Check your outliers! An accessible introduction to identifying statistical outliers with R and *easystats*

(Alt title:) Statistical Outliers: Univariate, Multivariate, and Model-Based

Abstract

xyz

Introduction

Statistical outliers need to be addressed because they can significantly affect our statistical models.

Leys et al. (2019) distinguish between error outliers, interesting outliers, and random outliers.

Current recommendations

Current guidelines recommend the following:

We note also that ideally the handling of outliers should be preregistered a priori (Leys et al., 2019).

Univariate Outliers

When using t-tests for example.

For univariate outliers, it is recommended (Leys et al., 2013, 2019) to use the Median Absolute Deviation (MAD). In easystats' `check_outliers()`,one can use this approach with `method = "zscore_robust"`.

Example:

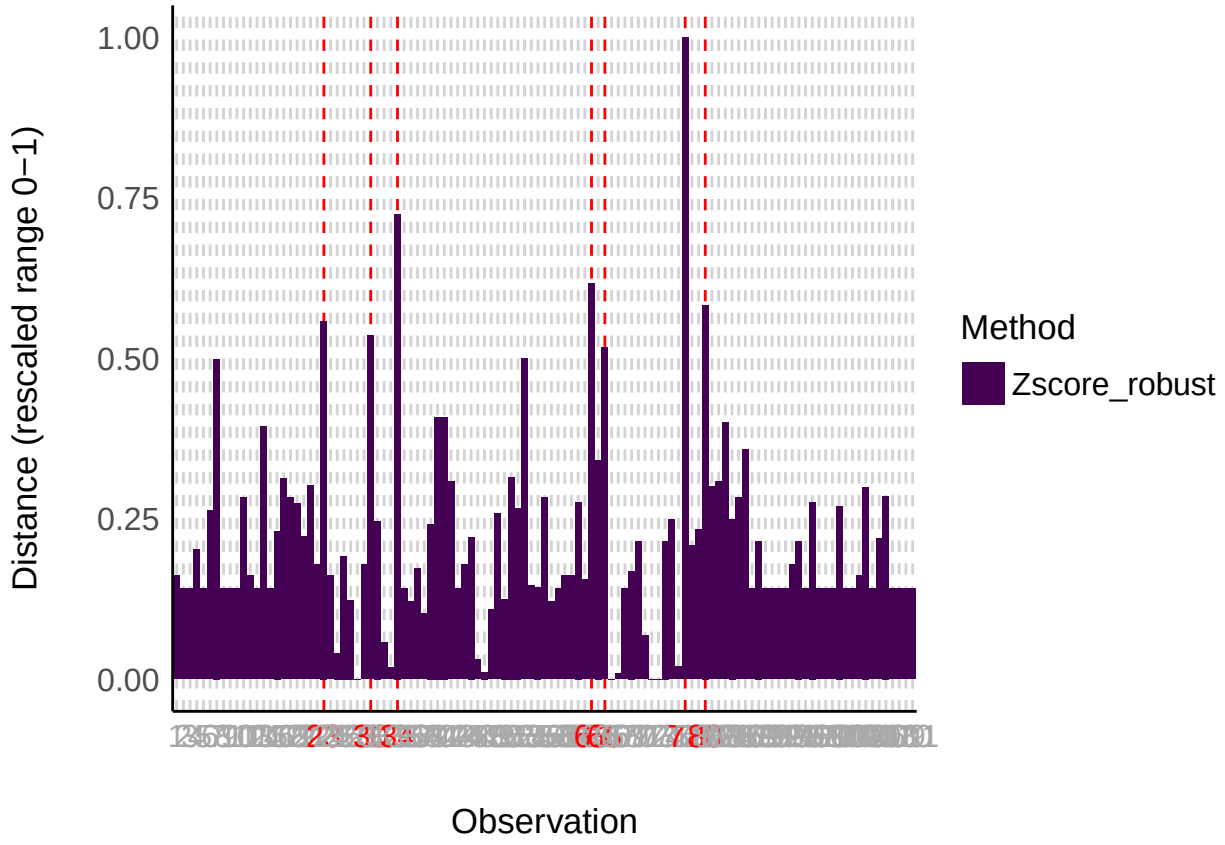
```
library(performance)
data <- na.omit(airquality)

x <- check_outliers(data, method = "zscore_robust")
x
```

```
## 7 outliers detected: cases 23, 30, 34, 63, 65, 77, 80.
## - Based on the following method and threshold: zscore_robust (3.09).
## - For variables: Ozone, Solar.R, Wind, Temp, Month, Day.
##
## ----------------------------------------------------------------------------
## Outliers per variable (zscore_robust):
##
## $Ozone
##    Row Distance_Zscore_robust
## 23  23               3.332778
## 34  34               4.126296
## 63  63               3.610509
## 65  65               3.134398
## 77  77               5.435602
```

```
## 80  80               3.451806
##
## $Wind
##    Row Distance_Zscore_robust
## 30  30               3.225825
```

```
library(see)
plot(x)
```



Multivariate Outliers

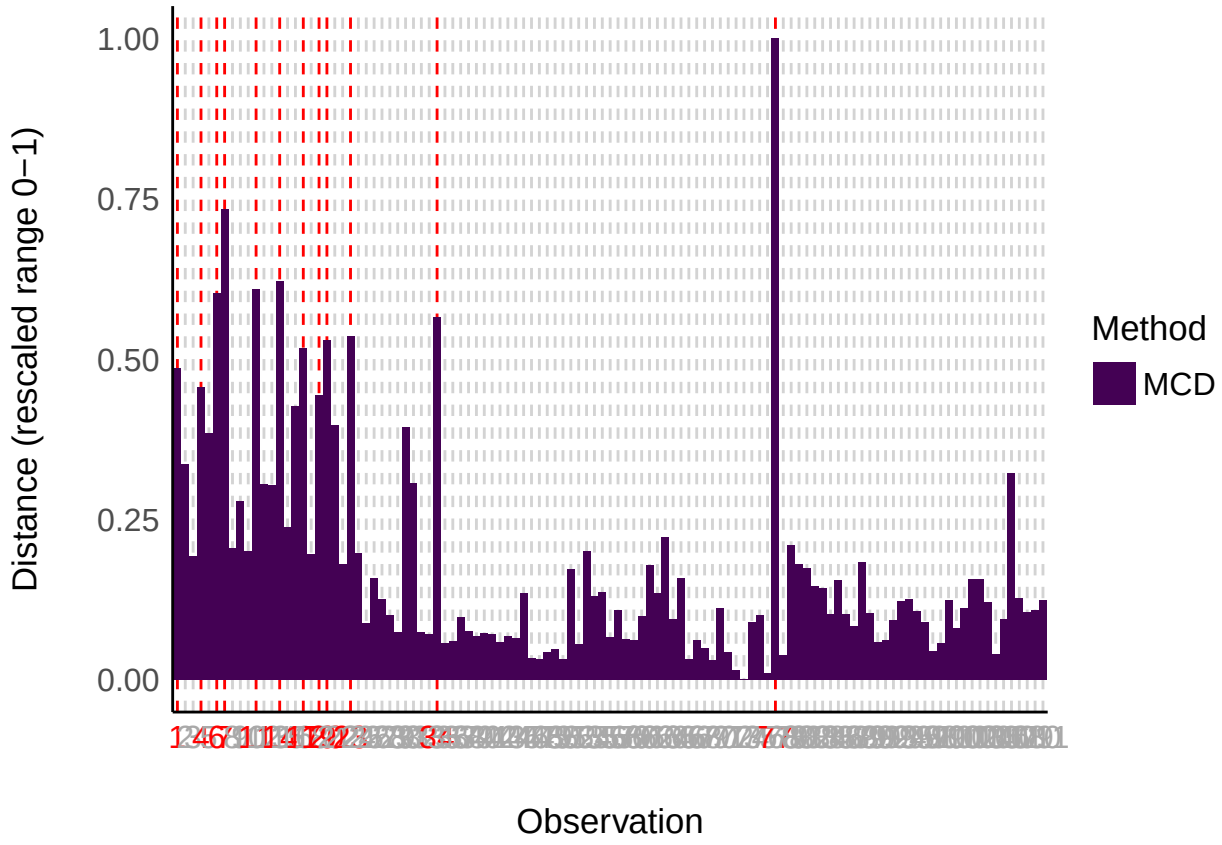
For multivariate outliers, it is recommended (Leys et al., 2019) to use the Minimum Covariance Determinant (MCD).

Example:

```
x <- check_outliers(data, method = "mcd")
x
```

```
## 12 outliers detected: cases 1, 4, 6, 7, 11, 14, 17, 19, 20, 23, 34, 77.
## - Based on the following method and threshold: mcd (22.46).
## - For variables: Ozone, Solar.R, Wind, Temp, Month, Day.
```

```
plot(x)
```



Model-Based Outliers

When using linear regression models for example.

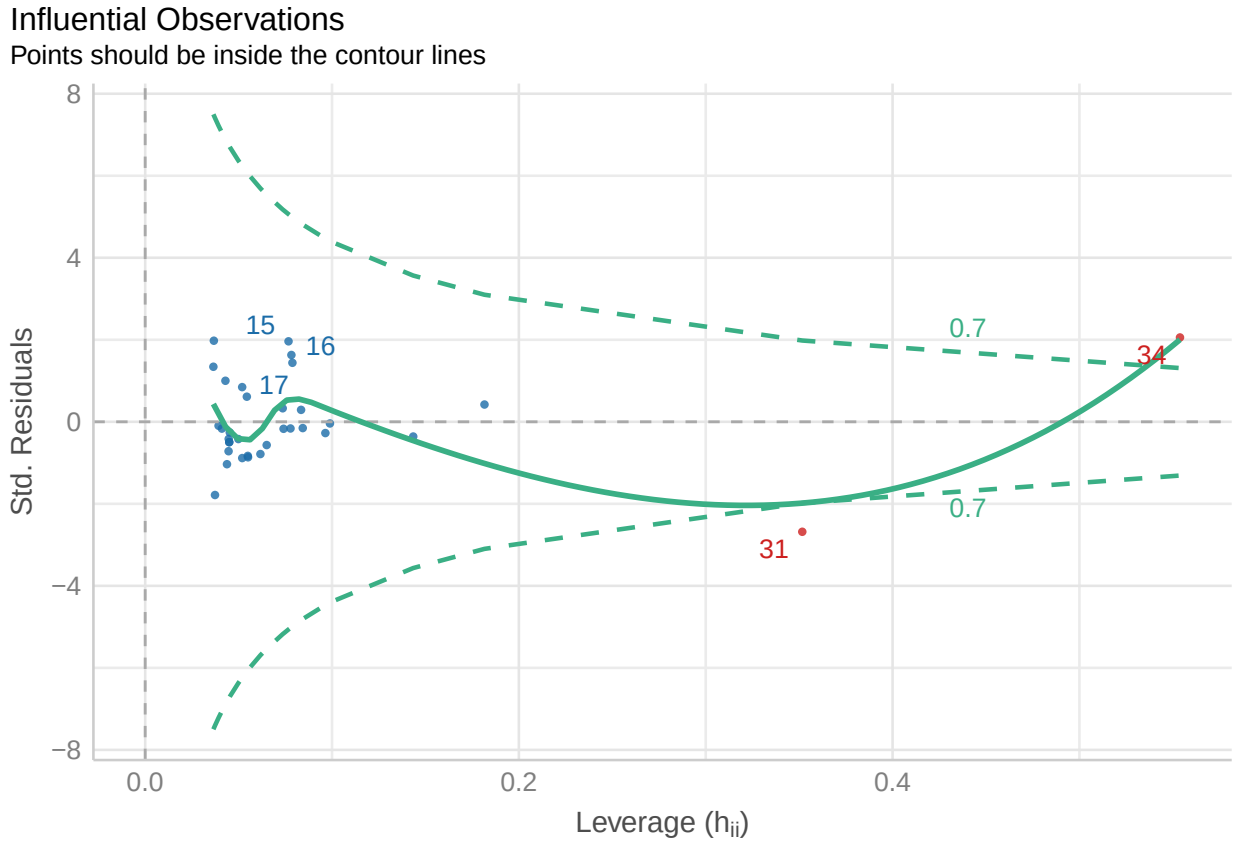
Example:

```
# create some fake outliers
data <- rbind(mtcars, 42, 55)

# fit model with outliers
model <- lm(disp ~ mpg + hp, data = data)
x <- check_outliers(model, method = "cook")
x
```

```
## 2 outliers detected: cases 31, 34.
## - Based on the following method and threshold: cook (0.81).
## - For variable: (Whole model).
```

```
plot(x)
```



Multiple methods

An alternative approach suggested by easystats is to combine several methods (model-based, univariate, and multivariate). This approach computes a composite outlier score, made of the average of the binary (0 or 1) results of each method. It represents the probability of each observation to be classified as an outlier by at least one method. The default decision rule classifies rows with composite outlier scores superior or equal to 0.5 as outlier observations (i.e., that were classified as outliers by at least half of the methods).

Example:

```
x <- check_outliers(data, method = c(
  "zscore", "IQR", "mahalanobis", "mahalanobis_robust", "mcd", "ics", "optics", "lof"))
x
```

```
## 2 outliers detected: cases 33, 34.
## - Based on the following methods and thresholds: zscore (3.09),
##   mahalanobis (31.26), mahalanobis_robust (31.26), mcd (31.26), ics
##   (0.001), optics (22), lof (0.001).
## - For variables: mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb.
## Note: Outliers were classified as such by at least half of the selected methods.
##
## ----------------------------------------------------------------------------
## The following observations were considered outliers for two or more variables
## by at least one of the selected methods:
##
##    Row n_Zscore n_Mahalanobis_robust          n_MCD          n_ICS
## 1   34        9       (Multivariate) (Multivariate) (Multivariate)
```

```
## 2   33        7       (Multivariate) (Multivariate) (Multivariate)
## 3    9        0       (Multivariate) (Multivariate)              0
## 4   19        0       (Multivariate)              0              0
## 5   27        0       (Multivariate) (Multivariate)              0
## 6   28        0       (Multivariate) (Multivariate)              0
## 7   29        0       (Multivariate)              0              0
## 8   30        0       (Multivariate) (Multivariate)              0
## 9   31        0       (Multivariate) (Multivariate) (Multivariate)
## 10   8        0                    0 (Multivariate)              0
## 11  21        0                    0 (Multivariate)              0
##             n_LOF
## 1  (Multivariate)
## 2  (Multivariate)
## 3               0
## 4               0
## 5               0
## 6               0
## 7               0
## 8               0
## 9               0
## 10              0
## 11              0
```

An example sentence for reporting the usage of the composite method could be:

> Based on a composite outlier score (see the 'check_outliers' function in the 'performance' R package; Lüdecke et al., 2021) obtained via the joint application of multiple outliers detection algorithms (Z-scores, Iglewicz, 1993; Interquartile range (IQR); Mahalanobis distance, Cabana, 2019; Robust Mahalanobis distance, Gnanadesikan and Kettenring, 1972; Minimum Covariance Determinant, Leys et al., 2018; Invariant Coordinate Selection, Archimbaud et al., 2018; OPTICS, Ankerst et al., 1999; Isolation Forest, Liu et al. 2008; and Local Outlier Factor, Breunig et al., 2000), we excluded one participant that was classified as outlier by at least half of the methods used.

Conclusion

This is the conclusion.

References

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*. https://doi.org/10.5334/irsp.289

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/https://doi.org/10.1016/j.jesp.2013.03.013