bib REVISTA BRASILEIRA DE INFORMAÇÃO BIBLIOGRÁFICA EM CIÊNCIAS SOCIAIS

ANPOCS

Balanços Bibliográficos

Living with outliers: How to detect extreme observations in data analysis

Dalson Figueiredo Filho[I]
https://orcid.org/0000-0001-6982-2262

Lucas Silva[II]
https://orcid.org/0000-0002-5013-6278

Antônio Pires[III]
https://orcid.org/0000-0001-5468-3407

Caio Malaquias[IV]
https://orcid.org/0000-0003-3189-2024

DOI: 10.17666/bib9906/2023

Submetido em: 11/02/2023
Ressubmetido em: 09/09/2023
Aceito em: 12/09/2023

1. INTRODUCTION[1]

Outliers can cause a range of problems in data analysis, including biased and inefficient estimates (Seo, 2006), incorrect coefficient signs (Fox, 1972; Verardi & Croux, 2009), and ineffective visualization (Atkinson & Mulira, 1993). However, available evidence suggests that scholars seldom report checking for extreme observations of any sort (Iglewicz & Banerjee, 2001; Weber, 2010). According to Osborne et al. (2001), the practice of verifying statistical testing assumptions - such as checking for outliers - is infrequent, occurring only 8% of the time. This lack of verification is a concerning issue as it can have dangerous consequences. Bollen (1988) reported that as few as six outliers in a sample of 100 observations could lead to a false negative result. More recently, Imai et al. (2016) showed that the primary findings from De la O (2013) vanish after removing two outliers from the original sample.

[1] Replication materials are available on: <https://osf.io/cbwjv/>.

[I] Universidade Federal de Pernambuco (UFPE). E-mail: dalson.figueiredofo@ufpe.br

[II] Universidade Estadual de Ciências da Saúde de Alagoas (UNCISAL). E-mail: lucas.eosilva@academico.uncisal.edu.br

[III] Universidade Federal de Pernambuco (UFPE). E-mail: antonio.hps26@gmail.com

[IV] Universidade Federal de Pernambuco (UFPE). E-mail: caio.malaquias@ufpe.br

Publicada em novembro de 2023 | pp. 1-24

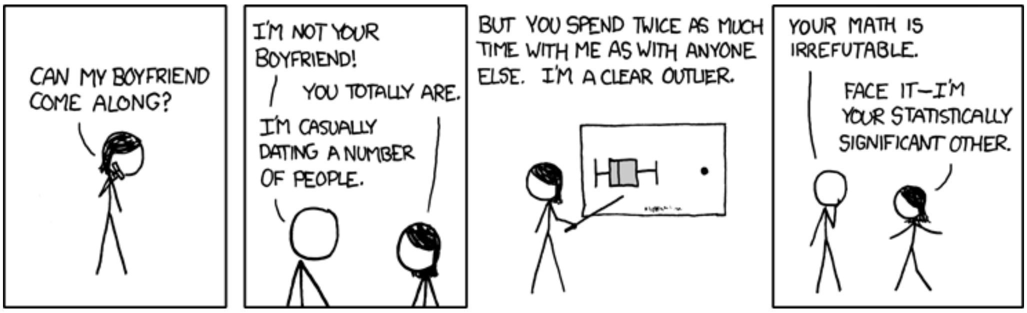
This paper outlines five statistical methods that can be used to identify outliers: (1) standardized scores; (2) interquartile range; (3) standardized residuals; (4) Cook's distance and (5) Mahalanobis distance. Our article offers four key advantages compared to other available resources. First, we focus on the comprehension of theoretical concepts rather than mathematical equations. Second, we employ examples that are familiar to social science researchers, making the content more engaging for them. Third, we make both the original data and R scripts readily accessible, allowing scholars to apply these methods to their own research. Fourth, accompanying the paper, we also offer a Q&A section that can help readers to grasp the core concepts.

The rest of the paper is structured as follows. Section 2 introduces the notion of outliers and their possible sources. Section 3 uses bibliometric data from Brazilian journals to explore how scholars in Political Science and International Relations scholars are handling outliers using bibliometric data from Brazilian journals. Section 4 discusses provides detailed explanation on the impact of outliers on statistical estimates. the outlier effect. Section 5 describes the step by step of five statistical techniques designed to detect extreme observations, and Section 6 concludes the paper.

2. OUTLIERS: What and Why?

An outlier is widely considered a unique instance, but what exactly makes it unique and to what degree? In this paper, we follow Hawkins' (1980) definition as it is the shared accepted notion of what an outlier is and highlights the role of underlying causal mechanisms in the data generation process (Bunge, 2016; Gerring, 2004; Mahoney, 2001). Figure 1 shows the outlier labeling rule, which involves using a boxplot to spot atypical cases (Hoaglin et al., 1986).

FIGURE 1 – *Outlier boyfriend*


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

Source: XKCD. Available on: https://xkcd.com/539

In this scenario, a graphical analysis indicates that a specific case stands out as significantly different from the other observations in the sample. Scholarly literature usually identifies univariate, bivariate, and multivariate outliers (Fox, 1972; Lewis & Barnett, 1994). Univariate outliers stand out when considering a single variable, and they are identified by looking for extremely low or high values. Bivariate outliers occur when two variables are combined and are typically identified using scatter plots. In a bivariate regression, an outlier is a case that has a highly unusual value in Y, given its value in X. Lastly, multivariate outliers represent cases that exhibit uncommon combinations of values across a group of variables.

From a technical perspective, an outlier has a minimal probability of being generated by the same statistical distribution that produced the other observations (Hawkins, 1980; Walfish, 2006). As such, it is crucial to understand the source of these exceptional cases. Chandola et al. (2007) propose four possible explanations for the presence of outliers: (1) malicious activity; (2) instrument malfunction; (3) abrupt changes in the environment; and (4) human error. Malicious activity typically refers to illegal actions that result in patterns that significantly diverge from theoretical expectations. For instance, if a credit card transaction appears to be highly discrepant based on a cardholder's typical spending habits, a credit card operator may suspect suspicious behavior.

Instrument error is more prevalent in the Natural Sciences compared to the Humanities due to the fact that measurements in these fields often rely on specialized devices. For example, a physicist measuring radiation levels might use a Geiger-Muller counter, while a chemist studying water evaporation may use a thermometer. Poorly calibrated or inadequate instruments can lead to unreliable and invalid measurements (Blalock, 1979; Walfish, 2006; Zeller et al., 1980). Depending on the situation, the instrument can yield significantly different readings from what would be obtained with proper calibration. In the Social Sciences, the questionnaire used in survey research is an example of an instrument that can negatively impact the validity and reliability of data if it is poorly designed[2].

Outliers caused by sudden changes in the environment are commonly seen in natural disasters. For example, heavy rainfall and flooding on one hand, and prolonged drought and subsequent water shortages on the other, can affect the consistency of the estimates. Consider a public safety study that tracks the daily number of homicides. Typically, the death toll is higher on weekends and holidays when the risk of being killed is elevated. However, if there is a sudden increase in rainfall, it can lead to fewer people being on the streets, resulting in a drastic decrease in the number of homicides.

The fourth and final explanation of outliers is human error (Belsley et al., 2005). This problem particularly relevant in the Social Sciences, where manual data collection and coding are still commonly used (Hopkins & King, 2010). Manual methods are often slower, more time-consuming, and less reliable than automated data extraction and tabulation procedures. Even a single mistake in a spreadsheet can result in the introduction of extreme cases into the sample, leading to incorrect conclusions (Hawkins, 1980; Walfish, 2006).

3. WHERE ARE OUTLIERS IN OUR FIELD?

To examine the use of outlier detection and treatment methods in the field of Political Science and International Relations, we employed a content analysis of 2,292 academic journals available on the Sucupira Platform under the Political Science and International Relations evaluation area (2013 - 2016). Then, we performed a search on SciELO using keywords related to outliers[3], which returned

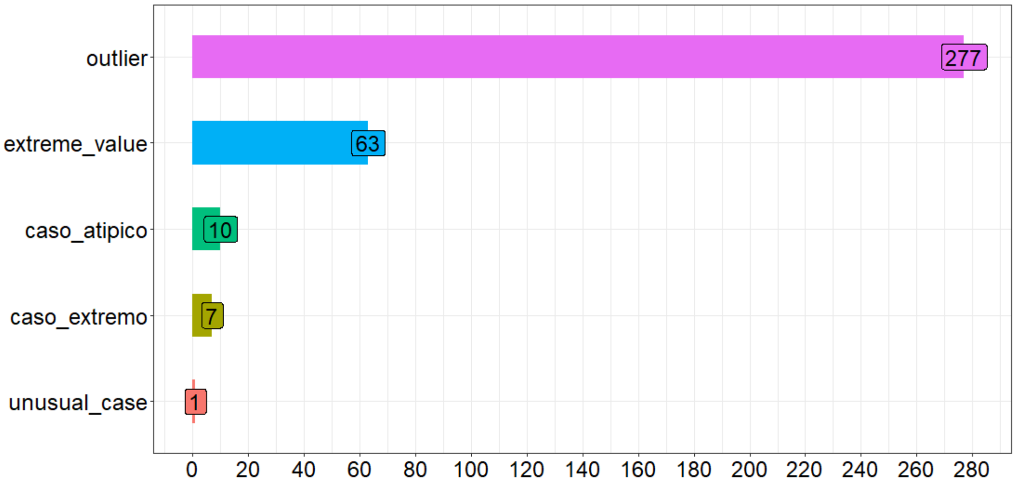
[2] Nunally and Bernstein (1994) define validity as the degree to which an instrument measures exactly what it is supposed to measure.

[3] The search field was filled in as follows: (outlier) OR (outliers) OR ("caso atipico") OR ("casos atipicos") OR ("unusual value") OR ("unusual values") OR ("caso extremo") OR ("casos extremos") OR ("extreme value") OR ("extreme values") OR ("observacao aberrante") OR ("observacoes aberrantes") OR ("aberrant observation") OR ("aberrant observations")

294 papers. Out of the 294 articles found on SciELO, 40 were published in the identified journals.

Using the *rvest* and *quanteda* R packages, we extracted the full texts of the 40 articles from their URLs and conducted content analysis to grasp the context of usage of the key terms. Initially, we analyzed the frequency of the search keywords in the articles. We found that mentions of "outlier" and its grammatical variations were predominant compared to expressions that may indirectly refer to it, such as "extreme value" or "unusual case" Thus, from this point forward, we focused our analysis solely on mentions of the term "outlier." Figure 2 displays the frequency of the keywords.

FIGURE 2 – Keywords frequency in 40 papers in the target fields



Source: The authors

In our keyword analysis, we neither translated nor applied any specific procedure to compare texts from different languages. It is also important to note that the term "outlier" and its variations appear in 15 of the 40 analyzed articles, despite being the most frequently keyword searched for on SciELO. Finally, the term ranks as the sixth most frequent among the words present in the 40 articles. This indicates that while mentions of outliers are numerous, they are only found in less than half of the articles.

To understand its context of the use of "outlier", we can ask the following question: What are the words that commonly appear alongside this term? Figure 3 displays a co-occurrence network that visualizes the term "outlier", the 15 words that co-occur most frequently with it, and the main words that co-occur in relation to these 15 words.

We can observe that references to outliers are made in the context of outlier detection procedures, as was our intention when selecting the keywords to search in SciELO. To ensure that the terms in Figure 3 were indeed referring to outlier detection and treatment methods, we conducted a qualitative categorization in five ideal types.

a) Qualitative Classification: mentions that highlight the relevance of a case or event.

b) Case Study: articles that focus on a single case study that is considered atypical or extreme, without this being necessarily relying on statistical analysis.

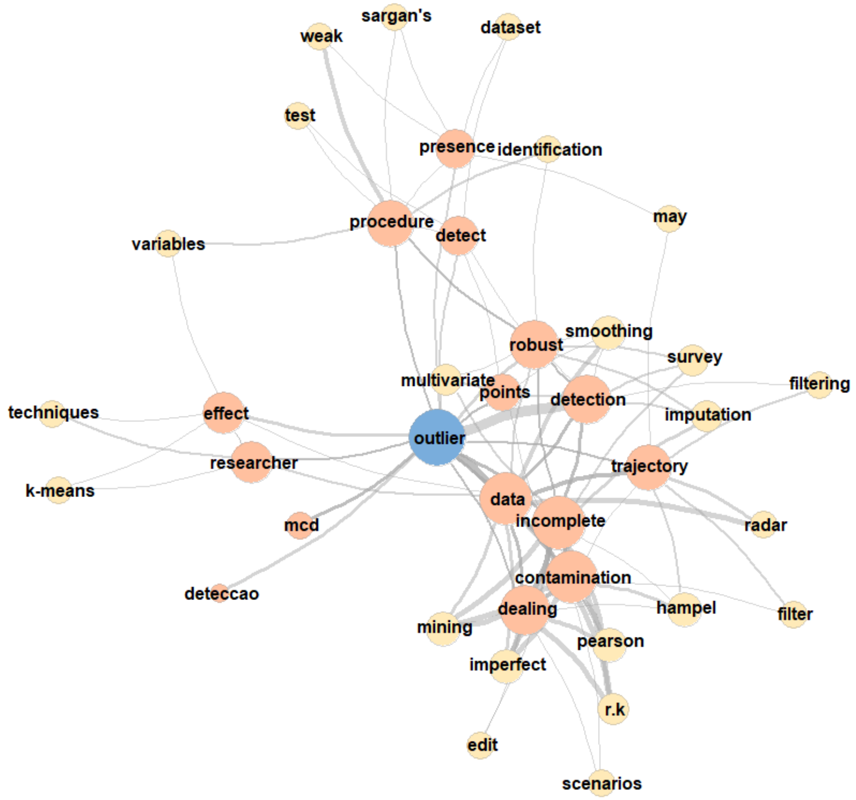
c) Outlier as a Theme: studies that analyze outliers from a methodological perspective.

d) Outliers in Results: articles that use outlier detection and treatment methods to enhance the robustness of their analysis.

e) Technical Term: articles in which the key term is part of a technical term, theory or concept, such as "Generalized Extreme Value", "Extreme Value Theory", "Preference Outliers", "Extreme Value Distribution".

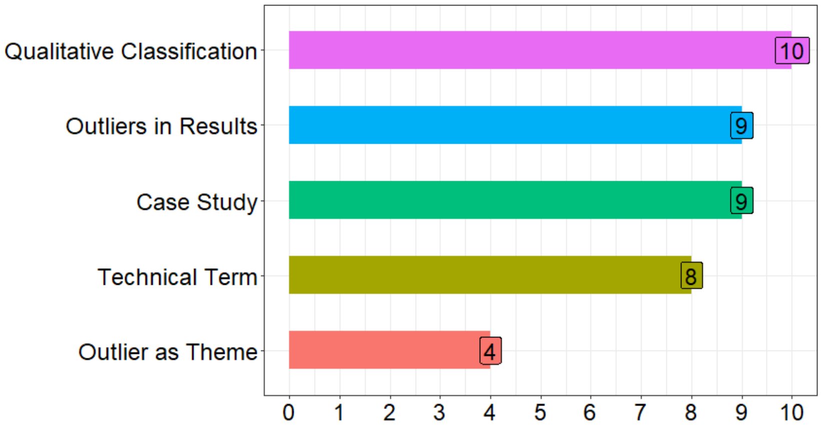
Figure 4 shows the distribution of articles in each category.

FIGURE 3 – Network of terms that co-occur with ouliers



Source: The authors

FIGURE 4 – Types of Outlier Mentions



Source: The authors

The most common use of the term outlier is for Qualitative Classification (10), followed closely by its mention in Results (9) and Case Study (9). After gaining understanding of the concept of outlier and how Political Science and International Relations articles in Brazil cite extreme cases, our next step was to describe how to detect the impact outliers in data analysis.

4. THE IMPACT OF OUTLIERS IN DATA ANALYSIS

As illustrated in Section 3, outliers occur in our field and can be categorized into various types. As such, it is important to understand the impact of outliers in data analysis. Osborne and Overbay (2019) argue that atypical cases increase sample variance and reduce the power of statistical tests. Outliers can also lead to assumption violations, affecting the chance of making type I and type II errors. Finally, extreme cases may even lead to wrong coefficients signs and biased estimates[4]. To grasp how outliers can affect statistical estimates, let's take a look at the distribution of the Human Development Index (HDI) across Brazilian states.

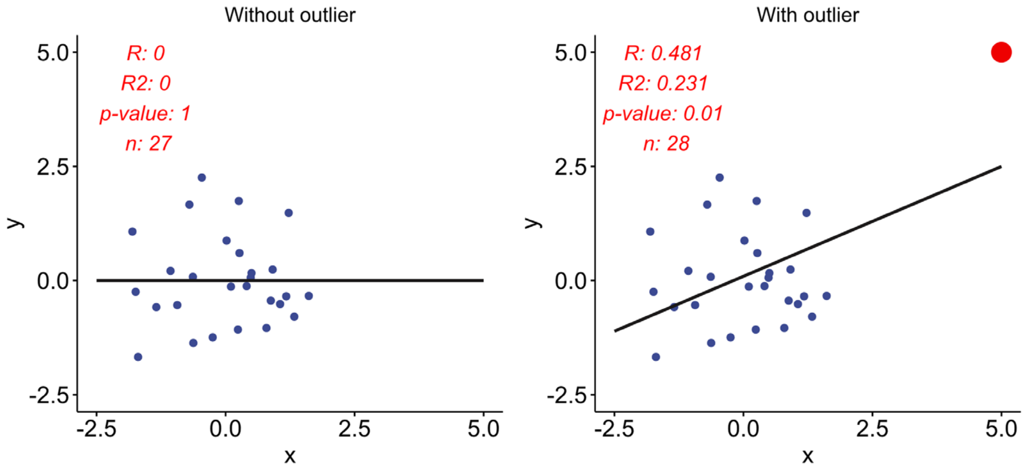
TABLE 1 – **HDI in Brazil (2010)**

	N	MEAN	STANDARD DEVIATION
With outlier	27	.7045	.049
Without outlier	26	.6999	.044

Source: The authors

Considering all cases, the HDI average is .7045, which would include Brazil among developed countries (.700 to .899). Without the Federal District, the mean lowers to .6999 which would leave Brazil in the middle category (.55 to .699). Additionally, extreme observations can lead to wrong coefficient signs. Figure 5 shows how outliers may affect correlational analyses.

FIGURE 5 – Comparing correlations with and without outliers



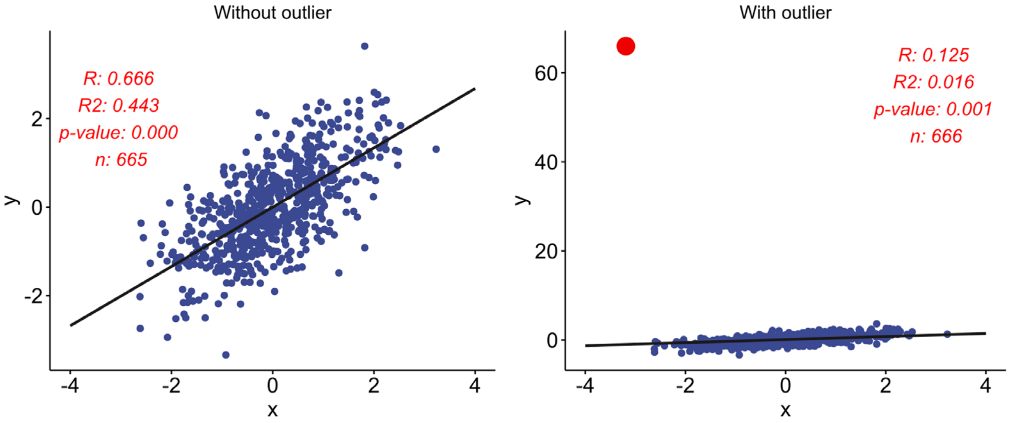
Source: The authors

[4] Type I error is the incorrect rejection of the null hypothesis, that is, a false positive result. Type II error is an inability to reject a false null hypothesis, meaning a false negative result.

In Figure 5, we see that both variables are orthogonal (no linear correlation). With the inclusion of a unique extreme case (5, 5), we reach a moderate positive correlation (r = .481) which has statistical significance at conventional levels. In a bivariate regression model, we would conclude that the variation in the independent variable explains 23.1% of the variance of the dependent variable. In short, these results would lead us to incorrectly reject the null hypothesis (type I error).

Another problem caused by outliers is meaningless graphical visualization. Figure 6 shows an example.

FIGURE 6 – Visualization problems caused by outliers



Source: The authors

In Figure 6, the outlier is so extreme that it hides the true relationship between X and Y. The observed correlation is positive, moderate (r = .666), and statistically significant (p-value<.01). With the atypical case, the relationship remains positive, but we would reach a different conclusion about the strength of the association between the variables. These examples show how outliers can adversely affect statistical estimates and lead to wrong conclusions.

5. HOW TO DETECT OUTLIERS?

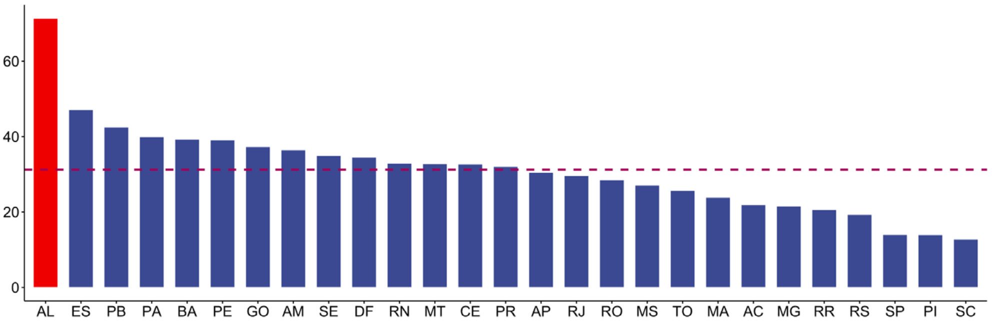
Having discussed the origins and effects of the outliers, the next step is to explain how to detect them. Cateni et al. (2008) suggest different identification methods, including informal tests, graphical analysis, hypothesis testing, distance measures, cluster analysis, and artificial intelligence (e.g., neural networks, fuzzy inference system, support vector machine). In this section, so that scholars in the field of political science and international relations can learn how to detect outliers, we will now present the nuts and bolts of the following statistical techniques: (1) standardized scores; (2) interquartile range; (3) standardized residuals; (4) Cook's distance, and (5) Mahalanobis distance.

a) Standardized scores

In addition to examining the measures of format (kurtosis and skewness) and variability (variance and standard deviation), a straightforward procedure to identify cases that are too distant from the central tendency of the distribution is to analyze the standardized scores. According to Iglewicz and Hoaglin (1993, p. 10), "this approach has an attractive simplicity, and most statistical software packages and even simple calculators make it easy to obtain the Z-scores". To estimate a standardized score, we must subtract the value of each observation from the mean and divide the result by the standard deviation. The new distribution has zero mean and the distance between the cases is now given in standard deviation units. The higher the Z score, in absolute values, the higher the distance between a particular observation and the general mean.

Scholarly literature defines outlier cases as those with values above 3 and lower than -3 standard deviation units (Atkinson & Mulira, 1993; Lewis & Barnett, 1994; Walfish, 2006). Under the standard distribution assumption, the reasoning is that 68%, 95% e 99% of the cases should lie between, one, two, and three standard deviations from the mean, respectively. In particular, less than 1% of the cases would show values as extreme as four standard deviations from the mean. To illustrate how standardized scores may be used to detect outliers, Figure 7 shows the variation of homicide rates across Brazilian states in 2011.

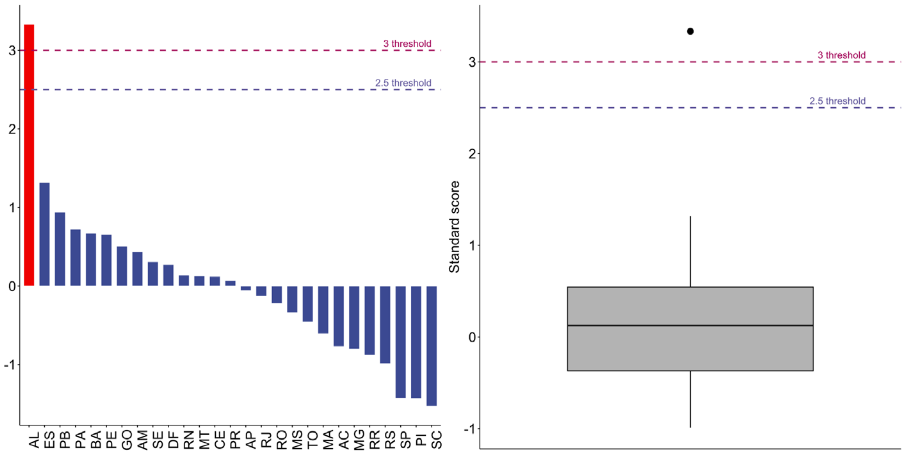
FIGURE 7 – Homicide rate in Brazil (2011)



Source: The authors

Assuming no measurement error and that the data collection procedures are the same, we observe that states of Alagoas (71.39), Espírito Santo (47.14), and Paraíba (42.57) show the highest rates of homicides per 100,000 habitants. For a scholar wishing to standardize the distribution, the first step is to subtract each value from the mean (31.25) and then divide the result by the sample's standard deviation (12). For example, in the state of Paraíba, the calculation would be as follows: (42.57 - 31.25)/12, producing a Z score of .94. Interpretation: Paraíba is .94 standard deviation above the mean. Figure 8 shows two different ways to graphically display standardized data to detect outliers.

FIGURE 8 – Standardized homicide rate



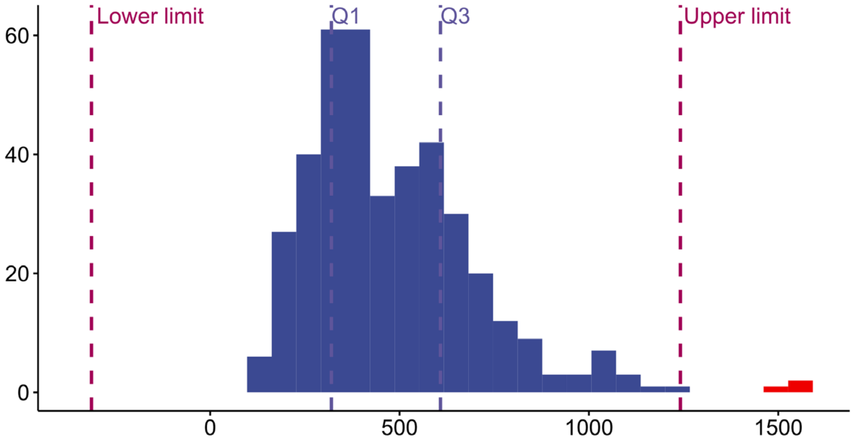
Source: The authors

In the bar graph, dotted lines represent the thresholds (3 and -3 standard deviations). Alagoas has a Z score of 3.4 which means that this case is more than three standard deviations above the mean. The boxplot also highlights Alagoas as an atypical observation in the sample of Brazilian states.

b) Interquartile range

The second procedure to identify atypical cases is uniquely suited to deal with univariate outliers in approximately normal distributions. The interquartile range informs the difference between the third and the first quartiles. To show how it works, we replicated data from the California Department of Education (*API 2000 dataset*). This dataset provides detailed aggregate information for 400 schools, including academic performance index, average class size, and proportion of pupils eligible for subsidized meals. Figure 9 shows the distribution of the number of students per school.

FIGURE 9 – Number of students per school

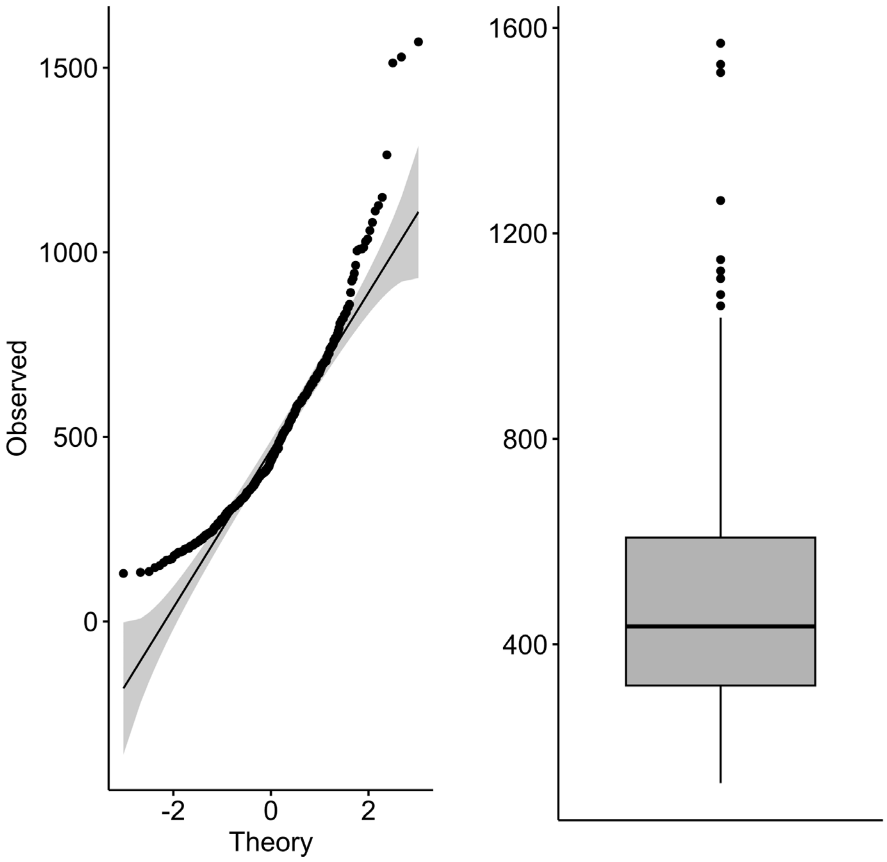


Source: The authors, based on *API 2000 dataset*

In Figure 9, the dotted blue lines represent the first (Q1 = 320) and the third quartiles (Q3 = 608). These thresholds are used to calculate the minimum and maximum limits which will be employed to detect unusual cases. Thus, values below the lower limit or above the upper limit are considered outliers. When applying this procedure to data from the California Department of Education, no outlier below the lower bound is found. At the upper limit, this method detected four atypical observations shown in red. However, before drawing substantive conclusions, it should be noted to what extent the variable of interest is normal. Figure 10 shows two normality diagnostics tests.

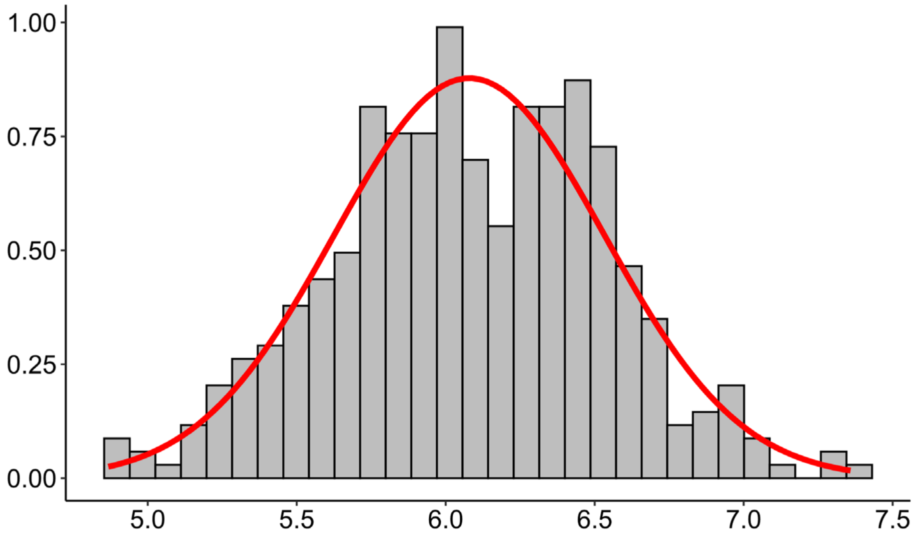
In Figure 10, the *Q-Q plot* compares the observed distribution with a normal theoretical assumption. The higher the similarity between the expected and the observed, the higher the evidence in favor of normality. As we can see, the variable is not normal (Kolmogorov-Smirnov Z = 1,941; p-value = 0,001). Therefore, before concluding which cases are extreme, researchers analyzing this data should first should first transform the original variable to reach normality (Davies & Gather, 1993). The fundamental rationale underlying the application of the logarithmic transformation lies in its ability to attenuate the scale variability of a variable. This transformation achieves this by taking the logarithm of the variable, effectively constraining the impact of outliers and extreme values. Consequently, this manipulation serves to normalize the data distribution, rendering it more closely aligned with a Gaussian distribution (Benoit, 2011). Taking these steps into account, Figure 11 shows the natural logarithm of the number of students.

FIGURE 10 – Normality tests



Source: The authors, based on *API 2000 dataset*

FIGURE 11 – Number of students after logarithmic transformation



Source: The authors, based on *API 2000 dataset*

As expected, Figure 11 shows that after logarithmic transformation, the variable reaches normality (p-value = 0,623). After applying interquartile range estimation, no outlier cases were identified. In short, when the focus is purely descriptive, we suggest using the original variable without transformation. However, when the focus is model building, we advise using the transformed variable once that the logarithmic transformation reduces the variance in the sample and lessens the impact of extreme cases on coefficient estimation.

c) Standardized residuals

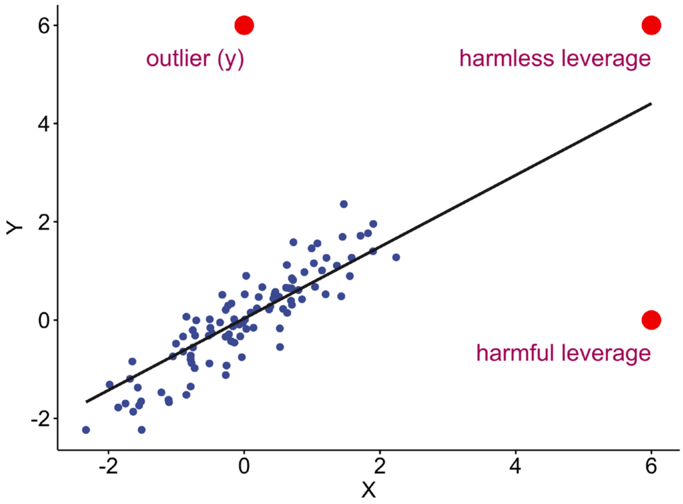
Different from both standardized scores and interquartile range, the standardized residuals technique is suitable to identify outliers in regression models (Belsley et al., 2005). Residuals (ε_i) represent the difference between observed values of the dependent variable (y_i) and the predicted values from the estimated model ($\hat{y}_i$). Residual analysis can be used to detect problems such as heteroskedasticity, autocorrelation, lack of linearity, among others.

Similarly, standardized residuals are the discrepancies between observed data points and their corresponding model-based predictions, expressed in terms of standard deviations. These residuals provide a standardized measure of the extent to which the model's predictions deviate from the actual data. By converting the raw residuals into a common scale of standard deviations, standardized residuals facilitate the identification of observations that exhibit substantial deviations from the model's expected behavior (Cook & Weisberg, 1982). In less technical parlance, standardized residuals are a way to measure how far off scholars' predictions are from the actual outcomes and they are adjusted to make it easier to compare across different situations. The use of this method is like looking at how much guesses are different from reality, but these differences are put into a common scale, which helps researchers see if some guesses are far off base while others are more on track.

It is important for researchers using these techniques to understand the conceptual difference between outlier, leverage, and influence points. Within the context of regression analysis, an outlier represents a data point with a large residual, signifying a substantial disparity between observed and predicted values. The notion of leverage pertains to extreme values in the independent variable. An observation with elevated leverage is prone to inducing biased parameter estimates. An influential observation, on the other hand, is characterized as a data point whose removal exerts substantial effects on statistical estimates. The magnitude of this influence is positively correlated with the extent of variance associated with that particular observation. Figure 12 shows the difference between outliers, leverage, and influence in a bivariate framework.

As illustrated in Figure 12, an outlier in the dependent variable will show an average value in X but an unexpected value in Y, distant from the regression line. As ordinary least squares (OLS) use the squares of the residuals, outliers can dramatically change estimates' magnitude and sign. In harmless leverage, the correlation between X and Y is only marginally affected. Observations close to regression line do not distort statistical estimates. Influential points, which are in the independent variable and far away from the regression line, will produce dramatic variations in the estimated coefficients (harmful leverage) (Fox, 1972).

FIGURE 12 – Outlier, leverage, and influence



Source: The authors

To illustrate the effect of outliers on statistical estimates, we reproduce data from Agresti and Finlay (1997) regarding crime in the United States[5]. The original dataset has 51 observations and the following information: state, crime, homicide, percentage of people living in the metropolitan area (pct metropolitan), percentage of white people (pct whites), percentage of people with undergraduate degrees (pct graduated), percentage of people living in poverty (pct poverty), and percentage of single parents (pct single parents). Table 2 displays the summary statistics of all variables.

TABLE 2 – **Descriptive statistics**

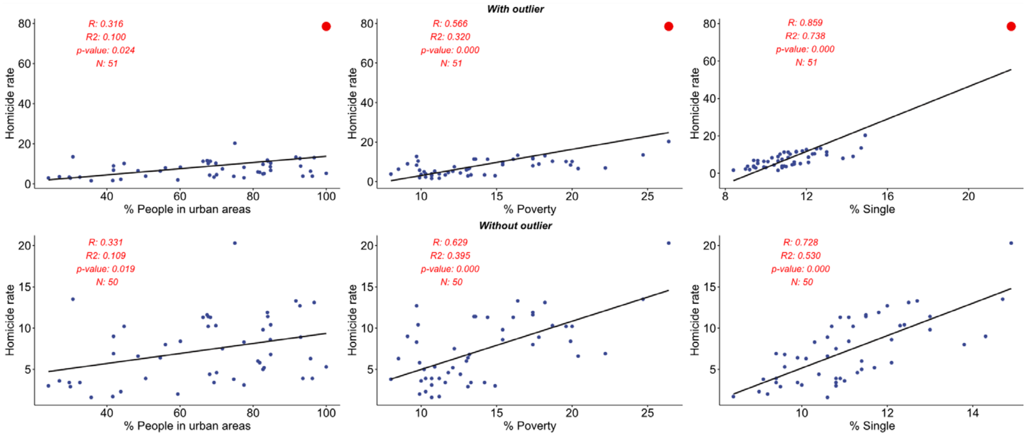
VARIABLE	$\underline{X}$	σ	MIN	MAX
crime	612.84	441.10	82	2,999
homicide	8.72	10.71	1.6	78.5
pct metropolitan	67.39	21.95	24	100
pct whites	84.11	13.5	31.8	98.5
pct graduates	76.22	5.59	64.3	86.6
pct poverty	14.25	4.58	8	26.4
pct single parents	11.32	2.12	8.4	22.1

Source: The authors, based on Agresti and Finlay (1997)

Using this data, a researcher might have the goal of estimating a linear regression model to explain the variation in homicide rates (y) from the percentage of people living in urban areas (pct metropolitan), the percentage of people in poverty (pct poverty) and the percentage of single parents (pct single parents). Figure 13 shows the bivariate correlations between homicide rates and these variables with and without the outlying observation.

[5] Original data available at: http://www.ats.ucla.edu/stat/stata/webbooks/reg/crime.

FIGURE 13 – Bivariate correlations with and without outliers



Source: The authors, based on Agresti and Finlay (1997)

The graphical analyses, as seen in Figure 13, suggest that the red dot (District of Columbia) is an abnormal case. To estimate how different a case is, a common procedure is run a correlation with and without the outlier. The higher the difference between the coefficients, the higher the detrimental statistical effects of a specific observation. Another typical step is to run the full model with and without the extreme case to determine what occurs with the regression's slopes and standard errors. Table 3 displays this information.

TABLE 3 **– OLS coefficients with and without outlier case**

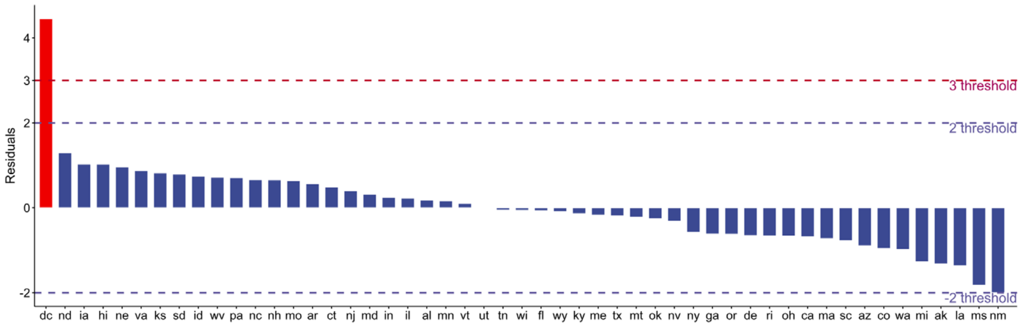
ESTIMATES	WITH	WITHOUT
Intercept	-43.24*** (-9.98)	-17.06*** (-7.43)
Percentage of people living in metropolitan area	.067 (1.83)	.061*** (4.31)
Poverty	.410* (2.02)	.444*** (5.69)
Percentage of single parents	3.672*** (8.08)	1.271*** (5.60)
R^2	.76	.75
N	51	50

***** p < .05, ** p < .01, *** p < .001

Source: The authors, based on Agresti and Finlay (1997)

As seen in Table 3, there is significant variation in the intercept between the two models. Moreover, the proportion of people living in metropolitan areas does not reach statistical significance when all cases are included. The effect of the proportion of single parents is three times higher for the full sample compared to its impact without the extreme case. Figure 14 shows the distribution of standardized residuals by state.

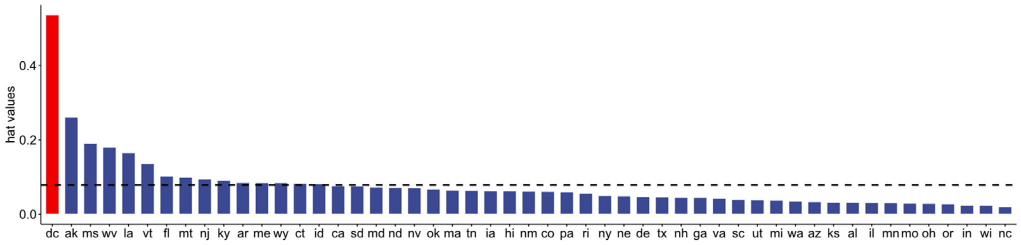
FIGURE 14 – Residuals by state



Source: The authors, based on Agresti and Finlay (1997)

Scholars should look carefully for residuals with absolute values above 2 and be extremely concerned about cases with residuals higher than 3 (Atkinson, 1994; Seo, 2006; Weber, 2010). In our working example, the District of Columbia has a standardized residual of 4,318. Figure 15 shows the leverage statistic by state.

FIGURE 15 – Leverage by state



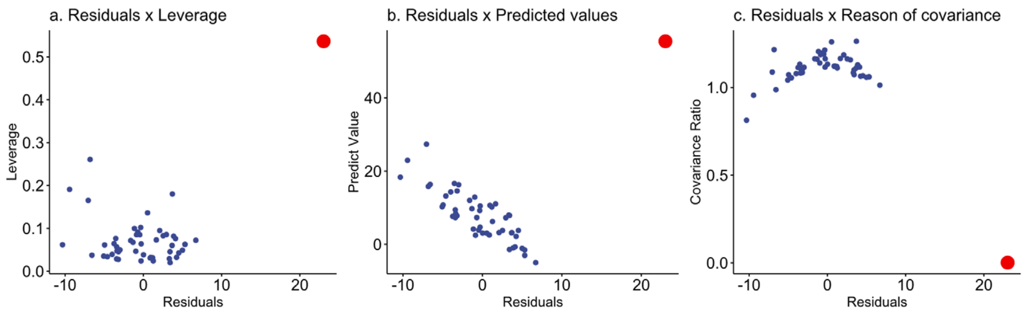
Source: The authors, based on Agresti and Finlay (1997)

Leverage values are as also known as hat values. The leverage mean is defined by k + 1/n, where k represents the number of independent variables and n indicates the sample size. The values vary between zero (where the case has no influence) and one (where the observation strongly distorts the predictive capability of the model). Hoaglin and Welsch (1978) argue that cases above (2(k + 1)/n) must be cautiously observed, and Stevens (1984) suggests three times above the mean (3(k + 1)/n) as a threshold to identify cases with a disproportionate influence. Again, results indicate that the District of Columbia is an odd case (see Figure 15). It had a large residual and now demonstrates strong influence (lev = 0,517). On the whole, these elements suggest that this case may affect the consistency of the estimates.

Furthermore, it is possible to observe the relationship between large residuals and strong leverage. Observations with both characteristics are considered influence points and may have devasting effects on the consistency of the estimates. Figure 16 illustrates that idea.

Overall, observations with larger residuals and larger leverages affect the consistency of the estimated coefficients (as shown in Figure 16.A). Residuals can also be examined as a function of the values predicted by the model (Figure 16.B). The ideal scenario would be to observe a random distribution with most cases close to zero, but that is made impossible by a case that is too different from the rest (DC). We can observe the relationship between the residuals and the ratio of covariance (Figure 16.C)[6]. The smaller that ratio, the more atypical is the observation and the larger is the expected variance in the regression coefficients.

FIGURE 16 – Residuals, leverage, predicted values, and covariance ratio



Source: The authors

d) Cook's Distance

Another measure for identifying atypical cases is Cook's distance. Introduced by Dennis Cook in the paper "*Detection of Influential Observation in Linear Regression*" (Cook, 1977), it indicates the expected variation of the regression coefficients in the absence of a given observation (Stevens, 1984). Thus, the larger its value, the larger the expected variation in the magnitude of the regression coefficients.

The estimation consists of excluding a given case from the sample and re-estimating the model. Then, the result is compared to a model estimated with all observations. Since Cook's distance measures the difference between both estimations, it is possible to evaluate the relative influence of each case on the magnitude of the coefficients. Schematically, the literature related to Cook´s distance has different criteria to classify a case as deviant: a) Cook and Weisberg (1982) state that values above one must be observed with caution; b) Cook > 4/n, in which n = number of cases; c) Cook > 4/(n - k - 1), in which n = number of cases and k = number of variables; and d) comparatively examine the value of the cases with a graphic analysis.

To illustrate how Cook's distance can be used to detect outliers, we replicate the data reported by Williams (2016) (see Table 4).

TABLE 4 – **Descriptive statistics of Cook's distance for Williams's (2016) database**

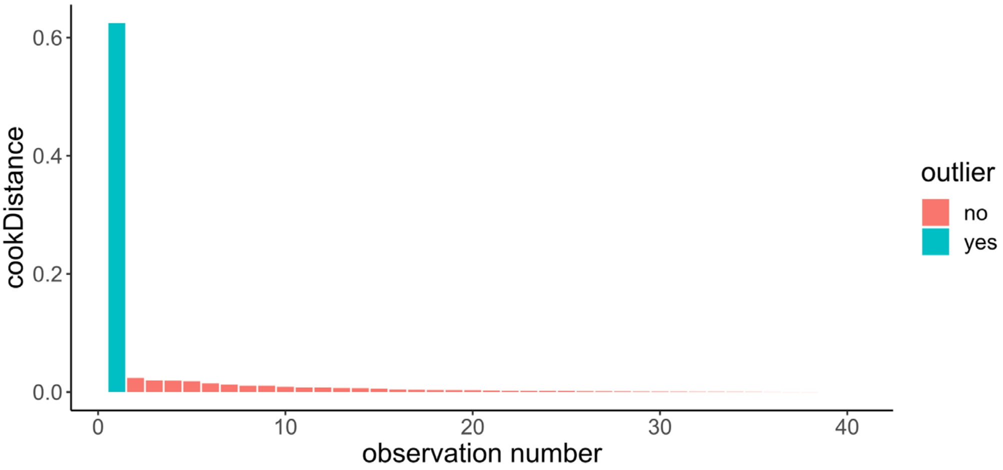
VARIABLE	MEAN	STANDARD DEVIATION	MINIMUM	MAXIMUM
Cook's Distance	0.0209	0.0981	0.0000	0.6246

Source: The authors, based on Williams (2016)

[6] That statistic is calculated from the determinant of the covariance matrix when a certain case is excluded from the analysis. The closer it is to one, the smaller the effect of one specific case.

As can be seen in Figure 17, there is no observation with Cook > 1. Using the Cook > 4/n rule of thumb, we find an observation with score 0.6246 that supersedes the threshold of 0.4 (4/40), suggesting a more careful inspection in the data.

FIGURE 17 – Graphic representation of Cook's distance



Source: The authors, based on Williams (2016)

In short, Cook's distance is a valuable metric for evaluating the impact of individual data points on a regression model. By identifying influential observations, scholars can make informed decisions about data cleaning and model selection.

e) Mahalanobis distance

The last procedure to detect atypical observations is Mahalanobis distance. Introduced in 1936 by Prasanta Chandra Mahalanobis in the paper "*On the generalised distance in statistics*" (Mahalanobis, 2018), it is one of the most used statistics to measure extreme cases in multivariate distributions. Mahalanobis distance informs the distance between the case and the centroid of the independent variables. In other words, as the centroid represents the mean of the means in a multidimensional space, Mahalanobis distance measures the similarity between a given observation and the mean of several distributions.

Mahalanobis distance is widely used in cluster analysis to classify observations by the level of similarity between the cases. Unlike the Euclidean distance, Mahalanobis considers the correlation between variables, which eliminates possible scale problems. The greater the distance, the higher the difference between a given case and the cluster center.

According to Hair et al. (2009, p. 75), the Mahalanobis distance "measures each observation's distance in multidimensional space from the mean center of all observations, providing a single value for each observation no matter how many variables are considered". The larger the D^2 is, the larger the distance of a given observation for the multidimensional space estimated by the variables' means. The literature suggests that observations with a D^2 above 2.5 must be considered multivariate outliers in small samples and cases above 3 or 4 might represent deviant observations in large samples (Hair et al., 2009).

We will use two examples to illustrate how scholars can use Mahalanobis distance to detect multivariate outliers. The first consists of a simulation of 100 observations and four variables. All have zero mean and standard deviation equal to one. Table 5 summarizes the descriptive statistics of the simulated variables.

TABLE 5 – **Descriptive statistics (x_1, x_2, x_3, and x_4)**

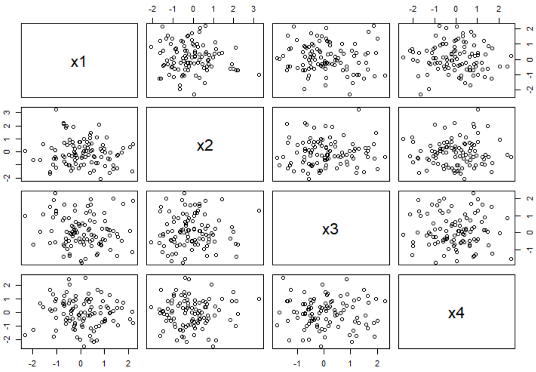
VARIABLE	MINIMUM	MAXIMUM	MEAN	STANDARD DEVIATION
X_1	-2,39	2,35	-0,17	1,03
X_2	-3,06	2,36	-0,03	1,04
X_3	-2,00	2,36	0,22	0,93
X_4	-2,45	2,58	0,11	0,87

Source: The authors

We examined the four distributions from the standardized scores' criteria and the interquartile range and did not detect any univariate outliers. We also examined the dispersion graphs and still did not find a deviant bivariate case. Since the variables were generated independently, it is expected that the correlation between them equals zero and is not statistically significant. That is, since the variables are random and were produced by different processes, it is expected that there is no dependency among them. Figure 18 illustrates this information.

FIGURE 18 – Correlation between the four random independent variables

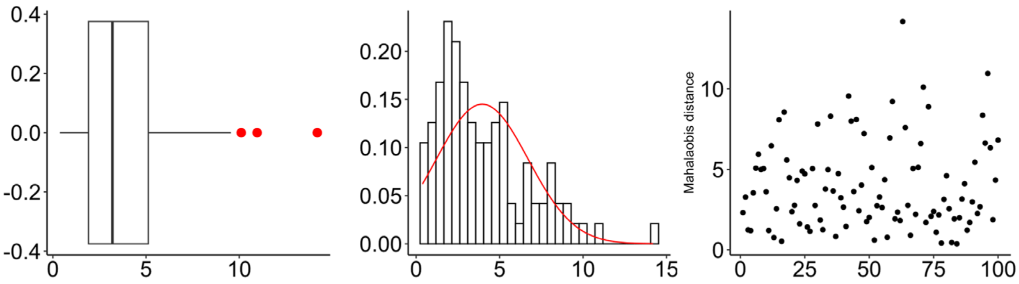
R P-VALUE	X_1	X_2	X_3	X_4
X_1	1	-0,041 (0,687)	0,001 (0,992)	-0,006 (0,953)
X_2		1	-0,168 (0,096)	-0,015 (0,884)
X_3			1	0,069 (0,492)
X_4				1



Source: The authors

Then, Figure 19 illustrates different ways of comparatively examining the value of D^2.

FIGURE 19 – Different ways of illustrating a multivariate outlier[7]



Source: The authors

[7] Another way to test the significance of Mahalanobis's distance for each observation is to calculate 1 - chi-square distribution, having as parameters the distance's magnitude and the degree of freedom equal to the number of variables used to calculate the centroid, k = 4, in our case.

All graphs in Figure 19 indicate that one case is very different from the remaining ones in the combination of the variables (x_1, x_2, x_3, and x_4). Thus, even if that case is not extreme in a univariate or bivariate way, it can be characterized as an extremely atypical combination of values. Faced with such an abnormality, researchers should verify possible causes and consequences for the consistency of the estimates.

In the sections above, five main criteria for identifying outliers were presented. These methods (summarized in Table 6) can be used by researchers in univariate, bivariate and multivariate data analysis.

TABLE 6 – **Identification criteria**

TECHNIQUE	CRITERION
Standardized scores	Standardized scores superior to 3 and inferior to -3 may be classified as atypical. In small samples (n<80), values above 2.5 also must be more carefully observed.
Interquartile range	Q1 and Q3 represent the values of the first and third quartiles, respectively, while the values for "g" are 1.5 for atypical cases, 2.2 for more extreme observations, and 3 for outliers that are worrisome
Standardized residuals	Model residuals larger than two are concerning and ones larger than three may be considered extreme cases.
Mahalanobis distance	D^2/df>2,5 for small samples and 3 or 4 for large samples (in which df = number of variables used; p-value = 0.005 or 0.001).
Cook's distance	Cook>1; Cook> 4/n, in which n = number of cases and Cook > 4/(n-k-1), in which n = number of cases and k = number of variables

Source: The authors

6. CONCLUSION

The identification of outliers is one of the earliest concerns in statistical analysis. As far back as 1778, Daniel Bernoulli criticized astronomers for their tendency to omit extreme observations and to analyze the remaining sample as if it were the original data. In 1964, the United States Supreme Court Justice Potter Stewart used the expression "*I know when I see it*" to establish the criterion for what is considered obscene (Jacobellis v. Ohio). But, unlike Stewart´s subjective criterion, data analysts have specific methods to identify unusual cases.

In this paper, we outlined five techniques for detecting univariate, bivariate, and multivariate outliers. For univariate cases, the Z score and the interquartile range are the most appropriate tools. In bivariate relationships or multi-variable relationships, scholars can use visual analysis to detect any patterns in the residuals and Cook's distance. As a specific technique for multivariate cases, we described the Mahalanobis distance. While the focus is on Political Science, these methods can be applied by data analysts across various fields of study. We believe that considerable progress in data analysis quality can occur if scholars follow the procedures presented in this article.

REFERENCES

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. **Journal of the American Statistical Association**, *89*(428), 1329–1339.

Atkinson, A. C., & Mulira, H.-M. (1993). The stalactite plot for the detection of multivariate outliers. **Statistics and Computing**, *3*(1), 27–35.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). **Regression diagnostics: Identifying influential data and sources of collinearity** (Vol. 571). John Wiley & Sons.

Benoit, K. (2011). Linear regression models with logarithmic transformations. **London School of Economics**, *London*, *22*(1), 23–36.

Blalock, H. M. (1979). Social Statistics. Revised. *New York, NY: McGraw-Hill. Box, GEP, & Cox, DR (1964). An analysis of transformations*. **Journal of the Royal Statistical Society**, *Series B (Methodology)*, *26*(2), 211–252.

Bluman, A. G. (2013). **Elementary statistics: A step by step approach: A brief version**. McGraw-Hill.

Bollen, K. A. (1988). "If You Ignore Outliers, will they Go Away?": A Response to Gasiorowski. **Comparative Political Studies**, *20*(4), 516–522. https://doi.org/10.1177/0010414088020004005

Bunge, M. (2016). Mechanism and Explanation: **Philosophy of the Social Sciences**. https://doi.org/10.1177/004839319702700402

Cateni, S., Colla, V., & Vannucci, M. (2008). Outlier Detection Methods for Industrial Applications. **Advances in Robotics, Automation and Control**. https://doi.org/10.5772/5526

Chandola, V., Banerjee, A., & Kumar, V. (2007). Outlier detection: A survey. **ACM Computing Surveys**, 14, 15.

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

Cook, R. D., & Weisberg, S. (1982). **Residuals and influence in regression**. New York: Chapman and Hall.

Davies, L., & Gather, U. (1993). The identification of multiple outliers. **Journal of the American Statistical Association**, *88*(423), 782–792.

De La O, A. L. (2013). Do conditional cash transfers affect electoral behavior? Evidence from a randomized experiment in Mexico. **American Journal of Political Science**, *57*(1), 1–14.

Finlay, B., & Agresti, A. (1997). **Statistical methods for the social sciences**. Dellen.

Fox, A. J. (1972). Outliers in Time Series. **Journal of the Royal Statistical Society**. *Series B (Methodological)*, *34*(3), 350–363. JSTOR.

Gerring, J. (2004). What Is a Case Study and What Is It Good for? **The American Political Science Review**, *98*(2), 341–354. JSTOR.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. **Technometrics**, *11*(1), 1–21.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). **Análise multivariada de dados** – 6ed. Bookman Editora.

Hawkins, D. M. (1980). **Identification of outliers** (Vol. 11). Springer.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. **Journal of the American Statistical Association**, *81*(396), 991–999.

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. **The American Statistician**, *32*(1), 17–22.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. **American Journal of Political Science**, *54*(1), 229–247.

Iglewicz, B., & Banerjee, S. (2001). A simple univariate outlier identification procedure. **Proceedings of the annual meeting of the american statistical association**, 5–9.

Iglewicz, B., & Hoaglin, D. C. (1993). **How to detect and handle outliers**. ASQC Quality Press.

Imai, K., King, G., & Rivera, C. V. (2016). Do nonpartisan programmatic policies have partisan electoral effects? Evidence from two large scale randomized experiments. **Unpublished manuscript**, *Princeton University and Harvard University Retrieved from* https://gking. harvard. edu/files/gking/files/progpol. pdf.

Johnson, R. A., & Wichern, D. W. (2002). **Applied multivariate statistical analysis** (Vol. 5). Prentice hall Upper Saddle River, NJ.

Lewis, T., & Barnett, V. (1994). **Outliers in statistical data**. John Wiley & Sons.

Mahalanobis, P. C. (2018). On the generalized distance in statistics. *Sankhyā:* **The Indian Journal of Statistics**, *Series A (2008-)*, *80*, S1–S7.

Mahoney, J. (2001). Path-Dependent Explanations of Regime Change: Central America in Comparative Perspective. **Studies in Comparative International Development**, *36*(1), 111–141. https://doi.org/10.1007/BF02687587

Mendenhall, W. (1982). **Statistics for management and economics** (4th edition). Duxbury Press.

Moore, D. S., & McCabe, G. P. (1999). **Introduction to the Practice of Statistics**. W.H. Freeman.

Nunnally, J., & Bernstein, I. (1994). **Psychometric Theory**: *3rd (Third) edition*.

Osborne, J., & Overbay, A. (2019). The power of outliers (and why researchers should ALWAYS check for them). **Practical Assessment, Research, and Evaluation**, *9*(1). https://doi.org/10.7275/qf69-7k43

Osborne, J. W., Christianson, W. R., & Gunter, J. S. (2001). **Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field**. https://eric.ed.gov/?id=ED463316

Pyle, D. (1999). **Data Preparation for Data Mining**. Morgan Kaufmann.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. **ACM SIGMOD Record**, *29*(2), 427–438. https://doi.org/10.1145/335191.335437

Ross, W. H. (1987). The Geometry of Case Deletion and the Assessment of Influence in Nonlinear Regression. **The Canadian Journal of Statistics / La Revue Canadienne de Statistique**, *15*(2), 91–103. JSTOR. https://doi.org/10.2307/3315198

Seo, S. (2006). **A review and comparison of methods for detecting outliers in univariate data sets** [Master's Thesis, University of Pittsburgh]. http://d-scholarship.pitt.edu/7948/

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. **Psychological Bulletin**, *95*(2), 334–344. https://doi.org/10.1037/0033-2909.95.2.334

Verardi, V., & Croux, C. (2009). Robust Regression in Stata: **The Stata Journal**. https://doi.org/10.1177/1536867X0900900306

Walfish, S. (2006). A review of statistical outlier methods. **A review of statistical outlier methods**, *30*(11), 82-88 [4 p.].

Weber, S. (2010). Bacon: An Effective way to Detect Outliers in Multivariate Data Using Stata (and Mata). **The Stata Journal**: *Promoting Communications on Statistics and Stata*, *10*(3), 331–338. https://doi.org/10.1177/1536867X1001000302

Williams, R. (2016). **Outliers**. University of Notre Dame. https://www3.nd.edu/~rwilliam/stats2/l24.pdf

Zeller, R. A., Zeller, R. A., Zeller, & Carmines, E. G. (1980). **Measurement in the Social Sciences**: *The Link Between Theory and Data*. CUP Archive.

Abstract

Living with outliers: How to detect extreme observations in data analysis

Data analysts often view outliers with skepticism due to their potential adverse effects, such as violating assumptions, hindering visualizations, and leading to biased estimates. In this paper, we present a practical guide for identifying outliers, which includes the step-by-step description of five statistical methods specifically designed to detect extreme observations: (1) standardized scores, (2) interquartile range, (3) standardized residuals, (4) Cook's distance, and (5) Mahalanobis distance. To enhance the learning experience, we provide both raw data and R scripts, empowering researchers to apply these techniques to their own data. By following the procedures outlined in this paper, scholars in a variety of fields will be able to make substantial progress in the quality of their data analysis.

Keywords: *Outliers; Extreme cases; Atypical observations; Anomaly detection.*

Resumo

Vivendo com *outliers*: Como detectar casos extremos em análise de dados

Os analistas de dados veem os outliers com ceticismo pelo potencial de efeitos adverso como, violação de pressupostos, dificuldade de visualização gráfica e estimativas enviesadas. Neste artigo, apresentamos um guia prático sobre como identificar outliers que inclui cinco técnicas estatísticas especialmente destinadas a detectar observações extremas: (1) escores padronizados; (2) intervalo interquartil; (3) resíduos padronizados; (4) distância de Cook e (5) distância de Mahalanobis. Para melhorar a experiência pedagógica, compartilhamos dados originais e scripts computacionais para R, que permitirão aos estudiosos implementar esses procedimentos para seus propósitos de pesquisa. Seguindo os procedimentos descritos neste trabalho, os acadêmicos de diversas áreas podem fazer um progresso significativo na qualidade de sua análise de dados.

Palavras-chave: *Outliers; Casos extremos; Observações atípicas; Detecção de anomalias.*

APPENDIX

Q1: What are outliers?

A1: Outliers are data points that significantly differ from the majority of other observations in a dataset. They can be unusually high or low values and may negatively affect statistical analyses and models if not appropriately handled. In what follows, we summarize different definitions of outliers.

AUTHOR (YEAR)	DEFINITION
Grubbs (1969)	*An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs*
Hawkins (1980)	*An observation that differs so much from other cases as to arouse suspicion that a different mechanism generated it*
Fox (1972)	*An outlier is an observation whose dependent variable value is unusual given the value of the independent variable*
Johnson and Wichern (2002)	*An observation in a dataset which appears to be inconsistent with the remainder of that set of data*
Mendenhall (1982)	*Observations whose values lie very far from the middle of the distribution in either direction*
Ross (1987)	*Outliers are data points that do not appear to follow the pattern of the other cases*
Pyle (1999)	*An outlier is a single occurrence, or one with very low frequency, of the value of a variable that is far away from the bulk of the values of the variable*
Moore and McCabe (1999)	*An outlier is an observation that lies outside the overall pattern of a distribution*
Ramaswamy et al. (2000)	*An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data*
Bluman (2013)	*An "outlier" is an extremely high or an extremely low data value when compared with the rest of the data values*
Hair et al. (2009)	*Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations*

Q2: Where do outliers come from?

A2: There are four main explanations for the presence of outliers in data: (1) malicious activity; (2) instrument malfunction; (3) abrupt changes in the environment; and (4) human error.

Q3: Why is it important to address outliers in data analysis?

A3: Outliers are often viewed with skepticism by data analysts due to their potential adverse effects, such as violating assumptions, hindering visualizations, leading to biased estimates, and altering the sign of coefficients.

Q4: What are some common techniques to identify outliers?

A4: Common techniques to identify outliers include visual inspection using box plots or scatter plots, statistical methods like the Z-score or the IQR (Interquartile Range). There are algo regression-based measures such as Cook´s Distance and Leverage values that may be useful to spot extreme cases. More recently, machine learning algorithms have been developed to detect outliers.

Q5: How can outliers be handled or managed in data analysis?

A5: Outliers can be handled in several ways:

Removal: You can choose to remove outliers from the dataset, but this should be done with caution, as it can lead to loss of valuable information.

Transformation: Applying mathematical transformations like logarithms can reduce the impact of outliers.

Imputation: Outliers can be replaced with more typical values using techniques like mean, median, or regression imputation.

Robust Models: Using models and algorithms that are less sensitive to outliers, such as robust regression or tree-based algorithms like Random Forests.