Robuste Statistik

Prof. Dr. Christine Müller
Technische Universität Dortmund

Sommersemester 2019

Inhaltsverzeichnis

Kapitel 1

Einleitung

Datensätze enthalten oft einige wenige "falsche" Daten y_n. Solche "falschen" Daten können dadurch entstehen, dass bei Messungen ein Fehler gemacht wurde oder dass Beobachtungseinheiten (z.B. Personen) zur Population dazu genommen wurden, die nicht dazu gehören. Z.B. wenn in einer Studie Personen mit einer bestimmten Krankheit untersucht werden, kann es passieren, dass in der Studie auch Personen aufgenommen werden, die gar nicht diese Krankheit besitzen. Die "falschen" Daten können dadurch auffallen, dass sie stark von den anderen Daten abweichen. Solche stark von den anderen abweichende Daten werden **Ausreißer** genannt. Aber falsche Daten müssen nicht immer als Ausreißer auffallen. Umgekehrt können Ausreißer auch "richtige" Daten sein. So ist es eine sehr schlechte Praxis, wenn per Hand die Ausreißer aussortiert werden. Das ist vor allem deshalb schlecht, weil willkürlich dann beschlossen wird, was Ausreißer sind.

Viel besser ist es daher, statistische Verfahren zu benutzen, die wenig durch einige wenige Ausreißer beeinflusst werden. Dann ist es egal, ob die auftretenden Ausreißer "falsche" oder "richtige" Daten sind. Und die falschen Daten, die wenig von den anderen Daten abweichen, spielen sowieso keine Rolle.

Mittelwert und Median sind bekannte Kennzahlen für die Lage. Dabei gilt der Median als robust gegenüber von Ausreißern. Allerdings stellt sich die Frage, wie Ausreißerrobustheit überhaupt definiert und gemessen werden kann. Wir werden hier verschiedene Definitionen von Ausreißerrobustheit kennenlernen.

Eine weitere Fragestellung ist, ob der Median mit seiner Ausreißerrobustheit auch auf andere Schätzprobleme der Statistik übertragbar ist. Was könnte z.B. der Median für multivariate Daten sein? Wir werden daher auch verschiedene Schätzmethoden kennenlernen und insbesondere verschiedene Möglichkeiten, den Median auf multivariate Daten und Regressionsprobleme zu verallgemeinern.

Wir untersuchen zuerst einmal, wie sich Ausreißer auf das arithmetische Mittel und den Median auswirken.

1.1 Mittelwert und Median im Vergleich

Es seinen $y_1, \ldots, y_N$ Beobachtungen eines quantitativen Merkmals Y und $y = (y_1, \ldots, y_N)^\top$ bezeichnet den Datenvektor. Bekannte Lageschätzfunktionen sind durch das arithmetisches Mittel

$$\overline{y} = \frac{1}{N} \sum_{n=1}^{N} y_n$$

und den Median gegeben. Da der Median $\text{med}(y)$ oft nicht eindeutig ist, wird er hier als Menge definiert, für die gilt:

$$\text{med}(y) := \widetilde{y}_{0.5} := \left\{ \mu \in \mathbb{R};\ \sharp\{n;\ y_n \leq \mu\} \geq \frac{N}{2} \leq \sharp\{n;\ y_n \geq \mu\} \right\}.$$

Ist $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$ der geordnete Datensatz, so gilt für den Median $\widetilde{y}_{0.5}$

$$\widetilde{y}_{0.5} = \begin{cases} \{y_{(k)}\}, & k = \frac{N+1}{2},\ N \textbf{ ungerade}, \\ [y_{(k)}, y_{(k+1)}], & k = \frac{N}{2},\ N \textbf{ gerade}. \end{cases}$$

Bei Eindeutigkeit des Medians werden aber oft die Mengenklammern weggelassen.

1.1.1 Beispiel

Wir betrachten folgenden geordneten Datensatz

$$2, 3, 5, 6, 9.$$

Der Wert 5 ist hier der Median und der Mittelwert. Wird ein Wert y_0 dazugefügt bzw. wird 5 durch y_0 ersetzt erhalten wir:

	Hinzufügen von y_0		Ersetzen von 5 durch y_0	
y_0	Mittelwert	Median	Mittelwert	Median
-10000	-1662.50	$[3, 5]$	-1996.0	3
-1000	-162.50	$[3, 5]$	-196.0	3
-100	-12.50	$[3, 5]$	-16.0	3
-10	2.50	$[3, 5]$	2.0	3
2	4.50	$[3, 5]$	4.4	3
3	$4.6\overline{6}$	$[3, 5]$	4.6	3
5	5.00	5	5.0	5
6	$5.1\overline{6}$	$[5, 6]$	5.2	6
9	$5.6\overline{6}$	$[5, 6]$	5.8	6
10	$5.8\overline{3}$	$[5, 6]$	6.0	6
100	$20.8\overline{3}$	$[5, 6]$	24.0	6
1000	$170.8\overline{3}$	$[5, 6]$	204.0	6
10000	$1670.8\overline{3}$	$[5, 6]$	2004.0	6

Es zeigt sich, dass der Median sowohl beim Hinzufügen als durch das Ersetzen von einem einzigen höchstens um 2 vom ursprünglichen Wert von 5 abweicht, während das arithmetische Mittel durch

einen Ausreißer beliebig stark verfälscht werden kann. Wir sehen außerdem, dass das Ersetzen die gleiche Auswirkung hat wie das Hinzufügen von x_0. Wir werden später sehen, dass in manchen Fällen das Ersetzen einfacher mathematisch zu handhaben ist. Wir betrachten daher jetzt nur das Ersetzen und ersetzen jetzt nicht nur die 5 sondern zusätzlich auch die 6 (und bei dreifachem Ersetzen die 9) durch y_0:

	Ersetzen von 5 und 6 durch y_0		Ersetzen von 5, 6 und 9 durch y_0	
y_0	Mittelwert	Median	Mittelwert	Median
-10000	-3997.2	2	-5999.0	-10000
-1000	-397.2	2	-599.0	-1000
-100	-37.2	2	-59.0	-100
-10	-1.2	2	-5.0	-10
2	3.6	2	2.2	2
3	4	3	2.8	3
5	4.8	5	4	5
6	5.2	6	4.6	6
9	6.4	9	6.4	9
10	6.8	9	7.0	10
100	42.8	9	61.0	100
1000	402.8	9	601.0	1000
10000	4002.8	9	6001.0	10000

Gleiches Verhalten würde man erhalten, wenn 5, 3 (und 2) ersetzt würden. Wir sehen, dass der Median nur beliebig stark verfälscht werden kann, wenn 3 Werte ersetzt werden. Das liegt daran, dass der Median sich immer nach der Mehrheit der Daten richtet.

1.2 Erste Robustheitsdefinitionen

Um statistische Methoden bezüglich ihrer Ausreißer-Robustheit vergleichen zu können, wird ein quantitatives Maß für die Ausreißer-Robustheit gebraucht. In der Vergangenheit wurden verschiedene Robustheitsmaße vorgeschlagen. Viele basieren auf dem Begriff der Verfälschung (Bias). Wir hatten beim obigen Beispile gesehen, dass der Mittelwert durch einen einzelnen Ausreißer beliebig verfälscht werden kann, während die Verfälschung des Median bei dem obigen Beispiel nicht größer als 2 wurde.

1.2.1 Definition (Verfälschungs-Funktion für einen Datensatz)
Die Verfälschungs-Funktion gibt für jeden Wert y_0 an, wie sich die Schätzung an einem Datensatz ändert, wenn der Wert y_0 hinzugefügt wird (Verfälschungs-Funktion durch Hinzufügen), bzw. wie sich die Schätzung an einem Datensatz maximal ändert, wenn ein beliebiger Beobachtungswert durch y_0 ersetzt wird (Verfälschungs-Funktion durch Ersetzen). Der maximale Wert der Verfälschungs-Funktion im Absolutbetrag ist die maximale Verfälschung.

Die Verfälschungs-Funktion hat den Nachteil, dass sie sehr vom Datensatz abhängt. Wir werden daher später noch eine asymptotische Verfälschungs-Funktion betrachten. Aber ein einfaches Robustheitmaß, der Bruchpunkt, hängt oft nicht vom Datensatz ab.

1.2.2 Definition
Der Bruchpunkt ist allgemein der kleinste Anteil von Ausreißer, der eine Schätzung beliebig fatal verfälschen kann.

Diese Definition von Bruchpunkt kann für beliebig komplexe Fragestellungen benutzt werden. Es muss jeweils nur geklärt werden, was

1. ein Ausreißer ist,
2. eine beliebige fatale Verfälschung ist.

Was eine fatale Verfälschung ist, ist schon bei Lage- und Streuungsschätzungen verschieden. Deshalb werden die formalen Definitionen später in Abschnitt 3.4 und Abschnitt 4.2 gegeben. Dort wird dann gezeigt, dass der Bruchpunkt des arithmetischen Mittels und der Standardabweichung $\frac{1}{N}$ ist, während es andere Schätzungen gibt, die einen Bruchpunkt von ca. $\frac{1}{2}$ besitzen.

Teil I

Univariate Daten

Kapitel 2

Äquivarianz- und Invarianz-Anforderungen an Schätzfunktionen

2.0.1 Definition (Schätzfunktion basierend auf univariaten Daten)
Eine Schätzfunktion für eine p-dimensionale Kennzahl θ basierend auf univariaten Daten ist eine Funktion

$$\hat{\theta} : \mathbb{R}^N \ni y = (y_1, \ldots, y_N)^\top \longrightarrow \hat{\theta}(y) \in \mathbb{R}^p.$$

Dabei heißt $\hat{\theta}(y)$ die Schätzung für θ beim Datensatz y.

In vielen Fällen wie dem Median sind Schätzungen nicht eindeutig definiert. Das bedeutet, dass zu einem Datensatz y mehrere Werte $\hat{\theta}$ in Frage kommen. Dann kann $\hat{\theta}$ keine Funktion nach $\mathbb{R}^p$ sein, sondern ist eine Funktion nach $\mathcal{P}(\mathbb{R}^p)$, der Potenzmenge von $\mathbb{R}^p$, d.h. der Menge aller Teilmengen von $\mathbb{R}^p$. Solche Schätzfunktionen werden wir mengenwertige Schätzfunktionen nennen.

2.0.2 Definition (Mengenwertige Schätzfunktion basierend auf univariaten Daten)
Eine mengenwertige Schätzfunktion für eine p-dimensionale Kennzahl θ basierend auf univariaten Daten ist eine Funktion

$$\hat{\theta} : \mathbb{R}^N \ni y = (y_1, \ldots, y_N)^\top \longrightarrow \hat{\theta}(y) \in \mathcal{P}(\mathbb{R}^p).$$

In diesem Fall ist $\hat{\theta}(y)$ die Menge aller möglichen Schätzwerte für θ beim Datensatz y.

2.0.3 Bemerkung
Mengenwertige Schätzfunktionen werden nur benötigt, wenn eventuell die Schätzung wie beim Median nicht eindeutig sein kann. Das muss insbesondere bei den Robustheitsmaßen (siehe z.B. Abschnitt 3.4 oder 4.2) berücksichtigt werden. Allerdings werden wir der einfachheitshalber oft nicht zwischen der Menge $\hat{\theta}(y)$ und deren Repräsentanten unterscheiden und den Repräsentanten auch mit $\hat{\theta}(y)$ bezeichnen.

Eine Schätzfunktion die immer den gleichen Wert c annimmt, d.h. es gilt $\hat{\theta}(y) = c$ für alle $y \in \mathbb{R}^N$, kann natürlich nicht durch Ausreißer verfälscht werden. Aber so eine Schätzfunktion ist nicht sinnvoll. Daher müssen wir festlegen, was vernünftige Schätzfunktionen sind, bevor wir deren Ausreißerrobustheit untersuchen können. Mit Äquivarianz- und Invarianz-Anforderungen werden sinnvolle Schätzfunktionen festgelegt.

Bei univariaten Daten sind die **Lage/Lokation** ($\rightarrow$ „Durchschnittswert“) und die **Streuung** die wesentlichen Kenngrößen, die von Interesse sind. Sie werden durch Lageparameter/Lageschätzungen/Lokationsschätzungen und Streuungsparameter/Streuungsschätzungen/Skalenschätzungen erfasst. Lageschätzungen dienen dabei zur Beschreibung des „Zentrums/Schwerpunktes“ des Datensatzes. Streuungsparameter ergänzen die in den Lageparametern enthaltene Information und dienen dazu, das Abweichungsverhalten (des Merkmals) in einer Population zu quantifizieren. Zur Interpretation von Streuungsparametern läßt sich festhalten: Je größer der Wert eines Streuungsparameters ist, desto mehr streuen die Beobachtungen. Ist der Wert klein, so sind die Beobachtungen um einen Wert konzentriert.

Will man die Lage des Datensatzes durch eine Kenngröße angeben, so sollte die zugehörige Lageschätzfunktion, auch Lokations-Schätzfunktion genannt, folgende Eigenschaft besitzen: Verschiebt man alle Daten um einen Betrag l, so sollte sich die Lage-Kenngröße auch um diesen Betrag ändern. Diese Eigenschaft wird **Lokations-Äquivarianz** genannt.

2.0.4 Definition (Lokations-Äquivarianz)
Eine (mengenwertige) Schätzfunktion $\hat{\theta} : \mathbb{R}^N \to \mathbb{R}$ (bzw. $\hat{\theta} : \mathbb{R}^N \to \mathcal{P}(\mathbb{R})$) heißt lokationsäquivariant, wenn für alle $y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ und alle $l \in \mathbb{R}$ gilt:

$$\hat{\theta}((y_1 + l, y_2 + l, \ldots, y_N + l)^\top) = \hat{\theta}(y) + l.$$

Dabei ist $A + l$ für eine Menge A und eine reelle Zahl l als

$$A + l = \{a + l;\ a \in A\}$$

definiert.

Soll dagegen die Kenngröße beschreiben, wie stark die Daten streuen, so sollte sich diese Kenngröße nicht ändern, wenn die Daten alle um l verschoben werden. Diese Eigenschaft wird **Lokations-Invarianz** genannt.

2.0.5 Definition (Lokations-Invarianz)
Eine (mengenwertige) Schätzfunktion $\hat{\theta} : \mathbb{R}^N \to \mathbb{R}$ (bzw. $\hat{\theta} : \mathbb{R}^N \to \mathcal{P}(\mathbb{R})$) heißt lokationsinvariant, wenn für alle $y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ und alle $l \in \mathbb{R}$ gilt:

$$\hat{\theta}((y_1 + l, y_2 + l, \ldots, y_N + l)^\top) = \hat{\theta}(y).$$

Werden alle Daten mit einem Faktor $s \in \mathbb{R}^+$ multipliziert, so sollte sowohl eine Lage-Kenngröße als auch eine Streuungs-Kenngröße entsprechend multipliziert werden. Diese Eigenschaft wird **Skalen-Äquivarianz** genannt.

2.0.6 Definition (Skalen-Äquivarianz)
Eine (mengenwertige) Schätzfunktion $\hat{\theta} : \mathbb{R}^N \to \mathbb{R}$ *(bzw.* $\hat{\theta} : \mathbb{R}^N \to \mathcal{P}(\mathbb{R})$*) heißt skalen-äquivariant, wenn für alle* $y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ *und alle* $s \in \mathbb{R}^+$ *gilt:*

$$\hat{\theta}((s\,y_1, s\,y_2, \ldots, s\,y_N)^\top) = s\,\hat{\theta}(y).$$

Dabei ist $s\,A$ für eine Menge A und eine reelle Zahl s als

$$s\,A = \{s\,a;\ a \in A\}$$

definiert.

Ebenso kann **Skalen-Invarianz** definiert werden.

2.0.7 Definition (Skalen-Invarianz)
Eine (mengenwertige) Schätzfunktion $\hat{\theta} : \mathbb{R}^N \to \mathbb{R}$ *(bzw.* $\hat{\theta} : \mathbb{R}^N \to \mathcal{P}(\mathbb{R})$*) heißt skalen-invariant, wenn für alle* $y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ *und alle* $s \in \mathbb{R}^+$ *gilt:*

$$\hat{\theta}((s\,y_1, s\,y_2, \ldots, s\,y_N)^\top) = \hat{\theta}(y).$$

Zusammenfassend werden folgende Forderungen gestellt: Die Schätzfunktion, die einen bestimmten Lageparameter ergibt, sollte lokations- und skalen-äquivariant sein. Die Schätzfunktion, die einen bestimmten Streuungsparameter ergibt, sollte lokations-invariant und skalen-äquivariant sein.

Kapitel 3

Lageschätzungen

3.1 Weitere bekannte Lageschätzungen

Bekannte Lageschätzungen neben dem arithmetischen Mittel und dem Median sind das Minimum $\min\{y_1, \ldots, y_N\}$, das Maximum $\max\{y_1, \ldots, y_N\}$ und das p-Quantil. Dabei ist das p-Quantil $\widetilde{y}_p$ oft nicht eindeutig, so dass es wieder als Menge definiert wird, für die gilt:

$$\widetilde{y}_p := \{\mu \in \mathbb{R};\ \sharp\{n;\ y_n \leq \mu\} \geq p\,N \text{ und } \sharp\{n;\ y_n \geq \mu\} \geq (1-p)\,N\}\,.$$

und es gilt

$$\widetilde{y}_p = \begin{cases} \{y_{(k)}\}, & Np < k < Np+1,\ Np \notin \mathbb{N}, \\ [y_{(k)}, y_{(k+1)}], & k = Np \in \mathbb{N}. \end{cases}$$

3.2 Charakterisierungen vom Median und arithmetischem Mittel

Um den Median auf multivariate Daten und Regressionsprobleme verallgemeinern zu können und weitere ausreißerrobuste Lageschätzungen zu gewinnen, die auch auf komplexere Situationen verallgemeinert werden können, sollen erstmal Charakterisierungen vom Median und dem arithmetischen Mittel hergeleitet werden.

Bei den folgenden Charakterisierungen wird die arg max-Schreibweise benutzt. Insbesondere, wenn die zu maximierende (bzw. zu minimierende) Funktion f von komplizierter Form ist, ist das eine sehr platzsparende Schreibweise. Im folgenden wird das oft von Vorteil sein.

3.2.1 Definition (arg max, arg min)
Sei $f : Z \to \mathbb{R}$ eine Funktion. Dann sei

$$\begin{aligned}\arg\max\{f(z);\ z \in Z\} &= \left\{z^*;\ \text{mit } f(z^*) = \max\{f(z);\ z \in Z\} = \max_{z\in Z} f(z)\right\},\\ \arg\min\{f(z);\ z \in Z\} &= \left\{z^*;\ \text{mit } f(z^*) = \min\{f(z);\ z \in Z\} = \min_{z\in Z} f(z)\right\}.\end{aligned}$$

D.h. $\arg\max\{f(z);\ z \in Z\}$ *(bzw.* $\arg\min\{f(z);\ z \in Z\}$*) ist die Menge aller Punkte $z \in Z$, bei denen die Funktion f maximal (bzw. minimal) wird.*

3.2.2 Satz
Für das arithmetische Mittel $\overline{y}$ gilt

$$\overline{y} = \arg\min_{\mu\in\mathbb{R}} \sum_{n=1}^{N} (y_n - \mu)^2,$$

d.h. die Funktion g definiert durch $g(\mu) = \sum_{n=1}^{N}(y_n - \mu)^2$ nimmt bei $\mu = \overline{y}$ ihr Minimum an.

3.2.3 Satz
Für den Median $med(y) = \widetilde{y}_{0.5}$ gilt

$$med(y) = \arg\min_{\mu\in\mathbb{R}} \sum_{n=1}^{N} |y_n - \mu|,$$

d.h. die Funktion g definiert durch $g(\mu) = \sum_{n=1}^{N} |y_n - \mu|$ nimmt bei $\mu = med(y) = \widetilde{y}_{0.5}$ ihr Minimum an.

Der folgende Satz 3.2.7 zeigt, dass der Median die größte **Lokations-Tiefe** besitzt.

3.2.4 Definition (Lokations-Tiefe (Location Depth))
Die Lokations-Tiefe $d_L(\mu, y)$ eines Lageparameters $\mu \in \mathbb{R}$ bezüglich des Datensatzes $y_1, \ldots, y_N$ ist definiert als

$$d_L(\mu, y) = \frac{1}{N} \min\left\{\sharp\{n;\ y_n \le \mu\},\ \sharp\{n;\ y_n \ge \mu\}\right\}.$$

Für spätere Betrachtungen ist es sinnvoll die Lokations-Tiefe über den sogenannten **Lokations-Nonfit** zu charakterisieren.

3.2.5 Definition (Lokations-Nonfit)
Ein Lageparameter $\mu \in \mathbb{R}$ ist ein Lokations-Nonfit bezüglich des Datensatzes $y_1, \dots, y_N$, falls es ein $\tilde{\mu}$ gibt mit $|y_n - \tilde{\mu}| < |y_n - \mu|$ für alle $n = 1, \dots, N$.

3.2.6 Satz
a) μ ist ein Nonfit bzgl. $y_1, \dots, y_N \iff \mu > y_n$ für alle $n = 1, \dots, N$ oder $\mu < y_n$ für alle $n = 1, \dots, N$, d.h. $d_L(\mu, y) = 0$.
b) Für die Lokations-Tiefe $d_L(\mu, y)$ gilt

$$d_L(\mu, y) = \frac{1}{N} \min \{M; \quad \text{Es existieren } n_1, \dots, n_M \in \{1, \dots, N\}, \text{ so dass } \mu \text{ ein Lokations-Nonfit bezüglich } \{y_1, \dots, y_N\} \setminus \{y_{n_1}, \dots, y_{n_M}\} \text{ ist. } \}.$$

3.2.7 Satz
a) Ein Lageparameter μ_* ist ein Median genau dann, wenn er die Lokations-Tiefe maximiert, d.h. wenn

$$med(y) = \arg \max_{\mu \in \mathbb{R}} d_L(\mu, y)$$

gilt.
b) Sind $y_1, \dots, y_N$ paarweise verschieden, so gilt

$$\max_{\mu \in \mathbb{R}} d_L(\mu, y) = \begin{cases} \frac{1}{2}, & \text{für } N \text{ gerade}, \\ \frac{N+1}{2N}, & \text{für } N \text{ ungerade}. \end{cases}$$

Beweis.
a) 1. Fall: $\text{med}(y)$ ist eindeutig. Dann ist entweder N ungerade oder es gibt mehr als eine Beobachtung bei $\text{med}(y)$, d.h. es gilt

$$d_1 = \sharp\{n;\ y_n \geq \text{med}(y)\} \geq \frac{N+1}{2}, \quad d_2 = \sharp\{n;\ y_n \leq \text{med}(y)\} \geq \frac{N+1}{2}$$

und für jedes $\mu \neq \text{med}(y)$ gilt

$$\tilde{d}_1 = \sharp\{n;\ y_n \geq \mu\} < \frac{N}{2} \quad \text{oder} \quad \tilde{d}_2 = \sharp\{n;\ y_n \leq \mu\} < \frac{N}{2}.$$

Daraus folgt

$$N\, d_L(\text{med}(y), y) = \min\{d_1, d_2\} \geq \frac{N}{2} > \min\{\tilde{d}_1, \tilde{d}_2\} = N\, d_L(\mu, y)$$

für alle $\mu \neq \text{med}(y)$.

2. Fall: med(y) ist nicht eindeutig, d.h. es gibt $\mu_1 < \mu_2$, so dass jedes $\mu_0 \in [\mu_1, \mu_2]$ ein Median ist. Dann muss aber N gerade sein und es gilt

$$\sharp\{n;\ y_n \le \mu_0\} = \frac{N}{2} \text{ für alle } \mu_0 \in [\mu_1, \mu_2), \qquad \sharp\{n;\ y_n \le \mu_0\} > \frac{N}{2} \text{ für alle } \mu_0 \in [\mu_2, \infty),$$
$$\sharp\{n;\ y_n \ge \mu_0\} = \frac{N}{2} \text{ für alle } \mu_0 \in (\mu_1, \mu_2], \qquad \sharp\{n;\ y_n \ge \mu_0\} > \frac{N}{2} \text{ für alle } \mu_0 \in (-\infty, \mu_1].$$

Also gilt

$$N\, d_L(\mu_0, y) = \min\{\sharp\{n;\ y_n \le \mu_0\}, \sharp\{n;\ y_n \ge \mu_0\}\} = \frac{N}{2}$$

für alle $\mu_0 \in [\mu_1, \mu_2]$. Ist $\mu < \mu_1$, so gilt $\sharp\{n;\ y_n \le \mu\} < \frac{N}{2}$ und somit $N\, d_L(\mu, y) < \frac{N}{2}$. Analog folgt $N\, d_L(\mu, y) < \frac{N}{2}$ für $\mu > \mu_2$.

b) Folgt aus a). □

3.3 Vom Median und Mittelwert abgeleitete Schätzfunktionen

Median und arithmetisches Mittel sind Spezialfälle der M-Schätzungen. Das arithmetische Mittel ist außerdem ein Spezialfall der LTS-Schätzungen.

3.3.1 Definition (Lokations-M-Schätzung (Huber 1964))
Die Lokations-M-Schätzung $\hat{\mu}_\rho(y)$ bezüglich $\rho : \mathbb{R} \to \mathbb{R}$ basierend auf den quantitativen Daten $y_1, \ldots, y_N$ ist definiert durch

$$\hat{\mu}_\rho(y) = \arg\min_{\mu \in \mathbb{R}} \sum_{n=1}^{N} \rho(y_n - \mu).$$

3.3.2 Bemerkung
Allgemeiner können M-Schätzungen für einen Parameter θ basierend auf Beobachtungen aus $\mathbb{R}^r$ bezüglich einer Score-Funktion $\rho : \mathbb{R}^r \times \Theta \to \mathbb{R}$ mittels

$$\hat{\theta}_\rho(y) = \arg\min_{\theta \in \Theta} \sum_{n=1}^{N} \rho(y_n, \theta)$$

definiert werden. Ist $\rho(z, \theta) = -\log(f_\theta(z))$, wobei $f_\theta(z))$ die Dichte einer Verteilung P_θ ist, so ist die M-Schätzung eine Maximum-Likelihood-Schätzung. Damit sind M-Schätzungen Verallgemeinerungen von ML-Schätzungen. Daher auch der Name.

3.3.3 Satz
Sei $\rho : \mathbb{R} \to \mathbb{R}$ stetig und bis auf endlich viele Punkte differenzierbar, so dass ψ gegeben durch $\psi(z) = \rho'(z)$ monoton wachsend ist. Außerdem sei

$$\begin{aligned} \hat{\mu}^*(y) &= \sup\{\mu;\ \sum_{n=1}^{N} \psi(y_n - \mu) > 0\}, \\ \hat{\mu}^{**}(y) &= \inf\{\mu;\ \sum_{n=1}^{N} \psi(y_n - \mu) < 0\}. \end{aligned}$$

Dann ist $\hat{\mu}_\rho(y)$ eine Lokations-M-Schätzung bezüglich ρ genau dann, wenn $\hat{\mu}^*(y) \le \hat{\mu}_\rho(y) \le \hat{\mu}^{**}(y)$ gilt.

3.3.4 Beispiel (Huber-M-Schätzung, Huber 1964)
Einen Kompromiss zwischen arithmetischen Mittel und Median bildet die Huber-M-Schätzung mit folgender Score-Funktion ρ und dessen Ableitung ψ:

$$\rho(z) = \begin{cases} \frac{1}{2} z^2, & \text{für } |z| \le c \\ c|z| - \frac{c^2}{2}, & \text{für } |z| > c \end{cases} \implies \psi(z) = \begin{cases} -c, & \text{für } z < -c \\ z, & \text{für } |z| \le c \\ c, & \text{für } z > c \end{cases} .$$

3.3.5 Definition (Lokations-LTS-Schätzung (LTS = Least Trimmed Squares))
Die Lokations-LTS-Schätzung $\hat{\mu}_{k,h}(y)$ bezüglich k und h basierend auf den quantitativen Daten $y_1, \dots, y_N$ ist definiert durch

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

Dabei ist $r_n(y, \mu) = |y_n - \mu|$ das sogenannte absolute **Residuum** *der n'ten Beobachtung und $r_{(1)}(y, \mu) \le r_{(2)}(y, \mu) \le \ldots \le r_{(N)}(y, \mu)$ sind die geordneten absoluten Residuen.*

3.3.6 Bemerkung
• Nach Satz 3.2.3 ist die Lokations-M-Schätzung $\hat{\mu}_\rho$ mit $\rho(z) = |z|$ der Median und nach Satz 3.2.2 ist die Lokations-M-Schätzung $\hat{\mu}_\rho$ mit $\rho(z) = z^2$ das arithmetische Mittel.
• Mit $k = 1$ und $h = N$ ist die Lokations-LTS-Schätzung $\hat{\mu}_{k,h}$ das arithmetische Mittel.
• Ist $k = \lfloor \frac{N+1}{2} \rfloor$ und $h = \lceil \frac{N+1}{2} \rceil$ so wird die LTS-Schätzung auch Lokations-LMS-Schätzung genannt (LMS=least median of squares).

3.3.7 Bemerkung (Berechnung der LTS-Schätzung)
Die Lokations-LTS-Schätzung bezüglich $k = 1$ und $h < N$ kann wie folgt berechnet werden. Man bilde alle h-elementigen Teilmengen $M = \{n_1, \dots, n_h\}$ von $\{1, \dots, N\}$ und damit alle Teildatensätze $y(M) := (y_{n_1}, \dots, y_{n_h})^\top$ von $y = (y_1, \dots, y_N)^\top$ und betrachte

$$Q(\mu, y(M)) = \sum_{i=1}^{h} (y_{n_i} - \mu)^2.$$

Nach Satz 3.2.2 wird $Q(\mu, y(M))$ minimal, wenn μ das arithmetische Mittel $\overline{y(M)} = \frac{1}{h}\sum_{i=1}^{h} y_{n_i}$ von $(y_{n_1}, \ldots, y_{n_h})$ ist. Die Lokations-LTS-Schätzung $\hat{\mu}_{1,h}$ ist dann das arithmetische Mittel des Teildatensatzes, bei dem $Q(\overline{y(M)}, y(M))$ minimal wird, d.h.

$$\hat{\mu}_{1,h}(y) = \overline{y(M_0)} \text{ mit } M_0 \in \arg\min\{Q(\overline{y(M)}, y(M));\ M \subset \{1, \ldots, N\} \text{ h-elementig}\}.$$

Das Betrachten aller h-elementigen Teilmengen von $\{1, \ldots, N\}$ ist aber sehr aufwendig. Da wir die Daten nach der Größe ordnen können, so dass wir $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(N)}$ erhalten, reicht es, nur die Teildatensätze

$$(y_{(1)}, \ldots, y_{(h)})^\top,\ (y_{(2)}, \ldots, y_{(h+1)})^\top,\ \ldots,\ (y_{(N-h+1)}, \ldots, y_{(N)})^\top \tag{3.1}$$

zu betrachten. Das liegt daran, dass $Q(\overline{y(M)}, y(M))$ größer wird, wenn die Streuung des Teildatensatzes $y(M)$ größer wird (siehe auch Abschnitt 4.1). Somit muss also zur Berechnung der Lokations-LTS-Schätzung nur $N - h + 1$-mal das arithmetische Mittel $\overline{y(M)}$ der Teildatensätze (3.1) und das zugehörige $Q(\overline{y(M)}, y(M))$ berechnet werden.

3.3.8 Definition (Getrimmter Mittelwert)
Sei $y_{(1)} \leq \ldots \leq y_{(N)}$ die geordnete Stichprobe und $\beta N \in \mathbb{N}$ für $\beta \in (0, \frac{1}{2})$. So ist das β-getrimmte Mittel durch

$$\overline{y}_\beta = \frac{1}{N - 2\,\beta N} \sum_{n=\beta N+1}^{N-\beta N} y_{(n)}$$

gegeben.

3.3.9 Satz
a) Die Schätzfunktionen, die durch das Minimum, das Maximum, die p-Quantile, die Lokations-LTS-Schätzungen und durch das β-getrimmte Mittel gegeben sind, sind lokations- und skalen-äquivariant.
b) Schätzfunktionen, die durch Lokations-M-Schätzungen gegeben sind, sind lokations-äquivarinat aber in der Regel nicht skalen-äquivariant.

Die in Satz 3.3.9 genannten Schätzfunktionen sind also sinnvolle Schätzungen für die Lage eines univariaten Datensatzes.

3.4 Bruchpunkte von Lageschätzungen

Um die Ausreißer-Robustheit messen zu können, wurde in Abschnitt 1.2 der Bruchpunkt allgemein definiert. Für Lageschätzungen muss danach nur noch geklärt werden, was Ausreißer und was beliebige fatale Verfälschungen sind.

Bei der Bestimmung der Lage eines univariaten Datensatzes liegt eine beliebige Verfälschung vor, wenn die Schätzung durch Ausreißer dazu gebracht wird, gegen ∞ oder $-\infty$ zu streben. Ausreißer sind in diesem Fall sehr große oder sehr kleine Werte. Im Datensatz $\overline{y} = (1.5, 2.2, 3.1, 3.2, 3.5, 100)$

ist 100 ein Ausreißer und das arithmetische Mittel ist 18.92. Wird der Ausreißer 100 durch noch größere Werte ersetzt, so wird das arithmetische Mittel immer größer. Strebt dieser Ausreißer gegen ∞, so strebt auch das arithmetische Mittel gegen ∞ und wir sagen, es bricht zusammen. Das arithmetische Mittel kann also durch einen Ausreißer beliebig verfälscht werden. Damit ist der kleinste Anteil von Ausreißern, der das arithmetische Mittel verfälschen kann, $\frac{1}{N}$. Somit ist der Bruchpunkt des arithmetischen Mittels $\frac{1}{N}$. Konvergiert der Stichprobenumfang N gegen ∞, so konvergiert dieser Bruchpunkt gegen 0. Wir sehen auch, dass der Bruchpunkt gar nicht vom speziellen Ausgangsdatensatz $y = (1.5, 2.2, 3.1, 3.2, 3.5, 4.2)$ abhängt.

Ganz anders verhält sich der Median. Das Austauschen des Wertes 4.2 durch 100 ändert den Median überhaupt nicht. Auch wenn wir zusätzlich noch 3.5 durch 100 austauschen, so dass wir den Datensatz $\overline{y} = (1.5, 2.2, 3.1, 3.2, 100, 100)$ erhalten, ändert der Median sich nicht. Das bleibt auch der Fall wenn die beiden Werte 100 durch noch größere Werte ausgetauscht werden. Tauschen wir zwei andere Daten aus dem Datensatz $y = (1.5, 2.2, 3.1, 3.2, 3.5, 4.2)$ durch beliebig große oder kleine Werte aus, so wird der Median immer im Intervall $[1.5, 4.2]$ liegen, so dass er nie beliebig groß oder klein werden kann. Das bedeutet, dass zwei Ausreißer den Median in diesem Fall nicht beliebig verfälschen können. Erst drei Ausreißer können den Median in diesem Fall beliebig verfälschen, denn jeder Punkt im Intervall $[3.1, 100]$ ist ein Median des Datensatzes $\overline{y} = (1.5, 2.2, 3.1, 100, 100, 100)$. Gehen die drei Ausreißer gegen ∞ so wird das Intervall, dass den Median ergibt, immer größer und enthält damit Werte, die auch gegen ∞ streben. Bei sechs Daten können also drei Ausreißer den Median beliebig verfälschen und somit ist der Bruchpunkt dann $\frac{3}{6} = \frac{1}{2}$. Als Übungsaufgabe soll gezeigt werden, dass der Bruchpunkt des Median im allgemeinen $\frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor$ ist, wobei $\lfloor z \rfloor = \max\{n \in \mathbb{N};\ n \leq z\}$ definiert wird, d.h. $\lfloor z \rfloor$ ist die nach unten gerundete Zahl z.

3.4.1 Definition (Explosionspunkt)
Der Explosionspunkt $\epsilon^+(\hat{\theta}, y)$ einer Schätzfunktion $\hat{\theta}$ beim Datensatz y ist definiert als

$$\epsilon^+(\hat{\theta}, y) = \frac{1}{N} \min \left\{ M;\ \sup_{\tilde{y} \in \mathcal{Y}_M(y)} |\hat{\theta}(\tilde{y})| = \infty \right\},$$

wobei $\mathcal{Y}_M(y) = \left\{ (\tilde{y}_1, \ldots, \tilde{y}_N)^\top \in \mathbb{R}^N;\ \sharp\{n;\ y_n \neq \tilde{y}_n\} \leq M \right\}$ die Menge aller Datensätze $\tilde{y}$ ist, die sich in höchstens M Einzelwerten vom Datensatz y unterscheiden. Ist $\hat{\theta}$ eine mengenwertige Schätzfunktion, so sei $\sup_{\tilde{y} \in \mathcal{Y}_M(y)} |\hat{\theta}(\tilde{y})| := \sup_{\tilde{y} \in \mathcal{Y}_M(y)} \sup_{\theta \in \hat{\theta}(\tilde{y})} |\theta|$.

3.4.2 Definition (Bruchpunkt einer Lageschätzung)
Der Bruchpunkt $\epsilon^*(\hat{\mu}, y)$ einer Lageschätzfunktion $\hat{\mu}$ beim Datensatz y ist der Explosionspunkt, d.h.

$$\epsilon^*(\hat{\mu}, y) = \epsilon^+(\hat{\mu}, y).$$

3.4.3 Satz
Ist die Schätzfunktion $\hat{\theta}$ lokations-äquivariant, so gilt für den Explosionspunkt

$$\epsilon^+(\hat{\theta}, y) \leq \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor.$$

Beweis. Sei $M = \left\lfloor \frac{N+1}{2} \right\rfloor$. Angenommen es gibt $K \in \mathbb{R}$ mit $|\hat{\theta}(\tilde{y})| \leq K$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$. Betrachte $\tilde{y}^{1l}$ mit $\tilde{y}_n^{1l} = y_n + l$ für $n = 1, \ldots, M$ und $\tilde{y}_n^{1l} = y_n$ für $n = M+1, \ldots, N$. Dann gilt $\tilde{y}^{1l} \in \mathcal{Y}_M(y)$ und somit $|\hat{\theta}(\tilde{y}^{1l})| \leq K$ für alle $l \in \mathbb{R}$. Betrachte ebenso $\tilde{y}^{2l}$ mit $\tilde{y}_n^{2l} = y_n$ für $n = 1, \ldots, M$ und $\tilde{y}_n^{2l} = y_n - l$ für $n = M+1, \ldots, N$. Wegen $N - M \leq M$ gilt auch $\tilde{y}^{2l} \in \mathcal{Y}_M(y)$ und somit $|\hat{\theta}(\tilde{y}^{2l})| \leq K$ für alle $l \in \mathbb{R}$. Außerdem gilt $\tilde{y}_n^{1l} = \tilde{y}_n^{2l} + l$ für alle $n = 1, \ldots, N$, so dass die Lokations-Äquivarianz $\hat{\theta}(\tilde{y}^{1l}) = \hat{\theta}(\tilde{y}^{2l}) + l$ für alle $l \in \mathbb{R}$ liefert. Daraus folgt der Widerspruch

$$2K \geq |\hat{\theta}(\tilde{y}^{1l})| + |\hat{\theta}(\tilde{y}^{2l})| \geq |\hat{\theta}(\tilde{y}^{1l}) - \hat{\theta}(\tilde{y}^{2l})| = |l|$$

für alle $l \in \mathbb{R}$. Also gilt $\sup_{\tilde{y} \in \mathcal{Y}_M(y)} |\hat{\theta}(\tilde{y})| = \infty$ und somit $\epsilon^+(\hat{\theta}, y) \leq \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor$. □

3.4.4 Satz (Siehe auch Huber (1981))
Sei $\rho : \mathbb{R} \to \mathbb{R}$ stetig und bis auf endlich viele Punkte differenzierbar, so dass ψ gegeben durch $\psi(z) = \rho'(z)$ monoton wachsend ist. Für den Bruchpunkt der Lokations-M-Schätzfunktion $\hat{\mu}_\rho$ bezüglich ρ gilt dann:
a) Ist ψ unbeschränkt, dann gilt für alle $y \in \mathbb{R}^N$

$$\epsilon^*(\hat{\mu}_\rho, y) = \frac{1}{N}.$$

b) Aus $\psi(\infty) := \lim_{z\to\infty} \psi(z) < \infty$ und $\psi(-\infty) := \lim_{z\to-\infty} \psi(z) > -\infty$ folgt für alle $y \in \mathbb{R}^N$

$$\epsilon^*(\hat{\mu}_\rho, y) \geq \frac{1}{N}\left\lfloor \frac{\eta}{\eta+1} N \right\rfloor \quad \textit{mit} \quad \eta = \min\left\{ -\frac{\psi(-\infty)}{\psi(\infty)}, -\frac{\psi(\infty)}{\psi(-\infty)} \right\}.$$

Beweis.
Nach Satz 3.3.3 gilt $\hat{\mu}^*(y) \leq \hat{\mu}_\rho(y) \leq \hat{\mu}^{**}(y)$ mit

$$\begin{aligned} \hat{\mu}^*(y) &= \sup\{\mu;\ \sum_{n=1}^{N} \psi(y_n - \mu) > 0\}, \\ \hat{\mu}^{**}(y) &= \inf\{\mu;\ \sum_{n=1}^{N} \psi(y_n - \mu) < 0\}. \end{aligned}$$

a) O.B.d.A. sei $\lim_{z\to\infty} \psi(z) = \infty$. Angenommen es gibt $K \in \mathbb{R}$ mit $\hat{\mu}^*(\tilde{y}) < K$ für alle $\tilde{y} \in \mathcal{Y}_1(y)$. Dann gilt

$$\sum_{n=1}^{N} \psi(\tilde{y}_n - \mu) \leq 0 \ \text{ für alle } \ \tilde{y} \in \mathcal{Y}_1(y) \ \text{ und } \ \mu > K.$$

Für $\tilde{y}^l \in \mathcal{Y}_1(y)$ mit $\tilde{y}^l_N = l$ gilt aber

$$\sum_{n=1}^{N} \psi(\tilde{y}^l_n - \mu) = \sum_{n=1}^{N-1} \psi(y_n - \mu) + \psi(l - \mu) \stackrel{l\to\infty}{\longrightarrow} \infty,$$

was ein Widerspruch ist. Also ist mit $\hat{\mu}^*(\tilde{y})$ auch $\hat{\mu}_\rho(\tilde{y})$ unbeschränkt für $\tilde{y} \in \mathcal{Y}_M(y)$.
b) Sei $M < \left\lfloor \frac{\eta}{\eta+1} N \right\rfloor$. Dann gilt $M < \frac{\eta}{\eta+1} N$ und somit $M < \eta(N - M)$. Sei o.B.d.A. $y_1 \geq y_2 \geq \ldots \geq y_N$. Dann gilt für alle $\tilde{y} \in \mathcal{Y}_M(y)$ und $\mu \in \mathbb{R}$

$$\sum_{n=1}^{N} \psi(\tilde{y}_n - \mu) \leq \sum_{n=1}^{N-M} \psi(y_n - \mu) + M\,\psi(\infty).$$

Außerdem gilt

$$\begin{aligned}\lim_{\mu\to\infty} \sum_{n=1}^{N-M} \psi(y_n - \mu) + M\,\psi(\infty) &= (N-M)\psi(-\infty) + M\psi(\infty) \\ < \quad (N-M)\psi(-\infty) + \eta\,(N-M)\,\psi(\infty) &\leq (N-M)\psi(-\infty) - \frac{\psi(-\infty)}{\psi(\infty)}\,(N-M)\,\psi(\infty) = 0.\end{aligned}$$

Also gibt es $\mu_0 < \infty$ mit

$$\sum_{n=1}^{N} \psi(\tilde{y}_n - \mu_0) < 0 \quad \text{für alle} \quad \tilde{y} \in \mathcal{Y}_M(y).$$

Damit gilt aber $\hat{\mu}_\rho(\tilde{y}) \leq \hat{\mu}^{**}(\tilde{y}) \leq \mu_0$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$. Analog zeigt man $\hat{\mu}_\rho(\tilde{y}) \geq \hat{\mu}^*(\tilde{y}) \geq \mu_1$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$. □

Aus Satz 3.4.4 folgt, dass das arithmetische Mittel einen Bruchpunkt von $\frac{1}{N}$ hat, da $\psi(z) = 2z$ unbeschränkt ist. Da $\psi(z) = \text{sign}(z)$ beschränkt ist, hat der Median nach diesem Satz einen Bruchpunkt, der nicht kleiner als $\frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$ ist. In der Tat ist der Bruchpunkt des Medians $\frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor$, was eine Übungsaufgabe ist.

3.4.5 Satz
Für den Bruchpunkt des p-Quantils gilt

$$\epsilon^*(\tilde{y}_p, y) = \frac{1}{N} \min\left\{\lceil p\,N \rceil, \lceil (1-p)\,N \rceil\right\}.$$

Dabei ist $\lceil z \rceil$ die nach oben gerundete Zahl, d.h. $\lceil z \rceil = \min\{n \in \mathbb{N};\ n \geq z\}$.

Beweis. Übung.

3.4.6 Satz
Für den Bruchpunkt einer Lokations-LTS-Schätzfunktion $\hat{\mu}_{k,h}$ mit $h = \left\lceil \frac{N+1}{2} \right\rceil$ gilt für alle $y \in \mathbb{R}^N$ und alle $k = 1, \ldots, h$:

$$\epsilon^*(\hat{\mu}_{k,h}, y) = \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor.$$

Beweis. Wegen Satz 3.4.3 muss nur noch $\epsilon^*(\hat{\mu}_{k,h}, y) \geq \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor$ gezeigt werden, was bedeutet, dass bei $M \leq \left\lfloor \frac{N+1}{2} \right\rfloor - 1$ Ausreißern noch kein Zusammenbruch erfolgt. Sei also $M \leq \left\lfloor \frac{N+1}{2} \right\rfloor - 1$ und

$$L = \max\{y_n;\ n = 1, \ldots, N\} - \min\{y_n;\ n = 1, \ldots, N\}.$$

Wir zeigen, dass für alle $\tilde{y} \in \mathcal{Y}_M(y)$ gilt

$$\hat{\mu}_{k,h}(\tilde{y}) \in A := \left[\min\{y_n;\ n = 1, \ldots, N\} - 2\,L\,\sqrt{h}\ ,\ \max\{y_n;\ n = 1, \ldots, N\} + 2\,L\,\sqrt{h}\right],$$

was bedeutet, dass $|\hat{\mu}_{k,h}(\tilde{y})|$ nicht beliebig groß wird. Sei dazu $\tilde{y} \in \mathcal{Y}_M(y)$ beliebig und $\mu \notin A$. O.B.d.A. sei $r_{(n)}(\tilde{y}, \mu) = r_n(\tilde{y}, \mu) = |\tilde{y}_n - \mu|$. Wegen $h \geq M + 1$ gibt es $l \in \{1, \ldots, h\}$ mit $r_{(l)}(\tilde{y}, \mu) = |y_l - \mu| \geq 2\,L\,\sqrt{h}$, d.h. dieses Residuum betrifft eine ursprüngliche Beobachtung. Für $n = l, \ldots, h$ folgt $r_{(n)}(\tilde{y}, \mu) \geq 2\,L\,\sqrt{h}$ und somit

$$\sum_{n=k}^{h} r_{(n)}(\tilde{y}, \mu)^2 \geq (2\,L\,\sqrt{h})^2 = 4\,L^2\,h.$$

Wegen $M \leq \left\lfloor \frac{N+1}{2} \right\rfloor - 1$ gilt $N - M \geq N - \left\lfloor \frac{N+1}{2} \right\rfloor + 1 \geq \left\lceil \frac{N+1}{2} \right\rceil = h$, so dass es mindestens h Beobachtungen y_n gibt mit $\tilde{y}_n = y_n$. Also gibt es $n_1, \ldots, n_h \in \{1, \ldots, N\}$ mit $\tilde{y}_{n_l} = y_{n_l}$ für $l = 1, \ldots, h$. Für beliebiges $\mu_0 \in [\min\{y_n;\ n = 1, \ldots, N\}, \max\{y_n;\ n = 1, \ldots, N\}]$ gilt dann $|\tilde{y}_{n_l} - \mu_0| = |y_{n_l} - \mu_0| \leq L$ für $l = 1, \ldots, h$. Daraus folgt

$$h\,L^2 \geq \sum_{n=k}^{h} r_{(n)}(\tilde{y}, \mu_0)^2.$$

Damit kann $\mu \notin A$ nicht die LTS-Schätzung bezgl. $\tilde{y}$ sein. □

Kapitel 4

Streuungsschätzungen

4.1 Bekannte und neue Streuungsschätzungen

Streuungsparameter/Streuungsschätzungen werden unterschieden in diejenigen, die

- auf der Differenz zwischen zwei Lageparametern beruhen (etwa Differenz zwischen Maximum und Minimum = Spannweite),
- die Abweichung zwischen den beobachteten Werten und einem Lageparameter (etwa quadratische Abweichung zwischen Beobachtungen und arithmetischem Mittel = Varianz) zur Berechnung nutzen, oder
- auf Differenzen zwischen den einzelnen beobachteten Werten y_n (z.B. Q-Schätzung) basieren.

Zudem unterscheidet man Streuungsparameter, die die gleiche Dimension wie die Daten haben (z.B. Spannweite, Quartilsabstand, Standardabweichung), und solche, die dimensionslos sind (z.B. Quartilskoeffizient und Variationskoeffizient).

4.1.1 Definition (Einige Streuungsschätzungen)
Seien $y_1, \dots, y_N$ Beobachtungswerte eines Merkmals Y. Dann heißt:

1. *die Differenz zwischen Maximum und Minimum* ***Spannweite (Range)****:* $R(y) = y_{(N)} - y_{(1)}$,
2. *die Differenz zwischen oberem und unterem Quartil* ***Quartilsabstand****:* $Q(y) = \widetilde{y}_{0.75} - \widetilde{y}_{0.25}$,
3. *die* ***kürzeste Hälfte (shortest half)****:* $\hat{\sigma}_{SH}(y) = \min\left\{b - a;\ b \geq a,\ \sharp\{n;\ y_n \in [a,b]\} \geq \frac{N}{2}\right\}$,
4. *der* ***mittlere absolute Abweichung (deviation)****:* $d(y) = \frac{1}{N}\sum_{n=1}^{N} |y_n - \widetilde{y}_{0.5}|$,
5. *der* ***Median der absoluten Abweichungen (MAD)****:* $d_{MAD}(y) =$ *Median von* $|y_1 - \widetilde{y}_{0.5}|, \dots, |y_N - \widetilde{y}_{0.5}|$,

6. *die* ***Standardabweichung****:* $s(y) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(y_n - \overline{y})^2} = \sqrt{\frac{1}{N-1}\left(\left(\sum_{n=1}^{N} y_n^2\right) - N\,\overline{y}^2\right)}$.

4.1.2 Definition (Q-Schätzung (Rousseeuw/Croux 1992/1993))
Die Q-Schätzung (Q von Quartil) $\hat{\sigma}_Q(y)$ *basierend auf den Daten* $y_1, \ldots, y_N$ *ist definiert durch*

$$\hat{\sigma}_Q(y) = H_{N,y}^{-1}\left(\frac{1}{4}\right) := \inf\left\{x;\ H_{N,y}(x) \geq \frac{1}{4}\right\},$$

wobei

$$H_{N,y}(x) = \frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{m=n+1}^{N} 1_{(-\infty,x]}(|y_n - y_m|) = \frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m=1,\, m\neq n}^{N} 1_{(-\infty,x]}(|y_n - y_m|)$$

die Verteilungsfunktion der Abstände zwischen den Beobachtungen ist. Die Q-Schätzung ist also ein $\frac{1}{4}$*-Quantil der absoluten Abstände, d.h. von* $\{|y_n - y_m|; n \neq m\}$.

4.1.3 Definition (Skalen-LTS-Schätzung)
Die Skalen-LTS-Schätzung $\hat{\sigma}_{k,h}(y)$ *bezüglich* k *und* h *basierend auf den Daten* $y_1, \ldots, y_N$ *ist definiert durch*

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

Dabei gilt wieder $r_n(y,\mu) = |y_n - \mu|$ *und* $r_{(1)}(y,\mu) \leq r_{(2)}(y,\mu) \leq \ldots \leq r_{(N)}(y,\mu)$ *(vergleiche Definition 3.3.5).*

4.1.4 Satz
Die Schätzfunktionen, die durch Größen in den Definitionen 4.1.1, 4.1.2 und 4.1.3 gegeben sind, sind lokations-invariant und skalen-äquivariant und damit sinnvolle Schätzfunktionen für die Streuung eines univariaten Datensatzes.

4.1.5 Definition (Von Streuungsparametern abgeleitete Größen)
Seien $y_1, \ldots, y_N$ *Beobachtungswerte eines Merkmals* Y*. Dann sei:*

1. ***(Empirische) Varianz*** *(mittlere quadratische Abweichung):*
   $s(y)^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \overline{y})^2 = \frac{1}{N-1}\sum_{n=1}^{N} y_n^2 - \frac{N}{N-1}\,\overline{y}^2$.

2. ***Quartilskoeffizient****:* $Q_{koeff}(y) = \frac{2Q(y)}{\widetilde{y}_{0.25}+\widetilde{y}_{0.75}} = \frac{2(\widetilde{y}_{0.75}-\widetilde{y}_{0.25})}{\widetilde{y}_{0.25}+\widetilde{y}_{0.75}}$*, wobei* Y *nur nicht-negative Ausprägungen besitzt.*

3. ***Variationskoeffizient****:* $V(y) = \frac{s(y)}{\overline{y}}$*, wobei* Y *nur nicht-negative Ausprägungen besitzt.*

Die Ähnlichkeit der Q-Schätzung zur empirischen Varianz sieht man aus folgendem Lemma.

4.1.6 Lemma
Es gilt

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

Beweis. Übung.

4.1.7 Korollar
Die Schätzfunktionen, die durch den Quartilskoeffizient und den Variationskoeffizient gegeben sind, sind skalen-invariant und dimensionslos.

4.1.8 Bemerkung

1. Statt $s(y)$ und $s(y)^2$ wird auch s_y und s_y^2 verwendet. Ist es klar, auf welchen Datensatz sich die Schätzung bezieht, wird der Datensatz oft nicht angegeben. So wird statt $s(y)$ und $s(y)^2$ oft einfach nur s und s^2 benutzt.

4.2 Bruchpunkte von Streuungsschätzungen

Die Ausreißer-Robustheit von Streuungsschätzungen kann wieder mit dem Bruchpunkt gemessen werden. Nach Definition 1.2.2 muss dafür wieder geklärt werden, was Ausreißer sind und was eine beliebige Verfälschung ist. Da der Wertebereich einer Streuungsschätzung $[0, \infty)$ ist, liegt eine beliebige Verfälschung nicht nur dann vor, wenn die Schätzung $\hat{\sigma}$ gegen ∞ geht, sondern auch, wenn sie gegen 0 konvergiert, wenn sie beim Ausgangsdatensatz $\hat{\sigma}(y) > 0$ erfüllt. Eine Konvergenz gegen 0 wird in der Regel dadurch erreicht, dass die abgeänderten Beobachtungen gleiche Werte annehmen. So etwas wollen wir auch als „Ausreißer“ bezeichnen. Die Konvergenz gegen 0 wird als Implosion bezeichnet.

4.2.1 Definition (Implosionspunkt)
Der Implosionspunkt $\epsilon^-(\hat{\theta}, y)$ einer Schätzfunktion $\hat{\theta}$ beim Datensatz y ist definiert als

$$\epsilon^-(\hat{\theta}, y) = \frac{1}{N} \min \left\{ M;\ \inf_{\tilde{y} \in \mathcal{Y}_M(y)} |\hat{\theta}(\tilde{y})| = 0 \right\},$$

wobei wieder $\mathcal{Y}_M(y) = \left\{ (\tilde{y}_1, \ldots, \tilde{y}_N)^\top \in \mathbb{R}^N;\ \sharp\{n|\ y_n \neq \tilde{y}_n\} \leq M \right\}$ *und* $\inf_{\tilde{y} \in \mathcal{Y}_M(y)} |\hat{\theta}(\tilde{y})| := \inf_{\tilde{y} \in \mathcal{Y}_M(y)} \inf_{\theta \in \hat{\theta}(\tilde{y})} |\theta|$ *für eine mengenwertige Schätzfunktion $\hat{\theta}$ gilt.*

4.2.2 Definition (Bruchpunkt einer Streuungsschätzung)
Der Bruchpunkt $\epsilon^*(\hat{\sigma}, y)$ einer Streuungsschätzfunktion $\hat{\sigma}$ beim Datensatz y ist das Minimum von Explosions- und Implosionspunkt, d.h.

$$\epsilon^*(\hat{\sigma}, y) = \min \left\{ \epsilon^+(\hat{\sigma}, y), \epsilon^-(\hat{\sigma}, y) \right\}.$$

4.2.3 Satz
Ist die Streuungsschätzfunktion $\hat{\sigma}$ skalen-äquivariant, so gilt:

$$\epsilon^*(\hat{\sigma}, y) \leq \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor.$$

Beweis. Sei $M = \left\lfloor \frac{N+1}{2} \right\rfloor$. Angenommen, es gibt $B > 0$ und $L < \infty$ mit

$$B < \hat{\sigma}(\tilde{y}) < L \tag{4.1}$$

für alle $\tilde{y} \in \mathcal{Y}_M(y)$. Betrachte:

$$\begin{array}{llll} \tilde{y}_n^{k,1} = k\, y_n & \text{für } n = 1, \dots M, & \tilde{y}_n^{k,1} = y_n & \text{für } n = M+1, \dots, N, \\ \tilde{y}_n^{k,2} = y_n & \text{für } n = 1, \dots M, & \tilde{y}_n^{k,2} = \frac{1}{k}\, y_n & \text{für } n = M+1, \dots, N. \end{array}$$

Dann gilt $\tilde{y}^{k,1} \in \mathcal{Y}_M(y)$ und wegen $N - M = N - \left\lfloor \frac{N+1}{2} \right\rfloor \leq \left\lfloor \frac{N+1}{2} \right\rfloor$ auch $\tilde{y}^{k,2} \in \mathcal{Y}_M(y)$. Wegen $\tilde{y}_n^{k,1} = k\, \tilde{y}_n^{k,2}$ für $n = 1, \dots, N$ und der Skalen-Äquivarianz gilt $\hat{\sigma}(\tilde{y}^{k,1}) = k\, \hat{\sigma}(\tilde{y}^{k,2})$. Aus (4.1) folgt

$$\begin{aligned} 2L &\geq |\hat{\sigma}(\tilde{y}^{k,1})| + |\hat{\sigma}(\tilde{y}^{k,2})| \geq |\hat{\sigma}(\tilde{y}^{k,1}) - \hat{\sigma}(\tilde{y}^{k,2})| \\ &= |k\, \hat{\sigma}(\tilde{y}^{k,2}) - \hat{\sigma}(\tilde{y}^{k,2})| = (k-1)\hat{\sigma}(\tilde{y}^{k,2}) \geq (k-1)B \end{aligned}$$

für alle $k > 1$. Für $k - 1 > \frac{2L}{B}$ liegt aber ein Widerspruch vor, so dass die Annahme (4.1) nicht gilt. □

4.2.4 Satz
Ist die Streuungsschätzfunktion $\hat{\sigma}$ skalen-äquivariant und lokations-invariant, so gilt:

$$\epsilon^*(\hat{\sigma}, y) \leq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor.$$

4.2.5 Satz
Für die Bruchpunkte der Streuungs-Schätzfunktionen $\hat{\sigma}$, die auf der Spannweite, der mittleren absoluten Abweichung und der Standardabweichung basieren, gilt für alle $y \in \mathbb{R}^N$:

$$\epsilon^*(\hat{\sigma}, y) = \frac{1}{N}.$$

Beweis. Der Beweis ist eine Übungsaufgabe. Dabei reicht es zu zeigen, dass für den Explosionspunkt $\epsilon^+(\hat{\sigma}, y) = \frac{1}{N}$ gilt.

4.2.6 Satz
Für den Bruchpunkt der Streuungs-Schätzfunktion $\hat{\sigma}_{SH}$ basierend auf der kürzesten Hälfte gilt für alle $y \in \mathbb{R}^N$, deren Komponenten paarweise verschieden sind:

$$\epsilon^*(\hat{\sigma}_{SH}, y) = \frac{1}{N}\left\lfloor \frac{N-1}{2} \right\rfloor .$$

Beweis.
1. Behauptung: Für den Explosionspunkt gilt $\epsilon^+(\hat{\sigma}_{\mathrm{SH}}, y) \geq \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor$.
Beweis: Sei $M \leq \left\lfloor \frac{N+1}{2} \right\rfloor - 1 = \left\lfloor \frac{N-1}{2} \right\rfloor$, $A = [\min\{y_n;\ n = 1, \ldots, N\}\ ,\ \max\{y_n;\ n = 1, \ldots, N\}\ ]$ und $B = \max\{y_n;\ n = 1, \ldots, N\} - \min\{y_n;\ n = 1, \ldots, N\}$. Für jedes $\tilde{y} \in \mathcal{Y}_M(y)$ sind dann $N - M \geq \left\lfloor \frac{N+1}{2} \right\rfloor \geq \frac{N}{2}$ Beobachtungen nicht verändert und liegen somit in A. Daraus folgt $\hat{\sigma}_{\mathrm{SH}}(\tilde{y}) \leq B$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$.

2. Behauptung: Für den Implosionspunkt gilt $\epsilon^-(\hat{\sigma}_{\mathrm{SH}}, y) \leq \frac{1}{N}\left\lfloor \frac{N-1}{2} \right\rfloor$.
Beweis: Sei $M = \left\lfloor \frac{N-1}{2} \right\rfloor$. Setze $\tilde{y}_n = y_{M+1}$ für $n = 1, \ldots, M$ und $\tilde{y}_n = y_n$ für $n = M+1, \ldots, N$. Dann gilt $\tilde{y} \in \mathcal{Y}_M(y)$ und $M + 1 = \left\lfloor \frac{N+1}{2} \right\rfloor \geq \frac{N}{2}$ Komponenten von $\tilde{y}$ nehmen den Wert y_{M+1} an. Damit gilt $\hat{\sigma}_{\mathrm{SH}}(\tilde{y}) = 0$.

3. Behauptung: Für den Implosionspunkt gilt $\epsilon^-(\hat{\sigma}_{\mathrm{SH}}, y) \geq \frac{1}{N}\left\lfloor \frac{N-1}{2} \right\rfloor$.
Beweis: Sei $B = \min\{|y_n - y_m|;\ n, m = 1, \ldots, N,\ n \neq m\}$. Da die Komponenten von y paarweise verschieden sind, folgt $B > 0$. Ist $M \leq \left\lfloor \frac{N-1}{2} \right\rfloor - 1 = \left\lfloor \frac{N+1}{2} \right\rfloor - 2$, dann nehmen von je $\left\lfloor \frac{N+1}{2} \right\rfloor$ Komponenten von $\tilde{y}$ mindestens zwei Komponenten verschiedene Werte aus $\{y_1, \ldots, y_N\}$ an, z.B. y_{n_k} und y_{n_l} mit $y_{n_k} \neq y_{n_l}$. Daraus folgt $\hat{\sigma}_{\mathrm{SH}}(\tilde{y}) \geq |y_{n_k} - y_{n_l}| \geq B$. Also gilt $\hat{\sigma}_{\mathrm{SH}}(\tilde{y}) \geq B > 0$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$.□

Damit hat eine Schätzfunktion basierend auf der kürzesten Hälfte nicht den größt möglichen Bruchpunkt von Satz 4.2.4, obwohl sie skalen-äquivariant und lokations-invariant ist. Allerdings liefert eine leichte Modifikation der Definition der kürzesten Hälfte den größt möglichen Bruchpunkt.

4.2.7 Satz
Für den Bruchpunkt der Streuungs-Schätzfunktion $\hat{\sigma}_{MSH}$ basierend auf einer Modifikation der Schätzfunktion basierend auf der kürzesten Hälfte gegeben durch

$$\hat{\sigma}_{MSH}(y) = \min\left\{ b - a;\ b \geq a,\ \sharp\{n;\ y_n \in [a, b]\} \geq \frac{N+1}{2} \right\},$$

gilt für alle $y \in \mathbb{R}^N$, deren Komponenten paarweise verschieden sind:

$$\epsilon^*(\hat{\sigma}_{MSH}, y) = \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor .$$

Beweis.
Nach Satz 4.2.4 muss nur $\epsilon^*(\hat{\sigma}_{\mathrm{MSH}}, y) \geq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$ gezeigt werden.

1. Behauptung: Für den Explosionspunkt gilt $\epsilon^+(\hat{\sigma}_{\mathrm{MSH}}, y) \geq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$.
Beweis: Sei $M \leq \left\lfloor \frac{N}{2} \right\rfloor - 1 = \left\lfloor \frac{N-2}{2} \right\rfloor$, $A = [\min\{y_n;\ n = 1, \ldots, N\}\ ,\ \max\{y_n;\ n = 1, \ldots, N\}\ ]$

und $B = \max\{y_n;\ n = 1,\dots,N\} - \min\{y_n;\ n = 1,\dots,N\}$. Für jedes $\tilde{y} \in \mathcal{Y}_M(y)$ sind dann $N - M \geq N - \left\lfloor \frac{N}{2} \right\rfloor + 1 = \left\lfloor \frac{N+1}{2} \right\rfloor + 1 \geq \frac{N+1}{2}$ Beobachtungen nicht verändert und liegen somit in A. Daraus folgt $\hat{\sigma}_{\mathrm{MSH}}(\tilde{y}) \leq B$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$.

2. Behauptung: Für den Implosionspunkt gilt $\epsilon^-(\hat{\sigma}_{\mathrm{SH}}, y) \geq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$.
Beweis: Sei $B = \min\{|y_n - y_m|;\ n, m = 1,\dots,N,\ n \neq m\}$. Da die Komponenten von y paarweise verschieden sind, folgt $B > 0$. Ist $M \leq \left\lfloor \frac{N}{2} \right\rfloor - 1 = \left\lfloor \frac{N+2}{2} \right\rfloor - 2$, dann nehmen von je $\left\lfloor \frac{N+2}{2} \right\rfloor$ Komponenten von $\tilde{y}$ mindestens zwei Komponenten verschiedene Werte aus $\{y_1,\dots,y_N\}$ an, z.B. y_{n_k} und y_{n_l} mit $y_{n_k} \neq y_{n_l}$. Daraus folgt $\hat{\sigma}_{\mathrm{MSH}}(\tilde{y}) \geq |y_{n_k} - y_{n_l}| \geq B$. Also gilt $\hat{\sigma}_{\mathrm{MSH}}(\tilde{y}) \geq B > 0$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$.□

4.2.8 Satz
Für den Bruchpunkt der Streuungs-Schätzfunktion d_{MAD} basierend auf dem Median der absoluten Abweichungen gilt für alle $y \in \mathbb{R}^N$, deren Komponenten paarweise verschieden sind:

$$\epsilon^*(d_{MAD}, y) = \frac{1}{N}\left\lfloor \frac{N-1}{2} \right\rfloor.$$

Beweis. Der Beweis ist eine Übungsaufgabe.

4.2.9 Satz
Für den Bruchpunkt der Q-Schätzfunktion $\hat{\sigma}_Q$ gilt für alle $y \in \mathbb{R}^N$, deren Komponenten paarweise verschieden sind:

$$\epsilon^*(\hat{\sigma}_Q, y) = \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor.$$

Beweis.
Nach Satz 4.2.4 muss nur $\epsilon^*(\hat{\sigma}_Q, y) \geq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$ gezeigt werden.
Im ganzen gibt es $\binom{N}{2} = \frac{N(N-1)}{2}$ verschiedene Differenzen $|y_n - y_m|$ von einem Datensatz $(y_1,\dots,y_N)^\top$.

1. Behauptung: Für den Explosionspunkt gilt $\epsilon^+(\hat{\sigma}_Q, y) \geq \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor$.
Beweis: Sei $M \leq \left\lfloor \frac{N+1}{2} \right\rfloor - 1 = \left\lfloor \frac{N-1}{2} \right\rfloor$ und $\tilde{y} \in \mathcal{Y}_M(y)$ beliebig. Dann gibt es höchstens $\left\lfloor \frac{N-1}{2} \right\rfloor$ abgeänderte Komponenten, so dass $\left\lceil \frac{N+1}{2} \right\rceil$ Komponenten unverändert sind. Damit gibt es mindestens

$$\binom{\left\lceil \frac{N+1}{2} \right\rceil}{2} = \frac{\left\lceil \frac{N+1}{2} \right\rceil \left(\left\lceil \frac{N+1}{2} \right\rceil - 1 \right)}{2} \geq \frac{\frac{N+1}{2}\left(\frac{N+1}{2} - 1 \right)}{2} = \frac{(N+1)(N-1)}{8}$$

unverfälschte Differenzen $|y_n - y_m|$. Für diese Differenzen gilt $|y_n - y_m| \leq \max\{|y_n - y_m|;\ n, m = 1,\dots,N\} =: B$. Da $\frac{(N+1)(N-1)}{8} > \frac{1}{4}\frac{N(N-1)}{2}$ gilt, sind mehr als ein Viertel aller möglichen Differenzen, die von $\tilde{y}$ gebildet werden können, unverfälscht. Es gilt somit

$$H_{N,\tilde{y}}(B) \geq \frac{2}{N(N-1)}\frac{(N+1)(N-1)}{8} > \frac{1}{4}$$

und damit $\hat{\sigma}_Q(\tilde{y}) = H^{-1}_{N,\tilde{y}}\left(\frac{1}{4}\right) \leq B$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$.

2. Behauptung: Für den Implosionspunkt gilt $\epsilon^-(\hat{\sigma}_Q, y) \geq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$.
Beweis: Sei $B = \frac{1}{3}\min\{|y_n - y_m|;\ n, m \in 1, \ldots, N,\ n \neq m\}$. Da die Komponenten von y paarweise verschieden sind, folgt $B > 0$. Ist $M \leq \left\lfloor \frac{N}{2} \right\rfloor - 1 = \left\lfloor \frac{N-2}{2} \right\rfloor$, dann können von $\tilde{y} \in \mathcal{Y}_M(y)$ höchstens $\left\lfloor \frac{N}{2} \right\rfloor$ Komponenten Werte besitzen, die sich alle nicht mehr als B unterscheiden. Somit gibt es höchstens

$$\binom{\left\lfloor \frac{N}{2} \right\rfloor}{2} = \frac{\left\lfloor \frac{N}{2} \right\rfloor \left(\left\lfloor \frac{N}{2} \right\rfloor - 1\right)}{2} \leq \frac{\frac{N}{2}\left(\frac{N}{2} - 1\right)}{2} = \frac{N\,(N-2)}{8}$$

Differenzen, die kleiner als B sind. Für alle anderen Differenzen $|\tilde{y}_n - \tilde{y}_m|$ gilt $|\tilde{y}_n - \tilde{y}_m| \geq B$. Damit gilt $H_{N,\tilde{y}}(x) \leq \frac{2}{N(N-1)}\frac{N\,(N-2)}{8} < \frac{1}{4}$ für alle $x \in [0, B)$ und somit $\hat{\sigma}_Q(\tilde{y}) = H^{-1}_{N,\tilde{y}}\left(\frac{1}{4}\right) \geq B$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$.

□

4.2.10 Bemerkung
Man könnte sich fragen, ob der Median der absoluten Differenzen auch einen hohen Bruchpunkt besitzt. Dazu sei

$$\widehat{\sigma}_{Med}(y) := \text{ Median von } \{|y_n - y_m|;\ n \neq m\}.$$

Zur Bestimmung des Explosionsbruchpunktes ändere M Beobachtungen zu $y_{(N)} + l, \ldots, y_{(N)} + Ml$ ab. Dann sind $\binom{M}{2} + M(N-M)$ differenzen größer gleich l und $\binom{N-M}{2}$ Differenzen sind unverändert kleiner gleich $\max\{|y_n - y_m|;\ n \neq m\}$. Wir suchen jetzt ein M mit $\frac{1}{2} > \binom{N-M}{2}/\binom{N}{2}$:

$$\frac{\binom{N-M}{2}}{\binom{N}{2}} = \frac{(N-M)(N-M-1)}{N(N-1)} = \left(1 - \frac{M}{N}\right)\left(1 - \frac{M}{N-1}\right) < \left(1 - \frac{M}{N}\right)^2 < \frac{1}{2}$$
$$\iff \ (N-M)^2 < N^2 \cdot \frac{1}{2} \iff N - M < \frac{N}{\sqrt{2}} \iff M > N - \frac{N}{\sqrt{2}} = N\left(1 - \frac{1}{\sqrt{2}}\right) = N \cdot 0.2929.$$

Damit folgt

$$\epsilon^*(\widehat{\sigma}_{Med}, y) < \frac{1}{N}\left\lceil N\left(1 - \frac{1}{\sqrt{2}}\right)\right\rceil < \frac{1}{N}\lceil N\,0.3\rceil\,.$$

4.2.11 Satz
Für den Bruchpunkt der Skalen-LTS-Schätzfunktion $\hat{\sigma}_{k,h}$ mit $h = \left\lceil \frac{N+1}{2} \right\rceil$ gilt für alle $k = 1, \ldots, h$ und alle $y \in \mathbb{R}^N$, deren Komponenten paarweise verschieden sind:

$$\epsilon^*(\hat{\sigma}_{k,h}, y) = \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor.$$

Beweis.
Nach Satz 4.2.4 muss nur $\epsilon^*(\hat{\sigma}_{k,h}, y) \geq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$ gezeigt werden.

1. Behauptung: Für den Explosionspunkt gilt $\epsilon^+(\hat{\sigma}_{k,h}, y) \geq \frac{1}{N}\left\lfloor \frac{N+1}{2} \right\rfloor$.

Beweis: Die Behauptung folgt ähnlich wie der Explosionspunkt für die Lokations-LTS-Schätzfunktion.

2. Behauptung: Für den Implosionspunkt gilt $\epsilon^-(\hat{\sigma}_{k,h}, y) \geq \frac{1}{N}\left\lfloor \frac{N}{2} \right\rfloor$.
Beweis: Sei $B = \frac{1}{2}\min\{|y_n - y_m|;\ n, m \in 1, \ldots, N,\ n \neq m\}$. Da die Komponenten von y paarweise verschieden sind, folgt $B > 0$. Außerdem sei $M \leq \left\lfloor \frac{N}{2} \right\rfloor - 1 = \left\lfloor \frac{N-2}{2} \right\rfloor$, und $\tilde{y} \in \mathcal{Y}_M(y)$ beliebig. Wegen $M \leq h-2$ gibt es von h Komponenten von $\tilde{y}$ immer mindestens zwei Komponenten mit $\tilde{y}_n = y_n$. Damit gilt $r_{(h)}(\tilde{y}, \mu) \geq B$ für alle $\mu \in \mathbb{R}$ und somit $\hat{\sigma}_{k,h}(\tilde{y}) = \min_{\mu\in\mathbb{R}} \sqrt{\frac{1}{h-k+1}\sum_{n=k}^{h} r_{(n)}(\tilde{y}, \mu)^2} \geq \sqrt{\frac{1}{h-k+1}}\, B$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$. □

4.2.12 Satz
Der Bruchpunkt des Quartilsabstandens Q ist

$$\epsilon^*(Q, y) = \frac{1}{N}\left\lceil \frac{N}{4} \right\rceil,$$

wenn alle Komponenten von y paarweise verschieden sind.

Beweis. Übung.

Teil II

Multivariate Daten und Regression

Kapitel 5

Multivariate Daten

Sind die Daten mehrdimensional (multivariate), so sind die einzelnen Beobachtungen durch

$$y_1 = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1p} \end{pmatrix}, \; y_2 = \begin{pmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2p} \end{pmatrix}, \; \ldots, \; y_N = \begin{pmatrix} y_{N1} \\ y_{N2} \\ \vdots \\ y_{Np} \end{pmatrix} \in \mathbb{R}^p$$

gegeben, wobei $p \geq 2$ ist. Die Datenmatrix y ist dann gegeben durch

$$y = \begin{pmatrix} y_1^\top \\ y_2^\top \\ \vdots \\ y_N^\top \end{pmatrix} = \left(y_{(1)}, y_{(2)}, \ldots, y_{(p)}\right) \in \mathbb{R}^{N \times p},$$

wobei $y_{(1)}, y_{(2)}, \ldots, y_{(p)}$ die Spalten dieser Matrix bezeichnen. Diese Spalten sind die Beobachtungen zur j'ten Variable, d.h. $y_{(j)} = (y_{1j}, y_{2j}, \ldots, y_{Nj})^\top \in \mathbb{R}^N$.

5.1 Lageschätzungen

5.1.1 Definition (Multivariater Mittelwert (komponentenweiser Mittelwert))
Der multivariate Mittelwert von $y \in \mathbb{R}^{N \times p}$ ist der Vektor bestehend aus den komponentenweisen Mittelwerten für jede Variable, d.h.

$$\overline{y} = \begin{pmatrix} \overline{y}_{(1)} \\ \overline{y}_{(2)} \\ \vdots \\ \overline{y}_{(p)} \end{pmatrix} = \frac{1}{N} \, y^\top 1_N.$$

5.1.2 Definition (Komponentenweiser Median)
Der komponentenweise Median $med_k(y)$ von $y \in \mathbb{R}^{N\times p}$ ist der Vektor bestehend aus den komponentenweisen Medianen für jede Variable, d.h.

$$med_k(y) = \begin{pmatrix} med(y_{(1)}) \\ med(y_{(2)}) \\ \vdots \\ med(y_{(p)}) \end{pmatrix}.$$

5.1.3 Satz
a) Der multivariate Mittelwert ist affin äquivariant, d.h. ist $z = y\,A^\top + 1_N\,b^\top = (A\,y_1 + b, \ldots, A\,y_N + b)^\top$ mit $A \in \mathbb{R}^{q\times p}$ und $b \in \mathbb{R}^q$, so gilt

$$\overline{z} = A\overline{y} + b.$$

b) Der komponentenweise Median ist selbst für orthogonale Matrizen nicht äquivaraint, d.h. es gilt nicht immer

$$med_k(y\,A^\top) = med_k((A\,y_1, \ldots, A\,y_N)^\top) = A\ med_k(y)$$

mit orthogonaler Matrix $A \in \mathbb{R}^{p\times p}$ für $y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^{N\times p}$.

5.1.4 Bemerkung
Die Eigenschaft in Satz 5.1.3 a) ist für $p = 1$ die Lokations- und Skalen-Äquivarianz, die für Lageparameter gefordert wurde. Damit erfüllt der multivariate (komponentenweise) Mittelwert die Anforderungen an einen Lageparameter. Da der komponentenweise Median diese Eigenschaft nicht erfüllt, ist er keine gute Schätzung für die Lage.

Nach Satz 3.2.3 gilt für den Median für univariate Daten

$$\text{med}(y) \in \arg\min_{\mu\in\mathbb{R}} \sum_{n=1}^{N} |y_n - \mu|.$$

Dies kann für den multivariaten Fall wie folgt verallgemeinert werden. Dabei sei für $\mu = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ und $y_n = (y_{n1}, \ldots, y_{np})^\top \in \mathbb{R}^p$

$$\|y_n - \mu\| := \sqrt{\sum_{i=1}^{p} (y_{ni} - \mu_i)^2} = \sqrt{(y_n - \mu)^\top\,(y_n - \mu)}$$

der euklidische Abstand.

5.1.5 Definition (l_1-Median (oder räumlicher Median))
Ein Parametervektor $med_1(y)$ heißt l_1-Median, falls gilt

$$med_1(y) \in \arg\min_{\mu\in\mathbb{R}^p} \sum_{n=1}^{N} \|y_n - \mu\|,$$

d.h. die Funktion g definiert durch $g(\mu) = \sum_{n=1}^{N} \|y_n - \mu\|$ nimmt bei $\mu = med_1(y)$ ihr Minimum an.

Die Bezeichnung l_1-Median kommt daher, dass $\sum_{n=1}^{N} \|y_n - \mu\|$ auch als l_1-Norm bezeichnet wird, während $\sqrt{\sum_{n=1}^{N} \|y_n - \mu\|^2}$ als l_2-Norm bezeichnet wird.

In der Regel ist der l_1-Median eindeutig (siehe z.B. Milasevic und Ducharme 1987) und kann nur per Computer ausgerechnet werden, z.B. mittels `spatial.median` des `R`-Paketes `ICSNP` (siehe z.B. Vardi und Zhang 1999). Nur in wenigen Fällen kann er per Hand bestimmt werden. Das folgende Lemma liefert ein Beispiel, wo der l_1-Median direkt angegeben werden kann.

5.1.6 Lemma
Der einzige l_1-Median von $z = \left(\binom{-1}{0}, \binom{1}{0}, \binom{0}{b}\right)^\top$ ist $\begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix}$ für alle $b \in [\sqrt{3}, \infty)$, d.h. es gilt für alle $b \in [\sqrt{3}, \infty)$:
$med_1(z) = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix}$ und $med_1(z) \neq \mu$ für alle $\mu \neq \begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix}$.

Beweis. Sei

$$G_z(\mu) := \sum_{n=1}^{N} \|z_n - \mu\|.$$

Sei $b \in (0, \infty)$ beliebig. Um zu zeigen, dass $\text{med}_1(z) = \left\{ \begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix} \right\}$ gilt, müssen wir zeigen, dass

$$G_z(\mu) > G_z\left(\begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix} \right) \tag{5.1}$$

für alle $\mu \neq \begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix}$ gilt. Dazu zeigen wir zuerst, dass $G_z\left(\binom{a}{m}\right) > G_z\left(\binom{0}{m}\right)$ für alle $m \in \mathbb{R}$ und $0 \neq a \in \mathbb{R}$ gilt. Danach müssen wir (5.1) nur für μ's der Form $\mu = \binom{0}{m}$ beweisen.

Sei also $m \in \mathbb{R}$ und $0 \neq a \in \mathbb{R}$ beliebig. Dann gilt

$$\begin{aligned}
G_z\left(\begin{pmatrix} a \\ m \end{pmatrix}\right) &= \left\|\begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} a \\ m \end{pmatrix}\right\| + \left\|\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} a \\ m \end{pmatrix}-\right\| + \left\|\begin{pmatrix} 0 \\ b \end{pmatrix} - \begin{pmatrix} a \\ m \end{pmatrix}-\right\| \\
&= \left\|\begin{pmatrix} -1-a \\ -m \end{pmatrix}\right\| + \left\|\begin{pmatrix} 1-a \\ -m \end{pmatrix}-\right\| + \left\|\begin{pmatrix} -a \\ b-m \end{pmatrix}-\right\| \\
&= \sqrt{(1+a)^2+m^2} + \sqrt{(1-a)^2+m^2} + \sqrt{a^2+(b-m)^2} \\
&> \sqrt{(1+a)^2+m^2} + \sqrt{(1-a)^2+m^2} + \sqrt{(b-m)^2} \\
&= \sqrt{a^2+2a+1+m^2} + \sqrt{a^2-2a+1+m^2} + \sqrt{(b-m)^2}
\end{aligned}$$

und analog

$$G_z\left(\begin{pmatrix} 0 \\ m \end{pmatrix}\right) = \sqrt{1+m^2} + \sqrt{1+m^2} + \sqrt{(b-m)^2} = 2\sqrt{1+m^2} + \sqrt{(b-m)^2}$$

Wenn wir auch noch zeigen können, dass $\sqrt{a^2+2a+1+m^2} + \sqrt{a^2-2a+1+m^2} \geq 2\sqrt{1+m^2}$ gilt, sind wir für diesen Teil fertig. Setzen wir $c = 1+m^2$, so müssen wir also $\sqrt{a^2+2a+c} + \sqrt{a^2-2a+c} \geq 2\sqrt{c}$ zeigen. Es gilt nun mit mehrmaliger Anwendung der Binomischen Formel

$$\begin{aligned}
&\sqrt{(a^2+2a+c)} + \sqrt{(a^2-2a+c)} \geq 2\sqrt{c} \\
&\iff \\
&\left(\sqrt{(a^2+2a+c)} + \sqrt{(a^2-2a+c)}\right)^2 \geq 4\,c \\
&\iff \\
&a^2+2a+c+2\sqrt{(a^2+2a+c)}\sqrt{(a^2-2a+c)} + a^2-2a+c \geq 4\,c \\
&\iff \\
&2\,a^2+2\,c+2\sqrt{(a^2+2a+c)}\sqrt{(a^2-2a+c)} \geq 4\,c \\
&\iff \\
&2\,a^2+2\sqrt{(a^2+2a+c)}\sqrt{(a^2-2a+c)} \geq 2\,c \\
&\iff \\
&a^2+\sqrt{(a^2+2a+c)}\sqrt{(a^2-2a+c)} \geq c.
\end{aligned}$$

Ist $a^2 > c$, so ist klar, dass die letzte Aussage richtig ist, da die Wurzel-Terme nicht negativ sind.

Ist $a \leq c$, so ist $c - a^2 \geq 0$ und es gilt

$$
\begin{aligned}
& a^2 + \sqrt{(a^2+2a+c)}\sqrt{(a^2-2a+c)} \geq c \\
\iff & \\
& \sqrt{(a^2+2a+c)}\sqrt{(a^2-2a+c)} \geq c - a^2 \\
\iff & \\
& (a^2+2a+c)(a^2-2a+c) \geq (c-a^2)^2 \\
\iff & \\
& a^4 - 2a^3 + ca^2 + 2a^3 - 4a^2 + 2ac + ca^2 - 2ac + c^2 \geq c^2 - 2ca^2 + a^4 \\
\iff & \\
& 4ca^2 - 4a^2 \geq 0 \\
\iff & \\
& 4a^2(c-1) \geq 0.
\end{aligned}
$$

Die letzte Ungleichung ist richtig, da $a \neq 0$ und $c - 1 = 1 + m^2 - 1 = m^2 \geq 0$ gilt. Somit gilt $G_z\left(\binom{a}{m}\right) > G_z\left(\binom{0}{m}\right)$ für alle $m \in \mathbb{R}$ und $0 \neq a \in \mathbb{R}$. Es bleibt zu zeigen, dass (5.1) für alle μ's der Form $\mu = \binom{0}{m}$ gilt. Dabei ist klar, dass wir nur $m \leq b$ betrachten müssen. Wir hatten schon gezeigt, dass

$$G_z(\mu) = 2\sqrt{(1+m^2)} + \sqrt{(b-m)^2} = 2\sqrt{(1+m^2)} + b - m$$

gilt. Sei nun $m = \frac{1}{\sqrt{3}} + \epsilon$ mit $\epsilon \leq b - \frac{1}{\sqrt{3}}$ beliebig. Dann erhalten wir mit der binomischen Formel

$$
\begin{aligned}
& G_z(\mu) \\
= \; & 2\sqrt{\left(1 + \left(\frac{1}{\sqrt{3}} + \epsilon\right)^2\right)} + b - \left(\frac{1}{\sqrt{3}} + \epsilon\right) = 2\sqrt{\left(1 + \frac{1}{3} + 2\frac{1}{\sqrt{3}}\,\epsilon + \epsilon^2\right)} + b - \frac{1}{\sqrt{3}} - \epsilon \\
= \; & 2\sqrt{\left(\frac{4}{3} + 2\frac{1}{\sqrt{3}}\,\epsilon + \epsilon^2\right)} + b - \frac{1}{\sqrt{3}} - \epsilon \;\geq\; 2\sqrt{\left(\frac{4}{3} + 2\frac{1}{\sqrt{3}}\,\epsilon + \frac{1}{4}\,\epsilon^2\right)} + b - \frac{1}{\sqrt{3}} - \epsilon \\
= \; & 2\sqrt{\left(\frac{4}{3} + 2\sqrt{\frac{4}{3}}\sqrt{\frac{1}{4}}\,\epsilon + \frac{1}{4}\,\epsilon^2\right)} + b - \frac{1}{\sqrt{3}} - \epsilon = 2\sqrt{\left(\sqrt{\frac{4}{3}} + \sqrt{\frac{1}{4}}\,\epsilon\right)^2} + b - \frac{1}{\sqrt{3}} - \epsilon \\
= \; & 2\left(\frac{2}{\sqrt{3}} + \frac{1}{2}\,\epsilon\right) + b - \frac{1}{\sqrt{3}} - \epsilon = \frac{4}{\sqrt{3}} + b - \frac{1}{\sqrt{3}} \\
= \; & 2\sqrt{\left(1 + \frac{1}{3}\right)} + b - \frac{1}{\sqrt{3}} = G_z\left(\begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix}\right).
\end{aligned}
$$

Dabei gilt nur Gleichheit, wenn $\epsilon = 0$ gilt. $\square$

5.1.7 Satz
a) Für den multivariaten Mittelwert $\overline{y}$ (komponentenweisen Mittelwert) gilt

$$\overline{y} \in \arg\min_{\mu\in\mathbb{R}^p} \sum_{n=1}^{N} \|y_n - \mu\|^2.$$

b) Für den komponentenweisen Median $med_k(y)$ gilt im allgemeinen ***nicht***

$$med_k(y) \in \arg\min_{\mu\in\mathbb{R}^p} \sum_{n=1}^{N} \|y_n - \mu\|,$$

was bedeutet, dass der komponentenweiser Median und der l_1-Median verschieden sein können.

5.1.8 Satz
a) Der l_1-Median ist äquivariant bzgl. beliebiger orthogonaler Transformationen, d.h. es gilt $med_1((A\,y_1,\ldots,A\,y_N)^\top) = A\,med_1((y_1,\ldots,y_N)^\top)$ für alle $(y_1,\ldots,y_N)^\top \in \mathbb{R}^{N\times p}$ und alle orthogonalen Matrizen $A \in \mathbb{R}^{p\times p}$

b) Der l_1-Median ist nicht äquivariant bzgl. beliebiger regulärer Transformationen, d.h. es gilt **nicht** *$med_1((A\,y_1,\ldots,A\,y_N)^\top) = A\,med_1((y_1,\ldots,y_N)^\top)$ für alle $(y_1,\ldots,y_N)^\top \in \mathbb{R}^{N\times p}$ und alle regulären Matrizen $A \in \mathbb{R}^{p\times p}$.*

Um eine Verallgemeinerung des univariaten Medians zubekommen, der auch äquivariant bezüglich regulärer Transformationen ist, kann man die Charakterisierung des Medians über die Lokations-Tiefe benutzen. Diese Lokations-Tiefe wird im multivariaten Fall zur Halbraum-Tiefe verallgemeinert (man überlege sich, warum die folgende Definition wirklich eine Verallgemeinerung der Lokations-Tiefe ist).

5.1.9 Definition (Halbraum-Tiefe)
Die Halbraum-Tiefe $d_H(\mu, y)$ eines Lageparameters $\mu \in \mathbb{R}^p$ bezüglich des Datensatzes $y = (y_1,\ldots,y_N)^\top$ ist definiert als

$$d_H(\mu, y) = \frac{1}{N} \min_{u\in\mathbb{R}^p\setminus\{0\}} \sharp\{n;\ u^\top y_n \geq u^\top \mu\}.$$

5.1.10 Beispiel
Abbildung 5.1 zeigt für 5 Datenpunkte im $\mathbb{R}^2$, wie die Halbraum-Tiefe eines Parameter $\mu \in \mathbb{R}^2$ gewonnen wird: Man betrachtet alle Halbräume, die durch eine Gerade begrenzt sind, die durch μ geht. Das kann dadurch geschehen, dass diese Gerade in μ rotiert werden. Der Halbraum, der die wenigsten Datenpunkte enthält, ergibt dann die Halbraum-Tiefe. Diese Situation ist rechts in der Abbildung 5.1 gegeben, wo der Halbraum nur einen Datenpunkt enthält. Die Halbraum-Tiefe des Parameters μ ist somit

$$d_H(\mu, (y_1,\ldots,y_5)) = \frac{1}{5}$$

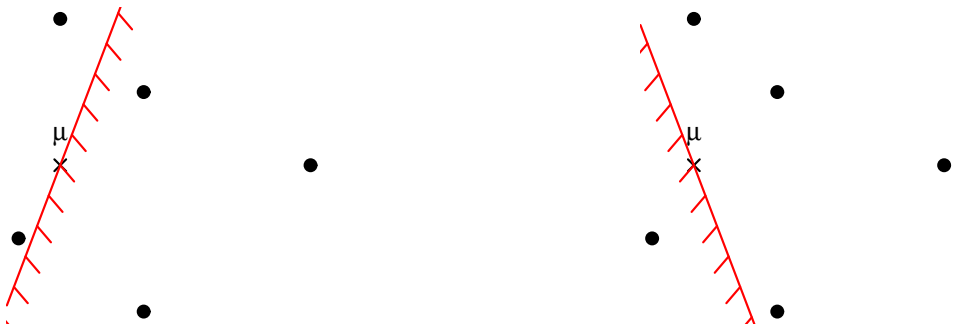


Abbildung 5.1: Halbraum-Tiefe von μ bezüglich 5 Datenpunkten im $\mathbb{R}^2$. Links: Halbraum enthält 3 Punkte. Rechts: Halbraum enthält 1 Punkt.

5.1.11 Definition (Halbraum-Median oder Tukey-Median (Tukey 1975))
Ein Parametervektor $med_H(y)$ heißt Halbraum-Median, falls gilt

$$med_H(y) \in \arg\max_{\mu \in \mathbb{R}^p} d_H(\mu, y).$$

Die Halbraum-Tiefe kann effizient mit dem `R` Paket `ddalpha` von Pokotylo, Mozharovskyi und Dyckerhoff berechnet werden. Das `R` Paket `mrfDepth` von Segaer, Hubert, Rousseeuw und Vakili enthält zusätzlich noch den Halbraum-Median.

5.1.12 Beispiel
Abbildung 5.2 zeigt für 8 Datenpunkte im $\mathbb{R}^2$ die Bereiche der Parameter $\mu \in \mathbb{R}^2$ mit verschiedener Tiefe. Mögliche Tiefen sind hier 0, $\frac{1}{8}$, $\frac{2}{8}$ und $\frac{3}{8}$. Außerhalb der konvexen Hülle der 8 Datenpunkte ist die Tiefe 0. Dann kommt ein ringförmiger Bereich mit der Tiefe $\frac{1}{8}$, dann ein ringförmiger Bereich mit Tiefe $\frac{1}{8}$. Und der innerste Bereich hat die Tiefe $\frac{3}{8}$. Damit wird dort die größte Tiefe angenommen, so dass dieser Bereich der Halbraum-Median bzw. Tukey-Median ist. Dieser Median ist wie der eindimensionale Median somit nicht eindeutig.

5.1.13 Satz
Der Halbraum-Median ist affin äquivariant bezüglich regulärer Transformationen, d.h. für alle $(y_1, \ldots, y_N)^\top \in \mathbb{R}^{N \times p}$, alle regulären Matrizen $A \in \mathbb{R}^{p \times p}$ und alle $b \in \mathbb{R}^p$ gilt

$$med_H((A\,y_1 + b, \ldots, A\,y_N + b)^\top) = A\; med_H((y_1, \ldots, y_N)^\top) + b.$$

Wird der Bruchpunkt wie im univariaten Lokationsfall als Explosionspunkt definiert, so gilt folgende allgemeine Abschätzung.

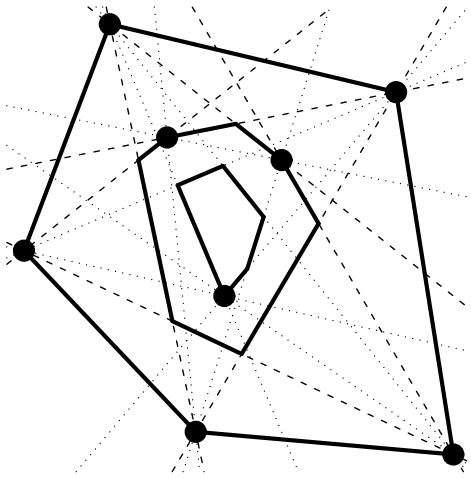


Abbildung 5.2: Bereiche mit verschiedener Halbraum-Tiefe bei 8 Datenpunkten im $\mathbb{R}^2$

5.1.14 Satz
Für den Bruchpunkt des Halbraum-Medians gilt für $N \to \infty$

$$\epsilon^*(med_H, y) \geq \frac{1}{p+1}.$$

Beweis. Siehe
Mizera, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* **30**, 1681-1736.

5.1.15 Definition (Simplex-Tiefe (Liu 1988,1990))
Die Simplex-Tiefe $d_S(\mu, y)$ eines Lageparameters $\mu \in \mathbb{R}^p$ bezüglich des Datensatzes $y = (y_1, \ldots, y_N)^\top$ ist definiert als

$$d_S(\mu, y) = \frac{1}{\binom{N}{p+1}} \sum_{1 \leq n_1 < n_2 < \ldots < n_{p+1} \leq N} \mathbb{1}\{\mu \in \text{Simplex aufgespannt durch } y_{n_1}, y_{n_2}, \ldots, y_{n_{p+1}}\}.$$

Dabei bezeichnet $\mathbb{1}\{\mu \in A\} := \mathbb{1}_A(\mu)$ die Indikatorfunktion bzgl. der Menge A.

Die Simplex-Tiefe kann für zweidimensionale Daten mit dem `R` Paket `mrfDepth` von Segaer, Hubert, Rousseeuw und Vakili berechnet werden.

5.1.16 Beispiel
Abbildung 5.1 zeigt 5 Datenpunkte im $\mathbb{R}^2$ und einen Parameter $\mu \in \mathbb{R}^2$. Das rote Dreieck enthält μ nicht. Insgesamt gibt es 7 Dreicke, die von drei Datenpunkten aufgespannt werden, die μ nicht enthalten, während 3 von den Datenpunkten aufgespannte Dreiecke μ enthalten. Die Simplex-Tiefe dieses Parameters μ ist somit

$$d_S(\mu, (y_1, \ldots, y_5)) = \frac{1}{\binom{5}{3}} \cdot 3 = \frac{3}{10}$$

5.1.17 Definition (Simplex-Median)
Ein Parametervektor $med_S(y)$ heißt Simplex-Median, falls gilt

$$med_S(y) \in \arg\max_{\mu \in \mathbb{R}^p} d_S(\mu, y).$$

5.1.18 Lemma

$$d_S(\mu, y) = \frac{1}{\binom{N}{p+1}} \sum_{1 \leq n_1 < n_2 < \ldots < n_{p+1} \leq N} \mathbb{1}\{d_H(\mu, (y_{n_1}, y_{n_2}, \ldots, y_{n_{p+1}})^\top) > 0\}.$$

Dabei bezeichnet $\mathbb{1}\{h(z) > 0\}$ die Indikatorfunktion $\mathbb{1}_A(z)$ mit $A = \{\tilde{z};\ h(\tilde{z}) > 0\}$ für eine beliebige Funktion h.

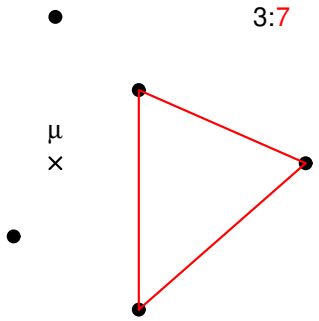


Abbildung 5.3: Simplex-Tiefe von μ bezüglich 5 Datenpunkten im $\mathbb{R}^2$

Beweis. Für eine beliebige Teilmenge $\{n_1, n_2, \dots n_{p+1}\}$ von $\{1, 2, \dots, N\}$ gilt:

$$\mu \in \text{Simplex aufgespannt durch } \; y_{n_1}, y_{n_2}, \dots, y_{n_{p+1}}$$
$$\Longleftrightarrow$$
$$\text{Jeder Halbraum, der } \mu \text{ enthält, enthält mindestens einen Datenpunkt von } y_{n_1}, y_{n_2}, \dots, y_{n_{p+1}}$$
$$\Longleftrightarrow$$
$$0 < d_H(\mu, (y_{n_1}, y_{n_2}, \dots, y_{n_{p+1}})^\top) = \tfrac{1}{p} \min_{u \in \mathbb{R}^p} \sharp\{n \in \{n_1, n_2, \dots n_{p+1}\};\ u^\top (y_n - \mu) \geq 0\}. \ \square$$

5.1.19 Korollar
Der Simplex-Median ist affin äquivariant bezüglich regulärer Transformationen, d.h. für alle $(y_1, \dots, y_N)^\top \in \mathbb{R}^{N \times p}$*, alle regulären Matrizen* $A \in \mathbb{R}^{p \times p}$ *und alle* $b \in \mathbb{R}^p$ *gilt*

$$med_S((A\, y_1 + b, \dots, A\, y_N + b)^\top) = A\ med_S((y_1, \dots, y_N)^\top) + b.$$

5.2 Streuungsschätzungen

5.2.1 Definition ((Empirische) Kovarianzmatrix)
Die (empirische) Kovarianzmatrix $S = S_y$ von $y \in \mathbb{R}^{N\times p}$ ist die $p \times p$-Matrix der (empirischen) Kovarianzen von $y_{(i)}$ und $y_{(j)}$, $i, j = 1, \ldots, p$, d.h.

$$S_y = \begin{pmatrix} s_{y_{(1)}y_{(1)}} & s_{y_{(1)}y_{(2)}} & \cdots & s_{y_{(1)}y_{(p)}} \\ s_{y_{(2)}y_{(1)}} & s_{y_{(2)}y_{(2)}} & \cdots & s_{y_{(2)}y_{(p)}} \\ \vdots & \vdots & \ldots & \vdots \\ s_{y_{(p)}y_{(1)}} & s_{y_{(p)}y_{(2)}} & \cdots & s_{y_{(p)}y_{(p)}} \end{pmatrix},$$

wobei $s_{y(i)y(j)} = \frac{1}{N-1}\sum_{n=1}^{N}(y_{ni} - \overline{y}_{(i)})(y_{nj} - \overline{y}_{(j)})$ für $i, j = 1, \ldots, p$ gilt.

5.2.2 Satz
Für die (empirische) Kovarianzmatrix gilt:

$$\begin{aligned} S_y &= \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \overline{y})(y_n - \overline{y})^\top = \frac{1}{N-1}\left(\sum_{n=1}^{N} y_n\, y_n^\top - N\,\overline{y}\,\overline{y}^\top\right) \\ &= \frac{1}{N-1}\left(y^\top y - N\,\overline{y}\,\overline{y}^\top\right) = \frac{1}{N-1}\, y^\top \left(I_{N\times N} - \frac{1}{N}\, 1_{N\times N}\right) y, \end{aligned}$$

wobei $I_{N\times N}$ die $N \times N$-Einheitsmatrix und $1_{N\times N}$ die $N \times N$-Matrix bestehend aus lauter Einsen ist.

Beweis.
Wir zeigen, dass die (i, j)'te Komponente von $\frac{1}{N-1}\sum_{n=1}^{N}(y_n-\overline{y})(y_n-\overline{y})^\top$ gerade $s_{y_{(i)}y_{(j)}}$ ist. Seien dazu $e_i, e_j \in \mathbb{R}^p$ die i'ten und j'ten Einheitsvektoren. Wegen $e_i^\top y_n = y_{ni}$, $e_i^\top \overline{y} = e_i^\top \frac{1}{N}\, y^\top 1_N = \frac{1}{N}\,(ye_i)^\top 1_N = \frac{1}{N}\, y_{(i)}^\top 1_N = \overline{y_{(i)}}$ und analoger Aussage für e_j ist die (i, j)'te Komponente gegeben durch

$$\begin{aligned} e_i^\top \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \overline{y})(y_n - \overline{y})^\top e_j &= \frac{1}{N-1}\sum_{n=1}^{N}(e_i^\top y_n - e_i^\top \overline{y})(y_n^\top e_j - \overline{y}^\top e_j) \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(y_{ni} - \overline{y_{(i)}})(y_{nj} - \overline{y_{(j)}}) = s_{y_{(i)}y_{(j)}}. \end{aligned}$$

Die zweite und dritte Darstellung sind Übungsaufgabe.

Wegen $y^\top I_{N\times N}\, y = (y_1, \ldots, y_N)\begin{pmatrix} y_1^\top \\ \vdots \\ y_N^\top \end{pmatrix} = \sum_{n=1}^{N} y_n\, y_n^\top$ und

$N\,\overline{y}\,\overline{y}^\top = N\,\frac{1}{N}\,y^\top 1_N\,(\frac{1}{N}\,y^\top 1_N)^\top = \frac{1}{N}\,y^\top 1_N\,1_N^\top\,y = \frac{1}{N}\,y^\top 1_{N\times N}\,y$ folgt die letzte Darstellung aus der dritten Darstellung von S_y. □

5.2.3 Satz
Für die (empirische) Kovarianzmatrix gilt:

$S_z = A\,S_y\,A^\top$ *für* $z = y\,A^\top + 1_N\,b^\top$ *mit* $A \in \mathbb{R}^{q\times p}$ *und* $b \in \mathbb{R}^q$.

Beweis.
Aus Satz 5.2.2 folgt mit

$$\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,1_N = 0_N$$

die Behauptung

$$\begin{aligned}
S_z &= \frac{1}{N-1}\,z^\top\,\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,z \\
&= \frac{1}{N-1}\,(y\,A^\top + 1_N\,b^\top)^\top\,\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,(y\,A^\top + 1_N\,b^\top) \\
&= \frac{1}{N-1}\,(A\,y^\top + b\,1_N^\top)\,\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,(y\,A^\top + 1_N\,b^\top) \\
&= \frac{1}{N-1}\,A\,y^\top\,\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,y\,A^\top + \frac{1}{N-1}\,A\,y^\top\,\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,1_N\,b^\top \\
&\quad + \frac{1}{N-1}\,b\,1_N^\top\,\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,y\,A^\top + \frac{1}{N-1}\,b\,1_N^\top\,\left(I_{N\times N} - \frac{1}{N}\,1_{N\times N}\right)\,1_N\,b^\top \\
&= A\,S_y\,A^\top.\square
\end{aligned}$$

5.2.4 Bemerkung
a) Die Eigenschaft in Satz 5.2.3 entspricht für $p = 1$ der Lokations-Invarianz und der Skalen-Äquivarianz, falls die Wurzel aus S_y gezogen wird.
b) Gilt $z = y\,a$ mit $a = A^\top \in \mathbb{R}^p$ so gilt nach Satz 5.2.2 $S_z = a^\top S_y\,a \in \mathbb{R}$. Das ist aber die (empirische) Varianz von z. Denn es gilt $z_n^\top = y_n^\top\,a \in \mathbb{R}$ und $a^\top\overline{y} = a^\top\,\frac{1}{N}\,y^\top 1_N = \frac{1}{N}\,z^\top 1_N = \overline{z} \in \mathbb{R}$ und somit mit Satz 5.2.2

$$\begin{aligned}
S_z &= a^\top S_y\,a \\
&= \frac{1}{N-1}\sum_{n=1}^N a^\top(y_n - \overline{y})(y_n - \overline{y})^\top\,a = \frac{1}{N-1}\sum_{n=1}^N (a^\top y_n - a^\top\overline{y})(y_n^\top\,a - \overline{y}^\top\,a) \\
&= \frac{1}{N-1}\sum_{n=1}^N (z_n - \overline{z})^2 = \mathrm{var}(z).
\end{aligned}$$

c) Ist $A = \begin{pmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_q^\top \end{pmatrix} \in \mathbb{R}^{q\times p}$, so ergeben die q Zeilen $a_1^\top, \ldots, a_q^\top$ von A neue Variablen, nämlich q neue Variablen. Die Beobachtungen $z_{(1)}, \ldots, z_{(q)}$ zu den q neuen Variablen sind durch $z_{(i)} = z\, a_i$ für $i = 1, \ldots, q$ gegeben.

5.2.5 Korollar
Die Kovarianzmatrix S_y ist eine symmetrische und positiv semidefinite Matrix, d.h. es gilt $S_y^\top = S_y$ (Symmetrie) und $a^\top S_y\, a \geq 0$ für alle $a \in \mathbb{R}^p$ (positive Semidefinitheit).

Beweis. Die Symmetrie folgt aus $s_{y_{(i)}y_{(j)}} = s_{y_{(j)}y_{(i)}}$ für alle $i, j = 1, \ldots, p$. Die positive Semidefinitheit gilt wegen $a^\top S_y\, a = \mathrm{var}(y\, a) \geq 0$. □

Auch wenn die Kovarianz-Matrix eine Verallgemeinerung der Varianz darstellt, ist sie als **Maß für die Streuung der Daten** ungeeignet, da sie eine Matrix ist. Ein Maß für die Streuung sollte aber eindimensionale positive Größe sein. So eine Größe kann man zum Beispiel über das Volumen der Ellipse (des Ellipsoides), das durch die Kovarianzmatrix $S_y \in \mathbb{R}^{p\times p}$ und einem Lokationsvektor $\mu \in \mathbb{R}^p$ gegeben ist, gewinnen. Das Volumen einer Ellipse (eines Ellipsoides)

$$\mathcal{E}(\mu, \Sigma) := \{z_0 \in \mathbb{R}^p;\ (z_0 - \mu)^\top\, \Sigma^{-1}\, (z_0 - \mu) \leq c\}.$$

ist proportional zu $\det(\Sigma)$. Also ist das Volumen von

$$\mathcal{E}(\mu, S_y) = \{z_0 \in \mathbb{R}^p;\ (z_0 - \mu)^\top\, S_y^{-1}\, (z_0 - \mu) \leq c\} \tag{5.2}$$

proportional zur Determinante $\det(S_y)$ der Kovarianzmatrix S_y. Somit kann $\sqrt{\det(S_y)}$ als Maß für die Streuung und als Verallgemeinerung der Standardabweichung benutzt werden.

5.2.6 Definition
Eine Streuungsschätzfunktion $\widehat{s} : \mathbb{R}^{N\times p} \ni y \to \widehat{s}(y) \in \mathbb{R}^+$ heißt
a) skalen-äquivariant, falls $\widehat{s}(y\, \mathrm{diag}(s_1, \ldots, s_p)) = |s_1|\cdot|s_2|\cdot \ldots |s_p|\cdot \widehat{s}(y)$ für alle $y \in \mathbb{R}^{N\times p}$ und alle $s_1, \ldots, s_p \in \mathbb{R}$ gilt, wobei $\mathrm{diag}(s_1, \ldots, s_p)$ die Diagonalmatrix mit Diagonalelementen $s_1, \ldots, s_p$ ist,
b) invariant bezüglich orthogonaler Transformationen und Verschiebungen, falls $\widehat{s}(y\, A^\top + 1_N\, b^\top) = \widehat{s}(y)$ für alle $y \in \mathbb{R}^{N\times p}$, alle orthogonalen Matrizen $A \in \mathbb{R}^{p\times p}$ und alle $b \in \mathbb{R}^p$ gilt.

5.2.7 Satz
Die Streuungsschätzfunktion gegeben durch $\sqrt{\det(S_y)}$ ist skalen-äquivariant und invariant bezüglich orthogonaler Transformationen und Verschiebungen.

Beweis.

Die Skalen-Äquivarianz folgt aus der letzten Darstellung in Satz 5.2.2. Aus Satz 5.2.3 ergibt sich für den transformierten Datensatz $z = yA^\top + 1_N\, b^\top$, wobei $A \in \mathbb{R}^{p\times p}$ eine orthogonale Matrix ist,

$$\sqrt{\det(S_z)} = \sqrt{\det(A\, S_y\, A^\top)} = \sqrt{\det(A)^2\ \det(S_y)} = \sqrt{\det(S_y)},$$

weshalb die Invarianz bezüglich orthogonaler Transformationen und Verschiebungen vorliegt. □

$\sqrt{\det(S_y)}$ hat als Verallgemeinerung der Standardabweichung aber einen Bruchpunkt von $\frac{1}{N}$ und damit hat auch die empirische Kovarianzmatrix wie die empirische Varianz und Kovarianz einen Bruchpunkt von $\frac{1}{N}$, d.h. ein Ausreißer kann die Schätzung beliebig verfälschen. Um mehr ausreißer-robuste Schätzungen zu bekommen, kann man die Methode der kürzesten Hälfte zur Schätzung der Streuung bei univariaten Daten verallgemeinern. Dazu ist anzumerken, dass die kürzeste Hälfte $\hat\sigma_{SH}$ auch folgendermaßen definiert werden kann:

$$\hat\sigma_{SH}(y) = \min\left\{2\sigma;\ (\mu,\sigma)\in\mathbb{R}\times\mathbb{R}^+ \text{ mit } \sum_{n=1}^{N} 1_{[\mu-\sigma,\mu+\sigma]}(y_n) \geq \frac{N}{2}\right\}.$$

Dabei ist das Intervall $[\mu-\sigma,\mu+\sigma]$ ein Spezialfall des Ellipsoides (5.2) mit $p=1$ und $c=1$. Sei $\mathbb{R}^{p\times p}_{pos.def.}$ die Menge aller $p\times p$-Matrizen, die symmetrisch und positive definit sind.

5.2.8 Definition (Minimum-Volumen-Ellipsoid-Schätzung oder MVE-Schätzung (Rousseeuw 1985))
Für $y\in\mathbb{R}^{N\times p}$ ist die MVE-Schätzung für den Lokationsvektor μ und die Kovarianzmatrix Σ gegeben durch

$$(\hat\mu_{MVE}(y),\hat\Sigma_{MVE}(y)) \in \arg\min\left\{\det(\Sigma);\ \mu\in\mathbb{R}^p,\Sigma\in\mathbb{R}^{p\times p}_{pos.def.}\ \textit{mit}\ \sum_{n=1}^{N} 1_{\mathcal{E}(\mu,\Sigma)}(y_n)\geq\frac{N+p+1}{2}\right\}.$$

Die Größe $\det(\hat\Sigma_{MVE}(y))$ ist dann ein Maß für die Streuung und die Verallgemeinerung von $\hat\sigma_{SH}$. Zusätzlich erhält man mit $\hat\mu_{MVE}(y)$ auch eine neue Lageschätzung. Manchmal wird allerdings in der Definition von $(\hat\mu_{MVE}(y),\hat\Sigma_{MVE}(y))$ auch $\sum_{n=1}^{N} 1_{\mathcal{E}(\mu,\Sigma)}(y_n)\geq\frac{N+p+1}{2}$ gefordert. Siehe Agulló Candela (1996): Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm. Proceedings in Computational Statistics. Ed. Prat, Physica-Verlag, 175-180.

Eine weitere robuste Streuungsschätzung ist die Minimum-Kovarianz-Determinant-Schätzung.

5.2.9 Definition (Minimum-Kovarianz-Determinant-Schätzung oder MCD-Schätzung (Rousseeuw 1985))
Sei für $h \in \{\lceil (N+p+1)/2 \rceil, \ldots, N\}$

$$\mathcal{Y}^h(y) := \left\{ (y_{n_1}, \ldots, y_{n_h})^\top;\ \{n_1, \ldots, n_h\} \text{ ist } h\text{-elementige Teilmenge von } \{1, \ldots, N\} \right\}$$

die Menge aller möglichen Beobachtungsmatrizen basierend auf h *Beobachtungsvektoren aus den* N *Beobachtungsvektoren* $y_1, \ldots, y_N \in \mathbb{R}^p$. *Die MCD-Schätzung für den Lokationsvektor* μ *und die Kovarianzmatrix* Σ *ist dann gegeben durch* $(\hat{\mu}_{MCD}(y), \hat{\Sigma}_{MCD}(y)) = (\overline{y^h_*}, S_{y^h_*})$, *wobei*

$$y^h_* \in \arg\min \left\{ \det(S_{y^h});\ y^h \in \mathcal{Y}^h(y) \right\}$$

gilt.

Minimum-Volumen-Ellipsoid-Schätzung und Minimum-Kovarianz-Determinant-Schätzung können mit dem `R` Paket `MASS` von Ripley et al. mittels `cov.rob` berechnet werden. Das `R` Paket `robustbase` von Maechler et al. liefert mittels `covMcd` auch die Minimum-Kovarianz-Determinant-Schätzung.

5.2.10 Satz
Die Streuungsschätzungen $\widehat{s}$ *gegeben durch die Minimum-Volumen-Ellipsoid-Schätzung* $\widehat{s}(y) = \det(\widehat{\Sigma}_{MVE}(y))$ *und durch die Minimum-Kovarianz-Determinant-Schätzung* $\widehat{s}(y) = \det(\widehat{\Sigma}_{MCD}(y))$ *sind skalen-äquivariant und invariant bezüglich orthogonaler Transformationen und Verschiebungen.*

5.2.11 Satz
Ist eine Streuungsschätzfunktion $\widehat{s} : \mathbb{R}^{N\times p} \ni y \to \widehat{s}(y) \in \mathbb{R}^+$ *skalen-äquivariant und invariant bezüglich orthogonaler Transformationen und Verschiebungen, so gilt*

$$\epsilon^*(\widehat{s}, y) \leq \frac{1}{N} \left\lfloor \frac{N-p+1}{2} \right\rfloor .$$

Beweis.
Sei $M = \left\lfloor \frac{N-p+1}{2} \right\rfloor$. Wir führen wieder die Annahme $\widehat{s}(\tilde{y}) \in [a, b]$ mit $a > 0$, $b < \infty$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$ zum Widerspruch.

Weil immer p Punkte in einer Hyperebene liegen, kann wegen der Invarianz bezüglich orthogonaler Transformationen und Verschiebungen ein Koodinatensystem gewählt werden, mit

$$y_1 = \begin{pmatrix} v_1 \\ 0 \end{pmatrix}, \ldots, y_p = \begin{pmatrix} v_p \\ 0 \end{pmatrix}$$

mit $v_1, \ldots, v_p \in \mathbb{R}^{p-1}$. Sei $A_k = \mathrm{diag}(1, \ldots, 1, k) \in \mathbb{R}^{p\times p}$ für $k \in \mathbb{R}$. Setze

$$\begin{aligned}\tilde{y}_n^{1k} &= A_k y_n \text{ für } n = 1,\dots,p+M, \quad \tilde{y}_n^{1k} = y_n \text{ für } n = p+M+1,\dots,N,\\ \tilde{y}_n^{2k} &= y_n \text{ für } n = 1,\dots,p+M, \quad \tilde{y}_n^{2k} = A_k^{-1} y_n \text{ für } n = p+M+1,\dots,N.\end{aligned}$$

Wegen $\tilde{y}_n^{1k} = y_n$ für $n = 1,\dots,p$ gilt $\tilde{y}^{1k} \in \mathcal{Y}_M(y)$. Wegen $M \geq \frac{N-p}{2}$ und wegen $N-(M+p) \leq N - \frac{N-p}{2} - p = \frac{N-p}{2} \leq M$ gilt auch $\tilde{y}^{2k} \in \mathcal{Y}_M(y)$. Außerdem gilt $\tilde{y}_n^{1k} = A_k \tilde{y}_n^{2k}$ für alle $n = 1,\dots,N$, so dass die Skalen-Äquivarianz $\widehat{s}(\tilde{y}^{1k}) = |k|\,\widehat{s}(\tilde{y}^{2k})$ für alle $k \in \mathbb{R}$ liefert. Somit kann nicht $\widehat{s}(\tilde{y}) \in [a,b]$ mit $a > 0$, $b < \infty$ für alle $\tilde{y} \in \mathcal{Y}_M(y)$ gelten. □

5.2.12 Satz
Für die Streuungsschätzungen $\widehat{s}$ gegeben durch die Minimum-Volumen-Ellipsoid-Schätzung $\widehat{s}(y) = \det(\widehat{\Sigma}_{MVE}(y))$ und durch die Minimum-Kovarianz-Determinant-Schätzung $\widehat{s}(y) = \det(\widehat{\Sigma}_{MCD}(y))$ gilt

$$\epsilon^*(\widehat{s}, y) = \frac{1}{N}\left\lfloor \frac{N-p+1}{2} \right\rfloor,$$

für $y = (y_1,\dots,y_N)^\top \in \mathbb{R}^{N\times p}$, wobei $y_1,\dots,y_N$ in allgemeiner Lage liegen, d.h. es gibt keine Hyperebene, die $p+1$ Beobachtungen enthält.

Beweis.
Vorbemerkung: Statt S_y kann bei der Definition der Minimum-Volumen-Ellipsoid-Schätzung und Minimum-Kovarianz-Determinant-Schätzung auch

$$S(\{y_1,\dots,y_N\}) := \sum_{n=1}^{N}(y_n - \overline{y})(y_n - \overline{y})^\top$$

betrachtet werden. D.h. die Normierung durch $N-1$ wird weggelassen. Dann gilt

$$S(\{y_{m_1},\dots,y_{m_k}\}) \geq S(\{y_{n_1},\dots,y_{n_l}\})$$

für beliebige Mengen $\{y_{n_1},\dots,y_{n_l}\} \subset \{y_{m_1},\dots,y_{m_k}\}$ im Sinne der Löwner-Ordnung, da nach Korollar 5.2.5 $S(\{y_{m_1},\dots,y_{m_k}\} \setminus \{y_{n_1},\dots,y_{n_l}\})$ eine positiv semidefinite Matrix ist. Damit gilt aber auch

$$\det(S(\{y_{m_1},\dots,y_{m_k}\})) \geq \det(S(\{y_{n_1},\dots,y_{n_l}\})). \tag{5.3}$$

1. Behauptung: Für den Implosionspunkt gilt $\epsilon^-(\widehat{s}, y) \geq \frac{1}{N}\left\lfloor \frac{N-p+1}{2} \right\rfloor$.
Da $y_1,\dots,y_N$ in allgemeiner Lage liegen, gibt es $B > 0$, so dass für alle Teilmengen $\{y_{n_1},\dots,y_{n_{p+1}}\} \subset \{y_1,\dots,y_N\}$ folgendes gilt:

$$\det(S(\{y_{n_1},\dots,y_{n_{p+1}}\})) \geq B \tag{5.4}$$

und

$$\det(\Sigma) > B \text{ für alle Ellipsoide } \mathcal{E}(\mu, \Sigma) \supset \{y_{n_1}, \ldots, y_{n_{p+1}}\}. \tag{5.5}$$

Werden nun $M \leq \left\lfloor \frac{N-p+1}{2} \right\rfloor - 1 = \left\lfloor \frac{N+p+1}{2} \right\rfloor - p - 1$ Beobachtungen abgeändert, so müssen immer $L := \left\lfloor \frac{N+p+1}{2} \right\rfloor$ Beobachtungen noch $p+1$ ursprüngliche Beobachtungen enthalten. D.h. es gibt immer $\{y_{n_1}, \ldots, y_{n_{p+1}}\} \subset \{\tilde{y}_{m_1}, \ldots, \tilde{y}_{m_L}\}$. Nach (5.3) und (5.4) folgt

$$\det(S(\{\tilde{y}_{m_1}, \ldots, \tilde{y}_{m_L}\})) \geq \det(S(\{y_{n_1}, \ldots, y_{n_{p+1}}\})) \geq B$$

und jedes Ellipsoid $\mathcal{E}(\mu, \Sigma)$ enthält mit $\{\tilde{y}_{m_1}, \ldots, \tilde{y}_{m_L}\}$ auch $\{y_{n_1}, \ldots, y_{n_{p+1}}\}$, weshalb nach (5.5) dieses Ellipsoid $\det[\Sigma) \geq B$ erfüllt. Somit ist eine Implosion nicht möglich.

2. Behauptung: Für den Explosionspunkt gilt $\epsilon^+(\widehat{s}, y) \geq \frac{1}{N} \left\lfloor \frac{N-p+1}{2} \right\rfloor$.
Für $M \leq \left\lfloor \frac{N-p+1}{2} \right\rfloor - 1 \leq \frac{N-p+1}{2} - 1 = \frac{N-p-1}{2}$ gilt $N - M > N - \frac{N-p-1}{2} = \frac{N+p+1}{2} \geq \left\lfloor \frac{N+p+1}{2} \right\rfloor$. Damit sind mindestens $\left\lfloor \frac{N+p+1}{2} \right\rfloor$ Beobachtungen unverändert. Da sowohl die Minimum-Volumen-Ellipsoid-Schätzung als auch die Minimum-Kovarianz-Determinant-Schätzung nur auf $\left\lfloor \frac{N+p+1}{2} \right\rfloor$ Beobachtungen basieren, bewirken diese $\left\lfloor \frac{N+p+1}{2} \right\rfloor$ unveränderten Beobachtungen, dass die Schätzung nicht explodieren kann. □

Eine weitere robuste Kovarianzmatrix-Schätzung ist durch die Räumliche Vorzeichen-Kovarianz-Matrix (spatial sign covariance matrix) aus dem `R` Paket `sscor` von Dürre und Vogel gegeben. Diese hat den Vorteil, dass man nicht entscheiden muss, wie viel weggetrimmt werden soll.

5.2.13 Definition (Räumliche Vorzeichen-Kovarianzmatrix-Schätzung (spatial sign covariance matrix), Croux, Ollila, Oja 2002, Oja 2010, Dürre, Tyler, Vogel 2016)
Räumliche Vorzeichen-Kovarianzmatrix-Schätzung ist gegeben durch

$$\widehat{\Sigma}_{SS}(y) = \frac{1}{N} \sum_{n=1}^{N} ssign(y_n - med_1(y))\, ssign(y_n - med_1(y))^\top,$$

wobei $med_1(y)$ wieder der l_1-Median ist und $ssign(z_n) := \frac{z_n}{\|z_n\|}$ die räumlichen Vorzeichen sind.

Alle robusten Kovarianzmatrix-Schätzungen können dazu benutzt werden, robuste Varianten von Verfahren zu gewinnen, die auf Kovarianzmatrix-Schätzungen basieren wie die Hauptkomponenten- und Diskriminanz-Analyse. Eine robuste Klassifikation kann aber auch auf dem Konzept der Datentiefe erfolgen.

5.3 Klassifikation

Jede Tiefe d kann für die Analyse, wie gut zwei Gruppen getrennt sind, und für die Klasifikation einer neuer Beobachtung zu einer von zwei Gruppen genutzt werden. Dazu seien die beiden Gruppen, wie folgt gegeben:

Gruppe 0: $(y_1, \ldots, y_N)$,

Gruppe 1: $(y_{N+1}, \ldots, y_{N+M})$,

$y_n \in \mathbb{R}^p$ für $n = 1, \ldots, N + M$.

5.3.1 Definition (DD-Plot, Liu, Parelius, und Singh 1999, Zhang et al. 2013)
Sei d eine Tiefe. Ein Streudiagramm der Punkte $(d(y_n, (y_1, \ldots, y_N)), d(y_n, (y_{N+1}, \ldots, z_{N+M})))$ für $n = 1, \ldots, N + M$ wird DD-Plot (Depth-Depth-Plot) genannt.

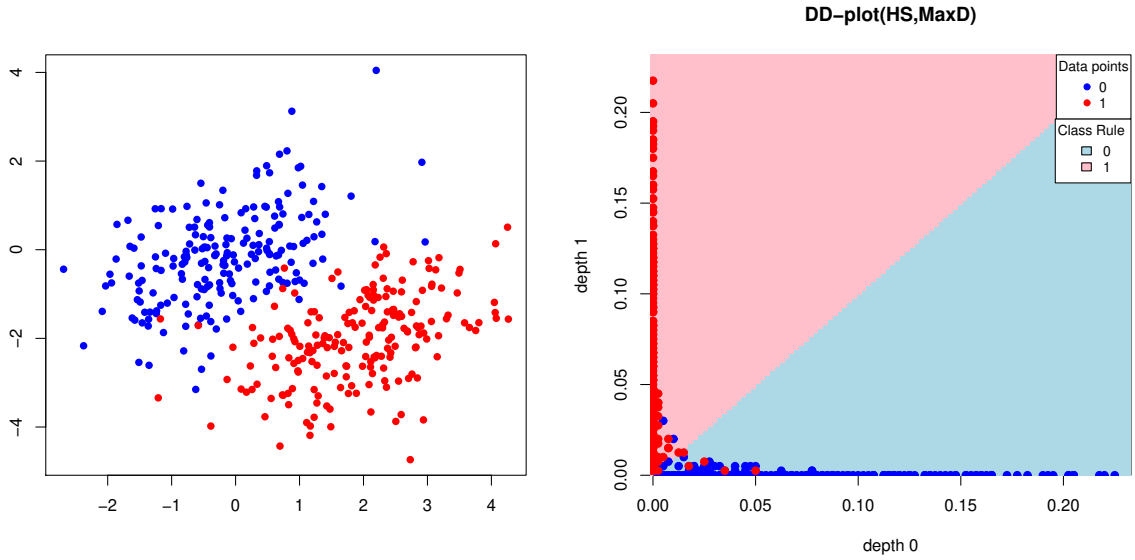


Abbildung 5.4: DD-Plot mittels Halbraum-Tiefe für fast vollständig getrennte Gruppen von zweidimensionalen Daten

Abbildung 5.4 zeigt den DD-Plot basierend auf der Halbraum-Tiefe für zwei stark getrennte Gruppen und Abbildung 5.5 den für überlappende Gruppen. Eine einfache Regel, eine neue Beobachtung y_0 kann zu einer Gruppe zu klassifizieren, steht darin sie der Gruppe zuzuordnen, wo die Tiefe größer ist. Diese Regel wird auch Maximale-Tiefe-Klassifikation (MaxD) genannt. Andere Regeln sind aber auch möglich.

DD-Plots und die zugehörige Klassifikation können mit den `R` Paketen `fda.usc` von Bande, de la Fuente et al. und `ddalpha` von Pokotylo, Mozharovskyi und Dyckerhoff durchgeführt werden. Beide Pakete erlauben auch die Anwendung auf funktionale Daten.

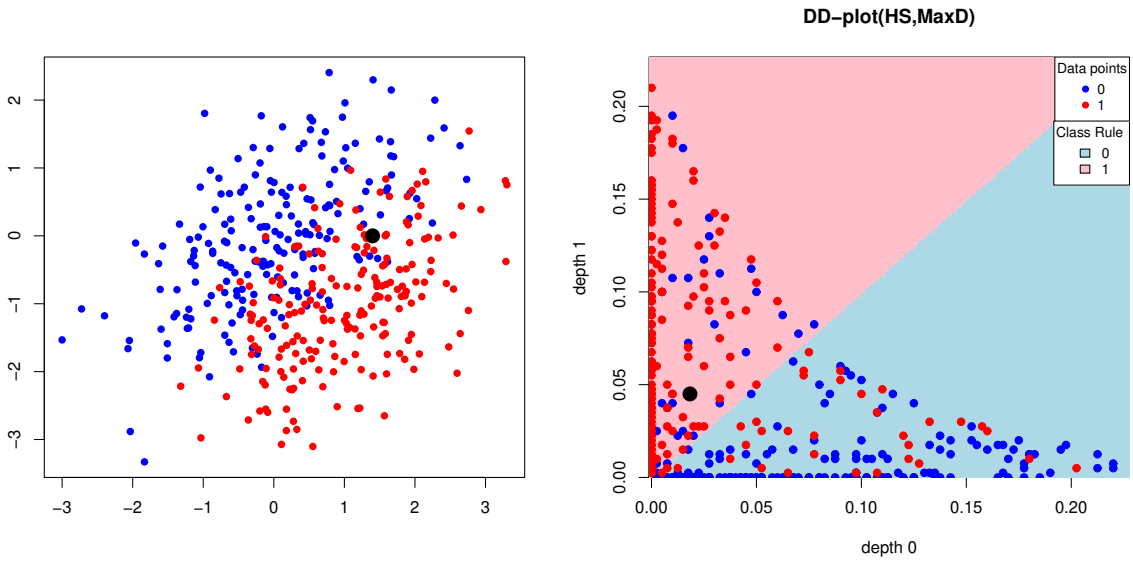


Abbildung 5.5: DD-Plot mittels Halbraum-Tiefe für überlappende Gruppen von zweidimensionalen Daten und Klassifikation einer neuen Beobachtung gegeben durch schwarzen Punkt

5.3.2 Definition (Maximale-Tiefe-Klassifikation)
Eine neue Beobachtung y_0 *wird*

Gruppe 0 zugeordnet, falls $d(y_0,(y_1,\ldots,y_N)) > d(y_0,(y_{N+1},\ldots,y_{N+M}))$ *gilt,*
und
Gruppe 1 zugeordnet, falls $d(y_0,(y_1,\ldots,y_N)) < d(y_0,(y_{N+1},\ldots,y_{N+M}))$ *gilt.*

Abbildung 5.6 zeigt die Trennung zweier Gruppen mit zweidimensionalen Daten mit 15% Kontamination, d.h. 15% der Daten einer Gruppe liegen auf der anderen Seite der anderen Gruppe. Die Vorhersagefehler, d.h. der Anteil der Beobachtungen, die zur falschen Gruppe klassifiziert werden, betragen:
Maximale-Tiefe-Klassifikation mittels Halbraum-Tiefe: 0.511,
Maximale-Tiefe-Klassifikation mittels Simplex-Tiefe: 0.424,
Klassifikation mittels klassischer linearer Diskrimanzanalyse: 0.836,
Klassifikation mittels klassischer quadratischer Diskrimanzanalyse: 0.554,
Klassifikation mittels Random Forest: 0.184.

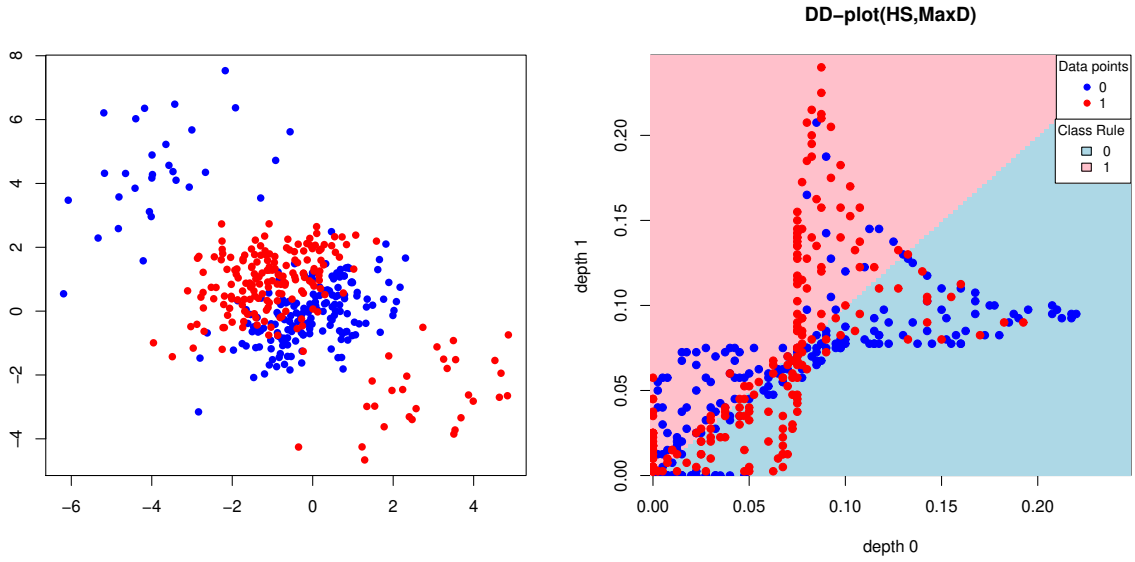


Abbildung 5.6: DD-Plot mittels Halbraum-Tiefe für überlappende Gruppen von zweidimensionalen Daten mit 15% Kontamination

Kapitel 6

Allgemeine lineare Regression

6.1 Klassische Regressionsschätzungen

Bei der allgemeinen linearen Regression werden Beobachtungen $y_1, \dots, y_N \in \mathbb{R}$ bei unabhängige Variablen $t_1, \dots, t_N \in \mathcal{T} \subset \mathbb{R}^q$ gemacht. Die Variablen $t_1, \dots, t_N \in \mathcal{T} \subset \mathbb{R}^q$ sind dabei in der Regel vom Experimentator vorgegeben und werden deshalb auch **Versuchsbedingungen** genannt. $d = (t_1, \dots, t_N)^\top \in \mathbb{R}^{N \times q}$ wird **Versuchsplan** genannt. Ein allgemeines lineares Regressionsmodell liegt dann vor, wenn es eine bekannte **Regressorfunktion** $x : \mathcal{T} \to \mathbb{R}^R$ und einen unbekannten **Regressionsparameter** $\beta \in \mathbb{R}^R$ gibt, so dass der Zusammenhang zwischen den $y_1, \dots, y_N$ und $t_1, \dots, t_N$ durch die Regressionsfunktion $f : \mathcal{T} \to \mathbb{R}$ mit $f(t) = x(t)^\top \beta$ gegeben ist. Dabei wird ein additiver Fehler e_n angenommen angenommen, so dass $y_n = x(t_n)^\top \beta + e_n = f(t_n) + e_n$ für $n = 1, \dots, N$ gilt.
Das allgemeine Regresionsmodel für alle Beobachtungen $y = (y_1, \dots, y_N)^\top \in \mathbb{R}^N$ kann dann wie folgt geschrieben werden

$$y = X_d \beta + e,$$

wobei $X_d = (x(t_1), \dots, x(t_N))^\top$ die **Planungsmatrix**, $d = (t_1, \dots, t_N)^\top$ der Versuchsplan und $e = (e_1, \dots, e_N)^\top$ der **Fehlervektor** ist. Schätzungen, sogenannte **Regressionsschätzungen**, für β hängen dann von y und X_d ab und können wieder mit der Kleinsten-Quadrate-Methode bestimmt werden.

6.1.1 Definition (Kleinste-Quadrat-Summen-Schätzung)
$\hat{\beta}(y, X_d)$ heißt Kleinste-Quadrat-Summen-Schätzung (KQSS) für β bei y und X_d, falls gilt

$$\hat{\beta}(y, X_d) \in \arg\min_{\beta \in \mathbb{R}^R} \sum_{n=1}^{N} (y_n - x(t_n)^\top \beta)^2 = \arg\min_{\beta \in \mathbb{R}^R} (y - X_d \beta)^\top (y - X_d \beta).$$

Die Kleinste-Quadrat-Summen-Schätzung kann auch wieder explizit angegeben werden. Dazu brauchen wir die **verallgemeinerte Inverse, kurz g-Inverse** einer Matrix.

6.1.2 Definition (g-Inverse)
Ist $A \in \mathbb{R}^{n\times m}$, so heißt $A^- \in \mathbb{R}^{m\times n}$ eine g-Inverse von A, falls gilt

$$A\,A^-\,A = A. \tag{6.1}$$

Ist A keine reguläre Matrix, dann ist die g-Inverse von A nicht eindeutig, d.h. es gibt mehrere Matrizen A^-, die die Eigenschaft (6.1) besitzen. Für die g-Inverse von $A^\top A$ gilt das folgende Lemma.

6.1.3 Lemma
Sei $(A^\top A)^-$ eine g-Inverse von $A^\top A$. Dann gilt
a) $((A^\top A)^-)^\top$ ist g-Inverse von $A^\top A$.
b) $A^\top A(A^\top A)^- A^\top = A^\top$ und $A(A^\top A)^- A^\top A = A$.
c) $A(A^\top A)^- A^\top$ ist idempotent, d.h. es gilt $A(A^\top A)^- A^\top\; A(A^\top A)^- A^\top = A(A^\top A)^- A^\top$.
d) $A(A^\top A)^- A^\top$ ist unabhängig von der Wahl der g-Inversen.
e) $A(A^\top A)^- A^\top$ ist symmetrisch.

Beweis.
a) $A^\top A((A^\top A)^-)^\top A^\top A = (A^\top A(A^\top A)^- A^\top A)^\top = (A^\top A)^\top = A^\top A$.
b) Allgemein gilt, dass aus $BD^\top D = CD^\top D$ die Gleichheit $BD^\top = CD^\top$ folgt. Denn aus $BD^\top D = CD^\top D$ folgt

$$\begin{aligned} 0 &= (BD^\top D - CD^\top D)(B-C)^\top = (BD^\top - CD^\top)D(B-C)^\top \\ &= (BD^\top - CD^\top)\,((B-C)D^\top)^\top = (BD^\top - CD^\top)\,(BD^\top - CD^\top)^\top \end{aligned}$$

Wird der letzte Ausdruck von rechts und links mit einem belibigen Vektor x der passenden Dimension multipliziert, so gilt $0 = x^\top(BD^\top - CD^\top)\,(BD^\top - CD^\top)^\top x$, was $0 = (BD^\top - CD^\top)^\top x$ für alle x und damit $0 = BD^\top - CD^\top$ bedeutet.
Auf Grund der Definition der g-Inversen gilt $A^\top A(A^\top A)^- A^\top A = A^\top A$. Setzen wir nun $B = A^\top A(A^\top A)^-$, $C = I$ die Einheitsmatrix und $D = A$, so folgt der erste Teil der Behauptung b). Der zweite Teil ist der erste Teil transponiert, wobei a) eingeht.
c) folgt sofort aus b).
d) Sei $(A^\top A)^\sim$ ebenfalls eine g-Inverse von $A^\top A$. Aus b) folgt $A(A^\top A)^\sim A^\top A = A = A(A^\top A)^- A^\top A$. Setzen wir $B = A(A^\top A)^\sim$, $C = A(A^\top A)^-$ und $D = A$, so folgt aus der Betrachtung in b) $A(A^\top A)^\sim A^\top = A(A^\top A)^- A^\top$. Also hängt $A(A^\top A)^- A^\top$ nicht von der Wahl der g-Inversen ab.
e) Aus a) folgt $(A(A^\top A)^- A^\top)^\top = A((A^\top A)^-)^\top A^\top = A(A^\top A)^- A^\top$.□

6.1.4 Satz
$\hat\beta(y, X_d)$ ist Kleinste-Quadrat-Summen-Schätzung für β bei y und X_d genau dann, wenn eins der folgenden Gleichungen gilt:

$$a) \quad X_d\,\hat\beta(y, X_d) = X_d\,(X_d^\top X_d)^-\,X_d^\top y,$$
$$b) \quad X_d^\top X_d\,\hat\beta(y, X_d) = X_d^\top\,y.$$

6.1.5 Bemerkung
Damit ist $\hat{\beta}(y, X_d)$ Kleinste-Quadrat-Summen-Schätzung, wenn $\hat{\beta}(y, X_d) = (X_d^\top X_d)^- X_d^\top y$ gilt.

Beweis.
a) Sei $\hat{\beta} = (X_d^\top X_d)^- X_d^\top y$. Für alle $\beta \in \mathbb{R}^R$ gilt wegen $X_d(X_d^\top X_d)^- X_d^\top X_d = X_d$ (Lemma 6.1.3 b))

$$\begin{aligned}
(y - X_d\beta)^\top (y - X_d\beta) &= (y - X_d\hat{\beta} + X_d\hat{\beta} - X_d\beta)^\top (y - X_d\hat{\beta} + X_d\hat{\beta} - X_d\beta) \\
&= (y - X_d\hat{\beta})^\top (y - X_d\hat{\beta}) + (y - X_d\hat{\beta})^\top (X_d\hat{\beta} - X_d\beta) \\
&\quad + (X_d\hat{\beta} - X_d\beta)^\top (y - X_d\hat{\beta}) + (X_d\hat{\beta} - X_d\beta))^\top (X_d\hat{\beta} - X_d\beta) \\
&= (y - X_d\hat{\beta})^\top (y - X_d\hat{\beta}) + (y - X_d(X_d^\top X_d)^- X_d^\top y)^\top (X_d\hat{\beta} - X_d\beta) \\
&\quad + (X_d\hat{\beta} - X_d\beta)^\top (y - X_d(X_d^\top X_d)^- X_d^\top y) + (X_d\hat{\beta} - X_d\beta))^\top (X_d\hat{\beta} - X_d\beta) \\
&= (y - X_d\hat{\beta})^\top (y - X_d\hat{\beta}) + (y^\top - y^\top X_d(X_d^\top X_d)^- X_d^\top)\, X_d\,(\hat{\beta} - \beta) \\
&\quad + (\hat{\beta} - \beta)^\top X_d^\top (y - X_d(X_d^\top X_d)^- X_d^\top y) + (X_d\hat{\beta} - X_d\beta))^\top (X_d\hat{\beta} - X_d\beta) \\
&= (y - X_d\hat{\beta})^\top (y - X_d\hat{\beta}) + y^\top (X_d - X_d(X_d^\top X_d)^- X_d^\top X_d)\,(\hat{\beta} - \beta) \\
&\quad + (\hat{\beta} - \beta)^\top (X_d^\top - X_d^\top X_d(X_d^\top X_d)^- X_d^\top) y + (X_d\hat{\beta} - X_d\beta))^\top (X_d\hat{\beta} - X_d\beta) \\
&= (y - X_d\hat{\beta})^\top (y - X_d\hat{\beta}) + (\hat{\beta} - \beta)^\top X_d^\top X_d\,(\hat{\beta} - \beta) \\
&\geq (y - X_d\hat{\beta})^\top (y - X_d\hat{\beta}).
\end{aligned}$$

Dabei gilt genau dann Gleichheit, wenn $X_d\,\beta = X_d\,\hat{\beta} = X_d\,(X_d^\top X_d)^- X_d^\top y$ gilt.

b) Eine notwendige Bedingung für ein Minimum von $g(\beta) = (y - X_d\beta)^\top (y - X_d\beta)$ ist, dass die Ableitung von g gleich 0 ist. Mit den Ableitungsregeln

$$\frac{\partial(x^\top A\,x)}{\partial\,x} = 2\,A\,x, \qquad \frac{\partial(a^\top x)}{\partial\,x} = a = \frac{\partial(x^\top a)}{\partial\,x}$$

für $x, a \in \mathbb{R}^R$ und $A \in \mathbb{R}^{R\times r}$ gilt

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\, g(\beta) = \frac{\partial}{\partial\beta}\left(y^\top y - \beta^\top X_d^\top y - y^\top X_d\beta + \beta^\top X_d^\top X_d\beta\right) = -2X_d^\top y + 2X_d^\top X_d\,\beta = 0 \\
&\iff \quad X_d^\top X_d\,\beta = X_d^\top y.
\end{aligned}$$

Da

$$\frac{\partial^2}{(\partial\beta)^2}\, g(\beta) = 2X_d^\top X_d$$

positiv semidefinit ist, ist jede Lösung β von $X_d^\top X_d\,\beta = X_d^\top y$ eine Minimum-Stelle von g und damit Kleinste-Quadrat-Summen-Schätzung. □

Ist X_d von vollem Rang, dann ist $X_d^\top X_d$ regulär und damit deren Inverse eindeutig. In diesem Fall ist die Kleinste-Quadrat-Summen-Schätzung eindeutig. Ist aber X_d nicht von vollem Rang, so ist die Kleinste-Quadrat-Summen-Schätzung nicht eindeutig. Es gibt dann einen affinen Unterraum von $\mathbb{R}^R$, so dass jedes Element von diesem Unterraum eine Kleinste-Quadrat-Summen-Schätzung ist. Das ist eine Konsequenz aus folgendem Satz.

Auch wenn die Kleinste-Quadrat-Summen-Schätzung nicht eindeutig ist, kann es also sein, dass bestimmte **Aspekte** des Parameters β eindeutig bestimmt sind. Wir konzentrieren uns hier auf die **linearen Aspekte**.

6.1.6 Definition (Linearer Aspekt)
Ist $L \in \mathbb{R}^{S\times R}$ vom Rang S, so heißt $\varphi(\beta) = L\beta$ linearer Aspekt von $\beta \in \mathbb{R}^R$.

6.1.7 Definition (Gauß-Markoff-Schätzung für linearen Aspekt)
$\hat{\varphi}(y, X_d)$ heißt Gauß-Markoff-Schätzung oder auch Kleinste-Quadrat-Summen-Schätzung für den linearen Aspekt $\varphi(\beta) = L\beta$ bei y und X_d, falls gilt $\hat{\varphi}(y, X_d) = L\hat{\beta}(y, X_d)$, wobei $\hat{\beta}(y, X_d)$ eine Kleinste-Quadrat-Summen-Schätzung für β bei y und X_d ist.

Die folgende Definition liefert eine Bedingung, wann die Gauß-Markoff-Schätzung eindeutig ist.

6.1.8 Definition (Lineare Identifizierbarkeit)
Ein linearer Aspekt $\varphi(\beta) = L\beta$ heißt linear identifizierbar bei d, wenn für alle $\beta \in \mathbb{R}^R$ gilt

$$X_d\beta = 0_N \Longrightarrow L\beta = 0_S.$$

6.1.9 Satz
a) Der lineare Aspekt $\varphi(\beta) = L\beta$ ist linear identifizierbar bei d genau dann, wenn es ein $K \in \mathbb{R}^{S\times N}$ gibt mit $L = K\,X_d$.

b) Ist der lineare Aspekt $\varphi(\beta) = L\beta$ linear identifizierbar bei d, dann ist die Gauß-Markov-Schätzung für den linearen Aspekt $\varphi(\beta) = L\beta$ bei y und X_d eindeutig.

c) Ist der lineare Aspekt $\varphi(\beta) = L\beta$ nicht linear identifizierbar bei d, dann gibt es einen affinen Unterraum U von $\mathbb{R}^S$, so dass jedes Element von U eine Gauß-Markov-Schätzung für den linearen Aspekt $\varphi(\beta) = L\beta$ bei y und X_d ist.

Beweis.
a) Da die Rückrichtung klar ist, muss nur die Hinrichtung der Äquivalenz gezeigt werden. Sei also $\varphi(\beta) = L\beta$ linear identifizierbar bei d. Sei $b \in \mathbb{R}^R$ beliebig und setze $\beta = b - (X_d^\top X_d)^- X_d^\top X_d\, b$. Dann gilt nach Lemma 6.1.3b) $X_d\beta = (X_d - X_d(X_d^\top X_d)^- X_d^\top X_d)\, b = 0_N$. Aus der linearen Identifizierbarkeit folgt $0_s = L\beta = Lb - L(X_d^\top X_d)^- X_d^\top X_d\, b$ und damit $Lb = L(X_d^\top X_d)^- X_d^\top X_d\, b$ für alle $b \in \mathbb{R}^R$. Daraus folgt $L = L(X_d^\top X_d)^- X_d^\top X_d = K\,X_d$ mit $K = L(X_d^\top X_d)^- X_d^\top$.
b) Aus a) folgt $L = K\,X_d$, so dass nach Satz 6.1.4 die Gauß-Markoff-Schätzung die Form $L\hat{\beta}(y, X_d) = K\,X_d(X_d^\top X_d)^- X_d^\top y$ hat. Da $X_d(X_d^\top X_d)^- X_d^\top$ nach Lemma 6.1.3 d) nicht von der Wahl der g-Inversen abhängt, ist die Gauß-Markoff-Schätzung eindeutig.
c) Ist $\varphi(\beta) = L\beta$ nicht linear identifizierbar, so gibt es $\tilde{\beta} \in \mathbb{R}^R$ mit $X_d\tilde{\beta} = 0_N$ und $L\tilde{\beta} = l \neq 0_S$. Sei $\hat{\beta}$ der Parameter einer Kleinste-Quadrat-Summen-Schätzung bei y und X_d. Wir zeigen nun, dass jedes Element des affinen Unterraumes $U = \{L\hat{\beta} + kl;\ k \in \mathbb{R}\}$ eine Gauß-Markoff-Schätzung ist. Sei dazu $L\hat{\beta} + kl \in U$ beliebig. Nach Satz 6.1.4 gilt $X_d^\top X_d\hat{\beta} = X_d^\top y$ und damit auch wegen $X_d\tilde{\beta} = 0_N$

$$X_d^\top X_d(\hat{\beta} + k\,\tilde{\beta}) = X_d^\top y.$$

Nach Satz 6.1.4 ist dann $\hat{\beta} + k\,\tilde{\beta}$ auch Kleinste-Quadrat-Summen-Schätzung und damit $L(\hat{\beta} + k\,\tilde{\beta}) = L\hat{\beta} + kl$ eine Gauß-Markoff-Schätzung. □

6.2 Alternative Regressionsschätzungen

Die Gauß-Markoff-Schätzfunktion ist als Verallgemeinerung des arithmetischen Mittel sehr ausreißerempfindlich. Um alternative Regressionsschätzungen zu bestimmen, formulieren wir wieder Anforderungen an diese Schätzungen. Diese Anforderungen sind wie für Lage- und Streuungsschätzungen die Skalen-Äquivarianz und alternativ zur Lokations-Äquivarianz die **Regressions-Äquivarinaz**. Dabei fassen wir die Schätzfunktionen für $\varphi(\beta) = L\beta$ als Abbildungen von $\mathbb{R}^N \times \mathbb{R}^{N\times r}$ nach $\mathbb{R}^s$ auf.

6.2.1 Definition (Regressions-Äquivarianz)
Eine Schätzfunktion $\hat{\varphi} : \mathbb{R}^N \times \mathbb{R}^{N\times r} \ni (y, X_d) \to \hat{\varphi}(y, X_d) \in \mathbb{R}^s$ für den linearen Aspekt $\varphi(\beta) = L\beta$ heißt regressions-äquivariant, falls für alle $y \in \mathbb{R}^N$, $X_d \in \mathbb{R}^{N\times r}$, $\gamma \in \mathbb{R}^R$ gilt

$$\hat{\varphi}(y + X_d\gamma, X_d) = \hat{\varphi}(y, X_d) + L\gamma.$$

6.2.2 Satz
Die Gauß-Markoff-Schätzfunktion für $\varphi(\beta) = L\beta$ ist regressions-äquivariant, falls $\varphi(\beta) = L\beta$ bei d linear identifizierbar ist.

Beweis. Der Beweis ist eine Übungsaufgabe.

Da aus jeder Schätzung $\hat{\beta}$ für β durch $\hat{\varphi} = L\hat{\beta}$ eine Schätzung für den linearen Aspekt $\varphi(\beta) = L\beta$ gewonnen werden kann, werden im folgenden nur Schätzungen für β definiert.

Eine Verallgemeinerung der Lokations-M-Schätzung ist die Regressions-M-Schätzung.

6.2.3 Definition (Regressions-M-Schätzung)
Die Regressions-M-Schätzung $\hat{\beta}_\rho(y, X_d)$ bezüglich ρ basierend auf y und X_d ist definiert durch

$$\hat{\beta}_\rho(y, X_d) \in \arg\min_{\beta\in\mathbb{R}^R} \sum_{n=1}^{N} \rho(y_n - x(t_n)^\top \beta).$$

Ist $\rho(z) = |z|$, so erhält man eine Verallgemeinerung des Medians. Da auch andere Verallgemeinerungen des Medians möglich sind (siehe unten), wird diese Verallgemeinerung auch **L_1-Regressionsschätzung** oder kurz **L_1-Schätzung** genannt.

Eine Verallgemeinerung der Lokations-LTS-Schätzung ist die Regression-LTS-Schätzung.

6.2.4 Definition (Rgeression-LTS-Schätzung (Rousseeuw und Leroy 1987))
Die Regression-LTS-Schätzung $\hat{\beta}_{k,h}(y, X_d)$ bezüglich k und h basierend auf y und X_d ist definiert durch

$$\hat{\beta}_{k,h}(y, X_d) \in \arg\min_{\beta \in \mathbb{R}^R} \sum_{n=k}^{h} r_{(n)}(y, X_d, \beta)^2.$$

Dabei ist $r_n(y, X_d, \beta) = |y_n - x(t_n)^\top \beta|$ das sogenannte absolute **Residuum** *der n'ten Beobachtung und $r_{(1)}(y, X_d, \beta) \leq r_{(2)}(y, X_d, \beta) \leq \ldots \leq r_{(N)}(y, X_d, \beta)$ sind die geordneten absoluten Residuen.*

Sowohl die Regressions-M-Schätzungen als auch die Regressions-LTS-Schätzungen enthalten als Spezialfall die Kleinste-Quadrat-Summen-Schätzung, nämlich für $\rho(z) = z^2$ bzw. für $k = 1$ und $h = N$.

Regressions-M-Schätzungen und Regressions-LTS-Schätzungen können mit dem `R` Paket `MASS` von Ripley et al. mittels `rlm` bzw. `lqs` (`lmsreg` und `ltsreg`) berechnet werden. Das `R` Paket `robustbase` von Maechler et al. liefert auch eine Berechnung der Regressions-LTS-Schätzung mittels `ltsReg`.

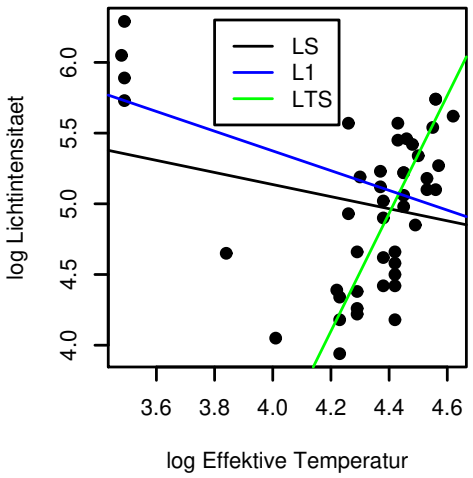


Abbildung 6.1: Hertzsprung-Russell-Daten mit Geraden-Schätzungen

6.2.5 Beispiel (Hertzsprung-Russell-Daten)
Die Abbildung 6.1 zeigt die Ergebnisse der Kleinsten-Quadrat-Summen-Schätzung (LS), der L_1-Regressionsschätzung (L1) und der LTS-Schätzung mit $k = 1$ und $h = 24$ für die Hertzsprung-Russell-Daten. Diese Daten bestehen aus dem Logarithmus der Temperatur (t) und dem Logarithmus der Lichtintensität (y) von 46 Sternen. Es wird eine lineares Regressionsmodell $f(t) = \beta_0 + \beta_1 t$ angenommen. Dieser Datensatz enthält einige Ausreißer, nämlich die Werte von sogenannten Giganten. Es zeigt sich deutlich, dass die LTS-Schätzung von den Ausreißern fast gar

nicht beeinflusst wird, während die Beeinflussung bei den beiden anderen beiden Schätzungen sehr stark ist. Dies ist auch bei der L_1-Schätzung der Fall, obwohl diese den Median verallgemeinert.

Eine Regressions-M-Schätzung mit $\rho(z) = |z|$ verallgemeinert die Charakterisierung des Medians, die durch Satz 3.2.3 gegeben ist. Eine andere Verallgemeinerung des Medians ist aber auch über den Satz 3.2.7 möglich. Die Verallgemeinerung dieser Charakterisierung liefert ganz andere Regressionsschätzungen, die von **Rousseeuw und Hubert (1999)** vorgeschlagen wurden. Dazu muss erstmal die **Regressionstiefe** über den **Regressions-Nonfit** definiert werden.

6.2.6 Definition (Regressions-Nonfit)
Ein Regressionsparameter $\beta \in \mathbb{R}^R$ ist ein Regressions-Nonfit bezüglich $(y_1, t_1^\top)^\top, \ldots, (y_N, t_N^\top)^\top$, falls es ein $\tilde{\beta}$ gibt mit $|y_n - x(t_n)^\top \tilde{\beta}| < |y_n - x(t_n)^\top \beta|$ für alle $n = 1, \ldots, N$.

6.2.7 Beispiel (Beispiele von Regressions-Nonfits)
Abbildung 6.2 zeigt Biespiele von Nonfits und keinen Nonfits.

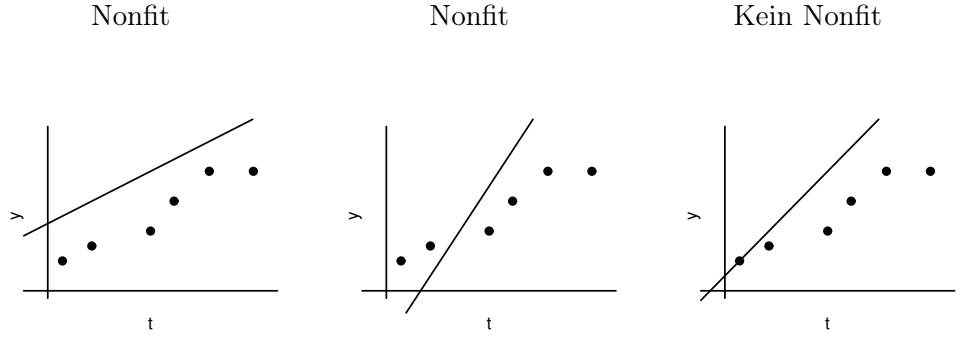


Abbildung 6.2: Beispiele von Nonfits und keinen Nonfits

6.2.8 Definition (Regressions-Tiefe)
Die Regressions-Tiefe $d_R(\beta, y, X_d)$ von β bezüglich y und X_d ist definiert als

$$d_R(\beta, y, X_d) = \frac{1}{N} \min \{M; \quad \text{es existieren } n_1, \ldots, n_M \in \{1, \ldots, N\}, \text{ so dass } \beta \text{ ein Regressions-Nonfit bezüglich } (y_n, t_n^\top)^\top \text{ mit } n \in \{1, \ldots, N\} \setminus \{n_1, \ldots, n_M\} \text{ ist} \Big\}.$$

6.2.9 Definition (Maximum-Regressions-Tiefe-Schätzung)
Die Maximum-Regressions-Tiefe-Schätzung $\hat{\beta}_R(y, X_d)$ *ist definiert als*

$$\hat{\beta}_R(y, X_d) \in \arg\max_{\beta \in \mathbb{R}^R} d_R(\beta, y, X_d).$$

Die Regressions-Tiefe und die Maximum-Regressions-Tiefe-Schätzung können mit dem `R` Paket `mrfDepth` von Segaer, Hubert, Rousseeuw und Vakili berechnet werden.

6.2.10 Beispiel (Hertzsprung-Russell-Daten: Fortsetzung von Beispiel (6.2.5))
Die Abbildung 6.3 zeigt die Geradenschätzungen und deren Regressionstiefe für die Kleinste-Quadrat-Summen-Schätzung (LS), die L_1-Regressionsschätzung (L1) und die Catline (Cat). Dabei ist die Catline eine Methode, um möglichst tiefe Geraden zu finden. Diese Methode liefert wie die LTS-Schätzung eine Gerade, die sich nach der Mehrheit der Daten richtet und die Ausreißer ignoriert.

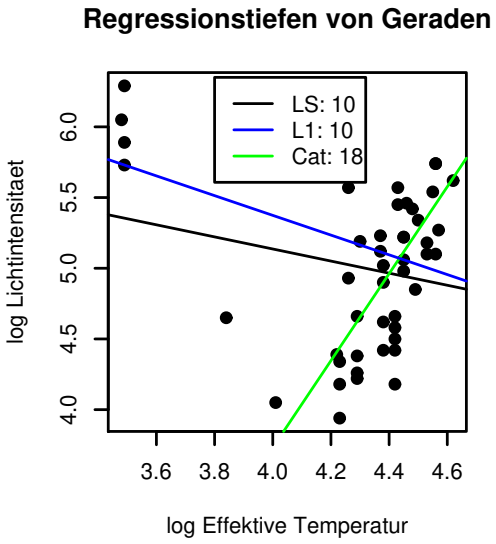


Abbildung 6.3: Regressionstiefen von Geraden-Schätzungen

6.2.11 Lemma
Im allgemeinen gilt

$$\max_{\beta \in \mathbb{R}^R} d_R(\beta, y, X_d) \geq \frac{1}{R+1}.$$

Beweis. Siehe Mizera (2002). □

6.2.12 Lemma
Gilt $\{t_1,\dots,t_N\} \subset \{t_1^*,\dots,t_R^*\}$ *für* $d = (t_1,\dots,t_N)$, *wobei* β *nicht identifizierbar bei* $\{t_1^*,\dots,t_{r-1}^*,t_{r+1}^*,\dots,t_R^*\}$ *für alle* $r = 1,\dots,R$ *gilt, d.h.* $\{t_1^*,\dots,t_R^*\}$ *ist minimaler Träger eines Plans, bei dem* β *identifizierbar ist, dann gilt*

$$\max_{\beta\in\mathbb{R}^R} d_R(\beta, y, X_d) \geq \frac{1}{2}.$$

Beweis. Sei $\widehat{y}_r \in \text{med}\{y_n;\ t_n = t_r^*\}$. Dann existiert ein $\beta^* \in \mathbb{R}^R$ mit $\widehat{y}_r = x(t_r^*)^\top\beta^*$ für alle $r = 1,\dots,R$. Gemäß der Definition von $\widehat{y}_1,\dots,\widehat{y}_R$ gilt für $r = 1,\dots,R$

$$\frac{\sharp\{n\ t_n = t_r^*,\ y_n - x(t_r^*)^\top\beta^* \geq 0\}}{\sharp\{n\ t_n = t_r^*\}} \geq \frac{1}{2} \leq \frac{\sharp\{n\ t_n = t_r^*,\ y_n - x(t_r^*)^\top\beta^* \leq 0\}}{\sharp\{n\ t_n = t_r^*\}}$$

Das bedeutet, dass bei jedem t_r^* mindestens die Hälfte der Beobachtungen y_n mit $t_n = t_r^*$ entfernt werden muss, so dass ein $\tilde{\beta}$ gibt mit $|y_n - x(t_n)^\top\tilde{\beta}| < |y_n - x(t_n)^\top\beta^*|$. □

6.2.13 Satz
Die Regressions-M-Schätzfunktionen, Regressions-LTS-Schätzfunktionen und die Maximum-Regressions-Tiefe-Schätzfunktionen sind regressions-äquivariant.

Beweis. Wir zeigen die Behauptung nur für die Maximum-Regressions-Tiefe-Schätzung.
1.z.z.: Die Regressions-Tiefe ist regressions-invariant, d.h. es gilt

$$d_R(\beta, y, X_d) = d_R(\beta+\gamma, y + X_d\,\gamma,\ X_d)$$

für alle $\beta,\ \gamma \in \mathbb{R}^R,\ y \in \mathbb{R}^N$.
Bew.: Sei $\beta,\ \gamma \in \mathbb{R}^R,\ y \in \mathbb{R}^N$ beliebig und

$$M_0 \in \arg\min\{M;\ \text{ es existieren } n_1,\dots,n_M \in \{1,\dots,N\}, \text{ so dass } \beta \text{ ein Regressions-Nonfit bezüglich } (y_n,t_n) \text{ mit } n \in \{1,\dots,N\}\setminus\{n_1,\dots,n_M\} \text{ ist }\}.$$

Sei o.B.d.A. β Regressions-Nonfit bezüglich $(y_1,t_1),\dots,(y_{N-M_0},t_{N-M_0})$. Dann existiert $\tilde{\beta}$ mit $|y_n - x(t_n)^\top\tilde{\beta}| < |y_n - x(t_n)^\top\beta|$ für alle $n = 1,\dots,N-M_0$. Für $\beta^* = \tilde{\beta}+\gamma$ gilt

$$|y_n + x(t_n)^\top\gamma - x(t_n)^\top\beta^*| = |y_n - x(t_n)^\top\tilde{\beta}| < |y_n - x(t_n)^\top\beta| = |y_n + x(t_n)^\top\gamma - x(t_n)^\top(\beta+\gamma)|$$

für alle $n = 1,\dots,N-M_0$. Das bedeutet, dass $\beta+\gamma$ ein Regressions-Nonfit bezüglich $(y_1 + x(t_1)^\top\gamma,\ t_1),\dots,(y_{N-M_0} + x(t_{N-M_0})^\top\gamma,\ t_{N-M_0})$ ist. Also gilt

$$d_R(\beta+\gamma, y + X_d\gamma, X_d) \leq \frac{1}{N}\,M_0 = d_R(\beta, y, X_d).$$

Ist umgekehrt $M_0 = d_R(\beta+\gamma, y + X_d\gamma, X_d)$, so kann wieder o.B.d.A. angenommen werden, dass $\beta+\gamma$ ein Regressions-Nonfit bzgl. $(y_1 + x(t_1)^\top\gamma,\ t_1),\dots,(y_{N-M_0} + x(t_{N-M_0})^\top\gamma,\ t_{N-M_0})$

ist. Also existiert $\tilde{\beta}$ mit $|y_n + x(t_n)^\top\gamma - x(t_n)^\top\tilde{\beta}| < |y_n + x(t_n)^\top\gamma - x(t_n)^\top(\beta+\gamma)|$ für alle $n = 1, \ldots, N - M_0$. Für $\beta^* = \tilde{\beta} - \gamma$ gilt dann

$$|y_n - x(t_n)^\top(\tilde{\beta}-\gamma)| = |y_n + x(t_n)^\top\gamma - x(t_n)^\top\tilde{\beta}| < |y_n + x(t_n)^\top\gamma - x(t_n)^\top(\beta+\gamma)| = |y_n - x(t_n)^\top\beta|$$

für alle $n = 1, \ldots, N-M_0$. Also ist β ein Regressions-Nonfit bezüglich $(y_1, t_1), \ldots, (y_{N-M_0}, t_{N-M_0})$, woraus

$$d_R(\beta, y, X_d) \leq \frac{1}{N} M_0 = d_R(\beta + \gamma, y + X_d\gamma, X_d)$$

folgt.
2.z.z.: $\hat{\beta}_R(y + X_d\,\gamma,\, X_d) = \hat{\beta}_R(y,\, X_d) + \gamma$.
Bew.: Wegen der Regressions-Invarianz der Regressions-Tiefe gilt:

$$\begin{aligned} d_R(\hat{\beta}_R(y + X_d\gamma, X_d), y + X_d\gamma, X_d) &= \max_{\beta\in\mathbb{R}^R} d_R(\beta, y + X_d\gamma, X_d) \\ &= \max_{\beta\in\mathbb{R}^R} d_R(\beta + \gamma, y + X_d\gamma, X_d) = \max_{\beta\in\mathbb{R}^R} d_R(\beta, y, X_d) \\ &= d_R(\hat{\beta}_R(y, X_d), y, X_d) = d_R(\hat{\beta}_R(y, X_d) + \gamma, y + X_d\gamma, X_d). \end{aligned}$$

Also gilt $\hat{\beta}_R(y + X_d\,\gamma,\, X_d) = \hat{\beta}_R(y,\, X_d) + \gamma$. □

In Abschnitt 5.1 wurde gezeigt, dass die Simplex-Tiefe sich mittels der Halbraum-Tiefe darstellen lässt, d.h. dass

$$d_S(\mu, y) = \frac{1}{\binom{N}{p+1}} \sum_{1\leq n_1<\ldots<n_{p+1}\leq N} \mathbb{1}\{d_H(\mu, (y_{n_1}, \ldots, y_{n_{p+1}})) > 0\}.$$

gilt. Benutzt man statt der Halbraum-Tiefe die Regressions-Tiefe, so erhält man die Simplex-Regressions-Tiefe.

6.2.14 Definition (Simplex-Regressions-Tiefe (Müller 2005))
Die Simplex-Regressions-Tiefe $d_{SR}(\beta, y, X_d)$ von β bezüglich y und X_d mit $d = (t_1, \ldots, d_N)^\top$ ist definiert als

$$d_{SR}(\beta, y, X_d) = \sum_{1\leq n_1<n_2<\ldots<n_{p+1}\leq N} \mathbb{1}\{d_R(\beta, (y_{n_1}, \ldots, y_{n_{p+1}})^\top, X_{(t_{n_1},\ldots,t_{n_{p+1}})^\top}) > 0\}.$$

Die Simplex-Regressions-Tiefe ist besonders gut zum Testen von Hypothesen geeignet, da sie eine sogenannte U-Statistik ist und von U-Statistiken im Prinzip die asymptotische Verteilung bekannt ist, siehe Kapitel 7.

6.3 Bruchpunkte von Regressionsschätzungen

Die Ausreißer-Robustheit von Regressionsschätzungen wird wieder mit dem Bruchpunkt gemessen. Dabei kann der Bruchpunkt diesmal auf zwei Weisen definiert werden, je nachdem ob die

unabhängigen erklärenden Variablen (Regressoren) Ausreißer enthalten können oder nicht. Sind sie zufällig wie die Temperatur in Beispiel 6.2.5, dann können sie auch Ausreißer enthalten. Werden sie aber vom Experimentator vorgegeben, dann können sie keine Ausreißer enthalten.

6.3.1 Definition (Regressions-Bruchpunkt bei zufälligen Regressoren)
Der Bruchpunkt $\epsilon_Z^*(\hat{\varphi}, y, X_d)$ einer Regressionsschätzfunktion $\hat{\varphi}$ für $\varphi(\beta) = L\beta$ bei y und X_d mit zufälligen Regressoren ist definiert als

$$\epsilon_Z^*(\hat{\varphi}, y, X_d) = \frac{1}{N} \min \left\{ M; \sup_{(\tilde{y},\tilde{d}) \in \mathcal{Y}_Z^M(y,d)} \|\hat{\varphi}(\tilde{y}, X_{\tilde{d}})\| = \infty \right\},$$

wobei $\mathcal{Y}_Z^M(y,d) = \left\{ (\tilde{y},\tilde{d}) \in \mathbb{R}^N \times \mathbb{R}^{N\times q};\ \sharp\{n|\ (y_n,t_n) \neq (\tilde{y}_n,\tilde{t}_n)\} \leq M \right\}$ die Menge aller Datensätze $(\tilde{y},\tilde{d})$ ist, die sich in höchstens M Einzelwerten vom Datensatz (y,d) unterscheiden.

6.3.2 Definition (Regressions-Bruchpunkt bei festen Regressoren)
Der Bruchpunkt $\epsilon_F^*(\hat{\varphi}, y, X_d)$ einer Regressionsschätzfunktion $\hat{\varphi}$ für $\varphi(\beta) = L\beta$ bei y und X_d mit festen Regressoren ist definiert als

$$\epsilon_F^*(\hat{\varphi}, y, X_d) = \frac{1}{N} \min \left\{ M; \sup_{\tilde{y} \in \mathcal{Y}_F^M(y)} \|\hat{\varphi}(\tilde{y}, X_d)\| = \infty \right\},$$

wobei $\mathcal{Y}_F^M(y) = \left\{ \tilde{y} \in \mathbb{R}^N;\ \sharp\{n|\ y_n \neq \tilde{y}_n\} \leq M \right\}$ die Menge aller Beobachtungsvektoren $\tilde{y}$ ist, die sich in höchstens M Einzelwerten vom Beobachtungsvektor y unterscheiden.

6.3.3 Satz
Es gilt für alle $y \in \mathbb{R}^N$ und alle $X_d \in \mathbb{R}^{N\times r}$

$$\epsilon_Z^*(\hat{\varphi}, y, X_d) \leq \epsilon_F^*(\hat{\varphi}, y, X_d).$$

Beweis. Klar.

6.3.4 Satz
Sei $\hat{\varphi}$ die Gauß-Markoff-Schätzfunktion für $\varphi(\beta) = L\beta$. Für alle $y \in \mathbb{R}^N$ und alle $X_d \in \mathbb{R}^{N\times r}$ gilt dann

$$\epsilon_Z^*(\hat{\varphi}, y, X_d) \leq \frac{1}{N} \geq \epsilon_F^*(\hat{\varphi}, y, X_d).$$

Ist $\varphi(\beta) = L\beta$ bei d nicht identifizierbar, dann gilt sogar

$$\epsilon_Z^*(\hat{\varphi}, y, X_d) = 0 = \epsilon_F^*(\hat{\varphi}, y, X_d).$$

Beweis. Der erste Teil ist Übungsaufgabe. Der zweite Teil folgt aus Satz 6.1.9. Nach diesem Satz gibt es einen affine Unterraum U, so dass jedes Element aus U eine Gauß-Markoff-Schätzung ist.

Damit gilt dann

$$\sup\{\|\hat{\varphi}\|;\ \hat{\varphi} \text{ ist Gauß-Markoff-Schätzung bei } y \text{ und } X_d\} = \sup\{\|\gamma\|;\ \gamma \in U\} = \infty. \ \square$$

Damit ein Bruchpunkt größer als 0 erreicht werden kann, ist eine Mindestvoraussetzung, dass $\varphi(\beta) = L\beta$ bei d identifizierbar ist. Tatsächlich hängt der Bruchpunkt vom sogenannten **Identifizierung-Parameter** ab.

6.3.5 Definition (Identifizierungs-Parameter)
Der Identifizierungs-Parameter $\mathcal{N}_\varphi(d)$ ist definiert als

$$\mathcal{N}_\varphi(d) = \max\left\{\sum_{n=1}^{N} 1_D(t_n);\ D \subset \{t_1,\dots,t_N\} \textit{ und } \varphi(\beta) = L\beta \textit{ ist nicht identifizierbar bei } D\right\}.$$

6.3.6 Satz (Müller 1995, 1997)
Ist $\hat{\varphi}$ eine regressions-äquivariante Schätzfunktion für $\varphi(\beta) = L\beta$, dann gilt für alle $y \in \mathbb{R}^N$ und alle $X_d \in \mathbb{R}^{N\times r}$

$$\epsilon_F^*(\hat{\varphi}, y, X_d) \le \frac{1}{N}\left\lfloor \frac{N - \mathcal{N}_\varphi(d) + 1}{2} \right\rfloor.$$

Beweis. O.B.d.A sei $\varphi(\beta) = L\beta$ bei $t_1,\dots,t_{\mathcal{N}_\varphi(d)}$ nicht identifizierbar. Dann gibt es $\tilde{\beta} \in \mathbb{R}^R$ mit $x(t_n)^\top \tilde{\beta} = 0$ für $n = 1,\dots,\mathcal{N}_\varphi(d)$ und $L\tilde{\beta} = l \neq 0$. Sei $M = \left\lfloor \frac{N-\mathcal{N}_\varphi(d)+1}{2} \right\rfloor$. Angenommen es gilt $\epsilon_F^*(\hat{\varphi}, y, X_d) > \frac{1}{N}M$. Dann gibt es $B \in \mathbb{R}$ mit $\|\hat{\varphi}(\tilde{y}, X_d)\| < B$ für alle $\tilde{y} \in \mathcal{Y}_F^M(y)$.
Sei $\tilde{y}^{1k}$ gegeben durch

$$\begin{aligned} \tilde{y}_n^{1k} &= y_n && \text{für } n = 1,\dots,\mathcal{N}_\varphi(d),\\ \tilde{y}_n^{1k} &= y_n + kx(t_n)^\top\tilde{\beta} && \text{für } n = \mathcal{N}_\varphi(d)+1,\dots,\mathcal{N}_\varphi(d)+M,\\ \tilde{y}_n^{1k} &= y_n && \text{für } n = \mathcal{N}_\varphi(d)+M+1,\dots,N, \end{aligned}$$

und $\tilde{y}^{2k}$ gegeben durch

$$\begin{aligned} \tilde{y}_n^{2k} &= y_n && \text{für } n = 1,\dots,\mathcal{N}_\varphi(d)+M,\\ \tilde{y}_n^{2k} &= y_n - kx(t_n)^\top\tilde{\beta} && \text{für } n = \mathcal{N}_\varphi(d)+M+1,\dots,N, \end{aligned}$$

mit $k \in \mathbb{R}$. Dann gilt $\tilde{y}^{1k} \in \mathcal{Y}_F^M(y)$ und wegen $N - \mathcal{N}_\varphi(d) - M \le N - \mathcal{N}_\varphi(d) - \frac{N-\mathcal{N}_\varphi(d)}{2} = \frac{N-\mathcal{N}_\varphi(d)}{2} \le M$ auch $\tilde{y}^{2k} \in \mathcal{Y}_F^M(y)$. Wegen der Wahl von $\tilde{\beta}$ gilt

$$\tilde{y}_n^{1k} = \tilde{y}_n^{2k} + k\,x(t_n)^\top\tilde{\beta} \text{ für alle } n = 1,\dots,N$$

und damit $\tilde{y}^{1k} = \tilde{y}^{2k} + k\,X_d\,\tilde{\beta}$. Aus der Regressions-Äquivarianz von $\hat{\varphi}$ folgt für alle $k \in \mathbb{R}$

$$\hat{\varphi}(\tilde{y}^{1k}, X_d) = \hat{\varphi}(\tilde{y}^{2k}, X_d) + k\,L\,\tilde{\beta} = \hat{\varphi}(\tilde{y}^{2k}, X_d) + k\,l$$

und damit der Widerspruch

$$2B \ge \|\hat{\varphi}(\tilde{y}^{1k}, X_d)\| + \|\hat{\varphi}(\tilde{y}^{2k}, X_d)\| \ge \|\hat{\varphi}(\tilde{y}^{1k}, X_d) - \hat{\varphi}(\tilde{y}^{2k}, X_d)\| = \|k\,l\| = |k|\,\|l\| \overset{k\to\infty}{\longrightarrow} \infty. \square$$

6.3.7 Satz
Für eine Regressions-LTS-Schätzfunktion für $\varphi(\beta) = L\beta$ *gegeben durch* $\hat{\varphi}_{k,h}(y, X_d) = L\hat{\beta}_{k,h}(y, X_d)$ *mit*

$$\left\lfloor \frac{N + \mathcal{N}_\varphi(d) + 1}{2} \right\rfloor \leq h \leq \left\lfloor \frac{N + \mathcal{N}_\varphi(d) + 2}{2} \right\rfloor$$

gilt für alle $y \in \mathbb{R}^N$ *und alle* $X_d \in \mathbb{R}^{N \times r}$

$$\epsilon_F^*(\hat{\varphi}_{k,h}, y, X_d) = \frac{1}{N} \left\lfloor \frac{N - \mathcal{N}_\varphi(d) + 1}{2} \right\rfloor .$$

Beweis. Siehe
Müller, Ch.H. (1997). *Robust Planning and Analysis of Experiments*. Lecture Notes in Statistics **124**, Springer, New York,
oder
Müller, Ch.H. (1995). Breakdown points for designed experiments. *J. Statist. Plann. Inference.* **45**, 413-427.

6.3.8 Satz
Ist $\mathcal{N}_\varphi(d) = \mathcal{N}_\beta(d)$*, dann gilt für eine Regressions-M-Schätzfunktion für* $\varphi(\beta) = L\beta$ *gegeben durch* $\hat{\varphi}_\rho(y, X_d) = L\hat{\beta}_\rho(y, X_d)$ *mit* $\rho(z) = \log(1 + z^2)$ *für alle* $y \in \mathbb{R}^N$ *und alle* $X_d \in \mathbb{R}^{N \times r}$

$$\epsilon_F^*(\hat{\varphi}_\rho, y, X_d) = \frac{1}{N} \left\lfloor \frac{N - \mathcal{N}_\varphi(d) + 1}{2} \right\rfloor .$$

Beweis. Siehe
Mizera, I. and Müller, Ch.H. (1999). Breakdown points and variation exponents of robust M-estimators in linear models. *Ann. Statist.* **27**, 1164-1177.

Nicht nur Regressions-M-Schätzfunktion mit $\rho(z) = \log(1+z^2)$, sogenannte Cauchy-M-Schätzfunktionen, erreichen den höchsten Bruchpunkt. In Mizera und Müller (1999) wird gezeigt, dass das für alle Regressions-M-Schätzfunktion gilt, die $\lim_{t\to\infty} \frac{\rho(tu)}{\rho(t)} = u^0 = 1$ für alle $u > 0$ und einige Regularitätsanforderungen erfüllen.

6.3.9 Satz
Für eine Maximum-Regressions-Tiefe-Schätzfunktion für $\varphi(\beta) = \beta$ gilt für alle $y \in \mathbb{R}^N$ und alle $X_d \in \mathbb{R}^{N\times r}$

$$\epsilon_Z^*(\hat{\varphi}_R, y, X_d) \geq \frac{1}{R+1} - \frac{1}{2}\frac{\mathcal{N}_\varphi(d)}{N}.$$

Gilt $\{t_1, \ldots, t_N\} \subset \{t_1^*, \ldots, t_R^*\}$ für $d = (t_1, \ldots, t_N)$, wobei β nicht identifizierbar bei $\{t_1^*, \ldots, t_{r-1}^*, t_{r+1}^*, \ldots, t_R^*\}$ für alle $r = 1, \ldots, R$ gilt, d.h. $\{t_1^*, \ldots, t_R^*\}$ ist minimaler Träger eines Plans, bei dem β identifizierbar ist, dann gilt

$$\epsilon_F^*(\hat{\varphi}_R, y, X_d) \geq \frac{1}{2}\left(1 - \frac{\mathcal{N}_\varphi(d)}{N}\right).$$

Beweis.
Nach Theorem 4.2 in [Mizera, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* **30**, 1681-1736] gilt

$$\epsilon_Z^*(\hat{\varphi}_R, y, X_d) \geq \eta_G - d_R(\infty, y, X_d),$$

wobei

$$d_R(\infty, y, X_d) := \limsup_{\|\beta\|\to\infty} d_R(\theta, y, X_d)$$

gilt und η_G mit $G = Z, F$ durch

$$\max\{d_R(\beta, \tilde{y}, X_d);\ \beta \in \mathbb{R}^R\} \geq \eta_G \text{ für alle } \tilde{y} \in \mathcal{Y}_G^M(y)$$

gegeben ist. Ist β bei $D \subset \{t_1, \ldots, t_N\}$ bei D nicht identifizierbar, so kann $(\beta_m)_{m\in\mathbb{N}}$ mit $\lim_{m\to\infty} \|\beta_m\| = \infty$ gewählt werden, so dass β_m kein Nonfit für y_n mit $t_n \in D$ ist. Um einen Nonfit von β_m zu erhalten, müssen maximal die Hälfte der y_n mit $t_n \in D$ entfernt werden. Wegen

$$\mathcal{N}_\varphi(d) = \max\left\{\sum_{n=1}^N 1_D(t_n);\ D \subset \{t_1, \ldots, t_N\} \text{ und } \varphi(\beta) = L\beta \text{ ist nicht identifizierbar bei } D\right\}.$$

gilt also $d_R(\infty, y, X_d) = \frac{1}{N}\frac{1}{2}\mathcal{N}_\varphi(d)$. Aus Lemma 6.2.11 und 6.2.12 folgt dann die Behauptung. □

6.4 Ausreißer-Erkennung

Ausreißer sind nicht nur störend, sondern können auch interessante Informationen enthalten. So verbergen sich in den Ausreißern oft interessante Phänomene wie die Giganten im Beispiel 6.2.5 über die Sterne. So ist die Ausreißer-Erkennung eine wichtige Aufgabe. Die Ausreißer-Erkennung geht am besten mit ausreißer-robusten Verfahren, denn bei nichtrobusten Verfahren werden die Ausreißer oft versteckt, man sagt auch maskiert.

Um die Ausreißer in einem Regressionsmodel zu entdecken, trägt man die geschätzten absoluten Residuen $\|y_n - x(t_n)^\top \hat{\beta}(y, X_d)\|$ gegen die Beobachtungsnummer n auf. Bei ausreißer-robusten Regressionsschätzungen werden dann diese Residuen bei Ausreißern besonders groß. Das zeigt das folgende Beispiel.

6.4.1 Beispiel (Fortsetzung von Beispiel 6.2.5 und Beispiel 6.2.10)
Abbildung 6.4 zeigt die absoluten geschätzten Residuen der Kleinsten-Quadrat-Summen-Schätzung (LS), der LTS-Schätzung (LTS) und der Catline, die schon in den Beispielen 6.2.5 und 6.2.10 betrachtet wurden. Dabei bezeichnen die Beobachtungsnummern 1 bis 4 die Giganten. Man sieht deutlich, dass bei der LTS-Schätzung und der Catline die Residuen an diesen Beobachtungsnummern sehr groß sind, während bei der Kleinsten-Quadrat-Summen-Schätzung dort nichts besonderes festzustellen ist.

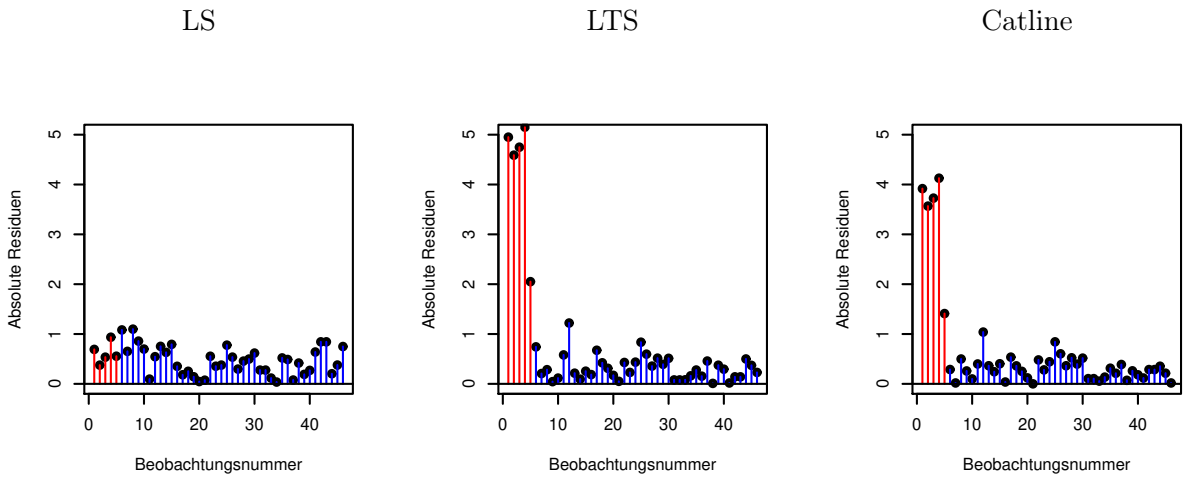


Abbildung 6.4: Absolute geschätzte Residuen der Kleinsten-Quadrat-Summen-Schätzung (LS), der LTS-Schätzung (LTS) und der Catline

Kapitel 7

Robuste Tests für Regressionsmodelle

Dieses Kapitel stellt ganz aktuelle Forschungsergebnisse der Arbeitsgruppe Müller vor.

7.1 Tests bei univariaten Regressoren

Wir betrachten hier ein allgemeines Regressions-Modell der Form

$$y_n = g(x_n, \beta) + \mathrm{e}_n, \quad \text{für } n \in \{1, \ldots, N\},$$

wobei $\beta \in \mathbb{R}^R$ der unbekannte Parametervektor ist, g eine bekannte Regressionsfunktion ist, die auch nichtlinear sein kann, und $z_n = (y_n, x_n)^\top \in \mathbb{R}^2$, $n \in \{1, \ldots, N\}$, die Datenpunkte sind. Die Fehler $\mathrm{e}_1, \ldots, \mathrm{e}_N$ sind Realisierungen von unabhängigen und identisch verteilten Zufallsgrößen $E_1, \ldots, E_N$, so dass $y_1, \ldots, y_N$ und $z_1, \ldots, z_N$ Realisierungen von Zufallsgrößen $Y_1, \ldots, Y_N$ bzw. $Z_1, \ldots, Z_N$ sind. Um autoregressive Modelle mitbetrachten zu, nehmen wir an dass die Regressoren $x_1, \ldots, x_N \in \mathbb{R}$ Realisierungen von Zufallsgrößen $X_1, \ldots, X_N$ sind, für die

$$X_1 < X_2 < \ldots < X_N \tag{7.1}$$

fast sicher gilt. Insbesondere haben wir $X_n = Y_{n-1}$ für einen autoregressiven Prozess, so dass die Bedingung (7.1) bedeutet, dass der autoregressive Prozess streng monoton wachsend ist, d.h. ein Wachstumsprozess ist. Auch wenn in der klassischen Regression die Regressoren $x_1, \ldots, x_N$ fest vorgegeben sind, interpretieren wir sie hier als Realisierungen von $X_1, \ldots, X_N$.

Als Verallgemeinerung der Halbraum-Tiefe und der Regressions-Tiefe führte Mizera (2002) die Globale Tiefe und die Tangent-Tiefe eine. Diese basieren auf einer allgemeinen Qualitätsfunktion. Eine naheliegende Qualitätsfunktion ist einfach das quadrierte Residuum, so dass hier nur quadrierte Residuen für die Globale Tiefe und Tangent-Tiefe betrachtet werden. Dazu seien $\mathrm{res}(z_n, \beta) := y_n - g(x_n, \beta)$, $n \in \{1, \ldots, N\}$, die Residuen.

7.1.1 Definition (Nonfit)
Ein Parameter $\beta \in \mathbb{R}^R$ ist ein Nonfit bezüglich $z_1, \ldots, z_N$, falls es ein $\tilde{\beta}$ gibt mit $\mathrm{res}(z_n, \tilde{\beta})^2 < \mathrm{res}(z_n, \beta)^2$ für alle $n = 1, \ldots, N$.

7.1.2 Definition (Globale Tiefe)
Die Globale Tiefe $d_G(\beta, z)$ von β bezüglich $z = (z_1, \ldots, z_N)^\top$ ist definiert als

$$d_G(\beta, z) = \frac{1}{N} \min \{M; \quad \text{es existieren } n_1, \ldots, n_M \in \{1, \ldots, N\}, \text{ so dass } \beta \text{ ein Nonfit bezüglich } z_n \text{ mit } n \in \{1, \ldots, N\} \setminus \{n_1, \ldots, n_M\} \text{ ist } \}.$$

7.1.3 Definition (Tangent-Tiefe)
Die Tangent-Tiefe von β in $z = (z_1, \ldots, z_N)^\top$ ist definiert als

$$d_T(\beta, z) := \frac{1}{N} \min_{u \in \mathbb{R}^R} \sharp \left\{ n \in \{1, \ldots, N\};\ u^\top \frac{\partial}{\partial \beta} \operatorname{res}(z_n, \beta)^2 \geq 0 \right\}.$$

Da u mit -1 multipliziert werden kann, gilt auch

$$d_T(\beta, z) = \frac{1}{N} \min_{u \in \mathbb{R}^R} \sharp \left\{ n \in \{1, \ldots, N\};\ u^\top \frac{\partial}{\partial \beta} \operatorname{res}(z_n, \beta)^2 \leq 0 \right\},$$

so dass dies auch zur Definition benutzt werden kann.

Da für $\operatorname{res}(z_n, \beta + \tilde{h}u)^2 < \operatorname{res}(z_n, \beta)^2$ aber auch $u^\top \frac{\partial}{\partial \beta} \operatorname{res}(z_n, \beta)^2 \geq 0$ gelten kann, sind im Allgemeinen die Globale Tiefe und die Tangent-Tiefe nicht identisch und es gilt nur

$$d_T(\beta, z) \geq d_G(\beta, z).$$

Wie die Simplex-Regressions-Tiefe kann auch eine Simplex-Tangent-Tiefe definiert werden.

7.1.4 Definition (Simplex-Tangent-Tiefe)
Die Simplex-Tangent-Tiefe $d_{ST}(\beta, z)$ von $\beta \in \mathbb{R}^R$ bezüglich $z = (z_1, \ldots, z_N)^\top$ ist definiert als

$$d_{ST}(\beta, z) = \frac{1}{\binom{N}{R+1}} \sum_{1 \leq n_1 < n_2 < \ldots < n_{R+1} \leq N} \mathbb{1}\{d_T(\beta, (z_{n_1}, \ldots, z_{n_{R+1}})^\top) > 0\}.$$

7.1.5 Beispiel
Für die lineare Regression mit $g(x_n, \beta) = \beta_0 + \beta_1 x_n$, $x_n \in \mathbb{R}$, $\beta = (\beta_0, \beta_1)^\top \in \mathbb{R}^2$ und $R = 2$ ist die Tangent-Tiefe die Regressions-Tiefe und Rousseeuw und Hubert (1999) hatten schon bemerkt, dass die Tangent-Tiefe von drei Beobachtungen nur dann größer Null ist, wenn die Residuen alternierende Vorzeichen besitzen. Abbildung 7.1 zeigt, dass bei einem Fit (linke Seite der Abbildung) die blau eingezeichneten Residuen alternierende Vorzeichen haben, während bei einem Nonfit (rechte Seite der Abbildung) die Residuen keine alternierenden Vorzeichen besitzen. In diesem Fall ergibt die rote Gerade Residuen, die alle kleiner als bei der schwarzen Gerade sind, weshalb die schwarze Gerade hier ein Nonfit ist.

Damit stellt sich insbesondere die Frage, wann $d_T(\beta, (z_1, \ldots, z_{R+1})^\top) > 0$ für $R + 1$ Beobachtungen gilt. Der folgende Satz zeigt, dass unter der Bedingung (7.1) und naheliegenden weiteren Bedingungen dies gilt, wenn die Residuen bei $z_1, \ldots, z_{R+1}$ alternierende Vorzeichen besitzen.

Abbildung 7.1: Fits und Nonfits bei der linearen Regression. Links: Fit der schwarzen Gerade. Rechts: Nonfit der schwarzen Gerade.

Um zu zeigen, dass nicht nur bei der linearen Regression sondern in viel allgemeineren Situationen die Tangent-Tiefe genau dann größer Null ist, wenn die Residuen alternierende Vorzeichen haben, nutzen wir aus, dass wegen der speziellen Gestalt der Residuen die Tangent-Tiefe auch wie folgt geschrieben werden kann:

$$d_T(\beta, (z_{n_1}, \ldots, z_{n_{R+1}})^\top) = \frac{1}{R+1} \min_{u \in \mathbb{R}^R} \sharp \left\{ n \in \{1, \ldots, R+1\};\ u^\top v(x_n, \beta)\, \mathrm{res}(z_n, \beta) \leq 0 \right\},$$

wobei

$$v(x_n, \beta) := \frac{\partial}{\partial \beta} g(x_n, \beta)$$

gilt.

7.1.6 Definition

a) Sei sign(y) das Vorzeichen von $y \in \mathbb{R}$, d.h. $sign(y) = 1$ falls $y > 0$, $sign(y) = -1$ falls $y < 0$, und $sign(y) = 0$ falls $y = 0$.
b) Ein Vektor $s = (s_1, \ldots, s_{R+1})^\top \in \mathbb{R}^{R+1}$ hat alternierende Vorzeichen, falls $sign(s_r) = -sign(s_{r+1}) \neq 0$ für alle $k \in \{1, \ldots, R\}$ erfüllt ist. Hat s alternierende Vorzeichen, dann hat es R Vorzeichenwechsel.
c) Eine Funktion $f : [a, b] \to \mathbb{R}$ hat R Vorzeichenwechsel im Interval $[a, b] \subset \mathbb{R}$, falls $x_1 < x_2 < \ldots < x_{R+1}$ existieren mit $x_r \in [a, b]$ für $k \in \{1, \ldots, R+1\}$ und $sign(f(x_r)) = -sgn(f(x_{r+1})) \neq 0$ für $k \in \{1, \ldots, R\}$ und es keine $x_1 < x_2 < \ldots < x_{L+1}$ für $L > R$ mit $x_l \in [a, b]$ für $l \in \{1, \ldots, L+1\}$ gibt, so dass $sign(f(x_l)) = -sgn(f(x_{l+1})) \neq 0$ für $l \in \{1, \ldots, L\}$ gilt.

7.1.7 Satz (Kustosz, Müller, Wendler 2016)
Sei $x_1 < x_2 < \ldots < x_{R+1} \in \mathbb{R}$ und nehme die folgende Bedingungen für $w_u : [x_1, x_{R+1}] \to \mathbb{R}$ gegeben durch $w_u(x) = u^\top v(x, \beta)$ an:

A) w_u hat höchstens $R-1$ Vorzeichenwechsel in $[x_1, x_{R+1}]$ für alle $u \in \mathbb{R}^R$.

B) Für jeden Vektor $s \in \{-1, 1\}^{R+1}$ mit höchstens $R-1$ Vorzeichenwechsel existiert ein $u_0 \in \mathbb{R}^R$ mit $sign(w_{u_0}(x_n)) = s_n$ für $n \in \{1, \ldots, R+1\}$.

Dann gilt $d_T(\beta, (z_1, \ldots, z_{R+1})^\top) > 0$ genau dann, wenn $(res(z_1, \beta), \ldots, res(z_{R+1}, \beta))^\top$ alternierende Vorzeichen hat oder mindestens eins der Residuen Null ist.

Beweis.
Es ist klar, dass $d_T(\beta, (z_1, \ldots, z_{R+1})^\top) > 0$ gilt, falls $\text{res}(z_n, \beta) = 0$ für ein $n \in \{1, \ldots, R+1\}$ erfüllt ist. Daher müssen wir nur den Fall betrachten, in dem $\text{res}(z_n, \beta) \neq 0$ für alle $n \in \{1, \ldots, R+1\}$ gilt. Beide Richtungen der Äquivalenz werden per Widerspruchsbeweis gezeigt bzw. es wird die Äquivalenz der gegenteiligen Eigenschaften gezeigt.

Dazu nehme als erstes an, dass $(\text{res}(z_1, \beta), \ldots, \text{res}(z_{R+1}, \beta))^\top$ keine alternierenden Vorzeichen besitzt. Das bedeutet, dass es ein $k \in \{1, \ldots, R\}$ mit $\text{sign}(\text{res}(z_r, \beta)) = \text{sign}(\text{res}(z_{r+1}, \beta))$ gibt. Setze $s_n = \text{sign}(\text{res}(z_n, \beta))$ für $n \in \{1, \ldots, R+1\}$ und $s = (s_1, \ldots, s_{R+1})^\top$. Dann gilt $s \in \{-1, 1\}^{R+1}$ und s hat höchstens $R-1$ Vorzeichenwechsel. Gemäß Bedingung B) existiert dann ein $u_0 \in \mathbb{R}^R$ mit $\text{sign}(w_{u_0}(x_n)) = s_n$ für $n \in \{1, \ldots, R+1\}$. Aber dies impliziert

$$\text{sign}(w_{u_0}(x_n)) \,\text{sign}(\text{res}(z_n, \beta)) = 1 \;\text{ for }\; n \in \{1, \ldots, R+1\}$$

and somit

$$u_0^\top v(x_n, \beta)\, \text{res}(z_n, \beta) = w_{u_0}(x_n)\, \text{res}(z_n, \beta) > 0 \;\text{ for }\; n \in \{1, \ldots, R+1\},$$

so dass $d_T(\beta, (z_1, \ldots, z_{R+1})^\top) = 0$ gilt.

Umgekehrt nehme $d_T(\beta, (z_1, \ldots, z_{R+1})^\top) = 0$ an. Dann existiert $u \in \mathbb{R}^R$ mit

$$u^\top v(x_n, \beta)\, \text{res}(z_n, \beta) = w_u(x_n)\, \text{res}(z_n, \beta) > 0 \;\text{ for }\; n \in \{1, \ldots, R+1\}. \tag{7.2}$$

Da w_u höchstens $R-1$ Vorzeichenwechsel auf $[x_1, x_{R+1}]$ gemäß Bedingung A) hat, existiert ein $k \in \{1, \ldots, R\}$ mit

$$\text{sign}(w_u(x_r)) = \text{sign}(w_u(x_{r+1})).$$

Das bedeutet mit (7.2), dass $\text{sign}(\text{res}(z_r, \beta)) = \text{sign}(\text{res}(z_{r+1}, \beta))$ gilt, so dass $(\text{res}(z_1, \beta), \ldots, \text{res}(z_{R+1}, \beta))^\top$ keine alternierenden Vorzeichen besitzt. □

In Kustosz, Müller, Wendler (2015) werden zahlreiche Beispiele gegeben, in denen die Bedingungen von Satz 7.1.7 erfüllt sind. Dies gilt z.B. für die polynomiale Regression beliebiger Ordnung, zahlreiche nichtlineare Wachstumsmodelle, das Michaelis-Menten-Modell sowie entsprechende

autoregressive Modelle und insbesondere für lineare und nichtlineare AR(1)-Modelle. In all diesen Fällen reduziert sich die Simplex-Tangent-Tiefe zu

$$d_{ST}(\beta,z)=\frac{1}{\binom{N}{R+1}}\sum_{1\le n_1<n_2<\ldots<n_{R+1}\le N}\left(\prod_{r=1}^{R+1}\mathbb{1}\left\{\mathrm{res}(z_{n_r},\beta)(-1)^r>0\right\}\right. \tag{7.3}$$
$$\left.+\prod_{r=1}^{R+1}\mathbb{1}\left\{\mathrm{res}(z_{n_r},\beta)(-1)^{r+1}>0\right\}+\left(1-\prod_{r=1}^{R+1}\mathbb{1}\left\{\mathrm{res}(z_{n_r},\beta)\neq 0\right\}\right)\right).$$

Da in der Regel eine stetige Verteilung der Fehler E_n angenommen wird, gilt $\mathrm{res}(z_n,\beta)\neq 0$ fast sicher, so dass statt (7.3) auch

$$\tilde{d}_{ST}(\beta,z)=\frac{1}{\binom{N}{R+1}}\sum_{1\le n_1<n_2<\ldots<n_{R+1}\le N}\left(\prod_{r=1}^{R+1}\mathbb{1}\left\{\mathrm{res}(z_{n_r},\beta)(-1)^r>0\right\}\right. \tag{7.4}$$
$$\left.+\prod_{r=1}^{R+1}\mathbb{1}\left\{\mathrm{res}(z_{n_r},\beta)(-1)^{r+1}>0\right\}\right)$$

benutzt werden kann.

Die Simplex-Tangent-Tiefe ist in ihrer ursprünglichen Form eine U-Statistik, so dass im Prinzip die asymptotische Verteilung nach einem Satz von Hoeffding (siehe z.B. Lee 1990, S. 76, oder Witting und Müller-Funk 1995, p. 635) für eine normierte Version $T_N(\beta,z)$ der Simplex-Tangent-Tiefe $d_{ST}(\beta,z)$ bekannt ist. Die normierte Version ist durch

$$T_N(\beta,z):=N\left(\tilde{d}_{ST}(\beta,z)-\left(\frac{1}{2}\right)^R\right)$$

gegeben. Die asymptotische Verteilung von $T_N(\beta,Z)$ hängt dann nicht von β ab. Kennt man diese asymptotische Verteilung, so können beliebige Hypothesen mit folgender Testvorschrift getestet werden:

$$\text{Lehne } H_0:\beta\in\beta_0 \text{ ab, falls } \sup_{\beta\in\mathcal{B}_0}T_N(\beta,z)<q_\alpha, \tag{7.5}$$

wobei q_α das α-Quantil der asymptotischen Verteilung von $T_N(\beta,\cdot)$ ist. Da $T_N(\beta,z)$ monoton von $d_{ST}(\beta,z)$ abhängt, bedeutet dies, dass $H_0:\beta\in\mathcal{B}_0$ abgelehnt wird, wenn die Tiefe aller Parameter β in der Hypothesenmenge $\mathcal{B}_0$ im Datensatz z zu klein ist.

7.1.8 Satz
Die durch (7.5) gegebene Entscheidungsregel ist ein asymptotischer Test zum Niveau α für $H_0:\beta\in\mathcal{B}_0$ gegen $H_1:\theta\notin\mathcal{B}_0$.

Beweis. Es gilt

$$P_{\beta_0}\left(\sup_{\beta\in\mathcal{B}_0}T_N(\beta,Z)<q_\alpha\right)\le P_{\beta_0}\left(T_N(\beta_0,Z)<q_\alpha\right)\overset{N\to\infty}{\longrightarrow}\alpha.$$

für alle $\beta_0 \in \mathcal{B}_0$. □

Allerdings ist die Simplex-Tangent-Tiefe in den meisten Fällen eine entartete U-Statistik, so dass eine Spekraldarstellung der bedingten Erwartungswerte gefunden werden muss. Diese Spektraldarstellungen wurden in Müller (2005) für die lineare und quadratische Regression, in Wellmann, Harmmand und Müller (2009) für die polynomiale Regression beliebiger Ordnung, in Wellmann und Müller (2010a) für multiple Regression und in Wellmann und Müller (2010b) für orthogonale Regression bestimmt. In diesen Arbeiten wird nicht die Annahme (7.1) gemacht, sondern nur angenommen, dass die Regressoren durch eine Verteilung gegeben sind.

Die vereinfachten Formen (7.3) und (7.4) basierend auf alternierenden Vorzeichen der Residuen bilden für $R > 1$ keine U-Statistik mehr, weil die Reihenfolge der $z_{n_1}, \ldots, z_{n_{R+1}}$ wichtig ist und somit keine symmetrische Kernfunktion vorliegt. Aber mit Methoden, die denen für U-Statistiken ähneln, konnte in Kustosz, Leucht und Müller (2016) auch die asymptotische Verteilung der Form (7.4) für $R = 2$ hergeleitet werden. Diese ist durch die Verteilung eines integrierten zweidimensionalen Gauss-Prozess gegeben. Im Fall $R = 1$ ist die asymptotische Verteilung durch eine Zufallsvariable gegeben, die eine χ^2-Verteilung mit einem Freiheitsgrad besitzt, und wurde schon in Müller (2005) für die lineare Regression durch den Ursprung bestimmt, siehe auch Kustosz und Müller (2014) für den Fall eines AR(1)-Modells.

Statt der "vollen" Simplex-Tangent-Tiefe können auch Vereinfachungen von (7.4) betrachtet werden, von denen es einfacher ist, die asymptotische Verteilung herzuleiten.

7.1.9 Definition (Vereinfachte Simplex-Tangent-Tiefen (Simplified Simplicial Depth), Kustosz, Müller, Wendler 2016)

$$d_S^1(\beta, z) := \frac{1}{\left\lfloor \frac{N}{R+1} \right\rfloor} \sum_{n=1}^{\left\lfloor \frac{N}{R+1} \right\rfloor} \left(\prod_{r=1}^{R+1} \mathbb{1}\left\{ res(z_{(R+1)(n-1)+r}, \beta)(-1)^r > 0 \right\} + \prod_{r=1}^{R+1} \mathbb{1}\left\{ res(z_{(R+1)(n-1)+r}, \beta)(-1)^{r+1} > 0 \right\} \right), \tag{7.6}$$

$$d_S^2(\beta, z) := \frac{1}{N-R} \sum_{n=1}^{N-R} \left(\prod_{r=1}^{R+1} \mathbb{1}\left\{ res(z_{n-1+r}, \beta)(-1)^r > 0 \right\} + \prod_{r=1}^{R+1} \mathbb{1}\left\{ res(z_{n-1+r}, \beta)(-1)^{r+1} > 0 \right\} \right), \tag{7.7}$$

sowie für $R = 2$

$$\begin{aligned} & d_S^3(\beta, z) \qquad (7.8) \\ := \quad & \frac{1}{\left\lfloor \frac{N-1}{2} \right\rfloor} \sum_{n=1}^{\left\lfloor \frac{N-1}{2} \right\rfloor} \Bigg(\mathbb{1}\{res(z_n, \beta) > 0\}\ \mathbb{1}\left\{ res(z_{\left\lfloor \frac{N+1}{2} \right\rfloor}, \beta) < 0 \right\}\ \mathbb{1}\{res(z_{N-n+1}, \beta) > 0\} \\ & + \mathbb{1}\{res(z_n, \beta) < 0\}\ \mathbb{1}\left\{ res(z_{\left\lfloor \frac{N+1}{2} \right\rfloor}, \beta) > 0 \right\}\ \mathbb{1}\{res(z_{N-n+1}, \beta) < 0\} \Bigg). \end{aligned}$$

Die Tiefe $d_S^1(\beta, z)$ benutzt nur nichtüberlappende Teilmengen von aufeinanderfolgenden Residuen, während die Teilmengen in $d_S^2(\beta, z)$ überlappend sind.

7.1.10 Satz (Kustosz, Müller, Wendler 2016)
Ist β der zugrunde liegende Parameter und gilt $P_\beta(res(Z_n, \beta) > 0) = P_\beta(res(Z_n, \beta) < 0) = \frac{1}{2}$ für alle $n \in \{1, \dots, N\}$, dann gilt

$$a) \qquad T_N^1(\beta, Z) := \sqrt{\left\lfloor \frac{N}{R+1} \right\rfloor} \, \frac{d_S^1(\beta, Z) - \left(\frac{1}{2}\right)^R}{\sqrt{\left(\frac{1}{2}\right)^R \left(1 - \left(\frac{1}{2}\right)^R\right)}} \longrightarrow \mathcal{N}(0,1),$$

$$b) \qquad T_N^2(\beta, Z) := \sqrt{N-R} \, \frac{d_S^2(\beta, Z) - \left(\frac{1}{2}\right)^R}{\sqrt{(\frac{1}{2})^R \cdot [3 - (\frac{1}{2})^{R-1} \cdot K - 3 \cdot (\frac{1}{2})^R]}} \longrightarrow \mathcal{N}(0,1),$$

$$c) \qquad T_N^3(\beta, Z) := \sqrt{\left\lfloor \frac{N-1}{2} \right\rfloor} \, \frac{d_S^3(\beta, Z) - \frac{1}{4}}{\sqrt{\frac{3}{16}}} \longrightarrow \mathcal{N}(0,1),$$

in Verteilung für $N \to \infty$.

Im Beweis von Satz 7.1.10 b) wird ein Grenzwertsatz für m-abhängige Zufallsvariablen genutzt.

7.1.11 Definition
Eine Folge von Zufallsvariablen $X_1, X_2, \dots$ heißt m-abhängig für $m \geq 0$, falls $(X_i, \dots, X_{i+n})$ unabhängig von $(X_{i+n+j}, \dots, X_{i+n+j+l})$ für alle $j > m$ und $i, n, l \in \mathbb{N}$ ist.

7.1.12 Satz (Hoeffding, Robbins 1948, S. 774)
Seien $X_1, X_2, \dots$ m-abhängig mit
(a) $E(X_i) = 0$, $E(|X_i|^3) \leq R < \infty$ für alle $i \in \mathbb{N}$,
(b) $\lim_{p\to\infty} \frac{1}{p} \sum_{h=1}^{p} A_{i+h} = A$ gleichförmig für alle $i \in \mathbb{N}_0$,
wobei $A_i = E(X_{i+m}^2) + 2 \sum_{j=1}^{m} E(X_{i+m-j} \, X_{i+m})$ für $i \in \mathbb{N}$ gilt.
Dann gilt

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} X_n \longrightarrow \mathcal{N}(0, A).$$

Beweis von Satz 7.1.10.
Beachte als erstes, dass $res(\beta, Z_n) = E_n$ gilt, falls β der zugrundeliegende Parameter ist.
a) Setze

$$V_n := \prod_{r=1}^{R+1} \mathbb{1}\left\{\text{res}(Z_{(R+1)(n-1)+r}, \beta)(-1)^r > 0\right\} + \prod_{r=1}^{R+1} \mathbb{1}\left\{\text{res}(Z_{(R+1)(n-1)+r}, \beta)(-1)^{r+1} > 0\right\}.$$

Dann sind V_n, $n \in \{1, \ldots, \left\lfloor \frac{N}{R+1} \right\rfloor\}$ unabhängige Zufallsgrößen mit Bernoulli-Verteilung mit $P(V_n = 1) = (1/2)^R$, so dass die Aussage aus dem Zentralen Grenzwertsatz folgt.
b) Setze

$$V_n := \prod_{r=1}^{R+1} \mathbb{1}\left\{\text{res}(Z_{n-1+r}, \beta)(-1)^r > 0\right\} + \prod_{r=1}^{R+1} \mathbb{1}\left\{\text{res}(Z_{n-1+r}, \beta)(-1)^{r+1} > 0\right\}.$$

Dann sind $V_n, n \in \{1, ..., N-R\}$ auch Bernoulli-verteilte Zufallsgrößen mit $P(V_n = 1) = (1/2)^R$. Durch Zentrierung mittels $X_n = V_n - (\frac{1}{2})^R$ erhält man stationäre Zufallsgrößen mit $E[X_n] = 0$ und $E[|X_n|^3] < \infty$. Damit kann der Grenzwertsatz von Hoeffding und Robbins (1948) für m-abhängige Zufallsvariablen angewendet werden, da V_n und V_m nur abhängig sind, wenn die zugehörigen Index-Mengen sich überlappen. Das impliziert nämlich, dass $X_1, X_2, ...$ R-abhängig sind. Um die Varianz in der Grenzverteilung zu bestimmen, müssen wir $E(X_1 X_d)$ für $d \in \{1, ..., R+1\}$ berechnen, um

$$A = E[X_1^2] + \sum_{d=2}^{R+1} 2 \cdot E[X_1 X_d]$$

zu erhalten. Für $d > R+1$ sind aber diese Erwartungswerte Null, da die Zufallsgrößen unabhängig und zentriert sind.
Für $d \in \{1, ..., R+1\}$ gilt

$$\begin{aligned}
E[X_1 X_d] &= E\left[\left(V_1 - \left(\frac{1}{2}\right)^R\right)\left(V_d - \left(\frac{1}{2}\right)^R\right)\right] \\
&= E[V_1 V_d] - \left(\frac{1}{2}\right)^R E[V_d] - \left(\frac{1}{2}\right)^R E[V_1] + \left(\frac{1}{2}\right)^{2R} \\
&= E[V_1 V_d] - \left(\frac{1}{2}\right)^R \cdot \left(\frac{1}{2}\right)^R - \left(\frac{1}{2}\right)^R \cdot \left(\frac{1}{2}\right)^R + \left(\frac{1}{2}\right)^{2R} \\
&= E[V_1 V_d] - \left(\frac{1}{2}\right)^{2R} \\
&= \left(\frac{1}{2}\right)^{R+d-1} - \left(\frac{1}{2}\right)^{2 \cdot R}.
\end{aligned}$$

Durch Einsetzen dieser expliziten Ausdrücke für die Erwartungswerte erhalten wir

$$\begin{aligned}
A &= \sum_{d=2}^{R+1} 2 \cdot \left[\left(\frac{1}{2}\right)^{R+d-1} - \left(\frac{1}{2}\right)^{2R} \right] + \left(\frac{1}{2}\right)^{R} \left(1 - \left(\frac{1}{2}\right)^{R}\right) \\
&= \sum_{d=2}^{R+1} \left(\frac{1}{2}\right)^{R+d-2} - R\left(\frac{1}{2}\right)^{2R-1} + \left(\frac{1}{2}\right)^{R} - \left(\frac{1}{2}\right)^{2R} \\
&= \left(\frac{1}{2}\right)^{R} \left[\sum_{d=2}^{R+1} \left(\frac{1}{2}\right)^{d-2} - R\left(\frac{1}{2}\right)^{R-1} + 1 - \left(\frac{1}{2}\right)^{R} \right] \\
&= \left(\frac{1}{2}\right)^{R} \left[\sum_{d=0}^{R-1} \left(\frac{1}{2}\right)^{d} - R\left(\frac{1}{2}\right)^{R-1} + 1 - \left(\frac{1}{2}\right)^{R} \right] \\
&= \left(\frac{1}{2}\right)^{R} \left[2 - \left(\frac{1}{2}\right)^{R-1} - R\left(\frac{1}{2}\right)^{R-1} + 1 - \left(\frac{1}{2}\right)^{R} \right] \\
&= \left(\frac{1}{2}\right)^{R} \left[3 - R\left(\frac{1}{2}\right)^{R-1} - \left(\frac{1}{2}\right)^{R-1} \left(1 + \frac{1}{2}\right) \right] \\
&= \left(\frac{1}{2}\right)^{R} \left[3 - R\left(\frac{1}{2}\right)^{R-1} - \left(\frac{1}{2}\right)^{R-1} \left(\frac{3}{2}\right) \right] \\
&= \left(\frac{1}{2}\right)^{R} \left[3 - \left(\frac{1}{2}\right)^{R-1} \cdot R - 3 \cdot \left(\frac{1}{2}\right)^{R} \right].
\end{aligned}$$

c) Hier ist

$$\begin{aligned}
V_n = &\mathbb{1}\{\mathrm{res}(Z_n, \beta) > 0\} \, \mathbb{1}\left\{\mathrm{res}(Z_{\lfloor \frac{N+1}{2} \rfloor}, \beta) < 0\right\} \, \mathbb{1}\{\mathrm{res}(Z_{N-n+1}, \beta) > 0\} \\
&+ \mathbb{1}\{\mathrm{res}(Z_n, \beta) < 0\} \, \mathbb{1}\left\{\mathrm{res}(Z_{\lfloor \frac{N+1}{2} \rfloor}, \beta) > 0\right\} \, \mathbb{1}\{\mathrm{res}(Z_{N-n+1}, \beta) < 0\}.
\end{aligned}$$

Wieder sind die V_n Bernoulli-Variabeln, hier aber mit $P(V_n = 1) = 1/4$. Um den zentralen Grenzwertsatz anwenden zu können, müssen wir die Unabhängigkeit von $V_1, ..., V_{\lfloor \frac{N-1}{2} \rfloor}$ zeigen. Dazu beobachte als erstes, dass

$$\begin{aligned}
&P\left(V_n = 1 \mid E_{\lfloor \frac{N+1}{2} \rfloor} > 0\right) \\
=&P\left(E_n < 0, E_{N-n+1} < 0 \mid E_{\lfloor \frac{N+1}{2} \rfloor} > 0\right) \\
=&P\left(E_n < 0, E_{N-n+1} < 0\right) = \frac{1}{4} = P(V_n = 1)
\end{aligned}$$

gilt, da $E_1, ..., E_N$ unabhängig sind. Analog erhalten wir

$$P\left(V_n = 1 \mid E_{\lfloor \frac{N+1}{2} \rfloor} < 0\right) = \frac{1}{4} = P(V_n = 1)$$

und

$$P\left(V_n = 0 \mid E_{\lfloor \frac{N+1}{2} \rfloor} < 0\right) = P\left(V_n = 0 \mid E_{\lfloor \frac{N+1}{2} \rfloor} > 0\right) = \frac{3}{4} = P(V_n = 0).$$

Die Unabhängig der $E_1, ..., E_N$ impliziert ferner, dass V_n und V_m mit $n < m < \lfloor \frac{N+1}{2} \rfloor$ bedingt unabhängig sind, wenn $E_{\lfloor \frac{N+1}{2} \rfloor}$ gegeben ist, so dass

$$\begin{aligned}
&P(V_n = k, V_m = l)\\
=&P\left(V_n = k, V_m = l | E_{\lfloor \frac{N+1}{2} \rfloor} > 0\right) P\left(E_{\lfloor \frac{N+1}{2} \rfloor} > 0\right)\\
+&P\left(V_n = k, V_m = l | E_{\lfloor \frac{N+1}{2} \rfloor} < 0\right) P\left(E_{\lfloor \frac{N+1}{2} \rfloor} < 0\right)\\
=&P\left(V_n = k | E_{\lfloor \frac{N+1}{2} \rfloor} > 0\right) P\left(V_m = l | E_{\lfloor \frac{N+1}{2} \rfloor} > 0\right) \cdot \frac{1}{2}\\
+&P\left(V_n = k | E_{\lfloor \frac{N+1}{2} \rfloor} < 0\right) P\left(V_m = l | E_{\lfloor \frac{N+1}{2} \rfloor} < 0\right) \cdot \frac{1}{2}\\
=&P(V_n = k)P(V_m = l),
\end{aligned}$$

für $k, l \in \{0, 1\}$ folgt. Damit sind aber V_n und V_m unabhängig. Genauso erhalten wir die Unabhängigkeit von $V_1, ..., V_{\lfloor \frac{N-1}{2} \rfloor}$. □

7.1.13 Bemerkung
Auch wenn für $R > 1$ nur die Tiefen d_S^i, $i = 1, 2, 3$ mit gleichem R vereinfachte Simplex-Tiefen sind, kann immer auch die Tiefe benutzt werden, wo in (7.6) bzw. (7.7) einfach nur $R = 1$ benutzt wird. D.h. auch wenn der unbekannte Parameter nicht eindimensional ist, können nur Teilmengen mit zwei Elementen betrachtet werden. Die asymptotischen Verteilungen aus Satz 7.1.10 mit $R = 1$ sind dann weiterhin gültig. Ebenso kann immer (7.4) mit $R = 1$ benutzt werden. Hier ist aber die asymptotische Verteilung durch eine transformierte χ^2-Verteilung gegeben, die in Müller (2005) hergeleitet wurde.

Allerdings hat der Test basierend auf der vollen Simplex-Tiefe aus (7.4) mit $R = 1$ große Ähnlichkeiten mit dem wenig effizienten Vorzeichen-Test. Daher ist es wahrscheinlich sinnvoller eher in (7.4), (7.6), (7.7) bzw. (7.8) mit $R = 2$ zu arbeiten, unabhängig davon, welche Dimension der Parameter hat.

Dabei ist aber die Berechnung der vollen Tiefe mit $R = 2$ mittels (7.4) sehr aufwendig und sollte daher mit dem `R` Paket `rexpar` von Kustosz und Szugat gemacht werden. Dieses kann folgendermaßen installiert werden:

```
> install.packages("rexpar_1.1.zip", repos= NULL, type="binary")
Paket 'rexpar' erfolgreich ausgepackt und MD5 Summen abgeglichen
> install.packages("matrixcalc_1.0-3.zip", repos= NULL, type="binary")
Paket 'matrixcalc' erfolgreich ausgepackt und MD5 Summen abgeglichen
> library(matrixcalc)
> library(parallel)
> library(MASS)
> library(rexpar)
> library(help=rexpar)
```

Dieses Paket ist vor allem für autoregressive Modelle geschrieben worden. Es hat aber die Option, dass beliebige Residuen eingegeben werden können. Die volle Tiefe mit $R = 2$ mittels (7.4) für den Residuenvektor $(1, -1, 1, -1, 1)^\top$ erhält man dann wie folgt:

```
> dS_lin2(resy=c(1,-1,1,-1,1))
```

```
[1] 0.5
```

Die vereinfachte Tiefe mit $R = 2$ mittels (7.7), d.h. mit überlappenden Teilmengen, erhält man für den Residuenvektor $(-1, 1, -1, -1, -1, -1, -1, -1, -1, -1)^\top$ wie folgt:

```
> dS3_lin2(resy=c(-1,1,-1,-1,-1,-1,-1,-1,-1,-1))
[1] 0.125
```

Die α-Quantile q_α der asymptotischen Verteilung der vollen Tiefe mit $R = 2$ mittels (7.4) erhält man für $\alpha = 0.05,\ 0.01,\ 0.001$ wie folgt:

```
> SimQuants[round(SimQuants[, 1], digits = 3) == round((0.05), digits = 3), 2]
    qvals
-1.254541
> SimQuants[round(SimQuants[, 1], digits = 3) == round((0.01), digits = 3), 2]
    qvals
-2.240396
> SimQuants[round(SimQuants[, 1], digits = 3) == round((0.001), digits = 3), 2]
   qvals
-3.71403
```

7.2 Test bei multivariaten Regressoren

Wir betrachten hier wieder das allgemeines Regressions-Modell der Form

$$y_n = g(x_n, \beta) + \mathrm{e}_n, \quad \text{für } n \in \{1, \dots, N\},$$

wobei $\beta \in \mathbb{R}^R$ der unbekannte Parametervektor und g eine bekannte Regressionsfunktion ist. Hier nehmen wir aber an, dass die Regressoren q-dimensional sind, d.h. dass $x_1, \dots, x_N \in \mathbb{R}^q$ mit $q > 1$ gilt. In diesem Fall ist eine Ordnung der Regression in der Form (7.1) nicht möglich. Man kann aber wie in Bemerkung 7.1.13 vorgeschlagen wieder Zweier-Teilmengen auf verschiedene Vorzeichen der Residuen untersuchen. Dazu können alle Zweierteilmengen von $\{1, \dots, N\}$ betrachtet werden oder nur spezielle Zweiermengen. Spezielle Zweiermengen erhält man durch Triangulation der Punktemenge $x_1, \dots, x_N$.

Zum Beispiel mit dem Verfahren der Delaunay-Triangulation werden Punkte $x_1, \dots, x_N \in \mathbb{R}^2$ so zu Dreiecken vernetzt, dass innerhalb des Kreises, auf dem die drei Dreieckspunkte liegen (Umkreis des Dreiecks), keine anderen Punkte enthalten sind, siehe Abbildung 7.2. Wegen der Beziehung zu Voronoi-Diagrammen gibt es auch Triangulationen für $x_1, \dots, x_N \in \mathbb{R}^q$ mit $q > 2$. In diesen Fällen werden die Dreicke zu Simplicies verallgemeinert.

R-Pakete für die Delaunay-Triangulation
geometry `http://cran.r-project.org/web/packages/geometry/geometry.pdf`
tripack `http://cran.r-project.org/web/packages/tripack/tripack.pdf`

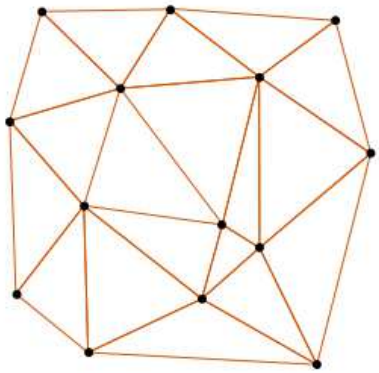


Abbildung 7.2: Delaunay-Triangulation.
Aus Wikipedia `http://de.wikipedia.org/wiki/Delaunay-Triangulation`

Kapitel 8

Nichtparametrische Regression

Kennt man den funktionalen Zusammenhang zwischen x und y bis auf wenige Parameter, so spricht man von parametrischer Regression und die Aufgabe ist, diese Parameter zu schätzen. Oft hat man aber so wenig Information über den Zusammenhang, dass eine parametrische Funktion nicht angesetzt werden kann. In solchen Situationen kann man versuchen, anhand der Daten eine erste Idee über den funktionalen Zusammenhang zu bekommen. Dazu gibt es mehrere Möglichkeiten.

Hier werden wir nur eine Methode behandeln, die Daten, typischerweise Zeitreihen, glättet. Das sind die sogenannten Kernschätzungen. Ein Spezialfall der Kernschätzungen ist das gleitende Mittel.

8.0.1 Definition (Gleitender Mittelwert (Moving Average))
Gegeben sei ein bivariater Datensatz $(x_1, y_1), \ldots, (x_N, y_N)$ *mit* $x_1 < x_2 < \ldots < x_N$ *und eine Bandweite* $B \in \mathbb{N}$. *Die durch den gleitenden Mittelwert geschätzten Funktionswerte* $\hat{f}(x_{B+1}), \ldots, \hat{f}(x_{N-B})$ *bei* $x_{B+1}, \ldots, x_{N-B}$ *sind gegeben durch*

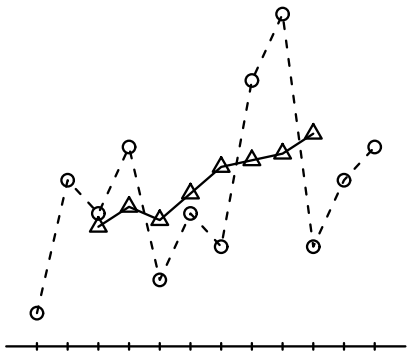
$$\hat{f}(x_n) = \frac{1}{2B+1} \sum_{m=n-B}^{n+B} y_m$$

für $n = B+1, \ldots, N-B$.

Der gleitende Mittelwert ist insbesondere sinnvoll, wenn die Regressoren $x_1, \ldots, x_N$ gleiche Abstände haben, wie das im folgenden Beispiel der Fall ist.

8.0.2 Beispiel
Die folgende Grafik zeigt eine Zeitreihe mit 12 Beobachtungen, die durch Kreise dargestellt sind. Wird als Bandweite $B = 2$ genommen, erhält man die durch Dreiecke gegebenen Punkte. Die durch die Dreiecke gegebene Kurve ist deutlich glatter als die ursprüngliche Kurve. Deshalb wird

der gleitende Mittelwert auch als **M**ittelwert-Glätter bezeichnet.



Indem man gewichtete Mittelwerte benutzt, kann man auch Funktionsschätzungen bestimmen, wenn die $x_1, \ldots, x_N$ keine gleichen Abstände haben. Die Gewichte werden durch eine **Kernfunktion** k gegeben und die Schätzungen heißen dann **Kernschätzungen**. Mit den Kernschätzungen können auch Funktionswerte an Stellen geschätzt werden, die zwischen den $x_1, \ldots, x_N$ liegen.

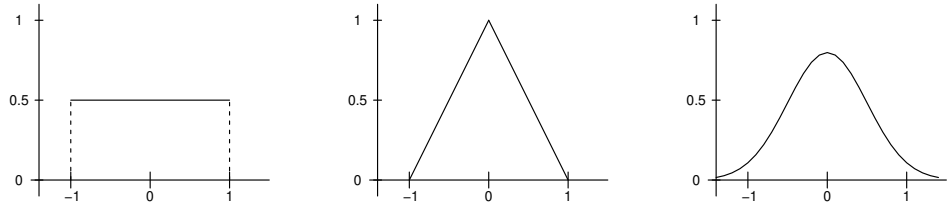
8.0.3 Definition (Kernschätzungen)
Gegeben sei ein bivariater Datensatz $(x_1, y_1), \ldots, (x_N, y_N)$ *mit* $x_1 < x_2 < \ldots < x_N$, *eine Bandweite* $b \in \mathbb{R}^+$ *und eine Kernfunktion* $k : \mathbb{R} \to \mathbb{R}$. *Dann heißt* $\hat{f}(x)$ *eine Kernschätzung bei* x *zur Bandweite* b *bzgl.* k, *wenn gilt*

$$\hat{f}(x) = \sum_{n=1}^{N} y_n \frac{k\left(\frac{x-x_n}{b}\right)}{\sum_{m=1}^{N} k\left(\frac{x-x_m}{b}\right)}.$$

Durch die Bandweite b wird bestimmt, wie glatt die geschätzte Funktion wird. Je größer b ist, desto glatter wird die Funktion. In der Regel ist die Kernfunktion k eine Funktion, die

$\int k(x)\,dx = 1$ erfüllt. Zum Beispiel können folgende Kernfunktionen benutzt werden:

Rechteckkern Dreieckskern Normalkern



8.0.4 Satz
Für die Kernschätzung $\hat{f}(x)$ bei x zur Bandweite b bzgl. k gilt

$$\hat{f}(x) = \arg\min_{\mu\in\mathbb{R}} \sum_{n=1}^{N} (y_n - \mu)^2\, k\left(\frac{x - x_n}{b}\right).$$

Beweis. Setze $\hat{\mu} = \hat{f}(x)$ und $a_n = \frac{k\left(\frac{x-x_n}{b}\right)}{\sum_{n=1}^{N} k\left(\frac{x-x_n}{b}\right)}$. Dann gilt

$$\sum_{n=1}^{N} a_n = 1$$

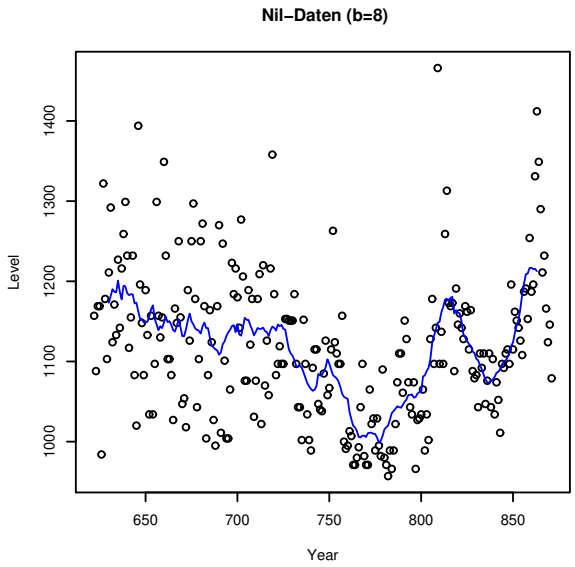
und somit für alle $\mu \in \mathbb{R}$

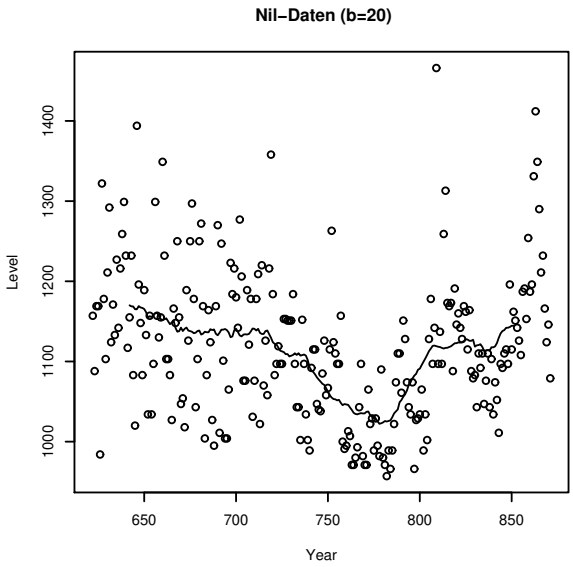
$$\begin{aligned}
\sum_{n=1}^{N} (y_n - \mu)^2\, a_n &= \sum_{n=1}^{N} (y_n - \hat{\mu} + \hat{\mu} - \mu)^2\, a_n \\
&= \sum_{n=1}^{N} (y_n - \hat{\mu})^2\, a_n + 2 \sum_{n=1}^{N} (y_n - \hat{\mu})(\hat{\mu} - \mu)\, a_n + \sum_{n=1}^{N} (\hat{\mu} - \mu)^2\, a_n \\
&= \sum_{n=1}^{N} (y_n - \hat{\mu})^2\, a_n + (\hat{\mu} - \mu)^2 \\
&\geq \sum_{n=1}^{N} (y_n - \hat{\mu})^2\, a_n.
\end{aligned}$$

Dabei gilt Gleichheit genau dann, wenn $\hat{f}(x) = \hat{\mu} = \mu$ gilt. Da bei der Minimierung es keine Rolle spielt, ob durch $\sum_{n=1}^{N} k\left(\frac{x-x_n}{b}\right)$ geteilt wird, folgt die Behauptung.□

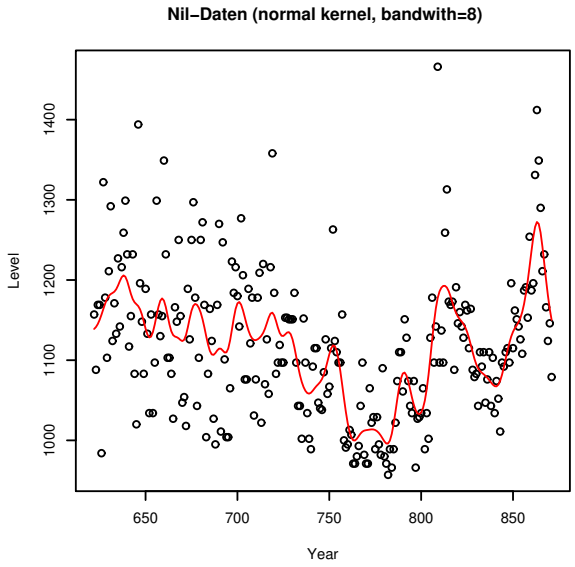
8.0.5 Beispiel
Die Wasserstände des Nils wurden über Jahrhunderte gemessen. Mit dem gleitenden Mittelwert

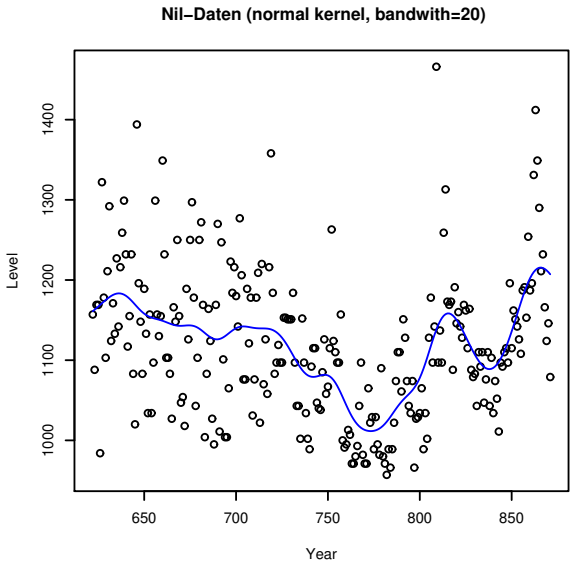
bzw. der Kernschätzung bezüglich des Rechteckkernes erhält man folgende geschätzte Funktionen:





Je größer die Bandweite b ist, desto glatter wird die geschätzte Funktion. Diese wird noch glatter, wenn der Normalkern benutzt wird:





In Anologie zu den Lokations-M-Schätzungen können **M-Kernschätzungen** definiert werden.

8.0.6 Definition (M-Kernschätzungen (Härdle, Gasser 1984))
Gegeben sei ein bivariater Datensatz $(x_1, y_1), \ldots, (x_N, y_N)$ *mit* $x_1 < x_2 < \ldots < x_N$, *eine Bandweite* $b \in \mathbb{R}^+$, *eine Kernfunktion* $k : \mathbb{R} \to \mathbb{R}$ *und eine Gewichtsfunktion* $\rho : \mathbb{R} \to \mathbb{R}^+$. *Dann heißt* $\hat{f}(x)$ *eine M-Kernschätzung bei* x *zur Bandweite* b *bzgl.* k *und* ρ, *wenn gilt*

$$\hat{f}(x) \in \arg\min_{\mu \in \mathbb{R}} \sum_{n=1}^{N} \rho(y_n - \mu) \; k\left(\frac{x - x_n}{b}\right).$$

Teil III

Asymptotische Robustheitskriterien

Kapitel 9

Einflussfunktion und asymptotische Bruchpunkte

9.1 Statistische Funktionale

Die in Definition 1.2.1 definierte Biasfunktion für einen Datensatz hat den Nachteil, dass sie im Gegensatz zum Bruchpunkt sehr von den Daten abhängig ist. Was man aber haben möchte, ist ein Maß für die Robustheit einer Schätzfunktion, das nicht von den Daten abhängt, da die ganze Funktion beurteilt werden soll. Um trotzdem den Einfluss einzelner Beobachtungen (Ausreißer) erfassen zu können, wird der asymptotische Wert einer Schätzfunktion benutzt. Dieser kann mittels eines statistischen Funktionals angegeben werden.

Wir wissen nach dem starken Gesetz der großen Zahlen, dass für jedes $z \in \mathbb{R}$ die empirische Verteilungsfunktion

$$F_{(y_1,y_2,\ldots,y_N)}(z) = \frac{1}{N}\sum_{n=1}^{N} 1_{(-\infty,z]}(y_n)$$

an der Stelle z für $\mathbb{P}$-fast alle $y_1, y_2, \ldots, y_N$ gegen die theoretische Verteilungsfunktion

$$F(z) = P((-\infty, z])$$

für $N \to \infty$ konvergiert, wenn P die zugrunde liegende Verteilung ist, d.h. wenn $y_1, y_2, \ldots, y_N$ Realisierungen von unabhängigen und identisch verteilten Zufallsgrößen $Y_1, Y_2, \ldots, Y_N$ sind mit $\mathbb{P}^{Y_n} = P$ für alle $n = 1, \ldots, N$. Diese Konvergenz ist nach dem Glivenko-Cantelli-Lemma (s. z.B. Billingsley (1995), Probability and Measure, S. 269) sogar gleichmäßig in z, d.h. es gilt

$$\mathbb{P}(\lim_{N\to\infty} \sup_{z\in\mathbb{R}} |F_{(Y_1,\ldots,Y_N)}(z) - F(z)| = 0) = 1.$$

Setze

$$y_{*N} = (y_1, y_2, \ldots, y_N) \text{ und } Y_{*N} = (Y_1, Y_2, \ldots, Y_N).$$

In der Regel können Schätzungen $\hat{\theta}_N(y_{*N})$ als Funktional-Werte $T(P_{y_{*N}})$ eines Funktionals $T : \mathcal{P} \to \mathbb{R}$ der empirischen Verteilung

$$P_{y_{*N}} = \frac{1}{N}\sum_{n=1}^{N} e_{y_n}$$

aufgefasst werden, wobei $\mathcal{P}$ der Raum aller Wahrscheinlichkeitsmaße auf $\mathbb{R}$ und e_a das Ein-Punkt-Maß bei a ist, d.h. es gilt $e_a(A) = 1_A(a)$ für jede (messbare) Menge A. Die Verteilungsfunktion der empirischen Verteilung $P_{y_{*N}}$ ist gerade die empirische Verteilungsfunktion $F_{y_{*N}}$. Oft überträgt sich die Konvergenz von $F_{Y_{*N}}$ gegen F bzw. $P_{Y_{*N}}$ gegen P dann auch auf die Funktional-Werte $T(P_{Y_{*N}})$, d.h. es gilt

$$T(P_{Y_{*N}}) \longrightarrow T(P) \;\; \mathbb{P}-\text{fast sicher (oder in Wahrscheinlichkeit) für } N \to \infty. \tag{9.1}$$

9.1.1 Beispiel (Arithmetisches Mittel, Erwartungswert)
Wir wissen nach dem starken Gesetz der großen Zahlen, dass das arithmetische Mittel fast sicher gegen den Erwartungswert konvergiert. Arithmetisches Mittel und Erwartungswert können mittels des Funktionals

$$T : \mathcal{P} \ni P \longrightarrow T(P) = \int x\, P(dx) \in \mathbb{R}$$

dargestellt werden. Dazu müssen wir erst definieren, was $T(P) = \int x\, P(dx)$ bedeutet.

9.1.2 Definition
Sei $h : \mathbb{R} \to \mathbb{R}$ eine Funktion von $\mathbb{R}$ nach $\mathbb{R}$.

a) Ist P ein diskretes Wahrscheinlichkeitsmaß mit diskreter Dichte p, d.h. $P = \sum_{k=1}^{\infty} p(x_k)\, e_{x_k}$, so ist

$$\int h(x)\, P(dx) := \sum_{k=1}^{\infty} h(x_k)\, p(x_k).$$

b) Ist P ein (absolut) stetiges Wahrscheinlichkeitsmaß mit Dichte f, so ist

$$\int h(x)\, P(dx) := \int h(x)\, f(x)\, dx.$$

c) Ist P eine Konvexkombination aus einem diskreten Wahrscheinlichkeitsmaß P_D mit diskreter Dichte p und einem (absolut) stetigem Wahrscheinlichkeitsmaß P_S mit Dichte f, d.h. $P = \alpha\, P_D + (1-\alpha)\, P_S$ mit $\alpha \in [0,1]$, so ist

$$\begin{aligned} \int h(x)\, P(dx) &:= \alpha \int h(x)\, P_D(dx) + (1-\alpha) \int h(x)\, P_S(dx) \\ &= \alpha \sum_{k=1}^{\infty} h(x_k)\, p(x_k) + (1-\alpha) \int h(x)\, f(x)\, dx. \end{aligned}$$

9.1.3 Beispiel (Arithmetisches Mittel, Erwartungswert, Fortsetzung von Beispiel 9.1.1)
Da die empirische Verteilung eine diskrete Verteilung mit $P_{y_{*N}} = \sum_{n=1}^{N} \frac{1}{N}\, e_{y_n}$ ist, gilt

$$\int x\, P_{y_{*N}}(dx) := \sum_{n=1}^{N} y_n \frac{1}{N} = \overline{y_{*N}},$$

d.h. wir erhalten das arithmetische Mittel. Ist Y_0 ein Zufallsgröße mit Verteilung $\mathbb{P}^{Y_0} = P = \mathbb{P}^{Y_n}$ und ist P ein diskretes Wahrscheinlichkeitsmaß mit diskreter Dichte p, so gilt nach der Definition

des Erwartungswertes

$$E_P(Y_0) = \sum_k x_k \, \mathbb{P}(Y_0 = x_k) = \sum_{k=1}^{\infty} x_k \, \mathbb{P}^{Y_0}(\{x_k\}) = \sum_{k=1}^{\infty} x_k \, P(\{x_k\}) = \sum_{k=1}^{\infty} x_k \, p(x_k) = \int x \, P(dx).$$

Ist P ein (absolut) stetiges Wahrscheinlichkeitsmaß mit Dichte f, so gilt nach der Definition des Erwartungswertes

$$E_P(Y_0) = \int x \, f(x) \, dx = \int x \, P(dx).$$

Ebenso kann die Verteilung von Y_0 auch durch eine Konvexkombination eines diskreten und eines stetigen Wahrscheinlichkeitsmaßes gegeben sein, d.h. $\mathbb{P}^{Y_0} = P = \alpha \, P_D + (1-\alpha) \, P_S$ mit $\alpha \in [0,1]$. In diesem Fall ist der Erwartungswert die Konvexkombination der Erwartungswerte für den diskreten und den stetigen Anteil (siehe Maßtheorie bzw. Statistik V), d.h.

$$E_P(Y_0) := \alpha \int x \, P_D(dx) + (1-\alpha) \int x \, P_S(dx) = \int x \, P(dx).$$

Das starke Gesetz der großen Zahlen liefert in allen Fällen dann

$$T(P_{Y_{*N}}) = \overline{Y_{*N}} \longrightarrow E_P(Y_0) = T(P) \;\; \mathbb{P} - \text{fast sicher für } N \to \infty.$$

9.1.4 Beispiel (p-Quantile)
Mit

$$T(P) = F^{-1}(p) := \inf\{z \in \mathbb{R};\; F(z) \geq p\} = \inf\{z \in \mathbb{R};\; P((-\infty, z]) \geq p\}$$

wird in der Regel das p-Quantil für Verteilungen definiert. Angewendet auf die empirische Verteilung $P_{y_{*N}}$ liefert es

$$T(P_{y_{*N}}) = \inf\{z \in \mathbb{R};\; F_{y_{*N}}(z) \geq p\} \in \tilde{y}_p$$

und somit eine eindeutig Version des empirischen p-Quantils. Der folgende Satz 9.1.6 zeigt, dass das p-Quantil auch die Bedingung (9.1) erfüllt, falls $F^{-1}(p)$ eindeutig ist. Insbesondere gilt es für $p = \frac{1}{2}$, also dem Median. Um Satz 9.1.6 zeigen zu können, wird folgendes Lemma aus Statistik II benötigt.

9.1.5 Lemma (Charakterisierung der Fast-sicheren-Konvergenz)
X_N konvergiert fast sicher gegen X

$$\Leftrightarrow \bigwedge_{\varepsilon > 0} \lim_{M \to \infty} \mathbb{P}\left(\bigcap_{N \geq M} \{|X_N - X| < \varepsilon\} \right) = 1.$$

$$\Leftrightarrow \bigwedge_{\varepsilon > 0} \lim_{M \to \infty} \mathbb{P}\left(\bigcup_{N \geq M} \{|X_N - X| \geq \varepsilon\} \right) = 0.$$

9.1.6 Satz (Starkes Gesetz der großen Zahlen für p-Quantile)
Sei $(Y_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger Zufallsgrößen mit Verteilungsfunktion $F_{Y_n} = F$ für $n \in \mathbb{N}$, $p \in (0,1)$ und $F(F^{-1}(p)+\varepsilon) > p$ für alle $\varepsilon > 0$. Dann konvergiert

$$F^{-1}_{Y_{*N}}(p) \text{ fast sicher gegen } F^{-1}(p).$$

Beweis. Sei $\varepsilon > 0$ beliebig und setze $F = F_{Y_n}$, $F_N = F_{Y_{*N}}$, $q = F^{-1}(p)$ und $q_N = F^{-1}_{Y_{*N}}(p)$. Nach Voraussetzung und Definition von $q = F^{-1}(p)$ gibt es $\delta > 0$ mit

$$F(q-\varepsilon) \le p-\delta, \quad F(q+\varepsilon) \ge p+\delta.$$

Weil für jedes $z \in \mathbb{R}$ die empirische Verteilungsfunktion $F_N(z) = F_{Y_{*N}}(z)$ fast sicher gegen $F(z)$ konvergiert, gilt mit Lemma 9.1.5

$$\lim_{M\to\infty} \mathbb{P}\left(\bigcup_{N\ge M} \{|F_N(q-\varepsilon) - F(q-\varepsilon)| \ge \delta/2 \text{ oder } |F_N(q+\varepsilon) - F(q+\varepsilon)| \ge \delta/2\} \right)$$

$$\le \lim_{M\to\infty} \mathbb{P}\left(\bigcup_{N\ge M} \{|F_N(q-\varepsilon) - F(q-\varepsilon)| \ge \delta/2\} \right)$$

$$+ \lim_{M\to\infty} \mathbb{P}\left(\bigcup_{N\ge M} \{|F_N(q+\varepsilon) - F(q+\varepsilon)| \ge \delta/2\} \right) = 0$$

und somit

$$1 = \lim_{M\to\infty} \mathbb{P}\left(\bigcap_{N\ge M} \{|F_N(q-\varepsilon) - F(q-\varepsilon)| < \delta/2 \text{ und } |F_N(q+\varepsilon) - F(q+\varepsilon)| < \delta/2\} \right)$$

$$\le \lim_{M\to\infty} \mathbb{P}\left(\bigcap_{N\ge M} \{F_N(q-\varepsilon) < F(q-\varepsilon) + \delta/2 \text{ und } F_N(q+\varepsilon) > F(q+\varepsilon) - \delta/2\} \right)$$

$$\le \lim_{M\to\infty} \mathbb{P}\left(\bigcap_{N\ge M} \{F_N(q-\varepsilon) < p - \delta/2 \text{ und } F_N(q+\varepsilon) > p + \delta/2\} \right)$$

$$\le \lim_{M\to\infty} \mathbb{P}\left(\bigcap_{N\ge M} \{|q_N - q| < \varepsilon\} \right).$$

Die Behauptung folgt also mit Lemma 9.1.5. □

9.1.7 Beispiel (Lokations-M-Schätzung, Lokations-M-Funktional)
Die Lokations-M-Schätzung $\hat{\mu}_\rho(y)$ (siehe Definition 3.3.1) kann in vielen Fällen über die Gleichung

$$\sum_{n=1}^{N} \psi(y_n - \hat{\mu}_\rho(y)) = 0,$$

gewonnen werden, wobei $\psi(z) = \rho'(z)$ gilt. Damit löst $\hat{\mu}_\rho(y) = \hat{\mu}_\rho(y_{*N})$ die Gleichung

$$\int \psi(z - \hat{\mu}_\rho(y_{*N}))\, P_{y_{*N}}(dz) = 0.$$

Das Funktional T, bei dem der Wert $T(P)$

$$\int \psi(z - T(P))\, P(dz) = 0.$$

erfüllt, wird daher Lokations-M-Funktional genannt. Diese Herleitung ist nicht nur auf den Lokationsfall beschränkt, sondern kann für beliebige Funktionen ψ definiert werden.

9.1.8 Definition (M-Funktional)
Sei $\psi : \mathbb{R}^r \times \Theta \to \mathbb{R}^s$. Das Funktional T, bei dem für jedes Wahrscheinlichkeitsmaß P auf $\mathbb{R}^r$ der Wert $T(P) \in \Theta$ durch

$$\int \psi(z, T(P))\, P(dz) = 0$$

gegeben ist, wird M-Funktional bezüglich ψ genannt.

9.2 Bruchpunkte und Einflussfunktionen für statistische Funktionale

9.2.1 Definition (Kontaminierte Verteilung)
Sei P eine Verteilung, Q eine weitere Verteilung und $\varepsilon \in (0, 1)$. Dann ist die durch Q ε-kontaminierte Verteilung gegeben durch

$$P_{Q,\varepsilon} = (1 - \varepsilon)P + \varepsilon Q.$$

Oft wird für Q das Einpunkt-Maß e_{x_0} benutzt, was einem Ausreißeranteil von ε bei x_0 darstellt. Für $P_{e_x,\varepsilon}$ wird kurz $P_{x,\varepsilon}$ geschrieben.

9.2.2 Definition (Verfälschung und Verfälschungsfunktion)
Die Verfälschung (Bias) einer Verteilung P durch ε-Kontamination bei Q ist gegeben durch

$$Bias(T, P, Q, \varepsilon) = |T(P_{Q,\varepsilon}) - T(P)|$$

und die maximale Verfälschung ist

$$maxBias(T, P, \varepsilon) = \sup_{Q \in \mathcal{P}} |T(P_{Q,\varepsilon}) - T(P)|.$$

Die Verfälschungsfunktion ist definiert als

$$B(T, P, \varepsilon) : \mathbb{R} \ni x \to B(T, P, \varepsilon)(x) = Bias(T, P, e_x, \varepsilon) \in \mathbb{R}.$$

9.2.3 Definition (Explosionspunkt für statistische Funktionale)

$$\epsilon^+(T, P) = \inf\{\varepsilon \in (0, 1);\ maxBias(T, P, \varepsilon) = \infty\}$$

heißt Explosionspunkt des statistischen Funktionals T.

Ebenso könnte auch ein Implosionspunkt für statistische Funktionale $T : \mathcal{P} \to \Theta = [0, \infty)$ defi-

niert werden. Um aber noch algemeinere Parameterräume Θ zu erfassen, kann folgende Definition eines Bruchpunktes benutzt werden.

9.2.4 Definition (Bruchpunkt für statistische Funktionale)

$$\epsilon^*(T,P) = \inf\{\varepsilon \in (0,1);\ \textit{für alle kompakten Mengen } C \subset \textit{int}(\Theta) \\ \textit{gibt es } Q \in \mathcal{P} \textit{ mit } T(P_{Q,\varepsilon}) \notin C\}$$

heißt Bruchpunkt des statistischen Funktionals T.

Dabei bezeichnet int(Θ) das Innere der Menge Θ. Z.B. ist $(0,\infty)$ das Innere von $[0,\infty)$ und in diesem Fall reicht es, als kompakte Menge abgeschlossene Intervalle $[a,b]$ mit $0 < a < b$ für C zu betrachten. Ist $\Theta = \mathbb{R}$, so ist $\mathbb{R}$, d.h. Θ selber, das Innere von Θ und es reicht, abgeschlossene Intervalle $[a,b]$ mit $-\infty < a < b < \infty$ für C zu betrachten. Insbesondere ist der Explosionspunkt ein Bruchpunkt, wenn $\Theta = \mathbb{R}$ gilt.

9.2.5 Definition (Einflussfunktion, Grobe-Fehler-Empfindlichkeit)
Sei T ein Funktional und P eine Verteilung.
a) Dann heißt die Abbildung $IF(T,P,\cdot) : \mathbb{R} \to \mathbb{R}$, gegeben durch

$$IF(T,P,x) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)P + \varepsilon e_x) - T(P)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(P_{x,\varepsilon})\Big|_{\varepsilon=0} \quad ,$$

Einflussfunktion (influence function) von T bei P.
b) Die Grobe-Fehler-Empfindlichkeit (gross error sensitivity) $\gamma(T,P)$ ist definiert als

$$\gamma(T,P) = \sup_{x \in \mathbb{R}} |IF(T,P,x)|.$$

9.2.6 Bemerkung
a) Ist die Grobe-Fehler-Empfindlichkeit endlich, so gilt das Funktional T und die zugehörige Schätzfunktion als ausreißerrobust. Die Endlichkeit der Grobe-Fehler-Empfindlichkeit ist dabei gleichbedeutend mit der Beschränktheit der Einflussfunktion.
b) Die obige Definition ist der einfachheitshalber nur für Wahrscheinlichkeitsmaße auf $\mathbb{R}$ und damit für $x \in \mathbb{R}$ gegeben, da wir in diesem Abschnitt nur für diesen Fall konkrete Beispiele betrachten werden. Der Raum $\mathcal{P}$ aller Wahrscheinlichkeitsmaße kann aber auch der Raum aller Wahrscheinlichkeitsmaße auf $\mathbb{R}^r$ mit $r > 1$ sein, so dass dann für einen Ausreißer $x \in \mathbb{R}^r$ gilt. Auch der Bildraum von T muss nicht $\mathbb{R}$ sein, sondern kann auch $\mathbb{R}^s$ mit $s > 1$ sein. Die obige Definition gilt dann ganz analog mit $IF(T,P,\cdot) : \mathbb{R}^r \to \mathbb{R}^s$.
c) Unter Regularitätsvoraussetzungen gilt (siehe Huber 1981)

$$\lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)P + \varepsilon Q) - T(P)}{\varepsilon} = \int IF(T,P,x)\, Q(dx), \tag{9.2}$$

d.h. die Einflussfunktion liefert für alle möglichen Verunreinigungen Q das Änderungsverhalten.

9.2.7 Satz
Für die Einflussfunktion des Erwartungswertes (arithmetischen Mittels) gilt

$$IF(T, P, x) = x - T(P).$$

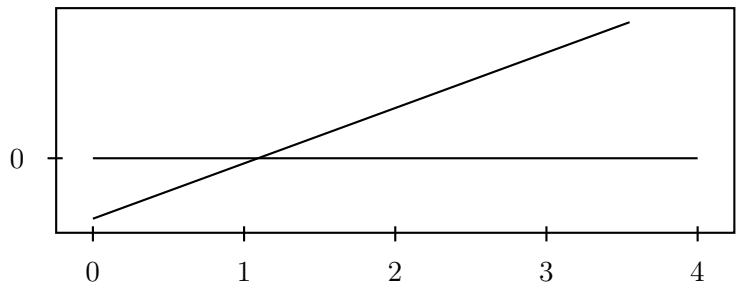


Abbildung 9.1: Einflussfunktion des normalen Mittelwertes

Der Mittelwert ist somit auch in diesem Sinne nicht robust, da die Einflussfunktion unbeschränkt ist, siehe Abb. 9.1.

9.3 Einflussfunktionen von Lokations-Funktionalen

Die Einflussfunktion kann und wurde für viele statistische Fragestellungen bestimmt. Wir beschränken uns hier auf den einfachsten Fall, der Bestimmung der Lage von univariaten Daten. Als erstes betrachten wir das Funktional, dass durch das p-Quantil gegeben ist.

Sei $F_{x,\varepsilon} = (1-\varepsilon)F + \varepsilon\, 1_{[x,\infty)}$ die zu $P_{x,\varepsilon}$ gehörende Verteilungsfunktion.

9.3.1 Lemma
Sei F stetig und streng monoton auf $\mathbb{R}$. Dann gilt für das p-Quantil der mit x ϵ-kontaminierten Verteilung

$$F_{x,\varepsilon}^{-1}(p) = \begin{cases} F^{-1}\left(\frac{p}{1-\varepsilon}\right), & p < (1-\varepsilon)F(x) \\ x, & (1-\varepsilon)F(x) \le p \le (1-\varepsilon)F(x) + \varepsilon \\ F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right), & (1-\varepsilon)F(x) + \varepsilon < p. \end{cases}$$

Beweis. 1.Fall: $p < (1-\varepsilon)F(x)$

$$F_{x,\varepsilon}^{-1}(p) = \inf\{z; F_{x,\varepsilon}(z) \ge p\} = \inf\{z; (1-\varepsilon)F(z) + \varepsilon 1_{[x;\infty)}(z) \ge p\}.$$

Wegen der Stetigkeit von F und der Anforderung an p, gilt

$$\begin{aligned} F_{x,\varepsilon}^{-1}(p) &= \inf\{z \in (-\infty, x); (1-\varepsilon)F(z) \geq p\} \\ &= \inf\left\{z \in (-\infty, x); F(z) \geq \frac{p}{(1-\varepsilon)}\right\} = F^{-1}\left(\frac{p}{1-\varepsilon}\right). \end{aligned}$$

2.Fall:$(1-\varepsilon)F(x) \leq p \leq (1-\varepsilon)F(x) + \varepsilon$
Da F streng monoton wachsend ist, gilt für alle $z < x$

$$F_{x,\varepsilon}(z) = (1-\varepsilon)F(z) < (1-\varepsilon)F(x) \leq p.$$

Desweiteren gilt nach Vorraussetzung $F_{x,\varepsilon}(x) = (1-\varepsilon)F(x) + \varepsilon \geq p$. Somit ergibt sich für das Quantil

$$F_{x,\varepsilon}^{-1}(p) = x.$$

3.Fall:$(1-\varepsilon)F(x) + \varepsilon < p$
Für alle $z < x$ gilt

$$F_{x,\varepsilon}(z) = (1-\varepsilon)F(z) < (1-\varepsilon)F(x) < p.$$

Somit wird in diesen Fall das Quantil im Interval $[x, \infty)$ gesucht:

$$\begin{aligned} F_{x,\varepsilon}^{-1}(p) &= \inf\{z \in [x, \infty); (1-\varepsilon)F(z) + \varepsilon \geq p\} \\ &= \inf\left\{z \in [x, \infty); F(z) \geq \frac{p-\varepsilon}{1-\varepsilon}\right\} = F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right). \square \end{aligned}$$

9.3.2 Satz (Einflussfunktion für Quantile)
Sei F streng monoton und stetig auf $\mathbb{R}$, f die zugehörige Dichtefunktion, $p \in (0,1)$ und $T(P) = F^{-1}(p)$ das p-Quantile von P. Dann ist

$$IF(T, P, x) = \begin{cases} \frac{p-1}{f(F^{-1}(p))}, & x < F^{-1}(p) \\ 0, & x = F^{-1}(p) \\ \frac{p}{f(F^{-1}(p))}, & x > F^{-1}(p) \end{cases}$$

die Einflussfunktion des Quantils.

Beweis. Da F stetig und streng monoton auf $\mathbb{R}$ ist, gilt $F^{-1}(F(x)) = x$ und $F(F^{-1}(p)) = p$.
1.Fall: $x < F^{-1}(p)$
Offensichtlich gilt $F(x) < p$. Somit existiert ein ε_0, so dass $p > F(x) + \varepsilon(1 - F(x)) = (1-\varepsilon)F(x) + \varepsilon$ für alle $0 < \varepsilon < \varepsilon_0$ erfüllt ist. Es gilt

$$\begin{aligned} IF(T, P, x) &= \lim_{\varepsilon \downarrow 0} \frac{F_{x,\varepsilon}^{-1}(p) - F^{-1}(p)}{\varepsilon} \\ &\overset{Lem 9.3.1}{=} \lim_{\varepsilon \downarrow 0} \frac{F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right) - F^{-1}(p)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right)\Big|_{\varepsilon=0}. \end{aligned}$$

Mit der Kettenregel ergibt sich sofort

$$IF(T,P,x) = \frac{\partial}{\partial\varepsilon}F^{-1}(z)\Big|_{z=p} \cdot \left(\frac{\partial}{\partial\varepsilon}\frac{p-\varepsilon}{1-\varepsilon}\Big|_{\varepsilon=0}\right).$$

Anwendung der Umkehrregel, mit der Tatsache das F die Dichtefunktion f besitzt, und der Quotientenregel liefert das gesuchte Ergebnis:

$$\begin{aligned} IF(T,P,x) &= \frac{1}{f(F^{-1}(p))}\left[\frac{-1(1-\varepsilon)-(-1(p-\varepsilon))}{(1-\varepsilon)^2}\right]\Big|_{\varepsilon=0} \\ &= \frac{1}{f(F^{-1}(p))}\left[\frac{p-1}{(1-\varepsilon)^2}\right]\Big|_{\varepsilon=0} = \frac{p-1}{f(F^{-1}(p))}. \end{aligned}$$

2.Fall: $x = F^{-1}(p)$
Da F stetig und streng monton auf $\mathbb{R}$ ist, gilt $p = F(x)$.
Somit folgt sofort $p \geq (1-\varepsilon)F(x)$ für alle $\varepsilon > 0$. Desweitern gilt

$$(1-\varepsilon)F(x) + \varepsilon = (1-\varepsilon)p + \varepsilon = p + \varepsilon(1-p) \geq p$$

für alle $\varepsilon > 0$. Mit diesen Eigenschaften von p kann nun Lemma 9.3.1 angewendet werden, es ergibt sich $F^{-1}_{x,\varepsilon}(p) = x$ für alle $\varepsilon > 0$ und

$$IF(T,P,x) = \lim_{\varepsilon\downarrow 0}\frac{F^{-1}_{x,\varepsilon}(p) - F^{-1}(p)}{\varepsilon} = \lim_{\varepsilon\downarrow 0}\frac{x-x}{\varepsilon} = 0.$$

3.Fall: $x > F^{-1}(p)$
Da F stetig und streng monton auf $\mathbb{R}$ ist, gilt $p < F(x)$. Also gibt es ein $\varepsilon_0 > 0$, so dass $(1-\varepsilon)F(x) > p$ für alle $0 < \varepsilon < \varepsilon_0$ gilt. Mit dieser Eigenschaft von p kann nun Lemma 9.3.1 angewendet werden. So ergibt sich $F^{-1}_{x,\varepsilon}(p) = F^{-1}(\frac{p}{1-\varepsilon})$ für alle $0 < \varepsilon < \varepsilon_0$:

$$\begin{aligned} IF(T,P,x) &= \lim_{\varepsilon\downarrow 0}\frac{F^{-1}_{x,\varepsilon}(p) - F^{-1}(p)}{\varepsilon} = \lim_{\varepsilon\downarrow 0}\frac{F^{-1}(\frac{p}{1-\varepsilon}) - F^{-1}(p)}{\varepsilon} \\ &= \frac{\partial}{\partial\varepsilon}F^{-1}(\frac{p}{1-\varepsilon})\Big|_{\varepsilon=0} = \frac{1}{f(F^{-1}(p))}\left[\frac{\partial}{\partial\varepsilon}\frac{p}{1-\varepsilon}\Big|_{\varepsilon=0}\right] \\ &= \frac{1}{f(F^{-1}(p))}\cdot\frac{p}{(1-\varepsilon)^2}\Big|_{\varepsilon=0} = \frac{p}{f(F^{-1}(p))}. \square \end{aligned}$$

Wir sehen, dass die Einflussfunktion des p-Quantils beschränkt ist. Das bedeutet, dass Ausreißer keinen beliebig großen Einfluss haben können. Insbesondere gilt das für den Median, dem $\frac{1}{2}$-Quantil.

Als nächstes betrachten wir die Einflussfunktion des nach oben getrimmten Mittelwertes. Dieser ist insbesondere dann von Interesse, wenn keine negativen Werte beobachtet werden können. Die Einflussfunktion des nach unten und nach oben getrimmten Mittelwertes ist etwas schwieriger herzuleiten, da die Quantils-Definition über $F^{-1}(p)$ immer den kleinsten möglichen Wert aller möglichen p-Quantile liefert.

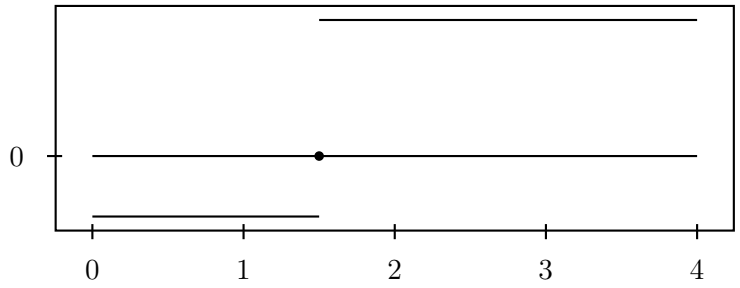


Abbildung 9.2: Einflussfunktion des p-Quantils

9.3.3 Definition
Sei $\beta \in (0,1)$. Das Funktional zum nach oben getrimmten Mittelwert ist gegeben durch

$$T_\beta(P) = \frac{1}{1-\beta} \int 1_{[0,F^{-1}(1-\beta)]}(y)\, y\, P(dy).$$

9.3.4 Bemerkung
a) Es gilt für alle $N \in \mathbb{N}$

$$T_\beta(P_{*N}) = \frac{1}{\lceil N(1-\beta)\rceil} \sum_{n=1}^{\lceil N(1-\beta)\rceil} y_{(n)},$$

falls $y_{(1)} \le \ldots \le y_{(N)}$ der gordnete Datensatz ist.
b) Im Buch von Staudte und Sheather (1990) wird die Einflussfunktion für den nach oben getrimmten Mittelwert angegeben, aber nicht in der korrekten Form.

9.3.5 Satz
Sei F streng monoton und differenziertbar auf $[0,\infty]$, $F(x) = 0$ für $x < 0$ und T_β das Funktional zum nach oben getrimmten Mittelwert. Dann gilt

$$IF(T_\beta, P, x) = \begin{cases} \frac{x - \beta F^{-1}(1-\beta)}{1-\beta} - T_\beta(P), & 0 \le x < F^{-1}(1-\beta), \\ \frac{x}{1-\beta} - T_\beta(P), & x = F^{-1}(1-\beta), \\ F^{-1}(1-\beta) - T_\beta(P), & F^{-1}(1-\beta) < x. \end{cases}$$

Beweis. Sei f die Ableitung von F, d.h. f ist die Dichte von P. Dann gilt

$$\begin{aligned} T_\beta(P_{x,\varepsilon}) &= \frac{1}{1-\beta} \int 1_{[0,F_{x,\varepsilon}^{-1}(1-\beta)]}(y)\, y\, P_{x,\varepsilon}(dy) \\ &= \frac{1}{1-\beta} \left((1-\varepsilon) \int 1_{[0,F_{x,\varepsilon}^{-1}(1-\beta)]}(y)\, y\, P(dy) + \varepsilon \int 1_{[0,F_{x,\varepsilon}^{-1}(1-\beta)]}(y)\, y\, e_x(dy) \right) \\ &= \frac{1}{1-\beta} \left((1-\varepsilon) \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} y\, f(y)\, dy + \varepsilon\, x\, 1_{[0,F_{x,\varepsilon}^{-1}(1-\beta)]}(x) \right). \end{aligned} \tag{9.3}$$

1. Fall: $x < F^{-1}(1-\beta)$.
Ist $x < F^{-1}(1-\beta)$, so gilt $F(x) < 1-\beta$. Wegen der Stetigkeit von F gibt es $\varepsilon_0 > 0$ mit

$$(1-\varepsilon)F(x) + \varepsilon < 1-\beta \;\text{ und }\; x < F^{-1}\left(\frac{1-\beta-\varepsilon}{1-\varepsilon}\right)$$

für alle $\varepsilon < \varepsilon_0$. Mit Lemma 9.3.1 folgt für alle $\varepsilon < \varepsilon_0$

$$x < F^{-1}\left(\frac{1-\beta-\varepsilon}{1-\varepsilon}\right) = F_{x,\varepsilon}^{-1}(1-\beta)$$

und somit mit (9.3)

$$T_\beta(P_{x,\varepsilon}) = \frac{1}{1-\beta}\left((1-\varepsilon)\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} y\, f(y)\, dy + \varepsilon\, x\right). \tag{9.4}$$

Es folgt mit der Produkt- und Kettenregel und mit Satz 9.3.2

$$\begin{aligned}
IF(T_\beta, P, x) &= \frac{\partial}{\partial\varepsilon} T_\beta(P_{x,\varepsilon})\Big|_{\varepsilon=0} \\
&= \frac{1}{1-\beta}\left(-\int 1_{[0,F_{x,\varepsilon}^{-1}(1-\beta)]}(y)\, y\, P(dy)\right)\Big|_{\varepsilon=0} \\
&\quad + \frac{1}{1-\beta}\left((1-\varepsilon)F_{x,\varepsilon}^{-1}(1-\beta)\, f(F_{x,\varepsilon}^{-1}(1-\beta))\big|_{\varepsilon=0}\right)\; \frac{\partial}{\partial\varepsilon}F_{x,\varepsilon}^{-1}(1-\beta)\Big|_{\varepsilon=0} \\
&\quad + \frac{x}{1-\beta} \\
&= -T_\beta(P) + \frac{1}{1-\beta}\left(F^{-1}(1-\beta)\; f(F^{-1}(1-\beta))\; \frac{(1-\beta)-1}{f(F^{-1}(1-\beta))} + x\right) \\
&= \frac{x - \beta\, F^{-1}(1-\beta)}{1-\beta} - T_\beta(P).
\end{aligned}$$

2. Fall: $x = F^{-1}(1-\beta)$.
Aus $x = F^{-1}(1-\beta)$ folgt $F(x) = 1-\beta$ wegen der Stetigkeit von F. Außerdem gilt $(1-\varepsilon)F(x)+\varepsilon = F(x) + \varepsilon(1-F(x)) \geq 1-\beta$, so dass nach Lemma 9.3.1 $x = F_{x,\varepsilon}^{-1}(1-\beta)$ folgt. $T_\beta(P_{x,\varepsilon})$ hat also

wieder die Gestalt (9.4). Nach Satz 9.3.2 gilt hier aber $\frac{\partial}{\partial\varepsilon}F_{x,\varepsilon}^{-1}(1-\beta)\big|_{\varepsilon=0}=0$, so dass wir

$$\begin{aligned}
IF(T_\beta,P,x) &= \frac{\partial}{\partial\varepsilon}T_\beta(P_{x,\varepsilon})\Big|_{\varepsilon=0}\\
&= \frac{1}{1-\beta}\left(-\int 1_{[0,F_{x,\varepsilon}^{-1}(1-\beta)]}(y)\,y\,P(dy)\right)\Big|_{\varepsilon=0}\\
&\quad+\frac{1}{1-\beta}\left((1-\varepsilon)F_{x,\varepsilon}^{-1}(1-\beta)\,f(F_{x,\varepsilon}^{-1}(1-\beta))\big|_{\varepsilon=0}\right)\,\frac{\partial}{\partial\varepsilon}F_{x,\varepsilon}^{-1}(1-\beta)\Big|_{\varepsilon=0}\\
&\quad+\frac{x}{1-\beta}\\
&= -T_\beta(P)+\frac{1}{1-\beta}\,(0+x)\\
&= \frac{x}{1-\beta}-T_\beta(P)
\end{aligned}$$

erhalten.
3. Fall: $x>F^{-1}(1-\beta)$.
Ist $x>F^{-1}(1-\beta)$, so gilt $F(x)>1-\beta$ wegen der strengen Monotonie von F. Damit gibt es $\varepsilon_0>0$ mit

$$(1-\varepsilon)F(x)>1-\beta \;\text{ und }\; x>F^{-1}\left(\frac{1-\beta}{1-\varepsilon}\right)$$

für alle $\varepsilon<\varepsilon_0$. Mit Lemma 9.3.1 folgt für alle $\varepsilon<\varepsilon_0$

$$x>F^{-1}\left(\frac{1-\beta}{1-\varepsilon}\right)=F_{x,\varepsilon}^{-1}(1-\beta)$$

und somit mit (9.3)

$$T_\beta(P_{x,\varepsilon})=\frac{1}{1-\beta}\left((1-\varepsilon)\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} y\,f(y)\,dy+0\right).$$

Satz 9.3.2 liefert hier $\frac{\partial}{\partial\varepsilon}F_{x,\varepsilon}^{-1}(1-\beta)\big|_{\varepsilon=0}=\frac{1-\beta}{f(F^{-1}(1-\beta))}$ und somit

$$\begin{aligned}
IF(T_\beta,P,x) &= \frac{\partial}{\partial\varepsilon}T_\beta(P_{x,\varepsilon})\Big|_{\varepsilon=0}\\
&= \frac{1}{1-\beta}\left(-\int 1_{[0,F_{x,\varepsilon}^{-1}(1-\beta)]}(y)\,y\,P(dy)\right)\Big|_{\varepsilon=0}\\
&\quad+\frac{1}{1-\beta}\left((1-\varepsilon)F_{x,\varepsilon}^{-1}(1-\beta)\,f(F_{x,\varepsilon}^{-1}(1-\beta))\big|_{\varepsilon=0}\right)\,\frac{\partial}{\partial\varepsilon}F_{x,\varepsilon}^{-1}(1-\beta)\Big|_{\varepsilon=0}\\
&= -T_\beta(P)+\frac{1}{1-\beta}\left(F^{-1}(1-\beta)\;f(F^{-1}(1-\beta))\;\frac{1-\beta}{f(F^{-1}(1-\beta))}\right)\\
&= F^{-1}(1-\beta)-T_\beta(P).\,\square
\end{aligned}$$

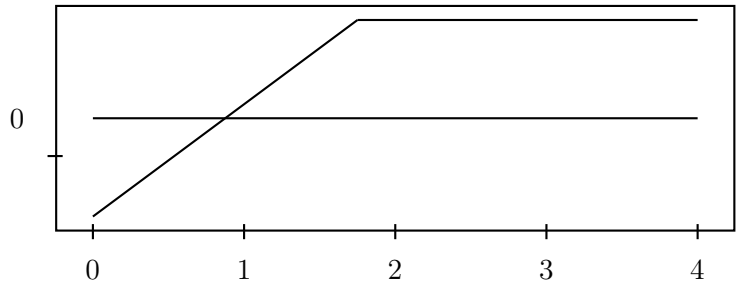


Abbildung 9.3: Einflussfunktion des nach oben getrimmten Mittelwertes

9.4 Einflussfunktion für M-Funktionale

Die Einflussfunktion des in Definition 9.1.8 definierten M-Funktionals hat eine einfache Form. Da die Herleitung auch für komplexere Fragestellungen nicht komplizierter wird, betrachten wir hier einen sehr allgemeinen Fall für Funktionale auf dem Raum der Wahrscheinlichkeitsmaße auf $\mathbb{R}^r$ mit Werten in $\mathbb{R}^s$. Gemäß Bemerkung 9.2.6b) kann die Einflussfunktion ganz analog definiert werden und ist dann eine Funktion vom $\mathbb{R}^r$ nach $\mathbb{R}^s$. Für die Herleitung dieser Einflussfunktion wird der Satz über implizite Funktionen aus der Analysis benötigt.

9.4.1 Satz (Satz über die Differenzierbarkeit von impliziten Funktionen, s. z.B. Heuser 1981, S. 292, 295)
Die Mengen $G \subset \mathbb{R}^p$ und $H \subset \mathbb{R}^s$ seien offen, ξ sei ein Punkt aus G, η ein Punkt aus H und $W : G \times H \to \mathbb{R}^s$ sei eine stetig differenzierbare Funktion mit folgenden Eigenschaften:

$$W(\xi,\eta) = 0 \;\; \textit{und} \;\; \frac{\partial}{\partial y} W(x,y)\Big|_{(x,y)=(\xi,\eta)} \;\; \textit{ist invertierbar.}$$

Dann gibt es eine Umgebung $U \subset G$ von ξ und eine Umgebung $V \subset H$ von η und genau eine stetige Funktion

$$w : U \to V \;\; \textit{mit} \;\; w(\xi) = \eta \;\; \textit{und} \;\; W(x, w(x)) = 0 \;\; \textit{für alle} \;\; x \in U$$

und diese Funktion w ist bei ξ differenzierbar mit der Ableitung

$$\frac{\partial}{\partial x} w(x)\Big|_{x=\xi} = -\left(\frac{\partial}{\partial y} W(x,y)\Big|_{(x,y)=(\xi,\eta)}\right)^{-1} \frac{\partial}{\partial x} W(x,y)\Big|_{(x,y)=(\xi,\eta)}.$$

9.4.2 Satz (Einflussfunktion für M-Funktionale)
Sei T ein M-Funktional bezüglich $\psi : \mathbb{R}^r \times \Theta \to \mathbb{R}^s$ mit $T(P) \in \Theta \subset \mathbb{R}^s$ und $x \in \mathbb{R}^r$. Ist

$$\frac{\partial}{\partial\theta} \int \psi(z,\theta)\, P(dz)\Big|_{\theta=T(P)}$$

invertierbar und existieren $\frac{\partial}{\partial\theta}\int \psi(z,\theta)\, P(dz)$ und $\frac{\partial}{\partial\theta}\psi(x,\theta)$ in einer offenen Umgebung $H \subset \Theta$ von $T(P)$, so gilt

$$IF(T,P,x) = -\left(\frac{\partial}{\partial\theta} \int \psi(z,\theta)\, P(dz)\Big|_{\theta=T(P)}\right)^{-1} \psi(x,T(P)).$$

Beweis. Betrachte $W : \mathbb{R} \times H \to \mathbb{R}^s$ mit

$$W(\varepsilon,\theta) = \int \psi(z,\theta)\, P_{x,\varepsilon}(dz) = (1-\varepsilon)\int \psi(z,\theta)\, P(dz) + \varepsilon\, \psi(x,\theta).$$

Da W linear in ε ist es insbesondere differenzierbar bezüglich ε für alle $\varepsilon \in \mathbb{R} = G$. Nach Voraussetzung ist W auch nach θ für $\theta \in H$ differenzierbar und es gilt

$$W(0,T(P)) = W(0,T(P_{x,0})) = 0.$$

Nach Satz 9.4.1 gibt es eine Umgebung $U \subset \mathbb{R}$ von 0, eine Umgebung $V \subset H$ von $T(P)$ und eine stetige Funktion $w : U \to V$ mit

$$w(0) = T(P) \;\text{ und }\; \int \psi(z,w(\varepsilon))\, P_{x,\varepsilon}(dz) = W(\varepsilon,w(\varepsilon)) = 0 \;\text{ für alle }\; \varepsilon \in U.$$

Damit existiert $T(P_{x,\varepsilon}) = w(\varepsilon)$ für alle $\varepsilon \in U$. Nach Satz 9.4.1 ist w auf U differenzierbar mit

$$\frac{\partial}{\partial\varepsilon} T(P_{x,\varepsilon})\Big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon} w(\varepsilon)\Big|_{\varepsilon=0} = -\left(\frac{\partial}{\partial\theta} W(\varepsilon,\theta)\Big|_{(\varepsilon,\theta)=(0,T(P))}\right)^{-1} \frac{\partial}{\partial\varepsilon} W(\varepsilon,\theta)\Big|_{(\varepsilon,\theta)=(0,T(P))}.$$

Dabei gilt

$$\frac{\partial}{\partial\theta} W(\varepsilon,\theta)\Big|_{(\varepsilon,\theta)=(0,T(P))} = \frac{\partial}{\partial\theta}\int \psi(z,\theta)\, P(dz)\Big|_{\theta=T(P)}$$

und

$$\frac{\partial}{\partial\varepsilon} W(\varepsilon,\theta)\Big|_{(\varepsilon,\theta)=(0,T(P))} = \left(\psi(x,\theta) - \int \psi(z,\theta)\, P(dz)\right)\Big|_{\theta=T(P)} = \psi(x,T(P)) - 0.$$

Also folgt

$$IF(T,P,x) = \frac{\partial}{\partial\varepsilon} T(P_{x,\varepsilon})\Big|_{\varepsilon=0} = -\left(\frac{\partial}{\partial\theta}\int \psi(z,\theta)\, P(dz)\Big|_{\theta=T(P)}\right)^{-1} \psi(x,T(P)).\ \square$$

9.4.3 Bemerkung
Die Einflussfunktion ist also beschränkt, sobald die Score-Funktion ψ beschränkt ist.

9.4.4 Korollar

Ist T ein Lokations-M-Funktional bezüglich $\psi : \mathbb{R} \to \mathbb{R}$ für univariate Daten, d.h. es gehört zur Lokations-M-Schätzung und es ist ein M-Funktional bezüglich $\tilde{\psi}$ mit $\tilde{\psi}(z, \theta) = \psi(z - \mu)$, dann gilt für alle $x \in \mathbb{R}$

$$IF(T, P, x) = \frac{-\psi(x - T(P))}{\frac{\partial}{\partial \mu} \int \psi(z - \mu)\, P(dz)\Big|_{\mu = T(P)}}.$$

Kapitel 10

Literatur

Bücher über Robuste Statistik

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics - The Approach Based on Influence Functions*. John Wiley, New York.

Huber, P.J. (1981). *Robust Statistics*. John Wiley, New York.

Jurečková, J. and Sen, P.K. (1996). *Robust Statistical Procedures. Asymptotics and Interrelations*. Wiley, New York.

Maronna, R.A., Martin, R.D., Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. John Wiley, New York.

Müller, Ch.H. (1997). *Robust Planning and Analysis of Experiments*. Lecture Notes in Statistics 124, Springer, New York.

Oja, H. (2010). *Multivariate nonparametric methods with R. An approach based on spatial signs and ranks*. Lecture Notes in Statistics 199, Springer, New York.

Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer, New York.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley, New York.

Staudte, R.G. and Sheather, S.J. (1990). *Robust Estimation and Testing*. Wiley, New York.

Weitere Litaratur

Croux, C., Ollila, E., Oja, H. (2002). Sign and rank covariance matrices: statistical properties and application to principal components analysis, in: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1-Norm and Related Methods (Papers of the 4th international conference on statistical analysis on the L_1-norm and related methods, Neuchâtel, Switzerland, August 4-9, 2002). Birkhäuser, Basel, pp. 257-269.

Dürre, A., Tyler, D. E., Vogel, D. (2016). On the eigenvalues of the spatial sign covariance matrix in more than two dimensions, to appear in *Statistics and Probability Letters*. arvix 1512.02863

Hoeffding, W., and Robbins, H. (1948). The Central Limit Theorem for Dependent Random Variables. *Duke Mathematical Journal*, 15, 773-780.

Huber, P. (1964). Robust estimation of a location parameter. *Ann. Statist.*, 35, 73-101.

Hubert, M. and Rousseeuw, P.J. (1998). The catline for deep regression. *J. Multivariate Anal.*, 66, 270-296.

Kustosz, Ch.P. and Müller, Ch.H. (2014). Analysis of crack growth with robust, distributionfree estimators and tests for nonstationary autoregressive processes. *Statistical Papers* 55, 125-140.

Kustosz, Ch.P., Leucht, A. and Müller, Ch.H. (2016). Tests based on simplicial depth for AR(1) models with explosion. *Journal of Time Series Analysis* 37, 763-784.

Kustosz, Ch.P., Müller, Ch.H. and Wendler, M. (2016). Simplified simplicial depth for regression and autoregressive growth processes. *Journal of Statistical Planning and Inference* 173, 125-146.

Lee, A.J. (1990). *U-Statistics. Theory and Practice*. Marcel Dekker, New York.

Liu, R.Y. (1988). On a notion of simplicial depth. *Proc. Nat. Acad. Sci. USA* 85, 1732-1734.

Liu, R.Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* 18, 405-414.

Liu, R.Y., Parelius, J.M., und Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *Annals of Statistics* 27, 783-858.

Milasevic, P. and Ducharme, G.R. (1987). Uniqueness of the spatial median. *Ann. Statist.* 15, 1332–1333.

Mizera, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* 30, 1681-1736.

Müller, Ch.H. (1995). Breakdown points for designed experiments. *J. Statist. Plann. Inference.* 45, 413-427.

Müller, Ch.H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis* 95, 153-181.

Müller, Ch.H. (2013). Upper and lower bounds for breakdown points. In: *Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather.* Eds. C. Becker, R. Fried and S. Kuhnt, Springer, Berlin, Heidelberg, 67-84.

Müller, Ch.H. and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *J. Statist. Plann. Inference.* 116, 503-519.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, und W. Wertz (Eds.), *Mathematical statistics and applications*, Vol. B, Dordrecht: Reidel, 283-297.

Rousseeuw, P.J. and Hubert, M. (1999). Regression depth (with discussion). *J. Amer. Statist. Assoc.* 94, 388-433.

Tukey, J.W. (1975). Mathematics and the picturing of data. In *Proc. International Congress of Mathematicians*, Vancouver 1974, 2, 523-531.

Vardi, Y. and Zhang, C.-H. (1999). The multivariate L1-median and associated data depth. *PNAS* 97, 1423–1426.

Wellmann, R., Harmand, P. and Müller, Ch.H. (2009). Distribution-free tests for polynomial regression based on simplicial depth. *Journal of Multivariate Analysis* 100, 622-635.

Wellmann, R., and Müller, Ch.H. (2010a). Tests for multiple regression based on simplicial depth. *Journal of Multivariate Analysis* 101, 824-838.

Wellmann, R., and Müller, Ch.H. (2010b). Depth notions for orthogonal regression. *Journal of Multivariate Analysis* 101, 2358-2371.

Witting, H. and Müller-Funk, U. (1995). *Mathematische Statistik II*. Teubner, Stuttgart.

Zhang, Z., Cui, X., Jeske, D.R., und Borneman, J. (2013). Biclustering scatter plots using data depth measures. *Statistical Analysis and Data Mining* 6, 102-115.