中央大学経済研究所年報　第44号(2013) pp.271-306

社会経済ミクロデータ研究会

法人企業統計のデータ・リンケージとその有効性の検証

――統計的マッチング手法の比較から――

坂　田　幸　繁
栗　原　由紀子

本稿では，モデルベースのパラメトリックなマッチング手法を法人企業統計の調査票情報に適用し，マッチングデータから得られる統計量の精度を評価することで，統計的マッチングの有効性を検証した。検証対象とするマッチング手法は，通常の回帰補定法に加え，ベイズ理論に応じて展開した回帰補定法である。また，キー変数の組合せや補助情報の効果に検討を加えるとともに，シミュレーションにより正規性の成立・不成立の影響を計測した。その結果，適切な補助情報を用いた通常の回帰補定法が，相関係数および回帰係数ともに比較的精度のよい推定量を与えることが示された。

1．はじめに

今日，統計情報をめぐる状況は一段と深刻さと過酷さを増しつつある。統計的精神が根付かないまま醸成されたプライバシー保護意識とその偏重，その帰結としての統計調査システムへの消極的，否定的意識形成，および被調査者としての統計調査への直接的非協力（調査拒否をはじめとする無回答や虚偽，不誠実回答など），これらの事象はこれまでになく統計調査情報に致命的な歪みを与え，調査票情報とその利用の信頼性，正確性に大きな影を落としている。それにもかかわらず，エビデンス重視の時代に情報ニーズはこれまでになく高まり，従来型のマクロ的な集計値情報では事足りず，任意の詳細分類，小地域，小集団といった部分母集団レベルでの推定を可能にする高コストのデータセットの作成と提供が求められる[1)]。しかし緊縮財政下，統計予算自体も縮小を余儀なくされ，調査環境も厳しい現状では，新規調査や追加調査によってそのようなニーズに応えるには限界がある。

統計調査情報をめぐるこのような諸条件や要因間の背反や矛盾の克服にはまだ多くの時間

1)　近年の統計法改正とその施行はこのような問題状況を改善する大きな制度変更であり，とくに調査票情報の有効活用の観点から，その実効性が期待される。

を要する。そこでは，パラダイム転換とでも形容すべき新たな統計システムの提示が求められることになる。とはいえ日々の意思決定は続けられねばならないのだから，根本的な改善策がない以上は，応急の，繋ぎの解法でも受け入れられねばならない。すなわち，統計調査情報の歪みや欠損（バイアス）に対しては補正や補定が，既存データの制約（上限）を超える情報ニーズに対しては，情報拡張的手法が喫緊の検討課題である。本稿はそのような方法の中で既存データセットの複合的統合的利用を可能にする代表的手法のひとつである統計的マッチングについて検討を加える。とりわけリンケージ手法による企業データの情報拡張の有効性を，企業を対象にした代表的な基幹統計調査の調査票情報を素材に吟味することにしたい。

統計的マッチングは，データ特性やそれがおかれる条件，および適用するマッチング手法などに応じて，拡張されたデータセットを用いた推定値の精度に違いが生じる。我が国における研究蓄積は必ずしも多くはないが，近年に限れば，荒木・美添（2007）において，家計調査と貯蓄動向調査（総務省統計局）との統計的マッチングによる接合が試みられ[2)]，最近隣距離法（制約付き，制約無し）を適用したリンケージデータの特性が検討されている。また栗原（2012）では，マハラノビス法を用いた統計的マッチングにより中小企業景況調査（中小企業整備基盤機構）から疑似パネルデータを作成し，景況調査のパネル分析を試みている。いずれもノンパラメトリックなマッチング手法に関する研究事例である。これに対して，先行する欧米における研究蓄積の到達点のひとつに Rässler（2002）がある。モデルベースのパラメトリックな統計的マッチング手法（回帰補定，ベイジアン回帰補定など）を体系的に比較したものであり，個人の消費支出データとテレビの視聴記録との接合実験の結果から手法の精度評価を行っている。その結果，周辺分布に関しては補助情報の利用はマッチング精度に影響しないこと，相関関係に関しては補助情報の利用により推定結果の精度改善が期待できること，そして手法としてはベイジアン回帰補定の使用がマッチングデータの柔軟な推定を可能にし，より安定的な結果が得られることなどを指摘している。

本稿の具体的な研究関心もここにあり，これまでの先行研究を踏まえつつ，とくに Rässler（2002）の議論をさらに進めて，モデルベースのパラメトリックなマッチング手法が日本における企業データに対しても適用可能であるのか，その条件は何なのか，統計的な検討を加えていくことにする。検討の素材は，法人企業統計調査（財務省）の調査個票データセットであり，必要に応じてシミュレーション・データも援用しながら，マッチング特性

2） この 2 調査は，重複標本調査部分については ID により完全照合が可能であるが，それ以外に完全照合できないケースも含まれており，それらの有効利用という観点から統計的マッチングを試みている。

を析出し，その精度を検証する。

2．統計的マッチングの方法概要

まず統計的マッチングの方法概要を簡単に整理しておこう。マッチングの関心となる変数Xと変数Y（X, Yを目標変数と呼ぶ）は同時には観察されておらず，したがってデータセット[X, Y]としては存在していない。しかし，各変数はそれぞれ2つのデータファイルAおよびBに分離されて観察されており，しかもAおよびBはマッチングのために利用可能な共通変数Zを含むものとする。これをA＝[X, Z]，B＝[Y, Z]と表すことにする。統計的マッチングでは，このようなファイルAおよびBから共通のキー変数Zを利用して，拡張データセット[X, Y, Z]を作成しようとする[3)]。通常では，どちらか一方のデータセットを基準情報としてもう1つのデータセットからの情報を追加するという処理をする。このときマッチングのベースとなるデータをrecipient，それに情報を提供し接合される側のデータをdonorと呼ぶ。以下ではデータセットAにrecipientファイル，Bにdonorファイルの役割を付与している。なお，モデルベースのマッチングにおいては，A，B相互に役割を換えて，AとB，それぞれ合併した標本サイズの拡張データセットを生成する場合もあることを付記しておく。また変数X, Y, Zは一般に多変量のベクトルであるが，論点を明確にするため，ここでは目標変数X，Yを単変量としている。

2-1　統計的マッチングの目標レベル

統計的マッチングの目標にはいくつかの階層レベルがある。Rässler（2002）にしたがって単純化すれば，下記の4つのレベルに分類できる。本研究の目標とするのは，そのうちレベル3の共分散構造の推定のためのマッチング手法である。相関，回帰といった手法を中心とする計量分析にはこの目標レベルでの検証によりマッチングアプローチの利用可能性を示すことができる。

レベル1：真のデータセットそのものを再構築する。個体識別子（ID）による完全マッチングの結果としてのデータセットの作製と同等の目標である。i番目の個体のマッチングによる推定値にチルダーを付すことにすれば，任意のiに対して，$\tilde{x}_i=x_i$, $\tilde{y}_i=y_i$の成立を目指していることになる。x_i, y_iはi番目の個体の真値（実際の観測値）である。ここでは，個体レベルでのマッチングによるヒット率や一致度を

3)　データセットに同一の標本が含まれ，かつキー変数Zが個体識別子（ID）の場合には完全マッチングが可能となる。

最大化することが目標となる。

レベル２：真の同時分布モデル（関数）を最もよく再現するマッチングデータの作製を目指す。ここでは，マッチングデータから推定される同時分布モデル$\hat{f}_{XYZ}$が，真の同時分布f_{XYZ}にほぼ一致する（$\hat{f}_{XYZ}=f_{XYZ}$）ことが望ましい。

レベル３：XとYとの関係を同時に含む相関特性などの把握のために，共分散構造の近似を可能とするデータセットを作製する。X, Y, Zからなる共分散行列に対して$\widetilde{Cov}(X, Y, Z)=Cov(X, Y, Z)$を目標とする。

レベル４：XあるいはYの周辺分布，もしくはキー変数との同時分布の推定を可能にし，$\hat{f}_X=f_X$, $\hat{f}_{XZ}=f_{XZ}$, $\hat{f}_Y=f_Y$, $\hat{f}_{YZ}=f_{YZ}$を近似的に実現するデータセットを作製する。このレベルの目標は，統計的マッチングを適用せずともＡまたはＢどちらかのデータセットから得られるものであるから，通常はレベル３以上の条件を達成させるときに，同時に満たされるべき制約条件としてレベル４が設定される。

2-2　マッチング精度を規定する条件

統計的マッチングの精度に影響する主な要因には，条件付き独立性の仮定，マッチング手法，目標変数とキー変数との相関，補助情報の良し悪し，および donor の標本サイズや重複標本率などが挙げられる。それぞれについて精度検証に必要な理論的要点を整理しておこう。

⑴　条件付き独立性（CIA；Conditional Independence Assumption）

Zをキー変数としてマッチングする場合，XとYに関するZの条件付き分布の独立性が成立していることが前提となる。

$$\hat{f}(X, Y|Z)=f(X|Z)f(Y|Z) \tag{1}$$

検証に際してはその成否の程度を$Cov(E(X|Z), E(Y|Z))$により確認しておかねばならない。そのために下記のように目標変数に対して，キー変数を説明変数として回帰した残差ε_Xとε_Yとの相関係数によって条件付き従属性（CID）の程度を計測する。CID がゼロに近ければ，CIA が成立していると判断する。

$$X=Z'\beta+\varepsilon_X, \qquad Y=Z'\beta+\varepsilon_Y \tag{2}$$

⑵　マッチング手法

様々なマッチング手法が提案されているが，大別すればノンパラメトリックなアプローチとパラメトリックな（あるいはモデルベースの）アプローチがある。前者の典型は，キー変

数に関する何らかの距離定義の下でA，B，２つのファイルに含まれる個体間の類似性を測り，距離が最も近い個体同士のX，あるいはYの値をリンクさせようというものである。これに対して，キー変数と目標変数の間に，例えば回帰モデルを想定し，その推定値や予測値を利用して拡張データセットを構成しようとするのがパラメトリックなマッチング法である。後者は想定したモデル（仮定）の妥当性が問題となるが，他方で仮定が明示的にモデルに具体化され，また過去の経験や知識などの補助情報をモデルに導入できるという利点をもつ。採用したモデルアプローチについては，次章で改めて説明する。

⑶　目標変数（X, Y）と接合のために利用するキー変数との相関

recipient側の目標変数Xとキー変数Zとの相関，またはdonor側の目標変数Yとキー変数Zとの相関が強ければ，マッチング精度は高まる。マッチングの実際においては，逆に，目標変数との相関を高めるような適切なキー変数セットの選択が重要となる。

⑷　補助情報の利用とその質

パラメトリックなアプローチでは，過去の調査，異なる地域（国）での結果，小グループでの事例的な経験といった情報をマッチングのための補助情報として，推定精度の改善のために利用できる。補助情報の良し悪しが推定精度の向上に影響する場合もあるであろう。本研究では，例えば本格的調査の前に実施される小規模のパイロット調査などの情報が利用できるような状況を想定している。具体的には，条件付き独立性の仮定の成否に関する事前情報が利用できることにして，A，Bファイルからそれぞれ抽出した小規模標本からプライアー（事前分布）を複数作成し，これらを使用して補助情報に関する評価を行う。

⑸　その他

２つのデータファイルに重複標本があれば，類似性を頼りにする統計的マッチングでは同一個体（要素）を接合する可能性が高まり，他方で重複標本がなければ，Donorのデータファイルのサイズが大きいほど一般には類似性が高い候補を引っ張る可能性が高まる。本稿では，問題を複雑化しないように，これらの要因は検証範囲に含めず，重複標本はないものとして，また標本サイズも固定している。

３．マッチング手法

統計的マッチングの精度は，適用するマッチングの手法に強く依存するが，それはデータセットの統計的分布特性によっても異なる。そのため現時点では，どのようなデータセットに対しても，いつでも精度のよい推定値を与えるマッチング手法が確立されているわけではない。本稿では，法人企業統計調査データに対して，ノンパラメトリックな手法の代表としてはマハラノビス法を，パラメトリックな手法の代表としては回帰補定法を取り上げる。

3-1 ノンパラメトリック・マッチング法

ノンパラメトリックな手法では，ある距離関数の下で要素間の距離を測定し，距離が最小となる要素同士を類似個体と想定して接合する。最近隣法による統計的マッチングの精度は，接合に使用する距離関数と変数セット（キー変数）に大きく左右される。本稿では，複数のキー変数セットを利用したマハラノビス距離関数（Mahalanobis Distance；以下 MHL と略称）を検討対象とした。

まず，キー変数ベクトルを z としたとき，接合する 2 つのデータセット A，B について，データセット A の i 番目の要素のキー変数の値 z_i^A と，データセット B の任意の j 番目の要素の値 z_j^B との距離ベクトル $(z_i^A - z_j^B)$ を以下の式で定義する。

$$\text{マハラノビス距離関数：} d_{AB} = (z_i^A - z_j^B)^T \Sigma_{zz}^{-1} (z_i^A - z_j^B) \tag{3}$$

なお Σ_{ZZ} は，A と B のキー変数をマージした変数セットから計算される分散共分散行列である。A の任意の i 番目の個体に対しては，このように定義されるマハラノビス距離が最短となる donor B の j 番目の要素の値を接合することで，拡張データセットを作製する。

3-2 パラメトリック・マッチング法

パラメトリック・マッチングでは，ノンパラメトリック法とは異なり，donor が提供する値を欠損部分にそのまま代入するのではなく，分布形を仮定して recipient および donor から推定したパラメータ・モデルから recipient の要素に対応するシミュレーション値を発生させ，それを補定値として用いる。以下では，古くから利用されている欠損値補定法から発展して統計的マッチング手法へと展開した回帰補定法を中心に，その概要を説明する。

(1) RIEPS（Regression Imputation with Random Residuals）

欠損値処理のための回帰補定法[4]は古くから利用されているが，通常は，donor の値をそのまま欠損値として用いる代わりに，recipient と donor から回帰推定した理論値で代用する。これに対して，補定値に若干のゆらぎを含め，より自然な振る舞いをもたせるために，回帰モデルによる理論値に確率誤差（random residuals）を付加したものが RIEPS である[5]。推定のためのモデル式は以下のようである。

$$X|y \sim N(\bar{\mu}_{X|ZY}\,;\, \hat{\Sigma}_{X|ZY} \otimes I_{na}) \tag{4}$$

$$Y|x \sim N(\bar{\mu}_{Y|ZX}\,;\, \hat{\Sigma}_{Y|ZX} \otimes I_{nb}) \tag{5}$$

---

4) これは予測平均マッチング（predictive mean matching）と呼ばれている。

5) RIEPS および次項の NIBAS の理論的詳細は Rässler（2002），pp.96-107を参照されたい。