Robuste Statistik

Musterlösung zu Blatt 6

Aufgabe 6.1:

a) Lese zunächst die Daten ein und schreibe eine Funktion, die die kürzeste Hälfte 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))

# shortest_half: Funktion fuer die kuerzste Haelfte
# Eingabe: y (numeric): Datenvektor, fuer den die Schaetzung berechnet werden soll
# Ausgabe: (numeric): Schaetzwert
shortest_half <- function(y){
   y <- sort(y)
  h <- ceiling(length(y)/2)
  y_diff <- min(diff(y, lag=h-1))
  return(y_diff)
}
```

Nun berechne die kürzeste Hälfte und die Standardabweichung auf den fünf Datensätzen:

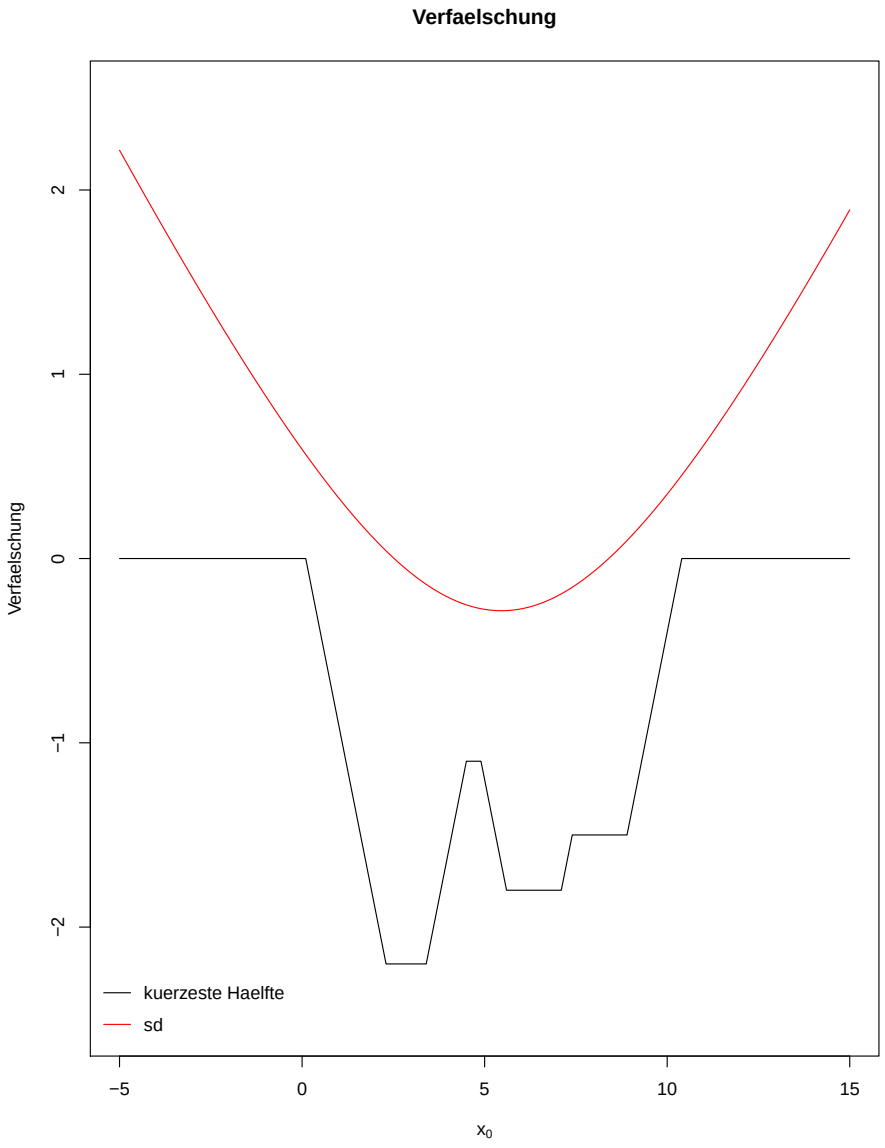
```
sapply(data, shortest_half)
# y1  y2  y3  y4  y5
# 3.3 3.3 3.3 0.0 0.0
sapply(data, sd)
#       y1       y2       y3       y4       y5
# 2.681977 7.542215 9.492997 9.879524 8.452337
```

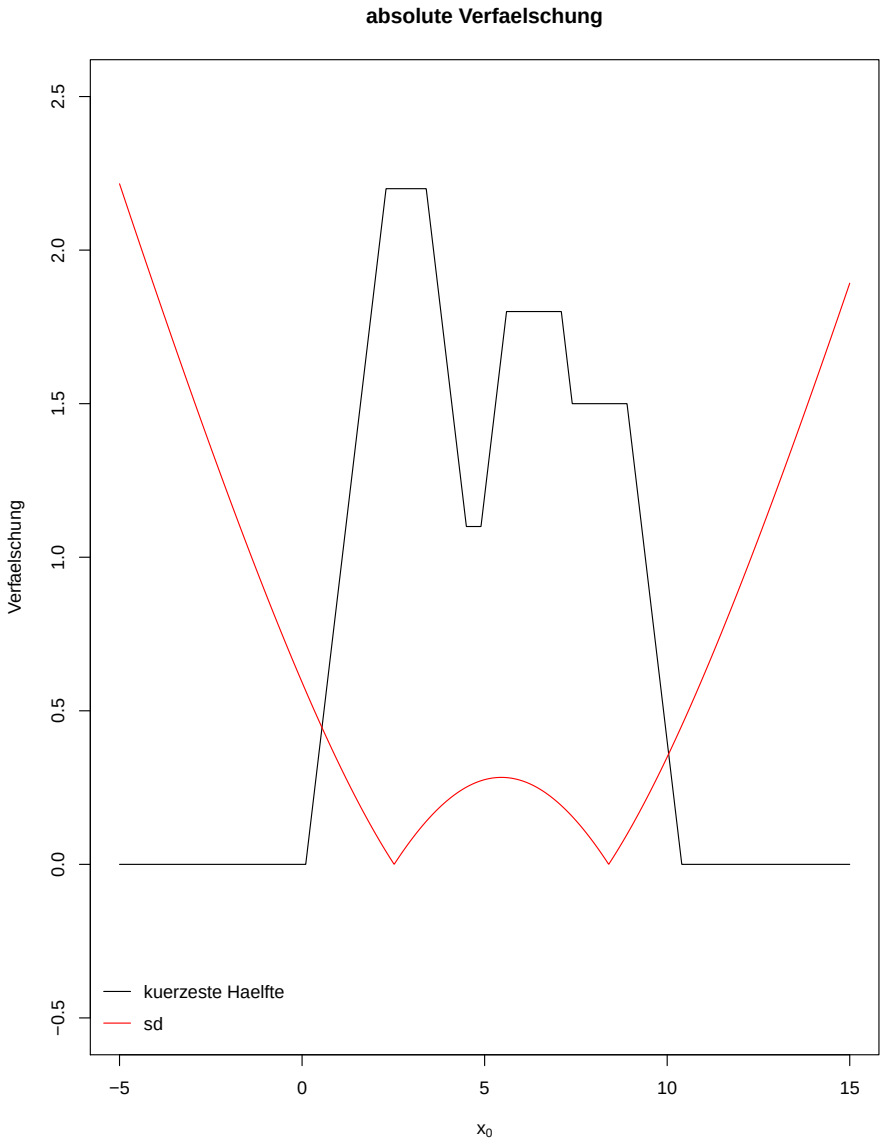
Es fällt auf, dass die kürzeste Hälfte bei y^4 implodiert, da sie einen Streuungsschätzer von 0 liefert. Auffällig ist, dass der Ausreißer bis zum Implosionspunkt die Schätzung für die kürzeste Hälfte gar nicht verfälscht. Hingegen steigt die Streuungsschätzung mittels der Standardabweichung von y^1 bis y^4 stets an. Schon ein einzelner Ausreißer verdreifacht dabei ungefähr die Streuungsschätzung. Hingegen ist der Streuungsanstieg zwischen zwei und drei Ausreißern nicht mehr wirklich groß. Ab vier Ausreißern nimmt die Streuungsschätzung wieder ab, da der Großteil des Datensatzes aus dem Ausreißer besteht.

b) Verfälschungsfunktion:

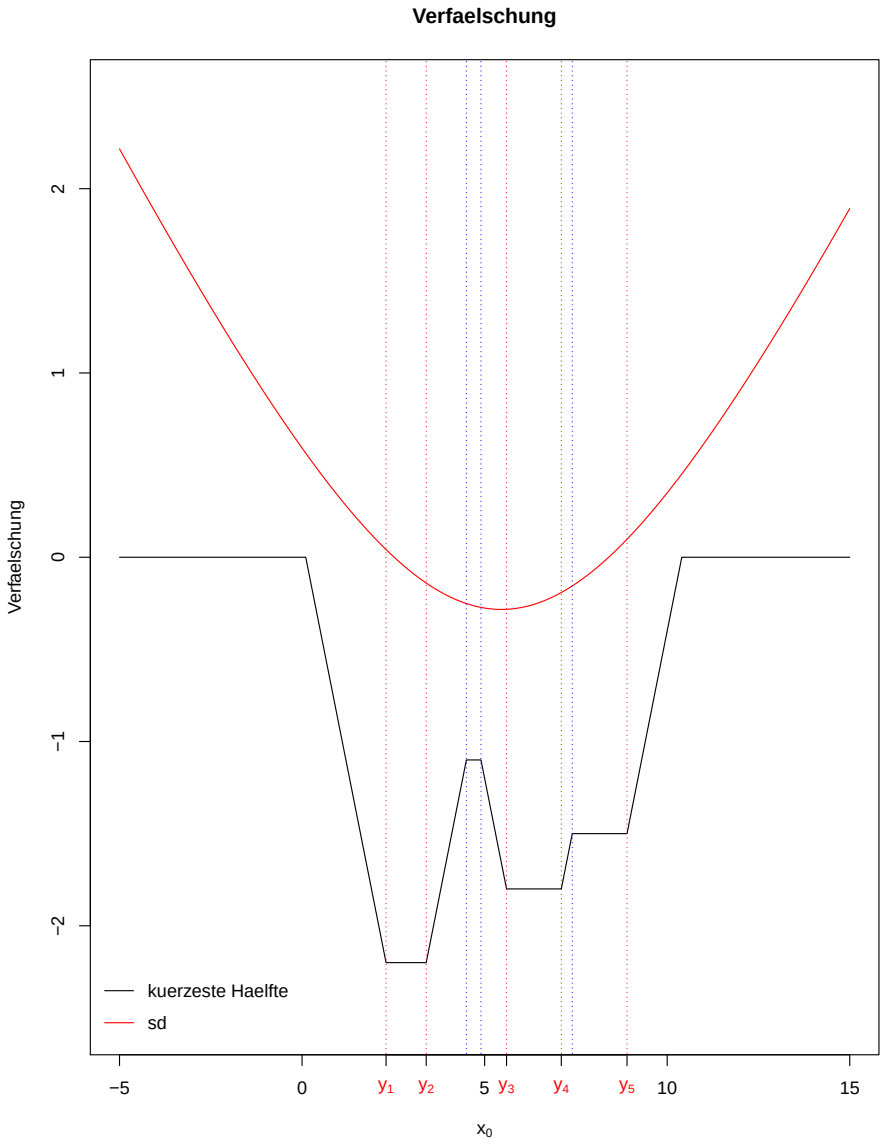
```
# verfaelschung: Funktion, welche die Verfaelschung durch Hinzufuegen berechnet
# Eingabe: data (numeric): Datenvektor, auf dessen Grundlage die Verfaelschung
#                          berechnet wird
#          fun (function): Schaetzfunktion, fuer die die Verfaelschung berechnet wird
#          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:





Es zeigt sich, dass die Verfälschungsfunktion der Standardabweichung eine Parabel darstellt. Diese ist nach oben unbeschränkt. Hingegen ist die Verfälschung der kürzesten Hälfte beschränkt und stückweise linear und konstant. Im mittleren Bereich sehen wir, dass die Verfälschung negativ ist, d.h. die neuen Schätzungen verkleinern sich stets, da wir mit Werten verfälschen, die innerhalb des Datensatzes liegen (sogenannte Inlier). Außerhalb des Datensatzes bezeichnen wir die Werte als Ausreißer (Outlier). Bei der Standardabweichung führen diese zu einer positiven und unbeschränkt wachsenden Verfälschung. Die kürzeste Hälfte reagiert gar nicht auf bei **einem** vorliegenden Ausreißer und besitzt bei betragsmäßig großen Ausreißern (außerhalb der Intervallgrenzen $\pm$ hinreichendem Abstand) eine Verfälschung von Null. Man kann sich überlegen, dass man die kürzeste Hälfte zum Explodieren bringen kann, indem man die Abstände zwischen den Ausreißern wachsen lässt. Bei gleichen Ausreißern würde sie sonst implodieren.



Ein tüchtiger Student hat in seiner Abgabe die Idee gebracht, die Steigungswechsel der Verfälschungsfunktion zu charakterisieren. Einerseits bilden die gegebenen Datenpunkte verschiedene Positionen mit Steigungswechsel. Andererseits hängt die Struktur der Verfälschungsfunktion zwischen den Daten von den Abständen der Datenpunkte im geordneten Datensatz ab. Besitzen die benachbarten Datenpunkte größere Abstände zum übernächsten Datenpunkt, so bleibt die Verfälschungsfunktion bis zum nächsten Datenpunkt konstant. Gibt es benachbarte Datenpunkte mit kleineren Abständen zum übernächsten Datenpunkt, so können sich Peaks wie zwischen y_2 und y_3 ergeben (Details in der Übung zur Bestimmung der blauen Linien).

Aufgabe 6.2:

zu zeigen: Die Lokations-Invarianz und Skalen-Äquivarianz des Skalen-LTS-Schätzers aus Definition 4.1.3.

Der Skalen-LTS-Schätzer ist gegeben durch:

$$\hat{\sigma}_{k,h}(y) = \min_{\mu \in \mathbb{R}} \sqrt{\frac{1}{h-k+1} \sum_{n=k}^{h} r_{(n)}(y,\mu)^2}.$$

- Zeige zunächst die Lokations-Invarianz: Sei dazu $z_n = y_n + l,\ n = 1, ..., N,\ l \in \mathbb{R}$.

Es ist bekannt bzw. leicht zu sehen, dass

$$\min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(z,\mu)^2 = \min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(y,\mu)^2,$$

da

$$|y_n - \mu| = |y_n + l - (\mu + l)| = |z_n - (\mu + l)|$$

und sich somit zwar die Stelle des Minimums um l verschiebt, wenn auf jede Beobachtung den Wert l draufaddiert, nicht jedoch der Wert des Minimums.

Somit gilt:

$$\begin{aligned}
\hat{\sigma}_{k,h}(z) &= \min_{\mu\in\mathbb{R}} \sqrt{\frac{1}{h-k+1} \sum_{n=k}^{h} r_{(n)}(z,\mu)^2} \\
&= \sqrt{\frac{1}{h-k+1} \min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(z,\mu)^2} \\
&= \sqrt{\frac{1}{h-k+1} \min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(y,\mu)^2} \\
&= \min_{\mu\in\mathbb{R}} \sqrt{\frac{1}{h-k+1} \sum_{n=k}^{h} r_{(n)}(y,\mu)^2} \\
&= \hat{\sigma}_{k,h}(y)
\end{aligned}$$

Somit ist $\hat{\sigma}_{k,h}(y)$ lokations-invariant.

- Zeige nun die Skalen-Äquivarianz: Sei dazu $z_n = s \cdot y_n,\ s > 0,\ n = 1, ..., N$

  Es gilt

  $$|z_n - s\mu| = |sy_n - s\mu| = s|y_n - \mu|,$$

  woraus folgt:

  $$\min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(z,\mu)^2 = \min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} s^2 r_{(n)}(y,\mu)^2 = s^2 \min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(y,\mu)^2.$$

  Somit gilt:

$$\begin{aligned}
\hat{\sigma}_{k,h}(z) &= \min_{\mu\in\mathbb{R}} \sqrt{\frac{1}{h-k+1} \sum_{n=k}^{h} r_{(n)}(z,\mu)^2} \\
&= \sqrt{\frac{1}{h-k+1} \min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(z,\mu)^2} \\
&= \sqrt{\frac{1}{h-k+1} s^2 \min_{\mu\in\mathbb{R}} \sum_{n=k}^{h} r_{(n)}(y,\mu)^2} \\
&= s \min_{\mu\in\mathbb{R}} \sqrt{\frac{1}{h-k+1} \sum_{n=k}^{h} r_{(n)}(y,\mu)^2} \\
&= s\hat{\sigma}_{k,h}(y)
\end{aligned}$$

Somit ist $\hat{\sigma}_{k,h}(y)$ skalen-äquivariant.

Aufgabe 6.3:

zu zeigen: $s(y)^2 = \frac{1}{2N(N-1)} \sum_{n=1}^{N} \sum_{m=1}^{N} (y_n - y_m)^2$.

Erste Variante (mit dem Kopf durch die Wand):

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

Zweite Variante (mit Nulladdition):

$$\begin{aligned}
&\frac{1}{2N(N-1)} \sum_{n=1}^{N} \sum_{m=1}^{N} (y_n - y_m)^2 \\
=&\frac{1}{2N(N-1)} \sum_{n=1}^{N} \sum_{m=1}^{N} \left((y_n - \bar{y})^2 + (y_m - \bar{y})^2 - 2(y_n - \bar{y})(y_m - \bar{y}) \right) \\
=&\frac{1}{2N(N-1)} N \left(\sum_{n=1}^{N} (y_n - \bar{y})^2 + \sum_{m=1}^{N} (y_m - \bar{y})^2 \right) + \frac{1}{N(N-1)} \Big(\underbrace{\sum_{n=1}^{N} (y_n - \bar{y})}_{=0} \underbrace{\sum_{m=1}^{N} (y_m - \bar{y})}_{=0} \Big) \\
=&\frac{2N}{2N(N-1)} \sum_{n=1}^{N} (y_n - \bar{y})^2 = \frac{1}{(N-1)} \sum_{n=1}^{N} (y_n - \bar{y})^2 = s(y)^2
\end{aligned}$$