ロジットモデル構築におけるWeight of Evidenceを用いた変数変換と欠損値処理方法の提案

木村和央
株式会社金融工学研究所 技術統括部

Proposals for Variable Transformation and Missing Value Complementing Method using Weight of Evidence in Logit Model Building

Kazuo Kimura
Technology Management Div., Financial Technology Research Institute Inc.

要旨：

本稿では，ソブリン（国）のデフォルトと格付の推計を例に，効率的なロジットモデル構築を図る手段として，Weight of Evidence（WoE）のさらなる活用法を提案した.

キーワード：ソブリン，外れ値処理，Yeo-Johnson変換，欠損値補完，重回帰モデル，同時推計，ブートストラップ，NLP，FMM，LOGISTIC，SURVEYSELECT，GMAPプロシジャ

本稿の内容は筆者に属し，所属組織の見解ではない.

1. はじめに

ロジットモデルは，金融機関実務のなかでも広く活用

コストをかけず，スキルによらず，一定精度のモデルを得たいという要請

大勢待（2008，SASユーザー会）は，Weight of Evidence（WoE）を説明変数に用いて，効率的なデフォルト確率モデルの構築方法を提案

本稿では，WoE（あるいは対数オッズ，Zスコア）をソブリン（国）のデフォルト確率および格付該当確率モデルの構築に利用する方法を提案

1. はじめに （WoE，対数オッズ，Zスコアの関係について）

架空の例：GDP成長率（x）	全体①	非デフォルト先数②	デフォルト先数③	デフォルト率④	対数オッズ⑤	WoE⑥
$x \leq -3\%$	500	480	20	4.00%	-3.178	0.714
$-3\% < x \leq 0\%$	900	882	18	2.00%	-3.892	0.000
$0\% < x \leq 3\%$	800	788	12	1.50%	-4.185	-0.293
$3\% < x$	600	594	6	1.00%	-4.595	-0.703
全体	2,800	2,744	56	2.00%	-3.892	0.000

対数オッズ⑤＝ln{④／（1－④）}，　WoE⑥＝対数オッズ⑤－全体の⑤

上表をもとに，単変数にてロジットモデルを推計すると，

p = 1／{1+exp（-z）}

z = 対数オッズ⑤ = 全体の⑤ + WoE⑥

WoE：部分集合の対数オッズ（Zスコア）と全体の対数オッズ（Zスコア）の差

2. データと分析の枠組み

分析対象：世界銀行加盟国（188），2001-2014

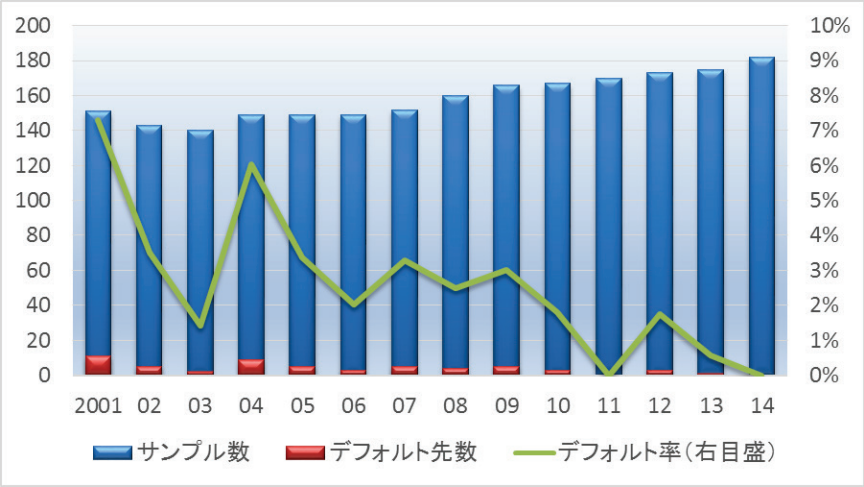
経済関連指標データ：
・World Development Indicators（WDI）

ガバナンス指標データ：
・The Worldwide Governance Indicators（WGI）

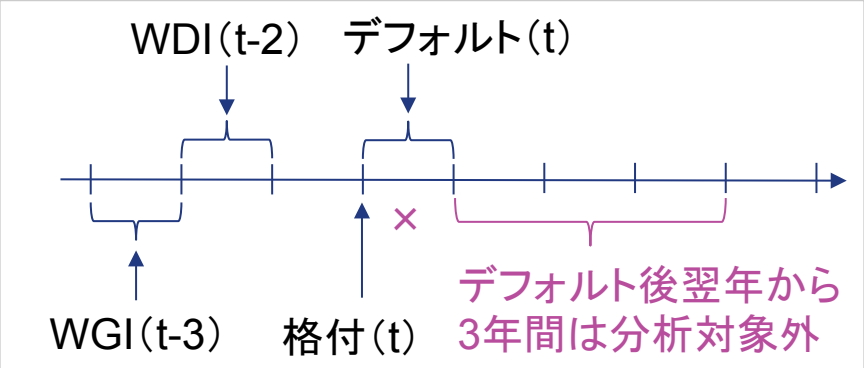
対外債務残高データ（WDIの欠損値を補完）：
・Quarterly External Debt Statistics（QEDS）
・The World Factbook from CIA

デフォルト認定データ：
・主要債権国会議（パリクラブ）へ持ち込まれた年
・債務リストラ年（Cruces and Trebesch，2013）

格付データ：
・格付投資情報センター（R&I）の前年12月末格付

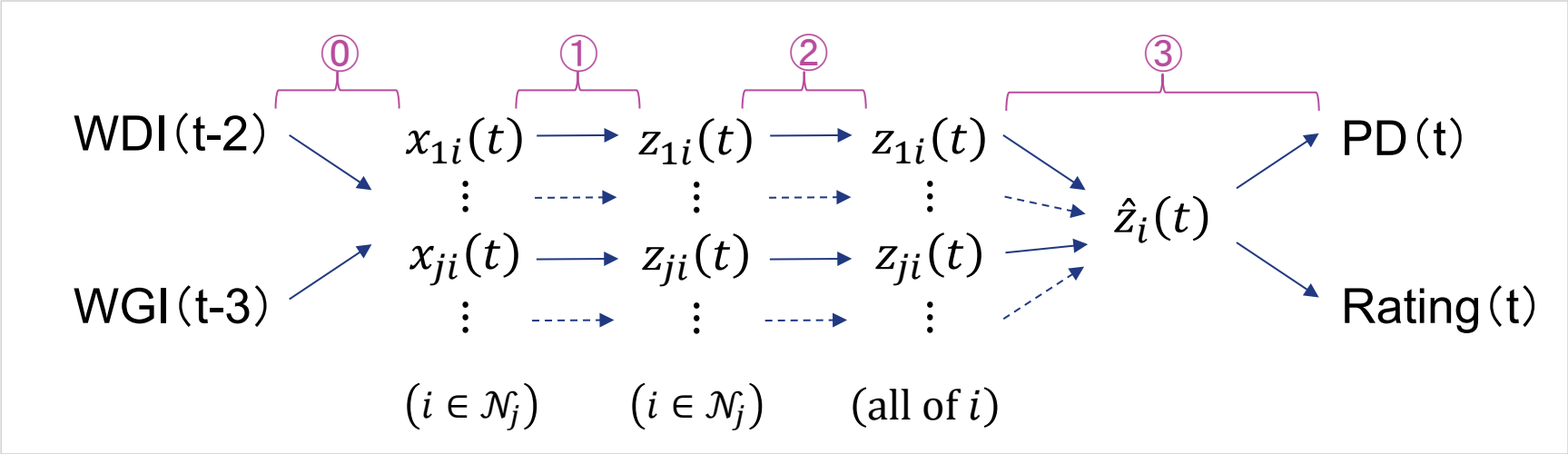


デフォルト率の時系列遷移（2001-2014）



分析の枠組みと分析対象期間の関係

2. データと分析の枠組み（モデル構築のプロセス）



⓪（変数加工）2期前のWDI，3期前のWGIデータから，当期を評価するための指標を作成．
フロー関連指標は，過去2年算術ないし3年幾何平均を実施．

①（変数変換）個々の指標（原数値）に対し，適当な外れ値処理とYeo-Johnson変換パラメータを選択し，Zスコアへ変換（UNIVARIATE，NLPプロシジャ）．

②（欠損処理）個々の指標（Zスコア）の欠損値部分について，欠損値パターンに応じた重回帰モデルを繰り返し用いた単一代入法により補完処理（FMMプロシジャ）．

③（同時推計）個々の指標（Zスコア）の加重和である中間変数（$\hat{z}_i$）を説明変数としたデフォルト確率モデルと格付該当確率モデルを同時推計（NLPプロシジャ）．

3. 幾つかの数学記号の事前準備（論文説明のための参考資料）

x_{ji}	$j=1$	$j=2$	$j=3$
$i=1$	123	456	.
$i=2$	789	.	135
$i=3$	254	.	.
$i=4$	.	369	444

原数値（「.」はSASの欠損値）

$$\mathcal{N}_j = \{i \mid (1 \le i \le N) \cap (u_{ji} = 1)\}$$

変数 j についてx_{ji}が欠損値でないサンプル番号の集合

$\mathcal{N}_1 = \{1,2,3\}$, $\mathcal{N}_2 = \{1,4\}$, $\mathcal{N}_3 = \{2,4\}$

u_{ji}	$j=1$	$j=2$	$j=3$
$i=1$	1	1	0
$i=2$	1	0	1
$i=3$	1	0	0
$i=4$	0	1	1

データ有無フラグ

$$\mathcal{M}_i = \{j \mid (1 \le j \le M) \cap (u_{ji} = 1)\}$$

サンプル i についてx_{ji}が欠損値でない変数番号の集合

$\mathcal{M}_1 = \{1,2\}$, $\mathcal{M}_2 = \{1,3\}$, $\mathcal{M}_3 = \{1\}$, $\mathcal{M}_4 = \{2,3\}$

4. 外れ値処理とYeo-Johnson変換を用いた変数変換

外れ値を適当な上下限値に丸め
→グリッドサーチ

$$y_{jki} = \max\left(\min\left(x_{ji}, x_{j,k_2}^{\max}\right), x_{j,k_1}^{\min}\right)$$

⓪下限値候補1, 2, …, 20%タイル値
上限値候補99, 98, …, 80%タイル値を計算.

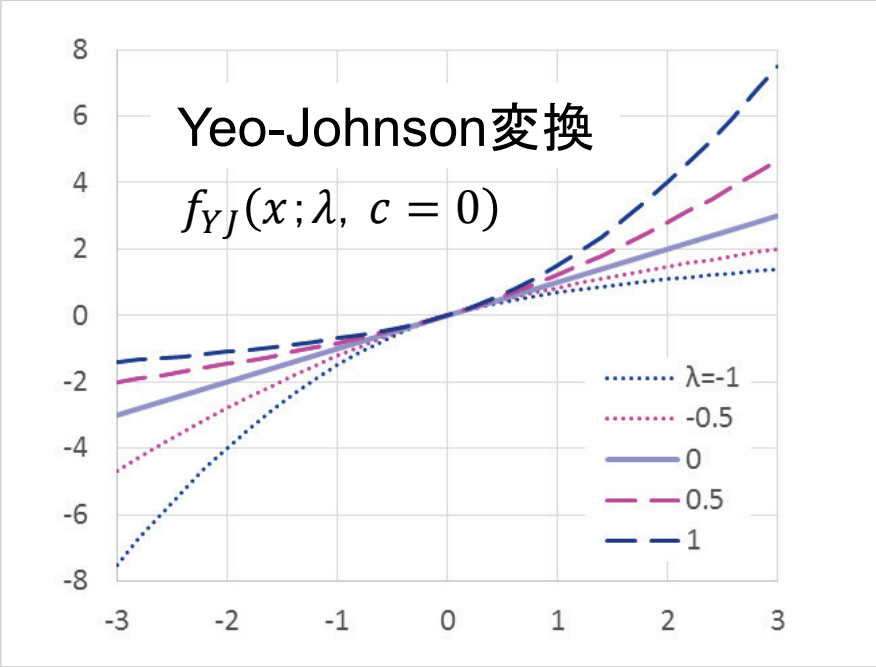
①4%間隔の格子点を選択し, 変数変換を考慮したロジット推計を行い, 尤度最大時の%タイル値を候補格子点として記帳.

②上記の候補格子点の周囲のみ, 2%間隔の格子点を選択し, 上記同様の推計を実施.

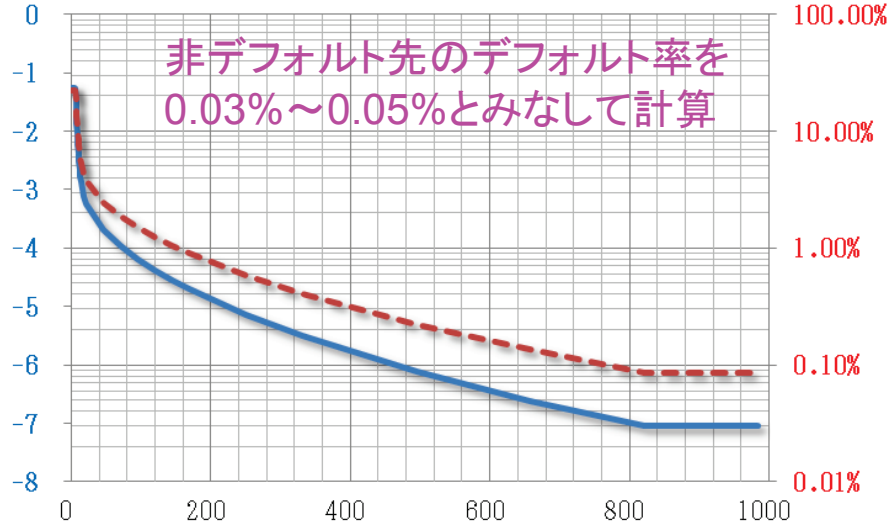
③上記の候補格子点の周囲のみ, 1%間隔の格子点を選択し, 上記同様の推計を実施.
得られた格子点で上下限値を確定.

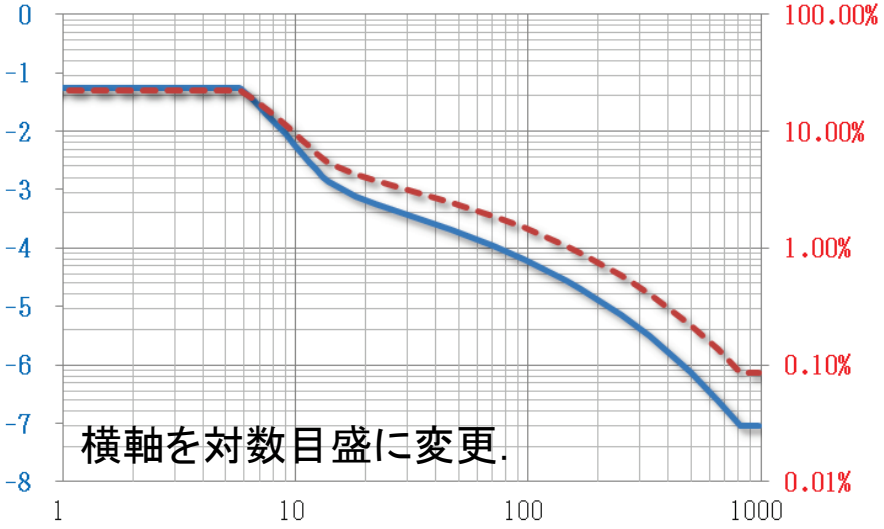
原数値を適当な関数形に変換
→Yeo-Johnson変換+基準点移動

$$z_{jki} = \alpha_{jk} + \beta_{jk} \cdot f_{YJ}\left(y_{jki}; \lambda_{jk}, c_{jk}\right)$$



4. 外れ値処理とYeo-Johnson変換を用いた変数変換（例）





1人当たりGDP/世界平均[%]（横軸）とZスコア（左軸），デフォルト確率換算（右軸）の関係

$$z_{ji} = \alpha_j + \beta_j \cdot \left[\frac{\{1 + \max(y_{ji} - c_j, 0)\}^{(1+\lambda_j)} - 1}{\underbrace{1 + \lambda_j}_{0.48}} - \frac{\{1 - \min(y_{ji} - c_j, 0)\}^{(1-\lambda_j)} - 1}{\underbrace{1 - \lambda_j}_{1.52}} \right]$$

$\alpha_j = -2.94$ $\beta_j = -0.08$ $c_j = 14.58$ $\lambda_j = -0.52$ $x_j^{\min} = 5.73$ $x_j^{\max} = 821.00$

5. 重回帰モデルによる欠損値補完処理

重回帰モデルによる単一代入法は，「標準誤差が小さくなるとともに，相関が誇張される（丹後ほか，2013）」との指摘があり，後段はそのとおり

しかし，本稿での欠損値補完処理は，目的ではなく手段であることに注意
目的は，デフォルト確率や推計格付を得ること，それに至るための手段

欠損値であるZスコアが，他の変数のZスコアと相関高く補完されてしまうことは，逆に捉えれば，想定外のZスコアが埋まるリスクを減少

ところで，多重代入法（MIプロシジャ）で，外れ値処理の反映は可能か？
補完値が外れ値範囲なら丸めるという内部ロジック組込みが必要では？

5. 重回帰モデルによる欠損値補完処理

例	データ有無フラグ				モデル推計用サンプル				欠損値補完適用モデル			
	z_{1i}	z_{2i}	z_{3i}	z_{4i}	z_{1i}	z_{2i}	z_{3i}	z_{4i}	z_{1i}	z_{2i}	z_{3i}	z_{4i}
$t=1$	0	1	0	1		A,C		A,C	A		C	
$t=2$	0	1	0	1		A,B,C,D		A,B	A		C	
$t=3$	0	1	1	0		A,B	A,C		A			D
$t=4$	0	0	1	0			A,B,C,D		A	B		D
$t=5$	0	0	1	1			A,B	A	A	B		

モデルA　推計：欠損値以外の全データ　　適用：時系列すべて欠損値の場合

モデルB　推計：1期前があるデータ　　適用：当該時点以前に1時点以上データが存在

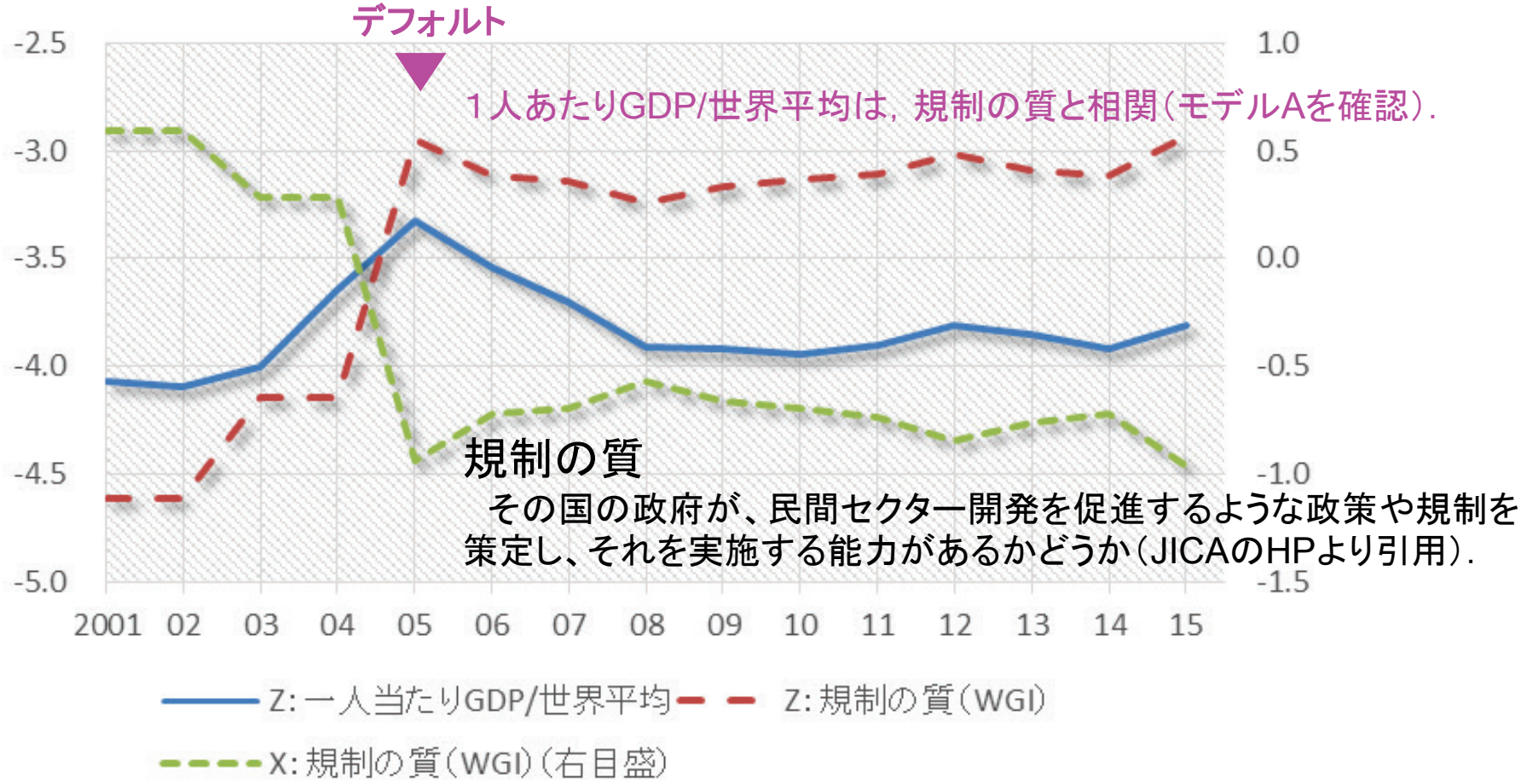
モデルC　推計：1期後があるデータ　　適用：当該時点以降に1時点以上データが存在

モデルD　推計：1期前後があるデータ　　適用：当該時点以前と以降に1時点以上データが存在

$$\hat{z}_{ji}^{M} = \max\left(\min\left(\alpha_j^M + \sum_{j' \neq j} \beta_{j,j'}^M \cdot z_{j'i} + \underbrace{\gamma_{-j}^M \cdot z_{ji^-(l,t-1)}}_{M=B \text{ or } D} + \underbrace{\gamma_{+j}^M \cdot z_{ji^+(l,t+1)}}_{M=C \text{ or } D}, z_j^{\max}\right), z_j^{\min}\right)$$

（制約条件：$\beta_{j,j'}^M, \gamma_{-j}^M, \gamma_{+j}^M \geq 0$）　　FMMプロシジャのRESTRICTを利用

5. 重回帰モデルによる欠損値補完処理（例）

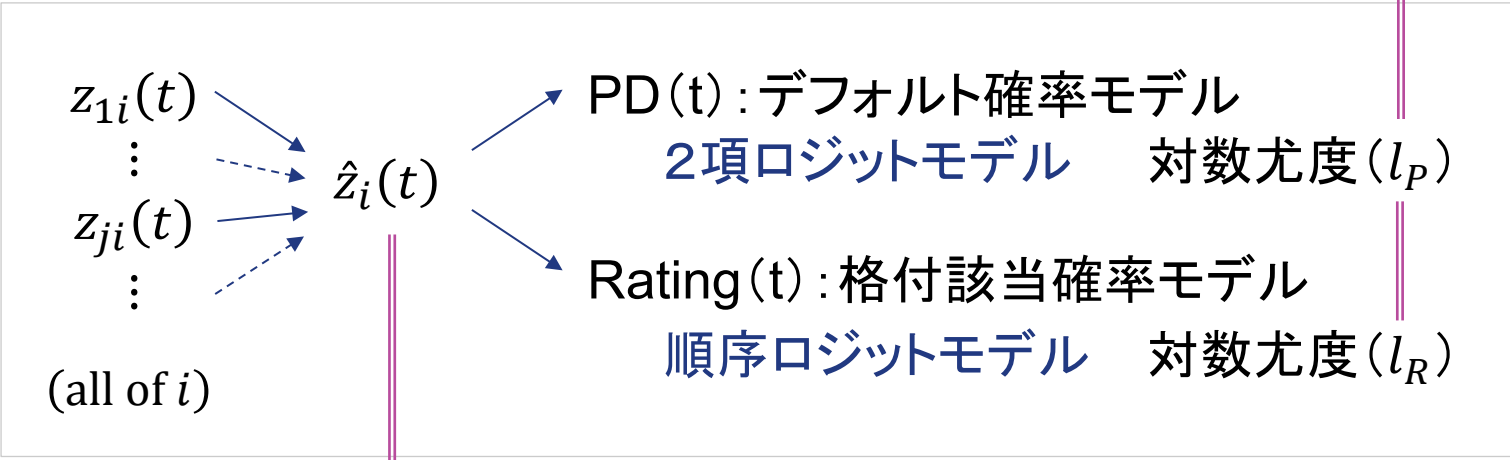


欠損値補完されたアルゼンチンの1人当たりGDP/世界平均のZスコアと規制の質（WGI）の時系列推移（2005年にデフォルト）

6. デフォルト確率モデルと格付該当確率モデルの同時推計

対数尤度の加重和を最大化するパラメータを決定

$(l_T = a \cdot l_P + (1-a) \cdot l_R \rightarrow maximize,\ \ a = 0.8)$



中間変数は説明変数Zスコアの加重和(制約条件：$\sum w_j = 1,\ w_j \geq 0$)

$$\hat{z}_i = \sum_j w_j (z_{ji} - \mu) = \sum_j w_j \cdot \mathrm{WoE}_{ji}$$

NLPプロシジャで中間変数を確定し，LOGISITICプロシジャで検算

7. 最終結果と精度確認

A. 中間変数を構成する説明変数の重みパラメータと説明力

B. デフォルト確率モデルにおける各種検証

C. 推計－実績格付マトリクスと一致率，順位相関係数

D. 2015年推計格付の世界地図による確認

A. 中間変数を構成する説明変数の重みパラメータと説明力

説明変数名称	推計値	*p*値	説明力	
			共分散	自分散
GDP成長率	0.072	0.136	0.3%	0.2%
債務利息支払/歳入	0.098	0.000	4.9%	2.3%
総国民貯蓄率	0.094	0.003	7.6%	1.8%
消費者物価上昇率の3年標準偏差	0.078	0.034	4.2%	0.5%
為替レート変化率の3年標準偏差	0.154	0.001	2.5%	0.7%
1人当たりGDP/世界平均	0.203	0.000	31.9%	12.9%
GDP/世界合計	0.054	0.000	9.2%	2.0%
対外債務/GDP÷√(GDP/世界合計)	0.021	0.075	2.9%	0.4%
政府の有効性(WGI)	0.091	0.005	19.3%	5.1%
法の支配(WGI)	0.135	0.007	17.3%	4.4%

政府の有効性

行政サービスの質、政治的圧力からの自立度合い、政府による政策策定・実施への信頼度、政府による(改革への)コミットメント。

法の支配

公共政策に携わる者が社会の法にどれだけ信頼を置いて順守しているか。特に契約の履行、警察、裁判所の質や、犯罪・暴力の可能性など。

(JICAのHPより引用)

B. デフォルト確率モデルにおける各種検証

		AR値	KS値	Div.
	統計量	0.726	0.572	2.367
	標準誤差	0.043	—	—
ブートストラップによる統計量 N=1,000	平均値	0.723	0.591	2.378
	標準偏差	0.044	0.055	0.369
	1.0%	0.620	0.468	1.639
	2.5%	0.636	0.492	1.718
	5.0%	0.650	0.506	1.822
	10.0%	0.667	0.522	1.930
	25.0%	0.695	0.550	2.120
	50.0%	0.725	0.590	2.357
	75.0%	0.753	0.627	2.613
	90.0%	0.780	0.662	2.870
	95.0%	0.794	0.682	3.012
	97.5%	0.805	0.704	3.205
	99.0%	0.822	0.723	3.378

非デフォルト先のデフォルト率を0.03%～0.05%とみなして計算

順位	サンプル数	デフォルト数	
		実績	推計
1	224	0.08	0.09
2	223	0.10	0.24
3	223	0.10	0.66
4	223	0.11	1.24
5	223	3.11	1.98
6	223	4.11	2.89
7	223	4.11	4.21
8	223	5.11	5.92
9	223	7.11	9.63
10	218	33.09	30.16

χ^2値=3.910　自由度=8　p値=0.865

SURVEYSELECTプロシジャにてブートストラップ（SAMPRATE=1）

C. 推計－実績格付マトリクスと一致率，順位相関係数

推＼実	1 AAA	2 AA	3 A	4 BBB	5 BB	6 B	合計	全体
1 AAA	173	34	2	←Korea	2013,14		209	266
2 AA	23	22	17	5	←Spain Ireland Iceland	2013,14 2011,14 2009	67	155
3 A		28	52	21		Brazil 2004 ↓	101	196
4 BBB		2	27	93	27	1	150	470
5 BB		↑ Greece 2001,02		16	14	7	37	472
6 B					8	7	15	667
合計	196	86	98	135	49	15	579	2,226

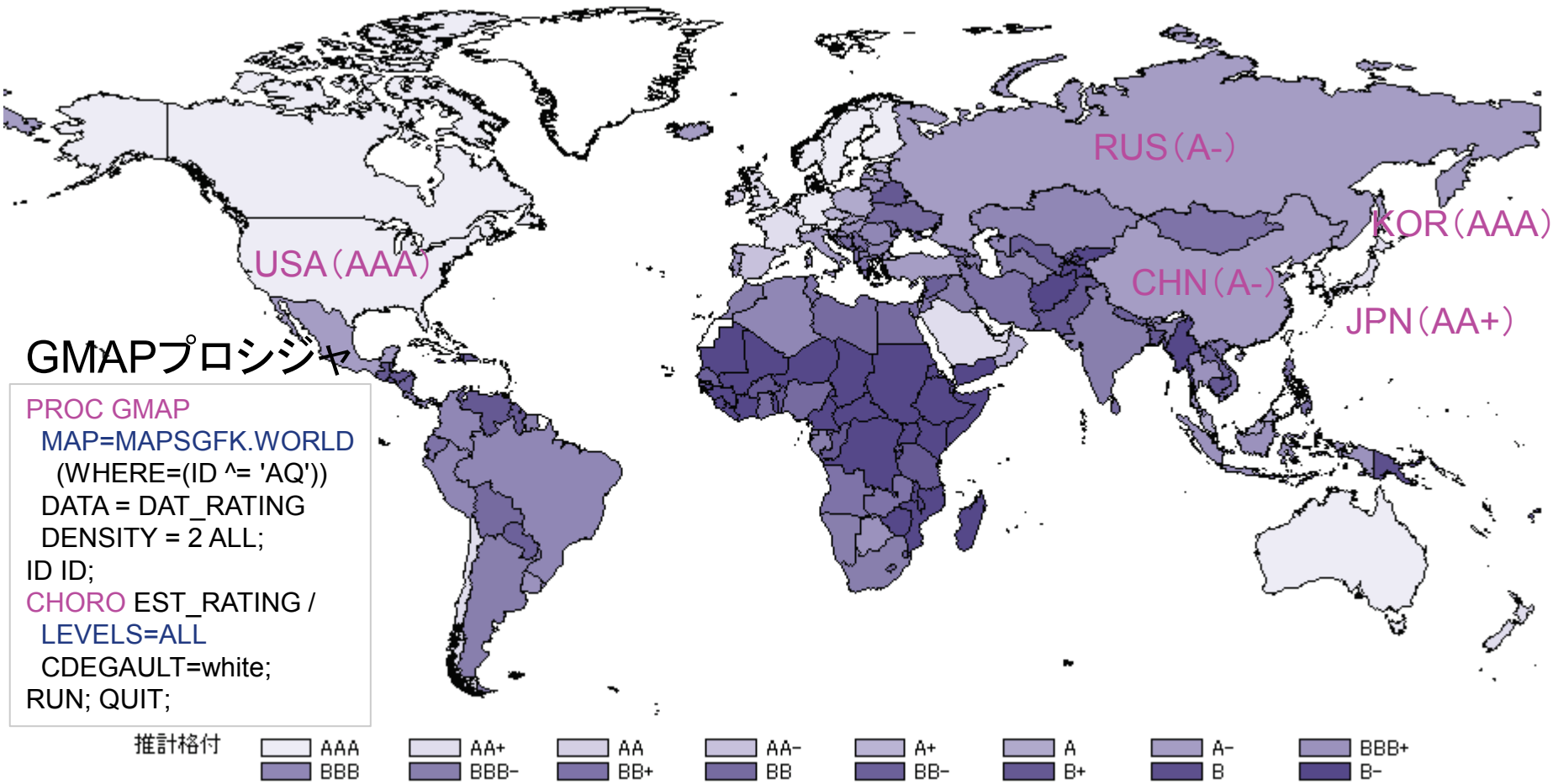
格付の一致率
=62.3%

κ 係数
=0.508

順位相関係数
=0.924

全体（最右列）は，R&Iが格付を付与していない国を含むすべて．

D. 2015年推計格付の世界地図による確認①全世界の概観

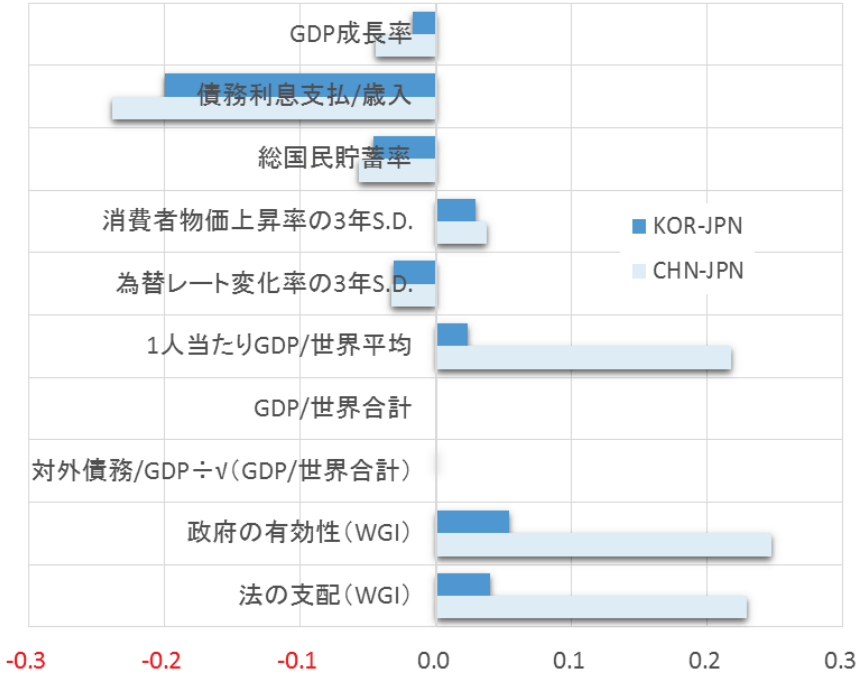


GMAPプロシジャ

```
PROC GMAP
  MAP=MAPSGFK.WORLD
    (WHERE=(ID ^= 'AQ'))
  DATA = DAT_RATING
  DENSITY = 2 ALL;
ID ID;
CHORO EST_RATING /
  LEVELS=ALL
  CDEGAULT=white;
RUN; QUIT;
```

D. 2015年推計格付の世界地図による確認②日本を基準とした韓中の比較

（良）　←（対日本基準）→　（悪）



韓国と中国における各説明変数の寄与度
（対日本基準，2015年）

韓国（KOR）：推計AAA　実績A+

債務利息支払/歳入が少ないと推定されたことが上方推定の要因．欠損補完値であるものの，水準感からすると間違いではなさそう．

中国（CHN）：推計A-　実績A+

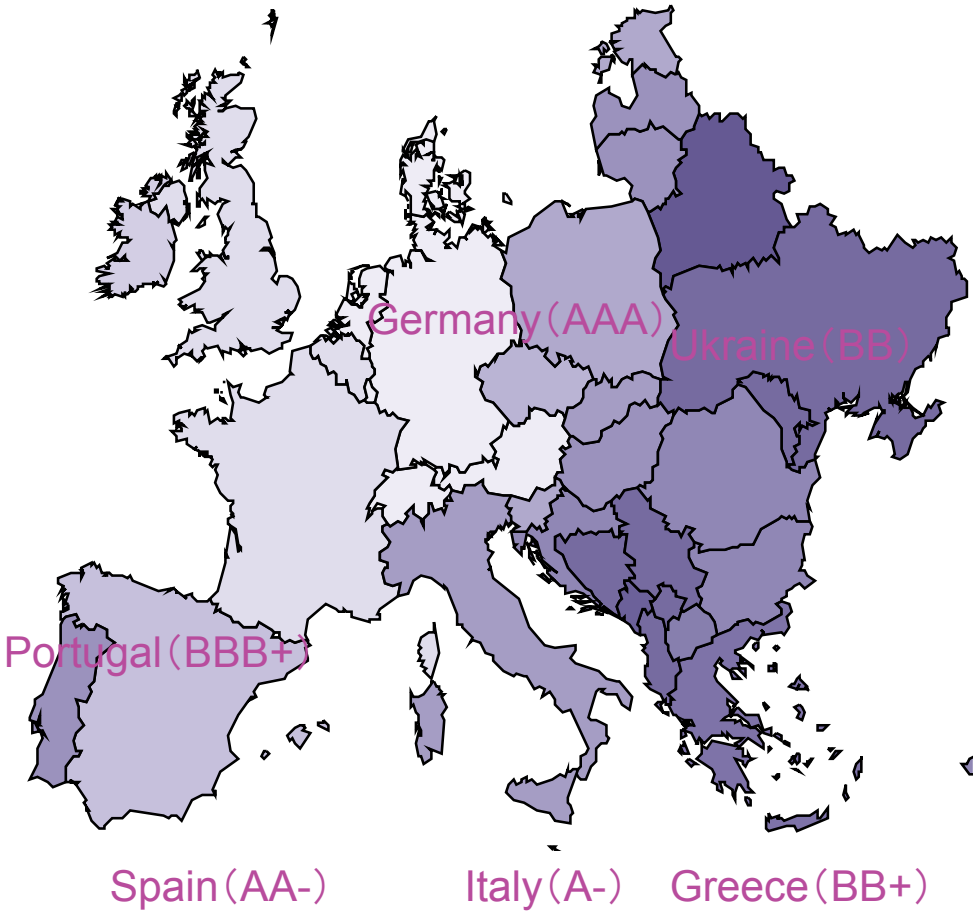
WGIの2指標を重くみていることが下方推定の要因．さらに一人当たりGDP/世界平均も足かせ．GDP成長率の寄与は小さい．

寄与度（s_{ji}）の定義

$$s_{ji} = w_j \cdot \left(\mathrm{WoE}_{ji} - \overline{\mathrm{WoE}_j}\right)$$

$\overline{\mathrm{WoE}_j}$:各変数の基準となるWoE（日本）

D. 2015年推計格付の世界地図による確認③EUおよび注目国の格付



Rank	Country Name	PD	Est.	R&I
1	Norway	0.02%	AAA	AAA
14	United States	0.06%	AAA	AAA
16	Korea, Rep.	0.07%	AAA	A+
18	France	0.08%	AA+	AAA
23	United Kingdom	0.09%	AA+	AAA
25	Japan	0.10%	AA+	AA+
28	Spain	0.15%	AA-	BBB
37	China	0.26%	A-	A+
39	Russian Feder.	0.27%	A-	—
40	Italy	0.29%	A-	A
49	Portugal	0.44%	BBB+	BB+
76	Argentina	0.84%	BBB-	—
88	Tunisia	1.09%	BB+	BBB-
97	Greece	1.28%	BB+	B-
115	Ukraine	1.64%	BB	CCC
188	Somalia	32.1%	B-	—

Greeceはデフォルト後の対象外期間中につき，PDと推計格付は参考値

8. おわりに

本稿では，WoE（あるいは対数オッズ，Zスコア）をソブリン（国）のデフォルト確率および格付該当確率モデルの構築に利用する方法を提案

提案は，一般のロジットモデル構築時も適用可能な基本的手法の集合体

モデル構築時に陥りがちな多重共線性などの問題は自動的に回避
さらに，モデルは極めて高い頑健性を保有（本稿では説明を割愛）

NLPプロシジャをOPTMODELプロシジャに置換するのは今後の課題

Contact : kkimura [at] ftri.co.jp

参考文献

[1] 大勢待利明 (2008). デフォルト確率推定モデル作成におけるWOE変換の役割とその利用方法. 2008SASユーザー総会論文集, 298-305.

[2] 丹後俊郎, 山岡和枝, 高木晴良 (2013). 新版ロジスティック回帰分析－SASを利用した統計解析の実際－. 朝倉書店.

[3] 山下智志, 三浦翔 (2011). 信用リスクモデルの予測精度－AR値と評価指標－. 朝倉書店.

[4] Cruces, Juan J. and Trebesch, Christoph (2013). Sovereign Defaults: The Price of Haircuts. American Economic Journal: Macroeconomics, 5(3), 85-117.

[5] Yeo, In-Kwon and Johnson, Richard (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87, 954-959.