Nichtparametrische, Semiparametrische und SUR-Modelle für stetige Longitudinaldaten

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik

der Ludwig–Maximilians–Universität München

zur Erlangung des Grades Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt von

Michael Meinel

eingereicht am 29.09.2005

1. Berichterstatter:	PD Dr. Christian Heumann, LMU München
2. Berichterstatter:	Prof. Dr. Helmut Küchenhoff, LMU München
3. Berichterstatter:	Prof. Dr. Peter Bühlmann, ETH Zürich
Prüfungsdatum:	12. Januar 2006

Inhaltsverzeichnis

Vorwort

Diese Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Statistik der Ludwig-Maximilian-Universität in München (2001, 2002) und später als Doktorand an der Eidgenössischen Technischen Hochschule Zürich (ETH) in den Jahren 2004 und 2005.

An dieser Stelle möchte ich mich bei allen bedanken, die mich während des Entstehens dieser Doktorarbeit unterstützt haben. Zunächst sei Prof. Dr. Tutz erwähnt, der zu Beginn der Arbeit die Betreuung übernahm. Nach meiner Rückkehr aus Australien und dem Umzug nach Zürich wurde ich von PD Dr. Christian Heumann betreut, der maßgeblich an der schnellen Fertigstellung der Arbeit und auch des zugehörigen Artikels beteiligt war.

Ein ganz besonderer Dank gilt Prof. Dr. Frey von der ETH Zürich, der mir die Möglichkeit gab, die Doktorarbeit während meiner Tätigkeit am Institut für Verhaltensforschung in Zürich zu beenden, und auch PD Dr. Joachim Fischer, der mir vor allem bei medizinischen Fragestellungen unter die Arme griff und die Cortisol-Datensätze, die Basis meines Artikels waren, zur Verfügung stellte. Desweiteren seien meine früherer Kollegen Dr. Stefan 'Puizi' Pilz und Dr. Thomas Nittner lobend erwähnt, die bei statistischen Problemen und bei der Suche nach geeigneter Literatur zu jeder Zeit ansprechbar waren und deren Kommentare und Anmerkungen sehr hilfreich waren bei der Fertigstellung meiner Arbeit und dem zugehörigen Artikel.

Kapitel 1

Einleitung

Die Grundlagen dieser Arbeit wurden in den Jahren 2001 und 2002 während meiner Tätigkeit am Institut für Statistik in München gelegt. Nach einem zusätzlichen Studienjahr in Sydney, Australien, wurde die Arbeit als Doktorand an der ETH Zürich im Jahr 2004 fortgesetzt und im Folgejahr 2005 fertiggestellt.

Die Arbeit beschäftigt sich hauptsächlich mit (Pseudo)-Likelihood-basierten Schätz-Methoden zur Analyse von marginalen Regressions-Modellen für Longitudinaldaten mit stetigem Response. Es werden drei neuartige Modelle entwickelt, die aufeinander aufbauen. Zunächst wird ein Regressions-Modell entwickelt, das neben einem nichtparametrischen Term auch multiplikative zeitabhängige Parameter und Varying-Coefficients-Terme beinhaltet. Im Anschluss wird das Modell um einen parametrischen Term erweitert, so dass ein semiparametrisches Modell entsteht. Desweiteren wird das Modell auf den Bereich der 'Seemingly-Unrelated-Regression'-Modelle (SUR) ausgeweitet. Das SUR-Modell stellt somit das allgemeinste Modell dar, aus dem das nicht- und semiparametrische und auch andere Modelle bei geeigneter Wahl der Parameter abgeleitet werden können.

Basis der Analysen sind die oft zitierten Artikel über IEE und GEE-Schätzer von Liang und Zeger (1986) und Zeger und Liang (1986). Darin werden anwenderfreundli-

che, relativ unkomplizierte, (Pseudo)-Likelihood-basierte Methoden zur Analyse von zeit- oder clusterkorrelierten Daten präsentiert, in die vergleichsweise wenige Modellannahmen einfließen. Die Korrelationen der Cluster oder Messwiederholungen gehen über eine 'Working-Kovarianz' in das Modell ein, die als 'Nuisance'-Parameter betrachtet wird. Hauptaugenmerk wird auf die korrekte Schätzung der Parameter des Erwartungsmodells und nicht auf die korrekte Schätzung der Working-Kovarianz gelegt, obwohl es von Vorteil ist, wenn auch diese korrekt spezifiziert ist. Methoden zur Schätzung der Working-Kovarianz sind unter der Bezeichnung 'Momentenmethode' bei Liang und Zeger (1986) und bei Aerts et al. (2002) detailliert beschrieben. Die GEE-Methoden werden in dieser Arbeit auf nichtparametrische Regressionsansätze angewendet. B-Splines, so wie sie von de Boor (1978) definiert wurden, dienen als Basis für die nichtparametrischen Schätzungen. Die Penalisierung der nichtparametrischen Schätzungen erfolgt über Differenzenpenalties. Diese wurden in Kombination mit nichtparametrischer Regression von Eilers und Marx (1996) unter dem Namen 'P-Splines' vorgestellt. Im zugehörigen Artikel befinden sich detaillierte Anweisungen zur computertechnischen Umsetzung der Schätzungen. Bei Tutz (2003) ist die Theorie der Differenzenpenalties in aller Kürze zusammengefasst.

Im folgenden soll auf jedes der drei Modelle kurz eingegangen werden, zunächst auf das nichtparametrische Modell. Neben einem nichtparametrischen Term werden auch zeitabhängige Varying-Coefficients-Terme und zeitabhängige multiplikative Parameter in das Modell integriert. Varying-Coefficients-Modelle wurden bei Hastie und Tibshirani (1993) erstmals vorgestellt. Diese gehen in dieser Arbeit als zeitabhängige Intercepts ein. Die multiplikativen Parameter werden in Kombination mit den nichtparametrischen Termen geschätzt. Restriktionen müssen eingeführt werden, um die Terme des Modells inhaltlich abzutrennen und damit die Identifizierbarkeit der Parameter zu gewährleisten. Dies stellt eine der Hauptschwierigkeiten der Arbeit dar, weil nicht nur die Schätzgleichungen, sondern auch die Hat-Matrizen, Fisher-Matrizen und die Working-Kovarianz, sowie die Sandwichschätzer und viele weitere Größen sensibel auf die Restriktionen reagieren. Details dazu werden in den folgenden Kapiteln und im Appendix erläutert. Die Schätzung der Modelle er-

folgt über einen iterativen Algorithmus, der nahezu immer konvergiert. Zunächst wird das Modell ausführlich simuliert und später an einem Datensatz überprüft. Die Simulations-Studie ist so angelegt, dass ein möglichst großes Spektrum an Konstellationen abgedeckt wird. Der Datensatz wurde von der ETH Zürich im Rahmen einer Arbeits- und Stress-Studie zur Verfügung gestellt. Ziel der Analysen war es, Stress, der sich in Form von Cortisol im Speichel messen lässt, in Beziehung zum Body-Mass-Index (BMI) zu setzen.

Im folgenden Teil der Arbeit wird das nichtparametrische Modell in ein semiparametrisches Modell überführt, indem zusätzlich zu den vorhandenen Termen ein parametrischer Term beliebiger Dimension eingebaut wird. Trotz der steigenden Modell-Komplexität werden nach wie vor sehr gute Schätzungen erzielt. Dies zeigt eine ausführliche Simulationsstudie. Anregungen zur semiparametrischen Modellierung finden sich bei Carroll (2003), Lin und Carroll (2000), Lin und Carroll (2001a, b), Lin et al. (2004), Ruppert et al. (2003), Wang (2003) und Wang et al. (2005).

Literatur zur Modellierung von Seemingly-Unrelated-Regression-Modellen (SUR) wurde unter anderem von Zellner (1962, 1963) und Carroll (2003) veröffentlicht. Zellner definiert darin die grundlegenden Eigenschaften von SUR-Modellen, während sich Carroll in seinem Manuskript eher auf ein spezielles marginales SUR-Modell konzentriert, das in Ansätzen mit dem in dieser Arbeit behandelten Modell übereinstimmt. Allerdings geht er dabei nicht auf Modelle mit longitudinaler Datenstruktur ein, sondern nur auf zwei korrelierte Responses, die durch ein semiparametrisches Modell vorhergesagt werden sollen. Recherchen zu Methoden und Ansätzen zur Modellierung von SUR-Modellen in Kombination mit Longitudinal-Daten führten zu keinen Ergebnissen.

In dieser Arbeit wird ein SUR-Modell konstruiert, das zwei longitudinale Responses in Beziehung zu einer unabhängigen Variablen setzt. Zunächst wird dabei auf zwei unabhängige, parallel geschaltetete Regressionsmodelle mit multiplikativen Para-

metern, Varying-Coefficients und nichtparametrischen Termen eingegangen. Im Anschluss werden Korrelationen zwischen den zwei Responses zugelassen. Zwei verschiedene Kovarianzstrukturen müssen modelliert werden. Zum einen ist dies die schon angesprochene Working-Kovarianz-Struktur innerhalb der scheinbar unabhängigen Modelle, zum anderen die zu modellierende SUR-Struktur. Die Schätzung der Working-Kovarianzen erfolgt wie zuvor über die Momentenmethode. SUR-Korrelationen werden auch in Form einer Working-Kovarianz-Dekomposition in das Modell integriert. Schätzer für nicht- und semiparametrische SUR-Modelle für Longitudinaldaten werden im Anschluss abgeleitet. Nach der theoretischen Herleitung der Modelle werden diese in einer weiteren Simulationsstudie überprüft.

Die Dissertation setzt sich wie folgt zusammen. Nach den Danksagungen und der Einleitung in Kapitel 1, beginnt der Theorieteil der Arbeit. Darin werden die Grundlagen für die statistischen Modelle gelegt, die im Anschluss definiert und formuliert werden. In Kapitel 2 wird sehr detailliert auf 'Generalisierte Lineare Modelle' (GLM) eingegangen. Zunächst werden grundlegende Definitionen besprochen, später werden beispielsweise Dispersion-Parameter, Quasi- und Pseudo-Log-Likelihoods diskutiert. Ziel des Abschnittes ist es, die einzelnen Likelihood-basierten Ansätze klar voneinander abzutrennen. Im anschließenden Kapitel 3 werden 'Marginale Modelle' behandelt. Basis dieser Sektion sind die schon zuvor angesprochenen Artikel von Liang und Zeger, die die 'Independence Estimation Equations' (IEE) und 'Generalized Estimation Equations' (GEE-1) formulieren. Hinzu kommen noch die um das dritte und vierte Moment erweiterten GEE-2-Schätzer, die Schätzung der Sandwich-Matrizen und verschiedene erweiterte Methoden zur Schätzung des Dispersion-Parameters. Ein kurze Diskussion zeigt die Schwächen der IEE/GEE-Ansätze auf. Nachdem nun die grundlegenden Schätzmethoden dargestellt wurden, können diese auf verschiedene nichtparametrische Verfahren angewandt werden. Diese sind in Kapitel 4 zusammengefasst. Zu den nichtparametrischen Verfahren zählen simple lokale Schätzer wie der Running-Mean oder der LOESS-Schätzer genauso wie kompliziertere Verfahren wie 'Kernel-Estimation'-Ansätze, 'B'- und 'P-Spline'-Ansätze, sowie 'Smoothing-Splines' und die 'Truncated Power Series Basis'. Modelloptimierungskriterien werden

am Ende dieses Kapitels besprochen. Anschließend wird in Kapitel 5 auf semiparametrische Modelle eingegangen. Darin sind klassische Ansätze wie z.B. der von Speckman (1988) und auch modernere Ansätze der Autoren Carroll, Lin, Wang, Ruppert und Welsh enthalten, die sich alle mit der Konstruktion eines effizienten Schätzer für longitudinale Kernel-Estimators beschäftigten. In Kapitel 6 werden die SUR-Modelle in aller Kürze zusammengefasst. Hier endet der Theorieteil und die Konstruktion der zuvor angesprochenen Modelle beginnt.

Kapitel 7 stellt das nichtparametrische Modell mit zeitabhängigen multiplikativen Parametern und Varying-Coefficients vor. In Kapitel 8 werden die Simulationen durchgeführt. Im Anschluss folgt in Kapitel 9 die Anwendung des Modells auf einen Cortisoldatensatz. Danach folgt die theoretische Darstellung des semiparametrischen Modells in Kapitel 10 mit anschließender Simulationsstudie in Kapitel 11. Die beiden Kapitel 12 und 13 beschäftigen sich mit den SUR-Modellen und den zugehörigen Simulationen. In Kapitel 14 folgt eine kurze Zusammenfassung der Ergebnisse. Im Appendix sind alle wichtigen Informationen zur Herleitung der restringierten Schätzer und auch die Ergebnisse der Simulationen in Tabellenform dargestellt.

Kapitel 2

Generalized Linear Models (GLM)

Die Grundlagen zur Theorie generalisierter linearer Modelle wurden 1972 von Nelder und Wedderburn gelegt. Sie erweiterten Finneys Prozeduren zur Maximum-Likelihood-Schätzung von Probit-Modellen aus dem Jahr 1952 und die klassische Regressionstheorie.

Die klassische lineare univariate Regressionstheorie basiert auf einer abhängigen, endogenen Variable y_i, $i = 1, \ldots, n$, die als normalverteilt und metrisch angenommen wird und einer unabhängigen, exogenen Variable x_i, die im einfachsten Fall skalar ist, in den meisten Fällen aufgrund mehrerer Einflussvariablen als $(1 \times p)$-Vektor dargestellt werden kann. Die unabhängigen Variablen können sowohl metrisch, binär als auch als Dummies kodiert in das Modell einfließen. Eine Kombination verschiedener Variablentypen ist zulässig.

Damit lässt sich folgendes lineares Regressionsmodell formulieren.

$$y_i = z_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim N(\mu, \sigma^2) \tag{2.1}$$

Der Designvektor z_i=$(1, x_{i1}, \ldots, x_{ip})'$ setzt sich aus einem Intercept (nicht verpflichtend) und den Einflussvariablen x_i zusammen. Die Fehler ε_i werden als normalverteilt mit Erwartungswert $\mu = 0$ und Varianz σ^2 und als unabhängig angenommen. Die Parameter β werden mittels Ordinary-Least-Square-Estimation (OLSE) geschätzt.

2.1 Generalisierung des klassischen Modells

Die getroffenen Annahmen bezüglich der Verteilung des Response und der Art der Variablen schränken die Modellbildung stark ein. Beispielsweise würde das klassische Regressionsmodell (2.1) bei Poisson-verteilten Zähldaten oder Binomial-verteilten Responses ineffektiv sein. Im Poisson-Modell, weil der Wertebereich des Erwartungswerts $E(y_i) = \mu$ von Natur aus gleich Null oder positiv ist, d.h. $0 \leq \mu < +\infty$, im Binomial-Modell, weil $0 \leq \mu \leq 1$. Eine lineare Abbildung von $z_i'\beta$ genau auf den gewünschten Wertebereich von y_i ist in solchen Fällen im klassischen Modell meistens nicht gegeben. Daher müssen Funktionen gefunden werden, die den unabhängigen Part mit dem abhängigen Part des Modells geeignet in Verbindung bringen. Diese Funktionen, in der Literatur als Link- oder Responsefunktionen bezeichnet, können für verschiedene Arten von Verteilungen und auch für verschiedene Arten von Responses, beispielsweise kategorialem und ordinalem Response, definiert werden.

2.2 Definition GLM

Eine Vielzahl an Definitionen und Erweiterungen für GLM's ist in der statistischen Literatur verfügbar. Zunächst führten Nelder und Wedderburn (1972) die GLM's im Jahr 1972 ein. Folgepublikationen erschienen von Kotz und Johnson (1983), McCullagh und Nelder (1989), Green und Silverman (1994), Jörgensen (1997) und Fahrmeir und Tutz (2001). Letztgenannte Publikation soll Basis der folgenden Ausführungen sein.

Im Vergleich zum klassischen Modell werden nun folgende Annahmen etwas genereller dargestellt:

1. Verteilungsannahme:

   - Die Beobachtungen $y_1, \ldots, y_n$ seien bei gegebenem $x_1, \ldots, x_n$ unabhängig
   - Die Verteilung von y_i gehört zur Gruppe der Exponentialfamilien mit Erwartungswert $E(y_i|x_i) = \mu_i$ und einem möglichen Scale-Parameter ϕ.

2. Strukturannahmen:

   - Der lineare Prädiktor $z_i'\beta$, im folgenden als $\eta_i = z_i'\beta$ bezeichnet, wird über eine Funktion mit dem Erwartungswert μ_i in Verbindung gebracht, so dass gilt

$$\mu_i = h(\eta_i) = h(z_i'\beta), \quad \text{bzw.} \quad \eta_i = g(\mu_i).$$

Die Funktion h wird dabei als ausreichend glatte Response-Funktion bezeichnet. Sie ist bekannt und verknüpft die beiden Teile des Modells in einer eineindeutigen Beziehung. Die Funktion g wird als 'Link-Funktion' bezeichnet und stellt die Inverse von h dar. Der Parameter β $(p \times 1)$ ist unbekannt und z_i' ist der Designvektor, der durch mögliche Transformationen von x_i entsteht. Auch er hat die Dimension $(1 \times p)$.

Ein spezifisches GLM ist also vollständig charakterisiert durch die drei Komponenten.

1. Typ der Exponentialfamilie (zufällige Komponente)
2. Response oder Link-Funktion
3. Design-Vektor

Vor allem mit der Exponentialfamilie soll sich nun genauer befasst werden.

2.3 Definition Exponentialfamilie

Beschäftigt man sich eingehend mit statistischer Literatur zum Thema Exponentialfamilien, wird man schnell merken, dass sich die formale Notation von Autor zu Autor und auch über die Jahrzehnte hinweg stets verändert hat. Generell lässt sich die einparametrische Exponentialfamilie gemäß Egge und Krowne (2000) folgendermaßen definieren:

Eine Dichtefunktion $f_X(x|\theta)$ mit Parameter θ gehört zur einparametrischen Exponentialfamilie, falls sie in folgender Notation ausgedrückt werden kann:

$$\begin{aligned} f_X(x|\theta) &= a(x)b(\theta)exp\{c(x)d(\theta)\}, \quad \text{bzw.,} \\ f_X(x|\theta) &= exp\{a(x) + b(\theta) + c(x)d(\theta)\}, \end{aligned} \tag{2.2}$$

wobei es keinen Unterschied machen soll, ob in die Funktionen $a(x)$ und $b(x)$ der Logarithmus integriert ist oder nicht im Vergleich beider Definitionen. Die vier Funktionen a, b, c and d sind bekannt, a und b werden als für die Verteilung spezifische Funktionen bezeichnet. Weitere Informationen erhält man bei Wales (2001). Hier wird die einparametrische Exponentialfamilie wie folgt definiert, wobei sich die Funktionsnamen verändern:

$$f(x|\theta) = h(x)exp\{\theta' T(x) - A(\theta)\} \tag{2.3}$$

und $d(\theta) = \theta$ (vgl. (2.2)).

Die einzelnen Komponenten der Exponentialfamilie werden hier genauer erläutert. So wird $h(x) = a(x)$ (aus (2.2)), als 'reference density' und θ als natürlicher Parameter bezeichnet, aber nur falls $T(x) = c(x) = x$ gilt (aus (2.2)). Das Produkt $\theta' T(x)$ is der einzige Berührungspunkt zwischen θ und x in der Dichtefunktion. $A(\theta) = b(\theta)$ ist die 'Normalizing Constant', d.h. Normalisierungskonstante, die gleichzeitig auch 'Cumulant Generating Function (CGF)' ist. Cumulants ist der englische Begriff für die Bildung von Momenten, d.h. $A(\theta)$ ist diejenige Funktion aus der durch Ableiten die jeweiligen ersten, zweiten und alle folgenden Momente gebildet werden. Die erste Ableitung der CGF entspricht dem 'Mean', d.h. $E(y) = A'(\theta) = \partial A/\partial\theta = \mu$, die zweite Ableitung der Varianzfunktion $V(\mu) = A''(\theta) = \partial^2 A/\partial\theta^2$. Auf diese und weitere Ableitungen wird zu einem späteren Zeitpunkt genauer eingegangen.

Wird den obengenannten Definitionen ein von μ unabhängiger 'Scale'-, 'Dispersion'- oder 'Nuisance'-Parameter $\alpha(\phi)$[1] angefügt, spricht man nicht mehr von Exponentialfamilien, sondern von 'Exponential Dispersion Models' (Toutenburg, 1994). Mehrere

[1] von nun an wird dieser Parameter als $\alpha(\phi)$ bezeichnet und nicht mehr als $a(\phi)$.

Artikel sind zu den Exponential Dispersion Models in Journalen veröffentlicht worden, wie z.B. der Basisartikel zu GLM's von Nelder und Wedderburn (1972). Nelder und Wedderburn definieren die Exponential Dispersion Models wie folgt[2]:

- Suppose our observations z come from a distribution with density function

$$\pi(z;\theta,\phi) = exp[a(\phi)\{z\theta - g(\theta) + h(z)\} + \beta(\phi,z)]. \tag{2.4}$$

In ihren Artikel gehen sie lediglich auf den Scale-Parameter $a(\phi)$ genauer ein, vernachlässigen aber den Rest oder setzen ihn als bekannt voraus. Aus heutiger Sicht hat sich wohl folgende (modernere) univariate Notation durchgesetzt (McCullagh und Nelder, 1983, Hastie und Tibshirani, 1990, Fahrmeir und Tutz, 2001):

$$f(y;\theta,\phi) = exp[\frac{y'\theta - b(\theta)}{a(\phi)} + c(\phi,y)] \tag{2.5}$$

Beide Definitionen (2.4) und (2.5) unterscheiden sich nur bezüglich der Funktion $h(z)$, die in der neueren Version wohl gleich in $c(\phi,y)$ integriert oder vernachlässigt wurde, sowie in der Umkehrung von $a(\phi)$ in dessen Kehrwert.

Die CGF ist, wie auch in (2.2), mit $b(\theta)$ bezeichnet. Große Unterschiede von (2.5) zu den Definitionen aus (2.2) bzw. (2.3) bestehen nicht. So entspricht $y'\theta$ dem Produkt $c(x)d(\theta)$ bzw. $\theta'T(x)$ und $c(\phi,y)$ den Termen $a(x)$ bzw. $h(x)$.

Zusätzliche Veränderungen der Definitionen treten auf, falls Gewichte eingeführt werden. Fahrmeir und Tutz (2001) und Jörgensen (1993, 1997) definieren die Exponential Dispersion Models dann wie folgt.

$$f(y;\theta,\phi,\omega) = exp[\frac{y'\theta - b(\theta)}{a(\phi)}\omega + c(\phi,y,\omega)] \tag{2.6}$$

Jörgensen (1993) verallgemeinert Definitionen (2.5) und (2.6) etwas und definiert eine erweiterte Klasse für Exponentialfamilien mit folgender Dichte

$$f(y;\theta,\kappa) = c(y,\kappa)exp\{a(\kappa)t(y,\theta)\}$$

[2]es sei darauf hingewiesen, dass die Definitionen der Artikel exakt übernommen werden und daher die Beobachtungen von nun an als z bzw. y und nicht mehr als x bezeichnet werden.

und zeigt, dass Verteilungen, die in diese Form zerlegbar sind, in vielerlei Hinsicht ähnliche (Schätz)-Eigenschaften besitzen wie generalisierte lineare Modelle basierend auf einparametrischen Exponentialfamilien.

Auch Zhao, Prentice und Self (1992) erweitern bzw. verallgemeinern den Begriff der Exponentialfamilie. Sie definieren eine neue Familie von Verteilungen und bezeichnen sie als 'partly exponential', d.h. eine Familie von Verteilungen mit teilweise exponentieller Form. Verallgemeinert lässt sich die Dichtefunktion schreiben als

$$f(y, \theta, \lambda) = \Delta^{-1} exp\{y'\theta + c(y, \lambda)\}$$

wobei die Autoren $\Delta = \Delta(\theta, \lambda)$ als 'integration constant' und c als 'shape-function' bezeichnen. Desweiteren wird in ihrem Artikel gezeigt, dass die neue Familie von Dichtefunktion sowohl die Klasse von generalisierten linearen Modellen, so wie sie Nelder und Wedderburn (1972) definieren, als auch die Klasse von diskreten generalisierten linearen Modellen, wie sie Jörgensen (1997) definiert, umfasst. Zahlreiche andere multivariate Verteilungen, wie die multivariate Normalverteilung, die multivariate Exponential-Verteilung, die negative Multinomial-Verteilung, die Dirichlet-Verteilung, die invertierte Dirichletverteilung, u.v.a. gehören auch zur Familie der teilweise exponentiellen Verteilungen.

Allgemein bekannte Mitglieder der einparametrischen Exponentialfamilie können beispielsweise die Normal-, Gamma-, Poisson-, Bernoulli- und Binomial-Verteilung. Im weiteren wird vor allem auf normalverteilte Daten eingegangen.

2.4 Response- und Linkfunktion

Grundsätzlich verbindet man den abhängigen und unabhängigen Part im Regressionsmodell mit einer Linkfunktion, damit der Response nicht transformiert werden muss. Im Fall eines binären Response wäre - wie schon angedeutet - eine Transformation in den Bereich zwischen 0 und 1 notwendig, um den Modellannahmen gerecht zu werden. Ein transformierter Response ist im Normalfall schwer zu interpretieren. So benutzt man Linkfunktionen, um zu gewährleisten, dass das Modell

linear in $g(\mu)$ ist und vor allem, um zu vermeiden, mit einem transformierten Mean arbeiten zu müssen.

Die Wahl der Link- bzw. deren Inversen, der Responsefunktion, ist von substantieller Bedeutung im GLM. Sie hängt von zwei verschiedene Faktoren ab. Zunächst sollte die Linkfunktion passend zum jeweiligen Response gewählt werden, andererseits sollte man eine Linkfunktion aussuchen, die für die Anwendung geeignet erscheint. Natürlich gibt es für jede Verteilung mehrere Möglichkeiten. So kann man bei binärem Response, der durch die Wahrscheinlichkeit π ausgedrückt wird, beispielsweise aus folgenden Funktionen auswählen.

- Probit: $\pi = \Phi(\eta) = \Phi(z'\beta)$[3]
- Logit: $\pi = h(\eta) = \frac{exp(\eta)}{1+exp(\eta)}$
- Complementary Log-Log: $\pi = h(\eta) = 1 - exp(-exp(\eta))$
- Complementary Log: $\pi = h(\eta) = 1 - exp(-\eta)$, falls $\eta > 0$, 0 sonst

Wie bereits angedeutet, bildet die jeweilige Linkfunktion die Linearkombination von unabhängigen Variablen auf den jeweiligen Wertebereich des Response ab. Im Fall des binären Response entspricht der Wertebereich dem Bereich zwischen Null und Eins, d.h. $0 \leq \pi \leq 1$.

Jede Verteilung besitzt eine natürliche oder kanonische Linkfunktion. Diese Funktion verknüpft den natürlichen, im englischen 'canonical' Parameter, direkt mit dem linearen Prädiktor, so dass gilt:

$$\theta = \theta(\mu) = \eta = z'\beta$$

Der natürliche Parameter ist von Verteilung zu Verteilung verschieden. Er lässt sich durch Zerlegen der Verteilung in die einzelnen Funktionen der Exponentialfamilie bestimmen. Im folgenden sind die natürlichen Parameter von fünf Verteilungen angegeben.

[3] Φ ist die Verteilungsfunktion der Normalverteilung

- Normalverteilung: μ (Mittelwert μ)
- Bernoulli: $log(\frac{\pi}{1-\pi})$ (Auftrittswahrscheinlichkeit π)
- Poisson: $log\lambda$ (Poisson-Rate λ)
- Gamma: $-\frac{1}{\mu}$
- Inverse Normalverteilung: $\frac{1}{\mu^2}$

Falls der natürliche Parameter im Fall der Normalverteilung ausgewählt wird, d.h. $E(y) = \mu = \eta = z'\beta$, spricht man vom 'identischen Link', d.h. man geht vom generalisierten linearen Modell in das klassische Regressionsmodell aus Sektion 2.1 über. Damit vereinfacht sich der Schätzvorgang erheblich, da kein iterativer Algorithmus zum Schätzen von β mehr notwendig ist.

2.5 Bildung der ersten und zweiten Momente

Wie bereits in Abschnitt 2.3 erwähnt spielt die 'Cumulant Generating Function' $b(\theta)$ bei der Bildung der Momente von Exponentialfamilien ein wichtige Rolle. Durch Ableiten von (2.5) nach θ werden die ersten beiden Momente, der Mean und die Varianz, bestimmt. Sie lassen sich durch folgende gut bekannte Relationen darstellen (McCullagh und Nelder, 1983)

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{2.7}$$

für den Erwartungswert und

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0 \tag{2.8}$$

für die Varianz, wobei in der zugehörigen Literatur generell von einer Likelihood ausgegangen wird, die mit l oder L bezeichnet wird. Es gelte

$$l(\theta; y) = log(f(\theta; y)) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

mit

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \tag{2.9}$$

und

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)} \tag{2.10}$$

wobei $b'(\theta)$ und $b''(\theta)$ die jeweiligen ersten und zweiten Ableitungen der CGF darstellen. Aus (2.7) und (2.9) ergibt sich folgender Zusammenhang

$$0 = E\left(\frac{\partial l}{\partial \theta}\right) = E\left(\frac{y - b'(\theta)}{a(\phi)}\right) \tag{2.11}$$

so dass gilt

$$E(y) = \mu = b'(\theta).$$

Im Falle der Varianz lässt sich folgende Gleichung aus (2.7), (2.8) und (2.10) ableiten.

$$\begin{aligned} 0 &= E\left(-\frac{b''(\theta)}{a(\phi)}\right) + \left(E\left(\frac{y - b'(\theta)}{a(\phi)}\right)^2\right) \\ &= E\left(-\frac{b''(\theta)}{a(\phi)}\right) + E\left(\frac{(y - b'(\theta))(y - b'(\theta))^T}{a(\phi)^2}\right) = \\ &= -\frac{b''(\theta)}{a(\phi)} + \frac{var(y)}{a(\phi)^2} \end{aligned}$$

Nach Umformung gilt

$$var(y) = \frac{\partial \mu}{\partial \theta} = a(\phi)b''(\theta) = a(\phi)V = a(\phi)V(\mu). \tag{2.12}$$

Somit sind die ersten und zweiten Momente von y bestimmt. Es sei darauf hingewiesen, dass damit die Varianzfunktion des GLMs eindeutig bestimmt ist. Keine unbekannten Parameter sind darin enthalten. Da die Varianzfunktion durch Ableiten von μ bestimmt wird, kann sie auch als Funktion von μ ausgedrückt werden, d.h. $V = V(\mu)$.

2.6 Schätzung der Parameter β

Generell wird die Maximum-Likelihood-Methode angewandt, um die unbekannten Parameter β im GLM zu schätzen. Falls die Annahmen, definiert in Sektion 2.2, verletzt sind oder etwas gelockert werden sollen, kann auf Quasi-Likelihood-Ansätze (QLE) und auch andere Ansätze zurückgegriffen werden. So gelten bei QLE's nur die Voraussetzungen, dass die Mean- bzw. die Varianzfunktion vor der Schätzung spezifiziert werden und keine zusätzlichen Annahmen beispielsweise bezüglich der Verteilung getroffen werden müssen.

Ziel der Maximum-Likelihood-Schätzung ist es, die unbekannten Parameter β im Modell $E(y_i|x_i) = \mu_i = h(z_i'\beta)$ zu schätzen, indem die Likelihoodfunktion maximiert wird. Größtenteils wird in der Literatur auf den Buchstaben 'L' bzw. 'l' als Symbol für die Likelihood zurückgegriffen. Im folgenden gelten nun die Annahmen aus Sektion 2.2, die Beobachtungen werden als y_i, $i = 1, \ldots, n$, die Kovariaten als $x_i = (x_{i1}, \ldots, x_{ip})$ bezeichnet mit Designvektor z_i. Die 'Gesamt-Designmatrix' $Z = (z_1, \ldots, z_n)'$ sei von vollem Rang p. Aus Gründen der Einfachheit beim Ableiten wird die Likelihoodfunktion logarithmiert.

Damit definiert sich die individuelle Log-Likelihoodfunktion für die Beobachtung y_i über die Definition der Exponentialfamilie (2.5).

$$l_i(\theta) = log f(y_i|\theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \tag{2.13}$$

Da

$$\theta_i = \theta(\mu_i) \quad \text{und} \quad \mu_i = h(z_i'\beta) = h(\eta_i) \text{ mit } \eta_i = z_i'\beta$$

gilt und folgt, dass

$$\theta_i = \theta(h(z_i'\beta)) \tag{2.14}$$

kann die Loglikelihoodfunktion aus (2.13) als von β abhängige Funktion betrachtet werden, mit

$$l_i(\beta) = log(l_i(h(z_i'\beta))) = \frac{y_i\theta(h(z_i'\beta)) - b(\theta(h(z_i'\beta))}{a(\phi)} + c(y_i, \phi). \tag{2.15}$$

Es sei angefügt, dass die Log-Likelihood, bei gegebenem y, als eine Funktion von θ (vgl. 2.13) und somit auch als Funktion von β (vgl. 2.15) betrachtet wird. Der Scale-Parameter $a(\phi)$ wird im folgenden als bekannt und konstant vorausgesetzt. Aufgrund der Unabhängigkeitsannahme der Beobachtungen y_i folgt, dass sich aus der Summe der individuellen Log-Likelihoodfunktionen eine Gesamt-Log-Likelihood bilden lässt.

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta) \tag{2.16}$$

Die Gesamt-Log-Likelihood wird nun in β maximiert. Nullsetzen der Ableitung und Auflösen nach β ist eine notwendige Bedingung für 'Maximum-(Log)-Likelihood-Estimators' (MLE). Die Ableitung der Gesamt-Log-Likelihood wird auch als 'Score-Funktion' $s(\beta)$ bezeichnet.

$$s(\beta) = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta) = \sum_{i=1}^{n} z_i D_i(\beta) \sigma^{-2}(\beta)[y_i - \mu_i(\beta)] \tag{2.17}$$

Die Scorefunktion (2.17) setzt sich aus der Summe der individuellen Scorefunktionen $s_i(\beta)$ zusammen und hat die Dimension $(p \times 1)$. Für die übrigen Terme gilt

$$\begin{aligned} \mu_i &= h(z_i'\beta), \\ \sigma_i^2(\beta) &= a(\phi) V(h(z_i'\beta)), \qquad (2.18) \\ D_i(\beta) &= \frac{\partial h(\eta_i)}{\partial \eta_i} = \frac{\partial h(z_i'\beta)}{\partial \eta_i}. \qquad (2.19) \end{aligned}$$

Es sei noch angemerkt, dass Gleichung (2.17) in der Literatur üblicherweise in Matrixnotation dargestellt wird,

$$s(\beta) = Z' D(\beta) \Sigma^{-1} [y - \mu(\beta)] \tag{2.20}$$

wobei $Z_{n \times p}$ die Gesamt-Designmatrix ist. $D(\beta)_{n \times n}$ ist die erste Ableitung der Responsefunktion $h(\eta)$, ausgewertet an der Stelle $\eta_i = z_i'\beta$. Die Diagonalmatrix $\Sigma^{-1}_{n \times n}$ ist die Inverse der Varianzfunktion und $[y - \mu(\beta)]_{n \times 1}$ sind die Residuen.

2.7 Fisher-Informations-Matrix

Die erwartete Fisher-Informations-Matrix ist definiert als

$$F(\beta) = cov(s(\beta)) = \sum_{i=1}^{n} F_i(\beta)$$

mit $F_i(\beta) = z_i z_i' w_i(\beta)$. Der Term $w_i(\beta)$ ist die Gewichtsmatrix, die sich aus zwei Termen der Scoregleichung (2.17) zusammensetzt.

$$w_i(\beta) = D_i^2(\beta)\sigma_i^{-2}(\beta) = \left(\frac{\partial h(\eta_i)}{\partial \eta_i}\right)^2 \frac{1}{v(h(z_i'\beta))a(\phi)} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{a(\phi)V(\mu_i)} \quad (2.21)$$

Als beobachtete Fisher-Matrix bezeichnet man

$$F_{obs}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'}.$$

Außerdem gilt

$$F(\beta) = E(F_{obs}(\beta)).$$

2.8 Ableitung der Log-Likelihoodfunktion

Der Übergang der Log-Likelihoodfunktion (2.16) zur Scorefunktion (2.17) erweist sich als nicht ganz trivial. Dies hat den Grund, dass nicht direkt nach dem Parameter θ, sondern nach dem Parameter β abgeleitet werden muss, um die unbekannten Parameter zu bestimmen. Nützlich erweist sich hierbei Gleichung (2.14), die zeigt, dass θ als Funktion von β dargestellt werden kann. In Gleichung (2.15) wird die Log-Likelihood als Funktion von β dargestellt.

Von Interesse ist also nun folgende Ableitung (ohne Index i)

$$\frac{\partial l}{\partial \beta_j}, \quad \text{gebildet aus} \quad l = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi).$$

Da $\theta = \theta(\mu)$, $\mu = h(\eta)$ und $\eta_i = z_i'\beta = \sum_{j=1}^{p} x_{ij}\beta_j$ mit $\eta = (\eta_1, \ldots, \eta_n)'$ und z_i ein aus den verschiedenen x_{ij} zusammengesetzter Designvektor ist, gilt nach der

Kettenregel beim Ableiten (McCullagh und Nelder, 1983)

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_j}. \tag{2.22}$$

Allgemein gilt

$$\frac{\partial a}{\partial b} = \frac{1}{\frac{\partial b}{\partial a}}, \quad \text{und somit gilt} \quad \frac{\partial \theta}{\partial \mu} = \frac{1}{\frac{\partial \mu}{\partial \theta}} = \frac{1}{a(\phi)V(\mu)}, \quad \text{nach (2.12).}$$

Damit folgt aus (2.11), (2.12) und $\eta = \sum_{j=1}^{p} x_j\beta_j$ zunächst

$$\frac{\partial l}{\partial \beta_j} = \left(\frac{\partial l}{\partial \theta}\right)\left(\frac{\partial \theta}{\partial \mu}\right)\left(\frac{\partial \mu}{\partial \eta}\right)\left(\frac{\partial \eta}{\partial \beta_j}\right) = \left(\frac{y-\mu}{a(\phi)}\right)\left(\frac{1}{a(\phi)V(\mu)}\right)\left(\frac{\partial \mu}{\partial \eta}\right)x_j. \tag{2.23}$$

Nach (2.21) gilt allgemein:

$$w = \left(\frac{\partial \mu}{\partial \eta}\right)^2 \frac{1}{a(\phi)V(\mu)}, \quad \text{so dass} \quad a(\phi)V(\mu) = \left(\frac{\partial \mu}{\partial \eta}\right)^2 \frac{1}{w}. \tag{2.24}$$

Einsetzen von Gleichung (2.24) in (2.23) führt zu

$$\frac{\partial l}{\partial \beta_j} = \left(\frac{y-\mu}{a(\phi)}\right)\frac{w}{\left(\frac{\partial \mu}{\partial \eta}\right)^2}\left(\frac{\partial \eta}{\partial \mu}\right)x_j = \frac{w}{a(\phi)}(y-\mu)\left(\frac{\partial \eta}{\partial \mu}\right)x_j.$$

Es folgt

$$\frac{\partial l_{ges}}{\partial \beta_j} = \sum_{i=1}^{n}\left(\frac{\partial \eta}{\partial \mu}\right)x_{ij}\frac{w}{a(\phi)}(y-\mu) = 0 \tag{2.25}$$

mit $w = (\partial\mu/\partial\eta)^2 V^{-1}a(\phi)^{-1}$. Gleichung (2.25) führt nach trivialen Umformungen zu Scorefunktion aus (2.17). Es ist zu beachten, dass in Gleichung (2.17) ein Designvektor an die Stelle von x_{ij} tritt.

2.9 Algorithmus zur Schätzung von β

Weil die Likelihood-Schätzgleichungen in der Regel nicht linear sind, wird ein iterativer Algorithmus zur Schätzung von β benötigt. Der meistgenutzte Lösungsansatz

in solchen Fällen ist die als 'Fisher-Scoring' bezeichnete 'Iterative Weighted Least Squares Estimate (IWLSE)'-Methode. Ausgehend von einem initialen Schätzer $\hat{\beta}^{(0)}$ wird die Schätzgleichung solange iterativ verbessert, bis keine großen Veränderungen bezüglich eines Abbruchkriteriums mehr entstehen.

Als Startwerte für das Fisher-Scoring lassen sich beispielsweise die Schätzer $\hat{\beta}^{(0)} = (Z'Z)^{-1}Z'y$ oder auch $\hat{\beta}^{(0)} = (Z'Z)^{-1}Z'g(y)$ verwenden (Toutenburg, 1994). Die Fisher-Scoring-Iterationen werden definiert als

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F^{-1}(\hat{\beta}^{(k)})s(\hat{\beta}^{(k)}). \tag{2.26}$$

Falls sich die Schätzer $\hat{\beta}^{(k+1)}$ und $\hat{\beta}^{(k)}$ nach einigen Iterationen nicht mehr stark voneinander unterscheiden, wird ein Threshold ε unterschritten und der Algorithmus terminiert. Das Abbruchkriterium ist definiert als

$$\frac{||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||}{||\hat{\beta}^{(k)}||} \leq \varepsilon. \tag{2.27}$$

Durch geschicktes Umformen kann der Schätzer (2.26) auf folgende Form gebracht werden (Fahrmeir und Tutz, 2001)

$$\hat{\beta}^{(k+1)} = (Z'W^{(k)}Z)^{-1}Z'W^{(k)}\tilde{y}^{(k)} \tag{2.28}$$

wobei $W^{(k)}$ und $\tilde{y}^{(k)}$ ausgewertet werden an der Stelle $\beta = \hat{\beta}^{(k)}$. Die Übereinstimmung zum IWLSE und auch zum Aitkenschätzer (Toutenburg, 1993) wird deutlich.

Ist ein Lösungsvektor für (2.28) gefunden, so ist $\hat{\beta}$ ein konsistenter Schätzer für β und zugleich asymptotisch normalverteilt und effizient mit $\hat{\beta} \overset{a}{\sim} N(\beta, V(\hat{\beta}))$ (Fahrmeir und Kaufmann, 1985, Wedderburn, 1976).

Für $k \to \infty$ lässt sich die asymptotische Kovarianzmatrix wie folgt berechnen.

$$cov(\hat{\beta}) = (Z'\hat{W}^{(k)}Z)^{-1} = F^{-1}(\hat{\beta}) \tag{2.29}$$

Wie allgemein bekannt, werden mit (k) bezeichnete Matrizen jeweils an der Stelle $\beta = \hat{\beta}^{(k)}$ berechnet.

Andere Lösungsansätze können auch zur Schätzung der Likelihood-Gleichungen verwendet werden. Ein weiteres Verfahren wird als 'Newton-Raphson' bezeichnet. Es unterscheidet sich lediglich dadurch vom Fisher-Scoring, dass die erwartete durch die beobachtete Fisher-Informations-Matrix ersetzt wird (Fahrmeir und Tutz, 2001). Bei kanonischen Links gleichen sich beide Methoden (McCullagh und Nelder, 1983). Quasi-Newton-Methoden sind alternativ zu verwenden. Sie sind teilweise effizienter als die zuvor besprochenen Ansätze.

2.10 Separation Principle

Geht man davon aus, dass $a(\phi) = \phi$, so gilt

$$var(y) = a(\phi)V(\mu) = \phi V(\mu).$$

Damit ist nur die Varianzfunktion $V(\mu)$ abhängig von μ, während der Dispersion Parameter ϕ unabhängig von μ ist. Box (1988) bezeichnet die Trennung vom Dispersion Parameter ϕ und der Varianzfunktion $V(\mu)$ als 'Separation Principle', d.h. beide Terme sind unabhängig voneinander und werden, falls unbekannt, separat voneinander geschätzt.

2.11 Quasi-Likelihood-Schätzung

Quasi-(Log)-Likelihoods sind den Log-Likelihoods aus (2.13) generell sehr ähnlich, mit dem Unterschied, dass die Verteilungsannahme fallengelassen und die bisher angenommene Struktur zwischen Erwartungswert und Varianz aufgebrochen wird, so dass gilt (Wedderburn, 1974, McCullagh und Nelder 1983, Fahrmeir und Tutz, 2001)

$$E(y|z) = \mu = h(z'\beta), \qquad \text{var}(y|z) = \phi V_{QL}(\mu). \tag{2.30}$$

Die Varianzfunktion $V_{QL}(\mu)$ sei eine bekannte, geeignet gewählte Funktion. Sie wird so spezifiziert, dass sie entweder dem Erwartungswert $E(y|z)$ entspricht oder in einem proportionalen Verhältnis zum Erwartungswert $E(y|z)$ steht (vgl. 2.38). Das

Verhältnis zwischen Varianzfunktion und Varianz wird mittels eines 'Dispersion Parameters' $\phi > 0$ angegeben. Dieser kann auf eins normiert sein (Log-Likelihood), als bekannt und konstant (Log-Likelihood), unbekannt und konstant angesehen werden (Quasi-Likelihood), sogar als ϕ_i in Abhängigkeit von z_i bzw. x_i oder Teilen von x_i separat geschätzt werden (Quasi-Likelihood), wie sich im folgenden zeigen wird.

Es werden in (2.30) also nur Annahmen über die ersten und zweiten Momente getroffen, die Verteilung des Response y entspricht nicht der Form einer Exponentialfamilie. Damit hat der Quasi-Likelihood-Ansatz zwei Vorteile im Vergleich zum ursprünglichen GLM-Ansatz zu bieten. Einerseits wird nun der Dispersion Parameter nicht mehr als fix angesehen, wie beispielsweise im Poisson-Modell, wo er auf $\phi = 1$ festgelegt ist, sondern z.B. von i abhängig oder unbekannt. Im QLE-Ansatz ist ein frei wählbares ϕ also zugelassen. Andererseits kann die Varianzfunktion $V(\mu)$ eine Form annehmen, die nicht der Standardform eines GLM entspricht (Lee und Nelder, 1992).

Wedderburn (1974) definiert, unter eben angesprochenen Annahmen, die Quasi-Loglikelihood mittels folgender Relation

$$\frac{\partial Q_i(y;\mu)}{\partial \mu} = \frac{y_i - \mu_i}{\phi V_{QL}(\mu_i)}. \tag{2.31}$$

Wedderburn zeigt im Abschnitt 4 des zugehörigen Artikels im Falle von normalen Log-Likelihoods, wie sich Gleichung (2.31) ableiten lässt. Er geht dabei von der üblichen einparametrischen Exponentialfamilie (2.5) aus und bildet die Ableitung $\partial l/\partial \mu$ mit $\theta = \theta(\mu)$, so dass gilt[4]

$$\frac{\partial l}{\partial \mu} = (y - b'(\theta))\frac{\partial \theta}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)}. \tag{2.32}$$

In die Ableitung (2.32) gehen die ersten beiden Ableitungen der Kettenregel (2.23) ein und $\partial\theta/\partial\mu = 1/(\partial\mu/\partial\theta)$. Ansonsten ist Ableitung (2.32) direkt mit (2.9) zu vergleichen. Ausgehend von (2.31) werden bei Wedderburn (1974) nun einige Eigenschaften und auch die Quasi-Log-Likelihood-Schätzung abgeleitet. So lässt sich die

[4]Indizes werden hier und auch im Artikel von Wedderburn vernachlässigt.

Quasi-Scorefunktion $s_{QL}(\beta)$ in derselben Form wie die ursprüngliche Scoregleichung (2.17) schreiben

$$s_{QL}(\beta) = \sum_{i=1}^{n} z_i D_i(\beta)\sigma^{-2}(\beta)[y_i - \mu_i(\beta)]. \tag{2.33}$$

Der Term $\sigma^2(\beta)$ verändert sich im Vergleich zu (2.18) leicht, man erhält

$$\sigma^2(\beta) = \phi V_{QL}(\mu).$$

Allgemein gilt nun

$$y \sim (\mu, \phi V_{QL}(\mu)),$$

wobei $V_{QL}(\mu)$ eine Matrix ist, deren Einträge bekannte Funktionen sind, und ϕ ein unbekannter Dispersion Parameter ist, der konsistent geschätzt werden muss. Der Gesamtausdruck $\phi V_{QL}(\mu)$ wird von McCullagh und Nelder (1983) als 'Working Variance' (Arbeitsvarianz) bezeichnet.

Werden die Komponenten von y als unabhängig vorausgesetzt, so muss die Kovarianzstruktur diagonale Gestalt annehmen. Dabei schlagen McCullagh und Nelder (1989) drei Möglichkeiten zur Wahl der Working Variance vor[5].

1. $V_{Ges,1} = V_{Ges,1}(\mu) = \text{diag}(V_{QL,1}(\mu), \ldots, V_{QL,n}(\mu))$
2. $V_{Ges,2} = V_{Ges,2}(\mu) = \text{diag}(V_{QL,1}(\mu_1), \ldots, V_{QL,n}(\mu_n))$
3. $V_{Ges,3} = V_{Ges,3}(\mu) = \text{diag}(V_{QL}(\mu_1), \ldots, V_{QL}(\mu_n))$

Die Wahl der Varianzfunktion $V_{Ges}(\mu)$ ist abhängig vom jeweiligen Modell. Allgemein ist keine der drei Funktionen zu präferieren.

Aufgrund der Unabhängigkeit der Responsevariablen y_i setzt sich auch in diesem Fall die Gesamt-Quasi-Likelihood aus der Summe der 'individual contributions',

[5]Zur Erläuterung: in Punkt 1 ist die Varianzfunktion von i abhängig, in Punkt 2 μ und die Varianzfunktion, in Punkt 3 nur μ.

wie es McCullagh und Nelder (1989) nennen, d.h. den individuellen Beiträgen zur Likelihood, zusammen.

$$Q(\mu; y) = \sum_{i=1}^{n} Q_i(\mu_i; y_i)$$

In Matrixschreibweise lässt sich die Quasi-Scoregleichung (2.33) wie folgt ausdrücken

$$s_{QL}(\beta) = \phi^{-1} D' V_{Ges}^{-1} (y - \mu), \tag{2.34}$$

wobei $D = \partial\mu_i/\partial\beta_j = \partial h(z_i'\beta)/\partial\beta_j$. Die Designmatrix z_i aus (2.33) ist hierbei in D integriert. Lösung von (2.34) ist $\hat{\beta}$, mit folgender asymptotischer Kovarianzmatrix.

$$cov(\hat{\beta}) = \phi (D' V_{Ges}^{-1} D')^{-1} \tag{2.35}$$

Die Schätzung von $\hat{\beta}$ wird nach dem zuvor beschriebenen Fisher-Scoring-Algorithmus (2.26) iterativ vorgenommen. Bei der Schätzung für β spielt der Dispersion Parameter ϕ keine Rolle, da er sich im Term $F^{-1}(\hat{\beta}^{(k)}) s(\hat{\beta}^{(k)})$ herauskürzt (Fahrmeir und Tutz, 2001). Dies bedeutet, dass man den Parameter β unabhängig davon, ob man über Information über den Dispersion Parameter besitzt, schätzen kann. Der Schätzer für β ist auch in diesem Fall approximativ unverzerrt (unbiased) und asymptotisch normalverteilt (Wedderburn, 1974). Allerdings hat der Dispersion Parameter Einfluss auf Folgeanalysen, wie z.B. bei Hypothesentests. Nur bei der Punktschätzung von β kann er vernachlässigt werden.

Um die Verteilungsannahme nicht ganz zu vernachlässigen kann die Quasi-(Log)-Likelihood durch Normalisierung in eine Verteilung umgeformt werden. Nelder and Lee (1992) bezeichnen diese Verteilung dann als 'Quasi-Distribution'

$$f_Q = \frac{exp(Q)}{\omega}, \quad \text{mit} \quad \omega = \int exp(Q) dy$$

wobei ω der Normalisierung-Term ist. Damit entspricht die Quasi-Log-Likelihood bezüglich der Quasi-Verteilung folgendem Ausdruck.

$$l_Q = log(\frac{exp(Q)}{\omega}) = Q - log(\omega) \tag{2.36}$$

Aus (2.36) werden die ML-Gleichungen bzw. die Scoregleichung für die Quasi-Verteilung entwickelt, die allerdings nicht den ML-Equations bzw. der Scoregleichung aus (2.34) entsprechen, da sie ja für die Quasi-Distribution gebildet werden. Der Unterschied zu (2.34) besteht nur im Term $-log(\omega)$ und dessen Ableitung (Nelder und Lee, 1992)

$$\frac{\partial \omega}{\partial \beta} = \frac{\partial \mu}{\partial \beta} \frac{\mu^* - \mu}{\phi V(\mu)}, \tag{2.37}$$

wobei μ^* der 'Quasi-Mean' $\int y f_Q dy$ ist. Falls die Differenz $\mu^* - \mu$ klein im Vergleich zu $y - \mu$ ist, kann erwartet werden, dass die Maximum-Quasi-Likelihoodschätzungen (2.34) approximativ den ML-Schätzungen der Quasi-Distribution entsprechen.

Weitere Anmerkungen und Eigenschaften von Quasi-Likelihood-Modellen sind bei Wedderburn (1974) und McCullagh und Nelder (1983) zu finden. Erweiterte Quasi-Likelihood-Theorien werden bei McCullagh und Nelder (1983) im Kapitel 'Quasi-likelihood functions' und bei Nelder und Pregibon (1987) behandelt.

Die letztgenannten Autoren erweitern die Theorie der Quasi-Likelihood-Schätzung etwas, indem sie annehmen, dass zusätzlich zu den bisherigen Vorgaben auch die Varianz als Funktion von θ dargestellt werden kann, so dass gilt

$$var(y) = \phi V_{QL}(\mu; \theta).$$

Damit generalisieren Nelder und Pregibon (1987) Wedderburns Vorgaben, der die Varianzfunktion V_{QL} als bekannte Funktion abhängig von μ beschreibt. Noch zu bestimmen ist nun, welche Form die Varianzfunktionen haben. So geben die Autoren eine Familie von Funktionen an, die assoziiert sind mit den Varianzfunktionen der Normal-, Poisson-, Gamma- und der inversen Gaussverteilung für $\theta = 0, 1, 2, 3$. Diese lassen sich wie folgt ausdrücken.

$$V_{QL} = \mu^\theta, \quad \theta = 0, 1, 2, 3 \tag{2.38}$$

Eine Übersicht zu Quasi-Likelihoods und zugehörigen Varianzfunktionen ist bei McCullagh und Nelder (1989) auf Seite 326 nachzuschlagen.

2.12 Schätzung des Dispersion Parameter

Obwohl der Dispersion Parameter keinen Einfluss auf die Schätzung von β hat, ist es trotzdem notwendig ϕ konsistent zu schätzen, weil er beispielsweise bei Hypothesentests als 'Scale Parameter' auftaucht. Wedderburn (1974) empfiehlt dafür die 'Bias-korrigierte' χ^2-Statistik[6]

$$\hat{\phi} = \frac{\chi^2}{n-p} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V_{QL}(\mu_i)}. \tag{2.39}$$

Gleichung (2.39) schätzt einen konsistenten Dispersion Parameter, der für alle Beobachtungen y_i konstant ist. Davidian (1995, p. 365) weist darauf hin, dass die Varianzfunktion, falls nicht bekannt, durch eine Schätzung, der sogenannten 'Working-(Co)-Variance'-Schätzung, ersetzt werden kann.

McCullagh (1983) schlägt einen anderen Schätz-Ansatz für einen konstanten Dispersion Parameter vor, der in einigen Fällen besser ist als (2.39), beispielsweise wenn die Log-Likelihood zur Exponentialfamilie gehört. Generell ist der Schätzer verschieden vom ML-Schätzer (2.39) für $\hat{\phi} = \hat{\sigma}^2$ und er wird berechnet, indem die beobachtete Devianz $d(y, \hat{\mu}) = 2l(y; y) - 2l(\hat{\mu}; y)$ gleichgesetzt wird mit dem approximativ erwarteten Wert der Devianz und dann nach dem Dispersion-Parameter aufgelöst wird.

Falls es gewisse Modelle erfordern oder es wünschenswert wäre einen nicht konstanten Dispersion Parameter in die Schätzung zu integrieren, müssen andere Schätzer angewandt werden. Pregibon (1984) (später übernommen von Nelder und Pregibon, 1987, McCullagh und Nelder 1989) schildert eine Prozedur, die diese Vorgaben erfüllt. Er geht hierbei von folgendem Quasi-Likelihood-Modell aus.

$$E(y_i) = \mu_i, \quad \eta_i = g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j, \quad var(y_i|x) = \phi_i V_{QL}(\mu_i) \tag{2.40}$$

Das Modell aus (2.40) unterscheidet sich lediglich durch den von i abhängigen Dispersion Parameter vom ursprünglichen Quasi-Likelihood-Modell (2.30). Damit wird

[6] Die Bias-korrigierte χ^2-Statistik entspricht der üblichen Schätzung der Varianz σ^2 im KQ-Modell. Im Falle einer Normalverteilung entspricht der Dispersion-Parameter daher der Varianz, so dass $\phi = \sigma^2$. Der Varianzfunktion entspräche in diesem Falle der Einheitsmatrix $I = V_{QL}$.

ϕ nicht länger als konstant angesehen, sondern verändert sich systematisch nach folgendem Prinzip.

$$E(d_i) = \phi_i, \quad \zeta_i = h(\phi_i) = \sum_{j=1}^{p} u_{ij}\gamma_j, \quad var(d_i) = \tau V_D(\phi_i) \tag{2.41}$$

Man geht also von einem weiteren GLM aus, das sich auf den Dispersion Parameter bezieht. Alle notwendigen Faktoren für ein GLM sind in diesem Fall auch in das Modell integriert, wie z.B. die Linkfunktion, die mit h bezeichnet wird[7]. Weitere Faktoren sind die Varianzfunktion V_D, der lineare Prädiktor ζ, die Regressionsgewichte γ_j, die zugehörigen unabhängigen Variablen u_{ij} und das Maß für Dispersion d_i, das im Modell die Rolle der abhängigen Variablen übernimmt. Das Maß für Dispersion kann auf zwei Arten geschätzt werden, einerseits nach der üblichen Methode über die Pearson-Residuen (2.42), andererseits nach den individuellen Beiträgen zur Devianz (2.43).

$$d_i = (r_i)_P^2 = \left(\frac{y_i - \hat{\mu}_i}{V_{QL}(\hat{\mu}_i)^{\frac{1}{2}}} \right)^2 \tag{2.42}$$

$$d_i = (r_i)_D^2 = \left(d(y_i - \hat{\mu}_i)^{\frac{1}{2}} \text{ sign } (y_i - \hat{\mu}_i) \right)^2 \tag{2.43}$$

mit $d(y, \mu) = 2\left(t(y, y) - t(y, \mu)\right)$ und $t(y, \mu) = y\theta - b(\theta)$. Sind die Dispersions-Maße d_i normalverteilt, entsprechen sich beide Schätzungen. Bei nicht normalverteilten Responses müssen 'Adjustments' bzgl. der Kurtosis und / oder den Freiheitsgraden vorgenommen werden. Genauer beschrieben sind derartige Adjustments bei McCullagh und Nelder (1989) in Kapitel 10.5.

Als Link-Funktionen werden die Identität $h(\phi_i) = \phi_i$ und der Logarithmus $h(\phi_i) = log(\phi_i)$ empfohlen. Die abhängigen Variablen u_{ij} setzen sich normalerweise, aber nicht notwendigerweise, aus einem Subset der Variablen x_{ij} zusammen. Weil beide GLMs miteinander zusammenhängen ist es offensichtlich, dass ein iterativer Algorithmus zum Anpassen beider Modelle verwendet wird, wobei jeweils einer der beiden Parameter $\hat{\phi}_i$ oder $\hat{\mu}_i$ zur Schätzung des einen bzw. des anderen Modells festgehalten

[7] h ist die Linkfunktion des GLMs für ϕ, nicht die Responsefunktion aus Sektion 2.2.

wird.

Diggle, Liang und Zeger (1994) weisen in ihrem Buch 'Analysis of Longitudinal Data' darauf hin, dass andere Schätzer bzw. Modelle für ϕ Verwendung finden könnten. In einem Modell, in dem zwischen Treatment- und Kontrollgruppe unterschieden wird, könnte man beispielsweise zwischen ϕ_1 für die Treatmentgruppe und ϕ_2 für die Kontrollgruppe unterscheiden. Bei longitudinaler Datenstruktur könnten z.B. ein Dispersion Parameter pro Zeitpunkt t geschätzt werden. Allerdings führen sie nicht genauer aus, wie solche Schätzungen durchzuführen wären.

2.13 Pseudo-(Log)-Likelihood

Gong und Samaniego (1981) definieren die Pseudo-Likelihood wie folgt:

- 'In general, pseudo maximum likelihood estimation consists of replacing all nuisance parameters in a model by estimates and solving a reduced system of likelihood equations.'

Explizit bedeutet dies, dass zunächst die Nuisance Parameters, die im Modell enthalten sind, konsistent geschätzt werden. Nach Breslow (1990) müssen sich zwei oder mehr davon in der Varianzfunktion $V(.)$ befinden, damit die Likelihood als Pseudo-Likelihood bezeichnet werden kann. Dann erst wird das reduzierte Likelihood-Gleichungssystem gelöst und das Ergebnis wird nun als Pseudo-Maximum-Likelihood-Schätzer bezeichnet. In der statistischen Literatur verschwimmen teilweise die Grenzen zwischen Likelihood, Quasi-Likelihood, extended Quasi-Likelihood und Pseudo-Likelihood. Zusammengefasst sind die unterschiedlichen Definitionen in einem Artikel von Nelder (2000), der sehr detailliert erklärt, was die Unterschiede zwischen den einzelnen Ansätzen sind und auch auf Fehler in der zugehörigen Literatur hinweist.

2.14 Erweiterung der GLMs

Natürlich kann die in diesem Kapitel beschriebene Theorie noch erweitert werden. Vor allem im binären Fall wurde in der statistischen Literatur viel veröffentlicht, so wie in der multivariate Erweiterung des binären Falles, im ordinalen und kategorialen Bereich. Auch im Bereich der Longitudinal-Daten ist viel geforscht worden. Modelle für Longitudinal-Daten sollen nun im folgenden dargestellt werden, wobei das Hauptaugenmerk auf Marginalen Modellen liegt. Den Grundstein dieser Theorie legten Liang und Zeger mit zwei Artikeln im Jahre 1986.

Kapitel 3

Marginale Modelle

Das Kapitel über generalisierte lineare Modelle bildet zwar die Grundlagen für folgenden Abschnitte, allerdings ist bisher nur von einem skalaren Response y_i und dem zugehörigen Prädiktoren-Vektor $x_{ij} = (x_{i1}, \ldots, x_{ip})$, $j = 1, \ldots, p$ ausgegangen worden. Von nun an wird die Theorie erweitert, indem ein zusätzlicher Index t eingeführt wird. Es verändern sich $(y_i)_{1\times 1}$ zu y_{it} mit $y_i = (y_{i1}, \ldots, y_{iT})'$, $t = 1, \ldots, T$ und $(x_{ij})_{1\times 1}$ zu x_{itj} mit $x_{it} = (x_{it1}, \ldots, x_{itp})$. Damit sind der Response y und auch die Prädiktor-Variablen x von einem zweiten bzw. dritten Index t abhängig, der beispielsweise bei Messwiederholungen oder bei Daten mit zeitlichem, longitudinalem Verlauf eingeführt wird. Folglich entstehen Korrelationen zwischen den einzelnen Messwiederholungen, die beispielsweise an einer Person durchgeführt wurden. Auch vorstellbar ist, dass innerhalb einer Gruppe ('Cluster') mit ähnlichen Merkmalen Korrelationen entstehen, wie beispielsweise innerhalb einer Familie. Allerdings muss im Falle der Familie der Index t erweitert werden zu $t = 1, \ldots, T_i$, weil die beobachteten Familien nicht (immer) von gleicher Größe sind. Um die Notation möglichst einfach zu halten, wird der erweiterte Index T_i in der Literatur, wie auch in dieser Arbeit, vernachlässigt, so dass gilt $T_i = T$.

Die Notation kann also folgendermaßen zusammengefasst werden

$$\begin{aligned} y &= (y_1', \ldots, y_n')', \quad i = 1, \ldots, n \\ y_i &= (y_{i1}, \ldots, y_{iT})', \quad t = 1, \ldots, T \\ x_{it} &= (x_{it1}, \ldots, x_{itp}), \quad j = 1, \ldots, p. \end{aligned}$$

Man beachte hierbei, dass y von nun an ein Vektor der Länge $nT \times 1$ ist. Der Vektor y setzt sich aus den einzelnen Vektoren y_i zusammen. Diese werden der Reihenfolge nach in y integriert. Es sei noch anzumerken, dass die Vektoren y_i nicht vergleichbar sind mit den skalaren y_i aus dem vorherigen Kapitel, außer im Falle $T = 1$, in dem das longitudinale Modell in das 'normale' GLM übergeht.

Generell gesehen lassen sich verschiedene Modellierungsansätze für korrelierten Response unterscheiden. Zum einen gibt es 'population-averaged models', d.h. Modelle, die den Durchschnitt über eine ganze Population gesehen ermitteln, zum anderen 'subject-specific models', d.h. Modelle, die sich mit individuellen Effekten beschäftigen.

Den 'population-averaged models' kann man die marginalen und auch die Konditional-Modelle zuordnen. Subjekt-spezifischen Modellen sind Random-Effects-Modelle. Nur marginale Modelle sollen im Folgenden von Interesse sein.

3.1 Definition Marginales Modell

Vor den Definitionen soll zunächst anhand zweier kurzer Beispiele auf marginale Modelle eingegangen werden. Beide Beispiele zeigen für welche Art von Fragestellungen marginale Modelle geeignet sind.

3.1.1 Beispiel 1 - Klinische Studien

Bei klinischen Studien ist oft die Wirkungsweise eines bestimmten Medikaments von Interesse. Dazu wird meist eine Control- und eine Treatment-Gruppe gebildet. Die Treatment-Gruppe nimmt ein neues Medikament ein (einmal oder öfter), während

die Control-Gruppe entweder ein Placebo oder ein bekanntes Standardmedikament einnimmt. Dann werden die Effekte des Medikaments bei beiden Gruppen gemessen. Bei der Control-Gruppe sollten keine Effekte auftreten, während im Treatment Effekte zu erwarten wären, so dass beispielsweise bestimmte medizinische Werte besser ausfallen als vor Einnahme des Medikaments. Entweder werden die Effekte cross-sectional oder über die Zeit hinweg gemessen, so dass eine oder mehrere Messungen pro Person ausgewertet werden können. Bei Mehrfach-Messungen $T \geq 2$ treten Korrelationen innerhalb der Messungen auf, die Personen untereinander werden als unabhängig betrachtet. Im besten Fall treten nun signifikante Unterschiede zwischen Control-und Treatment-Gruppe auf. Um diese Unterschiede statistisch zu verifizieren werden Modelle benötigt, die zwischen Control und Treatment unterscheiden können. Dazu sind marginale Modelle perfekt geeignet weil sie die Populationsdurchschnitte berechnen, zum einen von Control, zum anderen von Treatment, so dass getestet werden kann, ob sich ein Treatment-Effekt zeigt. In diesem Fall wären individuelle Effekte nicht von Interesse, da nicht die Differenzen zwischen einzelnen Individuen wichtig sind, sondern nur die Differenz zwischen Control und Treatment.

3.1.2 Beispiel 2 - Indonesian Children's Health Study

Die Indonesian Children Health Study ist eine gut bekannte Longitudinal-Studie zum Thema marginaler Modelle. Sie wird bei Diggle, Liang und Zeger (1994) zitiert. Dabei treten beispielsweise folgende Fragen auf:

- Wie gross ist der Anteil an Kindern in einem bestimmten Alter, die mit einer bestimmten Lungenkrankheiten infiziert sind?
- Ist diese Lungenerkrankung häufiger festzustellen bei Kindern mit Vitamin A Mangel?
- Wie verändert sich der Anteil an Kindern mit dieser Lungenkrankheit im Verlauf der Jugend?

Wieder sind Populationsdurchschnitte von Interesse. Zusatzmerkmale, wie Vitamin A Mangel, meist in binärer Form kodiert, werden dazu benötigt, um Sub-Popula-

tionen zu identifizieren, zu charakterisieren und in Kontrast zu setzen zu Populationen von Interesse.

3.1.3 Definition

- 'Marginal Models are natural analogues for correlated data of GLMs for independent data.' (Diggle, Liang und Zeger, 1994)

Laut dieser Definition bestehen keine großen Unterschiede zwischen 'normalen' generalisierten linearen Modellen und marginalen Modellen. Der einzige Unterschied besteht also darin, dass von nun an die 'within'-Korrelationsstruktur, d.h. die Korrelationsstruktur innerhalb eines Clusters (Person) zusätzlich modelliert werden muss. Daher erfolgt die Modellierung eines marginalen Modells in zwei separaten Teilen. Dies ist zum einen die Modellierung der 'Marginal Means', zum anderen die Modellierung der Kovarianzstruktur. Von primärem Interesse ist allerdings die Modellierung des Regressionsmodells bzw. der marginalen Mittelwerte μ_{it} mit

$$E(y_{it}|z_{it}) = \mu_{it} = h(z'_{it}\beta).$$

Nur von sekundären Bedeutung ist die Bestimmung der Kovarianzstruktur. Fahrmeir und Tutz (2001) beschreiben dies wie folgt:

- '... the scientific question of interest is to analyze the influence on the covariates,... whereas the spatial correlation ... is regarded as nuisance.'

Damit ist klar, dass die Korrelation nicht von Interesse ist, aber trotzdem in das Modell miteinbezogen und geschätzt werden muss. Die Korrelationsstruktur wird über die Score-Funktion, wie sich später zeigen wird, direkt mit den marginalen Wahrscheinlichkeiten in Verbindung gebracht (Aerts et al., 2002).

Diggle, Liang und Zeger (1994) definieren das marginale Modell, wie folgt. Hierbei zeigen sich Parallelen zur Definition des GLM aus Sektion 2.2.

- Der marginale Erwartungswert des Response, $E(y_{it}) = \mu_{it}$, ist abhängig von den Kovariablen x_{it} oder z_{it}[1] über eine Linkfunktion h mit $h(\mu_{it}) = z'_{it}\beta$.

[1] Vektor z setzt sich aus einer oder mehreren Kovariablen x zusammen.

- Die marginale Varianz ist abhängig vom marginalen Mean über die Relation $var(y_{it}) = \phi V(\mu_{it})$. V ist eine bekannte Varianzfunktion, der Dispersion-Parameter ϕ ist bekannt oder muss wie in Sektion 2.12 geschätzt werden.
- Die Korrelation $\text{corr}(y_{it}, y_{ik})$ ist eine Funktion vom marginalen Mean und eventuell eines zusätzlichen Parameters $\tilde{\alpha}$, so dass gilt $\text{corr}(y_{it}, y_{ik}) = \rho(\mu_{it}, \mu_{ik}, \tilde{\alpha})$, mit $\rho(\cdot)$ als bekannter Funktion (siehe auch Chaganty, 2004).

Bei genauer Betrachtung wird deutlich, dass die Regressionsparameter β und auch die 'Nuisance'-Parameter $\tilde{\alpha}$, über die gesamte (Sub)-Population gesehen, nicht variieren, d.h. β und $\tilde{\alpha}$ sind unabhängig von i und auch von t. Zum besseren Verständnis sei noch einmal erwähnt, dass nur Korrelationen innerhalb der Cluster zulässig sind, nicht zwischen den Clustern. Aufgrund dieser Tatsache ist es natürlich möglich, dass die Korrelationen bzw. die Korrelationsfunktion ρ abhängig vom Parameter i geschätzt wird, d.h. pro Cluster wird eine eigene Korrelationsstruktur geschätzt. Andererseits ist klar, dass auch gelten kann $\rho = \rho_i$, so dass eine generelle Kovarianzstruktur für die gesamte Population angenommen wird. Wird die Kovarianzstruktur nicht in das Modell miteinbezogen, kann dies zu erheblicher Unterschätzung der Varianzen führen, was überhöhte Teststatistiken zur Folge hätte (Aerts et al., 2002). Wie die Korrelationsfunktion ρ bzw. deren Parameter $\tilde{\alpha}$ geschätzt werden, zeigt sich in Kürze.

Die marginalen Regressionsparameter β sind genauso zu interpretieren wie die Koeffizienten einer Querschnittsstudie (Cross-Section), in der kein zeitlicher Verlauf untersucht wurde oder keine Cluster gebildet wurden. Besonders einfach ist die Interpretation der Regressionsparameter im logistischen Modell mit binärem x_{it} und $p = 1$.

- $logit(\mu_{it}) = log\frac{\mu_{it}}{1-\mu_{it}} = log\frac{Pr(y_{it}=1)}{Pr(y_{it}=0)} = \beta_0 + \beta_1 x_{it}$
- $var(y_{it}) = \mu_{it}(1 - \mu_{ij})$
- $corr(y_{it}, y_{ik}) = \rho(\tilde{\alpha}) = \tilde{\alpha}$

In diesem Modell gilt $\rho(\tilde{\alpha}) = \rho_i(\tilde{\alpha})$ und zusätzlich, dass die Korrelationen zwischen den einzelnen y_{it} und y_{ik} konstant auf einem Wert $\tilde{\alpha}$ liegen. Diese Struktur wird als

'Compound-Korrelationsstruktur' bzw. als 'Compound-Correlation' im Englischen bezeichnet.

Bei so vielen Gemeinsamkeiten zwischen marginalen Modellen und GLMs (natürlich gehören marginale Modelle zur Klasse der GLMs) ist es offensichtlich, dass auch die Maximum- bzw. Quasi-Likelihoodschätzungen in gewisser Weise strukturell ähnlich aufgebaut sind. Der nächste Abschnitt wird dies zeigen.

3.2 Independence Estimation Equations - IEE

Folgende Abschnitte beziehen sich zum Großteil auf die beiden Artikel von Liang und Zeger (1986) und Zeger und Liang (1986), die kurz nacheinander erschienen sind. Hauptaugenmerk liegt dabei zunächst auf erstgenanntem Artikel. Teil dessen sind auch die Independence Estimation Equation.

Der Unterschied zum GLM aus Kapitel 2 besteht nun darin, dass Liang und Zeger eine 'Working-Covariance-Matrix' einführen, die die Kovarianzstruktur in die Schätzung der Regressionsparameter miteingehen lässt. Somit erweitern sie den Begriff 'GLM' zu 'Working generalized linear model for the marginal distribution of y_{it}'. Die marginale Verteilung, d.h. die 'Randverteilung' von y_{it}, ist Basis für die Modellierung der 'marginal expectation', d.h. des marginalen Erwartungswerts $E(y_{it})$. Damit ist $E(y_{it})$ unabhängig von y_{it-1}, wie es bei bedingten Modellen der Fall wäre. Nur $E(y_{it})$ wird modelliert, nicht $E(y_{it}|y_{it-1})$. Identische x_{it} werden also genau demselben $E(y_{it})$ zugeordnet. Dies wäre bei bedingten Modellen nicht der Fall, da der Erwartungswert ja immer noch zusätzlich vom Vorgängerwert y_{it-1} abhängig ist. Diggle, Liang und Zeger (1994) schreiben hierzu:

- By marginal expectation, we mean the average response over the sub-population that shares a common value of x.

Falls der Response y_{it} multivariat normalverteilt ist reduzieren sich die Schätzer zu den bekannten Schätzern, basierend auf Pseudo-Likelihood-Ansätzen. Es wird in marginalen Modell also weder eine bedingte Verteilung benutzt, noch wird die

gemeinsame Verteilung der Messwiederholungen spezifiziert.

Nach Liang und Zeger (1986) sollte das Modell folgende Voraussetzungen erfüllen (vgl. Kapitel 2).

- Die marginale Verteilung sollte folgende Form annehmen können.

$$f(y_{it}) = exp[\frac{\{y_{it}\theta_{it} - b(\theta_{it}) + c(y_{it})\}}{\phi}]$$

- Zudem sollte gelten

$$\theta_{it} = h(\eta_{it}), \quad \eta_{it} = z'_{it}\beta.$$

- Die ersten beiden Momente seien gegeben durch

$$E(y_{it}) = b'(\theta_{it}), \quad var(y_{it}) = \phi b''(\theta_{it}).$$

Aus diesen drei Annahmen, die größtenteils mit den Annahmen zur ML-Schätzung übereinstimmen, werden dann die Scoregleichungen abgeleitet, unter der Annahme, dass die Messwiederholungen einer Person unabhängig voneinander sind. Daher wurde die Bezeichnung 'independence estimation equations' von Liang und Zeger eingeführt.

$$\begin{aligned} s_I(\beta) &= \sum_{i=1}^{n} z_i D_i [y_i - \mu_i(\beta)] \\ &= \sum_{i=1}^{n} z_i D_i I^{-1} [y_i - \mu_i(\beta)] \end{aligned} \tag{3.1}$$

Anstelle der inversen Varianz $\sigma^{-2}(\beta)$ aus Scoregleichung (2.17) tritt nun die Einheitsmatrix $I^{-1}_{T\times T} = I$, die vernachlässigt werden kann. $(z_i)_{p\times T}$ repräsentiert die Designmatrix und $(D_i)_{T\times T} = D_i(\beta) = \text{diag}(\partial h(z'_i\beta)/\partial\eta_{it})$. Der Schätzer β_I ist der $(p\times 1)$-Lösungsvektor der Scoregleichungen (3.1). Unter milden Regularitätskriterien ist β_I ein konsistenter Schätzer von β und $n^{\frac{1}{2}}(\hat{\beta}_I - \beta)$ ist asymptotisch multivariat normalverteilt für $n \to \infty$ mit Erwartungswert 0 und folgender Kovarianzmatrix in

Sandwichform.

Es sei nun $A_i = \text{diag}\{b''(\theta_{it})\}$ eine $T \times T$ Diagonalmatrix mit den Varianzen als Diagonalelementen. Damit ist

$$\begin{aligned} \hat{V}_i &= \hat{F}_1(\hat{\beta}_I)^{-1}\hat{F}_0\hat{F}_1(\hat{\beta}_I)^{-1} = \left(\sum_{i=1}^{n} z_i' D_i A_i D_i z_i\right)^{-1} \\ &\times \left(\sum_{i=1}^{n} z_i' D_i (y_i - \mu_i(\hat{\beta}))(y_i - \mu_i(\hat{\beta}))' D_i z_i\right)\left(\sum_{i=1}^{n} z_i' D_i A_i D_i z_i\right)^{-1} \end{aligned} \tag{3.2}$$

eine konsistente Schätzung der Kovarianzmatrix (vgl. Liang und Zeger, 1986, S. 15). Bei hoher Autokorrelation sind die Schätzer für β_I nicht sehr effizient. Gleiches gilt auch für die Kovarianzschätzung.

3.3 Generalized Estimation Equations - GEE-1

Liang und Zeger (1986) erweitern das Modell nun, indem Korrelationen in das Modell miteinbezogen werden und generieren konsistente Schätzungen für die Regressionsparameter β_{G1} und unter bestimmten Umständen auch für die Varianz. Die Annahme, dass die Fehler multivariat normalverteilt sind, wird wie auch bei den 'Independence Estimation Equations' beibehalten.

Generell wird eine Annahme für die Kovarianzstruktur zu Beginn der Modellierung gemacht, es spielt dabei keine Rolle, ob die Annahme korrekt ist oder nicht. Die Annahme wird dann in Matrixstruktur dargestellt, d.h. es wird eine symmetrische Matrix $R(\tilde{\alpha})$ eingeführt, die alle Eigenschaften einer Korrelationsmatrix erfüllt. Zudem wird $\tilde{\alpha}$ als $s \times 1$-Vektor definiert, der $R(\tilde{\alpha})$ vollständig charakterisiert. Damit enthält der Vektor s alle $\frac{T(T-1)}{2} \times 1$-Nebendiagonal-Elemente der Matrix $R(\tilde{\alpha})$. Diese wird im folgenden als 'Working-Correlation-Matrix' bezeichnet, das Gesamtkonstrukt

$$V_i = \phi A_i^{\frac{1}{2}} R(\tilde{\alpha}) A_i^{\frac{1}{2}} \tag{3.3}$$

als 'Working-Covariance-Matrix'. Die Korrelationsmatrix $R(\tilde{\alpha})$ wird also von beiden Seiten mit den Wurzeln der geschätzten Varianzen $A_i^{\frac{1}{2}} = \text{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_T)$, d.h. den

Standardabweichungen, multipliziert. Zusätzlich spielt der Dispersion Parameter ϕ eine Rolle[2]. Damit ergeben sich strukturell ähnliche Scoregleichungen. Es gelte

$$\begin{aligned} U_{G1}(\beta) &= \sum_{i=1}^{n} \tilde{D}_i' V_i^{-1}(y_i - \mu_i(\beta)) = \\ &= \sum_{i=1}^{n} z_i' D_i' A_i V_i^{-1}(y_i - \mu_i(\beta)) = \\ &= \sum_{i=1}^{n} z_i' D_i' A_i (A_i^{\frac{1}{2}} R(\tilde{\alpha}) A_i^{\frac{1}{2}})^{-1}(y_i - \mu_i(\beta)) = \\ &= \sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta} (A_i^{\frac{1}{2}} R(\tilde{\alpha}) A_i^{\frac{1}{2}})^{-1}(y_i - \mu_i(\beta)), \end{aligned} \tag{3.4}$$

mit $\tilde{D}_i' = z_i' D_i' A_i$ und D_i wie in Gleichung (2.19). Es ist offensichtlich, dass im Falle $R(\tilde{\alpha}) = I$ sich die Generalized Estimation Equations, kurz 'GEE-1', zu den Independence Estimation Equations reduzieren würden. Beim Vergleich mit den Quasi-Likelihood-Scoregleichungen (2.33) von Wedderburn fällt auf, dass in der Varianz bei den GEE-1 ein zusätzlicher Parameter $\tilde{\alpha}$ enthalten ist.

- Quasi-Likelihood: $var(y) = \phi V_{QL}(\mu) = \phi V_{QL}(\mu(\beta))$ [y univariat]
- GEE-1: $var(y) = \phi V_{G1}(\mu, \tilde{\alpha}) = \phi V_{G1}(\mu(\beta), \tilde{\alpha})$ [y multivariat]

Die Indizes wurden hier ohne Einschränkung der Allgemeingültigkeit kurzfristig vernachlässigt, genauso wie der Dispersion-Parameter in Gleichung (3.4), da er bei der Schätzung von β, wie erwähnt, keine Rolle spielt.

Zusätzliche Parameter in der Varianz führen dazu, dass die Likelihood nicht mehr vollständig spezifiziert ist. Daher müssen diese Parameter ersetzt werden durch konsistente Schätzungen. Parameter $\tilde{\alpha}$ wird, gemäß Liang und Zeger (1986), ersetzt durch $\hat{\alpha}(y, \beta, \phi)$, ein $n^{\frac{1}{2}}$-konsistenter Schätzer, falls β und ϕ bekannt sind. Der Dispersion-Parameter ϕ wird ersetzt durch $\hat{\phi}(y, \beta)$, ein $n^{\frac{1}{2}}$-konsistenter Schätzer,

[2]Der Dispersion Parameter wird im Originaltext als die Inverse der bisher bekannten Dispersion Parameters ϕ definiert. Allerdings wird in dieser Arbeit die gewohnte Definition beibehalten.

falls β bekannt ist (vgl. Sektion 2.12). Damit lässt sich (3.4) formulieren als Summe der individuellen Scorefunktionen U_i

$$\sum_{i=1}^{n} U_i[\beta, \hat{\alpha}(\beta, \hat{\phi}(\beta))]_{G1} = 0. \tag{3.5}$$

Die Lösung von (3.5) ist $\hat{\beta}_{G1}$. Die Schätzung der Kovarianzmatrix in Sandwich-Form entspricht exakt der Schätzung der Kovarianzmatrix der IEE, so dass gilt

$$\begin{aligned} \hat{V}_i &= \hat{F}_1(\hat{\beta}_{G1})^{-1}\hat{F}_0\hat{F}_1(\hat{\beta}_{G1})^{-1} = \left(\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} \tilde{D}_i\right)^{-1} \\ &\times \left(\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1}(y_i - \mu_i(\hat{\beta}))(y_i - \mu_i(\hat{\beta}))' V_i^{-1} \tilde{D}_i\right) \left(\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} \tilde{D}_i\right)^{-1}. \end{aligned} \tag{3.6}$$

Es sei angefügt, dass die Kovarianzmatrix unter den Annahmen gebildet wurde, dass die Anzahl der Cluster $n \to +\infty$ geht und die Anzahl an individuellen Messungen als konstanter Wert angesehen werden. Diese Annahmen entsprechen den natürlichen Rahmenbedingungen für Longitudinalstudien, in denen meist viele Fälle, aber wenig Wiederholungen gemessen werden (z.B. 250 Teilnehmer bei vier Messungen in der 'Indonesian Childrens Health Study').

Auch in diesem Zusammenhang ist die Konsistenz der Schätzung nur dann gesichert, falls der Erwartungswert des Modells korrekt spezifiziert ist. Allerdings ist es von Vorteil, dass die Konsistenz der Schätzungen nicht von der korrekten Wahl von $R(\tilde{\alpha})$ abhängt. Falls $\tilde{\alpha}$ und ϕ $n^{\frac{1}{2}}$-konsistent geschätzt werden, ist die asymptotische Varianz aus (3.6) unabhängig von beiden Parametern.

3.4 Schätzalgorithmus für die Regressionsparameter β

Die Standardprozedur zum iterativen Schätzen der IEE und GEE-1 wird bei Aerts et al. (2002) genau beschrieben. Sie gleicht in etwa der Prozedur aus Sektion 2.9:

1. Berechne die Startschätzer für die Regressionsparameter β, beispielsweise über ein univariates GLM mit Einheitsmatrix als Kovarianzmatrix.

2. Berechne alle Parameter, die gebraucht werden, um ein Update von β zu berechnen.

   - Berechne die Pearson-Residuen (vgl. nächster Abschnitt)
   - Berechne die Schätzer für $\tilde{\alpha}$ (vgl. nächster Abschnitt)
   - Berechne $R(\tilde{\alpha})$
   - Berechne eine Schätzung für ϕ
   - Berechne daraus die Working-Kovarianz-Matrix pro Cluster (Person)

$$V_i(\beta, \tilde{\alpha}) = \phi(A_i^{\frac{1}{2}} R(\tilde{\alpha}) A_i^{\frac{1}{2}})^{-1}$$

3. Berechne ein Update von β (vgl. (2.26))

$$\begin{aligned} \beta^{(k+1)} &= \beta^{(k)} + \left(\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} \tilde{D}_i \right)^{-1} \left(\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} (y - \mu_i) \right) \\ &= \beta^{(k)} + \hat{F}_1(\hat{\beta}^{(k)})^{-1} s(\hat{\beta}^{(k)}) \end{aligned} \tag{3.7}$$

Iteriere solange bis ein Threshold unterschritten wird, beispielsweise Threshold ε aus Sektion 2.9.

3.5 Möglichkeiten zur Schätzung von $\tilde{\alpha}$

3.5.1 Momentenschätzer

Liang und Zeger (1986) verwenden Momentenschätzer, um die Kovarianzparameter $\tilde{\alpha}$ zu schätzen. Für die vorgeschlagenen Kovarianzstrukturen wird je ein anderer Momenten-Schätzer abgeleitet. Ist der oder sind die notwendigen Parameter geschätzt, können sie an den passenden Stellen der Matrix $R(\tilde{\alpha})$ eingetragen werden.

Die Hauptdiagonale enthält nur Einsen, da die Varianzparameter (zunächst) separat geschätzt werden. Zur Bildung der Momentenschätzer müssen in den meisten Fällen die Pearson-Residuen

$$\hat{r}_{it} = \frac{y_{it} - \mu_{it}}{var(\mu_{it})^{\frac{1}{2}}} = \frac{y_{it} - b'(\hat{\theta}_{it})}{b''(\hat{\theta}_{it})^{\frac{1}{2}}}$$

berechnet werden, wobei $'$ und $''$ die jeweiligen Ableitungen angeben. Damit ergeben sich folgende Möglichkeiten zur Schätzung der Parameter für die Working-Kovarianz-Matrix $R(\tilde{\alpha})$.

1. Independence: $corr(y_{it}, y_{it'}) = 0 \qquad (t \neq t')$
   Es werden keine Parameter geschätzt.

2. Tridiagonale Working-Kovarianz: Sei $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{T-1})'$ und $\tilde{\alpha}_t = corr(y_{it}, y_{i,t+1})$ für $t = 1, \ldots, T-1$. Dann folgt

   $$\hat{\alpha}_t = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\hat{r}_{it}\hat{r}_{i,t+1}}{n - \tilde{p}}$$

   ist ein Schätzer für $\tilde{\alpha}_t$ und $\tilde{p}$ ist die Anzahl der Modellparameter. Falls gelten soll $\tilde{\alpha} = \tilde{\alpha}_t$ und damit die Nebendiagonalelemente alle gleich sind, ergibt sich folgende Schätzung

   $$\hat{\alpha} = \sum_{t=1}^{T-1} \frac{\hat{\alpha}_t}{T-1}.$$

3. Exchangeable oder Compound-Struktur: Es gelte $corr(y_{it}, y_{it'}) = \tilde{\alpha}$ für alle $t \neq t'$. Daraus ergibt sich

   $$\hat{\alpha} = \frac{1}{\phi \frac{n}{2} T(T-1) - \tilde{p}} \sum_{i=1}^{n} \sum_{t>t'} \hat{r}_{it}\hat{r}_{it'}. \tag{3.8}$$

   Formel (3.8) verändert sich, falls die Clustergröße variiert (vgl. Liang und Zeger, 1986).

4. AR(1)-Struktur: $corr(y_{it}, y_{i,t+s}) = \tilde{\alpha}^s \ (s = 1, \ldots, T-t)$

   - Vorschlag von Aerts et al. (2002):

$$\hat{\alpha} = \frac{1}{n\phi} \sum_{i=1}^{n} \frac{1}{T-1} \sum_{t \leq T-1} \hat{r}_{it} \hat{r}_{i,t+1}$$

   - Vorschlag von Liang und Zeger (1986) für stetige Daten:
     Regression von $log(\hat{r}_{it}\hat{r}_{i,t+s})$ auf $log(|t-s|)$. Es entstehen Schwierigkeiten bei negativen Residualprodukten wegen des Logarithmus.
   - Vorschlag von Toutenburg (1992) für Zeitreihen:

$$\hat{\alpha} \quad = \quad \frac{\sum_{t=2}^{T} \hat{r}_{it}\hat{r}_{i,t-1}}{(\sum_{t=2}^{T} \hat{r}_{it}^2)^{\frac{1}{2}} (\sum_{t=2}^{T} \hat{r}_{i,t-1}^2)^{\frac{1}{2}}}$$

5. unstrukturierte Working-Kovarianz: $corr(y_{it}, y_{it'}) = \tilde{\alpha}_{tt'} \ (t \neq t')$

$$\hat{\alpha}_{tt'} = \frac{1}{n\phi} \sum_{i=1}^{n} \hat{r}_{it}\hat{r}_{it'}$$

6. unspezifizierte Working-Kovarianz: $R = R(\tilde{\alpha})$

$$R = \frac{1}{n\phi} \sum_{i=1}^{n} A_i^{-\frac{1}{2}} (y_i - \mu_i(\hat{\beta}))(y_i - \mu_i(\hat{\beta}))' A_i^{-\frac{1}{2}}$$

Bei Liang und Zeger (1986) wird das Verhalten von drei dieser Working-Kovarianz-Matrizen in zwei kleinen Simulationen studiert. In einem weiteren Artikel aus dem Jahr 1986 erweitern Zeger und Liang (1986) die Working-Kovarianz-Theorie für Quasi-Likelihoodansätze. Bisher galt immer die Normalverteilungsannahme, nun wird diese fallengelassen, d.h. die Fehler folgen einer Verteilung aus einer einparametrischen Exponentialfamilie. Allerdings ändert dies nichts an den Schätzgleichungen. Trotzdem werden sie in diesem Artikel nochmals zusammengefasst. Wiederum wird auf die genannten günstigen Eigenschaften der Schätzmethode hingewiesen und

nochmals betont, dass auch bei Fehlspezifikation der Working-Kovarianz konsistente Schätzungen erzielt werden. Effizienzgewinne entstehen durch korrekte Spezifikation der Working-Kovarianz. Es sei noch erwähnt, dass die Momentenschätzung aus heutiger Sicht wohl teilweise den Pseudo-Likelihood-Ansätzen zuzuordnen wären, weil in manchen Fällen zwei oder mehr Varianzparameter geschätzt werden müssen (vgl. Gong und Samaniego, 1981, Breslow, 1990, Nelder, 2000).

Zusätzlich wird eine neue Kovarianzstruktur namens 'stationary m-dependence' eingeführt, die bekannt ist aus 'Random-Effects'-Ansätzen (Zeger und Liang, 1986). Es gilt

$$(R_i)_{tt'} = \begin{cases} \tilde{\alpha}^{|t-t'|}, & |t-t'| \leq m \\ 0, & |t-t'| > m \end{cases} .$$

Wird die absolute Differenz der Zeitpunkte t und t' größer als ein bestimmter Wert m, wird eine Korrelation von 0 angenommen. Dies ist eine Verallgemeinerung der AR(1)-Schätzung. Die Schätzung von $\tilde{\alpha}$ läuft daher wohl gleich ab.

Davidian (1995) bezeichnet die oberhalb definierten Gleichungen als 'linear estimation equations'. Im folgenden Abschnitt wird die Theorie nun auf quadratische Schätzgleichungen ausgeweitet. Diese werden im Englischen als 'quadratic estimation equations' (Davidian, 1995) oder 'second order estimation equations' (Aerts et al., 2002) bezeichnet. Der Unterschied zu den linearen Gleichungen liegt darin, dass von nun anstelle der Momentenschätzungen ein weiteres, zusätzliches GLM für die Kovarianzstruktur modelliert wird. Beide Modelle werden gemeinsam geschätzt. Der Ansatz wird in der Literatur als 'joint modelling'-Ansatz bezeichnet. In einem weiteren Schritt erfolgt die gemeinsame Modellierung der Varianzen zusammen mit der Kovarianzstruktur. Prentice legte die Grundlagen hierfür im Jahre 1988 und erweiterte seine Theorien 1990 in Zusammenarbeit mit Zhao.

3.5.2 Schätzung für binäre Responses nach Prentice

Speziell für binäre Responses erweiterte Prentice (1988) die Theorie zur Schätzung der Working-Kovarianz. Er geht dabei von einem ähnlichen Ansatz wie Pregibon (1984) aus, der den Dispersion-Parameter mittels eines zusätzlichen GLMs schätzt. Die Parameter β und $\tilde{\alpha}$ werden also in zwei inhaltlich getrennten GLMs gemeinsam geschätzt, wobei im ersten GLM die marginalen Response-Wahrscheinlichkeiten und im zweiten GLM die paarweisen Korrelationen modelliert werden. Es wird also ein zusätzliches GLM angefügt, das die Korrelationen $\tilde{\alpha}$ in einem $s \times 1$ Responsevektor mit $\frac{T(T-1)}{2} \times 1$ Elementen zusammenfasst. Hierbei spielt die Reihenfolge, in der die Korrelationen angeordnet sind keine Rolle, allerdings hat sich folgende Notation durchgesetzt (lexikographische Ordnung).

$$\begin{aligned} u_i &= (\tilde{\alpha}_{i12}, \tilde{\alpha}_{i13}, \ldots, \tilde{\alpha}_{i1T}, \tilde{\alpha}_{i23}, \ldots, \tilde{\alpha}_{i,T-1,T})'_{\frac{T(T-1)}{2}\times 1} \\ u &= (u_1', \ldots, u_n')' \end{aligned}$$

Die 'sample correlation' u (Stichprobenkorrelation), Response-Vektor des Zusatz-GLMs, wird im binären Fall wie folgt geschätzt.

$$u_{itt'} = u_{itt'}(\beta) = \frac{(y_{it} - \mu_{it})(y_{it'} - \mu_{it'})}{(\mu_{it}(1-\mu_{it})\mu_{it'}(1-\mu_{it'}))^{\frac{1}{2}}}.$$

Neue Scoregleichungen werden aufgestellt.

$$\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1}(y_i - \mu_i) = 0 \tag{3.9}$$

$$\sum_{i=1}^{n} \tilde{E}_i' W_i^{-1}(u_i - \delta_i) = 0. \tag{3.10}$$

Gleichung (3.9) entspricht genau der ersten Scoregleichung der GEE-1 (3.4), Gleichung (3.10) wurde von Prentice neu definiert mit den Nebendiagonal-Elementen der Working-Korrelations-Matrix $R(\tilde{\alpha})$ als neuem Responsevektor u_i und Mean-Vektor δ_i. Gleichung (3.10) beinhaltet also die Regression über die $n\frac{T(T-1)}{2}$ Nebendiagonal-Elemente $\tilde{\alpha}_{tt'}$, $t \neq t'$, der Working-Kovarianz-Matrix $R(\tilde{\alpha})$. Es gilt somit für den

Meanvektor δ_i

$$\begin{aligned} \delta_i &= (\delta_{i12}, \delta_{i13}, \ldots, \delta_{i1T}, \delta_{i23}, \ldots, \delta_{i,T-1,T})'_{\frac{T(T-1)}{2}\times 1} \\ \delta &= (\delta_1', \ldots, \delta_n')'_{\frac{nT(T-1)}{2}\times 1} \\ \delta_{itt'} &= E(u_{itt'}) \end{aligned} \tag{3.11}$$

Außerdem gilt

$$\begin{aligned} \tilde{E}_i &= \frac{\partial \delta_i}{\partial \tilde{\alpha}} \\ W_i &= \text{diag}(w_{i12}, w_{i13}, \ldots, w_{i23}, \ldots, w_{i,T-1,T})_{\frac{T(T-1)}{2}\times\frac{T(T-1)}{2}}. \end{aligned}$$

Die Nebendiagonal-Elemente der Working-Kovarianz-Matrix $R(\tilde{\alpha})$ werden im zusätzlichen GLM als Mean modelliert. Von nun an gibt es also zwei Working-Kovarianz-Matrizen, die eine, V_i, enthält die Parameter $\tilde{\alpha}$ auf den Nebendiagonalen, die andere, W_i, enthält die Kovarianzstruktur der Parameter $\tilde{\alpha}$. Die Working-Kovarianz-Matrix W_i hat im Falle der Unabhängigkeit Diagonalgestalt mit $\frac{T(T-1)}{2}$ Elementen auf der Hauptdiagonalen. Andere Working-Kovarianz-Matrizen sind vorstellbar, Prentice gibt aber nicht genauer an, wie diese geschätzt werden. Eine Möglichkeit sind sicher Momentenschätzer. Ansonsten stimmen beide GLMs überein. Die Diagonal-Elemente der Working-Kovarianz W_i entsprechen den Varianzen $var(u_{itt'}) = w_{itt'}$. Diese werden im binären Fall über eine vom Erwartungswert δ_i abhängige Funktion geschätzt, mit

$$w_{itt'} = 1 + (1-2\mu_{it})(1-2\mu_{it'})[(\mu_{it}(1-\mu_{it})\mu_{it'}(1-\mu_{it'}))^{-\frac{1}{2}}]\delta_{itt'} - \delta^2_{itt'}.$$

Die Matrix $\tilde{E}$ entspricht der Matrix $\tilde{D}$, während die Parameter δ_i den Means μ_i entsprechen. Eine Zusammenfassungen der Scoregleichungen in Matrizenform ist in Gleichung (3.12) dargestellt.

$$\sum_{i=1}^{n} \begin{pmatrix} \tilde{D}_i' & 0 \\ 0 & \tilde{E}_i' \end{pmatrix} \begin{pmatrix} V_i^{-1} & 0 \\ 0 & W_i^{-1} \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i(\beta) \\ u_i - \delta_i(\tilde{\alpha}) \end{pmatrix} = 0 \tag{3.12}$$

Gleichung (3.12) lässt sich zur bekannten Score-Notation umformen

$$\sum_{i=1}^{n} G_i(\xi)\tilde{V}_i^{-1}(\xi)(s_i(\xi) - m_i(\xi)) = 0, \tag{3.13}$$

wobei $\xi = (\beta', \tilde{\alpha}')'$, $G_i(\xi) = \text{diag}(\tilde{D}_i', \tilde{E}_i')$, $\tilde{V}_i^{-1}(\xi) = \text{diag}(V_i^{-1}, W_i^{-1})$, $s_i(\xi) = (y_i', u_i')'$ und $m_i(\xi) = (\mu_i(\beta)', \delta_i(\tilde{\alpha})')'$ ist. Obwohl die Gleichungen im Vergleich zu den Momentenschätzern nun nicht mehr als linear sondern als quadratisch bezeichnet werden, ist die Lösung der Scoregleichung (3.13) immer noch ein GEE-1-Schätzer (Davidian, 1995). Auch die Miteinbeziehung der Varianzen (vgl. folgende Matrizen) in das Modell mit u_i als Response ändert daran nichts. Es verlängert sich lediglich der Vektor u_i um T Elemente von $\frac{T(T-1)}{2} \times 1$ zu $\frac{T(T+1)}{2} \times 1 = T + \frac{T(T-1)}{2} \times 1$. Folgende Matrizen sollten mehr Aufschluss darüber geben. Beide Matrizen enthalten nur die relevanten Informationen, daher auch die Dreiecksform. In der rechten Matrix sind zusätzlich die Varianzen $\tilde{\alpha}_{i11}, \ldots, \tilde{\alpha}_{iTT}$ auf der Hauptdiagonalen zu erkennen.

$$\begin{pmatrix} 1 & \tilde{\alpha}_{i12} & \tilde{\alpha}_{i13} & \tilde{\alpha}_{i14} & \ldots & \tilde{\alpha}_{i1T} \\ & 1 & \tilde{\alpha}_{i23} & \tilde{\alpha}_{i24} & \ldots & \tilde{\alpha}_{i2T} \\ & & 1 & \tilde{\alpha}_{i34} & \ldots & \tilde{\alpha}_{i3T} \\ & & & 1 & \ldots & \tilde{\alpha}_{i4T} \\ & & & & \ddots & \vdots \\ & & & & & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} \tilde{\alpha}_{i11} & \tilde{\alpha}_{i12} & \tilde{\alpha}_{i13} & \tilde{\alpha}_{i14} & \ldots & \tilde{\alpha}_{i1T} \\ & \tilde{\alpha}_{i22} & \tilde{\alpha}_{i23} & \tilde{\alpha}_{i24} & \ldots & \tilde{\alpha}_{i2T} \\ & & \tilde{\alpha}_{i33} & \tilde{\alpha}_{i34} & \ldots & \tilde{\alpha}_{i3T} \\ & & & \tilde{\alpha}_{i44} & \ldots & \tilde{\alpha}_{i4T} \\ & & & & \ddots & \vdots \\ & & & & & \tilde{\alpha}_{iTT} \end{pmatrix}$$

Damit lässt sich u_i erweitern zu

$$u_i = (\tilde{\alpha}_{i11}, \tilde{\alpha}_{i12}, \tilde{\alpha}_{i13}, \ldots, \tilde{\alpha}_{i1T}, \tilde{\alpha}_{i22}\tilde{\alpha}_{i23}, \ldots, \tilde{\alpha}_{iTT})'_{\frac{T(T+1)}{2} \times 1}. \tag{3.14}$$

Die Schätzung verläuft exakt gleich, mit dem Unterschied, dass nun noch T zusätzliche Parameter über die um T Zeilen erweiterte Scoregleichung (3.13) geschätzt werden. Algorithmisch wird das Minimierungsproblem so umgesetzt, dass abwechselnd ein Parameter β oder $\tilde{\alpha}$ geschätzt wird, während jeweils der andere Parameter festgehalten wird. Es wird gemäß dem Algorithmus aus Sektion 3.4 solange iteriert, bis ein Threshold ε unterschritten wird. Eine Möglichkeit zur Generalisierung des Modells ist gegeben. So müssen nur die 'Off-Block-Diagonal'-Elemente in (3.12) durch Matrizen, die ungleich Null sind, ersetzt werden.

Ein Review zu binären GEE-1-Schätzern findet man bei Fitzmaurice, Laird und Nan (1993). Auch GEE-2 Ansätze werden darin behandelt. Genaueres dazu im nächsten Teilabschnitt.

3.6 Generalized Estimation Equations - GEE-2

Der Begriff 'GEE-2' wurde von Liang, Zeger und Qaqish (1992) eingeführt. Die zugehörigen GEE-2-Gleichungen wurden von Zhao und Prentice (1990) zwei Jahre zuvor veröffentlicht, allerdings nur für binären Response. Die Erweiterung des binären Ansatzes hin zu multivariat-diskreten und kontinuierlichen Responses ist beschrieben bei Prentice und Zhao (1991). Kategoriale Daten wurden erstmals im Jahre 1992 von Liang, Zeger und Qaqish behandelt. Eine sehr gute Zusammenfassung des Themas GEE-2 findet man bei Fitzmaurice, Laird und Rotnitzky (1993) für binäre Longitudinal-Daten und auch bei Davidian (1995) in Kapitel 14. In diesem Kapitel wird der Unterschied von GEE-1 zu GEE-2 anhand ausführlicher Kommentare und Darstellungen verdeutlicht. Prinzipiell besteht der Unterschied darin, dass nun die Korrelationsmatrix $\tilde{V}_i^{-1}(\xi)$ erweitert wird, so dass auch Korrelationen der beiden GLMs untereinander zulässig sind, und zudem eine 'Gradienten Matrix des Means' B_i (Davidian, 1995) in $G_i(\xi)$ eingeführt wird (vgl. (3.12), (3.13)). Damit ergeben sich die Score-Gleichungen

$$\sum_{i=1}^{n} \begin{pmatrix} \tilde{D}_i' & B_i' \\ 0 & \tilde{E}_i' \end{pmatrix} \begin{pmatrix} V_i^{-1} & C_i \\ C_i' & W_i^{-1} \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i(\beta) \\ u_i - \delta_i(\tilde{\alpha}) \end{pmatrix} = 0, \tag{3.15}$$

wobei $B_i = \partial\delta_i(\tilde{\alpha})/\partial\beta$ und $C_i = cov(y_i, u_i)$. Es ist offensichtlich, dass nun Annahmen über die dritten bzw. vierten Momente in das Modell einfließen müssen. Die ersten und zweiten Momente sind der Mean μ_i bzw. die Varianzschätzung, aus der die Working-Kovarianz-Matrix V_i gebildet wird. Die Varianzschätzung für V_i erfolgt im zweiten GLM über den Mean δ_i, den dritten Moment. Er ist sozusagen der marginale Mean der Varianzparameter $\tilde{\alpha}$, die in u_i als Response zusammengefasst werden. Der vierte Moment wird bestimmt über die Varianzschätzungen des zweiten GLMs, zusammengefasst in der Working-Kovarianz-Matrix W_i. Von entscheidender Bedeutung ist die richtige Spezifikation zumindest der ersten beiden Momente. Allerdings sollten die Working-Kovarianz W_i, genauso wie die Kovarianz C_i zwischen beiden GLMs auch korrekt spezifiziert werden. Denn falls alle Momente korrekt spezifiziert würden, würden die Scoregleichungen (3.15) eine optimale Lösung für die Schätzer finden. Allerdings ist es nicht einfach, sondern nahezu unmöglich, alle Momente kor-

rekt zu spezifizieren. Man betrachte zunächst die ursprüngliche Working-Assumption für die Kovarianz-Matrix V_i. Falls Fehlspezifikationen schon in diesen Bereich auftreten – und dies ist sehr wahrscheinlich – kann man nicht davon ausgehen, dass höhere Momente korrekt spezifiziert sind.

3.7 Diskussion

Es besteht also die Möglichkeit zur Wahl zwischen linearen und quadratischen Schätzungen. Wenn die Kovarianz-Matrix V_i nicht korrekt spezifiziert ist, erhält man im Fall der linearen Schätzung unverzerrte Schätzungen. Falls die Varianzen und die Korrelationsstruktur zusammen geschätzt werden (vgl. (3.14)), wird eine Fehlspezifikation der Kovarianzmatrix V_i sehr wahrscheinlich. Die quadratischen Gleichung sind aber nur dann unverzerrt, solange die Kovarianz V_i korrekt spezifiziert ist. Falls der Analytiker also Vertrauen hat in die richtige Spezifikation der Kovarianz V_i, eröffnen die quadratischen Gleichungen eine Alternative zu den linearen Gleichungen. Allerdings müssen dann noch die Matrizen C_i bzw. W_i korrekt spezifiziert werden. Falls dies so ist, erhält man einen effizienteren Schätzer als im linearen Lösungsansatz. Bei Fehlspezifikation von C_i oder W_i bleiben die Schätzungen unverzerrt, allerdings ist es schwer zu sagen, ob dies immer noch zu einem Effizienzgewinn gegenüber dem linearen Ansatz führt. Trotzdem liegt das Hauptaugenmerk insgesamt immer noch auf den Schätzungen der marginalen Means.

Problematisch am quadratischen Ansatz ist, dass die Schätzer nicht unverzerrt sind, falls die Kovarianz V_i falsch spezifiziert wurde. Inkonsistente Schätzungen sind die Folge. Generell ist es also ein Risiko quadratische Schätzgleichungen anzuwenden, weil man nicht weiß, ob die Kovarianz-Matrix fehlspezifiziert ist. Daher ist es grundsätzlich sicherer zur allgemeinen Anwendung GEE-1 Schätzungen zu gebrauchen. Zusammenfassend kann gesagt werden, dass GEE-2 Schätzer eher theoretisches Potential haben, um zu zeigen, dass unter bestimmten Voraussetzungen die Effizienz der Schätzer noch verbessert werden kann im Vergleich zu anderen Schätzern. Im täglichen Gebrauch sind wohl einfachere, sogar bei partieller Fehlspezifikation

mit Sicherheit unverzerrte Schätzungen erwünscht. Allein die Tatsache, dass keine Statistik-Software GEE-2 Schätzungen berechnen kann, zeigt dies.

Allerdings bringt auch das Gesamtkonstrukt 'GEE' und die damit verbundene (Quasi)-Likelihood-Theorie, so wie sie Gilmour, Anderson und Rae (1985), Liang und Zeger (1986) und Zeger und Liang (1986) definieren, umfangreiche Nachteile mit sich. Diese wurden in einem Artikel von Lindsey und Lambert (1998) in Kürze zusammengefasst.

1. Bekanntlich setzt sich die Kovarianzmatrix aus den Ableitungen der Scorefunktion s nach Parameter β zusammen. Hierbei wird im Log-Likelihood- sowie im Quasi-Likelihoodansatz vorausgesetzt, dass die Kovarianzmatrix symmetrisch ist und gleichzeitig positiv definit ist, so dass nach McCullagh und Nelder (1989) für $i, j = (1, \ldots, n)$, $i \neq j$, gelten soll

$$\frac{\partial s_i(\beta)}{\partial \beta_j} = \frac{\partial s_j(\beta)}{\partial \beta_i}, \tag{3.16}$$

   d.h. die Elemente der Kovarianzmatrix über der Hauptdiagonalen entsprechen den Elementen unter der Hauptdiagonalen. Allerdings gilt dies nur für ganz bestimmte Kovarianzmatrizen und Verteilungsformen, verallgemeinern lässt sich diese Aussage nicht (Lindsey und Lambert, 1998, McCullagh und Nelder, 1989). Wird also wie im Quasi-Likelihoodansatz die Verteilungsannahme vernachlässigt, kann es vorkommen, dass (3.16) nicht mehr gültig ist. Eine eindeutige Rücktransformation der Kovarianzmatrix zur Scorefunktion und damit auch zur Likelihoodfunktion ist nicht mehr gegeben. Damit würden die Definitionen statistischer Modelle allgemein verletzt werden, so wie sie in Kapitel 2 beschrieben wurden. Ein Beispiel dazu findet man bei McCullagh und Nelder (1989) auf den Seiten 336-339.

2. Auch wenn die Anforderungen aus Punkt 1 erfüllt sind und die marginalen Verteilungen in der Exponentialfamilie sind, ist die gemeinsame Verteilung nur in Einzelfällen auch in der Exponentialfamilie enthalten. Damit unterscheiden

sich die Quasi-Likelihoodansätze nicht nur bezüglich der Kovarianzstruktur (vgl. (2.30)), sondern auch bezüglich der marginalen Verteilungen von Log-Likelihoodansätzen, obwohl die Schätzgleichungen doch sehr ähnlich wirken.

3. In allen GEE-Ansätzen wird der Mean modelliert. Allerdings ist es bei der Modellierung des Means von entscheidender Bedeutung, wie er sich bei Veränderungen der Kovariaten verhält. Diese Veränderung wird durch die Kovarianzmatrix angegeben. Ist diese nicht bekannt, können die Schätzer nicht korrekt interpretiert werden.

4. Da die GEE-Ansätze nicht mehr von Likelihoodfunktionen ausgehen, sondern von Quasi-Likelihoodfunktionen, wurde die gesamte Inferenzstatistik verändert. Daher ist es fraglich, ob die gesamte Likelihoodtheorie eins zu eins auf Quasi-Likelihoodansätze übertragbar ist.

5. Es ist nicht mehr möglich Modelle über Kriterien, wie z.B. dem Akaike Information Criterion (AIC) zu vergleichen, weil keine Likelihoodfunktion mehr vorhanden ist. Nur noch Modellvergleiche über zusätzliche Regressionsterme sind möglich.

6. Es existiert keine Wahrscheinlichkeitsverteilung zu der allgemein Integrale formuliert werden könnten, die eine Rücktransformation zur Likelihood ermöglichen (vgl. Punkt 1)

7. Die Eigenschaft, dass die Schätzungen konsistent bleiben bei Fehlspezifikation der Kovarianzmatrix, solange der Mean korrekt spezifiziert ist, ist irrelevant. Denn kein Modell kann als korrekt angenommen werden.

8. Spezielle Problem ergeben sich für Longitudinal-Daten mit zeitlich abhängigen Kovariaten (vgl. Pepe und Anderson, 1994).

Die Modellierung von höheren Momenten (GEE-2) löst diese Art von Problemen nicht, weil auch diese Momente marginaler Art sind.

3.8 Gemeinsame Kovarianz-Matrix

Wie es so oft ist im mathematisch/statistischen Bereich, können gewisse Gleichungen als Spezialfälle anderer Gleichungen dargestellt werden. In diesem Fall kann die GEE-1 Schätzung als Spezialfall der GEE-2 Schätzung mit $B_i = \partial\delta_i(\tilde{\alpha})/\partial\beta = 0$ und $C_i = cov(y_i, u_i) = 0$ aufgefasst werden. Nach trivialen Umformungen verändern sich die Scoregleichungen von (3.15) zu (3.12). Dies gilt auch für die Schätzung der gemeinsamen Kovarianz-Matrix von $\hat{\beta}$ und $\hat{\tilde{\alpha}}$. Allgemein gelte für die GEE-2 Kovarianz-Matrix

$$\hat{V} = n \begin{pmatrix} \Phi_1 & 0 \\ \Phi_2 & \Phi_3 \end{pmatrix} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \begin{pmatrix} \Phi_1 & \Phi_2' \\ 0 & \Phi_3 \end{pmatrix}. \tag{3.17}$$

Matrix (3.17) schätzt die Kovarianzmatrix konsistent. Sie ist asymptotisch normalverteilt mit Erwartungswert 0 und setzt sich zusammen aus

$$\begin{aligned}
\Phi_1 &= \left(\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} \tilde{D}_i \right)^{-1}, \\
\Phi_2 &= \left(\sum_{i=1}^{n} \tilde{E}_i' V_i^{-1} \tilde{E}_i \right)^{-1} \left(\sum_{i=1}^{n} \tilde{E}_i' W_i^{-1} \frac{\partial u_i}{\partial \beta} \right) \left(\sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} \tilde{D}_i \right)^{-1}, \\
\Phi_3 &= \left(\sum_{i=1}^{n} \tilde{E}_i' V_i^{-1} \tilde{E}_i \right)^{-1}, \\
\Lambda_{11} &= \sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} cov(y_i) V_i^{-1} \tilde{D}_i, \\
\Lambda_{12} &= \sum_{i=1}^{n} \tilde{D}_i' V_i^{-1} cov(y_i, u_i) W_i^{-1} \tilde{E}_i, \\
\Lambda_{22} &= \sum_{i=1}^{n} \tilde{E}_i' W_i^{-1} cov(u_i) W_i^{-1} \tilde{E}_i, \\
\Lambda_{12} &= \Lambda_{21}.
\end{aligned}$$

Es ist offensichtlich, dass die einzelnen Matrizen nach Ende der Iterationen ausgewertet werden an der Stelle $(\beta, \tilde{\alpha})$ und dass die Kovarianzen $cov(y_i)$, $cov(u_i)$ und

$cov(y_i, u_i)$ geschätzt werden, wie es Liang und Zeger (1986) vorschlagen. Damit gilt

$$\begin{aligned} cov(y_i) = var(y_i) &= E(y_i - \mu_i)(y_i - \mu_i)', \\ cov(y_i, u_i) &= E(y_i - \mu_i)(u_i - \delta_i)', \\ cov(u_i) = var(u_i) &= E(u_i - \delta_i)(u_i - \delta_i)'. \end{aligned}$$

Um nun Gleichung (3.17) auf GEE-1 Niveau zu reduzieren, werden nur die Matrizen Φ_2 und Λ_{12} bzw. Λ_{21} gleich Null gesetzt. Damit ergibt sich für die GEE-1 Kovarianz-Matrix

$$\hat{V} = n \begin{pmatrix} \Phi_1 & 0 \\ 0 & \Phi_3 \end{pmatrix} \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix} \begin{pmatrix} \Phi_1 & 0 \\ 0 & \Phi_3 \end{pmatrix}. \tag{3.18}$$

Kapitel 4

Nicht-Parametrische Regressionsansätze

Im folgenden Kapitel wird die bisher lineare Modellstruktur etwas gelockert. Dabei geht man vom linearen Modell $y_i = \alpha + \beta x_i$ bzw. generalisierten linearen Modell $E(y_i) = \eta_i$ über zur nichtlinearen Kleinst-Quadrat-Schätzung mittels nicht-parametrischer penalisierter Schätzmethoden. Es wird zunächst von folgendem Zusammenhang ausgegangen

$$y_i = \gamma(x_i) + \varepsilon_i,$$

wobei $\gamma(x_i)$ eine beliebige glatte Funktion ist. Die Fehler ε_i sollten einen Erwartungswert von 0 aufweisen und zumindest approximativ normalverteilt sein.

Sehr viel Literatur ist zum Thema nicht-parametrischer Regressionsansätze veröffentlicht worden. Nur das wichtigste sei hier genannt. Dies ist zum einen das Buch 'Generalized Additive Models' von Hastie und Tibshirani (1990), das in den ersten Kapiteln die Möglichkeiten zur Glättung von Kurven zusammenfasst und Teile dieser Ansätze in den weiteren Kapiteln zunächst auf einfache lineare Modelle, dann auf GLMs und Generalisierte Additive Modelle (GAM) anwendet. Härdle (1990) beschäftigt sich vor allem mit nicht-parametrischen Ansätzen für stetige Responses,

während Green und Silverman (1994) sehr detailliert auf Regression mittels verschiedener Arten von Splines, Kernel-Methoden und Roughness-Penalties in Verbindung mit GLMs eingehen. Fahrmeir und Tutz (2001) fassen kurz verschiedenste Ansätze zur nicht-parametrischen Schätzmethoden inklusive Penalisierung zusammen.

4.1 Grundlagen

Im Kapitel 'Smoothing' ganz am Anfang ihres Buches erläutern Hastie und Tibshirani (1990) die Grundlagen und Ideen, die hinter nicht-parametrischen Regressionsansätzen mit Penalisierung stecken. Zunächst definieren sie, was Glätten (Smoothing) ist:

- A smoother is a tool for summarizing the trend of a response variable .... It produces an estimate of the trend that is less variable than Y itself.

Dabei geben sie zwei Hauptziele von Glättungsmethoden an:

1. Beschreibung der Daten: Die geschätzte Kurve verbessert die Darstellung der Daten im Vergleich zum Scatterplot

2. Die geschätzte Kurve ist ein Maß für die Abhängigkeit des Responses y von den Prädiktoren x.

Ganz ähnlich sehen dies auch Green und Silverman (1994), allerdings gehen sie nur kurz darauf ein und widmen ihre Aufmerksamkeit eher den Roughness Penalties für nicht-parametrische Regressionsansätze. Die nicht-parametrische Natur von Schätzern wird wie folgt beschrieben (Hastie und Tibshirani, 1990):

- An important property of a smoother is its nonparametric nature: it doesn't assume a rigid form for the dependence of y on $x_1, \ldots, x_p$.

Damit ist der Unterschied zwischen parametrischen und nicht-parametrischen Regressionsansätzen klar aufgezeigt. Während im linearen parametrischen Modell ein strikter linearer Zusammenhang angenommen wird, werden im nicht-parametrischen

Ansatz keine Annahmen über den Verlauf der Kurve gemacht, falls gewisse Voraussetzungen wie z.B. Stetigkeit der Kurve erfüllt sind.

Grundsätzlich kann man nicht-parametrische Regression als eine Art 'Local Averaging'-Prozedur (lokale Durchschnittsbildung) bezeichnen. Falls die Prädiktoren x_i beispielsweise in Kategorien gemessen wurden, wie z.B. Geschlecht (binär) oder Autotyp (kategoriell), ist der Prozess des Local Averaging sehr einfach zu bewältigen, indem der Mittelwert der jeweiligen Kategorie bzgl. des stetigen Responses y zur Schätzung berechnet wird (vgl. Abbildung 4.1). Bei stetigen Daten sind im Normal-

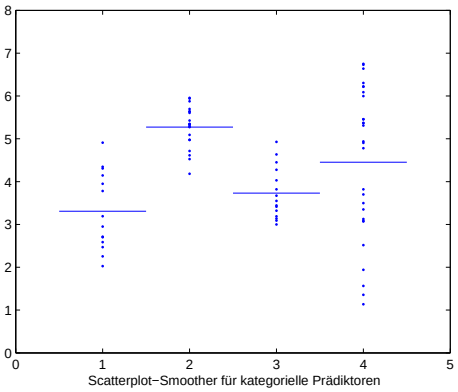

Abbildung 4.1: Beispiel für 'Local-Averaging'.

fall nicht mehr so viele oder gar keine Wiederholungen pro Prädiktor vorhanden wie bei kategorieller Datenstruktur. Daher wird nun ein Bereich pro Datenpunkt definiert, in dem der Local-Average, d.h. ein lokaler Schätzer in einen Bereich um den Datenpunkt, gebildet wird. Dieser Bereich wird im englischen als 'Neighborhood' von x_i bezeichnet. Folgende Fragestellungen sind nun von Interesse:

1. Wie bildet man den Local-Average in der Neighborhood?
2. Wie gross wählt man die Neighborhood?

Dabei bezieht sich die erste Fragestellung darauf, welche Art von Schätzer zum Local-Averaging verwendet wird, die zweite Fragestellung, welche Punkte in der

Umgebung von x_i mit welchem Gewicht ≥ 0 in die Schätzung miteingehen. Wählt man die Neighborhood größer kann man davon ausgehen, dass die Schätzung glatter wird und somit die Varianz der Schätzung abnimmt, aber sich der Bias erhöht, während bei kleinen Neighborhoods das Gegenteil der Fall ist, mit großen Varianzen und kleinem Bias. Einige der einfachsten Schätzmethoden sollen nun im Anschluss kurz dargestellt werden, später wird auf kompliziertere Ansätze eingegangen.

4.2 Einfache Schätzmethoden - Lokale Schätzer

4.2.1 Bin-Smoother

Der Bin-Smoother ist die Erweiterung der Schätzers aus Abbildung 4.1 für stetige Prädiktoren. Dabei werden die Prädiktoren in K Kategorien eingeteilt. Dann werden die Responses y pro Kategorie gemittelt und der resultierende Mittelwert wird als Schätzer für diese Kategorie angesehen. Im Plot erscheint der Mittelwert als waagrechte Linie mit Jumps (Unstetigkeitsstellen) am Anfang und Ende der jeweiligen Kategorie (vgl. Abbildung 4.1).

4.2.2 Running-Mean

Jeweils k Punkte links und rechts vom Datenpunkt x_i (symmetric nearest neighborhood) gehen in die Schätzung mit ein. Dabei wird einfach der Mittelwert über alle Punkte in der Neighborhood gebildet. Für alle x_i setzt sich die Neighborhood aus unterschiedlichen Datenpunkten zusammen.

Der Running-Mean-Smoother zeigt generell eine starke lokale Variabilität und hat einen starken Bias.

4.2.3 Running-Line

Um den Bias des Running-Mean-Smoothers etwas zu verringern, kann der Running-Line-Smoother eingesetzt werden. In jeder Neighborhood wird nun anstelle des Mit-

telwerts eine KQ-Schätzung berechnet. Die resultierende Gerade wird als Schätzer angesehen.

4.2.4 Running-Median

Der Running-Mean-Smoother wird einfach durch den Median ersetzt. Damit ist die Schätzung unanfälliger gegen Ausreißer.

4.2.5 LOESS

Auch der Running-Line-Smoother kann sehr stark hin-und her schwanken. Um dies zu verhindern, können zusätzlich zur KQ-Methode noch Gewichte eingeführt werden, die Datenpunkte nahe x_i und auch x_i selbst mit hohen Gewichten belegen, während weiter entfernte Datenpunkte mit sehr kleinen Gewichten oder einem Gewicht von 0 belegt werden. Man bezeichnet diese Methode 'Locally Weighted Running-Line Smoother (LOESS)'. Diese wurde im Jahre 1979 von Cleveland veröffentlicht. Zusätzlich werden nach Cleveland (1979) noch Gewichte für die Residuen eingeführt, so dass Ausreißern in der Schätzung deutlich weniger Gewicht zugewiesen wird. Damit gewichtet Cleveland sowohl die Punkte um x_i in der Neighborhood als auch die Residuen $y_i - \hat{y}_i$. Da die Residuen zu Beginn der Analyse unbekannt sind, wird zunächst eine Weighted Least Squares Regression durchgeführt und die Residuen bestimmt, dann werden die Residuen je nach Abstand zur Schätzung gewichtet und das Modell nochmals mit beiden Gewichtungen gerechnet. Allgemein werden Gewichte / Gewichtsfunktionen W nach einem bestimmten Prinzip ausgewählt. Die Gewichtsfunktionen müssen dabei folgende Eigenschaften erfüllen:

1. $W(x) > 0$, für $|x| < 1$
2. $W(-x) = W(x)$
3. W fällt ab $x \geq 0$
4. $W(x) = 0$ für $|x| \geq 1$

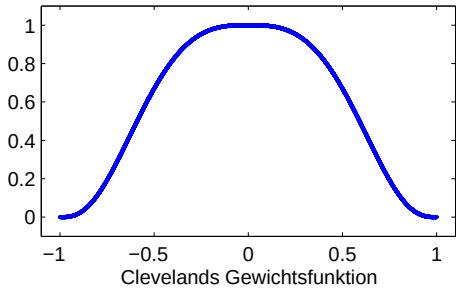

Abbildung 4.2: Beispiel: Gewichtsfunktion (Cleveland, 1979)

Cleveland gibt als Beispiel eine kubische Funktion an (vgl. Abbildung 4.2), die ausschließlich positiv und symmetrisch ist und nur auf einem bestimmten Intervall definiert ist.

$$W(x) = \begin{cases} (1 - |x|^3)^3 & \text{falls } |x| < 1 \\ 0 & \text{sonst} \end{cases}$$

Sowohl die Residuen als auch die Prädiktoren werden mit Funktionen dieser Art gewichtet. Damit ist die Basis für die Kern-Dichte-Schätzungen, die im Englischen als Kernel-Estimators bezeichnet werden, gelegt.

4.3 Kern-Dichte-Schätzung / Kernel-Estimation

Kernel-Schätzer werden bei Hastie und Tibshirani (1990) und Fahrmeir und Tutz (2001) kurz angesprochen, bei Härdle (1990) erfolgt eine detaillierte Beschreibung des Themas.

Gemäß Clevelands Definition gelten die vier oben genannten Eigenschaften auch für Kernel, hinzu kommt noch, dass gilt

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

d.h. der Kernel $K(u)$ ist definiert als kontinuierliche, symmetrische, begrenzte Funktion, deren Fläche sich zu 1 integriert. Damit lässt sich die Weight-Sequenz für

eindimensionale Prädiktoren x_i definieren als

$$W_i(x) = \frac{K_\lambda(x - x_i)}{\hat{f}_\lambda(x)} \tag{4.1}$$

wobei

$$\hat{f}_\lambda(x) = \frac{1}{n}\sum_{i=1}^{n} K_\lambda(x - x_i) \tag{4.2}$$

und

$$K_\lambda(u) = \frac{1}{\lambda}K(\frac{u}{\lambda}). \tag{4.3}$$

Gleichung (4.1) gibt das Gewicht an, mit dem Datenpunkt x_i multipliziert wird. Die Größe des Gewichts ist abhängig von der Differenz eines beliebigen Punktes x_i zu x. Je weiter beide Punkte voneinander entfernt sind, desto kleiner ist das Gewicht. Die Größe des Gewichts wird mittels der Kernelfunktion $K_\lambda(u)$ bestimmt. Die Funktion $\hat{f}(\cdot)$ aus (4.1) und (4.2) wird als Normierungsterm oder 'Rosenblatt-Parzen kernel density estimator' bezeichnet (Rosenblatt, 1956, Parzen, 1962). Sie gleicht die Gewichte der lokalen Intensität der Prädiktoren an und garantiert damit, dass sich die Gewichte in der Neighborhood zu 1 aufsummieren. Funktion $\hat{f}(x)$ hängt damit von der Anzahl an Punkten in der Neighborhood ab. Die Anzahl an Punkten in der Neighborhood, d.h. die Punkte, die mit Gewichten > 0 belegt werden, wird durch die Bandbreite λ bestimmt. Die Bandbreite λ geht in Gleichung (4.3) direkt in die Kernelfunktion $K(u)$ mit ein. Es ist offensichtlich, dass bei grosser Bandbreite mehr Datenpunkte Gewichte > 0 besitzen als bei kleiner Bandbreite. Bekannte Kernelfunktionen sind der Epanechnikov-Kernel (Epanechnikov, 1969, Bartlett, 1963)

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{falls } |u| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

und der Minimum-Varianz-Kernel (hauptsächlich bei Splines angewendet wegen der negativen Gewichte)

$$K(u) = \begin{cases} \frac{3}{8}(3 - 5u^2) & \text{falls } |u| \leq 1 \\ 0 & \text{sonst.} \end{cases}$$

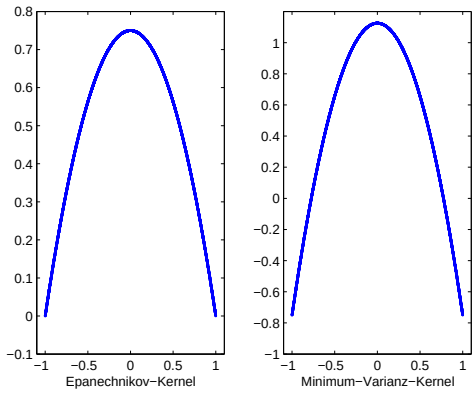

Abbildung 4.3: Beispiel: Epanechnikov- und Minimum-Varianz-Kernel

Abbildung 4.3 zeigt beide Kernel. Die Schätzung des nichtlinearen Zusammenhangs zwischen x und y erfolgt über den Nadaraya-Watson-Schätzer (Nadaraya, 1964, Watson, 1964)

$$\hat{\gamma}_\lambda(x) = \sum_{i=1}^{n} \frac{K_\lambda(x - x_i)}{\sum_{k=1}^{n} K_\lambda(x - x_k)} y_i. \tag{4.4}$$

Beide Autoren gehen dabei vom einfachsten polynomischen Fit aus, einem Fit eines Polynoms nullten Grades, d.h. einer Konstante α_0. In der zugehörigen Literatur wird ein Fit mit einer Konstante als 'Average Kernel Estimation' bezeichnet (Lin und Carroll, 2000). Es gilt somit für das Minimierungsproblem mit Gewichtung über den Kernel $K_\lambda(u)$

$$\sum_{i=1}^{n} (y_i - \alpha_0)^2 K_\lambda(x - x_i) \to \min_{\alpha_0}.$$

Nach Ableiten und Nullsetzen führt dies zum linearen Schätzer

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

mit

$$s(x, x_i) = \sum_{i=1}^{n} \frac{\frac{1}{n\lambda} K_\lambda(x - x_i)}{\frac{1}{n\lambda} \sum_{k=1}^{n} K_\lambda(x - x_k)} = \sum_{i=1}^{n} \frac{K_\lambda(x - x_i)}{\sum_{k=1}^{n} K_\lambda(x - x_k)}. \tag{4.5}$$

Man vergleiche (4.5) mit (4.4). Polynomiale Anpassung höherer Ordnung lassen sich äquivalent ableiten. Allgemein gilt folgende Minimierungsfunktion für Kern-Dichte-Schätzer

$$\sum_{i=1}^{n}(y_i - \alpha_0 - \sum_{l=1}^{q}\alpha_l(x_i - x)^q)^2 K_\lambda(x - x_i) \rightarrow \min_{\alpha_0,\alpha_l}. \tag{4.6}$$

Falls $q = 1$, wird die Schätzung als 'local linear', d.h. als lokal lineare Kernel-Estimation bezeichnet, weil im linearen Modell nur ein Intercept α_0 und die Steigung der Gerade α_1 lokal, d.h. begrenzt durch die Kernelfunktion in Kombination mit der Bandbreite, geschätzt werden. Natürlich sind auch lokale Polynome höherer Ordnung zulässig ($q = 2, 3, \ldots$). Allerdings wird in den Standardliteratur größtenteils auf 'average' bzw. 'local linear' Kernel-Estimation eingegangen.

Für multidimensionale Prädiktorvariablen $x_{i1}, \ldots, x_{ip}$ kann die multidimensionale Produkt-Kernel-Funktion

$$W_i(x) = \frac{\prod_{j=1}^{p} K_\lambda(x_j - x_{ij})}{\hat{f}_\lambda(x)}$$

angewandt werden, wobei im Normierungsterm auch ein Produkt-Kernel verwendet wird. Die anschließende Schätzung der Kurve läuft exakt gleich ab wie im eindimensionalen Fall.

4.3.1 Kernel-Estimation und Longitudinaldaten

In der vergleichsweise neuen statistischen Literatur zum Thema generalisierter nicht-parametrischer Regression in Kombination mit Kernelschätzern und Longitudinaldaten ging es in den letzten Jahren vor allem darum, die Korrelationsstruktur so in die Kernelschätzung miteinzubeziehen, dass konsistente und effiziente Schätzer entstehen. Dies gelang erst im Jahre 2003 in einem Artikel von Wang (2003). Zunächst hatten Lin und Carroll (2000) überraschenderweise festgestellt, dass es die beste Strategie sei, beim Modellieren die Korrelationsstruktur zu vernachlässigen und die Daten so zu behandeln als wären sie unabhängig. Miteinbeziehung der Struktur

würde zu einer ineffizienteren Schätzung führen und keine $n^{\frac{1}{2}}$-konsistenten Schätzer ergeben. Lin und Carroll gehen dabei von einem generalisierten nichtparametrischen Modell aus, das auf longitudinale Datenstrukturen und Kovariaten auf 'Observation-Level'-Basis, d.h. Abhängigkeit der Kovariaten x von i und t, mit und ohne Messfehler gleichermaßen eingeht. Es sei dabei y_{it} der Response und x_{it} die Kovariaten, die eventuell mit einem Messfehler m_{it} behaftet sind. Allerdings spielt der Messfehler in den folgenden Ausführungen eine eher untergeordnete Rolle. Es gelten die üblichen Modell-Voraussetzungen

$$g(\mu_{it}) = \gamma(x_{it})$$

und

$$\begin{aligned} E(y_{it}|x_{it}) &= \mu_{it}, \\ var(y_{it}|x_{it}) &= \phi_t w_{it}^{-1} V(\mu_{it}), \end{aligned}$$

wobei $\gamma(\cdot)$ eine unbekannte glatte Funktion ist, $g(\cdot)$ eine differenzierbare Linkfunktion ist, $V(\cdot)$ eine Varianzfunktion ist, ϕ_t ein zeit- bzw. cluster-abhängiger Scaleparameter ist und w_{it} ein Gewicht darstellt. Die GEE-Theorie wird nun auf die nichtparametrische Kernel-Schätzung für Longitudinaldaten einfach übertragen. Dabei muss Gleichung (4.6) in Matrixschreibweise überführt werden, resultierend in einer leicht veränderten Version von Gleichung (3.4)

$$\sum_{i=1}^{n} G_{i(q)}(x)' D_i(x) V_i(x)^{-1} K_{i(\lambda)}(x) \left(y_i - \mu_i(x)\right) = 0. \tag{4.7}$$

Der Term $(G_{i(q)})_{T\times q}$ tritt nun an die Stelle der bisherigen Designmatrix z_i'. Er enthält die Einzelteile der Polynome mit

$$\begin{aligned} G_{i(q)}(x) &= (G_{(q)}(x_{i1} - x), \ldots, G_{(q)}(x_{iT} - x))', \\ G_{(q)}(z) &= (1, z, z^2, \ldots, z^q)'. \end{aligned}$$

Zusätzlich werden die Daten lokal gewichtet über die Kernelmatrix $K_{i(\lambda)}(x)$, wobei λ wieder die Bandbreite darstellt. Die Matrix $K_{i(\lambda)}(x)$ ist eine Diagonalmatrix mit

$$K_{i(\lambda)}(x) = \mathrm{diag}(K_\lambda(x_{i1} - x), \ldots, K_\lambda(x_{iT} - x)).$$

Die Einträge der Matrix sind normalisierte Kernelfunktionen, definiert wie in (4.3). Die restlichen Terme sind genau wie in Gleichung (3.4) definiert. Es ist offensichtlich, dass auch im Falle der Kernel-Estimation eine Zerlegung der Working-Kovarianzmatrix erfolgt mit $V_i = A_i^{\frac{1}{2}} R(\tilde{\alpha}) A_i^{\frac{1}{2}}$. Die Matrix D_i ist definiert als

$$D_i = (D_i)_{T\times T} = \operatorname{diag}\left(\mu^{(1)}(G_{(q)}(x_{it})'\alpha_K)\right),$$

wobei $\mu^{(1)}(\cdot)$ die erste Ableitung von $\mu(\cdot)$ nach $\alpha_K = (\alpha_0, \alpha_1, \ldots, \alpha_q)'$ angibt. Zudem gilt allgemein und speziell im lokal-linearen Fall

$$\mu_{it} = \mu(\gamma(x_{it})) = \mu(G_{(q)}(x_{it})'\alpha_K) = \mu\left(\alpha_0 + \alpha_1 \frac{(x_{it} - x)}{\lambda}\right)_{LL}. \tag{4.8}$$

Im letzten Teil von (4.8) wurden die Vektoren $G_{(q)}$ und α_K für den lokal-linearen Fall zerlegt. Bei der average-kernel-estimation würde nur $\mu(\alpha_0)_{AK}$ übrig bleiben. Eine andere Möglichkeit, Gleichung (4.7) zu formulieren, zeigt Gleichung (4.9).

$$\sum_{i=1}^{n} G_{i(q)}(x)' D_i(x) \left(K_{i(\lambda)}^{\frac{1}{2}}(x) V_i(x)^{-1} K_{i(\lambda)}^{\frac{1}{2}}(x)\right) (y_i - \mu_i(x)) = 0 \tag{4.9}$$

Hierbei wird die Kernelmatrix $K_{i(\lambda)}(x) = K_{i(\lambda)}^{\frac{1}{2}}(x) K_{i(\lambda)}^{\frac{1}{2}}(x)$ in einer ähnlichen Weise wie die Working-Kovarianzmatrix aufgespalten und jeweils zu gleichen Teilen links und rechts an diese angefügt. Die restlichen Terme bleiben unverändert. Falls $R(\tilde{\alpha}) = I$, entsprechen sich beide Gleichungen (4.7) und (4.9). Nun werden die Gleichungssysteme mittels Fisher-Scoring aus Kapitel 2.9 gelöst. Man beachte, dass die Größe des Lösungsvektors α_K abhängig ist von der Wahl von q. Bei 'average kernel estimation' erhält man nur ein Skalar α_0 als Lösung, während man bei 'local linear estimation' zwei Parameter α_0 und α_1 erhält. Damit hat der Lösungsvektor die Größe $q \times 1$. Die Herleitung der Sandwich-Varianz-Schätzungen findet man bei Lin und Carroll (2000). Im weiteren Verlauf dieses Artikels wird nun auf die asymptotische Theorie beide Schätzer detailliert eingegangen und bewiesen, dass im Unterschied zum parametrischen GEE-Schätzer die effizientesten Schätzungen in diesem Falle für $R(\tilde{\alpha}) = I$, d.h. Working-Independence entstehen. Dies gilt sowohl für 'average-' als auch für 'local linear estimation'. Falls allerdings die Kovariaten auf 'Cluster-Level'-Basis in das Modell mit eingehen, d.h.

$$g(\mu_{it}) = \gamma(x_i),$$

sind die Schätzer effizient und $n^{\frac{1}{2}}$-konsistent, wenn die Working-Kovarianz der wahren Korrelationsmatrix entspricht (Lin und Carroll, 2001a). Schätzen dieses Modells mit Working-Independence würde die Effizienz des Schätzers verringern.

4.3.2 Miteinbeziehen der Korrelationsstruktur

Wang (2003) und auch Lin et al. (2003) erweitern das Modell aus (4.7) bzw. (4.9), so dass die Within-Cluster-Variation von nun an ohne Effizienzverlust in die Kernel-Schätzung miteinbezogen werden kann. Sie gehen dabei von folgender Idee aus: Sobald ein Datenpunkt x_{ij} innerhalb der Bandbreite $\pm\lambda$ eines beliebigen Punktes x liegt und daher zur Schätzung von $\gamma(x)$ verwendet wird, wird nun nicht nur Punkt x_{it} sondern auch die restlichen $T-1$ Punkte innerhalb des Clusters i verwendet, um $\gamma(x)$ zu schätzen. Allerdings gehen die restlichen Punkte nicht voll, sondern über deren Residuen in die Schätzung mit ein. Die Residuen lassen sich über die Differenz der Responses y und deren geschätzten Mittelwerten $\mu(\gamma(x))$ bilden. Offensichtlich stehen die Residuen nicht vor der Schätzung fest und daher schlägt Wang (2003) vor das Modell zunächst mit $R(\tilde{\alpha}) = I$ (Working-Independence) zu schätzen und daraus die Residuen zu berechnen. In den folgenden Iterationen werden die jeweiligen zuletzt gebildeten Schätzungen von $\gamma(x)$ zur Berechnung der Residuen verwendet. Der Schätzer von $\gamma(x)$ wird als $\hat{\gamma}(x)$ bezeichnet und geht direkt in die Schätzgleichung mit ein. Im zweiten und auch allen folgenden Schritten wird das Modell nun mit Within-Cluster-Variation geschätzt. Im Grunde genommen ändert sich bis auf die Designmatrix nicht viel im Vergleich zu (4.7). Damit lässt sich das verbesserte Modell wie folgt schreiben.

$$\sum_{i=1}^{n}\sum_{t=1}^{T} K_\lambda(x_{it}-x)\mu^{(1)}\left(\alpha_0+\alpha_1(x_{it}-x)\right) G^*_{iT(q)}(x)' \times V_i^{-1}\left(y_i-\mu^*_{ij}(x_{it},\alpha_K,\hat{\gamma}(x_{it}))\right) \tag{4.10}$$

Die Matrix $G^*_{iT(q)}$ hat die Dimensionen $T \times (q+1)$ und enthält nur Nullen außer an der t-ten Zeile mit $(1,(x_{it}-x),\ldots,(x_{it}-x)^q)$. Matrix $G^*_{it(q)}$ garantiert, dass nur Datenpunkt x_{it} und die restlichen $T-1$ Residuen (über die nichtdiagonale Working-Kovarianz-Matrix V_i) innerhalb des Clusters i in die Schätzung miteingehen. Die

Schätzung muss für alle Datenpunkte innerhalb des Clusters i separat berechnet werden, weil für jeden Datenpunkte in Cluster i die Residuen unterschiedlich sind. Daher wird auch die zusätzliche Summe $\sum_{t=1}^{T}$ in das Modell miteingefügt. Offensichtlich ist daher auch, dass $\mu_{ij}^{*}\left(x_{it}, \alpha_K, \hat{\gamma}(x_{it})\right)$ an dem jeweiligen Punkt $t = (1, \ldots, T)$ angepasst werden muss, so dass gilt

$$\mu_{ij}^{*}\left(x_{it}, \alpha_K, \hat{\gamma}(x_i)\right) = \left(\hat{\gamma}(x_{i1}), \ldots, \hat{\gamma}(x_{i,t-1}), \sum_{f=0}^{q}(\alpha_K)_f(x_{it}-x)^f, \hat{\gamma}(x_{i,t+1}), \ldots, \hat{\gamma}(x_{iT})\right)'.$$

Nur am Punkt x_{it} wird die gewöhnliche Schätzung, bekannt aus (4.9), durchgeführt. Die restlichen Punkte $x_{i1}, \ldots, x_{i,t-1}$ und $x_{i,t+1}, \ldots, x_{i,T}$ gehen über die Residuen und die nicht-diagonale Gestalt der Working-Kovarianz V_i in das Modell mit ein. $K_\lambda(x_{it}-x)$ und $\mu^{(1)}\left(\alpha_0 + \alpha_1(x_{it}-x)\right)$ sind in diesem Fall skalar, während die restlichen Matrizen und Vektoren von der Größe her entsprechend (4.9) gebildet werden. Auch in diesem Fall erhält man einen q-dimensionalen Lösungsvektor. Im weiteren Verlauf des Artikels geht Wang (2003) auf die asymptotischen Eigenschaften und die Effizienz detailliert ein. Es zeigt sich, dass durch die Einführung einer Korrelationsstruktur die Varianz reduziert wird und kein zusätzlicher Bias entsteht. Dies führt zum Effizienzgewinn und damit auch zu effizienten und konsistenten Schätzungen. Bei Lin et al. (2003) wird diese Form der Kernel-Estimation dann generell mit Smoothing-Splines verglichen.

'Piecewise polynomial fits', d.h. polynomische Fits in vielen kleinen aneinandergereihten Intervallen, sollen im Anschluss genauer besprochen werden.

4.4 B-Splines

B(asis)-Splines wurden von de Boor (1977, 1978) vorgestellt. De Boor geht dabei in seinem Buch 'A practical guide to Splines' nicht nur auf die Theorie ein, sondern auch darauf wie sich die B-Splines algorithmisch umsetzen lassen. Unter anderem beschäftigen sich Dierckx (1993), Stoer (1999) und Marx und Eilers (1992, 1996, 1998) auch eingehend mit demselben Thema. Von Dierckx wurden neben Kurven

auch dreidimensionale Oberflächendiagramme (Surfaceplots) mittels B-Splines gefittet, Stoer betrachtet die numerischen Eigenschaften von B-Splines und Marx und Eilers führten penalisierte B-Splines ein. Sie bezeichnen diese als P-Splines, 'Penalized Splines'. Dabei präsentieren sie in ihren Artikeln aus dem Jahre 1996 und 1998, die B-Splines-Theorie, zusammengefasst auf wenigen Seiten und erweitern diese Theorie bis hin zur Penalisierung. Die Penalisierung erfolgt über Differenzen-Penalties.

4.4.1 Definition B-Splines

B-Splines werden der Klasse der 'Regression-Splines' zugeordnet. Regressions-Splines bauen darauf auf, den Wertebereich von x in Intervalle zu zerlegen und innerhalb der Intervalle Polynomstücke an die Daten anzupassen. Zunächst wird der Wertebereich der Prädiktoren x_i also in K meist gleich grosse Intervalle zerlegt. Die Punkte an denen zwei Intervalle aneinander angrenzen werden als Knotenpunkte, im englischen als 'Nodes' oder 'Knots', bezeichnet. Die Knotenpunkte werden als ξ_k bezeichnet, so dass für die Intervalle I gilt

$$I_1 = [\xi_0, \xi_1], I_2 = [\xi_1, \xi_2], \ldots, I_K = [\xi_{K-1}, \xi_K], \quad k = 0, \ldots, K$$

mit

$$\xi_0 < \xi_1 < \ldots < \xi_K.$$

B-Splines sind, ähnlich Kernelfunktionen, nur in einem gewissen Bereich, beispielsweise zwischen zwei Knotenpunkten, verschieden von Null. Sie setzen je nach Grad des B-Splines aus 1, 2, 3 und mehr Polynom-Stücken zusammen. Abbildung 4.4 zeigt B-Splines verschiedenen Grades. Ganz oben in Abbildung 4.4 sind B-Splines 0-ten Grades zu erkennen. Sie entsprechen Indikatorfunktionen und bestehen genau aus einem Polynom-Stück, einem Polynom 0-ten Grades, d.h. einer Konstante. Zwischen zwei Knotenpunkten - in diesem Fall zwischen ξ_0 und ξ_1 - nimmt der Spline also einen Wert von 1 an, ansonsten 0. Ein Spline 0-ten Grades erstreckt sich genau über ein Intervall $[\xi_k, \xi_{k+1}]$, ein Spline 1. Grades hat Dreiecksform und erstreckt sich hingegen über genau zwei Intervalle. Damit entstehen, wie man rechts in Abbildung 4.4 erkennen kann, erste Überschneidungen zwischen Splines. Ein Spline 1. Grades

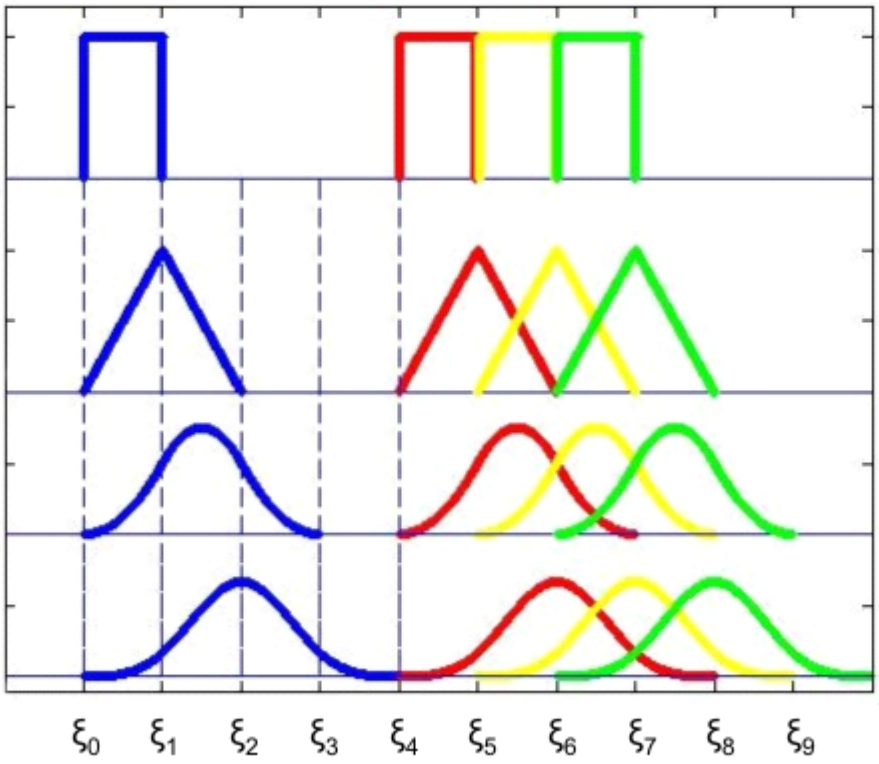

Abbildung 4.4: B-Splines verschiedenen Grades. (linke Seite, von oben nach unten: B-Spline 0-ten Grades, B-Spline 1. Grades, B-Spline 2. Grades, B-Spline 3. Grades. Rechte Seite: sich überlagernde B-Splines von Grad 0 bis 3.)

besteht aus genau zwei Polynom-Stücken 1. Grades, eines liegt zwischen $[\xi_0, \xi_1]$ linear aufsteigend, das zweite erstreckt sich von $[\xi_1, \xi_2]$ linear absteigend. Splines 2. Grades bestehen aus drei Polynom-Stücken, je eines zwischen zwei Knotenpunkten. Der Spline 2. Grades erstreckt sich über genau 3 Intervalle mit vier Knotenpunkten, der Spline 3. Grades genau über 4 Intervalle mit 5 Knotenpunkten, mit jeweils einem Polynom-Stück zwischen den Knoten. In diesen vier Intervallen ist der Spline außer an den Randpunkten von Null verschieden, an den Randpunkten - ξ_0 und ξ_4 im Fall des Splines 3. Grades - beginnt bzw. endet der Spline. Bei Punkt ξ_4 beginnt

zugleich ein neuer Spline (vgl. Abbildung 4.4), so dass keine Unstetigkeitsstellen auftreten. Dasselbe gilt natürlich auch für Punkt ξ_0. Hier kann ein Vorgänger-Spline enden, falls der Wertebereich dementsprechend gewählt wurde. Zusammenfassend kann man daher sagen, dass Splines höheren Grades sich über mehr Intervalle erstrecken, dafür nimmt die Höhe der Splines mit zunehmenden Anzahl an Graden ab.

Damit lassen sich folgende Eigenschaften von Splines verallgemeinern und ableiten. Ein B-Spline des Grades q

- besteht aus $q + 1$ Polynom-Stücken des Grades q, d.h. ein Polynom-Stück befindet sich zwischen zwei Knoten.
- besteht aus $q + 1$ Polynom-Stücken, die an den inneren Knoten miteinander verbunden sind.
- hat kontinuierliche Ableitungen vom Grad $q - 1$ an den inneren Knoten.
- ist größer als Null in einem Bereich von $q + 2$ Knoten.
- überschneidet sich außer an den Randpunkten mit $2q$ Polynom-Stücken der Nachbar-Splines (sich überschneidende Splines).
- ist nur an den Randpunkten gleich Null. Bei gegebenem x sind $q+1$ B-Splines ungleich Null.

Nachdem der Wertebereich von x in Intervalle mit Knoten ξ_k eingeteilt wurde, werden nun die Basisfunktionen, bspw. 2. Grades, angefügt. Gewichte δ_j, $j = 1, \ldots, m$ bestimmen die Höhe des Basisfunktion $B_j(x)$. Aus der gewichteten Summe der Basisfunktionen wird die Schätzung der Kurve gebildet. Die Gewichte sind die zu schätzenden Parameter, d.h. es gilt

$$\hat{\gamma}(x) = \sum_{j=1}^{m} \hat{\delta}_j B_j(x).$$

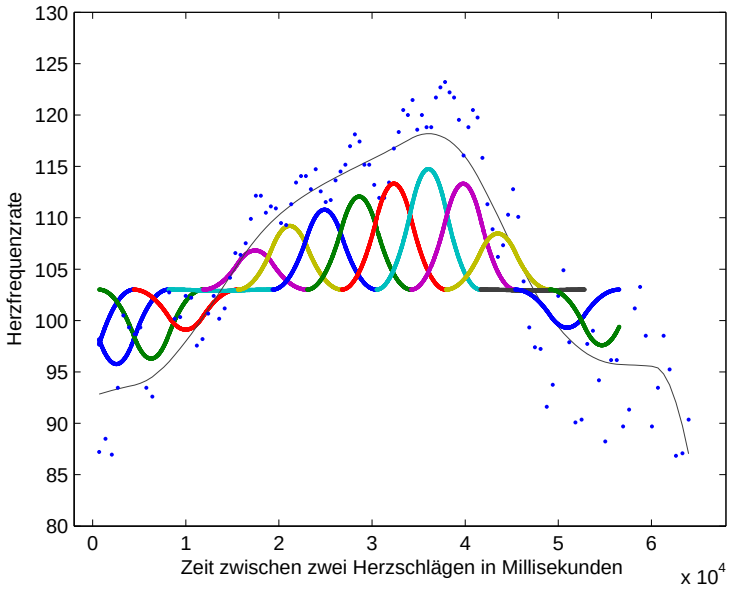

Abbildung 4.5: B-Spline-Schätzung (Grad 2) im nichtlinearen AR(1)-Modell für Herzfrequenzdaten.

Abbildung 4.5 verdeutlicht das Prinzip der Basisfunktions-Schätzung. Aus der Summe der Basisfunktionen an einem beliebigen Punkt x ergibt sich die Schätzkurve. Im Fall von B-Splines 2. Grades werden die y-Werte von sich überschneidenden Splines aufsummiert. Als Resultat erhält man die Schätzkurve (vgl. Abbildung 4.5). Auch Gewichte kleiner gleich Null können auftreten. Es sei noch erwähnt, dass die Basislinie der B-Splines auf Höhe des Intercepts (ca. 103) liegt. Dies kann nur durch geeignete Restriktionen erreicht werden, da der Intercept klar von den Gewichten δ in der Schätzung abgetrennt werden muss. Ansonsten ist die Identifizierbarkeit des Modells nicht gegeben. Eine Trennung kann beispielsweise durch die Restriktion

$$\delta_m = -\sum_{j=1}^{m-1} \delta_j$$

erreicht werden. Dies führt unweigerlich zu Veränderungen in den Design- und Splinematrizen.

Die Schätzung der Kurve erfolgt wie gewöhnlich mittels der KQ- oder Ordinary-Least-Squares-Methode, so dass gilt

$$S = \sum_{i=1}^{n} \left(y_i - (\alpha_0 + \sum_{j=1}^{m} \delta_j B_j(x_i)) \right)^2 \rightarrow \min_{\alpha_0, \delta_j}. \tag{4.11}$$

Dies ist lediglich 'ein' Vorschlag zum Fitten von B-Spline-Modellen. Dieses Modell ist Basis von Abbildung 4.5. Es ist offensichtlich, dass eine Vielzahl anderer Modelle auch über Least-Squares-Funktionen modelliert werden können.

4.4.2 Algorithmus zur Berechnung von B-Spline Basisfunktionen

Bisher wurde im Detail darauf eingegangen, was B-Splines sind und wie sie definiert werden, allerdings wurde noch nicht beschrieben, wie man sie berechnet. De Boor (1978) gibt zur Berechnung einen rekursiven Algorithmus (Cox-de Boor Recursion), der zunächst B-Splines 0-ten Grades berechnet und auf Basis dieser dann B-Splines 1. Grades berechnet, und so weiter. Eilers und Marx (1996) fassen dies im Appendix zusammen, geben aber sowohl für S-Plus und auch für Matlab einen Programmier-Code an.

Die rekursiven Formeln für B-Splines $B_{k,q}(x)$ folgen. Dabei ist $B_{k,q}(x)$ diejenige Basisfunktion, die am Punkt ξ_k beginnt und Grad q hat. Man beachte, dass mit wachsendem q sich die Länge der Basisfunktion verändert. Falls alle Intervalle I gleiche Länge haben, gilt für B-Splines 0-ten Grades

$$B_{k,0}(x) = \begin{cases} 1 & \text{falls } \xi_k \leq x \leq \xi_{k+1} \\ 0 & \text{sonst} \end{cases}$$

und die rekursiven Formeln für die B-Splines höheren Grades lauten

$$B_{k,q}(x) = \frac{x - \xi_k}{\xi_{k+q} - \xi_k} B_{k,q-1}(x) + \frac{\xi_{k+q+1} - x}{\xi_{k+q+1} - \xi_{k+1}} B_{k+1,q-1}(x). \tag{4.12}$$

Damit kann jeder Wert der Prädiktorvariablen x in Basisfunktionen ausgedrückt werden. In Matlab kann dies anhand folgenden Codes geschehen.

- Algorithmus zur Berechnung von B-Splines (Matlab)

```
1  function B = bspline(x, ξ0, ξK, int, q)[1]
2  l = (ξK − ξ0)/int;
3  vec = a + l ∗ [−q : int − 1];
4  mat = (0 ∗ x + 1) ∗ vec;
5  mat1 = x ∗ (0 + vec ∗ 1);
6  mat2 = (mat1 − mat)/l;
7  B = (mat <= mat1)&(mat1 < (mat + l));
8  r = [2 : length(t), 1];
9  for k = 1 : q
10   B = (P. ∗ B + (k + 1 − P). ∗ B(:, r))/k;
11   end;
12   end;
```

Zeile 1 ist der Matlab-Befehl für einen Funktionsaufruf, Zeile 2 berechnet die Länge der äquidistanten Intervalle I und Zeile 3 fasst die Knotenpunkte in einem Vektor *vec* zusammen. In Zeile 4 wird der Vektor *vec* in Matrixform umgeschrieben, um soviele Sequenzen an Knotenpunkten zu generieren wie es Datenpunkte gibt. Zeile 5 vervielfältigt den Datenvektor, so dass in Zeile 7 die B-Splines 0-ten Grades berechnet werden können. Diese Information ist Grundlage für die folgenden rekursiven Berechnungen. In Zeile 6 und 8 werden vorbereitende Maßnahmen für die nun folgende Schleife (Zeile 9-11) getroffen. Dabei wird in Zeile 6 der erste Bruch aus (4.12) berechnet, der nur noch leicht verändert werden muss, um den zweiten Bruch zu bilden. Dies erfolgt automatisch in Zeile 10 bei $(k + 1 - P)$. Zeile 8 bildet einen Vektor r, der in Zeile 10 bei $B(:, r)$ auf eine bestimmte Sequenz von Punkten innerhalb der Matrix B zugreift. In der ersten der k Iterationen werden nun die B-Splines 1. Grades, in der zweiten Iteration die B-Splines 2. Grades, usw., berechnet und in

[1] 'int' ist die Anzahl an Intervallen I

Form der Matrix B, die alle $B_{k,q}(x)$ enthält, ausgegeben (Zeile 1). Matrix B hat für B-Splines 2. Grades folgende Band-Gestalt (Eubank, 1999), falls die Werte von x aufsteigend sortiert sind.

$$B = \begin{pmatrix} 0.4592 & 0.5399 & 0.0009 & 0 & 0 \\ 0.1467 & 0.7483 & 0.1050 & 0 & 0 \\ 0 & 0.2509 & 0.7066 & 0.0425 & 0 \\ 0 & 0 & 0.0672 & 0.7322 & 0.2006 \end{pmatrix} \tag{4.13}$$

Es zeigt sich, dass $q + 1 = 2 + 1 = 3$ B-Splines zeilenweise verschieden von Null sind. In Matrix B ist $n = 4$ und die Anzahl an Splines beträgt 5. Jedem x_i wird eine Zeile zugeordnet. Die Spalten geben die Anzahl an Splines an (hier: 5). Die Matrix B enthält also die Werte der Basisfunktionen an den Stellen $x_i, i = 1, \ldots, n$. Wie man sieht ist x_1 in Zeile 1 nur in den ersten drei B-Splines enthalten, daher sind nur für diese drei B-Splines positive Werte zu erkennen, die restlichen zwei B-Splines nehmen Werte von Null an. Im Falle von x_4 sind die letzten drei B-Splines positiv, d.h. $x_4 > x_1$, da die B-Splines der Reihenfolge nach in Matrix B eingeordnet sind. Die Zeilensummen betragen immer 1, unabhängig vom Grad des B-Splines.

4.5 P-Splines

Zunächst sollte der Begriff Penalisierung definiert werden, bevor genauer auf penalisierte B-Splines eingegangen werden kann.

4.5.1 Penalisierung allgemein

Man betrachte zunächst ein nichtlineares 'Ordinary-Least-Squares (OLS)'-Modell, ohne irgendwelche Restriktionen an die Kurve γ zu stellen. In einem derartigen Modell ist es ohne weiteres möglich die Kurve γ so durch die Daten zu legen, dass jeder Datenpunkt genau von der Schätzung getroffen wird. Dies wird allgemein als Interpolation bezeichnet. Allerdings ist die Interpolation von Daten nicht das Ziel des Analysten, vielmehr soll nach der Schätzung ein Trend in den Daten erkennbar sein, der die lokale Variation als 'Noise' auffasst. Damit sind die beiden Gegenspieler

bei nicht-linearen statistischen Modellen schon definiert. Das OLS-Modell versucht zu interpolieren, während dessen Gegenspieler der 'Penalty' oder 'Roughness Penalty', wie Green und Silverman (1990) ihn bezeichnen, versucht, die lokale Variation aus dem Modell zu nehmen und einen Trend zu Vorschein zu bringen, um die Abhängigkeit zwischen Einflussgröße und Response gut interpretieren zu können.

Nun gilt es die 'Roughness' einer Schätzung zu bewerten. Vorstellbar sind natürlich mehrere Möglichkeiten, intuitiv würde man jedoch daran denken, die Roughness einer Kurve danach zu beurteilen, wie sie hin- und herschwankt. Eine Möglichkeit dazu bietet die zweite Ableitung der Kurve. Green und Silverman (1990) begründen dies wie folgt. Jedes Maß an Roughness sollte nicht von der Addition einer Konstante oder eines linearen Trends beeinflusst werden, d.h. falls zwei Funktionen sich nur bzgl. einer Konstante und einem linearen Term unterscheiden, sollte die Roughness der Kurve gleich sein. Dies wäre nur gegeben, falls nach zweifacher Ableitung die lineare und konstante Komponente aus dem Modell entfernt worden sind. Es lässt sich folgern, dass die zweite Ableitung als Maß für die Roughness sehr gut geeignet ist. Auch höhere Ableitungen können zur Bewertung der Roughness Anwendung finden

Generell formuliert man Roughness-Penalties nach folgendem Prinzip:

$$P = \int_a^b (\gamma''(x))^2 dx \tag{4.14}$$

Dazu mehr in den folgenden Kapiteln. Eilers und Marx (1996, 1998) führen in ihren Artikeln eine andere Art der Penalisierung ein, die allgemein als 'Difference Penalties' (Differenzen-Penalties) bezeichnet werden. Sie schreiben:

- The difference penalty is a good approximation to the integrated square of the kth derivative.

Im Unterschied zum Penalty aus (4.14) gibt es beim Differenzen-Penalty keine Ableitungen, sondern nur Approximationen der Ableitung, die in Graden gemessen

werden. So entspricht ein Differenzen-Penalty 2. Grades einer Approximation für einen Roughness-Penalty aus (4.14), d.h. einem Penalty gebildet mittels der zweiten Ableitung. Dies gilt synonym auch für Penalties höherer Ordnung, bzw. höheren Grades. Der Grad des Differenzen-Penalties wird mit k angegeben. Tutz (2003) fasst die Differenzenpenalties in Kürze zusammen und führt einen Differenzenoperator Δ_j für die B-Spline-Gewichte δ_j ein. Für diesen gilt

$$\Delta_j \delta_j = \delta_j - \delta_{j-1}.$$

Auch die zweiten Differenzen können penalisiert werden, d.h.

$$\Delta_j^2 \delta_j = \Delta_j(\delta_j - \delta_{j-1}) = \delta_j - 2\delta_{j-1} + \delta_{j-2}.$$

Höhere Differenzen können nach dem selben Schema rekursiv abgeleitet werden. Allgemein kann ein Differenzen-Penalty in Kombination mit einem OLS-Ansatz wie folgt zusammengefasst werden (Eilers und Marx, 1996, Tutz 2003)

$$\begin{aligned} S &= \sum_{i=1}^{n} (y_i - \gamma(x_i))^2 + \lambda \sum_{j=k+1}^{m} (\Delta_j^k \delta_j)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \delta_j B_j(x_i) \right)^2 + \lambda \sum_{j=k+1}^{m} (\Delta_j^k \delta_j)^2. \end{aligned} \tag{4.15}$$

In (4.15) wurde ein Glättungsparameter λ eingeführt. Dieser regelt den Einfluss des Differenzen-Penalties auf das Modell. Je größer Parameter λ gewählt wird, desto glatter wird die geschätzte Kurve, bei $\lambda = 0$ kann es vorkommen, dass die 'Residual Sum of Square' auf Null fällt. Dies ist jedoch recht unwahrscheinlich, falls die Anzahl an Knotenpunkten im Vergleich zur Anzahl an Datenpunkten niedrig gewählt wird. Bei hoher Anzahl an Knoten ist auch bei B-Spline-Schätzungen eine Interpolation möglich.

4.5.2 Differenzenpenalties in Matrixschreibweise

Durch simple Umformungen kann sowohl der Term der OLS-Schätzung als auch der Term der Differenzen-Penalties in Matrixschreibweise formuliert werden. Es gilt

$$S = (y - B\delta)'(y - B\delta) + \lambda \delta' P_k \delta, \tag{4.16}$$

wobei B der abgeleiteten B-Spline-Matrix aus (4.13) entspricht und

$$\begin{aligned} \delta' &= (\delta_1, \ldots, \delta_m) \\ y' &= (y_1, \ldots, y_n) \\ P_k &= D_k' D_k. \end{aligned} \tag{4.17}$$

Die Penalty-Matrix P_k setzt sich aus zwei Differenzenmatrizen $(D_k)_{(m-1)\times m}$ zusammen. Im Falle von Differenzen 1. Grades sind diese folgendermaßen aufgebaut.

$$D_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & \ldots & 0 \\ 0 & -1 & 1 & 0 & \ldots & 0 \\ 0 & 0 & -1 & 1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & -1 & 1 \end{pmatrix}$$

Multipliziert man nur die rechte bzw. linke Hälfte der Matrixprodukte $\delta' P_1 \delta = \delta' D_1' D_1 \delta$, erhält man die ersten Differenzen.

$$\delta' D_1' = (\delta_2 - \delta_1, \delta_3 - \delta_2, \ldots, \delta_m - \delta_{m-1})$$

Nun muss noch gewährleistet werden, dass keine negativen Differenzen entstehen. Daher gehen die Differenzen quadratisch in das Modell ein.

4.5.3 OLS-Schätzung und Hat-Matrix

Um eine geglättete OLS-Schätzung zu erhalten wird das Gleichungssystem aus (4.15) minimiert. Man erhält damit den KQ-Schätzer

$$\delta = (B'B + \lambda P_k)^{-1} B' y. \tag{4.18}$$

Ohne Glättung reduziert sich der Schätzer aus (4.18) zum allgemein bekannten Gauss-Markov-Schätzer. Die Schätzung kann nach dem Prinzip aus Kapitel 3 für

Longitudinaldaten mit Working-Kovarianz-Matrizen erweitert werden, so dass für den Longitudinal-Schätzer δ_{LT} gilt

$$\delta_{LT} = (B'V^{-1}B + \lambda P_k)^{-1}B'V^{-1}y. \tag{4.19}$$

Alle Parameter δ_{LT}, B, V^{-1}, P_k und y aus (4.19) werden entsprechend den Vorgaben für Longitudinaldaten modelliert.

Die Hatmatrizen unterscheiden sich nur bzgl. der Working-Kovarianzmatrix W.

$$\begin{aligned} H &= B(B'B + \lambda P_k)^{-1}B', \text{ bzw.} \\ H_{LT} &= B(B'V^{-1}B + \lambda P_k)^{-1}B'V^{-1} \end{aligned}$$

Auf Basis der Hatmatrix kann die Spur, d.h. die Anzahl an Freiheitsgraden im Modell, bestimmt werden (Hastie und Tibshirani, 1990). Falls $\lambda = 0$ entspricht die Spur der Hatmatrix $tr(H)$ der Anzahl an Freiheitsgraden. Werden Penalties in das Modell mitaufgenommen, reduziert sich die Anzahl an Freiheitsgraden. Dann kann die Spur der Hatmatrix als Approximation für die Anzahl an Freiheitsgraden aufgefasst werden (Hastie und Tibshirani, 1990).

4.6 Truncated Power Series Basis

Eine andere Methode, Polynom-Stücke in Intervallen an die Daten anzupassen, ist mittels der 'Truncated Power Basis Series'

$$\gamma(x) = \delta_0 + \delta_1 x + \delta_2 x^2 + \delta_3 x^3 + \sum_{k=1}^{K} \delta_{3+k}(x - \xi_k)_+^3, \tag{4.20}$$

wobei gilt, dass

$$(x - \xi_k)_+ = \max\{0, x - \xi_k\}.$$

Es ist offensichtlich, dass (4.20) ein kubisches Polynom auf dem Intervall $[\xi_k, \xi_{k+1}]$ mit kontinuierlicher erster und zweiter Ableitung darstellt. Es kommt zustande über

die Taylorentwicklung (Eubank, 1999)

$$y_i = \sum_{j=1}^{m=4} \delta_j x_i^{j-1} + \text{Rem}(x_i) + \varepsilon_i$$

mit

$$\text{Rem}(t) = \frac{1}{(m-1)!} \int_0^1 \gamma^{(m)}(x)(x-\xi)_+^{m-1} d\xi. \tag{4.21}$$

Der Summenterm aus (4.20) ist eine Approximation von (4.21). Lösungsalgorithmen und detaillierte Beschreibungen der Eigenschaften der truncated power series basis findet man bei Eubank (1999) aus statistischer Sichtweise und auch bei Stoer (1999) aus mathematisch-numerischer Sichtweise.

4.7 Smoothing Splines

Eine detaillierte Beschreibung von Smoothing Splines findet man bei Green und Silverman (1994) und bei Pollock (1999). Fahrmeir und Tutz (2001) fassen das Thema auf wenigen Seiten zusammen.

Auch die Smoothing Splines werden in Intervallen I zwischen Knotenpunkten $\xi_0 < \xi_1 < \ldots < \xi_{K-1} < \xi_K$ gefittet. Man wählt dabei so viele Knotenpunkte wie Datenpunkte, d.h. $K = n$ und $x_i = \xi_i$ mit $x_0 < x_1 < \ldots < x_{n-1} < x_n$, $i = 0, \ldots, n$, und versucht die Lücke zwischen zwei benachbarten Datenpunkten (x_i, y_i) und (x_{i+1}, y_{i+1}) mittels kubischer Funktionen zu überbrücken und zwar so, dass eine kontinuierliche Kurve entsteht, d.h. dass an den Datenpunkten keine Unstetigkeitsstellen auftreten. Dies wird durch die Voraussetzung erfüllt, dass die aneinandergereihten Polynom-Stücke kontinuierliche erste und zweite Ableitungen aufweisen müssen. Zunächst geht man dabei von einer Interpolation aus, die dann später über Roughness-Penalties in die Schätzung eines Trends übergeht.

4.7.1 Interpolation

Auch bei Smoothing Splines geht man wieder von einem Likelihood-Ansatz aus, so dass gilt

$$l = \sum_{i=0}^{n} w_i \left(y_i - S(x_i)\right)^2 = (y - S)W^{-1}(y - S), \tag{4.22}$$

wobei w_i ein Gewicht ist und $S(x_i)$ eine glatte Funktion darstellt, die die Daten interpoliert. Sie ist für jedes der Intervalle I definiert, so dass für ein Polynom-Stück $S_i(x)$ zwischen ξ_i und ξ_{i+1} gilt

$$S_i(x) = a_i(x - \xi_i)^3 + b_i(x - \xi_i)^2 + c_i(x - \xi_i) + d_i, \tag{4.23}$$

wobei $x \in [\xi_i, \xi_{i+1})$ und a_i, b_i, c_i und d_i gegebene Konstanten sind. In diesem Fall hat das Polynomstück kubische Form. Damit lassen sich die ersten und zweiten Ableitungen nach x

$$\begin{aligned} S_i'(x) &= 3a_i(x - \xi_i)^2 + 2b_i(x - \xi_i) + c_i \\ S_i'(x) &= 6a_i(x - \xi_i) + 2b_i \end{aligned}$$

einfach bilden und sind kontinuierlich. Die obengenannte Voraussetzung eine Kurve ohne Unstetigkeitsstellen zu schätzen, bringt mit sich, dass zwei benachbarte Polynom-Stücke S_{i-1} und S_i sich am Knotenpunkt ξ_i treffen sollten, d.h. es gilt

$$S_{i-1}(\xi_i) = S_i(\xi_i) = y_i \tag{4.24}$$

mit

$$\begin{aligned} S_i(\xi_i) &= a_i(\xi_i - \xi_i)^3 + b_i(\xi_i - \xi_i)^2 + c_i(\xi_i - \xi_i) + d_i \\ &= d_i. \end{aligned} \tag{4.25}$$

Aus (4.24) und (4.25) folgt

$$a_{i-1}h_{i-1}^3 + b_{i-1}h_{i-1}^2 + c_{i-1}h_{i-1} + d_{i-1} = d_i = y_i$$

mit $h_{i-1} = \xi_i - \xi_{i-1}$. Gemäß den obigen Definitionen muss dies auch im Fall der ersten Ableitung

$$\begin{aligned} &S'_{i-1}(\xi_i) = S'_i(\xi_i), \text{ oder äquivalent dazu} \\ &3a_{i-1}h_{i-1}^2 + 2b_{i-1}h_{i-1} + c_{i-1} = c_i \end{aligned} \tag{4.26}$$

und zweiten Ableitung

$$\begin{aligned} &S''_{i-1}(\xi_i) = S''_i(\xi_i), \text{ oder äquivalent dazu} \\ &6a_{i-1}h_{i-1} + 2b_{i-1} = 2b_i \end{aligned} \tag{4.27}$$

gelten. Es sollten nun noch die Bedingungen für die Endpunkte ξ_0 und ξ_n gewählt werden. Hierfür werden bei Pollock (1999) zwei Möglichkeiten angegeben. Einerseits kann man nach der als 'Clamping' bezeichneten Methode vorgehen und den ersten Ableitungen der Splines an den Endpunkten den entsprechenden Wert der Ableitung von y zuweisen. Andererseits können die Endpunkte nicht restringiert werden ('ends are left free', Pollock (1999)), so dass gilt

$$S''_0(\xi_0) = 2b_0 = 0 \text{ und } S''_{n-1}(\xi_n) = 2b_n = 0. \tag{4.28}$$

Falls Voraussetzung (4.28) zutrifft, bezeichnet man die zugehörigen Splines als 'Natural Splines'. Nun sind b_0 und b_n also festgelegt, d.h. es sind nur noch die restlichen Parameter 2. Grades $b_1, \ldots, b_{n-1}$ aus den Daten zu bestimmen. Falls diese bestimmt sind, können die übrigen Parameter berechnet werden. Es gelten nun folgende vier Voraussetzungen für das i-te Intervall

1. $S_i(\xi_i) = y_i$ (vgl. (4.24))
2. $S_i(\xi_{i+1}) = y_{i+1}$ (vgl. (4.24))
3. $S''_i(\xi_i) = 2b_i$ mit $b_0 = 0$ und $b_K = 0$ (vgl. (4.27))
4. $S''_i(\xi_{i+1}) = 2b_{i+1}$ mit $b_K = 0$ (vgl. (4.27))

Aus Voraussetzung (1) folgt, dass $d_i = y_i$ (vgl. (4.25)), d.h. der 'Intercept' d_i der kubischen Splines entspricht dem zugehörigen Wert für y_i. Aus Voraussetzung (2) und (4.25) folgt

$$a_i h_i^3 + b_i h_i^2 + c_i h_i + d_i = y_{i+1} \Leftrightarrow c_i = \frac{y_{i+1} - y_i}{h_i} - a_i h_i^2 - b_i h_i. \tag{4.29}$$

Voraussetzung (3) kann wie Voraussetzung (1) als Identität behandelt werden und Voraussetzung (4) kann mit Hilfe von (4.27) wie folgt aufgelöst werden

$$6a_i h_i + 2b_i = 2b_{i+1} \Leftrightarrow a_i = \frac{b_{i+1} - b_i}{3h_i}. \tag{4.30}$$

Setzt man nun (4.30) in (4.29) ein, erhält man

$$c_i = \frac{y_{i+1} - y_i}{h_i} - \frac{1}{3}(b_{i+1} + 2b_i)h_i \tag{4.31}$$

Wie man sieht, sind nun die Parameter a_i in (4.30) und c_i in (4.31) nur noch abhängig von den Parametern 2. Grades b_i bzw. b_{i+1} und von den Datenpunkten (x_i, y_i). Mit Hilfe von (4.30) und (4.31) kann nun (4.26) umgeformt werden zu

$$b_{i-1}h_{i-1} + 2b_i(h_{i-1} + h_i) + b_{i+1}h_i = \frac{3}{h_i}(y_{i+1} - y_i) - \frac{3}{h_{i-1}}(y_i - y_{i-1}). \tag{4.32}$$

Gleichung (4.32) kann nun in ein tridiagionales Gleichungssystem

$$\begin{pmatrix} p_1 & h_1 & 0 & \dots & 0 & 0 \\ h_1 & p_2 & h_2 & \dots & 0 & 0 \\ 0 & h_2 & p_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & p_{n-2} & h_{n-2} \\ 0 & 0 & 0 & \dots & 0 & p_{n-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_{n-2} \\ q_{n-1} \end{pmatrix}$$

überführt werden, indem man den Index k von 1 nach $K-1$ laufen lässt. Zusätzlich gelte

$$p_i = 2(h_{i-1} + h_i) = 2(\xi_{i+1} - \xi_{i-1})$$

und

$$q_i = \frac{3}{h_i}(y_{i+1} - y_i) - \frac{3}{h_{i-1}}(y_i - y_{i-1}).$$

Zunächst wird das Gleichungssystem in ein bidiagonales System umgewandelt, anschließend wird es gelöst durch 'Backsubstitution'. Danach werden die Parameter a_i nach (4.30) berechnet, im Anschluss die Parameter c_i über eine Rekursionsformel

$$c_i = (b_i + b_{i-1})h_{i-1} + c_{i-1}$$

Der Start-Parameter c_0 wird mittels (4.31) bestimmt. Damit sind die Datenpunkte x_i interpoliert.

4.7.2 Penalisierung bei Smoothing Splines

Die Penalisierung erfolgt nach dem gleichen Prinzip aus Kapitel 4.5.1 (Green und Silverman, 1990). Dabei werden wieder die Penalties aus (4.14) verwendet. Somit gelte für ein penalisiertes Likelihood-basiertes Gleichungssystem für Smoothing Splines

$$l = \lambda \sum_{i=0}^{n} w_i \left(y_i - S(x_i)\right)^2 + (1 - \lambda) \int_{\xi_0}^{\xi_n} S''(x)^2 dx \tag{4.33}$$

wobei $\lambda \in [0, 1]$ einen Glättungsparameter darstellt und w_i der Dispersion-Parameter bzw. ein Gewicht ist (vgl. Kapitel 2.12). Glättungsparameter λ gibt hierbei an, wie stark beide Summanden aus (4.33) gewichtet werden. Die Wahl $\lambda = 1$ führt zur Interpolation der Daten, während $\lambda = 0$ im Falle der zweiten Ableitung im Penaltyterm zu einer Schätzung in Form einer Gerade führt. Dies ist offensichtlich, da die zweite Ableitung kubischer Funktionen lineare Gestalt hat (vgl. (4.27)).

Aufgrund der Aufteilung des Wertebereichs von x in Polynomstücke, kann der Penaltyterm ähnlich wie bei den Differenzenpenalties (vgl. Kapitel 4.5.1) in Summenform geschrieben werden, so dass gilt

$$\int_{\xi_0}^{\xi_n} S''(x)^2 dx = \sum_{i=0}^{n-1} \int_{\xi_i}^{\xi_{i+1}} S_i''(x)^2 dx. \tag{4.34}$$

Danach kann die zweite Ableitung von $S(x)$ gebildet und das Integral bestimmt werden. Damit folgt

$$\int_{\xi_i}^{\xi_{i+1}} S_i''(x)^2 dx = \frac{4h_i}{3}(b_i^2 + b_i b_{i+1} + b_{i+1}^2). \tag{4.35}$$

Zusammenfassend präsentiert sich die Likelihoodfunktion nun wie folgt

$$l = \sum_{i=0}^{n} w_i \left(y_i - d_i\right)^2 + 2\lambda_P \sum_{i=0}^{n} h_i(b_i^2 + b_i b_{i+1} + b_{i+1}^2), \tag{4.36}$$

wobei $\lambda_P = 2(1-\lambda)/3\lambda$ und $d_i = S_i(x_i)$. Im Falle der Interpolation galt nach Voraussetzung (1) aus Kapitel 4.7.1, dass $y_i = d_i$. Da nun Penalties in das Modell mitaufgenommen werden, gilt nun diese Beziehung nicht mehr, da die geschätzte Kurve nicht mehr zwangsläufig direkt durch die Datenpunkte (x_i, y_i) verläuft und Residuen $e_i = y_i - d_i$ auftreten. Daher werden nun die ersten beiden Voraussetzungen aus Kapitel 4.7.1 umgeformt zu

1. $S_i(x_i) = d_i$
2. $S_i(x_{i+1}) = d_{i+1}$

Die zwei restlichen Voraussetzungen behalten ihre Gültigkeit. Die Berechnungen und Umformungen verlaufen exakt nach dem Prinzip aus Kapitel 4.7.1, so dass man eine vergleichbare Formel für (4.32) erhält

$$b_{i-1}h_{i-1} + 2b_i(h_{i-1} + h_i) + b_{i+1}h_i = \frac{3}{h_i}(d_{i+1} - d_i) - \frac{3}{h_{i-1}}(d_i - d_{i-1}). \tag{4.37}$$

Unter der Voraussetzung, dass $b_0 = b_n = 0$ gilt, kann das Gleichungssystem aus Kapitel 4.7.1 auch in penalisierten Fall für (4.37) gebildet werden.

$$\begin{pmatrix} p_1 & h_1 & 0 & \dots & 0 & 0 \\ h_1 & p_2 & h_2 & \dots & 0 & 0 \\ 0 & h_2 & p_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & p_{n-2} & h_{n-2} \\ 0 & 0 & 0 & \dots & 0 & p_{n-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{pmatrix}$$

$$= \begin{pmatrix} r_0 & f_1 & r_1 & 0 & \dots & 0 & 0 \\ 0 & r_1 & f_2 & r_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & r_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \dots & f_{n-1} & r_{n-1} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}$$

Für die Matrixelemente gilt dann

$$\begin{aligned} p_i &= 2(h_{i-1} + h_i) = 2(\xi_{i+1} - \xi_{i-1}), \\ r_i &= \frac{3}{h_i}, \\ f_i &= -\left(\frac{3}{h_{i-1}} + \frac{3}{h_i}\right) = -(r_{i-1} + r_i) \end{aligned}$$

und in Matrixschreibweise gilt

$$Mb = Q'd.$$

Damit ist der bei Green und Silverman (1994) geschilderte Zusammenhang (Seite 12/13, Theorem 2.1) hergeleitet. Sowohl bei Pollock als auch bei Green und Silverman folgt nun die Ableitung der Gesamt-Schätzung mit Penalisierung. Es gilt mit $b = M^{-1}Q'd$

$$\begin{aligned} l &= (y-d)'V^{-1}(y-d) + \lambda_P b'Mb \\ &= (y-d)'V^{-1}(y-d) + \lambda_P d'QM^{-1}Q'd, \end{aligned} \tag{4.38}$$

wobei V ähnlich wie bei Liang und Zeger (1986) als Working-Kovarianz-Matrix oder als Gewichtsmatrix interpretiert werden kann. Nun folgt die übliche Prozedur zur Bildung der Schätzparameter. Als Lösung ergibt sich

$$d = (V^{-1} + \lambda_P Q' M^{-1} Q)^{-1} (V^{-1})' y. \tag{4.39}$$

Im Falle $V^{-1} = I$ reduziert sich (4.39) mit $K_P = Q' M^{-1} Q$ zu

$$d = (I + \lambda_P Q' M^{-1} Q)^{-1} y = (I + \lambda_P K_P)^{-1} y. \tag{4.40}$$

Zur numerisch effizienteren Lösung des Minimierungsproblems aus (4.38) wurde ein Algorithmus von Reinsch (1967) entwickelt. Mittels der Cholesky-Zerlegung von $V^{-1} + \lambda_P K_P = L' D L$ kann die (zur damaligen Zeit) aufwendige Invertierung verhindert werden. Die Wahl des Smoothing-Parameters λ_p erfolgt nach Green und Silverman (1994) entweder frei oder nach einer automatisierten Methode, d.h. über die Kreuzvalidierungskriterien CV oder GCV. Mit derartigen Ansätzen beschäftigt sich der nächste Abschnitt.

4.8 Modelloptimierung

Details zur Wahl des bzw. der Glättungsparameter für nichtparametrische Regressionsansätze findet man bei Green und Silverman (1994), als Zusammenfassung bei Fahrmeir und Tutz (2001). Generell sind die Ansätze auf alle Arten an nichtparametrischer Regression anwendbar, denn man geht prinzipiell von einer Fehlerquadratsummen-Minimierung aus. Hier seien nur die wichtigsten Ansätze beschrieben. Als einfachsten Ansatz kann man wohl den ASR-Ansatz (Average-Squared Residual) bezeichnen, der nur die Residuen aufsummiert und mittelt.

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

Der ASR-Ansatz gibt kein gutes Maß für die Optimierung des Modells ab, weil die Schätzung und Modellevaluierung zusammenfallen. Geeigneter erscheint das

Kreuzvalidierungs-Kriterium

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\gamma}_\lambda^{-i}(x_i)\right)^2, \tag{4.41}$$

wobei jeweils der i-te Datenpunkt bei der Schätzung ausgelassen wird. Um nicht n-mal ein Submodell mit einem fehlenden Datenpunkt bzw. t fehlenden Datenpunkten, im Longitudinalfall, schätzen zu müssen, kann (4.41) umgeformt werden zu

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\gamma}_\lambda(x_i)}{1 - H_{ii}(\lambda)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} DEV_i \left(\frac{1}{1 - H_{ii}(\lambda)}\right)^2, \tag{4.42}$$

wobei $DEV = \sum_{i=1}^{n}(y_i - \hat{\gamma}_\lambda(x_i))^2$ als Abkürzung für die Devianz steht. Dabei sind H_{ii} die Diagonalelemente der zur Schätzung gehörigen Hatmatrix H. Im Falle der Smoothing Splines ohne Gewichtsmatrix W ist die Hatmatrix wie folgt definiert (vgl. (4.40)).

$$H = (I + \lambda_P K_P)^{-1}$$

Auch für andere nichtparametrische Ansätze können Hatmatrizen abgeleitet werden und in das CV-Kriterium miteinbezogen werden. Craven und Wahba (1979) definieren ein weiterentwickeltes ähnliches Kriterium namens 'Generalized Cross-Validation' (GCV), das darauf basiert, dass im Nenner die individuellen Einträge der Hatmatrix H durch einen Mittelwert bzw. die Spur ersetzt werden. Es gilt

$$\begin{aligned} GCV(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\gamma}_\lambda(x_i)}{1 - \frac{1}{n}\sum_{i=1}^{n} H_{ii}(\lambda)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\gamma}_\lambda(x_i)}{1 - \frac{tr(H_{ii}(\lambda))}{n}}\right)^2 \\ &= \mathrm{DEV}\left(\frac{1}{1 - \frac{tr(H_{ii}(\lambda))}{n}}\right)^2 \end{aligned} \tag{4.43}$$

Das GCV kann in gewisser Weise als Spezialfall des AIC angesehen werden. Eine kurze Beschreibung der Ableitung des AIC findet man bei Fahrmeir und Tutz (2001). Allgemein gilt für Likelihood-basierte Modelle

$$AIC = -2l + 2tr(H).$$

Das Bayesian Information Criterion (BIC) ist wie folgt definiert (Schwarz, 1978)

$$BIC = -2l + log(n)tr(H).$$

Es ist offensichtlich, dass alle Kriterien beim Übergang zur gewichteten nichtparametrischen Least-Squares-Schätzung und auch für Longitudinaldaten dementsprechend verändert werden müssen. Green und Silverman (1994) führen dies für die gewichtete Schätzung in Kapitel 3.5 detailliert aus. Da man im Falle von Longitudinaldaten auch von einer (Pseudo)-Likelihood als Modell-Basis ausgeht, werden alle Kriterien übertragen.

Kapitel 5

Semiparametrische Modelle

Zunächst wurde in Kapitel 2 auf das lineare Modell eingegangen. Es folgte in Kapitel 4 der Übergang zur nichtlinearen nichtparametrischen Schätzung. Nun werden beide Modelle miteinander verbunden, indem p parametrische Terme und ein zusätzlicher nichtparametrischer Term gleichzeitig in das Modell mitaufgenommen werden. Diese Modelle werden im allgemeinen als semiparametrisch bezeichnet. Sehr viel Literatur ist in den letzten Jahren über derartige Modelle veröffentlicht worden. Green und Silverman (1994) besprechen semiparametrische Schätzmethoden für Smoothing Splines, während vor allem Lin, Wang und Carroll in mehreren Artikeln auf semiparametrische Kernelschätzung eingehen, in denen die nichtparametrischen Modelle aus Sektion 4.3 übertragen werden. Desweiteren seien noch Speckman (1988), Severini und Staniswalis (1994), Hastie und Tibshirani (1989), Zeger und Diggle (1994) und Ruppert et al. (2004) als Autoren der Standardliteratur zu nennen.

Allgemein kann ein semiparametrisches Modell wie folgt definiert werden

$$y_i = z_i'\beta + \gamma(t_i) + \varepsilon_i, \tag{5.1}$$

wobei der erste Term $z_i'\beta$ den parametrischen Teil des Modells mit p Prädiktoren und $\gamma(t_i)$ den nichtparametrischen Teil mit einem skalaren Prädiktor t_i definiert[1]

[1]Im folgenden kann die nichtparametrische Variable nun nicht mehr mit x_i bezeichnet werden, da die x_i schon in der Designmatrix z_i enthalten sind.

(Ruppert et al. (2004)). Die Fehler sind wie so oft mit ε_i bezeichnet.

In der Praxis erweisen sich semiparametrische Modelle laut Green und Silverman (1994) als 'surprisingly useful'. Die Autoren nennen hierfür Beispiele. Zunächst führen sie an, dass in Datensätzen immer wieder ein paar Variablen nichtlinear mit den Responses zusammenhängen und daher speziell in das Modell eingebaut werden können. Außerdem gibt es - nach Meinung vieler obengenannter Autoren - Situationen, in denen ein Modell bis auf eine mögliche Inhomogenität bezüglich der Zeit als linear angenommen wird. Modell (5.1) würde dann beispielsweise einen Zeit-variierenden Intercept beinhalten, der sich nichtlinear verändert, während die restlichen Prädiktoren linear in das Modell eingehen. Zudem können Sprungstellen in nichtlinearen Zusammenhängen leicht mittels parametrisch modellierter Indikatorfunktionen überwunden werden.

Zunächst soll nun auf Greens und Silvermans Smoothing Spline Methode eingegangen werden.

5.1 Smoothing Splines in semiparametrischen Modellen

Man geht dabei von einem vergleichbaren Ansatz wie in Sektion 4.7 aus und integriert den zusätzlichen parametrischen Term $z_i'\beta$ in das Modell

$$l = \sum_{i=1}^{n} \left(y_i - z_i'\beta - S(t_i)\right)^2 .$$

Die glatte Funktion γ wird wie in Sektion 4.7 mit S bezeichnet, die nichtlinearen Prädiktoren werden von nun an als t_i definiert, da $z_i = (1, x_{11}, \ldots, x_{1p})'$ (vgl. Kapitel 2). Um die Interpolation zu vermeiden wird analog zu Sektion 4.7 ein Roughness-

Penalty angehängt und Gewichte eingefügt

$$l = \sum_{i=1}^{n} w_i \left(y_i - z_i'\beta - S(t_i)\right)^2 + \lambda_P \int S(t)''dt. \tag{5.2}$$

Die Gewichte w_i entsprechen im Modell mit normalverteilten Fehlern mit $E(\varepsilon_i) = 0$ dem inversen Dispersion-Parameter bzw. der inversen Varianz (vgl. Kapitel 2).

Beim Modellieren mit Smoothing Splines ist es notwendig mit geordneten, d.h. $t_1 < t_2 < \ldots < t_n$, und eindeutigen Prädiktorwerten, d.h. $t_i \neq t_j$, $i \neq j$ zu rechnen. Um die Eindeutigkeit zu gewährleisten, wird nun eine 'Incidence'-Matrix N eingeführt, die die t_i in eine geordnete und eindeutige Prädiktorenvariable s überführt, so dass gilt

$$\begin{aligned} t &= Ns, \quad s = (s_1, \ldots, s_m), \quad l = (1, \ldots, m), \quad m \leq n \\ N_{il} &= 1, \text{ wenn } t_i = s_l, \text{ sonst } 0 \end{aligned}$$

Damit lässt sich Gleichung (5.2) in Matrixschreibweise überführen

$$l = (y - z'\beta - Ns)'W^{-1}(y - z'\beta - Ns) + \lambda_P \int S''(t)dt. \tag{5.3}$$

- Beispiel: falls $t = (2, 1, 2, 3)$ folgt $s = (1, 2, 3)$ mit

$$N = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ und } t = Ns$$

Von nun an ist die Funktion $S(t_i)$ abhängig von $S(s_l)$. Zusammengefasst werden alle $S(s_l)$ im Vektor s (vgl. (5.3)). Damit lässt sich Gleichung (5.3) mit Hilfe von (4.38) überführen in

$$l = (y - z'\beta - Ns)'W^{-1}(y - z'\beta - Ns) + \lambda_{Ps} K_{Ps}, \tag{5.4}$$

mit $K_P = QM^{-1}Q$. Offensichtlich ist, dass nun sowohl nach β als auch nach s abgeleitet werden muss, um die gesamte Scorefunktion mit $p + m$ Gleichungen zu erhalten. Löst man (5.4) auf und leitet dann ab erhält man

$$\begin{pmatrix} z'V^{-1}z & z'V^{-1}N \\ N'V^{-1}N & N'V^{-1}N + \lambda_P K_P \end{pmatrix} \begin{pmatrix} \beta \\ s \end{pmatrix} = \begin{pmatrix} z' \\ N' \end{pmatrix} V^{-1}y. \tag{5.5}$$

Aus (5.5) ergibt sich

$$z'V^{-1}z\beta = z'V^{-1}(y - Ns), \tag{5.6}$$

$$(N'V^{-1}N + \lambda_P K_P)s = N'V^{-1}(y - z'\beta). \tag{5.7}$$

Beide Gleichungen sind intuitiv verständlich. Gleichung (5.6) schätzt, falls s bekannt ist, eine gewöhnliche 'weighted least squares' Regression mit einem 'Offset' von Ns. Gleichung (5.7) schätzt, bei bekanntem β, einen Smoothing Spline mit dem Offset $z'\beta$. Generell lösen sich beide Gleichungssysteme durch Backfitting, indem solange alternierend geschätzt, bis ein Abbruchkriterium unterschritten wird (Breiman und Friedman (1985), Hastie und Tibshirani (1989)). In der Theorie konvergiert der Backfitting-Algorithmus immer (Green und Silverman (1994)).

5.1.1 Direkter Lösungsansatz

Gleichung (5.7) kann umgeformt werden zu

$$Ns = A(y - z'\beta), \tag{5.8}$$

wobei $A = N(N'V^{-1}N + \lambda_P K_P)N'V^{-1}$ die zugehörige Hatmatrix[2] ist. Falls man nun (5.8) in (5.6) einsetzt, ergibt sich nach trivialen Umformungen

$$z'V^{-1}(I - A)z'\beta = z'V^{-1}(I - A)y. \tag{5.9}$$

Simples Auflösen nach β führt zum gewünschten Schätzer und Einsetzen in (5.8) löst die nichtparametrische Schätzgleichung bezüglich s. Nachteilhaft an der direkten Methode ist, dass keine orthogonalen Zerlegungen durchgeführt werden können. Zudem können größere Rundungsfehler entstehen (Green und Silverman (1994)).

[2]Die Hatmatrix wird auch als Smoother-Matrix, Smoother oder Projector-Matrix bezeichnet.

5.1.2 Kreuzvalidierung

Basierend auf derselben Idee wie in Sektion 4.8, werden die Kriterien dem semiparametrischen Modell entsprechend abgeändert. Von nun an gilt

$$\hat{y} = Hy = z'\hat{\beta} + N\hat{s}. \tag{5.10}$$

Die beiden Kriterien CV und GCV entsprechen den Kriterien aus (4.41) und (4.43). Beim GCV verändert sich lediglich die Spur zu

$$tr(H) = tr(A) + tr\left((z'V^{-1}(I-A)z)^{-1}z'V^{-1}(I-A)^2z\right).$$

5.1.3 Speckmans Ansatz

Speckman (1988) geht auch von Modell (5.2) aus, schlägt aber einen anderen Lösungsweg vor. Zunächst eliminiert er die Variablen $t = (t_1, \ldots, t_n)'$ aus dem Modell, d.h. er adjustiert zunächst sowohl den Prädiktor $z = (z_1, \ldots, z_n)'$ als auch den Response y bezüglich des nichtlinearen Einflusses von t. Dabei nimmt Speckman an, dass gilt

$$\tilde{z} = (I-A)z = z - Az = z - \hat{z},$$
$$\tilde{y} = (I-A)y = y - Ay = y - \hat{y},$$

wobei A die zugehörige Hatmatrix ist. Die Hatmatrix A leitet sich aus nichtparametrischen Schätzung von t auf z bzw. y ab und ist für beide Vektoren gleich. Offensichtlich ist, dass $\tilde{z}$ und $\tilde{y}$ wegen der Differenzen $z - Az$ bzw. $y - Ay$ nun Residuen darstellen. Da damit das Modell vom nichtparametrischen Term bereinigt wurde, wird nun zunächst der parametrische Term über die Residuen $\tilde{z}$ und $\tilde{y}$ geschätzt mit

$$\tilde{\beta} = (\tilde{z}'\tilde{z})^{-1}\tilde{z}'\tilde{y}$$

bzw. für gewichtete Schätzungen mit

$$\tilde{\beta} = (\tilde{z}'V^{-1}\tilde{z})^{-1}\tilde{z}'V^{-1}\tilde{y}.$$

Dann wird $\tilde{\beta}$ in Gleichung (5.2) als Schätzwert eingesetzt. Nun kann der nichtparametrische Teil des Modells $\gamma(t)$ mit dem parametrischen Term als Offset nach Gleichung (5.7) geschätzt werden. Zur Schätzung ist genau wie beim direkten Lösungsansatz kein iteratives Vorgehen notwendig.

Matrix A wird in diesem Fall als Smoother Matrix für Smoothing Splines bezeichnet. Ersetzt man die Matrix durch einen Kernel-Smoothing Operator, bekannt aus (4.1), erhält man das nichtparametrische Kernel-Analogon zum Smoothing Spline. Vergleiche zwischen Kernel und Smoothing-Spline-Schätzern findet man u.a. bei Speckman (1988) und Wang et al. (2004).

5.2 Semiparametrische GLMs und Smoothing Splines

Semiparametrische GLMs werden bei Green und Yandell (1985), Green und Silverman (1994) und Hastie und Tibshirani (1990) besprochen. Green und Silverman folgen dabei dem Ansatz von Green und Yandell. Generell gelten im folgenden die Voraussetzungen aus Kapitel 2. Das ursprüngliche GLM wird folglich um einen nichtlinearen Term $\gamma(t_i)$ erweitert, so dass gilt

$$g(\mu_i) = z_i'\beta + \gamma(t_i). \tag{5.11}$$

Im semiparametrischen GLM wird wieder von einem Likelihoodansatz ausgegangen, der, um Identifizierbarkeit zu gewährleisten, penalisiert wird. Falls keine Penalisierung in das Modell mitaufgenommen würde, wären die Parameter β nicht mehr identifizierbar, da die Daten durch die nichtparametrische Funktion $\gamma(t_i)$ interpoliert würden. Folglich gilt für die penalisierte (Pseudo)-Log-Likelihood

$$\begin{aligned} \Pi &= l(\theta, \phi) - \frac{1}{2}\lambda \int \gamma''(t)^2 dt \\ &= l(\theta, \phi) - \frac{1}{2}\lambda \gamma' K_P \gamma. \end{aligned} \tag{5.12}$$

Ableiten von (5.12) nach β bzw. γ führt zu den Scoregleichungen, die vergleichbar sind mit (5.5). Einen iterativen Lösungsalgorithmus findet man bei Green und Silverman (1994). Im Grunde genommen ist er dem Algorithmus zur Lösung von GLMs ähnlich, allerdings muss ein (iterativer) Zwischenschritt eingebaut werden. Folgende Schritte sind dabei auszuführen.

1. Bestimme die Incidence-Matrix N nach obengenannten Vorgaben
2. Wähle einen Startwert und initialisiere die Iterationen, z.B. mit $\beta = 0$ und $s = (N'N)^{-1}N'(g(y_1), \ldots, g(y_n))'$
3. Berechne ein Update des Modells mit $\eta = z'\beta + Ns$
4. Berechne den Working-Responsevektor $\hat{z} = (y_i - \mu_i)g'(\mu_i) + z'\beta + Ns$ und die Gewichtsmatrix $V = \text{diag}\left([g'(\mu_i)^2 b''(\theta_i)]^{-1}\right)$
5. Berechne die neuen Schätzer β^{neu} und s^{neu} mittels der Methoden aus Sektion 5.1, beispielsweise über die Scoregleichungen (5.6) und (5.7) und verwende dabei den Working-Responsevektor $\hat{z}$ (iterativer Zwischenschritt)
6. Wird ein Abbruchkriterium unterschritten, fahre mit Schritt 7 fort. Falls nicht, kehre zu Schritt 3 zurück mit $\beta = \beta^{neu}$ und $s = s^{neu}$
7. Vervollständige das Glätten der Smoothing Splines

Kreuzvalidierung für semiparametrische GLMs ist ausführlich bei Green und Silverman (1994) in Kapitel 5 beschrieben. Übereinstimmungen zu den bisher beschriebenen Ansätzen bestehen durchaus.

5.3 Semiparametrische GLMs und Kernelestimation

Als Grundlage der folgenden Ausführungen dient der oft-zitierte Artikel von Severini und Staniswalis aus dem Jahre 1994 über Quasi-Likelihood-Estimation in Kombina-

tion mit Kernel-Schätzungen und Glätten. Severini und Staniswalis reduzieren dabei die ursprünglichen Likelihood-Methoden, die auf Verteilungsannahmen basieren, auf Quasi-Likelihoodansätze, die nur die Bestimmung der ersten beiden Momente verlangen (vgl. Kapitel 2). Dabei stellen die Autoren einen Ansatz vor, den sie als 'Generalized Profile Likelihood' bezeichnen. Der Ansatz ist sehr ähnlich zu verstehen wie der Ansatz aus Sektion 5.1 und wurde erstmals von Severini und Wong (1992) vorgeschlagen. Es geht darum während der Schätzung der einen Komponente im semiparametrischen Modell, die andere Komponente auf einem konstanten Wert zu halten ('held fixed'). Die parametrische Komponente wird also während der iterativen Schätzung der nichtparametrischen Komponente festgehalten. Dabei ist nicht von Bedeutung welche Schätz-Methode zur Glättung verwendet wird. Anschließend wird der nichtparametrische Term konstant gehalten, um eine Profile Likelihood für die parametrische Komponente zu konstruieren, natürlich unter der Verwendung eines Likelihood- oder Quasi-Likelihood-Ansatzes. Die Profile Likelihood Funktion wird beispielsweise mittels Maximum-Likelihood-Ansätzen geschätzt. Vorteilhaft an diesem Ansatz ist, dass das semiparametrische Schätzproblem in zwei Teile zerlegt wird. In Sektion 5.2 wurde ein etwas anderer Algorithmus vorgeschlagen, der in Schritt 5 zwischen beiden Komponenten hin- und herschwankt. Dies ist nun nicht mehr der Fall, da nun zuerst eine Komponente und im Anschluss die andere Komponente (iterativ) geschätzt wird. Diese Zweiteilung der Schätzung bewirkt, dass die parametrische Komponente mit einer Rate von $n^{\frac{1}{2}}$ gegen den wahren Parameter konvergiert, falls der Smoothing Parameter optimal bestimmt ist.

Severini und Staniswalis bezeichnen die Quasi-Likelihoodfunktion mit Q, so dass allgemein für die separaten Scorefunktionen gilt

$$\frac{\partial}{\partial s}\sum_{i=1}^{n} Q(h(T_i s + z_i'\beta); y_i) = 0, \tag{5.13}$$

$$\frac{\partial}{\partial \beta}\sum_{i=1}^{n} Q(h(T_i s + z_i'\beta); y_i) = 0, \tag{5.14}$$

wobei $T_i s$ den nichtparametrischen Teil des Modell darstellt mit T_i als $1 \times m$-Vektor und $s = (s_1, \ldots, s_m)$' (vergleichbare Zerlegung wie bei den Smoothing-Splines). Zunächst wird nun β festgehalten und s geschätzt, während dann $\hat{s}$ in Scorefunktion (5.14) eingesetzt wird und der parametrische Teil geschätzt wird. Severini und Staniswalis merken zudem an, dass nur diese zwei Schritte notwendig sind, um das Modell vollständig zu lösen, allerdings geben sie nicht an, wie β initialisiert wird im ersten Schritt bei der Schätzung von (5.13).

5.3.1 Konstruktion eines Kernel-Schätzers

Da zuerst die nichtparametrische Schätzung (5.13) durchgeführt wird und im Anschluss die parametrische Schätzung (5.14), muss nur die Scorefunktion (5.13) mit Kernelgewichten versehen werden, so dass für die veränderte Funktionen nun gilt

$$\sum_{i=1}^{n} K_\lambda \left(\frac{\tilde{t} - t_i}{\lambda} \right) \frac{\partial}{\partial s} Q(h(T_i s + z_i' \beta); y_i) = 0, \tag{5.15}$$

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{n} Q(h(T_i s + z_i' \beta); y_i) = 0. \tag{5.16}$$

K_λ entspricht einem Kernelgewicht bei festem $\tilde{t}$. Welche Art von Kernelschätzung nun Anwendung findet, bleibt dem User überlassen. Beispielsweise kann Average-Kernel-Estimation oder Local-Linear-Kernel-Estimation (vgl. Kapitel 4.3) zum Einsatz kommen. Die parametrische Teil eröffnet auch viele Möglichkeiten zur Schätzung. Severini und Staniswalis schlagen beispielsweise zunächst eine Average-Kernel-Estimation über den Nadaraya-Watson-Schätzer (4.4) mit Offset $z_i' \beta$ vor. Im Anschluss Schätzen sie den parametrische Teil mit dem von Speckman (1988) vorgeschlagenen Ansatz, wobei die Smoother-Matrix A mittels Kernelfunktionen gebildet wird (vgl. Sektion 5.1.3).

5.4 Longitudinaldaten und semiparametrische Modelle

Ausgehend von einem Artikel von Zeger und Diggle (1994) soll die semiparametrische Theorie für Longitudinaldaten im folgenden detailliert besprochen werden. Anschließend werden die nichtparametrischen Kernel-Ansätze von Lin und Carroll (2000), Lin et al. (2003) und Wang (2003) in das semiparametrische Modell übertragen werden.

Generell ändert sich wenig im Bezug auf die Modelldefinition, da jeweils nur die Cluster- bzw. Personen-spezifischen Indizes t angefügt werden, so dass gilt

$$y_{it} = z'_{it}\beta + \gamma(t) + \varepsilon_{it}, \tag{5.17}$$

wobei z_{it} ein $p \times 1$ Vektor und $\gamma(t)$ ein glatte Funktion ist, die entweder nur von i oder von i und t abhängig ist. Die Fehler ε_{it} werden als normalverteilt angenommen. Generell wird genauso vorgegangen wie zuvor beschrieben und die Schätzung in zwei Teile zerlegt. Zunächst wird eine Schätzmethode für den nichtparametrischen Teil konstruiert, im Anschluss wird der parametrische Term bearbeitet. Zeger und Diggle (1994) gehen dabei zunächst von nichtparametrischen Modell mit $\beta = 0$ aus und konstruieren eine Kernel-Methode für den nichtparametrischen Term $\gamma(t)$, der auf Observation-Level-Basis, d.h. abhängig von i und t, behandelt wird. Bei der Schätzung wird auf den Average-Kernel-Estimator von Nadaraya-Watson zurückgegriffen. Dabei wird die Within-Cluster-Korrelation vollkommen ignoriert. Aus diesem Grund gelingt es den Schätzer effizient und konsistent zu konstruieren. Es gilt für den Schätzer von γ

$$\hat{\gamma}(t) = \sum_{i=1}^{n}\sum_{t=1}^{T} c_{it}(t)y_{it} \tag{5.18}$$

mit

$$c_{it}(t) = \frac{c^*_{it}(t, \lambda(t))}{\sum_{i=1}^{n}\sum_{t=1}^{T} c^*_{it}(t, \lambda(t))}.$$

und

$$c_{it}^*(t) = \frac{1}{\lambda(t)} K\left(\frac{t - t_{it}}{\lambda(t)}\right)$$

Nach der Initialisierung iteriert der Algorithmus zwischen folgenden zwei Schritten:

1. Bei gegebenem (Generalized Least Squares)-Schätzer $\hat{\beta}^{[k]}$ während der k-ten Iteration, werden die Residuen $r_{it}^{[k]} = y_{it} - z_{it}'\hat{\beta}^{[k]}$ berechnet und dann die zuvor beschriebene Kernelmethode zur Schätzung des nichtparametrischen Terms $\gamma^{[k]}(t)$ angewandt (Offsetmethode, vgl. Sektion 5.2 und 5.3).

2. Bei gegebenem $\gamma^{[k]}(t)$, werden nun die Residuen $u_{it}^{[k]} = y_{it} - \gamma^{[k]}(t_{ij})$ gebildet und $\hat{\beta}^{[k+1]}$ mittels Generalized Least Squares-Schätzung bestimmt:

$$\hat{\beta}^{[k+1]} = \left(\sum_{i=1}^{n} z_i' V_i^{-1} z_i\right)^{-1} z_i' V_i^{-1} u_i^{[k]}, \tag{5.19}$$

   wobei $u_i^{[k]} = (u_{i1}^{[k]'}, \ldots, u_{iT}^{[k]'})'$ und V_i ein Working-Kovarianzmatrix für y_i ist.

Anders als Severini und Staniswalis (1994) werden nun wieder beide Schritte parallel und nicht getrennt voneinander ausgeführt. Es wird solange iteriert bis ein Abbruchkriterium unterschritten wird. Der zugrundeliegende Algorithmus wurde ursprünglich von Hastie und Tibshirani (1986) vorgeschlagen und ist unter dem Namen 'Backfitting' bekannt. Es folgen im Anschluss weitere semiparametrische Modelle, die wie bei Severini und Staniswalis (1994) auf GLMs ausgeweitet werden.

5.4.1 Semiparametrisches GLM für Longitudinaldaten auf Cluster-Level-Basis

Nahezu parallel sind im Jahre 2001 zwei Artikel von Lin und Carroll (2001a, 2001b) erschienen. Im Grunde genommen werden darin sehr ähnliche semiparametrische Modelle für Kernel-Schätzer vorgestellt:

$$g(\mu_{it}) = z_{it}'\beta + \gamma(t_i) + \varepsilon_{it}, \tag{5.20}$$

$$g(\mu_{it}) = z_{it}'\beta + \gamma(t_{it}) + \varepsilon_{it}, \tag{5.21}$$

Modell (5.20) ist auf Cluster-Level-Basis, d.h. die nichtparametrische Funktion $\gamma(t)$ ist nur von i und nicht von i und t abhängig, im Gegensatz zum Observation-Level-Modell (5.21). Zunächst soll Modell (5.20) konstruiert werden. Lin und Carroll (2001a) zeigen in ihrem Artikel, das im Cluster-Level-Modell die Ergebnisse von Severini und Staniswalis (1994, S.508-509) eins zu eins übernommen werden können, d.h dass die Schätzung am effizientesten ist, wenn die wahre Kovarianz-Matrix und Working-Kovarianz-Matrix zusammenfallen und dass der Schätzer mit der gewöhnlichen Rate von $n^{\frac{1}{2}}$ gegen den wahren Schätzer konvergiert unabhängig von falscher oder richtiger Spezifikation der Working-Kovarianz-Matrix. Treffen alle eben genannten Voraussetzungen in Modellen zu, so werden sie von Lin und Carroll (2001a) als semiparametrisch effizient bezeichnet. Wie auch schon in Sektion 4.3.1 angesprochen, trifft dies im Falle des Observation-Level-Modells nicht zu; dazu später mehr. Es ergibt sich für den nichtparametrischen Teil von Modell (5.20) folgende Scoregleichung, die im Grunde genommen bis auf $\mu_i(x,\beta)$ und $K_{i(\lambda)}(x)_{1\times 1}$ Gleichung (4.7) entspricht

$$\sum_{i=1}^{n} K_{i(\lambda)}(x) G_{i(q)}(x)' D_i(x) W_i(x)^{-1} \left(y_i - \mu_i(x,\beta)\right) = 0. \tag{5.22}$$

Der Unterschied besteht darin, dass der Offset in das Modell integriert werden muss. In (4.7) galt

$$\begin{aligned} \mu_i(x) &= \alpha_0, \text{ im Average-Kernel Fall und} \\ \mu_i(x) &= \alpha_0 + \alpha_1 \frac{(x - x_i)}{\lambda}, \text{ im lokal-linearen Fall.} \end{aligned}$$

Im semiparametrischen Fall muss der parametrische Offset $z_i'\beta$ angefügt werden, so dass gilt

$$\begin{aligned} \mu_i(x) &= z_i'\beta + \alpha_0, \text{ im Average-Kernel Fall und} \\ \mu_i(x) &= z_i'\beta + \alpha_0 + \alpha_1 \frac{(x - x_i)}{\lambda}, \text{ im lokal-linearen Fall.} \end{aligned}$$

Die restlichen Vektoren und Matrizen entsprechen den Matrizen aus Sektion 4.3.1. Bei der Schätzung des parametrischen Teils darf der passende Offset auch nicht fehlen. In diesem Artikel wird auf einen Generalized Least Square Ansatz zurückgegriffen,

so dass gilt

$$\sum_{i=1}^{n} \frac{\partial \mu \left(z_i'\beta + \hat{\gamma}(z_i, \beta)U\right)'}{\partial \beta} V_i^{-1}(x) \left(y_i - \mu \left(z_i'\beta + \hat{\gamma}(z_i, \beta)U\right)\right) = 0. \tag{5.23}$$

Nun ist $\hat{\gamma}(z_i, \beta)U$ als Offset in das Modell miteingefügt worden. Der Vektor $U = (1, \ldots, 1)'_{T\times 1}$ hat die Aufgabe die Prädiktoren z_i von Cluster-Level auf Observation-Level-Basis zu bringen, da ja nur ein $(z_i)_{1\times 1}$ für die T-fache Menge an Responses $(y_i)_{T\times 1}$ vorhanden ist. Die restlichen Vektoren und Matrizen verändern sich nicht und sind wie zuvor definiert worden. Lin und Carroll (2001a) bezeichnen die gemeinsame Lösung $(\hat{\beta}, \hat{\gamma})$ der Gleichungssysteme als 'Profile Kernel Estimator' (vgl. Sektion 5.3). Die semiparametrische Schätzung der Sandwich-Kovarianz-Matrix ist auf Seite 1181 des zugehörigen Artikels beschrieben. Es sei noch erwähnt, dass der Profile Kernel Estimator auch für semiparametrische Modelle auf Quasi-Likelihood-Basis angewandt werden kann (Severini und Staniswalis, 1994, Lin und Carroll, 2001a).

5.4.2 Semiparametrisches GLM für Longitudinaldaten auf Observation-Level-Basis

Auf die Grundlagen des Modell wurde in Gleichung (5.21) schon eingegangen. Die Konstruktion des nichtparametrischen Schätzers verläuft analog zu (4.9) mit

$$\sum_{i=1}^{n} G_{i(q)}(x)' D_i(x) \left(K_{i(\lambda)}^{\frac{1}{2}}(x) V_i(x)^{-1} K_{i(\lambda)}^{\frac{1}{2}}(x)\right) (y_i - \mu_i(x, \beta)) = 0 \tag{5.24}$$

bzw.

$$\sum_{i=1}^{n} G_{i(q)}(x)' D_i(x) V_i(x)^{-1} K_{i(\lambda)}(x) (y_i - \mu_i(x, \beta)) = 0. \tag{5.25}$$

im assymmetrischen Fall. Sämtliche Vektoren und Matrizen werden aus Sektion 4.3.1 übernommen, allerdings wird wieder der zusätzliche Offset $z_i'\beta$ bei μ_i eingefügt. Das parametrische Gleichungssystem setzt sich entsprechend (5.23) zusammen

$$\sum_{i=1}^{n} \frac{\partial \mu \left(z_i'\beta + \hat{\gamma}(z_i, \beta)\right)'}{\partial \beta} V_i^{-1}(x) \left(y_i - \mu \left(z_i'\beta + \hat{\gamma}(z_i, \beta)\right)\right) = 0. \tag{5.26}$$

Die Vektoren U aus (5.23) können nun vernachlässigt werden, da sich die Prädiktoren z_i bereits auf Cluster-Level-Basis befinden. Allerdings bleibt der Offset $\hat{\gamma}(z_i, \beta)$ auch hier bestehen.

Im zugehörigen Artikel werden noch einige zusätzliche Adjustierungen vorgenommen, die zu sehr in das Detail gehen. Es handelt sich dabei beispielsweise darum, dass die Working-Kovarianz-Matrix auf zwei verschieden Arten geschätzt werden kann, je nachdem ob gerade nichtparametrisch oder parametrisch geschätzt wird.

Auch im Observation-Level-Fall wird die Profile Kernel Methode (vgl. Sektion 5.3) zur Schätzung angewandt. Jedoch zeigt sich, dass nur bei Working-Independence mit $W = I$ $n^{\frac{1}{2}}$-konsistente Schätzungen entstehen. Jede andere Wahl der Working-Kovarianz, d.h. selbst bei korrekt spezifizierter Kovarianz-Matrix, führt zu inkonsistenten Schätzungen, es sei denn γ is 'undersmoothed', d.h. zu stark geglättet. Semiparametrisch ineffiziente Schätzer wurden konstruiert (Lin und Carroll, 2001b). Der Schätzer von Zeger und Diggle (1994) kann als ein Spezialfall dieses Schätzers angesehen werden, da normalverteilte Daten mit Working-Independence verwendet werden.

5.4.3 Effiziente semiparametrische Schätzung für Longitudinaldaten

Wang et al. (2004) übertragen den nichtparametrischen effizienten Schätzansatz aus Sektion 4.3.2 nun auch auf semiparametrische Modelle auf Observation-Level-Basis und zeigen, dass dieser Ansatz die Kovarianzstruktur effektiv miteinbezieht. Modell (5.21) dient wieder als Grundlage für die Berechnungen. Dabei gehen die Autoren bei der nichtparametrischen Schätzung exakt so vor, wie in Sektion 4.3.2 beschrieben.

Es gelte für die Scorefunktion

$$\sum_{i=1}^{n}\sum_{t=1}^{T} K_\lambda(x_{it}-x)\mu^{(1)}\left(\alpha_0+\alpha_1(x_{it}-x)\right) G^*_{iT(q)}(x)' \times W_i^{-1}\left(y_i-\mu^*_{it}(x_{it},\alpha_K,\hat{\gamma}(x_i),\beta)\right). \tag{5.27}$$

Nur der Offset $z_{it}\beta$ wird zusätzlich in $\mu^*_{it}(x_{it},\alpha_K,\hat{\gamma}(x_i),\beta)$ aus Gleichung (5.27) miteinbezogen, so dass gilt

$$\begin{aligned}&\mu^*_{it}\left(x_{it},\alpha_K,\hat{\gamma}(z_i),\beta\right) = \\ &= ((z'_{i1}\beta+\hat{\gamma}(z_{i1}),\ldots,z'_{i,t-1}\beta+\hat{\gamma}(z_{i,t-1}),z'_{it}\beta+\sum_{f=0}^{q}(\alpha_K)_f(x_{it}-x)^f, \\ &z'_{i,t+1}\beta+\hat{\gamma}(z_{i,t+1}),\ldots,z'_{iT}\beta+\hat{\gamma}(z_{iT}))'.\end{aligned}$$

Das weitere Vorgehen entspricht exakt dem aus Sektion 4.3.2 mit dem einzigen Unterschied, dass ein zusätzlicher Offset auf die Schätzung $z'_{it}\beta+\sum_{f=0}^{q}\alpha_f(x_{it}-x)^f$ für x_{it} und auch die zugehörigen Residuen $z'_{it}\beta+\hat{\gamma}(z_{it})$ aufaddiert wird. Nach der nichtparametrischen Schätzung folgt die parametrische Schätzung. Diese verläuft nach demselben Prinzip wie in Sektion 5.4.2 beschrieben. Wang et al. (2004) formulieren dies wie folgt:

- '... our algorithm differs from those previously proposed in step I by replacing the original kernel GEE estimator by one that utilizes the correlations, while step II of the algorithm is the same.'

Es wird also wieder nach der Profile Kernel Methode von Severini und Staniswalis (1994) vorgegangen, die zunächst in Schritt I das nichtparametrische Gleichungssystem löst und im Anschluss in Schritt II das parametrische Gleichungssystem, wobei die parametrische Schätzung sich nicht von der Schätzung aus Sektion 5.4.2 unterscheidet.

Kapitel 6

Seemingly Unrelated Regression

Generell sind zwei Veröffentlichungen von Arnold Zellner aus dem Jahre 1962 und 1963 richtungsweisend für die Entwicklung von 'Seemingly Unrelated Regression'-Modellen, kurz SUR-Modellen. In seinem ersten Paper 'An efficient method of estimating seemingly unrelated regressions ...' definiert Zellner die SUR-Modelle und im darauffolgenden Paper 'Estimations for seemingly unrelated regression equations ...' fügt er einige Zusatzbemerkungen zum ersten Paper an und geht auf die Momentenschätzung ein. Allerdings wird in beiden Veröffentlichungen nicht auf Longitudinal-Daten eingegangen. Für Longitudinaldaten im Kontext von semiparametrischen SUR-Modellen wurde sehr wenig Literatur veröffentlicht.

SUR-Modelle sind vergleichsweise einfach zu beschreiben und auch zu verstehen. Zellner geht von folgendem Modell aus. Es sei

$$y_i = Z_i\beta_i + \varepsilon_i$$

die i-te Gleichungssystem von n Gleichungssystemen mit jeweils T Gleichungen. Dabei sei y_i ein $(T \times 1)$-Response-Vektor und Z_i eine nichtstochastische $(T \times p)$-Matrix mit vollem Rang. Der Vektor β_i habe die Dimensionen $(p \times 1)$. Die Fehler $(\varepsilon_i)_{T\times 1}$ sind normalverteilt sind mit Erwartungswert $E(\varepsilon_i) = 0$. In Matrixschreibweise lassen

sich die n Gleichungen zusammenfassen zu

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & Z_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

$$y = Z\beta + \varepsilon,$$

mit $y = (y_1', y_2', \dots, y_n')'_{nT\times 1}$, $\beta = (\beta_1', \beta_2', \dots, \beta_n')'_{np\times 1}$, $\varepsilon = (\varepsilon_1', \varepsilon_2', \dots, \varepsilon_n')'_{nT\times 1}$ und $Z_{nT\times np}$ in blockdiagonaler Form mit $i = 1, \dots, n$ und $t = 1, \dots, T$. Bei Unkorreliertheit der y_i werden n parallele unabhängige Regressionen geschätzt, falls Korrelationen zwischen den einzelnen y_i bestehen, werden die einzelnen Regressionen zu SUR-Modellen, da sie über die Fehler ε korrelieren. Der Einfluss der Regressionen aufeinander wird über eine 'Varianz-Kovarianz-Matrix' Σ bestimmt. Diese ist ähnlich zu interpretieren wie die Working-Kovarianz-Matrizen von Liang und Zeger (1986) ohne Longitudinal-Struktur. Der Unterschied zur Working-Kovarianz-Matrix von Liang und Zeger besteht darin, dass die Varianz-Kovarianz-Matrix Σ, so wie sie Zellner bezeichnet, aus mehreren Blöcken besteht. Jeder dieser Blöcke hat Diagonalgestalt und gibt entweder die Varianzen σ_{ii} im Falle von (y_i, y_i) an oder die Kovarianzen σ_{ij}, $i \neq j$, im Falle von (y_i, y_j), $i \neq j$, $j = 1, \dots, n$. Es gelte

$$\begin{aligned} \Sigma &= cov(\varepsilon) = \begin{pmatrix} \sigma_{11}I & \sigma_{12}I & \dots & \sigma_{1n}I \\ \sigma_{21}I & \sigma_{22}I & \dots & \sigma_{2n}I \\ \vdots & \vdots & & \vdots \\ \sigma_{n1}I & \sigma_{n2}I & \dots & \sigma_{nn}I \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix} \otimes I = \\ &= \Sigma_{Cov} \otimes I \end{aligned}$$

mit I als $(T \times T)$-Einheitsmatrix. Die einzelnen σ_{ij} sind Skalare, die über die Einheitsmatrix in die Varianz-Kovarianz-Matrix miteinbezogen werden. Auch heterogene Varianzschätzungen sind zulässig mit einer zusätzlichen Unterscheidung nach t, so dass sich das Skalar σ_{ii} bzw. σ_{ij} in einen $(T \times 1)$-Vektor mit den Elementen $\sigma_{ii} = (\sigma_{ii1}, \dots, \sigma_{iiT})$ bzw. $\sigma_{ij} = (\sigma_{ij1}, \dots, \sigma_{ijT})$ umwandeln lässt.

Die Schätzung des Modells erfolgt über Aitkens 'Two-Stage'-Methode, die zunächst die Gleichungssysteme über den Gauss-Markov-Schätzer löst, dann die Varianz-Kovarianz-Matrix Σ schätzt und das Gleichungssystem nochmals löst mit dem Aitkenschätzer

$$\beta = (Z'\Sigma^{-1}Z)^{-1}Z'\Sigma^{-1}y.$$

Da damals noch keine (schnellen) Computer existierten, wird in Zellners Artikel viel Wert auf effiziente Möglichkeiten zur Berechnung vor allem bei der Invertierung der Matrizen gelegt. Außerdem geht Zellner noch zusätzlich auf den Effizienzgewinn des SUR-Modells im Vergleich zu n parallelen unabhängigen Regressionsmodellen ein. Es ergibt sich, dass sich die Varianz im SUR-Modell im Vergleich zum unabhängigen Modell mit Varianz $V_{ind} = \sigma^2(Z'_{ind}Z_{ind})^{-1}$ approximativ um einen Faktor $(1-\rho)$ verringert, da nach Zellner (1962) gilt

$$V_{SUR} = (1-\rho)V_{ind} = (1-\rho)\sigma^2(Z'_{ind}Z_{ind})^{-1}.$$

Die Definition von ρ ist im zugehörigen Artikel auf den Seiten 353 und 354 vorzufinden. Detaillierte Analysen zu diesem Thema findet man auch bei Zellner (1963). Der Beweis der Erwartungstreue der SUR-Modelle folgt bei Kakwani (1967).

6.1 Seemingly Unrelated Semiparametric Regression

Im folgenden soll kurz über ein noch unveröffentlichtes Manuskript von Carroll (2003) berichtet werden. Der Autor geht dabei von zwei Responses y_{i1} und y_{i2} aus. Zusätzlich zum parametrischen Term von Zellner (1962) wird nun eine nichtparametrische Komponente angefügt. Es gilt somit

$$\begin{aligned} y_{i1} &= z'_{i1}\beta_1 + \gamma(t_{i1}) + \varepsilon_{i1}, \\ y_{i2} &= z'_{i2}\beta_2 + \gamma(t_{i2}) + \varepsilon_{i2}. \end{aligned} \tag{6.1}$$

mit

$$\begin{aligned}
\tilde{\varepsilon}_i &= (\varepsilon_{i1}, \varepsilon_{i2})', \\
\tilde{y}_i &= (y_{i1}, y_{i2})', \\
\tilde{t}_i &= (t_{i1}, t_{i2})', \\
\tilde{z}_i &= (z_{i1}, z_{i2})', \\
cov(\tilde{\varepsilon}_i) &= \Sigma_\varepsilon.
\end{aligned}$$

Falls die nichtparametrischen Terme in (6.1) vernachlässigt werden und es gilt, dass $z_{i1} \neq z_{i2}$, reduziert sich das Modell zum ursprünglichen SUR-Modell von Zellner, und dieses Modell weist im Vergleich zu separat gefitteten Modellen, wie berichtet, Effizienzgewinne auf. Dies soll bei Carroll (2003) auch für das semiparametrische Modell gezeigt werden. Allerdings liegen in dem Manuskript keine theoretischen Beweise vor, die Simulations-Studien aber zeigen, dass sich in bestimmten Situationen starke Effizienzgewinne für das semiparametrische SUR ergeben. Im Manuskript wird zur Schätzung der nichtlinearen Komponente von B-Spline Basisfunktionen ausgegangen, die über eine Pseudo-Likelihood mit Penalisierung modelliert werden, d.h. für die Splines gilt

$$\gamma_j(t) = \sum_{l=1}^{L} \delta_{lj} B_{lj}(t).$$

und die Schätzvektoren werden zusammengefasst zu

$$\begin{aligned}
\Delta_1 &= (\beta_1', \delta_{11}, \delta_{21}, \ldots, \delta_{L1})', \\
\Delta_2 &= (\beta_2', \delta_{12}, \delta_{22}, \ldots, \delta_{L2})', \\
\Delta &= (\Delta_1', \Delta_2')', \\
B &= \begin{pmatrix} z_1' & B_{11}(t_1) & B_{21}(t_1) & \ldots & B_{L1}(t_1) & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 & z_2' & B_{12}(t_2) & B_{22}(t_2) & \ldots & B_{L2}(t_2) \end{pmatrix}.
\end{aligned}$$

Es verändert sich (6.1) zu

$$\tilde{y}_i = B(\tilde{z}_i, \tilde{t}_i)\Delta + \tilde{\varepsilon}.$$

Damit lautet die Pseudo-Likelihood

$$l_p = \sum_{i=1}^{n} \left(\tilde{y}_i - B(\tilde{z}_i, \tilde{t}_i)\Delta\right)' \Sigma_\varepsilon^{-1} \left(\tilde{y}_i - B(\tilde{z}_i, \tilde{t}_i)\Delta\right) + \Delta' K(\lambda_1, \lambda_2)\Delta$$

mit $K(\lambda_1, \lambda_2)$ als Penaltymatrix in Blockdiagonaler Form mit jeweils Nullwerten für die parametrischen Terme, da diese nicht geglättet werden können. Als Lösung ergibt sich mit

$$\begin{aligned} \hat{\Delta} &= \sum_{i=1}^{n} \left(B(\tilde{z}_i, \tilde{t}_i)\Sigma_\varepsilon^{-1} B'(\tilde{z}_i, \tilde{t}_i) + K(\lambda_1, \lambda_2)\right)^{-1} \sum_{i=1}^{n} B(\tilde{z}_i, \tilde{t}_i)\Sigma_\varepsilon^{-1}\tilde{y}_i = \\ &= \left(B\Sigma_\varepsilon^{-1}B' + K\right)^{-1} B\Sigma_\varepsilon^{-1}\tilde{y} \end{aligned}$$

der penalisierte Aitkenschätzer. Desweiteren sind die Herleitung des Kreuzvalidierungskriteriums GCV im Manuskript enthalten. Außerdem werden einige Details zur Simulations-Studie besprochen. Auf die Schätzmethode wurde im Manuskript nicht eingegangen, man kann allerdings davon ausgehen, dass auch in diesem Fall Aitkens 'Two-Stage'-Schätzmethode Anwendung findet, weil das Modell nicht in GLM-Form präsentiert wird.

Kapitel 7

Nichtparametrisches Modell

Im folgenden soll ein statistisches Modell entwickelt werden, das sich mit Zeit-variierenden multiplikativen Effekten für Longitudinaldaten beschäftigt (Tutz und Meinel, 2001). Die Modellierung erfolgt über Likelihood-basierte Ansätze und die Schätzung erfolgt nichtparametrisch. 'Varying Coefficients'-Terme, wie sie von Hastie und Tibshirani (1993) definiert wurden, werden zusätzlich in das Modell mitaufgenommen. Penalisierte B-Spline Basis-Funktionen werden zum nichtlinearen Schätzen verwendet. Die Validierung des Modells erfolgt über eine Simulations-Studie. Die Studie beschäftigt sich unter anderem mit den Nachteilen von Modellen, die keine Zeit-variierenden Effekte beinhalten und stellt gleichzeitig die Vorteile von Modellen mit Zeit-variierenden multiplikativen Parametern vor. Vier verschiedene Typen von Zeit-variierenden Effekten werden untersucht. Anschließend wird das Modell auf einen 'Arbeits- und Stress'-Datensatz angewandt. Dabei zeigt sich, dass die Ergebnisse mit früheren Studien im biologisch-medizinischen Forschungs-Bereich übereinstimmen.

7.1 Einleitung

In biologischen und klinischen Studien kann es vorkommen, dass Effekte von Kovariaten, über die Zeit hinweg gemessen, verschwinden. In der folgenden Studie soll sich nun mit derartigen Effekten genauer befasst werden. Als Grundlage dient ein

Datensatz, der Messungen von Cortisolwerten von Arbeitern aus der Flugzeugindustrie in longitudinaler Form enthält. Zudem enthält der Datensatz verschiedene Kovariate, von denen BMI, kurz für Body-Mass-Index, als nichtlinearer Prädiktor verwendet wird. Als Motivation für folgende Ausführungen soll nun ein allgemeinbekannter medizinischer Zusammenhang angeführt werden. Unter anderem dokumentieren Kirschbaum und Hellhammer (1990, 1994) den Verlauf des menschlichen Cortisolspiegels im zeitlichen Verlauf eines Tages. Cortisol ist ein Prädiktor für Stress im Menschen (Dressendorfer et al., 1992). Je höher die Cortisolwerte steigen, desto mehr ist man im Stress. Üblicherweise steigen die Cortisolwerte bei Menschen kurz nach dem Aufstehen stark an und fallen dann langsam über den Tag hinweg ab, bis sie das Tagesminimum nachts während des Schlafes annehmen. Abbildung 7.1 zeigt diesen Zusammenhang. Er wurde geschätzt für Arbeiter aus der Flugzeugindustrie, über einen Tag hinweg gesehen, wobei das Aufwachen als $t = 0$ definiert wurde und die restlichen fünf Messungen, zeitlich gesehen, als Differenz zum Aufwachzeitpunkt in die Schätzung mitaufgenommen werden. Da je nach Aufwachzeitpunkt die Differenzen der Messungen und auch die Messzeitpunkte stark variieren, kann eine glatte Kurve über den Tag hinweg gefittet werden. Circa 750 Personen gingen in die Schätzung mit ein.

7.2 Modellkurzbeschreibung

Generell geht man bei folgendem Modell davon aus, dass die Kurve der Relation BMI vs. Cortisol Zeit-invariant ist, aber beeinflusst wird durch einen zeitabhängigen Intercept (Varying Coefficients) und einen Zeit-variierenden multiplikativen Effekt, der die Kurve multiplikativ vergrößert oder verkleinert. Ein ähnliches Modell wurde von Hastie und Tibshirani (1993) vorgestellt.

Von nun an gelte für die t-te Beobachtung, $t = 1, \ldots, T$, des i-ten Subjekts, $i = 1, \ldots, n$ mit Response y_{it} und Kovariaten z_{it}

$$\eta_{it} = v_{it}\beta_{0t} + \gamma_t \alpha(z_{it}). \tag{7.1}$$

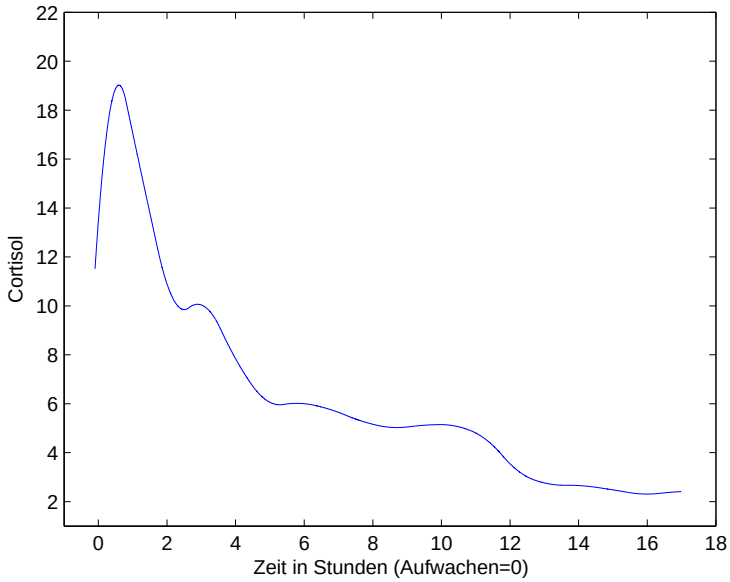

Abbildung 7.1: Schätzung der Cortisolkurve

Dabei stellt $\alpha(z_{it})$ eine unspezifizierte unbekannte Funktion dar und γ_t den multiplikativen Zeit-variierenden Effekt. Zusätzlich wird ein Term $v_{it}\beta_{0t}$ eingebaut, der als Zeit-variierender Intercept im Kontext von Varying-Coefficients-Modellen interpretiert werden kann. Da der Body-Mass-Index sich, über den Tag hinweg gesehen, nicht verändert, muss ein Sub-Modell von (7.1) gebildet werden, mit

$$\eta_{it} = v_{it}\beta_{0t} + \gamma_t\alpha(z_i), \tag{7.2}$$

so dass von nun an $\alpha(z_i) = \alpha(z_{it})$ gilt. Modell (7.2) wird in der folgenden Simulations-Studie überprüft. Zusätzlich werden weitere Analysen durchgeführt, die zeigen, dass Modell (7.2) in manchen Situationen einfachen additiven Modellen überlegen ist, beispielsweise Modellen ohne multiplikative Effekte mit $\gamma_t = 1$. Falls das reduzierte Modell $\eta_{it} = v_{it}\beta_{0t} + \alpha(z_i)$ in Situationen eingesetzt wird, in denen der multiplikative Effekt γ_t über die Zeit hinweg verschwindet, entstehen Artefakte. Im Gegensatz dazu sind Modelle mit Zeit-variierenden Parametern immer noch anwendbar, falls keine multiplikativen Effekte in den Daten vorkommen, indem annähernd konstante

Effekte $\gamma_t \approx 1$ geschätzt werden. Falls nun zusätzlich gelte, dass $\alpha(z_i)$ linear ist, stimmt Modell (7.2) exakt mit einem der von Hastie und Tibshirani (1993) vorgestellten Modelle überein, mit $\alpha(z_i) = z_i'\beta_t$ und $\eta_{it} = \beta_{0t} + z_i'\beta_t$, wobei gilt, dass $\beta_t = \gamma_t\beta$.

Zur Schätzung des Modells werden B-Spline Basis-Funktionen, wie sie von de Boor (1977, 1978) beschrieben wurden, mit Penalisierung verwendet. Die Penalisierung wurde im Jahre 1996 von Marx und Eilers unter dem Namen P-Splines, für 'Penalized B-Splines' eingeführt. Ein sehr gute Kurzzusammenfassung der Penalties findet man bei Tutz (2003). Um die Korrelationsstruktur in das Modell miteinzubeziehen, werden, nach den Vorgabe von Liang und Zeger (1986), Working-Kovarianz-Matrizen in das Modell mitaufgenommen. Die Working-Kovarianzen werden mittels der 'Method of Moments' geschätzt (Aerts et al., 2002).

7.3 Modell mit Zeit-variierenden multiplikativen Effekten

Die Beobachtungen der abhängigen Variablen seien durch $y_i = (y_{i1}, \ldots, y_{iT})'$, $i = 1, \ldots, n$, $t = 1, \ldots, T$ gegeben. Es wird angenommen, dass die Beobachtungen y_{it} über die Beziehung $y_{it} = \eta_{it} + \varepsilon_{it}$ mit normalverteilten Fehlern ε_{it} und

$$\eta_{it} = v_{it}\beta_{0t} + \gamma_t\alpha(z_{it}) \tag{7.3}$$

in das Modell eingehen. Die unbekannte Funktion $\alpha(z)$ wird durch Basisfunktionen $\{\Phi_l^\alpha\}$ approximiert. Die Basisfunktionen gehen über die lineare Beziehung

$$\alpha(z) \;=\; \sum_{l=1}^{r} \alpha_l \Phi_l^\alpha(z)$$

in das Modell ein. Alle Basisfunktionen $\Phi_1^\alpha, \Phi_2^\alpha, \ldots$ sind miteinander über die Knoten $\tau_1^\alpha < \tau_2^\alpha < \ldots$ verbunden. Die Knoten werden für den jeweiligen Wertebereich der

Variable, die in Basisfunktionen abgebildet werden soll, bestimmt. Durch geeignetes Umformen kann Gleichung (7.3) in Matrizenschreibweise überführt werden. Es ergibt sich mit $y_i = \eta_i + \varepsilon_i$ und $\eta_i = (\eta_{i1}, \ldots, \eta_{iT})'$ als linearem Prädiktor

$$\eta_i = V_i\beta_0 + \Gamma Z_i\alpha, \tag{7.4}$$

wobei $V_i = V_i(v_{it})$ und $Z_i = Z_i(z_{it})$ Designmatrizen sind, die im Appendix genauer beschrieben sind. Die restlichen Matrizen und Vektoren sind definiert durch die multiplikativen Parameter $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_T)$, die Gewichte der Basisfunktionen $\alpha' = (\alpha_1, \ldots, \alpha_r)$ und die Zeit-variierenden Intercepts $\beta_0' = (\beta_{01}, \ldots, \beta_{0T})$. Um das Modell zu vervollständigen, wird angenommen, dass gilt

$$y_i = \eta_i + \varepsilon_i,$$

mit normalverteilten Fehlern $\varepsilon_{it} \sim N(0, \Sigma)$ und unbekannter, potentiell parameterisierter Kovarianz-Matrix Σ. Um die Identifizierbarkeit zu gewährleisten sind zwei Restriktionen in das Modell einzubauen. Die erste Restriktion trennt die Varying Coefficients β_{0t} von den Spline-Gewichten α über $\alpha_r = -\sum_{l=1}^{(r-1)} \alpha_l$ und $\alpha = (\alpha_1, \ldots, \alpha_{r-1})'$. Die zweite Restriktion separiert die Parameter γ und α, indem postuliert wird, dass $\gamma_1 = 1$. Beide Restriktion führen unweigerlich zu Veränderungen in den Designmatrizen und Schätzgleichungen. Details dazu befinden sich im Appendix.

7.4 Schätzmethode

Die Schätzung basiert auf der 'Penalized Weighted Least-Squares'-Methode, bei der die Gewichte als Working-Kovarianzen interpretiert werden können. Derartige Ansätze sind im Theorieteil detailliert besprochen worden (Liang und Zeger 1986, Zeger und Liang 1986, Eilers und Marx, 1996). Ausgangsposition zur Schätzung ist die penalisierte Pseudo-Likelihoodfunktion

$$l_p = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \eta_i)'W_i^{-1}(y_i - \eta_i) - \frac{1}{2}\beta_0' K_{\beta_0}\beta_0 - \frac{1}{2}\gamma' K_\gamma \gamma - \frac{1}{2}\alpha' K_\alpha \alpha \to \max_{\beta_0,\gamma,\alpha},$$

wobei W_i eine Working-Kovarianz-Matrix darstellt und K_{β_0}, K_γ und K_α die jeweiligen Penalty-Matrizen. Die Penaltymatrizen penalisieren 'First Order', bzw. 'Higher Order'-Differenzen (Tutz, 2003) zwischen benachbarten Basisfunktionen. Detailliertere Informationen befinden sich im Appendix. In Vektor-Notation erhält man nun

$$l_p = -\frac{1}{2}(y-\eta)'W^{-1}(y-\eta) - \frac{1}{2}\beta_0' K_{\beta_0}\beta_0 - \frac{1}{2}\gamma' K_\gamma \gamma - \frac{1}{2}\alpha' K_\alpha \alpha,$$

wobei gilt dass

$$\begin{aligned} y &= (y_1', \ldots, y_n')', \\ \eta &= (\eta_1', \ldots, \eta_n')', \\ W &= \mathrm{diag}(W_1, \ldots, W_n). \end{aligned}$$

Fasst man nun alle zu schätzenden Parameter in einem Vektor $\nu = (\beta_0', \alpha', \gamma')'$ zusammen, kann man die entsprechende Scorefunktion $s_p = \partial l_p / \partial \nu$ ableiten. Die Schätzprozedur läuft hierbei in zwei iterierenden Schritten ab. Entweder β_0 und α oder β_0 und γ können gleichzeitig geschätzt werden. Der übrige bleibende Parameter (α oder γ) wird im Anschluss in einem weiteren Schritt geschätzt. Im folgenden sollen nun zunächst die Gleichungen $(\partial l_p/\partial\beta_0, \partial l_p/\partial\alpha) = 0$ gelöst werden, danach die Gleichungen $\partial l_p/\partial\gamma = 0$.

Zunächst sei nun der lineare Prädiktor $\eta_i = (V_i, \Gamma Z_i)\delta$ betrachtet, mit $\delta = (\beta_0', \alpha')'$. Damit ist der Vektor $\eta = (\eta_1', \ldots, \eta_n')'$ gegeben durch $\eta = \Phi_1 \delta$ mit Designmatrix

$$\Phi_1 = \begin{pmatrix} V_1 & \Gamma Z_1 \\ \vdots & \vdots \\ V_n & \Gamma Z_n \end{pmatrix}. \tag{7.5}$$

Die Designmatrix (7.5) besteht aus mehreren zusammengesetzten Matrizen. Im linken Teil befinden sich die Varying-Coefficients-Terme $V_i = I_{T\times T}$, im rechten Teil die Matrizen mit Basisfunktionen $Z_i' = (z_{i1}, \ldots, z_{iT})$ und $z_{it}' = (\tilde{z}_{it1}, \ldots, \tilde{z}_{itr})$ (vgl. Appendix). Damit ergibt sich das Gleichungssystem

$$\Phi_1' W^{-1}(y-\eta) - K_\delta \delta = 0 \tag{7.6}$$

zur Schätzung von δ. Außerdem gilt $K_\delta = \text{diag}(K_{\beta_0}, K_\alpha)$.

Anschließend soll nun der Parameter γ genauer untersucht werden. Seien z'_{it} die Zeilen der Designmatrix Z_i und γ die Zeit-variierenden Parameter in Vektorschreibweise. Dann lässt sich der lineare Prädiktor η_i verändern in

$$\eta_i = V_i\beta_0 + \tilde{Z}_i(\alpha)\gamma.$$

Es gilt dabei $\tilde{Z}_i(\alpha) = \text{diag}(z'_{i1}\alpha, \ldots, z'_{iT}\alpha)$. Im Matrixform ergibt sich $\eta = V\beta_0 + \Phi_2\gamma$ mit $V = (V_1, \ldots, V_n)'$ und $\Phi_2 = (\tilde{Z}'_1(\alpha), \ldots, \tilde{Z}'_n(\alpha))'$. Damit ist die Schätzgleichung äquivalent zu

$$\Phi'_2 W^{-1}(y - V\beta_0 - \Phi_2\gamma) - K_\gamma\gamma = 0. \tag{7.7}$$

Auflösen nach den Parametern δ und γ führt zu den Schätzern

$$\begin{aligned} \hat{\delta} &= (\Phi'_1 W^{-1}\Phi_1 + K_\delta)^{-1}\Phi'_1 W^{-1} y, \\ \hat{\gamma} &= (\Phi'_2 W^{-1}\Phi_2 + K_\gamma)^{-1}\Phi'_2 W^{-1}(y - V\hat{\beta}_0). \end{aligned} \tag{7.8}$$

Es ist anzumerken, dass die Matrizen Φ_1 und Φ_2 keine gewöhnlichen Designmatrizen sind, denn sie sind abhängig von den Parametern γ bzw. α, d.h. es gilt

$$\begin{aligned} \Phi_1 &= \Phi_1(\gamma), \\ \Phi_2 &= \Phi_2(\alpha). \end{aligned}$$

Falls keine Variation über die Zeit hinweg besteht, d.h. $\gamma_1 = \ldots = \gamma_T = 1$, dann ist Φ_1 unabhängig von γ und die Gleichungen aus (7.8) reduzieren sich auf das erste Gleichungssystem. Damit ist $\hat{\delta}$ eine eindeutige Lösung.

Algorithmisch gelöst werden beide Gleichungen aus (7.8) alternierend. Zunächst werden Startwerte beispielsweise für γ gewählt und im Anschluss solange iterativ geschätzt bis ein Abbruchkriterium unterschritten wurde. Der Algorithmus lässt sich im Kürze wie folgt zusammenfassen.

- Schritt 1: Seien $\hat{\gamma}^{(0)} = (1, \ldots, 1)'$ initielle Startwerte und $W_{(0)} = I$, wobei I die Einheitsmatrix ist.

- Schritt 2:

  1. Berechne

     $$\hat{\delta}^{(1)} = (\Phi_1(\hat{\gamma}^{(0)})' W_{(0)}^{-1} \Phi_1(\hat{\gamma}^{(0)}) + K_\delta)^{-1} \Phi_1(\hat{\gamma}^{(0)})' W_{(0)}^{-1} y.$$

  2. Schätze die Working-Kovarianz-Matrix $W_{(1)}$
  3. Mit $\hat{\delta}^{(1)} = (\hat{\beta}_0^{(1)'}, \hat{\alpha}^{(1)'})'$ und $W_{(1)}$ berechne

     $$\hat{\gamma}^{(1)} = (\Phi_2(\hat{\alpha}^{(1)})' W_{(1)}^{-1} \Phi_2(\hat{\alpha}^{(1)}) + K_\gamma)^{-1} \Phi_2(\hat{\alpha}^{(1)})' W_{(1)}^{-1} (y - V\hat{\beta}_0^{(1)}).$$

- Schritt 3: Wiederhole Schritt 2 und ersetze nun $\hat{\gamma}^{(c)}$ mit $\hat{\gamma}^{(c+1)}$ und $W_{(c)}$ mit $W_{(c+1)}$ bis das Abbruchkriterium erreicht ist, d.h.

  $$S = \frac{||\nu_c - \nu_{(c-1)}||^2}{||\nu_{(c-1)}||^2} < S_0.$$

  Parameter c bezeichnet hierbei den aktuellen Iterationsschritt und $\nu = (\beta_0', \alpha', \gamma')'$.

7.5 Inferenz

Beide Schätzgleichungen (7.6) und (7.7) wurden abgeleitet über eine Pseudo-Likelihood-Funktion l_p, die abhängig ist von den Gewichten W_i. Mit $\nu = (\beta_0', \alpha', \gamma')'$ als totalem Parameter-Vektor und $s_p(\nu) = \partial l_p(\nu)/\partial\nu$ als Pseudo-Score-Funktion erhält man über die 'First Order'-Taylor-Approximation $0 = s_p(\hat{\nu}) \approx s_p(\nu) + (\partial s_p/\partial\nu')(\nu' - \nu)$ die Approximation $\nu' - \nu \approx (-\partial s_p(\hat{\nu})/\partial\nu')^{-1} s_p(\nu)$. Daher kann die Kovarianz von ν über die 'Sandwich-Kovarianz-Matrix' (Liang und Zeger, 1986, Zeger und Liang, 1986). Die Schätzgleichung für die Sandwich-Matrix ist äquivalent zu

$$\hat{\text{cov}}(\hat{\nu}) = \hat{F}_p^{-1} \hat{F} \hat{F}_p^{-1}$$

mit

$$\begin{aligned}\hat{F}_p &= \sum_{i=1}^{n}\left(\frac{\partial\eta_i}{\partial\nu}\right)' W_i^{-1}\left(\frac{\partial\eta_i}{\partial\nu}\right) = \\ &= \begin{pmatrix} \hat{\Phi}_1' W^{-1}\hat{\Phi}_1 + K_\delta & \hat{\Phi}_1' W^{-1}\hat{\Phi}_2 \\ \hat{\Phi}_2' W^{-1}\hat{\Phi}_1 & \hat{\Phi}_2' W^{-1}\hat{\Phi}_2 + K_\gamma \end{pmatrix}\end{aligned}$$

und $\hat{F} = \mathrm{cov}(s_p(\nu))$, ausgewertet an der Stelle $\hat{\nu}$. Für $\hat{F}$ erhält man

$$\begin{aligned}\hat{F} &= \sum_{i=1}^{n}\left(\frac{\partial\eta_i}{\partial\nu}\right)' W_i^{-1}\Sigma W_i^{-1}\left(\frac{\partial\eta_i}{\partial\nu}\right) = \\ &= \begin{pmatrix} \hat{\Phi}_1' W^{-1}\Sigma W^{-1}\hat{\Phi}_1 + K_\delta & \hat{\Phi}_1' W^{-1}\Sigma W^{-1}\hat{\Phi}_2 \\ \hat{\Phi}_2' W^{-1}\Sigma W^{-1}\hat{\Phi}_1 & \hat{\Phi}_2' W^{-1}\Sigma W^{-1}\hat{\Phi}_2 + K_\gamma \end{pmatrix},\end{aligned}$$

wobei $\Sigma = \mathrm{cov}(\varepsilon)$. Die Kovarianz-Matrix Σ wird geschätzt durch

$$\hat{\Sigma} = (n - \mathrm{tr}(H))^{-1}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'.$$

Dabei gibt $\mathrm{tr}(H)$ die effektiven Freiheitsgrade an, basierend auf der approximativen Hatmatrix H, die im Appendix hergeleitet wurde (Hastie und Tibshirani, 1990).

Die Zerlegung der Working-Kovarianz-Matrix erfolgt nach dem allgemein bekannten Prinzip, wie es bei Fahrmeir und Tutz (2001) beschrieben ist, d.h. es gilt

$$W_i = A_i^{\frac{1}{2}} R(\theta) A_i^{\frac{1}{2}}$$

mit

$$A_i = \mathrm{diag}(\mathrm{var}(y_i)) = \mathrm{diag}(\sigma_{i1}^2, \ldots, \sigma_{iT}^2).$$

Die Working-Kovarianz-Matrix W wird mittels der Method of Moments, so wie bei Liang und Zeger (1986) beschrieben, geschätzt. In der einfachsten Form gilt $R(\theta) = I$, wobei I die Einheitsmatrix darstellt. Zwei weitere Korrelationsstrukturen fließen in die folgenden Simulationen mit ein. Dies ist zum eine das 'Exchangeable

Correlation Model', auch bekannt unter dem Namen 'Compound-Correlation', und das 'Autoregressive Model', meist abgekürzt mit 'AR(1)'. Die Schätzungen beider Möglichkeiten beinhalten die Pearson-Residuen

$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{\hat{\sigma}_{it}}}.$$

Der Parameter θ des Exchangeable Correlation Modells $\text{corr}(y_{ij}, y_{ik}) = \theta$, $j, k = 1, \ldots, T$ wird konsistent geschätzt mit

$$\hat{\theta} = \frac{1}{(\frac{n}{2}T(T-1) - \text{tr}(H))} \sum_{i=1}^{n} \sum_{k>j} \hat{r}_{ik}\hat{r}_{ij}.$$

Im Falle des autoregressiven Modells $\text{corr}(y_{it}, y_{it+k}) = \theta^k$ schätzen Aerts et al. (2002) θ wie folgt

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T-1} \sum_{t \leq T-1} \hat{r}_{it}\hat{r}_{i,t+1}.$$

Kapitel 8

Simulations-Studie Nichtparametrisches Modell

In einer Simulations-Studie soll das Regressions-Modell

$$y_{it} = \gamma_t sin(x_i) + \varepsilon_{it}$$

überprüft werden. Hierbei werden $n = 120$ und $T = 10$ gewählt. Desweiteren werden die x_i gleichverteilt aus dem Bereich $[0, 2\pi]$ gezogen. Die Basisfunktionen werden zwischen 15 äquidistanten Knoten aufgespannt. Als Gütekriterium zur Beurteilung der Schätzung wird der 'Mean Squared Error' (MSE) herangezogen. Für jede Simulation s definiert man den MSE über die Formel

$$\text{MSE}_s \quad = \quad \sum_{i=1}^{n} \sum_{t=1}^{T} (\hat{\eta}_{it} - \eta_{it})^2.$$

Zur generellen Bewertung wird ein Mittelwert des MSE über alle Simulationen berechnet.

8.1 Settings der Simulations-Studie

Zugrundeliegende Kovarianzstruktur $\mathbf{cov}(\varepsilon_i) = \Sigma$

- (I) Independence: $\Sigma = \sigma^2_{ind} I$

1. $\sigma^2_{ind,1} = 0.1$ (I1)
2. $\sigma^2_{ind,2} = 0.5$ (I2)

- (CS) Compound-Symmetry: $\Sigma = \sigma^2_0 I + \sigma^2_{comp} 1_{(t \times t)}$
  1. $\sigma_{0,1} = 0.05$, $\sigma^2_{comp,1} = 0.05$, resultierend in $\sigma_{0,1} + \sigma^2_{comp,1} = 0.1$ (CS1)
  2. $\sigma_{0,2} = 0.4$, $\sigma^2_{comp,2} = 0.1$, resultierend in $\sigma_{0,2} + \sigma^2_{comp,2} = 0.5$ (CS2)
- (AR) Autoregressive Korrelation AR(1): $\Sigma = (\sigma^2_{AR} \rho^k)$
  1. $\sigma_{AR,1} = 0.1$, $\rho = 0.4$ (AR1)
  2. $\sigma_{AR,2} = 0.5$, $\rho = 0.4$ (AR2)

Konfiguration der Parameter γ

- Additives Model: (konstant)
  (AD) $\gamma = (1, 1, \ldots, 1)$
- Multiplikatives Model 1: (linear ansteigend)
  (M1) $\gamma = (1, 1.1, 1.2, \ldots, 1.9)$
- Multiplikatives Model 2: (exponentiell fallend, 'time-vanishing')
  (M2) $\gamma = (1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{10})$

Working-Kovarianz Struktur

- Independence:
  (IW) $W_{ind} = \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}} I \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}}$
- Compound-Symmetry:
  (CSW) $W_{comp} = \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}} R(\hat{\theta}) \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}}$
- Autoregressive Korrelation AR(1):
  (ARW) $W_{AR} = \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}} R(\hat{\theta}) \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}}$

Es ist offensichtlich, dass alle drei Konfigurationen AD, M1, M2 gleichen Varianzen unterliegen. Dadurch wird Vergleichbarkeit zwischen den Konfigurationen garantiert. Die Working-Kovarianz wird jeweils so gewählt, dass sie der simulierten Situation entspricht. Im folgenden werden nun zwei zusätzliche Modelle eingeführt. Eines wird als 'Additives Modell (A)' ohne Zeit-variierende Modifikation bezeichnet, das andere als Multiplikatives Modell (M) mit Zeit-variierender Modifikation γ_t.

8.2 Wahl der Glättungsparameter

Das Additive Modell (A) beinhaltet zwei Glättungsparameter λ_α und λ_{β_0}, das Multiplikative Modell (M) drei Glättungsparameter λ_α, λ_{β_0} und λ_γ. Alle Glättungsparameter werden über einen Grid-Search ermittelt. Als Gütekriterium wird das 'Bayesian Information Criterion' (BIC) für Likelihood-basierte Modelle (Schwarz, 1978)

$$\text{BIC} \quad = \quad -2l_p + log(nT)tr(H). \tag{8.1}$$

verwendet. Das Gütekriterium BIC wird von der Likelihoodtheorie auf die Pseudolikelihoodtheorie übertragen.

Der Glättungsparameter λ_{β_0} wird während der Simulationen konstant auf 10^6 gehalten. Dies ist zum einen aufgrund der Vergleichbarkeit unter den verschiedenen Modellen notwendig, andererseits weil die diskreten, Zeit-variierenden Intercepts $\beta_{0t} = 0$ bei $\lambda_{\beta_0} = +\infty$ optimal geglättet sind. Da man den Glättungsparameter nicht auf 'unendlich' setzen kann, wurde ein sehr hoher Wert mit $\lambda_{\beta_0} = 10^6$ gewählt, um ausreichend glatte diskrete Schätzungen zu erhalten.

8.3 Additives Modell (A) ohne Zeit-variierende Modifikation

Der erste Teil der Simulations-Studie beschäftigt sich damit zu zeigen, dass das Additive Modell (A) mit dem Prädiktor $\eta_{it} = v_{it}\beta_{0t} + \gamma_t\alpha(z_i)$ und $\gamma_t = 1$ in einigen Situationen ungeeignet ist. Daher untersuchen wir zunächst das Additive Modell (A) in Situationen, in denen die simulierten Daten auf Konfiguration AD basieren. Im Anschluss wird das Additive Modell (A) für die Konfigurationen M1 und M2 geschätzt, d.h. multiplikative Effekte ungleich 1 werden in die Daten integriert, aber werden bei der Schätzung nicht berücksichtigt. Dies führt unweigerlich zu Artefakten.

8.3.1 Settings

Beim Fitten des Additiven Modells (A) muss nur das erste Gleichungssystem aus (7.8) gelöst werden. Damit reduziert sich das Modell zu einem generalisierten additiven Modell mit identischer Linkfunktion. Die Schätzung erfolgt in zwei Schritten nach dem Prinzip von Aitkens Two-Stage-Methode (vgl. Kapitel 6.1) . Zunächst werden alle Parameter mit $W = I$ geschätzt und die Working-Kovarianz berechnet, danach wird das Modell wieder geschätzt mit $W = W(\theta)$. Die Simulations-Studie basiert auf 100 Datensätzen. Das Additive Modell (A) wird mit verschiedenen Kovarianz- und Working-Kovarianz-Strukturen simuliert (vgl. Kapitel 8.1).

8.3.2 Ergebnisse - Additives Modell (A)

Die Simulations-Studie des Additiven Modells (A) ist in zwei Teile zerlegt worden. Der erste Teil enthält Schätzungen für die jeweiligen Konfigurationen mit niedriger Varianz (I1, AR1 und CS1), der zweite Teil für Konfigurationen mit hoher Varianz (I2, AR2 und CS2). Die Ergebnisse beider Teile sind in Tabelle 8.1 zusammengefasst. Die obere Tabelle zeigt die Schätzungen für niedrige Varianzen, die untere Tabelle für hohe Varianzen. Die generellen Settings bleiben unverändert in beiden Tabellen.

	Additives Model (LV) ungeglättet			Additives Model (LV) geglättet		
	Modell 1 (I1)	Modell 2 (AR1)	Modell 3 (CS1)	Modell 1 (I1)	Modell 2 (AR1)	Modell 3 (CS1)
AD	2.59	4.42	9.78	1.31	2.07	5.22
M1	53.31	55.15	60.42	52.58	53.64	56.94
M2	45.11	46.90	52.26	43.63	44.40	46.37
	Additives Model (HV) ungeglättet			Additives Model (HV) geglättet		
	Modell 1 (I2)	Modell 2 (AR2)	Modell 3 (CS2)	Modell 1 (I2)	Modell 2 (AR2)	Modell 3 (CS2)
AD	12.95	22.11	27.29	4.79	9.08	11.45
M1	63.64	72.74	77.94	56.44	63.89	63.87
M2	53.44	64.46	69.79	46.11	50.58	50.49

Tabelle 8.1: Additives Modell (A): Durchschnittliche geglättete und ungeglättete Mean Squared Errors für die Konfigurationen AD, M1 und M2 für niedrige und hohe Varianzen (niedrige Varianz (LV): Modell 1: zugrundeliegende Kovarianz-Struktur I1, zugrundeliegende Working-Kovarianz-Struktur IW, Modell 2: AR1, ARW, Modell 3: CS1, CSW, hohe Varianz (HV): Modell 1: I2, IW, Modell 2: AR2, ARW, Modell 3: CS2, CSW)

Aus Tabelle (8.1) ist ersichtlich, dass der MSE für korrelierte Datenstrukturen generell höher ausfällt. Zudem fällt auf, dass das Fitten eines Additiven Modells (A) in Situationen, in denen das zugrundeliegende Modell Zeit-variierende multiplikative Effekte enthält, zu deutlich erhöhten MSEs führt. Glätten reduziert die MSEs zwar, teilweise sogar stark, bis zu 58% (27.29 $\rightarrow$ 11.45), allerdings verbessern sich die Schätzungen nicht in dem Maße, dass die originale Sinus-Funktion vom Konfidenzbereich überdeckt wird.

Abbildung 8.1 zeigt die Effekte auf die Schätzkurve, wenn das zugrundeliegende Modell multiplikativ ist. Abbildung 8.1(a) zeigt die geschätzte Kurve für die Parameter γ, die konstant bei 1 liegen (Konfiguration AD, Modell 1). Dies ist die einzige Situation, in der das Additive Modell (A) passende Schätzungen liefert. In allen anderen Situationen werden Artefakte geschätzt, wie z.B. für linear ansteigende (M1) bzw. exponentiell abfallende (M2) Parameter γ. Die Schätzungen für die letzten beiden Parameterkonstellationen sind in Abbildung 8.1(b),(c) vorzufinden. Falls γ anwächst wird die Sinus-Funktion deutlich überschätzt (Abb. 8.1(b)), im Falle von fallenden Parametern γ ergibt sich eine starke Unterschätzung des Modells (Abb. 8.1(c)).

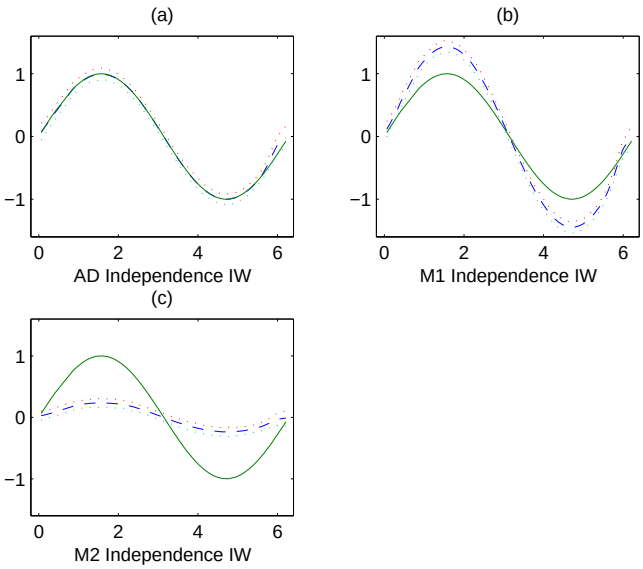

Abbildung 8.1: Geglättete Schätzungen des Additiven Modells (A) mit Independence-Working-Kovarianz-Struktur (IW, LV, Modell 1) für die Konfigurationen AD, M1 and M2 (gestrichelte Linie: Schätzung, gepunktete Linie: empirische Konfidenzintervalle, durchgezogene Linie: zugrundeliegende Sinus-Funktion, analog in allen folgenden Abbildungen).

Abbildung 8.2 zeigt Schätzungen der Sinus-Funktion für korrelierte und unkorre-

lierte Daten im additiven Modell (A) für die Konfiguration (AD). Dabei gehen verschiedene Korrelationsstrukturen in die Daten ein. Abbildung 8.2(a) basiert auf Schätzungen unkorrelierter Daten, während Abbildung 8.2(b) eine autoregressive Struktur aufweist. Die Abbildungen 8.2(c),(d) unterliegen der Compound-Struktur. Abbildung 8.2(a) zeigt ein AR(1)-Modell mit Independence Working-Kovarianz-Annahme und sehr nahe liegenden Konfidenzintervallen. Dies zeigt, dass sich bei niedrigen Varianzen generell sehr gute Schätzungen mit niedrigem MSE ergeben. Nach Einführung von AR-Korrelation und Schätzung mit ARW-Struktur verbreitern sich die Konfidenzintervalle nun etwas, wie Abbildung 8.1(b) zeigt. Dies war zu erwarten. Abbildung 8.1(c) zeigt die Schätzungen für Compound-korrelierte Daten. Auch in diesem Fall ist eine Verbreiterung der Konfidenzintervalle im Vergleich zu Abbildung 8.2(a) zu erkennen. Diese fällt sogar etwas stärker aus als im Falle des AR-Modells. Die MSEs aus Tabelle 8.1 bestätigen diese Analysen. Generell haben Compound-Schätzungen die höchsten MSEs, während die AR-Schätzungen kleinere und Modelle ohne Einfluss von Korrelation die kleinsten MSEs aufweisen. Abbildung 8.2(d) zeigt die geglätteten Schätzer des Compound-Modells aus Abbildung 8.2(c). Ein leichter Bias ist in Form einer minimalen Unterschätzung des Sinus zu erkennen, die Konfidenzintervalle liegen jedoch deutlich näher als noch in Abbildung 8.2(c). Glätten führt also zu kleineren Varianz-Schätzungen, allerdings auf Kosten eines (leichten) Bias.

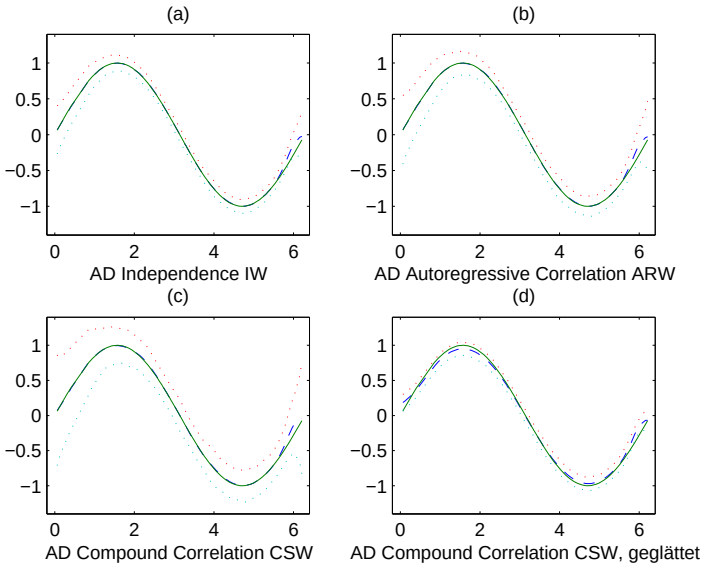

Abbildung 8.2: Einführung von Korrelationen in die Daten. Schätzungen für verschiedene Arten an Korrelations-Strukturen, geglättet und ungeglättet.

8.4 Multiplikatives Modell (M) mit Zeit-variierender Modifikation

Im folgenden wird das Multiplikative Modell (M) genauer betrachtet. Die Daten und auch die Konfigurationen ändern sich nicht. Parameter γ wird von nun an in die Schätzungen miteinbezogen und entweder konstant bei 1, linear ansteigend oder exponentiell fallend gewählt.

8.4.1 Settings

Das Fitten eines multiplikativen, Zeit-variierenden Modells (M) basiert auf dem iterativen Algorithmus, der in Sektion 7.5 beschrieben wurde. Beide Gleichungssysteme werden abwechselnd gelöst. Die Startwerte liegen bei $\gamma = (1, 1, \ldots, 1)'$ und die Working-Kovarianz-Matrix ist $W = I_{(nT \times nT)}$. Wie auch das Additive Modell (A) wird das Multiplikative Modell (M) mit verschiedenen Typen an Kovarianzen und Working-Kovarianzen geschätzt.

8.4.2 Ergebnisse - Multiplikatives Modell (M)

Tabelle 8.2, äquivalent zu Tabelle 8.1, führt die Ergebnisse der Schätzungen des Multiplikativen Modells (M) für niedrige und hohe Varianzen mit und ohne Glättung an. Die Schätzungen für Konfiguration AD bleiben nahezu unverändert, während eine Verkleinerung des MSE bei den Konfigurationen M1 und M2 beim Vergleich von Tabelle 8.1 und 8.2 beobachtet werden kann. Die Verbesserung begründet sich auf der Tatsache, dass das Multiplikative Modell (M) nicht nur für konstante Parameter γ geeignet ist, sondern auch für Zeit-variierende Parameter. Beim Vergleich aller MSEs aus Tabelle 8.2 zeigt sich, dass nun die MSEs der drei Konfigurationen AD, M1 und M2 pro Modell annähernd auf einem Niveau liegen.

Abbildung 8.3 präsentiert die entsprechenden Schätzungen für die Konfigurationen AD, M1 und M2. Im Vergleich mit Abbildung 8.1 zeigen sich die positiven

	Multiplikatives Modell (LV) ungeglättet			Multiplikatives Modell (LV) geglättet		
	Modell 1 (I1)	Modell 2 (AR1)	Modell 3 (CS1)	Modell 1 (I1)	Modell 2 (AR1)	Modell 3 (CS1)
AD	3.48	5.20	10.94	1.61	2.91	5.99
M1	3.52	5.16	9.77	2.15	3.26	6.19
M2	3.56	4.43	4.61	1.85	2.39	2.86
	Multiplikatives Modell (HV) ungeglättet			Multiplikatives Modell (HV) geglättet		
	Modell 1 (I2)	Modell 2 (AR2)	Modell 3 (CS2)	Modell 1 (I2)	Modell 2 (AR2)	Modell 3 (CS2)
AD	17.57	26.26	32.49	6.78	12.25	15.29
M1	17.55	25.91	31.26	7.69	16.36	18.99
M2	18.66	23.18	22.84	7.69	12.25	10.85

Tabelle 8.2: Multiplikatives Model (M): Durchschnittliche, ungeglättete und BIC-optimal geglättete MSEs für die Konfigurationen AD, M1 and M2 für niedrige und hohe Varianzen (vergleiche mit Tabelle 8.1).

Auswirkungen des Multiplikativen Modells (M) auf die Schätzungen bei den Konfigurationen M1 und M2. Während sich das Additive Modell (A) stark verschätzt, befinden sich beim Multiplikativen Modell (M) die geglätteten Schätzungen nahe am vorgegebenen Sinus. Die Konfidenzintervalle überdecken nun auch die vorgegebene Sinus-Kurve. Bei genauerer Betrachtung von Abbildung 8.3(a), (c) wird wieder eine leichte Unterschätzung des Sinus sichtbar. Diese ist auf die Glättung der Kurve zurückzuführen (Bias-Varianz-Tradeoff). Generell kann man anfügen, dass die Konfidenzintervalle im Multiplikativen Modell (M) etwas größer ausfallen als im Additiven Modell (A). Dies ist auf den zusätzlichen Parameter γ zurückzuführen.

Abbildung 8.4 beschäftigt sich mit den entsprechenden Parametern γ. Jede Ein-

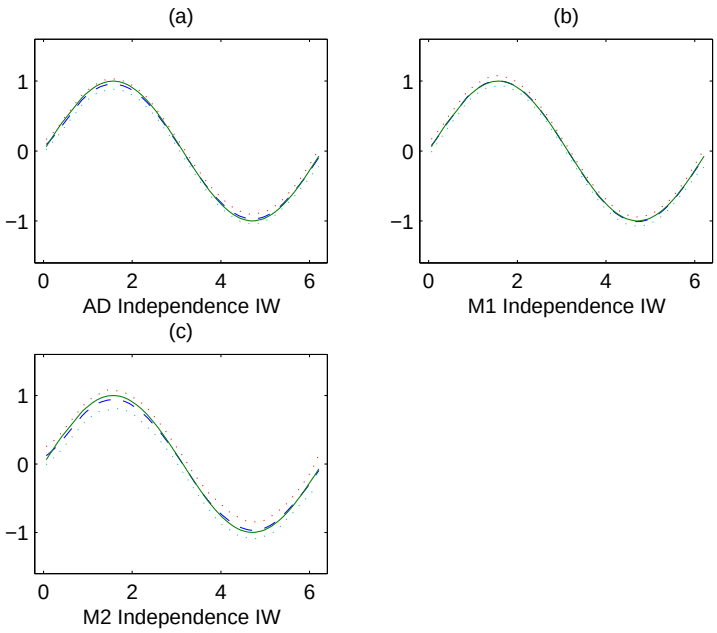

Abbildung 8.3: Multiplikatives, BIC-optimal geglättetes Modell (M) mit Independence-Working-Kovarianz-Struktur (IW, LV, Modell 1) für die Konfigurationen AD, M1 and M2 (vgl. Abbildung 8.1).

zelabbildung zeigt neun Boxplots, die die Variation von $\gamma = (\gamma_2, \ldots, \gamma_{10})'$ über die

100 Simulationen zusammenfassen. Parameter γ_1 wird hierbei aufgrund der zuvor eingeführten Restriktion als konstant angesehen.

Abbildung 8.4(a) zeigt die ungeglätteten Schätzungen von $\gamma_t = 1$ für Konfiguration AD. Die Schätzungen scheinen auf den ersten Blick besser zu passen als die geglätteten Schätzungen aus Abbildung 8.4(b). Allerdings wird erst bei der Betrachtung von Abbildung 8.3(a) klar, dass eine leichte Unterschätzung des Sinus zu einer Überschätzung der Parameter γ (oder umgekehrt) führt (vgl. Abbildung 8.4(b)). Damit wird die Unterschätzung des einen Parameters durch die Überschätzung des anderen ausgeglichen und der MSE reduziert sich deutlich von 3.48 auf 1.61 (vgl. Tabelle 8.2). Alle Abbildungen 8.3(b-d) zeigen geglättete Schätzungen. Abbildung 8.3(c) zeigt sehr gute Schätzungen, die nahe am Optimum liegen und die Schätzungen aus Abbildung 8.3(d) weisen, wie in Abbildung 8.3(b), einen leichten Bias auf.

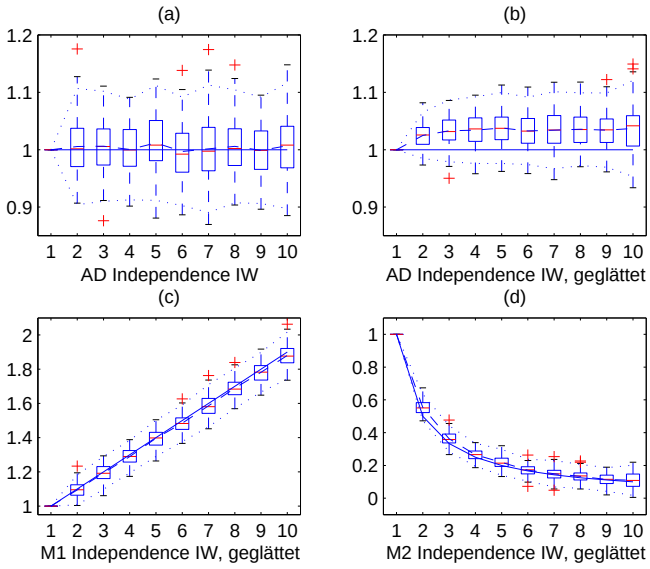

Abbildung 8.4: Schätzer und Konfidenz-Intervalle von γ für die Konfigurationen AD, M1 und M2, mit und ohne Glättung.

Abbildung 8.5 enthält acht Einzelabbildungen, die linear ansteigende und exponentiell abfallende Schätzer und auch die Kurvenschätzungen zeigen. Die Abbildungen 8.5(a-d) unterliegen demselben Modell mit gleicher Korrelationsstruktur, allerdings unterscheiden sie sich bezüglich der Varianz. Die Abbildungen 8.5(a),(c) stellen die Schätzungen der Kurven für niedrige und hohe Varianz dar, während sich die Abbildungen 8.5(b),(d) auf die entsprechenden Parameter γ konzentrieren. Beide Abbildungen 8.5(a),(b) zeigen sehr gute Schätzungen. Wird die Varianz allerdings erhöht, weisen die Schätzungen einen leichten Bias auf und die Konfidenzintervalle verbreitern sich stark (vgl. Abbildung 8.5(c),(d)). Der Bias ist damit begründet, dass das BIC-Kriterium nicht immer in dem Punkt optimal ist, an dem der MSE einen sehr niedrigen oder den niedrigsten Wert annimmt. Vor allem bei hohen Varianzen kommt es manchmal vor, dass Biases entstehen, die auf falsche Wahl der Glättungsparameter und damit indirekt auf das BIC-Kriterium zurückzuführen sind.

Die übrigen vier Abbildungen 8.5(e-h) illustrieren zwei Modelle mit exponentiell abfallenden Parametern γ, die sich nur bezüglich der Wahl der Working-Kovarianz-Struktur unterscheiden. Die Abbildungen 8.5(e),(f) wurden mit Independence-Annahme geschätzt, während bei den Abbildungen 8.5(g),(h) eine Compound-korrelierte Working-Kovarianz-Matrix Anwendung fand. Es sind kaum Unterschiede festzustellen, allerdings werden generell die Ergebnisse von Liang und Zeger (1986) bestätigt, da sich der MSE (marginal) reduziert falls die Working-Kovarianz-Matrix richtig spezifiziert wurde.

8.5 Multiplikative Parameter γ mit Richtungswechsel

Zusätzliche Simulationen wurden für die multiplikativen Parameter γ mit Richtungswechsel durchgeführt. Dies war notwendig, da der Datensatz, der im nächsten Abschnitt verwendet wird, sowohl Intercepts als auch multiplikative Effekte mit Rich-

tungswechsel beinhaltet. Die Anzahl an Zeitpunkten wurde von 10 auf 6 reduziert, gemäß dem Datensatz, und die multiplikativen Parameter als $\gamma = [1\ 2.5\ 1.8\ 1.2\ 0.9\ 0.6]$ (Konfiguration M3) gewählt. Die übrigen Settings wurden nicht verändert.

Beide oberen Abbildungen 8.6(a),(b) stellen die Schätzung der Kurve und die zugehörigen Parameter γ für das Compound-Modell mit niedriger Varianz dar. Selbst bei Richtungswechsel der Parameter mit einem starken Anstieg von 1 auf 2.5 und einem stetigen Abfall während der restlichen vier Zeitpunkte sind die Schätzungen aller Parameter nahezu optimal. Alle Schätzungen liegen sehr nahe am vorgegebenen Wert und der Sinus wird vollständig von den Konfidenzintervallen überdeckt. Ähnlich gute Schätzungen können in den anderen beiden Abbildungen 8.6(c),(d) für das AR(1)-Modell beobachtet werden. Auf das Glätten der Parameter γ musste allerdings verzichtet werden, da weder 'First Order'- noch 'Second Order'-Penalties an den Richtungswechsel angepasst werden konnten. Glättungsparameter λ_{β_0} wurde wieder konstant gehalten.

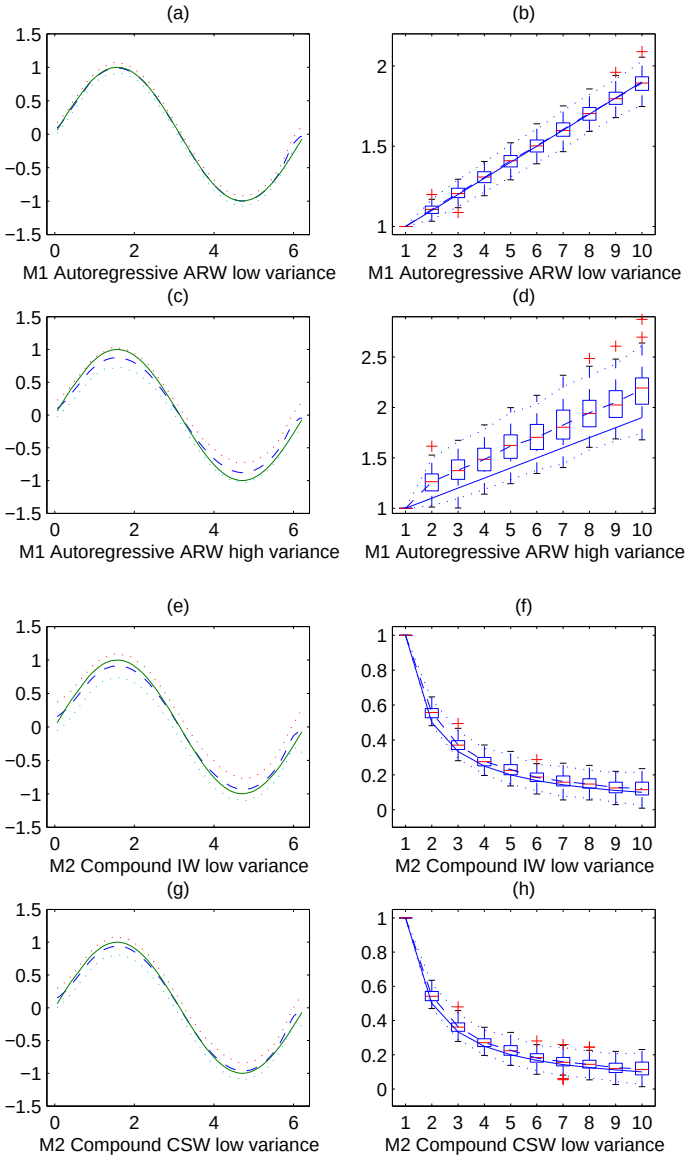

Abbildung 8.5: Schätzer und 95% Konfidenzintervalle für Parameter γ für die Konfigurationen M1 and M2, basierend auf niedrigen und hohen Varianzen und verschiedenen Working-Kovarianz-Strukturen.

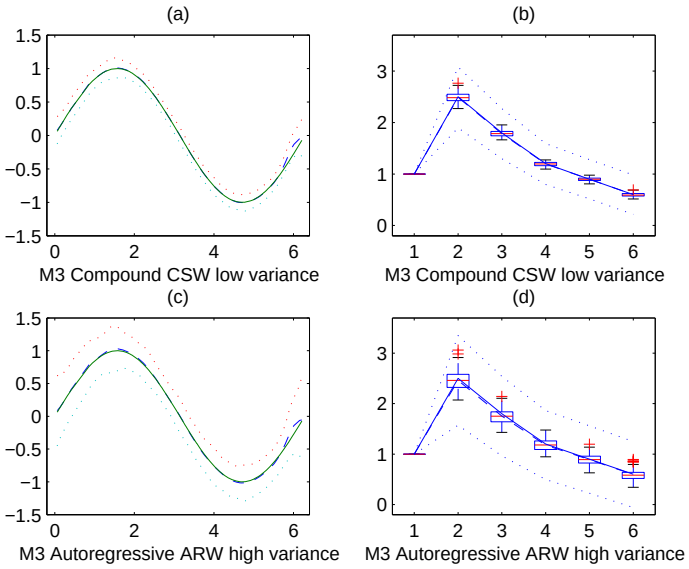

Abbildung 8.6: Parameter γ mit Richtungswechsel: BIC-optimale Schätzungen und Konfidenzintervalle für Konfiguration M3 für das Compound-korrelierte Modell mit niedriger Varianz (a-b) und das autoregressive Modell mit hoher Varianz (c-d), geglättet.

8.6 Abbruchkriterien, Thresholds und der Einfluss von Korrelation auf den MSE

Abbildung 8.7 liefert einige Zusatzanalysen. Zunächst wird in Abbildung 8.7(a) auf den Einfluss von steigender Korrelation auf den MSE eingegangen. Hierbei wurde ein autoregressives Modell mit Konfiguration AD und Independence-Working-Kovarianz-Matrix ausgewählt und die Korrelation in Schritten von 0.05 erhöht, bis die maximale Korrelation von 0.99 erreicht war. Die Ergebnisse wurden in Abbildung 8.7(a) geplottet und es fällt auf, dass sich der MSE mit steigender Korrelation exponentiell erhöht.

Desweiteren ist die Analyse der Abbruchkriterien von Interesse. Am Ende jeder Iteration wird bekanntlich das Abbruchkriterium neu berechnet. Verändern sich die geschätzten Parameter stark, so wird der Wert hoch ausfallen und eine weitere Iteration wird folgen, da eine Verbesserung der Schätzer zu beobachten war. Falls keine ausreichende Verbesserung eintritt, terminiert das Programm. Für verschiedene Parameterkonstellationen treten verschiedene Verläufe für die Abbruchkriterien auf. Zwei dieser Verläufe sind in Abbildung 8.7 zusammengefasst. Abbildung 8.7(b) zeigt einen typischen Verlauf für Konfiguration M2, mit exponentiell abfallenden Parametern γ. Zwei Minima können beobachtet werden. Das erste lokale Minimum würde die Iterationen bei zu hoch gewähltem Abbruchkriterium sehr frühzeitig beenden und eventuell in unzureichenden Schätzern resultieren. Das zweite globale Minimum liefert im allergrößten Teil der Fälle abgesicherte Schätzungen. Abbildung 8.7(c) konzentriert sich auf den typischen Verlauf von Abbruchkriteria-Kurven für die Konfigurationen AD und M1, mit einem starken Abfall während der ersten Iterationen, gefolgt von einem stetigen moderaten Sinken der Kurve. Wählt man in diesem Fall den Wert für das Abbruchkriterium zu hoch, kann es dazu führen, dass sich Schätzungen mit einem starken Bias ergeben (vgl. Abbildung 8.7(d)).

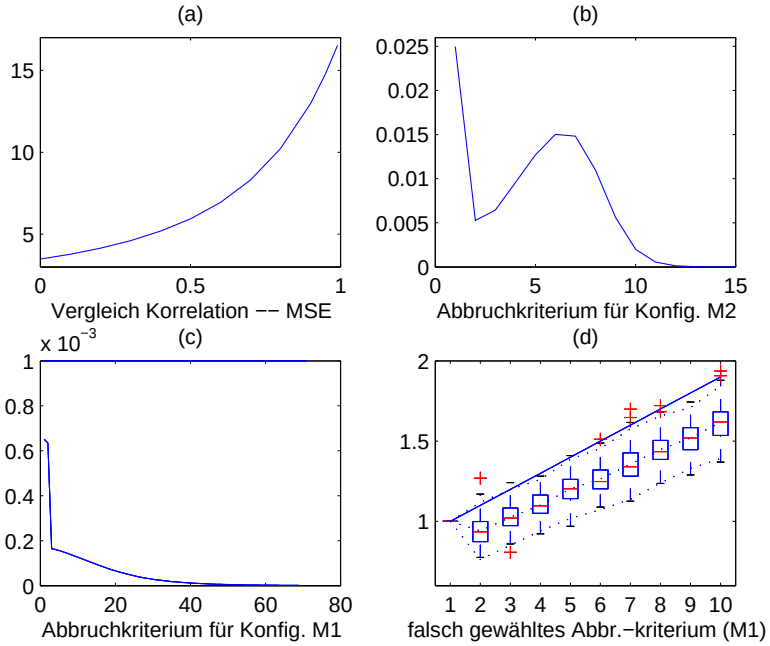

Abbildung 8.7: Zusatzanalysen

Kapitel 9

Anwendungsbeispiel - Cortisoldaten

Das im vorigen Kapitel beschriebene Modell soll im folgenden an einem Datensatz überprüft werden. Der Datensatz wurde in einer Studie der ETH Zürich erhoben. Die Studie wurde in der Flugzeugbau-Industrie durchgeführt und beschäftigt sich vor allem mit dem Thema Stress am Arbeitsplatz. Sie ist ausgelegt als Kohortenstudie und zielt darauf ab, psychologische Merkmale im Arbeitsbereich mit gesundheitlichen Merkmalen in Verbindung zu bringen und zu beurteilen, ob Veränderungen bei den Angestellten über die Zeit hinweg auftreten. Zunächst wurde ein Zeitpunkt erhoben. Weitere Messungen werden folgen.

Wie bereits erwähnt, wird ein nichtlinearer Zusammenhang zwischen Body-Mass-Index (BMI) und den Cortisolwerten, gemessen über den Tag hinweg, modelliert. Dazu muss zunächst erläutert werden, warum es von Interesse ist diese beiden Variablen zu modellieren. Außerdem soll kurz darauf eingegangen werden, was in bisherigen Studien im medizinisch-psychologischen Bereich zu diesem Thema veröffentlicht wurde.

Über den ganzen Tag hinweg werden von Menschen verschiedenste Arten an Reizen

aufgenommen. Dies können beispielsweise visuelle Reize, d.h. Reize, aufgenommen über die Augen, sein oder auditorische Reize, aufgenommen über die Ohren, sein. Zunächst wird ein Reiz im menschlichen Gehirn weiterverarbeitet. Dabei gibt es zwei verschiedene Arten der Weiterverarbeitung, eine direkte und eine indirekte. Direkte Weiterverarbeitung läuft auf der unbewussten Ebene ab, beispielsweise in Gefahrensituationen, in denen schnell reagiert werden muss. Vom Gehirn aus wird in solchen Situationen der Hypothalamus aktiviert, der der Dirigent des autonomen Nervensystems ist. Das autonome Nervensystem reguliert und aktiviert die inneren Organe eines Menschen, z.B. das Herz, die Muskeln und andere Organe, und versetzt den Körper in Gefahrensituationen in Alarmbereitschaft. Die Regulierung erfolgt über die Hirnanhangdrüse (Hypophyse), die Hormone ausschüttet, um die Regulation durchzuführen. Die Hypophyse erhält vom Hypothalamus Befehle. Beispielsweise könnte der Befehl lauten das Hormon ACTH auszuschütten. Das Hormon löst im menschlichen Körper die Ausschüttung von Cortisol aus. Die Ausschüttung von Cortisol erfolgt in der Nebennieren-Rinde. Falls Cortisol im Blut, bzw. Speichel des Menschen nachweisbar ist, spricht man von einer Stressreaktion. Der Verlauf des Cortisolspiegels ist in Abbildung 7.1 zusammengefasst und zeigt eine starke Stressreaktion kurz nach dem Aufstehen. Diese Stressreaktion wird allgemein als 'Morning-Peak' bezeichnet. Über den Tag hinweg fällt der Cortisolspiegel kontinuierlich ab. Detaillierte Informationen zum Thema findet man bei Schmidt und Thews (1995) und bei Kendal (2000).

Der Gesamtablauf oder -komplex, der hier auf sehr einfache Art und Weise zusammengefasst wurde, wird als 'Hypothalamus / Hypophysen-Nebennieren-Rinden'-Achse, im englischen als 'Hypothalamic-Pituitary-Adrenal-Axis' (HPA-Axis) bezeichnet. Von Kirschbaum und Hellhammer (1990,1994) wurde gezeigt, dass die Funktion der HPA-Achse noch weitere nachgelagerte biologische Maße beeinflusst. Der Hauptindikator für die Funktion der HPA-Achse ist die Messung von Cortisolwerten über den Tag hinweg. Cortisol kann sehr einfach über die Konzentration im Speichel gemessen werden, weil die Speichelcortisolwerte sehr hoch mit dem ungebundenen freien Cortisol im Plasma korrelieren (Dressendorfer et. al.,

1992). Wie bereits beschrieben, wird das Cortisol in der Nebenniere produziert. Durch die Ausschüttung werden eine Vielzahl an biologischen Funktionen beeinflusst. Beispielsweise treten Veränderungen von intrazellulären Vorgängen auf, die dann zu weiteren Veränderungen unter anderem bei der Protein-Produktion und Protein-Bildung führen. Die Menge des Cortisols im Blut wird über einen negativen Feedback-Mechanismus im Gehirn geregelt (Kirschbaum und Hellhammer, 1994).

Cortisol-Sekretion folgt dem beschriebenen täglich annähernd gleichen Ablauf (siehe Abbildung 7.1). Nachts erreicht der Cortisolwert ein Minimum. Kurz vor dem Aufwachen erhöht sich der Cortisolwert dann wieder leicht und erreicht circa eine halbe Stunde nach dem Aufwachen sein Maximum, bevor wieder der stetige, aber langsame Nachlassen einsetzt. Treten stressige Situationen während des Tages auf, z.B. Autofahren zur Rush Hour, wird die normale Cortisolkurve durch einen kurz andauernden Stress-Peak überlagert. Bei mehreren zusätzlichen Stressreaktionen kann die Cortisolkurve unregelmäßig verlaufen, mit je einem zusätzlichen Peak für jede Stresssituation. Sehr viel Literatur wurde über die Cortisol-Stress-Beziehung veröffentlicht. Allerdings gibt es sehr wenige Studien, die sich über mehrere Zeit- und Messpunkte hinweg mit dem Thema Stress im Allgemeinen befassen. Kleinere Studien zeigten sehr unterschiedliche Ergebnisse, beispielsweise im Bereich Cortisol vs. BMI. BMI ist definiert als das Verhältnis von Gewicht zu Körpergröße im Quadrat.

Um etwas genauer auf den zugrundeliegenden Datensatz einzugehen, sei erwähnt, dass 749 Angestellte aus der Flugzeugindustrie in Süddeutschland teilnahmen. Jeder Angestellte sollte 6-mal täglich eine Speichelcortisolprobe von sich selbst nehmen, und zwar morgens kurz nach dem Aufwachen, dann eine halbe Stunde nach dem Aufwachen, um den Morning-Peak zu erfassen. Die weiteren vier Zeitpunkte waren auf 9.00 Uhr, 11.00 Uhr, 15.00 Uhr und 20.00 Uhr festgelegt worden, um das stetige Nachlassen des Cortisolspiegels über den Tag hinweg zu dokumentieren. Insgesamt wurden 19213 Proben genommen, allerdings reduzierte sich die Anzahl für diese Studie, weil noch nicht alle Proben ausgewertet werden konnten. Damit verringer-

te sich die Anzahl auf 330 Personen, von denen diejenigen ausgeschlossen wurden, die abnormale Cortisolverläufe zu einem der Messzeitpunkte aufwiesen. Dies schloss beispielsweise Personen aus, die erhöhte Cortisolwerte hatten, weil sich während der Arbeitszeit eine nichtvorhersehbare Stresssituation ergab, z.B. ein Unfall oder ähnliches. Außerdem wurden Angestellte mit schweren Depressionen oder Personen in schlechtem medizinischen Zustand von der Analyse ausgeschlossen. Es musste zudem darauf geachtet werden, dass die sechs Messzeitpunkte bei allem Teilnehmern der Studie zeitlich ungefähr übereinstimmten, um Vergleichbarkeit zu gewährleisten. Dies senkte die Anzahl an Personen noch einmal zusätzlich auf 262.

Im folgenden soll sich nun detailliert damit beschäftigt werden, ob sich biologische Maße, d.h. im speziellen BMI, in irgendeiner Weise im Zusammenhang mit dem Cortisolverlauf über den Tag hinweg setzen lassen (Steptoe et al., 2004). Dabei ist aus wissenschaftlicher Sicht immoment noch nicht ganz klar, ob die Fettleibigkeit, d.h. BMI>30, mit der Ausschüttung / Sekretion an Cortisol in Zusammenhang steht (Marin et al., 1992, Marniemi et al., 2002) oder nicht (Andrew et al., 1998, Ljung et al., 2000).

Um dies zu untersuchen, wurde das zuvor beschriebene penalisierte Modell zur Schätzung verwendet. Nach Residuenanalysen zeigte sich, dass die Korrelation approximativ als autoregressiver Prozess (AR(1)) interpretiert werden kann. Natürlich wurde das Modell auch mit Working-Independence getestet und die Resultate waren annähernd gleich. Der Wertebereich von BMI wurde für die Bildung der B-Splines in 15 gleichgroße Intervalle zerlegt. Die Glättung wurde eindimensional durchgeführt, d.h. nur die Dimension 'Glattheit der Kurve' ($= \lambda_\alpha$) wurde mit Hilfe des BIC als Kriterium optimal geglättet. Die beiden anderen Dimensionen konnten wegen des Richtungswechsels nicht penalisiert werden. Abbildung 9.1(a) stellt den generellen Zeit-invarianten Schätzer der Kurve BMI vs. Cortisol dar. Hierbei zeigt sich, dass dicke Personen mit einem BMI von mehr als 30 deutlich niedrigere Cortisolwerte aufzuweisen haben als Personen mit Normalgewicht im Bereich von 22 bis 30. Per-

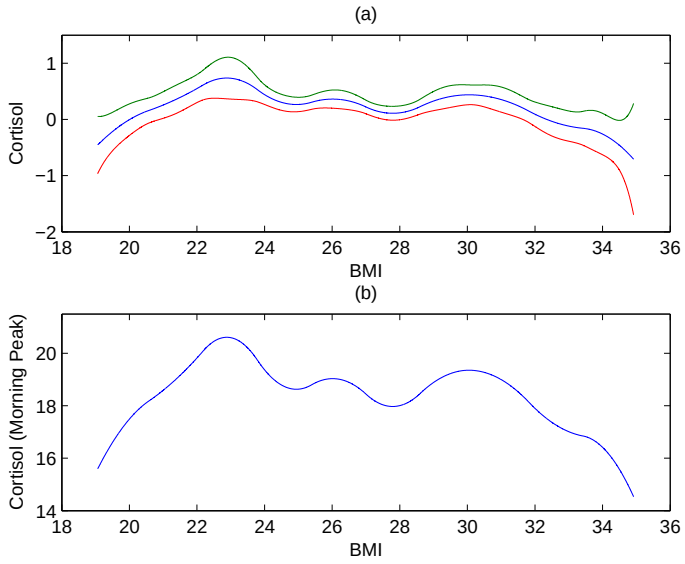

Abbildung 9.1: Cortisol-Schätzungen (Anwendungsbeispiel)

sonen in diesem Bereich haben nahezu konstante Cortisolwerte. Dies bestätigt die Analysen von Marin et al. (1992) und Marniemi et al. (2002). Sehr dünne Personen mit einem BMI von weniger als 22 weisen auch deutlich niedrigere Cortisolwerte auf. Die Parameterschätzungen sind in Tabelle 9.1 zusammengefasst. Die Intercepts β_0 haben den typischen Cortisolverlauf, über den Tag hinweg, mit dem Morning-Peak zum Zeitpunkt $t = 2$ und dem folgenden stetigen Nachlassen während der Zeitpunkte $t = 3, \ldots, 6$. Auch die multiplikativen Effekte sind ähnlich ausgeprägt. Wegen der Normierung auf $\gamma_1 = 1$ werden die anderen Effekte im Verhältnis angegeben. Auch hier erkennt man den Morning-Peak zum Zeitpunkt $t = 2$, der circa um das vierfache höher ausfällt als Zeitpunkt $t = 1$. Zum Zeitpunkt $t = 3$ ergibt sich ein ähnlicher Effekt (≈ 1) wie zum Zeitpunkt $t = 1$. Im weiteren Verlauf des Tages konvergieren die multiplikativen Effekte ($t = 3, \ldots, 6$) langsam gegen Null. Multiplikative Effekte ungefähr von Null werden nachts erwartet. Abbildung 9.1(b) zeigt die Kurve adjustiert für den Intercept und den multiplikativen Effekt zum Zeitpunkt $t = 2$. Auch in dieser Abbildung können die selben Zusammenhänge beobachtet werden. Damit zeigt sich, dass dicke Personen um bis zu 25% niedrigere Cortisolwerte haben als

	Zeit-variierende Effekte			
Zeit	Intercept		Mult. Effekt	
t	β_0	SD	γ	SD
1	11.34	0.45	1.00	–
2	17.51	0.98	4.21	0.44
3	10.34	0.64	1.14	0.34
4	5.17	0.38	0.87	0.26
5	4.61	0.25	0.61	0.17
6	2.65	0.16	0.21	0.09

Tabelle 9.1: Zeit-variierende Parameter-Schätzungen und Standardfehler.

Personen mit normalem BMI im Bereich von 22 bis 30.

Kapitel 10

Semiparametrisches Modell

10.1 Einleitung

In biologischen und medizinischen Analysen ist es oftmals notwendig Modelle auf verschiedene Einflussgrößen hin zu 'kontrollieren'. Dabei werden Zusatzvariablen additiv in das Modell mitaufgenommen, um zu zeigen, dass vorhandene Effekte von Einflussgrößen auf diese zurückzuführen sind und nicht auf die Kontrollvariablen. Typische Kontrollvariablen sind dabei beispielsweise Variablen wie Alter (in stetiger Form), Geschlecht (in binärer Form) und Bluttyp (in kategorieller Form), die in linearer Form bzw. Dummy-Kodierung in das Modell einfließen. Im englischsprachigen Raum werden derartige Variablen als 'Confounder' oder 'Confounding Variables' bezeichnet.

Das nichtparametrische Modell aus Kapitel 7 soll nun einen zusätzlichen parametrischen Term erhalten. Aus der Kombination an parametrischen und nichtparametrischen Schätzmethoden ergibt sich dann ein semiparametrisches Modell mit Varying Coefficients. Mehrere Beispiel für semiparametrische Modelle wurden ausführlich im Theorieteil behandelt. Zusätzlich zum Varying-Coefficients-Term $v_{it}\beta_{0t}$ und dem nichtlinearen Term $\gamma_t\alpha(z_{it})$ (vgl. Kapitel 7) wird nun ein parametrischer Term $x_{it}\beta$

in das ursprüngliche Modell miteingeführt. Damit gilt allgemein für den Prädiktor

$$\eta_{it} = \eta_{it}^{VC} + \eta_{it}^{P} + \eta_{it}^{NP},$$

mit

$$\begin{aligned} \eta_{it}^{VC} &= v_{it}\beta_{0t}, \\ \eta_{it}^{P} &= x_{it}\beta, \\ \eta_{it}^{NP} &= \gamma_t \alpha(z_{it}). \end{aligned}$$

Für die Einflussgrößen x_{it} gilt

$$x_{it} = (x_{it1}, \ldots, x_{itp})', \quad j = 1, \ldots, p.$$

Die restlichen Größen sind analog wie zuvor definiert. Damit lässt sich der Prädiktor des semiparametrischen Modells zusammenfassen zu

$$\eta_{it} = v_{it}\beta_{0t} + x_{it}\beta + \gamma_t \alpha(z_{it}). \tag{10.1}$$

Natürlich sind verschiedene Varianten des Modells denkbar, mit

$$\begin{aligned} \eta_{it} &= v_{it}\beta_{0t} + x_i\beta + \gamma_t \alpha(z_{it}), \\ \eta_{it} &= v_{it}\beta_{0t} + x_{it}\beta + \gamma_t \alpha(z_i), \\ \eta_{it} &= v_{it}\beta_{0t} + x_i\beta + \gamma_t \alpha(z_i). \end{aligned} \tag{10.2}$$

Version (10.2) ist somit eine Erweiterung des Modells (7.1) um einen parametrischen Term, der unabhängig von der Zeit ist. Variationen bezüglich des Varying Coefficients-Terms eröffnen weitere Möglichkeiten zur Modellierung. Auch mehrere additive nichtparametrische Terme können in das Modell mit einbezogen werden, so dass gilt:

$$\eta_{it} = v_{it}\beta_{0t} + x_{it}\beta + \sum_{k=1}^{m} \gamma_{kt}\alpha_k(z_{it}). \tag{10.3}$$

Alle bisherigen Vorschläge und auch der nichtparametrische Ansatz aus (7.1) können daher als Sub-Modelle von (10.3) betrachtet werden. Detailliert besprochen werden

soll im folgenden Modell (10.1). Zur Schätzung des Modells (10.1) wird wieder von denselben Voraussetzungen wie zuvor ausgegangen. B-Spline Basis-Funktionen mit Penalisierung in Form von Differenzenpenalties werden zur Schätzung des nichtparametrischen Terms eingesetzt. Desweiteren gehen die Korrelationen über die Working-Kovarianz-Matrix in das Modell mit ein. Die 'Method of Moments' findet Anwendung bei der Bestimmung der Korrelation.

10.2 Semiparametrisches Modell mit Zeit-variierenden multiplikativen Effekten

Die Beobachtungen der abhängigen Variablen seien durch $y_i = (y_{i1}, \ldots, y_{iT})'$, $i = 1, \ldots, n$, $t = 1, \ldots, T$ gegeben. Es wird angenommen, dass die Beobachtungen y_{it} über die Beziehung $y_{it} = \eta_{it} + \varepsilon_{it}$ mit normalverteilten Fehlern ε_{it} und

$$\eta_{it} = v_{it}\beta_{0t} + x_{it}\beta + \gamma_t \alpha(z_{it}) \tag{10.4}$$

in das Modell eingehen. Die unbekannte Funktion $\alpha(z)$ wird durch Basisfunktionen $\{\Phi_l^\alpha\}$ approximiert. Die Basisfunktionen gehen über die lineare Beziehung

$$\alpha(z) \;=\; \sum_{l=1}^{r} \alpha_l \Phi_l^\alpha(z)$$

in das Modell ein. Alle Basisfunktionen $\Phi_1^\alpha, \Phi_2^\alpha, \ldots$ sind miteinander über die Knoten $\tau_1^\alpha < \tau_2^\alpha < \ldots$ verbunden. Die Knoten werden für den jeweiligen Wertebereich der Variable, die in Basisfunktionen abgebildet werden soll, bestimmt. Durch geeignetes Umformen kann Gleichung (10.4) in Matrizenschreibweise überführt werden. Es ergibt sich mit $y_i = \eta_i + \varepsilon_i$ und $\eta_i = (\eta_{i1}, \ldots, \eta_{iT})'$ als linearem Prädiktor

$$\eta_i = V_i\beta_0 + X_i\beta + \Gamma Z_i \alpha, \tag{10.5}$$

wobei $V_i = V_i(v_{it})$, $X_i = X_i(x_{it})$ und $Z_i = Z_i(z_{it})$ Designmatrizen sind, die im Appendix genauer beschrieben sind. Die restlichen Matrizen und Vektoren sind definiert durch die multiplikativen Parameter $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_T)$, die Gewichte

der Basisfunktionen $\alpha' = (\alpha_1, \ldots, \alpha_r)$ und die Zeit-variierenden Intercepts $\beta_0' = (\beta_{01}, \ldots, \beta_{0T})$. Im Vergleich zum Modell aus (7.1) ist nun ein parametrischer Term angefügt worden. Um das Modell zu vervollständigen, wird angenommen, dass gilt

$$y_i = \eta_i + \varepsilon_i,$$

mit normalverteilten Fehlern $\varepsilon_{it} \sim N(0, \Sigma)$ und unbekannter, potentiell parameterisierter Kovarianz-Matrix Σ. Um die Identifizierbarkeit zu gewährleisten sind auch im erweiterten Modell nur zwei Restriktionen in das Modell einzubauen. Die erste Restriktion trennt die Varying Coefficients β_{0t} von den Spline-Gewichten α über $\alpha_r = -\sum_{l=1}^{(r-1)} \alpha_l$ und $\alpha' = (\alpha_1, \ldots, \alpha_{r-1})$. Die zweite Restriktion separiert die Parameter γ und α, indem postuliert wird, dass $\gamma_1 = 1$. Beide Restriktion führen unweigerlich zu Veränderungen in den Designmatrizen und Schätzgleichungen. Details dazu befinden sich im Appendix.

10.3 Schätzmethode

Die Schätzung basiert auf der 'Penalized Weighted Least-Squares'-Methode, bei der die Gewichte als Working-Kovarianzen interpretiert werden können. Ausgangsposition zur Schätzung ist die penalisierte Pseudo-Likelihoodfunktion

$$l_p = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \eta_i)' W_i^{-1}(y_i - \eta_i) - \frac{1}{2}\beta_0' K_{\beta_0}\beta_0 - \frac{1}{2}\gamma' K_\gamma \gamma - \frac{1}{2}\alpha' K_\alpha \alpha \rightarrow \max_{\beta_0,\gamma,\alpha},$$

wobei W_i eine Working-Kovarianz-Matrix darstellt und K_{β_0}, K_γ und K_α die jeweiligen Penalty-Matrizen. Die Penaltymatrizen penalisieren 'First Order', bzw. 'Higher Order'-Differenzen (Tutz, 2003) zwischen benachbarten Basisfunktionen. Detailliertere Informationen befinden sich im Appendix. In Vektor-Notation erhält man nun

$$l_p = -\frac{1}{2}(y - \eta)' W^{-1}(y - \eta) - \frac{1}{2}\beta_0' K_{\beta_0}\beta_0 - \frac{1}{2}\gamma' K_\gamma \gamma - \frac{1}{2}\alpha' K_\alpha \alpha,$$

wobei gilt dass

$$\begin{aligned} y &= (y_1', \ldots, y_n')', \\ \eta &= (\eta_1', \ldots, \eta_n')', \\ W &= \mathrm{diag}(W_1, \ldots, W_n). \end{aligned}$$

Fasst man nun alle zu schätzenden Parameter in einem Vektor $\nu' = (\beta_0', \beta', \alpha', \gamma')$ zusammen, kann man die entsprechende Scorefunktion $s_p = \partial l_p / \partial \nu$ ableiten. Die Schätzprozedur läuft auch in diesem Fall in zwei iterierenden Schritten ab. Allerdings bestehen mehr Möglichkeiten als zuvor, in welchem Schätzschritt welche Parameter bestimmt werden. Zuvor konnten entweder β_0 und α oder β_0 und γ gleichzeitig geschätzt werden. Der übrige bleibende Parameter (α oder γ) wurde im Anschluss in zweiten Schritt geschätzt. Nun wird wie folgt vorgegangen: Im ersten Schritt sollen nun zunächst die Gleichungen $(\partial l_p/\partial \beta_0, \partial l_p/\partial \beta, \partial l_p/\partial \alpha) = 0$ gelöst werden, danach die Gleichungen $\partial l_p/\partial \gamma = 0$. Teilt man die Parameter auf diese Weise auf, sind aus algorithmischer Sicht nur wenige Umformungen zur Schätzung des erweiterten Modells notwendig.

Zunächst sei nun der lineare Prädiktor $\eta_i = (V_i, X_i, \Gamma Z_i)\delta_{SP}$ betrachtet, mit $\delta_{SP} = (\beta_0', \beta', \alpha')'$. Damit ist der Vektor $\eta = (\eta_1', \ldots, \eta_n')'$ gegeben durch $\eta = \Phi_1 \delta_{SP}$ mit Designmatrix

$$\Phi_1 = \begin{pmatrix} V_1 & X_1 & \Gamma Z_1 \\ \vdots & \vdots & \vdots \\ V_n & X_n & \Gamma Z_n \end{pmatrix}. \tag{10.6}$$

Die Designmatrix (10.6) besteht aus mehreren zusammengesetzten Matrizen. Im linken Teil befinden sich die Varying-Coefficients-Terme $V_i = I_{T\times T}$, in der Mitte die parametrischen Terme X_i und im rechten Teil die Matrizen mit Basisfunktionen $Z_i' = (z_{i1}, \ldots, z_{iT})$ und $z_{it}' = (\tilde{z}_{it1}, \ldots, \tilde{z}_{itr})$ (vgl. Appendix). Damit ergibt sich das Gleichungssystem

$$\Phi_1' W^{-1}(y - \eta) - K_\delta \delta_{SP} = 0 \tag{10.7}$$

zur Schätzung von δ_{SP}. Außerdem gilt $K_\delta = \text{diag}(K_{\beta_0}, K_\beta, K_\alpha)$ mit $K_\beta = 0_{p\times p}$ (Nullmatrix). Die Matrix K_β ist keine Penaltymatrix im eigentlichen Sinne, sie wird eingefügt, um die Dimensionen anzugleichen.

Im folgenden soll Parameter γ_{SP} genauer untersucht werden. Seien z_{it}' die Zeilen der

Designmatrix Z_i und γ_{SP} die Zeit-variierenden Parameter in Vektorenschreibweise. Dann lässt sich der lineare Prädiktor η_i verändern in

$$\eta_i = V_i\beta_0 + X_i\beta + \tilde{Z}_i(\alpha)\gamma_{SP}.$$

Es gilt dabei $\tilde{Z}_i(\alpha) = \text{diag}(z'_{i1}\alpha, \ldots, z'_{iT}\alpha)$. Im Matrixform ergibt sich $\eta = V\beta_0 + X\beta + \Phi_2\gamma$ mit $V = (V_1, \ldots, V_n)'$, $X = (X_1, \ldots, X_n)'$ und $\Phi_2 = (\tilde{Z}'_1(\alpha), \ldots, \tilde{Z}'_n(\alpha))'$. Damit ist die Schätzgleichung äquivalent zu

$$\Phi'_2 W^{-1}(y - V\beta_0 - X\beta - \Phi_2\gamma_{SP}) - K_\gamma\gamma_{SP} = 0. \tag{10.8}$$

Auflösen nach den Parametern δ_{SP} und γ_{SP} führt zu den Schätzern

$$\begin{aligned} \hat{\delta}_{SP} &= (\Phi'_1 W^{-1}\Phi_1 + K_\delta)^{-1}\Phi'_1 W^{-1} y, \\ \hat{\gamma}_{SP} &= (\Phi'_2 W^{-1}\Phi_2 + K_\gamma)^{-1}\Phi'_2 W^{-1}(y - V\hat{\beta}_0 - X\hat{\beta}). \end{aligned} \tag{10.9}$$

Nur ein zusätzlicher Offset $X\beta$ wurde in die zweite Gleichung aus (10.9) integriert. Es sei wiederum angemerkt, dass die Matrizen Φ_1 und Φ_2 keine gewöhnlichen Designmatrizen sind, denn sie sind abhängig von den Parametern γ_{SP} bzw. α, d.h. es gilt

$$\begin{aligned} \Phi_1 &= \Phi_1(\gamma_{SP}), \\ \Phi_2 &= \Phi_2(\alpha). \end{aligned}$$

Falls keine Variation über die Zeit hinweg besteht, d.h. $\gamma_1 = \ldots = \gamma_T = 1$, dann ist Φ_1 unabhängig von γ_{SP} und die Gleichungen aus (10.9) reduzieren sich auf das erste Gleichungssystem. Damit ist $\hat{\delta}_{SP}$ eine eindeutige Lösung. Das semiparametrische Modell kann dann bei geeigneten Restriktionen innerhalb eines Schrittes geschätzt werden.

Algorithmisch gelöst werden beide Gleichungen aus (10.9) alternierend. Zunächst werden Startwerte beispielsweise für γ_{SP} gewählt und im Anschluss solange iterativ geschätzt bis ein Abbruchkriterium unterschritten wurde. Der Algorithmus lässt sich im Kürze wie folgt zusammenfassen.

- Schritt 1: Seien $\hat{\gamma}_{SP}^{(0)'} = (1, \ldots, 1)'$ initielle Startwerte und $W_{(0)} = I$, wobei I die Einheitsmatrix ist.

- Schritt 2:

  1. Berechne

     $$\hat{\delta}_{SP}^{(1)} = (\Phi_1(\hat{\gamma}_{SP}^{(0)})' W_{(0)}^{-1} \Phi_1(\hat{\gamma}_{SP}^{(0)}) + K_\delta)^{-1} \Phi_1(\hat{\gamma}_{SP}^{(0)})' W_{(0)}^{-1} y.$$

  2. Schätze die Working-Kovarianz-Matrix $W_{(1)}$
  3. Mit $\hat{\delta}_{SP}^{(1)'} = (\hat{\beta}_0^{(1)'}, \hat{\beta}^{(1)'}, \hat{\alpha}^{(1)'})$ und $W_{(1)}$ berechne

     $$\hat{\gamma}_{SP}^{(1)} = (\Phi_2(\hat{\alpha}^{(1)})' W_{(1)}^{-1} \Phi_2(\hat{\alpha}^{(1)}) + K_\gamma^{-1} \Phi_2(\hat{\alpha}^{(1)})' W_{(1)}^{-1} (y - V\hat{\beta}_0^{(1)} - X\hat{\beta}^{(1)}).$$

- Schritt 3: Wiederhole Schritt 2 und ersetze nun $\hat{\gamma}_{SP}^{(c)}$ mit $\hat{\gamma}_{SP}^{(c+1)}$ und $W_{(c)}$ mit $W_{(c+1)}$ bis das Abbruchkriterium erreicht ist, d.h.

  $$S = \frac{||\nu_c - \nu_{(c-1)}||^2}{||\nu_{(c-1)}||^2} < S_0.$$

  Parameter c bezeichnet hierbei den aktuellen Iterationsschritt und $\nu = (\beta_0', \beta', \alpha', \gamma')'$.

10.4 Inferenz

Beide Schätzgleichungen (10.7) und (10.8) wurden abgeleitet über eine Pseudo-Likelihood-Funktion l_p, die abhängig ist von den Gewichten W_i. Mit $\nu = (\beta_0', \beta', \alpha', \gamma')'$ als totalem Parameter-Vektor und $s_p(\nu) = \partial l_p(\nu)/\partial\nu$ als Pseudo-Score-Funktion erhält man über die 'First Order'-Taylor-Approximation $0 = s_p(\hat{\nu}) \approx s_p(\nu) + (\partial s_p/\partial\nu')(\nu' - \nu)$ die Approximation $\nu' - \nu \approx (-\partial s_p(\hat{\nu})/\partial\nu')^{-1} s_p(\nu)$. Daher kann die Kovarianz von ν über die 'Sandwich-Kovarianz-Matrix', im englischen auch als Sandwich-Matrix bezeichnet, geschätzt werden (Liang und Zeger, 1986, Zeger und Liang, 1986). Die Schätzgleichung für die Sandwich-Matrix ist äquivalent zu

$$\hat{\text{cov}}(\hat{\nu}) = \hat{F}_p^{-1} \hat{F} \hat{F}_p^{-1}$$

mit

$$\begin{aligned}\hat{F}_p &= \sum_{i=1}^{n}\left(\frac{\partial\eta_i}{\partial\nu}\right)' W_i^{-1}\left(\frac{\partial\eta_i}{\partial\nu}\right) = \\ &= \begin{pmatrix} \hat{\Phi}_1' W^{-1}\hat{\Phi}_1 + K_\delta & \hat{\Phi}_1' W^{-1}\hat{\Phi}_2 \\ \hat{\Phi}_2' W^{-1}\hat{\Phi}_1 & \hat{\Phi}_2' W^{-1}\hat{\Phi}_2 + K_\gamma \end{pmatrix}\end{aligned}$$

und $\hat{F} = \mathrm{cov}(s_p(\nu))$, ausgewertet an der Stelle $\hat{\nu}$. Für $\hat{F}$ erhält man

$$\begin{aligned}\hat{F} &= \sum_{i=1}^{n}\left(\frac{\partial\eta_i}{\partial\nu}\right)' W_i^{-1}\Sigma W_i^{-1}\left(\frac{\partial\eta_i}{\partial\nu}\right) = \\ &= \begin{pmatrix} \hat{\Phi}_1' W^{-1}\Sigma W^{-1}\hat{\Phi}_1 + K_\delta & \hat{\Phi}_1' W^{-1}\Sigma W^{-1}\hat{\Phi}_2 \\ \hat{\Phi}_2' W^{-1}\Sigma W^{-1}\hat{\Phi}_1 & \hat{\Phi}_2' W^{-1}\Sigma W^{-1}\hat{\Phi}_2 + K_\gamma \end{pmatrix},\end{aligned}$$

wobei $\Sigma = \mathrm{cov}(\varepsilon)$. Die Kovarianz-Matrix Σ wird geschätzt durch

$$\hat{\Sigma} = (n - \mathrm{tr}(H))^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)(y_i - \hat{y}_i)'$$

und an die Dimensionen angepasst. Dabei gibt $\mathrm{tr}(H)$ die effektiven Freiheitsgrade an, basierend auf der approximativen Hatmatrix H, die im Appendix hergeleitet wurde (Hastie und Tibshirani, 1990). Die Schätzung der Working-Kovarianz-Matrix erfolgt, wie in Kapitel 7.5 beschrieben.

10.5 Kreuzvalidierung

Zusätzlich zum AIC und BIC wurden zwei weitere Kreuzvalidierungs-Kriterien zur Bewertung der Modelle in die Analysen miteinbezogen. Dies ist zum einen das CV, zum anderen das GCV. Der Unterschied beider Kriterien liegt darin, dass die Devianz jeweils unterschiedlich gewichtet wird, beim CV individuell für jede Beobachtung, beim GCV über die Spur (vgl. Kapitel 7.5).

Die Schwierigkeit, die Kriterien anzuwenden, besteht beim semiparametrischen Modell nun darin, dass das Modell in zwei Schätzgleichungen zerlegt wird, die abwechselnd iterativ geschätzt werden. Da für jeden der beiden Schritte eine separate Hatmatrix berechnet werden kann, stellt sich nun die Frage, welche der beiden Matrizen nun in die Kreuzvalidierung mit einbezogen wird, d.h. es gilt

$$\begin{aligned} \hat{y}_\delta &= H_\delta y, \\ \hat{y}_\gamma &= H_\gamma y. \end{aligned}$$

Wendet man ein Kreuzvalidierungskriterium auf beide Hatmatrizen an, erhält man Werte, die ungefähr auf einem Niveau liegen. Tendenziell fällt das Kriterium von H_δ in den Simulationen etwas größer aus. Im folgenden wurden die Kriterien ausschließlich für H_δ berechnet, da H_γ über Approximationen zum Teil aus H_δ abgeleitet wurde (vgl. Appendix, semiparametrisches Modell, Hat-Matrizen).

Kapitel 11

Simulations-Studie Semiparametrisches Modell

In der folgenden Simulations-Studie soll das Regressions-Modell

$$y_{it} = \beta_{0t} + x_i\beta - \gamma_t sin(2z_i) + \varepsilon_{it} \tag{11.1}$$

mit $\beta_0 = (\beta_{01}, \ldots, \beta_{0T})' = (0, 0.1, 0.2, \ldots, 0.9)'$ und $\beta = 1.4$ überprüft werden. Wieder gilt $n = 120$ und $T = 10$. Desweiteren werden die z_i gleichverteilt aus dem Bereich $[0, 2\pi]$ und die x_i gleichverteilt aus dem Bereich $[0, 25]$ gezogen. Die Anzahl der Basisfunktionen wird von 15 auf 18 erhöht. Zur Bewertung der Güte des Modells wird der MSE betrachtet, so wie in Kapitel 8. Die Settings werden entsprechend Kapitel 8.1 gewählt. Die Wahl der Glättungsparameter wird über verschiedene Kriterien bestimmt. Zusätzlich zum Bayesian Information Criterion (BIC), werden das Akaike Information Criterion (AIC) und auch Kreuzvalidierungskriterien CV und GCV in die Analysen miteinbezogen. Im multiplikativen Modell (M) waren drei Glättungs-Parameter λ_{β_0}, λ_α und λ_γ im dreidimensionalen Grid-Search zu bearbeiten, während in additiven Modell (A) nur λ_α und λ_{β_0} optimal bestimmt werden mussten.

11.1 Vergleich Additives Modell (A) vs. Multiplikatives Modell (M)

Es folgt ein Vergleich des Additiven Modells (A) mit dem Multiplikativen Modell (M), so wie in Kapitel 8 beschrieben. Der Übersicht wegen, wird nur das Modell ohne Korrelation für niedrige und hohe Varianzen (I1) und (I2) mit Independence-Annahme bei der Working-Kovarianz (IW) begutachtet. Auf weitere Modelle wird im Anschluss detailliert eingegangen. Generell wurden wieder alle möglichen Modelle mit und ohne Glättung simuliert. Allerdings ließen sich die Gütekriterien AIC, BIC, CV und GCV zur Bestimmung des optimal geglätteten Modells nicht mehr in der Form anwenden, wie zuvor, da nun aufgrund des zusätzlichen Varying-Coefficients- und des parametrischen Terms die optimale Glättung bei $\lambda_{\beta_0} = 0$ und $\lambda_\alpha = 0$, bzw. $\lambda_\gamma = 0$ vorzufinden war. Betrachtet man die einzelnen Gütekriterien im Detail, erscheint dies logisch.

	MSE Add. Modell (A), (LV) (I1, IW)	MSE Add. Modell (A), (HV) (I2, IW)
AD	2.39	9.44
M1	56.01	77.61
M2	46.59	53.66
	MSE Mult. Modell (M), (LV) (I1, IW)	**MSE Mult. Modell (M), (HV) (I2, IW)**
AD	2.91	11.43
M1	3.17	13.31
M2	2.99	12.21

Tabelle 11.1: Vergleich: MSE Additives Modell (A) vs. MSE Multiplikatives Modell (M) (Abkürzungen: vgl. Tabelle 8.1 oder Appendix).

Zunächst soll der MSE im direkten Vergleich zwischen additivem und multiplikativem Modell analysiert werden. Dieser ist in Tabelle 11.1 in aller Kürze zusam-

mengefasst. Es zeigt sich wieder, dass in gewissen Situationen, d.h. bei linear ansteigenden und auch bei exponentiell abfallenden Parametern γ, keine optimalen Modellschätzung im Falle des additiven Modells zu erwarten sind. Deutlich bessere Schätzungen können hingegen beim multiplikativen Modell beobachtet werden. Zwar erhöhen sich die MSEs bei konstanten Parametern γ (Konf. AD) leicht, von 2.39 auf 2.91 bei niedriger Varianz und von 9.44 auf 11.43 bei hoher Varianz; dies ist aber aufgrund des zusätzlichen multiplikativen Parameters γ durchaus zu erwarten. Bei den Konfigurationen M1 und M2 reduzieren sich die MSEs im Vergleich beider Modelle stark, sowohl für niedrige als auch hohe Varianzen. Die Abbildungen 11.1 und 11.2 verdeutlichen die Unterschieden nochmals. Direkte Vergleiche zu den MSEs aus Tabelle 8.1 und 8.2 und den Abbildungen 8.1 und 8.3 sind zu vermeiden, da zwangsläufig andere Datensätze verwendet wurden.

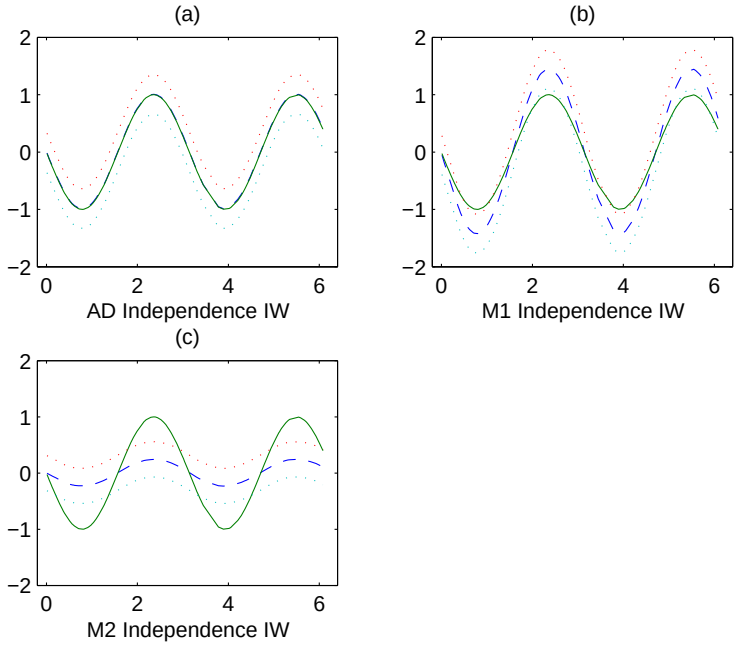

Abbildung 11.1: Schätzungen des Additiven Modells (A), (LV) und (I1, IW) für die Konfigurationen AD, M1 und M2 (vgl. Abbildung 11.2).

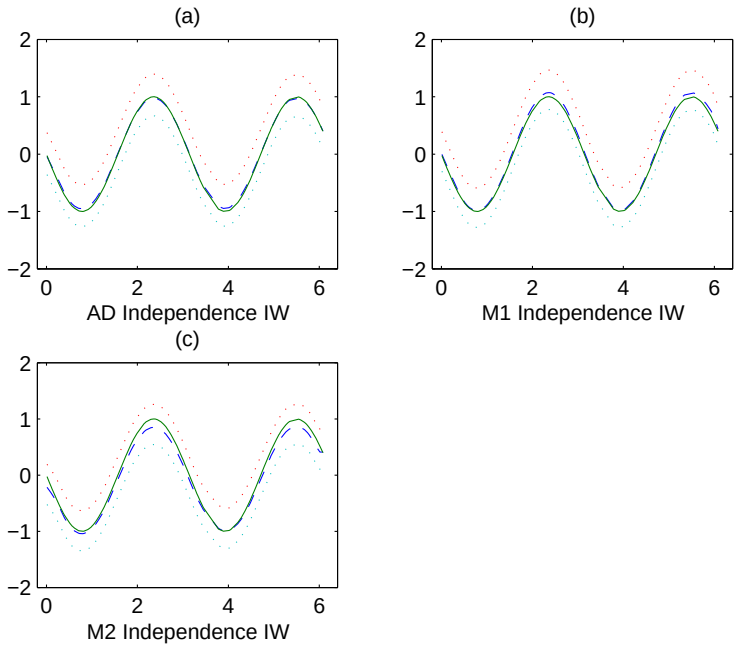

Abbildung 11.2: Schätzungen des Multiplikativen Modells (M), (LV) und (I1, IW) für die Konfigurationen AD, M1 und M2 (vgl. Abbildung 11.1).

11.2 Analysen Multiplikatives Modell (M)

Bisher wurden nur die nichtparametrischen Terme in Form von Abbildungen dargestellt. Der Varying-Coefficients-Term, der parametrische Term und auch die zeitvariierenden Parameter wurden vernachlässigt. Abbildung 11.3 zeigt alle vier Modell-Parameter. Dabei geht Abbildung 11.3(a) auf die Varying-Coefficients ein. Die Schätzung entspricht nahezu der Vorgabe, keine Ausreißer sind festzustellen. In Abbildung 11.3(b) wird der parametrische Term des Modells gezeigt. Die Konfidenzintervalle sind nicht mehr zu erkennen, da sie zu nahe an der Schätzung liegen, d.h. der parametrische Term wird sehr exakt geschätzt und weist daher einen sehr niedrigen Standardfehler von 0.002 auf. Damit liegen die 95%-Konfidenzintervalle um circa 0.004 von der Schätzgeraden entfernt und sind mit bloßem Auge nicht mehr zu erkennen. Auch bei Schätzungen, die hohe MSEs aufweisen, wie beispielsweise im

additiven Modell bei den Konfigurationen M1 und M2, wird der parametrische Term unverzerrt geschätzt, mit einem Standardfehler von 0.0014 im Falle des Modells (M1, I1, IW). Abbildung 11.3(c) geht auf den nichtparametrischen Term ein. Eine leichte Überschätzung des negativen Sinus mit erhöhter Frequenz ist im ungeglätteten Fall zu erkennen. Die Überschätzung wird durch eine leichte Unterschätzung der multiplikativen Parameter γ ausgeglichen (vgl. Abbildung 11.3(d)). Einige Ausreißer (mit '+' gekennzeichnet) sind zu erkennen, allerdings liegen diese vergleichsweise nahe an der Vorgabe. Bei hohen Varianzen können diese (im schlechtesten Fall) deutlich weiter entfernt liegen.

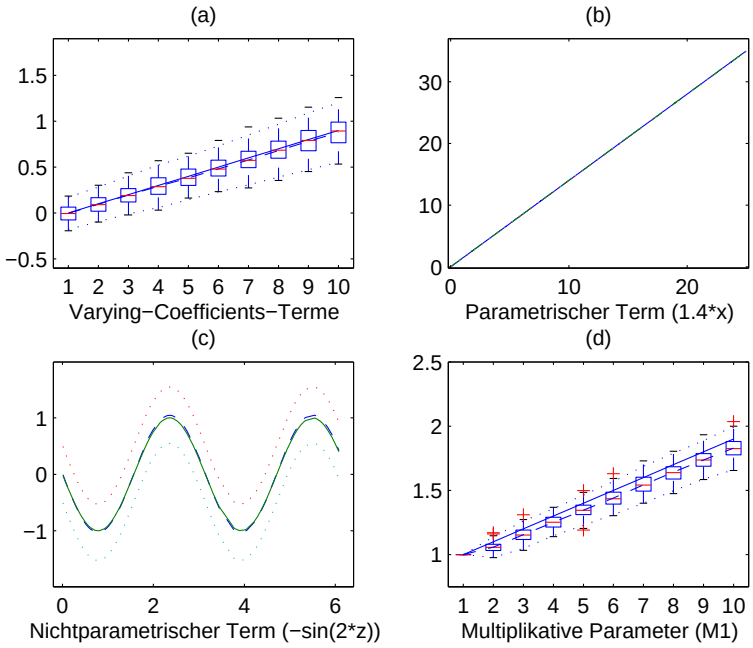

Abbildung 11.3: Parameter-Schätzungen des Multiplikativen Modells (M), (LV) und (AR1, ARW) für die Konfiguration M1, ungeglättet.

Abbildung 11.4 zeigt das autoregressive Modell aus Abbildung 11.3 für hohe Varianzen (AR2). Es ist zu erkennen, dass alle vier Parameter-Schätzungen in etwa mit

denen aus Abbildung 11.3 übereinstimmen, mit dem Unterschied, dass die Konfidenzintervalle sich im Falle des HV-Modells, wie zu erwarten, etwas weiter entfernen und die Anzahl an Ausreißern zunimmt. In Abbildung 11.4(b) sind wieder keine Konfidenzintervalle zu erkennen, da sich der Standardfehler in etwa verdoppelt von 0.002 auf 0.0046 und daher immer noch sehr nahe an der Schätzgeraden liegt. Die leichte Überschätzung der Kurve in Abbildung 11.4(c) und die leichte Unterschätzung der zeit-variierenden Parameter γ in Abbildung 11.4(d) bleiben bestehen.

Geglättete Parameter-Schätzungen sind in Abbildung 11.5 zusammengefasst. Alle drei Glättungsparameter wurden berücksichtigt. Es zeigt sich, dass die Konfidenzintervalle generell enger zusammenrücken und die leichte Unterschätzung der zeit-variierenden Parameter nahezu verschwindet und in eine leichte Überschätzung übergeht. Es entsteht ein leichter Bias bei den Varying Coefficients, keine Reduktion des Bias ist beim nichtparametrischen Term sichtbar. Der Standardfehler des parametrischen Terms sinkt von 0.0046 auf 0.0043.

Im Falle von Compound-Kovarianzstrukturen ergeben sich vergleichbare Resultate. Dies gilt auch für die beiden anderen Konfigurationen AD und M2.

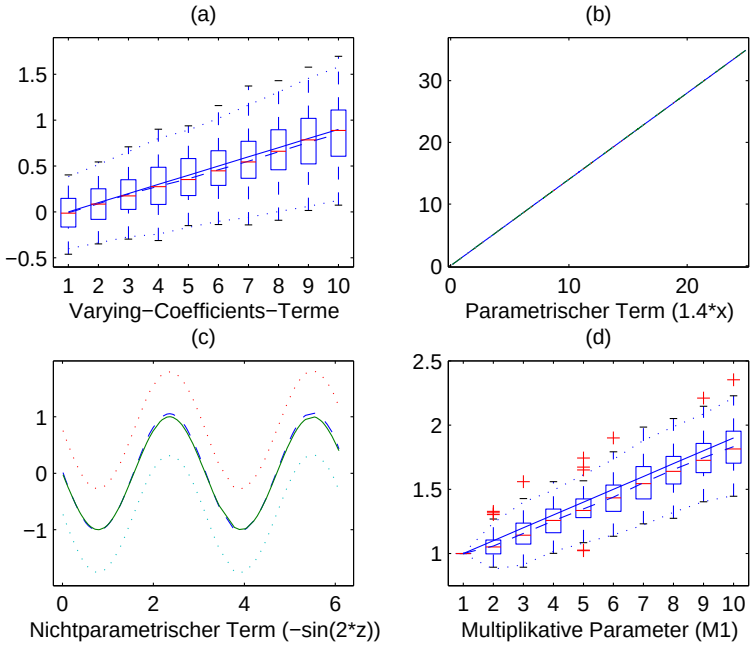

Abbildung 11.4: Parameter-Schätzungen des Multiplikativen Modells (M), (HV) und (AR2, ARW) für die Konfiguration M1, ungeglättet.

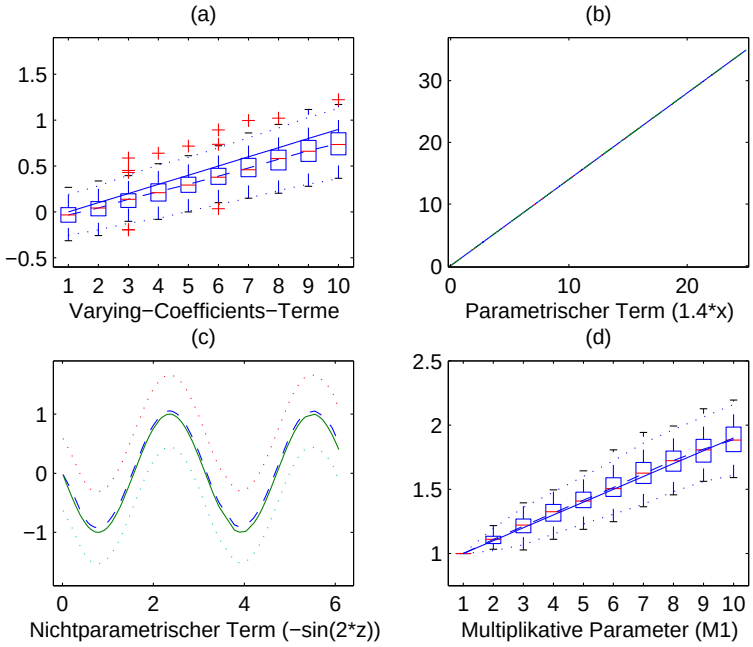

Abbildung 11.5: Parameter-Schätzungen des Multiplikativen Modells (M), (HV) und (AR2, ARW) für die Konfiguration M1, MSE-optimal geglättet.

11.3 Auswirkungen von korrekt spezifizierter Kovarianzstruktur auf den MSE

Erste Versuche und Simulationen mit integrierter Working-Kovarianz-Matrix bei Longitudinaldaten wurden 1986 von Liang und Zeger durchgeführt. Sie zeigten sowohl theoretisch als auch praktisch, dass die Einbeziehung der korrekt spezifizierten Kovarianz-Struktur vor allem bei hohen Korrelationen die Schätzung stark verbessert. Bei niedrigen Korrelationen treten marginale Verbesserungen auf. Im folgenden soll nun ein ähnlicher Vergleich beschrieben werden. Da alle Modelle für niedrige und hohe Varianzen inklusive verschiedener Korrelationsstrukturen und Working-Kovarianz-Strukturen simuliert wurden, sind derartige Vergleiche möglich. Tabelle 11.2 fasst die wichtigsten Ergebnisse zusammen.

	MSE Multiplikatives Modell (M), (LV)					
	(AR1, IW)	(AR1, ARW)	%	(CS1, IW)	(CS1, CSW)	%
AD	5.03	4.74	6.8	10.37	8.86	14.6
M1	5.14	4.96	3.5	9.82	10.64	17.7
M2	4.06	3.86	5.0	6.40	4.47	30.1
	MSE Multiplikatives Modell (M), (HV)					
	(AR2, IW)	(AR2, ARW)	%	(CS2, IW)	(CS2, CSW)	%
AD	24.37	22.56	7.4	25.74	21.38	16.9
M1	23.82	21.79	8.5	30.44	23.47	22.9
M2	19.15	17.54	8.4	20.00	16.75	16.3

Tabelle 11.2: Vergleich: MSE des Modells mit Independence-Working-Kovarianz vs. MSE des Modells mit korrekt spezifizierter Working-Kovarianz-Matrix für AR(1)- und Compound-Struktur, für niedrige und hohe Varianzen mit jeweiligem Verbesserungsgrad, jeweils gleiche Glättungsparameter.

Im oberen Teil von Tabelle 11.2 sind die MSEs für Modelle mit niedriger Varianz zusammengefasst, im unteren Teil Modelle mit hoher Varianz. Es zeigt sich, dass

der MSE bei korrekt spezifizierter Working-Kovarianz, unabhängig vom zugrundeliegenden Modell, stets kleiner ist als bei Independence-Annahme. Vor allem bei Compound-Struktur treten deutliche Verbesserungen von bis zu 30.1% auf. Die Reduktion des MSE und vor allem die Auswirkungen auf die Varianzen soll anhand folgender Abbildung demonstriert werden.

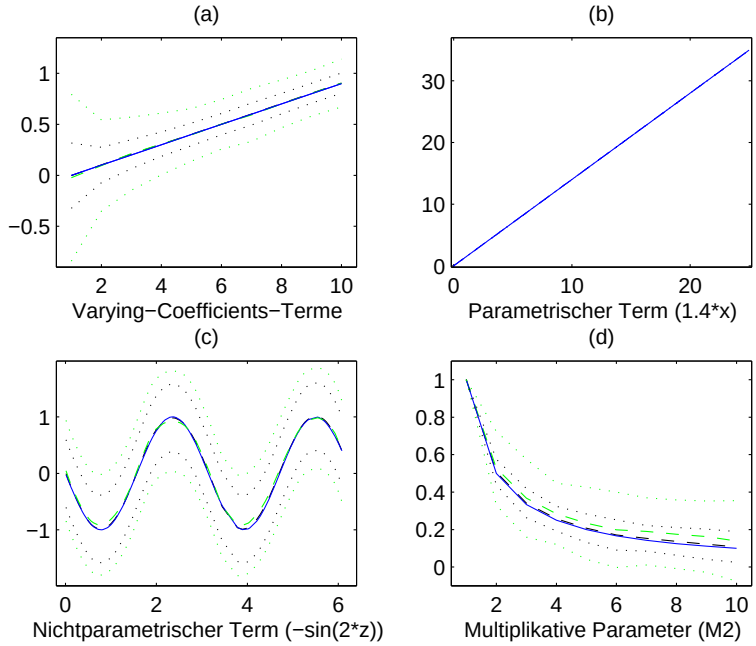

Abbildung 11.6: Vergleich: MSE des Modells mit Independence-Working-Kovarianz (grüne Linien) vs. MSE des Modells mit korrekt spezifizierter Working-Kovarianz (schwarze Linien) basierend auf (CS1), (LV) und (M2) ohne zugehörige Boxplots.

Es ist zumindest auf drei der vier Graphiken aus Abbildung 11.6 klar zu erkennen, dass die Schätzungen bei korrekt spezifizierter Working-Kovarianz-Matrix deutlich an Qualität gewinnen. Hierbei geben die grünen Linien (weiter entfernt liegend) die Schätzungen und Konfidenzintervalle für das Modell mit Independence-Annahme der Working-Kovarianz-Matrix an, die schwarzen Linien (näher liegend) die analo-

gen Schätzungen bei korrekt spezifizierter Working-Kovarianz-Matrix. Die jeweiligen Boxplots wurden der Übersicht halber vernachlässigt. Abbildung 11.6(a) zeigt, dass die Schätzungen der Varying-Coefficients-Terme für (CS1, IW) und (CS1, CSW) nahezu identisch sind, allerdings liegen die Konfidenzintervalle deutlich näher beieinander im Falle korrekt spezifizierter Working-Kovarianz-Matrix. Ähnliche Ergebnisse zeigen sich auch in den drei übrigen Graphiken aus Abbildung 11.6. Abbildung 11.6(b) enthält wie auch in den vorigen Abbildungen zwar Konfidenzintervalle; sie liegen aber wieder zu nahe an der Schätzungen, um die Intervalle selbst und auch Unterschiede zwischen (CS1, IW) und (CS1, CSW) erkennen zu lassen. Offensichtlich ergeben sich aber auch in diesem Fall Unterschiede mit einer Reduktion des Standardfehlers von 0.0047 auf 0.0032. In den übrigen beiden Abbildungen 11.6(c),(d) ist neben den enger liegenden Konfidenzintervallen auch ein reduzierter Bias bei korrekter Spezifikation der Working-Kovarianz-Matrix sichtbar. Sowohl der Sinus als auch die multiplikativen Parameter γ sind im Falle korrekter Spezifikation deutlich präziser geschätzt (schwarze Linien). Die jeweiligen Schätzungen stimmen mit dem vorgegebenen Sinus und den exponentiell fallenden Parametern γ nahezu überein. Abbildung 11.6(c) zeigt eine minimale Unterschätzung des Sinus mit Independence-Annahme in Kombination mit einer Überschätzung der multiplikativen Parameter in Abbildung 11.6(d) (grüne Linien). Insgesamt ergibt sich eine Verbesserung des MSE um 5.0% (vgl. Tabelle (11.2)).

11.4 Abbruchkriterien und die Auswirkungen von Glättung auf den MSE

Es bestehen keinerlei Unterschiede zum nichtparametrischen Modell aus Sektion 8.6 bezüglich des Verlaufs der Abbruchkriterien. Konfiguration (AD) konvergiert sehr schnell, meist innerhalb von 3-6 Iterationen, Konfiguration (M1) langsam und kontinuierlich und Konfiguration (M2) weist in den meisten Fällen den typischen Verlauf auf, so wie in Abbildung 11.7 dargestellt. Glättung führt allgemein zu einer Erhöhung der Anzahl an Iterationen.

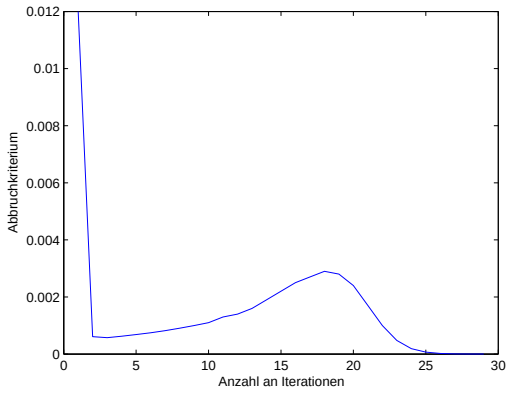

Abbildung 11.7: typischer Verlauf des Abbruchkriteriums für Konfiguration (M2).

Im folgenden soll sich nun auf die Auswirkungen von Glättung auf den MSE konzentriert werden. Tabelle 11.3 zeigt die MSEs verschiedener Modelle, geglättet und ungeglättet und die jeweiligen Verbesserungsgrade. In der oberen Tabelle sind die MSEs für Modelle mit niedrigen Varianzen dargestellt, in der unteren Tabelle die MSEs der Modelle mit hohen Varianzen. Generell fällt auf, dass sich durch Glätten deutliche Verbesserungen ergeben (um bis zu 45.9%). Alle Modelle wurden MSE-optimal geglättet, d.h. dass keines der obengenannten Gütekriterien zur Bewertung der Modelle geeignet war. Alle vier Kriterien beurteilten das ungeglättete Modell als das optimale. Dies steht in Gegensatz zum nichtparametrischen Modell aus Sektion 7, in dem zumindest das Bayesian Information Criterion (BIC) größtenteils von Null verschiedene Glättungsparameter lieferte (vgl. Appendix).

Abbildung 11.8 vergleicht ein geglättetes und ungeglättetes Modell für Konfiguration AR2 mit korrekt spezifizierter ARW-Working-Kovarianz-Matrix. Abbildung 11.8 ist nach dem gleichem Schema aufgebaut wie Abbildung 11.6. Alle Boxplots sind der Übersicht halber entfernt worden. Die schwarzen Linien geben die Schätzungen und die Konfidenzintervalle für das ungeglättete Modell an, die grünen Linien sind analog zu interpretieren für das geglättete Modell. Man erkennt, falls keine Glättung vorliegt, dass die (schwarzen) Konfidenzintervalle in Abbildung 11.8(a),(c) deutlich

	MSE Multiplikatives Modell (M), (LV)								
	(AR1, ARW)			(CS1, CSW)			(I1, IW)		
	NG	G	%	NG	G	%	NG	G	%
AD	6.16	4.75	22.9	13.77	8.86	35.7	3.95	2.91	26.3
M1	6.30	4.96	21.3	12.36	8.76	29.1	4.02	3.17	21.1
M2	5.08	3.86	24.0	5.69	4.47	21.4	5.34	2.99	44.0
	MSE Multiplikatives Modell (M), (HV)								
	(AR2, ARW)			(CS2, CSW)			(I2, IW)		
	NG	G	%	NG	G	%	NG	G	%
AD	28.94	22.56	22.0	39.52	21.38	45.9	20.02	11.43	42.9
M1	31.26	21.79	30.3	38.50	23.47	39.0	20.07	13.31	33.7
M2	26.81	17.54	34.6	27.07	16.75	38.1	20.95	12.21	41.7

Tabelle 11.3: Vergleich: MSE von nicht geglätteten Modellen vs. MSE von geglätteten Modellen mit jeweiligem Verbesserungsgrad für niedrige und hohe Varianzen (G=geglättet, NG=nicht geglättet).

weiter entfernt liegen als die analogen (grünen) Intervalle mit Glättung. Allerdings sind die ungeglätteten Schätzer sind deutlich präziser. Im geglätteten Fall weist die Schätzung einen beträchtlichen Bias auf (vgl. Abbildungen 11.8(a),(b),(d)), während die ungeglätteten Schätzungen im Schnitt sehr nahe am Optimalwert liegen ('Bias-Variance-Tradeoff'). Die Unterschätzung des Sinus in Abbildung 11.8(c) hat wie gewöhnlich eine leichte Überschätzung der Zeit-variierenden Parameter γ zur Folge (vgl. Abbildung 11.8(d)). Im parametrischen Term in Abbildung 11.8(b) reduziert sich der Standardfehler in geringem Maße von 0.0043 auf 0.0041. Die Reduktion ist allgemein auf die Glättung zurückzuführen, da der parametrische Term nicht geglättet werden kann. Bei der Analyse anderer Konfiguration und Modelle zeigen sich analoge Ergebnisse, für kleine Varianzen reduzieren sich die Abstände der Konfidenzintervalle und auch der Bias.

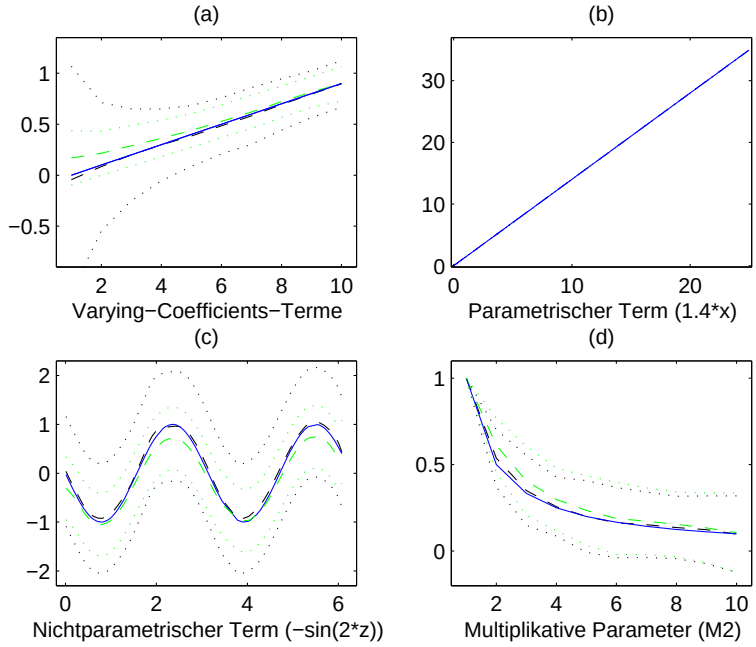

Abbildung 11.8: Vergleich: nicht geglättetes Modell (schwarz) vs. geglättetes Modell (grün) (AR2, ARW, HV).

Kapitel 12

Seemingly-Unrelated-Regression-Modell (SUR)

12.1 Einleitung

Folgende Studie entstand aus der Idee eine Kurve durch eine dreidimensionale Punktewolke zu legen (siehe Abbildung 12.2) und das nichtparametrische Modell aus Kapitel 7 dahingehend auszuweiten. Zunächst war angedacht, zwei unabhängige Modelle $y_1 = f(x)$ und $y_2 = g(x)$ zu fitten. Falls beide Responses y_1 und y_2 unabhängig und unkorreliert wären, würde man zwei parallele unabhängige Modelle fitten, die gemeinsam eine dreidimensionale Kurve ergäben. Dies wäre keine Erweiterung des beschriebenen nichtparametrischen Modells, außer dass nun dreidimensional, d.h. explizit '2-mal 2-dimensional = 3-dimensional' geschätzt und geglättet würde. Bei unabhängigen Responses ergäben sich also zwei unabhängige Schätzungen. Falls die Responses y_1 und y_2 korreliert sind, muss dies bei der Modellschätzung und -spezifizierung berücksichtigt werden. Parallele Regressionsmodelle mit Abhängigkeiten deuten stark auf 'Seemingly Unrelated Regression'-Modelle (SUR) hin. Diese sind bei Zellner (1962, 1963) detailliert beschrieben. Als Motivation für folgende Ausführungen seien nun Schätzungen eines Modells mit Compound-

korrelierten Fehlern gezeigt. Es gelte

$$\begin{aligned}
&y_1 = f(x) = sin(x) \text{ mit } \sigma_1^2 = 0.4, \sigma_{comp,1}^2 = 0.1,\\
&y_2 = g(x) = -sin(2x) \text{ mit } \sigma_2^2 = 0.4, \sigma_{comp,2}^2 = 0.1,\\
&Corr(y_{it1}, y_{it2}) = 0.35,\\
&Corr(y_{it1}, y_{is2}) = 0, \quad t \neq s.
\end{aligned}$$

Abbildung 12.1 zeigt zunächst beide zweidimensionalen Schätzungen für 100 Simulationen mit den jeweiligen Konfidenzintervallen. Beide Schätzungen sind ungeglättet, verlaufen aber trotzdem entlang der vorgegebenen Kurven. Die dreidimensionale Schätzung ist in Abbildung 12.2 geplottet. Auch in diesem Fall ist die Schätzung qualitativ hochwertig, obwohl vergleichsweise hohe Varianzen und Korrelationen in das Modell eingehen.

Ein ähnliches Modell wurde in einem bisher unveröffentlichten Manuskript von Carroll (2003) zur 'Seemingly-Unrelated-Regression' betrachtet. Carroll erweitert das von Zellner (1962) beschriebene SUR-Modell um eine nichtlineare Komponente, so dass ein semiparametrisches Modell entsteht. Es wird dabei allerdings nicht von Daten in longitudinaler Form ausgegangen. Allgemein geht Zellner (1962) davon aus, dass zwei oder mehrere abhängige Variablen in einem Modell zusammengefasst werden und über die Fehler, die in Form einer Working-Kovarianz-Matrix in das Modell eingehen, korrelieren. Dabei bleibt die unabhängige Variable immer gleich, so dass mehrere parallele Regressionen durchgeführt werden. Zusätzlich bleiben bei Longitudinaldaten natürlich die Korrelationen unter den Messwiederholungen bestehen. Damit ergibt sich für das neue Modell mit dem Response $y_{its} = \eta_{its} + \varepsilon_{its}$ der Prädiktor

$$\begin{aligned}
\eta_{it1} &= \eta_{it1}^{VC} + \eta_{it1}^{P} + \eta_{it1}^{NP} + \varepsilon_{it1}\\
\eta_{it2} &= \eta_{it2}^{VC} + \eta_{it2}^{P} + \eta_{it2}^{NP} + \varepsilon_{it2}\\
&\vdots\\
\eta_{itd} &= \eta_{itd}^{VC} + \eta_{itd}^{P} + \eta_{itd}^{NP} + \varepsilon_{itd}, \quad s = 1, \ldots, d
\end{aligned}$$

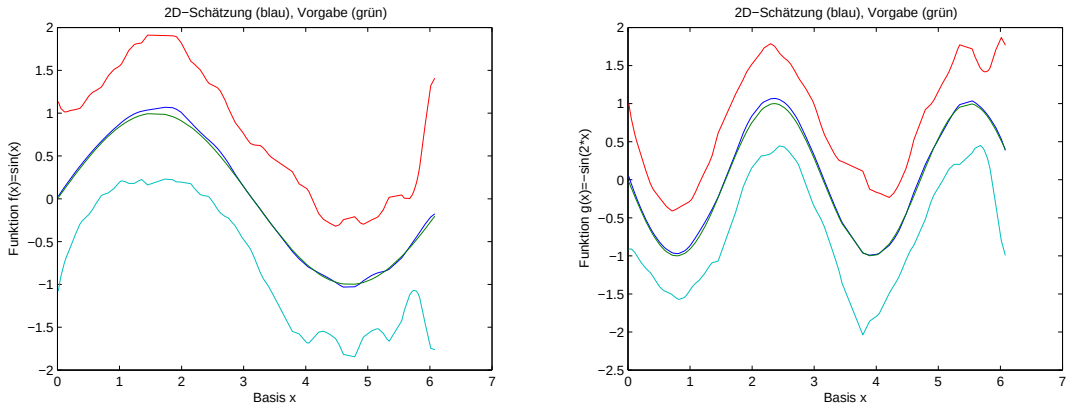

Abbildung 12.1: ungeglättete 2-dimensionale Plots $f(x) = sin(x)$ und $g(x) = -sin(2x)$ mit Konfidenzintervallen und gemeinsamer Basis x für Modell y_1 mit $\sigma_1^2 = 0.4$, $\sigma_{comp,1}^2 = 0.1$ und Modell y_2 mit $\sigma_2^2 = 0.4$, $\sigma_{comp,2}^2 = 0.1$ und $Corr(y_{it1}, y_{it2}) = 0.35$, $Corr(y_{it1}, y_{is2}) = 0$, $t \neq s$.

mit

$$\begin{aligned} \eta_{its}^{VC} &= v_{its}\beta_{0ts}, \\ \eta_{its}^{P} &= x_{it}\beta_s, \\ \eta_{its}^{NP} &= \gamma_{ts}\alpha_s(z_{it}), \quad s = 1, \ldots, d. \end{aligned}$$

Der Einfachheit halber soll wie auch bei Carroll (2003) im folgenden nur auf zwei Responses eingegangen werden, allerdings ist es problemlos möglich, die Theorie auf $d > 2$ Responses auszuweiten. Somit gilt für das Modell in genereller Form

$$\begin{aligned} \eta_{it1} &= v_{it1}\beta_{0t1} + x_{it}\beta_1 + \sum_{k=1}^{m} \gamma_{kt1}\alpha_{k1}(z_{it}) + \varepsilon_{it1}, \\ \eta_{it2} &= v_{it2}\beta_{0t2} + x_{it}\beta_2 + \sum_{k=1}^{m} \gamma_{kt2}\alpha_{k2}(z_{it}) + \varepsilon_{it2}, \end{aligned}$$

mit den Varying-Coefficients-Termen $v_{its}\beta_{0ts}$, den parametrischen Termen $x_{its}\beta_s$ und den nichtparametrischen Termen mit Zeit-variierenden Effekten in Summenform $\sum_{k=1}^{m} \gamma_{kts}\alpha_{ks}(z_{it})$. Davon können zahlreiche Submodelle abgeleitet werden. Setzt

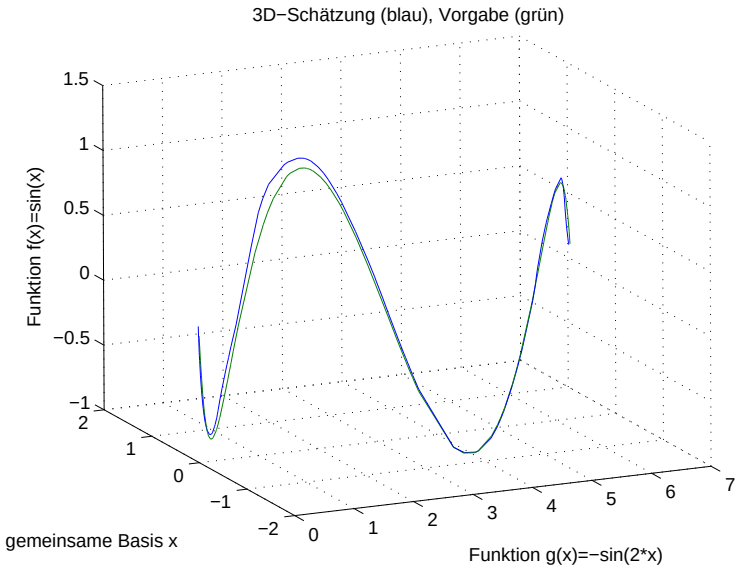

Abbildung 12.2: ungeglätteter 3-dimensionaler Plot für f(x) und g(x).

man $v_{its} = 0$, $m = 1$, $T = 1$ und $\gamma_{kts} = 1$ erhält man das von Carroll (2003) beschriebene Modell

$$\begin{aligned} y_{i1} &= x_{i1}\beta_1 + \alpha_1(z_i) + \varepsilon_{i1}, \\ y_{i2} &= x_{i2}\beta_2 + \alpha_2(z_i) + \varepsilon_{i2}. \end{aligned} \tag{12.1}$$

Führt man den Zeitindex $t = 1, \ldots, T$ bei (12.1) wieder ein, erhält man das Modell in longitudinaler Form.

$$\begin{aligned} y_{it1} &= x_{it1}\beta_1 + \alpha_1(z_{it}) + \varepsilon_{it1}, \\ y_{it2} &= x_{it2}\beta_2 + \alpha_2(z_{it}) + \varepsilon_{it2}. \end{aligned} \tag{12.2}$$

Nach Einführung des Zeit-variierenden Parameters γ und der Varying Coefficients $v_{its}\beta_{0ts}$ erhält man das Modell aus Kapitel 7 in SUR-Form. Es gilt

$$\begin{aligned} y_{it1} &= v_{it1}\beta_{01} + \gamma_{t1}\alpha_1(z_i) + \varepsilon_{it1}, \\ y_{it2} &= v_{it2}\beta_{02} + \gamma_{t2}\alpha_2(z_i) + \varepsilon_{it2}. \end{aligned} \tag{12.3}$$

Modell (12.3) soll im folgenden analysiert und auch simuliert werden. Es werden wieder dieselben Schätzmethoden angewandt, wie in Kapitel 3 und 7 beschrieben.

12.2 Nichtparametrisches SUR-Modell mit Varying-Coefficients für Longitudinaldaten

Der Unterschied zum von Zellner (1962) entwickelten Verfahren besteht darin, dass von nun an in der Kovarianz-Matrix auch Nebendiagonal-Elemente im Blockdiagonal-Bereich verschieden von Null zulässig sind. Zellner (1962) gibt folgende Struktur für ein Modell mit zwei Responses für $d = 2$, $T = 3$ (vgl. Kapitel 6) vor.

$$\mathrm{Cov}_{SUR} = \left(\begin{array}{ccc|ccc} \sigma_{11} & 0 & 0 & \sigma_{12} & 0 & 0 \\ 0 & \sigma_{11} & 0 & 0 & \sigma_{12} & 0 \\ 0 & 0 & \sigma_{11} & 0 & 0 & \sigma_{12} \\ \hline \sigma_{21} & 0 & 0 & \sigma_{22} & 0 & 0 \\ 0 & \sigma_{21} & 0 & 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{21} & 0 & 0 & \sigma_{22} \end{array}\right) = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{21} \\ \hline \Sigma_{12} & \Sigma_{22} \end{array}\right) \otimes I = \Sigma \otimes I. \quad (12.4)$$

Auf den Hauptdiagonal-Blöcken sind dabei die 'Wihtin'-Kovarianzen zu finden, die durch die Messwiederholungen entstehen. Dabei wird in Matrix (12.4) von Unabhängigkeit der Messwiederholungen ausgegangen, da nur die Diagonalelemente in den Hauptdiagonal-Blöcken verschieden von Null sind. In die Nebendiagonal-Blöcke werden die Korrelationen zwischen den scheinbar unabhängigen Modellen eingetragen ('Between'). Diese haben im Falle von Matrix (12.4) auch Diagonalgestalt.

Werden nun zusätzlich Korrelationen unter den Messwiederholungen eingeführt, gilt

$$\begin{aligned}\mathrm{Cov}_{SUR} &= \left(\begin{array}{ccc|ccc} \sigma_{1111} & \sigma_{1112} & \sigma_{1113} & \sigma_{12} & 0 & 0 \\ \sigma_{1121} & \sigma_{1122} & \sigma_{1123} & 0 & \sigma_{12} & 0 \\ \sigma_{1131} & \sigma_{1132} & \sigma_{1133} & 0 & 0 & \sigma_{12} \\ \hline \sigma_{21} & 0 & 0 & \sigma_{2211} & \sigma_{2212} & \sigma_{2213} \\ 0 & \sigma_{21} & 0 & \sigma_{2221} & \sigma_{2222} & \sigma_{2223} \\ 0 & 0 & \sigma_{21} & \sigma_{2231} & \sigma_{2232} & \sigma_{2233} \end{array}\right) \\ &= \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{21} \\ \hline \Sigma_{12} & \Sigma_{22} \end{array}\right) \otimes I = \Sigma \otimes I. \end{aligned} \tag{12.5}$$

Falls $\Sigma_{12} = \Sigma_{21} = 0$ werden zwei unabhängige Modelle geschätzt. Als Erweiterung wäre denkbar, dass auch die Nebendiagonal-Elemente von Σ_{12} bzw. Σ_{21} verschieden von Null gewählt werden könnten oder zumindest auch unterschiedliche Varianzen auf der Hauptdiagonalen zulässig sind mit

$$\mathrm{Cov}_{SUR} = \left(\begin{array}{ccc|ccc} \sigma_{1111} & \sigma_{1112} & \sigma_{1113} & \sigma_{121} & 0 & 0 \\ \sigma_{1121} & \sigma_{1122} & \sigma_{1123} & 0 & \sigma_{122} & 0 \\ \sigma_{1131} & \sigma_{1132} & \sigma_{1133} & 0 & 0 & \sigma_{123} \\ \hline \sigma_{211} & 0 & 0 & \sigma_{2211} & \sigma_{2212} & \sigma_{2213} \\ 0 & \sigma_{212} & 0 & \sigma_{2221} & \sigma_{2222} & \sigma_{2223} \\ 0 & 0 & \sigma_{213} & \sigma_{2231} & \sigma_{2232} & \sigma_{2233} \end{array}\right). \tag{12.6}$$

Falls die Elemente der Kovarianz-Matrix unbekannt sind, müssen sie geschätzt werden. Der genaue Ablauf der Schätzung wird in den folgenden Sektionen ausführlich diskutiert.

12.3 Modellspezifikation

Aufgrund des zusätzlichen Responses gelte nun, dass

$$\begin{aligned} y &= (y_1', y_2')', \\ y_1 &= (y_{11}', \ldots, y_{n1}')' \\ y_{i1} &= (y_{i11}, \ldots, y_{iT1})' \\ y_2 &= (y_{12}', \ldots, y_{n2}')' \\ y_{i2} &= (y_{i12}, \ldots, y_{iT2})' \end{aligned}$$

mit $i = 1, \ldots, n$, $t = 1, \ldots, T$, $s = 1, 2$ und $d = 2$. Es wird angenommen, dass die Beobachtungen y_{its} über die Beziehung $y_{its} = \eta_{its} + \varepsilon_{its}$ mit normalverteilten Fehlern ε_{its} und

$$\eta_{its} = v_{its}\beta_{0ts} + \gamma_{ts}\alpha_s(z_i) \tag{12.7}$$

in das Modell eingehen. Im reduzierten Modell entstehen somit zwei Gleichungen

$$\eta_{it1} = v_{it1}\beta_{01} + \gamma_{t1}\alpha_1(z_i), \tag{12.8}$$

$$\eta_{it2} = v_{it2}\beta_{02} + \gamma_{t2}\alpha_2(z_i). \tag{12.9}$$

Der nichtparametrische Term wird wie zuvor über Basisfunktionen Φ mit Knoten τ modelliert. In Matrizenschreibweise lässt sich (12.7) wie folgt zusammenfassen

$$\eta_{is} = V_{is}\beta_{0s} + \Gamma_s Z_i \alpha_s, \tag{12.10}$$

wobei V_{is} und Z_i Designmatrizen sind, die im Appendix näher definiert sind. Es ist dabei auffällig, dass die Matrix Z_i unabhängig von s ist, d.h. Z_i wird parallel auf y_1 und y_2 modelliert. Bei der Modellierung sind sowohl die Korrelationen von y_1 und y_2 als auch die Korrelationen über die verschiedenen Messungen hinweg von Interesse. Für die restlichen Matrizen gilt

$$\begin{aligned} \Gamma_s &= \text{diag}(\gamma_{1s}, \ldots, \gamma_{Ts}), \\ \alpha_s &= (\alpha_{1s}, \ldots, \alpha_{rs})', \\ \beta_{0s} &= (\beta_{01s}, \ldots, \beta_{0Ts})'. \end{aligned}$$

Allgemein gilt

$$y_s = \eta_s + \varepsilon_s$$

mit normalverteilten Fehlern $\varepsilon_{its} \sim N(0, \Sigma)$ und unbekannter, potentiell parametrisierter Kovarianzmatrix $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$. Keine zusätzlichen Restriktionen werden eingeführt, allerdings erhält jedes Regressionsmodell zwei Restriktionen. Die erste Restriktion trennt die Varying Coefficients β_{0ts} von den Spline-Gewichten α_s über $\alpha_{rs} = -\sum_{l=1}^{(r-1)} \alpha_{ls}$ und $\alpha_s' = (\alpha_{1s}, \ldots, \alpha_{r-1,s})$. Die zweite Restriktion separiert die Parameter γ_s und α_s, indem postuliert wird, dass $\gamma_{1s} = 1$. Beide Restriktion führen unweigerlich zu Veränderungen in den Designmatrizen und Schätzgleichungen. Details dazu befinden sich im Appendix.

12.4 Schätzmethode

Bei der Schätzung verändert sich im Vergleich zu beiden Vorgängermodellen wenig. Ausgangsbasis ist wieder eine Pseudo-Likelihoodfunktion. Es gelte

$$\begin{aligned} l_p &= -\frac{1}{2}\sum_{s=1}^{d}\sum_{i=1}^{n}(y_{is} - \eta_{is})' W_{is}^{-1}(y_{is} - \eta_{is}) \\ &- \frac{1}{2}\sum_{s=1}^{d}(\beta_{0s}' K_{\beta_{0s}} \beta_{0s} - \alpha_s' K_{\alpha_s} \alpha_s - \gamma_s' K_{\gamma_s} \gamma_s) \rightarrow \max_{\beta_{0s}, \alpha_s, \gamma_s} \end{aligned} \tag{12.11}$$

im folgenden. Die Working-Kovarianz-Matrix W_{is} wird dabei an das SUR-Modell angepasst und die Penalty-Matrizen $K_{\beta_{0s}}$, K_{γ_s} und K_{α_s} bleiben bis auf den zusätzlichen Index s unverändert. Die Anzahl der Glättungsparameter verdoppelt sich im Vergleich zu beiden Vorgängermodellen, falls $d = 2$. Zusammenfassend erhält man in Vektor-Notation

$$l_p = -\frac{1}{2}(y - \eta)' W^{-1}(y - \eta) - \frac{1}{2}\beta_0' K_{\beta_0} \beta_0 - \frac{1}{2}\alpha' K_\alpha \alpha - \frac{1}{2}\gamma' K_\gamma \gamma.$$

Die Working-Kovarianz-Matrix W unterzieht sich dabei großen Veränderungen. Für W gilt von nun an

$$W_{(2nT\times 2nT)} = \begin{pmatrix} W_{11} & W_{21} \\ W_{12} & W_{22} \end{pmatrix},$$

mit

$$\begin{aligned} (W_{11})_{nT\times nT} &= \text{blockdiag}(W_{111},\ldots,W_{11n}),\ \text{für } s=1, \\ (W_{22})_{nT\times nT} &= \text{blockdiag}(W_{221},\ldots,W_{22n}),\ \text{für } s=2, \\ (W_{12})_{nT\times nT} &= W_{21} = \text{blockdiag}(W_{121},\ldots,W_{12n}). \end{aligned}$$

Die Matrizen W_{11i} bzw. W_{22i} entsprechen dabei den nichtdiagonalen Working-Kovarianz-Matrizen, so wie sie in den Vorgängermodellen definiert wurden, oder Σ_{11} bzw Σ_{22} aus (12.4). W_{12i} bzw. W_{21i} haben Diagonalgestalt. Im homogenen Fall gilt

$$W_{12i} = \text{diag}(\sigma_{12i},\ldots,\sigma_{12i}),$$

im heterogenen Fall gilt

$$W_{12i} = \text{diag}(\sigma_{12i1},\ldots,\sigma_{12iT}).$$

Die Matrix $W_{12i} = W_{21i}$ ist vergleichbar mit $\Sigma_{12} = \Sigma_{21}$ aus (12.5) und (12.6). Als Erweiterung wäre denkbar, dass für $W_{12i} = W_{21i}$ auch Nebendiagonalelemente verschieden von Null zulässig sind. Für den Prädiktor η gilt

$$\begin{aligned} \eta &= (\eta_1', \eta_2')', \\ \eta_s &= (\eta_{1s}', \ldots, \eta_{ns}')', \\ \eta_{is} &= (\eta_{i1s}, \ldots, \eta_{iTs})'. \end{aligned}$$

Fasst man nun alle zu schätzenden Parameter in einem Vektor

$$\nu = (\beta_{01}', \alpha_1', \beta_{02}', \alpha_2', \gamma_1', \gamma_2')'$$

zusammen, kann man die entsprechende Scorefunktion $s_p = \partial l_p/\partial\nu$ ableiten. Die Schätzprozedur läuft auch im Falle des SUR-Modells in zwei iterierenden Schritten

ab. Es ist offensichtlich, dass je mehr Parameter in das Modell eingehen, desto mehr Variationsmöglichkeiten bestehen, wie die einzelnen Parameter auf welchen Schritt der Schätzung verteilt werden. Um die Schätzung möglichst identisch mit den Schätzungen der Vorgängermodelle ablaufen zu lassen, wird zunächst β_{01}, β_{02}, α_1 und α_2 im ersten Schritt geschätzt, im zweiten Schritt γ_1 und γ_2.

Zunächst sei nun der lineare Prädiktor $\eta_{is} = (V_{is}, \Gamma_s Z_i)\delta_s$ betrachtet, mit $\delta_s = (\beta_{0s}', \alpha_s')'$. Damit ist der Vektor $\eta = (\eta_1', \eta_2')'$ gegeben durch $\eta = \Phi_1 \delta$ mit $\delta = (\delta_1', \delta_2')'$ und Designmatrix

$$\Phi_1 = \begin{pmatrix} V_{11} & \Gamma_1 Z_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ V_{n1} & \Gamma_1 Z_n & 0 & 0 \\ 0 & 0 & V_{12} & \Gamma_2 Z_1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & V_{n2} & \Gamma_2 Z_n \end{pmatrix}. \tag{12.12}$$

Die Designmatrix (12.12) hat die typische SUR-Gestalt und enthält zweimal die Matrix aus (7.5) in blockdiagonaler Form für den jeweiligen Response y_1 bzw. y_2. Jeweils links werden pro Block die Varying-Coefficients $V_{is} = I_{T\times T}$, rechts die nichtparametrischen Terme $Z_i' = (z_{i1}, \ldots, z_{iT})$ mit $z_{it}' = (\tilde{z}_{it1}, \ldots, \tilde{z}_{itr})$ (vgl. Appendix) angegeben. Würde man noch einen parametrischen Term hinzufügen, sähe die Designmatrix wie folgt aus (vgl. (10.6))

$$\Phi_1 = \begin{pmatrix} V_{11} & X_{11} & \Gamma_1 Z_1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{n1} & X_{n1} & \Gamma_1 Z_n & 0 & 0 & 0 \\ 0 & 0 & 0 & V_{12} & X_{12} & \Gamma_2 Z_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & V_{n2} & X_{1n} & \Gamma_2 Z_n \end{pmatrix}.$$

Auch δ würde sich dementsprechend verändern, so dass gilt

$$\begin{aligned}\delta &= (\delta_1', \delta_2')',\\ \delta_s' &= (\beta_{0s}', \beta_{Ps}', \alpha_s').\end{aligned}$$

Für beide Modelle ergibt sich folgendes Gleichungssystem

$$\Phi_1' W^{-1}(y - \eta) - K_\delta \delta = 0 \tag{12.13}$$

zur Schätzung von δ mit $K_\delta = \text{diag}(K_{\beta_{01}}, K_{\alpha_1}, K_{\beta_{02}}, K_{\alpha_2})$ im nichtparametrischen Modell und $K_\delta = \text{diag}(K_{\beta_{01}}, K_{\beta_{P1}}, K_{\alpha_1}, K_{\beta_{02}}, K_{\beta_{P2}}, K_{\alpha_2})$ mit $K_{\beta_{Ps}} = 0_{p\times p}$ im semiparametrischen Modell. Die Matrix $K_{\beta_{Ps}}$ ist keine Penaltymatrix im eigentlichen Sinne, sie wird eingefügt, um die Dimensionen anzugleichen.

Im zweiten Schritt werden die übrigen Parameter γ begutachtet. Seien z_{it}' (unabhängig von s!) die Zeilen der Designmatrix Z_i und $\gamma = (\gamma_1', \gamma_2')'$ der Vektor der multiplikativen Parameter. Dann lässt sich der lineare Prädiktor η_{is} verändern in

$$\eta_{is} = V_{is}\beta_{0s} + \tilde{Z}_i(\alpha_s)\gamma_s.$$

und

$$\begin{aligned}\eta_s &= (\eta_{is}', \ldots, \eta_{ns}')',\\ \eta &= (\eta_1', \eta_2')'.\end{aligned}$$

Dabei gilt $\tilde{Z}_i(\alpha_s) = \text{diag}(z_{i1}'\alpha_s, \ldots, z_{iT}'\alpha_s)$. Im Matrix-Schreibweise ergibt sich

$$\eta = V\beta_0 + \Phi_2\gamma$$

mit $V = (V_1', V_2')'$, $V_s = (V_{1s}, \ldots, V_{ns})'$, $\beta_0 = (\beta_{01}', \beta_{02}')'$ und $\Phi_2 = (\tilde{Z}_i(\alpha_1)', \tilde{Z}_i(\alpha_2)')'$. Damit ist die Schätzgleichung äquivalent zu

$$\Phi_2' W^{-1}(y - V\beta_0 - \Phi_2\gamma) - K_\gamma\gamma = 0, \tag{12.14}$$

mit $K_\gamma = \text{diag}(K_{\gamma_1}, K_{\gamma_2})$. Auflösen nach δ und γ führt zu

$$\begin{aligned}\hat{\delta} &= (\Phi_1' W^{-1}\Phi_1 + K_\delta)^{-1}\Phi_1' W^{-1} y,\\ \hat{\gamma} &= (\Phi_2' W^{-1}\Phi_2 + K_\gamma)^{-1}\Phi_2' W^{-1}(y - V\hat{\beta}_0).\end{aligned} \tag{12.15}$$

Bei geeigneter Definition von Φ_2 lauten die semiparametrischen Schätzer

$$\begin{aligned}\hat{\delta}_{SP} &= (\Phi_1' W^{-1} \Phi_1 + K_\delta)^{-1} \Phi_1' W^{-1} y, \\ \hat{\gamma}_{SP} &= (\Phi_2' W^{-1} \Phi_2 + K_\gamma)^{-1} \Phi_2' W^{-1} (y - V\hat{\beta}_0 - X\hat{\beta}). \end{aligned} \tag{12.16}$$

Diese entsprechen exakt den Vorgaben aus Kapitel 10.2. Algorithmisch gelöst werden beide Gleichungen aus (12.15) alternierend. Der Algorithmus wurde bereits zweimal in den Kapiteln 7.3 und 10.2 beschrieben.

12.5 Inferenz

Beide Schätzgleichungen (12.13) und (12.14) wurden abgeleitet über eine Pseudo-Likelihood-Funktion l_p, die abhängig ist von den Gewichten W_i. Mit dem Vektor $\nu = (\beta_{01}', \alpha_1', \beta_{02}', \alpha_2', \gamma_1', \gamma_2')'$ als totalem Parameter-Vektor und $s_p(\nu) = \partial l_p(\nu)/\partial \nu$ als Pseudo-Score-Funktion erhält man über die 'First Order'-Taylor-Approximation $0 = s_p(\hat{\nu}) \approx s_p(\nu) + (\partial s_p/\partial \nu')(\nu' - \nu)$ die Approximation $\nu' - \nu \approx (-\partial s_p(\hat{\nu})/\partial \nu')^{-1} s_p(\nu)$. Daher kann die Kovarianz von ν über die 'Sandwich-Kovarianz-Matrix', im englischen auch als Sandwich-Matrix bezeichnet, geschätzt werden (Liang und Zeger, 1986, Zeger und Liang, 1986). Die Schätzgleichung für die Sandwich-Matrix ist äquivalent zu

$$\hat{\mathrm{cov}}(\hat{\nu}) = \hat{F}_p^{-1} \hat{F} \hat{F}_p^{-1}$$

mit

$$\begin{aligned}\hat{F}_p &= \sum_{i=1}^{n} \left(\frac{\partial \eta_i}{\partial \nu}\right)' W_i^{-1} \left(\frac{\partial \eta_i}{\partial \nu}\right) = \\ &= \begin{pmatrix} \hat{\Phi}_1' W^{-1} \hat{\Phi}_1 + K_\delta & \hat{\Phi}_1' W^{-1} \hat{\Phi}_2 \\ \hat{\Phi}_2' W^{-1} \hat{\Phi}_1 & \hat{\Phi}_2' W^{-1} \hat{\Phi}_2 + K_\gamma \end{pmatrix}\end{aligned}$$

und $\hat{F} = \mathrm{cov}(s_p(\nu))$, ausgewertet an der Stelle $\hat{\nu}$. Für $\hat{F}$ erhält man

$$\begin{aligned}\hat{F} &= \sum_{i=1}^{n} \left(\frac{\partial \eta_i}{\partial \nu}\right)' W_i^{-1} \Sigma W_i^{-1} \left(\frac{\partial \eta_i}{\partial \nu}\right) = \\ &= \begin{pmatrix} \hat{\Phi}_1' W^{-1} \Sigma W^{-1} \hat{\Phi}_1 + K_\delta & \hat{\Phi}_1' W^{-1} \Sigma W^{-1} \hat{\Phi}_2 \\ \hat{\Phi}_2' W^{-1} \Sigma W^{-1} \hat{\Phi}_1 & \hat{\Phi}_2' W^{-1} \Sigma W^{-1} \hat{\Phi}_2 + K_\gamma \end{pmatrix},\end{aligned}$$

wobei $\Sigma = \text{cov}(\varepsilon)$. Die Kovarianz-Matrix Σ wird geschätzt durch

$$\hat{\Sigma} = (n - \text{tr}(H))^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)(y_i - \hat{y}_i)'$$

und an die Dimensionen angepasst. Dabei gibt $\text{tr}(H)$ die effektiven Freiheitsgrade an, basierend auf der approximativen Hatmatrix H, die im Appendix hergeleitet wurde (Hastie und Tibshirani, 1990). Die Schätzung der Working-Kovarianz-Matrix erfolgt, wie in Kapitel 7.5 beschrieben. Carroll und Tutz (2003) leiten in ihrem Manuskript das GCV-Kreuzvalidierungskriterium für SUR-Modelle ab. Es weist keine Unterschiede zu dem bisher definierten GCV auf und es gilt

$$GCV_{SUR} = DEV \left(\frac{1}{1 - \frac{tr(H)}{2nT}} \right)^2$$

Auch in diesem Fall stellt sich die Frage, welcher der beiden Schätzschritte als Basis für die Kreuzvalidierung verwendet wird. Generell wird so vorgegangen, wie in Kapitel 7.5 beschrieben.

12.6 Working-Kovarianz-Matrix

Grundvoraussetzung für eine beliebige (Working)-Kovarianz-Matrix ist, dass sie positiv definit ist, d.h. dass entweder die Determinante größer als Null ist oder äquivalent dazu, dass die Eigenwerte alle größer als Null sind, damit Invertierbarkeit gewährleistet ist. Diese Grundvoraussetzung ist in einigen Konstellationen im SUR-Modell nicht erfüllt. Es müssen also beim Simulieren Kovarianz-Matrizen konstruiert werden, die diese Voraussetzungen erfüllen und damit zur Generierung von korrelierten Daten zulässig sind.

Eine wichtige Eigenschaft um die angesprochenen Voraussetzungen zu erfüllen, ist '(Haupt-)Diagonal-Dominanz'. Falls beispielsweise die Nebendiagonalelemente aus (12.4) größer sind als die Elemente auf den Hauptdiagonalen, werden die Eigenwerte zum Teil kleiner Null und verhindern eine Invertierung der Kovarianzmatrix.

Falls gilt $\Sigma_{11} = \Sigma_{12} = \Sigma_{21} = \Sigma_{22}$ sind zwar einige Eigenwerte gleich Null und eine Invertierung ist gerade nicht mehr möglich. Erhöht man nun die Kovarianzen in den Blöcken Σ_{11} und Σ_{22} minimal (vgl. Matrix der Form (12.5)), ergibt sich mindestens ein Eigenwert, der kleiner als Null ist. Damit ist keine Invertierbarkeit mehr gegeben. Je höher bei Matrizen der Form (12.5) die Kovarianzen in den Blöcken Σ_{11} und Σ_{22} gewählt werden, desto niedriger müssen die Kovarianzen in den Blöcken Σ_{21} und Σ_{12} gewählt werden, um Invertierbarkeit zu gewährleisten. Dabei spielt es keine Rolle, ob AR1- oder Compound-Modellannahmen zugrundeliegen. Auch falls sich die Kovarianzen Σ_{11} und Σ_{22} stark unterscheiden, muss auf eine geeignete Wahl der Matrizen Σ_{12} und Σ_{21} geachtet werden. In gewissen Situationen können dabei die Diagonal-Elemente aus Σ_{21} und Σ_{12} größer gewählt werden als die Diagonal-Elemente aus einer der beiden Matrizen Σ_{11} und Σ_{22}.

Die Schätzung der Working-Kovarianz-Matrix erfolgt nach dem vorgegebenen Prinzip von Liang und Zeger (1986). Allerdings bestehen zwei Möglichkeiten zur Schätzung. Entweder wird die komplette Kovarianz-Matrix, so wie in (7.5) geschätzt oder beide blockdiagonalen Matrizen separat. Die komplette Kovarianzmatrix Σ wird, wie schon zuvor, über

$$\hat{\Sigma} = (n - \text{tr}(H))^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)(y_i - \hat{y}_i)'$$

geschätzt mit $y_i = (y'_{i1}, y'_{i2})'$ und die Elemente der Hauptdiagonalen in die Matrizen A_{is} so eingesetzt, dass gilt

$$W_{is} = A_{is}^{\frac{1}{2}} R(\theta) A_{is}^{\frac{1}{2}}.$$

Die Working-Kovarianz für y_{i1} und y_{i2} separat zu schätzen mit

$$\hat{\Sigma}_1 = (n - \text{tr}(H_1))^{-1} \sum_{i=1}^{n} (y_{i1} - \hat{y}_{i1})(y_{i1} - \hat{y}_{i1})'$$

$$\hat{\Sigma}_2 = (n - \text{tr}(H_2))^{-1} \sum_{i=1}^{n} (y_{i2} - \hat{y}_{i2})(y_{i2} - \hat{y}_{i2})'$$

und angepassten Hatmatrizen H_1 und H_2, führt (zumindest in diesem Modell) zu deutlich besseren Schätzungen. Verwenden von Elementen aus $\hat{\Sigma}$ für A_{is} führt zu starkem Unterschätzen der (Ko)-Varianzen, Einsetzen von Elementen aus $\hat{\Sigma}_{i1}$ und $\hat{\Sigma}_{i2}$ führt in den meisten Fällen zur Überschätzung der (Ko)-Varianzen, so wie es teilweise auch im nicht- und semiparametrischen Modell festzustellen war.

Auch die Nebendiagonalblöcke können in Form einer Working-Kovarianz-Zerlegung ausgedrückt werden. Es gelte allgemein

$$U_{id,ie} = (B_i)_{de} R(\zeta) (B_i)_{de}$$

mit $d, e = (1, \ldots, s)$. Zwei Strukturen werden im Folgenden behandelt. Es wird dabei zwischen Working-Independence und Independence unterschieden. Der 'Independence'-Fall geht von völliger Unabhängigkeit zwischen beiden Modellen aus, so dass $R(\zeta) = 0_{T \times T}$ gilt. Bei Working-Independence geht man von $R(\zeta) = I_{T \times T}$ aus. Da in diesem Fall nur zwei Responses vorhanden sind, muss nur ein Nebendiagonalblock geschätzt werden. Auch in diesem Fall wird die Kovarianz über

$$\hat{\Sigma}_{12} = (n - \mathrm{tr}(H_3))^{-1} \sum_{i=1}^{n} (y_{i1} - \hat{y}_{i1})(y_{i2} - \hat{y}_{i2})'$$

geschätzt und die Diagonalelemente entsprechend in die Blöcke $\Sigma_{21} = \Sigma_{12}$ eingefügt. Andere Möglichkeiten zur Schätzung von $\Sigma_{21} = \Sigma_{12}$ sind denkbar. Werden die entsprechenden Elemente aus Σ entnommen, zeigen sich wieder deutliche Unterschätzungen.

Im Haupt-Bereich der Kovarianz-Matrix werden Independence, AR1- und Compound-Strukturen eingesetzt. Eine Mischung aller Kovarianz-Formen in einem Modell wird in den folgenden Simulationen diskutiert. Zusätzlich werden verschiedene Konfigurationen an Parametern γ diskutiert.

Kapitel 13

Simulations-Studie SUR-Modell

In der folgenden Simulations-Studie soll das Regressions-Modell

$$\begin{aligned} y_{it1} &= \beta_{0t1} + \gamma_{t1} sin(z_i) + \varepsilon_{it1} \quad \text{(SUR1)} \\ y_{it2} &= \beta_{0t2} - \gamma_{t2} sin(2z_i) + \varepsilon_{it2} \quad \text{(SUR2)} \end{aligned} \tag{13.1}$$

mit $\beta_{0t1} = \beta_{0t2} = (0, \ldots, 0)'$ überprüft werden. Aufgrund der SUR-Annahme gilt $\text{Cov}(y_1, y_2) \geq 0$. Desweiteren gilt $n = 120$ und $T = 10$. Die Datenpunkte z_i werden gleichverteilt aus dem Bereich $[0, 2\pi]$ gezogen. Die Anzahl der Basisfunktionen wird auf 18 festgelegt. Zur Bewertung der Güte des Modells wird der MSE betrachtet, so wie in Kapitel 8. Die Settings werden entsprechend Kapitel 8.1 gewählt. Die Wahl der Glättungsparameter wird über verschiedene Kriterien bestimmt. Zusätzlich zum Bayesian Information Criterion (BIC), werden das Akaike Information Criterion (AIC) und auch Kreuzvalidierungskriterien CV und GCV in die Analysen miteinbezogen. Nur noch das multiplikative Modell, kein additives Modell wird betrachtet. Darin enthalten sind sechs Glättungs-Parameter $\lambda_{\beta_0 1}$, $\lambda_{\alpha 1}$ und $\lambda_{\gamma 1}$ für SUR1 und $\lambda_{\beta_0 2}$, $\lambda_{\alpha 2}$ und $\lambda_{\gamma 2}$ für SUR2. Um alle Glättungs-Parameter optimal im Sinne der obengenannten Kriterien zu wählen, ist ein 'sechs'-dimensionaler Grid-Search notwendig. Dieser ist computertechnisch kaum zu bewältigen. Deswegen wurden nur wenige der Modelle optimal geglättet und vor allem bei den Parametern $\lambda_{\beta_0 1}$ und $\lambda_{\beta_0 2}$ eher ein grobes Gitter ausgewählt.

13.1 Settings der Simulations-Studie

Zugrundeliegende Kovarianzstruktur $\mathbf{cov}(\varepsilon_i) = \Sigma$

- (CS) Compound-Symmetry: $\Sigma = \sigma_0^2 I + \sigma_{comp}^2 1_{(t \times t)}$
  1. $\sigma_{0,1} = 0.05$, $\sigma_{comp,1}^2 = 0.05$, resultierend in: $\sigma_{0,1} + \sigma_{comp,1}^2 = 0.1$ (CS1)
  2. $\sigma_{0,2} = 0.4$, $\sigma_{comp,2}^2 = 0.1$, resultierend in: $\sigma_{0,2} + \sigma_{comp,2}^2 = 0.5$ (CS2)
- (AR) Autoregressive Korrelation AR(1): $\Sigma = (\sigma_{AR}^2 \rho^k)$
  1. $\sigma_{AR,1} = 0.1$, $\rho = 0.4$ (AR1)
  2. $\sigma_{AR,2} = 0.5$, $\rho = 0.4$ (AR2)

Konfiguration der Parameter γ

- Additives Model: (konstant)
  (AD) $\gamma = (1, 1, \ldots, 1)$
- Multiplikatives Model 1: (linear ansteigend)
  (M1) $\gamma = (1, 1.1, 1.2, \ldots, 1.9)$
- Multiplikatives Model 2: (exponentiell fallend, 'time-vanishing')
  (M2) $\gamma = (1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{10})$

Working-Kovarianz Struktur

- Independence:
  (IW) $W_{ind} = \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}} I \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}}$
- Compound-Symmetry:
  (CSW) $W_{comp} = \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}} R(\hat{\theta}) \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}}$
- Autoregressive Korrelation AR(1):
  (ARW) $W_{AR} = \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}} R(\hat{\theta}) \text{diag}(\hat{\sigma}_{it})^{\frac{1}{2}}$

SUR-Struktur:

$\sigma_{t12} = cov(y_{t1}, y_{t2})$, $\sigma_{12} = diag(\sigma_{112}, \ldots, \sigma_{T12})$, $t = 1, \ldots, T$

- Independence:
  (IV) $U_{ind} = 0_{T \times T}$
- Working-Independence:
  (IWV) $U_{wind} = \hat{\sigma}_{12} = \hat{\sigma}_{12}^{\frac{1}{2}} I_{T \times T} \hat{\sigma}_{12}^{\frac{1}{2}}$
  1. (IWV1) $\hat{\sigma}_{t12} = 0.04$
  2. (IWV2) $\hat{\sigma}_{t12} = 0.25$
  3. (IWV3) $\hat{\sigma}_{t12} = 0.35$

Es ist offensichtlich, dass alle drei Konfigurationen AD, M1, M2 gleichen Varianzen unterliegen. Dadurch wird Vergleichbarkeit zwischen den Konfigurationen garantiert.

13.2 Analysen der SUR-Simulationen

Zunächst soll ein SUR-Modell diskutiert werden, das zwei verschiedene Kovarianzstrukturen enthält. Im Modell SUR1 seien die Fehler in Compoundstruktur (CS), in Modell SUR2 seien die Fehler autoregressiv korreliert (AR). Die Working-Kovarianzen werden nach den jeweiligen Vorgaben geschätzt (CSW, ARW). Zudem wird eine SUR-Korrelation zwischen den Modellen eingeführt . Die SUR-Korrelation hat Diagonalgestalt (Working-Independence, IWV) oder ist nicht vorhanden (Independence, IV). Tabelle 13.1 fasst die MSEs für alle möglichen Parameterkonfigurationen zusammen. Dabei wird das Modell unabhängig (IV) geschätzt, d.h. zwei parallele, separate Modelle werden geschätzt, und zusätzlich wird auch ein Modell mit SUR-Korrelation geschätzt (IWV1). Vergleicht man die Modelle hinsichtlich dieses Gesichtspunktes zeigt sich, dass bei jeweils korrekt spezifizierten Kovarianzmatrizen (CSW und ARW) die MSEs im Modell mit SUR-Korrelationen etwas höher ausfallen. Dies war zu erwarten. Unerwartet niedrig fallen hingegen die MSEs der

	SUR-Modell (LV) SUR1: CS1, CSW SUR2: AR1, ARW		SUR-Modell (HV) SUR1: CS2, CSW SUR2: AR2, ARW	
	(IV)	(IWV1)	(IV)	(IWV2)
AD, AD	19.09	18.57	69.89	73.55
AD, M1	18.46	18.76	65.69	70.17
AD, M2	17.53	17.75	62.05	58.56
M1, AD	17.07	17.31	64.80	70.44
M1, M1	16.03	18.14	64.13	68.17
M1, M2	16.97	18.14	60.53	71.29
M2, AD	10.76	10.84	54.03	60.63
M2, M1	10.67	10.75	53.43	61.07
M2, M2	9.74	9.79	49.78	54.55

Tabelle 13.1: SUR-Modell: Mean Squared Errors für alle möglichen Konfigurationen für niedrige und hohe Varianzen mit zugrundeliegendem Compound-AR-Modell.

Modelle (M2, AD), (M2, M1) und (M2, M2) aus. Dies ist nur dadurch zu erklären, dass die Working-Kovarianz-Matrix im Falle der Konfiguration M2 mit Compoundstruktur sehr genau geschätzt wird und sich damit der MSE deutlich verringert. Diese Aussage lässt sich treffen, da die MSEs unabhängig von der Konfiguration des zweiten Modells SUR2 generell auf einem niedrigeren Niveau liegen. Dies gilt sowohl für hohe als auch niedrige Varianzen. Tendenziell wurden auch in den nicht- und semiparametrischen Simulationen niedrigere M2-MSEs beobachtet, allerdings nicht in diesem Ausmaß.

Aus obengenannten Gründen wurden nur zwei der Modelle aus Tabelle 13.1 geglättet. Die Ergebnisse sind in Tabelle 13.2 zusammengefasst und zeigen bei MSE-optimaler Glättung eine unerwartet hohe Reduktion des MSEs von bis zu 50%. Keines des vier Gütekriterien (AIC, BIC, CV, GCV) verbesserte sich durch Glättung eines der sechs Glättungsparameter. Damit ergibt sich eine optimale Glättung beider

Modelle in Tabelle 13.2 bei $\lambda_{\beta_01} = \lambda_{\alpha 1} = \lambda_{\gamma 1} = \lambda_{\beta_02} = \lambda_{\alpha 2} = \lambda_{\gamma 2} = 0$. Dies war bekanntlich schon zuvor im semiparametrischen Modell zu beobachten und überrascht daher auch nicht. In den Abbildungen 13.1 und 13.2 sind die Parameterschätzungen

	SUR-Modell (LV)		
	SUR1: CS1, CSW		
	SUR2: AR1, ARW		
	ungeglättet	geglättet	%
AD, AD (IWV1)	18.57	9.50	50.3
M2, M1 (IV)	10.67	6.91	35.3

Tabelle 13.2: SUR-Modell: Mean Squared Errors für Modelle mit Glättung inklusive Verbesserungsgrad.

zusammengefasst. Dabei zeigt Abbildung 13.1 die ungeglättete Version, während Abbildung 13.2 die glatten Schätzungen darstellt. Um Vergleichbarkeit zwischen den Schätzungen zu gewährleisten, wurden alle Plots aus den Abbildungen 13.1 und 13.2 bezüglich der Achsenskalierung normiert.

Die Ergebnisse sind im SUR-Modell ähnlich zu interpretieren wie im nicht- und semiparametrischen Ansatz. In der ungeglätteten Version 13.1 ergeben sich sehr exakte Schätzungen. Sowohl die nichtparametrischen Schätzungen als auch die Schätzungen der multiplikativen Parameter liegen sehr nahe an deren Vorgaben. Die Konfidenzintervalle liegen im Vergleich zu Abbildung 13.2 deutlich weiter entfernt und sind zum Teil sehr unruhig. An den Rändern erkennt man im ungeglätteten Modell eine leichte Verbreiterung der Konfidenzintervalle. Diese Verbreiterung ist im geglätteten Modell kaum mehr sichtbar. Die Schätzungen der multiplikativen Parameter können mit Ausnahme der wenigen Ausreißern als sehr gut bezeichnet werden. Nach Einführung der Glättungen reduziert sich die Anzahl der Ausreißer und auch die Konfidenzintervalle rücken sowohl bei den multiplikativen Parametern als auch bei der nichtparametrischen Schätzung deutlich näher. Allerdings führt die Glättung wieder zu einem erhöhten Bias, der sich in einer leichten, kaum sichtbaren Unterschätzung

beider sinusförmigen Kurven ausdrückt. Diese bedingt die leichte Überschätzung der multiplikativen Parameter. Die Überschätzung ist vor allem bei AD-Modellen stark ausgeprägt, bei M1- und M2-Modellen fallen die Unterschätzungen tendenziell kleiner aus.

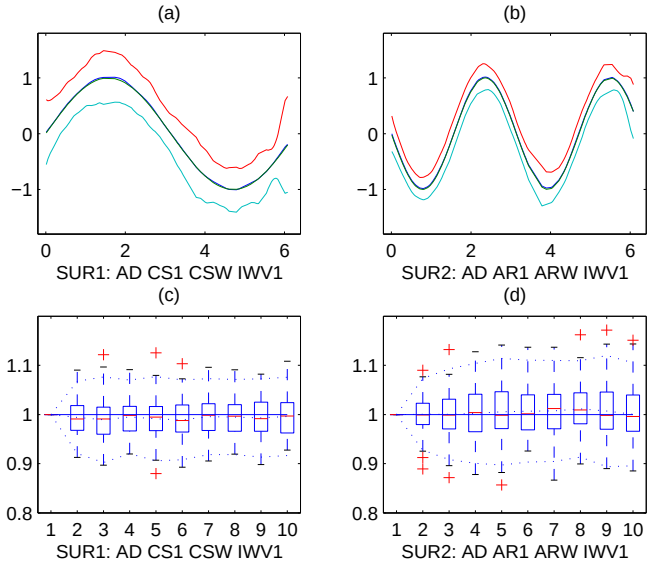

Abbildung 13.1: Nichtparametrische Schätzungen (a), (b) und Schätzungen der multiplikativen Parameter γ (c), (d) des SUR-Modells mit SUR1 in Compound-Struktur (CS1, CSW) und SUR2 in AR-Struktur (AR1, ARW) und SUR-Kovarianzstruktur (IVW1) für die Konfigurationen AD, AD, ungeglättet (MSE=18.57).

13.2.1 Vergleich zweier Modelle mit verschiedenen SUR-Strukturen – (IV) vs. (IWV)

Im folgenden soll sich mit der Frage beschäftigt werden, wie sich die Einführung einer SUR-Kovarianzstruktur im Vergleich zur parallelen Schätzung unabhängiger Modelle auf den 'Model-Fit' auswirkt. Erste Informationen dazu wurden schon in Tabelle 13.1 für MSEs präsentiert. Es zeigt sich bis auf eine Ausnahme, dass die Einführung von SUR-Kovarianzen zu einer Erhöhung des MSE führt. Bei niedrigen

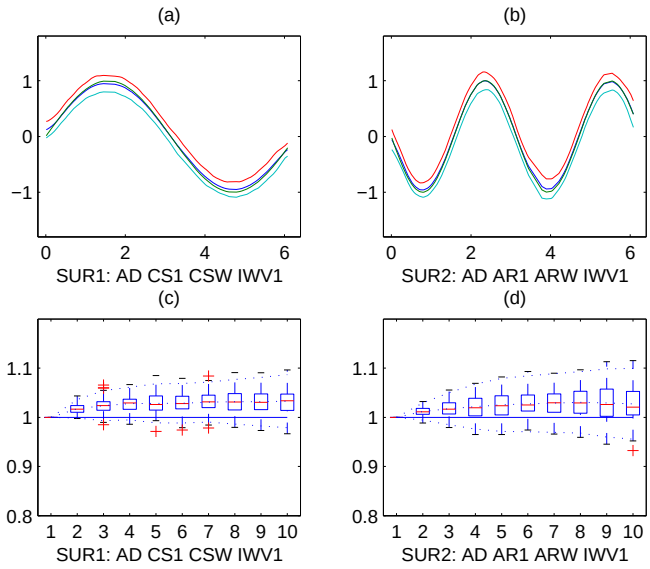

Abbildung 13.2: Nichtparametrische Schätzungen (a), (b) und Schätzungen der multiplikativen Parameter γ (c), (d) des SUR-Modells mit SUR1 in Compound-Struktur (CS1, CSW) und SUR2 in AR-Struktur (AR1, ARW) und SUR-Kovarianzstruktur (IVW1) für die Konfigurationen AD, AD, geglättet (MSE=9.50).

Varianzen (CS1, AR1) und niedrigen SUR-Kovarianzen (IWV1) treten nur marginale Unterschiede auf, bei hohen Varianzen (CS2, AR2) und SUR-Kovarianzen (IWV2) werden die Unterschiede deutlicher. Die Abbildungen 13.3 und 13.4 sollen exemplarisch die Unterschiede verdeutlichen. Wieder ist die Achsenskalierung normiert, um die Unterschiede sichtbar zu machen. Dies lässt die Konfidenzintervalle in Abbildung 13.3 klein erscheinen.

Abbildung 13.3 zeigt die ungeglätteten Schätzungen für das Compound-AR-Modell mit hohen Varianzen (CS2, AR2) mit Independence-Annahme der SUR-Struktur (IV). Es fallen auch in dieser Abbildung 13.3(a), (b) wieder die wackeligen, teilweise rauhen Konfidenzintervalle auf, die im Vergleich mit Abbildung 13.4(a), (b) trotzdem sehr glatt erscheinen. Die Schätzungen können insgesamt als sehr gut bezeichnet

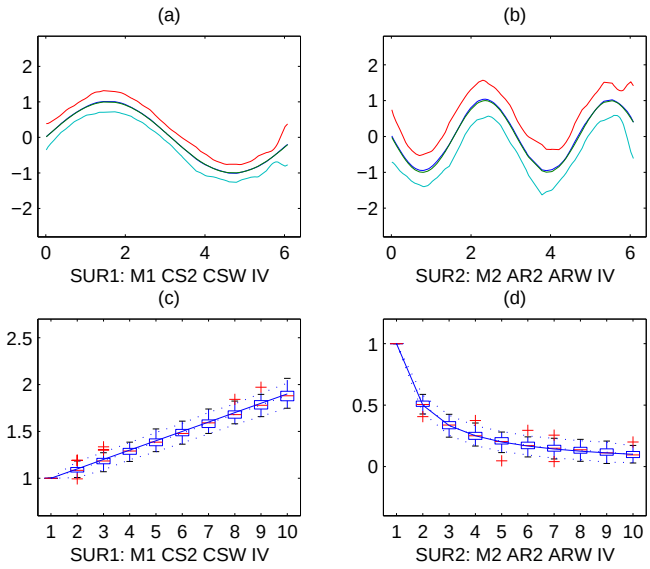

Abbildung 13.3: Nichtparametrische Schätzungen (a), (b) und Schätzungen der multiplikativen Parameter γ (c), (d) des SUR-Modells mit SUR1 in Compound-Struktur (CS2, CSW) und SUR2 in AR-Struktur (AR2, ARW) und SUR-Kovarianzstruktur (IV) für die Konfigurationen M1, M2, ungeglättet (MSE=60.53).

werden. Minimale Unterschiede bestehen zwischen Vorgabe und Schätzung. Partiell können im nichtparametrischen Bereich leichte, kaum sichtbare Unterschätzungen auftreten, speziell an den Maxima und Minima in Abbildung 13.3(b).

In Abbildung 13.4 ist das entsprechende Modell mit Working-Independence-Annahme der SUR-Struktur (IWV2) dargestellt. Die restlichen Parameter und Einstellungen wurden nicht verändert, um Vergleichbarkeit zu gewährleisten. Es zeigt sich, dass sich durch Einführung der SUR-Korrelation die Konfidenzintervalle aller Parameterschätzungen deutlich ausweiten. Vor allem im nichtparametrischen Bereich präsentieren sich die Konfidenzintervalle sehr rauh und unruhig. Auch an den Rändern entstehen sehr große Abweichungen, speziell in Abbildung 13.4(b). Generell können die gemittelten Schätzungen jedoch als sehr gut bezeichnet werden, da die

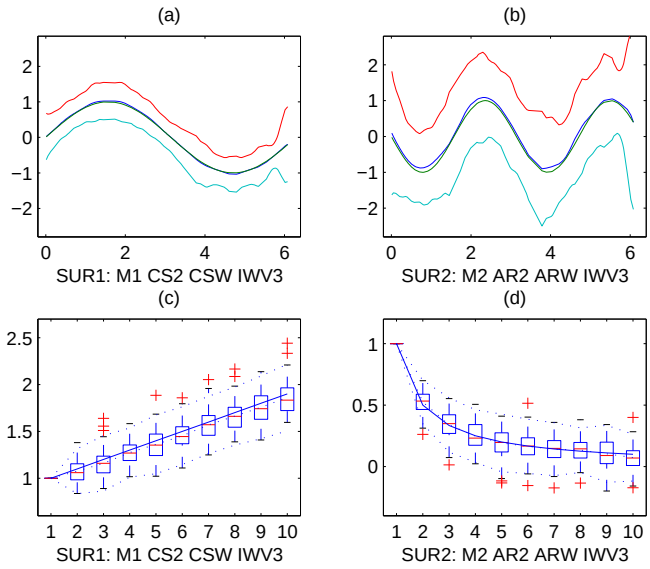

Abbildung 13.4: Nichtparametrische Schätzungen (a), (b) und Schätzungen der multiplikativen Parameter γ (c), (d) des SUR-Modells mit SUR1 in Compound-Struktur (CS2, CSW) und SUR2 in AR-Struktur (AR2, ARW) und SUR-Kovarianzstruktur (IWV2) für die Konfigurationen M1, M2, ungeglättet (MSE=71.29).

Diskrepanz zwischen Vorgabe und Schätzung auch im Falle hoher Varianzen minimal ist. Wieder sei auf die leichten Unterschätzungen an den Minima und Maxima von Abbildung 13.4(b) hingewiesen. Die Kombination aus linear ansteigenden multiplikativen Parametern γ im Compound-Modell und den exponentiell fallenden Parametern im AR-Modell wird vom Modell ohne die Einführung zusätzlicher Restriktionen akzeptiert und nahezu perfekte Schätzungen entstehen. Einige Ausreißer sind aber trotzdem in Abbildung 13.4(c), (d) zu erkennen. Auch für alle möglichen anderen Kombinationen von Konfigurationen AD, M1 und M2 zeigen sich ähnliche Ergebnisse für niedrige und hohe Varianzen.

13.2.2 Vergleich von Modellen mit unterschiedlichen Working-Kovarianz-Strukturen

Im folgenden soll ein Vergleich zweier Modelle durchgeführt werden, die sich nur aufgrund der Working-Kovarianz-Matrix unterscheiden. Hierbei geht man von zwei parallelen SUR-Modellen mit Compound-Struktur aus, die über die SUR-Annahme (IWV3) korrelieren oder unabhängig sind (IV). Die Working-Kovarianz-Matrizen für SUR1 und SUR2 werden gemäß den erwähnten Vorgaben geschätzt. Es wird dabei zwischen Working-Independence (IW) und voll spezifizierter Working-Kovarianz (CSW) unterschieden. Zum Vergleich wurde ein Compound-Compound-Modell mit hohen Varianzen (CS2) ausgewählt. Die Ergebnisse sind in Tabelle 13.3 zusammengefasst. Die beiden linken Spalten beinhalten die MSEs der Modelle mit SUR-Struktur (IV) simuliert mit Working-Independence-Annahme (IW) und voll spezifizierter Working-Kovarianz (CSW), die beiden rechten Spalten die MSEs der Modelle mit SUR-Struktur (IWV3) für (IW) und (CSW). Generell sollten bei korrekt spezifizierter Working-Kovarianz-Matrix (CSW) die MSEs deutlich niedriger sein als bei Working-Independence (IW). Dies ist in Tabelle 13.3 nur zum Teil der Fall, weil die jeweiligen Schätzungen der Working-Kovarianzen - und damit sind nicht nur die Blockdiagonalen Elemente der Matrix, sondern auch die SUR-Elemente gemeint - stark von den Vorgaben abweichen, teilweise völlig überschätzt sind. Folglich verändern sich die Einzel-Schätzungen stark und erhöhen den MSE. Im Mittel ergeben sich allerdings vernünftige Schätzungen, wie sich in den folgenden Abbildungen zeigen wird.

Daher stellt sich die Frage, ob die Momenten-Methode von Liang und Zeger (1986) zur Schätzung der Working-Kovarianz-Elemente in SUR-Modellen geeignet ist. Im nicht- und semiparametrischen Bereich wurden größtenteils Ergebnisse präsentiert, die die Schätzungen bei korrekt spezifizierter Working-Kovarianz verbesserten. Dies ist nun nicht mehr der Fall. Vor allem bei Modellen mit SUR-Struktur (IWV3) und korrekt spezifizierter Working-Kovarianz (CSW) werden, bis auf einen Fall, höhere MSEs errechnet als bei Independence-Annahme (IW). Bei Modellen mit

	SUR-Modell (HV) SUR1: CS2 SUR2: CS2		SUR-Modell (HV) SUR1: CS2 SUR2: CS2	
	(IV)(IW)	(IV)(CSW)	(IWV3)(IW)	(IWV3)(CSW)
AD, AD	70.81	79.09	76.20	95.86
AD, M1	70.07	89.74	75.51	104.31
AD, M2	66.29	63.98	74.35	81.97
M1, AD	70.26	79.10	75.95	106.03
M1, M1	69.29	77.51	74.46	98.72
M1, M2	65.61	61.99	74.51	88.00
M2, AD	66.12	63.40	74.38	79.47
M2, M1	65.22	61.84	74.03	85.84
M2, M2	61.57	46.25	71.23	54.46

Tabelle 13.3: SUR-Modell: Mean Squared Errors für alle möglichen Konfigurationen für hohe Varianzen mit zugrundeliegendem Compound-Compound-Modell und verschiedenen Working-Kovarianz-Strukturen (IW) und (CSW).

SUR-Struktur (IVW3) liegt das Verhältnis bei vier zu fünf. Nur bei Modellen, in denen exponentiell fallende Parameter γ im Modell enthalten sind, liefert die Momenten-Methode gute Working-Kovarianz-Schätzungen und als Folge auch niedrigere MSEs. Dies war auch in den Analysen zuvor schon festzustellen.

Jörgensen (1993) beschäftigt sich mit Ansätzen zur Verbesserung der Momenten-Methode, allerdings ist fraglich, ob dessen Ansätze zur Schätzung bei SUR-Modellen geeignet sind. Ansätze zur Schätzung, die sich direkt auf SUR-Modelle beziehen, konnten in der Literatur nicht gefunden werden. Damit lässt sich ein weiterer Kritikpunkt zu den im Theorieteil genannten acht Punkten von Lindsey und Lambert (1998) im Bezug auf die GEE-Schätzungen anfügen.

Drei der Modelle aus Tabelle 13.3 wurden geglättet. Die Ergebnisse sind in Tabelle

13.4 zusammengefasst. Auch in diesem Fall zeigt sich, dass der MSE durch geeignetes Glätten um über 50 % verringert werden kann.

	SUR-Modell (HV) SUR1: CS2 SUR2: CS2		
	ungeglättet	geglättet	%
AD, AD (IWV3)(IW)	76.20	37.08	51.3
AD, AD (IWV3)(CSW)	79.09	36.26	54.2
M2, M1 (IWV3) (CSW)	61.99	33.55	45.9

Tabelle 13.4: SUR-Modell: Mean Squared Errors für Modelle mit Glättung inklusive Verbesserungsgrad.

Die folgenden Abbildungen konzentrieren sich auf drei Modelle, die sich sehr ähnlich sind. Zunächst wird ein Modell ohne SUR-Korrelation betrachtet, im Anschluss dasselbe Modell mit SUR-Korrelation und zuletzt nochmals dasselbe Modell in geglätteter Form. Alle Abbildungen sind bezüglich der Achsenskalierung normiert. Abbildung 13.5 zeigt die nichtparametrischen Schätzungen ((a) und (b)) und die multiplikativen Parameter γ ((c) und (d)) für das Modell ohne SUR-Korrelation (IV). Als Basis wurde ein Compound-Compound-Modell mit hohen Varianzen (CS2, CS2) und korrekt spezifizierter Working-Kovarianz (CSW, CSW) für die Konfigurationen M1 und M2 ausgewählt. Gemittelt über 100 Simulationen zeigen sich trotz der hohen Varianzen gute Schätzungen, die von sehr wackeligen, weit entfernt liegenden Konfidenzintervallen eingerahmt werden. Erwartungsgemäß treten in Abbildung 13.5(b) an den Maxima und Minima der Sinusfunktionen wieder leichte Unterschätzungen auf. Diese Unterschätzungen sind auch in Abbildung 13.6 zu erkennen, allerdings in nicht ganz so ausgeprägter Form. Die Abbildungen 13.6(a), (b) zeigen auch wackelige und weit(er) entfernt liegende Konfidenzintervalle; die Schätzungen sind vergleichbar mit denen aus Abbildung 13.5. Abbildung 13.7 zeigt geglättete Parameter-Schätzungen. Man erkennt sofort, dass die Varianzen sich ver-

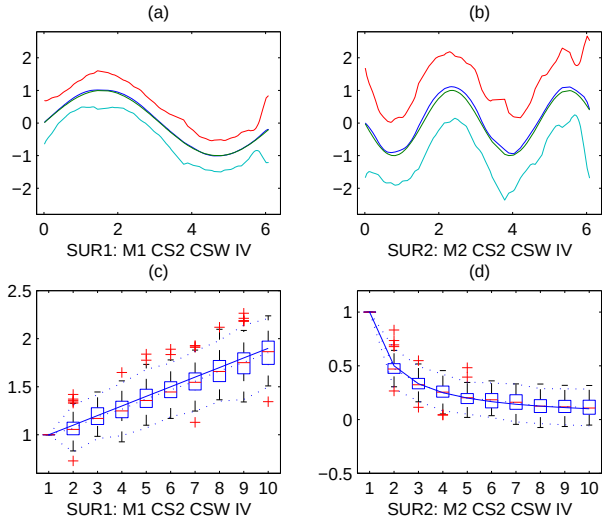

Abbildung 13.5: Nichtparametrische Schätzungen (a), (b) und Schätzungen der multiplikativen Parameter γ (c), (d) des SUR-Modells mit SUR1 in Compound-Struktur (CS2, CSW) und SUR2 in Compound-Struktur (CS2, CSW) und SUR-Kovarianzstruktur (IV) für die Konfigurationen M1, M2, ungeglättet (MSE=61.99).

kleinern und die Konfidenzintervalle deutlich näher liegen, allerdings auf Kosten eines leichten Bias, der sich in minimalen Unterschätzungen der Sinus-Kurven und den damit verbundenen Unterschätzungen der multiplikativen Parameter ausdrückt. Die Unterschätzungen der multiplikativen Parameter nehmen bei den Konfigurationen M1 und M2 bei weitem keine so großen Ausmaße an wie im Falle von AD. Dies zeigte sich auch schon im nicht- und semiparametrischen Bereich und auch in den SUR-Simulationen.

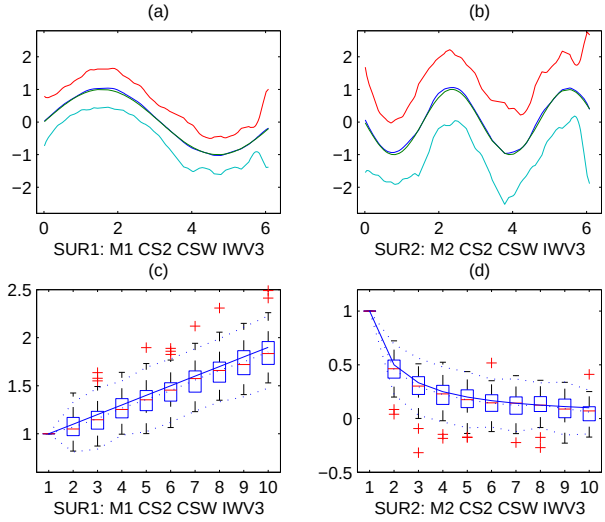

Abbildung 13.6: Nichtparametrische Schätzungen (a), (b) und Schätzungen der multiplikativen Parameter γ (c), (d) des SUR-Modells mit SUR1 in Compound-Struktur (CS2, CSW) und SUR2 in Compound-Struktur (CS2, CSW) und SUR-Kovarianzstruktur (IWV3) für die Konfigurationen M1, M2, ungeglättet (MSE=88.00).

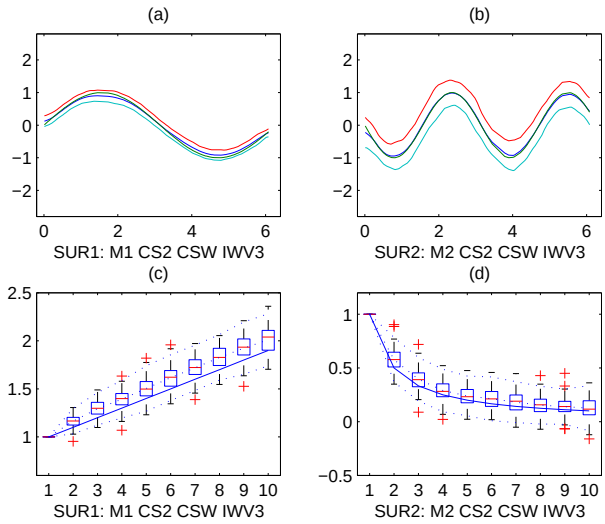

Abbildung 13.7: Nichtparametrische Schätzungen (a), (b) und Schätzungen der multiplikativen Parameter γ (c), (d) des SUR-Modells mit SUR1 in Compound-Struktur (CS2, CSW) und SUR2 in Compound-Struktur (CS2, CSW) und SUR-Kovarianzstruktur (IWV3) für die Konfigurationen M1, M2, geglättet (MSE=33.55).

Kapitel 14

Zusammenfassung

14.1 Zusammenfassung

Im Rahmen dieser Arbeit wurden drei statistische Modelle für Longitudinal-Daten ausführlich diskutiert. Dabei wurde ein nichtparametrisches Modell, später ein semiparametrisches Modell und zum Abschluss ein SUR-Modell hergeleitet und jeweils auch in Simulations-Studien überprüft.

Zunächst wurde das nichtparametrische Modell behandelt. Neben einem nichtparametrischen Term wurde ein Varying-Coefficients-Term und multiplikative Zeitvariierende Parameter in das Modell integriert. Die theoretische Herleitung folgte. Anschließend wurde das Modell für verschiedene Kovarianz- und Working-Kovarianz-Strukturen und Konfigurationen der Parameter simuliert. Dabei wurde im ersten Schritt der Studie zwischen einem 'Additiven Modell' ohne Zeit-variierende Parameter und einem 'Multiplikativen Modell' mit Zeit-variierenden Parametern γ unterschieden und beide Modelle miteinander verglichen. Es zeigte sich, dass additive Modelle in manchen Situationen nicht anwendbar sind, während die multiplikativen Modelle hingegen generell einsetzbar sind, auch falls keine Zeit-Variation besteht. Außerdem wurden verschieden Analysen und Vergleiche nur für das multiplikative Modell durchgeführt. Dabei war von Interesse, ob das Modell in allen möglichen

Kombinationen an Parameter-Konfigurationen einsetzbar ist. Nicht nur für hohe und niedrige Varianzen, sondern auch für konstante, linear ansteigende und exponentiell fallende multiplikative Parameter und verschiedene Kovarianz- und Working-Kovarianz-Strukturen wurden Simulationen durchgeführt. Zusätzlich wurde in Form von Differenzen-Penalties geglättet, um den Trend der Daten besser zu erfassen. Die Güte der Modelle wurde zunächst über das AIC und BIC-Kriterium erfasst. In den allermeisten Fällen zeigte sich, dass das BIC-Kriterium die bessere Wahl zur Beurteilung der Güte der Modelle ist, da sich der Einfluss der Devianz auf das Kriterium reduziert.

Generell wurden sehr gute Schätzungen erzielt. Im ungeglätteten Fall überlagerten sich vorgegebene Funktion und Schätzung im Mittel über 100 Simulationen meistens sehr exakt, während bei Modellen mit Glättung sehr oft ein leichter Bias zu erkennen war. Dieser drückte sich in einer Unterschätzung der vorgegebenen Funktion in Kombination mit einer Überschätzung der multiplikativen Parameter aus. Vor allem bei Modellen mit konstanten multiplikativen Parametern war der Bias sehr deutlich zu erkennen, allerdings reduzierte sich der MSE in solchen Fällen trotzdem, obwohl man aufgrund der Graphiken zunächst annimmt, die Modelle seien suboptimal geschätzt. Aus der Kombination an Unter- und Überschätzung, die sich sozusagen gegenseitig ausgleichen, ergaben sich in diesen Fällen bessere Schätzer mit deutlich besseren Varianzschätzungen und leicht erhöhtem Bias. Glätten verringerte den MSE bei hohen Varianzen um bis zu 50%.

Auch die korrekte Spezifikation der Working-Kovarianz-Matrix hat Auswirkungen auf den MSE. Beim Vergleich zu Modellen mit Independence-Annahme wurden größtenteils leichte Verbesserungen des MSE festgestellt. Speziell bei Modellen mit exponentiell fallenden multiplikativen Parametern wurden die deutlichsten Verbesserungen erzielt. In einigen Fällen konnten keine Verbesserungen erzielt werden. Diese sind auf schlechte Schätzungen der Working-Kovarianz zurückzuführen.

Desweiteren wurden Analysen für multiplikative Parameter mit Richtungswechsel durchgeführt. Dies war notwendig, da im Testdatensatz zunächst ein starker Anstieg mit darauf folgendem kontinuierlichen Abfall der multiplikativen Parameter zu erkennen war. Der Testdatensatz beinhaltet sechs Messzeitpunkte für Cortisol, über den Tag hinweg gemessen, und den jeweiligen Body-Mass-Index (BMI), der als konstant angenommen wird. Beide Größen wurden mit Hilfe des zuvor beschriebenen Modells in Beziehung zueinander gesetzt. Es zeigte sich, dass die Schätzungen und Ergebnisse mit denen anderer Forschergruppen übereinstimmten. Demnach haben vor allem dicke Menschen einen um maximal 25% verringerten Cortisolausstoss im Vergleich zu Menschen mit normalem BMI. Sehr dünne Teilnehmer wiesen gemäß der Schätzung auch deutlich niedrigere Werte auf.

Im nächsten Schritt wurde das nichtparametrische Modell in ein semiparametrisches Modell überführt. Dazu wurde ein zusätzlicher parametrischer Term in das Modell integriert. Neben dem Varying-Coefficients- und dem nichtparametrischen Term mit multiplikativem Parameter wurde also ein vierter Parameter angefügt, der zusätzlich zu schätzen war. An der Schätzmethode und auch an den Konfigurationen, sowie den zugrundeliegenden Kovarianz- und Working-Kovarianz-Matrizen wurde nichts verändert. Nach der theoretischen Ableitung des Modells wurde eine weitere Simulationsstudie durchgeführt. Diese zielte auf ähnlich Inhalte ab, wie die zuvor beschriebene Studie des nichtparametrischen Modells. Zunächst wurde wieder ein Vergleich zwischen additivem und multiplikativem Modell angestellt. Erwartungsgemäß zeigte sich, dass das multiplikative Modell in gewissen Situationen dem additiven überlegen ist. Desweiteren wurde der Einfluss von Glättung auf die einzelnen Parameter untersucht, sowie Modelle mit unterschiedlichen Kovarianzstrukturen und Working-Kovarianz-Matrizen. Zwei zusätzliche Gütekriterien wurden in die Analysen miteinbezogen, nachdem schnell klar geworden war, dass in diesem Fall auch das BIC-Kriterium nicht mehr funktionierte. Beide Kriterien, CV und GCV, basieren auf Kreuzvalidierungsmechanismen, sind aber im Grunde genommen den bisher verwendeten Informations-Kriterien sehr ähnlich. Auch diese beiden Kriterien konnten die Güte der Modelle nicht ausreichend bewerten.

Größtenteils wurden trotz steigender Komplexität des Modells sehr gute Schätzungen erzielt. Die Varying-Coefficients-Terme wurden linear ansteigend gewählt, der parametrische Term wurde eindimensional in das Modell integriert und die zugrundeliegende Funktion des nichtparametrischen Terms wurde verändert. Die Konfigurationen multiplikativen Parameter wurden beibehalten. Generell fiel auf, dass der parametrische Term auch bei hohen Varianzen sehr exakt geschätzt wurde, mit minimalen Standardabweichungen, und dass auch die Varying-Coefficients-Terme, die nicht mehr konstant gewählt wurden, in den allermeisten Fällen sehr gut geschätzt wurden. Ansonsten sind ähnliche Ergebnisse und Probleme wie im parametrischen Modell zu berichten.

Als drittes und letztes Modell wurde ein SUR-Modell für Longitudinaldaten eingehend betrachtet. Carroll (2003) betrachtet zwar ein ähnliches SUR-Modell, allerdings ohne Longitudinalstruktur. Zwei Kovarianz-Strukturen mussten geschätzt werden, zum einen die Working-Kovarianzen innerhalb der Einzelmodelle, zum anderen die SUR-Struktur zwischen den (beiden) Modellen. Dies wurde über die bekannten Momentenschätzer bewerkstelligt. Es zeigte sich, dass die Momentenschätzung in einigen Fällen ungeeignet war, um die die Working-Kovarianz und auch die SUR-Struktur ausreichend gut zu schätzen. Nur für Modelle, in denen zumindest eines der Modelle exponentiell fallende multiplikative Effekte aufwies, zeigten sich sehr gute Schätzungen und daher auch deutlich verringerte MSEs. Andere Möglichkeiten zur Schätzung der Kovarianzstruktur werden von Jörgensen (1993) vorgeschlagen. Es ist aber zweifelhaft, ob derartige Verfahren bei SUR-Modellen in Kombination mit Longitudinaldaten einsetzbar sind. Keine weiteren Methoden zur Schätzung beider Strukturen sind nach Meinung des Autors zu diesem Thema bisher veröffentlicht worden. Dies scheint ein Ansatzpunkt zu sein, der auch zukünftig diskutiert werden könnte.

Die Schätzungen fallen gemittelt über 100 Simulationen bei niedrigen und hohen

Varianzen und verschiedenen Modellstrukturen sehr gut aus. Es wurden Modelle mit unterschiedlichen Kovarianzstrukturen, d.h. Compound- und AR-Strukturen gemeinsam in einem Modell, simuliert. Verschiedene SUR-Working-Kovarianz-Strukturen wurden in die Simulationen miteinbezogen und Modelle mit und ohne SUR-Struktur verglichen. Die Verbesserung der Schätzungen durch Glättung konnte aufgrund der hohen Anzahl an Glättungsparametern nur anhand einiger Modelle demonstriert werden. Dabei zeigte sich, wie schon zuvor, dass Glättung den MSE stark reduziert und die Konfidenzintervalle deutlich glatter werden und folglich näher an der Schätzung liegen. Erwartungsgemäß traten bei den meisten Schätzungen wieder leichte Biases auf.

Zusammenfassend kann gesagt werden, dass vor allem das nicht- und semiparametrische Modell Praxisrelevanz haben. Bei geeigneten Datensätzen, wie dem, der in dieser Arbeit benutzt wurde, sind derartige Modelle problemlos anwendbar. Nach Einführung bestimmter Restriktionen sind alle Parameter identifizierbar. Dies kompliziert zwar die Analysen und Bestimmung einiger wichtiger Größen, hat aber den Vorteil, dass man das Modell mit seinem Annahmen sehr gut zu verstehen lernt. Dieses Verständnis ist vor allem dann von großem Nutzen, wenn zusätzliche Komponenten in das Modell eingebaut werden, wie z.B. die semiparametrische und SUR-Struktur, und die einzelnen Programmteile, wie z.B die Hatmatrizen, deutlich komplizierter und auch rechenaufwendiger werden. Eine weitere Hauptschwierigkeit bestand in einer effizienten algorithmischen Umsetzung der Modelle. Während die ersten Simulationen noch sehr viel Zeit brauchten, wurden Stück für Stück Verbesserungen vorgenommen, um die Rechendauer zu reduzieren. Hauptsächlich ist die Reduzierung auf bessere Programmiertechnik und auf den 'Sparse'-Befehl, der in Matlab (www.mathworks.com) integriert ist, zurückzuführen. Dieser Befehl reduziert die Rechenzeit um ein Vielfaches, indem die Nullelemente innerhalb der Matrizen bei Multiplikationen keine Beachtung mehr finden. Auch programmiertechnische Kleinigkeiten wie Klammersetzungen führten zu deutlichen Verkürzungen der Rechenzeit. So konnte beispielsweise die Dauer einer bestimmten Simulation beim SUR-Modell von circa vier Stunden auf eine Stunde reduziert werden. Daher waren

Grid-Searches mit sechs Glättungsparametern erst möglich. Für die Grid-Searches und alle weiteren Simulationen konnten zwei Server des Instituts für Verhaltensforschung an der ETH Zürich über Wochen hinweg genutzt werden.

Abschließend sei nochmals allen Personen gedankt, die mich während des Entstehens der Dissertation moralisch und statistisch unterstützt haben.

14.2 Summary

Three regression models were proposed in this dissertation. Firstly a nonparametric model with varying-coefficients and multiplicative, time-varying parameters was derived, followed by a semiparametric and a seemingly unrelated regression model (SUR).

The nonparametric part of the dissertation consists of the derivation of the model and a simulation-study. In this study the model was verified using different covariance and working-covariance structures and parameter configurations. Firstly comparisons of an additive model without time-varying modification and a multiplicative model with time-varying modification were carried out. The analyses showed that the additive models are not very useful in some situations, while the multiplicative model is applicable even if no variation across time is present in the underlying data. In addition different analyses and comparisons were made for the multiplicative model. The model was checked for low and high variances, for constant, linear increasing and exponential decreasing multiplicative parameters and different settings of the working-covariance, like compound- and autoregressive correlation. Difference-penalties on coefficients of adjacent B-Splines were introduced to the model for smoothing and the fit of the model was assessed by the AIC- and mainly by the BIC-criterion that appeared to be the best choice for the nonparametric model.

In general the results of the estimations were excellent. In the majority of cases the

underlying nonparametric function was completely superimposed by the estimation, averaged over 100 simulations, if no smoothing was used. Smoothed estimates often showed a small bias expressed by a barely visible underestimation of the underlying sinus. The underestimation of the sinus led to an overestimation of the multiplicative parameters or vice versa. Especially smoothed models with constant multiplicative parameters showed larger biases. But the combination of underestimation of the nonparametric function and the overestimation of the multiplicative parameters led to smaller MSEs for smoothed models even if the estimates seemed to be inferior at first sight. Smoothing reduced the MSEs up to 50 % for high variances.

The MSE is also influenced by the specification of the working-covariance. Comparisons between models with correct specification and models with independent working-covariance mainly show small improvements of the MSE. Models with exponential decreasing multiplicative parameters showed the strongest improvements. In some cases no improvements of the MSE could be noticed because of inappropriate estimates of the working-covariance.

Further analyses were carried out for multiplicative parameters changing direction. This was necessary because the application-data showed multiplicative parameters changing direction with a strong increase at the beginning followed by a continuous decrease. The dataset consists of 6 repeated measurements for cortisol, measured over the course of the day, and the associated Body-Mass-Index (BMI) per person. Both measurements were introduced to the nonparametric model described above. The analysis of the estimates corresponds with the results of related publications. According to our results obese people have an up to 25 % higher cortisol secretion than people with normal BMI. Very skinny people show signs of a lowered cortisol secretion as well.

Secondly the nonparametric model was expanded into a semiparametric model. An additional parametric term was introduced to the model that already consisted of a

nonparametric term with multiplicative parameters and a varying-coefficients term. The general settings concerning the configurations and the method of estimation remained unchanged. After the theoretical derivation of the semiparametric model another simulation-study was carried out. This study deals with similar problems and contents like the previous study. At first another comparison between an additive and a multiplicative semiparametric model was made, then the influence of smoothing on the parameters was analyzed and models with differing covariances and working-covariances were examined. Two additional model-selection-criteria, Cross-Validation (CV) and Generalized-Cross-Validation (GCV), were incorporated into the analyses, because neither the AIC nor the BIC-criterion worked properly. Both Cross-Validation-Criteria were unable to evaluate the model fit sufficiently, too.

The increased complexity of the semiparametric model did not have an impact on the quality of the estimation. In general the parametric terms were estimated very precise with standard deviations close to zero. The varying-coefficients that were chosen to be linear increasing also showed very good estimates as well as the nonparametric term even if correlation is introduced to the model.

The remaining model proposed is a seemingly unrelated regression model (SUR) in the context of longitudinal data. Carroll (2003) proposed a similar SUR-model but without time-dependence of the response. Two working-correlation structures have to be managed in this model: Firstly the working-covariance within the seemingly unrelated regressions, secondly the SUR-structure between both models. The estimations were carried out using the method of moments. This method of estimation appeared to be inappropriate in most cases. Just for models with time-vanishing multiplicative effects the estimation worked properly. Related approaches to estimate the working-covariance were published by Jörgensen (1993), but it is questionable whether these methods are appropriate for SUR-models in combination with longitudinal data. Further literature research was ineffective.

Another simulation study was accomplished. The estimates averaged over 100 simulations were better than expected even if the estimations were based on high variances and different configurations. Models with and without SUR-correlation were analyzed and grid-searches were carried out. Because of the comparably high number of smoothing parameters only a few models could be analyzed by grid-search. The results showed that the MSE is reduced by smoothing and the variances decrease at the expense of a mainly small bias.

Mainly the non- and semiparametric model are applicable to adequate longitudinal data. After the introduction of appropriate restrictions the derivations and analyses are definitely more complex and complicated, but it keeps the parameters identifiable and helps to better understand the model and the underlying assumptions. The gained knowledge is useful if parameters like the parametric or SUR-part are added to the model and the programming gets more complicated. Another difficulty was efficient coding in Matlab. At the beginning the simulations took a long time to finish, but the code was consistently improved mainly by more efficient programming and some special Matlab-commands. In some cases the computing time could be reduced from four to one hours. Therefore grid-searches with up to six smoothing parameters could be accomplished. For the grid-searches two servers of the Institute of Behavioural Science of the ETH Zurich were used.

Finally I would like to thank everybody who supported me morally and statistically over the past years.

Kapitel 15

Appendix

Nichtparametrisches Modell

Design-Matrizen

Der lineare Prädiktor hat die Form

$$\eta_i = V_i\beta_0 + \Gamma Z_i\alpha$$

mit

$$\begin{array}{ll} (V_i)_{T\times T} = I, & (\beta_0)_{T\times 1} = (\beta_{01},\ldots,\beta_{0T})', \\ (\Gamma)_{T\times T} = \mathrm{diag}(\gamma_1,\ldots,\gamma_T), & (Z_i)_{T\times r} = (z_{i1},\ldots,z_{iT})', \\ (z_{it})_{1\times r} = (\tilde{z}_{it1},\ldots,\tilde{z}_{itr}), & (\tilde{z}_{itl})_{1\times 1} = \Phi_l^{\alpha}(z_{itl}), \\ (\alpha)_{r\times 1} = (\alpha_1,\ldots,\alpha_r)', & (\gamma)_{T\times 1} = (\gamma_1\ldots,\gamma_T)'. \end{array}$$

Penalty-Matrizen

Die Pseudo-Log-Likelihood-Funktion l_p wird penalisiert über Differenzenpenalties. Generell ist die Penalty-Matrix benachbarter Koeffizienten von Basisfunktionen ζ definiert als

$$K = \sum_{l=1}^{r} (\Delta_l^d \zeta_l)^2 = \zeta' D' D \zeta.$$

Δ_l wird als Differenzenoperator bezeichnet, d.h. $\Delta_l \zeta_l = \zeta_l - \zeta_{l-1}$ für First Order Differenzen und $\Delta_l^2 \zeta_l = \Delta_l(\zeta_l - \zeta_{l-1}) = \zeta_l - 2\zeta_{l-1} + \zeta_{l-2}$ für Second Order Differenzen, usw., und D ist die zugehörige Differenzen-Matrix mit $\zeta = (\zeta_1, \ldots, \zeta_r)'$. Diese Penalty-Matrix wird auf die Koeffizienten des Modells β_0, α und γ übertragen. Die zugehörigen Penalty-Parameter sind λ_{β_0}, λ_α und λ_γ. Die Penalty Matrizen des Modells sind durch

$$\begin{aligned} K_\delta &= \text{diag}(\lambda_{\beta_0} K, \lambda_\alpha K) = \text{diag}(K_{\beta_0}, K_\alpha), \\ K_\gamma &= \lambda_\gamma K. \end{aligned}$$

vollständig definiert.

Hat-Matrizen

Aus (7.8) lassen sich die Hatmatrizen ableiten. Die Hatmatrix von δ und die abgeleitete Hat-Matrix von γ sind definiert durch

$$\begin{aligned} \hat{y}_\delta &= \Phi_1(\Phi_1' W^{-1} \Phi_1 + K_\delta)^{-1} \Phi_1' W^{-1} y \\ &= H_\delta y, \\ \hat{y}_\gamma &= \Phi_2(\Phi_2' W^{-1} \Phi_2 + K_\gamma)^{-1} \Phi_2' W^{-1}(y - V\beta_0) \\ &= \Phi_2(\Phi_2' W^{-1} \Phi_2 + K_\gamma)^{-1}(\Phi_2' W^{-1} y - \Phi_2' W^{-1} B[\Phi_1' W^{-1} \Phi_1 + K_\delta]^{-1} \Phi_1' W^{-1} y) \\ &= \Phi_2(\Phi_2' W^{-1} \Phi_2 + K_\gamma)^{-1}(\Phi_2' W^{-1} - \Phi_2' W^{-1} B(\Phi_1' W^{-1} \Phi_1 + K_\delta)^{-1} \Phi_1' W^{-1}) y \\ &= H_\gamma y \end{aligned}$$

Die Matrix B ist dabei eine erweiterte Version von V mit den Dimensionen $nT \times (T + r)$. Die angehängten r Spalten enthalten nur Nullen. Dies ist notwendig weil nur die Schätzungen von β_0 in die Hat-Matrix H_γ miteinbezogen werden, die Schätzungen von α können vernachlässigt werden. Die gemeinsame Hat-Matrix ist dann gegeben durch H=Diag(H_δ, H_γ). Die Spur von H gibt die approximative Anzahl an Freiheitsgraden an (Hastie and Tibshirani, 1990).

Zweite Ableitungen

Die Komponenten von $(\partial l_p/\partial\nu'\partial\nu) = (\partial\eta/\partial\nu)'W^{-1}(\partial\eta/\partial\nu')$ sind gegeben durch

$$\begin{aligned}
\left(\frac{\partial\eta}{\partial\delta}\right)' W^{-1}\left(\frac{\partial\eta}{\partial\delta}\right) &= \Phi_1' W^{-1}\Phi_1 + K_\delta, \\
\left(\frac{\partial\eta}{\partial\gamma}\right)' W^{-1}\left(\frac{\partial\eta}{\partial\gamma}\right) &= \Phi_2' W^{-1}\Phi_2 + K_\gamma, \\
\left(\frac{\partial\eta}{\partial\delta}\right)' W^{-1}\left(\frac{\partial\eta}{\partial\gamma}\right) &= \Phi_1' W^{-1}\Phi_2, \\
\left(\frac{\partial\eta}{\partial\gamma}\right)' W^{-1}\left(\frac{\partial\eta}{\partial\delta}\right) &= \Phi_2' W^{-1}\Phi_1.
\end{aligned}$$

Restringiertes Nichtparametrisches Modell

Zwei Restriktionen wurden in das Modell integriert um die Schätzer identifizierbar zu halten. Die erste Restriktion $\alpha_r = -\sum_{l=1}^{(r-1)} \alpha_l$ trennt β_0 und α und hat einen Einfluss auf die Design-Matrix Φ_1. Die Matrix verkleinert sich von $nT \times (T + r)$ auf die Größe $nT \times (T + r - 1)$, weil α_r durch die übrigen α_l ausgedrückt wird. Die zweite Restriktion $\gamma_1 = 1$ trennt α und γ und verkleinert die Design-Matrix Φ_2 von $nT \times T$ zu $n(T-1) \times (T-1)$, weil γ_1 als konstanter Wert definiert ist. Dies hat zur Folge, dass die Gleichungen mit γ_1 nicht geschätzt werden. Die verschiedene Größen der restringierten Vektoren und Matrizen erschweren die Analysen und verlängern die Rechenzeiten teilweise um ein Vielfaches.

First-Order Penalty-Matrizen und Schätzgleichungen

Die Block-Diagonale Form der Penalty-Matrix K_δ bleibt unverändert, aber K_α verändert sich aufgrund der ersten Restriktion. Folgende Matrixdarstellung zeigt

die Veränderungen für First Order-Penalties.

$$D_\alpha = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ -1 & -1 & -1 & \dots & -1 & 1 \end{pmatrix}$$

$$K_\alpha = D'_\alpha D_\alpha$$

Die erste Schätzgleichung von (7.8)

$$\hat{\delta} = (\Phi'_1 W_\delta^{-1} \Phi_1 + K_\delta)^{-1} \Phi'_1 W_\delta^{-1} y \tag{15.1}$$

bleibt unverändert, aber die Working-Kovarianz und einige Elemente von Φ_1 sind verschieden wegen der ersten Restriktion im Vergleich zum Modell ohne Restriktionen. Die zweite Restriktion verändert die Schätzgleichung und auch die Penalty-Matrix. Ein konstanter Term wird nun an die zweite Schätzgleichung von (7.8) angehängt

$$\hat{\gamma} \;=\; (\Phi'_2 W_\gamma^{-1} \Phi_2 + K_\gamma)^{-1} (\Phi'_2 W_\gamma^{-1} (y - V\hat{\beta}_0) - \lambda_\gamma D_\gamma e), \tag{15.2}$$

wobei

$$e = \begin{pmatrix} -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{und} \quad D_\gamma = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix}$$

mit $K_\gamma = D'_\gamma D_\gamma$.

Restringierte Working-Kovarianz-Matrix

Die restringierten Working-Kovarianz-Matrizen W_δ und W_γ haben unterschiedliche Größen je nach Schätzgleichung (15.1) oder (15.2). Die restringierte zweite Gleichung

(15.2) beinhaltet eine kleinere Working-Kovarianz-Matrix, weil bekanntlich γ_1 nicht geschätzt wird. In jeder Dimension wird $(W_i)_\gamma$ daher um eine Dimension kleiner, d.h. man schätzt eine eigene Kovarianzmatrix für $\gamma_{reduziert} = (\gamma_2, \ldots, \gamma_T)'$.

Die restringierte erste Gleichung (15.1) verliert keine Dimension(en), allerdings müssen einige Elemente der Working-Kovarianz-Matrix angepasst werden. Dies wird anhand der folgenden Matrix demonstriert. Die Matrix ist in Compound-Struktur.

$$(W_\delta)_i = \begin{pmatrix} 0.1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0.1 & 0.05 & 0.05 & \ldots & 0.05 \\ 0 & 0.05 & 0.1 & 0.05 & \ldots & 0.05 \\ 0 & 0.05 & 0.05 & 0.1 & \ldots & 0.05 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0.05 & 0.05 & 0.05 & \ldots & 0.1 \end{pmatrix}$$

Alle Elemente von γ_1 außer $(W_\delta)_{11}$ werden auf Null gesetzt. Erfolgt keine Anpassung der Working-Kovarianz-Matrix, erhält man entweder schlechte Schätzungen, im Extremfall kann es vorkommen, dass der Algorithmus nicht konvergiert. Die notwendige Anpassung lässt sich auf die Restriktion $\gamma_1 = 1$ zurückführen.

Penalisierte Hat-Matrix

Die Hat-Matrix wird ähnlich abgeleitet wie zuvor beschrieben. Allerdings tritt ein zusätzliches Problem auf. Wegen Restriktion 2 haben die Vektoren und auch die Gleichungssysteme unterschiedliche Dimensionen. Daher müssen die Matrizen Φ_1 und Φ_2 und Vektoren y_δ und y_γ aneinander angepasst werden. Das Problem wird durch die Einführung zweier zusätzlicher Designmatrizen B_1 und C gelöst. Eine dieser Matrizen, B_1, ist ähnlich aufgebaut wie B (unrestringiertes Modell). Der zusätzliche konstante Term $\lambda_\gamma D_\gamma e$ wird angepasst indem er mit $\frac{1}{n(T-1)}\hat{y}'_\gamma y_\gamma = 1$

multipliziert wird, mit $\hat{y}_\gamma = 1/y_\gamma$ und

$$\begin{aligned}
\hat{y}_\delta &= \Phi_1(\Phi_1' W_\delta^{-1}\Phi_1 + K_\delta)^{-1}\Phi_1' W_\delta^{-1} y_\delta \\
&= H_\delta y_\delta, \\
\hat{y}_\gamma &= \Phi_2(\Phi_2' W_\gamma^{-1}\Phi_2 + K_\gamma)^{-1}(\Phi_2' W_\gamma^{-1}(y_\gamma - V\beta_0) - \lambda_\gamma D_\gamma e) \\
&= \Phi_2(\Phi_2' W_\gamma^{-1}\Phi_2 + K_\gamma)^{-1}\Phi_2' W_\gamma^{-1} y_\gamma \\
&- \Phi_2(\Phi_2' W_\gamma^{-1}\Phi_2 + K_\gamma)^{-1}\Phi_2' W_\gamma^{-1} B_1(\Phi_1' W_\delta^{-1}\Phi_1 + K_\beta)^{-1}\Phi_1' W_\delta^{-1} C y_\gamma \\
&- \Phi_2(\Phi_2' W_\gamma^{-1}\Phi_2 + K_\gamma)^{-1}\lambda_\gamma D_\gamma e \frac{1}{n(T-1)} \hat{y}_\gamma' y_\gamma \\
&= H_\gamma y_\gamma.
\end{aligned}$$

Außerdem gilt $\hat{B}_1' = (0_{(T-1)\times 1}, I_{(T-1)\times(T-1)}, 0_{(T-1)\times(r-1)})$, $B_1 = (\hat{B}_1', \ldots, \hat{B}_1')'$, n-mal, und $y_\delta = C y_\gamma$. W_δ und W_γ sind abgeleitete Versionen von W, $H = \mathrm{diag}(H_\delta, H_\gamma)$.

Inferenz

Es ist deutlich komplizierter als erwartet die Sandwich-Matrix im restringierten Modell abzuleiten, weil die Dimensionen der Design-Matrizen $\Phi_1 = \Phi_1(\gamma)$ und $\Phi_2 = \Phi_2(\alpha)$ sich unterscheiden. Die Ableitungen

$$\begin{aligned}
\left(\frac{\partial \eta}{\partial \delta}\right)' W_\delta^{-1} \left(\frac{\partial \eta}{\partial \delta}\right) &= \Phi_1' W_\delta^{-1}\Phi_1 + K_\delta, \qquad (15.3) \\
\left(\frac{\partial \eta}{\partial \gamma}\right)' W_\gamma^{-1} \left(\frac{\partial \eta}{\partial \gamma}\right) &= \Phi_2' W_\gamma^{-1}\Phi_2 + K_\gamma
\end{aligned}$$

bleiben zwar unverändert, aber die Ableitungen, bei denen sich beide Schätzgleichungen aus (7.8) überschneiden, unterscheiden sich vom unrestringierten Modell. Es gilt

$$\left(\frac{\partial\eta}{\partial\gamma}\right)' W^{-1}_{\gamma,\beta_0}\left(\frac{\partial\eta}{\partial\beta_0}\right) = \begin{pmatrix} \sum_{i=1}^n z'_{it}\alpha w^{-1}_{i12} & \sum_{i=1}^n z'_{it}\alpha w^{-1}_{i22} & \dots & \sum_{i=1}^n z'_{it}\alpha w^{-1}_{iT'2} \\ \sum_{i=1}^n z'_{it}\alpha w^{-1}_{i13} & \sum_{i=1}^n z'_{it}\alpha w^{-1}_{i23} & \dots & \sum_{i=1}^n z'_{it}\alpha w^{-1}_{iT'3} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n z'_{it}\alpha w^{-1}_{i1T} & \sum_{i=1}^n z'_{it}\alpha w^{-1}_{i2T} & \dots & \sum_{i=1}^n z'_{it}\alpha w^{-1}_{iT'T} \end{pmatrix}_{(T-1)\times(p-1)}$$

$$\left(\frac{\partial\eta}{\partial\gamma}\right)' W^{-1}_{\gamma,\alpha}\left(\frac{\partial\eta}{\partial\alpha}\right) = \begin{pmatrix} \sum_{i=1}^n z'_{it}\alpha z'_{it}\sum_{t=1}^T \gamma_t w^{-1}_{i2t} \\ \sum_{i=1}^n z'_{it}\alpha z'_{it}\sum_{t=1}^T \gamma_t w^{-1}_{i3t} \\ \vdots \\ \sum_{i=1}^n z'_{it}\alpha z'_{it}\sum_{t=1}^T \gamma_t w^{-1}_{iTt} \end{pmatrix}_{(T-1)\times T}$$

$$\left(\frac{\partial\eta}{\partial\alpha}\right)' W^{-1}_{\alpha,\beta_0}\left(\frac{\partial\eta}{\partial\beta_0}\right) = \begin{pmatrix} \sum_{i=1}^n z'_{it}\sum_{t=1}^T \gamma_t w^{-1}_{i1t} \\ \sum_{i=1}^n z'_{it}\sum_{t=1}^T \gamma_t w^{-1}_{i2t} \\ \vdots \\ \sum_{i=1}^n z'_{it}\sum_{t=1}^T \gamma_t w^{-1}_{iTt} \end{pmatrix}_{T\times(p-1)}.$$

Die Matrizen W^{-1}_{γ,β_0}, $W^{-1}_{\gamma,\alpha}$ und W^{-1}_{α,β_0} sind abgeleitete Versionen von W^{-1}. Die erste und zweite Ableitung $(\partial\eta/\partial\gamma)'W^{-1}_{\gamma,\beta_0}(\partial\eta/\partial\beta)$ und $(\partial\eta/\partial\gamma)'W^{-1}_{\gamma,\alpha}(\partial\eta/\partial\alpha)$ verlieren je eine Dimension, weil die Ableitungen des konstanten Parameters γ_1 fehlen. Die erste Restriktion reduziert die erst- und drittgenannte Ableitung je um eine Dimension. Die Ableitung von $(\partial\eta/\partial\alpha)'W^{-1}_{\alpha,\beta_0}(\partial\eta/\partial\beta)$ ist ein Teil der Ableitung von (15.3) und kann verwendet werden um Teile der anderen Ableitungen zu überprüfen. Falls $z_i = z_{it}$ werden die Parameter z_i anstelle von z_{it} bei den Ableitungen verwendet. $w^{-1}_{it't}$ ist ein Element der Inversen der Block-Diagonalen Working-Kovarianz-Matrix. Generell läuft die Schätzung der Working-Kovarianz-Matrix genauso wie im unrestringierten Modell ab.

Semiparametrisches Modell

Design-Matrizen

Der lineare Prädiktor hat die Form

$$\eta_i = V_i\beta_0 + X_i\beta + \Gamma Z_i\alpha$$

mit

$$\begin{array}{ll}
(V_i)_{T\times T} = I, & (\beta_0)_{T\times 1} = (\beta_{01},\ldots,\beta_{0T})', \\
(X_i)_{T\times p} = (x_{i1},\ldots,x_{ip}), & (x_{ij})_{T\times 1} = (x_{i1j},\ldots,x_{iTj})', \\
(\beta)_{p\times 1} = (\beta_1,\ldots,\beta_p), & (\Gamma)_{T\times T} = \mathrm{diag}(\gamma_1,\ldots,\gamma_T), \\
(Z_i)_{T\times r} = (z_{i1},\ldots,z_{iT})', & (z_{it})_{1\times r} = (\tilde{z}_{it1},\ldots,\tilde{z}_{itr}), \\
(\tilde{z}_{itl})_{1\times 1} = \Phi_l^{\alpha}(z_{itl}), & (\alpha)_{r\times 1} = (\alpha_1,\ldots,\alpha_r)', \\
(\gamma)_{T\times 1} = (\gamma_1\ldots,\gamma_T)'. &
\end{array}$$

Penalty-Matrizen

$$\begin{aligned}
K_\delta &= \mathrm{diag}(\lambda_{\beta_0}K, 0\cdot K, \lambda_\alpha K) = \mathrm{diag}(K_{\beta_0}, K_\beta, K_\alpha), \\
K_\gamma &= \lambda_\gamma K.
\end{aligned}$$

Hat-Matrizen

Die Hatmatrizen werden nach demselben Schema abgeleitet wie im nichtparametrischen Modell. Es gilt

$$\begin{aligned}
\hat{y}_\delta &= \Phi_1(\Phi_1' W^{-1}\Phi_1 + K_\delta)^{-1}\Phi_1' W^{-1} y \\
&= H_\delta y, \\
\hat{y}_\gamma &= \Phi_2(\Phi_2' W^{-1}\Phi_2 + K_\gamma)^{-1}\Phi_2' W^{-1}(y - V\beta_0 - X\beta) \\
&= \Phi_2(\Phi_2' W^{-1}\Phi_2 + K_\gamma)^{-1} \times \\
&\times \ (\Phi_2' W^{-1} y - \Phi_2' W^{-1} B[\Phi_1' W^{-1}\Phi_1 + K_\delta]^{-1}\Phi_1' W^{-1} y - X\beta y^{-1} y) \\
&= \Phi_2(\Phi_2' W^{-1}\Phi_2 + K_\gamma)^{-1} \times \\
&\times \ (\Phi_2' W^{-1} - \Phi_2' W^{-1} B(\Phi_1' W^{-1}\Phi_1 + K_\delta)^{-1}\Phi_1' W^{-1} - X\beta y^{-1}) y \\
&= H_\gamma y,
\end{aligned}$$

mit $y^{-1}y = 1$.

Zweite Ableitungen

Alle Komponenten entsprechen den Vorgaben aus dem nichtparametrischen Modell.

Restringiertes Semiparametrisches Modell

Da keine zusätzlichen Restriktionen eingeführt wurden, ändert sich bis auf den zusätzlichen Term $X\beta y^{-1}$ in der Hatmatrix H_γ und eine zusätzliche Ableitung in der Fishermatrix nichts. Es gilt daher

$$\left(\frac{\partial \eta}{\partial \gamma}\right)' W_{\gamma,\beta}^{-1} \left(\frac{\partial \eta}{\partial \beta}\right) = \Phi_2' W X. \tag{15.4}$$

Zwar gehen noch andere Ableitungen ein, diese sind jedoch in

$$\begin{aligned}
\left(\frac{\partial \eta}{\partial \delta}\right)' W_\delta^{-1} \left(\frac{\partial \eta}{\partial \delta}\right) &= \Phi_1' W_\delta^{-1}\Phi_1 + K_\delta, \qquad (15.5) \\
\left(\frac{\partial \eta}{\partial \gamma}\right)' W_\gamma^{-1} \left(\frac{\partial \eta}{\partial \gamma}\right) &= \Phi_2' W_\gamma^{-1}\Phi_2 + K_\gamma
\end{aligned}$$

integriert. Auch die drei Ableitungen W^{-1}_{γ,β_0}, $W^{-1}_{\gamma,\alpha}$ und W^{-1}_{α,β_0} gehen unverändert in das semiparametrische Modell ein.

SUR-Modell (nichtparametrisch)

Designmatrizen

Der lineare Prädiktor hat die Form

$$\eta_{is} = V_{is}\beta_{0s} + \Gamma_s Z_i \alpha_s$$

mit

$$\begin{aligned}
&(V_{is})_{T\times T} = I, && (\beta_{0s})_{T\times 1} = (\beta_{01s},\ldots,\beta_{0Ts})',\\
&(\Gamma_s)_{T\times T} = \mathrm{diag}(\gamma_{1s},\ldots,\gamma_{Ts}), && (Z_i)_{T\times r} = (z_{i1},\ldots,z_{iT})',\\
&(z_{it})_{1\times r} = (\tilde{z}_{it1},\ldots,\tilde{z}_{itr}), && (\tilde{z}_{itl})_{1\times 1} = \Phi^{\alpha}_l(z_{itl}),\\
&(\alpha_s)_{r\times 1} = (\alpha_{1s},\ldots,\alpha_{rs})', && (\gamma_s)_{T\times 1} = (\gamma_{1s}\ldots,\gamma_{Ts})',\\
&(\delta_s)_{(T+r)\times 1} = (\beta_{0s}',\alpha_s')', && (\gamma)_{2T\times 1} = (\gamma_1',\gamma_2')'.
\end{aligned}$$

Penalty-Matrizen

Bei $d = 2$ enthält das Modell sechs Glättungsparameter, je drei pro Submodell. Die Glättungsparameter erhalten deswegen einen zusätzlichen Index $d = 1, 2$, der für das jeweiligen Submodell steht. Für die Penaltymatrizen gilt

$$\begin{aligned}
K_\delta &= \mathrm{diag}(K_{\delta_1}, K_{\delta_2})_{2(T+r)\times 2(T+r)},\\
K_{\delta_s} &= \mathrm{diag}(\lambda_{\beta_{0s}} K_{\beta_{0s}}, \lambda_{\alpha_s} K_{\alpha_s})_{(T+r)\times(T+r)},\\
K_\gamma &= \mathrm{diag}(K_{\gamma_1}, K_{\gamma_2})_{2T\times 2T},\\
K_{\gamma_s} &= (\lambda_{\gamma_s} K_{\gamma_s})_{T\times T}.
\end{aligned}$$

Hat-Matrizen

Beide Hatmatrizen entsprechen denjenigen aus dem nichtparametrischen Modell. Allerdings sind die zugehörigen Designmatrizen und auch die Working-Kovarianz-

Matrix stark verändert. Zusammengesetzt ergibt sich $H = \text{diag}(H_\delta, H_\gamma)$. Die Anzahl an Freiheitsgraden ergibt sich aus der Spur von H.

Restringiertes SUR-Modell

Keine zusätzlichen Restriktionen müssen eingeführt werden. Jedes der Modelle enthält somit zwei Restriktionen, die sich gegenseitig bei der Schätzung nicht beeinflussen. Allerdings muss beim Zusammensetzen der Working-Kovarianz-Matrizen genau auf die Zusammensetzung und auf die Position der einzelnen Elemente in der Matrix geachtet werden. Die Ableitungen der zweiten Ableitungen erweisen sich aufgrund der SUR-Struktur als deutlich komplizierter.

Simulations-Studie

Additives Modell - Nichtparametrisch

Konfigurationen:

- AD (konstante multiplikative Parameter)
- M1 (linear ansteigende multiplikative Parameter)
- M2 (exponentiell fallende multiplikative Parameter)
- M3 (multiplikative Parameter mit Richtungswechsel)

Kovarianz Struktur:

- I (Independence)
- AR (Autoregressive Correlation)
- CS (Compound Symmetry)

Working-Kovarianz Struktur:

- IW (Independence WC-Struktur)
- ARW (Autoregressive WC-Struktur)
- CSW (Compound-Symmetry-WC-Struktur)

SUR-Struktur:

- IV (Independence SUR-Struktur)
- IWV (Working-Independence SUR-Struktur)

Varianz:

- LV (Low Variance – niedrige Varianz, $\sigma^2 = 0.1$)
- HV (High Variance – hohe Varianz, $\sigma^2 = 0.5$)

	Additives Modell (LV) ungeglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	2.59	4.42	9.78	4.39	9.81
M1	53.31	55.15	60.42	55.15	60.50
M2	45.11	46.90	52.26	46.90	52.29
	Additives Modell (LV) geglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	1.31	2.07	5.22	2.45	5.36
M1	52.58	53.64	56.94	53.74	56.99
M2	43.63	44.40	46.37	44.41	46.36
	Additives Modell (HV) ungeglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
AD	12.95	22.11	27.29	21.93	27.40
M1	63.64	72.74	77.94	72.84	78.08
M2	53.44	64.46	69.78	64.51	69.86
	Additives Modell (HV) geglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
AD	4.79	9.08	11.45	9.07	11.46
M1	56.44	61.14	63.87	61.90	63.89
M2	46.11	49.02	50.49	49.12	50.58

Tabelle 15.1: Zusammenfassung der Simulations-Studie für das Additive Modell (A) für niedrige und hohe Varianzen, ungeglättet und geglättet (Modelle 1 bis 3 entsprechen den Modellen in der Ergebnis-Sektion).

Penalty-Parameter

	Glättungs-Parameter (LV)				
λ_α	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	5	5	10	5	10
M1	20	5	10	5	10
M2	5	30	40	25	50
	Glättungs-Parameter (HV)				
λ_α	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
AD	10	15	20	15	20
M1	5	10	10	10	15
M2	40	60	80	50	80

Tabelle 15.2: Penalty-Parameter für die additiven Simulations-Studien für niedrige und hohe Varianzen (Modelle 1 bis 3 stimmen mit dem Modellen aus der Ergebnis-Sektion überein).

Multiplikatives Modell - Nichtparametrisch

	Multiplikative Modell (LV) ungeglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	3.48	5.20	10.94	5.18	11.25
M1	3.52	5.16	9.77	5.20	10.02
M2	3.56	4.43	4.61	4.61	7.42
	Multiplikatives Modell (LV) geglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	1.61	2.91	5.99	2.99	6.21
M1	2.15	3.26	6.19	3.25	6.50
M2	1.85	2.39	2.86	2.58	3.92
	Multiplikatives Modell (HV) ungeglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
AD	17.57	26.26	32.49	26.36	33.06
M1	17.55	25.91	31.26	25.90	31.56
M2	18.66	23.18	22.84	24.44	26.63
	Multiplikatives Modell (HV) geglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
A	6.78	12.25	15.29	12.42	14.68
M1	7.69	16.36	18.99	13.35	15.27
M2	7.69	12.25	10.85	11.55	12.05

Tabelle 15.3: Zusammenfassung der Simulations-Studie für das Multiplikative Modell (M) für niedrige und hohe Varianzen, ungeglättet und geglättet (Modelle 1 bis 3 entsprechen den Modellen in der Ergebnis-Sektion).

Penalty-Parameter

	Glättungs-Parameters (LV)				
$\lambda_\alpha, \lambda_\gamma$	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	20, 350	20, 200	30, 15	25, 300	40, 500
M1	10, 50	20, 100	25, 20	30, 300	50, 375
M2	10, 50	15, 25	15, 25	15, 50	20, 25
	Glättungs-Parameter (HV)				
$\lambda_\alpha, \lambda_\gamma$	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
AD	10, 75	10, 40	10, 20	15, 75	20, 75
M1	10, 50	10, 0	10, 0	25, 50	30, 60
M2	5, 50	10, 75	5, 25	10, 25	10, 25

Tabelle 15.4: Penalty-Parameter für die multiplikativen Simulations-Studien für niedrige und hohe Varianzen (Modelle 1 bis 3 stimmen mit dem Modellen aus der Ergebnis-Sektion überein).

Multiplikatives Modell - Semiparametrisch

	MSE Multiplikative Modell (LV) ungeglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	3.95	6.16	13.77	6.18	13.02
M1	4.02	6.30	12.36	6.29	12.77
M2	3.96	5.07	5.69	5.34	9.53
	MSE Multiplikatives Modell (LV) geglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	2.91	4.75	8.86	5.05	9.12
M1	3.17	4.96	8.76	5.40	9.82
M2	2.99	3.86	4.47	4.01	6.08
	MSE Multiplikatives Modell (HV) ungeglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
AD	20.02	28.94	39.52	31.47	38.10
M1	20.07	31.26	38.50	31.44	37.60
M2	20.95	26.81	27.07	28.83	32.82
	MSE Multiplikatives Modell (HV) geglättet				
	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
A	11.43	22.56	21.38	24.37	23.13
M1	13.31	21.79	23.47	23.82	26.88
M2	12.21	17.54	16.75	19.12	20.00

Tabelle 15.5: Zusammenfassung der semiparametrischen Simulations-Studie für das Multiplikative Modell (M) für niedrige und hohe Varianzen, ungeglättet und geglättet.

Penalty-Parameter

	Glättungs-Parameters (LV)				
$\lambda_\beta, \lambda_\alpha, \lambda_\gamma$	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I1, IW)	(AR1, ARW)	(CS1, CSW)	(AR1, IW)	(CS1, IW)
AD	5,250,100	7.5,150,100	5,300,50	5,250,50	10,350,100
M1	5,200,150	10,150,100	10,100,10	5,150,100	5,150,50
M2	5,20,100	5,2.5,20	7,5,50	5,5,100	5,1,100
	Glättungs-Parameter (HV)				
$\lambda_\beta, \lambda_\alpha, \lambda_\gamma$	Modell 1	Modell 2	Modell 3	Modell 4	Modell 5
	(I2, IW)	(AR2, ARW)	(CS2, CSW)	(AR2, IW)	(CS2, IW)
AD	2,250,150	2,2,10	3,250,50	2,10,75	10,350,100
M1	2,150,150	3,50,50	5,150,50	5,100,50	5,100,50
M2	2,20,100	2,10,50	3,10,100	2,10,100	5,1,100

Tabelle 15.6: Penalty-Parameter für die multiplikativen semiparametrischen Simulations-Studien für niedrige und hohe Varianzen.

SUR-Modell

Es konnten aufgrund der hohen Rechenanforderung nicht alle Kombinationen simuliert werden. Alle Ergebnisse wurden in Sektion 13.2 präsentiert.

Kapitel 16

Literaturverzeichnis

AERTS, M., GEYS, H., MOLENBERGHS, G. & RYAN L.M. (2002). Topics in Modeling of Clustered Data. Chapman and Hall, CRC Press, Boca Raton.

ANDREW, R., PHILLIPS, D.I. & WALKER, B.R. (1998). Obesity gender influence cortisol secretion and metabolism in man. Journal of clinical endocrinal metabolism 83, 1806-1809.

BOX, G.E.P. (1988). Signal-to-noise ratios, performance criteria, and transformations (with discussions). Technometrics, 30(1), 1-40.

BARTLETT, M.S. (1963). Statistical estimation of density functions. Sankhya Series A, 25, 245-254.

BREIMAN, L. & FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, 580-598.

BRESLOW, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other Quasi-likelihood models. Journal of the American Statistical Association, 85, 410, 565-571.

BUJA, A., HASTIE, T. & TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). Annals of Statistics, 17, 453-555.

CARROLL R.J. (2003). Semiparametric seemingly unrelated regression model, unpublished manuscript.

CHAGANTY, N.R. (2004). Efficiency of generalized estimating equations for binary responses. Journal of the Royal Statistical Society, B 66, 851-860.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74, 368, 829-836.

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik, 31, 377-403.

DAVIDIAN, M. & GILTINAN, D. (1995). Nonlinear Models for Repeated Measurement Data. Chapman and Hall, London.

DE BOOR, C. (1977). Package for calculating with B-Splines. SIAM Journal of scientific statistical computation 14, 441-472.

DE BOOR, C. (1978). A practical guide to Splines, Springer, Berlin.

DIGGLE, P.J., LIANG, K.Y. & ZEGER, S.L. (1994). The analysis of Longitudinal Data. Oxford University Press, England.

DRESSENDORFER, R.A., KIRSCHBAUM, C., ROHDE W. et al. (1992). Synthesis of a cortisol-biotin conjugate and evaluation as a tracer in an immunoassay for salivary cortisol measurement. Journal of Steroid Biochemistry and Molecular Biology 43, 683-692.

EGGE, N. & KROWNE, A. (2000). Internet-Plattform www.planetmath.org.

EILERS, P.H.C. & MARX, B.D. (1992). Generalized linear models with P-splines. Proceedings of GLIM 92 and 7th International Workshop on Statistical Modeling, Munich, Germany. Lecture Notes in Statistics, 78, Advances in GLIM and Statistical Modeling, Eds. L. Fahrmeir, B. Francis, R. Gilchrist, G. Tutz. Springer, New York, 72-77.

EILERS, P.H.C. & MARX, B.D. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). Statistical Science 11(2), 89-121.

EILERS, P.H.C. & MARX, B.D. (1998). Direct generalized additive modeling with penalized likelihood. Computational Statistics and Data Analysis 28(2), 193-209.

EPANECHNIKOV, V.A. (1969). Nonparametric estimates of a multivariate probability density. Theory of Probability and its Application, 14, 153-158.

EUBANK, R. (1999). Spline smoothing and nonparametric regression. Marcel Dekker, New York.

FAHRMEIR, L. & KAUFMANN, H. (1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Model. Annals of Statistics, 13, 342-368. Correction, Annals of Statistics, 14, 1643.

FAHRMEIR, L. & TUTZ, G. (2001). Multivariate Statistical Modeling Based on Generalized Linear Models. Springer, New York.

FINNEY D.J. (1952). Probit analysis. A statistical treatment of the sigmoid response curve. Cambridge University Press.

FISCHER J.E., ANOUK, C., DETTLING, A., ZEIER, H. & FANCONI S. (2000). Experience and endocrine stress responses in neonatal and pediatric critical care nurses and physicians. Critical Care Medicine 28(9), 3281-3288.

FITZMAURICE, G.M., LAIRD, N.M. & ROTNITZKY, A.G. (1993). Regression models for discrete longitudinal responses. Statistical Science, 8, 3, 284-309.

GILMOUR, A.R., ANDERSON, R.D. & RAE, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. Biometrika, 72, 593-599.

GREEN, P.J. & SILVERMAN, B.W. (1994). Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

GREEN, P. & YANDELL, B. (1985). Semi-parametric generalized linear models. Proceedings 2nd International GLIM Conference, Lancaster, Lecture notes in Statistics No. 32, 44-55, Springer, New York.

GONG, G. & SAMANIEGO, F.J. (1981). Pseudo maximum likelihood estimation theory and applications. Annals of Statistics, 9, 4, 861-869.

HÄRDLE, W. (1990). Applied Nonparametric Regression, Econometric Society Monographs No. 19, Cambridge University Press.

HASTIE, T. & TIBSHIRANI, R. (1986). Generalized additive models (with discussion). Statistical Science 1, 2, 297-318.

HASTIE, T. & TIBSHIRANI, R. (1990). Generalized additive models. Chapman and Hall, London.

HASTIE, T. & TIBSHIRANI, R. (1993). Varying-coefficient models (with discussion). Journal of the Royal Statistical Society, B 55, 757-796.

JORGENSEN, B. (1993). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. Biometrika, 70, 1, 19-28.

JORGENSEN, B. (1997). The Theory of Dispersion Models. Chapman and Hall, London.

KAKWANI, N.C. (1967). The unbiasedness of Zellner's SUR Estimators, Journal of the American Statistical Association, 82, 141-142.

KENDAL, E. (2000). Principles of neural science. McGraw-Hill, New York.

KIRSCHBAUM, C. & HELLHAMMER D.H. (1989). Salivary cortisol in psychobiological research. Neuropsychobiology 22, 150-169.

KIRSCHBAUM, C. & HELLHAMMER D.H. (1994). Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. Psychoneuroendocrinology 19, 313-333.

KOTZ, S. & JOHNSON, N. (1985). Encyclopedia of statistical sciences, Volume 5. Wiley, New York.

LIANG, K.Y. & ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.

LIANG, K.Y., ZEGER, S.L. & QAQISH, B. (1992). Multivariate Regression analyses for categorical data. Journal of the Royal Statistical Society, B 54, 1, 3-40.

LIN, X. & CARROLL, R. (2000). Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error. Journal of the American Statistical Association, 95, 450, 520-534.

LIN, X. & CARROLL, R. (2001a). Semiparametric regression for clustered data using generalized estimation equations. Journal of the American statistical association 96, 1045-1056.

LIN, X. & CARROLL, R. (2001b). Semiparametric regression for clustered data. Biometrika 88, 1179-1185.

LIN X., WANG, N., WELSH., A.H. & CARROLL, R. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. Biometrika 91, 177-193.

LINDSEY, J.K. & LAMBERT, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. Statistics in Medicine, 17, 447-469.

LJUNG, T., HOLM, G., FRIBERG, P., ANDERSSON, B., BENGTSSON, B.A., SVENSSON, J., DALLMAN, M., MCEWEN, B. & BJORNTORP P.

(2000). The activity of the hypothalamic-pituitary-adrenal axis and the sympathetic nervous system in relation to waist/hip circumference ratio in men. Obesity Research 8, 498-495.

MARIN, P., DARIN, N., AMEMIYA, T., ANDERSSEN, B., JERN, S. & BJORNTORP, P. (1992). Cortisol secretion in relation to body fat distribution in obese premenopausal women. Metabolism 41, 882-886.

MARNIEMI, J., KRONHOLM, E., AUNOLA, S., TOIKKA, T., MATTLAR, C.E., KOSKENVUO, M. & RONNEMAA, T. (2002). Visceral fat and psychosocial stress in identical twins discordant for obesity. Journal of internal medicine 251, 35-43.

MATHWORKS INC. (1994-2005). Matlab, Natick, USA.

MCCULLAGH, P. (1983). Quasi-likelihood functions. Annals of Statistics, 11, 1, 59-67.

MCCULLAGH, P. & NELDER, J.A. (1983, 1989). Generalized Linear Models. Chapman and Hall, London.

NADARAYA, E.A. (1964). On Estimating Regression. Theory of Probability Applied, 10, 186-190.

NELDER, J.A. (2000). Quasi-likelihood and Pseudo-likelihood are not the same thing. Journal of Applied Statistics, 27, 8, 1007-1011.

NELDER, J.A. & LEE, Y. (1992). Likelihood, Quasi-likelihood and Pseudo-likelihood: some comparisons. Journal of the Royal Statistical Society, B 54, 1, 273-284.

NELDER, J.A. & PREGIBON, D. (1987). An extended quasi-likelihood function. Biometrika, 74, 2, 221-232.

NELDER, J.A. & WEDDERBURN, R.W.M. (1972). Generalized linear models. Journal of the Royal Statistical Society A 135(3), 370-384.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). Statistical science 1, 505-527.

O'SULLIVAN, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. SIAM Journal of scientific statistical computation 9, 363-379.

PARZEN, E. (1962). On estimating of a probability density and mode. Annals of Mathematical Statistics, 35, 1065-1076.

PEPE, M.S. & ANDERSON, G.L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics, 23, 939-951.

POLLOCK, D.S.G. (1999). Time-Series Analysis Signal Processing and Dynamics. Academic Press Ltd., London.

PREGIBON, D. (1984). Data Analytic Methods for Matched Case-Control Studies. Biometrics, 40, 639-651.

PRENTICE, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. Journal of the American Statistical Association, 81, 394, 321-327.

PRENTICE, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. Biometrics, 44, 1033-1048.

PRENTICE, R.L. & ZHAO, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics, 47, 825-839.

REINSCH, C. (1967). Smoothing by spline functions. Numerische Mathematik 10, 177-183.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics, 27, 832-837.

RUPPERT, D. & CARROLL, R. (1997). Penalized Regression Splines, unpublished manuscript, http://www.orie.cornell.edu/∼davidr/papers/ .

RUPPERT, D. & CARROLL, R. (1999). Spatially-adaptive penalties for spline fitting. Australian and New Zealand Journal of Statistics 42(2), 205-223.

RUPPERT, D., WAND, M.P. & CARROLL, R. (2003). Semiparametric Regression. Cambridge University Press.

SCHWARZ, G. (1978). Estimating the dimension of a model. Annals of statistics, 6(2), 461-464.

SCHMIDT, R.F., THEWS, G. (1995). Physiologie des Menschen. Springer, Berlin.

SEVERINI, T.A. & STANISWALIS, J.G. (1994). Quasi-Likelihood Estimation in Semiparametric Models. American Statistical Association, 89, 426, 501-511.

SEVERINI, T.A. & WONG, W.H. (1992). Profile likelihood and conditionally parametric models. Annals of Statistics, 20, 4, 1768-1802.

SPECKMAN, P. (1988). Kernel Smoothing in Partial Linear Models. Journal of the Royal Statistical Society, B 50, 3, 413-436.

STEPTOE, A., KUNZ-EBRECHT, S.R., BRYDON, L. & WARDLE, J. (2004). Central adiposity and cortisol responses to waking in middle-aged men and women. International Journal of Obesity 28, 1168-1173.

STOER, J. (1999). Numerische Mathematik 1 (8. Auflage). Springer, Berlin.

TOUTENBURG, H. (1992). Lineare Modelle. Physica Verlag, Heidelberg.

TOUTENBURG, H. (1994). Versuchsplanung und Modellwahl. Physica-Verlag, Heidelberg.

TUTZ, G. (2003). Generalized semiparametrically structured ordinal models. Biometrics 59, 263-273.

TUTZ, G. & MEINEL, M. (2002). Working-Paper: Nichtparametrisches Modell mit Zeit-variierenden multiplikativen Effekten.

WALES, J. (2001). Internet-Plattform www.wikipedia.org, Wikimedia Foundation Inc., St. Petersburg, USA.

WANG, N. (2003). Marginal Nonparametric kernel regression accounting for within-subject correlation, Biometrika, 90, 1, 43-52.

WANG, N., CARROLL, R.J. & LIN, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. Journal of the American Statistical Association, 100, 147-157.

WATSON, G.S. (1964). Smooth Regression Analysis. Sankhya, Series A, 26, 359-372.

WEDDERBURN, R.M.W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, 61, 3, 439-447.

WEDDERBURN, R.M.W. (1976). On the existence and uniqueness of the maximum likelihood estimates. Biometrika, 63, 27-32.

WELSH, A.H., LIN, X. & CARROLL, R.J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. Journal of the American Statistical Association, 97, 482-493.

ZEGER, S.L. & DIGGLE, P.J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics, 50, 689-699.

ZEGER, S.L. & LIANG, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42, 121-130.

ZEGER, S.L. & LIANG, K.Y. & ALBERT, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics, 44, 1049-1060.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Society 57, 348-368.

ZELLNER, A. (1963). Estimators for seemingly unrelated regression equations: Some exact finite sample results. Journal of the American Statistical Society 58, 977-992.

ZHAO, L.P. & PRENTICE, R.L. (1990). Correlated binary regression using a quadratic exponential model. Biometrika, 77, 642-648.

ZHAO, L.P., PRENTICE, R.L. & SELF, S.G. (1992). Multivariate mean parameter estimation using a partly exponential model. Journal of the Royal Statistical Society, B 54, 805-811.

Eidesstattliche Versicherung

Hiermit versichere ich eidesstattlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe. Die aus fremden Quellen wörtlich oder nahezu wörtlich übernommenen Inhalte sowie mir gegebene Anregungen sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch in ganzen Teilen noch nicht veröffentlicht.

München, 29.09.2005