**How to Detect Outliers**

**N.B A group of used methods for detecting outliers in this work (multivariate methods) such as residual analysis, Cook’s distance, are all based on linear model, that assume in advance a linear relation between the predictor and the target. If the data is not linear, these methods may be not appropriate and my not be effective to detect outliers.**

**In such cases, it may be better to use other methods such as Local Outlier Factor (LOF), DBSCAN, or isolation forest that are designed specifically to detect outliers in non-linear data.**

Real-world data, often contain **wrong data types**, **missing values**, **irrelevant features** or **outliers**.

When it comes to **anomalies**, or **outliers**, they’re values in the data set that exhibit some abnormality and deviate significantly from the normal data. They can be a serious issue when training a machine learning algorithms or applying statistical techniques. They’re often the result of errors in measurements or exceptional system conditions. There’s no definitive rule to identify them and it depends on subject-area knowledge and an understanding of data collection process.

In this guide, we’ll go through univariate and multivariate methods to help us detect outliers.

1. **The Dataset:**

In this work, we generated a synthetic dataset where the relation between the two variables is linear. To explore the effect of unusual observations, we also included some outliers in the dataset.

1. **Detect Outliers:**
   1. **Univariate Techniques:**

A univariate method is a technic that operates on a single feature (also known as the target variable). In the context of identifying outliers, a univariate method uses only a single feature of the data to identify points that are significantly different from the data.

In the next steps in univariate analysis, I focused on analyzing the target ‘y’.

* + 1. **Kernel Density Estimate:**

Kernel density estimate is a useful tool for understanding the distribution of numerical variables. By examining the KDE, it can help identifying any skewness in the data. Skewness can indicate the presence of outliers, which are data points that lie significantly outside the range of the rest of the data.

In this plot, we used a kernel density estimate to visualize the distribution of the target. We marked the mean and three standard deviations from it with vertical lines. Upon examining the distribution, we noticed that there is slight skewness present in both tails of the distribution and that some data points lie significantly outside the range of three standard deviation units from the mean. These data points may be considered outliers and require further investigation.

* + 1. **Box plot:**

A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that displays the distribution of the data. It can be used to visualize the spread and skewness of the data, and it can also be used to detect outliers.

In a box plot, the box represents the interquartile range (IQR) of the data, which is the range between the first quartile (Q1) and the third quartile (Q3). The horizontal line inside the box represents the median of the data. The "whiskers" extending from the box represent the minimum and maximum values of the data, excluding any outliers. Outliers are represented by individual points on the plot.

To detect outliers in a box plot, you can look for points that lie significantly outside the range of the box and whiskers. In general, points that lie more than 1.5 times the IQR (Q3 - Q1) away from the first quartile or the third quartile may be considered outliers.

Unlike KDE plot, box plots statistically identify outliers according to **IQR**, which mean the identification of an outlier doesn’t rely only on our visual interpretation.

* + 1. **Interquartile Range:**

The interquartile range (IQR) is a descriptive statistic, often used to identify outliers. It’s a measure of the difference between the third quartile (Q3) and first quartile (Q1) of the data and an outlier is a data point that lie outside the IQR range:

Xi > Q3 + k(IQR) ∨ xi < Q1 − k(IQR) where IQR = Q3 − Q1 and k ≥ 0

From the plot above we can determine that only 6 outliers were detected out of 10 using IQR method.

* + 1. **Z-scores:**

The z-score is a parametric outlier detection method that work only on one dimensional space, and it assume a Gaussian distribution of the data. The outliers are data points that lie in tails of a distribution and therefore have an important deviation from the mean. How far depends on a set threshold Zthr for the normalized data points Zi calculated with the formula:

Graphical user interface, text, application

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

where **xi** is a data point, **μ** is the mean of all xi and σ is the standard deviation of all xi. An outlier is then a normalized data point which has an absolute value greater than **zthr**. That is: **|zi|>zthr**.

Commonly used zthr values are 2.5, 3.0 and 3.5.

* 1. **Multivariate Techniques:**
     1. **Scatter Plot:**

A scatter plot is an essential technique for analyzing the relationship between two variables. It’s useful to visualize the distribution of data points by plotting the independent variable on the x-axis and the dependent variable on the y-axis. This visualization makes it easy to identify data points that deviate significantly from the general pattern of the data.

As a result, the relation between the feature and the target is observed to be linear, with some data points deviating from the overall trend (here, I used the indices of the generated outliers to mark them). These few points that lie outside the general distribution of the data are identified as outliers.

For more than one feature, we cannot identify outliers with visualization, because the relationship between the features and the target equation is described by a hyperplane, but we can use residual analysis instead.

* + 1. **Residual Analysis:**

The basic idea behind residual analysis is to check the distribution of predictor against residual (the difference between the observed and predicted values) and look for any residuals that are far from zero. Often, these residuals indicate that the model doesn’t fit the data well, or this residual can be considered simply as an outlier.

As we can see from the result, the residuals are generally distributed around 0 with few exceptions. Specifically, there are 10 residuals that are far from the rest of the distribution.

N.B: Residual plots are primarily used to identify outliers in linear regression models, where the relationship between the independent and dependent variables is assumed to be linear. If an outlier is present in the independent variable, it will have a large impact on the line of best fit and may skew the overall results of the model or the outliers that present in the independent variable may deviate significantly from the fitted line, as the line of best fit is only determined by the independent variable, thus their residuals will be large and easily detected by using residual plots.

In multiple regression models, where there are more than one independent variable, the relationship between the independent and dependent variables **may not be as straightforward as in linear regression**. As a result, residual plots may not be as effective in detecting outliers in multiple regression models. The multiple regression model has more parameters than the simple linear regression model, so there are more assumptions to check, the presence of outliers in one predictor may have less influence than in the simple linear regression model (The presence of multiple independent variables allows for more "buffering" of the effect of an outlie).

Therefore, additional indicators like Cook's distance are often used to identify outliers in multiple regression models.

* + 1. **Cook’s distance:**

Cook’s distance is a powerful method to detect outliers. The basic idea behind this method, each time it measure the influence of each observation on model’s prediction.

For example, to compute Cook’s distance for a given observation, a linear model is fit with and without the observation, and the calculated sum of squared residuals are compared. If the RSS with the observation is much higher than the RSS without the observation, it means that the observation has a large effect on the model's predictions, and the Cook's distance will be high, and thus, the observation is an outlier.

As shown above, there’re 4 observations present a high cook’s distance with respect to the rest of the data (higher Cook’s distance means the observation has an important influence).

There are a few different ways to determine the cutoff value for Cook's distance to indicate that an observation may be influential, one common approach is to use the rule of thumb that observations with Cook's distance greater than 4/n are considered influential, where n is the number of observations. It is a common rule of thumb, but it should be used with caution and be verified by the domain expert.

The following stem plot shows that 6 observations have a Cook's distances higher than the threshold.

1. **How to deal with outliers**

Removing outliers is one option for dealing with them in a machine learning model, but it's not the only option and it's not always the best choice. It's important to be cautious when removing outliers, because they might be valid observations that are important for understanding the underlying relationship in the data. If removing outliers is not an option, then using robust models that can handle outliers or applying transformations to the data (such as logarithmic or scaling) are other options that can help make the model more robust to the presence of outliers.

**This work demonstrate different univariate and multivariate techniques to detect outliers on continuous variables with linear relationship.**