学位論文要旨

Model Selection Criteria in Generalized Linear Models and their Extensions

(一般化線形モデル及びその拡張におけるモデル選択規準)

氏名 伊森 晋平

統計解析において, 候補のモデルの集合から最も良いモデルを選択することは, 事象の予測や要因等を特定するための重要な役割の一端を担っている. 本研究では, Nelder and Wedderburn (1972, *Journal of Royal Statistical Society, series A*) により提案され, 実データ解析において有用なモデルのクラスとして知られる一般化線形モデル (Generalized Linear Models; GLM) 及びその拡張モデルに対して, 「規準量のバイアス補正」と「真のモデルの選択」という 2 つの異なる観点から, 3 つのモデル選択規準量を提案している.

本論文の前半は「規準量のバイアス補正」について議論している. モデルの良さはリスク関数によって定義され, 例えば期待 Kullback-Leibler 情報量 (Kullback and Leibler, 1954, *Annals of Mathematical Statistics*) を基に構築されたリスク関数がモデル選択に用いられる. このリスク関数は未知パラメータを含むため, 実解析での使用には推定が不可欠である. Akaike (1973, In *Second International Symposium on Information Theory*; 1974, *IEEE Trans. Automatic Control*) は, リスク関数を, "-2 × 最大対数尤度" で推定した際のバイアスを "2 × パラメータ数" で補正することによって, 赤池情報量規準 (Akaike Information Criterion; AIC) を導出した. AIC はリスク関数の漸近不偏な推定量であり, そのバイアスのオーダーはサンプル数 n に対し, $O(n^{-1})$ であることが知られている. AIC は最も有名なモデル選択規準の一つであり, その定義の簡便さによって広く用いられ, AIC に関連する研究も豊富に行われている.

しかしながら, サンプル数が十分に大きくない場合には, そのバイアスが無視できずモデル選択結果に悪影響を及ぼすことがある. このような規準量のバイアスによる問題点の解決方策として, バイアス補正されたモデル選択規準を用いることはしばしば有効である. 実際に, バイアス補正された AIC(Corrected AIC; CAIC) はこれまでにいくつかのモデルで提案されており, その有用性が確認されている. しかしながら, CAIC の導出は各モデルに依存しており, さらにその導出は容易ではないため, CAIC が既に導出されているモデル以外で CAIC を用いることは難しく, 不便であった.

そこで本研究では, GLM における CAIC を計算するためのシンプルな公式を導出した. GLM は指数型分布族とリンク関数によって構成され, これらを変化させることで多くのモデルを表現することが可能である. そのため, GLM における CAIC の公式は, 実データ解析において広く応用・適用することが出来る. さらに GLM の一部であるロジスティック回帰モデルを, 複数のカテゴリーを目的変数として持つ多変量データに拡張した, 多項ロジスティック回帰モデルにおいても CAIC を導出している. これらの CAIC のバイアスのオーダーは $O(n^{-2})$ であり, シミュレーション及び実データ解析例を通して, 通常の AIC に比べてモデル選択結果の改善が確認された. GLM に

おける CAIC と多項ロジスティック回帰モデルにおける CAIC の導出結果は, Imori, Wakaki and Yanagihara (2014, *Scandinavian Journal of Statistics*) 及び Yanagihara, Kamo, Imori and Satoh (2012, *Linear Algebra and its Applications*) において公表されている.

次に本論文の後半では, Liang and Zeger (1986, *Biometrika*) によって提案された, 回帰パラメータの推定方程式である, 一般化推定方程式 (Generalized Estimating Equations; GEE) の枠組みにおけるモデル選択問題を考えている. GEE の枠組みでは, 目的変数の周辺分布に GLM を仮定しており, さらに作業相関行列を与えることによって同一個体内の相関を考慮している. そのため, GEE は GLM におけるスコア方程式の経時データ等の相関を持つデータに対する拡張と捉えることが出来る. この方程式の解として得られる GEE 推定量は適当な条件の下で一致性や漸近正規性等の有用な漸近性質を持つことが知られており, さらにこの性質は, 例え作業相関行列が真の相関行列と異なっていても保たれるため, 経時データ解析において GEE はしばしば用いられている. 実際に GEE を用いて解析する場合には, まず作業相関構造を仮定し, 次に作業相関構造を構成する相関パラメータをデータから推定することで作業相関行列を決定する. しかしながら, 真の相関行列によっては, 仮定した作業相関構造ではその推定量が構築できない場合があり, このような場合には GEE 推定量の漸近性質が失われる. 作業相関構造に独立性, つまり単位行列を仮定すれば, 相関パラメータの推定が必要なくなるためこのような問題を避けることが出来る. しかしこの状況では, 作業相関行列と真の相関行列が一致している場合に比べ, 回帰パラメータの推定量の効率が下がることがある. 一方で, 作業相関構造に具体的な構造を仮定しないことで真の作業相関の一致推定量を定めることは可能だが, サンプル数が経時測定の時点数に比べて十分に大きくないと, この推定量は安定しない.

このような GEE 特有の問題を解決するために, 作業相関構造の選択問題はこれまでにしばしば議論されており, 実際に選択規準量もいくつか提案されている. しかしながら, これらの先行研究では, 規準量の性質はシミュレーションを通してのみ確認され, それらの理論的性質の導出にはあまり関心が向けられていなかった. そこで, 本研究では, 真の作業相関構造を選択するための規準量を提案し, その選択確率が 1 に収束するための条件を導出する.

Nishii (1984, *Biometrika*) は, AIC のバイアス補正項を一般化した規準量, Generalized Information Criterion (GIC) を考え, 説明変数の選択に対して真の説明変数の組み合わせを選択する確率が 1 に収束するための条件を示している. 本研究では, この結果を GEE における作業相関構造の選択に拡張することを試みる. しかし, GEE は周辺分布の仮定のみで定義されているため, 尤度を用いる GIC を直接適用することができない. そこで, 本研究では行列同士の距離を測るための関数として, 尤度の代わりに Stein のロス関数 (James and Stein, 1961, In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*) を基にした規準量を構成した. 回帰パラメータが $\sqrt{n}-$consistency を持つとき, 適当な緩い正則条件を仮定することで, この規準量による真の相関構造の選択確率が 1 に収束する. この結果は, Imori (2014, *Hiroshima Mathematical Journal*, to appear) で示されている.