$$\tilde{\mu}_{X|ZY}=Z_A\hat{\beta}_{XZ}+(Y-Z_A\hat{\beta}_{YZ})\hat{\Sigma}_{YY|Z}^{-1}\hat{\Sigma}_{YX|Z} \tag{6}$$

$$\tilde{\mu}_{Y|ZX}=Z_B\hat{\beta}_{YZ}+(X-Z_B\hat{\beta}_{XZ})\hat{\Sigma}_{XX|Z}^{-1}\hat{\Sigma}_{XY|Z} \tag{7}$$

X は q 個，Y は p 個，Z は h 個の変数から構成されるベクトルとして，$\tilde{\mu}_{X|ZY}$ および $\tilde{\mu}_{Y|ZX}$ は回帰補定による理論値を意味する。$\hat{\beta}_{XZ}$ および $\hat{\beta}_{YZ}$ は，X を Z_B に回帰したときのパラメータ・ベクトルと，Y を Z_A に回帰したときのパラメータ・ベクトルをそれぞれ示す。

また，$\hat{\Sigma}_{XX|Z}$ および $\hat{\Sigma}_{YY|Z}$ は，Z の条件つき X の分散共分散行列，および Z の条件つき Y の分散共分散行列をそれぞれ示す。回帰補定による理論値 $\tilde{\mu}_{X|ZY}$ および $\tilde{\mu}_{Y|ZX}$ との誤差を $\hat{\Sigma}_{X|ZY}$ および $\hat{\Sigma}_{Y|ZX}$ で計測し，これを正規分布の分散に関するパラメータ計算に利用する。⊗ はクロネッカー積を，I_{na}，I_{nb}は A，B それぞれのデータサイズn_a，n_b を次数とする単位行列である。さらに，Z の条件付き相関係数$\rho_{XY|Z}$ と，Z の条件付き X または Y の分散 $\tilde{\sigma}_{XX|Z}$，$\tilde{\sigma}_{YY|Z}$ から $\hat{\Sigma}_{YX|Z}$ を定義する。

$$\hat{\Sigma}_{YX|Z}=\left\{\rho_{X_iY_j|Z}\sqrt{\tilde{\sigma}_{X_iX_i|Z}\tilde{\sigma}_{Y_jY_j|Z}}\right\},\ i=1,2,\ldots,q,\ j=1,2,\ldots,p. \tag{8}$$

ここで，$\rho_{X_iY_j|Z}$ は，完全データの場合に観測可能な値であり，マッチングの実際では未知である。したがって，ここには補助情報から計測した数値を適用することになるが，補助情報がない（無情報の）場合にはゼロを投入せざるを得ない。RIEPS はベイズ理論に沿ったモデルではないため，この値はベイズ的意味での事前情報とは異なるが，事前に得た補助情報を投入するという意味でこれをプライアーと呼ぶことにする。

RIEPS モデルでは，多変量正規分布を仮定してランダムに補定値を発生させる。しかし，実際には現実の企業データがとり得る範囲を超える危険性があることから，本稿では法企データの最大値と最小値を閾値とした切断された多変量正規分布から補定値を発生させている。

⑵　NIBAS（Non-iterative Bayesian-based Imputation）

NIBAS は，RIEPS の発想をベイズ理論の枠組で再構成し，明示的に事前分布を設定して得られる事後分布によって実現されるマッチング手法である。展開後に得られる推定式をまとめると，RIEPS で推定した $\hat{\beta}_{XZ}$，$\hat{\beta}_{YZ}$，または $\hat{\Sigma}_{XX|Z}$，$\hat{\Sigma}_{YY|Z}$ に基づいて，β_{XZ} および β_{YZ} は正規分布から発生させた値を使用し，分散共分散行列 Σ_{XX}^{-1} および Σ_{YY}^{-1} は逆ウィシャート分布からそれぞれ発生させた値を利用する。

$$X|y,\beta,\Sigma\ \sim N(\mu_{X|ZY};\ (\Sigma_{XX|Z}-\Sigma_{XY|Z}\Sigma_{YY|Z}^{-1}\Sigma_{YX|Z})\otimes I_{na}) \tag{9}$$

$$Y|x,\beta,\Sigma\ \sim N(\mu_{Y|ZX};\ (\Sigma_{YY|Z}-\Sigma_{YX|Z}\Sigma_{XX|Z}^{-1}\Sigma_{XY|Z})\otimes I_{nb}) \tag{10}$$

$$\mu_{X|ZY}=Z_A\beta_{XZ}+(Y-Z_A\beta_{YZ})\Sigma_{YY|Z}^{-1}\Sigma_{YX|Z} \tag{11}$$

$$\mu_{Y|ZX}=Z_B\beta_{YZ}+(X-Z_B\beta_{XZ})\Sigma_{XX|Z}^{-1}\Sigma_{XY|Z} \tag{12}$$

ベイズ理論に沿って展開すれば，事前分布情報が $\Sigma_{YX|Z}$ として残るため，実質的に，統計的マッチングをする際のプライアーは $\Sigma_{YX|Z}$ および $\Sigma_{XY|Z}$ を使用することになる。理論的導出は異なるが，RIEPS および NIBAS ともに同じ事前情報 $\Sigma_{YX|Z}$ および $\Sigma_{XY|Z}$ を使用するため，既に述べたようにこれをプライアーと呼んでいる。RIEPS と同様に，NIBAS のプライアーに関しても，事前に何らかの補助情報が得られている場合にはそれを利用するが，補助情報がない場合にはゼロを設定する。また NIBAS についても，多変量正規分布をシミュレートするが，RIEPS と同じ理由から使用する法企データの最大値と最小値を用いて，切断された多変量正規分布から補定値を発生させている。

3-3 MI 推定量

回帰補定法では，仮定した分布から確率的に補定値を発生させるため，補定後のデータから得られる統計量も確率的に変動する。また，マハラノビス距離関数においても，最近隣法に基づく接合は最短距離に位置する要素が複数個観測されるケースが多く，そのような場合には，確率的に接合個体を割り当てる方法がとられる。そのため推定値はマッチング試行ごとに確率的なぶれを含むことになる。

この種の変動は，統計的マッチングによりデータを作成することの不確実性を表すものであり，このような不確実性まで含めて統計的に評価するために，複数回の統計的マッチングの結果得られたデータから，複数個の推定値を算出し，例えばその平均値を統計的マッチングの推定値とする。このような多重補定（Multiple Imputation：MI）的方法による結果を MI 値と略称する（これに対して 1 回限りの補定による統計量を Single Imputation と呼ぶ）。本稿で使用する MI 値の特性を整理しておこう。

まず，統計的マッチングを m 回試行したときの MI 値 $\bar{\theta}^{MI}$ を，m 回分の推定値の平均値(13)式により計測することにする。ここで，$\hat{\theta}_k$ は k 回目，$k=1,2,\ldots,m$ の推定量（例えば平均，相関係数，回帰係数など）を意味する。

$$\bar{\theta}^{MI}=\frac{\Sigma_k\hat{\theta}_k}{m} \tag{13}$$

MI 値の分散は，各回の推定値毎のばらつき W（within variance）と，推定値間のばらつき B（between variance）から構成される総分散 T（total variance）で与えられる。具体的には W は，各回のマッチングから得られる推定値の分散 $\hat{V}(\hat{\theta}_k),k=1,2,\ldots,m$ の平均であり，

B は，m 回分の $\hat{\theta}_k$ の分散を意味している。

$$W=\frac{\Sigma_k \hat{V}(\hat{\theta}_k)}{m} \tag{14}$$

$$B=\frac{\Sigma_k(\hat{\theta}_k-\hat{\theta}^{MI})^2}{m-1} \tag{15}$$

$$T=\left(1+\frac{1}{m}\right)B+W \tag{16}$$

MI 値とその分散推定値から計算される次の統計量は，自由度 ν の t 分布に従う。この性質を利用して，検定や推定を行うことができる。

$$\frac{\hat{\theta}^{MI}-\theta}{\sqrt{T}} \sim t(\nu) \tag{17}$$

$$\nu=(m-1)\left(1+\frac{W}{\left(1+\frac{1}{m}\right)B}\right)^2 \tag{18}$$

なお，本稿ではマッチングによる推定量のバイアスや平均平方誤差（あるいは RMSE）を，下記のように完全データから得られる真値 θ に対する相対比率として実際の推定値から算出している。これを手掛かりにマッチング手法やキー変数の組み合わせ，あるいは補助情報の効果などを評価していく。なお図表中の表記は簡略化して，相対バイアスを bias，相対 RMSE を mse と以下では表記している。

$$\text{bias}=\left[\frac{\Sigma_k \hat{\theta}_k}{m}-\theta\right]/|\theta| \tag{19}$$

$$\text{mse}=\sqrt{\frac{\Sigma_k(\hat{\theta}_k-\theta)^2}{m}}/|\theta| \tag{20}$$

4．精度検証の方法

4-1　検証内容と検証方法

法人企業統計調査（四半期調査）――以下では簡単に法企データとも呼ぶことにする――の2000年第 1 四半期の調査票情報から，製造業・大企業を対象として検証用に $n_a=n_b=500$ の標本をランダムに抽出し利用する。キー変数には，従業員数（Z1），平均給与（Z2），資本金（Z3），目標変数として Recipient には売上高（Y），Donor には資産合計（X）を設定する。以下では，変数をそれぞれ Z1，Z2，Z3，X，Y と表記する。

検証の方法は，図 4-1 にまとめている。まず，検証用のデータセット（標本サイズ

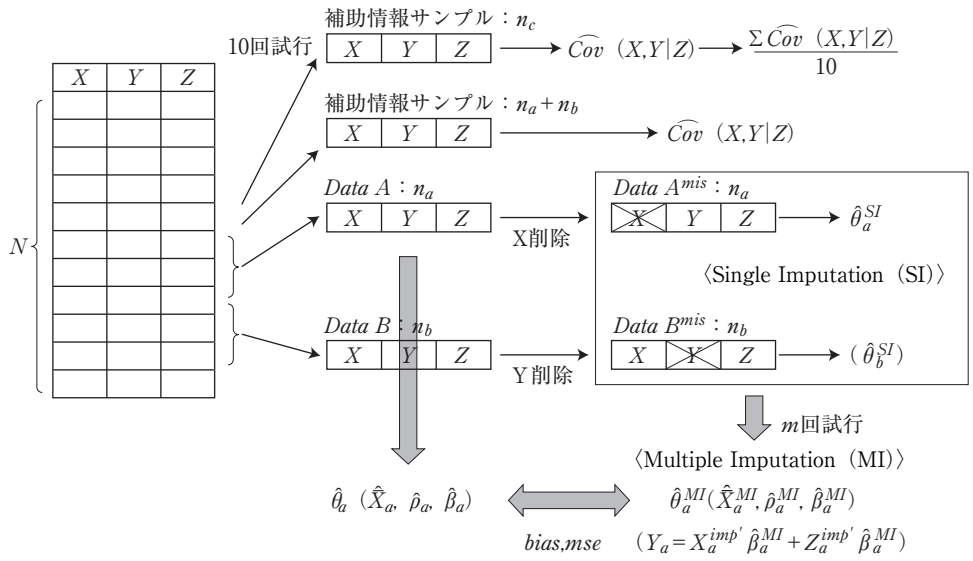
n＝1000）を用意し，ランダムにサイズ500のサンプルずつに切り分け，それぞれデータセットＡおよびＢとおく。ＡおよびＢには変数Ｘ，Ｙ，Ｚすべてが含まれている。ここで，データセットＡから変数Ｘのみを削除して，新たにデータセットA^{mis}を用意する。同様に，Ｂからはのみを削除したデータセットB^{mis}を用意する。この２つのデータセットA^{mis}およびB^{mis}を統計的マッチングにより接合することで，Ｘ,Ｙ,Ｚが揃ったデータセットを改めて作製する。そして最後に，マッチングによりＸが補定されたデータセットA^{mis}から必要な統計量を算出する。この１回限りのマッチングから得られた推定結果は単一補定（Single Imputation）による推定値$\hat{\theta}_a^{SI}$である。このようなプロセスを100回繰り返して得られる推定値集合から，多重補定（Multiple Imputation）による推定値$\hat{\theta}_a^{MI}$が得られる。統計的マッチングの精度検証のために，本稿ではＸ，Ｙ，Ｚが揃った本来の完全データセットＡから計算した推定値$\hat{\theta}_a$を利用して，多重補定による推定値$\hat{\theta}_a^{MI}$の相対バイアスや相対RMSEを算出する。

ここで，検証のために算出する統計量は，平均値（前出の目標レベル４），相関係数（目標レベル３）および回帰係数（$Y=\beta_0+\beta_1 Z1+\beta_2 Z2+\beta_3 Z3+\beta_4 X$，目標レベル３）である。また本稿では，以下の３点に絞って，マッチングの有用性を検証している。

(a) マッチング手法について，RIEPS，NIBAS，MHLの３手法の精度比較を行う。同時に，補助情報が使える手法については，補助情報による改善の程度を検証する。（後述）

(b) キー変数の組み合わせの違いによるマッチング精度の違いを評価する。

図 4-1 統計的マッチングの検証方法



(c) 最後に，正規性からの逸脱度がどの程度マッチング精度に影響するかを検討する。多変量での正規性の仮定を満たしたシミュレーション・データからのマッチング実験による推定値と，実際の法企データによるマッチングデータからの推定値を用いて，統計的マッチングの精度に違いがあるかどうかを検証する。シミュレーション・データは，法企データ A の対数変換値から，平均値および分散共分散行列を算出し，これを使って同じ平均と分散・共分散を有する多変量正規分布から1000個のデータを発生させ，これを逆変換してシミュレーション・データ A′ を作成している。

4-2　データセットの特性

(1) 基本統計量

表 4-1 から 4-4 は，検証に使用するデータ A および A′ の基本統計量と相関行列を示している。基本統計量に関しては，実際の法企データ A よりも，シミュレーション・データ A′

表 4-1　Data A の基本統計量

統計量	Z1（人）	Z2（百万）	Z3（億円）	X（億円）	Y（億円）
下位 3 %平均	39	0.63	10	25	4
中央値	623	1.44	33	311	67
上位 3 %平均	15236	2.64	1597	14241	2950
平均値	1476	1.47	120	1056	219
標準偏差	3511	0.42	358	3034	664

表 4-2　Data A′ の基本統計量

統計量	Z1（人）	Z2（百万）	Z3（億円）	X（億円）	Y（億円）
下位 3 %平均	34	0.73	4	17	4
中央値	695	1.41	44	369	76
上位 3 %平均	8869	2.83	531	6137	1384
平均値	1260	1.48	80	757	168
標準偏差	1760	0.45	107	1212	302

表 4-3　Data A の相関行列

	Z1	Z2	Z3	X	Y
Z1	1				
Z2	0.146	1			
Z3	0.650	0.285	1		
X	0.853	0.314	0.825	1	
Y	0.852	0.337	0.724	0.910	1

表 4-4　Data A′ の相関行列

	Z1	Z2	Z3	X	Y
Z1	1				
Z2	0.154	1			
Z3	0.642	0.290	1		
X	0.845	0.306	0.814	1	
Y	0.851	0.318	0.725	0.909	1

の方が，平均値が低めに出ており，また標準偏差とレンジが示すようにばらつきも小さい。上位３％平均が示すように，対数化しても実データは正規分布に対していくらか右にすそ野が長い分布となっており，その歪みが逆変換して元に戻したときに上記のような違いとなって現れている。

相関行列については際立った差異は見当たらず，どちらも類似した数値を示している。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 と Z3 はよいキー変数であり，これに対して Z2 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。

⑵ プライアー

マッチング精度を改善するためのプライアーとなるＺの条件付ＸおよびＹの共分散 Cov（X, Y | Z）を，補助情報の入手可能性を考慮して，３段階に分けて設定した。ひとつは情報がない場合であり，このときプライアーはゼロとする（モデル表記は null)。次に小サンプルのプライアーが得られた場合を想定し，完全データからランダムに抽出した50サンプル（抽出率５％）から Cov（X, Y | Z）を計算し，これを10回繰り返した場合の平均値を，２つ目のプライアー（表記は sample）としている。また，比較のために完全データをそのまま利用して計算した値を３つ目のプライアー（表記は full）として用意する。補助情報として，最善のプライアーが入手できたというケースを想定している。これによって補助情報の極限効果が検討できる（表 4-5，4-6 参照)。

表 4-5 Data A のモデル別プライアー

Model	Z123	Z12	Z13	Z23	Z1	Z2	Z3
sample	0.373	0.330	0.397	0.599	0.302	0.200	0.650
full	0.420	0.360	0.436	0.663	0.321	0.209	0.650

表 4-6 Data A′ のモデル別プライアー

Model	Z123	Z12	Z13	Z23	Z1	Z2	Z3
sample	0.377	0.337	0.406	0.590	0.325	0.168	0.601
full	0.424	0.365	0.448	0.650	0.345	0.175	0.648

⑶ 正規性の確認

図 4-2 は，法企データ A の原数値(a)とその対数変換値(b)の Q-Q プロットをそれぞれ示している。法企データの原数値のままでは，全く正規性の条件は満たされないが，対数変換によってある程度正規分布に近付いている様子は確認できる。ただし，RIEPS と NIBAS は多変量正規分布を仮定した手法であるのに対して，対数変換によって保障されるのは単変量で

図 4-2(a)　Data A の変数の Q-Q プロット

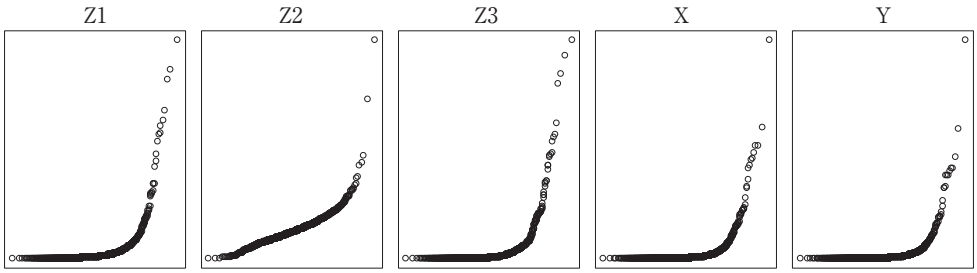


図 4-2(b)　Data A の対数変換した変数の Q-Q プロット

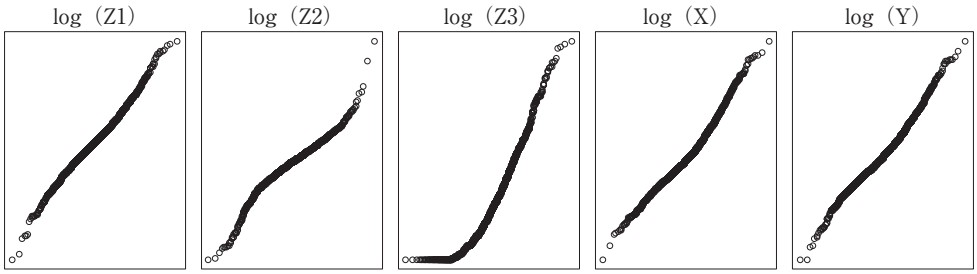


図 4-3(a)　Data A′ の変数の Q-Q プロット

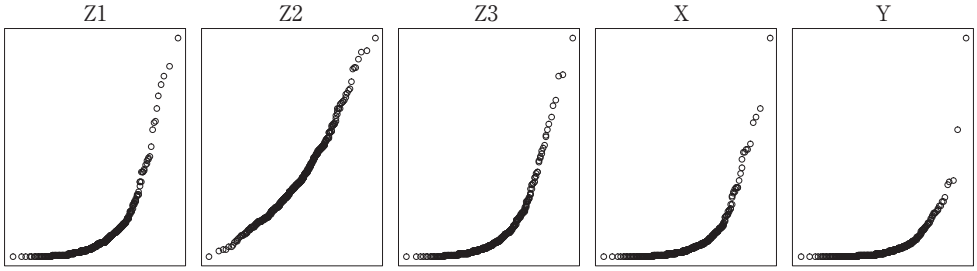
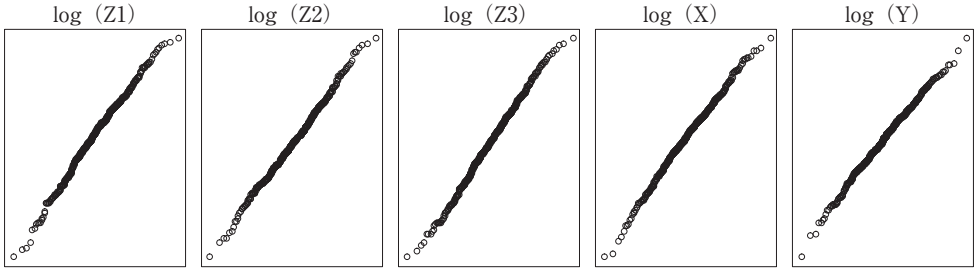


図 4-3(b)　Data A′ の対数変換した変数の Q-Q プロット



の周辺正規性であるから，この点がどのようにマッチング精度に影響を及ぼすかは，シミュレーション・データによる推定結果との比較により確認する。

図 4-3 には，シミュレーション・データ A′ の原数値(a)とその対数変換値(b)の Q-Q プロットが示されている。シミュレーション・データの作成方法からも明らかなように，A′ の対数変換値は当然，正規分布に従う。

5．検 証 結 果

検証手順としては，平均値，相関係数，回帰係数の順に，それぞれ(1)手法比較，(2)モデル（キー変数セット）比較，(3)シミュレーション・データを用いた正規性の仮定チェックを行っている。以下で使用する図表中のモデル表記などをここで改めて整理しておく。なお，結果数値表は本稿末尾の〈資料〉に掲載している。

［手法（Method）］ ［プライアー］ ［キー変数セット(Model)］

$$\begin{bmatrix}\text{RIEPS}\\ \text{NIBAS}\\ \text{MHL(Mahalanobis)}\end{bmatrix} \times \begin{bmatrix}\text{null}\\ \text{sample}\\ \text{full}\end{bmatrix} \times \begin{bmatrix}\text{Z123(Z1, Z2, Z3を使用)}\\ \text{Z12（Z1, Z2を使用）}\\ \text{Z13（Z1, Z3を使用）}\\ \text{Z23（Z2, Z3を使用）}\\ \text{Z1（Z1を使用）}\\ \text{Z2（Z2を使用）}\\ \text{Z3（Z3を使用）}\end{bmatrix}$$

5-1 平均値（周辺分布の点推定値）

(1) マッチング手法の比較

統計的マッチングの手法や補助情報の違いによる X（対数変換値）の点推定値（＝平均値）と信頼区間の相違をまとめたものが図 5-1 である。また，表 5-1 には，それらのバイアスと mse を整理している。計算結果は，すべてのキー変数（Z1, Z2, Z3）を使用したモデル Z123のケースである。

図 5-1 より，各手法の点推定値（○印）は完全データ A の真値（TRUE）とそれほど大きな違いはみられないが，モデルベースのようにデータを発生させて推定値を補定するものとは異なり，既存の donor データの要素を接合するマハラノビス法（MHL）が最もよい近似を示している。なおいずれも，統計的マッチングにより加わる不確実性のため，95％信頼区間（実線）は完全データによる真の区間幅と比較して，約2倍以上に拡大している。

表 5-1 からは，マッチング手法や補助情報の違いにかかわらず，bias も mse も0.5％以下であり，誤差は極めて小さく，十分実用的なデータセットがマッチングにより提供されると判断できる。最もよい結果を示す手法はマハラノビス法であり，次に完全データの補助情報

図 5-1　手法別 logX の平均値および信頼区間（Data A，モデル Z123）

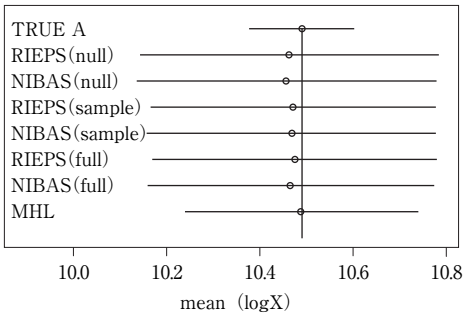


表 5-1　手法別 logX 平均値の bias および mse（Data A，モデル Z123）

Method	bias	mse
RIEPS（null）	−0.0026	0.0035
NIBAS（null）	−0.0033	0.0050
RIEPS（sample）	−0.0018	0.0026
NIBAS（sample）	−0.0021	0.0037
RIEPS（full）	−0.0015	0.0025
NIBAS（full）	−0.0023	0.0036
MHL	−0.0001	0.0001

を利用した RIEPS（full）となっている。

（2）　モデルの比較

キー変数の組み合わせの違いによるマッチング・パフォーマンスを比較するために表 5-2 を作成した。マッチング手法は RIEPS に固定して，相対バイアス（bias と表記）に関して，使用するキー変数の影響や，プライアーを使用した場合の変化率[6]を検討する。変化率は，プラスであればプライアー利用によりバイアスが改善したことを，マイナスのとき悪化したことを示す。Z の条件付き X と Y の従属性（CID），および RIEPS（full）で使用したプライアー（prior）値も参考数値として掲げている。

RIEPS（null）においてモデル間の bias 値を比較すると，最悪の場合（Z2で0.1）を除けば，他のケースはほぼ0に近く，実用上は問題のないレベルと判断できる。補助情報がなくキー変数を頼りにマッチングを行う場合には，周辺分布に関しては適切なキー変数の設定が重要といえる。また CID が高く条件付独立性が成立していなくとも（Z23），バイアスに関しては問題の少ない点推定値を算出することができる。

それでは，補助情報は役に立つのだろうか。プライアーを使用したときにバイアスが改善するかどうかを変化率で確認すると，多くのケースで変化率がプラスである。相対バイアスだから，RIEPS（null）がゼロに近いケースの変化率は無視してよい。むしろ，Z2を使用するケースで，プライアー使用により2割程度のバイアスの縮小がみられる点に注目すべきであろう。

（3）　シミュレーション・データによる結果との比較

図 5-2 は，実際の法企データ A を使用した場合の統計的マッチングの平均値と，シミュレーション・データ A′ から統計的マッチングにより得られた平均値を比較している。○印

6）　RIEPS（full）と RIEPS（null）の差を RIEPS（null）で除した値。