International Journal of Fuzzy Systems

OUTLIER DETECTION ALGORITHMS OVER FUZZY DATA WITH WEIGHTED LEAST SQUARES

--Manuscript Draft--

Manuscript Number:	IJFS-D-19-00829R2
Full Title:	OUTLIER DETECTION ALGORITHMS OVER FUZZY DATA WITH WEIGHTED LEAST SQUARES
Article Type:	S.I. : Fuzzy models for Business Analytics
Funding Information:	Hungarian National Research, Development and Innovation Office (NKFIH) (129528, KK) — Prof. Krasimir Kolev Higher Education Institutional Excellence Programme of the Ministry of Human Capacities in Hungary (Molecular Biology thematic programme of Semmelweis University (KK)) — Prof. Krasimir Kolev University of Tasmania (RT.112222) — Prof. Kiril Tenekedjiev
Abstract:	Abstract In the classical leave-one-out procedure for outlier detection in regression analysis, we exclude an observation, then construct a model on the remaining data. If the difference between predicted and observed value is high we declare this value an outlier. As a rule, those procedures utilize single comparison testing. The problem becomes much harder when the observations can be associated with a given degree of membership to an underlying population and the outlier detection should be generalized to operate over fuzzy data. We present a new approach for outlier that operates over fuzzy data using two inter-related algorithms. Due to the way outliers enter the observation sample, they may be of various order of magnitude. To account for this, we divided the outlier detection procedure into cycles. Furthermore, each cycle consists of two phases. In Phase 1 we apply a leave-one-out procedure for each non-outlier in the data set. In Phase 2, all previously declared outliers are subjected to Benjamini-Hochberg step-up multiple testing procedure controlling the false discovery rate, and the non-confirmed outliers can return to the data set. Finally, we construct a regression model over the resulting set of non-outliers. In that way we ensure that a reliable and high-quality regression model is obtained in Phase 1 because the leave-one-out procedure comparatively easily purges the dubious observations due to the single comparison testing. In the same time, the confirmation of the outlier status in relation to the newly obtained high-quality regression model is much harder due to the multiple testing procedure applied hence only the true outliers remain outside the data sample. The two phases in each cycle are a good trade-off between the desire to construct a high-quality model (i.e. over informative data points) and the desire to use as much data points as possible (thus leaving as much observations as possible in the data sample). The number of cycles is user-defined, but the procedures can finalize the analysis in case a cycle with no new outliers is detected. We offer one illustrative example and two other practical case studies (from real-life thrombosis studies) that demonstrate the application and strengths of our algorithms. In the concluding section, we discuss several limitations of our approach and also offer directions for future research. Keywords: regression analysis, leave-one-out method, degree of membership, multiple testing, Benjamini-Hochberg step-up multiple testing, false-discovery rate Highlights: - We develop algorithms for outlier rejection over fuzzy samples using weighted least squares that operate in a given number of cycles - Each cycle has two phases – use single testing leave-one-out procedure for initial purging of data, then confirm the previous outlier status with multiple testing - We offer one illustrative example and two examples from a case study in thrombosis research to show the strength of our cycle-based approach
Corresponding Author:	Natalia Nikolova, PhD University of Tasmania Australian Maritime College

OUTLIER DETECTION ALGORITHMS OVER FUZZY DATA WITH WEIGHTED LEAST SQUARES

Natalia Nikolova[1,2*], Rosa M. Rodríguez[3], Mark Symes[4], Daniela Toneva[5], Krasimir Kolev[6], Kiril Tenekedjiev[7,8]

[1] Australian Maritime College, University of Tasmania
1 Maritime Way, Launceston
7250 TAS, Australia
E-mail: Natalia.Nikolova@utas.edu.au
* Corresponding author

[2] Nikola Vaptsarov Naval Academy – Varna, Faculty of Engineering
73 Vasil Drumev Street
Varna 9026, Bulgaria
E-mail: natalianik@gmail.com

[3] University of Jaén
Campus las lagunillas s/n
23071, Jaén (Spain)
E-mail: rmrodrig@ujaen.es

[4] Australian Maritime College, University of Tasmania
1 Maritime Way, Launceston
7250 TAS, Australia
E-mail: Mark.Symes@utas.edu.au

[5] Technical University – Varna
Faculty of Marine Sciences and Ecology
10 Studentska Str., Varna 9010, Bulgaria
E-mail: dtoneva@abv.bg

[6] Department of Medical Biochemistry, Semmelweis University,
1085 Budapest, Üllői út 26., Hungary
E-mail: Kolev.Krasimir@med.semmelweis-univ.hu

[7] Australian Maritime College, University of Tasmania
1 Maritime Way, Launceston
7250 TAS, Australia
E-mail: Kiril.Tenekedjiev@utas.edu.au

[8] Nikola Vaptsarov Naval Academy – Varna, Faculty of Engineering
73 Vasil Drumev Street
Varna 9026, Bulgaria
E-mail: Kiril.Tenekedjiev@fulbrightmail.org

ORCID Nikolova: http://orcid.org/0000-0001-6160-6282
ORCID Rodriguez: https://orcid.org/0000-0002-1736-8915
ORCID Symes: https://orcid.org/0000-0003-2241-4995
ORCID Toneva: https://orcid.org/0000-0003-1599-395X
ORCID Kolev: https://orcid.org/0000-0002-5612-004X
ORCID Tenekedjiev: https://orcid.org/0000-0003-3549-0671

Abstract In the classical leave-one-out procedure for outlier detection in regression analysis, we exclude an observation, then construct a model on the remaining data. If the difference between predicted and observed value is high we declare this value an outlier. As a rule, those procedures utilize single comparison testing. The problem becomes much harder when the observations can be associated with a given degree of membership to an underlying population and the outlier detection should be generalized to operate over fuzzy data. We present a new approach for outlier that operates over fuzzy data using two inter-related algorithms. Due to the way outliers enter the observation sample, they may be of various order of magnitude. To account for this, we divided the outlier detection procedure into cycles. Furthermore, each cycle consists of two phases. In Phase 1 we apply a leave-one-out procedure for each non-outlier in the data set. In Phase 2, all previously declared outliers are subjected to Benjamini-Hochberg step-up multiple testing procedure controlling the false discovery rate, and the non-confirmed outliers can return to the data set. Finally, we construct a regression model over the resulting set of non-outliers. In that way we ensure that a reliable and high-quality regression model is obtained in Phase 1 because the leave-one-out procedure comparatively easily purges the dubious observations due to the single comparison testing. In the same time, the confirmation of the outlier status in relation to the newly obtained high-quality regression model is much harder due to the multiple testing procedure applied hence only the true outliers remain outside the data sample. The two phases in each cycle are a good trade-off between the desire to construct a high-quality model (i.e. over informative data points) and the desire to use as much data points as possible (thus leaving as much observations as possible in the data sample). The number of cycles is user-defined, but the procedures can finalize the analysis in case a cycle with no new outliers is detected. We offer one illustrative example and two other practical case studies (from real-life thrombosis studies) that demonstrate the application and strengths of our algorithms. In the concluding section, we discuss several limitations of our approach and also offer directions for future research.

Keywords: regression analysis, leave-one-out method, degree of membership, multiple testing, Benjamini-Hochberg step-up multiple testing, false-discovery rate

Highlights:

- We develop algorithms for outlier rejection over fuzzy samples using weighted least squares that operate in a given number of cycles
- Each cycle has two phases – use single testing leave-one-out procedure for initial purging of data, then confirm the previous outlier status with multiple testing
- We offer one illustrative example and two examples from a case study in thrombosis research to show the strength of our cycle-based approach

1. INTRODUCTION

Regression analysis aims to construct a linear model of a given process that relates to the relationship between a dependent (response) variable and one or more independent (explanatory, predictor) variables. Using that linear model, we can then make inferences regarding the process. Regression analysis serves for a wide range of predictions and forecasting, as well as to infer causal connections between the predictor and the response variables. The adequacy of the model benefits from a preliminary screening of the input sample for non-informative, misleading and/or erroneous data points, known as outliers [63]. In fact, the proper identification and rejection of outliers contributes to the quality of the regression model a lot more than the size of the input sample. Therefore, a procedure that associates with higher rejection rate (thus improving the quality of the sample) should be preferred over a procedure that has insufficient rejection of outliers (thus aiming to maintain somewhat larger sample size) [29].

Assume the initial sample has *n* observations. A typical procedure to detect an outlier is to exclude the *i*-th observation, construct a model on the remaining *n*–1 data points and measure the difference between the predicted and the observed value. If that difference is above a given threshold, the *i*-th observation is declared an outlier. Then *n* models test *n* hypotheses regarding the outlier status of an *i*-th observation. This procedure is in fact a leave-one-out (LOO) routine to test the performance of a model [46, 73]. It suffers from several drawbacks:

- the errors in the observations may vary significantly in scale and order of magnitude;
- once an observation is detected as an outlier it has no chance to return into the sample;
- all tests have the same significance level α_{crit} , yet multiple testing procedures require to change α_{crit} of the tests.

In a classical setup, we apply regression analysis over crisp data to study crisp relationships between the predictor and response variables [10]. Often, though, we deal with fuzzy data where the data points have some sort of an associated degree of membership to a given Population. Various proposals discuss how fuzzy data enters real-life data analysis [21; 56]. Viertl [70] claims that real-life data rarely comes as precise numbers, but as some form of fuzzy data so statistical analysis needs to be adapted to such data. Coppi in [12] introduces the information paradigm to interpret uncertainty and accommodate fuzzy-possibilities and probabilities approaches within traditional statistical paradigms. Coppi suggests that uncertainty may be associated with the relationship between: a) the response and predictor variables; b) the data sample and the underlying population; c) the sample data points themselves. Probabilistic tools are then suggested for the cases where uncertainty factors appear in isolation. For the cases where a combination of uncertainty factors appear, Coppi suggests the use of fuzzy-possibilistic tools. Those follow the ideas in [23, 24] about possibility theory as the bridging concept between fuzzy sets and probability theory. Uncertainty relevant to the second source of uncertainty in [12] is explored from a fuzzy perspective in other works. For example, Ruspini in [62] measures the resemblance between two worlds by a generalized similarity relation that then allows to interpret the main aspects of fuzzy logic.

There are many discussions on the way fuzzy data enters and modifies the regression analysis procedures specifically (see the work of Chachi and Taheri in [8], as well as the discussion in Section 3). This aspect is greatly emphasized in the review work [10], which shows that within regression analysis, the fuzzy uncertainty may be measured by possibility as per Dubois, Prade [24] and Klir [39]. It further comments that in traditional fuzzy regression models, both predictor and response variables are fuzzy variables, hence their relationship is interpreted by a fuzzy function, using possibilistic tools to construct its distribution. Other studies explore the ways to construct a linear regression model that interprets the connection between fuzzy response and crisp predictor variables [13]. The work [14] further explores this setting to construct an interactive least square based estimation method. Regression analysis is utilized as demonstration of the information paradigm under various scenarios of complexity by Coppi in [12]. Suggestions focuses on regression analysis are also offered by Gao and Gao in [30]. They postulate that the deviation between observed and estimated values in regression analysis originating from either indefiniteness of the system structure or incompleteness of data should be treated as fuzziness and has to be handled by regression analysis with fuzzy data, as proposed by Tanaka et al. in [65, 66]. Tanaka's works pioneered the fuzzy regression analysis domain and explored the degree of the fitting and the vagueness of the model in the fuzzy linear regression process. They did not stress the issues of best fit by residuals in the fuzzy regression setup, which were developed by Diamond's fuzzy least square approach [22] as the fuzzy version of traditional ordinary least squares, and introduced a new distance measure on the space of fuzzy numbers. Jinn et al. [35] elaborated on the impact of outliers (and influential observations) on the model and how the fitting process may hide flaws in the regression model. They developed outlier detection approaches adaptable to fuzzy linear regression that utilized the Cook distance [11; 55]. Their suggestions were best adapted to the general fuzzy linear model, yet for more elaborate cases such as the doubly linear adaptive fuzzy regression model of D'Urso and Gataldi [15], and for D'Urso's fuzzy regression model [14], they only arrived at a procedure that rejects outliers and recalculate residuals.

Our scope of work is to improve the way outlier detection is handles in the presence of fuzzy data. In this paper, we develop procedures to adapt the LOO approach over fuzzy data and improve its performance in outlier detection. Our scope of analysis refers to fuzzy data originating from setups similar to the second uncertainty type of Coppi [12], where each response-predictor pair has a degree of membership to the underlying process of analysis (or to analyzed object). The procedures we propose will run in multiple cycles so that we can deal with the various order of magnitude, scale and diversity of outliers. Furthermore, each cycle runs in two phases:

- Phase 1 constructs a LOO model for each observation and uses single testing procedure to declare the outliers. At the end of Phase 1, we use the non-outlier data to construct an intermediate regression model.
- In Phase 2, we test the status of all current outliers based on the intermediate model using multiple testing procedures. If an outlier is not confirmed as such in Phase 2, it returns to the non-outlier sample. Finally we use the confirmed non-outliers to construct a final model.

The procedure stops either after a predefined number of cycles are performed (usually defined by the user), or until the procedure reaches a cycle that does not reject new outliers. Our proposed approach has two key features:

- Phase 1 of each cycle relies on a single testing approach, hence it is comparatively easy to purge the dubious observations so that to construct a reliable regression model of good quality (i.e. over reliable data points)
- Phase 2 of each cycle relies on multiple testing procedure, hence it is comparatively hard to keep a previously confirmed outlier out of the data sample, hence only those points that are indeed dubious will be left out

Our proposed methodology incorporates a simple way to reject dubious observations, and a robust way to keep away from further analysis only those observations that are indeed dubious and non-informative. At the same time, the methodology is flexible in that it adapts to different order of magnitude of outliers, while it also allows observations to leave and return the data set as the model develops. Finally, our procedure operates over fuzzy data. These properties are a highly desired trade-off between the quality and quantity of sample data that impacts positively the adequacy of the regression model.

We shall develop formalized algorithms that realize our methodology. The first algorithm uses least square (LS) method to solve the regression task with fixed outliers. The second algorithm uses the weighted LS (WLS) method for the regression task with varying outliers and employs the first algorithm in its steps. We shall use examples of various type to demonstrate the applicability of our approaches. Those examples help us demonstrate the strengths and benefits of our proposed algorithms for the proper purging of data and the creation of adequate regression models.

Our paper is organized as follows. In Section 2 we formalize the setup of linear regression analysis. This is then extended to the case with fuzzy samples in Section 3 as per our specific setup. Section 4 presents our first algorithm that uses the LS solution of the linear regression analysis problem with fixed outliers. It is later utilized within the second algorithm, developed in Section 5, which finds the WLS solution of the linear regression analysis problem with varying outliers. Section 6 presents three examples – a simple illustrative example and two examples from real-life thrombosis studies, where we apply our methodology. Some final discussions and conclusions are offered in Section 7.

2. CLASSICAL SETUP OF LINEAR REGRESSION ANALYSIS

In the linear regression analysis, we aim to predict the value y of the response (dependent) random variable (r.v.) Y given the observed values $z_1, z_2,\ldots, z_p$ of the directly measurable independent (explanatory, predictor) variables $Z_1, Z_2,\ldots, Z_p$. The linearity only concerns the q unknown parameters of the dependence, organized in a coefficient vector $\vec{\beta}=\left(\beta_1,\beta_2,\ldots,\beta_q\right)^T$, but not the values x_j of the q predictor variables $X_1, X_2,\ldots,X_q$ that could be arbitrary complex functions of $z_1, z_2,\ldots, z_p$:

$$y=\beta_1 x_1+\beta_2 x_2+\cdots+\beta_q x_q+u \quad , \tag{1}$$

Here, $x_j = F_j(z_1, z_2,\ldots, z_p)$, for j=1, 2, …, q, while u is an unobserved instance of the r.v. U, known as unexplained error. The classical linear regression assumption claims that U is normally distributed with mean zero and unknown standard deviation σ : $U \sim N\left(0,\sigma^2\right)$ [26].

Assume we have a sample of n observations (n>>q), where the i^{th} observation contains the measured value y_i^{mes} of Y dependent on the corresponding p observed values $z_{i,1}, z_{i,2},\ldots, z_{i,p}$ of the explanatory variables $Z_1, Z_2,\ldots, Z_p$. Those can be recalculated into the values of the q independent predictor variables $x_{i,j} = F_j(z_{i,1}, z_{i,2},\ldots, z_{i,p})$ for j=1, 2,…, q, organized in a vector $\vec{x}_i = \left(x_{i,1}, x_{i,2},\ldots,x_{i,q}\right)^T$. According to (1), the data in the i^{th} observation should satisfy (2):

$$y_i^{mes} = \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_q x_{i,q} + \hat{u}_i = \vec{x}_i^T \hat{\vec{\beta}} + \hat{u}_i = \hat{y}_i + \hat{u}_i \text{ , for } i = 1, 2, \ldots \tag{2}$$

Here, $\hat{\vec{\beta}} = \left(\hat{\beta}_{1,} \hat{\beta}_{2,},\ldots,\hat{\beta}_q\right)^T$ is the estimate of the vector of coefficients; the estimated value of the response variable is $\hat{y}_i = \vec{x}_i^T \hat{\vec{\beta}} = E\left(Y \middle| X_1 = x_{i,1} \vee X_2 = x_{i,2} \vee \cdots \vee X_q = x_{i,q}\right)$ and represents the expected value of Y conditioned on the values of the predictors. The quantity $\hat{u}_i$ is called residual and is an estimate of the unobserved realization u_i of U for the i^{th} observation. The quantities in (2) can be organized in the following data structures:

- n-dimensional vector of the measured response values $\vec{y}^{mes} = \left(y_1^{mes}, y_2^{mes},\ldots,y_n^{mes}\right)^T$;
- n-dimensional vector of the estimated response values $\hat{\vec{y}} = \left(\hat{y}_1, \hat{y}_2,\ldots,\hat{y}_n\right)^T$;
- n-dimensional vector of the residuals $\hat{\vec{u}} = \left(\hat{u}_1, \hat{u}_2,\ldots,\hat{u}_n\right)^T$;
- n x q dimensional matrix (a.k.a. *design matrix*) denoted with X, whose i-th row is $\vec{x}_i^T$.

Then we can represent (2) in a matrix form as follows:

$$\vec{y}^{mes} = X\hat{\vec{\beta}} + \vec{u} = \hat{\vec{y}} + \vec{u} \tag{3}$$

Classical linear regression has several other assumptions, as follows [45]:

- the variables $Z_1, Z_2,\ldots, Z_p$ are directly measured without errors;
- the predictor variables $X_1, X_2,\ldots, X_q$ are linearly independent, hence q is the rank of the matrix X;
- the residual u_i at any data point does not depend on the values of the unexplained error at all other data points.

A widely adopted method to identify the coefficients of the regression is the LS method (see [73]). It aims to identify the coefficient estimates $\hat{\vec{\beta}}^{LS,cur}$ so that to minimize the sum of squared differences between the measured and the estimated response values as per (4):

$$\hat{\vec{\beta}}^{LS} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\chi^2\left(\hat{\vec{\beta}}\right)\right\} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\sum_{i=1}^{n}\left(y_i^{mes} - \vec{x}_i^T\hat{\vec{\beta}}\right)^2\right\} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\sum_{i=1}^{n}\left(y_i^{mes} - \hat{y}_i\right)^2\right\} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\sum_{i=1}^{n}\hat{u}_i^2\right\} \tag{4}$$

3. LINEAR REGRESSION ANALYSIS PROBLEM OVER FUZZY DATA

Many works recognize the necessity to operate with fuzzy data samples in statistical (and regression) analysis, with varying interpretations of the degree of membership, caused by the measurement or interpretation process (see Gao and Gao in [30]). Practical reasons may cause fuzzy samples to emerge. Viertl in [70] discusses that sometimes the membership of a given observation to a given subpopulation is defined using a given classificator (probabilistic, metric, neural network, or subjective). Then the confidence in the correctness of the result of the classificator for a given observation may be interpreted as a degree of membership to the sample of the subpopulation. In medical analysis, a given parameter may be measured in t spatial points (e.g. measurements of different parts of a thrombus) for a given patient. Then the degree of membership to the sample for such measurements may be accepted as 1/t so that to provide equal weight for each patient. Other examples are also reported, such as the case described by Nikolova et al. in [36], where objects are assigned to several groups with a given degree of membership based on how similar they were to the descriptors of those groups, utilizing techniques of Denoeux [21] to interpret fuzzy data in experiments with uncertain outcomes.

As the variables in a regression analysis are assumed correlated, if there is a variable represented by a fuzzy sample entering the regression analysis, its degree of membership applies to the overall multi-dimensional observation of regression variables. Once fuzzy data is present, we need to utilize fuzzy regression analysis to explore dependencies between variables (see Section 1 for further discussion on fuzzy regression analysis). A valuable study on fuzzy regression analysis is by Chachi and Taheri in [8], who summarized three classes of approaches to fuzziness in regression analysis. The first class is the possibilistic class that stems from the works of Tanaka [65, 66], where the fuzzy regression problem is formulated as a mathematical programming problem. Tanaka's approaches and their sensitivity to outliers was discussed in [58], which lead to development of new algorithms of outlier detection for fuzzy regression models as for example the ones offered in [71]. The methods in this group investigate and develop ways to minimize the spread of the fuzzy parameters under certain constraints. The works [51, 52] developed fuzzy linear regression models linked to varying risk relations to minimize the difference between observed and estimates spreads of the output, relying on the concepts of necessity and possibility. The possibilistic regression modelling was reviewed in detail in [6]. The second class comprises least squares and least absolute methods, which aim to estimate the parameters of the model based on a distance on the space of fuzzy numbers. Some of the main works in this class are of D'Urso [14], who proposed regression models for crisp/fuzzy input-output data, and D'Urso et al. [16] that proposed robust fuzzy linear regression model using least median squares-WLS estimation procedure to deal with data that contains outliers. The work of Dehghan et al. [20] may also be attributed to this group as it uses LS and least absolute deviations methods to compare classical and fuzzy regressions using numerical examples of geographical data with symmetric fuzzy observations. Another notable work of Coppi et al. [13] proposed an iterative procedure with LS estimations for a regression mode to study the connection between crisp inputs and fuzzy output observations. D'Urso and Massari [17] explored the iterative WLS domain over a general linear regression model that includes a general class of fuzzy response variables on a set of crisp or LR_2 fuzzy explanatory variables. Bargiela et al. explored iterative techniques to study the coefficients of multiple regressions with fuzzy variables in [2]. DUrso et al. [19] also adopt an exploratory approach to find the best fit of a fuzzy linear regression using a new coefficient of determination and the Mallows index. For the case of imprecise responses, Ferrano et al. [28] proposed ways to construct linear regression models with accompanying hypothesis testing procedures. The third class is a heuristic class and it collects hybrid approaches to construct fuzzy regression models. Such are the works of Kao and Chyu [36, 37], who employed crisp coefficients and fuzzy error terms in a two-stage LS based procedure for fuzzy regression analysis. Lu and Wang [44] developed and improved fuzzy linear regression models that can avoid the spreads increasing problem, while Chachi et al. [7] demonstrated various practical applications of hybrid fuzzy regression models. The third class may also be expanded with the discussions on how to expand clusterwise regressions (see the works of Jajuga [34] and Yank and Ko [74] that introduced and developed the concept) in the fuzzy domain. The work [64] discussed this by combining fuzzy clustering and ridge regressions so that to deal with multicollinearity. The works [18, 25] developed a model for fuzzy linear regression utilizing fuzzy clusterwise linear regression that combined symmetrical crisp predictor and fuzzy response variables.

Another similar classification of fuzzy regression methods was also proposed by Jinn et al [35], who outlined a class of approaches utilizing Tanaka's ideas, and another class that combines the fuzzy least square approaches, pioneered by Diamond [22]. The review work of Chukhrova and Johannssen [10] relied on almost 500 sources to encapsulate thoroughly the various trends, approaches and applications of fuzzy regression analysis. The work identified the major and minor fields of fuzzy regression analysis, and then explored possibilistic approaches as well as fuzzy least square approaches. It also stressed the application of machine learning techniques in fuzzy regression analysis. The work also outlined several minor fields such as fuzzy probabilistic approaches, fuzzy clusterwise regressions, simulation techniques in fuzzy regressions, etc. One of the research questions in [10] was also to investigate the areas of reported practical implementations of fuzzy regression analysis. While engineering and environmental research prevail as implementation areas, the second largest group is that of business administration and economics. Case studies in that respect range from workforce forecasting [43] to project evaluations [33], analysis of macroeconomic parameters [58; 42], analysis of gross domenstic product [75], and stock price forecasting [38].

Other works discuss the impact of outliers in fuzzy regression models. As argued by Gao and Gao [30], sometimes outliers or influence points enter the data set for regression due to various unavoidable causes. This impedes the practical implications and adequacy of the fuzzy regression methods. Therefore, they investigated outlier detection procedures for fuzzy regressions using type-2 trapezoidal fuzzy numbers. Jinn et al. [35] elaborated on the importance of influential observations (that includes outliers) in regression analysis, highlighting the hidden flaws from the fitting process affecting the quality of the regression model. They further developed procedures for fuzzy regressions using the Cook distance, but only adapted those to the general fuzzy regression model, while in other cases their procedures only ended in a simple rejection of outliers with no subsequent analysis. Kwong et al. [41] also discuss the inherited fuzziness of experimental data and show an application of fuzzy regression in manufacturing processes, accounting for outlier rejection for the model accuracy through Peter's fuzzy regression. The works of Chan et al [9] and Gladysz and Kuchta [31] also demonstrate the necessity to reject outliers for fuzzy regression analysis in real-world applications from engineering and manufacturing. The proposal introduced by Nasrabadi et al. [54] discusses ways to apply linear programming and fuzzy least squares to the outlier detection in fuzzy regression analysis. The importance of outlier detection was also raised by Mashinchi et al. [48], where the authors developed two stage LS approach with no user defined variables. The first stage detects outliers, while the second stage uses the purged sample to fit a regression model with the model fitting measure minimized with a hybrid optimization technique. Yet, this two-stage procedure did not assume that some of the outliers may return to the data set as the model develops. All those works show that regression modelling over fuzzy data is highly applicable to many data analysis problems, and the proper rejection of outliers in those procedures are of paramount importance for the adequacy and validity of the constructed models in practice.

When fuzzy data is present, our setup is similar to what we presented in Section 2. Yet, the i^{th} observation $\left(\vec{x}_i;y_i^{mes}\right)$ of the sample now belongs to the Population with a degree of membership $\mu_i \in [0,1)$. We can organize the degrees of membership of all the observations in an n-dimensional membership vector $\vec{\mu} = \left(\mu_1, \mu_2, \ldots, \mu_n\right)^T$.

A suitable method to identify the coefficients of (3) is the WLS method (see [73]). It aims to derive the estimate $\hat{\vec{\beta}}^{WLS,cur}$ of the coefficients such that to minimize the weighted sum of square differences between the measured and the estimated response values, as follows:

$$\hat{\vec{\beta}}^{WLS} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\chi^2\left(\hat{\vec{\beta}}\right)\right\} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\sum_{i=1}^{n}\mu_i\left(y_i^{mes}-\vec{x}_i^T\hat{\vec{\beta}}\right)^2\right\} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\sum_{i=1}^{n}\mu_i\left(y_i^{mes}-\hat{y}_i\right)^2\right\} = \underset{\hat{\vec{\beta}}}{arg\,min}\left\{\sum_{i=1}^{n}\mu_i\hat{u}_i^2\right\} \tag{5}$$

Such a setup imposes several difficulties that we need to address, namely:

- how to identify and reject the outliers in the sample;
- how to find the following parameters based on the purged sample:
    - the estimates $\hat{\vec{\beta}} = \left(\hat{\beta}_{1,}\, \hat{\beta}_{2,},\ldots,\hat{\beta}_q\right)^T$ of the unknown coefficients, their confidence intervals and covariance matrix;
    - the estimate $\hat{\sigma}_u$ of the standard deviation of the unexplained error and its confidence interval;
    - the p_{value} of the hypotheses for nullity of $\hat{\beta}_j$ for j=1, 2,…, q and the p_{value} of the hypotheses for adequacy of the model;
    - the adjusted coefficient of multiple determination R_{adj}^2, which shows what part of the initial variance of Y is explained by the model, considering the count of determined parameters.

4. LEAST SQUARE SOLUTION OF THE LINEAR REGRESSION PROBLEM WITH KNOWN OUTLIERS OVER FUZZY DATA

Let us first construct a linear model, based on part of the observations (the non-outliers, the others being treated as outliers). If flag variable f_i for the i^{th} observation equals to 1, then that i^{th} observation $\left(\vec{x}_i;y_i^{mes}\right)$ is included in the model, whereas if f_i equals to 0 then that i^{th} observation $\left(\vec{x}_i;y_i^{mes}\right)$ is considered an outlier. The flag variables could be organized in an n-dimensional data flag vector $\vec{f} = \left(f_1, f_2, \ldots, f_n\right)^T$. All initially available observations can be denoted as the quadruplet $\left(\vec{y}^{mes};\, X;\, \vec{\mu};\, \vec{f}\right)$. In fact, the model is constructed on $n_{in}^{cur} = \sum_{i=1}^{n} f_i$ observations, which shall be referred to as *in-observations*. The absolute numbers of the in-observations can be organized in n_{in}^{cur}-dimensional vector $\vec{\delta}_a^{cur} = \left(\delta_{a,1}^{cur}, \delta_{a,2}^{cur}, \ldots, \delta_{a,n_{in}^{cur}}^{cur}\right)^T$. The count of the outliers is $n_{out}^{cur} = n - n_{in}^{cur}$ (which form the set of out-observations) and their absolute numbers can be organized in an n_{out}^{cur}-dimensional vector: $\vec{\delta}_b^{cur} = \left(\delta_{b,1}^{cur}, \delta_{b,2}^{cur}, \ldots, \delta_{b,n_{out}^{cur}}^{cur}\right)^T$. We can solve the optimization task (4) using singular value decomposition (SVD) [60]. The procedure is realized in Algorithm 1 below, based on the quadruplet $\left(\vec{y}^{mes};\, X;\, \vec{\mu};\, \vec{f}\right)$.

ALGORITHM 1: CONSTRUCTION OF THE LINEAR REGRESSION MODEL WITH KNOWN OUTLIERS

STEP A. Separate the initial observations in $\left(\vec{y}^{mes};\, X;\, \vec{\mu};\, \vec{f}\right)$ into outliers and in-observations (non-outliers) according to the values of $\vec{f}$ based on the following steps:

A1. Set: i=1, ia=1, and ib=1.

A2. If f_i =1 then set $\delta_{a,ia}^{cur}$=i, and ia=ia+1.

A3. If f_i = 0 then set $\delta_{b,ib}^{cur}$=i, and ib=ib+1.

A4. Set i=i+1.

A5. If $i \le n$ then go to **A2**.

A6. Set: $n_{in}^{cur} = ia - 1$, and $n_{out}^{cur} = ib - 1$.

A7. Set: $\vec{\delta}_a^{cur} = \left(\delta_{a,1}^{cur}, \delta_{a,2}^{cur}, \ldots, \delta_{a,n_{in}^{cur}}^{cur}\right)^T$, and $\vec{\delta}_b^{cur} = \left(\delta_{b,1}^{cur}, \delta_{b,2}^{cur}, \ldots, \delta_{b,n_{out}^{cur}}^{cur}\right)^T$.

A8. Define the current $n_{in}^{cur} \times q$ dimensional design matrix X^{cur}, whose rows are all $\sqrt{\mu_i}\,\vec{x}_i^T$ for which f_i=1: $\sqrt{\mu_i}\,\vec{x}_i^{cur} = \sqrt{\mu_{\delta_{a,i}^{cur}}}\,x_{\delta_{a,i}^{cur}}$, for i=1, 2, …, n_{in}^{cur}.

A9. Define the current n_{in}^{cur}-dimensional vector of response values $\vec{y}^{mes,cur}$, whose elements are all $\sqrt{\mu_i}\,y_i^{mes}$, for which f_i=1: $y_i^{mes,cur} = \sqrt{\mu_{\delta_{a,i}^{cur}}}\,y_{\delta_{a,i}^{cur}}^{mes}$, for i=1, 2, …, n_{in}^{cur}.

A10. Define the current $n_{out}^{cur} \times q$ dimensional outlier matrix X_{out}^{cur}, whose rows are all $\sqrt{\mu_i}\,\vec{x}_i^T$ for which f_i=0: $\sqrt{\mu_i}\,\vec{x}_{out,i}^{cur} = \sqrt{\mu_{\delta_{b,i}^{cur}}}\,\vec{x}_{\delta_{b,i}^{cur}}$, for i=1, 2, …, n_{out}^{cur}.

A11. Define the current n_{out}^{cur}-dimensional outlier vector of response values $\vec{y}^{mes,cur}$, whose elements are all $\sqrt{\mu_i}\,y_i^{mes}$, for which f_i=1: $y_{out,i}^{mes,cur} = \sqrt{\mu_{\delta_{b,i}^{cur}}}\,y_{\delta_{b,i}^{cur}}^{mes}$, for i=1, 2, …, n_{out}^{cur}.

Step B. The X^{cur} could be factored to the product of three matrices using the SVD decomposition:

$$X^{cur} = W \times S \times V^T \tag{6}$$

Here, W is an $n_{in}^{cur} \times q$ dimensional column-orthonormal matrix with columns $\vec{w}_j$ (for j=1, 2, …, q); S is an q x q dimensional diagonal matrix with non-negative elements s_j (for j=1, 2, …, q) on the main diagonal, called *singular values*; V is a q x q dimensional orthonormal matrix with columns $\vec{v}_j$ (for j=1, 2, …, q).

Step C. The SVD decomposition (6) is usually executed by a computer program and is subject to round-off errors. We need to define which singular values are in fact small positive real values and which are in fact zeros (but estimated as small positive real values due to round-off errors). Therefore we correct the singular values s_j to s_j^{cor} (as per [57, 67]) in four steps:

C1. If s_j is non-positive, then $s_j^{cor} = 0$.

C2. If s_j is positive, then we compute an estimate of the unit vector $\vec{w}_j$ as $\vec{w}_j^{nonzero} = X^{cur}\vec{v}_j / s_j$.

C3. If the angle between $\vec{w}_j$ and $\vec{w}_j^{nonzero}$ is less or equal to 1 *deg*, and the length of $\vec{w}_j^{nonzero}$ is within [0.99; 1.01], then $s_j^{cor} = s_j$.

C4. If the angle between $\vec{w}_j$ and $\vec{w}_j^{nonzero}$ is greater than 1 *deg* or if the length of $\vec{w}_j^{nonzero}$ sits outside of the interval [0.99; 1.01], then $s_j^{cor} = 0$.

Step D. According to [60] we can solve (4) using the current coefficient vector $\hat{\vec{\beta}}^{WLS,cur} = \sum_{\substack{i=1 \\ s_j^{cor}>0}}^{q} \left(\frac{\vec{w}_j^T\,\vec{y}^{mes,cur}}{s_j^{cor}}\right)\vec{v}_j$

Step E. We can calculate consecutively the following parameters that relate to outlier rejection:

E1. The current vector of estimated response values $\hat{\vec{y}}^{cur} = X^{cur}\hat{\vec{\beta}}^{WLS,cur}$.

E2. The current vector of the WLS residuals $\hat{\vec{u}}^{cur} = \vec{y}^{mes,cur} - \hat{\vec{y}}^{cur}$.

E3. The current residual sum of squares: $RSS^{cur} = \sum_{i=1}^{n_{in}}\left(\hat{u}_i^{cur}\right)^2$.

E4. The current estimate of the standard deviation of the unexplained error: $\hat{\sigma}_u^{cur} = \sqrt{RSS^{cur}\Big/\left(n_{in}^{cur} - q\right)}$.

E5. The current $q \times q$ dimensional covariance matrix of the parameters:

$$K^{cur} = \left(\hat{\sigma}_u^{cur}\right)^2 \sum_{\substack{j=1 \\ s_j^{cor}>0}}^{q} \vec{v}_i\vec{v}_i^T \,/\, s_j^{cor} = \left[k_{i,j}^{cur}\right]_{1\le i\le q, 1\le j\le q} .$$

E6. The current standard errors of the parameters: $\hat{\sigma}_{\beta_j}^{cur} = \sqrt{k_{j,j}^{cur}}$ for j=1, 2, …, q.

E7. The current (1-α)-confidence intervals for each of the parameters $\beta_j \in \left(\hat{\beta}_j^{LS,cur} - t_{1-\frac{\alpha}{2};n_{in}^{cur}-q}\hat{\sigma}_{\beta_j}^{cur}, \hat{\beta}_j^{LS,cur} + t_{1-\frac{\alpha}{2};n_{in}^{cur}-q}\hat{\sigma}_{\beta_j}^{cur} \right)$ for j=1, 2, …, q. Here, $t_{1-\frac{\alpha}{2};n_{in}^{cur}-q}$ is the $\left(1-\alpha/2\right)$ quantile of the t-distribution with $n_{in}^{cur}-q$ degrees of freedom [32].

E8. The current p-values of the hypotheses for nullity of $\hat{\beta}_j$: $p_{value,\beta_j}^{cur} = 2 \times CDF_{t,n_{in}^{cur}-q}\left(-\left|\beta_j^{cur} / \hat{\sigma}_{\beta_j}^{cur}\right|\right)$, for j=1, 2, …, q. Here, $CDF_{t,n_{in}^{cur}-q}(.)$ is the cumulative distribution function of the t-distribution with $n_{in}^{cur}-q$ degrees of freedom (see [32]).

E9. The current (1-α)-confidence intervals of the standard deviation of the unexplained error: $\sigma_u^{cur} \in \left(\frac{\sqrt{n_{in}^{cur}-q}\hat{\sigma}_u^{cur}}{\sqrt{\chi_{1-\frac{\alpha}{2},n_{in}^{cur}-q}}}, \frac{\sqrt{n_{in}^{cur}-q}\hat{\sigma}_u^{cur}}{\sqrt{\chi_{\frac{\alpha}{2},n_{in}^{cur}-q}}} \right)$. Here, $\chi_{\gamma,n_{in}^{cur}-q}$ is the γ-quantile of the χ^2-distribution with $n_{in}^{cur}-q$ degrees of freedom (see [32]).

E10. The current total sum of squares: $TSS^{cur} = \sum_{i=1}^{n_{in}^{cur}} \left(y_i^{mes,cur} - \sum_{i=1}^{n_{in}^{cur}} y_i^{mes,cur} / n_{in}^{cur} \right)^2$.

E11. The current adjusted coefficient of multiple determination: $R_{adj}^{2,cur} = 1 - \frac{\left(n_{in}^{cur}-1\right)RSS}{\left(n_{in}^{cur}-q\right)TSS}$.

E12. The current ANOVA test p_{value} for overall adequacy of the model: $p_{value,ad}^{cur} = 1 - CDF_{F,q-1,n_{in}^{cur}-q}\left(\frac{\left(n_{in}^{cur}-q\right)\left(TSS-RSS\right)}{\left(q-1\right)RSS} \right)$. Here, $CDF_{F,q-1,n_{in}^{cur}-q}(.)$ is the cumulative distribution function of the F-distribution with q-1 and $n_{in}^{cur}-q$ degrees of freedom (see [32]).

STEP F. To test if any of the in-observations is an outlier, we need to calculate the externally Studentized residuals. The p_{value} for the hypothesis H_0 that a particular observation is not an outlier can be calculated since given H_0 its externally Studentized residual follows a t-distribution with $n_{in}^{cur}-q-1$ degrees of freedom. The required p_{value} may be found using ideas from [53] by calculating consecutively:

F1. The current $n_{in}^{cur} \times n_{in}^{cur}$ dimensional hat matrix of the parameters:
$H^{cur} = X^{cur}K^{cur}\left(X^{cur}\right)^T / \left(\hat{\sigma}_u^{cur}\right)^2 = \left[h_{i,j}^{cur}\right]_{1 \le i \le n_{in}^{cur}, 1 \le j \le n_{in}^{cur}}$.

F2. The current predicted residuals: $\hat{r}_i^{pred,cur} = \hat{u}_i^{cur} / \left(1-h_{i,i}^{cur}\right)$, for i=1, 2, …, n_{in}^{cur}.

F3. The current estimate of the standard deviation of the predicted residuals:
$\hat{s}_i^{pred,cur} = \sqrt{\left(RSS^{cur} - \frac{\left(\hat{u}_i^{cur}\right)^2}{\left(1-h_{i,i}^{cur}\right)} \right)} / \left(n_{in}^{cur}-q-1\right) / \sqrt{\left(1-h_{i,i}^{cur}\right)}$, for i=1, 2, …, n_{in}^{cur}.

F4. The externally Studentized residuals: $\hat{t}_i^{pred,cur} = \hat{r}_i^{pred,cur} / \hat{s}_i^{pred,cur}$, for i=1, 2, …, n_{in}^{cur}.

F5. The current p_{value} of the hypotheses that the observations $\left(\vec{x}_i^{cur}; y_i^{mes,cur}\right)$ are not outliers: $p_{value,in,i}^{cur} = 2 \times CDF_{t,n_{in}^{cur}-q-1}\left(-\left|\hat{t}_i^{pred,cur}\right|\right)$, for i=1, 2, …, n_{in}^{cur}. Here, $CDF_{t,n_{in}^{cur}-q-1}(.)$ is the cumulative distribution function of the t-distribution with $n_{in}^{cur}-q-1$ degrees of freedom (see [53]).

STEP G. To confirm that the outliers (i.e. the data points, where f_i equals to 0) are indeed outliers, we need to calculate their Studentized residuals. We can calculate the p_{value} for H_0: “A particular observation initially declared outlier is not an outlier”, since, given H_0 its Studentized residual follows a t-distribution with $n_{in}^{cur}-q$ degrees of freedom. The count of observations declared outliers is $n_{out}^{cur} = n - n_{in}^{cur}$. If n_{out}^{cur}>0, we can find the p_{value} by calculating consecutively [53]:

G1. The current estimated predicted response values: $\hat{y}_{out,i}^{cur} = \vec{x}_{out,i}^T \hat{\vec{\beta}}^{LS,cur}$, for i=1, 2, …, n_{out}^{cur}.

G2. The current estimate of the standard deviation of the predicted response values: $\hat{s}_{out,i}^{pred,cur} = \sqrt{\left(\hat{\sigma}_u^{cur}\right)^2 + \vec{x}_{out,i}^T K^{cur} \vec{x}_{out,i}}$, for i=1, 2, …, n_{out}^{cur} .

G3. The current p_{value} for the hypothesis H_0 that observations $\left(\vec{x}_{out,i}^{cur};y_{out,i}^{mes,cur}\right)$ are not outliers: $p_{value,out,i}^{cur} = 2\times CDF_{t,n_{in}^{cur}-q}\left(-\left|\frac{\hat{y}_{out,i}^{cur} - y_{out,i}^{mes,cur}}{\hat{s}_{out,i}^{pred,cur}}\right|\right)$, for i=1, 2, …, n_{out}^{cur} . Here, $CDF_{t,n_{in}^{cur}-q}(.)$ is the cumulative distribution function of the t-distribution with $n_{in}^{cur}-q$ degrees of freedom (see [32]).

Algorithm 1 allows to reject outliers from the initial sample (in step F), but also in step G it tests to confirm the outlier status of those data points identified as such in Step F. This makes Algorithm 1 flexible and allows it to contribute to the proper quality and quantity of data for fuzzy regression analysis. The model is constructed only on the predefined set of in-observations (i.e. non-outliers). Importantly, Algorithm 1 gives as an output the p_{value} of a hypothesis test for outliers of each measurement in the dataset. We use different approaches to calculate the p_{value} for the in-observations (i.e. the data points, where f_i=1) and for the outliers (i.e. the data points, where f_i=0). This algorithm will later become a working engine in a large procedure that works on varying outliers (see Section 5).

The rationale of steps C1-C4 comes from the basic property of SVD decomposition $X^{cur}\vec{v}_j = s_j\ \vec{w}_j$ (see [67]). If the singular value is positive, then both sides of the equation could be divided to s_j, and $\vec{w}_j^{nonzero}$ would be almost a perfect estimate of $\vec{w}_j$. If the singular value is zero, then the basic property of the SVD decomposition degenerates to equality of two null vectors ($X^{cur}\vec{v}_j = \vec{0} = 0\ \vec{w}_j$). Then obviously the $\vec{w}_j^{nonzero}$ would be an estimate of a null vector divided by zero, hence it should be quite different from $\vec{w}_j$.

5. LEAST SQUARE SOLUTION OF THE LINEAR REGRESSION PROBLEM WITH UNKNOWN OUTLIERS OVER FUZZY DATA

In Section 4 we offered a procedure that constructed a regression model by finding the coefficients (5) using the WLS method. However, this algorithm did not change the status of the data points, i.e. the initial separation to in-observations and out-observations (outliers) remained unchanged. In this section, we shall develop another approach that relies and uses Algorithm 1, yet its task is to assess the outliers and to find the characteristics of the model for different samples iteratively in several cycles. The main reason to use cycles is that as a rule the outliers appear in the initial sample due to some sort of errors in the measurement process (e.g. equipment failure, human error, etc.). As those errors are of miscellaneous nature, the outliers may have varying order of magnitude. Those outliers with higher order can then hide (mask) the ones of smaller order making the latter look as legitimate in-observations in the outliers detection procedure. Therefore, in each cycle we remove only the outliers with the highest order of deviation from the data sample. Initial results from this approach for the case of crisp data samples is reported by Tenekedjiev and Radoinova in [68].

Assume that the maximum count of cycles permitted is C_{max}. The first cycle would start by declaring all data points as in-observations. Alternatively, we can think that the 0th cycle has identified no outliers in the data set. The cth cycle would then start by constructing a model using the in-observations after the (c–1)th cycle. We will calculate the adjusted coefficient of multiple determination for this model. All the outliers can be purged with independent statistical tests for each of the in-observations. The purged observations are added to those declared outliers from the (c–1)th cycle. Then an intermediate model is constructed, based on the current in-observations. At the end of the cth cycle, outliers are only those observations, where the null hypothesis was rejected. All the other observations are added back to the in-observations after the cth cycle.

We can perform the Benjamini-Hochberg step-up multiple testing procedure controlling the false discovery rate (FDR) for independent test statistics to confirm the status of each observation declared outlier (see [4, 5] for discussion of the method). The FDR should be less or equal to a predetermined FDR_{max} (see also [3] for discussion on FDR). Unlike the p-value in the single testing procedure, which measures the number of false rejections out of all cases where H_0 was true, the FDR measures the number of false rejections out of all rejections of H_0. The procedures that utilize FDR deal with the expected proportion of false discoveries and were elaborated as an alternative to extremely demanding controlling procedures in multiple testing, such as the Bonferroni correction (for discussion on the Bonferroni approach see [49, 50, 61]). The following considerations justify the deployment of FDR-based multiple testing procedure:

a) The outliers rarely exceed 20% of the data points;

b) If a small amount (e.g 10%) of the declared outliers are in fact legitimate, then the in-observation set will change insignificantly (e.g. from 80% to 78%, i.e. 80% – 10% × 20%=78%);

c) The derived regression model will not be drastically affected although losing any legitimate measurement is not too desired;

d) The application of a Bonferroni-like method implies that a substantial percentage of the true outliers (at least 50%) will not be identified;

e) The resulting regression model will use almost all legitimate in-observation, but also a substantial amount of outliers (e.g. 10%, which is 50% of the 20%);

f) If the regression model is constructed over a data set with substantial amount of outliers, then its quality would be substandard and will eventually make it useless.

In our approach, the c^{th} cycle will be considered incomplete if either the count of observations in the intermediate model is less or equal to the count of regressors, or if after the cycle the outliers were the same as those from any of the previous cycles. If the c^{th} cycle was unsuccessful, then the true count of cycles $C_{true}=c-1$. If the C_{max}-th cycle was successful, then a model based on the in-observations is constructed and the adjusted coefficient of multiple determination is calculated. Then $C_{true}=C_{max}$. Each cycle gives different set of outliers. The cycle that maximized the adjusted coefficient of multiple determination is considered to give the correct answer for the outliers. The resulting model along with its group of outliers should be adequate and with significant regression coefficients. If any of the coefficients was not significant, then the regressor corresponding to the coefficient with maximal p_{value} is deleted from the model and we repeat the whole procedure again. All these ideas are incorporated into Algorithm 2 below, which runs in three main steps.

ALGORITHM 2: CONSTRUCTION OF THE LINEAR REGRESSION MODEL WITH UNKNOWN OUTLIERS

STEP A. Initiation of the first cycle

A1. Define the data flag vector before the first cycle $\vec{f}^{\,0}=(1,1,\ldots,1)^T$ that consists of n values of 1.

A2. Initialize the set of data flag vectors $F=\{\vec{f}^{\,0}\}$.

A3. Initialize the set of coefficients of multiple determination $A=\{\ \}$.

A4. Define the initial cycle to be executed: $c=1$.

STEP B. Execution of the c^{th} cycle

B1. Define the count of in-observations from the previous cycle: $n_{in}^{cur}=\sum_{i=1}^{n} f_i^{c-1}$.

B2. Construct the linear model with data $\left(\vec{y}^{mes};X;\vec{\mu};\vec{f}^{\,c-1}\right)$ using *Algorithm 1.*

B3. Find the adjusted coefficient of multiple determination: $R^2_{adj,c-1}=R^{2,cur}_{adj}$.

B4. Update the set of coefficients of multiple determination: $A=A\cup\left\{R^2_{adj,c-1}\right\}$.

B5. Initialize $\vec{f}^{\,c}=\vec{f}^{\,c-1}$.

B6. Perform n_{in}^{cur} independent statistical tests, with H_0: "Each of the in-observations is not an outlier" at a significance level α : if $p^{cur}_{value,in,i}\le\alpha$ then set $f^{c}_{\delta^{cur}_{a,i}}=0$, for i=1, 2, …, n_{in}^{cur}.

B7. Count the in-observations in the intermediate model of the c^{th} cycle: $n_{in}^{c,int}=\sum_{i=1}^{n} f_i^{c}$.

B8. Declare the cycle unsuccessful if there is not enough in-observations: if $n_{in}^{c,int}<=q$, set $C_{true}=c-1$, go to **C1**.

B9. Count the outliers in the intermediate model of the c^{th} cycle: $n_{out}^{c,int}=n-n_{in}^{c,int}$.

B10. Finish the c^{th} cycle if there are no outliers: if $n_{out}^{c,int}=0$ then go to **B16**.

B11. Construct the linear model with data $\left(\vec{y}^{mes};X;\vec{\mu};\vec{f}^{\,c}\right)$ using *Algorithm 1.*

B12. Sort the current p_{value} values $p^{cur}_{value,out,i}$ (for i=1, 2, …, n_{out}^{cur}) for the H_0 that observations $\left(\vec{x}^{cur}_{out,i};y^{mes,cur}_{out,i}\right)$ are not outliers. Let [$s_{out}(1)$, $s_{out}(2)$, …, s_{out}(n_{out}^{cur})] be a permutation of the set {1,2,…, n_{out}^{cur} } such that:

$$p^{cur}_{value,out,s_{out}(1)}\le p^{cur}_{value,out,s_{out}(2)}\le\ldots\le p^{cur}_{value,out,s_{out}\left(n_{out}^{cur}\right)}.$$

B13. Perform the Benjamini-Hochberg step-up multiple testing procedure controlling FDR for independent test statistics, to test the n_{out}^{cur} hypothesis that each observation declared outlier is in fact not an outlier at a maximum false discovery rate FDR_{max}: find the maximum $i_{\max}=i$ where $p^{cur}_{value,out,s_{out}(i)}\le\frac{i}{n_{out}^{cur}}FDR_{max}$, for i=1, 2, …, n_{out}^{cur}. If such i does not exist, then set $i_{max}=0$.

B14. Declare as an in-observation any outlier for which H_0 ("the observation is not in fact an outlier") is not rejected: set $f^{c}_{\delta^{cur}_{b,s_{out}(i)}}=1$, for $i=i_{max}+1, i_{max}+2,\dots, n^{cur}_{out}$.

B15. Declare the cycle unsuccessful if the estimated data flag vector after the c^{th} cycle coincides with the estimated data flag vector after any former cycle: if $\vec{f}^{c-k}=\vec{f}^{c}$ for any k=1,2,…, c then set $C_{true}=c-1$ and go to step **C1**.

B16. Update the set of data flag vectors $F=F\cup\left\{\vec{f}^{c}\right\}$.

B17. If $c<C_{max}$ then set $c=c+1$ and go to step **B1**.

B18. Construct the linear model with quadruplets $\left(\vec{y}^{mes};X;\vec{\mu};\vec{f}^{C_{max}}\right)$ using *Algorithm 1.*

B19. Find the adjusted coefficient of multiple determination: $R^2_{adj,C_{max}}=R^{2,cur}_{adj}$.

B20. Update the set of coefficients of multiple determination: $A=A\cup\left\{R^2_{adj,C_{max}}\right\}$.

B21. Set: $C_{true}=C_{max}$.

STEP C. Selection and construction of the model

C1. Find the cycle number with the maximum adjusted coefficient of multiple determination: find c_{result} such that $R^2_{adj,c_{results}}\geq R^{2,cur}_{adj,c}$, for c=1, 2, …, C_{true}.

C2. Find the "best" data flag vector $\vec{f}^{best}=\vec{f}^{c_{results}}$.

C3. Construct the "best" linear model with quadruplet $\left(\vec{y}^{mes};X;\vec{\mu};\vec{f}^{best}\right)$ using *Algorithm 1.*

C4. Define the following parameters:

a) the count of outliers is $n_{out}=n^{cur}_{out}$ with absolute numbers in $\vec{\delta}_b=\vec{\delta}^{cur}_b$;

b) the count of in-observations is $n_{in}=n^{cur}_{in}$ with absolute numbers in $\vec{\delta}_a=\vec{\delta}^{cur}_a$;

c) the estimate of the unknown coefficients is $\hat{\vec{\beta}}^{LS}=\hat{\vec{\beta}}^{LS,cur}$;

d) the confidence interval for the j^{th} coefficient (for j=1, 2, …, q) is

$$\beta_j\in\left(\hat{\beta}^{LS}_j-t_{1-\frac{\alpha}{2};n_{in}-q}\hat{\sigma}_{\beta_j},\hat{\beta}^{LS}_j+t_{1-\frac{\alpha}{2};n_{in}-q}\hat{\sigma}_{\beta_j}\right)$$, where $\hat{\sigma}_{\beta_j}=\hat{\sigma}^{cur}_{\beta_j}$, for j=1, 2, …, q;

e) the covariance matrix of the unknown coefficients is $K=K^{cur}$;

f) the estimate of the standard deviation of the unexplained error is $\hat{\sigma}_u=\hat{\sigma}^{cur}_u$;

g) the confidence interval of the standard deviation of the unexplained error is:

$$\sigma_u\in\left(\frac{\sqrt{n_{in}-q}\hat{\sigma}_u}{\sqrt{\chi_{1-\frac{\alpha}{2},n_{in}-q}}},\frac{\sqrt{n_{in}-q}\hat{\sigma}_u}{\sqrt{\chi_{\frac{\alpha}{2},n_{in}-q}}}\right);$$

h) the p_{value} of the hypotheses for nullity of $\hat{\beta}_j$ (for j=1, 2, …, q) is $p_{value,\beta_j}=p^{cur}_{value,\beta_j}$;

i) the p_{value} of the hypothesis for adequacy of the model is $p_{value,ad}=p^{cur}_{value,ad}$;

j) the adjusted coefficient of multiple determination is $R^2_{adj}=R^{2,cur}_{adj}$.

The Benjamini-Hochberg step-up multiple testing procedure is implemented in steps B12, B13 and B14 of Algorithm 2. In steps B and C, Algorithm 2 uses Algorithm 1 as a working engine. As a result, Algorithm 2 runs only in 3 steps and is simpler (unlike Algorithm 1 that runs in seven steps and multiple sub-procedures). It realizes the construction of the fuzzy regression model in several cycles, thus accounting for the order of magnitude of the outliers. The number of cycles may be predefined by the user. Alternatively, the procedure will end once there were no new outliers detected. In such way, sometimes the cycle will finish before the maximum allowed number of cycles are executed.

6. NUMERICAL EXAMPLES

In this section, we will demonstrate how we apply Algorithms 1 and 2 in three examples. We shall start from an illustrative numerical example in Section 6.1. Afterwards, in Section 6.2 we shall offer two other examples related to data analysis in thrombosis case studies.

6.1. Illustrative numerical example

Table 1 presents the values of the predictor variable X, response variable Y and the degrees of membership of 12 records.

INITIAL ANALYSIS

Initially, without any outlier rejection, the model was constructed using the 12 in-observations as $y= -0.1823x+20.74+e$. In the terminology of Section 2, we have the response variable Y, while the directly measured independent variable Z_1 is X (obviously p=1). The q=2 predictor variables are: X_1=1, X_2= X (that is, $F_1(z_1)$=1, $F_2(z_1)$= z_1). The 95%-confidence intervals of the model's coefficients are from –3.093 to 2.728 and from –7.447 to 48.92. The estimated standard error is in a 95%-confidence interval from 13.43 to 33.74. The adjusted coefficient of multiple determination is calculated to be $R^2_{adj} = -0.09568$. The regression parameters are not significant (all tests with p_{value}>>0.05) and the model is not adequate (ANOVA with p_{value}=0.8467). The results of the initial analysis are presented in Fig. 1.

The following results were achieved using Algorithms 1 and 2 with maximal number of cycles C_{max}= 6. Table 1 presents the results from the outlier detection through the three sub-columns of the parameter "Outliers after each cycle".

CYCLE 1

A total of 2 outliers were identified and rejected from the original sample (rows 3 and 7 of Table 1). This can be traced in the first sub-column of columns 2 and 7 in Table 1 (shaded rows), where those two observations have a value of 1 (indicating they were identified as outliers in cycle 1). The model was constructed using the remaining 10 in-observations as $y= -1.371x+23.65 +e$. The 95%-confidence intervals of the model's coefficients are from –3.649 to 0.9064 and from 4.310 to 43.00. The estimated standard error is in a 95%-confidence interval from 6.597 to 18.71. The adjusted coefficient of multiple determination is R^2_{adj} = 0.09370. The regression parameters are with varying significance (first parameter not significant with p_{value}=0.2025, second parameter significant with p_{value}=0.02250). The model is not adequate (ANOVA with p_{value}=0.2022). The results of the analysis after cycle 1 are presented in Fig. 2.

CYCLE 2

A new outlier was identified and rejected from the original sample (row 10 of Table 1). This can be traced in the second sub-column of columns 2 and 7 in Table 1 (shaded rows), where this observation has a value of 2 (indicating it was identified as an outlier in cycle 2). The previous two outliers were confirmed again as outliers (their second sub-column in columns 2 and 7 has the value of 1 indicating their status from cycle 1 was confirmed in cycle 2). The model was constructed using the remaining 9 in-observations as $y= -1.37x+26.30+e$. The 95% confidence intervals of the model's coefficients are from –2.526 to –0.2145 and from 16.40 to 36.19. The estimated standard error is in a 95% confidence interval from 3.204 to 9.863. The adjusted coefficient of multiple determination is R^2_{adj} =0.4616. The regression parameters are statistically significant (all tests with p_{value}<0.05). The model is adequate (ANOVA with p_{value}=0.0264). The results of the analysis after cycle 2 are presented in Fig. 3.

CYCLE 3

A new outlier was identified and rejected from the original sample (row 12 of Table 1). This can be traced in the third sub-column of columns 2 and 7 in Table 1 (shaded rows), where this observation has a value of 3 (indicating it was identified as an outlier in cycle 3). Two previous outliers were confirmed (rows 3 and row 10 of Table 1). Their third sub-column of columns 2 and 7 has the value of either 1 or 2 indicating their status from cycle 1 or 2 was confirmed in cycle 3. One observation returned to the original sample (row 7 of Table 1, from cycle 1). Its third sub-column of "Outliers after each cycle" has the value of 0 indicating the outlier status was rejected in cycle 3. The model was constructed using the newly identified 9 in-observations as y=1.958x+5.128+e. The 95% confidence intervals of the model's coefficients are from 1.825 to 2.09 and from 3.917 to 6.34. The estimated standard error is in a 95% confidence interval from 0.4439 to 1.367. The adjusted coefficient of multiple determination is R^2_{adj} =0.9937. The regression parameters are statistically significant (all tests with p_{value}<<0.05). The model is adequate (ANOVA with p_{value}<<0.0005). The results of the analysis after cycle 3 are presented in Fig. 4.

CYCLE 4

No new outliers were identified so the 4th cycle is considered incomplete. The true number of cycles is C_{true}=3 and there is no need for cycles 5 and 6 either. The algorithm exits the cycle loop.

"BEST" LINEAR MODEL SELECTION

As expected, the "best" linear model is the one after cycle 3 because it has the largest adjusted coefficient of multiple determination $R^2_{adj} = 0.9937$ (larger than the second cycle $R^2_{adj} = 0.4616$, and the first cycle $R^2_{adj} = 0.09370$).

The example results show that throughout the cycles we observed new outliers being identified, and some of the initially identified ones returning to the original sample as the model developed. The results from the initial analysis were unacceptable overall. We can compare cycle 1 (which is the result of procedure for outlier rejection with no cycles) and cycle 3 (which is the result of the multi-cycle procedures proposed earlier in the paper) and see that the significance of the regression parameters improved substantially, and the adequacy of the model became very high, with R^2_{adj} going from 0.0937 up to 0.9937, which is more than a ten-fold increase. This demonstrates the importance of using cycles in outlier rejection.

6.2. Applications to Real Life Medical Data

Thrombosis is a major cause of death, killing 1 person every 6 seconds world-wide. The evolution of thrombi in the arteries causes tissue damage. For example, thrombotic occlusion of the coronary arteries causes acute myocardial infarction. Fibrin formed from the blood plasma protein fibrinogen is a long-known structure founding the scaffold of thrombi [1] and the presence of cells (platelets, white blood cells – primarily neutrophils) significantly impacts the course of the disease (e.g. affecting the size of myocardial infarction [47]). Medical research needs to investigate if the cellular composition of thrombi could be predicted from easily accessible biomarkers circulating in the blood. We shall apply our algorithms to this medical prediction task.

In a recent work [27] we analyzed the connection between routinely available clinical data and structural characteristics of thrombi (with other similar analyses also reported in Kovacs et al. [40]). Particularly, we tried to find the statistical connection between the relative surface occupied by platelets in the thrombi (*sPlt*, in %) and the fibrinogen blood levels (*Fibrinogen*, in g/l). We also tried to predict the relative surface occupied by white blood cells in the thrombi (*sWBC*, in %) based on the white blood cell count (*WBC*, in 1000/μl) in the blood and the C-reactive protein (*CRP*) as one of the soluble inflammatory markers. However, due to the high degree of heterogeneity of the thrombus structure (for discussion see [69; 72]), the *sPlt* and the *sWBC* were measured on up to 6 regions of the retrieved thrombi with scanning electron microscope, whereas the other parameters (*Fibrinogen*, *WBC*, *CRP*) are characteristics of the patient. If for a patient, we have *m* regions measured for the thrombi, we have in fact created *m* records for each patient, where the patient-specific parameters are the same.

If we assume that we deal with a crisp sample of records, then the patient-specific parameters will be heavily distorted because the patients with more regions measured on the thrombi will participate with higher weight into the estimated sample characteristics. In other words, we have a practical need to fuzzify the sample to avoid biased analysis. Therefore we apply a fuzzy sample approach and assume that the *m* records for a given patient belong to the general population with degree of membership 1/*m*. For example, the 5 records for patient 07 (see Table 2) obtain a degree of membership of 1/5=0.20, while the 4 records for patient 06 obtain a degree of membership of 1/4=0.25. If other considerations are present, each record may receive different degrees of membership, so this is only one possible approach, chosen here for simplicity.

6.2.1. Practical example 1

We shall first study the dependence of *sPlt* from *Fibrinogen*. Table 2 presents the measurement data, with a total of 59 records for a total of 13 patients (patient codes given in columns 1, 6, 11, and 16). Patients were selected so that for all records, *Fibrinogen*>4.2 g/l. Columns 5, 10, 15 and 20 of Table 2 show the degrees of membership of each patient record.

INITIAL ANALYSIS

Initially, without any outlier rejection, we construct a quadratic regression model for *sPlt* on the predictor *Fibrinogen*, using the 59 in-observations as follows: *sPlt*=218.9–76.5×*Fibrinogen*+6.885×*Fibrinogen*^{2}+*e*. In the terminology of Section 2, the response variable *Y* is *sPlt* and the directly measured independent variable Z_1 is *Fibrinogen* (evidently, p=1). The q=3 predictor variables are: X_1=1, X_2= Z_1 and X_3= Z_1^2 (that is, $F_1(z_1)$=1, $F_2(z_1)$= z_1, and $F_3(z_1)$= z_1^2). The 95%-confidence intervals of the model's coefficients are from –81.98 to 519.8, from –192.6 to 39.6, and from –4.13 to 17.9. The estimated standard error is in a 95%-confidence interval from 17.46 to 25.37. The adjusted coefficient of multiple determination is $R^2_{adj} = 0.02073$. The regression parameters are not significant (all tests with p_{value}>>0.05) and the model is not adequate (ANOVA with p_{value}=0.2082).

The following results were achieved using Algorithms 1 and 2 with maximal number of cycles C_{max}=3. Table 2 also presents the results from the outlier detection through the three sub-columns of "Outliers after each cycle" (columns 2, 7, 12 and 17).

CYCLE 1

A total of 5 outliers were identified and rejected from the original sample: observation 1 for patient 05, observation 2 for patient 34, observations 4 and 5 for patient 35, and observation 5 for patient 56. This can be traced in the first sub-column of "Outliers after each cycle" in Table 2 (rows shaded), where those five observations have a value of 1 (indicating they were identified as outliers in cycle 1). The model was constructed using the remaining 54 in-observations as: *sPlt*=165.7 – 60.37×*Fibrinogen*+5.539×*Fibrinogen*^{2}+*e*. The 95%-confidence intervals of the model's coefficients are from 77.06 to 254.4, from –94.41 to –26.32, and from 2.321 to 8.756. The estimated standard error is in a 95%-confidence interval from

4.947 to 7.321. The adjusted coefficient of multiple determination is $R^2_{adj} = 0.2095$. The regression parameters are significant (all tests with p_{value}<<0.05). The model is adequate (ANOVA with p_{value}<<0.05).

CYCLE 2

A new outlier was identified and rejected from the original sample: observation 2 for patient 05. This can be traced in the second sub-column of "Outliers after each cycle" in Table 2 (row shaded), where this observation has a value of 2 (indicating it was identified as an outlier in cycle 2). The previous 5 observations were confirmed again as outliers (their second sub-column of "Outliers after each cycle" has the value of 1 indicating their status from cycle 1 was confirmed in cycle 2). The model was constructed using the remaining 53 in-observations: $sPlt$=95.84 – 35.82×$Fibrinogen$+3.401×$Fibrinogen^2$+e. The 95% confidence intervals of the model's coefficients are from 49.43 to 142.2, from –53.55 to –18.08, and from 1.731 to 5.07. The estimated standard error is in a 95% confidence interval from 2.535 to 3.766. The adjusted coefficient of multiple determination is $R^2_{adj} = 0.2246$. The regression parameters are statistically significant (all tests with p_{value}<<0.005). The model is adequate (ANOVA with p_{value}<<0.05).

CYCLE 3

Two new outliers were identified and rejected from the original sample: observation 1 for patient 10 and observation 1 for patient 18. This can be traced in the third sub-column of "Outliers after each cycle" in Table 2 (rows shaded), where these observations have a value of 3 (indicating they were identified as outliers in cycle 3). The previous 6 observations were confirmed again as outliers (their third sub-column of "Outliers after each cycle" has the value of either 1 or 2 indicating their status from either cycle 1 or 2 was confirmed in cycle 3). The model was constructed using the remaining 51 in-observations as $sPlt$=128.7 – 48.92×$Fibrinogen$+4.683×$Fibrinogen^2$+e. The 95% confidence intervals of the model's coefficients are from 84.75 to 172.6, from –65.83 to –32.01, and from 3.079 to 6.288. The estimated standard error is in a 95% confidence interval from 2.217 to 3.321. The adjusted coefficient of multiple determination is $R^2_{adj} = 0.3985$. The regression parameters are statistically significant (all tests with p_{value}<<0.0005). The model is adequate (ANOVA with p_{value}<<0.05). The third cycle is complete, and the true number of cycles is C_{true}=C_{max}=3. The algorithm exits the cycle loop.

"BEST" LINEAR MODEL SELECTION

As expected, the "best" linear model is the one after the cycle 3 because it has the largest adjusted coefficient of multiple determination $R^2_{adj} = 0.3985$ (larger than the second cycle $R^2_{adj} = 0.2246$, and larger than the first cycle $R^2_{adj} = 0.2095$).
The results from the example show that throughout the cycles we observed new outliers being identified, but none of the initially identified ones returned to the original sample. We would usually expect that with each cycle, less outliers will be identified. On the contrary – in this example we saw that the rejection of the sixth outlier in cycle 2 was crucial for the rejection of the last two outliers in cycle 3, which were somewhat hidden up until after cycle 2. This well demonstrates the benefits of having cycles in our approach to allow thorough analysis and purging of the data from non-representative measurements.
The importance of using cycles in outlier rejection can also be seen by comparing cycle 1 (which is the result of no-cycle procedure for outlier rejection) and cycle 3 (which is the result of multiple cycle procedure developed in this paper). The results from the initial analysis were unacceptable. The significance of the regression parameters improved substantially between cycles 1 and 3, and the adequacy of the model increased substantially, with R^2_{adj} going from 0.2095 up to 0.3985, which is almost a two-fold increase. The final model after cycle 3 is presented in Fig. 5, demonstrating a strong dependence of $sPlt$ on $Fibrinogen$. This example illustrates that the fuzzy sample approach in combination with outlier rejection in cycles is a helpful tool in medical studies of biological systems that helps us minimize the effect of intra-individual sample heterogeneity on the conclusions based on inter-individual diversity.

6.2.2. Practical example 2

In this example, we explore the dependence of $sWBC$ on WBC and CRP. Table 3a and Table 3b present the measurement data, with a total of 296 records for 61 patients (patient codes given in columns 1, 7, 13, and 19). There is no limitation that we imposed on which patients to include in the data set. Columns 6, 12, 18 and 24 of Table 3a and Table 3b show the degrees of membership of each patient record.

INITIAL ANALYSIS

Initially, without any outlier rejection, we constructed a partial quadratic regression model (step-wise regression analysis was utilized to identify that only 2 of the 6 coefficients are not zero) for $sWBC$ on the predictors WBC and CRP, using the 296 in-observations as follows: $sWBC$=0.005376×WBC^2+0.0002298×CRP^2+e. In the terminology of Section 2, the response variable Y is $sWBC$ and the directly measured independent variables are the following: Z_1 is WBC and Z_2 is CRP (obviously, p=2). The q=2 predictor variables are: $X_1 = Z_1^2$, and $X_2 = Z_2^2$ (that is, $F_1(z_1, z_2) = z_1^2$, and $F_2(z_1, z_2) = z_2^2$). The

95%-confidence intervals of the model's coefficients are from 0.004057 to 0.006695, and from 0.000172 to 0.0002877. The estimated standard error is in a 95%-confidence interval from 2.125 to 2.498. The adjusted coefficient of multiple determination is $R_{adj}^2 = 0.1966$. The regression parameters are significant (all tests with p_{value}<<0.05) and the model is adequate (ANOVA with p_{value}=0).

The following results were achieved using Algorithms 1 and 2 with C_{max}=3. Table 3a and Table 3b present the results from the outlier detection through the two sub-columns of the parameter "Outliers after each cycle".

Cycle 1

A total of 16 outliers were identified and rejected from the original sample: observation 1 for patient 15, observation 5 for patient 15, observation 1 for patient 29, observation 2 for patient 38, observations 3 and 5 for patient 42, observations 3 and 4 for patient 43, observations 1, 3 and 5 for patient 46, observation 4 for patient 47, observation 5 for patient 49, observation 4 for patient 54, and observations 1 and 2 for patient 55. This can be traced in the first sub-column of "Outliers after each cycle" in Table 3a and Table 3b (shaded rows), where those 16 observations have a value of 1 (indicating they were identified as outliers in cycle 1). The model was constructed using the remaining 280 in-observations as: $sWBC$=0.003075×WBC^2+0.0002514×CRP^2+e. The 95%-confidence intervals of the model's coefficients are from 0.002393 to 0.003757 and from 0.0002167 to 0.0002862. The estimated standard error is in a 95%-confidence interval from 1.046 to 1.235. The adjusted coefficient of multiple determination is $R_{adj}^2 = 0.4282$. The regression parameters are significant (all tests with p_{value}<<0.05). The model is adequate (ANOVA with p_{value}=0).

Cycle 2

A total of 21 new outliers were identified and rejected from the original sample: observation 4 for patient 03, observation 4 for patient 12, observation 1 for patient 17, observation 2 for patient 21, observation 2 for patient 22, observation 4 for patient 32, observation 3 for patient 34, observation 3 for patient 35, observations 1 and 4 for patient 38, observation 2 for patient 43, observation 1 for patient 49, observation 5 for patient 50, observation 4 for patient 51, observation 2 for patient 52, observation 1 for patient 53, observation 5 for patient 55, observation 5 for patient 56, observation 2 for patient 57, observation 2 for patient 62, and observation 5 for patient 64. This can be traced in the second sub-column of "Outliers after each cycle" in Table 3a and Table 3b (shaded rows), where those 21 observations have a value of 2 (indicating they were identified as outliers in cycle 2). The previous 16 observations were confirmed again as outliers (their second sub-column of "Outliers after each cycle" has the value of 1 indicating their status from cycle 1 was confirmed in cycle 2). The model was constructed using the remaining 259 in-observations as: $sWBC$=0.001997×WBC^2+0.0002191×CRP^2+e. The 95%-confidence intervals of the model's coefficients are from 0.001501 to 0.002494 and from 0.0001915 to 0.0002467. The estimated standard error is in a 95%-confidence interval from 0.7193 to 0.8553. The adjusted coefficient of multiple determination is $R_{adj}^2 = 0.4942$. The regression parameters are significant (all tests with p_{value}<<0.05). The model is adequate (ANOVA with p_{value}=0).

Cycle 3

No new outliers were identified, and the third cycle is incomplete. The true number of cycles is C_{true}=2. The algorithm exits the cycle loop.

"Best" linear model selection

As expected, the "best" linear model is the one after cycle 2 because it has the largest adjusted coefficient of multiple determination $R_{adj}^2 = 0.4942$ (larger than the first cycle $R_{adj}^2 = 0.4282$).

The model after cycle 2 is final. In this way, we demonstrate another benefit of our algorithms, namely that sometimes we will achieve ultimate outlier detection that does not depend from the permitted count of cycles. In the example we also observed that throughout the two cycles, we identified new outliers, but none of the initially identified ones returned to the original sample. Hence, with the rejection of outliers in cycle 1 we were able to better detect new outliers in cycle 2. Even though the results from the initial analysis were acceptable (coefficients were significant and the model was adequate), we can see the importance of using cycles in outlier rejection by comparing cycle 1 (which is the result of no-cycle procedure for outlier rejection) and cycle 2 (which is the result of multiple cycle procedure developed in this paper), as 21 new outliers were purged from the data set and the adequacy of the model increased with R_{adj}^2 going from 0.4282 to 0.4942 (and being 0.1966 for the initial model with no outlier rejection).

The final model after cycle 2 is presented in Fig. 6, demonstrating a strong dependence of $sWBC$ on WBC and CRP. Thus, this case again illustrates that the fuzzy sample approach in combination with outlier rejection in cycles based on the multiple testing paradigm in medical studies of biological systems helps us minimize the effect of intra-individual sample heterogeneity on conclusions based on inter-individual diversity.

7. DISCUSSION AND CONCLUSIONS

In this paper, we focused on ways of improving detection and rejection of outliers in cycles over fuzzy data with subsequent construction of a linear regression model. To do so, we have defined two algorithms. The Algorithm 1 builds a linear regression model using n fuzzy in-observations. It relied on SVD decomposition to improve the quality of the linear regression model as it allowed eliminating multicollinearity in the predictor variables. Algorithm 1 helps identify the true zero values among the singular values automatically and in such a way is self-protected against singularities in the data. We employed a LOO approach with single testing procedure for each of the in-observations to calculate the p_{value} of statistical hypotheses that each of the in-observations is an outlier. The Algorithm 2 operated over fuzzy data and utilizes repeatedly Algorithm 1 as a working engine to identify the outliers in the data set. Algorithm 2 operated in predefined number of cycles, where some in-observations became outliers, and some outliers could then return to the set of in-observations. In such a way, Algorithm 2 accounted for the varying order of magnitude that outliers may have which could hide some of the outliers from the detection procedure. Each cycle in Algorithm 2 ran in two phases:

a) The first phase developed a high-quality intermediate model on a given set of in-observations by easily purging the potential outliers using single testing LOO procedure.

b) The second phase sought stringent confirmation for the outlier status of each current outlier from all cycles using Benjamini-Hochberg step-up multiple testing procedure controlling the FDR. All observations with non-confirmed outlier status were brought back to the in-observations data set. A final model was then constructed as an outcome of each cycle.

Algorithm 2 stops either when the predefined number of cycles is completed or when a cycle did not add new outliers. Our approach has an advantage in that it allowed in phase 2 of each cycle for the earlier declared outliers to potentially return to the data sample. In light of the discussions in Sections 1 and 3 regarding fuzzy regression models, we can see this is a rare feature of fuzzy regression procedures, yet desired one as it contributes to the quality and quantity of information for adequate regression analysis.

The usefulness of the proposed algorithms is demonstrated by means of three examples with fuzzy data. The first is a small illustrative case. The benefits of using cycles in our procedure was well demonstrates in this example, since in each cycle we had new outliers identified, while some returned to the data sample. We observed a ten-fold improvement of the significance of regression parameters between the first and the last cycle in the procedure.

Our other two examples are both related to a case study in medical research (namely, thrombosis research). The medical setup of those two examples necessitated the use of fuzzy data in order to avoid distortions in characteristics calculated from the data. Therefore, we constructed two quadratic linear regression models over the dependence between routinely available clinical data and structural characteristics of thrombi. In the first example, we observed that some outliers may hide or prevent other data points to be revealed as outliers (see Section 6.2.1), which was a particularly strong point in favor of our cycle-based approach. The second practical example in Section 6.2.2 demonstrates that the algorithms avoid calculations (cycles) that would not add up to the overall quality of the model and constructs the final model in a smaller number of cycles.

In all three examples, we observed that initial analysis constructed mostly unacceptable models, which subsequently are improved significantly (up to ten-fold) after the implementation of our algorithms, as indicated by their R^2_{adj}. We observed that each cycle in our examples added new outliers, with only one cycle in the illustrative example returning an outlier to the original data set (so the outliers from previous steps were in overall confirmed).

Our approach has certain limitations. First of all, we relied in our constructions on the classical linear regression assumption that the unexplained error U is normally distributed: $U \sim N\left(0,\sigma^2\right)$ (see [45]), which may not always be the case. Simulation techniques may allow to explore other setups for U. Secondly, our procedures (both Algorithm 1 and 2) are computationally complex and without software implementation they may impose significant challenge to implement. All proposed algorithms from this work, along with the example results were performed using original software codes in MATLAB R2019a © (that are available free of charge upon request from the authors). In such a way we have alleviated this limitation of our work. Finally, the strength of Algorithm 2 (i.e. that it allows to reject outliers in layers, accounting for the order of magnitude of outliers) may sometimes become problematic. If we perform way too many cycles, the procedure may (in some limited cases) lead to rejecting way too many of the initial observations as outliers. This is particularly true if the underlying model is quadratic or cubic, but the user employs a linear model in the procedure. To overcome this, it is recommended to run a default count of 3-4 cycles (which was demonstrated in the examples) and predefine the user input on the count of cycles.

We can outline several directions of future research. To overcome the limitation of our procedure only assuming normality of the unexplained error, we shall aim to utilize simulation techniques (e.g. Bootstrap simulations, see [26; 59] for discussion on simulation modelling with Bootstrap) to explore the nature of U. Another direction of future studies is to seek implementation of our algorithms in other practical case studies in management, economics, supply chain management, environmental research and engineering. Also, in our future research we shall aim to explore case studies and/or numerical examples where we will be able to compare the performance of our method versus other similar cycle-based outlier detection procedures over fuzzy data.

ACKNOWLEDGMENTS

This work was supported in part by the Hungarian National Research, Development and Innovation Office (NKFIH) (129528, KK) and the Higher Education Institutional Excellence Programme of the Ministry of Human Capacities in Hungary for the Molecular Biology thematic programme of Semmelweis University (KK). The research is also supported by the University of Tasmania's internal research and development fund RT.112222.

REFERENCES

1. Ariens, R.A.: Fibrin(ogen) and thrombotic disease. J Thromb Haemost. 11 Suppl1: 294-305 (2013)
2. Bargiela, A., Pedrycz, W., Nakashima, T.: Multiple regression with fuzzy data, Fuzzy Sets Syst. 158 2169-2188 (2007)
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 57 289–300 (1995)
4. Benjamini, Y., Yekutieli, D.: The Control of the False Discovery Rate in Multiple Testing under Dependency, The Annals of Statistics, 29(4) 1165-1188 (2001).
5. Benjamini, Y.: Discovering the false discovery rate, Journal of the Royal Statistical Society, Series B. 72(4) 405–416 (2010)
6. Bisserier, A., Boukezzoula, R., Galichet, S.: A revisited approach to linear fuzzy regression using trapezoidal fuzzy intervals, Inf. Sci., 180 3653-3673 (2010)
7. Chachi, J., Taheri, S. M., Arghami, N. R.: A hybrid fuzzy regression model and its application in hydrology engineering, Applied Soft Comput., 25 149-158 (2014)
8. Chachi, J., Taheri, S.: Multiple fuzzy regression model for fuzzy input-output data, Iranian Journal of Fuzzy Systems, 13 (4) 63-78 (2016)
9. Chan, K. Y., Kwong, C. K., Fogarty, T. C.: Modelling manufacturing processes using a genetic programming-based fuzzy regression with detection of outliers, Information Sciences, 180 506-518 (2010)
10. Chukhrova, N., Johannssen, A.: Fuzzy regression analysis: Systematic review and bibliography, Applied Soft Computing Journal 84: 105708 (2019)
11. Cook, R. D.: Influential observations in linear regression, Journal of the American Statistical Association 74(365): 169–174 (1979)
12. Coppi, R.: Management of uncertainty in statistical reasoning: The case of regression analysis, Internat. J. Approx. Reason. 47 (3): 284–305 (2008)
13. Coppi, R., D'Urso, P., Giordani, P., Santoro, A.: Least squares estimation of a linear regression model with LR fuzzy response, Comp. Stat. Data Anal., 51 267-286 (2006)
14. D'Urso, P.: Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data, Computational Statistics & Data Analysis, 42 47-72 (2003)
15. D'Urso, P., Gastaldi, T.: A least-squares approach to fuzzy linear regression analysis, Computational Statistics and Data Analysis 34: 427-440 (2000)
16. D'Urso, P., Massari, R., Santoro, A.: Robust fuzzy regression analysis, Information Science, 181 4154-4174 (2011)
17. D'Urso, P., Massari, R.: Weighted Least Squares and Least Median Squares estimation for the fuzzy linear regression analysis, Metron, 71, 279-306 (2013)
18. D'Urso, P., Santoro, A.: Fuzzy clusterwise linear regression analysis with symmetrical fuzzy output variable computational statistics & Data Analysis, 51(1), 287-313 (2006)
19. D'Urso, P., Santoro, A.: Goodness of fit and variable selection in the fuzzy multiple linear regression, Fuzzy Sets and Systems, 157: 2627-2647 (2006)
20. Dehghan, M., Hamidi, F., Salajegheh, H.: Study of Linear Regression Based on Least Squares and Fuzzy Least Absolutes Deviations and its Application in Geography, 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) 1-6 (2015)
21. Denoeux, Th.: Maximum likelihood estimation from fuzzy data using the EM algorithm. Fuzzy Sets and Systems, 183 72–91 (2011)
22. Diamond, P.: Fuzzy least squares. Information Sciences, 46, 141-157 (1988)
23. Dubois, D., Nguyen, H.T., Prade, H.: Possibility theory, probability and fuzzy sets misunderstandings, bridges and gaps. In: Dubois D., Prade H. (eds) Fundamentals of fuzzy sets. The Handbooks of Fuzzy Sets Series, vol 7: 343-438, Springer, Boston, MA (2000)

24. Dubois, D., Prade, H: Fuzzy sets and probability: misunderstandings, bridges and gaps, Second IEEE International Conference on Fuzzy Systems, San Francisco, CA, USA, vol 2: 1059-1068 (1993)
25. D'Urso, P, Massari, R., Santoro, A.: A class of fuzzy clusterwise regression models, Information Sciences, 180, 4737-4762 (2010)
26. Efron, B., Tibshirani, R.: An introduction to the Bootstrap, New York, NY, USA: Chapman&Hall, pp. 45-57 (1993)
27. Farkas, A., Farkas, V.J., Gubucz, I., Szabó, L., Bálint, K., Tenekedjiev, K., Nagy, A.I., Sótonyi, P., Hidi, L., Nagy, Z., Szikora, I., Merkely, B., Kolev, K.: Neutrophil extracellular traps in thrombi retrieved during interventional treatment of ischemic arterial diseases. Thromb Res 175:46-52 (2019)
28. Ferraro, M. B., Coppi, R., Gonzalez Rodriguez, G., Colubi, A.: A linear regression model for imprecise response, Int. J. Approx. Reason., 51 759-770 (2010)
29. Freedman, D.: Statistical Models: Theory and Practice. Cambridge University Press (2009)
30. Gao, P., Gao, Y.: Quadrilateral Interval Type-2 Fuzzy Regression Analysis for Data Outlier Detection, Mathematical Problems in Engineering, Volume 2019, Article ID 4914593, 9 pages, https://doi.org/10.1155/2019/4914593 (2019)
31. Gladysz, B., Kuchta, D.: Outliers detection in selected fuzzy regression models, in WILF '07: Proceedings of the 7th international workshop on Fuzzy Logic and Applications, (Berlin, Heidelberg), 211-218, Springer-Verlag (2007)
32. Gujarati D. N., Porter, D.: Basic Econometrics, McGraw-Hill, Fifth Edition (2008)
33. Imoto, S., Yabuuchi, Y., Watada, J.: Fuzzy regression model of R & D project evaluation, Appl. Soft Comput. 8 (3) 1266–1273 (2008)
34. Jajuga, K.: Linear fuzzy regression, Fuzzy Sets and Systems 20 (3) 343–353 (1986)
35. Jinn, J.H., Song, C., Chao, J.C.: A study of fuzzy linear regression. In: InterStat, (6) (accessed on 08 November 2020), http://interstat.statjournals.net/YEAR/2008/articles/0807006.pdf (2008)
36. Kao, C., Chyu, C.: A fuzzy linear regression model with better explanatory power, Fuzzy Sets and Systems, 126 401-409 (2002)
37. Kao, C., Chyu, C.: Least-square estimates in fuzzy regression analysis, European Journal of Operations Research, 148 426-435 (2003)
38. Khashei, M., Hejazi, S.R., Bijari, M.: A new hybrid artificial neural networks and fuzzy regression model for time series forecasting, Fuzzy Sets and Systems 159 (7) 769–786 (2008)
39. Klir, G.: Foundations of fuzzy set theory and fuzzy logic: a historical overview, Int J Gen Syst, 30(2) 91–131 (2001)
40. Kovács, A., Sótonyi, P., Nagy, A.I., Tenekedjiev, K., Wohner, N., Komorowicz, E., Kovács, E., Nikolova, N.D., Szabó, L., Kovalszky, I., Machovich, R., Szelid, Z., Becker, D., Merkely, B., Kolev, K.: Ultrastructure and Composition of Thrombi in Coronary and Peripheral Artery Disease: Correlations with Clinical and Laboratory Findings, Thrombosis Research, 135(4). 760-766 (2015)
41. Kwong, C.K., Chen, Y., Wong, H.: Modeling manufacturing processes using fuzzy regression with the detection of outliers, Int J Adv Manuf Technol 36: 547–557 (2008)
42. Lee, H., Tanaka, H.: Fuzzy approximations with non-symmetric fuzzy parameters in fuzzy regression analysis, J. Oper. Res. Soc. Japan 42 (1): 98–112 (1999)
43. Lee, H.T., Chen, S.H.: Fuzzy regression model with fuzzy input and output data for manpower forecasting, Fuzzy Sets and Systems 119 (2) 205–213 (2001)
44. Lu, J., Wang, R.: An enhanced fuzzy linear regression model with more flexible spreads, Fuzzy Sets Syst. 160 2505-2523 (2009)
45. Maddala, G. S.: Introduction to Econometrics, 2nd ed., New York: MacMillan (1992)
46. Magnusson, M., Andersen, M., Jonasson, J., Vehtari, A.: Bayesian leave-one-out cross-validation for large data, Proceedings of the 36th International Conference on Machine Learning, PMLR 97: 4244-4253 (2019)
47. Mangold, A., Alias, S., Scherz, T., Hofbauer, T., Jakowitsch, J., Panzenböck, A., Simon, D., Laimer, D., Bangert, C., Kammerlander, A., Mascherbauer, J., Winter, M.P., Distelmaier, K., Adlbrecht, C., Preissner, K.T., Lang, I.M.: Coronary neutrophil extracellular trap burden and deoxyribonuclease activity in ST-elevation acute coronary syndrome are predictors of ST-segment resolution and infarct size. Circ Res. 116(7):1182-92 (2015)
48. Mashinchi, M. H., Orgun, M. A., Mashinchi, M. R.: A Least Square Approach for the Detection and Removal of Outliers for Fuzzy Linear Regressions, Second World Congress on Nature and Biologically Inspired Computing Dec. 15-17, 2010 in Kitakyushu, Fukuoka, Japan 134-139 (2010)
49. McCloskey, A.: Bonferroni-based size-correction for nonstandard testing problems, Journal of Econometrics 200 17–35 (2017)
50. Mittelhammer, R.C., Judge, G., Miller, D.: Econometric Foundations. Cambridge University Press (2000)
51. Modarres, M., Nasrabadi, E., Nasrabadi, M.: Fuzzy linear regression analysis from the point of view risk, Int. J. Uncertain., Fuzziness Knowledge-Based Syst. 12 635-649 (2004)
52. Modarres, M., Nasrabadi, E., Nasrabadi, M.: Fuzzy linear regression with least squares errors, Appl. Math. Comput., 163 977-989 (2005)
53. Montgomery, D., Peck, E., Vining, G.: Introduction to linear regression analysis, Wiley (2001)

54. Nasrabadi, E., Hashemi, S.M., Ghatee, M.: An LP-based approach to outliers detection in fuzzy regression analysis, International journal of uncertainty fuzziness and knowledge-based systems, 15(4) 441-456 (2007)
55. Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W.: Applied linear statistical models, 4th ed. Chicago: Irwin (1996)
56. Nikolova, N., Panayotov, P., Panayotova, D., Ivanova, S., Tenekedjiev, K.: Using fuzzy sets in surgical treatment selection and homogenizing stratification of patients with significant chronic ischemic mitral regurgitation. International Journal of Computational Intelligence Systems, 12 (2019; in print)
57. Nikolova, N.D., Toneva-Zheynova, D., Naydenov, D., Tenekedjiev, K.: Imputing Missing Values of Environment Multi-Dimensional Vectors Using a Modified Roweis Algorithm, Proc. IFAC Workshop on Dynamics and Control of Agriculture and Food Processing, Plovdiv, Bulgaria 119-205, (2012)
58. Peters, G.: Fuzzy linear regression with fuzzy intervals, Fuzzy Sets and Systems, 63 45-55 (1994)
59. Politis, D.: Computer-intensive methods in statistical analysis, IEEE Signal Processing Magazine, vol. 15(1): 39-55 (1998)
60. Press, W. H., Teukolski, S. A., Vetterling, W. T., Flannery, B. P.: Numerical Recipes – The Art of Scientific Computing, 3rd edition, Cambridge University Press (2007)
61. Romano, J.P., Shaikh, A.M., Wolf, M.: A practical two-step method for testing moment inequalities. Econometrica 82 (5) 1979–2002 (2014)
62. Ruspini, E.: Possibility as similarity; the semantics of fuzzy logic, UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, MIT, Cambridge, MA, USA, July 27-29 (1990)
63. Selvanathan, SA, Selvanathan, S, Keller, G.: Business Statistics: Australia New Zealand, Seventh Edition, Cengage Learning Australia (2017)
64. Suk, H.W., Hwang, H., Regularized fuzzy clusterwise ridge regression, Adv. Data Analy. Classif. 4 (1) 35–51 (2010)
65. Tanaka, H, Hayashi, I., Watada, J.: Possibilistic linear regression analysis for fuzzy data, European Journal of Operations Research, 40 389-396 (1989)
66. Tanaka, H, Vejima, S, Asai, K.: Linear regression analysis with fuzzy model, IEEE Transactions on Systems, Man and Cybernetics, 12 903-907 (1982)
67. Tenekedjiev, K., Karakatsanis, N., Bekiaris, A.: Fictitious Covariance Matrices, Proc. Forth International Conference, Adaptive Computing in Design and Manufacture ACDM'2000, 23-26, Plymouth, UK (2000)
68. Tenekedjiev, K., Radoinova D.: Numeral procedures for stature estimating according to length of limb long bones in Bulgarian and Hungarian populations. Acta morphologica et anthropologica (6) 90-97 (2001)
69. Varjú, I., Sótonyi, P., Machovich, R., Szabó, L., Tenekedjiev, K., Silva, M.M., Longstaff, C., Kolev, K.: Hindered dissolution of fibrin formed under mechanical stress. J Thromb Haemost 9:979–86 (2011)
70. Viertl, R.: Statistical Methods for Fuzzy Data. John Wiley (2011)
71. Wang, G., Guo, P.: Outlier detection approaches in fuzzy regression models, 2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), Edmonton, AB, 980-985 (2013)
72. Wohner, N., Sótonyi, P., Machovich, R., Szabó, L., Tenekedjiev, K., Silva, M.M., Longstaff, C., Kolev, K.: Lytic resistance of fibrin containing red blood cells. Arterioscler Thromb Vasc Biol 31:2306–13 (2011)
73. Yan, X., Gang Su, X.: Linear regression analysis: Theory and computing, World Scientific (2009)
74. Yang, M.-S., Ko, C.-H.: On cluster-wise fuzzy regression analysis, IEEE Trans. Syst. Man Cybern. B 27 (1) 1–13 (1997)
75. Yang, Z., Yin, Y., Chen, Y.: Robust fuzzy varying coefficient regression analysis with crisp inputs and gaussian fuzzy output, J. Comput. Sci. Eng. 7 (4): 263–271 (2013)

77.

Table 1. Set of 12 observations of the predictor variable *X* and the response variable *Y* for the illustrative example in Section 6.1. The degrees of membership are given in columns 5 and 10 for each record. The sub-columns of "Outliers after each cycle" indicate whether the observation was declared outlier (shaded rows), and at which cycle.

No.	**Outliers after each cycle**	***x***	***y***	**Degree of membership *μ***	**No.**	**Outliers after each cycle**	***x***	***y***	**Degree of membership *μ***
1	0 0 0	7.1	18.6	1	**7**	1 1 0	17	38.6	1
2	0 0 0	7.6	19.7	0.75	**8**	0 0 0	4.4	14.5	0.4
3	1 1 1	3.64	70.7	0.5	**9**	0 0 0	5.7	15.5	0.25
4	0 0 0	7.6	18.7	0.5	**10**	0 2 2	7.9	–9.8	0.75
5	0 0 0	6.2	18.1	1	**11**	0 0 0	7.8	20.5	1
6	0 0 0	3.49	12.1	0.5	**12**	0 0 3	15	2	1

Table 2. Set of 59 observations of *Fibrinogen* and *sPlt* for a total of 13 patients for practical example 1 from Section 6.2.1. The degrees of membership are given in columns 5, 10, 15 and 20 for each record. The sub-columns of "Outliers after each cycle" indicate whether the observation was declared outlier (shaded rows), and at which cycle.

Patient code	**Outliers after each cycle**	***sPlt***	***Fibrinogen***	**Degree of membership**	**Patient code**	**Outliers after each cycle**	***sPlt***	***Fibrinogen***	**Degree of membership**	**Patient code**	**Outliers after each cycle**	***sPlt***	***Fibrinogen***	**Degree of membership**	**Patient code**	**Outliers after each cycle**	***sPlt***	***Fibrinogen***	**Degree of membership**
05	1 1 1	65	4.26	0.50	**18**	0 0 3	8.9	5.6	0.20	**34**	0 0 0	2.4	5.3	0.20	**54**	0 0 0	0	5.3	0.20
05	0 2 2	31.1	4.26	0.50	**18**	0 0 0	3.7	5.6	0.20	**34**	1 1 1	63.5	5.3	0.20	**54**	0 0 0	0.6	5.3	0.20
06	0 0 0	7.5	5.95	0.25	**18**	0 0 0	0.7	5.6	0.20	**34**	0 0 0	3.6	5.3	0.20	**54**	0 0 0	0.3	5.3	0.20
06	0 0 0	0	5.95	0.25	**18**	0 0 0	0	5.6	0.20	**34**	0 0 0	0.5	5.3	0.20	**54**	0 0 0	0.6	5.3	0.20
06	0 0 0	0.6	5.95	0.25	**18**	0 0 0	0.7	5.6	0.20	**34**	0 0 0	0.7	5.3	0.20	**54**	0 0 0	2	5.3	0.20
06	0 0 0	0	5.95	0.25	**20**	0 0 0	6.1	4.3	0.20	**35**	0 0 0	3.1	5.95	0.20	**56**	0 0 0	0.2	4.42	0.20
07	0 0 0	9.8	4.26	0.20	**20**	0 0 0	1.5	4.3	0.20	**35**	0 0 0	3.2	5.95	0.20	**56**	0 0 0	0	4.42	0.20
07	0 0 0	3.8	4.26	0.20	**20**	0 0 0	9	4.3	0.20	**35**	0 0 0	4.6	5.95	0.20	**56**	0 0 0	6.7	4.42	0.20
07	0 0 0	5	4.26	0.20	**20**	0 0 0	0.3	4.3	0.20	**35**	1 1 1	81	5.95	0.20	**56**	0 0 0	2.8	4.42	0.20
07	0 0 0	6.5	4.26	0.20	**20**	0 0 0	5.8	4.3	0.20	**35**	1 1 1	73.7	5.95	0.20	**56**	1 1 1	73.6	4.42	0.20
07	0 0 0	3.6	4.26	0.20	**33**	0 0 0	0.9	6	0.20	**49**	0 0 0	5.9	5.27	0.20	**62**	0 0 0	0.7	4.64	0.20
10	0 0 3	1.5	6.58	0.33	**33**	0 0 0	1.4	6	0.20	**49**	0 0 0	6.9	5.27	0.20	**62**	0 0 0	0.5	4.64	0.20
10	0 0 0	10.6	6.58	0.33	**33**	0 0 0	2.2	6	0.20	**49**	0 0 0	1.9	5.27	0.20	**62**	0 0 0	0.2	4.64	0.20
10	0 0 0	12.6	6.58	0.33	**33**	0 0 0	3.4	6	0.20	**49**	0 0 0	1.5	5.27	0.20	**62**	0 0 0	7.8	4.64	0.20
					33	0 0 0	0.1	6	0.20	**49**	0 0 0	1.5	5.27	0.20	**62**	0 0 0	5	4.64	0.20

Table 3a. Set of 296 observations of *WBC*, *CRP* and *sWBC* for a total of 61 patients for practical example 2 from Section 6.2.2. The degrees of membership are given in columns 6, 12, 18 and 24 for each record. The sub-columns of "Outliers after each cycle" indicate whether the observation was declared outlier (shaded rows), and at which of the two cycles (*cont.*)

Patient	**Outliers after each cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**	**Patient**	**Outliers after each cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**	**Patient**	**Outliers after each cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**	**Patient**	**Outliers after each cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**
01	0 0	0	10.7	1.35	0.20	**13**	0 0	1.4	16.4	5.5	0.20	**23**	0 0	0.6	10.9	3.65	0.20	**33**	0 0	0	11.1	4.2	0.20
01	0 0	0.8	10.7	1.35	0.20	**13**	0 0	0.3	16.4	5.5	0.20	**23**	0 0	0	10.9	3.65	0.20	**33**	0 0	0	11.1	4.2	0.20
01	0 0	0.6	10.7	1.35	0.20	**13**	0 0	0	16.4	5.5	0.20	**23**	0 0	0	10.9	3.65	0.20	**33**	0 0	0	11.1	4.2	0.20
01	0 0	0	10.7	1.35	0.20	**13**	0 0	0	16.4	5.5	0.20	**23**	0 0	0	10.9	3.65	0.20	**33**	0 0	0	11.1	4.2	0.20
01	0 0	0.3	10.7	1.35	0.20	**13**	0 0	0	16.4	5.5	0.20	**23**	0 0	0	10.9	3.65	0.20	**33**	0 0	0	11.1	4.2	0.20
02	0 0	1.4	10.8	16.7	0.20	**14**	0 0	0	9.7	15	0.25	**24**	0 0	0	9.3	4.3	0.20	**34**	0 0	0	5	9.4	0.20
02	0 0	0.9	10.8	16.7	0.20	**14**	0 0	0	9.7	15	0.25	**24**	0 0	0	9.3	4.3	0.20	**34**	0 0	0	5	9.4	0.20
02	0 0	0	10.8	16.7	0.20	**14**	0 0	0	9.7	15	0.25	**24**	0 0	0	9.3	4.3	0.20	**34**	0 2	2.8	5	9.4	0.20
02	0 0	0	10.8	16.7	0.20	**14**	0 0	0	9.7	15	0.25	**24**	0 0	0	9.3	4.3	0.20	**34**	0 0	0	5	9.4	0.20
02	0 0	0	10.8	16.7	0.20	**15**	1 1	5.6	11.6	0.22	0.20	**24**	0 0	0	9.3	4.3	0.20	**34**	0 0	0	5	9.4	0.20
03	0 0	0.9	9.9	8.3	0.20	**15**	0 0	0	11.6	0.22	0.20	**25**	0 0	0	8.09	0.46	0.20	**35**	0 0	0	19.5	44.6	0.20
03	0 0	0	9.9	8.3	0.20	**15**	0 0	1.6	11.6	0.22	0.20	**25**	0 0	0	8.09	0.46	0.20	**35**	0 0	0	19.5	44.6	0.20
03	0 0	0	9.9	8.3	0.20	**15**	0 0	0	11.6	0.22	0.20	**25**	0 0	0	8.09	0.46	0.20	**35**	0 2	4.3	19.5	44.6	0.20
03	0 2	3.4	9.9	8.3	0.20	**15**	1 1	6.4	11.6	0.22	0.20	**25**	0 0	0	8.09	0.46	0.20	**35**	0 0	0	19.5	44.6	0.20
03	0 0	0	9.9	8.3	0.20	**16**	0 0	0	11.6	1.9	0.20	**25**	0 0	0	8.09	0.46	0.20	**35**	0 0	0	19.5	44.6	0.20
04	0 0	3.1	10	58.6	0.25	**16**	0 0	0	11.6	1.9	0.20	**26**	0 0	0	8.4	17	0.25	**37**	0 0	0	8.03	0.42	0.20
04	0 0	0	10	58.6	0.25	**16**	0 0	0	11.6	1.9	0.20	**26**	0 0	0	8.4	17	0.25	**37**	0 0	0	8.03	0.42	0.20
04	0 0	1.5	10	58.6	0.25	**16**	0 0	0	11.6	1.9	0.20	**26**	0 0	0	8.4	17	0.25	**37**	0 0	0	8.03	0.42	0.20
04	0 0	0	10	58.6	0.25	**16**	0 0	0	11.6	1.9	0.20	**26**	0 0	0	8.4	17	0.25	**37**	0 0	0	8.03	0.42	0.20
05	0 0	0	16.7	1.22	0.50	**17**	0 2	2.5	6.8	1.3	0.20	**27**	0 0	0	7.6	3.4	0.20	**37**	0 0	0	8.03	0.42	0.20
05	0 0	0	16.7	1.22	0.50	**17**	0 0	0	6.8	1.3	0.20	**27**	0 0	1.2	7.6	3.4	0.20	**38**	0 2	3.5	14.57	4.26	0.20
06	0 0	0	16.9	14.5	0.25	**17**	0 0	0	6.8	1.3	0.20	**27**	0 0	0	7.6	3.4	0.20	**38**	1 1	5.8	14.57	4.26	0.20
06	0 0	0	16.9	14.5	0.25	**17**	0 0	0	6.8	1.3	0.20	**27**	0 0	0	7.6	3.4	0.20	**38**	0 0	0	14.57	4.26	0.20
06	0 0	0	16.9	14.5	0.25	**17**	0 0	0	6.8	1.3	0.20	**27**	0 0	0	7.6	3.4	0.20	**38**	0 2	3.5	14.57	4.26	0.20
06	0 0	0	16.9	14.5	0.25	**18**	0 0	0	15.9	5.12	0.20	**28**	0 0	0	9.5	24.9	0.20	**38**	0 0	2.1	14.57	4.26	0.20
07	0 0	0	15.4	7.48	0.20	**18**	0 0	0	15.9	5.12	0.20	**28**	0 0	0	9.5	24.9	0.20	**39**	0 0	0	12.8	4.9	0.17
07	0 0	0	15.4	7.48	0.20	**18**	0 0	0	15.9	5.12	0.20	**28**	0 0	0	9.5	24.9	0.20	**39**	0 0	0	12.8	4.9	0.17
07	0 0	0	15.4	7.48	0.20	**18**	0 0	0	15.9	5.12	0.20	**28**	0 0	1.4	9.5	24.9	0.20	**39**	0 0	0	12.8	4.9	0.17
07	0 0	0	15.4	7.48	0.20	**18**	0 0	0.8	15.9	5.12	0.20	**28**	0 0	1	9.5	24.9	0.20	**39**	0 0	0	12.8	4.9	0.17
07	0 0	0	15.4	7.48	0.20	**19**	0 0	0	12.8	21.3	0.20	**29**	1 1	9.6	8.8	4.3	0.20	**39**	0 0	0.9	12.8	4.9	0.17
09	0 0	0	17.9	23.8	0.20	**19**	0 0	0	12.8	21.3	0.20	**29**	0 0	1.3	8.8	4.3	0.20	**39**	0 0	0.7	12.8	4.9	0.17
09	0 0	1.5	17.9	23.8	0.20	**19**	0 0	0	12.8	21.3	0.20	**29**	0 0	0	8.8	4.3	0.20	**41**	0 0	1.2	16.7	34.8	0.20
09	0 0	0	17.9	23.8	0.20	**19**	0 0	0	12.8	21.3	0.20	**29**	0 0	0.5	8.8	4.3	0.20	**41**	0 0	1.5	16.7	34.8	0.20
09	0 0	0.8	17.9	23.8	0.20	**19**	0 0	0	12.8	21.3	0.20	**29**	0 0	0	8.8	4.3	0.20	**41**	0 0	2.1	16.7	34.8	0.20
09	0 0	1.6	17.9	23.8	0.20	**20**	0 0	0	21.3	4.2	0.20	**30**	0 0	0	10.8	3.8	0.20	**41**	0 0	1.7	16.7	34.8	0.20
10	0 0	0	9.01	12	0.33	**20**	0 0	0	21.3	4.2	0.20	**30**	0 0	0	10.8	3.8	0.20	**41**	0 0	0	16.7	34.8	0.20
10	0 0	0	9.01	12	0.33	**20**	0 0	0	21.3	4.2	0.20	**30**	0 0	1	10.8	3.8	0.20	**42**	0 0	0	10.9	1.3	0.20
10	0 0	0	9.01	12	0.33	**20**	0 0	0	21.3	4.2	0.20	**30**	0 0	0	10.8	3.8	0.20	**42**	0 0	0.6	10.9	1.3	0.20
11	0 0	2.3	12.3	1.33	0.20	**20**	0 0	0	21.3	4.2	0.20	**30**	0 0	0	10.8	3.8	0.20	**42**	1 1	8.1	10.9	1.3	0.20
11	0 0	2.7	12.3	1.33	0.20	**21**	0 0	0	12.3	2.57	0.20	**31**	0 0	0	7.96	1.25	0.20	**42**	0 0	2.1	10.9	1.3	0.20
11	0 0	0	12.3	1.33	0.20	**21**	0 2	2.8	12.3	2.57	0.20	**31**	0 0	0	7.96	1.25	0.20	**42**	1 1	8.1	10.9	1.3	0.20
11	0 0	0	12.3	1.33	0.20	**21**	0 0	1.7	12.3	2.57	0.20	**31**	0 0	0	7.96	1.25	0.20	**43**	0 0	5.4	11.1	188	0.17
11	0 0	1.7	12.3	1.33	0.20	**21**	0 0	0	12.3	2.57	0.20	**31**	0 0	0	7.96	1.25	0.20	**43**	0 2	13.3	11.1	188	0.17
12	0 0	0	13.3	22.1	0.20	**21**	0 0	0	12.3	2.57	0.20	**31**	0 0	0	7.96	1.25	0.20	**43**	1 1	0.9	11.1	188	0.17
12	0 0	0	13.3	22.1	0.20	**22**	0 0	0	8.9	6.08	0.20	**32**	0 0	0	19.24	40	0.20	**43**	1 1	15.4	11.1	188	0.17
12	0 0	1.2	13.3	22.1	0.20	**22**	0 2	2.8	8.9	6.08	0.20	**32**	0 0	1.2	19.24	40	0.20	**43**	0 0	12	11.1	188	0.17
12	0 2	3	13.3	22.1	0.20	**22**	0 0	0	8.9	6.08	0.20	**32**	0 0	0	19.24	40	0.20	**43**	0 0	6.5	11.1	188	0.17
12	0 0	0	13.3	22.1	0.20	**22**	0 0	0	8.9	6.08	0.20	**32**	0 2	4.1	19.24	40	0.20	**44**	0 0	0	17.7	0.65	0.25
						22	0 0	0	8.9	6.08	0.20	**32**	0 0	3.4	19.24	40	0.20	**44**	0 0	0.8	17.7	0.65	0.25
																		44	0 0	0.9	17.7	0.65	0.25
																		44	0 0	2.5	17.7	0.65	0.25

Table 3b. Set of 296 observations of *WBC*, *CRP* and *sWBC* for a total of 61 patients for practical example 2 from Section 6.2.2. The degrees of membership are given in columns 6, 12, 18 and 24 for each record. The sub-columns of "Outliers after each cycle" indicate whether the observation was declared outlier (shaded if so), and at which of the two cycles

Patient	**Outliers after a cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**	**Patient**	**Outliers after a cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**	**Patient**	**Outliers after a cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**	**Patient**	**Outliers after a cycle**	***sWBC***	***WBC***	***CRP***	**Degree of membership**
45	0 0	1.3	9.97	4.94	0.20	**51**	0 0	1.3	6.85	4.4	0.20	**56**	0 0	0	10.23	24.3	0.20	**61**	0 0	2.5	19.5	7.6	0.20
45	0 0	0.7	9.97	4.94	0.20	**51**	0 0	2.4	6.85	4.4	0.20	**56**	0 0	0	10.23	24.3	0.20	**61**	0 0	0.6	19.5	7.6	0.20
45	0 0	0	9.97	4.94	0.20	**51**	0 0	0.7	6.85	4.4	0.20	**56**	0 0	0	10.23	24.3	0.20	**61**	0 0	2.3	19.5	7.6	0.20
45	0 0	0	9.97	4.94	0.20	**51**	0 2	3.2	6.85	4.4	0.20	**56**	0 0	0	10.23	24.3	0.20	**61**	0 0	1.5	19.5	7.6	0.20
45	0 0	0	9.97	4.94	0.20	**51**	0 0	0	6.85	4.4	0.20	**56**	0 2	3.9	10.23	24.3	0.20	**61**	0 0	2.3	19.5	7.6	0.20
46	1 1	6.8	3.3	19.5	0.20	**52**	0 0	0	5.2	7.7	0.20	**57**	0 0	0.9	15.4	4.8	0.20	**62**	0 0	0	10.2	9.7	0.20
46	0 0	0.1	3.3	19.5	0.20	**52**	0 2	3.6	5.2	7.7	0.20	**57**	0 2	3.6	15.4	4.8	0.20	**62**	0 2	2.8	10.2	9.7	0.20
46	1 1	6.1	3.3	19.5	0.20	**52**	0 0	0	5.2	7.7	0.20	**57**	0 0	0	15.4	4.8	0.20	**62**	0 0	0	10.2	9.7	0.20
46	0 0	0	3.3	19.5	0.20	**52**	0 0	0	5.2	7.7	0.20	**57**	0 0	0	15.4	4.8	0.20	**62**	0 0	0	10.2	9.7	0.20
46	1 1	7.8	3.3	19.5	0.20	**52**	0 0	0	5.2	7.7	0.20	**57**	0 0	0	15.4	4.8	0.20	**62**	0 0	0	10.2	9.7	0.20
47	0 0	0	5.8	1.8	0.20	**53**	0 2	3.1	9.7	1.56	0.20	**58**	0 0	0	9	1.8	0.20	**63**	0 0	0	11	5.6	0.25
47	0 0	0	5.8	1.8	0.20	**53**	0 0	0.5	9.7	1.56	0.20	**58**	0 0	1.5	9	1.8	0.20	**63**	0 0	0	11	5.6	0.25
47	0 0	0	5.8	1.8	0.20	**53**	0 0	2	9.7	1.56	0.20	**58**	0 0	2.4	9	1.8	0.20	**63**	0 0	1.6	11	5.6	0.25
47	1 1	12.5	5.8	1.8	0.20	**53**	0 0	0.8	9.7	1.56	0.20	**58**	0 0	0	9	1.8	0.20	**63**	0 0	0	11	5.6	0.25
47	0 0	1.4	5.8	1.8	0.20	**53**	0 0	0	9.7	1.56	0.20	**58**	0 0	0	9	1.8	0.20	**64**	0 0	0	14.97	4.1	0.20
49	0 2	3.4	8.7	6.43	0.20	**54**	0 0	0	11	4.9	0.20	**59**	0 0	0	11.6	59	0.20	**64**	0 0	1.5	14.97	4.1	0.20
49	0 0	0	8.7	6.43	0.20	**54**	0 0	0.7	11	4.9	0.20	**59**	0 0	2.3	11.6	59	0.20	**64**	0 0	0	14.97	4.1	0.20
49	0 0	0	8.7	6.43	0.20	**54**	0 0	0	11	4.9	0.20	**59**	0 0	3.2	11.6	59	0.20	**64**	0 0	2.9	14.97	4.1	0.20
49	0 0	0	8.7	6.43	0.20	**54**	1 1	10.6	11	4.9	0.20	**59**	0 0	0	11.6	59	0.20	**64**	0 2	3.2	14.97	4.1	0.20
49	1 1	5.3	8.7	6.43	0.20	**54**	0 0	0.6	11	4.9	0.20	**59**	0 0	1.6	11.6	59	0.20	**65**	0 0	0.6	15.9	9.2	0.20
50	0 0	0	16.11	5.82	0.20	**55**	1 1	20.8	25.9	16.6	0.20	**60**	0 0	0	11	0.31	0.20	**65**	0 0	0	15.9	9.2	0.20
50	0 0	0	16.11	5.82	0.20	**55**	1 1	13.8	25.9	16.6	0.20	**60**	0 0	0	11	0.31	0.20	**65**	0 0	0	15.9	9.2	0.20
50	0 0	0	16.11	5.82	0.20	**55**	0 0	3.7	25.9	16.6	0.20	**60**	0 0	1.2	11	0.31	0.20	**65**	0 0	0	15.9	9.2	0.20
50	0 0	0	16.11	5.82	0.20	**55**	0 0	2.4	25.9	16.6	0.20	**60**	0 0	0.6	11	0.31	0.20	**65**	0 0	1.7	15.9	9.2	0.20
50	0 2	3.5	16.11	5.82	0.20	**55**	0 2	7.3	25.9	16.6	0.20	**60**	0 0	0	11	0.31	0.20						

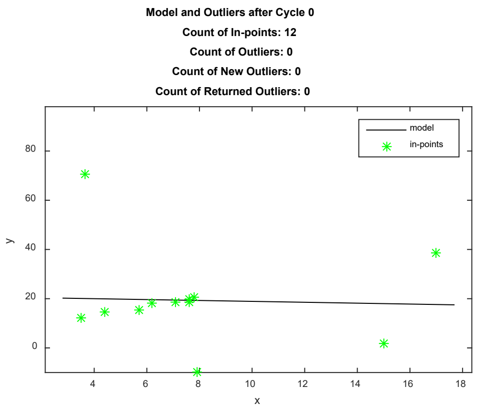

Fig. 1. Regression model $y=-0.1823x+20.74+e$ and outliers (no such identified) from the initial cycle in the illustrative example from Section 6.1, the parameters are not significant (all tests with $p_{value}>>0.05$) and the model is inadequate (ANOVA with $p_{value}=0.8467$)

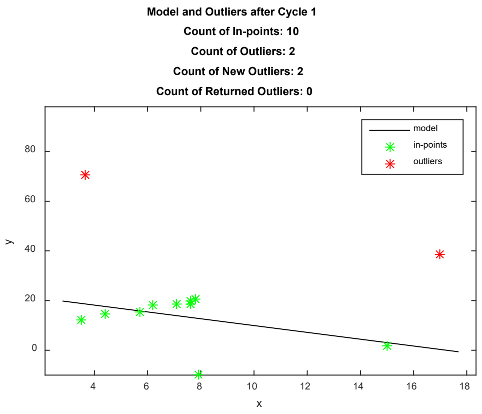

Fig. 2. Regression model $y=-1.371x+23.65+e$ and outliers (2 identified) from cycle 1 in the illustrative example in Section 6.1, the parameters are with varying significance (first with $p_{value}=0.2025$ and insignificant, second with $p_{value}=0.02250$ and significant), and the model is not adequate (ANOVA with $p_{value}=0.2022$)

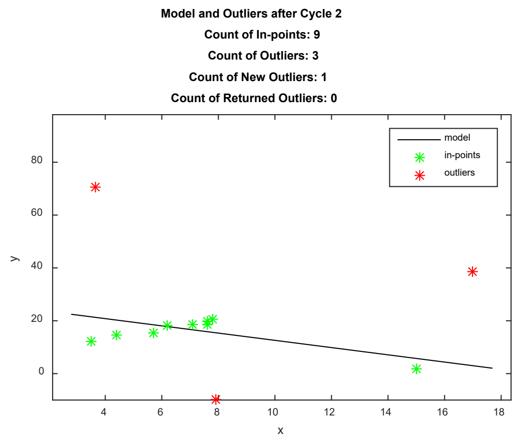

Fig. 3. Regression model $y=-1.37x+26.30+e$ and outliers (1 new and 2 from cycle 1) from cycle 2 of the illustrative example in Section 6.1, the parameters are significant (all tests with $p_{value}<0.05$), and the model is adequate (ANOVA with $p_{value}=0.0264$)

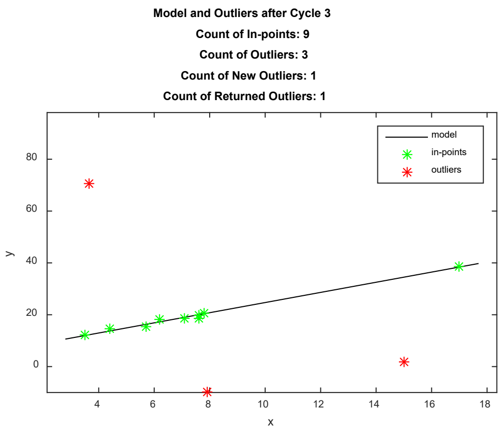

Fig. 4. Regression model $y=1.958x+5.128+e$ and outliers (1 new, 2 from cycle 1 and 2, and 1 from cycle 1 returned to original sample) from cycle 3 of the illustrative example in Section 6.1, the parameters are significant (all tests with $p_{value}<<0.05$), and the model is adequate (ANOVA with $p_{value}<<0.0005$)

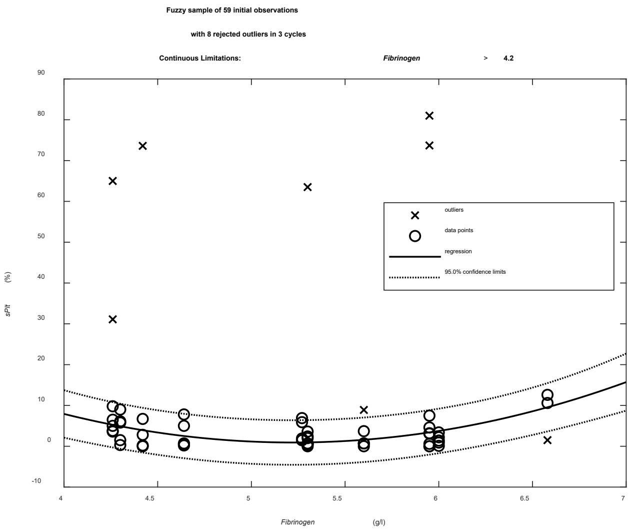

Fig. 5. Regression model for practical example 1 in Section 6.2.1 with 59 initial observations in a fuzzy sample, with a total of 8 utliers rejected in 3 cycles

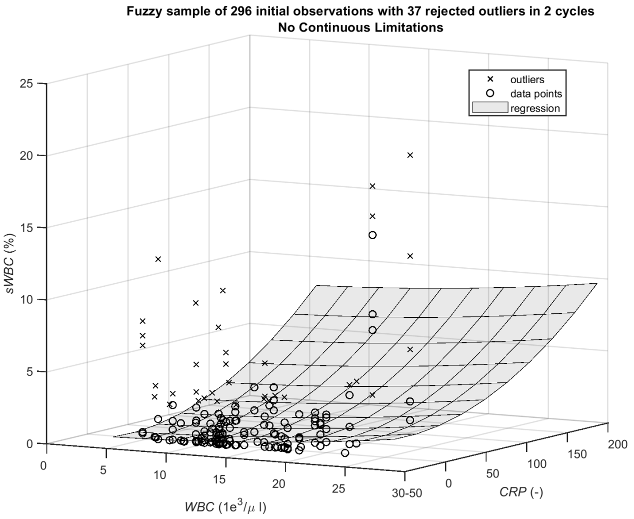

Fig. 6. Regression model for practical example 2 in Section 6.2.2 with 296 initial observations in a fuzzy sample, with a total of 37 utliers rejected in 2 cycles