات

فصل اول 1

1-1مقدمه 2

2-1تعاریف و مفاهیم اولیه 3

1-2-1سازوکارهای گمشدگی 3

1-1-2-1گمشده‌های غیر تصادفی 3

2-1-2-1گمشده‌های تصادفی 4

3-1-2-1گمشده‌های کاملا تصادفی 4

3-1بیان مساله و ضرورت موضوع 5

1-3-1تحلیل موارد کامل 5

2-3-1جانهی منفرد 6

1-2-3-1جانهی میانگین 6

1-3-2-2جانهی رگرسیونی 7

1-3-3روش‌های جانهی چندگانه 7

1-3-3-1جانهی چندگانه با معادلات زنجیره‌ای 8

1-3-3-2جانهی چندگانه با معادلات زنجیره‌ای بدون حضور اثر متقابل در مدل جانهی جورسازی میانگین پیش‌بینی (PMM) 10

1-3-3-2-1مفهوم دنر درPMM 11

1-3-3-2-2معیارهای جورسازی برای اندازه گیری فاصله ی دو مقدار درPMM 11

1-3-4روش‌های جانهی در MICE برای هدایت مقادیر گمشده‌ی پاسخ در حضور اثر متقابل 12

1-3-4-1به کارگیری روش PMM در MICE در حضور اثر متقابل 13

1-3-4-2 جانهی چندگانه با معادلات زنجیره‌ای با حضور اثر متقابل در مدل جانهیPMM 13

1-3-4-3روش جانهی چندگانه با معادلات زنجیره‌ای رده‌بندی شده 14

1-3-4-4جانهی چندگانه با معادلات زنجیره‌ای با استفاد از درخت رده‌بندی و رگرسیونی 15

1-3-4-5جانهی چندگانه با معادلات زنجیره‌ای با استفاد از جنگل تصادفی 16

1-4هدف کلی 17

1-5اهداف جزئی 17

1-6هدف کاربردی 17

1-7فرضیات پژوهش 18

فصل دوم 19

2-1بررسی متون 20

2-1-1مقایسه‌ی دو روش پارامتری FCS وJM 20

2-1-2مقایسه‌ی دو روش جانهی پارامتری و روش ناپارامتریMIRF 21

2-1-3معرفی روش CART در MICE و مقایسه‌ی آن با یک روش جانهی چندگانه پارامتری 22

2-1-4مقایسه‌ی سه روش جانهی (نیمه) پارامتریPMM ، مجهول وJAV 24

2-1-5مقایسه‌ی دو روش پارامتری و نیمه پارامتری با دو روش بر پایه‌ی جنگل تصادفی در MICE بر مشاهدات کوهورت بقا 25

2-1-6معرفی روش جانهی جنگل تصادفی در MICE و مقایسه‌ی آن با روش (نیمه) پارامتری PMM و روش ناپارامتری CARTدرMICE 26

12--7مقایسه‌ی روش‌های جانهی پارامتری با و بدون اثر متقابل در مدل جانهی 27

12--8مقایسه‌ی روش نیمه پارامتری) PMM با و بدون اثر متقابل در مدل جانهی) و روش‌های ناپارامتری CARTو RF درMICE 28

فصل سوم 34

3-1داده‌ی واقعی 35

3-2جورسازی میانگین پیش‌بینی در حضور اثر متقابل 36

3-2-1روشMICE-Stratified 36

3-2-2روشMICE-Interaction 37

3-3روش CART درMICE 37

3-3-1الگوریتم به کارگیری روش CART درMICE 38

3-4روش RF درMICE 39

3-4-1الگوریتم به کارگیری روش RFدرMICE 40

3-5معیارهای اعتبار سنجی (پیامدهای ارزیابی) 41

3-5-1اریبی و اریبی نسبی 41

3-5-2احتمال پوشش (پوشش) برای فاصله اطمینان 95٪ 42

3-5-3خطای معیار تجربی 43

3-5-4خطای معیار بر پایه‌ی مدل 43

3-5-5نسبت واریانس (تغییرات) منتسب به مقادیر گمشده(λ) 44

3-5-6معنی‌داری ضرائب 45

3-6طراحی مطالعات شبیه سازی 46

3-6-1طراحی شبیه‌سازی برای ایجاد مقادیر گمشده در داده‌ی واقعی 46

3-6-2تولید داده‌ی شبیه‌سازی بر مبنای داده‌ی واقعی (پاسخ دو حالتی و اثر متقابل بین متغیر دو حالتی و کمی) 47

3-6-3-1شبیه سازی داده با پاسخ دو حالتی 48

3-6-3-2شبیه سازی داده با پاسخ کمی 50

3-7حذف مشاهدات 51

3-8تجزیه و تحلیل 51

3-9توضیحاتی در مورد نرم افزارهای استفاده شده 51

فصل چهارم 53

4-1نتایج شبیه سازی‌ها با پاسخ دو حالتی 54

4-1-1داده‌ی واقعی 55

4-1-2مشاهدات شبیه سازی 57

4-1-2-1نتایج شبیه سازی داده با پاسخ دو حالتی و یک اثر متقابل بین دو متغیر دو حالتی 57

4-1-2-2نتایج شبیه سازی مشاهدات با پاسخ دو حالتی و یک اثر متقابل بین دو متغیر کمی 61

4-1-2-3نتایج شبیه سازی مشاهدات با پاسخ دو حالتی و اثرمتقابل بین یک متغیر دو حالتی و یک متغیر کمی 63

4-2نتایج شبیه سازی‌های با پاسخ کمی 67

4-2-1نتایج شبیه سازی داده با پاسخ کمی و یک اثر متقابل بین دو متغیر دو حالتی 67

2-2-4نتایج شبیه سازی داده با پاسخ کمی و یک اثر متقابل بین دو متغیر کمی 72

3-2-4نتایج تولید یا شبیه سازی داده با پاسخ کمی و یک اثر متقابل بین متغیر کمی و متغیر دو حالتی 74

فصل پنجم 80

1-5بحث 81

1-1-5سناریوهایی که در آن‌ها پاسخ دوحالتی بودند 84

2-1-5سناریوهایی که در آن‌ها پاسخ کمی بودند 87

3-1-5نقاط قوت 89

4-1-5محدودیت‌های مطالعه 90

5-1-5پیشنهادات برای مطالعات آینده 91

2-5نتیجه‌گیری 91

فهرست منابع 93

پیوست 98

(1 توضیحاتی در مورد روش‌های جانهی 99

I روش‌های سنتی جانهی انواع متغیرها99

I .I متغیرهای کمی 99

I .I متغیرهای دو حالتی 100

I I جانهی چندگانه 100

II.Iسه منبع پراکندگی در جانهی چندگانه 100

II. I I چند جانهی لازم است؟ 103

IIIشبیه سازی 104

III.I طرح شبیه‌سازی که تنها شامل سازوکار نمونه‌گیری است. 104

III.IIطرح شبیه‌سازی که شامل ترکیب سازوکارهای نمونه‌گیری و گمشدگی است.105

III.IIIطرح شبیه سازی که تنها شامل سازوکار گمشدگی است. 105

(2نتایج شبیه سازی داده با پاسخ دو حالتی و یک اثر متقابل بین دو متغیر دو حالتی 107

(3 نتایج شبیه سازی مشاهدات با پاسخ دو حالتی و یک اثر متقابل بین دو متغیر کمی 117

(4 نتایج شبیه سازی برمبنای داده‌ی واقعی (با پاسخ دوحالتی و اثرمتقابل بین متغیر دو حالتی و متغیر کمی) 125

(5 نتایج شبیه سازی داده با پاسخ کمی و یک اثر متقابل بین دو متغیر دو حالتی137

(6 نتایج شبیه‌سازی داده با پاسخ کمی و یک اثر متقابل بین دو متغیر کمی 148

(7 نتایج شبیه سازی داده با پاسخ کمی و یک اثر متقابل بین متغیر کمی و متغیر دو حالتی 156

(8نتایج شبیه‌سازی برای ایجاد مقادیر گمشده بر داده‌ی واقعی (پاسخ دو حالتی و اثر متقابل بین یک متغیر دو حالتی و یک متغیر کمی) 165

(9 فهرست کوتاه نوشته‌ها 209