ロジットモデルを用いた複数企業データベースの結合方法

総合研究大学院大学　高部　　勲
統計数理研究所教授　山下　智志

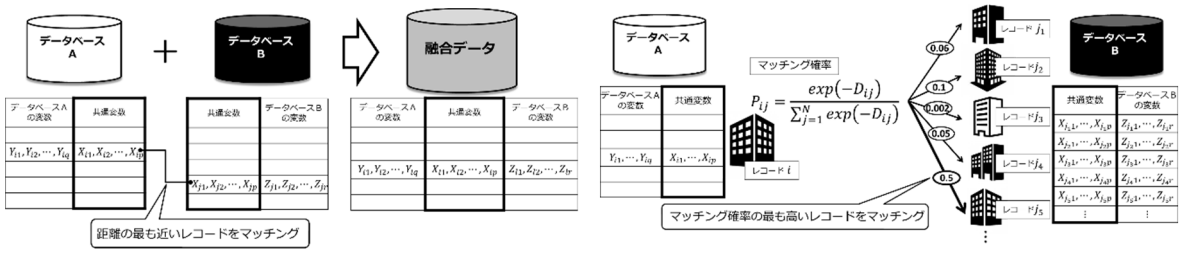
１．統計的マッチングの概要

統計的マッチングは，異なるデータを結合して単一のデータを構築する技術であり，新たなデータ収集を行うことなく有用なデータの作成が可能となる(D'Orazio et al.(2006)，Rassler(2002))．企業データに関しては秘匿性の観点から十分な情報を利用できない場合が多く，売上高や従業者数などの歪んだ分布を持つ変数が多く含まれることもあり，統計的マッチングの応用例は少なく，実務に適用可能な手法が求められている.

２．多項ロジットモデルに基づく統計的マッチング手法の提案

本研究では，多項ロジットモデルとウエイト付き距離関数を組み合わせた新たな統計的マッチング手法を提案する．２種類の企業データの共通フィールドから算出したレコード間の距離を説明変数として多項ロジットモデルを構築することにより，ウエイトの合理的な決定及びマッチング精度の確率的な評価が可能となる．この方法には従来の手法にはない以下のような利点がある.

・距離関数における各変数のウエイトを統計的に決定することが可能.
・連続変数とカテゴリ変数が混在している場合でも問題なく推定が可能.
・マッチングの精度を確率の形で表現することが可能.
・p値や疑似決定係数などにより，モデルの当てはまりを評価することが可能.



３．最適マッチング法に基づく改良

マッチング確率に基づき機械的にマッチングを行った場合，同一のマッチング先レコードに複数のレコードを結びつけてしまう可能性がある．そこで，マッチング確率を重みとした２部グラフのマッチングの問題を考えて，重複のない，重みの和が最大となるような最適マッチングを行う（例えば，２部グラフのマッチング問題において非常に効率が良いとされているハンガリー法など）.

提案手法を，公的統計ミクロデータを含む企業データベースに適用し，従来のマッチング手法等の結果との比較を行った結果については，当日報告する.

参考文献：

[1] D'Orazio, M., M. Di Zio & M. Scanu (2006), *Statistical Matching: Theory and Practice*, Wiley
[2] Rässler, S. (2002), *Statistical Matching*, Springer
[3]高部,山下(2018) 多項ロジットモデルを用いた新たな統計的マッチング手法の提案（forthcoming）

※本研究は科研費（16H02013及び15H03390）の助成を受けている.