1.3.5.17. Detection of Outliers

Introduction

An outlier is an observation that appears to deviate markedly from other observations in the sample.

Identification of potential outliers is important for the following reasons.

1. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

2. In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation. However, if the data contains significant outliers, we may need to consider the use of robust statistical techniques.

Labeling, Accomodation, Identification

Iglewicz and Hoaglin distinguish the three following issues with regards to outliers.

1. outlier labeling - flag potential outliers for further investigation (i.e., are the potential outliers erroneous data, indicative of an inappropriate distributional model, and so on).

2. outlier accomodation - use robust statistical techniques that will not be unduly affected by outliers. That is, if we cannot determine that potential outliers are erroneous observations, do we need modify our statistical analysis to more appropriately account for these observations?

3. outlier identification - formally test whether observations are outliers.

This section focuses on the labeling and identification issues.

Normality Assumption

Identifying an observation as an outlier depends on the underlying distribution of the data. In this section, we limit the discussion to univariate data sets that are assumed to follow an approximately normal distribution. If the normality assumption for the data being tested is not valid, then a determination that there is an outlier may in fact be due to the non-normality of the data rather than the prescence of an outlier.

For this reason, it is recommended that you generate a normal probability plot of the data before applying an outlier test. Although you can also perform formal tests for normality, the prescence of one or more outliers may cause the tests to reject normality when it is in fact a reasonable assumption for applying the outlier test.

In addition to checking the normality assumption, the lower and upper tails of the normal probability plot can be a useful graphical technique for identifying potential outliers. In particular, the plot can help determine whether we need to check for a single outlier or whether we need to check for multiple outliers.

The box plot and the histogram can also be useful graphical tools in checking the normality assumption and in identifying potential outliers.

Single Versus Multiple Outliers

Some outlier tests are designed to detect the prescence of a single outlier while other tests are designed to detect the prescence of multiple outliers. It is not appropriate to apply a test for a single outlier sequentially in order to detect multiple outliers.

In addition, some tests that detect multiple outliers may require that you specify the number of suspected outliers exactly.

Masking and Swamping

Masking can occur when we specify too few outliers in the test. For example, if we are testing for a single outlier when there are in fact two (or more) outliers, these additional outliers may influence the value of the test statistic enough so that no points are declared as outliers.

On the other hand, swamping can occur when we specify too many outliers in the test. For example, if we are testing for two or more outliers when there is in fact only a single outlier, both points may be declared outliers (many tests will declare either all or none of the tested points as outliers).

Due to the possibility of masking and swamping, it is useful to complement formal outlier tests with graphical methods. Graphics can often help identify cases where masking or swamping may be an issue. Swamping and masking are also the reason that many tests require that the exact number of outliers being tested must be specified.

Also, masking is one reason that trying to apply a single outlier test sequentially can fail. For example, if there are multiple outliers, masking may cause the outlier test for the first outlier to return a conclusion of no outliers (and so the testing for any additional outliers is not performed).

Z-Scores and Modified Z-Scores

The Z-score of an observation is defined as

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

with $\bar{Y}$ and $\boldsymbol{s}$ denoting the sample mean and sample standard deviation, respectively. In other words, data is given in units of how many standard deviations it is from the mean.

Although it is common practice to use Z-scores to identify possible outliers, this can be misleading (partiucarly for small sample sizes) due to the fact that the maximum Z-score is at most $(n-1)/\sqrt{n}$

Iglewicz and Hoaglin recommend using the modified Z-score

$$M_i = \frac{0.6745(x_i - \tilde{x})}{\mathrm{MAD}}$$

with MAD denoting the median absolute deviation and $\tilde{x}$ denoting the median.

These authors recommend that modified Z-scores with an absolute value of greater than 3.5 be labeled as potential outliers.

Formal Outlier Tests

A number of formal outlier tests have proposed in the literature. These can be grouped by the following characteristics:

- What is the distributional model for the data? We restrict our discussion to tests that assume the data follow an approximately normal distribution.
- Is the test designed for a single outlier or is it designed for multiple outliers?
- If the test is designed for multiple outliers, does the number of outliers need to be specified exactly or can we specify an upper bound for the number of outliers?

The following are a few of the more commonly used outlier tests for normally distributed data. This list is not exhaustive (a large number of outlier tests have been proposed in the literature). The tests given here are essentially based on the criterion of "distance from the mean". This is not the only criterion that could be used. For example, the Dixon test, which is not discussed here, is based a value being too large (or small) compared to its nearest neighbor.

1. Grubbs' Test - this is the recommended test when testing for a single outlier.

2. Tietjen-Moore Test - this is a generalization of the Grubbs' test to the case of more than one outlier. It has the limitation that the number of outliers must be specified exactly.

3. Generalized Extreme Studentized Deviate (ESD) Test - this test requires only an upper bound on the suspected number of outliers and is the recommended test when the exact number of outliers is not known.

Lognormal Distribution

The tests discussed here are specifically based on the assumption that the data follow an approximately normal disribution. If your data follow an approximately lognormal distribution, you can transform the data to normality by taking the logarithms of the data and then applying the outlier tests discussed here.

Further Information

Iglewicz and Hoaglin provide an extensive discussion of the outlier tests given above (as well as some not given above) and also give a good tutorial on the subject of outliers. Barnett and Lewis provide a book length treatment of the subject.

In addition to discussing additional tests for data that follow an approximately normal distribution, these sources also discuss the case where the data are not normally distributed.