Quantilregression mit SAS

Thomas Bruckner
Universität Heidelberg
Im Neuenheimer Feld 305
69120 Heidelberg
bruckner@imbi.uni-heidelberg.de

Lorenz Uhlmann
Universität Heidelberg
Im Neuenheimer Feld 305
69120 Heidelberg
uhlmann@imbi.uni-heidelberg.de

Andreas Deckert
Universität Heidelberg
Im Neuenheimer Feld 364
69120 Heidelberg
a.deckert@uni-heidelberg.de

Zusammenfassung

Die klassische lineare Regressionsanalyse dient dazu, den linearen Zusammenhang einer unabhängigen Einflussgröße mit dem bedingten Erwartungswert der abhängigen Variablen zu modellieren. Wenn allerdings die Vorrausetzungen der Regressionsanalyse nicht erfüllt sind (z. B. bei Heteroskedastizität) oder man nicht an dem bedingten Erwartungswert interessiert ist, weil z.B. bei stark linksschiefen Verteilungen vor allem der Einfluss auf die Werte oberhalb eines bestimmten Quantils der abhängigen Variable untersucht werden soll, benötigt man andere Methoden, um entsprechende Zusammenhänge zu modellieren. Eine mögliche Methode dafür ist die Quantilregression [1].
Am Beispiel mehrerer Funktionsparameter der oberen Extremitäten bei gesunden Probanden (DASH-Score und Handkraft [2]) wird die Anwendung der Quantilregression anschaulich demonstriert und die Unterschiede zur klassischen linearen Regression erläutert. Es wird gezeigt, wie die Programmierung der Quantilregression in SAS mit den Prozeduren QUANTREG und SGPLOT erfolgt.

Schlüsselwörter: Dash-Score, Regression, Quantilregression, QUANTREG-Prozedur, SGPLOT-Prozedur

1 Einleitung

Die klassische lineare Regressionsanalyse ist ein unverzichtbares Tool im statistischen Werkzeugkasten. Sie dient dazu, den Zusammenhang zwischen einer (oder mehreren) erklärenden Variablen x auf eine interessierende Zielvariable y zu untersuchen. Die Anfänge der Regressionsanalyse gehen auf Sir Francis Galton (1822-1911) zurück, der den Zusammenhang zwischen der Körpergröße von Kindern und dem Durchschnitt der Größe beider Elternteile untersuchte. Die mathematische Weiterentwicklung wurde unter anderem von Karl Pearson (1857-1936), Francis Edgeworth (1845-1926) und George Yule (1871-1951) betrieben.

Das einfachste Regressionsmodell beschreibt einen linearen Zusammenhang zwischen der Zielvariablen und einer erklärenden Variablen, wobei zufällige Schwankungen in einer Störgröße zusammengefasst werden:

$$y=\beta_0 + \beta_1 x + \varepsilon,$$

wobei die Parameter β_0 und β_1 (Achsenabschnitt und Steigung der Regressionsgerade) mit der Methode der kleinsten Quadrate aus einer Stichprobe (x_i, y_i), i=1,…, n, geschätzt werden, deren mathematische Formulierung auf den Mathematiker Adrien Legrendre (1752-1823) zurückgeht.
Ein wesentliches Merkmal der Regressionsanalyse ist, dass der Zusammenhang zwischen der Zielgröße y und der (oder den) erklärenden Variable(n) nicht exakt als Funktion f(x) angegeben werden kann, sondern durch zufällige Störgrößen „verrauscht" ist. Die Zielvariable y ist somit eine Zufallsvariable deren Verteilung von der oder den erklärenden Variablen abhängt.
Ein Hauptziel der Regressionsanalyse besteht darin, den Einfluss der erklärenden Variablen auf den Mittelwert der Zielgröße zu untersuchen. Es wird also der (bedingte) Erwartungswert $E(y|x_1,\ldots,x_n)$ von y in Abhängigkeit der sogenannten Kovariablen x_i modelliert.

$$E(y|x_1,\ldots,x_n) = f(x_1,\ldots,x_n)$$
$$y= E(y|x_1,\ldots,x_n) + \varepsilon= f(x_1,\ldots,x_n) + \varepsilon$$

Die klassische Regressionsanalyse fokussiert also auf den Erwartungswert. Allerdings sind die Modellannahmen nicht immer gegeben, z.B. ist die Annahme der Homoskedastizität oft verletzt, auch schiefe Verteilungen sind in der Praxis nicht selten. Auch der Fokus auf den Erwartungswert kann Informationen, die in den erklärenden Variablen vorhanden sind, nicht immer genügend beachten. Diese Informationen kann man nur ausschöpfen, wenn man die gesamte Verteilung der erklärenden Variablen betrachtet.
Die Methode der Quantilregression erweitert die Modellierung der bedingten Erwartungswerte zu einem Regressionsmodell für die bedingten Quantile der Zielvariablen. Sie wurde als erstes von Koenker und Bassett 1978 veröffentlicht [1].
Zur Einführung folgende Definitionen:

Das p-te Quantil $Q(p)$ einer Verteilungsfunktion F ist das Minimum aller Werte x derart, dass $F(x) \geq p$. Die Funktion $Q(p)$ (als eine Funktion von p, $0<p<1$) heißt Quantilfunktion von F.

Sei $x_1,\ldots, x_n$ eine zufällige Stichprobe vom Umfang n. Wir definieren das p-te Stichprobenquantil $\hat{Q}(p)$ als das p-te Quantil der empirischen Verteilungsfunktion $\hat{F}$, $\hat{Q}(p)= Q(p)(\hat{F})$. Die zugehörige Quantilfunktion heißt Stichprobenquantilfunktion.

Wenn $x_1,\ldots, x_n$ eine Stichprobe vom Umfang n ist, dann ist $x_{(1)},\ldots, x_{(n)}$ die geordnete Stichprobe mit $x_{(1)} \leq x_{(2)} \leq ,\ldots, \leq x_{(n)}$. Für eine Stichprobe vom Umfang n ist das (k/n)te Stichprobenquantil gerade $x_{(k)}$.

Die Quantilfunktion ist also definiert als

$$Q(p)=\inf\{y:F(y) \geq p\} \text{ mit } 0<p<1.$$

Es ist bekannt, dass der Stichprobenmedian das Minimum der Summe der absoluten Differenzen ist

$$\text{Median}=\min_q \sum_{i=1}^{n} |y_i - q|$$

mit derselben Argumentation kann man das allgemeine α Stichprobenquantil formulieren als die Lösung eines Optimierungsproblems

$$q(\alpha)= \min_q \sum_{i=1}^{n} \varphi_\alpha (y_i - q)$$

wobei $\varphi_\alpha(z)=z(\alpha - I(z<0))$, $0< \alpha<1$. I(.) ist dabei die Indikatorfunktion.

Die bedingte lineare Quantilfunktion $Q(\alpha|X=x)=x'\beta(\alpha)$ kann durch Lösung der Gleichung

$$\widehat{\beta}(\alpha) = \text{argmin}_{\beta \epsilon R^p} \sum_{i=1}^{n} \varphi_\alpha (y_i - x_i'\beta)$$

für alle α aus (0,1) geschätzt werden. $\widehat{\beta}(\alpha)$ heißt das α-te Regressionsquantil. Obige Gleichung kann z.B. mit Methoden der linearen Optimierung gelöst werden. Weitere Details finden sich unter [3].

2 Material und Methoden

Im klinischen Bereich, vor allem bei der Behandlung von Erkrankungen, welche die Funktion der Hand und Handgelenke betreffen, werden die Parameter Handbeweglichkeit und Kraft gemessen, um Behandlungsergebnisse zu objektivieren. Auch ein subjektiver Beschwerdefragebogen, der DASH-Fragebogen (Disabilities of the Arm, Shoulder and Hand), wird oft zur Überprüfung der Behandlungsergebnisse benutzt. Aus diesem Fragebogen wird ein Score berechnet (im Folgenden DASH-Score genannt), der bei Beschwerdefreiheit den Wert Null annimmt und bei maximalen Beschwerden den Wert 100 annehmen kann, siehe Anhang. Da besonders die arbeitende Bevölkerung häufig von Handgelenksbeschwerden betroffen ist (und Normdaten aus Deutschland fehlen) sollte eine Studie Normdaten zu Beweglichkeit und DASH-Score liefern. Dazu wurden an einem beschwerdefreien Kollektiv (387 Männer und 363 Frauen) unter anderem diese Parameter gemessen. Allerdings weiß man, dass der DASH-Score stark al-

tersabhängig ist und eine große Heteroskedastizität mit wachsendem Alter aufweist (siehe Abbildung 1).

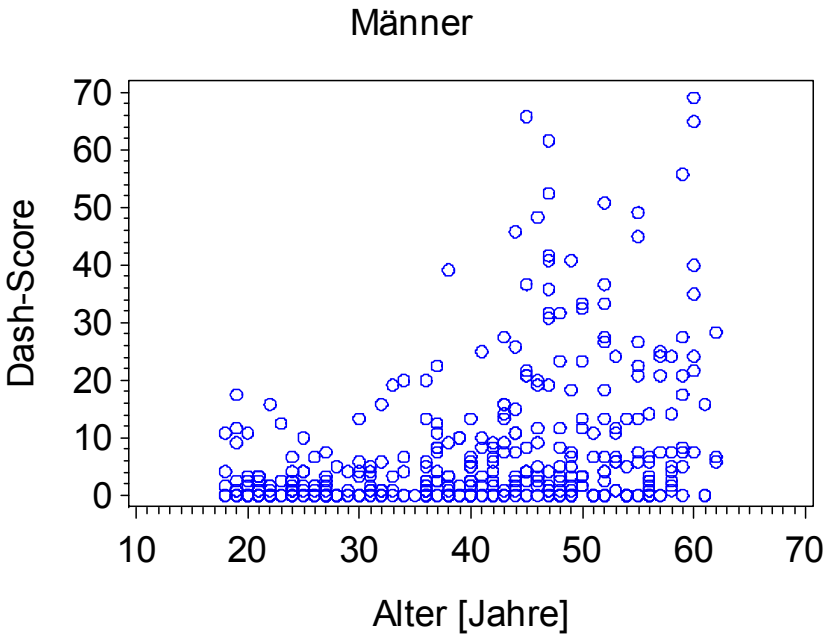

Abbildung 1: DASH-Score in Abhängigkeit vom Alter (am Beispiel Männer)

Auch die Handkraft nimmt mit dem Alter zu, erreicht ihr Maximum etwa im Alter von 45 und nimmt mit zunehmendem Alter wieder ab (siehe Abbildung 2). Sie zeigt eine näherungsweise negative Parabel.

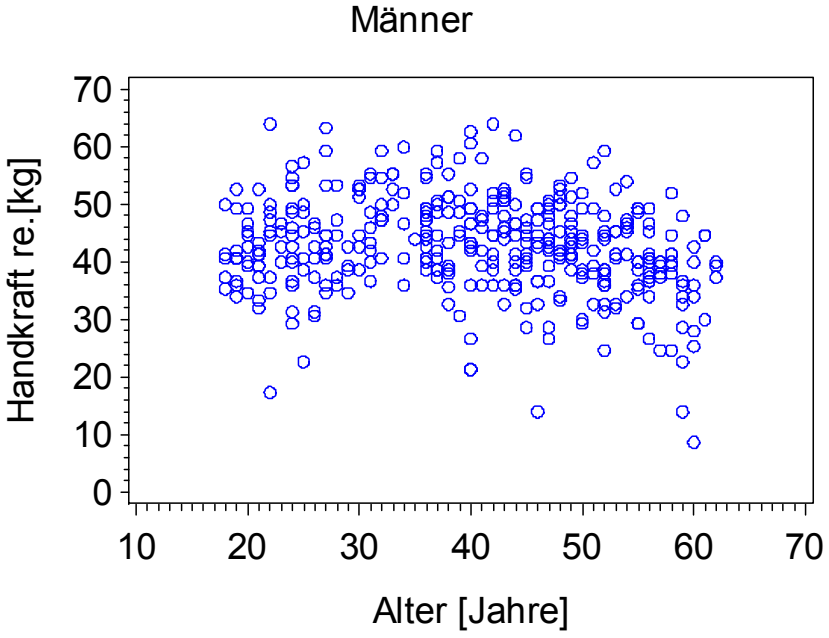

Abbildung 2: Handkraft in Abhängigkeit vom Alter (am Beispiel Männer)

In beiden Beispielen sieht man, dass wesentliche Voraussetzungen für eine lineare Regressionsanalyse verletzt sind. Deshalb wurde als nichtparametrische Alternative die Quantilregression gewählt, um Normbereiche für dieses Kollektiv zu bilden. Dazu wurden die SAS-Prozeduren QUANTREG und SGPLOT verwendet.

3 Ergebnisse

3.1 DASH-Score

Die Werte des DASH-Scores und damit auch Handgelenksbeschwerden nehmen mit dem Alter zu, allerdings auch die Streuung. Deshalb wurden als Referenzwerte die 95% und 97,5% Quantilfunktionen aus der empirischen Verteilung geschätzt und als Illustration der Veränderung über das Alter zusätzlich die mediane Veränderung gezeigt.

Die folgenden SAS-Programmzeilen berechnen die Quantilfunktionen und erzeugen die Graphiken der Abbildung 3.

```
ods listing gpath="\studien\klum\output" image_dpi=200 ;
ods graphics on / imagename='men_dashakt';

title 'Men';
proc quantreg data=men ci=sparsity;
      model dash_aktivitaet=alter
       / quantile=0.5,0.95,0.975 plot=fitplot;
        output out=outp pred=p  /columnwise;
run;
ods graphics off;
```

Als Graphikschiene wurde die ODS Graphik gewählt, um die graphischen Möglichkeiten der Prozedur Quantreg auszunutzen. Nach Aufruf der Prozedur wird das Modell spezifiziert, und die zu schätzenden Quantile angegeben. Die Plot-Option erzeugt die unten stehenden Graphiken. Zusätzlich gibt die Prozedur die gefundenen Schätzwerte im Output-Fenster (oder optional in einer mit der ODS angegebenen Datei) aus. Unten sieht man den Output am Beispiel des 0.5 Quantils (d.h. des Medians).

Quantile and Objective Function

Quantile	0.5
Objective Function	1323.6378
Predicted Value at Mean	3.4192

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	-3.0449	0.7352	-4.4903	-1.5994	-4.14	<.0001
alter	1	0.1603	0.0310	0.0994	0.2212	5.17	<.0001

Die folgende Tabelle zeigt die Ergebnisse für den DASH Score:

Tabelle 1: Geradengleichungen nach Geschlecht

Quantil	Geradengleichung (Männer)	Geradengleichung (Frauen)
Median	0,16*Alter – 3,05	0,29*Alter – 4,46
95% Perentil	1,06*Alter – 10,9	0,89*Alter – 1,0
97,5% Perzentil	1,29*Alter-12,6	1,06*Alter – 0,38

Die Abbildung 3 zeigt die Veränderungen des DASH-Scores über das Alter für Frauen und Männer am Beispiel des rechten Arms.

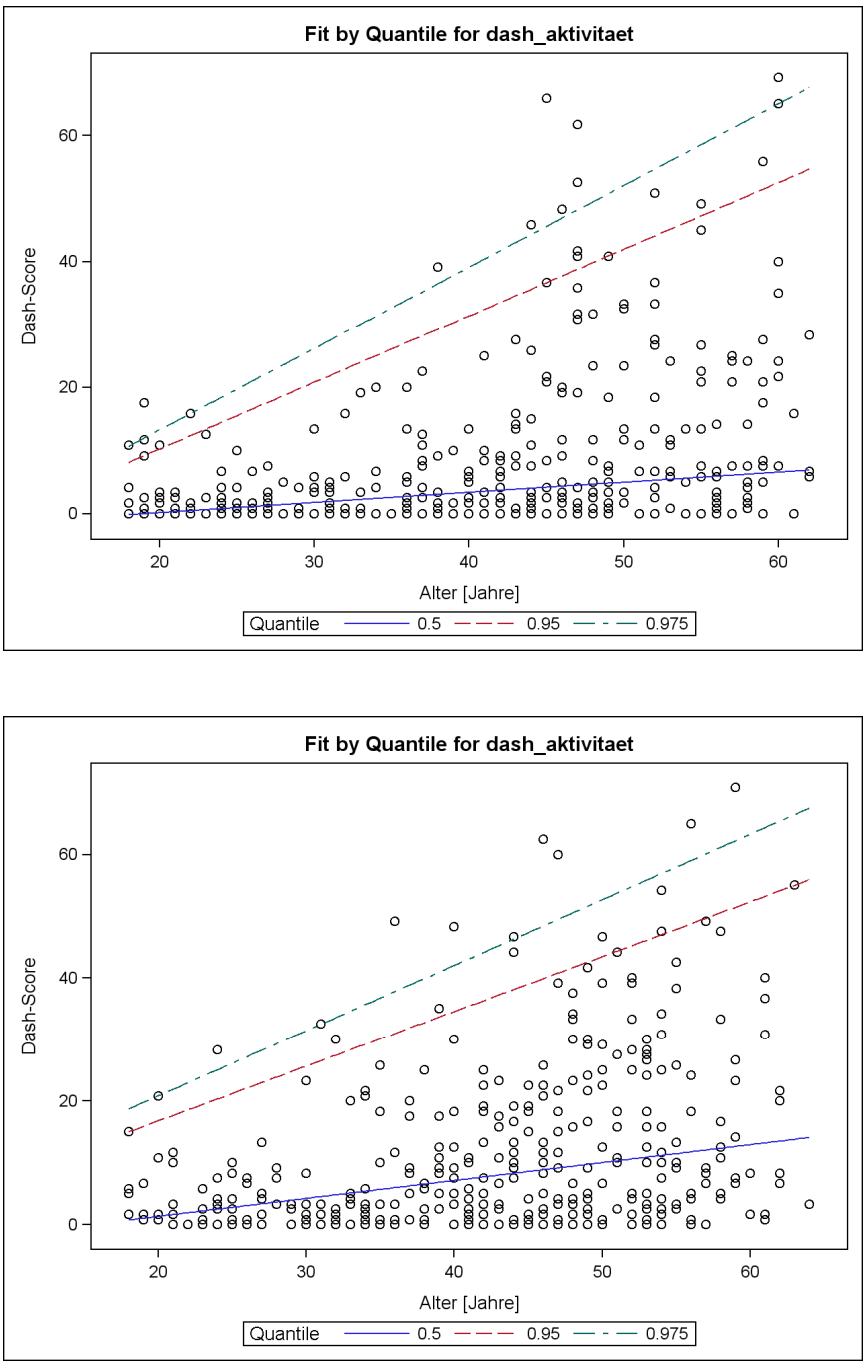

Abbildung 3: Veränderungen des DASH-Scores mit zunehmendem Alter (oben Männer, unten Frauen).

3.2 Handkraft

Die Handkraft wird mit Hilfe einer speziellen hydraulischen Presse gemessen, dem sog. Jamar-Hand Dynamometer. Die Stärke der Handkraft mit steigendem Alter folgt näherungsweise einer quadratischen Funktion, deren Maximum etwa bei 45 Jahren liegt.
Die folgenden SAS-Programmzeilen berechnen die Quantilfunktionen und erzeugen die Graphiken der Abbildung 4. Da die Prozedur QUANTREG nur lineare Funktionen graphisch darstellen kann, wurde zur Visualisierung der Ergebnisse die Prozedur SGPLOT herangezogen.

```
proc quantreg data=men ci=sparsity;
      model jamarre=alter alter_2
       / quantile=0.025 0.05 0.5,0.95,0.975 ;* plot=fitplot;
        output out=outp pred=p  /columnwise;
run;

proc sort data=outp;
    by alter quantile;
run;


proc sgplot data=outp;
      title 'Percentile (Männer)';
      yaxis label='Handkraft [kg] re.' min=0 max=70
            values=(0 10 20 30 40 50 60 70);
      xaxis label='Alter [Jahre]' min=10 max=70
            values=(10 20 30 40 50 60 70);
      scatter x=alter y=jamarre / markerattrs=(size=4);
      series  x=alter y=p/group=QUANTILE;
run;
```

Auch hier wird ein Ausschnitt des Outputs (am Beispiel des 95% Perzentils) dargestellt:

```
    Quantile and Objective Function

Quantile                          0.95
Objective Function              295.0732
Predicted Value at Mean          54.9055


                      Parameter Estimates

                     Standard   95% Confidence
Parameter DF Estimate    Error       Limits      t Value Pr > |t|

Intercept  1  31.1991   6.7825  17.8636  44.5347    4.60   <.0001
alter      1   1.5411   0.4122   0.7306   2.3516    3.74   0.0002
alter_2    1  -0.0216   0.0053  -0.0320  -0.0113   -4.11   <.0001
```

Abbildung 4 veranschaulicht den Verlauf der Handkraft mit steigendem Alter.

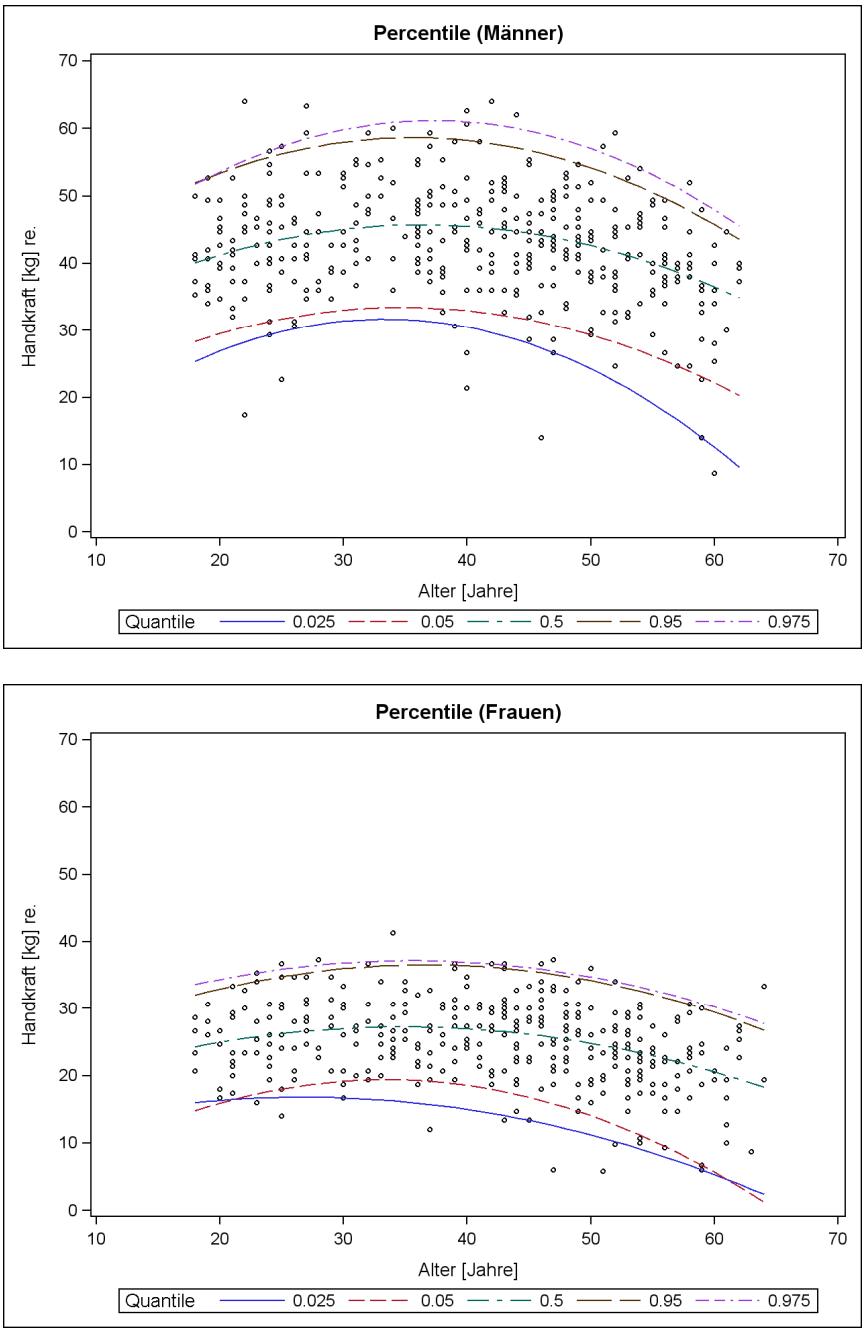

Abbildung 4: Veränderungen der Handkraft mit zunehmendem Alter (oben Männer, unten Frauen).

Auch hier sind die Ergebnisse der Quantilregression tabellarisch zusammengefasst.

Tabelle 2: Gleichungen 2. Grades nach Geschlecht

Quantil	Gleichung (Männer)	Gleichung (Frauen)
2,5% Perzentil	-0,027*Alter²+1,79*Alter+1,83	-0,010*Alter²+0,56*Alter+9,29
5% Perzentil	-0,018*Alter²+1,25*Alter+11,73	-0,019*Alter²+1,30*Alter-2,28
Median	-0,017*Alter²+1,21*Alter+23,65	-0,011*Alter²+0,73*Alter+14,53
95% Perentil	-0,022*Alter²+1,54*Alter+31,20	-0,013*Alter²+0,96*Alter+19,00
97,5% Perzentil	-0,026*Alter²+1,90*Alter+25,77	-0,012*Alter²+0,83*Alter+22,50

4 Diskussion

Die Quantilregression erweist sich als nützliche Alternative zur linearen Regressionsanalyse, insbesondere, wenn deren Vorrausetzungen verletzt sind oder man andere Bereiche der Verteilungsfunktion als den Erwartungswert im Fokus hat. Die Umsetzung der Methode in SAS ist recht benutzerfreundlich, das gewählte Modell kann analog zur Syntax in PROC REG angegeben werden. Auch die graphischen Möglichkeiten der Prozedur sind bei der Schätzung linearer Quantilfunktionen ausreichend, bei komplexen Graphiken muss man die Ergebnisse als SAS-Datei abspeichern und mit der Prozedur SGPLOT weiterbearbeiten. Auch diese erweist sich als recht benutzerfreundlich, allerdings kommt man bei ihr, wie bei allen Graphik-Prozeduren von SAS, schnell ins Suchen, denn die Fülle der Möglichkeiten der Prozedur SGPLOT schlägt sich in ca. 100 Seiten Handbuch nieder.

Literatur

[1] Koenker, R., Bassett, G.W. (1978) Regression Quantiles. Econometrica, 46, 33-50

[2] Klum, M., Wolf, M.B., Hahn, P., Leclère, F.M., Bruckner, T., Unglaub, F. (2012). Normative Data on Wrist Function. J Hand Surg Am 132(12):1807-11

[3] Chen, C. An Introduction to Quantile Regression and the QUANTREG Procedure. SUGI 30 (2005), Paper 213-30

Anhang A: Zusammensetzung des DASH-Scores:

Der DASH-Fragebogen besteht aus 30 Fragen mit möglichen Werten von 1-5, die summiert werden und den Rohwert bilden. Dieser Rohwert hat eine Spannweite von 30 bis 150 Punkten. Die Rohwerte werden in eine Skala von 0-100 konvertiert, diese bilden den DASH-Score, wobei der Wert 0 keine Einschränkung und der Wert 100 sehr hohe Einschränkung bedeutet.

DASH=(Rohwert - 30) / 1.2

Bei lückenhaft beantworteten Fragebögen gilt es Folgendes zu beachten: Wurden weniger als 10% (3 Fragen) nicht beantwortet, so darf der Mittelwert aller anderen Fragen für den fehlenden Wert bzw. die fehlenden Werte verwendet werden. Sind 3 oder mehr Fragen nicht beantwortet worden, so darf dieser Fragebogen nicht verwendet werden.