Juliane Hiltraud KÖRBLER, BSc

Quantile Regression

Eine Anwendung auf Versicherungsleistungsdaten

MASTERARBEIT
zur Erlangung des akademischen Grades einer
Diplom-Ingenieurin

Masterstudium Finanz- und Versicherungsmathematik

Technische Universität Graz

Betreuer:
Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig Friedl
Institut für Statistik

Graz, März 2014

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am _____ _____

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Zusammenfassung

Die folgende Arbeit beschäftigt sich mit *Quantiler Regression* (QR). Diese Art der Modellierung von Zusammenhängen zwischen Response- und Prädiktorvariablen stellt eine sehr gute Alternative zu herkömmlicher Regression dar. Auf restriktive Annahmen, wie beispielsweise eine bestimmte Verteilung für die Fehlerterme, kann bei dieser Modellklasse verzichtet werden. Dazu liefert die Schätzung der bedingten Quantilsfunktion der Responseverteilung gegeben einer Prädiktormenge ein ausführlicheres Bild über die Verteilung der abhängigen Variable, als dies bei Erwartungswertmodellen der Fall ist. In dieser Arbeit liegt das Hauptaugenmerk auf den zugrundeliegenden theoretischen Überlegungen dieser Modellklasse. Wichtige Punkte sind dabei die Formulierung des QR-Problems als Lineares Programm und die daraus ableitbaren Eigenschaften des resultierenden Schätzers. Weiters wird im Detail auf Inferenz und in Verbindung damit auf die asymptotische Verteilung des Schätzers eingegangen. Auch die Robustheit gegenüber Ausreißern wird näher erläutert und mit jener von Erwartungswertschätzern verglichen. Während zu Beginn der Arbeit einige einfache Beispiele zur Illustration gezeigt werden, liefert ein ausführliches Anwendungsbeispiel über einen realen Datensatz der Klientel einer privaten Krankenversicherung einen Einblick in die praktische Umsetzung der zuvor erarbeiteten Methoden.

Abstract

The following paper deals with *Quantile Regression* (QR), which states a good alternative to common regression methods for modelling relationships between responses and predictors. There is no need for such restrictive assumptions like in classical regression models, for example choosing a designate error distribution. Another advantage is the more complete view of the whole distribution one gets by estimating the conditional quantile functions of the response distribution, given some set of predictors. In this master thesis the main part is taking a deeper look into the basic theoretical principals of this model class. Principal points are the reformulation of the QR-problem into a linear program and the deducible properties of the resulting estimator. Another important part is inference and therefore computing the estimator's asymptotic distribution. Also the robustness against outliers will be treated and compared to mean model estimation. While looking at some simple examples for illustration in the beginning, there will be a detailed application of the QR-methods, wich deals with a real dataset containing observations for the clientele of a private health insurance coverage.

Inhaltsverzeichnis

1. Motivation

Um Zusammenhänge zwischen der *Response* und unterschiedlichen *Prädiktoren* zu analysieren, verwendet man Methoden der *Regression*. Dabei gibt es viele verschiedene Möglichkeiten, solche Modelle zu erstellen und deren Parameter zu schätzen.

Die einfachste Form der Analyse solcher Zusammenhänge ist die klassische *Lineare Regression*. Eine detailierte Betrachtung dieser Methoden findet sich in Fahrmeir et al. (2013).

Zur Erstellung solcher Modelle trifft man einige fundamentale Annahmen:

1. Für jeden einzelnen Wert des *Prädiktors* x ist die *Response* Y eine Zufallsvariable und ihr Erwartungswert hängt von x ab.
2. Der Erwartungswert der abhängigen Variable Y lässt sich als deterministische Funktion in x schreiben.

Eine typische Situation sieht dann wie folgt aus:

Gegeben sind Daten $(x_{i1}, x_{i2}, \ldots, x_{i,p-1}, Y_i)$ für $i = 1, \ldots, n$. Ein **Multiples Lineares Regressionsmodell** (MLR) hat die Form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,$$

wobei für den *statistischen Fehler* ε_i angenommen wird, dass

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

gilt. Die Parameter $\beta_0, \beta_1, \ldots, \beta_{p-1}$, sowie σ^2 müssen in weiterer Folge geschätzt werden, während die $x_{i1}, \ldots, x_{i,p-1}$ bekannte Konstanten sind.

In Vektor- beziehungsweise Matrixschreibweise ergibt sich das MLR durch

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Dabei steht $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ für den *Responsevektor*, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$ repräsentiert den *Parametervektor* und $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ bezeichnet den *Fehlervektor*. Die $(n \times p)$-

Matrix $\mathbf{X}$ heißt *Designmatrix* des Modells und hat die Form

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}.$$

Mit $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$ folgt für den Responsevektor $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n)$.

Nun sollen die unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 geschätzt werden, um das Modell an die Daten anzupassen.

Um einen Schätzer $\hat{\boldsymbol{\beta}}$ zu bestimmen, minimiert man das Kleinste-Quadrate Kriterium, das heißt, man sucht jenen Wert von $\boldsymbol{\beta}$ der

$$\mathrm{SSE}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

minimiert. Durch Differenzieren und Nullsetzen dieser Funktion erhält man den *Kleinsten-Quadrate Schätzer*

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Unter der Normalverteilungsannahme für $\mathbf{Y}$ kann man leicht zeigen, dass der *Kleinste-Quadrate Schätzer* für $\boldsymbol{\beta}$ mit dem *Maximum-Likelihood Schätzer* übereinstimmt (Fahrmeir et al., 2013).

Für den Erwartungswert des Schätzers erhält man

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$$

und für die Varianz folgt

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}.$$

Da jede Komponente von $\hat{\boldsymbol{\beta}}$ eine Linearkombination der normalverteilten Zielvariablen ist, gilt für den Schätzer

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right).$$

Die *geschätzten Erwartungswerte* ergeben sich durch

$$\hat{\boldsymbol{\mu}} := \widehat{\mathbb{E}}(\mathbf{Y}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y},$$

mit der $(n \times n)$ *Hat-Matrix* $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$. Auch dieser Prognosevektor $\hat{\boldsymbol{\mu}}$ ist wieder normalverteilt mit $\mathbb{E}(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu}$ und $\mathrm{Var}(\hat{\boldsymbol{\mu}}) = \sigma^2 \mathbf{H}$.

Als Maximum-Likelihood Schätzer (MLE) für die *Responsevarianz* erhält man

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_i\right)^2 = \frac{1}{n}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}\right).$$

Da dieser jedoch verzerrt ist, korrigiert man den MLE, damit er unbiased wird und bekommt dann

$$S^2 = \frac{1}{n-p}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}\right)$$

als erwartungstreuen Schätzer für σ^2. Für weitere Eigenschaften und Details zur Linearen Regressionsanalyse siehe Fahrmeir et al. (2013).

Der Vorteil solcher Modelle liegt darin, dass sie einfach zu verstehen, zu schätzen und zu interpretieren sind, da nur eine Änderung in den Prädiktorvariablen zu einer Änderung im Erwartungswert der Response führt.

Schwachstellen dieser Form der Modellierung sind die sehr restriktiven und teilweise künstlichen Annahmen. Beispiele für dieses Versagen sind Modelle für absolute Häufigkeiten oder relative Anteile, sowie jene, die für größere Erwartungswerte auch größere Variabilität implizieren, wie beispielsweise Modelle für konstante Variationskoeffizienten.

Aufgrund dieser Probleme geht man zur Klasse der *Generalisierten Linearen Modelle* (GLM) über. Details zu dieser Art der Datenanalyse wurden aus Aitkin et al. (2009) übernommen.

Solche Modelle stellen eine flexible Verallgemeinerung des bereits Gezeigten dar: Statt der Normalverteilung wird für die Responsevariable eine Verteilung aus der *einparametrigen linearen Exponentialfamilie* angenommen und dadurch die Varianz als Funktion des Erwartungswertes modelliert. Eine weitere Relaxierung der Annahmen ist die Tatsache, dass der Erwartungswert nicht notwendigerweise direkt linear modelliert werden muss.

Genauer bedeutet das: Der Zufallsvektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ habe eine Verteilung aus der einparametrigen Exponentialfamilie mit unabhängigen Komponenten Y_i, $\mathbb{E}(Y_i) = \mu_i$ und $\mathrm{Var}(Y_i) = a(\phi)\mathrm{V}(\mu_i)$. Für die *Dispersion* $a(\phi)$ gilt $a(\phi) > 0$ und ϕ nennt man *Dispersionsparameter*. Die Funktion $\mathrm{V}(\mu)$ heißt *Varianzfunktion*.

Ein *Generalisiertes Lineares Modell* geht nun aus folgender Parametrisierung hervor:

$$\begin{aligned} Y_i &\overset{ind}{\sim} \text{Exponentialfamilie}(\theta_i) \\ \mathbb{E}(Y_i) = \mu_i &= \mu(\theta_i) \\ \eta_i &= \mathbf{x}_i^\top \boldsymbol{\beta} \\ g(\mu_i) &= \eta_i. \end{aligned}$$

Die Funktion $g(.)$ heißt *Linkfunktion* und wird als bekannt vorausgesetzt. Auch hier

bezeichnet $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top$ den Vektor der Kovariablen und $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ die Designmatrix. Der Vektor der unbekannten Parameter ist $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$ und $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ steht für den Vektor der linearen Prädiktoren.

Die wichtigsten Unterschiede zum klassischen linearen Modell sind:

1. Die Additivität der nicht-beobachtbaren Fehlerterme ε_i ist im Allgemeinen nicht mehr gegeben.
2. Die Varianz kann nun auch in Abhängigkeit des Erwartungswertes modelliert werden.
3. Eine **Funktion des Erwartungswertes** wird linear modelliert, das heißt, es ist keine simple Transformation der abhängigen Variable.

In der Klasse der GLMe können für die unbekannten Parameter ebenfalls sehr leicht Maximum-Likelihood Schätzer angegeben werden, was einer Verallgemeinerung der Ergebnisse der Linearen Modelle entspricht. Genauere Ausführungen finden sich in Aitkin et al. (2009).

Auch die Modellklasse der GLMe lässt sich weiter ausdehnen: Was tun, wenn der funktionale Einfluss einzelner Prädiktoren auf die Response im Vorfeld nicht bekannt ist? Hier kommt ein sogenanntes *Generalisiertes Additives Modell* (GAM) zum Einsatz. Dabei wird das lineare Modell (parametrischer Teil) um einen nichtparametrischen Teil erweitert, der die nichtlinearen, glatten Effekte einzelner Prädiktoren enthält. Seien $\mathbf{x}_1, \ldots, \mathbf{x}_n$ mit $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top$ jene Kovariablen, die einen linearen Einfluss auf den Responsevektor $\mathbf{Y}$ ausüben und $\mathbf{z}_1, \ldots, \mathbf{z}_n$ mit $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})^\top$ solche, deren Effekt nichtparametrisch modelliert und analysiert werden soll (Fahrmeir et al., 2013).

Dies führt zu folgendem Modell:

$$Y_i = f_1(z_{i1}) + \ldots + f_m(z_{im}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{i,p-1} + \varepsilon_i.$$

Im Standardfall werden hier für den statistischen Fehler ε_i die gleichen Eigenschaften wie im klassischen linearen Modell angenommen: Es handelt sich um identisch, normalverteilte Zufallsvariablen mit $\mathbb{E}(\varepsilon_i) = 0$ und $\mathrm{Var}(\varepsilon_i) = \sigma^2$, für $i = 1, \ldots, n$.

Um den nichtparametrischen Teil zu schätzen, kann man Penalisierungsansätze für Basisfunktionen anwenden, wie sie beispielsweise bei Fahrmeir et al. (2013) vorgestellt werden. Eine sehr detailierte Betrachtung der Generalisierten Additiven Modelle findet sich in Wood (2006).

Die bisher betrachteten Methoden haben eines gemein: Sie modellieren alle den Erwartungswert beziehungsweise eine Funktion des Erwartungswertes der Responsevariable Y. Die Verteilung der Zielvariable ist dann in Abhängigkeit dieses Erwartungswertparameters eindeutig charakterisiert.

Im folgenden wird nun ein gänzlich **verteilungsfreier** Ansatz vorgestellt, der es ermöglicht, Effekte der Prädiktoren auf die *Quantile der Responseverteilung* zu schätzen.

Diese Art der Modellierung nennt man **Quantile Regression**. Es ist dadurch möglich, die Effekte der Kovariablen auf die gesamte (konditionale) Verteilung der Response zu modellieren und somit ein vollständigeres Bild als bei den üblichen „Erwartungswertmodellen“ zu erhalten.

Durch diese neue Form der Analyse von Zusammenhängen zwischen erklärenden und abhängigen Variablen können auch viele restriktive Annahmen der klassischen Regressionsmodelle relaxiert werden. Im Detail verzichtet man auf die Bedingung der Homoskedastizität und es wird auch keine bestimmte Verteilung für den Zufallsfehler angenommen (Fahrmeir et al., 2013).

Als Grundlage für die folgenden theoretischen Ausführungen wurden vor allem Koenker (2005) und Kapitel 10 in Fahrmeir et al. (2013) verwendet.

2. Einführung

Bevor man direkt in die Tiefen der Quantilen Regression eintaucht, ist es zuerst sinnvoll, den Begriff des Quantils näher zu erläutern.

2.1. Quantile

Sei Y eine reellwertige Zufallsvariable mit Verteilungsfunktion $F(y) = \mathbb{P}[Y \leq y]$.

Definition 2.1.1 (Theoretisches Quantil)**.** *Das* ***theoretische Quantil*** $F^{-1}(\tau)$*, für* $\tau \in (0,1)$*, einer Zufallsvariablen* Y *mit Verteilungsfunktion* $F(y)$ *ist definiert durch*

$$F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}.$$

Es gilt daher für das Quantil weiters:

$$\mathbb{P}\left[Y \leq F^{-1}(\tau)\right] \geq \tau \ \wedge \ \mathbb{P}\left[Y \geq F^{-1}(\tau)\right] \geq 1 - \tau.$$

Die Berechnung des theoretischen τ-Quantils lässt sich mithilfe eines Optimierungsproblems durchführen. Ziel ist es

$$\mathbb{E}\left[\rho_\tau(Y - q_\tau)\right] \tag{2.1}$$

bezüglich q_τ zu minimieren, wobei $\rho_\tau(u)$ eine stückweise lineare Funktion ist, die man *Loss-Funktion* nennt. Sie wird wie folgt definiert:

Definition 2.1.2 (Loss-Funktion)**.** *Für ein* $\tau \in (0,1)$ *und* $u \in \mathbb{R}$ *ist*

$$\rho_\tau(u) = u(\tau - \mathbb{1}_{\{u<0\}}).$$

Es handelt sich also um eine *asymmetrisch gewichtete Betragsfunktion*, bezüglich der das Kriterium minimiert werden soll. Ein Beispiel einer solchen Loss-Funktion wird in Abbildung 2.1 gezeigt, wobei dafür $\tau = 0.25$ gewählt wird. Für $\tau = 0.5$ erhält man eine symmetrisch gewichtete Betragsfunktion und berechnet damit den Median $F^{-1}(0.5)$ der Verteilung, welcher genau das 50%-Quantil ist.

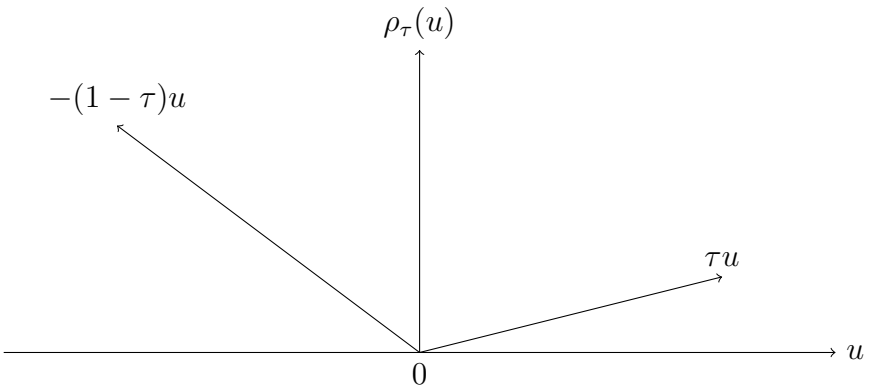

Abbildung 2.1.: Loss-Funktion mit $\tau = 0.25$

Einsetzen von $\rho_\tau(Y - q_\tau)$ in (2.1) liefert nun

$$\begin{aligned}\mathbb{E}\left[\rho_\tau\left(Y - q_\tau\right)\right] &= \mathbb{E}\left[(Y - q_\tau)\tau - (Y - q_\tau)\mathbb{1}_{\{(Y-q_\tau)<0\}}\right] \\ &= (\tau - 1)\int_{-\infty}^{q_\tau}(y - q_\tau)\mathrm{d}F(y) + \tau\int_{q_\tau}^{\infty}(y - q_\tau)\mathrm{d}F(y).\end{aligned}$$

Um eine Lösung des entsprechenden Minimierungsproblems zu erhalten leitet man nun diese Funktion bezüglich q_τ ab und setzt dies dann Null:

$$\begin{aligned}0 &\stackrel{!}{=} -(\tau - 1)\int_{-\infty}^{q_\tau}\mathrm{d}F(y) - \tau\int_{q_\tau}^{\infty}\mathrm{d}F(y) \\ &= (1 - \tau)F(q_\tau) - \tau + \tau F(q_\tau) \\ &= F(q_\tau) - \tau.\end{aligned}$$

Falls es nur ein q_τ gibt, welches $q_\tau = F^{-1}(\tau)$ erfüllt, so hat man ein eindeutiges τ-Quantil gefunden. Dies ist der Fall, wenn $F(y)$ streng monoton wachsend und stetig ist. Andernfalls gibt es ein „Intervall von τ-Quantilen“, da jedes y aus der Menge $\{y : F(y) = \tau\}$ aufgrund der Monotonie von F ein Minimum ist. Von diesen soll das kleinste Element gewählt werden. Die zweite Möglichkeit kann jedoch vernachlässigt werden, indem man annimmt, dass $F(y)$ streng monoton wachsend ist. Die Einschränkung ist jedoch nicht sehr restriktiv, da diese nur impliziert, dass es keine Lücken im Wertebereich der Response gibt, in denen die Dichte $f(y)$ gleich Null ist (Fahrmeir et al., 2013).

Ersetzt man die theoretische Verteilung $F(y)$ durch die empirische Verteilungsfunktion

$$F_n(y) = n^{-1}\sum_{i=1}^{n}\mathbb{1}_{\{Y_i \leq y\}},$$

wird ebenfalls ein q_τ gesucht, das den erwarteten Verlust minimiert. Dies führt zur Ziel-

funktion

$$\min_{q_\tau} \int \rho_\tau (y - q_\tau) \, \mathrm{d}F_n(y) = \min_{q_\tau} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau (Y_i - q_\tau),$$

und die Lösung dieses Minimierungsproblems heißt dann **empirisches τ-Quantil** oder **τ-Stichprobenquantil**. Eine etwas genauere Betrachtung erfordert später noch der Fall, wenn $n\tau$ ganzzahlig ist, da man dann tatsächlich ein Lösungsintervall $\{y : F_n(y) = \tau\}$ erhält.

Die Berechnung eines τ-Stichprobenquantils ist ein Problem, welches eng mit der Idee verbunden ist, die Ordnungsstatistik der Stichprobe zu bilden, das heißt, die Beobachtungen der Größe nach zu *ordnen*. Durch die Umformulierung in ein Minimierungsproblem wurde jedoch das Sortieren durch *Optimieren* ersetzt. Dies ist die Kernidee der Quantilen Regression und wird im Folgenden weiter ausgeführt (Koenker, 2005).

2.2. Das Stichprobenquantil - Eine erste Idee

Das Problem, ein **τ-Stichprobenquantil** $\hat{q}_\tau$ zu finden, das heißt

$$\hat{q}_\tau = \operatorname*{argmin}_{q_\tau \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau (Y_i - q_\tau),$$

kann in ein *Lineares Programm* (LP) umformuliert werden, indem man zwei nichtnegative Variablen $\{u_i, v_i : i = 1, \ldots, n\}$ einführt. Diese repräsentieren den positiven und den negativen Teil des Residuenvektors $\mathbf{r}$, wobei beide aus $\mathbb{R}^n_+$ sind. Der Residuenvektor ist allgemein durch $\mathbf{r} = \mathbf{Y} - \mathbf{1}_n q_\tau$ definiert und weiters gilt wegen der Definition der beiden Vektoren $\mathbf{u}, \mathbf{v}$, dass $\mathbf{r} = \mathbf{u} - \mathbf{v}$ ist, mit $\mathbf{u} = (u_1, \ldots, u_n)^\top$ und $\mathbf{v} = (v_1, \ldots, v_n)^\top$. Hier bezeichnet $\mathbf{1}_n$ den n-dimensionalen Einsvektor. Aufgrund dieser Tatsachen kann weiters gefolgert werden, dass

$$\mathbf{Y} = \mathbf{r} + \mathbf{1}_n q_\tau = \mathbf{u} - \mathbf{v} + \mathbf{1}_n q_\tau$$

ist und für $i = 1, \ldots, n$

$$\begin{aligned}
\rho_\tau(r_i) &= r_i \left(\tau - \mathbb{1}_{\{r_i < 0\}} \right) \\
&= (u_i - v_i)\tau - (u_i - v_i)\mathbb{1}_{\{u_i = 0\}}, \quad \text{da } r_i < 0 \Leftrightarrow u_i = 0 \\
&= u_i\tau - v_i\tau + v_i \\
&= \tau u_i + (1 - \tau)v_i
\end{aligned}$$

gilt. Mit Hilfe dieser Umformungen kann nun ein Lineares Programm formuliert werden.

Definition 2.2.1 (Formulierung des LPs)**.** *Seien $\mathbf{u}, \mathbf{v}, q_\tau$ wie zuvor. Dann ist das zu*

$\min\limits_{q_\tau \in \mathbb{R}} \sum\limits_{i=1}^{n} \rho_\tau \left(Y_i - q_\tau\right)$ *äquivalente LP wie folgt definiert:*

$$\min_{(q_\tau, \mathbf{u}, \mathbf{v}) \in \mathbb{R} \times \mathbb{R}_+^{2n}} \left\{ \tau \mathbf{1}_n^\top \mathbf{u} + (1-\tau) \mathbf{1}_n^\top \mathbf{v} \mid \mathbf{1}_n q_\tau + \mathbf{u} - \mathbf{v} = \mathbf{Y} \right\}.$$

Es wird also eine *lineare Zielfunktion* bezüglich einer *polyedrischen Restriktionsmenge* minimiert. Diese besteht aus den Schnitten der $(2n+1)$-dimensionalen Hyperebenen, welche durch die linearen Gleichheitsbedingungen und $\mathbb{R} \times \mathbb{R}_+^{2n}$ festgelegt werden. Um eine solche Optimierungsaufgabe zu lösen, verwendet man Methoden der *Linearen Programmierung* auf die in Anhang A noch im Detail eingegangen wird.

Nun betrachtet man die Zielfunktion

$$Z(q_\tau) = \sum_{i=1}^{n} \rho_\tau (Y_i - q_\tau),$$

deren Wert minimiert werden soll. $Z(\hat{q}_\tau)$ ist dann minimal und damit optimal, falls sich der Zielfunktionswert erhöht, wenn man sich von $\hat{q}_\tau$ wegbewegt. Das bedeutet, dass die beiden Richtungsableitungen von $Z(.)$ im Punkt $\hat{q}_\tau$ nichtnegativ sein müssen (weitere Details folgen in Kapitel 3). Dazu betrachtet man

$$\begin{aligned} Z'(q_\tau^+) &= \lim_{h \to 0} \frac{Z(q_\tau + h) - Z(q_\tau)}{h} \\ &= \lim_{h \to 0} \left(\frac{1}{h} \sum_{i=1}^{n} \left[\rho_\tau (Y_i - q_\tau - h) - \rho_\tau (Y_i - q_\tau) \right] \right). \end{aligned}$$

Durch Einsetzen von $\rho_\tau(u) = u(\tau - \mathbb{1}_{\{u<0\}})$ und Kürzen folgt dann

$$\begin{aligned} Z'(q_\tau^+) =& \lim_{h \to 0} \left(\frac{1}{h} \sum_{i=1}^{n} \left[-h\tau - (Y_i - q_\tau - h) \mathbb{1}_{\{Y_i - q_\tau - h < 0\}} - (Y_i - q_\tau) \mathbb{1}_{\{Y_i - q_\tau < 0\}} \right] \right) \\ =& \sum_{i=1}^{n} \left[\lim_{h \to 0} \left(-\tau - \frac{1}{h} (Y_i - q_\tau) \mathbb{1}_{\{Y_i - q_\tau < h\}} \right) \right] + \\ & \sum_{i=1}^{n} \left[\lim_{h \to 0} \left(\mathbb{1}_{\{Y_i - q_\tau < h\}} + \frac{1}{h} (Y_i - q_\tau) \mathbb{1}_{\{Y_i - q_\tau < 0\}} \right) \right] \\ =& \sum_{i=1}^{n} \left[-\tau + \mathbb{1}_{\{Y_i < q_\tau + 0\}} + (Y_i - q_\tau) \lim_{h \to 0} \frac{\mathbb{1}_{\{Y_i - q_\tau < h\}} - \mathbb{1}_{\{Y_i - q_\tau < 0\}}}{h} \right]. \end{aligned}$$

Da

$$\lim_{h \to 0} \frac{\mathbb{1}_{\{Y_i - q_\tau < h\}} - \mathbb{1}_{\{Y_i - q_\tau < 0\}}}{h}$$

die linksseitige Richtungsableitung von $\mathbb{1}_{\{Y_i - q_\tau < 0\}}$ ist, verschwindet der letzte Term. Es

ergibt sich also folgende Darstellung für die linksseitige Ableitung von $Z(q_\tau)$:

$$Z'(q_\tau^+) = \sum_{i=1}^{n} \left[\mathbb{1}_{\{Y_i < q_\tau + 0\}} - \tau \right].$$

Für die rechtsseitige Ableitung folgt analog:

$$Z'(q_\tau^-) = \sum_{i=1}^{n} \left[\tau - \mathbb{1}_{\{Y_i < q_\tau - 0\}} \right].$$

Aufgrund der geforderten Nichtnegativität der Richtungsableitungen erhält man dann die Bedingung

$$\sum_{i=1}^{n} \mathbb{1}_{\{Y_i < q_\tau - 0\}} \leq n\tau \leq \sum_{i=1}^{n} \mathbb{1}_{\{Y_i < q_\tau + 0\}}.$$

Das bedeutet, dass $n\tau$ im Intervall $[R^-, R^+]$ liegt, wobei R^+ gleich der Anzahl jener Residuen $Y_i - q_\tau$ ist, die kleiner oder gleich Null sind und R^- gibt an, wie viele von ihnen streng kleiner Null sind.

Falls $n\tau$ nicht ganzzahlig ist, gibt es einen eindeutig bestimmten Wert $\hat{q}_\tau$, der diese Bedingung erfüllt. Nimmt man an, dass es keine Bindungen gibt, das heißt $Y_{(i-1)} < Y_{(i)}$, wobei $Y_{(i)}$ das i-te Element der geordneten Stichprobe bezeichnet, dann gehört $\hat{q}_\tau$ zu einer bestimmten Ordnungsstatistik. Falls dies jedoch nicht zutrifft, ist zwar $\hat{q}_\tau$ immer noch eindeutig, aber es gibt möglicherweise mehrere zugehörige $Y_{(i)}$.

Ist $n\tau$ ein Element aus $\mathbb{Z}$, dann liegt $\hat{q}_\tau$ zwischen zwei benachbarten Ordnungsstatistiken. Damit ist die Lösung nur dann eindeutig, wenn die beiden Ordnungsstatistiken zusammenfallen. Das Auftreten solcher Bindungen geschieht jedoch mit Wahrscheinlichkeit 0 und kann daher vernachlässigt werden.

Aufgrund dieser Beobachtungen kann man nun mit Hilfe der *Dualität Linearer Programme* folgenden Zusammenhang erklären: Das *primale Problem* ist das Finden eines Stichprobenquantils, während das dazu *duale Problem* das Erstellen einer Ordnungsstatistik, genauer das Berechnen der *Ränge* der Beobachtungen, darstellt. Mit Hilfe dieser Tatsache erhält man elegante Verallgemeinerungen linearer Rangtests für das lineare QR-Modell (Koenker, 2005).

2.3. Das QR-Problem

Die bisher gefundenen Ergebnisse können nun zur Berechnung von *bedingten (konditionalen) Quantilsfunktionen* weiter verallgemeinert werden. Dazu liefern die Methoden zur Kleinsten-Quadrate-Schätzung brauchbare Vorlagen:

Es ist bekannt, dass der empirische Erwartungswert der Stichprobe durch Lösen von

$$\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (Y_i - \mu)^2$$

berechnet wird. Möchte man also den *bedingten Erwartungswert* von $\mathbf{Y}$ gegeben $\mathbf{x}$, i.e. $\mu(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$, angeben, so sollte $\boldsymbol{\beta}$ mit Hilfe der Berechnung von

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

geschätzt werden.

Ähnlich geht man nun bei der Bestimmung der *bedingten τ-Quantilsfunktion* vor: Das τ-Stichprobenquantil $\hat{q}_\tau$ löst das Minimierungsproblem

$$\min_{q_\tau\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - q_\tau).$$

Außerdem wird die konditionale τ-Quantilsfunktion durch $Q_{\mathbf{Y}}(\tau \mid \mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}(\tau)$ definiert und in Folge der vorherigen Überlegungen ergibt sich dann der Schätzer für $\boldsymbol{\beta}(\tau)$, das heißt $\hat{\boldsymbol{\beta}}(\tau)$, als jener Wert, der

$$\min_{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}(\tau)) \tag{2.2}$$

löst. Dies ist die grundlegende Idee Quantiler Regression, die von Koenker & Bassett (1978) entwickelt wurde.

Das Minimierungskriterium (2.2) wird als *Quantiles-Regressions-Problem (QR-Problem)* bezeichnet und kann analog zur Berechnung des Stichprobenquantils in ein Lineares Programm umgeschrieben werden.

Definition 2.3.1. *Sei* $\mathbf{X}$ *die* $(n \times p)$*-Designmatrix des Regressionsproblems. Dann ist*

$$\min_{(\boldsymbol{\beta}(\tau),\mathbf{u},\mathbf{v})\in\mathbb{R}^p\times\mathbb{R}^{2n}_+} \left\{\tau \mathbf{1}_n^\top \mathbf{u} + (1-\tau)\mathbf{1}_n^\top \mathbf{v} \mid \mathbf{X}\boldsymbol{\beta}(\tau) + \mathbf{u} - \mathbf{v} = \mathbf{Y}\right\}$$

eine zu (2.2) *äquivalente Darstellung des QR-Problems. Die Lösung dieses Linearen Programms,* $\hat{\boldsymbol{\beta}}(\tau)$*, heißt* ***quantilsspezifischer Regressionskoeffizient****.*

Wieder wurde der Residuenvektor $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}(\tau)$ in seinen positiven und negativen Teil zerlegt. Damit wird erneut eine lineare Zielfunktion bezüglich einer polyedrischen Restriktionsmenge minimiert und daher gelten auch für die Lösung $\hat{\boldsymbol{\beta}}(\tau)$ die bekannten Eigenschaften von Lösungen Linearer Programme.

Im Folgenden werden zwei einfache Beispiele betrachtet, die Quantilsfunktionen für ein

homoskedastisches und ein *heteroskedastisches* Modell illustrieren. In beiden Fällen handelt es sich um ein einfaches Regressionsmodell. Beide Beispiele stammen aus Koenker (2005).

Beispiel 2.3.1 (Homoskedastisches Modell)**.**

$$Y_i = \beta_0 + x_i\beta_1 + \varepsilon_i,$$

mit $\varepsilon_i \overset{iid}{\sim} F_\varepsilon$, *für* $i = 1, \ldots, n$.
Dann lautet die bedingte Quantilsfunktion von Y:

$$Q_Y(\tau \mid x) = \beta_0 + x\beta_1 + F_\varepsilon^{-1}(\tau).$$

Gesucht wird dann ein Schätzer $\hat{\boldsymbol{\beta}}(\tau) = \left(\hat{\beta}_0(\tau), \hat{\beta}_1\right)$, *für Populationsparameter der Form*

$$\boldsymbol{\beta}(\tau) = (\beta_0 + F_\varepsilon^{-1}(\tau), \beta_1)^\top.$$

In diesem Modell sind die Quantilsfunktionen für unterschiedliche τ *parallel zueinander, da die Fehler* ε_i *alle iid-verteilt sind. Dies kann man auch daran erkennen, dass der Slope-Parameter* β_1 *konstant für alle* τ *ist. Die Intercepts der einzelnen Funktionen sind aufgrund der Abhängigkeit von* τ *unterschiedlich.*

Im diesem Fall ist es in der Regel nicht notwendig, Methoden der Quantile Regression anzuwenden, da hier Least-Squares Modelle ausreichend sind.

Es wird jedoch etwas komplizierter, wenn die Annahme konstanter Varianz auf den Datensatz nicht zutrifft.

Beispiel 2.3.2 (Heteroskedastisches Modell)**.**

$$Y_i = \beta_0 + x_i\beta_1 + \sigma(x_i)\varepsilon_i,$$

wobei $\sigma(x) = \gamma x^2$, $\varepsilon_i \overset{iid}{\sim} F_\varepsilon$, *für* $i = 1, \ldots, n$.
Dann lautet die bedingte Quantilsfunktion von Y:

$$\begin{aligned} Q_Y(\tau \mid x) &= \beta_0 + x\beta_1 + \sigma(x)F_\varepsilon^{-1}(\tau) \\ &= \beta_0 + x\beta_1 + x^2\beta_2(\tau), \end{aligned}$$

mit $\beta_2(\tau) = \gamma F_\varepsilon^{-1}(\tau)$. *Sie kann weiters durch Minimierung von*

$$\sum \rho_\tau(Y_i - \beta_0 - x_i\beta_1 - x_i^2\beta_2(\tau))$$

geschätzt werden. $\hat{\boldsymbol{\beta}}(\tau)$ *ist dann konsistenter Schätzer für den Populationsparametervektor* $(\beta_0, \beta_1, \gamma F_\varepsilon^{-1}(\tau))^\top$.

Damit liefert Quantile Regression eine gute Möglichkeit, Datensätze zu modellieren, deren „Heteroskedastizität“ in Zusammenhang mit den Prädiktoren steht.

2.4. Beispiel: Münchner Mietspiegel

Beispiel 10.1 aus Fahrmeir et al. (2013) behandelt die Zusammenhänge zwischen Münchner Nettomieten von Wohnungen und verschiedenen Einflussfaktoren, wie etwa Wohnfläche oder Baujahr. Es wurden die Daten von 3082 Appartements in München erfasst.

Hier werden nun drei verschiedene Modelle betrachtet, die Auskunft darüber geben sollen, inwiefern die Response *Miete* von dem Prädiktor *Fläche* abhängig ist.

Bei den ersten beiden Modellen unterstellt man einen linearen Zusammenhang zwischen den beiden Variablen

$$\text{Miete}_i = \beta_0 + \text{Fläche}_i \cdot \beta_1 + \varepsilon_i,$$

mit $\varepsilon_i \overset{iid}{\sim} F_\varepsilon$.

Die Quantilsfunktion sieht in diesem Fall wie folgt aus:

$$Q_{\text{Miete}}\left(\tau \mid \text{Fläche}\right) = \left(\beta_0\left(\tau\right) + F_\varepsilon^{-1}\left(\tau\right)\right) + \beta_1\left(\tau\right) \cdot \text{Fläche}.$$

Mit Hilfe der Methoden der Quantilen Regression sucht man also einen Schätzer für den Parametervektor

$$\boldsymbol{\beta}(\tau) = \left(\beta_0(\tau) + F_\varepsilon^{-1}\left(\tau\right), \beta_1(\tau)\right)^\top,$$

wobei hier die einzelnen Parameter alle abhängig von τ sind und somit für jedes Quantilsniveau unterschiedliche Werte annehmen. Dies führt, wie man in der ersten Spalte von Abbildung 2.2 erkennen kann, zu nicht parallelen Geraden.

Als zweites Modell betrachtet man ein einfaches lineares Regressionsmodell, das heißt man nimmt im obigen Modell an, dass für die Fehlerverteilung $F_\varepsilon \equiv \mathcal{N}\left(0, \sigma^2\right)$ gilt. Dann werden wie in diesen Fällen üblich Kleinste-Quadrate Schätzer für die Regressionskoeffizienten berechnet.

Das τ-Quantil der Response ist in diesem Fall gegeben durch

$$Q_{\text{Miete}}\left(\tau \mid \text{Fläche}\right) = \beta_0 + \text{Fläche} \cdot \beta_1 + z_\tau \sigma^2,$$

wobei z_τ das τ-Quantil der Standardnormalverteilung ist. Wie man anhand des mittleren Plots in Abbildung 2.2 sieht, sind die Quantilsfunktionen wieder linear und aufgrund der Unabhängigkeit des Slopeparameters vom Quantilsniveau τ parallel zueinander.

Zu guter Letzt wird noch ein heteroskedastisches Modell gefittet. Dazu wählt man den Ansatz

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \exp\left(\mathbf{x}_i^\top \boldsymbol{\alpha}\right) \varepsilon_i, \quad i = 1, \ldots, n,$$

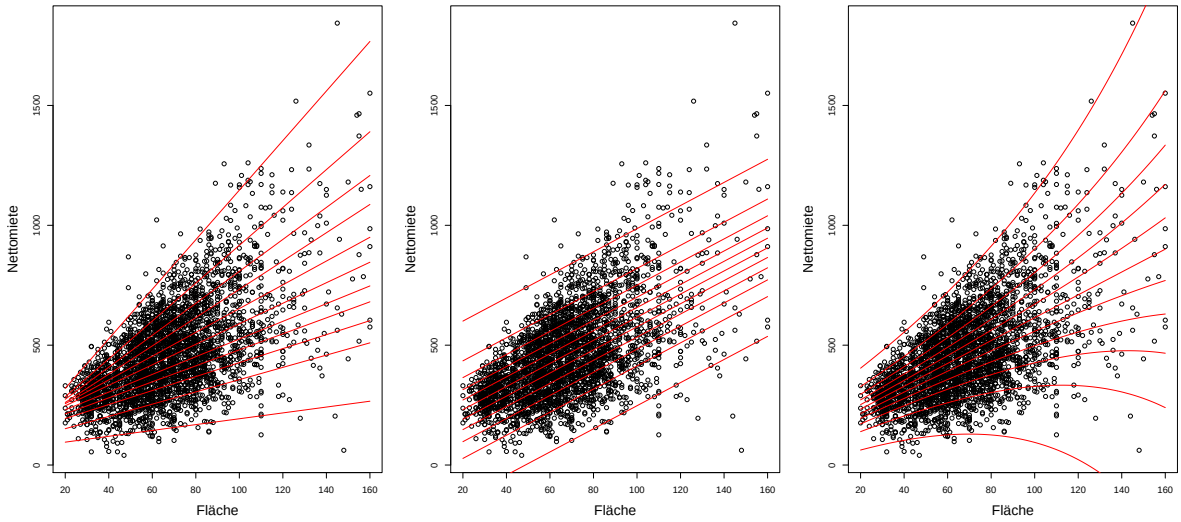

Abbildung 2.2.: Scatterplots von *Miete in Euro* gegen *Wohnfläche* zusammen mit 11 gefitteten Quantilen erzeugt durch ein Quantiles Regressionsmodell (*links*), ein lineares, homoskedastisches Modell (*mitte*) und ein heteroskedastisches Regressionsmodell (*rechts*).

mit $\varepsilon_i \sim \mathcal{N}(0,1)$. Y_i steht hier für die Response *Miete*, $\mathbf{x}_i = (1, \text{Fläche}_i)^\top$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ und $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$.

Dieses Modell wird mit Hilfe des `R`-Pakets `GAMLSS` von Rigby & Stasinopoulos (2005) gefittet und die Quantilsfunktionen, die in Abbildung 2.2 ganz rechts geplottet wurden, ergeben sich durch schätzen der unbekannten Parameter von

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + z_\tau \exp\left(\mathbf{x}^\top \boldsymbol{\alpha}\right).$$

Auch hier bezeichnet z_τ wieder das τ-Quantil der Standardnormaverteilung.

In Abbildung 2.2 werden jeweils 11 (geschätzte) Quantile der bedingten Responseverteilung für $\tau \in \{0.01, 0.1, 0.2, \ldots, 0.8, 0.9, 0.99\}$ geplottet.

Man sieht sofort, dass die verschiedenen Quantile beim QR-Modell einen deutlichen Einfluss auf die Kovariableneffekte haben und die Datenlage sehr gut widerspiegeln. Während bei niedrigen Preisen (kleine Werte für τ) der Einfluss von *Fläche* beinahe verschwindet, wächst der Slope der einzelnen Quantilsfunktionen mit steigendem τ stetig an. Dies bedeutet, dass der Prädiktor *Fläche* bei hochpreisigen Wohnungen eine deutlich größere Rolle spielt, als bei billigen Wohnungen. Vor allem die nahezu waagerechte Gerade bei $\tau = 0.01$ zeigt, dass dort die unabhängige Variable die Response kaum beeinflusst.

Auch in Abbildung 2.3 sieht man erneut deutlich den steigenden Trend der geschätzten Regressionskoeffizienten. Eine weitere Aussage dieses Diagnoseplots ist, dass falls der graue Bereich zur Gänze innerhalb der roten, gestrichelten Linien liegt, der QR-Ansatz

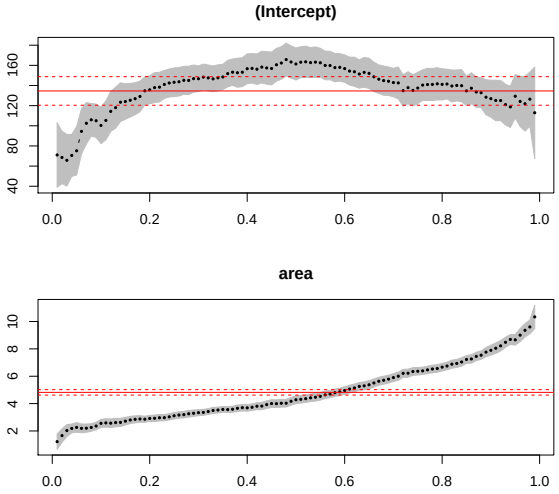

Abbildung 2.3.: Geschätzte Quantilskoeffizienten (schwarze Punkte) zu den Niveaus $\tau \in \{0.01, \ldots, 0.99\}$ mit 95%-Konfidenzintervallen (hellgrau) der einzelnen Schätzungen. Die horizontalen roten Linien zeigen die Kleinste-Quadrate Schätzung für diesen Ansatz und deren 95%-Konfidenzintervall.

überflüssig ist. Dies ist hier jedoch nicht der Fall und somit ist es durchaus sinnvoll Methoden der Quantilen Regression auf diesen Datensatz anzuwenden.

Beim homoskedastischen, linearen Modell sind die einzelnen Quantilsfunktionen wie bereits erwähnt parallel zueinander und werden immer nur durch die verschiedenen τ-Quantile der Standardnormalverteilung geshiftet. Das hat zur Folge, dass hier der Einfluss von *Fläche* in jedem Preissegment gleich groß eingeschätzt wird.

Die Quantile, die durch das heteroskedastische Modell erhalten werden, fitten die Daten ebenfalls sehr gut und sind jenen des Quantilen Regressionsmodells ähnlich. Dadurch scheinen hier sowohl die Methoden der Quantilen Regression, als auch jene der Generalisierten Additiven Modellen gut anwendbar zu sein, um diese Datenlage zu beschreiben. Für weitere Details wird auf Fahrmeir et al. (2013) verwiesen.

2.5. Beispiel: Datensatz von Engel

Ein weiteres Beispiel, wo Methoden Quantiler Regression zur Beschreibung der Datenlage eingesetzt werden stammt aus Koenker (2005) und ist als *Datensatz von Engel* bekannt. Analysen dieser Daten dienen dazu, den Zusammenhang zwischen dem Haushaltseinkommen und den Ausgaben für Essen belgischer Haushalte der Arbeiterklasse im 19. Jahrhundert zu erläutern. Im Datensatz befinden sich 235 Beobachtungen. Die abhängige Variable (Response) in diesem Modell sind die Ausgaben für Essen (`foodexp`)

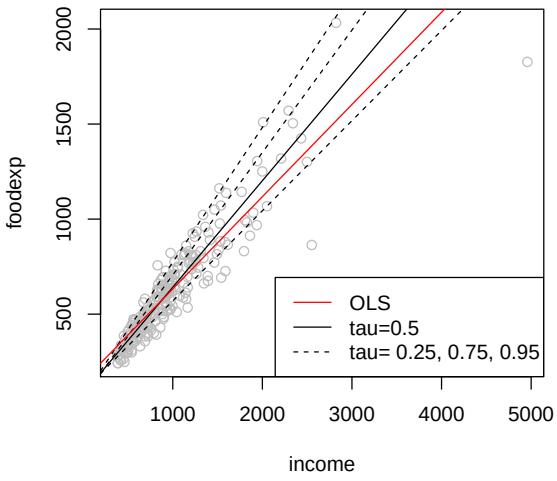

Abbildung 2.4.: Geschätzte Quantilsfunktionen für $\tau \in \{0.25, 0.75, 0.95\}$ (gestrichelt) und $\tau = 0.5$ (schwarz). Weiters stellt die rote Gerade die Least-Squares Schätzung dar.

und der Prädiktor ist Haushaltseinkommen (`income`), beides gemessen in belgischen Francs.

Durch einen Scatterplot der Daten lässt sich deutlich ein linearer Zusammenhang zwischen den beiden Variablen erkennen. Weiters kann man zwei Ausreißer feststellen, deren Ausgaben für Essen im Verhältnis zum Einkommen eher gering sind. Wie sich zeigt, haben diese Punkte vor allem einen nachteiligen Einfluss auf die Least-Squares Schätzung. In Abbildung 2.4 werden geschätzte Quantilsfunktionen zu den Niveaus $\tau \in \{0.25, 0.5, 0.75, 0.95\}$ dargestellt, sowie die Kleinsten-Quadrate Schätzung.

Anhand von 2.4 erkennt man, dass die Gerade des Kleinsten-Quadrate Schätzers nicht durch den Koordinatenursprung verläuft. Dies würde aber bedeuten, dass auch Haushalte, die kein Einkommen erhalten, Ausgaben für Essen tätigen, was die Realität absolut nicht widerspiegelt! Diese Fehleinschätzung ist durch den Einfluss der oben genannten Ausreißer entstanden, die die lineare Schätzung für den Erwartungswert nach unten hebeln und so verhindern, dass der Ursprung auf der gefitteten Gerade liegt.

Bei den geschätzten linearen Quantilsfunktionen ist dieses Verhalten nicht festzustellen, was auf die Robustheit gegenüber Ausreißern dieser Klasse von Schätzern zurückzuführen ist. Diese Tatsache spricht für die Verwendung Quantiler Regression zur Modellierung dieser Datenlage.

In Abbildung 2.5 werden noch zusätzlich die geschätzten Quantilskoeffizienten für $\tau \in \{0.01, 0.02, \ldots, 0.99\}$ gezeigt. Auch hier stellt man fest, dass die Schätzung des Intercepts für den Kleinsten-Quadrate Schätzer im gesamten Bereich über der QR-Schätzung liegt.

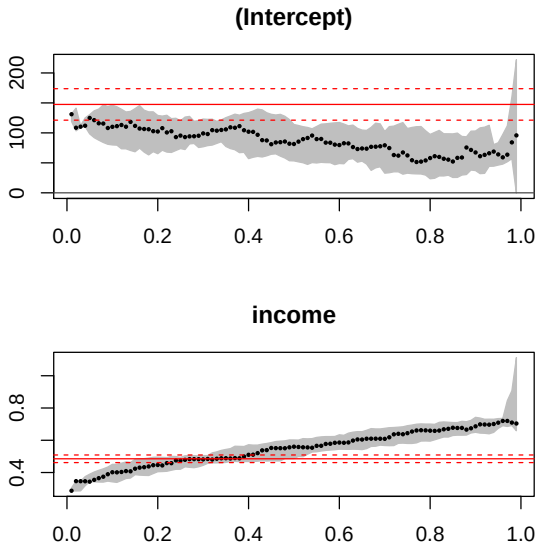

Abbildung 2.5.: Geschätzte Quantilskoeffizienten für $\tau \in \{0.01, \ldots, 0.99\}$ (schwarze Punkte) mit ihren jeweiligen geschätzten 95%-Konfidenzintervallen (grau schattierter Bereich). In rot ist die Least-Squares Schätzung und deren 95%-Konfidenzintervall dargestellt.

Weiters sieht man, dass das 95%-Konfidenzintervall des Intercepts (rot gestrichelt) die der QR-Schätzungen nicht überdeckt. Wäre dies der Fall kann man davon ausgehen, dass auf die Anwendung Quantiler Regression verzichtet werden kann und gewöhnliche Least-Squares Methoden ausreichen um die Datenlage abzubilden, da dann die Quantile der Verteilung unabhängig vom Prädiktor wären.

In der unteren Zeile von 2.5 zeigt sich, dass mit zunehmendem Quantilsniveau die geschätzten Koeffizienten größer werden. Dies erkennt man auch daran, dass die geschätzten Quantilsfunktionen nicht parallel verlaufen. Das heißt: Bei höheren Ausgaben für Essen ist der Einfluss von Einkommen größer, als im niedrigen Ausgabenbereich. Bei ansteigendem Gehalt tendieren Menschen in der Regel dazu, sich teurere Lebensmittel zu kaufen, weswegen die Essensaufwendungen im oberen Segment mit steigendem Gehalt schneller anwachsen.

Weiters kann man erkennen, dass die Konfidenzintervalle für keinen Quantilskoeffizienten die Null überdecken und daher angenommen werden kann, dass diese Parameter signifikant von Null verschieden sind. Das bedeutet wiederum, dass Einkommen auf alle Quantile der Verteilung von Essensaufwendung einen positiven Einfluss hat (alle Koeffizienten > 0). Mit steigendem Einkommen wachsen also die Ausgaben für Essen, egal ob grundsätzlich weniger oder mehr für Lebensmittel aufgewendet wird. Dies erkennt man auch an den monoton ansteigenden Geraden in Abbildung 2.5.

In diesem Fall eignen sich also Methoden der Quantilen Regression sogar besser zur

Modellierung der Datenlage, als die gewöhnlichen Least-Squares Verfahren.

2.6. Der Quantile-Treatment Effekt

Das einfachste Beispiel eines Regressionsmodells ist das klassische *Zweistichproben Treatment-Control Problem*. Dieses Modell liefert eine erste Idee zur Interpretation Quantiler Regression. Lehmann (1974) formulierte dieses Problem wie folgt:

Definition 2.6.1 (Treatment-Control Modell)**.** *Man nimmt an, dass die Behandlung (Treatment) den Wert $\Delta(x)$ hinzufügt, wenn die Response des unbehandelten Objektes x war. Dann ist die Verteilung G der Treatment-Response (Zielvariable des behandelten Objektes) jene einer Zufallsvariable $X + \Delta(X)$, während X eine Verteilung F hat.*

Spezialfälle sind zum Beispiel das *Location-Shift Modell* $\Delta(X) = \Delta_0$ und das *Scale-Shift Modell* $\Delta(X) = \Delta_0 X$. Die Behandlung ist erfolgreich, wenn $\Delta(X) \geq 0 \quad \forall X$ und damit ist dann die Verteilung G der Treatment-Response stochastisch größer als jene der Kontrollgruppe F. Dies bedeutet dann beispielsweise für klinische Studien, dass durch die Einnahme eines bestimmten Medikaments eindeutig eine Besserung eintritt und die Behandlung somit erfolgreich ist.

In Doksum (1974) wird $\Delta(x)$ durch

$$F(x) = G\left(x + \Delta(x)\right)$$

definiert. Dies entspricht dem „horizontalen Abstand“ zwischen den beiden Verteilungen F und G. Damit folgt dann

$$G^{-1}\left(F(x)\right) = x + \Delta(x) \Leftrightarrow \Delta(x) = G^{-1}\left(F(x)\right) - x. \tag{2.3}$$

Setzt man nun $\tau = F(x)$ beziehungsweise $x = F^{-1}(\tau)$ in Gleichung (2.3) ein, erhält man den *Quantile-Treatment Effekt* (QTE)

$$\delta(\tau) = \Delta\left(F^{-1}(\tau)\right) = G^{-1}(\tau) - F^{-1}(\tau).$$

Abbildung 2.6 zeigt ein Beispiel eines *Location-Scale Shift* Modells. Dort wurde jeweils die Normalverteilung für F und G angenommen, wobei sich diese sowohl im Erwartungswert, als auch in ihrer Varianz unterscheiden.

Beim Median $\tau = 0.5$ kann man deutlich einen positiven Quantile-Treatment Effekt erkennen, da $G^{-1}(0.5)$ größer als $F^{-1}(0.5)$ ist und somit $\delta(0.5) > 0$ folgt. Je weiter man sich nun nach rechts (hin zum rechten Tail der Verteilungen) bewegt, desto größer wird dieser Effekt.

Geht man jedoch in die andere Richtung, das heißt, nähert man sich dem linken Tail, wird der Treatment-Effekt sogar negativ, da dort $F^{-1}(\tau) > G^{-1}(\tau)$ gilt. Somit hat bei

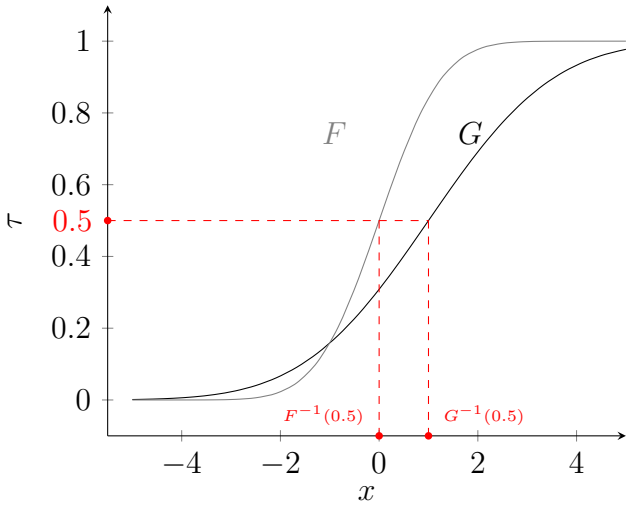

Abbildung 2.6.: Verteilung der behandelten Gruppe, G, und jene der Kontrollgruppe, F. Beim Median $\tau = 0.5$ erkennt man aufgrund der Tatsache, dass $[G^{-1}(0.5) - F^{-1}(0.5)] > 0$ ist, einen *positiven* QTE $\delta(\tau)$.

den niedrigeren Quantilen die Behandlung sogar einen *nachteiligen* Einfluss. Dies zeigt die grundlegende Herangehensweise bei der Interpretation Quantiler Regressionsmodelle.

Im einfachen Zweistichproben-Problem lässt sich der QTE $\delta(\tau)$ durch

$$\hat{\delta}(\tau) = G_n^{-1}(\tau) - F_m^{-1}(\tau)$$

schätzen, wobei $G_n(.)$ und $F_m(.)$ die empirischen Verteilungsfunktionen der beiden Stichproben sind. Die Indizes m und n geben dabei die betrachteten Stichprobenumfänge an.

Das entsprechende Quantile-Regressionsproblem lässt sich nun wie folgt formulieren (Koenker, 2005):

$$Q_{Y_i}(\tau|\mathbb{1}_i) = \alpha(\tau)(1 - \mathbb{1}_i) + \beta(\tau)\mathbb{1}_i,$$

mit Indikatorvariablen

$$\mathbb{1}_i = \begin{cases} 1, & i\text{-te Beobachtung in Treatmentgruppe,} \\ 0, & i\text{-te Beobachtung in Kontrollgruppe.} \end{cases}$$

Damit erhält man dann die Schätzer $\hat{\alpha}(\tau) = F_m^{-1}(\tau)$ und $\hat{\beta}(\tau) = G_n^{-1}(\tau)$ und mit ihrer Hilfe wird dann der geschätzte Quantile-Treatment Effekt berechnet. Durch

$$Q_{Y_i}(\tau|\mathbb{1}_i) = \alpha(\tau) + \delta(\tau)\mathbb{1}_i$$

kann der QTE auch direkt geschätzt werden.

Dieser Ansatz kann auch leicht auf Modelle mit mehreren Behandlungsmöglichkeiten

erweitert werden, siehe dazu Koenker (2005).

Weiters lässt sich ein enger Zusammenhang zwischen dem Quantile-Treatment Effekt und dem klassischen Q-Q-Plot für das Zweistichprobenproblem aufzeigen: Die Funktion die im üblichen Zweistichproben Q-Q-Plot gezeichnet wird ist

$$\hat{\Delta}(x) = G_n^{-1}\left(F_m(x)\right) - x.$$

und durch Reparametrisierung entspricht dies erneut dem geschätzten QTE. Damit kann der Quantil-Treatment Effekt nach Lehmann und Doksum als Verallgemeinerung des traditionellen Zweistichproben Q-Q-Plots und verwandter Methoden angesehen werden (Koenker, 2005).

3. Die Funktionsweise Quantiler Regression

Um in der klassischen Regressionsanalyse den Schätzer $\hat{\boldsymbol{\beta}}$ zu berechnen, minimiert man

$$\mathrm{SSE}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

indem man nach $\boldsymbol{\beta}$ ableitet und dann Null setzt. Dies liefert die sogenannten *Normalgleichungen* und deren Lösung ist dann der Schätzer

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y},$$

das heißt, man erhält eine elegante Lösung in geschlossener Form.

Ähnlich möchte man nun auch in der Quantilen Regression vorgehen, nur das hier nicht die Euklidische Distanz $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|$ minimiert wird, sondern eine Zielfunktion der Form

$$Z(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right).$$

Hier ist jedoch bei der Differentation Vorsicht geboten, da es sich bei $Z(\boldsymbol{\beta})$ um eine stückweise lineare und stetige Funktion handelt, welche überall differenzierbar ist, außer in jenen Punkten wo ein oder mehrere Residuen $Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$ Null sind. In diesen Punkten existieren jedoch alle möglichen *Richtungsableitungen*.

Allgemein ist eine Richtungsableitung in Richtung $\mathbf{w}$ mit $\|\mathbf{w}\| = 1$ wie folgt definiert:

$$\nabla f(\mathbf{x}_0, \mathbf{w}) = \lim_{t \to 0^+} \frac{f(\mathbf{x}_0 + t\mathbf{w}) - f(\mathbf{x}_0)}{t}.$$

Hier wird nun als Funktion $f(.)$ die Zielfunktion $Z(.)$ ausgewertet und deren Richtungs-

ableitung nach $\mathbf{w}$ in den nicht differenzierbaren Punkten $Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} = 0$ hat die Form

$$\begin{aligned}\lim_{t\to 0^+} \frac{1}{t}\left[Z(\boldsymbol{\beta}+t\mathbf{w}) - Z(\boldsymbol{\beta})\right] &= \lim_{t\to 0^+} \frac{1}{t}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{x}_i^\top t\mathbf{w}\right) \\ &\quad - \lim_{t\to 0^+} \frac{1}{t}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \mathbf{x}_i^\top\boldsymbol{\beta}\right) \\ &= \lim_{t\to 0^+} \frac{1}{t}\sum_{i=1}^{n} -\mathbf{x}_i^\top t\mathbf{w}\left[\tau - \mathbb{1}_{\left\{-\mathbf{x}_i^\top t\mathbf{w}<0\right\}}\right].\end{aligned}$$

Falls $-\mathbf{x}_i^\top t\mathbf{w} < 0$ muss $-\mathbf{x}_i^\top \mathbf{w} < 0$ gelten, da t von rechts gegen 0 strebt und daher nicht negativ ist beziehungsweise wird.

Damit folgt dann

$$\lim_{t\to 0^+} \frac{1}{t}\sum_{i=1}^{n} -\mathbf{x}_i^\top t\mathbf{w}\left[\tau - \mathbb{1}_{\left\{-\mathbf{x}_i^\top t\mathbf{w}<0\right\}}\right] = \sum_{i=1}^{n} -\mathbf{x}_i^\top \mathbf{w}\left[\tau - \mathbb{1}_{\left\{-\mathbf{x}_i^\top \mathbf{w}<0\right\}}\right].$$

Also ergibt sich für den Gradienten von $Z(\boldsymbol{\beta})$ in den Punkten $Y_i - \mathbf{x}_i^\top\boldsymbol{\beta} = 0$ die Darstellung

$$\nabla Z(\boldsymbol{\beta}, \mathbf{w}) = \sum_{i=1}^{n} -\mathbf{x}_i^\top \mathbf{w}\left[\tau - \mathbb{1}_{\left\{-\mathbf{x}_i^\top \mathbf{w}<0\right\}}\right].$$

Für $Y_i - \mathbf{x}_i^\top\boldsymbol{\beta} \neq 0$ berechnet man die Ableitung in Richtung $\mathbf{w}$ durch

$$\begin{aligned}\nabla Z(\boldsymbol{\beta}, \mathbf{w}) \equiv& \frac{\mathrm{d}}{\mathrm{d}t} Z(\boldsymbol{\beta}+t\mathbf{w}) \mid_{t=0} \\ =& \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{x}_i^\top t\mathbf{w}\right)\left[\tau - \mathbb{1}_{\left\{(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}-\mathbf{x}_i^\top t\mathbf{w})<0\right\}}\right]_{t=0} \\ =& \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{n} -\left(Y_i - \mathbf{x}_i^\top\boldsymbol{\beta}\right)\mathbb{1}_{\left\{(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}-\mathbf{x}_i^\top t\mathbf{w})<0\right\}} \mid_{t=0} \\ &+ \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{n}\left(-\mathbf{x}_i^\top t\mathbf{w}\right)\left[\tau - \mathbb{1}_{\left\{(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}-\mathbf{x}_i^\top t\mathbf{w})<0\right\}}\right]_{t=0}\end{aligned}$$

Da die Ableitungen im Punkt $t = 0$ betrachtet werden, gilt in diesem Fall $Y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{x}_i^\top t\mathbf{w} \neq 0$, womit die Unstetigkeitsstellen ausgeschlossen werden und die Indikatorfunktion differenzierbar ist. Somit ist der erste Ableitungsterm Null und mit Hilfe der

Produktregel folgt für den verbleibenden zweiten Term

$$\nabla Z(\boldsymbol{\beta}, \mathbf{w}) = \sum_{i=1}^{n} \left(-\mathbf{x}_i^\top \mathbf{w}\right) \left[\tau - \mathbb{1}_{\left\{(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top t\mathbf{w}) < 0\right\}}\right]_{t=0} - \sum_{i=1}^{n} \mathbf{x}_i^\top t\mathbf{w} \frac{\mathrm{d}}{\mathrm{d}t} \left[\tau - \mathbb{1}_{\left\{(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top t\mathbf{w}) < 0\right\}}\right]_{t=0}.$$

Auch hier verschwindet analog zu vorher der hintere Ableitungsterm und man erhält dann

$$\nabla Z(\boldsymbol{\beta}, \mathbf{w}) = \sum_{i=1}^{n} -\mathbf{x}_i^\top \mathbf{w} \left[\tau - \mathbb{1}_{\left\{(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) < 0\right\}}\right].$$

Zusammengefasst hat man nun folgende Darstellung der Richtungsableitung der Zielfunktion $Z(\boldsymbol{\beta})$ (Koenker, 2005):

$$\begin{aligned} \nabla Z(\boldsymbol{\beta}, \mathbf{w}) &\equiv \frac{\mathrm{d}}{\mathrm{d}t} Z(\boldsymbol{\beta} + t\mathbf{w}) \mid_{t=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top t\mathbf{w}\right) \left[\tau - \mathbb{1}_{\left\{(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top t\mathbf{w}) < 0\right\}}\right]_{t=0} \\ &= -\sum_{i=1}^{n} \psi_\tau^*(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}, -\mathbf{x}_i^\top \mathbf{w}) \mathbf{x}_i^\top \mathbf{w}, \end{aligned} \tag{3.1}$$

mit der Funktion

$$\psi_\tau^*(u, v) = \begin{cases} \tau - \mathbb{1}_{\{u<0\}}, & \text{falls } u \neq 0, \\ \tau - \mathbb{1}_{\{v<0\}}, & \text{falls } u = 0. \end{cases}$$

Im Folgenden kann mit Hilfe des Gradienten $\nabla Z(\boldsymbol{\beta}, \mathbf{w})$ eine Optimalitätsbedingung für die Lösung des Minimierungsproblems aufgestellt werden, was zu Begriffen wie *Basislösungen* und *Subgradienten-Bedingung* führt.

3.1. Basislösungen und die Subgradienten-Bedingung

Existiert nun ein $\hat{\boldsymbol{\beta}}$, für das alle Richtungsableitungen von $Z(\boldsymbol{\beta})$ nichtnegativ sind, so minimiert dies die Zielfunktion. Dies ist analog dazu $\nabla Z(\boldsymbol{\beta}, \mathbf{w})$ Null zu setzen, wenn die Funktion glatt wäre, das heißt in allen Punkten ausreichend oft differenzierbar sei. Mit anderen Worten: Man verlangt, dass die Funktion monoton ansteigt, wenn man sich von dem Punkt $\hat{\boldsymbol{\beta}}$ wegbewegt, unabhängig davon, welche Richtung man wählt (Koenker, 2005).

Eine geometrische Interpretation dieser Lösungen sieht wie folgt aus: Die Restriktionsmenge dieses Minimierungsproblems kann man sich als Polyeder vorstellen, welches

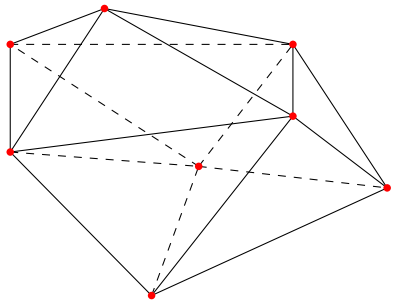

Abbildung 3.1.: Unregelmäßiges Polyeder mit glatten Begrenzungsflächen, geraden Verbindungskanten, die sich in einzelnen Eckpunkten schneiden.

durch glatte Flächen erzeugt wird, die durch gerade Kanten miteinander verbunden sind und sich in einzelnen Ecken treffen.

Minimiert man nun eine Zielfunktion bezüglich dieser Restriktionsmenge, so folgt aufgrund der Theorie über Lineare Programme, dass sich eine optimale Lösung im Allgemeinen in einer der Ecken befinden muss. Möchte man also ihre Optimalität zeigen, so genügt es zu beweisen, dass die Zielfunktion entlang aller Kanten, die aus dieser Ecke entspringen, zunimmt.

Weiters erfüllen diese Lösungen die Eigenschaft des „Exakten-Fits". Diese besagt, dass in diesen Eckpunkten immer genau p Beobachtungen interpoliert werden, wenn man ein Modell mit p Parametern schätzen möchte. Interpolieren bedeutet in diesem Zusammenhang, dass diese p Beobachtungen exakt gefittet werden. Die anderen werden dabei jedoch nicht einfach ignoriert, sondern dienen dazu, die zu interpolierenden Punkte festzulegen. Der durch die Regression geschätzte Median beispielsweise, identifiziert p Beobachtungen, die eine Hyperebene bilden, welche die bedingte Medianfunktion am besten repräsentiert (Koenker, 2005).

In der Theorie der Linearen Programme heißen diese $p-$elementigen Untermengen *Basislösungen*. Es ist durchaus möglich, dass es nicht nur eine Lösung des Linearen Programms gibt, sondern dass eine ganze Lösungsmenge auftritt. Dies ist der Fall, wenn die Zielfunktionsebene die polyedrische Restriktionsmenge nicht exakt in einem Eckpunkt berührt, sondern beispielsweise entlang einer ganzen Kante des Polyeders schneidet. Hier spielen Basislösungen eine entscheidende Rolle: Jedes Element der Lösungsmenge ist als Linearkombination der Basislösungen darstellbar.

Sei nun $h \in \mathcal{H}$ eine $p-$elementige Teilmenge aus $\mathcal{N} = \{1, \ldots, n\}$ und $\mathbf{X}(h)$ bezeichne die Untermatrix von $\mathbf{X}$ mit Zeilen $\{\mathbf{x}_i : i \in h\}$. Analog dazu ist $\mathbf{Y}(h)$ ein $p-$dimensionaler Vektor mit Koordinaten $\{Y_i : i \in h\}$. Das Komplement von h bezüglich $\mathcal{N}$ wird mit $\bar{h}$ bezeichnet. Mit Hilfe dieser Notation kann nun der Begriff Basislösung formal definiert werden.

Definition 3.1.1. ***Basislösungen****, welche die Punktmenge* $\{(\mathbf{x}_i, Y_i), i \in h\}$ *interpolie-*

ren, sind von der Form

$$\mathbf{b}(h) = \mathbf{X}(h)^{-1}\mathbf{Y}(h),$$

unter der Annahme, dass die $(p \times p)$ *Matrix* $\mathbf{X}(h)$ *nicht singulär ist.*

Da es aber $\binom{n}{p} = \mathcal{O}(n^p)$ verschiedene solcher Lösungen gibt, wäre es zu aufwendig alle willkürlich durchzusehen. Daher verwendet man den *Simplex-Algorithmus* um die Optimallösung zu bestimmen. Dabei geht man von einer Ecke der Restriktionsmenge zur nächsten, immer in Richtung des steilsten Abstiegs.

Soll nun eine Basislösung $\mathbf{b}(h)$ der oben genannten Form optimal für das Minimierungsproblem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$$

sein, so muss dort die Nichtnegativität der Richtungsableitungen erfüllt sein. Das heißt es muss

$$\nabla Z\left(\mathbf{b}(h), \mathbf{w}\right) \geq 0$$

gelten.

Reparametrisiert man die Richtung, sodass

$$\mathbf{v} = \mathbf{X}(h)\mathbf{w} \Leftrightarrow \mathbf{w} = \mathbf{X}(h)^{-1}\mathbf{v}$$

gilt, dann ist $\mathbf{b}(h)$ genau dann Optimallösung des Problems, wenn

$$0 \leq -\sum_{i=1}^{n} \psi_\tau^*\left(Y_i - \mathbf{x}_i^\top \mathbf{b}(h), -\mathbf{x}_i^\top \mathbf{X}(h)^{-1}\mathbf{v}\right) \mathbf{x}_i^\top \mathbf{X}(h)^{-1}\mathbf{v}, \ \forall\, \mathbf{v} \in \mathbb{R}^p \tag{3.2}$$

gilt.

Unter bestimmten Voraussetzungen an die Reihenfolge der Elemente von h gilt natürlich $\mathbf{e}_i^\top = \mathbf{x}_i^\top \mathbf{X}(h)^{-1}$ für $i \in h$, wobei $\mathbf{e}_i^\top$ der i-te Basisvektor von $\mathbb{R}^p$ ist. Dadurch folgt dann weiters

$$\mathbf{x}_i^\top \mathbf{X}(h)^{-1}\mathbf{v} = \mathbf{e}_i^\top \mathbf{v} = v_i, \text{ für } i \in h.$$

Setzt man dies nun in Forderung (3.2) ein, so erhält man

$$0 \leq -\sum_{i\in h} \psi_\tau^*(0, -v_i)v_i - \sum_{i\in\bar{h}} \psi_\tau^*\left(Y_i - \mathbf{x}_i^\top \mathbf{b}(h), -\mathbf{x}_i^\top \mathbf{X}(h)^{-1}\mathbf{v}\right) \mathbf{x}_i^\top \mathbf{X}(h)^{-1}\mathbf{v}.$$

Durch die Substitution

$$\boldsymbol{\xi}(h)^\top = \sum_{i\in\bar{h}} \psi_\tau^*(Y_i - \mathbf{x}_i^\top \mathbf{b}(h), -\mathbf{x}_i^\top \mathbf{X}(h)^{-1}\mathbf{v})\mathbf{x}_i^\top \mathbf{X}(h)^{-1}$$

wird dies schließlich zu

$$0 \leq -\sum_{i \in h} \psi_\tau^*(0, -v_i) v_i - \boldsymbol{\xi}(h)^\top \mathbf{v} = -\sum_{i \in h} (\tau - \mathbb{1}_{\{-v_i < 0\}}) v_i - \boldsymbol{\xi}(h)^\top \mathbf{v}. \tag{3.3}$$

Der Raum $\mathbb{R}^p$ wird von den Richtungsvektoren $\mathbf{v} = \pm \mathbf{e}_l$ mit $l = 1, \ldots, p$ aufgespannt und daher folgt, dass die oben gezeigte Bedingung an die Richtungsableitung nur genau dann $\forall\, \mathbf{v} \in \mathbb{R}^p$ erfüllt ist, wenn sie für alle $2p$ Einheitsvektoren hält.

Setzt man also $\mathbf{v} = \mathbf{e}_i$, dann ist

$$\begin{aligned} v_i &= 1, \\ v_j &= 0 \text{ für } j \neq i, \\ \boldsymbol{\xi}(h)^\top \mathbf{e}_i &= \xi_i(h) \end{aligned}$$

und somit vereinfacht sich die rechte Seite in (3.3) zu

$$\begin{aligned} -\sum_{i \in h} (\tau - \mathbb{1}_{\{-v_i < 0\}}) v_i - \boldsymbol{\xi}(h)^\top \mathbf{v} &= -(\tau - \mathbb{1}_{\{-1 < 0\}}) 1 - \boldsymbol{\xi}(h)^\top \mathbf{e}_i \\ &= -(\tau - 1) - \xi_i(h). \end{aligned}$$

Damit ergeben sich also p Ungleichungen der Form (Koenker, 2005)

$$0 \leq -(\tau - 1) - \xi_i(h), \;\; i = 1, \ldots, p.$$

Für $\mathbf{v} = -\mathbf{e}_i$ erhält man analog

$$0 \leq \tau + \xi_i(h), \;\; i = 1, \ldots, p.$$

Geht man davon aus, dass die Responsevariable $\mathbf{Y}$ eine Dichtefunktion bezüglich eines Lebesquemaßes hat, dann sind alle Residuen für $i \in \bar{h}$ mit Wahrscheinlichkeit 1 ungleich Null, das heißt $(\mathbf{X}, \mathbf{Y})$ befinden sich in *allgemeiner Lage*, was von Rousseeuw & Leroy (1987) wie folgt definiert wird:

Definition 3.1.2. *Das Paar $(\mathbf{X}, \mathbf{Y})$ befindet sich in **allgemeiner Lage**, wenn es für jede Wahl von p ein einziger exakter Fit resultiert. Das bedeutet, dass*

$$Y_i - \mathbf{x}_i^\top \mathbf{b}(h) \neq 0$$

für jedes $i \notin h$.

Durch diese Tatsache verschwindet die Abhängigkeit von $\boldsymbol{\xi}$ bezüglich $\mathbf{v}$ und die oben angeführten Ungleichungen können zur *Optimalitätsbedingung*

$$-\tau \mathbb{1}_p \leq \boldsymbol{\xi}(h) \leq (1 - \tau) \mathbb{1}_p$$

zusammengefasst werden.

Mit Hilfe dieser Beobachtungen kann nun Theorem 3.3 aus Koenker & Bassett (1978) neu formuliert werden:

Satz 3.1.1 (Existenz einer Optimallösung)**.** *Falls sich* $(\mathbf{X}, \mathbf{Y})$ *in allgemeiner Lage befinden, dann existiert eine optimale Basislösung* $\mathbf{b}(h) = \mathbf{X}(h)^{-1}\mathbf{Y}(h)$ *des QR-Problems* (2.2) *genau dann, wenn für ein* $h \in \mathcal{H}$

$$-\tau \mathbb{1}_p \leq \boldsymbol{\xi}(h) \leq (1-\tau)\mathbb{1}_p$$

gilt. Der Vektor $\boldsymbol{\xi}(h)^\top$ *ist von der Form*

$$\boldsymbol{\xi}(h)^\top = \sum_{i \in \bar{h}} \psi_\tau(Y_i - \mathbf{x}_i^\top \mathbf{b}(h))\mathbf{x}_i^\top \mathbf{X}(h)^{-1}$$

mit $\psi_\tau(u) = \tau - \mathbb{1}_{\{u<0\}}$*. Weiters ist* $\mathbf{b}(h)$ *genau dann eindeutige Lösung, falls die Ungleichungen strikt erfüllt sind, andernfalls erhält man als Lösungsmenge die konvexe Hülle verschiedener Basislösungen.*

Dieses Optimalitätskriterium impliziert eine weitere wichtige Eigenschaft für endliche Stichproben, welche besagt, dass es unter der Annahme einer Designmatrix mit Intercept ungefähr $n\tau$ negative und $n(1-\tau)$ positive Residuen gibt, wobei n den Stichprobenumfang angibt (Koenker, 2005).

Satz 3.1.2. *Man bezeichne mit* P *und* M *die Anzahl der positiven und negativen Residuen* $r_i = Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$*, während* N *angibt, wie viele davon Null sind. Falls für die Designmatrix* $\mathbf{X}$ *ein* $\boldsymbol{\alpha} \in \mathbb{R}^p$ *existiert, sodass* $\mathbf{X}\boldsymbol{\alpha} = \mathbf{1}_n$*, dann gilt für jede Optimallösung* $\hat{\boldsymbol{\beta}}(\tau)$ *von Problem* (2.2)*:*

$$\begin{aligned} M \leq n\tau \leq M + N, \\ P \leq n(1-\tau) \leq P + N. \end{aligned}$$

Beweis. Optimalität hält in $\hat{\boldsymbol{\beta}}(\tau)$ genau dann, wenn

$$0 \leq -\sum_{i=1}^{n} \psi_\tau^*(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau), -\mathbf{x}_i^\top \mathbf{w})\mathbf{x}_i^\top \mathbf{w}$$

gilt. Die Funktion $\psi_\tau^*(u, v)$ ist wie zuvor durch

$$\psi_\tau^*(u,v) = \begin{cases} \tau - \mathbb{1}_{\{u<0\}}, & \text{falls } u \neq 0, \\ \tau - \mathbb{1}_{\{v<0\}}, & \text{falls } u = 0 \end{cases}$$

festgelegt. Setze man jetzt $\mathbf{w} = \boldsymbol{\alpha}$, sodass $\mathbf{X}\boldsymbol{\alpha} = \mathbf{1}_n$ gilt, dann liefert dies die Bedingung

$$0 \leq -\sum_{i=1}^{n} \psi_\tau^*(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau), -1).$$

Mit Hilfe der Fallunterscheidung

- $r_i > 0$: $\psi_\tau^*(r_i, -1) = \tau$,
- $r_i < 0$: $\psi_\tau^*(r_i, -1) = \tau - 1$,
- $r_i = 0$: $\psi_\tau^*(r_i, -1) = \tau - 1$ (da $-1 < 0$)

und der Tatsache, dass es genau P positive, M negative und N Residuen mit Wert Null gibt, folgt nun

$$0 \leq -\sum_{i=1}^{n} \psi_\tau^*(r_i, -1) = -P\tau - M(\tau - 1) - (\tau - 1)N$$

und dies ist gleichbedeutend mit

$$0 \geq P\tau - (1 - \tau)M - (1 - \tau)N. \tag{3.4}$$

Analog bekommt man für $\mathbf{w} = -\boldsymbol{\alpha}$:

$$0 \leq \sum_{i=1}^{n} \psi_\tau^*(r_i, 1) = \tau P + (\tau - 1)M + \tau N,$$

was äquivalent ist zu

$$0 \geq -\tau P + (1 - \tau)M - \tau N. \tag{3.5}$$

Weiters gilt $n = P + M + N$ und damit folgt dann für (3.4)

$$0 \geq \tau(P + M + N) - M - N \Leftrightarrow \tau n \leq M + N.$$

Analog ergibt sich für Ungleichung (3.5)

$$\tau n \geq M.$$

Kombiniert man nun diese beiden Ergebnisse so erhält man die Ungleichungskette

$$M \leq \tau n \leq M + N$$

und subtrahiert man dann überall n so folgt nach Multiplikation der gesamten Ungleichungskette mit -1

$$P \leq n(1 - \tau) \leq P + N.$$

Dies sind genau jene Ungleichungsketten, die es zu beweisen galt. □

Geht man von einer nicht-degenerierten Basislösung aus, das heißt es gibt genau p Residuen die den Wert Null haben, so kann man $N = p$ setzen und man sieht sofort aufgrund

von Satz 3.1.2 und nach Division durch die Stichprobengröße n, dass

$$\frac{M}{n} \leq \tau \leq \frac{M+p}{n}$$

folgt und somit der Anteil negativer Residuen approximativ τ ist.

Analog ergibt sich für den relativen Anteil der positiven Residuen ein Wert von circa $(1-\tau)$, da

$$\frac{P}{n} \leq (1-\tau) \leq \frac{P+p}{n}$$

gilt.

Setzt man $p = 1$, das heißt man betrachtet ein Modell, welches nur einen Intercept enthält, dann spiegelt dieses Ergebnis genau das Stichprobenquantil wider. Falls $n\tau$ ganzzahlig ist, sind die Ungleichungen nur schwach erfüllt und man erhält ein Intervall von τ-Stichprobenquantilen, welches zwischen zwei benachbarten Ordnungsstatistiken liegt. Ist andernfalls $n\tau$ nicht ganzzahlig, dann gibt es ein eindeutiges empirisches τ-Quantil. Folgendes Beispiel soll diesen Sachverhalt illustrieren.

Beispiel 3.1.1 ($n = 6$, $p = 1$)**.** *Wähle zuerst $\tau = 1/3$. Damit ist dann $n\tau = 2$ und $n(1-\tau) = 4$, also ist $n\tau$ beziehungsweise $n(1-\tau) \in \mathbb{Z}$. Daraus ergeben sich folgende Ungleichungen:*

$$1 \leq M \leq 2,$$
$$3 \leq P \leq 4.$$

Aufgrund dessen kann M die Werte 1 und 2 annehmen, P passend dazu die Werte 4, 3. Damit ist die Lösung nicht eindeutig, da dies für das gesamte Intervall zwischen der zweiten und dritten Ordnungsstatistik erfüllt wird (inklusive der Ordnungsstatistiken selbst).

Sei nun $\tau = 1/5$, d.h. $n\tau = 1.2$ und $n(1-\tau) = 4.8$. Damit erhält man die Ungleichungen

$$0.2 < M < 1.2 \Rightarrow M = 1,$$
$$3.8 < P < 4.8 \Rightarrow P = 4.$$

Dadurch ist die Lösung eindeutig festgelegt.

3.2. Eigenschaften des Schätzers $\hat{\beta}(\tau)$

Nun werden einige *Equivarianz-Eigenschaften* des Schätzers $\hat{\boldsymbol{\beta}}(\tau)$ angegeben, die eine einheitliche Interpretation der Regressionsergebnisse garantieren sollen.

Im Allgemeinen geht man davon aus, dass beispielsweise eine Änderung der Messskala keinen Einfluss auf die Schätzung hat. Falls die Daten also einer bekannten Transformation unterzogen werden, nimmt man an, dass sich die Regressionsschätzer in einer Weise ändern, sodass die Interpretation der Ergebnisse gleich bleibt und somit *invariant* bezüglich der Transformation ist.

Im Folgenden wird das τ-Regressionsquantil, dass von den Daten $(\mathbf{X}, \mathbf{Y})$ festgelegt wird, explizit mit $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ gekennzeichnet. Die folgenden Equivarianz-Eigenschaften stammen aus Koenker & Bassett (1978).

Satz 3.2.1 (Equivarianz-Eigenschaften)**.** *Sei* $\mathbf{A}$ *eine nichtsinguläre,* $(p \times p)$ *Matrix,* $\boldsymbol{\gamma} \in \mathbb{R}^p$ *und* $a > 0$*. Dann gilt für jedes* $\tau \in [0, 1]$

1. $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, a\mathbf{Y}) = a\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$,
2. $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, -a\mathbf{Y}) = -a\hat{\boldsymbol{\beta}}(1 - \tau; \mathbf{X}, \mathbf{Y})$,
3. $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y} + \mathbf{X}\boldsymbol{\gamma}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) + \boldsymbol{\gamma}$,
4. $\hat{\boldsymbol{\beta}}(\tau; \mathbf{XA}, \mathbf{Y}) = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$.

Die ersten beiden Eigenschaften implizieren eine Art *Skaleninvarianz*, während die dritte im Allgemeinen *Shift-* beziehungsweise *Regressions-Invarianz* heißt. Nummer vier kann man als Equivarianz gegenüber Reparametrisierung des Design verstehen (Koenker, 2005).

Um den Beweis durchführen zu können, muss man zuerst eine alternative Darstellung der Zielfunktion $Z(\boldsymbol{\beta})$ angeben.

Satz 3.2.2 (Äquivalente Darstellung)**.** *Für die Zielfunktion* $Z(\boldsymbol{\beta})$ *gilt:*

$$Z(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left[\tau - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\right] \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)$$

mit

$$\operatorname{sgn}(u) = \begin{cases} 1, & u > 0, \\ -1, & u < 0, \\ 0, & u = 0. \end{cases}$$

Beweis. Siehe Anhang B! □

Die zweite Darstellung stammt ebenfalls aus Koenker & Bassett (1978).

Mit Hilfe dieser neuen Schreibweise für die Zielfunktion kann nun der Beweis zu Satz 3.2.1 durchgeführt werden, wobei analog zu $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ auch die zu minimierende Funktion genauer durch $Z(\boldsymbol{\beta}; \tau, \mathbf{Y}, \mathbf{X})$ charakterisiert wird.

Beweis zu Satz 3.2.1. Um die zuvor angeführten Equivarianz-Eigenschaften zu zeigen, genügt es die folgenden Eigenschaften der Zielfunktion zu beweisen, da $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ diese Funktion minimiert:

1. $Z\left(a\boldsymbol{\beta}; \tau, \mathbf{X}, a\mathbf{Y}\right) = aZ\left(\boldsymbol{\beta}; \tau, \mathbf{X}, \mathbf{Y}\right)$ mit $a > 0$,
2. $Z\left(-a\boldsymbol{\beta}; 1-\tau, \mathbf{X}, -a\mathbf{Y}\right) = aZ\left(\boldsymbol{\beta}; \tau, \mathbf{X}, \mathbf{Y}\right)$ mit $a > 0$,
3. $Z\left(\boldsymbol{\beta}; \tau, \mathbf{X}, \mathbf{Y}\right) = Z\left(\boldsymbol{\beta} + \boldsymbol{\gamma}; \tau, \mathbf{X}, \mathbf{Y} + \mathbf{X}\boldsymbol{\gamma}\right)$,
4. $Z\left(\boldsymbol{\beta}; \tau, \mathbf{X}, \mathbf{Y}\right) = Z\left(\mathbf{A}^{-1}\boldsymbol{\beta}; \tau, \mathbf{X}\mathbf{A}, \mathbf{Y}\right)$.

Eine wichtige Tatsache, die in weiterer Folge zur Anwendung kommt, ist

$$\begin{aligned} \operatorname{sgn}(u) &= \operatorname{sgn}(a \cdot u) \text{ für } a > 0, \\ -\operatorname{sgn}(u) &= \operatorname{sgn}(-u). \end{aligned}$$

Damit kann die erste Gleichheit wie folgt gezeigt werden:

$$\begin{aligned} aZ\left(\boldsymbol{\beta}; \tau, \mathbf{X}, \mathbf{Y}\right) &= a \cdot \sum_{i=1}^{n} \left[\tau - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\right]\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) \\ &= \sum_{i=1}^{n} \left[\tau - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(aY_i - \mathbf{x}_i^\top a\boldsymbol{\beta})\right]\left(aY_i - \mathbf{x}_i^\top a\boldsymbol{\beta}\right) \\ &= Z\left(a\boldsymbol{\beta}; \tau, \mathbf{X}, a\mathbf{Y}\right). \end{aligned}$$

Ähnlich geht man auch vor, um Äquivalenz zwei zu zeigen:

$$\begin{aligned} aZ\left(\boldsymbol{\beta}; \tau, \mathbf{X}, \mathbf{Y}\right) &= a \cdot \sum_{i=1}^{n} \left[\tau - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\right]\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) \\ &= \sum_{i=1}^{n} (-1)\left[\tau - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(a(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}))\right]\left(-a(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\right) \\ &= \sum_{i=1}^{n} \left[(1-\tau) - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(-a(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}))\right]\left(-a(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\right) \\ &= Z\left(-a\boldsymbol{\beta}; 1-\tau, \mathbf{X}, -a\mathbf{Y}\right). \end{aligned}$$

Die dritte Gleichung beweist man, indem man zu den Residuen $\pm\mathbf{x}_i^\top \boldsymbol{\gamma}$ addiert und somit

$$\begin{aligned} &Z\left(\boldsymbol{\beta}; \tau, \mathbf{X}, \mathbf{Y}\right) = \\ &= \sum_{i=1}^{n} \left[\tau - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(Y_i + \mathbf{x}_i^\top \boldsymbol{\gamma} - \mathbf{x}_i^\top (\boldsymbol{\beta} + \boldsymbol{\gamma}))\right]\left(Y_i + \mathbf{x}_i^\top \boldsymbol{\gamma} - \mathbf{x}_i^\top (\boldsymbol{\beta} + \boldsymbol{\gamma})\right) \\ &= Z\left(\boldsymbol{\beta} + \boldsymbol{\gamma}; \tau, \mathbf{X}, \mathbf{Y} + \mathbf{X}\boldsymbol{\gamma}\right) \end{aligned}$$

erhält.

Die letzte Identität ist klar, da $\mathbf{X}\mathbf{A}\mathbf{A}^{-1}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$ ist. □

Diese Eigenschaften haben auch die Kleinsten-Quadrate Schätzer, obwohl sie nicht einfach auf alle Arten von Schätzern verallgemeinert werden dürfen. Quantile jedoch besitzen eine weitere, noch viel wichtigere Eigenschaft: die *Equivarianz gegenüber monotoner Transformation*!

Satz 3.2.3. *Sei $f(.)$ eine monoton wachsende Funktion in $\mathbb{R}$. Dann gilt für eine beliebige Zufallsvariable Y*

$$Q_{f(Y)}(\tau) = f\left(Q_Y(\tau)\right).$$

Diese Eigenschaft der Quantile nennt man ***Equivarianz gegenüber monotoner Transformation****.*

Beweis. Es gilt, dass

$$y = \inf\left\{x \in \mathbb{R} : \mathbb{P}\left[Y \leq x\right] \geq \tau\right\} \Leftrightarrow f(y) = \inf\left\{x \in \mathbb{R} : \mathbb{P}\left[f(Y) \leq f(x)\right] \geq \tau\right\},$$

da $f(.)$ eine monoton wachsende Funktion ist. Daher ist y genau dann τ-Quantil der Verteilung von Y wenn $f(y)$ jenes der Verteilung der transformierten Zufallsvariable $f(Y)$ ist, das heißt

$$\begin{aligned} Q_Y(\tau) &= y, \\ Q_{f(Y)}(\tau) &= f(y). \end{aligned}$$

Aufgrund dieser Tatsache folgt dann

$$Q_{f(Y)}(\tau) = f(y) = f\left(Q_Y(\tau)\right)$$

und damit ist die Aussage bewiesen. □

Wegen dieser Tatsache ist die Interpretation von Transformationen bei Quantilen-Regressionsmodellen einleuchtender als bei Methoden, die den Erwartungswert der Response modellieren, da er diese Eigenschaft im Allgemeinen nicht erfüllt.

Hat man also beispielsweise ein lineares QR-Modell $\mathbf{x}^\top \hat{\boldsymbol{\beta}}$ für den bedingten Median von $f(\mathbf{Y})$ gegeben $\mathbf{x}$ gefittet, so kann $f^{-1}(\mathbf{x}^\top \hat{\boldsymbol{\beta}})$ sofort als Schätzer für den Median von $\mathbf{Y}$ gegeben $\mathbf{x}$ verwendet werden.

Ein Anwendungsgebiet dieser Equivarianz-Eigenschaften ist zum Beispiel die Betrachtung von *zensierten Variablen*, das heißt die Responsevariable wird auf einen bestimmten Bereich eingeschränkt, wie etwa $Y_i^* = \max\left\{0, Y_i\right\}$. Für eine detailiertere Behandlung dieses Themas siehe Kapitel 2.2.4 in Koenker (2005).

3.3. Einflussfunktion und Bruchpunkt als Maße der Robustheit

Schon Gauß und Laplace setzten sich mit den verschiedenen Vorzügen von Erwartungswert und Median als Lageparameterschätzungen auseinander. Während Gauß zeigte, dass der Erwartungswert als Schätzer in gewissem Sinn optimal ist, wenn die Fehler eine Verteilung haben, deren Dichte proportional zu $\exp^{-x^2}$ ist, hob Laplace hervor, dass der Median bei Beobachtungen mit großen Fehlern ein besseres Verhalten aufweist.

Im Folgenden bezeichnet das Funktional $\hat{\theta}(F)$ einen Parameter der Verteilungsfunktion $F(\cdot)$. Ein Maß um die Robustheit von $\hat{\theta}(F)$ anzugeben ist die *Einflussfunktion*, welche von Hampel (1974) eingeführt wurde. Mit ihrer Hilfe kann man feststellen, wie es den Parameter $\hat{\theta}(F)$ beeinflusst, wenn die Verteilung $F(\cdot)$ bezüglich der man ihn berechnet „verunreinigt“ wird. Möchte man im Speziellen die Robustheit von Schätzern genauer betrachten, so wertet man das Funktional $\hat{\theta}(F)$ bezüglich der empirischen Verteilungsfunktion aus und erhält so den empirischen Schätzer.

Die Verteilung $F(\cdot)$ wird nun wie folgt kontaminiert: Man ersetzt einen kleinen Teil $\omega \geq 0$ von $F(\cdot)$ durch ein äquivalentes Maß, das um y konzentriert ist. Damit ergibt sich die „verunreinigte“ Verteilung $F_\omega(\cdot)$ wie folgt:

$$F_\omega(u) = \omega\delta_y(u) + (1-\omega)\,F(u), \tag{3.6}$$

wobei δ_y eine Verteilungsfunktion ist, die dem Punkt y Masse Eins zuweist, auch bekannt als *Dirac-Maß*. Dies bedeutet, dass einerseits für $y \leq u$ das Maß $\delta_y(u) = 1$ ist, während andernfalls $\delta_y(u) = 0$ gilt. Man bezeichnet solche Verteilungen im Allgemeinen auch noch als *Einpunktverteilungen*. Mit Hilfe dieser Voraussetzungen kann nun die Einflussfunktion definiert.

Definition 3.3.1 (Einflussfunktion des Parameters $\hat{\theta}(F)$)**.** *Die Einflussfunktion $EF\left(y;\hat{\theta}(F)\right)$ eines Parameters $\hat{\theta}(F)$, welcher als Funktional einer Verteilung $F(\cdot)$ betrachtet wird, ist durch*

$$EF\left(y,\hat{\theta}(F)\right) = \lim_{\omega\downarrow 0}\frac{\hat{\theta}\left(F_\omega\right) - \hat{\theta}\left(F\right)}{\omega}$$

definiert.

Einflussfunktion des Erwartungswertes

Um die Einflussfunktion des Erwartungswertes herzuleiten, setzt man zuerst

$$\hat{\theta}_E\left(F_\omega\right) = \mathbb{E}_{F_\omega}\left[Y\right] = \int y\mathrm{d}F_\omega.$$

Dann verwendet man Darstellung (3.6) für $F_\omega(\cdot)$ und erhält damit

$$\int y \mathrm{d}F_\omega = \int y \mathrm{d}\left(\omega\delta_y + (1-\omega)F\right) = \omega\int y\mathrm{d}\delta_y + (1-\omega)\int y\mathrm{d}F.$$

Da das Dirac-Maß δ_y dem Punkt y Wahrscheinlichkeitsmasse Eins zuweist, reduziert sich das erste Integral auf $\omega\int y\mathrm{d}\delta_y = \omega y$ und im zweiten Term gilt $\int y\mathrm{d}F = \mathbb{E}_F[Y] = \hat{\theta}_E(F)$. Dies liefert nun

$$\hat{\theta}_E(F_\omega) = \omega\int y\mathrm{d}\delta_y + (1-\omega)\int y\mathrm{d}F = \omega y + (1-\omega)\hat{\theta}_E(F).$$

Dieses Ergebnis verwendet man jetzt um die Einflussfunktion von $\hat{\theta}_E(F_\omega)$ herzuleiten:

$$\begin{aligned}
EF(y;\hat{\theta}_E(F)) &= \lim_{\omega\downarrow 0}\frac{\hat{\theta}_E(F_\omega) - \hat{\theta}_E(F)}{\omega} \\
&= \lim_{\omega\downarrow 0}\frac{\omega y + (1-\omega)\hat{\theta}_E(F) - \hat{\theta}_E(F)}{\omega} \\
&= \lim_{\omega\downarrow 0}\frac{\omega\left(y - \hat{\theta}_E(F)\right) + \cancel{\hat{\theta}_E(F)} - \cancel{\hat{\theta}_E(F)}}{\omega} \\
&= \lim_{\omega\downarrow 0}\frac{\cancel{\omega}\left(y - \hat{\theta}_E(F)\right)}{\cancel{\omega}} \\
&= y - \hat{\theta}_E(F)
\end{aligned}$$

Einflussfunktion der Quantile zum Niveau τ

Im nächsten Schritt soll die Einflussfunktion für ein τ-Quantil angegeben werden. Dazu definiert man

$$\hat{\theta}_\tau(F_\omega) = F_\omega^{-1}(\tau).$$

Dies ist aber äquivalent zu

$$F_\omega\left(\hat{\theta}_\tau(F_\omega)\right) = \tau.$$

Setzt man dann $u = \hat{\theta}_\tau(F_\omega)$ in Darstellung (3.6) von F_ω ein, folgt damit

$$F_\omega(u) = \omega\delta_y\left(\hat{\theta}_\tau(F_\omega)\right) + (1-\omega)F\left(\hat{\theta}_\tau(F_\omega)\right) = \tau, \tag{3.7}$$

wobei aufgrund der Definition der Dirac-Funktion beziehungsweise Einpunktverteilung im Punkt y

$$\delta_y\left(\hat{\theta}_\tau(F_\omega)\right) = \begin{cases} 1, & y \le \hat{\theta}_\tau(F_\omega) \\ 0, & y > \hat{\theta}_\tau(F_\omega) \end{cases}$$

gilt. Nun führt man eine Fallunterscheidung durch um eine Darstellung für das τ-Quantil bezüglich der Verteilung $F_\omega(\cdot)$ zu finden.

• Dabei betrachtet man zuerst den Fall $y > \hat{\theta}_\tau(F_\omega)$, was gleichbedeutend mit $\delta_y\left(\hat{\theta}_\tau(F_\omega)\right) = 0$ ist, und es folgt für (3.7)

$$(1-\omega)\, F\left(\hat{\theta}_\tau(F_\omega)\right) = \tau.$$

Umformen liefert weiters

$$F\left(\hat{\theta}_\tau(F_\omega)\right) = \frac{\tau}{1-\omega}$$

und damit ergibt sich für das Quantil zum Niveau τ bezüglich der Verteilung $F_\omega(\cdot)$ in diesem Fall die Darstellung

$$\hat{\theta}_\tau(F_\omega) = F^{-1}\left(\frac{\tau}{1-\omega}\right).$$

• Im zweiten Fall gilt $y \leq \hat{\theta}_\tau(F_\omega)$ und daher automatisch $\delta_y\left(\hat{\theta}_\tau(F_\omega)\right) = 1$. Setzt man dies nun wieder in (3.7) ein, führt das zu

$$\omega + (1-\omega)\, F\left(\hat{\theta}_\tau(F_\omega)\right) = \tau.$$

Formt man erneut um, bekommt man

$$\hat{\theta}_\tau(F_\omega) = F^{-1}\left(\frac{\tau-\omega}{1-\omega}\right),$$

was bedeutet, dass das τ-Quantil bezüglich der „verunreinigten“ Verteilung nicht eindeutig festgelegt ist.

Zusammenfassend erhält man also für das theoretische τ-Quantil der Verteilung $F_\omega(\cdot)$

$$\hat{\theta}_\tau(F_\omega) = \begin{cases} F^{-1}\left(\frac{\tau}{1-\omega}\right), & y > \hat{\theta}_\tau(F_\omega) \\ F^{-1}\left(\frac{\tau-\omega}{1-\omega}\right), & y \leq \hat{\theta}_\tau(F_\omega). \end{cases}$$

Für die beiden Argumente in der Quantilsfunktion $F^{-1}(\cdot)$ wird im Folgenden eine Taylorreihenentwicklung durchgeführt.

Definition 3.3.2 (Taylorreihe)**.** *Die Taylorreihenentwicklung für eine unendlich oft differenzierbare Funktion $f : I \to \mathbb{R}$, $I \subset \mathbb{R}$ in einem Entwicklungspunkt ω_0 ist durch*

$$T(f(\omega), \omega_0) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\omega_0)}{n!} (\omega - \omega_0)^n$$

festgelegt.

Als Entwicklungspunkt verwendet man in beiden Fällen $\omega_0 = 0$ und für $f_1(\omega) = \frac{\tau}{1-\omega}$ gilt

$$f_1(\omega_0) = \tau \text{ und } f_1^{(1)}(\omega_0) = \frac{\tau}{(1-\omega_0)^2} = \tau,$$

womit dann die Reihenentwicklung

$$T(f_1(\omega), 0) = \tau + \tau\omega + \mathcal{O}\left(\omega^2\right)$$

folgt.

Im Fall $f_2(\omega) = \frac{\tau-\omega}{1-\omega}$ erhält man mit Entwicklungspunkt $\omega_0 = 0$

$$f_2(\omega_0) = \tau \text{ und } f_2^{(1)}(\omega_0) = \frac{\tau-1}{(1-\omega_0)^2} = \tau - 1,$$

was weiters zur Taylorreihenentwicklung

$$T(f_2(\omega), 0) = \tau + (\tau-1)\omega + \mathcal{O}\left(\omega^2\right)$$

führt.

In beiden Fällen kann der Term $\mathcal{O}(\omega^2)$ vernachlässigt werden, da ω sehr klein gewählt wird, wie man zu Beginn bereits bemerkt hat, und damit wird dies verschwindend klein.

Weiters bekommt man mit Hilfe dieser Taylorreihenentwicklung für $y > \hat{\theta}_\tau(F_\omega)$ folgende Schranke für y:

$$\hat{\theta}_\tau(F_\omega) = F^{-1}(\tau + \tau\omega) \geq F^{-1}(\tau) = \hat{\theta}_\tau(F),$$

da wegen $\tau \in (0,1)$ und $\omega \geq 0$ die Abschätzung $\tau + \tau\omega \geq \tau$ hält, und die Quantilsfunktion $F^{-1}(\cdot)$ monoton wachsend ist. Dies führt dann zu

$$\begin{aligned} & y > \hat{\theta}_\tau(F_\omega) \geq \hat{\theta}_\tau(F) \\ \Rightarrow\ & y > \hat{\theta}_\tau(F). \end{aligned}$$

Andererseits kann man analog für $y \leq \hat{\theta}_\tau(F_\omega)$ feststellen, dass

$$\hat{\theta}_\tau(F_\omega) = F^{-1}(\tau + (\tau-1)\omega) \leq F^{-1}(\tau) = \hat{\theta}_\tau(F),$$

denn es gilt $(\tau - 1) < 0$ für $\tau \in (0,1)$ und $\omega \geq 0$, was zu $\tau + (\tau-1)\omega \leq \tau$ führt. Somit kann man für y folgern:

$$\begin{aligned} & y \leq \hat{\theta}_\tau(F_\omega) \leq \hat{\theta}_\tau(F) \\ \Rightarrow\ & y \leq \hat{\theta}_\tau(F). \end{aligned}$$

Zusammenfassend lässt sich mit Hilfe dieser Erkenntnisse das theoretische Quantil zum Niveau τ der Verteilung $F_\omega(\cdot)$ schreiben als

$$\hat{\theta}_\tau\left(F_\omega\right) = \begin{cases} F^{-1}\left(\tau + \tau\omega\right), & y > \hat{\theta}_\tau\left(F\right) \\ F^{-1}\left(\tau + \left(\tau - 1\right)\omega\right), & y \leq \hat{\theta}_\tau\left(F\right). \end{cases}$$

Diese Darstellung von $\hat{\theta}_\tau\left(F_\omega\right)$ setzt man nun in Definition 3.3.1 der Einflussfunktion ein, wobei auch hier wieder eine Fallunterscheidung vorgenommen werden muss.

An dieser Stelle ist es sinnvoll den Begriff des Differentialquotienten näher zu erläutern, da er später ein wichtiges Hilfsmittel zur Ermittlung von $EF(y;\hat{\theta}_\tau(F))$ sein wird.

Definition 3.3.3 (Differentialquotient)**.** *Der Differentialquotient einer Funktion $f : I \to \mathbb{R}$, $I \subset \mathbb{R}$, im Punkt x_0 ist wie folgt definiert:*

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x}|_{x=x_0} = \lim_{h\to 0^+} \frac{f(x_0+h) - f(x_0)}{h}.$$

Betrachtet man also den Fall $y - \hat{\theta}_\tau\left(F\right) > 0$, dann erhält man

$$EF(y;\hat{\theta}_\tau(F)) = \lim_{\omega\downarrow 0} \frac{1}{\omega}\left[F^{-1}\left(\tau + \tau\omega\right) - F^{-1}\left(\tau\right)\right]$$

und durch Multiplikation mit $1 = \frac{\tau}{\tau}$ kommt man zu

$$EF(y;\hat{\theta}_\tau(F)) = \tau \lim_{\omega\downarrow 0} \frac{1}{\tau\omega}\left[F^{-1}\left(\tau + \tau\omega\right) - F^{-1}\left(\tau\right)\right].$$

Wählt man nun $h = \tau\omega \Leftrightarrow \frac{1}{h} = \frac{1}{\tau\omega}$ und $x_0 = \tau$ erkennt man mit Hilfe der Definition des Differentialquotienten, dass es sich bei letzterem Grenzwert um die Ableitung von $F^{-1}(x)$ an der Stelle $x_0 = \tau$ handelt, das heißt die Einflussfunktion vereinfacht sich zu

$$EF(y;\hat{\theta}_\tau(F)) = \tau \frac{\mathrm{d}F^{-1}(x)}{\mathrm{d}x}\,|_{x=\tau}.$$

Weiters betrachtet man noch die Ableitung der Quantilsfunktion genauer: Aufgrund der Definition der Verteilungsfunktion und deren Umkehrfunktion gilt

$$F\left(F^{-1}(x)\right) = x.$$

Leitet man dies nun mit Hilfe der Kettenregel nach x ab, erhält man

$$f\left(F^{-1}(x)\right)\frac{\mathrm{d}}{\mathrm{d}x}F^{-1}(x) = 1,$$

wobei $f(\cdot)$ die Dichtefunktion zur Verteilung $F(\cdot)$ ist. Damit ergibt sich für die Ableitung

der inversen Funktion weiters

$$\frac{\mathrm{d}}{\mathrm{d}x}F^{-1}(x) = \frac{1}{f\left(F^{-1}(x)\right)}.$$

Also kann die Einflussfunktion des τ-Quantils in diesem Fall wie folgt dargestellt werden:

$$EF(y;\hat{\theta}_\tau(F)) = \frac{\tau}{f\left(F^{-1}\left(\tau\right)\right)}, \text{ für } y - \hat{\theta}_\tau\left(F\right) > 0.$$

Ähnlich geht man jetzt vor, um den Fall $y - \hat{\theta}_\tau\left(F\right) \leq 0$ abzuhandeln:

$$EF(y;\hat{\theta}_\tau(F)) = \lim_{\omega\downarrow 0}\frac{1}{\omega}\left[F^{-1}\left(\tau + (\tau-1)\omega\right) - F^{-1}\left(\tau\right)\right]$$

und durch Multiplikation mit $1 = \frac{\tau-1}{\tau-1}$ erhält man

$$EF(y;\hat{\theta}_\tau(F)) = (\tau-1)\lim_{\omega\downarrow 0}\frac{1}{(\tau-1)\omega}\left[F^{-1}\left(\tau + (\tau-1)\omega\right) - F^{-1}\left(\tau\right)\right].$$

Wählt man nun $h = (\tau-1)\omega \Leftrightarrow \frac{1}{h} = \frac{1}{(\tau-1)\omega}$ und $x_0 = \tau$, dann liefert die Definition des Differentialquotienten erneut

$$\begin{aligned} EF(y;\hat{\theta}_\tau(F)) &= (\tau-1)\frac{\mathrm{d}F^{-1}(x)}{\mathrm{d}x}|_{x=\tau} \\ &= \frac{\tau-1}{f\left(F^{-1}\left(\tau\right)\right)}, \text{ für } y - \hat{\theta}_\tau\left(F\right) \leq 0. \end{aligned}$$

Zusammenfassend kann man also die Einflussfunktion des τ-Quantils durch

$$EF(y;\hat{\theta}_\tau(F)) = \frac{\tau - \mathbb{1}_{\{y\leq\hat{\theta}_\tau(F)\}}}{f\left(F^{-1}\left(\tau\right)\right)} \tag{3.8}$$

darstellen, wobei für den Nenner dessen Existenz und Positivität, das heißt $f\left(F^{-1}\left(\tau\right)\right) > 0$, vorausgesetzt wird.

Vergleicht man nun $EF(y;\hat{\theta}_E(F))$ und $EF(y;\hat{\theta}_\tau(F))$ miteinander, dann sieht man sofort einen deutlichen Unterschied:

Da

$$\begin{aligned}\left|EF(y;\hat{\theta}_\tau(F))\right| &= \underbrace{\frac{1}{f\left(F^{-1}\left(\tau\right)\right)}}_{>0}\left|(-1)\left(\mathbb{1}_{\left\{y\leq\hat{\theta}_\tau(F)\right\}}-\tau\right)\right| \\ &= \frac{1}{f\left(F^{-1}\left(\tau\right)\right)}\underbrace{\left|\mathbb{1}_{\left\{y\leq\hat{\theta}_\tau(F)\right\}}-\tau\right|}_{<1,\text{ da }\tau\in(0,1)} \\ &< \frac{1}{f\left(F^{-1}\left(\tau\right)\right)} < \infty\end{aligned}$$

gilt, ist die Einflussfunktion des τ-Quantils durch die Konstante

$$s\left(\tau\right) = \frac{1}{f\left(F^{-1}\left(\tau\right)\right)}$$

beschränkt, das heißt eine Veränderung der Verteilung im Punkt y hat immer nur einen begrenzten Einfluss auf das Quantil. Das ist gleichbedeutend damit, dass die τ-Quantile **robuste** Schätzer für einen Lageparameter sind.

Betrachtet man hingegen die Einflussfunktion des Erwartungswertes

$$EF(y;\hat{\theta}_E(F)) = y - \hat{\theta}_E(F),$$

so stellt man fest, dass der Einfluss einer Änderung von $F(\cdot)$ in y einen beliebig großen, also unbegrenzten Einfluss auf den Erwartungswert haben kann, da auch schon eine geringe „Kontaminierung“ der Verteilung in einem Punkt y weit genug von $\hat{\theta}_E(F)$ entfernt, den Wert von $EF(y;\hat{\theta}_E(F))$ sogar unendlich groß werden lassen kann.

Die Funktion

$$s\left(\tau\right) = \frac{1}{f\left(F^{-1}\left(\tau\right)\right)},\ \tau\in(0,1)$$

wird bei Tukey (1965) als *Sparsity Funktion* oder *Sparsity zum Niveau* τ und bei Parzen (1979) als *Dichtefunktion der Quantile* bezeichnet.

Sie ist ein Maß für die Genauigkeit der Quantilsschätzer, denn sie spiegelt die Dichte in der Nähe des betrachteten τ-Quantils wieder: Sind die Daten in der Nähe des Quantils eher dünn *(engl.: sparse)*, was gleichbedeutend mit einer hohen Sparsity ist, dann wird es schwierig zu diesem Niveau τ eine Schätzung anzugeben. Andererseits erhält man bei einer dichten Datenlage rund um das gewünschte Quantil, das heißt bei einem niedrigen Wert der Sparsity Funktion, deutlich präzisere Schätzer. Später wird noch genauer darauf eingegangen, wie diese Funktion mit Hilfe einer Stichprobe geschätzt werden kann.

Vergleich der Einflussfunktionen von Median und Erwartungswert

Ersetzt man in (3.8) τ durch $\frac{1}{2}$, so liefert dies die Einflussfunktion des Medians, $EF(y;\hat{\theta}_M(F))$:

$$EF(y;\hat{\theta}_M(F)) = \frac{\frac{1}{2} - \mathbb{1}_{\{y\leq\hat{\theta}_M(F)\}}}{f\left(F^{-1}\left(\frac{1}{2}\right)\right)} = \frac{1}{2}\,\mathrm{sgn}^*\left(y - \hat{\theta}_M(F)\right)\frac{1}{f\left(F^{-1}\left(\frac{1}{2}\right)\right)},$$

wobei hier

$$\mathrm{sgn}^*(u) = \begin{cases} -1, & u \leq 0 \\ 1, & u > 0 \end{cases}$$

gilt.

In Abbildung 3.2 wird der Unterschied bezüglich der Einflussfunktionen von Erwartungswert und Median graphisch analysiert, wobei für die Verteilung F die Standardnormaverteilung angenommen wurde. Dadurch erhält man für die Einflussfunktion des Medians

$$EF(y;\hat{\theta}_M(F)) = \frac{2.507}{2}\,\mathrm{sgn}^*\left(y - 0\right) = 1.253\,\mathrm{sgn}^*\left(y\right),$$

da $\hat{\theta}_M(F) = 0$ und $\frac{1}{f\left(F^{-1}\left(\frac{1}{2}\right)\right)} = \sqrt{2\pi} = 2.507$ gilt.

Für den Erwartungswert liefert diese Wahl der Verteilungsfunktion

$$EF(y;\hat{\theta}_E(F)) = y,$$

da $\hat{\theta}_E(F) = 0$ ist.

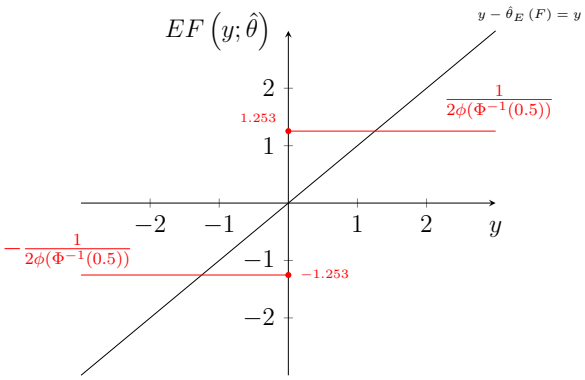

Abbildung 3.2.: Vergleich der Einflussfunktionen von Median und Erwartungswert, wobei hier $F \equiv \mathcal{N}(0,1)$ gewählt wurde.

Die Folge dieser Erläuterungen ist, dass der Median im Gegensatz zum Erwartungswert ein robusterer Schätzer für einen Lageparameter ist.

Einflussfunktion der Regressionsquantile

Die erhaltenen Ergebnisse kann man auch auf Quantile Regression übertragen. Hier muss F die gemeinsame Verteilung der Paare $(\mathbf{x}, Y)$ beschreiben. Dazu verwendet man die Definition der bedingten Dichte

$$f(y \mid x) = \frac{f_{X,Y}(x,y)}{g_X(x)}.$$

Statt $\hat{\theta}_\tau(F)$ wird nun das geschätzte Regressionsquantil zum Niveau τ verwendet werden, welches die Form $\mathbf{x}^\top \hat{\boldsymbol{\beta}}_F(\tau)$ hat.

Unter diesen Voraussetzungen und der Annahme, dass die Dichte $f(\cdot|x)$ stetig und strikt positiv ist, ergibt sich laut Koenker (2005) für die Einflussfunktion der Regressionsquantile die Darstellung

$$EF\left((y,\mathbf{x}), \hat{\boldsymbol{\beta}}_F(\tau)\right) = Q^{-1}\mathbf{x}\,\mathrm{sgn}^*\left(y - \mathbf{x}^\top \hat{\boldsymbol{\beta}}_F(\tau)\right),$$

mit

$$Q = \int \mathbf{x}\mathbf{x}^\top f\left(\mathbf{x}^\top \hat{\boldsymbol{\beta}}_F(\tau)|\mathbf{x}\right) g(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Auf die genaue Herleitung dieser Funktion wird an dieser Stelle verzichtet. Auch hier sieht man, dass der Einfluss einer Änderung bezüglich der Response Y nur einen beschränkten Einfluss auf die Schätzung hat, was wieder gleichbedeutend damit ist, dass die Regressionsquantile robuste Schätzungen darstellen.

Da die Einflussfunktion hier nur vom Vorzeichen des Residuums $r_i = Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_F(\tau)$ abhängig ist, kann man die Responsebeobachtung Y_i beliebig in y-Richtung verschieben, solange man dadurch das Vorzeichen des Residuums nicht verändert, ohne dass sich die Lage des geschätzten Regressionsquantils $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_F(\tau)$ ändert.

Dies wiederum verdeutlicht eine bereits zuvor gemachte Bemerkung: Bei QR-Modellen werden keine Beobachtungen als Ausreißer verworfen, sondern sie helfen dabei die repräsentativen Punkte für die Schätzung auszuwählen.

Diese Eigenschaft Quantiler Regression kann man auch formal wie folgt angeben, wobei nun wieder auf die explizite Kennzeichung der Abhängigkeit des Schätzers $\hat{\boldsymbol{\beta}}_F(\tau)$ von der Verteilung F verzichtet werden kann:

Satz 3.3.1. *Sei* $\mathbf{D}$ *eine Diagonalmatrix mit nichtnegativen Einträgen* d_i, $i = 1, \ldots, n$. *Dann gilt*

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Dr} + \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})),$$

mit dem Residuenvektor $\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$.

Beweis. Da man annimmt, dass $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ Lösung des Optimierungsproblems ist, gilt für den Gradienten der Zielfunktion (3.1) in diesem Punkt, $\nabla Z(\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}), \mathbf{w})$, für alle

Richtungen $\mathbf{w} \in \mathbb{R}^p$ mit $\|\mathbf{w}\| = 1$

$$-\sum_{i=1}^{n} \psi_\tau^* \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \geq 0,$$

wobei hier der Übersicht wegen für $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ nur $\hat{\boldsymbol{\beta}}$ geschrieben wird. Weiters ist die Funktion $\psi_\tau^*(u, v)$ erneut wie folgt definiert:

$$\psi_\tau^*(u, v) = \begin{cases} \tau - \mathbb{1}_{(u<0)} & u \neq 0, \\ \tau - \mathbb{1}_{(v<0)}, & u = 0. \end{cases}$$

Nun muss bewiesen werden, dass $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ auch Lösung des „kontaminierten" Problems ist und daher gleich $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Dr} + \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}))$ ist. Dazu setzt man der Übersicht halber $\widetilde{\mathbf{Y}} = \mathbf{Dr} + \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ und zeigt im Folgenden, dass

$$\begin{aligned} \nabla Z\left(\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \widetilde{\mathbf{Y}}), \mathbf{w}\right) &= -\sum_{i=1}^{n} \psi_\tau^* \left(\widetilde{Y}_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \\ &\overset{z.z.}{\geq} -\sum_{i=1}^{n} \psi_\tau^* \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \geq 0 \end{aligned} \tag{3.9}$$

gilt, wobei $\widetilde{Y}_i = d_i \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\right) + \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ ist.

Für den Gradienten $\nabla Z\left(\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \widetilde{\mathbf{Y}}), \mathbf{w}\right)$ ergibt sich durch Einsetzen von $\widetilde{Y}_i$ die Darstellung

$$\begin{aligned} \nabla Z\left(\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \widetilde{\mathbf{Y}}), \mathbf{w}\right) &= -\sum_{i=1}^{n} \psi_\tau^* \left(\widetilde{Y}_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \\ &= -\sum_{i=1}^{n} \psi_\tau^* \left(\cancel{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}} + d_i \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\right) - \cancel{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \\ &= -\sum_{i=1}^{n} \psi_\tau^* \left(d_i \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\right), -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w}. \end{aligned}$$

Weiters hält die zu zeigende Ungleichung in (3.9), wenn für sämtliche Glieder der beiden Summen

$$-\psi_\tau^* \left(d_i \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\right), -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \geq -\psi_\tau^* \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \tag{3.10}$$

gilt.

Um die weiteren Betrachtungen etwas übersichtlicher zu gestalten, setzt man

$$r_i := Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$$
$$v_i := \mathbf{x}_i^\top \mathbf{w},$$

wodurch Ungleichung (3.10) zu

$$-v_i \cdot \psi_\tau^* (d_i r_i, -v_i) \geq -v_i \cdot \psi_\tau^* (r_i, -v_i) \tag{3.11}$$

wird.

Für den Fall, dass $r_i = 0$ ist, gibt es keine Unklarheiten, denn es folgt sofort

$$-v_i \cdot \psi_\tau^* (0, -v_i) = -v_i \cdot \psi_\tau^* (0, -v_i)$$

und somit ist die Aussage in diesem Fall erfüllt, unabhängig davon welche Werte v_i oder d_i annehmen.

Deshalb geht man nun auch zu $r_i \neq 0$ über.

Weiters betrachtet man im ersten Schritt $v_i > 0$, womit dann durch Division von $-v_i < 0$ in (3.11)

$$\psi_\tau^* (d_i r_i, -v_i) \overset{!}{\leq} \psi_\tau^* (r_i, -v_i)$$

folgt.

Wählt man jetzt $d_i > 0$ so haben die beiden Argumente $d_i r_i \neq 0$ und $r_i \neq 0$ immer dasselbe Vorzeichen und aufgrund der Definition von $\psi_\tau^*()$ erhält man die wahre Aussage

$$\tau - \mathbb{1}_{\{d_i r_i < 0\}} \leq \tau - \mathbb{1}_{\{r_i < 0\}},$$

denn $\mathbb{1}_{\{d_i r_i < 0\}} = 1 \Leftrightarrow \mathbb{1}_{\{r_i < 0\}} = 1$.

Nun wird gezeigt, warum die Annahme nichtnegativer Diagonaleinträge zwingend notwendig ist. Falls $d_i < 0$ betrachtet man die selbe Ungleichung wie für $d_i > 0$, aber $d_i r_i \neq 0$ und $r_i \neq 0$ haben jetzt genau das gegengleiche Vorzeichen, was für negative Residuen r_i zu einer falschen Aussage

$$\tau \leq \tau - 1$$

führt, denn für τ aus $(0, 1)$ gilt $(\tau - 1) \in (-1, 0)$. Also können negative Werte für d_i ausgeschlossen werden, da in diesem Fall Ungleichung (3.10) nicht erfüllt wird.

Die Wahl $d_i = 0$ liefert

$$\psi_\tau^* (0, -v_i) \leq \psi_\tau^* (r_i, -v_i),$$

was wiederum durch Einsetzen der Definition von $\psi_\tau^*()$ gleichbedeutend mit

$$\tau - \mathbb{1}_{\{-v_i < 0\}} \leq \tau - \mathbb{1}_{\{r_i < 0\}}$$

ist. Da man bisher vorausgesetzt hat, dass $v_i > 0$ und somit $-v_i < 0$ ist, vereinfacht sich dies zu

$$\tau - 1 \leq \tau - \mathbb{1}_{\{r_i<0\}} = \begin{cases} \tau, & r_i \geq 0 \\ \tau - 1, & r_i < 0. \end{cases}$$

Dies ist also für alle $r_i \neq 0$ erfüllt.

Im Fall von $v_i \leq 0$ wird (3.11) durch Division von $-v_i \geq 0$ zu

$$\psi_\tau^* \left(d_i r_i, -v_i\right) \overset{!}{\geq} \psi_\tau^* \left(r_i, -v_i\right).$$

Für die Fälle $r_i = 0 \ \wedge \ d_i$ beliebig, sowie $r_i \neq 0 \ \wedge \ d_i \geq 0$ kann man analog zu den vorherigen Betrachtungen folgern, dass die Ungleichung hält. Auch hier würde das Zulassen negativer Diagonaleinträge d_i zu Problemen führen, denn es gilt

$$\tau - \mathbb{1}_{\{d_i r_i<0\}} \geq \tau - \mathbb{1}_{\{r_i<0\}}$$

und durch die ungleichen Vorzeichen von $d_i r_i$ und r_i folgt dann für $r_i > 0$

$$\tau - 1 \geq \tau.$$

Diese falsche Aussage zeigt noch einmal deutlich, dass für d_i nur Werte aus $\mathbb{R}_+$ in Frage kommen und die Annahme im Satz daher gerechtfertig ist.

Mit Hilfe dieser Überlegungen hat man also gezeigt, dass für alle $i = 1, \ldots, n$ die Ungleichung

$$-\psi_\tau^* \left(d_i \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\right), -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \geq -\psi_\tau^* \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w}$$

für alle $d_i \geq 0$ und jedes mögliche $r_i = Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ hält. Weiters wurden auch alle möglichen Richtungen überprüft, indem man jede mögliche Wahl von $v_i = \mathbf{x}_i^\top \mathbf{w}$ betrachtet hat. Damit gilt nun auch

$$\begin{aligned} &- \sum_{i=1}^{n} \psi_\tau^* \left(\widetilde{Y}_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \\ \geq &- \sum_{i=1}^{n} \psi_\tau^* \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, -\mathbf{x}_i^\top \mathbf{w}\right) \mathbf{x}_i^\top \mathbf{w} \geq 0 \end{aligned}$$

und somit ist der Gradient der Zielfunktion des „kontaminierten“ Problems, $\nabla Z \left(\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \widetilde{\mathbf{Y}}), \mathbf{w}\right)$ für $d_i \geq 0$, sowie für alle Richtungen $\mathbf{w} \in \mathbb{R}^p$ in $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ nichtnegativ. Also minimiert $\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ die Zielfunktion des verschmutzten Problems und ist somit auch dessen Optimallösung, das heißt

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Dr} + \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})),$$

solange die Diagonalmatrix $\mathbf{D}$ nur nichtnegative Einträge hat. □

Unter Verwendung der Equivarianz-Eigenschaften aus Satz 3.2.1 kann man die eben bewiesene Aussage noch einmal umformulieren. Dazu betrachtet man erneut die Satzaussage

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Dr} + \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}))$$

und addiert auf beiden Seiten $-\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) \in \mathbb{R}^p$, was zu

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) - \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Dr} + \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})) - \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$$

führt. Aufgrund des dritten Unterpunktes von Satz 3.2.1 und der Parametrisierung

$$\boldsymbol{\gamma} := -\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) \in \mathbb{R}^p$$

folgt auf beiden Seiten

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Dr} + \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y}) - \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})).$$

Da $\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ gilt, erhält man zuletzt

$$\hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{r}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Dr}).$$

Durch diese Umformung der Aussage von Satz 3.3.1 sieht man noch deutlicher seine praktische Bedeutung: Da für die Diagonaleinträge der Matrix $\mathbf{D}$ nur nichtnegative Einträge in Frage kommen, ändert sich das geschätzte Regressionsquantil $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ nicht, solange die Responsevariablen Y_i nur derart verändert werden, dass die Residuen bezüglich dieser Beobachtungen ihre Vorzeichen nicht ändern, das heißt vereinfacht ausgedrückt, solange Y_i auf der selben Seite des gefitteten Regressionsquantils $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau; \mathbf{X}, \mathbf{Y})$ bleibt.

Eine genauere Betrachtung erfordert auch das Vorkommen des Prädiktorvektors $\mathbf{x}$ in der Einflussfunktion

$$EF\left((y, \mathbf{x}), \hat{\boldsymbol{\beta}}_F(\tau)\right) = Q^{-1}\mathbf{x}\,\mathrm{sgn}^*\left(y - \mathbf{x}^\top \hat{\boldsymbol{\beta}}_F(\tau)\right),$$

da eine Störung von $(\mathbf{x}, Y)$ durch ein stark abweichendes $\mathbf{x}$ zu einem beliebig großen Wert von $EF\left((y, \mathbf{x}), \hat{\boldsymbol{\beta}}_F(\tau)\right)$ führen kann. Das bedeutet, dass Schätzer Quantiler Regression durch den beschränkten Einfluss von Y zwar sehr robust bezüglich Response-Ausreißern sind, sich aber gegenüber Ausreißern in den Prädiktoren eher sensibel verhalten. Ein illustratives Beispiel dafür bekommt man, indem man einen einzigen Ausreißer $(\mathbf{x}_i, Y_i)$ immer weiter von der Masse der Beobachtungen entfernt, sodass die QR-Funktionen schließlich gezwungen sind durch diesen Punkt zu verlaufen. Dieses Verhalten kennt man bereits von den Kleinsten-Quadrate Schätzern und ist daher nicht überraschend.

Die Einflussfunktion ist ein unverzichtbares Werkzeug um die Sensitivität eines Schätzers

bezüglich infinitesimalen Störungen im Ausgangsmodell zu messen. Es besteht jedoch die Möglichkeit, dass verschiedene Methoden bezüglich dieser nahezu unendlich kleinen Kontaminierungen robust sind, aber auf kleine, endliche Störungen (nicht mehr infinitesimal klein) höchst sensibel reagieren. Dies ist beispielsweise beim α-getrimmten Mittel der Fall. Das α-getrimmte Mittel basiert auf dem zentralen $(1-2\alpha)$ Bereich der Verteilung und kann daher wie folgt angegeben werden:

$$\mu_\alpha = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y\mathrm{d}F(y),$$

für $0 < \alpha < 1/2$.

Dadurch sieht man, dass dieser Lokationsparameter auf Störungen der Größenordnung $\omega < \alpha$ nicht reagiert, aber ab einem $\omega > \alpha$ *zusammenbrechen* kann, da dann die Distanz zum α-getrimmten Mittel der „ungestörten“ Verteilung F sogar unendlich werden kann.

Ein Maß für die globale Robustheit, dass in diesem Zusammenhang auftaucht, ist der **Bruchpunkt**. Er gibt den kleinsten Prozentsatz an Kontaminierung an, sodass der Schätzer zusammenbricht, das heißt die Schätzung bezüglich der veränderten Stichprobe beziehungsweise der kontaminierten Verteilung weit entfernt vom ursprünglichen Ergebnis liegt. Formal wird er in Donoho & Huber (1983) für endliche Stichproben wie folgt definiert:

Definition 3.3.4 (Bruchpunkt für endliche Stichproben)**.** *Bezeichne $\mathbf{v}_0$ die Ausgangsstichprobe mit Umfang n und $\mathbf{v}_m$ jene Stichprobe, bei der m Einträge ersetzt wurden. Der **Bruchpunkt** eines Schätzers $\hat{\theta}_n$ basierend auf $\mathbf{v}_0$ wird durch*

$$b_n^*\left(\hat{\theta}_n, \mathbf{v}_0\right) = \frac{m^*(\mathbf{v}_0)}{n}$$

festgelegt, wobei $m^*(\mathbf{v}_0)$ die kleinste nichtnegative Zahl ist, für die

$$\sup_{\mathbf{v}_m} \left\|\hat{\theta}_n(\mathbf{v}_m) - \hat{\theta}_n(\mathbf{v}_0)\right\| = \infty$$

gilt. Also gibt $m^*(\mathbf{v}_0)$ die Anzahl der Werte an, durch deren Ersetzung es zu einem Zusammenbruch des betrachteten Schätzers kommt.
Ist $b_n^*\left(\hat{\theta}_n, \mathbf{v}_0\right)$ unabhängig von der Ausgangsstichprobe $\mathbf{v}_0$ so kann man den Bruchpunkt auch als

$$b^*(\hat{\theta}) = \lim_{n\to\infty} b_n^*\left(\hat{\theta}_n, \mathbf{v}_0\right)$$

definieren.

Beispiel 3.3.1. *Der Bruchpunkt für das Stichprobenmittel*

$$\overline{V}_n = \frac{1}{n}\sum_{i=1}^{n} V_i$$

ist unabhängig von der gewählten Ausgangsstichprobe, denn

$$b_n^*\left(\overline{V}_n, \mathbf{v}_0\right) = \frac{1}{n}$$

und damit ergibt sich

$$b^*(\overline{V}_n) = 0.$$

Geht man beim QR-Problem davon aus, dass es sich bei Kontaminierung um eine Störung der $(\mathbf{x}, Y)$-Paaren handelt, so ergibt sich, unabhängig von der gewählten Stichprobe, für den Schätzer $\hat{\boldsymbol{\beta}}(\tau)$ ein $m^* = 1$. Dies kommt daher, dass schon Verschieben eines einzigen Paares gleichzeitig in $\mathbf{x}$ und y Richtung hin zu „$\pm\infty$“ dazu führen kann, dass alle Quantilsfunktionen durch diesen einen Ausreißer verlaufen müssen und somit die Schätzung zusammenbricht. Diese Sensibilität bezüglich Störungen der Designbeobachtungen ist ein wohlbekanntes Problem für alle *M-Schätzer*.

Eine alternative Dartstellung, die besser zu den bisherigen Betrachtungen passt findet man in Hampel et al. (1968) oder auch in Maronna et al. (2006):

Definition 3.3.5 (Bruchpunkt bezüglich der gesamten Population)**.** *Sei $F(u)$ die Ausgangs- oder Modellverteilung und $F_\omega(u) = \omega H(u) + (1-\omega)\, F(u)$ die gestörte Verteilung, wobei hier $H(u)$ die kontaminierende Verteilung ist. Der Bruchpunkt ist unter diesen Voraussetzungen gegeben durch*

$$\omega^*\left(\hat{\theta}(F)\right) = \inf\left\{\omega : B\left(\omega; \hat{\theta}(F)\right) = \infty\right\}$$

mit

$$B\left(\omega; \hat{\theta}(F)\right) = \sup_{H(u)} \left|\hat{\theta}(F_\omega) - \hat{\theta}(F)\right|.$$

$B\left(\omega; \hat{\theta}(F)\right)$ *wird als maximaler Bias bezeichnet.*

Der Bruchpunkt gibt also an, inwieweit ein Schätzer von einer schlechten Datenlage beeinflusst wird, das heißt wie widerstandsfähig er gegenüber Ausreißern ist. Mit anderen Worten: Bekommt man noch sinnvolle Schätzungen, wenn die tatsächliche Verteilung der Daten stark von der angenommenen Modellverteilung abweicht! Als kontaminierende Verteilung $H(u)$ wurde zuvor immer die Einpunktverteilung $\delta_y(u)$ gewählt. Hier spricht man nun von Kontaminierung der Verteilung und nicht mehr von einer Störung der Designbeobachtungen.

Beispiel 3.3.2. *Man erhält als Bruchpunkte für den Erwartungswert $\hat{\theta}_E(F)$ und den Median $\hat{\theta}_{0.5}(F) = F^{-1}(0.5)$ bezüglich der Verteilung F*

$$\omega^*\left(\hat{\theta}_E(F)\right) = 0 \quad \textit{und} \quad \omega^*\left(\hat{\theta}_{0.5}(F)\right) = \frac{1}{2}.$$

Durch den höheren Bruchpunkt des Medians sieht man erneut, dass er ein robusterer Schätzer für einen Lokationsparameter ist als der Erwartungswert.

In Koenker (2005) und bei He et al. (1990) wird diese zweite Definition des Bruchpunktes dazu verwendet, um den Zusammenhang zwischen Maßen für das *Tailverhalten von Regressionsschätzern*, die ebenfalls die Güte eines Schätzers angeben, und deren Bruchpunkten analysiert. Hier wird auf eine genauere Betrachtung verzichtet.

3.4. Quantile Crossing

Eine der attraktivsten Eigenschaften Quantiler Regression ist die *Verteilungsfreiheit*. Es ist also möglich ohne Annahme einer bestimmten globalen Verteilung einzelne Teile der bedingten Responseverteilung näher zu betrachten beziehungsweise zu modellieren. Dazu genügt bereits lokale Information nahe dem zu schätzenden Quantil. Dies zeigte schon Satz 3.3.1, der die Aussage postulierte, dass geschätzte Quantilsfunktion nicht auf Störungen der Daten oberhalb oder unterhalb der geschätzten Quantilsfunktion reagieren. Auch das asymptotische Verhalten der quantilsspezifischen Regressionskoeffizienten, das später noch eingeführt wird, untermauert diese Feststellung.

Leider führt genau dies zu einem Problem bei der Modellierung: Der Vorteil eine Familie von Quantilsfunktionen unabhängig von globalen Annahmen zu schätzen, kann das Phänomen des *Quantile Crossings* hervorrufen. Dabei schneiden sich zwei gefittete Quantile, was den Grundsatz, dass Verteilungsfunktionen sowie deren Umkehrfunktionen monoton wachsend sein müssen, verletzt.

Wie jedoch der folgende Satz zeigt, treten solche Kreuzungen in der Regel nur im Randbereich des Designraums auf, da im Zentrum der Daten

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

die geschätzten Quantilsfunktionen

$$\hat{Q}_Y\left(\tau|\overline{\mathbf{x}}\right) = \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau)$$

monoton wachsend in τ sind.

Satz 3.4.1. *Die aus der Stichprobe resultierenden Pfade von* $\hat{Q}_Y\left(\tau|\overline{\mathbf{x}}\right)$ *sind monoton wachsend in* $\tau \in (0,1)$.

Beweis. Um die Aussage des Satzes zu beweisen zeigt man, dass aus $\tau_1 < \tau_2$

$$\overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_1) \leq \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_2)$$

folgt.

Im ersten Schritt betrachtet man dazu für ein beliebiges $\boldsymbol{\beta} \in \mathbb{R}^p$ und $u_i = Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$ die Differenz

$$\rho_{\tau_2}(u_i) - \rho_{\tau_1}(u_i) = u_i \left(\tau_2 - \mathbb{1}_{\{u_i<0\}}\right) - u_i \left(\tau_1 - \mathbb{1}_{\{u_i<0\}}\right)$$

und durch Zusammenfassen erhält man

$$\begin{aligned}\rho_{\tau_2}(u_i) - \rho_{\tau_1}(u_i) =& u_i \left(\tau_2 - \mathbb{1}_{\{u_i<0\}} - \tau_1 + \mathbb{1}_{\{u_i<0\}}\right) \\ =& u_i (\tau_2 - \tau_1).\end{aligned}$$

Durch Summation und mit Hilfe der Definition des Mittelwertes kommt man zu

$$\begin{aligned}\sum_{i=1}^{n} \left[\rho_{\tau_2}(u_i) - \rho_{\tau_1}(u_i)\right] &= \sum_{i=1}^{n} (\tau_2 - \tau_1) u_i \\ &= (\tau_2 - \tau_1) \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) \\ &= (\tau_2 - \tau_1) n \left(\overline{Y} - \overline{\mathbf{x}}^\top \boldsymbol{\beta}\right).\end{aligned}$$

Das ist also gleichbedeutend mit

$$\sum_{i=1}^{n} \rho_{\tau_2}\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) - \sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) = n (\tau_2 - \tau_1) \left(\overline{Y} - \overline{\mathbf{x}}^\top \boldsymbol{\beta}\right), \tag{3.12}$$

für ein beliebiges $\boldsymbol{\beta} \in \mathbb{R}^p$ und jedes $\tau_1, \tau_2 \in (0,1)$.

Im nächsten Schritt ersetzt man $\boldsymbol{\beta}$ durch die quantilsspezifischen Regressionskoeffizienten $\hat{\boldsymbol{\beta}}(\tau_j),\ j \in \{1,2\}$. Da $\hat{\boldsymbol{\beta}}(\tau_1)$ die Zielfunktion zur Schätzung des τ_1-Quantils minimiert gilt

$$\sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_1)\right) \leq \sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right)$$

und damit kann nun folgende Abschätzung durchgeführt werden:

$$\begin{aligned}&\sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_1)\right) + n (\tau_2 - \tau_1) \left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) \\ \leq &\sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) + n (\tau_2 - \tau_1) \left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_2)\right).\end{aligned}$$

Durch Umformen von Gleichung (3.12) kann man den rechten Term der Ungleichung wie folgt darstellen:

$$\sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) + n (\tau_2 - \tau_1) \left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) = \sum_{i=1}^{n} \rho_{\tau_2}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right)$$

und daher gilt dann

$$\sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) + n\left(\tau_2 - \tau_1\right)\left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right)\right) \leq \sum_{i=1}^{n} \rho_{\tau_2}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right). \quad (3.13)$$

Da $\hat{\boldsymbol{\beta}}\left(\tau_2\right)$ optimaler Koeffizientenschätzer für das τ_2-Quantil ist hält

$$\begin{aligned}\sum_{i=1}^{n} \rho_{\tau_2}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) &\leq \sum_{i=1}^{n} \rho_{\tau_2}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_1)\right) \\ &\overset{(3.12)}{=} \sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) + n\left(\tau_2 - \tau_1\right)\left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right).\end{aligned}$$

Verwendet man nun dieses Ergebnis in (3.13), so erhält man die Ungleichung

$$\begin{aligned}&\sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) + n\left(\tau_2 - \tau_1\right)\left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right)\right) \\ \leq &\sum_{i=1}^{n} \rho_{\tau_1}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) + n\left(\tau_2 - \tau_1\right)\left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right).\end{aligned}$$

Die Summen der rechten und der linken Seite der Ungleichung stimmen überein und können daher weggelassen werden und man bekommt

$$n\left(\tau_2 - \tau_1\right)\left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right)\right) \leq n\left(\tau_2 - \tau_1\right)\left(\overline{Y} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right),$$

was durch Umformen zu

$$n\left(\tau_2 - \tau_1\right)\left(\cancel{\overline{Y}} - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right) - \cancel{\overline{Y}} + \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) \leq 0$$

wird. Da laut Annahme $\tau_1 < \tau_2$ gilt, ist $\tau_2 - \tau_1 > 0$ und somit muss

$$\overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right) - \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right) \leq 0 \Leftrightarrow \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right) \leq \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right)$$

sein, da sonst die Ungleichung nicht erfüllt ist. □

Natürlich bedeutet dies nicht automatisch, dass die geschätzten Pfade der Quantilsfunktion auch in allen anderen Punkten $\mathbf{x} \neq \overline{\mathbf{x}}$ monoton wachsend sind. Ganz im Gegenteil: Nimmt man eine lineare Form des Modells an, so müssen sich die einzelnen $\hat{Q}_Y(\tau|\mathbf{x})$ sogar irgendwann weiter weg von $\overline{\mathbf{x}}$ schneiden, außer man erhält tatsächlich für alle Werte τ parallele Schätzungen. Findet man durch Überprüfen einzelner $\mathbf{x}$-Werte eine große Anzahl an Punkten in denen die Voraussetzung einer monoton wachsenden Quantilsfunktionschätzung verletzt ist, so spricht dies für falsch spezifizierte Effekte der Kovariablen und die gewählte Form des Modells sollte überdacht werden. Mehr zum Umgang mit einem falsch spezifiziertem Modell findet sich in Koenker (2005).

4. Inferenz in QR-Modellen

Im folgenden Kapitel werden verschiedene Inferenz-Methoden für QR-Modelle betrachtet und deren Vor- und Nachteile gegenübergestellt. Ideal wäre es, wenn man bereits bei endlichem Stichprobenumfang Inferenz über die entstandenen QR-Schätzer betreiben könnte, wie es beispielsweise in der klassischen Linearen Regression aufgrund der Annahme unabhängiger und identisch verteilter Gaußscher-Fehlerterme der Fall ist. Aber das alleine reicht noch nicht aus, denn sowie auch bei den Least-Squares Schätzern muss man zu asymptotischen Annäherungen übergehen, sobald man erkennt, dass die idealisierten Bedingungen einer Gaußschen Fehlerverteilung nicht mehr erfüllt sind.

4.1. Die endliche Stichproben Verteilung von Regressionsquantilen

Zuerst startet man mit einer kurzen Einführung in die endliche Stichproben Theorie von QR-Schätzern. Man betrachtet also Zufallsvariablen $Y_1, \ldots, Y_n$, die iid-verteilt sind mit allgemeiner Verteilungsfunktion F, wobei man annimmt, dass F in der Nähe von $q_\tau = F^{-1}(\tau)$ eine stetige Dichte f mit $f(q_\tau) > 0$ hat. Die Zielfunktion zur Berechnung des τ-Stichprobenquantil hat die Form

$$\hat{q}_\tau = \inf_q \left\{ q \in \mathbb{R} : \min_q \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - q \right) \right\}$$

und ist somit eine Summe konvexer Funktionen und daher ebenfalls konvex. Um nun den Gradienten $g_n(q)$ der Zielfunktion zu bestimmen betrachtet man zuerst die Ableitung der Funktion

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}q} \rho_\tau \left(Y_i - q \right) =& \frac{\mathrm{d}}{\mathrm{d}q} \left(Y_i - q \right) \left(\tau - \mathbb{1}_{\{Y_i < q\}} \right) \\ =& \frac{\mathrm{d}}{\mathrm{d}q} Y_i \left(\tau - \mathbb{1}_{\{Y_i < q\}} \right) - \frac{\mathrm{d}}{\mathrm{d}q} q \left(\tau - \mathbb{1}_{\{Y_i < q\}} \right), \end{aligned}$$

wobei der Fall $Y_i = q$ aufgrund der Annahme einer stetigen Verteilung der Y_i, das heißt es gilt $\mathbb{P}\left[Y_i = q \right] = 0$, vernachlässigt wird. Andernfalls würde hier auch die bereits zu Beginn der Arbeit vorgestellte Subgradienten-Interpretation zum Einsatz kommen. Mit Hilfe der Produktregel und der Tatsache, dass die Ableitungen der Indikatorfunktionen

verschwinden, da die Unstetigkeitsstellen ausgeschlossen wurden, erhält man

$$\frac{\mathrm{d}}{\mathrm{d}q}\rho_\tau\left(Y_i - q\right) = -\left(\tau - \mathbb{1}_{\{Y_i<q\}}\right).$$

Durch diese Beobachtungen kann die Ableitung der Zielfunktion ermittelt werden:

$$g_n(\tau) = \frac{\mathrm{d}}{\mathrm{d}q}\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i - q\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Y_i<q\}} - \tau\right).$$

Aufgrund der Tatsache, dass der Gradient monoton wachsend in q ist, kann $\hat{q}_\tau > q$ nur genau dann gelten, wenn $g_n(q) < 0$ ist. Somit kann nun die Wahrscheinlichkeit $\mathbb{P}\left[\hat{q}_\tau > q\right]$ durch $\mathbb{P}\left[g_n(q) < 0\right]$ ausgedrückt werden. Genauere Betrachtung dieses Terms und Einsetzen der Darstellung von $g_n(q)$ liefert:

$$\begin{aligned}\mathbb{P}\left[\hat{q}_\tau > q\right] = \mathbb{P}\left[g_n(q) < 0\right] =& \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Y_i<q\}}\right) - \tau < 0\right] \\ =& \mathbb{P}\left[\sum_{i=1}^{n}\left(\mathbb{1}_{\{Y_i<q\}}\right) < n\tau\right].\end{aligned}$$

Die Summe entspricht einer binomialverteilten Zufallsvariable $B\left(n, F(q)\right)$, womit

$$\mathbb{P}\left[\hat{q}_\tau > q\right] = \mathbb{P}\left[B\left(n, F(q)\right) < n\tau\right]$$

folgt.

Sei nun $m = \lceil n\tau \rceil$ die kleinste ganze Zahl die größer oder gleich $n\tau$ ist. Im nächsten Schritt betrachtet man

$$\begin{aligned}G(q) = \mathbb{P}\left[\hat{q}_\tau \leq q\right] =& 1 - \mathbb{P}\left[\hat{q}_\tau > q\right] \\ =& 1 - \mathbb{P}\left[B\left(n, F(q)\right) < n\tau\right].\end{aligned}$$

Verwendet man nun die Wahrscheinlichkeitsfunktion der Binomialverteilung mit Parametern n und $F(q)$ so erhält man

$$\begin{aligned}G(q) =& 1 - \sum_{j=0}^{m-1}\binom{n}{j}F(q)^j\left(1 - F(q)\right)^{n-j} \\ =& \sum_{j=m}^{n}\binom{n}{j}F(q)^j\left(1 - F(q)\right)^{n-j}.\end{aligned}$$

Um weiter Überlegung anzustellen, benötigt man zuerst die Definition der *Unvollständigen-Beta-Funktion.*

Definition 4.1.1 (Unvollständige-Beta-Funktion)**.** *Die* ***Unvollständige-Beta-***

Funktion *ist für beliebige* $x > 0$ *durch*

$$Be\,(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}, \quad a, b > 0$$

definiert und die ***normierte Unvollständige-Beta-Funktion*** *ergibt sich durch*

$$I_x(a, b) = \frac{Be\,(x; a, b)}{Be\,(1; a, b)}.$$

Eine weitere nützliche Darstellungsform dieser Funktion ist

$$I_x(a, b) = \sum_{j=a}^{a+b-1} \binom{a+b-1}{j} x^j (1-x)^{a+b-1-j}. \tag{4.1}$$

Mit Hilfe von Darstellung (4.1) erkennt man, dass durch Wahl der Parameter $a = m$ und $b = n - m + 1$

$$\begin{aligned} G(q) =& I_{F(q)}(m, n-m+1) \\ =& \frac{\int_0^{F(q)} t^{m-1}(1-t)^{n-m+1-1}\mathrm{d}t}{\int_0^1 t^{m-1}(1-t)^{n-m+1-1}\mathrm{d}t} = \frac{\int_0^{F(q)} t^{m-1}(1-t)^{n-m}\mathrm{d}t}{\int_0^1 t^{m-1}(1-t)^{n-m}\mathrm{d}t} \end{aligned} \tag{4.2}$$

folgt.

Nun berechnet man seperat den Nenner dieses Terms. Im ersten Schritt erhält man durch partielle Integration

$$\int_0^1 t^{m-1}(1-t)^{n-m}\mathrm{d}t = \frac{t^m}{m}(1-t)^{n-m}|_{t=0}^1 + \int_0^1 (n-m)(1-t)^{n-m-1}\frac{t^m}{m}\mathrm{d}t.$$

Der erste Term der rechten Seite verschwindet sowohl für Ober- als auch Untergrenze und somit erhält man

$$\begin{aligned} \int_0^1 t^{m-1}(1-t)^{n-m}\mathrm{d}t &= \frac{n-m}{m}\int_0^1 (1-t)^{n-m-1}t^m\mathrm{d}t \\ &= \frac{n-m}{m}\int_0^1 (1-t)^{(n-m)-1}t^{(m-1)+1}\mathrm{d}t. \end{aligned}$$

Im j-ten Schritt der Integration bekommt man durch analoges Vorgehen

$$\begin{aligned} &\int_0^1 t^{m-1}(1-t)^{n-m}\mathrm{d}t = \\ &\frac{(n-m)(n-m-1)\cdots(n-m-(j-1))}{m(m+1)\cdots(m+(j-1))}\int_0^1 (1-t)^{(n-m)-j}t^{(m-1)+j}\mathrm{d}t. \end{aligned}$$

Dies wird solange fortgeführt bis $n - m - j = 0$ was gleichbedeutend mit $j = n - m$ ist und dies liefert dann

$$\int_0^1 t^{m-1}(1-t)^{n-m}\mathrm{d}t = \frac{(n-m)(n-m-1)\cdots 1}{m(m+1)\cdots(n-1)} \int_0^1 (1-t)^0 t^{n-1}\mathrm{d}t.$$

Das letzte Integral lässt sich nun einfach berechnen und man erhält die Darstellung

$$\begin{aligned}\int_0^1 t^{m-1}(1-t)^{n-m}\mathrm{d}t =& \frac{(n-m)(n-m-1)\cdots 1}{m(m+1)\cdots(n-1)} \frac{t^n}{n}\Big|_{t=0}^1 \\ =& \frac{(n-m)!}{m(m+1)\cdots(n-1)n} \\ =& \frac{(n-m)!(m-1)!}{(n-1)!n}.\end{aligned}$$

Der letzte Gleichheit hält, da $m \leq n$ gilt. Somit kommt man schließlich zur endgültigen Darstellung des Nenners

$$\begin{aligned}\frac{1}{\int_0^1 t^{m-1}(1-t)^{n-m}\mathrm{d}t} =& n\frac{(n-1)!}{(n-m)!(m-1)!} \\ =& n\frac{(n-1)!}{((n-1)-(m-1))!(m-1)!} \\ =& n\binom{n-1}{m-1}.\end{aligned}$$

Setzt man dies nun in (4.2) ein, so liefert dies schlussendlich das Ergebnis

$$G(q) = n\binom{n-1}{m-1} \int_0^{F(q)} t^{m-1}(1-t)^{n-m}\mathrm{d}t.$$

Durch Ableiten erhält man dann die Dichtefunktion für $\hat{q}_\tau$

$$g(q) = n\binom{n-1}{m-1} F(q)^{m-1} \left(1 - F(q)\right)^{n-m} f(q),$$

wobei $f(q)$ als innere Ableitung des Integrals dazukommt.

Diesen Ansatz kann man auch dazu verwenden um approximative Konfidenzintervalle für das theoretische Quantil q_τ der Form

$$\mathbb{P}\left[\hat{q}_{\tau_1} < q_\tau < \hat{q}_{\tau_2}\right] = 1 - \alpha$$

zu generieren, wobei τ_1 und τ_2 so gewählt werden müssen, dass sie

$$\mathbb{P}\left[n\tau_1 < B(n,\tau) < n\tau_2\right] = 1 - \alpha$$

erfüllen. Diese Konfidenzintervalle haben den großen Vorteil, dass sie im Falle einer stetigen Funktion $F(\cdot)$ *verteilungsfrei* sind, das heißt sie halten ungeachtet der Verteilung $F(\cdot)$.

In Koenker & Bassett (1978) wird die endliche Stichproben Dichte der QR-Koeffizientenschätzer $\hat{\boldsymbol{\beta}}(\tau)$ hergeleitet und in folgendem Satz zusammengefasst.

Satz 4.1.1. *Man betrachtet das lineare Modell*

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

für $i = 1, \ldots, n$, wobei die Fehlerterme $\{\varepsilon_i\}$ iid-verteilt sind und eine allgemeine Verteilungsfunktion $F(\cdot)$ besitzen. Diese hat eine strikt positive Dichte $f(\cdot)$ im Punkt $F^{-1}(\tau)$. Dann hat die Dichte von $\hat{\boldsymbol{\beta}}(\tau)$ die Form

$$g(\mathbf{b}) = \sum_{h \in \mathcal{H}} \mathbb{P}\left[\boldsymbol{\xi}_h(\mathbf{b}) \in \mathcal{C}\right] |\mathbf{X}(h)| \prod_{i \in h} f\left(\mathbf{x}_i^\top (\mathbf{b} - \boldsymbol{\beta}(\tau)) + F^{-1}(\tau)\right),$$

mit

$$\boldsymbol{\xi}_h(\mathbf{b}) = \sum_{i \in \overline{h}} \psi_\tau\left(Y_i - \mathbf{x}_i^\top \mathbf{b}\right) \mathbf{x}_i^\top \mathbf{X}(h)^{-1}, \quad \mathbf{b} \in \mathbb{R}^p.$$

Weiters bezeichnet $\mathcal{C}$ den Würfel $[\tau - 1, \tau]^p$.

Der Beweis nützt die Subgradienten-Bedingung aus, wobei hier auf eine detailierte Ausführung verzichtet wird. Diese findet sich beispielsweise in Koenker & Bassett (1978) oder etwas weniger detailiert in Koenker (2005).

Von einem praktischen Standpunkt aus gesehen, muss man aber leider anhand dieser Darstellung erkennen, dass die $\binom{n}{p}$ Summanden bei den meisten Anwendungen ineffizient sind und daher der Übergang zu asymptotischen Approximationen von Vorteil ist. Bei Methoden, die sich ausschließlich auf endliche Stichproben beziehen, hat man auch eine Menge gewagter Annahmen zu treffen, was ebenfalls für die Betrachtung einer „asymptotischen Theorie“ spricht.

4.2. Asymptotik von QR-Schätzern

Um die Güte beziehungsweise den Erfolg neuer Techniken zu messen, stellen sich drei Fragen:

1. Liefert die verwendete Methode ein Ergebnis, welches in gewissem Sinn gegen ein bestimmtes Objekt konvergiert unter der Annahme, dass die zugrunde liegenden Daten bestimmten Voraussetzungen unterliegen?
2. Unter welchen genauen Voraussetzungen funktioniert die Prozedur überhaupt?
3. Wie „gut“ arbeitet sie im Vergleich zu anderen vergleichbaren Ansätzen?

Zum Einstieg wird die asymptotische Verteilung des τ-Stichprobenquantils genauer betrachtet, das das Quantil der Verteilung einer Zufallsvariable schätzt und darauf aufbauend wird dann jene der quantilsspezifischen Regressionskoeffizienten näher erläutert. Diese Erkenntnisse können in weiterer Folge dazu verwendet werden, die zuvor gestellten Fragen (großteils) zu beantworten.

4.2.1. Asymptotische Verteilung des τ-Stichprobenquantils

Auch hier beginnt man mit der Betrachtung des Gradienten

$$g_n(q) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Y_i<q\}} - \tau\right)$$

und wegen seiner Monotonie in q gilt erneut, dass $\hat{q}_\tau > q$ nur genau dann halten kann, wenn $g_n(q) < 0$ gilt. Somit kann man analog zu obigen Überlegungen folgende Wahrscheinlichkeit ermitteln, wobei q_τ das wahre Quantil von $F(\cdot)$ zum Niveau τ ist:

$$\begin{aligned}\mathbb{P}\left[\sqrt{n}\left(\hat{q}_\tau - q_\tau\right) > \delta\right] =& \mathbb{P}\left[\hat{q}_\tau > q_\tau + \frac{\delta}{\sqrt{n}}\right] \\ =& \mathbb{P}\left[g_n\left(q_\tau + \frac{\delta}{\sqrt{n}}\right) < 0\right]. \end{aligned} \tag{4.3}$$

Durch die Definition des Gradienten erhält man

$$\mathbb{P}\left[\sqrt{n}\left(\hat{q}_\tau - q_\tau\right) > \delta\right] = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\left\{Y_i<q_\tau+\delta/\sqrt{n}\right\}} - \tau\right) < 0\right].$$

Die Summanden sind nichts anderes als Bernoulli-Variablen, die mit Wahrscheinlichkeit $F\left(q_\tau + \delta/\sqrt{n}\right)$ den Wert $(1-\tau)$ und $(-\tau)$ mit Wahrscheinlichkeit $\left(1 - F\left(q_\tau + \delta/\sqrt{n}\right)\right)$ annehmen. Der Erwartungswert des Gradienten berechnet sich also durch

$$\mathbb{E}\left[g_n\left(q_\tau + \frac{\delta}{\sqrt{n}}\right)\right] = F\left(q_\tau + \frac{\delta}{\sqrt{n}}\right)(1-\tau) + \left(1 - F\left(q_\tau + \frac{\delta}{\sqrt{n}}\right)\right)(-\tau).$$

Durch Ausmultiplizieren und Wegkürzen kommt man schließlich zu

$$\mathbb{E}\left[g_n\left(q_\tau + \frac{\delta}{\sqrt{n}}\right)\right] = F\left(q_\tau + \frac{\delta}{\sqrt{n}}\right) - \tau. \tag{4.4}$$

Nun gilt weiters, dass

$$q_\tau = F^{-1}(\tau) \Leftrightarrow \tau = F(q_\tau)$$

und damit ergibt sich für den Erwartungswert (4.4)

$$F\left(q_\tau + \frac{\delta}{\sqrt{n}}\right) - \tau = F\left(q_\tau + \frac{\delta}{\sqrt{n}}\right) - F(q_\tau).$$

Erweitert man diesen Ausdruck nun durch

$$1 = \frac{\delta}{\sqrt{n}} / \frac{\delta}{\sqrt{n}},$$

so liefert dies den Term

$$\mathbb{E}\left[g_n\left(q_\tau + \frac{\delta}{\sqrt{n}}\right)\right] = \frac{\delta}{\sqrt{n}} \frac{F\left(q_\tau + \frac{\delta}{\sqrt{n}}\right) - F(q_\tau)}{\frac{\delta}{\sqrt{n}}}$$

und man erkennt, dass es sich bei dem zweiten Bruch um einen Differenzenquotienten handelt, der, sofern man $\delta/\sqrt{n}$ gegen Null gehen lässt, zum Differentialquotienten (vergleiche Definition (3.3.3)) wird und damit erhält man

$$\mathbb{E}\left[g_n\left(q_\tau + \frac{\delta}{\sqrt{n}}\right)\right] = \frac{\delta}{\sqrt{n}} \frac{F\left(q_\tau + \frac{\delta}{\sqrt{n}}\right) - F(q_\tau)}{\frac{\delta}{\sqrt{n}}} \longrightarrow \frac{\delta}{\sqrt{n}} f(q_\tau). \tag{4.5}$$

Nun folgt die Berechnung der Varianz des Gradienten:

$$\begin{aligned}\mathrm{Var}\left[g_n\left(q_\tau + \frac{\delta}{\sqrt{n}}\right)\right] =& \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\left\{Y_i < q_\tau + \delta/\sqrt{n}\right\}}\right) - \tau\right] \\ \overset{\tau \text{ konst.}}{=}& \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}\left(\mathbb{1}_{\left\{Y_i < q_\tau + \delta/\sqrt{n}\right\}}\right)\right].\end{aligned}$$

Die Summanden sind iid-verteilte Bernoulli-Zufallsvariablen und daher ist die Summe binomialverteilt mit Parametern $(n, F\left(q_\tau + \delta/\sqrt{n}\right))$. Aufgrund der Varianz einer Bino-

mialverteilung erhält man

$$\begin{aligned}\operatorname{Var}\left[g_n\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)\right] =& \frac{n}{n^2}F\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)\left(1-F\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)\right)\\ =& \frac{1}{n}F\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)\left(1-F\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)\right).\end{aligned}$$

Für $\delta/\sqrt{n}\to 0$ strebt $F\left(q_\tau+\delta/\sqrt{n}\right)$ gegen $F\left(q_\tau\right)=\tau$ und somit bekommt man

$$\operatorname{Var}\left[g_n\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)\right]\longrightarrow\frac{\tau\left(1-\tau\right)}{n}. \tag{4.6}$$

Jetzt setzt man die beiden Ergebnisse (4.5) und (4.6) in die zu betrachtende Wahrscheinlichkeit (4.2) ein, was zu

$$\mathbb{P}\left[g_n\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)<0\right]=\mathbb{P}\left[\frac{g_n\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)-\frac{\delta}{\sqrt{n}}f(q_\tau)}{\sqrt{\tau(1-\tau)/n}}<\frac{-\frac{\delta}{\sqrt{n}}f(q_\tau)}{\sqrt{\tau(1-\tau)/n}}\right]$$

führt. Nach Wegkürzen von $\sqrt{n}$ auf der rechten Seite und mit der Substitution

$$\omega^2:=\frac{\tau(1-\tau)}{f^2\left(q_\tau\right)}\quad\Leftrightarrow\quad\omega^{-1}=\frac{f\left(q_\tau\right)}{\sqrt{\tau(1-\tau)}} \tag{4.7}$$

erhält man

$$\mathbb{P}\left[g_n\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)<0\right]=\mathbb{P}\left[\frac{g_n\left(q_\tau+\frac{\delta}{\sqrt{n}}\right)-\frac{\delta}{\sqrt{n}}f(q_\tau)}{\sqrt{\tau(1-\tau)/n}}<-\omega^{-1}\delta\right]\longrightarrow\Phi\left(-\omega^{-1}\delta\right),$$

wobei Letzteres durch die Anwendung des *Satzes von Moivre-Laplace*, einem Spezialfall des Zentralen Grenzwertsatzes, folgt. Dies ist aber gleichbedeutend mit

$$\mathbb{P}\left[\sqrt{n}\left(\hat{q}_\tau-q_\tau\right)>\delta\right]\longrightarrow\Phi\left(-\omega^{-1}\delta\right)=1-\Phi\left(\omega^{-1}\delta\right).$$

Weiters kann man daraus schließen, dass

$$\mathbb{P}\left[\sqrt{n}\left(\hat{q}_\tau-q_\tau\right)\leq\omega\delta\right]\longrightarrow 1-\left(1-\Phi\left(\omega^{-1}\omega\delta\right)\right)=\Phi\left(\delta\right)$$

gilt, was äquivalent ist mit

$$\mathbb{P}\left[\frac{\sqrt{n}}{\omega}\left(\hat{q}_\tau-q_\tau\right)\leq\delta\right]\longrightarrow\Phi\left(\delta\right),$$

also

$$\frac{\sqrt{n}}{\omega}\left(\hat{q}_\tau - q_\tau\right) \xrightarrow{d} \mathcal{N}\left(0,1\right)$$

impliziert.

Das heißt zusammengefasst kommt man zu dem Schluss, dass

$$\sqrt{n}\left(\hat{q}_\tau - q_\tau\right) \xrightarrow{d} \mathcal{N}\left(0,\omega^2\right)$$

gilt, wobei man hier von *Konvergenz in Verteilung* spricht. Man kann also feststellen, dass die Präzision beziehungsweise die Genauigkeit des Schätzers $\hat{q}_\tau$ (= Stichprobenquantil als Schätzer für Populationsquantil q_τ zum Niveau τ) zwei Einflussfaktoren hat: Zum einen den $\tau(1-\tau)$-Effekt, der dazu tendiert den Schätzer des Quantils $\hat{q}_\tau$ in den Tails der Verteilung genauer zu machen, da der Nenner $h(\tau) = \tau(1-\tau) = -\tau^2+\tau$ als Funktion in τ betrachtet, seine niedrigsten Werte für sehr kleine und sehr große Werte von τ annimmt, das heißt also im rechten und linken Randbereich der Verteilung. Andererseits gibt es noch den Einfluss des Dichteterms auf die asymptotische Varianz des Schätzers ω^2, der $\hat{q}_\tau$ in Bereichen mit geringer Dichte unpräziser macht. Der zweite Faktor dominiert in der Regel den $\tau(1-\tau)$-Effekt.

Konfidenzintervall des τ-Quantils

Kennt man einen Schätzer des Nuisance-Parameters ω in (4.7), so können sehr leicht Konfidenzintervalle für q_τ angegeben werden. In diesem Fall gibt es aber auch einen direkteren Zugang, der keine Betrachtung von ω benötigt.

Angenommen man möchte folgende Hypothese testen:

$$H_0 : q_\tau = q_\tau^0.$$

Unter dieser Nullhypothese ist die Teststatistik

$$Z_n(q_\tau^0) = \sum_{i=1}^{n} \mathbb{1}_{\{Y_i < q_\tau^0\}}$$

binomialverteilt mit Parametern (n,τ), denn in diesem Fall ist q_τ^0 das τ-Quantil unter H_0 und somit ist $\mathbb{1}_{\{Y_i<q_\tau^0\}} = 1$ mit Wahrscheinlichkeit τ und $\mathbb{1}_{\{Y_i<q_\tau^0\}} = 0$ mit Wahrscheinlichkeit $(1-\tau)$.

Die Teststatistik berechnet die Anzahl der Realisierungen von Y_i, die unter dem angenommen τ-Quantil q_τ^0 liegen. Diese Summe sollte aufgrund der Definition des τ-Stichprobenquantils also nahe ihres Erwartungswertes $n\tau$ liegen und somit wird H_0

verworfen, wenn

$$\left|Z_n(q_\tau^0) - n\tau\right| \quad \Leftrightarrow \quad T_n(q_\tau^0) = \left|\frac{1}{n}Z_n(q_\tau^0) - \tau\right|$$

ausreichend groß wird.

Ein großer Vorteil dieses Tests ist zum Einen die bereits erwähnte Unabhängigkeit der asymptotischen Varianz ω^2 von $\sqrt{n}\hat{q}_\tau$ und zum anderen, dass die Verteilung der Teststatistik unter der Nullhypothese vollkommen unabhängig von der ursprünglichen Verteilung $F(\cdot)$ der Beobachtungen ist. T_n entspricht auch dem zuvor ausführlich betrachteten Gradienten $g_n(q_\tau^0)$ ausgewertet an der Stelle q_τ^0.

Um nun Konfidenzintervalle zum Niveau $1-\alpha$ zu konstruieren muss man die Menge

$$K_\alpha = \{q_\tau : T_n(q_\tau) \text{ verwirft zum Niveau } \alpha \text{ nicht}\}$$

betrachten. Da $T_n(q_\tau)$ stückweise konstant und monoton wachsend ist, ist man auf eine endliche Anzahl verschiedener Konfidenzintervalle der Form

$$K_\alpha = \left[Y_{(U)}, Y_{(O)}\right]$$

beschränkt, welche auf den beiden Ordnungsstatistiken $Y_{(U)}$ und $Y_{(O)}$ beruhen. Nun kommt es auf die Wahl der Indizes U und O an: Sie können direkt von der Tatsache abgeleitet werden, dass $T_n(q_\tau)$ binomialverteilt ist oder man kann sie im Fall großer Stichproben auch durch Approximation mit Hilfe der Normalverteilung angeben. Diese so approximierten Indizes haben dann die Form

$$\{U, O\} = n\tau \pm z_{1-\frac{\alpha}{2}}\sqrt{n\tau(1-\tau)},$$

wobei $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ das $\left(1-\frac{\alpha}{2}\right)$-Quantil der Standardnormalverteilung ist.

Zhou & Portnoy (1996) behandeln in ihrer Arbeit eine Erweiterung dieses Ansatzes zur Konstruktion von Prädiktionsintervallen für allgemeine QR-Modelle. Diese Intervalle basierend auf geschätzten interquantilen Bereichen eignen sich besonders gut zur Modellierung von Fällen in denen von einer asymmetrischen Dichte ausgegangen wird.

4.2.2. Asymptotische Verteilung von QR-Koeffizientenschätzern

Man betrachtet nun eine allgemeine Form eines linearen QR-Modells. Seien dazu $Y_1, Y_2, \ldots$ unabhängige Zufallsvariablen, wobei jede von ihnen eine Verteilungsfunktion $F_1, F_2, \ldots$ besitzt. Weiters nimmt man an, dass die bedingte Quantilsfunktion zum Niveau τ die Form

$$Q_{Y_i}(\tau|\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$$

hat, das heißt linear im Prädiktor $\mathbf{x}$ ist. Die bedingte Verteilungsfunktion der Y_i wird im Folgenden durch

$$\mathbb{P}\left[Y_i < y|\mathbf{x}_i\right] = F_{Y_i}(y|\mathbf{x}_i) = F_i(y)$$

ausgedrückt und damit erhält man für die bedingte Quantilsfunktion von Y_i die Darstellung

$$Q_{Y_i}(\tau|\mathbf{x}_i) = F_{Y_i}^{-1}(\tau|\mathbf{x}_i) =: q_i(\tau),$$

das heißt, sie steht für das theoretische Quantil der bedingten Verteilung der Zufallsvariable Y_i gegeben $\mathbf{x}_i$ zum Niveau τ.

Untersucht wird nun das asymptotische Verhalten der Schätzer für die quantilsspezifischen Regressionskoeffizienten $\hat{\boldsymbol{\beta}}(\tau)$, welche durch

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)$$

berechnet werden. Um eine asymptotische Verteilung des Schätzers ermitteln zu können, müssen zuerst einige Regularitätsbedingungen aufgestellt werden.

Bedingung I: Die Verteilungsfunktionen $\{F_i\}$ sind absolut stetig mit stetigen Dichten $\{f_i(q)\}$, welche in den Punkten $q_i(\tau)$ deutlich von 0 und ∞ verschieden sind.

Bedingung II: Es existieren positiv definierte Matrizen $\mathbf{D}_0$ und $\mathbf{D}_1(\tau)$, sodass

1. $\lim_{n\to\infty} n^{-1} \sum \mathbf{x}_i\mathbf{x}_i^\top = \mathbf{D}_0$
2. $\lim_{n\to\infty} n^{-1} \sum f_i(q_i(\tau))\mathbf{x}_i\mathbf{x}_i^\top = \mathbf{D}_1(\tau)$
3. $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|_2 / \sqrt{n} \to 0$

gilt.

Die Bedingungen **II**.1. und **II**.3. garantieren, dass der Zentrale Grenzwertsatz (ZGWS) für Zufallsvariablen, die nicht identisch verteilt sind, anwendbar ist. Man spricht in diesem Zusammenhang von *Lindeberg-Bedingungen* und dem *Lindeberg-Feller-ZGWS*, der eben ein Spezialfall des Zentralen Grenzwertsatzes für iid-verteilte Zufallsvariablen ist.

Nun kann also die Hauptaussage getätigt werden:

Satz 4.2.1 (Asymptotische Verteilung von $\hat{\boldsymbol{\beta}}(\tau)$)**.** *Unter den Bedingungen **I** und **II** gilt:*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) \to \mathcal{N}\left(0, \tau(1-\tau)\mathbf{D}_1^{-1}(\tau)\mathbf{D}_0\mathbf{D}_1^{-1}(\tau)\right).$$

Beweis. Um das Verhalten von $\hat{\boldsymbol{\delta}}_n := \sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right)$ herzuleiten betrachtet man die

Zielfunktion

$$Z_n(\boldsymbol{\delta}) = \sum_{i=1}^{n} \left(\rho_\tau \left(u_i - \mathbf{x}_i^\top \boldsymbol{\delta}/\sqrt{n} \right) - \rho_\tau(u_i) \right), \quad \boldsymbol{\delta} \in \mathbb{R}^p \tag{4.8}$$

mit $u_i = Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$. Der Index n bei $\hat{\hat{\boldsymbol{\delta}}}_n$ soll verdeutlichen, dass der Schätzer aus einer Stichprobe mit Umfang n hervorgeht. $Z_n(\boldsymbol{\delta})$ ist als Summe konvexer Funktionen wieder konvex.

Weiters folgt durch Einsetzen von $\hat{\hat{\boldsymbol{\delta}}}_n$ in (4.8) und mit $u_i = Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$

$$\begin{aligned}
Z_n \left(\hat{\hat{\boldsymbol{\delta}}}_n \right) &= \sum_{i=1}^{n} \left(\rho_\tau \left(u_i - \mathbf{x}_i^\top \hat{\hat{\boldsymbol{\delta}}}_n/\sqrt{n} \right) - \rho_\tau(u_i) \right) \\
&= \sum_{i=1}^{n} \left(\rho_\tau \left(u_i - \mathbf{x}_i^\top \sqrt{n} \left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau) \right)/\sqrt{n} \right) - \rho_\tau(u_i) \right) \\
&= \sum_{i=1}^{n} \left(\rho_\tau \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau) + \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) \right) - \rho_\tau(u_i) \right) \\
&= \sum_{i=1}^{n} \left(\rho_\tau \left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau) \right) - \rho_\tau(u_i) \right).
\end{aligned}$$

Da $\hat{\boldsymbol{\beta}}(\tau)$ aber genau jener Parameter ist, der den ersten Teil dieser Zielfunktion

$$\sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)$$

minimiert und weil der zweite Teil $\rho_\tau(u_i)$ unabhängig von $\boldsymbol{\delta}$ ist, erkennt man, dass

$$\hat{\hat{\boldsymbol{\delta}}}_n = \underset{\boldsymbol{\delta} \in \mathbb{R}^p}{\operatorname{argmin}} \, Z_n(\boldsymbol{\delta})$$

gilt. Mit Hilfe der sogenannten *Knight-Identität* (Knight, 1998)

$$\rho_\tau (u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v \left(\mathbb{1}_{\{u \le s\}} - \mathbb{1}_{\{u \le 0\}} \right) \mathrm{d}s,$$

mit $\psi_\tau(u) = \tau - \mathbb{1}_{\{u \le 0\}}$ und der Parameterwahl

$$u := u_i, \quad v = v_{ni} := \frac{\mathbf{x}_i^\top \boldsymbol{\delta}}{\sqrt{n}}$$

lässt sich $Z_n(\boldsymbol{\delta})$ in (4.8) wie folgt darstellen:

$$\begin{aligned} Z_n(\boldsymbol{\delta}) &= \sum_{i=1}^{n} \left(-\frac{\mathbf{x}_i^\top \boldsymbol{\delta}}{\sqrt{n}} \psi_\tau (u_i) + \int_0^{v_{ni}} \left(\mathbb{1}_{\{u_i \leq s\}} - \mathbb{1}_{\{u_i \leq 0\}} \right) \mathrm{d}s \right) \\ &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\delta} \psi_\tau (u_i) + \sum_{i=1}^{n} \int_0^{v_{ni}} \left(\mathbb{1}_{\{u_i \leq s\}} - \mathbb{1}_{\{u_i \leq 0\}} \right) \mathrm{d}s. \end{aligned}$$

Bezeichnet man die erste Summe mit $Z_{1n}(\boldsymbol{\delta})$ und die zweite analog mit $Z_{2n}(\boldsymbol{\delta})$ so ergibt dies

$$Z_n(\boldsymbol{\delta}) = Z_{1n}(\boldsymbol{\delta}) + Z_{2n}(\boldsymbol{\delta})$$

und die beiden Summanden werden im Folgenden getrennt betrachtet.

Im ersten Schritt wird nun die asymptotische Verteilung von $Z_{1n}(\boldsymbol{\delta})$ ermittelt. Dazu berechnet man Erwartungswert und Varianz dieser Summe von unabhängig verteilten Zufallsvariablen

$$Z_{1n}(\boldsymbol{\delta}) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\delta} \psi_\tau (u_i) .$$

Für den Erwartungswert von $-Z_{1n}(\boldsymbol{\delta})$ erhält man mit Hilfe der zweiten Darstellung

$$\begin{aligned} \mathbb{E}\left[-Z_{1n}(\boldsymbol{\delta})\right] &= \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\delta} \psi_\tau (u_i) \right] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\delta} \mathbb{E}\left[\psi_\tau (u_i)\right]. \end{aligned}$$

Setzt man nun $\psi_\tau (u_i) = \tau - \mathbb{1}_{\{u_i < 0\}}$ ein, liefert dies

$$\begin{aligned} \mathbb{E}\left[-Z_{1n}(\boldsymbol{\delta})\right] &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\delta} \left(\tau - \mathbb{E}\left[\mathbb{1}_{\{u_i < 0\}} \right] \right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\delta} \left(\tau - \tau \right) = 0, \end{aligned}$$

denn

$$\mathbb{E}\left[\mathbb{1}_{\{u_i < 0\}} \right] = \mathbb{P}\left[u_i < 0\right] = \mathbb{P}\left[Y_i < \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) | \mathbf{x}_i^\top \right] = \tau,$$

weil $\mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$ das theoretische τ-Quantil der bedingten Verteilung von Y_i ist.

Nun wird die Varianz von $-Z_{1n}(\boldsymbol{\delta})$ näher betrachtet:

$$\begin{aligned}\text{Var}\left[-Z_{1n}(\boldsymbol{\delta})\right] =& \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^\top\boldsymbol{\delta}\psi_\tau\left(u_i\right)\right]\\ =& \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^\top\boldsymbol{\delta}\left(\tau-\mathbb{1}_{\{u_i<0\}}\right)\right].\end{aligned}$$

Da $\mathbf{x}_i^\top\boldsymbol{\delta}\tau$ konstant ist, reduziert sich dies zu

$$\begin{aligned}\text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top\boldsymbol{\delta}\tau-\mathbf{x}_i^\top\boldsymbol{\delta}\mathbb{1}_{\{u_i<0\}}\right)\right] =& \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^\top\boldsymbol{\delta}\mathbb{1}_{\{u_i<0\}}\right]\\ =& \left(\frac{1}{\sqrt{n}}\right)^2\boldsymbol{\delta}^\top\left\{\sum_{i=1}^{n}\mathbf{x}_i\text{Var}\left[\mathbb{1}_{\left\{Y_i<\mathbf{x}_i^\top\boldsymbol{\beta}(\tau)\right\}}\right]\mathbf{x}_i^\top\right\}\boldsymbol{\delta}.\end{aligned}$$

Varianz und Summation durften im letzten Schritt vertauscht werden, da es sich um eine Summe unabhängig verteilter Zufallsvariablen handelt. Aufgrund der Identität

$$\text{Var}\left[\mathbb{1}_{\{A\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{A\}}^2\right] - \mathbb{E}\left[\mathbb{1}_{\{A\}}\right]^2 = \mathbb{P}\left[A\right] - \mathbb{P}\left[A\right]^2 = \mathbb{P}\left[A\right]\left(1-\mathbb{P}\left[A\right]\right)$$

gilt also für die Varianz-Terme in der Summe

$$\text{Var}\left[\mathbb{1}_{\left\{Y_i<\mathbf{x}_i^\top\boldsymbol{\beta}(\tau)\right\}}\right] = \mathbb{P}\left[\mathbb{1}_{\left\{Y_i<\mathbf{x}_i^\top\boldsymbol{\beta}(\tau)\right\}}\right]\left(1-\mathbb{P}\left[\mathbb{1}_{\left\{Y_i<\mathbf{x}_i^\top\boldsymbol{\beta}(\tau)\right\}}\right]\right) = \tau\left(1-\tau\right).$$

Mit dieser Erkenntnis und Bedingung **II**.1. folgt also nun für die Varianz von $-Z_{1n}(\boldsymbol{\delta})$

$$\text{Var}\left[-Z_{1n}(\boldsymbol{\delta})\right] = \frac{1}{n}\boldsymbol{\delta}^\top\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\delta}\tau(1-\tau) \overset{n\to\infty}{\longrightarrow} \tau(1-\tau)\boldsymbol{\delta}^\top\mathbf{D}_0\boldsymbol{\delta}$$

Da es sich bei $Z_{1n}(\boldsymbol{\delta})$ um eine Summe unabhängiger (nicht identisch verteilter!) Zufallsvariablen handelt, erhält man nun durch Anwendung des ZGWS unter der Voraussetzung, dass die Bedingungen **II**.1. und **II**.3. erfüllt sind (Lindeberg-Feller-ZGWS), schließlich folgendes Ergebnis:

$$\frac{-Z_{1n}(\boldsymbol{\delta})-0}{\sqrt{\tau(1-\tau)\boldsymbol{\delta}^\top\mathbf{D}_0\boldsymbol{\delta}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Dies ist aber gleichbedeutend mit

$$Z_{1n}(\boldsymbol{\delta}) \xrightarrow{d} -\boldsymbol{\delta}^\top\mathbf{W}, \tag{4.9}$$

mit $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \tau(1-\tau)\mathbf{D}_0)$.

Im nächsten Schritt betrachtet man das asymptotische Verhalten von $Z_{2n}(\boldsymbol{\delta})$ genauer.

Dazu verwendet man die Darstellung

$$\begin{aligned} Z_{2n}(\boldsymbol{\delta}) =& \sum_{i=1}^{n} \left(Z_{2ni}(\boldsymbol{\delta}) + \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta}) \right] - \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta}) \right] \right) \\ =& \sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta}) \right] + \sum_{i=1}^{n} \left(Z_{2ni}(\boldsymbol{\delta}) - \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta}) \right] \right), \end{aligned} \tag{4.10}$$

mit

$$Z_{2ni}(\boldsymbol{\delta}) = \int_0^{v_{ni}} \left(\mathbb{1}_{\{u_i \leq s\}} - \mathbb{1}_{\{u_i \leq 0\}} \right) \mathrm{d}s.$$

Nun berechnet man zuerst

$$\sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta}) \right] = \sum_{i=1}^{n} \int_0^{v_{ni}} \left(\mathbb{E}\left[\mathbb{1}_{\{u_i \leq s\}} \right] - \mathbb{E}\left[\mathbb{1}_{\{u_i \leq 0\}} \right] \right) \mathrm{d}s.$$

Dies wird mit Hilfe der Definition des Erwartungswertes über Indikatorfunktionen zu

$$\begin{aligned} & \sum_{i=1}^{n} \int_0^{v_{ni}} \left(\mathbb{E}\left[\mathbb{1}_{\{u_i \leq s\}} \right] - \mathbb{E}\left[\mathbb{1}_{\{u_i \leq 0\}} \right] \right) \mathrm{d}s \\ =& \sum_{i=1}^{n} \int_0^{v_{ni}} \left(\mathbb{P}\left[Y_i \leq \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) + s | \mathbf{x}_i \right] - \mathbb{P}\left[Y_i \leq \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) | \mathbf{x}_i \right] \right) \mathrm{d}s \\ =& \sum_{i=1}^{n} \int_0^{v_{ni}} \left[F_i\left(q_i(\tau) + s \right) - F_i\left(q_i(\tau) \right) \right] \mathrm{d}s. \end{aligned}$$

Nun führt man die Substitution

$$\begin{aligned} t = s\sqrt{n} \quad &\Leftrightarrow \quad s = \frac{t}{\sqrt{n}} \\ \frac{\mathrm{d}s}{\mathrm{d}t} = \frac{1}{\sqrt{n}} \quad &\Leftrightarrow \quad \mathrm{d}s = \frac{\sqrt{n}}{n} \mathrm{d}t \\ 0 \leq s \leq \frac{\mathbf{x}_i^\top \boldsymbol{\delta}}{\sqrt{n}} = v_{ni} \quad &\Leftrightarrow \quad 0 \leq t \leq \mathbf{x}_i^\top \boldsymbol{\delta} \end{aligned}$$

durch, was wiederum zur Darstellung

$$\begin{aligned} & \sum_{i=1}^{n} \int_0^{v_{ni}} \left[F_i\left(q_i(\tau) + s \right) - F_i\left(q_i(\tau) \right) \right] \mathrm{d}s \\ =& \frac{1}{n} \sum_{i=1}^{n} \int_0^{\mathbf{x}_i^\top \boldsymbol{\delta}} \sqrt{n} \left[F_i\left(q_i(\tau) + \frac{t}{\sqrt{n}} \right) - F_i\left(q_i(\tau) \right) \right] \mathrm{d}t \end{aligned}$$

führt. Umformen und Erweitern mit $1 = t/t$ liefert

$$\sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})\right] = \frac{1}{n}\sum_{i=1}^{n} \int_0^{\mathbf{x}_i^\top \boldsymbol{\delta}} t \frac{F_i\left(q_i(\tau) + \frac{t}{\sqrt{n}}\right) - F_i\left(q_i(\tau)\right)}{\frac{t}{\sqrt{n}}} \mathrm{d}t.$$

Man erkennt, dass es sich bei dem Integrand um einen Differenzenquotient handelt mit $x_0 = q_i(\tau)$ und $h = \frac{t}{\sqrt{n}}$. Für $h \to 0$ beziehungsweise gleichbedeutend für $n \to \infty$ erhält man dann den Differentialquotienten, was zu

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^{\mathbf{x}_i^\top \boldsymbol{\delta}} t \frac{F_i\left(q_i(\tau) + \frac{t}{\sqrt{n}}\right) - F_i\left(q_i(\tau)\right)}{\frac{t}{\sqrt{n}}} \mathrm{d}t \to \frac{1}{n}\sum_{i=1}^{n} \int_0^{\mathbf{x}_i^\top \boldsymbol{\delta}} \left(t f_i(q_i(\tau))\right) \mathrm{d}t$$

führt. Nun kann die Integration durchgeführt werden und man erhält

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \int_0^{\mathbf{x}_i^\top \boldsymbol{\delta}} t f_i(q_i(\tau)) \mathrm{d}t =& \frac{1}{n}\sum_{i=1}^{n} \frac{t^2}{2} f_i\left(q_i(\tau)\right) \Big|_{t=0}^{\mathbf{x}_i^\top \boldsymbol{\delta}} \\ =& \frac{1}{2}\boldsymbol{\delta}^\top \frac{1}{n}\sum_{i=1}^{n} f_i(q_i(\tau)) \mathbf{x}_i \mathbf{x}_i^\top \boldsymbol{\delta}.\end{aligned}$$

Aufgrund von Bedingung **II**.2. kann man schlussendlich für $n \to \infty$

$$\sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})\right] \longrightarrow \frac{1}{2}\boldsymbol{\delta}^\top \mathbf{D}_1(\tau)\boldsymbol{\delta}$$

folgern.

Im nächsten Schritt betrachtet man nun die Varianz von $Z_{2n}(\boldsymbol{\delta})$ genauer um zu zeigen, dass die zweite Summe in (4.10) gegen Null strebt:

$$\begin{aligned}\mathrm{Var}\left[Z_{2n}(\boldsymbol{\delta})\right] =& \mathrm{Var}\left[\sum_{i=1}^{n} Z_{2ni}(\boldsymbol{\delta})\right] \overset{\text{unabh.}}{=} \sum_{i=1}^{n} \mathrm{Var}\left[Z_{2ni}(\boldsymbol{\delta})\right] \\ =& \sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})^2\right] - \underbrace{\sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})\right]^2}_{\geq 0} \\ \leq& \sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})^2\right].\end{aligned}$$

Die rechte Seite der Ungleichung lässt sich nun wegen $0 \leq s$ für beliebiges i mit Hilfe

von

$$Z_{2ni}(\boldsymbol{\delta}) = \int_0^{v_{ni}} \underbrace{\left(\mathbb{1}_{\{u_i \leq s\}} - \mathbb{1}_{\{u_i \leq 0\}}\right)}_{\leq 1} \mathrm{d}s$$
$$\leq \int_0^{v_{ni}} 1 \mathrm{d}s = v_{ni} \leq \max |v_{ni}| = \frac{1}{\sqrt{n}} \max \left|\mathbf{x}_i^\top \boldsymbol{\delta}\right| \quad \forall i$$

weiter abschätzen und man erhält dadurch

$$\mathrm{Var}\left[Z_{2n}(\boldsymbol{\delta})\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})^2\right] \leq \frac{1}{\sqrt{n}} \max \left|\mathbf{x}_i^\top \boldsymbol{\delta}\right| \sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})\right].$$

Für den ersten Term der rechten Seite gilt aufgrund der Hölder Ungleichung für L_2-Normen

$$\max |\mathbf{x}_i^\top \boldsymbol{\delta}| \leq \max \left\|x_i^\top\right\|_2 \|\boldsymbol{\delta}\|_2$$

die Abschätzung

$$\mathrm{Var}\left[Z_{2n}(\boldsymbol{\delta})\right] \leq \frac{1}{\sqrt{n}} \max \left\|\mathbf{x}_i^\top\right\|_2 \|\boldsymbol{\delta}\|_2 \sum_{i=1}^{n} \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})\right]$$

und da wegen Bedingung **II**.3. $\frac{1}{\sqrt{n}} \max \left\|\mathbf{x}_i^\top\right\|_2 \to 0$ gilt, folgt schlussendlich

$$\mathrm{Var}\left[Z_{2n}(\boldsymbol{\delta})\right] \longrightarrow 0.$$

Dies ist wegen der Definition der Varianz einer Zufallsvariable gleichbedeutend mit

$$\mathbb{E}\left[\sum_{i=1}^{n} \left(Z_{2ni}(\boldsymbol{\delta}) - \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})\right]\right)^2\right] \longrightarrow 0.$$

Daraus kann wiederum gefolgert werden, dass

$$Z_{2ni}(\boldsymbol{\delta}) - \mathbb{E}\left[Z_{2ni}(\boldsymbol{\delta})\right] \xrightarrow{f.s.} 0$$

gilt. Setzt man diese Erkenntnisse nun in (4.10) ein, erhält man letztlich

$$Z_{2n}(\boldsymbol{\delta}) \longrightarrow \frac{1}{2}\boldsymbol{\delta}^\top \mathbf{D}_1(\tau)\boldsymbol{\delta}$$

Schließlich bekommt man durch Zusammensetzen der Ergebnisse:

$$Z_n(\boldsymbol{\delta}) = Z_{1n}(\boldsymbol{\delta}) + Z_{2n}(\boldsymbol{\delta}) \xrightarrow{d} Z_0(\boldsymbol{\delta}) = -\boldsymbol{\delta}^\top \mathbf{W} + \frac{1}{2}\boldsymbol{\delta}^\top \mathbf{D}_1(\tau)\boldsymbol{\delta}.$$

Für die begrenzende Zielfunktion $Z_0(\boldsymbol{\delta})$ gilt:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\delta}} Z_0(\boldsymbol{\delta}) = -\mathbf{W} + \mathbf{D}_1(\tau)\boldsymbol{\delta}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}^2\boldsymbol{\delta}} Z_0(\boldsymbol{\delta}) = \mathbf{D}_1(\tau) \geq 0,$$

und somit handelt es sich aufgrund der Positivität der zweiten Ableitung unabhängig vom Argument $\boldsymbol{\delta}$ erneut um eine konvexe Funktion, die in

$$-\mathbf{W} + \mathbf{D}_1(\tau)\boldsymbol{\delta} \stackrel{!}{=} \mathbf{0}$$
$$\Rightarrow \hat{\boldsymbol{\delta}}_0 = \mathbf{D}_1(\tau)^{-1}\mathbf{W}$$

ihr Minimum annimmt.

Aufgrund des „Basic Corollary“ in Hjort & Pollard (1993) folgt nun aus $Z_n(\boldsymbol{\delta}) \xrightarrow{d} Z_0(\boldsymbol{\delta})$, dass

$$\operatorname{argmin} Z_n(\boldsymbol{\delta}) = \hat{\boldsymbol{\delta}}_n \xrightarrow{d} \operatorname{argmin} Z_0(\boldsymbol{\delta}) = \hat{\boldsymbol{\delta}}_0 \quad \Leftrightarrow \quad \sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) \xrightarrow{d} \mathbf{D}_1^{-1}(\tau)\mathbf{W}$$

gilt, mit $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \tau(1-\tau)\mathbf{D}_0)$. Weiters weiß man, dass $\mathbf{D}_1^{-1}(\tau)$ eine symmetrische Matrix ist.

Zusammengefasst erhält man nun die Aussage des Satzes:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \tau(1-\tau)\mathbf{D}^{-1}(\tau)\mathbf{D}_0\mathbf{D}^{-1}(\tau)\right).$$

□

Die Form der asymptotischen Varianz bezeichnet man auch als *Huber-Sandwich* (Huber, 1967). Im Fall iid-verteilter Fehlerterme „kollapiert“ dieses Sandwich und man erhält einen viel einfacheren Ausdruck für die asymptotische Varianz.

Korollar 4.2.1. *Für den Fall iid-verteilter Fehlerterme im QR-Modell reduziert sich die asymptotische Varianz zu*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \omega^2\mathbf{D}_0^{-1}\right),$$

mit $\omega^2 = \tau(1-\tau)/\left[f(q(\tau))\right]^2$.

Beweis. Aufgrund der Annahme iid-verteilter Fehlerterme folgt, dass auch die Responsevariablen iid-verteilt sind und daher gilt

$$f_i\left(q_i(\tau)\right) = f\left(q(\tau)\right),$$

für alle $i = 1, 2, \ldots$. Damit folgt nun für Bedingung **II**.2

$$\lim_{n\to\infty} \frac{1}{n} \sum f(q_i(\tau)) \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{D}_1(\tau) = f\left(q(\tau)\right) \lim_{n\to\infty} \frac{1}{n} \sum \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{D}_1(\tau).$$

Bedingung **II**.1 liefert dann

$$f\left(q(\tau)\right) \mathbf{D}_0 = \mathbf{D}_1(\tau).$$

Setzt man dies nun in die zuvor erhalten Darstellung der asymptotischen Verteilung ein, erhält man

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \tau(1-\tau) f\left(q(\tau)\right)^{-1} \underbrace{\mathbf{D}_0^{-1}\mathbf{D}_0}_{I} f\left(q(\tau)\right)^{-1} \mathbf{D}_0^{-1}\right)$$

und dies vereinfacht sich zu

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \omega^2 \mathbf{D}_0^{-1}\right),$$

mit $\omega^2 = \tau(1-\tau) f\left(q(\tau)\right)^{-2}$. □

Mit Hilfe dieser Erkenntnisse über die asymptotische Verteilung des QR-Schätzers kann auch eine lineare Darstellung, die sogenannte *Bahadur-Darstellung* angegeben werden. Sie ist eine direkte Folgerung aus den zuvor getätigten Überlegungen.

Korollar 4.2.2 (Bahadur-Darstellung)**.** *Die Bahadur-Darstellung ist eine lineare Repräsentation des Schätzers $\hat{\boldsymbol{\beta}}(\tau)$ und hat die Form*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) = \mathbf{D}_1^{-1}(\tau) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i \psi_\tau\left(Y_i - q_i(\tau)\right) + R_n,$$

mit $R_n = o_{\mathbb{P}}(1)$.

Beweis. Wie bereits zuvor gezeigt wurde gilt

$$\hat{\boldsymbol{\delta}}_n - \hat{\boldsymbol{\delta}}_0 = \hat{\boldsymbol{\delta}}_n - \mathbf{D}_1^{-1}(\tau)\mathbf{W} \xrightarrow{d} \mathbf{0},$$

die Differenz strebt also in Verteilung gegen eine Konstante und damit folgt aus der Verteilungskonvergenz die Konvergenz in Wahrscheinlichkeit, das heißt

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\hat{\boldsymbol{\delta}}_n - \hat{\boldsymbol{\delta}}_0\right| > \varepsilon\right] = 0,$$

für alle $\varepsilon > 0$. Aufgrund dessen gilt nun

$$\hat{\boldsymbol{\delta}}_n - \hat{\boldsymbol{\delta}}_0 = o_{\mathbb{P}}(1) \quad \Leftrightarrow \quad \hat{\boldsymbol{\delta}}_n = \mathbf{D}_1^{-1}(\tau)\mathbf{W} + R_n', \tag{4.11}$$

mit $R_n^{'} = o_{\mathbb{P}}(1)$. Weiters wurde bereits in (4.9) bewiesen, dass

$$Z_{1n}(\boldsymbol{\delta}) - (-\boldsymbol{\delta}^\top)\mathbf{W} = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^\top\boldsymbol{\delta}\psi_\tau\left(u_i\right) - (-\boldsymbol{\delta}^\top)\mathbf{W} \xrightarrow{d} 0$$

gilt, was erneut gleichbedeutend ist mit

$$Z_{1n}(\boldsymbol{\delta}) = -\boldsymbol{\delta}^\top\mathbf{W} + o_{\mathbb{P}}(1).$$

Einsetzen der Darstellung von $Z_{1n}(\boldsymbol{\delta})$ liefert dann

$$-\boldsymbol{\delta}^\top\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\psi_\tau\left(u_i\right) = -\boldsymbol{\delta}^\top\mathbf{W} + o_{\mathbb{P}}(1),$$

das heißt

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\psi_\tau\left(u_i\right) \xrightarrow{d} \mathbf{W}$$

Setzt man dies nun in (4.11) für $\mathbf{W}$ ein, so erhält man

$$\hat{\boldsymbol{\delta}}_n = \sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) = \mathbf{D}_1^{-1}(\tau)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\psi_\tau\left(u_i\right) + o_{\mathbb{P}}(1)\right) + R_n^{'}.$$

Zusammenfassen der $o_{\mathbb{P}}(1)$-Terme liefert dann

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) = \mathbf{D}_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\psi_\tau(Y_i - q_i(\tau)) + R_n.$$

□

Der Vorteil dieser Form ist, dass ein eher komplizierter, nichtlinearer Schätzer durch eine normalisierte Summe iid-verteilter Zufallsvariablen dargestellt werden kann.

4.2.3. Asymptotik bei nichtlinearen QR-Modellen

In manchen Fällen ist es auch von Interesse Modelle zu betrachten, die nicht linear in den Parametern sind, das heißt

$$Q_{Y_i}\left(\tau|\mathbf{x}_i\right) = g\left(\mathbf{x}_i, \boldsymbol{\beta}(\tau)\right).$$

Der nichtlineare Schätzer ist wie folgt definiert:

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\beta\in\mathcal{B}}{\operatorname{argmin}}\sum\rho_\tau\left(Y_i - g(\mathbf{x}_i, \boldsymbol{\beta})\right),$$

mit der kompakten Menge $\mathcal{B} \subset \mathbb{R}^p$. Die folgenden Bedingungen und Eigenschaften werden an dieser Stelle jedoch nicht näher motiviert oder hergeleitet. Detailierte Betrachtungen finden sich in Barnett et al. (1991) beziehungsweise in Koenker (2005).

Vom linearen Modell wird Bedingung **I** für die Verteilungsfunktionen $\{F_i\}$ mit theoretischen τ-Quantilen $q_i(\tau) = g(\mathbf{x}_i, \boldsymbol{\beta}(\tau))$ und deren Dichten $\{f_i(q)\}$ übernommen. An die *quantile Responsefunktion* $g(\cdot, \cdot)$ müssen jedoch noch weitere Anforderungen gestellt werden. Also trifft man folgende Annahmen:

Bedingung II*: Es existieren positiv definite Matrizen $\mathbf{D}_0$ und $\mathbf{D}_1(\tau)$, sodass mit

$$g_i' = \frac{\partial g(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}(\tau)}$$

gilt:

1. $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n g_i' g_i'^\top = \mathbf{D}_0$
2. $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n f(q_i(\tau)) g_i' g_i'^\top = \mathbf{D}_1(\tau)$
3. $\max_{i=1,\ldots,n} \|g_i'\|_2 / \sqrt{n} \longrightarrow 0.$

Bedingung III: Es existieren Konstanten κ_0, κ_1 und n_0, sodass für $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}$ und $n > n_0$

$$\kappa_0 \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 \leq \left(\frac{1}{n} \sum_{i=1}^n \left(g(\mathbf{x}_i, \boldsymbol{\beta}_1) - g(\mathbf{x}_i, \boldsymbol{\beta}_2) \right)^2 \right)^{1/2} \leq \kappa_1 \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2$$

gilt.

Bedingung **III** garantiert, dass die Zielfunktion ein eindeutiges Minimum in $\boldsymbol{\beta}(\tau)$ hat und genügend Glätte aufweist. Unter diesen Voraussetzungen kann man auch im nichtlinearen Fall eine Bahadur-Darstellung angeben:

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau) \right) = \mathbf{D}_1^{-1}(\tau) \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i' \psi_\tau \left(Y_i - g(\mathbf{x}_i, \boldsymbol{\beta}(\tau)) \right) + o_{\mathbb{P}}(\tau).$$

Weiters lässt dies auch auf die asymptotische Verteilung schließen, welche von der Form

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau) \right) \xrightarrow{d} \mathcal{N} \left(0, \tau(1-\tau) \mathbf{D}_1^{-1}(\tau) \mathbf{D}_0 \mathbf{D}_1^{-1}(\tau) \right)$$

ist.

Mit Hilfe dieser Ergebnisse kann man nun ähnlich zu den gewöhnlichen Erwartungswertmodellen verschiedenen Tests zur Überprüfung der Signifikanz von Parametern im Modell angeben.

Aufgrund der Tatsache, dass der QR-Schätzer asymptotisch normalverteilt ist, liefert Division durch die geschätzte Standardabweichung, also Standardisieren des Schätzers,

unter der Nullhypothese

$$H_0 : \beta_j(\tau) = 0$$

eine *Student-t*-verteilte Teststatistik für $j = 1, \ldots, p-1$. Dadurch kann analog zu den Erwartungswertmodellen für einzelne Parameter des QR-Modells deren Signifikanz überprüft werden kann. So einen Test nennt man auch *asymptotischen Wald Test*.

Koenker (2005) und Davino et al. (2013) haben sich noch ausführlich mit weiteren möglichen Tests auseinandergesetzt. Für mehr Information zu diesem Thema wird an dieser Stelle also auf deren Arbeiten verwiesen.

4.3. Konsistenz

Durch die Betrachtung der asymptotischen Verteilungen wurde bereits die Frage nach der Konvergenzrate geklärt. Dabei wurde aber insgeheim vorausgesetzt, dass der Schätzer *konsistent* ist. Grob gesprochen ist ein Schätzer konsistent, falls eine Vergrößerung des Stichprobenumfangs zu einer Annäherung des Schätzers an den wahren Parameter führt.

Man nimmt an, dass die Form der bedingten Quantilsfunktion der Response Y gegeben $\mathbf{x}$ zum Niveau τ die parametrische Form

$$Q_Y(\tau|\mathbf{x}) = g(\mathbf{x}, \boldsymbol{\beta}(\tau))$$

hat. Die Frage, die sich nun stellt ist: Welche Bedingungen müssen erfüllt sein, dass der Schätzer

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau (Y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))$$

in Wahrscheinlichkeit gegen den wahren Parameter $\boldsymbol{\beta}(\tau)$ konvergiert, das heißt

$$\left\| \hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau) \right\| \xrightarrow{P} 0,$$

für $n \to \infty$?

4.3.1. Konsistenz des Stichprobenquantils

Der einfachste Fall ist die Betrachtung der Konsistenz des Stichprobenquantils zum Niveau τ. Basierend auf einer Stichprobe $(Y_1, \ldots, Y_n)$ aus der Verteilung $F(\cdot)$ erhält man das geschätzte Stichprobenquantil durch

$$\hat{q}(\tau) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau (Y_i - q).$$

Nimmt man nun an, dass die Verteilung F ein eindeutiges (theoretisches) τ-Quantil $q_\tau = F^{-1}(\tau)$ hat, dann gilt

$$\hat{q}_\tau \xrightarrow{P} q_\tau.$$

Dies kann direkt durch den *Satz von Glivenko-Cantelli* gefolgert werden, der besagt, dass die empirische Verteilungsfunktion für *iid*-verteilte Zufallsvariablen gleichmäßig gegen die theoretische Verteilungsfunktion strebt und dies impliziert auch die Konvergenz der empirischen gegen die theoretischen Quantile.

Mizera & Wellner (1998) leiten in ihrer Arbeit für unabhängige, aber nicht identisch-verteilte Zufallsvariablen $\{Y_i\}_{i=1}^n$ mit Verteilungsfunktionen $\{F_i\}_{i=1}^n$, die alle gerade **ein gemeinsames** τ-Quantil $F_i^{-1}(\tau) = q_\tau$ für $i = 1, \ldots, n$ haben, *notwendige und hinreichende Bedingungen* für die Konsistenz des τ-Stichprobenquantils her.

Die empirische Verteilungsfunktion der Y_i und die gemittelte Verteilungsfunktion werden dabei wie folgt festgelegt:

$$\mathbf{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}},$$
$$\overline{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} F_i(y).$$

Es ergeben sich dann die Zusammenhänge

$$\mathbb{E}\left[\mathbf{F}_n(y)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\{Y_i \leq y\}}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left[Y_i \leq y\right] = \frac{1}{n}\sum_{i=1}^{n} F_i(y) = \overline{F}_n(y)$$

und

$$q_\tau = \overline{F}_n^{-1}(\tau), \text{ denn } \overline{F}_n(q_\tau) = \frac{1}{n}\sum_{i=1}^{n} F_i(q_\tau) = \frac{n}{n}\tau = \tau$$

da gerade $F_i(q_\tau) = \tau$ für alle $i = 1, \ldots, n$ gilt.

Weiters benötigt man dazu die Definition

$$a_n(\varepsilon) := \mathbb{E}\left[\mathbf{F}_n\left(\overline{F}_n^{-1}(\tau) + \varepsilon\right)\right] = \overline{F}_n\left(q_\tau + \varepsilon\right)$$
$$b_n(\varepsilon) := \mathbb{E}\left[\mathbf{F}_n\left(\overline{F}_n^{-1}(\tau) - \varepsilon\right)\right] = \overline{F}_n\left(q_\tau - \varepsilon\right),$$

wobei die jeweilig letzte Gleichheit wegen den zuvor angeführten Eigenschaften gilt. Zusätzlich liefert dies

$$b_n(\varepsilon) \leq \tau \leq a_n(\varepsilon),$$

für alle $\varepsilon > 0$.

Mizera & Wellner (1998) haben schließlich gezeigt, dass das Stichprobenquantil genau dann ein konsistenter Schätzer ist, wenn

$$\sqrt{n}\,(a_n(\varepsilon) - \tau) \longrightarrow \infty \quad \text{und} \quad \sqrt{n}\,(\tau - b_n(\varepsilon)) \longrightarrow \infty.$$

4.3.2. Konsistenz von QR-Schätzern

Bei Regressionsmodellen führt das Vorhandensein der Kovariablen zu zusätzlichen Komplikationen im Vergleich zu den Überlegungen für das einfache Stichprobenquantil. Für die Quantilsfunktion der Response Y gegeben $\mathbf{x}$ nimmt man an, dass sie von der Form

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}(\tau)$$

ist und die bedingten Verteilungsfunktionen $F_i(\cdot)$ von Y_i sollen die Bedingungen

$$\sqrt{n}\,(a_n(\varepsilon) - \tau) \longrightarrow \infty \quad \text{und} \quad \sqrt{n}\,(\tau - b_n(\varepsilon)) \longrightarrow \infty$$

für $i = 1, \ldots, n$ erfüllen, wobei hier entsprechend

$$a_n(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} F_i\left(\mathbf{x}_i^\top \boldsymbol{\beta}(\tau) + \varepsilon\right)$$
$$b_n(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} F_i\left(\mathbf{x}_i^\top \boldsymbol{\beta}(\tau) - \varepsilon\right)$$

gewählt wird.

Bantli & Hallin (1999) haben gezeigt, dass die gestellten Bedingungen an die Verteilungsfunktion hinreichend und notwendig sind, sodass der QR-Schätzer konsistent ist, das heißt es gilt dann $\hat{\boldsymbol{\beta}}(\tau) \xrightarrow{P} \boldsymbol{\beta}(\tau)$, wobei man noch folgende Annahmen bezüglich der Prädiktorvariablen $\mathbf{x}_i$ treffen muss:

Bedingung X1: Es existiert ein $d > 0$, sodass

$$\liminf_{n\to\infty} \inf_{\|\mathbf{u}\|=1} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{|\mathbf{x}_i^\top \mathbf{u}| < d\}} = 0.$$

Bedingung X2: Es existiert ein $D > 0$, sodass

$$\limsup_{n\to\infty} \sup_{\|\mathbf{u}\|=1} \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{x}_i^\top \mathbf{u}\right)^2 \leq D.$$

Die erste Bedingung **X1** garantiert, dass die $\{\mathbf{x}_i\}$ nicht auf einem Unterraum von $\mathbb{R}^p$ konzentriert sind und wird zur Identifizierbarkeit verwendet. **X2** kontrolliert die Wachs-

tumsrate der Prädiktoren und ist unter der Bedingung

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{D}_0,$$

die man bereits bei der Analyse der Konvergenzrate kennengelernt hat, immer erfüllt.

4.4. Sparsity-Schätzung

Ein weitere Aufgabe, die sich durch Betrachtung der asymptotischen Verteilung der QR-Koeffizienten ergibt, ist die Notwendigkeit einer Schätzung der Kovarianzmatrix.

4.4.1. Skalare Sparsity-Schätzung

Im Fall iid-verteilter Fehlerterme im QR-Modell muss aufgrund von

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right) \xrightarrow{d} \mathcal{N}\left(0, \omega^2 \mathbf{D}_0^{-1}\right)$$

mit $\omega^2 = \tau(1-\tau)/f\left(q(\tau)\right)^2$ die Funktion

$$s(\tau) = \frac{1}{f\left(q_\tau\right)} = \frac{1}{f(F^{-1}(\tau))}$$

geschätzt werden, die bereits in Kapitel 3.3 auf Seite 39 behandelt wurde und dort die Bezeichnung *Sparsity Funktion* erhalten hat. Wie ebenfalls in Kapitel 3.3 gezeigt wurde, gilt

$$s(\tau) = \frac{\mathrm{d}F^{-1}(\tau)}{\mathrm{d}\tau}$$

und damit erscheint die Verwendung von Schätzern, die auf Differenzenquotienten basieren, als sinnvoll. Diese sind von der Form

$$\hat{s}_n(\tau) = \frac{\hat{F}_n^{-1}(\tau + h_n) - \hat{F}_n^{-1}(\tau - h_n)}{2h_n},$$

wobei $\hat{F}_n^{-1}(\cdot)$ die empirische Quantilsfunktion bezüglich $F^{-1}(\cdot)$ ist. Weiters bezeichnet man h_n als *Bandbreite*, die für $n \to \infty$ gegen Null streben soll. Aufgrund der Überlegungen von Bofinger (1975) wäre eine mögliche Wahl für die Bandbreite h_n

$$h_n = n^{-1/5}\left(4.5\left(\frac{s(\tau)}{s''(\tau)}\right)^2\right)^{1/5}. \tag{4.12}$$

Für $s''(\tau)$ gilt die Darstellung

$$s''(\tau) = \frac{-f''\left(F^{-1}(\tau)\right) f\left(F^{-1}(\tau)\right) + 3\left(f'\left(F^{-1}(\tau)\right)\right)^2}{f^5\left(F^{-1}(\tau)\right)}$$

und dies liefert dann

$$\begin{aligned}
\frac{s(\tau)}{s''(\tau)} =& \frac{\frac{1}{f(F^{-1}(\tau))}}{\frac{-f''(F^{-1}(\tau))f(F^{-1}(\tau))+3\left(f'(F^{-1}(\tau))\right)^2}{f^5(F^{-1}(\tau))}} \\
=& \frac{\left(f\left(F^{-1}(\tau)\right)\right)^4}{-f''\left(F^{-1}(\tau)\right) f\left(F^{-1}(\tau)\right) + 2\left(f'\left(F^{-1}(\tau)\right)\right)^2 + \left(f'\left(F^{-1}(\tau)\right)\right)^2} \\
=& \frac{\left(f\left(F^{-1}(\tau)\right)\right)^2}{\frac{2\left(f'(F^{-1}(\tau))\right)^2}{\left(f(F^{-1}(\tau))\right)^2} + \frac{\left(f'(F^{-1}(\tau))\right)^2}{\left(f(F^{-1}(\tau))\right)^2} - \frac{f''(F^{-1}(\tau))f(F^{-1}(\tau))}{\left(f(F^{-1}(\tau))\right)^2}} \\
=& \frac{\left(f\left(F^{-1}(\tau)\right)\right)^2}{2\left(\frac{f'(F^{-1}(\tau))}{f(F^{-1}(\tau))}\right)^2 + \left[\left(\frac{f'(F^{-1}(\tau))}{f(F^{-1}(\tau))}\right)^2 - \frac{f''(F^{-1}(\tau))}{f(F^{-1}(\tau))}\right]}.
\end{aligned}$$

Die Berechnung von h_n ist aber nur durch Kenntnis von $s(\tau)$ möglich, was wiederum aber bedeutet, dass man h_n gar nicht braucht, da kein Schätzer für $s(\tau)$ berechnet werden muss. Glücklicherweiße ist $s(\tau)/s''(\tau)$ nicht sehr sensibel gegenüber der Wahl von $F(\cdot)$ und so verliert man wenig, wenn man für $F(\cdot)$ beispielsweise die Standardnormalverteilung annimmt.

In diesem Fall ist $F(\cdot) = \Phi(\cdot)$ und es gilt

$$\begin{aligned}
\frac{f'\left(F^{-1}(\tau)\right)}{f\left(F^{-1}(\tau)\right)} &= \frac{-F^{-1}(\tau) f\left(F^{-1}(\tau)\right)}{f\left(F^{-1}(\tau)\right)} = -\Phi^{-1}(\tau), \\
\frac{f''\left(F^{-1}(\tau)\right)}{f\left(F^{-1}(\tau)\right)} &= \frac{\left(\left(F^{-1}(\tau)\right)^2 - 1\right) f\left(F^{-1}(\tau)\right)}{f\left(F^{-1}(\tau)\right)} = \left(\Phi^{-1}(\tau)\right)^2 - 1,
\end{aligned}$$

was

$$\left(\frac{f'\left(F^{-1}(\tau)\right)}{f\left(F^{-1}(\tau)\right)}\right)^2 - \frac{f''\left(F^{-1}(\tau)\right)}{f\left(F^{-1}(\tau)\right)} = 1$$

liefert.

Setzt man dies nun in (4.12) ein, so erhält man schließlich die optimale Bandbreite

$$h_n = n^{-1/5} \left(\frac{4.5\phi^4\left(\Phi^{-1}(\tau)\right)}{\left(2\left(\Phi^{-1}(\tau)\right)^2 + 1\right)^2} \right)^{1/5}.$$

Eine weitere Möglichkeit zur Schätzung der Bandbreite wurde von Hall & Sheater (1988) angegeben und ist von der Form:

$$h_n = n^{-1/2} z_{1-\alpha/2}^{2/3} \left(1.5 \frac{s(\tau)}{s''(\tau)} \right)^{1/3},$$

mit $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$ und dem gewünschtem Testlevel α.

Nach der Wahl der Bandbreite stellt sich auch die Frage nach der Form der empirischen Quantilsfunktion $\hat{F}_n^{-1}(\tau)$. Der einfachste Ansatz ist die Verwendung der Residuen der QR-Anpassung:

$$r_i = Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau),$$

für $i = 1, \ldots, n$. Die empirische Quantilsfunktion dieser Stichprobe von Residuen ist

$$\hat{F}_n^{-1}(\tau) = r_{(i)},$$

für $\tau \in [(i-1)/n, i/n)$, wobei $r_{(i)}$ die i-te Ordnungsstatistik der empirischen Residuen bezeichnet.

Ein Nachteil dieser Wahl der empirischen Verteilungsfunktion ist, dass die Bandbreite groß genug gewählt werden muss, um die p Residuen mit Wert Null zu vermeiden, die sich unweigerlich durch die QR-Schätzung ergeben. Dies ist nur ein Problem, wenn p/n im Vergleich zu h_n relativ groß ist. Eine Möglichkeit ist analog zur „Degrees-of-Freedom“ Korrektur bei der Schätzung von σ^2, das heißt man vernachlässigt die p Stück Residuen mit Wert Null und verwendet als effektive Stichprobengröße zur Sparsity-Schätzung $(n - p)$.

Eine weitere Möglichkeit für $\hat{F}_n^{-1}(\tau)$ bei kleinem Stichprobenumfang ist die Wahl von

$$\hat{F}_n^{-1}(\tau) = \overline{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau),$$

mit $\overline{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$.

4.4.2. Schätzung der Kovarianzmatrix

Im Fall von nicht zwingend iid-verteilten Fehlertermen muss

$$\mathbf{D}_1(\tau) = \lim_{n\to\infty} \sum_{i=1}^n f_i\left(q_i(\tau)\right) \mathbf{x}_i \mathbf{x}_i^\top$$

geschätzt werden, da die asymptotische Kovarianzmatrix von $\sqrt{n}\hat{\boldsymbol{\beta}}(\tau)$ die Form

$$\tau(1-\tau)\mathbf{D}_1^{-1}(\tau)\mathbf{D}_0\mathbf{D}_1^{-1}(\tau),$$

hat, was einem Huber-Sandwich entspricht. Die Matrix $\mathbf{D}_0$ kann direkt berechnet werden, da es dort keinen Nuisanceparameter gibt.

Im Folgenden werden zwei Möglichkeiten für eine Schätzung von $\mathbf{D}_1(\tau)$ angegeben.

Die erste, das *Hendricks-Koenker Sandwich*, ist eine Verallgemeinerung der zuvor gezeigten skalaren Sparsity-Schätzung. Hier nimmt man wieder an, dass es sich um ein lineares Modell handelt. Für $h_n \to 0$ werden die theoretischen Parameter der Quantilsfunktionen zum Niveau $\tau \pm h_n$ konsistent durch $\hat{\boldsymbol{\beta}}(\tau \pm h_n)$ geschätzt. Dies liefert den Schätzer

$$\hat{s}_n(\tau) = \frac{\hat{F}_n^{-1}(\tau + h_n) - \hat{F}_n^{-1}(\tau - h_n)}{2h_n} = \frac{\mathbf{x}_i^\top \left(\hat{\boldsymbol{\beta}}(\tau + h_n) - \hat{\boldsymbol{\beta}}(\tau - h_n)\right)}{2h_n}.$$

Da $s(\tau) = \left(f\left(F^{-1}(\tau)\right)\right)^{-1}$ gilt, erhält man als Schätzer für $f_i(.)$ an der Stelle $q_i(\tau) = Q_{Y_i}(\tau|\mathbf{x}_i)$

$$\hat{f}_i(q_i(\tau)) = \frac{2h_n}{d_i},$$

mit $d_i = \mathbf{x}_i^\top \left(\hat{\boldsymbol{\beta}}(\tau + h_n) - \hat{\boldsymbol{\beta}}(\tau - h_n)\right)$.

Ein Nachteil dieses Schätzers ist die Tatsache, dass er nicht zwingend für alle Beobachtungen in der Stichprobe positiv sein muss. Dies sollte aber der Fall sein, da eine positive Dichte $f_i\left(q_i(\tau)\right)$ angenommen wurde. Das Problem tritt nur dann auf, wenn Quantile Crossing vorliegt, was in der Praxis nicht sehr häufig der Fall ist. Bei der Implementation dieses Schätzers wurde $\hat{f}_i(q_i(\tau))$ einfach durch den positiven Teil ersetzt, das heißt

$$\hat{f}_i^+(q_i(\tau)) = \max\left\{0, \frac{2h_n}{d_i - \varepsilon}\right\},$$

wobei ein Tolleranzparameter $\varepsilon > 0$ hinzugefügt wurde um Division durch Null zu vermeiden.

Es gibt jedoch noch einen einfacheren Weg, das QR-Sandwich zu schätzen: Durch die Form von $\mathbf{D}_1(\tau)$ erkennt man, dass man auf der Suche nach einem Matrix-gewichteten Dichteschätzer ist. Zu diesem Zweck führte Barnett et al. (1991) den Kern-Schätzer

$$\hat{\mathbf{D}}_1(\tau) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{r_i}{h_n}\right) \mathbf{x}_i \mathbf{x}_i^\top$$

ein, mit $r_i = Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$ und einer Bandbreite h_n, die die Bedingungen $h_n \to 0$ und $\sqrt{n}h_n \to \infty$ erfüllt. Barnett et al. (1991) verwendet

$$\hat{\mathbf{D}}_1(\tau) = \frac{1}{2nc_n} \sum_{i=1}^{n} \mathbb{1}_{\{|r_i| < c_n\}} \mathbf{x}_i \mathbf{x}_i^\top,$$

mit $c_n \to 0$ und $\sqrt{n}c_n \to \infty$. In der R-Implementation dieses Ansatzes verwendet Koenker (2005)

$$c_n = \kappa \left(\Phi^{-1}(\tau + h_n) - \Phi^{-1}(\tau - h_n) \right),$$

wobei κ ein robuster Scale-Schätzer ist. Weitere Details zu dieser Form der Schätzung der Kovarianzmatrix findet sich auch in Koenker (2005).

4.5. Modellwahl

Es stellt sich nun die Frage nach einer Auswahlstrategie, um aus Modellen mit unterschiedlicher Komplexität, das optimale auszuwählen. Dazu hebt Machado (1993) die Notwendigkeit von Location-Scale invarianten Prozeduren hervor, was für $\tau = 0.5$ (Median-Regression) zu

$$\mathrm{SIC}(j) = n \log\left(\hat{\sigma}_j^2\right) + \frac{1}{2} p_j \log(n),$$

mit $\hat{\sigma}_j^2 = n^{-1} \sum_{i=1}^{n} \rho_{0.5}\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(0.5)\right)$. Der Parameter p_j bezeichnet die Dimension des j-ten Modell. Gewählt wird dann jenes Modell bezüglich dem der Wert $\mathrm{SIC}(j)$ minimiert wird.

Alternativ kann auch das *Akaike-Informations-Kriterium* verwendet werden:

$$\mathrm{AIC}(j) = n \log\left(\hat{\sigma}_j^2\right) + p_j.$$

Modellwahl basierend auf der Minimierung von $\mathrm{AIC}(j)$ führt mit positiver Wahrscheinlichkeit zu einer Überschätzung der Modelldimension. Diese beiden Kriterien liefern eine konsistente Modellauswahlstrategie unter der Annahme, dass eines der p_j-dimensionalen Modell korrekt ist.

4.5.1. Strafmethoden

Eine weitere Möglichkeit zur Modellwahl ist der Einsatz sogenannter *Strafmethoden*. Man betrachtet dazu den Schätzer

$$\hat{\boldsymbol{\beta}}(\tau, \lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum \rho_\tau\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) + \lambda \left\| \boldsymbol{\beta} - \boldsymbol{\beta}_0 \right\|_1,$$

wobei hier der *Lasso*-Strafterm von Tibshirani (1996) verwendet wird, das heißt die L_1-Norm $\|\mathbf{v}\|_1 = \sum_j |v_j|$. Ein λ zwischen $(0, \infty)$ balanziert zwei Ideen: Zum einen soll ein lineares Modell von $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ gegeben $\mathbf{X} \in \mathbb{R}^{n \times p}$ geschätzt werden, zum anderen die Anzahl der Parameter möglichst verringert werden, das heißt einige der β_j werden durch

die Verwendung von Lasso-Strafmethoden gleich Null gesetzt, das heißt es gilt dann $\boldsymbol{\beta}_0 = \mathbf{0}$, und somit findet tatsächlich eine Variablenselektion statt.

Für $\tau = 0.5$ ist die Berechnung dieses Schätzers kein Problem, da es sich bei beiden Termen um Summen von symmtrisch gewichteten Betragsfunktionen handelt. Für $\tau \neq 0.5$ fordert man jedoch asymmetrische Gewichtung der Residuen und symmetrische des Strafterms. Doch das Problem behält die Struktur eines linearen Programms und kann somit einfach angepasst werden. Im allgemeinen können Lasso-Schätzer nur approximativ berechnet werden und deshalb wird im Folgenden näher auf ihre asymptotische Verteilung eingegangen.

Knight & Fu (2000) haben sich eingehend mit der Asymptotik von Lasso-basierten Schätzern befasst. Um die asymptotische Verteilung von $\hat{\boldsymbol{\beta}}(\tau, \lambda)$ zu ermitteln, verwenden sie ähnliche Methoden wie bei der Herleitung der Asymptotik von unpenalisierten Schätzern (Satz 4.2.1) und betrachten dazu die Zielfunktion

$$Z'_n(\boldsymbol{\delta}) = \sum_{i=1}^{n} \left(\rho_\tau \left(u_i - \frac{\mathbf{x}_i^\top \boldsymbol{\delta}}{\sqrt{n}} \right) - \rho_\tau (u_i) \right) + \lambda_n \sum_{j=1}^{p} \left(\left| \beta_j + \frac{\delta_j}{\sqrt{n}} \right| - |\beta_j| \right).$$

Die Parameter β_j und δ_j sind die jeweilig j-ten Komponenten der Vektoren $\boldsymbol{\beta}$ und $\boldsymbol{\delta}$. Weiters wurde $\beta_0 = \mathbf{0}$ gesetzt und es gilt $u_i = Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$. Die Funktion $Z'_n(\boldsymbol{\delta})$ wird in $\hat{\boldsymbol{\delta}} = \sqrt{n} \left(\hat{\boldsymbol{\beta}}(\tau, \lambda_n) - \boldsymbol{\beta}(\tau) \right)$ minimiert und wie bereits im Beweis zu Satz 4.2.1 gezeigt wurde, gilt auch hier

$$Z'_{1n}(\boldsymbol{\delta}) := \sum_{i=1}^{n} \left(\rho_\tau \left(u_i - \frac{\mathbf{x}_i^\top \boldsymbol{\delta}}{\sqrt{n}} \right) - \rho_\tau (u_i) \right) \xrightarrow{d} -\boldsymbol{\delta}^\top \mathbf{W} + \frac{1}{2} \boldsymbol{\delta}^\top \mathbf{D}_1(\tau) \boldsymbol{\delta},$$

mit $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \tau(1-\tau)\mathbf{D}_0)$. Nun soll weiters $\frac{\lambda_n}{\sqrt{n}} \to \lambda_0$ gelten. Der Strafterm lässt sich umformen zu

$$\begin{aligned} Z'_{2n}(\boldsymbol{\delta}) =& \lambda_n \sum_{j=1}^{p} \left(\left| \beta_j + \frac{\delta_j}{\sqrt{n}} \right| - |\beta_j| \right) \\ =& \lambda_n \sum_{j=1}^{p} \left(\left| \beta_j + \frac{\delta_j}{\sqrt{n}} \right| - |\beta_j| \right) \mathbb{1}_{\{\beta_j \neq 0\}} + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} |\delta_j| \, \mathbb{1}_{\{\beta_j = 0\}}. \end{aligned}$$

Für die zweite Summe erhält man wegen $\lambda_n / \sqrt{n} \to \lambda_0$ als Grenzwert

$$\frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} |\delta_j| \, \mathbb{1}_{\{\beta_j = 0\}} \longrightarrow \lambda_0 \sum_{j=1}^{p} |\delta_j| \, \mathbb{1}_{\{\beta_j = 0\}}.$$

Die erste Summe wird wie folgt zerlegt:

$$\begin{aligned}\lambda_n \sum_{j=1}^{p} \left(\left| \beta_j + \frac{\delta_j}{\sqrt{n}} \right| - |\beta_j| \right) \mathbb{1}_{\{\beta_j \neq 0\}} =& \lambda_n \sum_{j=1}^{p} \left(\beta_j + \frac{\delta_j}{\sqrt{n}} \right) \mathbb{1}_{\left\{\beta_j \geq -\delta_j/\sqrt{n}\right\}} \mathbb{1}_{\{\beta_j \neq 0\}} \\ &+ \lambda_n \sum_{j=1}^{p} - \left(\beta_j + \frac{\delta_j}{\sqrt{n}} \right) \mathbb{1}_{\left\{\beta_j \leq -\delta_j/\sqrt{n}\right\}} \mathbb{1}_{\{\beta_j \neq 0\}} \\ &- \lambda_n \left[\sum_{j=1}^{p} \beta_j \mathbb{1}_{\{\beta_j > 0\}} + \sum_{j=1}^{p} (-\beta_j) \mathbb{1}_{\{\beta_j < 0\}} \right] \mathbb{1}_{\{\beta_j \neq 0\}} \\ =& : Z'_{2n,1}(\boldsymbol{\delta}).\end{aligned}$$

Herausheben liefert weiters

$$\begin{aligned}Z'_{2n,1}(\boldsymbol{\delta}) =& \lambda_n \sum_{j=1}^{p} \left[\beta_j \left(\mathbb{1}_{\left\{\beta_j \geq -\delta_j/\sqrt{n}\right\}} - \mathbb{1}_{\{\beta_j > 0\}} \right) - \beta_j \left(\mathbb{1}_{\left\{\beta_j \leq -\delta_j/\sqrt{n}\right\}} - \mathbb{1}_{\{\beta_j < 0\}} \right) \right] \mathbb{1}_{\{\beta_j \neq 0\}} \\ &+ \lambda_n \sum_{j=1}^{p} \left[\frac{\delta_j}{\sqrt{n}} \left(\mathbb{1}_{\left\{\beta_j \geq -\delta_j/\sqrt{n}\right\}} - \mathbb{1}_{\left\{\beta_j \leq -\delta_j/\sqrt{n}\right\}} \right) \right] \mathbb{1}_{\{\beta_j \neq 0\}}.\end{aligned}$$

Die erste Summe in $Z'_{2n,1}(\boldsymbol{\delta})$ strebt für $n \to \infty$ gegen Null und für den zweiten Term bekommt man

$$Z'_{2n,1}(\boldsymbol{\delta}) \overset{n \longrightarrow \infty}{\longrightarrow} \lambda_0 \sum_{j=1}^{n} \left[\delta_j \left(\mathbb{1}_{\{\beta_j > 0\}} - \mathbb{1}_{\{\beta_j < 0\}} \right) \right] \mathbb{1}_{\{\beta_j \neq 0\}} = \lambda_0 \sum_{j=1}^{n} \operatorname{sgn}(\beta_j) \mathbb{1}_{\{\beta_j \neq 0\}}.$$

Insgesamt gilt also:

$$Z'_{2n}(\boldsymbol{\delta}) \to \lambda_0 \left[\sum_{j=1}^{n} \operatorname{sgn}(\beta_j) \mathbb{1}_{\{\beta_j \neq 0\}} + |\delta_j| \mathbb{1}_{\{\beta_j = 0\}} \right].$$

Das heißt die Zielfunktion $Z'_n(\boldsymbol{\delta})$ strebt dann zusammengefasst in Verteilung gegen

$$Z'_n(\boldsymbol{\delta}) \xrightarrow{d} Z_0(\boldsymbol{\delta}) = -\boldsymbol{\delta}^\top \mathbf{W} + \frac{1}{2} \boldsymbol{\delta}^\top \mathbf{D}_1(\tau) \boldsymbol{\delta} + \lambda_0 \left[\sum_{j=1}^{n} \operatorname{sgn}(\beta_j) \mathbb{1}_{\{\beta_j \neq 0\}} + |\delta_j| \mathbb{1}_{\{\beta_j = 0\}} \right].$$

Da auch $Z_n(\boldsymbol{\delta})$ eine konvexe Funktion ist, kann das selbe Argument wie im Beweis von Satz 4.2.1 angewendet werden und es folgt

$$\operatorname*{argmin}_{\boldsymbol{\delta}} (Z_n(\boldsymbol{\delta})) = \hat{\boldsymbol{\delta}}_n = \sqrt{n} \left(\hat{\boldsymbol{\beta}}(\tau, \lambda_n) - \boldsymbol{\beta}(\tau) \right) \xrightarrow{d} \operatorname*{argmin}_{\boldsymbol{\delta}} Z_0(\boldsymbol{\delta}).$$

Knight & Fu (2000) haben sich auch noch mit einer breiteren Klasse von Straftermen basierend auf L_p-Normen beschäftigt und gezeigt, dass das Akaike-Informations-Kriterium ein Grenzfall dieses Ansatzes ist, wobei an dieser Stelle für mehr Details auf deren Arbeit

verwiesen wird.

5. QR-Modelle für Versicherungsleistungsdaten

Die bisher gezeigten Methoden und Modellierungsansätze werden nun auf Leistungsdaten einer Versicherung angewendet, um deren Verhalten genauer zu analysieren und daraus fundierte Aussagen über zukünftig benötigte Aufwendungen der Versicherung ableiten zu können. Für diesen Fall muss noch eine Klasse von QR-Modellen eingeführt werden, nämliche die der *nichtparametrischen QR-Modelle*.

5.1. Nichtparametrische Quantile Regression

Obwohl parametrische Modelle eine sehr große Rolle spielen, gibt es einige Situationen, in denen sie versagen und man zu einem flexibleren Ansatz übergehen muss. In diesen Fällen kommen *nichtparametrische Modelle* ins Spiel. Das Ziel hier ist es die optimale Regressionsfunktion $f(\mathbf{x})$ zu finden, die die Form der aus der Verteilung der Daten resultierenden Quantilsfunktionen am besten widerspiegelt. Im Gegensatz zum parametrischen Fall, wo die Form der Quantilsfunktion durch

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}(\tau)$$

genau vorgegeben wurde, trifft man nun folgende Modellannahme:

$$f(\mathbf{x}) = Q_Y(\tau|\mathbf{x}).$$

Dadurch erhält man das Minimierungsproblem

$$\min_{f(.)\in\mathcal{C}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - f(\mathbf{x}_i)\right),$$

wobei $\mathcal{C}$ ein passend gewählter Funktionenraum ist.

Für die Funktion $f(.)$ können beispielsweise *B-Splines* verwendet werden, das heißt die Funktion ist von der Gestalt

$$f(x) = \sum_{j=1}^{m} B_j^k(x)\beta_j,$$

wobei diese im Anhang C noch genauer erläutert werden. Da in der folgenden Anwendung nur eine Prädiktorvariable vorkommt, wird nun nur mehr x_i statt dem vektorwertigen $\mathbf{x}_i$ verwendet.

Eine weitere Möglichkeit zur Schätzung nichtparametrischer QR-Modelle ist das Hinzufügen eines Strafterms in die Zielfunktion, was zu *Quantile-Smoothing-Splines* führt. Ein ähnliches Vorgehen kennt man bereits von Generalisierten Linearen Modellen. Diese Form der Quantilsschätzer werden beispielsweise bei Hendricks & Koenker (1992) vorgestellt und sind auch in Davino et al. (2013) zu finden.

Als Minimierungsproblem ergibt sich in diesem Fall

$$\min_{f(.)\in\mathcal{C}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - f(x_i)\right) - \lambda \int \left(|f''(x)^p|\right)^{1/p} \mathrm{d}x,$$

mit $p \geq 1$. $\mathcal{C}$ ist der Sobolev-Raum der Funktionen $f(.)$, mit quadratisch integrierbarer zweiter Ableitung. Der Parameter λ bestraft zu hohe Schwankung der geschätzten Funktion, das heißt er kontrolliert die Glattheit der geschätzten Kurve. Je größer λ gewählt wird, desto glatter wird die Funktion. Dieses Problem setzt weiter voraus, dass die vierte Ableitung der Funktion fast überall Null ist. So eine Bedingung nennt man *Euler-Bedingung ersten Grades*. Dies garantiert, dass es sich bei $f(.)$ um ein stückweises Polynom dritten Grades handelt, welches jeweils in den Beobachtungen x_i springt. Weiters wird angenommen, dass es sich bei $f(.)$ um eine stetige Funktion handelt.

Für $p = 2$ geben Bosch et al. (1995) einen Inneren-Punkte-Algorithmus zur Lösung des Minimierungsproblems an und Koenker et al. (1994) geben detailierte Methoden zur Lösung solcher Probleme an. Die Lösungen dieser Minimierungsprobleme sind wieder Polynome dritten Grades, die nun Quantilsfunktionen der Responseverteilung gegeben dem Prädiktor zum jeweiligen Niveau τ schätzen.

Eine weitere mögliche Wahl des Strafterms ist die Verwendung der *Totalen Variation* einer Verteilung. Für eine Funktion $f : [a, b] \to \mathbb{R}$ ist die Totale Variation wie folgt definiert:

$$V(f) = \sup_{a=x_1\leq\ldots\leq x_n=b} \sum_{i=1}^{n-1} \left|f(x_{i+1}) - f(x_i)\right|.$$

Falls so ein reelles Supremum nicht gefunden werden kann, dann setzt man $V(f) = \infty$. Es gilt, dass für eine Funktion $f(.)$, deren erste Ableitung absolut stetig ist,

$$V(f') = \int_a^b |f''(x)| \,\mathrm{d}x.$$

In Koenker (2005) wird gezeigt, dass diese Strafterme zu folgendem Minimierungspro-

blem führen:

$$\min_{f(.)\in\mathcal{C}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - f(x_i)\right) - \lambda V(f'),$$

wobei dadurch eine Funktion $f : [a, b] \to \mathbb{R}$ gefunden wird, deren erste Ableitung absolut stetig ist. Die Lösung entspricht einer stückweisen linearen Funktion mit Knotenpunkten x_i. Dieser Strafterm wird bei der in `R` implementierten Funktion `rqss()` aus dem Paket `quantreg` zur Berechnung solcher Modelle als Default-Wert verwendet. Mehr Details zu dieser Methoden und Ansätzen finden sich in Koenker (2005).

5.2. Praktische Anwendung

Die zuvor vorgestellten Methoden werden nun auch in die Tat umgesetzt und auf reale Daten angewendet. Dazu betrachtet man einen Datensatz der die Klientel einer Versicherung über drei Jahre (2010, 2011 und 2012) beinhaltet. Dieser enthält 184930 Beobachtungen und jede wird durch die folgenden Variablen charakterisiert:

- `LNR` ... Leistungsnummer zur Identifizierung der Person, die die Leistung in Anspruch genommen hat
- `Jahr` ... Betrachtet werden die Jahre 2010, 2011 und 2012
- `TKZ` ... Tarifklasse
  - `TKZ` = 1 ... Ambulante Kosten
  - `TKZ` = 2 ... Spitalskosten
- `Alter`
- `Geschlecht` ... W (Frauen) und M (Männer)
- `Leistung` ... Enthält die beanspruchten Leistungen der Versicherungsnehmer in jedem der drei Jahre und für die beiden Tarifklassen, das heißt maximal 6 Kostenstellen pro Person im Datensatz

Modelliert wird nun die Responsevariable `Leistung` in Abhängigkeit vom Prädiktor `Alter`. Geschlecht beziehungsweise Tarifklassen, sowie die einzelnen Jahre sollen vorerst getrennt voneinander analysiert werden, das heißt der Datensatz muss dementsprechend gesplittet werden.

Anhand von Tabelle 5.1 sieht man, dass das maximale Alter in der zweiten Tarifklasse deutlich höher ist als in der ersten. Bei den Spitalskosten sind auch die Maxima der beanspruchten Leistungen bei weitem höher, was darauf zurückzuführen ist, dass diese in der Regel höher sind als die reinen Behandlungskosten. In der ersten Tarifklasse befinden sich beispielsweise Kosten für Medikamente oder Besuche bei praktischen Ärzten, während in der zweiten Leistungsarten wie Geburtenkostenbeihilfe und Krankenhaus-Pflegegebühren einfließen.

Jahr = 2010	**Anzahl**	**Max**	**Mittel**	**Median**	**Alter**
TKZ = 1 ∧ W	6552	1785.39	192.63	62.27	0-62
TKZ = 1 ∧ M	5522	1770.00	116.90	0	0-66
TKZ = 2 ∧ W	24933	39813.00	878.70	0	0-104
TKZ = 2 ∧ M	18691	47539.90	601.80	0	0-102

Jahr = 2011	**Anzahl**	**Max**	**Mittel**	**Median**	**Alter**
TKZ = 1 ∧ W	9869	2345.00	210.20	80.40	0-65
TKZ = 1 ∧ M	8374	1835.00	127.72	4.05	0-69
TKZ = 2 ∧ W	24756	41100.90	893.00	0	0-105
TKZ = 2 ∧ M	18668	32682.00	600.70	0	0-103

Jahr = 2012	**Anzahl**	**Max**	**Mittel**	**Median**	**Alter**
TKZ = 1 ∧ W	12987	2668.00	209.86	70.75	0-70
TKZ = 1 ∧ M	11531	1858.00	123.40	0	0-70
TKZ = 2 ∧ W	24438	39578.00	866.00	0	0-106
TKZ = 2 ∧ M	18609	41006.60	624.90	0	0-104

Tabelle 5.1.: Anzahl der Versicherungsnehmer in den einzelnen Klassen, deren maximale, mittlere, sowie mediane Leistung und die Altersbereiche

Das Leistungsminimum ist bei allen Gruppen Null, da es immer Personen gibt, die keine Zahlungen in Anspruch genommen haben. Weiters ist die Anzahl der Frauen in beiden Tarifklassen in allen Jahren größer als die der Männer. Die mittleren Leistungen schwanken in den einzelnen Teildatensätzen über die Jahre kaum, was auch für den Median gilt, der beinahe in allen Fällen den Wert Null hat. Das zeigt, dass in diesen Fällen zumindest die Hälfte der betrachteten Personen keine Leistung benötigt hat.

Als Beispiel zeigt Abbildung 5.1 die Leistungsdaten für Frauen aus dem Jahr 2012 unterteilt nach den beiden Tarifklassen. Aufgrund des Scatterplots ist eine unterschiedliche Altersabhängigkeit in den beiden Gruppen zu erkennen, was man auch anhand von Abbildung 5.2 deutlicher feststellen kann. Im Alter zwischen 20 und 40 Jahren ist bei den Spitalskosten eine deutliche Erhebung zu sehen, die sich dadurch begründen lässt, dass bei diesem Alter viele Geburten eintreten.

Betrachtet man alternativ dazu Abbildung 5.3, die Leistungen für die Männer unterteilt nach den beiden Tarifklassen für dieses Jahr zeigt, sieht man deutlich, dass hier dieser Geburtenhügel nicht vorhanden ist.

Da die Datenlage für hohe Altersstufen sehr dünn wird und man nur mehr vereinzelte Beobachtungen hat, betrachtet man bei den Spitalskosten nur Personen bis zu einem Alter von 85 Jahren und bei Ambulanten Kosten schneidet man bei 60 Jahren ab.

Diese Erkenntnisse sind auch auf die Daten der Jahre 2010 und 2011 übertragbar und daher werden dafür keine Abbildungen angegeben. In weiterer Folge betrachtet man bei-

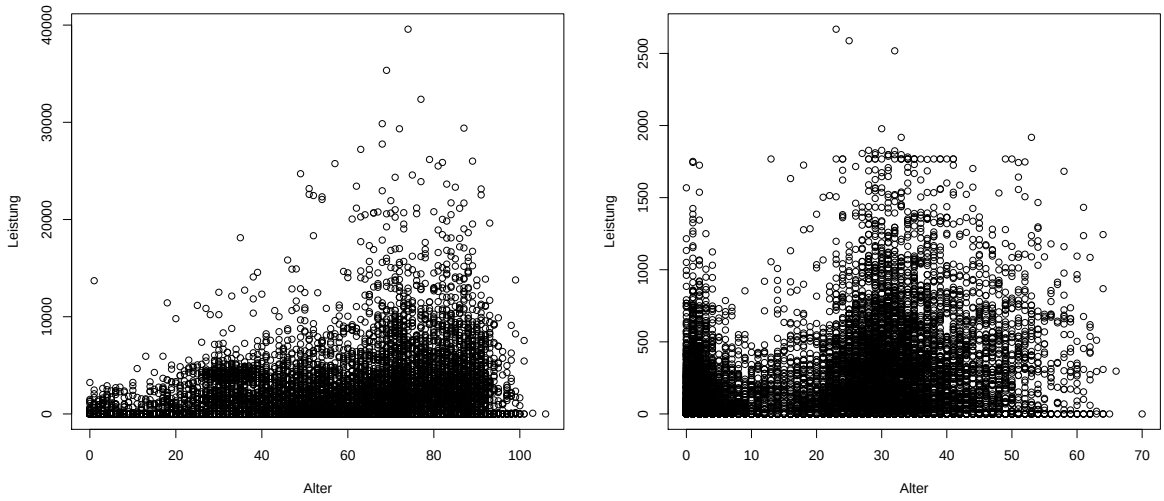

Abbildung 5.1.: Spitalskosten (links) und Ambulante Kosten (rechts) für Frauen im Jahr 2012.

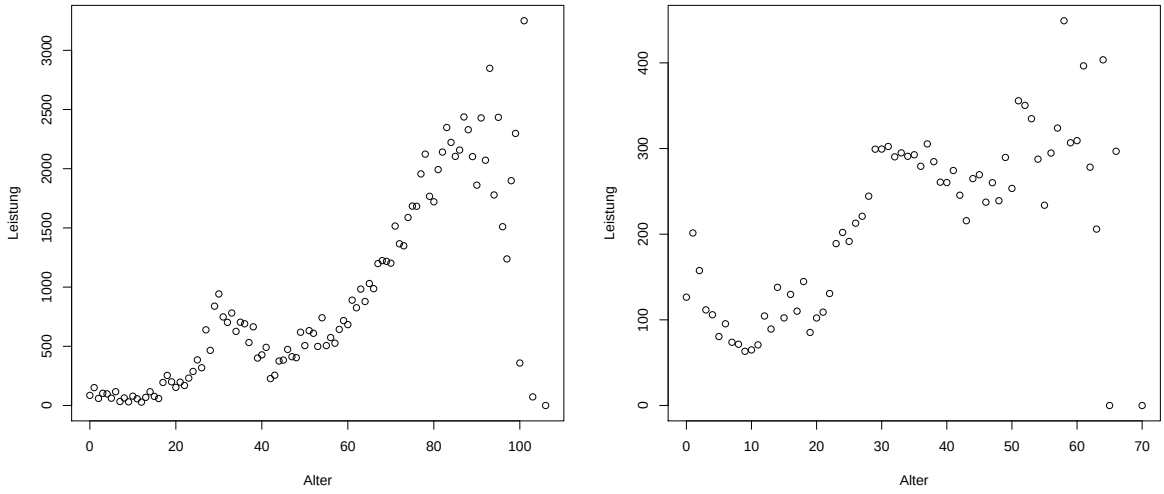

Abbildung 5.2.: Mittlere Leistungen für Frauen im Jahr 2012; Spitalskosten (links) und Ambulante Kosten (rechts).

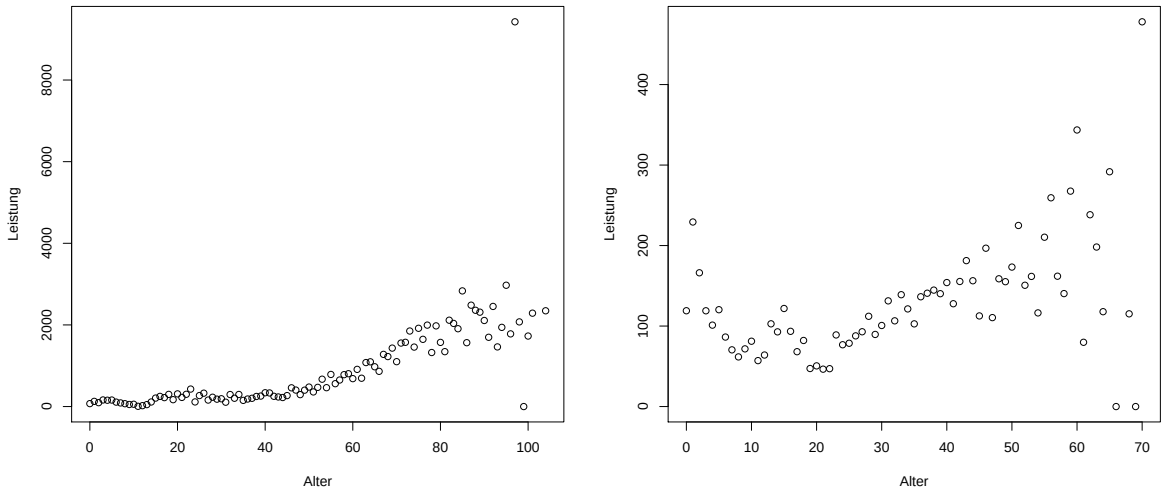

Abbildung 5.3.: Mittlere Leistungen für Männer im Jahr 2012; Spitalskosten (links) und Ambulante Kosten (rechts).

spielhaft den Subdatensatz für Frauen aus dem Jahr 2012 und erstellt für jede Tarifklasse ein *QR-Modell* mit Hilfe des `R`-Paket `quantreg`. Auch sollen alle drei Jahre bezüglich der beiden Tarifklassen gegenüber gestellt werden, um so zu erkennen, wie sich die Kosten der Versicherung über die Jahre verhalten.

Zur praktischen Anwendung der Methoden Quantiler Regression stellt Koenker (2005) sein `R`-Paket `quantreg` vor. Eine der wichtigsten Funktionen in diesem Paket ist die Funktion `rq`(). Sie ist das Gegenstück zu `lm`() bei Linearen Modellen beziehungsweise zu `glm`() bei den Generalisierten Linearen Modellen mit `R`. Mit Hilfe dieser Funktion wird ein QR-Modell angepasst.

Auch altbekannte Funktion wie beispielsweise `summary`(), `plot`() oder `anova`() sind in diesem Paket verfügbar. Mit folgendem `R`-Kommando erhält man eine vollständige Auflistung aller verfügbaren Befehle:

Hilfe zu Paket quantreg

```
1 help(package="quantreg")
```

5.3. Modellierung

Als erste Anwendung betrachtet man nun die Spitalskosten für Frauen im Jahr 2012. Aufgrund der Struktur der Daten kann die Annahme eines linearen Zusammenhangs zwischen der Response *Leistung* und dem Prädiktor *Alter* verworfen werden.

Für den ausgewählten Datensatz (W, TKZ=2, 2012) möchte man also ein *nichtparametrisches* QR-Modell schätzen und verwendet dazu *B-Splines*, die im Anhang C näher erläutert werden.

Um dies in die Tat umzusetzen, wendet man folgenden R-Code an:

Erstes QR-Modell

```
1 X <- model.matrix(Leistung~bs(Alter, degree=3, df=11))
2 Fit <- rq(Leistung~bs(Alter, degree=3, df=11), tau=tau)
3 QRfit <- X %*% Fit$coef
```

In Zeile 1 generiert man die Design- oder Modellmatrix `X`. Für `tau` wird das gewünschte Quantilsniveau eingetragen, je nachdem welche Quantilsfunktion man schätzen möchte. Es kann auch ein Vektor angegeben werden, wenn mehrere Funktionen auf einmal angepasst werden sollen. Die Verwendung der Funktion `bs()` zeigt an, dass mit B-Splines gearbeitet wird und der Eingabeparameter `degree=3` zeigt an, dass es sich um kubische Splines handelt. Die Anzahl der Knotenpunkte wird durch `df-degree` festgelegt, ist in diesem Fall also 8. In Zeile 4 werden noch die gefitteten Werte berechnet.

In Abbildung 5.4 sind nun einige so generierte Quantilsfunktionen dargestellt, wobei rechts nur eine detailiertere Darstellung der linken Graphik gezeigt wird.

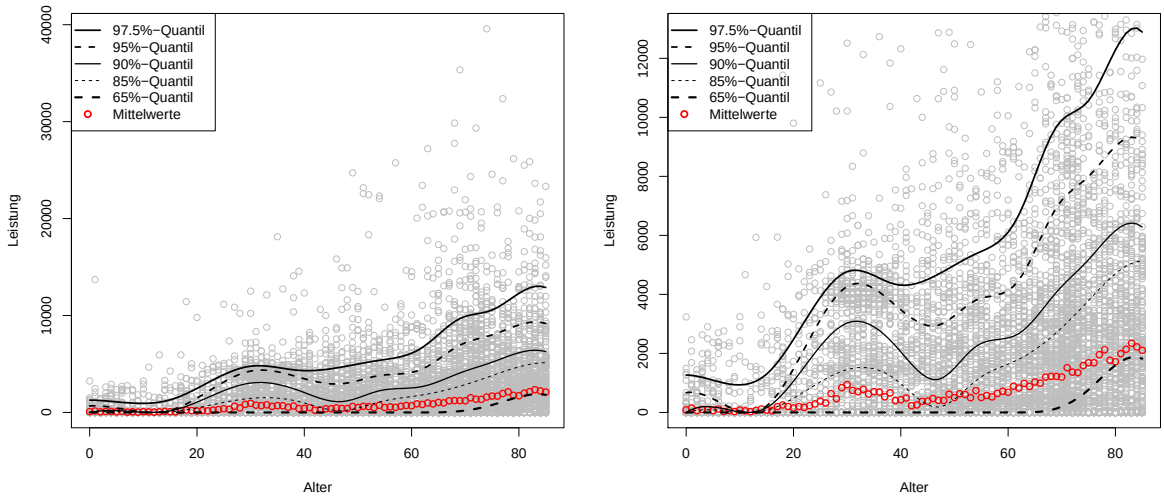

Abbildung 5.4.: Geschätzte Quantilsfunktionen der Spitalskosten für Frauen im Jahr 2012, $\tau = \{0.65, 0.85, 0.90, 0.95, 0.975\}$ (links); Detailansicht (rechts).

Anhand von Abbildung 5.4 sieht man, dass die einzelnen Quantilsfunktionen die Datenlage sehr gut modellieren. Der Geburtenhügel im Bereich zwischen 20 und 40 Jahren ist bei den hohen Quantilen gut zu erkennen, was daran liegt, dass für Geburten größere Leistungen anfallen. Am deutlichsten wird die Erhebung für $\tau = 0.95$, $\tau = 0.90$ und $\tau = 0.85$.

Bei kleinerem τ erhält man einen eher konstanten Verlauf, der nur im hohen Alter eine Erhebung aufweist. Dies liegt daran, dass sehr viele Leistungen der Höhe Null vorliegen.

Man kann im Allgemeinen zu jedem Quantilsniveau τ einen steigenden Verlauf der Kosten feststellen, das heißt im höheren Alter fallen für die Versicherung auch höhere Kosten an.

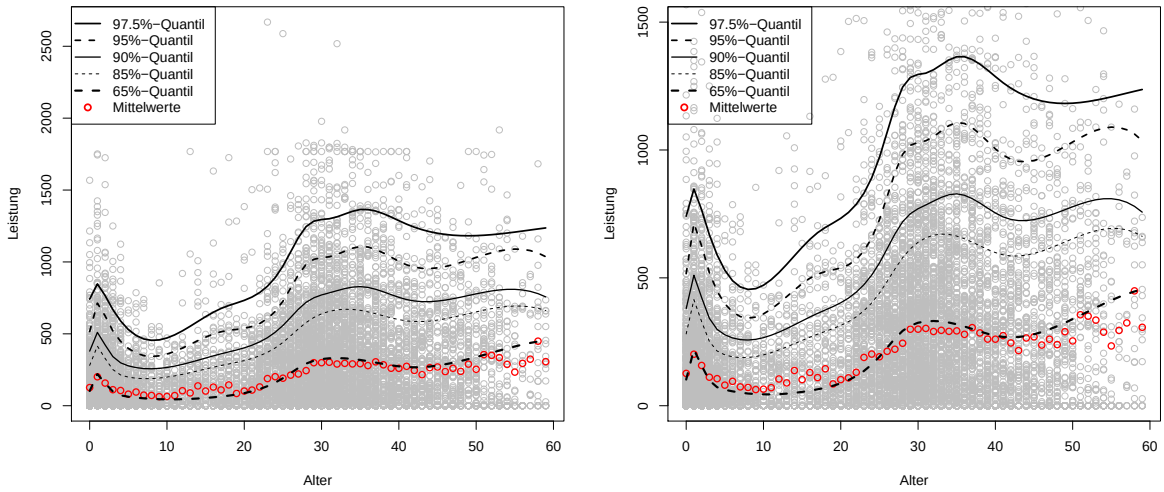

Abbildung 5.5.: Geschätzte Quantilsfunktionen der Ambulanten Kosten für Frauen im Jahr 2012, $\tau = \{0.65, 0.85, 0.90, 0.95, 0.975\}$ (links); Detailansicht (rechts).

Im Vergleich dazu zeigt Abbildung 5.5 QR-Modelle für die Ambulanten Kosten von Frauen im Jahr 2012, die sich in der ersten Tarifklasse befinden. Hier ist die Geburten-Erhebung nicht zu erkennen, da diese Kosten in dieser Gruppe keinen Einfluss haben. Anhand dieser Abbildung sieht man, dass bei den Ambulanten Leistungen ebenfalls mit dem Alter ansteigende Quantilsfunktionen vorliegen.

Bei den Ambulanten Kosten ist bereits das 65%-Quantil in allen Altersbereichen von Null verschieden. Das bedeutet, dass es in dieser Klasse weniger Personen gibt, die keine Leistung in Anspruch genommen haben. Dies dürfte an der Art der Kosten liegen, die in diese Klasse fallen, denn dabei handelt es sich um Behandlungskosten und Ambulante Kosten. Diese sind in der Regel niedriger (deutlich niedrigeres Maximum), aber dafür schneller beziehungsweise öfter auszuzahlen. Es kommt beispielsweise öfter vor, dass jemand ein Medikament verschrieben bekommt oder seinen Hausarzt aufsucht, als ein Spitalsbett oder eine teure Operation in Anspruch zu nehmen.

Abbildung 5.6 gibt nun zum Vergleich die Quantilsfunktionen der Männer im Jahr 2012 für beide Tarifklassen an. Um ein besseres Bild zu bekommen, wurde bei der rechten Graphik die y-Achse abgeschnitten.

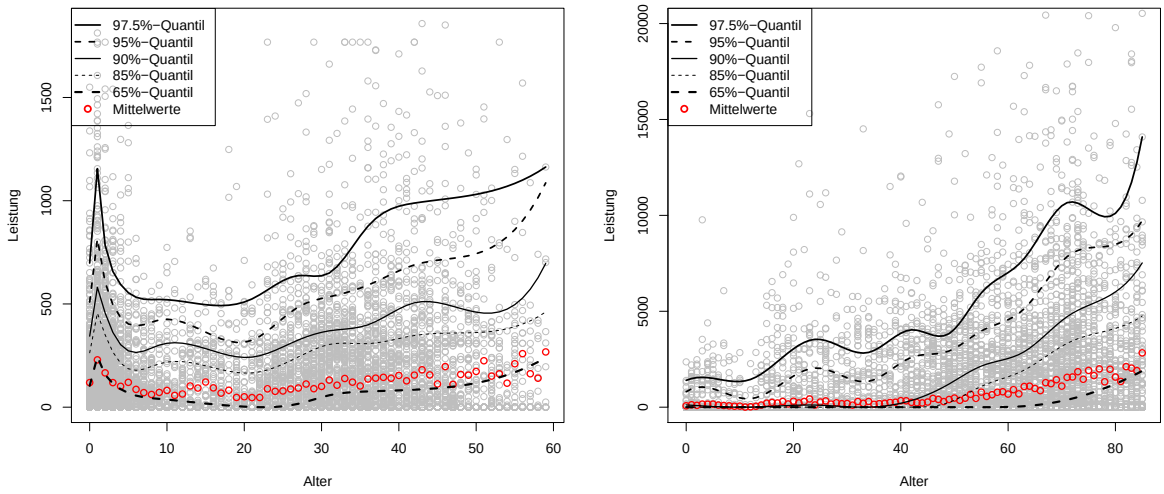

Abbildung 5.6.: Detailansicht der geschätzten Quantilsfunktionen der Ambulanten Kosten (links) und der Spitalskosten (rechts) für Männer im Jahr 2012, $\tau = \{0.65, 0.85, 0.90, 0.95, 0.975\}$.

Man sieht sofort, dass sich die Leistungen der beiden Geschlechter bezüglich der Ambulanten Kosten bei allen Quantilsniveaus sehr ähnlich verhalten.

Einen großen Unterschied jedoch gibt es in Tarifklasse zwei: Bei den Männern fehlt natürlich der Geburtenhügel im Alter zwischen 20 und 40 Jahren. Die Modelle der höheren Quantilsniveaus schwanken ein wenig mehr, als jene im unteren Bereich. Bei kleinem τ beginnt das Quantilsmodell zwar erst ab einem Alter von circa 40 Jahren anzusteigen, es lässt sich jedoch bei allen feststellen, dass die Kosten mit zunehmenden Alter anwachsen.

Analog zum Jahr 2012 werden nun die Plots für die Frauen in beiden Tarifklassen für die Jahre 2011 und 2010 angegeben. Um einen detailierteren Eindruck über den Verlauf und das Verhalten der Quantilsfunktionen zu erhalten, wurden bei den Abbildungen die y-Achsen oben abgeschnitten und so eine „Detailansicht“ erstellt.

Anhand der Abbildungen 5.7 unf 5.8 lässt sich auch in diesen Jahren ein sehr ähnliches Verhalten so wie im (ausführlich betrachteten) Jahr 2012 feststellen.

Würde man alle gegebenen Datenpunkte in die Modellierung mit einbeziehen, das heißt, man würde auch Personen mit einem Alter > 85 in der Klasse der Spitalskosten und > 60 für Ambulante Kosten betrachten, so tritt bei manchen Modellen das Phänomen des *Quantile Crossings* auf. Dies liegt daran, das in diesem Bereich die Datenlage sehr dünn ist und somit dort keine gute Schätzung angegeben werden kann. Dies ist hier aber vernachlässigbar, da die Versicherung kein Interesse daran hat, diese hohen Altersbereiche zu modellieren.

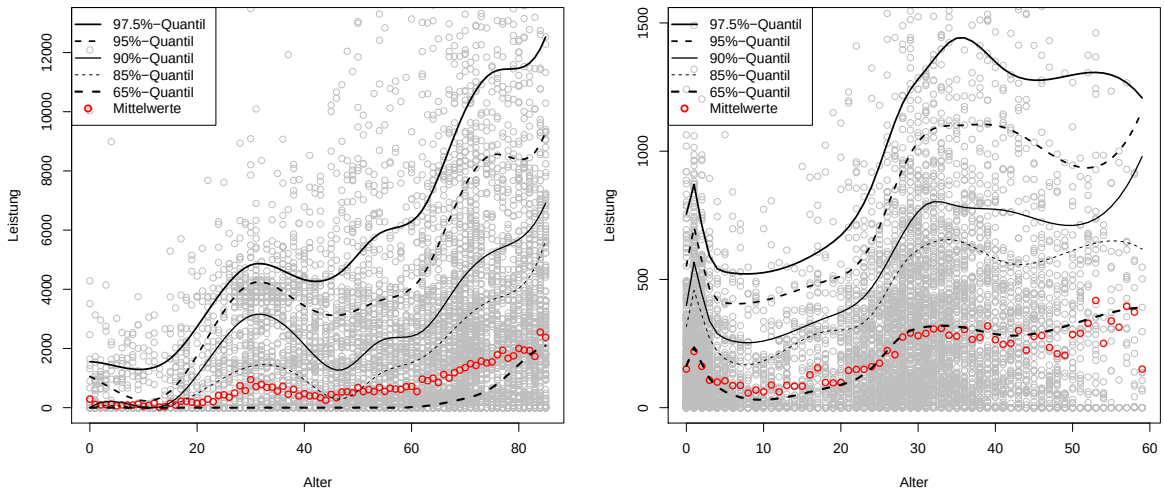

Abbildung 5.7.: Detailansicht der geschätzten Quantilsfunktionen für Frauen im Jahr 2011 bezüglich Spitalskosten (links) und Ambulante Kosten (rechts), $\tau = \{0.65, 0.85, 0.90, 0.95, 0.975\}$.

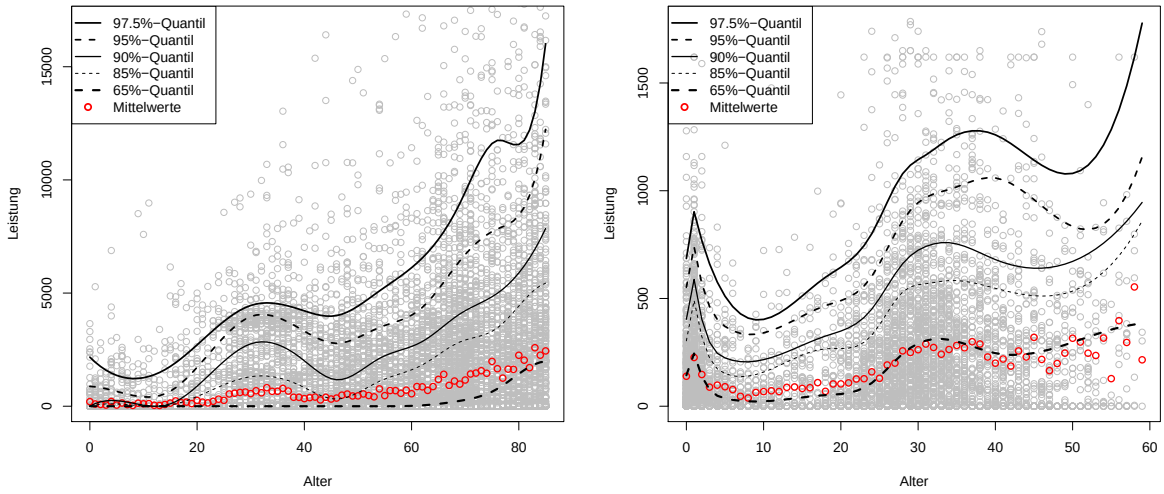

Abbildung 5.8.: Detailansicht der geschätzten Quantilsfunktionen für Frauen im Jahr 2010 bezüglich Spitalskosten (links) und Ambulante Kosten (rechts), $\tau = \{0.65, 0.85, 0.90, 0.95, 0.975\}$.

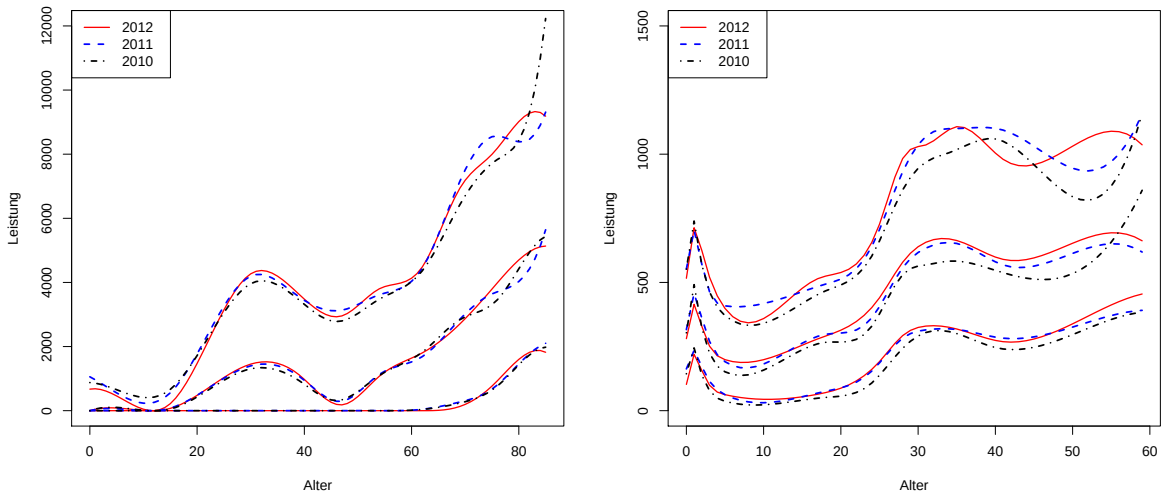

Abbildung 5.9.: Geschätzte Quantilsfunktionen für Frauen bezüglich Spitalskosten (links) und Ambulante Kosten (rechts) über die Jahre 2010, 2011 und 2012, $\tau = \{0.65, 0.85, 0.95\}$.

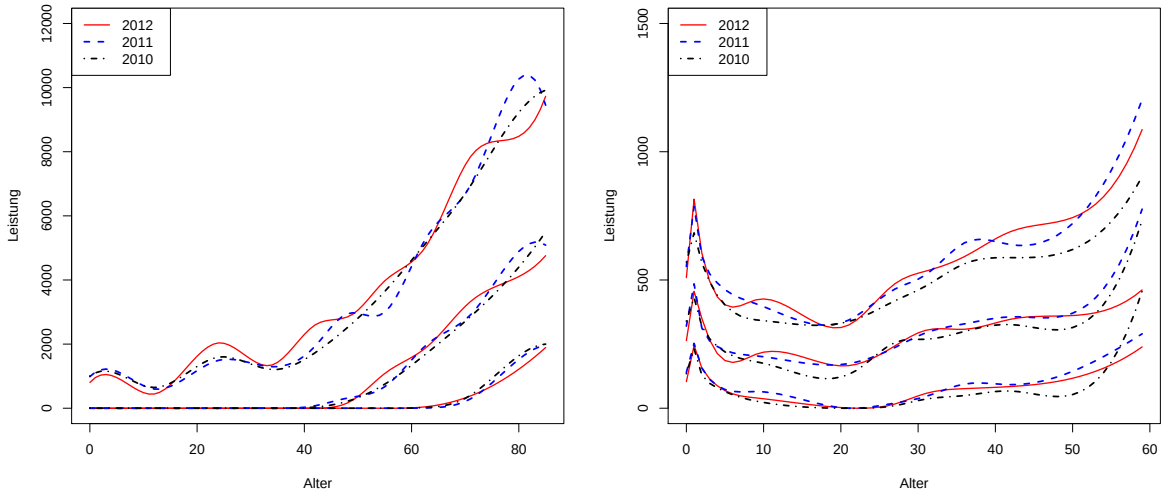

Abbildung 5.10.: Geschätzte Quantilsfunktionen für Männer bezüglich Spitalskosten (links) und Ambulante Kosten (rechts) über die Jahre 2010, 2011 und 2012, $\tau = \{0.65, 0.85, 0.95\}$.

Abbildung 5.9 zeigt nun geschätze Quantilsfunktionen für die Frauen aus allen drei Jahren für $\tau \in \{0.65, 0.85, 0.95\}$ beider Tarifklassen. Es lässt sich keine eindeutige Aussage über den Verlauf der Kosten treffen, das heißt man kann weder von einer Zu- noch einer Abnahme über die Jahre hinweg ausgehen. Gleiches gilt auch bei den Männern, deren Modelle in Abbildung 5.11 dargestellt sind.

5.4. Fazit

Im Allgemeinen kann man sagen, dass durch die Verwendung der Methoden Quantiler Regression und durch die Wahl von B-Splines zur Beschreibung des Zusammenhangs zwischen den Variablen `Leistung` und `Alter`, sehr gute Modelle generiert werden, um den vorliegenden Datensatz zu beschreiben.

Ein Problem stellt die Schätzung von Quantilsfunktionen mit einem $\tau < 0.60$ dar, da aufgrund der vielen Responsevariablen mit Wert Null das Programm `R` keine Fit generieren kann.

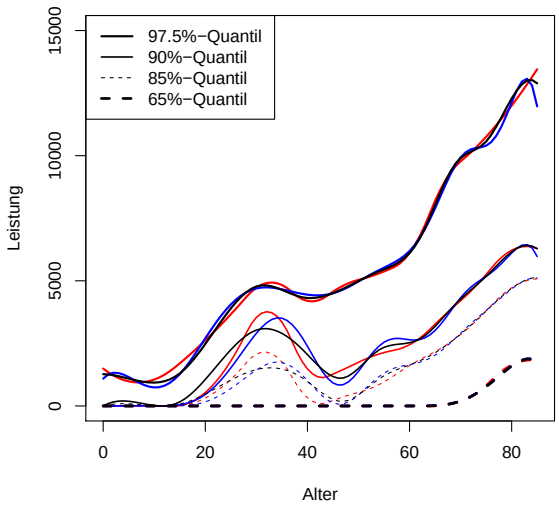

Abbildung 5.11.: Vergleich der geschätzten Quantilsfunktionen für Spitalskosten der Frauen im Jahr 2012, $\tau = \{0.65, 0.85, 0.90, 0.975\}$ generiert durch B-Splines mit `degree` $= 3$ (schwarz), mit `degree` $= 2$ (rot) und mit `degree` $= 4$ (blau).

Wie man anhand der Abbildung 5.11 erkennen kann, wäre auch eine andere Wahl des B-Spline-Grades in Ordnung gewesen. Für niedrige und sehr hohe Quantilsniveaus ($\tau = 0.65$ und $\tau = 0.975$) ist kaum ein Unterschied zu erkennen, was auch in den Randbereichen für alle geschätzten Modelle gilt. Unterschiedliche Schätzungen ergeben sich

beim Geburtenhügel für die dazwischenliegenden Quantile, beispielsweise für $\tau = 0.85$ und $\tau = 0.90$. Die Wahl eines niedrigeren Grades (rot) führt zu einer glatteren Schätzung während ein höherer Grad (blau) größere Schwankungen hervorruft.

Auch die AIC-Werte liefern keinen eindeutig besten Grad der stückweisen Polynome, wie man auch anhand der exemplarisch angegebenen AIC-Werte in Tabelle 5.2 erkennen kann.

Frauen, `TKZ` $= 2$, 2012	$\tau = 0.95$	$\tau = 0.65$
`degree` $= 2$	447900.20	396710.00
`degree` $= 3$	447875.00	396713.10
`degree` $= 4$	447861.60	396714.20

Tabelle 5.2.: AIC-Werte für die geschätzten Quantilsfunktionen zum Niveau $\tau \in \{0.65, 0.95\}$ für Frauen im Jahr 2012 in Tarifklasse 2

Da Wood (2006) gezeigt hat, dass die kubischen Splines in gewissem Sinn die optimale Wahl zur Modellierung darstellen, wurde zuvor zur Analyse der Daten der Eingabeparameter `degree` auf den Wert Drei festgesetzt, was auch gleichzeitig dem Defaultwert der Funktion entspricht. Auch die graphische Analyse der Abbildung 5.11 zeigt, dass diese Wahl ein gutes Mittelmaß für die Modellschätzung zu sein scheint.

Weiters hat sich gezeigt, dass die Anwendung von QR-Modellen in `R` keine großen Schwierigkeiten bereitet und somit in der Praxis eine sehr gute Alternative zur herkömmlichen Regression darstellen.

A. Lineare Programme

Optimierungsprobleme, die eine lineare Funktion bezüglich einer linearen Restriktionsmenge minimieren wollen, nennt man *Lineare Programme* (LP). Das Ziel solcher Methoden ist es beispielsweise Kosten zu minimieren beziehungsweise Profit zu maximieren. In diesem Abschnitt wird nun eine kurze Einführung zu diesem Thema vorgenommen. Als Literaturgrundlage wurde hier Davino et al. (2013) und Koenker (2005) herangezogen.

Die Variablen $Y_i \geq 0$ mit $i = 1, \ldots, n$ nennt man *Entscheidungsvariablen* oder *Zielvariablen*. Sie sollen in gewissem Sinn optimal gewählt werden. Grundsätzlich ist es in einem linearen Programm so, dass ein Vektor $\hat{\mathbf{Y}} \in \mathbb{R}_+^n$ unter allen Vektoren $\mathbf{Y} \in \mathbb{R}_+^n$, die die Bedingungen einer vorgegebene *lineare Restriktionsmenge* erfüllen, ausgewählt werden soll, der den Wert einer vorgegebenen *Zielfunktion* minimiert beziehungsweise maximiert. Bei der Restriktionsmenge handelt es sich um lineare Gleichungen oder Ungleichungen.

Beim QR-Problem handelt es sich eigentlich um ein nichtlineares Optimierungsproblem, dass jedoch durch geeignete Transformationen zu einem linearen Programm umformuliert werden kann. Die daraus resultierende lineare Form garantiert, dass es effiziente Methoden zur Lösung dieses Problems gibt.

Folgende Bedingungen werden von einem typischen LP erfüllt:

Die n Entscheidungsvariablen Y_i sind nichtnegativ. Das heißt, die Lösung wird auf die Menge $\mathbb{R}_+^n$ eingeschränkt. Falls die Zielvariablen jedoch, wie beispielsweise im QR-Fall, unbeschränkte Vorzeichen haben, das bedeutet, die Variablen können positiv, negativ oder gleich Null sein, führt man nichtnegative Variablen $\left\{u_i, v_i \in \mathbb{R}_+^n : i = 1, \ldots, n\right\}$ ein. Diese repräsentieren den positiven und negativen Teil von $\mathbf{Y}$; gleichbedeutend mit

$$\mathbf{Y} = \mathbf{u} - \mathbf{v},$$

wobei $u_i = 0$, falls die Komponente $Y_i \leq 0$ und $v_i = 0$, wenn $Y_i \geq 0$. Das heißt jede unbeschränkte Variable wird als Differenz zweier nichtnegativer Variablen ausgedrückt.

Das Kriterium zur optimalen Wahl der Zielvariable $\mathbf{Y}$, also die Zielfunktion, ist von der Form

$$Z(\mathbf{Y}) = \min_{Y_i \in \mathbb{R}_+} \sum_{i=1}^{n} c_i Y_i = \min_{\mathbf{Y} \in \mathbb{R}_+^n} \mathbf{c}^\top \mathbf{Y}.$$

Die p Restriktionen, die den gesamten Prozess beschränken, können sowohl lineare Glei-

chungen als auch Ungleichungen sein:

$$a_1Y_1 + \ldots + a_nY_n \begin{cases} \leq b, \\ = b, \\ \geq b. \end{cases}$$

Durch die Einführung sogenannter *Schlupfvariablen* $s \geq 0$ können Ungleichungen zu Gleichungen umgeschrieben werden, das heißt

$$a_1Y_1 + \ldots + a_nY_n \leq b \quad \Leftrightarrow \quad a_1Y_1 + \ldots + a_nY_n + s = b.$$

Weiters gilt

$$a_1Y_1 + \ldots + a_nY_n \geq b \quad \Leftrightarrow \quad -a_1Y_1 - \ldots - a_nY_n \leq -b.$$

Im Gegensatz dazu können auch Gleichheitsbedingungen zu Ungleichungen umformuliert werden:

$$a_1Y_1 + \ldots + a_nY_n = b \quad \Leftrightarrow \quad \begin{aligned} a_1Y_1 + \ldots + a_nY_n &\leq b \\ a_1Y_1 + \ldots + a_nY_n &\geq b \end{aligned}$$

Daraus resultiert, dass es keinen Unterschied macht, wie die Restriktionsmenge eines LPs gegeben ist, da sie sowieso immer in Standardform transformiert werden kann.

Betrachtet man dies nun von einem geometrischen Standpunkt aus, so legen die linearen Gleichungen Hyperebenen fest, während eine Ungleichung den n-dimensionalen Raum in zwei Teile schneidet. In einem Bereich liegen die Punkte, die die Restriktion erfüllen und im anderen jene, für die die Ungleichung nicht erfüllt ist. Eine *Hyperebene* ist ein Unterraum eines Vektorraums, der von genau einem Basisvektor weniger aufgespannt wird, als der Gesamtraum. Im zweidimensionalen Raum sind dies beispielsweise Geraden.

Fasst man alle Bedingungen an die Variablen Y_i zu einer Restriktionsmenge zusammen, so ist die *Lösungsmenge* ein Schnittpunkt der $n + p$ Unterräume, die durch die Nichtnegativitätsbedingungen der n Zielvariablen und der p Gleichungen beziehungsweise Ungleichungen, die den gesamten Prozess regulieren, festgelegt werden.

Die Restriktionsmenge kann geometrisch als *konvexer Polyeder* interpretiert werden, wobei seine Dimension nach oben durch die Anzahl der Parameter beschränkt ist. Die Minimierung der Zielfunktion entspricht dann einer Verschiebung der Hyperebene, die durch die Zielfunktion definiert wird, in Richtung des Vektors $\mathbf{c}$ solange bis das Polyeder gerade noch berührt wird, das heißt eine zulässige Lösung des Problems liegt immer in einer Ecke des konvexen Restriktionsvielecks.

Die Standardform eines Linearen Programms in Matrixschreibweise sieht nun wie folgt

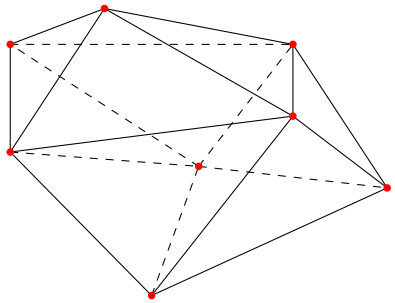

Abbildung A.1.: Unregelmäßiges Polyeder mit glatten Begrenzungsflächen, geraden Verbindungskanten, die sich in einzelnen Eckpunkten schneiden.

aus:

$$\begin{aligned} \min\ & \mathbf{c}^\top \mathbf{Y} \\ \text{subject to}\ & \mathbf{AY} \leq \mathbf{b} \\ & \mathbf{Y} \geq 0, \end{aligned}$$

mit $\mathbf{Y} \in \mathbb{R}^{n\times 1}$, $\mathbf{c} \in \mathbb{R}^{n\times 1}$, $\mathbf{b} \in \mathbb{R}^{p\times 1}$ und $\mathbf{A} \in \mathbb{R}^{p\times n}$. Der Vektor $\mathbf{c}$ enthält die Gewichte beziehungsweise Kosten der Zielvariablen und durch die Matrix $\mathbf{A}$ und dem Vektor $\mathbf{b}$ werden die p Restriktionen festgelegt. Eine *zulässige Lösung* $\mathbf{Y}$ dieses Problems erhält man, wenn $\mathbf{Y}$ alle Restriktionen erfüllt. Minimiert $\hat{\mathbf{Y}}$ auch noch den Zielfunktionswert, so hat man eine *Optimallösung* gefunden. Der Vektor $\hat{\mathbf{Y}}$ ist also in der Menge der zulässigen Lösungsvektoren jener mit minimalen Kosten.

Zu jedem linearen Programm gibt es auch sein dazu passendes *duales Programm*. Das zuvor vorgestellte LP nennt man in diesem Zusammenhang *primales Programm*. Die duale Formulierung enthält die gleiche Koeffizientematrix $\mathbf{A}$ und den Vektor $\mathbf{b}$, dreht aber alles um. Der primale Kostenvektor $\mathbf{c}$ wird zum dualen Restriktionsvektor und umgekehrt wird $\mathbf{b}$ zum dualen Profitvektor. Ein duales Programm ist also von der Form

$$\begin{aligned} \max\ & \mathbf{b}^\top \mathbf{Y} \\ \text{subject to}\ & \mathbf{YA} \geq \mathbf{c} \\ & \mathbf{Y} \geq 0. \end{aligned}$$

Es zeigt sich schließlich, dass jede zulässige Lösung des primalen Programms eine Schranke des optimalen Zielfunktionswerts des dualen Problems ist und umgekehrt.

Der Satz von der *Schwachen Dualität* besagt dann, dass das duale Programm obere Schranken für das primale liefert, das heißt

$$\mathbf{c}^\top \mathbf{Y} \leq \mathbf{b}^\top \mathbf{Y}.$$

Laut dem Theorem der *Starken Dualität* gilt weiters, dass wenn das primale Programm

eine Optimallösung $\hat{\mathbf{Y}}$ besitzt

$$\mathbf{c}^\top \hat{\mathbf{Y}} = \mathbf{b}^\top \hat{\mathbf{Y}}$$

hält.

Es zeigt sich, dass es oft leichter ist eine Optimallösung des primalen LPs zu finden, indem man bei dem dualen Programm beginnt.

A.1. Methoden zur Bestimmung einer Optimallösung

Eine sehr häufig verwendete Methode zur Lösung solcher Linearen Programme ist der **Simplex-Algorithmus**, der von Dantzig (1963) vorgestellt wurde. Die Idee, die dahinter steckt, ist einfach: Man beginnt in einem Punkt der Menge der zulässigen Lösungen und sucht dann nach einer neuen, besseren Lösung. Besser bedeutet in diesem Zusammenhang, dass sich der Zielfunktionswert verkleinert (primal) beziehungsweise vergrößert (dual). Der Prozess iteriert solange, bis schlussendlich keine Verbesserung möglich ist.

Eine intuitive Erklärung des Simplex-Algorithmus ergibt sich aus der geometrischen Interpretation des Problems: Gibt es eine zulässige Lösung des Problems, so liegt die Optimallösung, wie bereits zuvor beschrieben, in einer Ecke des konvexen Polyeders, welches die $(n + p)$ Restriktionen beschreibt. Der Simplex-Algorithmus besteht also aus Bewegungen entlang der Kanten der Menge der zulässigen Lösungen oder äquivalent entlang der Kanten des Restriktionspolyeder, wobei in einer Ecke der Menge begonnen wird. Typischerweise gibt es in so einer Ecke n Kanten, in deren Richtung der Algorithmus fortgesetzt werden kann, manche davon verbessern sukzessive den Zielfunktionswert andere wiederum verschlechtern ihn. Dantzig (1963) schlägt vor immer eine neue Ecke mit kleinerem Zielfunktionswert zu wählen und zeigt, dass es dann keine Möglichkeit mehr gibt, zu etwas schlechterem zurückzukehren. Die Optimallösung ist dann ein Punkt, aus dem nur mehr Kanten wegführen, die den Zielfunktionswert veschlechtern würden.

Weitere Möglichkeiten zur Lösung solcher Programme wären beispielsweise **Innere-Punkte Verfahren**. Dort nähert man sich der Optimallösung aus dem inneren der Menge der zulässigen Lösungen, statt der Bewegung entlang der Kanten. Diese Methode ist vor allem bei Problemen mit großem Stichprobenumfang effizienter. Mehr Details dazu und zur Anwendung auf das QR-Problem finden sich in Koenker (2005).

B. Beweis zu alternativer Darstellung der Zielfunktion

Hier wird nun die Richtigkeit der alternativen Darstellung der Zielfunktion gezeigt.

Satz 1 (Äquivalente Darstellung)**.** *Für die Zielfunktion $Z(\boldsymbol{\beta})$ gilt:*

$$Z(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left[\tau - \frac{1}{2} + \frac{1}{2}\operatorname{sgn}(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\right] \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)$$

mit

$$\operatorname{sgn}(u) = \begin{cases} 1, & u > 0, \\ -1, & u < 0, \\ 0, & u = 0. \end{cases}$$

Beweis. Zuerst zeigt man, dass

$$Z(\boldsymbol{\beta}) = \sum_{\left\{i\,:\,Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} > 0\right\}} \left|Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right| \cdot \tau + \sum_{\left\{i\,:\,Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} < 0\right\}} \left|Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right| \cdot (1 - \tau).$$

Dazu betrachtet man

$$Z(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) \left(\tau - \mathbb{1}_{\left\{\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) < 0\right\}}\right)$$

und zerlegt dies dann in zwei Summen:

$$\begin{aligned} &\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) \left(\tau - \mathbb{1}_{\left\{\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) < 0\right\}}\right) = \\ &= \sum_{\left\{i\,:\,Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} > 0\right\}} \underbrace{\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)}_{>0} \tau + \sum_{\left\{i\,:\,Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} < 0\right\}} \underbrace{\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)}_{=(-1)\cdot\left|Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right|} (\tau - 1) \\ &= \sum_{\left\{i\,:\,Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} > 0\right\}} \left|Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right| \tau + \sum_{\left\{i\,:\,Y_i - \mathbf{x}_i^\top \boldsymbol{\beta} < 0\right\}} \left|Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right| (1 - \tau). \end{aligned}$$

Nun gilt weiters, dass

$$\sum_{\{i\,:\,Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}>0\}} \tau\cdot\left|Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right| + \sum_{\{i\,:\,Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}<0\}} (1-\tau)\cdot\left|Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right| =$$
$$\sum_{i=1}^{n}\left[\tau-\frac{1}{2}+\frac{1}{2}\mathrm{sgn}(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\right]\left[Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right],$$

wobei die Funktion $\mathrm{sgn}(u)$ wie folgt definiert ist:

$$\mathrm{sgn}(u)=\begin{cases}1, & u>0,\\ -1, & u<0,\\ 0, & u=0.\end{cases}$$

Um diesen letzten Schritt zu zeigen führt man eine Fallunterscheidung durch:

Fall 1: $Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}>0$

$$\sum_{\{i\,:\,Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}>0\}}\left[\tau-\frac{1}{2}+\frac{1}{2}\right]\underbrace{\left(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right)}_{=1\cdot\left|Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right|} = \sum_{\{i\,:\,Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}>0\}}\tau\cdot\left|Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right|\checkmark$$

Fall 2: $Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}<0$

$$\sum_{\{i\,:\,Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}<0\}}\left[\tau-\frac{1}{2}-\frac{1}{2}\right]\underbrace{\left(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right)}_{=(-1)\cdot\left|Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right|} = \sum_{\{i\,:\,Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}<0\}}[1-\tau]\cdot\left|Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right|\checkmark$$

Fall 3: $Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}=0\rightarrow$ Klar ✓

Setzt man nun alles zusammen erhält man die Aussage von Satz 3.2.2. □

C. B-Splines

Ein nichtparametrisches Modell ist allgemein von der Form

$$Y_i = f(x_i) + \varepsilon_i, \text{ für } i = 1, \ldots, n.$$

Die Funktion $f(.)$ wird dabei als *Glättungsfunktion* bezeichnet und kann mit Hilfe von *Basisfunktionen* angegeben werden. Dies liefert dann folgende Darstellung von $f(.)$:

$$f(x) = \sum_{j=1}^{p+1} b_j(x)\beta_j,$$

wobei $b_j(x)$ die j-te Basisfunktion bezeichnet. Als Funktion $f(x)$ kann dabei ein Polynom-Spline verwendet werden, der wie folgt bei Fahrmeir et al. (2013) definiert wird

Definition C.0.1 (Polynom-Spline)**.** *Eine Funktion $f : [a, b] \to \mathbb{R}$ heißt **Polynom-Spline** vom Grad $p \geq 0$ zu den Knoten $a = x_1 < \ldots < x_m = b$, falls sie die folgenden Bedingungen erfüllen:*

- *$f(x)$ ist $(p-1)$ mal stetig differenzierbar.*
- *$f(x)$ ist auf den durch die Knotenpunkte gebildeten Intervallen $[x_j\ , x_{j+1})$ ein Polynom vom Grad p.*

Ein Spezialfall dieser Polynom-Splines, ist der sogenannte B-Spline, der von Wood (2006) durch folgende Definition charakterisiert wird:

Definition C.0.2 (B-Spline)**.** *Ein **B-Spline** $f(x)$ ist ein Polynom-Spline der Ordnung $(k+1) = p$*

$$f(x) = \sum_{j=1}^{m} B_j^k(x)\beta_j,$$

mit rekursiv definierten Basisfunktionen der Form

$$B_j^k(x) = \frac{x - x_j}{x_{j+k+1} - x_j} B_j^{k-1}(x) + \frac{x_{j+k+2} - x}{x_{j+k+2} - x_{j+1}} B_{j+1}^{k-1}(x), \quad j = 1, \ldots, m$$

und

$$B_j^{-1} = \begin{cases} 1, & x_j \leq x \leq x_{j+1} \\ 0, & sonst. \end{cases}$$

Zur Konstruktion dieser Splines müssen also $m + k + 2$ Knoten $x_1 < \cdots < x_{m+k+2}$ gewählt werden. Der Spline wird dann über dem Intervall $[x_{k+2}, x_m]$ ausgewertet und somit ist die Wahl der äußeren $2(k + 1)$ Knoten mehr oder weniger beliebig.

Diese Basisfunktion sind immer nur in Intervallen zwischen $(k+3)$ benachbarten Knotenpunkte ungleich Null und somit handelt es sich bei B-Splines um *lokale* Polynom-Splines, was auch deren größter Vorteil ist.

Um B-Splines in `R` verwenden zu können kann das Paket `splines` von Bates & Venables (2013) geladen werden. Dort gibt es die Funktion `bs()`, die genau diese Art von Spline generiert.

Literaturverzeichnis

Aitkin, M., Francis, B., Hinde, J. & Darnell, R. (2009). *Statistical Modelling in R.* Oxford Statistical Science Series.

Bantli, F. E. & Hallin, M. (1999). L_1 estimation in linear models with heterogeneous white noise. *Statistics and Probability Letters*, *45*, 305-315.

Barnett, W., Powell, J. L. & Tauchen, G. (1991). Nonparametrics and Semiparametrics Methods in Econometrics. *Metrika*, *41*, 309-311.

Bates, D. M. & Venables, W. N. (2013). *Regression Spline Functions and Classes.* R Documentation. Zugriff auf `http://127.0.0.1:20742/library/splines/html/splines-package.html`

Bofinger, E. (1975). Estimation of a density function using order statistics. *Australian Journal of Statistics*, *17*, 1-7.

Bosch, R., Ye, Y. & Woodworth, G. (1995). A convergent algorithm for quantile regression with smoothing splines. *Computational Statistics & Data Analysis*, *19*, 613-630.

Dantzig, G. (1963). *Linear Programming and Extensions.* Princeton University Press.

Davino, C., Furno, M. & Vistocco, D. (2013). *Quantile Regression.* Wiley Series in Probability and Statistics.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Annals of Statistics*, *2*, 267-277.

Donoho, D. & Huber, P. (1983). The Notion of Breakdown Point. In P. J. Bickel, Kjell Doksum & J. Hodges Jr (Hrsg.), *Eine Festschrift für Erich L. Lehmann* (S. 157-184).

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013). *Regression.* Springer.

Hall, P. & Sheater, S. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society, Series B*, *50*, 381-391.

Hampel, F. (1974). The influence function curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383-393.

Hampel, F., Ronchetti, E., Rousseeuw, P. & Stahel, W. (1968). *Robust Statistics: The Approach Based on Influence Functions.* Wiley.

He, X., Jureckova, J., Koenker, R. & Portnoy, S. (1990). Tail behavior of regression estimators and their breakdown points. *Econometrica*, *58*, 1195-1214.

Hendricks, W. & Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, *87*, 58-68.

Hjort, N. & Pollard, D. (1993). *Asymptotics for Minimizers of Convex Processes* (Bericht). University of Oslo and Yale University.

Huber, P. (1967). Behavior of maximum likelihood estimates under nonstandard conditions. In L. M. Le Cam & Jerzy Neyman (Hrsg.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Bd. 1, S. 221-233). University of California Press, Berkeley.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Annals of Statistics*, *26*, 755-770.

Knight, K. & Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics*, *28*, 1356-1378.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33-50.

Koenker, R., Ng, P. & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*, 673-680.

Lehmann, E. (1974). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day.

Machado, J. A. F. (1993). Robust model selection and M-estimation. *Econometric Theory*, *9*, 478-493.

Maronna, R., Martin, D. & Yohai, V. (2006). *Robust Statistics: Theory and Methods*. Wiley.

Mizera, I. & Wellner, J. (1998). Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables. *Annals of Statistics*, *26*, 672-691.

Parzen, E. (1979). Nonparametric statistical data modelling. *Journal of the American Statistical Association*, *74*, 105-121.

Rigby, R. A. & Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, *54*, 507-554.

Rousseeuw, P. & Leroy, A. (1987). *Robust Regression and Outlier Detection*. Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267-288.

Tukey, J. (1965). Which Part of the Sample Contains the Information. *Proceedings of the National Academy of Sciences*, *53*, 127-134.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Zhou, K. & Portnoy, S. (1996). Direct use of regression quantiles to construct confidence sets in linear models. *Annals of Statistics*, *24*, 287-306.