13 Allgemeine Regressionsmodelle

13.1 Allgemeines Lineares Regressions-Modell

a **Modell.** Im Verallgemeinerten Linearen Modell wurde angenommen, dass der Erwartungswert μ_i der Verteilung der Zielgrösse Y_i über die Link-Funktion g linear von den Ausgangsgrössen $x_i^{(j)}$ abhängt. Durch den Erwartungswert und allenfalls einen weiteren Parameter, der für alle Beobachtungen gleich war, war jeweils die Verteilung bestimmt. Zusätzlich wurde dort vorausgesetzt, dass die Verteilung einer Exponential-Familie angehören müsse. Das führt in der Theorie und bei der Berechnung zu gewissen Vereinfachungen. Wenn man diese Voraussetzung fallen lässt, wird es aber nicht wesentlich komplizierter.

Wir wollen also etwas allgemeiner schreiben

$$Y_i \sim \mathcal{F}\langle \mu_i, \gamma \rangle \;, \qquad g\langle \mu_i \rangle = \eta_i = \underline{x}_i^T \underline{\beta}$$

wobei μ_i nicht notwendigerweise der Erwartungswert sein muss.

b **Weibull-Verteilung.** Die Weibull-Verteilung bewährt sich für die Untersuchung von **Ausfalls-Zeiten** oder **Überlebenszeiten**. Sie hat die Dichte

$$f\langle x \rangle = \frac{\alpha}{\sigma}(x/\sigma)^{\alpha-1} \exp\langle -(x/\sigma)^{\alpha} \rangle$$

und den Erwartungswert $\sigma\Gamma\langle 1/\alpha + 1 \rangle$, wobei Γ die Gamma-Funtion bezeichnet. Die Dichte lässt sich nicht in der Form schreiben, die für die Exponential-Familien vorausgesetzt wird.

Der Parameter σ ist ein **Skalen-Parameter**, das heisst, wenn $Y \sim \mathcal{W}\langle \sigma, \alpha \rangle$ ist, dann ist ein Vielfaches $c \cdot Y$ ebenfalls Weibull-verteilt, und zwar multipliziert sich einfach σ entsprechend, $c \cdot Y \sim \mathcal{W}\langle c \cdot \sigma, \alpha \rangle$.

Demgegenüber charakterisiert α die Form der Verteilung. Abbildung 13.1.b zeigt einige Dichtekurven für verschiedene α.

c ▷ **Beispiel**. Für **Kohlenstoff-Fasern** verschiedener Länge (1, 10, 20, 50 mm) wurde je für 57 bis 70 Fasern die Kraft gemessen, die zum Reissen der Fasern führte. (Quelle: Crowder, Kimber, Smith and Sweeting (1991, Abschnitt 4.8).) Abbildung 13.1.c zeigt die Verteilung der Reissfestigkeit für die vier verschiedenen Längen.

Längere Fasern reissen eher. Wie hängt die Reissfestigkeit von der Länge ab? Eine quantitative Antwort liefert ein einfaches Regressionsmodell. Zielgrösse ist die Reissfestigkeit, Eingangsgrösse die Faserlänge. ◁

d **Weibull-Regression.** Um die Abhängigkeit der Zielgrösse von den Ausgangsgrössen zu modellieren, nehmen wir an, dass nur der Skalen-Parameter von den Ausgangsgrössen abhängt, während der Form-Parameter für alle Beobachtungen gleich bleibt. Da der Skalen-Parameter positiv sein muss, liegt als Link-Funktion der Logarithmus nahe. Das führt zu

$$Y_i \sim \mathcal{W}\langle \sigma_i, \alpha \rangle \;, \qquad \log\langle \sigma_i \rangle = \underline{x}_i^T \underline{\beta} \;.$$

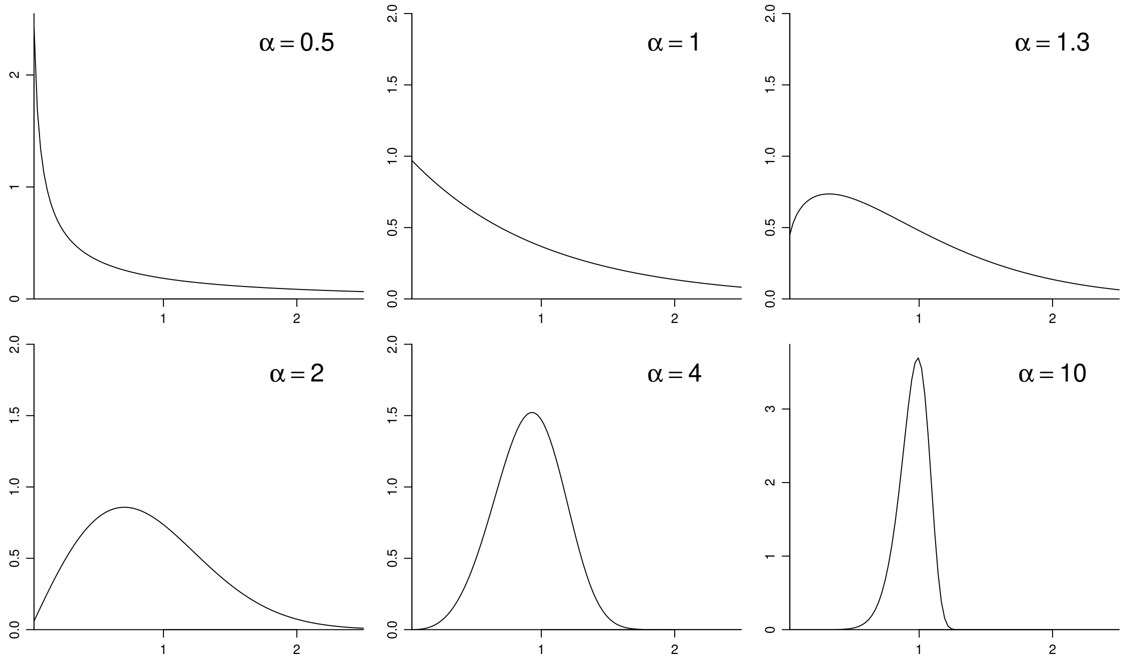

Abbildung 13.1.b: Dichten von sechs Weibull-Verteilungen

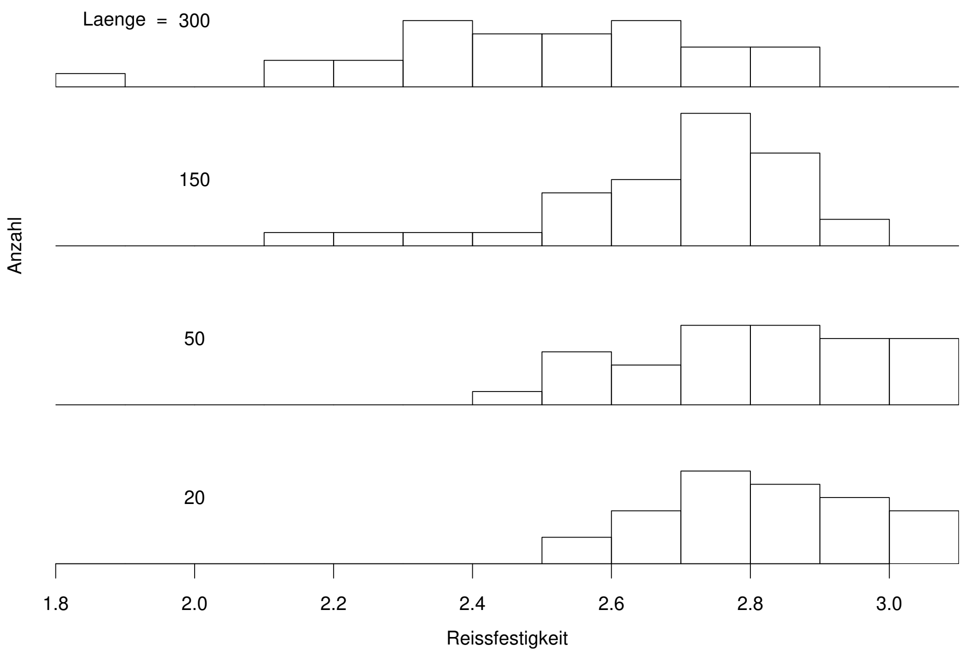

Abbildung 13.1.c: Verteilung der Reissfestigkeit für 4 verschiedene Fasernlängen

e **Gumbel-Regression.** Das Weibull-Modell lässt sich auch anders ausdrücken, indem die Zielgrösse zuerst logarithmiert wird. Diese Transformation macht aus der Weibull-Verteilung eine so genannte **umgedrehte Gumbel-Verteilung** mit der Dichte

$$f\langle x\rangle = \tau^{-1} e^{z} \exp\langle -e^{z}\rangle \ , \qquad z = \frac{x-\mu}{\tau}$$

wobei $\mu = \log\langle\sigma\rangle$ und $\tau = 1/\alpha$. Der Erwartungswert ist $\mu + \gamma\tau \approx \mu + 0.577\,\tau$.

Diese Dichte ist auf die unübliche Seite schief, wie Abbildung 13.1.e zeigt. Unüblich ist eine solche Schiefe für ursprüngliche Daten, aber nach Logarithmus-Transformation ist damit schon zu rechnen, und für logarithmierte Ausfall- und Überlebenszeiten bewährt sich das Modell, da sich ja die Weibull-Verteilung für die untransformierten Daten eignet.

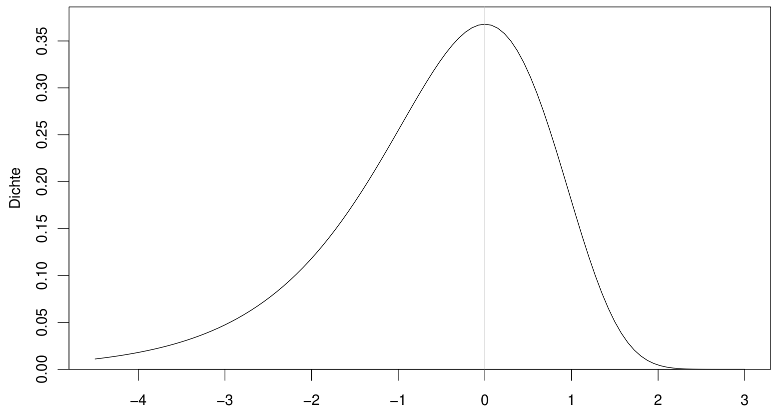

Abbildung 13.1.e: Dichte der umgedrehten Gumbel-Verteilung

Das Regressionsmodell wird dann zu

$$\widetilde{Y}_i = \log Y_i \sim \mathcal{G}\langle \mu_i, \tau\rangle \ , \qquad \mu_i = \underline{x}_i^T \underline{\beta} \ .$$

f* **Zensierte Daten.** In der **Zuverlässigkeits-Theorie** und bei **Überlebenszeiten** kommt oft als Komplikation dazu, dass einige dieser Zeiten nicht zu Ende verfolgt werden, weil beispielsweise einige Maschinen noch keinen Defekt hatten oder einige Kranke am Ende der Studiendauer glücklicherweise noch am Leben sind. Die beiden Gebiete werden deshalb oft in spezialisierten Büchern behandelt, in denen solche „zensierte Daten" behandelt werden und in denen die Weibull-, Gumbel- und andere Verteilungen und die entsprechenden Regressionsmodelle eine wichtige Rolle spielen.

g **Lineares Modell mit nicht-normalen Fehlern.** Damit sind wir wieder bei einem linearen Modell ohne Link-Funktion angelangt. Der einzige Unterschied zum gewöhnlichen linearen Regressionsmodell besteht darin, dass für die Verteilung der Zufallsabweichungen keine Normalverteilung vorausgesetzt wird. Wir können das Modell wieder in der Form „Zielgrösse = Regressionsfunktion + Zufallsabweichung" schreiben,

$$Y_i = \underline{x}_i^T \underline{\beta} + E_i \ , \qquad E_i/\sigma \sim \mathcal{F}_1 \ .$$

Wenn wir $\mathcal{F}_1 = \mathcal{G}\langle 0, 1\rangle$ einsetzen, erhalten wir die Gumbel-Regression (mit Y_i statt $\widetilde{Y}_i$ geschrieben). In diesem Modell kann man natürlich auch eine andere als die Gumbel-Verteilung verwenden.

h **Langschwänzige Fehler.** Die Praxis lehrt, dass Daten eigentlich nie wirklich normalverteilt sind. Wenn sie schief sind, kann man mit Transformation oft eine genäherte Symmetrie erreichen. Aber auch in diesem Fall ist es meistens so, dass extreme Beobachtungen häufiger vorkommen als gemäss der Normalverteilung zu erwarten wäre. Bei extremen Beobachtungen, die als Ausreisser angesprochen werden, kann es sich um „grobe Fehler“ handeln oder um Beobachtungen, die eigentlich nicht dem Modell folgen, das für den Grossteil der Daten gilt. Dafür kann man Gründe suchen. Andererseits kann es sein, dass sich „die Natur“ einfach nicht an die Normalverteilung hält – welche Frechheit! Dann sollte man wohl eine andere Verteilung zur Beschreibung der zufälligen Abweichungen verwenden, nämlich eine, die eben eine höhere Wahrscheinlichkeit für extreme Beobachtungen festlegt. Solche Verteilungen heissen **langschwänzig** oder dickschwänzig oder kurtotisch.

i **t-Verteilung.** Für diesen Zweck wird oft die Familie der t-Verteilungen benützt, die ja eigentlich nicht als Verteilung von Beobachtungen, sondern als Verteilung der Teststatistik des t-Tests eingeführt wurde. Damit sie für Daten taugt, muss man sie zunächst verallgemeinern, indem man Skalenänderungen „einbaut“: X ist dann t-verteilt mit Lage-Parameter μ, Skalenparameter σ und Formparameter ν, wenn die standardisierte Variable $(X - \mu)/\sigma$ eine t-Verteilung mit ν **Freiheitsgraden** hat. Die Dichte ist deshalb

$$f_{\mu,\sigma,\nu}\langle x\rangle = c(1 + x^2/\nu)^{-(\nu+1)/2} ,$$

wobei die Normierungskonstante $c = \Gamma\left\langle(\nu + 1)/2\right\rangle\big/(\Gamma\langle\nu/2\rangle\sqrt{\pi\nu})$ und Γ die Gamma-Funktion ist. Dabei muss für ν keine ganze Zahlen eingesetzt werden. Wenn ν gegen Unendlich geht, geht die t-Verteilung in die Normalverteilung über. ($1/\nu$ könnte also als Mass für die Langschwänzigkeit benützt werden.)

Für $\nu = 1$ erhält man die so genannte **Cauchy-Verteilung**, die so extrem langschwänzig ist, dass sie nicht einmal einen Erwartungswert hat, da das Integral, das den Erwartungswert ja definiert, nicht bestimmt werden kann! Der Parameter μ ist dann immer noch Symmetriezentrum und Median der Verteilung, aber nicht mehr Erwartungswert. Eine Varianz hat diese Verteilung noch weniger, und der Zentrale Grenzwertsatz gilt (deshalb) nicht. Es zeigt sich, dass das arithmetische Mittel von Beobachtungen dieser Verteilung nicht genauer ist als jede einzelne Beobachtung – es hat sogar genau die gleiche Verteilung wie jede einzelne Beobachtung. Dieses Modell widerspricht also sozusagen dem gesunden Menschenverstand, aber kann gerade deshalb als Warnung dienen, dass allzu langschwänzige Verteilungen zu völlig unerwarteten Effekten führen können!

Realistische Verteilungen ergeben sich für $\nu > 2$. Für diese existieren Erwartungswert und Varianz. In unserem Zusammenhang werden wir die t-Verteilung mit $\nu = 3$ oder $\nu = 5$ ins Regressionsmodell als einsetzen, um die Zufallsabweichungen zu modellieren. Genauer wird $\mathcal{F}_1 = t\langle\mu = 0, \sigma = 1, \nu\rangle$ sein mit einem festgesetzten ν.

j **Maximum Likelihood.** Wie sollen die Parameter geschätzt werden? Wie üblich benützt man das Prinzip der Maximalen Likelihood. Für das oben erwähnte Modell (13.1.g) führt dies für die Schätzung der Koeffizienten β_j zur Minimierung der negativen log-Likelihood

$$\begin{aligned} \ell\ell\langle\underline{\beta}, \sigma\rangle &= \sum\nolimits_i \rho\left\langle\frac{Y_i - \underline{x}_I^T\underline{\beta}}{\sigma}\right\rangle + n\log\langle\sigma\rangle \\ \rho\langle r\rangle &= -\log\langle f_1\langle r\rangle\rangle \ . \end{aligned}$$

(* Wenn $E_i/\sigma \sim \mathcal{F}_1$ mit Dichte f_1 gilt, dann hat E_i die Dichte $\frac{1}{\sigma}f_1\langle e/\sigma\rangle$. Von da kommt der Term $n\log\langle\sigma\rangle$.)

Wenn man für f_1 die Standard-Normalverteilungs-Dichte wählt, erhält man $\rho\langle r\rangle = r^2/2$. Für die t-Verteilung ist

$$\rho\langle r\rangle = \frac{\nu+1}{2}\,\log\big\langle 1 + r^2/\nu\big\rangle\ .$$

(Die Konstante $\log\langle c\rangle$ kann man weglassen.)

k **Normalgleichungen.** Statt die Minimalstelle zu finden, kann man wie üblich ableiten und null setzen. Die Ableitung von ρ bezeichnen wir mit ψ; diejenige von $R_i = (Y_i - \underline{x}_I^T\underline{\beta})/\sigma$ nach den Komponenten von $\underline{\beta}$ ergibt den Vektor $-\underline{x}_i/\sigma$. Ableiten und null Setzen führt dadurch zu

$$\sum\nolimits_i \psi\left\langle\frac{Y_i - \underline{x}_i^T\underline{\widehat{\beta}}}{\sigma}\right\rangle \underline{x}_i = \underline{0}\ , \qquad \psi\langle r\rangle = \rho'\langle r\rangle$$

Für die Standard-Normalverteilungs-Dichte wird $\psi\langle r\rangle = r$ und daraus die Gleichung $\sum_i (Y_i - \underline{x}_i^T)\underline{x}_i = \underline{0}$. Die letzte Vektorgleichung wird als „die Normalgleichungen" bezeichnet. Bekanntlich kann man sie mit linearer Algebra explizit lösen. Wenn eine andere Verteilung angenommen wird, muss man einen iterativen Algorithmus zur Lösung der Gleichung oder zur Minimierung der negativen log-Likelihood einsetzen.

Für die t-Verteilungen wird

$$\psi\langle r\rangle = (1+1/\nu)\,\frac{r}{1+r^2/\nu}\ .$$

Abbildung 13.1.k zeigt diese Funktionen für 4 verschiedene Freiheitsgrade ν.

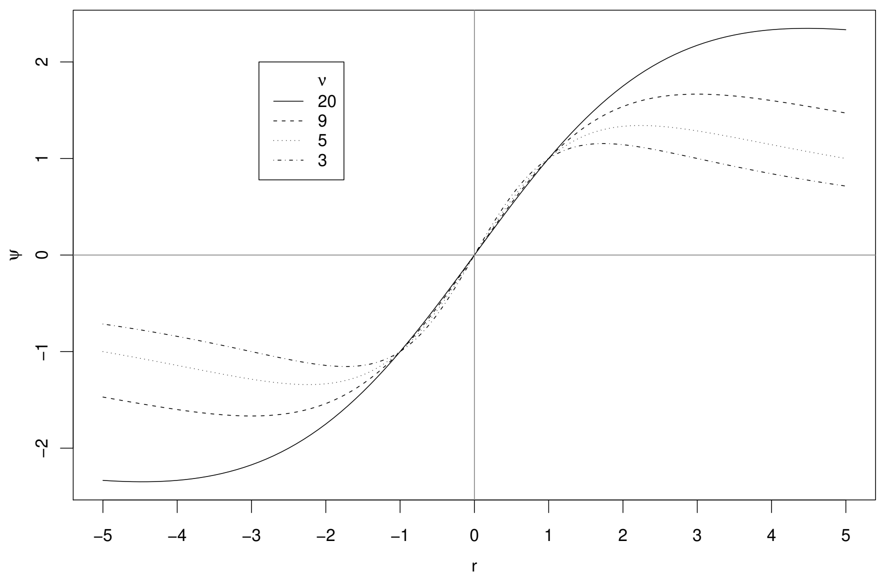

Abbildung 13.1.k: ψ-Funktionen für t-Verteilungen mit vier Freiheitsgraden.

l **Gewichtete Kleinste Quadrate.** Die verallgemeinerten Normalgleichungen kann man auch schreiben als

$$\sum_i w_i R_i \underline{x}_i = \underline{0}\ , \qquad R_i = \frac{Y_i - \underline{x}_I^T \widehat{\underline{\beta}}}{\sigma}\ , \qquad w_i = \psi\langle R_i\rangle / R_i\ .$$

Das sind die Normalgleichungen für Gewichtete Kleinste Quadrate. Im Unterschied zur Gewichteten Linearen Regression hängen hier die Gewichte aber von den geschätzten Parametern ab, und damit werden die Gleichugen zu impliziten Gleichungen. Immerhin kann man diese für einen iterativen Algorithmus brauchen: Mit einer vorläufigen Schätzung für $\underline{\beta}$ bestimmt man die Gewichte w_i und löst dann das Problem der Gewichteten Kleinsten Quadrate. Das führt zu einer verbesserten Schätzung von $\underline{\beta}$. Diese beiden Schritte können wiederholt werden, bis die Schätzung sich nicht mehr ändert.

Die Tatsache, dass die Lösung schliesslich wie eine gewichtete gewöhnliche Regressions-Schätzung aussieht, kann auch der Anschauung helfen.

m ▷ Im **Beispiel der Reissfestigkeit von Fasern** soll eine Regression der Zielgrösse Reissfestigkeit (strength) und die Länge (length) als Eingangsgrösse angepasst werden. Die Weibull-Regression ist im R-package `survivial` enthalten, das auf die Analyse von zensierten Daten ausgerichtet ist. Deshalb muss die Zielgrösse als Objekt der Klasse `Surv` „eingepackt“ werden. Aufruf und Resultat sind in Tabelle 13.1.m enthalten. Die Grösse `scale` entspricht τ (13.1.e) und ist damit der Reziprokwert des Form-Parameter α. Es wird also $\widehat{\alpha} = 1/0.0588 = 17.0$. In der Tabelle wird auf der letzten Zeile `Log(scale)`=0 getestet mit dem P-Wert, der hier 0 wird. Das zu testen ist nicht ganz sinnlos, da diese Hypothese $\alpha = 1$ gleichkommt, und das würde eine Exponential-Verteilung der Abweichungen für die unlogarithmierte Zielgrösse bedeuten.

```
survreg(formula = Surv(strength, rep(1, nrow(dd))) ~ length,
    data = dd)
                Value Std. Error      z        p
(Intercept)  1.068937   8.53e-03 125.28 0.00e+00
length      -0.000343   4.99e-05  -6.87 6.31e-12
Log(scale)  -2.833522   7.24e-02 -39.11 0.00e+00

Scale= 0.0588

Weibull distribution
Loglik(model)= 31.5   Loglik(intercept only)= 13.4
 Chisq= 36.1 on 1 degrees of freedom, p= 1.8e-09
Number of Newton-Raphson Iterations: 6
n= 119
```

Tabelle 13.1.m: Numerische Ergebnisse der Weibull-Regression im Beispiel der Reissfestigkeit von Fasern

Für die grafische Darstellung (Abbildung 13.1.m) benützen wir die logarithmierte Zielgrösse, da für sie der Zusammenhang mit der Eingangsgrösse linear ist. ◁

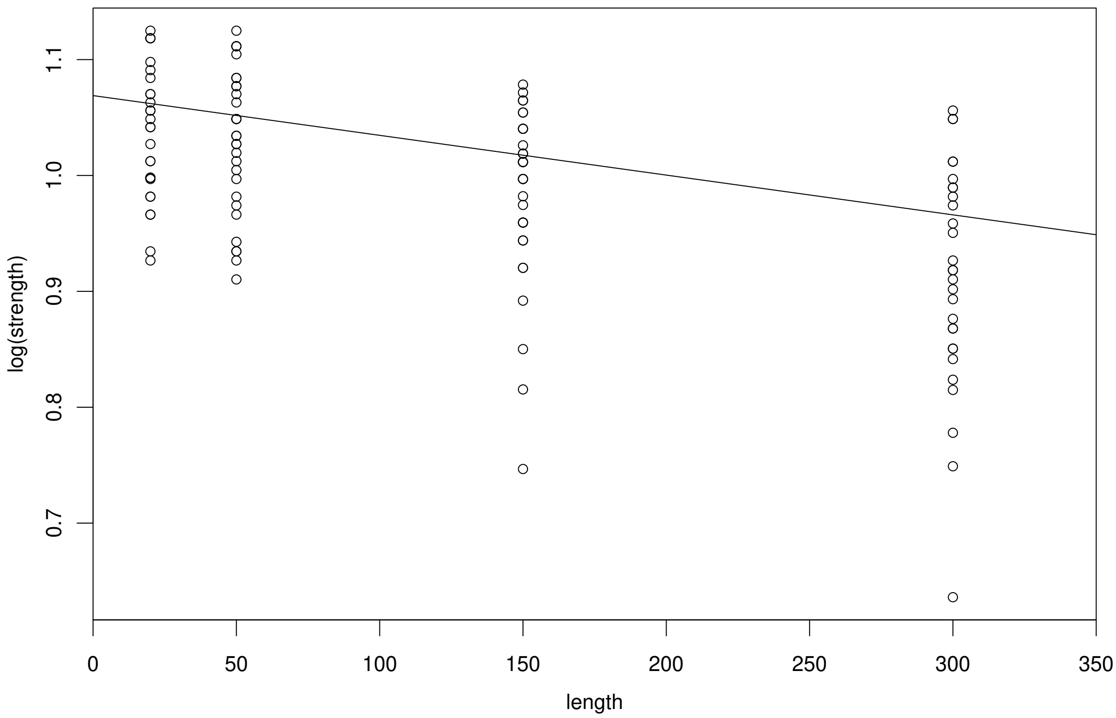

Abbildung 13.1.m: Weibull- oder Gumbel-Regression im Beispiel der Reissfestigkeit von Fasern

n **Verteilung der Schätzung.** Wie bei den Verallgemeinerten Linearen Modellen kann die Verteilung der geschätzten Koeffizienten $\underline{\widehat{\beta}}$ nicht exakt angegeben werden, sondern nur genähert. Man benützt die Näherung, die sich aus dem Zentralen Grenzwertsatz ergibt, die „asymptotische Verteilung". Sie ist eine mehrdimensionale Normalverteilung mit Erwartungswerts-Vektor $\underline{\beta}$ und einer Kovarianzmatrix, die mit der Kovarianzmatrix im Fall der gewöhnlichen linearen Regression (also der Kleinste-Quadrate-Schätzung bei normalverteilten Zufallsabweichungen, siehe 3.5) bis auf einen Faktor übereinstimmt. So wird

$$\mathrm{var}\left\langle\underline{\widehat{\beta}}\right\rangle \approx \sigma^2 \cdot \kappa \boldsymbol{C}^{-1} , \qquad \boldsymbol{C} = \sum\nolimits_i \underline{x}_i \underline{x}_i^T ,$$

wobei $\kappa = \int \psi^2 \langle u \rangle \, f_1 \langle u \rangle \, du$ ist. Für die Standard-Normalverteilung erhält man $\kappa = 1$ und damit die Kovarianzmatrix, die in 3.5 angegeben wurde.

o **Tests, Vertrauensbereiche.** Aus dieser Näherungs-Verteilung erhält man in der üblichen Weise Tests und Vertrauensintervalle für einzelne Koeffizienten. Die Kovarianzmatrix enthält in ihrer Diagonalen die Standardfehler der einzelnen Koeffizienten, die man dazu braucht. Auch für Tests von mehreren Koeffizienten, also beispielsweise für den Test betreffend den Einfluss eines Faktors, und für die entsprechenden Vertrauensbereiche enthält die Kovarianzmatrix die nötige Information.

13.2 Tobit-Regression

a Bei der Messung von Schadstoffen kann man unterhalb der so genannten *Nachweisgrenze* eines Messgeräts die Konzentration nicht mehr angeben. Solche Messungen einfach als fehlend zu betrachten, wäre aber ein Missgriff, denn man weiss je etwas Entscheidendes über die entsprechende Schadstoff-Konzentration: Sie ist kleiner als die Nachweisgrenze. Wenn man aber einfach mit der Nachweisgrenze rechnet, obwohl man weiss, dass die Konzentration kleiner ist, macht man ebenfalls einen Fehler.

Solche **nach unten begrenzte Zufallsvariable** gibt es auch in anderen Anwendungsgebieten:

- Die Regenmenge kann nicht negativ sein, null aber glücklicherweise schon. Das ist nicht genau der gleiche Fall, da in dann die wirkliche Regenmenge ja gemessen werden kann, eben 0.
- Das unten besprochene Modell wurde von einem Oekonomen names Tobin eingeführt, der beschreiben wollte, wofür die Leute ihr Geld ausgeben. Ausgabenposten in einem Haushaltbudget können auch im Prinzip nicht negativ werden (die Wenigsten haben negative Ausgaben für das Vergnügen).
- In Versicherungspolicen kann es Schadenfälle geben, die zwar angemeldet werden, aber aus verschiedenen Gründen schliesslich doch nicht zu einem Schaden führen. (Hier werden die seltenen Fälle weggelassen, für die die Versicherung aus einem Schadenfall einen Gewinn zieht, zum Beispiel wegen Regressionsmöglichkeiten.)

In allen diesen Beispielen beobachten wir eine Variable, deren Verteilung aus zwei Teilen besteht: Einer Wahrscheinlichkeit p_0, dass der Minimalwert erhalten wird, und einer kontinuierlichen Verteilung für Werte, die grösser sind.

b $\triangleright$ Als **Beispiel** betrachten wir die Daten, die **Tobin** zur Illustration seines Modells dienten. Zielgrösse ist die Menge von „haltbaren Gütern“ (durable goods, Variable `durable`), die 20 Individuen kauften; Eingangsvariable sind das Alter (`age`) und ein Index (`quant`) für die Liquidität der Personen. Abbildung 13.2.b zeigt die Daten. $\triangleleft$

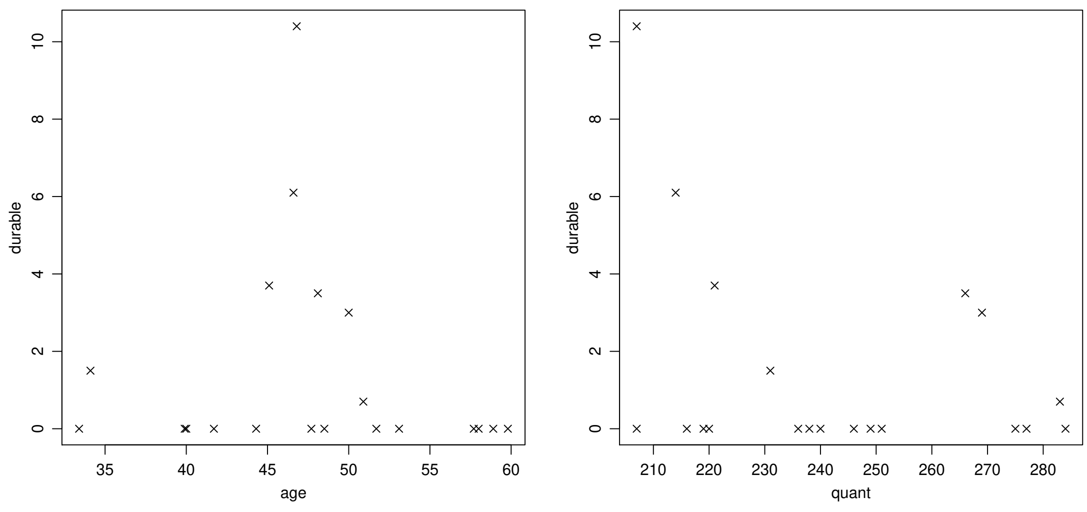

Abbildung 13.2.b: Die Daten des Beispiels von Tobin

c **Separate Modelle.** Wie soll eine solche Variable als Zielgrösse Y in ein Regressionsmodell einfliessen? Eine recht verbreitete Art der Modellierung besteht darin, zunächst eine Regression aufzusetzen mit der binären Zielgrösse Y^*, die gleich 1 ist, wenn die eigentliche Zielgrösse Y positiv ist – beispielsweise eine logistische Regression. Als zweite Stufe stellt man ein Modell auf für die Beobachtungen, für die $Y_i > 0$ ist. Diesen Ansatz wollen wir hier nicht weiter verfolgen.

d **Tobit-Regression.** Im Beispiel, und allgemein für Situationen, in denen ein Messinstrument zur Begrenzung der Werte führt, ist es konsequent, zunächst ein Regressionsmodell für die Zielgrösse „ohne Begrenzung" anzusetzen und in einem zweiten Schritt zu formulieren, dass Werte unter der Nachweisgrenze nicht quantitativ erfasst werden können. Wir setzen also zunächst eine gewöhnliche lineare Regression für die (logrithmierte) „wahre" Schadstoffbelastung Z an,

$$Z_i = \underline{x}_i^T \beta + E_i \;, \quad E_i \sim \mathcal{N}\left\langle 0, \sigma^2 \right\rangle$$

Die Beobachtungen sind

$$Y_i = \begin{cases} y^* & \text{falls } Z_i \le y^* \\ Z_i & \text{falls } Z_i > y^* \end{cases}$$

Die Variable Z ist damit eine „teilweise latente" Variable. Wir brauchen sie zu Modellierungszwecken.

Abbildung 13.2.d veranschaulicht dieses Modell für eine einfache Regression mit simulierten Daten.

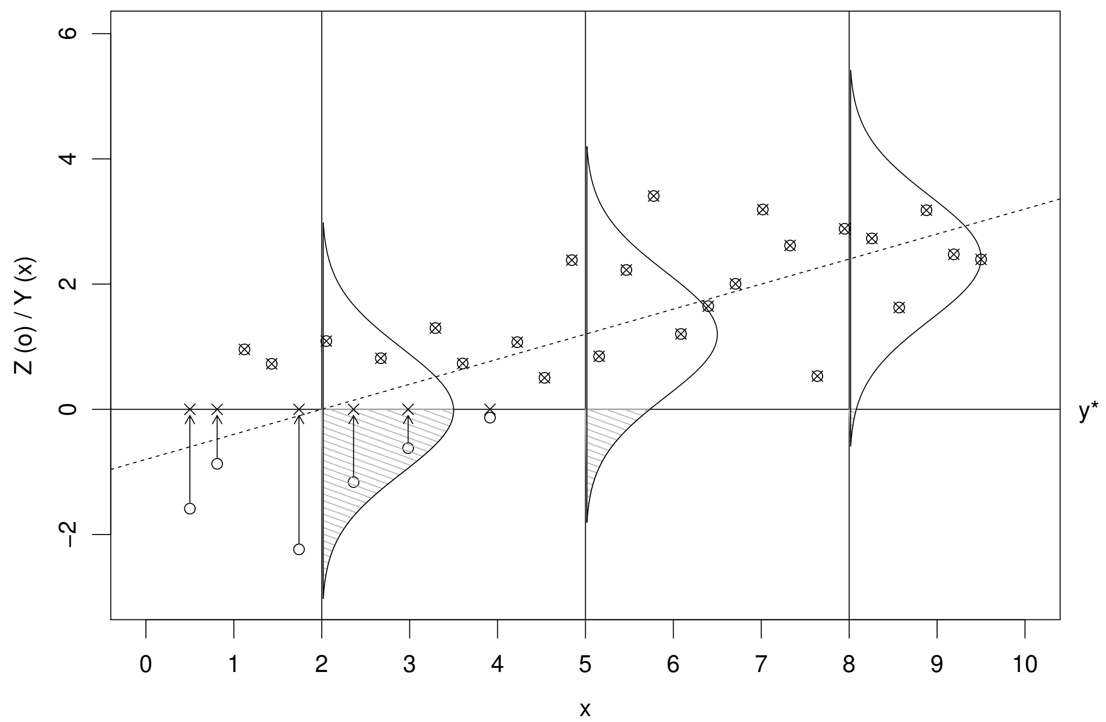

Abbildung 13.2.d: Das Tobit-Modell

e **Interpretation.** Im Falle einer technischen Begrenzung der gemessenen Werte durch ein Messinstrument entspricht die Einführung der latenten Variablen Z den Gegebenheiten. In den anderen eingangs erwähnten Situationen (13.2.a) ist das nicht zwingend, kann aber auch sinnvoll interpretiert werden:

- Beim Regen kann man sich die latente Variable als „Regenpotential" vorstellen, das auch negativ sein kann (trocken oder sehr trocken), wobei der Regen $= 0$ ist, wenn das Potential negativ ist, während Regen und Regenpotential das Gleiche sind, wenn sie positiv sind.
- Ausgaben für gewisse Angebote hängen vom verfügbaren freien Einkommen ab. Wenn dieses unter ein Niveau sinkt, das einen gewissen Komfort erlaubt, werden für nicht lebensnotwendige Angebote wie Ferien immer weniger Haushalte überhaupt etwas ausgeben, und wenn schon, werden es immer kleinere Beträge sein.

Wie erwähnt, können solche Phänomene auch mit zwei Regressionsmodellen, einem binären und einem quantitativen, beschrieben werden. Das hat zwei Nachteile:

- Es könnten sich unplausible Ergebnisse zeigen, indem für bestimmte Situationen zwar eine kleine Wahrscheinlichkeit $P\langle Y > 0\rangle$ für ein positives Y geschätzt wird, aber ein grosser Erwartungswert $E\langle Y \mid Y > 0\rangle$, gegeben dass Y positiv ist, resultiert.
- Die beiden Modelle haben insgesamt etwa doppelt so viele Parameter, die zu schätzen sind.

Beide Punkte können auch Vorteile bieten im Sinne der Flexibilität.

f **Schätzung.** Die Schätzung der Parameter erfolgt auch hier über das Prinzip der Maximalen Likelihood. Die meisten wundern sich zunächst, dass hier Likelihoods, die aus diskreten Wahrscheinlichkeiten $P\langle Y_i = 0\rangle$ entstehen, mit solchen, die Dichten $f\langle y_i\rangle$ entsprechen, gemischt werden können. Das geht; man gewöhnt sich an den Gedanken.

Tests und Vertrauensintervalle ergeben sich auch wie üblich aus der asymptotischen Verteilung der geschätzten Grössen.

g ▷ **Beispiel** der Daten von **Tobin**. Eigentlich müsste man sagen, dass der Datensatz zu klein sei, um ein Modell anzupassen. Tun wir es trotzdem, dann wird im einfachsten Modell `r1 <- regr(Tobit(durable) ~ age + quant, data=tobin)` weder ein Koeffizient noch die Gesamt-Regression signifikant. Aus der Darstellung der Daten kann man auf eine quadratische Abhängigkeit vom Alter schliessen. Wir erhalten die in Tabelle 13.2.g enthaltenen Resultate.

Die Liquidität hat keinen nachweisbaren Einfluss. Da auch das Gesamt-Modell immer noch keine Signifikanz zeigt, sind die beiden signifikanten Koeffizienten eigentlich auch nicht ernst zu nehmen – besonders, wenn wir noch daran denken, dass der lineare Term nicht direkt interpretierbar ist, weil der quadratische Term da ist, und dass dieser ins Modell genommen wurde, weil die Daten das nahelegen; wir müssen also mit dem Selektions-Fehler rechnen, der bei explorativer Modell-Entwicklung immer vorhanden ist. (Ein zufälliges Muster in den Daten führt zu einem formal signifikanten Testergebnis, wenn man das Modell nach ihm ausrichtet und dann den entsprechenden Term testet.)

Trotz diesem negativen Ergebnis soll das Beispiel noch dazu benützt werden, deutlich zu machen, dass naive Auswertungen in die Irre führen. In Abbildung 13.2.g wird nicht nur die nach der Tobit-Regression angepasste quadratische Funktion gezeigt, sondern auch

```
Call:
regr(formula = Tobit(durable) ~ age + I(age^2) + quant, data = tobin)
Fitting function:  survreg

Terms:
                coef stcoef signif  R2.x df p.value
(Intercept) -88.2971     NA -0.722    NA  1      NA
age           4.5248  34.51  1.057 0.992  1   0.038
I(age^2)     -0.0505 -36.47 -1.088 0.992  1   0.033
quant        -0.0494  -1.28 -0.496 0.060  1   0.331
log(scale)    1.5292     NA  2.576    NA  1   0.000

      deviance df p.value
Model     5.65  3    0.13
Null     53.33 20      NA

Distribution:  gaussian.  Shape parameter ('scale'):  4.61
AIC:  5.0063.33
```

Tabelle 13.2.g: Ergebnisse der Tobit-Regression für das Beispiel der Tobin-Daten mit quadratischer Abhängigkeit vom Alter.

diejenigen, die man erhält, indem man

- die Nullen wie gewöhnliche Beobachtungen behandelt und eine gewöhnliche Regression durchführt,
- die Nullen weglässt und dann eine gewöhnliche Regression rechnet. $\triangleleft$

h **Zensierte Beobachtungen.** Die Situation, dass Variable manchmal nicht exakt erfasst werden können, entsteht nicht nur aus nach unten begrenzten Messbereichen von Messinstrumenten.

- Bei Überlebenszeiten oder Ausfallzeiten (engl. *survival* oder *failure time data)* kennt man oft für einige Beobachtungen nur eine Obergrenze: Für Patient/innen wird verfolgt, wie lange sie nach einem Startereignis wie einer Ansteckung, einem Unfall, einer Operation krank sind oder überleben. Da eine Studie nicht ewig dauern kann und einige Patient/innen wegziehen oder aus anderen Gründen nicht mehr weiter verfolgt werden können, weiss man für diese Personen nur, dass die Zeitspanne sicher länger war als bis zum letzten Kontakt. Analog kann man für technische Geräte die Zeitspanne bis zum ersten Fehler erfassen. Auch da will man nicht so lange untersuchen, bis alle Geräte einen Fehler gezeigt haben.
- Versicherungs-Schäden sind jeweils nur bis zu einem vereinbarten Höchstbetrag versichert. Wenn der Schaden höher ist, wird seine Höhe bei der Versicherung oft nicht genauer erfasst.
- Es gibt auch Situationen, in denen man weiss, dass eine „Überlebenszeit" in einem Intervall liegt, zum Beispiel, dass das fragliche Ereignis zwischen zwei Arztbesuchen stattgefunden hat.

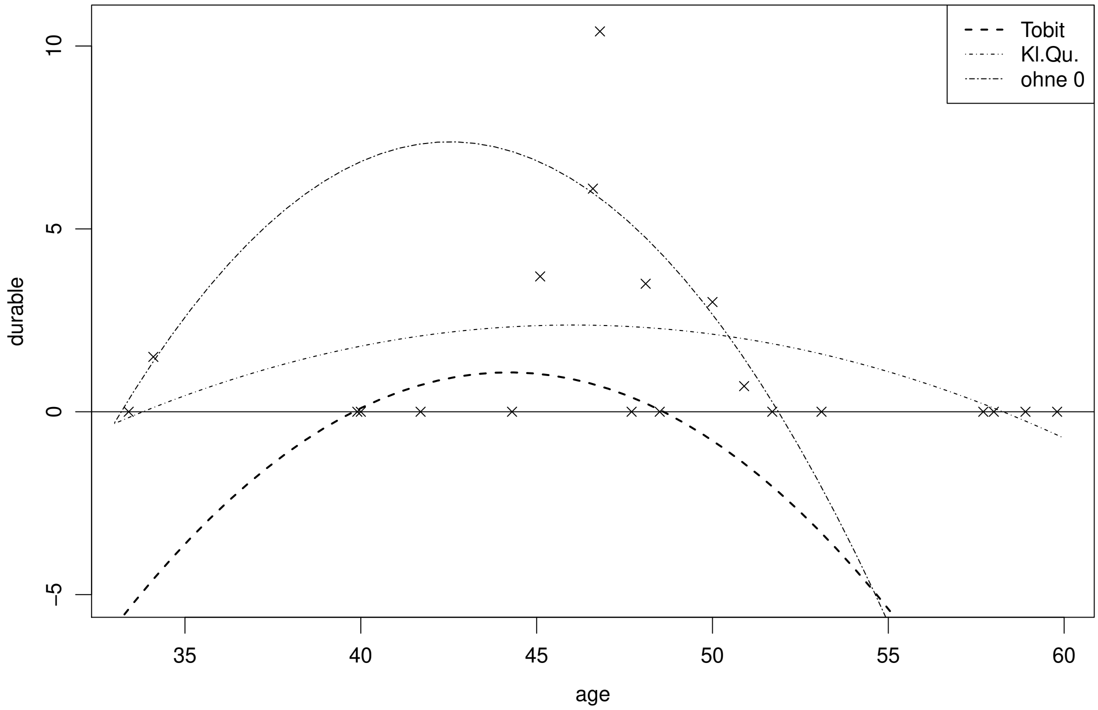

Abbildung 13.2.g: Angepasste Modelle im Beispiel von Tobin

All diese Situationen führen zu (teilweise) **zensierten Daten**, bei Nachweisgrenzen zu „links zensierten“, bei Überlebens- und Ausfallzeiten meist zu „rechts zensierten“ und im letzten Fall zu „Intervall-zensierten“ Daten.

L **Literatur:** Zensierte Daten und Überlebenszeiten. Die Statistik für solche Daten füllt ganze Bücher. Da zensierte Daten oft im Zusammenhang mit Überlebenzeiten auftreten, behandeln die Bücher die Kombination dieser beiden Themen. Beispiele sind Collet (1994), Tableman and Kim (2003), Therneau and Grambsch (2000).

Von Ausfallzeiten handelt beispielsweise Crowder et al. (1991). Hier wird auch die Weibull-Regression (13.1.d) gut beschrieben.

14 Robuste Methoden

14.1 Einfluss und Robustheit

a Robuste Methoden im engeren Sinn sind solche, die sich durch grob falsche Beobachtungen nicht stark verschlechtern. In diesem Kapitel sollen robuste Methoden für die Anpassung von Regressionsmodellen eingeführt werden. Dazu brauchen wir zunächst Grundbegriffe, die die Robustheit allgemein charakterisieren.

b **Sensitivität.** Als ein Teil der Residuenanalyse wurde in 4.10.b der Einfluss einer einzelnen Beobachtung auf die Resultate der Regression untersucht. Die Idee war, reihum eine einzelne Beobachtung wegzulassen, die Regression neu zu schätzen und die Differenz der erhaltenen Werte zu den ursprünglichen zu bilden, die mit dem ganzen Datensatz gewonnen wurden. Diese Differenz bildete ein Mass für den Einfluss jeder einzelnen Beobachtung.

Statt eine Beobachtung wegzulassen, wollen wir nun eine hinzufügen und in der gleichen Art untersuchen, welchen Einfluss sie hat – in Abhängigkeit von ihren Werten für die Regressoren und die Zielgrösse. So wird beispielsweise in einer gewöhnlichen, einfachen Regression der Einfluss einer Beobachtung mit den Werten $[x,\, y]$ auf die geschätzte Steigung gleich

$$\Delta\beta = \widehat{\beta}^{+}\langle x, y\rangle - \widehat{\beta} = c \cdot \widetilde{x}(y - \beta x) \ ,$$

wobei $\widehat{\beta}^{+}\langle x, y\rangle$ die Schätzung von β aus dem um die potentielle Beobachtung $[x,\, y]$ erweiterten Datensatz bezeichnet, $\widetilde{x} = x - \bar{x}$ die Abweichung des x-Wertes der zusätzlichen Beobachtung vom Stichproben-Mittelwert bedeutet und $c^{-1} = \widetilde{x}^2 + \sum_i (x_i - \bar{x})^2$ ist.

Die Differenz ist also proportional zu $x - \bar{x}$ und $y - \widehat{\beta}x$ bis auf den Umstand, dass x auch in c vorkommt.

c ▹ Für das **Beispiel der Reissfestigkeit von Fasern** kann der Einfluss als Funktion von y für die vier untersuchten Faserlängen gezeigt werden (Abbildung 14.1.c). (Man kann ihn auch für andere Faserlängen ausrechnen.)
◃

d **Sensitivitäts-Kurve.** Diese Differenz ist natürlich umso kleiner, je grösser der Datensatz ist. Es ist deshalb sinnvoll, sie mit der Beobachtungszahl n zu multiplizieren, um ein Mass des Einflusses zu erhalten, das vom Umfang des Datensatzes nicht abhängt. Man definiert daher allgemein die Sensitivitäts-Kurve einer Schätzung eines Parameters θ als Funktion des potentiellen Beobachtungsvektors $\underline{x}$ durch

$$\mathrm{SC}\langle\underline{x}\rangle = (n+1)\left(\widehat{\theta}\langle\underline{x}, \underline{x}_1, \underline{x}_2, ..., \underline{x}_n\rangle - \widehat{\theta}\langle\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\rangle\right) \ .$$

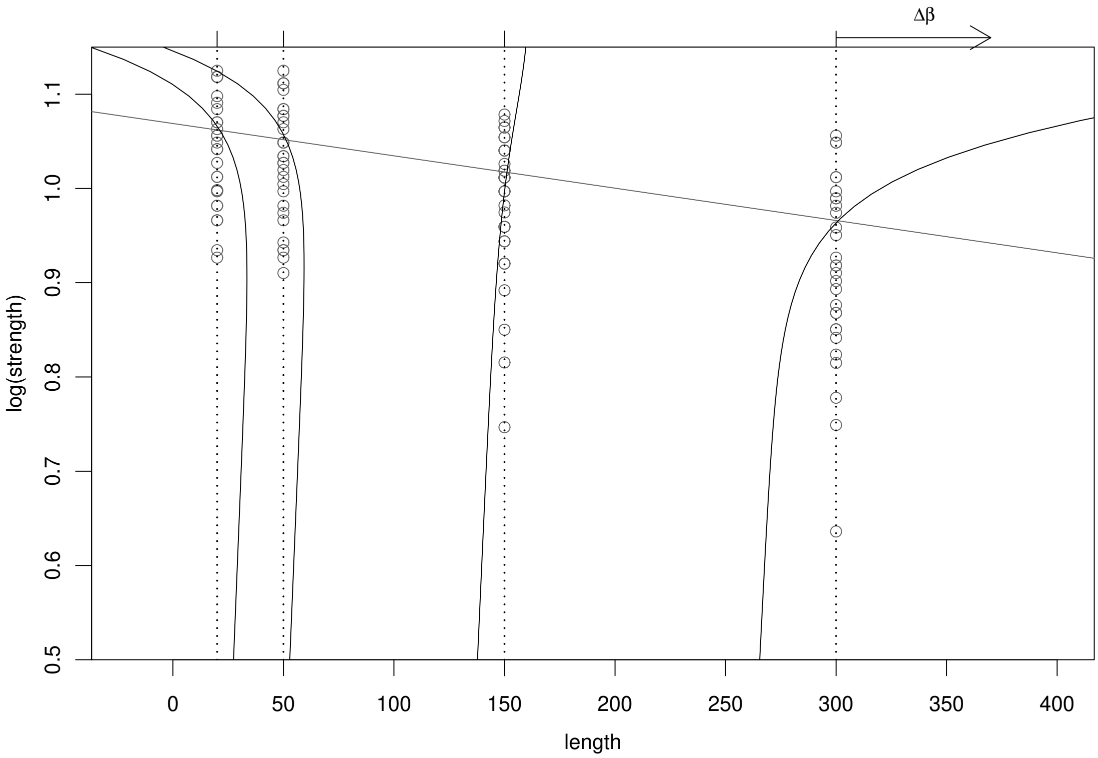

Abbildung 14.1.c: Veränderung der Steigung im Beispiel der Reissfestigkeit von Fasern als Funktion von y für die vier realisierten x-Werte. Die Kurven beziehen sich jeweils auf die entsprechende vertikale Linie und zeigen die Abweichungen als Funktion von y in horizontaler Richtung an, siehe Skala am oberen Rand.

e **Schätzungen für einen Lage-Parameter.** Um die Bedeutung dieser Überlegungen noch weiter zu verfolgen, wenden wir sie zunächst auf das wohl geläufigste Schätzproblem an: Es soll der Erwartungswert μ einer Normalverteilung geschätzt werden. Für das arithmetische Mittel wird

$$\begin{aligned} \mathrm{SC}\langle \underline{x}\rangle &= (n+1)\left(\tfrac{1}{n+1}(x+x_1+x_2+...+x_n)-\tfrac{1}{n}(x_1+x_2+...+x_n)\right) \\ &= x+\left(1-\frac{n+1}{n}\right)(x_1+x_2+...+x_n)=x-\frac{1}{n}\cdot n\bar{x} \\ &= x-\bar{x}\;. \end{aligned}$$

Es muss ja nicht das arithmetische Mittel sein. Wenn Ausreisser zu befürchten sind, verwendet man besser den Median oder ein gestutztes Mittel. Das letztere kommt zustande, indem man für einen festgelegten Stutzungsprozentsatz $\alpha \cdot 100\%$ die αn kleinsten und ebensoviele grösste Beobachtungen weglässt und von den übriggebliebenen das arithmetische Mittel berechnet. Abbildung 14.1.e zeigt die Sensitivitätskurven für diese Schätzungen, und zwar für das **Beispiel der Schlafdaten**, die aus den 10 Werten 1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4 bestehen (siehe W.Stahel, „Einführung in die statistische Datenanalyse“, 1.b).

Für das gestutzte Mittel bleibt der Einfluss begrenzt, wenn die Beobachtung gross wird. Wenn ein grober Fehler auftritt, beispielsweise eine Verschiebung des Dezimalpunktes beim Abschreiben, dann wirkt sich das nur beschränkt auf das gestutzte Mittel aus. Man könnte zunächst vermuten, dass so eine Beobachtung gar keinen Einfluss auf das Resultat hat,

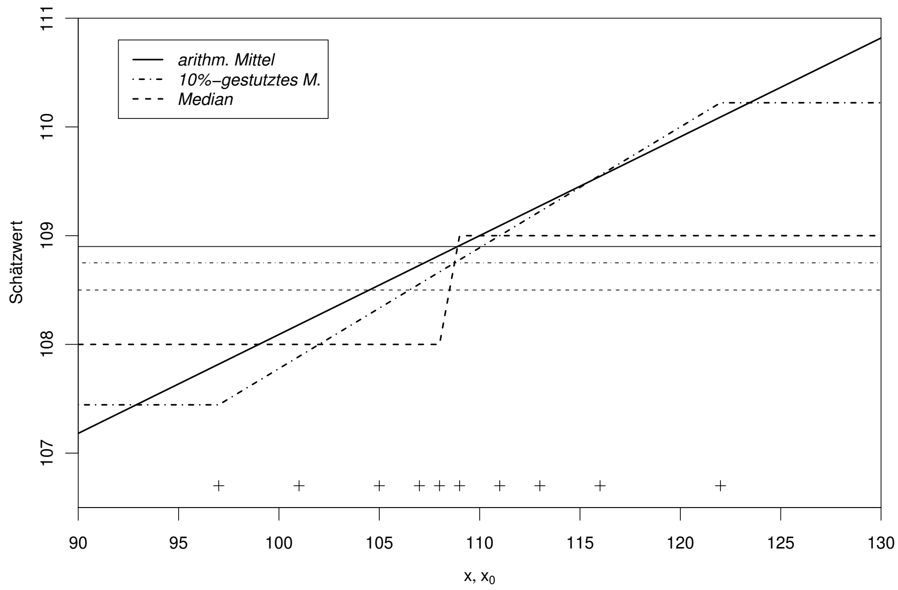

Abbildung 14.1.e: Sensitivitätskurven für drei Schätzungen einer Lage. Die horizontalen Linien sind jeweils auf der Höhe des Schätzwertes für die gegebene Stichprobe gezeichnet. [y-Skala noch zu korrigieren]

da sie ja zur Berechnung des arithmetischen Mittels der „inneren“ Beobachtungen nicht benützt wird. Wieso hat sie trotzdem einen Einfluss? Die Antwort überlassen wir Ihnen.

f Eine Beobachtung, die über dem arithmetischen Mittel liegt, hat einen positiven Einfluss, d.h., wenn man sie weglassen würde, würde das arithmetische Mittel kleiner. Mittelt man den „Einfluss“ über die Beobachtungen, so gilt offenbar

$$\tfrac{1}{n}\sum\nolimits_i \mathrm{SC}\langle x_i\rangle = \frac{1}{n}\sum\nolimits_i (x_i - \bar{x}) = \tfrac{1}{n}\sum\nolimits_i x_i - n\bar{x} = 0 \; .$$

Das gilt approximativ auch für andere Schätzungen: Die Summe der Einflüsse der einzelnen Beobachtungen ist null.

g **Einfluss-Funktion.** Die Sensitivitätskurve hängt offensichtlich von der Stichprobe $[x_1, x_2, ..., x_n]$ ab – wenn auch nur unwesentlich. Für die genaue mathematische Untersuchung von Eigenschaften einer Schätzung ist das ein Hindernis. Um es zu vermeiden, könnte man Stichproben gemäss einer Verteilung simulieren, jeweils die SC berechnen und dann die Kurven mitteln. Das heisst, man würde den Erwartungswert der Kurven bilden für eine gegebene Verteilung der Stichprobe. Wenn man mit einigen Begriffen der mathematischen Statistik vertraut ist, gibt es eine einfachere Definition, die im Wesentlichen das Gleiche macht, aber ausserdem einen Grenzwert für einen unendlichen Stichprobenumfang bildet. Diese „asymptotische Variante“ der Sensitivitäts-Kurve heisst Einfluss-Kurve oder Einfluss-Funktion. Sie hängt ab von der angenommenen Verteilung der Beobachtungen, die durch die kumulative Verteilungsfunktion F bestimmt ist, und charakterisiert das Schätz-Verfahren T. Wir können sie deshalb mit

$$\mathrm{IF}\langle \underline{x}; T, F\rangle$$

bezeichnen.

Die Eigenschaft des Mittels der Einflüsse wird für die IF sogar exakt:

$$\mathcal{E}\langle \mathrm{IF}\langle \underline{X}; T, F\rangle\rangle = \underline{0}$$

(wobei $\mathcal{E}$ den Erwartungswert für $X \sim F$ bedeutet).

h* Man braucht die folgenden Begriffe: Eine Schätzung ist ja eine Funktion T der Beobachtungen $x_1, x_2, ..., x_n$. Die Reihenfolge der Beobachtungen spielt in aller Regel keine Rolle. Aus der empirischen Verteilungsfunktion

$$\widehat{F}\langle x_1, x_2, ..., x_n\rangle = \tfrac{1}{n}\mathrm{Anzahl}\langle i \mid x_i \leq x\rangle$$

kann man die Beobachtungen bis auf ihre Reihenfolge zurückgewinnen. Deshalb kann man die Schätzung auch als Funktion $T\left\langle \widehat{F}\right\rangle$ der empirischen Verteilungsfunktion schreiben.

Und was soll das? Lieber kompliziert als einfach? – Wenn der Stichprobenumfang n immer grösser wird, nähert sich die empirische Verteilungsfunktion immer mehr der (theoretischen) kumulativen Verteilungsfunktion der angenommenen Vereilung an – und die „vernünftigen“ Schätzungen haben einen sinnvollen Grenzwert. Das einfachste Beispiel ist das arithmetische Mittel, das nach dem Gesetz der Grossen Zahl gegen den Erwartungswert der Verteilung konvergiert, der ja definiert ist als

$$E\langle X\rangle = \sum\nolimits_x xP\langle X = x\rangle$$

für diskrete Verteilungen. Fassen wir nun die empirische Verteilungsfunktion als Verteilungsfunktion einer diskreten Verteilung auf mit $P\langle X = x_i\rangle = 1/n$ (für Stichproben mit Bindungen muss man das etwas komplizierter aufschreiben), dann ist

$$\overline{X} = T\left\langle \widehat{F}\right\rangle \ .$$

Damit können wir das arithmetische Mittel T als Erwartungswert $T\langle F\rangle = E\langle F\rangle$ definieren und erhalten das Stichprobenmittel mit der gleichen Formel, wenn wir statt der theoretischen Verteilungsfunktion die empirische einsetzen – eben $\overline{X} = T\left\langle \widehat{F}\right\rangle$. Wir betrachten also nicht mehr Funktionen der Beobachtungen, sondern solche von Verteilungsfunktionen – theoretischen wie empirischen. Solche Funktionen heissen *Funktionale* (auf dem Raum der Verteilungsfunktionen).

Für kontinuierliche Verteilungen ist

$$E\langle X\rangle = \int_{-\infty}^{\infty} xf\langle x\rangle\ dx\ ,$$

wobei f die Dichte der Verteilung ist. Die Mathematik verallgemeinert nun auch das *Integral über eine Verteilung* so, dass die letzte Gleichung für diskrete Verteilungen als die oben aufgeschriebene Summe zu interpretieren ist. Man schreibt dann $E\langle X\rangle = \int_{-\infty}^{\infty} x\ dF\langle x\rangle$ für den allgemeinen Fall.

Diese ganze Maschinerie erlaubt es, Grenzübergänge einfacher zu betrachten: Es gilt nach dem Gesetz der Grossen Zahl, dass

$$T\left\langle \widehat{F}\right\rangle \to T\langle F\rangle\ ,$$

wenn der Stichprobenumfang $n \to \infty$ geht. Eine andere Version des gleichen Gesetzes sagt $\widehat{F} \to F$, also: Die empirische Verteilungsfunktion geht gegen die theoretische.

In der Differentialrechnung werden Funktionen $g\langle u\rangle$ *stetig* genannt, wenn folgendes gilt: Wenn eine Folge $u_1, u_2, ...$ von Argumentwerten einen Grenzwert u besitzt, geht $g\langle u_i\rangle \to g\langle u\rangle$. Im gleichen Sinne kann man Funktionale stetig nennen, wenn für $F_n \to F$ immer auch $T\langle F_n\rangle \to T\langle F\rangle$ gilt. Da $\widehat{F} \to F$ gilt, haben stetige Funktionale für immer grösser werdende Stichproben den „richtigen“ Grenzwert $T\langle F\rangle$.

Wenn wir nun schliesslich zur Sensitivitätskurve zurückkehren, dann müssen wir zunächst die „gemischte“ Stichprobe aus den Beobachtungen und der zugefügten Beobachtung x betrachten. Wir notieren das als „Mischverteilung“

$$\widehat{F}_n^* = \frac{n}{n+k}\widehat{F}_n + \frac{k}{n+k}\delta_x$$

mit $k = 1$, wobei δ_x die Verteilungsfunktion der „Verteilung" ist, die dem Wert x die Wahrscheinlichkeit 1 gibt. Die Sensitivitätskurve lässt sich dann schreiben als

$$\mathrm{SC}\langle x; T, x_1, ..., x_n\rangle = \left(T\left\langle(1-\varepsilon)\widehat{F}_n + \varepsilon\delta_x\right\rangle - T\left\langle\widehat{F}_n\right\rangle\right)\Big/\varepsilon$$

mit $\varepsilon = k/(n+k)$, $k = 1$. Nun machen wir den Grenzübergang $n \to \infty$ und erhalten

$$\mathrm{SC}\langle x; T, x_1, x_2, ...\rangle = (T\langle(1-\varepsilon)F + \varepsilon\delta_x\rangle - T\langle F\rangle)/\varepsilon$$

(Dabei muss k proportional zu n steigen, da sonst der Zähler und der Nenner nach 0 gehen.) Lässt man nun $\varepsilon \to 0$ gehen, dann erhält man die Definition der Einfluss-Funktion,

$$\mathrm{IF}\langle x; T, F\rangle = \lim_{\varepsilon\to 0}\left(T\langle(1-\varepsilon)F + \varepsilon\delta_x\rangle - T\langle F\rangle\right)/\varepsilon$$

Das ist eine Art Ableitung nach ε für die Stelle $\varepsilon = 0$.

Das waren viele neue Begriffe auf ein Mal. Sie können die Einflussfunktion auch anschaulich verstehen als Sensitivitätskurve für grosse Stichproben.

i **Gross Error Sensitivity, Robustheit.** Wenn Ausreisser zu befürchten sind, ist es wünschenswert, dass ihr Einfluss auf die Resultate beschränkt bleibt. Verfahren, die das leisten, werden als **robuste Verfahren** bezeichnet. Als quantitatives Mass für die Robustheit bietet sich deshalb der maximale Wert der Einflussfunktion an,

$$\gamma\langle T, F\rangle = \sup_{\underline{x}}\langle\,|\,\mathrm{IF}\langle\underline{x}; T, F\rangle\;|\,\rangle$$

genannt „Gross Error Sensitivity". (sup heisst supremum und ist der mathematisch präzise Ausdruck für das Maximum.)

Es gibt natürlich noch weitere Aspekte von Robustheit, die mit anderen Massen quantifiziert werden. Eines davon wird in 14.4.a beschrieben.

j **Ziel.** Da wir nun ein Mass für die Robustheit einer Schätzung haben, ist es, mindestens für mathematisch veranlagte Leute, naheliegend, dieses zu optimieren, also nach der robustesten Schätzung zu fragen. Man kann zeigen, dass der **Median** die Schätzung mit der kleinsten Gross Error Sensitivity für einen Lage-Parameter ist.

Allerdings ist auch bekannt, dass der Median ungenauer ist (grössere Varianz hat) als das arithmetische Mittel, wenn die Daten normalverteilt sind – aber das Mittel hat einen unbegrenzten Einfluss und ist also gar nicht robust. Man sollte daher nicht nur **Robustheit** oder nur **kleine Varianz** fordern, sondern einen „**optimalen Kompromiss**" zwischen den beiden Zielen. Eine mögliche Formulierung lautet:

> *Suche unter allen Schätzungen mit einer Gross Error Sensitivity, die kleiner als eine gegebene Schranke ist, diejenige, die die kleinste Varianz hat.*

Das ist das Optimalitätsproblem von Hampel (1974).

14.2 Robuste Schätzungen

a **M-Schätzung.** Die so genannten M-Schätzungen spielen in der Robusten Statistik eine grosse Rolle. Es sind im Wesentlichen Maximum-Likelihood-Schätzungen, aber sie werden nicht direkt mit einer Verteilungsannahme für die Daten verknüpft und etwas allgemeiner definiert.

Eine M-Schätzung für einen Parameter (-Vektor) $\underline{\theta}$ ist gegeben durch eine Funktion $\rho\langle\underline{x},\underline{\theta}\rangle$. Sie ist definiert als

$$\widehat{\underline{\theta}} = \arg\min_{\underline{\theta}}\left\langle\sum\nolimits_i \rho\langle\underline{x}_i,\underline{\theta}\rangle\right\rangle ,$$

also als Argument $\underline{\theta}$, das die Summe der $\rho\langle\underline{x}_i,\underline{\theta}\rangle$ minimiert. (Das ist nur gut definiert, wenn dieses Minimum existiert und eindeutig ist.)

Das ist die gleiche Definition wie diejenige der Maximum-Likelihood-Schätzung (13.1), ausser dass hier nicht vorausgesetzt wird, dass ρ die logarithmierte Dichte der Verteilung der Beobachtungen ist. Meistens kann man zwar zu einem sinnvollen ρ eine Verteilung finden, für die $\rho\langle\underline{x},\underline{\theta}\rangle = -\log\langle f\langle\underline{x},\underline{\theta}\rangle\rangle$ gilt, aber **wir lösen uns von der Idee, dass die Beobachtungen genau dieser Verteilung folgen.**

Die Minimierung hat dann eine eindeutige Lösung, wenn $\sum_i \rho\langle\underline{x}_i,\underline{\theta}\rangle$ eine konvexe Funktion von θ ist. Das ist zwar nicht identisch mit der Forderung, dass ρ selbst konvex ist bezüglich θ, aber diese Eigenschaft ist doch sehr nützlich für die Eindeutigkeit der Schätzwerte.

b **Nullstellen-Form.** Wie für die Maximum-Likelihood-Schätzungen kann man statt der Minimierung die Gleichung lösen, die durch Ableiten und Null-Setzen zustande kommt,

$$\sum\nolimits_i \underline{\psi}\langle\underline{x}_i,\underline{\theta}\rangle = \underline{0} , \qquad \underline{\psi}\langle\underline{x},\underline{\theta}\rangle = \frac{d}{d\theta}\,\rho\langle\underline{x}_i,\underline{\theta}\rangle .$$

Normalerweise lässt sich diese Gleichung nicht explizit nach θ auflösen, und man hat ein Minimierungsproblem durch eine Nullstellen-Suche ersetzt, was einfacher sein kann, aber nicht muss.

c **M-Schätzung für einen Lage-Parameter.** Von einer sinnvollen Schätzung eines Lage-Parameters wird man erwarten, dass sie sich um Δ vergrössert, wenn man zu jeder Beobachtung Δ hinzuzählt,

$$\widehat{\mu}\langle x_1+\Delta, x_2+\Delta, ..., x_n+\Delta\rangle = \widehat{\mu}\langle x_1, x_2, ..., x_n\rangle + \Delta .$$

Wenn das für M-Schätzungen gelten soll, dann darf ρ nur von der Differenz $x-\mu$ abhängen. Also bestimmt man die Schätzung als

$$\widehat{\mu} = \arg\min_{\mu}\left\langle\sum\nolimits_i \rho\langle x_i-\mu\rangle\right\rangle$$

oder als Lösung der Gleichung

$$\sum\nolimits_i \psi\langle x_i-\mu\rangle = 0 , \qquad \underline{\psi}\langle r\rangle = \rho'\langle r\rangle .$$

($'$ bezeichnet die gewöhnliche Ableitung.) Wenn ρ konvex ist, so ist ψ monoton nicht-abnehmend (und umgekehrt).

Es lässt sich leicht zeigen, dass die Sensitivity Curve dieser Schätzung eng mit der ψ-Funktion verbunden ist,

$$\mathrm{SC}\langle x\rangle \approx c\cdot\psi\langle x-\widehat{\mu}\rangle , \qquad c^{-1} = \sum\nolimits_i \psi'\langle x_i-\widehat{\mu}\rangle .$$

Die Einfluss-Funktion ist sogar exakt proportional zu ψ:

$$\mathrm{IF}\langle x; \widehat{\mu}, F\rangle = c \cdot \psi\langle x - \widehat{\mu}\rangle \ , \qquad c^{-1} = \mathcal{E}\big\langle \psi'\langle X - \widehat{\mu}\rangle\big\rangle$$

(wobei $\mathcal{E}$ wieder den Erwartungswert für $X \sim F$ bezeichnet).

d **ψ-Funktionen.** Dieses Resultat ermöglicht es, mit der Wahl der ψ-Funktion direkt festzulegen, wie der Einfluss als Funktion der Differenz $r = x - \mu$ aussehen soll. Für das arithmetische Mittel ist er ja proportional zu r. Da das für normalverteilte Daten optimal ist und die Verteilung der meisten Datensätze „in der Mitte" recht gut mit der Normalverteilung übereinstimmt, ist es sinnvoll, die ψ-Funktion für kleine bis mittlere r gleich r zu setzen. Für extremere Beobachtungen soll davon abgewichen werden, damit diese keinen ungebührlich grossen Einfluss erhalten. Abbildung 14.2.d zeigt einige Möglichkeiten:

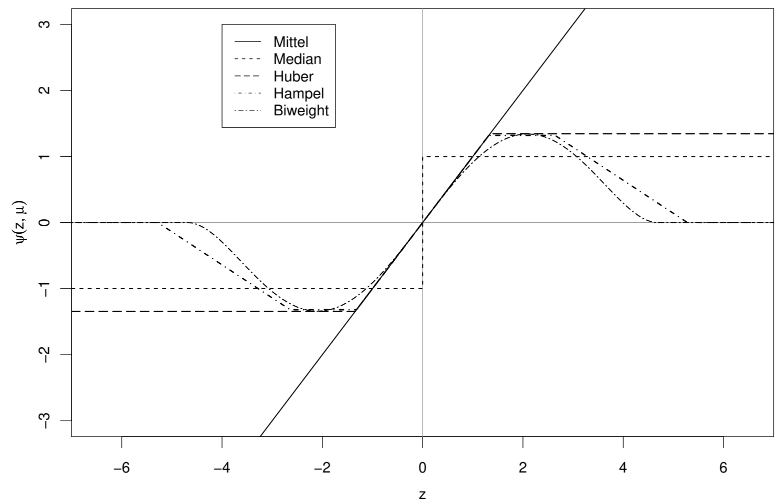

Abbildung 14.2.d: Gebräuchliche ψ-Funktionen

- Wir können den Einfluss „abschneiden" mit der Funktion

$$\psi\langle r\rangle = \min\langle\max\langle r, -c\rangle, c\rangle = \begin{cases} -c & r < -c \\ r & -c \le r \le c \\ c & r > c \end{cases} \ .$$

  Dabei ist c eine „Wahlkonstante", die es erlaubt, den Einfluss mehr oder weniger stark zu beschränken. Diese Funktion heisst in der Statistik **Huber-Funktion** nach Peter Huber, der sie in seiner grundlegenden Arbeit über Robuste Statistik eingeführt hat und der etliche in Zürich gewirkt hat.

- Was geschieht für kleine Wahlkonstanten c? Der Maximalwert der Funktion wird ebenfalls klein. Wir können aber die Funktion mit irgendeiner Konstanten multiplizieren, ohne dass die resultierende Schätzung sich ändert, wie man der Gleichung 14.2.b ansieht. Multipliziert man das vorhergehende ψ mit $1/c$ und lässt c gegen 0 gehen, dann erhält man die „Vorzeichen-Funktion" $\psi\langle r\rangle = 1$ für $r > 0$ und $= -1$ für $r < 0$. Sie ist in der Abbildung mit „Median" bezeichnet, da man sich leicht überzeugen kann, dass die entsprechende Schätzung gleich dem Median ist. (Das gilt für ungerade Beobachtungszahl, bei gerader Zahl löst jeder Wert zwischen den beiden mittleren Beobachtungen die Gleichung 14.2.b, und man muss die Lösung auf geeignete Art eindeutig machen.) Der Median ist also ein Extremfall der Huber-Schätzung. Am anderen Ende der Skala, für $c \to \infty$, erhalten wir $\psi\langle r\rangle = r$, also das arithmetische Mittel.

- Man kann auch dafür sorgen, dass klare Ausreisser keinen Einfluss auf die Schätzung haben, indem man die ψ-Funktion von einem bestimmten r weg null setzt. Ein solcher Wert wird „**Verwerfungs-Punkt**“ oder **rejection point** genannt. Da es von Vorteil ist, wenn die ψ-Funktion keine Sprungstellen aufweist und auch nirgends zu steil ist, hat **Hampel** (1974) die folgende „**three part redescending**“ Funktion eingeführt:

$$\psi\langle r\rangle = \begin{cases} r & |r| \le c \\ \operatorname{sign}\langle r\rangle \; c & c < |r| \le b \\ \operatorname{sign}\langle r\rangle \; (c - (|r| - c)/(d - c) & b < |r| < d \\ 0 & |r| > d \end{cases} .$$

Sie enthält drei Wahlkonstanten $c \le b < d$ und ist am einfachsten von der grafischen Darstellung her zu verstehen.

Eine populäre ψ-Funktion mit Ausreisser-Verwerfung ist die „bisquare“- oder **„bi-weight“-Funktion**, die J.W. Tukey erfunden hat. Ihre Formel lautet

$$\psi\langle r\rangle = \begin{cases} r\left(\left(1 - \frac{r}{c}\right)^2\right)^2 & |r| \le c \\ 0 & |r| > c \end{cases} .$$

- Die ψ-Funktionen, die der **t-Verteilung** entsprechen, wurden schon in 13.1.k besprochen und grafisch dargestellt. Sie fallen für grosse r ab, aber nur wenn die Zahl ν der Freiheitsgrade klein ist, geschieht das im Bereich, in dem überhaupt Daten zu erwarten sind, und auch dann kommt die Funktion nicht so bald in die Nähe von 0.

e **Zusammenhang zwischen Maximum-Likelihood und M-Schätzungen.** Wenn man für die Beobachtungen eine langschwänzige Verteilung annimmt, beispielsweise eine t-Verteilung mit 5 Freiheitsgraden, dann wird für die Maximum-Likelihood-Schätzung die ψ-Funktion und damit die Einflussfunktion beschränkt und die Schätzung in diesem Sinne robust.

Allerdings stellt sich die Frage, wie man die Wahl der Verteilung rechtfertigen soll. Wieso 5 Freiheitsgrade? Eine Möglichkeit besteht darin, die Freiheitsgrade offen zu lassen und aus den Daten zu schätzen. Das bewährt sich nur für grosse Stichproben, da dadurch die Schätzung der Standardfehler (s. unten) variabler wird.

Eine Änderung der Haltung führt hier weiter: Wir bleiben bei einem einfachen, allenfalls optimistischen Modell betreffend die Ausreisser, aber wir stellen sicher, dass der Einfluss einzelner Beobachtungen auf die Schätzung beschränkt bleibt, indem eine M-Schätzung mit beschränkter ψ-Funktion verwendet wird. Wir verzichten damit bewusst auf ein wenig Effizienz (denn die Maximum-Likelihood-Schätzung ist ja asymptotisch die effizienteste) und sichern uns damit gegen übermässige Einflusswerte ab.

Das entspricht dem Gedanken einer **Versicherung**, mit der Verfälschung des Resultates durch Ausreisser als Risiko und dem Effizienzverlust als Prämie.

f **Tests und Vertrauensintervalle.** Für M-Schätzungen gilt der Zentrale Grenzwertsatz; sie sind also näherungsweise normalverteilt. Damit müssen wir nur den (oder die) Standardfehler bestimmen und erhalten die üblichen Tests und Vertrauensintervalle mit den „Plus-Minus-Zwei-Standardfehler-Regeln".

Wie der Standardfehler berechnet wird, werden wir in einem anderen Block diskutieren. Als Verteilungsannahme dient meist die Normalverteilung! Das scheint etwas schizophren, aber es ist ungefährlich, weil für robuste Schätzungen auch der Standardfehler meist „automatisch" robust herauskommt, und sogar die Verteilung der Schätzwerte besser einer Normalverteilung folgt als die Verteilung der nicht-robusten Schätzungen.

* Die letzte Aussage muss man genauer fassen. Wenn die Beobachtungen exakt normalverteilt sind, kann sie nicht stimmen, da ja dann der Mittelwert auch exakt normalverteilt ist – und die robusten Schätzungen sind es nicht. Aber sobald die „wahre Verteilung" leicht langschwänzig ist, dreht sich das ins Gegenteil.

g **Wahl der ψ-Funktion.** Die Flexibilität der M-Schätzungen erlaubt es auch, das **Optimalitätsproblem** von Hampel (14.1.j) zu lösen. Die optimale M-Schätzung ist gegeben durch die Huber-Funktion. Die Wahlkonstante muss so gewählt werden, dass die gewünschte Schranke für die Gross Error Sensitivity gerade eingehalten wird. Das geht allerdings nur, wenn man nicht zu viel verlangt. Der Median hat die kleinste Gross Error Sensitivity, die für die Schätzung eines Lageparameters erreicht werden kann.

Wenn man zusätzlich eine **Ausreisser-Verwerfung** wünscht, wird man eine entsprechende ψ-Funktion wählen.

h* Wenn man nochmals Abbildung 14.1.e betrachtet, fällt auf, dass die Sensitivitätskurve des gestutzten Mittels eine (verschobene) Huber-Funktion ist. Die Einfluss-Funktion dieser Schätzung ist genau identisch mit der einer M-Schätzung mit Huber-ψ-Funktion. Da die Optimalität nur mit asymptotischen Eigenschaften zu tun hat und ein arithmetisches Mittel für grosse Stichproben immer genauer mit der ensprechenden M-Schätzung übereinstimmt, löst auch ein gestutztes Mittel das Optimalitätsproblem.

M-Schätzungen bilden eine flexible Klasse, die es erlaubt, gewünschte Eigenschaften der Einfluss-Funktion direkt zu wählen. Es gibt aber auch andere Klassen von robusten Schätzungen, beispielsweise

- die R-Schätzungen, die auf den Rängen der Beobachtungen beruhen,
- die L-Schätzungen, die die geordneten Beobachtungen benützen, wie dies das gestutzte Mittel tut.

i **Wahlkonstanten.** Die Wahlkonstanten steuern den Kompromiss zwischen Robustheit und statistischer Genauigkeit (Effizienz). Da die Zahl 5% von der Irrtums-Wahrscheinlichkeit von Tests her für Statistiker offenbar etwas Magisches hat, werden die Wahlkonstanten oft so gesetzt, dass die (asymptotische) Effizienz der Schätzung gegenüber dem arithmetischen Mittel 95% beträgt, wenn die Daten der Normalverteilung folgen. Das führt zu $c = 1.345$ für die Huber-Funktion.

14.3 M-Schätzung für Regression

a ▷ **Beispiel NO_2-Mittelwerte.** Kehren wir zur Regression zurück. Zunächst ein Beispiel, das zeigt, wofür robuste Methoden in der Regression gebraucht werden.

Schadstoffe in der Luft werden durch automatische, fest installierte Messstationen im Stundentakt gemessen. Wie gut können für Orte zwischen den Stationen die Werte geschätzt werden? Für längerfristige Mittelwerte gibt es physikalische Modelle und billige Messmethoden. Diese können dadurch flächendeckend oder wenigstens für ein wesentlich dichteres Netz von Messpunkten bestimmt werden. Man kann sie dann verwenden, um mittels Regression auch kurzfristige Werte zu schätzen, beispielsweise Tagesmittelwerte als Zielgrösse aus Jahresmittelwerten als Prädiktor. Grundlage dazu bildet der Zusammenhang zwischen diesen Grössen, der aus den Daten der Messstationen der Region ermittelt wird – möglichst spezifisch für einzelne Wetterlagen.

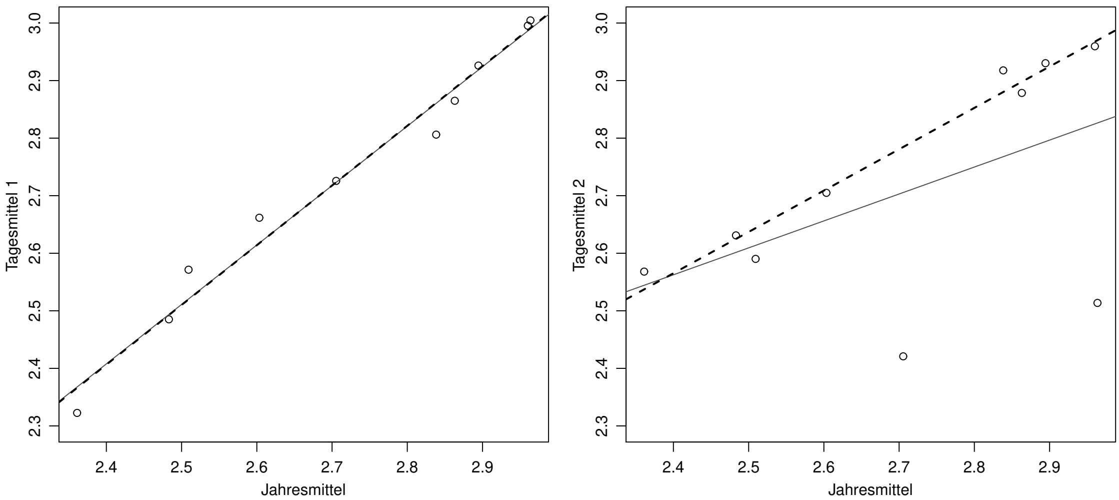

Abbildung 14.3.a: Kleinste-Quadrate- (——) und robuste Regression im Beispiel der NO_2-Mittelwerte

Die beiden Diagramme in Abbildung 14.3.a zeigen den Zusammenhang des Jahresmittelwerts und des Tagesmittelwerts für zwei verschiedene Tage. Im zweiten Fall gibt es offensichtlich zwei Beobachtungen, die dem üblichen Zusammenhang nicht folgen. Eine gewöhnliche Regression führt zu einer „Kompromiss-Geraden“, die robust geschätzte Gerade zeigt den Zusammenhang recht unverfälscht. Wenn man die Residuen für die robuste Variante betrachtet, dann sind sie so klein wie für den ersten Tag, bis auf die zwei Beobachtungen, die nicht passen, und die damit noch etwas klarer als Ausreisser identifiziert werden. ◁

b **M-Schätzungen.** Gemäss dem oben erwähnten Gedanken (14.2.a) minimieren wir für die Schätzung der Koeffizienten einer linearen Regression die Summe $\sum_i \rho\left\langle (Y_i - \underline{x}_I^T \underline{\beta})/\sigma \right\rangle$ (siehe 13.1) mit einer ρ-Funktion, die zu einem beschränkten Einfluss führt. Nach 13.1.k können wir auch die entsprechenden Normalgleichungen lösen.

c **Einflussfunktion von M-Schätzern.** Die Einflussfunktion für die Regressionskoeffizienten ist, analog zum Lage-Problem, proportional zur ψ-Funktion, aber zusätzlich auch zum Vektor der Regressoren $\underline{x}$,

$$\mathrm{IF}\left\langle \underline{x}, y; \underline{\widehat{\beta}}, F\right\rangle = \psi\left\langle \frac{Y - \underline{x}^T\underline{\widehat{\beta}}}{\sigma}\right\rangle (\widetilde{\kappa}\, \boldsymbol{C})^{-1}\underline{x} \ ,$$

wobei die Matrix $\boldsymbol{C}$ die Kovarianzmatrix der Regressoren ist und die Konstante $\widetilde{\kappa} = \int \psi'\langle u\rangle\, f_1\langle u\rangle\, du$.

d* **Asymptotik für Regression.** Die Einfluss-Funktion und später die Angabe einer Verteilung der Schätzungen beruhen auf einer asymptotischen Betrachtung. Wie soll ein Datensatz, der mit Regression modelliert wird, wachsen? Welche weiteren Versuchsbedingungen sollen gewählt werden oder wie werden die Werte der Eingangsgrössen sein, wenn wir ja nicht annehmen, dass dies zufällig seien? Es gibt zwei Möglichkeiten:

- Wir fassen die Eingangsgrössen für diese Betrachtung doch als Zufallsgrössen auf. Eine spezielle (gemeinsame) Verteilung müssen wir nicht spezifizieren.
- Wir machen Asymptotik in grösseren Schritten und denken uns den Datensatz verdoppelt, dann verdreifacht und so weiter, indem alle x-Werte wiederholt werden und neue Beobachtungen von Y hinzu kommen.

Man kann asymptotische Resultate auch noch unter etwas allgemeineren Bedingungen herleiten.

e **Gross Error Sensitivity.** Wenn ψ beschränkt ist, ist der Einfluss für gegebenes $\underline{x}$ auch beschränkt. Aber wenn $\underline{x}$ auch gross werden kann, wächst der Einfluss, wie man in der Formel sieht, unbegrenzt. Wenn also auch Ausreisser in den Eingangsvariablen möglich sind, genügt es nicht, ψ als beschränkt zu wählen, um die Gross Error Sensitivity zu beschränken.

f* **Verallgemeinerte M-Schätzung.** Wir brauchen also eine allgemeinere Idee. Wenn wir zulassen, dass die ψ-Funktion auch vom Vektor $\underline{x}$ der Eingangsgrössen abhängt, dann gelingt die Begrenzung des gesamten Einflusses. Eine verallgemeinerte M-Schätzung ist definiert als Lösung von

$$\sum\nolimits_i \eta\left\langle \frac{Y_i - \underline{x}_i^T\underline{\widehat{\beta}}}{\sigma}, \underline{x}_i\right\rangle \underline{x}_i = \underline{0}$$

Meist hängt $\eta\langle r_i, \underline{x}_i\rangle$ von $\underline{x}_i$ über die „leverage“ h_i ab, die misst, wie weit die Gesamtheit der Eingangsgrössen für die $i$te Beobachtung von deren Mittelwerten entfernt ist – und dabei Zusammenhänge unter den Eingangsgrössen berücksichtigt (siehe 4.3.g und 4.3.h). Da diese wiederum von Ausreissern in den Eingangsgrössen stark beeinflusst sein kann, sollten auch dafür robuste Alternativen verwendet werden.

Zwei übliche Varianten solcher η-Funktionen sind

- der Mallows-Typ: Die Abhängigkeit von $\underline{x}_i$ wird durch eine Gewichtung ausgedrückt: $\eta\langle r_i, \underline{x}_i\rangle = \omega\langle h_i\rangle\, \psi\langle r_i\rangle$. Oft wird für ψ die Huber-Funktion und für ω die entsprechenden Gewichte $\psi\langle h_i\rangle / h_i$.
- der Schweppe-Typ: Man verkleinert die Wahlkonstante der ψ-Funktion für grosse h_i, $\eta\langle r_i, \underline{x}_i\rangle = \psi_{c\langle h_i\rangle}\langle r_i\rangle$, wobei wieder ψ_c die Huber-Funktion sein kann – und $c\langle h_i\rangle$ ebenfalls.

Die verallgemeinerten M-Schätzungen haben einen gesamthaft beschränkten Einfluss – also auch für Ausreisser in den Eingangsgrössen –, wenn man die Funktion η geeignet wählt und die h_i robust bestimmt.

g **Verteilung der Schätzungen, Tests und Vertrauensintervalle.** Wie für die Verallgemeinerten Linearen Modelle und für Maximum-Likelihood-Schätzungen im Allgemeinen kann man auch für die M-Schätzungen eine asymptotische Normalverteilung herleiten. Da die Schätzung nicht mehr die Maximum-Likelihood-Schätzung ist, die der angenommenen Verteilung der Beobachtungen entspricht, wird der Ausdruck für den Faktor κ in der Kovarianz-Matrix etwas komplizierter als in 13.1.n,

$$\kappa = \frac{\int \psi^2 \langle u \rangle \, f_1 \langle u \rangle \, du}{\left(\int \psi' \langle u \rangle \, f_1 \langle u \rangle \, du\right)^2}$$

Es gibt einige Vorschläge, um diese Näherung an die Kovarianzmatrix für gegebene Beobachtungszahl n oder sogar für gegebene Matrix der Regressoren zu verbessern. Besonders soll wieder die Verwendung einer unrobusten Matrix $\boldsymbol{C}$ vermieden werden. In den verschiedenen packages von R zur robusten Statistik sind verschiedene solche Korrekturterme implementiert. Die Untersuchungen dazu sind noch immer nicht abgeschlossen.

Aus der genäherten Verteilung von $\underline{\widehat{\beta}}$ ergeben sich in der üblichen Weise Tests und Vertrauensintervalle (vergleiche 13.1.o).

* Für Verallgemeinerte M-Schätzungen wird die Kovarianzmatrix noch einiges komplizierter, siehe Maronna, Martin and Yohai (2006).

14.4 Bruchpunkt und weitere Schätzmethoden

a **Bruchpunkt.** Die Einfluss-Funktion zeigt den Effekt einer einzelnen zusätzlichen Beobachtung zu einer (grossen) Stichprobe auf eine Schätzung, und die Gross Error Sensitivity gibt an, wie gross dieser Effekt für Ausreisser werden kann. Nun soll untersucht werden, was passiert, wenn mehrere „wilde" Beobachtungen dazu kommen.

Wir gehen von einer Stichprobe $x_1, x_2, .., x_n$ aus und fügen weitere q beliebige Werte $x_1^*, x_2^*, ..., x_q^*$ als Beobachtungen hinzu. Nun fragen wir, wie gross die dadurch verursachte Abweichung, der „Bias" $|\, T\left\langle x_1, x_2, ..., x_n, x_1^*, x_2^*, ..., x_q^*\right\rangle - T\langle x_1, x_2, ..., x_n\rangle \,|$ werden kann. Wir müssen uns damit zufrieden geben, dass dieser Bias wenigstens beschränkt bleibt, also auch dann nicht gegen unendlich geht, wenn die Ausreisser dorthin wandern. Deshalb führen wir das folgende weitere Mass für Robustheit ein:

Der **(empirische) Bruchpunkt** ist $q/(n+q)$ wobei q die maximale Anzahl zusätzlicher Beobachtungen ist, für die der Bias $|\, T\left\langle x_1, x_2, ..., x_n, x_1^*, x_2^*, ..., x_q^*\right\rangle - T\langle x_1, x_2, ..., x_n\rangle \,|$ beschränkt bleibt.

Für grössere q kann der Betrag beliebig gross werden, und man spricht vom „**Zusammenbruch**" der Schätzung.

Im Prinzip hängt der empirische Bruchpunkt von der Stichprobe ab. Häufig ist das aber dann doch nicht so. Beispielsweise ist anschaulich klar, dass das 10% **gestutzte Mittel** zusammenbricht, wenn der Anteil beliebig falscher Beobachtungen mehr als 10% beträgt, und zwar unabhängig von den gegebenen Beobachtungen $x_1, x_2, .., x_n$. (Genauer: Für ein festes $n+q$ kann der Bruchpunkt nur ein Vielfaches von $1/(n+q)$ sein. Der empirische Bruchpunkt ist das kleinste Vielfache, das $\leq 10\%$ ist.)

* Es gibt auch einen asymptotischen Bruchpunkt, den wir hier nicht formell einführen wollen. Er hat, wie die Einflussfunktion, den Vorteil, dass er nicht von der konkreten Stichprobe abhängt, dafür aber von einer angenommenen Verteilung.

b **Bruchpunkt vom M-Schätzungen.** Es zeigt sich, dass für M-Schätzungen in der Regression der Bruchpunkt kleiner als $1/p$ ist, wobei p die Anzahl Koeffizienten bedeutet. Das ist für $p > 10$ nicht gerade viel, vor allem, wenn man bedenkt, dass ein grober Fehler in einer der p Eingangsgrössen oder in der Zielgrösse eine ganze Beobachtung zum Ausreisser macht. Deshalb hat man nach Schätzungen mit höherem Bruchpunkt gesucht.

c **M-Schätzung mit abfallendem ψ.** Es ist anschaulich klar, dass die M-Schätzungen eigentlich robust sein müssten, wenn man eine ψ-Funktion mit Verwerfung wählt (14.2.d), da diese ja den Einfluss von starken Ausreissern völlig eliminieren. In der Tat gibt es für die Gleichung 14.2.b dann in der Regel eine Lösung, die die Ausreisser „richtig" identifiziert – aber eben auch andere; die Gleichung definiert die M-Schätzung dann nicht eindeutig. Je nach Startpunkt für einen Algorithmus zur Berechnung der Schätzung wird man bei verschiedenen Lösungen landen. Wenn man einen geeigneten – robusten – Startpunkt wählt, erhält man die „richtige" Lösung! Aber damit sind wir noch nicht weiter: Wir brauchen eine Schätzung mit hohem Bruchpunkt!

d **S-Schätzungen.** Generell ist ja eine Regressionsfunktion dann an die Daten gut angepasst, wenn die Residuen klein sind. Das kann man so formalisieren, dass eine *robuste* Skalenschätzung der Residuen möglichst klein sein soll.

Eine robuste Skalen-M-Schätzung ist gegeben als Auflösung der impliziten Gleichung

$$\frac{1}{n-p} \sum_i \rho\left\langle \frac{r_i}{s} \right\rangle = \kappa$$

nach s. Dabei ist κ durch die Funktion ρ bestimmt und sorgt dafür, dass für normalverteilte Beobachtungen asymptotisch die Standardabweichung σ als Lösung herauskommt. Damit dies eine robuste Schätzung ergibt, muss ρ beschränkt sein.

Da die Residuen $r_i = y_i - \underline{x}_i^T \underline{\beta}$ von $\underline{\beta}$ abhängen, ist die Lösung s eine Funktion von $\underline{\beta}$. Die S-Schätzung ist nun gegeben durch die Minimierung von $s\langle \underline{\beta} \rangle$ über $\underline{\beta}$.

e Man kann sich vorstellen, dass die Berechnung einer solchen Schätzung nicht einfach ist. Die Funktion $s\langle \underline{\beta} \rangle$ hat unangenehme Eigenschaften. Der Rechenaufwand wächst mit der Dimension p so rasch, dass man auf Verfahren zurückgreifen muss, die die Lösung nicht sicher finden, sondern nur mit einer bestimmten Wahrscheinlichkeit. Je nach Dimension und Computer-Kapazität kann man wenigstens diese Wahrscheinlichkeit hoch ansetzen – oder auch nicht.

f **MM-Schätzungen.** Ein weiterer Nachteil der S-Schätzungen ist ihre statistische Ineffizienz: Die Schätzungen sind immerhin asymptotisch normalverteilt, aber mit einer wesentlich grösseren Varianz (resp. Kovarianzmatrix) als die M-Schätzungen.

Eine nahe liegende Idee ist die Kombination von S- und M-Schätzungen, so dass von beiden Verfahren die Vorteile ausgenützt werden. Wenn wir die Idee von oben (14.4.c) aufgreifen, dann gelangen wir zur Klasse der MM-Schätzungen: Man berechnet eine S-Schätzung und verwendet sie als Startpunkt für eine M-Schätzung mit Ausreisser-Verwerfung.

g **Fortsetzung folgt.** In diesem Kapitel haben wir eine Einführung in die Robuste Statistik und die Robuste Regression gegeben. Genaueres folgt im Weiterbildungslehrgang in den Blöcken über Resampling und Asymptotik und über Robuste Statistik.

L Literatur zu robusten Methoden

a Zur Robusten Statistik gibt es ein sehr empfehlenswertes neues Buch von Maronna et al. (2006). Grundlegende Bücher sind Huber and Ronchetti (2009) (erste Auflage 1981) und Hampel, Ronchetti, Rousseeuw and Stahel (1986). Die Grundsteine zu diesen Büchern findet man in Huber (1964) und Hampel (1974).