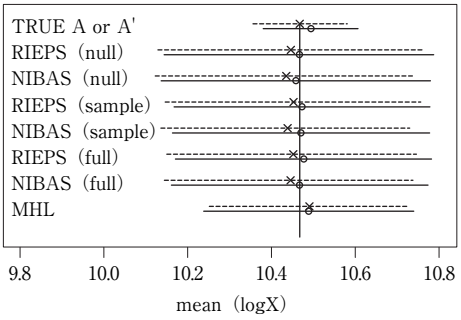
表 5-2 モデル別 logX 平均値の bias の比較（Data A，RIEPS）

Model	RIEPS（null）	RIEPS（full）	変化率	CID	prior
Z123	−0.0026(3)	−0.0015(1)	0.4220	−0.025	0.420
Z12	−0.0125(5)	−0.0105(5)	0.1567	−0.058	0.360
Z13	−0.0030(4)	−0.0017(3)	0.4231	−0.006	0.436
Z23	0.0000(1)	0.0019(4)	−53.2838	0.753	0.663
Z1	−0.0181(6)	−0.0134(6)	0.2633	−0.042	0.321
Z2	−0.0919(7)	−0.0740(7)	0.1950	0.861	0.209
Z3	−0.0004(2)	0.0017(2)	−2.9265	0.727	0.650

図 5-2 手法別 logX の平均値と信頼区間（Data A および Data A′，モデル Z123）



（注） ○印と実線が Data A による統計的マッチングの結果，×印と破線が Data A′（シミュレーション・データ）による結果を示す。

表 5-3 手法別 logX の平均値の bias（Data A および Data A′，モデル Z123）

Method	Data A	Data A′	変化率
RIEPS（null）	−0.0026	−0.0022	0.201
NIBAS（null）	−0.0033	−0.0037	−0.092
RIEPS（sample）	−0.0018	−0.0014	0.329
NIBAS（sample）	−0.0021	−0.0030	−0.314
RIEPS（full）	−0.0015	−0.0010	0.562
NIBAS（full）	−0.0023	−0.0023	0.018
MHL	−0.0001	0.0027	−0.947

（実線）と×印（破線）は，それぞれデータ A による結果とデータ A′による結果を示している。キー変数の組み合わせモデルは，Z123に限定している。

法企データによる結果（○印と実線）とシュミレーション・データによる結果（×印と破線）が非常に類似していることがわかる。シミュレーション・データ A′からも，完全データからの真値の近傍に統計的マッチングによる推定値が得られている。また表 5-3 からは，正規分布を想定して発生させたデータ A′の方がバイアスに関しては若干よい特性を示しているようにみえるが，大した差異ではない。むしろ周辺分布の特性値の推定については，法企データの対数変換によって，正規分布をシミュレートした結果と同程度のマッチング成果が得られていることがわかる。

5-2 相 関 係 数

(1) マッチング手法の比較

対数変換した X と Y の相関係数の精度を手法別に比較するために，図 5-3 には相関係数

図 5-3　手法別，相関係数および信頼区間（Data A，モデル Z123）

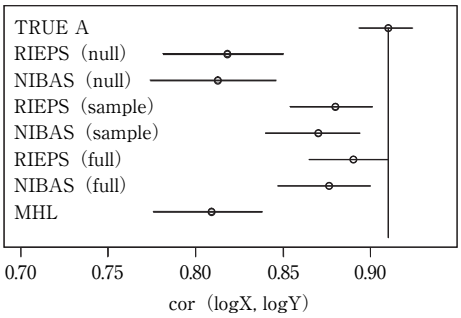


表 5-4　手法別，相関係数の bias および mse（Data A，モデル Z123）

Method	bias	mse
RIEPS（null）	−0.060	0.060
NIBAS（null）	−0.063	0.064
RIEPS（sample）	−0.020	0.020
NIBAS（sample）	−0.026	0.027
RIEPS（full）	−0.013	0.014
NIBAS（full）	−0.022	0.023
MHL	−0.066	0.066

の信頼区間を，表 5-4 には推定される相関係数（Z 変換値）の bias と mse を整理している。

図 5-3 より，マッチングにより得られた相関係数は，完全データ A からの相関係数（真値）に対して下方バイアスをもつ。RIEPS（full）のみ，95％信頼区間内に完全データ A の相関係数の真値を含んでいる。

表 5-4 からわかるように，プライアーの使用の有無にかかわらず，どのケースをとっても，NIBAS よりも，若干ではあるが RIEPS の手法が bias も mse も小さい。MHL は最も精度が悪い結果となった。また，RIEPS および NIBAS ともに，プライアーを利用した推定の方が，bias は小さくなる傾向にあり，とくに完全データから得られたプライアーを利用すると精度の改善が顕著である。

(2)　モデルの比較

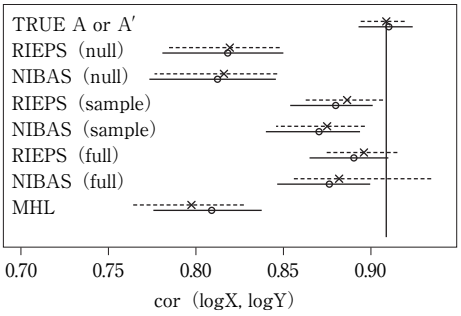
表 5-5 は，RIEPS（null）と RIEPS（full）の bias を，各モデルについて示し，カッコ内にはモデル間で比較した場合のバイアスの低さを順位で示している。さらに，プライアーによるバイアスの改善度を，RIEPS（null）と RIEPS（full）の変化率として示している。

無情報の RIEPS（null）の場合，キー変数の組み合わせによって bias は大きく異なり，最良のモデル（Z123）で −0.06，最悪のモデル（Z2）では −0.44となった。CID が小さく

表 5-5　モデル別，相関係数の bias の比較（Data A，RIEPS）

Model	RIEPS（null）	RIEPS（full）	変化率	CID	prior
Z123	−0.060(1)	−0.013(1)	0.7785	−0.025	0.420
Z12	−0.083(3)	−0.035(5)	0.5813	−0.058	0.360
Z13	−0.070(2)	−0.018(2)	0.7491	−0.006	0.436
Z23	−0.178(5)	−0.021(3)	0.8801	0.753	0.663
Z1	−0.106(4)	−0.066(6)	0.3758	−0.042	0.321
Z2	−0.448(7)	−0.407(7)	0.0910	0.861	0.209
Z3	−0.179(6)	−0.024(4)	0.8664	0.727	0.650

図 5-4 手法別，相関係数と信頼区間（Data A および Data A′，モデル Z123）



（注） ○印と実線が Data A による統計的マッチングの結果，×印と破線が Data A′（シミュレーション・データ）による結果を示す。

表 5-6 手法別，相関係数の bias（Data A および Data A′，モデル Z123）

Method	Data A	Data A′	変化率
RIEPS（null）	−0.060	−0.058	0.033
NIBAS（null）	−0.063	−0.062	0.018
RIEPS（sample）	−0.020	−0.016	0.262
NIBAS（sample）	−0.026	−0.022	0.174
RIEPS（full）	−0.013	−0.009	0.457
NIBAS（full）	−0.022	−0.017	0.339
MHL	−0.066	−0.073	−0.098

条件付従属性が低いキー変数の組み合わせの場合（Z123, Z12, Z13, Z1）にバイアスが低くなる傾向にある。すなわち，キー変数の組み合わせは，マッチング精度を規定する重要な問題であり，CID が低くなるように適切にキー変数を選択する必要がある。

プライアーを利用した RIEPS（full）ではバイアスの順位が若干入れ替わり，特に CID が高い値を示す Z23，Z3において変化率は大きく改善している。CID が高く，無情報のままでは精度が悪い状態にある場合には，質のよいプライアーを使用することで，精度向上が見込める可能性を示している。

⑶ シミュレーション・データによる結果との比較

図 5-4 は，法企データ A およびシミュレーション・データ A′を利用して，手法別に相関係数と信頼区間を算出した結果である。表 5-6 には，データ A とデータ A′について，手法別に相関係数（Z 変換値）の bias を比較したものである。

RIEPS および NIBAS では，シミュレーション・データの結果の方が完全データによる真値に近く，その傾向はプライアーを使用した方が強くなる。つまり，多変量正規分布の成立は，飛躍的に精度を改善するものではないが，プライアーによる精度改善の効果をより大きなものにしている。

5-3 回帰パラメータ

⑴ マッチング手法の比較

表 5-7 は，マッチング手法別の回帰分析の結果である。実際の法企完全データでは，Z1，Z2，Z3の係数が5％水準で有意であり，X の係数は有意ではない。これと同じ結果が得られ

表 5-7　手法別，回帰係数（Data A，モデル Z123）

Method	β(Z1)	β(Z2)	β(Z3)	β(X)
TRUE A	0.652***	0.339***	0.451***	−0.041
RIEPS（null）	0.008	0.717***	0.762***	0.258
NIBAS（null）	−0.005	0.725***	0.768***	0.267
RIEPS（sample）	0.427**	0.454**	0.606**	−0.018
NIBAS（sample）	0.361**	0.496***	0.634**	0.031
RIEPS（full）	0.503**	0.407**	0.577**	−0.068
NIBAS（full）	0.411**	0.471***	0.619**	−0.004
MHL	−0.007	0.740***	0.783***	0.267

（注）　表中の「***」，「**」，および「*」は，有意確率0.01以下，0.05以下，0.1以下をそれぞれ示す。

る手法は，補助情報を利用した RIEPS（full, sample）と NIBAS（full, sample）である。推定値の大きさも真の値にかなり近いものとなった。適切な補助情報の利用がマッチングの成否を左右することがわかる。

図 5-5 が示すように，適切な補助情報を利用するほど（null → sample → full の順），マッチング・データによる回帰係数の推定量の分布は，完全データによる回帰係数の真の値に近付いていく。その傾向は，NIBAS よりも RIEPS においていくらか顕著である。またマッチングにより Donor として接合した変数 X の回帰係数に関しては，補助情報を与えた場合，推定値の分布が完全データによる真の値（縦線の位置）の周辺に重なることになった。

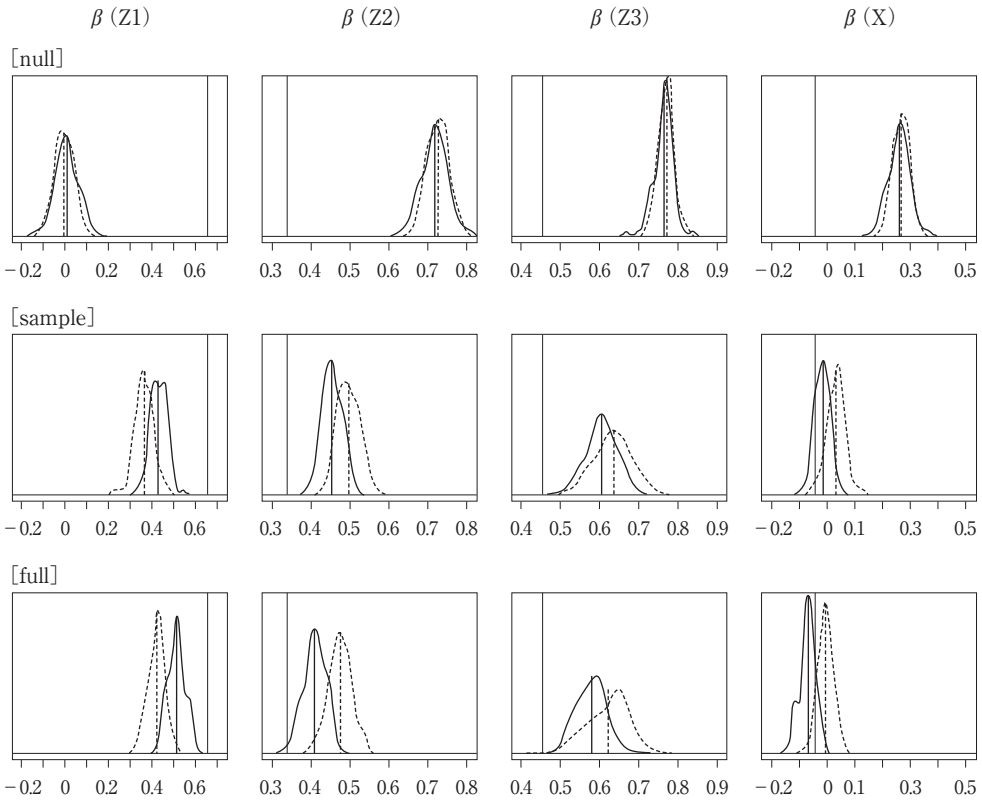
⑵　モデルの比較

キー変数の組合わせの違いによる効果をみるために，RIEPS（full）を用いて推定した回帰係数の結果を表 5-8 に整理している。パラメータの有意性が完全データ A と同様の結果となったモデルは，Z123，Z13，Z23，Z3である。CID が低くなくとも，プライアーとして適切な情報を組み込むことができれば，有意性検定という点では統計的マッチングによる分析の有用性を示唆している。また係数の推定値の大きさも完全データの特徴をうまく再現していると言ってよい。

⑶　シミュレーション・データによる結果との比較

図 5-6 は，法企データ A およびシミュレーション・データ A′を用いて，マッチングデータを回帰分析した結果（推定値の分布）を，モデル Z123に関して示している。法企データ A に対して，A′はモデルベースのマッチング手法の前提である多変量正規性をシミュレートしている。したがって 2 つを比較すれば，実データである A を用いたマッチング結果が，多変量正規性の仮定によってどの程度歪められるのか，その程度を推し量ることができる。なお，回帰パラメータの真値（縦線の位置）は当然 A，A′で異なることに注意されたい。

図 5-5 手法別，回帰係数の分布と推定値



（注） 実線は RIEPS, 破線は NIBAS による推定結果（100回分）をカーネル密度関数により推定した結果を示している。

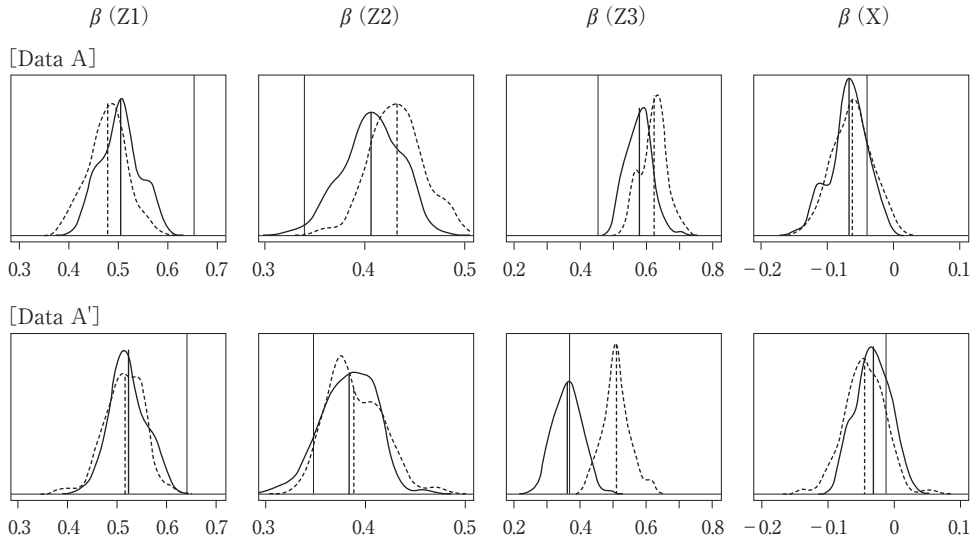
表 5-8 モデル別，回帰係数の比較（Data A，RIEPS（full））

Model	β (Z1)	β (Z2)	β (Z3)	β (X)	CID	prior
TRUE A	0.652***	0.339***	0.451***	−0.041	—	—
Z123	0.503**	0.407**	0.577**	−0.068	−0.025	0.420
Z12	0.232*	0.446**	0.514*	0.236*	−0.058	0.360
Z13	0.477**	0.432**	0.621**	−0.063	−0.006	0.436
Z23	0.619***	0.464***	0.676***	−0.249	0.753	0.663
Z1	0.136	0.538***	0.728***	0.250*	−0.042	0.321
Z2	0.013	0.719***	0.674**	0.262*	0.861	0.209
Z3	0.587***	0.491***	0.529**	−0.217	0.727	0.650

（注） 表中の「***」，「**」，および「*」は，有意確率0.01以下，0.05以下，0.1以下をそれぞれ示す。

図 5-6 の各分布（折れ線）を［Data A］と［Data A′］のように縦に比較すれば明らかなように，予想通り，A′ による推定値の分布が，A のそれに比べ，僅かではあるが真値のよい近似を与えていることがわかる。逆に A の分布の近似度の悪さは，多変量正規性の不成

図 5-6　データおよびモデル別，回帰係数の分布と推定値



（注）　実線は RIEPS での Z123モデル，破線は RIEPS での Z13モデルによる推定結果（100回分）をカーネル密度関数により推定した結果を示している。

立の程度に起因しているとみなすことができる。しかしながら，本稿の検証の枠組みにおいては，法企データ A による推定量は際立って劣るという程ではない。今回使用した変数セットについては対数化によって，推定に支障がない程度には多変量正規性の想定が満たされたと考えてよいであろう。

6．おわりに

本稿は，日本における企業データセットの情報拡張の方法として統計的マッチングがどのように有効であるのか，それを企業データの代表ともいえる法人企業統計調査（財務省）の調査票情報を利用して，具体的に検証しようという試みである。2000年第Ⅰ四半期の製造業・大企業を対象に，従業員数，平均給与，資本金をキー変数として，Recipient 側の売上高変数に対して Donor 側の資産合計情報を接合する実験を行った。

実際には，先行する欧米では近年とくに存在感を増しつつあるモデルベースのパラメトリック・マッチングの手法を中心に比較を試みた。通常の回帰補定法からマッチング技法として定着した RIEPS，およびそのベイズモデル・バージョンともいえる NIBAS を取り上げている。参考数値として，従来型のノンパラメトリックなマッチング手法の代表としてマハラノビス距離関数を用いたマッチング（MHL）も試しているが，法人企業統計調査を用いた今回の検証の枠組みの中では，MHL の結果が最も悪く，対照的にモデルベースの手法であ

るRIEPS，およびNIBASが好結果を示している。またベイズタイプのNIBASよりRIEPSの方が総合的には若干パフォーマンスが高めに現れている。

NIBASやRIEPSタイプのモデリックなアプローチの良さは，個体レベルの確率変動の組み込みが比較的容易であること，補助情報を明示的にモデルに導入し評価可能である点にあるが，これに対して分布形を仮定せざるを得ないことが欠点となる。多変量正規分布を仮定するNIBASとRIEPSについては，法人企業統計の各変数の対数化によって，これらのモデルでマッチングを実行する条件が実用的な程度には整うことを検証結果は示している。また，キー変数の組合せや補助情報のクオリティーの違いによってマッチング・パフォーマンスは異なることから，適切なキー変数を採用し，質の良い補助情報を確保する方策の有無が，これらのテクニックの成否には決定的といえる。

それにしても個人や世帯とは比較にならない異質性の集団ともいえる製造業・大企業グループに対して，条件付きではあれ，相関や回帰分析というレベルではマッチング・データによる解析の有効性，あるいは可能性が検証された意義は大きい。ノンパラメトリックな手法とは異なるモデルベースのアプローチ特性がそれを支えている。前者では，距離関数が近い類似した（一般には異なる）企業同士の実データを接合するが，後者では変数間の共分散特性から発生させた推定値を補定する。企業データにおける相関や回帰といった目標レベルには，このようなマッチング・ポリシーが適合的であると考えてよいであろう。しかし，マッチングの目標レベルを上げ，より汎用性のあるマッチングデータを作製する場合はどうであろうか。そこでは，異質ではあっても実データを接合するノンパラメトリックな特性も活かしたアプローチが不可欠となる。ノンパラメトリック＋パラメトリック混合型マッチングモデルについては次稿の課題としたい。

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；企業の業績および財務内容と賃金構造の関係に関する計量経済分析」（研究代表者：上智大学　出島敬久，平成24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～2011年10-12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。

参考文献

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，175-210ページ。

井出満（2000）「個別データのリンケージに関する研究」『大阪産業大学経済論集』第１巻第２号，1-6ページ。

伊藤研一・道明義弘・井澤裕司（1999）「日・米・加産業（業種）別自己資本経常利益率規定要因の推

計―企業財務分析データにもとづくパネルデータ分析―」『立命館経済学』第48巻第１号，7-33ページ。

栗原由紀子（2012）「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』第43号，中央大学経済研究所，489-551ページ。

栗原由紀子（2012）『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.35，1-38ページ。

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

Goel, P. K. & T. Ramalingam (1980), *The Matching Methodology: Some Statistical Properties*, Springer-Verlag.

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

Little, R. J. A. & D. B. Rubin (2002), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics.

Ranser, A., J. R. Frick, and M. M. Grabka (2011), "Extending the Empirical Basis for Wealth Inequality Research Using Statistical Matching of Administrative and Survey Data," *SOEPpapers on Multidisciplinary Panel Data Research*, German Institute for Economic Research Berlin, pp. 1-42.

Rässler, S. (2002), *Statistical Matching*, Springer.

〈資　料〉

付表 1-1　Data A の平均値（周辺分布特性）

Model	Mean						
	logX	lower	upper	W	B	bias	mse
TRUE A	10.489	10.376	10.602	1.657	—	—	—
[Z123]							
RIEPS(null)	10.461	10.140	10.783	2.57	0.026	−0.0026	0.0035
NIBAS(null)	10.454	10.133	10.775	2.470	0.025	−0.0033	0.0050
RIEPS(sample)	10.470	10.163	10.777	2.357	0.024	−0.0018	0.0026
NIBAS(sample)	10.467	10.158	10.777	2.326	0.023	−0.0021	0.0037
RIEPS(full)	10.473	10.167	10.779	2.331	0.023	−0.0015	0.0025
NIBAS(full)	10.464	10.158	10.771	2.305	0.023	−0.0023	0.0036
MHL	10.487	10.237	10.737	1.588	0.016	−0.0001	0.0001
[Z12]							
RIEPS(null)	10.358	9.953	10.762	4.023	0.040	−0.0125	0.0130
NIBAS(null)	10.410	10.034	10.787	3.336	0.033	−0.0075	0.0090
RIEPS(sample)	10.374	10.001	10.746	3.387	0.034	−0.0110	0.0115
NIBAS(sample)	10.406	10.045	10.768	3.045	0.030	−0.0079	0.0093
RIEPS(full)	10.378	10.008	10.748	3.343	0.033	−0.0105	0.0111
NIBAS(full)	10.406	10.046	10.767	3.007	0.030	−0.0078	0.0094
MHL	10.459	10.204	10.714	1.651	0.017	−0.0028	0.0028
[Z13]							
RIEPS(null)	10.457	10.130	10.785	2.657	0.027	−0.0030	0.0039
NIBAS(null)	10.448	10.124	10.772	2.530	0.025	−0.0039	0.0053
RIEPS(sample)	10.470	10.161	10.779	2.374	0.024	−0.0018	0.0028
NIBAS(sample)	10.463	10.152	10.773	2.326	0.023	−0.0025	0.0042
RIEPS(full)	10.471	10.165	10.776	2.338	0.023	−0.0017	0.0025
NIBAS(full)	10.467	10.158	10.776	2.306	0.023	−0.0020	0.0039
MHL	10.494	10.237	10.752	1.681	0.017	0.0005	0.0006
[Z23]							
RIEPS(null)	10.489	10.149	10.830	2.826	0.028	0.0000	0.0033
NIBAS(null)	10.475	10.135	10.815	2.685	0.027	−0.0013	0.0049
RIEPS(sample)	10.507	10.205	10.809	2.266	0.023	0.0018	0.0028
NIBAS(sample)	10.495	10.176	10.814	2.329	0.023	0.0006	0.0049
RIEPS(full)	10.509	10.210	10.808	2.230	0.022	0.0019	0.0027
NIBAS(full)	10.512	10.197	10.827	2.286	0.023	0.0022	0.0051
MHL	10.544	10.273	10.816	1.877	0.019	0.0053	0.0053
[Z1]							
RIEPS(null)	10.299	9.851	10.746	4.856	0.049	−0.0181	0.0187
NIBAS(null)	10.406	10.008	10.804	3.717	0.037	−0.0079	0.0095
RIEPS(sample)	10.333	9.926	10.740	4.024	0.040	−0.0149	0.0154
NIBAS(sample)	10.396	10.010	10.782	3.366	0.034	−0.0089	0.0108
RIEPS(full)	10.349	9.943	10.755	3.984	0.040	−0.0134	0.0140
NIBAS(full)	10.393	10.014	10.773	3.323	0.033	−0.0091	0.0106
MHL	10.439	10.184	10.693	1.647	0.016	−0.0049	0.0049
[Z2]							
RIEPS(null)	9.525	8.632	10.418	17.913	0.179	−0.0919	0.0931