ARGUMENTA
OECONOMICA
CRACOVIENSIA

No 1–2(26–27) • 2024
ISSN 1642-168X
e-ISSN 2545-3866
AOC, 2024; 1–2(26–27): 73–83
https://doi.org/10.15678/AOC.2024.2604

Klaudia Lenart

ANOMALY DETECTION BASED ON MEASURES OF INFLUENCE FOR MODELLING ECONOMIC PHENOMENA

Abstract

Objective: Anomalies are data points (or sequences of points) for which relationships between variables are significantly different to those that can be observed under normal circumstances. Their presence in data used for estimating an econometric model may significantly influence the values of the parameter estimates. The result is a skewed projection of the real world and less accurate forecasts. The purpose of this study is to propose a method of identifying anomalies in data based on their influence on the regression function parameter estimates.
Research Design & Methods: This paper proposes a method of detecting anomalies by identifying data points with the most significant influence on the estimates of the model parameters using permutations of the dataset. The method was applied to data generated using copula functions, and anomalies were generated by changing the marginal distribution of the dependent variable. A fixed percentage of data points was identified as anomalies and removed. This method was compared with one based on distance to *k*-nearest neighbours.
Findings: The exclusion of the anomalies identified by the proposed method resulted in models with a significantly lower prediction error. Additionally the method based on influence of the observations was more accurate in identifying anomalies.
*Implications/Recommendation*s: Excluding anomalies can be an important stage in data preparation for estimating an econometric model, particularly when one aims to predict. Nevertheless, it is important to keep in mind the risk of deleting valid observations from the dataset.
Contribution: In the conducted simulation study removing the observations identified as anomalies resulted in models with a significantly lower prediction error, even when

Klaudia Lenart, University of Economics in Katowice, Doctoral School, 1 Maja 50, 40-287 Katowice, Poland, e-mail: klaudia.lenart@edu.uekat.pl, ORCID: https://orcid.org/0000-0001-8135-9362.

some typical observations were incorrectly classified as anomalies. The method based on influence on the model parameter estimates allowed for accurate identification of anomalies although it was dependent on correct prediction of the percentage of anomalous observations that would appear in the data.

Keywords: anomaly detection, influential observations, econometric model, outliers.
JEL Classification: C15, C18, C51.

1. Introduction

Econometric models should reflect the relationships between variables that can be observed in the real world. For this to be possible, data quality needs to be ensured, and, in particular, any data that was corrupted or observed under unusual circumstances needs to be excluded from the dataset. A good example of this is a model predicting crop yield based on the soil quality and fertiliser used. If data from regions affected by a flood is included in the analysis, the effect of the independent variables on the crop yield will become much less clear. Moreover, it is likely that the observations from those regions will be influential and, as such, will have a significant impact on the values of the model parameter estimates. This amplifies the issue of the estimated model providing a skewed projection of the real world but can also allow identification of the anomalous data.

In this paper a simulation study was used to test how accurately anomalies in data can be identified based on the influence on the model parameter estimates. This method was compared with a method based on the measurement of distance.

2. Anomalies and Influential Observations

The range of applications of anomaly detection means it is impossible to formulate a universal and unambiguous definition of an anomaly. The general definition of an anomaly is an observation (or in certain cases a group of observations) that is unusual for a given dataset (Chandola, Banerjee & Kumar, 2009), but what constitutes an unusual observation will be heavily dependent on the goal of the analysis. Additionally, the cost of making an error by classifying an observation as a false positive or false negative will differ depending on the anomaly detection application.

Hawkins (1980, p. 1) defines an outlier as “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”. This definition is especially important for the subject of this paper as it points to the reason why anomaly

detection is important as a stage in data preparation during econometric model estimation. The foundation of multivariate statistical methods is the assumption that the observations all come from the same multivariate distribution. This is not true if the dataset contains anomalous observations that are an effect of unusual circumstances or have been corrupted (Trzęsiok, 2014). It should be emphasised that, as noted by Hawkins (1980), an unusual observation may either be generated from a different distribution or just be an unlikely realisation of a variable. Excluding the latter from the dataset may lead to incorrect analysis. Most researchers treat the terms "outlier" and "anomaly" as synonyms (Chandola, Banerjee & Kumar, 2009; Aggarwal, 2017; Mehrotra, Mohan & Huang, 2017). For the sake of clarity in this paper the term "anomaly" will be used when referring to observations generated from a different multivariate distribution while the word "outlier" will refer to all untypical observations. This distinction is especially important because, during the simulation study the desired result is detection observations generated from a different distribution and classifying them as anomalies.

This paper focuses on anomaly detection as a stage of data preparation during estimation of an econometric model. This is done to exclude from the model observations that misrepresent the relationships between variables which can be observed under normal circumstances, usually due to some factor that is not included in the available set of variables. To simulate this during the data generation the multivariate distribution of the dependent and independent variables will be changed.

Influential observation is defined by Belsley, Kuh and Welsch (1980) as an observation that either individually or together with several other observations has a demonstrably larger impact on the calculated values of various estimates than most other observations. An outlier will not necessarily be an influential observation. An observation may have an unusually high or low value of all variables but without significant differences in the proportions between variables, thus not impacting the estimates of model parameters (Draper & John, 1981). Similarly there may be influential observations for which all values of all variables considered separately are typical for the dataset – it is only the skewed relationships between the variables that impact the parameters' estimates. The fact that disrupted relationships between the variables are often the cause of an observation having an unusual level of influence on model parameter estimates should allow the identification of anomalies by identifying the influential observations.

There are several statistics proposed for measuring the influence observations have on the estimated model. Cook (1977) proposes statistics based on differences in the *y* estimates made by the models. A different approach is shown in the DFBETAS measure (Belsley, Kuh & Welsch, 1980) that is based on the differences in model parameter estimates. A variation of this approach was used during this study.

An important thing to note about the measurements mentioned above is that they are not statistical tests allowing verification of a hypothesis. Several methods for identifying the cut-off points for these measures have been proposed, but there is no unambiguous way of pinpointing the level of influence an observation must have on the model estimation to be classified as influential.

The terms "influential observation" and "outlier" are often connected. It is a typical approach to identify outliers based on the distance between observations (Mehrotra, Mohan & Huang, 2017). To compare a distance and influence based approach distance to *k*-nearest neighbours method will be used.

3. Generating Multidimensional Data Using Copulas

In order to carry out a simulation study a method for generating a multidimensional dataset with fixed relationships between variables is needed. Copula functions can be used for this purpose (Heilpern, 2007).

An m-dimensional copula is a function C with domain $[0, 1]^m$ when the following conditions are met (Nelsen, 1998):

– $C(1, \ldots, 1, a_n, 1, \ldots, 1) = a_n$,
– $C(a_1, \ldots, a_m) = 0$ if $a_i = 0$ for every $i \leq m$,
– C is m-increasing.

The foundation of the theory of copulas as well as its applications in statistics can be found in Sklar's theorem which was first published in (Sklar, 1959).

Let H be a joint distribution function with margins F and G. Then there exists a copula C such that for all x, y in $\overline{R}$,

$$H(x, y) = C(F(x), G(y)). \tag{1}$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $RanF \times RanG$. Conversely, if C is a copula and F and G are distribution functions, then the function H is a joint distribution function with margins F and G.

Over the years many copulas have been proposed differing in dependence structure and having unique properties. The normal copula, also known as the Gaussian copula was first described in (Lee, 1983). An important characteristic of this copula is the fact that the values of the correlation parameter θ, can be positive or negative, as long as it meets the condition $-1 < \theta < 1$. The normal copula can be written as:

$$C(u_1, u_2, \theta) = \Phi_G\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \theta\right), \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution (Trivedi & Zimmer, 2007).

4. The Proposed Method

The proposed method allows identification of the observations which have the biggest influence on the values of the estimated parameters, so that they can be classified as anomalies.

The use of the method requires that for a *k*-dimensional dataset consisting of *n* observations a linear regression function form is known as:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_1 x_{1i} + \ldots + \alpha_k x_{ki} + \xi_i. \tag{3}$$

The method relies on determining by how much removing each observation from the dataset influences the values of the estimated parameters of the linear regression function. This influence is quantified using the proposed *T* statistic, the calculation of which is described below.

The procedure of calculation of the *T* statistic consists of the following steps:

1. Estimation of the parameters of the linear regression function based on the entire dataset. The estimates are denoted as $a_{00}, \ldots, a_{0k}$.
2. For each $i \in [1, 2, \ldots, n]$ a subset of the data is created by excluding the *i*th observation.
3. Estimation of the parameters of the linear regression function based on the created subsets of data. The estimates are denoted as $a_{i0}, \ldots, a_{ik}$.
4. Calculation of the differences between the values of the parameters estimated based on the entire dataset and the *i*th subset:

$$R_{ij} = a_{ij} - a_{0j}. \tag{4}$$

5. Standardising the differences for each $j \in [0, 1, \ldots, k]$:

$$RS_{ij} = \frac{R_{ij} - \overline{R_i}}{S_{R_j}}, \tag{5}$$

where S_{R_j} is a standard deviation of the jth parameter estimates calculated on the subsets of data.

6. Calculation of the T statistic values for each $i \in [1, 2, \ldots, n]$ using the formula:

$$T_i = \sum_{j=0}^{k} \left| RS_{ij} \right|. \tag{6}$$

After the values of the T statistic are calculated the T_p can be determined by finding the quantile of the T statistic's vector corresponding to the percentage of the observations that will be identified as anomalies, which equals p. In this paper the value of p will always be equal to 5% although the question of determining p based on the values of the calculated T statistic is worth further research.

The ith observation is classified as an anomaly if the following condition is met:

$$T_i \geq T_p. \tag{7}$$

5. The Simulation Study

To test if the proposed method allows for accurate classification of the anomalies in the dataset a simulation study was conducted. The data was generated using copulas, implemented in the R programme (Hofert *et al.*, 2018), and the anomalies were added by changing the marginal distribution of the dependent variable. Two variants were considered, as shown in Table 1. In variant A beta distribution B(s_1, s_2) was used as the marginal distribution of the variables. In variant B the marginal distribution of the variables was a normal distribution N(μ, σ^2), where the mean of the variable was equal to μ and the variance was equal to σ^2. For the anomalous observations in variant B the distribution of y was changed to an exponential distribution with the λ parameter equal to 0.5.

Table 1. Marginal Distributions of the Generated Data

Variable	Variant A		Variant B	
	Typical Observations	Anomalies	Typical Observations	Anomalies
y	B(2, 6)	B(6, 2)	N(2, 2)	Exp(0.5)
x_1	B(2, 6)	B(2, 6)	N(4, 4)	N(4, 4)
x_2	B(2, 6)	B(2, 6)	N(10, 4)	N(10, 4)
x_3	B(2, 6)	B(2, 6)	N(10, 7)	N(10, 7)
x_4	B(2, 6)	B(2, 6)	N(3, 1.2)	N(3, 1.2)

Source: author's own work.

For each variant a total of 1,000 observations was generated three times. Each time the percentage of the anomalous observations was different (consecutively 2%, 5% and 10%) so that the consequences of the assumed percentage of the anomalies being over or underestimated could be tested. This data was then classified using the proposed method and the method using distance to *k*-nearest neighbours. The method described by Mehrotra, Mohan and Huang (2017) requires that for each *i*th datapoint the *k*-nearest neighbours are identified. We define *Near*(*i*, *j*) as the *j*th nearest neighbour of the *i*th observation and *d*(*a*, *b*) as a function of distance between two datapoints, *a* and *b*. The statistic used in this method is calculated as:

$$\alpha(p) = \sum_{j=1}^{k} d(p, Near(p, j)). \tag{8}$$

Similarly to the proposed method, a quantile of the $\alpha(p)$ vector can be used as a cut-off point. Then all observations for which the sum of distance to *k*-nearest neighbours is greater or equal to that quantile is identified as an anomaly.

As shown in Figure 1, the proposed method consistently achieved higher accuracy, the only exception being variant A with 2% of the anomalies, where both methods detected all anomalies correctly (note that, as the assumed percentage of the anomalous observations was different from the actual proportions, the maximal accuracy that could be achieved in this case was 97%). It is worth noting that when the assumed and real percentage of anomalies was equal for both variants, the proposed method correctly classified all observations.

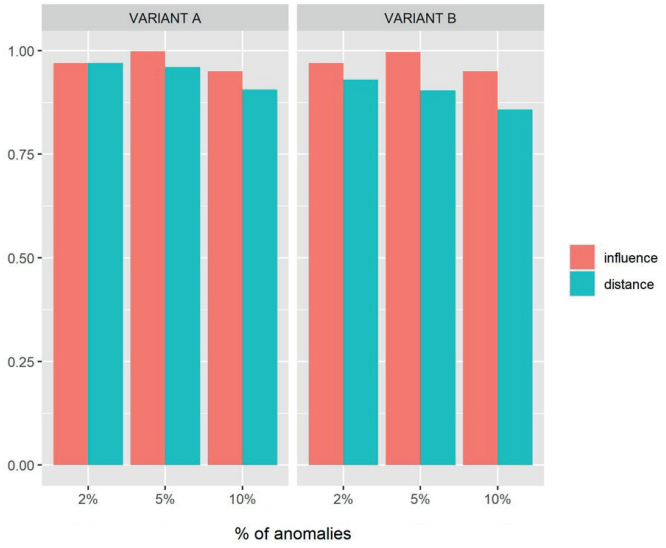

Fig. 1. Accuracy of the Proposed Method and the Method Using Distance to 20 Nearest Neighbours

Source: author's own work in the R programme.

The generated datasets were unbalanced, as the anomalies constituted only between 2–10% of the data instead of the observations being evenly spread between both classes. For this reason it is important to analyse not only accuracy but also recall. As shown in Figure 2, in most cases the proposed method correctly classified more than 90% of the anomalous observations, with the exception of datasets with 10% of the anomalies, where the maximal recall that could be achieved was 50%.

For each dataset, observations identified as anomalies were excluded. So prepared data was used to estimate parameters of the linear regression function. 500 new typical observations were generated for each dataset, so that performance of the models can be compared. The results are shown in Table 2.

As shown in Figure 2, excluding anomalies from the dataset had a positive impact on the precision of the prediction. Additionally, better accuracy of the proposed permutation method allowed for a bigger reduction in value of RMSE (root mean square error).

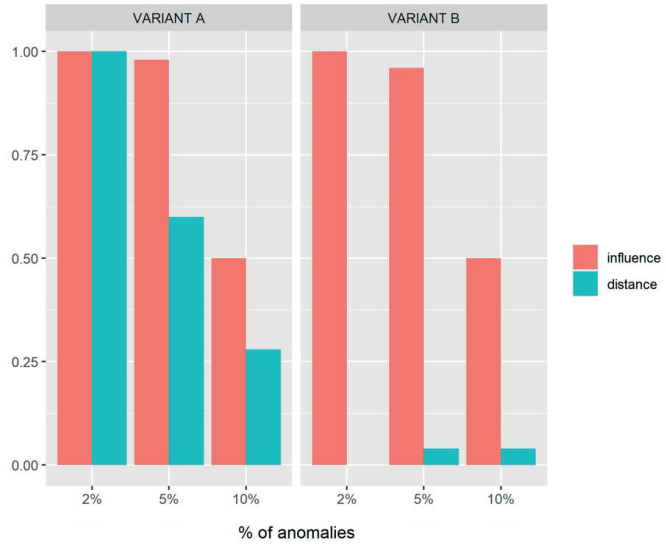

Fig. 2. Recall of the Proposed Method and the Method Using Distance to 20 Nearest Neighbours

Source: author's own work in the R programme.

Table 2. RMSE Values for Models with Anomalies Identified Using Different Methods Excluded from the Training Dataset

Data	Percent of Anomalies	Method		
		None	Distance	Influence
Variant A	2	0.0303	0.02846	0.02968
	5	0.03765	0.0313	0.0297
	10	0.0580	0.0493	0.0297
Variant B	2	0.0063	0.0070	0.0000
	5	0.0079	0.0105	0.0000
	10	0.0247	0.0287	0.0000

Source: author's own work.

6. Conclusion

The simulation study showed that removal of the anomalies based on the influence the observations had on the model parameter estimates allowed a significant reduction in the value of RMSE. Additionally, the proposed method was identifying observations for which the multivariate

distribution of the dependent and independent variables was changed more accurately than the method based on distance. It is worth noting that even the anomalies not identified by the proposed method would not have a large impact on the model parameter estimates.

There is no definitive method of identifying the cut-off point of the level of influence on the model parameter estimates an observation has to have to be considered influential. Because of this, cases where the assumed percentage of anomalies is larger and smaller than the real percentage were considered. In both cases, no significant increase of RMSE occurred when compared to the model estimated using the entire dataset. When more observations than necessary were deleted, there was no significant loss in precision of prediction information. However, it is important to note that, as this study is using generated data, the homogeneity of the data (excluding added anomalies) as well as the large number of observations could help reduce the consequences of deleting additional observations.

This paper shows that identifying anomalies based on the influence the observations had on the model parameter estimates may be an important step in model estimation, especially if the researcher expects that the data may contain anomalous observations. Additionally, it is often more important for the precision of the prediction than the more widely known issue of identifying outliers based on measurements of distance between observations. Nevertheless, it is necessary to keep in mind that outliers can occur naturally, particularly in the case of fat-tailed distributions. Because of this, the observations identified as anomalies should be carefully examined before their removal from the dataset.

The simulation study examined the accuracy of the proposed method when applied to a simple linear regression model. Further research could investigate applying this method to more complex models. In addition, more research to examine the sensitivity of the proposed method's accuracy to changes in the homogeneity of the generated data is needed.

References

Aggarwal, C. C. (2017). *Outlier Analysis*. Springer. https://doi.org/10.1007/978-3-319-47578-3

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons. https://doi.org/10.1002/0471725153

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15–18. https://doi.org/10.2307/1268249

Draper, N. R., & John, J. A. (1981). Influential Observations and Outliers in Regression. *Technometrics*, *23*(1), 21–26. https://doi.org/10.2307/1267971

Hawkins, D. M. (1980). *Identification of Outliers* (Vol. 11). Springer. https://doi.org/10.1007/978-94-015-3994-4

Heilpern, S. (2007). *Funkcje łączące*. Wydawnictwo Akademii Ekonomicznej im. Oskara Langego.

Hofert, M., Kojadinovic, I., Mächler, M., & Yan, J. (2018). *Elements of Copula Modeling with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-89635-9

Lee, L.-F. (1983). Generalized Econometric Models with Selectivity. *Econometrica: Journal of the Econometric Society*, *51*(2), 507–512. https://doi.org/10.2307/1912003

Mehrotra, K. G., Mohan, C. K., & Huang, H. (2017). *Anomaly Detection Principles and Algorithms* (Vol. 1). Springer. https://doi.org/10.1007/978-3-319-67526-8

Nelsen, R. B. (1998). *An Introduction to Copulas*. Springer Science & Business Media.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP*, *8*(3), 229–231.

Trivedi, P. K., & Zimmer, D. M. (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends® in Econometrics*, *1*(1), 1–111. https://doi.org/10.1561/0800000005

Trzęsiok, M. (2014). O jakości danych w kontekście obserwacji oddalonych w wielowymiarowej analizie regresji. *Studia Ekonomiczne*, 191, 75–88.