Robuste Statistik

Musterlösung zu Blatt 7

Aufgabe 7.1:

a) Lese zunächst die Daten ein und schreibe eine Funktion, die die Q-Schätzung berechnet.

```
# Datensaetze
data <- list(y1 = c(2.3, 3.4, 5.6, 7.1, 8.9),
  y2 = c(-10, 3.4, 5.6, 7.1, 8.9),
  y3 = c(-10, -10, 5.6, 7.1, 8.9),
  y4 = c(-10, -10, -10, 7.1, 8.9),
  y5 = c(-10, -10, -10, -10, 8.9))

# q.estimator: Funktion fuer die Q-Schaetzung
# Eingabe: y (numeric): Datenvektor, fuer den die Schaetzung berechnet werden soll
# Ausgabe: (numeric): Schaetzwert
q.estimator <- function(y){
  grid <- expand.grid(v1 = 1:length(y), v2 = 1:length(y))
  index <- which(apply(grid, 1, diff) == 0)
  grid <- grid[-index, ]
  diffs <- apply(grid, 1, function(x) abs(diff(y[x])))
  return(as.numeric(quantile(diffs, 0.25, type = 1)))
}
```

Nun berechne die Q-Schätzung auf den fünf Datensätzen:

```
sapply(data, q.estimator)
#  y1  y2  y3  y4  y5
# 1.8 2.2 1.8 0.0  0.0
```

Es fällt auf, dass die Q-Schätzung bei y^4 und y^5 wie die kürzeste Hälfte implodiert, da sie einen Streuungsschätzer von 0 liefert. Für die anderen drei Datensätze scheint die Q-Schätzung vergleichsweise robust zu sein, alle drei Schätzwerte liegen in der gleichen Größenordnung.

In einem anderen schönen Code von zwei Studentinnen aus dem Kurs wird die Funktion `dist()` mit Parameter `manhattan` benutzt, um eine Distanzmatrix zu generieren und diese dann zu verdoppeln:

```
q.estimator <- function(y) {
  pairwAbsDiff <- rep(as.numeric(dist(y,"manhattan")),2)
  N <- length(pairwAbsDiff)
  k <- ceiling(N/4)
  res <- sort(pairwAbsDiff)[k]

  return(res)
}
```

Auch dieser Code eines anderen Studenten verwendet die Funktion `rep()` geschickt:

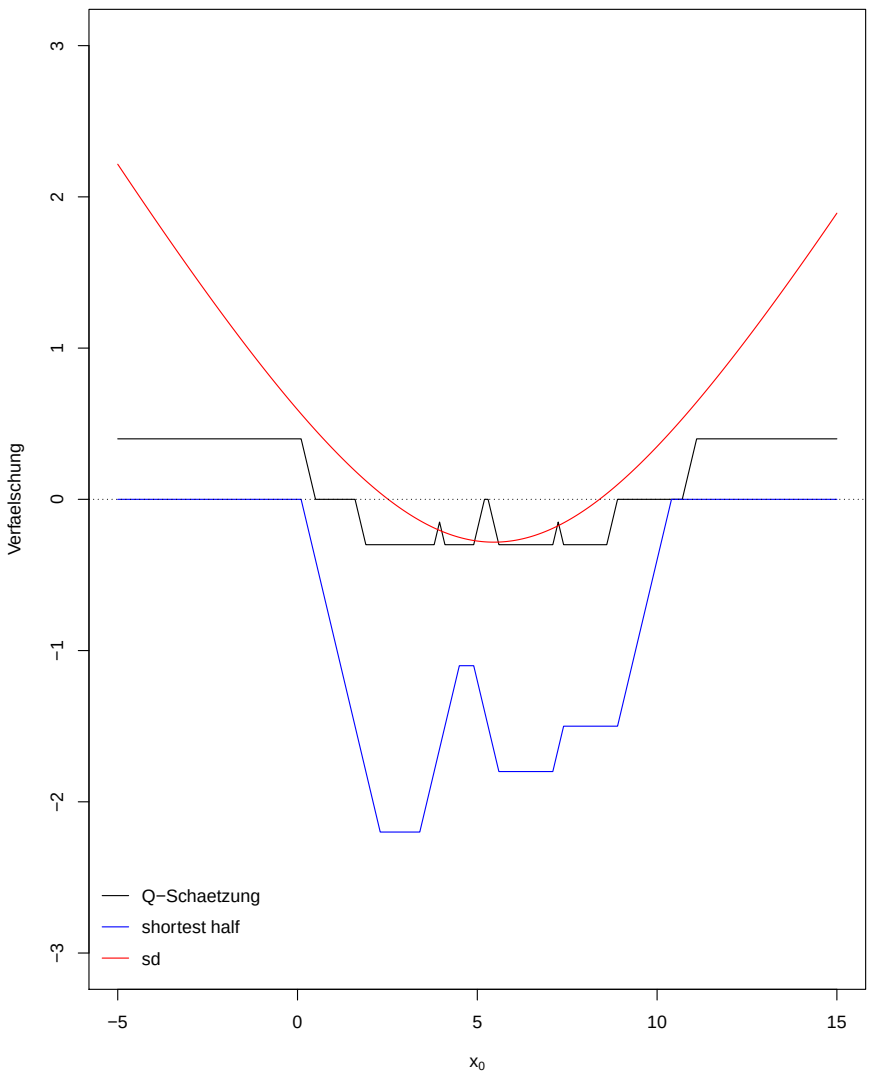
```
q.estimator <- function(data){
  n <- length(data)
  v1 <- rep(data, n)
  v2 <- rep(data, each = n)
  erg <- abs(v1 - v2)
  sorted <- sort(erg)
  sorted <- sorted[-(1:n)]
  as.numeric(quantile(sorted, 0.25))
}
```

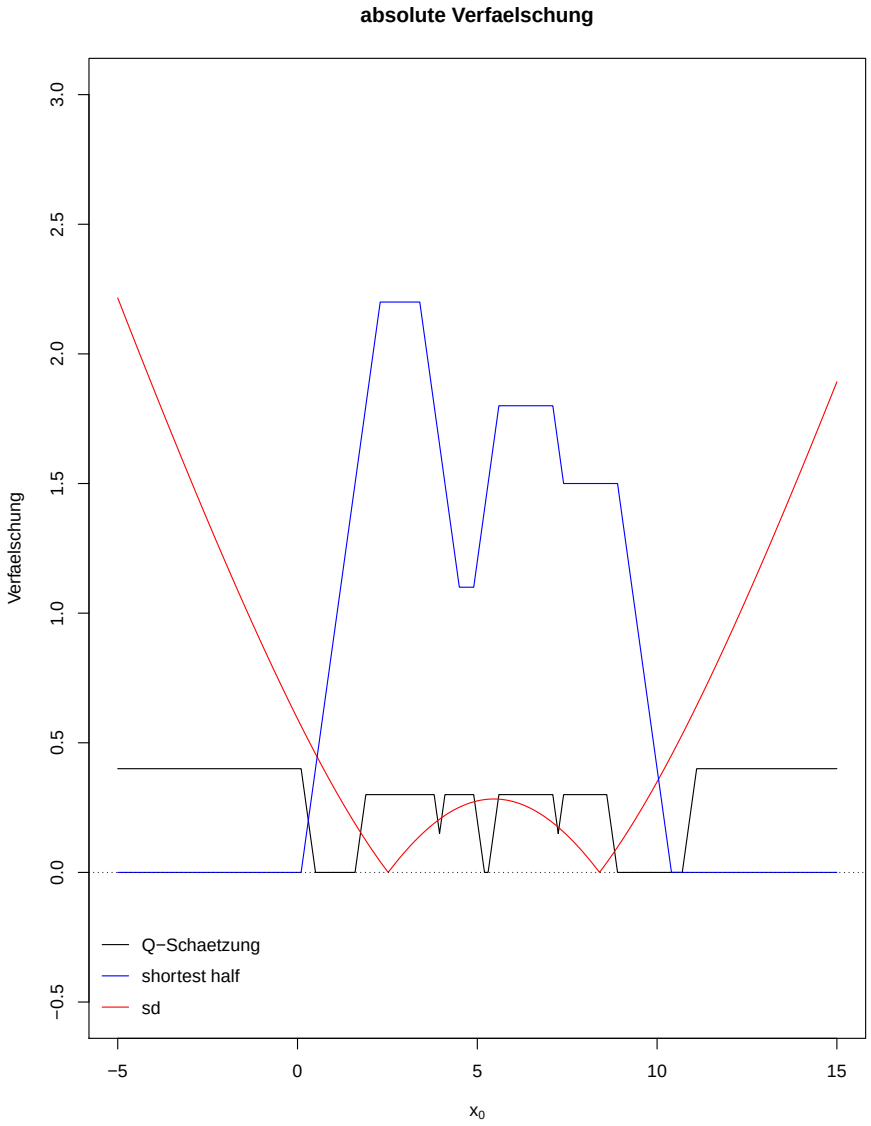
Alle Varianten haben allerdings eine quadratische Laufzeit, da $\binom{N}{2} = \frac{1}{2}N(N-1)$ Möglichkeiten berechnet werden. In dem Paper *Alternatives to the Median Absolute Deviation* von Rosseeuw und Croux von 1992 wird ein anderer Schätzer Q_n angegeben, auf dem der Q-Schätzer aus der Vorlesung beruht. Ein Schätzung mit Q_n kann sogar in Zeit von $O(N \log(N))$ (wegen der Sortierung) statt in quadratischer Zeit ermöglicht werden. In dem Paper *Time efficient algorithms for two highly robust estimators of scale* von Rosseeuw und Croux von 1992 wird ein Algorithmus mit verkürzter Laufzeit vorgestellt.

b) Verfälschungsfunktion:

```
# verfaelschung: Funktion, welche die Verfaelschung durch Hinzufuegen berechnet
# Eingabe: data (numeric): Datenvektor, auf dessen Grundlage die Verfaelschung
#                          berechnet wird
#          fun (function): Schaetzfunktion, fuer die die Verfaelschung berechnet wird
#          x0 (numeric): Vektor mit hinzuzufuegenden Werten zu data
# Ausgabe: (numeric): Verfaelschung
verfaelschung <- function(data, fun, x0){
  true <- fun(data)
  schaetzung <- sapply(x0, function(x){
    dat <- c(data, x)
    fun(dat)
  })
  return(schaetzung - true)
}
```

Dies führt zu folgenden grafischen Darstellungen:


Verfaelschung
Verfaelschung
3
2
1
0
−1
−2
−3
−5
0
5
10
15
x_0
Q−Schaetzung
shortest half
sd



Es zeigt sich, dass die Verfälschungsfunktion der Standardabweichung eine Parabel darstellt. Diese ist nach oben unbeschränkt. Hingegen ist die Verfälschung der Q-Schätzung beschränkt. Diese ist stückweise konstant mit einzelnen Peaks. Für die inliers sind die Verfälschungen für die Q-Schätzung und Schätzung der Standardabweichung am geringsten und die kürzeste Hälfte deutlich höher. Für Ausreißer explodiert die Standardabweichung, während die anderen beiden Schätzer robust sind. Im Gegensatz zur kürzesten Hälfte hat die Q-Schätzung bei sehr extremen Werte keine Verfälschung von 0.

Aufgabe 7.2:

zu zeigen: Die Bruchpunkte der Spannweite, der absoluten Abweichung und der Standardabweichung betragen jeweils $\frac{1}{N}$.

Ein Bruchpunkt ist dabei definiert als das Minimum des Explosionspunkts und des Implosionspunkts. Ein Bruchpunkt weist immer mindestens den Wert $\frac{1}{N}$ auf (untere Schranke). Dieser minimale Bruchpunkt wird angenommen, wenn es ausreicht nur eine Beobachtung zu verfälschen.

a) Die Spannweite $y_{(N)} - y_{(1)}$.

Verfälsche eine Beobachtung. Setze dazu $y_{(N)}$ auf $l \geq y_{(N)}$. Somit ändert sich die Reihenfolge der Sortierung nicht. Nun gilt:

$$\lim_{l\to\infty}(y_{(N)} - y_{(1)}) = \lim_{l\to\infty}(l - y_{(1)}) = \lim_{l\to\infty} l - y_{(1)} = \infty - y_{(1)} = \infty$$

Somit kann schon mit einer Beobachtung eine beliebige Verfälschung kreiert werden. Da der Explosionspunkt schon den minimal möglichen Wert angenommen hat, braucht der Implosionspunkt nicht berechnet zu werden, um festzustellen, dass der Bruchpunkt der Spannweite $\frac{1}{N}$ beträgt.

b) Die absolute Abweichung $\frac{1}{N}\sum_{n=1}^{N}|y_n - \tilde{y}_{0.5}|$.

Verfälsche eine Beobachtung. Setze dazu $y_{(N)}$ auf $l \geq y_{(N)}$. OBdA wird angenommen, dass $N \geq 3$ gilt. Somit bleibt der Median von einem Ausreißer in der größten Beobachtung unberührt. Dennoch gilt:

$$\lim_{l\to\infty}\frac{1}{N}\sum_{n=1}^{N}|y_n - \tilde{y}_{0.5}| = \lim_{l\to\infty}\frac{1}{N}\sum_{n=1}^{N}|y_{(n)} - \tilde{y}_{0.5}| = \lim_{l\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N-1}|y_{(n)} - \tilde{y}_{0.5}| + \frac{1}{N}|l - \tilde{y}_{0.5}|\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N-1}|y_{(n)} - \tilde{y}_{0.5}| + \frac{1}{N}\lim_{l\to\infty}|l - \tilde{y}_{0.5}| = \frac{1}{N}\sum_{n=1}^{N-1}|y_{(n)} - \tilde{y}_{0.5}| + \frac{1}{N}|\infty - \tilde{y}_{0.5}| = \infty$$

Somit kann schon mit einer Beobachtung eine beliebige Verfälschung kreiert werden. Da der Explosionspunkt schon den minimal möglichen Wert angenommen hat, braucht der Implosionspunkt nicht berechnet zu werden, um festzustellen, dass der Bruchpunkt der absoluten Abweichung $\frac{1}{N}$ beträgt.

c) Die Standardabweichung $\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2}$. Wir können auch das Quadrat der Standardabweichung betrachten.

Verfälsche eine beliebige Beobachtung y_n.

$$\begin{aligned}
\frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2 &= \frac{1}{2N(N-1)}\sum_{m=1}^{N}\sum_{n=1}^{N}(y_m - y_n)^2 \\
&= \frac{1}{2N(N-1)}\left(\sum_{m=1}^{N-1}\sum_{n=1}^{N-1}(y_m - y_n)^2 + 2\sum_{n=1}^{N-1}(y_n - l)^2\right) \\
&= \frac{1}{2N(N-1)}\left(\sum_{m=1}^{N-1}\sum_{n=1}^{N-1}(y_m - y_n)^2 + 2\sum_{n=1}^{N-1}(y_n^2 - 2y_n l + l^2)\right) \\
&= \frac{1}{2N(N-1)}\left(\sum_{m=1}^{N-1}\sum_{n=1}^{N-1}(y_m - y_n)^2 + 2\sum_{n=1}^{N-1}y_n^2 + 2(N-1)l^2 - \left(4\sum_{n=1}^{N-1}y_n\right)l\right)
\end{aligned}$$

Für $l \to \infty$ ist $al^2 + bl \to \infty$ für $a > 0$ und $b \in \mathbb{R}$, sodass wir für die Varianz eine Explosion erhalten. Ziehen wir die Wurzel, bleibt die Aussage für die empirische Standardabweichung korrekt.