Article

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

Rémi Thériault[1,*], Mattan S. Ben-Shachar[2], Indrajeet Patil[3], Daniel Lüdecke[4], Brenton M. Wiernik[5], Dominique Makowski[6]

1 Department of Psychology, Université du Québec à Montréal, Montréal, Québec, Canada;
2 Independent Researcher;
3 Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany;
4 Institute of Medical Sociology, University Medical Center Hamburg-Eppendorf, Germany;
5 Independent Researcher, Tampa, FL, USA;
6 School of Psychology, University of Sussex, Brighton, UK;
* Correspondence: theriault.remi@courrier.uqam.ca.

Version April 17, 2023

Simple Summary: The *{performance}* package from the *easystats* ecosystem makes it easy to diagnose outliers in R and according to current best practices thanks to the `check_outiers()` function.

Abstract: Beyond the challenge of keeping up-to-date with current best practices regarding the diagnosis and treatment of outliers, an additional difficulty arises concerning the mathematical implementation of the recommended methods. In this paper, we provide an overview of current recommandations and best practices and demonstrate how they can easily and conveniently be implemented in the R statistical computing software, using the *{performance}* package of the *easystats* ecosystem. We cover univariate, multivariate, and model-based statistical outlier detection methods, their recommended threshold, standard output, and plotting methods. We conclude with recommendations on the handling of outliers: the different theoretical types of outliers, whether to exclude or winsorize them, and the importance of transparency.

Keywords: univariate outliers; multivariate outliers; robust detection methods; R; easystats

---

1. Introduction

Real-life data often contain observations that can be considered *abnormal* when compared to the main population. The cause of it—be it because they belong to a different distribution (originating from a different generative process) or simply being extreme cases, statistically rare but not impossible—can be hard to assess, and the boundaries of "abnormal" are hard to define.

Nonetheless, the improper handling of these outliers can substantially affect statistical model estimations, biasing effect estimations and weakening the models' predictive performance. It is thus essential to address this problem in a thoughtful manner. Yet, despite the existence of established recommendations and guidelines, many researchers still do not treat outliers in a consistent manner, or do so using inappropriate strategies [1,2].

One possible reason is that researchers are not aware of the existing recommendations, or do not know how to implement them using their analysis software. In this paper, we show how to follow current best practices for automatic and reproducible statistical outlier detection (SOD) using R and the *{performance}* package [3], which is part of the *easystats* ecosystem of packages that build an R framework for easy statistical modeling, visualization, and reporting [4].