様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　２０　日現在

機関番号：　１４４０１
研究種目：　挑戦的萌芽研究
研究期間：　2013～2015
課題番号：　２５５４００１１

研究課題名（和文）欠損値データ解析の新展開: NMARness and APB

研究課題名（英文）New developments of missing data analysis: NMARness and APB

研究代表者

狩野　裕（Yutaka, Kano）

大阪大学・基礎工学研究科・教授

研究者番号：２０２０１４３６

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究は，研究代表者と研究分担者である岩崎学教授（成蹊大学）を中心とする二つの研究グループによる共同研究として進められた．小規模および中規模の研究集会を各年度数回開催した．研究目的は欠損値データ解析の理論再構築とその応用である．具体的な研究成果としては，条件MAR の数学的緩和，NMARness の定義，APB の導出と理論的性質の検討である．これらの理論を応用し，いくつかの統計モデルにおいて補助変数導入の有効性を検討した．具体的なモデルとして，エンドポイントに欠損がある場合，代替特性を補助変数として取り入れることによって，興味あるパラメータの推定バイアスが減少するための数学的条件が導かれた．

研究成果の概要（英文）：This research project has been completed by the two research groups conducted by Professor Yutaka Kano and Professor Manabu Iwasaki. We have offered research colloquiums several times for each year to advance the research project. The aim of the research project is to re-structure the theory of missing data analysis and to apply them to some statistical models for the analysis with missing data. Results of the project include mathematically weakening the MAR condition, defining NMARness and Approximate population Bias (APB) and studying mathematical properties of the NMARness and APB. Applying these theoretical results, we studied effectiveness of introducing auxiliary variables in several statistical models for the analysis of missing data. One particular result is to derive mathematical conditions under which introducing surrogate endpoints can reduce the bias of the MLE for data with possibly missing data at the endpoint.

研究分野：　統計科学

キーワード：　Missing at random　APB　NMARness　bias of the MLE　sarrogate endpoint

１．研究開始当初の背景

欠損値データの解析においては，欠損がいかなるメカニズムによって生じたか，すなわち，欠損メカニズム(missing-data mechanism)の理解が重要である．このことを指摘したのが1970 年代の Donald Rubin らの一連の研究であった．得られるべき完全にデータ $Y = [Y_{ij}]$ に対して，確率変数 R_{ij} を Y_{ij} が欠損(観測) のとき $R_{ij} = 0(1)$ と定義する．$R = [R_{ij}]$ を欠損指標とよぶ．彼らは R と Y の関係が欠損値データ解析に重要な役割を果たすことを指摘した．Rubin らは，Y の観測部分を Y_{obs}，欠損部分を Y_{mis} と書いた．このとき，$Y = [Y_{obs}; Y_{mis}]$ である．この簡明な記号は，欠損値データ解析手法を適用する応用研究者・実務家への普及を助けたと思われる．しかし，数学的に曖昧なこれらの記号が多くの誤解を生み，理論発展の阻害要因になってきたことは看過できない．$P(R|Y_{obs},Y_{mis}) = P(R|Y_{obs})$ のとき，欠損メカニズムは MAR であると言う．MAR の仮定の下では，欠損メカニズムを指定するという困難な問題を避けて，$f(Y_{obs};\theta)$ にもとづく最尤法を実行することができる(Little and Rubin 2002; 岩崎 2002)．しかし，MAR の条件は相当に制限的で満たされない状況が多い．MAR が満たされないとき，推定にバイアスが生じるが，それがどの程度重大な問題なのか，定量的な評価方法が存在しない．

２．研究の目的

欠損値を含むデータを適切に解析することは相当の困難を伴う．近年はビッグデータや超高次元データの解析需要が増大し，欠損値データ解析の方法論は益々重要になっている．本研究の目的は，欠損値データ解析の理論を再整備しその拡張を行うこと，そして，その理論結果を今まで想定し得なかった新規の統計的推測問題へ応用することである．具体的には，かなり強い条件と言われている MAR(missing at random) を緩和し，そして，MAR が成立しないときの推定量のバイアスを評価する方法論を構築する．これらの理論研究をベースにした応用的研究として，(i) 補助変数の追加が推定量のバイアスを減じるための条件の導出，(ii) ロバスト推測の包括的な理論体系の構築，(iii) 実質科学の応用研究，等を実施する．

３．研究の方法

本研究は，研究代表者と研究分担者である岩崎学教授（成蹊大学）を中心とするいくつかの研究グループによる共同研究として進める．初年度は主に理論研究を進める．具体的には，欠損値データ解析の理論再構築，MAR の拡張，NMARness の定義，Approximate Population Bias(APB) の導出である．二年目と三年目は，理論研究に加えて，いくつかの具体的な統計的推測問題への応用研究を進め，研究の幅を広げる．具体的な応用課題は，補助変数導入の有効性の検討，ロバスト推定法の開発，CAT の研究であるが，これらへの応用研究に留まらず，さらに応用対象を広げていく．実質科学における実データ解析も実施する．

これらの研究を進めるため，小規模および中規模の研究集会を各年に数回開催した．主な講演者は，麻生英樹（産業技術総合研究所，研究員），鈴木　譲（大阪大学，准教授）高井啓二（関西大学，准教授），馬場崇充（九州大学，修士課程）であった．

４．研究成果

(1) 既存の欠損値データ解析の方法論に関してサーベイと理論の再整備を行った．その結果を「NMAR の下での尤度法」として日本統計学会誌に出版した．
(2) 欠損メカニズムを指定せずとも一致推定を可能にするためのより弱い条件，すなわち数学的に緩和された MAR 条件を整理し，それらの数学的・統計学的性質を検討した．
(3) MAR でない程度を量的に表現する NMARness の統計的性質を検討した．それは不偏性が成立しない程度と定義することができる．
(4) 近似理論を用いて定義された Approximate Population Bias(APB)の統計的性質を検討した．
(5) 平成 25 年 9 月に大阪大学で開催された統計関連学会連合大会において「不完全データ解析と潜在変数モデル」なる演題で研究成果を招待講演として発表した．
(6) 平成 25 年 9 月に大阪大学において開催された国際会議「Incomplete Data Analysis and Causal Inference」において「Approximate population bias and nonignorable missingness」なる演題で研究成果を招待講演として発表し意見交換した．
(7) 欠損原因がユニットごとに異なる状況における統計的推測を発展させた．本研究成果を平成 26 年 7 月に台湾で開催された国際会議 IMS-APRM2014 において口頭発表した．
(8) 経時測定データにおいてエンドポイントに欠損がある場合，その代替特性を観測することがある．代替特性導入の有効性についての理論研究を進め，数値実験により結果を補強した．本研究成果を平成 26 年 9 月に東京大学で開催された統計関連学会連合大会（口頭）および，平成 26 年 11 月に京都で開催された国際会議 Kyoto International Conference on Modern Statistics in the 21st Century（ポスター）において発表した．
(9) 二値の結果変数をもつ経時測定データにおける欠損値問題（ドロップアウト）の識別性の数学的構造を明らかにし，識別可能で且つ現実的な統計モデルを複数提

案した．この理論結果を実際のデータの解析に応用した．本研究成果を平成 26 年 11 月にマレーシアで開催された国際会議 ISI Regional Statistics Conference 2014 において口頭発表した．

(10) 欠損率が 90%を超える大量の欠損値を含むデータを因子分析するための新しい方法論を提案し実証した論文が査読を経て学術雑誌 Journal of Statistical Computation and Simulation に出版された．

(11) シンプルなセットアップにおいて，欠測が NMAR のとき，欠測メカニズムを用いない尤度に基づく最尤推定量のバイアスの公式を明示的に導出し，それが減少するための数学的条件を提出した．臨床統計においてエンドポイントのデータが得られないとき（欠測）しばしば代替特性が用いられる．代替特性を補助変数とした最尤法のバイアスを評価することに成功し，バイアスが減少するための十分条件を提出した．本研究成果は平成 27 年度統計関連学会連合大会において 2 件の講演として発表された．

(12) 欠測が NMAR のとき，逆回帰のアイデアに基づく最尤法において，共変量の分布をノンパラメトリックに推定する新たな推定方法を提案し，対応する推定量(MLE)の漸近的性質を検討した．統計的に望ましい性質である MLE の一致性と漸近正規性を証明した．本研究成果は平成 27 年度統計関連学会連合大会で発表された．

(13) 脱落は欠測の代表例である．脱落の伴う経時カテゴリーデータに対する多重代入法の有効性を数値的に検討した．

(14) 統計的因果推論は欠測値問題の応用分野である．統計的因果推論における傾向スコアマッチング後の個体差の評価方法を提案した．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

① Hirose, K., Kim, S., Kano, Y., Imada, M., Yoshida, M. and Matsuo, M. (2015). Full information maximum likelihood estimation in factor analysis with a large number of missing values. Journal of Statistical Computation and Simulation, 86(1), 91-104. (査読有)
DOI: 10.1080/00949655.2014.995656

② 狩野裕(2014). NMAR の下での尤度法. 日本統計学会誌, 13(2), 359-377. (査読有)

③ Hojo, S., Yamamoto, M. and Kano, Y. (2014). Effect of violation of the normal assumption on MI and ML estimators in the analysis of incomplete data. Communications in Statistics - Theory and Methods, 4(15), 3234-3250. (査読有)
DOI: 10.1080/03610926.2013.819920

④ Takai, K. and Kano, Y.(2013). Asymptotic inference with missing data. Communications in Statistics - Theory and Methods, 42(17), 3174-3190. (査読有)
DOI: 10.1080/03610926.2011.621577

⑤ Yoshida, K. and Iwasaki, M. (2013). A new testing procedure for the probability of rare events. Journal of the Faculty of Science and Technology, 50(2), 11-16. (査読有)

〔学会発表〕（計６件）

① Morikawa, K. and Kano, Y.(2015/12). Doubly Robust Estimation under Nonignorable Nonresponse. The 9th Conference of the Asian Regional Section of the IASC, Singapore (Singapore). (招待講演)

② Morikawa, K. and Kano, Y. (2015/11). Doubly Robust Estimation under Nonignorable Nonresponse. AMBN2015. Keio University, Kanagawa, Yokohama.

③ Abe, T., Shiosakai, K., Sano, F., Roberts, R., Sato, Y. and Iwasaki, M. (2015/8). Multiple imputation for longitudinal count data with dropouts; A methodological evaluation. Joint Statistical Meetings. Seattle (USA).

④ Kano, Y. (2015/7). Developments in multivariate missing data analysis. IMPS2015. Beijing (China). (Keynote Lecture).

⑤ Takagi, Y. and Kano, Y.(2014/11). Parameter estimation using auxiliary variables under NMAR. Kyoto International Conference on Modern Statistics in the 21st Century. The Kyoto International Conference Center. Kyoto.

⑥ Kano, Y. (2013/9). Approximate population bias and nonignorable missingness. Incomplete Data Analysis and Causal Inference, Osaka University, Osaka, Toyonaka. (Invited Talk)

〔図書〕（計２件）

① 岩崎 学(2015). 統計的因果推論. pp.204 朝倉書店.

② 高井啓二, 星野崇宏, 野間久史(2016). 欠測データの統計科学. pp.232. 岩波書店.

〔産業財産権〕
○出願状況（計０件）
なし

○取得状況（計１件）
名称：対人感情推定装置，対人感情推定方法及び対人感情推定プログラム
発明者：金順映，今田美幸，吉田学，狩野裕，廣瀬慧
権利者：日本電信電話株式会社，
国立大学法人大阪大学
種類：特許
番号：第５８７５００５号
取得年月日：平成２８年1月２９日
国内外の別： 国内

〔その他〕
なし

６．研究組織
(1)研究代表者
狩野　裕（KANO, Yutaka）
大阪大学・基礎工学研究科・教授
研究者番号：２０２０１４３６

(2)研究分担者
岩崎　学（IWASAKI, Manabu）
成蹊大学・理工学部・教授
研究者番号：４０２５５９４８

(3)連携研究者
高井　啓二（TAKAI, Keiji）
関西大学・商学部・准教授
研究者番号：２０５７２０１９

大津　起夫（OTSU, Tatsuo）
大学入試センター・研究開発部・教授
研究者番号：１０２０３８２９

廣瀬　慧（HIROSE, Kei）
大阪大学・基礎工学研究科・助教
研究者番号：４０６０９８０６

鎌谷　研吾（KAMATANI, Kengo）
大阪大学・基礎工学研究科・講師
研究者番号：００５６９７６７

菊地　賢一（KIKUCHI, Kenichi）
東邦大学・理学部・教授
研究者番号：５０２７０４２６

(4) 海外研究協力者
Michael E. Sobel, Professor
Columbia University

Ke-Hai Yuan, Professor
University of Notre Dame

Ricardo Silva, Lecturer
University College London,

Mortaza Jamshidian, Professor
California State University,
Fullerton

Aapo Hyvarinen, Professor
University of Helsinki

付録　　シンポジウムの詳細

科研費シンポジウム 2013
International Symposium
at Osaka University
Incomplete Data Analysis and
Causal Inference

Dates: September 22-23, 2013
Place: Osaka University, Toyonaka Campus, Graduate School of Engineering Science, Building A, Room A304
Organizers: Yutaka Kano (Osaka U), Shohei Shimizu (Osaka U)
Sponsor: Japan Society for the Promotion of Science (JSPS)

September 22, Sunday, 2013, 13:00-18:40
[English Session]

Chair: E. Kumagai
13:00-13:30 Yutaka Kano (Osaka University), Approximate population bias and nonignorable missingness

13:30-14:20 Yoshio Takane (McGill University, University of Victoria), PCA with Missing Data

14:30-15:20 Kei Hirose (Osaka University), Full Information Maximum Likelihood Estimation in Factor Analysis with a Large Number of Missing Values

Break

Chair: M. Yoshimori
15:50-16:40 Mortaza Jamshidian (California State University, Fullerton), Examining Missing Data Mechanism via Homogeneity of Parameters, Homogeneity of Distributions, and Multivariate Normality

16:50-17:40 Kentaro Fukumoto (Gakushuin University), Blocking Reduces, if not Removes, Attrition Bias

17:40-18:40　　Discussion

September 23, Monday, 2013, 10:00-12:00
[English Session]

Chair: S. Shimizu
10:00-10:50 Ricardo Silva (University College London), On Factors and Residuals: Searching for Latent Structure at Two Levels of Detail

11:00-11:50 Aapo Hyvarinen (University of Helsinki), Determining Causal Direction Between Two Variables Based on non-Gaussianity

September 23, Monday, 2013, 13:00-18:00
[Japanese Session]

座長：狩野　裕
13:00-13:45　清水昌平（大阪大学 産業科学研究所), LiNGAM による因果構造推定：潜在交絡変数がある場合

13:45-14:30 星野崇宏（名古屋大学 経済学研究科), Semiparametric Bayesian Estimation for Marginal Potential Outcome Modeling: Application to Causal Inference

休憩

座長：高井啓二
14:45-15:15 森川耕輔（大阪大学 基礎工学研究科 M2), 脱落のある二値データに対する識別性の問題

15:15-16:00 松山　裕（東京大学 医学系研究科), Semiparametric Estimation of Treatment Effect in a Randomized Clinical Trial with Missing Data

休憩

座長：廣瀬　慧
16:15-17:00 加藤裕一（島根大学 総合理工学研究科), 統計的ラフ集合手法による if-then ルール導出に関する考察とシミュレーション実験－条件属性に欠損値・決定属性値にノイズを含む場合－

17:00-17:45 石岡恒憲（大学入試センター研究開発部), Imputation of Missing Values for Unsupervised or Semi-supervised Data Using the Proximity in Random Forests

17:45-18:00 総合討論

以上