Imputation and outlier detection in banking datasets

Andrea Pagano, Domenico Perrotta and Spyros Arsenis

Abstract Data from the banking balance sheets can be used to analyse the financial stability of the banking sector. Occasionally, it may occur that some data values are either incorrect or missing, which would have an important effect on the results of the analyses. Thus, incorrect values should be detected and removed or corrected, while missing values should be imputed. This contribution addresses the two problems using a robust data analysis approach, known as Forward Search. In particular, the Forward Search is used to address the presence of high data collinearity, which may give rise to many irrelevant outliers. In recent years a MATLAB toolbox, the Forward Search for Data Analysis (FSDA), has been applied to similar problems in official statistics. The contribution extends the application to the banking sector.

Key words: Imputation, Banking datasets, Forward Search, Regression Outliers

1 Introduction

In many cases missing or incorrect data complicate dramatically the work of the analysts. These problems may occur in many disciplines, varying from environmental studies to economic analysis. It is therefore a good practice for researchers to include, when the study design is planned, techniques able to address these issues.

In our work we deal with banks' balance sheets. We analyse about 3000 banks across the European Union in order to asses the probability of a systemic financial crisis and the consequent impact on public finances. We use Bankscope database, a commercial source of information about banks' annual reports developed by Bureu van Dijk (`http://www.bvdinfo.com/`). We use the data stored in Bankscope as input for a model called SYMBOL (*SYstemic Model of Banking Originated*

Andrea Pagano, Domenico Perrotta and Spyros Arsenis
European Commission, Joint Research Centre e-mail: `andrea.pagano@jrc.ec.europa.eu`, `domenico.perrotta@ec.europa.eu`, `spyros.arsenis@jrc.ec.europa.eu`

Asset PD	computed from balance sheet variables
Total Assets	taken from the balance sheet
Capital Requirement	taken either from the balance sheet or reconstructed
Customer Deposit	elaborated using data from balance sheet
Inter-bank Exposure	elaborated using data from balance sheet

Table 1 Input variables used by the SYMBOL model.

Losses) that we develop at the Joint Research Centre of the European Commission in view of monitoring the current financial crisis. SYMBOL simulates potential crises in the banking sector under various assumptions, and it allows assessing the cumulative effects of different regulatory measures (e.g. higher capital requirements, strengthened deposit insurance and introduction of resolution funds) and their most effective combinations.

SYMBOL uses items in bank's balance sheet to estimate the potential losses for a given banking system via a Monte Carlo analysis. The model is flexible and can be deployed either on a single country or on a set of financial institutions sharing common features. The basic idea is to simulate enough random scenarios and compare the bank assets with the asset probability of defaults (AssetPD). Then the event of a bank default is estimated by comparing the bank asset probability of default with the capital (actual or envisaged). The details on SYMBOL model can be found in De Lisa *et al.*, 2011.

The SYMBOL model uses the variables in Table 1. An important variable, capital requirement, for many banks is not directly available from Bankscope. This is due to many reasons, for example the fact that for some countries the legislation does not oblige banks to report this information in their annual report. Moreover, since different aspects of bank's activities contribute to the calculation of AssetPD, each one subject to approximations or recording mistakes, a check on data coherence is necessary in order to have reliable results.

Therefore, we are facing with two issues: imputation of missing values and detection of anomalies in the data. In this paper we address both problems with a single robust regression technique, based on the Forward Search approach of Atkinson and Riani (2000), which is introduced in Section 4. The dataset used to demonstrate the approach and the connected imputation issue are described in the next two sections. Results and some final remarks conclude the paper (Sections 5 and 6).

2 Dataset description

We focus our analysis on data from Bankscope relative to year 2010, which is the last complete data set available at the present date. We are interested in data for all European Union Member States (27 countries).

We start with a data set containing 8893 banks with 28 fields. Bankscope lists banks with respect to their activities, which are: Bank Holding & Holding Com-

panies; Central Bank; Clearing Institutions & Custody; Commercial Banks; Cooperative Bank; Finance Companies (Credit Card, Factoring & Leasing); Group Finance Companies; Investment & Trust Corporations; Investment Banks; Islamic Banks; Micro-Financing Institutions; Multi-Lateral Government Banks; Other Non Banking Credit Institution; Private Banking & Asset Mgt Companies; Real Estate & Mortgage Bank; Savings Bank; Securities Firm; Specialized Governmental Credit Institution.

Our main interest is to quantify the impact of the financial crisis on the public finances of the Member States, which may be called to cover losses to protect depositors. Therefore, for this purpose we only select banks and institutions listed under the following categories: Commercial Banks, Cooperative Bank, Savings Bank. This reduces the database extraction to 6500 banks.

After having done some standard data coherence checks on the basis of the accounting rules, we select only banks for which data for both Total Assets and Equity are available. These two variables, which can only be found in the banks balance sheets, are necessary for the statistical analysis and the Montecarlo simulation in SYMBOL. After this selection, our dataset finally reduces to about 3580 banks.

3 Imputation approach

One of the key elements in running the SYMBOL model is the capital requirement. This is needed first for computing the AssetPD, then for estimating potential losses and evaluating net losses in the case of bank's default. In our database we have two different variables related to the capital requirements, which in Bankscope are either both available or both missing: (a) Total Regulatory Capital (TRC); (b) Tier1 Capital (Tier1). They refer to slightly different notions of capital requirements and in SYMBOL we normally use Tier1.

When the information on capital is missing, we can use the fact that bank's capital and equity are strongly correlated. In fact, in the majority of the cases extracted from Bankscope, we observe that Tier1 and Equity pairs lie very close to a single line (the case study of Figure 1 exemplifies the situation). Therefore an approach to impute missing Tier1 values is to fit the subset of banks for which both variables are available, with Equity as explanatory variable, and estimate the capital requirement from the fit.

Two aspects must be carefully considered with this approach. One is that the fit must be robust to the presence of outlying values in the data. The second has to do with the fact that in presence of high data collinearity even minor deviations from the regression line, which from the operational point of view may be totally irrelevant, become statistically significant and are therefore detected as outliers. These two aspects are discussed in more details in Section 4.

The approach can be deployed at different levels: on the entire dataset, within different bank categories, or within each single country. As case study for this paper we use the country level approach for Austria.

4 Robust regression through the Forward Search

The goal of robust statistics is to build estimators independent from model assumption deviations and identify outliers, i.e. observations which are distant from the bulk of the observed data and can hardly comply with model assumptions. The discipline has grown considerably in the last two decades and many robust methods are available in the literature (Maronna *et al.*, 2006, is an excellent introduction to the field). Among such methods, the Forward Search of Atkinson and Riani (2000) has shown superior properties in terms of size and power (Torti *et al.* 2012).

For a regression problem with p explanatory variables, the Forward Search (FS) builds subsets of increasing size m, starting from $m_0 = p$, until all observations are included. The subsets are built using simple ordering criteria: at step m, the traditional least squares is used for fitting the m observations in the current subset and the next subset is built with the $m+1$ units with smaller residuals of the fitted model.

During the process, as m goes from p to n, we can monitor the evolution of model estimates, the residuals of the fitted model, or other test regression statistics. In absence of outliers we expect that during the search process all these statistics remain rather constant or show smooth increases. On the contrary the entry of outliers, which by construction will happen in the last subsets, will be revealed by appreciable changes of the monitored statistics. For an important statistic, the minimum deletion residual among observations not in the subset, distributional results and confidence bands can be used to identify precisely the outliers (see e.g. Atkinson and Riani, 2006).

As anticipated in Section 3, the majority of the Tier1 and Equity pairs are almost perfectly aligned on one single line. In such case, the estimated value of the variance of the errors of the regression line, σ^2, will be very close to zero and a small difference between the capital and equity Bankscope sources may lead to very large residuals, being standardized by the estimated values of σ. Of course, for the same reason also the p-values will be very small. In robust statistics this problem is known as "perfect fit" (Maronna et al., 2006). The Forward Search offers a very natural way to keep into account the potential presence of perfect fit cases, by monitoring the value of the coefficient of determination (R^2) during the search. A value of R^2 that during the progression of the search stays constantly close to 1 is an indication of almost perfect fit. In such case, we disregard outlier signals based on the standard diagnostic regression statistics, such as the minimum deletion residual, and we increase the confidence level to declare observations as anomalous.

In addition, datasets for which the estimated R^2 value was very small for most of the search were also collected and studied separately, denoting cases where the supposed correlation between capital and equity does not hold, which for example may happen in presence of multiple groups in the data. Under this scenario, the data should be segmented differently or studied using robust clustering approaches (see for example Garca-Escudero *et al.*, 2010).

This approach has been implemented using routines contained in the FSDA toolbox for Matlab, developed jointly by the University of Parma and the Joint Research

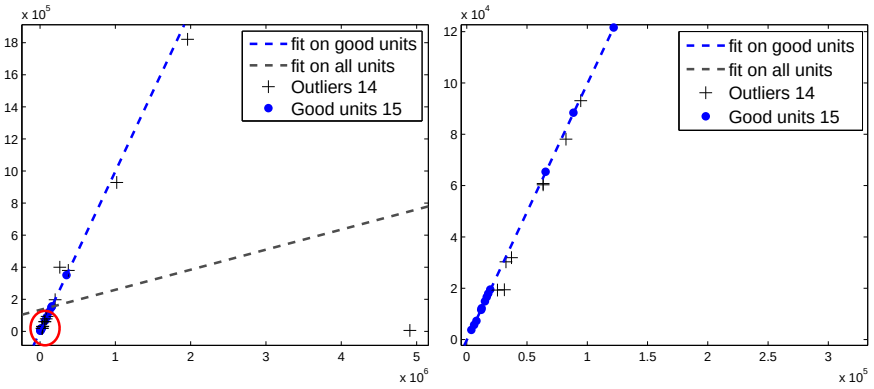

Fig. 1 Outliers detected by the FS with the default simultaneous 99% confidence level. The right plot magnifies the area highlighted with an ellipse in the left plot. The strong collinearity and the consequent "perfect fit" problem are clear.

Centre of the European Commission (Riani, Perrotta and Torti, 2012). FSDA is freely available for non commercial use from `http://www.riani.it/MATLAB` or `http://fsda.jrc.ec.europa.eu`.

5 Results

We discuss the application of the methods to the case of Austria. For this country Bankscope stores 423 banks (369 of them being unconsolidated) including all specializations. Focusing only on Commercial, Cooperative and Savings banks we reduce the dataset to 322 financial institutions. Among them 257 carry information on capital, i.e. the fields of both Total Assets and Equity are filled. Within this subset, banks for which we have data associated to Tier1 capital are 29.

First of all, we note that the correlation coefficient between Equity and Tier1 is very low: 0.3254. Also, if we test the hypothesis of no correlation, we obtain a p-value of 0.085. These results, which contradict the expectation of strong linear relation between the two variables, depend on the presence of outliers in the data.

Figure 1 shows with symbol '+' the outliers detected by the Forward Search with the default simultaneous 99% confidence level. This means that in presence of normally distributed data without contamination, we expect to find outliers in 1% of the datasets which are analyzed. Figure 2 also plots the trajectory of the minimum deletion residual and its 1%, 50% and 99% confidence bands (dotted lines). The search has started in the area where the majority of the points are collinear, i.e. the area highlighted in the left plot of Figure 1 and zoomed in the right plot of the same Figure. Then, the inclusion of observations which deviate just slightly from the alignment produce the early exceedances from the 99% band of Figure 2 (the upper dotted line) and are therefore detected as outliers.

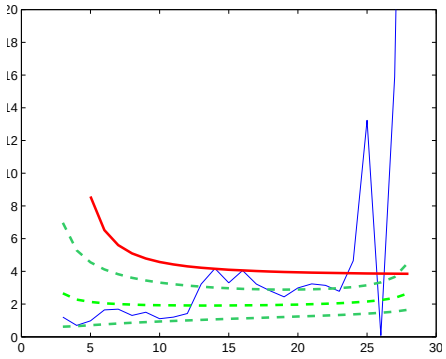

Fig. 2 Monitoring of the minimum deletion residual among observations outside the subset and the theoretical simultaneous 1%, 50% and 99% confidence level bands (dotted lines). The 99% simultaneous band is compared with the corresponding 99% Bonferroni band (plain thick line).

From these plots it is quite clear that many of the 14 outliers detected with this standard approach should not be excluded from the model fitting. We can therefore think to relax the algorithm and declare an observation as an outlier only if, by including it in the subset, the R^2 becomes (or stays) smaller than a reasonable threshold associated to the 'perfect fit' problem, say 0.95. The result of this relaxed algorithm is shown in the left plot of Figure 3, where the outliers identified are now just two[1]. To be even more selective and focus only on the very extreme outliers, one may also relax the confidence level of the outlier tests by looking at the sig-

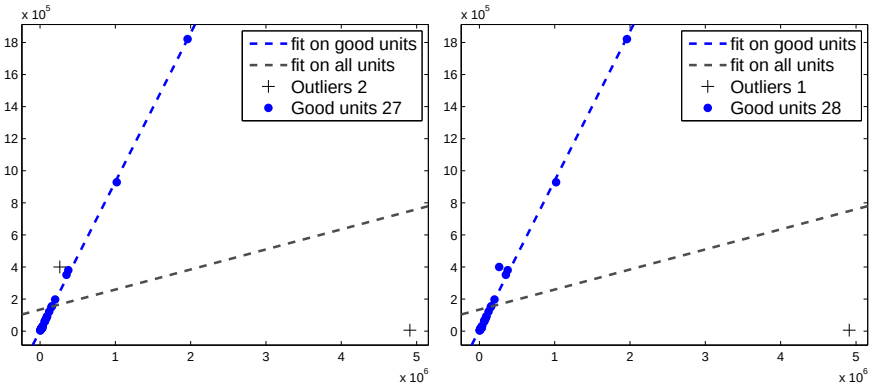

Fig. 3 Left plot: outliers detected by the FS with simultaneous 99% confidence level, but only if by including them the R^2 becomes smaller than 0.95. Right plot: as on the left, but now the confidence level is not simultaneous; it is relaxed to be 99% corrected with Bonferroni throughout the search.

[1] To be strict, one should distinguish between 'outliers detected' and 'outliers excluded from the fit'. To simplify the wording, we just talk about 'outliers detected'.

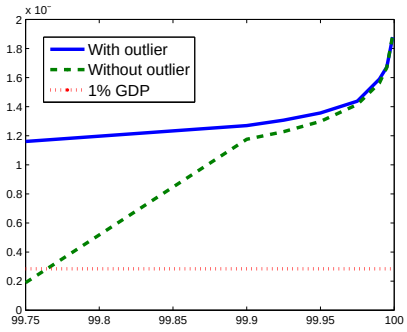

Fig. 4 Losses predicted by SYMBOL with outliers (continuous line) and without (dashed line) for various percentiles. The horizontal dotted line identifies 1% of Austria GDP.

nals exceeding a Bonferronised 99% confidence level, instead of the standard 99% simultaneous level. In Figure 2, this more conservative confidence band is the flat plain thick line. Details on the theoretical and the Bonferroni bands can be found in Atkinson and Riani (2006).

The R^2 and p-value associated to testing the hypothesis of no correlation for the 29 records in the original dataset are respectively 0.1059 and 0.0850. Once the 14 outliers detected by the standard Forward Search algorithm are excluded from the fit, the R^2 raises to almost 1 and the hypothesis of no correlation is drastically rejected (p-value is in the order of 10^{30}). Similar, but less extreme results, are obtained with the two more conservative methods: with the R^2-relaxed method we get 2 outliers and a final $R^2 = 0.994$; with the method further relaxed with Bonferroni bands, we get 1 outlier and final $R^2 = 0.9935$. In both cases, as for the default algorithm, the hypothesis of no correlation is still drastically rejected.

Since our final goal is to have reliable input datasets for the SYMBOL model, we report briefly the results obtained on the Austrian dataset with and without the outliers, which we expect potentially responsible of unreliable results. SYMBOL was run under the hypotheses that

- all banks have a capitalization satisfying Basel II requirements;
- contagion effect between banks takes place.

With the full contaminated dataset, SYMBOL predicts a much higher level of losses, which does not match with other analyses we have carried out. Figure 4 compares this manifestly wrong prediction with that obtained after excluding the most extreme outlier in the Austrian dataset.

In the absence of an automatic detection method for such anomalies in the SYMBOL input datasets, the analyst would be able to identify the problematic cases at the cost of thousands of runs. Then, to find the source of the problems, he would be forced to scrutinize manually the datasets for incoherent values. For large countries (e.g. Germany has about 1400 banks), this is practically infeasible.

6 Final remarks

The approach described in the paper, will be included in the creation of a banking database for the European Commission Member States, on which we are currently working on. Our final objective is to produce a plausible picture of the banking systems in the case of financial crises. The problem of missing values and outliers (or incorrect data, in general) may have a dramatic impact on results of a simulation exercise, which in our case is done using our SYMBOL model.

We have found that the Forward Search approach is able to efficiently detect significant and operationally relevant data anomalies. Compared to other robust methods that can be applied similarly, the Forward Search has the advantage to naturally deal with the perfect fit problem, hence avoiding to remove false/marginal outliers in presence of high data collinearity. We have seen that outlier detection has also an impact on imputation: databases where outliers have been ruled out are used to reconstruct missing values, giving stronger importance to the more reliable and representative dataset values.

It is worth mentioning that, once outliers are detected, it would be important to go directly to the source (i.e. the annual report of the banks for which the problem has occurred) and check the coherence of the cases detected with the values entered in Bankscope. If the values in the source documents differ from those in Bankscope, a simple data correction will fix the situation. If this is not the case (i.e. the data are identical), then a deeper problem arise: should we keep the values as they are and let the model use them, or these kind of unexpected/extreme values should be used to reconsider the way we model the banking system?

References

Atkinson A. C., Riani M. (2000), *Robust Diagnostic Regression Analysis*, Springer, New York.

Atkinson A. C., Riani M. (2006), Distribution theory and simulations for tests of outliers in regression, *Journal of Computational and Graphical Statistics*, 15, 460–476.

Atkinson, A. C., Riani, M., Cerioli, A., (2010). The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society* 39, 19 117-134.

De Lisa, R., Zedda, S., Vallascas, F., Campolongo, F., Marchesi, M., (2011), Modelling Deposit Insurance Scheme Losses in a Basel 2 Framework. *Journal of Financial Services Research* vol. 40, n. 3, p. 123-141.

Garca-Escudero, L.A., Gordaliza, A., Mayo-Iscar, A., San Martin, R., (2010). Robust clusterwise linear regression through trimming. *Computational Statistics and Data Analysis* 54, 3057-3069. doi:10.1016/j.csda.2009.07.002.

Maronna, R.A., Martin, D.R., Yohai, V.J., (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Riani M., Perrotta D. and Torti F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration, *Chemometrics and Intelligent Laboratory Systems*, in press doi 10.1016/j.chemolab.2012.03.017

Torti F., Perrotta D., Atkinson A.C. and Riani M. (2012). Benchmark testing of algorithms for very robust regression: FS, LMS and LTS, *Computational Statistics and Data Analysis*, 56, 2501-2512.