様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ５ 月 １９ 日現在

機関番号：３４５０４
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１２７５６

研究課題名（和文）口頭剖検に基づく死因のベイズ予測分析

研究課題名（英文）Bayesian analysis of cause of death assignment using verbal autopsy data

研究代表者
國濱　剛（KUNIHAMA, Tsuyoshi）

関西学院大学・経済学部・准教授

研究者番号：４０７７９７１６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では，故人の病歴や症状に関する調査データの特徴を取り入れながら，口頭剖検による死因予測を行うための統計手法を新たに開発した．特に，質問のタイプとしてYes・Noを答える２値型だけでなく，カウント型，カテゴリー型，連続型を一つの枠組みの中で同時に表現することで，従来のアプローチに比べて仮定の弱い統計モデルを考案した．加えて，新たな統計手法に対して，実際の調査データを使って死因予測を行うための計算アルゴリズムも提案した．

研究成果の学術的意義や社会的意義
提案手法は口頭剖検のための統計手法として，既存のアプローチに比べて聞き取り調査データの特徴を柔軟に捉えることができるため，発展途上国の死因分布の推定精度の向上や既存分析の妥当性の検証のために役立つと考える．加えて，一般の多変量データ解析のための統計手法としても，提案手法は複雑な相関構造や様々な観測尺度を有する高次元データの分析を目的とするため，部分的に他分野への応用も可能である．例えば，多くの社会調査は多数の質問項目から構成され，項目間の複雑な関係性，頻出する欠損値の存在，異なる観測尺度など，本研究で取り組んだデータの特徴との類似点があると考えられる．

研究成果の概要（英文）：This research develops a new statistical method for prediction of causes of death with verbal autopsy data, taking into account characteristics of surveys about health history and symptoms of the deceased. The proposed statistical model relies on less assumptions than existing methods, by incorporating count, categorical, continuous variables as well as binary ones in the framework. Moreover, for the proposed method, a new algorithm is developed for prediction of causes of death using real survey data.

研究分野：経済統計

キーワード：ベイズ統計学　マルコフ連鎖モンテカルロ法

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

死因，出生率，乳幼児死亡率などの人口学情報は，各国の公衆衛生政策の根幹を成すものであるが，多くの発展途上国では戸籍制度や人口動態統計が完備されていないため，国民の健康状態を把握した上で適切な保健戦略や支援策を適用することが困難な状況にある．死因に関しては，世界全体で 3 分の 2 以上の死は，その原因が特定されていないと言われ，特に貧困に苦しむ地域では保健システムが不完全であるため死因情報が乏しい．例えば，大規模調査を行うことで，各地域における死因情報を多く集めることはできるが，死因を特定するためには医学知識が必要であるため遺族に死因を直接尋ねることはできない．また，医学の専門家を調査員として多数雇うのは費用面から現実的ではない．そこで，実行可能な方法として，家族，保健ケア提供者，コミュニティの構成員に故人の死亡状況，症状，病歴などについて聞き取り調査を行い，その情報に基づいて死因を予測する口頭剖検が広く用いられている．

近年，死因を統計的に予測するアプローチが多く提案されてきたが，遺族への聞き取り調査データには標準的な統計手法では捉えることが難しい特徴が含まれており，予測精度が十分高いとは言えないのが現状である．例えば，遺族への聞き取り調査票には多数の質問項目が含まれていることに加え，項目間に複雑な相関関係が存在する．米ワシントン大学の研究機関である Institute for Health Metrics and Evaluation における Population Health Metrics Research Consortium プロジェクトが提供する口頭剖検データでは 100 以上に及ぶ質問項目が死因との関連付けのために使われている．ところが，既存の統計手法では簡略化のために，調査データが持つ重要な特徴を統計モデルに十分に反映できておらず，大幅な改善の余地が残っている．

２．研究の目的

本研究の目的は，故人の病歴・症状等に関する聞き取り調査データの特徴を十分に考慮に入れた上で，口頭剖検に基づく死因予測を行うための統計手法を新たに開発することである．まず，調査データは複雑な相関構造を持つ高次元多変量データと考えることができるが，実務において広く用いられている標準的なアプローチでは，死因が与えられた下での条件付き独立性という強い仮定が置かれている．これは極めて制約的なものであるため，この仮定に依らずに質問項目の関係性を柔軟に記述できる統計モデルが必要となる．また，別の重要な特徴として，聞き取り調査票は様々な観測尺度を持つ質問項目から構成される点が挙げられる．例えば，特定の病歴や症状の有無を Yes・No で答える質問は２値変数，体重は連続変数，年齢はカウント変数，学歴など３個以上の選択肢がある質問はカテゴリー変数と考えることができる．一般に，複数の異なる観測尺度を持つ多変量データに対して，柔軟な統計モデルを構築することは容易ではなく，口頭剖検の関連分野における既存の統計手法では，すべての質問項目を人為的に２値変数に変換することで混合する観測尺度の問題を避けている．ところが，このような２値変換はデータが持つ情報の一部を失うことに加えて，閾値の設定方法に推定結果が影響を受ける可能性があるため，調査データに含まれる情報すべてを反映させながら安定的な死因予測を行うためにも調査データの混合尺度を保ったまま統計分析を行うことが望ましいと考えられる．

また，死の発生と聞き取り調査の実施に時間的なラグが生じるため，遺族が故人の症状をすべて詳細に覚えているとは限らず，その結果として多くの欠損値が調査データの中に発生する．標準的な統計手法は分析の際に欠損値の存在を許しておらず，何らかの特別な統計処理が別途必要となる．他にも構造上起こりえない組合せの存在も一つの特徴として考えられる．例えば，既婚の幼児や，出産経験のある男性など，構造的に発生することがあり得ない組合せが聞き取り調査票の中に存在する．既存の統計手法において見過ごされているこれらの特徴を取り入れることで，死因の予測分析により適した新たな統計モデルの提案を目指す．

３．研究の方法

質問項目の相関関係について，聞き取り調査票に含まれる質問数が多いことから，その相関構造に何も制約を課さない標準的なアプローチを取ると，利用できるサンプル数と比較して統計モデル内のパラメータ数が多くなり過ぎてデータに過適合してしまい，死因の予測精度が低くなると考えられる．例えば，100 個の質問項目に対しては，その相関の組み合わせは約 5000 個ある一方で，サンプル数は高々数千程度であるため，安定的な予測結果を得るためにはパラメータ数を少なくする必要がある．そこで，各々の相関を直接表現する代わりに，多変量解析において広く用いられている因子分析の考えを応用し，各サンプルに対して低次元な因子を導入する．これらの因子はすべての質問項目の分布に影響を与え，質問項目は因子を共有することで間接的に相関構造を構築する．実際の分析では，データが持つ相関関係の複雑さを反映させながら因子数を調整することで，適切なパラメータ数に基づいて死因予測を行うことができる．

採用する因子モデリングは一般に連続型の多変量データに対する統計手法であり，この手法を多様な混合尺度が存在する聞き取り調査データに応用するために，個々の質問項目に対して因子と紐付けられた潜在的な連続変数を導入し，質問の観測尺度に応じて潜在変数の変数変換を行うアプローチを考案した．具体的には，２値変数に対しては 0 を閾値とする指示関数，カテ

ゴリー変数に対しては多項プロビット関数を用いた変換を考えた．さらに，カウント変数に関しては，log 変換を行なって擬似的に連続変数とみなす方法と，潜在変数とポワソン分布のパラメータを組み合わせる方法を開発した．これにより混合尺度を持ちながら高次元である質問項目に対して，その相関構造をデータの情報を失わずに柔軟に捉えることが可能となる．

聞き取り調査において頻発する欠損データについて，標準的な統計アプローチでは欠損値を何らかの方法で補完する必要があるが，一般に，高次元かつ混合尺度を持つデータに対する欠損値の補完は容易ではなく，採用する補完方法に予測結果が大きく依存する可能性がある．そこで本研究では，missing-at-random という欠損過程に関する標準的かつ現実的な仮定に基づき，死因が与えられた下での質問項目の条件付き分布を構築することで欠損値を補完する必要をなくし，この問題を回避している．また，構造的制約に関しては，制約に関わる変数を組み合わせ，起こり得ないカテゴリーを除去した上で新たな変数として統合することで，制約問題を避ける方法を考案した．

４．研究成果

本研究の主な成果として，頻出する欠損値の存在，変数間の複雑な相関構造，質問項目により異なる観測尺度等，故人の病歴や症状に関する調査データの特徴を柔軟に取り入れながら，口頭剖検による死因予測を行うためのベイズ統計手法を新たに開発したことが挙げられる．特に，因子モデリングを応用した連続型潜在変数を導入し，これを変換することにより，Yes・No を答える２値型だけでなく，カウント型，カテゴリー型，連続型を一つの枠組みの中で同時に表現することを可能とした．加えて，新たに提案した統計手法に対して，マルコフ連鎖モンテカルロ法を用いた効率的なパラメータ推定方法を考案し，実際の聞き取り調査データを使って死因予測を行うための計算アルゴリズムの開発も行った．現在，実務に広く用いられている統計手法は，質問項目の観測尺度とその相関関係に対して強い条件を仮定している一方で，提案手法では聞き取り調査データの特徴を比較的弱い仮定の下で表現しているため，今後実施される発展途上国の死因分布の推定や既存分析の妥当性の検証において活用される可能性がある．また，関連する研究成果について統計学会連合大会や International Society for Bayesian Analysis 等の国内・国際学会において発表を行い，関連論文が応用統計学の国際雑誌に掲載された．

提案手法は口頭剖検による死因予測のために考案されたものであるが，遺族への聞き取り調査データが持つ欠損値や混合尺度等の特徴は，他の社会科学分野の高次元データにおいても時折見られるものであり，本研究で新たに開発した統計モデリングが異なる学術分野のデータ分析において部分的に応用できると考えられる．例えば，公的機関が実施する大規模調査は回答者の個人特定の恐れのある情報を含むためプライバシー保護の観点からそのまま公開することが難しい．その一方で，社会の現状に対する理解を深めるためには様々な研究者・実務家に広く分析されることが望ましい．そこで現実的な解決策として，元の調査データから個人特定が困難な擬似データを発生させ，それを公開することが考えられるが，その際に用いる統計手法が調査データの特徴を柔軟に捉えられるかどうかが重要となる．多くの社会調査は多数の質問項目から構成され，項目間の複雑な関係性，頻出する欠損値の存在，複数の観測尺度など，本研究で取り組んだデータの特徴との類似点があると考えられる．

今後の展望として，混合尺度を持つ質問項目に対して２値変換を行うアプローチの妥当性の検証が考えられる．死因予測を行う既存の統計手法の多くでは，簡潔化のためにすべての質問項目を２値変数に変換して分析を行うが，その妥当性・有効性について実際のデータによる検証はまだ行われていない．この変換によりデータが持つ重要な情報が失われる可能性がある一方で，医療専門家が定めた閾値に基づく２値変換では死因予測に必要な情報は保持されるのかもしれない．本研究において，Population Health Metrics Research Consortium による実際の死因データを用いて提案手法と既存手法との比較分析を行なったところ，現在に至るまで死因分布の推定精度に大きな違いは観測されていない．提案手法の事前分布の設定など，まだ調整すべき点も残っており，比較分析を慎重に進める必要があるが，死因予測の観点からは従来の２値変換でもそれほど多くの情報を失っていない可能性を示唆するのかもしれない．

また，別の展望として，症状の相関関係が年齢に依存して緩やかに変化するような拡張が挙げられる．本研究期間中に関連文献の中で，故人の症状・病歴と年齢とを異質な説明変数として区別し，前者の共分散を年齢の非線形関数として明示的に統計モデル内で表現することにより死因の予測精度を向上させることができるという指摘があった．そこで，この点を踏まえて本研究で考案した統計手法の枠組みを拡張し，年齢や性別等の人口学的情報を症状・病歴に対する共変量として解釈する方向性が考えられる．その際に，因子モデリングを応用した潜在変数の中にどのように年齢を組み込んで，その複雑な相関関係を年齢に依存させるかが重要な問題になると予想される．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Tsuyoshi Kunihama, Zehang Richard Li, Samuel Clark and Tyler McCormick	４．巻 14
２．論文標題 Bayesian factor models for probabilistic cause of death assessment with verbal autopsies	５．発行年 2020年
３．雑誌名 The Annals of Applied Statistics	６．最初と最後の頁 241-256
掲載論文のDOI（デジタルオブジェクト識別子） 10.1214/19-aoas1253	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計2件（うち招待講演　1件／うち国際学会　1件）

１．発表者名 Tsuyoshi Kunihama, Zehang Richard Li, Samuel Clark and Tyler McCormick
２．発表標題 Bayesian Factor Models for Probabilistic Cause of Death Assessment with Verbal Autopsies
３．学会等名 THE 4TH EASTERN ASIA MEETING ON BAYESIAN STATISTICS, EAC-ISBA 2019（招待講演）（国際学会）
４．発表年 2019年

１．発表者名 國濱剛
２．発表標題 Bayesian analysis of verbal autopsy data with high dimensional mixed-scale variables
３．学会等名 統計関連学会連合大会，中央大学
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関