Outler Detection using Regression in R

Ludwig-Maximilians-Universität München

Fakultät für Mathematik, Informatik und Statistik

Seminar "'Ausreisser- und Anomaliedetektion"' von Prof. Dr. Christian Heumann

Hausarbeit

30. Juni 2018

Autor
Samuel Lippl

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

Samuel Lippl, 27.08.2018

Abstract

This report will provide an overview over outlier detection in R. It starts by discussing some general principles of outlier detection. Linear methods and their nonlinear extensions are presented next. Along the general methods, examples of application and an appropriate methodology in R are introduced. The report will be concluded by a discussion of method evaluation as well as a comparison of the different introduced algorithms.

Contents

1 | Introduction: neurons, models and outliers

Statistics, like all science, are man-made. Acceptance of their presuppositions depends on human intuition of how to model connections between variables, what distributional assumptions make sense – and what constitutes an outlier. In many cases, these presuppositions are straightforward and we seldomly question them. This makes sense, of course. It is impossible to make progress if we questioned our most basic assumptions all the time. Before going on to outlier analysis, however, let us briefly revisit the beginnings of statistics. Francis Galton, an early pioneer of regression analysis, discovered a curious connection when analyzing body height of parents and their children.

Whereas taller parents had taller children, the offspring of particularly tall parents was smaller than them wheareas smaller parents' children had a relatively large body height. In an attempt to characterize the connection he drew a line with a slope of 2/3. (Galton, 1889, p. 96f.) Regression analysis has come a far way since then. His successors determined the more rigorous method of least squares to compute linear models. The slope determined by this method is 0.65; Galton's visual estimation[1] had therefore been quite accurate.

[1] reproduced from Galton (1889, p. 96)

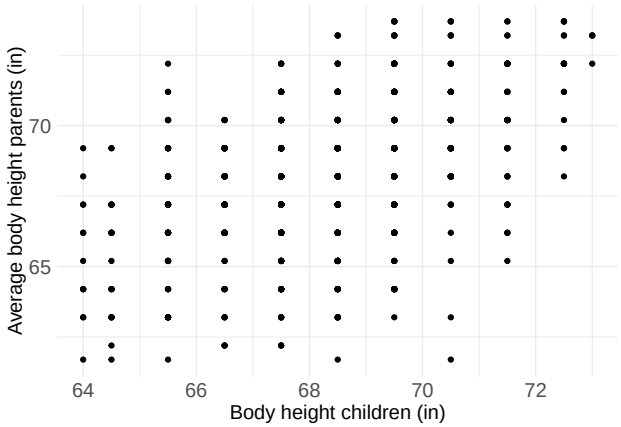

Figure 1.1: Galton's heredity data in the `psych` package (Revelle, 2018, dataset `galton`)

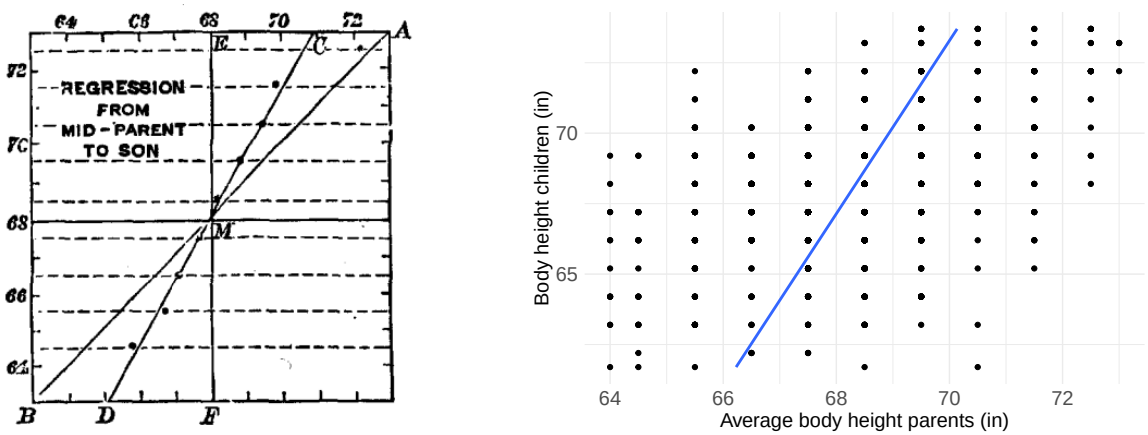

Figure 1.2: Left: Galton's original estimation (the line connecting C and D). Right: Least Squares linear model. Note that the x-axis now represents the parents' height wheareas the childrens' height is drawn on the y-axis.

2 | Literature

Here is a review of existing methods.

3 | Methods

We describe our methods in this chapter.

4 | Applications

Some significant applications are demonstrated in this chapter.

4.1 Example one

4.2 Example two

5 | Final Words

We have finished a nice book.

Bibliography

Galton, F. (1889). Natural Inheritance. Natural Inheritance. Macmillan.

Revelle, W. (2018). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois.