Ludwig Maximilian Universität

Institut für Statistik

Bachelor-Arbeit

Expektile Regression

Autor:
Barbara Habereder

Betreuer:
Prof. Dr. Göran Kauermann

Februar 2015

Inhaltsverzeichnis

Abstract

Quantile werden durch Optimierung einer asymmetrisch gewichtetetn L_1 Norm berechnet. Es wird die Summe der Absolutbeträge der Residuen minimiert. Im Gegensatz hierzu werden Expektile durch Optimierung einer asymmetrisch gewichteten L_2 Norm bestimmt. Bei der Berechnung der Expektile wird die Quadratsumme der Residuen zur Minimierung verwendet. Der zusätzlich eingeführte Asymmetrieparameter wird über eine Gewichtsfunktion in die Schätzung mit einbezogen. Die Berechnung ist einfach, denn ein globales Minimum lässt sich durch die gewichtete Regression in wenigen Schritten bestimmen. Die asymmetrisch gewichtete Kleinste-Quadrate Methode ermöglicht mit Kombination von P-Splines die Schätzung von geglätteten Expektilkurven.

1 Einleitung

Ein großes Aufgabengebiet der Statistik ist es Zusammenhänge und Einflüsse von unterschiedlichen Variablen zu erklären. Bei der Regression wird eine abhängige Variable, auch Zielgröße genannt, durch eine oder mehrere unabhängige Kovariablen mittels eines mathematisch - statistischen Prozesses erklärt. Doch nicht immer liegt ein linearer Zusammenhang zwischen der Zielgröße und der erklärenden Größe vor. In einem solchen Fall ist es nicht ausreichend, wenn die Kovariable(n) linear in das Modell eingehen. Aus diesem Grund findet die semiparametrische Regression häufig Anwendung. Dort werden die Kovariablen mittels einer unbekannten Funktion in das Modell aufgenommen und die Funktion geschätzt. Am häufigsten findet die Mittelwertregression Anwendung, hierbei ist die Varianzhomogenität eine der Voraussetzungen, wobei sich hier die Streuung für alle Ausprägungen gleich verhält. Ist diese Annahme verletzt, kann die normale Regression nicht problemlos durchgeführt werden. Ebenso will man unter Umständen mehr Informationen, als nur den Mittelwert, aus den Daten erhalten, z.B. Charakteristiken der Verteilung, Verhalten an den Rändern, Informationen zur Schiefe und Symmetrie der Verteilung. Die Expektilregression ist eine Erweiterung der Mittelwertregression, die es ermöglicht solche Informationen zu gewinnen. Hierbei wird ein Asymmetrieparameter mit in die Gleichung der Kleinste-Quadrate-Schätzung aufgenommen. Dies führt dann zur Kleinsten-Asymmetrischen-Gewichteten-Quadrate-Methode (LAWS). Ein Schwachpunkt der Expektile ist die mangelnde intuitive Interpretierbarkeit. Jedoch sind die Expektile über eine bijektive Funktion mit den Quantilen verknüpft. Die Quantilsregression als eine Art der Medianregression bietet eine gute Interpretierbarkeit. Diese versucht man zu nutzen, indem man Quantile aus Expektilen berechnet. Hierzu wurden bereits Methoden entwickelt, um die ermittelten Quantile für die Interpretation zu nutzen.

In dieser Arbeit wird zuerst die semiparametrische Regression kurz vorgestellt, da die Expektilregression in semiparametrischen Modellen gute Ergebnisse erziehlt. Nachfolgend werden die Expektile eingeführt und erklärt. In Kapitel 4 wird die Regression mit Expektilen einschließlich der Wahl des optimalen Glätteparameters ausgeführt und das Problem kreuzender Kurven mit Lösungsvorschlägen vorgestellt. Das Vorgehen Quantile aus Expektilen zu berechnen und die Berechnung von Konfidenzintervallen sind auch Teil dieses Kapitels. Zuletzt wird die Methodik am Beispiel der Münchner Mietspiegel Daten aus dem Jahr 2013 beispielhaft angewandt. Hierbei wird die Nettomiete pro Quadratmeter als Zielgröße u.a. durch die Kovariablen Wohnfläche und Baujahr erklärt.

2 Semiparametrische Regression

Bei der Mittelwertregression wird in den meisten Fällen ein parametrisches Modell der Form

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \Leftrightarrow \eta = \boldsymbol{x}^T \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

mit $i = 1, ..., n$ Beobachtungen und $\boldsymbol{y}$ als unabhängiger Zielgrösse verwendet. Hiermit wird ein Intercept (β_0), sowie je ein eigener Parameter für jede Kovariable (x_1 bis x_p) geschätzt.
Da der lineare Einfluss nicht immer vorhanden ist, bedarf es Alternativen. Ein anderer Ansatz zur Schätzung ist die Betrachtung einer allgemeinen, unbekannten funktionalen Form des Prädiktors. Hierbei handelt es sich um die semiparametrische Regression. Die Bestimmungsgleichung sieht hierbei wie folgt aus: [1]

$$\boldsymbol{y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta_1} + f_2(\boldsymbol{z}_2) + ... + f_r(\boldsymbol{z}_r) + \boldsymbol{\epsilon}$$

$$\Leftrightarrow \eta = \beta_0 + \boldsymbol{X}\boldsymbol{\beta_1} + \sum_{j=2}^{r} f_j(\boldsymbol{z}_i) + \boldsymbol{\epsilon}$$

Weiterhin ist ein Intercept im Modell enthalten. Die parametrischen Effekte sind analog dem gewöhnlichen Modell in der Matrix $\boldsymbol{X}$ erfasst. Mittels der Funktionen f_j werden die unbekannten Effekte mit in das Modell integriert. Dies können beispielsweise nichtlineare, räumliche oder zufällige Effekt sein, vgl. [Sobotka, 2012, Kapitel 1]. Zusammen mit jeder Funktion f_j wird ein Bestrafungsterm $\lambda_j \; pen(f_j)$, welcher spezielle Eigenschaften erzwingen soll, mit berechnet. $\lambda_j \geq 0$ ist der jeweils zugehörige Glätteparameter über den der Einfluss der Bestrafung vorgenommen wird. Ein breites Spektrum an Funktionstypen erhält man, wenn man für die Schätzung der unbekannten Funktionen folgende Annahmen befolgt: Die Funktionen f_j werden durch K Basisfunktionen $f_j(z) = \sum_{k=1}^{K} \beta_{jk} B_k(z)$ approximiert, wobei $B_k(z)$ die Basisfunktion und β_{jk} den jeweiligen zugehörigen Basiskoeffizienten angibt. Die Bestrafung wird quadratisch in den Vektor der Basiskoeffizienten $\boldsymbol{\beta_j} = (\beta_{j2}, ..., \beta_{jK})^T$, z.B. als $pen(f_j) = \boldsymbol{\beta_j^T K_j \beta_j}$ mit Bestrafungsmatrix $\boldsymbol{K_j}$ aufgenommen. Die Matrix $\boldsymbol{K}$ wird so gewählt, dass die gewünschten Regularisierungseigenschaften erreicht werden, vgl. [Sobotka, 2012, Kapitel 2].
Durch Zusammenfassen zu einer Basis-Matrix $f_j = \boldsymbol{Z}_j \boldsymbol{\beta}_j$ lässt sich obige Gleichung umschreiben und wie folgt darstellen, vgl. [Sobotka, 2012, Kapitel1]:

$$\boldsymbol{y} = (\mathbb{1}, \boldsymbol{X}, \boldsymbol{Z}_2, .., \boldsymbol{Z}_r)(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_r)^T + \boldsymbol{\epsilon}$$

In der nachfolgenden Tabelle (1) ist eine Übersicht der jeweiligen Basismatrizen $\boldsymbol{Z}$ und Bestrafungsmatrizen $\boldsymbol{K}$ aufgelistet. Die P-Splines werden in Kapitel 4.2 näher

[1] im Fall von P-Splines gilt: $f(z) = f(x)$

Effekt	Baismatrix $\boldsymbol{Z}$	Bestrafungsmatrix $\boldsymbol{K}$
parametrisch	$\boldsymbol{X}$	0
P-Splines	Splines geschätzt anhand Beobachtungen	Quadrierte Matrix der Differenzen 2.Ordnung
Bivariate Splines	ausgewertetes Splines Tensor Produkt	$\boldsymbol{K_1} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{K_2}$
Markov random field	Indikator Matrix	Matrix mit Nachbarschaftsbezug
Radial Basis	ermittelte Distanz zwischen den Beobachtungen	Distanz zwischen den Knoten
Kriging	basierend auf emp. Korrelation	Korrelationsmatrix
Zufall	Indikatormatrix	$\boldsymbol{I}$

Tabelle 1: Mögliche Effekte der Kovariablen in einem semiparametrischen Modell, übernommen aus [Sobotka, 2012, Kapitel1].

erläutert.

3 Expektile

3.1 Einführung

Da in manchen Fällen nicht nur der Mittelwert von Interesse ist, sondern mehr Informationen aus den Daten gewonnen werden sollen, führten Newey und Powell eine gewichtete Methode der Kleinsten-Quadrate-Schätzung als Alternative zu Quantilen ein,vgl. [Newey and Powell, 1987]. Statt dem Absolutbetrag wird hier ein quadratischer Term mit in das Modell aufgenommen. Für diese neu definierte Formel wählten sie den Namen **Expektile**, vgl. [Schnabel, 2011, Kapitel 1]. Mit diesem Verfahren kann man eine Expektilkurve durch iterative Gewichtung der kleinsten-Quadrate fitten. Ebenso lässt sich eine Verteilung sowohl durch Expektile als auch durch Quantile eindeutig beschreiben, vgl.[Schnabel, 2011, Kapitel 1]. Die Quantile haben den Vorteil, dass ihnen eine intuitive Interpretation zugrunde liegt. Bei den Expektilen ist dies Interpretation ein Nachteil, hier kann man nicht von einer eingängigen Aussage der Kurven sprechen. Es wird aber nachfolgend eine Möglichkeit der Interpretation vorgestellt. Andererseits sind die Expektile jedoch effizienter als Quantile und Newey und Powell nannten in ihrem Artikel drei weitere wichtige Vorteile:

1. Die einfache Berechnung stellt einen großen Vorteil dar.

2. Der Schätzer ist effizienter gegenüber der Quantilregression, denn aus den Daten wird mehr Information mit einbezogen.

3. Die Kovarianzmatrix kann berechnet werden, ohne vorheriger Bestimmung der genauen Dichte der Daten.

Die Expektile finden bereits in einigen Bereichen stetig Anwendung, z.B. im Fall von asymmetrische Einflüssen, denn diese werden hierbei durch die unterschiedlichen Gewichte für positive und negative Residuen berücksichtigt. Ebenso im Bereich von Risikomaßen bei Finanzanlagen und im Fall von Heteroskedastizität, vgl. [Schulze-Waltrup et al., 2014a].
In den nachfolgenden Kapiteln werden die Expektile theoretisch und praktisch eingeführt und anschließend die Regression mit Expektilen vorgestellt.

3.2 Anwendungsbezogene Einführung der Expektile

Bei der bekannten, einfachen kleinste-Quadrate Schätzung (ordinary least squares, OLS) liegt das folgende quadratische Minimierungsproblem zugrunde.

$$S_{OLS} = \sum_{i=1}^{n}(y_i - \mu_i)^2 \Leftrightarrow \hat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)^2 \tag{3.1}$$

Wobei hier n die Anzahl der Beobachtungen, y_i die unabhängige Zielgrösse und μ_i den erwarteten geschätzten Wert des Modells darstellen. Oft ist das hinterlegte Modell als univariates lineares Modell mit $\mu_i = \beta_0 + \beta_1 x_i$ gegeben.
Da aber nicht immer, wie bereits erwähnt, der Mittelwert von Interesse ist, sondern z.B. die bedingte Verteilung des Response, wurden weitere statistische Methoden entwickelt. Um mehr Informationen zu erhalten, wird in der Literatur die Quantilsregression als generalisiertes Modell der Mittelwertregression genannt. Bei der Quantilregression werden, anders als bei der OLS die Absolutbeträge der Residuen minimiert. In Formelnotation lässt sich dies darstellen als:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - (\boldsymbol{X}\boldsymbol{\beta})_i| \tag{3.2}$$

Zusätzlich zur Minimierung der Absolutbeträge wurde von Koenker und Bassett noch folgende Gewichtsfunktion eingeführt, vgl. [Koenker and Bassett Jr, 1978]

$$\omega_i(\alpha) = \begin{cases} \alpha & \text{wenn } y_i > \mu_i(\alpha) \\ 1-\alpha & \text{wenn } y_i \leq \mu_i(\alpha). \end{cases} \tag{3.3}$$

Das heißt, liegt der Datenpunkt oberhalb von α, wird er mit α gewichtet. Für den Fall, dass er unterhalb liegt, wird ihm das Gewicht $1-\alpha$ zugeordnet, vgl.[Schnabel, 2011, Zusammenfassung]. Über diese Funktion fließt der Asymmetrieparameter $\alpha \in (0,1)$ in das Modell ein. Dieser bewirkt, dass positive und negative Residuen unterschiedliche Gewichte erhalten. Für ein festes, vom Anwender gewähltes α ergibt sich

die Schätzung eines p-Quantils aus:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \arg\min_{\beta} \sum_{i=1}^{n} \omega_i(\alpha)|y_i - (\boldsymbol{X}\boldsymbol{\beta})_i| \tag{3.4}$$

Allerdings besteht bei der Anwendung von (3.4) die Problematik, dass die Betragsfunktion nicht differenzierbar in Null ist und somit eine Lösung nur numerisch bestimmt werden kann. Dies stellt einen großen Nachteil zu anderen Verfahren dar. Kombiniert man hingegen das Verfahren OLS mit der definierten Gewichtsfunktion (3.3), so erhält man die gewichtete Alternative zu OLS: die kleinste asymmetrisch gewichtete Quadrate-Methode (Least asymmetrically weighted squares, LAWS) und die Bestimmungsgleichung ergibt sich nun wie folgt, vgl. [Sobotka, 2012, Kapitel 1]:

$$\begin{aligned} S_{LAWS} &= \sum_{i=1}^{n} \omega_i(\alpha)(y_i - \mu_i(\alpha))^2 \\ \Leftrightarrow \hat{\boldsymbol{\beta}}_{\alpha} &= \arg\min_{\beta} \sum_{i=1}^{n} \omega_i(\alpha)(y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)^2 \end{aligned} \tag{3.5}$$

Das Ergebnis des oben genannten Minimierungsproblems liefert das α-Expektil. $\mu_i(\alpha)$ entspricht dem population expectile für die unterschiedlichen Werte des Asymmetrieparameters $\alpha \in (0, 1)$. [Schnabel, 2011, Kapitel 1] Die Gleichung ist analytisch lösbar und die bedingte Verteilung des Response kann damit vollständig geschätzt werden. Dies verdeutlicht nochmals den rechnerischen Vorteil der Expektile gegenüber den Quantilen. Weiterhin bleibt noch die schlechte Interpretierbarkeit der Expektile bestehen. Den Expektilen liegt leider keine intuitive Interpretation zugrunde, während die Quantile die Inverse der Verteilungsfunktion sind, vgl. [Schulze-Waltrup et al., 2014a]. „Bedingt durch die Konstruktion des Gewichtsvektors wird also ein Expektil durch ein iteratives Verfahren ermittelt, in dem in jedem Schritt die Gewichte neu bestimmt werden. Anschließend wird das Expektil selbst neu berechnet. Diese Schritte werden bis zur Konvergenz wiederholt“, [Schnabel, 2011, Zusammenfassung]. Das iterative Verfahren ist notwendig, da die Gewichte vom Parametervektor abhängen und der Parameterverktor von den Gewichten. Als Startwert dient $\alpha = 0.5$. Als Ergebnis des Verfahrens erhält man eine Expektilkurve zum vorher festgelegten Asymmetrieparameter α. Die OLS-Methode stellt einen Sonderfall von LAWS für den Asymmetrieparameter $\alpha = 0.5$ dar. In der nachfolgenden Abbildung 1 ist das $0.2-$Expektil dargestellt. Die Abstände der Datenpunkte unterhalb der Kurve zur Kurve sind rot und die Abstände der Datenpunkte oberhalb der Kurve zur Kurve sind in schwarz eingezeichnet. Es kann folgende Aussage bzw. Interpretation des 0.2-Expektils angegeben werden: 20% der mittleren Abstände zwischen m_α und y sind durch die Masse unterhalb der Kurve bestimmt. Demnach liegen hier 80% (100% − 20%) der Masse überhalb der Kurve. Die Formel, die der

Interpretation zugrundeliegt, ist in der Grafik angeben, vgl. [Schulze-Waltrup et al., 2014b].

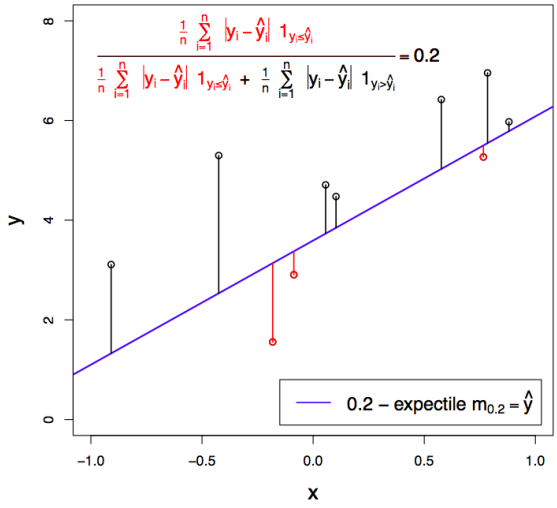


Abbildung 1: Interpretation des 0.2-Expektils. Grafik übernommen aus [Schulze-Waltrup et al., 2014b]

Mit den bisherigen Formeln 3.3 und 3.5 werden nur gerade, starre Expektilkurven modelliert. In der Praxis sind aber deutlich flexiblere Kurven nötig. Daher wird die Methode LAWS häufig beispielsweise mit P-splines, oder anderen Methoden zur Anpassung an die Daten kombiniert. Diese Vorgehensweise wurde von Eilers und Marx vorgestellt, vgl. [Eilers and Marx, 1996] Die Methode der P-Splines kombiniert die Methode der B-Splines und eine Bestrafung, um die Glattheit und Anpassung der Kurven an die Daten zu verbessern. Der Bestrafungsparameter λ reguliert dabei die Glätte der Kurve. Durch die Kombination von P-Splines und LAWS ist es möglich flexible Kurven in jeder Region der Daten zu schätzen. Mehr zur P-Spline Anwendung ist in Kapitel 4.2 zu lesen.
Wie der optimale Glätteparameter oder Bestrafungsparameter λ gewählt wird, ist in Kapitel 4.3 näher ausgeführt.

3.3 Theoretische Betrachtungsweise der Expektile

Im nächsten Abschnitt werden die Expektile theoretisch betrachtet und hergeleitet. Für jede gegebene Verteilung mit Verteilungsfunktion F und endlichen Erwartungswerten können alle theoretischen Expektile bestimmt werden, vgl. [Sobotka et al., 2014]. Wie bereits erwähnt, ist es möglich Expektile für jede beliebige Verteilung theroetisch zu ermitteln. Ein analytisches Ergebnis ist jedoch nur für ausgewählte Verteilungsfamilien möglich. Eine numerische Näherung kann jedoch immer gegeben werden. Umgekehrt lässt sich eine Verteilung aber komplett durch ihre Expektile beschreiben. Für eine Wahrscheinlichkeits- oder Dichtefunktion $f(x)$ und ihre Ver-

teilungsfunktion $F(x)$, lässt sich z.B. die kumulative Dichtefunktion und die partial (first) moment function $G(x)$ wie folgt definieren, vgl. [Schnabel, 2011, Kapitel 1]:

$$F(x) = \int_{-\infty}^{x} f(u)du \text{ and } G(x) = \int_{-\infty}^{x} uf(u)du. \tag{3.6}$$

Das theoretische Expektil wird mit m_α bezeichnet, wobei $\alpha \in (0, 1)$ wieder der Asymmetrieparameter ist.
m_α ist definiert als:

$$\arg\min_{m_\alpha} M = (1-\alpha) \int_{-\infty}^{m_\alpha} (u - m_\alpha)^2 f(u)du + \alpha \int_{m_\alpha}^{\infty} (u - m_\alpha)^2 f(u)du \tag{3.7}$$

Die Minimierung obiger Formel führt zu:

$$(1-\alpha) \int_{-\infty}^{m_\alpha} (u - m_\alpha) f(u)du + \alpha \int_{m_\alpha}^{\infty} (u - m_\alpha) f(u)du = 0. \tag{3.8}$$

Nach einigen algebraischen Umformungen und einsetzen von $F(x)$ und $G(x)$ in Formel 3.8 ergibt sich eine Gleichung für das theoretische Expektil m_α.

$$m_\alpha = \frac{(1-\alpha)G(m_\alpha) + \alpha(m_{0.5} - G(m_\alpha))}{(1-\alpha)F(m_\alpha) + \alpha(1 - F(m_\alpha))} \tag{3.9}$$

In Formel 3.9 ist $m_{0.5}$ der Mittelwert der zugrundelegenden Verteilung F und $G(\infty) = m_{0.5}$. Löst man die Formel 3.9 nach α auf, so erhält man

$$\alpha = \frac{G(m_\alpha) - m_\alpha F(m_\alpha)}{2(G(m_\alpha) - m_\alpha F(m_\alpha)) + (m_\alpha - \mu)} \tag{3.10}$$

mit μ als Erwartungswert von $F(x)$ und $G(\infty) = \mu$. Für den Fall, dass m_α gegeben ist, ist es einfach den Asymmetrieparameter α zu bestimmen. Um m_α für ein gegebenes α zu bestimmen, muss die letzte Gleichung mit einigen numerischen Schritten invertiert werden, vgl. [Schnabel, 2011, Kapitel 1]. Nun gibt es nach der Einführung der Expektile weitere statistische Herausforderungen zu lösen. Zum einen muss der optimale Glätteparameter bestimmt werden, um flexible glatte Kurven zu erhalten. Zum anderen gilt es kreuzende Expektilkurven zu vermeiden. Die Glätte der Kurven hängt von der richtigen Wahl des Parameters ab. Wie bereits erwähnt werden in Kapitel 4.3 Verfahren zur richtigen Wahl vorgestellt. Das Problem kreuzender Expektilkurven wird in Kapitel 4.4 behandelt.

4 Expektile Regression

4.1 Allgemein

Bei der (linearen) Regression ist es das Ziel den Erwartungswert einer unabhänigigen Zielgröße anhand mehrerer Kovariablen zu schätzen. Viel mehr Information, lässt sich durch die Quantilregression aus den Daten holen. Diese liefert nicht nur den Median, sondern auch das untere und das obere Quantil. Hiermit lassen sich wiederum der Wertebereich und die Streuung der Zielgröße untersuchen. Mithilfe der Gewichtsfunktion, die die gewünschte Asymmetrie gewährleistet und des quadratischen Terms lassen sich diese Eigenschaften auch auf die so definierten Expektile übertragen. Expektile können in das bekannte Regressions-Modell mit aufgenommen werden. Zuerst wird ein einfaches Modell,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{\alpha} + \boldsymbol{\epsilon}_{\alpha} \tag{4.1}$$

mit unabhängiger stetiger Zielgröße $\boldsymbol{y}$, Designmatrix $\boldsymbol{X}$ und unabhängigem Fehlerterm $\boldsymbol{\epsilon}_{\alpha}$ betrachtet. α bestimmt den Randbereich der bedingten Verteilung, welche von Interesse ist, vgl. [Sobotka, 2012, Kapitel 2]. Anders als bei der Mittelwertregression, bei der für den Fehlerterm die Annahmen $\mathbb{E}(\epsilon_i) = 0, \forall i$; $Var(\epsilon_i) = \sigma_i, \forall i$ und $Cov(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$ getroffen werden, wird bei der Expektilregression folgende Annahme für den Fehlerterm getätigt:

$$0 = \arg\min_{m} \mathbb{E}(\omega_{\alpha}(\epsilon_{\alpha i})|\epsilon_{\alpha i} - m|^2) \tag{4.2}$$

Das heißt, $\boldsymbol{X}\boldsymbol{\beta}_p$ ist gleich dem α-Expektil des Response y_i, vgl. [Sobotka, 2012, Kapitel 2]. Zu erwähnen ist auch, dass keine weiteren Annahmen für den Fehlerterm getroffen werden. Die Fehler können heteroskedastisch sein und können auch eine andere Verteilung als die geforderte Normalverteilung im klassischen Modell haben. Der Hauptvorteil der Expektilregression gegenüber der Quantilregression ist wie bereits ausgeführt, dass 3.5 differenzierbar ist. Daraus lässt sich ein einfacher Prozess ableiten, mit dem expektilspezifische Koeffizienten auch in komplexeren Modellen mit nichtlinearen, räumlichen oder Zufallseffekten bestimmt werden können, vgl. [Sobotka, 2012, Kapitel 2]. Expektile nutzen mehr Informationen aus den vorhandenen Daten als Quantile. Denn sie beziehen die Distanz der Beobachtung zum Prädiktor mit ein, wobei Quantile nur die Lage des Datenpunktes ober- oder unterhalb des Prädiktors berücksichtigen. Aus diesem Grund sind Expektile allerdings Ausreißern gegenüber anfälliger. Während bei Quantilen die geschätzte Linie genau durch durch dieselbe Anzahl an Punkten gehen muss, wie es Regressionskoeffizienten gibt, gibt es bei der Expektilregression keine solche Einschränkung, vgl.[Sobotka, 2012, Kapitel 2]

4.2 Penalisierte Schätzverfahren

Bei der Expektilregression erhält man die Koeffizientenschätzer durch Minimierung von 3.5. Die Schätzung kann mittels einer abgewandelten iterativen gewichteten kleinste Quadrate Methode durchgeführt werden.

$$\hat{\boldsymbol{\beta}}_p^{[b]} = (\boldsymbol{X'W}_p^{[b-1]}\boldsymbol{X})^{-1}\boldsymbol{XW'}_p^{[b-1]}\boldsymbol{y} \tag{4.3}$$

Hierbei ist $\hat{\boldsymbol{\beta}}_p^{[b]}$ der dem Modell entsprechende Koeffizientenvektor der b-ten Iteration. Die Iteration muss so lange durchgeführt werden, bis Konvergenz mit der Gewichtsmatrix $\boldsymbol{W}_p^{[b]} = diag(w_p(y_1, \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{[\boldsymbol{b}]}), ..., w_p(y_n, \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{[\boldsymbol{b}]})$ herrscht,vgl. [Sobotka et al., 2014].

Nun wird ein flexibles nichtlineares Modell der Form

$$\boldsymbol{y} = f_p\boldsymbol{x} + \boldsymbol{\epsilon}_p \tag{4.4}$$

betrachtet. Für die Expektilkurve f_p gibt es mehrere Möglichkeiten die Funktion zu wählen. Bei der Einführung der Expektile durch Newey und Powell wurde ein lineares Modell bevorzugt, doch wie bereits erwähnt, ist es in der Praxis oft erforderlich flexible Kurven zu modellieren. In der Literatur wird häufig die Kombination mit P-Splines vorgeschlagen. Die Grundidee von Glättungssplines ist es, die unbekannte Funktion $f(x)$, und somit den Einfluss einer stetigen Variable, durch Polynome Splines vom Grad l zu approximieren, vgl. [Sobotka et al., 2014]. Eine Funktion $f : [a, b] \to \mathbb{R}$ heißt Polynom-Spline vom Grad $l > 0$ zu den Knoten $a = K_1 < ... < K_m = b$, falls folgende Bedingungen erfüllt sind, vgl. [Fahrmeir et al., 2009, Kapitel 7].

1. $f(x)$ ist $(l-1)$ mal stetig differenzierbar.

2. $f(x)$ ist auf den durch die Knoten gebildeten Intervallen $[K_j, K_{j+1})$ ein Polynom vom Grad l.

Es kann nun $f(x)$ als $\sum_{k=1}^{K} u_k B_k^{(l)}(x)$ geschrieben werden. $B_k^{(l)}(x)$ sind die B-Spline Basisfunktionen, u_k die dazugehörigen Amplituden und K gibt die Dimension der Basis an. Die B-Splines werden zur Charakterisierung der Menge der Polynom-Splines benötigt, vgl. [Fahrmeir et al., 2009, Kapitel 7]. Diese lassen sich in einer Basismatrix $\boldsymbol{B}$ zusammenfassen, wobei die Amplituden im Koeffizientenvektor $\boldsymbol{u}$ notiert werden. Die Güte der Polynomsplines hängt stark von der Wahl der Anzahl der Knoten ab. Um diese Schwierigkeit zu umgehen, werden penalisierte Splines durch eine bestimmte Auswahl gleichmäßig verteilter Knoten in Kombination mit Penalisierung bestimmt. Diese Idee lässt sich wie folgt zusammenfassen: Es wird die zu schätzende Funktion $f(x)$ durch einen Polynom-Spline mit einer großen Zahl von Knoten approximiert (üblich sind hier 20 bis 40), vgl. [Fahrmeir et al., 2009,

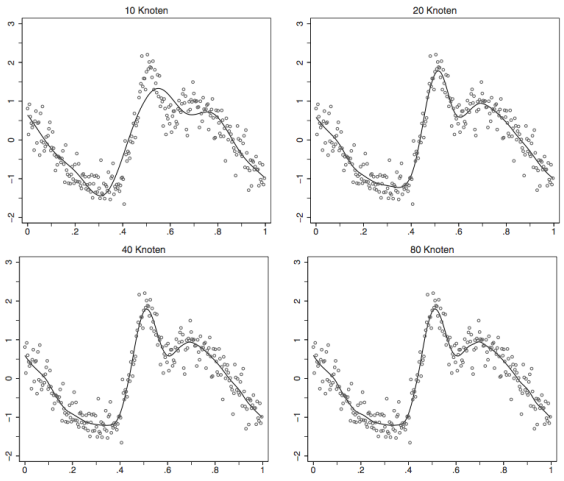


Abbildung 2: Einfluss der Knotenzahl auf die Schätzung von P-Splines.
Grafik übernommen aus [Fahrmeir et al., 2009, Kapitel 7]

Kapitel 7]. Dies gewährleistet, dass die Funktion flexibel genug ist. In Abbildung 2, ist der Einfluss der Knotenzahl auf die Schätzung von P-Splines dargestellt. Hat man zu wenig Knoten gewählt, ist die Schätzung zu unflexibel, das Maximum der Datenpunkte wird nicht erreicht. Je mehr Knoten man wählt, umso flexibler wird die Kurve und die Anpassung wird genauer. Es kann allerdings auch vorkommen, dass bei zu vielen Knoten die Schätzung in einigen Bereichen sehr rau wird.
Zusätzlich zur Approximation wird ein Strafterm eingeführt, der eine zu große Variabilität der Schätzung bestraft, vgl.[Fahrmeir et al., 2009, Kapitel 7]. Im Folgenden wird die einfachere Approximation von Eilers und Marx (1996), die auf den Differenzen benachbarter Koeffizienten basiert, beschrieben. $\boldsymbol{D}$ stellt die Differenz-Matrix r-ter Ordnung dar, die Bestrafungsmatrix $\boldsymbol{P} = \lambda \boldsymbol{D'D} = \lambda \boldsymbol{K}$ bringt eine Bestrafung über den Glätteparameter λ und quadrierter Differenzen r-ter Ordnung der Basisfunktionen beispielsweise wie folgt ein. $\boldsymbol{u'Pu} = \sum_{k=r+1}^{K} (\Delta_r(u_k))^2$, wobei Δ_r die Differenz r-ter Ordnung ist (üblich ist die Wahl von $r = 2$), vgl. [Sobotka et al., 2014]. Die Schätzung der Koeffizienten ergibt sich nun durch Iteration zwischen der Berechnung von

$$\hat{\boldsymbol{u}}_p^{[b]} = (\boldsymbol{B'W}_p^{[b-1]}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B'W}_p^{[b-1]}\boldsymbol{y} \tag{4.5}$$

und der Neuberechnung der Gewichte wie oben geschrieben, vgl. [Sobotka et al., 2014].

4.3 Optimaler Glätteparameter

Um die flexiblen, an die Daten angepassten Kurven zu erhalten und eine optimale Schätzung der Effekte durchzuführen, ist die Bestrafung oder Penalisierung notwendig. Im Folgenden werden zwei Methoden dargestellt, um den optimalen Glätteparameter zu bestimmen.

4.3.1 Asymmetrische-Kreuzvalidierung

Die Formel für die Asymmetrische Kreuzvalidierung (ACV) lässt sich aus der bekannten Variante der Kreuzvalidierung herleiten. Im Fall der OLS-Methode ist die Kreuzvalidierung (ordinary cross validation, OCV) wie folgt definiert:

$$CV_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i)^2 \tag{4.6}$$

mit μ_i als Erwartungwert des Modells, welches alle Beobachtungen außer (x_i, y_i), mit $i = 1, ..., n$, enthält. Dieses Verfahren ist auch als leave-one-out Kreuzvalidierung bekannt. Aus der minimierten Funktion 4.6 kann der optimale Parameter bestimmt werden. Eine Alternative zu diesem Verfahren, bei dem n Modelle gerechnet werden müssen, bietet die Generalisierte Kreuzvalidierung (generalized cross validation, GCV):

$$CV_g = \frac{n\sum_{i=1}^{n}(y_i - \mu_i)^2}{[tr(\boldsymbol{I} - \boldsymbol{H})]^2} \tag{4.7}$$

mit $\boldsymbol{H}$ als Hat-Matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X} + \boldsymbol{P})^{-1}\boldsymbol{X}^T$.
Um nun eine asymmetrische Kreuzvalidierung (asymmetric ordinary cross validation, AOCV) berechnen zu können, werden wieder die Gewichte ω_i, wie in Formel 3.3 definiert, und mit in die Gleichung aufgenommen. Es ergibt sich nun eine gewichtete Version von OCV:

$$CV_o^{\omega} = \frac{1}{n}\sum_{i=1}^{n}\frac{\omega_i(y_i - \mu_i)^2}{(1 - h_{ii}^{\omega})^2} \tag{4.8}$$

Hierbei bezeichnet h_{ii}^{ω} das $i-te$ Diagonalelement der Hatmatrix $\boldsymbol{H}^{\omega}$ aus Formel 4.10. Analog lässt sich die asymmetrische generalisierte Kreuzvalidierung (asymmetric generalized cross validation, AGCV) herleiten.

$$CV_g^{\omega} = \frac{n\sum_{i=1}^{n}\omega_i(y_i - \mu_i)^2}{[tr(\boldsymbol{I} - \boldsymbol{H}^{\omega})]^2} \tag{4.9}$$

Die in 4.9 verwendete Hatmatrix $\boldsymbol{H}^{\omega}$, ist wie folgt definiert, vgl.[Schnabel, 2011, Kapitel 2]:

$$\boldsymbol{H}^{\omega} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{P})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2} \tag{4.10}$$

Diese Methode ist in dem R package `expectreg` implementiert und kann dort bei der Schätzung zur Bestimmung des optimalen Glätteparameters verwendet werden.

4.3.2 Schall-Algorithmus

Als Alternative zur asymmetrischen Krezuvalidierung wird ein weiteres Vorgehen zur Bestimmung des optimalen Glätteparameters in einem LAWS-Modell dargestellt. Hierfür wird der von Schall 1991 eingeführte Algorithmus modifiziert. Ursprünglich

wurde dieser Algorithmus als Werkzeug für die Anwendung in generalisierten linearen Modellen entworfen,vgl. [Schnabel, 2011, Kapitel 2].
„Aufgrund der Korrespondenz von penalisierten Splines und generalisierten gemischten Modellen kann dieser Algorithmus auf Expektile angewendet werden“[Schnabel, 2011, Zusammenfassung]
Folgende Definition wird für die Einführung des Algorithmus genutzt: $\boldsymbol{Ba} = \boldsymbol{X\beta} + \boldsymbol{Zu}$, mit einer Matrix $\boldsymbol{X}$, deren erste Spalte die Eins-Spalte ist und in deren restlichen Spalten die Variablen $\boldsymbol{x}$ abgetragen sind. Die Matrix $\boldsymbol{Z}$ hat eine spezielle Struktur, sodass $\boldsymbol{u} = \boldsymbol{D}$a gilt.
Der Schall Alghorithmus lässt sich in einem Modell gemeinsam mit normalen P-Splines anwenden. Er besteht aus 2 alternierenden Schritten. Im ersten Schritt wird der P-Spline Koeffizient a für ein gegebenes λ geschätzt. Im zweiten Schritt werden die Varianzkomponenten geschätzt:
σ_ϵ^2 für die Fehler $y - \mu$ und
σ_u^2 für die Kontraste $u = \boldsymbol{D}a$.
Hieraus wird ein neuer Wert für den Glätteparameter λ mit $\sigma_\epsilon^2/\sigma_u^2$ berechnet. Im Anschluss wird eine neue Iteration gestartet.
Wird der ursprünglichen Schall Algortihmus an das LAWS-Verfahren angepasst, berechnet man die Parameter a_ω unter Einbezug der Gewichte ω aus Funktion 3.3. Ein Startwert für λ_ω wird vorgegeben. Im zweiten Schritt werden $\sigma_{\epsilon\omega}^2$, der gewichteten Fehler, und die Kontraste $u_\omega = \boldsymbol{D}a_\omega$ wie folgt ermittelt:

$$\sigma_{\epsilon\omega}^2 = \frac{(y-\hat{\mu})^T \boldsymbol{W}(y-\hat{\mu})}{n - ED}, \sigma_{u\omega}^2 = \frac{\|\hat{u_\omega}\|^2}{ED} \tag{4.11}$$

n bezeichnet den Stichprobenumfang und ED ist als $tr(\boldsymbol{H}^\omega)$ definiert. Nun wird λ_ω als $\lambda_\omega = \sigma_{\epsilon\omega}^2/\sigma_{u\omega}^2$ neu berechnet und zwischen der Glätte- und Varianzberechnung bis zur Konvergenz mit λ_ω iteriert, vgl. [Schnabel, 2011, Kapitel 2]. Diese Methode zur Bestimmung des optimalen Glätteparameters ist ebenso im R-Package `expectreg` als itereativer Prozess implementiert. vgl. [Sobotka et al., 2014].
Bei den verschiedenen durchgeführten Simulationen kam es zu folgendem Ergebnis: Die Anwendung von Schall's Algorithmus ist etwas schneller, als die Wahl von λ durch ACV. Die Ergebnisse sind qualitativ vergleichbar, es ergaben sich annähernd identische Ergebnisse, vgl. [Schnabel, 2011, Kapitel 2]

4.4 Kreuzende Expektilkurven

Aufgrund der Daten und der Modellierung kann es vorkommen, dass sich die geschätzten Kurven kreuzen. Dies ist der Fall, wenn $\hat{m}_\alpha > \hat{m}_{\alpha'}$ für $\alpha < \alpha'$ für ein x geschätzt wird, vgl. [Schulze-Waltrup et al., 2014a]. Bei Quantilen tritt dieses Problem häufiger auf als bei Expektilen. Theoretisch ist es bei Expektilen nicht möglich, dass es zur

Kreuzung der Kurven kommt, doch in der Praxis hat man mit diesem Problem dennoch zu tun. Vor allem bedingt durch die zufällige Streuung in kleinen Stichproben kann es zu kreuzenden Kurven kommen, vgl. [Schnabel, 2011, Zusammenfassung]. Es gibt Ansätze und Möglichkeiten dieses Problem zu lösen. Nachfolgend werden zwei Methoden zum Umgang mit kreuzenden Kurven vorgestellt.

4.4.1 Restringierte Expektilregression

Um das Problem einzudämmen wurde ein Verfahren zur Vermeidung kreuzender Kurven eingeführt. Da es bei der Quantilregression öfter zu kreuzenden Kurven kommt, wurde das Verfahren einer restringierten Regression hierfür implementiert, es lässt sich aber, wie unten aufgezeigt, auch bei der Expektilregression anwenden, vgl. [He, 1997]. Dieses Verfahren basiert auf einem location-scale Modell. He verwendete folgende Formel für die nonparametrische bedingte Quantilfunktion $q(x, \alpha) = t(x)+s(x)c_\alpha$. Das Verfahren der restringierten Kurven erstreckt sich über drei Schritte. Im ersten Schritt wird die bedingte Median-Funktion $t(x)$ bestimmt, anschließend erfolgt im zweiten Schritt die Schätzung der glatten nicht-negativen scale-Funktion $s(x)$. Im letzten Schritt wird schrittweise der Asymmetrieparameter c_α für jede $\alpha-$Quantilkurve extra berechnet, vgl. [Sobotka et al., 2014].
Der Nachteil dieses Verfahrens ist, dass eine Menge an Flexibilität aufgegeben wird und nicht alle Informationen, vor allem in heteroskedastischen Fällen, genutzt werden. Trotz allem liefert ein Modell mit restringierten Kurven bessere Ergebnisse als ein Modell indem nichts gegen kreuzende Kurven unternommen wird. Hat man zudem den Fall, dass sich viele Kurven kreuzen, ist der Verlust an Flexibiltät durch restringierte Kurven auf jeden Fall zu bevorzugen. Wendet man diese Methode auf die Expektile an, so sind auch hier drei Schritte von Nöten. Im ersten Schritt wird die Mittelwertsfunktion $t(x)$ durch die Mittelwertregression bestimmt, anschließend werden die Residuen genutzt, um $s(x)$ auf dieselbe Weise wie bei Quantilen zu schätzen. Abschließend wird c_α als Regressionskoeffizient einer Expektilregression mit Response $y - t(x)$ und s(x) als Kovariable geschätzt,vgl. [Sobotka et al., 2014]. Durch die Wahl einer nicht-negativen scale-Funktion wird das Problem kreuzender Expekilkurven vermieden.

4.4.2 Expektilbündel

Ein weiterer Ansatz um kreuzende Kurven zu vermeiden, ist das Expektilbündel. Hierbei wird nicht wie bei LAWS jede Kurve einzeln mit Hilfe von P-Splines für jeden Parameter α geschätzt, sondern mittels eines location-scale-model Ansatzes. Das Expektilbündel ist gegeben als $\mu_i(x, \alpha) = t(x) + c(\alpha)s(x)$. Ein solches Bündel besteht aus drei Komponenten: einer glatten Trendkurve $t(x)$, einer ebenfalls glatten Kurve $s(x)$, die die lokale Streuung des Bündels beschreibt, sowie einer Asymme-

triefunktion $c(\alpha)$. Die Kurven des Expektilbündels bauen auf P-Splines auf und werden mit Hilfe von LAWS geschätzt. Die Glätteparameter werden durch Kreuzvalidierung bestimmt, vgl. [Schnabel, 2011, Kapitel 4]. Bei den Bündeln besteht eine Ähnlichkeit mit den restringierten Expektilkurven. Ebenso bildet die für die restringierte Regression vorgestellte Formel die Basis für die drei Schritte. Der große Unterschied zwischen der restringierten Expektilregression und den Expektilbündeln liegt hier allerdings auf der zusätzlichen Iteration im zweiten und dritten Schritt . Nun wird im zweiten Schritt die optimale Residuenkurve $s(x)$ für alle berechneten Expektile geschätzt. Für die Schätzung der Residuenkurve wird gewöhnlich die LAWS-Methode verwendet. Im Anschluss an die Schätzung der Residuenkurve wird der Asymmetrieparameter c_α neu berechnet. Diese beiden Schritte werden bis zur Konvergenz durchgeführt, vgl. [Sobotka et al., 2014].
Als Ergebnis erhält man mit beiden Verfahren nicht kreuzende Expektilkurven.

4.5 Quantile aus Expektilen

Die Expektilregression ist in vielen Bereichen besser ist als die Quantilregression, allerdings ist die Expektilregression schwieriger zu interpretieren. Aus diesem Grund gibt es bereits Methoden die es möglich machen, aus den leicht zu schätzenden Expektilen gut interpretierbare Quantile zu berechnen. Bei vorhandenen bzw. gegebenen Expektilen ist es möglich die zugehörigen Verteilungsfunktion zu bestimmen, denn diese ist eindeutig durch Expektile charakterisiert. Zwar besteht hier kein intuitiver Zusammenhang wie bei Quantilen, denn dort gilt, dass $F(y)$ eindeutig durch $q_\alpha = q(\alpha) = F^{-1}, \alpha \in \ (0,1)$ definiert ist, aber $q(\alpha)$ lässt sich numerisch aus $m(\alpha)$ herleiten, vgl. [Schulze-Waltrup et al., 2014a]. Ein Verfahren, um aus einer geschätzten Expektilfunktion die Quantilfunktion zu berechnen, wird nachfolgend vorgestellt. Grundlage hierfür bildet ein feines Raster von nah beieinander liegenden Expektilkurven mit zugehörigem Asymmetrieparameter, $\alpha_t \in (0,1)$ mit $t = 1, ..., T$ und zusätzlicher Vorgabe: $\alpha_1 < \alpha_2 < ... < \alpha_T$. Die Expektile sind durch eine bijektive Funktion $h(.)$ mit den Quantilen verknüpft. Eine Methode die Quantile zu berechnen ist nachfolgend skizziert. Als erstes wird $F_t := F(m_{\alpha t})$ defininiert, diese Funktion soll aus $m_{\alpha t}$ bestimmt werden. Nun wird F_t spezifiert und zwar so, dass gilt: $F_t = \sum_{j=1}^{t} \xi_j, t = 1, ..., T$, ξ_t lässt sich aus $m_{\alpha t}$ berechnet. Zu beachten ist, dass $\xi_t > 0$ für $t = 1, ...T$ gelten muss und $\sum \xi_t \leq 0, \forall t$ erfüllt ist, denn ansonsten liegen die Voraussetzungen einer Verteilungsfunktion nicht vor. Gestartet wird der Prozess der Berechnung indem man $F_0 \equiv 0$ und m_0 als kleinstes Expektil definiert, vgl. [Sobotka et al., 2014]. Hat man Zugriff auf die Originaldaten, so kann man für m_0 das Minimum der beobachteten Werte y_i wählen. Analog kann dieses Vorgehen für das Maximum der beobachteten Werte y_i angewandt werden, vgl. [Schulze-Waltrup

et al., 2014a] Sinnvoll ist es, extreme Expektile für diese Berechnung zu verwenden, damit die Schätzung möglichst genau erfolgt. Wie bereits in dieser Arbeit ausgeführt, gilt: $G(m) = \int_{-\infty}^{\infty} yf(y)dy$. Die Funktion f wird nun durch die Approximation $\widetilde{f}(y)$ ersetzt. Diese ist wie folgt definiert:

$$\widetilde{f}(y) = \begin{cases} \frac{\xi_t}{m_{\alpha t} - m_{\alpha t-1}} & \text{wenn } y_i \in [m_{\alpha t}, m_{\alpha t-1}) \\ 0 & \text{sonst} \end{cases} \tag{4.12}$$

Durch Einsetzen der approximierten Funktion erhält man

$$\widetilde{G}(m_{\alpha t}) := \int_{-\infty}^{m_{\alpha t}} y \, \widetilde{f}(y)dy = \sum_{j=1}^{t} \frac{m_{\alpha j} + m_{\alpha j-1}}{2} \xi_j \tag{4.13}$$

Anschließend wird die Funktion

$$m_\alpha = \frac{(1-\alpha)G(m_\alpha) + \alpha(m_{0.5} - G(m_\alpha))}{(1-\alpha)F(m_\alpha) + \alpha(1 - F(m_\alpha))} \tag{4.14}$$

verwendet, um das Minimierungsproblem

$$g_t(\xi) := m_{\alpha t} - \frac{(1-\alpha_t)G(m_{\alpha t}) + \alpha(m_{0.5} - G(m_{\alpha t}))}{(1-\alpha_t)F(m_{\alpha t}) + \alpha(1 - F(m_{\alpha t}))}, \tag{4.15}$$

unter der Nebenbedingung $\xi_t > 0 \ \forall t = 1, ..., T$, mit $\boldsymbol{\xi} = (\xi_1, .., \xi_T)$, zu lösen, vgl. [Sobotka et al., 2014]. In R benutzt man das package `quadprog`, um die Funktion $\frac{1}{2}\sum_{t=1}^{T} d_t^2(\boldsymbol{\xi})$ mittels eines iterativen Prozesses zu minimieren. Als Lösung wird $\widetilde{\boldsymbol{\xi}}$ ausgegeben. Diese Verfahrensweise lässt sich nicht nur auf das population expectile anwenden, sondern bietet auch die Möglichkeit die bedingte Verteilung eines Response $\boldsymbol{y}$ durch eine gegebene Kovariable x zu bestimmen. In R gibt es weiterhin die Möglichkeit nicht nur die Verteilung zu schätzen, sondern auch Quantile auszugeben, vgl. [Sobotka et al., 2014]. Diese können dann zur Interpretation genutzt werden.
Ein weiterer Ansatz eine Dichte aus Expektilen zu berechnen, besteht darin, dass die Gesamtdichte aus einer Expektilregression, basierend auf einer restringierten Regression oder Bündel-Schätzung, berechnet wird, vgl [Sobotka et al., 2014]. Dieses Verfahren wird in dieser Arbeit nicht näher ausgeführt.

4.6 Konfidenzintervalle

4.6.1 Herleitung asymptotischer Normalverteilung

Auch bei der Expektilregression ist es möglich, Konfidenzintervalle anzugeben. Es kann zur Kleinste-Quadrate-Punktschätzung eine asymptotische Normalverteilung hergeleitet werden, die der Konstruktion von Konfidenzintervallen zugrundeliegt,

vgl. [Sobotka, 2012, Zusammenfassung]. Sobotka hat im Rahmen seiner Arbeit (2012) hierzu die Formeln vorgestellt. Unter anderem zeigte er, dass die Schätzung unter Verwendung von LAWS mit festen Gewichten asymptotisch normalverteilt ist, beispielsweise gilt:

$$\hat{\boldsymbol{\beta}}_{\alpha}^{0} \overset{a}{\sim} N\left(\boldsymbol{\beta}_{\alpha}^{0}, Var(\hat{\boldsymbol{\beta}}_{\alpha}^{0})\right)$$

Ausserdem bewies er, dass die LAWS-Schätzung mit geschätzten Gewichten ebenso einer asymptotischen Normalverteilung folgt:

$$\hat{\boldsymbol{\beta}}_{\alpha} \overset{a}{\sim} N\left(\boldsymbol{\beta}_{\alpha}^{0}, Var(\hat{\boldsymbol{\beta}}_{\alpha}^{0})\right)$$

Die Einzelheiten zu den jeweiligen zugrundeliegenden Definitionen und Kovarianzstrukturen wurden hier aus Vereinfachungsgründen weggelassen. Diese sind in [Sobotka, 2012, Kapitel 3] nachzulesen. Die Eigenschaften lassen sich auch auf ein semiparametrisches Modell, welches mit LAWS geschätzt wird, übertragen. Für feste Glätteparameter gilt auch hier die asymptotische Normalverteilung. Die Schätzer eines semiparametrischen Modells mit festem Glätteparameter können als $\boldsymbol{\theta}_{\alpha} = \left(\sum_{i=1}^{n} \boldsymbol{u}_i^T w_{i,\alpha} \boldsymbol{u}_i + \boldsymbol{P}\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{u}_i^T w_{i,\alpha} y_i\right)$ geschrieben werden, wobei in $\boldsymbol{\theta}_{\alpha} = (\boldsymbol{\beta}_{\alpha}^T, \boldsymbol{\gamma}_{1,\alpha}^T, ..., \boldsymbol{\gamma}_{r,\alpha}^T)^T$ und $\boldsymbol{u}_i = (\boldsymbol{x}_i^T, \boldsymbol{b}_{i,1}^T, ..., \boldsymbol{b}_{i,r}^T)^T$ alle Regressionskoeffizienten und Designvektoren wiedergegeben werden. Hieraus folgt, dass

$$\hat{\boldsymbol{\theta}}_{\alpha} \overset{a}{\sim} N\left(\boldsymbol{\theta}_{\alpha}^{0}, Var(\hat{\boldsymbol{\theta}}_{\alpha}^{0})\right) \tag{4.16}$$

gilt, wobei $\boldsymbol{\theta}_{\alpha}^{0}$ analog zu $\boldsymbol{\beta}_{\alpha}^{0}$ definiert ist.
Ersetzt man die in dieser Kovarianzmatrix hinterlegten Residuen durch die gefitteten Residuen und nimmt einige weitere Umformungen vor, so erhält man die asymptotische Kovarianzmatrix für den gesamten Schätzer $\hat{\boldsymbol{\theta}}_{\alpha}$. Daraus lässt sich die Kovarianzmatrix für den interessierenden Koeffizienten ableiten. Die Kovarianzmatrix ist im Anhang abgedruckt. Zusammen mit der asymptotischen Normalverteilung der LAWS-Schätzung kann ein Konfidenzintervall für die wahre Funktion $f_{j,\alpha}(\boldsymbol{z}_i)$ angegeben werden.

$$KI(\hat{f}_{j,\alpha}(\boldsymbol{z}_i)) = \left[\hat{f}_{j,\alpha}(\boldsymbol{z}_i) \pm z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{f}_{j,\alpha}(\boldsymbol{z}_i))}\right] \tag{4.17}$$

$z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$, ist das $(1-\frac{\alpha}{2})$- Quantil der Standardnormalverteilung.
Durch die Verwendung der Normalverteilung statt der t-Verteilung besteht eine gewisse Ungenauigkeit, vgl. [Sobotka, 2012, Kapitel 3].

4.6.2 Berechnung mittels Bootstrapverfahren

Ein weiteres Verfahren um ein $(1-\alpha)-$Konfidenzintervall zu erhalten ist die Anwendung des Bootstrap-Verfahrens. Hierbei liegt nicht das theoretische Verteilungsmodell zugrunde, da dies auch häufig nicht bekannt ist, sondern die empirische Verteilung der Stichprobe bildet die Basis. Grundlegend werden aus der vorhandenen Stichprobe mit Umfang n, n unabhängige Stichproben mittels Ziehen mit Zurücklegen gezogen. Diese neuen Stichproben werden zur weiteren Berechnung verwendet. Wendet man dieses Verfahren an, um ein Konfidenzintervall für geschätzte Expektilkurven zu berechnen, kann man wie folgt vorgehen. Hier wird die Verteilung der geschätzten Expektile approximativ bestimmt. Als Erstes werden B- Bootstrap-Stichproben aus den Originaldaten gezogen, $b = 1, ..., B$. Die Ziehung beinhaltet hierbei den Responsevektor $\boldsymbol{y}$ mit zugehöriger Designmatrix $\boldsymbol{X}$. Die Expektile werden nun unabhängig für alle B-Bootstrap-Stichproben aus der unbekannten Verteilung der wahren Expektile $m_\alpha(x_i)$ geschätzt.
Die punktweisen Intervalle werden aus dem $\frac{\alpha}{2}$-B'ten und $(1-\frac{\alpha}{2})$-B'ten Element der geschätzten Expektile für jeden Effekt f_j aus den Bootstrap-Stichproben und i=1,..,n berechnet. Die Expektile werden für diese Berechnung vorher sortiert, vgl. [Schulze-Waltrup, 2014]

5 Praxisteil anhand der Münchner Mietspiegel Daten aus dem Jahr 2013

5.1 Datengrundlage

Im Folgenden wird anhand der Münchner Mietspiegel Daten aus dem Jahr 2013 beispielhaft die Theorie der Expektile angewendet. Die Daten stammen aus einer Mieterbefragung in München. Hierfür wurde eine Stichprobe unter speziellen Kriterien gezogen. Zusätzlich zu den Mietern wurden auch die Vermieter zu Informationen die das Haus betreffen befragt, vgl. [Sozialreferat-München, 2013]. Insgesamt umfasst der Datensatz 3.080 Beobachtungen und 743 Variablen. Für diese Arbeit wird der Datensatz reduziert und nur eine geringe Auswahl an Variablen betrachtet, diese sind: Nettomiete, Nettomiete pro Quadratmeter, Lage der Wohnung, Baujahr, Wohnfläche, Anzahl der Zimmer, Terrasse und Heizung. Für die Wohnfläche wurde die Variable verwendet, bei der die Kappung berücksichtigt wurde. Es werden nur noch die Wohnungen berücksichtigt deren Wohnfläche zwischen 20 qm und 160 qm liegt. Die Kappung wurde durchgeführt, da an den Rändern nur sehr wenig Beobachtungen vorlagen. Für die Variable Baujahr wird die Variable benutzt bei der die fehlenden Werte aus der Mieterbefragung über bedingte Mittelwerte imputiert wurden und der Abgleich zwischen Mieter- und Vermieterangaben durchgeführt wurde.

Das Merkmal Terrasse ist mit Ja kategorisiert, falls die Wohnung eine Terrasse in Süd- oder Westrichtung, mit einer Fläche von mindestens 5 qm besitzt. Die Variable Heizung ist ebenso binär und enthält die 1, falls eine Zentralheizung in der Wohnung vorhanden ist, ansonsten 0, vgl. [Sozialreferat-München, 2013]. Bei der Variable Anzahl Zimmer wurden für diese Arbeit alle Beobachtungen die mehr als 6 Zimmer aufwiesen zu der Ausprägung 6 Zimmer zusammengefasst. Es lag in dem Bereich nur eine minimale Anzahl von Beobachtungen vor.

5.2 Deskriptive Auswertung

Zunächst wurde der reduzierte Datensatz deskriptiv ausgewertet, um sich einen ersten Überblick über die Daten zu verschaffen. Dabei war auffällig, dass die Variable Nettomiete linksschief verteilt ist, die Variable Nettomiete pro Quadratmeter hingegen ist annähernd normalverteilt.

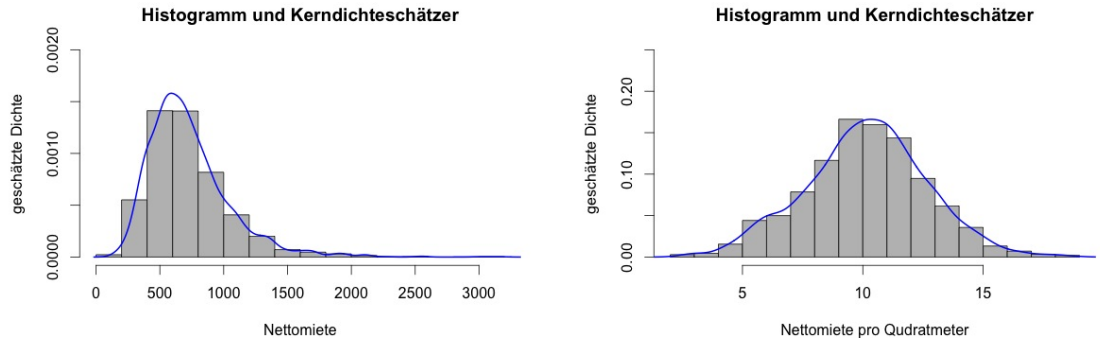


Abbildung 3: Histogramm mit Kerndichteschätzer: Nettomiete und Nettomiete pro Quadratmeter

Bei der Betrachtung des Histogramms für die Variable Wohnfläche ist zu erkennen, dass diese auch linksschief verteilt ist. Die Histogramme der Variablen Wohnfläche und Baujahr sind nachfolgend in Abbildung 4 dargestellt. Bei der Variable Baujahr sind deutliche Schwankungen zu erkennen.

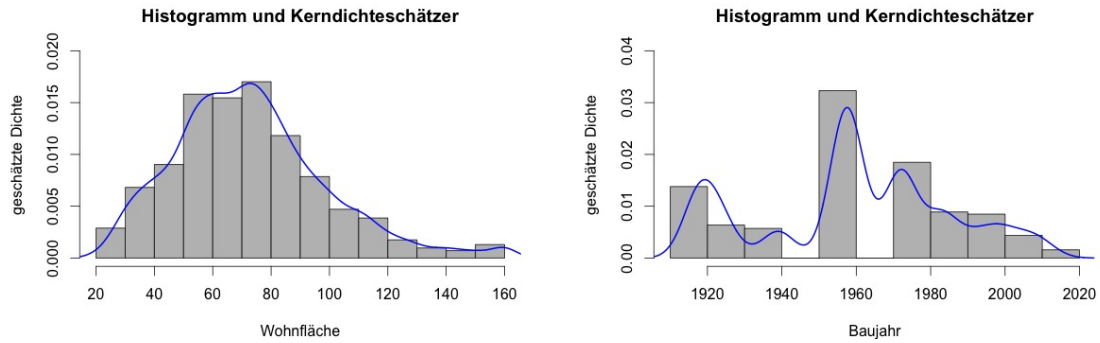


Abbildung 4: Histogramm mit Kerndichteschätzer: Wohnfläche und Baujahr

Im Anschluss wurde der Einfluss der Variablen Lage, Terrasse, Heizung und Anzahl der Wohnräume untersucht. Hierbei war auffallend, dass mit besserer Lage

der Median der Nettomiete pro Quadratemeter einen höheren Wert annimmt, siehe Abbildung 12. Ein analoges Verhalten lässt sich für die Variablen Terrasse und Heizung beobachten. Gehört der Wohnung eine Terrasse an liegt der Median über dem Wert für keine Terrasse, siehe Abbildung 12, bzw. verfügt die Wohnung über eine Zentralheizung so ist auch hier der Median höher, Abbildung siehe 13. Bei der Untersuchung der Anzahl der Zimmer ist zu erkennen, dass je mehr Zimmer die untersuchte Wohnug hat, der Median der Nettomiete pro Quadratmeter geringer wird, siehe Abbildung 13. Die erwähnten Boxplots sind im Anhang abgebildet, diese wurden jeweils proportional zur Größe erstellt, d.h. die Box für 6-Zimmer ist beispielsweise deutlich kleiner, als die für 3 Zimmer, dort sind viel mehr Beobachtungen vorhanden.

5.3 Implementierung der Expektile in R

In dem package `expectreg` wurden diverse Funktionen implementiert, die die Anwendung der Expektile in R ermöglichen. So können auch hier für die gängisten Verteilungen die theoretischen Expektile hergeleitet werden. Bei der Implementierung wurde der bekannte Funktionsname durch den Anfangsbuchstaben „e“ ersetzt. Eine Übersicht der Verteilungen mit der entsprechenden Funktion und den dazugehörigen Parametern ist in Tabelle 2 angegeben.

Verteilung	Expektil Funktion	Parameter
Normal	`enorm`	m, sd
Student t	`et`	df
χ^2	`echisq`	df
Gamma	`egamma`	shape,rate,scale
Exponential	`eexp`	rate
Beta	`ebeta`	a, b
Gleich	`eunif`	min, max
Log-Normal	`elnorm`	meanlog, sdlog
emq	`eemg \ qemq`	m,s

Tabelle 2: Tabelle mit Übersicht der Verteilungen und Funktionen übernommen aus [Sobotka et al., 2014]

Nachfolgend werden der Expektil-Plot sowie der Q-Q-Plot, Abbildung 5, für die Variable Nettomiete pro Quadratmeter gegenübergestellt. Die Vermutung der Normalverteilung wird durch beide Plots bestätigt. Jedoch sind im Q-Q-Plot noch leichte Abweichungen und Schwankungen zu erkennen.
Im weiteren Verlauf werden die Verfahren und Funktionen die den Einsatz der Expektile in der Regression ermöglichen vorgestellt. Ebenso werden gefittete Modelle vorgestellt.

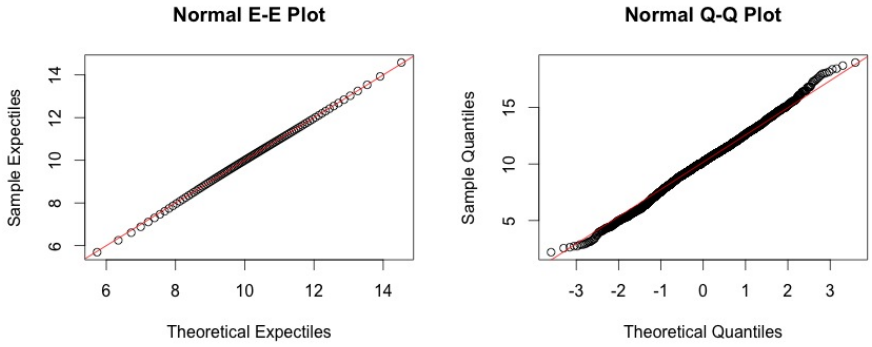


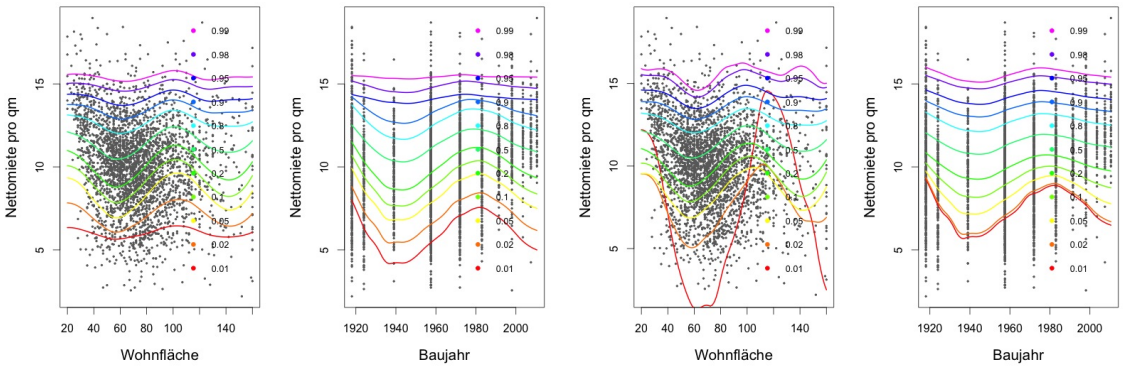
Abbildung 5: Gegenüberstellung des Expektil-Plots und QQ-Plots der Variablen Nettomiete pro Quadratmeter

5.4 Regression[2]

Die LAWS-Methode kann mittels der implementierten Funktion `expectreg.ls` angewendet werden, diese basiert auf der bereits implementierten Funktion `lsfit` aus dem package `stats`. Hierbei wird der Schätzungsprozess solange iterativ durchgeführt bis Konvergenz bei den Gewichten herrscht, vgl. [Sobotka et al., 2014]. Variablen können mittels der Funktion `rb` als flexible Funktion, die den Einfluss der Variable auf die Zielgröße beschreibt, in das Modell mit aufgenommen werden. Verwendet man den P-Spline Ansatz, so wird standardmäßig eine Anzahl von 20 Knoten und die Bestrafungsmatrix, analog dem package `splines`, welche mittels Differenzen 2. Ordnung ermittelt wird, verwendet, vgl. [Sobotka et al., 2014]. Die bereits in dieser Arbeit vorgestellten Methoden um den optimalen Glätteparameter zu bestimmen, sind beide im package `expectreg` implementiert. Zusätzlich zur Schätzung mittels LAWS sind die restringierte Expektilregression, Expektilbündel und Sheets in der Funktion `expectreg.ls` verfügbar.
Anschließend wird ein Modell in sechs unterschiedlichen Varianten gefittet. Es wird jeweils die Nettomiete pro Quadratmeter durch den Einfluss der Wohnfläche und des Baujahres geschätzt. Weitere Variablen werden aus Vereinfachungsgründen nicht mit aufgenommen. Es werden die Methoden LAWS, Bündelschätzung und restringierte Expektilregression angewendet. Zur Wahl des optimalen Glätteparameters wird jeweils GCV und Schall's Algorithmus verwendet. Die Ergebnisse werden nachfolgend in Grafiken gegenüber gestellt.
Als Erstes wurde ein Modell mit **LAWS** und GCV gefittet. Bei der LAWS-Methode wird jede Expektilkurve einzeln geschätzt und das Problem kreuzender Kurven bleibt hier unberücksichtigt. Das Ergebnis ist in Abbildung 6 dargestellt. Es finden

[2] Alle Auswertungen bzgl. der Expektile wurden mit Version 0.39 des package expectreg erstellt. Da die Funktion rb in dieser Version noch fehlerhaft in R implementiert ist, wurde für die Auswertung der Mietspiegeldaten die von Frau Schulze-Waltrup zur Verfügung gestellte Version der Funktion rb verwendet.

keine Überschneidungen der Kurven statt. Der Verlauf der Kurven ist annähernd parallel, dies lässt auf Homogenität der Daten schließen. Bei allen Kurven ist ein Einbruch der Nettomiete pro Quadratmeter bei ungefähr 60 qm Wohnfläche zu erkennen. Allerdings folgt darauf ein erneuter Anstieg zwischen 80 und 100 qm bis schließlich die Nettomieten pro Quadratmeter wieder sinken. Dieser Verlauf ist bei allen sechs Modellen gleichermaßen zu erkennen. Die Spreizung zwischen dem 0.2- und 0.9-Expektil ist ebenso in allen Modellen ersichtlich. Bei der Betrachtung der Grafiken bzgl. des Baujahres lässt sich erkennen, dass für Nachkriegsbauten und Altbauten bis zu einem ungefähren Baujahr 1980 die Nettomieten pro Quadratmeter steigen. Dieses Verhalten ist ebenso in allen Modellen gegeben und über alle Expektile verteilt. Bei den größeren Expektilen, 0.9-, 0.99-Expektil, ist der Verlauf der Kurve deutlich ruhiger und flacher. Für Neubauten ab 1980 sinkt die Nettomiete pro Qudratmeter zum Teil stark ab. Im Bereich der unteren Kurven sind kaum Beobachtungen für die Jahre ab 1980 vorhanden. Ebenso nimmt die Anzahl der Beobachtungen über einer Nettomiete von 15 Euro pro Quadratmeter ab, daher werden die Kurven dort flacher. Bei der Kombination der LAWS-Methode mit



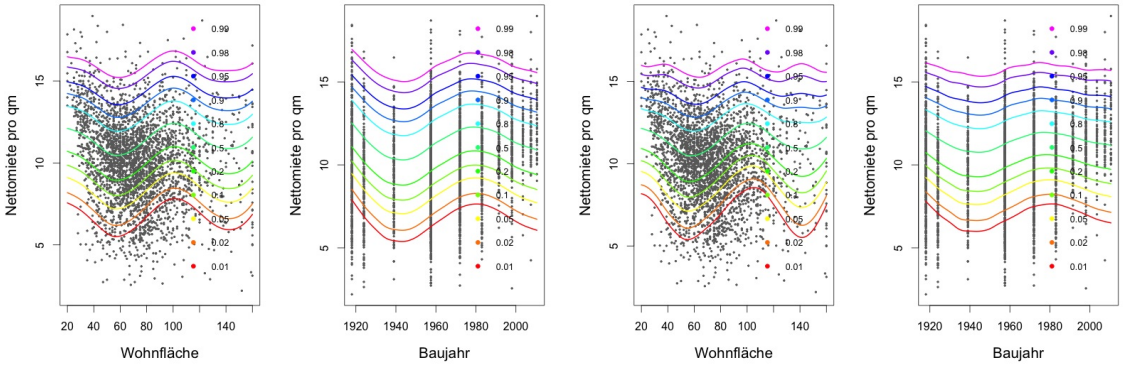
(a) smooth:GCV (b) smooth:Schall

Abbildung 6: geschätze Expektilkurven für Nettomiete pro m^2, Methode:LAWS

Schall's Algorithmus tritt das Problem kreuzender Kurven auf, siehe Abbildung 6. Hierbei ist die Kurve des 0.01-Expektils sehr unruhig und kreuzt sich mit anderen Kurven. Es kreuzt sich auch die Kurve des 0.02-Expektils mit der des 0.05-Expektils im Bereich knapp über 100 Quadratmeter. Bei der Variable Wohnfläche scheint das Problem stärker zu sein, als bei der Variable Baujahr. Hier findet eine Kreuzung nur zwischen dem 0.01- und 0.02-Expektil statt. Die restlichen Kurven verlaufen auch hier nahezu parallel.

Als nächstes werden die beiden Modelle dargestellt, die mit der **Bündelmethode** gefittet wurden. Beim Betrachten der Grafiken, siehe Abbildung 7 fällt auf, dass hierbei keine kreuzenden Kurven auftreten. Dies liegt an der gewählten Methode, denn wie bereits in dieser Arbeit geschrieben, ist die Schätzung unter Einbezug der

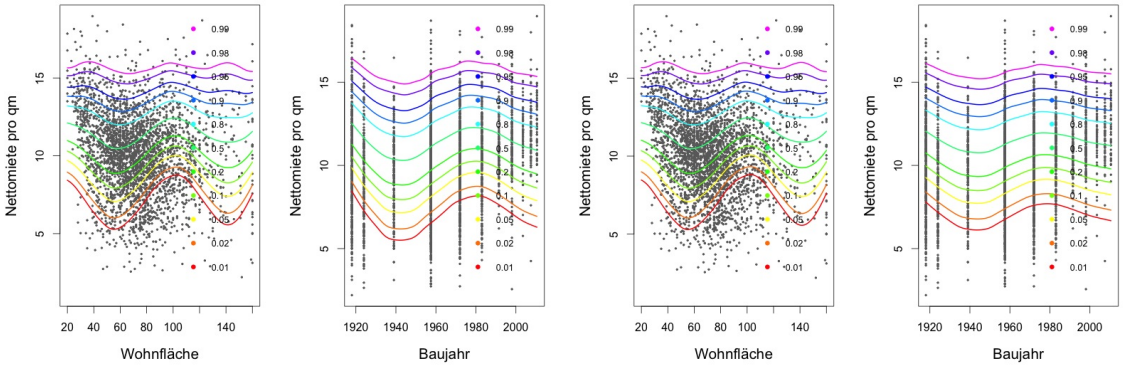
Expektilbündel eine Maßnahme um gegen kreuzende Kurven vorzugehen.



(a) smooth:GCV (b) smooth:Schall

Abbildung 7: geschätze Expektilkurven für Nettomiete pro m^2, Methode:bundle

Zuletzt werden die beiden Modelle grafisch dargestellt, die mittels einer **restringierten Expektilregression**, wieder für beide Verfahren zur Glättung, modelliert wurden. Die Methode der restringierten Expektilregression ist die zweite vorgestellte Variante zur Vermeidung kreuzender Kurven. Dies tritt auch hier in beiden Fällen, Abbildung 8, nicht auf. Vergleicht man die linke Abbildung 8 mit der linken Abbildung 7 so ist zu erkennen, dass der Verlauf der oberen Expektilkurven bei der Wohnfläche unter der Methode der restringierten Expektilregression einen insgesamt ruhigeren Verlauf hat. Sie weist allerdings betragsmäßig mehr Auf- und Abbewegungen auf, allerdings ist die Höhe der Schwankung kleiner. Der Verlauf in der linken Abbildung 7 stellt eher eine Ausnahme dar. Hierbei haben alle Expektilkurven einen parallelen Verlauf.



(a) smooth:GCV (b) smooth:Schall

Abbildung 8: geschätze Expektilkurven für Nettomiete pro m^2, Methode:restricted

Der Verlauf von nahezu parallelen Kurven über alle Ausprägungen hinweg ist nicht immer gegeben. Liegt das Problem der Heteroskedastizität vor, so sehen die Kurven

anders aus. Ein Beispiel mit Kurven zu hetereoskedastischen Daten ist in Abbildung 9 zu sehen. Dieses Beispiel wurde mithilfe des Testdatensatzes `lidar` (light detection and ranging aus dem package `semipar` erstellt. Der Datensatz umfasst 221 Beobachtungen. Als Zielvariable wurde die Entfernung (ratio), die das gestreute Licht zurückgelegt hat, bevor es wieder zur Quelle reflektiert wurde,verwendet. Die Kovariable ist der logarithmierte Anteil des zurückgestreuten Lichts(logratio).

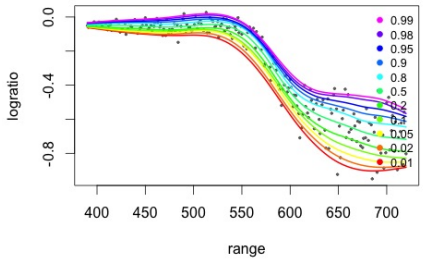


Abbildung 9: geschätzte Expektilkurven für Lidar-Daten,
Methode:LAWS, smooth:GCV

Quantile aus Expektilen

In Kapitel 4.5 wurde erläutert, dass aus den geschätzten Expektilen Quantile berechnet werden können. Hierfür fand eine nochmalige Vereinfahung des Modells statt. Es wird nun nur noch die Variable Wohnfläche als einizigste Kovariable verwendet. Die Anzahl der geschätzten Expektile wurde hingegen erhöht, um eine ausreichende Menge an Expektilen als Grundlage für die weitere Berechnung zu erhalten. Es wurde der Vektor `expectiles=c(0.0001,0.0001,seq(0.01,0.99,0.01),0.999,0.9999)` übergeben. Aus dem hieraus berechneten Modellergebnis wurden anschließend mittels der `predict`-Funktion spezifizierte Kovariablenwerte ausgegeben. Für jede dieser spezifizierten Kombinationen werden Expektile ausgegeben. Diese Expektile wurden zur Umrechnung in Quantile verwendet. Für die Berechnung wurde die Funktion `expectile.2.cdf.final`[3] verwendet. Hierbei werden Expektile und Asymmetrien bestimmt und nicht wie bei der implementierten Funktion `cdf.qp` direkt das Objekt der Klasse `exepectreg` verwendet. Für die Grafik in Abbildung 10 wurden Quantile für folgende ausgewählte Werte berechnet: Wohnfläche 25, 35, ..., 155 und α= 0.01, 0.02, 0.1, 0.2, ..., 0.9, 0.98, 0.99. Für die 0.02- und 0.1-Quantile ist ein unruhiger Verlauf zu erkennen. Die Quantile ab 0.7 weisen einen nahezu konstanten Verlauf über alle Wohnflächen hinweg auf. Bei der Umrechnung wird auf die Funktion `solve.QP` zurückgegriffen. Diese Funktion wird auch beim Schätzvorgangg mittels der Funktion `expectreg.qp` verwendet. Der Vorgang wird hier analog der Schätzung mit

[3]Diese Funktion wurde auch freundlicherweise von Frau Schulze-Waltrup zur Verfügung gestellt.

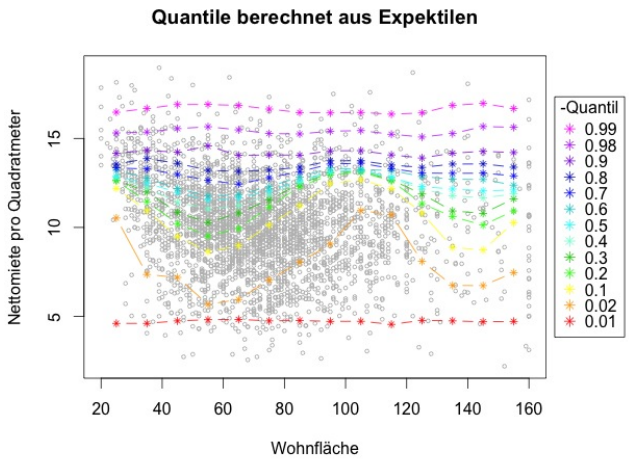


Abbildung 10: Quantile berechnet aus Expektilen. Einzelne Werte sind als Stern dargestellt und durch Striche miteinander verbunden

LAWS als iterativer Prozess durchgeführt, allerdings ist er hier so konzipiert, dass kreuzende Kurven vermieden werden, vgl. [Sobotka et al., 2014]. Bei der Berechnung der Quantile aus den Expektilen erhält man ebenso die bedingte Verteilungsfunktionen der gewählten Kombinationen. In dieser Arbeit wurde mehrfach erwähnt, dass die Expektile als auch die Quantile eine Verteilung eindeutig charakterisieren und sich die Funktion daraus berechnen lässt. Als Beispiel hierzu ist die bedingte Verteilung der Nettomiete pro Quadratmeter bedingt auf eine Wohnfläche von 75 Quadratmetern in Abbildung 11 dargestellt.

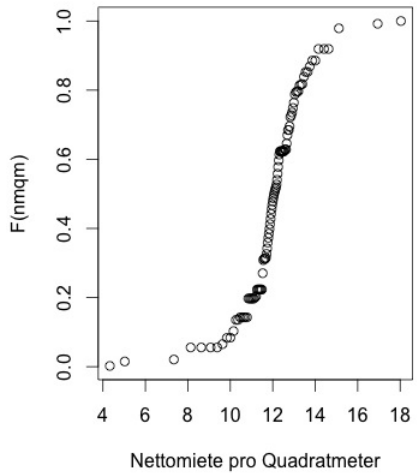


Abbildung 11: bedingte Verteilung der Nettomiete pro Quadratmeter, gegeben 75 Quadratmeter

6 Zusammenfassung

In dieser Arbeit wurden vorwiegend die Expektile und die Regression mit Expektilen vorgestellt.
Durch die Verankerung in der L^2 Welt können die Expektile in viele bereits vorhandene Ansätze mit eingebaut werden. Denn sie sind schnell und gut zu berechnen. Aus diesem Grund sind sie auch ein geeignetes Mittel um mehr Informationen aus den Daten zu gewinnen, als dies bei der gewöhnlichen Mittelwertregression üblich ist. Eine der Schwachstellen ist jedoch, wie bereits mehrfach erwähnt, die fehlende intuitive Interpretation. Allerdings gibt es die Möglichkeit aus den geschätzten Expektilen die zugehörigen Quantile zu berechnen und diese zur Interpretation zu nutzen. In der Arbeit wurde knapp die Aussage bzw. Interpretation eines bestimmten $\alpha-$Expektils beschrieben. Die einzelnen Kurven können in gewissem Maße interpretiert werden.
Um die auftretenden Probleme durch kreuzende Kurven in der Praxis einzudämmen wurden die Verfahren der restringierten Expektilregression und die Expektilbündel vorgestellt. Überwiegend finden die Expektile bis jetzt in semiparametrischen Modellen Anwendung, da sie sich hier auch sehr gut einbinden lassen. Die praktische Anwendung wurde beispielhaft an den Münchner Mietspiegel Daten gezeigt.
Abschließend lässt sich sagen, dass die Expektile stetig an Bedeutung zulegen und aufgrund ihrer guten Recheneigenschaften mehr zur Anwendung kommen. Sie bleiben aber weiterhin hinter den Quantilen zurück.

A Notation

α	Asymmetrieparameter
m_α	$\alpha-$Expektil
q_α	$\alpha-$Quantil
n	Stichprobenumfang
y	Zielgröße
$\boldsymbol{y}$	Responsevektor
x	Kovariable
$\boldsymbol{x}$	Kovariablenvektor
$\boldsymbol{X}$	Designmatrix (parametrische Effekte)
$\boldsymbol{\beta}$	Parametervektor (parametrische Effekte)
$\boldsymbol{B}$	B-Spline Basismatrix
$\boldsymbol{u}$	Koeffizientenvektor zur B-Spline Basis
$\boldsymbol{Z}$	Basismatrix
$\boldsymbol{\epsilon}$	Residuen
$\boldsymbol{H}$	Hatmatrix
$\boldsymbol{I}$	Einheitsmatrix
ω_α	Gewichte
$\boldsymbol{W}$	Gewichtsmatrix
$\boldsymbol{D}$	Differenzmatrix
$\boldsymbol{K} = \boldsymbol{D'D}$	Bestrafungsmatrix
λ	Bestrafungsparameter oder Glätteparameter
$\boldsymbol{P} = \lambda \boldsymbol{K}$	Bestrafungsmatrix
η	linearer Prädiktor
$F(.)$	kumulative Verteilungsfunktion
$f(.)$	Wahrscheinlichkeits- oder Dichtefunktion
$G(.)$	partial (first) moment function
$h(.)$	bijektive Funktion, die Expektile und Quantile verbindet
$\hat{\xi}$	geschätze Schritte der Verteilungsfunktion
$t(.)$	Mittelwerts- oder Medianfunktion
$s(.)$	scale Funktion
σ^2_ϵ	Varianzkomponente für die Fehler $y - \mu$
σ^2_u	Varianzkomponente für die Kontraste $\boldsymbol{u} = \boldsymbol{D}a$

B Abkürzungsverzeichnis

OLS	Ordinary least squares (gewöhnliche kleinste Quadrate-Methode)
LAWS	Least asymmetrically weighted squares (Kleinste asymmetrisch gewichtete Quadrate-Methode)
ACV	asymmetric cross validation (asymmetrische Kreuzvalidierung)
OCV	oridnary cross validation (gewöhnliche Kreuvalidierung)
GCV	generalized cross validation (generalisierte Kreuzvalidierung)
AOCV	asymmetric ordinary cross validation (asymmetrisch gewöhnliche Kreuzvalidierung)
AGCV	asymmetric generalized cross validation (asymmetrisch generalisierte Kreuzvalidierung)

C Erläuterungen

Homogenität	Varianz der Störgrößen bleibt für alle Beobachtungen konstant d.h. $Var(\epsilon_i) = \sigma^2$
Heteroskedastizität	Varianz variiert von Beobachtung zu Beobachtung d.h. $Var(\epsilon_i) = \sigma_i^2$
B-Splines	reiner Fit von Basisfunktion, d.h. $\lambda = 0$
location scale model	Verteilungsparameter werden in diesem Modell separat geschätzt (z.B. Lagparameter, Varianz)

Kovarianzmatrix $\boldsymbol{\theta}_\alpha^0$:

$$Var(\hat{\boldsymbol{\theta}}_\alpha^0) = \left(\sum_{i=1}^{n} \bar{\omega}_{i,\alpha}^0 \boldsymbol{u}_i \boldsymbol{u}_i^T + \boldsymbol{P}\right)^{-1} \left\{\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i^T Var\left((\omega_{i,\alpha}^0)(y_i - \boldsymbol{u}_i^T \boldsymbol{\Theta}_\alpha^0)\right)\right\} \left(\sum_{i=1}^{n} \bar{\omega}_{i,\alpha}^0 \boldsymbol{u}_i \boldsymbol{u}_i^T + \boldsymbol{P}\right)^{-1}$$

Differenzen r-ter Ordnung:

Definition Differenzenmatrix 1. Ordnung:

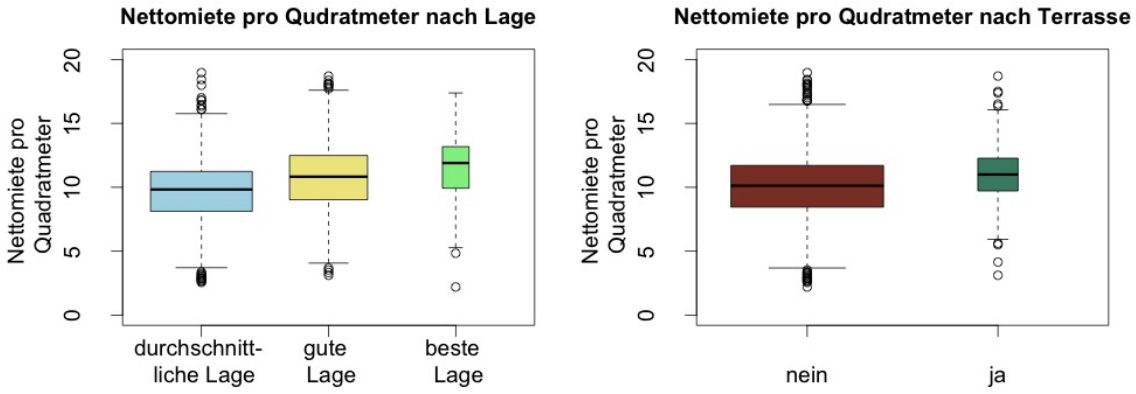
$$\boldsymbol{D}_1 = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

Differenzen höherer Ordnung können mit Hilfe von Differenzematrizen der Form $\boldsymbol{D}_k = \boldsymbol{D}_1 \boldsymbol{D}_{k-1}$ bestimmt werden.
Für $r = 2$ ergibt sich:

$$\boldsymbol{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}$$

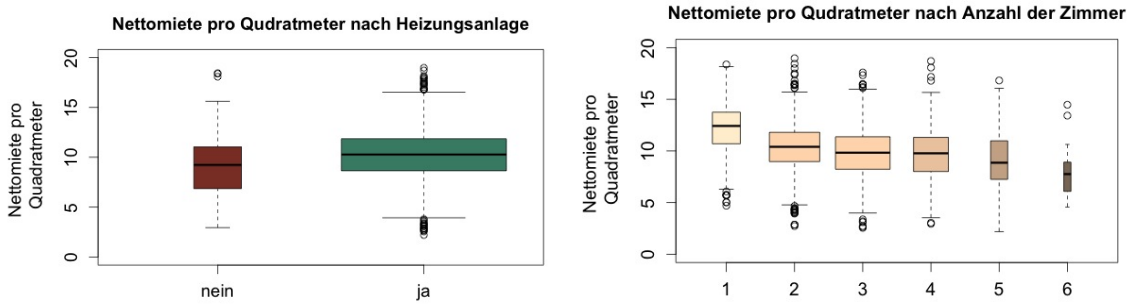
Der Strafterm kann als $\lambda \sum_{k=r+1}^{K} (\Delta_r(u_k))^2 = \boldsymbol{\lambda u}^T \boldsymbol{D}_r^T \boldsymbol{D}_r \boldsymbol{u} = \boldsymbol{\lambda u}^T \boldsymbol{K}_r \boldsymbol{u}$ dargestellt werden.

D Grafiken



(a) (b)

Abbildung 12: proportionale Boxplots der Variablen Lage und Terrasse



(a) (b)

Abbildung 13: proportionale Boxplots der Variablen Heizung und Anzahl Zimmer

Erklärung zur Bachelorarbeit

Hiermit erkläre ich, dass ich die Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe.
Die Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vorgelegt. Sie wurde bisher nicht veröffentlicht.

Ort, Datum	Unterschrift

Abbildungsverzeichnis

Tabellenverzeichnis

Literatur

Eilers, P. H. and Marx, B. D. [1996]. Flexible smoothing with b-splines and penalties, *Statistical science* pp. 89–121.

Fahrmeir, L., Kneib, T. and Lang, S. [2009]. *Regression, Modelle, Methoden und Anwendungen*, Berlin Heidelberg: Springer.

He, X. [1997]. Quantile curves without crossing, *The American Statistician* **51**: 186–192.

Kneib, T. [2013]. Beyond mean regression, *Statistical Modelling* **13**(4): 275–303.

Koenker, R. and Bassett Jr, G. [1978]. Regression quantiles, *Econometrica: journal of the Econometric Society* pp. 33–50.

Newey, W. K. and Powell, J. L. [1987]. Asymmetric least squares estimation and testing, *Econometrica: Journal of the Econometric Society* pp. 819–847.

Schnabel, S. K. [2011]. *Expectile smoothing: new perspectives on asymmetric least squares*, PhD thesis, Universiteit Utrecht.

Schulze-Waltrup, L. [2014]. *Extensions of Semiparametric Expectile Regression*, PhD thesis, Ludwig Maximilian Universität.

Schulze-Waltrup, L., Sobotka, F., Kneib, T. and Kauermann, G. [2014a]. Expectile an quantile regression - david and goliath?, pp. 1–24.

Schulze-Waltrup, L., Sobotka, F., Kneib, T. and Kauermann, G. [2014b]. Semiparametric regression and expectile regression, pp. 1–16.

Sobotka, F. [2012]. *Semiparametric Expectile Regression*, PhD thesis, Carl von Ossietzky Universität Oldenburg.

Sobotka, F., Schnabel, S., Kauermann, G. and Kneib, T. [2014]. expectreg: An r package for expectile regression, pp. 70–100.

Sozialreferat-München [2013]. Mietspiegel für münchen 2013: Statistik, dokumentation und analysen, *Technical report*, Landeshauptstadt München und TNS Infratest GmbH.