偏りのある不完全大規模データに基づく中小企業信用リスクマイニング

リスク解析戦略研究センター
外来研究員　　宮本　道子　（秋田県立大学）　[1]

1　はじめに

情報化社会の進展にともなってデータの大量化、高次元化という新しい傾向が顕著になっている。その中で、中小企業に関するデータでも属性数が非常に大きい超高次元の多変量時系列データが得られるようになってきた。大規模データセットに対しては、データセットを種々の観点から探索的に分析する要求が出てくるであろう。その場合、データセットに欠測が含まれている不完全データの場合では扱いがきわめて難しくなる(岩崎　2002)。これらの巨大データに対する情報抽出・知識発見や予測・制御のために、欠測データに配慮した普遍的手法の開発が必要と思われる。

このように、実際のデータ解析では、研究対象となる変数が多くなるにしたがい欠測データの発生は回避できない問題となる。信用リスクにおける従来の研究では、欠測が多い項目は削除され、せっかくの情報が活かしきれない。そこで本研究では、反復法を用いて欠測データの擬似的な完全データを生成し、数学的に扱いが容易な完全データの枠組みで最尤推定値を求めるEMアルゴリズム（Dempster, Laird, and Rubin 1977）など、 不完全データの統計的推論に用いられる数値計算法を取り入れ、より多くの情報を活かした信用リスク推定のための統計モデルへの適用方法を示す。さらに、Selection Model や　Pattern Mixture Modelなどで欠測の原因のモデリングも行う。

また変数の数が増えることで多重共線性の問題も不可避となる。信用リスクの計測にはロジットモデルが多く使われているが、本研究では、宮本・椿（2001）の改良主成分回帰分析を一般化線形モデルに拡張する。これは、わが国ではあまり利用されていないが、主成分回帰分析(Massy, 1965)は、主成分分析で得られた主成分スコアを説明変数として回帰分析を行う方法である。オリジナルの説明変数より少ない特徴を抽出して回帰分析を行うことになるので計算も簡単になり、主成分が互いに無相関であることから多重共線性の問題がない。しかし、主成分回帰は、説明変数である主成分スコアを各説明変数と直接関連付けて解釈できないという短所がある。宮本・椿（2001）の方法は、主成分回帰分析を改良し発展させたものであり、回帰に寄与する主成分スコアをバリマックス回転し、オリジナルの変数と直接関連付けて回帰の結果を解釈するものである。

2　使用するデータ

1995年から2007年まで累計で総数約860万件の決算書数（債務者合計1.6百万人）が含まれたCRD中小企業データベースを用いる。

本研究ではCRDデータベースに収められた貴重なデータをできるだけ多く生かすことに焦点を当て、EMアルゴリズム等を用いた欠測への処置法等を検討しながら、それらのデータを使って中小企業における信用リスクの計測を行っていく。また欠測の原因についてのモデリングも行う。

不完全データに配慮した上で、大規模中小企業データベースを活用した信用リスクの計測ならびに中小企業金融の諸側面に関する研究を推進することは、日本の金融市場全体に対して、1つのベンチマークとして必要かつ有用ではないかと思われる。まず、CRDデータベースから欠測値

[1] 本研究は統計数理研究所の安藤雅和特任研究員と逸見昌之助教との共同研究である。

の傾向を分析し、欠測データのメカニズムを探る。

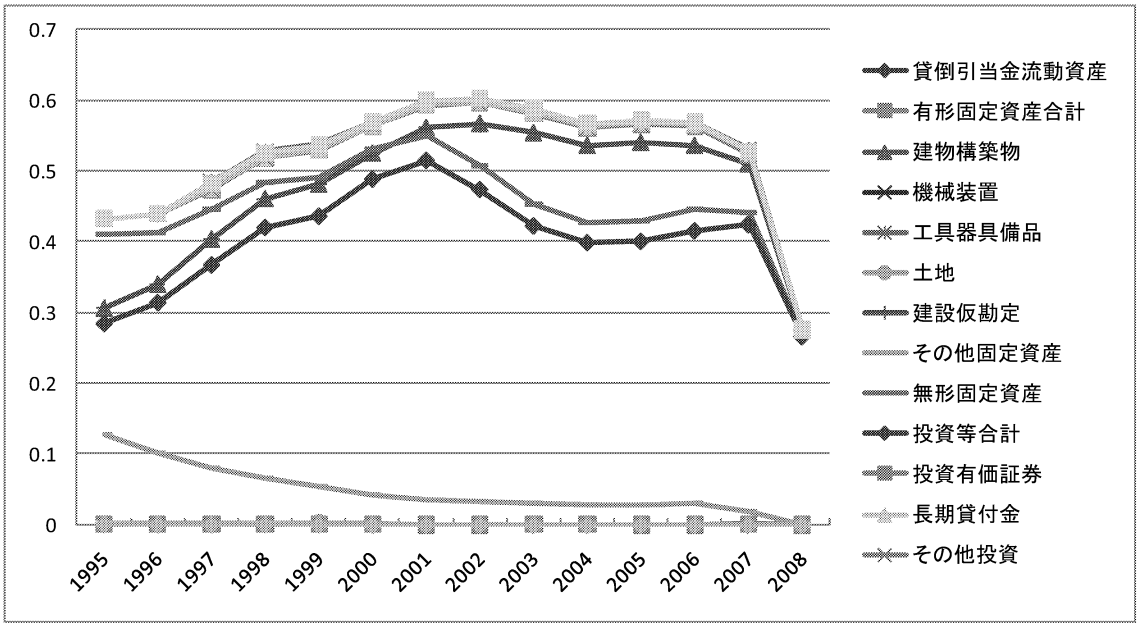
2　欠測値の観察

データのうち欠測がない変数は、表 1 に掲載されているとおりである。それ以外は欠測が観測されている。表 2 は、1995 年から 2008 年の間で、欠測が見られるデータ（一部）の欠測率の推移を表す。

表 1　すべて観測されている変数

流動資産合計	資産合計	負債合計	受取利息割引料配当金
現金預金	流動負債合計	資本合計	支払利息割引料
受取手形	支払手形	資本金	経常利益
売掛金	買掛金	その他の資本	当期利益
棚卸資産合計	短期借入金	負債資本合計	減価償却実施額
その他流動資産合計	その他流動負債合計	売上高営業収益	期末従業員数人
固定資産合計	固定負債合計	売上原価営業原価	デフォルト
有形固定資産合計	社債長期借入金	売上総利益	
土地	その他固定負債	販売費および一般管理費	
繰延資産	長短借入金合計	営業利益	

表 2　変数別欠測率の推移



これらのデータを用いた分析結果は当日発表する。

参考文献

A.P.Dempster, N.M.Laird, and D.B.Rubin,(1977) “Maximum Likelihood from Incomplete Data via the EM Algorithm” Journal of The Royal Statistical Society (B), vol.39, no.1, pp.1-38.

岩崎学　(2002) 『不完全ﾃﾞｰﾀの統計解析』 ｴｺﾉﾐｽﾄ社。

Miyamoto, M. and Tsubaki, H.,” Measuring Technology and Pricing Differences in the Digital Still Camera Industry Using Improved Hedonic Price Estimation,” *Behaviormetrika*, Vol.28, No.2. 111-152 (2001)