

令和 5 年 6 月 22 日現在

機関番号：12601

研究種目：基盤研究(C)（一般）

研究期間：2019～2022

課題番号：19K11853

研究課題名（和文）経時データに対して有効性の高い回帰診断法の開発

研究課題名（英文）Research on regression diagnostic methods for longitudinal data

研究代表者

倉田 博史（Kurata, Hiroshi）

東京大学・大学院総合文化研究科・教授

研究者番号：50284237

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）：本研究は、経時データに対する回帰モデルの推測に有効な回帰診断法の理論的基礎を明らかにし、併せて新しい診断方法を提案しようとするものである。経時データは観測期間が長い場合、外れ値やデータの欠測、統計手法が前提とする各種の仮定が途中から成立しなくなるなどといった問題が起こり易く、それらを回帰診断によって効率的に検出することの重要性は増している。ところが、既存の回帰診断の理論は必ずしも経時データの特徴を取り入れたものとなっておらず、理論と応用が乖離した状態となっている。本研究では、この間隙を埋めるべく、経時データに対して有効な回帰診断法の理論的基礎を提供することを狙う。

研究成果の学術的意義や社会的意義

本研究は、経時データに対する回帰モデルの推測に有効な回帰診断法の理論的基礎を明らかにし、併せて新しい診断方法を提案しようとするものである。経時データは、官庁統計、医療統計、経済統計など広く社会で蓄積されている一方で、観測期間が長さによって起こる問題、すなわち外れ値やデータの欠測、統計手法が前提とする各種の仮定が途中から成立しなくなるなどが起こり易く、それらを回帰診断によって効率的に検出することの重要性は増している。また、学術的にも、80年代以降のロバスト統計学の発展により、回帰診断的アプローチの理論研究が停滞する傾向にある。本研究はこれらの空白を埋めようとするものである。

研究成果の概要（英文）：This research aims to derive some useful theoretic results for inference of regression models with longitudinal data and to propose some novel diagnostic methods. As is widely recognized, analyzing longitudinal data has various difficulties including outlier problem, missing value problem, violation of assumptions that are indispensable for standard methods, and hence the regression diagnostics that can settle such difficulties have been getting more and more important. However, the existing theory does not fully take these characteristics of longitudinal data into account. The significance of this research lies in this facts.

研究分野：数理統計学

キーワード：経時データ ロバスト統計学 回帰診断 外れ値 不均一分散 正規性

1. 研究開始当初の背景

回帰モデルの推測を行う際、予測値や残差を事後的に検証することによって、自分の行った推定や検定の結果の妥当性について評価を行うことを回帰診断という。統計理論の多くは、データの発生以前すなわち標本空間上の可測関数としての確率変数の性質を考察対象とするものがほとんどであるのに対し、回帰診断は確率変数の実現値をも理論のフレームワークに含める点が独自のであり、それゆえに応用性にも富む理論体系である。しかし、80年代以降のロバスト統計学の進展に伴い、回帰診断の理論研究は現在ほとんどなされていないと言ってよい。

一方、経時データが近年広く利用されるようになっており、その経時データを用いた回帰モデルの推測も盛んにおこなわれているが、それらに対して適切な回帰診断の手法が存在していないというのが研究開始当初の状況であった。

2. 研究の目的

本研究は、経時データに対する回帰モデルの推測に有効な回帰診断法の理論的基礎を明らかにし、併せて新しい診断方法を提案しようとするものである。回帰診断は、回帰モデルを推測する上での前提となる諸仮定(正規性、均一分散性、外れ値が存在しないことなど)が成り立っているとみてよいかをデータによって検証する方法である。経時データは観測期間が長い場合、外れ値やデータの欠測、統計手法が前提とする各種の仮定が途中から成立しなくなるなどといった問題が起こり易く、それらを回帰診断によって効率的に検出することの重要性は増していると言える。この視点に立って、本研究では回帰診断の枠組みを整備し拡張することによって、経時データに有効な診断の理論的基礎と具体的な診断手法の提案することを目的とする。

3. 研究の方法

申請者の過去の研究のうち距離行列の集合の順序構造や逆行列に関する数学的研究を応用することによって、説明変数相互の距離の時点間比較法などについての新たな知見を得ることを狙う。また、経時データにおいては、時点間や個体間で分散が異なると考えられる場合が多い。これらのモデルにおける各種の回帰診断統計量の構造を明らかにし、既存の分散均一性の検定統計量との関係を調べてゆく。さらに、誤差項の空間における外れ値の検出のための理論的基礎についても研究を進める。外れ値は誤差項の空間において発生する場合もあり、これは通常スチューデント化残差によって評価される。しかし、経時データの場合、誤差項の分散共分散行列により一般の構造を仮定する必要があるため、スチューデント化残差の数学的性質はより複雑なものとなり得る。スチューデント化残差の分布は多変量対称分布の研究の一環として幾つか結果が得られており、本研究では、それらの研究を経時データの回帰診断の文脈で整理・拡張することを目指す。特に、時点情報も利用したカットオフポイントとしてより効率的なものを導出する。そしてそれらを利用して、Cookの距離やDFBETAS統計量などの改良を行う。また、経時的均一分散性の検定あるいは診断の問題：経時データにおいては、時点間や個体間で分散が異なると考えられる場合が多い。これらのモデルにおける各種の回帰診断統計量の構造を明らかにし、既存の分散均一性の検定統計量との関係を調べてゆく。経時データモデルにおける多重共線性の影響についても研究対象とする。多重共線性の研究は多数あるが、経時データモデルにおいてどのように作用するかについては詳しく調べられていない。特にランダム効果を含むモデルにおける、効果の有無の検定における影響は未知である。これについて研究する。

4. 研究成果

令和元年度は、説明変数の空間における外れ値の検出のための理論基礎を考察することを通して、より効率的でかつ既存のロバスト統計学とも整合的な回帰診断の手法を得ることを狙って進めた。説明変数の空間において平均値からマハラノビス距離の意味で離れた観測値は推定値に大きな影響を与え得る。このことは、誤差項の分散共分散行列が単位行列の定数倍である場合については詳しく調べられているが、分散共分散行列がより一般の構造を持つ場合については余り多くは知られていない。そこで、元年度は、最も単純な場合として、2時点において分散が未知でかつ不均一であるような経時データのモデルを考え、

分散の推定値を代入して得られるランダムなマハラノビス距離に基づく説明変数の外れ度の評価について考察した。この評価において、分散の推定誤差の影響を如何に取り除くかが問題であり、何らかの上下限に基づく評価方法を模索中である。関連する研究として SUR(seemingly unrelated regression)モデルにおける推定問題や距離行列の性質に関する結果も得ている。

2年目である令和2年度は、上記の評価の問題を、分散の推定誤差の影響を如何に取り除くかという観点から継続し、加えて、誤差項の空間における外れ値の検出のための理論的基礎の問題についても考察を始めた。外れ値は誤差項の空間において発生する場合もあり、これは通常スチューデント化残差によって評価される。しかし、経時データの場合、誤差項の分散共分散行列により一般の構造を仮定する必要があるため、本研究では、それらの研究を経時データの回帰診断の文脈で整理・拡張すべく、まず通常の推定方法と結果的に一致するような分散構造の範囲の導出問題について考察するところから始めている。ここでも、関連する研究として SUR(seemingly unrelated regression)モデルにおける推定問題や距離行列の性質に関する結果も得ている。

3年度は、誤差項の空間における外れ値の検出のための理論的基礎について考察した。外れ値は誤差項の空間において発生する場合もあり、これは通常スチューデント化残差によって評価される。しかし、経時データの場合、誤差項の分散共分散行列により一般の構造を仮定する必要があるため、スチューデント化残差の数学的性質はより複雑なものとなり得る。スチューデント化残差の分布は多変量対称分布の研究の一環として幾つか結果が得られており、本研究では、それらの研究を経時データの回帰診断の文脈で整理・拡張することを試みた。また、経時データモデルにおける多重共線性の影響についての研究が僅少であることに気付き、研究へ如何に取り込んでゆくかについて考察した。

最終年度である本年度は経時的均一分散性の検定あるいは診断の問題として、これらのモデルにおける各種の回帰診断 統計量の構造特に既存の分散均一性の検定統計量との関係を調べた。多重共線性の影響についても考察した。これらについては紀要などで発表すべく準備中である。

5. 主な発表論文等

〔雑誌論文〕 計7件（うち査読付論文 6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 Shun Matsuura and Hiroshi Kurata	4. 巻 1
2. 論文標題 Statistical estimation of quantization for probability distributions: best equivariant estimator of principal points	5. 発行年 2022年
3. 雑誌名 Machine learning, optimization, and data science	6. 最初と最後の頁 430-441
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/978-3-030-95467-3_31	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 -
1. 著者名 Shun Matsuura and Hiroshi Kurata	4. 巻 63
2. 論文標題 Optimal estimator under risk matrix in a seemingly unrelated regression model and its generalized least squares expression	5. 発行年 2022年
3. 雑誌名 Statistical Papers	6. 最初と最後の頁 13-141
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s00362-021-01232-5	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 -
1. 著者名 倉田博史	4. 巻 -
2. 論文標題 エグゼクティブのための統計学(全6回連載)	5. 発行年 2021年
3. 雑誌名 先端教育	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 無
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 -
1. 著者名 Shun Matsuura and Hiroshi Kurata	4. 巻 -
2. 論文標題 Optimal estimator under risk matrix in a seemingly unrelated regression model and its generalized least squares expression	5. 発行年 2021年
3. 雑誌名 Statistical Papers	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s00362-021-01232-5	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 -

1. 著者名 Ravindra B. Bapat and Hiroshi Kurata	4. 巻 562
2. 論文標題 On Cartesian product of Euclidean distance matrices	5. 発行年 2019年
3. 雑誌名 Linear Algebra and its Applications	6. 最初と最後の頁 135-153
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.laa.2018.10.008	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

1. 著者名 Shun Matsuura and Hiroshi Kurata	4. 巻 29
2. 論文標題 Covariance matrix estimation in a seemingly unrelated regression model under Stein's loss	5. 発行年 2020年
3. 雑誌名 Statistical Methods and Applications	6. 最初と最後の頁 79-99
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s10260-019-00473-x	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 -

1. 著者名 Koji Tsukuda and Hiroshi Kurata	4. 巻 61
2. 論文標題 Covariance structure associated with an equality between two general ridge estimators	5. 発行年 2020年
3. 雑誌名 Statistical Papers	6. 最初と最後の頁 1069-1084
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s00362-017-0975-8	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 -

〔学会発表〕 計4件 (うち招待講演 0件 / うち国際学会 2件)

1. 発表者名 Shun Matsuura and Hiroshi Kurata
2. 発表標題 Statistical estimation of quantization for probability distributions: Best equivariant estimator of principal points
3. 学会等名 7th International Conference on Machine Learning, Optimization, and Data Science (Grasmere, United Kingdom, オンライン参加) (国際学会)
4. 発表年 2021年

1. 発表者名 Shun Matsuura and Hiroshi Kurata
2. 発表標題 Optimal covariance matrix estimators in a seemingly unrelated regression model
3. 学会等名 63rd World Statistics Congress (オンライン開催 (国際学会))
4. 発表年 2021年

1. 発表者名 松浦 峻、倉田 博史
2. 発表標題 非対称な1次および2次損失関数の下でのprincipal pointsの推定について
3. 学会等名 統計関連連合大会
4. 発表年 2020年

1. 発表者名 松浦峻・倉田博史
2. 発表標題 SURモデルにおける偏回帰係数ベクトルのリスク行列の下での最適な推定量について
3. 学会等名 統計関連連合大会
4. 発表年 2019年

〔図書〕 計3件

1. 著者名 倉田博史 (監修)	4. 発行年 2021年
2. 出版社 ニュートンプレス	5. 総ページ数 304
3. 書名 文系のためのめっちゃやさしい確率	

1. 著者名 倉田博史	4. 発行年 2021年
2. 出版社 ニュートンプレス	5. 総ページ数 304
3. 書名 文系のためのめっちゃやさしい 統計	

1. 著者名 倉田博史	4. 発行年 2019年
2. 出版社 KADOKAWA	5. 総ページ数 95
3. 書名 図解大学4年間の統計学が10時間でざっと学べる	

〔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
---------	---------