Die approbierte Originalversion dieser Diplom-/Masterarbeit ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this diploma or master thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

DIPLOMARBEIT

Master Thesis

Die Entwicklung einer Softwareumgebung zur Strukturoptimierung

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Diplom-Ingenieurs

unter der Leitung von

Univ.Prof. Dipl.-Ing. Dr.techn. *Christian Bucher*
E206, Institut für Hochbau und Technologie
Forschungsbereich für BAUMECHANIK und BAUDYNAMIK

eingereicht an der Technischen Universität Wien
Fakultät für Bauingenieurwesen

von

Christopher Riemel
0425187
Karl-Tornay-G. 45-47 6/8
1230 Wien

Wien im März 2011

Deutsch Ziel dieser Arbeit war es, für eine gegebene diskretisierte Struktur einer Finite-Elemente Berechnung, eine zeiteffiziente und speicherunkritische Optimierungsstrategie zu entwickeln. Es sollte gezeigt werden, dass dies durch eine Schätzung der rechenaufwendigen Zielfunktion mit Hilfe einer leichter optimierbaren, klassischen, linearen Regressionsfläche möglich ist. Zu diesem Zweck wurden die theoretischen Grundlagen zur schrankenrestringierten Optimierung, zur Parameterschätzung und zur Diagnose von linearen Regressionsmodellen erläutert. Schließlich wurden diese Grundlagen in einer Softwareumgebung umgesetzt und die Arbeitshypothese mit Hilfe eines einfachen Falltestbeispiels getestet. Dabei stellte sich leider heraus, dass klassische lineare Regressionsmodelle für solche komplexen Aufgabenstellungen ungeeignet sind.

Englisch The object of this thesis was to develop a time- and memory-efficient optimization strategy for a discretized structure of a finite element analysis. It should be shown that this can be done by estimating the computational expensive objective function with a more easily optimizable classical linear regression surface. To this end, the theoretical basis for bounded optimization, regressor parameter estimation and diagnosis of linear regression models were explained. Finally, these principles were implemented in a software environment and the working hypothesis has been tested using a simple drop test case example. Unfortunately it turned out that classical linear regression models are inappropriate for such complex tasks.

Typografische Konventionen

Symbole - allgemein

$\kappa(\cdot)$	Konditionszahl
$\|\cdot\|$	Euklidische Norm
$\mathrm{rg}(\boldsymbol{X})$	Rang der Matrix $\boldsymbol{X}$
$\mathbb{R}^n$	n-dimensionaler Raum der reellen Zahlen
x	skalare Größe
$\boldsymbol{x}$	Spaltenvektor
x_i	i-tes Element des Vektors $\boldsymbol{x}$
$\boldsymbol{X}$	Matrix
x_{ij}	Matrixelement der i-ten Zeile und der j-ten Spalte
$\boldsymbol{X}'$	Transponierte Matrix
$\boldsymbol{x}'$	Zeilenvektor

Symbole - Optimierung

δ_{ij}	Kroneckersymbol
$\nabla f(\boldsymbol{x})$	Gradient des Vektors $\boldsymbol{x}$
$\nabla^2 f(\boldsymbol{x})$	Hessematrix des Vektors $\boldsymbol{x}$
$\boldsymbol{H}(\boldsymbol{x})$	Hessematrix des Vektors $\boldsymbol{x}$

Symbole - Lineare Regression

α	Signifikanzniveau
$\boldsymbol{\beta}$	p-dimensionaler reeller Regressorkoeffizientenvektor
$\mathrm{Cov}(X, Y)$	Kovarianz der stochastischen Größen X und Y
$\mathbb{E}Y$	Erwartungswert der stochastischen Größe Y
$F_{m,n}$	F-Verteilung mit m und n Freiheitsgraden
$\Phi(0, \sigma^2)$	Standardnormalverteilung mit Erwartungswert Null und Varianz σ^2
p	Anzahl der Regressorkoeffizienten β_i
n	Anzahl der Beobachtungen einer stochastischen Größe Y
t_n	Studenten-Verteilung
$t_{n,p}$	p-Fraktil der Studentenverteilung
$\bar{x}$	arithmetisches Mittel
$\mathrm{Var}(Y)$	Varianz der stochastischen Größe Y
Y	stochastische Größe
$\hat{Y}$	Schätzwert für Y
$\sim$	verteilt nach

Akronyme

ANOVA	Analysis of variance
LM	Lineares Regressionsmodell
MSE	Mean square error
REML	Restringierter Maximum-Likelihood-Schätzer
SSE	Sum of squared errors
SSPE	Sum of squared prediction error
SSR	Sum of squared residuals
SST	Sum of squared total

Vorwort

Zugegeben, um der Entwicklung dieser Arbeit größten Freiraum zu lassen, wurde das Thema sehr allgemein gewählt. Allerdings verspricht es deshalb vielleicht zu viel. An dieser Stelle soll daher zunächst eine Themenabgrenzung vorgenommen werden.

Ziel dieser Arbeit sollte es sein, für eine gegebene diskretisierte Struktur einer Finite-Elemente Berechnung, eine vor allem *zeiteffiziente* und *speicherunkritische* Optimierungsstrategie zu entwickeln. Da Optimierungsalgorithmen grundsätzlich iterativ arbeiten, ist im Zuge einer Optimierung eine mehrfache Auswertung der Zielfunktion erforderlich. Im Falle einer Finite-Elemente Formulierung für ein kinetisches System, entspricht die Zielfunktion dem folgenden Differentialgleichungssystem (Bathe, 2002, S.816)[1]

$$\boldsymbol{M\ddot{U}} + \boldsymbol{C\dot{U}} + \boldsymbol{KU} = \boldsymbol{R} \tag{0.1}$$

Bei einer Finite-Elemente Formulierung stellt aber gerade die Lösung dieser Bewegungsgleichungen den, im Hinblick auf das zuvor genannte Ziel, kritischen Schritt dar. Aus diesem Grund wurden von einer möglichen Lösungsstrategie zwei grundsätzliche Forderungen abverlangt:

- eine Minimierung der erforderlichen Anzahl an Lösungen der Bewegungsgleichungen, und
- die Lösung dieser Gleichungssysteme sollte parallel erfolgen können.

In dieser Arbeit wurde versucht, beide Forderungen mit Hilfe einer, der Optimierung vorgelagerten *Regressionsanalyse* zu erfüllen.

Softwareumgebung Die gewählte Strategie führt im wesentlichen zu einem dreistufigen Arbeitsablauf, der sich auch in der Kapitelgliederung der gegenständlichen Arbeit wiederfindet:

1. Die Generierung einer im gegebenen Parameterraum möglichst gleichmäßig verteilten Menge an Eingabedaten für die Zielfunktion in (0.1) und die Berechnung der Zielgröße.
2. Basierend auf diesen Eingabedaten und den daraus resultierenden Zielgrößen, die Schätzung einer Regressionsfläche, und schließlich
3. das Auffinden eines lokalen (globalen) Minimums auf dieser Regressionsfläche.

[1] Eine sehr umfassende Einführung in die Finite-Elemente Methode findet sich ebenda.

Anders als es der Arbeitstitel vielleicht vermuten lassen würde, werden sich in den jeweiligen Kapiteln aber keine Implementierungsdetails zu den Arbeitsabläufen finden, sondern ausschließlich theoretische Erläuterungen. Dies deswegen, da eine solche Vorgehensweise einerseits den Rahmen der Arbeit sprengen würde und andererseits nicht notwendig ist, da es für jeden Arbeitsablauf bereits zahlreiche, bewährte und kostenlose open source Implementierungen gibt. Durch deren entsprechende Kombination lässt sich eine Softwareumgebung zur Strukturoptimierung bauen. Anhand der für die gegenständliche Arbeit getroffenen Wahl, soll dies nun beispielhaft skizziert werden.
Unter dem Betriebssystem Linux kamen im Wesentlichen folgende Softwareprodukte zum Einsatz:

MATLAB® Die Softwareumgebung MATLAB® beinhaltet bereits zahlreiche Algorithmen zur Regressionsanaylse und zur Optimierung, und bietet auch die Möglichkeit, innerhalb ihrer Umgebung systemspezifische Befehle ausführen zu lassen. Da sie darüber hinaus als kostengünstige Studentensoftware beim Zentralen Informatikdienst (ZID) der TU-Wien verfügbar ist und der Verfasser dieser Arbeit mit der Verwendung des Produkts auch schon ein wenig vertraut war, wurden innerhalb der MATLAB®-Umgebung in der Version R2009a sämtliche Arbeitsschritte eingebettet.
Eine kostenlose open source Alternative zu MATLAB® mit einer sehr ähnlichen Syntax bietet GNU octave[2]. Da zum Zeitpunkt der Erstellung der gegenständlichen Arbeit GNU octave aber leider kaum Algorithmen zur Regressionsanalyse beinhaltete, wäre bei dieser Wahl der Implementierungsaufwand größer gewesen.
Darüber hinaus sei noch das ebenso kostenlose wie umfangreiche daktoa toolkit[3] erwähnt.

speedyne Da die gegenständliche Arbeit in Kooperation mit der Dynardo GmbH[4] verfasst wurde, konnte die Lösung der Zielfunktion in (0.1) mit Hilfe des dort entwickelten Hybridsolvers speedyne in Version 1.0.1 vorgenommen werden. Dieser Gleichungslöser unterscheidet sich insofern von anderen Finite-Elemente Lösern, als dass die Zeitintegration mit Hilfe einer *modalen Reduktion* (Bucher, 2001) deutlich beschleunigt werden kann und somit dem Effizienzprinzip dieser Arbeit entspricht.

bash Zur Kommunikation zwischen den beiden oben genannten Softwareprodukten wurde die GNU Bourne-Again-SHell (bash)[5] in Version 3.2.39 verwendet. Weiters wurde mit Hilfe

[2] http://www.gnu.org/software/octave/
[3] http://dakota.sandia.gov/
[4] http://www.dynardo.at/
[5] http://www.gnu.org/software/bash/

dieser Skriptsprache eine parallelisierte Jobverwaltung für speedyne realisiert, sodass für mehrere unterschiedliche Eingabedaten gleichzeitig die Zielgröße in (0.1) bestimmt werden konnte.

Die Kombination dieser Softwareprodukte wurde entsprechend dem bereits beschriebenen Arbeitsablauf vorgenommen. Die wesentlichen der dabei verwendeten MATLAB®-Funk-tionen, wie `haltonset`, `regstats`, `fmincon`, u. a. werden an geeigneten Stellen dieser Arbeit noch näher erläutert werden.
Sowohl zur MATLAB®-Umgebung wie auch zur speedyne-Software werden umfangreiche und detaillierte Handbücher und Dokumentationen mitgeliefert. Zwar bietet auch die bash eine gute Hilfefunktion zum raschen Nachschlagen zur Verfügung gestellter Befehle, als gute Einstiegsliteratur seien hier aber beispielhaft die Bücher Albing et al. (2007) und Wolf (2010) erwähnt.

Inhaltsverzeichnis

Typografische Konventionen V
Vorwort VII

1 Optimierung **1**
1.1 Einführung in die Optimierung 2
1.2 Grundlagen der linearen, unrestringierten Optimierung 4
1.3 Das Newton-Verfahren 7
1.4 Methode der kleinsten Fehlerquadratsumme 12

2 Lineare Regression **15**
2.1 Einführung in die Regression 17
2.2 Das Lineare Regressionsmodell 19
2.3 Parameterschätzung 21
2.4 Statistische Eigenschaften der Schätzer 24
2.5 Modellauswahl 29
2.5.1 Einige Modellwahlkriterien 31
2.5.2 Selektionsverfahren 33
2.6 Modelldiagnostik 38
2.6.1 Residuenanalyse 38
2.6.2 Ausreißeranalyse 48
2.6.3 Kollinearitätsanalyse 50
2.6.4 Maßnahmen 52

3 Anwendung und Resümee **57**

1 Optimierung

Vorbemerkungen

In diesem Kapitel wird unter anderem die Kenntnis folgender Begriffe vorausgesetzt

Gradient Der Gradient einer mindestens einfach partiell differenzierbaren Funktion $f(\boldsymbol{x})$ im Punkt $\boldsymbol{x} \in \mathbb{R}^n$ ist definiert als

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)'$$

mit ∇ als sogenannter *Nabla*-Operator.

Hessematrix Die Hessematrix ist insbesondere zur Bestimmung der Art eines stationären Punktes von Bedeutung und ist als quadrierte Nablaoperation definiert. Für eine mindestens zweifach partiell differenzierbare Funktion $f(\boldsymbol{x})$ im Punkt $\boldsymbol{x} \in \mathbb{R}^n$ gilt

$$\boldsymbol{H} = \nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

Konditionszahl Die Konditionszahl ist ein Maß für die Stabilität eines numerischen Problems. Sie ist unabhängig vom verwendeten Lösungsalgorithmus. Für eine reguläre Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ und eine beliebige *Matrixnorm* $\|\cdot\| : \mathbb{R}^{n \times n} \to \mathbb{R}_+$ ist die *Konditionszahl* $\kappa(\boldsymbol{A})$ wie folgt definiert

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\| \|\boldsymbol{A}\|^{-1}$$

Euklidische Norm Unter einer *Norm* versteht man jede beliebige Abbildung $\|\cdot\|$ eines beliebigen Vektorraums über $\mathbb{R}^n$ auf eine nichtnegative reelle Zahl, $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$, die gewisse Bedingungen erfüllt. In diesem Kapitel wird vordergründig von der *Euklidischen*

Vektornorm Gebrauch gemacht, die für einen Vektor $\boldsymbol{x} \in \mathbb{R}^n$ wie folgt definiert ist

$$\|\boldsymbol{x}\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} = (\boldsymbol{x}'\boldsymbol{x})^{\frac{1}{2}}$$

Neben den Vektornormen mit $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$ definiert man auch *Matrixnormen* mit $\|\cdot\| : \mathbb{R}^{n \times n} \to \mathbb{R}_+$.

Eine sehr gute Einführung in die deterministische und auf lokale Lösungen beschränkte Optimierung findet sich in Nocedal und Wright (2006). Das Buch wird in diesem Kapitel intensiv zitiert werden. Zum Nachschlagen der ein oder anderen darin zu kurz gekommenen Erklärung numerischer Probleme und Definitionen, empfiehlt sich z. B. Plato (2010).

1.1 Einführung in die Optimierung

In der Optimierung versucht man für ein gegebenes Problem jene Lösung zu finden, die im Rahmen bestimmter Optimalitätsbedingungen und eventueller Restringierungen als *optimal* bezeichnet werden kann.
Solch ein Optimierungsproblem lässt sich im Zuge der *mathematischen Modellierung* wie folgt formulieren (Nocedal und Wright, 2006, S.3)

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathscr{H}_{ad}} f(\boldsymbol{x}) \quad \text{mit} \quad \mathscr{H}_{ad} = \{ \boldsymbol{x} \in \mathbb{R}^n \mid \; & c_i(\boldsymbol{x}) = 0, \; i \in \mathscr{E} \\ & c_i(\boldsymbol{x}) \leq 0, \; i \in \mathscr{I} \} \end{aligned} \tag{1.1}$$

Das Optimierungsproblem besteht also im allgemeinsten Fall aus der *Minimierung* einer *Zielfunktion* $f(\boldsymbol{x})$ eines unbekannten *Parametervektors* $\boldsymbol{x}$, unter gegebenen *Restringierungen* oder *Nebenbedingungen* c_i der zulässigen *Lösungsmenge* $\mathscr{H}_{ad}$, die eine eventuell existierende, konkrete Lösung $\boldsymbol{x}^* \in \mathscr{H}_{ad}$ dieses Parametervektors $\boldsymbol{x}$ erfüllen muss. Ist hingegen – wie im Zuge dieser Arbeit – die *Maximierung* einer *Zielfunktion* $f(\boldsymbol{x})$ gewünscht, lässt sich dieses Problem in Form einer Minimierungsaufgabe wie folgt notieren

$$\arg \max_{\boldsymbol{x} \in \mathscr{H}_{ad}} f(x) = \arg \min_{\boldsymbol{x} \in \mathscr{H}_{ad}} -f(x)$$

Entscheidend für eine *eindeutige* und allgemein auch einfachere Lösbarkeit des in (1.1) definierten Optimierungsproblems, ist die fundamentale Eigenschaft der *Konvexität*.

Konvexe Menge Eine Menge $\mathscr{H} \in \mathbb{R}^n$ wird dann als *konvexe Menge* bezeichnet, wenn eine gerade Verbindungsline zwischen zwei beliebigen Punkten $x \in \mathscr{H}$ und $y \in \mathscr{H}$ der Menge $\mathscr{H}$, zur Gänze in selbiger enthalten ist. Formal schreibt man (Nocedal und Wright, 2006, S.8)

$$\alpha\, x + (1-\alpha)\, y \;\in\; \mathscr{H} \quad \forall \quad \alpha \in [0,1] \tag{1.2}$$

Konvexe Funktion Weiters wird eine Funktion $f(\cdot)$ als *konvexe Funktion* bezeichnet, wenn ihr Definitionsbereich $\mathscr{H} \subseteq \mathbb{R}^n$ eine konvexe Menge ist, also $f(\cdot) : \mathscr{H} \to \mathbb{R}^n$, und alle Funktionswerte $f(\cdot)$ zwischen zwei beliebigen Punkten $x \in \mathscr{H}$ und $y \in \mathscr{H}$ unterhalb oder auf der geraden Verbindungslinie dieser beiden Punkte liegen. Formal muss also für zwei beliebige Punkte $x \in \mathscr{H}$ und $y \in \mathscr{H}$ die folgende Bedingung erfüllt sein (Nocedal und Wright, 2006, S.8)

$$f(\alpha\, \boldsymbol{x} + (1-\alpha)\, \boldsymbol{y}) \leq \alpha\, \boldsymbol{x} + (1-\alpha)\, \boldsymbol{y} \quad \forall\ \alpha \in [0,1] \tag{1.3}$$

Man spricht von *strikt konvexen Funktionen* falls

$$f(\alpha\, \boldsymbol{x} + (1-\alpha)\, \boldsymbol{y}) \leq \alpha\, \boldsymbol{x} + (1-\alpha)\, \boldsymbol{y} \quad \forall\ \alpha \in (0,1),\ \boldsymbol{x} \neq \boldsymbol{y}$$

erfüllt ist und von *(strikt) konkaven Funktionen* falls $-f(\cdot)$ (strikt) konvex ist.

Konvexes Optimierungsproblem Ein *Optimierungsproblem* gilt als *konvex*, wenn in (1.1) sowohl die Zielfunktion $f(\boldsymbol{x})$ als auch die Nebenbedingungen $c_i \leq 0$ zumindest konvex sind und die Nebenbedingungen $c_i = 0$ linear sind.

Zu den im Laufe dieses Kapitels wiederkehrenden Beispielen konvexer Funktionen zählen unter anderem

- die *linear konvexe Funktion* $f(\boldsymbol{x}) = \boldsymbol{c}'\boldsymbol{x} + \alpha$, für jeden beliebigen konstanten Vektor $\boldsymbol{c} \in \mathbb{R}^n$ und jeden Skalar α, und
- die *quadratisch konvexe Funktion* $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$, für jede symmetrisch und positiv semidefinite Matrix $\boldsymbol{A}$.

Je nach Konfiguration der Zielfunktion $f(\boldsymbol{x})$, dem Wertebereich des unbekannten Parametervektors $\boldsymbol{x}$ und eventuell vorhandener Restringierungen c_i in (1.1) auf der vorherigen Seite, unterscheidet man auch unterschiedliche Ausprägungen des Optimierungsproblems. Entsprechend der Aufgabenstellung dieser Arbeit, beschränken sich die nachfolgenden Ausführungen zunächst auf ein

- *lineares* – bezüglich der Zielfunktion
- *deterministisches* und *kontinuierliches* – bezüglich des deterministischen Wertebereichs des Parametervektors – und
- *unrestringiertes*

Optimierungsproblem, für das ein *lokal* gültiges Optimum gesucht werden soll. Es wird sich zeigen, dass die eigentlich für diese Arbeit erforderliche *schrankenrestringierte* Optimierungsaufgabe, ein einfacher Spezialfall der restringierten Optimierung ist und bereits mit wenigen Modifizierung einer unrestringierten Methode gelöst werden kann.

1.2 Grundlagen der linearen, unrestringierten Optimierung

Im einfachsten Fall, dem *linearen*, *deterministischen*, *kontinuierlichen* und *unrestringierten* Optimierungsproblem, degeneriert das mathematische Modell in (1.1) zu (Nocedal und Wright, 2006, S.10)

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \tag{1.4}$$

mit $\boldsymbol{x} \in \mathbb{R}^n$ als deterministischen, reellen Vektor mit $n \geq 1$ und $f : \mathbb{R}^n \to \mathbb{R}$ als stetige Zielfunktion.

Für die Entwicklung eines Lösungsalgorithmus zu diesem Problem, müssen natürlich zunächst jene Bedingungen definiert werden, unter denen eine für (1.4) gefundene Lösung $\boldsymbol{x}^*$ als *optimal* bezeichnet werden darf. Darüber hinaus unterscheidet man noch zwischen einer *globalen* und *lokalen* Lösungen (Nocedal und Wright, 2006, S.7f)

Globale Lösung Als *globalen Minimierer* versteht man jenen Punkt der zulässigen Lösungsmenge $\boldsymbol{x}^* \in \mathscr{H}_{ad}$ – im Falle unrestringierter Aufgaben der gesamte euklidische Vektorraum $\boldsymbol{x}^* \in \mathbb{R}^n$ –, an dem die Zielfunktion ihren kleinsten Funktionswert annimmt. Formal muss also für einen globalen Minimierer $\boldsymbol{x}^*$

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \mathbb{R}^n$$

gelten. Da die geometrische Form der Zielfunktion aber meistens nur lokal bekannt ist, kann es sehr aufwendig sein ein globales Minimum der Zielfunktion zu finden. Im Falle nicht konvexer Optimierungsprobleme liefern die Algorithmen daher meistens lokale Lösungen.

Lokale Lösung Unter einer *lokalen Lösung* versteht man somit jene Punkte der zulässigen Lösungsmenge $\boldsymbol{x}^* \in \mathscr{H}_{ad}$ – im Falle unrestringierter Aufgaben der gesamte euklidische Vektorraum $\boldsymbol{x}^* \in \mathbb{R}^n$ –, welche die Zielfunktion in hinreichend kleiner und offener *Nachbarschaft* $\mathscr{N} = \{x \in \mathscr{H}_{ad} \mid \|\boldsymbol{x} - \boldsymbol{x}^*\| < \epsilon\}$ von $\boldsymbol{x}^*$ minimieren. Formal unterscheidet man hier weiters in

- einen *lokalen Minimierer* $\boldsymbol{x}^*$ für den $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \mathscr{N}$ gelten muss, mit $\mathscr{N}$ als hinreichend kleine Nachbarschaft von $\boldsymbol{x}^*$.
- und einen *strikten lokalen Minimierer* $\boldsymbol{x}^*$ für den $f(\boldsymbol{x}^*) < f(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \mathscr{N}$ gelten muss, mit $\mathscr{N}$ als hinreichend kleine Nachbarschaft von $\boldsymbol{x}^*$, *und* $\boldsymbol{x}^* \neq \boldsymbol{x}$.

Konvexes Problem Für den Spezialfall des *konvexen Optimierungsproblems* gilt als wesentliche Eigenschaft, dass jedes lokale Minimum gleichzeitig das eindeutige globale Minimum der Minimierungsaufgabe darstellt.

Im Folgenden sollen nun die *Optimalitätsbedingungen* eingeführt werden, die einerseits dem Auffinden eines der eben beschriebenen Optima dienen, gleichzeitig aber auch die Gültigkeit dessen Kriteriums sicherstellen sollen. Für die Optimalitätsbedingungen sei neben der Stetigkeit der Zielfunktion $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, nun auch deren zweifach stetige Differenzierbarkeit als gegeben vorausgesetzt. Mit diesen Annahmen können folgende Optimalitätsbedingungen formuliert werden[1]

Notwendige Bedingungen Ist $\boldsymbol{x}^*$ ein lokales Minimum der Zielfunktion $f(\boldsymbol{x})$, dann muss $f(\boldsymbol{x}^*)$ ein *stationärer* Punkt sein. Formal schreibt man für diese *Notwendige Bedingung 1. Ordnung* $\nabla f(\boldsymbol{x}^*) = 0$. Diese Bedingung wird allerdings von jedem Extremum einer – nicht notwendigerweise konvexen – Funktion $f(\boldsymbol{x})$ erfüllt, also sowohl von einem Minimum, als auch von einem Maximum, oder einem Sattelpunkt. Um zumindest ein Maximum ausschließen zu können, wird für ein gültiges Optimum zusätzlich die positive Semidefinitheit der *Hessematrix* $\nabla^2 f(\boldsymbol{x}^*) \geq 0$ gefordert.
Zusammenfassend ergebend sich also die *notwendigen Optimalitätsbedingungen* für ein lokales Minimum $\boldsymbol{x}^* \in \mathbb{R}^n$ einer zweimal stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ aus den folgenden Forderungen

$$\begin{aligned} \nabla f(\boldsymbol{x}^*) &= 0 \quad \text{und} \\ \nabla^2 f(\boldsymbol{x}^*) &\geq 0 \end{aligned} \tag{1.5}$$

Hinreichende Bedingungen Hinreichende Optimalitätsbedingungen für ein lokales Minimum

[1] Die Beweise zu diesen Optimalitätsbedingungen findet man u. a. in Nocedal und Wright (2006, Kap.2).

$\boldsymbol{x}^* \in \mathbb{R}^n$, welche die Existenz dieses Minimums sogar *garantieren*, sind für eine zweimal stetig differenzierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ folgendermaßen definiert

$$\begin{aligned} \nabla f(\boldsymbol{x}^*) &= 0 \quad \text{und} \\ \nabla^2 f(\boldsymbol{x}^*) &> 0 \end{aligned} \tag{1.6}$$

mit $\nabla^2 f(\boldsymbol{x}^*) > 0$ als *positiv definite Hessematrix*.

Auch hier lässt sich für ein *konvexes Optimierungsproblem* wieder vereinfachend feststellen, dass jeder stationäre Punkt $\boldsymbol{x}^*$ der einfach differenzierbaren konvexen Zielfunktion $f(\boldsymbol{x})$ ein garantiertes Minimum ist.

Da in den wenigsten Fällen weder der Gradient noch die Hessematrix analytisch bestimmt werden können, wurden zur Lösung eines linearen, unrestringierten Optimierungsproblems zahlreiche Iterationsverfahren entwickelt. All diesen Verfahren ist gemein, dass sie ausgehend von einem gegebenen Startpunkt $\boldsymbol{x}_0$ mit $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}^*)$, den nächsten Iterationspunkt $\boldsymbol{x}_k$ durch eine geeignete Kombination aus Suchrichtung und zugehöriger Schrittweite bestimmen. Diese Wahl wird meistens unter Berücksichtigung der Forderung $f(\boldsymbol{x}_0) > f(\boldsymbol{x}_k) \geq f(\boldsymbol{x}^*)$ getroffen. Dabei unterscheiden sich die Algorithmen im wesentlichen in der Strategie zur Bestimmung der Schrittweite und der Suchrichtung. Aufgrund der nicht speicherkritischen Größe des Optimierungsproblems dieser Arbeit, konnte das *Newton-Verfahren* gewählt werden, welches im nächsten Kapitel auch beschrieben werden wird.

Ein wesentliches Hilfsmittel zur Analyse von Optimierungsaufgaben, ist die Approximation der Zielfunktion $f(\boldsymbol{x})$ an einem Iterationspunkt $\boldsymbol{x}_k$ durch ein Polynom, dessen Eigenschaften an diesem Iterationspunkt ähnlich denen der Zielfunktion sein sollen. Eine solche Approximation stellt die Formel von Taylor und MacLaurin dar, das sogenannte *Taylor Polynom*. Da von dieser Approximation öfters Gebrauch gemacht werden wird, soll sie hier kurz angegeben werden (Nocedal und Wright, 2006, S.14).
Ist die Zielfunktion $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $\boldsymbol{p} \in \mathbb{R}^n$, so gelten die folgenden Beziehungen

$$f(\boldsymbol{x} + \boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x} + \alpha \boldsymbol{p})' \boldsymbol{p} \quad \text{mit} \quad \alpha \in (0, 1) \tag{1.7}$$

für zumindest einfach stetig differenzierbare Zielfunktionen, und

$$f(\boldsymbol{x} + \boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})' \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}' \nabla^2 f(\boldsymbol{x} + \alpha \boldsymbol{p}) \boldsymbol{p} \quad \text{mit} \quad \alpha \in (0, 1) \tag{1.8}$$

für zumindest zweifach stetig differenzierbare Zielfunktionen.

1.3 Das Newton-Verfahren

Das *Newton-Verfahren* gehört zu den *line-search* Methoden die in jedem Iterationsschritt k zunächst eine Suchrichtung $\boldsymbol{p}_k$ wählen, und dann entlang dieser Suchrichtung jene Schrittweite α_k fixieren, die den Funktionswert $f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k)$ reduziert.

Schrittweite Bei als gegeben angenommener Suchrichtung $\boldsymbol{p}_k$ soll zunächst eine optimale Wahl der Schrittweite erläutert werden. Hierzu sei angemerkt, dass die nun folgenden Ausführungen nicht nur für das Newton-Verfahren gültig sind, sondern für alle Optimierungsalgorithmen die innerhalb des line-search Rahmens operieren.
Wie in der Einleitung bereits angedeutet, stellt die Lösung des Schrittweitenproblems ein dem eigentlichen Optimierungsproblem untergelagertes, skalares Optimierungsproblem dar. Die optimale Lösung wäre natürlich der globale Minimierer von

$$\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k), \quad \alpha > 0 \tag{1.9}$$

Für eine quadratisch konvexe Funktion wie $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}'\boldsymbol{x}$ lässt sich das globale Optimum auch analytisch mit

$$\alpha_k = -\frac{\nabla f_k' \boldsymbol{p}_k}{\boldsymbol{p}_k' \boldsymbol{Q} \boldsymbol{p}_k}$$

angeben. Für allgemein nichtlineare Funktionen wäre die exakte numerische Lösung des Problems allerdings zu rechenaufwendig. Praktikable Algorithmen versuchen daher mit minimalem numerischem Aufwand, aus einer Reihe von Kandidatenpunkten α_k – die sogenannte *backeting phase* – jenen auszuwählen – die *interpolation phase* – der zu einer adäquaten Reduktion der Zielfunktion führt. Zu den Bedingungen die solch eine adäquate, namentlich hinreichende, Reduktion der Zielfunktion sicherstellen, gehören

- die Wolfe-Bedingungen
- die Goldstein Bedingungen
- und das backtracking line-search Verfahren

Da die Goldstein Bedingungen eventuell alle Kandidatenpunkte ausschließen kann, beschränken sich die weiteren Ausführungen auf die Wolfe-Bedinungen und das insbesondere in Newton-Verfahren verwendete backtracking line-search Verfahren[2].

[2] Für Quasi-Newton Verfahren wird die einfachere backtracking Methode hingegen nicht empfohlen (Nocedal und Wright, 2006, S.36f)

Wolfe-Bedingungen Die Wolfe-Bedingungen setzten sich aus der *Abstiegsbedingung*[3]

$$\underbrace{f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k)}_{\phi(\boldsymbol{\alpha})} \leq \underbrace{f(\boldsymbol{x}_k)}_{\phi(0)} + c_1 \underbrace{\alpha \nabla f_k' \boldsymbol{p}_k}_{\phi(0)'}, \quad c_1 \in (0,1) \tag{1.10}$$

und der *Krümmungsbedingung*

$$\nabla f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k)' \boldsymbol{p}_k \geq c_2 \alpha \nabla f_k' \boldsymbol{p}_k, \quad c_2 \in (c_1, 1) \tag{1.11}$$

zusammen.
Die Abstiegsbedingung (1.10) stellt nichts anderes als eine Gerade durch den Punkt $\phi(0)$ mit negativer Steigung $c_1 \phi(0)'$ dar. Alle Funktionspunkte $\phi(\boldsymbol{\alpha})$ die unterhalb dieser definierten Gerade liegen, würden als Lösung von (1.9) akzeptiert werden. Für sehr kleine Werte c_1 liegt diese Gerade deutlich über dem Zielfunktionsverlauf. Üblicherweise wird daher c_1 mit $c_1 = 10^{-4}$ gewählt (Nocedal und Wright, 2006, S.33).
Um gleichzeitig zu kleine Schrittlängen auszuschließen, soll die Krümmungsbedingung (1.11) sicherstellen, dass die negative Steigung am Punkt $\phi(\alpha_k)$ mindestens das c_2-fache der Anfangskrümmung aufweist. Für das Newton-Verfahren und Quasi-Newton Verfahren, wie dem BFGS Algorithmus, wird c_2 üblicherweise mit $c_2 = 0.9$ gewählt (Nocedal und Wright, 2006, S.34).
Formuliert man die Krümmungsbedingung in der etwas verschärften Form

$$|\nabla f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k)' \boldsymbol{p}_k| \leq c_2 |\alpha \nabla f_k' \boldsymbol{p}_k|, \quad c_2 \in (c_1, 1) \tag{1.12}$$

lässt sich darüber hinaus sicherstellen, dass sich die Lösung des Problems (1.9) zumindest in der Nähe eines lokalen Minimums befindet. In Kombination mit der Abstiegsbedingung (1.10) spricht man in diesem Fall von den *strengen Wolfe-Bedingungen*.

Ein Pseudoalgorithmus zur iterativen Lösung des line-search Optimierungsproblems in (1.9), unter Beachtung der strengen Wolfe-Bedingungen, findet sich z. B. in Nocedal und Wright (2006, S.60ff). Wie alle line-search Algorithmen startet auch dieser mit einer Anfangslösung und generiert eine monoton ansteigende Lösungssequenz $\{\alpha_k\}$. Unter der Annahme, dass $\boldsymbol{p}_k$ in (1.9) eine negative Steigung hat und die Zielfunktion $f(\cdot)$ von unten beschränkt ist, garantiert dieser Algorithmus das Auffinden einer gültigen Lösung α^*.

backtracking line-search In der backtracking Strategie kann durch eine geeignete Auswahl

[3] auch als *Armijo Bedingung* bekannt

der Kandidatenpunkte auf die Krümmungsbedingung (1.11) verzichtet werden. Die Strategie beschränkt sich also alleine auf die Abstiegsbedingung.
Auch hierfür findet sich z. B. in Nocedal und Wright (2006, S.37) ein Pseudoalgorithmus, der ausgehend von einer, für Newton-Verfahren mit 1 gewählten initialen Schrittweite, selbige iterativ verkürzt bis die Abstiegsbedingung erfüllt ist.

Suchrichtung Wie bereits erwähnt, muss in einem line-search Verfahren zur Bestimmung des nächsten Iterationspunktes $\boldsymbol{x}_{k+1}$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k \tag{1.13}$$

neben der Schrittweite α_k auch die Suchrichtung $\boldsymbol{p}_k$ bestimmt werden. Die intuitivste und einfachste Möglichkeit bietet hierfür das *Gradientenverfahren*, das am Iterationspunkt $\boldsymbol{x}_k$ als Suchrichtung schlicht den negativen Gradienten $\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)/\|\nabla f(\boldsymbol{x}_k)\|$ verwendet. Mit dieser Wahl steht $\boldsymbol{p}_k$ in jedem Iterationspunkt $\boldsymbol{x}_k$ orthogonal zur Höhenlinie $f(\boldsymbol{x}_k) = \text{konst.}$ und stellt somit an diesem Iterationspunkt den steilst möglichen Abstieg im Funktionswert dar.[4] Für gut konditionierte Probleme konvergiert diese Vorgehensweise sehr schnell zur optimalen Lösung und erfordert dafür auch lediglich Gradienteninformationen. Der wesentliche Nachteil dieses Verfahrens liegt allerdings darin, dass es bei schlecht konditionierten oder schlecht skalierten Problemen, äußerst langsam werden kann.
Eine mögliche Alternative stellt die *Newton Richtung* dar. Zu Herleitung dieser Suchrichtung wird die Zielfunktion am Iterationspunkt $\boldsymbol{x}_k$ durch ein Taylorpolynom zweiter Ordnung (1.7) auf Seite 6 approximiert (Nocedal und Wright, 2006, S.22)

$$f(\boldsymbol{x}_k + \boldsymbol{p}) \approx f(\boldsymbol{x}_k) + \boldsymbol{p}'\nabla f(\boldsymbol{x}_k) + \frac{1}{2}\boldsymbol{p}'\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p} \tag{1.14}$$

Unter der Annahme einer positiv definiten Hessematrix $\nabla^2 f(\boldsymbol{x}_k)$ führt einfache Ableitung nach $\boldsymbol{p}$ und Nullsetzung dieses konvexen quadratischen Polynoms zur Newton Richtung

$$\boldsymbol{p}_k = -(\nabla^2 f(\boldsymbol{x}_k))^{-1}\nabla f(\boldsymbol{x}_k) \tag{1.15}$$

Diese Richtung ist insbesondere dann zuverlässig, wenn die Taylorapproximation am Iterationspunkt $\boldsymbol{x}_k$ hinreichend genau ist.
Im Falle positiver Definitheit der Hessematrix konvergieren Verfahren mit Hilfe der Newton Richtung üblicherweise quadratisch zur optimalen lokalen Lösung. Für quadratische Optimie-

[4] weswegen dieses Verfahren auch als *stepptest descent* Verfahren bekannt ist

rungsprobleme konvergiert dieses Verfahren sogar unabhängig vom Startpunkt $\boldsymbol{x}_0$ innerhalb von einem Iterationsschritt. Der wesentliche Nachteil der Newton Formulierung liegt allerdings in der Erfordernis der Hessematrix, deren Bestimmung sehr rechenintensiv sein kann. Ist eine exakte Bestimmung der Jacobi-Matrix und der Hessematrix zu aufwendig oder nicht möglich, bieten sogenannte *finite-Differenzenapproximationen* und *automatische Differenzierungsalgorithmen* numerische Approximationsmöglichkeiten (näheres siehe z. B. Nocedal und Wright (2006, Kap.8)). *Quasi-Newton* Verfahren, wie die populäre *BFGS* Methode, versuchen diesen Nachteil der Newton Suchrichtung durch eine Approximation der Hessematrix zu vermeiden. Da das in der gegenständlichen Arbeit zu optimierende mathematische Modell allerdings eine einfache lineare Regressionsfläche ist, können sowohl der Gradient als auch die Hessematrix analytisch bestimmt werden. Demzufolge stellt die Newton-Richtung für den Optimierungsalgorithmus eine logische Wahl dar. Zur Berückstichtgung von Schrankenrestrinktionen muss sie allerdings ein wenig modifziert werden. Die dafür notwendigen Schritte werden im nächsten Abschnitt dargelegt.

Schrankenrestringierte Optimierung Das mathematische Modell der schrankenrestringierten Optimierung ist als einfacher Spezialfall des allgemein restringierten Optimierungsproblemes (1.1) wie folgt definiert (Hintermüller, S.59)

$$\min_{\boldsymbol{x}\in\mathscr{H}_{ad}} f(\boldsymbol{x}) \quad \text{mit} \quad \mathscr{H}_{ad} = \{\boldsymbol{x} \in \mathbb{R}^n \mid L_i \leq x \leq U_i\} \tag{1.16}$$

$L_i \in \mathbb{R}$ und $U_i \in \mathbb{R}$ definieren hier die *endlichen, skalaren* Schranken. Falls $x_i = L_i$ oder $x_i = U_i$ gilt, mit $i = 1, \ldots, n$, bezeichnet man diesen Index $i \in \mathscr{A}(\boldsymbol{x})$ als *aktiv*, andernfalls als *inaktiv*. Die Menge $\mathscr{A}(\boldsymbol{x})$ heißt *aktive Menge* und beinhaltet alle aktiven Indizes.
Die in 1.2 auf Seite 5 eingeführten Optimalitätsbedingungen, müssen für einen schrankenrestringierten Algorithmus nun neu formuliert werden.

Notwendige Bedingungen Sei wieder $\boldsymbol{x}^* \in \mathbb{R}^n$ ein lokales Minimum der zweimal stetig differenzierbaren Funktion $f : [L, U] \to \mathbb{R}$ auf dem Intervall $-\infty < L_i < U_i < \infty$ so stellt sich die *notwendige Bedingung 1. Ordnung* wie folgt dar (Hintermüller, S.59)

$$\nabla f(\boldsymbol{x}^*)'(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0 \quad \forall\, \boldsymbol{x} \in \mathscr{H}_{ad} \tag{1.17}$$

Zur Angabe der notwendigen Bedingung 2. Ordnung wird die *reduzierte Hessematrix* ∇^2_R eingeführt. Für eine zweimal stetig differenzierbare Funktion $f : [L, U] \to \mathbb{R}$ ist

diese wie folgt definiert (Hintermüller, S.60)

$$(\boldsymbol{H}_R)_{ij} = (\nabla_R^2 f(\boldsymbol{x}))_{ij} = \begin{cases} \delta_{ij} & \text{für } i \in \mathscr{A}(\boldsymbol{x}) \text{ oder } j \in \mathscr{A}(\boldsymbol{x}), \\ (\nabla_R^2 f(\boldsymbol{x}))_{ij} & \text{sonst} \end{cases} \tag{1.18}$$

mit δ_{ij} als *Kroneckersymbol*. Dabei gilt $\delta_{ij} = 0$ für $i \neq j$ und $\delta_{ij} = 1$ für $i = j$. Mit Hilfe der reduzierten Hessmatrix lässt sich die *notwendige Bedingung 2. Ordnung* wie folgt angeben

$$\nabla_R^2 f(\boldsymbol{x}^*) \geq 0 \tag{1.19}$$

Hinreichende Bedingung Zur Angabe der hinreichenden Bedingung wird der Begriff des *nichtdegenerierten statinären Punkts* benötigt. Für solch einen Punkt muss (Hintermüller, S.61)

$$\nabla f(\boldsymbol{x}^*)_i \neq 0 \quad \forall\, i \in \mathscr{A}(\boldsymbol{x}^*) \tag{1.20}$$

gelten. Anders ausgedrückt erfüllt ein *nichtdegeneriertes lokales Minimum* $\boldsymbol{x}^*$ die folgenden *strikten komplimentaritäts* Beziehungen

$$\begin{aligned} (x_i = L_i \vee x_i = U_i) \wedge \nabla f(\boldsymbol{x})_i \neq 0 \\ L_i < x_i < U_i \wedge \nabla f(\boldsymbol{x})_i = 0 \end{aligned}$$

Im Folgenden sollen nun die Grunzüge des *projezierten Newton-Verfahrens* angegeben werden (vgl. Hintermüller, S.61ff).
Anstelle des in (1.13) auf Seite 9 angegebenen neuen Iterationspunktes $\boldsymbol{x}_{k+1}$ wird für das projezierte Verfahren

$$\boldsymbol{x}_{k+1} = P(\boldsymbol{x}_k - \alpha(\nabla_R^2 f(\boldsymbol{x}_k))^{-1} \nabla f(\boldsymbol{x}_k)) \tag{1.21}$$

verwendet. Mit ∇_R^2 als reduzierte Hessematrix gemäß (1.18) und $P : \mathbb{R}^n \to \mathscr{H}_{ad}$ als Projektionsmatrix auf die Definitionsmenge $\mathscr{H}_{ad}$. Diese ist gegeben durch

$$P(\boldsymbol{x})_i = \begin{cases} L_i & \text{falls } x_i < L_i, \\ x_i & \text{falls } L_i < x_i < U_i \text{ und} \\ U_I & \text{falls } x_i > U_i \end{cases} \tag{1.22}$$

Zur Bestimmung einer adäquaten Schrittweite α im schrankenrestringierten Fall, müssen neben der Suchrichtung auch noch die Schrittweitenstrategien aus 1.3 auf Seite 7 angepasst werden. Die z. B. im *backtracking-line search* Verfahren verwendete Abstiegsbedingung, lau-

tet im Falle projezierter Gradientenverfahren

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -\frac{c}{\alpha}\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\|^2 \tag{1.23}$$

mit $c \in (0, 1)$.
Ein Pseudoalgorithmus für das projezierte Newton-Verfahren findet sich z. B. in Hintermüller.

Da in MATLAB® bereits zahlreiche restringierte und unrestringierte Optimierungsalgorithmen zur Verfügung gestellt werden, die u. a. auf der hier vorgestellten line-search Methode aufbauen, wurde in der gegenständigen Arbeit auf diese Implementierungen zurückgegriffen. Konkret wurde die Funktion `fmincon` mit einem active-set Algorithmus verwendet. Obgleich es möglich gewesen wäre, sowohl den Gradienten als auch die Hessematrix analytisch zu bestimmen und der Funktion zur Verfügung zu stellen, wurde die Berechnung der Ableitungen an die Funktion selbst delegiert. MATLAB® verwendet hierfür die finite Differenzenapproximation.

1.4 Methode der kleinsten Fehlerquadratsumme

Speziell für das Kapitel der linearen Regression sollen in diesem letztem Teilkapitel der Optimierung noch die Grundlagen zur Lösung des folgenden unrestringierten Optimierungsproblems (Nocedal und Wright, 2006, S.246)

$$f(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \mathbb{R}^m \tag{1.24}$$

mit $r_i : \mathbb{R}^m \to \mathbb{R}$ als eine stetige Funktion und $n \geq m$, erläutert werden. Die Funktion r_i, das sogenannte *Residuum*, könnte z. B. ein Maß für die Diskrepanz zwischen dem Ergebnis eines mathematischen Modells $\hat{\boldsymbol{y}}(\boldsymbol{x})$ und der Daten $\boldsymbol{y}$ der Realität sein. Fasst man die Residuen r_i zu einem Vektor $\boldsymbol{r}(\boldsymbol{x}) : \mathbb{R}^m \to \mathbb{R}^n$ zusammen, und setzt für das mathematische Modell – wie im Falle der linearen Regression[5] – $\hat{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{X}\,\boldsymbol{\beta}$ an, so ergibt sich dieser Residuenvektor z. B. zu $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{y}$. Das Optimierungsproblem in (1.24) lässt sich in diesem Fall wie folgt anschreiben

$$f(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{y}\|^2 \tag{1.25}$$

mit $\|\cdot\|$ als euklidische Norm. Eine wesentliche Eigenschaft dieser Zielfunktion $f(\boldsymbol{\beta})$ und auch ganz allgemein der Zielfunktion in (1.24) liegt darin, dass die Hessematrix $\nabla^2 f(\boldsymbol{x})$ einfach

[5] näheres siehe Kapitel 2

als Gradientenprodukt dargestellt werden kann. Es lässt sich also folgender Zusammenhang zeigen, (Nocedal und Wright, 2006, S.247)

$$\begin{aligned} \nabla f(\boldsymbol{\beta}) &= \sum_{i=1}^{n} r_i(\boldsymbol{\beta})\, \nabla r_i(\boldsymbol{\beta}) = \boldsymbol{J}(\boldsymbol{\beta})'\, \boldsymbol{r}(\boldsymbol{\beta}) \\ \nabla^2 f(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \nabla r_i(\boldsymbol{\beta})\, \nabla r_i(\boldsymbol{\beta})' + \sum_{i=1}^{n} r_i(\boldsymbol{\beta})\, \nabla^2 r_i(\boldsymbol{\beta}) \\ &= \boldsymbol{J}(\boldsymbol{x})'\, \boldsymbol{J}(\boldsymbol{x}) + \approx 0 \end{aligned} \tag{1.26}$$

mit $\nabla f(\boldsymbol{\beta})$ als Gradienten, $\nabla^2 f(\boldsymbol{\beta})$ als Hessematrix und $\boldsymbol{J}(\boldsymbol{\beta})$ als Jacobi-Matrix. Einsetzten des Residuums aus (1.25) führt auf die folgenden Ausdrücke

$$\nabla f(\boldsymbol{\beta}) = \boldsymbol{X}'(\boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{y}), \quad \text{und} \quad \nabla^2 f(\boldsymbol{\beta}) = \boldsymbol{X}'\,\boldsymbol{X} \tag{1.27}$$

Da für die *lineare* Regression sogar $\nabla^2 r_j = 0$ gilt, verschwindet der zweite Ausdruck in (1.26) für die Hessematrix exakt. Da darüber hinaus die Zielfunktion in (1.25) konvex ist[6], entspricht gemäß 1.2 auf Seite 6 jeder stationäre Punkt dem globalen Minimum. Die eindeutige, globale Lösung ergibt sich somit aus der folgenden *Normalgleichung*

$$\boldsymbol{X}'\,\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}'\,\boldsymbol{y} \tag{1.28}$$

Unter der Annahme einer regulären Matrix $\boldsymbol{X}$ mit $n \geq m$ lässt sich dieses lineare Gleichungssystem mit einem der folgenden Algorithmen lösen (vgl. Nocedal und Wright, 2006, S.251ff)

1. der Cholesky-Faktorisierung
2. der QR-Faktorisierung, oder
3. der singular-value-decomposition, dem SVD-Algorithmus

Die *Cholesky-Faktorisierung* (vgl. z. B Plato, 2010, S.71ff) wird sehr häufig verwendet und ist insbesondere für den Fall sehr vieler Beobachtungen n, also $n \gg m$, und einfach zu speicherndem Matrixprodukt $\boldsymbol{X}'\,\boldsymbol{X}$ geeignet. Sie hat aber den entscheidenden Nachteil, dass der relative Fehler sehr groß werden kann. Dies deswegen, da der relative Fehler üblicherweise proportional zur Konditionszahl κ ist, die im Falle des Matrixprodukts dem Quadrat von $\kappa(\boldsymbol{X})$ entspricht. Darüber hinaus, kann das Cholesky-Verfahren bei schlecht konditionierten Matrizen zur Gänze versagen.
Die *QR-Faktorisierung* (vgl. z. B Plato, 2010, S.88ff) ist üblicherweise etwas robuster als die

[6] vgl. hierzu auch die Beispiele in 1.2 auf Seite 4

Cholesky-Faktorisierung. Darüberhinaus ist der relative Fehler des Optimums $\boldsymbol{\beta}^*$ in dieser Formulierung nur proportional zur Konditionszahl $\kappa(\boldsymbol{X})$. Die in dieser Arbeit verwendete `regstats`-Funktion der MATLAB® -Umgebung verwendet diese Faktorisierung.
Der *SVD-Algorithmus* ist zwar der rechenaufwendigste unter den eben vorgestellten Algorithmen, ist aber auch gleichzeitig der robusteste, mit den verlässlichsten Ergebnissen. Zusätzlich lässt sich mit seiner Hilfe feststellen, wie sensitiv die Lösung $\boldsymbol{\beta}^*$ auf kleine Störungen der Eingangsgrößen $\boldsymbol{X}$ reagiert.

2 Lineare Regression

„All models are wrong, but some are useful.“ (George E. P. Box)

Vorbemerkungen

In diesem Kapitel wird die Kenntnis folgender statistischer Grundbegriffe vorausgesetzt:

Ereignisfeld Ein Ereignisfeld ist ein System von Teilmengen eines *Merkmalraums*, das bestimmte Eigenschaften erfüllt. Die leere Menge ist das unmögliches Ereignis. Ein Ereignisfeld mit nur einem Element heißt elementares Ereignis.

Erwartungswert Der Erwartungswert oder das 1. Moment einer stochastischen Größe X, ist ein Zahlenwert, der das Zentrum der Wahrscheinlichkeitsverteilung dieser Größe anzeigt. Sofern das Integral existiert gilt im Falle einer kontinuierlichen Dichtefunktion $\mathrm{f}(\cdot)$

$$\mathbb{E}X = \int_{-\infty}^{\infty} x\,\mathrm{f}(x)\;dx$$

Die Berechnung des Erwartungswertes ist eine lineare Operation.

Korrelationskoeffizient Der Korrelationskoeffizient $\rho_{X,Y}$ ist eine standardisierte Maßzahl für die lineare Beziehung zweier stochastischer Größen X und Y. Er ist wie folgt definiert

$$\rho_{X,Y} = \frac{\mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right]}{\sigma_X \sigma_Y} \quad \text{mit} \quad -1 \leq \rho_{X,Y} \leq 1$$

Ein Korrelationskoeffizienten $\rho_{X,Y} = 0$ signalisiert unkorrelierte stochastische Größen, während bei $|\rho_{X,Y}| = 1$ eine lineare Abhängigkeit mit Wahrscheinlichkeit 1 besteht. Bei einem Korrelationskoeffizienten $\rho_{X,Y} > 0$ spricht man von positiver Korrelation, bei $\rho_{X,Y} < 0$ von negativer Korrelation.

Kovarianz Im Gegensatz zum *Korrelationskoeffizienten* handelt es sich bei der Kovarianz um eine nicht standardisierte Maßzahl für die lineare Beziehung zweier stochastischer Größen X und Y. Sie ist von Null verschieden, wenn eine lineare Beziehung besteht,

ansonsten identisch Null.
Im Falle der Existenz der *Varianzen* $\mathrm{Var}(X)$ und $\mathrm{Var}(Y)$ gilt

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)]$$

Merkmalraum Der Merkmalraum beinhaltet alle möglichen Werte die eine stochastische Größe annehmen kann.

Varianz Die Varianz oder das 2. zentrale Moment einer stochastischen Größe X ist ihr wichtigster Streuparameter. Im Falle der Existenz des Erwartungswertes gilt

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2]$$

Wahrscheinlichkeitsraum Ein Wahrscheinlichkeitsraum wird aus dem Tripel *Merkmalraum*, einem darin befindlichen *Ereignisfeld* und der darauf definierten *Wahrscheinlichkeitsfunktion* aufgebaut.

Wahrscheinlichkeitsverteilung Die Wahrscheinlichkeitsverteilung ist eine Funktion auf dem *Ereignisfeld*, die jedem Ereignis des Ereignisfeldes eine reelle Zahl zuordnet.

Eine grundlegende, mathematisch orientierte Einführung in das Gebiet der Stochastik mit zahlreichen Definitionen und auch einer kurzen Einführung in die Regression, findet sich z. B. in Viertl (1990). Eine vertiefende, eher anwendungsorientierte Einführung in die derzeit zur Verfügung stehenden stochastischen Regressionsmodelle, bietet z. B. Fahrmeir et al. (2009). Auf lineare Regressionsmodelle beschränkt und speziell auf deren Regressionsanalyse ausgerichtete Bücher können u. a. Fox (2008), Xin und Xiao (2009) und Urban und Mayerl (2008) emfpohlen werden. Besonders Fox (2008) scheint auf diesem Gebiet zur Standardliteratur zu zählen. Das Werk von Miller (2002), dass sich nahezu ausschließlich dem Thema der Modellselektion widmet, dürfte in diesem Gebiet ebenso zur Standardliteratur gehören.
Als ein Werk mit direktem Bezug zum Bauingenieurwesen sei das Buch von Bucher (2009) erwähnt, welches stochastische Methoden im Kontext von strukturmechanischen Aufgabenstellungen erläutert.

Beispiele und Abbildungen Um textlich beschriebene Eigenschaften und Effekte des klassischen linearen Regressionsmodells kontrolliert visualisieren zu können, wurde das Modell in

(2.1) für $n = 50$ Beobachtungen *exakt* bestimmt.

$$\boldsymbol{y} = \beta_0 + \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{x}_1^2\beta_3 + \boldsymbol{\epsilon} \tag{2.1}$$

Zu diesem Zweck wurden in umgekehrter Weise zunächst sämtliche Modellparameter festgelegt und erst mit deren Hilfe die abhängige Größe $\boldsymbol{y}$ berechnet. Die Regressoren $x \in [0, 1]$ wurden mit Hilfe der `haltonset`-Klasse der MATLAB®-Umgebung als Halton Folge generiert, der Parametervektor $\boldsymbol{\beta}$ und auch die Störgröße $\boldsymbol{\epsilon}$ mittels der `rand`-Funktion einer standardisierten Normalverteilung entnommen. Die Werte des Parametervektors wurden dann so angepasst, dass der quadratische Anteil des Regressor x_1 signifikant größer als sein linearer Anteil ist. Mit den derart bekannten Parametern wurde schließlich die abhängige Größe $\boldsymbol{y}(\boldsymbol{X})$ festgelegt. Weiters wurde in analoger Weise mit einem Teil der Parameter aus (2.1) das folgende Regressionsmodell exakt festgelegt

$$\boldsymbol{y} = \exp(\beta_0 + \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\epsilon}) \tag{2.2}$$

Sämtliche der in diesem Kapitel 2 angeführten Beispielausgaben und Beispielabbildungen, beziehen sich nun auf eines der nachfolgend angeführten geschätzten linearen Regressionsmodelle. Wenn nicht anders angegeben, bezieht sich die Schätzung stets auf das klassische lineare Regressionsmodell in (2.1).

$$\boldsymbol{y} = \hat{\beta}_0 + \boldsymbol{x}_2\hat{\beta}_1 \tag{2.3}$$

$$\boldsymbol{y} = \hat{\beta}_0 + \boldsymbol{x}_1\hat{\beta}_1 + \boldsymbol{x}_2\hat{\beta}_2 \tag{2.4}$$

$$\boldsymbol{y} = \hat{\beta}_0 + \boldsymbol{x}_1\hat{\beta}_1 + \boldsymbol{x}_2\hat{\beta}_2 + \boldsymbol{x}_1^2\hat{\beta}_3 \tag{2.5}$$

2.1 Einführung in die Regression

Mathematische Größen die durch ein Modell beschrieben werden können, welches mit den selben Eingabewerten wiederholt das selbe Ergebnis liefert, bezeichnet man als *deterministische Größen*. Im Gegensatz dazu werden Größen, die sich nicht durch solch ein deterministisches Modell beschreiben lassen, als *stochastische Größen* bezeichnet. Dabei geht man davon aus, dass den stochastischen Größen eine wahre, jedoch unbekannte Wahrscheinlichkeitsverteilung zugrunde liegt, die durch einen endlichdimensionalen Parameter charakterisierbar ist. Ziel der *klassischen parametrischen Inferenzstatistik* ist es nun, diesen Parameter für konkrete Daten möglichst gut zu schätzen und damit ein *stochastisches Modell* zu konstruieren. Formal betrachtet man stochastische Größen also als Abbildung eines Wahrscheinlichkeitsraumes in den

Euklidischen Raum.
Das in diesem Kapitel zu beschreibende *Regressionsmodell* ist ein solches stochastisches Modell.

Ziel eines Regressionsmodells ist es somit, den kausalen Zusammenhang zwischen einer *abhängigen Größe* $\boldsymbol{y}$ – auch Zielfunktion genannt – und ihren dazu gehörenden *Regressoren* $\boldsymbol{X}$ – auch Kovariablen, erklärende oder unabhängige Variablen genannt – geeignet zu beschreiben. Im Sinne eines stochastischen Modells, entspricht die abhängige Variable $\boldsymbol{y}$ einer stochastischen Größe. Existiert deren Erwartungswert $\mathbb{E}(\boldsymbol{y})$ für alle x aus einem bestimmten Bereich $x \in [a, b]$ gilt (Viertl, 1990, S.132f)

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{X}) = \mathrm{f}(\boldsymbol{X}) \tag{2.6}$$

mit $\mathrm{f}(\boldsymbol{X})$ als *deterministische Regressionsfunktion*[1] der stochastischen Schätzfunktion $\boldsymbol{y}$ bezüglich $\boldsymbol{X}$. Die Verteilung der stochastischen Größe $\boldsymbol{y}$ hängt im Allgemeinen von den Regressoren ab. Das *stochastische Regressionsmodell* wird nun in folgender Weise festgelegt

$$\boldsymbol{y}(\boldsymbol{X}) = \mathrm{f}(\boldsymbol{X}) + \boldsymbol{\epsilon} \tag{2.7}$$

mit $\boldsymbol{\epsilon}$ als stochastische Größe mit Erwartungswert $\mathbb{E}(\boldsymbol{\epsilon}) = 0$. Die stochastische Größe $\boldsymbol{\epsilon}$ wird als *Störgröße* bezeichnet und ist die zufällige, nicht von den Regressoren erklärte Abweichung vom Erwartungswert $\mathbb{E}(\boldsymbol{y})$ der abhängigen Größe. Das charakteristische an stochastischen Modellen wie diesem ist, dass der durch (2.7) beschriebene Zusammenhang eben nicht exakt gilt, sondern durch die zufälligen Einflüsse $\boldsymbol{\epsilon}$ überlagert wird.
Die auf (2.7) aufbauenden Regressionsmodelle unterscheiden sich in der Definition der deterministischen Regressionsfunktion, wobei sämtliche zur Verfügung stehende Regressionsmodelle Spezialfälle des *Strukturiert-Additiven Regressionsmodells* (STAR) sind. Diese Spezialfälle unterscheiden sich einerseits im Typ der verwendeten Größen und andererseits in ihrer Komplexität. Für den einfachsten Fall des *linearen Regressionsmodells* (LM) müssen die verwendeten Größen metrisch oder kategorial und die zu beschreibenden Effekte linear sein. Sind hingegen viele der metrischen Regressoren von nichtlinearer Natur, ermöglichen *nicht- und semiparametrische Regressionsmodelle* die automatisierte Schätzung dieser Effekte. Ist die abhängige Größe binär, stehen z. B. sogenannte *generalisierte lineare Regressionsmodelle* (GLM) und zur zusätzlichen Berücksichtigung von nichtlinearen Effekten der Regressoren die *Generalisierte additive Regression* (GAMM) zur Verfügung. Soll die räumliche Anordnung der konkreten Daten oder deren Gruppierungen in der Modellierung berücksichtigt werden, stehen *Geoadditive*

[1] auch als *systematische Komponente* bekannt

oder *Gemischte Modelle* zur Verfügung. Eine Einführung in die hier aufgelisteten Regressionsmodelle findet sich z. B. in Fahrmeir et al. (2009).
Im Zuge dieser Arbeit wurde ausschließlich das klassische lineare Normalregressionsmodell behandelt.

In der *Regressionsanalyse* versucht man nun die systematische Komponente $\mathrm{f}(\boldsymbol{X})$ aus den konkreten Daten effizient zu schätzen und von der stochastischen Größe $\boldsymbol{\epsilon}$ zu trennen. Im Allgemeinen verfolgt man dabei folgende Ziele:

- eine möglichst genaue Beschreibung und Vorhersage der abhängigen Größe $\boldsymbol{y}$, und
- eine Quantifizierung der Beziehung zwischen der abhängigen und den unabhängigen Größen.

2.2 Das Lineare Regressionsmodell

Das bekannteste und grundlegendste Modell, das *lineare Regressionsmodell* (LM), „[...] *ist insbesondere dann sinnvoll einsetzbar, wenn die Zielvariable y stetig und wenn möglich approximativ normalverteilt ist.*“ (Fahrmeir et al., 2009, S.20). Es ist mit den folgenden Annahmen A1 bis A6 definiert, deren Gültigkeit im Zuge der Regressionsanalyse zu überprüfen ist:

A1 - Linearität Im LM wird die systematische Komponente $\mathrm{f}(\boldsymbol{X})$ als *Linearkombination* der Regressoren modelliert. Damit ist die Regressionsfunktion in (2.6) im LM wie folgt festgelegt

$$\begin{aligned} \mathrm{f}(\boldsymbol{X}) &= \boldsymbol{X}\boldsymbol{\beta}, \quad \mathrm{f} \in \mathbb{R}^n \\ \boldsymbol{X} &= \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times p} \quad \vee\, x_{ij} \in [a, b] \quad \text{mit} \\ \mathrm{rg}(\boldsymbol{X}) &= p \quad \text{und} \quad p = (k+1) \\ \boldsymbol{\beta} &= \begin{pmatrix} \beta_0 & \beta_1 & \dots & \beta_k \end{pmatrix}' \in \mathbb{R}^p \end{aligned} \tag{2.8}$$

Die $n \times p$ Matrix $\boldsymbol{X}$ bezeichnet man als *Designmatrix* und ihre i-te Zeile als *Designstelle*. Die Forderung ihres vollen Spaltenranges $\mathrm{rg}(\boldsymbol{X}) = p$, also der linearen Unabhängigkeit ihrer Spalten, ist eine Folgerung der verlangten Eindeutigkeit des unbekannten und effizient zu schätzenden p-dimensionalen Parametervektors $\boldsymbol{\beta}$ (vgl. 1.4 auf Seite 12). Der

Parameter β_0, als *Intercept* bezeichnet, dient als Modellkonstante, die eine eventuell vorhandene systematische Modellabweichung von Null berücksichtigt.
Die Voraussetzung der Linearkombination erscheint zunächst recht restriktiv. Einfache nichtlineare Beziehungen lassen sich aber durch geschickt gewählte Regressor- respektive Modelltransformationen dennoch beschreiben, solange (2.8) als Funktion des unbekannten p-dimensionalen Parametervektors $\boldsymbol{\beta}$ linear bleibt. Eine solche Auflistung möglicher linearisierbarer nichtlinearer Funktionen findet sich z. B. in Urban und Mayerl (2008, S.211ff) und in Fox (2008, Kap. 4.3). Diese einfachen Techniken haben allerdings den entscheidenden Nachteil, dass die prinzipielle funktionale Form der Beziehung bekannt sein muss. Wie im Laufe der Arbeit gezeigt wird, kann diese Erkenntnis z. B. aus den in 2.6 vorgestellten *Component-plus-residual plots* gewonnen werden, die gleichzeitig zur Überprüfung der Gültigkeit der Linearitätsannahme dienen.

A2 - Additivität Weiters wird von der Additivität der Störgröße ausgegangen. Das LM lässt sich mit den unter (2.8) eingeführten Bezeichnungen somit wie folgt anschreiben:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \in \mathbb{R}^p \tag{2.9}$$

Für $k > 1$, also mehreren Regressoren, spricht man von *multipler linearer Regression* und für den Fall $k = 1$ erhält man das *Standardmodell der linearen Einfachregression*.
Die Annahme der Additivität der Störgröße, ist in einer Vielzahl von praktischen Anwendungen zumindest annähernd erfüllt (vgl. Fahrmeir et al. (2009, S.60) und Nocedal und Wright (2006, S.249)). Selbst bei einem exponentiellen Wachstumsmodell, in welchem eine multiplikative Störgröße plausibler erscheint als eine additive, kann durch einfaches Logarithmieren ein lineares Modell mit additiver Störgröße gewonnen werden (vgl. auch hier Urban und Mayerl (2008, S.211ff) oder Fahrmeir et al. (2009, S.71)).

A3 - Mittelwert Gemäß (2.7) auf Seite 18 wird der Erwartungswert der Störgröße mit Null angenommen, $\mathbb{E}(\boldsymbol{\epsilon}) = 0$. Dies ist natürlich nur dann erfüllt, wenn eine eventuell vorhandene systematische Modellabweichung von Null bereits durch den Intercept berücksichtigt wird.

A4 - Homoskedastizität Die Varianzen der Störgröße werden als identisch angenommen, d. h. $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2$. Man spricht in diesem Fall von *Homoskedastizität*, andernfalls von *Heteroskedastizität*. Die Gültigkeit dieser Annahme kann mit den in 2.6 vorgestelltem *Residualplot* oder wieder mit dem *Component-plus-residual plot* überprüft werden.

A5 - Unkorreliertheit Darüber hinaus werden die Störgrößen ϵ_i im LM als unkorreliert ange-

nommen. Im einfachsten Fall autokorrelierter Störungen, dem autoregressiven Prozess erster Ordnung (AR(1)), besteht also für zwei aufeinander folgende Störungen ϵ_i und ϵ_{i-1} ein linearer Zusammenhang $\epsilon_i = \rho\epsilon_{i-1} + u_i$, mit $-1 \leq \rho \leq 1$ als zugehöriger Korrelationskoeffizient und u_i als unabhängige und identisch verteilte stochastische Größe. Diese Autokorrelation 1. Ordnung lässt sich grafisch z. B. mit Hilfe eines sogenannten *Lag-1 plot* oder rechnerisch mittels *Durbin-Watson Test* aufdecken (vgl. 2.6 auf Seite 38).

A6 - Normalverteilung Für Hypothesentests und zur Konstruktion von Konfidenzintervallen geht man davon aus, dass die Störgröße $\boldsymbol{\epsilon}$ zumindest approximativ normalverteilt ist[2], d. h. $\boldsymbol{\epsilon} \sim \Phi(\mathbf{0}, \sigma^2\boldsymbol{I})$

Aus diesen Definitionen ergeben sich für das LM zusammenfassend folgende Eigenschaften

$$\begin{aligned} \mathbb{E}(\boldsymbol{y}) &= \boldsymbol{X\beta} \\ \mathrm{Var}(y_i) &= \mathrm{Var}\left(\boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i\right) = \mathrm{Var}(\epsilon_i) = \sigma^2 \\ \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{y}) &= \mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}, \\ \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \sim \Phi(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}) \end{aligned} \tag{2.10}$$

2.3 Parameterschätzung

Zur Bestimmung des LM in (2.9) auf der vorherigen Seite müssen die unbekannten Parameter $\boldsymbol{\beta}$ und $\boldsymbol{\epsilon}$ noch optimal geschätzt werden. In diesem Kapitel sollen die dafür erforderlichen Formeln angeführt werden. Zur eindeutigen Kennzeichnung, dass es sich bei den Ergebnissen dieser Formeln nicht um die wahre Größe handelt, sondern um Schätzwerte, werden selbige mit einem Dach ˆ versehen. Da es sich bei der Lösung dieses Problems eigentlich um ein klassisches Optimierungsproblem handelt, sei an dieser Stelle auf die bereits erarbeiteten Grundlagen in 1.4 auf Seite 12 verwiesen. In diesem Kapitel werden daher nur mehr die für die Regression wesentlichen Schritte und Ergebnisse angeführt.

Gemäß dem berühmten *Gauß-Markov-Theorem* besitzen die *Kleinsten Quadrate*-Schätzer[3], KQ-Schätzer, in der speziellen Klasse der linearen und erwartungstreuen Schätzer minimale Varianzen und ermöglichen somit eine optimale Prognose der abhängigen Größe $\boldsymbol{y}$. Darüber hinaus kann man sich für die Bestimmung der KQ-Schätzer von der Einschränkung der normalverteilten stochastischen Komponente befreien. Gilt diese Annahme allerdings, lässt sich

[2] bei Gültigkeit dieser Annahme spricht man auch von der *klassischen linearen Normalregression.*
[3] auch als *Least-square* Schätzer, LS-Schätzer, oder *Best linear unbiased estimator* Schätzer, BLUE-Schätzer, bekannt

zeigen, dass die KQ-Schätzer für den unbekannten Regressionsparametervektor $\boldsymbol{\beta}$ mit den *Maximum-Likelihood-Schätzer* übereinstimmen (vgl. z. B. Fahrmeir et al., 2009, S.92f). Aus diesen Gründen findet die *Methode der kleinsten Fehlerquadratsumme* in der linearen Regressionsanalyse auch die meiste Beachtung.

Dem Prinzip dieser Methode folgend, erhält man eine Hyperebene als Schätzung des unbekannten Regressionsparametervektors $\boldsymbol{\beta}$ durch Minimierung der Fehlerquadratsumme. Naheliegenderweise definiert man den zu minimierenden Fehler, das sogenannte *Residuum* $\hat{\boldsymbol{\epsilon}}$, als Abweichung der wahren stochastischen Größe $\boldsymbol{y}$ von ihrem Schätzwert $\hat{\boldsymbol{y}}$

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{2.11}$$

Mit Hilfe der Residuen lässt sich die Fehlerquadratsumme nun wie folgt ausdrücken

$$\hat{\boldsymbol{\beta}} = \arg\min \frac{1}{2}\left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right] = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{2}\|\hat{\boldsymbol{\epsilon}}\|^2 = \frac{1}{2}\,\mathrm{SSE} \tag{2.12}$$

wobei mit $\|\cdot\|$ die euklidische Norm gemeint ist und SSE für die Residuenquadratensumme steht. Durch Nullsetzen der partiellen Ableitung von (2.12) nach $\boldsymbol{\beta}$ erhält man dann die bekannte *Normalgleichung*

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{2.13}$$

Unter der Annahme einer positiv definiten Designmatrix $\boldsymbol{X}$, welche gleichbedeutend mit der unter 2.2 auf Seite 20 angeführten Forderung des vollen Spaltenrangs ist, ist die Normalgleichung eindeutig lösbar. Für den geschätzten Parametervektor $\boldsymbol{\beta}$ ergibt sich somit die folgende eindeutige Lösung

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{2.14}$$

Ausgehend von diesem Schätzer für den wahren Parameter $\boldsymbol{\beta}$, lässt sich nun ein, im Sinne des Gauß-Markov-Theorems, optimaler, bedingter Erwartungswert der abhängigen Größe $\hat{\boldsymbol{y}}$ folgendermaßen schätzen

$$\mathbb{E}(\hat{\boldsymbol{y}}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{2.15}$$

Die $n \times n$ Matrix $\boldsymbol{H}$ bezeichnet man als *Prädiktionsmatrix* oder *hat-matrix*. Sie weist einige interessante Eigenschaften auf (vgl. Fahrmeir et al. (2009, S.93) und Xin und Xiao (2009, S.50f)):

- Die Prädiktionsmatrix $\boldsymbol{H}$ ist symmetrisch.

- Für die symmetrische Prädiktionsmatrix $\boldsymbol{H}$ gilt: $\boldsymbol{H}^2 = \boldsymbol{H}$. Allgemein bezeichnet man eine $n \times n$ symmetrische Matrix die diese Gleichheit erfüllt als *idempotent*.

- Für eine idempotente Matrix gilt, dass ihre Eigenwerte entweder 1 oder 0 sind.

- Vor allem im Kapitel der Modelldiagnostik 2.6 sind die Diagonalelemente h_{ii} der Prädiktionsmatrix, die sogenannten *Hebelwerte* oder *leverage*-Werte, von Bedeutung. Unter der Voraussetzung, dass alle Designstellen verschieden sind, ist ihr Wertebereich durch $\frac{1}{n} \leq h_{ii} \leq 1$ beschränkt.

- Die Matrix $\boldsymbol{I} - \boldsymbol{H}$ ist ebenfalls symmetrisch und idempotent mit Rang $n - p$

Mit Hilfe der Prädiktionsmatrix $\boldsymbol{H}$ lassen sich die Residuen auch wie folgt beschreiben

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y} \tag{2.16}$$

Wie in 1.4 bereits ausgeführt, kann die Lösung der Normalgleichung (2.13) auf der vorherigen Seite mittels Cholesky Faktorisierung zu numerischen Ungenauigkeiten führen. Eine Verbesserung schafft hier die QR-Faktorisierung der Designmatrix mit einer orthogonalen Matrix $\boldsymbol{Q}$ und einer oberen Dreiecksmatrix $\boldsymbol{R}$. Im Folgenden werden daher auch die Schätzwerte bei Verwendung der QR-Faktorisierung angegeben. Auch, oder insbesondere, da die in dieser Arbeit zur Parameterbestimmung und Auswertung zahlreicher statistischer Methoden verwendete `regstats`-Funktion der MATLAB® -Umgebung eine QR-Faktorisierung durchführt.
Ausgehend von der Normalgleichung führt eine QR-Faktorisierung der Designmatrix $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ zu folgendem geschätzten Parametervektor $\hat{\boldsymbol{\beta}}$

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{R}\hat{\boldsymbol{\beta}} &= \boldsymbol{R}'\boldsymbol{Q}\boldsymbol{y} \\ \boldsymbol{R}\hat{\boldsymbol{\beta}} &= \boldsymbol{Q}\boldsymbol{y} \\ \hat{\boldsymbol{\beta}} &= \boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{y} \end{aligned} \tag{2.17}$$

Für die Schätzwerte der bedingten abhängigen Größe $\hat{\boldsymbol{y}}$ und der Residuen $\hat{\boldsymbol{\epsilon}}$ gilt dann

$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{y}}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{y} \quad \text{und} \\ \hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \left(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}'\right)\boldsymbol{y} \end{aligned}$$

2.4 Statistische Eigenschaften der Schätzer

Ausgehend von dem in den bisherigen Regressionskapiteln 2.1 bis 2.3 auf Seiten 17–21 aufgebauten stochastischen Modell, sollen nun die darin geschätzten stochastischen Parameter mit Mitteln der klassischen parametrischen Inferenzstatistik untersucht werden. Dazu gehören die Angabe ihrer Erwartungswerte, statistische Tests und Bereichsschätzungen.

KQ-Schätzer Wie in 2.3 bereits erwähnt, sind die KQ-Schätzer *unverzerrt*, d. h. der Erwartungswert der Regressionskoeffizienten fällt mit dem zu schätzenden Parametervektor zusammen

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{y}) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Für die geschätzte Kovarianzmatrix der Regressionskoeffizienten gilt (Fahrmeir et al., 2009, S.101f)

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \hat{\sigma}^2\boldsymbol{R}^{-1}\left(\boldsymbol{R}'\right)^{-1} \tag{2.18}$$

Da die wahre *Modellvarianz*, die mit der Varianz σ^2 der Störgröße $\boldsymbol{\epsilon}$ zusammenfällt, unbekannt ist, muss diese ebenso geschätzt werden mit (Xin und Xiao, 2009, S.60f)

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \tag{2.19}$$

Diesen Schätzer bezeichnet man als *Restringierter Maximum-Likelihood-Schätzer* (REML). Die Wurzel aus (2.19) liefert die *Modellstandardabweichung* $\hat{\sigma}$.
Entsprechend der Definition der Kovarianzmatrix handelt es sich bei ihren Diagonalelementen um die geschätzten Varianzen der Regressionskoeffizienten. Wie (2.18) zeigt, sind diese direkt proportional zur Modellvarianz σ^2. Je kleiner also die Modellvarianz ist, desto kleiner fallen auch die Varianzen der Regressionskoeffizienten aus. Weiters lässt sich zeigen (vgl. Fahrmeir et al., 2009, S.101), dass größere Varianzen der Regressoren oder geringere Kollinearitäten derselbigen ebenso die Varianzen der Regressionskoeffizienten minimieren. Insbesondere bei geplanten Experimenten versucht man daher, durch die Wahl möglichst orthogonaler Designs geringe lineare Abhängigkeiten der Regressoren zu erzielen.

Residuen Der Erwartungswert der in (2.11) eingeführten Residuen als Schätzer der wahren stochastischen Komponente $\boldsymbol{\epsilon}$, sollte gemäß (2.7) auf Seite 18 Null sein; wie die nachfolgende

Rechnung auch bestätigt (Fahrmeir et al., 2009, S.107)

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\epsilon}}) = \mathbb{E}\left((\boldsymbol{I}) - \boldsymbol{H})\boldsymbol{y}\right) = \mathbb{E}(\boldsymbol{y}) - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{y}) = \\ = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}\end{aligned} \tag{2.20}$$

Unter Beachtung der Eigenschaften Symmetrie und Idempotenz der Matrix $(\boldsymbol{I} - \boldsymbol{H})$[4] gilt für die Kovarianzmatrix der Residuen allerdings

$$\mathrm{Cov}(\hat{\boldsymbol{\epsilon}}) = \mathrm{Cov}\left((\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}\right) = (\boldsymbol{I} - \boldsymbol{H})\,\mathrm{Var}(\boldsymbol{y})\,(\boldsymbol{I} - \boldsymbol{H})' = (\boldsymbol{I} - \boldsymbol{H})\,\sigma^2 \tag{2.21}$$

Mit den beiden Eigenschaften aus (2.20) und (2.21) lässt sich für die Residuen somit folgendes feststellen:

- Die Residuen sind wie die wahren Störgrößen im Mittel Null.
- Sie sind aber im Gegensatz zu den wahren Störgrößen *korreliert*.
- Und besitzen im Gegensatz zu den wahren Störgrößen *heteroskedastische Varianzen*, wobei die Varianz umso kleiner ist, je größer der Hebelwert h_{ii} ist.

Statistische Tests und Bereichsschätzungen Für die nun folgenden statistischen Tests und Bereichsschätzungen sei die Normalverteilungsannahme der Störgröße $\epsilon_i \sim \Phi(0, \sigma^2)$ als zutreffend vereinbart. Mit Hilfe des *Zentralen Grenzverteilungssatzes* (vgl. Viertl (1990, S.84) und Fahrmeir et al. (2009, S.105f)) lässt sich jedoch zeigen, dass die Aussagen dieser statistischen Tests und Bereichsschätzungen für annähernd viele Beobachtungen n, mit $n \to \infty$, gültig bleiben. Darüber hinaus sind die Tests relativ robust gegenüber geringen Abweichungen von der Normalverteilung.

Mittels sogenannter Modellwahlkriterien wird in 2.5 auf Seite 29 gezeigt werden, dass die Aufnahme aller verfügbarer Regressoren in die stochastische Modellierung selbige nicht zwingend verbessert. *Statistische Tests* können hier bereits helfen, fundierte Entscheidungen über die Signifikanz eines Regressors im stochastischen Modell zu treffen. Dazu formuliert man für jeden einzelnen Regressionskoeffizienten eine sogenannte *Nullhypothese* $\mathscr{H}_0$ die dessen Merkmalraum in zwei Teilmengen zerlegt, z. B. $\mathscr{H}_0 : \beta_i = 0$. Diese Nullhypothese besagt, dass die Aufnahme des i-ten Regressors das Regressionsmodell nicht verbessert und der Regressor somit ausgeschlossen werden kann. Man erhielte also ein *reduziertes Regressionsmodell*, im Gegensatz zum *vollen Regressionsmodell* in dem alle möglichen Regressoren enthalten sind.

[4] also der Gültigkeit von $(\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

Die *Alternativhypothese* lautet dann $\mathscr{H}_1 : \beta_i \neq 0$. Eine *Testfunktion* ordnet nun jeden Regressorkoeffizienten der Nullhypothese oder der Alternativhypothese zu. Ein Regressorkoeffizient der die Alternativhypothese stützt, führt zu einer *Verwerfung* der Nullhypothese.
Für dieses Testproblem gibt es grundsätzlich zwei Fehlerarten:

Fehler 1. Art Die Nullhypothese ist zwar richtig, wird durch die Testfunktion aber verworfen. Die Wahrscheinlichkeit für diesen Fehler, das sogenannte *Signifikanzniveau*, bezeichnet man mit α. Üblicherweise wird α mit $\alpha = 5\%$ festgelegt.

Fehler 2. Art Die Annahme einer falschen Nullhypothese.

Die in der Regressionsanalyse verwendete Testfunktion für allgemeine lineare Nullhypothesen der Form $\mathscr{H}_0 : \beta_i = 0$ ist die F-Teststatistik (Fahrmeir et al., 2009, S.113ff)

$$F = \frac{n-p}{r} \frac{\Delta \, \mathrm{SSE}}{\mathrm{SSE}} \sim F_{r,n-p} \tag{2.22}$$

r entspricht hier der Anzahl der linearen Restriktionen der Form $\mathscr{H}_0 : \beta_i = 0$ mit $i, r \in \{\mathbb{N} \mid i = [1, r],\ r \leq p\}$ und SSE wieder der in (2.12) auf Seite 22 eingeführten Residuenquadratensumme. In dieser Teststatistik betrachtet man also die relative Differenz der Residuenquadratensumme zwischen dem restringierten und dem vollen Modell. Je größer diese Differenz ist, desto eher wird die Nullhypothese verworfen, d. h. desto signifikanter sind die Regressoren. Die Teststatistik in (2.22) ist unter $\mathscr{H}_0$ F-verteilt mit r und $n-p$ Freiheitsgraden. Die Nullhypothese wird nun zu einem Signifikanzniveau α verworfen, falls die Teststatistik größer ist, als das $(1-\alpha)$-Quantil der F-Verteilung

$$F > F_{r,n-p}(1-\alpha)$$

Zwei in der Regressionsanalyse wesentliche Teststatistiken sollen nun ein wenig näher betrachtet werden (vgl. Fahrmeir et al. (2009, S.115ff) und Pruscha (2006, S.108ff)).

Partielle F-Test Der *partielle F-Test* testet die Signifikanz eines einzelnen Regressionskoeffizienten mit der Nullhypothese $\mathscr{H}_0 : \beta_i = 0$ für $i = 1, \ldots, p$. In diesem Fall gilt für die Teststatistik

$$F_k = \frac{\hat{\beta}_i^2}{\mathrm{Var}(\hat{\beta}_i)} \sim F_{1,n-p}$$

Die Teststatistik F_k wird auch mit *F-to-enter* bezüglich der Regressorvariable $\boldsymbol{x}_k$ bezeichnet. Dieser Wert wird in den sogenannten Screening-Techniken in 2.5 auf Seite 29 wesentlich werden.

Statt der F-Teststatistik kann für die Testfunktion auch die t-Teststatistik verwendet

werden. Hier wird die Nullhypothese verworfen, falls der Betrag der Teststatistik größer ist, als das $(1-\frac{\alpha}{2})$-Quantil einer Studenten-Verteilung t_{n-p} mit $n-p$ Freiheitsgraden

$$|t| = \frac{\hat{\beta}_i}{\mathrm{Var}(\hat{\beta}_i)^{\frac{1}{2}}} > t_{1-\frac{\alpha}{2},n-p} \tag{2.23}$$

mit $\mathrm{Var}(\hat{\beta}_i)^{\frac{1}{2}}$ als *Standardabweichung* von $\hat{\beta}_i$.
Statt einem Größenvergleich zwischen den Werten der Testfunktion und dem F-Test oder t-Test auszugeben, wird allerdings üblicherweise das zur Testfunktion gehörende minimale Signifikanzniveau α_p ausgegeben. Unter Beachtung der Symmetrie der Dichtefunktion $f(x|n)$ der Studenten-Verteilung (vgl. Viertl, 1990, S.124) ergibt sich das minimale Signifikanzniveau α_p zu

$$\begin{aligned} 1 - \frac{\alpha_p}{2} &= f(t, n-p) \\ \frac{\alpha_p}{2} &= f(-|t|, n-p) \\ \alpha_p &= 2 * f(-|t|, n-p) \end{aligned} \tag{2.24}$$

Globale F-Test Der *globale F-Test* testet gleichzeitig die Signifikanz *aller* möglichen Regressionskoeffizienten mit der Nullhypothese $\mathscr{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Hier wird also unterstellt, dass keiner der Regressoren $\boldsymbol{x}$ einen Einfluss auf die Zielvariable $\boldsymbol{y}$ besitzt. Nur wenn diese Nullhypothese verworfen wird, kann überhaupt von einer linearen Abhängigkeit ausgegangen werden! Die Alternativhypothese bedeutet allerdings nicht, dass alle Regressoren signifikant sind, sondern lediglich, dass mindestens einer der k-Regressoren einen Einfluss besitzt.
Bevor nun die Teststatistik angegeben wird, werden zunächst die Bezeichnungen der *fundamentalen Streuungszerlegungsformel* für die empirischen Varianzen der abhängigen Größe $\boldsymbol{y}$ eingeführt. Die Steuungszerlegung ist gegeben durch (Fahrmeir et al., 2009, S.98)

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \mathrm{SSE} = \\ \mathrm{SST} = \mathrm{SSR} + \mathrm{SSE} \quad | \quad \times \frac{1}{n-1} \\ s^2_{\boldsymbol{y}} = s^2_{\hat{\boldsymbol{y}}} + s^2_{\hat{\boldsymbol{\epsilon}}} \end{aligned} \tag{2.25}$$

mit $\bar{y}$ als arithmetisches Mittel der abhängigen Größe $\boldsymbol{y}$. Mit diesen Bezeichnungen folgt

für die Teststatistik

$$F = \frac{n-p}{k}\frac{\text{SSR}}{\text{SSE}} = \frac{n-p}{k}\frac{R^2}{1-R^2} \tag{2.26}$$

mit R als sogenanntes Bestimmtheitsmaß (näheres siehe 2.5 auf Seite 29). Mit Hilfe der in (2.25) eingeführten Streuungsformel lassen sich nun die relevanten Größen des globalen F-Tests in der viel zitierten *analysis of variance* (ANOVA) Tafel angeben (vgl. Tab. 2.1 auf dieser Seite).

Sogenannte Bereichsschätzungen, oder *Konfidenzintervalle*, sollen für eine stochastische Größe jene Teilmengen ihres Parameterraumes[5] quantifizieren, in denen der wahre Parameter mit großer, vorgegebener Wahrscheinlichkeit liegt. In linearen Regressionsmodellen sind diese Konfidenzintervalle zum Signifikanzniveau $(1-\alpha)$ grundsätzlich folgendermaßen aufgebaut

$$\hat{Y} - S\,\text{Var}(\hat{Y})^{\frac{1}{2}} \leq Y \leq \hat{Y} + S\,\text{Var}(\hat{Y})^{\frac{1}{2}} \tag{2.27}$$

Mit S als vorgegebenes $(1-\alpha)$-Quantil einer geeigneten Verteilung. Üblicherweise verwendet man mit $\alpha = 0.05$ ein 95%-Konfidenzintervall.
Ein Konfidenzintervall für z. B. β_i ist gegeben durch (Fahrmeir et al., 2009, S.121)

$$\left[\hat{\beta}_i - t_{n-p,1-\frac{\alpha}{2}}\,\text{Var}(\hat{\beta}_i)^{\frac{1}{2}} \quad \hat{\beta}_i + t_{n-p,1-\frac{\alpha}{2}}\,\text{Var}(\hat{\beta}_i)^{\frac{1}{2}}\right]$$

Überdeckt dieses Konfidenzintervall eines Parameters β_i nicht die Null, so muss auch der statistische Test in (2.23) auf der vorherigen Seite die Nullhypothese $\beta_i = 0$ verwerfen.
Neben der unter (2.15) auf Seite 22 eingeführten optimalen Punktschätzung für eine zukünftige Beobachtung $\boldsymbol{x}_0$, könnte man auch an einem Konfidenzintervall für dessen Erwartungswert, oder einem *Prognoseintervall* interessiert sein. Das Prognoseintervall soll dabei den Wert der zukünftigen Beobachtung $\boldsymbol{x}_0$ mit einer Wahrscheinlichkeit von $(1-\alpha)$ überdecken und ist wie

[5] der *Parameterraum* beinhaltet alle möglichen Parameter

Variationsursache	SS	df	F	p
Regressoren	SSE	k	$\frac{n-p}{k}\frac{\text{SSR}}{\text{SSE}}$	1 - $F_{k,n-p}$(F)
Residuen	SSR	$n-p$		
Total	SST	$n-1$		

Tabelle 2.1: ANOVA-Tafel des globalen F-Tests

folgt definiert (Fahrmeir et al., 2009, S.122)

$$\begin{bmatrix} \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}(1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{\frac{1}{2}} & \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}(1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{\frac{1}{2}} \\ \left[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}(1+\boldsymbol{x}_0'(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{x}_0)^{\frac{1}{2}} & \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}(1+\boldsymbol{x}_0'(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{x}_0)^{\frac{1}{2}}\right] \end{bmatrix} \tag{2.28}$$

Dieses Intervall ist allerdings nur für eine individuelle zukünftige Beobachtung $\boldsymbol{x}_0$ gültig. Zur gleichzeitigen Schätzung mehrerer zukünftiger Beobachtungen, ist die Konstruktion eines sogenannten *simultanen* Konfidenzintervalls erforderlich. Dies ist z. B. mit Hilfe des *Scheffé-schen Quantil* oder der etwas konservativen *Bonferroni-Korrektur* möglich. Bei Verwendung des Schefféschen Quantil S wird dieses anstelle der Studentenverteilung $t_{n-p,1-\frac{\alpha}{2}}$ in (2.28) eingesetzt (Pruscha, 2006, S.115)

$$S = \sqrt{p \cdot F_{p,\, n-p,\, 1-\alpha}}$$

Die Bonferroni-Korrektur teilt das Konfidenzniveau jedes einzelnen Parameters α durch die Anzahl der betrachteten Parameter n, sodass die Vereinigung der Konfidenzregionen dieser Parameter wiederum das gewünschte Konfidenzniveau $(1-\alpha)$ ergeben. Anstelle des $(1-\frac{\alpha}{2})$-Quantils verwendet man in (2.28) also das $(1-\frac{\alpha}{2n})$-Quantil für $\boldsymbol{x}_0 \in \mathbb{R}^n$.

2.5 Modellauswahl

Wie zur Einleitung der statistischen Tests in 2.4 bereits erwähnt wurde, geht mit der Aufnahme aller verfügbaren Regressoren in das LM nicht zwingend eine Verbesserung desselbigen einher. In diesem Kapitel soll dies nun näher erläutert werden. Aufbauend auf der klassischen Bias-Varianz-Trade Off Definition werden einige Kriterien aufgezählt, die danach trachten den Trade Off zu minimieren. Anschließend werden einfache automatisierte Verfahren beschrieben, die von diesen Kriterien Gebrauch machen um ein, in deren Sinne, möglichst gutes LM zu konstruieren.

Im Folgenden wird davon ausgegangen, dass für die Konstruktion eines LM eine Menge $\mathscr{K}$ mit k potentiellen Regressoren $\{x_1, \ldots, x_k\} \in \mathscr{K}$ zur Verfügung steht. Schätzer eines LM, welches aus einer beliebigen Teilmenge A aus $\mathscr{K}$, $A \subseteq \mathscr{K}$, mit a Regressoren konstruiert wird, werden mit dem Teilmengensymbol A im Index versehen, z. B. $\hat{\boldsymbol{y}}_A$.
Wird anstelle des korrekt spezifizierten Modells $K \subset \mathscr{K}$ ein davon verschiedenes Teilmodell $A \subset \mathscr{K}$ zur Modellschätzung verwendet, hat dies auf die Modelleigenschaften folgende Auswirkungen (vgl. Fahrmeir et al. (2009, S.157ff) und Xin und Xiao (2009, Kap.5)):

Regressionsparameter

- Das korrekt spezifizierte Modell verlangt alle k-Regressoren, die abhängige Größe wird allerdings mit einem Teilmodell mit a-Regressoren geschätzt. Dies führt grundsätzlich zu verzerrt geschätzten Regressionsparametern $\hat{\boldsymbol{\beta}}_A$. Allerdings ist es auch möglich, dass die Varianzen der Regressionsparameter des fehlspezifizierten Modells kleiner sind, als jene des korrekt spezifizierten Modells, d. h. $\mathrm{Var}(\hat{\boldsymbol{\beta}}_A) < \mathrm{Var}(\hat{\boldsymbol{\beta}})$.

- Das korrekt spezifizierte Modell verlangt nur die Teilmenge A der potentiellen Regressoren, die abhängige Größe wird allerdings mit einem vollen Modell aus allen k-Regressoren geschätzt. In diesem Fall sind zwar die geschätzten Regressionsparameter des fehlspezifizierten Modells unverzerrt, allerdings gilt für deren Varianzen nun ebenso $\mathrm{Var}(\hat{\boldsymbol{\beta}}_A) < \mathrm{Var}(\hat{\boldsymbol{\beta}})$.
  Man verliert also mit der Aufnahme irrelevanter Regressoren an Schätzgenauigkeit!

Prognosegüte Ähnliche Aussagen lassen sich auch zur Prognosegüte machen. Wie in 2.3 bereits beschrieben, dienen Regressionsmodelle auch der Prognose von zukünftigen Beobachtungen y_{n+i}. Unter Verwendung der in der Statistik „widely used“(vgl. Xin und Xiao, 2009, S.164) Summe der *erwarteten quadrierten Abweichungen*[6] (MSE), lässt sich zunächst die Qualität des Schätzers $\hat{y}_A$ quantifizieren

$$\begin{aligned}
\mathrm{MSE} &= \sum_{i=1}^{n} \mathbb{E}(\hat{y}_{i,A} - \mathbb{E}(y_i))^2 \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\hat{y}_{i,A} - \mathbb{E}(\hat{y}_{i,A}) + \mathbb{E}(\hat{y}_{i,A}) - \mathbb{E}(y_i)\right]^2 \\
&= \sum_{i=1}^{n} \mathbb{E}\left(\hat{y}_{i,A} - \mathbb{E}(\hat{y}_{i,A})\right)^2 + \sum_{i=1}^{n} \left(\mathbb{E}(\hat{y}_{i,A}) - \mathbb{E}(y_i)\right]^2 \\
&= a\,\sigma^2 + \sum_{i=1}^{n} \left(\mathbb{E}(\hat{y}_{i,A}) - \mathbb{E}(y_i)\right)^2
\end{aligned} \tag{2.29}$$

Je kleiner der MSE, desto besser also der Schätzwert. Mit Hilfe dieses Gütemaßes lässt sich nun auch die Qualität des *erwarteten quadrierten Prognosefehlers* (SSPE) quantifizieren (Fahrmeir

[6] im engl. als *mean square error* bezeichnet

et al., 2009, S.158)

$$\begin{aligned}\text{SSPE} = \sum_{i=1}^{n} \mathbb{E}\left(y_{n+i} - \hat{y}_{i,A}\right)^2 &= \sum_{i=1}^{n} \mathbb{E}\left(y_{n+i} - \mathbb{E}(y_i)\right)^2 + \sum_{i=1}^{n} \mathbb{E}(\hat{y}_{i,A} - \mathbb{E}(y_i))^2 \\ &= n\,\sigma^2 + a\,\sigma^2 + \sum_{i=1}^{n} \left(\mathbb{E}(\hat{y}_{i,A}) - \mathbb{E}(y_i)\right)^2\end{aligned}$$

Die Herleitung erfolgt ähnlich der in (2.29).
Der SSPE besteht also aus den folgenden drei Termen:

1. Dem *irreduziblen Prognosefehler* $n\,\hat{\sigma}^2$ der sich selbst der Beeinflussung von Modellauswahlverfahren entzieht.

2. Der *Summe der Varianzen* $a\,\sigma^2$ die umso kleiner wird, je weniger Regressoren in das Modell aufgenommen werden. Dieser Term ist durch Modellauswahlverfahren beeinflussbar.

3. und dem *quadrierten Bias*. Dieser Term ist ebenso beeinflussbar und wird umso kleiner je mehr Regressoren in das Modell aufgenommen werden.

In den letzten beiden Aufzählungspunkten ist der *klassische Bias-Varianz-Trade Off* bereits deutlich erkennbar: Kleinere Teilmodelle $A \subset \mathscr{K}$ verringern zwar die Prognosevarianzen, sie erhöhen allerdings den Bias der Regressionskoeffizienten β_i und der Prognose $\hat{y}_i$. Je eher das Teilmodell an das volle Modell $\mathscr{K}$ herankommt, desto kleiner wird nun der Bias, desto größer werden im Gegenzug aber wieder die Varianzen.
Die im nächsten Abschnitt behandelten Modellwahlkriterien sollen nun helfen, innerhalb dieses Widerspruchs das beste LM auszuwählen.

2.5.1 Einige Modellwahlkriterien

Modellwahlkriterien dienen allgemein zur Messung der Güte der Anpassung verschiedener LM an die gleichen konkreten Daten. Die meisten in diesem Kapitel vorgestellten Modellwahlkriterien versuchen dazu den in 2.5 definierten Bias-Varianz-Trade Off zu minimieren, indem sie komplexere Modelle stärker bestrafen als einfache.

Bei einer großen Anzahl konkreter Daten unterscheidet man diese Kriterien in asymptotisch effiziente und konsistente. Das Ziel *asymptotisch effizienter Kriterien* liegt in der Wahl jenes Modells mit dem geringsten MSE, während *konsistente Kriterien* das wahre Modell mit Wahrscheinlichkeit eins auswählen. Zu letzterer Gruppe gehört nur das hier vorgestellte Bayesiani-

sche Informationskriterium. Die anderen hier angeführten sind den asymptotische effizienten Kriterien zuzuordnen (vgl. Xin und Xiao, 2009, S.170).

Bestimmtheitsmaß Zu den sehr einfachen Kriterien zählt das Bestimmtheitsmaß[7] R^2, welches mit Hilfe der in (2.25) auf Seite 27 eingeführten fundamentalen Streuungszerlegungsformel wie folgt definiert ist (Fahrmeir et al., 2009, S.98)

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \quad \text{mit} \quad 0 \leq R^2 \leq 1 \tag{2.30}$$

Das Bestimmtheitsmaß beschreibt den Anteil der durch die Regressoren erklärten Varianz an der Gesamtvarianz der abhängigen Größe y. Je näher es bei eins liegen soll, desto kleiner muss die Residuenquadratensumme SSE sein und desto besser muss sich das Modell somit den konkreten Daten anpassen. Im Extremfall $R^2 = 1$ müssen alle Werte der abhängigen Größe auf einer p-dimensionalen Hyperebene des $\mathbb{R}^p$ liegen. Umgekehrt gilt jedoch, dass ein kleineres Bestimmtheitsmaß auf ein schlechtes LM hindeutet. Im Extremfall $R^2 = 0$ muss SSR identisch 0 gelten, d. h. sämtliche Prognosen der abhängigen Größe y_i sind gleich ihrem Erwartungswert $\mathbb{E}(y)$. Daraus folgert, dass die Regressoren keinerlei Erklärungsgehalt für $\boldsymbol{y}$ besitzen. Vor dieser Schlussfolgerung muss allerdings noch eine Fehlspezifikation des verwendeten Modells ausgeschlossen werden.
Obgleich seiner einfachen Interpretierbarkeit, eignet sich das Bestimmtheitsmaß nur bedingt bis gar nicht zum Vergleich verschiedener LM. Insbesondere da es mit jedem neu aufgenommenen Regressor automatisch größer wird. Dieser Umstand sollte mit dem *korrigierten Bestimmtheitsmaß* vermieden werden (Pruscha, 2006, S.119)

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p}(1 - R^2) = 1 - \frac{\hat{\sigma}^2}{\text{SST}} \tag{2.31}$$

mit $\hat{\sigma}^2$ als REML gemäß (2.19) auf Seite 24. Mit dieser Korrektur erhöht sich der R^2_{adj} eines Modells nicht notwendigerweise bei der Aufnahme neuer Regressoren. Da allerdings auch für dieses Kriterium gezeigt werden kann, dass es selbst bei der Aufnahme eines Regressors mit einem p-Fraktil von circa 0.3 ansteigen würde[8], wird von dessen Verwendung abgeraten (vgl. Fahrmeir et al. (2009, S.161) und Xin und Xiao (2009, S.166)).

[7] auch als *coefficient of determination* (COD) bekannt
[8] Wie auf auf Seite 27 beschrieben, würden somit mit relativ hoher Wahrscheinlichkeit auch Regressoren in das Modell aufgenommen werden, die eigentlich keinen signifikanten Erklärungsgehalt für die abhängigen Variable aufweisen. Ein Umstand der mit dem in 2.5 beschriebenen Bias-Varianz-Trade Off nicht wünschenswert ist.

Informationskriterium nach Akaike Das aus der Informationstheorie von Akaike hergeleitete Kriterium, hat für das klassische LM die folgende Form (Miller, 2002, S.156)

$$\mathrm{AIC} = n \log(\hat{\sigma}^2) + 2p \tag{2.32}$$

wobei hier für $\hat{\sigma}^2$ nicht der REML eingesetzt wird, sondern der verzerrte Maximum-Likelihood-Schätzer $\frac{1}{n}\,\mathrm{SSE}$. Der erste Teil quantifiziert die Anpassung des Modells an die konkreten Daten, während der zweite die Modellkomplexität A bestraft. Ein gutes LM ist also durch ein kleines AIC gekennzeichnet.
Da dieses Kriterium für weniger Beobachtungen n, eher dazu tendiert größere LM zu akzeptieren, bevorzugt man in diesen Fällen das unverzerrte AIC_c Kriterium (Miller, 2002, S.157)

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1} \tag{2.33}$$

Bayesianische Informationskriterium Das Bayesianische Informationskriterium, oder auch *Schwartz-Kriterium* genannt, hat bei Normalverteilungsannahme folgende Gestalt (Fahrmeir et al., 2009, S.162)

$$\mathrm{BIC} = n \log(\hat{\sigma}^2) + \log(n)p$$

Obwohl die beiden Informationskriterien BIC und AIC sehr ähnlich aufgebaut sind, bestraft das BIC komplexere Modelle stärker als das AIC. Wieder ist ein gutes LM durch ein kleines BIC gekennzeichnet.

Im Zuge dieser Arbeit wurde das Bestimmtheitsmaß R^2 und dessen korrigierte Version R^2_{adj} lediglich informativ ausgegeben. Zur Modellauswahl wurden die beiden Varianten des AIC herangezogen.
Grundsätzlich wird empfohlen, bei wenigen vorhandenen Beobachtungen der abhängigen Größe $\boldsymbol{y}$ das AIC_c-Kriterium, für eine etwas größer Anzahl das AIC-Kriterium und für eine große Anzahl von Beobachtungen das BIC-Kriterium zu verwenden (vgl. Xin und Xiao, 2009, S.170).

2.5.2 Selektionsverfahren

Selektionsverfahren dienen der Auswahl und Aufnahme von Regressoren in ein Modell nach bestimmten Kriterien. Im Zusammenhang mit einer Regressionsanalyse muss man allerdings aufpassen, dass man dadurch die Datenbasis erheblich modifizieren kann und sich somit vom ursprünglichen, auf theoretischen Zusammenhang zu überprüfenden, Regressionsmodell entfernt. Der Ausschluss eines Regressors oder einer Designstelle sollte daher stets überdacht

werden (vgl. z. B. Urban und Mayerl (2008, S.184) und Xin und Xiao (2009, S.180)). Diese Warnung trifft allerdings nicht nur auf die in diesem Abschnitt behandelten Selektionsverfahren zu, sondern auch auf Maßnahmen die im Zuge der in 2.6.4 vorgestellten Modelldiagnostiken möglich sind.

Die hier vorgestellten Selektionsverfahren lassen sich in zwei Gruppen einteilen

- Best-Subset-Selection und
- Stepwise-Selection.

In der *Best-Subset-Selection* versucht man mit Hilfe von einem der in 2.5.1 auf Seite 31 vorgestellten Kriterien, aus allen möglichen Kombinationen der k-Regressoren das im Sinne dieses Kriteriums beste LM auszuwählen. Schließt man das Nullmodell aus, welches nur den Intercept beinhalten würde, müssen somit $(2^k - 1)$ verschiedene Kombinationen berücksichtigt werden. Im Falle von sieben Regressoren, wie in dieser Arbeit, wären dies bereits 127 zu untersuchende Modelle. Man kann also leicht erkennen, dass diese Art der Vorgehensweise nur bei einer moderaten Anzahl an Regressoren vertretbar ist.
Um dennoch mit vertretbarem Rechenaufwand bei einer etwas höheren Anzahl an Regressoren eine Best-subset-selection durchführen zu können, wurden von Furnival und Wilson (1974) eine besonders effiziente *leaps-and-bound*-Technik entwickelt. Diese 1974 vorgestellte Arbeit, wurde von Gunst (2000) als eine der revolutionärsten Arbeiten im Gebiet der Regression gewürdigt. Da der Algorithmus einerseits die Berechnung aller kombinatorischen Modelle vermeidet und andererseits die zahlreichen Normalgleichung (2.13) besonders handhabt, kann er bis zu 40 (vgl. Fahrmeir et al., 2009, S.164) potentielle Regressoren berücksichtigen – immerhin wären dies bereits $1.1 \cdot 10^{12}$ mögliche Modelle. Wenn auch beabsichtigt, konnte der leaps-and-bounds Algorithmus im Zuge dieser Arbeit aus Zeitmangel leider nicht mehr implementiert werden.

Bei einer großen Anzahl an potentiellen Regressoren besteht die Möglichkeit, mit Hilfe sogenannter *Variablen-Screening* Techniken zunächst ein kleineres Submodell herauszufiltern. Insbesondere für diese Vorgehensweise sollen die im Folgenden vorgestellten *stepwise-selection* Verfahren geeignet sein. Im Hinblick auf die eingangs erwähnt Problematik, empfehlen Xin und Xiao (2009, S.180) unterschiedliche Verfahren zu kombinieren und wirklich nur jene Regressoren aus dem LM auszuschließen, die auch von allen verwendeten Techniken ausgeschlossen wurden.
All den hier vorgestellten Screening Techniken ist gemein, dass sie in jedem Iterationsschritt gemäß einem Kriterium entweder einen Regressor in das Modell aufnehmen und/oder einen im Modell bereits vorhandenen Regressor ausschließen. In diesem Zusammenhang ist erwähnenswert, dass die in einem Iterationsschritt zu vergleichenden Modelle alle gleich komplex

sind und somit ausnahmslos alle in 2.5.1 vorgestellten Modellwahlkriterien als Gütemaß gültig sind. Um die Submodelle der jeweiligen Iterationsschritte auch später untereinander vergleichen zu können, empfiehlt es sich jedoch ein globales Gütemaß wie das AIC oder das BIC zu verwenden. Die Screening Techniken brechen den Iterationsprozess ab, sobald durch Aufnahme und/oder Ausschluss eines Regressors, das LM im Sinne des gewählten Kriteriums nicht mehr verbessert werden kann. Der dadurch gewonnene Regressorensatz ist im Allgemeinen zwar nicht der Beste im Sinne des Modellwahlkriteriums, aber dennoch ein guter. Auch wenn eventuell vermutet, lässt die Reihenfolge der Aufnahme und/oder des Ausschlusses der Regressoren ab dem zweiten Iterationsschritt im Allgemeinen keinen Rückschluss auf die Rangfolge der Regressoren untereinander zu.

Vorwärts-Selektion Ausgehend von einem Startmodell[9] $A \subset \mathscr{K}$ wird für alle $(k - a_i)$ verbleibenden Regressoren jeweils ein Regressionsmodell geschätzt. In jedem Iterationsschritt i wird nun jener Regressor in das Startmodell A_i aufgenommen, der die größte Reduktion des Modellwahlkriteriums liefert. Der Algorithmus bricht ab, sobald die Aufnahme einer der $(k-a_i)$ verbleibenden Regressoren, keine weitere Reduktion im Vergleich zum Iterationsschritt $(i-1)$ bringen würde. Ein diesem Problem anhaftendes Problem liegt darin, dass ein einmal in das Modell aufgenommener Regressor, auch auf jeden Fall in diesem bleibt; selbst wenn in einem späteren Iterationsschritt ein Regressor aufgenommen werden sollte, der mit diesem hoch korreliert wäre.
Der Vorwärts Algorithmus muss also höchstens $k(k+1)/2$ Modelle schätzen, bis er in jedem Fall abbricht.

Rückwärts-Selektion Dieses Verfahren entspricht in etwa der umgekehrten Vorwärts-Selektion. Ausgehend von dem vollen Modell $A \subseteq \mathscr{K}$ werden in jedem Iterationsschritt i alle noch im Modell A_i verbliebenen Regressoren einzeln ausgeschlossen und die dazu gehörenden Regressionsmodelle einzeln geschätzt. In jedem Iterationsschritt i wird nun jener Regressor aus dem Modell A_i ausgeschlossen, der die größte Reduktion des Modellwahlkriteriums liefert. Der Algorithmus bricht wieder ab, sobald der Ausschluss einer der a_i im Modell verbliebenen Regressoren, keine weitere Reduktion im Vergleich zum Iterationsschritt $(i-1)$ bringen würde. Das an diesem Algorithmus haftende Problem ist das Gleiche wie bei der Vorwärts-Selektion. Ein einmal aus dem Modell ausgeschlossener Regressor wird nicht mehr in dieses zurückfinden. Die Rückwärts-Selektion muss somit ebenso wie die Vorwärts-Selektion höchstens $k(k+1)/2$ Modelle schätzen.

[9] Dieses kann einerseits das Nullmodell sein, welches nur den Intercept beinhaltet, oder ein beliebig definiertes.

Schrittweise-Selektion Dieses Verfahren stellt eine Kombination der beiden oben genannten dar. Ähnlich der Vorwärts-Selektion startet der Algorithmus mit einem Startmodell A, dass entweder alle Regressoren oder nur einen Teil davon enthalten kann. In jedem Iterationsschritt i wird allerdings zunächst überprüft, ob ein im Startmodell A_i des Iterationsschrittes i enthaltener Regressor entfernt werden kann. Erst im zweiten Schritt wird dann versucht, das Modellgütemaß durch die Aufnahme einer der $(k - a_i)$ verbliebenen Regressoren weiter zu reduzieren. Einmal ausgeschlossene Regressoren haben somit die Möglichkeit wieder in das Modell zu gelangen. Der Algorithmus bricht ab, sobald der Ausschluss einer der a_i im Modell verbliebenen Regressoren, oder die Aufnahme einer der $(k - a_i)$ Regressoren, im Vergleich zum Iterationsschritt $(i - 1)$ keine weitere Reduktion bringen würde.
Unter den hier vorgestellten Variable-Screening Techniken liefert die Schrittweise-Selektion einen, im Sinne des gewählten Modellwahlkriteriums, besseren Regressorensatz als die anderen beiden.

Die eben beschriebenen Selektionsverfahren fanden bereits 1960 in leicht leicht abgewandelter Form, innerhalb der von *Efroymson* vorgeschlagenen Methodik Anwendung. In jedem Iterationsschritt i wird dort jener Regressor aufgenommen und/oder ausgeschlossen, der den höchsten oder niedrigsten auf Seite 26 definierten *F-to-enter* Wert aufweist. Das Verfahren bricht ab, sobald der F-to-enter Wert eines aufzunehmenden oder auszuschließenden Regessors ein vorgegebenes α-Fraktil über- oder unterschreitet, oder alle Regressoren bereits aufgenommen oder entfernt wurden. Üblicherweise wird für diese Schranke in der Rückwärts-Selektion ein α_{stay} Wert von 0.10 oder 0.05, und in der Vorwärts-Selektion ein α_{entry} Wert von 0.05 festgelegt (vgl. Xin und Xiao, 2009, S.171f). Das Hauptproblem dieses Auswahlverfahrens ist die wiederholte Anwendung der F-Teststatistik zum Test, ob ein Regressionskoeffizient von Null verschieden ist. Diese Teststatistik ist nämlich aufgrund der Wahl jenes Regessors, der den F-Wert maximiert *nicht* F-verteilt. Dies wäre sie nur dann, wenn in jedem Iterationsschritt ein beliebiger Regressor ausgewählt werden würde (vgl. Fahrmeir et al. (2009, S.164f) und Miller (2002, S.42f)).
Da die Rückwärts-Selektion im Rahmen der Efroymson Methodik allerdings höchsten k Modellschätzungen in k Iterationsschritten braucht, wird diese Vorgangsweise zufolge Xin und Xiao auch heute noch gerne angewandt. Selbst wenn in dem von diesem Verfahren ausgewählten Regressorensatz, mit sehr hoher Wahrscheinlichkeit nicht signifikante Regressoren enthalten sind und signifikante fehlen.

Ein Standardwerk zur Modellauswahl in der Regressionsanalyse scheint Miller (2002) zu sein, in dem, neben den hier genannten, auch noch zahlreiche andere Verfahren detailliert beschrieben sind.

Im Zuge dieser Arbeit wurden folgende in der MATLAB®-Umgebung zur Verfügung gestellte Funktionen verwendet:

`stepwisesearch` Diese Funktion entspricht der Efroymson Methodik im Rahmen der Schrittweisen-Selektion. Als Startmodell wurden sowohl das Nullmodell als auch das volle Modell gewählt. MATLAB® verwendet für α_{entry} und α_{stay} standardmäßig die 0.05 und 0.10 Fraktile. Eine besondere Behandlung hierarchischer Terme, wie Polynome, ist mit dieser Funktion allerdings nicht möglich. Es kann also vorkommen, dass zwar der quadratische Term in das Modell aufgenommen wird, der lineare Anteil hingegen verworfen wird.

In Listing 2.1 findet sich eine leicht modifzierte Beispielausgabe zur `stepwisesearch`-Funktion für die Modellschätzung aus (2.5). Obwohl es sich dabei um ein korrekt spezifiziertes Modell handelt, zeigt sich, dass der Selektionsalgorithmus lediglich den signifikanteren quadratischen Term des Regressors x auswählt.

Listing 2.1: Beispielausgabe der `stepwisesearch`-Funktion für das Teilmodell (2.5).

```
Start stepwise forward search:
Initial columns included: none
Step 1, added column x^2, p=5.42e-28 (AICc=22.2913)
Step 2, added column y, p=2.65e-09 (AICc=-101.902)
Final columns included: y x^2 (AICc=-137.701)
'Label' 'Coeff'     'Std.Err.'   'Status'   'P'
x       [-0.6055]   [ 0.4775]   'Out'      [   0.2112]
y       [ 0.8735]   [ 0.1193]   'In'       [2.6471e-09]
x^2     [ 3.9696]   [ 0.1156]   'In'       [6.0165e-35]

Model: y(x) = b(0) + b(1) y + b(2) x^2

Mean squared error = 0.0595
R^2 = 0.9627
Adjusted R^2 = 0.9611
```

`sequentialfs` Eigentlich für Kreuzvalidierungen[10] vorgesehen, ermöglicht diese Funktion auch eine Vorwärts- und Rückwärts-Selektion über die Definition eigener Kriterien. Darüber hinaus könnten hier Regressoren auch zwingend im Modell belassen oder ausgeschlossen werden. Hier wurden ebenso jeweils das Nullmodell und das volle Modell als Startmodell gewählt. Als Modellwahlkriterium wurde das AIC_c verwendet.

Da sich weder das Ergebnis noch die Ausgabe dieser Funktion von jener in Listing 2.1 wesentlich unterscheidet, wird an dieser Stelle auf eine Beispielausgabe verzichtet.

[10] ebenso ein Modellselektionsverfahren

2.6 Modelldiagnostik

Wie in 2.2 bereits festgehalten, wird das LM mit einer Reihe *regressionstheoretischer*, A3 bis A6, und *modellbezogener* Annahmen, A1 und A2, definiert. In 2.3 und 2.4 wurden für solch ein LM Methoden der Parameterschätzung und der Inferenzstatistik vorgestellt, deren Aussagekraft auf der, zumindest approximativen, Gültigkeit dieser Annahmen aufbaut. Ein wesentliches Ziel der Modelldiagnostik liegt in der Überprüfung der, einem erfolgreich geschätzten LM zugrundeliegenden Annahmen. Dazu werden vor allem bei einer großen Anzahl an Beobachtungen n, *visuelle Diagnosemethoden* empfohlen (vgl. u. a. Pruscha (2006, S.124) oder itl (2006, Kap.5.2.4)). Die Analyse des für eine Schätzung verwendeten Regressorensatzes, oder ganz allgemein der potentiellen Regressoren, gehört ebenso zum Gebiet der Modelldiagnostik. Hier versucht man statistisch auffallende Regressoren, sogenannte Ausreißer, und lineare Abhängigkeiten der Regressoren untereinander aufzudecken.

2.6.1 Residuenanalyse

In 2.3 auf Seiten 22–23 zur Parameterschätzung das erste Mal eingeführt, stellen Residuen eines der wesentlichsten Hilfsmittel dar, um die Annahmen A1 bis A6 des LM zu verifizieren. Wie die wahre Störgröße sind auch die Residuen im Mittel Null. Wie (2.21) aber zeigt, sind sie weder unkorreliert noch homoskedastisch und können somit nicht uneingeschränkt in der Modelldiagnostik verwendet werden. Während ihre Korrelation bei korrekt spezifiziertem Modell in der Regel vernachlässigbar ist (vgl. Fahrmeir et al., 2009, S.108), gilt dies nicht für ihre Heteroskedastizität. Da sie nämlich per Definition mit $\text{Var}(\hat{\boldsymbol{\epsilon}}) = (\boldsymbol{I} - \boldsymbol{H})\sigma^2$ heteroskedastisch sind, können sie nicht zur Überprüfung der Homoskedastizitätsannahme $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2$ verwendet werden. Dafür bedarf es einer geeigneten Transformation der Residuen.
Darüber hinaus sind die klassischen Residuen auch nur bedingt zur Überprüfung der Linearitätsannahme geeignet. Für aussagekräftige Informationen bedarf es auch hier einer Transformation.
Neben dem klassischen Residuum sind für die Modelldiagnostik folgende transformierte Varianten von Interesse

Standardisierte Residuen Der standardisierten Residuen bedarf es vor allem zur Überprüfung der Homoskedastizitätsannahme (A4). Man erhält sie durch Division der Wurzel ihrer geschätzten Varianzen gemäß (2.21) auf Seite 25

$$\hat{r}_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{2.34}$$

mit h_{ii} als Hebelwerte gemäß (2.15) auf Seite 22. Unter Voraussetzung der Gültigkeit von A4, müssten die standardisierten Residuen nun auch homoskedastisch sein.

Studentisierte Residuen Die studentisierten Residuen, *leave-one-out* Residuen oder auch PRESS genannt, zählen ebenso zu den standardisierten Residuen. Allerdings handelt es sich bei dieser Standardisierung um eine speziell gewichtete, weshalb sich die studentisierten Residuen neben der Überprüfung der Annahme A4 vor allem zur sogenannten Ausreißeranalyse[11] eignen. Zur Schätzung des i-ten studentisierten Residuums $\hat{r}_i^*$ wird jeweils auf die i-te Designstelle $\boldsymbol{x}_i$ bei den Paramterschätzungen verzichtet[12] und es gilt (Fahrmeir et al., 2009, S.110)

$$\hat{r}_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \tag{2.35}$$

mit $\hat{\sigma}_{(i)}$ als Schätzer jener Modellvarianz, die nicht auf der i-ten Beobachtung beruht. Um zur Bestimmung der studentisierten Residuen nun nicht wie im Zuge der Modellauswahl erneut $i = 1, \ldots, k$ Regressionsmodelle schätzen zu müssen, lassen sich folgende Zusammenhänge zeigen (Xin und Xiao, 2009, S.108)

$$\hat{\sigma}^2_{(i)} = \frac{1}{n-p-1}\left((n-p)\hat{\sigma}^2 - \frac{\hat{\epsilon}_i^2}{1-h_{ii}}\right) \tag{2.36}$$

für den leave-one-out Schätzer der Modellvarianz[13] und damit (Fahrmeir et al., 2009, S.110)

$$\hat{r}_i^* = \hat{r}_i\left(\frac{n-p-1}{n-p-\hat{r}_i^2}\right)$$

der einfach aus den standardisieren Residuen berechenbare Zusammenhang. Unter Voraussetzung der Gültigkeit von A4 müssten die studentisierten Residuen nun auch homoskedastisch sein.

partielle Residuen Die erstmals von Wayne und McCleary (1972) definierten Residuen, ermöglichen für einen spezifischen Regressor des multiplen LM, die Art seines Einflusses auf die abhängige Variable $\boldsymbol{y}$ zu beschreiben. Während die klassischen Residuen zwar einen nichtlinearen Zusammenhang aufdecken könnten, fehlt ihnen allerdings die Möglichkeit aufzuzeigen, ob es sich dabei um monotone oder nichtmonotone Nichtlinearität handelt. Um ein korrektes LM spezifizieren zu können, ist aber gerade diese Information

[11] Zur Erläuterung des Begriffs Ausreißers und zu dessen zahlreichen Analysemethoden siehe 2.6.2 auf Seite 48. Hier sollen die studentisierten Residuen lediglich im Zusammenhang der Annahmenanalyse Erwähnung finden.

[12] daher auch die übliche Bezeichnung eines leave-one-out Schätzers

[13] Im Falle vorhandener Ausreißer, wäre der REML σ^2 verzerrt

vonnöten.
Die partiellen Residuen eignen sich damit vor allem zur Überprüfung der Linearitätsannahme. Sie sind wie folgt definiert[14]

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{x_j}} = \hat{\boldsymbol{\epsilon}} + \hat{\beta}_j \boldsymbol{x_j} \tag{2.37}$$

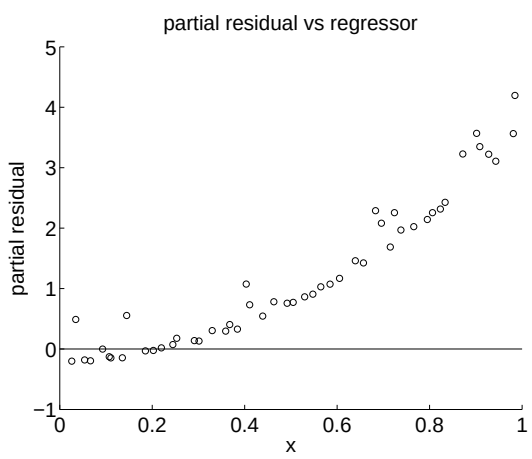
Damit wird der Einfluss aller Regressoren $\boldsymbol{x}$ bis auf den betrachteten $\boldsymbol{x}_j$ entfernt. Wie Abb. 2.1 auf der nächsten Seite für das korrekt spezifizierte Modell (2.5) deutlich zeigt, behalten die partiellen Residuen auch die Information ihres Regressionsparameters. Sie spiegeln somit auch ein wenig die Signifikanz des betrachteten Regressors wieder.

Mit Hilfe dieser Transformationen sind die Residuen nun geeignet, die regressionstheoretischen Annahmen A4 bis A6 und die beiden modellbezogenen Annahmen zu überprüfen. Eine Überprüfung der regressionstheoretischen Annahme A3, die für den Mittelwert der Residuen $\mathbb{E}(\boldsymbol{\epsilon}) = 0$ verlangt, ist nicht möglich. Dies deswegen, da die wahre stochastische Größe $\boldsymbol{\epsilon}$ unbekannt ist, und der Mittelwert der Residuen aufgrund der verwendeten Schätztechnik immer identisch Null ist. Würde diese Annahme verletzt werden, wäre aber lediglich die Schätzung des Intercept β_0 verzerrt (vgl. Urban und Mayerl, 2008, S.201).
Im Folgenden sollen nun die in dieser Arbeit verwendeten visuellen[15] und formalen Diagnostikmethoden vorgestellt werden. Die visuelle Diagnostik der empirischen Verteilung der Residuen stützt sich dabei hauptsächlich auf sogenannte *Streudiagramme*, oder auch *Scattergramme* genannt. In einem solchen vorwiegend 2-dimensional verwendeten Streudiagramm, werden Wertepaare entlang der Ordinate und der Abszisse als Punkt abgetragen.

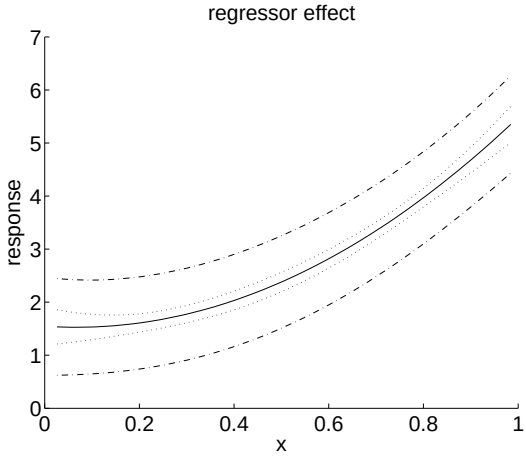
Residualplot Im klassischen Residualplot werden die standardisierten oder studentisierten Residuen $\hat{\epsilon}_i$ gegen die geschätzte abhängige Größe $\hat{y}_i$ abgetragen. Die Verwendung von $\hat{y}_i$ dient dabei als eine Art Notlösung, da in der multivariaten linearen Regression die Residuen eigentlich gegen jeden Regressor einzeln abgetragen werden sollten. In Modellen mit zahlreichen Regressoren ist dies verständlicherweise unpraktikabel. Da in dieser Arbeit allerdings höchstens 7 unabhängige Regressoren in einer Schätzung vorkommen konnten, soll im weiteren Verlauf mit dem klassischen Residualplot tatsächlich ein Plot der studentisierten Residuen $\hat{r}_i^*$ gegen einen Regressor x_j assoziiert werden.
Prinzipiell können die Residualplots dem Analysten dabei helfen

- Heteroskedastizität (A4) zu diagnostizieren,

[14] Sollte die Beziehung zwischen dem Regressor $\boldsymbol{x}_j$ und der unabhängigen Variable $\boldsymbol{y}$ polynomial modelliert worden sein (siehe 2.6.4 auf Seite 52), so muss (2.37) um die enstprechenden Terme ergänzt werden.
[15] Eine gute Zusammenstellung zahlreicher visueller Diagnostikmethoden findet sich z. B. in Fox (2008).



(a) Das partielle Residuum des Regressors $\boldsymbol{x}_1$



(b) Der Effekt des Regressors $\boldsymbol{x}_1$ auf die abhängige Größe $\boldsymbol{y}$. Die punktierte Linie entspricht dem 95%-Konfidenzintervall des prognostizierten Mittelwertes einer Beobachtung und die strichtlierte Linie dem 95%-Prognoseintervall des Beobachtungswertes selbst (vgl. Seite 29).

Abbildung 2.1: Der Effekt eines Regressors am Beispiel des Teilmodells (2.4)

- Ausreißer aufzudecken,
- und grundsätzlich auch um die Linearitätsannahme (A1) zu verifizieren.

Werden die Annahmen vom untersuchten LM eingehalten, müssten die studentisierten Residuen wie in Abb. 2.2(a) auf Seite 43 mit gleich bleibender Varianz (A4) regellos (A1) streuen. Folgen die studentisierten Residuen allerings in etwa einem gekrümmten Verlauf, wie in 2.2(b), lässt dies einen nichtlinearen Zusammenhang zwischen dem betrachteten Regressor und der abhängigen Variable y vermuten. Damit wäre A1 von diesem LM verletzt[16].

Weist der Residualplot hingegen eine z. B. trompetenförmige Verteilung der Punktewolke auf, würde dies auf eine Verletzung von A4 hindeuten. Zur Aufdeckung von Heteroskedastizität existieren zwar auch einige wenige formale Tests, wie der in Fahrmeir et al. (2009, S.131f) vorgestellte *Breusch-Pagan-Test*, deren Aussagekraft ist allerdings aufgrund der ihnen zugrundeliegenden Annahmen beschränkt. Wenn überhaupt sollten solche formalen Tests somit in Kombination mit visuellen Methoden verwendet werden.

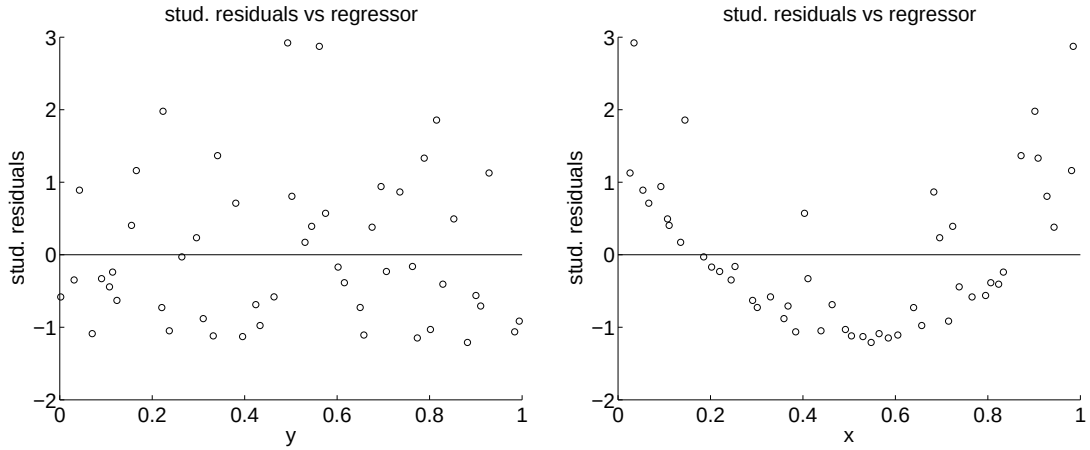
[16] Werden die studentisierten Residuen im Residualplot durch partielle Residuen ersetzt, wie dies im folgenden *Component-plus-residual plot* der Fall ist, könnten genauere Aussagen über die Art der Nichtlinearität gemacht werden.

Component-plus-residual plot Wie bei der Definition der partiellen Residuen auf Seite 40 bereits erwähnt wurde, eignen sich diese insbesondere zur Überprüfung der Linearitätsannahme (A1). Der Component-plus-residual plot wird somit genauso konstruiert wie der eben beschriebene Residualplot, mit dem wesentlichen Unterschied, dass nun die partiellen Residuen $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{x}_j}$ gegen den betrachteten Regressor $\boldsymbol{x}_j$ abgetragen werden. Zum Vergleich des Componente-plus-residual plot mit dem Residualplot werden in beiden Abbildungen, Abb. 2.2(b) für den Residualplot und Abb. 2.3 auf der nächsten Seite für den Component-plus-residual plot, der Regressor $\boldsymbol{x}_1$ des Teilmodells (2.4) dargestellt. Wie der Residualplot zeigt auch der Component-plus-residual plot für den betrachteten Regressor einen gekrümmten Verlauf des partiellen Residuums und damit eine Verletzung der Linearitätsannahme. Nun kann aus der Art dieses Verlaufes, aber auch auf die erforderliche Transformation des *Regressors* x_j geschlossen werden (vgl. 2.6.4 auf Seite 52).

run-sequence-order plot Ähnlich dem klassischen Residualplot unterstützt der run-sequence-order plot den Analysten

- Heteroskedastizität (A4) zu diagnostizieren,
- Ausreißer aufzudecken, und vor allem
- zur Diagnostik des einfachsten Falles *autokorrelierter Störungen*, der Autokorrelation erster Ordnung.

Beim run-sequence-order plot werden die Residuen $\hat{\epsilon}_i$ – oder eine ihrere standadisierten Transformationen – der Reihe nach gegen ihre Reihenzahl i abgetragen. Man spricht in diesem Zusammenhang auch vom „Zeitablauf“. Folgt einem positiven Residuum tendenziell wieder ein positives, könnte positive Autokorrelation erster Ordnung vorliegen. Alternierende Residuen, also Residuenverläufe mit häufig wechselndem Vorzeichen, deuten auf eine negative Autokorrelation erster Ordnung hin (vgl. Abb. 2.4(a) auf Seite 45). In beiden Fällen wäre somit die Annahme unkorrelierter Störgrößen (A5) verletzt.
Abbildung 2.4(b) zeigt den bei den Residualplots bereits erwähnten Fall von Heteroskedastizität. Entgegen der Annahme A4 streuen hier die studentisierten Residuen nicht mit gleichem Abstand um deren Mittelwert, sondern folgen von links nach rechts in etwa einem trompetenförmigen Verlauf. Anders als die anderen beiden Abbildungen in 2.4, wurde für diesen run-sequence-order plot das Modell (2.2) mit Hilfe des Teilmodells (2.4) geschätzt.



(a) Der Regressor $\boldsymbol{x}_2$ zeigt wie zu erwarten einen unauffälligen Residualplot

(b) Ganz anders sieht der Plot allerdings für den Regressor $\boldsymbol{x}_1$ aus. Der gekrümmte Verlauf lässt hier korrekter Weise einen nichtlinearen Zusammenhang zwischen Regressor und abhängiger Variable vermuten.

Abbildung 2.2: Residualplots am Beispiel des Teilmodells (2.4)

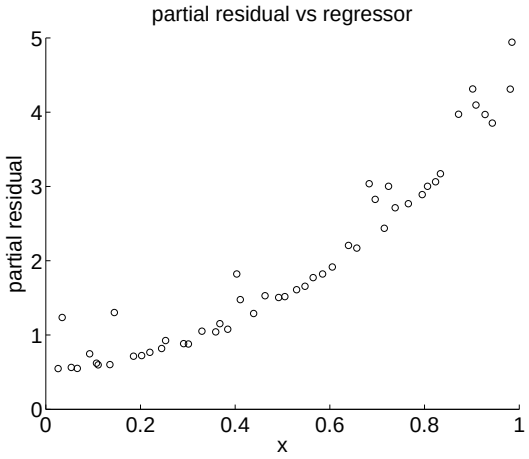


Abbildung 2.3: Im Vergleich zur Abb. 2.2(b) auf dieser Seite zeigt der Component-plus-residual für den Regressor $\boldsymbol{x}_1$ eindeutig einen *monotonen, nichtlinearen Zusammenhang* auf.

Ähnlich dem run-sequence-order plot dient der lag-1 plot der Diagnose autokorrelierter Störungen. Hier werden im Streudiagramm das i-te Residuum $\hat{\epsilon}_i$ gegenüber dem $(i-1)$-tem Residuum $\hat{\epsilon}_{i-1}$ abgetragen. Schmiegt sich die dabei ergebende Punktewolke tendenziell an eine steigende Gerade an, deutet dies auf postive Korrelation erster Ordnung hin; also $\rho > 0$. Schmiegt sie sich eher an eine fallende Gerade an, dürfte negative Autokorrelation vorliegen; also $\rho < 0$ (vgl. Abb. 2.4(c) auf der nächsten Seite)[17].
Weitere grafische Möglichkeiten stellen z. B. Visualisierungen der empirischen Autokorrelationsfunktionen dar (vgl. Fahrmeir et al., 2009, S.137ff). Neben diesen visuellen Methoden wird zur Feststellung serieller Korrelation auch sehr häufig der im Folgenden beschriebene *Durbin-Watson-Test* verwendet.
Autokorrelierte Störungen treten häufig dann auf, wenn das verwendete Regressionsmodell fehlspezifiziert ist. Dies kann z. B. der Fall sein, wenn der Einfluss eines Regressors nicht korrekt modelliert wurde, oder ein einflussreicher Regressor im Regressionsmodell unberücksichtigt blieb. Genauso wie die in 2.6.3 auf Seite 50 erläuterte Kollinearität der Regressoren, verschlechtern diese Störungen die Prognosegenauigkeit der abhängigen Größe y (vgl. Fahrmeir et al., 2009, S.68ff).

Durbin-Watson-Test Als formale Alternative zu den eben beschriebenen visuellen Methoden zum Testen der Korrelationsannahme A5, kann der Durbin-Watson Test[18] verwendet werden. Basierend auf der Teststatistik (Fahrmeir et al., 2009, S.141)

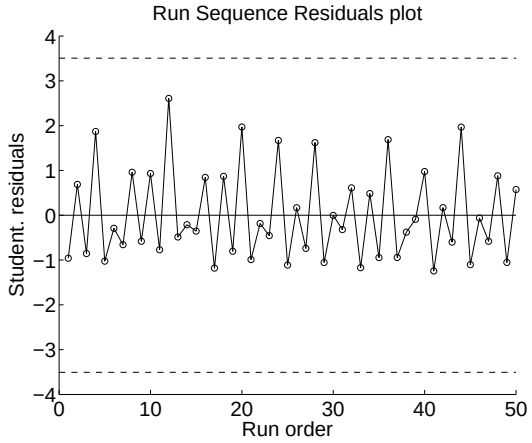
$$d = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\epsilon}_{i-1}^2} \approx 2(1-\hat{\rho}) \quad \text{mit} \quad -1 < \hat{\rho} < 1 \tag{2.38}$$

soll hierfür die folgende Nullhypothese verifiziert werden

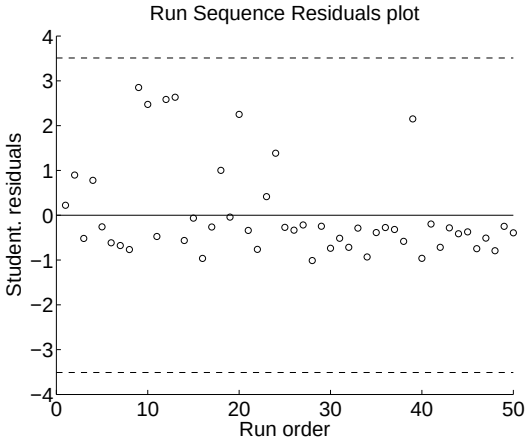
$$\mathscr{H}_0 : \rho = 0$$

mit ρ als Korrelationskoeffizient. Da die Bestimmung der Verteilung der Teststatistik unter $\mathscr{H}_0$ von der Designmatrix $\boldsymbol{X}$ abhängt und relativ schwierig ist, soll hier nicht näher darauf eingegangen werden. Grundsätzlich gilt, dass für unkorrelierte Residuen $\hat{\boldsymbol{\epsilon}}$ in etwa $\rho \approx 0$ gelten muss. Aus (2.38) ist somit ersichtlich, dass zufolge der Intervallgrenzen von $\rho \Rightarrow 0 < d < 4$ gelten muss. Die Nullhypothese wird also eher für $d \approx 2$ beibehalten.
In der MATLAB®-Umgebung wird der Durbin-Watson Test in der `dwtest`-Funktion zur Verfügung gestellt.

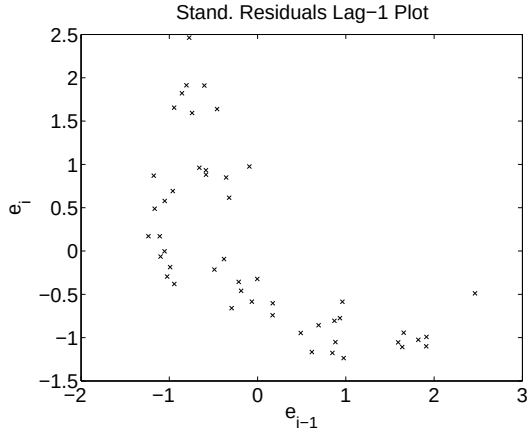
[17] Eine an die Punktewolke angepasste lineare Einfachregression $\boldsymbol{y} = a + b\boldsymbol{x}$ erleichtert eventuell die Beurteilung.
[18] eine kurze Einführung und einige Grundbegriffe zu statistischen Tests, findet sich in 2.4 auf Seite 24ff.



(a) Wie zu erwarten führt der fehlende Regressor $\boldsymbol{x}_1$ im betrachteten Teilmodell (2.3) zu einer deutlich negativen Autokorrelation



(b) Der von links nach rechts in etwa trompetenförmige Verlauf der studentisierten Residuen, könnte auf eine Verletzung der Homoskedastizitätsannahme hinweisen. Anders als für die anderen beiden Abbildungen (a) und (c), wurde für diesen run-sequence-order plot das Modell (2.2) mit Hilfe des Teilmodells (2.4) geschätzt.



(c) Genauso wie der run-sequence-order plot in Abb. (a), deutet auch der betrachtete lag-1-plot für das Teilmodell (2.3) auf eine negative Autokorrelation hin. Der monoton fallende nichtlineare Verlauf der Punktwolke lässt sogar eine Autokorrelation höherer Ordnung vermuten.

Abbildung 2.4: Autokorrelation und Heteroskedastizität

quantil-quantil plot Die Gültigkeit der Normalverteilungsannahme (A6) kann grafisch mit Hilfe eines quantil-quantil plots überprüft werden. Diese Visualiserungen eignen sich besonders für eine größere Anzahl an Beobachtungen n, da in solchen Fällen formale Tests ihre Nullhypothese tendenziell verwerfen.
quantil-quantil plots werden wie folgt konstruiert (siehe Pruscha, 2006, Kap.1)

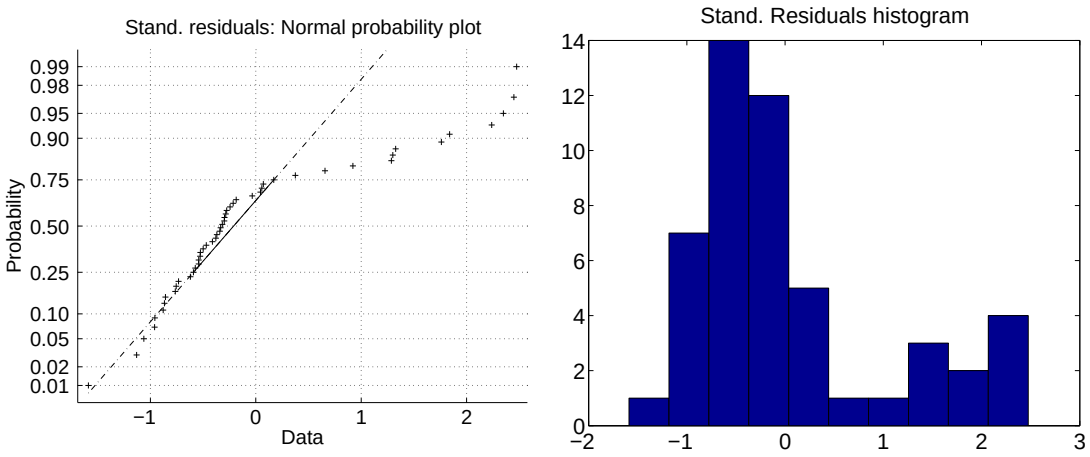
1. Zunächst werden die *standardisierten Residuen* $\hat{r}_i$ (2.34) auf Seite 38 mit $i = 1, \ldots, n$ aufsteigend geordnet und neu durchnummeriert, sodass die Ordnungszahl i dem Rang R_i des Residuums entspricht. Diese Transformation führt auf den Residuenvektor $\boldsymbol{e}$ mit der gewünschten Eigenschaft $e_{i-1} \leq e_i \; \forall \; i \in [1, n]$.

2. Im nächsten Schritt wertet man die inverse Funktion der Standardnormalverteilung Φ^{-1} an den Stellen
$$\Phi^{-1}\left(\frac{i - 1/3}{n + 1/3}\right) \quad i = 1, \ldots, n$$
aus und trägt diese Werte gegen die geordneten Residuen e_i im Streudiagramm ab.

3. Schließlich legt man eine Gerade durch die Punkte $\begin{bmatrix}\mathbb{E}(\hat{\boldsymbol{r}}) - \mathrm{Var}(\hat{\boldsymbol{r}}) & -1\end{bmatrix}$ und $\begin{bmatrix}\mathbb{E}(\hat{\boldsymbol{r}}) + \mathrm{Var}(\hat{\boldsymbol{r}}) & 1\end{bmatrix}$.

Wird die Normalverteilungsannahme (A6) vom untersuchten LM eingehalten, müssten die im Punkt 2 abgetragenen standardisierten Residuen $\hat{\boldsymbol{r}}$, zumindest näherungsweise an der Gerade liegen.
Die MATLAB®-Umgebung stellt zur Konstruktion solcher plots die `qqplot`-Funktion und speziell für die hier beschriebenen Quantile der Standardnormalverteilung die `normplot`-Funktion zur Verfügung. Abbildung 2.5 auf der nächsten Seite zeigt für das Teilmodell (2.5) sowohl den quantil-quantil plot also auch das dazugehörige Histogramm. Obwohl dieses LM eigentlich korrekt spezifiziert sein sollte, weisen beide plots eine deutliche Verletzung von A6 auf.

χ^2-Anpassungstest Bei Vorliegen einer ausreichenden Anzahl an Beobachtungen n kann die Normalverteilungsannahme der stochastischen Größe $\boldsymbol{\epsilon}$, neben der eben beschriebenen visuellen Methode, auch mit Hilfe des χ^2-Anpassungstest verifiziert werden (vgl. Viertl, 1990, S.128f). Mit Hilfe ihrer Teststatistik muss dazu die folgende Nullhypothese getestet werden
$$\mathscr{H}_0 : \hat{\boldsymbol{\epsilon}} \sim \Phi(0, \sigma^2)$$
Mit $\Phi(0, \sigma^2)$ als Standardnormalverteilung. Zur Konstruktion der χ^2-Teststatistik wird



(a) Residuen quantil-quantil plot

(b) Residuen Histogramm

Abbildung 2.5: Residuenverteilung am Beispiel des Teilmodells (2.5)

der Merkmalraum der Standardnormalverteilung in Ereignisse A_i mit $i = 1, \ldots, l$ zerlegt. Diesen Ereignissen ordnet man anschließend ihre entsprechenden Wahrscheinlichkeiten $p_i = \Phi(A_i) - \Phi(A_{i-1})$ zu. Die Anzahl m_i bezeichnet nun die Anzahl der Residuen, die einem solchen Ereignis entsprechen – deren Wert also im Ereignisfeld A_i liegt. Damit das Signifikanzniveau α hinreichend genau approximiert werden kann, wird empfohlen den Merkmalraum derart zu zerlegen, dass $m_i \; p_i \geq 5 \; \forall \; i = 1, \ldots, l$ gilt. Unter diesen Voraussetzung kann gezeigt werden, dass die Teststatistik

$$\sum_{i=1}^{l} \frac{(m_i - n \; p_i)^2}{n \; p_i} \sim \chi^2_{l-1} \quad \text{mit} \quad \sum_i m_i = n$$

χ^2_{l-1} verteilt ist mit $l - 1$ Freiheitsgraden. Die Nullhypothese $\mathscr{H}_0$ wird nun zu einem Signifikanzniveau α verworfen, falls die Teststatistik größer ist als das $(1 - \alpha)$-Quantil der χ^2_{l-1}-Verteilung

$$\chi^2_{l-1} > \chi^2_{l-1,1-\alpha}$$

Die Normalverteilungsannahme (A6) wäre somit verletzt.
Neben dem χ^2_{l-1}-Test kann auch der *Kolmogorov-Smirnov Test* verwendet werden (vgl. Pruscha, 2006, S.25f). Im Zuge dieser Arbeit wurden beide Tests ausgewertet. MATLAB® stellt hierfür die `chi2gof`-Funktion für den χ^2-Anpassungstest, und die `lillietest`-Funktion für den Kolmogorov-Smirnov Test zur Verfügung

2.6.2 Ausreißeranalyse

Betraf das vorherige Kapitel die Analyse der Residuen eines bereits geschätzten LM, betreffen dieses und das folgende Kapitel die Analyse eines *Regressorensatzes*.
Beinhaltet nämlich der zur Schätzung eines LM verwendete Regressorensatz in Frage kommende *Ausreißer* – auch *outlier* genannt – könnten die in 2.3 und 2.4 vorgestellten Schätz- und Inferenzstatistikmethoden in unerwünscht starkem Ausmaß von diesem beeinflusst worden sein. Dieser Einfluss ist umso stärker ausgeprägt, je mehr Regressorenwerte als Ausreißer in Frage kämen - man spricht von *multiplen Ausreißern* - und je eher diese am Rand des Wertebereichs liegen (vgl. Fahrmeir et al., 2009, S.173). Problematisch an einer Ausreißeranalyse ist allerdings, dass es keine mathematisch exakte Definition des Begriffs gibt, und folglich auch keine eindeutigen Analyseaussagen.
Grundsätzlich versteht man unter einem Ausreißer einen Regressor $\boldsymbol{x}_i$, der vom bedingten Erwartungswert der Modellprognose $\mathbb{E}(\hat{y}_i) = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ um einen wesentlichen Betrag Δ_i abweicht. Man spricht in diesem Zusammenhang auch von einem *mean shift* der stochastischen Größe $\mathbb{E}(\epsilon_i) = \Delta_i$ der die Mittelwertannahme (A3) der Definition des LM verletzt. In einem ersten Schritt könnte man also versuchen, einen Ausreißer i mit Hilfe der Größe seines Residuumbetrages $|\hat{\epsilon}_i|$ zu bestimmen. Hierzu gibt es zwei akzeptierte Möglichkeiten. So spricht man von einem m-fachen Ausreißer, falls der Betrag des standardisierten Residuum $\hat{r}_i$ eines Regressors i größer ist als die m-fache Modellstandardabweichung

$$\hat{r}_i > m\ \hat{\sigma}$$

mit $\hat{\sigma}$ aus 2.19 auf Seite 24 (vgl. etwa Pruscha (2006, S.126) und Urban und Mayerl (2008, S.185)). Das Problem an dieser Herangehensweise wurde bereits in der Einleitung des Kapitels angesprochen. Da ein eventuell am Rand des Regressorenwertebereichs liegender Ausreisser dazu führt, dass die Modellschätzung in Richtung dieses Ausreißers gezogen wird, könnte dessen Residuum zu klein werden um aufzufallen. Mit den in (2.35) auf Seite 39 transformierten *studentisierten Residuen* könnte dieses Problem vermieden werden. Dies deswegen, da die Schätzung der studentisierten Residuen auf allen, bis auf der i-ten Beobachtung beruht, welche ein potentieller Ausreißer sein könnte. Da sich zeigen lässt, dass diese Residuen bei korrekt spezifizierten LM t_{n-p-1}-verteilt sind, mit $n - p - 1$ Freiheitsgraden, lässt sich mit ihnen folgende Nullhypothese testen

$$\mathscr{H}_0 : \Delta_i = 0$$

mit Δ_i als mean-shift. Die Nullhypothese $\mathscr{H}_0$ wird nun zu einem Signifikanzniveau α verworfen, falls der Wert des studentisierten Residuums größer ist, als das $(1-\frac{\alpha}{2})$-Quantil der t-Verteilung. Damit wäre der Regressor i ein potentieller Ausreißerkandidat. Möchte man wieder simultan mehrere Regressoren auf mögliche Ausreißerkandidaten testen, bedarf es entweder der Verwendung des Schefféschen Quantils oder der konservativen Bonferroni-Korrektor; die auf Seite 29 bereits beschrieben wurden. Allgemein kann es allerdings relativ schwierig sein, multiple Ausreißer zu identifizieren (vgl. Fahrmeir et al., 2009, S.175).
In Abb. 2.4(a) auf Seite 45 wird für die studentisierten Residuen des Teilmodells (2.3) ein run-sequence-order plot dargestellt. Die beiden strichlierten Linien entsprechen dem 0.95-Quantil der $t_{50,\ 1-0.05/2}$-Studentenverteilung unter Anwendung der Bonferroni Korrektur. Da alle Beobachtungen innerhalb der beiden Intervallgrenzen zu finden sind, dürfte der zur Modellschätzung verwendete Regressorensatz keine Ausreirßer beinhalten – wie aufgrund der Konstruktion des Modells auch zu erwarten war.

Neben der – wiedereinmal – zentralen Residuenanalyse, können Ausreißerkandidaten auch mit Hilfe einiger Maßzahlen der *Einflussanalyse* aufgedeckt werden. Dabei geht man davon aus, dass jene Regressoren welche eine Schätzung besonders beeinflussen, potentielle Ausreißer sein könnten. Im Folgenden sollen einige übliche Maßzahlen vorgestellt werden.[19]

Hebelwerte Neben dem in der Residuenanalyse verwendeten mean-shift Δ_i zur Diagnose von Ausreißern, kann auch die Änderung der Varianz $\mathrm{Var}(\hat{\epsilon}_i) = \hat{\sigma}^2(1-h_{ii})$ als Kriterium herangezogen werden. Die als *Hebelwert* oder *leverage* bezeichneten Wert h_{ii} wurden bereits als Diagonalelemente der in (2.15) auf Seite 22 definierten Prädiktionsmatrix $\boldsymbol{H}$ eingeführt. Diese Werte sind durch $1/n < h_{ii} < 1$ beschränkt. Ein Hebelwert nahe Eins lässt die Varianz $\mathrm{Var}(\hat{\epsilon}_i)$ somit gegen Null gehen. Die Hyperebene verläuft somit praktisch sicher durch den Punkt $\boldsymbol{X}_i$, der damit als potentieller Ausreißerkandidat gilt (vgl. Fahrmeir et al., 2009, S.177f).

Im Idealfall sollten die Hebelwert h_{ii} möglichst gleich verteilt sein. Beobachtungen mit einem Hebelwert $h_{ii} > \frac{2p}{n}$, bei zahlreichen Stichproben, bzw. $h_{ii} > \frac{3p}{n}$ (vgl. Urban und Mayerl, 2008, S.188), bei kleineren Stichproben, gelten als allgemein anerkannte, ungefähre Schwellenwerte, die eine Überprüfung dieser Beobachtungen als mögliche Ausreißer nach sich ziehen sollte.

Analog zum zuvor beschriebenen statistischen Test der studentisierten Residuen, werden in Abb. 2.6(a) auf Seite 51 die Hebelwerte des Teilmodells (2.3) dargestellt. Die Abbildung zeigt eine unauffällige und gleichmäßige Verteilung der Hebelwerte. Da der

[19] Neben vielen anderen können auch diese Maßzahlen von der MATLAB®-Funktion `regstats` im Zuge einer Regressionsschätzung ausgewertet werden

Schwellenwert $h_{ii} > \frac{2p}{n} = 0.08$ von keinem Hebelwert überschritten wird, schließt auch dieses Kriterium die Existenz von Ausreißern im betrachteten Regressorensatz aus.

Cook-Distanz Die *Cook-Distanz* entspricht genauso wie die folgende DFFITS Maßzahl einem globalen Einflusskriterium, welches die Veränderung der Schätzer bei leave-one-out Schätzungen misst. Bei der Cook-Distanz wird der euklidische Abstand zwischen der Prognose $\hat{\boldsymbol{y}}$ und der leave-one-out Prognose $\hat{\boldsymbol{y}}_{(\boldsymbol{i})}$ gemessen und gewichtet (Fahrmeir et al., 2009, S.178)

$$D_i = \frac{\|\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(\boldsymbol{i})}\|^2}{p\ \hat{\sigma}^2}$$

Wieder nur als Faustregel gelten Beobachtungen mit $D_i > 0.5$ als auffällig und sollten spätestens ab $D_i > 1$ näher betrachtet werden.
Abbildung 2.6(b) auf der nächsten Seite bildet die Cook-Distanzen des Teilmodells (2.3) ab. Genauso wie der statistische Test der studentisierten Residuen und die Analyse der Hebelwerte schließt auch eine Überprüfung der Cook-Distanzen die Existenz von Ausreißern aus.

DFFITS Der standardisierte DFFITS-Wert ist eine kombinierte Funktion aus den zuvor vorgestellten Hebelwerten h_{ii} und den studentisierten Residuen $\hat{r}_i^*$ (Urban und Mayerl, 2008, S.189)

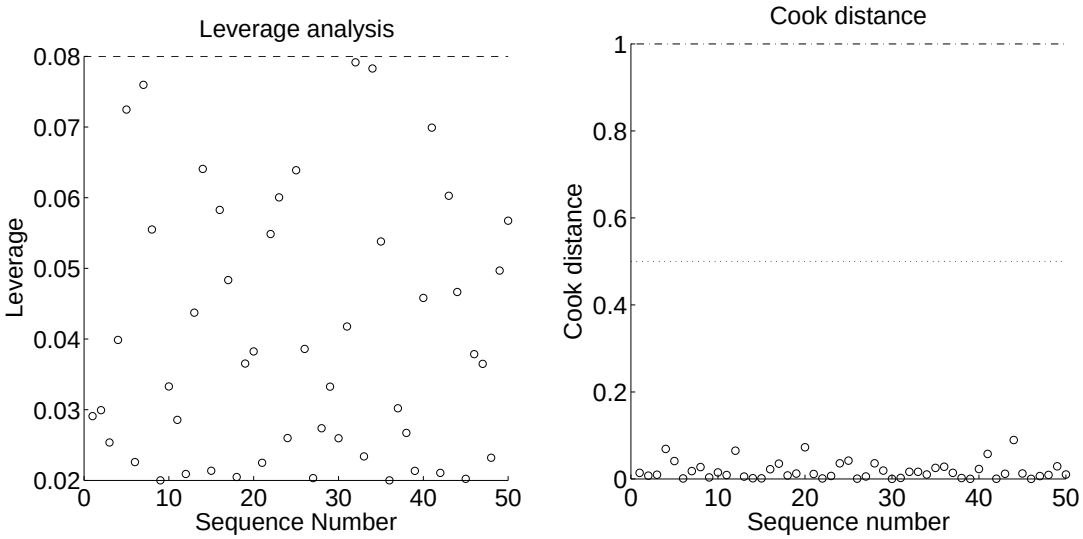
$$\text{DFFITS}_i = \hat{r}_i^* \sqrt{\frac{h_{ii} + 1/n}{1 - h_{ii} - 1/n}}$$

Als grobe Faustregel für auffällige Werte werden hier die Schranken $|2\sqrt{\frac{p}{n}}|$ bei zahlreichen Beobachtungen und $|1|$ bei wenigen Beobachtungen definiert.

2.6.3 Kollinearitätsanalyse

Unter *Kollinearität* versteht man die lineare Abhängigkeit von Regressoren. Für die stochastische Größe $\boldsymbol{y}$ erhofft, können hingegen kollineare Regressoren $\boldsymbol{x}$, genauso wie korrelierte Residuen, die Schätzung der Regressionsparamter stark verzerren. Im Extremfall perfekter linearer Abhängigkeit eines Regressors $\boldsymbol{x}_j$ von anderen Regressoren, explodiert die Varianz des jeweilgen Regressionsparamter β_j (vgl. Xin und Xiao, 2009, S.84f).
Der einfachsten Fall, nämlich das Vorliegen linearer Abhängigkeit zwischen zwei Regressoren, kann mit Hilfe der Korrelationsmatrix entdeckt werden (Bucher, 2009)

$$\rho_{ij} = \frac{\text{Cov}(x_i, x_j)}{\sigma_i\ \sigma_j} \quad \text{und} \quad -1 < \rho < 1 \tag{2.39}$$



(a) Die unauffällige und gleichmäßige Verteilung der Hebelwerte unterhalb des Schwellenwertes 0.08 schließt die Existenz von Ausreißern aus.

(b) Auch der plot der Cook-Distanzen signalisiert keine Auffälligkeiten

Abbildung 2.6: Ausreißeranalyse mit Hilfe von Einlusszahlen am Beispiel des Teilmodells (2.3)

mit ρ_{ij} als sogenanntem *Korrelationskoeffizienten*.
Für den Fall *multipler Korrelation*, also der linearen Abhängigkeit eines Regressors x_j von mehreren anderen, schätzt man hingegen für jeden Regressor x_j mit $j = 1, \ldots, p$ das folgende LM

$$x_j = \beta_0 + x_1 \cdot \beta_1 + \cdots + x_{j-1} \cdot \beta_{j-1} + x_{j+1} \cdot a_j + \cdots + x_p \cdot \beta_{p-1}$$

und berechnet dessen zugehöriges Bestimmtheitsmaß R_j^2 gemäß (2.30) auf Seite 32. Der darauf aufbauende *Varianzinflationsfaktor* (Fahrmeir et al., 2009, S.171)

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad \text{mit} \quad 0 \leq R_j^2 \leq 1 \tag{2.40}$$

beschreibt dann jenen Faktor, um den die Varianz von $\hat{\beta}_j$ zufolge des Kollinearitätseffektes vergrößert wird. Je größer die multiple Korrelation zwischen $\boldsymbol{x}_j$ und den anderen Regressoren ist, desto größer ist sein Bestimmtheitsmaß R_j und desto größer fällt der Varianzinflationsfaktor VIF_j aus. Als Faustregel zum Erkennen eines eventuellen Kollinearitätsproblems gilt ein Variationsinflationsfaktor $\text{VIF}_j > 10$ (vgl. Fahrmeir et al. (2009, S.171) und Xin und Xiao (2009, S.86)). Der Kehrwert von (2.40) wird als *Toleranz* bezeichnet. Demgemäß signalisieren hier Toleranzwerte kleiner 0.1 ein mögliches Kollinearitätsproblem.

Ein weiteres Anzeichen für unerwünschte Kollinearitäten stellt eine sehr große Konditionszahl des in (2.14) auf Seite 22 verwendeten Matrixprodukts $\boldsymbol{X}'\boldsymbol{X}$ der Designmatrix $\boldsymbol{X}$ dar. Im Extremfall würde das Matrizenprodukt singulär werden - also nicht mehr invertiert werden können (vgl. Xin und Xiao, 2009, S.84f).

2.6.4 Maßnahmen

Wurde von den in 2.6.1 bis 2.6.3 auf Seiten 38–50 vorgestellten Diagnostikmethoden im zu untersuchenden LM ein Modellverstoß entdeckt, können weder die Schätzwerte dieses LM, noch die darauf aufbauende Inferenzstatistik als verlässlich angesehen werden. In diesem Teilkapitel sollen nun einfache Transformationstechniken der Regressoren oder des gesamten LM vorgestellt werden, mit deren Hilfe versucht werden kann, diese Modellverstöße zu beseitigen.

Nichtlinearität Signalisiert der Component-residual-plot durch einen gekrümmten Verlauf der partiellen Residuen eine nichtlineare Beziehung zwischen dem Regressor $\boldsymbol{x}_j$ und der abhängigen Variable $\boldsymbol{y}$ (vgl. Abb. 2.3 auf Seite 43), kann in Abhängigkeit dieses Krümmungsverlaufes versucht werden, den Regressor geeignet zu transformieren und den Modellverstoß somit zu beseitigen. Wie schon des öfteren angemerkt, findet sich eine Übersicht linearisierbarer nichtlinearer Beziehungen z. B. in Urban und Mayerl (2008, S.211).

Eine der einfachsten und im Zuge dieser Arbeit auch angewandten Transformationstechnik zur Berücksichtigung monotoner oder parabelförmiger Nichtlinearität metrischer Regressoren, ist die Verwendung einer *polynomischen Regressionsfunktion*. In der Regel werden hierfür höchstens Polynome dritten Grades angesetzt, da die resultierenden Schätzungen sonst zu instabil werden. Den einfachen aber notwendigen Transformationsschritt zur Berücksichtigung des nichtlinearen Einflusses eines Regressors x_1 auf die abhängige Variable, soll anhand eines Polynoms 2. Grades gezeigt werden

$$\begin{aligned} &\mathrm{f}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2, \quad \text{mit} \\ &\quad z_1 = x_1, \quad z_2 = x_1^2, \quad z_3 = x_2, \quad \text{folgert schließlich} \\ &\mathrm{f}(x) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 = \boldsymbol{\beta}'\boldsymbol{z} \end{aligned} \tag{2.41}$$

Mit Hilfe dieser Transformationstechnik lassen sich auch sogenannte *Interaktionsterme* in die Regression miteinbeziehen. Diese Terme sind immer dann notwendig, wenn der Effekt einer oder mehrerer Regressoren vom Wert von mindestens einem anderen Regressor abhängt. Berücksichtigt man also in (2.41) zusätzlich zu den sogenannten

Haupteffekten $\beta_1 x_1$ und $\beta_2 x_2$ den Interaktionsterm $x_1 x_2$, erhält man schließlich folgende Regressionsfunktion

$$\begin{aligned} \mathrm{f}(x) &= \cdots + \beta_4 x_1 x_2, \quad \text{mit} \\ z_4 &= x_1 x_2 \quad \text{folgert wieder} \\ \mathrm{f}(x) &= \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 = \boldsymbol{\beta}' \boldsymbol{z} \end{aligned}$$

Bei der Berücksichtigung von Interaktionen zwischen zwei metrischen Regressoren die polynomial in die Regressionsfunktion eingehen, stößt man mit den einfachen linearen Regressionsfunktionen allerdings an die Grenze. Hierfür müssten dann 2-dimensionale Funktionen z. B. durch mehrdimensionale Polynome modelliert werden. In diesem Fall sind die automatisierten Schätztechniken der nicht- und semiparametrischen Regressionsmodelle, den hier gezeigten Möglichkeiten deutlich überlegen.
Mit der hier vorgestellten Maßnahme ist allerdings auch ein erwähnenswerter Nachteil verbunden. Da nämlich der neu definierte Regressorensatz aus einer Transformation der in der ursprünglichen Regressionsfunktion bereits enthaltenen Regressoren hervorgeht, weisen dessen Regressoren auch eine höhere *Multikollinearität* auf (vgl. Urban und Mayerl (2008, S.210) und Listing 3.2 auf Seite 65).

Multiplikative Störgröße und Heteroskedastizität Signalisiert einer der entsprechenden Streudiagramme in 2.6.1 heteroskedastische Varianzen, verliert das Gauß-Markov-Theorem seine Gültigkeit, und die KQ-Schätzer stellen in der speziellen Klasse der linearen und erwartungstreuen Schätzer nicht mehr die minimalvarianten Schätzer dar (vgl. Urban und Mayerl, 2008, S.242).
Im einfachsten Fall stellt Heteroskedastizität ein Anzeichen für ein fehlspezifiziertes Modell dar. Das betrachtete Teilmodell vermag in diesem Fall nicht, einen bedeutenden systematischen Variationsanteil der abhängigen Variable zu beschreiben. Schafft eine Neuspezifikation des Modells dennoch keine Verbesserung, kann wie im Falle von (2.2) versucht werden, das Regressionsmodell geeignet zu transformieren. Auslöser für die heteroskedastischen Varianzen in (2.2) ist nämlich dessen multiplikative Störgröße $\boldsymbol{\epsilon}$ (vgl. Fahrmeir et al., 2009, S.70)

$$\begin{aligned} \boldsymbol{y} &= \exp(\beta_0 + \boldsymbol{x}_1 \beta_1 + \boldsymbol{x}_2 \beta_2 + \boldsymbol{\epsilon}) \\ &= \exp(\beta_0) \exp(\boldsymbol{x}_1 \beta_1) \exp(\boldsymbol{x}_2 \beta_2) \exp(\boldsymbol{\epsilon}) \end{aligned}$$

also eine Verletzung der Additivitätsannahme A2. In diesem Fall hilft ein einfaches

Logarithmieren der abhängigen Variable $\boldsymbol{y}$. Anstelle des Teilmodells (2.4) würde also das folgende Teilmodell zur Schätzung verwendet werden

$$\log(\boldsymbol{y}) = \beta_0 + \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\epsilon}$$

Liegt die Ursache der Heteroskedastizität in einem anderen Grund, oder gibt es eventuell sogar Argumente, welche die Notwendigkeit der Heteroskedastizität begründen (vgl. Urban und Mayerl, 2008, S.243ff), kann z. B. mit Hilfe der *gewichteten Methode der kleinsten Fehlerquadrate* versucht werden, die Residuenvarianz konstant zu halten (vgl. Fahrmeir et al., 2009, S.133f).

Korrelation Signalisieren der run-sequence-order plot oder der lag-1-plot für die *Störgröße* $\boldsymbol{\epsilon}$ eine Autokorrelation erster Ordnung, so liegt eine fehlspezifizierte Modellschätzung vor. Schafft auch hier eine Neuspezifikation des Modells keine Verbesserung, kann wie im Falle von Heteroskedastizität mit Hilfe der gewichteten Methode der kleinsten Fehlerquadrate versucht werden, einen verbesserten Parametervektor $\boldsymbol{\beta}$ zu schätzen (vgl. auch hier Fahrmeir et al., 2009, S.142f).
Wird im Zuge der Korrelationsanalyse für einen *Regressor* eine hohe multiple Korrelation mit anderen Regressoren des verwendeten Satzes festgestellt, so sind mehrere Maßnahmen denkbar. Die einfachste und gleichzeitig fragwürdigste Variante ist eine erneute Modellschätzung unter Vernachlässigung des betreffenden Regressors. Die Konsequenz dieser Vorgehensweise wurde bereits in der Einleitung zur Modellselektion 2.5.2 auf Seite 33 erläutert. Eine weitere einfache und häufig angewandte Maßnahme bei diagnostizierte Kollinearität, ist die Zentrierung der Designmatrix. Mit Ausnahme des Intercept wird der Schätzwert des Parametervektors durch diese numerisch stabilisierende Maßnahme nicht verändert (vgl. Xin und Xiao, 2009, S.87ff). Für ein beliebiges lineares Regressionsmodell ergibt sich seine zentrierte Variante zu

$$\begin{aligned} \boldsymbol{y} &= \beta_0 + \boldsymbol{x}_1 \cdot \beta_1 + \boldsymbol{x}_2 \cdot \beta_2 + \cdots + \boldsymbol{x}_p \cdot \beta_p + \boldsymbol{\epsilon} \\ &= \alpha + \beta_1(\boldsymbol{x}_1 - \bar{x}_1) + \cdots + \beta_p(\boldsymbol{x}_p - \bar{x}_p) + \boldsymbol{\epsilon} \\ \text{also} \quad &= \alpha - \bar{\boldsymbol{x}} \cdot \boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned}$$

Damit ergibt sich der modifizierte Intercept zu $\beta_0 = \alpha - \bar{\boldsymbol{x}} \cdot \boldsymbol{\beta}$.
Weitere Möglichkeiten um Kollinearitäten zwischen Regressoren zu berücksichtigen oder gänzlich zu vermeiden, stellen die *Ridge-Regression* (vgl. Fahrmeir et al., 2009, S.171f) und die *Hauptkomponentenregression* (vgl. ebenda und Xin und Xiao (2009, S.238ff)) dar.

Ausreißer Wurden im Zuge der Ausreißeranalyse potentielle Ausreißerkandidaten gefunden, empfiehlt es sich zunächst die Ursache des Ausreißers zu klären. So könnte z. B. gerade dieser auffallende Wert für die Regressionsanalyse von großer Bedeutung sein. Ist dies nicht der Fall, oder kann die Ursache nicht festgestellt werden, gibt es wieder mehrere mögliche Vorgehensweisen. Neben der bereits öfters erwähnten und in der Literatur gleichsam kritisierten Variante des kommentarlosen Ausschlusses, kann auch hier auf spezielle Schätzverfahren, wie die *robuste Regression*, zurückgegriffen werden. Diese Schätzverfahren werden durch einzelne Ausreißer nicht oder wesentlich geringer als die KQ-Schätzer beeinflusst (vgl. z. B. Fahrmeir et al., 2009, S.90ff und S.175f).

3 Anwendung und Resümee

Wie im Vorwort auf Seite VII bereits beschrieben, wurden die in den Kapiteln 1 bis 2 erarbeiteten theoretischen Grundlagen, innerhalb der Software MATLAB®, soweit nicht bereits vorhanden, implementiert. In Kombination mit der Skriptsprache bash und dem Finite-Elemente Hybridsolver speedyne wurde mit der derart gestalteten Softwareumgebung anschließend ein Testbeispiel ausgewertet. Die dabei gewonnenen Ergebnisse sollen nun in diesem letzten Kapitel gezeigt und bewertet werden.

Bei dem dieser Arbeit zugrunde liegenden Testbeispiel handelt es sich um einen Falltest, bei dem ein C-Profil aus Stahl aus genau definierter Anfangslage auf eine ebene Unterlage fällt. Ziel dieses Testbeispiels ist es, die Vergleichsspannung in einem einzelnen, genau definierten Knotenpunkt des C-Profils, durch Variation der Startposition des desselbigen zu maximieren. Da speedyne die Rotation eines Körpers mit Hilfe der *Rodrigues-Formel* beschreibt, konnte zur Optimierung der folgende Parametervektor variiert werden

$$\boldsymbol{x} = \begin{pmatrix} x & y & z & k_x & k_y & k_z & \alpha \end{pmatrix} \tag{3.1}$$

mit $x,\ y,\ z$ als *Translationsvektor* des Massezentrums des Körpers, $k_x,\ k_y,\ k_z$ zur Definition einer *Rotationsachse*, und α als jener *Drehwinkel* mit dem der Körper um die Rotationsachse gedreht wird.
Für dieses Testbeispiel wurde im Sinne der Arbeit zunächst eine lineare Regressionsfläche konstruiert, mit dem Parametervektor $\boldsymbol{x}_i$ aus (3.1) als Regressoren und dem Wert der Vergleichsspannung als abhängige Größe. Zu diesem Zweck wurden die Vergleichsspannungen für drei Designs gemäß (3.1) mit $n = 20,\ 50$ und 100 Parametervektoren $\boldsymbol{x}_i$ durch speedyne mit einer Simulationsdauer von $t = 0.04s$ ausgewertet. Die Designstellen wurden dazu mit Hilfe der MATLAB® Funktion `haltonset` einer Halton-Folge entnommen[1] und auf den gewünschten, schrankenrestringierten Bereich abgebildet. Dieser Definitionsbereich wurde durch Schranken für die Elemente des Translationsvektors mit $0 \leq \boldsymbol{x}_m \leq 50\text{cm}$, jene zur Definition

[1] Dabei handelt es sich um eine besonders gute, „zufällig“ verteilte Punktmenge auf dem sogenannten Einheitswürfel. Für näheres vgl. z. B. Niederreiter (1992).

der Rotationsachse mit $0 \leq \boldsymbol{x}_r \leq 1$ und den Schranken für den Winkel α mit $0 \leq \alpha \leq 360°$ begrenzt. Die über die Simulationsdauer oszillierenden Vergleichsspannungen jeder einzelnen Designstelle wurden in MATLAB® importiert und vor einer weiteren Verarbeitung mit einem sogenannten *moving-average Filter* geglättet. Im Zuge dieser Filterung wird der Wert jeder Vergleichsspannung σ zum Zeitpunkt t, durch einen wie folgt definierten Mittelwert $\bar{\sigma}_t$ ersetzt

$$\bar{\sigma}_t = \frac{1}{2N+1}\left(\sigma_{(t+N)} + \sigma_{(t+N-1)} + \cdots + \sigma_{(t-N)}\right)$$

Die Variable N bezeichnet dabei die Anzahl der für die Mittelwertbildung zu berücksichtigenden Nachbarpunkte des betrachteten Zeitpunktes t. Im Falle von Randwerten, bei denen die gewünschte Anzahl an Nachbarpunkten nicht zur Verfügung steht, muss N entsprechend reduziert werden. Die Filterung wurde mit Hilfe der `smooth`-Funktion der MATLAB®-Umgebung mit einem Wert $N = 10$ vorgenommen.
Für die gefilterten Eingangsdaten wurde schließlich für jede der i Designstellen das Spannungsmaximum über t bestimmt und als abhängige Größe y_i in das Regressionsmodell aufgenommen (vgl. Abb. 3.1 auf der nächsten Seite). Mit den derart gewonnenen Werten wurde anschließend für jedes Design, sofern möglich[2], alle fünf in Tab. 3.1 auf der nächsten Seite angeführten linearen Regressionsmodelle geschätzt. Die Tabelle stellt für jeden der in die LM 1 bis 5 eingegangenen Regressoren x_i dar, ob und in welcher Art er zuvor transformiert wurde. Die *ID*-Spalte soll dabei helfen, die in dieser Arbeit geschätzten LM kompakt zu beschreiben. Mit Hilfe dieser ID's lässt sich das mit dem Parametervektor $\boldsymbol{x}$ in (3.1) konstruierte LM

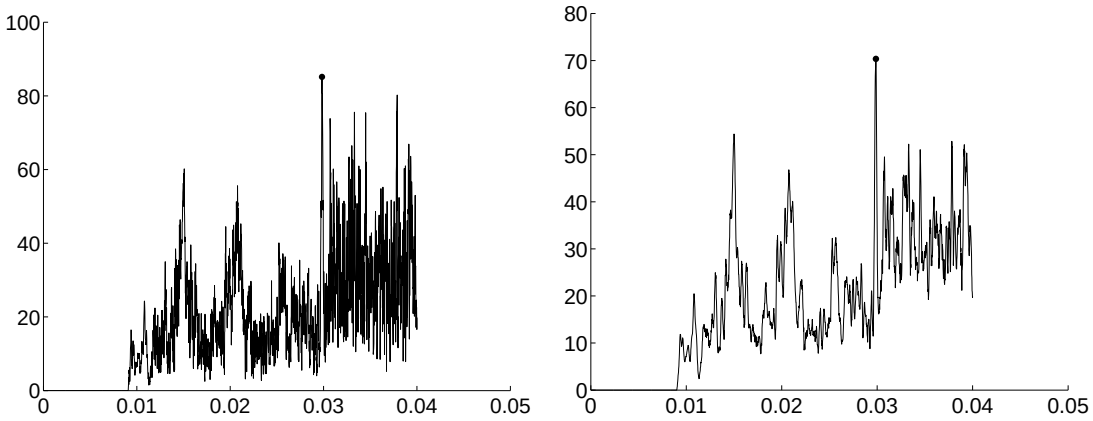
$$\begin{aligned} y &= \boldsymbol{x}\boldsymbol{\beta} \quad \text{mit} \\ \boldsymbol{x} &= \begin{pmatrix} x & y + yx + yz + yk_x + yk_y + yk_z + y\alpha + y^2 & 0 & k_x & 0 & k_y^2 & k_z + k_z^2 \end{pmatrix} \end{aligned}$$

beispielsweise dergestalt beschreiben

$$\boldsymbol{x} = \begin{pmatrix} 1 & 7 & 0 & 1 & 0 & 4 & 5 \end{pmatrix}$$

Die Spalte Selektion zeigt an, ob mit Hilfe der in 2.5.2 auf Seite 33 angeführten Algorithmen vor der Modellschätzung eine Variablenselektion vorgenommen wurde. Wie bereits erwähnt, wurde diese mit den MATLAB® Funktionen `stepwisesearch` und `sequentialfs` sowohl für das Nullmodell als auch für das volle Modell durchgeführt. Aus der weiteren Modellschätzung wurden dann jene Regressoren ausgeschlossen, die von allen vier Modellselektionsverfahren

[2] Für das erste Design mit $n = 20$ Parametervektoren sind zur Schätzung des Parametervektors gemäß (2.17) auf Seite 23 der LM 4 und 5 der Tab. 3.1 zu wenige Designstellen verfügbar. Für dieses Design musste daher auf die Schätzung dieser beiden Modelle verzichtet werden.



(a) Die Abbildung zeigt den ungefilterten Vergleichsspannungsverlauf einer Designstelle über die Zeit t

(b) Eine moving-average Filterung mit $N = 10$ führt im Vergleich zur Abb. (a) zu einer deutlichen Glättung. Die Punkte markieren die Maxima der beiden Verläufe.

Abbildung 3.1: Beispielausgabe für eine moving-average Filterung

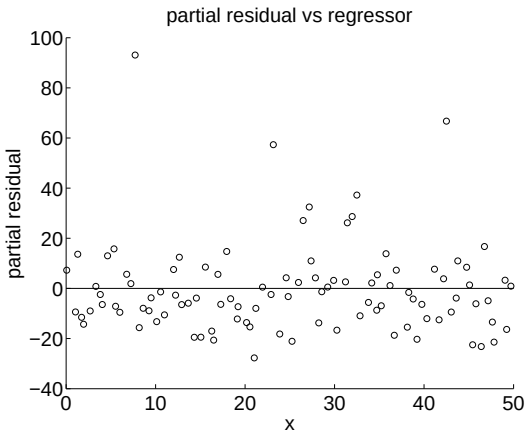
gleichsam ausgeschlossen worden sind.

Ergebnisse Im Folgenden sollen nun zwei der im Zuge dieser Arbeit ausgewerteten LM im Detail vorgestellt werden. Zur Demonstration des kausalen Zusammenhangs zwischen den Regressoren und der abhängigen Variable, sollen zunächst die Ergebnisse des LM 1 erörtert werden. Daran anschließend werden die Ergebnisse des, im Sinne des Modellwahlkriteriums AIC_c, besten gefundenen Regressionsmodells vorgestellt. Ein Ausblick über mögliche die Modelle verbessernde Maßnahmen schließt diese Arbeit dann ab.

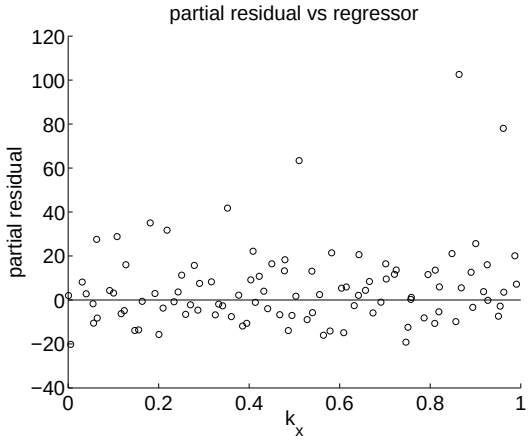
LM 1 Abbildung 3.2 auf der nächsten Seite zeigt den component-plus-Residualplot beispielhaft für drei Regressoren des LM 1 mit $n = 100$ Designstellen. Wie zu erkennen ist,

LM Nr.	ID	Grad	Polynom	Selektion
1	1	linear	x_i	nein
2	1	linear	x_i	ja
3	4	quadratisch	x_i^2	ja
4	7	quadratisch gemischt	$x_i + x_i x_j + x_i^2$	ja
5	15	kubisch gemischt	$x_i + x_i x_j + x_i^2 + x_i^3$	ja

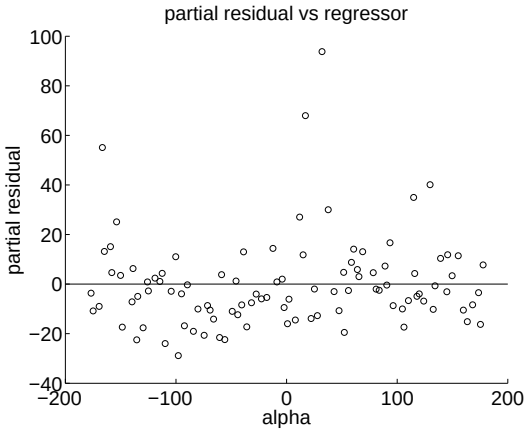
Tabelle 3.1: Regressionsmodelle im Testbeispiel



(a) Die partiellen Residuen der x-Komponente des Translationsvektors streuen unauffällig



(b) Die partiellen Residuen der x-Komponente des Rotationsvektors zeigen leicht heteroskedastische Eigenschaften



(c) Der Drehwinkel α

Abbildung 3.2: Component-plus-residual plots am Beispiel des LM 1 mit $n = 100$ Designstellen

streuen die partiellen Residuen der x-Komponente des Translations- und Rotationsvektor in 3.2(a) und (b) regellos um ihren Mittelwert und unterstützen damit die Annahme eines linearen Zusammenhangs. Einzig die partiellen Residuen in Abbildung 3.2(c) folgen einem leicht parabelförmigen Verlauf und weisen somit auf eine Beziehung höherer Ordnung hin. Der marginal trompetenförmige Verlauf in 3.2(b) könnte eventuell auf eine Verletzung der Homoskedastizitätsbedingung hindeuten. Weiters finden sich in Listing 3.1 einige der zu diesem Modell gehörenden Modellselektionskriterien und Auswertungen statistischer Tests. Wie sich zeigt, weist mit Ausnahme des Intercept nur die z-Komponente des Rotationsvektors einen signifikanten Werte auf. Diese Aussage wird im Prinzip auch vom globalen F-Test bestätigt, weshalb sich jede weitere Ausführung zu diesem Modell erübrigt.

Listing 3.1: Einige wesentliche Kenngrößen des LM1 für $n = 100$ Designs

```
Akaike Information Criteria (AIC) = 595.1935
AIC bias-corrected = 596.7759

Mean squared error = 356.1375
R^2 = 0.0933
Adjusted R^2 = 0.0243

Coefficients in the Equation
                                                Conf. inter.
Coeff        B      SE B        T     Sig T    lower    upper R(j)^2    VIF
const   43.9822    8.3513   5.2665   < 0.001  27.3957  60.5686
x       -0.0221    0.1303  -0.1693    0.8660  -0.2809   0.2368   0.01    1.01
y       -0.1347    0.1316  -1.0235    0.3087  -0.3961   0.1267   0.01    1.01
z        0.0458    0.1300   0.3520    0.7256  -0.2124   0.3039   0.00    1.00
k_x     10.8166    6.7125   1.6114    0.1105  -2.5151  24.1483   0.03    1.04
k_y      3.9283    6.6763   0.5884    0.5577  -9.3315  17.1881   0.04    1.04
k_z     15.1501    6.6490   2.2785    0.0250   1.9445  28.3556   0.01    1.01
alpha    0.0166    0.0182   0.9090    0.3657  -0.0196   0.0528   0.01    1.01

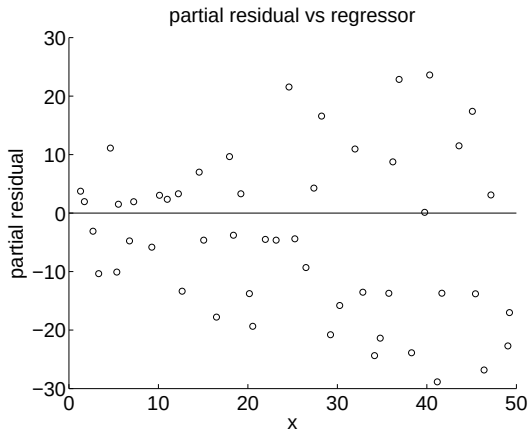
Regression ANVOA
Source       SS       DF       MS   F-value  p-value
Regr    3370.65     7.00   481.52      1.35     0.24
Resid  32764.65    92.00   356.14
-----------------------------------------------------
Tot  36135.30    99.00
```

LM 5 In Tab. 3.2 auf Seite 64 findet sich eine Zusammenstellung aller im Zuge dieser Arbeit

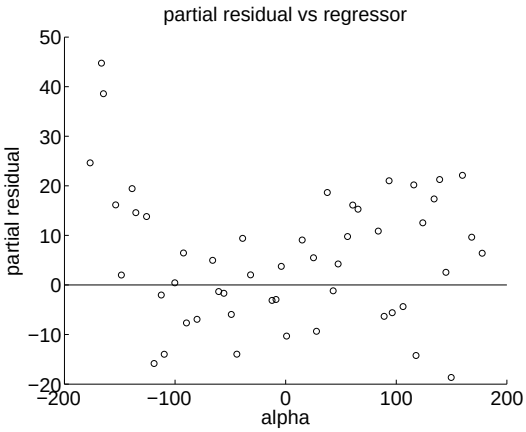
geschätzten LM. Dabei handelt es sich bei den LM 2 bis 5 bereits um jene Modelle, welche von den Selektionsverfahren ausgewählten wurden. Die in der ID-Spalte angegebenen Werte entsprechen der in Tab. 3.1 auf Seite 59 eingeführten kompakten Modelldarstellung. Diesbezüglich muss leider eingeräumt werden, dass der teilweise Ausschluss von Interaktionstermen eines Regressors x_i im Zuge der Modellselektion, mit Hilfe dieser kompakten Darstellung nicht korrekt abgebildet werden kann. Dennoch vermittelt sie ein hinreichend genaues Bild des tatsächlich geschätzten Modells und ist auch völlig ausreichend, um den im Kapitel 2.5 angeführten Sachverhalt aufzuzeigen. Wie Tab. 3.2 nämlich deutlich zeigt, wurde von den Selektionsverfahren für das selbe Startmodell, für jedes Design ein anderes zu schätzendes LM ausgewählt.

Von all diesen Modellen erwies sich das für $n = 50$ Designs gefundene LM 5 als das, im Sinne des in 2.5.1 auf Seite 32 angeführten Modellwahlkriteriums R^2_{adj}, beste Regressionsmodell. Ebendieses soll nun im Weiteren näher betrachtet werden. Abbildung 3.3 auf der nächsten Seite zeigt für zwei ausgewählte Regressoren zunächst den Component-plus-residual plot. Wie zu erkennen ist, wird die Homoskedastizitätsannahme (A4) deutlich verletzt. Dies ist auch für die Component-plus-residual plots der übrigen Regressoren zutreffend. Demzufolge stellen die mit Hilfe der KQ-Schätzer ermittelten Regressionsparameter nicht mehr die minimalvarianten Schätzer dar (vgl. 2.3 auf Seite 21). Dieses Varianzproblem wird ebenso von den zahlreichen explodierenden Varianz-Inflations Faktoren und den sehr großen Konfidenzintervallen in Listing 3.2 auf Seite 65 deutlich aufgezeigt.

Die übrigen in Kapitel 2.6 vorgestellten visuellen und formalen Diagnosemethoden liefern hingegen keine Auffälligkeiten. In Abb. 3.4(a) auf der nächsten Seite und 3.4(b) finden sich als Beispiel der run-sequence-order plot und der quantil-quantil-plot. In Listing 3.2 sind wieder einige wesentliche formale Ergebnisse dargestellt.

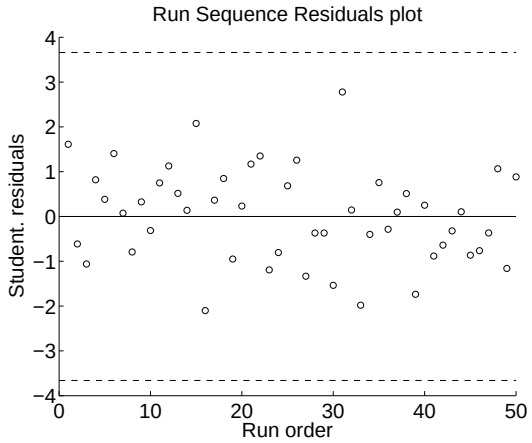


(a) Die partiellen Residuen der x-Komponente des Translationsvektors zeigen eindeutig heteroskedastische Eigenschaften

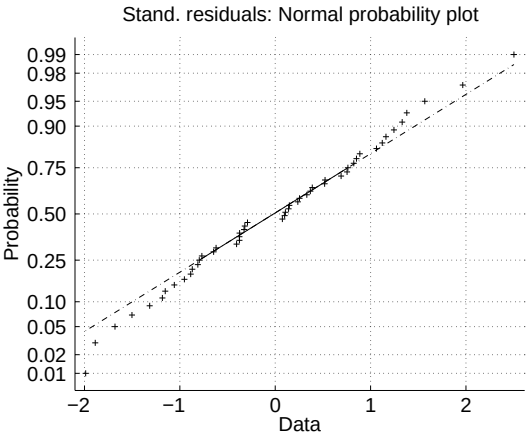


(b) Für den Drehwinkel α ist die Verletzung der Homoskedastizitätsannahme aufgrund der kubischen Modellierung nicht so offensichtlich

Abbildung 3.3: Component-plus-residual plots am Beispiel des LM 5 mit $n = 50$ Designstellen



(a) Ein unauffälliger run-sequence-order plot



(b) Genauso wie die formalen Tests bestätigt der quantil-quantil plot die Normalverteilungsannahme der Residuen

Abbildung 3.4: Zwei Diagnostikplots zum Beispiel des LM 5 mit $n = 50$ Designstellen

LM Nr.	ID							Designs	AIC_c	AIC	R^2	R^2_{adj}
1	1	1	1	1	1	1	1	20	133.72	120.63	0.32	-0.08
2	0	0	0	0	0	0	1	20	113.58	112.88	0.16	0.11
3	4	0	0	0	0	0	4	20	109.77	108.27	0.36	0.32
1	1	1	1	1	1	1	1	50	281.80	278.29	0.05	-0.11
2	1	0	0	0	0	0	0	50	268.06	276.81	0.02	0
3	0	0	0	0	0	0	4	50	267.07	266.81	0.04	0.02
4	2	2	3	7	7	2	6	50	262.61	254.18	0.50	0.35
5	2	15	3	15	11	6	15	50	257.88	224.86	0.81	0.67
1	1	1	1	1	1	1	1	100	596.77	595.19	0.09	0.02
2	0	0	0	1	0	1	0	100	587.97	587.72	0.07	0.05
3	0	0	0	4	0	4	0	100	587.79	587.54	0.07	0.05
4	0	5	0	1	5	1	0	100	580.67	579.46	0.21	0.16
5	12	0	0	8	12	1	8	100	580.51	576.28	0.32	0.23

Tabelle 3.2: Modellzusammenstellung

Listing 3.2: Einige wesentliche Kenngrößen des LM5 für $n = 50$ Designs

```
Akaike Information Criteria (AIC) = 224.8557
AIC bias-corrected = 257.8557

Mean squared error =  66.8092
R^2 = 0.8054
Adjusted R^2 = 0.6712

Coefficients in the Equation
                                        Conf. inter.
Coeff         B      SE B    Sig T    lower    upper R(j)^2    VIF
const    1.7646   13.1514   0.8942 -25.1331  28.6623
y        4.3816    0.9064 < 0.001    2.5277   6.2354   0.99 128.76
z       -0.3813    0.1886   0.0525  -0.7669   0.0044   0.82   5.56
k_x    123.0751   52.7125   0.0267  15.2660 230.8843   0.99 169.79
k_y     64.0319   18.0937   0.0014  27.0263 101.0376   0.95  19.41
alpha    0.1888    0.0483 < 0.001    0.0900   0.2876   0.95  19.11
x y     -0.0055    0.0030   0.0719  -0.0116   0.0005   0.47   1.87
x alpha  0.0036    0.0009 < 0.001    0.0017   0.0056   0.85   6.71
y alpha -0.0036    0.0009 < 0.001   -0.0055  -0.0017   0.81   5.31
z k_x    1.0761    0.3696   0.0069   0.3201   1.8320   0.92  12.95
z alpha -0.0034    0.0009   0.0010  -0.0053  -0.0015   0.83   5.72
k_x k_y 16.6302   17.2115   0.3419 -18.5713  51.8317   0.91  11.06
k_y k_z -56.2654 14.8170 < 0.001  -86.5696 -25.9612   0.86   6.99
y^2     -0.1781    0.0416 < 0.001   -0.2633  -0.0929   1.00 700.64
k_x^2 -291.8883 113.2565   0.0153 -523.5238 -60.2529  1.00 839.46
k_z^2   14.9204    8.0023   0.0724  -1.4461  31.2868   0.76   4.18
alpha^2  0.0005    0.0001   0.0012   0.0002   0.0008   0.22   1.29
y^3      0.0021    0.0006 < 0.001    0.0010   0.0033   1.00 273.41
k_x^3  160.2269   74.7977   0.0407   7.2485 313.2052   1.00 331.87
k_y^3  -46.4115   12.2428 < 0.001  -71.4508 -21.3721   0.88   8.64
alpha^3 -0.0000    0.0000 < 0.001   -0.0000  -0.0000   0.88   8.48

Regression ANVOA
Source       SS       DF       MS   F-value  p-value
  Regr  8017.39    20.00   400.87     6.00     0.00
 Resid  1937.47    29.00    66.81
  ------------------------------------------------------
  Tot   9954.85    49.00
```

Die auf diesem Modell aufbauende Optimierung wurde mit Hilfe der MATLAB®-Funktion `fmincon` durchgeführt. Zu diesem Zweck wurden wieder mittels `haltonset`-Funktion einige wenige über dem Definitionsbereich gleichmäßig verteilte Punkte generiert und

dem Optimierungsalgorithmus als Startpunkte übergeben. Die derart vom LM prognostizierten, unterschiedlichen lokalen (globalen) Minima werden in Tab. 3.3 auf dieser Seite aufgelistet und mit den zu diesen Punkten gehörenden tatsächlichen Werten verglichen. Die wahren Werte wurden wieder mit Hilfe von speedyne ermittelt und als Maxima einer über die Simulationsdauer geglätteten Vergleichsspannungskurve bestimmt. Wie leider eindeutig zu erkennen ist, liegen die prognostizierten Werte deutlich neben den wahren Werten. Dies ist zwar aufgrund des definitiven Kollinearitätsproblems nicht weiter verwunderlich, allerdings liegen die wahren Werte auch nicht in den Prognoseintervallen. Ferner dürfte auch die Geometrie des LM den tatsächlichen kausalen Zusammenhang zwischen dem Parametervektor 3.1 auf Seite 57 und den Vergleichsspannungen nicht korrekt abbilden. Dies wird im Vergleich der unterschiedlichen prognostizierten Minima mit dem Verlauf der wahren Werte ersichtlich. Während ersteres Problem eventuell mit Hilfe der in 2.6.4 auf Seite 52 vorgestellten Zentrierung, oder einer vielversprechenderen Hauptkomponentenanalyse in den Griff zu bekommen sein dürfte, dürfte zweiteres Problem in der Klasse der linearen Regressionsmodelle nicht gelöst werden können. Diesbezüglich scheint es eher ratsam zu sein, den unbekannten Zusammenhang mit Hilfe eines *nichtparametrischen Regressionsmodells* zu schätzen.

Resümee, Ausblick Mit Hilfe einer der Optimierung vorgeschalteten linearen Regressionsfläche sollte im Zuge dieser Arbeit gezeigt werden, dass für eine gegebene diskretisierte Finite-Elemente Struktur eine zeit- und speichereffiziente Methode entwickelt werden kann, die an die Ergebnisse einer direkten Optimierung der Bewegungsgleichungen in (0.1) auf Seite VII hinreichend genau heran kommt. Zu diesem Zweck wurden im Zuge dieser Arbeit die theoretischen Grundlagen zur schrankenrestringierten Optimierung, und zur Parameterschätzung sowie zur Diagnose von linearen Regressionsmodellen zusammengetragen. Vor allem im linearen Regressionskapitel wurden die erörterten Methoden auch mit kleinen Beispielen ergänzt. Schließlich wurde u. a. mit den Softwareprodukten MATLAB® und speedyne eine Softwareumgebung implementiert, mit deren Hilfe die Arbeitshypothese an einem Falltest-

Parameter							LM	Prognoseintervall		speedyne
								unten	oben	
0	50	50	1	0.76	0	-180	194.88	159.11	230.65	22.58
50	13.80	50	1	0.76	0	113.63	110.71	84.08	137.34	49.58

Tabelle 3.3: Die Tabelle zeigt einen Vergleich der prognostizierten Optima, mit den von speedyne ermittelten „exakten“ Vergleichsspannungswerten.

beispiel getestet werden sollte. Das dabei von Selektionsalgorithmen gefundene beste lineare Regressionsmodell konnte den kausalen Zusammenhang der Regressoren und der abhängigen Variable aber leider nicht korrekt abbilden. Zwar wurde die Arbeitshypothese nur an einem einzigen Beispiel getestet, dennoch scheint es so, als wären einfache lineare Regressionsmodell für diese Aufgabenstellung ungeeignet.

Eine noch zu untersuchende Möglichkeit, die auf der in dieser Arbeit verfolgten Strategie aufbaut, wäre die Verwendung eines *nicht- oder semiparametrischen Regressionsmodells*. Da dort die Effekte mit automatisierten und mathematisch raffinierteren Schätztechniken untersucht werden, besteht die Hoffnung damit den kausalen Zusammenhang aufzudecken.

Tabellenverzeichnis

2.1 ANOVA-Tafel des globalen F-Tests 28

3.1 Regressionsmodelle im Testbeispiel 59
3.2 Modellzusammenstellung 64
3.3 Vergleich der Optima 66

Abbildungsverzeichnis

2.1 Der Effekt eines Regressors am Beispiel des Teilmodells (2.4) 41
2.2 Residualplots am Beispiel des Teilmodells (2.4) 43
2.3 Component-plus-residual plot am Beispiel des Teilmodells (2.4) 43
2.4 Autokorrelation und Heteroskedastizität 45
2.5 Residuenverteilung am Beispiel des Teilmodells (2.5) 47
2.6 Ausreißeranalyse 51

3.1 Beispielausgabe für eine moving-average Filterung 59
3.2 Component-plus-residual plot des LM 1 60
3.3 Component-plus-residual plot des LM 5 63
3.4 Modelldiagnostik des LM 5 63

Literaturverzeichnis

NIST/SEMATECH e-handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/, 2006. Besucht am 13. Dezember 2010.

Albing, Carl, Vossen, JP und Newham, Cameron: *bash Cookbook - Solutions and Examples for bash Users*. O'Reilly Media, Inc., Sebastopol, erste Auflage, 2007.

Bathe, Klaus-Jürgen: *Finite Elemente Methoden*. Springer-Verlag, Berlin Heidelberg, zweite Auflage, 2002.

Bucher, Christian: Stabilization of explicit time imtegration by modal reduction. In: Wall, W., Bletzinger, K. und Schweizerhof, K. (Hg.) *Trends in Computational Structural Mechanics*. CIMNE, 2001, S. 429–437.

Bucher, Christian: *Computational analysis of randomness in structural mechanics*. CRC Press, 2009.

Fahrmeir, Ludwig, Kneib, Thomas und Lang, Stefan: *Regression - Modelle, Methoden und Anwendungen*. Springer, Berlin, Heidelberg, zweite Auflage, 2009.

Fox, John: *Applied regression analysis and generalized linear models*. Sage, Los Angeles, Calif. u.a., zweite Auflage, 2008.

Furnival, George M. und Wilson, Robert W. Jr.: Regression by leaps and bounds. In: *Technometrics*, Band 16(4):S. 499–511, 1974.

Gunst, Richard F.: Classical studies that revolutionized the practice of regression analysis. In: *Technometrics*, Band 42(1):S. 62–64, 2000.

Hintermüller, Michael: Numerische Verfahren in der Optimierung. Vorlesungsskript WS 2001/2002, URL http://www.math.uni-konstanz.de/fb_seiten/contrib/studium/vorlesungsverzeichnis/BA_SS09_17-03-2009.pdf.

Miller, Allan J.: *Subset selection in regression*, Band 95 von *Monographs on statistics and applied probability*. Chapman & Hall, CRC, Boca Raton, Fla. u. a., zweite Auflage, 2002.

Niederreiter, Harald: *Random number generation and quasi-Monte Carlo methods.* Society for Industrial and Applied Mathematics, Philadelphia, 1992.

Nocedal, Jorge und Wright, Stephen J.: *Numerical optimization.* Springer Science+Business Media, LLC., New York, zweite Auflage, 2006.

Plato, Robert: *Numerische Mathematik kompakt.* Vieweg+Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden, vierte Auflage, 2010.

Pruscha, Helmut: *Statistisches Methodenbuch - Verfahren, Fallstudien, Programmcodes.* Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2006.

Urban, Dieter und Mayerl, Jochen: *Regressionsanalyse: Theorie, Technik und Anwendung.* VS Verlag für Sozialwissenschaften, 2008. ISBN 978-3-531-91194-6.

Viertl, Reinhard: *Einführung in die Stochastik.* Springer-Verlag, Wien, 1990.

Wayne, A. Larsen und McCleary, Susan J.: The use of partial residual plots in regression analysis. In: *Technometrics*, Band 14(3):S. 781–790, 1972.

Wolf, Jürgen: *Shell-Programmierung - das umfassende Handbuch.* Galileo Press, Bonn, dritte Auflage, 2010.

Xin, Yan und Xiao, Gang Su: *Linear Regression Analysis - Theroy and Computing.* World Scientific, Singapore u.a., 2009.