【論文】（『統計学』第108号 2015年3月）

統計的マッチングにおける推定精度とキー変数選択の効果

― 法人企業統計調査ミクロデータを対象として ―

栗原由紀子*

要旨

本稿は，法人企業統計調査（財務省）に関する調査票情報の利活用範囲の拡大を目指して，統計的マッチングによるパネルデータの作成可能性を検討した。とくに統計的マッチング手法の比較とともに，精度の高いマッチング推定量（相関係数）を得るためのキー変数選択の条件とその効果について抽出実験により検証を行った。

結果は次の三点に整理できる。まず，マハラノビス法とベイジアン回帰補定法（NIBAS）の比較において，NIBASによる推定量のバイアスが相対的に小さいことを確認した。また，NIBASで適切な推定量を得るための条件としては，条件付従属性がゼロ近傍に位置することのみならず，目標変数との相関が可能な限り強いキー変数セットを用意することが求められる。最後に，NIBASに基づく多重代入法から構成される95％信頼区間については，高い比率で真値をカバーしており，マッチングによる不確実性が多重代入法によりかなりの程度捉えられていることを確認した。

キーワード

ベイジアン回帰補定法，多重代入法，マハラノビス法，正準相関係数，標本実験

1．はじめに

統計的マッチングは，異なるデータソースを個体ベースで融合（Data fusion）することで，情報資源の統合的活用を可能にするとともに，新たな分析枠組みを提供するものである。しかしながら，異なる標本から構成される2つのデータセットに対して，両者にたまたま共通して存在する変数セットを接着剤代わりに融合するため，作製されたマッチング・データに基づく統計量に関してはマッチング誤差が極めて大きな問題となる。

実際にマッチングの適用が必要な場面においては，当然，真値や完全データ[1)]による推定値は不明であるから，マッチング・データからのアウトプットとしての推定値が利用可能な精度を保持しているか否かの判断は困難といえる。そのため，統計的マッチングの利用可能性は，融合対象であるデータセットの特徴や条件を考慮しながら，完全データによる推定値（あるいは真値）が入手できるような特殊な状況をうまく利用して，推定値の分布や特性を詳細に吟味・検討し，その成果を現実の場に敷衍するという方法が有力である。本稿のアプローチもそのような方向に沿っている。

統計的マッチングに関する研究蓄積のなかで主要な成果のひとつにRässler（2002）が

* 弘前大学人文学部
E-mail：yukuri@hirosaki-u.ac.jp

挙げられる。パラメトリック・モデルによる統計的マッチングの手法を比較したものであり，異なる調査結果から得られた消費支出に関するデータとテレビの視聴時間データとの融合を行うことで，マッチング手法の精度比較を行っている。日本においては，荒木・美添（2007）が，家計調査と貯蓄動向調査（総務省統計局）に関して統計的マッチングを行い，ノンパラメトリック手法である各種最近隣距離法による結果の相違が検討されている。また，栗原（2012b）では，ノンパラメトリック手法のマハラノビス距離関数を用いて中小企業景況調査（中小企業整備基盤機構）から疑似パネルデータを作製し，景況調査のパネル分析を試みている。これに対して，坂田・栗原（2013）では，ノンパラメトリック手法およびパラメトリック手法を，法人企業統計調査（財務省）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バイアスや平均二乗誤差を比較することで，有効な推定量を得るためのマッチング手法を検証している。

マッチングの有効性を示すには，標本抽出による推定量の変動を考慮したうえで，マッチングによる推定量のブレを評価する必要がある。しかし先行研究では標本は固定されたままであり，標本抽出の影響に対して十分に注意が払われているとは言い難い。そこで本研究は，法人企業統計調査（財務省）の調査票情報（以下では，法企データとも呼称する）を用いて，抽出実験を行うことにより，法企データのパネル分析に向けたマッチングの精度検証を試みている。

法企データは，資本金規模10億円以上の大企業に限定すれば全数調査が行われており，その階層であれば原理的には識別子によりパネル化できる[2)]。しかし，中小・中堅企業は確率抽出によるサンプルであることから，識別子が利用できたとしても年度をまたがる（1年を超える）パネル化は困難である。したがって，法企データによるこの階層のパネル分析は，有効性が検証された統計的マッチングによって実現することができる。

法企データのパネル化では，同一調査の照合を行うのですべてが共通変数と思われがちであるが，標本も異なり観測時点も異なるのではキー変数の役割を果たさない。そのため，時間的に一定，もしくは変動が少ないと想定される調査項目の異時点データをキー変数に用いるという工夫も考えられるが，作製されたデータセットの有効性という点では疑義が残る。

しかしながら，法企データの一部項目については，当期の実績値に加え前期実績値も同時に記入されており，統計的マッチングにおいて問題となるキー変数の時点間のズレに関しては，これらの調査項目を利用すれば理論的には解消できる。いわば，統計的マッチングには比較的有利なデータセットの条件を法企データは有している。そこで本稿は，このような特性を活用して，法人企業統計調査から統計的マッチングにより作製した疑似パネルデータ分析の可能性を図るため，真値が把握可能な標本階層を検証範囲として，そこからリサンプルした異なる標本間のパネル的融合による推定値の特性を精査することを目的とする。これにより，統計的マッチング手法の選択と推定バイアスとの関係，およびマッチングに使用するキー変数の選択条件とその効果を明らかにしていく。

2．統計的マッチングの概要

統計的マッチングの基本概念を整理しておこう。分析目標は変数Xと変数Y（X，Yを目標変数と呼ぶ）との相関係数の推定に限定する。しかしXとYは同時に観察されておらず，2つのデータセットAおよびBに分離されて観察されているものとする。AおよびBにはマッチングのために利用可能なキー変数セットZが含まれており，A，Bそれぞれのデー

タセットの内容をA：[Y, Z]，B：[X, Z] と表すことにする。統計的マッチングは，このようなデータセット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] を作製するものである[3]。なお，マッチングにより拡張される側のデータセットをrecipientファイル，変数情報を提供し融合される側のデータセットをdonorファイルと呼び，以下ではAにrecipientファイル，Bにdonorファイルの役割を割り当てている。統計的マッチングの精度は，採用するマッチング手法，条件付き独立性の仮定の成否，目標変数とキー変数との相関特性に規定される。以下に，それらの理論的要点を整理しておく[4]。

2.1　マッチング手法

統計的マッチング手法は，ノンパラメトリック法とパラメトリック法の2つに大別できる。前者は，距離関数を定義して，キー変数に関して最も距離が近い個体同士を接合するものである。これに対して，後者は，キー変数と目標変数の間に統計モデルを想定し，その推定値や予測値を利用して理論分布のパラメータを求め，その分布から確率的に発生させた値を補定値とする。本稿では，マハラノビス法とベイジアン回帰補定法を，ノンパラメトリック法とパラメトリック法の代表的手法としてそれぞれとり挙げ，統計的マッチングを実行している[5]。

(a)　マハラノビス法

ノンパラメトリック手法の一つであるマハラノビス法は，キー変数をマハラノビス距離関数（Mahalanobis Distance，以下MHLと略称）に適用して，各要素の距離を測定し，最も距離が最小となる要素同士を接合するものである[6]。

その特徴としては，マッチング計算にはキー変数のみを利用し目標変数は利用しないこと，また補定される値はdonorファイルの値が直接使用され，新たに推定した値ではないことなどが挙げられる。

なお，接合後のマッチング・データから相関係数とその信頼区間を算出する方法は，通常の完全データを用いた方法と同様である。まず相関係数 $\hat{r}$ を算出し，それを(1)式により $\hat{\theta}$ へと変換し，$\hat{\theta}$ の分散推定値が $V(\hat{\theta})=1/(n_1-3)$ であることを用いて，(2)および(3)式により θ の信頼区間 $[\underline{\theta},\overline{\theta}]$ を算出する。ただし，n_1 はサンプルサイズである。

$$\hat{\theta}=\frac{1}{2}\log\frac{1+\hat{r}}{1-\hat{r}} \tag{1}$$

$$\underline{\theta}=\hat{\theta}-1.96\sqrt{V(\hat{\theta})} \tag{2}$$

$$\overline{\theta}=\hat{\theta}+1.96\sqrt{V(\hat{\theta})} \tag{3}$$

その後，(4)式に基づく逆変換（チルダで表示）により相関係数およびその信頼区間を算出する。

$$\tilde{r}=\frac{\exp(2\hat{\theta})-1}{\exp(2\hat{\theta})+1} \tag{4}$$

(b)　回帰補定法と多重代入法

回帰補定法は欠損値処理のために開発されたものであり，データセットに多変量正規分布を仮定して，そのパラメータを回帰モデルなどにより求めたうえで，推定に必要な分布のパラメータの値や目標変数への補定値を確率的に発生させるものである。本稿では，ベイズモデルを援用してパラメータ推定を行うベイジアン回帰補定法（NIBAS；Non-iterative Bayesian-based Imputation）を適用する[7]。マハラノビス法とは異なり，回帰補定法では，キー変数だけでなく目標変数も補定に利用され，また補定値はドナーファイルの値を直接用いるのではなくモデルからの推定値が利用される。なお，補助情報がある場合には，それをモデルに取り込み精度改善に役立てられる柔軟さも有している。

NIBASはある特定の分布から確率的にパ

ラメータや補定値を発生させるため，その補定値は変動し，同時に補定後のデータから得られる統計量も変動する。多重代入法（Multiple Imputation）では，このような確率分布に基づいて発生させた変動を，統計的マッチングによりデータを作製することの不確実性を表すものと捉え，この不確実性まで含めて推定値の評価を行う。そのために，統計的マッチングを複数回実行し，マッチング回毎に推定値を算出し，その推定値集合の平均値を統計的マッチングの推定値とする[8)]。以下では，多重代入法により得られた推定値をMI値と略称する。

MI値とその信頼区間は次のように求められる。まず，統計的マッチングを M 回繰り返すものとする。そのうちの任意の試行回を $m(m=1,\cdots,M)$ としたとき，マッチング・データから算出される相関係数の変換値は(1)式にしたがって $\hat{\theta}_m$ として与えられる。このとき，MI値は $\hat{\theta}_1,\cdots,\hat{\theta}_M$ の平均値として計測される。

$$\hat{\theta}^{MI}=\frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{5}$$

次に，MI値の分散は，1回の推定値に対する群内分散 W（Within Variance）と，推定値間のばらつきである群間分散 B（Between Variance）を複合的に考慮した総分散 T（Total Variance）で与えられる。W は，M 回のマッチングから得られる推定値の分散 $\hat{V}(\theta_m)$ の平均値を，B は M 回分の推定値 $\hat{\theta}_m$ の分散を意味している。

$$W=\frac{\Sigma_{m=1}^{M}V(\hat{\theta}_m)}{M} \tag{6}$$

$$B=\frac{\Sigma_{m=1}^{M}(\hat{\theta}_m-\hat{\theta}^{MI})^2}{M-1} \tag{7}$$

$$T=\left(1+\frac{1}{M}\right)B+W \tag{8}$$

MI値については，推定値の分散を総分散として，自由度 ν の t 分布に従うことが知られている。

$$\nu=(M-1)\left[1+\frac{W}{\left(1+\frac{1}{M}\right)B}\right]^2 \tag{9}$$

MI値による信頼区間 $[\underline{\theta}^{MI},\overline{\theta}^{MI}]$（信頼係数を $1-\alpha$ とする）は，この性質を利用して(10)および(11)式により求められる。

$$\underline{\theta}^{MI}=\theta^{MI}-t_{\alpha/2}(\nu)\sqrt{\hat{T}} \tag{10}$$

$$\overline{\theta}^{MI}=\theta^{MI}+t_{\alpha/2}(\nu)\sqrt{\hat{T}} \tag{11}$$

相関係数のMI値は，相関係数の変換値（M 回分）の平均値により算出している。そのため，相関係数の変換値に関するMI値や信頼区間の値についても，(4)式により逆変換した値を求めている。なお，NIBASによる推定値の算出には，Rässler（2002）のSPLUSコードを参考に，統計ソフトRのためのプログラムを作成し，分析に用いている[9)]。

2.2 条件付き独立性

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（CIA；Conditional Independence Assumption）が成立していることが前提となる。

$$f(X,Y|Z)=f(X|Z)f(Y|Z) \tag{12}$$

この条件の成否を捉えるには完全データが必要であるが，実際に統計的マッチングが必要とされる状況では観測不可能である。しかし本稿では検証の条件として，その成否の程度を確認しておかねばならない。そのために，完全データから目標変数XおよびYのそれぞれをキー変数に対して回帰した残差 ε_X と ε_Y との相関係数を求め，これに基づきCIAの成否を評価する[10)]。これは，いわば条件付き従属性（CID；Conditional Independence and Dependence Index）を示すものであり，CIDがゼロに近いほど，マッチングの精度が高いと期待できる。

$$X = Z'\beta + \varepsilon_X,\ Y = Z'\beta + \varepsilon_Y \quad (13)$$

2.3　目標変数とキー変数との相関

マッチング精度を高める条件のひとつとして，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との相関はできるだけ強いことが望ましい。当然，XとZおよびYとZの両方の相関が極めて強いことが理想的であるが，入手したデータセットがそのような都合のよい条件を満たすとは限らない。そこで，より現実的な場面を想定して，許容できる範囲の精度で推定量を得るには，XとZの相関またはYとZの相関のうち一方だけでも強ければよいのか[11]，あるいはやはり両方の相関がある程度強い必要があるのか，そのときその相関の強さはどの程度あればよいのか，といった実際的な問題への指針となるべく検証作業が設定される必要がある。

本稿では，目標変数と複数のキー変数との相関の強さを測るために，正準相関係数（CCE；Canonical correlation coefficient estimation）[12]を用いている。周知のように，これは2つの変数群の相関構造を探るための手法であり，とくに複数の変数の相関構造を1つの合成指標として捉えることができる。

3．検証方法

3.1　データセットの特徴

本稿では，法人企業統計調査（四半期調査）の2001年第1四半期と2000年第4四半期に関する調査票情報を用いて検証を進める。検証対象は，資本金10億円以上の製造業で識別子によりパネル化が可能である$n=622$社[13]を利用して，2001年第1四半期の収益性指標である総資本経常利益率と，その二期（半年）前の安全性指標である2000年第3四半期の自己資本比率との相関係数の算出を目標とする[14]。

マッチング検証用のデータセットは，表1に示すように，目標変数としてrecipientには総資本経常利益率（Y），donorには自己資本比率（X）を設定し[15]，キー変数はそれぞれZ1～Z8とする[16]。本稿では，donor側の［X, Z］データセットを用いて，recipient側のXを統計的マッチングにより補定することで，［X, Y］が揃ったデータセットを作製することを目標とする。

ここで，Z1，Z2，Z4，Z7については，同時点の情報をキー変数として利用することができる。ただし，標本が重複していれば，それら同時点の情報はほぼ識別子の役割を果たす可能性があるが，本研究では重複標本がないケースを検討するために，同時点であってもrecipientとdonorで異なる標本要素を割り

表1　データセット

［Recipient Data A：2001年Q1］		［Donor Data B：2000年Q4］	
X	missing	X	前期自己資本比率（2000年Q3）
Y	総資本経常利益率（2001年Q1）	Y	missing
Z1	前期流動比率（2000年Q4）	Z1	当期流動比率（2000年Q4）
Z2	前期自己資本比率（2000年Q4）	Z2	当期自己資本比率（2000年Q4）
Z3	従業員数	Z3	従業員数
Z4	前期資本金（2000年Q4）	Z4	当期資本金（2000年Q4）
Z5	売上高	Z5	売上高
Z6	経常利益	Z6	経常利益
Z7	前期総資本（2000年Q4）	Z7	当期総資本（2000年Q4）
Z8	従業員給与	Z8	従業員給与

表2　基本統計量

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Y	X
下位3％平均	33.0	−8.1	13.6	267.2	178.8	−218.5	931.6	17.2	−2.5	−7.2
中央値	114.4	30.2	218.0	494.0	1693.0	15.0	7327.0	272.0	0.3	30.0
平均値	123.5	32.3	258.0	554.8	2058.2	29.1	8291.5	312.9	0.4	32.1
上位3％平均	294.2	75.6	849.9	971.9	6499.6	382.3	25221.7	1012.2	3.9	75.0
標準偏差	53.7	18.8	185.2	190.7	1468.5	111.5	5429.5	223.9	1.3	19.0

（注）　キー変数のZ1からZ8は，Data Aの変数を用いた結果であるが，Data Bについても同様の傾向を示している。
（出所）　著者により作成。

当てており，これら同時点の変数が識別子と同等の役割を果たすものではないことに注意が必要である。

表2には，検証に使用するデータの基本統計量を示している。基本統計量に関しては，その多くが，右に裾野が長い分布形状を示していることが想定される。パラメトリック手法を適用する際には，各変数の正規性の成立が不可欠であることから，これをQ-Qプロットにより確認すると，図1(a)からはX，Y，Z1を除いて，正規性を満たしていないことが分かる。対数変換によりある程度正規化を

図1(a)　Q-Qプロット

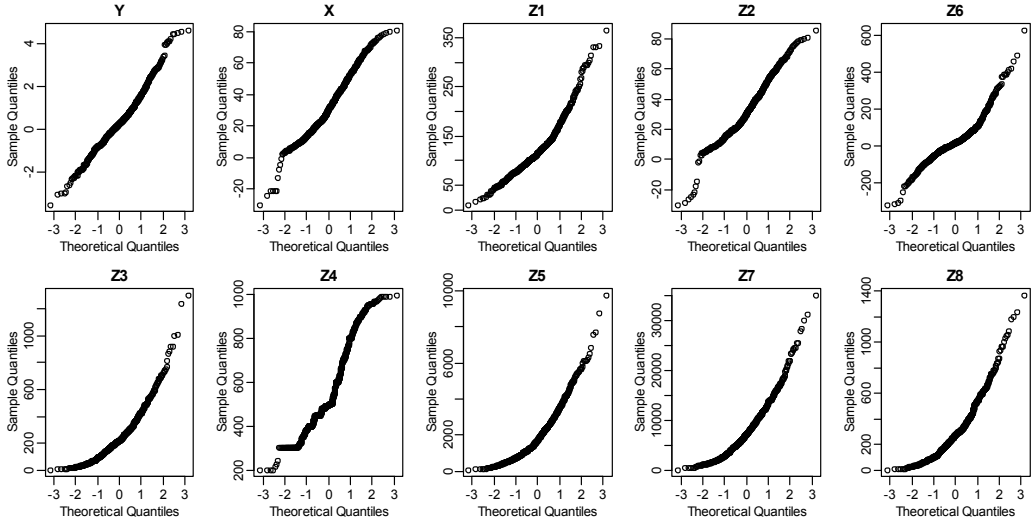
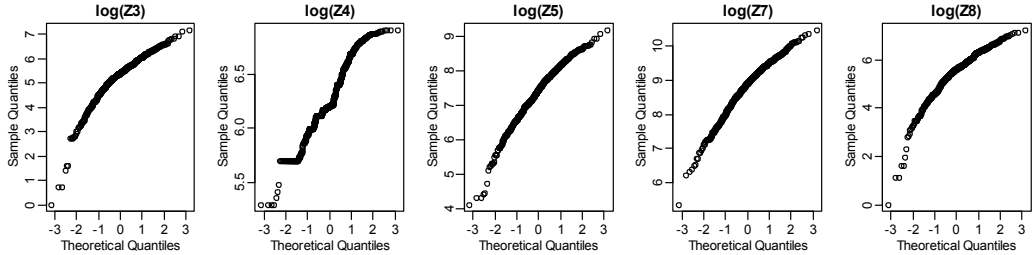


図1(b)　対数変換した変数のQ-Qプロット



（注）　Data Aの変数について分析したものであるが，Data Bについても同様の傾向を示している。
（出所）　著者により作成。

表３　相関行列

	Z1	Z2	log（Z3）	log（Z4）	log（Z5）	Z6	log（Z7）	log（Z8）
X	0.65	0.98	0.04	−0.17	−0.25	0.04	−0.12	−0.11
Y	0.17	0.21	0.00	−0.01	0.09	0.81	−0.06	−0.02

（出所）　著者により作成。

図ることは可能であるが，負の値を含む変数については処理が難しいため，本稿ではZ3，Z4，Z5，Z7，Z8のみ対数変換を行い，正規化を図った（図1(b)）。

相関行列の特徴としては（表3），キー変数ZはX（またはY）との相関が強いほどマッチング精度の改善が見込めるので，単純に比較すると，Z1，Z2，Z6はよいキー変数であり，そのほかのキー変数はマッチングに有効な情報をあまり含んでいないようにみえる[17]。

3.2　検証のプロセス

本稿では，6つの手順により検証を進める。

(1)　まず，母集団として，識別子により完全マッチングが可能な検証用のデータセットA，B（各データのサンプルサイズはそれぞれ$n=622$）を用意し，ここから相関係数の真値rを算出する。

(2)　母集団からサンプルサイズn_1（100）でランダムにサンプリングを行う。ただし，データAとBからはそれぞれ異なる要素を抽出する。データAのサンプリングデータには，Xが含まれないためデータA^{Xmis}とし，同様に，BからはYが得られないためデータB^{Ymis}と表記する。

(3)　この二つのデータA^{Xmis}およびB^{Ymis}を統計的マッチングにより融合することで，[X, Y, Z］が揃ったデータセットを作成する。

(4)　マッチングによりXが補定されたデータ（A^{Xmis}の補定済みデータ）から必要な統計量（相関係数）を算出する。この一回限りのマッチングから得られた推定結果は単一代入法（Single Imputation）による推定値$\tilde{r}_A^{SI}$となる。下付のAはデータセットAの欠損変数Xへの補定であることを示している。

(5)　NIBASについては，(3)と(4)を$M=30$回繰り返して得られる推定値の集合から，Multiple Imputationによる推定値$\tilde{r}_{A,k}^{MI}$およびその95%信頼区間$[\underline{\tilde{r}}_{A,k}^{MI}, \overline{\tilde{r}}_{A,k}^{MI}]$を算出する。

(6)　標本の違いによる影響を考慮するために，(2)から(5)の作業を$K=100$回繰り返し，$r_{A,k}^{MI}$の期待値の推定値$\hat{E}\left(\tilde{r}_{A,k}^{MI}\right)$およびカバレッジを算出する。

$$\hat{E}\left(\tilde{r}_{A,k}^{MI}\right)=\frac{1}{K}\sum_{k=1}^{K}\tilde{r}_{A,k}^{MI} \tag{14}$$

なお，カバレッジは$K=100$回の試行のうち，95%信頼区間$[\underline{\tilde{r}}_{A,k}^{MI}, \overline{r}_{A,k}^{MI}]$に真値が含まれる割合を示す。

4．検証結果

4.1　統計的マッチング手法とバイアス

まずはマッチング手法による結果の違いを評価するために，Z1～Z8の8個全てのキー変数を適用したケースから始めよう。表4には，完全データと統計的マッチング・データ，それぞれについて100回の抽出実験により算出された推定値の期待値（実際には，推定値の期待値に関する推定値であるが，簡略化して「推定値の期待値」と表現する）が示されている。母集団要素をすべて使った真値（TRUE）を基準としたとき，まず完全データの抽出実験により得られた推定値の期待値（COMP）は真値と一致している。これと比べてNIBASによる推定値の期待値は，COMPよりも精度は劣るが，ほぼ真値の近傍に位置している。ただし，MHLはNIBASよりさら

表4　$\hat{E}[\widehat{Cor}(X,Y)]$ とカバレッジ

推定方法	$\hat{E}[\widehat{Cor}(X,Y)]$	Coverage
TRUE（$n=622$）	0.213	
COMP（$n_1=100$）	0.213	98%
NIBAS（$n_1=100$）	0.192	97%
MHL（$n_1=100$）	0.160	92%

（注）　COMPは完全データについて標本抽出実験を行った結果である。なお，CIDは約0.029である。
（出所）　著者により作成。

に精度が悪く，下方にバイアスをもつ。

また，カバレッジについては，NIBASが97%とCOMPの結果に近い数値を示しており，95%信頼区間には100回の抽出実験で95回以上真値が含まれていることが分かる。ただし，MHLについては，カバレッジ95%を下回っており，マハラノビス法で得られた95%信頼区間を疑問視させる結果であった。マハラノビス法に対して求めた相関係数の標準誤差は，通常のデータに適用する標準誤差であり，マッチングによる不確実性が反映されていないことから，信頼区間が過小に設定されていることを示している。以上より，目標統計量を相関係数としてZ1～Z8の全てのキー変数を使用する場合，バイアスの観点からも，また統計的マッチングの精度を適切に評価しているという点でも，MHLよりNIBASが適切といえる。

4.2　キー変数の選択とバイアス

統計的マッチングの精度を規定する条件付き独立性やキー変数と目標変数との相関は，キー変数に左右されることから，キー変数の数やその組み合わせがマッチング精度に与える影響を明らかにしたうえで，利用可能な精度でマッチング・データから推定量を得るためのキー変数の条件を特定しておく必要がある。そこで，キー変数Z1～Z8に対して，1個だけをキー変数として利用した場合から，8個全てを利用した場合まで，全ての組み合わせ（全255通り）についてマッチング実験を行った。

その結果を，マッチングにより得られた推定値の期待値を縦軸，条件付き従属性CIDを横軸として，マッチング手法別に図2に示している。なお，傾向として5つの郡に分けられるため，それぞれA群からE群として大別している（マークについては図3とともに後述する）。

まず，NIBASおよびMHLともに，CIDがゼロ付近であるときバイアスが小さく，CIDの値が高い場合にはバイアスが大きくなる傾向がみてとれる。しかしながら，A群とB群のようにCIDがゼロ付近にあっても，バイアスが小さい場合と大きい場合の2群に分かれるケースがある。さらにNIBASでは，CIDが低いC群よりもCIDが高いD群が，バイアスが若干小さいケースもある。すなわち，キー変数の組み合わせによってCIDは異なるが，CIDとバイアスは直線的な関係で捉えることはできず，統計的マッチングの精度とCIAの関係に関する理論的条件が示すような「CIDがゼロ付近＝バイアスが小さい」という関係が必ずしも成立していないことが分かる。

そこで，マッチングによる推定量のバイアスを，目標変数X，Yそれぞれとキー変数との相関関係から捉え直してみよう。図3において，縦軸は目標変数Xとキー変数Zの相関の強さを示す正準相関係数（CCE），横軸はYとキー変数Zの相関の強さを示すCCEを示している。とくにNIBASにおいては，A群，B群・D群，C群・E群の順にバイアスは低かったが，図3の縦軸における目標変数Xとキー変数Zの相関が強さの順位が，バイアスの低さの順位と同じであることが分かる。すなわち，NIBASを用いて，recipientを固定しXの補定のみにより［X, Y］データセットを作成する場合には，YとZよりもXとZの相関が強いことが不可欠であると考えられる。これに対して，MHLでは，キー変数ZとX

図２　キー変数セット別，$\hat{E}[\widehat{Cor}(X,Y)]$ とCIDの関係

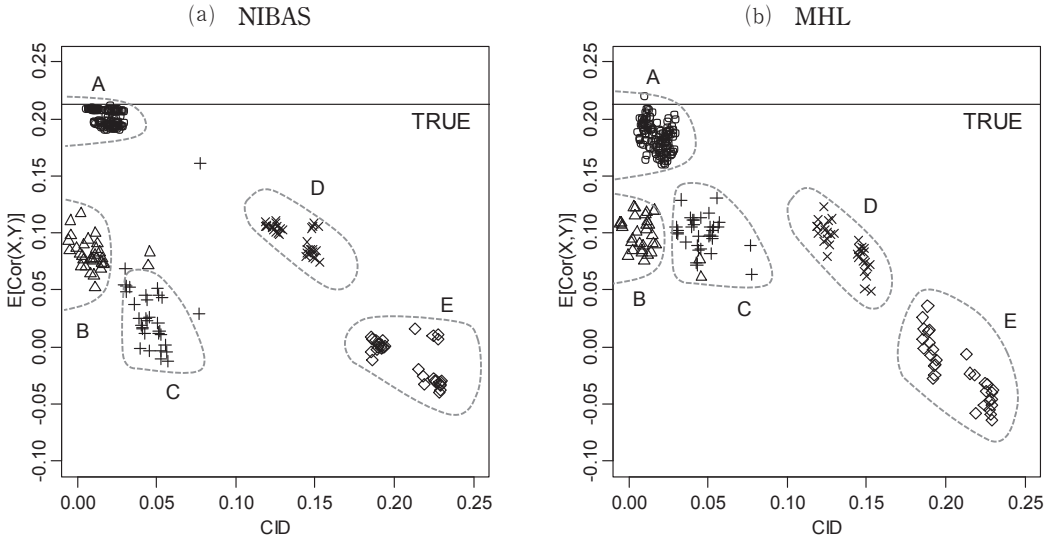


図３　データＡとＢの正準相関係数

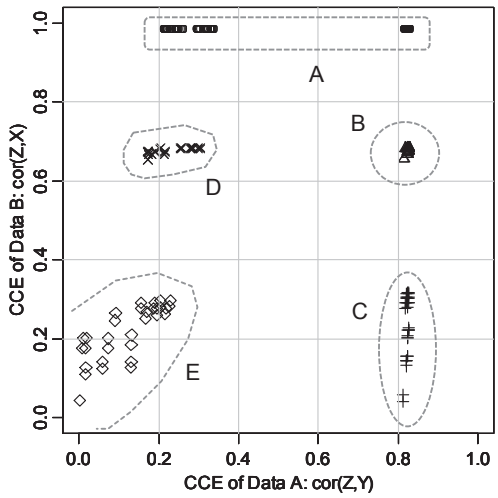
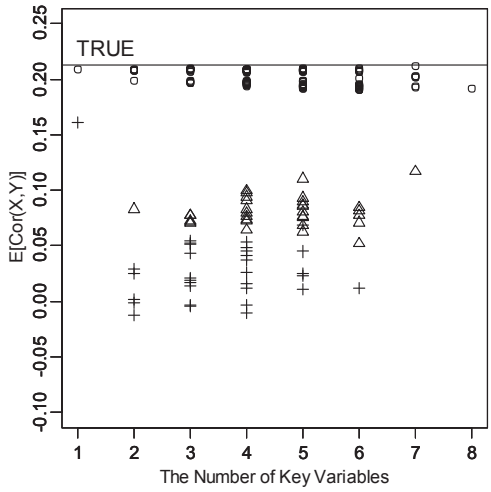


図４　キー変数の数と $\hat{E}[\widehat{Cor}(X,Y)]$ の関係
（NIBAS，Ａ・Ｂ・Ｃ群）



（注）　図２および図４のマークは，図３の結果をもとに分類している。
（出所）　著者により作成。

の相関のみではなく，ＺとＹの相関の強さも精度改善に寄与しており，とくにＹとＺの相関が強いＣ群については，NIBASよりもバイアスが軽減されている。

さらに，各群のキー変数セットの特徴を詳細に検討すると，表５のように，正準相関係数の大きさに応じて，各群に共通する特徴を抽出することができる。本稿での課題の場合，キー変数セットの中で目標変数と最も相関が強い変数によってマッチングの良し悪しのパターンが分類できる。逆にみれば，望ましいキー変数選択の基準として，正準相関係数がそのようなデータサイドの事情を適確に捉えているものと考えられる。

表５　各群と *Cor*(*Z, X*) および *Cor*(*Z, Y*) の最大値

群	*Cor*(*Z, X*) の最大値	*Cor*(*Z, Y*) の最大値	備考
A（○）	*Cor*(*Z2, X*) = 0.98	*Cor*(*Z6, Y*) = 0.82 または *Cor*(*Z2, Y*) = 0.21	Z2 を含む組み合わせ
B（△）	*Cor*(*Z1, X*) = 0.65	*Cor*(*Z6, Y*) = 0.82	Z1 と Z6 を含み Z2 は含まない組み合わせ
C（＋）	*Cor*(*Z6, X*) = 0.20	*Cor*(*Z6, Y*) = 0.82	Z6 を含み Z2 と Z1 は含まない組み合わせ
D（×）	*Cor*(*Z1, X*) = 0.65	*Cor*(*Z1, Y*) = 0.16	Z1 を含み Z2 と Z6 は含まない組み合わせ
E（◇）	上記以外		

なお，図 4 からキー変数の数とバイアスの関係（NIBAS）について確認することができ，特に変数の数の多寡で推定精度が決まるわけではないことがわかる。

4.3　キー変数の選択とカバレッジ

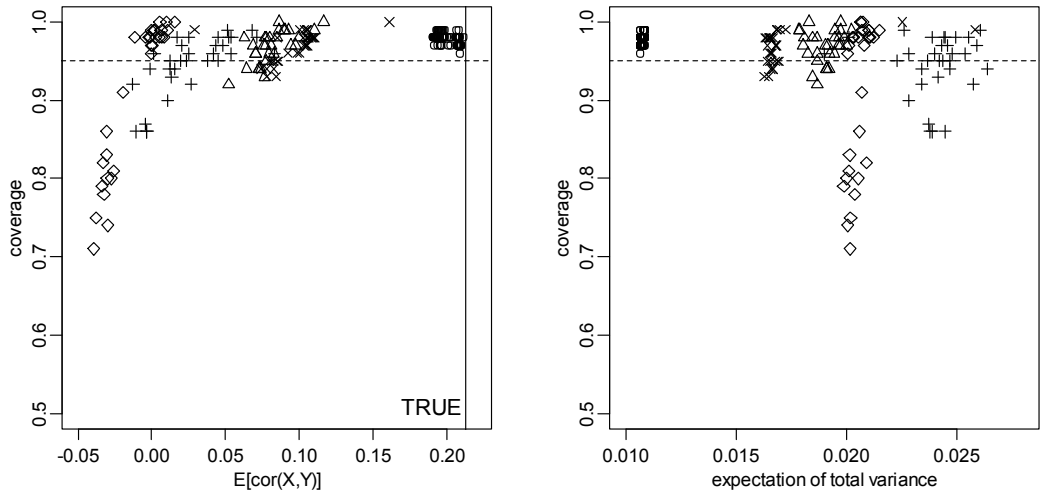
点推定量の特性を踏まえたうえで，統計的マッチングによる 95％信頼区間の特性を，キー変数の組み合わせとの関連で確認しておこう。

図 5 には，NIBAS の結果として，信頼区間のカバレッジ（縦軸）を(a)X と Y の相関係数の推定値の期待値との関連で，また(b)分散推定値の期待値との関連でグラフ化したものである。

図 5(a)によれば，真値の近傍にある A 群（○）はカバレッジも 95％以上であり，若干バイアスのある B，D 群（△，×）の 95％信頼区

図５　カバレッジの特徴（NIBAS）

(a)　*Cor*(*X, Y*) の期待値　　(b)　総分散の期待値



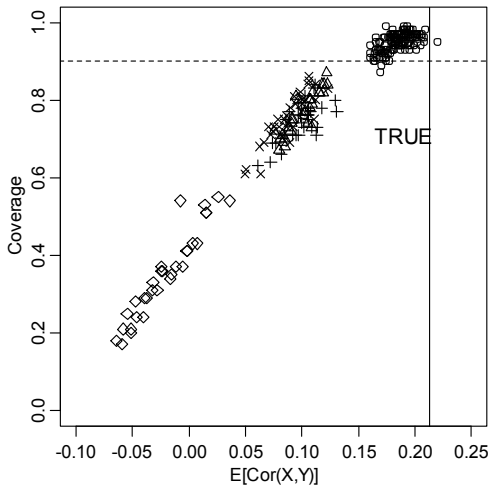
（注）　マークの種別は図 3 と同様である。また Total Variance は，相関係数の変換値に対する分散である。
（出所）　著者により作成。

間についても，多くが90％以上の比率で真値をカバーしている。図５(b)から推察できるように，バイアスが大きいB，D群については，推定量の分散（Total variance）が大きくなることでカバレッジが高く保たれていることが分かる。ただし，比較的バイアスの大きいC，E群（＋，◇）については，カバレッジが90％を下回るケースもある。

これらの結果を正準相関係数との関係から整理すれば，A，B，D群のようにある程度，XとZの正準相関係数が高ければ，CIDがゼロ付近でなくバイアスがあったとしても，もしくはCIDの確認が困難な場合でも，信頼区間を頼りに分析を進めることができる。しかしながら，C，E群のように，XとZの正準相関係数が低い場合には，信頼区間自体も妥当性を欠く。結論的にはXと強い相関を示すキー変数を改めて探すこと，もしくは補助情報の獲得と利用が求められる[18]。

なお，MHLから得られた推定量の期待値とカバレッジとの関係からは（図６），推定量のバイアスが大きくなるにつれカバレッジは低下しており，95％信頼区間とは名ばかりの結果である。とくに，本稿で適用したマハラノビス距離関数に基づく信頼区間に関しては，マッチングによる不確実性をその評価方法に反映させることができないため，そのまま分析に利用するのは問題である。マハラノビス法に関しては，マッチング誤差の評価方法を含めてさらなる検討が必要である。

図６　カバレッジの特徴（MHL）



（注）　マークの種別は図３と同様である。
（出所）　著者により作成。

５．おわりに

本稿では，法人企業統計調査の調査票情報を対象に，マッチング・データからの推定量（相関係数）とマッチング手法およびキー変数選択との関連について検証した。

法企データの一部の調査変数に関しては，調査票情報として前期と当期のデータが与えられているため，パネルデータを作製する際の障壁となるキー変数の時点間のズレに関する問題を，ある程度回避できる。そのため法企データは統計的マッチングによるパネル化という点では，他統計に比して有利な条件が揃っている。このような条件を活用しながら，とりわけ精度の高いマッチング推定量（相関係数）を得るための条件を抽出実験により明らかにすることを試みた。

その結果，バイアスおよびカバレッジにおいて，ノンパラメトリック手法であるマハラノビス法よりもパラメトリック手法であるNIBASのほうが，良い推定量を与えていること，またキー変数選択の際には，CIDがゼロ付近であり，かつキー変数と目標変数Xとの相関（正準相関係数）が極めて強いことが不可欠である。CIDの観測には完全データが必要だが，完全データに代わって補助的な小サンプルデータなどが入手できれば，これらの条件を満たすようキー変数の選択を行えばよい。なお，キー変数の数の多寡はマッチングの精度に強い作用を及ぼすものではないため，キー変数を増やすことよりも，可能なかぎり目標変数XおよびYとの相関が両者ともに強いキー変数を用意する方が効果的といえる。

さらに，NIBASの95％信頼区間に含まれる真値の割合を示すカバレッジ指標については，目標変数との相関が強いキー変数の組み合わせにおいて，高いパフォーマンスが示されており，マッチング誤差に起因する不確実性が，ある程度，多重代入法によりカバーされていることがわかる。もしCIDがゼロ付近にあるか否か確認できない場合には，正準相関係数がある程度高い水準にあることを確認のうえ，マッチング誤差も含めて推定値を評価する信頼区間を分析に利用すればよい。

統計的マッチングの実用化のためには，理論面からのアプローチだけでなく，具体的な統計調査データに即してより多くの検証事例，または適用事例を蓄積していくことが重要といえる。そのような経験の蓄積が，真値が不明な状況下で適切なキー変数セットを選択するための方法論の確立，およびマッチング誤差計測の精度向上に不可欠といえる。本稿の成果を用いた統計的マッチングによる法人企業統計調査の疑似パネルデータ分析については，稿を改めることにしたい。

謝辞

本研究は，「一橋大学経済研究所共同利用共同研究拠点事業プロジェクト研究：立地要因を考慮した企業・事業所活動の経時的特性に関する研究」（研究代表者：法政大学　森博美，平成26年度）の成果の一部である。また，本研究は，財務省から「法人企業統計調査1983年4-7月期～2014年1-3月期」の調査票情報の提供を受け，個票データに基づいて分析を行っている。記して関係諸機関への謝辞とします。

注

1）統計的マッチングを実行することなく，分析に必要な変数が全て揃ったデータを完全データと呼ぶことにする。

2）法人企業統計調査（財務省）には，年次別調査（1948年から実施）および四半期別調査（1950年から実施）があり，1983年以降の調査設計では，資本金10億円以上の企業は全数調査，10億円未満の企業は標本調査が行われている。また，四半期別調査の調査実施時期は，4～6月，7～9月，10～12月，および1～3月の仮決算計数を，それぞれ8月，11月，2月，および5月に調査している（財務省，2011）。なお，四半期別調査では，1年間は固定標本であるから，資本金規模によらず識別子（あるいは企業名，住所などの照合）により年度内については完全照合によるリンケージは可能である。ただし実際には，無回答などによりリンケージできない要素もある。

3）データAとBに同一の標本が含まれ，かつキー変数Zとして個体識別子（ID）が付与されている場合には完全マッチングが可能となる。

4）統計的マッチングの詳細は，Rässler（2002），pp.15-43およびD'Orazio et. al.（2006）pp.13-64を参照。

5）近年，傾向スコアを用いた手法（Propensity Score Matching; PSM）も多用されている（Guo & Fraser，2010，pp.127-210；星野，2009，pp.191-212）。マハラノビス法では，キー変数（共変量）をそのまま照合に用いて最近隣距離法によりマッチングを行うが，PSMは共変量を傾向スコアに集約してその近さでデータをマッチングするという違いがある。これに対して，NIBASは実際にはdonor ファイルのデータをrecipientファイルにリンケージしているのではなく，donorファイルとrecipientファイルからなる多変量分布を想定して，モデルベースでの補定値をマッチング・データとする点で，これらとは大きく異なる（注7を参照）。なお，Rässler（2002）pp.25-42には，3変量正規分布により発生させたシミュレーション・データをもとに，傾向スコアを用いた統計的マッチングの精度を検証し，マッチング後のXとYの相関係数のバイアスが大きいことを示している。

6）マハラノビス距離関数に基づくマッチングは，データAに属するi番目の要素のキー変数ベクト

ルをz_i^A，データBに属するj番目の要素のキー変数ベクトルをz_j^B，またAとBをマージしたキー変数の分散共分散行列をΣ_{ZZ}とする。このとき，これら任意の要素間の距離は以下のように定義でき，マッチングの際には，この距離が最小となるような要素同士を接合する。

$$d_{AB} = (z_i^A - z_j^B)^T \Sigma_{ZZ}^{-1} (z_i^A - z_j^B)$$

なお，MHLの理論的詳細はRässler（2002）p.56を参照のこと。

7）NIBASは，多変量正規分布のパラメータ（$\mu_{X|ZY}$, $\mu_{X|ZY}$, $\Sigma_{X|ZY}$, $\Sigma_{Y|ZX}$）をベイジアンベースにより展開し推定する方法である。

$$X|y, \beta, \Sigma \sim N(\mu_{X|ZY}; \Sigma_{X|ZY})$$
$$Y|x, \beta, \Sigma \sim N(\mu_{Y|ZX}; \Sigma_{Y|ZX})$$

$\mu_{X|ZY}$および$\mu_{Y|ZX}$はそれぞれ回帰モデルを想定して正規分布により発生させ，また$\Sigma_{X|ZY}$および$\Sigma_{Y|ZX}$は逆ウィシャート分布により発生させたうえで，上記モデルに適用し欠損値を確率的に発生させる。なお，NIBASの理論的詳細はRässler（2002）pp.96-107を参照のこと。

8）これに対して1回限りの補定を単一代入法（Single Imputation）と呼ぶ。

9）プログラムコードの詳細は，Rässler（2002）pp.214-221を参照のこと。なお，SPLUSとRのコマンドには相違がある場合もあるため注意が必要である。

10）CIAに関する計測方法は，荒木・美添（2007）に提示されており，栗原（2012a）では相関係数とCIDの理論的関係とともにモンテカルロ・シミュレーションによりその特性を検証している。

11）栗原（2012a）では，ノンパラメトリック法を用いたシミュレーション結果から，XとYの少なくとも一方がキー変数と相関が強ければ，統計的マッチングは利用可能であることを示している。

12）変数群のひとつが1変量で構成されている正準相関係数は重相関係数と一致するが，本稿では一般性を保つために正準相関係数として議論している。

13）検証用データセット（622サンプル）からは，マハラノビス距離にもとづき有意水準5％で外れ値を検出・除外している（奥野・山田，1995，pp.134-137）。

14）法企データの場合，1ファイル内に前期と当期の値が与えられていることから，統計的マッチングによりパネル化をせずとも，一期前の値との相関係数は容易に求められる。

15）統計的マッチングの基本は同時分布を捉えることにあるため，実際の分析に利用する変数が比率や合成値などの場合には，原データをマッチングした後に比率や合成値に変換するのではなく，変換後の値に対してマッチングを適用し，推定量を求めたほうが精度がよい。

16）キー変数には，目標変数との間に可能な限り多様な相関を示す変数を選択している。

17）なお，完全データによるXとYの相関係数は0.21であった。このことから，大企業・製造業（外れ値除外）サンプルに限れば，総資本経常利益率（Y）に対する相関は，1期前の自己資本比率（Z2）であっても2期前の自己資本比率（X）であっても0.21と不変である。

18）本稿の精度検証をもとに，資本金10億円未満の企業に関して，統計的マッチングを試行したところ，最も正準相関係数が高い（ZとXのCCEは0.98，ZとYのCCEは0.51）キー変数の組み合わせは全てのキー変数を使用したケースであり，目標変数の相関係数は0.055，信頼区間は［0.006, 0.104］であった。資本金10億円以上の企業では，0.21であったことから，資本金規模が小さい企業に関しては，当期収益性と2期前の安全性との相関は無い（または極めて小さい）ことが示されている。

参考文献

［1］ D'Orazio, M., M. Di Zio & M. Scanu（2006）, *Statistical Matching: Theory and Practice*, Wiley, West Sussex.

［2］ Goel, P.K. & T. Ramalingam（1980）, *The Matching Methodology: Some Statistical Properties*, Springer, Berlin.

［3］ Guo, S. & M.W. Fraser（2010）, *Propensity Score Analysius: Statistical Methods and Applications*, SAGE, California.

［4］ Haltiwanger, J.C., etc（1999）, *The Creation and Analysis of Employer-Employee Matched Data*, North-

Holland.
［5］ Little, R.J.A. & D.B. Rubin（2002）, *Statistical Analysis with Missing Data*, Wiley, New York.
［6］ Rässler, S.（2002）, *Statistical Matching*, Springer, New York.
［7］ 荒木万寿夫・美添泰人（2007），「家計データを利用した完全照合と統計的照合」，『青山経営論集』，第42巻第1号，pp.175-210.
［8］ 奥野忠一，山田文道（1995），『情報化時代の経営分析』，東京大学出版会.
［9］ 栗原由紀子（2012a），「相関特性推定における統計的マッチングの有効性について ― モンテカルロ・シミュレーションによる精度検証 ― 」，『中央大学経済研究所年報』，中央大学経済研究所，第43号，pp.489-551.
［10］ 栗原由紀子（2012b），『疑似景況パネルによる予測パフォーマンスの計測 ― マハラノビス・マッチングの適用から ― 』，法政大学日本統計研究所，オケージョナル・ペーパー，No. 35, pp.1-38.
［11］ 財務省（2011），「法人企業統計調査の変遷と概要」，『フィナンシャル・レビュー』，財務省財務総合政策研究所，通巻第107号.
［12］ 坂田幸繁･栗原由紀子（2013），「法人企業統計のデータ･リンケージとその有効性の検証」，『中央大学経済研究所年報』，中央大学経済研究所，第44号，pp.271-306.
［13］ 星野崇宏（2009），『調査観察データの統計科学』，岩波書店.
［14］ 間瀬茂（2007），『Rプログラミングマニュアル』，数理工学社.

Estimation Precision of Statistical Matching and Selection Effects of Common Variables

Yukiko KURIHARA*

Summary

This study verifies the precision of correlation coefficients based on statistical matching and multiple imputation under different matching methods and combinations of common variables. The matching methods for verification are a non-parametric approach based on Mahalanobis distance and the Bayesian regression imputation method (NIBAS)—a parametric method. Questionnaire data from the Financial Statements Statistics of Corporations by Industry (Ministry of Finance) were used to clarify the effectiveness of matching data created from different sample datasets.

The three main findings are as follows: First, NIBAS enables the estimation of correlation coefficients with lesser bias than those of the Mahalanobis matching method. Second, the primary condition for high-precision estimation is a combination of common variables with both low conditional dependence and strong correlation with target variables. Finally, the confidence interval computed by multiple imputation with NIBAS suitably covers the true value and measures the uncertainty inherent in statistical matching, except in the case of point estimates with extremely large bias.

Key Words

Bayesian regression imputation, Multiple imputation, Mahalanobis method, Canonical-correlation coefficient, Sampling experiment

* Faculty of Humanities, Hirosaki University