Skript zur Vorlesung Computerintensive Verfahren in der Statistik

Stefan Lang

Institut für Statistik

Ludwigstrasse 33

email: lang@stat.uni-muenchen.de

15. Januar 2004

Ich bedanke mich bei zahlreichen StudentInnen für Verbesserungsvorschläge und gefundene Fehler!!!

Mein besonderer Dank gilt Manuela Hummel, die das Skript unglaublich genau durchgelesen hat und neben zahlreichen Schreibfehlern selbst die entlegensten Fehler in Formeln gefunden hat.

Inhaltsverzeichnis

1

Einführung und Beispiele

1.1 Gegenstand der Vorlesung

Die Vorlesung behandelt eine Auswahl moderner computerintensiver Verfahren in der Statistik. Einen Schwerpunkt bilden dabei sogenannte nichtparametrische Verfahren, etwa nichtparametrische Dichteschätzer oder Ansätze zur nichtparametrischen Regression. Nichtparametrische Verfahren sind wesentlich flexibler als ihre parametrischen Pendants und werden heute vor allem zur explorativen Datenanalyse eingesetzt. Einen weiteren Schwerpunkt bilden Methoden zur Ziehung von Zufallszahlen, insbesondere moderne Markov Chain Monte Carlo (MCMC) Verfahren. MCMC Verfahren erlauben das Ziehen von Zufallszahlen aus hochdimensionalen Verteilungen und werden im Moment vor allem in Bayesianischen Ansätzen eingesetzt.

1.2 Datensätze

In diesem Abschnitt beschreiben wir die Datenbeispiele, die während der Vorlesung zur Illustration der verwendeten Verfahren herangezogen werden.

1.2.1 Motorcycledaten

Mit Hilfe dieses bekannten Literaturdatensatzes soll die Beschleunigung des Kopfes im Zeitablauf während eines simulierten Unfalls mit einem Dummy untersucht werden. Der Datensatz besteht aus folgenden Variablen:

Variable	**Beschreibung**
intnr	Beobachtungsnummer
zeit	Zeit in Millisekunden seit dem Aufprall
besch	Beschleunigung des Kopfes

Abbildung 1.1 zeigt einen Scatterplot (Streudiagramm) zwischen Beschleunigung des Kopfes und der Zeit seit dem Aufprall. Man erkennt, dass die Abhängigkeit der Beschleunigung von der Zeit eindeutig nichtlinear ist. Es wird ein Ziel der Vorlesung sein, Schätzmethoden kennenzulernen, mit denen man derartige nichtlineare Beziehungen schätzen kann.

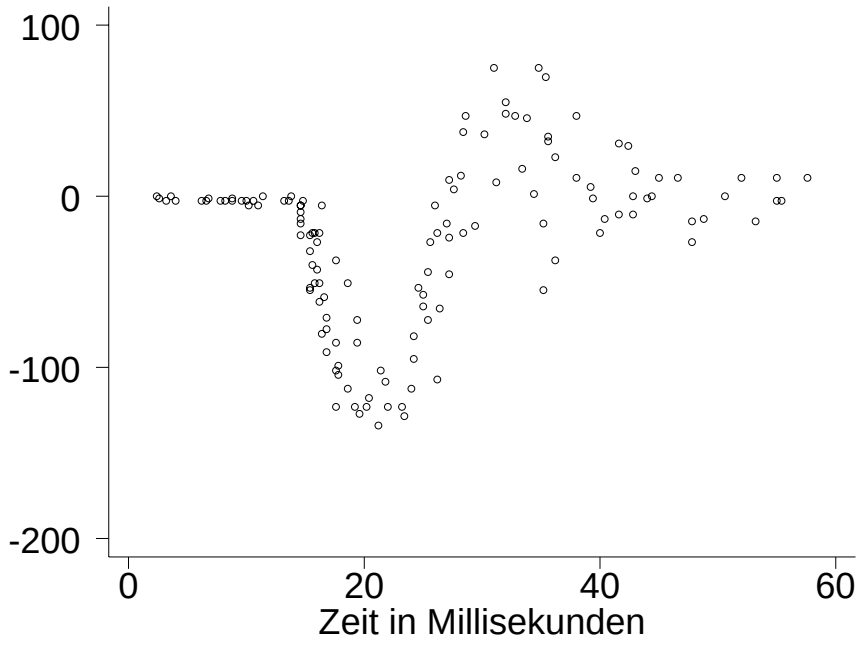

Abbildung 1.1. *Motorcycledaten: Scatterplot zwischen Beschleunigung und der Zeit.*

1.2.2 Mietspiegel für München 1999/2000

Nach dem Gesetz zur Regelung der Miethöhe kann der Vermieter die Zustimmung zu einer Erhöhung des Mietzinses verlangen, wenn "der Mietzins die üblichen Entgelte nicht übersteigt, die in der Gemeinde für nicht preis- gebundenen Wohnraum vergleichbarer Art, Größe, Ausstattung, Beschaffenheit und Lage in den letzten vier Jahren vereinbart oder Erhöhungen geändert worden sind".

Zur Feststellung der "üblichen Entgelte" erstellen die meisten Städte und viele Gemeinden Mietspiegel. Diese ermöglichen die Berechnung der "durchschnittlichen" Miete, die pro Quadratmeter und Monat für eine Wohnung mit einer bestimmten Wohnfläche (in Quadratmeter), dem Baujahr und Merkmalen, welche die Ausstattung der Wohnung, den Haustyp und die Lage der Wohnung in der Gemeinde charakterisieren, bezahlt wird.

Da in größeren Städten wie München eine Erfassung aller Mietpreise schon aus Zeit- und Kostengründen nicht möglich ist, werden Daten zu Miethöhen und zugehörigen Merkmalen

über eine repräsentative Stichprobe gesammelt. In München wird diese Erhebung von Interviewern der Firma Infratest durchgeführt. Für den Mietspiegel 1999 wurden ca. 3000 Wohnungen ausgewählt.

Zur Schätzung der durchschnittlichen Nettomiete (d.h. die Miete ohne Betriebs- und Nebenkosten) wird ein statistisches Modell, ein sogenanntes Regressionsmodell, verwendet. Folgende Variablen sind unter anderen von Interesse:

Interessierende Variable

Nettomiete pro Monat (NM) bzw.
Nettomiete pro Monat und qm ($NMqm$)

Erklärende Variablen

- Wohnfläche in qm (Wfl)
- Baujahr der Wohnung (Bj)
- Lage der Wohnung, auch Verkehrsbelastung
- Zahlreiche Ausstattungsmerkmale, z.B.
  - Ausstattung des Bades
  - Küchenausstattung
  - Art des Hauses (Hochhaus, Mehrfamilienhaus, etc.)

Der Einfluss der Einflussvariablen $Wfl, Bj, \ldots$ auf die Nettomiete pro Quadratmeter $Nmqm$ wurde im offiziellen Mietspiegel mit Hilfe eines *nichtparametrischen Regressionsmodells* modelliert:

$$Nm = f_1(Wfl) + f_2(Bj) + f_{1,2}(Wfl, Bj) + x'\beta + \epsilon$$

Die Funktion f_1 beschreibt dabei die Abhängigkeit von der Wohnfläche Wfl, die Funktion f_2 die Abhängigkeit vom Baualter Bj und die Funktion $f_{1,2}(Wfl, Bj)$ zusätzliche Effekte auf die Miethöhe durch bestimmte Kombinationen (Wfl, Bj) von Wohnfläche und Baujahr.

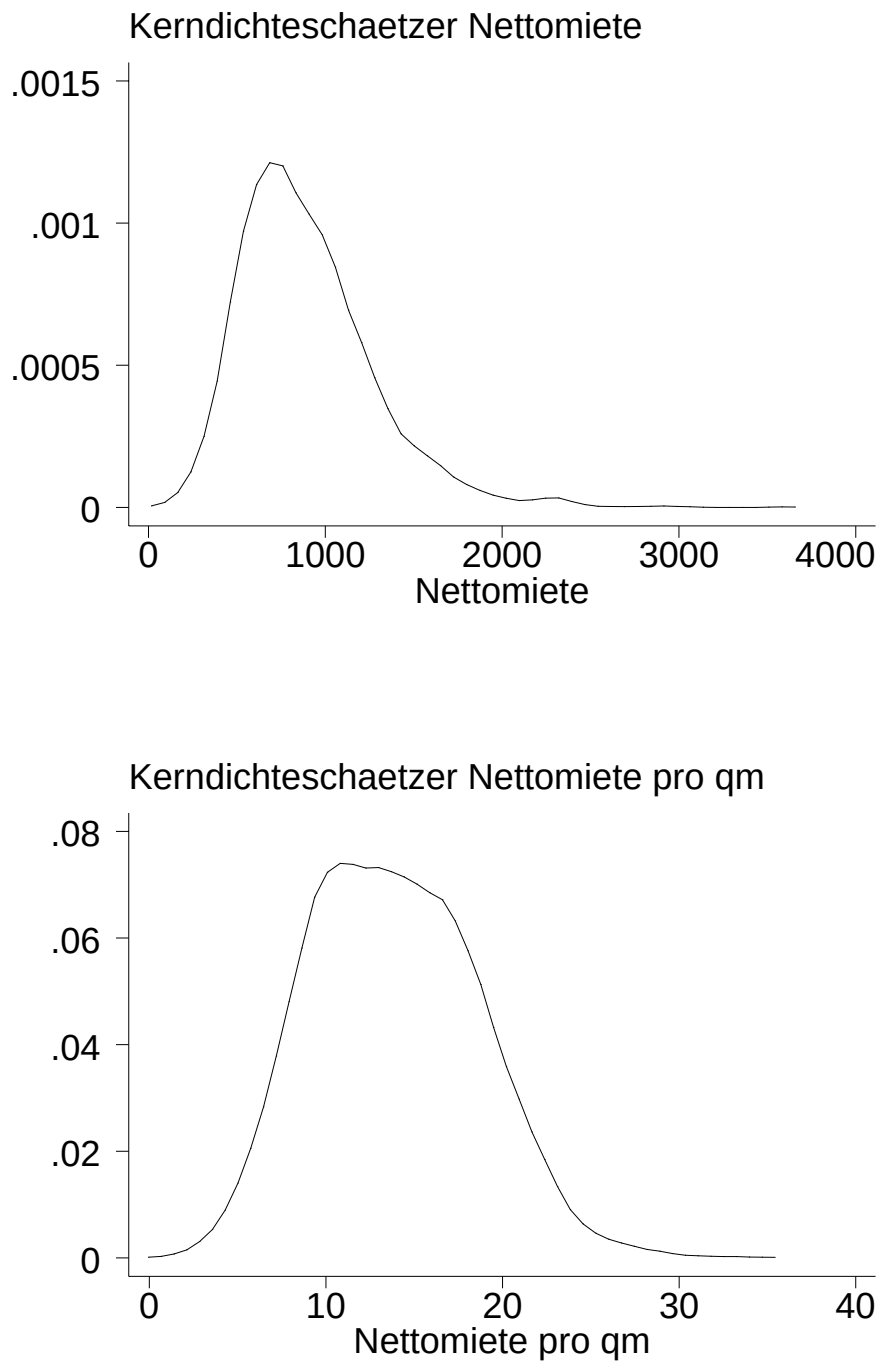

Abbildung 1.2. *Kerndichteschätzer für die Nettomiete und die Nettomiete pro Quadratmeter.*

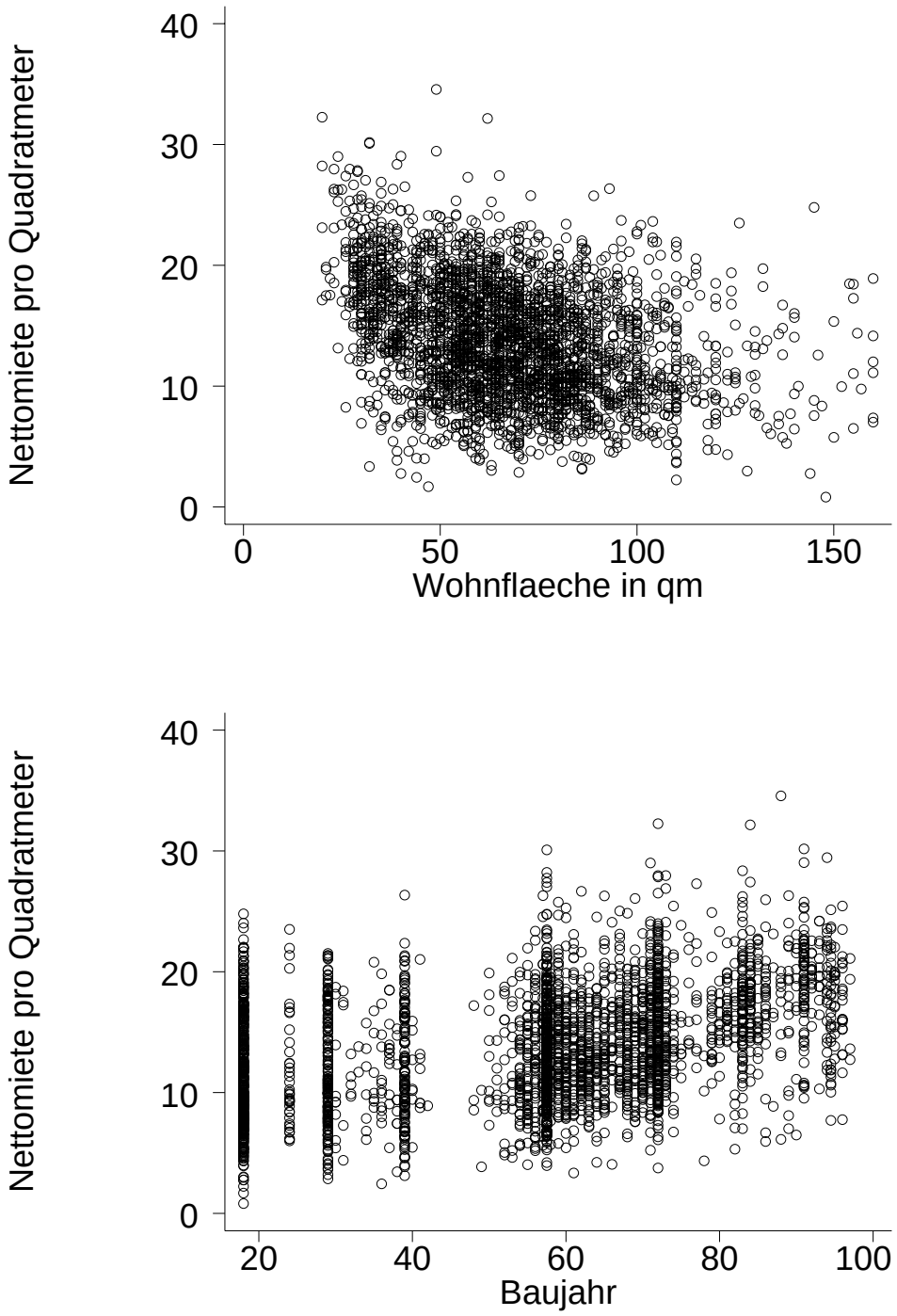

Abbildung 1.3. *Scatterplots: Nettomiete pro qm gegen Wohnfläche und Baujahr.*

1.2.3 Unterernährung von Kindern in Afrika

Unterernährung bei Kindern gilt als eines der schwerwiegensten Gesundheitsprobleme in Entwicklungsländern. Zur Untersuchung von Unterernährung bei Kindern und anderen Fragestellungen (Bevölkerungsentwicklung, Fertilitätsverhalten, AIDS usw.) werden seit ca. 10 Jahren von der Firma ,,Macro International" zahlreiche Studien durchgeführt und unter dem Namen ,,Demographic and Health Surveys" öffentlich zugänglich gemacht. In dieser Vorlesung soll ein Teildatensatz des Demographic and Health Surveys für Zambia von 1992 untersucht werden. Ziel der Untersuchung ist die Bestimmung von Determinaten für Unterernährung bei Kindern. Als Maß für die Unterernährung wird üblicherweise ein sogenannter Z-score herangezogen, der wie folgt definiert ist:

$$Z_i = \frac{AI_i - MAI}{\sigma}$$

Hier bezeichnet AI_i die Größe eines Kindes in einem bestimmten Alter, MAI den Median der Größe für eine Referenzpopulation und σ die Standardabweichung der Referenzpopulation.

Der Datensatz beinhaltet folgende Variablen:

Variable	**Beschreibung**
Z	Z-score
bmi	Body mass Index der Mutter des Kindes
alter	Alter des Kindes in Monaten
rcw	Erwerbsstatus der Mutter 1 = Mutter arbeitet -1 = Mutter arbeitet nicht
edu	Ausbildungsstatus der Mutter 0 = keine Ausbildung 1 = incomplete primary education 2 = complete primary education 3 = incomplete secundary education 4 = complete secondary education 5 = higher education
tpr	Stadt/Land 1 = Mutter lebt in der Stadt -1 = Mutter lebt auf dem Land
sex	Geschlecht des Kindes 1 = männlich -1 = weiblich
reg	Wohnort der Mutter 1 = Central 2 = Copperbelt 3 = Eastern 4 = Luapula 5 = Lusaka 6 = Northern 7 = North-Western 8 = Southern 9 = Western
dist	Wohnort der Mutter (genauere geographische Unterteilung)

1.2.4 Kreditscoring

Um die zukünftige Bonität eines potentiellen Kreditnehmers abschätzen zu können, wurden von einer großen deutschen Bank Daten von früheren Kreditkunden erhoben. Der vorliegende Datensatz enthält neben der Bonität der Kunden auch Kovariablen wie die Kreditdauer oder die Laufzeit des Kredits, von denen angenommen wird, dass sie einen Einfluss auf die Kreditwürdigkeit des Kunden haben, vergleiche Tabelle 1.1. Insgesamt liegt eine geschichtete Stichprobe mit 1000 Beobachtungen vor, von denen 300 aus „schlechten“ Krediten und 700 aus „guten“ Krediten bestehen. Ziel der statistischen Analyse ist die Untersuchung von Art und Umfang des Zusammenhangs zwischen Bonität und den erklärenden Variablen. Ein Besonderheit (im Vergleich zu den zuvor betrachteten Datensätzen) ist, dass die interessierende Variable nur zwei verschiedene Ausprägungen besitzt, 1 für nicht kreditwürdig und 0 für kreditwürdig. Geeignete statistische Modelle zur Analyse dieser Daten werden in Kapitel 4 im Rahmen von generalisierten linearen und additiven Modellen behandelt.

Variable	**Beschreibung**
boni	Bonität des Kunden 1 = Kunde nicht kreditwürdig, d.h. Kredit wurde nicht zurückbezahlt 0 = Kunde kreditwürdig
laufz	Laufzeit des Kredits in Monaten
moral	Frühere Zahlungsmoral des Kunden 1 = gute Moral -1 = schlechte Moral
zweck	Verwendungszweck 1 = privat -1 = geschäftlich
hoehe	Kredithöhe in tausend DM
geschl	Geschlecht 1 = männlich -1 = weiblich
famst	Familienstand -1 = ledig 1 = verheiratet
ko1	laufendes Konto 1 = gutes Konto 0 = kein Konto -1 = schlechtes Konto
ko2	laufendes Konto 1 = schlechtes Konto 0 = kein Konto -1 = gutes Konto

Table 1.1. *Beschreibung der Variablen des Kreditscoringdatensatzes.*

1.2.5 Ein simulierter Datensatz

Ein Ziel der Vorlesung ist es auch Methoden kennenzulernen, mit denen Zufallszahlen aus beliebigen Verteilungen gezogen werden können. Methoden der Zufallszahlengewinnung sind unter anderem aus zwei Gründen von Bedeutung:

1. Zur Beurteilung der Eignung und zur Bestimmung von Eigenschaften statistischer Verfahren durch Simulationsstudien. In diesem Zusammenhang spielt auch der Vergleich konkurrierender Verfahren eine Rolle.
2. Als Hilfsmittel zur Schätzung komplexer statistischer Verfahren (z.B. Bayesianische Regression).

Die Abbildungen 1.4 und 1.5 zeigen Scatterplots zwischen einer abhängigen Variable y und 5 Einflussvariablen x_1, x_2, x_3, x_4 und x_5. Die Daten wurden gemäß folgendem Modell erstellt:

$$y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + f_5(x_{i5}) + \epsilon_i, \quad \epsilon_i \sim N(0, 0.6^2)$$

Die Funktionen f_1-f_5 sind dabei verschiedene stetige Funktionen. Offenbar ist es aufgrund der Scatterplots für einige Einflussgrößen unmöglich auf den Verlauf der Funktionen zu schließen. Oberflächlich betrachtet würde man sogar zu der Vermutung gelangen, dass die Kovariable x_5 überhaupt keinen Einfluss auf y besitzen.

Die Abbildungen 1.6 und 1.7 zeigen Schätzungen der Funktionen $f_1 - f_5$, die mit Hilfe eines Verfahrens der nichtparametrischen Regressions (Kapitel 3) gewonnen wurden. Die bei der Simulation tatsächlich verwendeten Funktionen sind zum Vergleich in die Grafiken mitaufgenommen worden. Das verwendete Schätzverfahren beruht übrigens auch auf der Ziehung von Zufallszahlen.

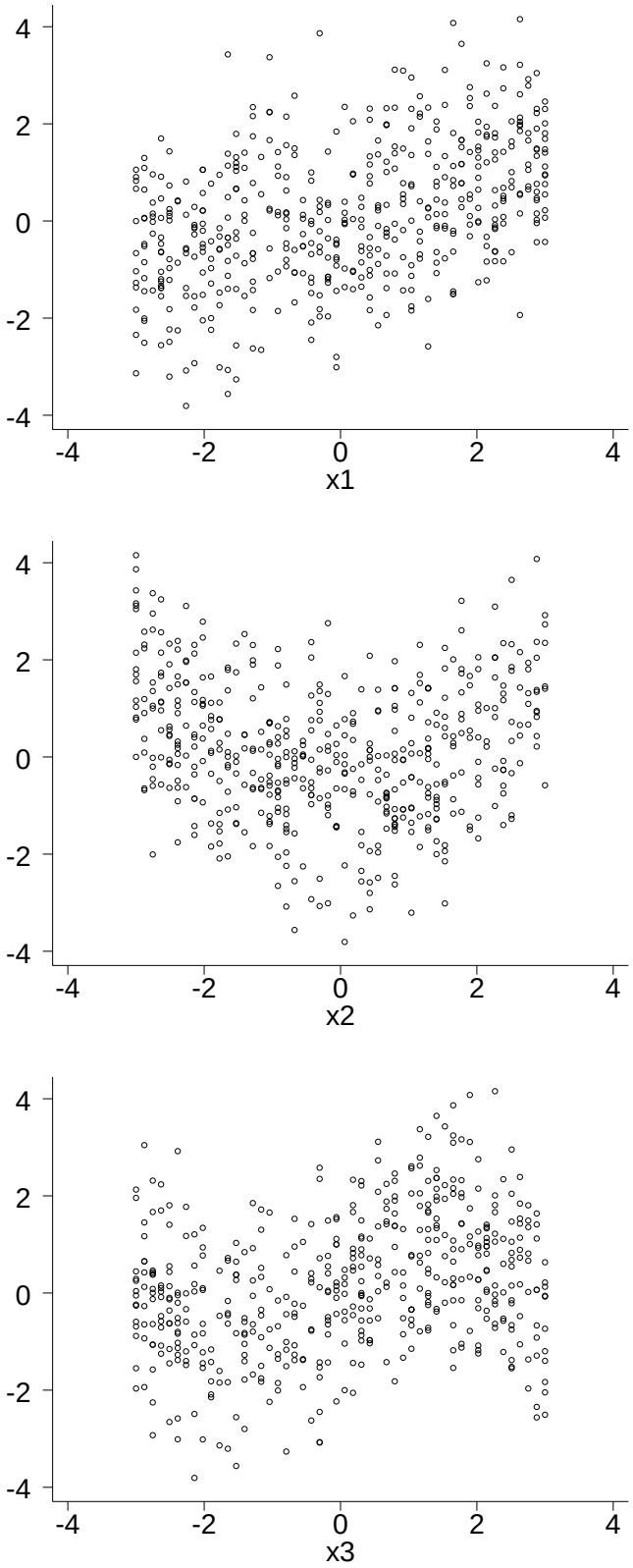

Abbildung 1.4. *Simulierter Datensatz: Scatterplots zwischen abhängiger Variable und den Kovariablen.*

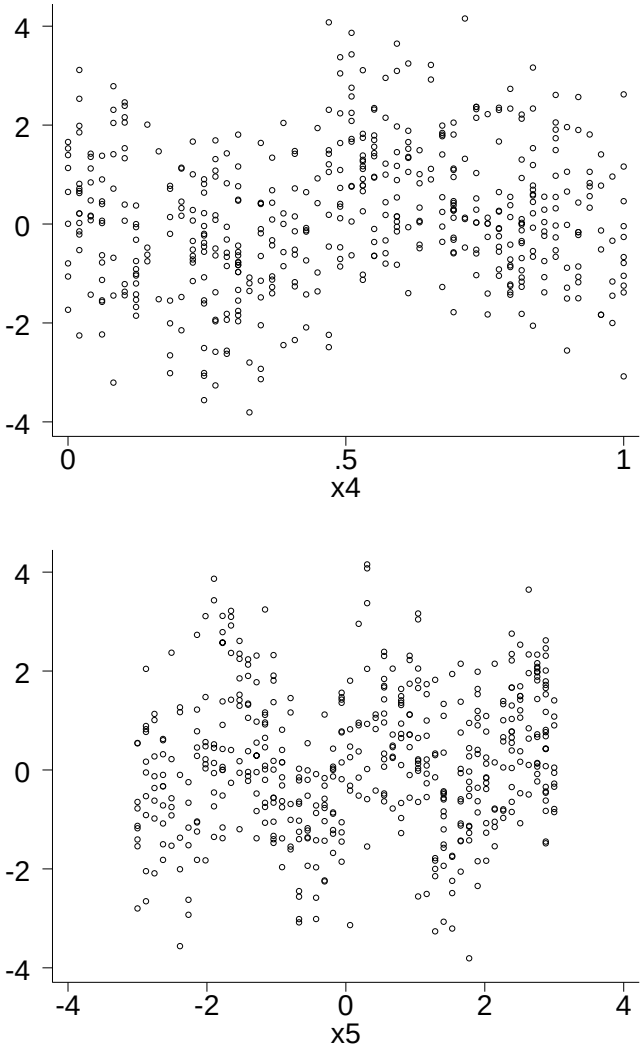

Abbildung 1.5. *Simulierter Datensatz: Scatterplots zwischen abhängiger Variable und den Kovariablen.*

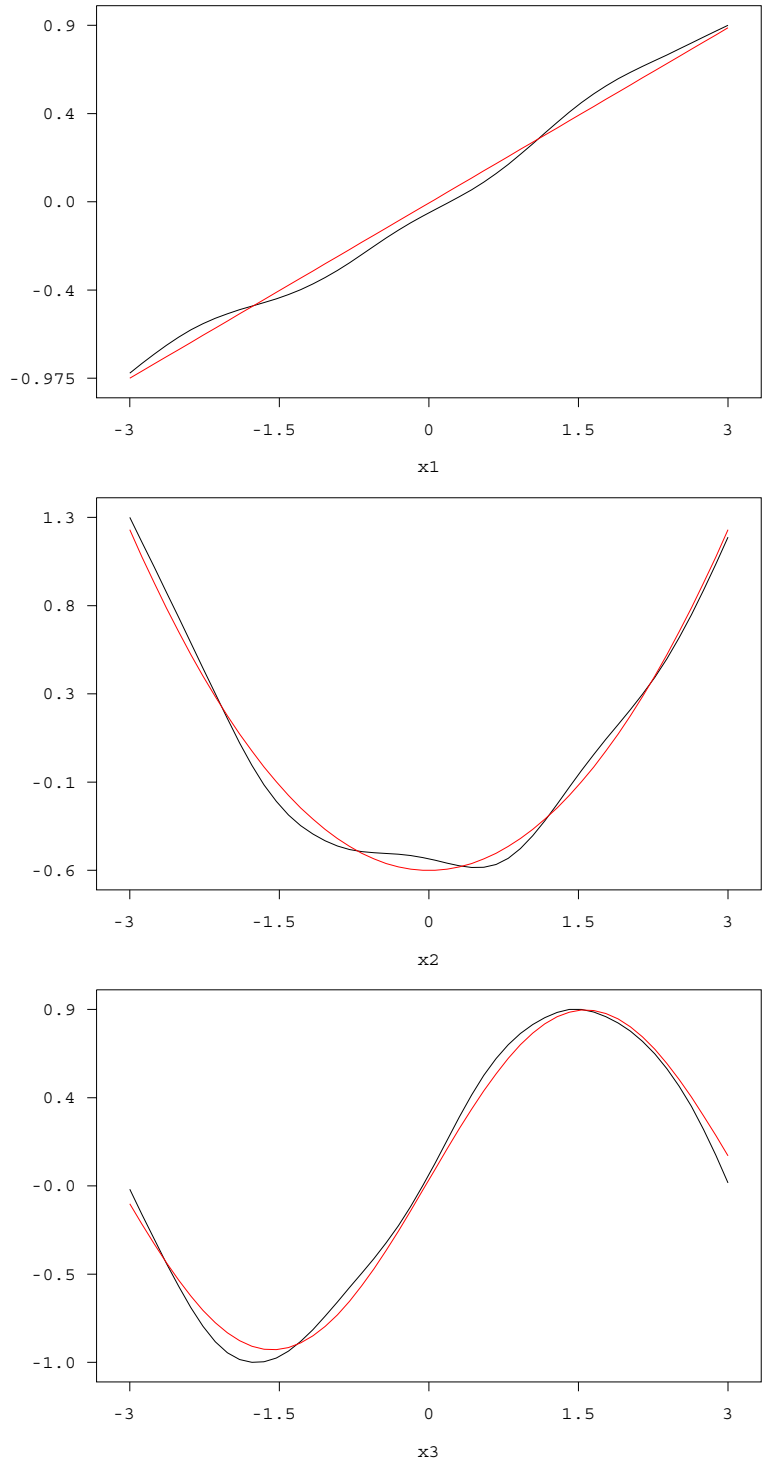

Abbildung 1.6. *Simulierter Datensatz: Geschätzte Funktionen f_1,f_2 und f_3. Die wahren Funktionen sind in den Grafiken rot eingezeichnet.*

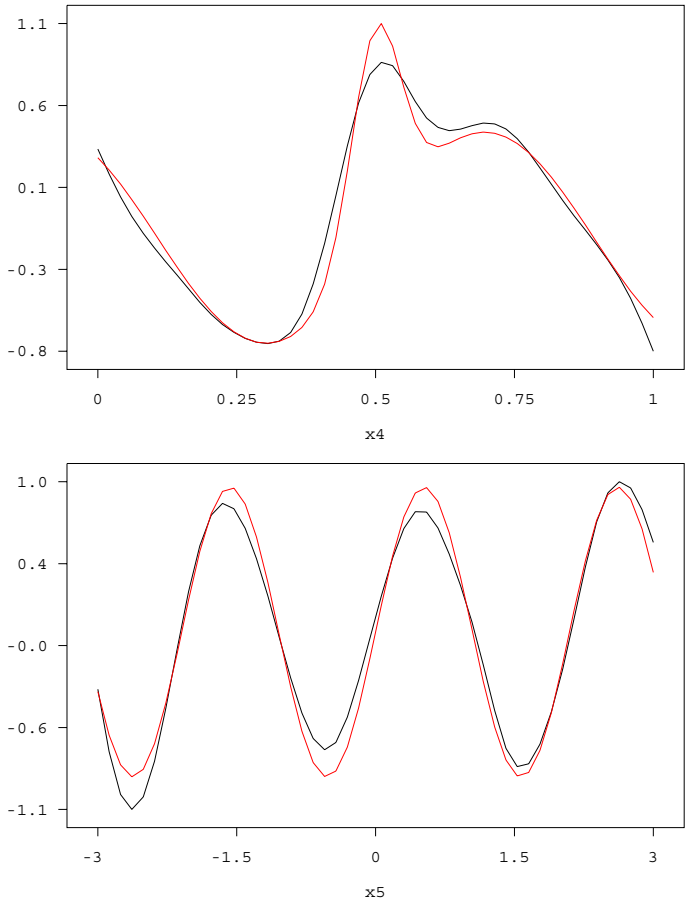

Abbildung 1.7. *Simulierter Datensatz: Geschätzte Funktionen f_4 und f_5. Die wahren Funktionen sind in den Grafiken rot eingezeichnet.*

2

Nichtparametrische Dichteschätzung - Scatterplotsmoother

2.1 Einführung

Gegeben sei eine i.i.d. Stichprobe $x_1, ..., x_n$ einer stetigen Zufallsvariable X mit Dichtefunktion $f(x)$. Ziel ist die Schätzung von f durch $\hat{f}$. Zur Schätzung der Dichte unterscheiden wir grundsätzlich zwei Konzepte:

– **Parametrische Dichteschätzung**
  Hier nehmen wir an, dass die Verteilungsfamilie bekannt ist (z.B. Normalverteilung) und lediglich einige Parameter der Verteilung (z.B. Erwartungswert und Varianz bei der Normalverteilung) unbekannt sind und geschätzt werden müssen. Es gilt also

$$f(x) \in \{f(x \mid \theta), \theta \subset \mathbb{R}^p\},$$

  wobei f nach Schätzung von θ durch $\hat{\theta}$ eindeutig festgelegt ist. Die parametrische Dichteschätzung wird in der Vorlesung Statistik I/II und insbesondere in der Vorlesung Test- und Schätztheorie behandelt. Das Hauptproblem der parametrischen Dichteschätzung ist, dass die Verteilungsklasse (z.B. Normalverteilung) bekannt sein muss. In der Praxis ist diese leider oft nicht bekannt.

– **Nichtparametrische Dichteschätzung**
  Hier wird im wesentlichen nur vorausgesetzt, dass X eine stetige Zufallsvariable ist und die Dichte f eine "glatte" Funktion. Eine bestimmte Verteilungsklasse wird *nicht* vorausgesetzt. Im folgenden sollen das Histogramm und sogenannte Kerndichteschätzer behandelt werden.

2.2 Das Histogramm

Dem Histogramm liegt folgende Idee zugrunde: Zerlege den Datenbereich beginnend im Ursprung x_0 (z.B. $x_0 = 0, x_0 = x_{min} = x_{(1)}$) in Intervalle (sogenannte Bins) gleicher Länge h (sogenannte Binweite). Für den j-ten Bin

$$B_j := [x_0 + (j-1)h, x_0 + jh]$$

gilt

$$P(X \in B_j) = \int\limits_{x_0+(j-1)h}^{x_0+jh} f(x)\,dx. \tag{2.1}$$

Ein naheliegender Schätzer für (2.1) ist die relative Häufigkeit der x_i im Intervall B_j, d.h.

$$P(\widehat{X \in} B_j) = \frac{1}{n}(\#x_i \; in \; B_j) = \frac{1}{n}\sum_{i=1}^{n} I_{B_j}(x_i). \tag{2.2}$$

Weiter folgt nach dem Mittelwertsatz der Integralrechnung (Voraussetzung f stetig)

$$\int\limits_{x_0+(j-1)h}^{x_0+jh} f(x)dx = f(\xi) \cdot h$$

für $\xi \in B_j$. Approximiert man nun f in B_j durch einen konstanten Wert, so erhält man unter Verwendung von (2.2)

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} I_{B_j}(x_i),$$

für $x \in B_j$. Damit erhalten wir folgende Definition:

Definition 2.1 (Histogramm)
Sei $x_1, \ldots, x_n$ eine i.i.d. Stichprobe einer stetigen Zufallsvariable X mit Dichte f. Dann heißt der Schätzer

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j\in Z} I_{B_j}(x_i) I_{B_j}(x)$$

Histogramm mit Klassenbreite (Bandweite) $h > 0$ und Ursprung x_0.

Das Histogramm besitzt folgende Vor- und Nachteile:

Vorteile des Histogramms:

– Einfach zu berechnen und zu präsentieren.
– In jedem Statistikprogramm implementiert.

Nachteile des Histogramms:

– Unstetiger Schätzer für eine stetige Dichte.
– Graphische Darstellung ist abhängig von x_0.
– In ungünstigen Situationen hängt $\hat{f}_h(x)$ mehr von Beobachtungen ab, die weiter von x entfernt sind als von Beoabchtungen, die nahe bei x liegen, vergleiche Abbildung 2.1.

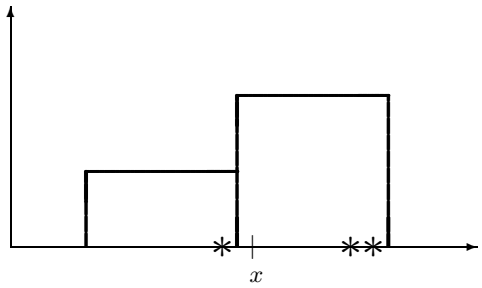

Abbildung 2.1. *Die Grafik zeigt, dass es Fälle geben kann, bei denen weiter entfernte Beobachtungen ein größeres Gewicht bei der Schätzung von f and der Stelle x erhalten als näher liegende Beobachtungen.*

Der Einfluss der Bandweite h lässt sich wie folgt zusammenfassen:

Einfluss der Bandweite h

$h \to 0$	Nadelplot
h klein	sehr rauhe Darstellung, große Datentreue
h groβ	glatte Darstellung, wenig Datentreue
$h \to \infty$	Gleichverteilung

In vielen Programmpaketen wird nicht die Bandweite h spezifiziert, sondern die *Anzahl der Intervalle* (Bins). Diese Anzahl induziert dann eine bestimmte Bandweite.

Zum Einfluss der Bandweite h bzw. der Anzahl der Intervalle vergleiche die beiden folgenden Beispiele.

Beispiel 2.1 (Mietspiegel)
Abbildung 2.2 zeigt verschiedene Dichteschätzer für die Nettomiete pro Quadratmeter im Mietspiegeldatensatz. Der Einfluss der Bandweite auf die Schätzungen ist hier relativ gering.

△

Beispiel 2.2 (Mischung aus Normalverteilungen)
Abbildung 2.3 zeigt für einen simulierten Datensatz Dichteschätzer mit unterschiedlichen Bandweiten. Es wurden 100 Beobachtungen simuliert aus der Dichte

$$f(x) = 0.6 \cdot f_1(x) + 0.4 \cdot f_2(x).$$

Dabei ist f_1 die Dichte einer Normeilverteilung mit $\mu = -1$ und $\sigma^2 = 1$ und f_2 die Dichte einer Normalverteilung mit $\mu = 2$ und $\sigma^2 = 1$. Es handelt sich bei f also um eine Mischung aus zwei Normalverteilungsdichten. Die wahre Dichte ist in Abbildung a) zu finden, die Abbildungen b) - f) zeigen Histogramme mit unterschiedlicher Klassenbreite. Hier ist der Einfluss der Bandweite auf die Schätzungen relativ groß.

△

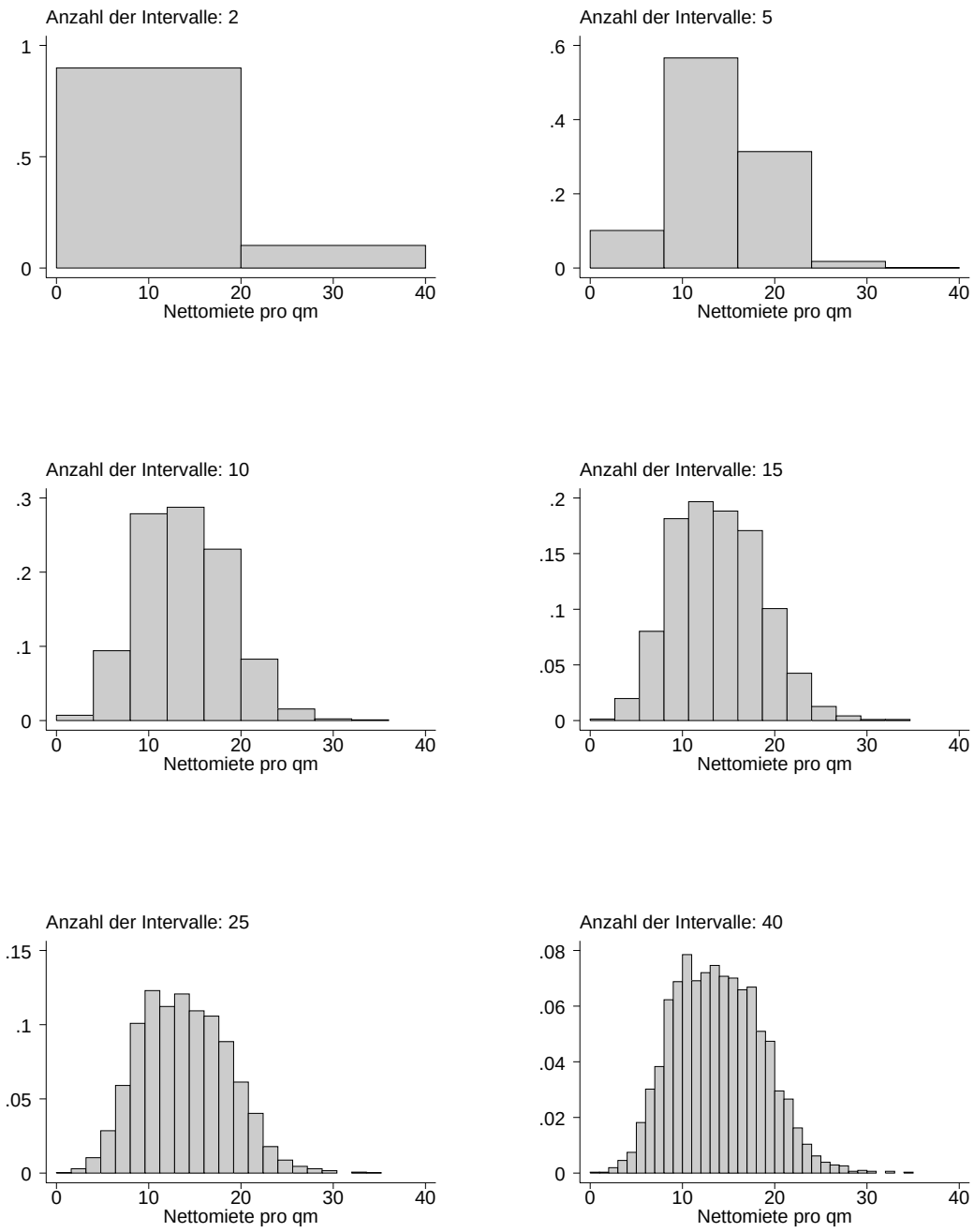

Abbildung 2.2. *Einfluß der Bandweite beim Histogramm: Die Grafiken a) - f) zeigen Histogramme mit unterschiedlichen Bandweiten für die Nettomiete pro qm.*

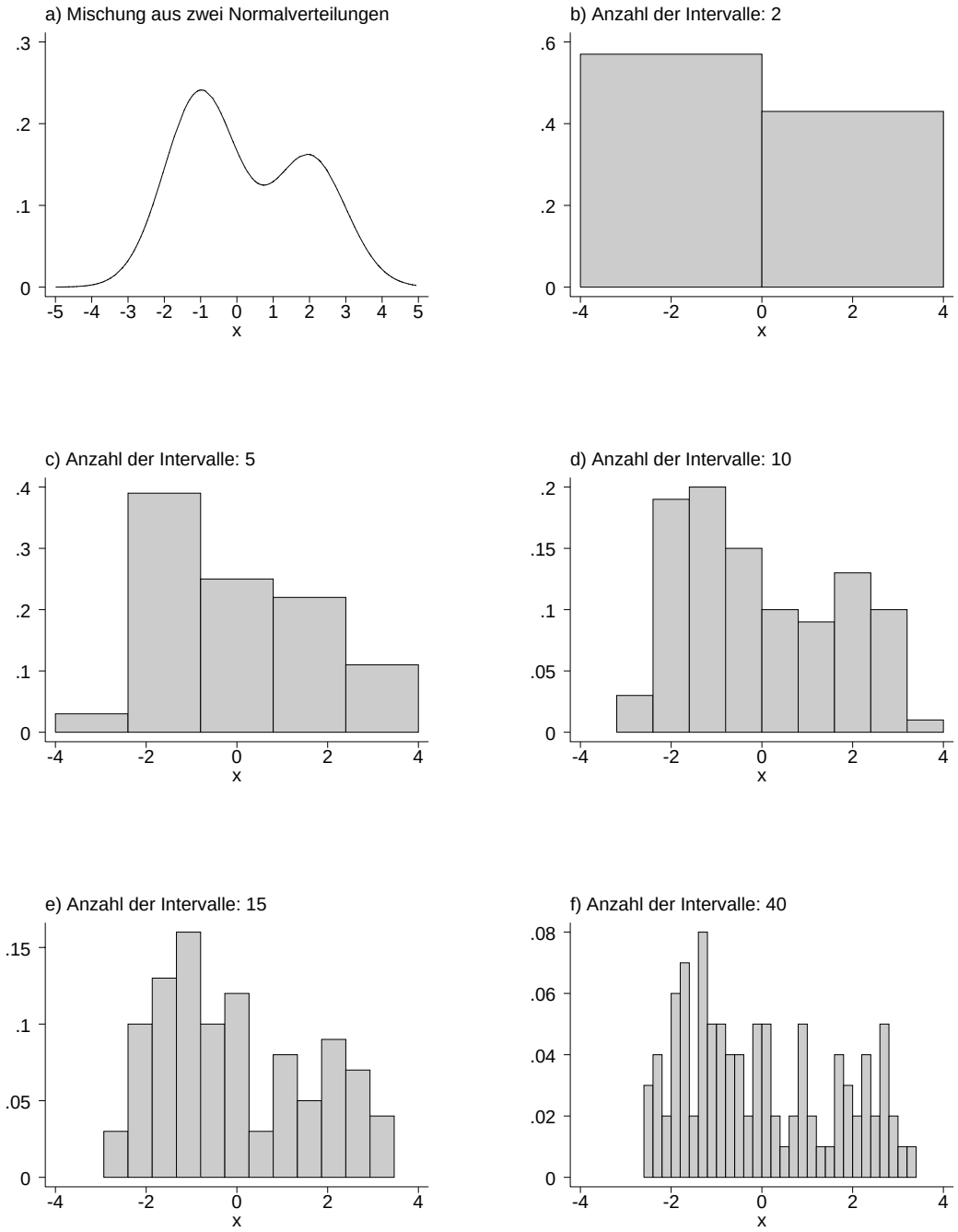

Abbildung 2.3. *Einfluss der Bandweite beim Histogramm: Grafik a) zeigt die wahre Dichte. Die Grafiken b) - f) zeigen Histogramme mit unterschiedlichen Bandweiten. Grundlage der Schätzungen sind 100 simulierte Beobachtungen gemäß der wahren Dichte in a).*

2.3 Kerndichteschätzer

Die im letzten Abschnitt genannten Probleme beim Histogramm können wir durch sogenannte *gleitende Histogramme* umgehen:

Definiere Intervalle $[x-h; x+h]$ der Breite $2h$ und lasse diese über die x-Achse "gleiten". Damit erhalten wir als Schätzer für f das gleitende Histogramm

$$\hat{f}_h(x) = \frac{1}{2nh}(\#x_i \text{ im Intervall } [x-h; x+h]) \tag{2.3}$$

Mit der " Kernfunktion"

$$K(u) = \begin{cases} \frac{1}{2} & |u| \leq 1 \\ 0 & \text{sonst} \end{cases}, \quad \text{(Rechteckskern)}$$

erhalten wir für (2.3)

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Eine naheliegende Verallgemeinerung des gleitenden Histogramms erhalten wir, indem wir andere Kernfunktionen als den Rechteckskern zulassen. Wir ersetzen also den Rechteckskern durch allgemeine Kernfunktionen, die folgende Eigenschaften besitzen sollen:

1. $K(u) = K(-u)$ (Symmetrie um Null);
2. $\arg\max K(u) = 0$ (Maximum bei $u = 0$);
3. $\int K(u)du = 1$;
4. $K(u) \geq 0$.
5. $|u|K(u) \to 0$ für $|u| \to \infty$
6. $K(u)$ beschränkt.
7. $\int u^2 K(u)\, du < \infty$

Die Eigenschaften 5-7 sind eher technischer Natur und werden bei asymptotischen Aussagen zum Kerndichteschätzer benötigt, vergleiche Abschnitt 2.4.2. Beispiele für Kernfunktionen neben dem Rechteckskern sind (vgl. auch Abbildung 2.7):

– Dreieckkern: $K(u) = (1 - |u|)I_{[-1,1]}(u)$

– Epanechnikovkern: $K(u) = \frac{3}{4}(1 - u^2)I_{[-1,1]}(u)$

– Normalkern: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$

Damit erhalten wir

Definition 2.2 (Kerndichteschätzer)
Der Schätzer

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

mit

$$K_h(u) := \frac{1}{h} K\left(\frac{u}{h}\right)$$

heißt Kerndichteschätzer mit Kern K (bzw. K_h) und Bandweite $h > 0$.

Die Abbildungen 2.4 - 2.6 illustrieren die Berechnung des Kerndichteschätzers. Abbildung 2.4 enthält 5 Beoachtungen (dargestellt als Kreise) und die dazugehörigen (normierten) Kernfunktionen $1/5K_h(x - x_i)$. Der Kerndichteschätzer $\hat{f}_h$ an einer Stelle x ist nichts anderes als die Summe der 5 (normierten) Kernfunktionen an dieser Stelle. Dabei gehen Kernfunktionen, deren zugehörige Beobachtung näher an x liegt, mit höherem Gewicht ein. Die Abbildungen 2.5 und 2.6 veranschaulichen, wie sich der Kerndichteschätzer ändert, wenn die Bandweite variiert wird. Im Vergleich zu Abbildung 2.4 sind die Kernfunktionen in Abbildung 2.5 wegen der kleineren Bandweite enger und höher. Die geschätzte Dichte wird rauher. in Abbildung 2.6 sind im Vergleich zu Abbildung 2.4 die Kernfunktionen wegen der größeren Bandweite weiter und flacher. Die geschätzte Dichte wird glatter.

Bemerkungen:

– Aus

$$\int K(u)du = 1$$

folgt auch

$$\int \hat{f}_h(x)dx = 1,$$

d.h. $\hat{f}_h$ erfüllt die Voraussetzungen an einen Dichteschätzer.

– Der Schätzer $\hat{f}_h(x)$ "erbt" die Eigenschaften des verwendeten Kerns. D.h. wenn K stetig (stetig differenzierbar etc.) ist, übertragen sich diese Eigenschaften auf $\hat{f}_h(x)$.

Den Einfluss der Bandweite h können wir wie folgt zusammenfassen:

$h \to 0$	Nadelplot
h klein	rauhes Bild, relativ datentreu
h groß	glattes Bild, weniger datentreu
$h \to \infty$	sehr glatte Schätzung, etwa Form von K

Beispiel 2.3 (Mietspiegel)
In Abbildung 2.8 sind Kerndichteschätzer für die Nettomiete pro Quadratmeter im Mietspiegelbeispiel für verschiedene Bandweiten abgebildet. Als Kernfunktion wurde der Epanechnikovkern verwendet. Die in Abbildung 2.8 d) verwendete Bandweite ist in gewissem Sinne optimal, vgl. Abschnitt 2.4.3.

△

Beispiel 2.4 (Mischung aus Normalverteilungen)
Abbildung 2.9 zeigt Kerndichteschätzer für den simulierten Datensatz (Mischung aus zwei Normalverteilungen) aus Beispiel 2.2. Als Kernfunktion wurde der Epanechnikovkern verwendet. Ähnlich zum Histogramm hängen die Schätzer in erheblichem Maß von der verwendeten Bandweite ab. Die in Abbildung 2.9 d) verwendete Bandweite ist in gewissem Sinne optimal, vgl. Abschnitt 2.4.3.

△

Zur Bestimmung von möglichst optimalen Bandweiten, bestimmen wir im nächsten Abschnitt zunächst statistische Eigenschaften von Kerndichteschätzern.

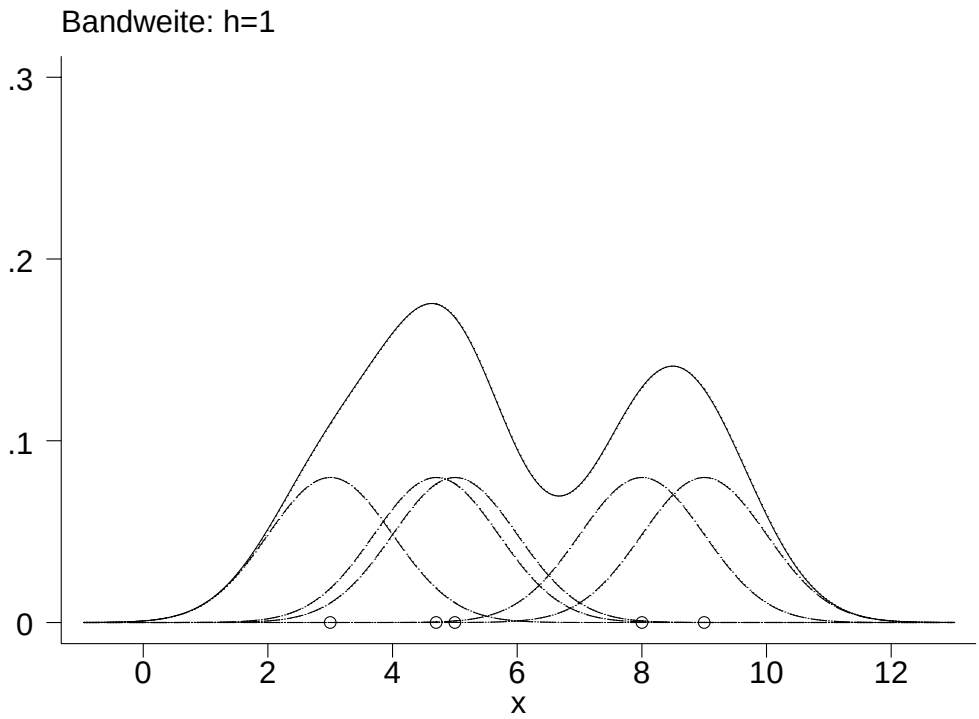

Abbildung 2.4. *Illustration zur Berechnung des Kerndichteschätzers.*

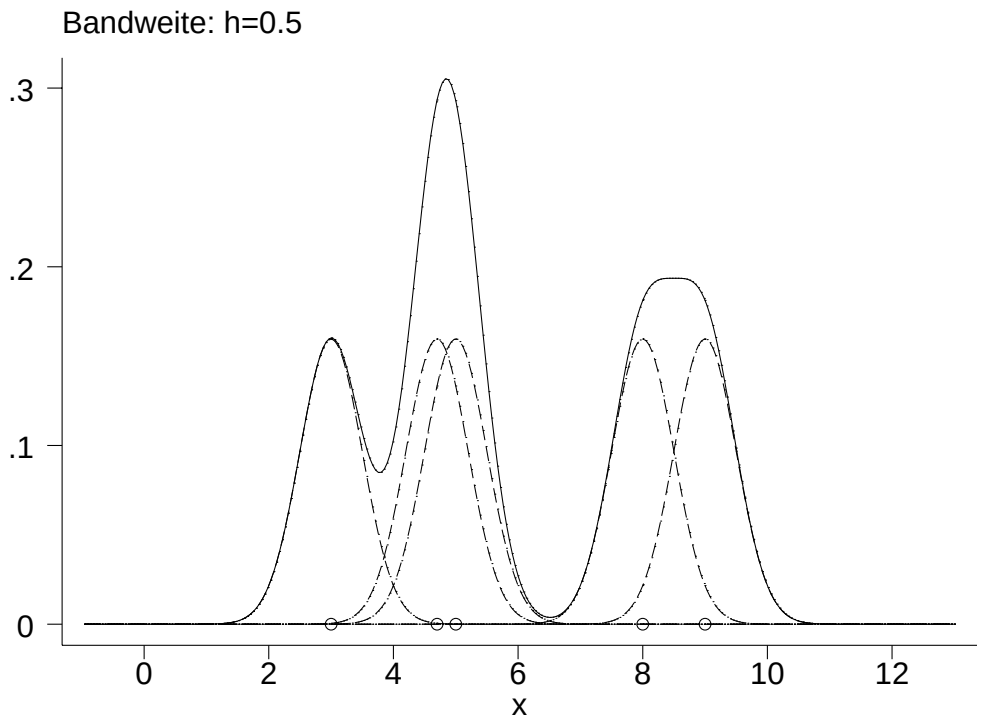

Abbildung 2.5. *Illustration zur Berechnung des Kerndichteschätzers. Im Vergleich zu Abbildung 2.4 sind die Kernfunktionen wegen der kleineren Bandweite enger und höher. Die geschätzte Dichte wird rauher.*

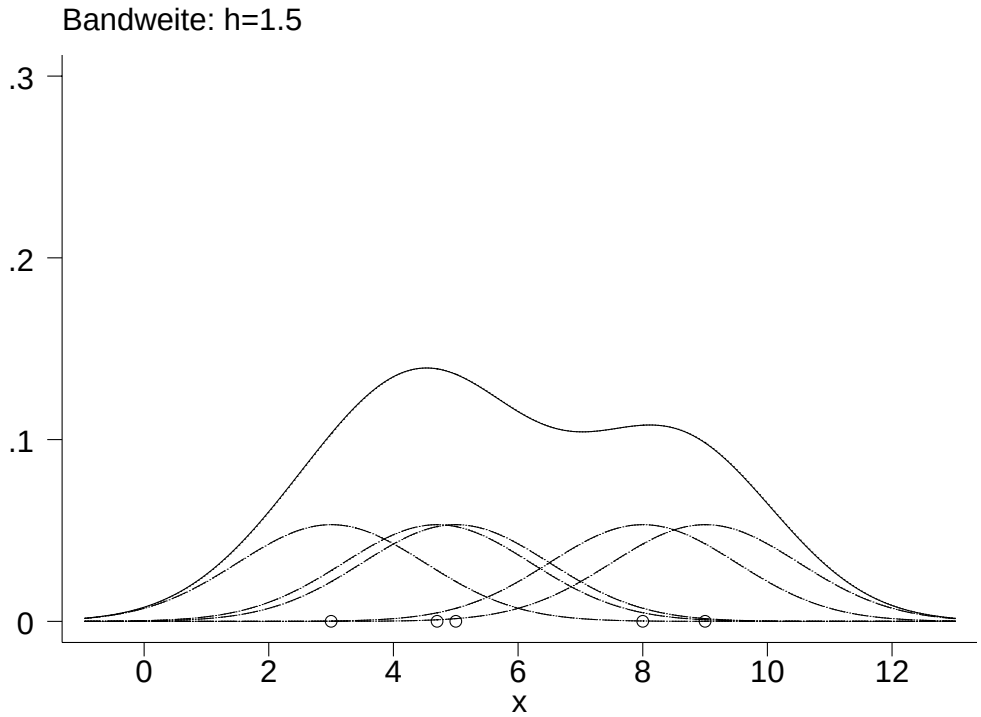

Abbildung 2.6. *Illustration zur Berechnung des Kerndichteschätzers. Im Vergleich zu Abbildung 2.4 sind die Kernfunktionen wegen der größeren Bandweite weiter und flacher. Die geschätzte Dichte wird glatter.*

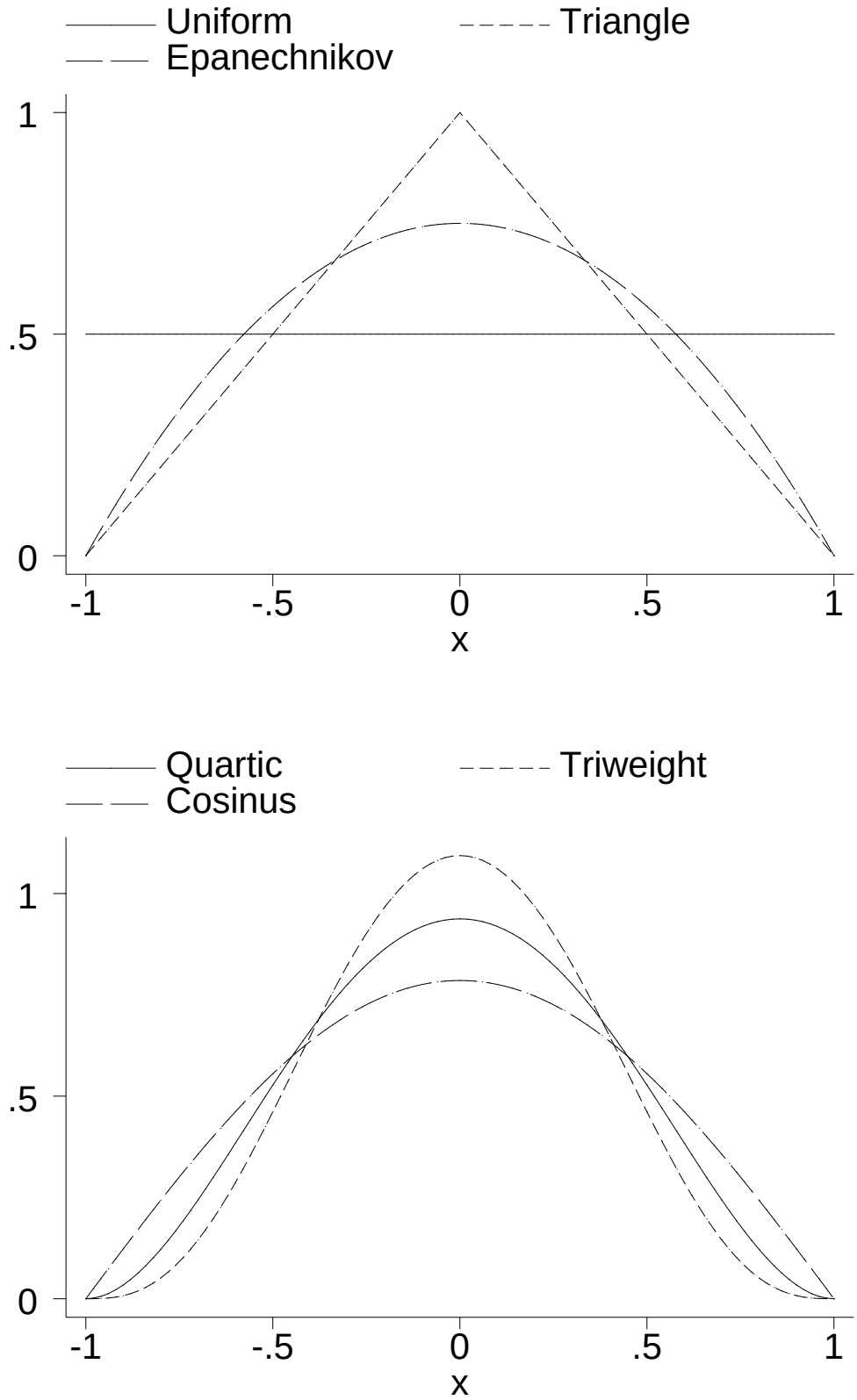

Abbildung 2.7. *Grafische Darstellung verschiedener Kerne.*

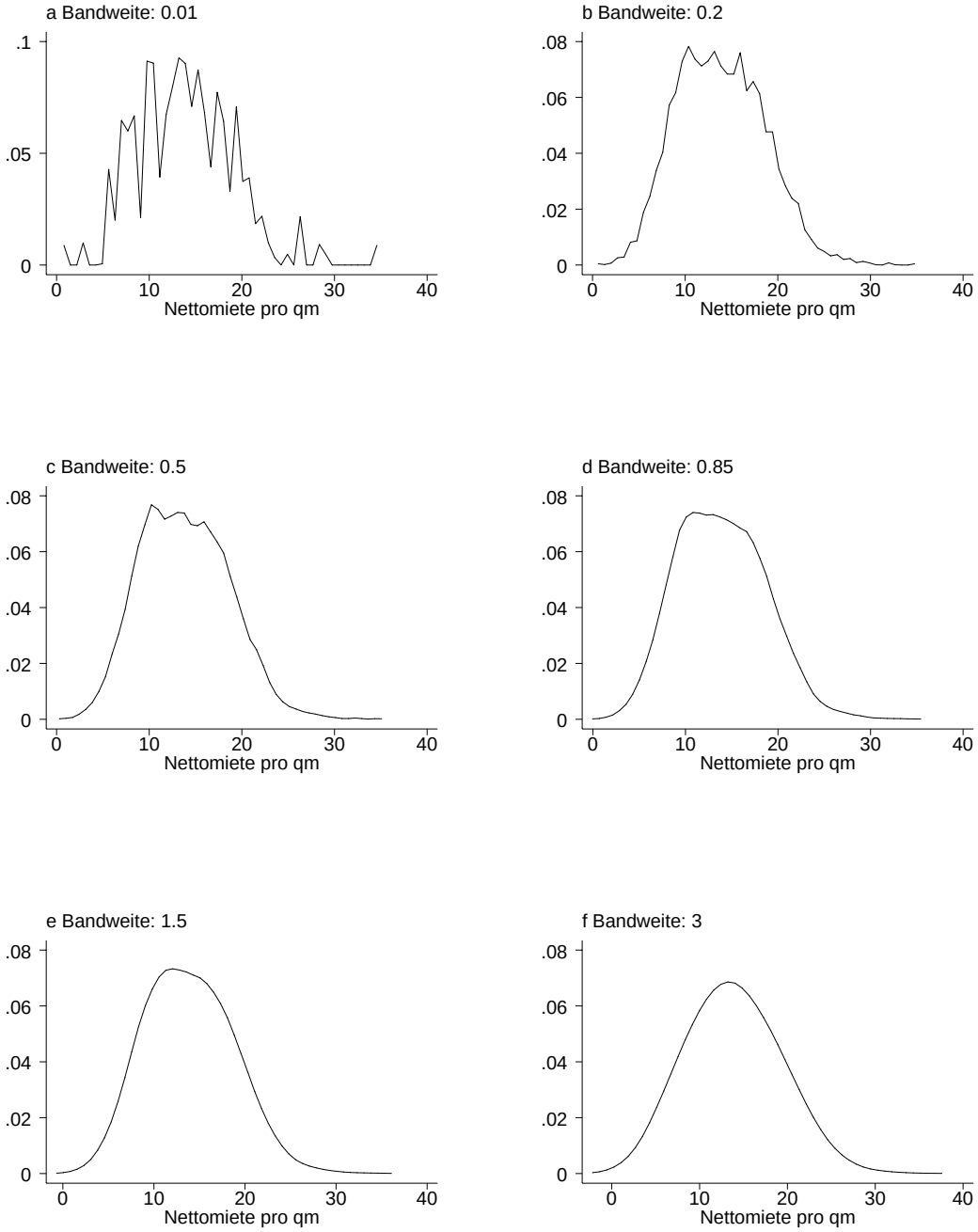

Abbildung 2.8. *Einfluss der Bandweite beim Kerndichteschätzer: Die Grafiken a) - f) zeigen Kerndichteschätzer mit unterschiedlichen Bandweiten für die Nettomiete pro qm. Die AMISE ,,optimale" Bandweite ist ungefähr* $h = 0.85$. *Als Kernfunktion wurde der Epanechnikovkern verwendet.*

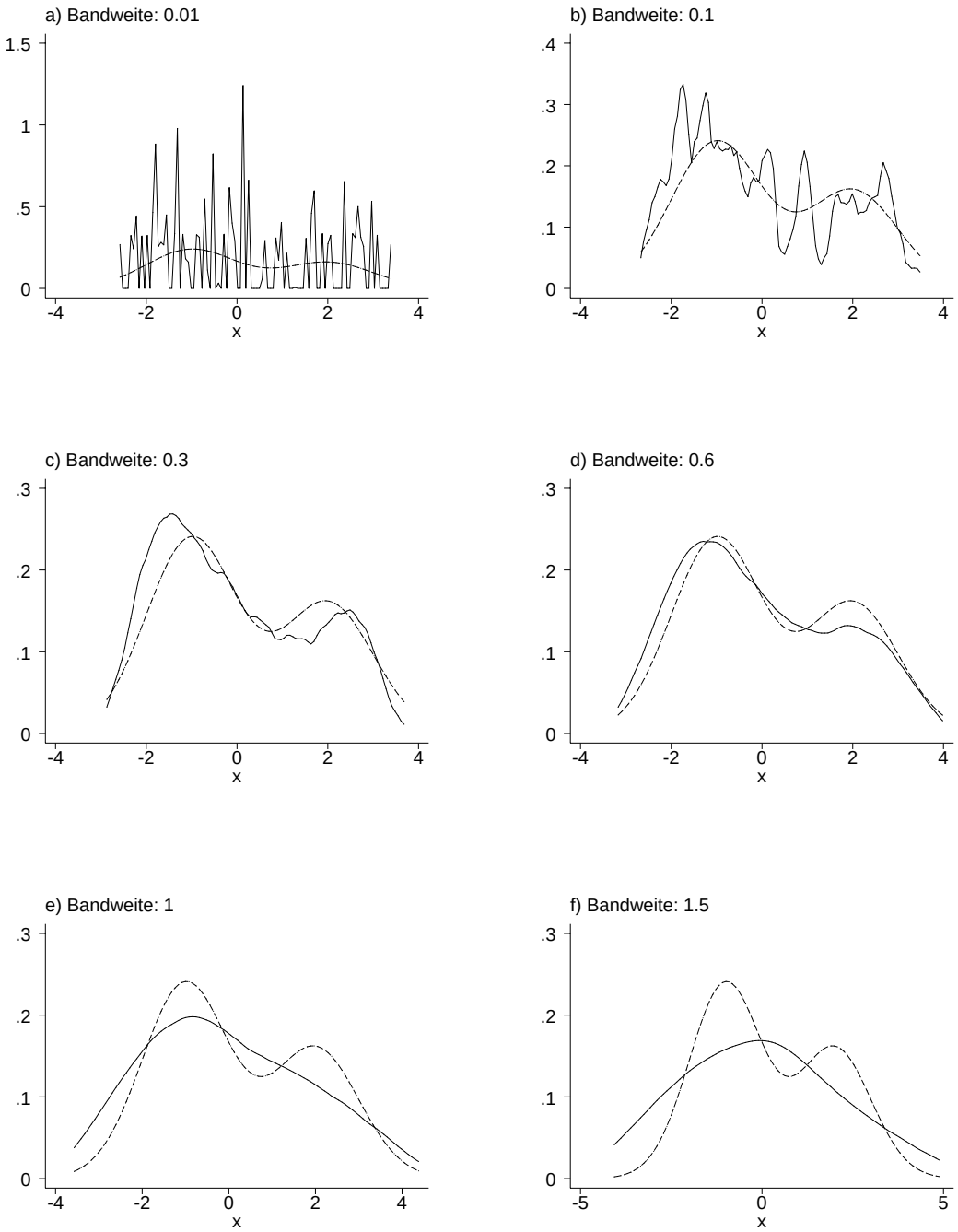

Abbildung 2.9. *Einfluss der Bandweite beim Kerndichteschätzer: Die Grafiken a) - f) zeigen Kerndichteschätzer für x mit unterschiedlichen Bandweiten. Grundlage der Schätzungen sind 100 simulierte Beobachtungen gemäß der wahren Dichte (gestrichelte Linien). Die AMISE ,,optimale" Bandweite ist ungefähr $h = 0.6$. Als Kernfunktion wurde der Epanechnikovkern verwendet.*

2.4 Statistische Eigenschaften des Kerndichteschätzers

2.4.1 Erwartungswert, Varianz und MSE

Wir berechnen zunächst den Erwartungswert von $\hat{f}_h$ für festes x:

$$\begin{aligned} E(\hat{f}_h(x)) &= E\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} E\left(K\left(\frac{x-x_i}{h}\right)\right) \\ &\overset{iid}{=} \frac{1}{nh} nE\left(K\left(\frac{x-X}{h}\right)\right) \\ &= \frac{1}{h}\int_{\mathbf{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy. \end{aligned}$$

Wir bestimmen jetzt die Varianz des Kerndichteschätzers. Zunächst stellen wir fest, dass

$$E\left(K\left(\frac{x-X}{h}\right)\right) = hE(\hat{f}_h(x)),$$

vergleiche die Berechnungen zum Erwartungswert. Damit erhalten wir

$$\begin{aligned} Var(\hat{f}_h(x)) &= Var\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right) \\ &\overset{iid}{=} \frac{1}{n^2h^2}\sum_{i=1}^{n} Var\left(K\left(\frac{x-x_i}{h}\right)\right) \\ &\overset{iid}{=} \frac{1}{n^2h^2} nVar\left(K\left(\frac{x-X}{h}\right)\right) \\ &= \frac{1}{nh^2}\left[E\left(K^2\left(\frac{x-X}{h}\right)\right) - \left(EK\left(\frac{x-X}{h}\right)\right)^2\right] \\ &= \frac{1}{nh^2}E\left(K^2\left(\frac{x-X}{h}\right)\right) - \frac{1}{n}(E\hat{f}_h(x))^2 \\ &= \frac{1}{nh^2}\int_{\mathbf{R}} K^2\left(\frac{x-y}{h}\right) f(y)\,dy - \frac{1}{n}E(\hat{f}_h(x))^2. \end{aligned}$$

Mit Hilfe des Erwartungswerts und der Varianz können wir auch den Mean Squared Error (MSE) von $\hat{f}_h$ an der Stelle x berechnen. Zunächst erhalten wir für den Bias

$$Bias(\hat{f}_h(x)) = E(\hat{f}_h(x)) - f(x) = \frac{1}{h}\int_{\mathbf{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x).$$

Damit folgt

$$\begin{aligned} MSE(\hat{f}_h(x)) &= Var(\hat{f}_h(x)) + Bias^2(\hat{f}_h(x)) \\ &= \frac{1}{nh^2}\int_{\mathbf{R}} K^2\left(\frac{x-y}{h}\right) f(y)\,dy - \frac{1}{n}E(\hat{f}_h(x))^2 \\ &\quad + \left(\frac{1}{h}\int_{\mathbf{R}} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x)\right)^2 \end{aligned}$$

Bei den bisher betrachteten Größen handelt es sich ausschließlich um lokale Maße, d.h. sie hängen von x ab. Ein globales Maß ist der sogenannte Mean Integrated Squared Error (MISE). Der MISE ist definiert als

$$MISE(\hat{f}_h) = \int MSE(\hat{f}_h(x))\,dx.$$

Im Gegensatz zum MSE hängt der MISE nur noch von der Bandweite h (und der unbekannten Dichte) ab, jedoch nicht mehr von x. Damit erscheint der MISE als ein geeignetes Maß zur Bestimmung einer möglichst optimalen Bandweite. Bevor wir jedoch zur Bestimmung einer optimalen Bandweite kommen, beschäftigen wir uns im nächsten Abschnitt mit der Frage der Konsistenz von $\hat{f}_h(x)$.

2.4.2 Konsistenz des Kerndichteschätzers

Bekanntlich ist ein Schätzer dann konsistent, wenn der MSE gegen Null konvergiert. Wir müssen also zeigen, dass $\hat{f}_h$ asymptotisch erwartungstreu ist und die Varianz gegen Null konvergiert.

Wir benötigen folgenden

Satz 2.1 (von Parzen)
Sei $R(x)$, $x \in \mathbf{R}$ eine (messbare) Funktion mit den Eigenschaften

1. $\sup_{x\in\mathbf{R}} |R(x)| < \infty$ *(d.h. $R(x)$ ist beschränkt)*
2. $\int |R(x)|\,dx < \infty$
3. $|x|R(x) \to 0$.

Sei weiterhin $g(x)$, $x \in \mathbf{R}$, eine (messbare) Funktion mit $\int |g(x)|\,dx < \infty$. Betrachte die Folge

$$g_n(x) = \frac{1}{h_n}\int R\left(\frac{x-y}{h_n}\right) g(y)\,dy$$

wobei h_n eine Folge ist mit $\lim_{n\to\infty} h_n = 0$. Dann gilt für jeden Stetigkeitspunkt x von g

$$g_n(x) \to g(x) \int R(s)\, ds$$

falls $n \to \infty$.

Beweis:
Den vollständigen Beweis des Satzes findet man in Parzen (1962). Wenn man zusätzlich annimmt, dass g beschränkt ist, kann man den Beweis relativ leicht führen. Die Aussage folgt dann aus dem Satz von der majorisierten Konvergenz (vergleiche zum Beispiel Gänssler und Stute (1977)). Sei a_n eine Folge integrierbarer und beschränkter Funktionen mit integrierbarer Grenzfunktion. Dann kann man gemäß dem Satz von der majorisierten Konvergenz Integration und Grenzwertbildung vertauschen, d.h.

$$\lim_{n\to\infty} \int a_n(x)\, dx = \int \lim_{n\to\infty} a_n(x)\, dx.$$

Unter Zuhilfenahme dieser Aussage erhalten wir

$$\begin{aligned}
\lim_{n\to\infty} g_n(x) &= \lim_{n\to\infty} \frac{1}{h_n} \int R\left(\frac{x-y}{h_n}\right) g(y)\, dy \\
&= \lim_{n\to\infty} \int R\,(s)\, g(x - sh_n)\, ds \\
&= \int \lim_{n\to\infty} R\,(s)\, g(x - sh_n)\, ds \\
&= g(x) \textstyle\int R(s)\, ds.
\end{aligned}$$

Dabei haben wir in der zweiten Zeile die Substitution $s = (x-y)/h_n$ vorgenommen. Eine Voraussetzung für die Anwendbarkeit des Satzes von der majorisierten Konvergenz ist die Beschränktheit von $R\,(s)\, g(x - sh_n)$, was nach Vorraussetzung erfüllt ist.

Mit Hilfe des Satzes von Parzen erhalten wir folgenden

Satz 2.2 (Konsistenz des Kerndichteschätzers)
Sei f stetig. Dann gilt

$$E(\hat{f}_{h_n}(x)) \to f(x)$$

falls die Bandweite h_n für $n \to \infty$ gegen Null konvergiert. $\hat{f}_{h_n}(x)$ ist also erwartungstreu. Falls $nh_n \to \infty$ für $n \to \infty$, dann gilt

$$Var(\hat{f}_{h_n}(x)) \to 0.$$

Damit ist $\hat{f}_{h_n}(x)$ konsistent.

Beweis:
Zum Beweis der Erwartungstreue wenden wir Satz 2.1 an mit $R(x) = K(x)$ und

$$g_n(x) = E(\hat{f}_{h_n}(x)) = \frac{1}{h_n} \int K\left(\frac{x-y}{h_n}\right) f(y)\, dy.$$

Aufgrund des Satzes folgt

$$g_n(x) \to f(x) \int K(s)\, ds = f(x).$$

Zum Beweis der zweiten Aussage wenden wir wiederum Satz 2.1 an mit $R(x) = K^2(x)$ und

$$g_n(x) = \frac{1}{h_n} \int K^2\left(\frac{x-y}{h_n}\right) f(y)\, dy.$$

Es folgt

$$g_n(x) \to f(x) \int K^2(s)\, ds.$$

Wegen

$$Var(\hat{f}_{h_n}(x)) = \frac{1}{nh_n^2} \int K^2\left(\frac{x-y}{h_n}\right) f(y)\, dy - \frac{1}{n} E(\hat{f}_{h_n}(x))^2$$

erhalten wir

$$0 \leq Var(\hat{f}_{h_n}(x)) \leq \frac{1}{nh_n} \frac{1}{h_n} \int K^2\left(\frac{x-y}{h_n}\right) f(y)\, dy = \frac{1}{nh_n} g_n(x) \to 0.$$

2.4.3 Konvergenzordnung des MISE

Ein naheliegendes Optimalitätskriterium zur Wahl der Bandweite h beim Kerndichteschätzer ist der Mean Integrated Squared Error MISE. Der MISE ist definiert als

$$MISE(\hat{f}_h) = \int MSE(\hat{f}_h(x))\, dx = \int Var(\hat{f}_h(x))\, dx + \int Bias^2(\hat{f}_h(x))\, dx.$$

Zur Bestimmung der Konvergenzordnung des MISE benötigen wir zunächst die sogennanten Landau Symbole (bzw. die Notation groß O und klein o):

Definition 2.3 (Landau Symbole)
Gegeben seien die reellwertigen Reihen $\{a_n\}$ und $\{b_n\}$ mit $n \in \mathbb{N}$. Wir schreiben

$$a_n = O(b_n)$$

falls der Quotient

$$\left|\frac{a_n}{b_n}\right|$$

für $n \to \infty$ beschränkt ist. (Sprechweise: a_n ist groß O von b_n.) Die Reihe $\{a_n\}$ ist also ungefähr von der selben Größenordnung wie $\{b_n\}$. Offenbar bedeutet $a_n = O(1)$, dass a_n beschränkt ist.

Wir schreiben

$$a_n = o(b_n),$$

falls der Quotient

$$\left|\frac{a_n}{b_n}\right|$$

für $n \to \infty$ gegen Null konvergiert. (Sprechweise: a_n ist klein o von b_n.) Die Reihe $\{a_n\}$ ist also von geringerer Ordnung als $\{b_n\}$ (konvergiert schneller gegen Null). Offenbar bedeutet $a_n = o(1)$ nichts anderes als

$$\lim_{n\to\infty} a_n = 0.$$

Nach diesen Vorbemerkungen kommen wir jetzt wieder zurück auf die Bestimmung der Konvergenzordnung des MISE. Es gilt der

Satz 2.3
Sei f mindestens zweimal stetig differenzierbar, f'' beschränkt, f und f'' quadratintegrierbar. Sei h_n eine Folge mit $h_n \to 0$. Unter Verwendung der Abkürzungen $\int g^2(s)\,ds = ||g||_2^2$ und $\mu_2(g) = \int g(s)s^2\,ds$ für eine Funktion g gilt:

1. $Var(\hat{f}_{h_n}(x)) = \dfrac{1}{nh_n}||K||_2^2 f(x) + o\left(\dfrac{1}{nh_n}\right)$ *bzw.*

   $$\int Var(\hat{f}_{h_n}(x))\,dx = \frac{1}{nh_n}||K||_2^2 + o\left(\frac{1}{nh_n}\right).$$

2. $Bias(\hat{f}_{h_n}(x)) = \dfrac{{h_n}^2}{2}\mu_2(K)f''(x) + o({h_n}^2)$ *bzw.*

   $$\int Bias^2(\hat{f}_{h_n}(x))\,dx = \frac{{h_n}^4}{4}\mu_2^2(K)||f''||_2^2 + o({h_n}^4).$$

3. $MISE(\hat{f}_{h_n}) = \dfrac{1}{nh_n}||K||_2^2 + \dfrac{{h_n}^4}{4}\mu_2^2(K)||f''||_2^2 + o(\dfrac{1}{nh_n} + {h_n}^4).$

Beweis: Pruscha (2000).

Aufgrund des Satzes stellen wir also folgendes fest:

– Der Bias ist umso kleiner, je kleiner h gewählt wird. Andererseits wird die Varianz kleiner, je größer h wird. Es gibt also einen Zielkonflikt zwischen der Reduzierung der Varianz und des Bias (*Bias-Varianz Trade-off*).

– Der Bias hängt von $f''(x)$ ab, was ein Maß für die Krümmung von f ist. Je stärker die Krümmung, desto größer der Bias. Damit erhalten wir einen positiven Bias bei lokalen Minima der Dichte und einen negativen Bias bei lokalen Maxima der Dichte, vergleiche auch Abbildung 2.10.

– Bias und Varianz hängen auch vom gewählten Kern K ab, i.d.R. verändern andere Kerne den Bias aber nur unwesentlich.

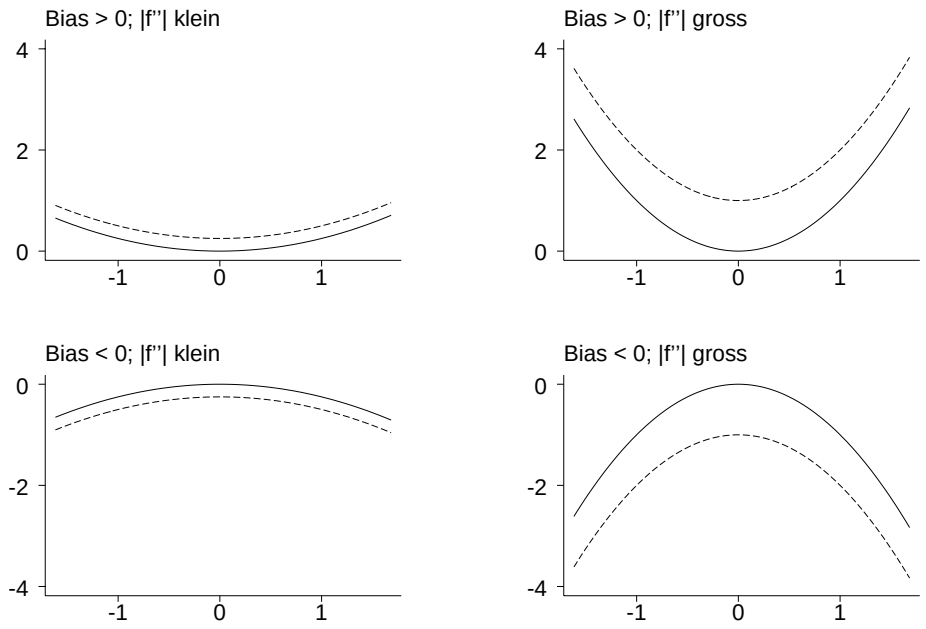

Abbildung 2.10. *Veranschaulichung des Bias in Abhängigkeit der Krümmung der Dichte. Wir erhalten einen positiven Bias bei lokalen Minima und einen negativen Bias bei lokalen Maxima der Dichte.*

Zur Berechnung einer optimalen Bandweite minimieren wir den sogenannten AMISE (Asymptotic Mean Integrated Squared Error), der aus dem MISE durch Streichung der o-Terme entsteht, d.h.

$$AMISE(\hat{f}_h) = \frac{1}{nh}||K||_2^2 + \frac{h^4}{4}\mu_2^2(K)||f''||_2^2 \tag{2.4}$$

Durch Differenzieren und Nullsetzen erhalten wir die AMISE optimale Bandweite

$$h_0 = \left(\frac{\|K\|_2^2}{\|f''\|_2^2\mu_2^2(K)n}\right)^{\frac{1}{5}} \tag{2.5}$$

Offensichtlich besteht das Problem, dass die optimale Bandweite zur Schätzung von f von Funktionalen von f abhängt (circulus virtuosis). In der Praxis (z.B. in STATA) setzt man daher eine Referenzdichte ein. Nehmen wir z.B. eine Normalverteilung an, dann können wir $\|f''\|_2^2$ schätzen (nachdem vorher die Varianz σ^2 durch den üblichen Schätzer $\hat{\sigma}^2$ geschätzt wurde). Unter Verwendung des Normalkerns erhalten wir als "optimale" Bandweite

$$\hat{h}_0 = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\,\hat{\sigma}\,n^{-\frac{1}{5}}.$$

Ein weniger ausreißeranfälliger Schätzer für σ basiert auf dem sogenannten Interquartilsabstand $\hat{R} = x_{(0.75n)} - x_{(0.25n)}$. Damit erhalten wir als neue Faustregel für h_0

$$\hat{h}_0 = 0.79\,\hat{R}\,n^{-\frac{1}{5}}.$$

Man beachte, dass $\hat{R} \approx 1.34\hat{\sigma}$ (falls als Referenzdichte eine Normalverteilung zugrundegelegt wird). Eine Kombination beider Regeln liefert

$$\hat{h}_0 = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-\frac{1}{5}}.$$

Unter Verwendung des Epanechnikov Kerns erhalten wir

$$\hat{h}_0 = 0.9 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-\frac{1}{5}}.$$

Als "Nebenprodukt" der AMISE optimalen Bandweitenwahl können wir die Konvergenzgeschwindigkeit bestimmen, mit welcher der AMISE gegen Null geht. Einsetzen von (2.5) in den AMISE (2.4) liefert

$$AMISE(\hat{f}_{h_0}) = \frac{5}{4}\|K\|_2^{2\frac{4}{5}} \left(\mu_2(K)\,\|f''\|_2^2\right)^{\frac{2}{5}} n^{-\frac{4}{5}}. \tag{2.6}$$

Für wachsendes n wird der AMISE mit der Rate $n^{-\frac{4}{5}}$ kleiner. Beim Histogramm wird der AMISE nur mit einer Rate von $n^{-\frac{2}{3}}$ kleiner, d.h. Kerndichteschätzer haben eine höhere Konvergenzgeschwindigkeit als Histogramme.

Wir stellen fest, dass im Ausdruck (2.6) ein Faktor

$$F(K) = (||K||_2^2)^{\frac{4}{5}} \mu_2(K)^{\frac{2}{5}}$$

vorkommt, der nur vom Kern K abhängt. Durch Minimierung dieses Faktors bezüglich K können wir einen in gewissem Sinne optimalen Kern bestimmen. Man kann zeigen, dass der Epanechnikov Kern den Faktor $F(K)$ minimiert.

2.4.4 Optimale Bandweite durch Kreuzvalidierung

Wir unterscheiden ML-Kreuzvalidierung (Härdle (1990), Seite 92 ff.) und Least-Squares Kreuzvalidierung. Hier beschränken wir uns auf die Least-Squares Kreuzvalidierung. Betrachte als Maß für den Unterschied zwischen $\hat{f}$ und f den Integrated Squared Error (ISE)

$$ISE(h) = \int (\hat{f}_h(x) - f(x))^2\,dx = \int \hat{f}_h^2(x)\,dx - 2\int \hat{f}_h(x) f(x)\,dx + \int f^2(x)\,dx.$$

Wir versuchen im folgenden $ISE(h)$ bzgl. h zu minimieren. Der erste Ausdruck $\int \hat{f}_h^2(x)\,dx$ kann leicht berechnet werden, den letzten Ausdruck können wir weglassen, weil er nicht von h abhängt. Für den mittleren Ausdruck gilt zunächst

$$\int \hat{f}_h(x) f(x)\, dx = E_X \hat{f}_h(X)$$

wobei der Erwartungswert bzgl. einer zusätzlichen und unabhängigen Beobachtung X gebildet wird. Zur Schätzung dieses Erwartungwerts verwenden wir den sogenannten "leave one out" Schätzer:

$$E_X \widehat{\hat{f}_h}(X) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{h,i}(x_i)$$

wobei

$$\hat{f}_{h,i}(x_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right)$$

der Kerndichteschätzer an der Stelle x_i ist, bei dem x_i nicht berücksichtigt wurde. Insgesamt wird also die Kreuzvalidierungsfunktion

$$CV(h) = \int (\hat{f}_h^2(x)) dx \; - \; \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{h,i}(x_i) \tag{2.7}$$

bzgl. h minimiert.

Das Integral in (2.7) kann analytisch berechnet werden. Dazu verwenden wir die Faltung einer Funktion f, die definiert ist als

$$f \star f(x) = \int f(x-y) f(y) dy.$$

Damit erhalten wir

$$\begin{aligned}
\int \hat{f}_h^2(x) dx &= \frac{1}{n^2 h^2} \int \left(\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \right)^2 dx \\
&= \frac{1}{n^2 h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int K\left(\frac{x - x_i}{h}\right) K\left(\frac{x - x_j}{h}\right) dx \\
&= \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} \int K(s)\, K\left(\frac{x_i - x_j}{h} + s\right) ds \\
&= \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} \int K(s)\, K\left(\frac{x_j - x_i}{h} - s\right) ds \\
&= \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} K \star K\left(\frac{x_j - x_i}{h}\right).
\end{aligned}$$

Mit Hilfe der Formel für das Integral können wir schließlich $CV(h)$ schreiben als

$$CV(h) = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} K \star K\left(\frac{x_j - x_i}{h}\right) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{h,i}(x_i).$$

Beispiel 2.5 (Mischung von Normalverteilungen)
Abbildung 2.11 zeigt für den simulierten Datensatz aus den Beispielen 2.2 und 2.4 die Kreuzvalidierungsfunktion. Als Kern wurde ein Gausskern verwendet. In diesem Fall gilt

$$K \star K(u) = \frac{1}{2\sqrt{\pi}} \exp(-u^2/4).$$

Das Minimum der Kreuzvalidierungsfunktion liegt ungefähr bei $h = 0.6$. Die Dichteschätzer mit der CV-optimalen Bandweite findet man in Abbildung 2.12. Zum Vergleich ist die wahre Dichte zusätzlich eingezeichnet (gestrichelte Linie).

△

Beispiel 2.6 (Mietspiegel)
Abbildung 2.13 zeigt für die Mietspiegeldaten die Kreuzvalidierungsfunktion der Nettomiete. Wie in Beispiel 2.5 wurde ein Gausskern verwendet. Die Abbildung zeigt ein typisches Phänomen der Kreuzvalidierung, die Kreuzvaldierungsfunktion besitzt kein eindeutiges Optimum.

△

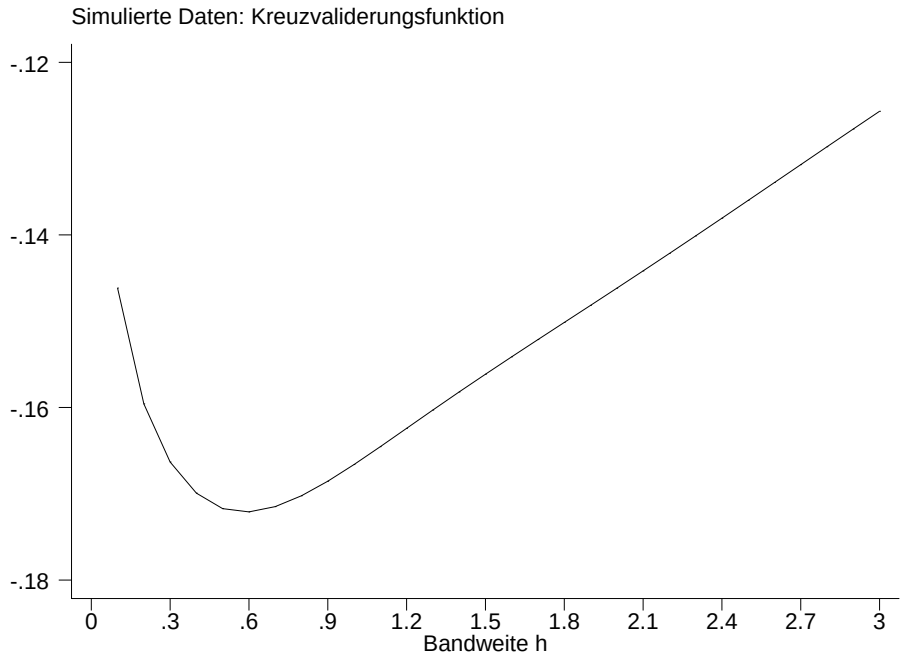

Abbildung 2.11. *Kreuzvalidierungsfunktion für die simulierten Daten aus Beispiel 2.2 (Mischung aus Normalverteilungen). Als Kern wurde ein Gausskern verwendet. Die optimale Bandweite ist* $h = 0.6$.

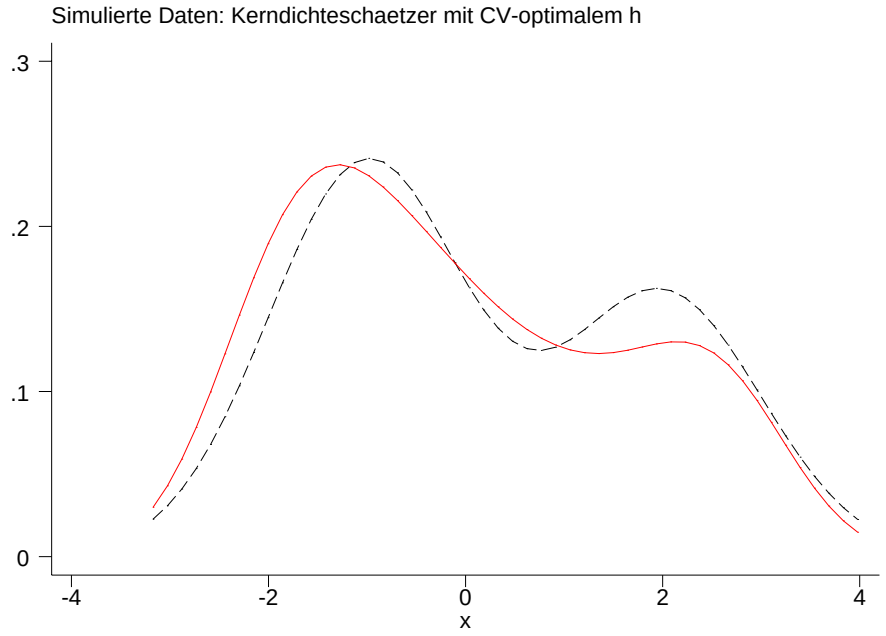

Abbildung 2.12. *Kerndichteschätzer für die simulierten Daten aus Beispiel 2.5, wobei die CV-optimale Bandweite $h = 0.6$ verwendet wurde. Als Kern wurde ein Gausskern verwendet. Zum Vergleich ist die wahre Dichte zusätzlich eingezeichnet (gestrichelte Linie).*

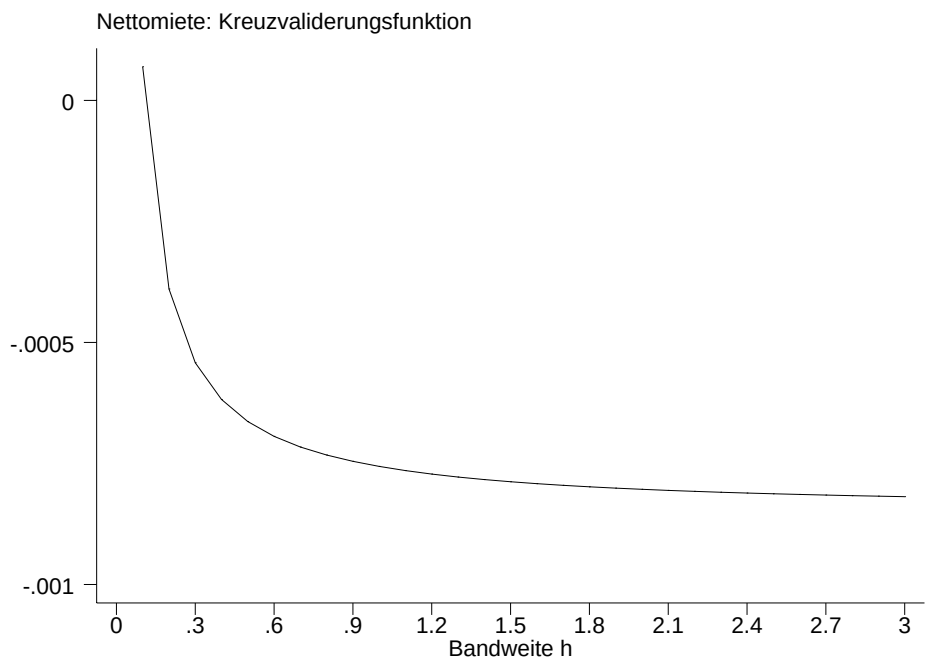

Abbildung 2.13. *Kreuzvalidierungsfunktion für Nettomiete pro Quadratmeter aus dem Mietspiegeldatensatz. Die Abbildung zeigt ein typisches Phänomen der Kreuzvalidierung, die Kreuzvaldierungsfunktion hat kein Optimum.*

2.5 Multivariate Kerndichteschätzer

Gegeben sei nun ein d - dimensionaler Zufallsvektor $X = (X_1, \ldots, X_d)$ mit Dichte $f(x_1, \ldots, x_d) = f(x)$ .Weiterhin sei eine iid. Stichprobe $\overrightarrow{x_1}$, ..., $\overrightarrow{x_n}$ gegeben, die wir in der Matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nd} \end{pmatrix}$$

zusammenfassen. Wir betrachten folgende multivariate Verallgemeinerungen von Kerndichteschätzern:

– Produktkerne:

$$\hat{f}_{\overrightarrow{h}}(\overrightarrow{x}) = \frac{1}{nh_1 \ldots h_d} \sum_{i=1}^{n} \left(\prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j} \right) \right),$$

mit $\overrightarrow{h} := (h_1, \ldots, h_d)'$.

– Multivariate Version univariater Kernfunktionen:

$$\hat{f}_h(\overrightarrow{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{(\overrightarrow{x} - \overrightarrow{x_i})' S^{-1} (\overrightarrow{x} - \overrightarrow{x_i})}{h^2} \right).$$

Beispiele für multivariate Kerne $K(u)$ sind gegeben durch:

– Rechteckskern

$$K(u) = \begin{cases} \frac{1}{h^d |S|^{\frac{1}{2}} c_0} & \text{für } u' S^{-1} u \leq h^2 \\ 0 & \text{sonst} \end{cases}$$

$$c_o = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{p}{2}+1)}$$

– Gausskern

$$K(u) = \frac{1}{(2\pi)^{\frac{d}{2}} h^d |S|^{\frac{1}{2}}} \exp\left(-\frac{u' S^{-1} u}{2h^2} \right)$$

Für die Wahl von S bestehen unter anderem folgende Möglichkeiten:

– $S = I$, d.h. gleiche Bandweiten in allen Dimensionen;

– $S = diag(s_1^2, \ldots, s_d^2)$, wobei $s_1^2, \ldots, s_d^2$ die empirischen Varianzen sind;

– $S =$ empirische Kovarianzmatrix (damit werden auch Abhängigkeiten berücksichtigt).

3

Nichtparametrische Regression I: Scatterplotsmoother

3.1 Wiederholung: lineare Modelle

In diesem Abschnitt wiederholen wir kurz das lineare Regressionsmodell. Das lineare Modell dient als Grundlage für die nichtparametrische Regression, viele Verfahren lassen sich auf lineare Regressionsmodelle zurückführen. Wir beginnen mit dem klassischen linearen Modell.

3.1.1 Das klassische lineare Regressionsmodell

Gegeben sei eine primär interessierende Variable Y und eine Menge $X = (X_0, \ldots, X_p)'$ von sogenannten *Kovariablen* (auch *unabhängige Variablen*). Y heißt *Responsevariable* (kurz: *Response*) oder auch *abhängige Variable*. Man nimmt an, dass zumindest approximativ ein funktionaler Zusammenhang zwischen Y und den Kovariablen besteht, d.h.

$$Y \approx f(X) = f(X_0, \ldots, X_p). \tag{3.1}$$

Im Rahmen der linearen Modelle wird speziell von einem additiven und linearen Zusammenhang zwischen Y und X ausgegangen, d.h.

$$Y \approx \beta_0 X_0 + \cdots + \beta_p X_p. \tag{3.2}$$

In der Regel gilt der Zusammenhang nicht exakt, sondern wird durch eine zufällige Störgröße ε kontaminiert/überlagert/gestört. Wir gehen im Folgenden davon aus, dass die Störung additiv ist, d.h. das Modell (3.2) wird zu

$$Y = \beta_0 X_0 + \cdots + \beta_p X_p + \varepsilon. \tag{3.3}$$

Aufgabe der Statistik ist es die Art und Weise des Zusammenhangs zu bestimmen. Dies ist gleichbedeutend mit der geeigneten Schätzung des Parametervektors $\beta = (\beta_0, \ldots, \beta_p)'$. Zu diesem Zweck werden Daten y_i und $x_i = (x_{i0}, \ldots, x_{ip})'$, $i = 1, \ldots, n$, erhoben, so dass man für jeden Beobachtungspunkt die Gleichung

$$y_i = \beta_0 x_{i0} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{3.4}$$

erhält. Definiert man die $n \times 1$ Vektoren

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{und} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

sowie die *Designmatrix* $\mathbf{X}$ der Dimension $n \times p + 1$

$$\mathbf{X} = \begin{pmatrix} x_{10} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n0} & \cdots & x_{np} \end{pmatrix},$$

so lassen sich die n Gleichungen aus (3.4) kompakt in Matrixnotation schreiben:

$$y = \mathbf{X}\beta + \varepsilon. \tag{3.5}$$

Im Rahmen des klassischen linearen Modells werden über den Vektor ε der Störgrößen folgende Annahmen getroffen:

– $E(\varepsilon) = 0$, d.h. die Störungen sind im Mittel Null;

– $E(\varepsilon\varepsilon') = \mathrm{Cov}(\varepsilon) = \sigma^2\mathbf{I}$, d.h. die Varianz der Störgrößen bleibt konstant (Homoskedastie) und die Störungen sind von Beobachtung zu Beobachtung unkorreliert;

– Zur Durchführung von Hypothesentests über die unbekannten Parameter und zur Berechnung von Konfidenzintervallen nehmen wir noch zusätzlich an, dass die Störgrößen normalverteilt sind, d.h. $\varepsilon \sim N(0, \sigma^2\mathbf{I})$.

Für die Designmatrix $\mathbf{X}$ nehmen wir zusätzlich an, dass

– $\mathbf{X}$ nichtstochastisch ist und

– $\mathrm{rg}(\mathbf{X}) = p + 1$, d.h. $\mathbf{X}$ hat vollen Spaltenrang bzw. ist spaltenregulär.

Insgesamt erhalten wir das *klassische lineare Regressionsmodell*:

1. $y = \mathbf{X}\beta + \varepsilon$
2. $E(\varepsilon) = 0$
3. $E(\varepsilon\varepsilon') = \sigma^2\mathbf{I}$
4. $\varepsilon \sim N(0, \sigma^2\mathbf{I})$

5. $\mathbf{X}$ ist nichtstochastisch und besitzt vollen Spaltenrang.

Als einfache Folgerungen erhält man

$$E(y) = E(\mathbf{X}\beta + \varepsilon) = \mathbf{X}\beta + E(\varepsilon) = \mathbf{X}\beta \tag{3.6}$$

und

$$\mathrm{Cov}(y) = \mathrm{Cov}(\mathbf{X}\beta + \varepsilon) = \mathrm{Cov}(\varepsilon) = \sigma^2 I. \tag{3.7}$$

Unter Berücksichtigung der Normalverteilungsannahme erhalten wir

$$y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

Beispiel 3.1 (univariates Regressionsmodell)
Einen wichtigen Spezialfall des linearen Modells stellt das univariate Regressionsmodell dar, das eine Konstante (sogenannter Intercept) und nur eine unabhängige Variable X enthält:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1, \ldots, n) \tag{3.8}$$

Die Designmatrix hat in diesem Fall die Gestalt

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Beispiel 3.2 (multiples Regressionsmodell mit Intercept)
Das multiple Regressionsmodell mit konstantem Glied (Intercept) ist gegeben durch

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \qquad (i = 1, \ldots, n). \tag{3.9}$$

Für die Designmatrix $\mathbf{X}$ gilt in diesem Fall

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Beispiel 3.3 (nichtlineare Beziehungen)
Im Rahmen der linearen Modelle können durchaus auch nichtlineare Beziehungen zwischen der abhängigen Variable und den Kovariablen behandelt werden. Betrachte zum Beispiel das folgende Modell

$$y_i = f(z_i) + \varepsilon_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 + \varepsilon_i,$$

in dem die Funktion f ein Polynom dritten Grades ist. Wir können dieses Modell auf ein einfaches lineares Modell zurückführen, indem wir die Variablen $x_{1i} := z_i$, $x_{2i} := z_i^2$ und $x_{3i} := z_i^3$ definieren. Damit erhalten wir wieder ein lineares Modell. In Abhängigkeit der Beobachtungen z_i ergibt sich die Designmatrix zu

$$\mathbf{X} = \begin{pmatrix} 1 & z_1 & z_1^2 & z_1^3 \\ \vdots & \vdots & & \vdots \\ 1 & z_n & z_n^2 & z_n^3 \end{pmatrix}.$$

Im Allgemeinen lassen sich alle nichtlinearen Beziehungen auf ein einfaches lineares Modell zurückführen, solange sie linear in den Parametern sind. Ein Beispiel für ein echtes nichtlineares Modell ist gegeben durch:

$$y_i = f(z_i) + \varepsilon_i = \beta_0 + \beta_1 \sin(\beta_2 z_i) + \varepsilon_i$$

3.1.2 Schätzungen

Wir befassen uns zunächst mit der Schätzung von β. Nach der Methode der kleinsten Quadrate (KQ-Methode) wird die Schätzung $\hat{\beta}$ von β so bestimmt, dass

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (y_i - \beta_0 x_{i0} - \cdots - \beta_p x_{ip})^2 \\ &= \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \\ &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \epsilon' \epsilon \\ &= (y - \mathbf{X}\beta)' (y - \mathbf{X}\beta) \end{aligned}$$

bezüglich β minimal wird. Differenzieren und Nullsetzen liefert den KQ - Schätzer

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' y.$$

Der KQ - Schätzer besitzt folgende Eigenschaften:

Satz 3.1 (Eigenschaften des KQ-Schätzers)

1. *$E(\hat{\beta}) = \beta$ d.h. $\hat{\beta}$ ist erwartungstreu.*
2. *$Cov(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$*
3. *$\hat{\beta}$ ist bester linearer unverzerrter Schätzer (BLUE), d.h. unter allen linearen unverzerrten Schätzern der Form $\tilde{\beta} = \mathbf{A}Y$ besitzt $\hat{\beta}$ minimale Varianz:*

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j) \qquad j = 0, \ldots, p.$$

(Gauss-Markov Theorem)

Diese Eigenschaften sind nicht an eine spezielle Verteilungsannahme gebunden, (d.h. die Normalverteilungsannahme der Störgrößen ist nicht erforderlich).

Unter der Normalverteilungsannahme für die Störgrößen erhält man als Likelihood

$$L(\beta) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)'(y - \mathbf{X}\beta)\right)$$

Maximierung der Likelihood bzgl. β liefert wieder den KQ-Schätzer, d.h.

$$\hat{\beta}_{ML} = \hat{\beta}.$$

Basierend auf der KQ-Schätzung $\hat{\beta}$ erhalten wir auch eine Schätzung von $E(y)$:

$$\hat{y} = \widehat{E(y)} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

Die Matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ heißt Hatmatrix.

In der Regel ist neben den unbekannten Regressionsparametern β auch die Varianz σ^2 der Störgrößen unbekannt. Eine erwartungstreue Schätzung ist gegeben durch

$$\hat{\sigma^2} = \frac{1}{n-p-1}\sum_{i=1}^{n}(\underbrace{y_i - x_i'\hat{\beta}}_{\hat{\varepsilon}_i})^2 = \frac{1}{n-p-1}\hat{\varepsilon}'\,\hat{\varepsilon}$$

Dies is die übliche Schätzung für σ^2. Die ML-Schätzung ist gegeben durch

$$\hat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\underbrace{y_i - x_i'\hat{\beta}}_{\hat{\epsilon}_i})^2 = \frac{1}{n}\hat{\varepsilon}'\,\hat{\varepsilon}.$$

Die ML-Schätzung für σ^2 ist also nicht erwartungstreu.

3.1.3 Das Bestimmtheitsmaß

Im Modell mit Intercept gilt folgende Streuungszerlegungsformel:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Total sum of squars (SST)}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Regression sum of squars (SSR)}} + \underbrace{\sum_{i=1}^{n}\hat{\epsilon}_i^2}_{\text{Error sum of squars (SSE)}}$$

Das Bestimmtheitsmaß basiert auf der Streungszerlegungsformel und ist definiert als

$$B = R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Es gilt $0 \leq R^2 \leq 1$. Das Bestimmtheitsmaß kann als Gütemaß zur Beurteilung der Anpassung des Regressionsmodells an die Daten herangezogen werden. Das Bestimmtheitsmaß ist nahe eins, wenn die Residuenquadratsumme möglichst klein ist. D. h. je näher R^2 bei 1 liegt, desto besser ist die Anpassung, je näher R^2 bei 0 desto schlechter die Anpassung.

Beispiel 3.4 (Mietspiegel für München)
Im Folgenden sollen nur die Wohnfläche (wfl) und das Baujahr (bj) als Kovariablen berücksichtigt werden. Wir passen zunächst ein Modell mit der Nettomiete (nm) als Responsevariable an und anschließend ein Modell mit der Nettomiete pro qm ($nmqm$).

Nettomiete als abhängige Variable

Abbildung 3.1 zeigt Scatterplots zur Verdeutlichung des Zusammenhangs zwischen Nettomiete und Wohnfläche bzw. Baujahr, die zumindest annähernd lineare Beziehungen vermuten lassen. Wir schätzen also das Modell

$$y_i = \beta_0 + \beta_1 wfl_i + \beta_2 bj_i + \epsilon_i.$$

mit Hilfe der KQ-Methode. Die Schätzungen findet man in Abbildung 3.2. Abbildung 3.3 zeigt sogenannte „added variable plots". Diese können bei einer Regression mit mehr als einer Kovariable dazu benutzt werden, den geschätzten Zusammenhang zwischen der Responsevariable und einer der Kovariablen grafisch darzustellen, wobei der Einfluss der übrigen Kovariablen herausgerechnet wird (vgl. dazu Weisberg (1985)). Added variable plots werden wie folgt erzeugt:

1. Betrachte allgemein das Regressionsmodell

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon. \tag{3.10}$$

   Ziel ist es den Zusammenhang zwischen y und x_k unter Eliminierung des Einflusses der übrigen Kovariablen grafisch darzustellen.

2. Schätze dazu ein Regressionsmodell zwischen y und den Kovariablen, wobei x_k nicht berücksichtigt wird. Bestimme die Residuen $\hat{\epsilon}_{-x_k}$ des geschätzten Modells. Die Residuen entsprechen dem Anteil von y der nicht von den Kovariablen erklärt wird außer x_k.

3. Schätze ein Regressionsmodell zwischen x_k und den restlichen Kovariablen. Bestimme die Residuen $\hat{\epsilon}_k$. Die Residuen entsprechen dem Anteil von x_k, der nicht von den restlichen Kovariablen erklärt wird.

4. Zeichne einen Scatterplot zwischen $\hat{\epsilon}_k$ und $\hat{\epsilon}_{-x_k}$. Würde man den Zusammenhang zwischen $\hat{\epsilon}_k$ und $\hat{\epsilon}_{-x_k}$ mit Hilfe einer univariaten Regression schätzen, so erhielte man eine Konstante von 0 und als Steigung der Regressionsgeraden den geschätzten Regressionskoeffizienten aus dem Modell (3.10).

Wir erkennen in Abbildung 3.3, dass die Anpassung an die Daten zufriedenstellend ist. Allerdings scheint die Annahme der Homoskedastizität verletzt, da die Variabilität der Residuen mit steigender Wohnfläche zunimmt.

Nettomiete pro qm als abhängige Variable

Abbildung 3.4 zeigt Scatterplots zur Verdeutlichung des Zusammenhangs zwischen Nettomiete pro Quadratmeter und Wohnfläche bzw. Baujahr. Zusätzlich ist bereits eine Regressionsgerade enthalten, die aus univariaten Regressionen zwischen abhängiger Variable $nmqm$ und jeweils einer der Kovariablen resultiert. Während beim Baualter ein linearer Zusammenhang zumindest einigermaßen gerechtfertigt ist, scheint bei der Wohnfläche ein nichtlinearer Zusammenhang zu bestehen. Wir transformieren die Variable wfl und erhalten als neue Kovariable $wfltr = 1/wfl$. Abbildung 3.5 zeigt einen Scatterplot zwischen Nettomiete pro qm und Wohnfläche, in den die geschätzte Funktion $\hat{\beta}_0 + \hat{\beta}_1 1/wfl$ zusätzlich eingezeichnet ist. Hier scheint die Anpassung an die Daten zufriedenstellend zu sein. Wir schätzen daher folgendes Regressionsmodell:

$$nmqm_i = \beta_0 + \beta_1 \frac{1}{wfl_i} + \beta_2 bj_i + \epsilon_i$$

Abbildung 3.6 zeigt die Schätzergebnisse. Das Bestimmtheitsmaß beträgt $R^2 = 0.259$ und ist deutlich höher als bei einer Regression, in der die Wohnfläche linear eingeht. Hier ist das Bestimmtheitsmaß $R^2 = 0.217$. Abbildung 3.7 zeigt wieder „added variable plots". Auch hier ist die Homoskedastizitätsannahme zumindest fraglich.

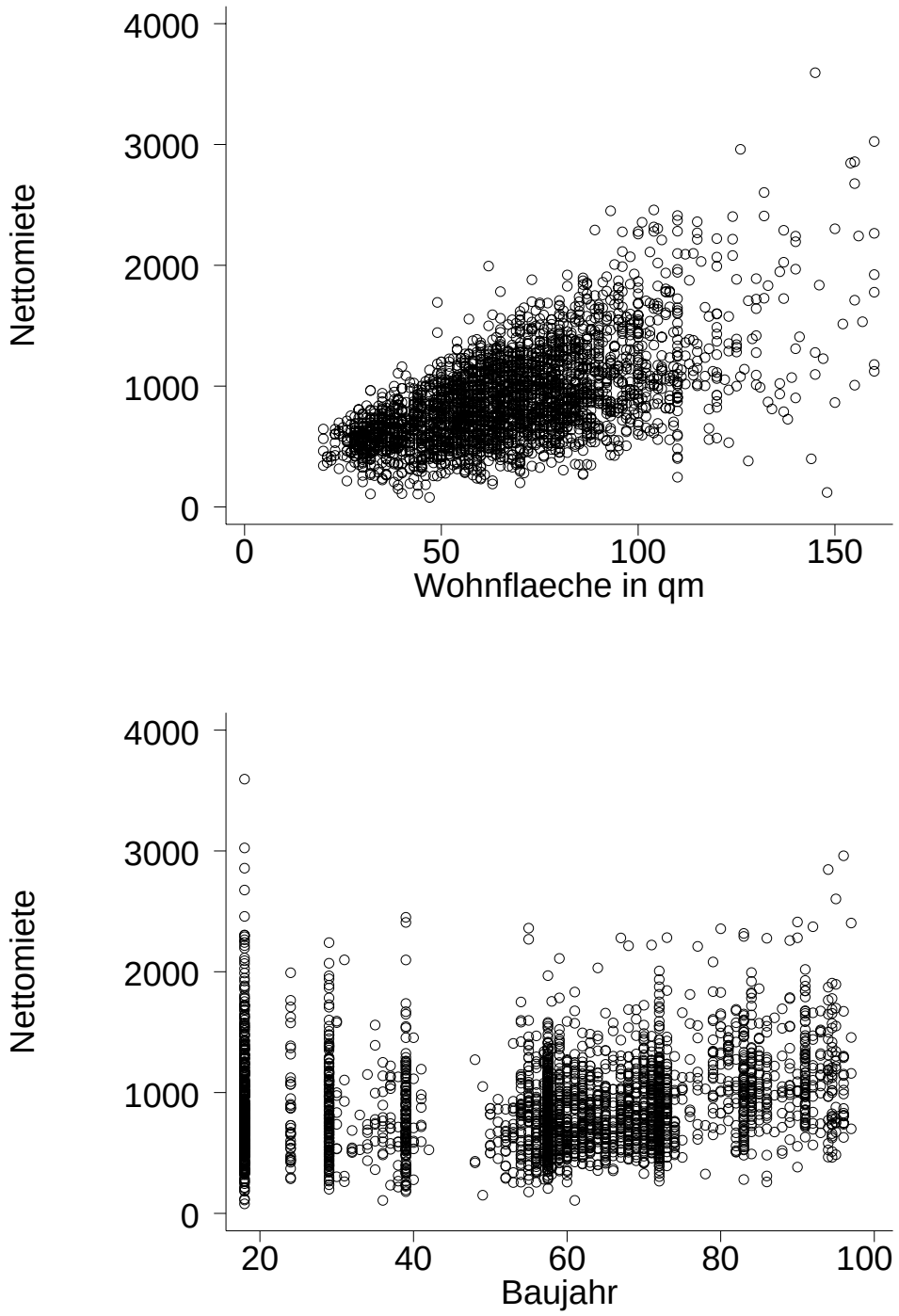

Abbildung 3.1. *Scatterplots: Nettomiete gegen Wohnfläche und Baujahr.*

```
1 . regress nm wfl bam

      Source |       SS       df       MS              Number of obs =    3082
-------------+------------------------------           F(  2,  3079) = 1106.03
       Model |  187501367     2  93750683.6           Prob > F      =  0.0000
    Residual |  260985873  3079  84763.1936           R-squared     =  0.4181
-------------+------------------------------           Adj R-squared =  0.4177
       Total |  448487240  3081  145565.479           Root MSE      =  291.14

------------------------------------------------------------------------------
          nm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wfl |   10.45556   .2272494    46.01   0.000     10.00998    10.90113
         bam |   4.858069   .2416284    20.11   0.000       4.3843    5.331838
       _cons |  -82.07799   23.30466    -3.52   0.000    -127.7722   -36.38374
------------------------------------------------------------------------------
```

Abbildung 3.2. *Schätzergebnisse für das Modell* $y_i = \beta_0 + \beta_1 wfl_i + \beta_2 bj_i + \epsilon_i$.

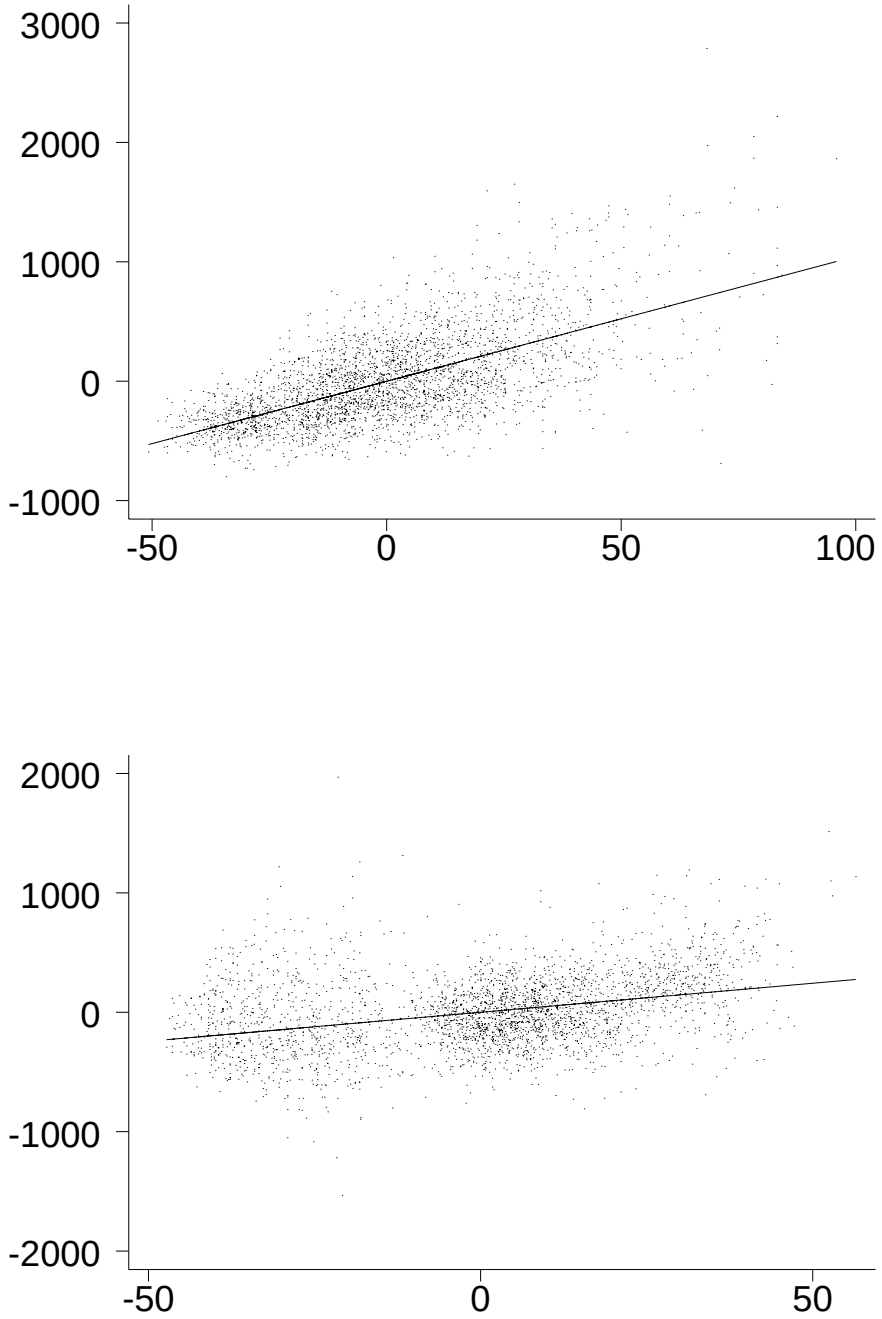

Abbildung 3.3. *Nettomiete: Added variable plots für die Wohnfläche (obere Grafik) und das Baujahr (untere Grafik).*

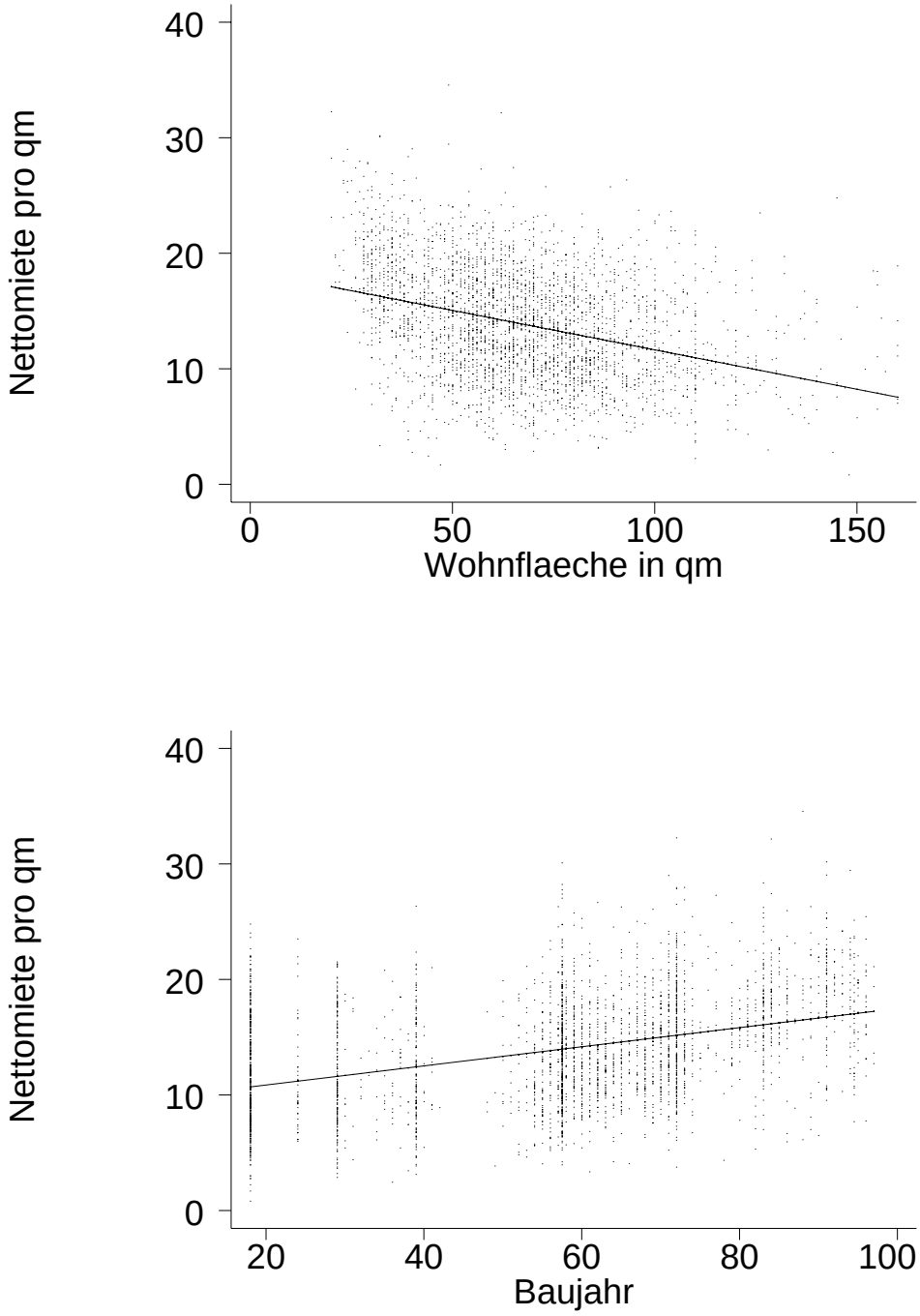

Abbildung 3.4. *Scatterplots: Nettomiete pro qm gegen Wohnfläche und Baujahr.*

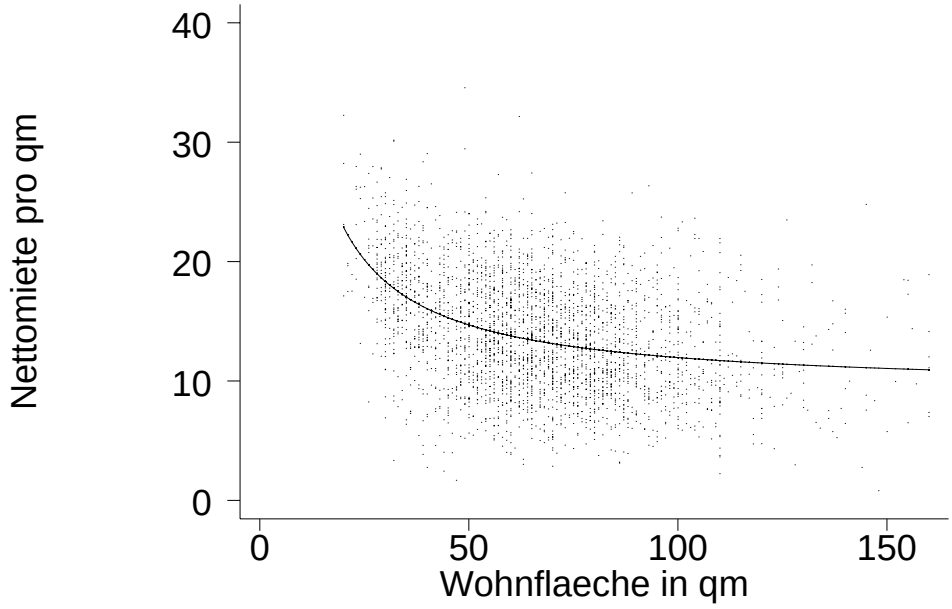

Abbildung 3.5. *Scatterplots: Nettomiete pro qm gegen Wohnfläche. Zusätzlich ist die geschätzte Funktion* $\hat{\beta}_0 + \hat{\beta}_1 1/wfl$ *abgebildet.*

```
regress nmproqm wfltr bam

      Source |       SS       df       MS              Number of obs =    3082
-------------+------------------------------           F(  2,  3079) =  538.45
       Model |  18014.8078     2  9007.40391           Prob > F      =  0.0000
    Residual |  51506.5845  3079  16.7283483           R-squared     =  0.2591
-------------+------------------------------           Adj R-squared =  0.2586
       Total |  69521.3924  3081  22.5645545           Root MSE      =    4.09

------------------------------------------------------------------------------
     nmproqm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       wfltr |   232.7534   10.99055    21.18   0.000     211.2039     254.303
         bam |   .0702642   .0033556    20.94   0.000     .0636848    .0768436
       _cons |   5.960297   .2519675    23.66   0.000     5.466256    6.454339
------------------------------------------------------------------------------
```

Abbildung 3.6. *Schätzergebnisse für das Modell* $yqm_i = \beta_0 + \beta_1 1/wfl_i + \beta_2 bj_i + \epsilon_i$.

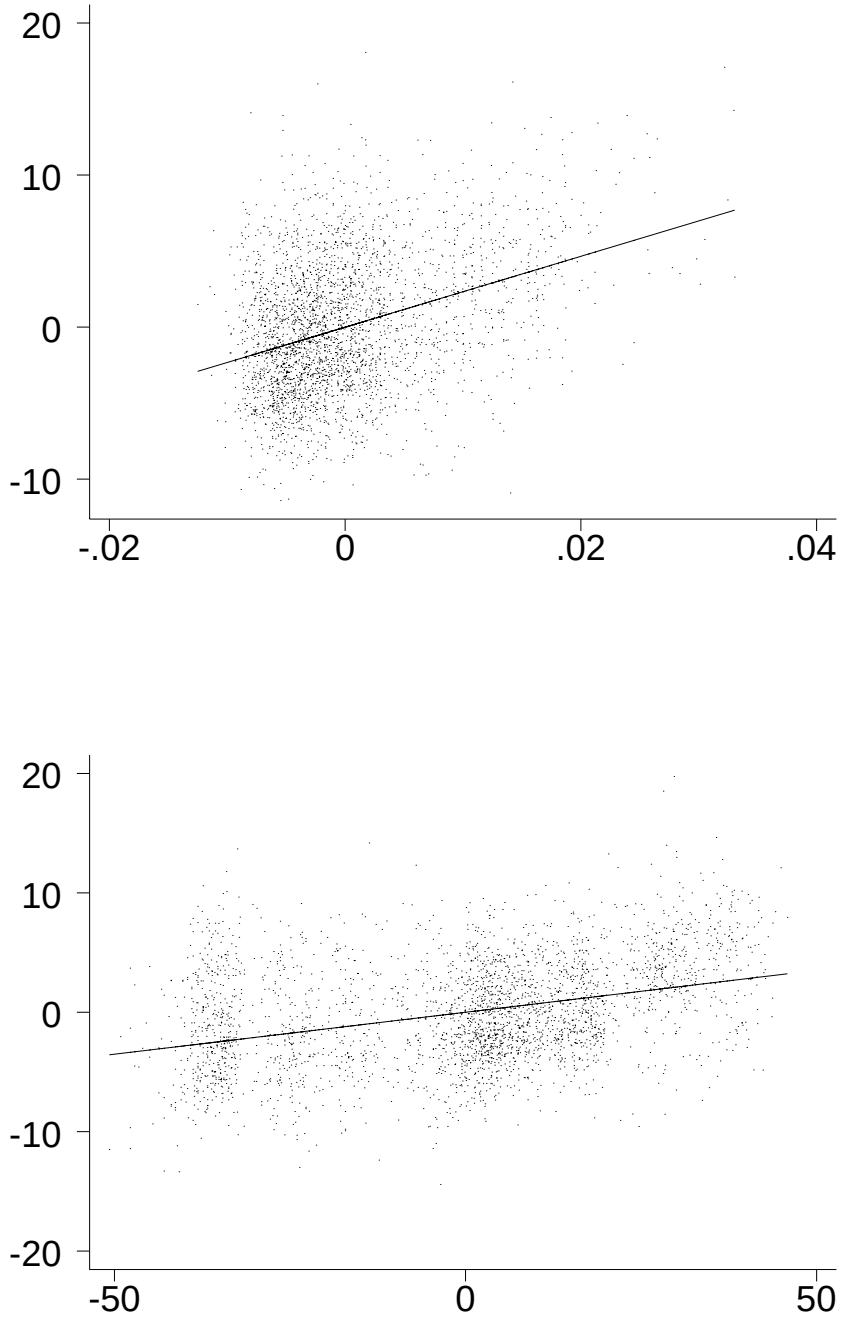

Abbildung 3.7. *Nettomiete pro Quadratmeter: Added variable plots für die Wohnfläche (obere Grafik) und das Baujahr (untere Grafik).*

3.1.4 Gewichtete Regression

In diesem Abschnitt ersetzen wir die Annahme $\text{cov}(\varepsilon) = \sigma^2\mathbf{I}$ durch

$$\text{cov}(\varepsilon) = \sigma^2\,\mathbf{W}^{-1} = \sigma^2 diag(\frac{1}{w_1}, \ldots, \frac{1}{w_n}),$$

wobei die Gewichtsmatrix $\mathbf{W}$ bekannt sei. Unter Berücksichtigung der Normalverteilungsannahme erhalten wir

$$\varepsilon \sim N(0, \sigma^2\mathbf{W}^{-1}) \quad \text{bzw.} \quad y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{W}^{-1}).$$

Wir sprechen von heteroskedastischen Fehlern bzw. von Heteroskedastie. Die Bestimmung einer Schätzung für β beruht wieder auf dem KQ-Prinzip, wobei im vorliegenden Fall die gewichtete Residuenquadratsumme

$$S(\beta) = \sum_{i=1}^{n} w_i(y_i - x_i'\beta)^2 = (y - \mathbf{X}\beta)'\mathbf{W}(y - \mathbf{X}\beta)$$

bezüglich β minimiert wird. Wir erhalten

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}y.$$

Der gewichtete KQ-Schätzer besitzt folgende Eigenschaften:

Satz 3.2 (Eigenschaften des gewichteten KQ-Schätzers)

1. *$E(\hat{\beta}) = \beta$, d.h. der Schätzer ist erwartungstreu.*
2. *$cov(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$*
3. *Analog zum einfachen KQ-Schätzer gilt das Gauß-Markov Theorem.*

Schließlich erhalten wir analog zum ungewichteten Fall eine erwartungstreue Schätzung von σ^2 durch

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\epsilon}'\mathbf{W}\hat{\epsilon}$$

wobei

$$\hat{\epsilon} = y - \hat{y} = y - \mathbf{X}\hat{\beta}.$$

Bei Vorliegen von heteroskedastischen Fehlern kann man in der Praxis wie folgt vorgehen:

1. Schätze mit Hilfe der ungewichteten KQ-Methode β durch $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$.
2. Bestimme die quadrierten Residuen $\hat{\epsilon}_i^2$ aus der ungewichteten Regression in 1. Schätze ein Regressionsmodell mit $\hat{\epsilon}_i^2$ als abhängiger Variable und erhalte $\widehat{\hat{\epsilon}_i^2}$.

3. Schätze das Regressionsmodell mit Hilfe der gewichteten KQ-Methode erneut und verwende als Gewichte
$$w_i \propto \frac{1}{\widehat{\hat{\epsilon}_i^2}} \qquad i = 1, \ldots, n.$$
Häufig wird so normiert, dass $\sum w_i = n$ oder $\sum w_i = 1$.

Beispiel 3.5 (Mietspiegeldaten)
Wir beschränken uns auf das Modell mit der Nettomiete (nm) als abhängiger Variable. Um die Gewichte w_i für die gewichtete KQ-Schätzung zu bestimmen, berechnen wir zunächst die Residuen $\hat{\epsilon}_i$, die aus der ungewichteten KQ-Schätzung resultieren. Anschließend schätzen wir ein lineares Modell mit den logarithmierten quadrierten Residuen $log(\hat{\epsilon}_i^2)$ als abhängiger Variable und der Wohnfläche und dem Baujahr als Kovariablen (Abbildung 3.8). Die Gewichte w_i ergeben sich dann zu
$$w_i = \frac{1}{\widehat{\hat{\epsilon}_i^2}}.$$
Abbildung 3.9 zeigt die Ergebnisse der gewichteten Regression.

```
regress res2l wfl bam

      Source |       SS       df       MS              Number of obs =    3082
-------------+------------------------------           F(  2,  3079) =  181.56
       Model |   1653.0446     2  826.522301           Prob > F      =  0.0000
    Residual |  14016.8375  3079  4.55239933           R-squared     =  0.1055
-------------+------------------------------           Adj R-squared =  0.1049
       Total |  15669.8821  3081  5.08597278           Root MSE      =  2.1336

------------------------------------------------------------------------------
       res2l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wfl |   .0284544   .0016654    17.09   0.000      .025189    .0317198
         bam |  -.0075614   .0017708    -4.27   0.000    -.0110334   -.0040894
       _cons |   8.418719   .1707888    49.29   0.000     8.083848    8.753591
------------------------------------------------------------------------------
```

Abbildung 3.8. *Schätzergebnisse für das Modell* $log(\hat{\epsilon}_i^2) = \beta_0 + \beta_1 wfl_i + \beta_2 bj_i + \epsilon_i$.

```
regress nm wfl bam [aweight=w]

      Source |       SS       df       MS              Number of obs =    3082
-------------+------------------------------           F(  2,  3079) = 1259.58
       Model |   132952804     2  66476402.2           Prob > F      =  0.0000
    Residual |   162499041  3079  52776.5641           R-squared     =  0.4500
-------------+------------------------------           Adj R-squared =  0.4496
       Total |   295451845  3081  95894.7891           Root MSE      =  229.73

------------------------------------------------------------------------------
          nm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wfl |   10.16611   .2160647    47.05   0.000     9.742469    10.58976
         bam |    4.87865    .210988    23.12   0.000     4.464959    5.292342
       _cons |  -65.56685   18.96973    -3.46   0.001    -102.7615   -28.37224
------------------------------------------------------------------------------
```

Abbildung 3.9. *Schätzergebnisse für die gewichtete Regression* $nm_i = \beta_0 + \beta_1 wfl_i + \beta_2 bj_i + \epsilon_i$.

3.2 Scatterplotsmoother

3.2.1 Definition

Wir wollen uns im Folgenden zunächst auf eine Kovariable konzentrieren. Gegeben seien also unabhängige Beobachtungen $(y_i, x_i),\ i = 1, \ldots, n$ wobei x_i Skalare seien. Im Rahmen der linearen Modelle ist es zwar möglich nichtlineare Beziehungen zwischen Y und X zu modellieren (vgl. die Beispiele 3.3 und 3.4), es können jedoch folgende Probleme auftreten:

– Häufig ist nicht klar, auf welche Weise die erklärende Variable X transformiert werden muss.

– Bei manchen Datensätzen gibt es keine "einfache" Transformation, vergleiche etwa die Motorcycledaten im folgenden Beispiel 3.6.

Beispiel 3.6 (Motorcycledaten)
Die Motorcyledaten sind ein gutes Beispiel für ein Regressionsproblem, bei dem eine einfache Transformation der unabhängigen Variable nicht existiert (vgl. Abbildung 3.10).

△

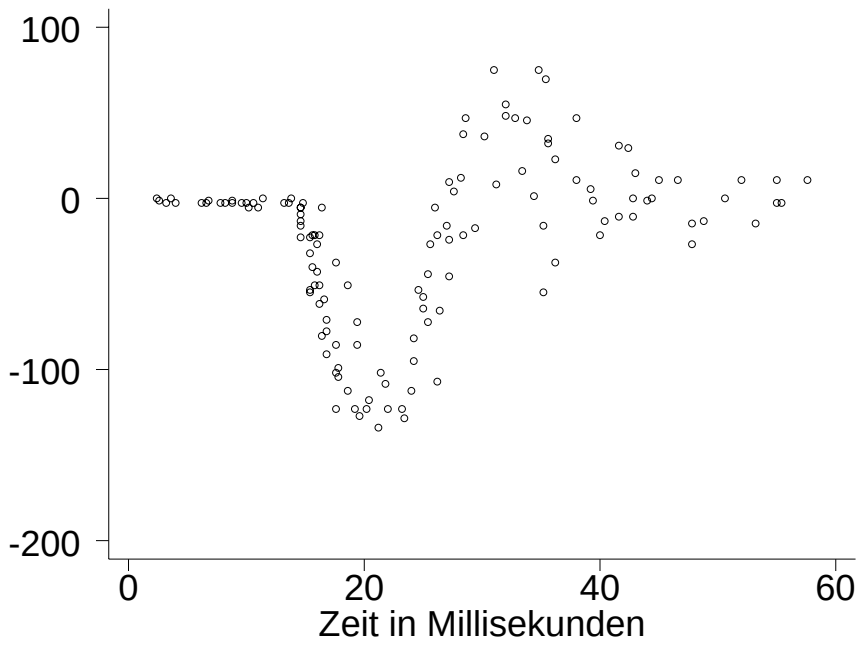

Abbildung 3.10. *Motorcycledaten: Scatterplot zwischen Beschleunigung und der Zeit.*

Um oben genannte Probleme zu lösen gehen wir wie folgt vor: Wir ersetzen die lineare Beziehung

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

durch

$$Y = f(x) + \varepsilon \tag{3.11}$$

wobei f eine zunächst unspezifizierte Funktion ist, die anhand der Daten geschätzt werden soll. Man fordert lediglich, dass f bestimmten Glattheitsforderungen genügt (etwa f stetig, stetig differenzierbar etc.). Für die Störgröße ε sollen dieselben Annahmen wie im klassischen linearen Regressionsmodell gelten.

Definition 3.1 (Scatterplotsmoother)
Ein Scatterplotsmoother ist eine Funktion

$$\hat{f}(z) = S(y, x)$$

wobei $\hat{f}(z)$ den geschätzten Funktionswert von f an der Stelle z angibt und S eine Funktion der Datenvektoren $y = (y_1, \ldots, y_n)'$ und $x = (x_1, \ldots, x_n)'$ ist.

Wir behandeln im Folgenden vor allem *lineare Schätzer* der Form

$$\hat{f} = \mathbf{S}y. \tag{3.12}$$

Hier ist $\hat{f} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))'$ und

$$\mathbf{S} = (s_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

ist eine $n \times n$ Matrix, die *Smoothermatrix* genannt wird. Die Schätzung $\hat{f}(x_i)$ erhält man also als eine gewichtete Summe der y_i, d.h.

$$\hat{f}(x_i) = s_{i1}y_1 + \cdots + s_{in}y_n.$$

Die Gewichte s_{ij} hängen dabei in der Regel von den beobachteten Kovariablenwerten x_i ab.

3.3 Basisfunktionenansätze

Basisfunktionsansätzen liegt folgende einfache Idee zugrunde: Approximiere die unbekannte Funktion f durch einen möglichst flexiblen Funktionenraum (z.B. Polynome p-ten Grades), so dass die Funktion f als Linearkombination einer endlichen Menge von Basisfunktionen darstellbar ist, d.h.

$$f(x) = \beta_0 B_0(x) + \beta_1 B_1(x) + \cdots + \beta_p B_p(x),$$

wobei $B_0, \ldots, B_p$ die Basisfunktionen sind. Wenn wir die neuen Variablen

$$\begin{aligned} z_{i0} &:= B_0(x_i) \\ z_{i1} &:= B_1(x_i) \\ &\vdots \\ z_{ip} &:= B_p(x_i) \end{aligned}$$

definieren, dann können wir das Modell (3.11) als einfaches lineares Modell

$$y_i = f(x_i) + \varepsilon_i = \beta_0 z_{i0} + \cdots + \beta_p z_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

schreiben und die unbekannten Regressionskoeffizienten $\beta = (\beta_0, \ldots, \beta_p)'$ wieder mit Hilfe der KQ-Methode schätzen. In Matrixschreibweise erhalten wir

$$y = \mathbf{X}\beta + \varepsilon$$

mit der Designmatrix

$$\mathbf{X} = \begin{pmatrix} z_{10} & \ldots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n0} & \ldots & z_{np} \end{pmatrix} = \begin{pmatrix} B_0(x_1) & \ldots & B_p(x_1) \\ \vdots & \ddots & \vdots \\ B_0(x_n) & \ldots & B_p(x_n) \end{pmatrix}.$$

Als Schätzung für β erhalten wir wieder

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

und damit als Schätzung für $f = (f(x_1), \ldots, f(x_n))'$

$$\hat{f} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y.$$

Offensichtlich handelt es sich um einen linearen Schätzer gemäß Definition 3.12 mit Smoothermatrix

$$\mathbf{S} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Beispiel 3.7 (Polynome vom Grad p)
Ein einfacher Basisfunktionenansatz basiert auf Polynomen vom Grad p, d.h. wir modellieren die unbekannte Funktion durch

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p.$$

Als Basisfunktionen verwenden wir also

$$B_0(x) = 1,\ B_1(x) = x,\ B_2(x) = x^2,\ \ldots\ , B_p(x) = x^p.$$

Die Designmatrix ergibt sich zu

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \ldots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^p \end{pmatrix}.$$

Üblicherweise verwendet man Polynome vom Grad $p \leq 3$, da Polynome höheren Grades häufig zu sehr instabilen Schätzungen führen, andererseits die Anpassung an die Daten nicht wesentlich verbessern. Abbildung 3.11 zeigt für die Mietspiegeldaten Schätzungen für f im einfachen Modell

$$nmqm_i = f(wfl_i) + \varepsilon_i.$$

Dabei wurden Polynome vom Grad 2, 3 und 5 verwendet. In diesem Beispiel scheint die Modellierung mit Polynomen zu brauchbaren Ergebnissen zu führen. Abbildung 3.12 zeigt Scatterplotsmoother für die Motorcycledaten, wobei auch hier Polynome vom Grad 2, 3 und 5 zur Modellierung verwendet wurden. Offensichtlich sind Polynome in diesem Beispiel weniger geeignet eine brauchbare Anpassung an die Daten zu gewährleisten. Auch eine weitere Erhöhung des Polynomgrads bringt keine entscheidende Verbesserung der Schätzungen wie Abbildung 3.13 zeigt. Hier wurden Polynome vom Grad 7, 8 und 9 verwendet. Die verbesserte Anpassung an die Daten beim Minimum wird erkauft durch deutlich instabilere Schätzungen an den linken und rechten Rändern.

△

Wie im Beispiel 3.7 gezeigt wurde, entstehen bei der Verwendung von Polynomen zur Modellierung der Funktion f die folgenden Probleme/Fragen:

– Welcher Polynomgrad soll verwendet werden?

– Häufig ergeben sich sehr rauhe und instabile Schätzer.

Zur Vermeidung der genannten Probleme verwenden wir im Folgenden sogenannte *polynomiale Splines* anstelle von einfachen Polynomen. Splines werden in etwa wie folgt konstruiert:

– Wir unterteilen die x-Achse (genauer den Bereich zwischen x_{min} und x_{max}) in *kleinere Teilintervalle* und schätzen in *jedem* Teilintervall ein *separates Polynom*.

– Damit die Gesamtfunktion wieder glatt wird, fordern wir zusätzlich, dass an den Intervallrändern (sogenannte Knoten) die einzelnen Polynome möglichst glatt (z.B. stetig differenzierbar) anschließen.

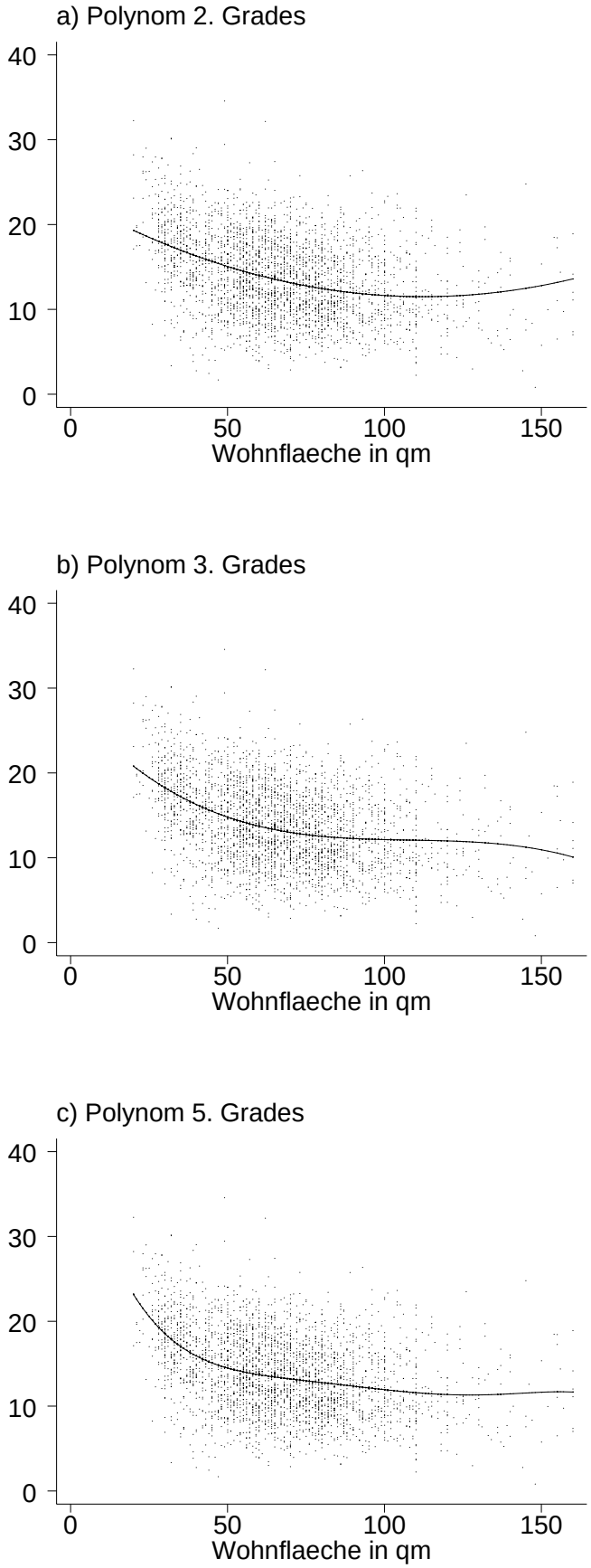

Abbildung 3.11. *Mietspiegeldaten: Polynomiale Regressionen zwischen Nettomiete pro qm und Wohnfläche.*

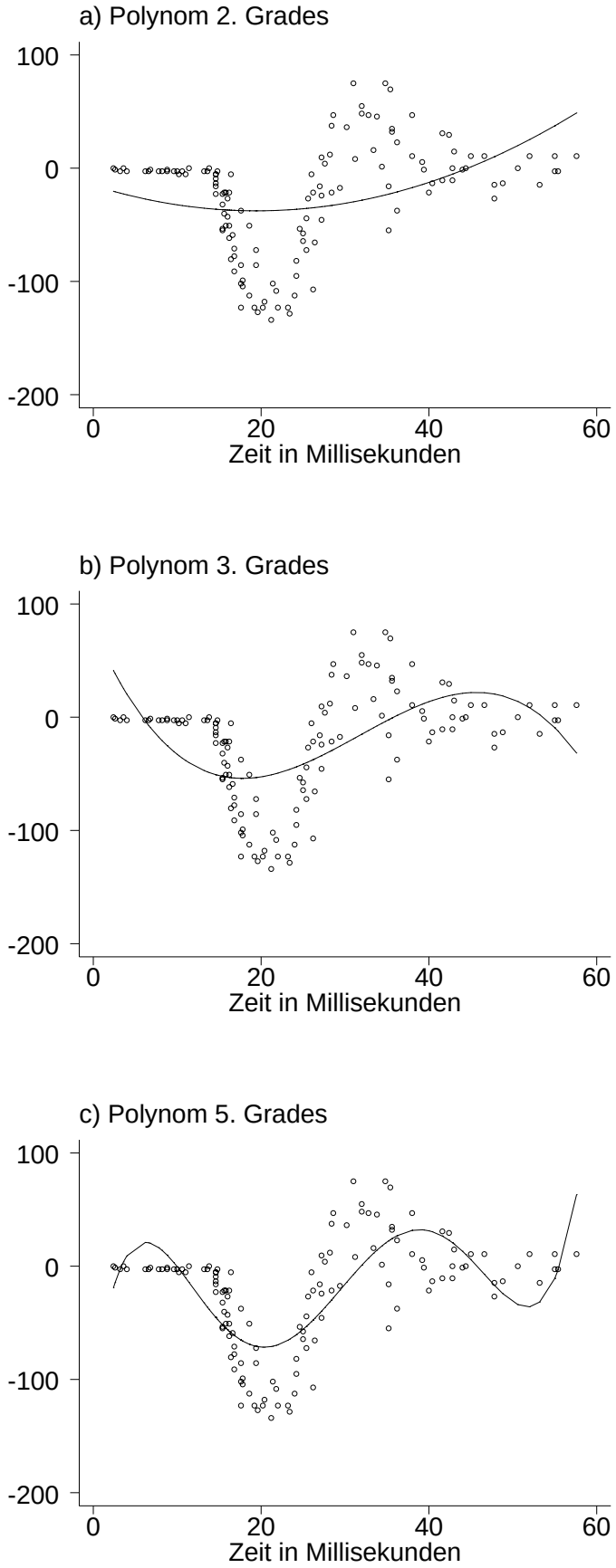

Abbildung 3.12. *Motorcycledaten: Polynomiale Regressionen zwischen Beschleunigung und der Zeit.*

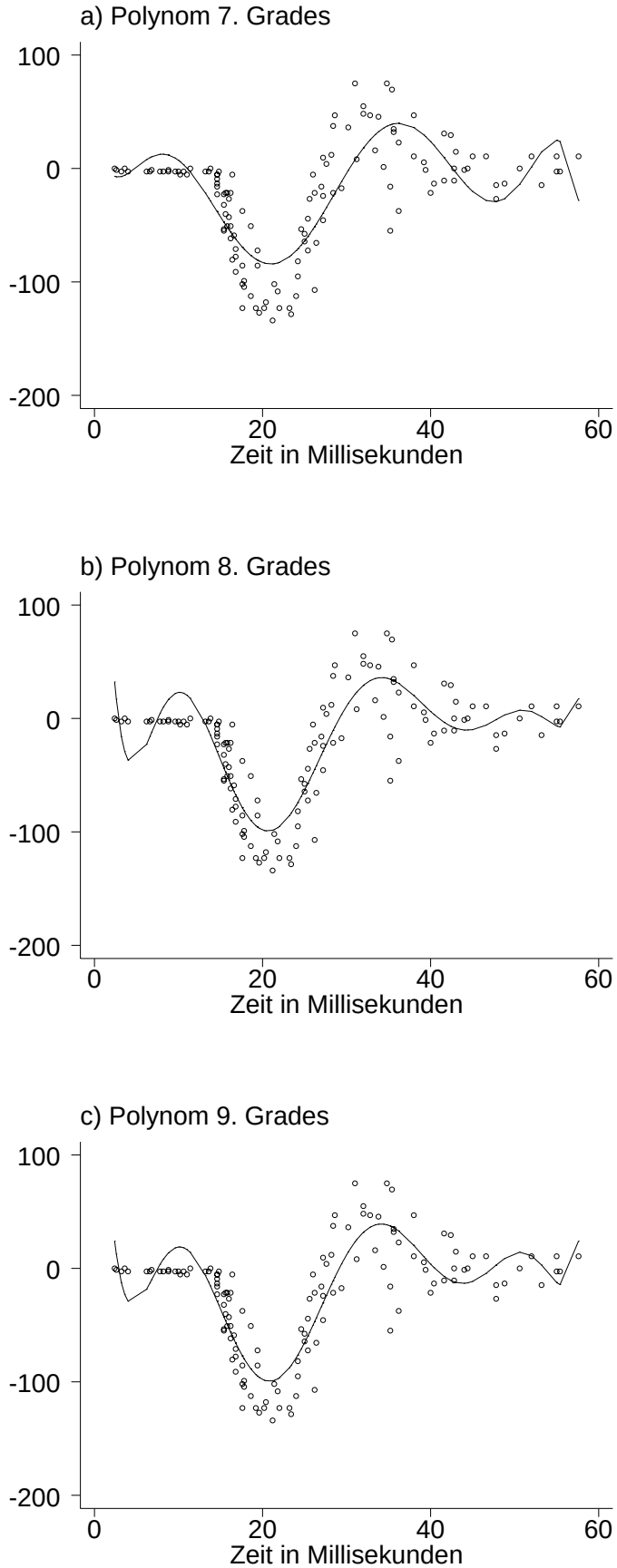

Abbildung 3.13. *Motorcycledaten: Polynomiale Regressionen zwischen Beschleunigung und der Zeit.*

Dies führt zu folgender

Definition 3.2 (Polynomsplines)
Sei $a = \xi_1 < \xi_2 < \ldots < \xi_m = b$ eine Unterteilung des Intervalls $[a, b]$. Eine Funktion $s : [a, b] \to \mathbb{R}$ heißt Polynom-Spline vom Grad l, wenn sie die folgenden Eigenschaften besitzt:

1. *$s(z)$ ist ein Polynom vom Grad l für $z \in [\xi_j, \xi_{j+1})$, $1 \leq j < m$.*
2. *s ist $(l-1)$-mal stetig differenzierbar.*

Die reellen Zahlen $\xi_1, \ldots, \xi_m$ heißen Knoten des Splines. Die Menge $\Omega_m = \{\xi_1, \ldots, \xi_m\}$ wird Knotenmenge genannt.

Beispiel 3.8
Wir betrachten das Intervall $[0, 1]$ und die Unterteilung $0 = \xi_1 < \ldots < \xi_5 = 1$ mit $\xi_2 = 0.25$, $\xi_3 = 0.5$ und $\xi_4 = 0.75$. Abbildung 3.14 zeigt für diese Knotenmenge jeweils ein Beispiel für einen Spline vom Grad 0 (Abbildung a), 1 (Abbildung b) und 2 (Abbildung c). Bei Splines vom Grad 0 handelt es sich um stückweise konstante Funktionen. Splines vom Grad 0 sind die einzigen Splines, die an den Knoten nicht stetig sind. Splines vom Grad 1 sind Polygonzüge durch die Punkte $(\xi_j, s(\xi_j))$, $j = 1, \ldots, m$. An den Knoten sind diese zwar stetig, jedoch nicht differenzierbar. Splines vom Grad 2 und höher sind dann mindestens einmal stetig differenzierbar.

$\triangle$

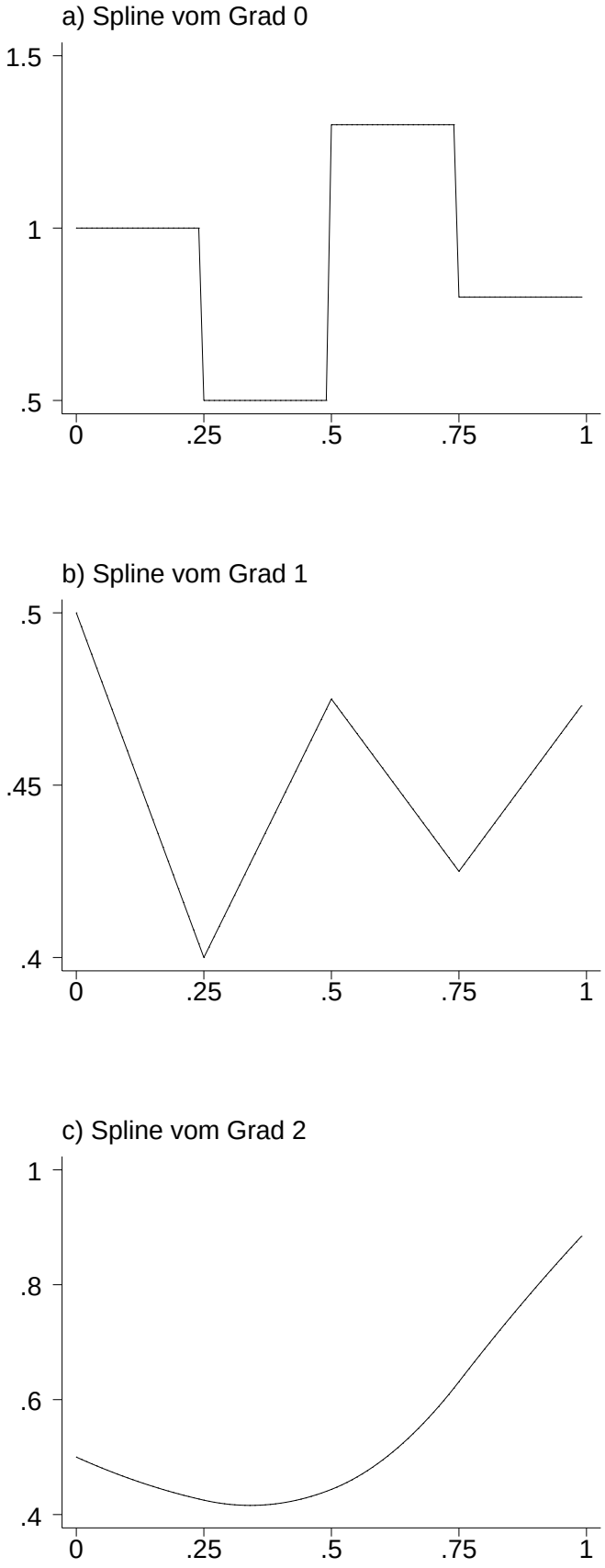

Abbildung 3.14. *Beispiele für Splines vom Grad 0 (Abbildung a), 1 (Abbildung b) und 2 (Abbildung c) zur Knotenmenge* $\{0,0.25,0.5,0.75,1\}$.

Wir befassen uns im Folgenden mit der geeigneten Darstellung von Splines durch Basisfunktionen, wobei wir uns im wesentlichen auf Hämmerlin und Hoffman (1990) beziehen. Man kann zeigen, dass es sich bei Polynomsplines um einen Vektorraum der Dimension $m + l - 1$ handelt. Wir bezeichnen den Vektorraum der Splines vom Grad l zur Knotenmenge Ω_m mit $S_l(\Omega_m)$. Offenbar handelt es sich um einen Unterraum des Vektorraums aller $l - 1$ mal stetig differenzierbaren Funktionen $C_{l-1}[a, b]$. Aufgrund der endlichen Dimension des Vektorraums, können wir jeden Spline als Linearkombination von $m + l - 1$ Basisfunktionen darstellen, d.h.

$$s(z) = \beta_0 B_0(z) + \cdots + \beta_{m+l-2} B_{m+l-2}(z).$$

Bei gegebenen Basisfunktionen ist ein Spline dann eindeutig durch die Koeffizienten $\beta = (\beta_0, \ldots, \beta_{m+l-2})'$ bestimmt. Die Basisfunktionen sind natürlich nicht eindeutig. Im Folgenden werden wir die zwei gebräuchlichsten Basen für Splines behandeln, die *truncated power series Basis* und die *B-spline Basis*. Die truncated power series Darstellung eines Splines ist gegeben durch

$$s(z) = \sum_{j=0}^{l} \beta_j z^j + \sum_{j=2}^{m-1} \beta_{l+j-1} (z - \xi_j)_+^l$$

wobei

$$(z - \xi_j)_+^l = \begin{cases} (z - \xi_j)^l & z \geq \xi_j \\ 0 & \text{sonst.} \end{cases}$$

Bei den Basisfunktionen handelt es sich also um Potenzfunktionen und abgeschnittene Potenzen, d.h.

$$\begin{aligned} B_0(z) &= 1 \\ B_1(z) &= z \\ &\vdots \\ B_l(z) &= z^l \\ B_{l+1}(z) &= (z - \xi_2)_+^l \\ &\vdots \\ B_{l+m-2}(z) &= (z - \xi_{m-1})_+^l. \end{aligned}$$

Beispiel 3.9

Wir betrachten die Knotenmenge {0,0.25,0.5,0.75,1} und die Splines aus Beispiel 3.8. Die Abbildung 3.15 zeigt die zur Berechnung benötigten Basisfunktionen für den Spline vom Grad 1, die Abbildung 3.16 die Basisfunktionen für den Spline vom Grad 2.

△

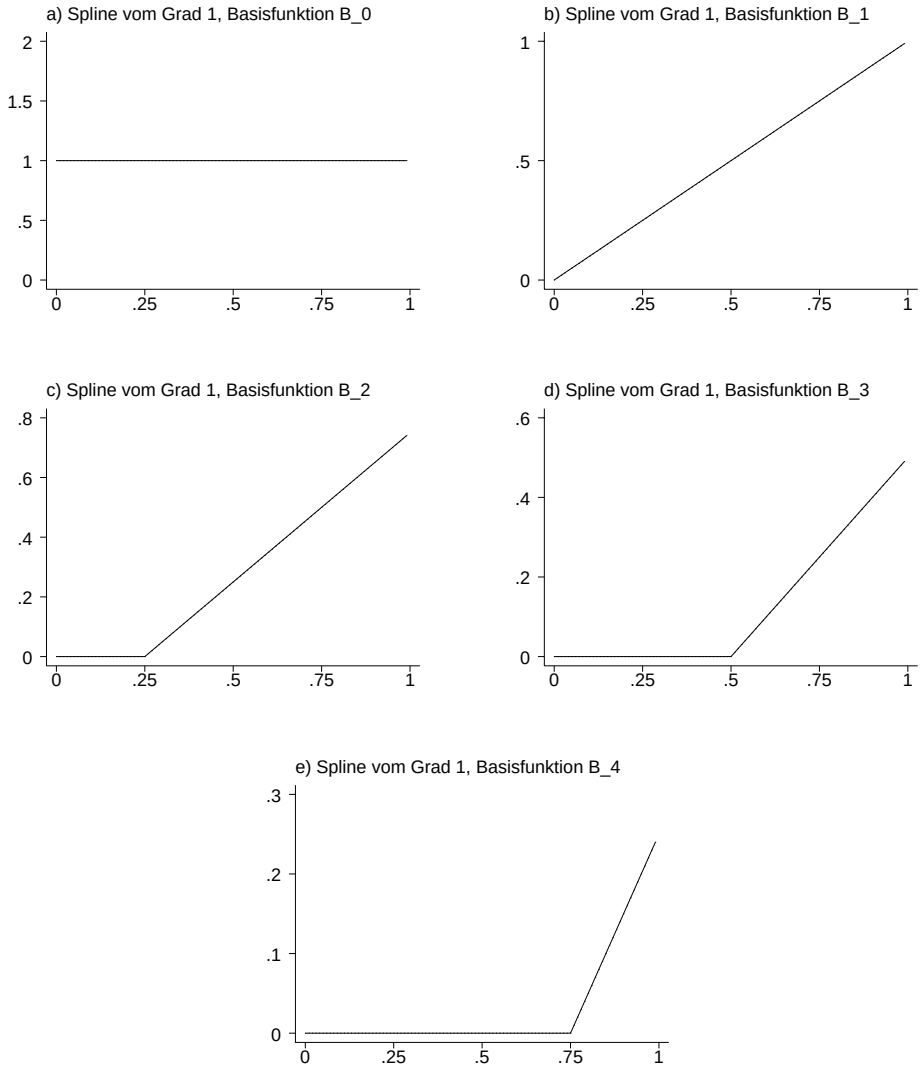

Abbildung 3.15. *Basisfunktionen B_0-B_4 für den Spline vom Grad 1 aus Beispiel 3.8.*

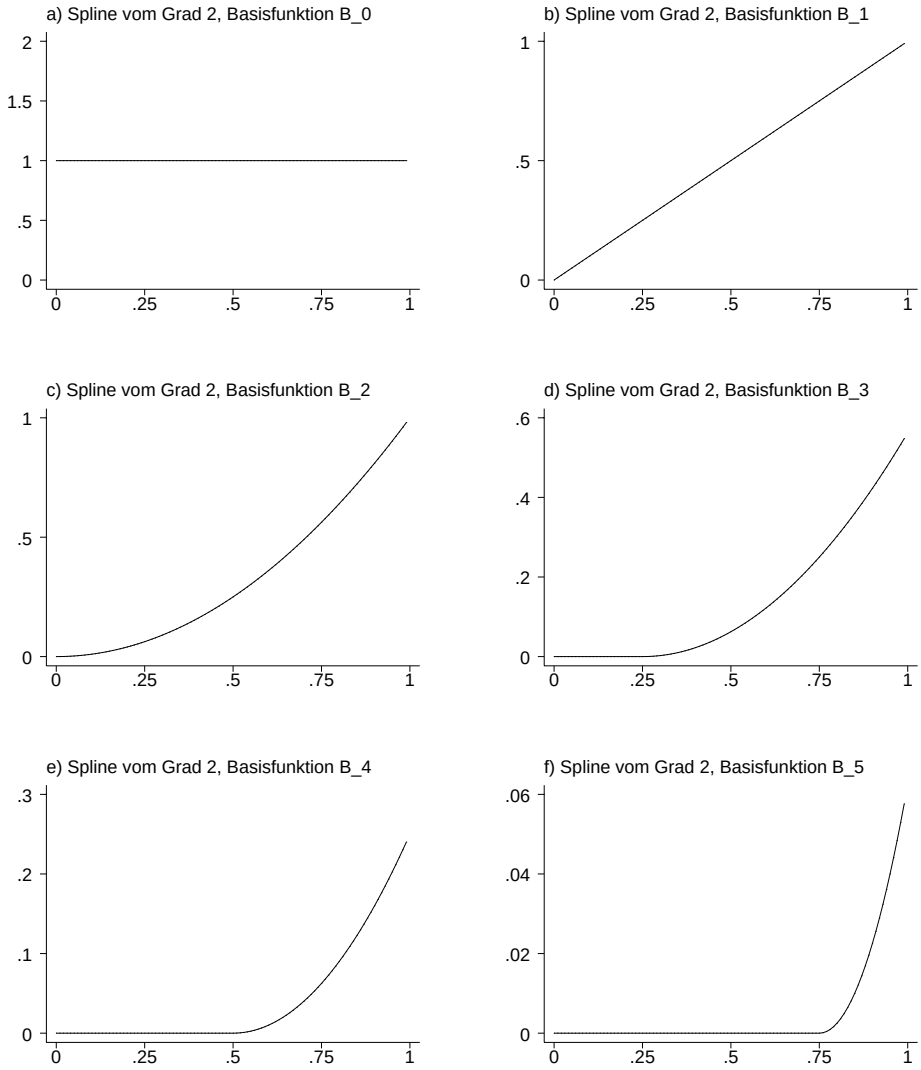

Abbildung 3.16. *Basisfunktionen B_0-B_5 für den Spline vom Grad 2 aus Beispiel 3.8.*

Unter der Annahme, dass die unbekannte Funktion f im Modell 3.11 durch einen Spline approximiert werden kann, und unter Verwendung einer truncated power series Basis erhalten wir

$$y_i = f(x_i) + \epsilon_i = \sum_{j=0}^{l} \beta_j x_i^j + \sum_{j=2}^{m-1} \beta_{l+j-1}(x_i - \xi_j)_+^l + \varepsilon_i$$

und für die Designmatrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^l & (x_1 - \xi_2)_+^l & \cdots & (x_1 - \xi_{m-1})_+^l \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^l & (x_n - \xi_2)_+^l & \cdots & (x_n - \xi_{m-1})_+^l \end{pmatrix}.$$

Beispiel 3.10

Wir betrachten wieder die Motorcycledaten und die Mietspiegeldaten. Abbildung 3.17 zeigt Scatterplotsmoother für die Abhängigkeit der Beschleunigung und der Zeit unter Verwendung von Splines mit truncated power series Basis. In Abbildung a) wurden zwischen der minimal und maximal beobachteten Zeit 3 äquidistante Knoten verwendet, in Abbildung b) 6 äquidistante Knoten und in Abbildung c) 10 Knoten. Offenbar kann im Gegensatz zur Modellierung mit einfachen Polynomen eine befriedigende Anpassung an die Daten erzielt werden, vorausgesetzt die Anzahl der Knoten ist groß genug. Offensichtlich spielt die geeignete Wahl der Anzahl (und auch der Position) der Knoten eine entscheidende Rolle bei der Splineregression. Abbildung 3.18 zeigt Scatterplotsmoother basierend auf Splines für den Zusammenhang zwischen Nettomiete pro Quadratmeter und Wohnfläche. Hier scheint die Abhängigkeit von der Anzahl der Knoten geringer. Es ist aber typisch, dass die Schätzungen mit wachsender Knotenzahl rauer werden.

△

Untersuchungen haben gezeigt, dass die Verwendung der truncated power series Basis unter Umständen zu numerischen Problemen führen kann. Eine numerisch stabilere Basis ist die B-Spline Basis. Diese kann rekursiv definiert werden ausgehend von den Basisfunktionen für Splines vom Grad 0 (stückweise konstante Funktionen). Wir definieren

$$B_j^0(z) = \begin{cases} 1 & \xi_j \leq z < \xi_{j+1} \\ 0 & \text{sonst} \end{cases} \tag{3.13}$$

und

$$B_j^l(z) = \frac{z - \xi_j}{\xi_{j+l} - \xi_j} B_j^{l-1}(z) + \frac{\xi_{j+l+1} - z}{\xi_{j+l+1} - \xi_{j+1}} B_{j+1}^{l-1}(z) \tag{3.14}$$

für $l \geq 1$. Dann erhalten wir für $z \in [a, b]$

$$s(z) = \sum_{j=-l+1}^{m-1} \beta_j B_j^l(z)$$

Zur Berechnung der Basisfunktionen benötigen wir $2l$ zusätzliche Knoten $\xi_{-l+1}, \ldots, \xi_{-1}, \xi_0$ und $\xi_{m+1}, \ldots, \xi_{m+l}$ mit

$$\xi_{-l+1} < \xi_{-l+2} < \ldots < \xi_1 < \ldots < \xi_m < \xi_{m+1} < \ldots < \xi_{m+l}.$$

Die Knotenmenge $\{\xi_{-l+1}, \ldots, \xi_0, \xi_1, \ldots, \xi_m, \xi_{m+1}, \ldots, \xi_{m+l}\}$ bildet die sogenannte *erweiterte Partition.*

Beispiel 3.11
Wir betrachten wieder die Knotenmenge {0,0.25,0.5,0.75,1} und die Splines aus Beispiel 3.8. Dazu berechnen wir zunächst die Basisfunktionen für Splines vom Grad 0. Dafür gilt

$$s(z) = \beta_1 B_1^0(z) + \beta_2 B_2^0(z) + \beta_3 B_3^0(z) + \beta_4 B_4^0(z).$$

Die Definition einer erweiterten Partition ist hier noch nicht nötig. Abbildung 3.19 zeigt die vier Basisfunktionen B_1^0-B_4^0. Zur Berechnung der Basisfunktionen für Splines vom Grad 1 benötigen wir zwei zusätzliche Knoten ξ_0 und ξ_6. Wir wählen $\xi_0 = -0.25$ und $\xi_6 = 1.25$. Es gilt

$$s(z) = \beta_0 B_0^1(z) + \beta_1 B_1^1(z) + \beta_2 B_2^1(z) + \beta_3 B_3^1(z) + \beta_4 B_4^1(z).$$

Abbildung 3.20 a) zeigt die 5 Basisfunktionen B_0^1-B_4^1. Zur Berechnung der Basisfunktionen für Splines vom Grad 2 benötigen wir zwei weitere Knoten ξ_{-1} und ξ_7. Wir wählen $\xi_{-1} = -0.5$ und $\xi_7 = 1.5$. Abbildung 3.20 b) zeigt die 6 Basisfunktionen B_{-1}^2-B_4^2.

△

Wie bei der truncated power series Basis können wir die unbekannten Koeffizienten

$$\beta = (\beta_{-l+1}, \ldots, \beta_{m-1})'$$

wieder im Rahmen der linearen Modelle schätzen. Die Designmatrix $\mathbf{X}$ besteht wieder aus den Basisfunktionen ausgewertet an den Beobachtungen x_i, d.h.

$$\mathbf{X} = \begin{pmatrix} B_{-l+1}^l(x_1) & \cdots & B_{m-1}^l(x_1) \\ \vdots & \ddots & \vdots \\ B_{-l+1}^l(x_n) & \cdots & B_{m-1}^l(x_n) \end{pmatrix}.$$

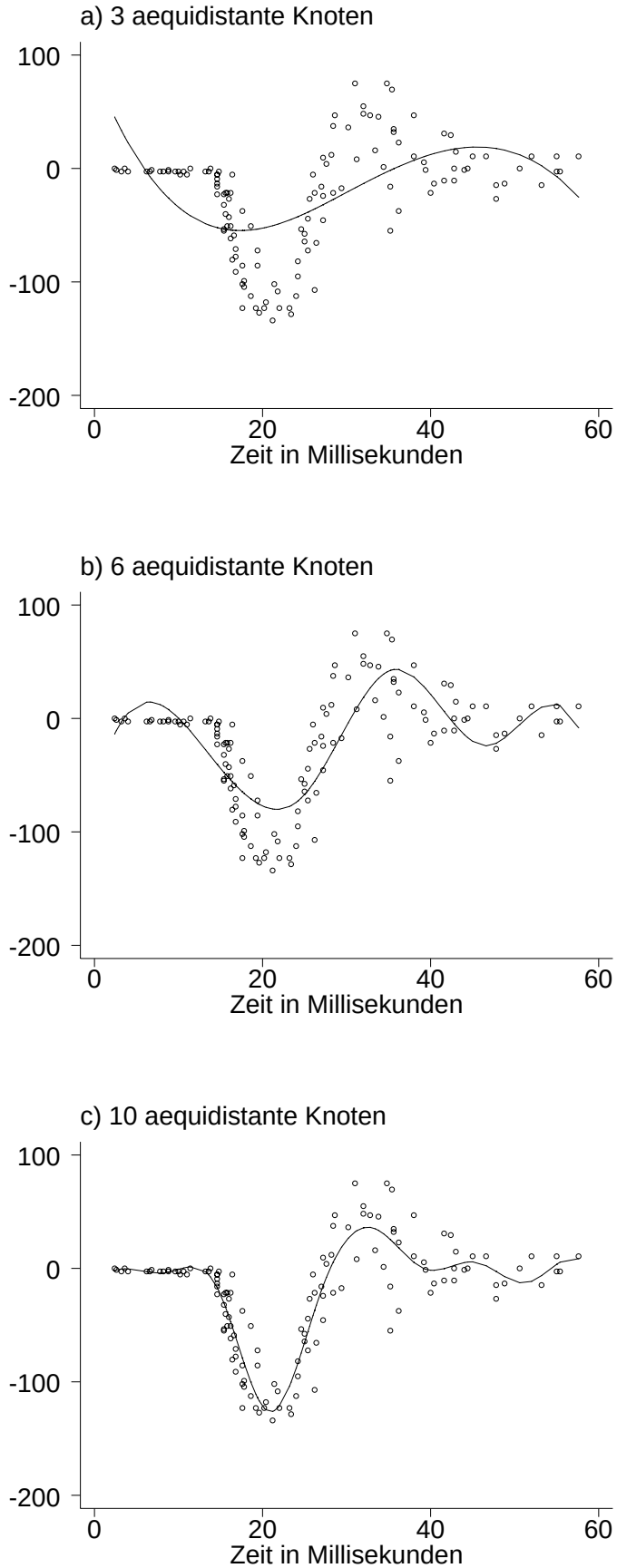

Abbildung 3.17. *Motorcycledaten: Glättungssplines mit truncated power series Basis.*

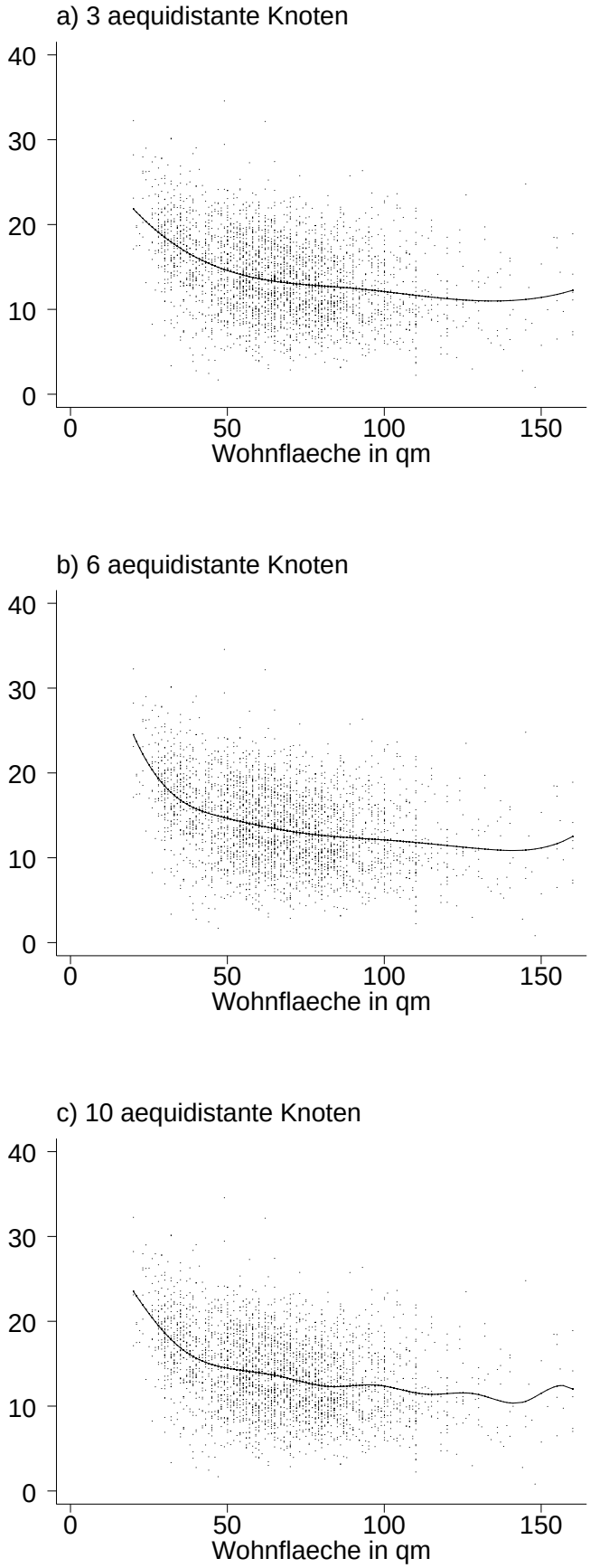

Abbildung 3.18. *Mietspiegeldaten: Glättungssplines mit truncated power series Basis.*

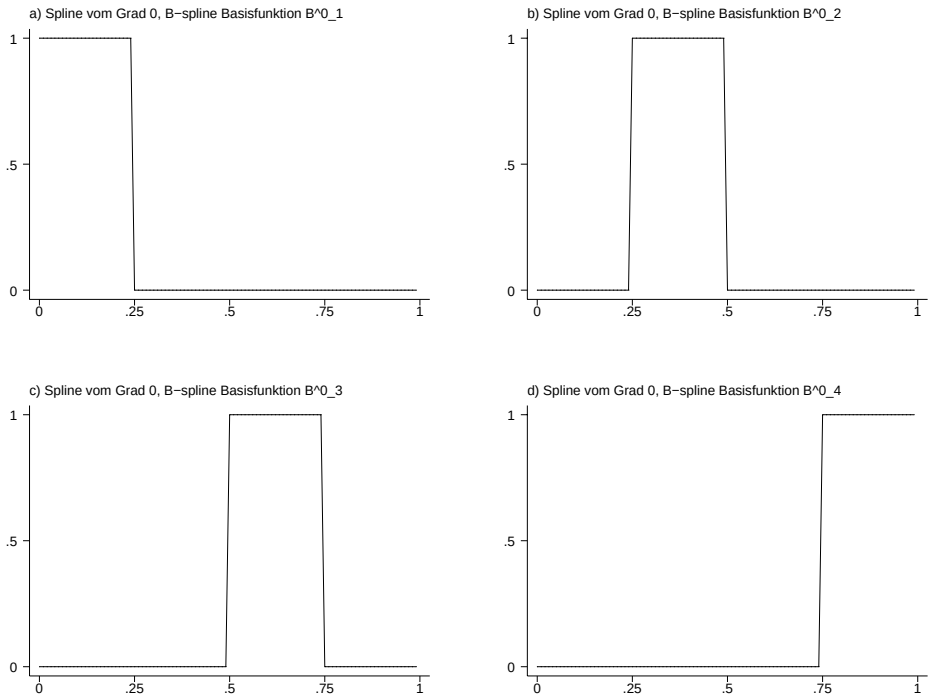

Abbildung 3.19. *B-spline Basisfunktionen* B_1^0*-*B_4^0 *für den Spline vom Grad 0 aus Beispiel 3.8.*

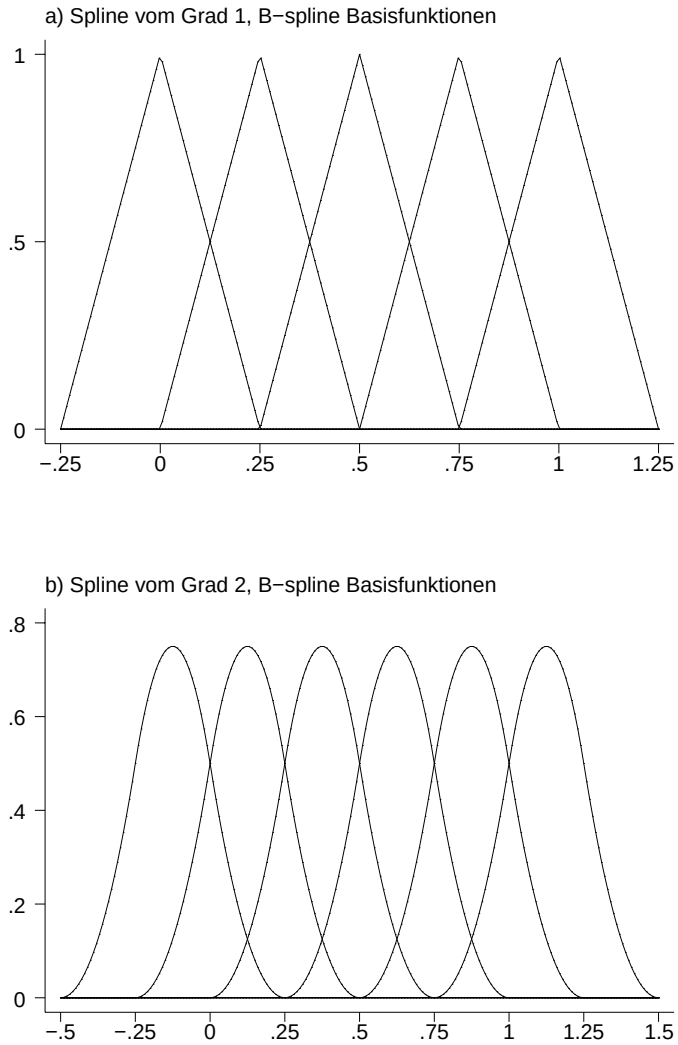

Abbildung 3.20. *Die erste Grafik zeigt die B-spline Basisfunktionen B_0^1-B_4^1 für den Spline vom Grad 1 aus Beispiel 3.8. Die zweite Grafik zeigt die Basisfunktionen B_{-1}^2-B_4^2 für den Spline vom Grad 2.*

Man macht sich leicht klar, dass B-spline Basisfunktionen folgende Eigenschaften besitzen, (vgl. auch die Abbildungen 3.19 und 3.20):

– Für $z \in [a, b]$ gilt $\sum_j B_j(z) = 1$, d.h. die Zeilensumme der Designmatrix ist eins.

– B-splines sind in einem Bereich von $2 + l$ Knoten größer als Null und überlappen mit $2 \cdot l$ benachbarten B-Splines.

– An jedem Punkt z sind $l + 1$ B-Splines größer als Null.

Wie man dem Beispiel 3.10 entnimmt, kommt der Anzahl und auch der Position der Knoten eine entscheidende Bedeutung bei der Anpassung an die Daten zu. Im Folgenden wollen wir zunächst einige einfache Möglichkeiten für die Wahl der Knoten behandeln. Wir unterscheiden die *äquidistante Knotenwahl* und die *Wahl gemäß der Quantile der unabhängigen Variable*. Bei der äquidistanten Knotenwahl unterteilen wir das Intervall $[x_{min}, x_{max}]$ in $m - 1$ äquidistante Intervalle der Länge

$$h = \frac{x_{max} - x_{min}}{m - 1}$$

und erhalten die Knoten

$$\xi_j = x_{min} + (j - 1) \cdot h$$

mit $j = 1, \ldots, m$ bei Verwendung der truncated power series Basis und $j = -l+1, \ldots, m+l$ bei Verwendung einer B-spline Basis. Werden die Knoten an den Quantilen der unabhängigen Variable definiert, dann erhalten wir die Knoten

$$\begin{aligned}
\xi_1 &= x_{min} \\
\xi_2 &= x_{(\frac{100}{m-1})} \\
\xi_3 &= x_{(2 \cdot \frac{100}{m-1})} \\
&\vdots \\
\xi_m &= x_{max}.
\end{aligned}$$

Für die erweiterte Partition bei B-splines können wir als Abstand der Knoten beispielsweise den Abstand der letzten beiden Knoten benutzen.

Obige einfache Regeln zur Wahl der Knoten können nur eine erste einfache Lösung sein. Die grundsätzlichen Probleme bleiben bestehen:

– *Wie viele* Knoten sollen spezifiziert werden?

– *Wo* sollen die Knoten platziert werden?

Zur Lösung des Problems werden wir in den folgenden Abschnitten grundsätzlich zwei Strategien verfolgen:

– *Penalisierungsansätze:* Hier wird eine relativ große Anzahl von Knoten bzw. Basisfunktionen definiert, um einen Funktionenraum zu erhalten, der flexibel genug ist eine gute Anpassung an die Daten zu gewährleisten. Zu rauhe Funktionsschätzungen werden verhindert, indem die Regressionskoeffizienten geeignet penalisiert werden. Insbesondere wird verhindert, dass benachbarte Regressionskoeffizienten zu starke Sprünge aufweisen. Im Folgenden werden wir *P(enalized)-Splines* und *Glättungssplines* behandeln.

– *Ansätze mit adaptiver Knotenwahl:* Ziel dieser Ansätze ist, durch geeignete (adaptive) Knotenwahl eine gute Anpassung an die Daten zu erhalten bei gleichzeitiger sparsamer Parameterzahl. In der Literatur existieren mittlerweile eine Fülle konkurierender Ansätze, z.B. das CART Verfahren (Classification And Regression Trees) oder MARS (Multivariate Adaptive Regression).

3.4 Penalisierungsansätze I: P-splines

3.4.1 Grundidee

P-Splines basieren grundsätzlich auf den folgenden drei Grundannahmen Eilers und Marx (1996):

– Wir nehmen an, dass die unbekannte Funktion f durch einen Spline vom Grad l (üblicherweise $l = 3$) approximiert werden kann, d.h.

$$f(x) = \sum_{j=0}^{p} \beta_j B_j(x)$$

wobei B_j eine B-Spline Basis ist.[1]

– Definiere eine relativ große Anzahl äquidistanter Knoten (ca. 20 - 40), um ausreichende Flexibilität des Splineraums zu gewährleisten.

– Zu starke Abweichungen benachbarter Regressionskoeffizienten β_j werden durch Strafterme basierend auf quadrierten Differenzen k-ter Ordnung bestraft.

Anstelle der bei einfachen Basisfunktionsansätzen verwendeten Residuenquadratsumme

[1] Um eine einfachere und mit vorherigen Abschnitten konsistentere Darstellung zu erhalten, numerieren wir im Folgenden die Basisfunktionen wieder von 0 bis p (anstatt von $-l+1$ bis $-m-1$) wie bei den linearen Modellen oder bei den einfachen Basisfunktionsansätzen. Außerdem unterdrücken wir die Indizierung der Basisfunktionen mit dem Grad des Splines l (außer wenn dies fürs Verständnis notwendig ist).

$$S(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j B_j(x_i) \right)^2$$

wird die *penalisierte* Residuenquadratsumme

$$SP(\beta) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j B_j(x_i) \right)^2 + \lambda \sum_{j=k+1}^{p} (\Delta^k \beta_j)^2 \tag{3.15}$$

bezüglich $\beta = (\beta_0, ..., \beta_p)'$ minimiert. Das Symbol Δ^k bezeichnet den Differenzenoperator k-ter Ordnung:

$$\begin{aligned}
\Delta^0 \beta_j &= \beta_j \\
\Delta^1 \beta_j &= \beta_j - \beta_{j-1} \\
\Delta^2 \beta_j &= \Delta\beta_j - \Delta\beta_{j-1} \\
&= \beta_j - \beta_{j-1} - \beta_{j-1} + \beta_{j-2} \\
&= \beta_j - 2\beta_{j-1} + \beta_{j-2} \\
&\vdots
\end{aligned}$$

Der Strafterm $\sum_{j=k+1}^{p} (\Delta^k \beta_j)^2$ verhindert eine zu starke Anpassung an die Daten und damit ein überfitten. Der Trade off zwischen Datentreue und Glattheit wird durch den zusätzlichen Parameter λ gesteuert. Der Parameter λ wird *Glättungsparameter* genannt.

3.4.2 Penalisierte KQ-Schätzung

Wir gehen im Folgenden zunächst davon aus, dass der Glättungsparameter gegeben ist. Zur Bestimmung der optimalen Regressionskoeffizienten, schreiben wir die penalisierte Residuenquadratsumme (3.15) in Matrixnotation. Dazu definieren wir zunächst die Matrix $\mathbf{D}_k$ der k-ten Differenzen rekursiv durch

$$\begin{aligned}
\mathbf{D}_0 &= \mathbf{I}_{p+1}, \\
\mathbf{D}_1 &= \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}, \\
\mathbf{D}_k &= \mathbf{D}_1 \mathbf{D}_{k-1} \qquad k > 1.
\end{aligned}$$

Die Differenzenmatrizen k-ter Ordnung wurden dabei so konstruiert, dass das Matrixprodukt $\mathbf{D}_k \beta$ als Resultat den Vektor der k-ten Differenzen ergibt. Beispielsweise erhalten wir für $k = 1$

$$\mathbf{D}_1\beta = \begin{pmatrix} \beta_1 - \beta_0 \\ \beta_2 - \beta_1 \\ \vdots \\ \beta_p - \beta_{p-1} \end{pmatrix}.$$

Mit Hilfe der Differenzenmatrizen $\mathbf{D}_k$ schreiben wir den Strafterm in Matrixnotation:

$$\sum_{j=k}^{p} (\Delta^k \beta_j)^2 = \beta' \mathbf{D}_k' \mathbf{D}_k \beta.$$

Schließlich erhalten wir unter Verwendung der Designmatrix

$$\mathbf{X} = \begin{pmatrix} B_0(x_1) & \cdots & B_p(x_1) \\ \vdots & & \vdots \\ B_0(x_n) & \cdots & B_p(x_n) \end{pmatrix}$$

die Residuenquadratsumme $SP(\beta)$ in Matrixnotation:

$$\begin{aligned} SP(\beta) &= \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j B_j(x_i) \right)^2 + \lambda \sum_{j=k}^{p} (\Delta^k \beta_j)^2 \\ &= (y - \mathbf{X}\beta)'(y - \mathbf{X}\beta) + \lambda \beta' \mathbf{D}_k' \mathbf{D}_k \beta \\ &= y'y - \beta'\mathbf{X}'y - y'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta + \lambda\beta'\mathbf{D}_k'\mathbf{D}_k\beta \\ &= y'y - 2y'\mathbf{X}\beta + \beta'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D}_k'\mathbf{D}_k)\beta \end{aligned}$$

Differenzieren nach β und Nullsetzen liefert

$$-2\mathbf{X}'y + 2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D}_k'\mathbf{D}_k)\beta = 0.$$

Auflösen nach β ergibt

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{D}_k'\mathbf{D}_k)^{-1}\mathbf{X}'y = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_k)^{-1}\mathbf{X}'y \tag{3.16}$$

wobei die Matrix $\mathbf{P}_k := \mathbf{D}_k'\mathbf{D}_k$ auch als *Strafmatrix* bezeichnet wird. Für $k = 0$ erhalten wir $\mathbf{P}_0 = \mathbf{I}_{p+1}$. Für $k = 1, 2$ ergibt sich

$$\mathbf{P}_1 = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix} \tag{3.17}$$

bzw.

$$\mathbf{P}_2 = \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}. \tag{3.18}$$

Bemerkungen:

– Bei der Herleitung des penalisierten KQ-Schätzers wurde die Ableitung einer Funktion $f(\beta)$ eines Vektors $\beta = (\beta_0, \ldots, \beta_p)'$ benutzt. Diese ist allgemein definiert als der Vektor der partiellen Ableitungen, d.h.

$$\frac{\delta f}{\delta \beta} = \left(\frac{\delta f}{\delta \beta_0}, \ldots, \frac{\delta f}{\delta \beta_p} \right)'.$$

Es wurden folgende (leicht beweisbare) Regeln benutzt:

– Für eine $p+1 \times p+1$ Matrix $\mathbf{A}$ gilt

$$\frac{\delta \beta' \mathbf{A} \beta}{\delta \beta} = \begin{cases} 2\mathbf{A}\beta & \text{falls } \mathbf{A} \text{ symmetrisch} \\ (\mathbf{A} + \mathbf{A}')\beta & \text{sonst.} \end{cases}$$

– Für den $p+1 \times 1$ Spaltenvektor a gilt

$$\frac{\delta a' \beta}{\delta \beta} = a.$$

– Bei der praktischen Berechnung von $\hat{\beta}$ muss das lineare Gleichungssystem

$$(\mathbf{X}'\mathbf{X} + \lambda \mathbf{P}_k)\beta = \mathbf{X}'y$$

nach β aufgelöst werden. Dabei kann man ausnützen, dass die Matrix $\mathbf{X}'\mathbf{X} + \lambda \mathbf{P}_k$ Bandmatrixgestalt hat mit Bandweite $\max\{l, k\}$.

Ausgehend von der Schätzung $\hat{\beta}$ für die Regressionskoeffizienten können wir auch eine Schätzung für den Vektor der Funktionswerte $f = (f(x_1), \ldots, f(x_n))'$ bestimmen. Wir erhalten

$$\hat{f} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{P}_k)^{-1}\mathbf{X}'y = \mathbf{S}y$$

mit der Smoothermatrix

$$\mathbf{S} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{P}_k)^{-1}\mathbf{X}'.$$

Bei P-Splines handelt es sich also wieder um einen linearen Smoother gemäß Definition 3.12.

Beispiel 3.12 (Motorcycledaten)
Wir betrachten wieder die Motorcycledaten. Die Abbildungen 3.21 und 3.22 zeigen 6 Scatterplotsmoother basierend auf P-splines vom Grad 3 (kubische Splines) und 2. Differenzen als Penalty. Dabei wurden verschiedene Glättungsparameter λ verwendet. Offensichtlich hängt die Güte der Schätzung entscheidend von der Wahl des Glättungsparameters ab. Wir werden deshalb im übernächsten Abschnitt Verfahren zur möglichst optimalen Wahl des Glättungsparameters behandeln. Das gebräuchlichste Verfahren beruht auf der Kreuzvalidierung, die uns bereits im Zusammenhang mit der nichtparametrischen Dichteschätzung begegnet ist. Abbildung 3.21 c) basiert auf einem kreuzvalidierungsoptimalen Glättungsparameter.

△

Aus obigem Beispiel ist bereits der starke Einfluss der Wahl des Glättungsparameters erkennbar. Allgemein gilt:

$\lambda \to 0$	relativ rauhe unpenalisierte Schätzung,
λ "klein"	relativ datentreu, eher rauher Schätzer,
λ "groß"	relativ glatte Schätzung,
$\lambda \to \infty$	bei einem Strafterm der Ordnung k, und $l \geq k$ ergibt sich ein Polynom vom Grad $k-1$.

Wir begründen den Grenzfall $\lambda \to \infty$ für den Fall $k = 2$ und $l \geq 2$. Zunächst gilt für die zweite Ableitung (de Boor (1978))

$$s''(z) = \sum_{j=0}^{p} \beta_j (B_j^l)''(z) = \frac{1}{h^2} \sum_{j=2}^{p} \Delta^2 \beta_j (B_j^{l-2})''(z). \tag{3.19}$$

wobei h der Abstand der (äquidistanten) Knoten ist. Für großes λ muss der Strafterm $\sum_{j=2}^{p} (\Delta^2 \beta_j)^2$ gegen Null streben und damit die zweiten Differenzen $\Delta^2 \beta_j$. Dies bedeutet aber, dass die zweite Ableitung (3.19) ebenfalls gegen Null strebt., d.h. konstant ist. Dies bedeutet, dass s nur ein Polynom vom Grad 1 sein kann.

Beispiel 3.13
Wir verdeutlichen den Grenzfall $\lambda \to \infty$ nochmal anhand der Motorcycledaten. Abbildung 3.23 zeigt kubische P-Spline Schätzungen mit Straftermen basierend auf Differenzen 3. Ordnung. Dabei wurde der Glättungsparameter λ schrittweise erhöht. Zusätzlich sind in die jeweiligen Graphiken geschätzte Polynome zweiten Grades eingezeichnet. Offensichtlich nähert sich der P-spline mit wachsendem λ immer mehr dem Polynom an.

△

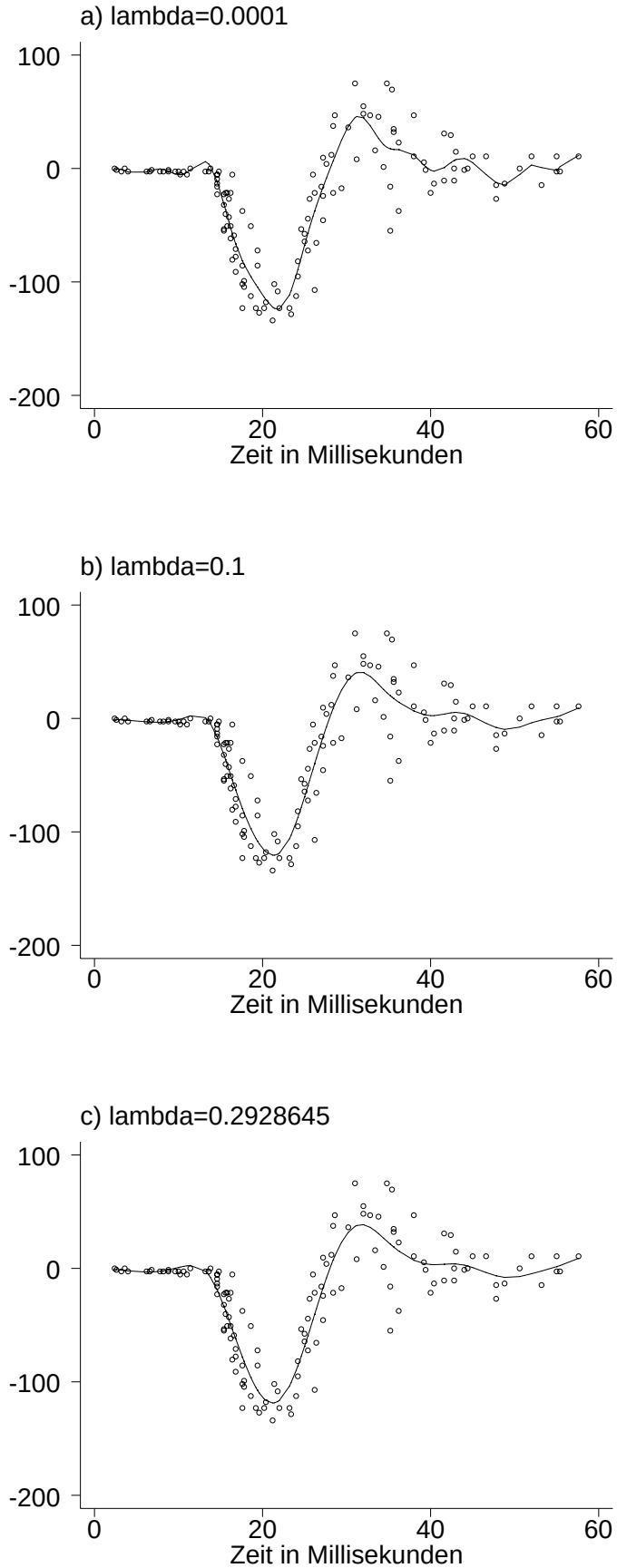

Abbildung 3.21. *Motorcycledaten: P-splines vom Grad 3 mit 2. Differenzen als Penalty für verschiedene Werte von λ.CV-optimales λ: 0.2928645*

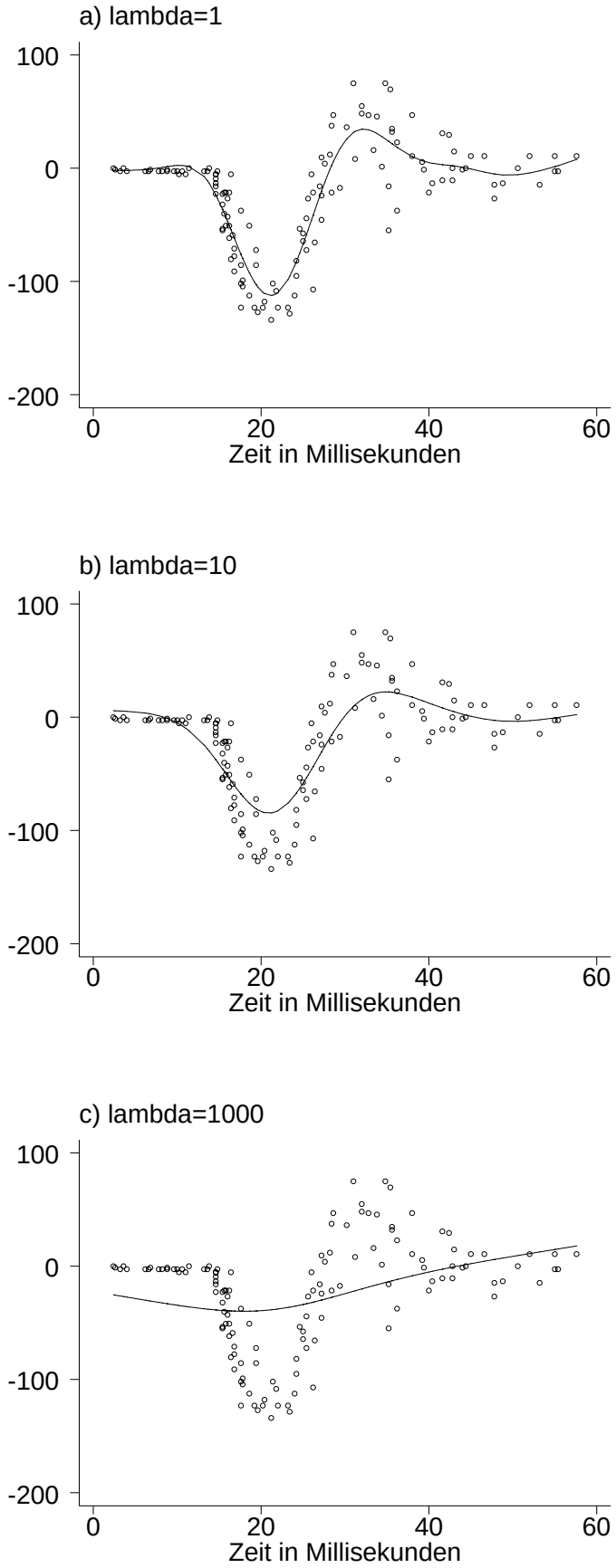

Abbildung 3.22. *Motorcycledaten: P-splines vom Grad 3 mit 2. Differenzen als Penalty für verschiedene Werte von λ.*

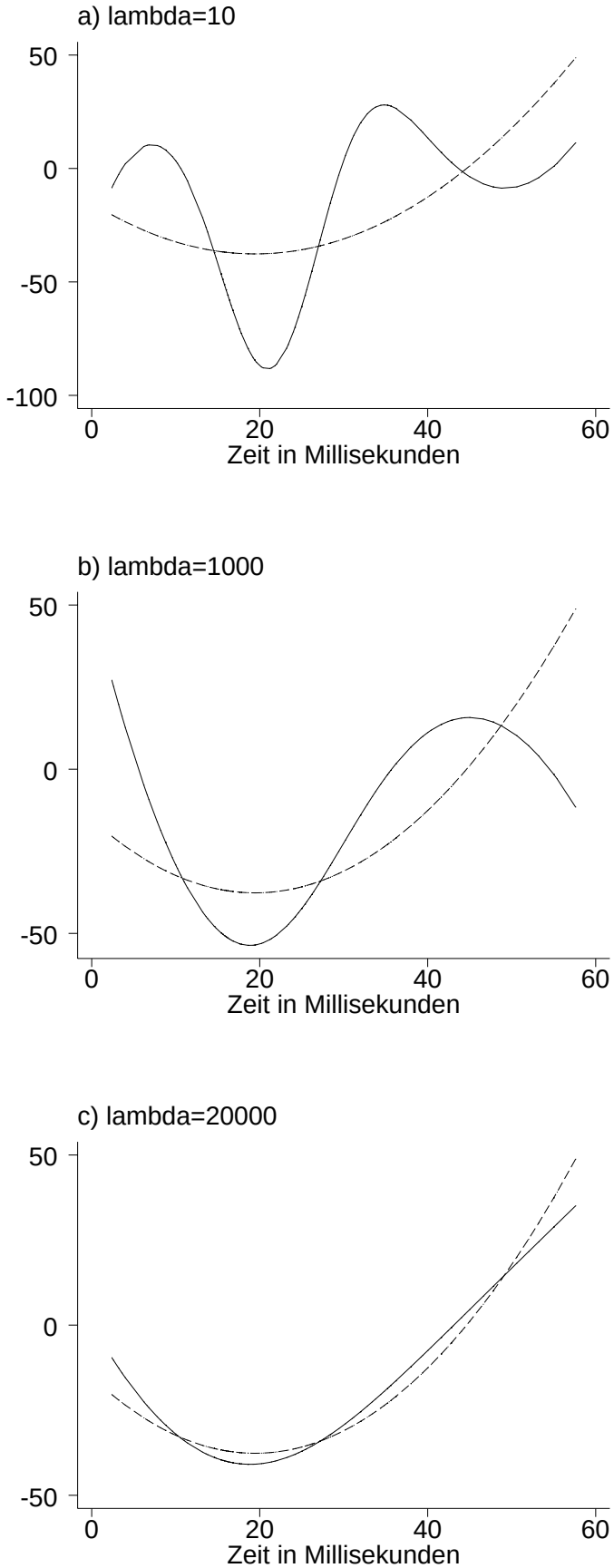

Abbildung 3.23. *Motorcycledaten: P-splines vom Grad 3 mit 3. Differenzen als Penalty. Für große Werte von lambda nähert sich die Schätzung einer polynomialen Regression vom Grad 2 (gestrichelte Linie) an.*

3.4.3 Bayesianische Variante von P-splines

Bayesianische Inferenz unterscheidet sich fundamental von der frequentistischen Sichtweise, die in der Likelihoodtheorie eingenommen wird. Während in der Likelihoodtheorie die unbekannten Parameter als feste, nichtstochastische Größen aufgefasst werden, sind diese im Bayes–Ansatz wie die beobachtbaren Größen ebenfalls zufällig. Ein bayesianischer Ansatz besteht daher im wesentlichen aus

– der Spezifikation der *bedingten Verteilung der beobachtbaren Größen* bei gegebenen Parametern, dem sogenannten *Beobachtungsmodell* und

– aus der Spezifikation einer sogenannten *a priori Verteilung der unbekannten Parameter*, in der das subjektive Vorwissen des Statistikers über die unbekannten Parameter einfließt.

Bevor wir uns der Bayesianischen Variante von P-splines zuwenden, demonstrieren wir die Konstruktion und anschließende Schätzung eines Bayesmodells anhand des linearen Regressionsmodells. Wenn wir vom klassischen linearen Regressionsmodell (3.5) ausgehen, erhalten wir als Beobachtungsmodell

$$y \,|\, \beta \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n). \tag{3.20}$$

Der Einfachheit halber nehmen wir an, dass die Varianz σ^2 bekannt ist. Als a priori Verteilung für die unbekannten Paramter β wählen wir eine multivariate Normalverteilung d.h.

$$\beta \sim N_{p+1}(\mu_0, \Sigma_0).$$

Nun folgt nach dem Satz von Bayes für die bedingte Verteilung von β bei gegebenen Daten y

$$P(\beta \,|\, y) = \frac{P(y \,|\, \beta) P(\beta)}{\int P(y \,|\, u) P(u)\, du} \tag{3.21}$$

Von dieser sogenannten *posteriori Verteilung* von β gehen nun im Bayes–Ansatz sämtliche Inferenzschlüsse bezüglich der unbekannten Parameter aus. Als Punktschätzer für β eignen sich beispielsweise der Erwartungswert oder der Modus (sofern eindeutig) der posteriori Verteilung. Das Analogon zum Konfidenzbereich ist im Bayes'schen Kontext ein Bereich des Parameterraums, der eine vorher festgelegte Wahrscheinlichkeitsmasse enthält (z.B. 90 Prozent). Zweckmäßigerweise wählt man einen Bereich aus, in dem die Dichte in jedem Punkt größer ist als in jedem Punkt außerhalb des Bereichs. Ein solcher Bereich heißt HPD (Highest Posterior Density) Bereich. Grundsätzlich ist man aber im Bayes'schen Kontext nicht nur an Punktschätzern oder Konfidenzbereichen interessiert, sondern vielmehr an

der Gewinnung von möglichst vielen Eigenschaften der posteriori Verteilung bis hin zu einer Schätzung der Dichte der Verteilung.

Wir bestimmen nun die posteriori Verteilung von β, d.h. Verteilungstyp, Erwartungswert und Varianz. Dazu betrachten wir zunächst allgemein die Dichte einer beliebigen multivariat normalverteilten Zufallsvariable z der Dimension p mit Erwartungswert μ und Kovarianzmatrix Σ. Für die Dichte gilt

$$p(z) = \frac{1}{\sqrt{2\pi}^p \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right). \tag{3.22}$$

Unter Vernachlässigung aller Faktoren in (3.22), die nicht von z abhängen, erhalten wir

$$\begin{aligned} p(z) &\propto \exp\left(-\tfrac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right) \\ &= \exp\left(-\tfrac{1}{2}z'\Sigma^{-1}z + z'\Sigma^{-1}\mu - \tfrac{1}{2}\mu'\Sigma^{-1}\mu\right) \\ &\propto \exp\left(-\tfrac{1}{2}z'\Sigma^{-1}z + z'\Sigma^{-1}\mu\right). \end{aligned} \tag{3.23}$$

D.h. die Dichte einer multivariat normalverteilten Zufallsvariable ist stets proportional zu obiger Darstellung (3.23). Wir zeigen im Folgenden, dass die posteriori Verteilung (3.21) ebenfalls die Darstellung (3.23) besitzt und daher normalverteilt ist. Es gilt

$$\begin{aligned} P(\beta \,|\, y) &= \frac{P(y \,|\, \beta)P(\beta)}{\int P(y \,|\, u)P(u)\,du} \\ &\propto P(y \,|\, \beta)P(\beta) \\ &= \tfrac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left(-\tfrac{1}{2\sigma^2}(y-\mathbf{X}\beta)'(y-\mathbf{X}\beta)\right) \\ &\quad \tfrac{1}{\sqrt{2\pi}^p\sqrt{|\Sigma_0|}} \exp\left(-\tfrac{1}{2}(\beta-\mu_0)'\Sigma_0^{-1}(\beta-\mu_0)\right) \\ &\propto \exp\left(-\tfrac{1}{2\sigma^2}(y-\mathbf{X}\beta)'(y-\mathbf{X}\beta)\right) \exp\left(-\tfrac{1}{2}(\beta-\mu_0)'\Sigma_0^{-1}(\beta-\mu_0)\right) \\ &\propto \exp\left(-\tfrac{1}{2\sigma^2}y'y + \tfrac{1}{\sigma^2}\beta'\mathbf{X}'y - \tfrac{1}{2\sigma^2}\beta'\mathbf{X}'\mathbf{X}\beta\right) \\ &\quad \exp\left(-\tfrac{1}{2}\beta'\Sigma_0^{-1}\beta + \beta'\Sigma_0^{-1}\mu_0\right) \\ &\propto \exp\left(\tfrac{1}{\sigma^2}\beta'\mathbf{X}'y - \tfrac{1}{2\sigma^2}\beta'\mathbf{X}'\mathbf{X}\beta - \tfrac{1}{2}\beta'\Sigma_0^{-1}\beta + \beta'\Sigma_0^{-1}\mu_0\right) \\ &= \exp\left(\beta'(\tfrac{1}{\sigma^2}\mathbf{X}'y + \Sigma_0^{-1}\mu_0) - \tfrac{1}{2}\beta'(\tfrac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \Sigma_0^{-1})\beta\right) \end{aligned}$$

Damit besitzt die Posterioriverteilung dieselbe Form wie (3.23), d.h. es handelt sich um eine multivariate Normalverteilung. Für die Kovarianzmatrix gilt

$$\Sigma_\beta = (\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \Sigma_0^{-1})^{-1}.$$

Der posteriori Erwartungswert ergibt sich als Lösung des Gleichungssystems

$$\Sigma_\beta^{-1}\mu_\beta = \frac{1}{\sigma^2}\mathbf{X}'y + \Sigma_0^{-1}\mu_0,$$

d.h. es gilt

$$\mu_\beta = \Sigma_\beta(\frac{1}{\sigma^2}\mathbf{X}'y + \Sigma_0^{-1}\mu_0).$$

Zusammenfassend erhalten wir

$$\beta|y \sim N(\mu_\beta, \Sigma_\beta).$$

Für $\Sigma_0^{-1} \to 0$, d.h. bei verschwindender priori Information, erhalten wir als Spezialfall den gewöhnlichen KQ-Schätzer.

Nach diesem Einführungsbeispiel wenden wir uns jetzt wieder den P-Splines zu. Bei P-Splines nehmen wir an, dass die unbekannte Funktion f durch einen Spline modelliert werden kann und minimieren die penalisierte Residuenquadratsumme

$$SP(\beta) = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j B_j(x_j)\right)^2 + \lambda\sum_{j=k}^{p}(\Delta^k\beta_j)^2$$

bezüglich β. Zur Definition bayesianischer P-splines benötigen wir ein Beobachtungsmodell und eine geeignete Prioriverteilung für β. Wenn wir analog zum einfachen linearen Modell annehmen, dass die Beobachtungen y_i bei gegebenen Regressionsparametern normalverteilt, erhalten wir

$$y|\beta \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}), \tag{3.24}$$

wobei die Designmatrix aus den B-spline Basisfunktionen besteht, ausgewertet an den Beobachtungen x_i. Zur Definition einer priori Verteilung für die Regressionskoeffizienten verwenden wir die stochastischen Analoga zu den Straftermen $(\Delta^k\beta_j)^2$, wobei wir uns beschränken auf $k = 1$ (Differenzen erster Ordnung) und $k = 2$ (Differenzen zweiter Ordnung). Die stochastischen Analoga zu Straftermen basierend auf ersten Differenzen sind *Random walks erster Ordnung*. Diese können über die bedingte Verteilung der Parameter β_j gegeben β_{j-1} definiert werden. Genauer nehmen wir an, dass für $j = 1, \ldots, p$

$$\beta_j = \beta_{j-1} + u_j \tag{3.25}$$

mit unabhängigen identisch verteilten Abweichungen $u_j \sim N(0, \tau^2)$. Der Einfachheit halber nehmen wir wieder an, dass die Varianzen σ^2 und τ^2 bekannt seien. Äquivalent zu (3.25) ist

$$\beta_j|\beta_{j-1} \sim N(\beta_{j-1}, \tau^2).$$

Für den ersten Parameter β_0 soll eine Gleichverteilung auf $\mathbb{R}$ angenommen werden. Wir sprechen auch von einer diffusen Verteilung für β_0 und schreiben

$$\beta_0 \propto konstant.$$

Analog zu ersten Differenzen sind die stochastischen Analoga zu zweiten Differenzen *Random walks zweiter Ordnung.* Diese werden über die bedingten Verteilungen von β_j gegeben β_{j-1} und β_{j-2} definiert als

$$\beta_j = 2\beta_{j-1} - \beta_{j-2} + u_j, \qquad j = 2, \ldots, p. \tag{3.26}$$

Hierzu ist

$$\beta_j | \beta_{j-1}, \beta_{j-2} \sim N(2\beta_{j-1} - \beta_{j-2}, \tau^2)$$

äquivalent. Für die beiden Anfangsparameter setzen wir wieder

$$\beta_0 \propto konstant$$

bzw.

$$\beta_1 \propto konstant.$$

Ausgehend von den Definitionen (3.25) bzw. (3.26) können wir die gemeinsame Verteilung der Regressionsparameter β unter Zuhilfenahme des Satzes von der totalen Wahrscheinlichkeit berechnen. Für Random walks erster Ordnung gilt

$$\begin{aligned} P(\beta) &= \prod_{j=1}^{p} P(\beta_j | \beta_{j-1}) \\ &= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\beta_j - \beta_{j-1})^2\right) \\ &= \frac{1}{(2\pi\tau^2)^{p/2}} \exp\left(-\frac{1}{2\tau^2}\sum_{j=1}^{p}(\beta_j - \beta_{j-1})^2\right) \\ &= \frac{1}{(2\pi\tau^2)^{p/2}} \exp\left(-\frac{1}{2\tau^2}\beta'\mathbf{P}_1\beta\right) \end{aligned}$$

wobei $\mathbf{P}_1$ die in (3.17) definierte Strafmatrix ist. Analog erhalten wir für Random walks zweiter Ordnung

$$P(\beta) = \prod_{j=2}^{p} P(\beta_j | \beta_{j-1}, \beta_{j-2}) = \frac{1}{(2\pi\tau^2)^{(p-1)/2}} \exp\left(-\frac{1}{2\tau^2}\beta'\mathbf{P}_2\beta\right)$$

mit der in (3.18) definierten Strafmatrix. Ausgehend vom Beobachtungsmodell (3.24) und den Prioriverteilungen für die Regressionsparameter erhalten wir für die Posterioriverteilung

$$P(\beta \,|\, y) \propto P(y \,|\, \beta)P(\beta) \propto \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)'(y - \mathbf{X}\beta) - \frac{1}{2\tau^2}\beta'\mathbf{P}_k\beta\right), \qquad k = 1, 2.$$

Durch völlig analoge Rechnung wie im linearen Modell kann man zeigen, dass β gegeben y normalverteilt ist mit Kovarianzmatrix

$$\Sigma_\beta = (\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \frac{1}{\tau^2}P_k)^{-1}$$

und Erwartungswert

$$\mu_\beta = \Sigma_\beta \mathbf{X}'y = (\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{\tau^2}P_k)^{-1}\mathbf{X}'y.$$

Definiert man $\lambda = \sigma^2/\tau^2$, so erkennt man, dass offensichtlich der posteriori Erwartungswert mit dem penalisierten KQ-Schätzer (3.16) übereinstimmt. Da bei der Normalverteilung Erwartungswert und Modus wegen der Symmetrie der Verteilung gleich sind, stimmt auch der posteriori Modus mit (3.16) überein.

Bisher haben wir Random walk prioris über die bedingten Verteilungen der Parameter β_j gegeben β_{j-1} bzw. β_{j-1}, β_{j-2} definiert. Ausgehend von der gemeinsamen Verteilung der Regressionsparameter β können wir auch die bedingten Verteilungen von β_j gegeben *alle* restlichen Parameter bestimmen. Wir bezeichnen diese Verteilung mit $\beta_j|\cdot$. Es wird sich herausstellen, dass die bedingte Verteilung lediglich von einigen wenigen Parametern abhängt, bei einem Randomwalk erster Ordnung von β_{j-1}, β_{j+1} und bei einem Randomwalk zweiter Ordnung von $\beta_{j-2}, \beta_{j-1}, \beta_{j+1}, \beta_{j+2}$.

Wir demonstrieren die Herleitung der bedingten priori Verteilungen von β_j gegeben die übrigen Parameter am Beispiel eines Randomwalks erster Ordnung. Dabei benutzen wir wieder dieselbe Technik wie bei der Herleitung der Posterioriverteilung selbst. Zunächst stellt man fest, dass die Verteilung von $\beta_j|\cdot$ proportional ist zur gemeinsamen Verteilung von β. Aus dieser streichen wir dann sämtliche Faktoren, die nicht von β_j abhängen. Für $j > 0$ und $j < p$ gilt

$$\begin{aligned}
P(\beta_j|\cdot) &\propto P(\beta) \\
&= \frac{1}{(2\pi\tau^2)^{p/2}} \exp\left(-\frac{1}{2\tau^2}\sum_{s=1}^{p}(\beta_s - \beta_{s-1})^2\right) \\
&\propto \exp\left(-\frac{1}{2\tau^2}(\beta_{j+1} - \beta_j)^2 - \frac{1}{2\tau^2}(\beta_j - \beta_{j-1})^2\right) \\
&= \exp\left(-\frac{1}{2\tau^2}(\beta_{j+1}^2 - 2\beta_{j+1}\beta_j + 2\beta_j^2 - 2\beta_j\beta_{j-1} + \beta_{j-1}^2)\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{2}{\tau^2}\beta_j^2 + \beta_j\frac{1}{\tau^2}(\beta_{j+1} + \beta_{j-1})\right).
\end{aligned}$$

Damit hat die bedingte Verteilung wieder die Gestalt (3.23), d.h. es handelt sich um eine Normalverteilung mit Varianz

$$\sigma^2_{\beta_j|\cdot} = \frac{\tau^2}{2}.$$

Der Erwartungswert ergibt sich aus der Gleichung

$$\frac{1}{\sigma^2_{\beta_j|\cdot}}\mu_{\beta_j|\cdot} = \frac{1}{\tau^2}(\beta_{j+1} + \beta_{j-1}).$$

Wir erhalten also

$$\mu_{\beta_j|\cdot} = \frac{1}{2}(\beta_{j+1} + \beta_{j-1}),$$

d.h. der bedingte Erwartungswert ist nichts anderes als das armithmetische Mittel des jeweils linken und rechten Nachbarn. Für die Spezialfälle $j = 0$ und $j = p$ gilt trivialerweise

$$\beta_0|\cdot \sim N(\beta_1, \tau^2)$$

und

$$\beta_p|\cdot \sim N(\beta_{p-1}, \tau^2).$$

Zusammenfassend erhalten wir für einen Randomwalk erster Ordnung:

$$\beta_j|\cdot \sim \begin{cases} N(\beta_1, \tau^2) & j = 0 \\ N(\frac{1}{2}(\beta_{j+1} + \beta_{j-1}), \frac{\tau^2}{2}) & j = 1, \ldots, p-1 \\ N(\beta_{p-1}, \tau^2) & j = p \end{cases} \tag{3.27}$$

Völlig analog berechnen wir für einen Randomwalk zweiter Ordnung

$$\beta_j|\cdot \sim \begin{cases} N(2\beta_1 - \beta_2, \tau^2) & j = 0 \\ N(\frac{4}{5}\beta_2 - \frac{1}{5}\beta_3 + \frac{2}{5}\beta_0, \frac{\tau^2}{5}) & j = 1 \\ N(-\frac{1}{6}\beta_{j+2} + \frac{2}{3}\beta_{j+1} + \frac{2}{3}\beta_{j-1} - \frac{1}{6}\beta_{j-2}, \frac{\tau^2}{6}) & j = 2, \ldots, p-2 \\ N(\frac{2}{5}\beta_p + \frac{4}{5}\beta_{p-2} - \frac{1}{5}\beta_{p-3}+, \frac{\tau^2}{5}) & j = p-1 \\ N(2\beta_{p-1} - \beta_{p-2}, \tau^2) & j = p. \end{cases} \tag{3.28}$$

Für die Erwartungswerte beider bedingter Verteilungen ergeben sich interessante Interpretationen. Es lässt sich nämlich zeigen, dass sich die Gewichte, mit denen die benachbarten Parameter in den bedingten Erwartungswerten gewichtet werden, aus einem KQ-Kalkül ableiten lassen. Wir demonstrieren die Vorgehensweise am Beispiel eines Randomwalks zweiter Ordnung. Ein naheliegendes Konstruktionsprinzip für den bedingten Erwartungswert von β_j gegeben die Nachbarparameter $\beta_{j-2}, \beta_{j-1}, \beta_{j+1}, \beta_{j+2}$ ergibt sich aus folgendem Regressionsansatz:

$$\begin{aligned}
\beta_{j-2} &= \alpha_0 + \alpha_1(j - (j-2)) + \alpha_2(j - (j-2))^2 + \varepsilon_{-2} = \alpha_0 + \alpha_1(-2) + \alpha_2(-2)^2 + \varepsilon_{-2} \\
\beta_{j-1} &= \alpha_0 + \alpha_1(j - (j-1)) + \alpha_2(j - (j-1))^2 + \varepsilon_{-1} = \alpha_0 + \alpha_1(-1) + \alpha_2(-1)^2 + \varepsilon_{-1} \\
\beta_{j+1} &= \alpha_0 + \alpha_1(j - (j+1)) + \alpha_2(j - (j+1))^2 + \varepsilon_{1} = \alpha_0 + \alpha_1(1) + \alpha_2(1)^2 + \varepsilon_{1} \\
\beta_{j+2} &= \alpha_0 + \alpha_1(j - (j+2)) + \alpha_2(j - (j+2))^2 + \varepsilon_{2} = \alpha_0 + \alpha_1(2) + \alpha_2(2)^2 + \varepsilon_{2}
\end{aligned}$$

Hier passen wir ein Polynom zweiten Grades an die Punkte $(\beta_{j-2}, -2)$, $(\beta_{j-1}, -1)$,$(\beta_{j+1}, 1)$ und $(\beta_{j+2}, 2)$ an. Als Designmatrix fungiert die Matrix

$$\begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}.$$

Der bedingte Erwartungswert von β_j gegeben die benachbarten Punkte ergibt sich nun durch einfaches Einsetzen in die Schätzgleichung, d.h.

$$E(\beta_j|\cdot) = \hat{\alpha}_0 + \hat{\alpha}_1(j-j) + \hat{\alpha}_2(j-j)^2 = \hat{\alpha}_0.$$

Die Durchführung der KQ-Methode liefert schließlich

$$\hat{\alpha}_0 = -\frac{1}{6}\beta_{j+2} + \frac{2}{3}\beta_{j+1} + \frac{2}{3}\beta_{j-1} - \frac{1}{6}\beta_{j-2}$$

und damit genau den in (3.28) erhaltenen Erwartungswert. Analog kann man zeigen, dass sich für einen Randomwalk erster Ordnung der bedingte Erwartungswert in (3.27) durch Anpassung eines Polynoms ersten Grades (also einer Regressionsgeraden) an die beiden Punkte $(\beta_{j-1}, -1)$ und $(\beta_{j-1}, 1)$ ergibt.

Die Konstruktionsprinzipien der bedingten Erwartungswerte sind nochmal in den Abbildungen 3.24 - 3.26 veranschaulicht.

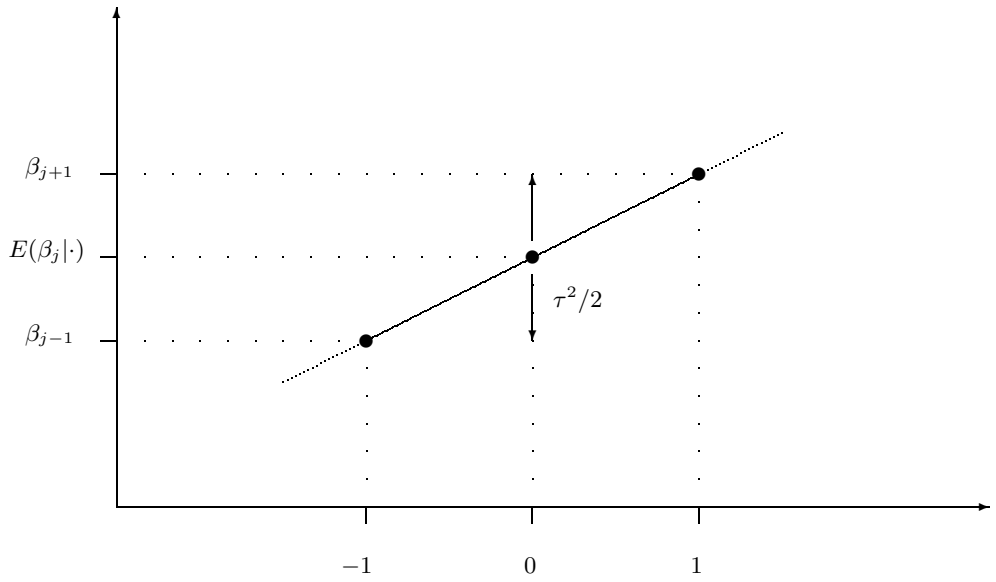

Abbildung 3.24. *Bayesianische P-splines: Veranschaulichung des bedingten Erwartungswerts von β_j gegeben β_{j-1}, β_{j+1} wenn ein Random Walk erster Ordnung verwendet wird.*

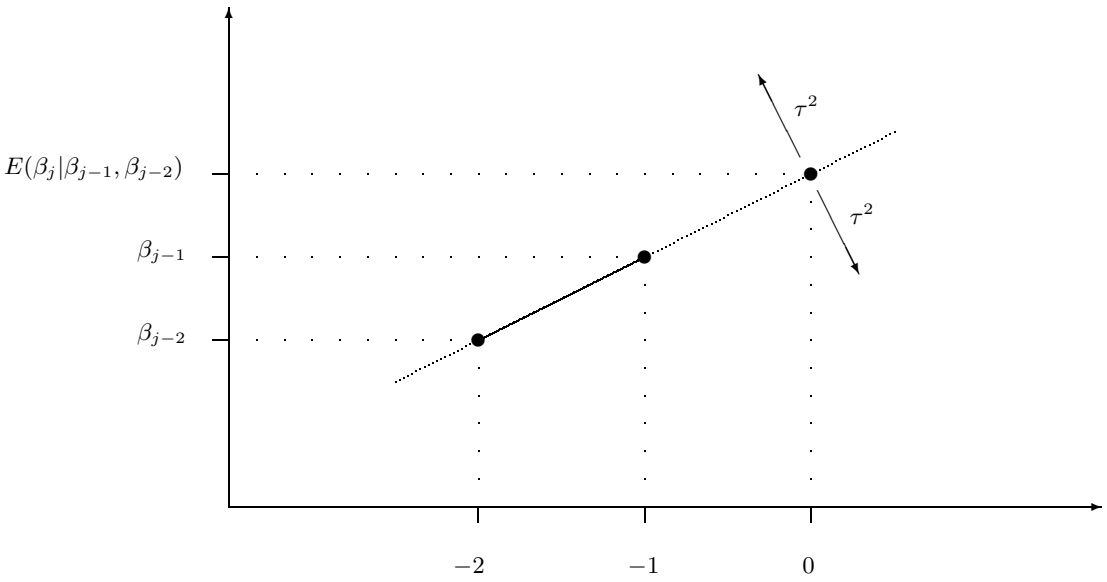

Abbildung 3.25. *Bayesianische P-splines: Veranschaulichung des bedingten Erwartungswerts von β_j gegeben β_{j-1}, β_{j-2} bei Verwendung eines Random Walks zweiter Ordnung.*

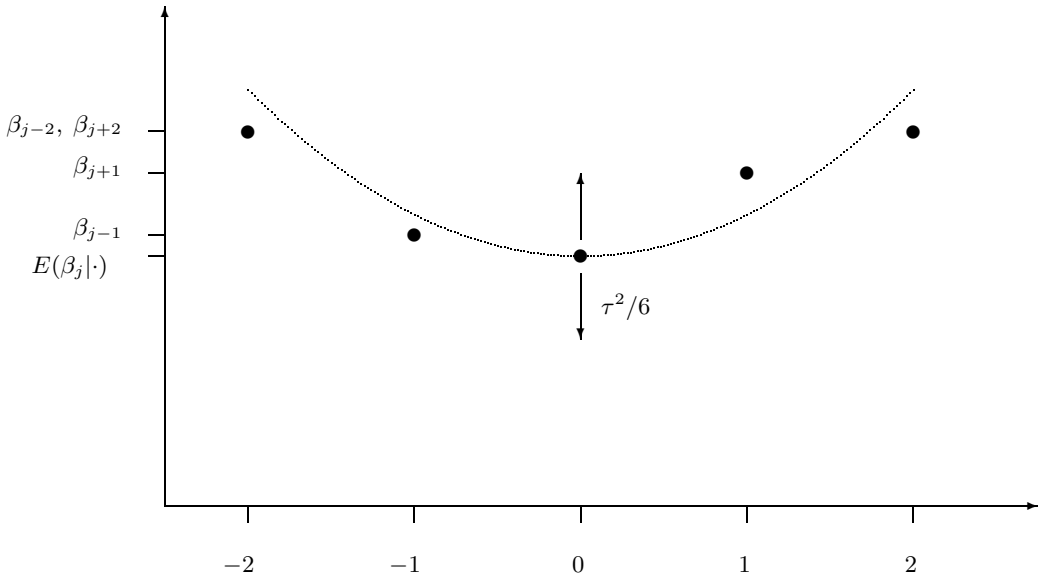

Abbildung 3.26. *Bayesianische P-splines: Veranschaulichung des bedingten Erwartungswerts von β_j gegeben $\beta_{j-1}, \beta_{j-2}, \beta_{j+1}, \beta_{j+2}$ bei Verwendung eines Random Walks zweiter Ordnung.*

3.4.4 Wahl des Glättungsparameters

Wir werden hier zunächst die Glättungsparameterwahl basierend auf *Kreuzvalidierung* behandeln. In den folgenden Abschnitten werden weitere Verfahren betrachtet.

Kreuzvalidierung basiert auf dem sogenannten mittleren Predicted Squared Error (PSE), der definiert ist als

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E(y_i^* - \hat{f}_\lambda(x_i))^2,$$

wobei y_i^* eine neue Beobachtung an der Stelle x_i ist. Eine Minimierung von PSE bezüglich λ ist nicht ohne weiteres möglich, da keine neuen Beoachtungen zur Verfügung stehen. Bei der Kreuzvalidierung wird versucht das Problem zu umgehen, indem neue Beobachtungen „imitiert“ werden. Die genaue Vorgehensweise ist wie folgt:

– Eliminiere für $i = 1, \ldots, n$ jeweils den i-ten Datenpunkt (y_i, x_i) aus dem Datensatz.

– Bestimme eine Vorhersage $\hat{f}_\lambda^{-i}(x_i)$ an der Stelle x_i basierend auf den verbleibenden $n-1$ Beobachtungen.

– Minimiere die Kreuzvalidierungsfunktion

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^{-i}(x_i))^2$$

bezüglich λ. Die Minimierung kann mit Hilfe einer einfachen Gittersuche durchgeführt werden, d.h. $CV(\lambda)$ wird für eine vorgegebene Menge von Glättungsparametern λ berechnet und derjenige Glättungsparameter bestimmt für den $CV(\lambda)$ minimal ist.

Dieses Vorgehen kann (grob) gerechtfertigt werden durch die Tatsache, dass

$$E(CV(\lambda)) \approx PSE(\lambda) \tag{3.29}$$

gilt. Zum „Nachweis“ von (3.29) berechnen wir zunächst

$$\begin{aligned} E(y_i - \hat{f}_\lambda^{-i}(x_i))^2 &= E((y_i - f(x_i)) + (f(x_i) - \hat{f}_\lambda^{-i}(x_i)))^2 \\ &= E(y_i - f(x_i))^2 + 2E(y_i - f(x_i))(f(x_i) - \hat{f}_\lambda^{-i}(x_i)) + \\ &\quad E(f(x_i) - \hat{f}_\lambda^{-i}(x_i))^2 \\ &= \sigma^2 + E(f(x_i) - \hat{f}_\lambda^{-i}(x_i))^2. \end{aligned}$$

Dabei haben wir in der zweiten Zeile augenützt, dass $y_i - f(x_i)$ und $f(x_i) - \hat{f}_\lambda^{-i}(x_i)$ unabhängig sind, weil $\hat{f}_\lambda^{-i}(x_i)$ nicht von y_i abhängt. Damit können wir den Erwartungswert des gemischten Glieds in der zweiten Zeile als Produkt der Erwartungswerte berechnen. Wegen $E(y_i - f(x_i)) = 0$ fällt dann das gemischte Glied weg. Analog können wir zeigen, dass

$$E(y_i^* - \hat{f}_\lambda(x_i)) = \sigma^2 + E(f(x_i) - \hat{f}_\lambda(x_i))^2.$$

Unter der Annahme $\hat{f}_\lambda^{-i}(x_i) \approx \hat{f}_\lambda(x_i)$ folgt schließlich $E(CV(\lambda)) \approx PSE(\lambda)$.

Zur möglichst schnellen Berechnung von $CV(\lambda)$ wäre es wünschenwert, dass wir $\hat{f}_\lambda^{-i}(x_i)$ aus den ursprünglichen Schätzungen $\hat{f}_\lambda(x_i)$ berechnen können. Auf diese Weise müssten wir nur *einmal* eine Schätzung $\hat{f}$ berechnen. Dazu zunächst der folgende

Satz 3.3
Sei für festes λ und i $\hat{f}_\lambda^{-i} = (\hat{f}_\lambda^{-i}(x_1), \ldots, \hat{f}_\lambda^{-i}(x_n))'$ der Vektor der Funktionsschätzungen an $x_1, \ldots, x_n$, wobei der Punkt (y_i, x_i) für die Schätzungen $\hat{f}_\lambda^{-i}(x_j)$, $j = 1, \ldots, n$, nicht berücksichtigt wurde. Sei der Vektor $y^* = (y_1^*, \ldots, y_n^*)'$ definiert durch

$$y_j^* = \begin{cases} y_j & i \neq j \\ \hat{f}_\lambda^{-i}(x_i) & i = j \end{cases}.$$

Dann gilt

$$\hat{f}_\lambda^{-i} = \mathbf{S}(\lambda) y^*,$$

wobei $\mathbf{S}(\lambda)$ die Smoothermatrix des Schätzers $\hat{f}_\lambda$ basierend auf allen Beobachtungen ist.

Beweis:

Sei g ein beliebiger B-Spline und seien $\beta = (\beta_0^g, \ldots, \beta_p^g)'$ die dazugehörigen Koeffizienten. Wie üblich bezeichnen wir die Koeffizienten von $\hat{f}_\lambda^{(-i)}$ mit β. Dann folgt

$$\begin{aligned}
\sum_{j=1}^n (y_j^* - g(x_j))^2 + \lambda \sum_{j=k}^p (\Delta^k \beta_j^g)^2 &\geq \\
\sum_{j \neq i} (y_j^* - g(x_j))^2 + \lambda \sum_{j=k}^p (\Delta^k \beta_j^g)^2 &\geq \\
\sum_{j \neq i} (y_j^* - \hat{f}_\lambda^{(-i)}(x_j))^2 + \lambda \sum_{j=k}^p (\Delta^k \beta_j)^2 &= \\
\sum_{j=1}^n (y_j^* - \hat{f}_\lambda^{(-i)}(x_j))^2 + \lambda \sum_{j=k}^p (\Delta^k \beta_j)^2, &
\end{aligned}$$

wobei das Gleichheitszeichen in der letzten Zeile wegen $y_i^* = \hat{f}_\lambda^{(-i)}(x_i)$ gilt. Damit haben wir gezeigt, dass $\hat{f}_\lambda^{(-i)}$ die penalisierte Residuenquadratsumme

$$\sum_{j=1}^n (y_j^* - g(x_j))^2 + \lambda \sum_{j=k}^p (\Delta^k \beta_j^g)^2$$

minimiert, d.h.

$$\hat{f}_\lambda^{(-i)} = \mathbf{S}(\lambda) y^*.$$

□

Aufgrund des Satzes folgt

$$\begin{aligned}
\hat{f}_\lambda^{(-i)}(x_i) - y_i &= \sum_{j=1}^{n} S_{ij}(\lambda) y_j^* - y_i \\
&= \sum_{i \neq j} S_{ij}(\lambda) y_j + S_{ii}(\lambda) \hat{f}_\lambda^{(-i)}(x_i) - y_i \\
&= \sum_{j=1}^{n} S_{ij}(\lambda) y_j - y_i + S_{ii}(\lambda) \left(\hat{f}_\lambda^{(-i)}(x_i) - y_i \right) \\
&= \hat{f}_\lambda(x_i) - y_i + S_{ii}(\lambda) \left(\hat{f}_\lambda^{(-i)}(x_i) - y_i \right).
\end{aligned}$$

Damit erhalten wir zunächst

$$(\hat{f}_\lambda^{(-i)}(x_i) - y_i)(1 - S_{ii}(\lambda)) = \hat{f}_\lambda(x_i) - y_i$$

bzw.

$$y_i - \hat{f}_\lambda^{(-i)}(x_i) = \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}(\lambda)}$$

und schließlich

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}(\lambda)} \right)^2. \tag{3.30}$$

Generalisierte Kreuzvalidierung

Betrachte das multiple Regressionsmodell

$$y_i = \beta_0 x_{i0} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad i = 1, \ldots, n.$$

Für $\hat{y}$ gilt

$$\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = \mathbf{H}y$$

wobei $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ die Hat-Matrix ist. Für die Spur der Hatmatrix erhalten wir

$$spur(\mathbf{H}) = spur(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = spur(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = spur(\mathbf{I}_{p+1}) = p + 1,$$

d.h. die Spur der Hatmatrix ist gleich der Anzahl der geschätzten Parameter bzw. der Anzahl der *Freiheitsgrade* des Modells. In Analogie dazu können wir allgemein für lineare Smoother

$$\hat{f} = \mathbf{S}y$$

sogenannte *äquivalenten Freiheitsgrade* eines Glätters definieren als

$$df_f = spur(\mathbf{S}).$$

Für df_f nahe 2 handelt es sich um eine annähernd lineare Beziehung zwischen Y und X. Je höher df_f desto größer die Nichtlinearität von f.

Bei der *generalisierten Kreuzvalidierung* wird der Faktor $1 - S_{ii}(\lambda)$ in (3.30) durch den Mittelwert

$$\frac{1}{n}\sum_{i=1}^{n}(1 - S_{ii}(\lambda)) = 1 - \frac{1}{n}sp(\mathbf{S}(\lambda))$$

ersetzt. Wir erhalten

$$GCV(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2}{\left(1 - \frac{1}{n}sp(\mathbf{S}(\lambda))\right)^2} = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2}{\left(1 - \frac{1}{n}df_f\right)^2} = n\frac{\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2}{(n - df_f)^2}.$$

Der Hauptgrund für die Verwendung der generalierten Kreuzvalidierung ist die mathematisch einfachere Berechenbarkeit von GCV.

Beispiel 3.14

Für die Motorcycledaten erhalten wir in Abbildung 3.27 a) die Kreuzvalidierungsfunktion $CV(\lambda)$. In der Abbildung b) findet man den dazugehörigen CV-optimalen P-Spline. Die entsprechenden Grafiken für die generalisierte Kreuzvalidierung sind in Abbildung 3.28 zu finden. Das CV-optimale λ ist ungefähr 0.35, der GCV-optimale Glättungsparameter ist ungefähr $\lambda = 0.55$. In beiden Fällen wurde das optimale λ auf einem Gitter mit 51 Punkten zwischen Werten von 10^{-8} und 10 gesucht mit $\lambda_j = 10^{-8+j\cdot 9/50}$, $j = 0, \ldots, 50$.

△

Beispiel 3.15

Für die Mietspiegeldaten erhalten wir in Abbildung 3.29 a) die Kreuzvalidierungsfunktion $CV(\lambda)$ für die Regression zwischen Nettomiete pro Quadratmeter und Wohnfläche. In der Abbildung b) findet man den dazugehörigen CV-optimalen P-Spline. Die entsprechenden Grafiken für die generalisierte Kreuzvalidierung sind in Abbildung 3.30 zu finden. Sowohl das CV-optimale als auch das GCV-optimale λ ist ungefähr 57. In beiden Fällen wurde das optimale λ auf einem Gitter mit 51 Punkten zwischen Werten von 10^{-4} und 10^4 gesucht.

△

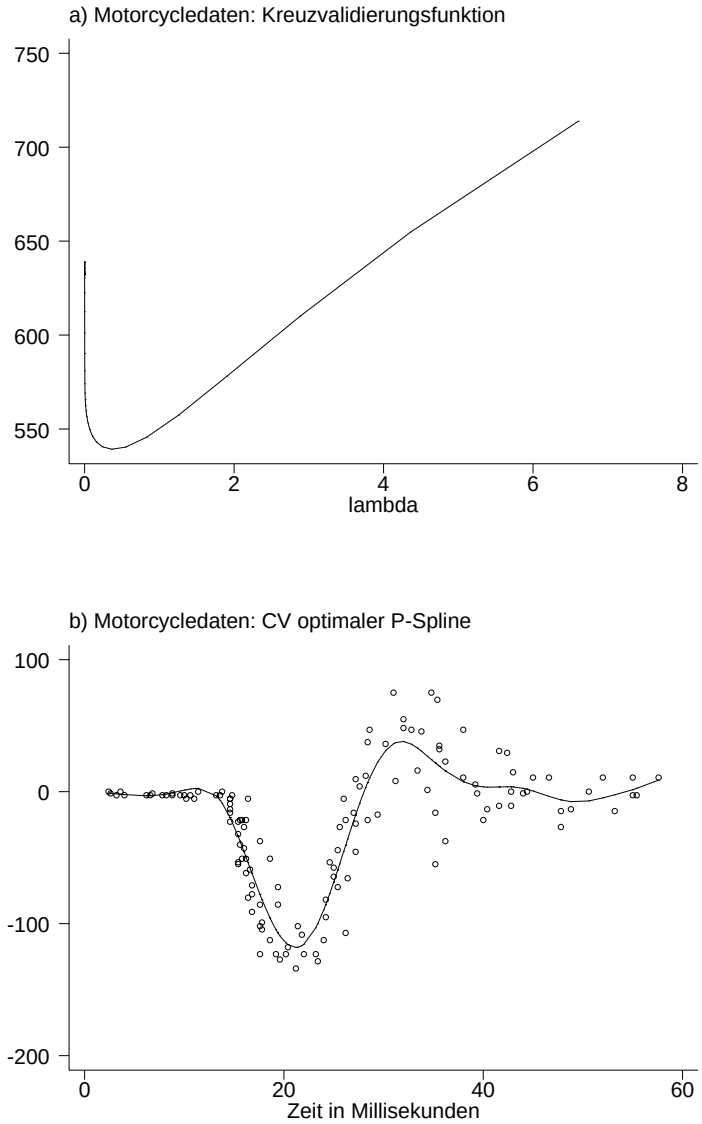

Abbildung 3.27. *Motorcycledaten: Kreuzvalidierungsfunktion $CV(\lambda)$ und geschätzter P-Spline basierend auf dem CV-optimalen $\lambda \approx 0.35$.*

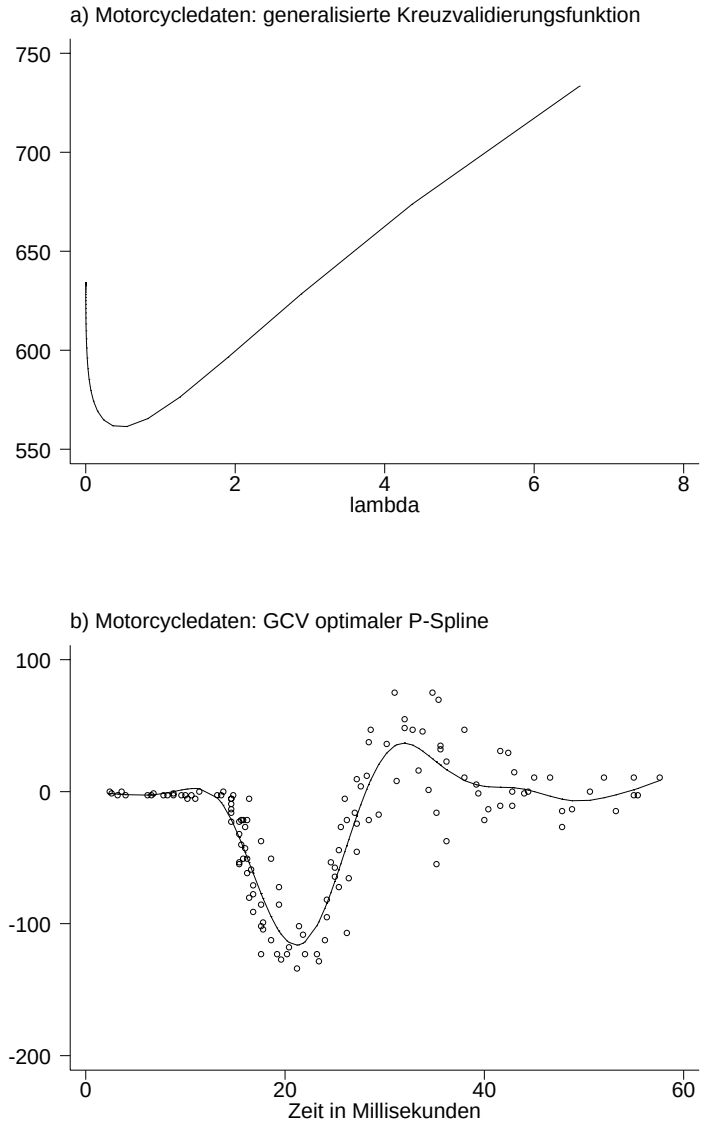

Abbildung 3.28. *Motorcycledaten: Generalisierte Kreuzvalidierungsfunktion $GCV(\lambda)$ und geschätzter P-Spline basierend auf dem GCV-optimalen $\lambda \approx 0.55$.*

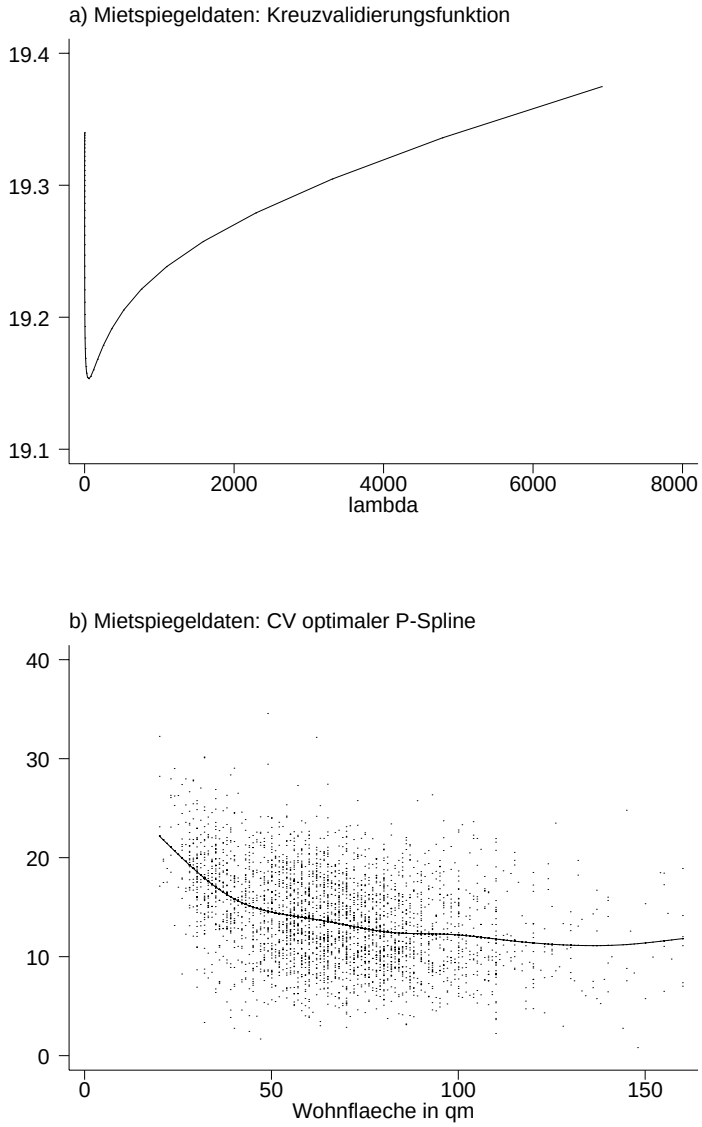

Abbildung 3.29. *Mietspiegeldaten: Kreuzvalidierungsfunktion $CV(\lambda)$ und geschätzter P-Spline für den Einfluss der Wohnfläche auf die Nettomiete pro qm basierend auf dem CV-optimalen $\lambda \approx 57$.*

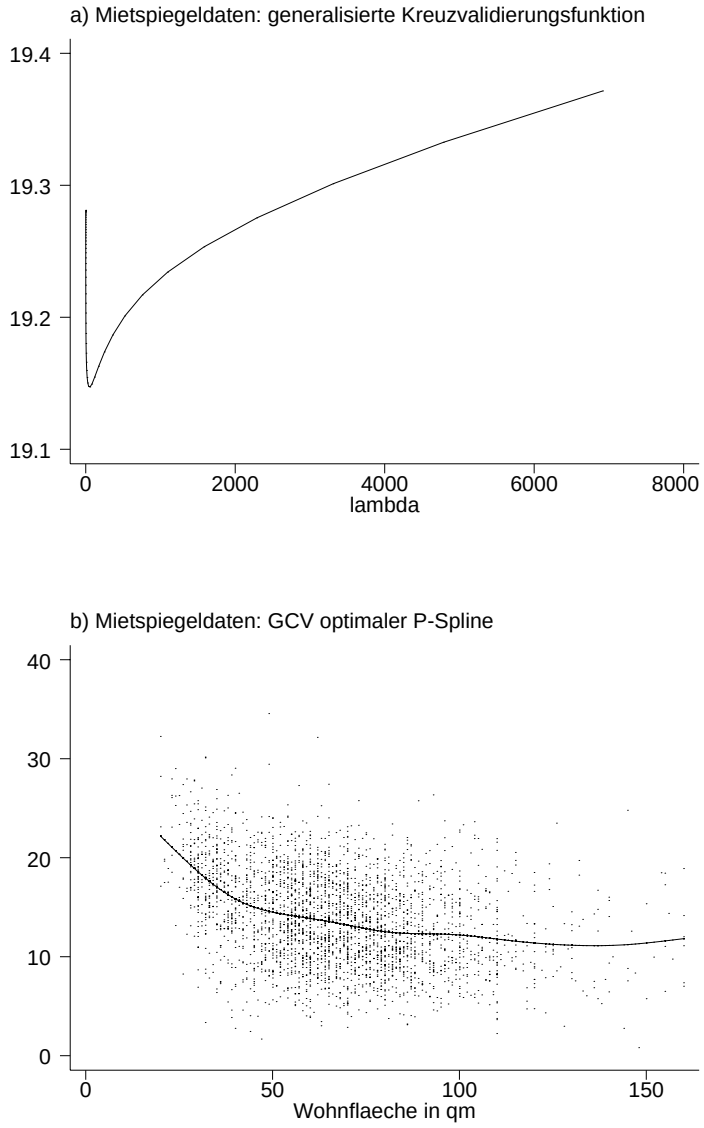

Abbildung 3.30. *Mietspiegeldaten: Generalisierte Kreuzvalidierungsfunktion $GCV(\lambda)$ und geschätzter P-Spline für den Einfluss der Wohnfläche auf die Nettomiete pro qm basierend auf dem GCV-optimalen $\lambda \approx 57$.*

3.5 Penalisierungsansätze II: Glättungssplines

3.5.1 Schätzansatz

Wir gehen in diesem Abschnitt zunächst davon aus, dass die beobachteten Kovariablenwerte alle verschieden und der Größe nach geordnet sind, d.h. $a < x_1 < x_2 < \ldots < x_n < b$. Wir suchen im Folgenden aus der Menge aller 2-mal stetig differenzierbaren Funktionen $f \in C_2[a, b]$ diejenige, welche das penalisierte KQ-Kriterium

$$SP(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx \tag{3.31}$$

bezüglich f minimiert. Im Vergleich zum P-Splineansatz aus dem vorangegangenen Abschnitt ergeben sich zwei wesentliche Unterschiede:

– Als Strafterm verwenden wir das Integral der quadrierten zweiten Ableitungen von f und keine quadrierten Differenzen.

– Wir haben im Gegensatz zu P-Splines nicht voraussgesetzt, dass f ein Spline ist.

Es wird sich herausstellen, dass es sich bei der gesuchten Funktion (obwohl nicht vorausgesetzt) um einen *natürlichen kubischen Spline* handelt. Natürliche kubische Splines sind spezielle Splines von Grad 3 mit zusätzlichen Bedingungen an den Rändern.

Definition 3.3 (natürliche kubische Splines)
Sei $a < x_1 < x_2 < \ldots < x_n < b$ eine Unterteilung des Intervalls $[a, b]$. Ein natürlicher kubischer Spline (NKS) s ist ein kubischer Spline bzgl. oben definierter Knotenmenge, für den zusätzlich die Randbedingungen

1. $s''(a) = 0$
2. $s''(b) = 0$

gelten, d.h. in den Intervallen $[a, x_1]$ und $[x_n, b]$ ist s linear.

Bevor wir uns mit der Minimierung des penalisierten KQ-Kriteriums (3.31) befassen, beschäftigen wir uns zunächst mit der Frage der Existenz interpolierender Splines:

Gegeben seine die Werte $z_1, \cdots, z_n$. Es stellt sich die Frage, ob es stets einen eindeutig bestimmten natürlichen kubischen Spline gibt, der die Werte z_i an der Stelle x_i interpoliert, d.h.

$$s(x_i) = z_i, \qquad i = 1, \ldots, n.$$

Dazu zunächst folgender

Satz 3.4 (Integralgleichung)
Sei $n \geq 2$ und sei s ein interpolierender NKS zu den Werten $z_1, \ldots, z_n$ an den Knoten $x_1, \ldots, x_n$. Sei weiterhin g eine beliebige Funktion in $C_2[a,b]$, die ebenfalls die Werte z_i interpoliert, d.h.

$$g(x_i) = z_i, \qquad i = 1, \ldots, n.$$

Dann gilt

$$\int (g''(x))^2 dx = \int_a^b (s''(x))^2 dx + \int_a^b (h''(x))^2 dx$$

wobei $h = g - s$.

Beweis:
Da sowohl g als auch s die Werte $z_1, \ldots, z_n$ interpolieren, gilt

$$h(x_i) = 0 \qquad i = 1, \ldots, n.$$

Weiter erhalten wir durch partielle Integration [2]

$$\begin{aligned}
\int_a^b s''(x)h''(x)dx &= \overbrace{[s''(x)h'(x)]_a^b}^{=0} - \int_a^b s'''(x)h'(x)dx \\
&= -\int_a^b s'''(x)h'(x)dx \\
&= -\sum_{j=1}^{n-1} s'''(x_j^+) \int_{x_j}^{x_{j+1}} h'(x)dx \\
&= -\sum_{j=1}^{n-1} s'''(x_j^+)(h(x_{j+1}) - h(x_j)) = 0.
\end{aligned}$$

Dabei haben wir von der zweiten zur dritten Zeile ausgenützt, dass $s''' = 0$ in (a, x_1) bzw. (x_n, b) und s''' eine Konstante ist in den restlichen Teilintervallen. Damit erhalten wir

$$\begin{aligned}
\int_a^b (g''(x))^2 dx &= \int_a^b (s''(x) + h''(x))^2 dx \\
&= \int_a^b (s''(x))^2 dx + 2 \underbrace{\int_a^b s''(x)h''(x)dx}_{=0} + \int_a^b (h''(x))^2 dx \\
&= \int_a^b (s''(x))^2 dx + \int_a^b (h''(x))^2 dx
\end{aligned}$$

□

[2] Allgemein gilt $\int_a^b f(x)g'(x)\,dx = [f(x)g(x)]_a^b - \int_a^b g(x)f'(x)\,dx$.

Aus der Integralgleichung können wir noch die folgende *Extremaleigenschaft* von interpolierenden NKS's folgern:

Satz 3.5 (Extremaleigenschaft)
Unter den Vorraussetzungen des Satzes 3.4 (Integraleigenschaft) gilt

$$\int (g''(x))^2 dx \geq \int (s''(x))^2 dx$$

Das Gleichheitszeichen gilt dabei genau dann, wenn g und s identisch sind.

Beweis:
Die Ungleichung folgt unmittelbar aus der Integralgleichung. Gleichheit gilt nur dann, wenn in der Integralgleichung $\int\limits_a^b (h''(x))^2 dx = 0$, d.h. h muss linear sein in $[a, b]$. Da aber $h(x_i) = 0$ gilt für $i = 1; \ldots, n$ und $n \geq 2$ muss $h = 0$ gelten. Das bedeutet aber $s = g$.

□

Unter Verwendung der Integralgleichung können wir folgende Existenz und Eindeutigkeitsaussage beweisen:

Satz 3.6
Für $n \geq 2$ existiert ein eindeutig bestimmter NKS s für den

$$s(x_i) = z_i, \qquad i = 1, \ldots, n,$$

gilt.

Beweis:
Unter Verwendung der truncated power series Basis für Splines kann die Interpolationsaufgabe als Lösung des folgenden linearen Gleichungssystems aufgefasst werden:

$$\begin{aligned}
s(x_i) &= \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3 + \sum_{j=2}^{n-1} \beta_j (x_i - x_j)_+^3 = z_i, \qquad i = 1, \ldots, n \\
s''(x_1) &= 0 \\
s''(x_n) &= 0
\end{aligned}$$

Es handelt sich um ein Gleichungssysstem mit $n+2$ Gleichungen und $n+2$ Unbekannten. Das Gleichungssystem ist genau dann eindeutig lösbar, wenn das zugehörige homogene System ($z_i = 0$) nur die triviale Lösung besitzt, d.h. $\alpha_0 = \alpha_1 = \ldots = \beta_{n-1} = 0$ bzw. $s = 0$. Angenommen es gäbe eine weitere Lösung g. Aufgrund der Integralgleichung muss $g'' = 0$ gelten, da das Integral der zweiten Ableitung von g nicht größer als das Ingetral von $s'' = 0$ sein kann. Damit folgt aus

$$g(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \sum_{j=2}^{n-1} \beta_j (x - x_j)_+^3$$

die Bedingung

$$g''(x) = 2\alpha_2 + 6\alpha_3 x + \sum_{j=2}^{n-1} 6\beta_j (x - x_j)_+ = 0$$

für alle $x \in [a, b]$. Bei $g''(x)$ handelt es sich offenbar um einen linearen Spline in truncated power series Darstellung. Da die Basisfunktionen $1, x, (x-x_2)_+, \ldots$ linear unabhängig sind, folgt

$$2\alpha_2 = 6\alpha_3 = 6\beta_2 = \ldots = 6\beta_{n-1} = 0$$

so dass $g(x)$ eine lineare Funktion sein muss. Aufgrund der Interpolationsbedingung $g(x_1) = \ldots = g(x_n) = 0$ folgt aber wieder $g = s = 0$.

□

Mit Hilfe der Interpolationseigenschaft und der Extremaleigenschaft können wir nun zeigen, dass das Minimum des penalisierten KQ-Kriteriums ein NKS ist.

Sei g eine beliebige 2 mal stetig differenzierbare Funktion. Sei s ein NKS mit $s(x_i) = g(x_i)$, $i = 1, \ldots, n$. Dann gilt

$$\sum_{i=1}^{n} (y_i - s(x_i))^2 = \sum_{i=1}^{n} (y_i - g(x_i))^2$$

und aufgrund der Extremaleigenschaft

$$\int (s''(x))^2 dx < \int (g''(x))^2 dx$$

Da g beliebig gewählt wurde, ist gezeigt, dass es sich bei der minimierenden Funktion um einen NKS handeln muss.

3.5.2 Penalisierte KQ-Schätzung

Bevor wir uns der Minimierung des penalisierten KQ-Kriteriums (3.31) widmen, befassen wir uns zunächst mit der Darstellung von natürlichen kubischen Splines durch B-Spline Basisfunktionen. Bei NKS's handelt es sich um einen Unterraum des Vektorraums der kubischen Splines mit Dimension n. Wir können einen NKS s wieder in B-Spline Basisdarstellung schreiben, d.h.

$$s(x) = \sum_{j=1}^{n} \beta_j B_j(x), \qquad a \leq x \leq b$$

Für $j = 3, \ldots, n-2$ können wir die üblichen B-Spline Basisfunktionen verwenden, wie bereits in (3.13) und (3.14) definiert. Für $j = 1, 2$ und $j = n-1, n$ müssen die Basisfunktionen wegen der zusätzlichen Randbedingungen natürlicher kubischer Splines angepasst werden.

Die Basisfunktionen lassen sich zum Beispiel mit Hilfe *dividierter Differenzen* berechnen. Diese sind für eine beliebige Funktion g definiert als

$$[x_j, \ldots, x_{j+l}]g = \frac{[x_{j+1}, x_{j+l}]g - [x_j, x_{j+l-1}]g}{x_{j+l} - x_j} \tag{3.32}$$

und

$$[x_j]g = g(x_j). \tag{3.33}$$

Damit erhalten wir die Darstellung:

$$B_j = \begin{cases} [x_1, x_2, x_3](t-x)_+^3 & j = 1 \\ [x_2, x_3, x_4](t-x)_+^3 & j = 2 \\ (x_{j+2} - x_{j-2})[x_{j-2}, \ldots, x_{j+2}](x-t)_+^3 & j = 3, \ldots, n-2 \\ [x_{n-3}, x_{n-2}, x_{n-1}](x-t)_+^3 & j = n-1 \\ [x_{n-2}, x_{n-1}, x_n](x-t)_+^3 & j = n \end{cases}$$

Beispiel 3.16
Wir betrachten die Knotenmenge $a = 0 < 0.1 < 0.2 < 0.3 < 0.4 < 0.5 < 0.6 < 0.7 < 0.8 < 0.9 < 1 = b$. Abbildung 3.31 zeigt die neun Basisfunktionen $B_1 - B_9$ für einen natürlichen kubischen Spline zu dieser Knotenmenge.

△

Wir kommen jetzt zur Lösung des Optimierungsproblems (3.31). Wir schreiben zunächst das penalisierte KQ-Kriterium in Matrixnotation. Da die Lösung ein NKS sein muss, den wir in B-spline-Darstellung schreiben können, folgt

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n}\beta_j B_j(x_i)\right)^2 = (y - \mathbf{X}\beta)'(y - \mathbf{X}\beta)$$

wobei $\mathbf{X}$ eine $(n \times n)$ Designmatrix ist, mit den Elementen $x_{ij} = B_j(x_i)$. Weiter folgt

$$\begin{aligned} \int (f''(x))^2 dx &= \int \left(\sum_{j=1}^{n}\beta_j B_j''(x)\right)^2 dx \\ &= \int \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j B_i''(x)B_j''(x)\right) dx \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j \underbrace{\int B_i''(x)B_j''(x)dx}_{=K_{ij}} \\ &= \beta'\mathbf{K}\beta \end{aligned}$$

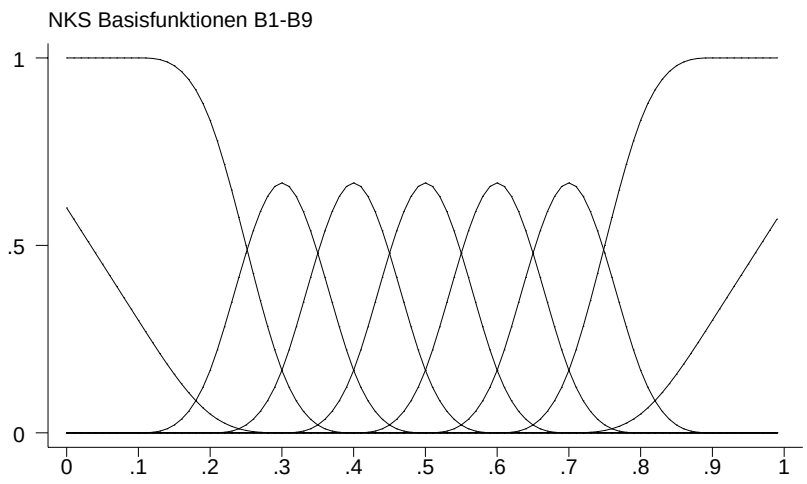

Abbildung 3.31. *Basisfunktionen für einen natürlichen kubischen Spline zu den Knoten* $a = 0 < 0.1 < 0.2 < 0.3 < 0.4 < 0.5 < 0.6 < 0.7 < 0.8 < 0.9 < 1 = b$.

wobei K eine $(n \times n)$ Strafmatrix ist mit den Elementen

$$k_{ij} = \int B_i''(x) B_j''(x) dx.$$

Wir erhalten also das penalisierte KQ-Kriterium (3.31) in Matrixnotation

$$SP(\beta) = (y - \mathbf{X}\beta)'(y - \mathbf{X}\beta) + \lambda \beta' \mathbf{K} \beta.$$

Die Minimierung bezüglich β kann völlig analog zu den P-Splines erfolgen. Wir erhalten

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{K})^{-1} \mathbf{X}' y$$

und

$$\hat{f} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))' = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{K})^{-1} \mathbf{X}' y = \mathbf{S} y,$$

mit der Smoothermatrix

$$\mathbf{S} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{K})^{-1} \mathbf{X}'.$$

Bei Glättungssplines handelt es sich also wieder um lineare Smoother.

Ähnlich wie bei P-Splines besitzt die Matrix $\mathbf{X}'\mathbf{X} + \lambda \mathbf{K}$ wieder Bandstruktur mit Bandweite 3. Diese Bandstruktur kann zur effizienten Berechnung von $\hat{\beta}$ ausgenutzt werden.

3.5.3 Einfluss und Wahl des Glättungsparameters

Die geschätzten Funktionen $\hat{f}$ sind wieder stark von der speziellen Wahl des Glättungsparameters λ abhängig. Für $\lambda \to 0$ verschwindet der Strafterm in (3.31). Da an jeder

Beobachtung ein Knoten definiert ist, werden genauso viele Parameter geschätzt wie Beobachtungen. Damit werden für $\lambda \to 0$ die Daten interpoliert, d.h. wir erhalten den interpolierenden natürlichen kubischen Spline. Für $\lambda \to \infty$ erhält der Strafterm großes Gewicht und muss daher gegen Null streben, d.h. die zweiten Ableitungen müssen Null werden. Damit ergibt sich im Grenzfall $\lambda \to \infty$ als Schätzung $\hat{f}$ ein Polynom ersten Grades, d.h. eine Regressionsgerade. Zusammenfassend erhalten wir:

$\lambda \to 0$	Interpolation der Daten
λ „klein“	sehr rauhe Schätzung
λ „groß“	sehr glatte Schätzung
$\lambda \to \infty$	Regressionsgerade

Die Wahl des Glättungsparameters kann wieder auf ähnliche Art und Weise erfolgen wie bei P-Splines. Wie bei P-splines minimieren wir die Kreuzvalidierungsfunktion

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^{-i}(x_i))^2.$$

Analog zu den P-splines zeigt man, dass

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}(\lambda)}\right)^2$$

gilt. Alternativ kann die generalisierte Kreuzvalidierungsfunktion

$$GCV(\lambda) = n\frac{\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2}{(n - df_f)^2}.$$

minimiert werden. Diese ensteht aus $CV(\lambda)$, indem wieder der Faktor $1 - S_{ii}(\lambda)$ durch den Mittelwert

$$\frac{1}{n}\sum_{i=1}^{n}(1 - S_{ii}(\lambda)) = 1 - \frac{1}{n}sp(\mathbf{S}(\lambda)) = 1 - \frac{1}{n}df_f$$

ersetzt wird.

Beispiel 3.17
Wir betrachten wieder die Motorcycledaten. Die Abbildungen 3.32 und 3.33 zeigen für verschiedene äquivalente Freiheitsgrade geschätzte Glättungssplines.

△

3.5.4 Gruppierte Daten

Im Vergleich zu P-Splines sind bei Glättungssplines viel mehr Basisfunktionen (n= Anzahl der Beobachtungen) notwendig. Bei sehr vielen *verschiedenen* Beobachtungen kann dies

zu Rechenzeitproblemen führen. In der Praxis werden die Daten daher oft gruppiert oder man geht zu folgendem leicht verändertem penalisierten KQ-Kriterium über:

$$SP(\beta) = \sum_{i=1}^{n} (y_i - s(x_i))^2 + \lambda \int (s''(x))^2 dx$$

wobei

$$s(x) = \sum_{j=1}^{p} \beta_j B_j(x) \qquad p << n$$

ein NKS in B-spline Darstellung ist. Dieser Ansatz ist fast identisch zu den P-splines aus Abschnitt 3.4. Im Unterschied dazu werden hier aber *natürliche* Splines und eine andere Penalisierung verwendet.

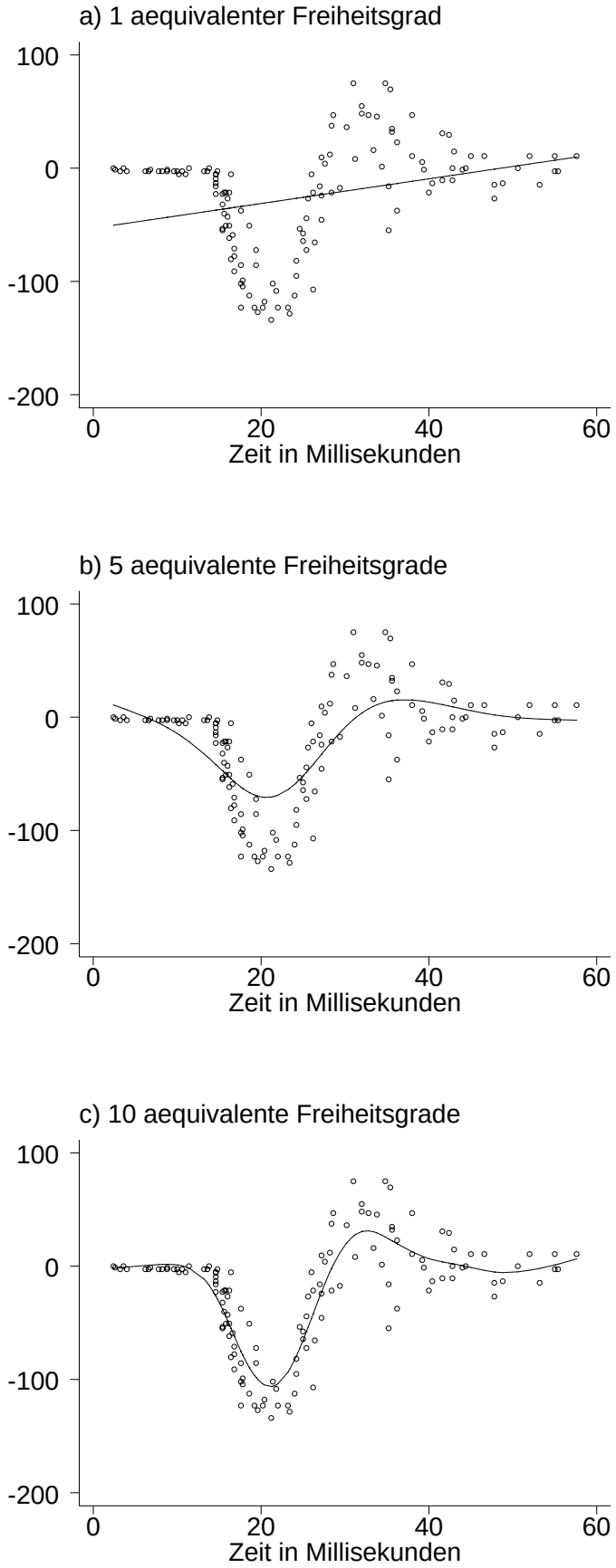

Abbildung 3.32. *Motorcycledaten: Glättungssplines mit verschiedenen äquivalenten Freiheitsgraden.*

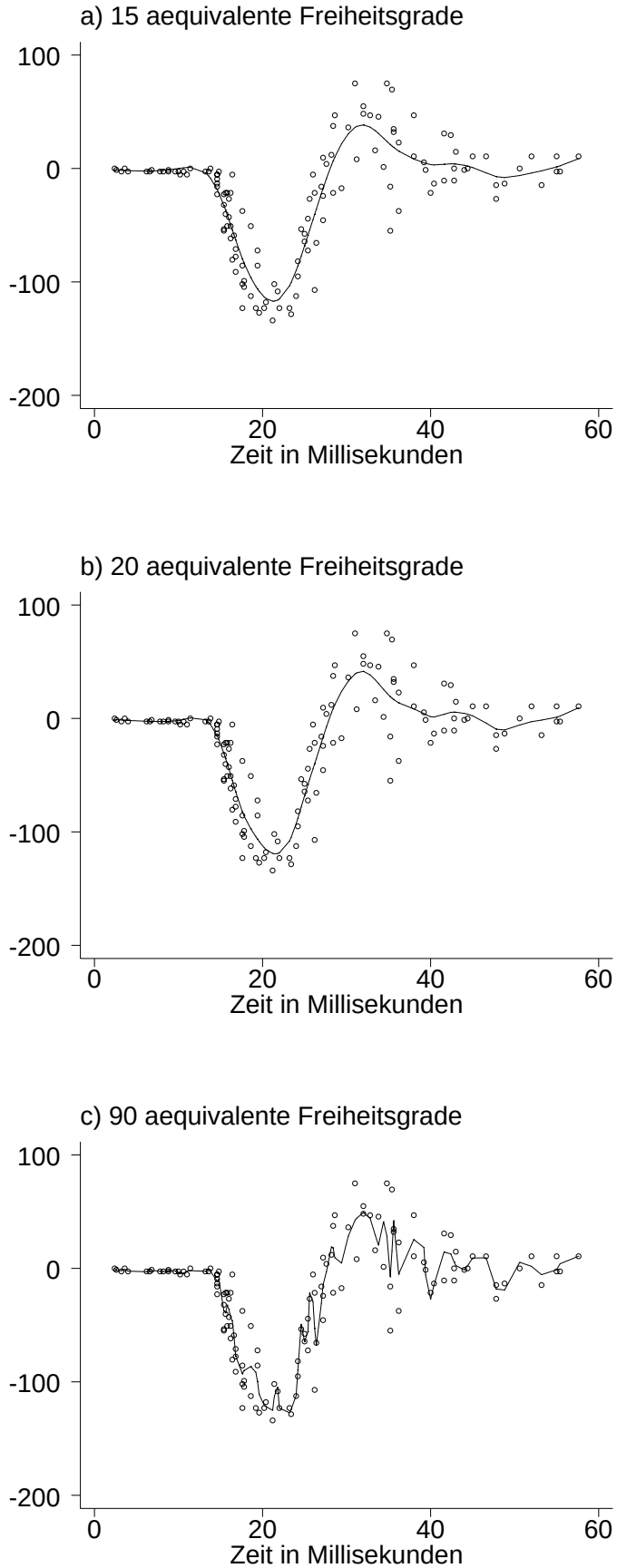

Abbildung 3.33. *Motorcycledaten: Glättungssplines mit verschiedenen äquivalenten Freiheitsgraden.*

3.6 Lokale Scatterplotsmoother

3.6.1 Nächste Nachbarn Schätzer

Nächste Nachbarn Schätzern liegt die einfache Idee zugrunde, als Schätzer $\hat{f}(x_i)$ an der Stelle x_i, $i = 1, \ldots, n$, den "Mittelwert" der Responsebeobachtungen in einer Nachbarschaft von x_i zu verwenden. Formal schreiben wir

$$\hat{f}(x_i) = Ave_{j \epsilon N(x_i)}(y_j)$$

wobei Ave ein Mittelwertoperator ist und $N(x_i)$ eine Nachbarschaft von x_i. Je nachdem wie Ave und $N(x_i)$ definiert sind, erhalten wir unterschiedliche Glätter. Im folgenden behandeln wir einige Beispiele für $N(x_i)$:

– *Symmetrische Nachbarschaft*
  Bei einer symmetrischen Nachbarschaft verwenden wir links *und* rechts von x_i gleich viele Beobachtungen. Sei $\omega \in (0, 1)$ und sei der ganzzahlige Anteil $[\omega n]$ von ωn ungerade. Dann definieren wir

$$N(x_i) = (max\left\{1, i - \frac{[\omega n] - 1}{2}\right\}, \ldots, i-1, i, i+1, \ldots, min\left\{i + \frac{[\omega n] - 1}{2}, n\right\}).$$

  Die Nachbarschaft $N(x_i)$ enthält also die *Indizes* der geordneten Daten $x_1, \ldots, x_n$, die bei der Mittelwertbildung berücksichtigt werden. Von den Rändern abgesehen, werden also jeweils $\frac{[wn]-1}{2}$ Beobachtungen links und rechts von x_i benutzt. Die Zahl ω bezeichnet somit den Prozentanteil der Daten, die zur Berechnung des Funktionswertes $\hat{f}(x_i)$ berücksichtigt werden. Problematisch an diesem Vorgehen ist natürlich, dass die Nachbarschaft an den Randpunkten unsymmetrisch und vor allem auch kleiner wird. Als Glättungsparameter fungiert ω. Je näher ω bei 0 liegt, desto rauher wird die Schätzung, je näher ω bei 1 liegt, desto glatter.

– *k nächste Nachbarn (unsymmetrische Nachbarschaft)*
  Zur Mittelwertbildung an der Stelle x_i werden hier die k nächsten Nachbarbeobachtungen herangezogen, d.h.

$$N(x_i) = \{j \,|\, x_j \text{ist einer der k nächsten Nachbarn von } x_i\}$$

  Als Glättungsparameter fungiert hier k. Je kleiner k desto rauher, je größer k desto glatter die Schätzung.

Unter anderen sind folgende Mittelwertoperatoren denkbar:

– *Running mean Schätzer*
  Hier wird das arithmetische Mittel der Beobachtungen in $N(x_i)$ zur Bestimmung von $\hat{f}(x_i)$ benutzt.

– *Running median Schätzer*
  Hier wird der Median der Beobachtungen in $N(x_i)$ zur Bestimmung von $\hat{f}(x_i)$ benutzt. Beim Running median Schätzer handelt es sich um einen *nichtlinearen* Glätter.

– *Running line Schätzer*
  Beim Running line Schätzer definieren wir

$$\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

  wobei $\hat{\beta}_0, \hat{\beta}_1$ die KQ-Schätzer basierend auf den Beobachtungen in $N(x_i)$ sind.

Beispiel 3.18
Wir betrachten wieder die Motorcycledaten. Die Abbildungen 3.34 und 3.35 zeigen für 6 verschiedene Bandweiten running mean Schätzer. Die Abbildungen 3.36 und 3.37 zeigen verschiedene running line Schätzer. Für kleine Bandweiten, d.h. für kleine Nachbarschaften werden die Schätzer datentreuer, jedoch auch rauher. Für große Bandweiten nähert sich der running mean Schätzer einer konstanten Funktion (dem arithmetischen Mittel der y_i's) an. Der running line Schätzer nähert sich für Bandweiten nahe 1 der Regressionsgeraden zwischen x und y an. Tendenziell ist der running mean Schätzer rauher als der running line Schätzer. Im vorliegenden Fall ist eine relativ geringe Bandweite nötig, um eine befriedigende Schätzung zu erhalten.

△

Beispiel 3.19
In diesem Beispiel betrachten wir die Mietspiegeldaten. Die Abbildungen 3.38 und 3.39 zeigen für 6 verschiedene Bandweiten running mean Schätzer für den Zusammenhang zwischen Nettomiete pro Quadratmeter und der Wohnfläche. Die Abbildungen 3.40 und 3.41 zeigen verschiedene running line Schätzer. Offensichtlich gelingt es mit dem running mean Schätzer nicht eine befriedigende Anpassung am linken Rand der Daten zu gewährleisten. Mit dem running line Schätzer erhält man eine bessere Anpassung in diesem Bereich. Insgesamt ist in diesem Beispiel die Abhängigkeit von der gewählten Bandweite geringer als bei den Motorcycledaten.

△

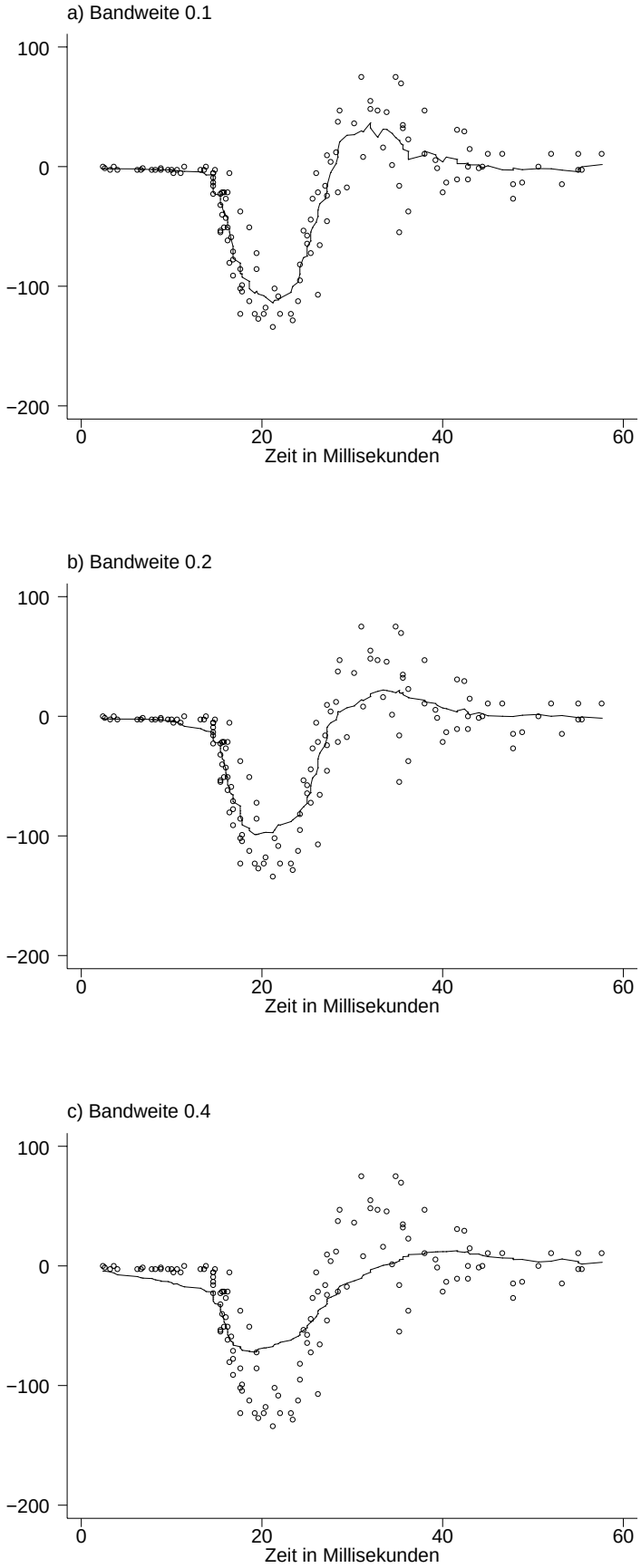

Abbildung 3.34. *Motorcycledaten: Running mean Schätzer für verschiedene Bandweiten.*

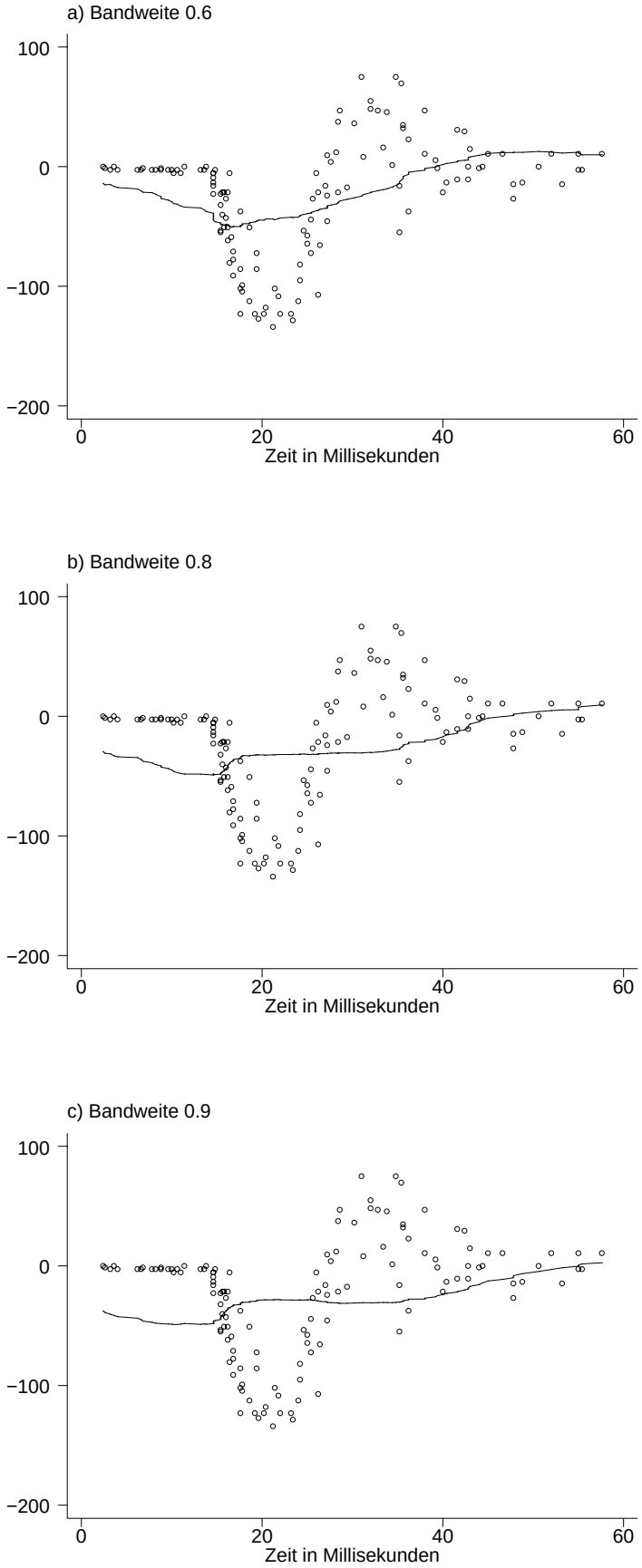

Abbildung 3.35. *Motorcycledaten: Running mean Schätzer für verschiedene Bandweiten.*

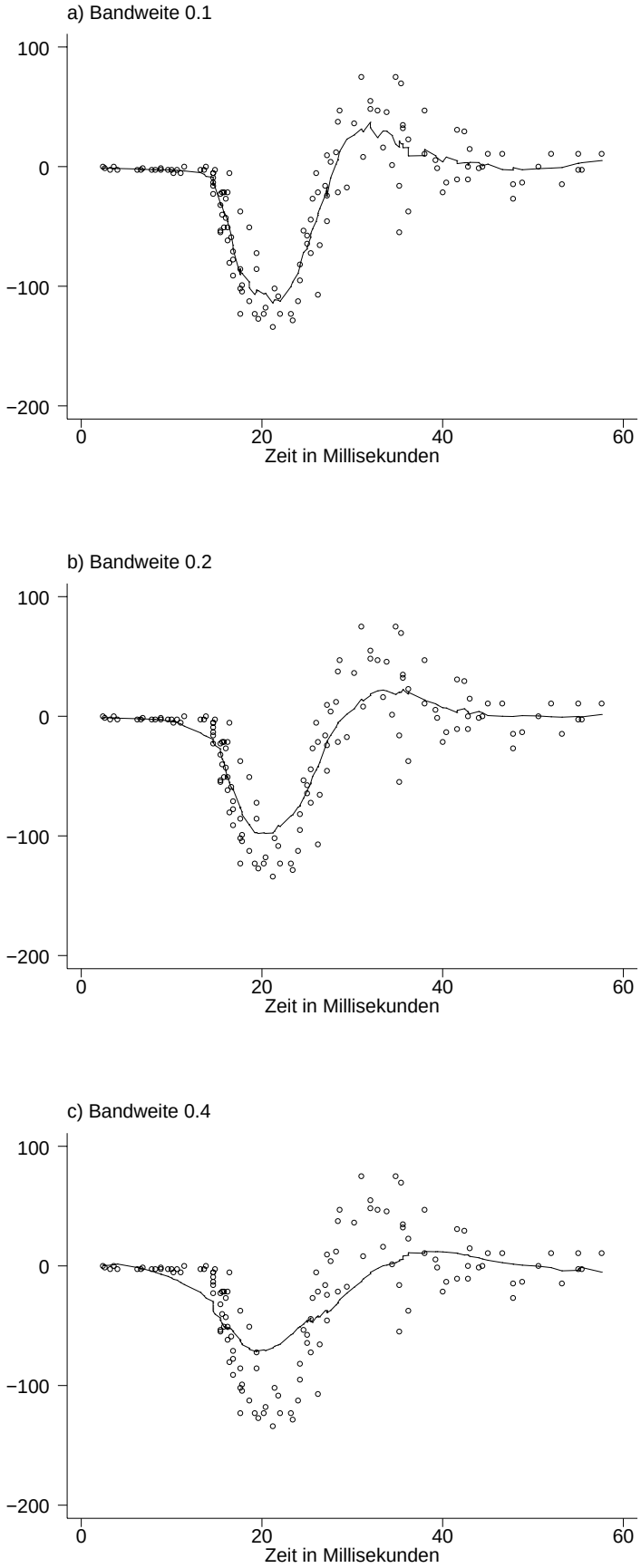

Abbildung 3.36. *Motorcycledaten: Running line Schätzer für verschiedene Bandweiten.*

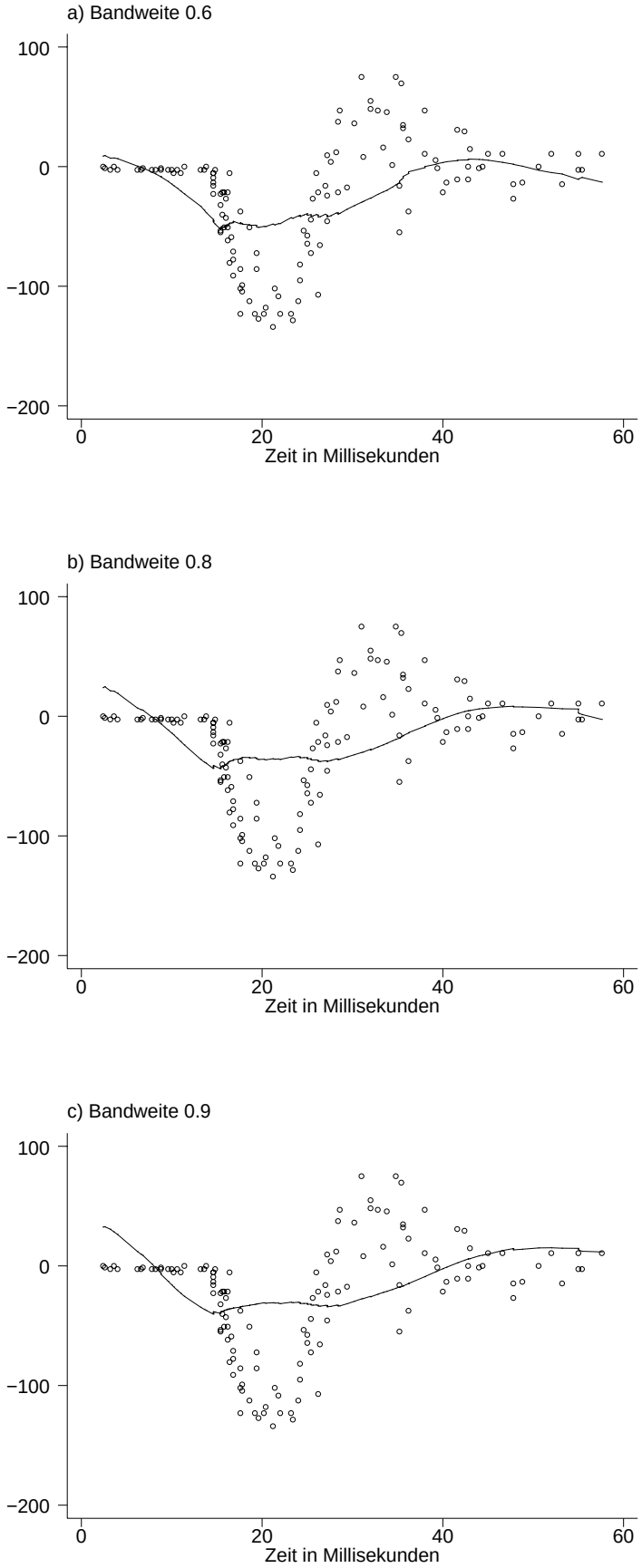

Abbildung 3.37. *Motorcycledaten: Running line Schätzer für verschiedene Bandweiten.*

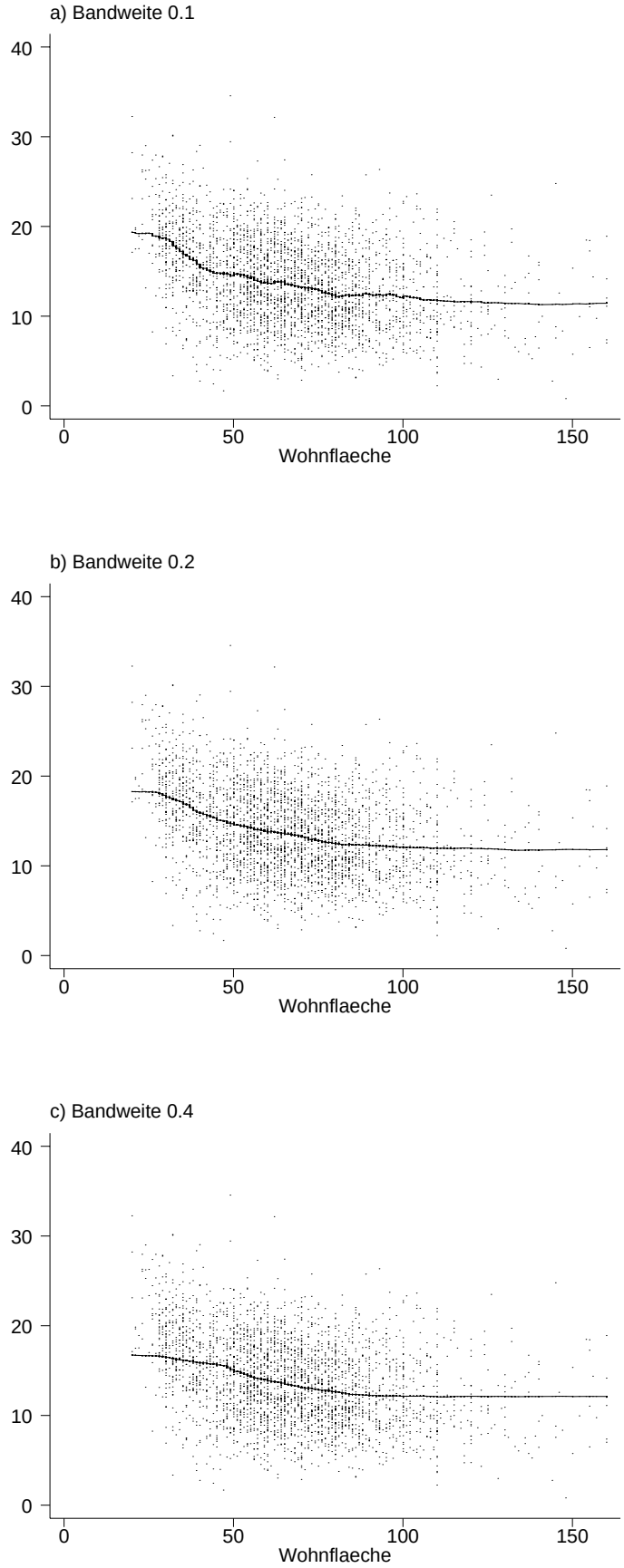

Abbildung 3.38. *Mietspiegeldaten: Running mean Schätzer für verschiedene Bandweiten für die Regression zwischen Nettomiete pro Quadratmeter und Wohnfläche.*

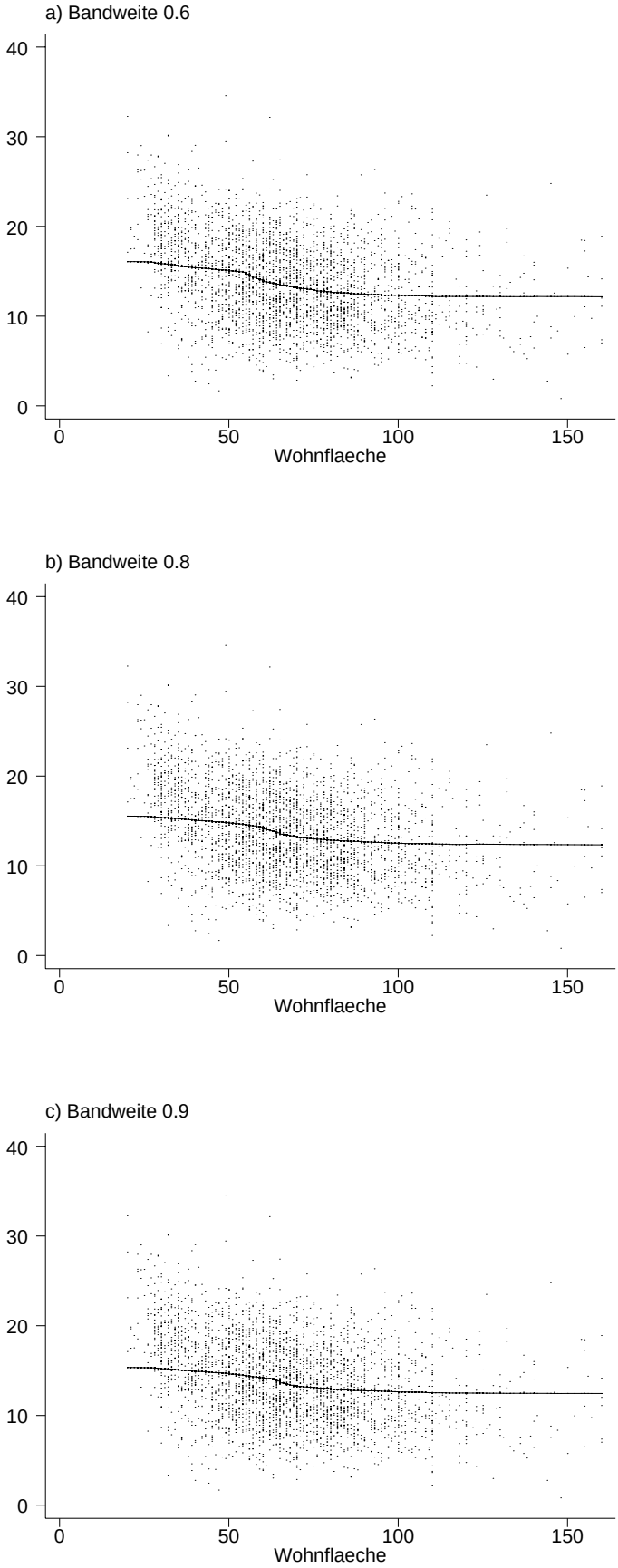

Abbildung 3.39. *Mietspiegeldaten: Running mean Schätzer für verschiedene Bandweiten für die Regression zwischen Nettomiete pro Quadratmeter und Wohnfläche.*

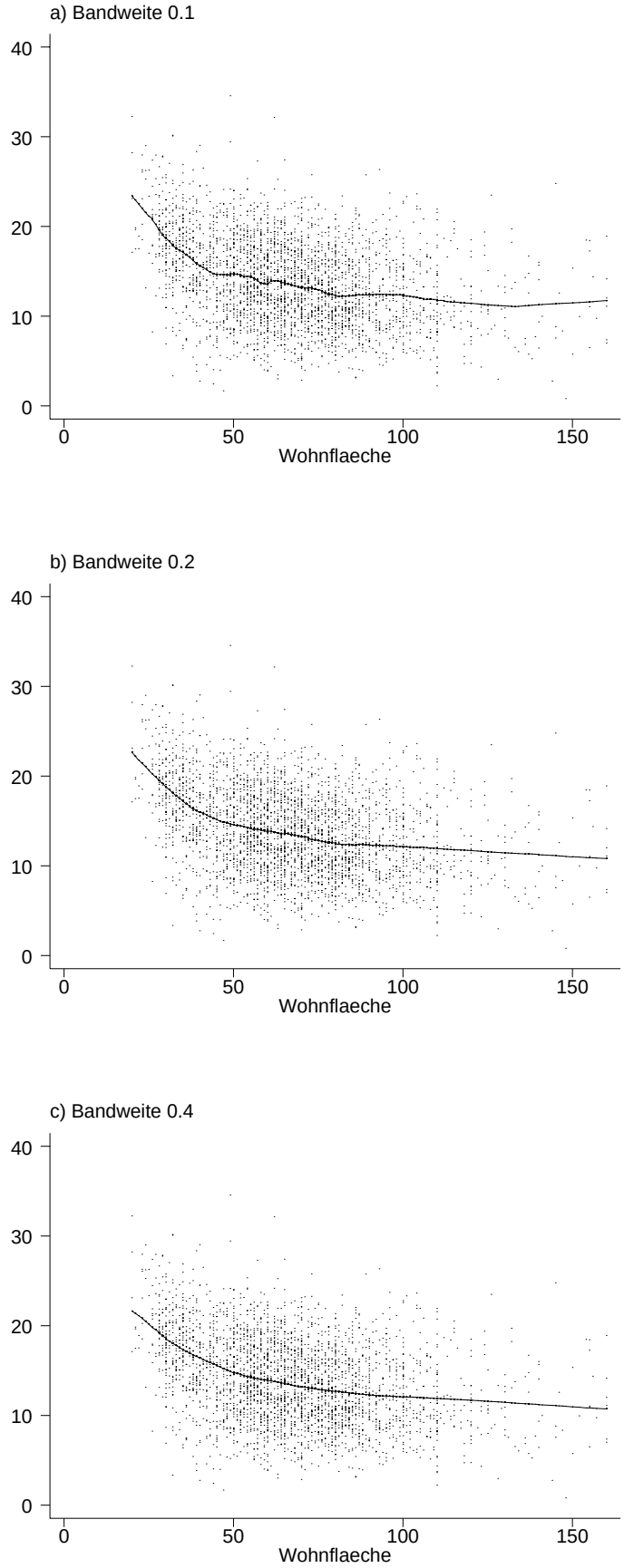

Abbildung 3.40. *Mietspiegeldaten: Running line Schätzer für verschiedene Bandweiten für die Regression zwischen Nettomiete pro Quadratmeter und Wohnfläche.*

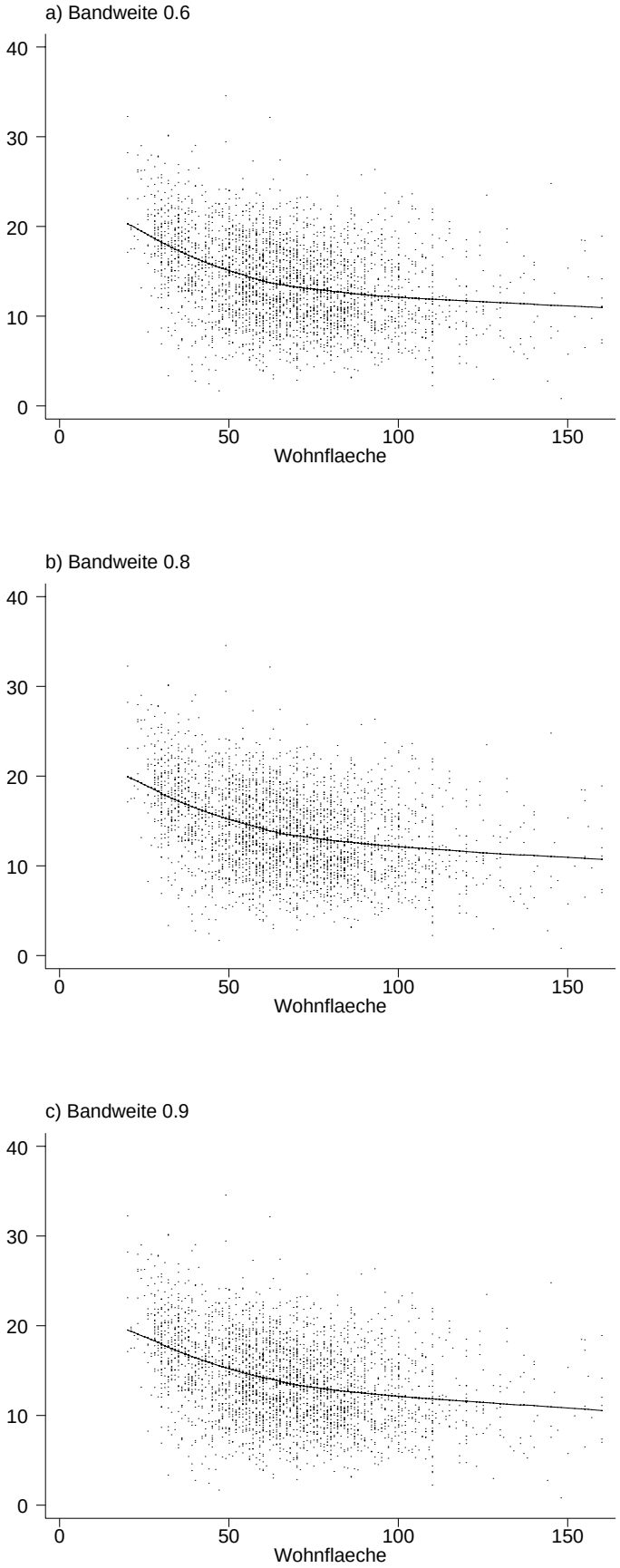

Abbildung 3.41. *Mietspiegeldaten: Running line Schätzer für verschiedene Bandweiten für die Regression zwischen Nettomiete pro Quadratmeter und Wohnfläche.*

3.6.2 Lokal polynomiale Regression

Obwohl $f(x)$ in der Regel kein Polynom ist, kann man häufig annehmen, dass f zumindest in einer *Umgebung* von x durch ein Polynom p-ten Grades approximiert werden kann. Dies lässt sich durch eine Taylorreihenapproximation rechtfertigen. Entwicklung von $f(x_i)$ mit Entwicklungspunkt x liefert

$$f(x_i) = \underbrace{f(x)}_{\beta_0} + \underbrace{f'(x)}_{\beta_1}(x_i - x) + \underbrace{\frac{f''(x)}{2}}_{\beta_2}(x_i - x)^2 + \ldots \approx \beta_0 + \sum_{j=1}^{p} \beta_j (x_i - x)^j$$

Für festes x können wir also die gewichtete Residuenquadratsumme

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j (x_i - x)^j \right)^2 w_\lambda(x, x_i)$$

bzgl. $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ minimieren. Die Gewichte $w_\lambda(x, x_i)$ sind dabei in der Regel umso kleiner, je größer der Abstand zwischen x und x_i ist. Zusätzlich hängen die Gewichte von einem Glättungsparameter λ ab, der steuert wie "schnell" die Gewichte kleiner werden. Für geschätzte Werte $\hat{\beta}$ erhalten wir dann

$$\hat{f}(x) = \hat{\beta}_0.$$

Aus der Taylorreihenabschätzung lassen sich zusätzlich Schätzungen für die Ableitungen von f bestimmen:

$$\hat{f}^{(j)} = j!\hat{\beta}_j$$

Die Gewichte $w_\lambda(x, x_i)$ hängen von x ab und damit auch die Regressionskoeffizienten $\hat{\beta}$. Man muss also für jeden x-Wert, für den f geschätzt werden soll, eine separate KQ-Schätzung $\hat{\beta}$ bestimmen. Als Gewichtsfunktionen wählt man üblicherweise Kernfunktionen K, die bereits aus der nichtparametrischen Dichteschätzung bekannt sind, d.h.

$$w_\lambda(x, x_i) = K\left(\frac{x_i - x}{\lambda}\right).$$

Zur praktischen Durchführung der Schätzungen schreiben wir zunächst das KQ - Kriterium in Matrixnotation

$$\sum \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j (x_i - x)^j \right)^2 K\left(\frac{x_i - x}{\lambda}\right) = (y - \mathbf{X}\beta)'\mathbf{W}(y - \mathbf{X}\beta)$$

mit der $n \times (p+1)$ Designmatrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{pmatrix}$$

und der Gewichtsmatrix

$$\mathbf{W} = diag\left(K\left(\frac{x_1 - x}{\lambda}\right), \ldots, K\left(\frac{x_n - x}{\lambda}\right)\right).$$

Damit erhält man

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}y$$

und mit $e_1 = (1, 0, \ldots, 0)'$

$$\hat{f}(x) = \hat{\beta}_0 = e_1'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}y$$

d.h. $\hat{f}(x)$ ist ein linearer Glätter.

Einen Spezialfall stellen lokal konstante Schätzer dar. Bei lokal konstanten Schätzern besteht ein interessanter Zusammenhang mit den in Kapitel 2.3 behandelten Kerndichteschätzern.

Für einen lokalen konstanten Fit ist für festes x

$$\sum_{i=1}^{n}(y_i - \beta_0)^2 K\left(\frac{x - x_i}{\lambda}\right)$$

bezüglich β_0 zu minimieren. Wir erhalten

$$\hat{f}(x) = \sum_{i=1}^{n} s(x, x_i) y_i$$

mit

$$s(x, x_i) = \frac{\frac{1}{n\lambda} K\left(\frac{x - x_i}{\lambda}\right)}{\frac{1}{n\lambda}\sum\limits_{j=1}^{n} K\left(\frac{x - x_j}{\lambda}\right)}.$$

Dieser Schätzer lässt sich auch aus Kerndichteschätzungen ableiten:

Es liegt nahe als Schätzung $\hat{f}(x)$ den bedingten Erwartungswert $E(Y|X = x)$ heranzuziehen, d.h.

$$\hat{f}(x) = E(Y \mid X = x) = \frac{\int y\, d(x, y) dy}{d(x)} \tag{3.34}$$

wobei $d(x, y)$ die Dichte von x, y ist und $d(x)$ die Dichte von x. Für die unbekannten Dichten $d(x, y)$ und $d(x)$ können wir Kerndichteschätzer einsetzen. Für die zweidimensionale Dichte $d(x, y)$ verwenden wir einen Kerndichteschätzer basierend auf Produktkernen (vergleiche Kapitel 2.5)

$$\hat{d}(x, y) = \frac{1}{n\lambda^2}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) K\left(\frac{y - y_i}{\lambda}\right).$$

Für die eindimensionale Dichte $d(x)$ können übliche Kerndichteschätzer verwenden werden, d.h.

$$\hat{d}(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right).$$

Damit erhalten wir zunächst für den Zähler in (3.34)

$$\begin{aligned}
\int y\,\hat{d}(x,y)dy &= \frac{1}{n\lambda^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) \int y\,K\left(\frac{y - y_i}{\lambda}\right) dy \\
&= \frac{1}{n\lambda^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) \int \lambda(s\lambda + y_i)K(s)ds \\
&= \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) \left[\lambda \underbrace{\int sK(s)ds}_{=0} + y_i \underbrace{\int K(s)ds}_{=1}\right] \\
&= \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) y_i
\end{aligned}$$

Dabei wurde in der zweiten Zeile die Substitution $s = \frac{y-y_i}{\lambda}$ verwendet. Es folgt

$$\hat{f}(x) = \frac{\int y\,\hat{d}(x,y)dy}{\hat{d}(x)} = \frac{\frac{1}{n\lambda} \sum\limits_{i=1}^{n} K\left(\frac{x-x_i}{\lambda}\right) y_i}{\frac{1}{n\lambda} \sum\limits_{i=1}^{n} K\left(\frac{x-x_i}{\lambda}\right)}$$

Dieser Schätzer heißt Nadaraya-Watson Schätzer und stimmt offensichtlich mit dem lokal konstanten Fit überein.

3.6.3 Lokal gewichteter running line Smoother (Loess)

Loess (Cleveland (1979)) ist eine Kombination von nächste Nachbarn Schätzern und lokal gewichteter Regression. Eine Schätzung von f an der Stelle x erhält man wie folgt:

i) Bestimme die Menge $N(x)$ der k nächsten Nachbarn von x.

ii) Bestimme den betragsmäßig größten Abstand $\Delta(x)$ zwischen x und den k nächsten Nachbarn.

iii) Definiere Gewichte

$$w_i = K\left(\frac{|x - x_i|}{\Delta(x)}\right)$$

mit

$$K(u) = \begin{cases} (1 - u^3)^3 & 0 \le u < 1 \\ 0 & \text{sonst.} \end{cases}$$

iv) Bestimme $\hat{f}(x)$ durch gewichtete lineare (quadratische, usw.) Regression.

Loess kann robust gemacht werden bezüglich Ausreißern. Dazu werden in einem zweiten Glättungsschritt Beobachtungen mit großen Residuen heruntergewichtet. Seien $r_i = y_i - \hat{y}_i$

$i = 1, \ldots, n$. Definiere Gewichte δ_i, die umso kleiner sind je größer $|r_i|$. Definiere die neuen Gewichte

$$w_i = \delta_i K\left(\frac{|x - x_i|}{\Delta(x)}\right)$$

und schätze erneut.

Beispiel 3.20
Wir betrachten wieder die Motorcycledaten. Die Abbildungen 3.42 und 3.43 zeigen für 6 verschiedene Bandweiten loess Schätzer. Wie in Beispiel 3.18 ist eine relativ kleine Bandweite nötig, um eine befriedigende Anpassung and die Daten zu gewährleisten.

△

Beispiel 3.21
Für die Mietspiegeldaten zeigen die Abbildungen 3.44 und 3.45 für 6 verschiedene Bandweiten loess Schätzer. In diesem Fall sind die Schätzungen relativ robust bezüglich der Wahl der Bandweiten.

△

3.6.4 Bias- Varianz Trade off am Beispiel lokaler polynomialer Regression

Die asymptotischen Eigenschaften lokal polynomialer Glätter sind relativ gut erforscht, während für Penalisierungsansätze weit weniger Resultate existieren.

Für $\lambda \to 0$ und $n\lambda \to \infty$ erhalten wir für lokal polynomiale Glätter

$$Bias\hat{f}(x) = E(\hat{f}(x) - f(x)) = \frac{\lambda^{p+1} f^{(p+1)}(x)}{(p+1)!}\mu_{p+1}(K) + o_{pr}(\lambda^{p+1}) \tag{3.35}$$

falls p ungerade ist und

$$Bias\hat{f}(x) = \left[\frac{\lambda^{p+2} f^{(p+1)}(x) d'(x)}{(p+1)! d(x)} + \frac{\lambda^{p+2} f^{(p+2)}(x)}{(p+2)!}\right]\mu_{p+2}(K) + o_{pr}(\lambda^{p+2}) \tag{3.36}$$

falls p gerade ist. Dabei bezeichnet $d(x)$ die Dichte von X, $f^{(q)}$ die q-te Ableitung von f und $\mu_q(K) = \int u^q K(u)\, du$. Die Terme o_{pr} sind die stochastischen Analoga zu den Termen klein o (vergleiche Kapitel 2.4.3). Diese sind definiert als

$$o_{pr}(\lambda^q) \Longleftrightarrow \frac{f(\lambda)}{\lambda^q} \xrightarrow{pr} 0$$

für eine Funktion f von λ.

Die asymptotischen Biasterme (3.35) und (3.36) lassen folgende Schlussfolgerungen zu:

– Der asymptotische $Bias$ wird kleiner, wenn der Glättungsparameter λ kleiner wird.

– Für $p = 0$ (lokal konstant) und $p = 1$ (lokal linear) besitzt der Bias die gleiche Ordnung, ebenso $p = 2, p = 3$ usw.

– Für p ungerade hängt der $Bias$ nicht vom Design, d.h. von der Dichte $d(x)$ ab. In diesem Sinn sind die Schätzungen unabhängig vom Design.

– Für p gerade hängt der $Bias$ von einem zusätzlichen Term ab, der insbesondere die $p+2$-te Ableitung von f enthält. Für $p = 0$ (lokal konstant) ist also der Bias umso stärker, je größer die erste und zweite Ableitung von f ist. Im Falle der ersten Ableitung bedeutet dies einen höheren Bias wenn f steil ist. Da die zweite Ableitung von f ein Maß für die Krümmung ist, erhalten wir einen positiven Bias (Überschätzung) bei lokalen Minima von f und einen negativen Bias (Unterschätzung) bei lokalen Maxima. Bei lokal linearen Schätzern ($p = 1$) hängt der asymptotische Bias zwar von der zweiten Ableitung, jedoch nicht von der ersten Ableitung ab.

Für die asymptotische Varianz erhalten wir

$$Var(\hat{f}_\lambda(x)) = \frac{\sigma^2(x)}{n\lambda d(x)} \int K^2(u)du + o_{pr}\left(\frac{1}{\lambda n}\right).$$

Dies lässt folgende Interpretation zu:

– Die Varianz wird umso kleiner je größer λ wird. Da der Bias umso kleiner wird, je kleiner der Glättungsparameter ist, ergibt sich ähnlich wie bei Kerndichteschätzern ein Trade off zwischen Minimierung des Bias und der Varianz.

– Die Varianz hängt von der Dichte $d(x)$ ab, je kleiner $d(x)$ desto größer ist die Varianz und umgekehrt.

3.7 Ergänzungen zu Scatterplotsmoothern

Äquivalente Kerne von Smoothern

Bis auf den running median Smoother sind alle Smoother, die wir behandelt haben, lineare Smoother, d.h.

$$\hat{f} = \begin{pmatrix} \hat{f}(x_1) \\ \vdots \\ \hat{f}(x_n) \end{pmatrix} = \mathbf{S}y$$

Wir können also verschiedene Glätter vergleichen, indem wir die Gewichte (in $\mathbf{S}$) vergleichen, mit denen die Responsebeobachtungen y_i an einem x-Wert x_j gewichtet

werden. Diese Gewichte heißen *äquivalenter Kern* an x_j.

Approximative F-Tests

Wir wollen im Folgenden einen (heuristischen) Test entwickeln zum Vergleich zweier Smoother $\hat{f}_1 = \mathbf{S}_1 y$ und $\hat{f}_2 = \mathbf{S}_2 y$. Häufig handelt es sich bei $\hat{f}_1$ um eine lineare Funktion während $\hat{f}_2$ nichtlinear ist. Die zugrundeliegende Frage in diesem Fall ist, ob eine nichtlineare Modellierung wirklich notwendig ist.

Im multiplen Regressionsmodell werden lineare Hypothesen der Form

$$\begin{array}{cccccc} H_0: & R & \beta & = & r \\ & (I \times p) & (p \times 1) & & (I \times 1) \end{array}$$

mit Hilfe der Teststatistik

$$F = \frac{\frac{1}{I}(SSE_{H_0} - SSE)}{\frac{1}{n-p}(SSE)} \sim F_{I,n-p}$$

getestet. Hier bezeichnet SSE_{H_0} die Residuenquadratsumme im Modell unter H_0 und SSE die Residuenquadratsumme im unrestringierten Modell. $F_{I,n-p}$ ist die F- Verteilung mit I und $n-p$ Freiheitsgraden. Analog dazu können wir einen heuristischen F-Test zum Vergleich zweier Glätter basierend auf der Teststatistik

$$F = \frac{\frac{1}{df_2 - df_1}(SSE_1 - SSE_2)}{\frac{1}{n-df_2} SSE_2} \overset{appr}{\sim} F_{df_2 - df_1, n-df_2}$$

verwenden. Hier ist SSE_j, $j = 1, 2$, die Residuenquadratsumme von $\hat{f}_j$ und df_j die äquivalenten Freiheitsgrade. Eine Voraussetzung für die Anwendbarkeit ist, dass $\hat{f}_2$ mehr äquivalente Freiheitsgrade, d.h. mehr Nichtlinearität, besitzt als $\hat{f}_1$.

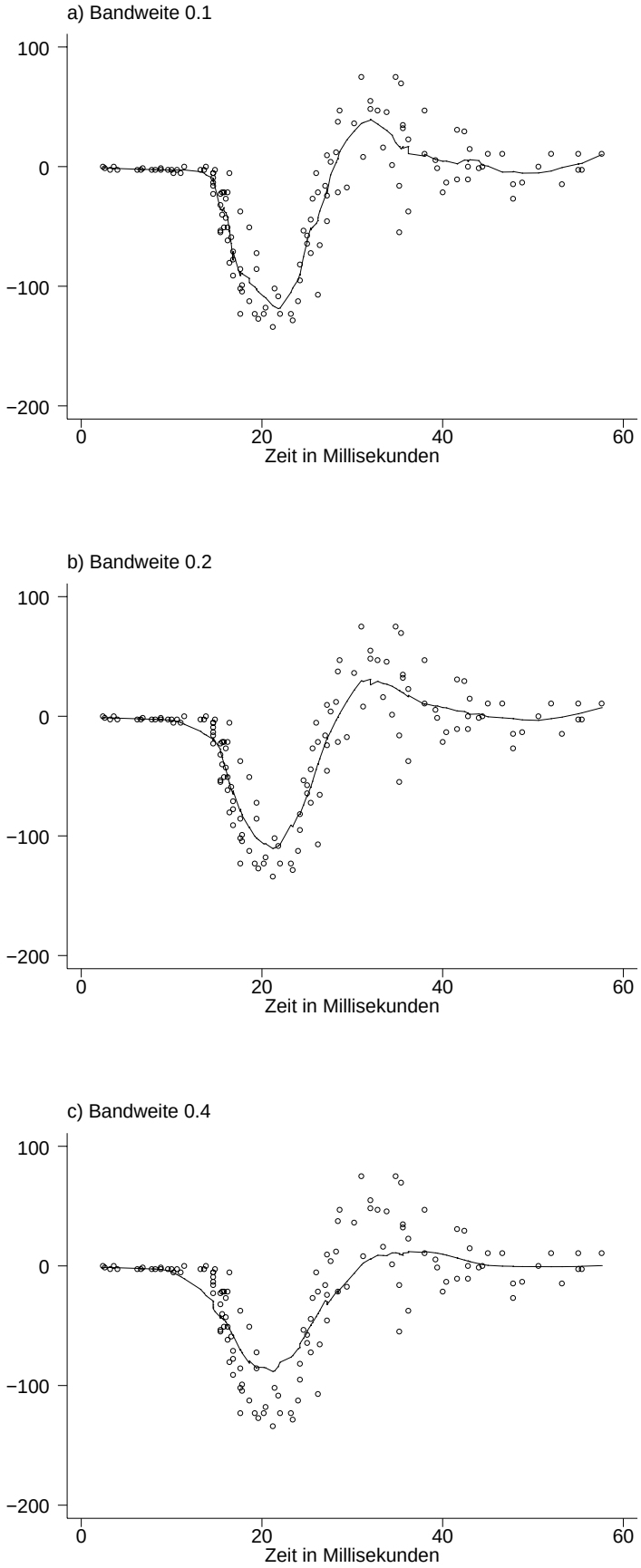

Abbildung 3.42. *Motorcycledaten: Loess Schätzer für verschiedene Bandweiten.*

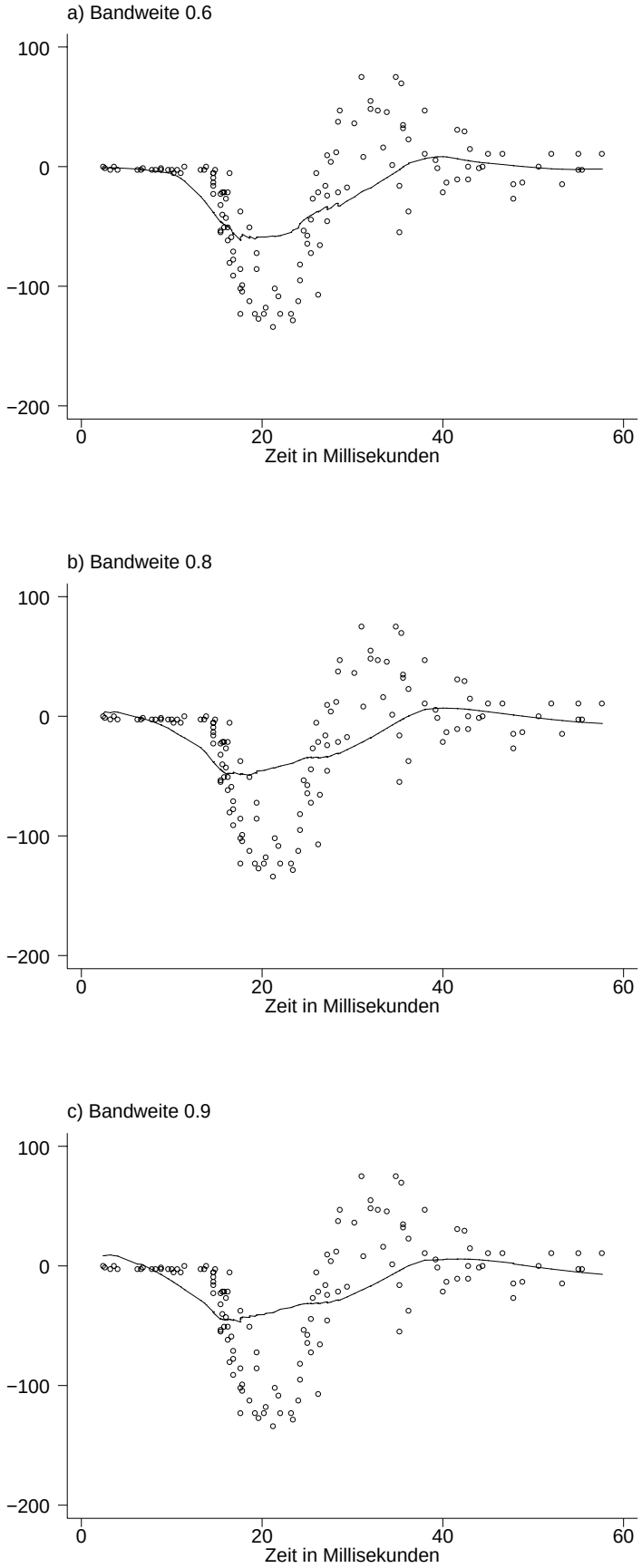

Abbildung 3.43. *Motorcycledaten: Loess Schätzer für verschiedene Bandweiten.*

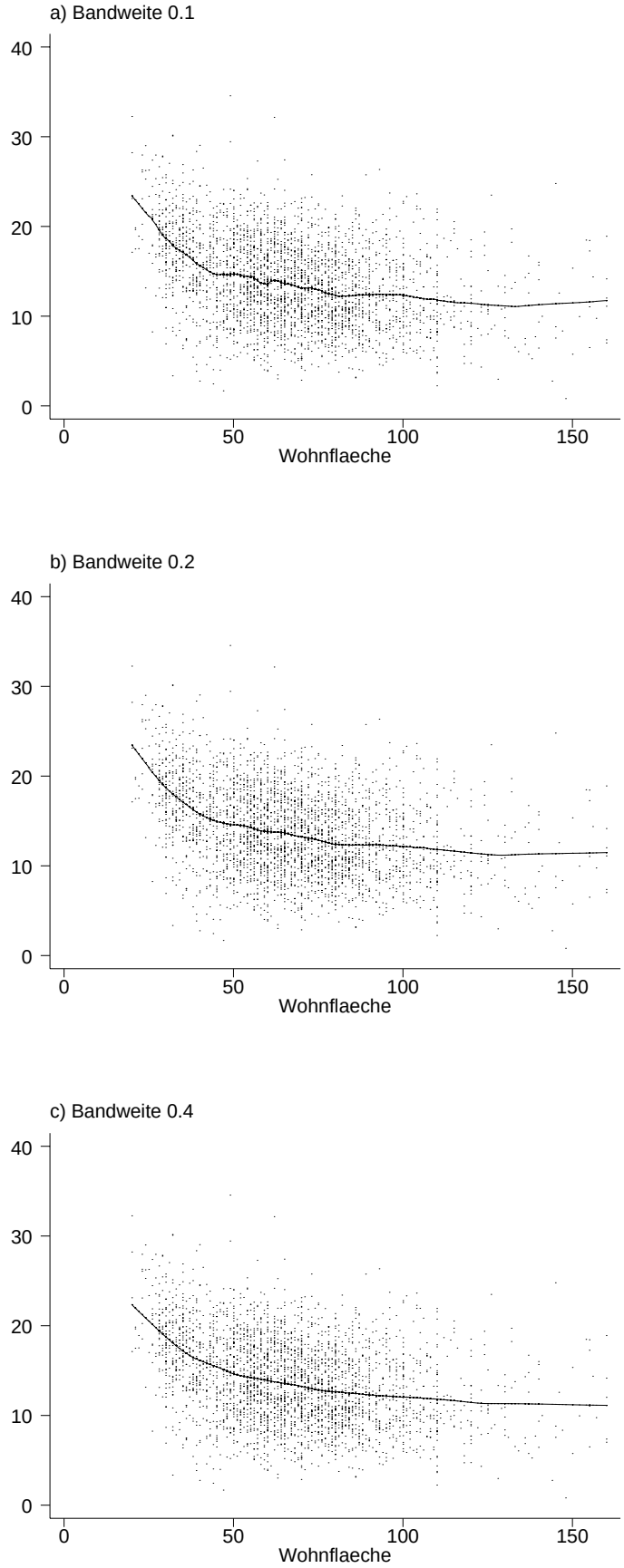

Abbildung 3.44. *Mietspiegeldaten: Loess Schätzer für verschiedene Bandweiten für die Regression zwischen Nettomiete pro Quadratmeter und Wohnfläche.*

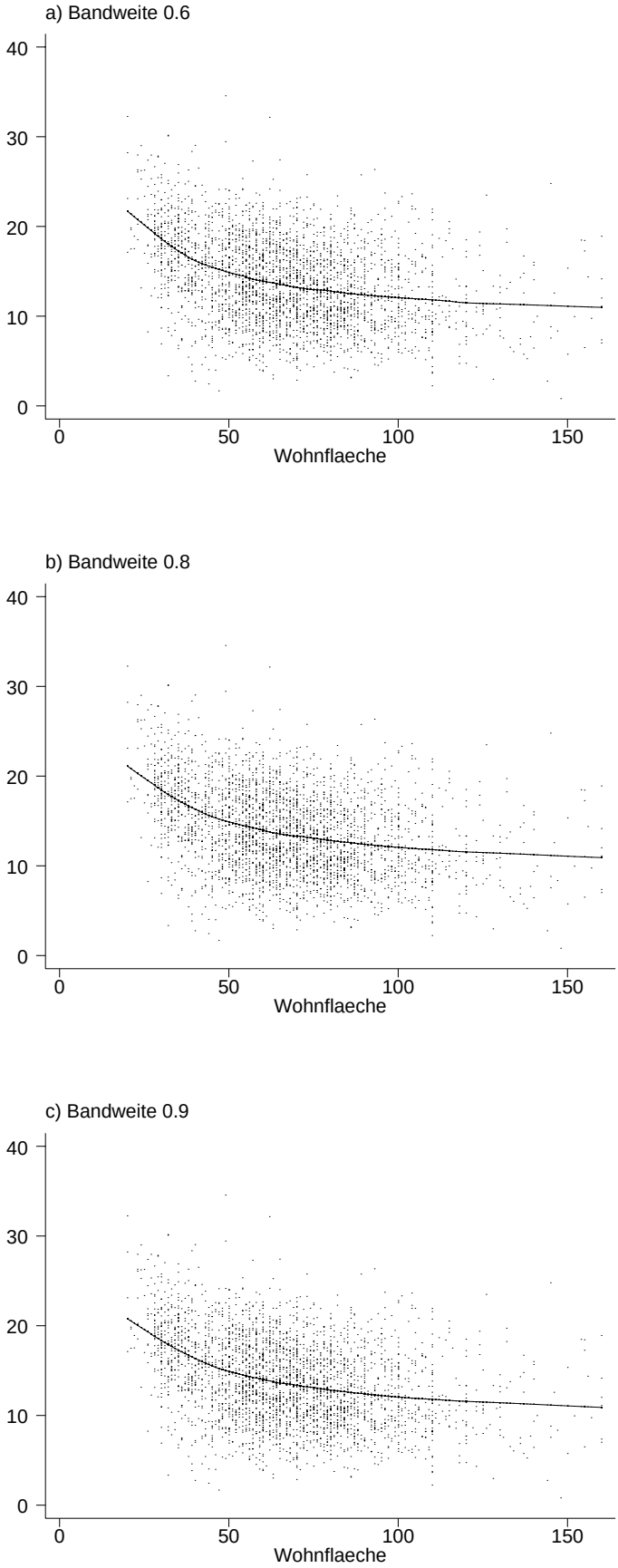

Abbildung 3.45. *Mietspiegeldaten: Loess Schätzer für verschiedene Bandweiten für die Regression zwischen Nettomiete pro Quadratmeter und Wohnfläche.*

4

Nichtparametrische Regression: Generalisierte Additive Modelle

4.1 Additive Modelle

4.1.1 Modelldefinition und Schätzalgorithmus

Im gesamten letzten Kapitel sind wir von der Situation ausgegangen, dass nur eine metrische Kovariable X vorhanden ist und haben das Modell

$$y_i = f(x_i) + \varepsilon_i \tag{4.1}$$

betrachtet. Wir wollen im folgenden das Modell erweitern und annehmen, dass wir p metrische Kovariablen $X_1, \ldots, X_p$ beobachten. Eine naheliegende Verallgemeinerung von (4.1) ist gegeben durch

$$y_i = f(x_{i1}, \ldots, x_{ip}) + \varepsilon_i \tag{4.2}$$

wobei $f : I\!R^p \to I\!R$ wieder eine möglichst "glatte" Funktion in nunmehr p Variablen ist. Folgende Probleme treten auf:

- Für $p > 2$ kann f nicht mehr visualisiert werden und die Schätzergebnisse sind daher schwer zu interpretieren.
- Fluch der Dimension ("curse of dimensionality"):
  Betrachte den Einheitswürfel der Dimension p, wobei die Beobachtungen im Würfel gleichverteilt seien. Wir stellen uns die Frage, welche Länge ein (Teil-)würfel besitzen muss, der $100 \times span\%$ der Daten enthält? Die folgende kleine Tabelle gibt darüber Aufschluss:

$p = 1$	$l = span$
$p = 2$	$l = span^{\frac{1}{2}}$
$p = 3$	$l = span^{\frac{1}{3}}$
$\vdots$	$\vdots$
Allgemein	$l = span^{\frac{1}{p}}$

  Als Beispiel betrachten wir in der nächsten Tabelle $span = 0.1$, d.h. 10% der Daten:

$$\begin{array}{ll} p = 1 & l = 0.1 \\ p = 2 & l = 0.3 \\ p = 3 & l = 0.47 \\ \vdots & \vdots \\ p = 10 & l = 0.8 \end{array}$$

Wir erkennen, dass in höheren Dimensionen entweder die Länge des Würfels, in dem lokal geglättet wird, erhöht werden muss, wodurch aber der Bias der Schätzungen ansteigt. Alternativ kann man die Länge gleich lassen und gleichzeitig den *span* verringern, d.h. weniger Daten benutzen. Dadurch wird aber die Varianz der Schätzung erhöht. Man spricht hier vom Fluch der Dimension!

Als Abhilfe betrachten wir im folgenden ein einfacheres Modell als (4.2). Im linearen Modell haben die Kovariablen einen *additiven* Einfluss und die funktionale Form des Einflusses ist *linear*, dh.

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon_i.$$

In sogenannten *additiven Modellen* behalten wir die Additivität der Einflussgrößen bei, während die Annahme eines linearen Einflusses fallengelassen wird. Wir erhalten

$$y_i = \eta_i + \varepsilon_i = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i.$$

Dabei sind $f_1, \ldots, f_p$ unbekannte "glatte" Funktionen von $x_1, \ldots, x_p$, die geschätzt werden müssen.

Additive Modelle haben ein *Identifizierbarkeitsproblem*, das Niveau der einzelnen Funktionen $f_1, \ldots, f_p$ ist nicht identifizierbar. Betrachte zur Illustration ein additives Modell mit zwei Kovariablen

$$y = \eta + \varepsilon = \beta_0 + f_1(x_1) + f_2(x_2) + \varepsilon.$$

Addiert man zu f_1 eine Konstante c und subtrahiert man diese von f_2, so bleibt der additive Prädiktor η unverändert. Wir fordern daher zusätzlich

$$E(f_j(x_j)) = 0 \qquad j = 1, \ldots, p,$$

d.h. die unbekannten Funktionen sind um Null zentriert. Diese Forderung sichert nicht nur die Identifizierbarkeit des Modells, sondern erleichtert auch die Interpretation. Positive Funktionswerte weisen im Vergleich zum Mittelwert auf tendenziell höhere Werte der Responsevariablen für den entsprechenden Kovariablenwert hin. Negative Funktionswerte auf tendenziell niedrigere Werte.

Die Schätzung additiver Modelle kann mit Hilfe des Backfitting-Algorithmus auf Scatterplotsmoother zurückgeführt werden. Beim Backfitting Algorithmus werden nacheinander

univariate Scatterplotsmoother auf die jeweiligen *partiellen Residuen* angewandt. Seien $\hat{\beta}_0$ und $\hat{f}_j = (\hat{f}_j(x_{1j}), \ldots, \hat{f}_j(x_{nj}))'$ Schätzungen von β_0 und den Funktionen f_j. Dann sind die partiellen Residuen r_i bezüglich Kovariable x_j definiert als

$$\begin{aligned} r_{ij} &= y_i - \hat{\beta}_0 - \hat{f}_1(x_{i1}) - \cdots - \hat{f}_{j-1}(x_{i,j-1}) - \hat{f}_{j+1}(x_{i,j+1}) - \cdots - \hat{f}_p(x_{ip}) \\ &= \hat{f}_j(x_{ij}) + y_i - \hat{\eta}_i \end{aligned}$$

wobei

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{f}_1(x_{i1}) + \cdots + \hat{f}_p(x_{ip}).$$

In Vektor- bzw. Matrixnotation erhalten wir

$$r_j = \hat{f}_j + y - \hat{\eta}.$$

Ausgehend von Startschätzungen $\hat{\beta}_0$ und $f_j^{(0)}$ erhält man im Backfitting Algorithmus verbesserte Schätzungen $f_j^{(1)}$ durch

$$\hat{f}_j^{(1)} = S_j \left(y - \hat{\beta}_0 - \sum_{k<j} \hat{f}_k^{(1)} - \sum_{k>j} \hat{f}_k^{(0)} \right),$$

wobei S_j ein beliebiger Glätter sei. Das Verfahren wird solange iteriert bis sich die resultiernden Funktionsschätzungen nicht mehr ändern. Insgesamt erhalten wir:

Algorithmus 4.1 (Backfitting)

i) Initialisierung:

Definiere Startwerte

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

und

$$\hat{f}_j^{(0)} = f_j^* \qquad j = 1 \ldots, p.$$

Setze $r = 1$.

ii) Für $j = 1, \ldots, p$ berechne

$$\hat{f}_j^{(r)} = S_j \left(y - \hat{\beta}_0 - \sum_{k<j} \hat{f}_k^{(r)} - \sum_{k>j} \hat{f}_k^{(r-1)} \right)$$

wobei S_j beliebige Scatterplotsmoother (z.B. P-Splines, Smoothingsplines etc.) sind.

iii) Zentrierung:

Für $j = 1, \ldots, p$ zentriere die Funktionen um Null, d.h.

$$\hat{f}_j^{(r)} = \hat{f}_j^{(r)} - \frac{1}{n} \sum_{k=1}^{n} \hat{f}_j^{(r)}(x_{kj}).$$

iv) Fahre fort mit ii) bis sich die geschätzten Funktionen nicht mehr ändern.

Häufig ist Analog zum gewichteten linearen Modell (vergleiche Abschnitt 3.1.4) der ungewichtete durch einen *gewichteten Backfitting Algorithmus* zu ersetzen, beispielsweise wenn die Fehler heteroskedastisch sind, vergleiche auch Beispiel 4.1. Insbesondere bei der Schätzung generalisierter additiver Modelle im übernächsten Abschnitt 4.3 wird eine gewichtete Version des Backfitting benötigt. Eine gewichtete Version des Backfitting Algorithmus erhält man, indem jeweils gewichtete Versionen der verwendeten Scatterplotsmoother verwendet werden.

Bemerkungen:

– Die Scatterplotsmoother müssen nicht identisch sein (z.B. ist eine Kombination von Glättungssplines und running mean Glättern erlaubt). Insbesondere können einzelne Funktionen auch linear sein, d.h.

$$f_j = \beta_j x_j$$

und damit

$$S_j = x_j(x_j' x_j)^{-1} x_j' y.$$

Damit kann man auch sogenannte *semiparametrische additive Modelle* schätzen:

$$y_i = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \beta' z_i + \varepsilon_i$$

– Die Konvergenz des Backfitting Algorithmus kann nicht in allen Fällen bewiesen werden, in der Praxis treten aber selten Probleme auf. Details zur Konvergenz des Algorithmus findet man in Hastie und Tisbshirani (1990) Kapitel 5.

– Der Backfitting Algorithmus minimiert folgendes penalisierte KQ-Kriterium

$$\sum_{i=1}^{n}(y_i - \beta_0 - f_1(x_{i1}) - \ldots - f_p(x_{ip}))^2 + \lambda_1 \int f_1''(x_1)^2 dx_1 + \ldots + \lambda_p \int f_p''(x_p)^2 dx_p$$

bezüglich $f_1, \ldots, f_p$, wenn als Scatterplotsmoother jeweils Glättungssplines verwendet werden.

– In manchen Fällen kann man effizienter ohne Backfitting schätzen, z.B. im einfachen multiplen Regressionsmodell oder bei Verwendung einfacher Basisfunktionsansätze. Backfitting ist dann am besten geeignet, wenn zur direkten Optimierung hochdimensionale Gleichungssysteme zu lösen sind.

– Im Backfitting Algorithmus sind auch zweidimensionale Scatterplotsmoother erlaubt. Sind zusätzlich die Haupteffekte im Modell enthalten, d.h. gilt

$$\eta = \ldots f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2) \ldots,$$

so müssen zusätzliche Restriktionen beachtet werden.

4.1.2 Interpretation additiver Modelle

Wir demonstrieren die Interpretation additiver Modelle anhand der Mietspiegeldaten. Wir gehen zunächst von dem einfachen additiven Modell

$$mproqm_i = \beta_0 + f_1(wfl_i) + f_2(bj_i) + \varepsilon_i$$

aus, in dem der Einfluss der Wohnfläche wfl und des Baujahrs bj auf die Nettomiete pro Quadratmeter durch nichtlineare Funktionen f_1 und f_2 modelliert wird. Die Schätzergebnisse für f_1 und f_2 findet man in Abbildung 4.1. Für den Intercept erhalten wir $\hat{\beta}_0 = 13.87$. Da beide Funktionen aus Gründen der Identifizierbarkeit um Null zentriert sind, kann der Intercept β_0 als Durchschnittsmiete für die Wohnfläche und Baujahrskombination aufgefasst werden, für die $f_1(wfl) = 0$ und $f_2(bj) = 0$ gilt. In unserem Beispiel erhalten wir $f_1 = 0$ für eine Wohnfläche von 58 Quadratmetern und $f_2 = 0$ für das Baujahr 1966, d.h. der geschätzte Parameter $\hat{\beta}_0 = 13.87$ entspricht der Durchschnittsquadratmetermiete (in DM) einer 1966 gebauten Wohnung mit 58 Quadratmetern Wohnfläche. Wohnungen mit einer größeren Wohnfläche als 58 Quadratmetern sind demnach billiger als die Referenzwohnung. Ebenso sind Wohnungen, die nach 1966 erbaut wurden, teuerer als die Referenzwohnung.

In einem zweiten Beispiel betrachten wir das semiparametrische additive Modell

$$mproqm_i = \beta_0 + f_1(wfl_i) + f_2(bj_i) + \beta_1 keingrraum_i + \varepsilon_i,$$

wobei $keingrraum$ eine Dummyvariable ist, die den Wert eins annimmt, wenn die Wohnung keinen Raum besitzt, der größer als 25 Quadratmetern ist. Wie im ersten Modell sind die beiden nichtlinearen Funktionen Null für 1966 gebaute Wohnungen mit 58 Quadratmetern Wohnfläche. Als Schätzung für den Intercept erhalten wir jetzt jedoch $\hat{\beta}_0 = 15.33$, der Intercept unterscheidet sich also relativ stark vom Intercept im ersten Modell. Dies ist aber nicht weiter verwunderlich, da der Intercept in diesem Modell eine andere Interpretation besitzt. Es handelt sich jetzt um die Durchschnittsmiete für eine 1966 erbaute Wohnung mit 58 Quadratmetern Wohnfläche und mindestens einem Raum, der größer als 25 Quadratmeter ist. Von dieser Referenzwohnung können alle weiteren Interpretationen des Modells erfolgen. Als Beispiel betrachten wir eine Wohnung, die 20 Jahre später erbaut wurde (alle anderen Merkmale halten wir fest). Es gilt $f_2(1986) = 3.75$, d.h. diese

Wohnung ist 3.75 DM teuerer als die Referenzwohnung. Sie hat einen Durchschnittspreis von $15.33 + 3.75 = 19.08$ DM.

Die Ergebnisse sind also stets bezüglich einer Referenzmerkmalskombination zu interpretieren. Der durschnittliche geschätzte Response kann im Intercept abgelesen werden. Je nach verwendeten Kovariablen und Parametrisierung der Kovariablen ändert sich aber unter Umständen die Referenzmerkmalskombination und damit die Interpretation des Intercepts.

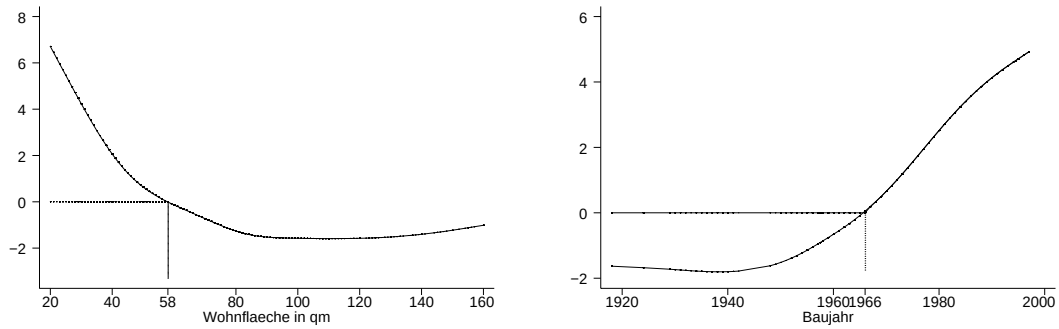

Abbildung 4.1. *Zur Interpretation additiver Modelle.*

4.1.3 Wahl der Glättungsparameter

Im Unterschied zu Scatterplotsmoothern sind bei additiven Modellen insgesamt p Glättungsparameter zu wählen. Diese können wieder mit generalisierter Kreuzvalidierung bestimmt werden, wobei die GCV Funktion

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\eta}_i)^2}{(1 - sp(\mathbf{R})/n)^2} \tag{4.3}$$

bezüglich $\lambda = (\lambda_1, \ldots, \lambda_p)'$ minimiert wird. Die Matrix R ist dabei die Gesamtsmoothermatrix, die gegeben ist durch

$$\hat{\eta} = \mathbf{R}y.$$

Die Funktion (4.3) kann als Approximation des Kreuzvalidierungskriteriums

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\eta}_i^{-i})^2$$

betrachtet werden. Darin ist $\hat{\eta}_i^{-i}$ eine Schätzung (eigentlich Prognose) des additiven Prädiktors für die i- te Beobachtung, wobei diese aber nicht zur Schätzung herangezogen wird. Die Minimierung von(4.3) ist technisch sehr kompliziert. Ein sehr effektives

Verfahren wird in Wood (2000) beschrieben und ist im Softwarepaket R implementiert, vergleiche auch Abschnitt 4.4.

Alternativ können Verfahren der Variablenselektion zur Bestimmung der Glättungsparameter herangezogen werden wie beispielsweise im Programmpaket S-plus implementiert. Dabei wird folgender Algorithmus benutzt:

Algorithmus 4.2 (Stepwise-Prozedur)

i) Bestimme ein Startmodell M_0. Beispielsweise könnte man ein einfaches lineares Modell verwenden, in dem alle Kovariableneffekte linear modelliert werden. Bestimme für jede Kovariable x_j, $j = 1, \ldots, p$ eine hierarchisch geordnete Liste $A_{x_j} = \{A^1_{x_j}, \ldots, A^q_{x_j}\}$ von Modellierungsalternativen. Diese Liste könnte für eine Kovariable x_j in etwa wie folgt aussehen:

1. *x_j ist nicht im Modell enthalten.*
2. *Modelliere den Effekt von x_j durch einen Glättungsspline mit einem äquivalenten Freiheitsgrad (entspricht einer linearen Funktion).*
3. *Modelliere den Effekt von x_j durch einen Glättungsspline mit $2, 3, \ldots$ oder 10 äquivalenten Freiheitsgraden.*

ii) Schätze ausgehend vom Startmodell M_0 eine Reihe weiterer Modelle, in denen jeweils für eine der Kovariablen die Modellierungsalternative getestet wird, die in der Hierarchie genau unter oder oberhalb des Startmodells liegt. Bestimme aus den getesteten Modellen das Modell M_1 mit dem besten Fit (bezüglich eines Modellwahlkriteriums).

1. *Falls das Modell M_1 besser ist als das Startmodel M_0, dann ersetze M_0 durch M_1 und fahre fort mit ii), andernfalls beende den Algorithmus.*

Als Modellwahlkriterien kommen prinzipiell mehrere in betracht, beispielsweise könnte das GCV Kriterium (4.3) verwendet werden. In S-Plus wird das Akaike Information Kriterium (AIC) benutzt, das definiert ist als

$$AIC = \sum_{i=1}^{n} \frac{(y_i - \hat{\eta}_i)^2}{\hat{\sigma}^2} + 2\hat{\sigma}^2 sp(\mathbf{R}).$$

ähnlich wie beim adjustierten R^2 im multiplen Regressionsmodell besteht das AIC aus zwei Termen. Der erste Term bestraft eine mangelnde Anpassung an die Daten, der zweite Term bestraft eine zu hohe Modellkomplexität.

Der Vorteil der Stepwise Prozedur besteht darin, dass neben der Glättungsparameterwahl auch eine Variablenselektion durchgeführt wird, da einzelne Kovariablen prinzipiell auch

ganz aus dem Modell verschwinden können. Nachteilig ist sicherlich, dass die "optimalen" Glättungsparameter in der Regel nicht gefunden werden.

Beispiel 4.1 (Mietspiegel für München)
Wir betrachten wieder den Mietspiegeldatensatz und schätzen das semiparametrische additive Modell

$$nmproqm_i = \beta_0 + f_1(wfl_i) + f_2(bj_i) + \beta' z_i + \varepsilon_i, \tag{4.4}$$

wobei z_i alle kategorialen Merkmale beinhaltet, welche die Aussattung und Lage der Wohnung beschreiben (z.B. Qualität der Küche usw.). Unter Verwendung des in R implementierten GCV Kriteriums zur Glättungsparameterwahl, erhalten wir die in Abbildung 4.2 gezeigten Schätzungen. Die Abbildungen zeigen die geschätzten Funktionen f_1 und f_2 mit punktweisen 95% Konfidenzbändern. In die Abbildungen a) und b) sind zusätzlich noch die jeweiligen partiellen Residuen eingezeichnet. Mit Hilfe der in S-plus implementierten Stepwise-Prozedur zur Bestimmung der Glättungsparameter erhalten wir sehr ähnliche Ergebnisse (vergleiche Abbildung 4.4). Sowohl bei den mit R als auch mit S-plus erzielten Ergebnissen wurden (Pseudo)glättungssplines verwendet. Unter Verwendung des loess Verfahrens und der Stepwise-Prozedur in S-plus ergeben sich die in Abbildung 4.5 gezeigten Schätzungen, die insbesondere für den Effekt der Wohnfläche etwas rauher sind.

Zur Überprüfung der Varianzhomogenität wurde mit R zusätzlich ein additives Modell mit dem Logarithmus der quadrierten Residuen $log(\hat{\varepsilon}_i^2)$ als abhängiger Variable geschätzt. Dabei wurden dieselben Kovariablen wie bei dem ursprünglichen Modell für die Nettomiete pro Quadratmeter verwendet. Wir erhalten die in Abbildung 4.3 abgebildeten Effekte für die Wohnfläche und das Baujahr. Offensichtlich sinken die Varianzen linear mit steigender Wohnfläche. Für das Baujahr ergibt sich ein eindeutig nichtlinearer Effekt mit einem Minimum um das Jahr 1970 herum. Mit Hilfe der geschätzten logarithmierten quadrierten Residuen $log(\widehat{\hat{\varepsilon}}_i^2)$ könnte man im Anschluss das Modell (4.4) erneut schätzen mit Gewichten

$$w_i = \frac{1}{\widehat{\hat{\varepsilon}}_i^2}.$$

△

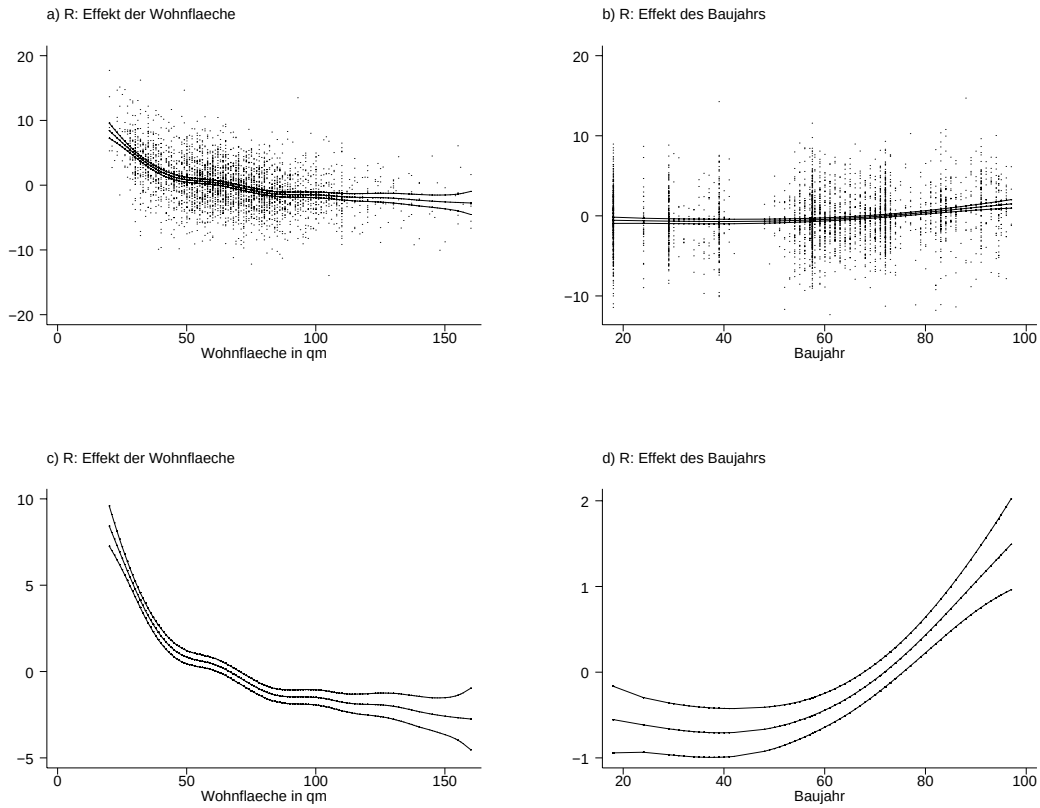

Abbildung 4.2. *Mietspiegeldaten: Geschätzte nichtlineare Effekte der Wohnfläche (Abbildungen a) und c)) und des Baujahrs (Abbildungen b) und d)) inklusive 95% punktweise Konfidenzbänder. In die Abbildungen a) und b) sind zusätzlich noch die jeweiligen partiellen Residuen eingezeichnet. Die Schätzungen basieren auf (Pseudo)glättungssplines. Die Glättungsparameter wurden durch generalisierte Kreuzvalidierung unter Verwendung von R geschätzt.*

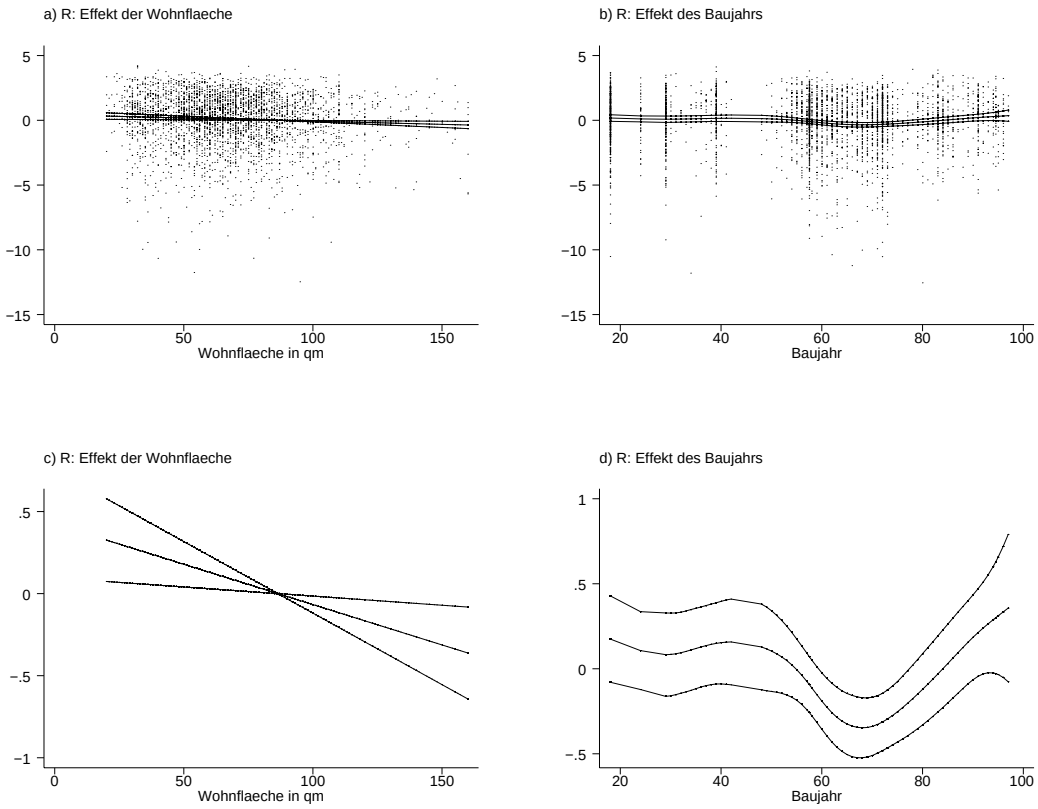

Abbildung 4.3. *Mietspiegeldaten: Geschätzte nichtlineare Effekte der Wohnfläche (Abbildungen a) und c)) und des Baujahrs (Abbildungen b) und d)) inklusive 95% punkweise Konfidenzbänder für die Regression mit den logarithmierten quadrierten Residuen* $log(\hat{\varepsilon}_i^2)$ *zur Überprüfung der Varianzhomogenität.*

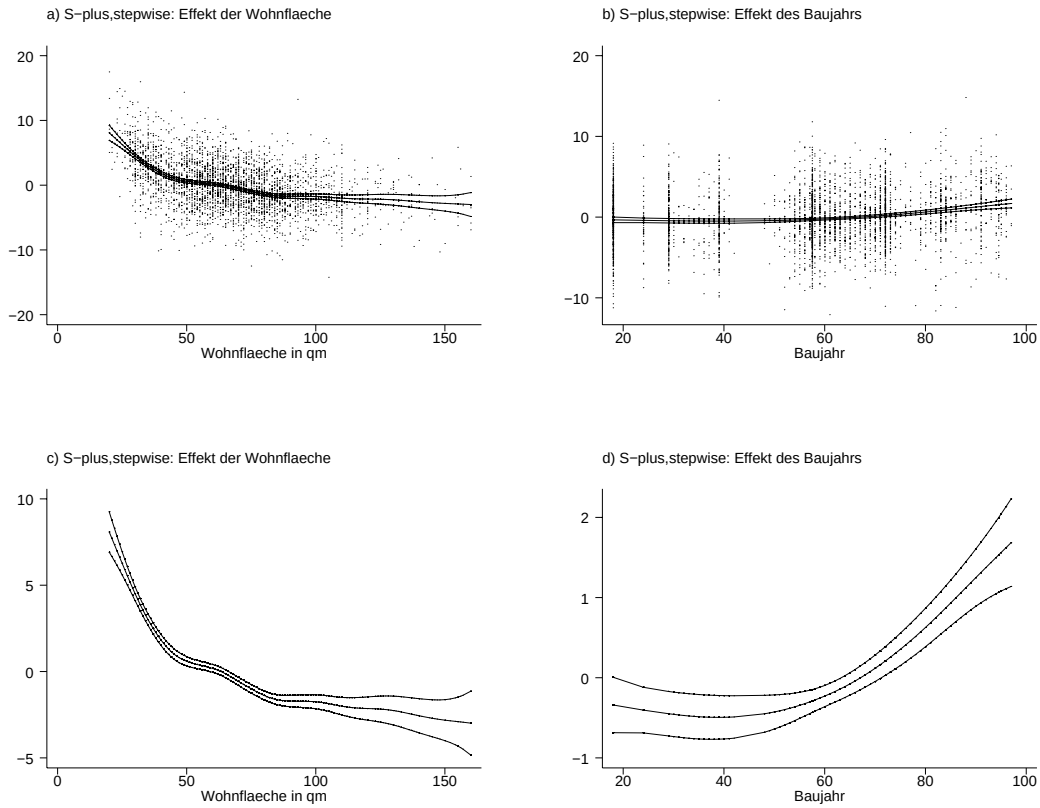

Abbildung 4.4. *Mietspiegeldaten: Geschätzte nichtlineare Effekte der Wohnfläche (Abbildungen a) und c)) und des Baujahrs (Abbildungen b) und d)) inklusive 95% punkweise Konfidenzbänder. In die Abbildungen a) und b) sind zusätzlich noch die jeweiligen partiellen Residuen eingezeichnet. Die Schätzungen basieren auf (Pseudo)glättungssplines. Die Glättungsparameter wurden unter Zuhilfenahme der Stepwise-Prozedur in S-plus geschätzt.*

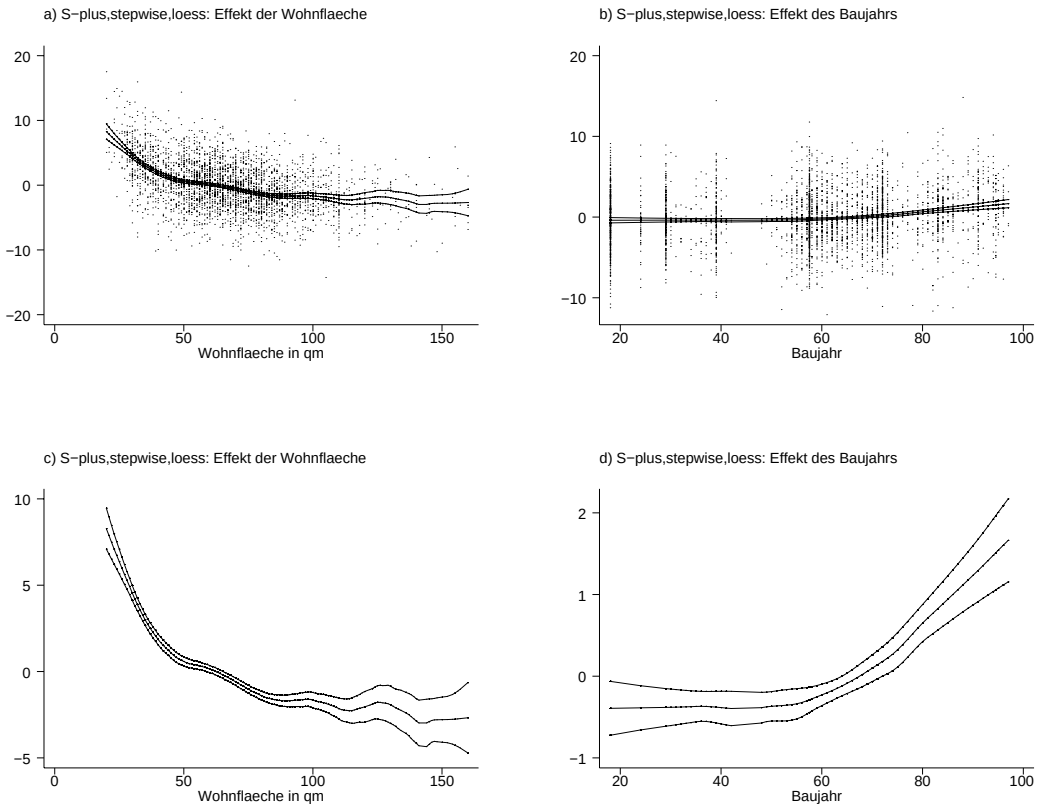

Abbildung 4.5. *Mietspiegeldaten: Geschätzte nichtlineare Effekte der Wohnfläche (Abbildungen a) und c)) und des Baujahrs (Abbildungen b) und d)) inklusive 95% punkweise Konfidenzbänder. In die Abbildungen a) und b) sind zusätzlich noch die jeweiligen partiellen Residuen eingezeichnet. Die Schätzungen basieren auf dem loess Verfahren. Die Glättungsparameter wurden unter Zuhilfenahme der Stepwise-Prozedur in S-plus geschätzt.*

4.2 Wiederholung: Generalisierte lineare Modelle

4.2.1 Definition

Bisher haben wir lediglich Modelle betrachtet, bei denen die Responsevariable Y metrisch ist. Zur Modellierung des Einflusses der Kovariablen haben wir in Kapitel 3.1 das lineare Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

und im letzten Abschnitt 4.1 das additive Modell

$$y_i = \eta_i + \varepsilon_i = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i.$$

behandelt.

Eine Voraussetzung für die Anwendbarkeit dieser Modelle ist, dass die abhängige Variable metrisch ist.

Um eine allgemeinere Modellklasse handelt es sich bei den *generalisierten linearen Modellen*, die auch kategorialen Response oder Häufigkeiten (etwa Schadenshäufigkeiten bei Versicherungen) als Response zulassen. Wir sprechen von einem generalisierten linearen Modell, wenn folgende Annahmen erfüllt sind:

1. **Verteilungsannahme:**
   Die bedingte Verteilung von y_i gegeben x_i gehört einer einfachen Exponentialfamilie an, d.h. die (bedingte) Dichte d_i von y_i hat die Gestalt

$$d_i(y_i|\theta_i, \phi, w_i, x_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)\right\}, \qquad i, \ldots, n, \tag{4.5}$$

   wobei

   $b(\cdot)$, $c(\cdot)$ gewisse (reellwertige) Funktionen sind,

   $\theta_i \in \Theta \subset \mathbb{R}^{p+1}$ der sogenannte natürliche Parameter ist,

   der zusätzliche (bekannte oder unbekannte) Parameter $\phi \in \Phi \subset \mathbb{R}$ einen Skalenparameter darstellt,

   und die w_i (bekannte) Gewichte sind.

   Für einfache Exponentialfamilien gilt allgemein (vgl. Mc Cullagh und Nelder (1989) S. 28)

$$E(y_i|x_i) = \mu_i = b'(\theta_i) \tag{4.6}$$

und

$$Var(y_i|x_i) = \phi \frac{b''(\theta_i)}{w_i}, \tag{4.7}$$

wobei $b''(\theta_i)$ auch Varianzfunktion heißt.

Wichtige Vertreter von einfachen Exponentialfamilien sind die Normalverteilung, die Gammaverteilung, die Binomialverteilung und die Poissonverteilung. In der folgenden Tabelle sind einige wichtige Verteilungen mit ihren Charakteristika zusammengefasst (vgl. auch Mc Cullagh und Nelder (1989) S. 30).

2. **Strukturelle Annahme:**
   Der (bedingte) Erwartungswert μ_i von y_i hängt über eine sogenannte Responsefunktion $h : \mathbb{R} \mapsto \mathbb{R}$ vom linearen Prediktor $\eta_i = x_i'\beta$ ab, d.h.

$$\mu_i = h(\eta_i) \qquad i = 1, \ldots, n. \tag{4.8}$$

   Bei dem $(p+1) \times 1$ Vektor β handelt es sich um einen zu schätzenden unbekannten Parameter. Gilt $\eta_i = \theta_i$, so heißt die dazugehörige Responsefunktion natürlicher Response.

Verteilung	Bezeichnung	Definitionsb.	θ	$b(\theta)$	ϕ
Normal	$N(\mu, \sigma^2)$	$\mathbb{R}$	μ	$\theta^2/2$	σ^2
Bernoulli	$B(1, \pi)$	$0, 1$	$\log(\frac{\pi}{(1-\pi)})$	$\log(1 + e^\theta)$	1
(skaliert) Binomial	$B(n, \pi)/n$	$0, 1/n, \ldots, 1$	$\log(\frac{\pi}{(1-\pi)})$	$\log(1 + e^\theta)$	$1/n$
Poisson	$Po(\mu)$	$0, 1, 2, \ldots$	$\log(\mu)$	e^θ	1
Gamma	$Ga(\mu, \nu)$	$\mathbb{R}_+$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}

Table 4.1. *Einfache Exponentialfamilien und ihre Charakteristika.*

4.2.2 Beispiele für generalisierte lineare Modelle

Metrischer Response

Gegeben sei das klassische lineare Regressionsmodell

$$y \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n), \qquad \sigma^2 > 0 \tag{4.9}$$

bzw. komponentenweise

$$y_i \sim N(x_i'\beta, \sigma^2) = N(\eta_i, \sigma^2). \tag{4.10}$$

Dieses erweist sich mit der Responsefunktion $h(\eta_i) = \eta_i = id$ als spezielles generalisiertes lineares Modell.

In den Rahmen der generalisierten linearen Modelle fallen aber zusätzlich auch einige häufig benutzte nichtlineare Modelle. Beispielsweise erfüllen die Modelle

$$y_i \sim N(\exp(x_i'\beta), \sigma^2) \tag{4.11}$$

oder

$$y_i \sim N((x_i'\beta)^2, \sigma^2) \tag{4.12}$$

mit den Responsefunktionen $h(\eta_i) = \exp(\eta_i)$ bzw. $h(\eta_i) = \eta_i^2$ die Vorraussetzungen eines generalisierten linearen Modells.

Handelt es sich bei Y um ein positives Merkmal, so dass die Beobachtungen y_i sämtlich größer Null sind, dann kann es von Vorteil sein, anstelle der Normalverteilung eine Gammaverteilung für die y_i anzunehmen. In der Regel nimmt man

$$y_i|x_i \sim Ga(\mu_i, \nu)$$

mit dem Erwartungswert $\mu_i = exp(\eta_i) = exp(x_i'\beta)$. Die Responsefunktion $h(\eta_i) = exp(\eta_i)$ stellt also sicher, dass der geschätzte Erwartungswert stets positiv ist. Bei ν handelt es sich in diesem Fall um einen Shapeparameter.

Binärer und binomialer Response

Im Gegensatz zum vorangegangenen Abschnitt, in dem Y metrisch war, gehen wir in diesem Abschnitt davon aus, dass das interessierende Merkmal Y kategorial ist. Speziell nehmen wir an, dass Y und damit die Beobachtungen y_i nur zwei Werte annehmen können, z.B. 1 und 0. Eins könnte dann zum Beispiel das Vorhandensein einer bestimmten Augenkrankheit bedeuten und der Wert Null das Fehlen dieser Krankheit. Ziel einer Analyse ist es dann herauszufinden, ob verschiedene Kovariablen einen Einfluss auf die Erfolgswahrscheinlichkeit $\pi_i = P(y_i = 1|x_i)$ besitzen. Denkbare Kovariablen (in diesem Zusammenhang auch Risikofaktoren) für obiges Beispiel wären z.B. das Alter, das Geschlecht oder auch das Vorliegen einer anderen Krankheit, welche das Auftreten der Augenkrankheit begünstigt.

Als Verteilung der y_i können wir eine Bernoulli– bzw. Binomialverteilung annehmen, d.h. $y_i \sim B(1, \pi_i)$. In diesem Fall stimmt der Erwartungswert μ_i von y_i mit der Erfolgswahrscheinlichkeit π_i überein. Durch die Wahl einer geeigneten Responsefunktion, über die der Erwartungswert π_i (also die Erfolgswahrscheinlichkeit) von den Kovariablen beeinflusst wird, befinden wir uns wieder im Rahmen der generalisierten linearen Modelle.

Eine natürliche Wahl der Responsefunktion ergibt die identische Abbildung, d.h.wir gehen davon aus, dass die Erfolgswahrscheinlichkeit in linearer Weise von den Kovariablen abhängt, also

$$\pi_i = \eta_i = x_i'\beta.$$

In diesem Fall müssten aber, da die Erfolgswahrscheinlichkeit π_i zwischen Null und Eins liegt, Restriktionen an die Parameter beachtet werden, was zu erheblichen Schätzproblemen führen kann. Aus diesem Grund wählt man als Responsefunktion eine streng monotone Verteilungsfunktion F, so dass die Werte des linearen Prediktors η_i auf das Intervall zwischen Null und Eins transformiert werden. Es gilt dann

$$P(y_i = 1|x_i) = \pi_i = F(\eta_i) = F(x_i'\beta)$$

bzw.

$$y_i \sim B(1, F(\eta_i)) = B(1, F(x_i'\beta)).$$

Spezielle Wahl von F führt zu verschiedenen Modellen. Wählt man den natürlichen Response

$$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

so erhält man ein sogenanntes Logitmodell, während die Wahl der Verteilungsfunktion der Standardnormalverteilung, d.h.

$$\pi_i = h(\eta_i) = \Phi(\eta_i)$$

zu einem Probitmodell führt. Ein weiteres Modell erhält man mit der sogenannten Extreme–minimal–value Verteilung, d.h. mit

$$\pi_i = h(\eta_i) = 1 - \exp(-\exp(\eta_i)). \tag{4.13}$$

Häufig kommt es im Zusammenhang mit Modellen für binären Response vor, dass sämtliche Kovariablen kategorial sind und damit nur wenige verschiedene Kovariablenvektoren vorliegen. Die Daten werden dann gruppiert. Nimmt man als Responsebeobachtung y_i den Mittelwert, d.h. die relative Häufigkeit mit der in der i–ten Gruppe der Wert Eins vorkommt, dann sind die y_i skaliert binomialverteilt, d.h. sie nehmen die Werte $0, 1/n_i, 2/n_i, \ldots, n_i/n_i = 1$ an und es gilt $E(y_i|x_i) = \pi_i$ und $Var(y_i|x_i) = \dfrac{\pi_i(1-\pi_i)}{n_i}$. Für die Dichte gilt:

$$f_i(y_i|\beta, \mathbf{X}) = \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1-\pi_i)^{n_i(1-y_i)}, \qquad y_i = 0, 1/n_i, \ldots, 1 \tag{4.14}$$

Wir befinden uns somit auch im gruppierten Fall im Rahmen der Generalisierten Linearen Modelle.

Zähldaten

Gegeben seien nunmehr Beobachtungen $y_1, \ldots, y_n$, welche die Häufigkeit angeben, mit der ein Ereignis in einer bestimmten Zeitperiode auftritt. Daten dieser Art heißen Zähldaten. Typische Beispiele sind das Auftreten von Versicherungsfällen innerhalb eines Jahres oder die Anzahl der Unfälle bei Kfz Haftpflichtversicherungen. Mögliche Kovariablen, welche z.B. die Anzahl der Unfälle beeinflussen, sind dann das Alter des Versicherungsnehmers das Alter des Autos, die Jahresfahrleistung. Ziel einer Analyse ist es also Faktoren zu bestimmen, welche die erwartete Häufigkeit mit der das Ereignis auftritt, möglichst gut determinieren.

Nehmen wir nun an, dass die Auftretenshäufigkeit des Ereignisses zu jedem Zeitpunkt innerhalb der Zeitperiode zumindest näherungsweise einem Poissonprozess[1] folgt, dann sind die Häufigkeiten y_i am Ende der Periode possionverteilt, d.h.

$$y_i \sim Po(\mu_i), \tag{4.15}$$

wobei der Parameter μ_i bekanntlich der Erwartungswert (und die Varianz) von y_i ist, also die erwartete Häufigkeit. Der Einfluss der Kovariablen x_i auf die erwartete Häufigkeit wird in der Regel multiplikativ angesetzt, d.h.

$$E(y_i|x_i) = \mu_i = \exp(x_i'\beta). \tag{4.16}$$

Geht man von der logarithmierten erwarteten Häufigkeit aus, dann gilt

$$\log(\mu_i) = x_i'\beta, \tag{4.17}$$

so dass obiges Modell vor allem unter dem Namen loglineares Poissonmodell bekannt ist.

4.2.3 Schätzungen von generalisierten linearen Modellen

Die Schätzung der Regressionsparameter β basiert auf dem Maximum Likelihood Prinzip.

Die (Log)likelihood der i-ten Beobachtung ist gegeben durch

$$l_i(\theta_i) = \log(d_i(y_i|\theta_i, \phi, w_i, x_i)) = \frac{y_i\theta_i - b(\theta_i)}{\phi} w_i.$$

Wegen $\theta_i = \theta(\mu_i)$ und $\mu_i = h(x_i'\beta)$ können wir l_i in Abhängigkeit von β schreiben, d.h.

$$l_i(\beta) = \frac{y_i\theta(h(x_i'\beta)) - b(\theta(h(x_i'\beta)))}{\phi}\omega_i$$

[1] Ein Zählprozess heißt Poissonprozess, wenn er unabhängige und stationäre Zuwächse besitzt und die Sprunghöhen der Trajektorien (mit Wahrscheinlichkeit 1) die Höhe 1 haben.

Die gesamte Loglikelihood ist die Summe der individuellen Likelihoods

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta).$$

Wir maximieren $l(\beta)$ bezüglich β durch differenzieren und null setzen

$$\frac{\delta l}{\delta \beta} = s(\beta) = \sum_{i=1}^{n} s_i(\beta) = 0.$$

Die Ableitung der Loglikelihood $s(\beta)$ bezüglich β heißt Scorefunction. Das Gleichungssystem ist in der Regel nicht linear und muss in einem iterativen Verfahren gelöst werden.

Der gebräuchlichste Schätzalgorithmus ist unter dem Namen Fisherscoring bekannt. Das Verfahren besteht im wesentlichen darin, dass ausgehend von einer Startschätzung $\hat{\beta}^{(0)}$, verbesserte Schätzer $\hat{\beta}^{(r+1)}$, $r = 0, 1, \ldots$ als gewichtete KQ-Schätzer gewonnen werden, wobei die jeweiligen Gewichte und sogenannte Arbeitsbeobachtungen $\tilde{y}_i$ vom gerade aktuellen Schätzwert $\hat{\beta}^{(r)}$ und natürlich von der jeweiligen Exponentialfamilie abhängen. Das Verfahren wird solange iteriert bis sich die Schätzungen nicht mehr verändern.

Algorithmus 4.3 (Iterativ gewichtete KQ-Schätzung)

i) Initialisiere $\hat{\beta}^{(0)} = \beta^*$ (z.B. $\beta^* = 0$). Setze $r = 0$.

ii) Die Schätzung $\hat{\beta}^{(r+1)}$ für β in der $r+1$-ten Iteration ist gegeben durch

$$\hat{\beta}^{(r+1)} = (\mathbf{X}'\mathbf{W}^{(r)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(r)}\tilde{y}^{(r)}$$

mit

$$\begin{aligned}
\mathbf{W}^{(r)} &= diag(d_1^{(r)}, \ldots, d_n^{(r)}) \\
\eta_i^{(r)} &= x_i'\hat{\beta}^{(r)} \\
\mu_i^{(r)} &= h(x_i'\hat{\beta}^{(r)}) \\
\theta_i^{(r)} &= \theta(h(x_i'\hat{\beta}^{(r)})) \\
d_i^{(r)} &= \left(\frac{\delta h(\eta_i^{(r)})}{\delta \eta}\right)^2 \cdot \left(\frac{\delta b(\theta_i^{(r)})}{\delta^2 \theta}\right)^{-1} \frac{w_i}{\phi} \\
\tilde{y}_i^{(r)} &= \eta_i^{(r)} + \left(\frac{\delta h(\eta_i^{(r)})}{\delta \eta}\right)^{-1} \cdot (y_i - \mu_i^{(r)}).
\end{aligned}$$

iii) Falls für $\epsilon > 0$

$$\frac{||\hat{\beta}^{(r+1)} - \hat{\beta}^{(r)}||}{||\hat{\beta}^{(r)x}||} \leq \epsilon$$

dann beende den Algorithmus, ansonsten setze $r = r + 1$ und fahre fort mit ii).

Zur Durchführung des Algorithmus werden für die jeweiligen Exponentialfamilienverteilungen die Ableitungen $\frac{\delta h(\eta)}{\delta \eta}$ und $\frac{\delta b(\theta)}{\delta^2 \theta}$ benötigt. Diese findet man in Tabelle 4.2.3.

Modell	$h(\eta)$	$\frac{\delta h(\eta)}{\delta \eta}$	$\frac{\delta b(\theta)}{\delta^2 \theta}$	Gewicht d
Normal	η	1	1	1
Logit	$\frac{\exp(\eta)}{1+\exp(\eta)}$	$\frac{\exp(\eta)}{(1+\exp(\eta))^2}$	$\pi(1-\pi)$	$\pi(1-\pi)$
logl. Poisson	$\exp(\eta)$	$\exp(\eta)$	$\exp(\eta)$	$\exp(\eta)$

Table 4.2. *Die Größen $h(\eta)$, $\frac{\delta h(\eta)}{\delta \eta}$, $\frac{\delta b(\theta)}{\delta^2 \theta}$ sowie die Gewichte d der Gewichtsmatrix* $\mathbf{W}$ *für einige einfache Exponentialfamilien.*

Beispiel 4.2 (Binäres Logitmodell)
Wir veranschaulichen den Schätzalgorithmus für ein binäres Logitmodell, d.h. $y_i \in \{0,1\}$. Für die Wahrscheinlichkeitsdichte gilt

$$\begin{aligned}
P(y_i|x_i) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
&= \exp(y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)) \\
&= \exp(y_i \log(\pi_i/(1-\pi_i)) + \log(1-\pi_i)) \\
&= \exp(y_i \log(\pi_i/(1-\pi_i)) - \log(1/(1-\pi_i))) \\
&= \exp(y_i \log(\pi_i/(1-\pi_i)) - \log(1+\pi_i/(1-\pi_i))) \\
&= \exp(y_i \log(\pi_i/(1-\pi_i)) - \log(1+\exp(\log(\pi_i/(1-\pi_i))))).
\end{aligned}$$

Damit ist gezeigt, dass es sich bei der Bernoulliverteilung um eine einfache Exponentialfamilie handelt mit natürlichem Parameter $\theta_i = \log(\pi_i/(1-\pi_i))$ und $b(\theta_i) = \log(1+\exp(\theta_i))$. Für die Erfolgswahrscheinlichkeiten π_i nehmen wir in einem Logitmodell

$$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{(1+\exp(\eta_i))} = \frac{\exp(x_i'\beta)}{(1+\exp(x_i'\beta))}$$

an. Damit gilt

$$h'(\eta_i) = \frac{(1+\exp(\eta_i))\exp(\eta_i) - \exp(\eta_i)\exp(\eta_i)}{(1+\exp(\eta_i))^2} = \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2} = \pi_i(1-\pi_i)$$

und wegen $Var(y_i \,|\, x_i) = b''(\theta_i)$

$$b''(\theta_i) = \pi_i(1-\pi_i).$$

Für die Arbeitsbeobachtungen $\tilde{y}_i$ erhalten wir somit

$$\tilde{y}_i = \eta_i + \frac{(y_i - \pi_i)}{\pi_i(1-\pi_i)}$$

und für die Elemente der Gewichtsmatrix $\mathbf{W}$

$$d_i = \frac{h'(\eta_i)^2}{b''(\theta_i)} = \frac{(\pi_i(1-\pi_i))^2}{\pi_i(1-\pi_i)} = \pi_i(1-\pi_i) = Var(y_i \,|\, x_i).$$

△

4.3 Generalisierte additive Modelle

Bei generalisierten additiven Modellen wird der lineare Prädiktor

$$\eta_i = x_i'\beta$$

ersetzt durch einen additiven Prädiktor

$$\eta_i = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

d.h.

$$\mu_i = h(\eta_i) = h(\beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip})).$$

Die unbekannten Funktionen in generalisierten additiven Modellen können durch eine Kombination des iterativen KQ-Algorithmus 4.3 und des Backfitting Algorithmus 4.1 geschätzt werden. Ausgehend von Startschätzungen $\hat{\beta}_0^{(0)}$ und $\hat{f}_j^{(0)}$, $j = 1, \ldots, p$, werden zunächst Arbeitsbeobachtungen $\tilde{y}_i$ und Gewichte d_i berechnet. Anschließend wird ein gewichtetes additives Modell mit Hilfe des Backfitting Algorithmus geschätzt woraus verbesserte Schätzungen $\hat{\beta}_0^{(1)}$ und $\hat{f}_j^{(1)}$ resultieren. Damit werden neue Arbeitsbeobachtungen und Gewichte bestimmt und erneut ein gewichtetes additives Modell geschätzt. Das Verfahren wird solange iteriert, bis sich die geschätzten Funktionen nicht mehr ändern. Zusammenfassend erhalten wir:

Algorithmus 4.4

i) Initialisierung:
   Setze (zum Beispiel)

$$\hat{\beta}_0^{(0)} = h^{-1}(\overline{y})$$

   und

$$\hat{f}_0^{(0)} = \cdots = \hat{f}_p^{(0)} = 0.$$

   Setze $r = 0$

ii) Definiere wie beim iterativen KQ-Algorithmus

$$\begin{aligned}
\mathbf{W}^{(r)} &= diag(d_1^{(r)}, \ldots, d_n^{(r)}) \\
\eta_i^{(r)} &= \beta_0^{(r)} + f_1^{(r)} + \cdots + f_p^{(r)} \\
\mu_i^{(r)} &= h(\eta_i^{(r)}) \\
\theta_i^{(r)} &= \theta(h(\eta_i^{(r)})) \\
d_i^{(r)} &= \left(\frac{\delta h(\eta_i^{(r)})}{\delta \eta}\right)^2 \cdot \left(\frac{\delta b(\theta_i^{(r)})}{\delta^2 \theta}\right)^{-1} \frac{w_i}{\phi} \\
\tilde{y}_i^{(r)} &= \eta_i^{(r)} + \left(\frac{\delta h(\eta_i^{(r)})}{\delta \eta}\right)^{-1} \cdot (y_i - \mu_i^{(r)}).
\end{aligned}$$

iii) Schätze ein gewichtetes additives Modell mit den Gewichten $d_1^{(r)}, \ldots, d_n^{(r)}$ und den Arbeitsbeobachtungen $\tilde{y}_i^{(r)}$ als abhängige Variable. Erhalte $f_1^{(r+1)}, \ldots, f_p^{(r+1)}, \eta^{(r+1)}, \mu^{(r+1)}$ usw.

iv) Berechne das Konvergenzkriterium

$$\Delta(\eta^{(r+1)}, \eta^{(r)}) = \frac{\sum_{j=1}^{p} ||f_j^{(r+1)} - f_j^{(r)}||}{\sum_{j=1}^{p} ||f_j^{(r)}||}$$

Falls $\Delta(\eta^{(r+1)}, \eta^{(r)})$ kleiner als ε beende den Algorithmus, ansonsten setze $r = r + 1$ und fahre fort mit ii).

Die Glättungsparameter $\lambda = (\lambda_1, \ldots, \lambda_p)'$ können wie beim additiven Modell entweder durch generalisierte Kreuzvalidierung oder durch eine Stepwise-Prozedur bestimmt werden. Für die Kreuzvalidierung muss das GCV Kriterium (4.3) entsprechend modifiziert werden. Die im Zähler von (4.3) vorkommende Residuenquadratsumme wird in generalisierten additiven Modellen durch die sogenante *Devianz* ersetzt:

Sei $l_i(\hat{\mu}_i)$ die Loglikelihood der i–ten Beobachtung in Abhängigkeit vom geschätzen Erwartungswert $\hat{\mu}_i = h(\hat{\eta}_i)$. Die maximal mögliche Loglikelihood erreicht man, wenn $\hat{\mu}_i$ in $l_i(\hat{\mu}_i)$ durch y_i ersetzt wird. Die Devianz D ist definiert als die mit -2 multiplizierte Summe der Abweichungen zwischen der tatsächlich realisierten Loglikelihood $l_i(\hat{\mu}_i)$ und der maximal erreichbaren $l_i(y_i)$:

$$D(y, \hat{\mu}) := -2 \sum_{i=1}^{n} (l_i(\hat{\mu}_i) - l_i(y_i)) \tag{4.18}$$

Je höher die Devianz, desto schlechter ist die Anpassung.

Damit erhalten wir das GCV Kriterium

$$GCV(\lambda) = \frac{1/nD(y, \hat{\mu})}{(1 - sp(\mathbf{R})/n)^2},$$

wobei die Matrix $\mathbf{R}$ durch

$$\hat{\eta} = \mathbf{R}\tilde{y}$$

gegeben ist. Die Optimierung ist wie auch bei additiven Modellen technisch anspruchsvoll, vergleiche Wood (2000). Eine Implementation für Glättungssplines findet man im Softwarepaket R, siehe auch Abschnitt 4.4.

Die Stepwise-Prozedur läuft völlig analog zu additiven Modellen ab. Im AIC Kriterium muss aber wie beim GCV Kriterium die Residuenquadratsumme durch die Devianz ersetzt werden, d.h.

$$AIC = D + 2sp(\mathbf{R})\phi.$$

Beispiel 4.3 (Kreditscoring)
Wir veranschaulichen die Schätzung generalisierter additiver Modelle anhand des Kreditscoringdatensatzes. Eine Beschreibung des Datensatzes findet man in Abschnitt 1.2.4. Da die interessierende Variable lediglich die zwei Werte 1 (nicht kreditwürdig) und 0 (kreditwürdig) annimmt, schätzen wir ein binäres Logitmodell mit zunächst linearem Prädiktor

$$\eta_i = \beta_0 + \beta_1 laufz + \beta_2 hoehe + \beta_3 moral + \cdots.$$

Wir erhalten die folgenden Ergebnisse:

boni	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
laufz	.0350267	.0078491	4.46	0.000	.0196426	.0504107
hoehe	.0324235	.0333474	0.97	0.331	-.0329361	.0977831
moral	-.4941854	.1264752	-3.91	0.000	-.7420722	-.2462985
zweck	-.2371991	.0802339	-2.96	0.003	-.3944547	-.0799436
geschl	.1117566	.1104147	1.01	0.311	-.1046523	.3281656
famst	-.1927104	.1096852	-1.76	0.079	-.4076894	.0222685
ko1	.8621074	.1085855	7.94	0.000	.6492837	1.074931
ko2	-1.089567	.1228731	-8.87	0.000	-1.330394	-.8487402
_cons	-1.227922	.1936844	-6.34	0.000	-1.607537	-.8483077

Ein Blick auf die 95% Konfidenzintervalle zeigt, dass neben dem Geschlecht (geschl) und dem Familienstand (famst)) des Kreditnehmers überraschenderweise auch die Kredithöhe (hoehe) keinen signifikanten Einfluss auf die Bonität des Kreditnehmers besitzt.

In einem zweiten Schritt ersetzen wir die lineare Modellierung der beiden metrischen Variablen Laufzeit und Kredithöhe und schätzen ein generalisiertes (semiparametrisch) additives Modell mit Prädiktor

$$\eta_i = \beta_0 + f_1(laufz) + f_2(hoehe) + \cdots.$$

Abbildung 4.6 zeigt Schätzungen der nichtlinearen Funktionen f_1 und f_2 basierend auf (Pseudo)glättungssplines. Die Glättungsparameter wurden GCV optimal mit dem Programm R bestimmt. Abbildung 4.7 zeigt die alternative Schätzung basierend auf der Stepwise Prozedur von S-plus. Beide Schätzungen sind sehr ähnlich. Auffallend ist der U-förmige Effekt der Kredithöhe. Die Form des Effekts erklärt auch, warum die Kredithöhe in einem linearen Modell keinen Einfluss hatte. In beiden Abbildungen sind neben den Schätzungen auch die jeweiligen partiellen Residuen eingezeichnet. Für eine beliebige Kovariable x_j sind die partiellen Residuen definiert als

$$r_{ij} = f_j(x_{ij}) + \tilde{y}_i - \eta_i.$$

△

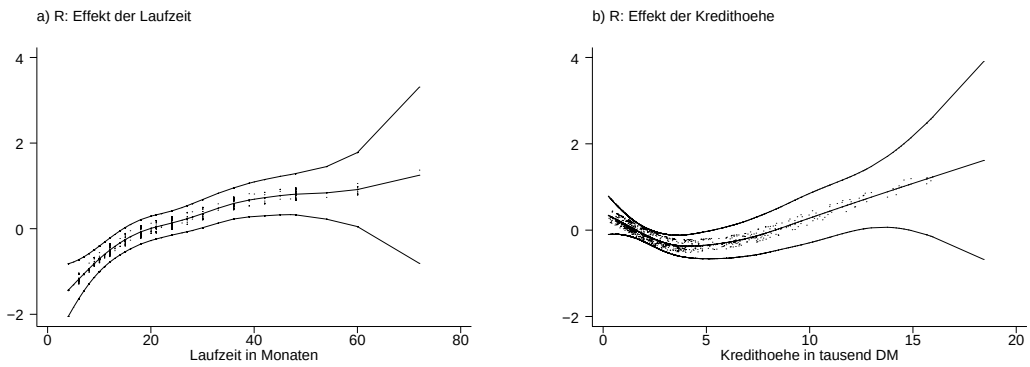

Abbildung 4.6. *Kreditscoringdaten: Geschätzte nichtlineare Effekte der Laufzeit (Abbildung a) und der Kredithöhe (Abbildung b) inklusive 95% punktweise Konfidenzbänder. In die Abbildungen sind zusätzlich noch die jeweiligen partiellen Residuen eingezeichnet. Die Schätzungen basieren auf (Pseudo)glättungssplines. Die Glättungsparameter wurden GCV optimal mit Hilfe des Programms R bestimmt.*

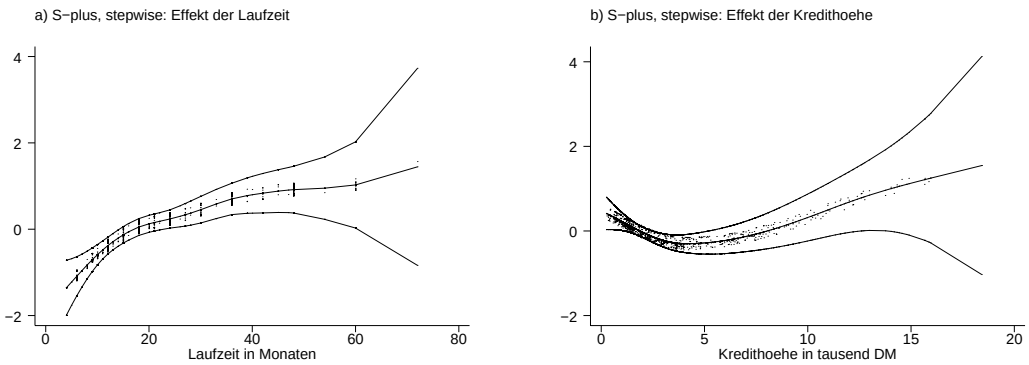

Abbildung 4.7. *Kreditscoringdaten: Geschätzte nichtlineare Effekte der Laufzeit (Abbildung a) und der Kredithöhe (Abbildung b) inklusive 95% punkweise Konfidenzbänder. In die Abbildungen sind zusätzlich noch die jeweiligen partiellen Residuen eingezeichnet. Die Schätzungen basieren auf (Pseudo)glättungssplines. Die Glättungsparameter wurden mit Hilfe der Stepwise-Prozedur des Programms S-plus bestimmt.*

4.4 Software zur Schätzung von GAM's

In diesem Abschnitt geben wir einen Überblick über die vorhandene Software zur Schätzung generalisierter additiver Modelle. Bis vor kurzem konnten GAM's lediglich von einigen wenigen Statistikprogrammen (vor allem S-plus) geschätzt werden. In letzter Zeit wurde diese Lücke aber weitgehend geschlossen, so dass nahezu alle wichtigen Statistikprogrammpakete Funktionen zur Schätzung von GAM's bereitstellen.

Im Folgenden demonstrieren wir die Benutzung der jeweiligen Sofware anhand des Motorcycledatensatzes und der Mietspiegeldaten. Dabei gehen wir davon aus, dass sämtliche Daten im Verzeichnis c:\texte\compstat\software gespeichert sind.

SAS proc gam

– *Installation*
  Die Prozedur gam zur Schätzung generalisierter additiver Modelle ist standardmäßig installiert.

– *Dokumentation*
  Die Funktion gam ist in der Onlinehilfe beschrieben sowie im SAS Handbuch SAS/STAT Software: Changes and Enhancements, Release 8.1.

– *Einlesen der Daten*
  Ascii Datensätze werden in SAS mit Hilfe eines Datasteps eingelesen. Beispielsweise die Motorcycledaten werden mit folgenden Befehlen eingelesen:

```
data mcycle;
infile "c:\texte\compstat\software\mcycle_sas.raw";
input times accel;
run;
```

– *Glättungsparameterwahl*
  Eine automatische Glättungsparameterwahl basierend auf dem GCV Kriterium ist implementiert. Eine genaue Beschreibung der Glättungsparameterwahl fehlt jedoch.

– *Vorhandene Glättungsverfahren*
  Die hier vorgestellte Prozedur 'proc gam' erlaubt lediglich die Schätzung von Glättungssplines wie allgemein in Abschnitt 3.5 beschrieben.

– *Beispiel Scatterplotsmoother: Motorcycledaten*
  Das folgende Beispielprogramm implementiert die Schätzung des Scatterplotsmoothers

und das Visualisieren der geschätzten Funktion. Kommentare zur Erleichterung der Lesbarkeit werden in SAS zwischen die Zeichen /* ... */ gesetzt.

```
/* Schaetzen eines Scatterplotsmoothers zwischen Beschleunigung */
/* und Zeit. Der Glaettungsparameter wird mit GCV geschaetzt */

proc gam data=mcycle;
model accel = spline(times) / method=gcv dist=gaussian;
output out = mcyclefit all;
run;

/* SAS zerlegt den geschaetzten Spline in einen linearen Anteil */
/* und einen nichtlinearen Anteil. Der folgende data step */
/* berechnet geschaetzten Spline */

data mcyclefit;
set mcyclefit;
fit_times = p_times+1.09068*times;
run;

/* Zeichnen des geschaetzten Splines inklusive Intercept */
/* zusammen mit den Daten. p_accel enthaelt also den  */
/* geschaetzten Praediktor */

proc gplot data=mcyclefit;
symbol value=DOT interpol=join;
plot (accel P_accel)* times / overlay;
run;

/* Zeichnen des geschaetzten Splines */

proc gplot data=mcyclefit;
symbol V=DOT interpol=join;
plot (P_times fit_times) * times / overlay;
run;
```

– *Beispiel additives Modell: Mietspiegel für München*

Das folgende Beispiel berechnet für die Mietspiegeldaten das additive Modell

$$nmproqm = \beta_0 + \beta_1 lagegut + \beta_2 lagesgut + f_1(wfl) + f_2(bam) + \varepsilon$$

und visualisiert die Effekte.

```
/* Schaetzen eines additiven Modells mit abhaengiger Variable    */
/* nmproqm und Kovariablen lagegut, lagesgut, wfl und bam.       */
/* Der Einfluss der kategorialen Variablen lagegut und           */
/* lagesgut wird linear modelliert. Die metrischen Variablen    */
/* wfl und bam werden durch Glaettungssplines modelliert. Die    */
/* Output Option all bewirkt, dass alle implementierten Groessen */
/* (Standardfehler, Konfidenzintervalle, etc.) berechnet werden. */

proc gam data=miete;
model nmproqm = param(lagegut lagesgut) spline(wfl) spline(bam)
/ method=gcv dist=gaussian;
output out = mietefit all;
run;

/* SAS zerlegt die geschaetzten Splines in einen linearen Anteil   */
/* und einen nichtlinearen Anteil. Der folgende data step          */
/* berechnet die tatsaechlich geschaetzten Splines fue wfl und bam */

data mietefit;
set mietefit;
fit_wfl = p_wfl-0.06186*wfl;
fit_bam = p_bam+0.07546*bam;
run;

/* Mit Hilfe der Prozedur proc gplot werden die geschaetzten */
/* Funktionen gezeichnet. */

proc gplot data=mietefit;
symbol1 V=DOT interpol=none;
plot (fit_wfl )* wfl  / overlay;
run;
```

```
proc gplot data=mietefit;
symbol V=DOT interpol=none;
plot (fit_bam )* bam /overlay;
run;
```

– *Beispiel gneralisiertes additives Modell: Kreditscoringdaten*

Das folgende Beispiel schätzt ein generalisiertes additives Modell (Logit Modell) mit der Bonität 'boni' als abhängiger Variable und den Kovariablen 'laufz', 'hoehe', 'zweck' und 'moral'.

```
/* Schaetzen eines generalisierten additiven Modells mit    */
/* abhaengiger Variable boni und Kovariablen laufz, hoehe,  */
/* moral und zweck. Der Einfluss der kategorialen Variablen */
/* moral und zweck  wird linear modelliert. Dier metrischen */
/* Variablen laufz und hoehe werden durch (pseudo)          */
/* Glaettungssplines modelliert.                            */

proc gam data=kredit;
model boni = param(moral zweck) spline(laufz) spline(hoehe)
/ method=gcv dist=binomial;
output out = kreditfit all;
run;


/* SAS zerlegt die geschaetzten Splines in einen linearen Anteil   */
/* und einen nichtlinearen Anteil. Der folgende data step          */
/* berechnet die tatsaechlich geschaetzten Splines fur laufz  */
/* und hoehe */

data kreditfit;
set kreditfit;
fit_laufz = p_laufz+0.043633*laufz;
fit_hoehe = p_hoehe-0.02624*hoehe;
run;

/* Mit Hilfe der Prozedur proc gplot werden die geschaetzten */
```

```
/* Funktionen gezeichnet. */

 proc gplot data=kreditfit;
symbol1 V=DOT interpol=none;
plot (fit_laufz )* laufz  / overlay;
run;

proc gplot data=kreditfit;
symbol1 V=DOT interpol=none;
plot (fit_hoehe )* hoehe  / overlay;
run;
```

STATA

– *Installation*

Standardmäßig können in STATA keine GAM's geschätzt werden. Jedoch können die ado-files (Funktionen) 'gam' und 'gamplot' von der Homepage der Vorlesung heruntergeladen werden (`http://www.stat.uni-muenchen.de/~lang/compstat/mat/gam.zip`). Zur Installation werden die Dateien (z.B.) im Verzeichnis c:\texte\compstat\software entpackt und STATA gestartet. Anschließend muss STATA mitgeteilt werden, in welchem Verzeichnis sich die Dateien befinden. Mit dem Befehl

> sysdir

erhält man Informationen, in welchen Verzeichnissen STATA ado-files/Funktionen sucht. Um beispielsweise das persönliche ado-Verzeichnis zu ändern schreiben wir:

> sysdir set PERSONAL c:\texte\compstat\software

Zuletzt schreiben wir

> global GAMDIR c:\texte\compstat\software\

um STATA mitzuteilen, wo die beiden externen Programme 'gamfit.exe' und 'gambit.exe' zu finden sind, die bei der Verwendung der GAM Funktionen benötigt werden.

– *Dokumentation*

Aufruf von

> help gam

bzw.

> help gamplot

in STATA.

– *Einlesen der Daten*
ASCII Datensätze werden mit dem *infile* Befehl eingelesen. Dieser Befehl besitzt folgende allgemeine Struktur:

infile varlist using myfile

Dabei wird von STATA angenommen, dass die Variablen (in *varlist*) in der ASCII-Datei spaltenweise angeordnet sind. Die Spezifizierung einer Variable in *varlist* hat folgende allgemeine Syntax:

[type] newvarname[:labelname]

Als Variablentypen sind ganze Zahlen (byte, int und long), reelle Zahlen (float und double) und Strings (str1 - str80) zugelassen.

Beispielsweise werden mit dem folgenden Befehl die Motorcycledaten in Stata eingelesen:

infile times accel using c:\texte\compstat\software\mcycle.raw

Nach dem Einlesen der Daten können die Variablen im STATA-Format (Dateiendung dta) durch Anklicken des Menüpunktes *File–SaveAs* abgespeichert werden. Durch Öffnen des Datenbrowsers oder des Dateneditors (*Window–Data Editor*) können die Daten visualisiert werden. Der Editor erlaubt auch das Editieren (Verändern) der Daten.

– *Glättungsparameterwahl*
Eine automatische Glättungsparameterwahl ist nicht implementiert.

– *Vorhandene Glättungsverfahren*
Die hier vorgestellte Funktion 'gam' erlaubt lediglich die Schätzung von Glättungssplines wie allgemein in Abschnitt 3.5 beschrieben. Aus numerischen Gründen werden aber lediglich die Pseudoglättungssplines aus Abschnitt 3.5.4 verwendet.

– *Beispiel Scatterplotsmoother: Motorcycledaten*
Das folgende Beispielprogramm implementiert die Schätzung des Scatterplotsmoothers und das Visualisieren der geschätzten Funktion. Kommentare zur Erleichterung der Lesbarkeit werden in STATA mit dem Zeichen * gekennzeichnet.

```
#delimit ;

* Schaetzen eines Scatterplotsmoothers mit abhaengiger
```

```
* Variable accel und erklaerender Variable times. Die
* Anzahl der aequivalenten Freiheitsgrade fuer die
* nichtparametrische Funktion ist 10 (insgesamt im
* Modell 11, da ein Intercept mitgeschaetzt wird).

gam accel times , family(gaussian) link(identity) df(10) big;


* Zeichnen des geschaetzten um Null zentrierten Splines
* inklusive Standardfehlerbaender (95%) und partieller
* Residuen. Abspeichern der Grafik als ps-file.

set textsize 120;
translator set Graph2eps scheme custom1;
gamplot times , t1title("Titel") b2title("Zeit in Millisekunden")
xlab ylab;
translate @Graph c:\texte\compstat\grafiken\mc_gam_stata1.eps ,
translator(Graph2eps) replace;

* Zeichnen des geschaetzten Splines, der Standardfehlerbaender
* und der Daten mit der graph Funktion als Alternative zu
* gamplot. Abspeichern der Grafik als ps-file.

set textsize 120;
translator set Graph2eps scheme custom1;
graph mu accel times , connect(l.) s(.o)  t1title("Titel")
b2title("Zeit in Millisekunden") xlab ylab;
translate @Graph c:\texte\compstat\grafiken\mc_gam_stata2.eps ,
translator(Graph2eps) replace;
```

– *Beispiel additives Modell: Mietspiegel für München*

Das folgende Beispiel berechnet für die Mietspiegeldaten das additive Modell

$$nmproqm = \beta_0 + \beta_1 lagegut + \beta_2 lagesgut + f_1(wfl) + f_2(bam) + \varepsilon$$

und visualisiert die Effekte.

```
#delimit ;
```

```
* Schaetzen eines additiven Modells mit abhaengiger Variable
* nmproqm und Kovariablen lagegut, lagesgut, wfl und bam.
* Der Einfluss der kategorialen Variablen lagegut und
* lagesgut wird linear modelliert. Dier metrischen Variablen
* wfl und bam werden durch (pseudo) Glaettungssplines mit
* 5 aequivalenten Freiheitsgraden modelliert.

gam nmproqm lagegut lagesgut wfl bam , family(gaussian) link(identity)
df(lagegut lagesgut: 1,wfl bam: 5) big;

* Zeichnen der geschaetzten um Null zentrierten Splines
* inklusive Standardfehlerbaender (95%) und partieller
* Residuen. Abspeichern der Grafiken als ps-file.

set textsize 120;
translator set Graph2eps scheme custom1;
gamplot wfl , t1title("Titel") b2title("Wohnflaeche in qm")
xlab ylab;
translate @Graph c:\texte\compstat\grafiken\miete_gam_stata_wfl.eps ,
translator(Graph2eps) replace;

gamplot bam , t1title("Titel") b2title("Baujahr")
xlab ylab;
translate @Graph c:\texte\compstat\grafiken\miete_gam_stata_bam.eps ,
translator(Graph2eps) replace;

* Zeichnen der geschaetzten Splines, der Standardfehlerbaender
* und der Daten mit der graph Funktion als Alternative zu
* gamplot. Abspeichern der Grafiken als ps-file.

sort wfl;
generate wfl_ki_o = s_wfl+1.96*e_wfl;
generate wfl_ki_u = s_wfl-1.96*e_wfl;
graph s_wfl wfl_ki_o wfl_ki_u wfl , connect(lll) s(...)  t1title("Titel")
b2title("Wohnflaeche in qm") xlab ylab;
translate @Graph c:\texte\compstat\grafiken\miete_gam_stata2_wfl.eps ,
```

```
translator(Graph2eps) replace;

sort bam;
generate bam_ki_o = s_bam+1.96*e_bam;
generate bam_ki_u = s_bam-1.96*e_bam;
graph s_bam bam_ki_o bam_ki_u bam , connect(lll) s(...)  t1title("Titel")
b2title("Baujahr") xlab ylab;
translate @Graph c:\texte\compstat\grafiken\miete_gam_stata2_bam.eps ,
translator(Graph2eps) replace;
```

– *Beispiel generalisiertes additives Modell: Kreditscoringdaten*
Das folgende Beispiel schätzt ein generalisiertes additives Modell (Logit Modell) mit der Bonität 'boni' als abhängiger Variable und den Kovariablen 'laufz', 'hoehe', 'zweck' und 'moral'.

```
#delimit ;

* Schaetzen eines generalisierten additiven Modells mit
* abhaengiger Variable boni und Kovariablen laufz, hoehe,
* moral und zweck. Der Einfluss der kategorialen Variablen
* moral und zweck  wird linear modelliert. Dier metrischen
* Variablen laufz und hoehe werden durch (pseudo) Glaettungs-
* splines mit 5 aequivalenten Freiheitsgraden modelliert.

gam boni moral zweck laufz hoehe , family(binomial) link(logit)
df(moral zweck: 1,laufz hoehe: 5) big;

* Zeichnen der geschaetzten um Null zentrierten Splines
* inklusive Standardfehlerbaender (95%) und partieller
* Residuen. Abspeichern der Grafiken als ps-file.

set textsize 120;
translator set Graph2eps scheme custom1;
gamplot laufz , t1title("Titel") b2title("Laufzeit in Monaten")
xlab ylab;
translate @Graph c:\texte\compstat\grafiken\kredit_gam_stata_laufz.eps ,
translator(Graph2eps) replace;
```

```
gamplot hoehe , t1title("Titel") b2title("Kredithoehe")
xlab ylab;
translate @Graph c:\texte\compstat\grafiken\kredit_gam_stata_hoehe.eps ,
translator(Graph2eps) replace;

* Zeichnen der geschaetzten Splines, der Standardfehlerbaender
* und der Daten mit der graph Funktion als Alternative zu
* gamplot. Abspeichern der Grafiken als ps-file.

sort laufz;
generate laufz_ki_o = s_laufz+1.96*e_laufz;
generate laufz_ki_u = s_laufz-1.96*e_laufz;
graph s_laufz laufz_ki_o laufz_ki_u laufz , connect(lll) s(...)
t1title("Titel") b2title("Laufzeit in Monaten") xlab ylab;
translate @Graph c:\texte\compstat\grafiken\kredit_gam_stata2_laufz.eps ,
translator(Graph2eps) replace;

sort hoehe;
generate hoehe_ki_o = s_hoehe+1.96*e_hoehe;
generate hoehe_ki_u = s_hoehe-1.96*e_hoehe;
graph s_hoehe hoehe_ki_o hoehe_ki_u hoehe , connect(lll) s(...)
t1title("Titel") b2title("Kredithoehe") xlab ylab;
translate @Graph c:\texte\compstat\grafiken\kredit_gam_stata2_hoehe.eps ,
translator(Graph2eps) replace;
```

R

– *Installation*
Die GAM Routinen sind standardmäßig installiert, müssen aber geladen werden. Dies geschieht durch Anklicken des Menüs *Packages–Load Package–mgcv*.

– *Dokumentation*
Die Dokumentation mgcv-manual.pdf kann von der Homepage der Vorlesung heruntergeladen werden
(`http://www.stat.uni-muenchen.de/~lang/compstat/mat/mgcv-manual.pdf`).

– *Einlesen der Daten*
Daten können in R (wie in S-plus) mit dem Befehl 'read.table' eingelesen werden. Die Motorcycledaten können beispielsweise mit

```
mcycle<-read.table("c:\\texte\\compstat\\software\\mcycle.raw",header=T)
```

eingelesen werden. Falls die erste Zeile des Datensatzes keine Variablennamen enthält, so muss in obigem Befehl 'header=F' gesetzt werden.

– *Glättungsparameterwahl*
Die Glättungsparameterwahl erfolgt mit generalisierter Kreuzvalidierung (GCV) wie in Abschnitt 3.4.4 beschrieben. Die numerische Berechnung basiert auf (relativ komplizierten) Verfahren, welche in Wood (2000) beschrieben sind.

– *Vorhandene Glättungsverfahren*
Die Funktion 'gam' erlaubt lediglich die Schätzung von Glättungssplines wie allgemein in Abschnitt 3.5 beschrieben. Aus numerischen Gründen werden aber lediglich die Pseudoglättungssplines aus Abschnitt 3.5.4 verwendet.

– *Beispiel Scatterplotsmoother: Motorcycledaten*
Das folgende Beispielprogramm implementiert die Schätzung des Scatterplotsmoothers und das Visualisieren der geschätzten Funktion. Kommentare zur Erleichterung der Lesbarkeit werden in R mit dem Zeichen # gekennzeichnet. Das Programm kann in jedem beliebigen Editor erstellt werden und mit source("dateiname") in R gestartet werden.

```
# Schaetzen eines Scatterplotsmoothers zwischen accel und times.
# Dabei werden k=20 Knoten verwendet, als Splinebasis werden
# B-Splines fuer NKS (bs="cr", alternativ bs="ts" Thin Plate Splines)

mcyclefit_gam(accel~s(times,k=20,bs="cr"),family=gaussian(),data=mcycle)

# Zeichnen des geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender

plot.gam(mcyclefit,se=T)

# Zeichnen des geschaetzten Splines (plus Intercept) inklusive der Daten.
```

```
plot(mcycle$times,mcyclefit$linear.predictor,type="l")
lines(mcycle$times,mcycle$accel,type="p")
```

– *Beispiel additives Modell: Mietspiegel für München*

Das folgende Beispiel berechnet für die Mietspiegeldaten das additive Modell

$$nmproqm = \beta_0 + \beta_1 lagegut + \beta_2 lagesgut + f_1(wfl) + f_2(bam) + \varepsilon$$

und visualisiert die Effekte.

```
# Schaetzen eines additiven Modells mit abhaengiger Variable
# nmproqm und Kovariablen lagegut, lagesgut, wfl und bam.
# Der Einfluss der kategorialen Variablen lagegut und
# lagesgut wird linear modelliert. Die metrischen Variablen
# wfl und bam werden durch (pseudo) Glaettungssplines modelliert.

mietefit_gam(nmproqm~lagegut+lagesgut+s(wfl,k=20,bs="cr")+
s(bam,k=20,bs="cr"),family=gaussian(),data=miete)

# Zeichnen des geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender

plot.gam(mietefit,se=T)
```

– *Beispiel gneralisiertes additives Modell: Kreditscoringdaten*

Das folgende Beispiel schätzt ein generalisiertes additives Modell (Logit Modell) mit der Bonität 'boni' als abhängiger Variable und den Kovariablen 'laufz', 'hoehe', 'zweck' und 'moral'.

```
# Schaetzen eines generalisierten additiven Modells mit
# abhaengiger Variable boni und Kovariablen laufz, hoehe,
# moral und zweck. Der Einfluss der kategorialen Variablen
# moral und zweck  wird linear modelliert. Die metrischen
# Variablen laufz und hoehe werden durch (pseudo)
# Glaettungssplines modelliert.

kreditfit_gam(boni~moral+zweck+s(laufz,k=20,bs="cr")+
s(hoehe,k=20,bs="cr"),family=binomial(),data=kredit)
```

```
# Zeichnen der geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender

plot.gam(kreditfit,se=T)
```

S-plus Funktionen gam und step.gam

– *Installation*
Die Funktionen gam und step.gam sind standardmäßig in S-plus installiert.

– *Dokumentation*
Die Funktionen sind in der S-plus Onlinehilfe, in den Handbüchern und in Chambers and Hastie (1992) Kapitel 7 beschrieben.

– *Einlesen der Daten*
Daten können in S-plus und R mit dem Befehl 'read.table' eingelesen werden. Die Motorcycledaten können beispielsweise mit

```
mcycle<-read.table("c:\\texte\\compstat\\software\\mcycle.raw",header=T)
```

eingelesen werden. Falls die erste Zeile des Datensatzes keine Variablennamen enthält, so muss in obigem Befehl 'header=F' gesetzt werden.

– *Glättungsparameterwahl*
Die Glättungsparameterwahl erfolgt durch die Funktion 'step.gam', die ähnlich funktioniert wie Variablenselektionsalgorithmen. Vergleiche auch Abschnitt 3.4.4.

– *Vorhandene Glättungsverfahren*
In der S-plus Funktion 'gam' sind standardmäßig (Pseudo) Glättungssplines (vergleiche die Abschnitte 3.5 und 3.5.4) und loess (vergleiche Abschnitt 3.6.3) implementiert. P-splines (Kapitel 3.4) sind nicht standardmäßig vorhanden, die entsprechende Funktion 'ps' (von Brian Marx) kann aber von der Homepage der Vorlesung heruntergeladen werden.

– *Beispiel Scatterplotsmoother: Motorcycledaten*
Das folgende Beispielprogramm implementiert die Schätzung des Scatterplotsmoothers und das Visualisieren der geschätzten Funktion. Kommentare zur Erleichterung der Lesbarkeit werden in S-plus mit dem Zeichen # gekennzeichnet. Das Programm kann in jedem beliebigen Editor erstellt werden und mit source("dateiname") in S-plus gestartet werden.

```
# Schaetzen eines Scatterplotsmoothers zwischen accel und times
# unter Verwendung von Glaettungssplines. Der Glaettungsparameter
# wird mit step.gam gewaehlt. Zur Auswahl stehen 1-15 aequivalente
# Freiheitsgrade sowie die Nichtaufnahme der Kovariable (d.h. das
# Modell besteht lediglich aus einem Intercept.

mcycleest.scope<-list(
    "times" = ~ 1+
                    times+
                    s(times,df=2) +
                    s(times,df=3) +
                    s(times,df=4) +
                    s(times,df=5) +
                    s(times,df=6) +
                    s(times,df=7) +
                    s(times,df=9) +
                    s(times,df=11) +
                    s(times,df=13) +
                    s(times,df=15))

mcycleest.start<-gam(accel~times,family=gaussian,data=mcycle,model=T)

mcycleest.step<-step.gam(mcycleest.start,mcycleest.scope,
trace=T,steps=500)

# Zeichnen des geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender

plot.gam(mcycleest.step,se=T)

# Zeichnen des geschaetzten Splines (plus Intercept) inklusive der Daten.

plot(mcycle$times,mcycleest.step$fitted.values,type="l")
lines(mcycle$times,mcycle$accel,type="p")
```

Im folgenden Beispiel wurden die Glättungssplines durch loess ersetzt:

```
mcycleest.scope<-list(
    "times" = ~ 1+
                    times+
                    lo(times,span=0.9) +
                    lo(times,span=0.8) +
                    lo(times,span=0.7) +
                    lo(times,span=0.6) +
                    lo(times,span=0.5) +
                    lo(times,span=0.4) +
                    lo(times,span=0.3) +
                    lo(times,span=0.2) +
                    lo(times,span0.1))

mcycleest.start<-gam(accel~times,family=gaussian,data=mcycle,model=T)

mcycleest.step<-step.gam(mcycleest.start,mcycleest.scope,
trace=T,steps=500)
```

– *Beispiel additives Modell: Mietspiegel für München*

Das folgende Beispiel berechnet für die Mietspiegeldaten das additive Modell

$$nmproqm = \beta_0 + \beta_1 lagegut + \beta_2 lagesgut + f_1(wfl) + f_2(bam) + \varepsilon$$

und visualisiert die Effekte.

```
# Schaetzen eines additiven Modells mit abhaengiger Variable
# nmproqm und Kovariablen lagegut, lagesgut, wfl und bam.
# Der Einfluss der kategorialen Variablen lagegut und
# lagesgut wird linear modelliert. Die metrischen Variablen
# wfl und bam werden durch (pseudo) Glaettungssplines modelliert.
# Die Glaettungsparameter werden mit step.gam bestimmt.

mieteest.scope<-list(
        "lagegut" = ~1 + lagegut,
        "lagesgut" = ~1 + lagesgut,
    "wfl" = ~ 1+
                    wfl+
                    s(wfl,df=2) +
```

```
                    s(wfl,df=3) +
                    s(wfl,df=4) +
                    s(wfl,df=5) +
                    s(wfl,df=6) +
                    s(wfl,df=7) +
                    s(wfl,df=8),
    "bam" = ~ 1+
                    bam+
                    s(bam,df=2) +
                    s(bam,df=3) +
                    s(bam,df=4) +
                    s(bam,df=5) +
                    s(bam,df=6) +
                    s(bam,df=7) +
                    s(bam,df=8)
                    )

mieteest.start<-gam(nmproqm~lagegut+lagesgut+wfl+bam,
family=gaussian,data=miete,model=T)

mieteest.step<-step.gam(mieteest.start,mieteest.scope,
trace=T,steps=500)

# Zeichnen der geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender

plot.gam(mieteest.step,se=T)
```

– *Beispiel generalisiertes additives Modell: Kreditscoringdaten*

Das folgende Beispiel schätzt ein generalisiertes additives Modell (Logit Modell) mit der Bonität 'boni' als abhängiger Variable und den Kovariablen 'laufz', 'hoehe', 'zweck' und 'moral'.

```
# Schaetzen eines generalisierten additiven Modells mit
# abhaengiger Variable boni und Kovariablen laufz, hoehe,
# moral und zweck. Der Einfluss der kategorialen Variablen
# moral und zweck  wird linear modelliert. Die metrischen
```

```
# laufz und hoehe werden durch (pseudo) Glaettungssplines
# modelliert. Die Glaettungsparameter werden mit step.gam
# bestimmt.

kreditest.scope<-list(
        "moral" = ~1 + moral,
        "zweck" = ~1 + zweck,
    "laufz" = ~ 1+
                    laufz+
                    s(laufz,df=2) +
                    s(laufz,df=3) +
                    s(laufz,df=4) +
                    s(laufz,df=5) +
                    s(laufz,df=6) +
                    s(laufz,df=7) +
                    s(laufz,df=8),
    "hoehe" = ~ 1+
                    hoehe+
                    s(hoehe,df=2) +
                    s(hoehe,df=3) +
                    s(hoehe,df=4) +
                    s(hoehe,df=5) +
                    s(hoehe,df=6) +
                    s(hoehe,df=7) +
                    s(hoehe,df=8)
                    )


kreditest.start<-gam(boni~moral+zweck+laufz+hoehe,family=binomial,
data=kredit,model=T)

kreditest.step<-step.gam(kreditest.start,kreditest.scope,
trace=T,steps=500)

# Zeichnen der geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender
```

```
plot.gam(kreditest.step,se=T)
```

S-Plus/R Funktion ggamm (von Thomas Kneib)

– *Installation*

Die S-plus/R Funktion `ggamm` und einige Hilfsfunktionen in `helpfunctions` findet man auf der Homepage der Vorlesung unter
`http://www.stat.uni-muenchen.de/~lang/compstat/mat/ggamm.zip`.

Die S-plus Implementation besteht aus den Dateien `ggamm.s`, `ggammc.s`, `mat.dll` und `helpfunctions.s`. Während die beiden Dateien `ggamm.s` und `ggammc.s` unterschiedliche Versionen der Funktion `ggamm` zur Schätzung beinhalten, werden in der Datei `helpfunctions.s` eine Reihe von Hilfsfunktionen definiert. Um die Geschwindigkeit, mit der das Programm ausgeführt wird zu erhöhen, werden in der in `ggammc.s` enthaltenen Implementation einige Berechnungen nicht mit Hilfe von in S-Plus programmierten Funktionen, sondern durch in der Programmiersprache C abgefasste Funktionen durchgeführt. Diese C-Funktionen sind in der Datei `mat.dll` enthalten.

Um die Funktionen in S-Plus zu definieren, müssen zunächst durch die Ausführung des Kommandos

```
> source("c:\\texte\\compstat\\software\\helpfunctions.s")
```

die Hilfsfunktionen installiert werden. Eventuell ist dabei noch zusätzlich der Pfad zu dem Verzeichnis, in dem sich diese Datei befindet, entsprechend zu verändern. Man beachte, dass durch den Aufruf des `source`-Kommandos auch die in `mat.dll` enthaltenen Funktionen in S-Plus eingelesen werden. Damit dies möglich ist, muss in der Datei `helpfunctions.s` der korrekte Pfad, unter dem `mat.dll` zu finden ist, in dem Aufruf

```
> dll.load("c:\\texte\\compstat\\software\\mat.dll",...
```

eingesetzt werden. Man beachte außerdem, dass die in C programmierten Funktionen nur temporär in S-Plus definiert werden. Das heißt, nach Beendigung von S-Plus sind die zugehörigen Funktionen beim nächsten Programmaufruf nicht mehr verfügbar und müssen wieder über den obigen source-Befehl definiert werden.

Je nachdem, welche der beiden `ggamm`-Implementationen gewählt wird, ist dann zusätzlich noch der Befehl

```
> source("c:\\texte\\compstat\\software\\ggamm.s")
```

beziehungsweise

```
> source("c:\\texte\\compstat\\software\\ggammc.s")
```

in S-plus auszuführen.

In R existiert lediglich eine Version von `ggamm`. Die R Implementation umfasst die Dateien `ggamm.r` und `helpfunctions.r`. Hier wird `ggamm` durch Aufruf der Befehle

```
> source("c:\\texte\\compstat\\software\\helpfunctions.r")
```

und

```
> source("c:\\texte\\compstat\\software\\ggamm.r")
```

installiert.

– *Dokumentation*
Eine ausführliche Dokumentation der Software findet man auf der Homepage der Vorlesung unter `http://www.stat.uni-muenchen.de/~lang/compstat/mat/ggamm.ps`. Die zugrundeliegende Methodik ist ausführlich in Kneib (2003) und kürzer in Fahrmeir et al. (2003) beschrieben.

– *Einlesen der Daten*
Daten können in S-plus und R mit dem Befehl 'read.table' eingelesen werden. Die Motorcycledaten können beispielsweise mit

```
mcycle<-read.table("c:\\texte\\compstat\\software\\mcycle.raw",header=T)
```

eingelesen werden. Falls die erste Zeile des Datensatzes keine Variablennamen enthält, so muss in obigem Befehl 'header=F' gesetzt werden.

– *Glättungsparameterwahl*
Die Glättungsparameter können simultan mitgeschätzt werden, indem das nonparametrische Regressionsmodell durch Parametertransformation in ein generalisiertes lineares gemischtes Modell überführt wird. Anschließend werden Standardschätzverfahren für generalisierte lineare gemischte Modelle angewandt.

– *Vorhandene Glättungsverfahren*
Die hier vorgestellte Funktion 'ggamm' erlaubt die Schätzung von P-splines wie allgemein in Abschnitt 3.4 beschrieben. Darüberhinaus können auch räumliche Glätter und zufällige Effekte im Modell verwendet werden, die aber nicht im Rahmen der Vorlesung behandelt werden. Eine ausführliche Beschreibung findet man in Kneib (2003).

– *Beispiel Scatterplotsmoother: Motorcycledaten*

Das folgende Beispielprogramm implementiert die Schätzung des Scatterplotsmoothers und das Visualisieren der geschätzten Funktion. Kommentare zur Erleichterung der Lesbarkeit werden in R mit dem Zeichen # gekennzeichnet. Das Programm kann in jedem beliebigen Editor erstellt werden und mit source("dateiname") in R gestartet werden.

```
# Schaetzen eines Scatterplotsmoothers zwischen accel und times.
# Dabei werden kubische P-splines mit nknot=20 Knoten und Straf-
# termen basierend auf 2. Differenzen verwendet.

mcyclefit<-ggamm(dep=mcycle$accel,smooth=mcycle$times,
nknot=20,ord=2,deg=3,family="normal",dispers=T)

# Zeichnen des geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender

plotf(mcyclefit)

# Zeichnen des geschaetzten Splines (plus Intercept) inklusive der Daten.

plot(mcycle$times,mcyclefit$predict[,1],type="l")
lines(mcycle$times,mcycle$accel,type="p")
```

– *Beispiel additives Modell: Mietspiegel für München*

Das folgende Beispiel berechnet für die Mietspiegeldaten das additive Modell

$$nmproqm = \beta_0 + \beta_1 lagegut + \beta_2 lagesgut + f_1(wfl) + f_2(bam) + \varepsilon$$

und visualisiert die Effekte.

```
# Schaetzen eines additiven Modells mit abhaengiger Variable
# nmproqm und Kovariablen lagegut, lagesgut, wfl und bam.
# Der Einfluss der kategorialen Variablen lagegut und
# lagesgut wird linear modelliert. Die metrischen Variablen
# wfl und bam werden durch kubische P-splines und 20 Knoten
# modelliert. Die Strafterme basieren auf Differenzen 2.
# Ordnung.
```

```
mietefit<-ggamm(dep=miete$nmproqm,smooth=cbind(miete$wfl,miete$bam),
fix=cbind(miete$lagegut,miete$lagesgut),nknot=cbind(20,20),
ord=cbind(2,2),deg=cbind(3,3),family="normal",dispers=T)

# Zeichnen der geschaetzten Splines (um Null zentriert) inklusive
# der Standardfehlerbaender

plotf(mietefit)
```

– *Beispiel generalisiertes additives Modell: Kreditscoringdaten*
Das folgende Beispiel schätzt ein generalisiertes additives Modell (Logit Modell) mit der Bonität 'boni' als abhängiger Variable und den Kovariablen 'laufz', 'hoehe', 'zweck' und 'moral'.

```
# Schaetzen eines generalisierten additiven Modells mit
# abhaengiger Variable boni und Kovariablen laufz, hoehe,
# moral und zweck. Der Einfluss der kategorialen Variablen
# moral und zweck  wird linear modelliert. Die metrischen
# Variablen moral und zweck werden durch kubische P-splines
# mit 20 Knoten modelliert. Die Strafterme basieren auf
Differenzen 2. Ordnung.

kreditfit<-ggamm(dep=kredit$boni,
smooth=cbind(kredit$laufz,kredit$hoehe),
fix=cbind(kredit$moral,kredit$zweck),nknot=cbind(20,20),
ord=cbind(2,2),deg=cbind(3,3),family="binomial",
link="logit")

# Zeichnen der geschaetzten Splines (um Null zentriert) inklusive der
# Standardfehlerbaender

plotf(kreditfit)
```

Literaturverzeichnis

Böhme, R. und Lang, S., 1997: Skript lineare Modelle.

Statistical Models in S. Wadsworth and Brooks, Pacific Grove.

Cleveland, W. S., Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74, 829-836.

De Boor, C., 1978: *A Practical Guide to Splines* Springer–Verlag, New York.

Eilers, P.H.C. and Marx, B.D., 1996: Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, 11 (2), 89-121.

Fahrmeir, L., Hamerle, A. und Tutz, G., 1995: *Multivariate statistische Verfahren.* de Gruyter, Berlin - New York.

Fahrmeir, L., Kneib, Th. and Lang, S., 2003: Penalized additive regression for space-time data: a Bayesian perspective Discussion paper 305, SFB 386, Universität München.

Fahrmeir, L. und Tutz, G., 2001: *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer–Verlag, New York.

Gänssler, P. und Stute, W. 1977: *Wahrscheinlichkeitstheorie.* Springer-Verlag.

George, A. and Liu, J.W.H., 1981: *Computer Solution of Large Sparse Positive Definite Systems.* Prentice-Hall, London.

Green, P. J. und Silverman, B.W., 1994: *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Hämmerlin G. und Hoffman, K.H., 1990: *Numerische Mathematik.* Springer-Verlag, Berlin.

Härdle, W., 1990: *Smoothing Techniques.* Springer Verlag, New York.

Hastie T. und Tibshirani R., 1990: *Generalized Additive Models.* Chapman and Hall, London.

Kneib, Th., 2003: Bayes Inferenz in generalisierten additiven gemischten Modellen. Diplomarbeit, Institut für Statistik, Universität München.

Mc Cullagh, P. und Nelder, J.A., 1989: *Generalized Linear Models.* New York: Chapman and Hall.

Parzen, E. 1962: On estimation of a probability density function and mode. *Annals Mathematical Statistics*, 33, 1065-1076.

Pruscha, H., 2000: Vorlesungen ber Mathematische Statistik (Kapitel 8). Teubner, Stuttgart.

Weisberg, S., 1985: *Applied Linear Regression Analysis.* Wiley, New York.

Wood, S., N. 2000: Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society B*, 62, 413-428.