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　６ 年　 ６ 月 １０ 日現在

機関番号：３３９１４
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ０１４３１

研究課題名（和文）回帰モデルにおける重共線性分析と変数・モデル選択法

研究課題名（英文）Multicollinearity Analysis and Variable/Model Selection in Regression

研究代表者
刈屋　武昭（Kariya, Takeaki）

名古屋商科大学・マネジメント研究科・教授

研究者番号：７００９２６２４

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：この研究では、重共線性を持つ伝統的回帰分析モデルｙ＝Xb＋uにおいて、各個別最小2乗推定値の非効率性リスク測度Iとモデルの不安定性（重共線性）リスク測度Cの2つを特定し、すべての説明変数に対して一様にそれらを制御する有効な回帰モデルの集合Hを導出する方法を構築する。非負の（I,C）リスクは、モデル比較に半順序を与え、統計的決定論の枠組みを与える。I=C=0の必要十分条件は、Xの列ベクトルが互いに直交することである。Hの中からのモデル選択はAIC などを利用する。
Xから有効な回帰モデルの集合を導出するアルゴリズムとして、変数増加法と、主成分分析を利用する変数減少法の2つを開発する。

研究成果の学術的意義や社会的意義
重共線性のもとでの有効な回帰モデル選択問題の研究課題の核心をなす学術的「問い」は、「現在の学術的状況では回帰分析の大きな狙いである因果実証性の検証可能性、実証的科学性をどこまで方法的に担保できるのか」、という問いであると考える。その意味で、この研究はこれまでの状況を異なる新しい代替的方法で大きく改善をしたことと考える。回帰分析の一つの教科書的な基礎を与えるものとなると考える。

研究成果の概要（英文）：　In traditional regression model y=Xb +u with collinearity, based on X only, this research first defines the inefficiency risk measure I and collinearity (instability) measure C of each individual OLSE and a model is defined to be effective if I<c and C<d are satisfied uniformly controlled for given (c,d). Then we develop a model selection process (MSP) of finding a class H of effective sub-models. The risk measure (I, C) gives a partial ordering on the set H and so it also gives a decision-theoretic framework in comparing models with such concept of inadmissibility. It is shown that I = C =0 hold if and only if the columns of X are mutually orthogonal. Once the class H is obtained, an optimal model is obtained by applying such model selection criteria as AIC.
　To get H, two algorithms are proposed: variable-increasing method and variable-decreasing method with principal component analysis.

研究分野：統計学

キーワード：linear regression model　model selection process　collinearity　OLS　effective modeling　VIF　principal component　decision theory

様　式　C－１９、F－１９－１（共通）

1. 研究開始当初の背景、研究の方法

⑴ 伝統的な回帰分析における重共線性の問題は、モデル選択、モデルの有効性･安定性、推定値の効率性、t-検定の有効性など、有効な線形回帰モデルに大きな影響を持つ実証分析の基本問題であるが、統計学界･計量経済学会において、その問題は依然として十分な解決を得ていない周知の重要問題である。この問題に対して、2019 年、2020 年に、日本統計学界で t-値の分解による重共線性とモデル選択へのアプローチを発表。

⑵　文献上の背景としては、重共線性の問題は、説明変数データの作る回帰行列 X は多くの場合所与であるので、X の列ベクトルの近似的1次従属関係が有効なモデル構築を阻む問題である。その統計上の問題の基本は、最小2乗推定値 $\hat{\beta}=(X'X)^{-1}X'y$ の逆行列 $(X'X)^{-1}$ が、$\det X'X=0$ の近傍で不安定となる問題として始まる。この問題への対応法としては、リッジ回帰推定値や、$X'X$ の固有値のばらつきを調整するもの、非線形縮小推定値、主成分回帰、LASSO 推定法の利用など、特定化されたモデルにおける推定値全体を安定化させる対応法がある。また変数選択では分散拡大係数 VIF の利用、などが基本であるが、これらの方法を混合した対応法を提案する論文など極めて多く、また多様である。しかし、これらは、モデル選択や変数選択の枠組みになじまず、推定量にだけ注目している。

経済モデル分析等では、モデルの定式化、説明変数の数とその変数形は事前に確定せず、経済理論･仮説等に基づいて一定の説明変数の集合(含ダミー変数)から、主として t 値や補正決定係数、符号条件等により変数を増減しながら、有効な回帰モデルを実証的に特定していくプロセスをとる。それゆえ、上記の推定値安定化法はこのモデル特定化のプロセスには不適である。一方、第k推定量の分散(安定性)は、誤差項に標準的仮定をおくと、

$$Var(\hat{\beta}_k^*)=\frac{\sigma^2}{EEF_k^2}=\frac{\sigma^2}{Ns_{xk}^2}\times VIF_k$$

となるので、推定量の分散を基準とする重共線性の影響度は、分散拡大係数 VIF で指数化できる。しかし、変数をひとつずつ増減していくプロセスでは、重共線性は変数全体に関わっているので、VIF の大きさの順序が変数入替の順序とはならない。それゆえ、個別変数ごとでなく、変数の組合せごとに判断することが必要となろう。

⑶　申請当時での具体的な目的は、最終的に得られる実証回帰分析モデルの有効性･信頼性･安定性を保証できる包括的な研究開発であることとした。そのため、個別係数の最小 2 乗推定量の分散拡大要因 (VIF)と t-値の関係を利用し、有効で安定的なモデル選択を確保するアルゴリズムの構築を狙いとした。その独自性と創造性は、「***t* 値**の分解式から重共線性に対する ***t* 値**の信頼率を定義し、その信頼率がすべての変数に対して一定以上となる「説明変数の組」のモデルの集合を選択する」、という重共線性問題への新しいアプローチを提案。

ここでは、有効な回帰モデルの集合を最初に補足するアプローチであることを認識していた。

⑷　そして、研究計画書に書いた重共線性指標としての VIF を利用し、すべての t-値の信頼率が一定以上になるモデル･変数選択する研究開発の方法としては次の方法を計画していると書いた。

①「重共線性の測度としての t 値の信頼率をもとに変数･モデル選択法」を扱う論文はない。
②そしてそのアイデアを実現するうえで、「事前に指定した0＜δ＜1に対して、説明変数の数pを持つモデルで、すべてのt 値が信頼率δ以上になる変数の組の集合 Q(p, δ)を識別する」、というアルゴリズムの構築方法を考察している論文はない、と判断している(ｸﾞｰｸﾞﾙ検索結果)。

本研究では、「これを説明変数の番号k=2,･･･,K の順に構築することで、計算量を減らし、モデルの重共線性の各変数のt 値への影響をδ以下にし、変数の組に対する信頼率をδ以上にするだけでなく、t 値の大きさの条件、符号条件を課し、補正決定係数等による説明力を担保する、変数･モデル選択アルゴリズム」を構築することを狙う。これら全体は創造性であろう。なお、重共線性へ測度としてt 値の信頼率を利用する議論については、2020 の日本統計学会で発表した。

(5) 共同研究者林高樹の役割は、当初の上記の視点から有効な実証分析を可能にする有効な R によるアルゴリズムの開発と実際のデータでの実証分析を行うことであった。

2. 研究目的

科研費申請書にもストレートに書いたのだが、「重共線性のもとでの有効な回帰モデル選択」 の『研究課題の核心をなす学術的「問い」は、「現在の学術的状況では回帰分析の大きな狙いである因果実証性の検証可能性、実証的科学性をどこまで方法的に担保できるのか」、という問いであり、この状況を異なる新しい代替的方法で少しでも改善をしたい』ことが、その最終的な研究の目的である。

3. 研究の方法

共同研究者(林高樹、慶應義塾大学)や協力者(倉田博史、東京大学 2022 年 1 月ー23 年 2 月まで参加)とのコミュニケーションはズームを利用し、3 か月に 2 回ほどのペースで議論した。

まず、定数項を含む線形回帰モデルを

(1.1) $$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u},\ E(\boldsymbol{u}) = \boldsymbol{0},\ \ Var(\boldsymbol{u}) = \sigma^2\boldsymbol{I},\ \boldsymbol{X}: N \times K$$

とする。このモデルにおいて、上に述べた研究目的の視点から、有効な回帰モデルの概念を明示的に定義し、実証回帰分析プロセス全体を見直し、モデル概念の再構成、重共線性のモデルとの関係、事後的な診断的分析の考え方の問題点などを整理した。研究費補助を受けた 2021 年 4 月からこのプロセスの理論的枠組みを考察していく過程で、被説明変数yは説明変数行列Xの重共線性を識別する能力がないことに気が付き、次の伝統的な分析プロセス

(1.1) のモデルの指定ーyによるモデルの推定ー変数モデル選択ー事後診断

では重共線性の扱い方は正しくないと判断した。したがって、いわゆるモデルが X の列ベクトルからなす部分行列の集合全体になることから、

モデルの集合全体{Xα}ー有効なサブモデル集合Hの選択(効率性･安定性条件を課す)ーyを利用した AIC,最尤法等による最適モデルの選択ー有効性の確認

というモデル選択プロセスに変更する方法を提案した。最終的に求めるモデルが実証的に有効な回帰モデルであることを最初から、有効性の基準を満たさないモデルを排除しておいて、

残りのモデルのクラスHから、通常のモデル推定プロセスにより最適モデルを選択する。したがって研究の基本的視点と制御対象とするモデル構造を次のように変更した。まず、有効性を毀損するK個の説明変数行列Xの構造にかかる2つのリスクとして、

1) 不安定性リスクとしてのマルチコ(Collinearity)リスク(C-リスク)

2) 最小2乗推定値$\widehat{\boldsymbol{\beta}}$の非効率性(Inefficiency)リスク(I-リスク)

を選択し、これらのリスク指標を制御する問題を設定した。ここで C-リスクは個別変数の VIF である。具体的には、各説明変数 k の N 個の平均予測分散

$$IndPSV_k \equiv \sum_{n=1}^{N} Var\left(x_{nk}\hat{\beta}_k\right)/N = \sigma^2 \frac{1}{N}[I_k \times C_k]$$

$$I_k \equiv \bar{x}_k^2/s_k^2 \quad , \quad C_k \equiv VIF_k = 1/(1-\check{R}_k^2) \geq 0$$

において、与えられた(c, d)に対して $I_k \leq c$、 $C_k \leq d$ がすべてのkについて一様に成立するように変数の組のモデルの集合をアルゴリズムによって選択する。このアイデアに気が付くまでに半年以上かかったが、その素案ができていたうえで、2022年1月より参加した協力者を含めて、内容を進化させていった。22 年の秋には当時の論文を投稿し、審査員からのコメントを利用する一方、ミネソタ大学の統計学スクール(School of Statistics)設立 50 周年記念大会で招待講演をしたり、東北大学でのコンファランスで発表するなどしたりして、概念や表現、内容を深化させていった。

一方、このモデル制御の目的を実行する方法として、1 節で述べたアルゴリズムと同様な考え方による変数増加的な方法と、上のモデル選択における制御問題を同時的に扱う、X の主成分分析による結果をモデル選択に利用する方法を発展させていった。

4. 研究成果

研究費を申請した内容に関しての成果は、次の3つである。

(1) T. Kariya, H. Kurata and T. Hayashi. A Modelling Framework for Regression with Collinearity. Journal of Statistical Planning and Inference. 228 (2024) 95-115 (Open access publication)

(2) T. Kariya. Ineffectiveness of Model Selection via t-Test in Regression with Collinearity. Currently submitted (2024 日本統計学会で発表、当時のタイトルとは少し異なる).

(3) 林高樹、刈屋武昭、倉田博史(2024)「線形回帰分析における 多重共線性を考慮した変数選択法の実装」未出版 (日本 OR 学会で報告)

(1)の論文の成果。

1) $IndPSV_k$がすべてのkに対して最小値σ^2/N をとるための必要十分条件は、Xの列ベクトルが互いに直行することであると、証明している。

2) 2 つのリスクのペア(I_k, C_k)は、説明変数の数が等しいモデルの集合ごとに、その上に部分順位を与えるので、モデルの比較可能性を与える。したがって、モデルの許容性(Admissibility)など、統計的決定理論の視点からのモデル比較の基盤を与えている。

3) 実際に与えられたモデル X から、すべてのkに対して、説明変数が $I_k \leq c$、 $C_k \leq d$ を満たすモデルの集合 H を具体的に特定する2つのアルゴリズムを開発した。一つは、変数増加法、もう一つは、主成分分析法を利用した変数減少的なアルゴリズムで、各モデルの候補の中の変数の組が内部的に相関が一定以

上になる変数の組み合わせをしないように制御する。

4）得られたモデルの集合から、被説明変数 y を用いて、AIC や決定係数などから最適なモデルを選択するのは、これまでのモデル選択のプロセスのとおりである。

5）Regression by Example (2012) の本からの実際のデータに、上のアルゴリズムを適用することで、効率性・有効性の条件をみたすモデルの集合を導出し、実証分析をしている。加えて、レフェリーの提案により、シミュレーション分析も行っている。

(2) の単著論文は、未出版であり、現在投稿準備中である。

上記と同じ伝統的な線形回帰モデルにおいて、誤差項が近似的に正規分布に従うという仮定の下に、OLSE の t-値に基づいて変数を選択しながら仮説検定を繰り返して、モデル選択を行う方法では、1) 同じモデル中で、有意水準、検出力の不確定性、 2) 検定対象とする変数の順序に結果は依存、の問題点が指摘されている。本稿では、重共線性構造がある場合、その影響を考慮しながら次の結果を得ている。

1）この問題を包括的・構造的に分析できる枠組みを構築し、t-統計量、決定係数、検出力等、を扱う統計量の分析基盤を作る。重共線性の t-値への影響は、VIF の逆数の平方根であるので、個別の t-検定が有効となる条件を求める。

2）特に、片側検定に比べて、両側検定の場合、帰無仮説の近傍では重共線性の影響が強く出るため、t-値の検出力の低下が大きく出ることをしめす。

3）任意の 2 つの t-統計量は、説明変数の一つが連続系の場合、(確率 1 で) 相関を持つことを示す。このことは、変数選択で有意でない t-値が小さいものから変数を落とす方式には妥当性がないことを意味する。

4）相関を持つ 2 つの t-統計量 (t1, t2) の平均値の恒等性に対する仮設検定問題を考察し、一様最強力不変 (両側・片側) 検定方式を導く。

5）t-値による変数選択を通して、モデル選択をすることは、予備検定推定量を最終的に選択することになるので、最終的なモデルは非線形推定量を持つバイアス推定量を選ぶことになる。

(3) 共同研究者の林は、統計言語 R により実装した (1) の論文が2つのモデル選択のアルゴリズムに対応するソフトを開発した。この論文では、変数増加法によるアルゴリズムと主成分分析を利用するアルゴリズムを解説し、利用可能なソフトウェアとして提供する結果を記述.

(4) 賃貸・住宅価格回帰分析の情報収集のために、*科研費を利用して*、22 年 8 月に東京で開催された American Real Estate and Urban Economics Association と Asian Real Estate Society, と Japanese Association of Real Estate Financial Engineering (日本不動産金融工学) の共同会議に参加した。この機会にまとめた論文を発表し、それを Asian Real Estate Society の機関紙に投稿し出版。Takeaki Kariya, Hideyuki Takada, Yoshiro Yamamura、Tenant Portfolio Selection for Managing a Shopping Center、International Real Estate Review 26、2023、143-171. 科研のテーマに直接関係ないが、科研費への Acknowledgement を載せてある。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Takeaki Kariya, Hiroshi Kurata, Takaki HayashiA	４．巻 228
２．論文標題 A modelling framework for regression with collinearity	５．発行年 2024年
３．雑誌名 Journal of Statistical Planning and Inference	６．最初と最後の頁 95-115
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jspi.2023.07.001	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Takeaki Kariya, Hideyuki Takada, Yoshiro Yamamura	４．巻 26
２．論文標題 Tenant portfolio Selection for managing a shopping center	５．発行年 2023年
３．雑誌名 International Real Estate Review	６．最初と最後の頁 143-171
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 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　1件）

１．発表者名 刈屋武昭
２．発表標題 A Modelling Framework for Regression with Collinearity
３．学会等名 Conference for School of Statistic’s 50th Anniversary Celebration University of Minnesota（国際学会）
４．発表年 2022年

１．発表者名 刈屋武昭
２．発表標題 A Modelling Framework for Regression with Collinearity
３．学会等名 日本統計学会
４．発表年 2022年

１．発表者名 刈屋武昭
２．発表標題 A Modelling Framework for Regression with Collinearity
３．学会等名 Risk and Statistics, 3rd Tohoku-ISM-UUlm Joint Workshop
４．発表年 2022年

１．発表者名 林高樹
２．発表標題 線形回帰分析における多重共線性を考慮した変数選択法の実装
３．学会等名 日本オペレーションズ・リサーチ学会
４．発表年 2024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
研究分担者	林　高樹 (Hayashi Takaki) (80420826)	慶應義塾大学・経営管理研究科（日吉）・教授 (32612)	

	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
研究協力者	倉田　博史 (Kurata Hiroshi)	東京大学・国際社会科学専攻・教授	

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