Deloitte.
Digital

Regression and Outliers

Trufa Science Inside – No. 4

Andreas Mielke
Deloitte Digital GmbH, Mannheim

11th February 2019

Standard regression often yields bad results if outliers are present. There are two ways out of this dilemma. The first is to remove the outliers. The second is to use robust methods for which the result does not depend on outliers. If the data set is large, the second possibility is often the only possible one. In this paper, we point out that robust methods can and should be used for outlier detection. The reason is that outliers often contain additional information and are thus important. They may for instance show that an additional factor is relevant to understand the behaviour of the full data set. A precise outlier detection is therefore mandatory.

1 Introduction

Regression is probably the most often used method to analyse empirical data. There are many methods to do a regression, from the classical linear regression to non-parametric regressions. Classical linear regression is implemented in any spread sheet component of standard office packages, therefore making it the most widely used form of regression. Its advantage is that it is easy to calculate. Mathematically, it is based on a least square estimator (for the reader who does not know what that is, just ignore it here). A linear fit contains two parameters, the slope of the line and its intercept with the y-axis. For each data point, one calculates the distance from the point to the line. One then takes the sum of the squares of all those distances. Now, the two parameters, slope and intercept, are determined such that this sum of squares takes a minimum. This method works well if the distances of observations from the fit are normally (Gaussian) distributed. Often, especially if the regression is done using a spread sheet program, this assumption is not even tested. The problem is that outliers strongly influence the result, simply because they are far away from the line and therefore yield a huge contribution to the sum of the squared distances. Even a single outlier can have a dramatic influence. We will illustrate that problem in the subsequent section.

One possibility of obtaining a valid result using a classical linear regression is to eliminate outliers first. The problem here is that there is no universally accepted definition of what an outlier is. A common understanding is that an outlier deviates significantly from or appears to be inconsistent with the other data in the data set. But this is not a definition in a mathematical sense. And since there is no common definition, the number of methods to detect outliers is huge and depends on the question one wants to answer, for a review see [1]. In the context of a classical linear regression, one could take the deviation of a given data point from the statistical model related to the standard deviation as a measure. We will come back to this idea later, but one problem is immediate: If the outlier changes the model parameters significantly, and this is often the case, the question, what the normal behaviour is, is influenced by the outlier.

A way out of that problem is to use robust methods, see [2]. For a short comprehensive mathematical overview on robust methods for outlier detection we refer to [3], for a text book to [4]. The advantage is that the robust statistical model does not depend significantly on the outliers and that, therefore, outliers can be detected easily.

It is important to keep in mind that the question whether some data are considered as outliers depends on the statistical model used for the analysis. Outliers are often classified as errors, and indeed, an outlier may be caused by some error. But often, esp. if there

is a larger subset of outliers, the statistical model used for the classification may be incomplete. The outliers may contain some additional information, and expanding the model to include that information may solve the problem, most of the outliers become regular data.

2 Influence and classification of outliers

Let us start with some simple linear regressions, depicted in Fig. 1. Fig. 1a) shows a simple linear regression. The points lie on a straight line which passes approximately through the origin, the slope is approximately 1. The deviation of the points from the line are distributed approximately normal and the deviations are small.

Figs. 1b-d) show the same points, only one of the points it shifted so that it becomes an outlier. In Figs. 1b-c) the outlier changes the regression line significantly. In both cases, the slope is much smaller and the line no longer touches the origin. Fig. 1d) shows an outlier which does not change the regression, but which nevertheless lies far away from the other points. The effect here is different: the outlier extends the region where the regression yields valid results. For the first two outliers, predictions become unreliable due to the presence of the outlier. In the latter, case a prediction may look reliable in a certain region although it is not reliable there. In Fig. 1d) for instance, one may doubt that the regression is valid above $y \gtrsim 12$.

The examples also illustrate a classification of regular data points and outliers. One distinguishes the orthogonal distance and the score distance of a data point to a data set. To be precise, these notions are used together with the so called principal component analysis (PCA), see e.g. [3] and the references therein. The outliers in Figs. 1b-c) have a large orthogonal distance to the regression line, the outlier in Fig. 1d) has a large score distance to the regular data. Outliers with a large orthogonal distance change the result of the regression significantly. Regular observations have a small score distance and a small orthogonal distance [3].

The reason for the large effect of outliers in standard linear regression is the least square estimate: the algorithm minimises the sum of the distances squared of the points from the regression line, the classical standard deviation. The use of the standard deviation is the direct consequence of the assumption that the distances are normally distributed. This assumption, although many people believe that it is true in most cases, is often violated and has to be tested in each serious analysis. The outliers in Figs. 1b-c) have a large distance to the original line depicted in Fig. 1a). Therefore, the weight of the outlier is large and the effect on the regression line is large.

Often, outliers are identified using rules. A simple rule is to compare the observed data with the statistical model. If the deviation is larger than some multiple of the standard deviation σ, a data point is considered to be an outlier. A typical threshold id 2.5σ. As discussed in [3], this is arbitrary and often fails. The problem is that such a threshold is not robust. To obtain a robust method, one needs to replace the statistical model by a robust statistical model and the standard deviation σ by a robust measure.

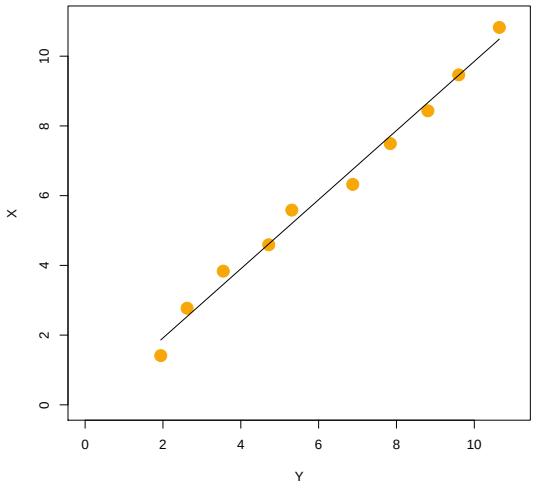
3 Robust regression and detection of outliers

The examples show that to obtain a reliable regression for the data, a suitable outlier detection may be necessary. Today, one often takes another route: one uses robust regression which yields results that do not change if a single outlier, or a small set of outliers is shifted around. The main difference between a robust regression and the standard regression is the use of a different measure for the deviation of the data from the regression curve. Concerning outliers, the main effect is that the weight of outliers is less than in the standard regression, it may even be 0, which means that the outlier is not taken into account. As a consequence, the outlier has only a small or even no effect on the regression curve.

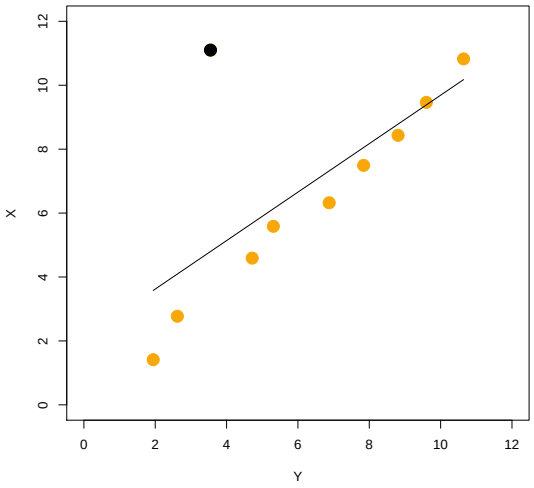
The main reason to use robust regression is clearly to obtain results which are valid even if there are outliers, or if the distances of the data to the regression curve are not normally distributed. But since the robust algorithm not only yields the regression curve but also the weights assigned to the data points, it can as well be used for outlier detection. We show a realistic example in Fig. 2. The example shows the data and the regression curve. The latter was calculated using a standard robust algorithm provided by the package 'robust' of the statistics software R [5]. Outliers are shown as well. Fig. 2 nicely depicts how a robust approach works. The algorithm induces a boundary of normal behaviour. Data within the boundary are considered normal, data outside the boundary are outliers. The boundary is not a hard boundary of in and out, but a weak boundary where a weight indicates how likely it is that a point is in or out.

There are many other ways to detect outliers using robust methods. For an overview we refer to [3].

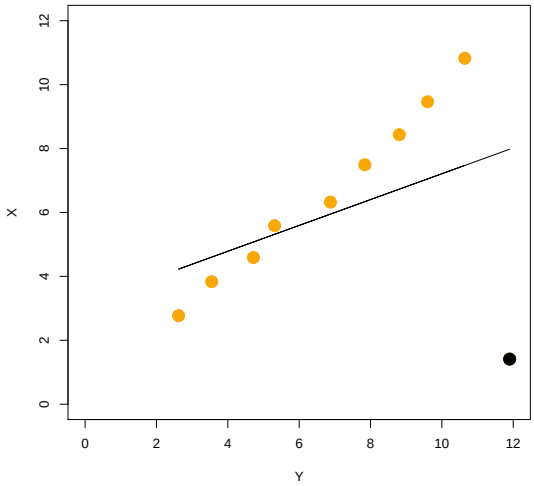
The important point we want to emphasise here is that outliers often yield additional information. The data points in the set of outliers may differ in some factor from the rest of the data. The outliers appear to be outliers because the additional factor is not taken into account. If this is the case, it is of interest to understand which factor is responsible for the difference. Today, in a world dominated by data, one often knows a lot more data than the ones shown in the diagram and in may be possible to find out what the differing



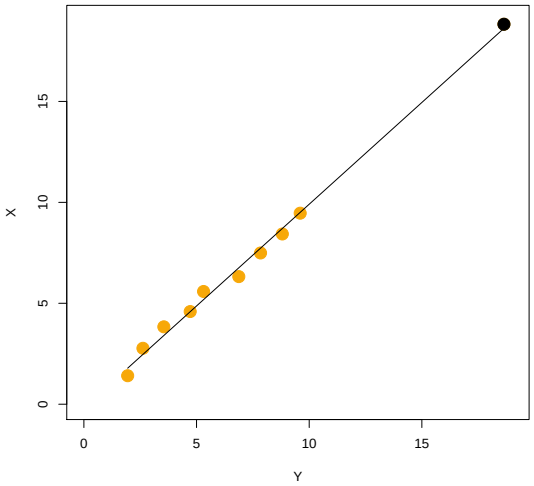
a) Regression without outliers.



b) One point shifted in y direction



c) One point shifted in x direction



d) One point shifted in both directions

Figure 1: Linear regression with outliers. The outliers are shown in black.

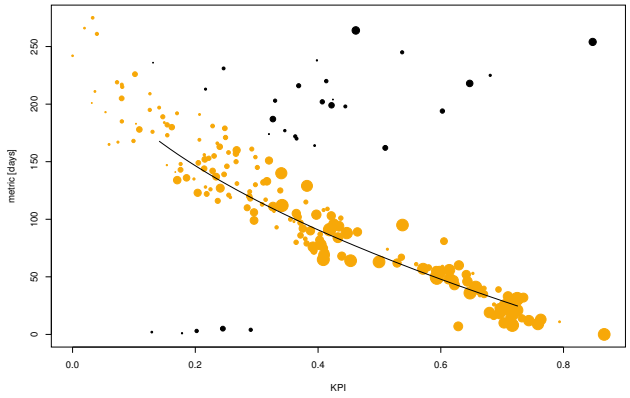


Figure 2: Robust regression and outlier detection. Outliers are shown in black.

factor is and what values it takes in the different subsets. In the example in Fig. 2 it was indeed possible to determine the factor from the data. The five data points around KPI≈ 0.2, metric≈ 0 in the figure belong to a special class, also most of the points with metric$\gtrsim 150$ and KPI$\gtrsim 0.3$. Including this factor eliminates roughly 90% of the outliers in this example.

In business data, the factor that differentiates the outliers from the regular data is often a different variant of the underlying business process. In easy cases it is caused by a different behaviour in a certain region, product group, plant, etc. In those cases the region, product group, or plant can be included as a factor in the statistical model. It becomes a relevant factor. For the new model, which includes the factor, the former outliers become regular data which can be explained by the model. In some of those easy cases an expert, who knows his business, would have included the additional factor from the start. But often, such knowledge is not hand or the data set is a bit more complex, or the differentiating factor is a combination of simple factors like a certain product group in a certain group of regions. In those cases, an automatic detection and a subsequent analysis of the outliers yields new insight. This new insight drives actions to improve, because one directly knows where to act. And this finally makes the outlier detection valuable.

References

[1] V.J. Hodge and J. Austin, A survey of outlier detection methodologies. Artificial Intelligence Review **22**, 85 (2004).

[2] A. Mielke. Robust Statistics. Trufa Science Inside – No. 1. (2016).

[3] P. J. Rousseeuw and M. Hubert, Robust statistics for outlier detection. WIREs Data Mining Knowl. Discov. **1**, 73–79 (2011).

[4] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection. Wiley-Interscience Paperback Series, John Wiley & Sons, Inc., Hoboken 2005.

[5] J. Wang, R. Zamar, A. Marazzi, V. Yohai, M. Salibian-Barrera, R. Maronna, E. Zivot, D. Rocke, D. Martin, M. Maechler, and K. Konis, robust: Robust Library. R package version 0.4-16. https://CRAN.R-project.org/package=robust (2014).

Deloitte.
Digital

This communication contains general information only not suitable for addressing the particular circumstances of any individual case and is not intended to be used as a basis for commercial decisions or decisions of any other kind. None of Deloitte GmbH Wirtschaftsprüfungsgesellschaft or Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collectively, the "Deloitte network") is, by means of this communication, rendering professional advice or services. No entity in the Deloitte network shall be responsible for any loss whatsoever sustained by any person who relies on this communication.

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. Please see www.deloitte.com/about for a more detailed description of DTTL and its member firms.

Deloitte provides audit, risk advisory, tax, financial advisory and consulting services to public and private clients spanning multiple industries; legal advisory services in Germany are provided by Deloitte Legal. With a globally connected network of member firms in more than 150 countries, Deloitte brings world-class capabilities and high-quality service to clients, delivering the insights they need to address their most complex business challenges. Deloitte's approximately 286,000 professionals are committed to making an impact that matters.