探索的財務ビッグデータ解析

地道 正行 *, 宮本 大輔 **, 阪 智香 *, 永田 修一 *
* 関西学院大学 商学部
** 奈良先端科学技術大学院大学 先端科学技術研究科

概要

本研究では, 世界の全上場企業の「規模の大きな」財務データに対して探索的データ解析 (Exploratory Data Analysis: EDA) (Tukey (1977)) を実行することによって新たな知見を得ることを目的としている.

まず, データとして, Bureau van Dijk (BvD) 社 [1)] から提供されるデータベース Osiris から抽出された世界 157 カ国の全上場企業 (一般事業会社, 上場廃止企業を含む 8 万社超) の 33 年分の財務データ (売上高, 営業利益, 総資産など 84 項目) を, Unix コマンド (`sed`, `grep` など) とデータ解析環境 R を利用することによって整形し, 近年注目されているクラスター・コンピューティング・システム Apache Spark^{TM} [2)] と SparkR パッケージを用いて R へ読み込んだ.

次に, このデータに対して探索的データ解析を行った. 具体的には, 2015 年に時点を固定し, データ可視化 (data visualization) を行った結果として, 売上高, 従業員数, 総資産のそれぞの (1 変量) 分布は極度に右に歪んだものであることがわかり, 各ペアの 2 変量分布も同様の結果となった. この結果に対して, 対数をとることによって歪みを修正したところ, 若干左に歪んでいることが分かった. この知見を利用して, Azzalini (1985), Azzalini and Capitanio (2014) によって提案された非対称分布 (非対称正規分, 非対称ティー分布) を売上高の対数に当てはめたところ, 非対称ティー分布の当てはまりが良いことがわかった. この結果から, 売上高を従業員数と総資産によって説明するためにコブ・ダグラス型生産関数 (Cobb and Douglas (1928)) を応用し, 両辺の対数をとった, いわゆる両対数モデル (double-log model) を利用して統計モデリング (statistical modeling) を行ったところ, 誤差項に非対称ティー分布を仮定したものが回帰診断の結果として最も良いことが分かった. さらに, 赤池情報量規準 (Akaike Information Criterion: AIC) を用いてモデル選択 (model selection) を行い, 交差確認法によって評価 (model evaluation) を行った結果も, 上記の結果が肯定されるものとなった.

本研究のデータ (ファイル) を処理する工程は Unix のシェルスクリプトと R スクリプトで一元管理し, 探索的データ解析の過程も意味のある結果を文書化する工程は, Sweave を利用して LaTeX ファイルに R コードを埋め込む形式で動的に生成した. さらに, これらの全工程を `Makefile` にスクリプトを記述し, Unix の `make` コマンドにより自動実行することによって, 再現可能研究 (reproducible research) を行うことを試みた.

[1)] `https://www.bvdinfo.com/en-gb/`

[2)] `http://spark.apache.org/docs/latest/index.html`

参考文献

[1] Azzalini, A. (1985) A class of distributions which includes the normal Ones, *Scandinavian Journal of Statistics*, Vol. 12, No. 2, pp. 171–178.

[2] Azzalini, A. with the collaboration of A. Capitanio (2014) *The Skew-Normal and Related Families*, Cambridge University Press, Institute of Mathematical Statistics Monographs.

[3] Cobb, C. W. and P. H. Douglas (1928) A Theory of Production, *American Economic Review*, Vol. 18, pp. 139–165.

[4] Gandrud, C. (2015) *Reproducible Research with R and RStudio*, Second Edition, CRC Press.

[5] Jimichi, M., Miyamoto, D., Saka, C. and Nagata, S. (2018) *Visualization and Statistical Modeling of Financial Big Data: Log-Linear Modeling with Skew Error*, SSRN: `https://ssrn.com/abstract=3166440`, submitted.

[6] 地道正行, 豊原法彦 (2018) 『景気先行指数の動的文書生成にもとづく再現可能研究』, 豊原法彦編著『関西経済の構造分析』, 第 5 章, pp. 77-111, 中央経済社.

[7] Konishi, S. and G. Kitagawa (2008) *Information Criteria and Statistical Modeling*, Springer.

[8] Leisch, F. (2002) *Sweave: Dynamic generation of statistical reports using literate data analysis*, In Wolfgang Härdle and Bernd Rönz, editors, Compstat 2002 - Proceedings in Computational Statistics, pp. 575-580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.

[9] Ryza, S., U. Laserson, S. Owen, and J. Wills (2016) *Advanced Analytics with Spark*, O'Reilly.
(玉川 竜司訳 (2016)『Spark による実践データ解析』, オライリー・ジャパン.)

[10] Saka, C. and M. Jimichi (2017) Evidence of inequality from accounting data visualisation, *Taiwan Accounting Review*, Vol. 13, No. 2, pp. 193–234.

[11] Tukey, J. W. (1977) *Exploratory Data Analysis*, Addison-Wesley Publishing Co.

謝辞

本研究の一部は以下の研究費より助成を得ていることに感謝の意を述べたい:

- 科学研究費 基盤研究 C:「グラフィカル・データ・アナリシスによる格差研究と社会環境会計による解決方法の提案」(2016 年 ~2018 年), 課題番号: 16K04022, 研究代表者: 阪 智香
- 平成 30 年度 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 課題:「財務ビッグデータの可視化と統計モデリング」, 課題番号: jh181001-NWJ, 研究代表者: 地道 正行
- 関西学院大学 図書館 図書費 B, 個人研究費

また, BvD 社の増田 歩氏にはデータの抽出に関して多大なるご協力いただいた. ここに感謝の意を述べる.