共同DBにおける欠損値解析法の利用

今井健太郎

日本リスク・データ・バンク株式会社　データベース統括部

Missing Data Analysis with Database Consortium

Imai Kentaro

Database Management Department

The Risk Data Bank of Japan, Limited

要旨:

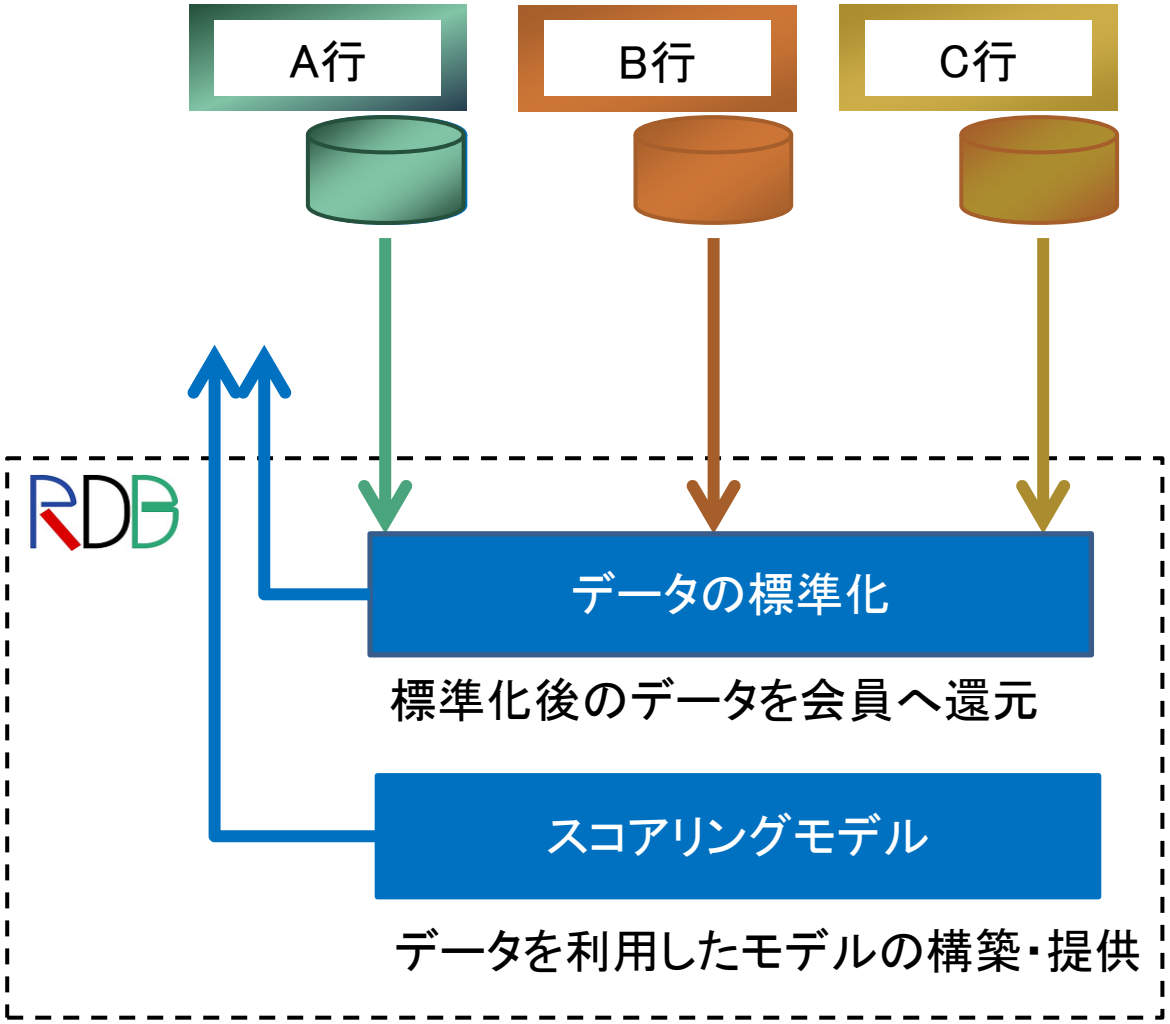
会員がデータを持ち寄る共同型DBにおいて、全会員が同じ項目を収集しているわけではなく、最大公約的なDBとなる。会員が独自で集めている項目を利用する際には、本報告の欠測値解析の1つである多重代入法が有効である。

キーワード: 欠損解析, 欠測解析, 多重代入法, missing data , proc mi, proc mianalyze, multiple imputation

□ 弊社の業務内容

全国65の金融機関が
参加するデータベース・
コンソーシアム

- 事業法人データ
- 個人事業者データ
- 債権回収データ
- オペレーショナルリスクデータ



信用スコアリング・モデルとは

財務の状況から、将来のデフォルト（債務不履行）の発生を予測するモデル

融資貸出審査、引当金の計算などを定量的行うことが可能

弊社ではlogisticモデルを利用。将来のデフォルト率pを財務指標xで回帰

$\log(p_i/(1-p_i))=b_0+b_1*x_{1i}+b2*x_{2i}$

```
proc logistic data = mydata;
      model status=x1 x2;
run;
```

債務者間の**序列**を知りたい　：　***格付けモデル***

興味：債務者iと債務者jのどちらが優良か？

債務者別に**デフォルト率**を知りたい　：　***デフォルト率推計モデル***

興味：債務者iの1年以内のデフォルト確率は？

共同データベースの構造

参加会員によってDB構造が異なるため、全会員が共通して保持する項目を使う方が好ましい(”最大公約数”的なデータベース)。

	売上高	自己資本	総資産	中間配当	経営者の車種		創業年数
A行	Available	Available	Available	Available	NA		Available
B行	Avaliable	Available	Available	NA	Available	・・・	NA
C行	Avaliable	Avaliable	Available	Available	NA		NA
Z社	Available	Available	Available	NA	NA		NA

使える変数が限定されても、大規模データを使うメリットの方が大きい

□ 会員ニーズ

当行独自で集めている情報を活かすことで、より精度の高いモデルを構築できないか？

分析データの構造

	共通項目	車種
B行データ	Obs	Obs
共通データ	Obs	Miss

B行データでだけでは、データ件数が乏しい

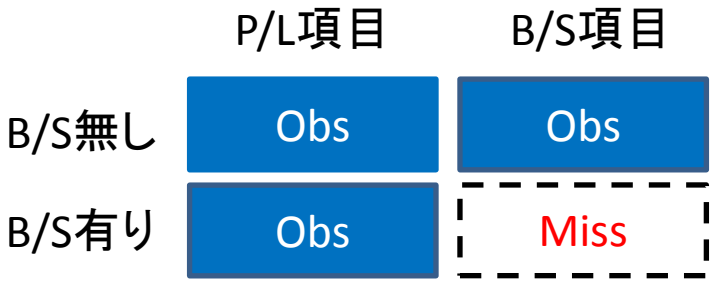
共同データでは、「車種」が集計対象外のため欠損値(Miss)になっている

有効な情報を持つ少数のデータセット、一部情報が欠落している大量データセット、という構造

※類似ケース
・事業法人の「連結,単体」「個人事業者のB/S,P/L」

今回の分析デザイン

分析データ：観測値と欠損値の関係が類似した個人事業者の財務データを利用



個人事業者は貸借対照表の作成は任意
RDBデータベースでは約50%が未作成

P/L項目とB/S項目を使って信用スコアリングモデルを構築する

- 応答変数：決算月から1年以内のデフォルト有無
  デフォルト発生(status=0)、正常な債務先(status=1)

- 説明変数：P/L財務項目から4指標、B/S項目から1指標
  P/Lの例：売上高営業利益率、 売上高支払利子率、・・・
  B/Sの例：自己資本比率

※実際に使用した項目はノウハウのため非公開

□ 欠損値の発生メカニズム

I. MCAR (Missing Complete At Random)

欠損は完全にランダムに発生

II. MAR (Missing At Random)

欠損は観測されたデータのみに依存する

自己資本比率の欠損になりやすさは売上高規模に依存する
（事業規模が小さい先はB/Sを作成しない）

III. NMAR (Non-MAR, non-ignorable)

欠損は欠損値に依存する

自己資本比率の欠損になりやすさは自己資本比率に依存する
（自己資本比率が悪い先はB/Sを作成しない）

□ 欠損値への対処方法の代表例

A) Complete Case Analsys, Listwise Deletion,

欠損データを削除して分析。慣用的な手法

削除でN数が減ってしまう。MCARの仮定が必要。

B) Single Imputation (単一値を代入)

欠損値に何らかの値（例:平均値）を代入して分析

C) Multiple Imputation(多重代入)

MIプロシージャ, MIANALYZEプロシージャ

欠損値に複数の異なる値を代入した複数データで分析

他：マッチング， FIML

MIとMIANALYZEプロシジャを用いた分析手順の概要

1. MIにより代入された複数のデータセットを作成

```
proc mi data=missdata out=impdata  nimpute=5;
   monotone reg(bx1=px6 px7 px8 px9 px10 px11);
   var px6 px7 px8 px9 px10 px11 bx1;
run;
```

欠損がある変数bx1について、欠損発生のモデルを指定
nimputeは発生させるデータセット数

2. データセット毎に分析の実行

```
proc logistic data=impdata  out=miout covout;
   model status=px1 px2 px3 px4 bx1;
   by _imputation_;
run;
```

発生させたデータセットごとに回帰結果がmioutに出力される

3. MIANALYZEで複数の結果を1つに統合

```
proc mianalyze data=miout mult;
   var intercept px1 px2 px3 px4 bx1;
run;
```

mioutの結果を統合することで、1つの推計結果を得ることができる

次項以降に詳細

Step0. B/S変数とP/L変数の関係を調査

観測可能データで、欠損指標bx1を重回帰

```
proc reg data=missdata;
   model bx1= px01 – px20/selection=stepwise;
run;
```

変数	パラメータ推定値	標準誤差	Type II 平方和	F 値	Pr > F
Intercept	-316.50673	18.71987	629366242	285.86	<.0001
px01	-0.00003084	0.00000398	132289591	60.09	<.0001
px04	0.00079243	0.00038516	9319432	4.23	0.0397
px06	32608	377.31852	16443044548	7468.60	<.0001
px07	595.56637	56.66627	243194700	110.46	<.0001
px08	-577.23088	56.45728	230145366	104.53	<.0001
px09	6178.84755	216.62496	1791187507	813.58	<.0001
px10	-17.33466	1.53423	281055905	127.66	<.0001
px11	521.09732	63.11831	150061748	68.16	<.0001
px12	-521.99240	63.11313	150602446	68.41	<.0001
px13	0.50590	0.27746	7319534	3.32	0.0683
px15	0.00404	0.00201	8926561	4.05	0.0441
px19	-0.00111	0.00031942	26485362	12.03	0.0005

bx1推計のために変数を5個(px06-px10)採用して、MIプロシージャの欠損値推計モデルに利用

Step1. Proc MI

債務者番号	ステータス	BS有無	px6	px7	px8	px9	bx1
1	0	有り	0.02	0.06	6.94	13.15	16.40
2	0	無し	0.07	0.38	12.01	4.73	.
3	1	無し	0.01	0.34	2.43	15.10	.

B/S未作成の場合指標は欠損

```
proc mi data=missdata out=impdata  nimpute=3;
   monotone reg(bx1 = px6 px7 px8 px9 px10);
   var px6 px7 px8 px9 px10 bx1;
run;
```

imputation	債務者番号	ステータス	BS有無	px6	px7	px8	px9	bx1
1	1	0	有り	0.02	0.06	6.94	13.15	16.40
1	2	0	無し	0.07	0.38	12.01	4.73	54.74
1	3	1	無し	0.01	0.34	2.43	15.10	44.78
2	1	0	有り	0.02	0.06	6.94	13.15	16.40
2	2	0	無し	0.07	0.38	12.01	4.73	14.55
2	3	1	無し	0.01	0.34	2.43	15.10	51.02
3	1	0	有り	0.02	0.06	6.94	13.15	16.40
3	2	0	無し	0.07	0.38	12.01	4.73	4.57
3	3	1	無し	0.01	0.34	2.43	15.10	25.50

同じ値

B/S変数の欠損値が穴埋めされるが、値はimputation毎に異なる

Step2. Proc Logistic

```
proc logistic data=impdata out=miout covout;
   model status(event='1') =  px1  px2  px3  px4  bx1;
   by _imputation_;
run;
```

by _imputation_ でデータセット毎にモデル推計

デフォルト(status=0)or非デフォルト(status=1)を予測するための説明変数は、欠損値穴埋めの変数と異なってよい。

Imputation Number	Intercept	bx1	px1	px2	px3	px4
1	-0.604	0.019	1.926	-4.004	1.10E-04	-0.027
2	-0.621	0.020	1.908	-4.038	1.12E-04	-0.028
3	-0.603	0.020	1.886	-3.889	1.06E-04	-0.029

_imputation_毎に推計結果が異なる

Step3. Proc MIAnalyze

代入毎に得られたパラメータ推計値Q_iと標準誤差w_iを、MIAnalyzeによって1つの結果にまとめる。mは代入したデータセットの回数

係数

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} Q_i$$

標準誤差

$$S.E. = \sqrt{\frac{1}{m}\sum_{i=1}^{m} w_i + \left(1+\frac{1}{m}\right)\left(\frac{1}{m-1}\right)\sum_{i=1}^{m}(Q_i-\bar{Q})^2}$$

```
proc mianalyze data=miout mult;
   var intercept px1 px2 px3 px4 bx5;
run;
```

MIanalyzeが計算してくれる

□ 本分析で検証したこと①

欠損値を含めたデータは使うにはどの方法が適当か?

5種類の方法を比較

- 非欠損データ=B/S有り100件（内デフォルト50件、非デフォルト50件）
- 欠損データ=B/S無しデータ（内デフォルト50件、非デフォルト50件）

の例

■ Method1 : 説明変数にB/S指標は使わず、全データを学習データに使用 (PL)

	デフォ	非デフォ	P/L指標	B/S指標
B/S無し	50	50	入力値	使わない
B/S有り	50	50		

■ Method2 : 説明変数にB/S指標も使うが、B/S無し先は学習データに使わない (CCA)

	デフォ	非デフォ	P/L指標	B/S指標
B/S無し				
B/S有り	50	50	入力値	入力値

■ Method3 : 説明変数にB/S指標も使い、B/S無し先は観測値の平均値を代入 (SI mean)

	デフォ	非デフォ	P/L指標	B/S指標
B/S無し	50	50	入力値	**平均値**
B/S有り	50	50	入力値	入力値

■ Method4 : 説明変数にB/S指標も使い、B/S無し先は観測値の最悪値を代入 (SI worst)

	デフォ	非デフォ	P/L指標	B/S指標
B/S無し	50	50	入力値	**最悪値**
B/S有り	50	50	入力値	入力値

■ Method5 : 説明変数にB/S指標も使い、B/S無し先は多重代入法を適用 (MI)

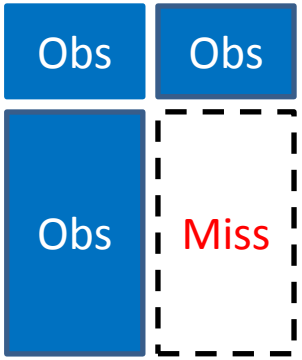
	デフォ	非デフォ	P/L指標	B/S指標	代入番号
B/S無し	50	50	入力値	**代入値1**	1
B/S有り	50	50	入力値	入力値	
B/S無し	50	50	入力値	**代入値2**	2
B/S有り	50	50	入力値	入力値	
B/S無し	50	50	入力値	**代入値3**	3
B/S有り	50	50	入力値	入力値	

□ 本分析で検証したこと②

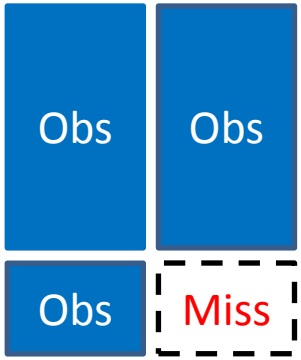
非欠損データと欠損データのN数が変わると、欠損値の対処方法に差が出るか?

非欠損データ:少
欠損データ:少

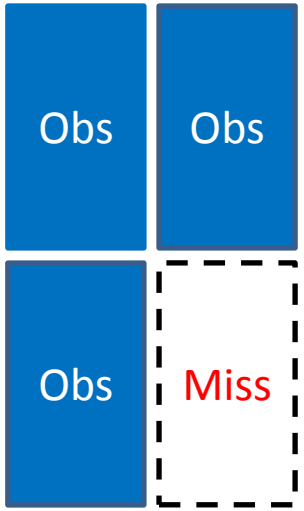
非欠損データ:少
欠損データ:多



非欠損データ:多
欠損データ:少



非欠損データ:多
欠損データ:多

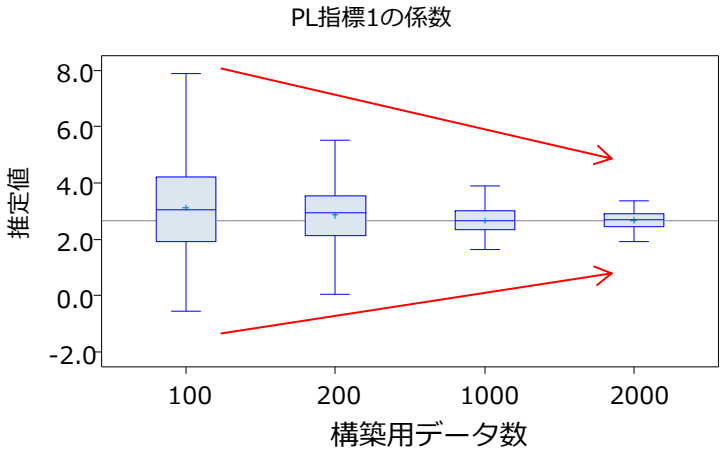


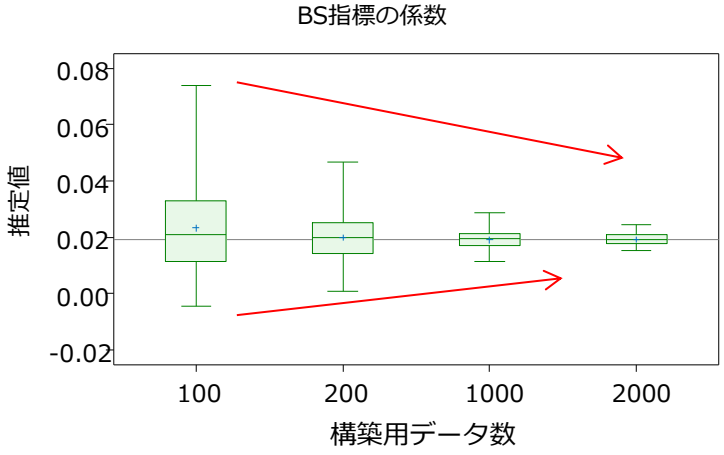
□ 手法CCAによる分析結果

CCA（BS有りデータに限定）による推計結果

	N=100(DF=50,ND=50)		N=200(DF=100,ND=100)		N=1000(DF=500,ND=500)		N=2000(DF=1000,ND=1000)	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	-0.8448	0.4795	-0.3274	0.3187	-0.5192 **	0.1480	-0.6481 **	0.1035
PL指標1	3.8245	1.6088	0.4848	1.0747	2.6248 **	0.4458	2.7277 **	0.3251
PL指標2	17.3331	14.0481	-14.8906	8.9485	-16.6700 **	4.0238	-18.5724 **	2.9398
PL指標3	0.0160 *	0.0063	-0.0004	0.0006	0.0010 **	0.0003	0.0009 **	0.0003
PL指標4	-0.0827	0.0418	0.0131	0.0236	-0.0164	0.0101	-0.0092	0.0075
BS指標	-0.0088	0.0111	0.0251 **	0.0078	0.0155 **	0.0032	0.0197 **	0.0023

サンプリングを繰り返し、CCAによる推計を100回(Iteration=100)行った結果

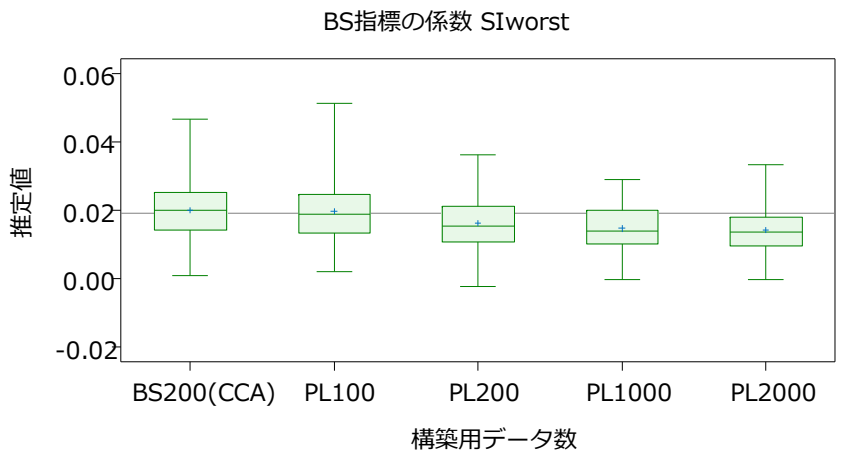


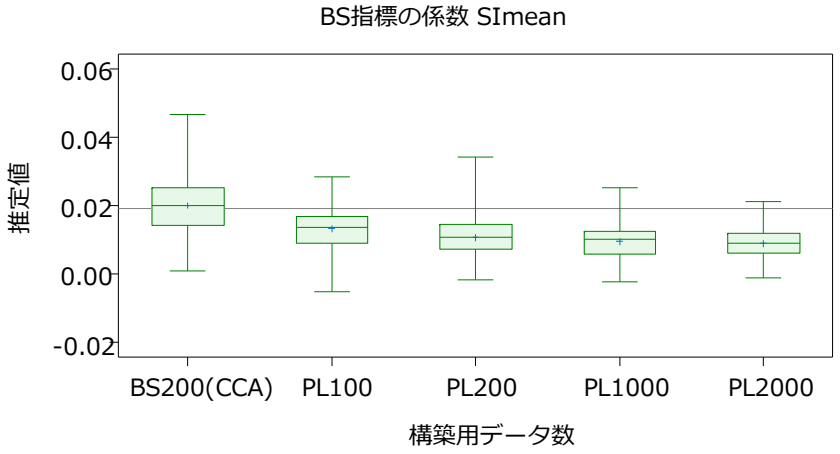


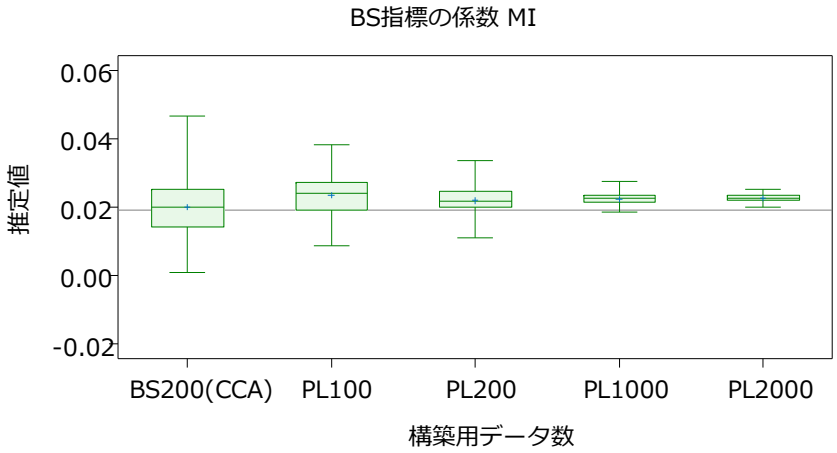
使用データが増えれば係数は安定する

□ 欠損値を補完することで係数は安定するか？

BS有りデータをN=100に固定し、BS無しデータを各手法に基づき欠損補完して、モデル構築を100回実施



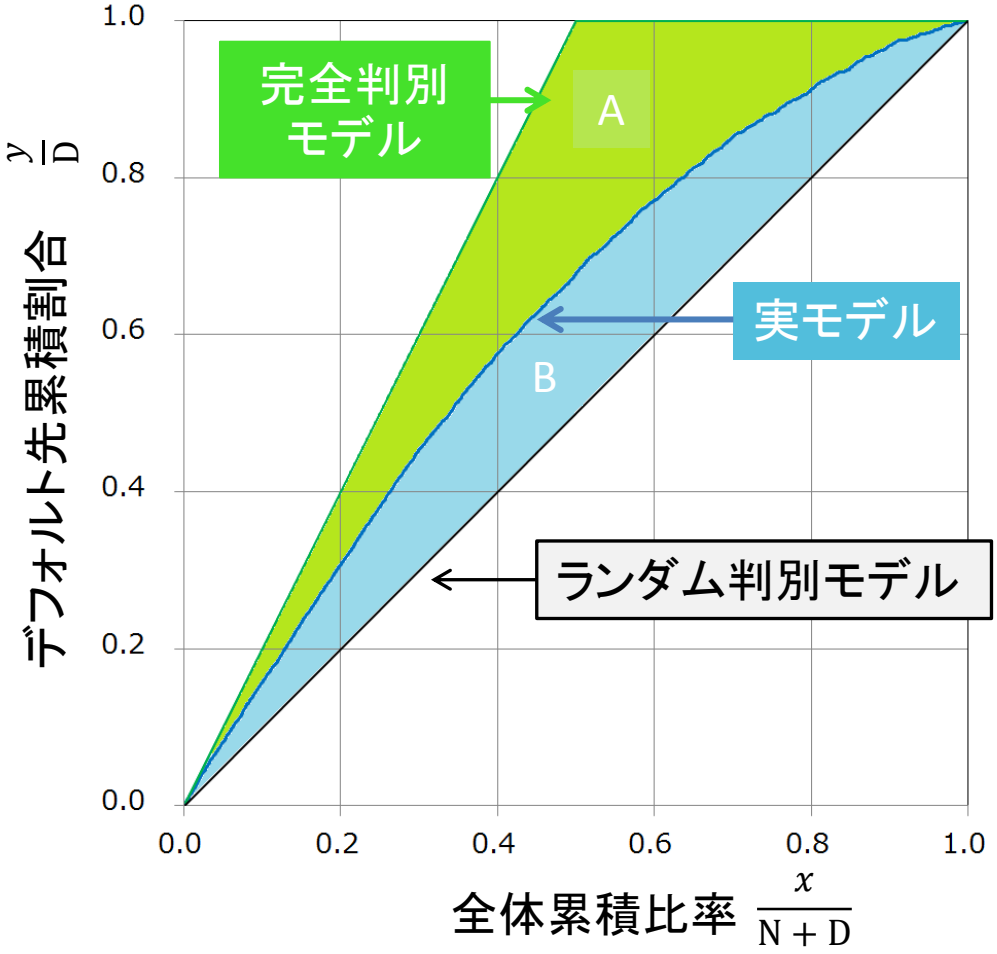




MI法の推計結果が安定

□ 推計モデルのパフォーマンス評価方法について

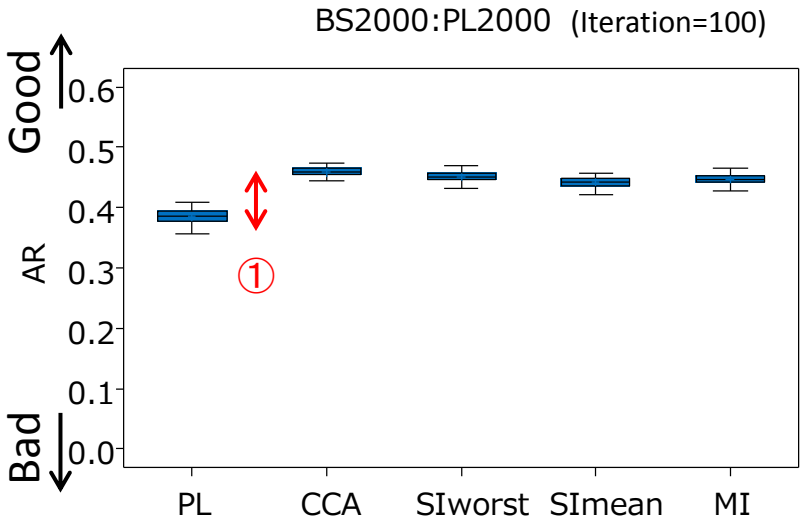
信用リスクモデルの序列推計性能の評価はARで行う



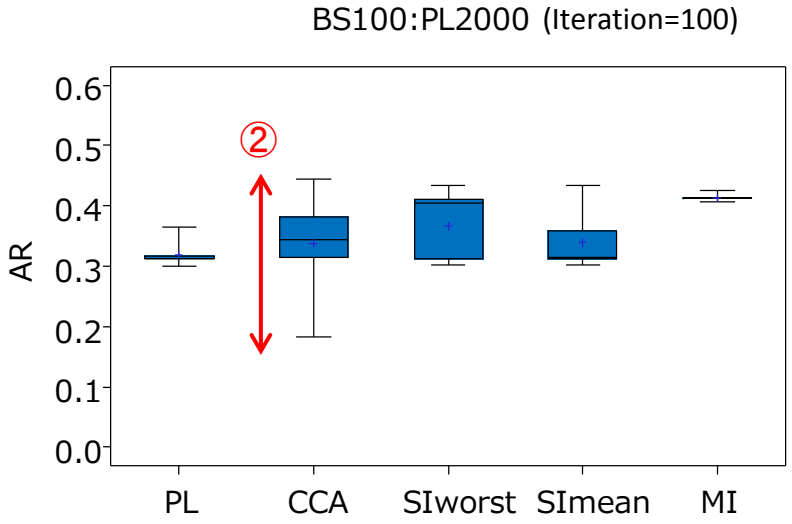
AR=Accuracy Ratio
対象となる貸出先のうち、デフォルト先の件数をD、非デフォルト先の件数をN とし、貸出先全てにスコアをつけて、スコアの悪い順（モデルがより「デフォルトしやすい」と評価する順番）に並べ、悪いほうからx 番目までの貸出先を取り出したときに、その中にデフォルト先がy件含まれていたとすると、両者の組み合わせ（x,y）は貸出先の数（D+N）だけ存在する。このとき、横軸をx の全体に対する比率（x/(N+D)）を累積全体比率、縦軸をy のデフォルト先全体に対する比率（y/D）を累積デフォルト先比率としてグラフに表したのが左図である。

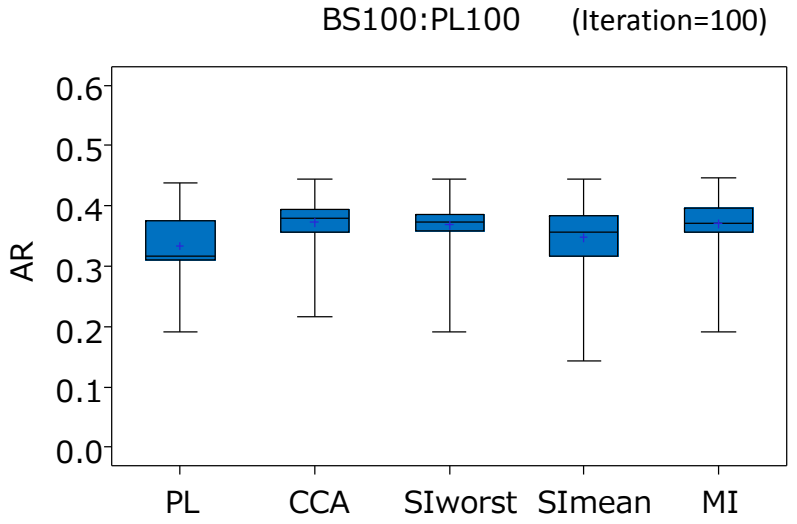
ROC曲線のAUC（Area Under the Curve）、ジニ係数、Somers'dと本質的に同じもの

□ 序列推計性能　アウトサンプルデータ(N=7220)

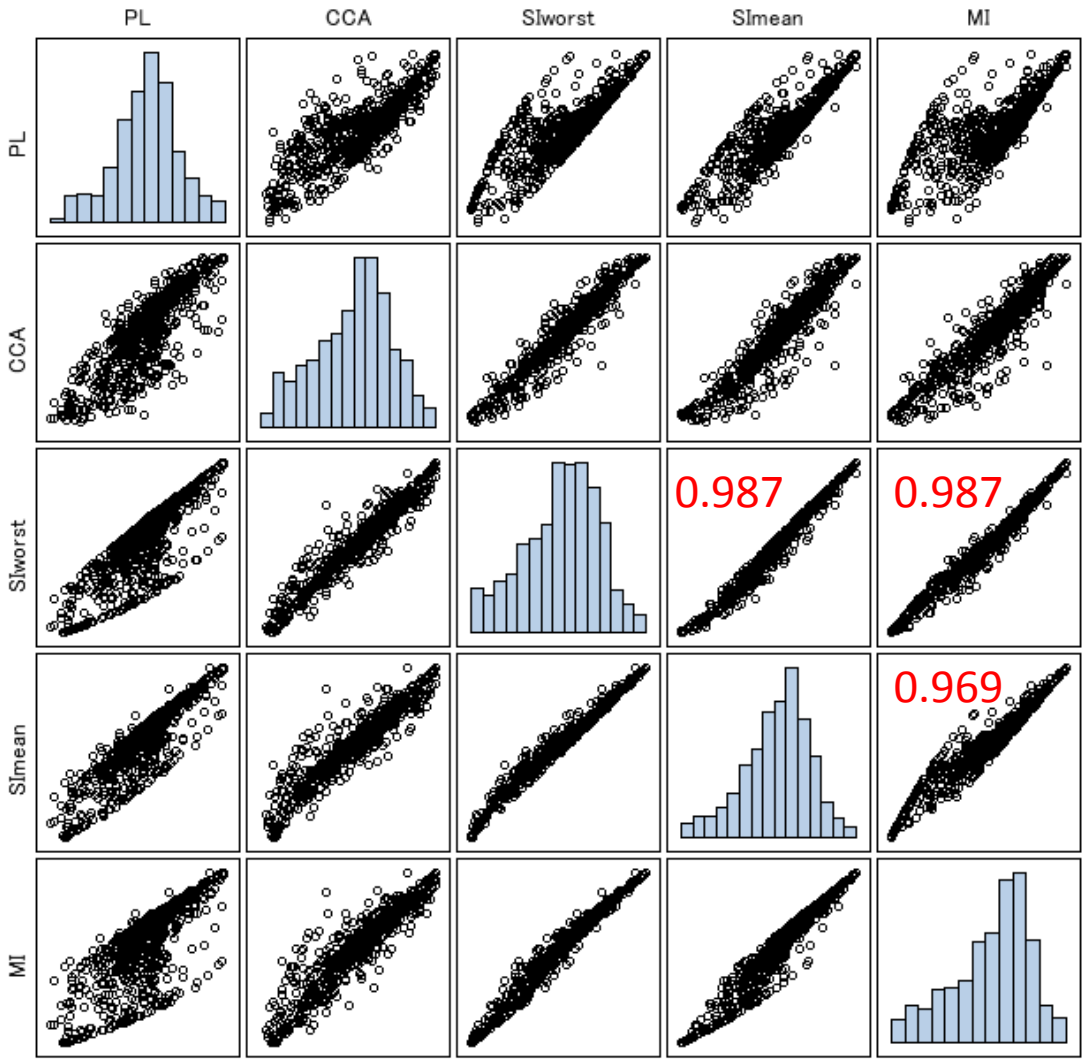


- 十分なデータ量がある場合、手法間で性能に大きな差は無く、BS情報を使用/未使用で差(①)が出る。
- B/S情報が少ない場合、CCAではモデル性能にバラツキが生じる(②)。悪いモデルを選択する可能性が高まる。
- P/L情報も少ないと、MIでもバラツキが発生する。





□ 各手法によって推計されたデフォルト確率の相関



左図はBS2000,PL2000データを使用して構築した各モデルが推計したprobability(デフォルト確率=PD)の散布図。

SIworst,SImean,MIで推計されたPDの相関は非常に高く、前項の序列性能=ARに差がでなかった。

モデルに使用した説明変数を固定したため、序列に差がつかなかった可能性がある。

□ 水準に注意

推計されたPDの要約統計量

要約統計量

変数	N	平均	標準偏差	中央値	最小値	最大値
PL	7220	0.50113	0.11710	0.51250	0.16936	0.79084
CCA	7220	0.46623	0.19316	0.49532	0.02302	0.85452
SIworst	7220	0.45828	0.15656	0.48472	0.06055	0.79892
SImean	7220	0.53250	0.13827	0.55541	0.14050	0.83366
MI	7220	0.50422	0.16704	0.54140	0.08146	0.82808

◘ 結論

- ☞モデル構築に使えるデータ数が十分であれば、慣用的な手法であるComplete Case Analysis ＝欠損レコードの削除を選択すれば良い(MCARの仮定が成り立つ必要あり）
- ☞MI法のメリットは、完全観測されたデータが少なく、かつ欠損データが大量にある場合に特に良好な結果であり、悪いモデルを選択する可能性が減少する
- ☞データ件数を確保を目的に、欠損値に単一値の代入を実施しても、大きなメリットはない。また推計値の水準にバイアスが発生

当行独自で集めている情報を活かすことで、より精度の高いモデルを構築できないか？

共同データベースと多重代入法を使うと、上手くいきますよ

□ 展望

☞ 指標探索の段階で、MI法を用いることで、少数データでも有効な変数の検出を行うことが可能と思われる。これによってモデル性能の向上は期待できる。

SAS9.2ではlogisticプロシージャにselectionを指定した場合は、MIAnalyzeに受け渡すことができなかった。

◘ 参考資料

[1]岩崎学「不完全データの統計解析」(2002)、エコノミスト社

[2]狩野裕「結測値データ解析の意味と有効性」(2013)、http://www.sigmath.es.osaka-u.ac.jp/~kano/research/seminar/others/KSPmissingWEB.pdf

[3]星野崇宏「調査観察データの統計科学　因果推論・選択バイアス・データ融合」(2009)、岩波書店

[4]村山航「欠損データ解析　完全情報最尤推定法と多重代入法」(2011),http://www4.ocn.ne.jp/~murakou/missing_data.pdf

[5]P.D.Allison「Missing Data」(2001),Sage Publications, Inc.

[6]Rubin,D.B.「Multiple Imputation for Nonresponse in Surveys」(1987),New York: Wiley