ETH zürich

ETH Library

Identifizierung von Ausreissern in eindimensionalen gewichteten Umfragedaten

Journal Article

Author(s):
Sandqvist, Anna P.

Publication date:
2016-06-16

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000121681

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
KOF Analysen 2016(2)

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

ETH zürich

KOF Analysen

Spezialanalysen:

Identifizierung von Ausreissern in eindimensionalen gewichteten Umfragedaten

2016, Nr. 2, Sommer – SA 2

KOF Konjunkturforschungsstelle

KOF

ETH Zürich
KOF Konjunkturforschungsstelle
LEE G 116
Leonhardstrasse 21
8092 Zürich

Tel. +41 44 632 42 39
Fax +41 44 632 12 18
www.kof.ethz.ch
#KOFETH

IDENTIFIZIERUNG VON AUSREISSERN IN EINDIMENSIONALEN GEWICHTETEN UMFRAGEDATEN

ANNA PAULIINA SANDQVIST

Zusammenfassung: Ausreisser und einflussreiche Beobachtungen sind bei verschiedenen Statistiken, Datenanalysen sowie auch bei Umfragedaten oft ein Problem. Vor allem, wenn die Verteilung der Daten asymmetrisch ist oder starke Ränder («Tails») hat, ist die Ausreisseruntersuchung nicht einfach. Ferner können sich für periodische Daten und für Datensätze mit mehreren Teilproben die (empirischen) Verteilungseigenschaften von einem Datensatz zum anderen verändern und daher kann die Erkennung der Ausreisser schwierig werden, da man die Identifizierungsmethode eventuell jedes Mal neu anpassen muss. In diesem Artikel untersuchen wir, wie sich verschiedene nicht-parametrische Ausreisser-Methoden für (grössen)gewichtete quantitative Umfragedaten (Veränderungsraten) verhalten. Wir schlagen neue bzw. angepasste Methoden vor, die besser mit verschiedenen Verteilungsmodellen funktionieren. Da das Erkennen und die Behandlung von Ausreissern oft mit viel Subjektivität verbunden sind, streben wir eine objektive Methode an, in der möglichst wenige Parameter im Voraus definiert werden müssen. Wir präsentieren eine Simulations-Analyse, um die Methoden unter verschiedenen Datensatzspezifikationen zu untersuchen.

Abstract: Outliers and influential observations are a frequent concern in all kind of statistics, data analysis and survey data. Especially, if data are asymmetrically distributed or heavy-tailed, outlier detection is not clear-cut. Even more, for periodic data and data with multiple subsets, in which the distributional characteristics of each data sample may differ, outlier detection is challenging as the method used may need to be adjusted each time. In this paper we examine various non-parametric outlier detection approaches for (size-) weighted growth rates from surveys and propose new respectively modified methods which can account better for non-normal data and particularly for altering levels of dispersion and asymmetry. As outlier detection (and treatment) involves in practice a lot of subjectivity, we pursue an approach in which as few as possible parameters need to be defined. We conduct a simulation study to compare these methods under various models.

JEL Classification Codes: C14

Keywords: Outlier detection, skewness, size-weight, periodic surveys

1 EINLEITUNG

Als Ausreisser betrachtet man üblicherweise extreme Beobachtungen, die weit weg von anderen Datenpunkten liegen (siehe zum Beispiel, Barnett und Lewis (1994)). Chambers (1986) war der erste, der zwischen *repräsentativen* und *nicht-repräsentativen* Ausreissern unterschied. Die ersteren sind Beobachtungen mit richtigen Werten und werden nicht als einzigartig betrachtet, während die letzteren als Elemente mit inkorrekten Werten oder aus anderen Gründen als einzigartig wahrgenommen werden. Die meisten Ausreisseruntersuchungen konzentrieren sich auf repräsentative Beobachtungen, da die nicht-repräsentativen Werte schon bei der Aufbereitung der Daten korrigiert bzw. bereinigt sein sollten.

Die Hauptmotivation für die Ausreisseridentifikation ist die Tatsache, dass Ausreisser Schätzungen verzerren können; die Schätzer können sehr unterschiedliche Ergebnisse liefern je nachdem, ob die Ausreisser berücksichtigt werden oder nicht. Datenpunkte

mit beachtlichem Einfluss auf die Schätzer werden als *einflussreiche Beobachtungen* bezeichnet. Sie sollten identifiziert und verarbeitet werden, um unverzerrte Ergebnisse sicherzustellen. Quantitative Daten aus Nicht-Zufallsstichproben, wie zum Beispiel Wachstumsraten, werden oft mit der Grösse der Beobachtungen gewichtet. Somit könnte es vorkommen, dass eine nicht gewichtete Beobachtung an sich kein Ausreisser ist, sie könnte aber durch ihre Gewichtung einflussreich werden.

Wenn der datengenerierende Prozess nicht bekannt ist, was meistens bei Umfragedaten der Fall ist, sind parametrische Ausreissertests nicht anwendbar. In dieser Analyse untersuchen wir nicht-parametrische Methoden, um einflussreiche Beobachtungen in quantitativen Umfragedaten, besonders im Falle grössengewichteter Wachstumsraten, zu erkennen. Solche Datenquellen sind zum Beispiel die KOF-Detailhandelsumfrage, die KOF-Investitionsumfrage und die KOF-Gastgewerbeumfrage. Da die Eigenschaften der Verteilung des Datensatzes (Wachstumsraten) von Periode zu Periode und/oder von Teilprobe zu Teilprobe ändern können, analysieren wir die Identifizierung von Ausreissern in Datensätzen mit unterschiedlichem Mass an Streuung und Schiefe. Weil das Erkennen und die Behandlung von Ausreissern oft mit viel Subjektivität verbunden ist, streben wir eine objektive Methode an, in der möglichst wenige Parameter im Voraus definiert werden müssen, die aber trotzdem mit Datensätzen aus verschiedenen Verteilungsmodellen angewendet werden können. Wir präsentieren zwei neue bzw. angepasste Methoden, die diese Punkte berücksichtigen.

Um zu untersuchen, wie sich unsere neuen Methoden im Vergleich zu der Methode von Hidiroglou und Berthelot (1986) (HB-Methode) verhalten, führen wir eine Simulationsstudie durch. Die Ergebnisse deuten darauf hin, dass unsere neuen Methoden mit schiefen Datensätzen besser funktionieren. Dies hängt damit zusammen, dass die HB-Methode dazu tendiert, am linken Rand zu sensitiv zu sein und daher zu viele (kleine) Ausreisser identifiziert. Dies könnte eventuell vermieden werden, wenn man die Parameterwerte der HB-Methode anpasst. Wir stellen jedoch auf eine Methode ab, die auch ohne Anpassungen der Parameter angewendet werden kann, wenn das Verteilungsmodell sich ändert.

2 IDENTIFIZIERUNG VON AUSREISSERN IN UNIVARIATEN DATEN

Zahlreiche Methoden zur Erkennung von Ausreissern sind bereits präsentiert worden. Jedoch existiert nach wie vor keine Generallösung für alle Fälle. Zudem ist die Wahl der Schätzer für Lagemass und Streuung für die meisten Ausreisser-Methoden entscheidend.

Das arithmetische Mittel und die Standardabweichung sind die am meisten benutzten Schätzer für Lagemass und Streuung. Jedoch sollte man mit diesen Schätzern vorsichtig umgehen, wenn mehrere Ausreisser potenziell vorhanden sind, da sie sehr anfällig für extreme Datenpunkte sind, wodurch das Erkennen von Ausreissern eventuell behindert wird. Die Alternative ist die Anwendung robusterer Schätzer wie zum Beispiel median und MAD (Median Absolute Deviation). Zudem haben Rousseeuw und Croux (1993) zwei Alternativen für MAD vorgeschlagen, welche MAD ähnlich sind, jedoch nicht auf der Annahme von symmetrisch verteilten Daten basieren. Ihr Q_n-Schätzer ist wie folgt definiert: $Q_n = d\{|x_i - x_j|; i < j\}_{(k)}$ wobei d eine Konstante ist und $k = \binom{h}{2} \approx \binom{n}{2}/4$, wobei $h = \left[\frac{n}{2}\right] + 1$.

Zahlreiche nicht-parametrische wie auch parametrische Ausreissertests sind bisher vorgeschlagen geworden; siehe dazu Barnett und Lewis (1994), Grubbs (1950), Dixon (1951). Zu den häufig verwendeten nicht-parametrischen Methoden gehören vor allem regelbasierte Methoden wie die relative Entfernung. *Die relative Entfernung* ist definiert als Betrag aus der Entfernung der Beobachtung y_i zu Lagemass l dividiert durch Skalenmass: $rd_i = \frac{|y_i - l|}{s}$. Wenn diese Teststatistik einen bestimmten Wert k, der vom Datenanalyst oder Forscher definiert ist, überschreitet, dann wird diese Beobachtung als Ausreisser identifiziert. Auch ein Toleranzintervall in Form von $[l - k_l s, l + k_h s]$ kann angewendet werden, wobei die Werte für k_l und k_h von sich aus unterscheiden können, um eine bessere Anpassung für die Schiefe des Datensatzes zu ermöglichen.

3 FRAGESTELLUNG UND METHODEN

Wenn y_j die Wachstumsrate und w_j das Grössengewicht der Firma i ist, kann der einfache Punktschätzer des gewichteten Mittelwerts einer Stichprobe mit n Beobachtungen wie folgt definiert werden

$$\mu_n = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

Unser Ziel ist es, Beobachtungen zu identifizieren, die einen starken Einfluss auf diesen Schätzer haben. Da sich die Eigenschaften der Verteilung von y_i sich von Periode zu Periode bzw. von Teilprobe zu Teilprobe verändern können, untersuchen wir die Identifizierung von Ausreissern mit unterschiedlichen Verteilungsmodellen.

Im Allgemeinen gibt es drei verschiedene Möglichkeiten, wie wir mit der Schiefe der Verteilung bei der Erkennung von Ausreissern umgehen können: Transformation der Beobachtungen, so dass sie ungefähr symmetrisch verteilt sind, verschiedene Skalenmasse auf der jeweiligen Seite des Medians anzuwenden oder einen Schätzer für Dispersion zu benutzen, der auch im Falle von asymmetrisch verteilten Daten robust ist.

3.1 Basismethode

Da mit der Methode von Hidiroglou und Berthelot (1986) die Grösse der Beobachtungen berücksichtigt werden kann, verwenden wir diese Methode als Basismethode. In einem ersten Schritt wenden sie folgende Transformation an

$$s_i = \begin{cases} 1 - q_{0.5}/r_i & \text{if } 0 < r_i < q_{0.5} \\ r_i/q_{0.5} - 1 & \text{if } r_i \geq q_{0.5} \end{cases}$$

wobei $r_i = x_{i,t}/x_{i,t-1}$ und $q_{0.5}$ der Median der r_i ist. Somit ist die Hälfte der transformierten Werte kleiner als null und die andere Hälfte grösser als null. Jedoch stellt diese Transformation keine symmetrische Verteilung der transformierten Beobachtungen sicher. Im zweiten Schritt fliesst die Grösse der Beobachtungen ein

$$E_i = s_i\{max([x_i(t), x_i(t-1)]\}^V$$

wobei der Parameter $0 \leq V \leq 1$ bestimmt, wie stark die Grösse der Beobachtungen berücksichtigt werden sollte. Danach wird das folgende Akzeptanzintervall angewendet

$$[E_{med} - CD_{Q1}, E_{med} + CD_{Q3}]$$

wobei $D_{Q1} = max\{E_{med} - E_{Q1}, |aE_{med}|\}$ und

$D_{Q3} = max\{E_{Q3} - E_{med}, |aE_{med}$ und $|aE_{med}|$

der Median von E_i's, E_{Q1} und E_{Q3} sind das erste und dritte Quartil und der Parameter a nehmen meistens den Wert von 0.05 an. Term $|aE_{med}|$ soll sicherstellen, dass das Intervall nicht zu klein wird, falls die Beobachtungen um einen bestimmten Wert angehäuft wären. Der Parameter C bestimmt die Breite des Intervalls. Jedoch müssen die Werte für die Parameter V, a und vor allem C mehr oder weniger willkürlich bestimmt werden. Wir benutzen diese Methode als Basismethode mit folgenden Parameterwerten a=0.05 und C=8.

3.2 Rechts-links Methode

Die erste neue Methode, die wir vorschlagen, basiert auf der Idee verschiedener Skalenmasse der jeweiligen Seite des Medians. Hubert und Van Der Veeken (2008) schlagen ein Mass für *adjusted outlyingness (AO)* vor, wobei sie individuelle Masse für die Streuung der jeweiligen Seite des Medians anwenden

$$AO_i = \begin{cases} \frac{y_i - q_{0.5}}{w_u - q_{0.5}} & \text{if } y_i > q_{0.5} \\ \frac{q_{0.5} - y_i}{q_{0.5} - w_l} & \text{if } y_i < q_{0.5} \end{cases}$$

wobei w_l und w_u der untere und obere whisker des angepassten Boxplots («adjusted Boxplot») von M. Hubert und Vandervieren (2008) sind

$$w_l = q_{0.25} - 1.5^{-4MC} IQR; w_u = q_{0.75} + 1.5^{3MC} IQR \quad \text{for } MC \geq 0$$

$$w_l = q_{0.25} - 1.5^{-3MC} IQR; w_u = q_{0.75} + 1.5^{4MC} IQR \quad \text{for } MC < 0$$

wobei MC für *medcouple* steht, ein Mass für die Schiefe; vorgeschlagen von Brys, Hubert und Struyf (2004)

$$MC_n = \underset{y_i < med_n < y_j}{med} h(y_i, y_j)$$

wobei med_n der Median (der Stichprobe) ist und für alle $y_i \neq y_j$

$$h(y_i, y_j) = \frac{(y_j - med_n) - (med_n - y_i)}{y_j - y_i}$$

Jedoch kann der Nenner in AO sehr klein werden, falls die Beobachtungen auf einer Seite des Medians zusammengeballt sind. Dann würden mehr Beobachtungen als Ausreisser identifiziert als es eigentlich der Fall ist. Der Nenner kann aber auch recht gross werden, falls ein (oder mehrere) sehr deutliche extreme Datenpunkte in der Stichprobe enthalten sind. In diesem Fall würden wahrscheinlich die Ausreisser nicht richtig identifiziert. Daher passen wir den Nenner wie folgt an, um diese Fälle vermeiden zu können.

$$rl_i = \begin{cases} \frac{y_i - q_{0.5}}{min(max(w_u - q_{0.5}, range(y)/n_y^{0.6}), IQR(y))} & \text{if } y_i \geq q_{0.5} \\ \frac{y_i - q_{0.5}}{min(max(q_{0.5} - w_l, range(y)/n_y^{0.6}), IQR(y))} & \text{if } y_i < q_{0.5} \end{cases}$$

Im nächsten Schritt wird die Grösse der Beobachtungen einbezogen

$$rl_i^w = rl_i w_i^V$$

wobei der Parameter $0 \leq V \leq 1$ den Einfluss des Grössengewichts bestimmt. Danach wird ein durch den Interquartilabstand definiertes Intervall verwendet: $[q_{0.25} - 3\ IQR, q_{0.75} + 3\ IQR]$. Wir beziehen uns auf diese Methode als *rl-methode*.

3.3 Q_n-Methode

Die zweite neue Methode, die wir präsentieren, basiert auf der relativen Entfernung. Wir verwenden den Q_n-Schätzer von Rousseeuw und Croux (1993) als Schätzer für die Streuung. In einem ersten Schritt wird die Entfernung zum Median durch Q_n (mit einem konstanten Faktor von 1) dividiert

$$Q_i = \frac{y_i - median(y)}{Q_n(y)}$$

Dieser Wert wird dann mit der Firmengewichtung multipliziert und der Parameter V integriert, um die Wichtigkeit der Gewichtung steuern zu können

$$Q_i^w = Q_i w_i^V$$

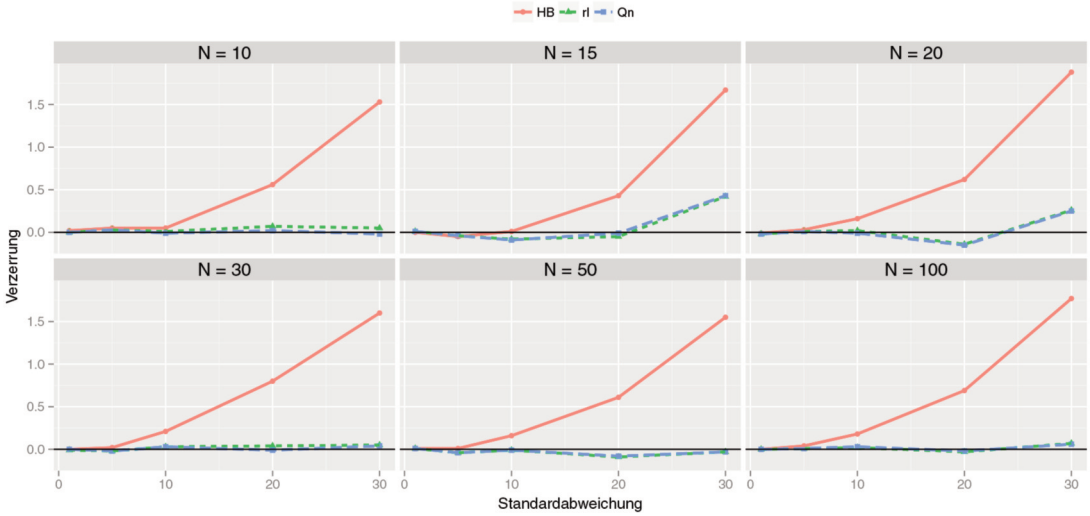
Um die Ausreisser erkennen zu können, wird ein auf dem Interquartilabstand $[q_{0.25} - 3\ IQR, q_{0.75} + 3\ IQR]$ basierendes Intervall angewendet.

4 SIMULATIONSSTUDIE

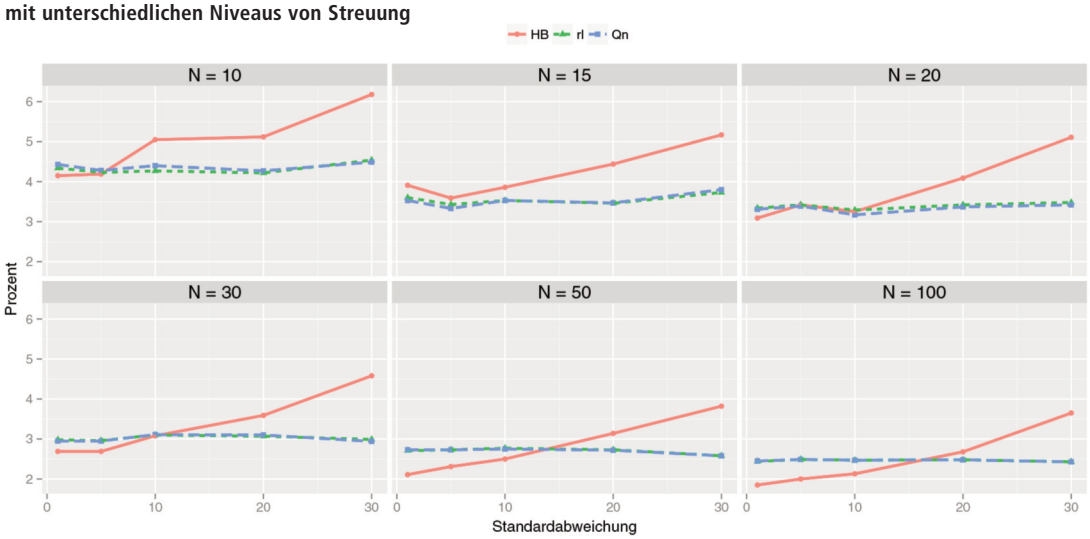
Um untersuchen zu können, wie diese Methoden unter verschiedenen Verteilungsmodellen funktionieren, führen wir eine Simulationsstudie durch. Für jede Simulation generieren wir 1000 Stichproben mit einer Anzahl von Beobachtungen gleich 10, 15, 20, 30, 50 und 100. Die Grössengewichte w_j werden aus einer Exponentialverteilung mit Parameter λ=1/20 generiert und gerundet auf ganze Zahlen. Diese Grössengewichte werden meistens aus Umsatz- oder Beschäftigtenzahlen abgeleitet. Insgesamt generieren wir drei verschiedene Datensätze aus der trunkierten Normalverteilung: ein mit unterschiedlichen Niveaus von Streuung sowie links und rechts schiefen Datensätzen. Wir setzen den Parameter V gleich 0.7 für alle Methoden.

Zuerst wird das Verhalten der Methoden bei unterschiedlich starker Streuung der Daten analysiert. Wir generieren Zufallszahlen aus der trunkierten Normalverteilung mit dem Mittelwert μ=5 und der Standardabweichung zwischen 1 und 30. Es wird die bei –100 trunkierte Normalverteilung verwendet, da wir davon ausgehen, dass die Wachstumsraten in Prozent (der Umsätze, Investitionen etc.) nicht kleiner sein können als –100. In Grafik 1 wird die systematische Verzerrung («Bias») definiert als $\hat{\mu}_m - \bar{\mu}$ wo $m = \{HB, rl, Qn\}$ dargestellt. Die Verzerrung für die HB-Methode steigt je grösser die Standardabweichung des Datensatzes ist, wobei für die zwei anderen Methoden die Verzerrung eher konstant bleiben. Die Entwicklung des durchschnittlichen Anteils der Beobachtungen, die als Ausreisser identifiziert sind, ist in Grafik 2 dargestellt. Für die rl- und die Qn-Methode scheint der Anteil recht unverändert zu bleiben, während die HB-Methode immer mehr Beobachtungen als Ausreisser identifiziert. Wenn wir den Anteil der Ausreisser getrennt für beide Ränder betrachten (nicht dargestellt), wird ersichtlich, dass die HB-Methode auf dem linken Rand immer mehr (kleine) Beobachtungen als Ausreisser erkennt. Ferner zeigt Tabelle 1 im Appendix, dass die rl- und Qn-Methoden zu kleineren Wurzeln der mittleren quadratischen Fehler führen, da sie sowohl niedrigere Varianz als auch Verzerrung aufweisen.

Grafik 1: Verzerrung für normalverteilte Datensätze mit unterschiedlichen Niveaus von Streuung

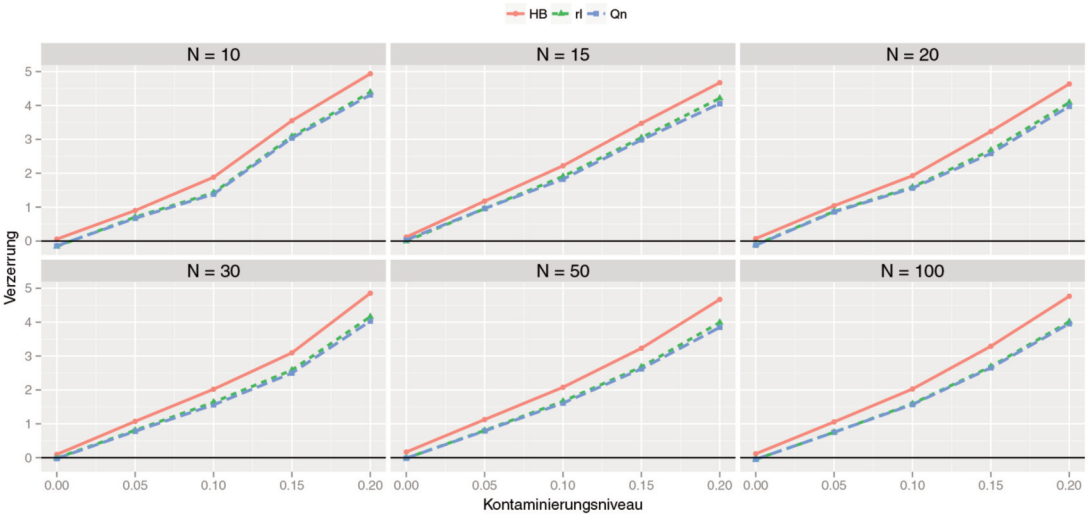


Grafik 2: Durchschnittlicher Anteil von Ausreissern für normalverteilte Datensätze mit unterschiedlichen Niveaus von Streuung

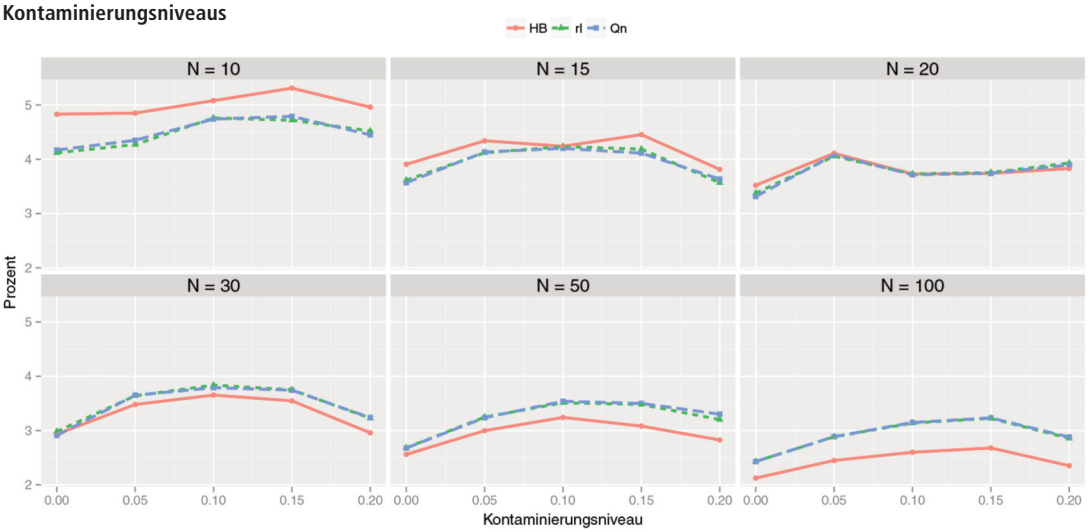


Um analysieren zu können, wie sich diese Methoden im Fall von Datensätzen aus rechtsschiefen Verteilungen verhalten, generieren wir kontaminierte Datensätze. Dafür mischen wir einen aus der trunkierten Normalverteilung generierten Datensatz (mit einem Mittelwert von 5 und einer Standardabweichung von 10) mit Zufallszahlen, die aus einer (trunkierten) Normalverteilung mit Mittelwert gleich 30 und Standardabweichung gleich 10 gezogen wurden. Wir variieren das Kontaminierungsniveau zwischen 0 und 20 Prozent. In der Grafik 3 ist die systematische Verzerrung wieder abhängig vom Niveau der Kontaminierung dargestellt. Wie erwartet steigt die Verzerrung umso mehr, je höher der Anteil der kontaminierten Beobachtungen ist. Allerdings steigt sie für die HB-Methode etwas mehr als bei den anderen zwei Methoden. Dies resultiert daraus, dass die HB-Methode mehr Ausreisser auf dem linken Rand als die rl- und die Qn-Methode identifiziert, obwohl der durchschnittliche Gesamtanteil der Ausreissern konstant bleibt (Grafik 4). Tabelle 2 zeigt, dass auch in diesem Fall die Varianz und die Verzerrung für die rl- und die Qn-Methode kleiner ist und daher auch die Wurzel aus dem mittleren quadratischen Fehler.

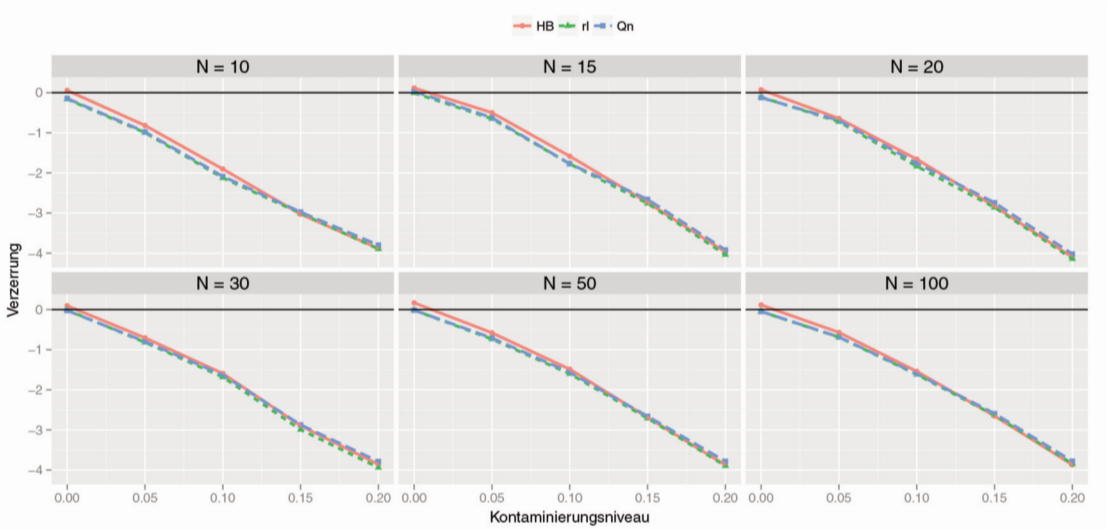
Grafik 3: Verzerrung für rechtsschiefe Datensätze mit unterschiedlichen Kontaminierungsniveaus



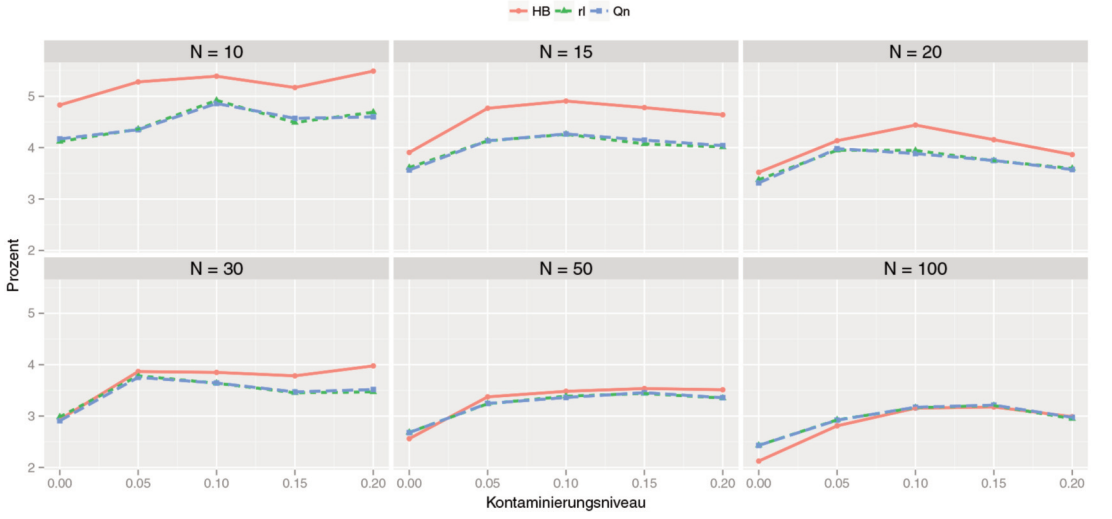
Grafik 4: Durchschnittlicher Anteil von Ausreissern für rechtsschiefe Datensätze mit unterschiedlichen Kontaminierungsniveaus



Grafik 5: Verzerrung für linksschiefe Datensätze mit unterschiedlichen Kontaminierungsniveaus



Grafik 6: Durchschnittlicher Anteil von Ausreissern für linksschiefe Datensätze mit unterschiedlichen Kontaminierungsniveaus



Wir generieren auch linksschiefe Datensätze auf die gleiche Weise, wie wir rechtsschiefe Daten generiert haben. Hier werden die Ausreisser (die Kontaminierung) aus einer trunkierten Normalverteilung mit dem Mittelwert –20 und der Standardabweichung von 10 gezogen. Erneut werden Kontaminierungsniveaus zwischen 0 und 20 Prozent berücksichtigt. In den Grafiken 5 wird die Entwicklung der Verzerrung in Abhängigkeit vom Kontaminierungsniveau dargestellt. Die Verzerrung scheint für alle Methoden etwa gleich gross zu sein. Der Prozentanteil der als Ausreisser identifizierten Beobachtungen ist für die HB-Methode etwas grösser als für die kleineren Datensätze der zwei anderen Methoden (siehe Grafik 6). Dies ist darauf zurückzuführen, dass die HB-Methode bei den kleinen Stichproben mehr Ausreisser auf dem linken Rand identifiziert. Wie in der Tabelle 3 zu sehen ist, führen die rl- und die Qn-Methode zu kleinerer Varianz als die HB-Methode.

Alles in allem deuten die Ergebnisse dieser Simulationsstudie darauf hin, dass die rl- und die Qn-Methode besser mit variierenden Niveaus von Streuung und Schiefe umgehen können. In allen drei Fällen führen die von uns vorgeschlagenen Methoden zu niedrigerer Varianz und gleich grosser oder niedrigerer Verzerrung. Ferner scheint die HB-Methode (mit fixem Wert für den Parameter *C*) zu sensitiv am linken Rand zu sein.

5 ZUSAMMENFASSUNG

In diesem Artikel untersuchen wir, wie sich verschiedene nicht-parametrische Ausreisser-Methoden für (grössen)gewichtete quantitative Umfragedaten (Veränderungsraten) verhalten. Da die Eigenschaften der Verteilung des Datensatzes (Wachstumsraten) von Periode zu Periode und/oder von Teilprobe zu Teilprobe ändern können, analysieren wir die Identifizierung von Ausreissern in Datensätzen mit unterschiedlichem Mass an Streuung und Schiefe. Wir schlagen neue bzw. angepasste Methoden vor, die besser mit verschiedenen Verteilungsmodellen funktionieren. Um zu untersuchen, wie sich unsere neuen Methoden in sich und im Vergleich zu der Methode von Hidiroglou und Berthelot (1986) unter verschiedenen Verteilungsmodellen verhalten, führen wir eine Simulationsstudie durch. Die Ergebnisse der Simulation deuten darauf hin, dass unsere neuen Methoden besser mit variierenden Niveaus von Streuung und Schiefe umgehen können. Dies hängt damit zusammen, dass die HB-Methode dazu tendiert, am linken Rand zu sensitiv zu sein und daher zu viele (kleine) Ausreisser identifiziert. Dies könnte eventuell vermieden werden, wenn man die Parameterwerte der HB-Methode anpasst. Wir stellen jedoch auf eine Methode ab, die auch ohne Anpassungen der Parameter angewendet werden kann, wenn das Verteilungsmodell sich ändert.

LITERATUR

Barnett, V. und Lewis, T. (1994). Outliers in statistical data (3. ed.). New York: Wiley.

Brys, G., Hubert, M. und Struyf, A. (2004). A robust measure of skewness. Journal of Computational and Graphical Statistics, 13(4), 996–1017.

Chambers, R. L. (1986). Outlier robust finite population estimation. Journal of the American Statistical Association, 81(396), 1063–1069.

Dixon, W. J. (1951). Ratios Involving Extreme Values. The Annals of Mathematical Statistics.

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. The Annals of Mathematical Statistics, 21(1), 27–58.

Hidiroglou, M. A. und Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12, 73–83.

Hubert, M. und Van Der Veeken, S. (2008). Outlier detection for skewed data. Journal of Chemometrics, 22(3-4), 235–246.

Hubert, M. und Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. Computational Statistics and Data Analysis, 52(12), 5186–5201.

Rousseeuw, P. J. und Croux, C. (1993). Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association, 88(424), 1273–1283.

APPENDIX

Tabelle 1: Ergebnisse der Simulation mit normalverteilten Datensätzen mit unterschiedlichen Niveaus von Streuung («sd»)

Sample Size	sd	HB_est	rl_est	Qn_est	HB_var	rl_var	Qn_var	HB_rmse	rl_rmse	Qn_rmse	HB_Bias	rl_Bias	Qn_Bias
10	1	4.99	4.97	4.97	0.18	0.16	0.16	0.28	0.28	0.28	0.02	0.00	0.00
15	1	4.98	4.99	4.99	0.13	0.11	0.12	0.24	0.24	0.24	0.00	0.01	0.01
20	1	5.01	5.00	5.00	0.10	0.09	0.09	0.20	0.20	0.20	-0.01	-0.02	-0.02
30	1	5.01	5.00	5.01	0.06	0.05	0.05	0.17	0.17	0.17	0.00	-0.01	0.00
50	1	5.00	5.00	5.00	0.04	0.04	0.04	0.14	0.14	0.14	0.01	0.01	0.01
100	1	5.00	5.00	5.00	0.02	0.02	0.02	0.10	0.10	0.10	0.00	0.00	0.00
10	5	5.02	5.00	5.00	4.69	4.18	4.22	1.39	1.37	1.39	0.05	0.03	0.03
15	5	4.95	4.96	4.96	3.52	2.93	2.94	1.21	1.18	1.20	-0.05	-0.04	-0.04
20	5	5.02	5.00	5.00	2.36	1.97	2.00	1.04	1.03	1.04	0.03	0.01	0.01
30	5	4.96	4.92	4.92	1.68	1.36	1.37	0.87	0.86	0.87	0.02	-0.02	-0.02
50	5	5.07	5.02	5.02	1.06	0.86	0.88	0.69	0.69	0.69	0.01	-0.04	-0.04
100	5	5.02	4.99	4.99	0.55	0.48	0.48	0.48	0.46	0.46	0.04	0.01	0.01
10	10	4.96	4.92	4.90	18.53	16.18	16.51	2.88	2.83	2.89	0.05	0.01	-0.01
15	10	4.93	4.84	4.83	12.24	10.09	10.23	2.44	2.33	2.36	0.01	-0.08	-0.09
20	10	5.16	5.02	4.99	9.90	8.61	8.69	2.08	1.99	1.99	0.16	0.02	-0.01
30	10	5.19	5.01	5.01	7.12	5.65	5.79	1.77	1.72	1.76	0.21	0.03	0.03
50	10	5.19	5.02	5.02	4.06	3.44	3.44	1.41	1.37	1.39	0.16	-0.01	-0.01
100	10	5.17	5.01	5.02	2.08	1.72	1.73	0.95	0.93	0.93	0.18	0.02	0.03
10	20	5.36	4.87	4.82	77.40	66.41	67.53	5.94	5.73	5.83	0.56	0.07	0.02
15	20	5.60	5.12	5.16	54.12	45.66	45.88	4.92	4.64	4.73	0.43	-0.05	-0.01
20	20	5.55	4.79	4.78	39.70	32.31	32.42	4.18	3.81	3.82	0.62	-0.14	-0.15
30	20	5.61	4.85	4.80	25.89	21.33	21.39	3.77	3.43	3.44	0.80	0.04	-0.01
50	20	5.61	4.91	4.92	18.04	14.56	14.76	2.83	2.59	2.61	0.61	-0.09	-0.08
100	20	5.82	5.10	5.11	7.85	6.83	6.84	2.08	1.88	1.88	0.69	-0.03	-0.02
10	30	5.83	4.35	4.28	169.32	151.30	153.25	9.38	8.56	8.61	1.53	0.05	-0.02
15	30	6.83	5.58	5.59	113.77	99.26	99.24	7.79	7.25	7.33	1.67	0.42	0.43
20	30	6.73	5.11	5.10	86.58	70.62	71.30	6.82	6.04	6.11	1.88	0.26	0.25
30	30	6.38	4.83	4.82	62.35	50.22	51.06	5.72	4.79	4.80	1.60	0.05	0.04
50	30	6.70	5.12	5.12	37.75	30.21	30.38	4.64	3.85	3.91	1.55	-0.03	-0.03
100	30	6.68	4.98	4.97	16.76	14.52	14.60	3.38	2.68	2.69	1.77	0.07	0.06

Tabelle 2: Ergebnisse der Simulation für rechtsschiefe Datensätze mit unterschiedlichen Kontaminierungsniveaus («LC»)

Sample Size	LC	HB_est	rl_est	Qn_est	HB_var	rl_var	Qn_var	HB_rmse	rl_rmse	Qn_rmse	HB_Bias	rl_Bias	Qn_Bias
10	0.00	5.06	4.84	4.85	18.25	15.47	15.83	4.27	3.93	3.98	0.06	-0.16	-0.15
15	0.00	5.12	4.99	5.03	13.20	11.40	11.66	3.63	3.37	3.41	0.12	-0.01	0.03
20	0.00	5.07	4.89	4.88	9.56	8.17	8.28	3.09	2.86	2.88	0.07	-0.11	-0.12
30	0.00	5.10	4.99	4.98	6.39	5.13	5.19	2.53	2.26	2.28	0.10	-0.01	-0.02
50	0.00	5.17	4.99	4.99	3.99	3.33	3.37	2.00	1.82	1.83	0.17	-0.01	-0.01
100	0.00	5.11	4.95	4.95	2.18	1.77	1.77	1.48	1.33	1.33	0.11	-0.05	-0.05
10	0.05	5.90	5.71	5.67	22.81	20.15	20.48	4.86	4.54	4.57	0.90	0.71	0.67
15	0.05	6.18	5.96	5.96	15.46	12.43	12.50	4.10	3.65	3.66	1.18	0.96	0.96
20	0.05	6.04	5.87	5.86	11.54	9.73	9.89	3.55	3.24	3.26	1.04	0.87	0.86
30	0.05	6.07	5.81	5.78	8.77	7.35	7.31	3.15	2.83	2.81	1.07	0.81	0.78
50	0.05	6.13	5.81	5.79	5.08	4.01	4.03	2.52	2.16	2.16	1.13	0.81	0.79
100	0.05	6.06	5.75	5.75	2.45	1.96	1.96	1.89	1.59	1.59	1.06	0.75	0.75
10	0.10	6.88	6.43	6.38	29.45	23.68	23.80	5.74	5.07	5.07	1.88	1.43	1.38
15	0.10	7.22	6.91	6.82	18.45	15.31	15.34	4.83	4.35	4.32	2.22	1.91	1.82
20	0.10	6.93	6.59	6.56	14.99	12.17	12.29	4.32	3.83	3.84	1.93	1.59	1.56
30	0.10	7.02	6.64	6.56	10.08	7.49	7.50	3.76	3.19	3.15	2.02	1.64	1.56
50	0.10	7.08	6.66	6.61	6.51	4.95	4.95	3.29	2.78	2.75	2.08	1.66	1.61
100	0.10	7.03	6.59	6.57	3.20	2.52	2.53	2.70	2.25	2.24	2.03	1.59	1.57
10	0.15	8.55	8.09	8.04	32.80	27.46	27.73	6.74	6.08	6.08	3.55	3.09	3.04
15	0.15	8.47	8.05	7.98	24.36	19.87	19.89	6.03	5.40	5.36	3.47	3.05	2.98
20	0.15	8.23	7.68	7.58	17.93	13.63	13.55	5.33	4.56	4.50	3.23	2.68	2.58
30	0.15	8.10	7.59	7.50	12.19	9.31	9.19	4.66	4.00	3.93	3.10	2.59	2.50
50	0.15	8.23	7.68	7.62	7.01	5.78	5.76	4.18	3.60	3.55	3.23	2.68	2.62
100	0.15	8.29	7.69	7.65	3.72	2.88	2.87	3.81	3.18	3.15	3.29	2.69	2.65
10	0.20	9.94	9.39	9.31	37.81	32.39	32.62	7.88	7.19	7.15	4.94	4.39	4.31
15	0.20	9.67	9.21	9.05	26.46	22.05	22.43	6.95	6.31	6.23	4.67	4.21	4.05
20	0.20	9.63	9.09	8.97	19.85	16.45	16.77	6.43	5.76	5.70	4.63	4.09	3.97
30	0.20	9.85	9.16	9.03	13.61	10.51	10.54	6.10	5.27	5.17	4.85	4.16	4.03
50	0.20	9.67	8.99	8.85	8.76	6.67	6.62	5.53	4.75	4.63	4.67	3.99	3.85
100	0.20	9.77	9.01	8.95	4.44	3.49	3.48	5.21	4.43	4.37	4.77	4.01	3.95

Tabelle 3: Ergebnisse der Simulation für linksschiefe Datensätze mit unterschiedlichen Kontaminierungsniveaus («LC»)

Sample Size	LC	HB_est	rl_est	Qn_est	HB_var	rl_var	Qn_var	HB_rmse	rl_rmse	Qn_rmse	HB_Bias	rl_Bias	Qn_Bias
10	0.00	5.06	4.84	4.85	18.25	15.47	15.83	4.27	3.93	3.98	0.06	-0.16	-0.15
15	0.00	5.12	4.99	5.03	13.20	11.40	11.66	3.63	3.37	3.41	0.12	-0.01	0.03
20	0.00	5.07	4.89	4.88	9.56	8.17	8.28	3.09	2.86	2.88	0.07	-0.11	-0.12
30	0.00	5.10	4.99	4.98	6.39	5.13	5.19	2.53	2.26	2.28	0.10	-0.01	-0.02
50	0.00	5.17	4.99	4.99	3.99	3.33	3.37	2.00	1.82	1.83	0.17	-0.01	-0.01
100	0.00	5.11	4.95	4.95	2.18	1.77	1.77	1.48	1.33	1.33	0.11	-0.05	-0.05
10	0.05	4.18	4.00	4.02	22.53	18.73	18.95	4.81	4.44	4.46	-0.82	-1.00	-0.98
15	0.05	4.50	4.35	4.38	15.63	12.70	13.10	3.98	3.62	3.67	-0.50	-0.65	-0.62
20	0.05	4.36	4.28	4.31	12.45	9.69	9.91	3.59	3.19	3.22	-0.64	-0.72	-0.69
30	0.05	4.30	4.18	4.20	8.50	7.05	7.07	3.00	2.78	2.77	-0.70	-0.82	-0.80
50	0.05	4.42	4.26	4.29	4.88	3.84	3.86	2.28	2.09	2.09	-0.58	-0.74	-0.71
100	0.05	4.43	4.31	4.32	2.39	1.91	1.93	1.65	1.54	1.55	-0.57	-0.69	-0.68
10	0.10	3.10	2.88	2.91	25.55	22.66	22.72	5.40	5.21	5.20	-1.90	-2.12	-2.09
15	0.10	3.42	3.22	3.23	19.09	16.44	16.31	4.65	4.42	4.41	-1.58	-1.78	-1.77
20	0.10	3.34	3.17	3.24	13.49	11.14	11.10	4.03	3.81	3.77	-1.66	-1.83	-1.76
30	0.10	3.41	3.32	3.37	9.90	7.99	7.99	3.53	3.29	3.26	-1.59	-1.68	-1.63
50	0.10	3.51	3.40	3.44	5.97	5.10	5.16	2.86	2.77	2.76	-1.49	-1.60	-1.56
100	0.10	3.46	3.38	3.40	2.99	2.36	2.36	2.32	2.23	2.22	-1.54	-1.62	-1.60
10	0.15	1.97	1.99	2.02	31.95	28.56	28.77	6.41	6.13	6.13	-3.03	-3.01	-2.98
15	0.15	2.26	2.24	2.34	23.70	19.10	19.16	5.58	5.17	5.12	-2.74	-2.76	-2.66
20	0.15	2.17	2.14	2.26	16.49	13.85	13.84	4.95	4.69	4.62	-2.83	-2.86	-2.74
30	0.15	2.10	2.01	2.13	11.26	9.14	9.01	4.43	4.25	4.15	-2.90	-2.99	-2.87
50	0.15	2.31	2.29	2.34	6.70	5.28	5.20	3.73	3.55	3.50	-2.69	-2.71	-2.66
100	0.15	2.35	2.36	2.41	3.74	3.08	3.05	3.28	3.17	3.13	-2.65	-2.64	-2.59
10	0.20	1.12	1.11	1.20	39.44	35.91	35.32	7.38	7.14	7.05	-3.88	-3.89	-3.80
15	0.20	1.03	0.96	1.08	25.48	21.82	22.04	6.42	6.17	6.11	-3.97	-4.04	-3.92
20	0.20	0.90	0.86	0.98	20.82	17.99	17.99	6.13	5.93	5.84	-4.10	-4.14	-4.02
30	0.20	1.14	1.06	1.21	14.15	11.20	11.14	5.39	5.17	5.05	-3.86	-3.94	-3.79
50	0.20	1.13	1.11	1.22	8.64	7.12	7.23	4.86	4.72	4.64	-3.87	-3.89	-3.78
100	0.20	1.13	1.16	1.22	4.49	3.67	3.67	4.41	4.30	4.24	-3.87	-3.84	-3.78