Journal of Probability and Statistical Science
22(1), 1-15 Sep. 2024

Analysis of Data Containing Outliers

David L. Farnsworth
School of Mathematics and Statistics
Rochester Institute of Technology, Rochester, New York, USA

ABSTRACT

A strategy for accommodating outlying observations, as well as non-representative, suspect, missing, or otherwise troubling observations, is described. Each unusual observation is decomposed into the sum of two components. One component is the value implied by the trusted observations in the data set. The other component is the unusual part. In this way, the fitting of the data set can then proceed, and, additionally, a numerical value can be ascribed to the unusual part. The method offers not only an antidote for observations with irregular numerical values, which often have the power to contaminate and alter analyses, but also a measure of the magnitudes of the unusual components of those observations. Univariate data and ordered pairs in least-squares fitting are presented as examples.

Keywords: Outlier; Least-squares fitting; Suspect observation; Additive fit; Linear regression.

1. Introduction

A method for addressing data sets containing outliers is presented. An outlier is an observation that is far from the data's subset in which there is confidence [1, 2, 12]. The method supplies an estimate of the size of each outlier, and, simultaneously, the data set is prepared for subsequent analyses without any potentially harmful influences of these outliers. All outliers are handled at once, not piecemeal, so that identifying and estimating an outlier does not contaminate the identifications and estimations of other outliers.

One of the method's desirable features is that it can be used when there are suspect observations that might not appear to be outliers, but are nonetheless problematical. Another is that it can be used for missing observations.

The main idea of the proposed method is to replace simultaneously each outlier or suspect value with its own placeholder, such that, when the altered data set is fitted, each placeholder's value gives a perfect fit and thus has a residual of zero. The difference between each placeholder's value and the actual data value is the size of the discordant or otherwise troubling part of the observation. The placeholders' values can remain in the data set for the next fitting or statistical procedure that might be favored for an ongoing analysis.

The strategy is not limited to a small number of troublesome observations. It does not break down, even in the presence of a large number of such observations. Other useful features are that it can be implemented without difficulty and be explained straightforwardly to colleagues and clients when consulting. It maintains the integrity of the data set. The method does not overwork the data. Although it might be difficult to quantify *overwork*, it is usually apparent when tens or hundreds of calculations per observation are performed, so that those computations may be deciding the outcome, instead of the data doing so. Also, calculations carry the baggage of round-off errors and of assumptions about the data, which might not be desirable.

The preferred way to determine that an observation is an outlier or simply suspect should be as simple as possible. Generally, graphical methods supply excellent ways to ascertain that an

- Received July 2024, in final form August 2024.
- David L. Farnsworth (corresponding author), is affiliated with the School of Mathematics and Statistics, Rochester Institute of Technology, Rochester, New York, USA.
dlfsma@rit.edu

observation is an outlier. Those allow the observations to speak for themselves with little massaging or manipulation. Also, each observation may be separately examined. Besides being a potential outlier, an observation may be suspect for many reasons, including the source being unreliable or the failure of equipment.

The strategy for accommodating outlying observations, as well as non-representative, suspect, missing, or otherwise troubling observations, can be summarized as follows. Each unusual observation is decomposed into the sum of two components. One component is to be given the value implied by the trusted observations in the data set along with the probable subsequent or planned analyses. The other component is the unusual part. In this way, the follow-up fitting of the data set can proceed using the first component, and a numerical value can be ascribed to the unusual part using the second component. This fitting and the estimation of the outlying portions are performed simultaneously on all suspect observations within one framework. The method offers not only an antidote for observations with irregular numerical values, which often have the power to contaminate and alter analyses, but also a measure of the magnitudes of the unusual components of those observations.

In Section 2, the relatively simple example of univariate observations introduces many of the main ideas. Data pairs are explored in Section 3. Conclusions are in Section 4. Least-squares fitting is emphasized, and a data-analysis stance is taken.

2. Univariate Observations

Consider a sample containing the $n_T + n_S = n$ values

$$x_1, x_2, \dots, x_{n_T}, x_{n_T+1}, x_{n_T+2}, \dots, x_{n_T+n_S} \tag{1}$$

which have been ordered for convenience in such a manner that it is believed that the first n_T values can be trusted (hence the subscript T) as probably sufficiently accurate or precise, but the remaining n_S values are suspect (hence the subscript S). Assume that there are good reasons for believing that the suspect data are outliers or problematical. Our basis for that decision might be from an examination of a dot diagram or a normal probability plot or from calculations of various statistics [5, 9]. There are other grounds for identifying an observation as suspect. For example, one is the observation's source, such as it being from a lab with unsanitary conditions or an instrument that might be mistrusted. Such data values may be included as suspect.

The goal is to estimate the sizes of the outliers and at the same time neutralize those observations' impact on the intended ensuing analyses of the data. There are at least three ways to proceed.

1. Use all n observations (1) as one set.
2. Use only the n_T trusted observations in (1) and delete the n_S suspect observations.
3. Use the n_T trusted observations along with the concordant part of each of the n_S suspect observations.

Choice 1 might be dismissed because it ignores that there is information or analyses that show that the n_S suspect observations are unusual. Those observations might harm ensuing analyses. For example, they could shift the mean and produce a large standard deviation, so that there is less precision and clarity. Outlying observations can contaminate, and even render useless, analyses, especially those that are least-squares procedures [1, 11, 12].

Choice 2 involves the deletion of data, which might be objected to on principle. It offers no guidance about putting the deleted observations to use.

Choice 3 is recommended. Each suspect observation is temporarily replaced by one value t, which is to be determined. Thus, the process uses the set

$$x_1, x_2, \dots, x_{n_T}, t, t, \dots, t, \tag{2}$$

containing n_S copies of t. The criterion for the value of t is that the mean of the set (2) is t. Through this process, each value x_i among the suspect observations is decomposed into the sum $t + (x_i - t)$, where the second addend is a measure of the discordant or outlying component of the observation. The first addend remains in the data set. In this way, the observations are set up for subsequent fitting and statistical analyses, and the sample size is not reduced beforehand. The discordant component is set aside as the outlying portion, i.e., the size of the outlying component is fitted thusly.

Theorem 1. *The criterion that the mean is t for the set (2), consisting of the n_T trusted observations and the n_S values t, implies that $t = \bar{x}_{n_T}$, where $\bar{x}_{n_T}$ is the mean of the n_T trusted values.*

Proof. The criterion is

$$\frac{n_T \bar{x}_{n_T} + n_S t}{n_T + n_S} = t,$$

whose unique solution is $t = \bar{x}_{n_T}$. □

In the set (2), the residual of each of the n_S suspect observations is zero from $t - \bar{x}_{n_T} = 0$. The sizes of the outlying components are $x_i - \bar{x}_{n_T} = x_i - t$ for $i = n_T + 1, n_T + 2, \dots, n_T + n_S = n$.

Theorem 1 says that the sample mean of the n_T trusted values is the center of the set (2) of n values composed of the trusted observations and the non-outlying components of the suspect observations. The sample variance of this set of n values is

$$s^2 = \frac{\sum_{i=1}^{n_T} (x_i - \bar{x}_{n_T})^2 + \sum_{i=n_T+1}^{n_T+n_S} (t - \bar{x}_{n_T})^2}{(n_T + n_S) - n_S - 1} = \frac{\sum_{i=1}^{n_T} (x_i - \bar{x}_{n_T})^2}{n_T - 1} = s_{n_T}^2,$$

where n_S degrees of freedom are subtracted in the denominator, because they have been used in the determinations of the outlying components of the n_S observations that had been deemed to be suspect. Therefore, in this case of a univariate data set with a least-squares analysis, the sample mean and variance of (2) can be computed as if the n_S suspect values have been discarded.

Statistical analyses can proceed using the n_T trusted values and n_S values $t = \bar{x}_{n_T}$ as the data set, i.e., (2), while the outliers have been fitted. One degree of freedom is lost for each of those fitted values. Thus, the outlying portions of the suspect observations have been estimated and can be set aside, leaving values that are the mean value and ready to be part of later analyses.

Unlike some ways of accommodating suspect data, there are no restrictions on the number n_S, except that at least two of the original data values must remain for the succeeding analysis from $(n_T + n_S) - n_S - 1 = n_T - 1 \geq 1$. The process does not break down when there is a large number of defective or unusual values. Of course, a very large number of suspect observations might threaten the integrity of an experiment and the usefulness of its statistical analysis. Indeed, if the researcher suspects that the data set is formed from a mixture of two or more distributions, designating observations as outliers may be incorrect. Designating a datum as suspect could take into consideration knowledge about the generation of the observations [3, 15]. The observations are not required to be far from the other observations or the mean in order for this process to be used, such as for problematic, but not necessarily outlying, values.

It is surprising that in this case of univariate observations this procedure gives a framework for justifying the deletion of the suspect observations, because the sample mean, sample variance, and the degrees of freedom are the same with this procedure and with deletion.

Investigating the potential causes of the suspect observations and the meanings of their sizes may be a follow-up to the procedure. Sometimes, instead of being errors, outliers can reveal something interesting and important about the data set or its source. As in any data set, it often happens that suspect or nonconforming observations have the most significance and information.

Example 1. Consider the ten values {2.6, 2.8, 2.85, 3.0, 3.2, 3.35, 3.4, 3.6, 6.0, 7.0}, which have been ordered for easier viewing. A dot diagram shows that the 6.0 and 7.0 are outliers. Thus, $n_T = 8, n_S = 2, n = 10,$ and $\bar{x}_{n_T} = (2.6 + 2.8 + \cdots + 3.6)/8 = 3.1$. The estimates of the sizes of the outlying components are 6.0 – 3.1 = 2.9 and 7.0 – 3.1 = 3.9. The sample variance is $s^2 = ((2.6 - 3.1)^2 + \cdots + (3.6 - 3.1)^2)/7 = 0.118.$

3. Ordered Pairs of Observations

Consider the set of n points (x_i, y_i), $i = 1, 2, \ldots, n$, which is composed of n_T trusted points for $i = 1, 2, \ldots, n_T$ and n_S suspect points for $i = n_T + 1, n_T + 2, \ldots, n_T + n_S = n$. Assume that there are no errors in the x-values. Suppose that, based upon a scatter diagram or some elementary analyses [4, 8, 10], the suspect points are believed to be outliers. The least-squares regression line for the n_T trusted points is

$$y = b_0 + b_1 x. \tag{3}$$

As in Section 2, each suspect point is decomposed, such that the y-value has two additive components. The x-value is not changed. In many applications, the explanatory x-value is assumed to be known and free of errors. The direction of prediction is parallel to the y-axis. This is analogous to Theorem 1, where the criterion leads to the unique mean for univariate data. Here, the unique regression line is

$$y = b_0' + b_1' x, \tag{4}$$

which fits the set consisting of the n_T trusted observations and the n_S suspect points' concordant components. The points associated with the concordant components of the suspect observations are

$$(x_{n_T+i}, t_i) \tag{5}$$

for $i = 1, 2, \ldots, n_S$. By design, the residuals for these points are zero. Recall that the residual for an observed point is its signed vertical distance from the regression line.

Theorem 2. *The criterion that the least-squares regression line (4) fits the n_T trusted observations and the n_S pairs in (5) implies that the line is (3).*

Proof. For the set of n pairs, the sum of squares is

$$\sum_{i=1}^{n} \left(y_i - (b_0' + b_1' x_i)\right)^2,$$

which is

$$\sum_{i=1}^{n_T} \left(y_i - (b_0' + b_1' x_i)\right)^2 + \sum_{i=1}^{n_s} \left(t_i - \left(b_0' + b_1' x_{n_T+i}\right)\right)^2.$$

Because the concordant portions of the suspect points are required to be on the line, each term in the second sum is zero, and this reduces to the minimization over the line's coefficients that is required for the trusted observations alone. Indeed, the criterion says that the concordant

components (5) of the suspect points lie on the line, so they have no influence on the fit [4, 6, 9], and the line is determined by the trusted points. □

As in Section 2, this process supplies n points for further fitting and statistical analysis. The measurement of the size of each outlying component is $y_i - b_0 - b_1 x_i$, which is the signed vertical distance of the suspect point from the line (3). Points with positive (negative) residuals are above (below) the line. Because of this fitting of the outlying components, n_S degrees of freedom are depleted.

The sample variance of the slope is the sum of squares of the residuals divided by the sum's degrees of freedom [4, 6]. Because the residuals are zero for the suspect points and the degrees of freedom for the set of n points is reduced by n_S, the variance of the slope is the same as for the slope of regression line for the trusted points,

$$s^2(b_1) = \frac{\sum_{i=1}^{n_T}(y_i - b_0 - b_1 x_i)^2}{n_T - 2}.$$

As is the case for the univariate set in Section 2, there are no restrictions on the number n_S of suspect data, except that at least three of the original data values must remain for the succeeding analysis from $(n_T + n_S) - n_S - 2 = n_T - 2 \geq 1$. This process offers a compelling argument for deletion of suspect observations in standard least-squares linear regression.

4. Conclusions

A procedure to address fitting in the presence of outliers and other discordant observations has been offered. It contains a method for assigning values to the outlying components of those observations. Univariate data with the mean and bivariate observations with the regression line are presented as examples.

As a first step, the set of observations needs to be examined for outliers and other suspect observations and for the suitability of the proposed additive fitting. In each of the examples, graphical techniques are favored [16], but many analytical methods are available [2]. For a univariate set, an elementary dot diagram may suffice. For bivariate observations, a scatterplot may be sufficient. It has been long held that Tukey's eye-to-paper or eye-smoothing technique for sighting along a proposed regression line is reliable [13].

By using graphical tools, the observations are viewed in a raw state, rather than being manipulated, altered, or reformulated. Such manipulations can possess assumptions, especially of the probability distribution that generated the data, that have no natural place in the analyses. The number of suspect observations is determined by this graphical search and by considerations about the quality of the observations, such as their source, instead of by limitations of the methodology for finding them or by uninformed guesses [2].

The core of the fitting procedure is that each outlier is decomposed into a sum of two components. One component is a perfectly fitting value that is determined from the trusted subset of the observations and the other is the outlying or discordant part of the observation. These two fitting procedures are performed concurrently in a complementary fashion. Because all the outliers are fitted in one process, instead of sequentially, the analysis of one outlier is not unduly influenced by other outliers.

The method works for any observation that is identified as suspect; it need not be distant from the other observations or the final fitted values. In those cases, the discordant parts may not be as meaningful as the discordant parts of outliers, but the advantage remains that the process leaves

them with little or no weight. The method can accommodate missing values, and it gives a value for further fitting [7, 8].

The method is feasible in the sense that the calculations are not difficult to implement or to explain to others. These simple choices and calculations can be performed with statistical computing programs in routine ways.

For cases of fitting a univariate data set with a mean and fitting pairs with a least-squares regression line, an additional benefit, which may be surprising, is that this method is equivalent to the deletion of the suspect values.

This method might be expanded to include errors or suspect values in the independent variables, such as x in Section 3. Research on errors in those values is an active area of inquiry; see [11] and [14] and their references.

References

[1]. Aggarwal, C.C. (2017). *Outlier Analysis* (2nd ed.). Berlin, Germany: Springer Nature. *doi.org/10.1007/978-3-319-47578-3*

[2]. American Society for Testing Materials Subcommittee E11.10 on Sampling/Statistics (2021). *ASTM E178-21: Standard Practice for Dealing with Outlying Observations.* West Conshohocken, PA, USA: ASTM International, *www.astm.org/e0178-21.html. doi.org/10.1520/E0178-21*

[3]. Cornell, J. (2002). *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data* (3rd ed.). New Yok, NY, USA: Wiley.

[4]. Harrell, Jr., F.E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2nd ed.). New York, NY, USA: Springer. *doi.org/10.1007/978-3-319-19425-7*

[5]. Hoaglin, D.C., and Moore, D.S. (1992). *Perspectives on Contemporary Statistics*. Washington, DC, USA: Mathematical Association of America.

[6]. Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). New York, NY, USA: McGraw-Hill/Irwin.

[7]. Little, R.J.A., and Rubin, D.B. (2020). *Statistical Analysis with Missing Data* (3rd ed.). New York, NY, USA: Wiley. *doi.org/10.1002/9781119482260*

[8]. Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A., and Verbeke, G. (2015). *Handbook of Missing Data Methodology*. Boca Raton, FL, USA: CRC Press. *doi.org/10.1201/b17622*

[9]. Montgomery, D.C., Runger, G.C., and Hubele, N.F. (2011). *Engineering Statistics* (5th ed.). New York, NY, USA: Wiley.

[10]. Rousseeuw, P.J., and Leroy, A.M. (2003). *Robust Regression and Outlier Detection*. New York, NY, USA: Wiley. *doi.org/10.1002/0471725382*

[11]. Shi, H., Zhang, X., Gao, Y., Wang, S., and Ning, Y. (2023). Robust total least squares estimation method for uncertain linear regression model. *Mathematics*, 11, 4354. *doi.org/10.3390/math11204354*

[12]. Suri, N.N.R.R., Murty, M.N., and Athithan, G. (2019). *Outlier Detection: Techniques and Applications*. Berlin, Germany: Springer. *doi.org/ 10.1007/978-3-030-05127-3*

[13]. Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley.

[14]. Vidal, R., Ma, Y., and Sastry, S.S. (2016). *Generalized Principal Component Analysis.* New York, NY, USA: Springer-Verlag. *doi.org/10.1007/978-0-387-87811-9*

[15]. Yao, W., and Xiang, S. (2024). *Mixture Models: Parametric, Semiparametric, and New Directions*. Boca Raton, FL, USA: CRC Press.
[16]. Young, F.W., Valero-Mora, P.M., and Friendly, M. (2006). *Visual Statistics: Seeing Data with Dynamic Interactive Graphics*. New York, NY, USA: Wiley.