Australian Journal of Crop Science AJCS

AJCS 11(04):459-463 (2017) ISSN:1835-2707
doi: 10.21475/ajcs.17.11.04.356

Software for the detection of outliers and influential points based on the HAT method

José Reinaldo da Silva Cabral de Moraes*, Glauco de Souza Rolim, Lucas Eduardo de Oliveira Aparecido

UNESP – São Paulo State University, Department of Exact Sciences, 14884-900, Jaboticabal, SP, Brazil

Abstract

We developed software in Visual Basic for application in Microsoft Excel that identifies outliers (OUTs) and influential datapoints (IPs) of scattered data using the HAT method (Hoaglin and Welsch). OUTs are commonly identified visually, which is susceptible to errors. The identification of IPs is not trivial, and using statistical tests is necessary. HAT is the most common statistical method to select OUTs and IPs in regression analyses and identifies four groups of data: 1) data within the standard range of variability, 2) OUTs, 3) IPs, and 4) both OUTs and IPs (OUT+IPs). The decision to remove or not remove data from the database depends on the researcher, and the HAT method helps to make these decisions. The removal of an OUT usually improves the accuracy of models. The removal of IPs, however, may or may not improve the accuracy. A small hypothetical data set of rainfall from automatic and conventional rain gauges was used to extensively test the software. The amount of data that can be used in the software is limited by the number of lines of the Excel spreadsheet (65518). The first step in identifying OUTs and IPs is to analyse all the data, which produced an R^2 for the raw data in our example of 0.11, indicating weak relationships between the variables. The HAT test identified two OUTs, three IPs, and one OUT+IP in the data. If all OUTs were removed, R^2 would increase to 0.19. If the OUT+IP was removed, R^2 would increase to 0.86. If all IPs were also removed, R^2 would decrease to 0.45. The software is free and can be requested by email from reinaldojmoraes@gmail.com.

Received 3 Oct 2016; Revised 19 Jan 2017; Accepted 1 March 2017.
Keywords: regression analysis, accuracy, model, dispersion.
Abbreviations: OUT_outlier; IP_influential datapoint; VBA_Visual Basic for Applications; EP_ standard error.

Introduction

Outliers (OUTs) are observations with a distribution differing from most of the other observations, displacing the sample mean and producing deceptive estimates (Raña et al., 2015). Gujarati and Porter (2009) defined an OUT as an observation with a large distance, or residual, between observed and estimated values. Influential datapoints (IPs) are those that remain within the variability of the points but have important weights in the slope of the fitted model (Genton and Gazen, 2010). Identifying OUTs and IPs in a database is very important in regression analysis. The removal of selected points in scattered graphs is one way to increase the accuracy of the adjusted models. Determining whether or not these datapoints should remain in the analysis, however, is not only a statistical option. This option should consider the validity of the points and if they can lead to new considerations and reformulations of the original model. Barnett and Lewis (1993) reported that OUTs and IPs can arise in correct measurements during data sampling, expressing the natural variability of the database. These datapoints can also appear due to human or sensor errors, such as incorrect data entry and failed execution. These error points, if identifiable, should be removed from the analysis. Some methods can select OUTs in regression analysis (Aucremanne et al., 2004; Kimber, 1990; Williams et al., 2015, Bedrick, 2001). Cook´s D statistic is a traditional method for identifying IPs (Cook, 1977). The calculation of the HAT matrix, also called the projection matrix, quantifies the influence of each value observed for each estimated value identifying OUTs and IPs. This method, proposed by Hoaglin and Welsch (1978), is more important in regression analysis than other methods, because it easily detects OUTs and IPs in the same analysis (Dufrenois and Noyer, 2013). The method, however, is restricted to linear relationships or performance analyses, i.e. the comparison of estimated and observed values in the same measurement units (Hoaglin and Welsch, 1978; Belsley et al., 1980). The new programming languages allow the use and development of computational programmes capable of performing analyses and applying concepts that were previously unfeasible due to the lack of processing power. Numerous discussions and studies in the search for new alternatives for analysing experimental data in temporal series have thus ensued. Many jobs using VBA in agricultural studies are found due to the ease of statistical analysis. Duvergel and Miranda (2014) have developed software in Microsoft (MS) Excel's Visual Basic environment, allowing multiple-proportion comparisons using the Wald method. This method is described in most basic statistical texts and is based on the normal asymptotic distribution of the difference between sample proportions. The software provides a fast and efficient way to perform a variety of treatments based on their contrasts, with precise estimates and interpreting existing errors. Stolf et al. (2014) developed a fast VBA programming-language tool that facilitated the computation of soil-resistance data and was capable of automatically composing resistance tables and graphs as input to the observed data. A routine capable of separating the effects of comparison between different experiments was thus created,

and the software simulated the objective proposed by the authors well.

We developed software to identify OUTs and IPs in a database using the HAT test in MS Excel. This environment was chosen to facilitate the input of data and the interpretation of the results.

Results and Discussion

Data input

After inserting the data in Table 1 into cells 'B18' and 'C18', the user must press the "Calculate HAT" button, which executes the algorithm developed in the VBA language. The results appear automatically in the broken areas and in the graphs. The final result for each pair of data is shown in column 'Q' if the answer is 'IP', indicating that the data pairs of lines 22, 30, 41, and 52 are IPs. The answers 'OUT' for outliers (lines 24 and 32) and 'OUT+IP' for both OUTs and IPs (line 49) can be checked in the same way.

Data analysis

Using Hc=0.0889 in the analysis, the HAT method identified two points as OUTs, three points as IPs, and one point as an OUT+IP. The other pairs of data remained within an appropriate range of dispersion (Fig 3). X and Y were not correlated in this initial situation, because the coefficient of determination (R^2) was low at 0.1101 (Fig 4.A).

If we considered all OUTs and OUT+IPs as casual measurement errors, we could remove these datapoints from the analysis. The removal of all OUTs increased R^2 to 0.1964, which was still considered very low (Fig 4.B). If the OUT+IP point was also removed, however, R^2 increased to 0.8583, suggesting that this option may be best for selecting the data in this analysis (Table 1). In this example, the model to calibrate the rain gauge based on the regression equation was Y=0.8971X + 111.19 (Fig 1).

This alternative of extracting data to improve the performance of the models was also applied by Menjoge and Welsch (2010), who proposed a method of simultaneous selection of OUT detection in linear regression models, improving the performance of their models. McCann and Welsch (2007) and Kim et al. (2008) constructed extensions of response variables, which consisted of a new algorithm for the selection of regression models, to identify OUTs in subsets of sample data, which improved the performance of the regression models. Shotwell and Slate (2011), however, proposed the detection of OUTs using the Dirichlet Process Mixture (DPM) method for clustering. The outlier procedure was run on a per-cluster basis from the estimated partitions from the DPM method, where each cluster was then subdivided into several non-OUT clusters and several OUT (or singleton) clusters.

The researcher can also choose to remove all or some of the IPs. IPs indicate values that have weight in the regression analysis, and their removal may or may not improve the adjustment (Genton and Gazen, 2010). The removal of all IPs in our analysis decreased R^2 to 0.4495 (Fig 4.D). Repeating the HAT method to identify new OUTs and IPs by pressing the 'Calculate HAT' button is necessary after the removal of any datapoint.

Nurunnabi et al. (2011) have shown that IPs can remain hidden in the presence of multiple OUTs, and their initial identification is not easy, because they are masked by the OUTs (Nurunnabi, Nasser and Imon, 2016). The simultaneous identification of OUTs, IPs, and OUT+IPs is thus necessary, because the presence of unusual values in a data set may impede the identification of the real influence of these observations (Peña and Yohai, 1995). The great advantage of using HAT, compared to other methods, is the joint identification of OUTs, IPs, and OUT+IPs.

Billor et al. (2000), Habshah et al. (2009), Nurunnabi, Imon, and Nasser (2011), and Nurunnabi, Imon, and Nasser (2016) used various methods of IP identification and observed that IPs as undesirable values should be characterised with caution, because the removal of these values may worsen the analysed adjustment.

Regression analysis and curve fitting are necessary agricultural tools for model development, so the use of the HAT test in the VBA environment of MS Excel is a robust methodology to jointly identify OUTs, IPs, and OUT+IPs in sample data, ensuring the application of any measure of evaluation.

Materials and Methods

Software development and statistical methods

The software was developed in the programming language Visual Basic for Applications (VBA) in Microsoft Excel (Fig 1), and the HAT method was codified as proposed by Hoaglin and Welsch (1978). The software allows the user to input up to 65 518 data pairs, consisting of an observed independent variable (Xobs) and an observed dependent variable (Yobs), into cells 'B18' and 'C18', respectively.

The programme calculates the estimated Y (Yest) by simple linear regression, sum of squared error (SSE), mean observed X (Mean Xobs), mean estimated Y (Mean Yest), total number of datapoints (N), mean square error (MSE, equation 1), value of the HAT scale (H) (equation 2), critical value of the HAT scale (Hc) (equation 3), linear and angular coefficients, and precision of the model by the coefficient of determination (R^2). These assessments are important and necessary for evaluating the Standard Deviation (SD, equation 4).

$$QME = \frac{\sum(Yobs-Yest)^2}{n-k-1} \quad (1)$$

where k is the number of independent variables in the original model, fixed in 1.

$$H = \frac{1}{n} + \frac{(Xobs-\bar{X})^2}{(n-1)\times S^2_{Xobs}} \quad (2)$$

$$hc = \frac{2\times(k+1)}{n} \quad (3)$$

$$EP = \frac{(Y-Yest)}{\sqrt{QME}} \quad (4)$$

Outlier Identification

Originally, the HAT method defined that any point in the scatter plot SE × H positioned above or below 3 (standard deviations) should be considered as an OUT (Fig 2). This limit can be modified in the cell 'D12'. Any point larger than Hc should be considered as an IP, and values larger than Hc, above 3, and less than -3, are considered OUT+IPs. Points within the areas limited by 3 and -3 and points smaller than Hc are consequently considered to be within the standard range of data (SR) (Fig 2).

Table 1. Hypothetical sample data of precipitation from two rain gauges, (X) Automatic and (Y) Conventional. The symbols OUT= Outlier, IP= Influential point, (OUT+IP) = outlier and influential points together, are the results of the analysis.

X	Y		X	Y		X	Y	
102.2	112		84.2	81		45	51	
200	203		90	100		1.2	1.200	OUT+IP
150	206		87.3	111		36	36	
132.7	144		44	115		35	26	
440	440	IIP	69.2	110		500	400	IP
103.4	129		85	91		50	73	
180	1.350	OUT	50	68		69.1	112	
140.8	300		95	96		82.4	114	
153.9	200		590	750	IP	98.2	199	
213.6	153		100	105		91.4	106	
108	141		38.7	90		97.6	116	
115.8	194		68	106		110.1	85	
0	0	IP	70	71		95	104	
110	127		75	62		181.9	108	
130	1.300	OUT	65	75		74.6	141	

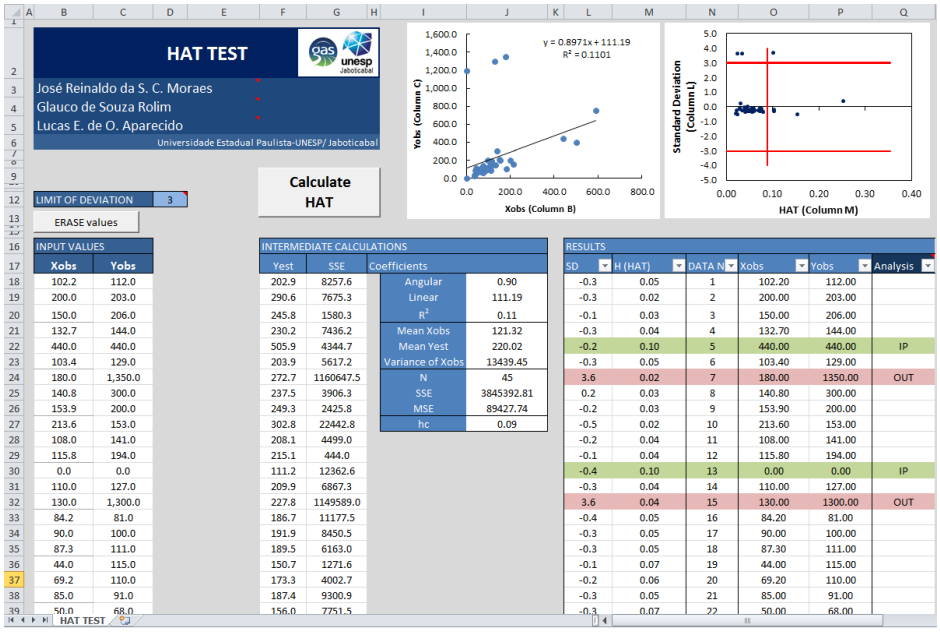

Fig 1. Main page of the HAT software implemented in MS-EXCEL environment.

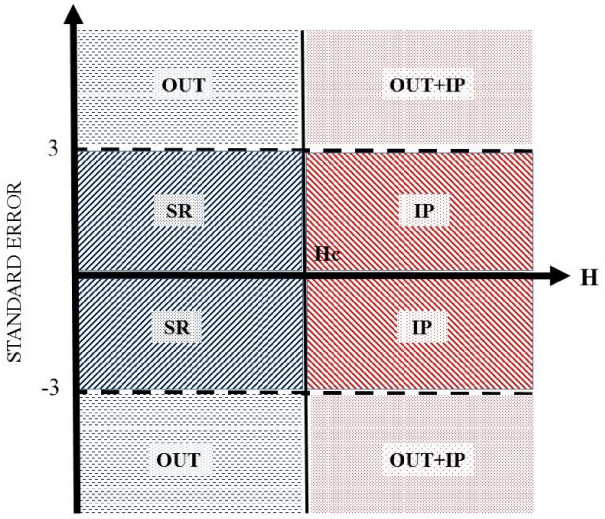

Fig 2. Areas of the HAT test to identify. (OUT) outlier, (IP) influential point, (SR) standard range of data.

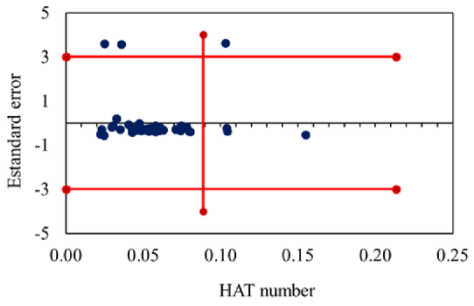

Fig 3. Points within the pattern, influential points, outlier in outlier and influential point observed in the dispersion of data from different weather stations.

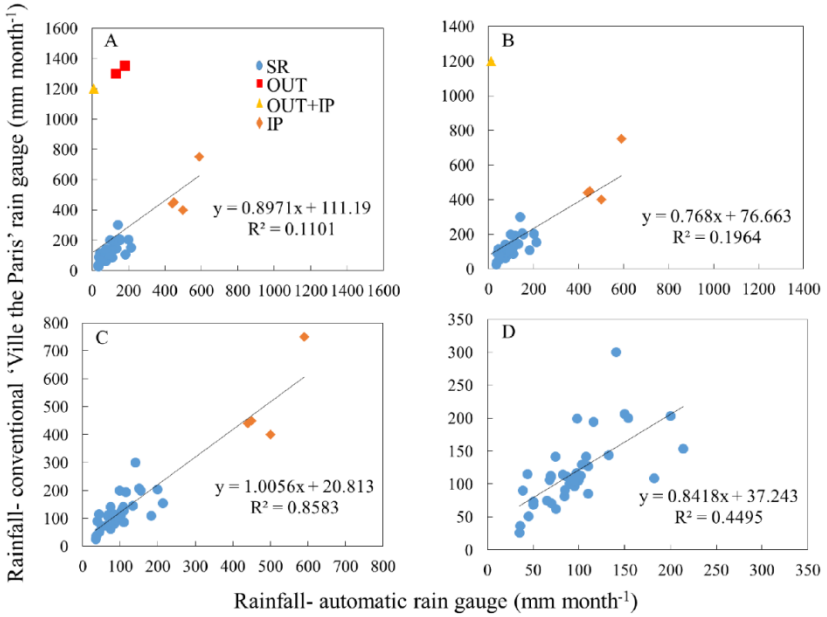

Fig 4. Rainfall from conventional 'Ville the Paris' and automatic rain gauges. A) original data points (all); B) regression analysis without OUT; C) regression analysis without OUT and (OUT+IP); D) regression analysis without OUT, IP and (OUT+IP).Legend: SR= standard region, OUT= outliers, IP= influential points, (OUT+IP) outliers and influential points together.

Software usage

Using the software: the user must first input the data in columns B (Xobs) and C (Yobs) from line 5 of the spreadsheet and then activate the "Calculate HAT" button, so all calculations and graphic operations will be generated automatically, identifying the points, OUTs, IPs, and OUT+IPs. The points not labelled are the SR along the database. The 'Delete' button prepares the spreadsheet to receive another database to be evaluated. We analysed hypothetical rainfall data to illustrate the use of the software (Table 1). We wished to know if an automatic rain gauge (X) was similar to a conventional 'Ville de Paris' rain gauge (Y), and if so, which calibration model should be used.

Conclusion

The software developed based on the HAT method in MS EXCEL is a tool for easily evaluating OUT, IP, and OUT+IP calculations. This environment simplifies analysis and is intuitive for users unfamiliar with more advanced statistical techniques. The HAT test can identify OUTs, IPs, and OUT+IPs, aiding the development of more robust regression models.

References

Aucremanne L, Brys G, Hubert M, Rousseeuw PJ, Struyf A (2004) A study of belgian inflation, relative prices and nominal rigidities using new robust measures of skewness and tail weight. Stat Ind Technol. 1:13-25.

Barnett V, Lewis T (1993) Outliers in statistics. New York: Wiley. 3rd edn.68.

Bedrick EJ (2001) An efficient score test for comparing several measuring devices. J Qual Technol. 20:96-103.

Belsley DA, Kuh E, Welsch R (1980) Regression diagnostics. Identifying influential data and sources of collinearity. 1rd edn. New York: Wiley.

Billor N, Hadi AS, Velleman F (2000) Blocked adaptive computationally-efficient outlier nominator. Comput Stat Data Anal. 34:279–298.

Cook RD (1977) Detection of Influential Observations in Linear Regression. Technom Am Stat Assoc.19: 15–18.

Dufrenois F, Noyer JC (2013) Formulating robust linear regression estimation as a one-class LDA criterion: discriminative hat matrix. IEEE T Neural Networ. 24:262-73.

Duvergel YC, Miranda I (2014) COMPAPROP: A system for multiple proportion comparisons. Rev. Protección Veg. 29: 231-234.

Genton MG, Gazen AR (2010) Visualizing Influential Observations in Dependent Data. J Comput Graph Stat. 19: 808–825.

Gujarati DN, Porter DC (2009) Basic Econometrics, 5rd edn. McGraw-Hill Publishing Companies, Inc. USA, New York.

Habshah M, Norazan R, Imon AHMR (2009) The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. J Appl Statist, 36:507–520.

Hoaglin D, Welsch R (1978) The hat matrix in regression and ANOVA. Am Stat, 32: 17-22.

Kim S, Park SH, Krzanowski WJ (2008) Simultaneous variables selection and outlier identification in linear regression using the mean-shift outlier model . J Appl Stat 35:283-291.

Kimber AC (1990) Exploratory data analysis for possibly censored data from skewed distributions. Ann Appl Stat. 39: 21–30.

McCann L, Welsch RE (2007) Robust variable selection using least angle regression and elemental set sampling. Comput Stat Data An, 52:249–257.

Menjogea RS, Welschb RE (2010) A diagnostic method for simultaneous feature selection and outlier identification in linear regression. Comput Stat Data An, 54:3181–3193.

Nurunnabi AAM , Imon AHMR, Nasser M (2011) A Diagnostic Measure for Influential Observations in Linear Regression. Commun Stat A-Theor, 40:1169-1183.

Nurunnabi AAM, Nasser M, Imon AHMR (2016) Identification and classification of multiple outliers, high leverage points and influential observations in linear regression. J Appl Stat, 43:509-525.

Pena D, Yohai VJ (1995) The detection of influential subsets in linear regression byusing an influence matrix. J. Roy. Statist Soc Series B, 57:145–156.

Ranâ P, Aneiros G, Vilar JM (2015) Detection of outliers in functional time series. Environmetrics.26:178–191.

Shotwell M, Slate E (2011) Bayesian Outlier Detection with Dirichlet Process Mixtures. Bayesian Analysis, 6: 665–690.

Stolf R, Murakami JH, Brugnaro C, Silva LG, Silva LCF, Margarido LACM (2014) Penetrômetro de impacto Stolf – programa computacional de dados em Excel-VBA. Rev Bras Cienc Solo, 38:774-782.

Williams JD, Birch JB, Abdel-Salamc ASG (2015) Outlier robust nonlinear mixed model estimation. Stat Med. 34:1304–1316.