

Analysis of complex data in official statistics

Analisi di dati complessi in statistiche ufficiali

Andrea Cerasa, Francesca Torti, Domenico Perrotta

Abstract International trade data are often affected by multiple linear populations and heteroscedasticity. An immediate consequence is the false declaration of outliers. We propose the monitoring of the White test statistic through the Forward Search as a new robust tool to test the presence of heteroscedasticity. We show that, if the data are considered on a monthly basis, the heteroscedastic problem can be often bypassed.

Abstract *I dati del commercio internazionale sono spesso affetti da popolazioni multiple ed eteroschedasticità. Una conseguenza diretta di questi comportamenti è la falsa dichiarazione di outlier. Proponiamo inoltre il monitoraggio del test di White attraverso la Forward Search come nuovo strumento robusto per verificare la presenza di eteroschedasticità. Si mostra inoltre come, analizzando i dati su base mensile, tale problema possa spesso essere evitato.*

Key words: trade data, multiple populations, heteroscedasticity, Forward Search.

1 Introduction

The international trade between EU Member States and third countries produces a huge amount of data which are firstly collected by the Customs and then monthly aggregated by national statistical offices (e.g. the Italian ISTAT). The analysis of the resulting dataset through suitable statistical procedures is usually focused on the

Andrea Cerasa

European Commission, Joint Research Centre, e-mail: andrea.cerasa@jrc.ec.europa.eu

Francesca Torti

European Commission, Joint Research Centre, e-mail: francesca.torti@jrc.ec.europa.eu

Domenico Perrotta

European Commission, Joint Research Centre e-mail: domenico.perrotta@ec.europa.eu

detection of anomalies of various kinds: recording errors, specific market price dynamics or more in general transactions which are not in line with the market. Monitoring such discrepancies is of primary interest for EU authorities as they could hide unfair or fraudulent commercial behaviors, such as trade-based money laundering, dumping, import duty evasion. All these can have a big negative impact on the EU economy, the EU budget and on the internal competition of EU market. The statistical treatment of such data requires the availability of robust methods, in order to obtain results that are not affected by the irregularities that we are looking for.

One of the robust statistical techniques most successfully applied in this context is the Forward Search (FS), “...a powerful general method for detecting unidentified subsets and masked outliers and for determining their effect on models fitted data” [4]. In [4], the FS is applied on imports of a fishery product that is now available in the FSDA toolbox [5]. The FS output of the regression of volume of trades on quantities traded clearly points out the presence of a point-wise outlier and of a group of transactions occurred at a price sensibly lower than the normal (“fair”) price of the product. Unfortunately, international trade dataset are sometimes characterized by two general problems that dramatically affect the good properties of FS and could lead to misleading conclusions: the presence of multiple linear structures and the heteroscedasticity. They can be considered as a direct consequence of the limits of products’ classification, whose categories are often not enough precise to distinguish the different quality levels of the products in the same category. This topic has been extensively debated in several official documents regarding the possibility of using custom data for the calculation of Import/Export Price Indexes. See [3].

A typical case of multiple populations is presented in the left panel of Figure 1, where the exchanged quantities of an ornamental fish and the corresponding volumes are plotted. It is possible to clearly detect at least three linear relations, representing three different prices. It is obvious that no robust method based on a linear relation between quantity and volume, like the FS (right panel), would be able to disentangle the three relations and to provide a plausible result.

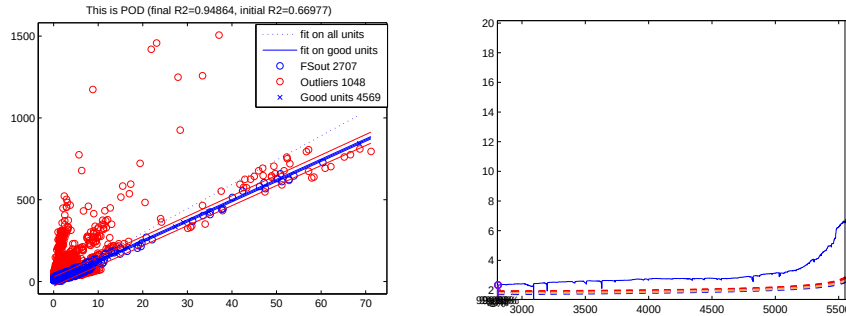


Fig. 1 EU imports of an ornamental fish. Data represent monthly aggregates for the period starting in January 2009 until December 2011. Left panel: scatterplot of VALUE against QUANTITY exchanged. Right panel: forward plot of the Minimum Deletion Residual

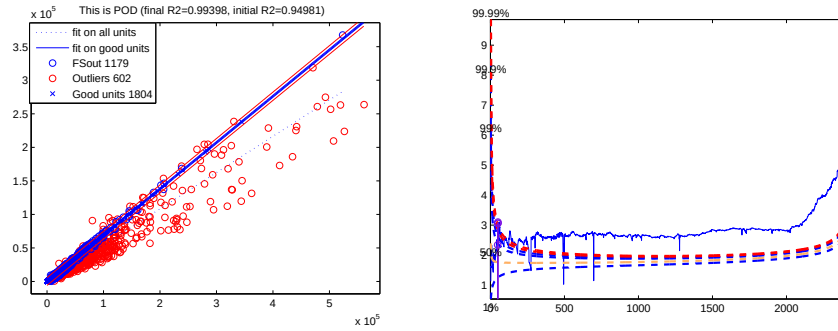


Fig. 2 EU imports of a mineral product. Data represent monthly aggregates for the period starting in January 2009 until December 2011. Left panel: scatterplot of VALUE against QUANTITY exchanged. Right panel: forward plot of the Minimum Deletion Residual

Left panel of Figure 2 represents instead a case of heteroscedastic trade data; the good considered here is a mineral product. As the data show, the assumption of homoskedasticity for the OLS residuals is clearly violated. Again, the results and the conclusions based on the FS regression (right panel) could be affected by the violation of such a fundamental assumption and should be evaluated carefully. Therefore, the availability of an instrument which is able to point out when we can trust the single normal model and therefore the FS output and when instead it has been probably contaminated by one of the mentioned problems, is of primary importance. It could help us indeed to distinguish the real discrepancies in trade data from the spurious ones.

The paper is organized as follows. In the next section, the consequences of the application of the FS on dataset characterized by multiple populations and heteroscedasticity are analyzed. Then, the use of the White test as a diagnostic tool for monitoring and evaluating the departure from the homoscedasticity assumption is presented. Finally, a monthly price approach is introduced as a possible solution to heteroscedasticity.

2 The Forward Search in presence of multiple populations and heteroscedasticity

The basic idea of the FS [1] is to start from a small, robustly chosen, subset of the data and to fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are clearly revealed by diagnostic monitoring. If there is only one population the increasing fitting from a few observations to all observations will be stable. Otherwise, if in the data there are outliers or groups, there will be a point where the stable progression of fits is interrupted. In the classical regression model we have one univariate response Y and v explanatory variables

$X_1, \dots, X_v$ satisfying

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_v x_{iv} \quad (1)$$

under the usual assumptions. Each subsample is obtained by looking at the n squared regression residuals

$$e_i^2(m) = [y_i - \{\hat{\beta}_0(m) + \hat{\beta}_1(m)x_{i1} + \dots + \hat{\beta}_v(m)x_{iv}\}]^2 \quad i = 1, \dots, n$$

computed from the OLS estimate of beta at step m . $S(m+1)$ is defined as the subset of observations corresponding to the $m+1$ smallest squared residuals $e_i^2(m)$. The search starts from an outlier-free subset of m_0 observations. Usually $m_0 = v+1$, with $S(m_0)$ chosen through the least median of squares criterion of [6].

In our context, the response variable is the value of the transactions whereas the explanatory one is the quantity of product exchanged. The output of the FS applied to the ornamental fish trades of Figure 1 and to the mineral product data of Figure 2 are represented in the corresponding right panels. In the case of multiple population, the FS automatically selects the linear structure with the highest leverage and declares outliers all the points of the other structures. However, this declaration may be misleading since the points detected could simply represent prices of different quality levels. Besides, in the case of heteroscedasticity the FS tends to over-declare the outliers. It seems indeed that most of them are actually coherent with the model, once heteroscedasticity is taken into account. Moreover, the line estimated by the method does not seem to represent the central tendency of the data.

Then the main consequence of applying the FS in presence of heteroscedasticity or multiple linear structures is an over-declaration of outliers. This represents a serious problem in the anti-fraud context because each outlier should be analyzed in detail in order to determine whether it may hide or not a fraudulent behavior and to initiate a costly investigation. In other words, declaring false outliers means wasting economic resources and should be thus avoided.

3 Forward plot of the White test

In order to verify the presence of heteroscedasticity, we can use the well known White test [2]. To avoid the bias introduced by the possible presence of outliers, we propose to robustify the test by monitoring the statistic with the FS. More precisely, the quantity monitored is the coefficient of determination of the following auxiliary regression, that we report for the simple case of one independent variable:

$$e_i^2(m) = \hat{\alpha}_0(m) + \hat{\alpha}_1(m)x_{i1} + \hat{\alpha}_2(m)x_{i1}^2 + u_i \quad i = 1, \dots, n \quad (2)$$

Figure 3 presents the forward plots of the White test statistic for the two datasets considered. For the mineral product (left panel) the plot shows that after about 500 steps the statistic exceeds the 90%, 95% and 99% confidence bands, obtained through montecarlo simulations, represented with dashed lines. This highlights very

clearly the presence of heteroscedasticity in the dataset. As a consequence, the forward plot of the minimum deletion residual, traditionally monitored in the regression context, exceeds the bands from the very first steps (right panel of Figure 2). In conclusion many outliers are wrongly declared. The right panel of Figure 3 demonstrates that the forward plot for the White test can be also used to detect situations where data contain different linear structures, like for the ornamental fish dataset.

As a result, in our operational context we can use the proposed approach to highlight cases when the standard FS outlier detection model is not reliable and more complex models have to be adopted. FS models for the heteroscedastic and multiple population cases are matter of future work. The next section proposes an approach to work around the problem, specifically applicable in our context.

4 Monthly fair prices

In trade data the latent factor which often explains the presence of heteroscedasticity is the time factor. The price of a product is indeed subject to changes during time due to several reasons (inflation, technical improvements, seasonal effects). As a result, considering simultaneously data related to different time periods could produce multiple groups. On the other hand, heteroscedasticity could also be seen as a particular case of multiple populations. In fact, if the linear structures are so close to almost overlap, it would be difficult to distinguish their graphical representation from one originating from a single population with heteroscedastic error.

But if we analyze each single time period independently, then we should be able to isolate every linear structure, provided that the product is well and univocally defined and that there are no other latent factors. As an example, Figure 4 shows in its left panel the scatterplot for the January data of the mineral product dataset. The data lie on an almost perfect line and there are no signs of heteroscedasticity. As a further confirmation, the right panel shows the corresponding forward White

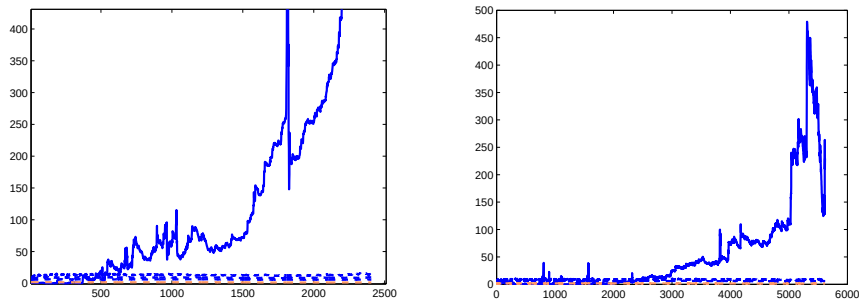


Fig. 3 Forward plot of the White test statistic for the mineral product (left panel) and the ornamental fishery (right panel) datasets.

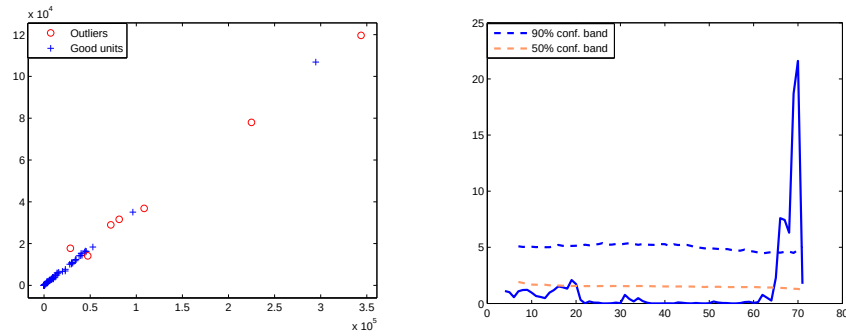


Fig. 4 EU imports of a mineral product, for January 2009 only. Left panel: scatterplot with outliers detected by the FS. Right panel: forward plot of the White test statistic.

test plot. The trajectory confirms indeed the absence of heteroscedasticity: the curve exits from the 90% band only in the last step of the FS, which is indication of the presence of some outliers but not of heteroscedasticity. Moreover, now the FS regression is able to capture the central tendency of the data and to give a reliable and robust estimate of the monthly price of the product. Scatterplots of other months show similar results.

Another advantage of the monthly approach is that it gives the possibility to study the trade price dynamics. This is relevant for detecting general patterns of economic interest such as unexpected peaks, seasonal components, markets trends and so on.

References

1. Atkinson A.C., Riani M., Robust Diagnostic Regression Analysis, Springer Verlag, New York (2000).
2. Greene, W.H., Econometric analysis, Prentice Hall, 7th Edition (2008).
3. ILO, IMF, OECD, Eurostat, United Nations Economic Commission for Europe, The World Bank: Export and Import Price Index Manual (2009).
4. Riani M., Cerioli A., Atkinson A.C., Perrotta D., Torti F. Fitting Mixtures of Regression Lines with the Forward Search. In: Mining Massive Data Sets for Security F. Fogelman-Souli et al. EDS. (pp. 271-286). IOS Press, Amsterdam (The Netherlands)(2008).
5. Riani M., Perrotta D., Torti F., FSDA: A MATLAB toolbox for robust analysis and interactive data exploration, Chemometrics and Intelligent Laboratory Systems, 116, p. 17-32, doi:10.1016/j.chemolab.2012.03.017.
6. Rousseeuw P. J., Leroy A. M., Robust regression and outlier detection. Wiley-Interscience, New York (1987).