Rではじめる信用リスク分析
—順序ロジットモデルを用いた格付けモデル構築—

山本　零

本稿では，代表的な信用リスク問題の一つである格付け推定問題を例として，順序ロジットモデルを用いた格付けモデルの構築方法について解説する．特に本稿では読者が実際に分析や研究を行う際の出発点になることを目的とし，実際の企業データを用いてその取り扱い方と R を用いた実装方法，結果の解釈やモデルの検証方法というモデル構築の一連の流れについて説明を行っていく．

キーワード：信用リスク，格付け推定，順序ロジットモデル，R

1. はじめに

本稿では本特集の中川先生の記事 [1] に引き続き信用リスクに関する解説として，信用リスク問題の一つである格付け推定に着目し，実際の企業データを用いたモデル構築の進め方について解説する．

格付けとは，企業の信用力に基づき付与された等級のことを指し，一般的に A+, BBB などのアルファベットと記号を組合せた簡単な表記で付与される[1]．格付けはさまざまなビジネスで利用されており，金融機関であれば与信判断や証券の購入判断，社債などの金融商品の価格付け，一般企業ではビジネスにおける相手先の信用判断が挙げられる．

特に近年，さまざまな金融危機が発生しリスク管理の高度化が求められる中で，金融機関は内部格付けという独自の格付けをもつことが求められており，その基本となるのは定量手法を用いた格付けモデルであるとされている [2].

多数ある企業群をその信用力に基づき複数の格付け（グループ）に判別する分析にはさまざまな OR 手法が適用できる．たとえば，倒産確率推計に利用されるロジットモデルを多群判別に拡張した順序ロジットモデル，データマイニングで利用される決定木分析，近年注目されている機械学習の手法の一つであるサポート・ベクター・マシンなどである．

本稿では，その中でもさまざまな研究や実務で利用されており，多くの統計ソフトウェアに組込まれているため実装も行いやすい順序ロジットモデルを利用した格付け推計モデルの構築について解説していく．

次節では，はじめに順序ロジットモデルの概要を説明する．そして分析に利用したデータや R を使った実装方法，結果の解釈について解説を行っていく．

2. 順序ロジットモデル

本節では次節のモデル構築で利用する順序ロジットモデルの概要について解説する[2].

分析対象の企業が n 社あり，それぞれの企業が信用力に基づき K 種類の格付けに分類されているものとする．このような序列をもった複数のグループを判別するためによく利用されるのが順序ロジットモデルである．順序ロジットモデルでは企業 i がもつ信用力を説明するためのリスクファクター（財務指標等）を $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{im})$ としたとき，その線形結合を信用力スコアとして用意する．

$$z_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}. \qquad (1)$$

さらに各企業を K 個の格付けに分類する閾値を以下のように定める．

$$-\infty = \tau_0 < \tau_1 < \cdots < \tau_{K-1} < \tau_K = \infty.$$

このとき誤差込みの信用力スコア $Z_i = z_i + \varepsilon_i$ と閾値との関係から格付けを表す確率変数 S_i を次のように対応させる．

$$S_i = s \quad \Longleftrightarrow \quad \tau_{s-1} < Z_i < \tau_s.$$

このように表現することで，リスクファクターが与え

やまもと　れい
武蔵大学経済学部
〒 176–8534 東京都練馬区豊玉上 1–26–1
rei.yamamoto@cc.musashi.ac.jp

[1] 国内では日本格付研究所 (JCR) や格付投資情報センター (R&I) が代表的な格付け企業である．
[2] 順序ロジットモデルだけでなく信用リスクモデル全般については木島と小守林 [3] が詳しい．

られたときの企業 i が格付け s に属する確率が次式で求められる．

$$P\{S = s \,|\, \boldsymbol{x_i}\} = P\{\tau_{s-1} - z_i < \varepsilon_i < \tau_s - z_i\}.$$

このとき誤差項 ε_i がロジスティック分布に従うことを仮定すれば，以下の式で企業 i がグループ s に属する確率 p_{is} を記述できる．

$$p_{is} = \frac{1}{1 + e^{-(\tau_s - z_i)}} - \frac{1}{1 + e^{-(\tau_{s-1} - z_i)}}. \quad (2)$$

順序ロジットモデルのパラメータ推計には最尤法を用いることが一般的である．最尤法は格付けが独立に付与されると仮定し，与えられた学習データで発生する事象を最も確からしく説明するパラメータを推計する方法である．ここで推計するパラメータを係数ベクトル $\boldsymbol{\beta}$，閾値ベクトル $\boldsymbol{\tau}$ と表記した場合，尤度関数は以下のように表される．

$$L(\boldsymbol{\beta}, \boldsymbol{\tau}) = \prod_{i=1}^{n} \prod_{s=1}^{K} p_{is}^{\delta_{is}}.$$

ここで δ_{is} は企業 i が格付け s に属すると 1 となり，そうでなければ 0 とする定数である．

最尤法では，この尤度関数 $L(\boldsymbol{\beta}, \boldsymbol{\tau})$ の対数をとった対数尤度関数を最大化することで順序ロジットモデルの係数 $\boldsymbol{\beta}$ と閾値 $\boldsymbol{\tau}$ を推計する．このようにして推計された係数は一致推定量であることが知られている．

これらのパラメータ推計に関しては，後で説明するように R などの統計ソフトウェアを利用することで簡単に行うことができる．

3. 実証分析例

本節では実際の企業データを用いて，前節で説明した順序ロジットモデルを用いた格付け推定モデルの構築手順を説明する．

3.1 データ

モデルを構築する学習データの被説明変数として 2011 年 3 月末の格付投資情報センター（以下 R&I 社）の格付けを利用し，テストデータとしては 2012 年 3 月末の格付けを利用した[3]．R&I 社の格付けは最も信用力の高い AAA から最も低い CCC+ まで 14 グループに分類されている．表 1 は分析に利用したデータの格付けごとのサンプル数を示したものである．

表 1 使用した格付けデータのサンプル数

格付け	2011	2012	クラス
AAA	3	0	6
AA+	16	10	
AA	23	20	
AA−	42	36	
A+	36	38	5
A	87	83	4
A−	78	82	3
BBB+	49	54	2
BBB	53	49	1
BBB−	13	12	
BB+	3	5	
BB	2	3	
BB−	0	0	
CCC+	0	1	
総数	405	393	

表 1 より，分類されている銘柄数は一様ではなく，格付けによっては極端に少ないことがわかる．定量分析を行う場合，極端にサンプル数が少ない場合には学習データにモデルを当てはめ過ぎてしまい，テストデータの説明力が極端に落ちるオーバーフィッティングという現象が起こりやすくなる．そのため，本稿では銘柄数のバランスと実務的な使いやすさを考慮して 6 分割した格付けを利用した．

次に格付けを説明する説明変数として，格付け取得日に最も近い時点で公表された本決算の財務データを利用した[4]．利用した財務指標は表 2 に示す 10 個であり，代表的な財務指標をその指標の意味するカテゴリ別にバランスよく選択した．「方向」はその指標単体で見た場合に，企業の信用力に対し有効であると想定される符号方向を表している．たとえば自己資本比率であれば，信用力が高い企業ほどプラスに高い値をもつと予想されることを意味する．

財務指標は指標によってオーダーが異なるため，係数の解釈が行いにくい場合がある．また極端な異常値がある場合，モデルがその異常値に合わせて係数の推計を行ってしまうことから，各指標を全銘柄で平均 0，標準偏差 1 になるよう正規化し ±4 で端数を丸める処理を行った．処理を行ったデータファイルのサンプル

[3] 最新の格付け情報は R&I 社のホームページ (https://www.r-i.co.jp/jpn/cfp/data/) から入手できる．

[4] 財務諸表は決算日から 2 カ月程度遅れて公表される．たとえば 3 月決算の企業の場合，一般的に 5 月上旬から中旬に財務諸表が公表される．先に起こる事象を反映した財務指標を使わないように，定量分析をする際には公表時点に留意する必要がある．

表 2　使用した財務指標一覧

番号	財務指標名称	カテゴリ	方向
F1	売上高営業利益率	収益性	+
F2	売上高税引後当期利益率	収益性	+
F3	自己資本比率	安全性	+
F4	負債回転期間	安全性	−
F5	インタレストカバレッジレシオ	償還能力	+
F6	棚卸資産回転期間	効率性等	−
F7	1 人当たり売上高	効率性等	+
F8	対数資本合計水準	規模	+
F9	総資本営業活動 CF 比率	現金	+
F10	売上高伸び率	成長性	+

表 3　データファイルサンプル (data2011.csv)

銘柄	格付け	F1	F2	···	F10
1	3	−0.20	0.12		0.36
2	6	4.00	2.01		−1.02
3	1	−0.43	−0.27		−0.21
4	3	−0.31	−0.01		−0.67
5	4	−1.48	−1.06		−0.88
6	4	−0.51	−0.39		−0.51
7	1	−0.09	−0.04		−0.60
8	2	−0.80	−0.15		−0.53
9	1	−0.94	−2.17		0.31
10	2	−0.60	−0.17		−0.55

を表 3 に示す.

3.2　R を用いた実装

本節では，用意したデータ（表 3）を利用した順序ロジット分析の実装について解説する．本稿ではフリーで利用することができ，さまざまな統計分析を可能にするパッケージが豊富に揃っている R を利用して実装・分析を行った[5].

具体的には，2 節で解説した順序ロジットモデルを用いて表 2 の財務指標から有効な財務指標をステップワイズ法で選択する．ステップワイズ法は，決められた手順に基づき変数の入れ替えを行いながら，AIC などのモデル説明力ができるだけ高い変数の組合せを探索する方法である[6].

つまり，分析結果としては探索した中で最も説明力の高い財務指標の組合せと，選択された財務指標を用いた順序ロジットモデルの推計結果が得られることになる.

次に分析を行ったプログラムについて解説を行う．はじめに分析の準備として，順序ロジットモデルやステップワイズ法を利用するために MASS パッケージの読み込みを行い，作成したデータ（表 3）を R に読み込ませる[7].

```
> library(MASS)
> data<-read.csv("data2011.csv",
                              header=T)
```

ここで read.csv 関数は csv ファイルを読み込ませる関数であり，ヘッダーの有無をオプションで設定する[8]. read.csv 関数によって表 3 の形式で data というデータフレームに入力される.

次に順序ロジットモデルの前処理として被説明変数の格付けを序列を付けた変数として R に認識させ，分析用のデータを作成する.

```
> y<-factor(data$格付け,
    levels=c("1","2","3","4","5","6"))
> t.data<-data.frame(y,
                  data[, 3:ncol(data)])
```

ここで factor 関数によって data の格付け列の序列を付ける．そして t.data というデータフレームに序列付けした格付けと説明変数（data の 3 列目以降）を追加したデータを作成する.

最後に作成したデータを用いて順序ロジットモデルの式を定義し，ステップワイズ法で変数選択と係数の推計を行う.

```
> eq<-polr(y~., data=t.data,
                    method="logistic")
> res<-step(eq, direction="both")
> print(summary(res))
```

ここで polr 関数が順序ロジットモデルの式を与える関数であり，“y~.” は格付けと全説明変数の線形結合を対応付けることを示す表記方法である[9]．step 関数は

[5] R は https://www.r-project.org/からダウンロード可能である．分析には本稿執筆時の最新版である R3.2.2 を利用した．また R の基本的な使い方は金 [4] を参照されたい.
[6] 近年，より説明力の高い変数選択が可能となるモデルとして，Sato et al. [5] がロジットモデルを用いた変数選択問題を整数計画問題として定式化することを提案している.
[7] MASS パッケージは GUI を通して CRAN などからダウンロードすることができる.
[8] R のカレントディレクトリが data2011.csv のフォルダと同じであることを前提として記載している.
[9] なお被説明変数が 2 グループの場合には，順序ロジットモデルではなくロジットモデルとなるため glm 関数で定義する.

ステップワイズ法を実行する関数であり，上で定義した順序ロジットモデルの式を用いて変数増減法で変数選択を行うことを意味している[10].

3.3 結果の解釈

本節ではRの分析から出力される結果の解釈について説明する．前節のプログラムを実行した結果は以下のようになる．

```
Coefficients:
    Value    Std. Error     t value
F2  0.2054     0.1345         1.527
F3  1.4977     0.1632         9.178
F4 -0.8967     0.1658        -5.410
F6  0.4952     0.1089         4.547
F8  2.9194     0.1856        15.727
F9  0.2761     0.1333         2.071

Intercepts:
    Value    Std. Error     t value
1|2 -3.4406    0.2259       -15.2282
2|3 -2.0846    0.1797       -11.5984
3|4 -0.2661    0.1473        -1.8059
4|5  1.8965    0.1812        10.4665
5|6  3.0374    0.2208        13.7589

Residual Deviance: 923.3191
AIC: 945.3191
```

この結果は最終的に選択された財務指標とその係数，五つの閾値を表しており，Value が係数，閾値の推計値，Std. Error が標準誤差，t value が t 検定の t 値を示している．また Residual Deviance は残差，AIC は赤池情報規準の統計量を表し，モデル全体の当てはまりを見るものである．

本結果の場合，指標 F2 の売上高税引後当期利益率の t 値の絶対値が 2 を下回っていることから統計的な有意性が若干低いが，それ以外の係数は統計的には問題ないと思われる．また F6 の棚卸資産回転期間に関しては，低いほど在庫が少なく効率的な経営を行っていることを示す指標であるが係数がプラスとなっており，定性的な解釈とは逆の結果が表れている．ほかの指標との相関が影響を与えていると考えられるが，中川 [1] でも指摘されているとおり，モデルを構築する際にはそのモデルが利用者に受け入れられる必要がある．そのためには定性的な解釈を行いやすいモデルを構築することが必要となるため，このような指標をモデルに組入れるかは検討が必要である．このほかにも表 2 にある財務指標のカテゴリのバランスが取れているか（ある特定のカテゴリに偏っていないか）も注意する項目の一つである．

また，分析の結果相関の高い類似した指標が複数選択される場合もある．変数間の高い相関は多重共線性の問題から推計値が不安定になる可能性がある．そのような場合には，選択された中でより説明力の低い変数を候補から除いて分析を行い直すなどの工夫も必要となる．

3.4 構築モデルの検証

構築したモデルの検証方法としては，モデルを構築した学習データで説明力などを検証する事前的検証と新たなテストデータを用意してモデルの有効性を検証する事後的検証がある．事前的検証では，モデル全体の検証として尤度比検定や AIC 基準の確認があり，各変数の検証として前節で説明した t 検定がある．

事後的検証は，モデルのロバスト性を判断するために行うものであり，学習データとは異なるテストデータが用いられる．その中でも最もわかりやすいものが的中率に関する分析である．これは推計したモデルにテストデータを入力値として推計した格付けと実際の格付けの的中率を評価するものである．本稿では，テストデータとして学習データの翌年の格付けと財務データを利用した．

テストデータを用いて検証を行う場合，テストデータの財務指標を用いて推計したモデルから格付けを推定する必要がある．以下に推計した係数，閾値を使用して格付けを推計するプログラムを示す．

はじめにテストデータ (data2012.csv) を読み込み，推計した係数，閾値を変数に入力する．

```
> data<-read.csv("data2012.csv",
                          header=T)
> beta<-c(0, 0.2054, 1.4977, -0.8967,
      0, 0.4952, 0, 2.9194, 0.2761, 0)
> tau<-c(-3.4406, -2.0846, -0.2661,
          1.8965, 3.0374)
```

ここで推計した係数 beta には採用されていない変数を 0 として入力することに注意が必要である．

[10]本節で利用したステップワイズ法は変数増減法と AIC 基準を用いた変数選択であるが，オプションの設定を変更することでほかの探索方法や探索基準を用いることも可能である．

次に計算用に説明変数だけのデータを作成し，各格付けの分布関数を計算して data の最終列に追加する．

```
> tmp<-data[, 3:ncol(data)]
> data<-cbind(data, "logit1"=
     apply(tmp, 1, function(x){
     1/(1+exp(-tau[1]+sum(beta*x)))}))
>...
```

ここで cbind 関数はデータフレームに列を追加する関数であり，data の最終列に logit1 という列を追加することを意味する．また apply 関数は与えたデータ tmp の行ごと，つまり銘柄ごとに function で定義したロジット関数を当てはめて計算することを示す．この処理を tau[1] から tau[5] まで変化させて繰り返し，それぞれ列名を “logit1” から “logit5” とする．

次に (2) 式より格付けの所属確率を計算し，data の最終列に追加する．

```
> data<-cbind(data,"prob1"=data$logit1)
> data<-cbind(data,
       "prob2"=data$logit2-data$logit1)
>...
> data<-cbind(data,
       "prob6"=1-data$logit5)
```

最後に所属確率の最大値を data の最終列に追加し，その最大値が示す所属確率に応じて data に予想格付けを追加する．

```
> data<-cbind(data, "max"=
   apply(data[,18:ncol(data)], 1, max))
> data<-cbind(data, "forecast"=
        ifelse(data$prob1==data$max, 1,
        ifelse(data$prob2==data$max, 2,
        ifelse(data$prob3==data$max, 3,
        ifelse(data$prob4==data$max, 4,
        ifelse(data$prob5==data$max, 5,
                                6))))))
```

ここで ifelse 関数は条件が満たされているかを判断し，処理を行う関数である．このプログラムを実行した結果，data に予想格付け forecast 列が追加される．

この予想格付けと実際の格付けの一致率を計測したものが表 4 である．

表 4　的中率分析結果

格付け	学習データ		テストデータ	
	完全一致	±1 以内	完全一致	±1 以内
全体	55.3%	86.4%	46.6%	77.6%
1	70.4%	81.7%	70.0%	75.7%
2	10.2%	87.8%	9.3%	79.6%
3	52.6%	88.5%	34.1%	76.8%
4	63.2%	86.2%	54.2%	72.3%
5	0.0%	88.9%	0.0%	78.9%
6	86.9%	86.9%	84.8%	84.8%

表 4 より，学習データの的中率が完全一致で 55.3%，±1 の誤差を許した場合 86.4%であるのに対し，テストデータでは完全一致が 46.6%，±1 の誤差を許した場合 77.6%と若干低下していることがわかる．学習データとテストデータの的中率が乖離して，テストデータの的中率が低い場合，モデルが学習データにオーバーフィットしている可能性が高い．そのような場合には変数やモデルそのものの見直しが必要になる．このほかにもさまざまな検証方法があるが，詳しくは山下ら [6] などを参照されたい．

4. おわりに

本稿では，読者が信用リスク問題を扱う際の教科書となることを目的とし，分析手法の概略からデータの取り扱い，R を用いた実装，結果の解釈について解説をしてきた．本稿を順を追って理解することで，順序ロジットモデルを用いた格付けモデルの構築が実現できると思われる．

ただし，本稿ではわかりやすさを重視したため，本来するべき説明を割愛したところも多い．たとえばほかのモデルとして，判別分析やサポート・ベクター・マシンを用いることもでき，さまざまな研究で取り上げられている．

一例として後藤と山本 [7] では，サポート・ベクター・マシンを格付け推定問題に適用して順序ロジットモデルと的中率を比較しており，テストデータにおいてより精度の高いモデルが構築できることを示している．

データの取り扱い方などは，どのような定量モデルを構築する際にも本稿で説明したものと大きくは変わらない．また R の利用方法，さまざまなデータ分析手法に関しては金 [4] に詳しく記載されている．本稿を出発点としてさまざまな研究に進んでいただければ幸いである．

参考文献

[1] 中川秀敏，“信用リスク入門，”オペレーションズ・リサーチ：経営の科学，**61**, pp. 359–364, 2015.

[2] 日本銀行金融機構局，“内部格付制度に基づく信用リスク管理の高度化，”リスク管理高度化と金融機関経営に関するペーパーシリーズ，2005.

[3] 木島正明，小守林克哉『信用リスク評価の数理モデル』，朝倉書店，1999.

[4] 金明哲『R によるデータサイエンス』，森北出版，2007.

[5] T. Sato, Y. Takano, R. Miyashiro and A. Yoshise, “Feature subset selection for logistic regression via mixed integer optimization,” *Computational Optimization and Applications*, 2016, to appear.

[6] 山下智志，敦賀智裕，川口昇，“信用リスクモデルの評価方法に関する考察と比較，”金融庁金融研究センターディスカッションペーパー，2003.

[7] 後藤順哉，山本零，“CVaR 最小化と信用リスク判別モデル，”MTEC ジャーナル，**24**, pp. 29–48, 2012.

