Ein Ausreißertest

Jürgen Grieser

22.09.1997

1 Warum Ausreißertests?

In einer Zeitreihe (wie in jeder Stichprobe) können Ausreißer vorkommen, d.h. Werte die wahrscheinlich nicht zum Rest der Stichprobe passen. Bei Zeitreihen von Beobachtungswerten, kommen dafür drei Gründe in Frage:

1. Es wurde ein zufälliges seltenes Ereignis beobachtet.
2. Es wurde ein spezielles seltenes Ereignis beobachtet.
3. Es wurde ein Beobachtungsfehler gemacht.

Da die Methode der kleinsten Quadrate, auf der zahlreiche Verfahren zur Statistischen Analyse von Zeitreihen beruhen, sehr empfindlich auf extreme Werte reagiert, stellt sich die Frage, ob der größte bzw. kleinste in der Zeitreihe vorkommende Wert in den statistischen Analysen der Zeitreihe mitverwendet werden soll, oder ob er sehr wahrscheinlich nicht aus der der Zeitreihe zugrundeliegenden Verteilung stammt. In diesem Fall sollte man ihn aus den weiteren Analysen ausschließen und getrennt dokumentieren.

Ausreißertests dienen nun dazu, zu berechnen, mit welcher Wahrscheinlichkeit der Ausreißer ein zufälliges seltenes Ereignis darstellt, d.h. mit welcher Wahrscheinlichkeit er verträglich mit der der Zeitreihe zugrundeliegenden Verteilung ist.

2 Anwendbarkeit des Ausreißertests

Für den hier durchgeführten Ausreißertest wird von der Hypothese ausgegangen, daß die restlichen Werte der Zeitreihe Gauß-verteilt sind. Diese Hypothese muß natürlich zunächst getestet werden. Falls die Zeitreihe nicht Gauß-verteilt ist, muß man entweder

- eine andere theoretische Verteilung anpassen (und die Teststrategie erneut lösen), oder
- die Zeitreihe in eine Gauß-verteilte Reihe transformieren oder
- etwaige Strukturen innerhalb der Zeitreihe (z.B. Trend, Saisonfigur) herausfiltern.

3 Theorie des Ausreißertests

Zunächst wird von einer Zeitreihe von identisch normalverteilten unabhängigen Variablen ausgegangen (Gaußsches Rauschen). Diese hat die Wahrscheinlichkeitsdichtefunktion

$$p(z) = \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{z^2}{2}\right). \tag{1}$$

Für die Wahrscheinlichkeitsfunktion $P(z \leq Z)$ folgt dann

$$P(z \leq Z) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{Z} \exp\left(-\frac{z'^2}{2}\right) dz'. \tag{2}$$

Für die Wahrscheinlichkeit eine Zahl zwischen $\pm Z$ anzutreffen, folgt dann

$$P(-Z \leq z \leq Z) = P(Z) - P(-Z). \tag{3}$$

Da die Wahrscheinlichkeitsdichtefunktion der Normalverteilung eine gerade Funktion ist, gilt weiter

$$P(-Z) = 1 - P(Z). \tag{4}$$

Daraus folgt sofort:

$$P(-Z \leq z \leq Z) = 2\,P(Z) - 1. \tag{5}$$

Aus dem gleichen Grund kann man für $Z > 0$ Glg. (2) umformulieren zu

$$P(z \leq Z) = \frac{1}{2} + \frac{1}{\sqrt{2\,\pi}} \int_{0}^{Z} \exp\left(-\frac{z'^2}{2}\right) dz'. \tag{6}$$

Dieses Integral ist nicht analytisch lösbar. Jedoch gibt es ein ähnliches Integral, für das z.B. in den Numerical Recipes Reihen- und Partialbruchnäherungen angegeben werden. Dieses "ähnliche" Integral ist die sogenannte Errorfunktion und hat folgende Gestalt:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \exp\left(-u^2\right) du. \tag{7}$$

Um das Integral in Glg. (6) in diese Form zu bringen wird eine einfache Variablentransformation durchgeführt. Dazu wird Glg. (6) zunächst umgeschrieben zu

$$P(z \leq Z) = \frac{1}{2} + \frac{1}{\sqrt{2\,\pi}} \int_0^Z \exp\left(- \left[\frac{z'}{\sqrt{2}} \right]^2 \right) dz'. \tag{8}$$

Transformiert wird nun $t = \frac{z'}{\sqrt{2}}$, d.h. $z' = \sqrt{2}\,t$, mit der Ableitung $\frac{dz'}{dt} = \sqrt{2}$. Damit folgt für $P(z \leq Z)$

$$P(z \leq Z) = \frac{1}{2} + \frac{1}{\sqrt{2\,\pi}} \int_0^{t(Z)} \exp\left(- t^2 \right) \frac{dz'}{dt} dt. \tag{9}$$

Setzt man die Ableitung und die transformierte Integralgrenze ein, so erhält man

$$P(z \leq Z) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{\frac{Z}{\sqrt{2}}} \exp\left(- t^2 \right) dt = \frac{1}{2} \left(1 + \mathrm{erf}\left(\frac{Z}{\sqrt{2}} \right) \right). \tag{10}$$

Aus den Gleichungen (5) und (10) folgt nun

$$P(-Z \leq z \leq Z) = \mathrm{erf}\left(\frac{Z}{\sqrt{2}} \right) \tag{11}$$

Dies ist die Wahrscheinlichkeit dafür, durch Zufall aus einer identisch normalverteilten Variable einen Wert zu ziehen, dessen Betrag kleiner Z ist. $1 - P(-Z \leq z \leq Z)$ ist dann die Wahrscheinlichkeit durch Zufall einen mindestens so großen Wert zu ziehen wie Z. Diese Wahrscheinlichkeit gilt für den Fall, daß man einmal zieht. Die Zeitreihe, die untersucht werden soll, besteht aber aus N Werten. Sie stellt damit gemäß den Annahmen eine Realisation dar, bei der N-mal hintereinander (unabhängig) eine solche Zufallszahl gezogen wurde. Dies wiederum ist ein Bernoulli-Experiment. Die Wahrscheinlichkeit bei N Realisationen k mal einen Wert mit der Eintrittswahrscheinlichkeit $1 - P(-Z \leq z \leq Z)$ zu erhalten folgt demnach einer Binomialverteilung. Für Werte von $N \geq 100$ und $1 - P(-Z \leq z \leq Z) \leq .05$ kann diese Verteilung durch die Poissonverteilung genähert werden. Somit ist die Wahrscheinlichkeit für das zufällige k-malige Auftreten eines solch großen (oder größeren) Wertes gegeben durch

$$p(k, N, 1 - P(-Z \leq z \leq Z)) = \frac{e^{-\lambda}\,\lambda^k}{k!} \tag{12}$$

mit $\lambda = N\,(1 - P(-Z \leq z \leq Z))$. Die Wahrscheinlichkeit dafür, daß in einer Zeitreihe der Länge N ein solch extremer Wert durch Zufall nicht auftritt ist demnach

$$p(k = 0, N, 1 - P(-Z \leq z \leq Z)) = \exp\left[-N\ \mathrm{erf}\left(\frac{Z}{\sqrt{2}} \right) \right] \tag{13}$$

Damit ist die Wahrscheinlichkeit dafür, daß ein Wert mit dem Abstand $\|z\| \geq Z$ vom Mittelwert der normierten Gaußverteilung in einer Zeitreihe des Umfangs N durch Zufall mindestens einmal auftritt, gegeben durch

$$1 - p(k = 0, N, 1 - P(-Z \leq z \leq Z)) = 1 - \exp\left[-N \operatorname{erf}\left(\frac{Z}{\sqrt{2}}\right)\right]. \quad (14)$$

4 Durchführung des Ausreißertests

Bei der Durchführung des Ausreißertests wird wie folgt vorgegangen:

Zunächst wird der am weitesten vom Mittelwert entfernte Wert der Verteilung gesucht und als möglicher Ausreißer ins Auge gefaßt. Aus den restlichen Werten werden Mittelwert und Standardabweichung der zugrundeliegenden Verteilung geschätzt. Mit Hilfe des Kolmogoroff-Smirnoff-Tests wird getestet, ob die Verteilung der restlichen Werte signifikant von der Gaußverteilung abweicht. Nur wenn dies nicht der Fall ist, darf der Test weiter durchgeführt werden. Im nächsten Schritt wird der normierte Abstand des möglichen Ausreißers vom Mittelwert der Verteilung der restlichen Werte berechnet. Daraufhin kann die Wahrscheinlichkeit dafür berechnet werden, daß ein Wert, der so weit oder weiter vom Mittelwert der Verteilung entfernt liegt, durch Zufall als eine Realisation der Verteilung auftritt (Gleichung (11)). Die Zeitreihe mit N Werten stellt dann ein Bernoulli-Experiment mit N Realisationen dar. Im letzten Schritt braucht man nun nur noch zu testen, wie wahrscheinlich kein solch großer Wert in einer Zeitreihe der gegebenen Länge auftritt (Gleichung (13)).