Statistik und ihre Anwendungen

Ludwig Fahrmeir
Thomas Kneib
Stefan Lang

Regression

Modelle, Methoden und Anwendungen

2. Auflage

Springer

Reihenherausgeber:
Prof. Dr. Holger Dette • Prof. Dr. Wolfgang Härdle

Statistik und ihre Anwendungen

Azizi Ghanbari, S.
Einführung in die Statistik für Sozial- und Erziehungswissenschaftler 2002

Bickeböller, H.; Fischer, C.
Einführung in die Genetische Epidemiologie 2007

Brunner, E.; Munzel, U.
Nichtparametrische Datenanalysen 2002

Dehling, H.; Haupt, B.
Einführung in die Wahrscheinlichkeitstheorie und Statistik
2. Auflage 2004

Dümbgen, L.
Stochastik für Informatiker 2003

Falk, M.; Becker, R.; Marohn, F.
Angewandte Statistik 2004

Franke, J.; Härdle, W.; Hafner, C.
Einführung in die Statistik der Finanzmärkte
2. Auflage 2004

Greiner, M.
Serodiagnostische Tests 2003

Handl, A.
Multivariate Analysemethoden 2003

Hassler, U.
Stochastische Integration und Zeitreihenmodellierung 2007

Hilgers, R. -D.; Bauer, R.; Scheiber, V.
Einführung in die Medizinische Statistik 2. Auflage 2007

Kohn, W.
Statistik Datenanalyse und Wahrscheinlichkeitsrechnung 2005

Kreiß, J. -P.; Neuhaus, G.
Einführung in die Zeitreihenanalyse 2006

Ligges, U.
Programmieren mit R 2. Auflage 2007

Meintrup, D.; Schäffler, S.
Stochastik Theorie und Anwendungen 2005

Plachky, D.
Mathematische Grundbegriffe der Stochastik 2002

Pruscha, H.
Statistisches Methodenbuch Verfahren, Fallstudien,
Programmcodes 2005

Schumacher, M.; Schulgen, G.
Methodik klinischer Studien 2. Auflage 2007

Steland, A.
Mathematische Grundlagen der empirischen Forschung 2004

Zucchini, W.; Schlegel, A.; Nenadic, O.; Sperlich, S.
Statistik für Bachelor- und Masterstudenten 2009

Ludwig Fahrmeir
Thomas Kneib
Stefan Lang

Regression

Modelle, Methoden und Anwendungen

Zweite Auflage

Springer

Professor Dr. Ludwig Fahrmeir
Institut für Statistik
Ludwig-Maximilians-Universität München
Ludwigstraße 33
80539 München
Deutschland
ludwig.fahrmeir@stat.uni-muenchen.de

Prof. Dr. Stefan Lang
Institut für Statistik
Leopold-Franzens-Universität Innsburg
Universitätsstraße 15
6020 Innsburg
Österreich
stefan.lang@uibk.ac.at

Prof. Dr. Thomas Kneib
Institut für Mathmatik
Carl von Ossietzky Universität Oldenburg
26111 Oldenburg
Deutschland
thomas.kneib@uni-oldenburg.de

ISBN 978-3-642-01836-7 e-ISBN 978-3-642-01837-4
DOI 10.1007/978-3-642-01837-4
Springer Heidelberg Dordrecht London New York

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

© Springer-Verlag Berlin Heidelberg 2007, 2009
Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland vom 9. September 1965 in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtsgesetzes.
Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Einbandentwurf: WMXDesign GmbH, Heidelberg

Gedruckt auf säurefreiem Papier

Springer ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Vorwort zur 1. Auflage

Regression ist die wohl am häufigsten eingesetzte statistische Methodik zur Analyse empirischer Fragestellungen in Wirtschafts-, Sozial- und Lebenswissenschaften. Dementsprechend existiert auch eine Vielfalt von Modellklassen und Inferenzmethoden, ausgehend von der klassischen linearen Regression bis hin zur modernen nicht- und semiparametrischen Regression. Zu den meisten speziellen Klassen von Regressionsmodellen gibt es bereits eigene Lehrbücher. Diese variieren zum Teil stark in Stil, mathematisch-theoretischem Niveau und Anwendungsorientierung. Warum nun noch ein Buch über Regression? Einer hohen Zahl von einführenden Texten zur linearen Regression, die sich vornehmlich an Studierende und Praktiker aus verschiedenen Anwendungsbereichen richten, steht eine vergleichsweise kleine Zahl von Texten zur modernen nicht- und semiparametrischen Regression gegenüber, die jedoch in mathematisch-formaler Hinsicht wesentlich anspruchsvoller und auch deutlich theoretischer angelegt sind.

Ziel dieses Buches ist eine anwendungsorientierte, einheitliche Einführung in die parametrische, nicht- und semiparametrische Regression, mit der diese bestehende Lücke zwischen Theorie und Praxis geschlossen wird. Wesentliches Auswahlkriterium für die behandelten Methoden ist dabei insbesondere die Verfügbarkeit geeigneter, benutzerfreundlicher Software gewesen. Auf solider formaler Basis werden die wichtigsten Modelle und Verfahren der Regressionsanalyse dargestellt und deren sachgerechte Anwendung vermittelt. Wir sehen dies sowohl für den Fortschritt in vielen Anwendungsdisziplinen als auch für die Entwicklung der methodischen Statistik, die ihre Motivation aus neuen praktischen Herausforderungen erhält, als wichtig an. Ein ähnliches Ziel, mit etwas anderen Schwerpunkten, verfolgen Ruppert, Wand & Carroll (2003) mit ihrem Buch „Semiparametric Regression“.

Damit wendet sich dieses Buch insbesondere an Studierende, Dozenten und Praktiker in den Wirtschafts-, Sozial und Lebenswissenschaften, an Studierende und Dozenten des Fachs Statistik, sowie an Mathematiker und Informatiker mit Interesse an statistischer Modellierung und Datenanalyse. Das Buch ist soweit wie möglich eigenständig lesbar und setzt lediglich Kenntnisse zur elementaren Wahrscheinlichkeitsrechnung und Statistik voraus, wie sie etwa in dem Einführungsbuch von Fahrmeir, Künstler, Pigeot & Tutz (2007) vermittelt werden. Teile des Buches, die kompliziertere Details behandeln oder zusätzliche Informationen beinhalten, die nicht unmittelbar zum Verständnis der vorgestellten Methoden notwendig sind und damit beim ersten Lesen übersprungen werden können, werden am Anfang durch das Symbol ◬ und am Ende durch das Symbol ⌧ am Rand gekennzeichnet. Die wichtigsten Definitionen und Aussagen werden in Kästen kompakt zusammengefasst. In zwei Anhängen werden die notwendigen Grundlagen zur Matrix-Algebra, sowie zur Wahrscheinlichkeitsrechnung und induktiven Statistik kompakt dargestellt.

Abhängig von Interessen und Schwerpunkten können Teile des Buches auch unabhängig von anderen Teilen und auch in veränderter Reihenfolge gelesen werden:

- Kapitel 2 bietet eine einführende Übersicht zu parametrischen sowie nicht- und semiparametrischen Regressionsmodellen, wobei auf die statistische Inferenz und technische Details bewusst verzichtet wird.
- Die Kapitel 1 – 3 sind als Einführung in lineare Modelle geeignet.
- Lineare gemischte Modelle (Kapitel 6.1 – 6.6) können als Erweiterung linearer Modelle direkt anschließend, ohne Vorkenntnisse aus den Kapiteln 4 und 5, gelesen werden.
- Die Kapitel 1 – 5 umfassen parametrische Regressionsmodelle für stetige und diskrete Zielvariablen.
- Schließlich können auch die Kapitel 1 – 3, 7 und die Abschnitte 8.1 – 8.3 als Einführung in die parametrische und semiparametrische Regression für stetige Zielvariablen studiert werden.
- Darauf aufbauend sind Erweiterungen für diskrete Zielvariablen dann in Kapitel 4 (generalisierte lineare Modelle), Kapitel 5 (kategoriale Regression), Abschnitt 6.7 – 6.8 (generalisierte lineare gemischte Modelle) und Abschnitt 8.4 (strukturiert-additive Regression) dargestellt.

Zahlreiche Anwendungsbeispiele aus unterschiedlichen Bereichen illustrieren die Modelle und Methoden. Die meisten zugehörigen Datensätze sind über die Homepage zum Buch

`http://www.statistik.lmu.de/ kneib/regressionsbuch/`

beziehungsweise über

`http://www.springer.de`

erhältlich und ermöglichen so auch das eigenständige Studium mit Hilfe realer Beispiele. Darüber hinaus enthält die Homepage auch Hinweise zu statistischer Software mit deren Hilfe die vorgestellten Verfahren angewendet werden können, neueste Informationen zum Buch und ausführlichere Versionen der Appendices zur Matrix-Algebra sowie zur Wahrscheinlichkeitstheorie und Inferenz.

Wie fast immer verbleiben auch in diesem Buch einige Lücken. Diese betreffen insbesondere Regressionsmodelle für Lebensdauern und multivariate Zielvariablen. Da unsere Vorgehensweise eher explorativ ausgerichtet ist, haben wir auch bewusst auf viele spezielle Tests, die insbesondere in der ökonometrischen Literatur populär sind, verzichtet.

Für die Hilfe und Unterstützung beim Schreiben von Teilen des Textes, bei der Ausarbeitung von Beispielen und beim Korrekturlesen bedanken wir uns insbesondere bei Kathrin Dallmeier, Oliver Joost, Franziska Kohl, Jana Lehmann, Cornelia Oberhauser, Sylvia Schmidt, Sven Steinert und Peter Wechselberger. Unser Dank gilt auch Lilith Braun und Christiane Beisel vom Springer Verlag für die stets freundliche, sehr gute und vor allen Dingen geduldige Zusammenarbeit.

München & Innsbruck,
Februar 2007

Ludwig Fahrmeir
Thomas Kneib
Stefan Lang

Vorwort zur 2., durchgesehenen Auflage

Bei der vorliegenden 2. Auflage handelt es sich um eine durchgesehene und verbesserte Version der ersten Auflage. Wir danken allen Kollegen, Freunden, Mitarbeitern und Studenten für Hinweise auf Fehler und Unstimmigkeiten ebenso wie für die positiven Rückmeldungen, die maßgeblich zur Fertigstellung dieser 2. Auflage beigetragen haben. Unser Dank gilt außerdem Niels Peter Thomas vom Springer Verlag für die Unterstützung bei der zügigen Erstellung der 2. Auflage.

Um eine dauerhaft zugängliche Internet-Präsenz für das Buch zu schaffen, ist die Homepage nun unter

`http://www.regressionbook.org/`

erreichbar.

München, Oldenburg & Innsbruck,
Mai 2009

Ludwig Fahrmeir
Thomas Kneib
Stefan Lang

Inhaltsverzeichnis

1 Einführung . . . 1
1.1 Anwendungsbeispiele . . . 4
1.2 Erste Schritte . . . 11
1.2.1 Beschreibung der Verteilung der Variablen . . . 11
1.2.2 Grafische Zusammenhangsanalyse . . . 13
Stetige erklärende Variablen . . . 13
Kategoriale erklärende Variablen . . . 16

2 Regressionsmodelle . . . 19
2.1 Einführung . . . 19
2.2 Lineare Regressionsmodelle . . . 20
2.2.1 Das einfache lineare Regressionsmodell . . . 20
2.2.2 Das multiple lineare Regressionsmodell . . . 24
2.3 Regression bei binären Zielvariablen: Das Logit-Modell . . . 30
2.4 Gemischte Modelle . . . 35
2.5 Einfache nichtparametrische Regression . . . 40
2.6 Additive Regression . . . 44
2.7 Generalisierte additive Regression . . . 47
2.8 Geoadditive Regression . . . 49
2.9 Modelle im Überblick . . . 55
2.9.1 Lineare Modelle (LM, Kapitel 3) . . . 55
2.9.2 Logit-Modell (Kapitel 4) . . . 56
2.9.3 Poisson-Regression (Kapitel 4) . . . 56
2.9.4 Generalisierte lineare Modelle (GLM, Kapitel 4, 5) . . . 56
2.9.5 Lineare gemischte Modelle (LMM, Kapitel 6) . . . 56
2.9.6 Additive Modelle und Erweiterungen (AM, Kapitel 7, 8) . . . 57
2.9.7 Generalisierte additive (gemischte) Modelle (GAMM, Kapitel 8) . . . 58
2.9.8 Strukturiert-additive Regression (STAR, Kapitel 8) . . . 58

3 Lineare Regressionsmodelle . . . 59
3.1 Das klassische lineare Modell . . . 59
3.1.1 Modelldefinition . . . 59
3.1.2 Modellparameter, Schätzungen und Residuen . . . 63
3.1.3 Diskussion der Modellannahmen . . . 64
Linearität des Einflusses der Kovariablen . . . 64
Homoskedastische Varianz der Störgrößen . . . 64

Unkorreliertheit der Störgrößen . . . 66
Additivität der Störgrößen . . . 70
3.1.4 Modellierung des Einflusses der Kovariablen . . . 72
Metrische Kovariablen . . . 72
Kategoriale Kovariablen . . . 80
Interaktionen zwischen Kovariablen . . . 83
3.2 Parameterschätzungen . . . 90
3.2.1 Schätzung der Regressionskoeffizienten . . . 90
Die Methode der kleinsten Quadrate . . . 90
Maximum-Likelihood-Schätzung . . . 92
Geschätzte Werte und Residuen . . . 93
3.2.2 Schätzung der Varianz der Störgrößen . . . 94
Maximum-Likelihood-Schätzung . . . 94
Restringierte Maximum-Likelihood-Schätzung . . . 94
3.2.3 Eigenschaften der Schätzungen . . . 95
Geometrische Eigenschaften des KQ-Schätzers . . . 95
Streuungszerlegung und Bestimmtheitsmaß . . . 98
Statistische Eigenschaften ohne spezielle Verteilungsannahmen . . . 101
Statistische Eigenschaften bei Normalverteilungsannahme . . . 103
Asymptotische Eigenschaften des KQ-Schätzers . . . 105
Statistische Eigenschaften der Residuen . . . 107
Standardisierte und studentisierte Residuen . . . 108
3.3 Hypothesentests und Konfidenzintervalle . . . 111
3.3.1 F-Test . . . 113
Zusammenhang mit dem Wald-Test . . . 115
F-Test für einige spezielle Testprobleme . . . 115
Asymptotische Eigenschaften des F-Tests . . . 119
3.3.2 Konfidenzbereiche und Prognoseintervalle . . . 119
Konfidenzintervalle und Ellipsoide für die Regressionskoeffizienten 119
Prognoseintervalle . . . 121
3.4 Das allgemeine lineare Regressionsmodell . . . 124
3.4.1 Modelldefinition . . . 124
3.4.2 Gewichtete Methode der kleinsten Quadrate . . . 125
Gruppierte Daten . . . 127
3.4.3 Heteroskedastische Fehler . . . 128
Diagnose heteroskedastischer Fehler . . . 129
Maßnahmen bei Heteroskedastizität . . . 132
3.4.4 Autokorrelierte Fehler . . . 136
Autokorrelation erster Ordnung . . . 137
Diagnose autokorrelierter Störungen . . . 139
Maßnahmen bei Autokorrelation erster Ordnung . . . 142

3.5 Bayesianische lineare Modelle . . . 147
3.5.1 Priori-Verteilungen . . . 147
3.5.2 Vollständig bedingte Dichten und MCMC-Inferenz . . . 149
3.5.3 Posteriori-Verteilung . . . 152
3.6 Modellwahl und Variablenselektion . . . 152
3.6.1 Auswirkunge Bias, Varianz und Prognosegüte . . . 155
Auswirkungen auf Bias und Varianz des KQ-Schätzers . . . 156
Auswirkung der Modellspezifikation auf die Prognosegüte . . . 157
3.6.2 Modellwahlkriterien . . . 159
Das korrigierte Bestimmtheitsmaß . . . 160
Mallows' C_p . . . 161
Informationskriterium nach Akaike AIC . . . 161
Kreuzvalidierung . . . 161
Bayesianisches Informationskriterium BIC . . . 162
3.6.3 Praktische Verwendung der Modellwahlkriterien . . . 163
3.6.4 Modelldiagnose . . . 168
Überprüfen der Modellannahmen . . . 168
Kollinearitätsanalyse . . . 170
Ausreißer- und Einflussanalyse . . . 173
Alternative Modellierungsansätze nach Modelldiagnose . . . 179
3.7 Bemerkungen und Ergänzungen . . . 180
3.7.1 Literaturhinweise . . . 180
3.7.2 Beweise . . . 181

4 Generalisierte lineare Modelle . . . 189
4.1 Binäre Regression . . . 189
4.1.1 Binäre Regressionsmodelle . . . 189
Logit-Modell . . . 190
Probit-Modell . . . 191
Komplementäres log-log-Modell . . . 191
Binäre Modelle als Schwellenwertmodelle latenter linearer Modelle 193
Parameterinterpretation . . . 194
Gruppierte Daten . . . 195
Überdispersion (Overdispersion) . . . 197
4.1.2 Maximum-Likelihood-Schätzung . . . 198
Vergleich mit ML- bzw. KQ-Schätzung im linearen Modell . . . 201
Iterative numerische Berechnung des ML-Schätzers . . . 202
Asymptotische Eigenschaften des ML-Schätzers . . . 203
4.1.3 Testen linearer Hypothesen . . . 204
4.1.4 Kriterien zur Modellanpassung und Modellwahl . . . 205
4.2 Regression für Zähldaten . . . 210

4.2.1 Modelle für Zähldaten . . . 210
Log-lineares Poisson-Modell . . . 210
Lineares Poisson-Modell . . . 210
Überdispersion . . . 210
4.2.2 Schätzen und Testen: Likelihood-Inferenz . . . 212
Maximum-Likelihood-Schätzung . . . 212
Testen linearer Hypothesen . . . 213
Kriterien zur Modellanpassung und Modellwahl . . . 213
Schätzung des Überdispersions-Parameters . . . 213
4.3 Modelle für positive stetige Zielvariablen . . . 215
Gamma-Regression . . . 217
Inverse Gauß-Verteilung . . . 217
4.4 Generalisierte Lineare Modelle . . . 217
4.4.1 Allgemeine Modelldefinition . . . 217
4.4.2 Likelihood-Inferenz . . . 220
Asymptotische Eigenschaften des ML-Schätzers . . . 223
Schätzung des Skalierungs- oder Überdispersionsparameters . . . 224
Testen linearer Hypothesen . . . 224
Kriterien zur Modellanpassung und Modellwahl . . . 225
4.5 Quasi-Likelihood-Modelle . . . 226
4.6 Bayesianische generalisierte lineare Modelle . . . 228
4.7 Bemerkungen und Ergänzungen . . . 233

5 Kategoriale Regressionsmodelle . . . 235
5.1 Einführung . . . 235
Multinomialverteilung . . . 236
Daten . . . 237
5.2 Modelle für ungeordnete Kategorien . . . 238
Nominale Modelle und latente Nutzenmodelle . . . 241
5.3 Ordinale Modelle . . . 242
Das kumulative oder Schwellenwert-Modell . . . 242
Das sequentielle Modell . . . 245
5.4 Schätzen und Testen: Likelihood-Inferenz . . . 247
Numerische Bestimmung des ML-Schätzers . . . 249
Asymptotische Eigenschaften und Tests linearer Hypothesen . . . 249
5.5 Bemerkungen und Ergänzungen . . . 252

6 Gemischte Modelle . . . 253
6.1 Lineare gemischte Modelle für Longitudinal- und Clusterdaten . . . 254
6.2 Das allgemeine lineare gemischte Modell . . . 259
6.3 Likelihood-Inferenz für LMM . . . 261
6.3.1 Schätzung bei bekannter Kovarianzstruktur . . . 261

6.3.2 Schätzung der Kovarianzstruktur . . . 263
6.3.3 Schätzung fixer und zufälliger Effekte . . . 264
6.3.4 Hypothesentests . . . 266
6.4 Likelihood-Inferenz für Longitudinal- und Clusterdaten-Modelle . . . 268
6.5 Bayesianische gemischte lineare Modelle . . . 271
Posteriori-Verteilung bei bekannter Kovarianzstruktur . . . 273
Empirische Bayes-Schätzung . . . 273
Volle Bayes-Schätzung . . . 274
6.6 Generalisierte lineare gemischte Modelle . . . 278
6.6.1 Definition und Eigenschaften von GLMM . . . 278
GLMM für Longitudinal- und Clusterdaten . . . 279
GLMM in allgemeiner Form . . . 279
Kategoriale gemischte Regressionsmodelle . . . 282
6.7 Likelihood- und Bayes-Inferenz in GLMM . . . 284
6.7.1 Penalisierte Likelihood- und empirische Bayes-Schätzung . . . 284
6.7.2 Volle Bayes-Inferenz mit MCMC . . . 287
6.8 Bemerkungen und Ergänzungen . . . 289

7 Nichtparametrische Regression . . . 291
7.1 Univariate Glättung . . . 292
7.1.1 Polynom-Splines . . . 293
Polynom-Splines und trunkierte Potenzen . . . 296
Einfluss der Knoten auf die Schätzung . . . 301
B-Splines . . . 303
7.1.2 Penalisierte Splines (P-Splines) . . . 306
P-Splines basierend auf der TP-Basis . . . 307
P-Splines basierend auf B-Splines . . . 309
Penalisierte KQ-Schätzung . . . 311
Bayesianische P-Splines . . . 316
7.1.3 Allgemeine Penalisierungsansätze . . . 320
7.1.4 Glättungssplines . . . 323
7.1.5 Random Walks . . . 326
7.1.6 Kriging . . . 327
Klassisches Kriging . . . 327
Kriging als Glättungsverfahren für Zeitreihen . . . 330
Kriging als Glättungsverfahren der nichtparametrische Regression 331
7.1.7 Lokale Glättungsverfahren . . . 333
Nächste-Nachbarn-Schätzer . . . 333
Lokal polynomiale Regression und Nadaraya-Watson-Schätzer . . . 335
Loess . . . 339
7.1.8 Allgemeine Streudiagramm-Glätter . . . 340

Lineare Glättungsverfahren . . . 340
Konfidenzintervalle und -bänder . . . 342
Äquivalente Freiheitsgrade (effektive Parameterzahl) . . . 345
Schätzung der Fehlervarianz . . . 347
Bias-Varianz-Trade Off . . . 348
7.1.9 Wahl des Glättungsparameters . . . 350
Glättungsparameterwahl basierend auf Optimalitätskriterien . . . 350
Repräsentation von Penalisierungsansätzen als gemischte Modelle . 353
Bayesianische Glättungsparameterwahl basierend auf MCMC . . . 357
7.1.10 Adaptive Verfahren . . . 359
Multivariate adaptive Regressions-Splines (MARS) . . . 359
Regressionsbäume . . . 361
Bayesianische adaptive Verfahren I: Model Averaging . . . 364
Bayesianische adaptive Verfahren II: Reversible Jump MCMC . . . 366
7.2 Bivariate Glättung und räumliche Effekte . . . 368
7.2.1 Tensorprodukt-P-Splines . . . 371
Tensorprodukt-Basen . . . 371
2D-Penalisierungsansätze . . . 375
7.2.2 Radiale Basisfunktionen . . . 379
7.2.3 Kriging . . . 381
Klassische Geostatistik . . . 382
Kriging als Basisfunktionenansatz . . . 384
Schätzung von Kriging-Modellen . . . 385
7.2.4 Markov-Zufallsfelder . . . 387
Nachbarschaften und penalisiertes KQ-Kriterium . . . 387
Bayesianische Modellformulierung . . . 389
Räumlich autoregressive Prozesse . . . 393
7.2.5 Fazit . . . 393
7.2.6 Lokale und adaptive Glättungsverfahren . . . 394
7.3 Höherdimensionale Glättung . . . 395
7.4 Bemerkungen und Ergänzungen . . . 397

8 Strukturiert-additive Regression . . . 399
8.1 Additive Modelle . . . 399
8.2 Geoadditive Regression . . . 404
8.3 Modelle mit Interaktionen . . . 407
8.3.1 Modelle mit variierenden Koeffizienten . . . 408
8.3.2 Interaktion zwischen zwei metrischen Kovariablen . . . 410
8.4 Strukturiert-additive Regression . . . 413
8.5 Inferenz . . . 419
8.5.1 Penalisierte KQ- bzw- Likelihood-Schätzung . . . 420

Backfitting . . . 420
Direkte Minimierung des penalisierten KQ-Kriteriums . . . 421
Generalisierte STAR-Modelle . . . 422
Schätzung der Glättungsparameter . . . 422
Modellwahl und Diagnose . . . 423
8.5.2 Inferenz basierend auf der Repräsentation als gemischtes Modell . . 423
Modellwahl und Diagnose . . . 425
8.5.3 Bayesianische Inferenz mit MCMC . . . 425
Normalverteilte Zielgrößen . . . 425
Latente normalverteilte Zielgrößen . . . 427
Nicht-normalverteilte Zielgrößen . . . 428
Modellwahl und Diagnose . . . 428
8.5.4 Software-Hinweise . . . 430
8.6 Fallstudie: Unterernährung in Sambia . . . 431
8.6.1 Hinweise zur grundsätzlichen Vorgehensweise . . . 431
Deskriptive Analyse der Rohdaten . . . 431
Datenaufbereitung . . . 431
Grafische zweidimensionale Zusammenhangsanalyse . . . 432
Schätzung erster Arbeitsmodelle . . . 432
Modelldiagnose und Verfeinerung der Arbeitsmodelle . . . 432
Darstellung der Ergebnisse . . . 434
8.6.2 Deskriptive Analysen . . . 435
8.6.3 Modellierungsvarianten . . . 437
8.6.4 Schätzergebnisse und Modellevaluation . . . 438
8.7 Bemerkungen und Ergänzungen . . . 443

A Matrix-Algebra . . . 445
A.1 Definition und elementare Operationen . . . 445
A.2 Der Rang einer Matrix . . . 449
A.3 Determinante und Spur einer Matrix . . . 451
A.4 Verallgemeinerte Inverse . . . 452
A.5 Eigenwerte und Eigenvektoren . . . 453
A.6 Quadratische Formen . . . 455
A.7 Differentiation von Matrixfunktionen . . . 457

B Wahrscheinlichkeitsrechnung und induktive Statistik . . . 459
B.1 Einige eindimensionale Verteilungen . . . 459
B.2 Zufallsvektoren . . . 461
B.3 Die multivariate Normalverteilung . . . 464
B.3.1 Definition und Eigenschaften . . . 464
B.3.2 Die singuläre Normalverteilung . . . 465
B.3.3 Verteilungen quadratischer Formen . . . 466

B.3.4 Multivariate t-Verteilung 467
B.4 Likelihood-Inferenz 467
B.4.1 Maximum-Likelihood-Schätzung 467
B.4.2 Numerische Berechnung des ML-Schätzers 473
B.4.3 Asymptotische Eigenschaften des ML-Schätzers 475
B.4.4 Likelihood-basierte Tests für lineare Hypothesen 475
B.4.5 Modellwahl 477
B.5 Bayes-Inferenz 478
B.5.1 Grundlagen der Bayes-Inferenz 478
B.5.2 Punkt- und Intervallschätzer 480
Punktschätzer 480
Intervallschätzung 481
B.5.3 MCMC-Methoden 482
Metropolis-Hastings-Algorithmus 483
Gibbs-Sampler und Hybrid-Algorithmen 486
B.5.4 Modellwahl 488

Literaturverzeichnis 491

Index 497

1 Einführung

Sir Francis Galton (1822–1911) war ein äußerst vielseitiger Forscher, der in zahlreichen Disziplinen bahnbrechende Arbeiten verfasste. Unter Statistikern ist er vor allem für die Entwicklung des nach ihm benannten Galtonbretts zur Veranschaulichung der Binomialverteilung bekannt.

Ende des 19. Jahrhunderts beschäftigte sich Galton vorwiegend mit Fragen der Vererbung. Sein primäres Interesse galt der Frage, wie bestimmte Eigenschaften der Eltern auf die Nachkommen übertragen werden. Dazu sammelte Galton umfangreiche Daten, unter Anderem auch zum Vergleich der Körpergröße der Eltern und deren erwachsenen Kindern. Er untersuchte den *Zusammenhang* zwischen der Körpergröße der Kinder und einem Durchschnitt der Größen beider Eltern. Als Ausgleich für die natürlichen Größenunterschiede wurden die Körpergrößen der Frauen jeweils mit dem Korrekturfaktor 1.08 multipliziert. Um den Zusammenhang besser untersuchen zu können, stellte er die Daten in Form einer Kreuztabelle dar (Tabelle 1.1). Durch die Inspektion der Tabelle konnte er folgende, zur damaligen Zeit bahnbrechende Beobachtungen machen:

- Zeilenweise, d.h. bei festgehaltener Durchschnittsgröße der Eltern, folgen die Größen der erwachsenen Kinder annähernd einer Normalverteilung.
- Die Varianz der jeweiligen Normalverteilungen bleibt von Zeile zu Zeile konstant.
- Bildet man zeilenweise die Durchschnittsgrößen der Kinder, so liegen diese annähernd auf einer Geraden mit Steigung 2/3. Eine Steigung kleiner als Eins ließ Galton schlussfolgern, dass Kinder besonders großer Eltern tendenziell kleiner sind als ihre Eltern und umgekehrt Kinder kleiner Eltern tendenziell größer. In jedem Fall besteht eine Tendenz zum Populationsmittelwert. Galton sprach von *Regression* (Rückkehr) zum Mittelwert.

Später stellte Galton die Daten in Form eines Streudiagramms zwischen der Größe der Kinder und der Durchschnittsgröße der Eltern dar (Abbildung 1.1). Zusätzlich zeichnete er die *Regressionsgerade* ein, auf der die zeilenweisen Durchschnittsgrößen der Kinder liegen. Die Steigung der Regressionsgerade bestimmte er zunächst visuell.

Mit seinen regressionsanalytischen Untersuchungen zur Vererbung gilt Galton als Pionier der Regressionsanalyse. Galtons mathematische Fähigkeiten waren aber begrenzt, so dass die mathematische Ausformulierung und Weiterentwicklung seinen Nachfolgern vorbehalten war, insbesondere dem Dreigespann Karl Pearson (1857–1936), Francis Ysidro Edgeworth (1845–1926) und George Udny Yule (1871–1951).

Heute sind lineare Regressionsmodelle Gegenstand jedes Einführungsbuchs zur Statistik. In moderner Notation untersuchte Galton den systematischen Einfluss der *erklärenden Variable* $x =$ „Durchschnittsgröße der Eltern“ auf die primär interessierende *Zielvariable* $y =$ „Größe des erwachsenen Kindes“. Erklärende Variablen werden auch als *Regressoren* oder *Kovariablen* bezeichnet. Synonyme für Zielvariable sind die Bezeichnungen *abhängige Variable* bzw. *zu erklärende Variable*. Charakteristisch für Regressionsfragestellungen ist die Beobachtung, dass der postulierte Zusammenhang nicht exakt gilt, son-

Durchschnitts-Größe der Eltern	**Größe der Kinder**														
	61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	73.7	**Gesamt**
64.0	1	-	2	4	1	2	2	1	1	0	0	0	0	0	14
64.5	1	1	4	4	1	5	5	0	2	0	0	0	0	0	23
65.5	1	0	9	5	7	11	11	7	7	5	2	1	0	0	66
66.5	0	3	3	5	2	17	17	14	13	4	0	0	0	0	78
67.5	0	3	5	14	15	36	38	28	38	19	11	4	0	0	211
68.5	1	0	7	11	16	25	31	34	48	21	18	4	3	0	219
69.5	0	0	1	16	4	17	27	20	33	25	20	11	4	5	183
70.5	1	0	1	0	1	1	3	12	18	14	7	4	3	3	68
71.5	0	0	0	0	1	3	4	3	5	10	4	9	2	2	43
72.5	0	0	0	0	0	0	0	1	2	1	2	7	2	4	19
73.0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	4
Gesamt	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928

Tabelle 1.1. *Kreuztabelle zwischen der Körpergröße von 928 erwachsenen Kindern und der Durchschnittsgröße ihrer 205 Elternpaare. Alle Angaben sind in der von Galton verwendeten Maßeinheit Zoll (1 Zoll entspricht 2.54 cm). Quelle: Galton (1889)*

Abb. 1.1. *Streudiagramm zwischen Größe der Kinder und Durchschnittsgröße der Eltern inklusive eingezeichneter Regressionsgerade.*

dern durch zufällige Einflüsse überlagert ist. Galton unterstellte das einfachst mögliche Regressionsmodell

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

in dem der systematische Einfluss $\beta_0 + \beta_1 x$ *linear* ist und die zufälligen Abweichungen in der sogenannten *Störgröße* ε zusammengefasst sind. Während Galton die Parameter β_0 und β_1 der Regressionsgerade noch mehr oder weniger ad hoc bestimmte, werden diese *Regressionsparameter* heute durch die *Methode der kleinsten Quadrate* geschätzt. Die Parameter β_0 und β_1 werden basierend auf Beobachtungen (y_i, x_i), $i = 1, \ldots, n$, so

geschätzt, dass die Summe der quadrierten Abweichungen

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

der Beobachtungen y_i von der Regressionsgeraden $\beta_0 + \beta_1 x_i$ minimal wird. Wendet man dieses Prinzip auf Galtons Datensatz an, so erhält man als Steigung der Regressionsgeraden den Wert 0.64. Galtons visuell bestimmte Steigung von 2/3 ist also nicht weit davon entfernt.

Die Methode der kleinsten Quadrate wurde bereits weit vor Galtons Untersuchungen zur Vererbung erfunden. Die erste Veröffentlichung der Methode im Jahr 1805 geht auf den Mathematiker Adrien Marie Legendre (1752–1833) zurück. Damit ist die Methode der kleinsten Quadrate eines der ältesten allgemeinen statistischen Schätzkonzepte. Die ersten Anwendungen dienten im 18. und 19. Jahrhundert hauptsächlich der Vorausberechnung von Asteroidenbahnen. Berühmt wurde die Berechnung der Bahn des Asteroiden Ceres durch Carl Friedrich Gauß (1777–1855). Der Asteroid Ceres wurde im Jahr 1801 durch den Astronom Giuseppe Piazzi entdeckt. Nach 40 Tagen Beobachtung verschwand der Asteroid hinter der Sonne und konnte zunächst nicht wieder gefunden werden, da die exakte Berechnung der Asteroidenbahn zur damaligen Zeit sehr kompliziert war. Eine brauchbare Vorausberechnung der Asteroidenbahn gelang schließlich dem 24-jährigen Gauß unter Verwendung der Methode der kleinsten Quadrate. In seinem 1809 erschienenen Werk „Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium“ reklamierte Gauß die Erfindung der Methode der kleinsten Quadrate für sich. Gauß behauptete später sogar, die Methode bereits seit 1795 (als 18-Jähriger) verwendet zu haben. Es kam daher zwischen Gauß und Legendre zum Streit, wer der Erfinder der Methode der kleinsten Quadrate sei. Fest steht, dass Gauß in seiner Arbeit die bis heute gültigen Grundlagen des linearen Regressionsmodells mit normalverteilten Fehlern legte.

Seit der Erfindung der Methode der kleinsten Quadrate durch Legendre und Gauß und der ersten Anwendung der Regressionsanalyse durch Francis Galton wurden die Methoden der Regressionsanalyse vielfältig bis in die heutige Zeit verfeinert und weiter entwickelt. Heutzutage finden Methoden der Regressionsanalyse breite Anwendung in nahezu allen Wissenschaftsdisziplinen. Ziel dieses Buches ist eine moderne Darstellung der wichtigsten Techniken und Modelle der Regressionsanalyse und deren kompetente Anwendung. Im einzelnen befassen wir uns mit folgenden Themen:

- *Regressionsmodelle:* Kapitel 2 stellt die im Weiteren Verlauf des Buches detaillierter beschriebenen unterschiedlichen Modellklassen ohne technische Details anhand ausgewählter Anwendungen vor.
- *Lineare Modelle:* Dieses Buch bietet in Kapitel 3 eine vollständige Einführung in das lineare Regressionsmodell inklusive neuester Entwicklungen.
- *Generalisierte Lineare Modelle:* In den Kapiteln 4 und 5 geben wir eine Abhandlung Generalisierter Linearer Modelle. Diese eignen sich insbesondere für Fragestellungen mit nicht normalverteilten Zielgrößen, darunter auch kategoriale Zielgrößen.
- *Gemischte Modelle:* In Kapitel 6 behandeln wir sogenannte gemischte Modelle (bzw. Modelle mit zufälligen Effekten) für Clusterdaten. Eine Hauptanwendung ist die Analyse von Panel- und Longitudinaldaten.

- *Univariate, bivariate und räumliche Glättung:* Kapitel 7 gibt eine Einführung in die uni- und bivariate Glättung (nichtparametrische Regression). Diese *semi- und nichtparametrischen Verfahren* sind geeignet, komplexe nichtlineare Regressionsbeziehungen automatisiert zu schätzen. Als Besonderheit werden auch Verfahren der räumlichen Statistik ausführlich beschrieben.
- *Strukturiert additive Regression:* In Kapitel 8 kombinieren wir die bis dahin beschriebenen Verfahren zu einer umfassenden Modellklasse. Als Spezialfall sind bekannte Modelle der nicht- und semiparametrischen Regression enthalten, insbesondere auch Additive Modelle, Geoadditive Modelle und Modelle mit variierenden Koeffizienten. Abschnitt 8.6 zeigt anhand einer detaillierten Fallstudie, wie diese Modelle in der Praxis eingesetzt werden können. Die Fallstudie vermittelt auch allgemeine Hinweise, wie bei Regressionsfragestellungen vorgegangen werden kann.

Damit gibt dieses Buch zum ersten Mal eine umfassende und anwendungsorientierte Abhandlung der wichtigsten Modelle und Verfahren der Regressionsanalyse. Eine Neuerung stellt auch Kapitel 2 dar. Dort werden sämtliche Modellklassen in einem einheitlichen Rahmen unter Auslassung der (oft komplizierten) Schätztechniken vorgestellt. Damit gibt dieses Kapitel dem Anwender einen Überblick über die modernen Verfahren der Regression und dient gleichzeitig als Leitfaden bei der Auswahl der für die jeweilige Fragestellung passenden Modellklasse.

Im folgenden Abschnitt zeigen wir anhand von Anwendungsbeispielen die Vielseitigkeit moderner Regressionsverfahren bei der Behandlung unterschiedlichster Fragestellungen.

1.1 Anwendungsbeispiele

In diesem Buch illustrieren wir die Modelle und Techniken der Regressionsanalyse durch Anwendungsbeispiele aus den unterschiedlichsten Disziplinen. Einen Überblick gibt die nachfolgende Aufstellung:

- *Entwicklungsökonomie:* Analyse sozio-ökonomischer Determinanten der Unterernährung neugeborener Kinder in Entwicklungsländern.
- *Hedonische Preise:* Analyse der Verkaufspreise von Golf-Modellen.
- *Innovationsforschung:* Untersuchungen zur Einspruchswahrscheinlichkeit bei der Erteilung von Patenten durch das europäische Patentamt.
- *Kredit-Scoring:* Analyse der Kreditwürdigkeit von privaten Bankkunden.
- *Marktforschung:* Zusammenhang zwischen dem Absatz eines Produktes und bestimmten Verkaufsförderungsmaßnahmen.
- *Mietspiegel:* Abhängigkeit der Miethöhe von Art, Lage und Beschaffenheit der Mietwohnung.
- *Prämienkalkulation:* Analyse der Schadenshäufigkeit und Schadenshöhe bei Kfz-Versicherungen zur Kalkulation der Versicherungsprämie.
- *Ökologie:* Analyse des Waldzustands.
- *Neurowissenschaften:* Bestimmung der Gehirnareale, die bei bestimmten kognitiven Aufgaben aktiv sind.

- *Medizinische und klinische Studien:*
  - Wirkung von Testosteron auf das Wachstum von Ratten.
  - Analyse der Wahrscheinlichkeit einer Infektion nach einer Kaiserschnittgeburt.
  - Studie zur Beeinträchtigung der Lungenfunktion.
  - Analyse der Lebensdauer von Leukämie-Patienten.
- *Psychologie:* Wortschatztests im Rahmen von Intelligenztests.

Einige der genannten Anwendungsbeispiele werden in diesem Buch eine zentrale Rolle spielen und sollen nachfolgend detaillierter beschrieben werden.

Beispiel 1.1 Mietspiegel

In vielen Städten und Gemeinden werden Mietspiegel erstellt. Sie sollen Mietern und Vermietern eine Marktübersicht zu „ortsüblichen Vergleichsmieten“ bieten. Grundlage dafür ist in Deutschland ein Gesetz, das die ortsübliche Vergleichsmiete definiert als „die üblichen Entgelte, die in der Gemeinde (...) für nicht preisgebundenen Wohnraum vergleichbarer Art, Größe, Beschaffenheit und Lage in den letzten vier Jahren vereinbart oder (...) geändert worden sind“. Sinngemäß bedeutet dies, dass die durchschnittliche Miete in Abhängigkeit von erklärenden Merkmalen wie Art, Größe, Beschaffenheit usw. der Wohnung zu schätzen ist. Somit liegt ein Regressionsproblem vor. Als Zielvariable verwenden wir die sogenannte Nettomiete, d.h. den monatlichen Mietpreis, der nach Abzug aller Betriebs- und Nebenkosten übrig bleibt. Alternativ kann auch die Nettomiete pro Quadratmeter (qm) als Zielvariable verwendet werden.

Im Rahmen dieses Buches beschränken wir uns aus Datenschutzgründen auf einen Teil der Daten und Variablen, die 1999 im Mietspiegel für München eingesetzt wurden. Wir verwenden Daten von 1999, da aktuellere Daten entweder nicht öffentlich zugänglich oder zur Illustration weniger gut geeignet sind. Den aktuellen Mietspiegel für München findet man inklusive Dokumentation unter: `http://www.mietspiegel.muenchen.de`

Tabelle 1.2 enthält für ausgewählte Variablen Kurzbezeichnungen, die später in den Analysen verwendet werden, sowie eine knappe Beschreibung. Die zugehörigen Daten von über 3000 Wohnungen wurden in einer repräsentativen Zufallsstichprobe erhoben.

Ziel einer Regression zur Analyse von Mietspiegeldaten ist eine möglichst realitätsnahe Erfassung des Einflusses der erklärenden Variablen (Wohnfläche, Baujahr, Wohnlage usw.) auf die Zielvariable Nettomiete (*miete*) bzw. Nettomiete pro qm (*mieteqm*). Letztendlich soll der Effekt der erklärenden Variablen in vereinfachter Form durch geeignete Tabellen in einer Mietspiegelbroschüre bzw. im Internet dargestellt werden.

In diesem Buch verwenden wir die Mietspiegeldaten vorwiegend zur Illustration von Regressionsmodellen mit metrischer Zielgröße, vergleiche die Kapitel 3 und 8. Dabei werden zum Großteil vereinfachte Modelle verwendet, so dass die Ergebnisse nicht immer mit dem offiziellen Mietspiegel übereinstimmen.

△

Beispiel 1.2 Unterernährung in Sambia

In Abstimmung mit der Weltgesundheitsorganisation (WHO) werden in Entwicklungsländern regelmäßig repräsentative Haushaltsbefragungen (Demographic and Health Surveys) durchgeführt. Sie enthalten unter anderem Informationen zu Unterernährung, Sterblichkeit und Krankheitsrisiken für Kinder. Die Daten werden vom amerikanischen Institut Macro International für über 50 Länder erhoben und sind im Internet unter `http://www.measuredhs.com/` kostenlos erhältlich. In diesem Buch betrachten

Variable	Beschreibung	Mittelwert/ Häufigkeit in %	Std.- abw.	Min/Max
miete	Nettomiete pro Monat (in DM)	895.90	381.53	79/3594.6
mieteqm	Nettomiete pro Monat und qm (in DM)	13.87	4.75	0.81/34.56
flaeche	Wohnfläche in qm	67.37	23.72	20/160
bjahr	Baujahr (in Jahren)	1956.31	22.31	1918/1997
lage	Lagekategorie gemäß Einschätzung durch Gutachter			
	1 = normale Lage	58.21		
	2 = gute Lage	39.26		
	3 = beste Lage	2.53		
bad	Ausstattung des Bades			
	0 = normal	93.80		
	1 = gehoben	6.20		
kueche	Ausstattung der Küche			
	0 = normal	95.75		
	1 = gehoben	4.25		
zh	Zentralheizung			
	0 = ohne Zentralheizung	10.42		
	1 = mit Zentralheizung	89.58		
bez	Bezirksviertel in München			

Tabelle 1.2. *Beschreibung der Variablen im Mietspiegel für München 1999. Zusätzlich sind für jede Variable einige statistische Kennzahlen aufgeführt.*

wir exemplarisch einen Querschnittsdatensatz für Sambia aus dem Jahr 1992 (insgesamt 4421 Beobachtungen). Die Republik Sambia liegt im südlichen Afrika und gehört zu den ärmsten und am wenigsten entwickelten Staaten der Erde.

Eines der drängendsten Probleme von Entwicklungsländern ist der schlechte, oft katastrophale Ernährungszustand weiter Teile der Bevölkerung. Unmittelbare Folgen der Unterernährung sind unter anderem eine hohe Sterblichkeit sowie eine verringerte Arbeitsproduktivität. Im Rahmen dieses Buches befassen wir uns speziell mit der Ernährungssituation von neugeborenen Kindern im Alter zwischen 0 und 5 Jahren. Der Ernährungszustand von Kindern wird üblicherweise durch eine anthropometrische Maßzahl, Z-Score genannt, gemessen. Der Z-Score vergleicht den anthropometrischen Status eines Kindes, z.B. die altersstandardisierte Körpergröße, mit Vergleichsgrößen aus einer Referenzpopulation. Bis zum Alter von 24 Monaten basiert die Referenzpopulation auf weißen US-amerikanischen Kindern aus wohlhabenden Familien mit hohem sozio-ökonomischem Status. Nach 24 Monaten wechselt die Referenzpopulation und besteht nunmehr aus einer repräsentativen Stichprobe aller US-amerikanischer Kinder. Unter mehreren denkbaren anthropometrischen Indikatoren verwenden wir hier eine Maßzahl für *chronische Unterernährung* („Stunting“), die auf der Körpergröße als Maß für die langfristige Entwicklung des Ernährungszustands basiert. Diese ist für ein Kind i definiert durch

$$zscore_i = \frac{g_i - mg}{\sigma},$$

wobei g_i die Körpergröße des Kindes ist, mg der Median der Größe von Kindern der Referenzpopulation im selben Alter und σ die entsprechende Standardabweichung für die Referenzpopulation.

Variable	Beschreibung	Mittelwert/ Häufigkeit in %	Std-abw.	Min/Max
zscore	Z-Score des Kindes	−171.19	139.34	-600/503
k_geschl	Geschlecht des Kindes			
	1 = männlich	49.02		
	0 = weiblich	50.98		
k_still	Stilldauer in Monaten	11.11	9.42	0/46
k_alter	Alter des Kindes in Monaten	27.61	17.08	0/59
m_alterg	Alter der Mutter bei der Geburt in Jahren	26.40	6.87	13.16/48.66
m_groesse	Größe der Mutter in cm	158.06	5.99	134/185
m_bmi	Body-Mass-Index der Mutter	21.99	3.32	13.15/39.29
m_bildung	Ausbildung der Mutter			
	1 = keine Ausbildung	18.59		
	2 = Grundschule	62.34		
	3 = Volksschule	17.35		
	4 = höherer Abschluss	1.72		
m_arbeit	Erwerbsstatus der Mutter			
	1 = Mutter arbeitet	55.25		
	0 = Mutter arbeitet nicht	44.75		
region	Wohnort (Region) in Sambia			
	1 = Central	8.89		
	2 = Copperbelt	21.87		
	3 = Eastern	9.27		
	4 = Luapula	8.91		
	5 = Lusaka	13.78		
	6 = Northern	9.73		
	7 = North-Western	5.88		
	8 = Southern	14.91		
	9 = Western	6.76		
district	Wohnort in Sambia, insgesamt 55 Distrikte			

Tabelle 1.3. *Variablenbeschreibung der Sambia Daten.*

Primäres Ziel der statistischen Analyse ist die Ermittlung des Einflusses bestimmter sozio-ökonomischer Variablen des Kindes, der Mutter und des Haushalts auf den Ernährungszustand des Kindes. Beispiele für sozio-ökonomische Variablen sind die Stilldauer (Variable *k_still*), das Alter des Kindes (*k_alter*), der Ernährungszustand der Mutter gemessen anhand des Body-Mass-Index (*m_bmi*) und das Bildungsniveau sowie der Erwerbsstatus der Mutter (*m_bildung* und *m_arbeit*). Zusätzlich enthält der Datensatz als geografische Information die Region bzw. den Distrikt, in dem der Wohnort der Mutter liegt. Eine Beschreibung aller zur Verfügung stehenden Variablen findet man in Tabelle 1.3.

Die genannten Ziele lassen sich mit den Regressionsmodellen dieses Buches verfolgen, wobei hier speziell sogenannte geoadditive Modelle (vergleiche Kapitel 8, insbesondere Abschnitt 8.2) zum Einsatz kommen. Diese erlauben zusätzlich die adäquate

Variable	Beschreibung	Mittelwert/ Häufigkeit in %	Std-abw.	Min/Max
einspruch	Einspruch gegen das Patent			
	1 = Ja	41.49		
	0 = Nein	58.51		
biopharm	Patent aus der Biotechnologie- / Pharma-Branche			
	1 = Ja	44.31		
	0 = Nein	55.69		
uszw	US Zwillingspatent			
	1 = Ja	60.85		
	0 = Nein	39.15		
patus	Patentinhaber aus den USA			
	1 = Ja	33.74		
	0 = Nein	66.26		
patdsg	Patentinhaber aus Deutschland, der Schweiz oder Großbritannien			
	1 = Ja	23.49		
	0 = Nein	76.51		
jahr	Jahr der Patenterteilung			
	1980	0.18		
	⋮	⋮		
	1997	1.62		
azit	Anzahl der Zitationen für dieses Patent	1.64	2.74	0/40
aland	Anzahl der Länder, für die Patentschutz gelten soll	7.8	4.12	1/17
ansp	Anzahl der Patentansprüche	13.13	12.09	1/355

Tabelle 1.4. *Beschreibung des Datensatzes zum Auftreten von Einsprüchen gegen Patente.*

Berücksichtigung *räumlicher Information* in den Daten. Die Analyse der Daten erfolgt im Rahmen einer umfassenden Fallstudie (vergleiche Kapitel 8.6), in der die praktische Anwendung der in diesem Buch vorgestellten Techniken und Verfahren ausführlich demonstriert wird.

△

Beispiel 1.3 Einsprüche gegen Patente

In Europa können Erfindungen durch das Europäische Patentamt für einen gewissen Zeitraum geschützt werden, so dass Wettbewerber zunächst von deren Verwertung ausgeschlossen werden. Aufgabe des Patentamts ist es, Erfindungen zu prüfen und ein Patent nur dann zu erteilen, wenn gewisse Voraussetzungen erfüllt sind. Insbesondere muss es sich bei der Erfindung um eine echte Neuerung handeln. Trotz sorgfältiger Prüfung kommt es in etwa 8–10 Prozent der Fälle zu Einsprüchen von Wettbewerbern gegen bereits erteilte Patente. In der neueren ökonomischen Literatur spielen die *Gründe* für Einsprüche gegen Patente eine wichtige Rolle, da damit indirekt eine Reihe ökonomischer Fragen untersucht werden können. Beispielsweise kann die Häufigkeit, mit der gegen Patente Einspruch erhoben wird, als Indikator für die Wettbewerbsintensität in verschiedenen Branchen verwendet werden.

Im Rahmen einer Analyse des Auftretens von Einsprüchen gegen Patente wurden die in Tabelle 1.4 angegebenen Merkmale für 4866 vom Europäischen Patentamt erteilte Patente aus den Branchen Biotechnologie/Pharma und Halbleiter/Computer erhoben. Ziel der Untersuchung ist es, für die binäre Zielvariable „Einspruch“ (ja/nein) die Wahrscheinlichkeit für einen Patenteinspruch in Abhängigkeit von Kovariablen zu modellieren. Somit liegt ein Regressionsproblem mit einer binären Zielvariablen vor.

Eine mögliche erklärende Variable ist die Variable *azit*, die angibt wie oft ein Patent in anderen, nachfolgenden Patenten zitiert wird. Zitationen von Patenten können mit Zitationen von wissenschaftlichen Arbeiten verglichen werden. Empirische Erfahrungen und ökonomische Argumente weisen darauf hin, dass die Wahrscheinlichkeit für einen Einspruch bei oft zitierten Patenten ansteigt. Diese und andere Hypothesen lassen sich mit Regressionsmodellen für binäre Zielvariablen formulieren und überprüfen.

Im Buch dient der Datensatz zur Illustration von Regressionsmodellen mit binärer Zielgröße, vergleiche die Kapitel 2 und 4.

△

Beispiel 1.4 Zustand des Waldes

Kenntnisse über den Zustand des Waldes und beeinflussende Faktoren sind aus ökologischer und ökonomischer Sicht wichtig. In Deutschland werden deshalb jährlich Waldzustandserhebungen im gesamten Bundesgebiet durchgeführt. Im Folgenden beschreiben wir ein spezielles Projekt im Forstgebiet Rothenbuch (Spessart), das von Axel Göttlein (TU München) seit 1982 durchgeführt wird. Im Vergleich zu den großflächigen offiziellen Erhebungen liegen die Beobachtungspunkte, d.h. die Standorte der untersuchten Bäume, wesentlich dichter zusammen. Abbildung 1.2 zeigt die Lage der 83 untersuchten Standorte im Forstgebiet Rothenbuch. Im Zentrum liegt der Ort Rothenbuch. Untersucht werden fünf Baumarten: Buche, Eiche, Fichte, Lärche und Kiefer. Im Weiteren beschränken wir uns auf die Buche. An jedem Beobachtungspunkt wird jedes Jahr der Zustand der Buchen durch die Zielvariable „Entlaubungsgrad“ in die neun ordinalen Kategorien 0%, 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, 87.5% und 100% Entlaubung eingestuft. Die Kategorie 0% bedeutet, dass die Buche gesund ist, während die Kategorie 100% bedeutet, dass sie abgestorben ist.

Neben der (ordinalen) Zielvariablen werden ebenfalls jährlich beeinflussende Faktoren erhoben. Tabelle 1.5 enthält eine Auswahl solcher Faktoren inklusive einiger deskriptiver Kennzahlen. Die Mittelwerte bzw. Häufigkeiten (in Prozent) sind über die Jahre 1983 – 2004 und die Beobachtungspunkte gemittelt.

Ziel von Analysen zum Waldzustand ist es, den Effekt beeinflussender Kovariablen auf den in geordneten Kategorien gemessenen Entlaubungsgrad zu schätzen. Zusätzlich sollen der zeitliche Trend für den Beobachtungszeitraum sowie räumliche Effekte der geografischen Lage der Standorte aus Abbildung 1.2 bei gleichzeitiger Adjustierung auf andere Kovariablen quantifiziert werden. Abbildung 1.2 zeigt außerdem den zeitlichen Trend der relativen Häufigkeiten für den in 3 Kategorien zusammengefassten Entlaubungsgrad. Für diese Problemstellung werden Regressionsmodelle für (mehr)kategoriale Zielvariablen benötigt, mit denen auch nichtlineare Einflüsse der metrischen Kovariablen sowie zeitliche und räumliche Trends in einem simultanen Ansatz modelliert und analysiert werden können. Wir verwenden die Daten zur Illustration kategorialer Regressionsmodelle in den Kapiteln 5 und 8.

△

Variable	Beschreibung	Mittelwert/ Häufigkeit in %	Std-abw.	Min/Max
id	Standort Identifikationsnummer			
jahr	Jahr der Erhebung	1993.59	6.34	1983/2004
buche	Entlaubung der Buchen, in 9 ordinalen Kategorien			
	0%	62.14		
	12.5%	24.22		
	25%	7.02		
	37.5%	3.79		
	50%	1.61		
	62.5%	0.89		
	75%	0.33		
	87.5%	0.00		
	100%	0.00		
x	x-Koordinate des Standorts			
y	y-Koordinate des Standorts			
alter	Bestandsalter, in Jahren	106.07	51.41	7/234
schirm	Beschirmungsgrad, d.h. Dichte der Laubdecke, in Prozent (0% – 100%)	77.29	23.70	0/100
hang	Hangneigung, in Prozent	15.45	11.27	0/46
hoehe	Höhe über dem Meeresspiegel, in Metern	386.99	58.88	250/480
grund	Gründigkeit, d.h. Bodentiefe, in der das Gestein beginnt, in cm	24.64	9.94	9/51
ph	pH-Wert in 0–2cm Tiefe	4.29	0.34	3.28/6.05
frische	Feuchtigkeitsstufe des Boden, in 3 Kategorien			
	1 = mäßig trocken	11.03		
	2 = mäßig frisch	55.12		
	3 = frisch oder mäßig wechselfeucht	33.85		
alkali	Anteil Alkali-/Erdalkali-Ionen im Boden, in 4 Kategorien			
	1 = sehr gering	19.60		
	2 = gering	55.18		
	3 = mäßig	17.15		
	4 = hoch	8.07		
humus	Dichte der Humusschicht in cm	1.57	1.38	0/9
artkat	Art des Waldes			
	0 = Mischwald	49.78		
	1 = Laubwald	50.22		
dueng	Düngung			
	0 = nicht gedüngt	80.90		
	1 = gedüngt	19.10		

Tabelle 1.5. *Beschreibung der Variablen des Datensatzes zum Waldzustand.*

Abb. 1.2. *Links: Beobachtungsstandorte. Im Zentrum befindet sich der Ort Rothenbuch. Rechts: Zeitlicher Trend der Schädigung.*

Im nächsten Abschnitt zeigen wir anhand der beiden Beispiele zum Mietspiegel und zur Unterernährung in Sambia, wie die ersten explorativen Schritte bei Regressionsanalysen aussehen.

1.2 Erste Schritte

1.2.1 Beschreibung der Verteilung der Variablen

Der erste Schritt bei der Durchführung einer Regressionsanalyse (und prinzipiell jeder statistischen Auswertung) besteht darin, sich einen *Überblick* über die Variablen des Datensatzes zu verschaffen. Im Zuge dieser ersten deskriptiven und grafischen univariaten Analyse werden folgende Ziele verfolgt:

- Beschreibung der Verteilung der Variablen,
- Auffinden von extremen Werten,
- Auffinden von Fehlkodierungen.

Zur Erreichung dieser Ziele können geeignete deskriptive Hilfsmittel (vor allem Lagemaße, Streuungsmaße sowie Minimum und Maximum) und grafische Darstellungsmöglichkeiten (Histogramme, Boxplots, etc.) herangezogen werden. Welche Hilfsmittel und Darstellungsmöglichkeiten geeignet sind, hängt vor allem vom jeweiligen Variablentyp ab. Wir können im Wesentlichen unterscheiden zwischen stetigen und kategorialen Variablen.

Einen ersten Überblick über stetige Variablen gewinnt man durch Bestimmung einiger deskriptiver Kennzahlen. Geeignet sind als Lagemaße insbesondere das arithmetische Mittel und der Median und als Streuungsmaß die Standardabweichung. Außerdem sind Minimum und Maximum der Daten von Interesse. Darüber hinaus sollte die Verteilung grafisch dargestellt werden. Geeignet sind Histogramme, Boxplots und Kerndichteschätzer. Kerndichteschätzer können als nichtparametrische Schätzungen für die Dichte einer stetigen Variable angesehen werden und stellen eine glatte Alternative zu Histogrammen dar. Eine leicht verständliche Darstellung findet man z.B. in dem Einführungsbuch von Fahrmeir et al. (2007).

Abb. 1.3. *Mietspiegel: Histogramme und Kerndichteschätzer für die metrischen Variablen miete, mieteqm, flaeche und bjahr.*

Einfacher lässt sich ein Überblick über die Verteilung kategorialer Variablen gewinnen. Hier genügen einfache Häufigkeitstabellen oder deren grafische Darstellung in Form von Säulen- oder Balkendiagrammen.

Beispiel 1.5 Mietspiegel – Univariate Verteilungen

Die wichtigsten deskriptiven Kennzahlen der stetigen Variablen *miete*, *mieteqm*, *flaeche* und *bjahr* findet man bereits in Tabelle 1.2 (Seite 6). Histogramme und Kerndichteschätzer für diese Variablen sind in Abbildung 1.3 abgedruckt. Exemplarisch interpretieren wir die Kennzahlen und grafischen Darstellungen für die beiden Variablen *miete* und *bjahr*:

Die Nettomiete schwankt im Datensatz zwischen 79 und annähernd 3600 DM. Die Durchschnittsmiete beträgt circa 900 DM. Für die überwiegende Mehrzahl der Wohnungen im Datensatz liegt die Miete zwischen 100 und 2400 DM, nur sehr wenige Wohnungen weisen eine Miete von mehr als 2400 DM auf. Für die späteren Regressionsanalysen könnte diese Beobachtung bedeuten, dass über diese sehr teueren Mietwohnungen nur sehr ungenaue Aussagen getroffen werden können, da die vorhandene Datenbasis zu dünn ist. Insgesamt handelt es sich um eine deutlich unsymmetrische, linkssteile Verteilung.

Die Verteilung des Baujahrs ist (historisch bedingt) sehr ungleichmäßig und daher multimodal. Für die Jahre der Wirtschaftskrise in der Weimarer Republik und des 2. Weltkriegs liegen nur sehr wenige Wohnungen vor, während für die späteren Aufbaujahre

relativ viele Wohnungen vorliegen (Modus circa im Jahr 1960). Ab Mitte der 1970er Jahre flacht die Bautätigkeit dann wieder ab. Insgesamt liegen Informationen für die Jahre 1918 bis 1997 vor. Offensichtlich lässt der Mietspiegel für 1999 keine Schlüsse auf Neubauten nach 1997 zu. Der Grund hierfür liegt in der verhältnismäßig großen zeitlichen Differenz von mehr als einem Jahr zwischen Datenerhebung und Veröffentlichung des Mietspiegels. Auffallend ist auch die relative Häufung von Wohnungen mit Baujahr 1918. Hier sind die Daten ungenau, da alle vor 1918 gebauten Wohnungen auf das Jahr 1918 datiert wurden.

Die Interpretation der Verteilungen der beiden anderen metrischen Variablen im Datensatz überlassen wir dem Leser.

Häufigkeitstabellen für die kategorialen Variablen findet man wieder in Tabelle 1.2. Hier stellen wir beispielsweise fest, dass sich die meisten Wohnungen in normaler Wohnlage befinden (58%) und nur circa 3% in bester Wohnlage.

△

Beispiel 1.6 Unterernährung in Sambia – Univariate Verteilungen

Einen Überblick über die Verteilung ausgewählter Variablen im Datensatz gibt neben Tabelle 1.3 (Seite 7) die Abbildung 1.4, die Histogramme und Kerndichteschätzer der Zielgröße und der metrischen erklärenden Variablen enthält. Eine ausführliche Interpretation im Hinblick auf die Regressionsfragestellung geben wir im Rahmen der Fallstudie in Kapitel 8.6.

△

1.2.2 Grafische Zusammenhangsanalyse

In einem zweiten Schritt kann, zumindest bei stetigen Zielgrößen, grafisch der Zusammenhang zwischen der Zielgröße und den erklärenden Variablen untersucht werden. Damit wird ein erster Überblick über die Art (z.B. linearer versus nichtlinearer Zusammenhang) und die Stärke des Zusammenhangs gewonnen. In den meisten Fällen wird man sich auf zweidimensionale Zusammenhangsanalysen zwischen Zielgröße und jeweils einer der erklärenden Variablen beschränken. Wir gehen im Folgenden stets von einer stetigen Zielgröße aus. Die geeigneten Darstellungsmöglichkeiten hängen vom Typ der erklärenden Variable ab. Wir unterscheiden stetige und kategoriale erklärende Variablen.

Stetige erklärende Variablen

Bei stetigen erklärenden Variablen bieten sich zunächst einfache Streudiagramme an, wie bereits von Galton Ende des 19. Jahrhunderts verwendet.

Beispiel 1.7 Mietspiegel – Streudiagramme

Für die Mietspiegeldaten findet man Streudiagramme zwischen Nettomiete bzw. Nettomiete pro qm und den erklärenden metrischen Variablen Wohnfläche und Baujahr in Abbildung 1.5. Bei großem Stichprobenumfang, wie hier mit über 3000 Beobachtungen, sind die Streudiagramme oft wenig informativ. Relativ gut erkennbar ist ein annähernd linearer Zusammenhang zwischen Nettomiete und Wohnfläche. Wir erkennen auch, dass die Streubreite der Mieten mit steigender Wohnfläche größer wird. Über den Zusammenhang zwischen Nettomiete pro qm und Wohnfläche lassen sich weniger genaue Aussagen treffen. Insgesamt scheinen die Mieten pro qm für größere Wohnungen

Abb. 1.4. *Unterernährung in Sambia: Verteilungen der metrischen Variablen.*

kleiner zu werden, über die Art des Zusammenhangs (linear oder nichtlinear) können wir jedoch keine Aussagen treffen. Auch der Zusammenhang der beiden Zielvariablen mit dem Baujahr ist (falls überhaupt vorhanden) kaum zu erkennen.

△

Das vorangegangene Beispiel zeigt, dass bei großem Stichprobenumfang der Informationsgehalt in einfachen Streudiagrammen häufig relativ gering ist. In diesem Fall kann es daher sinnvoll sein, die Daten zu *gruppieren*. Falls die Anzahl der *verschiedenen* Werte der erklärenden Variable im Vergleich zum Stichprobenumfang relativ klein ist, kann für jeden beobachteten Wert der Mittelwert der Zielgröße und die dazugehörige Standardab-

Abb. 1.5. *Mietspiegel: Streudiagramme zwischen Nettomiete bzw. Nettomiete pro qm und den erklärenden Variablen Wohnfläche und Baujahr.*

weichung bestimmt und in einem Streudiagramm visualisiert werden. Durch die auf diese Weise erzielte Datenreduktion lassen sich Zusammenhänge meistens besser erkennen. Ist die Anzahl der verschiedenen Werte im Vergleich zum Stichprobenumfang relativ groß, so kann es vorkommen, dass einige Gruppen sehr dünn besetzt sind. In diesem Fall kann der Wertebereich der erklärenden Variable in einem Zwischenschritt in kleine Intervalle unterteilt und anschließend Mittelwert und Standardabweichung der Zielgröße für jedes Intervall berechnet werden. Zuletzt werden Mittelwerte plus minus Standardabweichungen gegen die Gruppenmittelwerte in einem Streudiagramm abgetragen.

Beispiel 1.8 Mietspiegel – Streudiagramme nach Gruppierung

Im Falle der Wohnfläche und des Baujahrs liegen die Daten auf einen Quadratmeter genau bzw. jahresgenau vor. Wir können also ohne weiteres Mittelwerte und Standardabweichungen pro Quadratmeter Wohnfläche bzw. für jedes Jahr bestimmen und visualisieren, vergleiche Abbildung 1.6. Aussagen über mögliche Zusammenhänge lassen sich jetzt besser treffen. Wenn wir die Nettomiete pro qm als Zielgröße zugrunde legen, so erkennen wir einen deutlich nichtlinearen, monoton fallenden Zusammenhang mit der Wohnfläche. Für große Wohnungen ab 120 Quadratmeter Wohnfläche nimmt die Streuung um die Durchschnittsmiete deutlich zu. Auch zwischen dem Baujahr und der Nettomiete pro qm scheint ein (wenn auch deutlich schwächerer) Zusammenhang zu bestehen. Auch hier liegt eher eine nichtlineare Beziehung vor. Für vor 1940 gebaute Wohnungen schwanken die Mieten pro qm um einen konstanten Wert von etwa

Abb. 1.6. *Mittelwerte plus/minus eine Standardabweichung der Nettomiete bzw. Nettomiete pro qm versus Wohnfläche und Baujahr.*

12 DM. Für die wenigen Wohnungen in der Stichprobe aus den Kriegsjahren scheinen die Mieten im Durchschnitt etwas niedriger zu sein. Nach 1945 steigen dann die Durchschnittsmieten annähernd linear an.

△

Kategoriale erklärende Variablen

Die Visualisierung des Zusammenhangs zwischen metrischer Zielgröße und kategorialen erklärenden Variablen erfolgt durch die kategorienspezifische Darstellung der Verteilung der Zielgröße. Als Darstellungsmöglichkeiten kommen wieder Histogramme, Boxplots und Kerndichteschätzer in Frage. Boxplots sind häufig besonders geeignet, da hier Unterschiede im Mittelwert (genauer dem Median) am deutlichsten zu erkennen sind.

Beispiel 1.9 Mietspiegel – Zusammenhang bei kategorialen Variablen

Abbildung 1.7 zeigt die Verteilung der Nettomiete pro qm in Abhängigkeit von der Wohnlage. Die linke Grafik verwendet Boxplots zur Darstellung, die rechte Grafik Kerndichteschätzer. Anhand der Boxplots ist gut zu erkennen, dass die Durchschnittsmiete (und die Streuung) mit besser werdender Wohnlage zunimmt. Ähnliche Informationen liefern die Kerndichteschätzer, jedoch weniger deutlich sichtbar.

△

Abb. 1.7. *Verteilung der Nettomiete pro qm in Abhängigkeit von der Wohnlage.*

Abb. 1.8. *Unterernährung in Sambia: Verschiedene grafische Darstellungen des Zusammenhangs zwischen Z-Score und Alter des Kindes.*

Beispiel 1.10 Unterernährung in Sambia – Grafische Zusammenhangsanalysen

Grafische Darstellungen des Zusammenhangs zwischen dem Z-Score und ausgewählten erklärenden Variablen findet man in den Abbildungen 1.8 und 1.9. Am Beispiel des Alters des Kindes (Variable *k_alter*) lassen sich nochmal die Schwierigkeiten bei der grafischen Darstellung des Zusammenhangs zwischen Zielgröße und erklärenden Variablen in sehr großen Datensätzen veranschaulichen (Abbildung 1.8). Ähnlich wie bei den Mietspiegeldaten kann aus dem Streudiagramm zwischen Z-Score und dem Alter des Kindes in Abbildung a) nur unzureichend auf die Art des Zusammenhangs geschlossen werden. Als geeigneter erweist sich wieder die Visualisierung des durchschnittlichen Z-Scores inklusive Standardabweichung für jedes Alter zwischen 0 und 59 Monaten (Abbildung b). Diese Art der Darstellung wurde auch für die anderen stetigen Einflussvariablen in Abbildung 1.9 gewählt. Ausführlich gehen wir auf die gezeigten Grafiken in Kapitel 8.6 im Rahmen der Fallstudie zur Unterernährung ein.

△

Abb. 1.9. *Unterernährung in Sambia: Grafische Darstellung des Zusammenhangs zwischen Z-Score und ausgewählten erklärenden Variablen.*

2 Regressionsmodelle

2.1 Einführung

Alle im vorigen Kapitel beschriebenen Problemstellungen besitzen eine wesentliche Gemeinsamkeit: Eigenschaften einer *Zielvariablen* y sollen in Abhängigkeit von *Kovariablen* $x_1, \ldots, x_k$ beschrieben werden. Dabei werden die Zielvariable auch als *abhängige Variable* und die Kovariablen als *erklärende Variablen* oder *Regressoren* bezeichnet. Die behandelten Modelle unterscheiden sich im Wesentlichen durch unterschiedliche Typen von Zielvariablen (stetig, binär, kategorial oder Zählvariablen) und verschiedene Arten von Kovariablen, die ebenfalls stetig, binär oder kategorial sein können. In komplexeren Modellen können auch Zeitskalen, Variablen zur Beschreibung der räumlichen Anordnung der Daten oder Gruppierungsvariablen als Kovariablen auftreten.

Ein wesentliches Merkmal von Regressionsfragestellungen ist, dass der Zusammenhang zwischen Zielgröße y und den erklärenden Variablen nicht (wie beispielsweise häufig in der Physik) exakt als Funktion $f(x_1, \ldots, x_k)$ von $x_1, \ldots, x_k$ gegeben ist, sondern durch zufällige Störungen überlagert wird. Die Zielgröße y ist also eine Zufallsvariable, deren Verteilung von den erklärenden Variablen abhängt. Bei Galtons Daten zur Vererbung etwa kann bei gegebener Körpergröße der Eltern nicht exakt auf die Körpergröße der Kinder geschlossen werden. Wir können bei gegebener Größe der Eltern lediglich Aussagen über die *durchschnittliche Körpergröße* der Kinder und das Ausmaß der Streuung um den Durchschnitt treffen. Ganz ähnlich verhält es sich bei allen anderen in Kapitel 1 angesprochenen Fragestellungen. Ein Hauptziel der Regressionsanalyse besteht somit darin, den Einfluss der erklärenden Variablen auf den Mittelwert der Zielgröße zu untersuchen. Anders ausgedrückt modellieren wir den (bedingten) Erwartungswert $\mathrm{E}(y \mid x_1, \ldots, x_k)$ von y in Abhängigkeit der Kovariablen. Der Erwartungswert ist also eine Funktion der Kovariablen:

$$\mathrm{E}(y \mid x_1, \ldots, x_k) = f(x_1, \ldots, x_k)$$

Die Zielgröße lässt sich dann immer zerlegen in

$$y = \mathrm{E}(y \mid x_1, \ldots, x_k) + \varepsilon = f(x_1, \ldots, x_k) + \varepsilon,$$

wobei ε die zufällige, nicht von den Kovariablen erklärte Abweichung vom Erwartungswert ist. Häufig bezeichnet man $f(x_1, \ldots, x_k)$ auch als *systematische Komponente*. Die zufällige Abweichung ε wird auch als *stochastische Komponente*, *Störgröße* oder *Fehlerterm* bezeichnet. Ein Hauptziel der Regressionsanalyse besteht darin, die systematische Komponente f aus gegebenen Daten $y_i, x_{i1}, \ldots, x_{ik}$, $i = 1, \ldots, n$, zu schätzen und von der stochastischen Komponente ε zu trennen.

Am bekanntesten ist die Klasse der *linearen Regressionsmodelle*

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon,$$

in denen unterstellt wird, dass die Funktion f linear ist, so dass

$$\mathrm{E}(y \,|\, x_1, \ldots, x_k) = f(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

gilt. Wenn wir die Daten einsetzen, erhalten wir die n Gleichungen

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

mit den unbekannten Parametern bzw. Regressionskoeffizienten $\beta_0, \ldots, \beta_k$. Im linearen Modell wirkt also jede der Kovariablen linear auf y und die Effekte der einzelnen Kovariablen setzen sich additiv zusammen. Das lineare Regressionsmodell ist insbesondere dann sinnvoll einsetzbar, wenn die Zielvariable y stetig und wenn möglich approximativ normalverteilt ist. Allgemeinere Regressionsmodelle werden beispielsweise dann benötigt, wenn die Zielvariable binär ist, Effekte von Kovariablen flexibel und nichtlinear einzubeziehen sind oder die räumliche Verteilung der Daten mit analysiert werden soll.

Ausgehend vom klassischen linearen Regressionsmodell beschreiben die weiteren Abschnitte dieses Kapitels flexible Regressionsansätze, die zur Analyse der in Kapitel 1 beschriebenen, komplexen Problemstellungen geeignet sind. Unterstützt durch illustrierende Beispiele aus verschiedenen Anwendungsbereichen soll damit ein erster Überblick über die verschiedenen Modellierungsmöglichkeiten gegeben werden. Eingehendere Darstellungen der verschiedenen Regressionsmodelle und insbesondere der zugehörigen statistischen Inferenzverfahren folgen dann in den weiteren Kapiteln.

2.2 Lineare Regressionsmodelle

2.2.1 Das einfache lineare Regressionsmodell

Beispiel 2.1 Mietspiegel – Lineare Einfachregression

Wir greifen aus dem gesamten Datensatz die Wohnungen heraus, die seit 1966 gebaut wurden. Diese Teilstichprobe zerlegen wir in die Schichten „normale Lage", „gute Lage" und „beste Lage". Abbildung 2.1 (links) zeigt das Streudiagramm für die Wohnungen in normaler Lage mit der Zielgröße *miete* und der erklärenden Variable *flaeche*.

Das Streudiagramm legt einen annähernd linearen Einfluss der Wohnfläche auf die Miete nahe:

$$\mathit{miete}_i = \beta_0 + \beta_1 \cdot \mathit{flaeche}_i + \varepsilon_i. \tag{2.1}$$

Die Fehlervariablen ε_i werden als zufällige Abweichungen von der Geraden $\beta_0 + \beta_1 \mathit{flaeche}$ interpretiert. Da systematische Abweichungen von Null bereits durch den Parameter β_0 berücksichtigt werden, nimmt man $\mathrm{E}(\varepsilon_i) = 0$ an. Eine alternative Formulierung der Beziehung (2.1) ist

$$\mathrm{E}(\mathit{miete} \,|\, \mathit{flaeche}) = \beta_0 + \beta_1 \cdot \mathit{flaeche},$$

d.h der erwartete Mietpreis ist eine lineare Funktion der Wohnfläche.

△

Das Beispiel ist ein Spezialfall des *einfachen linearen Regressionsmodells*

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

bei dem die Funktion $f(x)$ bzw. der Erwartungswert $\mathrm{E}(y \,|\, x)$ in der allgemeineren Beziehung

Abb. 2.1. *Streudiagramm zwischen Nettomiete und Wohnfläche für nach 1966 gebaute Wohnungen in normaler Wohnlage (links). In der rechten Grafik ist zusätzlich die Regressionsgerade mit eingezeichnet.*

$$y = f(x) + \varepsilon = \mathrm{E}(y \,|\, x) + \varepsilon$$

als linear, d.h. $f(x) = \mathrm{E}(y \,|\, x) = \beta_0 + \beta_1 x$ angenommen wird.
Genauer werden für das *Standardmodell der linearen Einfachregression* folgende Annahmen getroffen: Es gilt

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.2}$$

wobei die Fehlervariablen ε_i unabhängig und identisch mit

$$\mathrm{E}(\varepsilon_i) = 0 \quad \text{und} \quad \mathrm{Var}(\varepsilon_i) = \sigma^2$$

verteilt sind. Die Eigenschaft gleich großer Varianzen σ^2 für alle Fehlervariablen wird auch als *Homoskedastizität* bezeichnet. Zur Konstruktion von Konfidenzintervallen und Teststatistiken ist es günstig, wenn darüber hinaus (zumindest approximativ) die *Normalverteilungsannahme*

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

gilt. Dann sind auch die Zielvariablen (bedingt) normalverteilt mit

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i, \quad \mathrm{Var}(y_i) = \sigma^2,$$

und die Zielvariablen sind bei gegebenen Kovariablenwerten x_i (bedingt) unabhängig. Die unbekannten Parameter β_0 und β_1 werden nach der Methode der kleinsten Quadrate (KQ-Methode) geschätzt. Dazu werden die Schätzwerte $\hat{\beta}_0$ und $\hat{\beta}_1$ so bestimmt, dass die Summe der quadratischen Abweichungen

$$\mathrm{KQ}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

bei gegebenen Daten (y_i, x_i), $i = 1, \ldots, n$, minimiert wird. Details zur Methode der kleinsten Quadrate behandeln wir in Kapitel 3.2.1. Setzt man $\hat{\beta}_0$, $\hat{\beta}_1$ in die Modellgerade ein, so erhält man die geschätzte Regressionsgerade $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. Die Regressionsgerade kann als Schätzung $\widehat{\mathrm{E}(y|x)}$ für den bedingten Erwartungswert von y bei gegebenem

Standardmodell der linearen Einfachregression

Daten

(y_i, x_i), $i = 1, \ldots, n$, zu metrischen Variablen y und x.

Modell

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \quad i = 1, \ldots, n\,.$$

Die Fehler $\varepsilon_1, \ldots, \varepsilon_n$ sind unabhängig und identisch verteilt (i.i.d.) mit

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

Die geschätzte Regressionsgerade $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ kann als Schätzung $\widehat{\mathrm{E}(y|x)}$ für den bedingten Erwartungswert von y bei gegebenem Kovariablenwert x angesehen und damit zur Prognose von y verwendet werden. Diese Prognose wird mit $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ bezeichnet.

Kovariablenwert x angesehen und damit zur Prognose von y verwendet werden. Diese Prognose wird mit $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ bezeichnet.

Beispiel 2.2 Mietspiegel – Lineare Einfachregression

Wir illustrieren die lineare Einfachregression mit den in Abbildung 2.1 gezeigten Daten und dem zugehörigen Modell (2.1). Ein Blick auf die Daten lässt dabei Zweifel an der Annahme gleich großer Varianzen $\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(y_i) = \sigma^2$ aufkommen, da die Variabilität mit wachsender Wohnfläche ebenfalls größer zu werden scheint. Vorerst ignorieren wir dieses Problem jedoch. In Kapitel 3.4.3 wird gezeigt, wie man dem Problem ungleicher Varianzen begegnen kann.

Für das Modell (2.1) ergeben sich nach der KQ-Methode die Schätzwerte $\hat{\beta}_0 = 253.95$, $\hat{\beta}_1 = 10.87$. Somit erhält man die geschätzte lineare Funktion

$$\hat{f}(\textit{flaeche}) = 253.95 + 10.87 \cdot \textit{flaeche}$$

in Abbildung 2.1 (rechts). Der Steigungsparameter $\hat{\beta}_1 = 10.87$ lässt sich wie folgt interpretieren: Nimmt die Wohnfläche um 1 qm zu, so erhöht sich die durchschnittliche Miete um 10.87 DM.

Wählt man statt der Miete selbst die Miete pro Quadratmeter als Zielvariable, so erhält man das Streudiagramm in Abbildung 2.2 (links). Offensichtlich ist die Beziehung zwischen *mieteqm* und *flaeche* eher nichtlinear. Dies wird auch deutlich durch die geschätzte Regressionsgerade

$$\hat{f} = 20.47 - 0.079 \cdot \textit{flaeche}.$$

Sie ist an die Daten zumindest für kleine und große Wohnflächen nicht gut angepasst.

Eine bessere Anpassung lässt sich erzielen, wenn man als neue erklärende Variable

$$x = \frac{1}{\textit{flaeche}}$$

definiert und eine Regression der Form

$$\textit{mieteqm}_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \beta_0 + \beta_1 \frac{1}{\textit{flaeche}_i} + \varepsilon_i \tag{2.3}$$

Abb. 2.2. *Streudiagramm zwischen Nettomiete pro qm und Wohnfläche und geschätzte Funktionen $\hat{f}$ bei Verwendung der Wohnfläche (links) und der inversen Wohnfläche (rechts) als erklärende Variable.*

ansetzt. Mit der transformierten Regressorvariablen ist (2.3) wieder in der Form einer linearen Einfachregression, so dass die Parameter β_0 und β_1 der Funktion

$$f(\mathit{flaeche}) = \beta_0 + \beta_1 \cdot \frac{1}{\mathit{flaeche}}$$

wieder mit der KQ-Methode geschätzt werden können. Man erhält die geschätzte Funktion

$$\hat{f}(\mathit{flaeche}) = 10.62 + 269.74 \cdot \frac{1}{\mathit{flaeche}}.$$

Die zugehörige Kurve in Abbildung 2.2 (rechts) ist besser an die Daten angepasst. Die Interpretation ist nun: Für einen gegebenen Wert der Wohnfläche, z.B. $\mathit{flaeche} = 30$ qm, ist

$$\widehat{\mathit{mieteqm}} = 10.62 + 269.74 \frac{1}{\mathit{flaeche}}$$

die geschätzte durchschnittliche Miete pro Quadratmeter. Nimmt die Wohnfläche um 1 qm auf $\mathit{flaeche} + 1$, z.B. auf 31 qm, zu, vermindert sich die durchschnittliche Miete auf

$$\widehat{\mathit{mieteqm}} = 10.62 + 269.74 \frac{1}{\mathit{flaeche} + 1}\,.$$

Wie auch aus Abbildung 2.2 (rechts) ersichtlich, ist die Verminderung nichtlinear. Sie kann durch Einsetzen der konkreten Werte (z.B. 30 qm und 31 qm) berechnet werden:

$$\widehat{\mathit{mieteqm}}(30) - \widehat{\mathit{mieteqm}}(31) = 269.74/30 - 269.74/31 \approx 0.29\,\mathrm{DM}\,.$$

Bei einer Wohnung mit 60 qm sinkt die Durchschnittsmiete pro qm um

$$\widehat{\mathit{mieteqm}}(60) - \widehat{\mathit{mieteqm}}(61) \approx 0.07\,\mathrm{DM}\,.$$

△

Allgemein gilt: Entscheidend für die Anwendung eines linearen Regressionsmodells ist eine *in den Regressionskoeffizienten* β_0 und β_1 *lineare Beziehung*. Die Regressorvariable x – und auch die Zielvariable y – dürfen dazu geeignet transformiert werden, so wie im obigen Beispiel die ursprüngliche Variable *flaeche*. Es verbleibt natürlich die Frage: Wie findet man eine geeignete Transformation? Eine flexible Möglichkeit bieten nichtparametrische Regressionsmodelle, die ausführlich in den Kapiteln 7 und 8 behandelt werden.

2.2.2 Das multiple lineare Regressionsmodell

Beispiel 2.3 Mietspiegel – Mieten in normaler und guter Lage

Wir nehmen nun Wohnungen mit guter Lage hinzu und markieren im Streudiagramm der Abbildung 2.3 Datenpunkte für Mieten in normaler und guter Lage entsprechend. Zusätzlich zur geschätzten Regressionsgeraden für Wohnungen in normaler Lage ist eine entsprechend separat geschätzte Regressionsgerade für Wohnungen in guter Lage im Streudiagramm eingezeichnet. Alternativ kann man beide Schichten gemeinsam mit einem Modell analysieren, bei dem die Geraden nur parallel verschoben sind. Dies lässt sich durch das Modell

$$\mathit{miete}_i = \beta_0 + \beta_1\, \mathit{flaeche}_i + \beta_2\, \mathit{glage}_i + \varepsilon_i \tag{2.4}$$

erreichen. Dabei ist *glage* eine binäre *Indikatorvariable*

$$\mathit{glage}_i = \begin{cases} 1 & \text{falls sich die } i\text{-te Wohnung in guter Lage befindet,} \\ 0 & \text{falls sich die } i\text{-te Wohnung in normaler Lage befindet.} \end{cases}$$

Mit der KQ-Methode erhält man als geschätzte Durchschnittsmiete

$$\widehat{\mathit{miete}} = 219.74 + 11.40 \cdot \mathit{flaeche} + 111.66 \cdot \mathit{glage}.$$

Äquivalent dazu ist wegen der 1/0-Kodierung der Lage die Darstellung

$$\widehat{\mathit{miete}} = \begin{cases} 331.4 + 11.40 \cdot \mathit{flaeche} & \text{für gute Lage,} \\ 219.74 + 11.40 \cdot \mathit{flaeche} & \text{für normale Lage.} \end{cases}$$

Diese beiden parallelen Geraden sind in Abbildung 2.4 eingetragen.

Die Koeffizienten lassen sich so interpretieren:

- In guter wie in normaler Lage führt die Erhöhung der Wohnfläche um 1 qm zur Erhöhung der durchschnittlichen Miete um 11.40 DM.
- Bei gleicher Wohnfläche ist die durchschnittliche Miete für eine Wohnung in guter Lage um 111.66 DM höher als für eine entsprechende Wohnung in normaler Lage.

△

Das Modell (2.4) ist ein Spezialfall des *multiplen linearen Regressionsmodells* für k Regressoren bzw. Kovariablen $x_1, \ldots, x_k$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i.$$

Dabei ist x_{ij} der Wert der j-ten Kovariable für die i-te Beobachtung, $i = 1, \ldots, n$. Die Kovariablen können metrisch, binär oder auch mehrkategorial (nach geeigneter Kodierung) sein. Ebenso wie bei der linearen Einfachregression können x-Variablen auch durch Transformation aus ursprünglichen Regressoren gewonnen werden. Für die Fehlervariablen ε_i werden die gleichen Annahmen wie für das einfache lineare Regressionsmodell getroffen. Bei Normalverteilungsannahme folgt dann wieder, dass die Zielvariablen bei gegebenen Kovariablenwerten (bedingt) unabhängig und normalverteilt sind:

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2),$$

mit

Abb. 2.3. *Die linke Grafik zeigt das Streudiagramm zwischen Nettomiete und Wohnfläche für Wohnungen in normaler (Kreise) und guter Lage (Pluszeichen). Die rechte Grafik zeigt separat geschätzte Regressionsgeraden für Wohnungen in normaler (durchgezogene Linie) und guter Lage (gestrichelte Linie).*

Abb. 2.4. *Nach Modell (2.4) geschätzte Regressionsgeraden für Wohnungen in normaler (durchgezogene Linie) und guter Lage (gestrichelte Linie).*

$$\mu_i = \mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

Die folgenden Beispiele illustrieren, wie flexibel das multiple lineare Regressionsmodell durch geeignete Transformation und Kodierung von Regressoren einsetzbar ist.

Beispiel 2.4 Mietspiegel – Nichtlinearer Einfluss der Wohnfläche

Wie in Beispiel 2.2 transformieren wir die Wohnfläche zu $x = \frac{1}{flaeche}$ und formulieren

$$mieteqm_i = \beta_0 + \beta_1 \cdot \frac{1}{flaeche_i} + \beta_2\, glage_i + \varepsilon_i \tag{2.5}$$

als gemeinsames Modell. Das geschätzte Modell für die Durchschnittsmiete pro qm ist

$$\widehat{mieteqm} = 10.74 + 262.70 \cdot \frac{1}{flaeche} + 1.75 \cdot glage.$$

Die beiden Kurven für die durchschnittlichen Quadratmetermieten

$$\widehat{mieteqm} = \begin{cases} 12.49 + 262.70 \cdot \frac{1}{flaeche} & \text{in guter Lage} \\ 10.74 + 262.70 \cdot \frac{1}{flaeche} & \text{in normaler Lage} \end{cases}$$

Abb. 2.5. *Links: Streudiagramm zwischen Nettomiete pro qm und Wohnfläche für Wohnungen in normaler (Kreise) und guter Lage (Pluszeichen). Rechts: Geschätzte Regressionskurven für Wohnungen in normaler (durchgezogene Linie) und guter Lage (gestrichelte Linie).*

sind in Abbildung 2.5 eingetragen. Der nichtlineare Einfluss der Wohnfläche ist wie in Beispiel 2.2 zu interpretieren.

△

In den Beispielen 2.3 und 2.4 hat die Lage einen rein additiven Effekt. In beiden Modellen ergibt eine gute Lage einen Zuschlag im Vergleich zu einer Wohnung mit gleicher Wohnfläche in normaler Lage. Dieser beträgt in Beispiel 2.3 111.66 DM und hier 1.75 DM pro Quadratmeter. Im Modell (2.4) folgt aus der Annahme eines rein additiven Effekts die Parallelität der Geraden in Abbildung 2.4. Vergleicht man dies mit Abbildung 2.3, so erscheint diese Annahme zweifelhaft. Durch das Einbeziehen einer *Interaktion* zwischen den beiden Regressoren *flaeche* und *lage* kann man sich von dieser Annahme lösen.

Beispiel 2.5 Mietspiegel – Interaktion zwischen Wohnfläche und Lage

Um eine Interaktion zwischen Wohnfläche und Lage in das Modell (2.4) einzubeziehen, definieren wir durch Multiplikation der Regressoren *flaeche* und *glage* die Interaktionsvariable *inter* mit den Werten

$$inter_i = flaeche_i \cdot glage_i \,.$$

Damit gilt

$$inter_i = \begin{cases} flaeche_i & \text{in guter Lage,} \\ 0 & \text{in normaler Lage.} \end{cases}$$

Wir erweitern Modell (2.4), indem wir neben den beiden *Haupteffekten flaeche* und *glage* auch den *Interaktionseffekt inter* = *flaeche* · *glage* einbeziehen, zu

$$miete_i = \beta_0 + \beta_1\, flaeche_i + \beta_2\, glage_i + \beta_3\, inter_i + \varepsilon_i \,. \tag{2.6}$$

Wegen der Definition von *glage* und *inter* ergibt sich

$$miete_i = \begin{cases} \beta_0 + \beta_1\, flaeche_i + \varepsilon_i & \text{für normale Lage,} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, flaeche_i + \varepsilon_i & \text{für gute Lage.} \end{cases}$$

Für $\beta_3 = 0$ ist kein Interaktionseffekt vorhanden und wir erhalten Modell (2.4) mit der Annahme paralleler Geraden, d.h. gleicher Steigung β_1 zurück. Für $\beta_3 \neq 0$ ist der Effekt

Abb. 2.6. *Basierend auf dem Interaktionsmodell (2.6) geschätzte Regressionsgeraden für normale (durchgezogene Linie) und gute Wohnlagen (gestrichelte Linie).*

der Wohnfläche, d.h. die Steigung der Geraden für Wohnungen in guter Lage, um den Wert β_3 im Vergleich zu Wohnungen in normaler Lage verändert.

Die KQ-Schätzung wird jedoch nicht wie in Abbildung 2.3 (rechts) separat für die beiden Schichten durchgeführt, sondern für das Modell (2.6) mit den Daten beider Schichten gemeinsam. Es ergibt sich

$$\hat{\beta}_0 = 253.95, \quad \hat{\beta}_1 = 10.87, \quad \hat{\beta}_2 = 10.15, \quad \hat{\beta}_3 = 1.60.$$

Die geschätzten Regressionsgeraden für gute und normale Wohnlagen findet man in Abbildung 2.6. Ob die Modellierung eines Interaktionseffekts notwendig ist, kann durch einen Test für die Hypothesen

$$H_0 : \beta_3 = 0 \;\text{ gegen }\; H_1 : \beta_3 \neq 0$$

geprüft werden, vergleiche Kapitel 3.3.

△

Wie in Beispiel 1.1 (Seite 5) beschrieben, wird die Wohnlage im gesamten Datensatz in die drei Kategorien

$$\begin{aligned} 1 &= \text{normale Lage} \\ 2 &= \text{gute Lage} \\ 3 &= \text{beste Lage} \end{aligned}$$

eingeteilt. Da die Lagevariable kategorial und nicht metrisch ist, kann der Effekt der Lage nicht in der Form $\beta \cdot lage$, mit den Werten 1, 2 oder 3 für *lage*, in einem linearen Regressionsmodell dargestellt werden. Das würde nämlich bedeuten, dass die willkürlich gewählte Kodierung der Lagevariable einen erheblichen Einfluss auf die Schätzergebnisse erhält. Die hier gewählte Kodierung würde dazu führen, dass Wohnungen in guter Lage einen doppelt so großen Effekt auf die Miete haben und Wohnungen in bester Lage einen dreimal so großen Effekt. Diese Relationen ändern sich automatisch bei veränderter Kodierung. Mit der Kodierung 1, 4, 6 für normale, gute und beste Lage hätten wir für gute bzw. beste Lagen einen viermal bzw. sechsmal so großen Effekt wie in normaler Lage.

Ähnlich wie die Lage in Beispiel 2.3 durch *eine* binäre Indikatorvariable kodiert wurde, ist jetzt eine Kodierung durch *zwei* binäre Variablen notwendig. Dazu wählt man eine der

drei Lagekategorien als *Referenzkategorie*. Wir wählen die normale Lage. Dann werden die beiden 1/0-Indikatorvariablen *glage* und *blage* für gute bzw. beste Lage durch

$$glage_i = \begin{cases} 1 & \text{falls sich Wohnung } i \text{ in guter Lage befindet,} \\ 0 & \text{sonst,} \end{cases}$$

$$blage_i = \begin{cases} 1 & \text{falls sich Wohnung } i \text{ in bester Lage befindet,} \\ 0 & \text{sonst,} \end{cases}$$

definiert. Eine Wohnung i der Referenzkategorie normale Lage ist somit durch

$$glage_i = blage_i = 0$$

definiert. Die Effekte der beiden binären Variablen *glage* und *blage* im Regressionsmodell werden dann stets mit Bezug auf die Referenzkategorie interpretiert, vergleiche auch das nachfolgende Beispiel.

Diese Art der 1/0-Kodierung einer *mehrkategorialen Variable* nennt man auch *Dummy-Kodierung*. Für eine Variable x mit c Kategorien, also $x \in \{1, \ldots, c\}$, ist diese Dummy-Kodierung folgendermaßen definiert: Man wählt eine Kategorie, zum Beispiel c, als *Referenzkategorie* und kodiert x durch $c-1$ *Dummy-Variablen* $x_1, \ldots, x_{c-1}$:

$$x_j = \begin{cases} 1 & \text{Kategorie } j \text{ liegt vor,} \\ 0 & \text{sonst,} \end{cases} \qquad j = 1, \ldots, c-1\,.$$

Für die Referenzkategorie c gilt dann

$$x_1 = 0, \ldots, x_{c-1} = 0.$$

Mehr Details zur Kodierung kategorialer Kovariablen findet man Kapitel 3.1.4.

Beispiel 2.6 Mietspiegel – Multiples Regressionsmodell

Zur Illustration analysieren wir den gesamten Datensatz mit allen in Beispiel 1.1 genannten erklärenden Variablen mit einem multiplen Regressionsmodell für die Miete pro Quadratmeter. Den nichtlinearen Effekt der Wohnfläche modellieren wir wieder durch die transformierte Variable $1/flaeche$ und die Lage durch die beschriebene Dummy-Kodierung. Da der Einfluss des Baujahrs vermutlich ebenfalls nichtlinear ist, setzen wir dazu ein einfaches Polynom vom Grad 2 an. Damit ergibt sich für ein Modell ohne Interaktionen der Ansatz

$$\begin{aligned} mieteqm_i = {} & \beta_0 + \beta_1 \cdot (1/flaeche_i) + \beta_2\, bjahr_i + \beta_3\, bjahr_i^2 + \beta_4\, glage_i + \beta_5\, blage_i \\ & + \beta_6\, bad_i + \beta_7\, kueche_i + \beta_8\, zh_i + \varepsilon_i. \end{aligned}$$

Die binären Regressoren *bad*, *kueche* und *zh* sind dabei wie in Tabelle 1.2 (Seite 6) kodiert. Abbildung 2.7 zeigt die geschätzten nichtlinearen Effekte von Wohnfläche und Baujahr. Die Kurven kommen dadurch zustande, dass in $\widehat{mieteqm}$ nur die Wohnfläche (bzw. nur das Baujahr) variiert und für die übrigen Kovariablen der jeweilige Mittelwert eingesetzt wird. Tabelle 2.1 enthält die geschätzten Koeffizienten $\hat{\beta}_4$ bis $\hat{\beta}_8$ der restlichen Regressoren.

Zur Interpretation eines Effektes hält man gedanklich die Werte der restlichen Regressoren fest. Für zwei Wohnungen mit den Wohnflächen 60 qm bzw. 100 qm, jedoch sonst gleichen Werten für das Baujahr sowie die Lage, Bad-, Küchen- und Zentralheizungsindikatoren, ist dann die Differenz $\hat{\beta}_1(1/60) - \hat{\beta}_1(1/100) = 268.13\,(1/60 - 1/100) = 1.79$

Klassisches lineares Regressionsmodell

Daten

$(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, zu einer metrischen Variablen y und metrischen oder binär kodierten kategorialen Regressoren $x_1, \ldots, x_k$.

Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i\,, \quad i = 1, \ldots, n\,.$$

Die Fehler $\varepsilon_1, \ldots, \varepsilon_n$ sind unabhängig und identisch verteilt (i.i.d.) mit

$$\mathrm{E}(\varepsilon_i) = 0\,, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

Die geschätzte lineare Funktion

$$\hat{f}(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

kann als Schätzung $\widehat{\mathrm{E}}(y|x_1, \ldots, x_k)$ für den bedingten Erwartungswert von y bei gegebenen Kovariablen $x_1, \ldots, x_k$ angesehen und damit zur Prognose von y verwendet werden. Diese wird wieder mit $\hat{y}$ bezeichnet.

Abb. 2.7. *Effekte der Wohnfläche (links) und des Baujahrs (rechts).*

der Durchschnittsmieten pro qm am eingezeichneten Wohnflächen-Effekt in der Abbildung 2.7 links direkt ablesbar. Wie erwartet, nimmt der Einfluss auf die Nettomiete pro qm mit wachsender Wohnfläche (nichtlinear) ab. Analog interpretiert man den Effekt des Baualters.

Die Effekte der Indikatorvariablen in Tabelle 2.1 interpretiert man als Zuschläge auf die Nettomiete pro qm im Vergleich zur jeweiligen Referenzkategorie. Beispielsweise erhöht sich die Durchschnittsmiete pro qm bei guter Lage um 1.32 DM gegenüber einer vergleichbaren Wohnung in normaler Lage.

△

Variable	geschätzter Koeffizient
1/flaeche	268.134
bjahr	-7.411
$bjahr^2$	0.002
glage	1.325
blage	2.961
bad	0.980
kueche	1.689
zh	3.647

Tabelle 2.1. *Geschätzte Koeffizienten des multiplen Modells.*

2.3 Regression bei binären Zielvariablen: Das Logit-Modell

Das lineare Regressionsmodell ist vor allem für stetige Zielvariablen geeignet, die – eventuell nach geeigneter Transformation – approximativ normalverteilt sind. In vielen Anwendungen treten jedoch binäre oder, allgemeiner, kategoriale Zielvariablen auf.

Beispiel 2.7 Einsprüche gegen Patente

Während der Prüfung eines Patentantrages kann es zu einem Einspruch kommen, vergleiche Beispiel 1.3 (Seite 8). Die Zielvariable (*einspruch*) ist binär und kodiert durch

$$einspruch_i = \begin{cases} 1 & \text{falls ein Einspruch gegen Patent } i \text{ erfolgt,} \\ 0 & \text{sonst.} \end{cases}$$

Die Entscheidung für einen Einspruch wird von verschiedenen Kovariablen beeinflusst, die teilweise metrisch sind, wie das Antragsjahr (Variable *jahr*), die Anzahl der Zitationen (*azit*) und die Anzahl der Länder (*aland*) und teilweise binär, siehe Tabelle 1.4 (Seite 8).

△

Der Erwartungswert einer binären Variable y ist gegeben durch

$$\mathrm{E}(y) = P(y = 0) \cdot 0 + P(y = 1) \cdot 1 = P(y = 1).$$

Ziel einer Regressionsanalyse mit binärer Zielvariable $y \in \{0, 1\}$ ist also die Modellierung und Analyse der Wahrscheinlichkeit

$$\mathrm{P}(y = 1) = \mathrm{P}(y = 1 \,|\, x_1, \ldots, x_k) = \pi$$

in Abhängigkeit von den Kovariablen. Ein übliches lineares Regressionsmodell

$$y_i = \mathrm{P}(y_i = 1) + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

mit $\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$ ist aus verschiedenen Gründen ungeeignet:

- Die rechte Seite ist – im Gegensatz zur linken – nicht binär.

- Auch wenn man auf die Normalverteilungsannahme für ε_i verzichtet, kann die Fehlervarianz $\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(y_i \mid x_i)$ nicht homoskedastisch, d.h. gleich σ^2 sein. Da y_i Bernoulliverteilt ist mit $\pi_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$, folgt, dass

$$\mathrm{Var}(y_i) = \pi_i(1 - \pi_i)$$

  ebenfalls von den Kovariablen und den Parametern $\beta_0, \ldots, \beta_k$ abhängt und somit nicht für alle i den gleichen Wert σ^2 besitzen kann.
- Das lineare Modell lässt für $\mathrm{P}(y_i = 1)$ auch Werte $\pi_i < 0$ und $\pi_i > 1$ zu, was für Wahrscheinlichkeiten nicht zulässig ist.

Diese Probleme lassen sich beseitigen, wenn man das Modell

$$\pi_i = \mathrm{P}(y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$$

annimmt, wobei der Wertebereich der Funktion F im Intervall $[0, 1]$ liegen soll. Da es aus interpretatorischen Gründen sinnvoll ist, dass F auch streng monoton wächst, bieten sich für F Verteilungsfunktionen an. Wählt man die logistische Verteilungsfunktion

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)},$$

so erhält man das Logit-Modell

$$\mathrm{P}(y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

mit dem *linearen Prädiktor*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

Analog zum linearen Regressionsmodell wird angenommen, dass die binären Zielvariablen y_i bei gegebenen Kovariablenwerten $x_i = (x_{i1}, \ldots, x_{ik})'$ (bedingt) unabhängig sind. Obwohl der Prädiktor linear ist, verändert sich die Interpretation im Vergleich zum linearen Modell: Erhöht sich der Wert des Prädiktors η um eine Einheit auf $\eta + 1$, so erhöht sich die Wahrscheinlichkeit für $y = 1$ *nichtlinear* von $F(\eta)$ auf $F(\eta + 1)$. Eine alternative Interpretation ergibt sich durch Auflösen der Modellgleichung mit Hilfe der Umkehrfunktion $\eta = \log\{\pi/(1 - \pi)\}$ der logistischen Funktion $\pi = \exp(\eta)/\{1 + \exp(\eta)\}$. Man erhält

$$\log\left(\frac{\mathrm{P}(y_i = 1)}{1 - \mathrm{P}(y_i = 1)}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{2.7}$$

bzw. wegen $\exp(a + b) = \exp(a) \cdot \exp(b)$

$$\frac{\mathrm{P}(y_i = 1)}{\mathrm{P}(y_i = 0)} = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik}). \tag{2.8}$$

Die linke Seite von (2.8), also der Quotient der Wahrscheinlichkeiten für $y = 1$ und $y = 0$, wird als *Chance (odds)* bezeichnet. Entsprechend ist die linke Seite von (2.7) die *logarithmierte Chance (log-odds)* für das Auftreten von $y = 1$ und $y = 0$. Für die *Chance* erhält man somit ein multiplikatives Modell: Wird z.B. der Wert x_{i1} der Variable x_1 um 1 erhöht, so wird der Quotient in (2.8) mit dem Faktor $\exp(\beta_1)$ multipliziert:

Das Logit-Modell für binäre Zielvariablen

Daten

$(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, zu einer binären Zielvariablen $y \in \{0, 1\}$ und metrischen oder binär kodierten Kovariablen $x_1, \ldots, x_k$.

Modell

Für die (bedingt) unabhängigen binären Zielvariablen $y_i \in \{0, 1\}$ wird für $\pi_i = \mathrm{P}(y_i = 1)$ im Logit-Modell der Ansatz

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

gewählt. Dazu äquivalent ist die Annahme

$$\frac{P(y_i = 1)}{P(y_i = 0)} = \frac{\pi_i}{1 - \pi_i} = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik})$$

eines multiplikativen Modells für die Chance $\pi_i / 1 - \pi_i$.

$$\begin{aligned} \frac{\mathrm{P}(y_i = 1 \,|\, x_{i1} + 1, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1} + 1, \ldots)} &= \exp(\beta_0) \exp(\beta_1 (x_{i1} + 1)) \cdot \ldots \cdot \exp(\beta_k x_{ik}) = \\ &= \frac{\mathrm{P}(y_i = 1 \,|\, x_{i1}, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1}, \ldots)} \exp(\beta_1). \end{aligned} \tag{2.9}$$

Ist x_1 speziell eine binäre Variable, so gilt

$$\frac{\mathrm{P}(y_i = 1 \,|\, x_{i1} = 1, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1} = 1, \ldots)} = \frac{\mathrm{P}(y_i = 1 \,|\, x_{i1} = 0, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1} = 0, \ldots)} \exp(\beta_1). \tag{2.10}$$

Für $\beta_1 > 0$ vergrößert sich also die Chance $\mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0)$, für $\beta_1 < 0$ verkleinert sie sich und für $\beta_1 = 0$ bleibt sie unverändert.

Für die *logarithmierte Chance* in (2.7) gilt wieder die übliche Interpretation des linearen Modells: Erhöht sich x_1 um 1, so verändert sich die logarithmierte Chance um β_1.

Da die Annahmen für das lineare Regressionsmodell nicht erfüllt sind, werden die Parameter nicht mit der KQ-Methode geschätzt, sondern nach der Maximum-Likelihood-(ML)-Methode, siehe Kapitel 4 bzw. Anhang B.4.1.

Abb. 2.8. *Histogramme und Kerndichteschätzer für die metrischen Kovariablen ansp (links) und azit (rechts).*

Beispiel 2.8 Einsprüche gegen Patente

Bevor wir uns der Analyse der Einspruchswahrscheinlichkeit widmen, werfen wir einen Blick auf Abbildung 2.8, in der Histogramme und Kerndichteschätzer für die beiden metrischen Kovariablen *ansp* und *azit* abgedruckt sind. Die Verteilungen beider Variablen sind extrem linkssteil. Der Großteil der Beobachtungen für *ansp* liegt zwischen 0 und 60 mit einigen wenigen Beobachtungen zwischen 61 und dem Maximalwert 355. Die Variable *azit* schwankt größtenteils zwischen 0 und 15. Einige wenige Beobachtungen sind größer als 15 mit dem Maximalwert bei 40. Aufgrund der sehr geringen Beobachtungszahl für $ansp > 60$ und $azit > 15$ sind in diesen Bereichen keine sinnvollen Aussagen über den Einfluss auf die Einspruchswahrscheinlichkeit zu erwarten. Daher sind diese extremen Beobachtungen von der nachfolgenden Analyse ausgeschlossen. Dieses Beispiel zeigt wie wichtig die deskriptive Analyse der Daten ist.

Wir unterteilen jetzt die Daten in die Teilschichten $biopharm = 0$ und $biopharm = 1$. Für die Teilschicht $biopharm = 0$, d.h. für Patente aus der Halbleiter-/Computer-Branche berechnen wir mit den restlichen Kovariablen aus Beispiel 2.7 ein Logit-Modell

$$\mathrm{P}(einspruch_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1\, jahr_i + \beta_2\, azit_i + \beta_3\, ansp_i + \beta_4\, uszw_i + \beta_5\, patus_i + \beta_6\, patdsg_i + \beta_7\, aland_i$$

für die Haupteffekte. Tabelle 2.2 enthält die geschätzten Koeffizienten $\hat{\beta}_j$, $j = 0, \ldots, 7$, sowie die jeweiligen relativen Chancen (odds ratios) $\exp(\hat{\beta}_j)$. In der multiplikativen Form (2.8) ergibt sich also

$$\frac{\mathrm{P(Einspruch)}}{\mathrm{P(kein\ Einspruch)}} = \exp(201.74) \cdot \exp(-0.102 \cdot jahr_i) \cdot \ldots \cdot \exp(0.097 \cdot aland_i).$$

Damit erhöht sich nach diesem Modell die Chance für einen Einspruch gegen ein Patent aus Deutschland, der Schweiz oder Großbritannien ($patdsg = 1$) um den Faktor $1.217 = \exp(0.196)$ im Vergleich zu einem Patent, das bei sonst identischen Kovariablenwerten nicht aus diesen Ländern oder den USA kommt. Durch Einsetzen der Kovariablenwerte für ein neu beantragtes Patent lässt sich dann die Chance P(Einspruch) / P(kein Einspruch) mit Hilfe des Modells prognostizieren.

Variable	Geschätzter Koeffizient	Geschätzte relative Chance
Konstante	$\hat{\beta}_0 = 201.74$	
jahr	$\hat{\beta}_1 =$ -0.102	$\exp(\hat{\beta}_1) = 0.902$
azit	$\hat{\beta}_2 =$ 0.113	$\exp(\hat{\beta}_2) = 1.120$
ansp	$\hat{\beta}_3 =$ 0.026	$\exp(\hat{\beta}_3) = 1.026$
uszw	$\hat{\beta}_4 =$ -0.402	$\exp(\hat{\beta}_4) = 0.668$
patus	$\hat{\beta}_5 =$ -0.526	$\exp(\hat{\beta}_5) = 0.591$
patdsg	$\hat{\beta}_6 =$ 0.196	$\exp(\hat{\beta}_6) = 1.217$
aland	$\hat{\beta}_7 =$ 0.097	$\exp(\hat{\beta}_7) = 1.102$

Tabelle 2.2. *Einsprüche gegen Patente: Geschätzte Koeffizienten und relative Chancen für das Logit-Modell.*

Abb. 2.9. *Einsprüche gegen Patente: Linearer und kubischer Effekt der Variable aland.*

Wie beim linearen Regressionsmodell ist fraglich, ob der Einfluss der metrischen Kovariablen linear oder nichtlinear ist. Wir modellieren exemplarisch den Effekt der Kovariable *aland* durch ein kubisches Polynom

$$\beta_7\, aland + \beta_8\, aland^2 + \beta_9\, aland^3\,.$$

Abbildung 2.9 zeigt das geschätzte Polynom im Vergleich zum linearen Effekt. Beim Zeichnen der Effekte wurden für die übrigen Kovariablen wieder die jeweiligen Mittelwerte eingesetzt. Die Schätzwerte für die Koeffizienten des Polynoms sind

$$\hat{\beta}_7 = 0.3938 \quad \hat{\beta}_8 = -0.0378 \quad \hat{\beta}_9 = 0.0014\,.$$

Sowohl die Abbildung wie auch die Koeffizienten deuten darauf hin, dass der Einfluss von *aland* in linearer Form bereits hinreichend gut modelliert wird. Diese Hypothese kann mit einem Test überprüft werden, vergleiche Kapitel 4.1.

△

Neben dem Logit-Modell existieren weitere Regressionsmodelle, die sich ergeben, wenn man die Verteilungsfunktion der logistischen Verteilung durch eine alternative Verteilungsfunktion ersetzt. Beispielsweise ergibt sich für $F = \Phi$, mit Φ als Verteilungsfunktion der Standardnormalverteilung, das sogenannte Probit-Modell, vergleiche Kapitel 4.

Darüber hinaus treten in Anwendungen neben binären Zielvariablen auch andere Typen diskreter Zielvariablen auf, für die lineare Regressionsmodelle nicht oder nur schlecht zur Analyse geeignet sind. Dazu gehören Regressionssituationen, in denen y eine Zählvariable mit Werten aus $\{0, 1, 2, \ldots\}$ ist, wie zum Beispiel die Anzahl von Schadensfällen eines Versicherungsunternehmers (vergleiche hierzu auch Beispiel 2.12), oder eine mehrkategoriale Variable, etwa mit den Kategorien schlecht, mittel, gut. Regressionsmodelle für solche Typen von diskreten Zielvariablen werden in den Kapiteln 4 und 5 beschrieben.

2.4 Gemischte Modelle

Die bisherigen Regressionsmodelle sind vor allem zur Analyse von Regressionsdaten geeignet, die bei Querschnittsstudien auftreten. Die Regressionskoeffizienten $\beta_0, \ldots, \beta_k$ werden dabei als unbekannte, aus den Daten zu schätzende Populationsparameter aufgefasst. Problemstellungen der Regression ergeben sich aber auch bei der Analyse von Longitudinaldaten, bei denen zeitlich wiederholte Beobachtungen von Individuen bzw. Objekten im Rahmen von Längsschnittstudien vorliegen. Dann lassen sich nicht nur feste Populationseffekte, sondern auch individuenspezifische Effekte modellieren und schätzen. Man fasst diese als „zufällige Effekte" auf, da sie zu Individuen gehören, die „zufällig" der Population entnommen wurden. Eng verwandt damit ist die Analyse von sogenannten Clusterdaten, wenn aus Primäreinheiten (Clustern) jeweils mehrere Individuen ausgewählt und dazu Beobachtungen zu interessierenden Variablen erhoben werden. Zum Beispiel können die Cluster ausgewählte Schulen sein, in denen für eine Teilstichprobe von Schülern Tests durchgeführt werden.

Gemischte Modelle (Mixed Models, Modelle mit zufälligen Effekten) beziehen in den Prädiktor neben den bisher betrachteten festen Populationseffekten $\beta_0, \ldots, \beta_k$ zusätzlich individuen- bzw. clusterspezifische zufällige Effekte mit ein. Deren Modellierung und Schätzung ermöglicht weitergehende Analysen auf individuenspezifischer Ebene. Dies wird im folgenden Beispiel für den Fall von Longitudinaldaten illustriert.

Beispiel 2.9 Hormontherapie bei Ratten

Um die Wirkung von Testosteron auf das Wachstum von Ratten zu untersuchen, wurde an der KUL (Katholieke Universiteit Leuven, Belgien) das im Folgenden beschriebene Experiment durchgeführt. Ausführlichere Beschreibungen und Datenanalysen finden sich bei Verbeke & Molenberghs (2000). Insgesamt 50 Ratten wurden zufällig einer Kontrollgruppe oder einer von zwei Therapiegruppen zugewiesen. Als Therapie wurde dabei eine niedrige oder hohe Dosis des Mittels Decapeptyl gegeben, mit dem die Testosteronproduktion bei Ratten gehemmt wird. Die Behandlung begann im Alter von 45 Tagen. Beginnend mit dem 50. Tag, wurde alle 10 Tage das Wachstum des Kopfes mittels Röntgenuntersuchung gemessen. Als Zielvariable diente dabei der Abstand (gemessen in Pixeln) zwischen zwei wohldefinierten Punkten des Kopfes, welche die Höhe des Kopfes charakterisieren. Die Anzahl n_i von wiederholten Messungen y_{ij}, $j = 1, \ldots, n_i$, dieser Zielvariable war für die Ratten $i = 1, \ldots, 50$, unterschiedlich. An 22 Ratten wurden insgesamt sieben Messungen bis zum Alter von 110 Tagen durchgeführt, während vier Ratten nur einmal zu Beginn im Alter von 50 Tagen untersucht wurden. Tabelle 2.3 beschreibt das so entstandene Beobachtungsdesign der Studie und Abbildung 2.10 zeigt die nach den drei Gruppen getrennten, individuellen Zeitreihen $\{y_{ij}, j = 1, \ldots, n_i\}$ für die Ratten $i = 1, \ldots, 50$.

Abb. 2.10. *Hormontherapie bei Ratten: Nach Dosierungsgruppe getrennte Zeitreihen.*

Alter (in Tagen)	Kontrolle	Niedrig	Hoch	Gesamt
50	15	18	17	50
60	13	17	16	46
70	13	15	15	43
80	10	15	13	38
90	7	12	10	29
100	4	10	10	24
110	4	8	10	22

Tabelle 2.3. *Anzahl der Beobachtungen pro Zeitpunkt und Dosierungsgruppe.*

Zur Formulierung von Regressionsmodellen bilden wir (wie Verbeke & Molenberghs (2000)) die metrische Kovariable *transformiertes Alter*

$$t = \log(1 + (alter - 45)/10).$$

Der Wert $t = 0$ entspricht dann dem Behandlungsbeginn (Alter = 45 Tage).
Für die drei Gruppen definieren wir die Indikatorvariablen C, N, H

$$C_i = \begin{cases} 1 & \text{Ratte } i \text{ in Kontrollgruppe,} \\ 0 & \text{sonst,} \end{cases}$$

$$N_i = \begin{cases} 1 & \text{Ratte } i \text{ in Niedrigdosisgruppe,} \\ 0 & \text{sonst,} \end{cases}$$

$$H_i = \begin{cases} 1 & \text{Ratte } i \text{ in Hochdosisgruppe,} \\ 0 & \text{sonst.} \end{cases}$$

Mit dem logarithmisch transformierten Alter t als Zeitskala und $t = 0$ als Behandlungsbeginn kann man nach Gruppen getrennte, einfache lineare Regressionsmodelle

$$y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} & i \text{ in Niedrigdosisgruppe,} \\ \beta_0 + \beta_2 t_{ij} + \varepsilon_{ij} & i \text{ in Hochdosisgruppe,} \\ \beta_0 + \beta_3 t_{ij} + \varepsilon_{ij} & i \text{ in Kontrollgruppe,} \end{cases}$$

formulieren. Dabei gilt für $t = 0$ in allen drei Gruppen $\mathrm{E}(y_{ij}) = \beta_0$, d.h. β_0 ist das *Populationsmittel* zu Behandlungsbeginn. Die Koeffizienten β_1, β_2 und β_3 entsprechen unterschiedlichen Steigungen der Variablen t, d.h. Effekten des (transformierten) Alters, in den drei linearen Modellen. Dies lässt sich auch in einem Gesamtmodell

$$y_{ij} = \beta_0 + \beta_1\, N_i \cdot t_{ij} + \beta_2\, H_i \cdot t_{ij} + \beta_3\, C_i \cdot t_{ij} + \varepsilon_{ij} \tag{2.11}$$

mit den 1/0-Indikatorvariablen N, H und C für die drei Gruppen zusammenfassen. Ebenso wie β_0 sind die Parameter β_1, β_2 und β_3 *Populationseffekte*, die keine individuellen Unterschiede zwischen den Ratten erfassen können. Aus Abbildung 2.10 wird aber bereits visuell deutlich, dass die individuellen Verlaufskurven offensichtliche Unterschiede im Niveau und möglicherweise auch in ihren Steigungen aufweisen. Zudem ist die Variabilität innerhalb der individuellen Kurvenverläufe deutlich geringer als die gesamte Variation der Daten in den jeweiligen Streudiagrammen der drei Gruppen. Die Berücksichtigung individuenspezifischer Information wirkt sich deshalb auch positiv auf die Qualitität der Schätzung aus.

Um die individuellen Effekte in einem Modell abzubilden, erweitern wir die obigen Regressionsansätze zu

$$y_{ij} = \begin{cases} \beta_0 + \gamma_{0i} + (\beta_1 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in Niedrigdosisgruppe,} \\ \beta_0 + \gamma_{0i} + (\beta_2 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in Hochdosisgruppe,} \\ \beta_0 + \gamma_{0i} + (\beta_3 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in Kontrollgruppe,} \end{cases}$$

bzw. zum Gesamtmodell

$$y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 N_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} + \gamma_{1i} \cdot t_{ij} + \varepsilon_{ij} \tag{2.12}$$

mit individuenspezifischen Abweichungen γ_{0i} von der Populationskonstanten β_0 und individuenspezifischen Abweichungen γ_{1i} von den Populationssteigungen β_1, β_2 und β_3. Im Gegensatz zu den „fixen" Effekten $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ werden die individuenspezifischen Effekte $\boldsymbol{\gamma}_i = (\gamma_{0i}, \gamma_{1i})'$ als zufällige Größen angesehen, da die Ratten eine Zufallsauswahl aus einer Population sind. Wir treffen dazu die spezifische Annahme, dass die zufälligen Effekte unabhängig und identisch normalverteilt sind mit

$$\gamma_{0i} \sim \mathrm{N}(0, \tau_0^2), \quad \gamma_{1i} \sim \mathrm{N}(0, \tau_1^2). \tag{2.13}$$

Die Erwartungswerte können dabei ohne Einschränkung gleich Null gesetzt werden, da die Populationsmittelwerte bereits in den fixen Effekten $\boldsymbol{\beta}$ enthalten sind.

Lineare gemischte Modelle für Longitudinal- und Clusterdaten

Daten

Für $i = 1, \ldots, m$ Individuen bzw. Cluster werden jeweils n_i zeitlich bzw. pro Cluster wiederholte Daten

$$(y_{ij}, x_{ij1}, \ldots, x_{ijk}), \qquad j = 1, \ldots, n_i,$$

für eine metrische Zielvariable y und metrische oder binär kodierte Kovariablen $x_1, \ldots, x_k$ erhoben.

Modell

Für ein lineares gemischtes Modell wird

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_k x_{ijk} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \ldots + \gamma_{li} u_{ijl} + \varepsilon_{ij},$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, angenommen. Dabei sind $\beta_0, \ldots, \beta_k$ feste *Populationseffekte* und $\gamma_{0i}, \gamma_{1i}, \ldots, \gamma_{li}$ *individuen- bzw. clusterspezifische Effekte*. Die zufälligen Effekte werden als unabhängig und identisch normalverteilt vorausgesetzt.

Für die Messfehler ε_{ij} nehmen wir in diesem Beispiel an, dass sie wie im klassischen linearen Modell unabhängig und identisch normalverteilt sind, d.h.

$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2). \tag{2.14}$$

Da das Modell (2.12) neben den festen Effekten des linearen Regressionsmodells (2.11) auch die zufälligen Effekte γ_{0i}, γ_{1i}, $i = 1, \ldots, 50$, enthält, spricht man von einem *linearen gemischten Modell* oder einem *Regressionsmodell mit zufälligen Effekten*.

△

Die Kovariablen $x_{ij1}, \ldots, x_{ijk}$ dürfen bei Longitudinaldaten zeitlich variieren (wie das transformierte Alter), können aber auch zeitkonstant sein (wie die Indikatorvariablen N_i, H_i und C_i). Für Clusterdaten bedeutet dies entsprechend, dass in Cluster i die Kovariablen von Objekt j abhängen oder auch nur clusterspezifische Information enthalten können.

In allgemeiner Notation lassen sich *lineare gemischte Modelle für Longitudinal- und Clusterdaten* für Beobachtungen zu den Zeitpunkten $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$ für Individuum i bzw. für Objekte $j = 1, \ldots, n_i$ im Cluster i in der Form

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_k x_{ijk} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \ldots + \gamma_{li} u_{ijl} + \varepsilon_{ij},$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, schreiben. Die festen Parameter $\beta_0, \ldots, \beta_k$ messen dabei Populationseffekte, während die zufälligen Parameter $\gamma_{0i}, \gamma_{1i}, \ldots, \gamma_{li}$ individuen- bzw. clusterspezifische Effekte beschreiben. Die zusätzlichen Designvariablen $u_{ij1}, \ldots, u_{ijl}$ bestehen oft aus einem Teil der Kovariablen $x_{ij1}, \ldots, x_{ijk}$, wie t_{ij} in Beispiel 2.9.

Für die Fehlervariablen werden im Standardfall die gleichen Annahmen getroffen wie in linearen Regressionsmodellen, d.h. dass die ε_{ij} unabhängig und identisch (normal-)

Abb. 2.11. *Links: Kerndichteschätzer (durchgezogene Linie) und Normalverteilungsdichte (gestrichelt). Rechts: Normal-Quantil-Diagramm.*

verteilt sind mit $\mathrm{E}(\varepsilon_{ij}) = 0$ und $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$. Es ist aber auch möglich, Korrelationen zwischen den Fehlern ε_{ij}, $j = 1, \ldots, n_i$, wiederholter Beobachtungen für ein Individuum bzw. Cluster zu modellieren, vergleiche Kapitel 6. Für die zufälligen Effekte wird oft wie in Beispiel 2.9 angenommen, dass sie unabhängig und identisch normalverteilt sind, jedoch sind auch dafür allgemeinere Modelle möglich, um Korrelationen zu berücksichtigen.

Analysen mit gemischten Modellen für Longitudinaldaten besitzen folgende Vorteile:

- Die Berücksichtigung individuenspezifischer Information kann im Vergleich zur Schätzung eines einfachen linearen Modells zu einer verbesserten Schätzgenauigkeit, d.h. verringerten Varianzen führen.
- Individuenspezifische Effekte können als Surrogat für die Effekte von Kovariablen dienen, die in den vorliegenden Daten nicht oder nur unzureichend gemessen wurden. Man spricht in diesem Fall vom Vorliegen *unbeobachteter Heterogenität*, da die Beobachtungen sich bezüglich der unbeobachteten Kovariablen unterscheiden.
- Die geschätzten individuellen Verlaufskurven erlauben auch individuelle Prognosen, die in einem herkömmlichen Regressionsmodell nicht möglich sind.

Die Schätzung der festen Effekte, der zufälligen Effekte sowie der Varianzparameter der Fehler und der zufälligen Effekte erfolgt mit Ansätzen der Likelihood- und Bayes-Inferenz, vergleiche Kapitel 6.

Beispiel 2.10 Hormontherapie bei Ratten

Wir verwenden zunächst das Modell (2.12), das sowohl individuenspezifische Abweichungen γ_{0i} von der Populationskonstanten β_0 als auch individuenspezifische Steigungsparameter γ_{1i} enthält. Wir schätzen die fixen Effekte, die Varianzparameter σ^2, τ_0^2, τ_1^2 und auch die zufälligen Effekte. Tabelle 2.4 enthält die Schätzwerte für die fixen Effekte und die Varianzparameter. Da der Schätzwert $\hat{\tau}_1^2$ für $\mathrm{Var}(\gamma_{1i})$ sehr klein ist, schätzen wir auch ein vereinfachtes Modell, das keine individuenspezifischen Terme $\gamma_{1i} t_{ij}$ enthält. Die Ergebnisse sind ebenfalls in Tabelle 2.4 zu finden. Es zeigt sich, dass die Schätzungen sehr ähnlich sind.

Für das vereinfachte Modell zeigt Abbildung 2.11 den für die Schätzwerte $\hat{\gamma}_{0i}$, $i = 1, \ldots, 50$ berechneten Kerndichteschätzer und den Normal-Quantil-Plot. Die Abweichungen von der angenommenen Normalverteilung sind nicht gravierend.

△

	Parameter	Modell (2.12) Schätzwert	Vereinfachtes Modell Schätzwert
Konstante	β_0	68.607	68.607
Niedrigdosis	β_1	7.505	7.507
Hochdosis	β_2	6.874	6.871
Kontrolle	β_3	7.313	7.314
$\mathrm{Var}(\gamma_{0i})$	τ_0^2	3.739	3.565
$\mathrm{Var}(\gamma_{1i})$	τ_1^2	<0.001	
$\mathrm{Var}(\varepsilon_{ij})$	σ^2	1.481	1.445

Tabelle 2.4. *Schätzergebnisse für das gemischte Modell (2.12) und das vereinfachte Modell ohne individuenspezifische Steigungsparameter.*

In Kapitel 6 präsentieren wir einige Erweiterungen von linearen gemischten Modellen. Dazu gehören insbesondere gemischte Modelle für binäre und diskrete Zielvariablen. Diese allgemeinere Modellklasse kann auch als Grundlage zur Inferenz für die in Kapitel 7 dargestellten nicht- und semiparametrischen Regressionsmodelle verwendet werden.

2.5 Einfache nichtparametrische Regression

Abbildung 2.2 (Seite 23) zeigt das Streudiagramm der beiden Variablen *mieteqm* und *flaeche* für die in Abschnitt 2.1 betrachteten Mietspiegel-Daten. Aus dem Streudiagramm ist ersichtlich, dass der Effekt der Wohnfläche auf die Nettomiete pro qm nichtlinear ist. In Beispiel 2.2 hatten wir daher den Effekt der Wohnfläche nichtlinear modelliert durch

$$f(\mathit{flaeche}) = \beta_0 + \beta_1/\mathit{flaeche}. \tag{2.15}$$

Abbildung 1.8 (Seite 17) zeigt das Streudiagramm zwischen Z-Score und Alter des Kindes für die Daten zur Unterernährung in Sambia (vergleiche Beispiel 1.2 auf Seite 5). Offensichtlich hängt der Z-Score nichtlinear vom Alter der Kinder ab.

Die meisten der in Kapitel 1 beschriebenen Problemstellungen aus verschiedenen Anwendungsbereichen enthalten ebenfalls nichtlineare Effekte, die oft schwierig durch parametrische *ad hoc*-Ansätze wie in (2.15) zu modellieren sind. Auch für die in Beispiel 2.2 betrachteten Mietspiegeldaten sind andere nichtlineare Transformationen vorstellbar z.B. $f(\mathit{flaeche}) = \beta_0 + \beta_1 \log(\mathit{flaeche})$ oder $f(\mathit{flaeche}) = \beta_0 + \beta_1(\mathit{flaeche})^{\frac{1}{2}}$. In komplexeren Anwendungen mit mehreren stetigen bzw. metrischen Kovariablen wird die Suche nach geeigneten Transformationen selbst mit großer Erfahrung extrem aufwändig bis undurchführbar.

Nicht- und semiparametrische Regressionsmodelle ermöglichen die Schätzung nichtlinearer Effekte in flexibler Weise, ohne zu restriktive Annahmen über eine bestimmte parametrische funktionale Form zu benötigen. Für den Fall *einer* stetigen Kovariablen x ist das *Standardmodell der nichtparametrischen Regression* durch

$$y_i = f(x_i) + \varepsilon_i \tag{2.16}$$

gegeben, wobei für die Fehlervariablen ε_i die gleichen Annahmen wie für das Modell der linearen Einfachregression (2.2) getroffen werden.

Standardmodell der nichtparametrischen Regression

Daten

(y_i, x_i), $i = 1, \ldots, n$, zu metrischen Variablen y und x.

Modell

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Für die Funktion f wird keine einfache parametrische Form angenommen, sondern lediglich bestimmte Glattheitseigenschaften wie Stetigkeit oder Differenzierbarkeit gefordert. Für die Fehler ε_i gelten die gleichen Annahmen wie bei der linearen Einfachregression.

Die Funktion f ist also eine „glatte" Funktion, die nicht durch eine starre parametrische Form a priori zu spezifizieren ist, sondern „nichtparametrisch" mittels der Daten (y_i,x_i) geschätzt werden soll. Kapitel 7 beschreibt verschiedene Techniken zur Schätzung der unbekannten Funktion f.

Ein leicht nachvollziehbares Schätzkonzept ist in den Abbildungen 2.12 und 2.13 veranschaulicht. Zur Illustration haben wir aus dem Datensatz zur Unterernährung in Sambia die Beobachtungen für einen speziellen Distrikt herausgegriffen (Abbildung 2.12 a). Ziel ist die Bestimmung von Schätzungen $\hat{f}(k_alter)$ für den Zusammenhang zwischen Z-Score und Alter des Kindes. Anhand von Abbildung 2.12 a) ist relativ klar, dass eine einfache Regressionsgerade keine zufriedenstellenden Schätzungen liefert. Wir stellen aber fest, dass *lokal*, d.h. wenn nur ein Ausschnitt der Daten betrachtet wird, die Annahme eines linearen Zusammenhangs gerechtfertigt ist, vergleiche die Abbildungen 2.12 b) und c). Basierend auf diesen Beobachtungen erhalten wir das folgende, unter dem Namen *Nächste-Nachbarn-Schätzung* bekannte Verfahren:

1. Bestimme aus dem Definitionsbereich von k_alter eine Menge von Werten
$$k_alter_1 < k_alter_2 < \ldots < k_alter_m,$$
für die Funktionsschätzungen $\hat{f}(k_alter_j)$, $j = 1, \ldots, m$, berechnet werden sollen.
2. Verwende zur Schätzung von f an den Stellen k_alter_j jeweils eine vorher festgelegte Zahl von Beobachtungen in einer *Nachbarschaft* von k_alter_j. In Abbildung 2.12 b) und c) wurden jeweils die nächsten 70 Beobachtungen rechts und links von $k_alter = 11$ (Grafik b) und $k_alter = 28$ (Grafik c) verwendet.
3. Bestimme eine lokale Regressionsgerade basierend auf den in Schritt 2 ausgewählten Beobachtungen und erhalte als Schätzung $\hat{f}(k_alter_j) = \hat{\beta}_0 + \hat{\beta}_1 k_alter_j$. Man beachte, dass für jeden Wert k_alter_j eine separate Regressionsgerade geschätzt wird, d.h. die Regressionskoeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$ variieren mit dem Alter.
4. Verbinde die erhaltenen Schätzungen $\hat{f}(k_alter_1), \ldots, \hat{f}(k_alter_m)$ und visualisiere die geschätzte Kurve.

Eine Illustration der Vorgehensweise findet man in Abbildung 2.13.

Abbildung 2.12 legt noch ein weiteres Verfahren nahe. Anstatt eine *globale* Regressionsgerade zu schätzen, könnte der Definitionsbereich der Alters in mehrere, sich nicht

Abb. 2.12. *Illustration globale und lokale Regression. Abbildung a) zeigt eine (globale) Regressionsgerade zur Schätzung des Zusammenhangs zwischen Z-Score und Alter des Kindes. Die Abbildung b) und c) zeigen lokale Regressionsgeraden basierend auf einem Ausschnitt der Daten.*

überlappende Intervalle zerlegt werden und anschließend in jedem Intervall eine separate Regressionsgerade an die Daten angepasst werden. Diese Vorgehensweise ist in Abbildung 2.14 a) illustriert. Dort wurde der Wertebereich in die drei Intervalle $[0, 19)$, $[19, 39)$ und $[39, 59]$ zerlegt. Im Gegensatz zur globalen Regressionsgerade erhalten wir eine zufriedenstellende Anpassung an die Daten. Es gibt jedoch einen Schönheitsfehler: An den Intervallgrenzen entstehen Sprungstellen. Eine naheliegende Zusatzforderung ist, dass die geschätzte Funktion insgesamt stetig ist, d.h. an den Intervallgrenzen gehen die Regressionsgeraden stetig ineinander über. Wird diese Zusatzbedingung berücksichtigt, erhalten wir die Schätzung in Abbildung 2.14 b). Der geschätzte Polygonzug kann als Spezialfall sogenannter *Polynom-Splines* angesehen werden. Splines sind stückweise Polynome, die an den Intervallgrenzen (Knoten genannt) gewisse Glattheitsbedingungen erfüllen. Die flexible Regression basierend auf Splines wird eine Hauptrolle in den Kapiteln 7 und 8 spielen. Dort werden wir uns ausführlich mit den wichtigsten Fragen und Problemen im Zusammenhang mit der Spline-Regression auseinandersetzen, etwa wie Splines dargestellt werden können oder wie viele Knoten gesetzt werden sollen. Als Vorgeschmack zeigt Abbildung 2.14 c) eine Schätzung basierend auf sogenannten *P-Splines*.

Nach erfolgter flexibler Schätzung kann man in einem zweiten Arbeitsschritt versuchen, einfache parametrische funktionale Formen zu finden, die den nichtparametrischen Funktionen gut angepasst sind, ohne wesentliche Charakteristika zu unterdrücken. In diesem

Abb. 2.13. *Veranschaulichung einer nichtparametrischen Nächste-Nachbarn-Schätzung für den Zusammenhang zwischen Z-Score und Alter des Kindes.*

Sinn lässt sich die nicht- und semiparametrische Regression auch als Werkzeug der explorativen Datenanalyse auffassen. Abbildung 2.14 d) zeigt einen Vergleich der nichtparametrischen Schätzung mit parametrisch angepassten Funktionen für die Modelle

$$zscore = \beta_0 + \beta_1 \log(alter + 1) + \varepsilon$$

und

$$zscore = \beta_0 + \beta_1/(alter + 1) + \varepsilon.$$

Trotz ähnlicher Kurvenverläufe ist die Anpassung der nichtparametrisch geschätzten Kurve vor allem im Bereich $0 - 20$ Monate besser.

Abb. 2.14. *Stückweise lineare Regression (Grafik a), linearer Spline (Grafik b) und kubischer P-Spline (Grafik c) zur Schätzung des Zusammenhangs zwischen Z-Score und Alter des Kindes, sowie Vergleich zwischen parametrischer und nichtparametrischer Regression (Grafik d).*

2.6 Additive Regression

In den meisten Anwendungen, wie in den Beispielen *Mietspiegel* und *Unterernährung in Sambia*, liegen mehrere oder sogar viele Kovariablen vor, die entweder stetig oder kategorial sind.

Beispiel 2.11 Unterernährung in Sambia

Die metrischen Kovariablen sind *k_alter* (Alter des Kindes), *k_still* (Stilldauer), *m_bmi* (Body Mass Index der Mutter), *m_groesse* (Größe der Mutter) und *m_alterg* (Alter der Mutter bei der Geburt). Wie im linearen Regressionsmodell werden die kategorialen Regressoren *m_bildung* (Bildung der Mutter), *m_arbeit* (Erwerbsstatus der Mutter) und *region* (Wohnort) dummy-kodiert. Für den Ausbildungsstatus wird die Kategorie 2 = „Grundschule“ als Referenzkategorie gewählt. Die Dummy-Variablen *m_bildung1*, *m_bildung3* und *m_bildung4* entsprechen den Bildungsniveaus „keine Ausbildung“, „Volksschule“ und „höherer Abschluss“. Beim Wohnort verwenden wir die Region Copperbelt (*region* = 2) als Referenzkategorie. Als Dummy-Variablen für die restlichen Regionen fungieren die Variablen *region1*, *region3*, ..., *region9*.

Da der Effekt der stetigen Kovariablen auf die Zielvariable *zscore* möglicherweise nichtlinear ist, wird statt eines linearen Regressionsmodells ein *additives Modell*

Abb. 2.15. *Geschätzte nichtlineare Funktionen inklusive 80% und 95% punktweisen Konfidenzintervallen. Die Punkte im unteren Teil der Grafiken deuten die Verteilung der Kovariablenwerte an.*

$$\begin{aligned} zscore = {} & f_1(k_alter) + f_2(m_bmi) + f_3(m_alterg) + f_4(m_groesse) \\ & + \beta_0 + \beta_1\, m_bildung1 + \ldots + \beta_{12}\, region9 + \varepsilon \end{aligned} \tag{2.17}$$

zugrunde gelegt. Die Stilldauer (Variable *k_still*) nehmen wir vorläufig nicht ins Modell auf, da diese hoch korreliert mit dem Alter des Kindes ist. Vergleiche hierzu auch die Fallstudie in Kapitel 8.6. Die Konstante β_0 und die Koeffizienten $\beta_1, \beta_2, \ldots$ der kategorialen Kovariablen *m_bildung*, *m_arbeit* und *region* sind wie im linearen Regressionsmodell zu interpretieren.

Die Funktionen f_1, f_2, f_3, f_4 bleiben wie die Funktion f im Basismodell (2.16) der nichtparametrischen Regression auf Seite 40 unspezifiziert und werden nichtparametrisch zusammen mit den Regressionskoeffizienten $\beta_0, \beta_1, \ldots$ der kategorialen Kovariablen geschätzt. Das Modell ist wegen der Effekte $f_1, \ldots, f_4$ nicht mehr linear, aber weiterhin additiv. Da es keine Interaktionen zwischen Kovariablen enthält, handelt es sich um ein *additives Haupteffekt-Modell*.

Bei additiven Modellen tritt folgendes *Identifikationsproblem* auf: Geht man durch Addition einer Konstanten $c \neq 0$ von $f_1\,(k_alter)$ über zu $\tilde{f}_1\,(k_alter) = f_1\,(k_alter) + c$ und gleichzeitig durch Subtraktion von c zu $\tilde{\beta}_0 = \beta_0 - c$, so ändert dies nichts am Wert der rechten Seite von (2.17). Das Niveau der nichtlinearen Funktionen ist daher nicht identifizierbar und muss durch Zusatzannahmen fixiert werden. Dies geschieht z.B. indem man sie „um Null zentriert" und verlangt, dass

$$\sum_{i=1}^{n} f_1\left(k_alter_i\right) = \ldots = \sum_{i=1}^{n} f_4\left(m_groesse_i\right) = 0$$

gilt. In Abbildung 2.15 sind die geschätzten Funktionen auf diese Weise zentriert.

Die Interpretation der geschätzten Effekte geschieht am einfachsten durch Visualisierung, vergleiche Abbildung 2.15. Beispielsweise lässt sich der Effekt des Alters des Kindes wie folgt interpretieren: Der durchschnittliche Z-Score nimmt mit wachsendem Alter zunächst annähernd linear ab. Anschließend stabilisiert sich der durchschnittliche Ernährungszustand für Kinder, die älter als 18 Monate sind. Ab 3 Jahren könnte sogar eine leichte Verbesserung eintreten. Auf die Interpretation der restlichen Effekte gehen wir im Rahmen einer detaillierten Fallstudie in Kapitel 8.6 ein.

△

Die allgemeine Form eines *additiven Modells* (ohne Interaktionen) ist

$$y_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \tag{2.18}$$

wobei für die Fehlervariablen die gleichen Annahmen wie im linearen Regressionsmodell gelten. Die unspezifizierten glatten Funktionen $f_1, \ldots, f_q$ stellen die (Haupt-)Effekte der metrischen Kovariablen $z_1, \ldots, z_q$ dar und werden mit nichtparametrischen Techniken geschätzt, siehe Kapitel 8. Die Kovariablen $x_1, \ldots, x_k$ sind kategorial oder auch metrisch mit üblichen linearen Effekten.

Additive Haupteffekt-Modelle der Form (2.18) können durch Interaktionsterme erweitert werden, um Wechselwirkungen zwischen Kovariablen zu berücksichtigen. In der Regel werden dabei nur paarweise Interaktionen betrachtet. Für zwei metrische Kovariablen z_1 und z_2 geschieht dies durch additives Einbeziehen einer glatten, zweidimensionalen Funktion $f_{1,2}(z_1, z_2)$. Damit wird der Prädiktor (2.18) zu

$$\eta_i = f_1(z_{i1}) + f_2(z_{i2}) + f_{1,2}(z_{i1}, z_{i2}) + \ldots$$

erweitert. Der Interaktionseffekt $f_{1,2}$ modifiziert somit die Haupteffekte f_1 und f_2 der beiden Kovariablen. Die Schätzung der glatten Oberfläche $f_{1,2}$ erfolgt durch Erweiterung der nichtparametrischen Techniken für eindimensionale Funktionen auf den bivariaten Fall, siehe Kapitel 7.2.

Eine paarweise Interaktion zwischen einer metrischen Kovariablen z_1 und einer binären Kovariablen x_1 wird durch additive Ergänzung des Prädiktors zu

$$\eta_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + f_x(z_{i1})\, x_{i1} + \ldots$$

modelliert. Der Interaktionsterm $f_x(z_1)\, x_1$, mit einer glatten Funktion f_x, kann auch als über z variierender Effekt von x interpretiert werden. Modelle mit parametrisch modellierten Interaktionen werden ausführlicher in Abschnitt 3.1.4 dargestellt. Nichtparametrisch modellierte Interaktionen sind Gegenstand von Abschnitt 8.3.

Die Schätzung additiver Modelle erfolgt iterativ, wobei für die Schätzung des linearen Teils $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ die KQ-Methode und für die Schätzung der Funktionen $f_1, \ldots, f_q$ nichtparametrische Methoden, z.B. basierend auf Splines, verwendet werden können.

Standardmodell der additiven Regression

Daten

$(y_i, z_{i1}, \ldots, z_{iq}, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, zu y und $x_1, \ldots, x_k$ wie in der linearen Regression und zu stetigen Kovariablen $z_1, \ldots, z_q$.

Modell

Es gilt

$$y_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i.$$

Für die Fehlervariablen ε_i werden die gleichen Annahmen getroffen wie im klassischen linearen Regressionsmodell. Die Funktionen $f_1(z_1)$, ..., $f_q(z_q)$ werden als „glatt“ vorausgesetzt und beschreiben nichtlineare Effekte der stetigen oder metrischen Kovariablen $z_1, \ldots, z_q$.

2.7 Generalisierte additive Regression

Nichtlineare Effekte von metrischen Kovariablen treten in analoger Weise auch bei Regressionsanalysen für binäre und andere nicht normalverteilte Zielvariablen auf. Wie für additive Modelle in Abschnitt 2.6 ist es oft wünschenswert, solche nichtlinearen Effekte nicht von vornherein durch die Annahme einer parametrischen funktionalen Form unnötig einzuschränken, sondern in nichtparametrischer Weise zu modellieren und zu analysieren.

Beispiel 2.12 Schadenshäufigkeiten bei Kfz-Versicherungen

Wir illustrieren die Verwendung *Generalisierter Additiver Modelle* durch die Analyse von Kfz-Versicherungsdaten für Belgien aus dem Jahr 1997. Die Kalkulation der Versicherungsprämie bei Kfz-Versicherungen basiert auf detaillierten statistischen Analysen der Risikostruktur der Versicherten. Ein wichtiger Schritt ist die Modellierung der *Schadenshäufigkeit*, die in der Regel abhängig ist von den Eigenschaften der Versicherten und des Fahrzeugs. Typische beeinflussende Charakteristika sind das Alter des Versicherungsnehmers (*alterv*), das Alter des Fahrzeugs (*alterkfz*), die Motorleistung gemessen in Pferdestärken (*ps*) und der bisherige Schadensverlauf. In Belgien wird der bisherige Schadensverlauf anhand eines 23-stufigen Bonus-Malus-Scores (*bm*) gemessen. Je höher der Score, desto schlechter die Vorgeschichte des Versicherten. Die statistische Analyse basiert auf Regressionsmodellen mit der Schadenshäufigkeit innerhalb eines Jahres als Zielgröße. Da die Schadenshäufigkeit nur die diskreten Werte $0, 1, 2, \ldots$ annehmen kann, sind Regressionsmodelle für metrische Zielgrößen nicht geeignet.

△

Eine häufig angemessene Verteilung für Zähldaten ist die Poisson-Verteilung. Wir können also annehmen, dass $y \sim \mathrm{Po}(\lambda)$ gilt, mit erwarteter Häufigkeit $\lambda = \mathrm{E}(y)$. Ziel ist die Modellierung der erwarteten Häufigkeit λ in Abhängigkeit von den Kovariablen. Die naheliegende Modellierung $\lambda = \eta$ durch einen linearen oder additiven Prädiktor η ist, ähnlich wie bei binären Zielgrößen, problematisch, da nicht sichergestellt ist, dass die geschätzten erwarteten Häufigkeiten $\hat{\lambda}$ positiv sind. Die Positivität lässt sich z.B. durch die Annahme

Additives Poisson-Modell

Allgemein ist ein additives Poisson-Modell $y_i \sim \mathrm{Po}(\lambda_i)$ durch

$$\lambda_i = \mathrm{E}(y_i) = \exp(\eta_i)$$

mit einem *additiven Prädiktor*

$$\eta_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + x_{ik}$$

gegeben. Additive Poisson-Modelle sind ein Spezialfall generalisierter additiver Modelle für nicht-normalverteilte Zielvariablen (Kapitel 8).

$\lambda = \exp(\eta)$ gewährleisten. Bei Verwendung eines linearen Prädiktors erhalten wir dann ein multiplikatives Modell

$$\lambda = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot \ldots \cdot \exp(\beta_k x_k)$$

für die erwartete Häufigkeit mit ähnlicher Interpretation wie im Logit-Modell. Eine Erhöhung beispielsweise der ersten Kovariable x_1 um eine Einheit verändert die erwartete Häufigkeit um den Faktor $\exp(\beta_1)$. Für einen additiven Prädiktor erhalten wir

$$\lambda = \exp(f_1(z_1) + \ldots + f_q(z_q) + \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k).$$

Je nach Verlauf der nichtlinearen Funktionen steigt oder fällt die erwartete Häufigkeit, wenn eine der Kovariablen erhöht wird. Im Gegensatz zum rein linearen Prädiktor hängt die Veränderung (wegen der Nichtlinearität) auch vom Vergleichswert ab. In der Regel wird eine Erhöhung von x_1 beispielsweise von 20 auf 21 eine andere Veränderung der erwarteten Häufigkeit bewirken als eine Erhöhung von 30 auf 31.

Beispiel 2.13 Schadenshäufigkeiten bei Kfz-Versicherung

Wir modellieren die Schadenshäufigkeiten der belgischen Versicherungsdaten durch den additiven Prädiktor

$$\eta_i = f_1(alterv_i) + f_2(alterkfz_i) + f_3(ps_i) + f_4(bm_i) + \beta_0 + \beta_1 geschl_i + \ldots$$

mit möglicherweise nichtlinearen Funktionen $f_1, \ldots, f_4$ der Variablen *alterv*, *alterkfz*, *ps* und *bm*. Die Punkte deuten an, dass neben den metrischen Variablen weitere kategoriale Variablen im Prädiktor enthalten sind. Exemplarisch haben wir das Geschlecht (*geschl*) aufgeführt. Die nichtlinearen Funktionen und die Regressionskoeffizienten können mit Methoden geschätzt werden, die wir ausführlich in Kapitel 8 darstellen. Abbildung 2.16 zeigt die damit erzielten Schätzungen $\hat{f}_1, \ldots, \hat{f}_4$ für die Versicherungsdaten. Besonders auffällig ist die hochgradig nichtlineare Funktion des Alters des Versicherungsnehmers. Zunächst sinkt die erwartete Häufigkeit annähernd linear bis zu einem Alter von ca. 40 Jahren. Anschließend bleibt die durchschnittliche Häufigkeit für einige Jahre annähernd konstant, um dann wieder kontinuierlich abzusinken. Für ältere Versicherungsnehmer steigt die durchschnittliche Schadenshäufigkeit wieder an. Insbesondere für sehr alte Versicherungsnehmer liegen nur noch wenige Daten vor, so dass die Schätzungen in diesem Bereich mit Vorsicht zu genießen sind. Dies spiegelt sich auch in dem sehr weiten Konfidenzintervall in diesem Bereich wieder.

△

Abb. 2.16. *Kfz-Versicherungsdaten: Geschätzte nichtlineare Funktionen inklusive 80% und 95% punktweisen Konfidenzintervallen. Die Punkte im unteren Teil der Grafik deuten die Verteilung der Kovariablenwerte an.*

2.8 Geoadditive Regression

In vielen Anwendungen ist für die Individuen oder Untersuchungseinheiten $i = 1, \ldots, n$ neben den Werten $(y_i, x_{i1}, \ldots, x_{ik}, z_{i1}, \ldots, z_{iq})$ von Ziel- und Kovariablen zusätzlich kleinräumige geografische Information vorhanden, z.B. durch den zugehörigen genauen Standort bzw. die Adresse, die Postleitzahl, den Wohnort oder den Landkreis. Für die Beispiele aus Kapitel 1 trifft dies auf die Daten zum Mietspiegel, zur Unterernährung in Sambia, zur Kfz-Versicherung und zum Waldzustand zu. In diesen Anwendungen ist es von inhaltlichem Interesse, räumliche Effekte, die nicht durch übliche Kovariablen erfasst werden, in geeigneter Form in Regressionsmodelle einzubeziehen und zu analysieren.

Beispiel 2.14 Unterernährung in Sambia – Geoadditives Modell

In Beispiel 2.11 (Seite 44) wurden geografische Effekte auf die Unterernährung in relativ grober Form durch die Effekte der Regionen in Sambia berücksichtigt. Dazu wurden die Regionen durch Dummy-Variablen binär kodiert und dann die entsprechenden regionenspezifischen Effekte genauso geschätzt wie übliche „feste" Effekte von kategorialen Kovariablen. Diese konventionelle Vorgehensweise hat zwei entscheidende Nachteile. Erstens: Die räumlichen Effekte werden separat modelliert und geschätzt, ohne dass vorhandene Information über die räumliche Nähe oder Nachbarschaft von Regionen benutzt wird. Zweitens: Möchte man die genauere kleinräumige Information nutzen, in welchem

Abb. 2.17. *Räumlicher Effekt für die Sambia-Daten.*

Distrikt von Sambia der Wohnort der Mutter liegt, wird die analoge distriktspezifische Vorgehensweise schwierig bis unmöglich. Da für jeden Distrikt ein separater, fester Effekt der entsprechenden Dummy-Variablen in den Prädiktor einbezogen wird, enthält das Modell eine sehr große Anzahl von Parametern, was eine hohe Schätzungenauigkeit verursacht. Es ist deshalb günstiger, den räumlichen Effekt der Variable *district* als unspezifizierte Funktion $f_{geo}(district)$ aufzufassen und bei der Modellierung und Schätzung von f_{geo} die geografische Nähe oder Nachbarschaft von Distrikten geeignet zu nutzen. Konzeptionell ähnelt dies der nichtparametrischen Schätzung einer glatten Funktion f einer metrischen Kovariablen, wie z.B. $f(alter)$. Wir reanalysieren somit die Daten mit folgendem *geoadditiven* Modell:

$$\begin{aligned} zscore = {} & f_1(k_alter) + f_2(m_bmi) + f_3(m_alterg) + f_4(m_groesse) + \\ & f_{geo}(district) + \beta_0 + \beta_1\, m_bildung1 + \ldots + \beta_4\, m_arbeit + \varepsilon \end{aligned}$$

Im Vergleich zu Beispiel 2.11 enthält der lineare Teil des Prädiktors nun keine regionenspezifischen Dummy-Variablen mehr. Abbildung 2.17 zeigt die in Distrikte unterteilte Karte von Sambia mit den entsprechend gefärbten Effekten. Der räumliche Effekt lässt sich dann analog zu nichtlinearen Effekten interpretieren. Beispielsweise bedeutet ein Effekt von 40 einen im Mittel um 40 Punkte erhöhten Z-Score im Vergleich zu einer Region mit Effekt bei Null.

Offensichtlich zeigt das distriktspezifische Muster, dass räumliche Effekte wenig mit den administrativen Grenzen der Regionen zu tun haben, sondern wohl auf andere Ursachen zurückzuführen sind. Im vorliegenden Fall deutet das ausgeprägte Nord-Süd Gefälle mit deutlich besserer Ernährungssituation im Süden unter anderem auf klimatische Ursachen hin. Im Süden des Landes herrschen nämlich schlechtere klimatische Bedingungen, da diese Landesteile eine deutlich niedrigere Höhenlage aufweisen als die Nördlichen. In diesem Sinn können geoadditive Modelle als exploratives Werkzeug der Datenanalyse angesehen werden: Gefundene geografische Muster helfen bei der Suche nach räumlichen Variablen, welche die räumliche Variation verursachen.

△

Allgemein liegt eine Problemstellung der *geoadditiven Regression* vor, wenn zusätzlich zu den Werten der Zielvariablen und metrischen bzw. kategorialen Kovariablen noch zu jeder Einheit i der Wert s_i einer *Lokationsvariablen* s vorliegt. Diese Lokationsvariable s kann

Geoadditive Modelle

Daten

Neben der stetigen Zielvariablen y, den stetigen Kovariablen $z_1, \ldots, z_q$ und den üblichen Kovariablen $x_1, \ldots, x_k$ liegen noch Beobachtungen zur räumlichen Lokation s vor.

Modell

$$y_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + f_{geo}(s_i) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i,$$

wobei für die Fehlervariablen ε_i die gleichen Annahmen wie für das klassische lineare Regressionsmodell getroffen werden. Die unbekannten glatten Funktionen $f_1, \ldots, f_q, f_{geo}$ und die parametrischen Effekte sind dabei mit Hilfe der Daten zu schätzen.

wie in Beispiel 2.14 ein *Lokationsindex* mit endlichem Definitionsbereich $s \in \{1, \ldots, S\}$ sein, der z.B. die Landkreise, Distrikte, usw. eines Landes indiziert. Zugleich ist, in Form einer Karte oder eines Grafen, die gesamte Nachbarschaftsinformation vorhanden. In anderen Anwendungen, wie etwa bei den Daten zum Waldzustand, ist s eine stetige Variable, die den Standort bzw. die Lokation z.B. in üblichen räumlichen Koordinaten (möglichst) genau angibt.

Für die flexible Modellierung und Schätzung der Funktion f_{geo} stehen verschiedene Möglichkeiten zur Verfügung, die sich unter anderem dadurch unterscheiden ob sie besser für eine diskrete oder stetige Lokationsvariable geeignet sind, siehe Kapitel 7.

Geoadditive Regressionsanalysen können auch für nicht-normalverteilte, insbesondere binäre, kategoriale oder diskrete Zielvariablen durchgeführt werden, wie etwa bei der Analyse des Waldzustands oder der Schadenshäufigkeiten in der Kfz-Versicherung. Dazu wird der Prädiktor η_i in additiven Logit- oder Poisson-Modellen bzw. – allgemeiner – in generalisierten additiven Modellen zum sogenannten *geoadditiven Prädiktor*

$$\eta_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + f_{geo}(s_i) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

erweitert, eventuell auch mit zusätzlichen Interaktionstermen.

Beispiel 2.15 Schadenshäufigkeiten bei Kfz-Versicherung

Es ist bekannt, dass die Schadensfälle bei Kfz-Versicherungen eine zum Teil erhebliche räumliche Variation aufweisen. In Deutschland werden daher sogenannte Regionalklassen mit regional unterschiedlicher Versicherungsprämie gebildet. Ähnlich sind die Tarifstrukturen in Belgien aufgebaut. Eine realistische Modellierung der Schadenshäufigkeiten erfordert daher die adäquate Berücksichtigung der räumlichen Variation. Dazu erweitern wir den rein additiven Prädiktor aus Beispiel 2.13 zu

$$\eta_i = f_1(alterv_i) + f_2(alterkfz_i) + f_3(ps_i) + f_4(bm_i) + f_{geo}(district_i) + \ldots,$$

wobei $f_{geo}(district)$ ein über die Landkreise in Belgien variierender räumlicher Effekt ist. Abbildung 2.18 zeigt den geschätzten Effekt. Je dunkler eine Region eingefärbt ist,

Abb. 2.18. *Geschätzter räumlicher Effekt für die Kfz-Versicherungsdaten.*

desto stärker ist der geschätzte Effekt. Schraffierte Flächen kennzeichnen Regionen, in denen keine Beobachtungen vorliegen. Ein Effekt von beispielsweise 0.3 bedeutet eine Erhöhung der erwarteten Häufigkeit um den Faktor $\exp(0.3) = 1.35$ im Vergleich zu einer Region mit einem Effekt von Null. Wir erkennen drei klar abgrenzbare Regionen mit höheren erwarteten Schadenshäufigkeiten. Hierbei handelt es sich um die Ballungsgebiete um Brüssel im Zentrum, Antwerpen im Norden und Lüttich im Osten von Belgien. Die dünn besiedelten Gegenden im Südosten weisen dagegen geringere durchschnittliche Häufigkeiten auf.

△

Im folgenden Beispiel betrachten wir eine Anwendung aus der Lebensdaueranalyse.

Beispiel 2.16 Überlebenszeiten von Leukämie-Patienten

Ziel der Anwendung ist die Analyse des Einflusses von Kovariablen auf die Überlebenszeit von Patienten nach der Diagnose eines speziellen Typs von Leukämie. Dabei interessiert insbesondere auch die räumliche Variation der Lebenserwartung, da diese einerseits Hinweise auf bisher unbekannte Risikofaktoren liefern kann und andererseits eine deutliche räumliche Variation auch ein Indiz für eine unterschiedliche Versorgungsqualität sein kann.

In diesem Beispiel wurden die Überlebenszeiten von 1043 Patienten in Nordwest-England untersucht, bei denen zwischen 1982 und 1998 akute myeloische Leukämie diagnostiziert wurde. Die Daten entstammen dem britischen North West Leukemia Register und enthalten neben der Überlebenszeit der Patienten Informationen zu den folgenden Kovariablen: dem Geschlecht des Patienten (Variable *sex*, 1=weiblich, 0=männlich), dem Alter des Patienten zum Zeitpunkt der Diagnose (Variable *alter*), der Anzahl weißer Blutkörperchen zum Zeitpunkt der Diagnose (Variable *wb*) sowie dem Townsend Index (Variable *ti*), einem Index der die Armut im Wohnbezirk der Patienten misst. Größere Werte des Townsend Index entsprechen dabei ärmeren Bezirken. Zusätzlich ist räumliche Information zu den Beobachtungen vorhanden. Zu jedem Patienten sind die genauen Koordinaten (in Längen- und Breitengrad) des Wohnorts innerhalb von

Abb. 2.19. *Verteilung der Beobachtungen in Nordwest-England. Jeder Punkt entspricht einer Beobachtung.*

Nordwest-England bekannt. Zusätzlich ergibt sich durch Aggregation dieser Information eine Zuordnung der Beobachtungen zu den Distrikten Nordwest-Englands. Die räumliche Verteilung der Beobachtungen ist in Abbildung 2.19 dargestellt. 16 Prozent der Überlebenszeiten sind nur unvollständig beobachtet, da die entsprechenden Patienten bis zum Ende der Studie überlebt haben.

Zur Schätzung des Einflusses von Kovariablen auf die Überlebenszeit oder Lebensdauer T_i eines Individuums i werden in der Lebensdaueranalyse sogenannte *Hazardratenmodelle* eingesetzt. Die *Hazardrate* $\lambda_i(t)$ der Überlebenszeitverteilung für ein Individuum i ist definiert als der Grenzwert

$$\lambda_i(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_i \leq t + \Delta t \,|\, T_i \geq t)}{\Delta t}.$$

Damit bezeichnet $\lambda_i(t)$ die bedingte Wahrscheinlichkeit, im Zeitintervall $[t, t + \Delta t]$ zu sterben, unter der Bedingung, dass das Individuum bereits bis zur Zeit t überlebt hat, relativ zur Breite des Intervalls Δt. In unserer Anwendung verwenden wir ein *geoadditives Hazardratenmodell*, das die Hazardrate $\lambda_i(t)$ über die Exponentialfunktion (in ähnlicher Weise wie im Poisson-Modell der Beispiele 2.13 und 2.15) mit einem geoadditiven Prädiktor verknüpft:

$$\lambda_i(t) = \exp\left[g(t) + f_1(alter_i) + f_2(ti_i) + f_{geo}(s_i) + \beta_0 + \beta_1 wb_i + \beta_2 sex\right]$$

Dieses Modell ist eine Erweiterung des populären Cox-Modells mit linearem Prädiktor auf Modelle mit geoadditivem Prädiktor. Es enthält nichtparametrische Effekte der stetigen Kovariablen *alter* und *ti*, lineare Effekte der Kovariablen *wb* und *sex*, sowie einen räumlichen Effekt, der entweder basierend auf den exakten Koordinaten oder basierend auf den Distrikten definiert werden kann (in Abschnitt 7.2 werden wir genauer auf die verschiedenen Modellierungsmöglichkeiten für unterschiedliche Typen räumlicher Effekte eingehen). Darüber hinaus besitzt das Modell aber auch noch eine zeitabhängige Komponente $g(t)$, die beschreibt, wie das Sterberisiko nach dem Zeitpunkt der Diagnose zeitlich variiert. Die Funktion $g(t)$ heißt *Log-Baseline-Hazardrate* und $\lambda_0(t) = \exp[g(t)]$ *Baseline-Hazardrate*.

Abb. 2.20. *Geschätzte räumliche Effekte basierend auf Distrikten (links) beziehungsweise den exakten Koordinaten der Beobachtungen (rechts).*

Abbildung 2.20 zeigt die geschätzten räumlichen Effekte und weist auf eine deutliche räumliche Variation des Mortalitätsrisikos hin. Vermutlich sind die räumlichen Effekte ein Surrogat für unbeobachtete Kovariablen, mit denen zumindest ein Teil der räumlichen Variation erklärbar würde. Abbildung 2.21 zeigt die geschätzten Funktionen $g(t)$, $f_1(alter)$ und $f_2(it)$. Der Verlauf der Log-Baseline-Hazardrate spiegelt ein bis etwa zum achten Jahr nach der Diagnose nichtlinear fallendes Sterberisiko wider, das aber später wieder ansteigt. Der Alterseffekt hat einen monotonen, fast linear ansteigenden Verlauf, während der Effekt des Townsend Index besagt, dass das Sterberisiko mit geringerem Wohlstand wächst und dann auf etwa gleichem Niveau verbleibt.

Der geschätzte Effekt der Anzahl weißer Blutkörperchen wb_i ist mit $\hat{\beta}_1 = 0.003$ positiv, aber scheinbar sehr niedrig. Da die Anzahl wb_i jedoch sehr groß ist, hat der Prädiktor $\hat{\beta}_1 wb_i$ tatsächlich einen relevanten Einfluss. Der geschätzte Geschlechtseffekt ist mit $\hat{\beta}_2 = 0.073$ sehr klein, so dass das Geschlecht kaum Einfluss auf die Hazardrate hat.

Regressionsmodelle zur Analyse von Lebens- oder Verweildauern spielen in vielen Bereichen eine wichtige Rolle, werden aber in diesem Buch nicht im Detail dargestellt. Literaturhinweise finden sich beispielsweise in Abschnitt 4.7, die Methodik zum hier vorgestellten Beispiel wird in Kneib & Fahrmeir (2007) beschrieben.

△

Abb. 2.21. *Geschätzte nichtlineare Kovariableneffekte mit 80% und 95% Konfidenzintervallen.*

2.9 Modelle im Überblick

Wir fassen die Regressionsmodelle dieses Kapitels nochmals in kompakter Form zusammen und geben dabei jeweils mit an, in welchen Kapiteln diese detaillierter behandelt werden. Dabei wird auch nochmals die gemeinsame, allgemeine Struktur der in diesem Buch behandelten Regressionsmodelle erkennbar werden.

2.9.1 Lineare Modelle (LM, Kapitel 3)

- *Zielgröße:* Die Beobachtungen y_i sind metrisch mit

$$y_i = \eta_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

  Die Fehler $\varepsilon_1, \ldots, \varepsilon_n$ sind unabhängig und identisch verteilt (i.i.d.) mit

$$\mathrm{E}\left(\varepsilon_i\right) = 0, \quad \mathrm{Var}\left(\varepsilon_i\right) = \sigma^2.$$

- *Erwartungswert:*

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \eta_i^{lin}.$$

- *Prädiktor:*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \eta_i^{lin}.$$

2.9.2 Logit-Modell (Kapitel 4)

- *Zielgröße:* Die Beobachtungen $y_i \in \{0, 1\}$ sind binär und $\mathrm{B}(1, \pi_i)$ verteilt.
- *Erwartungswert:*
$$\mathrm{E}\left(y_i\right) = P\left(y_i = 1\right) = \pi_i = \frac{\exp\left(\eta_i^{lin}\right)}{1 + \exp\left(\eta_i^{lin}\right)}.$$
- *Prädiktor:*
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \eta_i^{lin}.$$

2.9.3 Poisson-Regression (Kapitel 4)

- *Zielgröße:* Die Beobachtungen $y_i \in \{0, 1, 2, \ldots\}$ sind Zähldaten und geben an, wie häufig ein interessierendes Ereignis in einem bestimmten Zeitintervall auftritt. Im Poisson-Modell nimmt man $y_i \sim Po(\lambda_i)$ an.
- *Erwartungswert:*
$$\mathrm{E}\left(y_i\right) = \lambda_i = \exp\left(\eta_i^{lin}\right).$$
- *Prädiktor:*
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \eta_i^{lin}.$$

2.9.4 Generalisierte lineare Modelle (GLM, Kapitel 4, 5)

- *Zielgröße:* Die Beobachtungen y_i sind metrisch, kategorial oder Zähldaten und je nach Skalenniveau und Verteilungsannahme entweder normal-, binomial-, Poisson- oder gammaverteilt.
- *Erwartungswert:*
$$\mathrm{E}(y_i) = \mu_i = h\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\right) = h\left(\eta_i^{lin}\right),$$
wobei h eine (bekannte) Verknüpfungsfunktion ist, wie zum Beispiel $h(\eta) = \exp(\eta)/(1+\exp(\eta))$ im Logit-Modell.
- *Prädiktor:*
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \eta_i^{lin}.$$
- *Bemerkung:* Generalisierte lineare Modelle sind eine allgemeine Modellklasse, die Lineare Modelle, Logit-Modelle und Poisson-Modelle als Spezialfall enthält. Erweiterungen für kategoriale Zielvariablen werden in Kapitel 5 behandelt.

2.9.5 Lineare gemischte Modelle (LMM, Kapitel 6)

- *Zielgröße:* Die Beobachtungen y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ sind metrisch mit
$$y_{ij} = \eta_{ij} + \varepsilon_{ij}.$$
Es handelt sich um Longitudinal- oder Clusterdaten für m Individuen/Cluster mit n_i Beobachtungen für Individuum/Cluster i. Für die Fehler ε_{ij} trifft man die gleichen Annahmen wie im LM.

- *Erwartungswert:*

$$\begin{aligned}\mathrm{E}(y_{ij}) &= \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_k x_{ijk} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \ldots + \gamma_{li} u_{ijl} \\ &= \eta_{ij}^{lin} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \ldots + \gamma_{li} u_{ijl}.\end{aligned}$$

  Die zufälligen Effekte $\gamma_{0i}, \gamma_{1i}, \ldots$ werden als unabhängig und identisch verteilt betrachtet.
- *Prädiktor:*

$$\eta_{ij} = \eta_{ij}^{lin} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \ldots + \gamma_{li} u_{ijl}.$$

- *Bemerkung:* LMM mit korrelierten zufälligen Effekten und generalisierte lineare gemischte Modelle (GLMM) werden ebenfalls in Kapitel 6 behandelt.

2.9.6 Additive Modelle und Erweiterungen (AM, Kapitel 7, 8)

- *Zielgröße:* Die Beobachtungen y_i sind metrisch mit

$$y_i = \eta_i + \varepsilon_i.$$

  Für die Fehler ε_i werden die gleichen Annahmen wie im LM getroffen.
- *Erwartungswert:*

$$\mathrm{E}(y_i) = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin} = \eta_i^{add}.$$

- *Prädiktor:*

$$\eta_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin} = \eta_i^{add}.$$

- *Bemerkung:* Additive Modelle können um Interaktionen, räumliche Effekte und zufällige Effekte erweitert werden. Für Interaktionen wird der Prädiktor erweitert zu

$$\eta_i = \eta_i^{add} + f_1(z_1, z_2) + \ldots$$

  oder

$$\eta_i = \eta_i^{add} + f(z_1)\, x_1 + \ldots .$$

  Bei Geoadditiven Modellen wird der Prädiktor erweitert zu

$$\eta_i = \eta_i^{add} + f_{geo}(s_i)$$

  mit räumlichem Effekt $f_{geo}(s)$ der Lokationsvariable s. Durch Hinzunahme von zufälligen Effekten ergibt sich der Prädiktor

$$\eta_{ij} = \eta_{ij}^{add} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \ldots + \gamma_{li} u_{ijl}.$$

  Damit wird das Additive Modell zum Additiv Gemischten Modell (AMM).

2.9.7 Generalisierte additive (gemischte) Modelle (GAMM, Kapitel 8)

- *Zielgröße:* Die Beobachtungen y_i sind metrisch, kategorial oder Zähldaten und je nach Skalenniveau und Verteilungsannahme entweder normal-, binomial-, Poisson- oder gammaverteilt.
- *Erwartungswert:*
$$\mathrm{E}(y_i) = \mu_i = h\left(\eta_i^{add}\right)$$
mit der (bekannten) Verknüpfungsfunktion h.
- *Prädiktor:*
$$\eta_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin} = \eta_i^{add}.$$
- *Bemerkung:* Interaktionen, räumliche Effekte und zufällige Effekte können wie im AM ergänzt werden. Werden zufällige Effekte einbezogen, so erhält man Generalisierte Additive Gemischte Modelle (GAMM).

2.9.8 Strukturiert-additive Regression (STAR, Kapitel 8)

- *Zielgröße:* Die Beobachtungen y_i sind metrisch, kategorial oder Zähldaten und je nach Skalenniveau und Verteilungsannahme entweder normal-, binomial-, Poisson- oder gammaverteilt.
- *Erwartungswert:*
$$\mathrm{E}(y_i) = \mu_i = h(\eta_i)$$
mit der Verknüpfungsfunktion h.
- *Prädiktor:*
$$\eta_i = f_1(v_{i1}) + \ldots + f_q(v_{iq}) + \eta_i^{lin}$$
Dabei sind $v_1, \ldots, v_q$ ein- oder auch mehrdimensionale Variablen verschiedenen Typs, die aus den Kovariablen gebildet werden. Ebenso sind auch die Funktionen $f_1, \ldots, f_q$ von unterschiedlichem Typ. Beispiele sind:

$f_1(v_1) = f(z_1)$,	$v_1 = z_1$,	nichtlinearer Effekt von z_1.
$f_2(v_2) = f_{geo}(s)$,	$v_2 = s$,	räumlicher Effekt der Lokationsvariable s.
$f_3(v_3) = f(z)x$,	$v_3 = (z, x)$,	mit z variierender Effekt von x.
$f_4(v_4) = f_{12}(z_1, z_2)$,	$v_4 = (z_1, z_2)$,	nichtlineare Interaktion zwischen z_1 und z_2.
$f_5(v_5) = \gamma_i u$,	$v_5 = u$,	zufälliger Effekt von u.

- *Bemerkung:* Da der Prädiktor Effekte unterschiedlichen Typs in strukturiert-additiver Form enthält, sprechen wir von strukturiert-additiven Regressionsmodellen (STAR). Sämtliche bisher besprochenen Modellklassen sind als Spezialfall enthalten.

3 Lineare Regressionsmodelle

In diesem Kapitel befassen wir uns mit Theorie und Anwendung des *linearen Regressionsmodells*. Einige Beispiele für diese Modellklasse haben wir bereits in Kapitel 2.2 kennengelernt. Lineare Modelle spielen zweifellos eine Hauptrolle in der Statistik. Neben der direkten Anwendung in Regressionsfragestellungen dienen lineare Modelle auch als Grundlage für eine Vielzahl komplexerer Regressionsverfahren, etwa die generalisierten linearen Modelle in Kapitel 4 oder die semiparametrische Regression in den Kapiteln 7 und 8.

Dieses Kapitel besteht aus drei Teilen: Im ersten Teil (Abschnitte 3.1–3.3) befassen wir uns ausschließlich mit dem in Kapitel 2.2.2 eingeführten klassischen linearen Modell $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$ mit unabhängigen und identisch verteilten Fehlern. Abschnitt 3.1 definiert das Modell, diskutiert ausführlich dessen Annahmen und zeigt auf, wie der Einfluss der Kovariablen modelliert werden kann. Die folgenden beiden Abschnitte 3.2 und 3.3 beschreiben die klassische Test- und Schätztheorie im linearen Modell. Dabei spielt die Methode der Kleinsten-Quadrate (KQ-Methode) zur Schätzung der unbekannten Regressionskoeffizienten $\beta_0, \ldots, \beta_k$ eine Hauptrolle.

Im zweiten Teil dieses Kapitels (Abschnitte 3.4 und 3.5) werden alternative Modelldefinitionen bzw. Inferenzkonzepte diskutiert, um Schwächen des klassischen linearen Modells zu begegnen. Abschnitt 3.4 befasst sich mit dem allgemeinen linearen Modell und dessen Anwendungen. Dieses Modell erlaubt auch korrelierte Störgrößen und heteroskedastische Varianzen der Störgrößen. Abschnitt 3.5 beschreibt Bayesianische lineare Modelle als alternatives Inferenzkonzept. In der modernen Statistik spielen Bayesianische Ansätze eine immer wichtigere Rolle. Der Abschnitt über Bayesianische lineare Modelle kann daher (in Verbindung mit Anhang B.5) als Grundlage für die in den folgenden Kapiteln beschriebenen Bayesianischen Verfahren dienen.

Der dritte Teil dieses Kapitels (Abschnitt 3.6) behandelt Methoden und Strategien zur Modellwahl und Diagnose und die Anwendung der in den ersten beiden Teilen vorgestellten Modelle und Inferenzverfahren. Anhand einer Fallstudie zu Preisen von Gebrauchtwagen zeigen wir auf, wie lineare Modelle in der Praxis eingesetzt werden können.

In einem abschließenden Abschnitt 3.7 geben wir noch einige Literaturhinweise und liefern Herleitungen und Beweise für Aussagen im Haupttext, die dort aus Gründen der besseren Lesbarkeit zunächst nicht bewiesen werden.

3.1 Das klassische lineare Modell

3.1.1 Modelldefinition

Wie bisher interessieren wir uns für den Einfluss einer Reihe von erklärenden Variablen $x_1, \ldots, x_k$ auf eine primär interessierende Variablen y. Allgemein modellieren wir den

Zusammenhang zwischen y und $x_1, \ldots, x_k$ durch eine Funktion $f(x_1, \ldots, x_k)$. Die Beziehung gilt nicht exakt, sondern ist durch zufällige Störungen ε überlagert. In der Praxis werden in den meisten Fällen additive Störgrößen zu Grunde gelegt, so dass wir

$$y = f(x_1, \ldots, x_k) + \varepsilon$$

erhalten. Ziel ist die Schätzung der unbekannten Funktion f, d.h. die Trennung der systematischen Komponente f vom zufälligen Fehler ε.

Im Rahmen der linearen Modelle werden die folgenden speziellen Annahmen über die unbekannte Funktion f und den Fehler ε getroffen:

1. *Die systematische Komponente f ist eine Linearkombination der Kovariablen*

   Die unbekannte Funktion $f(x_1, \ldots, x_k)$ wird als Linearkombination der Kovariablen modelliert, d.h.

   $$f(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

   Die Parameter $\beta_0, \beta_1, \ldots, \beta_k$ sind unbekannt und müssen geschätzt werden. Der Parameter β_0 fungiert als Konstante (Intercept).

   Fassen wir die Kovariablen und die unbekannten Parameter zu den $p = (k+1)$ dimensionalen Vektoren $\boldsymbol{x} = (1, x_1, \ldots, x_k)'$ und $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ zusammen, so erhalten wir kompakter

   $$f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}.$$

   Man beachte, dass das erste Element des Vektors $\boldsymbol{x}$ gleich Eins sein muss, um die Konstante im Modell zu berücksichtigen.

   Auf den ersten Blick erscheint die Annahme einer linearen Funktion, d.h. eines linearen Zusammenhangs zwischen y und $\boldsymbol{x}$ sehr restriktiv. Wie bereits in Kapitel 2.2 (Seite 20) demonstriert, können im Rahmen der linearen Modelle auch nichtlineare Beziehungen modelliert werden. Auf diesen Aspekt gehen wir noch ausführlich in Abschnitt 3.1.4 und in Kapitel 7.1.1 ein.

2. *Additivität der Störgröße*

   Eine weitere Grundannahme des linearen Modells ist die Additivität der Störgröße, d.h.

   $$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon.$$

   Auch dies erscheint zunächst relativ restriktiv. Es zeigt sich jedoch, dass diese Annahme in einer Vielzahl von praktischen Anwendungen zumindest annähernd erfüllt ist. Darüberhinaus können Problemstellungen mit zunächst nicht additiven Störgrößen durch Transformation der Zielvariable y häufig auf Modelle mit additiven Störgrößen zurückgeführt werden (vergleiche in Abschnitt 3.1.3 die Seiten 70 ff.).

Zur Schätzung der unbekannten Parameter $\boldsymbol{\beta}$ werden Daten y_i und $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, $i = 1, \ldots, n$, erhoben, so dass man für jede Beobachtung die Gleichung

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{3.1}$$

erhält. Definieren wir die Vektoren

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

sowie die *Designmatrix* $\boldsymbol{X}$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}$$

so lassen sich die n Gleichungen aus (3.1) kompakt in Matrixnotation schreiben:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Für den Rest dieses Kapitels nehmen wir an, dass $\boldsymbol{X}$ vollen Spaltenrang besitzt, d.h. $\mathrm{rg}(\boldsymbol{X}) = k + 1 = p$. Die Spalten von $\boldsymbol{X}$ sind also linear unabhängig. Eine notwendige Bedingung hierfür ist, dass die Anzahl n der Beobachtungen größer oder gleich der Zahl p der Regressionskoeffizienten ist. Die Annahme ist beispielsweise dann verletzt, wenn eine der Kovariablen lediglich eine lineare Transformation einer zweiten Variable ist, etwa $x_1 = a + b\,x_2$. Die Variable x_1 könnte beispielsweise die Größe einer Person in Metern und x_2 die Größe in Zentimetern bedeuten, d.h. $a = 0$ und $b = 1/100$. Die beiden Variablen würden also die gleiche Information beinhalten und lediglich in unterschiedlichen Skalen vorliegen. Allgemein ist die Annahme linearer Unabhängigkeit immer dann verletzt, wenn mindestens eine der erklärenden Variablen als Linearkombination der übrigen darstellbar ist, d.h. wenn die Information bereits in den anderen Variablen enthalten ist. Wir werden später sehen, dass die lineare Unabhängigkeit der Spalten von $\boldsymbol{X}$ notwendig ist, um *eindeutige* Schätzungen für die Regressionsparameter $\boldsymbol{\beta}$ zu erhalten.

Das Modell wird komplettiert durch Annahmen über die Störgrößen. Im Rahmen des *klassischen linearen Modells* werden über den Vektor $\boldsymbol{\varepsilon}$ der Störgrößen folgende zusätzliche Annahmen getroffen:

1. *Erwartungswert der Störgrößen*

   Die Störungen sind im Mittel Null, d.h. $\mathrm{E}(\varepsilon_i) = 0$ bzw. $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$.

2. *Varianz und Korrelation der Störgrößen*

   Die Varianz der Störgrößen bleibt für alle Beobachtungen konstant, d.h. $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Wir sprechen auch von homoskedastischen Fehlern. Die Fehler heißen heteroskedastisch, wenn die Varianzen von Beobachtung zu Beobachtung variieren, d.h. $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. Neben homoskedastischen Varianzen nehmen wir zusätzlich an, dass die Störgrößen unkorreliert sind, also $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ für $i \neq j$. Die Annahme homoskedastischer und unkorrelierter Fehler führt schließlich zur Kovarianzmatrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}$.

3. *Annahmen über die Kovariablen bzw. die Designmatrix*

   Wir wollen zwei Situationen unterscheiden. Zum Einen können die erklärenden Variablen $x_1, \ldots, x_k$ deterministische, d.h. nicht stochastische, etwa im Rahmen von geplanten Experimenten vorgegebene Größen sein. Weitaus häufiger dürfte jedoch der Fall sein, dass neben den Beobachtungen der Zielvariablen auch die Kovariablen stochastisch sind. In diesem Fall spricht man auch von Beobachtungsdaten. In praktisch allen Anwendungsfällen in diesem Buch sind die Kovariablen wie auch die Zielgröße zufällig. Beim Münchner Mietspiegel etwa werden aus dem Mietwohnungsbestand der Stadt München die Wohnungen zufällig ausgewählt, so dass sämtliche Charakteristika der Wohnungen in der Stichprobe, beispielsweise die Wohnfläche oder das

Das klassische lineare Modell

Das Modell

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

heißt klassisches lineares Regressionsmodell, wenn folgende Annahmen gelten:

1. $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$.
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}$.
3. Die Designmatrix $\boldsymbol{X}$ besitzt vollen Spaltenrang, d.h. $\mathrm{rg}(\boldsymbol{X}) = k + 1 = p$.

Wir sprechen von klassischer Normalregression, wenn zusätzlich die Annahme

4. $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

getroffen wird.

Baujahr, als Realisierungen zufälliger Größen angesehen werden können. Allgemein können wir bei stochastischen Regressoren die Beobachtungen $(y_i, \boldsymbol{x}_i')$, $i = 1, \ldots, n$, als Realisationen eines Zufallsvektors $(y, \boldsymbol{x}')$ ansehen, und sämtliche Modellannahmen als bedingt auf die Designmatrix auffassen, also z.B. $\mathrm{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \boldsymbol{0}$ anstelle von $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$.

In jedem Fall nehmen wir an, dass die Designmatrix vollen Spaltenrang besitzt, d.h. $\mathrm{rg}(\boldsymbol{X}) = k + 1 = p$.

4. *Normalverteilte Störgrößen*

   Zur Konstruktion von Konfidenzintervallen und Hypothesentests wird in vielen Fällen zusätzlich angenommen, dass die zufälligen Fehler (zumindest approximativ) normalverteilt sind. Zusammen mit den ersten beiden Annahmen erhalten wir also $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ bzw. $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Bei stochastischen Kovariablen ergibt sich entsprechend $\varepsilon_i \mid \boldsymbol{x}_i \sim \mathrm{N}(0, \sigma^2)$ bzw. $\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Wenn normalverteilte Fehler angenommen werden, dann sprechen wir auch von der *klassischen linearen Normalregression.*

Aus Gründen der notationellen Vereinfachung unterdrücken wir für stochastische Kovariablen im Folgenden meist die Abhängigkeit von der Designmatrix $\boldsymbol{X}$.

Als erste einfache Folgerungen aus den Modellannahmen erhalten wir

$$\begin{aligned}
\mathrm{E}(y_i) &= \mathrm{E}(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \\
\mathrm{Var}(y_i) &= \mathrm{Var}(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2, \\
\mathrm{Cov}(y_i, y_j) &= \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0
\end{aligned}$$

für den Erwartungswert und die Varianz von y_i bzw. die Kovarianz zwischen y_i und y_j. In Matrixnotation gilt

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{und} \quad \mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}.$$

Wird zusätzlich angenommen, dass $\boldsymbol{\varepsilon}$ normalverteilt ist, so folgt

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

Man beachte, dass die Kovariablen ausschließlich den Erwartungswert von $\boldsymbol{y}$ beeinflussen. Die Varianz σ^2 der y_i bzw. die Kovarianzmatrix $\sigma^2\boldsymbol{I}$ von $\boldsymbol{y}$ sind dagegen unabhängig von den Kovariablen.

3.1.2 Modellparameter, Schätzungen und Residuen

Bevor wir mit einer Diskussion der Modellannahmen des klassischen linearen Regressionsmodells fortfahren, führen wir noch einige Bezeichnungen ein. Grundsätzlich unterscheiden wir die Modellparameter von ihren Schätzungen durch ein „Dach“, d.h. Schätzungen von $\boldsymbol{\beta}$ und σ^2 werden mit $\hat{\boldsymbol{\beta}}$ und $\hat{\sigma}^2$ bezeichnet. Diese Unterscheidung ist notwendig, da es wohl kaum gelingen wird, den „wahren“ Parametervektor $\boldsymbol{\beta}$ ohne Fehler zu schätzen, so dass stets $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ gilt. Die Regressionsparameter werden üblicherweise durch die Methode der kleinsten Quadrate geschätzt, die bereits in Kapitel 2.2 für den Fall der linearen Einfachregression skizziert wurde, vergleiche Seite 21. Ausführlich gehen wir auf die Schätzung von $\boldsymbol{\beta}$ und σ^2 in Abschnitt 3.2 ein.

Ausgehend von einer Schätzung $\hat{\boldsymbol{\beta}}$ erhalten wir eine (naheliegende) Schätzung des Erwartungswerts $\mathrm{E}(y_i)$ von y_i durch

$$\widehat{\mathrm{E}(y_i)} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}.$$

Diese Schätzung wird üblicherweise mit $\hat{y}_i$ bezeichnet, d.h. $\hat{y}_i = \widehat{\mathrm{E}(y_i)}$. Der Schätzfehler, d.h. die Abweichung $y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ des wahren Wertes y_i vom Schätzwert $\hat{y}_i$ heißt Residuum und wird mit $\hat{\varepsilon}_i$ bezeichnet. Es gilt also

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}.$$

Fassen wir sämtliche Residuen zum Vektor $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)'$ zusammen, so erhalten wir

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

Es ist wichtig zu verstehen, dass die Residuen $\hat{\varepsilon}_i$ nicht gleich den Störgrößen ε_i sind, die wie der Parametervektor $\boldsymbol{\beta}$ unbekannt sind. Vielmehr können die Residuen $\hat{\varepsilon}_i$ als Schätzungen für ε_i angesehen werden.

Die Residuen beinhalten die nicht durch die Kovariablen erklärte Reststreuung der Daten. Gelegentlich, insbesondere bei der Modelldiagnose, benötigen wir sogenannte partielle Residuen, bei denen der Einfluss der meisten jedoch nicht aller Kovariablen entfernt ist. Die partiellen Residuen bezüglich der j-ten Kovariable x_j sind definiert durch

$$\begin{aligned}
\hat{\varepsilon}_{x_j,i} &= y_i - \hat{\beta}_0 - \ldots - \hat{\beta}_{j-1} x_{i,j-1} - \hat{\beta}_{j+1} x_{i,j+1} - \ldots - \hat{\beta}_k x_{ik} \\
&= y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \hat{\beta}_j x_{ij} \\
&= \hat{\varepsilon}_i + \hat{\beta}_j x_{ij}.
\end{aligned}$$

Hier ist der Einfluss aller Kovariablen bis auf x_j entfernt. Wir werden später sehen, dass die partiellen Residuen bezüglich der Kovariable x_j bei der Beurteilung der gewählten Modellierung hilfreich sind.

3.1.3 Diskussion der Modellannahmen

Das klassische lineare Regressionsmodell enthält eine Reihe von Annahmen über die Art und Weise des Zusammenhangs zwischen der abhängigen Variable y und den erklärenden Variablen $x_1, \ldots, x_k$. Dieser Abschnitt verfolgt zwei Ziele: Zum Einen sollen die Annahmen nochmal vergegenwärtigt und einer kritischen Betrachtung unterzogen werden. Zum Anderen sollen Hinweise auf Abschnitte im Buch gegeben werden, in denen verschiedene restriktive Annahmen des klassischen linearen Modells durch flexiblere Modellierungen ersetzt werden.

Zur Illustration benutzen wir meist das einfache Regressionsmodell $y = \beta_0 + \beta_1 x + \varepsilon$ mit nur einer erklärenden Variable.

Linearität des Einflusses der Kovariablen

Auf den ersten Blick erscheint die lineare Modellierung des Einflusses der Kovariablen sehr restriktiv. Im Rahmen des linearen Modells können jedoch auch nichtlineare Beziehungen modelliert werden. Betrachten wir beispielsweise das Modell

$$y_i = \beta_0 + \beta_1 \log(z_i) + \varepsilon_i,$$

in dem der Einfluss der erklärenden Variable z_i logarithmisch ist. Dieses Modell kann problemlos im Rahmen der linearen Modelle behandelt werden. Dazu definieren wir einfach die neue Variable $x_i = \log(z_i)$ und erhalten wieder ein lineares Modell $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Im Allgemeinen lassen sich alle nichtlinearen Beziehungen auf ein einfaches lineares Modell zurückführen, solange diese *linear in den Parametern* sind. Ein Beispiel für ein in den Parametern nichtlineares Modell ist gegeben durch

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

In diesem Buch werden eine Vielzahl von Techniken zur Modellierung nichtlinearer Beziehungen zwischen der abhängigen Variable y und einer Einflussvariable x besprochen. Bereits im nächsten Abschnitt 3.1.4 werden wir auf zwei einfache Techniken genauer zu sprechen kommen, die Modellierung nichtlinearer Zusammenhänge durch Variablentransformation und durch Polynome. Ein Beispiel für die Modellierung durch Variablentransformation haben wir bereits in Beispiel 2.2 (Seite 22) kennen gelernt. Polynomiale Regression wurde in Beispiel 2.6 (Seite 28) zum ersten Mal verwendet. Beide einfache Techniken haben den Nachteil, dass die prinzipielle funktionale Form des Zusammenhangs bereits bekannt sein muss. Verfahren zur Schätzung nichtlinearer Zusammenhänge *ohne spezielle Kenntnis der funktionalen Form* behandeln wir ausführlich in den Kapiteln 7 und 8.

Homoskedastische Varianz der Störgrößen

Homoskedastische Varianz der Störgrößen bedeutet, dass die Fehler ε_i nicht in Abhängigkeit von y bzw. von einer oder mehreren Kovariablen x_j systematisch größer oder kleiner werden. Abbildung 3.1 illustriert den Unterschied zwischen homo- und heteroskedastischen Varianzen anhand simulierter Daten. Abbildung a) zeigt die idealtypische Situation

Abb. 3.1. *Illustration für homo- und heteroskedastische Varianzen: Die jeweils linken Abbildungen zeigen die (simulierten) Daten inklusive Regressionsgerade. In den jeweils rechten Abbildungen sind die dazu gehörigen Störgrößen dargestellt. Die Daten entsprechen dem Regressionsmodell* $y_i \sim \mathrm{N}(-1 + 2x_i, 1)$ *in a) und b) sowie* $y_i \sim \mathrm{N}(-1 + 2x_i, (0.1 + 0.3(x_i + 3))^2)$ *in c) und d).*

homoskedastischer Fehler. Die Beobachtungen schwanken mit gleich bleibender Variabilität um die eingezeichnete Regressionsgerade. In Abbildung b) findet man die zu a) gehörenden Störgrößen. Diese schwanken regellos und mit gleich bleibender Varianz um Null. Da es sich hier um simulierte Daten handelt, sind wir in der seltenen Situation, in der wir die wahren Regressionskoeffizienten und damit die Regressionsgerade und die Störgrößen kennen. In realen Datensituationen müssen wir auf Schätzungen für die Regressionsparameter zurückgreifen. Da die Störgrößen $\varepsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ von den unbekannten Regressionsparametern abhängen, werden diese durch $\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, also die Residuen, geschätzt. Bei der Überprüfung der Annahme homoskedastischer Fehler sind wir also üblicherweise auf die Untersuchung der Residuen angewiesen, vergleiche hierzu ausführlicher Abschnitt 3.6.4.

Die Abbildungen 3.1 c) und d) veranschaulichen ein typisches Beispiel für heteroskedastische Varianzen. Hier nimmt die Streuung um die Regressionsgerade mit wachsendem x zu. Die Störgrößen schwanken zwar um Null, werden aber mit wachsendem x immer größer.

Abb. 3.2. *Illustration heteroskedastischer Varianzen am Beispiel der Mietspiegeldaten. Die linke Grafik zeigt das Streudiagramm zwischen Nettomiete und Wohnfläche. Zusätzlich eingezeichnet ist die geschätzte Regressionsgerade. In der rechten Grafik sind die dazu gehörigen Residuen in Abhängigkeit von der Wohnfläche abgebildet.*

Beispiel 3.1 Mietspiegel – heteroskedastische Varianzen

Der trichterförmige Verlauf der Störgrößen in Abbildung 3.1 c) und d) ist typisch für viele reale Datensituationen, wie sich anhand der Mietspiegeldaten demonstrieren lässt. Abbildung 3.2 zeigt das Streudiagramm zwischen Nettomiete und Wohnfläche (linke Grafik). Zusätzlich eingezeichnet ist die geschätzte Regressionsgerade. Die beobachteten Nettomieten streuen mit immer größer werdender Varianz um die eingezeichnete Regressionsgerade. Die Mieten für große Wohnungen weisen also eine weitaus größere Bandweite auf als die Mieten für kleine Wohnungen. Die rechte Grafik von Abbildung 3.2 stellt die dazu gehörigen Residuen in Abhängigkeit von der Wohnfläche dar. Wir beobachten die oben beschriebene, trichterförmige Gestalt. Dieses Phänomen ist uns bereits im Einführungskapitel in Beispiel 2.1 (Seite 20) begegnet. Dort wurde der Zusammenhang zwischen Nettomiete und Wohnfläche für Wohnungen in normaler Lage und ab dem Baujahr 1966 untersucht, vergleiche Abbildung 2.1 (Seite 21). Im Vergleich zur Analyse der gesamten Daten in Abbildung 3.2 ergibt sich ein sehr ähnliches Bild, wenn auch die zunehmende Variabilität nicht so ausgeprägt ist. Ursächlich hierfür ist vor allem, dass in Abbildung 2.1 die Wohnungen aufgrund der Einschränkung auf normale Wohnlagen homogener sind.

△

Es stellt sich natürlich die Frage, welche Auswirkungen das Ignorieren heteroskedastischer Varianzen besitzt. Es wird sich herausstellen, dass insbesondere die Varianzen der geschätzten Regressionskoeffizienten $\hat{\boldsymbol{\beta}}$ nicht korrekt geschätzt werden. Dieser Fehler hat wiederum Auswirkungen auf Hypothesentests über die Regressionsparameter und deren Konfidenzintervalle.

Auf die Problematik heteroskedastischer Varianzen werden wir ausführlich in Abschnitt 3.4 (Seite 124) eingehen. Dort werden wir auch das klassische lineare Modell erweitern, so dass heteroskedastische Varianzen berücksichtigt werden können.

Unkorreliertheit der Störgrößen

Die Unkorreliertheit der Störgrößen ist in einigen Fällen, insbesondere bei Zeitreihen und Longitudinaldaten, eine unrealistische Annahme. In vielen Fällen treten sogenannte au-

Abb. 3.3. *Illustration für autokorrelierte Störgrößen: Die Grafiken a) und b) illustrieren positiv korrelierte Störungen, die Grafiken c) und d) negativ korrelierte Störungen. In den jeweils linken Grafiken findet man die (simulierten) Daten inklusive (wahrer) Regressionsgerade. In den jeweils rechten Abbildung sind die dazu gehörigen Störungen dargestellt. Im Falle negativer Autokorrelation sind die Störungen mit einer durchgezogenen Linie verbunden, um den Vorzeichenwechsel der Störungen besser zu verdeutlichen. Die Daten mit positiver Korrelation entsprechen dem Regressionsmodell* $y_i = -1 + 2x_i + \varepsilon_i$ *mit* $\varepsilon_i = 0.9\varepsilon_{i-1} + u_i$ *und* $u_i \sim \mathrm{N}(0, 0.5^2)$. *Die Daten mit negativer Korrelation sind gemäß dem Modell* $y_i = -1 + 2x_i + \varepsilon_i$ *mit* $\varepsilon_i = -0.9\varepsilon_{i-1} + u_i$ *und* $u_i \sim \mathrm{N}(0, 0.5^2)$ *erzeugt.*

tokorrelierte Störungen auf. Hierbei besteht im einfachsten Fall zwischen den Störungen ε_i zum Zeitpunkt i und den Störungen ε_{i-1} der Vorperiode die lineare Beziehung $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$, wobei u_i unabhängig und identisch verteilte Zufallsvariablen sind. Wir sprechen von Autokorrelation erster Ordnung. Autokorrelation zweiter bzw. allgemeiner l-ter Ordnung liegt vor, wenn $\varepsilon_i = \rho_1\varepsilon_{i-1} + \rho_2\varepsilon_{i-2} + u_i$ bzw. $\varepsilon_i = \rho_1\varepsilon_{i-1} + \ldots + \rho_l\varepsilon_{i-l} + u_i$ gilt.

Zur Illustration autokorrelierter Störungen betrachte man zunächst Abbildung 3.3. Die Grafiken a) und b) zeigen simulierte Daten zur Veranschaulichung von Störungen mit *positiver* Autokorrelation erster Ordnung. Positiv korrelierte Störungen bedeuten, dass nach einer positiven (negativen) Störung tendenziell wieder eine positive (negative) Störung folgt. Das Gegenteil, negative Autokorrelation der Störgrößen, findet man in den Grafiken c) und d). Hier folgen auf positive (negative) Störungen tendenziell negative (positive) Störungen, d.h. die Störungen wechseln häufig das Vorzeichen.

Abb. 3.4. *Illustration für korrelierte Residuen bei fehlspezifiziertem Regressionsmodell: Abbildung a) zeigt die (simulierten) Daten und die Funktion $E(y_i \mid x_i) = sin(x_i) + x_i$. In Abbildung b) ist eine Regressionsgerade an die Daten angepasst, d.h. die nichtlineare Beziehung wird ignoriert. Die dazu gehörigen Residuen findet man in Abbildung c).*

Autokorrelierte Störungen treten in der Regel dann auf, wenn das Regressionsmodell fehlspezifiziert ist. Ein Grund könnte sein, dass der Einfluss einer erklärenden Variable nicht korrekt modelliert wird. Eine solche Situation haben wir in Abbildung 3.4 illustriert. Grafik a) zeigt simulierte Beobachtungen aus dem Modell $y_i = \sin(x_i) + x_i + \varepsilon_i$. Der bedingte Erwartungswert von y_i ist gegeben durch $\mathrm{E}(y_i \mid x_i) = \sin(x_i) + x_i$. Es handelt sich also um eine nichtlineare Funktion von x, vergleiche die durchgezogene Linie in Abbildung 3.4 a). In Grafik b) wurde (fälschlicherweise) eine Regressionsgerade an die Daten angepasst, d.h. unser Modell ist fehlspezifiziert. Die dazu gehörigen Residuen in Abbildung c) sind deutlich positiv autokorreliert.

Bei weitem am häufigsten treten autokorrelierte Störgrößen bei der Analyse von Zeitreihen oder Longitudinaldaten auf. In vielen Fällen können nicht alle relevanten erklärende Variablen in das Modell mit aufgenommen werden, weil sie nicht beobachtbar sind oder nicht erhoben wurden. Weisen diese nicht berücksichtigten Kovariablen einen zeitlichen Trend auf oder sind saisonalen Schwankungen unterworfen, so enstehen korrelierte Fehler. Zur Illustration verwenden wir wieder simulierte Daten. Abbildung 3.5 zeigt in den Grafiken a) und b) den zeitlichen Verlauf zweier Variablen x_1 und x_2. Offenbar ist x_1 einem deutlichen zeitlichen Trend unterworfen, während x_2 im Zeitablauf verhältnismäßig regellos um Null schwankt. Wir betrachten jetzt das Regressionsmodell

Abb. 3.5. *Illustration für autokorrelierte Störungen bei Nichtberücksichtigung von Kovariablen, die einen zeitlichen Trend aufweisen. Die Grafiken a) und b) zeigen die Kovariablen x_1 und x_2 im Zeitablauf. In den Grafiken c)-e) sind die Residuen für die Regressionsmodelle $y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i$ (korrekt spezifiziertes Modell), $y_i = \beta_0+\beta_1 x_{i1}+\varepsilon_i$ (x_2 nicht berücksichtigt) und $y_i = \beta_0 + \beta_1 x_{i2} + \varepsilon_i$ (x_1 nicht berücksichtigt) zu sehen.*

$y_i = -1 + x_1 - 0.6\,x_2 + \varepsilon_i$, $i = 1, \ldots, 100$, mit unabhängigen, identisch normalverteilten Störgrößen $\varepsilon_i \sim \mathrm{N}(0, 0.5^2)$. Bei diesem Modell handelt es sich um ein klassisches lineares Regressionsmodell, das den im Kasten auf Seite 62 getroffenen Annahmen genügt. Als Schätzung erhalten wir $\hat{y}_i = -1.03 + 0.98x_{i1} - 0.61x_{i2}$. In Abbildung 3.5 c) sind die dazu gehörigen Residuen $\hat{\varepsilon}_i$ im Zeitablauf dargestellt. Wie nicht anders zu erwarten, lässt

sich kein auffälliges Verhalten feststellen. Insbesondere finden sich keinerlei Anzeichen für Autokorrelation. Wird die Variable x_2 nicht in das Regressionsmodell aufgenommen, so erhalten wir die Schätzung $\hat{y}_i = 1.01 + 0.95 \cdot x_{i1}$ und die in Abbildung 3.5 d) gezeigten Residuen im Zeitablauf. Obwohl die Variable x_2 „vergessen“ wurde, erkennen wir nach wie vor keine Besonderheiten. Die Schätzungen der Regressionskoeffizienten sind in guter Übereinstimmung mit den wahren Werten und die Residuen streuen im Zeitablauf annähernd regellos. Vergessen wir hingegen die trendbehaftete Variable x_1, so erhalten wir die Schätzung $\hat{y}_i = -2.22 - 0.60 \cdot x_{i2}$. Die dazu gehörigen Residuen sind in Abbildung 3.5 e) abgebildet und weisen deutlich auf Autokorrelation hin. Warum sind die Residuen bei Nichtberücksichtigung von x_2 eher unauffällig und zeigen bei Vernachlässigung von x_1 deutliche Autokorrelation? Die einfache Antwort ist, dass x_1 im Gegensatz zu x_2 einen ausgeprägten zeitlichen Trend aufweist. Wir können uns die Zusammenhänge wie folgt erklären: Die im Modell weggelassenen Ausdrücke $\beta_1 x_1$ bzw. $\beta_2 x_2$ können gedanklich in die Störgrößen integriert werden. Wenn wir diese neuen Störgrößen mit $\tilde{\varepsilon}$ bezeichnen, erhalten wir $\tilde{\varepsilon} = \beta_1 x_1 + \varepsilon$, wenn x_1 im Modell vergessen wird, und $\tilde{\varepsilon} = \beta_2 x_2 + \varepsilon$, wenn x_2 vergessen wird. Da x_2 im Zeitablauf regellos um Null streut, ergibt die Hinzunahme des vergessenen Ausdrucks $\beta_2\, x_2$ wieder annähernd unkorrelierte Störungen. Durch die Hinzunahme des Ausdrucks $\beta_1 x_1$ erhalten wir wegen des zeitlichen Trends von x_1 autokorrelierte Störungen, d.h. der Trend in x_1 spiegelt sich in den Störungen wieder.

Analog zu heteroskedastischen Fehlern stellt sich wieder die Frage, welche Auswirkungen ein eventuelles Ignorieren von korrelierten Fehlern hat. Ausführlich gehen wir darauf in Abschnitt 3.4 (Seite 124) ein. Dort behandeln wir auch Schätzmethoden zur Berücksichtigung korrelierter Fehler. Intuitiv macht man sich aber leicht klar, dass ein Ignorieren der Korrelation einen Informationsverlust bedeutet. Dazu betrachten wir die Aufgabe, basierend auf einer neuen Beobachtung $\boldsymbol{x}_{n+1}$ der erklärenden Variablen zum Zeitpunkt $n+1$ den Wert der abhängigen Variable zu prognostizieren. Bei unkorrelierten Fehlern können wir $\hat{y}_{n+1} = \boldsymbol{x}'_{n+1}\hat{\boldsymbol{\beta}}$ als nahe liegende Prognose verwenden. Bei korrelierten Fehlern verschenken wir durch dieses Vorgehen Information. Weisen beispielsweise die Residuen positive Autokorrelation erster Ordnung auf, dann folgen auf positive Residuen tendenziell wieder positive Residuen. Handelt es sich also bei $\hat{\varepsilon}_n$ um einen positiven Wert, so ist tendenziell das Residuum $\hat{\varepsilon}_{n+1}$ zum Zeitpunkt $n+1$ wieder positiv und weicht von $\hat{\varepsilon}_n$ umso weniger ab, je größer die Korrelation ist. Es ist also sinnvoll, für y_{n+1} einen *größeren* Wert als $\boldsymbol{x}'_{n+1}\hat{\boldsymbol{\beta}}$ zu prognostizieren. Umgekehrt ist es nahe liegend, bei negativem Residuum $\hat{\varepsilon}_n$ einen *kleineren* Wert als $\boldsymbol{x}'_{n+1}\hat{\boldsymbol{\beta}}$ zu prognostizieren. In Abschnitt 3.4 (Seite 124) werden wir sehen, dass $\hat{y}_{n+1} = \boldsymbol{x}'_{n+1}\hat{\boldsymbol{\beta}} + \hat{\rho}\hat{\varepsilon}_n$ eine optimale Prognose von y_{n+1} darstellt, wobei $\hat{\rho}$ der (geschätzte) Korrelationskoeffizient ist, mit dem $\hat{\varepsilon}_i$ und $\hat{\varepsilon}_{i-1}$ korreliert sind.

Additivität der Störgrößen

Prinzipiell sind für die Art des Fehlerprozesses viele verschiedene Modelle denkbar. In den meisten Anwendungen werden jedoch im Wesentlichen zwei Varianten betrachtet: zum einen additive Störungen und zum anderen multiplikative Störungen. Ein Beispiel für multiplikative Störungen ist das sogenannte exponentielle Modell

$$\begin{aligned} y_i &= \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i) \\ &= \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik})\exp(\varepsilon_i) \end{aligned} \tag{3.2}$$

Abb. 3.6. *Beispiel für ein multiplikatives Modell. Die Abbildungen a) und b) zeigen Streudiagramme zwischen y und x_1 bzw. x_2 im simulierten Modell $y_i = \exp(1 + x_{i1} - x_{i2} + \varepsilon_i)$. In den Abbildungen c) und d) findet man Streudiagramme zwischen* $\log(y)$ *und x_1 bzw. x_2.*

mit den multiplikativen Störungen $\tilde{\varepsilon}_i = \exp(\varepsilon_i)$. Modelle mit multiplikativen Fehlern sind bei exponentiellen Zusammenhängen plausibler als additive, da die Fehler proportional zum Durchschnittswert von y sind. Abbildung 3.6 zeigt simulierte Daten mit multiplikativen Fehlern. In den Abbildungen a) und b) sind Streudiagramme zwischen einer (simulierten) Zielgröße y und den beiden erklärenden Variablen x_1 und x_2 abgedruckt. Die Daten sind gemäß dem Modell

$$y_i = \exp(1 + x_{i1} - x_{i2} + \varepsilon_i) \tag{3.3}$$

entstanden. Anhand der Streudiagramme lassen sich nur schwer Aussagen über Art und Stärke des Zusammenhangs zwischen y und x_1 bzw. x_2 treffen.

Viele gängige Modelle mit multiplikativen Fehlern können durch einfache Variablentransformationen als lineare Modelle mit additiven Fehlern geschrieben werden. Logarithmieren von (3.2) liefert das lineare Modell

$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i.$$

Ein exponentielles Modell lässt sich also im Rahmen der linearen Modelle behandeln, indem die Zielgröße logarithmiert wird. Die Abbildungen 3.6 c) und d) zeigen für das simulierte Modell (3.3) Streudiagramme zwischen der logarithmierten Zielgröße $\log(y)$

Abb. 3.7. *Streudiagramme zwischen dem Absatz eines Produktes und seinem Preis (Grafik a) bzw. dem Preis eines Konkurrenzproduktes (Grafik b).*

und den Kovariablen x_1 und x_2. Hier ist der (lineare) Zusammenhang zwischen $\log(y)$ und x_1 bzw. x_2 deutlich erkennbar.

Unter der Voraussetzung normalverteilter Fehler ε ist die Zielgröße y *lognormalverteilt*, vgl. Anhang B, Seite 459. Der (bedingte) Erwartungswert der Zielgröße y ist dann gegeben durch

$$\mathrm{E}(y_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma^2/2).$$

Beispiel 3.2 Supermarkt-Scanner-Daten

Abbildung 3.7 zeigt zu den simulierten Daten ähnlich aussehende Streudiagramme für reale Daten. Bei den Daten handelt es sich um Scanner-Daten, die heute routinemäßig beim Bezahlen an der Supermarktkasse gewonnen werden. Im vorliegenden Fall geht es darum, den Zusammenhang zwischen dem wöchentlichen Absatz eines Produktes (hier einer Kaffeemarke) und seinem Preis bzw. dem Preis von Konkurrenzprodukten zu untersuchen. Abbildung a) zeigt das Streudiagramm zwischen dem Absatz und dem Preis des Produktes. In Abbildung b) findet man das Streudiagramm zwischen Absatz und dem Preis eines Konkurrenzproduktes. Die Hypothese, dass ein höherer eigener Preis zu geringerem Absatz und ein höherer Konkurrenzpreis zu erhöhtem Absatz führt, ist anhand der Streudiagramme nicht leicht zu überprüfen.

△

3.1.4 Modellierung des Einflusses der Kovariablen

Mit der Modellierung des Einflusses der Kovariablen haben wir uns bereits in Kapitel 2.2 (Seite 20) und im Zusammenhang mit der Diskussion der Modellannahmen im letzten Abschnitt 3.1.3 beschäftigt. In diesem Abschnitt wollen wir diesen wichtigen Aspekt weiter vertiefen.

Metrische Kovariablen

Wie bereits bei der Diskussion der Modellannahmen im letzten Abschnitt erwähnt, können im Rahmen der linearen Modelle auch nichtlineare Regressionsbeziehungen modelliert werden. Bei metrischen erklärenden Variablen ist eine nichtlineare Spezifikation

in vielen Fällen notwendig. In diesem Abschnitt wollen wir zwei einfache Verfahren behandeln, mit denen nichtlineare Beziehungen im Modell berücksichtigt werden können. Es handelt sich um einfache *Variablentransformationen* und um die *Modellierung durch Polynome*. Flexiblere und vor allem automatisierte Verfahren behandeln wir dann in Kapitel 7.

Die Modellierung durch Variablentransformation bzw. durch Polynome lässt sich am besten anhand von Beispielen demonstrieren. Zur Illustration verwenden wir die Mietspiegeldaten, wobei wir uns (zunächst) auf die Modellierung des Zusammenhangs zwischen Nettomiete und Wohnfläche konzentrieren.

Beispiel 3.3 Mietspiegel – Variablentransformation

Im Zusammenhang mit homo- bzw. heteroskedastischen Varianzen in Abschnitt 3.1.3 wurde bereits eine Regressionsgerade zwischen Nettomiete und Wohnfläche angepasst, vergleiche Abbildung 3.2. Hier erscheint die Annahme eines linearen Zusammenhangs gerechtfertigt. Wenn wir stattdessen die Nettomiete pro qm als Zielvariable verwenden, erhalten wir die Schätzung $\widehat{mieteqm}_i = 18.46 - 0.068\,flaeche_i$. Abbildung 3.8 zeigt in Grafik a) das Streudiagramm zwischen Nettomiete pro qm und Wohnfläche. Zusätzlich eingezeichnet ist die geschätzte Regressionsgerade. Die dazu gehörigen Residuen sind in Abbildung c) abgedruckt. Wir stellen fest, dass für kleinere Wohnungen ($flaeche < 30$) die Residuen größtenteils positiv sind. Zur Verdeutlichung haben wir in Abbildung e) für jeden Quadratmeterwert die durchschnittlichen Residuen gezeichnet. Beide Grafiken liefern einen deutlichen Hinweis darauf, dass ein nichtlinearer Zusammenhang zwischen Nettomiete pro qm und Wohnfläche vorliegt. Eine Möglichkeit zur Spezifikation einer nichtlinearen Beziehung ist durch Variablentransformation gegeben. Dazu gehen wir über zu dem Regressionsmodell

$$mieteqm_i = \beta_0 + \beta_1 \cdot f(flaeche_i) + \varepsilon_i, \tag{3.4}$$

in dem f eine beliebige Funktion ist. Die Funktion f muss vor der Schätzung geeignet spezifiziert werden, wird also nicht mitgeschätzt. Offensichtlich ist dieses Modell eine Verallgemeinerung des linearen Modells $mieteqm_i = \beta_0 + \beta_1 flaeche_i + \varepsilon_i$. Durch Definition der Variable $x_i = f(flaeche_i)$ erhalten wir jedoch wieder ein lineares Modell $mieteqm_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Zur Modellierung des Zusammenhangs zwischen Nettomiete pro qm und Wohnfläche scheint beispielsweise die Transformation $f(flaeche_i) = 1/flaeche_i$ geeignet. Diese Transformation wurde auch in Beispiel 2.2 für den Teildatensatz aller nach 1966 gebauten Wohnungen in normaler Wohnlage gewählt. Als Schätzung erhalten wir

$$\widehat{mieteqm}_i = 9.228 + 273.348 \cdot f(flaeche_i) = 9.228 + 273.348 \cdot 1/flaeche_i.$$

Die Designmatrix zu diesem Modell sieht wie folgt aus:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1/30 \\ 1 & 1/37 \\ \vdots & \vdots \\ 1 & 1/73 \\ 1 & 1/73 \end{pmatrix}.$$

Die zweite Spalte der Designmatrix enthält also die Beobachtungen $x_1 = f(flaeche_1) = 1/flaeche_1 = 1/30$, $x_2 = 1/37, \ldots$ der Variable x, die durch Transformation der Variable *flaeche* gebildet werden.

Zur Interpretation der Ergebnisse ist es zweckmäßig, den geschätzten Zusammenhang zu visualisieren. In Abbildung 3.8 b) ist daher der geschätzte Zusammenhang

Abb. 3.8. *Illustration zur Modellierung nichtlinearer Beziehungen durch Variablentransformation am Beispiel der Mietspiegeldaten. Die linke Spalte zeigt für eine lineare Modellierung die geschätzte Regressionsgerade inklusive Beobachtungen (Grafik a), die dazu gehörigen Residuen (Grafik c) sowie die durchschnittlichen Residuen für jeden Quadratmeterwert (Grafik e). Die rechte Spalte zeigt entsprechend den geschätzten nichtlinearen Zusammenhang* $\widehat{mieteqm}_i = 9.228 + 273.348 \cdot 1/flaeche_i$ *(Grafik b) und die dazu gehörigen Residuen (Grafiken d und f).*

$\widehat{mieteqm} = 9.228 + 273.348 \cdot 1/flaeche$ in Abhängigkeit von der Wohnfläche eingezeichnet (durchgezogene Linie). Man beachte, dass es sich dabei um den (bedingten) Erwartungswert der Zielvariable *mieteqm* handelt, d.h. um den *durchschnittlichen* Zusammenhang zwischen Nettomiete pro qm und Wohnfläche. Zusätzlich enthält die Abbildung die Be-

obachtungen $(mieteqm_i, flaeche_i)$. Mit Hilfe der Grafik lassen sich die Schätzergebnisse leicht interpretieren. Insgesamt nimmt die durchschnittliche Nettomiete pro qm mit größer werdenden Wohnungen ab. Für kleine Wohnungen bis ca. 40-45 qm ist die Abnahme am stärksten, anschließend flacht die Kurve zunehmend ab. Wohnungen ab einer Wohnfläche von etwa 100 qm und mehr haben annähernd dieselbe durchschnittliche Nettomiete pro qm. Ein Blick auf die Residuen $\hat{\varepsilon}_i = mieteqm_i - 9.228 - 273.348 \cdot 1/flaeche_i$ in Abbildung 3.8 d) bzw. die durchschnittlichen Residuen in Abbildung 3.8 f) zeigt, dass die gewählte nichtlineare Modellierung eine bessere Anpassung an die Daten liefert. Selbstverständlich sind auch andere Transformationen als $f(flaeche_i) = 1/flaeche_i$ denkbar, etwa $f(flaeche_i) = \log(flaeche_i)$. Häufig führen unterschiedliche Transformationen sogar zu sehr ähnlichen Schätzergebnissen. Bei der Wahl einer geeigneten Transformation gibt es also kein Patentrezept, so dass man auf die Inspektion der Streudiagramme bzw. der Residuen angewiesen ist, um vernünftige Variablentransformationen zu finden.

△

Eine weitere einfache Variante zur Spezifikation von nichtlinearen Beziehungen bietet die Modellierung durch Polynome. Hier approximieren wir den Effekt einer metrischen Kovariable durch ein Polynom. Wir illustrieren die Vorgehensweise wieder anhand der Mietspiegeldaten.

Beispiel 3.4 Mietspiegel – Polynomiale Regression

Bei Verwendung eines Polynoms 2. Grades erhalten wir

$$mieteqm_i = \beta_0 + \beta_1 \cdot flaeche_i + \beta_2 \cdot flaeche_i^2 + \varepsilon_i.$$

Ein Polynom 3. Grades führt zu

$$mieteqm_i = \beta_0 + \beta_1 \cdot flaeche_i + \beta_2 \cdot flaeche_i^2 + \beta_3 \cdot flaeche_i^3 + \varepsilon_i.$$

In der Regel werden Polynome niedrigen Grades verwendet. Polynomgrade größer als drei werden selten benutzt, da die resultierenden Schätzungen vor allem am Rand des Wertebereichs der Kovariable häufig sehr instabil sind und eine große Variabilität aufweisen.

Ähnlich wie bei der Modellierung durch einfache Variablentransformation können die resultierenden nichtlinearen Modelle in lineare Modelle überführt werden. Durch Definition der neuen Variablen $x_{i1} = flaeche_i$, $x_{i2} = flaeche_i^2$ und $x_{i3} = flaeche_i^3$ erhalten wir

$$mieteqm_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \varepsilon_i$$

bei Verwendung eines Polynoms 2. Grades und

$$mieteqm_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \varepsilon_i$$

für ein Polynom 3. Grades. Die jeweiligen Designmatrizen sind gegeben durch

$$\boldsymbol{X} = \begin{pmatrix} 1 & 30 & 30^2 \\ 1 & 37 & 37^2 \\ \vdots & \vdots & \vdots \\ 1 & 73 & 73^2 \\ 1 & 73 & 73^2 \end{pmatrix} = \begin{pmatrix} 1 & 30 & 900 \\ 1 & 37 & 1369 \\ \vdots & \vdots & \vdots \\ 1 & 73 & 5329 \\ 1 & 73 & 5329 \end{pmatrix}$$

bzw.

$$\boldsymbol{X} = \begin{pmatrix} 1 & 30 & 30^2 & 30^3 \\ 1 & 37 & 37^2 & 37^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 73 & 73^2 & 73^3 \\ 1 & 73 & 73^2 & 73^3 \end{pmatrix} = \begin{pmatrix} 1 & 30 & 900 & 27000 \\ 1 & 37 & 1369 & 50653 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 73 & 5329 & 389017 \\ 1 & 73 & 5329 & 389017 \end{pmatrix}.$$

Als Schätzungen erhalten wir

$$\begin{aligned} \widehat{mieteqm}_i &= 23.07 - 0.206 \cdot x_{i1} + 0.00092 \cdot x_{i2} \\ &= 23.07 - 0.206 \cdot \mathit{flaeche}_i + 0.00092 \cdot \mathit{flaeche}_i^2 \end{aligned}$$

bei Verwendung eines Polynoms 2. Grades und

$$\begin{aligned} \widehat{mieteqm}_i &= 27.838 - 0.424 \cdot x_{i1} + 0.0039 \cdot x_{i2} - 0.000012 \cdot x_{i3} \\ &= 27.838 - 0.424 \cdot \mathit{flaeche}_i + 0.0039 \cdot \mathit{flaeche}_i^2 - 0.000012 \cdot \mathit{flaeche}_i^3 \end{aligned}$$

bei Verwendung eines Polynoms 3. Grades.

Zur Interpretation der Ergebnisse visualisieren wir wieder die geschätzten Effekte. Abbildung 3.9 zeigt in Grafik a) den geschätzten durchschnittlichen Zusammenhang zwischen Nettomiete pro qm und der Wohnfläche für ein Polynom 2. Grades. Zusätzlich eingezeichnet sind wieder die Beobachtungen. Die zum geschätzten Modell passenden Residuen findet man in Abbildung c). Entsprechende Ergebnisse bei Verwendung eines Polynoms 3. Grades sind in den Abbildungen b) und d) abgedruckt. Wir erhalten ähnliche Ergebnisse wie bei Verwendung der Variablentransformation $f(\mathit{flaeche}_i) = 1/\mathit{flaeche}_i$. Welches der Modelle geeigneter ist, kann durch einfache grafische Hilfsmittel nicht entschieden werden. Hier benötigen wir Verfahren zur Modellwahl, vergleiche Abschnitt 3.6.

△

Bisher haben wir bei unseren Überlegungen nur eine Variable berücksichtigt. Die Erweiterung auf mehrere Kovariablen ist aber jederzeit möglich, wie wir wieder anhand der Mietspiegeldaten demonstrieren.

Beispiel 3.5 Mietspiegel – Additive Modelle

Zunächst können den Modellen natürlich Variablen hinzugefügt werden, deren Einfluss linear modelliert wird. Beispielsweise können wir Modell (3.4) durch die Hinzunahme des Baujahrs der Wohnung erweitern und erhalten

$$\mathit{mieteqm}_i = \beta_0 + \beta_1 \cdot f(\mathit{flaeche}_i) + \beta_2 \cdot \mathit{bjahr}_i + \varepsilon_i. \tag{3.5}$$

Unter Verwendung der Transformation $f(\mathit{flaeche}) = 1/\mathit{flaeche}$ gewinnen wir dann die Schätzung

$$\widehat{mieteqm}_i = -127.54 + 232.75 \cdot 1/\mathit{flaeche}_i + 0.07 \cdot \mathit{bjahr}_i.$$

Alternativ zu einer linearen Modellierung in (3.5) können wir auch den Effekt des Baujahrs nichtlinear schätzen. Eine Modellierung durch ein Polynom dritten Grades liefert beispielsweise

$$\mathit{mieteqm}_i = \beta_0 + \beta_1 \cdot 1/\mathit{flaeche}_i + \beta_2 \cdot \mathit{bjahr}_i + \beta_3 \cdot \mathit{bjahr}_i^2 + \beta_4 \cdot \mathit{bjahr}_i^3 + \varepsilon_i. \tag{3.6}$$

Wir erhalten die Schätzung

$$\widehat{mieteqm}_i = 56771.52 + 252.66 \cdot 1/\mathit{flaeche}_i - 83.53 \cdot \mathit{bjahr}_i + 0.041 \cdot \mathit{bjahr}_i^2 - 0.0000066 \cdot \mathit{bjahr}_i^3.$$

Abb. 3.9. *Illustration zur Modellierung nichtlinearer Beziehungen durch Polynome am Beispiel der Mietspiegeldaten. Die linke Spalte zeigt für ein quadratisches Polynom den geschätzten durchschnittlichen Zusammenhang inklusive Beobachtungen (Grafik a) und die dazu gehörigen Residuen (Grafik c). Die rechte Spalte zeigt entsprechend den geschätzten nichtlinearen Zusammenhang für ein Polynom 3. Grades.*

Fassen wir die Effekte der Wohnfläche und des Baujahrs in den Funktionen

$$f_1(\mathit{flaeche}) = \beta_1 \cdot 1/\mathit{flaeche}$$

und

$$f_2(\mathit{bjahr}) = \beta_2 \cdot \mathit{bjahr} + \beta_3 \cdot \mathit{bjahr}^2 + \beta_4 \cdot \mathit{bjahr}^3$$

zusammen, so erhalten wir

$$\mathit{mieteqm}_i = \beta_0 + f_1(\mathit{flaeche}) + f_2(\mathit{bjahr}) + \varepsilon_i.$$

Die zu f_1 und f_2 gehörigen Schätzungen sind dann gegeben durch

$$\hat{f}_1(\mathit{flaeche}) = 252.66 \cdot 1/\mathit{flaeche}$$

bzw.

$$\hat{f}_2(\mathit{bjahr}) = -83.53 \cdot \mathit{bjahr} + 0.041 \cdot \mathit{bjahr}^2 - 0.0000066 \cdot \mathit{bjahr}^3.$$

Damit wird klar, dass es sich bei (3.6) um ein additives Modell handelt, wie in Kapitel 2 in Abschnitt 2.6 eingeführt.

Die Interpretation der Effekte ist wieder am einfachsten durch Visualisierung möglich. Dazu tragen wir die geschätzten Effekte $\hat{f}_1(\mathit{flaeche})$ und $\hat{f}_2(\mathit{bjahr})$ gegen die Wohnfläche

Abb. 3.10. *Visualisierung des nichtlinearen Effekts $\hat{f}_2(bjahr)$ des Baujahrs.*

bzw. das Baujahr ab. Abbildung 3.10 a) zeigt exemplarisch den Baujahrseffekt. Dabei fällt zunächst die Beschriftung der Ordinate mit Werten zwischen -56764 und -56756 auf. Der Grund für die ungewöhnlich erscheinenden Zahlen liegt in der Skala der Variable Baujahr mit Werten zwischen 1918 und 1998 begründet. Wenn wir das Baujahr in den Werten $18, \ldots, 98$ angegeben (d.h. jeweils 1900 von den ursprünglichen Werten subtrahieren), erhalten wir die Schätzung

$$\widehat{mieteqm}_i = 10.56 + 252.66 \cdot 1/flaeche_i - 0.184 \cdot bjahr_i + 0.003 \cdot bjahr_i^2 - 0.0000067 \cdot bjahr_i^3.$$

Auf den ersten Blick scheint sich der geschätzte Baujahrseffekt durch die Umkodierung vom ursprünglichen Effekt deutlich zu unterscheiden. Abbildung 3.10 b) zeigt jedoch, dass der Effekt lediglich auf der Ordinate verschoben ist und ansonsten unverändert bleibt. Tatsächlich ist das Niveau der nichtlinearen Funktionen nicht identifizierbar. Addiert man beispielsweise zu $\hat{f}_2$ eine beliebige Konstante c und subtrahiert diese gleichzeitig von $\hat{\beta}_0$, so bleibt die geschätzte Miete pro qm $\widehat{mieteqm}$ unverändert. Das Niveau der nichtlinearen Funktionen kann also beispielsweise durch Transformation der erklärenden Variablen beliebig verschoben werden.

Um eine leichtere Interpretierbarkeit zu gewährleisten, ist es oft zweckmäßig, zusätzlich zu fordern, dass sämtliche Funktionen „*um Null zentriert*“ sind. Diese Bedingung ist automatisch erfüllt, wenn die Einflussvariablen um Null zentriert werden. Im Falle der Wohnfläche würde man also die Variable $flaecheinv = 1/flaeche$ ersetzen durch die zentrierte Variable $flaecheinvz = flaecheinv - \overline{flaecheinv}$, wobei $\overline{flaecheinv}$ das arithmetische Mittel der Variable *flaecheinv* ist. Analog könnte man beim Baujahr vorgehen, d.h. die Variablen $bjahr$, $bjahr2 = bjahr^2$ und $bjahr3 = bjahr^3$ werden ersetzt durch $bjahrz = bjahr - \overline{bjahr}$, $bjahr2z = bjahr2 - \overline{bjahr2}$ und $bjahr3z = bjahr3 - \overline{bjahr3}$.

Im Falle von Polynomen können wir noch einen Schritt weiter gehen durch die Verwendung von *orthogonalen Polynomen*. Hier wird die übliche Basis $bjahr, bjahr^2, bjahr^3$ zur Darstellung von Polynomen ersetzt durch orthogonale Basisfunktionen. Dies bedeutet insbesondere, dass die zu den Variablen $bjahr$, $bjahr^2$ und $bjahr^3$ gehörenden Spalten der Designmatrix jeweils zentriert und zueinander orthogonal sind, d.h. die Vektoren stehen senkrecht aufeinander. Orthogonale Polynome haben neben der gewünschten Zentrierungseigenschaft weitere wünschenswerte Eigenschaften. Beispielsweise ist die Berechnung der KQ-Schätzer numerisch stabiler. In der Regel erfolgt die Erzeugung orthogonaler Polynome in größeren Statistikprogrammen durch automatisierte Funktionen (in STATA etwa durch die Funktion `orthpoly`), so dass eine eigenständige Berechnung nicht notwendig ist. Wie orthogonale Polynome durch Anwendung von Eigenschaften der KQ-Methode leicht erzeugt werden können, zeigen wir in Beispiel 3.9.

Abb. 3.11. *Visualisierung der nichtlinearen zentrierten Effekte der Wohnfläche und des Baujahrs inklusive partielle Residuen.*

Mit Hilfe der orthogonalen Variablen *bjahro*, *bjahro2* und *bjahro3* und der zentrierten Variable *flaecheinvz* erhalten wir die Schätzung

$$\widehat{mieteqm}_i = 13.87 + 252.66 \cdot \mathit{flaecheinvz}_i - 1.54 \cdot \mathit{bjahro}_i + 0.97 \cdot \mathit{bjahro2}_i - 0.062 \cdot \mathit{bjahro3}_i.$$

und die in Abbildung 3.11 abgedruckten Effekte $\hat{f}_1(\mathit{flaeche})$ und $\hat{f}_2(\mathit{bjahr})$. Zusätzlich eingezeichnet sind die partiellen Residuen $\hat{\varepsilon}_{\mathit{flaeche},i}$ und $\hat{\varepsilon}_{\mathit{bjahr},i}$. Diese sind gegeben durch

$$\hat{\varepsilon}_{\mathit{flaeche},i} = \mathit{mieteqm}_i - \hat{\beta}_0 - \hat{f}_2(\mathit{bjahr}_i) = \hat{\varepsilon}_i + \hat{f}_1(\mathit{flaeche}_i)$$

und

$$\hat{\varepsilon}_{\mathit{bjahr},i} = \mathit{mieteqm}_i - \hat{\beta}_0 - \hat{f}_1(\mathit{flaeche}_i) = \hat{\varepsilon}_i + \hat{f}_2(\mathit{bjahr}_i).$$

Bei den partiellen Residuen $\hat{\varepsilon}_{\mathit{flaeche},i}$ handelt es sich um die Reststreuung, aus der die Einflüsse aller Variablen (hier nur des Baujahrs) bis auf den Einfluss der Wohnfläche eliminiert sind. Analog ist bei $\hat{\varepsilon}_{\mathit{bjahr},i}$ der Einfluss der Wohnfläche, jedoch nicht der Einfluss des Baujahrs herausgerechnet. Durch zusätzliches Einzeichnen der partiellen Residuen in Abbildung 3.11 können wir überprüfen, ob die gewählte nichtlineare Modellierung adäquat ist.

Die Effekte der Wohnfläche und des Baujahrs können nun wie folgt interpretiert werden: Der gezeichnete Effekt der Wohnfläche gibt an, wie sich die durchschnittliche Nettomiete pro Quadratmeter in Abhängigkeit von der Größe der Wohnung verändert, wenn alle anderen Einflussgrößen (in unserem Fall nur das Baujahr) gleich bleiben. Beispielsweise erhalten wir für Wohnungen mit 40 und 60 Quadratmetern Wohnfläche die Effekte $\hat{f}_1(40) = 252.66 \cdot 1/40 = 6.32$ und $\hat{f}_1(60) = 252.66 \cdot 1/60 = 4.21$. Wohnungen mit 60 Quadratmetern Wohnfläche sind also im Durchschnitt circa 2.11 DM billiger als Wohnungen mit 40 Quadratmetern, vorausgesetzt sie sind im selben Jahr erbaut. Vergleichen wir Wohnungen mit 60 und 80 Quadratmeter Wohnfläche, so ist der Unterschied nur noch 1.05 DM. Obwohl die Differenz in beiden Vergleichen 20 Quadratmeter ist, halbiert sich der Unterschied in der durchschnittlichen Nettomiete. Der Grund hierfür ist der deutlich nichtlineare Effekt, der mit zunehmender Wohnfläche immer flacher verläuft. Völlig analog kann der Effekt des Baujahrs interpretiert werden. Wohnungen, die vor dem zweiten Weltkrieg gebaut wurden, sind ungefähr gleich teuer (bei gleicher Wohnfläche). Ab dem Jahr 1945 erbaute Wohnungen werden annähernd linear teuerer.

△

Modellierung nichtlinearer Zusammenhänge durch Variablentransformation

Besitzt die metrische Kovariable z (approximativ) den nichtlinearen Einfluss $\beta_1 f(z)$ mit bekannter Transformation f, so kann das Regressionsmodell

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

durch Definition der Variable $x_i = f(z_i) - \bar{f}$ in ein gewöhnliches multiples Regressionsmodell

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \varepsilon_i$$

überführt werden. Durch die Subtraktion von

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} f(z_i)$$

ist der geschätzte Effekt $\hat{\beta}_1 x$ automatisch um Null zentriert. Die Schätzergebnisse können am besten durch grafische Darstellung des geschätzten Effekts $\hat{\beta}_1 x$ in Abhängigkeit von z interpretiert werden.

Modellierung nichtlinearer Zusammenhänge durch Polynome

Besitzt die metrische Kovariable z (approximativ) einen polynomialen Einfluss $\beta_1 z + \beta_2 z^2 + \ldots + \beta_l z^l$ mit Polynomgrad l, so kann das Regressionsmodell

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \ldots + \beta_l z^l + \ldots + \varepsilon_i$$

durch Definition der neuen Variablen $x_{i1} = z_i$, $x_{i2} = z_i^2$, ..., $x_{il} = z_i^l$ in ein gewöhnliches multiples Regressionsmodell

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \ldots + \beta_l x_{il} + \ldots + \varepsilon_i$$

überführt werden. Durch Zentrierung (und eventuelle Orthogonalisierung) der Vektoren $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^l$ zu $\boldsymbol{x}^1 - \bar{\boldsymbol{x}}_1, \ldots, \boldsymbol{x}^l - \bar{\boldsymbol{x}}_l$ mit den Mittelwertvektoren $\bar{\boldsymbol{x}}_j = (\bar{x}_j, \ldots, \bar{x}_j)'$ wird die Interpretation der geschätzten Effekte erleichtert. Die Schätzergebnisse können am besten durch grafische Darstellung des geschätzten durchschnittlichen Zusammenhangs in Abhängigkeit von z interpretiert werden.

Kategoriale Kovariablen

Bisher haben wir uns mit der Modellierung des Einflusses metrischer Kovariablen befasst. In diesem Abschnitt beschäftigen wir uns mit kategorialen Kovariablen und deren Besonderheiten. Wir illustrieren unser Vorgehen wieder anhand des Mietspiegeldatensatzes. Speziell befassen wir uns mit der Modellierung der Variable *lage* mit den drei Kategorien 1 = normale Lage, 2 = gute Lage, 3 = beste Lage. Es stellt sich die Frage,

wie diese drei-kategoriale Variable geeignet in einem Regressionsmodell als erklärende Variable berücksichtigt werden kann. Wir gehen zunächst völlig naiv vor und behandeln die Wohnlage wie eine metrische Variable. Wenn wir der Einfachheit halber nur die Wohnlage als Kovariable berücksichtigen, erhalten wir das Modell

$$mieteqm_i = \beta_0 + \beta_1 \cdot lage_i + \varepsilon_i$$

und mit Hilfe der KQ-Methode die Schätzung

$$\widehat{mieteqm}_i = 12.76 + 0.77 \cdot lage.$$

Aufgrund der gewählten Kodierung der Variable *lage* wäre bei dieser Vorgehensweise der Effekt einer guten Wohnlage doppelt so hoch wie der einer normalen Lage (0.77 DM versus $2 \cdot 0.77 = 1.54$ DM). Entsprechend wäre der Effekt der besten Wohnlage dreimal so hoch (0.77 DM versus $3 \cdot 0.77 = 2.31$ DM). Hätten wir die Wohnlage mit $2 =$ normale Lage, $6 =$ gute Lage, $8 =$ beste Lage kodiert, so würden Wohnungen in guter Lage drei mal so teuer und Wohnungen in bester Lage vier mal so teuer wie in normaler Wohnlage geschätzt werden. Die Schätzergebnisse hängen also bei dieser naiven Vorgehensweise von der willkürlich gewählten Kodierung der kategorialen Variable ab. Das Problem dabei ist, dass die Abstände zwischen den Kategorien nicht sinnvoll interpretierbar sind. Eine gute Wohnlage ist nicht doppelt so gut (oder dreimal so gut bei der zweiten Kodierung) wie eine normale Wohnlage. Aus diesem Grund ist es zweckmäßig, für jede Kategorie eine neue Kovariable in Null/Eins Form, eine sogenannte Dummy-Variable, zu definieren und für jede Kategorie einen separaten Effekt zu schätzen. Im Falle der Wohnlage definieren wir also die drei neuen Dummy-Variablen

$$nlage = \begin{cases} 1 & lage = 1 \text{ (normale Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

$$glage = \begin{cases} 1 & lage = 2 \text{ (gute Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

$$blage = \begin{cases} 1 & lage = 3 \text{ (beste Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

und erhalten das Modell

$$mieteqm_i = \beta_0 + \beta_1 \cdot nlage_i + \beta_2 \cdot glage_i + \beta_3 \cdot blage_i + \varepsilon_i.$$

Bei genauerem Hinsehen ergibt sich bei dieser Vorgehensweise eine weitere Schwierigkeit: Die Regressionsparameter sind nicht identifizierbar. Um uns das Identifizierbarkeitsproblem klar zu machen, bestimmen wir den Effekt der Wohnlage für die drei Lagekategorien. Für normale Wohnlagen ($nlage_i = 1$, $glage_i = 0$, $blage_i = 0$) erhalten wir den Effekt $\beta_0 + \beta_1$. Für gute und beste Wohnlagen ergeben sich entsprechend die Effekte $\beta_0 + \beta_2$ und $\beta_0 + \beta_3$. Wenn wir zur Konstante β_0 einen beliebigen Wert c hinzu addieren und gleichzeitig von β_1, β_2 und β_3 abziehen, erhalten wir in der Summe dieselben Effekte. Die Regressionsparameter sind also nicht eindeutig bestimmt. Mit Hilfe der KQ-Methode würden wir daher auch keine eindeutigen Schätzungen erhalten. Dieses Identifizierbarkeitsproblem lässt sich dadurch lösen, dass wir entweder ein Modell ohne Konstante verwenden oder eine der Dummy-Variablen *nlage*, *glage* oder *blage* aus dem Modell entfernen. Da

normalerweise eine Konstante im Modell enthalten sein soll, beschränken wir uns auf die letztgenannte Variante. Durch Nichtberücksichtigung der Variable *glage* erhalten wir das Modell

$$mieteqm_i = \beta_0 + \beta_1 \cdot nlage_i + \beta_2 \cdot blage_i + \varepsilon_i.$$

Damit ergibt sich für normale Wohnlagen der Effekt $\beta_0 + \beta_1$, für gute Wohnlagen der Effekt β_0 und für beste Wohnlagen der Effekt $\beta_0 + \beta_2$. Die Regressionsparameter sind jetzt eindeutig bestimmt, also identifizierbar. Mit Hilfe der KQ-Methode ergibt sich die Schätzung

$$\widehat{mieteqm}_i = 14.18 - 0.62 \cdot nlage_i + 1.75 \cdot blage_i.$$

Die Regressionsparameter für die normale und beste Wohnlage werden durch *Vergleich* mit der nicht berücksichtigten Kategorie, der guten Wohnlage, interpretiert. Für normale Wohnlagen ergibt sich der geschätzte Effekt $\hat{\beta}_0 + \hat{\beta}_1 = 14.18 - 0.62 = 13.56$, für gute Lagen der Effekt $\hat{\beta}_0 = 14.18$ und für beste Wohnlagen der Effekt $\hat{\beta}_0 + \hat{\beta}_2 = 14.18 + 1.75 = 15.93$. *Im Vergleich* zu Wohnungen in guter Lage sind also Wohnungen in normaler Wohnlage 0.62 DM billiger und Wohnungen in bester Lage 1.75 DM teurer. Die zum Vergleich herangezogene Wohnlagenkategorie wird als Referenzkategorie bezeichnet. Je nach gewählter Referenzkategorie ändert sich die Interpretation der geschätzten Regressionskoeffizienten. Es sei aber darauf hingewiesen, dass die jeweiligen Modelle äquivalent sind, d.h. die geschätzten Regressionskoeffizienten lassen sich bei veränderter Kodierung aus dem ursprünglichen Modell berechnen. Unter Verwendung der normalen Lage als Referenzkategorie erhalten wir die Schätzung

$$\widehat{mieteqm}_i = 13.56 + 0.62 \cdot glage_i + 2.37 \cdot blage_i.$$

Im Vergleich zu Wohnungen in normaler Lage sind Wohnungen in guter Lage um 0.62 DM teurer und Wohnungen in bester Lage sogar um 2.37 DM teurer.

Abschließend stellt sich noch die Frage, welche Kategorie einer kategorialen Variable als Referenzkategorie gewählt werden soll. Prinzipiell kommt jede Kategorie in Frage. Üblicherweise wird diejenige Kategorie gewählt, die als Vergleichsmaßstab inhaltlich am sinnvollsten erscheint, etwa die im Datensatz am häufigsten vorkommende Kategorie.

Für eine beliebige kategoriale Kovariable x mit c Kategorien fassen wir die Modellierung mit Hilfe von Dummy-Variablen im folgenden Kasten zusammen.

Die soeben beschriebene Kodierung ist keineswegs die einzig denkbare Kodierung kategorialer Variablen. Neben der Dummy-Kodierung ist vor allem noch die sogenannte *Effekt-Kodierung* verbreitet. Diese ist definiert durch

$$x_{i1} = \begin{cases} 1 & x_i = 1 \\ -1 & x_i = c \\ 0 & \text{sonst} \end{cases} \quad \ldots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c-1 \\ -1 & x_i = c \\ 0 & \text{sonst.} \end{cases}$$

Im Unterschied zur Dummy-Kodierung werden also bei Vorliegen der Referenzkategorie die neu erzeugten Variablen mit -1 kodiert. Wir illustrieren kurz den Unterschied zwischen beiden Kodierungsvarianten anhand des Mietspiegeldatensatzes.

Beispiel 3.6 Mietspiegel – Effekt-Kodierung

Wenn wir normale Wohnlagen als Referenzkategorie wählen, erhalten wir in Effekt-Kodierung die beiden Variablen

Dummy-Kodierung bei kategorialen Kovariablen

Zur Modellierung des Effekts einer c-kategorialen Kovariable $x \in \{1, \ldots, c\}$ mit Hilfe der Dummy-Kodierung werden die $c-1$ Dummy-Variablen

$$x_{i1} = \begin{cases} 1 & x_i = 1 \\ 0 & \text{sonst} \end{cases} \quad \ldots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c-1 \\ 0 & \text{sonst} \end{cases}$$

für $i = 1, \ldots, n$ definiert und als erklärende Variablen ins Regressionsmodell aufgenommen:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{i,c-1} x_{i,c-1} + \ldots + \varepsilon_i$$

Aus Gründen der Identifizierbarkeit wird für eine Kategorie von x, hier die c-te Kategorie, keine Dummy-Variable ins Modell mit aufgenommen. Diese Kategorie wird als Referenzkategorie bezeichnet. Die Schätzergebnisse werden dann jeweils im Vergleich zu der weggelassenen Kategorie interpretiert.

$$\mathit{glage} = \begin{cases} 1 & \mathit{lage} = 2 \text{ (gute Wohnlage)} \\ -1 & \mathit{lage} = 1 \text{ (normale Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

$$\mathit{blage} = \begin{cases} 1 & \mathit{lage} = 3 \text{ (beste Wohnlage)} \\ -1 & \mathit{lage} = 1 \text{ (normale Wohnlage)} \\ 0 & \text{sonst.} \end{cases}$$

Für das Regressionsmodell

$$\mathit{mieteqm}_i = \beta_0 + \beta_1 \cdot \mathit{glage}_i + \beta_2 \cdot \mathit{blage}_i + \varepsilon_i$$

ergibt sich dann die Schätzung

$$\widehat{\mathit{mieteqm}}_i = 14.56 - 0.38 \cdot \mathit{glage}_i + 1.37 \cdot \mathit{blage}_i.$$

Aufgrund der speziellen Kodierung kann bei Variablen in Effekt-Kodierung für die Referenzkategorie ein *eigener Regressionsparameter* bestimmt werden. Dieser kann als die negative Summe der übrigen Parameter berechnet werden, also $\hat{\beta}_3 = -\hat{\beta}_1 - \hat{\beta}_2 = 0.38 - 1.37 = -0.99$. Damit ergeben sich für normale, gute und beste Lage die (geschätzten) Effekte $\hat{\beta}_0 + \hat{\beta}_3 = 14.56 - 0.99 = 13.57$, $\hat{\beta}_0 + \hat{\beta}_1 = 14.56 - 0.38 = 14.18$ und $\hat{\beta}_0 + \hat{\beta}_2 = 14.56 + 1.37 = 15.93$.

△

Interaktionen zwischen Kovariablen

Interaktionen zwischen Kovariablen treten immer dann auf, wenn der Effekt einer Kovariable vom Wert mindestens einer anderen Kovariable abhängt. Wir beginnen mit dem einfachst möglichen Fall und betrachten das Regressionsmodell

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x z + \varepsilon \tag{3.7}$$

zwischen der Zielvariable y und den beiden erklärenden Variablen x und z. Der Term $\beta_3 x z$ heißt *Interaktion* zwischen x und z. Die nur von jeweils einer der Variablen

abhängigen Terme $\beta_1\, x$ und $\beta_2\, z$ werden *Haupteffekte* genannt. Die Bedeutung des Interaktionsterms machen wir uns am besten klar, indem wir überlegen wie sich $\mathrm{E}(y)$ ändert, wenn eine der Variablen, z.B. x, sich um den Betrag d ändert. Wir erhalten

$$\begin{aligned}\mathrm{E}(y \,|\, x+d, z) - \mathrm{E}(y \,|\, x, z) &= \beta_0 + \beta_1\,(x+d) + \beta_2\, z + \beta_3\,(x+d)\, z - \\ &\quad -\beta_0 - \beta_1\, x - \beta_2\, z - \beta_3\, x\, z \\ &= \beta_1\, d + \beta_3\, d\, z.\end{aligned}$$

Damit können wir die zwei Fälle $\beta_3 = 0$ und $\beta_3 \neq 0$ unterscheiden:

- Falls $\beta_3 = 0$, fällt die Interaktion aus dem Modell heraus und die erwartete Änderung $\beta_1\, d$ ist *unabhängig* vom speziellen Wert der zweiten Kovariable z.
- Falls $\beta_3 \neq 0$, hängt die erwartete Änderung $\beta_1\, d + \beta_3\, d\, z$ nicht nur von der Wertänderung d von x ab, sondern auch vom Wert der Kovariable z.

Einen Interaktionsterm benötigen wir daher immer dann, wenn die Auswirkung der Veränderung einer Variable zusätzlich von dem Wert einer anderen Variable abhängt. Die spezielle Modellierung von Interaktionen hängt von den Typen der beteiligten Variablen ab. Im Folgenden wollen wir die Modellierung von Interaktionen zwischen zwei kategorialen Variablen, einer metrischen und einer kategorialen Variable sowie zwei metrischen Variablen diskutieren.

Interaktionen zwischen kategorialen Variablen

Wir betrachten zunächst den einfachsten Fall einer Interaktion zwischen zwei binären Variablen x und z. Wenn wir an den Mietspiegel denken, könnten die Werte $x = 1$ bzw. $z = 1$ das Vorhandensein einer gehobenen Küche bzw. eines gehobenen Badezimmers bedeuten. In diesem Fall misst der Koeffizient β_1 in (3.7) den Effekt der gehobenen Küche auf die Nettomiete, β_2 misst den Effekt des gehobenen Badezimmers und β_3 den Zusatzeffekt, wenn gehobene Küche *und* gehobenes Bad gleichzeitig vorhanden sind. Je nachdem, welche Werte die Koeffizienten annehmen, ergeben sich unterschiedliche Interpretationen. Im vorliegenden Fall werden alle drei Koeffizienten höchstwahrscheinlich positiv sein wodurch sich folgende Interpretation ergibt: Sowohl das Vorhandensein einer gehobenen Küche ($x = 1$) als auch eines gehobenen Badezimmers ($z = 1$) erhöhen die durchschnittliche Nettomiete. Sind beide Ausstattungsmerkmale *gemeinsam* vorhanden, verstärkt sich der Effekt und die Miete wird zusätzlich verteuert.

Der Interaktionseffekt zwischen zwei binären Variablen misst also den Effekt, wenn die mit x und z assoziierten Eigenschaften *gemeinsam* auftreten. Im Fall zweier beliebiger kategorialer Variablen x und z mit c bzw. m Kategorien ist die Modellierung von Interaktionen aufwändiger. Wir illustrieren die Vorgehensweise im Falle zweier kategorialer Variablen mit jeweils drei Kategorien: Definiere die zu x gehörenden Dummy-Variablen x_1, x_2 und die zu z gehörenden Dummy-Variablen z_1, z_2. Als Referenzkategorie wählen wir jeweils die letzte Kategorie. Zur Modellierung der Interaktionseffekte müssen alle möglichen Kombinationen der Werte von x und z (außer die Referenzkategorien) beachtet werden, wir erzeugen also zusätzlich die Variablen $x_1 z_1$, $x_1 z_2$, $x_2 z_1$ und $x_2 z_2$. Damit erhalten wir das Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 x_1 z_1 + \beta_6 x_1 z_2 + \beta_7 x_2 z_1 + \beta_8 x_2 z_2 + \varepsilon.$$

Die Koeffizienten lassen sich dann wie folgt interpretieren:

β_1	Effekt bei Vorliegen von $x = 1$ und $z = 3$ (Referenz)
β_2	Effekt bei Vorliegen von $x = 2$ und $z = 3$ (Referenz)
β_3	Effekt bei Vorliegen von $x = 3$ (Referenz) und $z = 1$
β_4	Effekt bei Vorliegen von $x = 3$ (Referenz) und $z = 2$
$\beta_1 + \beta_3 + \beta_5$	Effekt bei Vorliegen von $x = 1$ und $z = 1$
$\beta_1 + \beta_4 + \beta_6$	Effekt bei Vorliegen von $x = 1$ und $z = 2$
$\beta_2 + \beta_3 + \beta_7$	Effekt bei Vorliegen von $x = 2$ und $z = 1$
$\beta_2 + \beta_4 + \beta_8$	Effekt bei Vorliegen von $x = 2$ und $z = 2$

Wie immer bei kategorialen Variablen sind diese Effekte als Differenz zu den Referenzkategorien $x = 3$ und $z = 3$ zu interpretieren. Beispielsweise misst $\beta_2 + \beta_4 + \beta_8$ den Effekt wenn $x = 2$ und $z = 2$ vorliegen im Vergleich zur Kombination $x = 3$ und $z = 3$.

Wir erkennen, dass die Modellierung der Interaktion zwischen x und z auch durch die Definition einer neuen kategorialen Variable w erfolgen kann, deren Kategorien aus allen möglichen Kombinationen der Werte von x und z bestehen:

$$w = \begin{cases} 1 & x = 1,\, z = 1 \\ 2 & x = 1,\, z = 2 \\ 3 & x = 1,\, z = 3 \\ \vdots & \vdots \qquad \vdots \\ 9 & x = 3,\, z = 3. \end{cases}$$

In vielen Fällen ist diese Vorgehensweise sogar die am besten interpretierbare Variante.

Beispiel 3.7 Mietspiegel – Interaktion mit Küchenausstattung

Da die für den Mietspiegel erhobenen Daten mit der Zeit durch Änderungen am Mietmarkt veralten, wird der Mietspiegel im Abstand von zwei Jahren aktualisiert. Im Zuge der Aktualisierung werden neue Daten erhoben. Aus Kostengründen wird jedoch nicht bei jeder Aktualisierung eine volle Erhebung (mit einem Stichprobenumfang von ca. 3000 Wohnungen) durchgeführt. So wurden bei der Aktualisierung des Mietspiegels im Jahr 2001 lediglich circa 1500 Wohnungen erhoben. Eine völlige Neugestaltung des Mietspiegels basierend auf dieser relativ kleinen Stichprobe konnte daher nicht vorgenommen werden. Stattdessen wurde folgende Vorgehensweise gewählt: Sämtliche Merkmale des Mietspiegel 1999 wurden für die Fortschreibung übernommen, d.h. die Struktur des Mietspiegels wurde nicht verändert. Bei sämtlichen Wohnungsmerkmalen wurde überprüft, ob sich im Vergleich zu 1999 Veränderungen ergeben haben. Zur Überprüfung wurden die beiden Datensätze für 1999 und 2001 gemeinsam analysiert und mit Hilfe von Interaktionen eventuelle Veränderungen untersucht. Wir illustrieren das Vorgehen anhand des Merkmals *kueche* mit den Kategorien „unterdurchschnittliche Küche“ (Referenzkategorie), „normale Küche“ (Dummy-Variable *nkueche*) und gehobene Küche (Dummy-Variable *gkueche*). Als Ausgangspunkt nehmen wir das in Beispiel 3.5 (Seite 76) entwickelte Modell (3.6). Dort wurde der Einfluss der Wohnfläche mit Hilfe der Transformation $1/\mathit{flaeche}$ und der Einfluss des Baujahrs durch ein Polynom dritten Grades modelliert. Wir erweitern dieses Modell um eine Interaktion zwischen der Küchenausstattung und dem Erhebungszeitpunkt und erhalten als Schätzergebnis

$$\widehat{\mathit{mieteqm}} = \cdots - 0.51\,\mathit{zeit01} + 1.79\,\mathit{nkueche} + 2.14\,\mathit{gkueche} +$$
$$0.79\,\mathit{nkueche} \cdot \mathit{zeit01} + 1.45\,\mathit{gkueche} \cdot \mathit{zeit01}.$$

Die Dummy-Variable *zeit01* gibt an, ob es sich um eine Beobachtung aus dem Jahr 2001 ($zeit01 = 1$) oder aus dem Jahr 1999 ($zeit01 = 0$) handelt.

Zur Interpretation gehen wir jeweils von Wohnungen mit gleicher Wohnfläche und gleichem Baujahr aus. Die Ergebnisse lassen sich dann wie folgt interpretieren:

- Wohnungen mit unterdurchschnittlicher Küche sind 2001 im Durchschnitt um 0.51 DM pro Quadratmeter billiger als 1999.
- Wohnungen mit normaler Küche sind im Durchschnitt um -0.51+0.79=0.28 DM teurer als 1999.
- Wohnungen mit gehobener Küche sind im Durchschnitt um -051+1.45=0.94 DM teurer als 1999.

Im Vergleich zum Originalmietspiegel sind die geschätzten Zuschläge für normale und gehobene Küchen sehr hoch. Dies liegt daran, dass in diesem illustrierenden Beispiel viele relevante Merkmale noch nicht berücksichtigt sind. Ob die Modellierung der Interaktion wirklich notwendig ist, lässt sich mit den bisherigen Mitteln nicht beantworten. Die hierzu benötigten Inferenzmethoden werden in den nächsten Abschnitten entwickelt, vergleiche insbesondere die Abschnitte 3.3 und 3.6.

△

Interaktionen zwischen metrischen und kategorialen Variablen

Ein Beispiel für die Modellierung einer Interaktion zwischen metrischen und kategorialen Variablen haben wir bereits in Beispiel 2.5 (Seite 26) in Kapitel 2 kennen gelernt. In diesem illustrierenden Beispiel wurden die seit 1966 erbauten Wohnungen in normaler und guter Wohnlage aus dem Datensatz herausgegriffen. Ziel war die Modellierung des Zusammenhangs der Nettomiete mit den erklärenden Variablen Wohnfläche und Wohnlage. Abbildung 2.3 (Seite 25) legt die Vermutung nahe, dass das Haupteffektemodell

$$miete = \beta_0 + \beta_1\, flaeche + \beta_2\, glage + \varepsilon$$

nicht adäquat ist, da die Regressionsgerade zwischen Nettomiete und Wohnfläche für Wohnungen in guter Lage ($glage = 1$) steiler verläuft. Das Haupteffektemodell wurde daher ersetzt durch

$$miete = \beta_0 + \beta_1\, flaeche + \beta_2\, glage + \beta_3\, inter + \varepsilon$$

mit der Interaktionsvariable $inter = flaeche \cdot glage$. Die darin enthaltenen Terme lassen sich wie folgt interpretieren:

- $\beta_1\, flaeche$: Effekt der Wohnfläche für Wohnungen in normaler Lage ($glage = 0$).
- $\beta_1\, flaeche + \beta_2 + \beta_3\, inter = \beta_2 + (\beta_1 + \beta_3) flaeche$: Effekt der Wohnfläche für Wohnungen in guter Lage ($glage = 1$). Die beiden linearen Wohnflächeneffekte für normale und gute Wohnlagen können am besten durch Visualisierung in einer gemeinsamen Grafik interpretiert werden, siehe Abbildung 2.6 (Seite 27).
- $\beta_2 + \beta_3\, inter = \beta_2 + \beta_3\, flaeche$: Differenzeffekt von Wohnungen in guter Lage im Vergleich zu Wohnungen in normaler Lage. Dieser Effekt variiert in Abhängigkeit von der Wohnfläche. Wegen der positiven Koeffizienten $\hat{\beta}_2 = 10.15$ und $\hat{\beta}_3 = 1.60$, sind Wohnungen in guter Lage stets teurer als Wohnungen in normaler Lage (bei gleicher Wohnfläche). Der Unterschied wird umso größer, je größer die Wohnfläche ist. Für jeden zusätzlichen Quadratmeter Wohnfläche erhöht sich der Unterschied zu Wohnungen in normaler Wohnlage um 1.1 DM.

In analoger Weise können Interaktionen für beliebige metrische und kategoriale Kovariablen x und z behandelt werden. Der Einfachheit halber gehen wir zunächst davon aus, dass $z \in \{0, 1\}$ binär ist. Wenn sowohl Haupteffekte als auch der Interaktionseffekt linear sind, erhalten wir das Modell

$$y = \beta_0 + \beta_1 z + \beta_2 x + \beta_3 xz + \ldots + \varepsilon.$$

Die Punkte deuten an, dass im Modell beliebige weitere Terme (anderer Kovariablen) enthalten sein können. Die im Modell enthaltenen Terme lassen sich sowohl mit Bezug zur metrischen Kovariable x als auch mit Bezug zur binären Variable z interpretieren. Mit Bezug auf x erhalten wir die folgende Interpretation:

- $\beta_2 x$: Linearer Effekt der metrischen Kovariable x, falls $z = 0$.
- $\beta_1 + (\beta_2 + \beta_3) x$: Linearer Effekt der metrischen Kovariable x, falls $z = 1$.

Umgekehrt ergibt sich folgende Interpretation mit Bezug auf z:

- $\beta_1 + \beta_3 x$: Differenzeffekt von Beobachtungen mit $z = 1$ zu Beobachtungen mit $z = 0$. Der Differenzeffekt variiert in Abhängigkeit von x (ist also nicht konstant wie bei Modellen ohne Interaktion).

Bisher sind wir davon ausgegangen, dass sowohl der Haupteffekt der metrischen Variable x als auch der Interaktionseffekt linear sind. Nichtlineare Modellierungen im Rahmen unserer bisherigen Möglichkeiten sind aber jederzeit möglich. Dazu schreiben wir das Modell in Abhängigkeit von Funktionen f_1 und f_2:

$$y = \beta_0 + \beta_1 z + f_1(x) + f_2(x) z + \ldots + \varepsilon. \tag{3.8}$$

Im Fall einer linearen Modellierung gilt $f_1(x) = \beta_2 x$ und $f_2(x) = \beta_3 x$. Nichtlineare Beziehungen können wieder durch Variablentransformation oder durch Polynome modelliert werden. Beispielsweise liefert die Verwendung von Polynomen 2. Grades die Funktionen $f_1(x) = \beta_2 x + \beta_3 x^2$ und $f_2(x) = \beta_4 x + \beta_5 x^2$ und damit das Modell

$$\begin{aligned} y &= \beta_0 + \beta_1 z + \beta_2 x + \beta_3 x^2 + (\beta_4 x + \beta_5 x^2) z + \varepsilon \\ &= \beta_0 + \beta_1 z + \beta_2 x + \beta_3 x^2 + \beta_4 xz + \beta_5 x^2 z + \varepsilon. \end{aligned}$$

Die Terme lassen sich analog zur rein linearen Modellierung wie folgt interpretieren:

- $f_1(x)$: Nichtlinearer Effekt der metrischen Kovariable x falls $z = 0$.
- $\beta_1 + f_1(x) + f_2(x)$: Nichtlinearer Effekt der metrischen Kovariable x falls $z = 1$. An der Gestalt der beiden Kurven $f_1(x)$ und $\beta_1 + f_1(x) + f_2(x)$ lässt sich grafisch ablesen, wie stark die Interaktion ist. Im Extremfall sind beide Kurven lediglich parallel verschoben. In diesem Fall liegt keine Interaktion vor.
- $\beta_1 + f_2(x)$: Nichtlinearer Differenzeffekt für Beobachtungen mit $z = 1$ im Vergleich zu Beobachtungen mit $z = 0$. Der Effekt variiert nichtlinear in Abhängigkeit von x. Wir sprechen in diesem Zusammenhang auch von einem Term mit variierenden Koeffizienten, weil der Effekt von z in Abhängigkeit einer weiteren Variable x variiert. Die den Effekt von z „modifizierende“ Variable x wird auch als Effektmodifizierer bezeichnet. In Kapitel 8.3.1 werden wir Verfahren kennen lernen, wie mögliche nichtlineare Effekte f_2 flexibel und automatisiert geschätzt werden können.

Zur Erleichterung der Interpretation ist es wieder zweckmäßig den Haupteffekt $f_1(x)$ zu zentrieren. Eine automatische Zentrierung wird durch Verwendung von zentrierten Designvektoren erreicht. Da die Funktion f_2 als *Differenz* zum Haupteffekt fungiert, muss diese nicht zentriert werden. Die geschätzten Effekte $\hat{f}_1$, $\hat{\beta}_1 + \hat{f}_1 + \hat{f}_2$ und $\hat{\beta}_1 + \hat{f}_2$ können gegen x abgetragen und auf diese Weise grafisch interpretiert werden. Zweckmäßigerweise werden die ersten beiden Effekte in eine gemeinsame Grafik eingezeichnet.

Beispiel 3.8 Mietspiegel – Interaktion zwischen Wohnfläche und Lage

Zur Illustration betrachten wir im Mietspiegeldatensatz nur die Wohnungen in normaler und bester Wohnlage und modellieren den Zusammenhang zwischen Nettomiete pro Quadratmeter und den erklärenden Variablen Wohnfläche und Lage. In vorangegangenen Beispielen (vergleiche etwa Beispiel 2.4 auf Seite 25) wurde der Einfluss der Wohnfläche mit Hilfe der Transformation $1/\mathit{flaeche}$ modelliert. Wenn wir eine lineare Interaktion zwischen Wohnfläche und Wohnlage unterstellen, erhalten wir das Modell

$$\begin{aligned} \mathit{mieteqm} &= \beta_0 + \beta_1\, \mathit{blage} + \beta_2\, \mathit{flaecheinvz} + \beta_3\, \mathit{flaeche} \cdot \mathit{blage} + \varepsilon \\ &= \beta_0 + \beta_1\, \mathit{blage} + f_1(\mathit{flaeche}) + f_2(\mathit{flaeche}) \cdot \mathit{blage} + \varepsilon, \end{aligned}$$

wobei *flaecheinvz* die um Null zentrierte Transformation $1/\mathit{flaeche}$ der Wohnfläche ist. Die Dummy-Variable *blage* ist gleich Eins für Wohnungen in bester Lage und Null sonst. Die automatisch um Null zentrierte Funktion $f_1(\mathit{flaeche}) = \beta_2\, \mathit{flaecheinvz}$ stellt den nichtlinearen Haupteffekt der Wohnfläche dar. Die Funktion $f_2(\mathit{flaeche}) = \beta_3\, \mathit{flaeche}$ ist der lineare Interaktionseffekt. Als Schätzung erhalten wir

$$\widehat{\mathit{mieteqm}} = 13.54 + 1.51\, \mathit{blage} + 279.09\, \mathit{flaecheinvz} + 0.02\, \mathit{flaeche} \cdot \mathit{blage}.$$

Zur Interpretation der Schätzergebnisse visualisieren wir

- den geschätzten Effekt $\hat{f}_1(\mathit{flaeche})$ der Wohnfläche für Wohnungen in normaler Lage (Abbildung 3.12 a), durchgezogene Linie),
- den Effekt $\hat{\beta}_1 + \hat{f}_1(\mathit{flaeche}) + \hat{f}_2(\mathit{flaeche})$ der Wohnfläche für Wohnungen in bester Wohnlage (Abbildung 3.12 a), gestrichelte Linie),
- den in Abhängigkeit von der Wohnfläche variierenden Effekt $\hat{\beta}_1 + \hat{f}_2(\mathit{flaeche})$ der besten Lage (Abbildung 3.12 b).

Die Grafiken lassen sich sowohl mit Bezug auf die Wohnfläche als auch mit Bezug auf die Wohnlage interpretieren:

- *Interpretation mit Bezug auf die Wohnfläche:* Die Effekte der Wohnfläche in normalen bzw. besten Lagen besitzen eine ähnliche Gestalt. Allerdings verlaufen die Kurven nicht parallel zueinander, ein gewisser Interaktionseffekt ist also vorhanden. Mit wachsender Wohnfläche sinkt die durchschnittliche Miete pro qm, wobei der Abfall zunächst steiler und ab einer Wohnfläche von circa 80 qm immer flacher verläuft. Wohnungen in bester Lage sind für jede Wohnungsgröße teurer als Wohnungen in normaler Wohnlage. Für kleine Wohnungen ist der Unterschied deutlich geringer als für große Wohnungen, d.h. mit wachsender Wohnfläche wächst die Differenz in der Nettomiete pro qm zwischen Wohnungen in normaler und bester Lage.
- *Interpretation mit Bezug auf die Lage:* Wohnungen in bester Lage sind unabhängig von der Wohnfläche teurer als Wohnungen in normaler Lage. Die Differenz nimmt mit wachsender Wohnfläche linear zu.

Die Frage, ob die Modellierung der Interaktion notwendig ist, kann hier nicht entschieden werden. Sie lässt sich mit den Inferenzmethoden in den folgenden Abschnitten, insbesondere mit Hilfe statistischer Tests (Abschnitt 3.3) und Methoden der Modellwahl (Abschnitt 3.6), beantworten. Dieses illustrierende Beispiel dient vorwiegend dazu, die Modellierung von Interaktionen und deren Interpretation zu veranschaulichen.

Abb. 3.12. *Grafik a: Darstellung der Wohnflächeneffekte für normale (durchgezogene Linie) und beste (gestrichelte Linie) Lagen. Grafik b: Die Abbildung visualisiert den in Abhängigkeit von der Wohnfläche variierenden Effekt der besten Lage.*

△

Im Folgenden illustrieren wir noch kurz den Fall einer Interaktion mit einer mehrkategorialen Variable. Für eine möglichst übersichtliche Notation, beschränken wir uns auf eine Variable mit drei Ausprägungen 1, 2 und 3. Die Übertragung auf mehr als drei Kategorien ist dann sehr einfach. Wir bezeichnen mit z_1 und z_2 die aus z gebildeten Dummy-Variablen mit $z_1 = 1$ falls $z = 1$ und $z_2 = 1$ falls $z = 2$. Als Referenzkategorie wählen wir die dritte Ausprägung $z = 3$. Analog zu (3.8) lässt sich ein Interaktionsmodell schreiben als

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + f_1(x) + f_2(x) z_1 + f_3(x) z_2 + \ldots + \varepsilon,$$

wobei f_1, f_2 und f_3 wieder lineare oder nichtlineare Funktionen sind. Das Modell enthält die folgenden Effekte:

- $f_1(x)$: Effekt der metrischen Variable x, falls $z = 3$ (Referenzkategorie).
- $\beta_1 + f_1(x) + f_2(x)$: Effekt der metrischen Variable x falls $z = 1$.
- $\beta_2 + f_1(x) + f_3(x)$: Effekt der metrischen Variable x falls $z = 2$.
- $\beta_1 + f_2(x)$: In Abhängigkeit von x variierender Effekt der Ausprägung $z = 1$. Wie immer bei kategorialen Variablen ist der Effekt als Differenz zur Referenzkategorie $z = 3$ zu interpretieren.
- $\beta_2 + f_3(x)$: In Abhängigkeit von x variierender Effekt der Ausprägung $z = 2$.

Die Schätzergebnisse können am einfachsten interpretiert werden, indem die geschätzten Effekte $\hat{f}_1(x)$, $\hat{\beta}_1 + \hat{f}_1(x) + \hat{f}_2(x)$ und $\hat{\beta}_2 + \hat{f}_1(x) + \hat{f}_3(x)$ gemeinsam in einer Grafik visualisiert werden. Analog können die variierenden Effekte $\hat{\beta}_1 + \hat{f}_2(x)$ und $\hat{\beta}_2 + \hat{f}_3(x)$ in einer Grafik gezeichnet und anschließend interpretiert werden.

Werden sämtliche Funktionen linear modelliert, so erhalten wir das lineare Modell

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 x + \beta_4 x z_1 + \beta_5 x z_2 + \ldots + \varepsilon.$$

Eine quadratische Modellierung ergibt

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 x + \beta_4 x^2 + \beta_5 x z_1 + \beta_6 x^2 z_1 + \beta_7 x z_2 + \beta_8 x^2 z_2 + \ldots + \varepsilon.$$

Interaktionen zwischen metrischen Variablen

Bei der Schätzung von Interaktionen zwischen zwei metrischen Kovariablen müssen zweidimensionale Funktionen, etwa durch mehrdimensionale Polynome, modelliert werden. Im Rahmen der linearen Modelle stößt man hier an die Grenze des Machbaren, da es allein durch Inspektion der (dreidimensionalen) Streudiagramme schwer bis unmöglich ist, geeignete Modellierungen zu finden. Hier sind automatisierte Verfahren deutlich überlegen und wir verweisen an dieser Stelle auf die Kapitel 7.2 und 8.3.2.

3.2 Parameterschätzungen

In diesem Abschnitt bestimmen wir Schätzungen und deren Eigenschaften für die unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 des linearen Modells. Der folgende Abschnitt 3.2.1 befasst sich mit der Schätzung der Regressionskoeffizienten $\boldsymbol{\beta}$. Im daran anschließenden Abschnitt 3.2.2 diskutieren wir Schätzungen für die Varianz σ^2. Im Abschnitt 3.2.3 untersuchen wir dann statistische Eigenschaften der vorher erhaltenen Schätzungen.

3.2.1 Schätzung der Regressionskoeffizienten

Die Schätztheorie für die Regressionskoeffizienten des linearen Modells ist eng mit der von Legendre bereits 1806 erfundenen Methode der kleinsten Quadrate (KQ-Methode) verknüpft. Neben der Methode der kleinsten Quadrate sind andere Konzepte zur Schätzung der Regressionskoeffizienten denkbar, insbesondere die Maximum-Likelihood-Methode. Wie wir sehen werden, stimmen aber die Maximum-Likelihood-Schätzer mit den KQ-Schätzern überein. Eine Alternative bieten sogenannte robuste Verfahren. Diese haben den Vorteil, dass Ausreißer weniger Einfluss auf die Schätzungen besitzen als bei der KQ-Methode.

Die Methode der kleinsten Quadrate

Gemäß dem Prinzip der kleinsten Quadrate erhalten wir eine Schätzung für die unbekannten Regressionskoeffizienten $\boldsymbol{\beta}$ durch Minimierung der *Summe der quadrierten Abweichungen*

$$\mathrm{KQ}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \tag{3.9}$$

bezüglich $\boldsymbol{\beta} \in \mathbb{R}^p$.

Eine zunächst näher liegende Vorgehensweise ist die Minimierung der Summe der *absoluten Abweichungen*, d.h. der Funktion

$$\mathrm{SM}(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| = \sum_{i=1}^{n} |\varepsilon_i|.$$

Tatsächlich ist diese Methode historisch gesehen deutlich älter als die Methode der kleinsten Quadrate. Sie wurde von Rudjer Joseph Boscovich (1711–1787) um 1760 zum ersten Mal vorgeschlagen. Die moderne Bezeichnung ist *Median-Regression*, weil als Lösung des

Minimierungsproblems der geschätzte Median der Zielgröße bei gegebenen Kovariablen resultiert. Die Median-Regression ist wiederum ein Spezialfall der sogenannten *Quantil-Regression*, bei der ein beliebiges Quantil $q \in (0,1)$ der Zielgröße in Abhängigkeit von den Kovariablen geschätzt werden soll. Es lässt sich zeigen, dass diese Forderung äquivalent ist zur Minimierung des Kriteriums

$$\mathrm{SQ}(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| h_i,$$

wobei h_i definiert ist durch

$$h_i = \begin{cases} 2q & y_i - \boldsymbol{x}_i'\boldsymbol{\beta} > 0 \\ 2(1-q) & \text{sonst.} \end{cases}$$

Das nach wie vor am häufigsten verwendete Verfahren zur Schätzung von $\boldsymbol{\beta}$ ist jedoch die Methode der kleinsten Quadrate. Dafür sind vor allem zwei Gründe ausschlaggebend: Zum Einen ist die Anwendung der KQ-Methode mathematisch verhältnismäßig einfach. Beispielsweise ist $\mathrm{KQ}(\boldsymbol{\beta})$ im Gegensatz zu $SM(\boldsymbol{\beta})$ und $SQ(\boldsymbol{\beta})$ bezüglich $\boldsymbol{\beta}$ differenzierbar. Zum Anderen besitzen Schätzungen, die auf der KQ-Methode beruhen, eine Reihe wünschenswerter statistischer Eigenschaften, vergleiche hierzu Abschnitt 3.2.3. Abbildung 3.13 veranschaulicht den Unterschied zwischen dem KQ-Prinzip und der Minimierung absoluter Differenzen anhand des einfachen Modells $y = \beta_0 + \beta_1 x + \varepsilon$. Wie die Abbildung zeigt, erhalten Beobachtungen mit großen Abweichungen durch die Quadrierung ein relativ starkes Gewicht, weshalb ein Nachteil des KQ-Prinzips die Anfälligkeit für Ausreißer ist. Im Gegensatz dazu erhalten Ausreißer bei Verwendung der absoluten Abweichungen ein deutlich geringeres Gewicht, weshalb diese Methode als *robust* gegenüber Ausreißern gilt.

Zur Bestimmung des Minimums von $\mathrm{KQ}(\boldsymbol{\beta})$ formen wir (3.9) zunächst um:

$$\begin{aligned} \mathrm{KQ}(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\,\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{3.10}$$

Dabei haben wir ausgenutzt, dass die Terme $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$ und $\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}$ (wie auch $\boldsymbol{y}'\boldsymbol{y}$ und $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$) Skalare sind. Damit stimmt die Transponierte von $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$ mit $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$ überein, es gilt also $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y})' = \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}$. Unter Verwendung zweier Regeln zur Differentiation von Matrixfunktionen, vergleiche Anhang A insbesondere Satz A.32.7 und A.32.10, erhalten wir

$$\frac{\partial\,\mathrm{KQ}(\boldsymbol{\beta})}{\partial\,\boldsymbol{\beta}} = -2\,\boldsymbol{X}'\boldsymbol{y} + 2\,\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \tag{3.11}$$

Nochmaliges Differenzieren liefert

$$\frac{\partial^2\,\mathrm{KQ}(\boldsymbol{\beta})}{\partial\,\boldsymbol{\beta}\partial\,\boldsymbol{\beta}'} = 2\,\boldsymbol{X}'\boldsymbol{X}.$$

Gemäß Annahme sind die Spalten der Designmatrix $\boldsymbol{X}$ linear unabhängig, d.h. $\mathrm{rg}(\boldsymbol{X}) = p$. Die Matrix $\boldsymbol{X}'\boldsymbol{X}$ ist also nach Satz A.27 (Seite 456) positiv definit, so dass wir die

Abb. 3.13. *Illustration der Methode der kleinsten Quadrate. Abbildung a) zeigt einige Beobachtungen (y_i, x_i), die zufällig um eine Gerade streuen. Gemäß der Methode der kleinsten Quadrate wird die eingezeichnete Regressionsgerade so gewählt, dass die Summe der quadrierten Abweichungen in Abbildung b) minimal werden. Wie in Abbildung b) verdeutlicht, erhalten große Abweichungen durch das Quadrieren überdurchschnittlich viel Gewicht. Zum Vergleich sind in Abbildung c) die absoluten Abweichungen eingezeichnet.*

Lösung unseres Minimierungsproblems durch Nullsetzen von (3.11) erhalten. Der KQ-Schätzer $\hat{\boldsymbol{\beta}}$ ergibt sich also als Lösung der sogenannten *Normalgleichungen*

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}. \tag{3.12}$$

Da $\boldsymbol{X}'\boldsymbol{X}$ positiv definit und damit invertierbar ist (vergleiche Satz A.25.1 auf Seite 456) sind die Normalgleichungen eindeutig lösbar und wir erhalten den KQ-Schätzer

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{3.13}$$

Maximum-Likelihood-Schätzung

Die KQ-Schätzer wurden ohne spezielle Annahmen über die Verteilung der Störgrößen $\boldsymbol{\varepsilon}$ und damit die Verteilung der Zielvariable $\boldsymbol{y}$ gewonnen. Unter der Annahme normalverteilter Störgrößen, d.h. $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, ist es naheliegend die unbekannten Parameter gemäß dem Maximum-Likelihood-Prinzip (ML) zu schätzen, vergleiche dazu in Anhang B Abschnitt B.4.1. Im Folgenden zeigen wir, dass der Maximum-Likelihood-Schätzer für $\boldsymbol{\beta}$ mit dem KQ-Schätzer übereinstimmt.

Mit normalverteilten Störgrößen gilt $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, und wir erhalten die Likelihood

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\right). \tag{3.14}$$

Durch Logarithmieren ergibt sich die Log-Likelihood zu

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{3.15}$$

Bei der Maximierung der Log-Likelihood bezüglich $\boldsymbol{\beta}$ können wir die ersten beiden Summanden in (3.15) vernachlässigen, da diese unabhängig von $\boldsymbol{\beta}$ sind. Die Maximierung des Ausdrucks $-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$ ist gleich bedeutend mit der Minimierung von $(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$, also des Zielkriteriums (3.9). Der Maximum-Likelihood-Schätzer für $\boldsymbol{\beta}$ stimmt also mit dem KQ-Schätzer (3.13) überein.

Geschätzte Werte und Residuen

Ausgehend von der KQ-Schätzung $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ für $\boldsymbol{\beta}$ können wir den (bedingten) Erwartungswert von $\boldsymbol{y}$ durch

$$\widehat{\mathrm{E}(\boldsymbol{y})} = \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

schätzen. Einsetzen des KQ-Schätzers liefert

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

mit der $n \times n$ Matrix

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Die Matrix $\boldsymbol{H}$ wird auch als *Prädiktionsmatrix* (prediction-matrix oder hat-matrix) bezeichnet. Sie besitzt die folgenden, leicht nachweisbaren Eigenschaften:

Eigenschaften der Prädiktionsmatrix

1. $\boldsymbol{H}$ ist symmetrisch.
2. $\boldsymbol{H}$ ist idempotent. Zur Definition idempotenter Matrizen vergleiche in Anhang A Definition A.11.
3. $\mathrm{rg}(\boldsymbol{H}) = \mathrm{sp}(\boldsymbol{H}) = p$. Dabei steht sp als Abkürzung für die Spur einer Matrix.
4. $\frac{1}{n} \leq h_{ii} \leq \frac{1}{r}$, wobei r die Anzahl der Zeilen in $\boldsymbol{X}$ mit identischen $\boldsymbol{x}_i$ ist. Falls alle Zeilen verschieden sind, gilt also $\frac{1}{n} \leq h_{ii} \leq 1$.
5. Die Matrix $\boldsymbol{I} - \boldsymbol{H}$ ist ebenfalls symmetrisch und idempotent mit $\mathrm{rg}(\boldsymbol{I} - \boldsymbol{H}) = n - p$.

Mit Hilfe der Prädiktionsmatrix $\boldsymbol{H}$ können auch die Residuen $\hat{\varepsilon}_i = y_i - \hat{y}_i$ in Matrixnotation dargestellt werden. Wir erhalten

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}.$$

3.2.2 Schätzung der Varianz der Störgrößen

Maximum-Likelihood-Schätzung

Es liegt nahe, die Varianz σ^2 durch eine Maximum-Likelihood-Schätzung zu schätzen. Bei der Bestimmung des ML-Schätzers für $\boldsymbol{\beta}$ wurden bereits die Likelihood $L(\boldsymbol{\beta}, \sigma^2)$ und Log-Likelihood $l(\boldsymbol{\beta}, \sigma^2)$ für das lineare Modell bestimmt, vergleiche (3.14) und (3.15). Partielle Differentiation der Log-Likelihood (3.15) nach σ^2 und Nullsetzen liefert

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0.$$

Einsetzen des ML- bzw. KQ-Schätzers $\hat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ ergibt

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \hat{\boldsymbol{y}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = 0$$

und damit

$$\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}.$$

Dieser Schätzer für σ^2 wird jedoch selten verwendet. Für den Erwartungswert der Residuenquadratsumme ergibt sich nämlich

$$\mathrm{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = (n - p) \cdot \sigma^2 \tag{3.16}$$

und damit

$$\mathrm{E}(\hat{\sigma}^2_{ML}) = \frac{n-p}{n}\sigma^2.$$

Der ML-Schätzer für σ^2 ist also nicht erwartungstreu. Einen Beweis für (3.16) findet man in Abschnitt 3.7.2 auf Seite 181.

Restringierte Maximum-Likelihood-Schätzung

Unter Beachtung von (3.16) können wir sofort einen erwartungstreuen Schätzer $\hat{\sigma}^2$ für σ^2 konstruieren:

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}. \tag{3.17}$$

Hierbei handelt es sich um die übliche Schätzung für σ^2. Der Schätzer (3.17) besitzt darüber hinaus eine interessante Charakterisierung als sogenannter *Restringierter Maximum-Likelihood-Schätzer* (REML). Es kann gezeigt werden, dass (3.17) die *marginale Likelihood*

$$L(\sigma^2) = \int L(\boldsymbol{\beta}, \sigma^2)\, d\boldsymbol{\beta}$$

maximiert. Bei der marginalen Likelihood wird der Vektor $\boldsymbol{\beta}$ aus der Likelihood „herausintegriert“. Im Allgemeinen ist der REML-Schätzer bei Varianzparametern weniger stark verzerrt als der ML-Schätzer und wird daher in der Regel bevorzugt. Im vorliegenden Fall ist der REML-Schätzer im Gegensatz zum ML-Schätzer sogar unverzerrt. Eine weitere Anwendung des REML-Prinzips findet man bei der Schätzung linearer gemischter Modelle, vergleiche Kapitel 6.

Parameterschätzer im klassischen linearen Modell

Schätzer für $\boldsymbol{\beta}$

Im klassischen linearen Modell minimiert der Schätzer

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

das KQ-Kriterium

$$\mathrm{KQ}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2.$$

Der KQ-Schätzer ist unter der Annahme normalverteilter Störgrößen gleichzeitig der ML-Schätzer für $\boldsymbol{\beta}$.

Schätzer für σ^2

Die Schätzung

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

ist erwartungstreu und kann als REML-Schätzer für σ^2 charakterisiert werden.

3.2.3 Eigenschaften der Schätzungen

Wir befassen uns jetzt mit den Eigenschaften der im letzten Abschnitt gewonnenen Schätzer. Die folgenden Unterabschnitte befassen sich zunächst mit den geometrischen Eigenschaften des KQ-Schätzers und anschließend mit statistischen Eigenschaften bei endlichem und unendlichem Stichprobenumfang. Zuletzt beschäftigen wir uns mit den Eigenschaften der Residuen.

Geometrische Eigenschaften des KQ-Schätzers

Geometrisch betrachtet ist der (bedingte) Erwartungswert $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ im linearen Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ein n-dimensionaler Vektor im $\mathbb{R}^n$. Da $\boldsymbol{\mu}$ eine Linearkombination der Spalten der Designmatrix $\boldsymbol{X}$ ist, muss $\boldsymbol{\mu}$ ein Element des Spaltenraums von $\boldsymbol{X}$ sein. Zum Spaltenraum einer Matrix vergleiche In Anhang A Definition A.15. Der Spaltenraum von $\boldsymbol{X}$ wird von den Spalten $\mathbf{1}, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$ aufgespannt, die annahmegemäß linear unabhängig sind ($\mathrm{rg}(\boldsymbol{X}) = p$). Damit handelt es sich um einen Vektorraum der Dimension p.

Bei den Beobachtungen $\boldsymbol{y}$ der Zielvariable handelt es sich geometrisch betrachtet ebenfalls um einen Vektor des $\mathbb{R}^n$. Dieser ist jedoch, ebenso wie $\boldsymbol{\varepsilon}$, *kein* Element des Spaltenraums von $\boldsymbol{X}$.

Wir veranschaulichen die Situation anhand des sehr einfachen Modells

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

mit nur zwei Beobachtungen und einem Parameter β_0. Da die Designmatrix nur aus dem Spaltenvektor $\boldsymbol{x}^0 = (1, 1)'$ besteht, handelt es sich beim Spaltenraum von $\boldsymbol{X}$ um alle

Abb. 3.14. *Veranschaulichung der geometrischen Eigenschaften des KQ-Schätzers.*

Punkte des $\mathbb{R}^2$, die auf der Gerade durch den Nullpunkt und den Punkt $(1,1)$ liegen, vergleiche Abbildung 3.14. Wird beispielsweise der Vektor $\boldsymbol{y} = (2,3)'$ beobachtet, so erhalten wir als KQ-Schätzer für β_0 den Wert $\hat{\beta}_0 = 2.5$ und damit $\hat{\boldsymbol{y}} = (2.5, 2.5)'$ als geschätzte Werte, vergleiche wieder Abbildung 3.14. Durch die Methode der kleinsten Quadrate minimieren wir $(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, d.h. den *euklidischen Abstand* zwischen $\boldsymbol{y}$ und $\boldsymbol{X\beta}$. Dieser Abstand ist in Abbildung 3.14 offenbar dann minimal, wenn $\hat{\beta}_0$ so gewählt wird, dass die Verbindungslinie zwischen $\boldsymbol{y}$ und $\hat{\boldsymbol{y}}$ *senkrecht* zu $\hat{\boldsymbol{y}}$ ist. Bei dieser Verbindungslinie handelt es sich um den Residuenvektor $\hat{\boldsymbol{\varepsilon}}$, so dass also die Residuen und die geschätzten Werte orthogonal zueinander sind, d.h. $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{y}} = 0$. Darüber hinaus stellen wir fest, dass offenbar auch $\boldsymbol{x}^0$ und $\hat{\boldsymbol{\varepsilon}}$ orthogonal zueinander sind.

Die soeben gemachten Beobachtungen lassen sich auf beliebige lineare Modelle verallgemeinern: Die Parameterschätzungen werden durch die KQ-Methode so bestimmt, dass die Residuen $\hat{\boldsymbol{\varepsilon}}$ und die geschätzten Werte $\hat{\boldsymbol{y}}$ orthogonal sind. Unter Zuhilfenahme von Eigenschaften der Prädiktionsmatrix $\boldsymbol{H}$ (vergleiche Seite 93) gilt nämlich wegen $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$ und $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ die Beziehung

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}'\boldsymbol{H}\,(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{Hy} - \boldsymbol{y}'\boldsymbol{HHy} = \boldsymbol{y}'\boldsymbol{Hy} - \boldsymbol{y}'\boldsymbol{Hy} = 0.$$

Darüberhinaus sind sämtliche Spalten der Designmatrix orthogonal zu den Residuen. Dies ist gleich bedeutend mit $(\boldsymbol{x}^j)'\hat{\boldsymbol{\varepsilon}} = 0$ bzw. $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$. Beim Beweis helfen uns wieder Eigenschaften der Prädiktionsmatrix:

$$\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Hy} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0}.$$

Aus der Tatsache, dass die Spalten der Designmatrix und die Residuen orthogonal zueinander sind, ergeben sich einige weitere interessante Implikationen, die wir in den Eigenschaften 3–5 des folgenden Kastens zusammengefasst haben. Einen Beweis findet man in Abschnitt 3.7.2 auf Seite 182.

Geometrische Eigenschaften des KQ-Schätzers

Die Methode der kleinsten Quadrate führt zu folgenden geometrischen Eigenschaften:

1. Die geschätzten Werte $\hat{\boldsymbol{y}}$ sind orthogonal zu den Residuen $\hat{\boldsymbol{\varepsilon}}$, d.h. $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$.
2. Die Spalten $\boldsymbol{x}^j$ von $\boldsymbol{X}$ sind orthogonal zu den Residuen $\hat{\boldsymbol{\varepsilon}}$, d.h. $(\boldsymbol{x}^j)'\hat{\boldsymbol{\varepsilon}} = 0$ bzw. $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$.
3. Die Residuen sind im Mittel Null, d.h.
$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \text{ bzw. } \bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$
4. Der Mittelwert der geschätzten Werte $\hat{y}_i$ ist gleich dem Mittelwert der beobachteten Werte y_i, d.h.
$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}$$
.
5. Die Regressionshyperebene geht durch den Schwerpunkt der Daten, d.h.
$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_k\bar{x}_k.$$

Beispiel 3.9 Orthogonalisierung

Die Tatsache, dass in einer durch die KQ-Methode geschätzten Regression die Residuen und die Spalten der Designmatrix orthogonal sind, kann benutzt werden zur Konstruktion einer Designmatrix, deren Spalten orthogonal sind. Eine Designmatrix mit orthogonalen Spalten bedeutet, dass die Kovariablen unkorreliert sind. Anwendung findet die Orthogonalisierung der Designmatrix beispielsweise bei der Konstruktion orthogonaler Polynome, die bei der Modellierung nichtlinearer Beziehungen verwendet werden, vergleiche Beispiel 3.5 (Seite 76).

Ausgangspunkt ist die Designmatrix $\boldsymbol{X}$ mit Spalten $\boldsymbol{x}^j$. Unser Ziel ist die Transformation der Spalten $\boldsymbol{x}^j$, so dass die resultierenden Spalten $\tilde{\boldsymbol{x}}^j$ der transformierten Designmatrix $\tilde{\boldsymbol{X}}$ orthogonal sind. Dazu verwenden wir für $j = 1, \ldots, k$ die Transformationen
$$\tilde{\boldsymbol{x}}^j = \boldsymbol{x}^j - \tilde{\boldsymbol{X}}_j(\tilde{\boldsymbol{X}}_j'\tilde{\boldsymbol{X}}_j)^{-1}\tilde{\boldsymbol{X}}_j'\boldsymbol{x}^j,$$
wobei die Matrix $\tilde{\boldsymbol{X}}_j$ die ersten $j-1$ transformierten Vektoren enthält. Die erste Spalte der Designmatrix $\boldsymbol{X}$, der Eins-Vektor, wird nicht transformiert. Bei $\tilde{\boldsymbol{x}}_j$ handelt es sich um den Vektor der Residuen einer Regression mit $\boldsymbol{x}^j$ als abhängiger Variablen und den j transformierten Vektoren $\tilde{\boldsymbol{x}}^0 = \boldsymbol{1}, \tilde{\boldsymbol{x}}^1, \ldots, \tilde{\boldsymbol{x}}^{j-1}$ als Kovariablen. Wegen der Orthogonalitätseigenschaft der Residuen ist $\tilde{\boldsymbol{x}}^j$ orthogonal zu den Spalten von $\tilde{\boldsymbol{X}}_j$, d.h. zu allen $j-1$ vorher konstruierten Variablen $\tilde{\boldsymbol{x}}^1, \ldots, \tilde{\boldsymbol{x}}^{j-1}$. Man beachte, dass die erste transformierte Variable $\tilde{\boldsymbol{x}}^1$ aus $\boldsymbol{x}^1$ durch einfache Zentrierung um den Spaltenmittelwert entsteht.

Das beschriebene Verfahren ist in der linearen Algebra auch als Orthogonalisierung nach Gram-Schmidt bekannt.

△

Streuungszerlegung und Bestimmtheitsmaß

Unter Zuhilfenahme der geometrischen Eigenschaften des KQ-Schätzers kann eine fundamentale Streuungszerlegungsformel für die empirische Varianz der Beobachtungen y_i der Zielgröße hergeleitet werden. Darauf aufbauend kann das sogenannte Bestimmtheitsmaß definiert werden. Dieses ist eng verwandt mit dem empirischen Korrelationskoeffizienten und kann als Maß für die Güte der Anpassung an die Daten herangezogen werden.

In Abschnitt 3.7.2 des Kapitels auf Seite 182 zeigen wir die Zerlegungsformel

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2. \tag{3.18}$$

Division auf beiden Seiten durch n (bzw. $n-1$) liefert die Streuungszerlegung

$$s_{\boldsymbol{y}}^2 = s_{\hat{\boldsymbol{y}}}^2 + s_{\hat{\boldsymbol{\varepsilon}}}^2,$$

wobei $s_{\boldsymbol{y}}^2$, $s_{\hat{\boldsymbol{y}}}^2$ und $s_{\hat{\boldsymbol{\varepsilon}}}^2$ die empirischen Varianzen der beobachteten und geschätzten Werte sowie der Residuen sind. Offenbar lässt sich die empirische Varianz der beobachteten Werte additiv zerlegen in die empirische Varianz der geschätzten Werte und die empirische Varianz der Residuen.

Als Anwendung der Streuungszerlegungsformel lässt sich das *Bestimmtheitsmaß*, definieren, welches die Güte der Anpassung an die Daten misst. Das Bestimmtheitsmaß wird mit R^2 bezeichnet und ist definiert durch

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Wegen der Zerlegungsformel (3.18) gilt $0 \leq R^2 \leq 1$.

Das Bestimmtheitsmaß kann unter Zuhilfenahme der Streuungszerlegungsformel wie folgt interpretiert werden: Je näher R^2 bei 1 liegt, desto kleiner ist die Residuenquadratsumme $\sum \hat{\varepsilon}_i^2$, d.h. desto besser ist die Anpassung an die Daten. Im Extremfall $R^2 = 1$ muss $\sum \hat{\varepsilon}_i^2 = 0$ gelten, d.h. sämtliche Residuen sind gleich Null und die Anpassung an die Daten ist perfekt. Ist umgekehrt R^2 nahe bei 0, so ist die Residuenquadratsumme groß und damit die Anpassung des Modells an die Daten gering. Im Extremfall $R^2 = 0$ muss $\sum(\hat{y}_i - \bar{y})^2 = 0$ gelten. Dies ist gleich bedeutend mit $\hat{y}_i = \bar{y}$, d.h. die Prognose von y_i ist stets gleich dem Mittelwert $\bar{y}$ der Zielgröße und damit *unabhängig* von den erklärenden Variablen. Die Kovariablen besitzen also keinerlei Erklärungsgehalt für den Erwartungswert von y. Man beachte jedoch, dass das Modell fehlspezifiziert sein kann, beispielsweise könnte für eine Variable ein nichtlinearer Zusammenhang vorliegen. In diesem Fall besitzen die Kovariablen durchaus einen Erklärungsgehalt trotz eines Bestimmtheitsmaßes nahe Null.

Im Spezialfall des einfachen Regressionsmodells

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

kann man

Bestimmtheitsmaß

Definition

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Interpretation

Je näher das Bestimmtheitsmaß bei 1 liegt, desto kleiner ist die Residuenquadratsumme und desto besser die Anpassung an die Daten. Falls $R^2 = 1$ ist, sind die Residuen Null und die Anpassung an die Daten ist perfekt.

Eigenschaften

1. Im einfachen linearen Modell $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ stimmt das Bestimmtheitsmaß mit dem quadrierten Korrelationskoeffizient nach Bravais-Pearson überein, d.h. $R^2 = r^2_{\boldsymbol{xy}}$.
2. Im multiplen Regressionsmodell kann das Bestimmtheitsmaß als quadrierter Korrelationskoeffizient zwischen den Beobachtungen $\boldsymbol{y}$ und den Schätzungen $\hat{\boldsymbol{y}}$ aufgefasst werden, d.h. $R^2 = r^2_{\boldsymbol{y}\hat{\boldsymbol{y}}}$.
3. Sei der Vektor $\boldsymbol{x}$ der erklärenden Variablen zerlegt in die beiden Vektoren $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$. Betrachte das volle Modell M1 $y_i = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}_{i1} + \boldsymbol{\beta}_2' \boldsymbol{x}_{i2} + \varepsilon_i$ und das in M1 enthaltene Teilmodell M2 $y_i = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}_{i1} + \varepsilon_i$. Dann gilt

   $$R^2_{M1} \geq R^2_{M2},$$

   d.h. bei hierarchisch geordneten Modellen ist das Bestimmtheitsmaß des Teilmodells stets kleiner oder gleich dem Bestimmtheitsmaß des vollen Modells.
4. Unterschiedliche Modelle können mit dem Bestimmtheitsmaß nur verglichen werden, wenn die Modelle die gleiche Zielgröße, die gleiche Zahl an Parametern und eine Konstante enthalten.

$$R^2 = r^2_{\boldsymbol{xy}}$$

zeigen, wobei $r^2_{\boldsymbol{xy}}$ den quadrierten empirischen Korrelationskoeffizienten (nach Bravais-Pearson) zwischen $\boldsymbol{x}$ und $\boldsymbol{y}$ bezeichnet. Diese Eigenschaft des Bestimmtheitsmaßes ist für die Bezeichnung mit R^2 verantwortlich.

Sowohl im einfachen als auch im multiplen Regressionsmodell kann das Bestimmtheitsmaß als quadrierter empirischer Korrelationskoeffizient zwischen den Beobachtungen $\boldsymbol{y}$ und den Schätzungen $\hat{\boldsymbol{y}}$ aufgefasst werden. Es gilt also

$$R^2 = r^2_{\boldsymbol{y}\hat{\boldsymbol{y}}}.$$

An dieser Stelle bereits wollen wir einer weit verbreiteten Praxis begegnen, nämlich der Jagd nach möglichst hohen Bestimmtheitsmaßen um auf diese Weise ein gutes Schätzergebnis zu erzielen. Bei der Beurteilung verschiedener Modelle mit Hilfe des Bestimmtheitsmaßes ist jedoch eine gewisse Vorsicht angebracht. Unterschiedliche Modelle können mit dem Bestimmtheitsmaß nur dann verglichen werden, wenn drei Voraussetzungen erfüllt sind:

- In allen Modellen wird die gleiche Zielgröße y verwendet, d.h. der Vergleich von Modellen mit den Zielgrößen y und $\log(y)$ ist nicht möglich.
- In allen Modellen ist die Zahl der Regressionskoeffizienten gleich.
- Alle Modelle besitzen eine Konstante β_0.

In der Regel können verschiedene Modelle mit dem Bestimmtheitsmaß also *nicht* vernünftig verglichen werden. Es lässt sich nämlich zeigen, dass das Bestimmtheitsmaß stets größer wird, wenn eine weitere erklärende Variable ins Modell mit aufgenommen wird. Wir veranschaulichen die Problematik anhand eines Beispiels:

Beispiel 3.10 Mietspiegel – Modellvergleich mit Bestimmtheitsmaß

In Abschnitt 3.1.4 (Seite 72) wurden verschiedene Varianten zur Modellierung des Zusammenhangs zwischen Nettomiete pro qm und Wohnfläche untersucht. Tabelle 3.1 fasst die dort geschätzten Modelle zusammen und gibt auch einen Überblick über das jeweilige Bestimmtheitsmaß.

Zunächst stellen wir fest, dass sämtliche Bestimmtheitsmaße relativ klein sind. Dies liegt zum einen an der doch beträchtlichen Streuung der Daten (vergleiche zum Beispiel das Streudiagramm in Abbildung 3.8 auf Seite 74). Zum anderen sind eine Vielzahl von relevanten Einflussvariablen in den Modellen noch nicht berücksichtigt. Das im offiziellen Mietspiegel 1999 verwendete Modell enthält mehr als 20 erklärende Variablen. Damit wird dann ein Bestimmtheitsmaß von 0.49 erreicht.

Vergleichen wir die Modelle M1 und M2, so stellen wir fest, dass M2 ein deutlich höheres Bestimmtheitsmaß besitzt. Da beide Modelle die selbe Anzahl an Parametern besitzen, ist ein Vergleich anhand des Bestimmtheitsmaßes sinnvoll und Modell M2 ist eindeutig zu bevorzugen. Zu diesem Schluss sind wir allerdings auch durch die Analyse der Residuen gekommen, vergleiche Abbildung 3.8 (Seite 74). Die Modelle M1, M3 und M4 sind hierarchisch geordnet, d.h. M3 und M1 sind in M4 enthalten und M1 in M3. Es ist daher klar, dass M1 den kleinsten R^2-Wert, M3 den zweitkleinsten R^2-Wert und M4 den größten R^2-Wert besitzt. Ein Vergleich anhand des Bestimmtheitsmaßes ist hier nicht sinnvoll, da die Anzahl der Parameter jeweils unterschiedlich ist. Es sei jedoch darauf hingewiesen, dass ein Modell mit größerer Anzahl an Regressionsparametern nicht notwendigerweise ein größeres Bestimmtheitsmaß besitzt. Dies kann dann der Fall sein, wenn die Modelle jeweils unterschiedliche erklärende Variablen enthalten. Exemplarisch vergleichen wir die Modelle M2 und M3. Obwohl M3 drei Regressionsparameter enthält und M2 nur zwei, besitzt M2 das größere Bestimmtheitsmaß. In diesem Fall ist es auch sinnvoll, M2 vor M3 zu bevorzugen, da dieses Modell weniger Parameter enthält *und* ein größeres Bestimmtheitsmaß besitzt. Das Bestimmtheitsmaß steigt also nur automatisch mit wachsender Anzahl an Parametern an, wenn die Modelle hierarchisch geordnet sind, d.h. die komplexeren Modelle enthalten die anderen Modelle als Spezialfall, wie dies bei den Modellen M1, M3 und M4 der Fall ist.

△

In Abschnitt 3.6 werden wir Maße kennenlernen, die geeignet sind, auch Modelle mit unterschiedlicher Anzahl an Parametern hinsichtlich ihrer Güte zu beurteilen.

Modell	Gleichung	R^2
M1	$\widehat{mieteqm} = 18.46 - 0.068 \cdot flaeche$	0.116
M2	$\widehat{mieteqm} = 9.23 - 273.35 \cdot 1/flaeche$	0.154
M3	$\widehat{mieteqm} = 23.07 - 0.206 \cdot flaeche + 0.00092 \cdot flaeche^2$	0.143
M4	$\widehat{mieteqm} = 27.84 - 0.42 \cdot flaeche + 0.0039 \cdot flaeche^2 - 0.000012 \cdot flaeche^3$	0.150

Tabelle 3.1. *Vergleich verschiedener Modelle zur Modellierung des Zusammenhangs zwischen Nettomiete pro qm (mieteqm) und Wohnfläche (flaeche).*

Statistische Eigenschaften ohne spezielle Verteilungsannahmen

Wir bestimmen jetzt die wichtigsten statistischen Eigenschaften des KQ-Schätzers. Dabei werden wir zunächst noch davon ausgehen, dass die Störgrößen einer beliebigen Verteilungsklasse angehören. Wir gehen also nicht notwendigerweise davon aus, dass die Störgrößen normalverteilt sind.

Erwartungswert und Verzerrung

Für den Erwartungswert des KQ-Schätzers erhalten wir

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

Der KQ-Schätzer ist also erwartungstreu für $\boldsymbol{\beta}$ und damit unverzerrt.

Kovarianzmatrix

Unter Verwendung von Satz B.1.5 (Seite 463) ist die Kovarianzmatrix des KQ-Schätzer gegeben durch

$$\begin{aligned}\mathrm{Cov}\,(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Cov}(\boldsymbol{y})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\ &= \sigma^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

Speziell für die Varianzen der geschätzten Regressionskoeffizienten $\hat{\beta}_j$ lässt sich noch eine besser interpretierbare Version herleiten, vergleiche zum Beispiel Wooldridge (2006). Es gilt

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2},$$

wobei R_j^2 das Bestimmtheitsmaß einer Regression zwischen x_j als abhängiger Variable und den restlichen erklärenden Variablen ist. Damit sehen wir sofort, von welchen Faktoren die Genauigkeit der Schätzung für die Regressionskoeffizienten abhängt:

- Je kleiner die Modellvarianz σ^2, desto kleiner die Varianz von $\hat{\beta}_j$ und damit desto genauer die Schätzung.

- Je geringer die lineare Abhängigkeit zwischen x_j und den anderen erklärenden Variablen, desto geringer die Varianz von $\hat{\beta}_j$. Offensichtlich werden die Varianzen $\mathrm{Var}(\hat{\beta}_j)$ minimiert, wenn $R_j^2 = 0$, d.h. wenn die Kovariablen unkorreliert sind. Dieses *orthogonale Design* wird bei geplanten Experimenten im Rahmen der Versuchsplanung angestrebt. Sind hingegen einige Kovariablen untereinander hoch korreliert, können die Schätzungen extrem ungenau sein. Im Extremfall $R_j^2 \to 1$ explodieren die Varianzen. Auf dieses sogenannte Multikollinearitätsproblem gehen wir noch genauer in Abschnitt 3.6 zur Modelldiagnose ein.
- Je größer die Streuung der Kovariable x_j um ihren Mittelwert, desto kleiner die Varianz von $\hat{\beta}_j$.

Die Kovarianz von $\hat{\boldsymbol{\beta}}$ ist in Anwendungen nicht berechenbar, da die Varianz σ^2 der Störgrößen unbekannt ist. Stattdessen muss $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ geschätzt werden indem σ^2 durch die Schätzung $\hat{\sigma}^2 = 1/(n-p)\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ ersetzt wird:

$$\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Bei den Diagonalelementen dieser Matrix handelt es sich um die geschätzten Varianzen der KQ-Schätzungen $\hat{\beta}_j$. Die Wurzeln der Diagonalelemente sind die geschätzten Standardfehler, die wir im Weiteren mit se_j abkürzen:

$$se_j = \widehat{\mathrm{Var}(\hat{\beta}_j)}^{\,1/2}, \quad j = 0, 1, \ldots, k.$$

Die geschätzte Kovarianzmatrix, insbesondere die daraus berechenbaren Standardfehler se_j, benötigen wir vor allem bei der Durchführung statistischer Tests über die Regressionskoeffizienten und bei der Berechnung von Konfidenzintervallen, vergleiche Abschnitt 3.3

Vergleich mit linearen Schätzern

Im Folgenden wollen wir den KQ-Schätzer mit der allgemeinen Klasse der *linearen Schätzer* hinsichtlich seiner "Güte" vergleichen. Ein linearer Schätzer $\hat{\boldsymbol{\beta}}^L$ hat dabei die Gestalt

$$\hat{\boldsymbol{\beta}}^L = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{y},$$

wobei $\boldsymbol{a}$ ein $p \times 1$ Vektor ist und $\boldsymbol{A}$ eine Matrix der Dimension $p \times n$. Die Komponenten β_j von $\boldsymbol{\beta}$ werden also durch eine Linearkombination der Beobachtungen y_i der Zielvariable geschätzt, d.h.

$$\hat{\beta}_j^L = a_j + A_{j1}y_1 + \cdots + A_{jn}y_n \qquad j = 0, \ldots, k.$$

Offensichtlich handelt es sich beim KQ-Schätzer $\hat{\boldsymbol{\beta}}$ um einen speziellen linearen Schätzer mit $\boldsymbol{a} = \boldsymbol{0}$ und $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Erwartungswert und Kovarianzmatrix linearer Schätzer können ähnlich wie beim KQ-Schätzer berechnet werden. Wir erhalten

$$\mathrm{E}(\hat{\boldsymbol{\beta}}^L) = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} \qquad \mathrm{Cov}\,(\hat{\boldsymbol{\beta}}^L) = \sigma^2\,\boldsymbol{A}\boldsymbol{A}'.$$

Lineare Schätzer sind also nicht notwendigerweise erwartungstreu. Ein Vergleich mit der speziellen Klasse der linearen und erwartungstreuen Schätzer $\hat{\boldsymbol{\beta}}^L$ zeigt, dass die KQ-Schätzer minimale Varianzen besitzen, d.h.

$$\mathrm{Var}(\hat{\beta}_j^L) \geq \mathrm{Var}(\hat{\beta}_j), \quad j = 0, \ldots, k.$$

Diese Eigenschaft überträgt sich auch auf eine beliebige Linearkombination

$$b_0\beta_0 + b_1\beta_1 + \ldots + b_k\beta_k = \boldsymbol{b}'\boldsymbol{\beta}$$

von $\boldsymbol{\beta}$, d.h.

$$\mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}^L) \geq \mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}).$$

Bei diesen Eigenschaften des KQ-Schätzers handelt es sich um das berühmte Gauß-Markov-Theorem. Einen Beweis findet man im Anhang des Kapitels auf Seite 183.

Das Gauß-Markov-Theorem lässt sich zur Bestimmung einer „optimalen“ Prognose einer neuen (zukünftigen) Beobachtung y_0 bei gegebenem Kovariablenvektor $\boldsymbol{x}_0$ anwenden. Als Prognose verwenden wir den bedingten Erwartungswert

$$\mathrm{E}(y_0 \,|\boldsymbol{x}_0) = \boldsymbol{x}_0{}'\boldsymbol{\beta}.$$

Eine optimale Schätzung des Erwartungswerts (im Sinne des Gauß-Markov-Theorems) ist dann gegeben durch

$$\hat{y}_0 = \boldsymbol{x}_0{}'\hat{\boldsymbol{\beta}}.$$

Statistische Eigenschaften bei Normalverteilungsannahme

Die bisher abgeleiteten statistischen Eigenschaften des KQ-Schätzers wurden ohne Annahme einer speziellen Verteilung für die Störgrößen $\boldsymbol{\varepsilon}$ gewonnen. Insbesondere wurde nicht vorausgesetzt, dass die Störgrößen normalverteilt sind. Unter der Annahme normalverteilter Störgrößen, d.h. $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, können wir weitere Eigenschaften des KQ-Schätzers gewinnen. Diese sind dann vor allem bei der Konstruktion von Hypothesentests und Konfidenzintervallen für $\boldsymbol{\beta}$ nützlich.

Da $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ und der KQ-Schätzer $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ eine lineare Transformation von $\boldsymbol{y}$ ist, folgt wegen Satz B.3 (Seite 464) sofort, das $\hat{\boldsymbol{\beta}}$ normalverteilt ist. Den Erwartungswert und die Kovarianzmatrix des KQ-Schätzers haben wir bereits vorher bestimmt, so dass wir unmittelbar

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

erhalten. Darüber hinaus liefert Satz B.6.1 (Seite 466) eine Aussage über den (mit der inversen Kovarianzmatrix gewichteten) Abstand zwischen dem KQ-Schätzer $\hat{\boldsymbol{\beta}}$ und $\boldsymbol{\beta}$:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi_p^2$$

Diese Eigenschaft ist bei der Herleitung der Verteilung der Teststatistik bei Hypothesentests nützlich.

Statistische Eigenschaften des KQ-Schätzers

Ohne spezielle Verteilungsannahme

1. *Erwartungswert:* $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, d.h. der KQ-Schätzer ist erwartungstreu und damit unverzerrt.
2. *Kovarianzmatrix:* $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$, speziell gilt

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2},$$

   wobei R_j^2 das Bestimmtheitsmaß einer Regression zwischen x_j als Zielgröße und den restlichen erklärenden Variablen ist.
   Eine Schätzung für die Kovarianzmatrix ist gegeben durch

$$\mathrm{Cov}\,(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

3. *Gauß-Markov-Theorem:* Unter allen linearen und erwartungstreuen Schätzern $\hat{\boldsymbol{\beta}}^L$ besitzen die KQ-Schätzer minimale Varianzen, d.h.

$$\mathrm{Var}(\hat{\beta}_j) \leq \mathrm{Var}(\hat{\beta}_j^L), \quad j = 0, \ldots, k.$$

   Darüber hinaus gilt auch für eine beliebige Linearkombination $\boldsymbol{b}'\boldsymbol{\beta}$

$$\mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}) \leq \mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}^L).$$

Mit Normalverteilungsannahme

1. *Verteilung der Zielgröße:*

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

2. *Verteilung des KQ-Schätzers:*

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

3. *Verteilung des gewichteten Abstands:*

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} \sim \chi_p^2.$$

Asymptotische Eigenschaften des KQ-Schätzers

Die Verteilungseigenschaften des KQ-Schätzers bilden die Grundlage für die im nächsten Abschnitt angegebenen Tests und Konfidenzintervalle. Damit diese exakt gültig sind, wird die Normalverteilungsannahme benötigt. Einige Aussagen bleiben jedoch auch ohne Normalverteilungsannahme asymptotisch bzw. approximativ gültig, wenn der Stichprobenumfang n gegen unendlich geht bzw. hinreichend groß ist. Zur Verdeutlichung indizieren wir das Modell in Matrixschreibweise mit n:

$$\boldsymbol{y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad \mathrm{E}(\boldsymbol{\varepsilon}_n) = \boldsymbol{0}, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}_n) = \sigma^2\boldsymbol{I}_n.$$

Analog indizieren wir den KQ-Schätzer $\hat{\boldsymbol{\beta}}_n$ und den Varianzschätzer $\hat{\sigma}_n^2$ mit n. Für die Gültigkeit asymptotischer Aussagen sind neben den Annahmen 1-3 auf Seite 62 noch Annahmen über das Grenzverhalten der Designmatrix $\boldsymbol{X}_n$ und damit über die Folge $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \ldots$ der Design- bzw. Regressionsvektoren nötig. Die oft getroffene Standardannahme ist, dass die mit n gemittelte Matrix $\boldsymbol{X}_n'\boldsymbol{X}_n$ gegen eine positiv definite Grenzmatrix $\boldsymbol{V}$ konvergiert, d.h.

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}_n'\boldsymbol{X}_n = \boldsymbol{V}, \quad \boldsymbol{V} \text{ positiv definit} \tag{3.19}$$

gilt. In diesem Fall erhalten wir die folgenden asymptotischen Aussagen:

Asymptotische Eigenschaften des KQ-Schätzers

1. Der KQ-Schätzer $\hat{\boldsymbol{\beta}}_n$ für $\boldsymbol{\beta}$ sowie die ML- bzw. REML-Schätzer $\hat{\sigma}_n^2$ für die Varianz σ^2 sind konsistent.
2. Der KQ-Schätzer ist asymptotisch normalverteilt, genauer gilt

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{V}^{-1}),$$

d.h. die mit $\sqrt{n}$ normierte Differenz $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$ konvergiert nach Verteilung gegen die Normalverteilung auf der rechten Seite.

Für hinreichend großen Stichprobenumfang n nutzt man diese asymptotischen Aussagen folgendermaßen: Zunächst ist $\hat{\boldsymbol{\beta}}_n$ approximativ normalverteilt mit

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \sigma^2\boldsymbol{V}^{-1}/n).$$

Ersetzt man nun σ^2 durch den konsistenten Schätzer $\hat{\sigma}_n^2$ und $\boldsymbol{V}$ durch die Approximation $\boldsymbol{V} \overset{a}{\sim} 1/n\boldsymbol{X}_n'\boldsymbol{X}_n$, so folgt

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \hat{\sigma}_n^2(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}),$$

d.h. auch ohne Normalverteilungsannahme für $\boldsymbol{\varepsilon}$ besitzt der KQ-Schätzer bei großem Stichprobenumfang approximativ die gleiche Normalverteilung wie mit Normalverteilungsannahme für $\boldsymbol{\varepsilon}$, falls die Annahme (3.19) erfüllt ist. Diese ist insbesondere dann erfüllt, wenn die beobachteten Regressorvektoren $\boldsymbol{x}_i$, $i = 1, \ldots, n$ unabhängige und identisch verteilte Realisierungen stochastischer Kovariablen $\boldsymbol{x} = (\boldsymbol{1}, x_1, \ldots, x_k)'$ sind, also wenn $(y_i, \boldsymbol{x}_i)$ aus einer zufälligen Stichprobe bezüglich $(y, \boldsymbol{x})$ stammen. Dies ist für viele

empirische Studien erfüllt, wie etwa in unseren Anwendungen zum Mietspiegel und zur Unterernährung. In diesem Fall gilt nach dem Gesetz der großen Zahlen

$$\frac{1}{n}\boldsymbol{X}_n'\boldsymbol{X}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' \to \mathrm{E}(\boldsymbol{x}\boldsymbol{x}') =: \boldsymbol{V}.$$

Typischerweise ist die Annahme jedoch verletzt bei deterministischen Regressoren mit Trend. Man sieht dies bereits an dem Beispiel eines linearen Trends $\boldsymbol{x}_i = i$, d.h.

$$y_i = \beta \cdot i + \varepsilon_i, \quad i = 1, \ldots, n,$$

(wobei wir zur Vereinfachung den konstanten Term $\beta_0 = 0$ gesetzt haben). Dann gilt

$$\frac{1}{n}\boldsymbol{X}_n'\boldsymbol{X}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^2 = \frac{1}{n}(1 + \cdots + i^2 + \cdots + n^2) \to \infty,$$

so dass die Annahme (3.19) nicht erfüllt ist.

Tatsächlich gelten Konsistenz und asymptotische Normalität jedoch auch unter den folgenden, wesentlich allgemeineren Voraussetzungen:

$$(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1} \to 0 \tag{3.20}$$

Informell bedeutet dies, dass die Information über die Kovariablen mit wachsendem Stichprobenumfang immer weiter zunimmt. Es lässt sich zeigen, dass die Bedingung (3.20) notwendig und hinreichend für die Konsistenz des KQ-Schätzers ist. Darüber hinaus sind dann auch die Schätzer für die Varianz konsistent.

Für die asymptotische Normalität wird noch zusätzlich die Bedingung

$$\max_{i=1,\ldots,n} \boldsymbol{x}_i'(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_i \to 0 \quad \text{für} \quad n \to \infty \tag{3.21}$$

benötigt. Informell besagt dies, dass der Einfluss jeder Beobachtung im Verhältnis zur Gesamtinformation $\boldsymbol{X}_n'\boldsymbol{X}_n = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$ verschwindend gering ist. Damit kommt der zentrale Grenzwertsatz (genauer: in der Form von Lindeberg-Feller) zum Tragen. Wenn die Bedingungen (3.20) und (3.21) erfüllt sind, so bleibt die für praktische Zwecke wichtige Normalverteilungsapproximation

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \hat{\sigma}_n^2(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}),$$

weiterhin gültig.

Beispiel 3.11 Lineare Einfachregression

Für das Modell der linearen Einfachregression (ohne Konstante β_0)

$$y_i = \beta x_i + \varepsilon_i$$

prüft man leicht Folgendes nach:

1. Für einen linearen Trend $x_i = i$ sind (3.20) und (3.21) erfüllt.
2. Für $x_i = 1/i$ sind sowohl (3.20) als auch (3.21) verletzt, d.h. der KQ-Schätzer ist weder konsistent noch asymptotisch normalverteilt. Der Grund dafür liegt darin, dass die Folge $x_i = 1/i$ der Regressionswerte zu schnell gegen Null konvergiert und somit nicht genügend Information liefert.
3. Für $x_i = 1/\sqrt{i}$ sind sowohl (3.20) als auch (3.21) erfüllt, d.h. obwohl ebenfalls $x_i \to 0$ für $n \to \infty$ gilt, wird noch genügend Information geliefert.

△

Statistische Eigenschaften der Residuen

Wir beenden diesen Abschnitt mit der Untersuchung der statistischen Eigenschaften der Residuen $\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$. In Abhängigkeit von der Prädiktions-Matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ lassen sich die Residuen schreiben als

$$\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

vergleiche Seite 93. Damit erhalten wir

$$\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathrm{E}(\boldsymbol{y}) - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

als Erwartungswert sowie

$$\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}) = \mathrm{Cov}((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}) = (\boldsymbol{I} - \boldsymbol{H})\sigma^2\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{I} - \boldsymbol{H}).$$

als Kovarianzmatrix der Residuen. Bei der Berechnung der Kovarianzmatrix haben wir neben Satz B.1.5 (Seite 463) in Anhang B ausgenutzt, dass die Matrix $\boldsymbol{I} - \boldsymbol{H}$ symmetrisch und idempotent ist. Speziell für die Varianzen der Residuen erhalten wir

$$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}),$$

wobei h_{ii} das i-te Diagonalelement der Prädiktionsmatrix ist. Wir stellen also fest:

- Die Residuen sind wie die Störgrößen im Mittel Null.
- Die Residuen sind im Gegensatz zu den Störgrößen *nicht* unkorreliert.
- Die Residuen besitzen im Gegensatz zu den Störgrößen *heteroskedastische Varianzen.* Wegen $\frac{1}{n} \leq h_{ii} \leq 1$ (vergleiche Seite 93) ist die Varianz des i-ten Residuums umso kleiner, je größer h_{ii} ist, d.h. desto näher h_{ii} bei Eins liegt.

Wenn wir zusätzlich von normalverteilten Störgrößen ausgehen, können wir auch die Verteilung der Residuen bestimmen. Wir erhalten

$$\hat{\boldsymbol{\varepsilon}} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})). \tag{3.22}$$

Wegen $\mathrm{rg}(\boldsymbol{H}) = p \leq n$ handelt es sich um eine singuläre Normalverteilung, vergleiche auch Abschnitt B.3.2 (Seite 465) in Anhang B.

Unter Verwendung von (3.22) lassen sich auch Aussagen über die Residuenquadratsumme ableiten. In Abschnitt 3.7.2 (Seite 184) zeigen wir

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2_{n-p}.$$

Äquivalent ist die Aussage

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Außerdem lässt sich zeigen, dass die Residuenquadratsumme und der KQ-Schätzer unabhängig sind. Beide Aussagen werden bei der Herleitung von Hypothesentests über die Regressionskoeffizienten benötigt.

Statistische Eigenschaften der Residuen

Ohne spezielle Verteilungsannahme

1. *Erwartungswert:* $\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$, d.h. die Residuen sind im Mittel Null.
2. *Varianzen:* Es gilt
$$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}),$$
d.h. die Residuen besitzen (im Gegensatz zu den Störgrößen ε_i) heteroskedastische Varianzen.
3. *Kovarianzmatrix:*
$$\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'),$$
d.h. die Residuen sind (im Gegensatz zu den Störgrößen) nicht unkorreliert.

Mit Normalverteilungsannahme

4. *Verteilung der Residuen:*
$$\hat{\boldsymbol{\varepsilon}} \sim \mathrm{N}(\mathbf{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$$
5. *Verteilung der Residuenquadratsumme:*
$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} = (n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$
6. *Unabhängigkeit:* Die Residuenquadratsumme $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ und der KQ-Schätzer $\hat{\boldsymbol{\beta}}$ sind unabhängig.

Standardisierte und studentisierte Residuen

In der Praxis dienen die Residuen hauptsächlich dazu, die Modellannahmen im linearen Modell zu überprüfen. Allerdings sind die Residuen dazu nicht ohne weiteres geeignet. Wie wir gesehen haben sind die Residuen selbst weder homoskedastisch noch unkorreliert. Die Korrelation ist in der Regel vernachlässigbar, jedoch nicht die Heteroskedastizität. Die Überprüfung der Annahme homoskedastischer Störgrößen ist daher problematisch, da heteroskedastische Residuen als Regelfall nicht auf heteroskedastische Störgrößen schließen lassen. Eine nahe liegende Lösung für das Problem der Heteroskedastizität ist die Standardisierung der Residuen. Durch Division mit der geschätzten Standardabweichung der Residuen erhalten wir die *standardisierten Residuen*

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}. \tag{3.23}$$

Unter der Voraussetzung, dass die Modellannahmen stimmen, sind die standardisierten Residuen homoskedastisch. Ist die Annahme homoskedastischer Varianzen verletzt, lässt

sich dies durch Analyse der standardisierten Residuen feststellen. In der Regel zeichnet man Streudiagramme zwischen den standardisierten Residuen und den geschätzten Werten bzw. den Kovariablen, vergleiche hierzu auch Abschnitt 3.6.4 zur Modelldiagnose.

Da die Residuen normalverteilt sind und der Ausdruck $(n-p)\hat{\sigma}^2/\sigma^2$ χ^2_{n-p}-verteilt ist, könnte man versucht sein, zu schließen, dass die standardisierten Residuen t-verteilt sind, vergleiche die Definition B.7 der t-Verteilung in Anhang B. Dieser Schluss ist allerdings nicht möglich, da $\hat{\varepsilon}_i$ in $\hat{\sigma}$ vorkommt und aus diesem Grund Zähler und Nenner in (3.23) nicht unabhängig sind.

Das Problem der Abhängigkeit kann jedoch auf elegante Weise umgangen werden. Dazu bestimmen wir Schätzungen, die auf allen Beobachtungen bis auf die i-te basieren. In der englischsprachigen Literatur spricht man von sogenannten „leave one out“ Schätzern. Anschließend definieren wir Residuen, die auf diesen „leave one out“ Schätzern beruhen und zeigen, dass diese nach Standardisierung t-verteilt sind.

Bezeichne $\boldsymbol{X}_{(i)}$ und $\boldsymbol{y}_{(i)}$ Designmatrix und Vektor der Zielgrößen, aus denen die i-te Zeile gelöscht wurde. Wir bezeichnen mit $\hat{\boldsymbol{\beta}}_{(i)}$ den auf diesen Größen beruhenden KQ-Schätzer, der auf allen Beobachtungen bis auf die i-te beruht. Es gilt

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

Basierend auf $\hat{\boldsymbol{\beta}}_{(i)}$ erhalten wir dann Schätzungen $\hat{y}_{(i)} = \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_{(i)}$ und Residuen

$$\hat{\varepsilon}_{(i)} = y_i - \hat{y}_{(i)} = y_i - \boldsymbol{x}'_i(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{y}_{(i)}.$$

Eine einfache Rechnung zeigt

$$\hat{\varepsilon}_{(i)} \sim \mathrm{N}(0, \sigma^2(1 + \boldsymbol{x}'_i(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i))$$

bzw.

$$\frac{\hat{\varepsilon}_{(i)}}{\sigma(1 + \boldsymbol{x}'_i(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}} \sim \mathrm{N}(0,1).$$

Gemäß Eigenschaft 5 des vorangegangenen Kastens gilt

$$(n-p-1)\frac{\hat{\sigma}^2_{(i)}}{\sigma^2} \sim \chi^2_{n-p-1},$$

wobei

$$\begin{aligned}\hat{\sigma}^2_{(i)} = \tfrac{1}{n-p-1}\Big[&\Big(y_1 - \boldsymbol{x}'_1\hat{\boldsymbol{\beta}}_{(i)}\Big) + \ldots + \Big(y_{i-1} - \boldsymbol{x}'_{i-1}\hat{\boldsymbol{\beta}}_{(i)}\Big) + \\ &+ \Big(y_{i+1} - \boldsymbol{x}'_{i+1}\hat{\boldsymbol{\beta}}_{(i)}\Big) + \ldots + \Big(y_n - \boldsymbol{x}'_n\hat{\boldsymbol{\beta}}_{(i)}\Big)\Big]\end{aligned}$$

eine Schätzung für σ^2 ist, die nicht auf der i-ten Beobachtung beruht. Jetzt können wir die Definition B.7 der t-Verteilung (Seite 461) verwenden und erhalten die sogenannten *studentisierten Residuen*

$$r^*_i = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}(1 + \boldsymbol{x}'_i(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}} \sim t_{n-p-1}.$$

Entscheidend für die Verteilungsaussage ist an dieser Stelle, dass $\hat{\varepsilon}_{(i)}$ und $\hat{\sigma}_{(i)}$ unabhängig sind, da die i-te Beobachtung y_i zur Berechnung von $\hat{\sigma}_{(i)}$ nicht verwendet wird. Eine

weitere Voraussetzung für obige Verteilungsaussage ist, dass das Modell korrekt ist. Die Kenntnis der Verteilung der studentisierten Residuen bei korrekt spezifiziertem Modell kann für die Modelldiagnose verwendet werden, um Beobachtungen zu identifizieren, die nicht zum angepassten Modell passen, vergleiche hierzu Abschnitt 3.6.4.

Die standardisierten und studentisierten Residuen lösen das Problem heteroskedastischer Residuen, sind aber ebenfalls nicht unkorreliert. In der Literatur existieren daher auch Vorschläge zur Definition unkorrelierter Residuen. Wie bereits erwähnt, ist bei korrekt spezifiziertem Modell die Korrelation der standardisierten und studentisierten Residuen zu vernachlässigen, so dass wir an dieser Stelle auf die Darstellung verzichten.

Im nachfolgenden Kasten fassen wir sämtliche bisher definierten Residuen zusammen:

Übersicht Residuen

Gewöhnliche Residuen

Die Residuen sind gegeben durch

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} \quad i = 1, \ldots, n.$$

Standardisierte Residuen

Die standardisierten Residuen sind definiert durch

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

wobei h_{ii} das i-te Diagonalelement der Prädiktionsmatrix ist.

Studentisierte Residuen

Die studentisierten Residuen sind definiert durch

$$r_i^* = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}(1 + \boldsymbol{x}_i'(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = r_i \left(\frac{n - p - 1}{n - p - r_i^2}\right)^{1/2}.$$

Die studentisierten Residuen finden Verwendung bei der Überprüfung der Modellannahmen und dienen zur Entdeckung von Ausreißern (Abschnitt 3.6.4).

Partielle Residuen

Die partiellen Residuen bezüglich Kovariable x_j sind definiert als

$$\hat{\varepsilon}_{x_j,i} = y_i - \hat{\beta}_0 - \ldots - \hat{\beta}_{j-1}x_{i,j-1} - \hat{\beta}_{j+1}x_{i,j+1} - \ldots - \hat{\beta}_k x_{ik} = \hat{\varepsilon}_i + \hat{\beta}_j x_{ij}.$$

Die partiellen Residuen $\hat{\varepsilon}_{x_j,i}$ sind vom Einfluss aller Kovariablen mit Ausnahme von x_j bereinigt. Damit sind sie vor allem geeignet zur Überprüfung der gewählten Modellierung des Einflusses von x_j (Abschnitt 3.6.4).

Hier finden sich auch zwei weitere Darstellungen der studentisierten Residuen. Diese zeigen insbesondere, dass die studentisierten Residuen aus den standardisierten Residuen

berechenbar sind, d.h. eine wiederholte Neuberechnung des KQ-Schätzers ist nicht notwendig. Darüberhinaus führen wir an dieser Stelle noch einmal die in Beispiel 3.5 (Seite 76) verwendeten partiellen Residuen auf.

3.3 Hypothesentests und Konfidenzintervalle

Ziel dieses Abschnitts ist die Konstruktion von statistischen Tests, mit denen Hypothesen über die unbekannten Regressionsparameter $\boldsymbol{\beta}$ überprüft werden können. Aufgrund der Dualität zwischen zweiseitigen statistischen Tests und Konfidenzintervallen lassen sich aus den konstruierten Tests auch Konfidenzintervalle für die Regressionsparameter $\boldsymbol{\beta}$ ableiten. Voraussetzung für die Konstruktion von (exakten) Tests- und Konfidenzintervallen ist die Gültigkeit der Normalverteilungsannahme der Störgrößen. In diesem Abschnitt setzen wir also zunächst unabhängige und identisch verteilte Störgrößen $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ voraus. Die Tests- und Konfidenzintervalle sind aber relativ robust gegenüber geringfügigen Abweichungen von der Normalverteilung. In Abschnitt 3.3.1 werden wir darüberhinaus sehen, dass die unter Normalverteilungsannahme hergeleiteten Tests- und Konfidenzintervalle asymptotisch, d.h. für großen Stichprobenumfang, gültig bleiben, auch wenn die Störgrößen nicht normalverteilt sind.

Beispiel 3.12 Mietspiegel – Hypothesentests

Wir illustrieren die Konzepte dieses Abschnitts wieder anhand der Daten zum Münchner Mietspiegel. Wir verwenden die Daten zum Mietspiegel 1999 zusammen mit den Fortschreibungsdaten aus dem Jahr 2001, vergleiche auch Beispiel 3.7 (Seite 85). Man betrachte das Regressionsmodell

$$\begin{aligned} \mathit{mieteqm}_i = \beta_0 + \beta_1\, \mathit{flaecheinvz}_i + \beta_2\, \mathit{bjahro}_i + \beta_3\, \mathit{bjahro2}_i + \\ \beta_4 \cdot \mathit{bjahro3}_i + \beta_5\, \mathit{nkueche} + \beta_6\, \mathit{gkueche} + \beta_7\, \mathit{zeit01} + \varepsilon_i, \end{aligned} \tag{3.24}$$

wobei *flaecheinvz* die um Null zentrierte Transformation $1/\mathit{flaeche}$ ist und *bjahro*, *bjahro2*, *bjahro3* orthogonale Polynome dritten Grades des Baujahrs sind. Die Dummy-Variable *zeit01* gibt an, ob eine Beobachtung aus dem Jahr 1999 ($\mathit{zeit01} = 0$) oder dem Fortschreibungsjahr 2001 ($\mathit{zeit01} = 1$) stammt. Als Schätzung erhalten wir

$$\begin{aligned} \widehat{\mathit{mieteqm}}_i = 13.56 + 243.19\, \mathit{flaecheinvz}_i + 1.43\, \mathit{bjahro}_i + 0.85\, \mathit{bjahro2}_i - \\ 0.02\, \mathit{bjahro3}_i + 2.04\, \mathit{nkueche} + 2.55\, \mathit{gkueche} - 0.37\, \mathit{zeit01}. \end{aligned}$$

Beim Blick auf die Schätzergebnisse stellen wir fest, dass die durchschnittliche Nettomiete 2001 im Vergleich zu 1999 um 0.37 DM sinkt. Diese Veränderung ist betragsmäßig relativ klein und darüber hinaus ist ein Absinken der Nettomieten zunächst überraschend. Es stellt sich hier die Frage, ob das in der Stichprobe beobachtete Absinken der Durchschnittsmiete auf die Grundgesamtheit übertragbar ist. Wir wollen also untersuchen, ob der Regressionsparameter β_7 signifikant von Null verschieden ist. Dies geschieht durch Überprüfung der Hypothesen

$$H_0 : \beta_7 = 0 \qquad \text{gegen} \qquad H_1 : \beta_7 \neq 0$$

mit einem geeigneten statistischen Test. Selbstverständlich wollen wir auch bei den anderen Variablen testen, ob deren Berücksichtigung im Regressionsmodell sinnvoll ist. Die Überprüfung der Signifikanz der Variable *kueche* mit den drei Kategorien „unterdurchschnittliche Küche“ (Referenzkategorie), „normale Küche“ (Dummy-Variable *nkueche*)

und „gehobene Küche“ (Dummy-Variable *gkueche*) ist komplizierter, da es sich um eine drei-kategoriale Variable handelt. Hier müssen die Hypothesen

$$H_0 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{gegen} \qquad H_1 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

statistisch überprüft werden.

Grundsätzlich besteht beim Mietspiegel das Ziel, mit möglichst wenigen Merkmalen auszukommen, um eine einfache Handhabung zu gewährleisten. Vor diesem Hintergrund stellt sich auch die Frage, ob eine Unterscheidung zwischen normalen und gehobenen Küchen überhaupt notwendig ist, da die geschätzten Regressionskoeffizienten für normale und gehobene Küchen sich nicht sehr stark unterscheiden. Die entsprechenden statistischen Hypothesen lauten

$$H_0 : \beta_5 = \beta_6 \qquad \text{gegen} \qquad H_1 : \beta_5 \neq \beta_6$$

oder äquivalent

$$H_0 : \beta_5 - \beta_6 = 0 \qquad \text{gegen} \qquad H_1 : \beta_5 - \beta_6 \neq 0.$$

△

Bei den betrachteten Fragestellungen handelt es sich um Beispiele der am häufigsten vorkommenden statistischen Hypothesen:

1. *Test auf Signifikanz einer Einflussvariable*

$$H_0 : \beta_j = 0 \qquad \text{gegen} \qquad H_1 : \beta_j \neq 0.$$

2. *Test eines Subvektors* $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)'$

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \qquad \text{gegen} \qquad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}.$$

3. *Test auf Gleichheit*

$$H_0 : \beta_j - \beta_r = 0 \qquad \text{gegen} \qquad H_1 : \beta_j - \beta_r \neq 0.$$

Die betrachteten Testprobleme 1-3 können als Spezialfälle des Tests allgemeiner *linearer Hypothesen*

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \qquad \text{gegen} \qquad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d} \tag{3.25}$$

aufgefasst werden. Dabei ist $\boldsymbol{C}$ eine $r \times p$ Matrix mit $\text{rg}(\boldsymbol{C}) = r \leq p$. Das bedeutet, dass unter H_0 insgesamt r linear unabhängige Bedingungen gelten. Beim Test auf Signifikanz einer Einflussvariable ist $\boldsymbol{d} = 0$ ein Skalar und $\boldsymbol{C}$ eine $1 \times p$ Matrix gegeben durch $\boldsymbol{C} = (0, \ldots, 0, 1, 0, \ldots, 0)$. Dabei befindet sich die Eins an der $(j+1)$-ten Spalte der Matrix. Beim Test der ersten r Komponenten erhalten wir speziell den r-dimensionalen Vektor $\boldsymbol{d} = \mathbf{0}$ und die $r \times p$ Matrix

$$\boldsymbol{C} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{pmatrix}.$$

Abb. 3.15. *Illustration des Unterschieds in der Anpassungsgüte für die KQ-Schätzung ohne und mit Beachtung der Restriktion* $0 \leq \beta \leq 1$. *Der (unrestringierte) KQ-Schätzer ist mit* $\hat{\beta}$ *gekennzeichnet, als restringierte Lösung ergibt sich die optimale Schätzung als* $\hat{\beta} = 1$.

Schließlich erhalten wir beim Test auf Gleichheit zweier Komponenten den Skalar $\boldsymbol{d} = 0$ sowie die $1 \times p$ Matrix $\boldsymbol{C} = (0, \ldots, 1, \ldots, -1, \ldots, 0)$. Die Eins befindet sich hier an der $(j+1)$-ten Stelle und die minus Eins an der $r+1$-ten Stelle.

Im nächsten Abschnitt entwickeln wir einen Test für allgemeine lineare Hypothesen (3.25). Die genannten Testprobleme 1–3 sind dann als Spezialfälle enthalten.

3.3.1 F-Test

Bei der Herleitung einer geeigneten Teststatistik für das allgemeine Testproblem (3.25) gehen wir wie folgt vor:

1. Bestimme die Residuenquadratsumme $\mathrm{SSE} = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ im vollen Modell.
2. Berechne die Residuenquadratsumme $\mathrm{SSE}_{H_0} = \hat{\boldsymbol{\varepsilon}}'_{H_0}\hat{\boldsymbol{\varepsilon}}_{H_0}$ im Modell unter der Nullhypothese, d.h. wenn die Restriktion $\boldsymbol{C\beta} = \boldsymbol{d}$ erfüllt ist.
3. Betrachte für die Teststatistik einen Ausdruck der Form
$$\frac{\Delta\mathrm{SSE}}{\mathrm{SSE}} = \frac{\mathrm{SSE}_{H_0} - \mathrm{SSE}}{\mathrm{SSE}}, \tag{3.26}$$
d.h. die relative Differenz zwischen den Residuenquadratsummen im restringierten und im vollen Modell.

Generell gilt: Bei Beachtung einer Restriktion ist die Anpassung an die Daten höchstens so gut wie ohne Restriktion. Dieser Sachverhalt ist in Abbildung 3.15 für den Fall eines linearen Modells $y = \beta x + \varepsilon$ mit nur einem Parameter β illustriert. Die Abbildung zeigt die Residuenquadratsumme $\mathrm{KQ}(\beta)$ in Abhängigkeit von β. Zusätzlich eingezeichnet ist der KQ-Schätzer $\hat{\beta} = 1.78$. Wenn die Restriktion $0 \leq \beta \leq 1$ beachtet werden muss, können die Schätzungen nur noch aus den Parametern gewählt werden, die innerhalb der beiden vertikalen Linien liegen. Für diesen eingeschränkten Parameterraum wird dann die Residuenquadratsumme minimal für $\hat{\beta} = 1$.

Damit wissen wir, dass die Differenz $\mathrm{SSE}_{H_0} - \mathrm{SSE}$ stets größer oder gleich Null ist, da die Anpassung an die Daten bei Beachtung der Restriktion $\boldsymbol{C\beta} = \boldsymbol{d}$ höchstens so gut wie

bei der unrestringierten KQ-Schätzung ist. Einen formalen Beweis für $\Delta\text{SSE} \geq 0$ liefern wir im Appendix auf Seite 184.

Nach diesen Überlegungen ist die Idee hinter der Teststatistik klar. Je geringer die Differenz zwischen SSE_{H_0} und SSE ist, desto näher liegen in Abbildung 3.15 die beiden Minima beieinander und die Nullhypothese wird eher beibehalten. Je größer die Differenz ist, desto eher werden wir die Nullhypothese verwerfen.

Die tatsächlich verwendete Teststatistik ist

$$F = \frac{\frac{1}{r}\Delta\text{SSE}}{\frac{1}{n-p}\text{SSE}} = \frac{n-p}{r}\frac{\Delta\text{SSE}}{\text{SSE}},$$

wobei r die Anzahl der Restriktionen ist, d.h. die Anzahl der Zeilen in $\boldsymbol{C}$. Die Hinzunahme des Vorfaktors $\frac{n-p}{r}$ spielt für die Interpretation keine Rolle. Dadurch wird lediglich sichergestellt, dass die Verteilung der Teststatistik unter H_0 leicht bestimmt und damit der Test durchgeführt werden kann.

Bei der Herleitung der Verteilung gehen wir in vier Schritten vor:

1. *Berechnung des KQ-Schätzers unter H_0*

   In Abschnitt 3.7.2 (Seite 184) des Kapitels bestimmen wir mit Hilfe eines Lagrange-Ansatzes den KQ-Schätzer $\hat{\boldsymbol{\beta}}^R$ unter H_0, d.h. unter Beachtung der Restriktion $\boldsymbol{C\beta} = \boldsymbol{d}$. Wir erhalten

   $$\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'(\boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\,)^{-1}(\boldsymbol{C}\,\hat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

2. *Bestimmung der Differenz der Residuenquadratsummen*

   Wiederum in Abschnitt 3.7.2 berechnen wir die Differenz ΔSSE der Residuenquadratsummen. Als Ergebnis erhalten wir

   $$\Delta\text{SSE} = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

3. *Stochastische Eigenschaften der Differenz der Residuenquadratsummen*

   Bei der Bestimmung der Verteilung der Teststatistik unter H_0 benötigen wir die folgenden stochastische Eigenschaften von ΔSSE:

   a) $\text{E}(\Delta\text{SSE}) = r\sigma^2 + (\boldsymbol{C\beta} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C\beta} - \boldsymbol{d})$ (unabhängig davon, ob die Restriktion erfüllt ist oder nicht).

   b) Unter H_0 gilt: $\dfrac{1}{\sigma^2}\cdot\Delta\text{SSE} \sim \chi^2_r$.

   c) ΔSSE und SSE sind stochastisch unabhängig.

   Beweise der Aussagen findet man in Abschnitt 3.7.2.

4. *Verteilung der Teststatistik*

   Unter Verwendung der stochastischen Eigenschaften in 3. können wir jetzt die Verteilung der Teststatistik unter der Nullhypothese bestimmen: Nach Eigenschaft 3b) gilt unter H_0

   $$\frac{1}{\sigma^2}\Delta\text{SSE} \sim \chi^2_r.$$

   Außerdem gilt

$$\frac{1}{\sigma^2}\text{SSE} \sim \chi^2_{n-p},$$

vergleiche den Kasten auf Seite 108. Darüberhinaus sind ΔSSE und SSE gemäß Eigenschaft 3c) stochastisch unabhängig. Damit ist die Teststatistik unter H_0 F-verteilt mit r und $n-p$ Freiheitsgraden, d.h. $F \sim F_{r,n-p}$. Dies folgt unmittelbar aus der Definition der F-Verteilung, vergleiche Seite 461 in Anhang B.

Damit kommen wir zu folgendem Test: Wenn α das Signifikanzniveau bezeichnet, wird die Nullhypothese abgelehnt, falls die Teststatistik größer ist als das $(1-\alpha)$-Quantil der entsprechenden F-Verteilung. Im vorliegenden Fall also, falls

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

Zusammenhang mit dem Wald-Test

Wir zeigen jetzt noch eine interessante Verbindung zum Wald-Test auf. Zur allgemeinen Idee des Wald-Tests vergleiche Abschnitt B.4.4 (Seite 475) in Anhang B.
Aus der Herleitung der Verteilung der Teststatistik des F-Tests ergibt sich die Beziehung

$$\Delta\text{SSE} = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

Mit $\text{SSE} = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (n-p)\hat{\sigma}^2$ erhalten wir dann

$$\begin{aligned} F &= \frac{(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\hat{\sigma}^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})}{r} \\ &= \frac{(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'\widehat{\text{Cov}(\boldsymbol{C}\hat{\boldsymbol{\beta}})}^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})}{r} \end{aligned}$$

Diese Darstellung liefert eine weitere interessante Interpretation der Teststatistik. Offenbar vergleicht F die Differenz zwischen der Schätzung $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ und dem hypothetischen Wert $\boldsymbol{d}$, wobei mit der Inversen der geschätzten Kovarianzmatrix

$$\widehat{\text{Cov}(\boldsymbol{C}\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'$$

von $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ gewichtet wird. Bei $\hat{\sigma}^2$ handelt es sich um die Schätzung von σ^2 basierend auf dem unrestringierten Modell. Eine analoge Vorgehensweise findet man beim Wald-Test, und wir erhalten den Zusammenhang

$$W = rF$$

für die Teststatistik des Wald-Tests.
Darüberhinaus lässt sich der F-Test auch als Likelihood-Quotienten-Test herleiten.

F-Test für einige spezielle Testprobleme

Im Folgenden wollen wir die Teststatistik des F-Tests für einige spezielle Testprobleme etwas genauer betrachten:

1. Test auf Signifikanz einer Einflussvariable (t-Test)

$$H_0 : \beta_j = 0 \qquad \text{gegen} \qquad H_1 : \beta_j \neq 0 \qquad j = 1, \ldots, p$$

In diesem speziellen Fall kann man zeigen, dass

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

gilt. Äquivalent dazu kann man den Test auch auf der „t-Statistik“

$$t = \frac{\hat{\beta}_j}{se_j} \tag{3.27}$$

aufbauen, wobei $se_j = \widehat{\mathrm{Var}(\hat{\beta}_j)}^{1/2}$ eine Abkürzung für die geschätzte Standardabweichung von $\hat{\beta}_j$ ist. Die ursprüngliche Teststatistik F kann als Quadrat von t aufgefasst werden. Die Teststatistik t ist t-verteilt mit $n-p$ Freiheitsgraden. Den kritischen Wert für den Ablehnbereich der Nullhypothese erhält man bei dieser Vorgehensweise als $(1-\alpha/2)$-Quantil einer t-Verteilung mit $n-p$ Freiheitsgraden. Die Nullhypothese wird also abgelehnt, wenn

$$|t| > t_{1-\alpha/2}(n-p)$$

gilt. Völlig analog können die etwas allgemeineren Hypothesen

$$H_0 : \beta_j = d_j \qquad \text{gegen} \qquad H_1 : \beta_j \neq d_j \qquad j = 1, \ldots, k$$

unter Verwendung der modifizierten Teststatistik

$$t = \frac{\beta_j - d_j}{se_j}$$

getestet werden.

2. Test eines r-dimensionalen Subvektors $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)'$:

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \qquad \text{gegen} \qquad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}.$$

In diesem Fall erhält man als Teststatistik

$$F = \frac{1}{r} \hat{\boldsymbol{\beta}}_1' \widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1} \hat{\boldsymbol{\beta}}_1 \sim F_{r,n-p}. \tag{3.28}$$

Die geschätzte Kovarianzmatrix der Teilvektors $\hat{\boldsymbol{\beta}}_1$ besteht aus den entsprechenden Elementen der geschätzten Gesamtkovarianzmatrix $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

3. Testen der Hypothese „kein funktionaler Zusammenhang“ :
Getestet werden soll die Hypothese

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

d.h. keine der Kovariablen besitzt einen Einfluss. Vorsicht: Die Gegenhypothese bedeutet nicht, dass *alle* Variablen einen Einfluss besitzen. Sie besagt lediglich, dass

mindestens eine der k Kovariablen einen Einfluss besitzt. Unter H_0 besteht die KQ-Schätzung nur aus einer Schätzung für β_0 mit $\hat{\beta}_0 = \bar{y}$. Folglich erhalten wir für die Residuenquadratsumme SSE_{H_0} unter der Nullhypothese

$$\mathrm{SSE}_{H_0} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Für die Differenz zwischen der Residuenquadratsumme unter H_0 und derjenigen im vollen Modell gilt unter Verwendung der Streuungszerlegung (3.18) (vergleiche Seite 98)

$$\Delta\mathrm{SSE} = \mathrm{SSE}_{H_0} - \mathrm{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

Damit folgt für die Teststatistik F:

$$\begin{aligned} F &= \frac{n-p}{k} \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum \hat{\varepsilon}_i^2} \\ &= \frac{n-p}{k} \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2} \\ &= \frac{n-p}{k} \frac{\sum (\hat{y}_i - \bar{y})^2 / \sum (y_i - \bar{y})^2}{1 - \sum (\hat{y}_i - \bar{y})^2 / \sum (y_i - \bar{y})^2} \\ &= \frac{n-p}{k} \frac{R^2}{1 - R^2} \end{aligned}$$

Die Teststatistik besitzt damit folgende interessante Interpretation: Bei einem kleinen Bestimmtheitsmaß R^2 wird die Hypothese „kein funktionaler Zusammenhang“ eher beibehalten (da F dann klein ist) als bei einem Bestimmtheitsmaß nahe bei Eins (in diesem Fall ist F vergleichsweise groß).

Beispiel 3.13 Mietspiegel – Standardoutput und Hypothesentests

Wir sind jetzt in der Lage den üblichen Output in statistischen Programmpaketen zu verstehen. Die folgende Tabelle 3.2 zeigt Schätzergebnisse für das Regressionsmodell (3.24) auf Seite 111.

Die Tabelle besteht aus sechs Spalten. Von links nach rechts sind abgedruckt: die Variablennamen, die geschätzten Regressionskoeffizienten $\hat{\beta}_j$, die geschätzten Standardfehler se_j, die Teststatistiken t der Tests für $H_0 : \beta_j = 0$ gegen $H_1 : \beta_j \neq 0$, die zu diesen Teststatistiken gehörenden p-Werte sowie die jeweiligen 95% Konfidenzintervalle (vergleiche hierzu den Abschnitt 3.3.2 auf Seite 119). Die t-Statistik ergibt sich dabei gemäß (3.27) als Quotient aus dem geschätzten Regressionskoeffizienten (2. Spalte) und dem Standardfehler (3. Spalte). Der p-Wert gibt bekanntlich das minimale Signifikanzniveau α an, zu dem die Nullhypothese $H_0 : \beta_j = 0$ verworfen werden kann.

Zunächst wird klar, dass die Verringerung der durchschnittlichen Nettomiete um 0.37 DM im Fortschreibungsjahr 2001 im Vergleich zu 1999 signifikant ist. Die Hypothese $H_0 : \beta_7 = 0$ kann zu jedem Signifikanzniveau $\alpha > 0.002$ abgelehnt werden. Tatsächlich

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	13.566	0.074	184.50	<0.001	13.422	13.710
flaecheinv	243.189	8.695	27.97	<0.001	226.142	260.236
bjahro	1.426	0.059	24.29	<0.001	1.311	1.541
bjahro2	0.852	0.059	14.53	<0.001	0.737	0.967
bjahro3	-0.022	0.057	-0.38	0.701	-0.134	0.090
nkueche	2.042	0.199	10.28	<0.001	1.653	2.432
gkueche	2.554	0.298	8.57	<0.001	1.970	3.138
zeit	-0.376	0.122	-3.08	0.002	-0.614	-0.137

Tabelle 3.2. *Ergebnistabelle für das Regressionsmodell (3.24)* .

gab es im Erhebungszeitraum für die Mietspiegelfortschreibung einen kleinen Einbruch im Mietwohnungsmarkt, so dass der geschätzte Abfall gut erklärbar ist.

Zur Überprüfung der Relevanz der Küchenausstattung testen wir die Hypothesen

$$H_0 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{gegen} \qquad H_1 : \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Zur Berechnung der Teststatistik (3.28) wird die geschätzte Kovarianzmatrix $\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}})}$ von $\hat{\boldsymbol{\beta}}$ benötigt. Sie ist gegeben durch

$$\begin{pmatrix} 0.00541 & & & & & & & \\ 0.02802 & 75.60869 & & & & & & \\ 0.00029 & -0.07314 & 0.00345 & & & & & \\ 0.00029 & 0.07990 & 0.00004 & 0.00344 & & & & \\ -0.00013 & 0.00788 & 0.00000 & 0.00002 & 0.00327 & & & \\ -0.00420 & -0.35190 & -0.00148 & -0.00152 & 0.00036 & 0.03947 & & \\ -0.00432 & 0.09103 & -0.00240 & -0.00279 & -0.00090 & 0.00538 & 0.08881 & \\ -0.00486 & 0.00993 & -0.00013 & -0.00009 & 0.00040 & 0.00021 & 0.00069 & 0.01486 \end{pmatrix}.$$

Wegen der Symmetrie der Kovarianzmatrix sind nur die Elemente unterhalb der Hauptdiagonalen angegeben. Damit erhalten wir als Teststatistik

$$F = \frac{1}{2} \begin{pmatrix} 2.0423 \\ 2.5540 \end{pmatrix}' \begin{pmatrix} 0.03947 & 0.00538 \\ 0.00538 & 0.08881 \end{pmatrix}^{-1} \begin{pmatrix} 2.0423 \\ 2.5540 \end{pmatrix} = 82.24.$$

Bei einem Signifikanzniveau von $\alpha = 0.05$ ist das entsprechende Quantil der F-Verteilung gegeben durch

$$F_{2,4559-8}(0.95) = 3.00.$$

Da $F = 82.24 > 3.00 = F_{2,4551}(0.95)$, kann die Nullhypothese zu diesem Niveau verworfen werden, die Ausstattung der Küche besitzt also einen signifikanten Einfluss auf die durchschnittliche Nettomiete. Man beachte, dass der Test keinesfalls impliziert, dass für beide Dummy-Variablen der Variable *kueche* die Regressionskoeffizienten von Null verschieden sind. Es reicht, wenn mindestens ein Koeffizient ungleich Null ist. Ein Blick auf die p-Werte der Variablen *nkueche* und *gkueche* in Tabelle 3.2 zeigt aber, dass auch beide Regressionskoeffizienten (hoch) signifikant von Null verschieden sind. Die praktische Durchführung des Tests ist in der Praxis nicht so aufwändig wie hier Schritt für Schritt demonstriert. Die meisten Statistikprogramme erlauben die einfache und schnelle Durchführung linearer Tests. In `STATA` etwa hätte der Befehl `test nkueche gkueche` zum gewünschten Ergebnis geführt.

Als Abschluss dieses Beispiels beantworten wir noch die Frage, ob eine Unterscheidung zwischen normaler und gehobener Küche notwendig ist. Wir wollen also die Hypothesen

$$H_0 : \beta_5 - \beta_6 = 0 \quad \text{gegen} \quad H_1 : \beta_5 - \beta_6 \neq 0$$

testen. Als Teststatistik erhalten wir $F = 2.23$. Für $\alpha = 0.05$ ist das entsprechende Quantil der F-Verteilung gegeben durch $F_{1,4551}(0.95) = 3.84$. Da $F = 2.23 < 3.84 = F_{1,4551}(0.95)$, kann die Nullhypothese nicht verworfen werden, der Unterschied zwischen normaler und gehobener Küchenausstattung ist also nicht signifikant. Der zu diesem Test gehörende p-Wert ist $p = 0.13$, die Nullhypothese könnte also lediglich für ein Signifikanzniveau $\alpha > 0.13$ verworfen werden.

△

Asymptotische Eigenschaften des F-Tests

Die Anwendbarkeit des bisher betrachteten F-Tests ist zunächst an normalverteilte Störgrößen gebunden. Unter Ausnutzung des Zusammenhangs mit dem Wald-Test können wir aber zeigen, dass der Test auch bei nicht normalverteilten Störgrößen benutzt werden kann. Wie bereits gezeigt, ist die Teststatistik des Wald-Tests gegeben durch $W = rF$. Darüber hinaus gilt $W \overset{a}{\sim} \chi^2_r$, vergleiche hierzu Abschnitt B.4.4 in Anhang B. Da außerdem die $F_{r,n-p}$ Verteilung für $n \to \infty$ in Verteilung gegen eine χ^2_r/r Verteilung konvergiert, können (für hinreichend großen Stichprobenumfang) die kritischen Werte bzw. p-Werte des F-Tests auch bei nicht normalverteilten Störgrößen benutzt werden.

Die Konvergenz der F-Verteilung gegen eine χ^2-Verteilung kann man sich auf zweierlei Art plausibel machen: Zum einen ist der Nenner der F-Teststatistik gegeben durch den erwartungstreuen Schätzer für die Fehlervarianz σ^2. Da dieser Schätzer konsistent ist, konvergiert der Nenner asymptotisch gegen den wahren, festen Wert, so dass der Quotient zweier χ^2-Verteilungen gegen die χ^2_r/r-Verteilung des Nenners konvergiert. Andererseits gilt nach dem Gesetz der großen Zahlen, dass χ^2_n/n für $n \to \infty$ gegen 1 konvergiert. Da sich die χ^2_n-Verteilung als Summe von n unabhängig und identisch χ^2_1-verteilten Zufallsvariablen auffassen lässt, ist χ^2_n/n der zugehörige Mittelwert, der für $n \to \infty$ gegen den Erwartungswert der χ^2_1-Verteilung und damit gegen 1 konvergiert.

3.3.2 Konfidenzbereiche und Prognoseintervalle

Konfidenzintervalle und Ellipsoide für die Regressionskoeffizienten

Aufgrund der Dualität zwischen zweiseitigen Tests und Konfidenzintervallen bzw. Konfidenzbereichen kann man basierend auf den betrachteten Tests sofort Konfidenzintervalle für einen einzelnen Parameter β_j, $j = 0, \ldots, k$, bzw. Konfidenzellipsoide für einen Subvektor $\boldsymbol{\beta}_1$ von $\boldsymbol{\beta}$ konstruieren.

Zur Konstruktion eines Konfidenzintervalls für β_j unter Normalverteilungsannahme verwenden wir die Teststatistik $t = (\hat{\beta}_j - d_j)/se_j$ des Tests der Hypothese $H_0 : \beta_j = d_j$. Die Nullhypothese wird abgelehnt, falls $|t| > t_{n-p}(1 - \alpha/2)$. Der Test ist so konstruiert, dass die Wahrscheinlichkeit, H_0 abzulehnen, wenn H_0 wahr ist, gleich α ist. Es gilt also unter H_0

$$\mathrm{P}(|t| > t_{n-p}(1 - \alpha/2)) = \alpha.$$

Testen linearer Hypothesen

Hypothesen

1. Allgemeine lineare Hypothesen

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \qquad \text{gegen} \qquad H_0 : \boldsymbol{C\beta} \neq \boldsymbol{d}$$

   wobei $\boldsymbol{C}$ eine $r \times p$ Matrix ist mit $\mathrm{rg}(\boldsymbol{C}) = r \leq p$ (r linear unabhängige Restriktionen).
2. Test auf Signifikanz einer Einflussvariable (t-Test):

$$H_0 : \beta_j = 0 \qquad \text{gegen} \qquad H_1 : \beta_j \neq 0.$$

3. Test eines Subvektors:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{gegen} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}.$$

4. „Kein funktionaler Zusammenhang“ :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ gegen}$$
$$H_1 : \beta_j \neq 0 \text{ für mindestens ein } j \in \{1, \ldots, k\}.$$

Teststatistiken

Unter H_0 gilt bei normalverteilten Störgrößen:

1. $F = 1/r\,(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\hat{\sigma}^2 \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim F_{r,n-p}$.
2. $t = \dfrac{\hat{\beta}_j}{se_j} \sim t_{n-p}$.
3. $F = \frac{1}{r}(\hat{\boldsymbol{\beta}}_1)' \widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1} (\hat{\boldsymbol{\beta}}_1) \sim F_{r,n-p}$.
4. $F = \dfrac{n-p}{k} \dfrac{R^2}{1-R^2} \sim F_{k,n-p}$.

Testentscheidungen

H_0 wird abgelehnt, falls

1. $F > F_{r,n-p}(1-\alpha)$.
2. $|t| > t_{n-p}(1-\alpha/2)$.
3. $F > F_{r,n-p}(1-\alpha)$.
4. $F > F_{k,n-p}(1-\alpha)$.

Die Tests sind relativ robust gegenüber geringen Abweichungen von der Normalverteilung. Darüber hinaus sind die Tests für großen Stichprobenumfang auch bei nicht normalverteilten Störgrößen anwendbar.

Die Wahrscheinlichkeit, H_0 nicht abzulehnen, ist damit gegeben durch

$$\mathrm{P}(|t| < t_{n-p}(1-\alpha/2)) = 1 - \alpha.$$

Dies ist gleich bedeutend mit

$$\mathrm{P}(\hat{\beta}_j - t_{n-p}(1-\alpha/2)\cdot se_j < \beta_j < \hat{\beta}_j + t_{n-p}(1-\alpha/2)\cdot se_j) = 1-\alpha.$$

Somit erhalten wir das Intervall

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\cdot se_j, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\cdot se_j]$$

als $(1-\alpha)$-Konfidenzintervall für β_j.
Analog konstruiert man einen $(1-\alpha)$-Konfidenzbereich für einen r dimensionalen Subvektor $\boldsymbol{\beta}_1$ von $\boldsymbol{\beta}$, vergleiche den nachfolgenden Kasten.

Beispiel 3.14 Mietspiegel - Konfidenzintervall

Wir illustrieren die Berechnung von Konfidenzintervallen anhand des Regressionsmodells aus dem letzten Beispiel, vergleiche Seite 117. Ein 95% Konfidenzintervall für den Regressionskoeffizienten β_7 des Zeit-Dummies lässt sich unter Verwendung von $se_7 = 0.122$ (vgl. Tabelle 3.2) und $t_{n-p}(1-\alpha/2) = t_{4551}(0.975) = 1.96$ berechnen als

$$-0.376 \pm 1.96 \cdot 0.122 = [-0.615, -0.136].$$

Der kleine Unterschied zu dem in Tabelle 3.2 angegebenen Konfidenzintervall ergibt sich durch die im Vergleich dazu ungenauere Rundung.

△

Prognoseintervalle

In Abschnitt 3.2.3 auf Seite 103 haben wir eine optimale Prognose $\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$ für eine neue (zukünftige) Beobachtung an der Stelle $\boldsymbol{x}_0$ bestimmt. Genau genommen handelt es sich bei $\hat{y}_0$ um eine Schätzung für den (bedingten) Erwartungswert $\mathrm{E}(y_0) = \boldsymbol{x}_0'\boldsymbol{\beta} = \mu_0$ der zukünftigen Beobachtung y_0. Neben der Punktschätzung ist man in der Regel auch an einer Intervallschätzung interessiert. Ein Konfidenzintervall für μ_0 lässt sich leicht konstruieren. Wegen $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ gilt für die Linearkombination

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{x}_0'\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0).$$

Durch Standardisieren erhalten wir

$$\frac{\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - \mu_0}{\sigma(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}} \sim N(0,1).$$

Wenn wir σ^2 durch die Schätzung $\hat{\sigma}^2$ ersetzen ist der resultierende Ausdruck t-verteilt mit $n-p$ Freiheitsgraden und es gilt

$$\mathrm{P}\left(-t_{n-p}(1-\alpha/2) \le \frac{\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - \mu_0}{\hat{\sigma}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}} \le t_{n-p}(1-\alpha/2)\right) = 1-\alpha.$$

Somit erhält man

$$[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{n-p}(1-\alpha/2)\hat{\sigma}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}, \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}]$$

als Konfidenzintervall für μ_0 zum Niveau $1-\alpha$.

Abb. 3.16. *Geschätzte Miete pro Quadratmeter in Abhängigkeit von der Wohnfläche inklusive 95% Konfidenzintervall (durchgezogene Linien) und 95% Prognoseintervall (gestrichelte Linien). Die Werte der übrigen Kovariablen wurden bei bjahr* $= 1918$*, nkueche* $= 0$*, gkueche* $= 0$ *und zeit01* $= 0$ *fixiert. Zusätzlich eingezeichnet sind die für diese Kovariablenkonstellation vorhandenen Beobachtungen.*

In vielen Fällen ist man darüber hinaus daran interessiert, ein Intervall anzugeben, in dem mit großer Wahrscheinlichkeit die zukünftige Beobachtung y_0 liegt. Beim Mietspiegel etwa möchte ein Mietinteressent vor allem wissen, in welchem Intervall sich die Miete für eine Wohnung mit den angestrebten Wohnungseigenschaften mit großer Wahrscheinlichkeit bewegt. Gesucht ist also ein Prognoseintervall für die zukünftige Beobachtung y_0. Dazu betrachten wir den Prognosefehler $\hat{\varepsilon}_0 = y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$. Dafür gilt

$$\hat{\varepsilon}_0 \sim N(0, \sigma^2 + \sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0).$$

Durch Standardisieren und Einsetzen der üblichen Schätzung $\hat{\sigma}^2$ für σ^2 erhalten wir

$$\frac{y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}}{\hat{\sigma}(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}} \sim t_{n-p}$$

und damit

$$P\left(-t_{n-p}(1-\alpha/2) \leq \frac{y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}}{\hat{\sigma}(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}} \leq t_{n-p}(1-\alpha/2)\right) = 1 - \alpha.$$

Somit liegt eine zukünftige Beobachtung mit Wahrscheinlichkeit $1-\alpha$ im Prognoseintervall

$$[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{n-p}(1-\alpha/2)\hat{\sigma}(1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}, \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\hat{\sigma}(1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}].$$

Per Konstruktion ist das Prognoseintervall stets breiter als das entsprechende Konfidenzintervall für μ_0. In Anwendungen mit großer Varianz der Störgrößen sind die Intervalle sogar erheblich breiter, vergleiche das nachfolgende Beispiel 3.15 zum Mietspiegel.

Obwohl die beiden konstruierten Intervalle auf den ersten Blick sehr ähnlich sind, handelt es sich um völlig unterschiedliche Konstrukte. Im ersten Fall haben wir ein Konfidenzintervall für $E(y_0) = \mu_0$ konstruiert. Das bedeutet, dass das zufällige Intervall mit Wahrscheinlichkeit $1-\alpha$ den unbekannten (festen, nicht zufälligen) Erwartungswert $\mathrm{E}(y_0)$

Konfidenzbereiche und Prognoseintervalle

Unter der Voraussetzung (zumindest approximativ) normalverteilter Störgrößen bzw. für großen Stichprobenumfang erhalten wir die folgenden Konfidenzintervalle bzw. Bereiche und Prognoseintervalle:

Konfidenzintervall für β_j

Ein Konfidenzintervall für β_j zum Niveau $1-\alpha$ ist gegeben durch

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\cdot se_j, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\cdot se_j].$$

Konfidenzellipsoid für einen Subvektor $\boldsymbol{\beta}_1$

Ein Konfidenzellipsoid für $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)'$ zum Niveau $1-\alpha$ ist gegeben durch

$$\left\{\boldsymbol{\beta}_1 \,:\, \frac{1}{r}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)'\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \leq F_{r,n-p}(1-\alpha)\right\}.$$

Konfidenzintervall für μ_0

Ein Konfidenzintervall für $\mu_0 = \mathrm{E}(y_0)$ einer zukünftigen Beobachtung y_0 an der Stelle $\boldsymbol{x}_0$ ist gegeben durch

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}.$$

Prognoseintervall

Ein Prognoseintervall für eine zukünftige Beobachtung y_0 and der Stelle $\boldsymbol{x}_0$ zum Niveau $1-\alpha$ ist gegeben durch

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}.$$

überdeckt. Im zweiten Fall haben wir ein Intervall konstruiert, in das die zukünftige Beobachtung y_0 mit großer Wahrscheinlichkeit (genauer $1-\alpha$) fallen wird.

Beispiel 3.15 Mietspiegel - Prognoseintervalle

Wir gehen wieder von dem Modell aus Beispiel 3.13 (Seite 117) aus. Abbildung 3.16 zeigt für Wohnungen aus dem Jahr 1918 mit durchschnittlicher Küche (*nkueche* = 0, *gkueche* = 0) und Erhebungszeitpunkt 1999 die geschätzte Miete pro Quadratmeter in Abhängigkeit von der Wohnfläche inklusive 95% Konfidenzintervalle und 95% Prognoseintervalle. Die einzelnen Konfidenzintervalle wurden mit Linien verbunden, so dass sich die gezeigten Bänder ergeben. Aufgrund der großen Streuung in den Mietspiegeldaten sind die Prognoseintervalle erheblich breiter als die Konfidenzintervalle für $E(y_0)$.

△

3.4 Das allgemeine lineare Regressionsmodell

In den ersten drei Abschnitten dieses Kapitels haben wir uns ausführlich mit dem klassischen linearen Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ mit unkorrelierten und homoskedastischen Störgrößen, d.h. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, befasst. Bei der Diskussion der Modellannahmen haben wir festgestellt, dass die Annahme unkorrelierter und homoskedastischer Störgrößen nicht immer erfüllt ist. In diesem Abschnitt wollen wir daher die Klasse der linearen Modelle dahingehend erweitern, dass auch heteroskedastische oder korrelierte Störgrößen zugelassen werden. Die entstehende Modellklasse wird als *allgemeines lineares Modell* bezeichnet. Das bisher betrachtete klassische lineare Modell ist dabei ein wichtiger Spezialfall. Es stellt sich nämlich heraus, dass viele Inferenzprobleme durch Rückführung auf das klassische lineare Modell gelöst werden können.

3.4.1 Modelldefinition

Im allgemeinen linearen Modell ersetzen wir

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$$

durch die allgemeinere Annahme

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{W},$$

wobei $\boldsymbol{W}$ eine positiv definite Matrix sei. Im Fall heteroskedastischer und (nach wie vor) unkorrelierter Störgrößen erhalten wir speziell

$$\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_n).$$

Für die heteroskedastischen Varianzen der Störgrößen gilt dann $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 w_i$.
Mit der Einführung einer allgemeineren Modellklasse werden in der Regel kompliziertere Inferenztechniken benötigt als im einfacheren Spezialfall. Es stellt sich daher immer die Frage, ob die Verwendung des allgemeineren Modells überhaupt notwendig ist. Wir wollen daher untersuchen, welche Auswirkungen die Verwendung der vergleichsweise einfachen Inferenztechniken des klassischen linearen Modells bei Vorliegen eines allgemeinen linearen Modells, d.h. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{W}$ anstatt $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, hat. Zunächst beschäftigen wir uns mit der Frage, welche Auswirkungen die Verwendung des gewöhnlichen KQ-Schätzers $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ hat, wenn ein verallgemeinertes lineares Modell vorliegt. Eine analoge Rechnung wie für das klassische lineare Modell (Seite 101) zeigt

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \qquad \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{3.29}$$

Der für das klassische lineare Modell entwickelte KQ-Schätzer ist also bei Vorliegen eines allgemeinen linearen Modells nach wie vor erwartungstreu für $\boldsymbol{\beta}$. Jedoch stimmt die Kovarianzmatrix im Allgemeinen nicht mehr mit der im klassischen Modell berechneten Kovarianzmatrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ überein. Damit sind alle Größen, die auf der Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ beruhen, falsch. Insbesondere erhalten wir keine korrekten Varianzen und Standardfehler für die geschätzten Regressionskoeffizienten und damit auch fehlerhafte Tests und Konfidenzintervalle.

Das allgemeine lineare Regressionsmodell

Das Modell

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

heißt allgemeines lineares Regressionsmodell, wenn folgende Annahmen gelten:

1. $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$.
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{W}$, wobei $\boldsymbol{W}$ eine bekannte positiv definite Matrix sei.
3. Die Designmatrix $\boldsymbol{X}$ besitzt vollen Spaltenrang, d.h. $\mathrm{rg}(\boldsymbol{X}) = p$.

Wir sprechen von allgemeiner Normalregression, wenn zusätzlich die Annahme

4. $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{W})$.

getroffen wird.

In den folgenden Unterabschnitten befassen wir uns mit verschiedenen Möglichkeiten, verbesserte Schätzer im allgemeinen linearen Modell zu erhalten. Abschnitt 3.4.2 entwickelt einen *gewichteten KQ-Schätzer* als Verallgemeinerung des gewöhnlichen KQ-Schätzers. Dieser besitzt dann dieselben Güteeigenschaften wie der ungewichtete KQ-Schätzer im klassischen linearen Modell. Voraussetzung für die Anwendbarkeit des gewichteten KQ-Schätzers ist jedoch, dass die Matrix $\boldsymbol{W}$ bekannt ist. In der Praxis wird das in den seltensten Situationen der Fall sein. Die Abschnitte 3.4.3 und 3.4.4 befassen sich daher mit Inferenztechniken, wenn $\boldsymbol{W}$ nicht bekannt ist. Dabei beschränken wir uns auf zwei wichtige Spezialfälle, heteroskedastische bzw. autokorrelierte Störgrößen.

3.4.2 Gewichtete Methode der kleinsten Quadrate

Im Folgenden entwickeln wir ein Schätzverfahren, das die oben angesprochenen Probleme bei Verwendung des gewöhnlichen KQ-Schätzers umgeht. Die einfache Idee dabei ist, abhängige Variable, Designmatrix und Störgrößen so zu *transformieren*, dass die transformierten Größen einem klassischen linearen Modell gehorchen. Zur Illustration betrachten wir zunächst ein Modell mit unkorrelierten heteroskedastischen Fehlern, d.h. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{W} = \sigma^2\mathrm{diag}(w_1, \ldots, w_n)$. Multiplikation der Störgrößen ε_i mit $1/\sqrt{w_i}$ liefert transformierte Störgrößen $\varepsilon_i^* = \varepsilon_i/\sqrt{w_i}$ mit konstanten Varianzen $\mathrm{Var}(\varepsilon_i^*) = \sigma^2$. Damit das Modell unverändert bleibt, müssen auch die Zielvariable und sämtliche Kovariablen (inklusive der Konstante) entsprechend transformiert werden. Wir erhalten $y_i^* = y_i/\sqrt{w_i}$, $x_{i0}^* = 1/\sqrt{w_i}, x_{i1}^* = x_{i1}/\sqrt{w_i}, \ldots, x_{ik}^* = x_{ik}/\sqrt{w_i}$ und damit das klassische lineare Modell

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_k x_{ik}^* + \varepsilon_i^*$$

mit homoskedastischen Fehlern ε_i^*. Formal entspricht die Transformation einer Multiplikation der Modellgleichung $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ von links mit $\boldsymbol{W}^{-1/2} = \mathrm{diag}(1/\sqrt{w_1}, \ldots, 1/\sqrt{w_n})$, d.h.

$$\boldsymbol{W}^{-1/2}\boldsymbol{y} = \boldsymbol{W}^{-1/2}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}^{-1/2}\boldsymbol{\varepsilon}.$$

Unter Verwendung der transformierten Größen $\boldsymbol{y}^* = \boldsymbol{W}^{-1/2}\boldsymbol{y}$, $\boldsymbol{X}^* = \boldsymbol{W}^{-1/2}\boldsymbol{X}$ und $\boldsymbol{\varepsilon}^* = \boldsymbol{W}^{-1/2}\boldsymbol{\varepsilon}$ erhält man schließlich

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*. \tag{3.30}$$

Durch die Transformationen befinden wir uns wieder im Rahmen des klassischen linearen Modells. Als KQ-Schätzer erhalten wir

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{y}^* \\ &= \left(\boldsymbol{X}'\boldsymbol{W}^{-1/2}\boldsymbol{W}^{-1/2}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1/2}\boldsymbol{W}^{-1/2}\boldsymbol{y} \\ &= \left(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{y}.\end{aligned}$$

Bei diesem Schätzer handelt es sich um den sogenannten *Aitken-Schätzer*. Man kann zeigen, dass der Aitken-Schätzer die „gewichtete“ Residuenquadratsumme

$$\mathrm{GKQ}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{1}{w_i}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

minimiert. Wir sprechen daher auch von gewichteter Regression. Offenbar erhalten Beobachtungen mit größerer Varianz (w_i groß) ein kleineres Gewicht (w_i^{-1} klein) als Beobachtungen mit kleinerer Varianz.

Unter Normalverteilungsannahme lässt sich darüber hinaus zeigen, dass der gewichtete KQ-Schätzer mit dem ML-Schätzer für $\boldsymbol{\beta}$ übereinstimmt, d.h. $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}$. Als ML-Schätzer für σ^2 ergibt sich

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\boldsymbol{W}^{-1}\hat{\boldsymbol{\varepsilon}}.$$

Wie im klassischen linearen Modell ist dieser Schätzer verzerrt. Ein unverzerrter Schätzer ist gegeben durch

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\boldsymbol{W}^{-1}\hat{\boldsymbol{\varepsilon}} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{1}{w_i}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2.$$

Dieser Schätzer kann wieder als restringierter Maximum-Likelihood-Schätzer aufgefasst werden, vergleiche Abschnitt 3.2.2 (Seite 94). Sämtliche Herleitungen und Beweise verlaufen völlig analog zum klassischen linearen Modell. Ebenso können die für das klassische lineare Modell entwickelten Tests- und Konfidenzintervalle übertragen werden.

Diese soeben am Beispiel heteroskedastischer Störgrößen demonstrierte Vorgehensweise lässt sich problemlos auf beliebige Kovarianzmatrizen $\sigma^2\boldsymbol{W}$ übertragen. Dazu machen wir Gebrauch von einer „Wurzel“ $\boldsymbol{W}^{1/2}$ mit $\boldsymbol{W}^{1/2}(\boldsymbol{W}^{1/2})' = \boldsymbol{W}$ der Matrix $\boldsymbol{W}$. Die Matrix $\boldsymbol{W}^{1/2}$ ist nicht eindeutig, lässt sich aber beispielsweise aus der Spektralzerlegung

$$\boldsymbol{W} = \boldsymbol{P}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\,\boldsymbol{P}'$$

von $\boldsymbol{W}$ gewinnen (vgl. Satz A.22 (Seite 455) in Anhang A):

$$\boldsymbol{W}^{1/2} = \boldsymbol{P}\,\mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})\,\boldsymbol{P}'$$

Der Vektor der Zielgröße, die Designmatrix und der Vektor der Störgrößen können dann mit der Matrix

Schätzer im allgemeinen linearen Modell

Gewichteter KQ- bzw. ML-Schätzer für $\boldsymbol{\beta}$

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

Eigenschaften des gewichteten KQ-Schätzers

1. *Erwartungswert:* $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, der gewichtete KQ-Schätzer ist also erwartungstreu.
2. *Kovarianzmatrix:* $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\right)^{-1}$.
3. *Gauß-Markov-Theorem:* Unter allen linearen, erwartungstreuen Schätzern $\hat{\boldsymbol{\beta}}^L = \boldsymbol{A}\boldsymbol{y}$ besitzt der gewichtete KQ-Schätzer minimale Varianzen, d.h.

$$\mathrm{Var}(\hat{\beta}_j) \leq \mathrm{Var}(\hat{\beta}_j^L), \quad j = 0, \ldots, k.$$

REML-Schätzer für σ^2

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\boldsymbol{W}^{-1}\hat{\boldsymbol{\varepsilon}}.$$

Der REML Schätzer ist erwartungstreu.

$$\boldsymbol{W}^{-1/2} = \boldsymbol{P}\,\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_n}}\right)$$

transformiert werden. Dadurch erhalten wir wieder ein Modell der Form (3.30). Hierbei handelt es sich um um ein klassisches lineares Modell, denn

$$\mathrm{E}(\boldsymbol{\varepsilon}^*) = \mathrm{E}(\boldsymbol{W}^{-1/2}\boldsymbol{\varepsilon}) = \boldsymbol{W}^{-1/2}\,\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$$

und

$$\mathrm{Cov}(\boldsymbol{\varepsilon}^*) = \mathrm{E}(\boldsymbol{W}^{-1/2}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{W}^{-1/2}) = \sigma^2\boldsymbol{W}^{-1/2}\boldsymbol{W}\boldsymbol{W}^{-1/2} = \sigma^2\boldsymbol{I}.$$

Schließlich stellen wir fest, dass für den gewichteten KQ-Schätzer die selben stochastische Eigenschaften wie für den gewöhnlichen KQ-Schätzer gelten, vergleiche den nachfolgenden Kasten. Der Beweis verläuft analog zum ungewichteten KQ-Schätzer.

Eine erste Anwendung des gewichteten KQ-Schätzers ergibt sich für gruppierte Daten:

Gruppierte Daten

Bislang sind wir davon ausgegangen, dass *Individualdaten* oder *ungruppierte Daten* vorliegen, d.h. zu jedem Individuum oder Objekt i aus einer Stichprobe vom Umfang n liegt eine Beobachtung $(y_i, \boldsymbol{x}_i)$ vor. Jeder Wert y_i der Zielvariable und jeder Kovariablenvektor $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ gehört dann zu genau einer Einheit i:

$$\begin{array}{c} \text{Einheit } 1 \\ \vdots \\ \text{Einheit } i \\ \vdots \\ \text{Einheit } n \end{array} \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{i1} & & x_{ik} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

Falls mehrere Kovariablenvektoren bzw. Zeilen der Kovariablen-Datenmatrix identisch sind, können die Daten *gruppiert* werden: Nach Umsortieren und Zusammenfassen enthält die Datenmatrix nur noch Zeilen mit *verschiedenen* Kovariablenvektoren $\boldsymbol{x}_i$. Dazu werden die Anzahl n_i der Wiederholungen von $\boldsymbol{x}_i$ in der Original-Stichprobe der Individualdaten und das arithmetische Mittel $\bar{y}_i$ der entsprechenden individuellen Werte der Zielvariablen, die zu $\boldsymbol{x}_i$ beobachtet wurden, angegeben:

$$\begin{array}{c} \text{Gruppe } 1 \\ \vdots \\ \text{Gruppe } i \\ \vdots \\ \text{Gruppe } G \end{array} \begin{bmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_G \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_i \\ \vdots \\ \bar{y}_G \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & & \vdots \\ 1 & x_{G1} & \cdots & x_{Gk} \end{bmatrix}$$

Dabei ist G die Anzahl *verschiedener* Kovariablenvektoren in der Stichprobe und oft deutlich kleiner als der Stichprobenumfang n. Dies ist insbesondere der Fall, wenn die Kovariablen binär bzw. kategorial sind.

Gruppierte Daten können leicht im Rahmen des allgemeinen linearen Modells behandelt werden, indem wir speziell $\boldsymbol{y} = (\bar{y}_1, \ldots, \bar{y}_G)'$ und $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{diag}(1/n_1, \ldots, 1/n_G)$ setzen.

Weitere Anwendungen gewichteter KQ-Schätzer werden wir in den Kapiteln 4 und 7 kennen lernen. In Kapitel 4 ist der gewichtete KQ-Schätzer ein Baustein bei der iterativen Bestimmung des ML-Schätzers in generalisierten linearen Modellen. In Kapitel 7 spielt die gewichtete KQ-Schätzung vor allem bei sogenannten *lokalen Glättern* eine wichtige Rolle, vergleiche hierzu insbesondere Abschnitt 7.1.7.

In den nächsten beiden Unterabschnitten behandeln wir Situationen, in denen die Gewichtsmatrix $\boldsymbol{W}$ zumindest teilweise unbekannt ist. Speziell befassen wir uns mit heteroskedastischen und autokorrelierten Störgrößen.

3.4.3 Heteroskedastische Fehler

Gegenstand dieses Abschnitts ist das lineare Modell mit heteroskedastischen Fehlern. Die Kovarianzmatrix der Störungen ist also gegeben durch $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{diag}(w_1, \ldots, w_n)$. In der Literatur, vor allem der ökonometrischen Literatur, existieren eine Fülle von Schätz- und Modellierungsvarianten, die hier nicht alle vorgestellt werden können. Einen guten Überblick findet man beispielsweise in Greene (2000). Im Rahmen dieses Buches behandeln wir ein auf dem KQ-Prinzip basierendes zweistufiges Schätzverfahren und skizzieren kurz den ML-Schätzer. Beide Varianten haben den Nachteil, dass Kenntnisse über die *Art* der Heteroskedastizität vorliegen müssen. Wir beschreiben daher auch ein von White (1980) entwickeltes Verfahren, bei dem die Art der Heteroskedastizität nicht bekannt

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	895.903	5.131	174.600	<0.001	885.842	905.963
flaecheo	237.544	5.347	44.430	<0.001	227.060	248.028
bjahro	105.956	5.277	20.080	<0.001	95.609	116.303
bjahro2	61.394	5.203	11.800	<0.001	51.193	71.595
bjahro3	-0.387	5.131	-0.080	0.940	-10.448	9.674

Tabelle 3.3. *Mietspiegel: Schätzergebnisse für die ungewichtete Regression.*

sein muss. Bevor jedoch die verschiedenen Schätz- und Modellierungsverfahren vorgestellt werden, zeigen wir im Folgenden, wie heteroskedastische Fehler überhaupt entdeckt werden können.

Diagnose heteroskedastischer Fehler

Ausgangspunkt zur Diagnose heteroskedastischer Fehler ist immer die Schätzung eines klassischen linearen Modells und die anschließende Untersuchung der Residuen. In der Literatur finden sich zwei unterschiedliche Ansätze, wie Heteroskedastizität aufgedeckt werden kann. In der statistischen und eher biometrisch orientierten Literatur werden vor allem grafische Hilfsmittel, insbesondere Residualplots, vorgeschlagen. Die ökonometrische Literatur entwickelt eine Vielzahl statistischer Tests zur Aufdeckung von Heteroskedastizität. Wir beschreiben im Folgenden beide Varianten.

Residualplots

Zur Aufdeckung heteroskedastischer Fehler können die standardisierten oder studentisierten Residuen (vergleiche Seite 108) in einem Streudiagramm gegen die geschätzten Werte $\hat{y}_i$ bzw. die Kovariablen x_{ij} abgetragen werden. Dabei sollten auch Kovariablen mit einbezogen werden, die nicht im Modell enthalten sind. Die standardisierten bzw. studentisierten Residuen sind den gewöhnlichen Residuen vorzuziehen, da diese selbst heteroskedastisch sind mit $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ und daher zur Überprüfung der Homoskedastizität weniger gut geeignet sind, vergleiche hierzu Seite 107. Bei Homoskedastizität sollten die standardisierten bzw. studentisierten Residuen jeweils regellos und mit konstanter Variabilität um Null streuen. Andernfalls bestehen Hinweise auf heteroskedastische Varianzen.

Beispiel 3.16 Mietspiegel – Diagnose heteroskedastischer Fehler

Wir illustrieren die Diagnose heteroskedastischer Fehler anhand der Daten zum Mietspiegel. Dabei beschränken wir uns auf ein Modell mit der Nettomiete als Zielgröße und den Kovariablen Wohnfläche und Baujahr. Die Streudiagramme zwischen Nettomiete und Wohnfläche bzw. Baujahr in Abbildung 1.5 (Seite 15) lassen einen linearen Effekt der Wohnfläche und einen leicht nichtlinearen Baujahrseffekt vermuten. Wir modellieren den Effekt des Baujahrs mit Hilfe orthogonaler Polynome 3. Grades (und den Effekt der Wohnfläche mit einem orthogonalen Polynom ersten Grades). Wir gehen also zunächst von dem klassischen linearen Modell

$$miete_i = \beta_0 + \beta_1\, flaeche_i + \beta_2\, bjahro_i + \beta_3\, bjahro2_i + \beta_4\, bjahro3_i + \varepsilon_i \qquad (3.31)$$

Abb. 3.17. *Die Abbildungen a) und b) zeigen den Wohnflächen- und Baujahrseffekt inklusive partieller Residuen. In den Grafiken c), d) und e) findet man die studentisierten Residuen in Abhängigkeit von der geschätzten Nettomiete, der Wohnfläche und des Baujahrs.*

aus. Die Schätzergebnisse findet man in Tabelle 3.3. Abbildung 3.17 a) und b) zeigt die geschätzten Effekte der Wohnfläche und des Baujahrs (jeweils inklusive partieller Residuen). Die studentisierten Residuen in Abhängigkeit von geschätzter Nettomiete, Wohnfläche und Baujahr sind in den Grafiken c), d) und e) abgebildet. Wir erhalten deutliche Hinweise auf heteroskedastische Varianzen, wobei die Varianzen vermutlich nicht nur von der Wohnfläche sondern auch vom Baujahr abhängen.

△

Tests auf Heteroskedastizität

Tests auf Heteroskedastizität werden ausführlich in der ökonometrischen Literatur beschrieben, vergleiche beispielsweise Greene (2000) und Judge, Griffith, Hill, Lütkepohl & Lee (1980). Wir beschränken uns hier illustrativ auf die Beschreibung eines Tests nach Breusch und Pagan. Dem Test liegt ein multiplikatives Modell für die Fehlervarianz zugrunde:

$$\sigma_i^2 = \sigma^2 \cdot h(\alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_q z_{iq})$$

Dabei ist h eine nicht von i abhängige Funktion und $z_1, \ldots, z_q$ sind Kovariablen, denen ein möglicher Einfluss auf die Varianz unterstellt wird. Die Hypothese homoskedastischer Varianzen ist äquivalent zu $\alpha_1 = \cdots = \alpha_q = 0$. Beim Breusch-Pagan Test werden somit die Hypothesen

$$H_0 : \alpha_1 = \ldots = \alpha_q = 0 \quad \text{gegen} \quad H_1 : \alpha_j \neq 0 \text{ für mindestens ein } j$$

getestet. Zur Durchführung des Tests wird eine Hilfsregression zwischen der Zielgröße

$$g_i = \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}_{ML}^2}$$

und den erklärenden Variablen $z_1, \ldots, z_q$ durchgeführt. Bei $\hat{\varepsilon}_i$ und $\hat{\sigma}_{ML}^2$ handelt es sich um die Residuen und den ML-Schätzer für σ^2 im zunächst zugrunde gelegten linearen Modell $y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$ mit homoskedastischen Fehlern. Als Teststatistik fungiert dann die Größe

$$T = \frac{1}{2} \sum_{i=1}^{n} (\hat{g}_i - \bar{g})^2.$$

Falls keine Variable einen Einfluss auf die Varianz hat, d.h. falls keine Heteroskedastizität vorliegt, wird $\hat{g}_i \approx \bar{g}$ gelten. Je größer also T, d.h. je größer die Summe der quadrierten Abweichungen der geschätzten Werte $\hat{g}_i$ vom Mittel $\bar{g}$, desto mehr spricht für die Gegenhypothese heteroskedastischer Varianzen. Unter H_0 ist die Verteilung von T unabhängig von der Funktion h und asymptotisch χ^2-verteilt mit q Freiheitsgraden, d.h. $T \overset{a}{\sim} \chi_q^2$. Bei einem Signifikanzniveau von α wird die Nullhypothese abgelehnt, falls der beobachtete Wert der Teststatistik größer als das $(1-\alpha)$-Quantil der χ_q^2-Verteilung ist.

Beispiel 3.17 Mietspiegel – Breusch-Pagan-Test

Da davon auszugehen ist, dass die Varianzen (unter Umständen nichtlinear) sowohl von der Wohnfläche als auch vom Baujahr abhängen, unterstellen wir das Varianzmodell

$$\sigma_i^2 = \sigma^2\, h(\alpha_0 + \alpha_1 \mathit{flaecheo}_i + \alpha_2 \mathit{flaecheo2}_i + \alpha_3 \mathit{flaecheo3}_i + \alpha_4 \mathit{bjahro}_i + \alpha_5 \mathit{bjahro2}_i + \alpha_6 \mathit{bjahro3}_i),$$

wobei *flaecheo*, *flaecheo2* und *flaecheo3* orthogonale Polynome dritten Grades der Wohnfläche sind. Basierend auf diesem Modell erhalten wir $T = 997.16$ als Teststatistik für den Breusch-Pagan-Test. Der dazu gehörige p-Wert ist praktisch Null, so dass auch der Breusch-Pagan-Test (neben den studentisierten Residuen) einen weiteren klaren Hinweis auf heteroskedastische Varianzen liefert.

△

Abschließend seien noch einige kritische Anmerkungen zu den Tests auf Heteroskedastizität angeführt. Dem einen oder anderen Leser mag die Durchführung eines formalen Tests exakter erscheinen als die Inspektion von Residuenplots. Heteroskedastizität wird aber in den meisten Fällen in explorativer Art und Weise diagnostiziert, in den wenigsten Fällen existieren substanzwissenschaftliche Theorien über Art und Ausmaß der Heteroskedastizität. Damit besteht bei der Modellierung der Fehlervarianzen im linearen Modell noch mehr Unsicherheit als bei der Modellierung des Erwartungswerts. Die Gültigkeit von Tests hängt aber entscheidend davon ab, dass das zugrunde gelegte Modell annähernd korrekt ist. Im Breusch-Pagan-Test wird beispielsweise eine multiplikative Varianz mit genau definierten Kovariablen unterstellt. Die Anzahl der Kovariablen im Varianzausdruck bestimmt unmittelbar die Verteilung der Teststatistik. Tests auf Heteroskedastizität sollten also ähnlich wie Residualplots eher als explorative (heuristische) Hilfsmittel angesehen werden. Auf keinen Fall sollten diese Tests als einzige Diagnosewerkzeuge zur Aufdeckung von Heteroskedastizität herangezogen werden. Diese Vorgehensweise legen einige Lehrbücher der Ökonometrie nahe, indem zwar eine Vielzahl von Tests auf Heteroskedastizität beschrieben werden, die Möglichkeit von Residualplots jedoch nicht einmal Erwähnung findet. Residualplots sollten jedoch in jedem Fall erstellt werden, oft geben erst diese Aufschluss über die Art der Heteroskedastizität, etwa darauf welche Variablen einen Einfluss auf die Fehlervarianz besitzen.

Maßnahmen bei Heteroskedastizität

Nachdem heteroskedastische Fehler diagnostiziert wurden, müssen geeignete Maßnahmen ergriffen werden, um Fehlschlüsse durch Nichtbeachtung der Heteroskedastizität zu vermeiden. Im Folgenden beschreiben wir die gängigsten Varianten.

Variablentransformation

Im Rahmen der Diskussion der Modellannahmen haben wir uns auch mit Modellen mit multiplikativen Fehlern beschäftigt, vergleiche Seite 60. Ein gängiges Modell mit multiplikativen Fehlern ist das exponentielle Modell

$$\begin{aligned} y_i &= \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i) \\ &= \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik})\exp(\varepsilon_i). \end{aligned}$$

Bei normalverteilten Störgrößen $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ sind die Größen $\exp(\varepsilon_i)$ und y_i logarithmisch normalverteilt. Unter Beachtung der Varianz logarithmisch normalverteilter Zufallsvariablen (vergleiche Seite 459 in Anhang B) erhalten wir

$$\mathrm{Var}(\exp(\varepsilon_i)) = \exp(\sigma^2) \cdot (\exp(\sigma^2) - 1)$$

und damit

$$\mathrm{Var}(y_i) = (\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^2 \exp(\sigma^2) \cdot (\exp(\sigma^2) - 1).$$

Wir stellen also fest, dass in Modellen mit multiplikativen Fehlern die Varianzen der y_i heteroskedastisch sind, obwohl die Varianzen der Störgrößen selbst homoskedastisch sind.

Wie bereits erwähnt, kann das exponentielle Modell durch Logarithmieren in ein gewöhnliches lineares Modell $\log(y_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$ mit homoskedastischen Varianzen überführt

werden, vergleiche auch Seite 70ff. Diese Erkenntnis liefert uns eine einfache Maßnahme bei diagnostizierten heteroskedastischen Varianzen. Falls ein multiplikatives Modell vorliegt, sollte die Zielgröße einfach logarithmiert werden und ein Regressionsmodell mit der transformierten Größe geschätzt werden.

Zweistufige Schätzung

Bei heteroskedastischen Varianzen sind im Allgemeinen neben den Regressionskoeffizienten $\boldsymbol{\beta}$ und dem Varianzparameter σ^2 auch die Gewichte w_i unbekannt. Eine nahe liegende Maßnahme bei heteroskedastischen Varianzen ist daher die *gemeinsame* Schätzung sämtlicher unbekannter Parameter. Im Folgenden entwickeln wir ein einfaches zweistufiges Schätzverfahren:
Wegen $\mathrm{E}(\varepsilon_i) = 0$ gilt $\mathrm{E}(\varepsilon_i^2) = \mathrm{Var}(\varepsilon_i) = \sigma_i^2$ und wir können ε_i^2 darstellen als

$$\varepsilon_i^2 = \sigma_i^2 + v_i,$$

wobei v_i die Abweichung der quadrierten Störung von ihrem Erwartungswert ist. In den meisten Fällen hängt σ_i^2 von einer oder mehreren Kovariablen ab. Es ist also nahe liegend,

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_{i1} + \ldots + \alpha_q z_{iq} = \boldsymbol{z}_i' \boldsymbol{\alpha}$$

anzunehmen. Der Vektor $\boldsymbol{z}$ beinhaltet alle Kovariablen, welche die Varianzen beeinflussen. In vielen Fällen wird der Vektor $\boldsymbol{z}$ mit dem Kovariablenvektor $\boldsymbol{x}$ identisch sein. Es können aber auch andere, nicht in $\boldsymbol{x}$ enthaltene Kovariablen aufgenommen werden. Zur Schätzung des unbekannten Parametervektors $\boldsymbol{\alpha}$ könnte man prinzipiell mit Hilfe der KQ-Methode ein lineares Modell mit den quadrierten Störgrößen ε_i^2 als abhängige Variable und $\boldsymbol{z}_i$ als unabhängige Variablen schätzen. Da die Störgrößen unbeobachtet sind, müssen diese durch die Residuen $\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$ aus einer zunächst ungewichteten Regression zwischen y und $\boldsymbol{x}$ ersetzt werden. Damit erhalten wir folgendes zweistufiges Verfahren:

1. Erhalte eine vorläufige Schätzung $\hat{\boldsymbol{\beta}}$ aus einer ungewichteten Regression zwischen y und $\boldsymbol{x}$. Bezeichne die Residuen mit $\hat{\varepsilon}_i$.
2. Erhalte Schätzungen $\hat{\boldsymbol{\alpha}}$ aus einer ungewichteten Regression zwischen den quadrierten Residuen $\hat{\varepsilon}_i^2$ und den die Varianz erklärenden Variablen $\boldsymbol{z}_i$. Schätze ein allgemeines lineares Modell mit geschätzten Gewichten

$$\hat{w}_i = \frac{1}{\boldsymbol{z}_i' \hat{\boldsymbol{\alpha}}}.$$

Dieses Verfahren ist jedoch in der Praxis nicht immer anwendbar, da nicht sichergestellt ist, dass $\boldsymbol{z}_i' \hat{\boldsymbol{\alpha}}$ stets größer als Null ist. Falls $\boldsymbol{z}_i' \hat{\boldsymbol{\alpha}} < 0$, würden aber negative Gewichte $\hat{w}_i$ und in der Folge negative erwartete Varianzen $\hat{\sigma}_i^2$ resultieren. Alternativ kann das Modell

$$\sigma_i^2 = \exp(\boldsymbol{z}_i' \boldsymbol{\alpha})$$

angenommen werden. Durch die Exponentialfunktion wird sichergestellt, dass die geschätzten erwarteten Varianzen positiv sind. Bei diesem Modell kann $\boldsymbol{\alpha}$ mit Hilfe der Regression

$$\log(\hat{\varepsilon}_i^2) = \boldsymbol{z}_i' \boldsymbol{\alpha} + v_i$$

geschätzt werden. Als Gewichte für die Regression zwischen y und $\boldsymbol{x}$ erhält man dann

$$\hat{w}_i = \frac{1}{\exp(\boldsymbol{z}_i'\hat{\boldsymbol{\alpha}})}.$$

Beispiel 3.18 Mietspiegel – zweistufige Schätzung bei Heteroskedastizität

Wir illustrieren das zweistufige Schätzverfahren mit den Daten zum Mietspiegel. In den Beispielen 3.16 und 3.17 wurden für das Modell (3.31) heteroskedastische Varianzen diagnostiziert. Die zweistufige Schätzung für dieses Modell erhalten wir in den folgenden drei Schritten:

1. Schritt: Klassisches lineares Modell

Im ersten Schritt schätzen wir ein klassisches lineares Modell, vergleiche die Ergebnisse aus Beispiel 3.16. Die Abbildungen 3.18 a) und b) zeigen jeweils Streudiagramme zwischen $\log(\hat{\varepsilon}_i^2)$ und Wohnfläche bzw. Baujahr. Beide Abbildungen geben Hinweise auf Varianzen σ_i^2, die in Abhängigkeit von Wohnfläche und Baujahr variieren.

2. Schritt: Hilfsregression

Zur Bestimmung der geschätzten Gewichte $\hat{w}_i$ für die gewichtete Regression schätzen wir mit Hilfe der Methode der kleinsten Quadrate das Regressionsmodell

$$\log(\hat{\varepsilon}_i^2) = \alpha_0 + \alpha_1\, \mathit{flaecheo}_i + \alpha_2\, \mathit{flaecheo2}_i + \alpha_3\, \mathit{flaecheo3}_i +$$
$$\alpha_4\, \mathit{bjahro}_i + \alpha_5\, \mathit{bjahro2}_i + \alpha_6\, \mathit{bjahro3}_i + v_i.$$

Die Abbildungen 3.18 c) und d) zeigen die geschätzten Effekte der Wohnfläche und des Baujahrs. Offenbar steigt die Variabilität der Nettomieten mit größerer Wohnfläche, wobei sich der Anstieg bei sehr großen Wohnungen etwas verlangsamt. Der Effekt des Baujahrs ist S-förmig, jedoch deutlich schwächer als der Effekt der Wohnfläche. Dieser Effekt ist gut interpretierbar: Die Nettomieten älterer Wohnungen schwanken stärker als bei moderneren Wohnungen.

3. Schritt: Gewichtetes lineares Modell

In einem letzten Schritt schätzen wir das Regressionsmodell (3.31) erneut mit den Gewichten

$$\hat{w}_i = \frac{1}{\exp(\hat{\eta}_i)}$$

mit $\hat{\eta}_i = \hat{\alpha}_0 + \hat{\alpha}_1\, \mathit{flaecheo}_i + \ldots + \hat{\alpha}_6\, \mathit{bjahro3}_i$. Wir erhalten die in Tabelle 3.4 abgedruckten Ergebnisse. Die geschätzten Standardfehler basieren auf der geschätzten Kovarianzmatrix

$$\widehat{\mathrm{Cov}\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\boldsymbol{X}'\mathrm{diag}(\hat{w}_1, \ldots, \hat{w}_n)\boldsymbol{X})^{-1}.$$

Diese sind insgesamt kleiner als bei der ungewichteten Regression. Unter Verwendung der korrigierten Standardfehler können die für das klassische lineare Modell entwickelten Tests und Konfidenzintervalle verwendet werden. Im vorliegenden Fall sind die Konfidenzintervalle etwas enger als bei Verwendung des ungewichteten KQ-Schätzers.

△

Prinzipiell ist eine *simultane Schätzung* von $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ und σ^2 im Vergleich zu dem beschriebenen zweistufigen Verfahren eleganter. Nahe liegend ist hier vor allem die vollständige ML-Schätzung oder die Kombination aus ML-Schätzung für $\boldsymbol{\beta}$ und REML-Schätzung für die Varianzparameter $\boldsymbol{\alpha}$ und σ^2. Wir verzichten an dieser Stelle auf die detaillierte Darstellung dieser Schätzansätze, da diese in den meisten Statistikprogrammen nicht standardmäßig implementiert sind. Ein Maximum-Likelihood-Ansatz für ein Modell mit multiplikativen Varianzen $\sigma_i^2 = \sigma^2 \exp(\boldsymbol{z}_i'\boldsymbol{\alpha})$ ist ausführlich in Greene (2000) Kapitel 12 beschrieben. Eine Implementation existiert beispielsweise für STATA (Funktion `regh`).

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	894.739	5.087	175.900	<0.001	884.766	904.712
flaecheo	229.585	4.645	49.430	<0.001	220.478	238.692
bjahro	94.959	5.180	18.330	<0.001	84.802	105.117
bjahro2	49.301	4.551	10.830	<0.001	40.377	58.225
bjahro3	-2.582	4.642	-0.560	0.578	-11.684	6.520

Tabelle 3.4. *Schätzergebnisse für die gewichtete Regression*

Abb. 3.18. *Mietspiegel für München mit heteroskedastischen Fehlern. Die Abbildungen a) und b) zeigen Streudiagramme zwischen* $\log(\hat{\varepsilon}_i^2)$ *und Wohnfläche bzw. Baujahr. In den Abbildungen c) und d) findet man Wohnflächen- und Baujahrseffekt inklusive partieller Residuen für die Varianzregression.*

White-Schätzer

Eine Voraussetzung für die Anwendbarkeit der bisher betrachteten Verfahren ist die Kenntnis der *Art* der Heteroskedastizität. Ein anderer Zugang wurde von White (1980) vorgeschlagen. Der Artikel schlägt ein Verfahren vor, wie bei Verwendung des *gewöhnlichen KQ-Schätzers* die Kovarianzmatrix (3.29) auf Seite 124 konsistent geschätzt werden kann, so dass zumindest asymptotisch korrekte Standardfehler, Konfidenzintervalle und Tests resultieren. White zeigt unter allgemeinen Bedingungen, dass

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	895.903	5.131	174.600	<0.001	885.842	905.963
flaecheo	237.544	7.178	33.090	<0.001	223.470	251.618
bjahro	105.956	6.043	17.530	<0.001	94.107	117.804
bjahro2	61.394	5.480	11.200	<0.001	50.649	72.139
bjahro3	-0.387	5.694	-0.070	0.946	-11.551	10.777

Tabelle 3.5. *Schätzergebnisse für die ungewichtete Regression mit korrigierten Standardfehlern.*

$$\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}})} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{diag}(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{3.32}$$

eine konsistente Schätzung für die Kovarianzmatrix (3.29) von $\hat{\boldsymbol{\beta}}$ ist. Die Varianzen $\sigma_i^2 = \sigma^2 w_i$ in (3.29) werden also durch die quadrierten Residuen ersetzt bzw. geschätzt. Aufgrund ihrer Form wird die Matrix (3.32) auch als „Sandwich"-Matrix bezeichnet. Ähnliche Formen für Kovarianzmatrizen werden wir im Rahmen gemischter Modell in Abschnitt 6.3.3 kennenlernen.

Die Schätzungen für die Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ können in die Teststatistik des F-Tests zum Test allgemeiner linearer Hypothesen eingesetzt werden. Auf diese Art und Weise erhält man zumindest asymptotisch korrekte Tests und über die Dualitätseigenschaft von Tests und Konfidenzintervallen auch asymptotische Konfidenzintervalle.

Das beschriebene Verfahren hat den Vorteil, dass die Matrix $\boldsymbol{W}$ im allgemeinen linearen Modell *nicht* bekannt sein muss. Dadurch erkauft man sich allerdings auch ungenauere Aussagen im Sinne erhöhter Varianzen der Schätzer.

Beispiel 3.19 Mietspiegel für München

Wir schätzen das Modell (3.31) auf Seite 129 erneut, verwenden aber (3.32) zur Konstruktion von Tests und Konfidenzintervallen. Wir erhalten die in Tabelle 3.5 abgedruckten Ergebnisse. Tendenziell sind die geschätzten Standardfehler deutlich größer im Vergleich zur Schätzung mit nicht adjustierter Kovarianzmatrix und auch im Vergleich zum zweistufigen Schätzverfahren. Dies ist jedoch nicht überraschend, schließlich haben wir insgesamt weniger Information, da über die Art der Heteroskedastizität keine Annahmen getroffen werden.

Insgesamt stellen wir fest: Die verschiedenen Varianten zur Berücksichtigung heteroskedastischer Fehler liefern nur geringe qualitative Unterschiede in der Interpretation im Vergleich zum klassischen linearen Modell. In allen geschätzten Modellen scheint der Einfluss der Wohnfläche und des Baujahrs gesichert. Außerdem bestehen Hinweise darauf, dass eine quadratische Modellierung des Baujahrseffekts ausreicht. Diese relative Robustheit der Ergebnisse wird in vielen Anwendungsfällen beobachtet. Trotz heteroskedastischer Varianzen ergeben sich durch deren Berücksichtigung häufig keine nennenswerten Änderungen in der Interpretation.

△

3.4.4 Autokorrelierte Fehler

Neben heteroskedastischen Fehlern sind in der Praxis korrelierte Fehler einer der Hauptgründe, warum die Annahmen des klassischen linearen Modells verletzt sind und Methoden für das allgemeine lineare Modell zur Schätzung herangezogen werden müssen.

In Abschnitt 3.1.3 (Seite 66) wurde angemerkt, dass autokorrelierte Störungen dann auftreten, wenn das Modell fehlspezifiert ist. Dabei wurden folgende Gründe für die Fehlspezifikation genannt:

- Der Einfluss einer erklärenden Variable wurde nicht korrekt modelliert. Bei autokorrelierten Fehlern sollte man also zunächst diese Fehlerquelle in Betracht ziehen. Dabei können zum Beispiel die in Abschnitt 3.1.4 diskutierten Modellierungsvarianten für den nichtlinearen Einfluss metrischer Kovariablen hilfreich sein.
- Relevante erklärende Variablen sind nicht beobachtbar und können daher im Modell nicht berücksichtigt werden. In Verbindung mit autokorrelierten Störungen tritt dieses Problem vor allem bei Zeitreihen sowie bei Panel- oder Longitudinaldaten auf, wenn die nicht berücksichtigten erklärenden Variablen einen zeitlichen Trend aufweisen. In diesem Fall können die Schätzmethoden dieses Abschnitts herangezogen werden, um zumindest zu genaueren Aussagen und verbesserten Prognosen zu gelangen.

Autokorrelation erster Ordnung

Wir beschränken uns auf den einfachsten und in der Praxis am häufigsten verwendeten Fehlerprozess, und zwar auf autokorrelierte Fehler erster Ordnung. Speziell nehmen wir an, dass die Störgrößen einem autoregressiven Prozess erster Ordnung ($AR(1)$) folgen, d.h.

$$\varepsilon_i = \rho\,\varepsilon_{i-1} + u_i$$

mit $-1 < \rho < 1$. Bezüglich der u_i treffen wir folgende Annahmen:

1. $\mathrm{E}(u_i) = 0$.
2. $\mathrm{Var}(u_i) = \mathrm{E}(u_i^2) = \sigma_u^2, \quad i = 1, \ldots, n$.
3. $\mathrm{Cov}(u_i, u_j) = \mathrm{E}(u_i u_j) = 0, \quad i \neq j$.

Zusätzlich nehmen wir an, dass der Prozess bereits weit in der Vergangenheit begonnen hat. Aus diesen Annahmen lässt sich ein spezielles lineares Modell ableiten mit $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ und der Kovarianzmatrix

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{W} = \frac{\sigma_u^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}. \tag{3.33}$$

Die genaue Herleitung findet man in Abschnitt 3.7.2 auf Seite 187.

Unter Zuhilfenahme der Kovarianzmatrix können wir leicht die Korrelationskoeffizienten zwischen ε_i und den um j Perioden verzögerten Störungen ε_{i-j} berechnen. Wir erhalten die sogenannte *Autokorrelationsfunktion*

$$\mathrm{ACF}(j) = \frac{\mathrm{Cov}(\varepsilon_i, \varepsilon_{i-j})}{\mathrm{Var}(\varepsilon_j)} = \rho^j \quad j = 0, 1, 2, \ldots$$

Eine charakteristische Eigenschaft autokorrelierter Störungen erster Ordnung ist also ein langsames Abklingen der Stärke der Korrelation zwischen ε_i und den verzögerten Störungen. Falls ρ positiv ist, nimmt die Korrelation geometrisch ab. Bei negativem ρ

Abb. 3.19. *Autokorrelationsfunktion und partielle Autokorrelationsfunktion bei autokorrelierten Störgrößen erster Ordnung. Abgebildet sind die Funktionen für* $\rho = 0.7, 0.9, -0.9$.

nimmt die Korrelation mit alternierendem Vorzeichen ab. Vergleiche hierzu auch die linke Spalte von Abbildung 3.19, in der die Autokorrelationsfunktionen für verschiedene Werte von ρ abgebildet sind. Grafische Darstellungen dieser Art werden auch als *Korrelogramme* bezeichnet.

Neben der Autokorrelationsfunktion ist die sogenannte partielle Autokorrelationsfunktion eine weitere charakteristische Kenngröße korrelierter Störungen bzw. allgemeiner von stochastischen Prozessen. Die partielle Autokorrelation $\mathrm{PACF}(j)$ zwischen ε_i und ε_{i-j} ist definiert als der Regressionskoeffizient α_j in dem Modell

$$\varepsilon_i = \alpha_1 \varepsilon_{i-1} + \ldots + \alpha_j \varepsilon_{i-j} + v_i. \tag{3.34}$$

In jedem Fall erhalten wir PACF(1) = ACF(1). Für autokorrelierte Störungen erster Ordnung gilt offensichtlich PACF(1) = ρ und PACF(j) = 0 für $j > 1$. Die rechte Spalte von Abbildung 3.19 zeigt die partiellen Autokorrelationsfunktionen für einige AR(1) Prozesse. Charakteristisch ist der abrupte Abfall der partiellen Autokorrelation für $j > 1$. Der Koeffizient α_j in (3.34) kann als Korrelationskoeffizient zwischen $\varepsilon_i - \alpha_1 \varepsilon_{i-1} - \ldots - \varepsilon_{i-j+1}\alpha_{j-1}$ und ε_{i-j} interpretiert werden. Die Bezeichnung als *partieller* Korrelationskoeffizient kommt also daher, dass nicht die Korrelation zwischen ε_i und ε_{i-j} berechnet wird, sondern die Korrelation zwischen ε_i und ε_{i-j} unter Ausschaltung des Einflusses der dazwischen liegenden Störungen.

Diagnose autokorrelierter Störungen

Wie bei der Diagnose heteroskedastischer Fehler basieren die Diagnoseinstrumente auf den Residuen eines zunächst unterstellten klassischen linearen Modells. Folgende Instrumente haben sich in der Praxis bewährt:

Grafische Darstellung der Residuen im Zeitablauf

Die naheliegendste Möglichkeit, Korrelationen in den Residuen aufzuspüren, bieten Residualplots, im vorliegenden Fall Streudiagramme der Residuen oder der studentisierten Residuen im Zeitablauf. Folgen auf positive Residuen tendenziell wieder positive Residuen, so können wir von positiver Autokorrelation ausgehen. Folgen auf positive Residuen tendenziell eher negative Residuen, so ist dies ein Hinweis auf negative Autokorrelation.

Empirische Autokorrelationsfunktionen

Hinweise auf die Art der Korrelation geben auch die *empirischen* Autokorrelationen und partiellen Autokorrelationen und deren Visualisierung in Korrelogrammen. Die empirischen Autokorrelationen sind Schätzungen für ACF(j) und können durch

$$\widehat{\mathrm{ACF}}(j) = \frac{\widehat{\mathrm{Cov}}(\varepsilon_i, \varepsilon_{i-j})}{\widehat{\mathrm{Var}}(\varepsilon_i)} \quad \text{mit} \quad \widehat{\mathrm{Cov}}(\varepsilon_i, \varepsilon_{i-j}) = \frac{1}{n} \sum_{i=j+1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-j}$$

berechnet werden. Im Wesentlichen handelt es sich bei den empirischen Autokorrelationen um die empirischen Korrelationskoeffizienten zwischen den Residuen und den um j Perioden verzögerten Residuen.

Die partiellen Autokorrelationen können durch mehrmalige Schätzung des Regressionsmodells (3.34) für $j = 1, 2, 3, \ldots$ und damit durch $\widehat{\mathrm{PACF}}(j) = \hat{\alpha}_j$ bestimmt werden. Dabei werden in (3.34) die Störgrößen ε_i durch die Residuen $\hat{\varepsilon}_i$ ersetzt. Durch Inspektion des Korrelogramms und des partiellen Korrelogramms erhält man Hinweise auf das Vorliegen von Autokorrelation. Sind sämtliche empirischen Autokorrelationen und partiellen Autokorrelationen annähernd Null, so können wir eher von unkorrelierten Störungen ausgehen, eine erneute Schätzung des Modells aufgrund von Autokorrelation ist dann nicht nötig. Sind die empirischen den theoretischen Autokorrelationen und partiellen

Abb. 3.20. *Residuen im Zeitablauf, empirische Autokorrelationsfunktion und partielle Autokorrelationsfunktion der Residuen für simulierte Daten mit positiver Autokorrelation erster Ordnung.*

Autokorrelationen von AR(1)-Störungen ähnlich, so ist das ein Hinweis auf das Vorliegen autokorrelierter Störungen erster Ordnung. Weisen die Korrelogramme nicht die für AR(1)-Störungen typische Gestalt auf, so ist dies ein Hinweis auf kompliziertere Korrelationsstrukturen. Diese werden im Rahmen der statistischen Zeitreihenanalyse untersucht und sprengen den Rahmen dieses Buches.

Beispiel 3.20 Simulierte Daten – Grafische Diagnose autokorrelierter Störgrößen

Wir betrachten zur Illustration das in Abbildung 3.3 (Seite 67) simulierte Regressionsmodell mit positiver Autokorrelation. Im Zuge der Diskussion der Modellannahmen des klassischen linearen Modells wurde das Modell $y_i = -1 + 2x_i + \varepsilon_i$ mit $\varepsilon_i = 0.9\varepsilon_{i-1} + u_i$ simuliert. Zur Illustration der Diagnose autokorrelierter Fehler anhand von Residualplots und Korrelogrammen schätzen wir für die Daten ein klassisches lineares Modell. Wir erhalten die Schätzung $\hat{y}_i = -1.08 + 2.13\,x_i$. Abbildung 3.20 zeigt die Residuen im Zeitablauf und die empirische bzw. partielle Autokorrelationsfunktion der Residuen. Man beachte, dass sich der Residualplot in Abbildung a) (wenn auch nur geringfügig) von den Störgrößen im Zeitablauf in Abbildung 3.3 b) auf Seite 67 unterscheidet, da die Residuen lediglich eine Schätzung der (in der Praxis unbekannten) Störgrößen sind. Die Grafiken zeigen nahezu idealtypische Hinweise auf Autokorrelation erster Ordnung. Die Residuen sind hochkorreliert, die empirischen Autokorrelationen fallen geometrisch ab und die partiellen Autokorrelationen sind für $j > 1$ annähernd Null.

△

Abb. 3.21. *Annahme- und Ablehnbereiche beim Durbin-Watson-Test.*

Tests auf Autokorrelation – Durbin-Watson-Test

Neben grafischen Hilfsmitteln können auch statistische Tests zur Aufdeckung von Autokorrelation verwendet werden. In diesem Zusammenhang am meisten Verwendung findet ein von Durbin und Watson entwickelter Test auf serielle Korrelation. Der Durbin-Watson-Test überprüft die Hypothesen

$$H_0 \ : \ \rho = 0 \quad \text{versus} \quad H_1 \ : \ \rho \neq 0$$

und basiert auf der Teststatistik

$$d = \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}.$$

Dieser Teststatistik liegt folgende Überlegung zugrunde: Insbesondere für großen Stichprobenumfang n gilt

$$d = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i^2 + \sum_{i=2}^{n} \hat{\varepsilon}_{i-1}^2 - 2\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2} \approx 1 + 1 - 2\hat{\rho} = 2(1 - \hat{\rho})$$

und wegen $-1 < \hat{\rho} < 1$ folglich $0 < d < 4$. Liegt d in der Nähe von 2, so ist $\hat{\rho} \approx 0$, und man wird die Nullhypothese beibehalten. Je näher die Teststatistik d bei 0 bzw. 4 liegt, desto näher liegt $\hat{\rho}$ bei 1 bzw -1, und die Nullhypothese wird tendenziell eher abgelehnt. Die Bestimmung der Verteilung der Teststatistik unter H_0 ist relativ schwierig, da diese von der Designmatrix abhängt. Unter Umständen kann daher eine Testentscheidung schwierig sein. Durbin und Watson haben eine Teillösung des Problems gefunden. Für bestimmte Intervalle des Wertebereichs von d kann eine Testentscheidung getroffen werden. Diese Intervalle hängen von den beiden unteren und oberen Schranken d_u und d_o

ab, die für verschiedene Stichprobenumfänge n und Anzahlen p an Regressoren vertafelt sind. Abbildung 3.21 liefert eine grafische Darstellung der Annahme- und Ablehnbereiche beim Durbin-Watson-Test.

Darüber hinaus können p-Werte für den Durbin Watson Test dank moderner Computer numerisch bestimmt werden. In einigen Programmpaketen haben diese Verfahren jedoch noch keinen Einzug gefunden. Eine Implementation liefert beispielsweise die Funktion `dwtest` des mathematischen Programmpakets Matlab. Eine gleichlautende Funktion im Paket `lmtest` des Statistikprogramms R stellte sich allerdings als fehlerhaft heraus.

Beispiel 3.21 Simulierte Daten – Durbin-Watson Test

Für die im letzten Beispiel 3.20 illustrativ verwendeten simulierten Daten erhalten wir als Teststatistik $d = 0.2619$ und mit der Matlab-Funktion 'dwtest' einen sehr kleinen p-Wert. Damit liefert auch der Durbin-Watson Test zusätzlich zu den vorher betrachteten grafischen Hilfsmitteln einen klaren Hinweis auf Autokorrelation erster Ordnung.

△

Maßnahmen bei Autokorrelation erster Ordnung

Wie bereits bei der Diskussion der Annahmen des klassischen linearen Modells erwähnt, deuten autokorrelierte Störgrößen auf eine (wie auch immer geartete) Fehlspezifikation des Modells hin. Bevor also die weiter unten beschriebenen Schätzverfahren für Modelle mit korrelierten Fehlern herangezogen werden, sollte zunächst untersucht werden, ob die Fehlspezifikation beseitigt werden kann. Mögliche Verbesserungen können etwa durch zusätzliche Aufnahme bisher nicht berücksichtigter Kovariablen oder durch nichtlineare Modellierung metrischer Kovariablen erfolgen.

Kann die Korrelation nicht beseitigt werden, können Schätzverfahren für Modelle mit autokorrelierten Fehlern verwendet werden. Im Modell mit autokorrelierten Störgrößen erster Ordnung benötigen wir Schätzungen für die Regressionskoeffizienten $\boldsymbol{\beta}$, den Varianzparameter σ^2 und den Korrelationskoeffizienten ρ. Zur Schätzung der Parameter existiert in der Literatur eine Vielzahl von Verfahren. Wir entwickeln zunächst ein zweistufiges Verfahren. Im ersten Schritt wird der Korrelationskoeffizient basierend auf einer gewöhnlichen KQ-Schätzung geschätzt. Anschließend werden die Regressionskoeffizienten mit Hilfe der gewichteten Methode der kleinsten Quadrate geschätzt. Zuletzt gehen wir auch kurz auf Maximum-Likelihood Schätzungen ein. Einen guten Überblick über weitere gängige Verfahren liefert Judge et al. (1980).

Zweistufige Schätzung

Eine Schätzung des Modells mit autokorrelierten Fehlern erhält man durch folgendes Vorgehen:

1. Erhalte eine vorläufige Schätzung $\hat{\boldsymbol{\beta}}$ aus einer ungewichteten Regression zwischen y und $\boldsymbol{x}$. Bezeichne die Residuen mit $\hat{\varepsilon}_i$.
2. Erhalte eine Schätzung $\hat{\rho}$ des Korrelationskoeffizienten durch

$$\hat{\rho} = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sqrt{\sum_{i=2}^{n} \hat{\varepsilon}_i^2} \sqrt{\sum_{i=2}^{n} \hat{\varepsilon}_{i-1}^2}}. \tag{3.35}$$

Bei $\hat{\rho}$ handelt es sich um den empirischen Korrelationskoeffizienten zwischen $\hat{\varepsilon}_i$ und $\hat{\varepsilon}_{i-1}$.

3. Setze $\hat{\rho}$ in die Gewichtsmatrix $\boldsymbol{W}$ ein und erhalte eine Schätzung $\hat{\boldsymbol{W}}$. Verwende $\hat{\boldsymbol{W}}$ zur erneuten Schätzung der Regressionskoeffizienten mit Hilfe der gewichteten Methode der kleinsten Quadrate.

Zur Verbesserung der Schätzungen sollten die Schritte 2 und 3 iteriert werden. Dieses Verfahren ist als Prais-Winsten-Schätzung bekannt. Unter recht allgemeinen Bedingungen lässt sich zeigen, dass die resultierende Schätzung für $\boldsymbol{\beta}$ konsistent ist. Neben dem ursprünglichen von Prais und Winsten vorgeschlagenen Verfahren existieren noch eine Fülle von Modifikationen, vergleiche dazu Greene (2000).

Maximum-Likelihood-Schätzung

Für die Maximum-Likelihood-Schätzung benötigen wir wieder die Zusatzannahme normalverteilter Fehler. Die Likelihood ist dann gegeben durch

$$L(\boldsymbol{\beta}, \sigma^2, \rho) = \frac{1}{(2\pi\sigma^2)^{n/2} \, |\boldsymbol{W}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{W}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right). \tag{3.36}$$

Zunächst berechnen wir

$$\boldsymbol{W}^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix}.$$

und

$$|\boldsymbol{W}| = \frac{1}{1-\rho^2}.$$

Damit erhalten wir die Log-Likelihood

$$l(\boldsymbol{\beta}, \sigma^2, \rho) = -\frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log(1-\rho^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{W}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Das Maximum der Log-Likelihood lässt sich nicht in geschlossener Form angeben, vielmehr muss der ML-Schätzer iterativ bestimmt werden. Details findet man in Greene (2000) oder in der dort zitierten Originalliteratur.

Beispiel 3.22 Simulierte Daten – ML-Schätzer

Für unsere simulierten Daten aus den vorangegangenen Beispielen enthält Tabelle 3.6 den Prais-Winsten Schätzer und Tabelle 3.7 den Maximum-Likelihood-Schätzer.

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	-1.084	0.091	-11.900	<0.001	-1.265	-0.903
x	2.132	0.051	41.810	<0.001	2.031	2.233

Tabelle 3.6. *Prais-Winsten Schätzer für die simulierten Daten.*

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	-0.982	0.307	-3.200	0.001	-1.583	-0.380
x	2.094	0.139	15.000	<0.001	1.821	2.368

Tabelle 3.7. *ML-Schätzer für die simulierten Daten.*

△

Beispiel 3.23 Supermarkt Scanner-Daten

In Beispiel 3.2 (Seite 72) haben wir als Illustration für multiplikative Störgrößen Scanner-Daten, die beim Bezahlen an der Supermarktkasse gewonnen werden, untersucht. Ziel ist die Modellierung des Zusammenhangs zwischen dem wöchentlichen Absatz eines Produktes (hier einer Kaffeemarke) und dem eigenen Preis bzw. dem Preis von Konkurrenzprodukten, vergleiche hierzu die Streudiagramme zwischen Absatz und Preis bzw. Konkurrenzpreis in Abbildung 3.6 (Seite 71). In diesem Beispiel analysieren wir die Daten zur Illustration des Vorgehens bei korrelierten Störungen. Der Einfachheit halber benutzen wir lediglich die Daten für ein Geschäft (in Abbildung 3.6 auf Seite 71 sind die Daten für fünf Geschäfte enthalten). Wegen der multiplikativen Störgrößen verwenden wir den logarithmierten Absatz anstelle des untransformierten Absatzes als Zielgröße. Abbildung 3.22 a) zeigt das Streudiagramm zwischen logarithmiertem Absatz und dem eigenen Preis. Zusätzlich eingezeichnet ist der mit Hilfe der gewöhnlichen KQ-Methode geschätzte logarithmierte Absatz wenn der Einfluss des Preises linear modelliert wird. Die dazu gehörenden Residuen sind in Abbildung b) im Zeitablauf dargestellt. Offensichtlich sind die Residuen deutlich korreliert. Die Teststatistik des Durbin-Watson-Tests ist $d = 1.1751$ (p-Wert $< 10^{-4}$). Der berechnete p-Wert basiert dabei auf der Matlab Funktion `dwtest`. Die Nullhypothese $H_0 : \rho = 0$ wird also sowohl zum Niveau $\alpha = 0.01$ als auch zum Niveau $\alpha = 0.05$ verworfen. Zur weiteren Absicherung untersuchen wir in Abbildung 3.23 noch die empirischen Autokorrelationsfunktionen bzw. partiellen Autokorrelationsfunktionen der Residuen. Die Grafiken enthalten zusätzlich punktweise 95% Konfidenzintervalle, auf deren Herleitung wir an dieser Stelle jedoch nicht eingehen. Die Korrelogramme geben deutliche Hinweise auf Autokorrelation. Die Autokorrelationsfunktionen nehmen mit zunehmender Verzögerung nur langsam ab. Die partiellen Autokorrelationsfunktionen sind ab einer Verzögerung von 3 relativ klein. Die Korrelogramme deuten damit nicht unbedingt auf Autokorrelation erster Ordnung hin, eventuell sollte eine komplizierte Korrelationsstruktur (Autokorrelation zweiter Ordnung) zugrunde gelegt werden. Hier zeigt sich ein Vorteil grafischer Hilfsmittel im Vergleich zu formalen Tests. Der Durbin-Watson Test gibt zwar Hinweise auf Autokorrelation, die Art der Korrelation kann aber nicht aufgedeckt werden.

Nachdem wir autokorrelierte Störgrößen mit einiger Sicherheit diagnostiziert haben, liegt folgende Strategie nahe: Zunächst sollte untersucht werden, ob durch verbesserte Modellspezifikation die Korrelationen verschwinden. Erst als letzter Ausweg sollte

Abb. 3.22. *Streudiagramm zwischen Log-Absatz und Preis inklusive dem geschätzten Log-Absatz bei linearer Modellierung des Preiseffekts (Grafik a). Die Grafik b) zeigt die Residuen im Zeitablauf. Grafik c) zeigt den Preis der Kaffeemarke im Zeitablauf.*

Abb. 3.23. *Empirische Autokorrelationsfunktion und partielle Autokorrelationsfunktion der Residuen in der Regression zwischen Log-Absatz und Preis.*

ein Modell mit autokorrelierten Fehlern geschätzt werden. Im vorliegenden Fall liefert ein Blick auf die Preise im Zeitablauf (Abbildung 3.22 c) den entscheidenden Hinweis, wie die Modellspezifikation verbessert werden kann. Offenbar lassen sich drei Perioden unterscheiden, in denen der „übliche“ Verkaufspreis der Kaffeemarke (in der Marke-

Abb. 3.24. *Streudiagramm zwischen Log-Absatz und dem Verhältnis zwischen dem aktuellen und regulären Preis inklusive dem geschätzten Log-Absatz (Grafik a). Die Grafik b) zeigt die Residuen im Zeitablauf. Grafik c) zeigt den Preisquotienten im Zeitablauf.*

tingliteratur spricht man von regulären Preisen) jeweils unterschiedlich war. Zunächst war der reguläre Preis bei 8.5 DM, anschließend wurde der reguläre Preis dauerhaft auf 8 DM und schließlich sogar auf 7.5 DM gesenkt. In jeder der Perioden wurden regelmäßige Preisaktionen mit erniedrigten Preisen durchgeführt. Es ist relativ klar, dass eine Nichtbeachtung der zeitlichen Entwicklung des regulären Preises problematisch ist. Ein Vorschlag in der Marketingliteratur, vergleiche z.B. Leeflang, Wittink, Wedel & Naert (2000), ist, den Preis als erklärende Variable im Modell zu ersetzen durch den Quotienten aus dem aktuellen und momentan regulären Preis. Verwende also die neue Variable

$$preisq = \frac{\text{aktueller Preis}}{\text{regulärer Preis}}.$$

Dieser Quotient liegt zwischen Null und Eins. Falls $preisq = 1$, ist der aktuelle Preis gleich dem regulären Preis, es gab also in dieser Woche keine Preissenkung. Abbildung 3.24 a) zeigt das Streudiagramm zwischen Log-Absatz und dem Quotienten aus dem aktuellen und regulären Preis. Zusätzlich eingezeichnet ist wieder der geschätzte Log-Absatz bei linearer Modellierung. Die Residuen im Zeitablauf in Abbildung b) sind bereits deutlich weniger korreliert als bei den Ursprungsregressionen mit dem aktuellen Preis als Einflussgröße (siehe nochmal die Abbildung 3.22 b). Diese Beobachtung wird auch durch den Durbin-Watson Test gestützt. Wir erhalten $d = 1.5946$ mit p-Wert $= 0.0348$. Die Nullhypothese wird also nur noch bei einem Niveau $\alpha = 0.05$ abgelehnt, auf dem Niveau $\alpha = 0.01$ wird die Nullhypothese beibehalten. Eine weitere Verringerung

der Autokorrelation wird erreicht durch Hinzunahme weiterer Kovariablen, etwa der Konkurrenzpreise, oder durch nichtlineare Modellierung.

△

3.5 Bayesianische lineare Modelle

Dieser Abschnitt behandelt das klassische lineare Modell aus Bayesianischer Sicht. Leser, die mit Bayesianischen Inferenzmethoden nicht vertraut sind, sollten zuerst Abschnitt B.5 in Anhang B durchlesen. Der Vorteil der Bayesianischen Sichtweise zur klassischen Inferenz besteht darin, dass a priori *Vorwissen* über die Parameter des Modells geeignet modelliert werden kann.

Ausgangspunkt dieses Abschnitts ist das klassische lineare Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Im Gegensatz zur klassischen Inferenz werden im Bayesianischen Konzept die unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 als Zufallsvariable angesehen. Damit kann die Verteilung der Zielgröße $\boldsymbol{y}$ als bedingte Verteilung in Abhängigkeit von $\boldsymbol{\beta}$ und σ^2 aufgefasst werden und wir erhalten

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

Da damit die (bedingte) Verteilung der beobachtbaren Größe $\boldsymbol{y}$ (im Gegensatz zu den unbeobachtbaren Parametern) spezifiziert wird, spricht man auch vom *Beobachtungsmodell*. Das Bayesianische lineare Modell wird komplettiert durch die Spezifikation einer geeigneten Priori-Verteilung für die unbekannten Parameter. Die am häufigsten verwendete Verteilungsklasse wird im folgenden Abschnitt vorgestellt.

3.5.1 Priori-Verteilungen

Die aus mathematischer Sicht am besten handhabbare Priori-Verteilung ist eine für $\boldsymbol{\beta}$ und σ^2 konjugierte Verteilungsfamilie (vergleiche Abschnitt B.5 in Anhang B). Wir können diese im linearen Modell am häufigsten verwendete Priori-Verteilung wegen

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta} \,|\, \sigma^2) p(\sigma^2)$$

definieren, indem wir die bedingte Verteilung von $\boldsymbol{\beta}$ gegeben σ^2 und die marginale Verteilung von σ^2 spezifizieren. Wir nehmen an, dass $\boldsymbol{\beta} \,|\, \sigma^2$ normalverteilt ist mit Priori-Erwartungswert $\boldsymbol{m}$ und Kovarianzmatrix $\sigma^2 \boldsymbol{M}$, d.h.

$$\boldsymbol{\beta} \,|\, \sigma^2 \sim \mathrm{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{M}).$$

Sowohl $\boldsymbol{m}$ als auch $\boldsymbol{M}$ seien dabei a priori bekannte, nicht stochastische Größen. Für σ^2 spezifizieren wir eine inverse Gammaverteilung mit Hyperparametern a und b, d.h.

$$\sigma^2 \sim \mathrm{IG}(a, b).$$

Mit diesen Spezifikationen handelt es sich bei der gemeinsamen Priori-Verteilung von $\boldsymbol{\beta}$ und σ^2 um eine sogenannte Normal-inverse Gammaverteilung mit Dichte

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2) &= p(\boldsymbol{\beta} \,|\, \sigma^2) p(\sigma^2) \qquad (3.37)\\
&= \frac{1}{(2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}} |\boldsymbol{M}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right)\\
&\quad \frac{b^a}{\Gamma(a)} \frac{1}{(\sigma^2)^{a+1}} \exp\left(-\frac{b}{\sigma^2}\right)
\end{aligned}
$$

und Hyperparametern $\boldsymbol{m}$, $\boldsymbol{M}$, a und b. Wir schreiben $\boldsymbol{\beta}, \sigma^2 \sim \mathrm{NIG}(\boldsymbol{m}, \boldsymbol{M}, a, b)$. Unter Vernachlässigung sämtlicher Faktoren in (3.37), die nicht von $\boldsymbol{\beta}$ und σ^2 abhängen, erhalten wir für die Dichte der Normal-inversen Gammaverteilung

$$
p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right) \frac{1}{(\sigma^2)^{a+1}} \exp\left(-\frac{b}{\sigma^2}\right). \qquad (3.38)
$$

Aus den Eigenschaften der multivariaten Normalverteilung und der inversen Gammaverteilung (vgl. Definition B.6 in Anhang B) folgt unmittelbar

$$
\mathrm{E}(\boldsymbol{\beta} \,|\, \sigma^2) = \boldsymbol{m} \qquad \mathrm{Cov}(\boldsymbol{\beta} \,|\, \sigma^2) = \sigma^2 \boldsymbol{M}.
$$

sowie

$$
\mathrm{E}(\sigma^2) = \frac{b}{a-1} \quad \text{falls a>1} \qquad \mathrm{Var}(\sigma^2) = \frac{b^2}{(a-1)^2(a-2)} \quad \text{falls a>2.}
$$

Im Folgenden bestimmen wir noch die *unbedingte* Priori-Verteilung von $\boldsymbol{\beta}$: Da der bedingte Erwartungswert von $\boldsymbol{\beta}$ unabhängig ist von σ^2 folgt (unter Verwendung des Satzes vom iterierten Erwartungswert) für den unbedingten Erwartungswert

$$
\mathrm{E}(\boldsymbol{\beta}) = \mathrm{E}(\mathrm{E}(\boldsymbol{\beta} \,|\, \sigma^2)) = \boldsymbol{m}.
$$

Für die unbedingte Kovarianzmatrix erhalten wir

$$
\mathrm{Cov}(\boldsymbol{\beta}) = \mathrm{E}(\mathrm{Cov}(\boldsymbol{\beta} \,|\, \sigma^2)) + \mathrm{Cov}(\mathrm{E}(\boldsymbol{\beta} \,|\, \sigma^2)) = \mathrm{E}(\sigma^2)\boldsymbol{M} = \frac{b}{a-1}\boldsymbol{M}.
$$

Zur Bestimmung der Verteilungstyps ist es zweckmäßig zunächst die bedingte Verteilung von $\sigma^2 \,|\, \boldsymbol{\beta}$ zu bestimmen. Die Dichte dieser Verteilung ist proportional zu (3.37) und es gilt

$$
\begin{aligned}
p(\sigma^2 \,|\, \boldsymbol{\beta}) &\propto p(\boldsymbol{\beta}, \sigma^2)\\
&\propto \tfrac{1}{(\sigma^2)^{a+\frac{p}{2}+1}} \exp\left(-\tfrac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right) \exp(-\tfrac{b}{\sigma^2}).
\end{aligned}
$$

Damit handelt es sich offenbar um die Form einer inversen Gammaverteilung und wir erhalten

$$
\sigma^2 \,|\, \boldsymbol{\beta} \sim IG\left(a + \frac{p}{2}, b + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right).
$$

Zur Bestimmung der unbedingten Verteilung von $\boldsymbol{\beta}$ nützen wir aus, dass diese durch Integration von (3.38) bezüglich σ^2 bestimmt werden kann. Da das Integral der Dichte von $\sigma^2 \,|\, \boldsymbol{\beta}$ eins sein muss, folgt

$$
\begin{aligned}
&\int \frac{1}{(\sigma^2)^{a+\frac{p}{2}+1}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right) \exp(-\frac{b}{\sigma^2})\, d\sigma^2\\
&= (b + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m}))^{-(a+\frac{p}{2})} \Gamma\left(a + \frac{p}{2}\right).
\end{aligned}
$$

Wir erhalten

$$\begin{aligned} p(\boldsymbol{\beta}) &\propto \left(b + \tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right)^{-(a+\frac{p}{2})} \Gamma\left(a+\tfrac{p}{2}\right) \\ &\propto \left(1 + \tfrac{1}{2a}(\boldsymbol{\beta}-\boldsymbol{m})'(\tfrac{b}{a}\boldsymbol{M})^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right)^{-(a+\frac{p}{2})}. \end{aligned}$$

Dies ist die Dichte einer multivariaten t-Verteilung mit $2a$ Freiheitsgraden, Lokationsparameter $\boldsymbol{m}$ und Dispersionsparameter $b/a\boldsymbol{M}$, d.h. $\boldsymbol{\beta} \sim t(2a, \boldsymbol{m}, b/a\boldsymbol{M})$. Zur multivariaten t-Verteilung vergleiche Seite 467 in Anhang B.

3.5.2 Vollständig bedingte Dichten und MCMC-Inferenz

Bayesianische Inferenz basiert auf den Charakteristika der Posteriori-Verteilung, d.h. der bedingten Verteilung der unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 bei gegebenen Daten $\boldsymbol{y}$. Die Dichte der Posteriori-Verteilung ist proportional zu dem Produkt aus Likelihood und Priori-Verteilung. Wir erhalten also

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) &\propto L(\boldsymbol{\beta}, \sigma^2) \cdot p(\boldsymbol{\beta}, \sigma^2) \qquad (3.39) \\ &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right) \\ &\quad \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right) \\ &\quad \frac{1}{(\sigma^2)^{a+1}} \exp(-\frac{b}{\sigma^2}). \end{aligned}$$

Technisch lassen sich Charakteristika der Posteriori-Verteilung im Wesentlichen auf zwei Arten bestimmen: Zum einen durch analytische Bestimmung des Posteriori-Verteilungstyps und damit der Charakteristika der Verteilung. Zum anderen durch Ziehung von Zufallszahlen aus der Posteriori-Verteilung basierend auf MCMC-Methoden. In der Mehrzahl der Fälle ist die analytische Bestimmung des Posteriori-Verteilungstyps unmöglich und daher nur der rechenzeitaufwändige Zugang über die Ziehung von Zufallszahlen möglich. Das lineare Modell ist eines der wenigen Beispiele, in dem die Posteriori-Verteilung analytisch zugänglich ist. Zur Illustration und zur Vorbereitung Bayesianischer Inferenztechniken in den nachfolgenden Kapiteln beschreiben wir jedoch zunächst den Zugang über die Ziehung von Zufallszahlen. Wir entwickeln einen Gibbs-Sampler (vergleiche Abschnitt B.5.3), bei dem nacheinander Zufallszahlen aus den vollständig bedingten Verteilungen von $\boldsymbol{\beta}$ und σ^2 gezogen werden. Im folgenden bestimmen wir zunächst die beiden vollständig bedingten Verteilungen, d.h. die bedingte Verteilung von $\boldsymbol{\beta}$ gegeben σ^2 und die Daten $\boldsymbol{y}$ und die bedingte Verteilung von σ^2 gegeben $\boldsymbol{\beta}$ und $\boldsymbol{y}$. Die Dichten beider Verteilungen sind jeweils proportional zur Dichte (3.39) der Posteriori-Verteilung. Durch Nichtbeachtung der nicht von $\boldsymbol{\beta}$ abhängigen Faktoren und durch Umstellen erhalten wir für die Dichte der vollständig bedingten Dichte von $\boldsymbol{\beta}$

$$
\begin{aligned}
p(\boldsymbol{\beta}\,|\cdot) &\propto \exp\left(-\tfrac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right)\exp\left(-\tfrac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right)\\
&\propto \exp\left(-\tfrac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{y}+\tfrac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}-\tfrac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}\right)\\
&\quad \exp\left(-\tfrac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{M}^{-1}\boldsymbol{\beta}+\tfrac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{M}^{-1}\boldsymbol{m}\right)\\
&\propto \exp\left(\tfrac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}-\tfrac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}-\tfrac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{M}^{-1}\boldsymbol{\beta}+\tfrac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{M}^{-1}\boldsymbol{m}\right)\\
&= \exp\left(\boldsymbol{\beta}'(\tfrac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{y}+\tfrac{1}{\sigma^2}\boldsymbol{M}^{-1}\boldsymbol{m})-\tfrac{1}{2}\boldsymbol{\beta}'(\tfrac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}+\tfrac{1}{\sigma^2}\boldsymbol{M}^{-1})\boldsymbol{\beta}\right).
\end{aligned}
$$

Damit besitzt die vollständig bedingte Dichte die Form einer multivariaten Normalverteilung, vergleiche in Anhang B Seite 464. Als Kovarianzmatrix erhalten wir

$$
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}+\frac{1}{\sigma^2}\boldsymbol{M}^{-1}\right)^{-1}. \tag{3.40}
$$

Der Erwartungswert $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ der Verteilung ergibt sich als

$$
\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{y}+\frac{1}{\sigma^2}\boldsymbol{M}^{-1}\boldsymbol{m}\right). \tag{3.41}
$$

Zusammenfassend erhalten wir $\boldsymbol{\beta}\,|\cdot \sim \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)$.
Durch analoge Vorgehensweise bestimmen wir die Dichte der vollständig bedingten Verteilung von σ^2. Zunächst gilt

$$
\begin{aligned}
p(\sigma^2\,|\,\cdot) \propto\ & \tfrac{1}{(\sigma^2)^{\frac{n}{2}}}\exp(-\tfrac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}))\\
& \tfrac{1}{(\sigma^2)^{\frac{p}{2}}}\exp(-\tfrac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m}))\tfrac{1}{(\sigma^2)^{a+1}}\exp(-\tfrac{b}{\sigma^2})\\
=\ & \tfrac{1}{(\sigma^2)^{a+\frac{n}{2}+\frac{p}{2}+1}}\\
& \exp\left(-\tfrac{1}{\sigma^2}\left(b+\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})+\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right)\right).
\end{aligned}
$$

Die vollständig bedingte Dichte von σ^2 besitzt damit die Form einer inversen Gammaverteilung mit Parametern

$$
a' = a+\frac{n}{2}+\frac{p}{2} \tag{3.42}
$$

und

$$
b' = b+\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m}). \tag{3.43}
$$

Damit können wir Charakteristika der Posteriori-Verteilung unter Verwendung des folgenden Gibbs-Samplers bestimmen:

1. Setze Startwerte $\boldsymbol{\beta}^{(0)}$ und $(\sigma^2)^{(0)}$.
2. Für $t = 1, 2, \ldots, T$ ziehe zunächst Zufallszahlen $\boldsymbol{\beta}^{(t)}$ aus der vollständig bedingten Dichte von $\boldsymbol{\beta}$, wobei in (3.40) und (3.41) der aktuelle Wert $(\sigma^2)^{(t-1)}$ für σ^2 eingesetzt wird. Ziehe anschließend eine Zufallszahl $(\sigma^2)^{(t)}$ aus der vollständig bedingten Dichte von σ^2, wobei in (3.43) der aktuelle Wert $\boldsymbol{\beta}^{(t)}$ für $\boldsymbol{\beta}$ eingesetzt wird.
3. Nach einer gewissen Konvergenzzeit t_0 können die gezogenen Zufallszahlen $\boldsymbol{\beta}^{(t_0+1)}$, $\ldots$, $\boldsymbol{\beta}^{(T)}$ bzw. $(\sigma^2)^{(t_0+1)}, \ldots, (\sigma^2)^{(T)}$ als Ziehungen aus den marginalen Posteriori-Verteilungen von $\boldsymbol{\beta}$ und σ^2 angesehen werden. Charakteristika der Posteriori können

dann durch ihre empirischen Analoga geschätzt werden, z.B. der Posteriori-Erwartungswert von $\boldsymbol{\beta}$ durch

$$\hat{\boldsymbol{\beta}} = \widehat{\mathrm{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y})} = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \boldsymbol{\beta}^{(t)}.$$

Bayesianisches lineares Modell

Beobachtungsmodell und Priori-Verteilung

1. *Beobachtungsmodell:* $y \,|\, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$.
2. *Priori-Verteilung:* $\boldsymbol{\beta} \,|\, \sigma^2 \sim \mathrm{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{M})$ und $\sigma^2 \sim \mathrm{IG}(a, b)$.

Posteriori-Verteilungen

1. *Vollständig bedingte Dichte von* $\boldsymbol{\beta}$*:* $\boldsymbol{\beta} \,|\, \cdot \sim \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)$ mit

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{M}^{-1}\right)^{-1} \qquad \boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\left(\frac{1}{\sigma^2}\boldsymbol{X}'y + \frac{1}{\sigma^2}\boldsymbol{M}^{-1}\boldsymbol{m}\right).$$

2. *Vollständig bedingte Dichte von* σ^2*:* $\sigma^2 \,|\, \cdot \sim \mathrm{IG}(a', b')$ mit

$$a' = a + \frac{n}{2} + \frac{p}{2}$$

und

$$b' = b + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}).$$

3. *Posteriori-Verteilung von* $\boldsymbol{\beta}, \sigma^2$*:* $\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y} \sim \mathrm{NIG}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{M}}, \tilde{a}, \tilde{b})$ mit Parametern

$$\tilde{\boldsymbol{M}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1} \qquad \tilde{\boldsymbol{m}} = \tilde{\boldsymbol{M}}(\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y})$$

sowie

$$\tilde{a} = a + \frac{n}{2} \qquad \tilde{b} = b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} - \tilde{\boldsymbol{m}}'\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{m}}\right).$$

Posteriori-Inferenz

1. *Gibbs-Sampler:* Verwende einen Gibbs-Sampler, bei dem nacheinander Zufallszahlen aus den vollständig bedingten Verteilungen $\boldsymbol{\beta} \,|\, \cdot$ und $\sigma^2 \,|\, \cdot$ gezogen werden.
2. *Charakteristika der Posteriori-Verteilung:* Erhalte Punkt- und Intervallschätzer durch Charakteristika der $\mathrm{NIG}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{M}}, \tilde{a}, \tilde{b})$ Verteilung, z.B. $\hat{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \tilde{\boldsymbol{m}}$.

3.5.3 Posteriori-Verteilung

Wie bereits erwähnt besteht im linearen Modell eine Alternative zur Inferenz basierend auf MCMC-Methoden, da die Posteriori-Verteilung analytisch zugänglich ist. Es lässt sich zeigen, dass die Posteriori-Verteilung wie die Priori-Verteilung eine Normal-inverse Gammaverteilung ist. Die Parameter $\tilde{\boldsymbol{m}}$, $\tilde{\boldsymbol{M}}$, $\tilde{a}$ und $\tilde{b}$ der Verteilung findet man im nachfolgenden Kasten und eine Herleitung in Abschnitt 3.7.2 auf Seite 188.

Unter Ausnutzung der Eigenschaften der Normal-inverse Gammaverteilung lassen sich interessierende Charakteristika der Posteriori-Verteilung in geschlossener Form gewinnen. Insbesondere erhalten wir als Punktschätzer für $\boldsymbol{\beta}$ den unbedingten Posteriori-Erwartungswert

$$\hat{\boldsymbol{\beta}}^B = \mathrm{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \tilde{\boldsymbol{m}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1}(\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y}).$$

Unter Verwendung der Matrix $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{X}$ können wir den Bayes-Schätzer als gewichteten Mittelwert aus Priori-Erwartungswert $\boldsymbol{m}$ und KQ-Schätzer $\hat{\boldsymbol{\beta}}$ schreiben:

$$\hat{\boldsymbol{\beta}}^B = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{m} + \boldsymbol{A}\hat{\boldsymbol{\beta}}$$

Je größer die Diagonalelemente von $\boldsymbol{M}$ sind, desto kleiner sind die Elemente von $\boldsymbol{M}^{-1}$ und $\boldsymbol{A}$ nähert sich einer Einheitsmatrix an. Umgekehrt bedeuten kleine Elemente in $\boldsymbol{M}$, dass sich die Matrix $\boldsymbol{A}$ der Nullmatrix bzw. $\boldsymbol{I} - \boldsymbol{A}$ der Einheitsmatrix annähert. Dies führt zu folgender Interpretation von $\hat{\boldsymbol{\beta}}^B$: Je geringer die Vorinformation über $\boldsymbol{\beta}$, d.h. je größer die Elemente von $\boldsymbol{M}$ sind, desto mehr nähert sich $\hat{\boldsymbol{\beta}}^B$ dem KQ-Schätzer an. Je größer die Vorinformation, d.h. je kleiner die Diagonalelemente von $\boldsymbol{M}$ sind, desto mehr wird $\hat{\boldsymbol{\beta}}^B$ vom Priori-Erwartungswert $\boldsymbol{m}$ dominiert.

3.6 Modellwahl und Variablenselektion

In vielen Anwendungsfällen stehen eine Vielzahl von potentiellen Einflussvariablen zur Verfügung und es stellt sich die Frage, welche der vorhandenen Variablen ins Regressionsmodell aufgenommen werden sollen. Beim Mietspiegel für München beispielsweise werden durch den Fragebogen circa 200 Variablen erfasst. Zwei naive (gleichwohl häufig praktizierte) Herangehensweisen sind die Folgenden:

- Strategie 1: Schätze ein möglichst komplexes Modell, in dem alle zur Verfügung stehenden Kovariablen ins Regressionsmodell aufgenommen werden.
- Strategie 2: Schätze zunächst ein Modell mit allen Variablen und entferne in einem zweiten Schritt die nicht signifikanten Variablen wieder aus dem Modell.

Wir diskutieren diese Vorgehensweisen zunächst anhand von zwei illustrierenden Beispielen mit simulierten Datensätzen.

Beispiel 3.24 Polynomiale Regression

Wir betrachten die in Abbildung 3.25 a) dargestellten simulierten Daten. Aufgrund des Streudiagramms erscheint eine polynomiale Modellierung des durchschnittlichen Zusammenhangs zwischen y und x sinnvoll. Wir verwenden also das Schätzmodell

Abb. 3.25. *Simulierte Trainingsdaten y_i (Abbildung a) und Validierungsdaten y_i^* (Abbildung b) zu 50 Designpunkten x_i, $i = 1, \ldots, 50$. Die Grafiken c)-e) zeigen geschätzte Polynome vom Grad $l = 1, 2, 5$ basierend auf den Trainingsdaten. Grafik f) zeigt die durchschnittliche quadratische Abweichung MSE(l) der Schätzungen von den Daten in Abhängigkeit vom Polynomgrad (durchgezogene Linie). Die gestrichelte Linie zeigt MSE(l), wenn die geschätzten Polynome zur Prognose der neuen Daten y_i^* verwendet werden.*

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_l x_i^l + \varepsilon_i.$$

Wie bei jeder polynomialen Regression muss der Polynomgrad l gewählt werden. Die Abbildungen 3.25 c) – e) zeigen den geschätzten Zusammenhang für $l = 1$ (Regressionsgerade), $l = 2$ und $l = 5$. Offenbar ist eine einfache Regressionsgerade, d.h. ein

Polynom vom Grad 1, den Daten nicht angemessen. Ab dem Polynomgrad $l > 2$ scheint eine zufrieden stellende Anpassung an die Daten gewährleistet. Abbildung 3.25 f) zeigt zusätzlich die durchschnittliche quadratische Abweichung

$$\mathrm{MSE}(l) = \frac{1}{50}\sum_{i=1}^{50}(y_i - \hat{y}_i(l))^2$$

der Schätzungen in Abhängigkeit vom Grad des Polynoms (durchgezogene Linie). Offenbar sinkt $\mathrm{MSE}(l)$ monoton mit wachsendem l, d.h. die Anpassung an die Daten wird mit größerem Polynomgrad immer besser. Die bisherigen Beobachtungen scheinen also die oben beschriebene erste Strategie zu bestätigen, möglichst viele Regressoren ins Modell aufzunehmen.

Als nächsten Schritt wollen wir uns ansehen, wie gut die geschätzten Funktionen neue gemäß dem selben Modell simulierte Daten prognostizieren. Abbildung 3.25 b) zeigt für jeden Designpunkt x_i, $i = 1, \ldots 50$, je eine zusätzlich simulierte Beobachtung y_i^*. Wir bezeichnen diesen Datensatz als Validierungsstichprobe. Den zur Schätzung verwendeten ersten Datensatz bezeichnen wir als Trainingsstichprobe. Abbildung 3.25 f) zeigt auch für den Validierungsdatensatz die durchschnittliche quadratische Abweichung der Schätzungen $\hat{y}_i^*$ von den Daten y_i^* (gestrichelte Linie). Offenbar wird die Anpassung an die neuen Daten zunächst mit steigendem Polynomgrad besser, ab dem Polynomgrad $l = 4$ wird die Anpassung aber wieder schlechter. Wir erkennen: Je komplexer das Modell, desto besser ist die Anpassung an die zur Schätzung benutzten Daten. Bei neuen, aus dem gleichen Modell stammenden Daten, führen zu komplexe Modelle jedoch zu einer schlechteren Anpassung.

△

Beispiel 3.25 Korrelierte Kovariablen

Betrachte die $n = 150$ Daten $(y_i, x_{i1}, x_{i2}, x_{i3})$, $i = 1, \cdots, 150$, in der Streudiagramm-Matrix von Abbildung 3.26. Die Daten sind wie folgt erzeugt: Die Variablen x_1 und x_3 sind unabhängig und gleichverteilt auf [0,1]. Die Variable x_2 ist definiert als $x_2 = x_1 + u$, wobei u auf [0,1] gleichverteilt ist. Die Variablen x_1 und x_2 sind also relativ stark korreliert. Die Zielgröße y ist dann gemäß dem Modell

$$y \mid x_1, x_2, x_3 \sim \mathrm{N}(-1 + 0.3x_1 + 0.2x_3, 0.2^2)$$

erzeugt. Der bedingte Erwartungswert von y hängt also in Wahrheit von x_1 und x_3 jedoch nicht von x_2 ab. Im Folgenden gehen wir aber davon aus, dass wir das wahre Modell nicht kennen. Wir schätzen zunächst ein Regressionsmodell mit allen zur Verfügung stehenden Kovariablen x_1, x_2 und x_3 und erhalten die in Tabelle 3.8 abgedruckten Schätzergebnisse. Offenbar sind x_1 und x_2 beide nicht signifikant. Würden wir nach Strategie 2 vorgehen und die nicht signifikanten Variablen aus dem Modell entfernen, so würden wir neben der nicht relevanten Kovariable x_2 auch die relevante Variable x_1 entfernen.

Wenn wir stattdessen ein korrekt spezifiziertes Modell mit den wahren Einflussvariablen x_1 und x_3 schätzen, erhalten wir die Ergebnisse in Tabelle 3.9. Bei korrekter Modellspezifikation ist also neben x_3 auch die vorher nicht signifikante Variable x_1 signifikant.

Wie stellen also fest: Wenn wir zunächst alle Variablen aufnehmen und anschließend die nicht signifikanten Variablen wieder aus dem Modell entfernen, kann es passieren, dass auch wichtige Variablen entfernt werden. Hauptgrund hierfür sind vorhandene Korrelationen der Kovariablen untereinander.

△

Abb. 3.26. *Streudiagramm-Matrix für die Variablen y, x_1, x_2 und x_3.*

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	-0.970	0.047	-20.46	<0.001	-1.064	-0.877
x_1	0.146	0.187	0.78	0.436	-0.224	0.516
x_2	0.027	0.177	0.15	0.880	-0.323	0.377
x_3	0.227	0.052	4.32	<0.001	0.123	0.331

Tabelle 3.8. *Ergebnisse für das Schätzmodell mit den Kovariablen x_1, x_2 und x_3.*

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	-0.967	0.039	-24.91	<0.001	-1.042	-0.889
x_1	0.173	0.055	3.17	0.002	0.065	0.281
x_3	0.226	0.052	4.33	<0.001	0.123	0.330

Tabelle 3.9. *Ergebnisse für das Schätzmodell mit den Kovariablen x_1 und x_3.*

3.6.1 Auswirkung der Modellspezifikation auf Bias, Varianz und Prognosegüte

Wir wollen jetzt die in den Beispielen gewonnenen Einsichten durch einige theoretische Überlegungen weiter festigen. Insbesondere interessieren wir uns für die folgenden Fragen:

1. *Irrelevante Variablen:* Was lässt sich über den Bias und die Varianz des KQ-Schätzers sagen, wenn wir auch Variablen ins Modell mit aufnehmen, die keinen Einfluss besitzen?
2. *Fehlende Variablen:* Was lässt sich über den Bias und die Varianz des KQ-Schätzers sagen, wenn wir relevante Variablen im Modell nicht berücksichtigen?
3. *Prognosegüte:* Welche Auswirkungen hat die Modellspezifikation, d.h. die Auswahl der zur Verfügung stehenden Variablen, auf die Prognosegüte?

Auswirkung der Modellspezifikation auf Bias und Varianz des KQ-Schätzers

Für die folgenden Überlegungen zerlegen wir die zur Verfügung stehenden erklärenden Variablen $\boldsymbol{x} = (x_0, x_1, \ldots, x_k)'$ mit $x_0 \equiv 1$ in die Teilvektoren $\boldsymbol{x}_1 = (x_0, x_1, \ldots, x_{k_1})'$ und $\boldsymbol{x}_2 = (x_{k_1+1}, \ldots, x_k)'$. Wir betrachten die beiden Schätzmodelle

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

und

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}.$$

Das erste Schätzmodell verwendet alle zur Verfügung stehenden Variablen, das zweite Schätzmodell nur eine Teilmenge. Als KQ-Schätzer erhält man

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad \text{und} \quad \tilde{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}.$$

Für den Schätzer $\tilde{\boldsymbol{\beta}}_1$ des Teilmodells ergeben sich die Eigenschaften

$$\begin{aligned}
\mathrm{E}(\tilde{\boldsymbol{\beta}}_1) &= \mathrm{E}((\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}) \\
&= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\mathrm{E}(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) \\
&= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2) \\
&= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2
\end{aligned}$$

sowie

$$\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_1) = \mathrm{Cov}((\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}) = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\sigma^2\boldsymbol{I}\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} = \sigma^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}.$$

Die statistischen Eigenschaften des Schätzers $\hat{\boldsymbol{\beta}}$ im vollen Modell haben wir bereits in Abschnitt 3.2.3 bestimmt, vergleiche den Kasten auf Seite 104.

Wir wollen jetzt die beiden folgenden Situationen unterscheiden:

- *Fehlende Variablen:* Das volle Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ist korrekt, wir schätzen aber fälschlicherweise das Submodell $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$. In diesem Fall haben wir die eigentlich relevanten Variablen in $\boldsymbol{x}_2$ vergessen.
- *Irrelevante Variablen:* Das Submodell $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$ ist korrekt, wir schätzen aber fälschlicherweise das volle Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In diesem Fall haben wir irrelevante Variablen ins Modell aufgenommen, die Variablen in $\boldsymbol{x}_2$ sind überflüssig.

Im ersten Fall fehlender Variablen folgt:

- $\hat{\boldsymbol{\beta}}_1$ ist verzerrt, außer für $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$. Die Bedingung $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$ bedeutet, dass jede Variable in $\boldsymbol{X}_1$ unkorreliert zu jeder Variable in $\boldsymbol{X}_2$ ist.
- Es lässt sich zeigen, dass die Differenz $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1) - \mathrm{Cov}(\tilde{\boldsymbol{\beta}}_1)$ der Kovarianzmatrizen positiv semi-definit ist. Dies bedeutet insbesondere, dass die Komponenten des Schätzers $\tilde{\boldsymbol{\beta}}_1$, basierend auf dem Teilmodell $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$ eine geringere Varianz aufweisen als die entsprechenden Komponenten des Schätzer $\hat{\boldsymbol{\beta}}_1$ basierend auf dem korrekten Modell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Es gilt also $\mathrm{Var}(\hat{\beta}_j) \geq \mathrm{Var}(\tilde{\beta}_j)$.
- Darüber hinaus lässt sich zeigen, dass es Situationen geben kann, in denen die Komponenten in $\tilde{\boldsymbol{\beta}}_1$ basierend auf dem eigentlich fehlspezifizierten Teilmodell einen kleineren MSE aufweisen, als die Komponenten in $\hat{\boldsymbol{\beta}}_1$ basierend auf dem vollen Modell, d.h. $\mathrm{MSE}(\hat{\beta}_j) \geq \mathrm{MSE}(\tilde{\beta}_j)$. Es ist also durchaus möglich, dass ein sparsames Modell mit nicht berücksichtigten Kovariablen bessere statistische Eigenschaften besitzt als das korrekt spezifizierte Modell.

Im zweiten Fall irrelevanter Variablen folgt:

- Obwohl irrelevante Variablen berücksichtigt wurden, ist $\hat{\boldsymbol{\beta}}$ (genauso wie der auf dem wahren Modell basierende Schätzer $\hat{\boldsymbol{\beta}}_1$) unverzerrt.
- Es lässt sich zeigen, dass Schätzungen für die Komponenten in $\hat{\boldsymbol{\beta}}_1$ basierend auf $\hat{\boldsymbol{\beta}}$ variabler sind als Schätzungen basierend auf $\tilde{\boldsymbol{\beta}}_1$. Es gilt also $\mathrm{Var}(\hat{\beta}_j) \geq \mathrm{Var}(\tilde{\beta}_j)$. Wenn wir irrelevante Variablen im Modell belassen, verlieren wir also bei der Schätzung an Genauigkeit.

Analoge Aussagen lassen sich über $\hat{\boldsymbol{y}}$ treffen.

Damit können wir folgendes Fazit ziehen: Das gewählte Modell sollte möglichst keine irrelevanten Kovariablen enthalten. Darüber hinaus ist eine möglichst sparsame Modellspezifikation anzustreben, so dass Bias und Varianz insgesamt, d.h. der MSE, möglichst klein sind.

Auswirkung der Modellspezifikation auf die Prognosegüte

Wir wollen jetzt einige Überlegungen zur Güte von Prognosen im linearen Modell anstellen. Dabei nehmen wir nicht notwendigerweise an, dass das Modell korrekt spezifiziert ist. Die Ergebnisse unserer Überlegungen erweisen sich dann bei der Konstruktion von Modellgütemaßen zum Vergleich unterschiedlicher Modelle als nützlich.

Wie bisher gehen wir von unabhängigen Beobachtungen y_i, $i = 1, \ldots, n$, mit Erwartungswert $\mathrm{E}(y_i) = \mu_i$ und $\mathrm{Var}(y_i) = \sigma^2$ aus. Als potentielle Regressoren stehen wieder die Variablen $x_0 = 1, x_1, \ldots, x_k$ zur Verfügung. In den weiteren Überlegungen gehen wir davon aus, dass eine Teilmenge der zur Verfügung stehenden Variablen zur Schätzung verwendet wird. Das Schätzmodell wird beschrieben durch die Teilmenge $M \subset \{0, 1, 2, \ldots, k\}$ der ins Modell aufgenommenen Variablen. Die zugehörige Designmatrix bezeichnen wir mit $\boldsymbol{X}_M$. Als KQ-Schätzer erhalten wir

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

Ein Schätzer $\hat{\boldsymbol{y}}_M$ für den Vektor $\boldsymbol{\mu}$ der Erwartungswerte $\mu_i = \mathrm{E}(y_i)$ ist dann gegeben durch

$$\hat{\boldsymbol{y}}_M = \boldsymbol{X}_M \hat{\boldsymbol{\beta}}_M.$$

Wir können die Schätzungen $\hat{y}_{iM}$ auch als Prognose für zukünftige Beobachtungen $y_{n+i} = \mu_i + \varepsilon_{n+i}$, $i = 1, \ldots, n$, bei gegebenen Kovariablen $x_{i1}, \ldots, x_{ik}$ ansehen. Wir leiten im Folgenden eine Formel für die Summe der erwarteten quadrierten Prognosefehler $\sum \mathrm{E}(y_{n+i} - \hat{y}_{iM})^2$ her. Dazu benötigen wir die folgenden leicht nachweisbaren Eigenschaften von $\hat{\boldsymbol{y}}_M$:

1. Erwartungswert:
$$\mathrm{E}(\hat{\boldsymbol{y}}_M) = \boldsymbol{X}_M (\boldsymbol{X}_M' \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M' \boldsymbol{\mu}.$$
2. Kovarianzmatrix:
$$\mathrm{Cov}(\hat{\boldsymbol{y}}_M) = \sigma^2 \boldsymbol{X}_M (\boldsymbol{X}_M' \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M'.$$
3. Summe der Varianzen:
$$\sum_{i=1}^n \mathrm{Var}(\hat{y}_{iM}) = |M|\, \sigma^2,$$
wobei $|M|$ die Mächtigkeit von M ist, d.h. die Anzahl der ins Modell aufgenommenen Kovariablen. Die Summe der Varianzen wird umso größer, je mehr Kovariablen im Modell aufgenommen werden.
4. Summe der erwarteten quadrierten Abweichungen (SMSE):
$$\mathrm{SMSE} = \sum_{i=1}^n \mathrm{E}(\hat{y}_{iM} - \mu_i)^2 = |M|\, \sigma^2 + \sum_{i=1}^n (\mu_{iM} - \mu_i)^2.$$
Dabei haben wir $\mu_{iM} = \mathrm{E}(\hat{y}_{iM})$ als Abkürzung für den Erwartungswert der Schätzung $\hat{y}_{iM}$ bezeichnet.

Mit diesen Eigenschaften erhalten wir für den erwarteten quadrierten Prognosefehler:

$$\begin{aligned}
\mathrm{SPSE} &= \sum_{i=1}^n \mathrm{E}(y_{n+i} - \hat{y}_{iM})^2 \\
&= \sum_{i=1}^n \mathrm{E}((y_{n+i} - \mu_i) - (\hat{y}_{iM} - \mu_i))^2 \\
&= \sum_{i=1}^n (\mathrm{E}(y_{n+i} - \mu_i)^2 - 2\mathrm{E}((y_{n+i} - \mu_i)(\hat{y}_{iM} - \mu_i)) + \mathrm{E}(\hat{y}_{iM} - \mu_i)^2) \\
&= \sum_{i=1}^n \mathrm{E}(y_{n+i} - \mu_i)^2 + \sum_{i=1}^n \mathrm{E}(\hat{y}_{iM} - \mu_i)^2 \\
&\qquad = n\,\sigma^2 + \mathrm{SMSE} \\
&\qquad = n\,\sigma^2 + |M|\, \sigma^2 + \sum_{i=1}^n (\mu_{iM} - \mu_i)^2
\end{aligned}$$

In Zeile vier haben wir ausgenutzt, dass der Erwartungswert für das gemischte Glied wegen der Unabhängigkeit von y_i und y_{n+i} als Produkt der Erwartungswerte geschrieben werden kann, so dass der gesamte Term Null wird. Der erwartete quadrierte Prognosefehler lässt sich also additiv in drei Terme zerlegen:

- *Irreduzibler Prognosefehler:* Der erste Term $n\,\sigma^2$ hängt von der Varianz des Fehlerterms ab und ist daher auch nicht durch noch so ausgeklügelte Inferenzmethoden und Modellwahlverfahren beeinflussbar. Daher kann dieser Term als irreduzibler Prognosefehler angesehen werden.
- *Varianz:* Beim zweiten Term handelt es sich um die Summe der Varianzen $\mathrm{Var}(\hat{y}_{iM})$ der Schätzungen $\hat{y}_{iM}$. Dieser Term ist durch Modellwahl beeinflussbar und wird umso kleiner, je *weniger* Variablen im Modell enthalten sind.
- *Quadrierter Bias:* Der letzte Term $\sum(\mu_{iM} - \mu_i)^2$ kann als Biasterm aufgefasst werden. Er erfasst den quadrierten Bias der Schätzung $\hat{y}_{iM}$ für den Erwartungswert μ_i. Dieser Term ist ebenfalls durch Modellwahl beeinflussbar und wird umso kleiner, je *mehr* Variablen im Modell enthalten sind.

Die Zerlegung des erwarteten Prognosefehlers in einen irreduziblen Fehler, einen Varianzterm und einen quadrierten Biasterm ist nicht auf die Klasse der linearen Modelle beschränkt. Es handelt sich um eine *fundamentale Eigenschaft von Prognosen* in allen statistischen Modellen.

Die Formel für SPSE zeigt einen klassischen Bias-Varianz-Trade Off auf. Je komplexer wir das Modell machen, desto geringer der quadrierte Bias und desto größer gleichzeitig die Varianz. Umgekehrt weisen einfachere Modelle einen größeren quadrierten Bias und im Ausgleich eine kleinere Varianz auf. Dieser Bias-Varianz-Trade Off findet sich nicht nur im linearen Modell wieder, sondern kann allgemein in statistischen Modellen beobachtet werden, vergleiche beispielsweise in Kapitel 7.1.8 (Seite 348) die Diskussion zum Bias-Varianz-Trade Off bei Glättungsverfahren.

3.6.2 Modellwahlkriterien

Die theoretischen Überlegungen zur Prognosegüte im letzten Abschnitt, insbesondere die fundamentale Formel zum erwarteten quadrierten Prognosefehler, liefern Werkzeuge zur Auswahl relevanter Variablen. Eine denkbare Vorgehensweise beruht auf der *Minimierung* des erwarteten quadratischen Prognosefehlers: Wir nehmen diejenigen Kovariablen in das Modell auf, die SPSE minimieren. Leider ist SPSE nicht direkt zugänglich, weil μ_i und σ^2 unbekannt sind. Wir benötigen also eine Schätzung für SPSE. Dazu sind zwei Strategien denkbar:

1. *Schätze SPSE mit Hilfe neuer, unabhängiger Daten*

   Wenn wir tatsächlich zusätzliche Beobachtungen y_{n+i} zur Verfügung haben, können wir $\mathrm{SPSE} = \sum \mathrm{E}(y_{n+i} - \hat{y}_{iM})^2$ einfach durch

   $$\widehat{\mathrm{SPSE}} = \sum_{i=1}^{n} (y_{n+i} - \hat{y}_{iM})^2$$

   abschätzen. In der Praxis ist diese Vorgehensweise natürlich nicht ohne weiteres umsetzbar, da in den seltensten Fällen zusätzliche Beobachtungen erhoben werden. Ein gangbarer Weg ist der Folgende:

   - Unterteile die Daten zufällig in zwei Teile, eine Test- und eine Validierungsstichprobe.
   - Verwende den Testdatensatz bei jeweils gegebener Modellspezifikation zur Schätzung der Modelle, d.h. zur Schätzung der Regressionskoeffizienten $\boldsymbol{\beta}$ und der Erwartungswerte μ_i.

- Verwende den Validierungsdatensatz zur Beurteilung der Modellgüte, d.h. zur Schätzung von SPSE.

Diese Vorgehensweise kann nur verwendet werden, wenn der zur Verfügung stehende Datensatz groß genug ist. Bei zu kleinen Datensätzen würde durch die Aufteilung der Daten zu viel Schätzgenauigkeit für die Schätzung der Regressionskoeffizienten verloren gehen. Wie groß der Stichprobenumfang sein muss, damit die vorgeschlagene Vorgehensweise vernünftig ist, hängt von der Problemstellung ab. Nach Kenntnis der Autoren gibt es hierzu auch keine Faustregeln oder allgemeine Hinweise in der Literatur. Aus diesen Gründen wird häufig von der zweiten im Folgenden beschriebenen Strategie Gebrauch gemacht.

2. *Schätze SPSE mit Hilfe der vorhandenen Daten*

Eine naive Schätzung für SPSE wäre die Verwendung der Residuenquadratsumme $\sum(y_i - \hat{y}_{iM})^2$. Diese *unterschätzt* jedoch im Mittel den erwarteten Prognosefehler, denn es lässt sich zeigen, dass

$$\mathrm{E}\left(\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2\right) = \mathrm{SPSE} - 2|M|\sigma^2.$$

Wir sollten also SPSE besser durch

$$\widehat{\mathrm{SPSE}} = \sum_{i=1}^{n}(y_i - y_{iM})^2 + 2|M|\hat{\sigma}^2$$

abschätzen und das Modell wählen, welches $\widehat{\mathrm{SPSE}}$ minimiert. Zu beachten ist dabei, dass wir immer die gleiche Schätzung für $\hat{\sigma}^2$ verwenden sollten. Diese sollte möglichst auf dem Modell mit allen Variablen beruhen, um die Verzerrung in $\hat{\sigma}^2$ klein zu halten. Das Kriterium $\widehat{\mathrm{SPSE}}$ besitzt eine für die meisten Modellwahlkriterien typische Struktur. Es besteht aus zwei Termen: Der erste Term, die Residuenquadratsumme, misst die Güte der Anpassung an die Daten und wird umso kleiner, je komplexer das Modell wird. Der zweite Term $2|M|\sigma^2$ ist ein Maß für die Modellkomplexität und wird umso kleiner, je sparsamer das Modell ist. Im Folgenden stellen wir die wichtigsten Modellwahlkriterien im linearen Modell zusammen. Diese sind durch ähnliche Überlegungen wie in diesem Abschnitt motiviert.

Das korrigierte Bestimmtheitsmaß

In Abschnitt 3.2.3 wurde das Bestimmtheitsmaß R^2 zur Beurteilung der Güte der Anpassung an die Daten definiert. Zum Vergleich verschiedener Modelle ist das Bestimmtheitsmaß jedoch nur bedingt geeignet, da es automatisch größer wird wenn eine neue Kovariable ins Modell aufgenommen wird. Dieses Problem soll mit dem sogenannten *korrigierten Bestimmtheitsmaß* verhindert werden, indem eine Korrektur für die Anzahl der Parameter angeführt wird, so dass das Gütemaß nicht notwendigerweise größer wird wenn eine zusätzliche Kovariable aufgenommen wird. Das korrigierte Bestimmtheitsmaß ist definiert durch

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2).$$

Das korrigierte Bestimmtheitsmaß ist sehr populär und wird in allen statistischen Programmpaketen standardmäßig ausgegeben. An dieser Stelle raten wir von der Verwendung eher ab, da die „Bestrafung“ für neu aufgenommene Kovariablen zu gering erscheint. Man kann nämlich zeigen, dass $\bar{R}^2$ bereits dann steigt, wenn eine Variable, deren t-Wert betragsmäßig größer als 1 ist, ins Modell aufgenommen wird. Wir würden also bereits Variablen mit einem p-Wert von circa 0.3 zusätzlich aufnehmen.

Mallows' C_p

Mallows' C_p („Complexity parameter“) beruht direkt auf den oben angestellten Überlegungen und ist definiert durch

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + 2|M|.$$

C_p kann als eine Schätzung für SMSE/σ^2 aufgefasst werden. Damit liefert eine Minimierung von C_p das selbe Modell wie eine Minimierung von $\widehat{\text{SPSE}}$.

Informationskriterium nach Akaike AIC

Das im Rahmen der Maximum-Likelihood-Inferenz am häufigsten verwendete Modellwahlkriterium ist das Informationskriterium nach Akaike (AIC), vergleiche in Anhang B Seite 477. Allgemein ist das AIC definiert durch

$$\text{AIC} = -2 \cdot l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) + 2\,|M+1|$$

wobei $l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2)$ der maximale Wert der Log-Likelihood ist, d.h. wenn in die Log-Likelihood die ML-Schätzer $\hat{\boldsymbol{\beta}}_M$ und $\hat{\sigma}^2$ eingesetzt werden. Man beachte, dass die Gesamtzahl an Parametern $M+1$ ist. Im linearen Modell mit normalverteilten Störgrößen erhält man

$$\begin{aligned} -2\,l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) &= n\,\log(\hat{\sigma}^2) + \tfrac{1}{\hat{\sigma}^2}(\boldsymbol{y} - \boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M)'(\boldsymbol{y} - \boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M) \\ &= n\,\log(\hat{\sigma}^2) + \frac{n\hat{\sigma}^2}{\hat{\sigma}^2} \\ &= n\,\log(\hat{\sigma}^2) + n \end{aligned}$$

und damit unter Vernachlässigung der Konstanten n

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2) + 2\,|M+1|.$$

Man beachte, dass hier der ML-Schätzer $\hat{\sigma}^2 = 1/n\,\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ eingesetzt wird und nicht der sonst übliche unverzerrte Varianzschätzer.

Kreuzvalidierung

Kreuzvalidierung *imitiert* die weiter oben beschriebene Aufteilung der Daten in einen Testdatensatz zur Schätzung der Modellparameter und einen Validierungsdatensatz zur Abschätzung der Prognosegüte. Der Kreuzvalidierung liegt folgende allgemeine Vorgehensweise zugrunde:

- Zerlege den Datensatz in K Teildatensätze $1, \ldots, K$ ungefähr gleicher Größe.
- Verwende zunächst den ersten Datensatz als Validierungsstichprobe und die restlichen $K-1$ Datensätze zum Schätzen der Parameter. Basierend auf den Schätzungen werden die Daten in der Validierungsstichprobe prognostiziert und die quadrierten Prognosefehler ermittelt und aufsummiert.
- Verwende jetzt nacheinander den zweiten, dritten usw. bis K-ten Datensatz als Validierungsstichprobe und jeweils die restlichen Datensätze zur Schätzung. Summiere die erhaltenen quadrierten Prognosefehler auf.
- Verwende das Modell mit der kleinsten Summe der quadrierten Prognosefehler, wobei die endgültigen Parameterschätzungen dann auf allen Daten beruhen. Die Aufteilung in Test- und Validierungsstichprobe dient nur zur Schätzung des erwarteten quadrierten Prognosefehlers.

Ein wichtiger Spezialfall ist die sogenannte „leave-one-out“ Kreuzvalidierung, bei der jeweils alle Beobachtungen bis auf *eine* zur Schätzung der Modellparameter verwendet werden. Dieser „leave-one-out“ Schätzer wird dann zur Prognose der nicht verwendeten Beobachtung verwendet und die quadrierte Abweichung der Prognose vom beobachteten Wert bestimmt. Wenn wir mit $\hat{y}_{iM}^{-i}$ den „leave-one-out“ Schätzer bezeichnen, erhalten wir die sogenannte Kreuzvalidierungsfunktion

$$\mathrm{CV} = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{y}_{iM})^{-i})^2.$$

Auf den ersten Blick erscheint die Berechnung der Kreuzvalidierungsfunktion CV äußerst aufwändig, da die KQ-Schätzer n mal berechnet werden müssen. Es lässt sich jedoch zeigen, dass die Kreuzvalidierungsfunktion unter Zuhilfenahme der ursprünglichen auf allen Daten basierenden Schätzungen $\hat{y}_{iM}$ berechenbar ist. Es gilt

$$\mathrm{CV} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_{iM}}{1 - h_{iiM}} \right)^2,$$

wobei h_{iiM} die Diagonalelemente der Prädiktionsmatrix $\boldsymbol{H}_M = \boldsymbol{X}_M (\boldsymbol{X}'_M \boldsymbol{X})^{-1} \boldsymbol{X}'_M$ sind.

Bayesianisches Informationskriterium BIC

Das BIC ist allgemein gegeben durch

$$\mathrm{BIC} = -2 \cdot l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) + \log(n)\, |M|,$$

vergleiche in Anhang B Seite 488. Das mit dem Faktor 1/2 multiplizierte BIC ist auch als Schwartz-Kriterium bekannt. Unter Normalverteilungsannahme erhalten wir

$$\mathrm{BIC} = n \cdot \log(\hat{\sigma}^2) + \log(n)\, |M|.$$

Das BIC besitzt damit eine ähnliche Gestalt wie das AIC. Man beachte jedoch, dass BIC und AIC auf sehr unterschiedliche Art und Weise motiviert sind. Der Hauptunterschied zum AIC aus praktischer Sicht ist, dass das BIC komplexe Modelle deutlich stärker bestraft als das AIC. Bei Verwendung des BIC resultieren also in aller Regel deutlich sparsamere Modelle.

Abb. 3.27. *AIC in Abhängigkeit vom Polynomgrad für die simulierten Daten aus Beispiel 3.24 (Seite 152).*

Beispiel 3.26 Polynomiale Regression – Modellwahl mit AIC

Abbildung 3.27 zeigt für die simulierten Daten aus Beispiel 3.24 (Seite 152) das AIC in Abhängigkeit vom Polynomgrad. Das AIC erreicht ein Minimum für $l = 2$. Damit erhalten wir ein vernünftiges Modell, wenn auch nicht den Polynomgrad des wahren Modells mit $l = 3$.

△

3.6.3 Praktische Verwendung der Modellwahlkriterien

Die genannten Modellwahlkriterien können verwendet werden, um aus einer Reihe zur Verfügung stehender Modelle die vielversprechendsten zu selektieren. Empfehlenswert ist die folgende Vorgehensweise:

- Durch substanzwissenschaftliche Vorüberlegungen bzw. durch Vorwissen aus vorangegangenen Analysen wird eine Vorauswahl potentieller Modelle getroffen. Die Modelle können sich in der Anzahl der Variablen aber auch in der Art der Modellierung (etwa linear versus nichtlinear) unterscheiden. Die Gesamtanzahl potentieller Modelle sollte dabei aber relativ klein sein.
- Die zur Verfügung stehenden Modelle können dann mit einem der oben genannten Modellwahlkriterien beurteilt werden. Bei der Beschreibung der Ergebnisse sollte man sich nicht auf das beste Modell beschränken. In der Regel haben mehrere Modelle annähernd das gleiche Gütemaß und unterscheiden sich auch nur in kleineren Teilaspekten des Modells. Hinsichtlich dieser Teilaspekte besteht dann Unsicherheit über die Schlussfolgerungen.

Diese Vorgehensweise ist leider nicht immer praktikabel, da in zahlreichen Anwendungen die Zahl der zur Verfügung stehenden Einflussvariablen und Modellierungsvarianten so groß ist, dass die Berechnung aller möglichen Modelle nicht ohne weiteres möglich ist. In diesem Fall finden die nachfolgend beschriebenen, zum Teil heuristischen Verfahren Anwendung:

- *Vollständige Modellselektion (All-Subset-Selection):* Falls die Anzahl der Kovariablen kleiner als 40 ist, können mit dem sogenannten „Leaps and bounds“ Algorithmus von Furnival & Wilson (1974) die im Sinne der Modellwahlkriterien besten Modelle bestimmt werden. Der Algorithmus vermeidet dabei die Berechnung aller Modelle. Eine Implementierung findet man im Programmpaket SAS und in R im Paket `leaps`.
- *Vorwärts-Selektion (Forward-Selection):* Ausgehend von einem Startmodell wird in der Vorwärts-Selektion in jedem Schritt des Verfahrens eine weitere Kovariable aufgenommen, und zwar diejenige, welche die größte Reduktion eines Modellwahlkriteriums (C_p, AIC, CV, BIC) liefert. Der Algorithmus ist beendet, wenn keine Reduktion mehr möglich ist.
- *Rückwärts-Selektion (Backward-Selection):* In der Rückwärts-Selektion wird mit dem vollen Modell, in dem alle Kovariablen enthalten sind, gestartet. Anschließend wird in jeder Iteration diejenige Kovariable aus dem Modell entfernt, welche die größte Reduktion eines Modellwahlkriteriums (C_p, AIC, CV, BIC) liefert. Der Algorithmus ist beendet, wenn keine Reduktion mehr möglich ist.
- *Schrittweise-Selektion (Stepwise-Selection):* Die Schrittweise-Selektion ist eine Kombination aus Vorwärts- und Rückwärts-Selektion. In jeder Iteration kann sowohl eine Variable aufgenommen werden, als auch eine Variable entfernt werden.

Die drei letztgenannten Verfahren Vorwärts-, Rückwärts- und Schrittweise-Selektion liefern im Allgemeinen nicht das im Sinne der Modellwahlkriterien beste Modell, jedoch in der Regel ein sehr gutes Modell.

Leider ist die zur Verfügung stehende Software zur Modellwahl häufig nicht gut durchdacht. Ein Hauptproblem ist, dass Dummy-Variablen mehrkategorialer Variablen in der Regel nicht *gemeinsam* in das Modell aufgenommen bzw. wieder aus dem Modell entfernt werden. Stattdessen kann es vorkommen, dass nur ein Teil der Dummy-Variablen einer kategorialen Variablen im Modell auftaucht, wodurch sich aber indirekt die Referenzkategorie und damit die Interpretation der Regressionskoeffizienten ändert. Darüber hinaus können weitere Schwierigkeiten auftreten. Meist existieren keine Verfahren zur Behandlung hierarchischer Terme. Verwendet man etwa Polynome zur Modellierung nichtlinearer Effekte, so kann es vorkommen, dass zwar der quadratische und kubische Anteil im Modell enthalten ist, jedoch nicht der lineare Anteil. Schließlich ist es meist auch nicht möglich bestimmte Terme von der Selektion auszuschließen und in jedem Fall ins Modell aufzunehmen.

Die genannten Verfahren sollten übrigens nicht mit einem von Efroymson in den 1960er Jahren vorgeschlagenen Algorithmus verwechselt werden, obwohl die prinzipielle Vorgehensweise ähnlich ist. Im Unterschied zu oben wird in Efroymsons Algorithmus diejenige Variable ins Modell aufgenommen bzw. entfernt, die den höchsten bzw. niedrigsten t-Wert besitzt. Das Verfahren wird beendet, wenn keine potentiell aufzunehmende Variable einen vorher festgelegten maximalen p-Wert (etwa 0.05) unterschreitet bzw. wenn keine zu entfernende Variable einen minimalen p-Wert (etwa 0.1) überschreitet. Dieses in nahezu allen größeren Softwarepaketen implementierte Verfahren gilt unter Statistikern (zu Recht) als obskur. Hierfür sind zwei Gründe verantwortlich:

- In der Regel liefern Vorwärts-, Rückwärts- und Schrittweise Selektion unterschiedliche Ergebnisse. Das passiert zwar auch unter Verwendung eines globalen Modellgütemaßes. Jedoch können die unterschiedlichen Selektionsergebnisse anhand des globalen Güte-

Modell	*kilstand* Grad 3	*alter* Grad 3	*sonderaus1* linear	*sonderaus2*	*tuev*	**AIC**
1	x	x				393.234 (1)
2	x	x	x			394.566 (2)
3	x	x		x		395.119 (4)
4	x	x			x	394.973 (3)
5	x	x	x	x		396.481 (6)
6	x	x	x		x	396.143 (5)
7	x	x		x	x	396.881 (7)
8	x	x	x	x	x	398.085 (8)

Tabelle 3.10. *Untersuchte Modelle. Die Werte in Klammern hinter dem AIC Wert geben den Rang des Modells an.*

maßes verglichen werden. Bei Efroymsons Algorithmus hingegen besteht keine Möglichkeit zwischen den unterschiedlichen Ergebnissen zu diskriminieren.

- Die wiederholte Verwendung der Teststatistik des t-Tests zum Test, ob ein Regressionskoeffizient von Null verschieden ist, suggeriert exakte Tests. Die Teststatistik des t-Tests ist in diesem Fall aber *nicht* t-verteilt unter der Nullhypothese. Der Grund hierfür ist, dass beim Auswahlprozess nicht eine beliebige Variable getestet wird, sondern diejenige mit dem *maximalen* t-Wert.

Beispiel 3.27 Preise von Gebrauchtwagen – Modellwahl

Wir illustrieren die mögliche Vorgehensweise bei der Wahl geeigneter Modelle anhand von Daten zum Verkaufspreis gebrauchter Golf-Modelle. Ziel ist die Modellierung des Zusammenhangs zwischen dem Verkaufspreis in 1000 Euro (Variable *preis*) und den fünf erklärenden Variablen „Alter des Autos in Monaten“ (*alter*), „Kilometerleistung in 1000 Km“ (*kilstand*), „Anzahl der Monate bis zum nächsten TÜV Termin“ (*tuev*), „ABS vorhanden ja/nein“ (*sonderaus1*) und „Schiebedach vorhanden ja/nein“ (*sonderaus2*). Untersuchungen dieser Art spielen bei der Ermittlung sogenannter hedonischer Preisindizes eine Rolle. Hedonische Preisindizes berücksichtigen im Gegensatz zu den gängigen Preisindizes Qualitätsunterschiede eines Produktes. Dies ist vor allem bei Produktgruppen, die einem raschen technologischen Wandel unterworfen sind, wichtig.

Abbildung 3.28 zeigt erste grafische Zusammenhangsanalysen zwischen Zielgröße und erklärenden Variablen. Die Grafiken lassen folgende Effekte vermuten: Bei den Variablen *alter* und *kilstand* ist ein linearer oder monoton fallender nichtlinearer Effekt zu vermuten. Hier bietet sich eine Modellierung durch (orthogonale) Polynome maximal 3. Grades an. Die Variable *tuev* scheint keinen oder höchstens einen schwach linearen Einfluss auf den durchschnittlichen Verkaufspreis zu besitzen. Eine lineare Modellierung ist hier ausreichend. Autos mit ABS (*sonderaus1*) scheinen etwas teuerer als ohne ABS. Der Effekt ist jedoch unsicher. Keine Unterschiede im durchschnittlichen Verkaufspreis sind für Modelle mit und ohne Schiebedach (*sonderaus2*) feststellbar. Insgesamt scheint ein Zusammenhang mit dem Alter und dem Kilometerstand relativ sicher. Bei den anderen Variablen ist ein Effekt jeweils fraglich.

Basierend auf diesen Überlegungen untersuchen wir zunächst acht Regressionsmodelle (vgl. Tabelle 3.10), die sich in der Modellierung der Variablen *alter* und *kilstand* nicht unterscheiden, während für die drei restlichen Variablen alle möglichen Modellierungs-

Abb. 3.28. *Streudiagramme zwischen dem Verkaufspreis und den metrischen erklärenden Variablen Alter, Kilometerstand und Monate bis zum nächsten TÜV. Boxplots für den Preis geschichtet nach den Werten der binären Variablen sonderaus1 und sonderaus2.*

varianten getestet werden. Verwenden wir als Gütekriterium das AIC, so erhalten wir das erste in Tabelle 3.10 angegebene Modell als vorläufig bestes Modell. Abbildung 3.29 visualisiert die erreichten AIC Werte für die acht Modelle.

Da wir lediglich fünf erklärende Variablen zur Verfügung haben können wir mit Hilfe des „Leaps and Bounds“ Algorithmus auch das AIC beste Modell unter allen möglichen Modellen bestimmen. Dieses Modell erreicht einen AIC Wert von 389.085 und unterscheidet sich vom bisherigen besten Modell lediglich dadurch,

Abb. 3.29. *AIC-Werte für die untersuchten Modelle.*

Abb. 3.30. Modell basierend auf vollständiger Selektion: Effekt des Alters und des Kilometerstands inklusive partieller Residuen.

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert	95% Konfidenz-intervall	
Konstante	3.397	0.056	60.220	<0.001	3.285	3.508
alterop1	-0.705	0.061	-11.470	<0.001	-0.826	-0.584
alterop2	0.187	0.057	3.270	0.001	0.074	0.300
kilstandop1	-0.439	0.061	-7.170	<0.001	-0.560	-0.318
kilstandop2	0.141	0.057	2.460	0.015	0.028	0.254

Tabelle 3.11. *Schätzergebnisse für das gemäß AIC beste Modell.*

dass für die Variablen *alter* und *kilstand* Polynome vom Grad 2 (und nicht vom Grad 3) verwendet werden. Abbildung 3.29 zeigt, dass der erzielte AIC Wert für dieses 9. Modell deutlich kleiner ist als der beste AIC Wert für die bisher untersuchten Modelle. Die Schätzergebnisse für Modell 9 finden sich in Tabelle 3.11. Abbildung 3.30 visualisiert zusätzlich die geschätzten Effekte des Alters und des Kilometerstands. Tabelle 3.11 zeigt, dass sämtliche Regressionskoeffizi-

enten bei einem Niveau von 5% signifikant von Null verschieden sind, jedoch nicht auf dem 1% Niveau. Grundsätzlich gilt: Die mit den hier besprochenen globalen Gütekriterien selektierten Modelle können durchaus nicht signifikante Koeffizienten enthalten. Das ist nur auf den ersten Blick verwunderlich. Bei sämtlichen Gütekriterien handelt es sich um globale Maße, die beispielsweise den Prognosefehler minimieren. Statistische Tests können eher als lokale Maße angesehen werden. Darüber hinaus sind Tests unsymmetrisch konstruiert, d.h. der Fehler 1. Art wird kontrolliert und klein gehalten, so dass die Nullhypothese im Vergleich zur Alternativhypothese bevorzugt wird.

△

3.6.4 Modelldiagnose

Nach erfolgter erster Schätzung sollte das gefundene Modell hinsichtlich seiner Tauglichkeit untersucht und gegebenenfalls durch besser geeignete ersetzt werden. Grundsätzlich werden mit der Modelldiagnose die folgenden Ziele verfolgt:

- *Überprüfung der Modellannahmen:* Kein Modell ist korrekt, jedoch sollten die hinter einem Modell stehenden Annahmen zumindest approximativ erfüllt sein, um Fehlschlüsse zu vermeiden. Im klassischen linearen Modell müssen die Annahmen homoskedastischer, unkorrelierter und eventuell normalverteilter Fehler sowie die Linearität des Prädiktors überprüft werden.
- *Auffinden ungewöhnlicher Beobachtungen:* Gelegentlich werden die erzielten Ergebnisse stark von einigen wenigen Beobachtungen beeinflusst. Diese Beobachtungen gilt es aufzuspüren und gegebenenfalls Auswirkungen auf die Schlussfolgerungen zu untersuchen.
- *Kollinearitätsanalyse:* Wir haben bereits durch theoretische Überlegungen gesehen, dass untereinander stark korrelierte Kovariablen einen erheblichen Einfluss auf die Schätzgenauigkeit besitzen können, vergleiche die Seiten 101ff. bzw. den Kasten auf Seite 104. Wichtig ist also eine Untersuchung der Korrelationsstruktur der verwendeten Kovariablen.

In den folgenden Abschnitten beschreiben wir einige Hilfsmittel zur Modelldiagnose, mit denen die genannten Fragestellungen untersucht werden können. Teilweise handelt es sich lediglich um eine kurze Zusammenfassung bereits vorher behandelter Konzepte.

Überprüfen der Modellannahmen

Homoskedastizität: Die Überprüfung der Annahme homoskedastischer Varianzen wurde bereits in Abschnitt 3.4.3 (Seite 128) behandelt. Wesentliche Hilfsmittel sind Residuenplots und Tests auf Heteroskedastizität. Maßnahmen bei vorhandener Heteroskedastizität sind entweder Transformationen der Zielgröße oder die Verwendung gewichteter Regression.

Unkorreliertheit der Störgrößen: Korrelierte Störgrößen lassen sich durch das Abtragen der Residuen im Zeitablauf oder durch statistische Tests entdecken, vergleiche hierzu Abschnitt 3.4.4 (Seite 136). Da korrelierte Fehler ein Hinweis auf Fehlspezifikation (insbesondere vergessene erklärende Variablen und unbeachtete Nichtlinearität) sind, sollte

Abb. 3.31. *Studentisierte Residuen in Abhängigkeit von* $\widehat{preis}$ *(links) und den metrischen Kovariablen. Die horizontalen Linien geben die Kritischen Grenzen für den in Abschnitt 3.6.4 (Seite 173) beschriebenen Ausreißertest.*

zunächst untersucht werden, ob die Modellspezifikation verbessert werden kann. Andernfalls kann eines der in Abschnitt 3.4.4 (Seite 136) diskutierten Schätzverfahren herangezogen werden.

Linearität: Eventuell übersehene nichtlineare Kovariableneinflüsse können durch ein Streudiagramm der (standardisierten bzw. studentisierten) Residuen gegen die geschätzten Werte entdeckt werden. Zusätzlich können die partiellen Residuen $\hat{\varepsilon}_{x_j,i} = \hat{\varepsilon}_i + \hat{\beta}_j x_{ij}$ gegen die Kovariablen x_{ij} abgetragen werden.

Normalverteilungsannahme: Verteilungsannahmen können generell mit Quantil-Quantil Diagrammen (Q-Q-Plots) grafisch überprüft werden. Bei Q-Q Diagrammen werden die empirischen Quantile gegen die theoretischen Quantile der Verteilung abgetragen. Entsprechen die Daten der Verteilung, so liegen die Punkte annähernd auf einer Geraden mit 45 Grad Steigung.

Beispiel 3.28 Preise von Gebrauchtwagen – Modelldiagnose

Abbildung 3.31 zeigt für das in Beispiel 3.27 geschätzte Modell die studentisierten Residuen in Abhängigkeit vom geschätzten Preis und den metrischen Kovariablen. Wir entdecken keine zusätzlichen nichtlinearen Effekte. Jedoch bestehen Hinweise auf (schwach) heteroskedastische Fehler. Die studentisierten Residuen streuen mit steigendem Kilometerstand mit geringer werdender Variabilität um Null. Auffällig ist auch die stärker

Abb. 3.32. *Q-Q Plot der standardisierten Residuen.*

werdende Streuung der Residuen für Golf Modelle mit 18 Monaten und mehr bis zum nächsten TÜV Termin. Hier scheint die Variable *tuev* zwar keinen Einfluss auf den bedingten Erwartungswert des Verkaufspreises zu besitzen, jedoch einen Einfluss auf die bedingte Varianz.

Abbildung 3.32 zeigt den Q-Q-Plot für die standardisierten Residuen. Größere Abweichungen von der Normalverteilung zeigen sich allenfalls am linken und rechten Rand.

△

Kollinearitätsanalyse

Wir illustrieren die mit der Kollinearitätsanalyse verbundenen Ziele zunächst anhand eines extremen Falls. Wir betrachten ein Modell

$$y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \varepsilon$$

mit drei Kovariablen x_1, x_2 und x_3, wobei zwischen x_1 und x_2 perfekte lineare Abhängigkeit $x_1 = a\, x_2$ herrsche. Beispielsweise könnte es sich bei x_1 um das Gewicht in Kilogramm und bei x_2 um das Gewicht in Tonnen handeln, d.h. x_1 und x_2 unterscheiden sich nur durch die Maßeinheit. In diesem extremen Fall sind die zu x_1 und x_2 gehörenden Regressionskoeffizienten β_1 und β_2 nicht eindeutig schätzbar, denn

$$y = \beta_0 + \beta_1 x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \varepsilon = \beta_0 + (\beta_1\, a + \beta_2)\, x_2 + \beta_3\, x_3 + \varepsilon.$$

In der Praxis sind die Kovariablen in der Regel nicht exakt linear abhängig, sondern hoch korreliert. Wir sprechen von Kollinearität. Wie wir in Abschnitt 3.2.3 auf Seite 101 gesehen haben, verursachen hoch korrelierte Kovariablen ungenaue Schätzungen der Regressionsparameter. Man vergleiche hierzu nochmal die Formel

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^n (x_{ij} - \overline{x}_j)^2},$$

für die Varianz von $\hat{\beta}_j$. Je größer die lineare Abhängigkeit einer Kovariablen x_j mit den übrigen erklärenden Variablen (gemessen durch das Bestimmtheitsmaß R_j^2), desto größer ist die Varianz $\mathrm{Var}(\hat{\beta}_j)$. Im Extremfall $R_j^2 \to 1$ explodiert die Varianz.

Die Formel für die Varianz liefert eine Kenngröße zur Messung des Ausmaßes der Multikollinearität, den sogenannten Varianzinflationsfaktor:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

Der Varianzinflationsfaktor gibt an, um welchen Faktor die Varianz von $\hat{\beta}_j$ durch die lineare Abhängigkeit vergrößert wird. Je größer die Korrelation zwischen x_j und den übrigen Kovariablen, desto größer R_j^2 und damit auch VIF_j. Als Faustregel für ein ernsthaftes Kollinearitätsproblem gilt ein Varianzinflationsfaktor größer als 10, d.h. $\mathrm{VIF}_j > 10$.

Beispiel 3.29 Preise von Gebrauchtwagen

Für das in Beispiel 3.27 erhaltene Regressionsmodell schwanken die Varianzinflationsfaktoren zwischen 1.03 und 1.19. Die Varianz wird also maximal um den Faktor 1.19 aufgebläht. Ein ernsthaftes Kollinearitätsproblem liegt somit nicht vor.

△

In der Literatur existieren eine Vielzahl von Vorschlägen zur Begegnung starker Multikollinearität:

- *Weglassen einiger betroffener Kovariablen:* Dieses Vorgehen wird in der Praxis wahrscheinlich am häufigsten verwendet und in der Literatur ebenso häufig kritisiert. In vielen Fällen existieren jedoch verschiedene Konstrukte zur Messung ein und des selben Sachverhalts. Als Beispiel betrachte man die verschiedenen Variablen zur Messung des Ernährungszustands der Mutter im Beispiel zur Unterernährung in Sambia. Im Originaldatensatz existiert neben dem Body Mass Index auch der sogenannte Rohrer Index als weiteres Ernährungsmaß. Beide Variablen sind hochkorreliert mit einem Korrelationskoeffizienten von über 0.95. In diesem und ähnlichen Fällen bleibt gar nichts anderes übrig als sich für ein Konstrukt zu entscheiden und ähnliche Maßzahlen von der weiteren Analyse auszuschließen.
- *Bilde aus den betroffenen Variablen eine gemeinsame (gut interpretierbare) Variable:* In Frage kommen hier insbesondere Linearkombinationen sowie Minimum oder Maximum der Variablen. Wir verdeutlichen die Idee anhand der Supermarkt-Scanner Daten zur Untersuchung verschiedener Verkaufsförderungsmaßnahmen auf den Absatz einer Kaffeemarke, vergleiche Beispiel 3.2 (Seite 72). Neben dem eigenen Preis der Kaffeemarke sind die Preise von Konkurrenzmarken die wichtigsten Kovariablen zur Modellierung des durchschnittlichen Absatzes. Für einige Konkurrenzprodukte liegen mehrere Marken vor, die sich nur durch die Packungsgröße unterscheiden. Die (auf die Packungsgröße normierten) Preise für diese Marken werden fast immer gleichzeitig und gleichgerichtet verändert, so dass bei separater Aufnahme der Preise dieser Konkurrenzmarken ins Modell hohe Kollinearität resultiert. In diesem Fall kann das Problem umgangen werden, indem beispielsweise der Durchschnittspreis (also eine spezielle Linearkombination) der Konkurrenzmarken gebildet und anstelle der Einzelmarken ins Modell aufgenommen wird. Alternativ ist auch die Verwendung des kleinsten Preises (Minimum) denkbar.
- *Ridge-Regression:* Der Ridge-Schätzer kann als Alternative zum KQ-Schätzer angesehen werden und ist definiert als

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

wobei $\lambda \geq 0$ ein geeignet zu wählender Tuning-Parameter ist. Offenbar ist der Ridge-Schätzer im Gegensatz zum KQ-Schätzer nicht erwartungstreu. Der Ridge-Schätzer wird in der Literatur vor allem bei Problemen mit Multikollinearität vorgeschlagen, da durch die Addition von $\lambda \boldsymbol{I}$ zu $\boldsymbol{X}'\boldsymbol{X}$ eine reguläre und damit invertierbare Matrix resultiert, auch wenn durch hohe Kollinearität $\boldsymbol{X}'\boldsymbol{X}$ fast singulär ist. Häufig besitzen daher die Komponenten des Ridge-Schätzers bei Kollinearität einen kleineren MSE als die entsprechenden Komponenten des KQ-Schätzers. Man beachte, dass der KQ-Schätzer nur in der Klasse der linearen erwartungstreuen Schätzer optimal ist. Lässt man auch verzerrte Schätzer zu, können andere Schätzer durchaus besser im Sinne des MSE sein als der KQ-Schätzer. Trotz intensiver Forschung verbunden mit einer Fülle an Publikationen wird der Ridge-Schätzer in der Praxis kaum verwendet. Ein Grund ist sicherlich, dass eine Trennung der Einflüsse der Kovariablen durch die hohe Kollinearität auch durch die Verwendung des Ridge-Schätzers nicht möglich ist. Schätzer mit ähnlicher Struktur wie der Ridge-Schätzer finden aber breite Anwendung in der nichtparametrischen Regression (vergleiche Kapitel 7) und bilden die Basis für Variablenselektionsverfahren, die für hochdimensionale Regressionsmodelle mit sehr großer Zahl von Einflussgrößen entwickelt wurden (vergleiche Abschnitt 3.7 und beispielsweise Bühlmann (2006)).

- *Hauptkomponentenregression:* Bei dieser Variante werden die zur Verfügung stehenden Kovariablen nicht direkt ins Modell aufgenommen, sondern geeignete Linearkombinationen der Kovariablen. Dabei geht man von metrischen Kovariablen aus. Grundlage des Verfahrens ist eine Hauptkomponenten- bzw. Spektralzerlegung der Matrix $\boldsymbol{X}'\boldsymbol{X}$ (vergleiche Satz A.22 (Seite 455)):

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{P}\mathrm{diag}(\lambda_1, \ldots, \lambda_p)\boldsymbol{P}'$$

Die Matrix $\boldsymbol{P}$ enthält die Eigenvektoren $\boldsymbol{p}_j$ bzw. Hauptkomponenten von $\boldsymbol{X}'\boldsymbol{X}$. Die erste Hauptkomponente $\boldsymbol{p}_1$ besitzt die Eigenschaft, dass die Linearkombination $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{p}_1$ die größte Varianz unter allen Linearkombinationen der Spalten von $\boldsymbol{X}$ besitzt. Für die zweite Hauptkomponente $\boldsymbol{p}_2$ hat die Linearkombination $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{p}_2$ die nächstgrößte Varianz unter allen Linearkombinationen, die orthogonal zu $\boldsymbol{z}_1$ sind. Entsprechend haben die Linearkombinationen $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{p}_j$ der j-ten Hauptkomponente höchste Varianz unter allen zu den ersten $j-1$ Hauptkomponenten orthogonalen Linearkombinationen. In der Hauptkomponentenregression verwendet man anstelle der Original-Einflussvariablen die abgeleiteten Linearkombinationen $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$ als Kovariablen, wobei nur die ersten $q \leq p$ der p Linearkombinationen berücksichtigt werden, während die restlichen entweder wegen fehlender Interpretierbarkeit oder wegen sehr geringer Varianz vernachlässigt werden.

Eine Anwendung der Hauptkomponentenregression ergibt sich im Zusammenhang mit der Modellierung der Determinanten von Unterernährung. Hier stehen häufig eine Reihe von Variablen zur Verfügung, die alle die ökonomische Situation des Haushalts beschreiben. Meist ist es nicht sinnvoll alle Kovariablen ins Modell aufzunehmen. Stattdessen werden basierend auf einer Hauptkomponentenzerlegung gut interpretierbare Linearkombinationen als Maße für den ökonomischen Status verwendet. Fast immer beschränkt man sich dabei auf die Linearkombination mit der größten Varianz basierend auf der ersten Hauptkomponente.

Ausreißer- und Einflussanalyse

Neben der Überprüfung der Modellannahmen und der Kollinearitätsanalyse besteht eine weitere wichtige Aufgabe der Modelldiagnose darin, den Einfluss einzelner Beobachtungen auf die Schätzergebnisse zu untersuchen. Einen besonders hohen Einfluss haben häufig sogenannte Ausreißer, die wir im Folgenden näher beleuchten wollen.

Ausreißer

Für den Begriff des Ausreißers gibt es keine exakte Definition. Als Ausreißer wollen wir hier Beobachtungen bezeichnen, die nicht dem angepassten Modell gehorchen. Speziell bezeichnen wir die i-te Beobachtung als Ausreißer, wenn der (bedingte) Erwartungswert von y_i nicht $\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ ist, wie vom Modell postuliert, sondern um den Betrag d verschoben ist, d.h. $\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta} + d$.

Abbildung 3.33 illustriert anhand von simulierten Datensätzen die Ausreißerproblematik. Abbildung a) zeigt simulierte Beobachtungen (y_i, x_i), die einen linearen Zusammenhang aufweisen. Eine geschätzte Regressionsgerade ist ebenfalls eingezeichnet (durchgezogene Linie). Offenbar handelt es sich bei dem ausgefüllten Punkt im Zentrum des Wertebereichs von x um einen Ausreißer. Jedoch verändert der Punkt die Regressionsgerade kaum. Schätzen wir nämlich die Gerade erneut ohne Berücksichtigung des Ausreißers, so ändert sich die Schätzung (gestrichelte Linie) nur unwesentlich. Die geschätzten Standardfehler und damit auch die Konfidenzintervalle ändern sich aber erheblich. Unter Verwendung aller Beobachtungen erhält man 0.241 als geschätzten Standardfehler für $\hat{\beta}_1$, ohne Ausreißer hingegen den Wert 0.069. Entsprechend ergibt sich für β_1 das Konfidenzintervall $[1.53, 2.50]$ basierend auf allen Beobachtungen und $[1.84, 2.11]$ ohne Berücksichtigung des Ausreißers. Obwohl in diesem Fall die geschätzte Regressionsgerade vom Ausreißer wenig verändert wird, können die Inferenzschlüsse erheblich beeinflusst werden. Abbildung 3.33 c) zeigt im Vergleich zu Abbildung a) eine leicht veränderte Situation. Hier befindet sich der Ausreißer nicht im Zentrum des Wertebereichs, sondern am Rand. In diesem Fall wirkt sich der Ausreißer nicht nur auf die geschätzten Standardfehler aus, sondern auch auf die geschätzte Regressionsgerade, die erheblich durch den Ausreißer beeinflusst ist. Noch gravierender sind die Auswirkungen auf die geschätzte Regressionsgerade wenn mehrere (sogenannte multiple) Ausreißer auftreten (Abbildung e).

Wir stellen also fest: Ausreißer können die Schätzungen und die gezogenen Inferenzschlüsse erheblich beeinflussen. Der Einfluss ist tendenziell größer im Randbereich des Wertebereichs und bei multiplen Ausreißern. Damit stellen sich zwei Fragen: Erstens, wie können Ausreißer entdeckt werden? Zweitens, welche Maßnahmen sind zu ergreifen wenn Ausreißer gefunden wurden? In der Literatur existieren eine Vielzahl von Vorschlägen zu beiden Fragen. Wir überlegen uns zunächst, wie wir Ausreißer finden können:

Es liegt nahe, zur Entdeckung von Ausreißern nach großen Residuen zu suchen. Die Entdeckung von Ausreißern wird jedoch dadurch erschwert, dass diese einen zum Teil erheblichen Einfluss auf die Schätzungen besitzen können. Vor allem am Rand des Wertebereichs der Kovariablen werden die Schätzungen in Richtung Ausreißer „gezogen“ , so dass die dazu gehörigen Residuen relativ klein sein können. Wenn wir also untersuchen wollen, ob die i-te Beobachtung ein Ausreißer ist, sollten wir diese bei der Schätzung unberücksichtigt lassen.

Abb. 3.33. *Illustration zur Ausreißerproblematik: Die jeweils linken Grafiken zeigen Streudiagramme zwischen y und x inklusive geschätzter Regressionsgerade (gestrichelte Linie) und geschätzter Regressionsgerade bei nicht Berücksichtigung der Ausreißer (durchgezogene Linie). Die jeweils rechten Grafiken zeigen die studentisierten Residuen in Abhängigkeit von x. Die Ausreißer sind durch ausgefüllte Punkte gekennzeichnet.*

In Abschnitt 3.2.3 auf Seite 108ff. haben wir die „leave one out“ Residuen

$$\hat{\varepsilon}_{(i)} = y_i - \hat{y}_{(i)} = y_i - \boldsymbol{x}_i'(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}'\boldsymbol{y}_{(i)},$$

definiert. Diese Residuen beruhen auf Schätzungen, bei denen die i-te Beobachtung unberücksichtigt ist. Durch Standardisieren haben wir schließlich die studentisierten Resi-

duen

$$r_i^* = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}(1 + \boldsymbol{x}_i'(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}}$$

erhalten. Diese sind bei korrekt spezifiziertem Modell t-verteilt mit $n - p - 1$ Freiheitsgraden. Als Ausreißerkandidat kommen somit alle Beobachtungen mit großem studentisiertem Residuum in Frage. Da die Verteilung der studentisierten Residuen bei korrekt spezifiziertem Modell bekannt ist, können wir formal testen, ob die i-te Beobachtung y_i ein Ausreißer ist. Dazu müssen wir bei gegebenem Signifikanzniveau α das i-te studentisierte Residuum r_i^* mit dem $\alpha/2$ bzw. $1 - \alpha/2$ Quantil der t-Verteilung mit $n - p - 1$ Freiheitsgraden vergleichen. Ist das Residuum kleiner als das $\alpha/2$ bzw. größer als das $1 - \alpha/2$-Quantil so handelt es sich um einen Ausreißerkandidaten.

In den meisten Fällen wollen wir nicht nur eine Beobachtung, sondern alle Beobachtungen *simultan* testen. Dann müssen wir für jede Beobachtung das Signifikanzniveau anpassen, um als Gesamtsignifikanzniveau den Wert α zu erhalten. Am einfachsten ist die Verwendung der Bonferroni-Korrektur, bei der wir in jedem Test das Niveau α/n verwenden. Die Bonferroni-Korrektur ist relativ konservativ, d.h. es werden (insbesondere bei großem Stichprobenumfang) tendenziell zu wenige Ausreißer entdeckt.

Die Abbildungen 3.33 b), d) und f) zeigen für die simulierten Datensätze die studentisierten Residuen in Abhängigkeit von x. Die jeweils eingezeichneten horizontalen Geraden markieren die 0.5 und 99.5 Prozent Quantile der t-Verteilung mit $n-p-1 = 40-2-1 = 37$ Freiheitsgraden. Beobachtungen unter- bzw. oberhalb der beiden Geraden sind potentielle Ausreißerkandidaten. Offenbar werden die beiden einzelnen Ausreißer im Zentrum und am Rand des Wertebereichs durch den Test erkannt. Die multiplen Ausreißer in Abbildung e) werden jedoch nicht erkannt. Generell kann es relativ schwierig sein multiple Ausreißer zu entdecken. Im vorliegenden Fall hilft uns die grafische Darstellung der studentisierten Residuen in Abbildung f). Bereits ein flüchtiger Blick auf die Grafik genügt um die unüblichen Beobachtungen zu erkennen.

Beispiel 3.30 Preise von Gebrauchtwagen – Ausreißer

Für das in Beispiel 3.27 (Seite 165) erhaltene Regressionsmodell zeigt Abbildung 3.31 die studentisierten Residuen in Abhängigkeit von den geschätzten Werten bzw. den Kovariablen. Zusätzlich eingezeichnet sind die 0.5 und 99.5 Prozent Quantile der t-Verteilung mit $n - p - 1 = 172 - 6 - 1 = 165$ Freiheitsgraden (horizontale Linien). Basierend auf einem Gesamtniveau von $\alpha = 0.01$ ergibt sich *eine* Beobachtung oberhalb der Geraden mit studentisiertem Residuum $r_i^* = 4.36$.

△

In der Praxis werden Ausreißerkandidaten häufig ohne weiteren Kommentar von der Analyse ausgeschlossen. Dieses Vorgehen ist meist nicht die beste Alternative. Stattdessen empfiehlt sich folgendes Vorgehen:

- *Ausschluss von Datenfehlern:* Ausreißer können einen Hinweis auf Datenfehler, z.B. Fehlkodierungen, sein. Diese Fehlerquelle sollte also zunächst ausgeschlossen werden.
- *Suche nach Erklärung für Ausreißer:* In manchen Anwendungen sind die Ausreißer des Modells die interessantesten Beobachtungen. Bekannt wurde der Fall des Satelliten Nimbus 7, der von der NASA betrieben wurde und Informationen zur Atmosphäre liefern sollte. Im Jahr 1985, nachdem der Satellit bereits mehrere Jahre in Betrieb war, beobachtete man einen extrem starken Abfall der Ozonkonzentration in der Atmosphäre,

Abb. 3.34. *Illustration zur robusten Regression: Geschätzte Regressionsgeraden basierend auf verschiedenen Schätztechniken (KQ-Schätzer mit allen Beobachtungen, KQ-Schätzer ohne Berücksichtigung der Ausreißer und Median-Regression). Links: Ausreißer am Rand des Wertebereichs. Rechts: Multiple Ausreißer am Rand des Wertebereichs.*

das Ozonloch war entdeckt. In diesem Fall gab es für die unüblichen Beobachtungen also eine Erklärung, die sogar von großer Bedeutung war.

- *Beschreibung der Unterschiede in den Ergebnissen:* Statt Ausreißer von der Analyse endgültig auszuschließen, kann das Modell auch mit und ohne Ausreißer geschätzt werden. Anschließend können die entstandenen Modellunterschiede beschrieben werden. Auf diese Art und Weise geht die Information über die Ausreißer nicht verloren.
- *Robuste Regression:* Seit den 1960er Jahren spielen robuste Verfahren eine immer wichtigere Rolle in der Statistik. Diese Verfahren sind weniger empfindlich gegenüber Ausreißern, d.h. die Schätzergebnisse werden durch einzelne Ausreißer nicht oder nur unwesentlich beeinflusst. Das älteste bereits von Boscovich um 1760 vorgeschlagene robuste Verfahren ist die Median-Regression (vergleiche auch Abschnitt 3.2.1), bei der anstelle der Summe der quadrierten Abweichungen die Summe der absoluten Abweichungen minimiert wird. Die Median-Regression ist im Vergleich zur KQ-Methode deutlich robuster bezüglich Ausreißern. Zur Illustration betrachte man Abbildung 3.34, die bei Vorliegen eines oder mehrerer Ausreißer am Rand des Wertebereichs verschiedene Schätzverfahren vergleicht. Die auf der KQ-Methode basierenden Ergebnisse sind nur dann unbeeinflusst von den Ausreißern, wenn diese bei der Schätzung unberücksichtigt bleiben. Werden stattdessen alle Beobachtungen bei der Schätzung verwendet, so werden die geschätzten Regressionsgeraden in Richtung Ausreißer „gezogen“. Die Median-Regression hingegen ist relativ unbeeinflusst von den Ausreißern, auch wenn diese bei der Schätzung nicht ausgeschlossen werden. Die Median-Regression besitzt also den Vorteil, dass eventuell vorhandene Ausreißer nicht mühsam entdeckt werden müssen, da das Verfahren unempfindlich gegenüber einzelnen Ausreißern ist. Problematisch wird die Verwendung der Median-Regression wenn verhältnismäßig viele Ausreißer vorliegen. Dann müssen weiter spezialisierte Verfahren verwendet werden, die aber den Rahmen dieses Buches sprengen würden, vergleiche hierzu etwa Rousseeuw & Leroy (2003).

  Es stellt sich natürlich die Frage, warum nach wie vor die KQ-Methode häufiger Verwendung findet als die deutlich Ausreißer unempfindlichere Median-Regression. Die Antwort ist, dass der KQ-Schätzer deutlich überlegene Schätzeigenschaften besitzt,

Abb. 3.35. *Die linke Grafik zeigt das Streudiagramm zwischen y und x inklusive geschätzter Regressionsgerade (gestrichelte Linie) und geschätzter Regressionsgerade wenn die einflussreiche Beobachtung nicht berücksichtigt wird (durchgezogene Linie). Die rechte Grafik zeigt die Hebelwerte in Abhängigkeit von x.*

wenn *keine* Ausreißer vorliegen. Bei normalverteilten Daten benötigt die Median-Regression circa 50% mehr Daten um die gleiche Schätzgenauigkeit wie die KQ-Methode zu erzielen. Es empfiehlt sich jedoch in jedem Fall bei der Analyse auch robuste Verfahren (soweit verfügbar) anzuwenden, um – im Zuge der Modelldiagnose – zu untersuchen wie robust die erzielten Schätzergebnisse sind.

Einflussanalyse

Die Einflussanalyse beschäftigt sich mit dem Einfluss, den einzelne Beobachtungen auf die Schätzergebnisse besitzen. Ziel ist insbesondere Beobachtungen, die einen großen Einfluss auf $\hat{\boldsymbol{\beta}}$ bzw. $\hat{\boldsymbol{y}}$ haben, ausfindig zu machen. Werden diese Beobachtungen nicht berücksichtigt, ändern sich die Schätzungen stark. In vielen Fällen handelt es sich bei diesen einflussreichen Beobachtungen zugleich um Ausreißer. In extremen Fällen wird der geschätzte Zusammenhang von einer einzelnen Beobachtung getragen. Wird die Beobachtung entfernt lässt sich kein Zusammenhang mehr feststellen. Ein Fall dieser Art ist in Abbildung 3.35 (linke Grafik) anhand von simulierten Daten illustriert.

In der Literatur existieren eine Vielzahl von Vorschlägen zur Entdeckung von besonders einflussreichen Beobachtungen. Unter anderem werden die folgenden Größen genannt:

- *Hebelwerte:* Die Diagonalelemente h_{ii} der Prädiktionsmatrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ heißen Hebelwerte (leverages). Die Hebelwerte schwanken zwischen $1/n$ und Eins, vergleiche den Kasten auf Seite 93. Ein großer Hebelwert nahe bei Eins bedeutet zweierlei. Zunächst bewirkt ein großer Hebelwert, dass $\mathrm{Var}(\varepsilon_i) = \sigma^2(1 - h_{ii})$ klein ist. Ein Hebelwert nahe bei Eins bedeutet sogar, dass die Varianz nahe Null ist. Im Fall von nur einer Kovariable geht damit die Regressionsgerade praktisch sicher durch den Punkt (y_i, x_i), unabhängig davon wie die restlichen Beobachtungen liegen. Damit besitzt diese Beobachtung einen erheblichen Einfluss auf die Schätzung. Diese Überlegungen können ohne weiteres auf den Fall mehrerer Kovariablen übertragen werden. Auch hier geht die Regressionshyperebene im Extremfall fast sicher durch den Punkt $(y_i, \boldsymbol{x}_i)$.

  Als zweite wichtige Eigenschaft weisen große Hebelwerte auf unübliche Kovariablenwerte $\boldsymbol{x}_i$ hin. Im Falle von nur einer Kovariable x_i sieht man diese Eigenschaft wieder

besonders gut, es gilt nämlich

$$h_{ii} = \frac{1}{n} + \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}.$$

Der Hebelwert h_{ii} wird also umso größer, je weiter entfernt die Beobachtung x_i vom Durchschnitt $\bar{x}$ ist. In diesem Sinne können Hebelwerte auch als Ausreißer in x-Richtung angesehen werden. Allgemein gilt

$$h_{ii} = \frac{1}{n} + (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$$

mit der Designmatrix der zentrierten Kovariablen

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} (x_{11} - \bar{x}_1) & \cdots & (x_{1k} - \bar{x}_k) \\ (x_{21} - \bar{x}_1) & \cdots & (x_{2k} - \bar{x}_k) \\ \vdots & \cdots & \vdots \\ (x_{n1} - \bar{x}_1) & \cdots & (x_{nk} - \bar{x}_k) \end{pmatrix}.$$

Idealerweise sollten die Hebelwerte h_{ii} möglichst gleich verteilt sein. Als grobe Faustregel sollten Beobachtungen mit $h_{ii} > 2p/n$ genauer untersucht werden. Dabei ist zu beachten, dass große Hebelwerte nicht zwangsläufig zu Problemen führen.

- *Cook-Distanz:* Bezeichne wieder $\hat{\boldsymbol{y}}_{(i)}$ eine Schätzung, die auf allen Beobachtungen bis auf der i-ten beruht. Eine nahe liegendes Maß für den Unterschied zwischen $\hat{\boldsymbol{y}}$ basierend auf allen Beobachtungen und $\hat{\boldsymbol{y}}_{(i)}$ ist der euklidische Abstand zwischen den beiden Schätzungen. Exakt ist die Cook-Distanz definiert als

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

d.h. es wird noch mit der geschätzten Varianz $\hat{\sigma}^2$ gewichtet. Als grobe Faustregel gelten Beobachtungen mit $D_i > 0.5$ als auffällig, Beobachtungen mit $D_i > 1$ sollten in jedem Fall untersucht werden.

Für die simulierten Daten in Abbildung 3.35 zeigt die rechte Grafik die Hebelwerte h_{ii} in Abhängigkeit von x_i. Die horizontale Linie markiert die Faustregel $h_{ii} > 2p/n = 2 \cdot 2/40 = 0.1$ für einflussreiche Beobachtungen. Die höchst einflussreiche Beobachtung bei $x = 12$ hat mit $h_{ii} = 0.6$ als einzige einen Hebelwert größer als die Faustregel. Auch die Cook-Distanz für diese Beobachtung ist mit $D_i = 2.55$ außerordentlich groß, so dass auch diese Kennzahl den erheblichen Einfluss der Beobachtung markiert. Auch handelt es sich um die einzige Beobachtung mit $D_i > 0.5$.

Beispiel 3.31 Preise von Gebrauchtwagen

Für das in Beispiel 3.27 (Seite 165) erhaltene Regressionsmodell sind die Cook-Distanzen sämtlich sehr niedrig mit dem größten Wert 0.12. Keine Beobachtungen hat also eine Cook-Distanz größer als 0.5. Für circa 10 Beobachtungen sind die Hebelwerte größer als der Schwellenwert $h_{ii} > 2p/n = 0.058$. Größtenteils handelt es sich um Beobachtungen mit eher kleinem ($kilstand < 50$) bzw. eher großem Kilometerstand ($kilstand > 230$). Der größte Hebelwert beträgt $h_{ii} = 0.23$. Es empfiehlt sich also die Auswirkungen auf die Schätzungen zu untersuchen, wenn Beobachtungen mit besonders kleinem bzw. großem Kilometerstand nicht berücksichtigt werden, vergleiche dazu das nächste Beispiel 3.32. △

Alternative Modellierungsansätze nach Modelldiagnose

Die Modelldiagnose liefert häufig Hinweise auf alternative Modellierungsansätze, die im Anschluss untersucht werden sollten. Hauptziel bei der Schätzung alternativer Modellansätze ist einerseits das *Auffinden noch besser geeigneter Modelle* und andererseits die *Evaluierung der Stabilität der Modelle*. In den allermeisten Fällen wird nicht ein alle anderen überragendes Modell gefunden. Stattdessen sind einige wenige Modelle annähernd gleich gut. Hier gilt es gemeinsame Kernaussagen zu extrahieren und auf Unterschiede hinzuweisen. In der Regel werden sich die Modelle, wenn überhaupt, nur in wenigen Teilaspekten stärker unterscheiden. Hinsichtlich dieser Aspekte herrscht dann mehr oder weniger starke Unsicherheit bezüglich der Schlussfolgerungen, die aus der statistischen Analyse gezogen werden.

Folgende alternativen Varianten kommen grundsätzlich in Frage:

- Insbesondere die Residuenplots und gelegentlich auch der Durbin-Watson-Test auf Autokorrelation geben Hinweise auf *alternative Modellierungen des Einflusses einer oder mehrerer Kovariablen*.
- Die Modelle können erneut geschätzt werden unter *Nichtberücksichtigung von auffälligen Beobachtungen* mit besonders großen Residuen, Hebelwerten und/oder Cook-Distanzen.
- Die Modelle können mit alternativen Schätzverfahren, insbesondere mit *robuster Regression*, erneut geschätzt werden.
- Unter Umständen erhält man durch die Modelldiagnose auch Hinweise auf *Transformationen der Zielgröße*. Insbesondere bei strikt positiver Zielgröße mit sehr großem Wertebereich, kann es sinnvoll sein die Daten zu Logarithmieren.
- Bei Hinweisen auf heteroskedastische Fehler kann alternativ eine *gewichtete Regression* geschätzt werden.

Beispiel 3.32 Preise von Gebrauchtwagen

Das in Beispiel 3.27 (Seite 165) erhaltene Regressionsmodell wurde in den Beispielen 3.28–3.31 einer ausführlichen Modelldiagnose unterzogen. Wir können folgendes Fazit ziehen:

- Es gibt keine Hinweise auf Fehlspezifikation des Einflusses der metrischen Variablen, vergleiche Abbildung 3.30 (Seite 167).
- Abbildung 3.31 gibt Hinweise auf eventuell vorhandene heteroskedastische Fehler.
- Die Einflussanalyse sowie auch der Residuenplot zwischen studentisierten Residuen und Kilometerstand in Abbildung 3.30 weisen auf unsichere Effekte für Golf-Modelle mit Kilometerstand kleiner als 50 bzw. größer als 230 hin.

Daher wurden folgende alternativen Modellspezifikationen getestet:

- *Gewichtete Regression zur Berücksichtigung heteroskedastischer Fehler:* Zum Einsatz kam das in Abschnitt 3.4.3 beschriebene zweistufige Verfahren. Für das Varianzmodell wurden die Kovariablen *alter*, *kilstand* und *tuev* verwendet, wobei der Einfluss des Alters und des Kilometerstands durch Polynome dritten Grades modelliert wurden. Insgesamt ergaben sich keine nennenswerten Unterschiede zur ungewichteten Regression. Dieses Phänomen ist im Zusammenhang mit heteroskedastischen Fehlern häufig zu beobachten. In vielen Anwendungen findet man zwar Hinweise auf Heteroskedastizität, jedoch sind die Unterschiede mit und ohne Berücksichtigung der Heteroskedastizität oft relativ gering.

Abb. 3.36. *Effekt des Kilometerstands für verschiedene alternative Modellierungen und Schätzverfahren.*

- *Logarithmierter Preis als Zielgröße:* Aufgrund der Positivität und der (leichten) Heteroskedastizität wurden auch Modelle mit dem logarithmierten Preis als Zielgröße geschätzt. Bei der Verwendung transformierter Zielgrößen ist zu beachten, dass Modellwahl und Diagnose erneut durchgeführt werden müssen. Beispielsweise können Ausreißer im Modell mit untransformierter Zielgröße im Modell mit logarithmierter Zielgröße verschwinden und umgekehrt. Mit der transformierten Zielgröße erhält man im Vergleich zur untransformierten Zielgröße ein insgesamt etwas sparsameres Modell. Der Effekt der Kilometerleistung kann hier linear modelliert werden. Darüber hinaus ist das Modell noch stabiler bezüglich Ausreißern und einflussreichen Beobachtungen. Die grundsätzlichen Schlussfolgerungen bleiben jedoch gleich.
- *Schätzung des Modells ohne Berücksichtigung von Beobachtungen mit kilstand* < 50 *bzw. kilstand* > 230 *sowie robuste Regression:* Abbildung 3.36 zeigt die Auswirkungen verschiedener Spezifikationen auf den Effekt des Kilometerstands. In allen Modellen beobachten wir einen monoton fallenden Effekt. Lediglich im Bereich *kilstand* < 50 besteht Unsicherheit über den Effekt.

Schließlich können wir folgendes Gesamtfazit ziehen:

- Wir erhalten schwach nichtlineare, monoton fallende Effekte des Alters und der Kilometerleistung.
- Die restlichen erklärenden Variablen haben keinen nachweisbaren (höchstens einen geringen) Effekt auf den Verkaufspreis.
- Die Modelle sind relativ stabil bezüglich Ausreißern und einflussreichen Beobachtungen. Der Effekt der Kilometerleistung für *kilstand* < 50 ist etwas instabil, wenn der untransformierte Preis als Zielgröße verwendet wird.
- Vorhandene Heteroskedastizität kann vernachlässigt werden.

△

3.7 Bemerkungen und Ergänzungen

3.7.1 Literaturhinweise

Nachdem lineare Modelle nach wie vor eine sehr wichtige Rolle in der Statistik einnehmen, existiert eine große Menge an Literatur. Empfehlenswerte neuere Standardmonographien

bzw. Lehrbücher sind Faraway (2004), Rawlings, Pantula & Dickey (2001) und Weisberg (2005). Die deutschsprachige Monographie von Toutenburg (2003) ist im Vergleich zur Behandlung linearer Modelle in diesem Kapitel wesentlich theorielastiger und setzt teilweise deutlich andere Schwerpunkte. Lineare Modelle aus Bayesianischer Sicht werden beispielsweise in O'Hagan (1994) behandelt.

Traditionell stark vertreten ist das lineare Modell auch in Einführungsbüchern zur Ökonometrie, wobei zum Teil andere Schwerpunkte gesetzt werden. Insbesondere explorative und grafische Hilfsmittel zur Modelldiagnose und Spezifikation werden wenig behandelt. Stattdessen spielen zahlreiche statistische Tests zur Modelldiagnose und Spezifikation eine wichtige Rolle. Ein internationales Standardlehrbuch ist Wooldridge (2006). Etwas schwerer zu lesen ist das ausführliche Kompendium von Greene (2000). Mittlerweile existiert die fünfte Auflage aus dem Jahr 2005. Wir empfehlen aber bewusst die dritte Auflage aus dem Jahr 2000, da nachfolgende Auflagen teilweise unübersichtlich und überladen wirken. Empfehlenswert in deutscher Sprache ist von Auer (2005) und nach wie vor Schneeweiß (1990).

Einigen speziellen Verfahren, die hier nur gestreift wurden, sind eigene Monographien gewidmet. Zu nennen ist hier insbesondere Koenker (2005) zur Quantil-Regression, Rousseeuw & Leroy (2003) zur robusten Regression und Belsley, Kuh & Welsch (2003) zur Identifikation von Ausreißern und einflussreichen Beobachtungen. Zur Modellierung korrelierter Fehler wurden hier nur autoregressive Prozesse erster Ordnung behandelt. Komplexere Fehlerstrukturen, insbesondere ARMA-Modelle (Autoregressive Moving Average), werden in der Literatur zur Zeitreihenanalyse ausführlich betrachtet. Standardbücher in englischer Sprache sind Brockwell & Davis (2002) und Hamilton (1994) und in deutsch Schlittgen & Streitberg (2001). Nicht behandelt wurden in diesem Buch Modelle mit multivariater Zielgröße, die beispielsweise in Fahrmeir, Hamerle & Tutz (1996) beschrieben werden. Sogenannte SUR-Modelle (Seemingly Unrelated Regression) und simultane Gleichungssysteme werden in den Lehrbüchern der Ökonometrie behandelt, vergleiche z.B. Greene (2000) und Schneeweiß (1990).

In jüngster Zeit rücken Methoden zur Modellwahl und Variablenselektion vor allem im Zusammenhang mit der Analyse von Gensequenzen vermehrt in den Blickpunkt des Interesses. Bei der Analyse von Gensequenzen müssen häufig aus einer extrem großen Zahl an Variablen (10000 und mehr) bei gleichzeitig geringem Beobachtungsumfang geeignete Kovariablen ausgewählt werden. Ein Standardwerk zur Variablenselektion in linearen Modellen ist Miller (2002). Neuere interessante Ansätze zur Variablenselektion werden in Efron, Hastie, Johnstone & Tibshirani (2004) und Bühlmann (2006) vorgeschlagen.

3.7.2 Beweise

Herleitung des Erwartungswerts der Residuenquadratsumme (Seite 94)

Für den Erwartungswert der Residuenquadratsumme erhalten wir unter Verwendung von Satz B.1.6 (Seite 463) und Eigenschaften der Matrix $\boldsymbol{I} - \boldsymbol{H}$ (vergleiche den Kasten auf Seiten 93)

$$\begin{aligned}
\mathrm{E}(\hat{\varepsilon}'\hat{\varepsilon}) &= \mathrm{E}(\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}) \\
&= sp((\boldsymbol{I} - \boldsymbol{H})\sigma^2\boldsymbol{I}) + \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} \\
&= \sigma^2(n-p) + \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{X}\boldsymbol{\beta} \\
&= \sigma^2(n-p) + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}
\end{aligned}$$

$$= \sigma^2(n-p) + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$
$$= \sigma^2(n-p).$$

□

Beweis der Eigenschaften 3-5 des Kastens auf Seite 97

zu 3) Da die erste Spalte $\boldsymbol{x}^0$ der Designmatrix der $\mathbf{1}$-Vektor ist, folgt die Behauptung unter Zuhilfenahme von Orthogonalitätseigenschaft 2)

$$0 = (\boldsymbol{x}^0)'\,\hat{\boldsymbol{\varepsilon}} = \mathbf{1}'\,\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i.$$

zu 4) Unter Verwendung von Eigenschaft 3) gilt

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (y_i - \hat{\varepsilon}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} y_i.$$

zu 5) Es gilt

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i + \hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} + \hat{\varepsilon}_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k.$$

Man beachte, dass zum Beweis der Eigenschaften entscheidend davon Gebrauch gemacht wurde, dass das Modell eine *Konstante* enthält. Die drei Aussagen sind also nur in üblichen Modellen mit Konstante gültig.

□

Beweis der Streuungszerlegungsformel (3.18) auf Seite 98

Bei der Herleitung der Streuungszerlegung machen wir Gebrauch von der speziellen $n \times n$ Matrix

$$\boldsymbol{C} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'.$$

Offenbar ist $\boldsymbol{C}$ symmetrisch und idempotent. Diese Matrix besitzt eine Reihe bemerkenswerter Eigenschaften. Multiplizieren wir beispielsweise $\boldsymbol{C}$ mit einem beliebigen Vektor $\boldsymbol{a} \in \mathbb{R}^n$, so erhalten wir

$$\boldsymbol{C}\boldsymbol{a} = \begin{pmatrix} a_1 - \bar{a} \\ \vdots \\ a_n - \bar{a} \end{pmatrix}. \tag{3.44}$$

Multiplikation mit $\boldsymbol{C}$ bewirkt also, dass der Vektor $\boldsymbol{a}$ zentriert wird. Für die quadratische Form $\boldsymbol{a}'\boldsymbol{C}\boldsymbol{a}$ erhalten wir

$$\boldsymbol{a}'\boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^{n} (a_i - \bar{a})^2. \tag{3.45}$$

In Definition A.11 (Seite 448) sind weitere Eigenschaften von $\boldsymbol{C}$ zusammengetragen. Für die Herleitung der Streuungszerlegung benötigen wir jedoch nur die Eigenschaften (3.44) und (3.45). Wir multiplizieren zunächst die Identität $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$ von links mit $\boldsymbol{C}$ und erhalten

$$\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\hat{\boldsymbol{y}} + \boldsymbol{C}\hat{\boldsymbol{\varepsilon}}.$$

Aufgrund von (3.44) und Eigenschaft 3) auf Seite 97 gilt $\boldsymbol{C}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$ und es folgt

$$\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\hat{\boldsymbol{y}} + \hat{\varepsilon} \quad \text{bzw} \quad \boldsymbol{y}'\boldsymbol{C} = \hat{\boldsymbol{y}}'\boldsymbol{C} + \hat{\varepsilon}'.$$

Damit erhalten wir

$$\begin{aligned} \boldsymbol{y}'\boldsymbol{C}\boldsymbol{C}\boldsymbol{y} &= (\hat{\boldsymbol{y}}'\boldsymbol{C} + \hat{\varepsilon}')(\boldsymbol{C}\hat{\boldsymbol{y}} + \hat{\varepsilon}) \\ &= \hat{\boldsymbol{y}}'\boldsymbol{C}\boldsymbol{C}\hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}'\boldsymbol{C}\hat{\varepsilon} + \hat{\varepsilon}'\boldsymbol{C}\hat{\boldsymbol{y}} + \hat{\varepsilon}'\hat{\varepsilon} \\ &= \hat{\boldsymbol{y}}'\boldsymbol{C}\hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}'\hat{\varepsilon} + \hat{\varepsilon}'\hat{\boldsymbol{y}} + \hat{\varepsilon}'\hat{\varepsilon}. \end{aligned} \tag{3.46}$$

Für die linke Seite von (3.46) gilt gemäß (3.45) $\boldsymbol{y}'\boldsymbol{C}\boldsymbol{C}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{C}\boldsymbol{y} = \sum(y_i - \bar{y})^2$. Unter Beachtung von $\bar{\hat{y}} = \bar{y}$ (vergleiche Eigenschaft 4) auf Seite 97) folgt $\hat{\boldsymbol{y}}'\boldsymbol{C}\hat{\boldsymbol{y}} = \sum(\hat{y}_i - \bar{y})^2$. Nach Eigenschaft 1) auf Seite 97 gilt $\hat{\boldsymbol{y}}'\hat{\varepsilon} = \hat{\varepsilon}'\hat{\boldsymbol{y}} = 0$ und wir erhalten

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

□

Beweis des Gauß-Markov Theorems (Seite 104)

Zum Beweis überlegen wir uns zunächst Bedingungen, unter denen lineare Schätzer erwartungstreu sind. Als Bedingung für die Erwartungstreue linearer Schätzer muss $\mathrm{E}(\hat{\boldsymbol{\beta}}^L) = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ für *alle* $\boldsymbol{\beta} \in \mathbb{R}^p$ gelten. Setzt man speziell $\boldsymbol{\beta} = \boldsymbol{0}$, so folgt $\boldsymbol{a} = \boldsymbol{0}$ als notwendige Bedingung für die Erwartungstreue von $\hat{\boldsymbol{\beta}}^L$. Umformen der Bedingung $\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ liefert $(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{0}$, so dass als weitere Bedingung für die Erwartungstreue von $\hat{\boldsymbol{\beta}}^L$ zwingend $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}_p$ folgt. Wegen $\mathrm{rg}(\boldsymbol{A}\boldsymbol{X}) = \min(\mathrm{rg}(\boldsymbol{X}), \mathrm{rg}(\boldsymbol{A})) = \mathrm{rg}(\boldsymbol{I}_p) = p$ muss darüber hinaus auch $\mathrm{rg}(\boldsymbol{A}) = p$ gelten.

Sei nun ohne Einschränkung der Allgemeinheit die Matrix $\boldsymbol{A}$ von der Form $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{B}$. Einsetzen in die Unverzerrtheitsbedingung $\boldsymbol{I}_p = \boldsymbol{A}\boldsymbol{X}$ liefert

$$\boldsymbol{I}_p = \boldsymbol{A}\boldsymbol{X} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}\boldsymbol{X} = \boldsymbol{I}_p + \boldsymbol{B}\boldsymbol{X}$$

und somit die Bedingung

$$\boldsymbol{B}\boldsymbol{X} = \boldsymbol{0}.$$

Damit folgt für die Kovarianzmatrix von $\boldsymbol{\beta}^L$:

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}^L) &= \sigma^2\,\boldsymbol{A}\boldsymbol{A}' \\ &= \sigma^2\,\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{B}\right\}\left\{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{B}'\right\} \\ &= \sigma^2\,\Big\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{B}' + \\ &\qquad \boldsymbol{B}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{B}\boldsymbol{B}'\Big\} \\ &= \sigma^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma^2\,\boldsymbol{B}\boldsymbol{B}' \\ &= \mathrm{Cov}(\hat{\boldsymbol{\beta}}) + \sigma^2\,\boldsymbol{B}\boldsymbol{B}' \end{aligned}$$

Nach Satz A.27 (Seite 456) ist $\boldsymbol{B}\boldsymbol{B}'$ nicht negativ definit, so dass wir schließlich durch Umstellen

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}^L) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{B}\boldsymbol{B}' \geq \boldsymbol{0},$$

erhalten. Aus dieser allgemeinen Aussage können wir jetzt die Minimalitätseigenschaften des KQ-Schätzers herleiten. Die Varianzen von $\boldsymbol{b}'\hat{\boldsymbol{\beta}}^L$ und $\boldsymbol{b}'\hat{\boldsymbol{\beta}}$ sind gegeben durch

$$\mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}^L) = \boldsymbol{b}'\mathrm{Cov}(\hat{\boldsymbol{\beta}}^L)\boldsymbol{b} \quad \text{und} \quad \mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}) = \boldsymbol{b}'\mathrm{Cov}(\hat{\boldsymbol{\beta}})\boldsymbol{b}.$$

Da die Differenz der Kovarianzmatrizen nicht negativ definit ist, ist somit die Differenz der Varianzen für jeden beliebigen Vektor $\boldsymbol{b}$ größer oder gleich Null (man beachte die Definition A.25 (Seite 455) definiter Matrizen). Damit ist die Aussage

$$\mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}^L) \geq \mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}})$$

bewiesen.

Da der Vektor $\boldsymbol{b}$ beliebig ist, können wir für $j = 0, \ldots, k$ jeweils $\boldsymbol{b} = (0, \ldots, 1, \ldots, 0)'$ mit der Eins an der $j+1$-ten Position definieren. Wir erhalten dann

$$\mathrm{Var}(\hat{\beta}_j^L) = \mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}^L) \geq \mathrm{Var}(\boldsymbol{b}'\hat{\boldsymbol{\beta}}) = \mathrm{Var}(\hat{\beta}_j).$$

□

Beweis der Aussagen 5 und 6 auf Seite 108

zu 5) Zum Beweis verwenden wir die Matrix

$$\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

wobei $\boldsymbol{H}$ die Prädiktionsmatrix aus Abschnitt 3.2.1 (Seite 93) ist. $\boldsymbol{Q}$ ist symmetrisch und idempotent mit $\mathrm{rg}(\boldsymbol{Q}) = n - p$. Multiplikation mit der Designmatrix $\boldsymbol{X}$ ergibt

$$\boldsymbol{Q}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{0}.$$

Unter Zuhilfenahme dieser Eigenschaft und der Symmetrie sowie Idempotenz von $\boldsymbol{Q}$ folgt

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}'\boldsymbol{Q}\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})'\boldsymbol{Q}\boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Q} + \boldsymbol{\varepsilon}'\boldsymbol{Q})(\boldsymbol{Q}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}'\boldsymbol{Q}\boldsymbol{\varepsilon}$$

Die Behauptung folgt nun unter Verwendung von $\boldsymbol{\varepsilon}/\sigma \sim N(\boldsymbol{0}, \boldsymbol{I})$ und durch Anwendung von Satz B.6 2) auf Seite 466.

zu 6) Die Behauptung ist bewiesen, wenn wir zeigen können, dass $\frac{1}{\sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ und $\frac{1}{\sigma^2}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ unabhängig sind. Unser Ziel ist es Satz B.6 2) auf Seite 466 mit $\boldsymbol{R} = \boldsymbol{Q}$ und $\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ anzuwenden. Die Voraussetzung $\boldsymbol{B}\boldsymbol{R} = \boldsymbol{0}$ ist wegen

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Q} = \boldsymbol{0}$$

erfüllt. Da $\boldsymbol{\varepsilon}/\sigma \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ folgt aus dem Satz, dass

$$\frac{\boldsymbol{\varepsilon}}{\sigma}'\boldsymbol{Q}\frac{\boldsymbol{\varepsilon}}{\sigma} = \frac{1}{\sigma^2}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \quad \text{und} \quad (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}/\sigma$$

unabhängig sind. Da

$$\begin{aligned}
\frac{1}{\sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \frac{1}{\sigma}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{\beta}\right\} \\
&= \frac{1}{\sigma}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta}\right\} \\
&= \frac{1}{\sigma}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\frac{\boldsymbol{\varepsilon}}{\sigma}
\end{aligned}$$

gilt, ist die Behauptung bewiesen.

□

Verteilung der Teststatistik unter H_0 bei linearen Hypothesen (Seite 113)

Berechnung des KQ-Schätzers unter H_0

Wir wollen die Residuenquadratsumme unter der Nebenbedingung $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$ minimieren. Die Minimierung erfolgt mit dem *Lagrange-Ansatz*

$$\begin{aligned}\mathrm{KQR}(\boldsymbol{\beta};\boldsymbol{\lambda}) &= \mathrm{KQ}(\boldsymbol{\beta}) - 2\boldsymbol{\lambda}'(\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d}) \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\,\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\,\boldsymbol{\lambda}'\boldsymbol{C}\boldsymbol{\beta} + 2\,\boldsymbol{\lambda}'\boldsymbol{d}.\end{aligned}$$

Dabei ist $\boldsymbol{\lambda}$ ein Spaltenvektor von Lagrange-Multiplikatoren der Dimension r. Unter Verwendung der Sätze A.32.7 und A.32.10 auf Seite 458 zur Differentiation von Matrixfunktionen erhalten wir:

$$\begin{aligned}\frac{\partial\, KQR(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial\,\boldsymbol{\beta}} &= -2\,\boldsymbol{X}'\boldsymbol{y} + 2\,\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\,\boldsymbol{C}'\boldsymbol{\lambda} \\ \frac{\partial\, KQR(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial\,\boldsymbol{\lambda}} &= -2\,\boldsymbol{C}\boldsymbol{\beta} + 2\,\boldsymbol{d}\end{aligned}$$

Nullsetzen liefert die beiden Gleichungen:

$$\boldsymbol{X}'\boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{C}'\boldsymbol{\lambda}$$

$$\boldsymbol{C}\,\boldsymbol{\beta} = \boldsymbol{d}$$

Wir lösen beide Gleichungen zuerst nach $\boldsymbol{\lambda}$ und anschließend nach $\boldsymbol{\beta}$ auf. Multiplikation der ersten Gleichung mit $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ von links liefert

$$\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\boldsymbol{\lambda}.$$

Multiplikation dieser Gleichung mit der Matrix $\boldsymbol{C}$ von links ergibt

$$\boldsymbol{C}\,\boldsymbol{\beta} - \boldsymbol{C}\,\hat{\boldsymbol{\beta}} = \boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\boldsymbol{\lambda}.$$

Durch Einsetzen der zweiten Gleichung $\boldsymbol{C}\,\boldsymbol{\beta} = \boldsymbol{d}$ erhalten wir

$$\boldsymbol{d} - \boldsymbol{C}\,\hat{\boldsymbol{\beta}} = \boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\boldsymbol{\lambda}$$

und damit

$$\boldsymbol{\lambda} = (\boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\,)^{-1}(\boldsymbol{d} - \boldsymbol{C}\,\hat{\boldsymbol{\beta}}).$$

Dabei haben wir ausgenutzt, dass die Matrix $\boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'$ gemäß Satz A.26.2 (Seite 456) positiv definit und damit invertierbar ist.
Einsetzen von $\boldsymbol{\lambda}$ in die erste Gleichung liefert

$$\boldsymbol{X}'\boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{C}'(\boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\,)^{-1}(\boldsymbol{d} - \boldsymbol{C}\,\hat{\boldsymbol{\beta}})$$

und schließlich den KQ-Schätzer

$$\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'(\boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\,)^{-1}(\boldsymbol{C}\,\hat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

Bestimmung der Differenz der Residuenquadratsummen

Wir schreiben zunächst den restringierten KQ-Schätzer $\hat{\boldsymbol{\beta}}^R$ als

$$\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}} - \Delta_{H_0},$$

wobei Δ_{H_0} definiert ist als

$$\Delta_{H_0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

Damit erhalten wir für die geschätzten Werte $\hat{\boldsymbol{y}}_{H_0}$ unter Beachtung der Restriktion (d.h. unter der Nullhypothese)

$$\hat{\boldsymbol{y}}_{H_0} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^R = \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \Delta_{H_0}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\Delta_{H_0} = \hat{\boldsymbol{y}} - \boldsymbol{X}\Delta_{H_0}$$

und für die Residuen $\hat{\boldsymbol{\varepsilon}}_{H_0}$ unter H_0

$$\hat{\boldsymbol{\varepsilon}}_{H_0} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{H_0} = \boldsymbol{y} - \hat{\boldsymbol{y}} + \boldsymbol{X}\Delta_{H_0} = \hat{\boldsymbol{\varepsilon}} + \boldsymbol{X}\Delta_{H_0}.$$

Damit berechnet sich die Residuenquadratsumme SSE_{H_0} unter H_0 als

$$\begin{aligned}
\mathrm{SSE}_{H_0} &= \hat{\boldsymbol{\varepsilon}}'_{H_0}\hat{\boldsymbol{\varepsilon}}_{H_0} \\
&= (\hat{\boldsymbol{\varepsilon}} + \boldsymbol{X}\Delta_{H_0})'(\hat{\boldsymbol{\varepsilon}} + \boldsymbol{X}\Delta_{H_0}) \\
&= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + \hat{\boldsymbol{\varepsilon}}'\boldsymbol{X}\Delta_{H_0} + \Delta'_{H_0}\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} + \Delta'_{H_0}\boldsymbol{X}'\boldsymbol{X}\Delta_{H_0} \\
&= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + \Delta'_{H_0}\boldsymbol{X}'\boldsymbol{X}\Delta_{H_0}.
\end{aligned}$$

In der vorletzten Zeile haben wir ausgenutzt, dass die Residuen des vollen Modells und die Spalten der Designmatrix orthogonal sind, d.h. $\hat{\boldsymbol{\varepsilon}}'\boldsymbol{X} = 0$. Vergleiche hierzu den Kasten auf Seite 97. Da $\boldsymbol{X}'\boldsymbol{X}$ und damit auch $\Delta'_{H_0}\boldsymbol{X}'\boldsymbol{X}\Delta_{H_0}$ positiv definit sind, haben wir auch den formalen Beweis erbracht, dass die Residuenquadratsumme SSE_{H_0} unter H_0 stets größer oder gleich der Residuenquadratsumme SSE im unrestringierten Modell ist, vergleiche hierzu Seite 113.

Schließlich erhalten wir für die Differenz $\Delta\mathrm{SSE}$ der Residuenquadratsummen:

$$\begin{aligned}
\Delta\mathrm{SSE} &= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + \Delta'_{H_0}\boldsymbol{X}'\boldsymbol{X}\Delta_{H_0} - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\
&= \left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})\right\}'\boldsymbol{X}'\boldsymbol{X}\cdot \\
&\quad \cdot\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})\right\} \\
&= (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}\cdot \\
&\quad \cdot\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \\
&= (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}).
\end{aligned}$$

Stochastische Eigenschaften der Differenz der Residuenquadratsummen

zu 1) Zum Beweis dieser Aussage verwenden wir Satz B.1.6 (Seite 463) über den Erwartungswert quadratischer Formen. Es gilt

$$\mathrm{E}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) = \boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d}$$

und

$$\mathrm{Cov}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) = \sigma^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'.$$

Wir wenden Satz B.1.6 auf $\boldsymbol{Z} = \boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}$ und $\boldsymbol{A} = \left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}$ an und erhalten

$$\begin{aligned}
\mathrm{E}(\Delta\mathrm{SSE}) &= \mathrm{E}\left\{(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})\right\} \\
&= \mathrm{sp}\left\{\sigma^2\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right\} + \\
&\quad (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d}) \\
&= \mathrm{sp}(\sigma^2\boldsymbol{I}_r) + (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d}) \\
&= r\sigma^2 + (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d})'\left(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d}).
\end{aligned}$$

zu 2) Die Behauptung ist eine Folgerung aus Satz B.6.1 (Seite 466). Wir definieren den Zufallsvektor $\boldsymbol{Z} = \boldsymbol{C}\hat{\boldsymbol{\beta}}$. Unter H_0 gilt

$$\mathrm{E}(\boldsymbol{Z}) = \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$$

und

$$\mathrm{Cov}(\boldsymbol{Z}) = \sigma^2 \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'.$$

Da $\hat{\boldsymbol{\beta}}$ normalverteilt ist, folgt weiterhin

$$\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{d}, \sigma^2 \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}').$$

Die Behauptung ergibt sich nun unmittelbar durch Anwendung von Satz B.6.1 auf den Zufallsvektor $\boldsymbol{Z}$.

zu 3) Die Differenz der Residuenquadratsumme ΔSSE hängt nur vom KQ-Schätzer $\hat{\boldsymbol{\beta}}$ ab. Damit folgt die Behauptung unmittelbar aus der Unabhängigkeit von $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ und $\hat{\boldsymbol{\beta}}$, vergleiche den Kasten auf Seite 108.

□

Autokorrelierte Fehler erster Ordnung (Seite 137)

Zunächst leiten wir die sogenannte MA(∞) Darstellung für die Störgrößen ε_i her. Anschließend berechnen wir unter Verwendung dieser Darstellung die Erwartungswerte, Varianzen und Kovarianzen der ε_i.

MA(∞) Darstellung der ε_i

Ziel ist die Darstellung von ε_i als gewichtete unendliche Summe der u_i. Dahinter steckt die Vorstellung, dass der Beginn des Prozesses weit in der Vergangenheit liegt. Wir setzen zunächst $\varepsilon_{i-1} = \rho\,\varepsilon_{i-2} + u_{i-1}$ in $\varepsilon_i = \rho\,\varepsilon_{i-1} + u_i$ ein und erhalten

$$\varepsilon_i = \rho\,(\rho\,\varepsilon_{i-2} + u_{i-1}) + u_i = \rho^2\,\varepsilon_{i-2} + \rho\,u_{i-1} + u_i.$$

Einsetzen von $\varepsilon_{i-2} = \rho\,\varepsilon_{i-3} + u_{i-2}$ liefert

$$\varepsilon_i = \rho^3\,\varepsilon_{i-3} + \rho^2\,u_{i-2} + \rho\,u_{i-1} + u_i.$$

Schließlich erhalten wir durch fortgesetztes Einsetzen die gewünschte Darstellung

$$\varepsilon_i = \sum_{k=0}^{\infty} \rho^k u_{i-k} = u_i + \rho u_{i-1} + \rho^2 u_{i-2} + \ldots. \tag{3.47}$$

Erwartungswert der Störgrößen

Unter Verwendung der Darstellung (3.47) erhalten wir

$$\mathrm{E}(\varepsilon_i) = \sum_{k=0}^{\infty} \rho^k\,\mathrm{E}(u_{i-k}) = 0.$$

Kovarianz der Störgrößen

Unter Verwendung der Darstellung (3.47) und von $\mathrm{E}(\varepsilon_i) = 0$ bzw. $\mathrm{E}(u_i u_j) = 0$ für $i \neq j$ erhalten wir

$$\begin{aligned}
\mathrm{Cov}(\varepsilon_i, \varepsilon_{i-j}) &= \mathrm{E}(\varepsilon_i \varepsilon_{i-j}) \\
&= \mathrm{E}\left(\left(u_i + \rho u_{i-1} + \ldots + \rho^j u_{i-j} + \ldots\right)\left(u_{i-j} + \rho u_{i-j-1} + \ldots\right)\right) \\
&= \mathrm{E}(\rho^j u_{i-j}^2) + \mathrm{E}(\rho^{j+2} u_{i-j-1}^2) + \ldots \\
&= \rho^j \sum_{k=0}^{\infty} \left(\rho^2\right)^k \mathrm{E}\left(u_{i-j-k}^2\right) \\
&= \rho^j \sum_{k=0}^{\infty} \left(\rho^2\right)^k \sigma_u^2 \\
&= \rho^j \frac{\sigma_u^2}{1-\rho^2}.
\end{aligned}$$

Im Spezialfall $j = 0$ erhalten wir schließlich noch

$$\mathrm{Var}(\varepsilon_i) = \frac{\sigma_u^2}{1-\rho^2}.$$

□

Herleitung der Posteriori-Verteilung im linearen Modell (Seite 151)
Zunächst formen wir den Ausdruck

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})$$

um. Unter Verwendung von

$$\begin{aligned}\tilde{\boldsymbol{M}} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1}\\ \tilde{\boldsymbol{m}} &= \tilde{\boldsymbol{M}}(\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y})\end{aligned}$$

erhalten wir

$$\begin{aligned}&(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}) =\\ &\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{M}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} =\\ &\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{M}^{-1}\boldsymbol{m}) + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} =\\ &\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}'\tilde{\boldsymbol{M}}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{M}}(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{M}^{-1}\boldsymbol{m}) + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} =\\ &\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}'\tilde{\boldsymbol{M}}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} =\\ &\boldsymbol{y}'\boldsymbol{y} + (\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{\boldsymbol{M}}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}}) - \tilde{\boldsymbol{m}}'\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m}\end{aligned}$$

Unter Verwendung von

$$\tilde{a} = a + \frac{n}{2}$$

und

$$\tilde{b} = b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} - \tilde{\boldsymbol{m}}'\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{m}}\right)$$

erhalten wir durch Einsetzen in die Posteriori-Verteilung (3.39)

$$P(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}) \propto \tfrac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left(-\tfrac{1}{2\sigma^2}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{\boldsymbol{M}}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})\right) \tfrac{1}{(\sigma^2)^{\tilde{a}+1}} \exp\left(-\tfrac{\tilde{b}}{\sigma^2}\right).$$

Ein Vergleich mit der Dichte einer Normal-inverse Gammaverteilung (vergleiche die Dichte (3.38) auf Seite 148) zeigt, dass die Posteriori-Verteilung wieder eine Normal-inverse Gammaverteilung ist mit Parametern $\tilde{\boldsymbol{m}}$, $\tilde{\boldsymbol{M}}$, $\tilde{a}$ und $\tilde{b}$.

□

4 Generalisierte lineare Modelle

Lineare Modelle eignen sich besonders für Regressionsanalysen, bei denen die Zielvariable stetig ist und – möglicherweise nach einer geeigneten Transformation – zumindest approximativ durch eine Normalverteilung modelliert werden kann. Zusätzlich muss sich der Erwartungswert der Zielvariablen durch eine Linearkombination von – möglicherweise ebenfalls transformierten – Kovariablen darstellen lassen. In vielen Anwendungen ist die Zielvariable jedoch nicht stetig, sondern binär bzw. kategorial oder eine Zählvariable. Beispiele sind:

- Einspruch gegen einen Patentantrag (ja/nein), vgl. Abschnitt 2.3 (Seite 30),
- Kunde kreditwürdig (ja/nein),
- Tumor gut- oder bösartig,
- Person ist arbeitslos, teilzeitbeschäftigt oder vollbeschäftigt,
- Baum ist sehr stark, stark, mittel, schwach, nicht geschädigt,
- Anzahl von Krankheitsfällen, Schadensfällen oder problematischen Krediten in bestimmten Zeitperioden.

Darüber hinaus lassen sich aber auch Regressionsanalysen für bestimmte Typen von stetigen Zielvariablen nicht immer zufriedenstellend mit linearen Modellen durchführen. Dazu gehören deutlich schief verteilte Variablen, wie etwa Lebensdauern, Schadenshöhen oder Einkommen. Obwohl Daten mit schiefer Verteilung sich oft in Daten mit annähernd symmetrischer Verteilung transformieren lassen, kann es günstiger sein ein Gamma-Regressionsmodell auf die nicht transformierte Zielvariable anzuwenden.

Generalisierte lineare Modelle (GLM) umfassen in einem methodisch einheitlichen Rahmen viele Regressionsansätze für nicht notwendigerweise normalverteilte Zielvariablen und enthalten zum Beispiel das Logit-Modell für binäre Zielvariablen (Abschnitt 2.3), aber auch das klassische lineare Modell mit Normalverteilungsannahme als Spezialfall. Wie in diesen beiden Modellen wird angenommen, dass der Effekt der Kovariablen durch einen linearen Prädiktor erfasst werden kann. Wir beschreiben zunächst in Abschnitt 4.1 Regressionsmodelle für binäre Zielvariablen, in Abschnitt 4.2 Regressionsmodelle für Zähldaten, insbesondere die Poisson-Regression und in Abschnitt 4.3 Modelle für positive stetige Zielvariablen. Dabei diskutieren wir jeweils auch die auf dem Likelihood-Prinzip beruhenden Methoden der statistischen Inferenz. Abschnitt 4.4 behandelt dann generalisierte lineare Modelle in einem allgemeinen, einheitlichen Rahmen.

4.1 Binäre Regression

4.1.1 Binäre Regressionsmodelle

Wie im letzten Kapitel gehen wir zunächst davon aus, dass zu n Objekten oder Individuen (ungruppierte) Daten $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, als Beobachtungen zu einer

binären, durch 0 und 1 kodierten Zielvariable y und zu Kovariablen $x_1, \ldots, x_k$, vorliegen. Wie bei linearen Modellen oder im Logit-Modell des Beispiels 2.8 können $x_1, \ldots, x_k$ aus den Original-Kovariablen eventuell durch geeignete Transformation oder Kodierung entstanden sein. Ziel einer binären Regressionsanalyse ist die Modellierung und Schätzung des Effekts der Kovariablen auf die (bedingte) Wahrscheinlichkeit

$$\pi_i = \mathrm{P}(y_i = 1 \,|\, x_{i1}, \ldots, x_{ik}) = \mathrm{E}(y_i \,|\, x_{i1}, \ldots, x_{ik})$$

für das Auftreten von $y_i = 1$, gegeben die Kovariablenwerte $x_{i1}, \ldots, x_{ik}$. Die Zielvariablen werden dabei als (bedingt) unabhängig angenommen.

Die Nachteile eines *linearen Wahrscheinlichkeitsmodells*

$$\pi_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{4.1}$$

für binäre Zielvariablen wurden bereits in Abschnitt 2.3 beschrieben. Insbesondere muss der *lineare Prädiktor*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta}, \tag{4.2}$$

mit $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ und $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ für alle Werte von $\boldsymbol{x}$ im Intervall $[0, 1]$ liegen. Dies erfordert schwierig handhabbare Restriktionen für die aus den Daten zu schätzenden Parameter $\boldsymbol{\beta}$. Alle üblichen binären Regressionsmodelle verknüpfen daher die Wahrscheinlichkeit π_i durch eine Beziehung der Form

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}), \tag{4.3}$$

mit dem linearen Prädiktor η_i. Dabei ist h eine auf der ganzen reellen Achse streng monoton wachsende Verteilungsfunktion, so dass immer $h(\eta) \in [0, 1]$ gilt und die Beziehung (4.3) mit Hilfe der Umkehrfunktion $g = h^{-1}$ in der Form

$$\eta_i = g(\pi_i) \tag{4.4}$$

geschrieben werden kann. In der Terminologie der generalisierten linearen Modelle heißt h *Responsefunktion* (Antwortfunktion) und $g = h^{-1}$ *Linkfunktion* (Verknüpfungsfunktion). Am bekanntesten sind Logit- und Probit-Modelle.

Logit-Modell

Das in Abschnitt 2.3 vorgestellte Logit-Modell ergibt sich durch die Wahl der logistischen Responsefunktion

$$\pi = h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} \tag{4.5}$$

bzw. der Logit-Linkfunktion

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{4.6}$$

Damit erhält man ein lineares Modell für die logarithmierten Chancen (log-odds) $\log(\pi/(1 - \pi))$. Transformation mit der Exponentialfunktion liefert

$$\frac{\pi}{1 - \pi} = \exp(\beta_0)\exp(\beta_1 x_1) \cdot \ldots \cdot \exp(\beta_k x_k), \tag{4.7}$$

so dass die Kovariableneffekte in exponentiell-multiplikativer Form auf die Chance $\pi/(1-\pi)$ wirken, siehe Abschnitt 2.3 für diese Interpretation. Eine weitere Interpretation, die auch für die beiden folgenden Modelle möglich ist, ergibt sich durch die anschließend beschriebene Verknüpfung mit latenten linearen Modellen, siehe Seite 193.

Abb. 4.1. *Responsefunktionen (links) und adjustierte Responsefunktionen im binären Regressionsmodell: Logit-Modell (—), Probit Modell (- - -), komplementäres log-log-Modell (· · ·).*

Probit-Modell

Für h wird die Verteilungsfunktion Φ der Standardnormalverteilung verwendet, d.h.

$$\pi = \Phi(\eta) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}). \tag{4.8}$$

Ein (geringfügiger) Nachteil ist die notwendig werdende numerische Auswertung von Φ bei der Likelihood-Schätzung der Parameter $\boldsymbol{\beta}$.

Komplementäres log-log-Modell

Dieses Modell (complementary log-log-model) besitzt die Extremwert-Verteilungsfunktion

$$h(\eta) = 1 - \exp(-\exp(\eta)) \tag{4.9}$$

als Responsefunktion und die Umkehrfunktion

$$g(\pi) = \log(-\log(1 - \pi))$$

als Linkfunktion. Im Vergleich zum Logit- und Probit-Modell ist das Modell für speziellere Anwendungen interessant, etwa zur Modellierung und Analyse von zeitdiskreten Verweildauern.

Abbildung 4.1 (links) zeigt die Responsefunktionen der drei Modelle, also die logistische Verteilungsfunktion (4.5), die Standard-Normalverteilungsfunktion (4.8) und die Extremwert-Verteilungsfunktion (4.9). Auf den ersten Blick erscheinen die drei Modelle recht unterschiedlich: Die Responsefunktionen des Logit- und Probit-Modells sind zwar beide symmetrisch um Null, aber die logistische Verteilungsfunktion geht für $\eta \to +\infty$ bzw. $\eta \to -\infty$ deutlich langsamer gegen 0 bzw. 1. Im Vergleich dazu ist die Responsefunktion des komplementären log-log-Modells asymmetrisch, hat für kleine η einen ähnlichen Verlauf, geht aber für $\eta \to +\infty$ deutlich schneller gegen 1. Somit scheinen statistische Analysen mit den drei Modellen zu recht unterschiedlichen Ergebnissen zu führen. Für einen adäquaten Vergleich der Modelle ist jedoch zu beachten: Verwendet man statt der Verteilungsfunktion Φ der $\mathrm{N}(0,1)$-Verteilung im Probit-Modell die Verteilungsfunktion

Binäre Regressionsmodelle

Daten

Die binären Zielvariablen y_i sind 0/1-kodiert und bei gegebenen Kovariablen $x_{i1}, \ldots, x_{ik}$ (bedingt) unabhängig.

Modelle

Die Wahrscheinlichkeit $\pi_i = \mathrm{P}(y_i = 1 \,|\, x_{i1}, \ldots, x_{ik})$ und der *lineare Prädiktor*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta}$$

sind durch eine *Responsefunktion* $h(\eta) \in [0, 1]$ miteinander verknüpft:

$$\pi_i = h(\eta_i).$$

Logit-Modell:

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)} \quad \Longleftrightarrow \quad \log \frac{\pi}{1 - \pi} = \eta.$$

Probit-Modell:

$$\pi = \Phi(\eta) \quad \Longleftrightarrow \quad \Phi^{-1}(\pi) = \eta.$$

Komplementäres log-log-Modell:

$$\pi = 1 - \exp(-\exp(\eta)) \quad \Longleftrightarrow \quad \log(-\log(1 - \pi)) = \eta.$$

$h(\eta)$ einer $\mathrm{N}(0, \sigma^2)$-Verteilung mit beliebig gewählter Varianz $\sigma^2 \neq 1$, so erhält man durch Standardisieren von h die Beziehung

$$\pi(\eta) = h(\boldsymbol{x}'\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma) = \Phi(\boldsymbol{x}'\tilde{\boldsymbol{\beta}})$$

mit $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$. Das bedeutet: Obwohl die beiden Responsefunktionen Φ (mit $\sigma^2 = 1$) und h (mit $\sigma^2 \neq 1$, etwa $\sigma^2 = 4$) voneinander verschieden sind, ist die Modellierung der Wahrscheinlichkeit $\pi(\eta)$ durch $h(\eta)$ mit $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ äquivalent zur Modellierung mit einem Probit-Modell mit reskalierten Parametern $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$. In diesem Sinn ist auch die Festlegung $\sigma^2 = 1$ im Probit-Modell willkürlich (man hätte genauso gut $\sigma^2 = 4$ wählen können). Zum gleichen Ergebnis kommt man auch bei der Herleitung von binären Regressionsmodellen aus latenten linearen Modellen (Seite 193).

Für den adäquaten Vergleich von Logit- und Probit-Modell bedeutet dies Folgendes: Da die logistische Verteilungsfunktion die Varianz $\pi^2/3$, mit $\pi = 3.141593\ldots$ als Kreiszahl besitzt, muss sie mit einer gegenüber Φ auf die Varianz $\sigma^2 = \pi^2/3$ adjustierten (reskalierten) Normalverteilungsfunktion $h(\eta)$ verglichen werden. In Abbildung 4.1 (rechts) erkennt man nun, dass die Logit- und die (adjustierte) Probit-Responsefunktion sehr ähnlich sind.

Statistische Analysen mit Logit- und Probit-Modellen führen deshalb zu sehr ähnlichen Resultaten für die aus Daten geschätzten Wahrscheinlichkeiten. Die Reskalierung $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$ bzw. $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}\sigma$ wird beim Schätzen der Parameter automatisch eingehalten. Somit

unterscheiden sich die geschätzten Koeffizienten $\tilde{\beta}_1, \tilde{\beta}_2, \ldots$ eines Logit-Modells von den entsprechenden Werten $\beta_1, \beta_2, \ldots$ eines Probit-Modells (mit $\sigma^2 = 1$) näherungsweise um den Faktor $\sigma = \pi/\sqrt{3} \approx 1.814$, aber die geschätzten Wahrscheinlichkeiten $\pi(\eta)$ sind sehr ähnlich. Da die Quotienten $\tilde{\beta}_1/\tilde{\beta}_2 = \beta_1/\beta_2$ usw. von σ unabhängig sind, sollte man deshalb nicht die absoluten (geschätzten) Koeffizienten, sondern besser die Verhältnisse β_1/β_2 etc. von Effekten interpretieren. Dies wird auch durch Beispiel 4.1 (Seite 194) illustriert.

Für einen Vergleich mit dem komplementären log-log-Modell gelten analoge Überlegungen. Da die Extremwertverteilung die Varianz $\sigma^2 = \pi^2/6$ und zudem den Erwartungswert -0.5772 besitzt, muss die Responsefunktion im Vergleich zur logistischen Verteilungsfunktion so adjustiert werden, dass sie die Varianz $\sigma^2 = \pi^2/3$ und den Erwartungswert 0 besitzt. Dies hat zusätzliche Auswirkungen auf den (geschätzten) Koeffizienten β_0 des Modells. Abbildung 4.1 (rechts) zeigt die so adjustierte Responsefunktion. Sie hat zwar für kleine η einen ähnlichen Verlauf wie Logit- und adjustierte Probit-Funktion, für größere η werden die Unterschiede aber deutlicher sichtbar. Entsprechend unterscheiden sich auch die Resultate statistischer Analysen deutlicher von denen mit Logit- bzw. Probit-Modellen.

Binäre Modelle als Schwellenwertmodelle latenter linearer Modelle

Binäre Regressionsmodelle lassen sich auch erklären durch die Annahme einer latenten (unbeobachteten) Zielvariablen, die mit der beobachtbaren binären Zielvariable durch einen Schwellenwertmechanismus verbunden ist. Sei $\tilde{y}$ eine latente stetige Variable, für die ein lineares Modell

$$\tilde{y} = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$$

gelte. Die Fehlervariable ε besitze die Verteilungsfunktion h. Dabei wird $\tilde{y}$ auch als latenter Nutzen (oder Schaden) interpretiert, der – etwa beim Kauf eines Produktes – zu einer binären Entscheidung y führt. Die beobachtbare binäre Variable y sei mit $\tilde{y}$ über den Schwellenwertmechanismus

$$y = \begin{cases} 1, & \tilde{y} > 0 \\ 0, & \tilde{y} \leq 0 \end{cases}$$

verbunden, wobei 0 der Schwellenwert ist. Damit ergibt sich

$$\pi = \mathrm{P}(y = 1) = \mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0) = 1 - h(-\boldsymbol{x}'\boldsymbol{\beta}) = h(\boldsymbol{x}'\boldsymbol{\beta}),$$

falls die Verteilungsfunktion h symmetrisch um 0 ist. Falls ε logistisch verteilt ist, erhält man das Logit-Modell, und für $\varepsilon \sim \mathrm{N}(0,1)$ das Probit-Modell $\pi = \Phi(\boldsymbol{x}'\boldsymbol{\beta})$. Falls $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ gilt, erhält man durch Standardisieren

$$\pi = h(\boldsymbol{x}'\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma) = \Phi(\boldsymbol{x}'\tilde{\boldsymbol{\beta}}),$$

mit $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$, d.h. mit einem Probit-Modell lassen sich die Regressionskoeffizienten $\boldsymbol{\beta}$ des latenten linearen Regressionsmodells nur bis auf den Faktor $1/\sigma$ identifizieren. Das Verhältnis von zwei Koeffizienten, z.B. β_1 und β_2, ist wegen $\beta_1/\beta_2 = \tilde{\beta}_1/\tilde{\beta}_2$ jedoch identifizierbar.

Interpretation des Logit-Modells

Mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta}$$

gilt für die Chance (odds)

$$\frac{\pi_i}{1-\pi_i} = \frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)}$$

das multiplikative Modell

$$\frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)} = \exp(\beta_0) \cdot \exp(x_{i1}\beta_1) \cdot \ldots \cdot \exp(x_{ik}\beta_k).$$

Wird z.B. x_{i1} um 1 auf $x_{i1} + 1$ erhöht, so gilt für das Verhältnis der Chancen

$$\frac{\mathrm{P}(y_i = 1 \,|\, x_{i1}, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1}, \ldots)} \Big/ \frac{\mathrm{P}(y_i = 1 \,|\, x_{i1} + 1, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1} + 1, \ldots)} = \exp(\beta_1).$$

$$\begin{aligned}
\beta_1 > 0: &\ \text{Chance } \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ wird größer,} \\
\beta_1 < 0: &\ \text{Chance } \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ wird kleiner,} \\
\beta_1 = 0: &\ \text{Chance } \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ bleibt gleich.}
\end{aligned}$$

Parameterinterpretation

Ein wesentlicher Grund für die Popularität des Logit-Modells ist die in Abschnitt 2.3 und durch (4.6) und (4.7) beschriebene Interpretation als lineares Modell für das logarithmierte relative Risiko bzw. als multiplikatives Modell für das relative Risiko $\pi/(1-\pi)$.

Für das Probit-Modell ist die Darstellung durch ein latentes lineares Modell nützlich: Die Kovariableneffekte lassen sich (bis auf einen gemeinsamen Faktor) als Effekte in einem latenten linearen Modell interpretieren. Allgemein denkt man am besten in zwei Stufen: Für den linearen Prädiktor $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ interpretiert man die Effekte wie im linearen Modell. Anschließend wird der für η lineare Effekt durch die Responsefunktion h in einen nichtlinearen Effekt für $\pi = h(\eta)$ transformiert.

Beispiel 4.1 Patenterteilung

In Beispiel 2.8 (Seite 33) wurde die Wahrscheinlichkeit für einen Einspruch bei der Patenterteilung durch ein Logit-Modell mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1 jahr_i + \beta_2 azit_i + \beta_3 ansp_i + \beta_4 uszw_i + \beta_5 patus_i + \beta_6, patdsg_i + \beta_7 aland_i$$

analysiert. Zum Vergleich wählen wir für eine Reanalyse der Daten ein Probit-Modell und ein komplementäres log-log-Modell mit dem selben Prädiktor. Tabelle 4.1 enthält die Parameterschätzungen für alle drei Modelle. Um die Ergebnisse der Probit- und

Variable	Logit	Probit	Probit (adj.)	Log-Log	Log-Log (adj.)
Konstante	201.740	119.204	216.212	164.519	211.744
jahr	-0.102	-0.060	-0.109	-0.083	-0.106
azit	0.113	0.068	0.123	0.088	0.113
ansp	0.026	0.016	0.029	0.021	0.027
uszw	-0.406	-0.243	-0.441	-0.310	-0.398
patus	-0.526	-0.309	-0.560	-0.439	-0.563
patdsg	0.196	0.121	0.219	0.154	0.198
aland	0.097	0.058	0.105	0.080	0.103

Tabelle 4.1. *Geschätzte Koeffizienten für das Logit-, das Probit- und das komplementäre log-log-Modell. Die sich durch Adjustierung ergebenden Koeffizienten des Probit- und des komplementären log-log-Modells sind zusätzlich angegeben.*

Logit-Analyse zu vergleichen, müssen gemäß unseren früheren Überlegungen die geschätzten Koeffizienten des Probit-Modells mit dem Faktor $\pi/\sqrt{3} \approx 1.814$ multipliziert werden. Man erhält zum Beispiel für den geschätzten Effekt der Variablen *jahr* $-0.060 \cdot 1.814 \approx -0.109$. Für andere Koeffizienten treten teilweise etwas größere Abweichungen auf (vgl. die vierte Spalte in Tabelle 4.1). Sie sind aber deutlich geringer als die Standardfehler der Schätzungen (vgl. Beispiel 4.4). Da gemäß der Interpretation binärer Modelle die Koeffizienten nur bis auf den gemeinsamen Faktor $1/\sigma$ interpretierbar sind, liefern das Logit- und das Probit-Modell praktisch gleiche Resultate. Für das komplementäre log-log-Modell ergeben sich nach der Reskalierung mit dem Faktor $\pi/\sqrt{6} \approx 1.283$ ebenfalls Koeffizienten, die nahe an denen des Logit-Modells liegen (Spalte 6 in Tabelle 4.1).

△

Gruppierte Daten

Bislang sind wir davon ausgegangen, dass *Individualdaten* oder *ungruppierte Daten* vorliegen, d.h. zu jedem Individuum oder Objekt i aus einer Stichprobe vom Umfang n liegt eine Beobachtung $(y_i, \boldsymbol{x}_i)$ vor. Jeder binäre, 0-1 kodierte, Wert y_i der Zielvariablen und jeder Kovariablenvektor $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$ gehört dann zu genau einer Einheit $i = 1, \ldots, n$.

Falls mehrere Kovariablenvektoren bzw. Zeilen der Kovariablen-Datenmatrix identisch sind, können die Daten wie in Kapitel 3.4.2 (Seite 127) *gruppiert* werden: Nach Umsortieren und Zusammenfassen enthält die Datenmatrix nur noch Zeilen mit *verschiedenen* Kovariablenvektoren $\boldsymbol{x}_i$. Dazu wird die Anzahl n_i der Wiederholungen von $\boldsymbol{x}_i$ in der Original-Stichprobe der Individualdaten und die relative Häufigkeit $\bar{y}_i$ der entsprechenden individuellen binären Werte der Zielvariablen angegeben:

$$\begin{array}{c} \text{Gruppe } 1 \\ \vdots \\ \text{Gruppe } i \\ \vdots \\ \text{Gruppe } G \end{array} \begin{bmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_G \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_i \\ \vdots \\ \bar{y}_G \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{i1} & & x_{ik} \\ \vdots & & \vdots \\ x_{G1} & \cdots & x_{Gk} \end{bmatrix}$$

Dabei ist G die Anzahl *verschiedener* Kovariablenvektoren in der Stichprobe und oft deutlich kleiner als der Stichprobenumfang n. Dies ist insbesondere der Fall, wenn die Kovariablen binär bzw. kategorial sind. Statt der *relativen* Häufigkeiten $\bar{y}_i$ kann man auch die *absoluten* Häufigkeiten $n_i\bar{y}_i$ angeben. Die Darstellung der gruppierten Daten erfolgt dann oft in komprimierter Form durch eine Kontingenztabelle, wie im folgenden Beispiel 4.2

Das Gruppieren von Individualdaten führt zu Rechenzeit- und Speicherplatzersparnis und wird auch aus Datenschutzgründen durchgeführt. Zudem sind einige Inferenzmethoden besser für gruppierte Daten geeignet, insbesondere im Bereich der Modellüberprüfung und Modelldiagnose, vergleiche Abschnitt 4.1.4 (Seite 205).

Beispiel 4.2 Kaiserschnitt-Geburten

Tabelle 4.2 enthält Daten zu Infektionen von Müttern nach Kaiserschnitt-Geburten im Klinikum der Universität München bereits in gruppierter Form. Die Zielvariable y „Infektion" ist binär mit

$$y = \begin{cases} 1 & \text{Infektion} \\ 0 & \text{keine Infektion.} \end{cases}$$

Dazu wurden bei jeder Geburt drei binäre Kovariablen erhoben:

$$NPLAN = \begin{cases} 1 & \text{Kaiserschnitt-Geburt war nicht geplant,} \\ 0 & \text{geplant,} \end{cases}$$

$$RISK = \begin{cases} 1 & \text{Risikofaktoren vorhanden,} \\ 0 & \text{nicht vorhanden,} \end{cases}$$

$$ANTIB = \begin{cases} 1 & \text{Antibiotika als Prophylaxe verabreicht,} \\ 0 & \text{nicht verabreicht.} \end{cases}$$

Nach Gruppieren der Individualdaten von 251 Müttern können die Daten in Form einer Kontingenztabelle dargestellt werden.

	Kaiserschnitt			
	geplant		**nicht geplant**	
	Infektion		**Infektion**	
	ja	nein	ja	nein
Antibiotika				
Risikofaktor	1	17	11	87
Kein Risikofaktor	0	2	0	0
Kein Antibiotika				
Risikofaktor	28	30	23	3
Kein Risikofaktor	8	32	0	9

Tabelle 4.2. *Gruppierte Infektionsdaten.*

Wenn wir die Wahrscheinlichkeit für eine Infektion durch ein Logit-Modell

$$\log \frac{\mathrm{P(Infektion)}}{\mathrm{P(keine\ Infektion)}} = \beta_0 + \beta_1\, NPLAN + \beta_2\, RISK + \beta_3\, ANTIB$$

modellieren, erhalten wir als geschätzte Koeffizienten

$$\hat{\beta}_0 = -1.89\,, \quad \hat{\beta}_1 = 1.07\,, \quad \hat{\beta}_2 = 2.03\,, \quad \hat{\beta}_3 = -3.25\,.$$

Wegen $\exp(\hat{\beta}_2) = 7.6$ bedeutet dies zum Beispiel, dass sich die „Chance" (odds) bzw. das Risiko für eine Infektion bei Vorliegen von Risikofaktoren um mehr als das Siebenfache erhöht. Diese Interpretation setzt natürlich voraus, dass das gewählte Modell, das keine Interaktionseffekte enthält, adäquat ist. Wir kommen in Beispiel 4.3 auf diese Frage zurück.

Wählt man zur Analyse ein Probit-Modell mit dem gleichen Prädiktor, ergeben sich als geschätzte Koeffizienten

$$\hat{\beta}_0 = -1.09\,, \quad \hat{\beta}_1 = 0.61\,, \quad \hat{\beta}_2 = 1.20\,, \quad \hat{\beta}_3 = -1.90\,.$$

Wie schon in Beispiel 4.1 sind die absoluten Werte scheinbar sehr verschieden; die relativen Effekte, d.h. die Verhältnisse $\hat{\beta}_1/\hat{\beta}_2$ etc. sind sehr ähnlich.

△

Die Individualdaten y_i sind Bernoulli-verteilt mit $\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i) = \pi_i$, d.h. $y_i \sim \mathrm{B}(1, \pi_i)$. Falls die Zielvariablen y_i (bedingt) unabhängig sind, folgt für die absoluten Häufigkeiten $n_i\bar{y}_i$ der gruppierten Daten

$$n_i\bar{y}_i \sim \mathrm{B}(n_i, \pi_i),$$

mit $\mathrm{E}(n_i\bar{y}_i) = n_i\pi_i$, $\mathrm{Var}(n_i\bar{y}_i) = n_i\pi_i(1-\pi_i)$. Die relativen Häufigkeiten sind dann „skaliert" binomialverteilt

$$\bar{y}_i \sim \mathrm{B}(n_i; \pi_i)/n_i,$$

d.h. der Wertebereich der Wahrscheinlichkeitsfunktion für relative Häufigkeiten ist $\{0, 1/n_i, 2/n_i, \ldots, 1\}$ statt $\{0, 1, 2, \ldots, n_i\}$, und es gilt

$$\mathrm{E}(\bar{y}_i) = \pi_i, \quad \mathrm{Var}(\bar{y}_i) = \frac{\pi_i(1-\pi_i)}{n_i}.$$

Für die Modellierung der Wahrscheinlichkeit π_i können die gleichen binären Regressionsmodelle wie für Individualdaten verwendet werden.

Überdispersion (Overdispersion)

Für gruppierte Daten lässt sich die Varianz innerhalb einer Gruppe abschätzen durch $\bar{y}_i(1-\bar{y}_i)/n_i$, da $\bar{y}_i$ der ML-Schätzer für π_i basierend auf den Daten der Gruppe i (und unter Vernachlässigung der Kovariableninformation) ist. In Anwendungen ist diese *empirische* Varianz oft deutlich größer als die durch ein binomiales Regressionsmodell vorhergesagte Varianz $\hat{\pi}_i(1-\hat{\pi}_i)/n_i$ mit $\hat{\pi}_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$. Man spricht dann von Überdispersion, da die Daten mehr Variabilität aufweisen, als durch das Modell unterstellt wird. Zwei Hauptursachen dafür sind: *Unbeobachtete Heterogenität*, die nicht durch die beobachteten Kovariablen erklärt wird, und *positive Korrelation* zwischen individuellen binären Beobachtungen der Zielvariablen, etwa wenn die individuellen Einheiten einem *Klumpen (cluster)*, wie z.B. dem gleichen Haushalt, angehören. In beiden Fällen sind dann die binären individuellen Zielvariablen innerhalb einer Gruppe (in der Regel positiv) korreliert. Die Summe ist dann nicht mehr binomialverteilt, sondern besitzt nach der Summenformel für korrelierte Variablen eine erhöhte Varianz. Die entsprechende Situation tritt in Abschnitt 4.2 für Poisson-verteilte Zielvariablen auf. Dort wird auch ein Beispiel behandelt, in dem Überdispersion vorhanden ist.

Die einfachste Möglichkeit, die vergrößerte Variabilität zu berücksichtigen, ist die Einführung eines multiplikativen Überdispersionsparameters $\phi > 1$ in der Varianzformel, d.h. man nimmt

$$\mathrm{Var}(y_i \,|\, \boldsymbol{x}_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$$

an. Da für die Maximum-Likelihood-Schätzung von $\boldsymbol{\beta}$ nur $\pi_i = \mathrm{E}(y_i \,|\, \boldsymbol{x}_i)$ und $\mathrm{Var}(y_i \,|\, \boldsymbol{x}_i)$ benötigt werden, nicht jedoch die Likelihood selbst, können $\boldsymbol{\beta}$ und ϕ formal genauso geschätzt werden wie im Fall einer Verteilung mit Skalierungsparameter ϕ, etwa bei Normal- oder Gammaverteilung (siehe Abschnitt 4.4.2). Tatsächlich führt die Einführung einer Überdispersion jedoch zur einfachsten Form der so genannten Quasi-Likelihood-Schätzung (vgl. Abschnitt 4.5): Zwar existieren Verteilungen mit Varianz $\phi\pi_i(1-\pi_i)/n_i$, etwa die Beta-Binomialverteilung, aber für die Schätzung wird deren tatsächliche Likelihood nicht verwendet. Andere Ansätze zur Modellierung von Überdispersion sind z.B. Modelle mit zufälligen Effekten oder mit korrelierten Zielvariablen, siehe Kapitel 6. Eine gute Darstellung zum Problem der Überdispersion findet man in Collett (1991).

4.1.2 Maximum-Likelihood-Schätzung

Primäres Ziel der statistischen Inferenz ist wie bei linearen Modellen in Kapitel 3 das Schätzen der Parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ und das Testen von Hypothesen über die Effekte. Die Methodik dieses Abschnitts beruht auf dem (vollständigen) Likelihood-Prinzip: Bei gegebenen Daten $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, erfolgt die Schätzung der Parameter durch Maximierung der (vollständig und korrekt spezifizierten) Likelihood. Hypothesen über die Parameter werden mit Likelihood-Quotienten-, Wald- und Score-Tests durchgeführt, siehe Abschnitt 4.1.3.

Wegen der (bedingten) Unabhängigkeit der Zielvariablen ergibt sich die Likelihood $L(\boldsymbol{\beta})$ als Produkt

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i \,|\, \boldsymbol{\beta}; \boldsymbol{x}_i) \tag{4.10}$$

der Dichten der $y_i \,|\, \boldsymbol{x}_i$, die über $\pi_i = \mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ vom zu schätzenden Parameter $\boldsymbol{\beta}$ abhängen. Maximierung von $L(\boldsymbol{\beta})$ bzw. der Log-Likelihood $l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$ liefert dann den ML-Schätzer $\hat{\boldsymbol{\beta}}$. Wir behandeln den Fall von Individualdaten und beschreiben dazu für das *binäre Logit-Modell* genauer die bei der Bestimmung des ML-Schätzers auftretenden Schritte (vergleiche auch Appendix B.4.1).

1. Bestimmung der Likelihood

Für *binäre Zielvariablen* $y_i \sim \mathrm{B}(1, \pi_i)$ mit $\pi_i = \mathrm{P}(y_i = 1) = \mathrm{E}(y_i) = \mu_i$ ist die (diskrete) Dichte durch

$$f(y_i \,|\, \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

gegeben. über die Beziehung $\pi_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ hängt sie damit, bei gegebenem $\boldsymbol{x}_i$, von $\boldsymbol{\beta}$ ab, so dass wir sie mit $f(y_i \,|\, \boldsymbol{\beta})$ bzw. als Likelihood-Beitrag $L_i(\boldsymbol{\beta})$ der i-ten Beobachtung bezeichnen. Wegen der (bedingten) Unabhängigkeit der y_i ist die *Likelihood* $L(\boldsymbol{\beta})$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

das Produkt der individuellen Likelihoodbeiträge $L_i(\boldsymbol{\beta})$.

2. Bestimmung der Log-Likelihood
Logarithmieren der Likelihood ergibt die Log-Likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(\pi_i) - y_i \log(1-\pi_i) + \log(1-\pi_i)\}, \tag{4.11}$$

mit den *Log-Likelihood-Beiträgen*

$$\begin{aligned} l_i(\boldsymbol{\beta}) = \log L_i(\boldsymbol{\beta}) &= y_i \log(\pi_i) - y_i \log(1-\pi_i) + \log(1-\pi_i) \\ &= y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \end{aligned}$$

Für das Logit-Modell gilt

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \quad \text{bzw.} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i.$$

Es folgt wegen $(1-\pi_i) = (1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^{-1}$

$$l_i(\boldsymbol{\beta}) = y_i(\boldsymbol{x}_i'\boldsymbol{\beta}) - \log(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})) = y_i\eta_i - \log(1+\exp(\eta_i)).$$

3. Berechnung der Score-Funktion
Zur Berechnung des ML-Schätzers als Maximierer der Log-Likelihood $l(\boldsymbol{\beta})$ bildet man die 1. Ableitung nach $\boldsymbol{\beta}$, die Score-Funktion

$$\boldsymbol{s}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}) \tag{4.12}$$

mit den individuellen Beiträgen $s_i(\boldsymbol{\beta}) = \partial l_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. Nach der Kettenregel folgt

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial l_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \left[y_i - \frac{1}{1+\exp(\eta_i)}\exp(\eta_i)\right]\boldsymbol{x}_i,$$

mit dem p-dimensionalen Vektor $\partial\eta_i/\partial\boldsymbol{\beta} = \boldsymbol{x}_i$. Einsetzen der logistischen Beziehung $\pi_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$ liefert

$$\boldsymbol{s}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i(y_i - \pi_i)$$

und die Score-Funktion

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \pi_i). \tag{4.13}$$

Dabei hängt $s(\boldsymbol{\beta})$ über $\pi_i = \pi_i(\boldsymbol{\beta}) = h(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$ nichtlinear von $\boldsymbol{\beta}$ ab. Wegen $\mathrm{E}(y_i) = \pi_i$ gilt die Gleichung

$$\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \boldsymbol{0}. \tag{4.14}$$

Nullsetzen der Score-Funktion liefert die *ML-Gleichung*

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \frac{\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}\right) = \boldsymbol{0}. \tag{4.15}$$

Dieses p-dimensionale, nichtlineare Gleichungssystem für $\hat{\boldsymbol{\beta}}$ wird üblicherweise iterativ durch den Newton-Raphson- bzw. Fisher-Scoring-Algorithmus gelöst, vgl. Seite 202.

4. Bestimmung der Informationsmatrizen

Zur Schätzung der Koeffizienten und der Kovarianzmatrix des ML-Schätzers $\hat{\boldsymbol{\beta}}$ benötigt man die *beobachtete Informationsmatrix*

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'},$$

mit den zweiten Ableitungen $\partial^2 l(\boldsymbol{\beta})/\partial\beta_j\partial\beta_k$ als Elementen der Matrix $\partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'$, oder die *Fisher-Matrix (erwartete Informationsmatrix)*

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) = \mathrm{Cov}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta})).$$

Die letzte Gleichung gilt wegen $\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathbf{0}$. Damit lässt sich die Fisher-Matrix $F(\boldsymbol{\beta})$ folgendermaßen herleiten: Es ist

$$\begin{aligned}\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta}) &= \left\{\sum_{i=1}^n \boldsymbol{x}_i(y_i-\pi_i)\right\}\left\{\sum_{j=1}^n (y_j-\pi_j)\boldsymbol{x}_j'\right\} \\ &= \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'(y_i-\pi_i)^2 + \sum_{i\neq j}\boldsymbol{x}_i\boldsymbol{x}_j'(y_i-\pi_i)(y_j-\pi_j).\end{aligned}$$

Wegen der Unabhängigkeit von y_i und y_j, $i \neq j$, ist $\mathrm{E}((y_i-\pi_i)(y_j-\pi_j)) = \mathrm{E}(y_i - \pi_i)\mathrm{E}(y_j-\pi_j) = 0$, so dass der Erwartungswert der zweiten Summe $\mathbf{0}$ ist. Für die erste Summe gilt

$$\mathrm{E}\left(\sum_i \boldsymbol{x}_i\boldsymbol{x}_i'(y_i-\pi_i)^2\right) = \sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\mathrm{E}(y_i-\pi_i)^2.$$

Wegen $\mathrm{E}(y_i-\pi_i)^2 = \mathrm{Var}(y_i) = \pi_i(1-\pi_i)$ erhält man somit

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta})) = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\pi_i(1-\pi_i).$$

Wegen $\pi_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ hängt auch die Fisher-Matrix von $\boldsymbol{\beta}$ ab. Die beobachtete Informationsmatrix $\boldsymbol{H}(\boldsymbol{\beta}) = -\partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = -\partial\boldsymbol{s}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$ erhält man durch nochmaliges Differenzieren von

$$-\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i(\pi_i(\boldsymbol{\beta})-y_i).$$

Mit Hilfe der Kettenregel erhält man nach kurzer Rechnung

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\frac{\partial\boldsymbol{s}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'} = \sum_{i=1}^n \boldsymbol{x}_i\frac{\partial\pi_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'} = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\frac{\partial\pi_i(\boldsymbol{\beta})}{\partial\eta_i} = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\pi_i(\boldsymbol{\beta})(1-\pi_i(\boldsymbol{\beta})).$$

Für das Logit-Modell sind also die erwartete und die beobachtete Informationsmatrix identisch, d.h. es gilt

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\partial^2 l(\boldsymbol{\beta})\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = \mathrm{E}\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) = \boldsymbol{F}(\boldsymbol{\beta}).$$

Diese Gleichung gilt jedoch *nicht* für andere Modelle wie das *Probit-* oder das *komplementäre log-log-Modell*. Üblicherweise verwendet man in diesen Modellen die Fisher-Matrix $\boldsymbol{F}(\boldsymbol{\beta})$, mit der in Abschnitt 4.4.2 angegebenen allgemeinen Form.

Wenn keine Individualdaten mit binären Zielvariablen $y_i \sim \mathrm{B}(1, \pi_i)$, sondern binomialverteilte Zielvariablen $y_i \sim \mathrm{B}(n_i, \pi_i)$ bzw. relative Häufigkeiten

$$\bar{y}_i \sim \mathrm{B}(n_i, \pi_i)/n_i \qquad i = 1, \ldots, n,$$

wie z.B. bei gruppierten Daten vorliegen, sind die Formeln für $l(\boldsymbol{\beta}), \boldsymbol{s}(\boldsymbol{\beta})$ und $\boldsymbol{F}(\boldsymbol{\beta})$ geeignet zu modifizieren. Die Log-Likelihood ist, von additiven Konstanten abgesehen,

$$\sum_{i=1}^{G} \{y_i \log(\pi_i) - y_i \log(1-\pi_i) + n_i \log(1-\pi_i)\}.$$

Daraus erhält man

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{G} \boldsymbol{x}_i (y_i - n_i \pi_i) = \sum_{i=1}^{G} n_i \boldsymbol{x}_i (\bar{y}_i - \pi_i)$$

und

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{G} \boldsymbol{x}_i \boldsymbol{x}_i' n_i \pi_i (1-\pi_i).$$

Vergleich mit der ML- bzw. KQ-Schätzung im linearen Regressionsmodell

Unter Normalverteilungsannahme gilt im linearen Regressionsmodell

$$y_i \sim \mathrm{N}(\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2).$$

Von konstanten Faktoren abgesehen, folgt für die Score-Funktion

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \mu_i),$$

wobei $\mathrm{E}(y_i) = \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ linear von $\boldsymbol{\beta}$ abhängt. Die Score-Funktion (4.13) für das Logit-Modell besitzt mit $\mathrm{E}(y_i) = \pi_i = \mu_i$ die gleiche Struktur, aber wegen $\pi_i = \mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}$ ist die Abhängigkeit von $\boldsymbol{\beta}$ nichtlinear. Die ML- bzw. KQ-Gleichung im linearen Modell ist

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0},$$

mit $\boldsymbol{y} = (y_1, \ldots, y_n)'$ und der Designmatrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. Falls die Designmatrix $\boldsymbol{X}$ vollen Rang p besitzt, erhält man als Lösung des linearen Gleichungssystems $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ den Schätzer

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

in einem Schritt, während im Logit-Modell die Lösung des nichtlinearen Gleichungssystems (4.15) in mehreren Iterationsschritten numerisch berechnet wird.

Die (beobachtete und erwartete) Informationsmatrix im linearen Modell ist

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{x}_i/\sigma^2 = \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}.$$

Die Struktur ist wieder ganz ähnlich wie beim Logit-Modell, jedoch hängt die Informationsmatrix nicht von $\boldsymbol{\beta}$ ab.

Iterative numerische Berechnung des ML-Schätzers

Zur Bestimmung des ML-Schätzers $\hat{\boldsymbol{\beta}}$ können verschiedene iterative numerische Algorithmen eingesetzt werden, die $\hat{\boldsymbol{\beta}}$ als Nullstelle der Likelihood-Gleichungen $\boldsymbol{s}(\hat{\boldsymbol{\beta}})$ berechnen. Am üblichsten ist das Fisher-Scoring-Verfahren, vgl. Abschnitt B.4.2 in Anhang B. Ausgehend von einem Startwert $\hat{\boldsymbol{\beta}}^{(0)}$, etwa einem Kleinste-Quadrate-Schätzer, sind die Iterationen durch

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{s}(\hat{\boldsymbol{\beta}}^{(k)}), \quad k = 0, 1, 2, \ldots \tag{4.16}$$

definiert. Falls ein Abbruchkriterium, etwa $||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}||/||\hat{\boldsymbol{\beta}}^{(k)}|| \leq \varepsilon$, erfüllt ist, wird gestoppt und $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}^{(k)}$ ist der ML-Schätzer. Da für das Logit-Modell $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{H}(\boldsymbol{\beta}) = -\partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'$ gilt, entspricht das Fisher-Scoring-Verfahren dem Newton-Verfahren. Die Iterationen (4.16) lassen sich auch in der Form einer iterativ gewichteten KQ-Schätzung schreiben, siehe Abschnitt 4.4.2 (Seite 220).

Damit die Fisher-Scoring-Iterationen gegen eine Lösung $\hat{\boldsymbol{\beta}}$ der ML-Gleichungen konvergieren, muss zumindest die Fisher-Matrix $\boldsymbol{F}(\boldsymbol{\beta})$ für alle $\boldsymbol{\beta}$ invertierbar sein. Wie im linearen Regressionsmodell muss dazu die Designmatrix vollen Rang p besitzen. Dann ist $\boldsymbol{F}(\boldsymbol{\beta})$ für die hier betrachteten Regressionsmodelle invertierbar. Für das Logit-Modell folgt dies z.B., weil dann auch $\boldsymbol{X}'\boldsymbol{X} = \sum_i \boldsymbol{x}_i\boldsymbol{x}_i'$ vollen Rang p besitzt und damit wegen $\pi_i(1-\pi_i) > 0$ für alle $\boldsymbol{\beta} \in \mathbb{R}^p$ auch $\boldsymbol{F}(\boldsymbol{\beta}) = \sum_i \boldsymbol{x}_i\boldsymbol{x}_i'\pi_i(1-\pi_i)$. Wir setzen somit wie im linearen Regressionsmodell

$$\mathrm{rg}(\boldsymbol{X}) = p$$

voraus.

In der Regel konvergiert dann der Algorithmus und stoppt nach einigen Iterationsschritten in der Nähe eines Maximums. Für die hier betrachteten Modelle lässt sich zeigen, dass das dann erreichte Maximum eindeutig und somit ein globales Maximum ist.

Dennoch kann der Fall auftreten, dass die Iterationen divergieren, d.h. dass die sukzessiven Differenzen $\|\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}\|$ anwachsen statt gegen Null zu konvergieren. Dieser Fall tritt auf, wenn die (Log)-Likelihood kein Maximum für endliches $\boldsymbol{\beta}$ besitzt und somit wenigstens eine Komponente von $\hat{\boldsymbol{\beta}}^{(k)}$ gegen $\pm\infty$ geht, also kein endlicher ML-Schätzwert existiert. In der Regel tritt die Nichtexistenz des ML-Schätzers bei sehr ungünstigen Datenkonstellationen auf, vor allem wenn der Stichprobenumfang n im Vergleich zur Dimension p klein ist.

Bedingungen zur Eindeutigkeit und Existenz des ML-Schätzers sind zwar für spezielle Modelle von verschiedenen Autoren angegeben. Diese Bedingungen sind jedoch teilweise recht komplex. Für praktische Zwecke ist es deshalb einfacher die Konvergenz bzw. Divergenz des Iterationsverfahrens empirisch zu überprüfen.

Beispiel 4.3 Kaiserschnitt-Geburten

In Beispiel 4.2 wurde für den linearen Prädiktor ein so genanntes Haupteffekt-Modell

$$\eta = \beta_0 + \beta_1\, NPLAN + \beta_2\, RISK + \beta_3\, ANTIB$$

gewählt, also ein Modell ohne Interaktionen zwischen den Kovariablen. Möchte man die Anpassung des Modells an die Daten durch Einführen von Interaktionstermen verbessern, so zeigt sich folgendes:

Bezieht man nur die Interaktion $NPLAN \cdot ANTIB$ ein, ergibt sich nur eine unwesentliche Verbesserung. Will man jedoch die Interaktionen $RISK \cdot ANTIB$ oder $NPLAN \cdot RISK$ einbeziehen, tritt das Problem der Nichtexistenz, d.h. der Divergenz der ML-Schätzung auf. Der Grund liegt darin, dass in den beiden Zellen $NPLAN = 0$, $RISK = 0$, $ANTIB = 1$ und $NPLAN = 1$, $RISK = 0$, $ANTIB = 0$ für die Zielvariable jeweils nur die Kategorie ,,keine Infektion" beobachtet wurde. Dies führt dazu, dass die Schätzwerte für die Effekte von $ANTIB$ und $RISK \cdot ANTIB$ bzw. $NPLAN$ und $NPLAN \cdot RISK$ gegen unendlich divergieren und die Interaktionen mit ungewöhnlich hohen Schätzwerten für diese Effekte abbrechen. Abhängig von der gewählten Software wird dabei eine Warnung ausgegeben oder auch nicht.

Dass dieses Problem von der besonderen Datenkonstellation abhängt sieht man auch, wenn man jeweils eine Beobachtung in den beiden Zellen in die ,,ja"-Kategorie überführt. Dann existieren endliche ML-Schätzungen und die Interaktionen konvergieren tatsächlich.

△

Asymptotische Eigenschaften des ML-Schätzers

Unter relativ schwachen Annahmen lässt sich zeigen, dass der ML-Schätzer für $n \to \infty$ asymptotisch existiert und sowohl konsistent als auch asymptotisch normalverteilt ist. Dabei müssen keineswegs die Messwiederholungen an jeder Messstelle gegen unendlich gehen, sondern es reicht aus, wenn nur der Gesamtstichprobenumfang $n \to \infty$ geht. Für hinreichend großen Stichprobenumfang n gilt dann, dass $\hat{\boldsymbol{\beta}}$ approximativ normalverteilt ist:

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})),$$

und die geschätzte Kovarianzmatrix

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$$

ist gleich der inversen Fisher-Matrix, ausgewertet für den ML-Schätzer $\hat{\boldsymbol{\beta}}$. Das Diagonalelement a_{rr} der inversen Fisher-Matrix $\boldsymbol{A} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$ ist somit ein Schätzer für die Varianz der r-ten Komponente $\hat{\beta}_j$ von $\hat{\boldsymbol{\beta}}$, d.h. es ist

$$\widehat{\mathrm{Var}}(\hat{\beta}_j) = a_{jj},$$

und $\sqrt{a_{jj}}$ ist ein Schätzer für die Standardabweichung $\sqrt{\mathrm{Var}(\hat{\beta}_j)}$.

4.1.3 Testen linearer Hypothesen

Lineare Hypothesen besitzen die gleiche Form wie im linearen Modell:

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \quad \text{gegen} \quad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d}, \tag{4.17}$$

wobei $\boldsymbol{C}$ vollen Zeilenrang $r \leq p$ hat. Zum Testen können die Likelihood-Quotienten-, die Score- und die Wald-Statistik verwendet werden, vgl. Anhang B.4.4.
Die *Likelihood-Quotienten-Statistik*

$$lq = -2\{l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\}$$

misst die Abweichung zwischen dem unrestringierten Maximum $l(\hat{\boldsymbol{\beta}})$ und dem unter H_0 restringierten Maximum $l(\tilde{\boldsymbol{\beta}})$, wobei $\tilde{\boldsymbol{\beta}}$ ML-Schätzer unter der Gleichungsrestriktion $\boldsymbol{C\beta} = \boldsymbol{d}$ ist. Für den Spezialfall

$$H_0 : \boldsymbol{\beta}_j = 0 \quad \text{gegen} \quad H_1 : \boldsymbol{\beta}_j \neq 0, \tag{4.18}$$

wobei $\boldsymbol{\beta}_j$ ein Teilvektor von $\boldsymbol{\beta}$ ist, testet man auf Signifikanz der zu $\boldsymbol{\beta}_j$ gehörigen Effekte. Die Berechnung von $\tilde{\boldsymbol{\beta}}$ erfordert dann die ML-Schätzung des entsprechenden Teilmodells. Für allgemeinere lineare Gleichungsrestriktionen ist der numerische Aufwand deutlich größer, da die Maximierung unter der Nebenbedingung $\boldsymbol{C\beta} = \boldsymbol{d}$ durchgeführt werden muss.
Die *Wald-Statistik*

$$w = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$$

misst die gewichtete Distanz zwischen $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ und $\boldsymbol{d} = \boldsymbol{C\beta}$, wobei mit der (inversen) asymptotischen Kovarianzmatrix $\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}'$ von $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ gewichtet wird.
Die *Score-Statistik*

$$u = \boldsymbol{s}'(\tilde{\boldsymbol{\beta}})\boldsymbol{F}^{-1}(\tilde{\boldsymbol{\beta}})\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$$

misst die gewichtete Distanz zwischen dem Wert $\boldsymbol{0} = \boldsymbol{s}(\hat{\boldsymbol{\beta}})$ der Score-Funktion, ausgewertet an der Stelle $\hat{\boldsymbol{\beta}}$, und dem Wert $\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$, ausgewertet für den restringierten ML-Schätzer $\tilde{\boldsymbol{\beta}}$.
Wald-Tests sind dann rechnerisch vorteilhaft, wenn zu einem bereits geschätzten Modell ein Teilmodell getestet werden soll. Dann braucht das Teilmodell nicht zusätzlich geschätzt werden. Umgekehrt sind Score-Tests günstig, wenn ein geschätztes Modell gegen ein Obermodell getestet werden soll. Mit Wald- und Score-Tests können deshalb Rückwärts- bzw. Vorwärtsselektionen von Variablen numerisch effizient durchgeführt werden.
Für die spezielle Hypothese (4.18) reduzieren sich Wald- und Score-Statistik zu

$$w = \hat{\boldsymbol{\beta}}_j' \hat{\boldsymbol{A}}_j^{-1} \hat{\boldsymbol{\beta}}_j$$

und

$$u = \tilde{\boldsymbol{s}}_j' \tilde{\boldsymbol{A}}_j \tilde{\boldsymbol{s}}_j,$$

wobei $\boldsymbol{A}_j$ die den Elementen von $\boldsymbol{\beta}_j$ entsprechende Teilmatrix von $\boldsymbol{A} = \boldsymbol{F}^{-1}$ und $\boldsymbol{s}_j$ der entsprechende Teilvektor der Score-Funktion $\boldsymbol{s}(\boldsymbol{\beta})$ ist. Die Notation „$\hat{}$" bzw. „$\tilde{}$" bedeutet den jeweiligen Wert an der Stelle $\hat{\boldsymbol{\beta}}$ bzw. $\tilde{\boldsymbol{\beta}}$. Falls β_j nur ein Element von $\boldsymbol{\beta}$ ist, ist die Wald-Statistik gleich dem quadrierten *„t-Wert"*

$$t_j = \frac{\hat{\beta}_j}{\sqrt{a_{jj}}},$$

mit a_{jj} als j-tem Diagonalelement der asymptotischen Kovarianzmatrix $\boldsymbol{A} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$.
Unter ähnlich schwachen Voraussetzungen wie für die asymptotische Normalität des ML-Schätzers sind die drei Teststatistiken unter H_0 asymptotisch äquivalent und asymptotisch bzw. approximativ χ^2-verteilt mit r Freiheitsgraden:

$$lq, w, u \overset{a}{\sim} \chi^2_r.$$

Kritische Werte oder p-Werte werden über diese asymptotische Verteilung berechnet. Für mittleren Stichprobenumfang ist die Approximation durch die χ^2-Grenzverteilung in der Regel ausreichend. Für kleinere Stichprobenumfänge, etwa ab $n \leq 50$, können sich die Werte der Teststatistiken jedoch deutlich unterscheiden.

4.1.4 Kriterien zur Modellanpassung und Modellwahl

Der Beurteilung der Modellanpassung eines geschätzten Modells liegt folgende Idee zugrunde: Sind die Daten soweit wie möglich gruppiert worden, so lässt sich der Parameter π_i in einer Gruppe durch den Mittelwert $\bar{y}_i$ schätzen. Die Verwendung der Mittelwerte als Schätzer entspricht nun einem sogenannten *saturierten* Modell, da das Modell für jede Gruppe einen separaten Parameter beinhaltet. Damit ist das saturierte Modell maximal an die Daten angepasst und dient als Maßstab zur Beurteilung der Modellanpassung geschätzter Regressionsmodelle. Formal kann man dann testen, ob die Abweichung in der Modellanpassung zwischen geschätztem Modell und saturiertem Modell signifikant ist. Zur Durchführung dieses Tests werden am häufigsten die Pearson-Statistik und die Devianz verwendet. Zur Definition dieser Test-Statistiken sollten die Daten soweit wie möglich gruppiert worden sein.
Die *Pearson-Statistik* ist gegeben durch die Summe der quadrierten, standardisierten Residuen

$$\chi^2 = \sum_{i=1}^{G} \frac{(\bar{y}_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)/n_i},$$

wobei G die Anzahl der Gruppen ist, $\bar{y}_i$ die relative Häufigkeit von Einsen in Gruppe i, $\hat{\pi}_i = h(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$ die durch das Modell geschätzte Wahrscheinlichkeit $\mathrm{P}(y_i = 1|\boldsymbol{x}_i)$ und $\hat{\pi}_i(1-\hat{\pi}_i)/n_i$ die geschätzte Varianz.
Die *Devianz* ist definiert durch

$$D = -2\sum_{i=1}^{G}\{l_i(\hat{\pi}_i) - l_i(\bar{y}_i)\},$$

wobei $l_i(\hat{\pi}_i)$ bzw. $l_i(\bar{y}_i)$ die Log-Likelihood der Gruppe i ist. Falls die Anzahl n_i in allen Gruppen $i = 1, \ldots, G$ hinreichend groß ist, sind beide Statistiken approximativ χ^2_{G-p}-verteilt, wobei p die Anzahl der geschätzten Koeffizienten ist. Damit lässt sich basierend auf dieser approximativen Verteilung ein formaler Modellanpassungstest durchführen, indem der vorliegende Wert der Test-Statistik mit dem entsprechenden Quantil der χ^2_{G-p}-Verteilung verglichen wird. Wenn jedoch n_i klein ist, insbesondere $n_i = 1$ wie bei ungruppierten Individualdaten, ist die Anwendung als Teststatistik problematisch. Große Werte von χ^2 oder D weisen dann nicht notwendig auf eine schlechte Anpassung hin.

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	201.740	22.321	9.04	<0.001
jahr	-0.102	0.011	-9.10	<0.001
azit	0.113	0.022	5.09	<0.001
ansp	0.026	0.006	4.49	<0.001
uszw	-0.402	0.100	-4.03	<0.001
patus	-0.526	0.113	-4.67	<0.001
patdsg	0.196	0.117	1.68	0.094
aland	0.097	0.015	6.55	<0.001

Tabelle 4.3. *Patenterteilung: Ergebnisse des Logit-Modells.*

Will man Modelle mit verschiedenen Prädiktoren und Parametern vergleichen, muss ein Kompromiss zwischen guter Datenanpassung durch hohe Parameteranzahl und zu großer Modellkomplexität getroffen werden. Am bekanntesten ist Akaikes Informationskriterium

$$\mathrm{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2p\,,$$

bei dem der Term $2p$ die Anzahl der Parameter in einem zu komplexen Modell bestraft. Bevorzugt werden bei der Wahl zwischen verschiedenen Modellen jene mit kleinsten AIC-Werten. Statt des AIC-Wertes selbst wird oft auch der mit dem Stichprobenumfang n normierte Wert AIC/n verwendet.

Beispiel 4.4 Patenterteilung

Tabelle 4.3 enthält für das Logit-Modell von Beispiel 4.1 (Seite 194) die geschätzten Koeffizienten zusammen mit den jeweiligen Standardabweichungen, t-Werten und p-Werten. Für die Log-Likelihood und das AIC-Kriterium des geschätzten Modells ergeben sich

$$l(\hat{\boldsymbol{\beta}}) = -1488.560\,, \quad \mathrm{AIC} = 2993.12\,.$$

Mit einem p-Wert von 0.094 ist der Einfluss der Variablen *patdsg* nur schwach signifikant. Wählt man als Signifikanzniveau $\alpha = 5\%$, so wird die Hypothese $H_0 : \beta_6 = 0$ nicht abgelehnt, d.h. die erhöhte Wahrscheinlichkeit für einen Einspruch, falls das Patent aus Deutschland, der Schweiz oder Großbritannien kommt, erscheint als nicht signifikant.

Tabelle 4.4 enthält die analogen Werte für das Probit-Modell, wobei sich zwar die geschätzten Koeffizienten und ihre Standardabweichungen wegen der fehlenden Adjustierung unterscheiden (vgl. Beispiel 4.1), die t- und p-Werte jedoch weitgehend gut übereinstimmen und zu den gleichen Schlussfolgerungen hinsichtlich der Signifikanz von Effekten führen. Für die Log-Likelihood und das AIC-Kriterium ergeben sich mit

$$l(\hat{\beta}) = -1488.407\,, \quad \mathrm{AIC} = 2992.815$$

sehr ähnliche Werte. Analysen der Patentdaten mit Logit- und Probit-Modellen führen also zu vergleichbaren Ergebnissen. Wir beschreiben deshalb die folgenden Resultate nur für das Logit-Modell.

Um zu untersuchen, ob der Effekt der metrischen Kovariablen *aland* linear oder nichtlinear ist, wurde in Beispiel 2.8 (Seite 33) ein kubisches Polynom

$$\beta_7\, aland + \beta_8\, aland^2 + \beta_9\, aland^3$$

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	119.204	13.191	9.04	<0.001
jahr	-0.060	0.007	-9.11	<0.001
azit	0.067	0.013	5.02	<0.001
ansp	0.016	0.004	4.46	<0.001
uszw	-0.243	0.060	-4.07	<0.001
patus	-0.309	0.065	-4.72	<0.001
patdsg	0.121	0.071	1.71	0.086
aland	0.058	0.009	6.51	<0.001

Tabelle 4.4. *Patenterteilung: Ergebnisse des Probit-Modells.*

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	198.131	22.739	8.71	<0.001
jahr	-0.100	0.011	-8.82	<0.001
azit	0.113	0.022	5.08	<0.001
ansp	0.026	0.006	4.45	<0.001
uszw	-0.409	0.100	-4.09	<0.001
patus	-0.539	0.113	-4.50	<0.001
patdsg	0.180	0.119	1.54	0.130
aland	0.394	0.184	2.14	0.032
$aland^2$	-0.038	0.024	-1.58	0.113
$aland^3$	0.001	0.001	1.50	0.134

Tabelle 4.5. *Patenterteilung: Ergebnisse des erweiterten Logit-Modells.*

in den linearen Prädiktor aufgenommen und dieses erweiterte Logit-Modell geschätzt. Tabelle 4.5 enthält die geschätzten Koeffizienten zusammen mit Standardabweichungen, t- und p-Werten. Die t- bzw. p-Werte für die Effekte von $aland^2$ und $aland^3$ weisen bereits darauf hin, dass die sparsamere lineare Modellierung wohl doch ausreicht und die Nichtlinearität überinterpretiert ist. Formal lässt sich dies durch einen Test der Hypothesen

$$H_0 : (\beta_8, \beta_9) = (0, 0) \qquad \text{gegen} \qquad H_1 : (\beta_8, \beta_9) \neq (0, 0)$$

prüfen. Für den Likelihood-Quotienten-Test ergibt sich für die Teststatistik der Wert

$$lq = -2\{-1488.56 - (-1487.23)\} = 2.66\,.$$

Das 95% Quantil der (approximativen) $\chi^2(2)$-Verteilung ist $\chi^2_{95\%}(2) = 5.99$, so dass H_0 nicht abgelehnt werden kann. Dies folgt auch aus dem entsprechenden p-Wert von 0.269. Somit kann die vereinfachte Annahme eines linearen Effekts der Kovariable *aland* nicht abgelehnt werden. Zum gleichen Ergebnis kommt man auch mit dem Wald-Test.

Für die Log-Likelihood und das AIC-Kriterium ergibt sich im erweiterten Modell

$$l(\hat{\beta}) = -1487.232\,, \quad \text{AIC} = 2994.463.$$

Damit erhält man auch hier eine Tendenz zum vereinfachten Modell mit linearer Modellierung des Effekts der Variable *aland*.

△

Variable	Beschreibung	Mittelwert/ rel. Häufigkeit
konto1	1 = kein Konto 0 = gutes oder schwaches Konto	0.274
konto2	1 = gutes Konto 0 = kein oder schwaches Konto	0.394
laufz	Laufzeit in Monaten	20.903
hoehe	Kredithöhe in DM,	3.271
moral	frühere Zahlungsmoral 1 = gut 0 = schlecht	0.911
zweck	Verwendungszweck 1 = privat 0 = geschäftlich	0.657

Tabelle 4.6. *Kredit-Scoring: Beschreibung der Variablen.*

Beispiel 4.5 Kredit-Scoring

Bei der Vergabe von Krediten prüfen Banken die ,,Bonität" oder ,,Kreditwürdigkeit" der Kunden, also deren Bereitschaft und Fähigkeit den Kredit ordnungsgemäß zurück zu zahlen. Bei der Bonitätsprüfung (Kredit-Scoring) mit Hilfe statistischer Verfahren werden Merkmale des Kreditnehmers erfragt bzw. erhoben, die seine persönliche und wirtschaftliche Situation kennzeichnen und die Wahrscheinlichkeit der Bonität beeinflussen. Binäre Regressionsmodelle eignen sich zur Bonitätsprüfung, da sie die Wahrscheinlichkeit eines Kreditausfalls in Abhängigkeit von Kundenmerkmalen modellieren.

Wir verwenden einen Datensatz von $n = 1000$ Privatkrediten einer deutschen Bank, der in Fahrmeir et al. (1996) abgedruckt ist. Zu jedem Kunden gibt es die binäre Zielvariable y ,,Bonität", mit

$$y = \begin{cases} 1 & \text{nicht kreditwürdiger Kunde} \\ 0 & \text{kreditwürdiger Kunde.} \end{cases}$$

Von den insgesamt 20 Merkmalen verwenden wir die in Tabelle 4.6 beschriebenen als Kovariablen.

Wir modellieren die Wahrscheinlichkeit $\mathrm{P}(y = 1 \,|\, \boldsymbol{x})$ für schlechte Bonität durch ein Logit-Modell mit linearem Prädiktor

$$\beta_0 + \beta_1\, konto1 + \beta_2\, konto2 + \beta_3\, laufz + \beta_4\, hoehe + \beta_5\, moral + \beta_6\, zweck$$

Tabelle 4.7 gibt zum AIC-Wert 1043.815 die geschätzten Koeffizienten, die zugehörigen Standardabweichungen, t- und p-Werte an. Der p-Wert zur Variablen *hoehe* deutet darauf hin, dass der Effekt nicht signifikant ist.

Wir nehmen in einem zweiten Analyseschritt für die Effekte der metrischen Variablen *laufz* und *hoehe* jeweils ein quadratisches Polynom an, um eventuelle Nichtlinearitäten zu entdecken. Tabelle 4.8 enthält die dazugehörigen Schätzresultate. Alle p-Werte, auch die für die quadratischen Effekte, weisen nun auf Signifikanz hin. Auch der geringere AIC-Wert von 1035.371 deutet auf eine verbesserte Modellanpassung hin.

Abbildung 4.2 zeigt zum Vergleich die geschätzten linearen Effekte von Kredithöhe und Laufzeit und die nichtlinearen quadratischen Effekte. Aus dem ,,badewannenförmigen" Verlauf des quadratischen Effekts der Kredithöhe wird deutlich, dass niedrige und hohe Kredite das Risiko vergrößern, während durch die lineare Modellierung dieser Effekt

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	-0.284	0.303	-0.94	0.347
konto1	0.618	0.175	3.53	<0.001
konto2	-1.338	0.201	-6.65	<0.001
laufz	0.033	0.008	4.29	<0.001
hoehe	0.023	0.033	0.72	0.474
moral	-0.986	0.251	-3.93	<0.001
zweck	-0.426	0.266	-2.69	0.007

Tabelle 4.7. *Kredit-Scoring: Ergebnisse des Logit-Modells.*

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	-0.488	0.390	-1.25	0.211
konto1	0.618	0.176	3.51	<0.001
konto2	-1.337	0.202	-6.61	<0.001
laufz	0.092	0.025	3.64	<0.001
$laufz^2$	-0.001	<0.001	-2.20	0.028
hoehe	-0.264	0.099	-2.68	0.007
$hoehe^2$	0.023	0.007	3.07	0.002
moral	-0.995	0.255	-3.90	<0.001
zweck	-0.404	0.160	-2.52	0.012

Tabelle 4.8. *Kredit-Scoring: Ergebnisse des erweiterten Logit-Modells.*

Abb. 4.2. *Geschätzte lineare (- - -) und quadratische (—) Effekte von Kredithöhe und Laufzeit.*

verlorengeht. Der nichtlineare Effekt der Laufzeit erscheint dagegen moderater, so dass wir noch formal das Modell mit quadratischen Effekten gegen das Teilmodell mit linearen Effekten für Kredithöhe und Laufzeit testen. Für die Likelihood-Quotienten- und Wald-Statistik ergeben sich

$$lq = 12.44\,, \qquad w = 11.47$$

bei 2 Freiheitsgraden, und die dazugehörigen p-Werte sind 0.0020 bzw. 0.0032. Damit wird das Modell mit linearen Effekten deutlich abgelehnt.

△

4.2 Regression für Zähldaten

Zähldaten treten vor allem auf, wenn die Anzahl bestimmter Ereignisse in einer Zeitperiode oder Häufigkeiten in Kontingenztabellen analysiert werden sollen. Manchmal, vor allem wenn die Ereignisse sehr häufig auftreten, reicht zur Analyse eine Normalverteilungsapproximation aus. Wenn jedoch nur wenige kleine Werte als Anzahlen vorkommen, sind auch Modelle für mehrkategoriale Zielvariablen (Kapitel 5) eine mögliche Alternative. In der Regel sind jedoch diskrete Verteilungen, die spezielle Eigenschaften von Zähldaten berücksichtigen, am geeignetsten. Die Poisson-Verteilung ist einfach und wird am häufigsten verwendet, es existieren jedoch einige Modifikationen und Alternativen, wie zum Beispiel die negative Binomialverteilung, die in der Spezialliteratur zu Zähldatenmodellen detaillierter beschrieben sind.

4.2.1 Modelle für Zähldaten

Log-lineares Poisson-Modell

In diesem bekanntesten Modell für Zähldaten wird die Rate $\lambda_i = \mathrm{E}(y_i \,|\, \boldsymbol{x}_i)$ der Poisson-Verteilung mit dem Prädiktor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ über

$$\lambda_i = \exp(\eta_i) = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik}) \tag{4.19}$$

bzw. in log-linearer Form durch

$$\log(\lambda_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{4.20}$$

verbunden. Der Effekt von Kovariablen auf die Rate λ ist also exponentiell-multiplikativ, ähnlich wie für das relative Risiko $\pi/(1-\pi)$ im Logit-Modell, während der Effekt auf die logarithmierte Rate in (4.20) linear ist.

Lineares Poisson-Modell

Die direkte Verknüpfung

$$\lambda_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

ist dann nützlich, wenn die Kovariablen additiv auf die Rate wirken. Da $\boldsymbol{x}_i'\boldsymbol{\beta}$ nicht-negativ sein muss, impliziert dies im Allgemeinen Restriktionen für $\boldsymbol{\beta}$.

Überdispersion

Falls die Poisson-Verteilung angenommen wird, impliziert dies

$$\lambda_i = \mathrm{E}(y_i) = \mathrm{Var}(y_i).$$

Aus ähnlichen Gründen wie für binomiale Daten wird in Anwendungen oft eine signifikant größere empirische Varianz beobachtet. Deshalb ist es häufig zweckmäßig, einen Überdispersions-Parameter ϕ durch die Annahme

$$\mathrm{Var}(y_i \,|\, \boldsymbol{x}_i) = \phi\lambda_i$$

Poisson-Regression für Zähldaten

Daten

Die Zielvariablen y_i nehmen Werte aus $\{0, 1, 2, \dots\}$ an und sind bei gegebenen Kovariablen $x_{i1}, \dots, x_{ik}$ (bedingt) unabhängig.

Modelle

Log-lineares Poisson-Modell: $y_i \,|\, \boldsymbol{x}_i \sim Po(\lambda_i)$ mit

$$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \quad \text{bzw.} \quad \log \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Modell mit Überdispersion:

$$\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad \mathrm{Var}(y_i \,|\, \boldsymbol{x}_i) = \phi\lambda_i$$

mit Überdispersions-Parameter ϕ.

einzuführen. Wie für binomiale Daten gibt es komplexere Ansätze, die zusätzliche Variabilität von Zähldaten berücksichtigen. Eine Möglichkeit ist die Verwendung der negativen Binomialverteilung; dazu verwandt sind Modelle mit zufälligen Effekten, siehe Kapitel 6.

Beispiel 4.6 Anzahl der Zitate von Patenten

Wir illustrieren die Anwendung von Regressionsmodellen für Zähldaten anhand der in Beispiel 1.3 (Seite 8) beschriebenen Patentdaten. Anders als in den Beispielen 4.1 und 4.4 wählen wir jedoch als Zielvariable die Anzahl der Zitationen für ein Patent, also die Variable *azit*, und verwenden den gesamten Datensatz, d.h. Patente, die entweder zur Biotechnologie oder zur Pharma-Branche gehören. Als Kovariablen beziehen wir die restlichen Merkmale aus Tabelle 1.4 (Seite 8) ein. Dabei zentrieren wir die metrischen Kovariablen *jahr*, *aland* und *ansp* wie in Kapitel 3.2 beschrieben um ihr arithmetisches Mittel d.h. wir verwenden *jahrc*, *alandc* und *anspc* als Kovariablen im linearen Prädiktor. Die Variable *ansp* weist eine Reihe von Ausreißern mit sehr großen Werten auf, die zu stark verzerrten Ergebnissen führen können. Wir entfernen daher alle Beobachtungen mit $ansp > 60$ aus dem Datensatz und erhalten damit eine Stichprobe vom Umfang $n = 4832$.

In einem ersten Analyseschritt betrachten wir ein log-lineares Poisson-Modell für die Rate $\lambda_i = \mathrm{E}(azit_i)$ mit rein linearen Effekten im Prädiktor

$$\begin{aligned}\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 jahrc_i + \beta_2 alandc_i + \beta_3 anspc_i + \beta_4 biopharm_i + \\ + \beta_5 uszw_i + \beta_6 patus_i + \beta_7 patdsg_i + \beta_8 einspruch_i\,.\end{aligned}$$

Dabei schätzen wir das Modell mit einem echten Poisson-Modell (ohne Überdispersionsparameter) und mit einem Modell, das einen Überdispersionsparameter ϕ in die Varianz $\mathrm{Var}(azit_i) = \phi\lambda_i$ einbezieht. Um mögliche nichtlineare Effekte der metrischen Kovariablen zu berücksichtigen, setzen wir für *jahrc*, *alandc* und *anspc* anschließend Polynome an und vergleichen die verschiedenen Modelle mit dem AIC-Kriterium.

△

4.2.2 Schätzen und Testen: Likelihood-Inferenz

Wir nehmen wieder an, dass die Zielvariablen y_i (bedingt) unabhängig sind. Die Herleitung der Likelihood, Score-Funktion und Informationsmatrizen erfolgt in vollständiger Analogie zum vorangehenden Abschnitt.

Maximum-Likelihood-Schätzung

Für *Poisson-verteilte Zielvariablen* ist die diskrete Dichte (bzw. die Likelihood $L_i(\boldsymbol{\beta})$ der i-ten Beobachtung) durch

$$f(y_i \,|\, \boldsymbol{\beta}) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}, \quad \mathrm{E}(y_i \,|\boldsymbol{x}_i) = \lambda_i$$

gegeben und hängt über $\lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ im linearen Poisson-Modell und über $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ im log-linearen Poisson-Modell von $\boldsymbol{\beta}$ ab. Somit ist die *Log-Likelihood* bis auf die additive, von $\boldsymbol{\beta}$ unabhängige Konstante $-n\log(y_i!)$ durch

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \log(\lambda_i) - \lambda_i)$$

gegeben. Daraus ergibt sich für das *log-lineare Poisson-Modell* mit $\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i(\boldsymbol{x}_i'\boldsymbol{\beta}) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i\eta_i - \exp(\eta_i)) \,.$$

Differenzieren nach der Kettenregel $\partial l_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = (\partial l_i/\partial\eta_i) \cdot \partial\eta_i/\partial\boldsymbol{\beta} = \partial l_i/\partial\eta_i \cdot \boldsymbol{x}_i$ ergibt die *Score-Funktion*

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \exp(\eta_i)) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \lambda_i)$$

und mit den gleichen Argumenten wie beim Logit-Modell die *Fisher-Information*

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta})) = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\lambda_i,$$

wobei man $\mathrm{E}(y_i - \lambda_i)^2 = \mathrm{Var}(y_i) = \lambda_i$ verwendet.
Für das *lineare Poisson-Modell* $\lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i$ erhalten wir

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \log(\eta_i) - \eta_i)$$

und daraus

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i\frac{1}{\eta_i} - 1) = \sum_{i=1}^{n} \boldsymbol{x}_i\frac{1}{\lambda_i}(y - \lambda_i),$$

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \frac{1}{\lambda_i}.$$

In beiden Poisson-Modellen ergibt sich für

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$$

wegen $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ bzw. $\lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ ein nichtlineares Gleichungssystem. Die Fisher-Matrix hängt ebenfalls von $\boldsymbol{\beta}$ ab.
Die numerische Bestimmung von $\hat{\boldsymbol{\beta}}$ erfolgt mit den entsprechenden Ausdrücken für $\boldsymbol{s}(\boldsymbol{\beta})$ und $\boldsymbol{F}(\boldsymbol{\beta})$ durch das Fisher-Scoring-Verfahren (4.16). Wie im Fall der linearen und binären Regression nehmen wir für die Designmatrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$

$$\mathrm{rg}(\boldsymbol{X}) = p$$

an. Bezüglich der Konvergenz bzw. Divergenz der Iterationen gelten die gleichen Bemerkungen wie in Abschnitt 4.1.2. Ebenso gilt unter schwachen Annahmen asymptotisch für $n \to \infty$ bzw. approximativ für großes n

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}))$$

mit der geschätzten Kovarianzmatrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$.

Testen linearer Hypothesen

Für das Testen linearer Hypothesen $\boldsymbol{C\beta} = \boldsymbol{d}$ werden die gleichen Teststatistiken wie für binäre Regressionsmodelle verwendet, wobei nur $l(\boldsymbol{\beta}), \boldsymbol{s}(\boldsymbol{\beta})$ und $\boldsymbol{F}(\boldsymbol{\beta})$ durch die entsprechenden Ausdrücke für Poisson-Modelle zu ersetzen sind. Ebenso gelten die Aussagen zur asymptotischen bzw. approximativen χ^2-Verteilung der Teststatistiken.

Kriterien zur Modellanpassung und Modellwahl

Auch die in Abschnitt 4.1.2 angegebenen Kriterien sind in Analogie zu binären Regressionsmodellen definiert. Da für die Poisson-Verteilung $\mathrm{Var}(y_i) = \lambda_i$ gilt, erhält man für die Pearson-Statistik

$$\chi^2 = \sum_{i=1}^{G} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i / n_i}.$$

Für die Devianz und das AIC-Kriterium ist die Poisson-Log-Likelihood zu verwenden, wobei $0 \cdot \log(0) = 0$ (für $y_i = 0$) gesetzt wird.

Schätzung des Überdispersions-Parameters

Falls eine mögliche Überdispersion durch die Annahme $\mathrm{Var}(y_i \,|\, x_i) = \phi\lambda_i$ berücksichtigt wird, so schätzt man ϕ durch die gemittelte Pearson-Statistik bzw. gemittelte Devianz:

$$\hat{\phi}_P = \frac{1}{n-p}\chi^2 \quad \text{bzw.} \quad \hat{\phi}_D = \frac{1}{n-p} D\,.$$

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	0.176	0.032	5.53	<0.001
jahrc	-0.075	0.003	-25.86	<0.001
alandc	-0.025	0.004	-6.12	<0.001
anspc	0.020	0.001	17.12	<0.001
biopharm	0.294	0.031	9.41	<0.001
uszw	-0.043	0.025	-1.77	0.077
patus	0.018	0.026	0.69	0.491
patdsg	-0.230	0.031	-7.31	<0.001
einspruch	0.404	0.024	16.55	<0.001

Tabelle 4.9. *Anzahl der Zitate von Patenten: Modell mit linearen Effekten,* $\phi = 1$.

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	0.176	0.060	2.91	0.004
jahrc	-0.075	0.005	-13.60	<0.001
alandc	-0.025	0.008	-3.22	0.001
anspc	0.020	0.002	9.01	<0.001
biopharm	0.294	0.059	4.95	<0.001
uszw	-0.043	0.047	-0.93	0.353
patus	0.018	0.050	0.36	0.717
patdsg	-0.230	0.060	-3.84	<0.001
einspruch	0.404	0.046	8.70	<0.001

Tabelle 4.10. *Anzahl der Zitate von Patenten: Modell mit linearen Effekten und Überdispersion.*

Dies entspricht der Schätzung der Fehlervarianz im linearen Modell, wobei χ^2 bzw. D die Rolle der Residuenquadratsumme einnehmen.

Die geschätzte Kovarianzmatrix ist dann mit $\hat{\phi}$ zu multiplizieren, d.h. es gilt $\widehat{\mathrm{Cov}}(\hat{\beta}) = \hat{\phi} F^{-1}(\hat{\beta})$. Diese Art der Berücksichtigung und Schätzung von Überdispersion entspricht streng genommen nicht mehr einem echten Likelihood-Ansatz, sondern einem Quasi-Likelihood-Modell (vgl. Abschnitt 4.5).

Beispiel 4.7 Anzahl der Zitate von Patenten

Für das log-lineare Poisson-Modell des Beispiels 4.6 mit rein linearen Effekten und ohne Überdispersion (d.h. $\phi = 1$) zeigt Tabelle 4.9 die Schätzergebnisse. (AIC = 20584.77, Devianz= 13465.093, Pearson-χ^2 = 17438.430)

Die p-Werte scheinen auf die Signifikanz aller Kovariablen mit Ausnahme von *uszw* und *patus* hinzudeuten. Da bei Poisson-Modellen für Zähldaten Überdispersion eher die Regel als die Ausnahme ist, reanalysieren wir das Modell, indem wir die Überdispersion durch die gemittelte Pearson-Statistik oder Devianz schätzen (vgl. auch Abschnitt 4.4). Wir erhalten

$$\hat{\phi}_P = \frac{1}{n-p}\chi^2 = 3.616 \quad \text{bzw.} \quad \hat{\phi}_D = \frac{1}{n-p}D = 2.792\,,$$

mit $n = 4832, p = 9$. Somit wird die geschätzte Varianz bzw. der Standardfehler der geschätzten Regressionsparameter im Vergleich zum Poisson-Modell mit $\hat{\phi}$ bzw. $\hat{\phi}^{1/2}$ multipliziert. Die Punktschätzer selbst sind die gleichen wie für das reine Poisson-Modell

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	0.516	0.075	6.88	<0.001
jahrc	39.081	5.750	6.80	<0.001
$jahrc^2$	-0.010	0.001	-6.81	<0.001
alandc	-0.244	0.098	-2.49	0.013
$alandc^2$	0.037	0.013	2.92	0.004
$alandc^3$	-0.002	<0.001	-3.42	0.001
anspc	0.040	0.007	5.59	<0.001
$anspc^2$	-0.001	<0.001	-2.91	0.004
biopharm	0.200	0.062	3.22	0.001
uszw	-0.048	0.048	-1.00	0.320
patus	-0.003	0.052	-0.06	0.952
patdsg	-0.234	0.062	-3.77	<0.001
einspruch	0.401	0.047	8.45	<0.001

Tabelle 4.11. *Anzahl der Zitate von Patenten: Erweitertes Modell mit Überdispersion.*

ohne Überdispersion. Tabelle 4.10 zeigt die Ergebnisse für $\hat{\phi}_P$. Im Vergleich zu Tabelle 4.9 vergrößern sich die Standardfehler um den Faktor $\hat{\phi}_P^{1/2} = 1.902$. Dies bewirkt eine Erhöhung der p-Werte, so dass der Effekt der Variable *uszw* nun deutlich nichtsignifkant ist, während die Analyse ohne Überdispersion einen p-Wert nahe an der Grenze zur Signifikanz ergab.

Um mögliche nichtlineare Effekte der zentrierten metrischen Kovariablen *jahrc*, *alandc* und *anspc* zu erkennen, wurden dafür Polynome verschiedenen Grades angesetzt. Die Zentrierung wird dabei wie in Abschnitt 3.1.4 vorgenommen. Das beste Modell, mit einem gegenüber dem rein linearen Modell deutlich reduzierten AIC ergab sich für einen Prädiktor mit quadratischen Polynomen für *jahrc* und *anspc* sowie einem kubischen Polynom für *alandc*. (AIC = 20196.62, Devianz= 13068.937, Pearson-χ^2 = 18335.606, $\hat{\phi}_D = 2.712$, $\hat{\phi}_P = 3.805$). Tabelle 4.11 zeigt die Schätzergebnisse. Die Variablen *uszw* und *patus* erweisen sich weiterhin als nichtsignifikant. Zugleich erhöhen sich durch die größere Anzahl zu schätzender Parameter auch die Standardfehler von einigen anderen geschätzten Effekten; die Variablen bleiben jedoch alle (auf dem 5%-Niveau) signifikant. Abbildung 4.3 vergleicht lineare und nichtlineare Effekte der metrischen Kovariablen. Die Interpretation der Modelle überlassen wir dem Leser.

△

4.3 Modelle für positive stetige Zielvariablen

Das klassische lineare Modell

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \;\; \mathrm{E}(\varepsilon_i) = 0, \;\; \mathrm{Var}(\varepsilon_i) = \sigma^2$$

ist besonders dann zur Analyse geeignet, wenn die Fehler ε_i (zumindest approximativ) normalverteilt sind. Dann gilt, dass die Zielvariablen y_i bei gegebenem Kovariablenvektor $\boldsymbol{x}_i$ (bedingt) unabhängig und normalverteilt sind mit

$$y_i \,|\, \boldsymbol{x}_i \sim \mathrm{N}(\mu_i, \sigma^2), \;\; \mu_i = \mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Abb. 4.3. *Anzahl der Zitate von Patenten: Lineare (- - -) und nichtlineare (—) Effekte der metrischen Kovariablen jahr, aland und ansp.*

In vielen Anwendungen, zum Beispiel bei der Analyse von Lebensdauern, Schadenshöhen, Kosten, usw., können die Zielvariablen keine negativen Werte annehmen und sind zudem schief, meist rechtsschief, verteilt. Damit sind die Zielvariablen deutlich nicht-normal verteilt. Um dennoch die Methodik linearer Modelle einsetzen zu können, wird die Zielvariable y häufig logarithmiert und für $\tilde{y} = \log(y)$ wird ein lineares Modell mit Normalverteilungsannahme angesetzt, d.h.

$$\tilde{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \text{bzw.} \quad \tilde{y}_i \,|\, \boldsymbol{x}_i \sim \mathrm{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2).$$

Für die ursprüngliche Variable y bedeutet dies, dass sie lognormalverteilt ist (vergleiche Definition B.2 Anhang B) mit

$$\mathrm{E}(y_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma^2/2), \;\; \mathrm{Var}(y_i) = \exp(2\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma^2)(\exp(\sigma^2) - 1) \tag{4.21}$$

Durch Einsetzen des KQ-Schätzers $\hat{\boldsymbol{\beta}}$ und der geschätzten Varianz $\hat{\sigma}^2$ für das lineare Modell kann man daraus „plug-in" Schätzer für (4.21) erhalten. Durch die Rücktransformation mit der Exponentialfunktion können dabei deutliche Verzerrungen bei der Schätzung $\hat{\mu}_i = \exp(\hat{\eta}_i) = \exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}^2/2)$ auftreten.

Gamma-Regression

Eine deshalb unter Umständen zu bevorzugende Alternative besteht darin für die Zielvariablen $y_i \mid \boldsymbol{x}_i$ eine Gammaverteilung mit Erwartungswert $\mathrm{E}(y_i) = \mu_i$ und Skalenparameter ν anzunehmen. Für die Varianz $\mathrm{Var}(y_i) = \sigma_i^2$ folgt dann (vgl. Appendix B.3)

$$\mathrm{Var}(y_i) = \mu_i^2/\nu.$$

Für die nichtnegativen, gammaverteilten Zielvariablen gilt $\mathrm{E}(y_i) = \mu_i > 0$. Daher ist ein direkter *linearer* Ansatz

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

wieder problematisch, da die Bedingung $\boldsymbol{x}_i'\boldsymbol{\beta} > 0$ eingehalten werden muss.
Geeigneter ist ein *multiplikativ-exponentielles Modell*

$$\mu_i = \exp(\eta_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik}) \tag{4.22}$$

mit Responsefunktion $h(\eta) = \exp(\eta)$.
Eine andere, auch in bekannten Programmpaketen mögliche Wahl der Response- bzw. Linkfunktion ist die *reziproke* Verknüpfung

$$\mu_i = \frac{1}{\eta_i} = \frac{1}{\boldsymbol{x}_i'\boldsymbol{\beta}}.$$

Da wieder $\boldsymbol{x}_i'\boldsymbol{\beta} > 0$ gelten muss, impliziert diese Wahl wieder Restriktionen an $\boldsymbol{\beta}$. Die *reziproke Responsefunktion* ist zwar die so genannte *natürliche* oder *kanonische Responsefunktion* für die Gammaverteilung (vgl. Abschnitt 4.4), zur Modellierung und Interpretation ist das multiplikativ-exponentielle Modell (4.22) jedoch meist besser geeignet.

Inverse Gauß-Verteilung

Diese Verteilung eignet sich ebenfalls für Regressionsanalysen mit unsymmetrischen und nichtnegativen stetigen Zielvariablen, siehe z.B. Jorgensen (1982).

4.4 Generalisierte Lineare Modelle

4.4.1 Allgemeine Modelldefinition

Lineare Modelle und die in den vorangehenden Abschnitten beschriebenen Regressionsansätze für nicht-normal verteilte Zielvariablen besitzen gemeinsame Eigenschaften, die sich in einheitlicher Darstellung beschreiben lassen:

1. Der Erwartungswert $\mu = \mathrm{E}(y \mid \boldsymbol{x})$ der Zielvariablen y wird mit dem linearen Prädiktor $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ durch eine Responsefunktion h bzw. eine Linkfunktion $g = h^{-1}$ verknüpft:

$$\mu = h(\eta) \quad \text{bzw.} \quad \eta = g(\mu).$$

2. Die Verteilungen der Zielvariablen (Normal-, Binomial-, Poisson- und Gammaverteilung) lassen sich alle in Form einer *einparametrischen Exponentialfamilie* schreiben:

Exponentialfamilie

Die Dichte einer einparametrischen Exponentialfamilie für die Zielvariable y ist durch

$$f(y\,|\,\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi}\omega + c(y,\phi,\omega)\right) \tag{4.23}$$

gegeben. Der Parameter θ heißt *natürlicher* oder *kanonischer* Parameter. Die Funktion $b(\theta)$ muss die Bedingung erfüllen, dass sich $f(y\,|\,\theta)$ normieren lässt und erste und zweite Ableitungen $b'(\theta)$ und $b''(\theta)$ existieren. Der zweite Parameter ϕ ist ein Dispersionsparameter, während ω ein bekannter Wert ist (üblicherweise ein Gewicht).

Damit können die Definition generalisierter linearer Modelle und die zugehörige statistische Inferenz in einem einheitlichen Rahmen erfolgen. Grundsätzlich lassen sich die Konzepte jedoch auch auf Regressionsprobleme mit Verteilungen, die nicht zur Exponentialfamilie gehören, übertragen. Um den Fall von Individualdaten und gruppierten Daten einheitlich zu behandeln, erweitern wir die übliche Definition durch einen Gewichtsfaktor ω. Dann gilt $\omega = 1$ für Individualdaten, während ω gleich dem Umfang einer Gruppe ist, wenn für y das arithmetische Mittel der individuellen Zielvariablen gewählt wird.

Die Bernoulli- und die Poisson-Verteilung besitzen keinen Dispersionsparameter; es gilt $\phi = 1$. Für die Normalverteilung gilt $\phi = \sigma^2$. In jedem Fall ist θ der Parameter von primärem Interesse, der mit dem linearen Prädiktor $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ verknüpft wird, während ϕ ein Störparameter („nuisance-Parameter") und von sekundärem Interesse ist. Der Term $c(y,\phi,\omega)$ hängt nicht von θ ab. Man kann zeigen, dass

$$\mathrm{E}(y) = \mu = b'(\theta), \quad \mathrm{Var}(y) = \phi\, b''(\theta)/\omega$$

gilt.

Beispiel 4.8 Bernoulli-, Poisson- und Normalverteilung

1. Bernoulliverteilung: Eine *Bernoulli-Variable* hat die Wahrscheinlichkeitsfunktion bzw. (diskrete) Dichte

$$f(y\,|\,\pi) = \mathrm{P}(Y = y) = \pi^y(1-\pi)^{1-y}, \quad y = 0, 1\,,$$

wobei $\mathrm{P}(Y = 1) = \pi = \mathrm{E}(Y) = \mu$ und $\mathrm{Var}(Y) = \pi(1-\pi)$ gilt. Logarithmieren ergibt

$$\log(f(y\,|\,\pi)) = y\log(\pi) - y\log(1-\pi) + \log(1-\pi).$$

Definiert man $\theta = \log(\pi) - \log(1-\pi) = \log(\pi/(1-\pi))$ als natürlichen Parameter und beachtet $\log(1-\pi) = -\log(1+\exp(\theta))$, so erhält man die Dichte in der Form einer Exponentialfamilie

$$f(y\,|\,\theta) = \exp(y\theta - \log(1+\exp(\theta)).$$

mit $b(\theta) = \log(1+\exp(\theta))$, $\phi = 1$ und $c = 0$. Differenzieren ergibt $b'(\theta) = \exp(\theta)/(1+\exp(\theta))$ und $b''(\theta) = \exp(\theta)/(1+\exp(\theta))^2$. Auflösen von $\theta = \log(\pi/(1-\pi))$ nach π ergibt

(a) Dichte

$$f(y \mid \theta, \phi, w) = \exp\left(\frac{y\theta - b(\theta)}{\phi}\omega + c(y, \phi, \omega)\right)$$

(b) Exponentialfamilien

Verteilung		$\theta(\mu)$	$b(\theta)$	ϕ
Normal	$\mathrm{N}(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2
Bernoulli	$\mathrm{B}(1, \pi)$	$\log(\pi/(1-\pi))$	$\log(1+\exp(\theta))$	1
Poisson	$\mathrm{Po}(\lambda)$	$\log(\lambda)$	$\exp(\theta)$	1
Gamma	$\mathrm{G}(\mu, \nu)$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}
Inverse Gauß	$\mathrm{IG}(\mu, \sigma^2)$	$1/\mu^2$	$-(-2\theta)^{1/2}$	σ^2

(c) Erwartungswert und Varianz

Verteilung	$\mathrm{E}(y) = b'(\theta)$	$b''(\theta)$	$\mathrm{Var}(y) = b''(\theta)\phi/\omega$
Normal	$\mu = \theta$	1	σ^2/ω
Bernoulli	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	$\pi(1-\pi)/\omega$
Poisson	$\lambda = \exp(\theta)$	λ	λ/ω
Gamma	$\mu = 1 - 1/\theta$	μ^2	$\mu^2\nu^{-1}/\omega$
Inverse Gauß	$\mu = (-2\theta)^{-1/2}$	μ^3	$\mu^3\sigma^2/\omega$

Tabelle 4.12. *Einfache Exponentialfamilien.*

$$\pi = \exp(\theta)/(1+\exp(\theta))\,,$$

so dass

$$\mathrm{E}(y) = b'(\theta) = \pi, \quad \mathrm{Var}(y) = b''(\theta) = \pi(1-\pi)$$

gilt.

2. Poisson-Verteilung: Eine *Poisson-Variable* hat die (diskrete) Dichte

$$f(y \mid \lambda) = \mathrm{P}(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}, \quad y = 0, 1, \ldots$$

Logarithmieren dieser Dichte ergibt

$$\log(f(y \mid \lambda)) = y\log(\lambda) - \lambda - \log(y!).$$

Mit $\theta = \log(\lambda)$ als natürlichem Parameter erhält man

$$\log(f(y\,|\,\theta)) = y\theta - \exp(\theta) - \log(y!).$$

Damit ist $b(\theta) = \exp(\theta) = \lambda$, $\phi = 1$ und $c(y, \phi) = -\log(y!)$. Mit $b'(\theta) = b''(\theta) = \exp(\theta) = \lambda$ erhält man

$$\mathrm{E}(y) = \mu = \lambda, \quad \mathrm{Var}(y) = \lambda,$$

also die für Poisson-Variablen charakteristische Gleichheit von Erwartungswert und Varianz.

3. Normalverteilung: Die Dichte der *Normalverteilung* ist

$$f(y\,|\,\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right],$$

wobei $\mu = \mathrm{E}(y)$ der Parameter von Interesse ist und $\sigma^2 = \mathrm{Var}(y)$ als Störparameter betrachtet wird. Die Dichte lässt sich mit

$$\begin{aligned} f(y\,|\,\mu) &= \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \\ &= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \end{aligned}$$

in die Form einer Exponentialfamilie mit $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \mu^2/2 = \theta^2/2$ und $c(y, \theta) = -y^2/(2\sigma^2) - 0.5\log(2\pi\sigma^2)$ bringen. Man erhält, wie es sein muss, wieder

$$b'(\theta) = \theta = \mu = \mathrm{E}(y), \quad b''(\theta) = 1$$

und somit

$$\mathrm{Var}(y) = \phi b''(\theta) = \sigma^2.$$

△

In analoger Weise erhält man die in Tabelle 4.12 zusammengefassten weiteren Beziehungen.

Ein spezifisches generalisiertes lineares Modell ist somit durch den Typ der Exponentialfamilie (Gauß-, Binomial-, Poisson-, Gammaverteilung usw.), durch die Wahl der Link- oder Responsefunktion und die Definition und Auswahl der Kovariablen vollständig charakterisiert.

Die Wahl einer geeigneten Response- oder Linkfunktion hängt wie in den vorangehenden Beispielen vom Typ der Zielvariablen ab. Zu jeder Exponentialfamilie gibt es eine so genannte *kanonische* (oder *natürliche*) Linkfunktion. Diese ist dadurch definiert, dass $\theta_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ gilt. Gemäß Tabelle 4.12 entspricht bei einer Normalverteilung das lineare Modell $\mu_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ der natürlichen Linkfunktion, bei binären Regressionsmodellen das Logit-Modell und bei Poisson-Modellen das log-lineare Modell.

4.4.2 Likelihood-Inferenz

In allgemeiner Form ist der Log-Likelihood-Beitrag einer Beobachtung $(y_i, \boldsymbol{x}_i)$ bis auf eine additive Konstante durch

$$l_i(\boldsymbol{\beta}) = \log(f(y_i\,|\,\boldsymbol{\beta})) = \frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i \tag{4.24}$$

gegeben, wobei der natürliche Parameter θ_i der Exponentialfamilie über

Generalisierte lineare Modelle (GLM)

Verteilungsannahme

Für gegebene Kovariablen $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$ sind die Zielvariablen (bedingt) unabhängig und die (bedingte) Dichte y_i gehört zu einer Exponentialfamilie mit

$$f(y_i \,|\, \theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i - c(y_i, \phi, \omega_i)\right).$$

Dabei ist θ_i der natürliche Parameter und ϕ ein gemeinsamer, von i unabhängiger Dispersionsparameter. Für $\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \mu_i$ und $\mathrm{Var}(y_i \,|\, \boldsymbol{x}_i)$ gilt somit

$$\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \mu_i = b'(\theta_i), \quad \mathrm{Var}(y_i \,|\, \boldsymbol{x}_i) = \sigma_i^2 = \phi b''(\theta_i)/\omega_i.$$

Das Gewicht ω_i ist 1 für ungruppierte Daten ($i = 1, \ldots, n$) und gleich n_i für gruppierte Daten ($i = 1, \ldots, G$), wenn das arithmetische Mittel der Zielvariablen von Gruppe i als deren Zielvariable y_i gewählt wird (bzw. $\omega_i = 1/n_i$, wenn die Summe der individuellen Zielvariablen gewählt wird).

Strukturannahme

Der (bedingte) Erwartungswert μ_i ist mit dem linearen Prädiktor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ durch

$$\mu_i = h(\eta_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta}) \quad bzw. \quad \eta_i = g(\mu_i)$$

verknüpft, wobei

h eine (eineindeutige und zweimal differenzierbare) *Responsefunktion* und g die *Linkfunktion*, d.h. die Umkehrfunktion $g = h^{-1}$ von h ist.

$$\mu_i = b'(\theta_i), \quad \mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$$

von $\boldsymbol{\beta}$ abhängt. Wegen der (bedingten) Unabhängigkeit der $y_i \,|\, \boldsymbol{x}_i$ ist auch allgemein

$$l(\boldsymbol{\beta}) = \sum l_i(\boldsymbol{\beta})$$

die Log-Likelihood der Stichprobe. Um Individualdaten ($i = 1, \ldots, n$) und gruppierte Daten ($i = 1, \ldots, G$) simultan zu behandeln, lassen wir hier n oder G als obere Grenze in den Summationszeichen weg. Für die *Score-Funktion* $\boldsymbol{s}(\beta) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ lässt sich mit den Eigenschaften von Exponentialfamilien

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i \frac{d_i}{\sigma_i^2}(y_i - \mu_i) \tag{4.25}$$

zeigen. Dabei hängen $\mu_i = \mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ und $\sigma_i^2 = \mathrm{Var}(y_i \,|\, \boldsymbol{x}_i)$ von $\boldsymbol{\beta}$ ab; ebenso die 1. Ableitung

$$d_i = \frac{dh(\eta_i)}{d\eta_i}$$

der Responsefunktion, ausgewertet an der Stelle $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. Da $\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \mu_i$, gilt $\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathbf{0}$, und die *ML-Gleichungen* sind durch

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{4.26}$$

gegeben. Für die Fisher-Matrix $\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta}))$ erhält man daraus

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i\boldsymbol{x}_i' w_i, \tag{4.27}$$

mit den von $\boldsymbol{\beta}$ abhängigen (skalaren) Gewichten

$$w_i = d_i^2/\sigma_i^2.$$

Man verifiziert leicht, dass sich für das Logit-Modell und das Poisson-Modell die abgeleiteten Formen als Spezialfall mit $\mu_i = \pi_i$ bzw. $\mu_i = \lambda_i$ ergeben.

Für eine kompaktere Schreibweise und für den Vergleich mit linearen Modellen ist eine Matrizendarstellung nützlich. Mit

$$\boldsymbol{y} = (y_1, \ldots, y_n)', \quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)',$$

der Designmatrix

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

und den Diagonalmatrizen

$$\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_n), \quad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2), \quad \boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_n)$$

ergibt sich

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \quad \boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X},$$

wobei $\boldsymbol{D}, \boldsymbol{\Sigma}$ und $\boldsymbol{W}$ von $\boldsymbol{\beta}$ abhängen. Für das lineare Regressionsmodell $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ erhält man speziell die bekannten Formen

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu}), \quad \boldsymbol{F}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}.$$

Für die Designmatrix $\boldsymbol{X}$ setzen wir wieder

$$\mathrm{rg}(\boldsymbol{X}) = p$$

voraus.

Die numerische Bestimmung des ML-Schätzers $\hat{\boldsymbol{\beta}}$ erfolgt üblicherweise durch den Fisher-Scoring-Algorithmus (4.16). Mit der obigen Matrizenschreibweise lassen sich die Iterationen auch in der Form einer *iterativ gewichteten KQ-Schätzung*

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(k)}\tilde{\boldsymbol{y}}^{(k)}, \quad k = 0, 1, 2, \ldots \tag{4.28}$$

schreiben. Dabei ist $\tilde{\boldsymbol{y}}^{(k)} = (\ldots, \tilde{y}_i(\hat{\boldsymbol{\beta}}^{(k)}), \ldots)'$ ein „Arbeitsvektor" mit den Elementen

Maximum-Likelihood-Schätzung in GLM

Definition

Der ML-Schätzer $\hat{\boldsymbol{\beta}}$ maximiert die (Log-) Likelihood und wird als Nullstelle

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

der Score-Funktion

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i \frac{d_i}{\sigma_i^2}(y_i - \mu_i) = \boldsymbol{X}'\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

bezeichnet. Die Fisher-Matrix ist

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i \boldsymbol{x}_i' w_i = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

mit $w_i = d_i^2/\sigma_i^2$ und der Gewichtsmatrix $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_n)$.

Numerische Berechnung

Der ML-Schätzer $\hat{\boldsymbol{\beta}}$ wird iterativ berechnet durch Fisher-Scoring in Form einer *iterativ gewichteten KQ-Schätzung*

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(k)}\tilde{\boldsymbol{y}}^{(k)}, \quad k = 0, 1, 2, \ldots$$

mit dem Vektor $\tilde{\boldsymbol{y}}^{(k)} = (\ldots, \tilde{y}_i(\hat{\boldsymbol{\beta}}^{(k)}), \ldots)'$ von *Arbeitsbeobachtungen* $\tilde{y}_i(\hat{\boldsymbol{\beta}}^{(k)}) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(k)} + d_i^{-1}(\hat{\boldsymbol{\beta}}^{(k)})(y_i - \hat{\mu}_i(\hat{\boldsymbol{\beta}}^{(k)}))$ und $\boldsymbol{W}^{(k)} = \boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})$.

$$\tilde{y}_i(\hat{\boldsymbol{\beta}}^{(k)}) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(k)} + d_i^{-1}(\hat{\boldsymbol{\beta}}^{(k)})(y_i - \hat{\mu}_i(\hat{\boldsymbol{\beta}}^{(k)})),$$

und $\boldsymbol{W}^{(k)}$ ist die Gewichtsmatrix, ausgewertet für $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(k)}$. Somit können in jedem Schritt effiziente Algorithmen zur Berechnung (gewichteter) KQ-Schätzer in linearen Modellen eingesetzt werden.

Dabei haben wir wieder vorausgesetzt, dass die Informationsmatrix $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ invertierbar ist. Da die Elemente w_i der Diagonalmatrix $\boldsymbol{W}$ von den Kovariablen $\boldsymbol{x}_i$ und von $\boldsymbol{\beta}$ abhängen, folgt allgemein aus dem vollen Rang von $\boldsymbol{X}$ und der Invertierbarkeit von $\boldsymbol{X}'\boldsymbol{X}$ nicht die Invertierbarkeit von $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$. In der Regel sind jedoch (fast) alle Gewichte positiv, so dass $\boldsymbol{F}(\boldsymbol{\beta})$ invertierbar ist. Dies setzen wir im Weiteren voraus. In der Regel konvergiert dann der Algorithmus und stoppt nach einigen Iterationsschritten gemäß dem Abbruchkriterium in der Nähe eines Maximums. Für natürliche Linkfunktionen lässt sich zeigen, dass das erreichte Maximum dann auch eindeutig ist. Allgemein gilt dies jedoch nicht, so dass man dann mit verschiedenen Anfangswerten starten sollte.

Asymptotische Eigenschaften des ML-Schätzers

Bei Regressoren mit kompaktem Wertebereich reicht bei natürlichen Linkfunktionen $(\boldsymbol{X}'\boldsymbol{X})^{-1} \to \mathbf{0}$ bzw. $\lambda_{min}(\boldsymbol{X}'\boldsymbol{X}) \to \infty$ für die schwache Konsistenz und asymptotische

Normalität aus, bei nicht-natürlichen Linkfunktionen muss $\boldsymbol{X}'\boldsymbol{X}$ noch etwas strengere Wachstumsbedingungen erfüllen. Dabei bezeichnet λ_{min} den kleinsten Eigenwert von $\boldsymbol{X}'\boldsymbol{X}$. Sind im Fall stochastischer Regressoren die Beobachtungen $(y_i, \boldsymbol{x}_i)$ unabhängig und identisch wie $(y, \boldsymbol{x})$ verteilt und erfüllen die Annahmen eines verallgemeinerten linearen Modells, so gilt unter schwachen Zusatzannahmen über die Randverteilung von $\boldsymbol{x}$ ebenfalls asymptotische Normalität.

Unter den gleichen Annahmen existiert $\hat{\boldsymbol{\beta}}_n$ asymptotisch mit Wahrscheinlichkeit 1, d.h. es gilt

$$\lim_{n\to\infty} \mathrm{P}(\hat{\boldsymbol{\beta}}_n \text{ existiert}) = 1,$$

Details und allgemeine Beweise finden sich in Fahrmeir & Kaufmann (1985).

Die Inverse der Fisher-Informationsmatrix $\boldsymbol{F}(\boldsymbol{\beta})$ (4.27), ausgewertet an der Stelle des ML-Schätzers $\hat{\boldsymbol{\beta}}$, ist wieder die asymptotische bzw. approximative Kovarianzmatrix $\boldsymbol{A} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$ des Schätzers $\hat{\boldsymbol{\beta}}$. Das Diagonalelement a_{jj} ist ein Schätzer für die Varianz $\sigma_j^2 = \mathrm{Var}(\hat{\boldsymbol{\beta}}_j)$ der r-ten Komponente und $\sqrt{a_{jj}}$ für die Standardabweichung σ_j.

Asymptotische Eigenschaften des ML-Schätzers

Deutlichkeitshalber schreiben wir $\hat{\boldsymbol{\beta}}_n$ für den ML-Schätzer zum Stichprobenumfang n. Unter Regularitätsbedingungen ist $\hat{\boldsymbol{\beta}}_n$ konsistent und asymptotisch normalverteilt:

$$\boldsymbol{F}^{\frac{1}{2}}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}),$$

bzw. in approximativer Form geschrieben

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\boldsymbol{\beta})).$$

Dieses Ergebnis bleibt richtig, wenn man $\boldsymbol{F}(\boldsymbol{\beta})$ durch den Schätzer $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$ ersetzt.

Schätzung des Skalierungs- oder Überdispersionsparameters

Falls der Dispersionsparameter ϕ unbekannt ist, kann er durch

$$\hat{\phi} = \frac{1}{G-p} \sum_{i=1}^{G} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/n_i}$$

konsistent geschätzt werden. Dabei sind p die Anzahl der Regressionsparameter, $\hat{\mu}_i = h(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$ der geschätzte Erwartungswert und $v(\hat{\mu}_i)$ die geschätzte Varianzfunktion, und die Daten sind so weit wie möglich zu gruppieren. In allen Ausdrücken, in denen ϕ vorkommt, etwa in $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$, ist ϕ durch $\hat{\phi}$ zu ersetzen.

Testen linearer Hypothesen

Für das Testen linearer Hypothesen

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{gegen} \qquad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d},$$

Testen linearer Hypothesen

Hypothese

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \quad \text{gegen} \quad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d}.$$

Teststatistiken

1. *Likelihood-Quotienten-Statistik:* $lq = -2\{l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\}$.
2. *Wald-Statistik:* $w = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{CF}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$
3. *Score-Statistik:* $u = \boldsymbol{s}'(\tilde{\boldsymbol{\beta}})\boldsymbol{F}^{-1}(\tilde{\boldsymbol{\beta}})\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$.

Dabei ist $\tilde{\boldsymbol{\beta}}$ der ML-Schätzer unter der Restriktion H_0.

Testentscheidungen

Für großes n gilt unter H_0 approximativ:

$$lq, w, u \overset{a}{\sim} \chi^2_r.$$

Dabei ist r der Rang von $\boldsymbol{C}$. H_0 wird abgelehnt, falls

$$lq, w, u > \chi^2_r(1-\alpha).$$

wobei $\boldsymbol{C}$ vollen Zeilenrang $r \leq p$ besitzt, werden die Likelihood-Quotienten-Statistik lq, die Wald-Statistik w und die Score-Statistik u eingesetzt. Dabei sind für $l(\boldsymbol{\beta})$, $s(\boldsymbol{\beta})$ und $F(\boldsymbol{\beta})$ die für das gewählte GLM spezifischen Formeln zu verwenden. Unter ähnlichen Voraussetzungen wie für die ML-Schätzung gilt $lq, w, u \overset{a}{\sim} \chi^2_r$. Damit können wieder Ablehnungsbereiche und (approximative) p-Werte berechnet werden.

Kriterien zur Modellanpassung und Modellwahl

Die zwei gebräuchlichsten globalen Statistiken, um die Anpassung eines Modells im Vergleich zum saturierten Modell zu überprüfen, sind die *Pearson-Statistik*

$$\chi^2 = \sum_{i=1}^{G} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/\omega_i}$$

und die *Devianz*

$$D = -2\phi \sum_{i=1}^{G} \{l_i(\hat{\mu}_i) - l_i(\overline{y}_i)\}.$$

Dabei sind $\hat{\mu}_i$ und $v(\hat{\mu}_i)$ die geschätzten Erwartungswert- und Varianzfunktionen und $l_i(\overline{y}_i)$ ist der i-te Log-Likelihood-Beitrag, wobei μ_i durch $\overline{y}_i$ ersetzt wird. Damit erhält man den maximal erreichbaren Wert der Log-Likelihood l_i. Für beide Modellwahl-Statistiken sollten die Daten so weit wie möglich gruppiert sein. Falls n_i, $i = 1, \ldots, G$,

in allen Gruppen hinreichend groß ist, so sind beide Statistiken approximativ (bzw. für $n \to \infty$ asymptotisch) $\phi\chi^2(G-p-1)$-verteilt, wobei p die Anzahl zu schätzender Parameter ist. In diesem Fall können beide Statistiken für formale Modellanpassungs-Tests benutzt werden, die die Modellanpassung des geschätzten Modells mit der Modellanpassung des saturierten Modells vergleichen. Für kleine n_i, insbesondere für $n_i = 1$, ist die Anwendung auch bei großem Gesamtumfang n problematisch. Große Werte von χ^2 oder D weisen dann nicht notwendig auf ein schlecht angepasstes Modell hin.

Das AIC-Kriterium zur Modellwahl ist allgemein definiert durch

$$\mathrm{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2p\,.$$

Enthält das Modell einen Dispersionsparameter ϕ wie bei der Normalverteilung, so ist dieser durch den Maximum-Likelihood-Schätzer im jeweiligen Modell zu ersetzen und die Parameterzahl entsprechend auf $p+1$ zu erhöhen.

4.5 Quasi-Likelihood-Modelle

Für generalisierte lineare Modelle impliziert die Annahme einer bestimmten Verteilung vom Exponentialfamilientyp, wie zum Beispiel Gauß-, Poisson- oder Binomialverteilung, und einer bestimmten Erwartungswert-Struktur $\mathrm{E}(\boldsymbol{y}\,|\,\boldsymbol{x}) = \mu = h(\boldsymbol{x}'\boldsymbol{\beta})$ eine spezifische Varianzstruktur $\mathrm{Var}(\boldsymbol{y}\,|\,\boldsymbol{x}) = \phi b''(\mu)$, wobei die Varianzfunktion $b''(\mu)$ durch die Wahl der Exponentialfamilie determiniert ist. Steht die empirische Varianz der Daten in deutlichem Widerspruch zu der geschätzten Varianz $\hat{\phi}b''(\hat{\mu})$, so ist die Verteilung der Daten durch die gewählte Exponentialfamilie falsch spezifiziert.

Quasi-Likelihood-Modelle erlauben die separate Spezifikation von Erwartungswert- und Varianzstruktur. Dazu ist es nicht notwendig, dass eine echte Likelihood-Funktion vorgegeben sein muss. Es reicht zum Beispiel aus, dass eine *Quasi-Likelihood-Funktion* $Q(\boldsymbol{\beta}, \phi)$ existiert, deren erste Ableitung die übliche Form einer Score-Funktion besitzt. Noch allgemeiner sind Quasi-Likelihood-Modelle, bei denen nur Erwartungswert- und Varianzfunktionen spezifiziert werden und die Parameter als Nullstellen von *Quasi-Score-Funktionen* oder *generalisierten Schätzgleichungen* (*generalized estimating equations, GEE*) bestimmt sind.

Man startet dann direkt mit der Vorgabe einer *generalisierten Schätzfunktion*

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{d_i}{\sigma_i^2}(y_i - \mu_i). \tag{4.29}$$

Wie in der Score-Funktion (4.29), die als Ableitung der Log-Likelihood eines GLM erhalten wird, ist $\mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ der korrekt spezifizierte Erwartungswert von y_i gegeben $\boldsymbol{x}_i$, d.h. es gilt

$$\mathrm{E}(y_i\,|\,\boldsymbol{x}_i) = \mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

Damit gilt wie bei echten Score-Funktionen

$$\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{d_i}{\sigma_i^2}(\mathrm{E}(y_i\,|\,\boldsymbol{x}_i) - \mu_i) = \boldsymbol{0},$$

eine für die Konsistenz der Parameterschätzung entscheidende Eigenschaft.

Dagegen muss die wahre Varianz $\mathrm{Var}(y_i \mid \boldsymbol{x}_i)$ nicht notwendig gleich

$$\sigma_i^2 = \sigma_i^2(\mu_i) = \phi b''(\mu_i)$$

mit $\mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ sein, sondern kann mit Hilfe einer vorgegebenen Quasi-Varianzfunktion $v(\mu)$ als $\sigma^2(\mu) = \phi v(\mu)$ spezifiziert werden. Man bezeichnet deshalb $\sigma^2(\mu) = \phi v(\mu)$ auch als *Arbeitsvarianz*.

Die einfachste Form einer (Arbeits-)Varianzfunktion ergibt sich bei der Behandlung von Überdispersion in Binomial- oder Poisson-Modellen. Dort setzt man

$$\sigma_i^2(\pi_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$$

bzw.

$$\sigma_i^2(\lambda_i) = \phi \lambda_i,$$

so dass die Quasi-Score-Funktion (4.29) zwar bis auf den konstanten Faktor $1/\phi$ identisch mit der Score-Funktion eines Binomial- oder Poisson-Modells ist, jedoch nicht mehr die Ableitung einer Log-Likelihood-Funktion.

Oft wird die (Arbeits-)Varianzfunktion durch einen weiteren Parameter θ parametrisiert und

$$\sigma^2(\mu) = \phi v(\mu; \theta)$$

gesetzt. Ein wichtiger Fall ist

$$v(\mu; \theta) = \mu^\theta.$$

Dieser Ansatz liefert für $\theta = 0, 1, 2, 3$ die Varianzfunktion der Gauß-, Poisson-, Gamma- und inversen Gauß-Verteilung.

Quasi-ML-Schätzer $\hat{\boldsymbol{\beta}}$ sind als Nullstelle(n) der Quasi-Score-Funktion, d.h. als Lösung der *generalisierten Schätzgleichung (GEE)*

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

definiert. Die Berechnung erfolgt iterativ wie bei der ML-Schätzung. Die Quasi-Fisher-Informationsmatrix $\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(-\partial \boldsymbol{s}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}')$ ergibt sich zu

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i' w_i,$$

wobei in die Gewichte $w_i = d_i^2/\sigma_i^2$ die Arbeitsvarianzen σ_i^2 eingehen. Im Allgemeinen ist nun jedoch $\boldsymbol{F}(\boldsymbol{\beta})$ verschieden von $\boldsymbol{V}(\boldsymbol{\beta}) = \mathrm{Cov}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta}))$. Tatsächlich ergibt sich

$$\boldsymbol{V}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i' w_i \frac{\sigma_{0i}^2}{\sigma_i^2}.$$

Nur wenn die Arbeitsvarianzen gleich den wahren Varianzen σ_{0i}^2 sind, ergibt sich wie bei der ML-Schätzung $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{V}(\boldsymbol{\beta})$.

4.6 Bayesianische generalisierte lineare Modelle

Die Ansätze der Bayes-Inferenz für lineare Modelle aus Kapitel 3.5 lassen sich konzeptionell auf generalisierte lineare Modelle übertragen. Die Durchführung wird jedoch rechnerisch komplizierter, und die volle Bayesianische Inferenz erfordert im Allgemeinen den Einsatz von MCMC-Simulationstechniken, die aufwändiger sind als für lineare Modelle. Dieser Abschnitt gibt einen kurzen Überblick zur Bayes-Inferenz in GLM. Wir beschränken uns dabei auf Modelle ohne Dispersionsparameter, d.h. insbesondere auf Binomial- und Poisson-Modelle. Für eine ausführliche Darstellung von Bayesianischen GLM und darauf basierenden Erweiterungen verweisen wir auf Dey, Gosh & Mallick (2000) sowie auf relevante Abschnitte in Fahrmeir & Tutz (2001).

In Bayesianischen GLM wird für den (als zufällig aufgefassten) Parametervektor $\boldsymbol{\beta}$ eine Priori-Dichte $p(\boldsymbol{\beta})$ angenommen. Die Inferenz beruht nach dem Satz von Bayes auf der Posteriori-Dichte $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$, gegeben die (bedingt unabhängigen) Zielvariablen $\boldsymbol{y} = (y_1, \ldots, y_n)'$ und die Kovariablen. Wenn wir zur Vereinfachung die Kovariablen $\boldsymbol{x}_i$, $i = 1, \ldots, n$, notationell unterdrücken, gilt also

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \frac{L(\boldsymbol{\beta} \,|\, \boldsymbol{y}) p(\boldsymbol{\beta})}{\int L(\boldsymbol{\beta} \,|\, \boldsymbol{y}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}} \propto L(\boldsymbol{\beta} \,|\, \boldsymbol{y}) p(\boldsymbol{\beta}) \,, \tag{4.30}$$

wobei

$$L(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i \,|\, \boldsymbol{\beta})$$

die Likelihood eines gewählten GLM ist, zum Beispiel die eines Logit-Modells oder eines log-linearen Poisson-Modells. Der Posteriori-Erwartungswert ist dann durch

$$\mathrm{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \int \boldsymbol{\beta} p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) d\boldsymbol{\beta}$$

definiert und

$$\mathrm{Cov}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \int (\boldsymbol{\beta} - \mathrm{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}))(\boldsymbol{\beta} - \mathrm{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}))' p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) d\boldsymbol{\beta}$$

ist die zugehörige Posteriori-Kovarianzmatrix als Maß für die Genauigkeit des Posteriori-Erwartungswerts. Auf den ersten Blick lässt sich somit die Bayes-Inferenz leicht durchführen. Das Problem sind jedoch die dabei auftretenden Integrationen, die analytisch nur in wenig praxisnahen Spezialfällen lösbar sind. Numerische Integrationsverfahren sind prinzipiell einsetzbar, solange die Dimension von $\boldsymbol{\beta}$ relativ klein (etwa ≤ 5) bleibt; jedoch sind Erweiterungen auf komplexere Modelle späterer Kapitel praktisch unmöglich. Somit verbleiben prinzipiell zwei Möglichkeiten, nämlich die Posteriori-Modus- oder MAP (Maximum A Posteriori)-Schätzung, für die der Zähler in (4.30) zu maximieren ist, und eine volle Bayes-Inferenz mit MCMC-Techniken.

Der *Posteriori-Modus-Schätzer* $\hat{\boldsymbol{\beta}}_p$ maximiert die Posteriori-Dichte $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ bzw. äquivalent dazu die Log-Posteriori-Likelihood

$$l_p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = l(\boldsymbol{\beta}) + \log\, p(\boldsymbol{\beta}) \,,$$

wobei $l(\boldsymbol{\beta})$ die Log-Likelihood des betrachteten GLM ist. Für die Normalverteilungs-Priori

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M}), \quad \boldsymbol{M} \text{ positiv definit}$$

ergibt sich speziell

$$l_p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = l(\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \boldsymbol{m}) \,,$$

wobei von $\boldsymbol{\beta}$ unabhängige Terme bereits weggelassen sind. Damit kann $l_p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ auch als *penalisierte (Log-) Likelihood* aufgefasst werden, wobei der Strafterm $(\boldsymbol{\beta} - \boldsymbol{m})' M^{-1} (\boldsymbol{\beta} - \boldsymbol{m})$ zu große Abweichungen vom Priori-Erwartungswert $\boldsymbol{m}$ penalisiert. Damit kann unter Umständen das Problem der Nichtexistenz bzw. Divergenz des ML-Schätzers behoben werden. Der Schätzer $\hat{\boldsymbol{\beta}}_p$ wird auch als penalisierter ML-Schätzer bezeichnet.

Für den Grenzfall $\boldsymbol{M}^{-1} \to \boldsymbol{0}$ einer nichtinformativen Priori

$$p(\boldsymbol{\beta}) \propto const$$

verschwindet der Strafterm und der Posteriori-Modus-Schätzer ist gleich dem ML-Schätzer.

Für den Spezialfall $\boldsymbol{m} = \boldsymbol{0}$ und $\boldsymbol{M} = \tau^2 \boldsymbol{I}$ ergibt sich der so genannte *Ridge Schätzer* mit *Shrinkage*-Parameter $\lambda = 1/(2\tau^2)$. Der Strafterm vereinfacht sich zu

$$\lambda(\boldsymbol{\beta}'\boldsymbol{\beta}) = \lambda(\beta_0^2 + \beta_1^2 + \ldots + \beta_k^2) \,,$$

und der Parameter λ regelt das „Schrumpfen" (Shrinkage) des ML-Schätzers $\hat{\boldsymbol{\beta}}$ gegen $\boldsymbol{0}$ und somit die Stabilisierung des ML-Schätzers bei zu hoher Variabilität.

Die Durchführung der Schätzung erfolgt in kompletter Analogie zur ML-Schätzung. Mit Normalverteilungs-Priori wird die Score-Funktion $\boldsymbol{s}(\boldsymbol{\beta})$ zur penalisierten Score-Funktion

$$\boldsymbol{s}_p(\boldsymbol{\beta}) = \frac{\partial l_p(\boldsymbol{\beta}'\boldsymbol{y})}{\partial \boldsymbol{\beta}} = \boldsymbol{s}(\boldsymbol{\beta}) - \boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})$$

und die Fisher-Informationsmatrix $\boldsymbol{F}(\boldsymbol{\beta})$ wird zu

$$\boldsymbol{F}_p(\boldsymbol{\beta}) = -\mathrm{E}\left(-\frac{\partial^2 l_p(\boldsymbol{\beta} \,|\, \boldsymbol{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{M}^{-1} \,.$$

Die numerische Berechnung erfolgt mit einem entsprechend modifizierten Fisher-Scoring-Algorithmus bzw. IWLS-Algorithmus, bei dem $\boldsymbol{s}(\boldsymbol{\beta})$ und $\boldsymbol{F}(\boldsymbol{\beta})$ durch $\boldsymbol{s}_p(\boldsymbol{\beta})$ und $\boldsymbol{F}_p(\boldsymbol{\beta})$ zu ersetzen sind.

Für $n \to \infty$ ist $\hat{\boldsymbol{\beta}}_p$ asymptotisch (bzw. approximativ) normalverteilt mit

$$\hat{\boldsymbol{\beta}}_p \sim \mathrm{N}\left(\boldsymbol{\beta}, \boldsymbol{F}_p^{-1}(\hat{\boldsymbol{\beta}}_p)\right) \,,$$

so dass der Posteriori-Modus $\hat{\boldsymbol{\beta}}_p$ und die (erwartete) Krümmung $\boldsymbol{F}_p^{-1}(\hat{\boldsymbol{\beta}}_p)$ gute Approximationen des Posteriori-Erwartungswerts $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ und der Posteriori-Kovarianzmatrix $\mathrm{Cov}(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ sind.

Die *volle Bayes-Inferenz* beruht auf MCMC-Techniken (vergleiche Anhang B.5) zum Ziehen von Zufallszahlen aus der Posteriori $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$. Posteriori-Erwartungswerte, -Mediane, -Quantile, -Varianzen etc. werden dann durch ihre empirischen Analoga approximiert. Für die Normalverteilungs-Priori $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M})$ und auch für den Grenzfall $\boldsymbol{M}^{-1} \to \boldsymbol{0}$ der nichtinformativen Priori $p(\boldsymbol{\beta}) \propto const$ gilt

Bayesianische GLM

Posteriori-Verteilung

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \frac{L(\boldsymbol{\beta} \,|\, \boldsymbol{y}) p(\boldsymbol{\beta})}{\int L(\boldsymbol{\beta} \,|\, \boldsymbol{y}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}} \propto L(\boldsymbol{\beta} \,|\, \boldsymbol{y}) p(\boldsymbol{\beta})$$

wobei $L(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ die Likelihood eines GLM ist und $p(\boldsymbol{\beta})$ die Priori-Verteilung.

Posteriori-Modus

Der *Posteriori-Modus-Schätzer* $\hat{\boldsymbol{\beta}}_p$ maximiert die Posteriori-Dichte $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$. Bei Normalverteilungs-Priori

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M})$$

ist diese äquivalent zur *penalisierten Log-Likelihood*

$$l_p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = l(\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \boldsymbol{m}) \,,$$

mit $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta} \,|\, \boldsymbol{y})$. Die iterative Berechnung von $\hat{\boldsymbol{\beta}}_p$ erfolgt mit Hilfe der penalisierten Score-Funktion und Fisher-Information.

Volle Bayes-Inferenz

Die volle Bayes-Inferenz wird mit MCMC zum Ziehen von Zufallszahlen aus der Posteriori-Dichte $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ durchgeführt.

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) \propto \exp\left(l(\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \boldsymbol{m}) \right).$$

Von Spezialfällen abgesehen existiert keine geschlossene analytische Form für die Normalisierungskonstante. Das Ziehen von Stichproben $\boldsymbol{\beta}^{(t)}, t = 1, \ldots, T$, aus $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ erfolgt deshalb mit MCMC-Techniken. Dellaportas & Smith (1993) schlagen einen Gibbs-Sampler mit adaptivem Rejection-Sampling vor, der im Programmpaket WinBUGS implementiert ist. Eine von uns bevorzugte Methode ist das Ziehen des gesamten Parametervektors $\boldsymbol{\beta}^{(t)}$ in jedem Iterationsschritt t mit einem Metropolis-Hastings (MH)-Algorithmus, bei dem eine so genannte IWLS-Vorschlagsdichte verwendet wird, siehe Gamerman (1997). Als Vorschlagsdichte $q(\boldsymbol{\beta}^* \,|\, \boldsymbol{\beta}^{(t)})$ für den nächsten Zufallsvektor $\boldsymbol{\beta}^{(t+1)}$ wird dabei eine Normalverteilung verwendet, deren Erwartungswert und Kovarianzmatrix eine (erste) Näherung des Posteriori-Modus-Schätzers und der zugehörigen Kovarianzmatrix sind.

Für den Grenzfall $\boldsymbol{m} = 0$, $\boldsymbol{M}^{-1} = 0$ einer nichtinformativen Priori $p(\boldsymbol{\beta}) \propto \mathit{const}$ wird der Vorschlag $\boldsymbol{\beta}^*$ aus einer Normalverteilungsdichte $q(\boldsymbol{\beta}^* | \boldsymbol{\beta}^{(t)})$ mit

$$\boldsymbol{\beta}^* \sim \mathrm{N}(\boldsymbol{\mu}^{(t)}, (\boldsymbol{X}' \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1})$$

gezogen. Dabei ist $\boldsymbol{F}^{(t)} = \boldsymbol{X}' \boldsymbol{W}^{(t)} \boldsymbol{X}$ die Fisher-Matrix ausgewertet an der Stelle $\boldsymbol{\beta}^{(t)}$ des aktuellen Iterationswertes $\boldsymbol{\beta}^{(t)}$ und

$$\boldsymbol{\mu}^{(t)} = (\boldsymbol{X}' \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{W}^{(t)} \tilde{\boldsymbol{y}}^{(t)},$$

Variable	Koeffizient	Standard-abweichung	2.5%-Quantil	97.5%-Quantil
Konstante	0.178	0.032	0.115	0.240
jahrc	-0.075	0.003	-0.081	-0.069
alandc	-0.025	0.004	-0.033	-0.016
anspc	0.020	0.001	0.018	0.023
biopharm	0.292	0.031	0.229	0.355
uszw	-0.044	0.026	-0.094	0.005
patus	0.016	0.026	-0.033	0.067
patdsg	-0.230	0.033	-0.294	-0.168
einspruch	0.402	0.025	0.357	0.451

Tabelle 4.13. *Anzahl der Zitate von Patenten: Poisson-Bayes-Modell.*

Variable	Koeffizient	Standard-abweichung	2.5%-Quantil	97.5%-Quantil
Konstante	0.514	0.039	0.438	0.591
jahrc	38.915	3.088	32.910	44.835
$jahrc^2$	-0.010	0.001	-0.011	-0.008
alandc	-0.247	0.048	-0.338	-0.147
$alandc^2$	0.037	0.006	0.025	0.049
$alandc^3$	-0.002	<0.001	-0.002	-0.001
anspc	0.040	0.004	0.033	0.047
$anspc^2$	-0.001	<0.001	-0.001	>-0.001
biopharm	0.199	0.031	0.142	0.259
uszw	-0.047	0.025	-0.097	0.001
patus	-0.003	0.027	-0.058	0.054
patdsg	-0.232	0.032	-0.294	-0.173
einspruch	0.400	0.024	0.352	0.446

Tabelle 4.14. *Anzahl der Zitate von Patenten: Erweitertes Poisson-Bayes-Modell mit nichtlinearen Effekten.*

mit $\boldsymbol{W}^{(t)} = \boldsymbol{W}(\boldsymbol{\beta}^{(t)})$ und den aktuellen Arbeitsbeobachtungen $\tilde{\boldsymbol{y}}^{(t)}$ (definiert wie beim Fisher-Scoring der ML-Schätzung). Die Akzeptanzwahrscheinlichkeit ist dann gegeben durch

$$\alpha(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{(t)}) = \min\left\{\frac{L(\boldsymbol{\beta}^*)q(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^*)}{L(\boldsymbol{\beta}^{(t)})q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{(t)})}\right\},$$

mit der Likelihood $L(\boldsymbol{\beta})$ des GLM ausgewertet am vorgeschlagenen bzw. aktuellen Wert $\boldsymbol{\beta}^*$ bzw. $\boldsymbol{\beta}^{(t)}$.

Beispiel 4.9 Anzahl der Zitate von Patenten

Wir illustrieren die Bayes-Inferenz durch eine Reanalyse des log-linearen Poisson-Modells von Beispiel 4.7 (Seite 214), wobei wir für $\boldsymbol{\beta}$ die flache Priori $p(\beta_j) \propto const$ wählen. Somit sollten bei der vollen Bayes-Inferenz der Posteriori-Erwartungswert und der ML-Schätzer, der zugleich Posteriori-Modus-Schätzer ist, nicht zu stark voneinander abweichen. Tabelle 4.13 enthält die Ergebnisse, zusammen mit Posteriori-Standardabweichungen und -Quantilen, für das Poisson-Bayes-Modell mit rein linearen Effekten, Tabelle 4.14 die entsprechenden Ergebnisse mit nichtlinearen Effekten für die

Abb. 4.4. *Anzahl der Zitate von Patenten: Geschätzte Posteriori-Dichten der Kovariablen uszw, patus und patdsg (durchgezogene Linie) und Normalverteilungsapproximationen (gestrichelte Linie).*

metrischen Kovariablen. Es zeigt sich tatsächlich ein recht gute Übereinstimmung. Man beachte jedoch, dass die vorhandene Überdispersion nicht berücksichtigt wurde!

Da sich die Ergebnisse der ML- und der Bayesschätzung in diesem Beispiel nur unwesentlich unterscheiden entsteht die Frage nach dem Vorteil der vergleichsweise computerintensiven Bayes-Inferenz mit MCMC-Methoden. Ein Vorteil ist, dass neben den Punktschätzern und Konfidenzintervallen mit Hilfe der gezogenen Zufallszahlen auch die gesamte Posteriori-Dichte $p(\beta \mid y)$ interessierender Effekte geschätzt werden kann. Abbildung 4.4 zeigt die Kerndichteschätzungen der Posteriori-Dichten $p(\beta_j \mid y)$ der Effekte der Kovariablen *uszw*, *patus* und *patdsg* sowie jeweils Normalverteilungsdichten mit gleichen Erwartungswerten und Varianzen. Die Posteriori-Dichten sind in guter Näherung normalverteilt, was bei dem im Vergleich zur Parameterzahl hohen Stichprobenumfang zu erwarten war. Generell gilt, dass die Bayes-Inferenz mit MCMC vor allem für komplexere Regressionsmodelle lohnend ist, bei denen die asymptotischen Normalverteilungsapproximationen der Likelihood-Inferenz zu ungenau werden.

△

Beispiel 4.10 Kredit-Scoring

Wir reanalysieren die Kredit-Scoring-Daten (vergleiche Beispiel 4.5) mit einem vollen Bayes-Ansatz, wobei für die Parameter eine nichtinformative Priori $p(\beta_j) = const$ gewählt wurde. Tabelle 4.15 zeigt die mit MCMC-Inferenz gewonnenen Schätzungen.

Variable	Koeffizient	Standard-abweichung	t-Wert	p-Wert
Konstante	-0.282	0.299	-0.871	0.299
konto1	0.633	0.176	0.301	0.981
konto2	-1.332	0.198	-1.709	-0.932
laufz	0.034	0.008	0.018	0.048
hoehe	0.024	0.033	-0.039	0.090
moral	-1.005	0.257	-1.503	-0.503
zweck	-0.433	0.153	-0.735	-0.144

Tabelle 4.15. *Kredit-Scoring: Bayes-Modell.*

Ein Vergleich mit den ML-Schätzungen in Beispiel 4.5 zeigt weitgehende Übereinstimmung. Dies lässt darauf schließen, dass die Stichprobe hinreichend groß ist und somit Posteriori-Erwartungswert und Posteriori-Modus (d.h. der ML-Schätzer) wegen der Normalverteilungsapproximation nahe beieinander liegen.

△

4.7 Bemerkungen und Ergänzungen

Generalisierte lineare Modelle wurden als allgemeine Modellklasse für Zielvariablen, deren Dichten einer Exponentialfamilie angehören, von Nelder & Wedderburn (1972) eingeführt. Damit konnten lineare, Logit-, Probit- und Poisson-Modelle in einem einheitlichen Konzept zusammengefasst werden, und die statistische Modellbildung und Methodenentwicklung erhielt neue und anwendungsrelevante Impulse. Eine ausführliche Darstellung gibt das Buch von McCullagh & Nelder (1989); kürzere Einführungen finden sich in Fahrmeir et al. (1996, Kap. 6), Fahrmeir & Tutz (2001, Kap. 2). Binäre Regressionsmodelle werden ausführlich von Collett (1991) und Tutz (2000, Kap. 2–5) beschrieben. Diese Bücher geben auch einen Überblick über auf Residuen beruhende Methoden der Modelldiagnose, die in Analogie zur linearen Regression von Kapitel 3 entwickelt wurden.

Nachfolgende Modifikationen und Erweiterungen von GLM wurden durch verschiedene Aspekte motiviert. So treten in Anwendungen Zielvariablen auf, deren Verteilungen nicht oder nur unzureichend durch univariate Exponentialfamilien wie in Abschnitt 4.4 modelliert werden können. Dies trifft insbesondere für die folgenden Regressionssituationen zu.

- *Kategoriale Regression:* Die Zielvariablen sind Merkmale mit geordneten oder ungeordneten Kategorien. Modelle der kategorialen Regression (siehe Kapitel 5) werden ausführlich von Tutz (2000) und Agresti (2002) beschrieben; knappere Fassungen finden sich in Fahrmeir et al. (1996, Kap. 6), Fahrmeir & Tutz (2001, Kap. 3).
- *Zählvariablen-Regression:* Die in Abschnitt 4.2 dargestellte Poisson-Regression ist zwar das Standardmodell der Zählvariablen-Regression, in Anwendungen ist die Poisson-Verteilung aber häufig zu einfach. Cameron & Trivedi (1998) und Winkelmann (1997) beschreiben viele darüber hinausgehende Konzepte der Zählvariablen-Regression.
- *Lebensdauer-(Survival-) und Verweildaueranalyse:* Lebensdauern, Verweildauern oder Wartezeiten bis zu einem bestimmten Ereignis treten in vielen Anwendungsbereichen

auf. Dabei liegen oft sogenannte zensierte Daten als Besonderheit vor, wenn Lebensdauern etwa bis zum Ende einer Studie noch andauern. Das populärste Regressionsmodell für (zensierte) Lebensdauern ist das Cox-Modell. Es weist formal enge Bezüge zur Poisson-Regression auf. Anwendungsorientierte Textbücher im Bereich der medizinischen Statistik sind Klein & Moeschberger (2005) und Collett (2003), während bei Blossfeld, Hamerle & Mayer (1986) und Blossfeld & Rohwer (2001) sozialwissenschaftliche Fragestellungen im Vordergrund stehen. Eine knappere Darstellung geben Fahrmeir et al. (1996, Kap. 7).

- *GLM für Lokations-, Skalen- und Formparameter:* Insbesondere für stetige Zielvariablen wird der Einfluss von Kovariablen nicht nur auf den Erwartungswert sondern auch auf Varianz-, Schiefe- oder Wölbungsparameter durch (lineare) Prädiktoren modelliert, siehe Rigby & Stasinopoulos (2005) sowie das R-Paket `GAMLSS`.
- *Mehrdimensionale Zielvariablen:* Wenn die Zielvariable $\boldsymbol{y} = (y_1, \ldots, y_q)$ aus mehreren skalaren Zielvariablen $y_1, \ldots, y_q$ besteht, spricht man auch von *multivariater Regression*. Fahrmeir et al. (1996, Kap. 4.2) führen in die multivariate lineare Regression als Erweiterung des linearen Regressionsmodell ein. Für nicht-normalverteilte Komponenten $y_1, \ldots, y_q$ besteht die Schwierigkeit darin, geeignete gemeinsame Verteilungen zu finden. Vor allem für stetige Komponenten bieten *Copula*-Konzepte eine geeignete Möglichkeit (Joe 1997). Andere Ansätze sind Quasi-Likelihood- oder marginale Modelle (Fahrmeir & Tutz, 2001, Kap. 3) und Modelle mit latenten Variablen (Skrondal & Rabe-Hesketh 2004).

Eine der wichtigsten Erweiterungen von GLM ist das Einbeziehen von *non- und semiparametrischen Ansätzen*, um nichtlineare Effekte von Kovariablen flexibel zu modellieren. Die resultierende Modellklassen, wie generalisierte additive Modelle (GAM), wurden bereits in Kapitel 2 eingeführt und werden ausführlicher in den Kapiteln 7 und 8 vorgestellt.

In ihrer ursprünglichen Definition sind GLM vor allem für Regressionsanalysen von Querschnittsdaten geeignet. Für die *Analyse von Zeitreihen- oder Longitudinaldaten* sind gemischte Modelle (Kapitel 6) ein möglicher Ansatz. Abhängig von den inhaltlichen Zielen einer Längsschnittstudie können auch *autoregressive* (oder *konditionale*) Modelle, bei denen zeitlich verzögerte Werte der Zielvariable als Kovariablen verwendet werden, oder *marginale Modelle*, die auf Quasi-Likelihood-Ansätzen basieren, eine sinnvolle Alternative zur Analyse sein, siehe dazu Diggle, Heagerty, Liang & Zeger (2002), Fahrmeir & Tutz (2001, Kap. 6) sowie die ergänzenden Bemerkungen des Abschnitts 6.8.

Bayesianische GLM und Erweiterungen erfuhren mit der Verbreitung von MCMC-Simulationstechniken zu Anfang der 90er Jahre eine rasante Entwicklung, siehe dazu Dey et al. (2000) sowie entsprechende Abschnitte in den nachfolgenden Kapiteln.

GLM mit Fehlern in den Variablen wurden für Datensituationen entwickelt, bei denen die Variablen nicht direkt, sondern nur mit Messfehlern behaftet, beobachtet werden können. Für Modelle dieser Art verweisen wir etwa auf Carroll, Ruppert, Stefanski & Crainiceanu (2006).

5 Kategoriale Regressionsmodelle

5.1 Einführung

In vielen Anwendungen ist die Zielvariable nicht binär wie in Abschnitt 4.1, sondern mehrkategorial. Die Zielvariable kann dabei *nominal* sein, d.h. die Kategorien $1, \ldots, c$ sind ungeordnet, oder sie kann *ordinal* sein, so dass die Kategorien der Größe nach geordnet werden können.

Beispiel 5.1 Infektion nach Kaiserschnitt-Geburt

	Kaiserschnitt					
	geplant			**nicht geplant**		
	Infektion			**Infektion**		
	I	II	nein	I	II	nein
Antibiotika						
Risikofaktor	0	1	17	4	7	87
Kein Risikofaktor	0	0	2	0	0	0
Kein Antibiotika						
Risikofaktor	11	17	30	10	13	3
Kein Risikofaktor	4	4	32	0	0	9

Tabelle 5.1. *Daten zu Infektionen bei 251 Kaiserschnitt-Geburten.*

Tabelle 5.1 zeigt Daten zu Infektionen nach Geburten mit Kaiserschnitt im Klinikum Großhadern in München. Die Zielvariable hat $c = 3$ ungeordnete Kategorien: Infektion vom Typ I, Infektion vom Typ II, keine Infektion. Dazu wurden drei binäre Kovariablen betrachtet:

NPLAN : Kaiserschnitt war geplant / nicht geplant
RISK : Risikofaktoren, wie z.B. Diabetes oder Übergewicht waren vorhanden / nicht vorhanden
ANTIB : Antibiotika wurden als Prophylaxe verabreicht/ nicht verabreicht

△

Beispiel 5.2 Zustand des Waldes

Für die Studie zum Zustand von Buchen in Beispiel 1.4 (Seite 9) liegt die Zielvariable „Belaubungsgrad“ zunächst in 9 geordneten Kategorien vor (0%, 12.5%,...,100% Schädigung). Da für einige dieser Kategorien nur wenige oder keine Beobachtungen vorliegen, fassen wir die Zielvariable in die drei Kategorien 1=keine (0%), 2=leichte (12.5%–37.5%), 3=schwere ($\geq$ 50%) Schädigung zusammen. Für jeden der Beobachtungspunkte in dieser Studie werden jährlich die Zielvariable zusammen mit den in Tabelle 1.4 aufgeführten metrischen und kategorialen Kovariablen erhoben. Anders als im Beispiel 5.1 liegen hier ungruppierte Individualdaten vor.

△

Alter	Rauchverhalten	Testergebnis		
		normal	grenzwertig	abnormal
< 40	kein Raucher	577	27	7
	früherer Raucher	192	20	3
	derzeitiger Raucher	682	46	11
40–59	kein Raucher	164	4	0
	früherer Raucher	145	15	7
	derzeitiger Raucher	245	47	27

Tabelle 5.2. *Lungenfunktionstest.*

Beispiel 5.3 Lungenfunktion

In einer Studie zur Beeinträchtigung der Lungenfunktion wurden als Einflussgrößen das Alter und das Rauchverhalten texanischer Industriearbeiter erhoben (siehe Forthofer & Lehnen (1981)). Als abhängige Variable werden die Ergebnisse eines Atmungstests mit den Kategorien „normal", „grenzwertig" und „abnormal" betrachtet. Man erhielt die Kontingenztabelle 5.2.

△

In den Beispielen und im Weiteren besitzt die Zielvariable Y jeweils c geordnete oder ungeordnete Kategorien, d.h. es gilt $Y \in \{1, \ldots, c\}$. Ziel kategorialer Regressionsmodelle ist die Modellierung und Schätzung der (Auftretens-) Wahrscheinlichkeiten

$$\pi_r = \mathrm{P}(Y = r), \qquad r = 1, \ldots, c,$$

in Abhängigkeit von Kovariablen. Im Weiteren erweist sich oft die Darstellung der Zielvariablen durch einen Vektor $\boldsymbol{y} = (y_1, \ldots, y_q)'$ von $q = c - 1$ Dummy-Variablen

$$y_r = \begin{cases} 1, & Y = r \\ 0, & \text{sonst} \end{cases} \qquad r = 1, \ldots, q$$

als zweckmäßig. Dabei haben wir Kategorie c als „Referenz"-Kategorie gewählt, so dass $y_c = 1 - y_1 - \ldots - y_q$ gilt. Somit erhält man

$$Y = r \Longleftrightarrow \boldsymbol{y} = (0, \ldots, 1, \ldots, 0)', \qquad r = 1, \ldots, q$$

mit einer 1 als r-te Komponente von $\boldsymbol{y}$, und

$$\pi_r = \mathrm{P}(Y = r) = \mathrm{P}(y_r = 1), \quad r = 1, \ldots, q.$$

Für die Referenzkategorie gilt entsprechend

$$Y = c \Longleftrightarrow \boldsymbol{y} = (0, \ldots, 0, \ldots, 0)', \qquad \mathrm{P}(Y = c) = 1 - \pi_1 - \ldots - \pi_q.$$

Multinomialverteilung

Als Verallgemeinerung der Binomialverteilung wird die Multinomialverteilung zur Modellierung mehrkategorialer Zielvariablen benötigt. Für eine individuelle Beobachtung

$\boldsymbol{y} = (y_1, \ldots, y_q)$ mit den 0/1-Variablen $y_1, \ldots, y_q$ und dem Vektor $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_q)$ der Auftretenswahrscheinlichkeiten, ist die Wahrscheinlichkeitsfunktion durch

$$f(\boldsymbol{y} \mid \boldsymbol{\pi}) = \pi_1^{y_1} \cdot \ldots \cdot \pi_q^{y_q} (1 - \pi_1 - \ldots - \pi_q)^{1 - y_1 - \ldots - y_q} \tag{5.1}$$

in Verallgemeinerung der Bernoulli-Verteilung gegeben. Für m unabhängige Wiederholungen bezeichne in Verallgemeinerung zur Binomialverteilung y_r, $r = 1, \ldots, c$ nun die Anzahl von Wiederholungen, in denen die Kategorie r aufgetreten ist. Dann besitzt $\boldsymbol{y} = (y_1, \ldots, y_q)'$ die Wahrscheinlichkeitsfunktion

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\pi}) &= \frac{m!}{y_1! \cdot \ldots \cdot y_q! (m - y_1 - \ldots - y_q)!} \pi_1^{y_1} \cdot \ldots \cdot \pi_q^{y_q} (1 - \pi_1 - \ldots - \pi_q)^{1 - y_1 - \ldots - y_q} \\ &= \frac{m!}{y_1! \cdot \ldots \cdot y_c!} \pi_1^{y_1} \cdot \ldots \cdot \pi_c^{y_c} \end{aligned} \tag{5.2}$$

einer Multinomialverteilung, kurz

$$\boldsymbol{y} \sim M(m, \boldsymbol{\pi}),$$

mit den Parametern m und $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_q)'$. Für die ersten beiden Momente ergibt sich

$$\mathrm{E}(\boldsymbol{y}) = m\boldsymbol{\pi} = \begin{pmatrix} m\pi_1 \\ \vdots \\ m\pi_q \end{pmatrix}, \quad \mathrm{Cov}(\boldsymbol{y}) = \begin{pmatrix} m\pi_1(1-\pi_1) & \cdots & -\pi_1\pi_q \\ \vdots & \ddots & \vdots \\ -\pi_q\pi_1 & \cdots & m\pi_q(1-\pi_q) \end{pmatrix}. \tag{5.3}$$

Verwendet man statt der absoluten Häufigkeiten y_r die relativen Häufigkeiten $\bar{y}_r = y_r/m$, so ist der Vektor $\bar{\boldsymbol{y}} = (\bar{y}_1, \ldots, \bar{y}_q)'$ der relativen Häufigkeiten skaliert multinomialverteilt, $\bar{\boldsymbol{y}} \sim M(m, \boldsymbol{\pi})/m$ mit

$$\mathrm{E}(\bar{\boldsymbol{y}}) = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_q \end{pmatrix}, \quad \mathrm{Cov}(\bar{\boldsymbol{y}}) = \frac{1}{m} \begin{pmatrix} \pi_1(1-\pi_1) & \cdots & -\pi_1\pi_q \\ \vdots & \ddots & \vdots \\ -\pi_q\pi_1 & \cdots & \pi_q(1-\pi_q) \end{pmatrix}. \tag{5.4}$$

Daten

Die Datenstruktur ist analog zu der von binären und anderen univariaten generalisierten linearen Modellen. Für ungruppierte Individualdaten wird zu jeder Einheit i der Wert der kategorialen Zielvariable $Y_i \in \{1, \ldots, c\}$ bzw. $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})'$ und der Kovariablenvektor $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})'$ angegeben:

$$\begin{array}{c} \text{Einheit } 1 \\ \vdots \\ \text{Einheit } i \\ \vdots \\ \text{Einheit } n \end{array} \begin{pmatrix} \boldsymbol{y}_1' = (y_{11}, \ldots, y_{1q}) \\ \vdots \\ \boldsymbol{y}_i' = (y_{i1}, \ldots, y_{iq}) \\ \vdots \\ \boldsymbol{y}_n' = (y_{n1}, \ldots, y_{nq}) \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

Bei gruppierten Daten werden Beobachtungen mit identischem Kovariablenvektor $\boldsymbol{x}_i$ zu einer Gruppe i von n_i Einheiten zusammengefasst und die absoluten Häufigkeiten

$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})'$ bzw. die relativen Häufigkeiten $\bar{\boldsymbol{y}}_i = (\bar{y}_{i1}, \ldots, \bar{y}_{iq})'$ der beobachteten Kategorien in der Gruppe i, $i = 1, \ldots, G$, notiert:

$$\begin{array}{c} \text{Gruppe } 1 \\ \vdots \\ \text{Gruppe } i \\ \vdots \\ \text{Gruppe } G \end{array} \begin{pmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_G \end{pmatrix} \begin{pmatrix} \bar{\boldsymbol{y}}_1' = (\bar{y}_{11}, \ldots, \bar{y}_{1q}) \\ \vdots \\ \bar{\boldsymbol{y}}_i' = (\bar{y}_{i1}, \ldots, \bar{y}_{iq}) \\ \vdots \\ \bar{\boldsymbol{y}}_G' = (\bar{y}_{G1}, \ldots, \bar{y}_{Gq}) \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} \\ \vdots & & \vdots \\ x_{G1} & \cdots & x_{Gk} \end{pmatrix}$$

Die Darstellung gruppierter Daten erfolgt häufig in Kontingenztabellen, wie in den Beispielen 5.1 und 5.3, unter Angabe der absoluten Häufigkeiten $y_{i1}, \ldots, y_{ic}$ statt der relativen Häufigkeiten.

Individualdaten $\boldsymbol{y}_i$ sind bei gegebenen Kovariablen $\boldsymbol{x}_i$ (bedingt) multinomialverteilt,

$$\boldsymbol{y}_i \,|\, \boldsymbol{x}_i \sim M(1, \boldsymbol{\pi}_i), \qquad \boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'.$$

Für gruppierte Daten sind die absoluten Häufigkeiten multinomialverteilt mit

$$\boldsymbol{y}_i \,|\, \boldsymbol{x}_i \sim M(n_i, \boldsymbol{\pi}_i),$$

während die relativen Häufigkeiten $\bar{\boldsymbol{y}}_i$ skaliert multinomialverteilt sind.

Ziel mehrkategorialer Regressionsmodelle ist die Modellierung und Schätzung der Wahrscheinlichkeiten $\pi_{ir} = \mathrm{P}(Y_i = r \,|\, \boldsymbol{x}_i) = \mathrm{P}(y_{ir} = 1 \,|\, \boldsymbol{x}_i)$ in Abhängigkeit von den Kovariablen $\boldsymbol{x}_i$.

Im Vergleich zu binären Regressionsmodellen existiert eine noch größere Vielfalt von Modellansätzen. Wir beschreiben in den nächsten beiden Abschnitten einige wichtige Modelle für nominale und ordinale Zielvariablen. Weitere Modelle finden sich bei Agresti (2002), Fahrmeir & Tutz (2001) und Tutz (2000).

5.2 Modelle für ungeordnete Kategorien

In diesem Abschnitt wird davon ausgegangen, dass die Kategorien $r \in \{1, \ldots, c\}$ entweder keine Ordnungsstruktur besitzen oder diese nicht benutzt wird. Dies trifft auf das Beispiel 5.1 und auf viele weitere Anwendungen in Biomedizin, Sozial- und Wirtschaftswissenschaften zu. Ein bekanntes Anwendungsfeld ist die Modellierung von Wahlsituationen. Dies gilt für die Wahl zwischen politischen Parteien ebenso wie bei der Wahl des Transportmittels auf dem Weg zum Arbeitsplatz (z.B. Auto, Bahn, Fahrrad). Die Kategorien der Zielvariablen stellen also nur unterschiedliche Alternativen ohne notwendige Ordnungsstruktur dar, d.h. die Zielvariable besitzt nur Nominalskalenniveau.

Für binäre Zielvariablen ($c = 2, q = 1$) wurde im Abschnitt 4.1 das logistische Modell

$$\mathrm{P}(Y_i = 1 \,|\, \boldsymbol{x}_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

betrachtet. Eine direkte Verallgemeinerung ist das *mehrkategoriale (multinomiale) Logit-Modell*:

Mehrkategoriales Logit-Modell

Daten

Die Zielvariable $Y_i \in \{1, \ldots, c\}$ ist kategorial und nominalskaliert. Zusätzlich sind Kovariablen $\boldsymbol{x}_i$ gegeben, die nicht von der Kategorie der Zielvariablen abhängen.

Modell

Die Auftretenswahrscheinlichkeit für Kategorie r ist bestimmt durch

$$\mathrm{P}(Y_i = r \,|\, \boldsymbol{x}_i) = \pi_{ir} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{q} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_s)} \qquad r = 1, \ldots, q. \tag{5.5}$$

Für die Referenzkategorie c gilt

$$\pi_{ic} = 1 - \pi_{i1} - \ldots - \pi_{iq} = \frac{1}{1 + \sum_{s=1}^{q} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_s)}$$

Eine äquivalente Darstellung ist

$$\log \frac{\pi_{ir}}{\pi_{ic}} = \boldsymbol{x}_i'\boldsymbol{\beta}_r \quad bzw. \quad \frac{\pi_{ir}}{\pi_{ic}} = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_r), \qquad r = 1, \ldots, c, \tag{5.6}$$

wobei aus $\pi_{ic}/\pi_{ic} = 1$ automatisch $\boldsymbol{\beta}_c = \mathbf{0}$ folgt.
Die Parameter $\boldsymbol{\beta}_r = (\beta_{r0}, \beta_{r1}, \ldots, \beta_{rk})'$ sind kategorienspezifisch, ebenso wie die linearen Prädiktoren $\eta_{ir} = \boldsymbol{x}_i'\boldsymbol{\beta}_r = \beta_{r0} + x_{i1}\beta_{r1} + \ldots + x_{ik}\beta_{rk}$, $r = 1, \ldots, q$.

In der Form (5.6) wird deutlich, dass der lineare Prädiktor $\eta_{ir} = \boldsymbol{x}_i'\boldsymbol{\beta}_r$ die (logarithmierte) *Chance* bzw. das *relative Risiko* zwischen Kategorie r und der Referenzkategorie c durch ein lineares bzw. exponentiell-multiplikatives Modell spezifiziert. Damit ist die Interpretation der Parameter analog zum binären Logit-Modell, mit dem einzigen Unterschied, dass nun nicht die Logits für das Auftreten ($Y = 1$) im Verhältnis zum Nichtauftreten ($Y = 0$), sondern die Logits für das Verhältnis von Kategorie r ($Y = r$) zur Referenzkategorie c ($Y = c$) modelliert werden.

Bei der Interpretation der Kovariableneffekte ist jedoch zu beachten, dass auch bei positivem Regressionskoeffizienten β_{rj} einer Kovariablen x_j die Wahrscheinlichkeit für Kategorie r nicht notwendigerweise ansteigt, wenn der Kovariablenwert von x_j auf $x_j + 1$ erhöht wird. Zwar vergrößert sich die Chance für Kategorie r im Vergleich zur Referenzkategorie, daraus folgt aber nicht, dass die Chance für Kategorie r im Vergleich zu allen anderen Alternativen ebenfalls anwächst. Besitzt beispielsweise Kategorie s einen größeren Regressionskoeffizienten als Kategorie r, das heißt gilt $\beta_{sj} > \beta rj$, so wächst die Chance für Kategorie s im Vergleich zur Referenzkategorie stärker als die Chance für Kategorie r, so dass sich insgesamt die Wahrscheinlichkeit für Kategorie r verkleinern kann, trotz positivem Koeffizienten β_{rj}. Um den Effekt sich ändernder Kovariablenausprägungen auf die Wahrscheinlichkeiten einfach abschätzen zu können, bietet es sich an, diese Wahrscheinlichkeiten in einer Grafik abzutragen, wobei alle übrigen Kovariablen bei typischen Werten festgehalten werden.

		Zielvariable		Kovariablen		
		y_1	y_2	*NPLAN*	*RISK*	*ANTIB*
Gruppe 1	n_1=40	4	4	0	0	0
Gruppe 2	n_2=58	11	17	0	1	0
Gruppe 3	n_3=2	0	0	0	0	1
Gruppe 4	n_4=18	0	1	0	1	1
Gruppe 5	n_5=9	0	0	1	0	0
Gruppe 6	n_6=26	10	13	1	1	0
Gruppe 7	n_7=98	4	7	1	1	1

Tabelle 5.3. *Kaiserschnitt-Geburten: Gruppierte Darstellung der Daten.*

Infektion vom Typ 1			Infektion vom Typ 2		
	β	$\exp(\beta)$		β	$\exp(\beta)$
Intercept	-2.621	0.072	Intercept	-2.560	0.077
NPLAN	1.174	3.235	*NPLAN*	0.996	2.707
ANTIB	-3.520	0.030	*ANTIB*	-3.087	0.046
RISK	1.829	6.228	*RISK*	2.195	8.980

Tabelle 5.4. *Kaiserschnitt-Geburten: Geschätzte Parameter.*

Beispiel 5.4 Infektion nach Kaiserschnitt-Geburt

Die Zielvariable Y „Infektion" besitzt die drei Kategorien „Infektion vom Typ 1" ($Y = 1$), „Infektion vom Typ 2" ($Y = 2$) und „keine Infektion" ($Y = 3$). Mit $Y = 3$ als Referenzkategorie kodieren wir die Zielvariable durch $\boldsymbol{y}' = (y_1, y_2)$, mit

$$\begin{array}{ll} \boldsymbol{y}' = (1,0) & \text{Infektion vom Typ 1,} \\ \boldsymbol{y}' = (0,1) & \text{Infektion vom Typ 2,} \\ \boldsymbol{y}' = (0,0) & \text{keine Infektion.} \end{array}$$

Als Kovariablen verwenden wir wie in Beispiel 4.2 (Seite 196) die binären Kovariablen *NPLAN*, *RISK* und *ANTIB*. Die in der Kontingenztabelle von Beispiel 5.1 wiedergegebenen Daten ergeben die gruppierte Darstellung in Tabelle 5.3.

Als multikategoriales Logit-Modell wählen wir ein Haupteffekt-Modell in der Form (5.6)

$$\log \frac{\mathrm{P}(\text{Infektion vom Typ } r)}{\mathrm{P}(\text{keine Infektion})} = \beta_{r0} + \beta_{rN} NPLAN + \beta_{rR} RISK + \beta_{rA} ANTIB, \qquad r = 1, 2$$

bzw. äquivalent dazu

$$\frac{\mathrm{P}(\text{Infektion vom Typ } r)}{\mathrm{P}(\text{keine Infektion})} = \exp(\beta_{r0}) \exp(\beta_{rN})^{NPLAN} \exp(\beta_{rR})^{RISK} \exp(\beta_{rA})^{ANTIB}.$$

Somit ist $\exp(\beta_r)$ der multiplikative Effekt einer binären Kovariablen für die Kategorie r, wenn diese den Wert 1 statt 0 hat. Alternativ kann man $\exp(\beta_r)$ auch als das Verhältnis von relativen Risiken (odds ratio) für die Werte 1 und 0 interpretieren. Zum Beispiel ist für $NPLAN = 1$ und $NPLAN = 0$

$$\exp(\beta_{rN}) = \frac{\mathrm{P}(\text{Typ } r \,|\, NPLAN = 1, R, A)}{\mathrm{P}(\text{keine} \,|\, NPLAN = 1, R, A)} \Big/ \frac{\mathrm{P}(\text{Typ } r \,|\, NPLAN = 0, R, A)}{\mathrm{P}(\text{keine} \,|\, NPLAN = 0, R, A)}.$$

Tabelle 5.4 gibt die aus den Daten geschätzten Parameter wieder.

Danach senken Antibiotika das relative Risiko für beide Infektionen, und zwar für den Typ 1 etwas stärker als für den Typ 2. Andererseits erhöht sich das relative Risiko bei Vorliegen von Risikofaktoren und nicht geplantem Kaiserschnitt anscheinend deutlich. Ob die Unterschiede in den Effekten für Typ 1 und Typ 2 tatsächlich signifikant sind, wird in Abschnitt 5.4 genauer untersucht.

△

Nominale Modelle und latente Nutzenmodelle

Das multikategoriale Logit-Modell sowie weitere kategoriale Regressionsmodelle, insbesondere Probit-Modelle, lassen sich mit Hilfe latenter Nutzenmodelle motivieren und ableiten. Dies ist nicht nur für alternative Interpretationen von Interesse, sondern auch nützlich für die statistische Inferenz, insbesondere für Bayes-Ansätze (vergleiche hierzu auch Abschnitt 8.5.3).

Man nimmt an, dass jeder der c Alternativen des Entscheidungsproblems ein bestimmter Nutzen (Utility) zugrundeliegt. Die zufällige Nutzenfunktion der r-ten Alternative besitze die Form

$$U_r = u_r + \varepsilon_r,$$

wobei u_r eine Konstante (der Nutzen der r-ten Alternative) und ε_r eine Zufallsvariable mit Verteilungsfunktion F ist. Nach dem Prinzip der Nutzenmaximierung wird die Alternative r beobachtet, wenn der (unbeobachtbare) Nutzen maximal ist, d.h.

$$Y = r \Longleftrightarrow U_r = \max_{s=1,\ldots,c} U_s.$$

Nimmt man unabhängige Störgrößen ε_r und für F die Extremwertverteilung $F(x) = \exp(-\exp(-x))$ an, so erhält man (McFadden 1973)

$$\mathrm{P}(Y = r) = \frac{\exp(u_r)}{\sum_{s=1}^{c} \exp(u_s)}, \qquad r = 1, \ldots, c.$$

Da nur Differenzen von Nutzenfunktionen identifizierbar sind, wählt man eine Alternative als Referenzkategorie. Wählt man die Referenzkategorie c, ergibt sich unmittelbar

$$\mathrm{P}(Y = r) = \frac{\exp(u_r - u_c)}{1 + \sum_{s=1}^{q} \exp(u_s - u_c)} = \frac{\exp(\tilde{u}_r)}{1 + \sum_{s=1}^{q} \exp(\tilde{u}_s)}, \tag{5.7}$$

wobei $\tilde{u}_r = u_r - u_c$. Man kann auch definitorisch den Nutzen U_c der Referenzkategorie als Null annehmen, so dass $U_c = u_c \equiv 0$ gilt. Dann ist (5.7) mit linearen Nutzenfunktionen

$$u_r = \boldsymbol{x}'\boldsymbol{\beta} = \beta_{r0} + \beta_{r1}x_1 + \ldots + \beta_{rk}x_k$$

bereits in der üblichen Form des kategorialen Logit-Modells mit der Nutzenfunktion u_r als linearem Prädiktor η_r darstellbar.

Eine Erweiterung dieses Konzepts ergibt sich, wenn der Nutzen der Alternative nicht nur von Variablen abhängt, die den Entscheidungsträger charakterisieren. Bezeichne $\boldsymbol{x}_i$ diese bisher schon einbezogenen Variablen und $\boldsymbol{w}_{ir}$ einen Vektor von Variablen, die spezifisch für die Alternative r sind. Bei der Wahl des Transportmodus kann dies der Fahrpreis oder die Fahrdauer sein, die durchaus von Person zu Person variieren kann (daher der Doppelindex in $\boldsymbol{w}_{ir}$, wobei i für den Entscheidungsträger, r für die Alternative steht). Die

Annahme der Unabhängigkeit der extremwertverteilten Störgrößen führt zum erweiterten Logit-Modell

$$\pi_{ir} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_r + (\boldsymbol{w}_{ir} - \boldsymbol{w}_{ic})'\boldsymbol{\gamma})}{1 + \sum_{s=1}^{q} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_s + (\boldsymbol{w}_{is} - \boldsymbol{w}_{ic})'\boldsymbol{\gamma})},$$

$r = 1, \ldots, q$. Das Modell besitzt *kategorienunspezifische* (*globale*) Einflussgrößen $\boldsymbol{x}_i$, die mit *kategorienspezifischen* Koeffizienten $\boldsymbol{\beta}_r$ eingehen, und *kategorienspezifische Einflussgrößen* $\boldsymbol{w}_{ir} - \boldsymbol{w}_{ic}$, die mit *kategorienunspezifischen* (*globalen*) Koeffizienten $\boldsymbol{\gamma}$ eingehen.

Andere Verteilungsannahmen für die Störgrößen im latenten Nutzenmodell führen zu alternativen Modellen. Nimmt man weiterhin unabhängige, jedoch standard-normalverteilte ε_r an, so erhält man das ("unabhängige") Probit-Modell. Wählt man für den Vektor $\boldsymbol{\varepsilon}$ der Störgrößen eine multivariate Normalverteilung (mit nichtdiagonaler Kovarianzmatrix), so erhält man das multivariate Probit-Modell.

5.3 Ordinale Modelle

In diesem Abschnitt werden Modelle behandelt, die dann geeignet sind, wenn die abhängige Variable Werte in geordneten Kategorien annimmt. Die explizite Nutzung der Ordinalskala von Y macht sparsamere Parametrisierungen möglich. Sowohl im Beispiel des Waldzustands (Beispiel 5.2) als auch im Beispiel der Lungenfunktion (Beispiel 5.3) ist die abhängige Variable ordinal.

Das kumulative oder Schwellenwert-Modell

Das am häufigsten angewandte Modell der ordinalen Regression beruht, wie im binären Fall, auf dem Schwellenwertkonzept, vergleiche Abschnitt 4.1.1 (Seite 193). Das Modell lässt sich motivieren durch die Annahme einer latenten (unbeobachtbaren) Variablen, die hinter den beobachteten Kategorien steht. Das Überschreiten einer Schwelle auf dem latenten Kontinuum drückt sich durch eine beobachtbare Kategorie der Zielvariablen aus.

Bezeichne U die zugrundeliegende, nicht beobachtbare Variable. Im Beispiel des Waldzustands oder des Lungenfunktionstests ist U vorstellbar als die nicht unmittelbar beobachtbare Schädigung eines Baumes bzw. der Lunge. Die latente Variable zur Kovariablenausprägung $\boldsymbol{x}_i$ sei bestimmt durch

$$U_i = -\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{5.8}$$

wobei $\boldsymbol{\beta}$ ein Parametervektor und ε_i eine Störvariable mit Verteilungsfunktion F ist. Die Verknüpfung zwischen Beobachtung und latenter Variable sei bestimmt durch das Schwellenwertkonzept

$$Y_i = r \Longleftrightarrow \theta_{r-1} < U_i \leq \theta_r, \qquad r = 1, \ldots, q, \tag{5.9}$$

wobei $-\infty = \theta_0 < \theta_1 < \ldots < \theta_c = \infty$ die auf dem latenten Kontinuum liegenden Schwellenwerte sind. Die Beobachtung erfolgt in Kategorie r, wenn die latente Variable zwischen den Schwellenwerten θ_{r-1} und θ_r liegt. In binärer Betrachtungsweise erhält man eine Zielvariable $Y_i \leq r$, wenn $U_i \leq \theta_r$, also wenn die latente Variable unter einer Schwelle θ_r bleibt, bzw. $Y_i > r$ wenn $U_i > \theta_r$. Man beachte, dass im binären Logit-Modell

Abb. 5.1. *Schwellenwerte und Dichten der latenten Variablen.*

in Abschnitt 4.1.1 (Seite 193) der Schwellenwert ohne Beschränkung der Allgemeinheit auf $\theta = 0$ gesetzt werden konnte, da dies lediglich eine Verschiebung der Konstanten bewirkt.

Zur Veranschaulichung des Schwellenwertkonzeptes betrachte man Abbildung 5.1. Dort ist die Dichte der latenten Variable $U_i = -\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$ dargestellt für zwei unterschiedliche Ausprägungen von $\boldsymbol{x}_i$. Die zugehörigen Dichten besitzen dieselbe Form ($\varepsilon_i \sim F$), sind aber auf dem latenten Kontinuum verschoben. Die Fläche zwischen zwei Schwellen gibt jeweils die Wahrscheinlichkeit einer bestimmten Kategorie wieder. In Abbildung 5.1 ist exemplarisch die Wahrscheinlichkeit für die zweite Kategorie grau eingefärbt. Man sieht deutlich, wie sich diese Wahrscheinlichkeit mit der durch $\boldsymbol{x}_i$ bewirkten Verschiebung verändert. Als Dichte wurde die logistische Verteilung gewählt.

Aus den beiden Annahmen (5.8) und (5.9) ergibt sich unmittelbar das *kumulative Modell mit Verteilungsfunktion F* durch

$$\mathrm{P}(Y_i \leq r \,|\, \boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i'\boldsymbol{\beta}), \qquad r = 1, \ldots, c. \tag{5.10}$$

Die Bezeichnung „kumulatives Modell“ bezieht sich darauf, dass auf der linken Seite die kumulierten Wahrscheinlichkeiten $\mathrm{P}(Y_i \leq r) = \mathrm{P}(Y_i = 1) + \ldots + \mathrm{P}(Y_i = r)$ stehen. Das Modell selbst enthält die latente Variable nicht mehr. Es lässt sich als ein Regressionsmodell verstehen mit den Regressoren $\boldsymbol{x}_i$ und den Parametern $\theta_1, \ldots, \theta_q$ und $\boldsymbol{\beta}$. Die Wahrscheinlichkeiten sind gegeben durch

$$\begin{aligned} \mathrm{P}(Y_i = 1 \,|\, \boldsymbol{x}_i) &= F(\theta_1 + \boldsymbol{x}_i'\boldsymbol{\beta}), \\ \mathrm{P}(Y_i = r \,|\, \boldsymbol{x}_i) &= F(\theta_r + \boldsymbol{x}_i'\boldsymbol{\beta}) - F(\theta_{r-1} + \boldsymbol{x}_i'\boldsymbol{\beta}), \quad r = 2, \ldots, q. \end{aligned} \tag{5.11}$$

Je nach Wahl von F erhält man verschiedene Modellvarianten. Für die logistische Verteilungsfunktion F erhält man das *kumulative Logit-Modell*

$$\mathrm{P}(Y_i \leq r \,|\, \boldsymbol{x}_i) = \frac{\exp(\theta_r + \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\theta_r + \boldsymbol{x}_i'\boldsymbol{\beta})} \tag{5.12}$$

oder äquivalent dazu

$$\log \frac{\mathrm{P}(Y_i \leq r \,|\, \boldsymbol{x}_i)}{\mathrm{P}(Y_i > r \,|\, \boldsymbol{x}_i)} = \theta_r + \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Kumulative Modelle

Daten

Die Zielvariable $Y_i \in \{1, \ldots, c\}$ ist kategorial und ordinalskaliert. Zusätzlich liegen Kovariablen $\boldsymbol{x}_i$ vor, die nicht von der Kategorie abhängen.

Modell

Die Zielvariable Y_i wird mit einer latenten Variable U_i durch

$$Y_i = r \quad \Longleftrightarrow \quad \theta_{r-1} < U_i \leq \theta_r, \qquad r = 1, \ldots, q$$

verknüpft, mit $-\infty = \theta_0 < \theta_1 < \ldots < \theta_c = \infty$. Für U_i gelte ein lineares Modell

$$U_i = -\boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

mit Verteilungsfunktion F. Dann gilt für Y_i das *kumulative Modell*

$$\mathrm{P}(Y_i \leq r \,|\, \boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i' \boldsymbol{\beta}), \qquad r = 1, \ldots, c$$

bzw. äquivalent dazu

$$\mathrm{P}(Y_i = r \,|\, \boldsymbol{x}_i) = F(\theta_r + \boldsymbol{x}_i' \boldsymbol{\beta}) - F(\theta_{r-1} + \boldsymbol{x}_i' \boldsymbol{\beta}).$$

Das Modell wird auch als *Modell der proportionalen kumulativen Chancen* (proportional odds model) bezeichnet. Diese Bezeichnung nimmt Bezug auf die Eigenschaft, dass das Verhältnis der kumulierten Chancen für zwei durch $\boldsymbol{x}_i$ bzw. $\tilde{\boldsymbol{x}}_i$ charakterisierte Subpopulationen gegeben ist durch

$$\frac{\mathrm{P}(Y_i \leq r \,|\, \boldsymbol{x}_i)/\mathrm{P}(Y_i > r \,|\, \boldsymbol{x}_i)}{\mathrm{P}(Y_i \leq r \,|\, \tilde{\boldsymbol{x}}_i/\mathrm{P}(Y_i > r \,|\, \tilde{\boldsymbol{x}}_i)} = \exp((\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)' \boldsymbol{\beta}). \tag{5.13}$$

Dieses Verhältnis hängt nicht von der Kategorie r ab, d.h. die kumulierten Chancen sind proportional über alle Kategorien hinweg.

Wählt man für F die Minimum-Extremwertverteilung $F(x) = 1 - \exp(-\exp(x))$, erhält man das *kumulative Extremwert-* oder *gruppierte Cox-Modell*

$$\mathrm{P}(Y_i \leq r \,|\, \boldsymbol{x}_i) = 1 - \exp(-\exp(\theta_r + \boldsymbol{x}_i' \boldsymbol{\beta})) \tag{5.14}$$

bzw.

$$\log(-\log \mathrm{P}(Y_i > r \,|\, \boldsymbol{x}_i)) = \theta_r + \boldsymbol{x}_i' \boldsymbol{\beta}.$$

Die Bezeichnung „gruppiertes Cox-" oder „proportional Hazards-Modell" leitet sich von einer zu (5.13) analogen Eigenschaft eines zugrunde liegenden Modells der Verweildauer-analyse ab. Durch die Umparametrisierung $\tilde{\theta}_r = \log(\exp(\theta_r) - \exp(\theta_{r-1}))$, $r = 1, \ldots, q$, lässt sich (5.14) auch darstellen durch

$$\mathrm{P}(Y = r \,|\, Y \geq r, \boldsymbol{x}_i) = 1 - \exp(-\exp(\tilde{\theta}_r + \boldsymbol{x}_i' \boldsymbol{\beta})) = F(\tilde{\theta}_r + \boldsymbol{x}_i' \boldsymbol{\beta}) \tag{5.15}$$

Auf der linken Seite findet sich hier die Wahrscheinlichkeit für die Zielvariable in Kategorie r, gegeben es tritt die Kategorie r oder eine kleinere Kategorie auf. In der Verweildaueranalyse (wenn Y für die diskrete Zeit steht) wird diese bedingte Wahrscheinlichkeit als diskrete Hazardfunktion bezeichnet. Das Modell (5.15) ist nun die diskrete Variante des stetigen Cox-Modells der Verweildaueranalyse.

Das sequentielle Modell

Eine Alternative zum kumulativen Modell, die ebenso das ordinale Niveau der Kategorien der Zielvariablen explizit benutzt, ist das sequentielle Modell. Es ist aufgrund seiner Konstruktion dann geeignet, wenn die Kategorien der abhängigen Variablen nur sukzessive erreichbar sind. Geht man in Beispiel 5.3 davon aus, dass die Lungenfunktion ursprünglich normal ist, so wird (bei gegebenem Rauchverhalten und Alter) die Kategorie abnormal nur erreicht, wenn vorher einmal ein Grenzzustand vorlag. Ein anderes Beispiel ist die Dauer der Arbeitslosigkeit in Jahren (und damit in Kategorien). Man kann nur zwei Jahre arbeitslos sein, wenn man zuvor mindestens ein Jahr arbeitslos war. Die der Modellbildung zugrundeliegende Annahme sind also sukzessive bzw. schrittweise erreichbare Kategorien der Variable $Y \in \{1, \ldots, c\}$.

Die schrittweise Erreichbarkeit wird explizit durch eine Abfolge von dichotomen Übergängen modelliert. Der Prozess startet in $Y_i = 1$ und der Übergang zu $Y_i > 1$ wird durch ein binäres Modell aus Kapitel 4 modelliert:

$$\mathrm{P}(Y_i = 1 \,|\, \boldsymbol{x}_i) = F(\beta_{01} + \boldsymbol{x}_i'\boldsymbol{\beta}).$$

Verbleibt die Zielvariable in Kategorie 1 stoppt der Prozess, ansonsten wird der folgende Übergang wieder dichotom modelliert, mit

$$\mathrm{P}(Y_i = 2 \,|\, Y_i \geq 2, \boldsymbol{x}_i) = F(\beta_{02} + \boldsymbol{x}_i'\boldsymbol{\beta})$$

als bedingter Wahrscheinlichkeit für den Verbleib in $Y_i = 2$ und der komplementären Wahrscheinlichkeit $1 - \mathrm{P}(Y_i = 2 \,|\, Y_i \geq 2, \boldsymbol{x}_i)$ für den Übergang nach $Y_i > 2$.

Entsprechend ist der r-te Schritt bestimmt durch

$$\mathrm{P}(Y_i = r \,|\, Y_i \geq r, \boldsymbol{x}_i) = F(\beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta}). \tag{5.16}$$

Der Prozess stoppt, sobald ein Übergang nicht erfolgt. Man erhält als Gesamtmodell

$$\mathrm{P}(Y_i = r \,|\, Y_i \geq r, \boldsymbol{x}_i) = F(\beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta}), \qquad r = 1, \ldots, q. \tag{5.17}$$

Es ist zu beachten, dass nicht einzelne Übergänge beobachtet werden, sondern immer nur die tatsächliche Kategorie der Zielvariablen. Der sequentielle Mechanismus ist eine Annahme, die der Modellbildung dient. Die Modellformulierung spiegelt unmittelbar dieses Konstruktionsprinzip wieder. Alternativ lässt sich (5.17) in unbedingten Wahrscheinlichkeiten ausdrücken durch

$$\mathrm{P}(Y_i = r \,|\, \boldsymbol{x}_i) = F(\beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta}) \prod_{s=1}^{r-1} (1 - F(\beta_{0s} + \boldsymbol{x}_i'\boldsymbol{\beta})), \qquad r = 1, \ldots, q, \tag{5.18}$$

und

Sequentielle Modelle

Daten

Die Zielvariable $Y_i \in \{1, \ldots, c\}$ ist kategorial und ordinalskaliert, wobei die Kategorien nur schrittweise aufsteigend erreicht werden können.

Modell

Der Übergang zwischen den Kategorien wird durch binäre Regressionsmodelle modelliert. Ist die Kategorie $r \in \{1, \ldots, q\}$ bereits erreicht, ist der (potentielle) Verbleib in r durch

$$\mathrm{P}(Y_i = r \,|\, Y_i \geq r, \boldsymbol{x}_i) = F(\beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta})$$

bestimmt. Damit ergibt sich

$$\mathrm{P}(Y_i = r \,|\, \boldsymbol{x}_i) = F(\beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta}) \prod_{s=1}^{r-1} (1 - F(\beta_{0s} + \boldsymbol{x}_i'\boldsymbol{\beta})), \qquad r = 1, \ldots, q.$$

Ist F die logistische Verteilungsfunktion,so erhält man das sequentielle Logit-Modell, ist F die Verteilungsfunktion der Standardnormalverteilung, so ergibt sich das sequentielle Probit-Modell. Mit dem erweiterten Prädiktor (5.19) können auch kategorienspezifische Effekte von Kovariablen berücksichtigt werden.

$$\mathrm{P}(Y_i = c \,|\, \boldsymbol{x}_i) = 1 - \sum_{s=1}^{q} \mathrm{P}(Y_i = s \,|\, \boldsymbol{x}_i)$$

für die Referenzkategorie.

Durch Spezifikation der Verteilungsfunktion F erhält man wiederum das *logistische sequentielle Modell* (für die logistische Funktion F) oder das *sequentielle Extremwertmodell* (für $F(x) = 1 - \exp(-\exp(\boldsymbol{x}))$). Ebenso sind natürlich Probit-Modelle oder andere binäre Regressionsmodelle möglich. Wie man unmittelbar sieht, lässt sich das sequentielle Logit-Modell darstellen durch die bedingten Chancen des Übergangs

$$\log \frac{\mathrm{P}(Y_i = r \,|\, Y_i \geq r, x_i)}{1 - \mathrm{P}(Y_i = r \,|\, Y_i \geq r, x_i)} = \beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Eine Erweiterung des sequentiellen Modells, die bei einer kleinen Zahl von Kategorien sinnvoll sein kann, erhält man, wenn die Koeffizienten von ausgewählten Variablen als spezifisch für den Übergang zugelassen werden. Anstatt des linearen Terms $\beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta}$ in (5.16) erhält man dann den allgemeineren linearen Term

$$\beta_{0r} + \boldsymbol{x}_i'\boldsymbol{\beta}_r + \boldsymbol{z}_i'\boldsymbol{\beta}, \tag{5.19}$$

wobei $\boldsymbol{x}_i$ und $\boldsymbol{z}_i$ die Variablengruppen mit kategorienspezifischen Effekten $\boldsymbol{\beta}_r$ bzw. globalen Effekten $\boldsymbol{\beta}$ darstellen.

5.4 Schätzen und Testen: Likelihood-Inferenz

Das Schätzen der Parameter erfolgt wie für binäre Regressionsmodelle und allgemein für GLM wie in Kapitel 4 nach dem Maximum-Likelihood-Prinzip. Da die Zielvariablen multinomial verteilt sind mit

$$\boldsymbol{y}_i \,|\, \boldsymbol{x}_i \sim M(n_i, \boldsymbol{\pi}_i)$$

und als (bedingt) unabhängig angenommen werden, ergibt sich die Likelihoodfunktion als Produkt

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i \,|\, \boldsymbol{\pi}_i)$$

von Dichten $f(\boldsymbol{y}_i \,|\, \boldsymbol{\pi}_i)$ der Form (5.2). Dabei hängt $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'$ je nach dem spezifischen Modell über lineare Prädiktoren vom Vektor $\boldsymbol{\beta}$ aller Parameter ab. Für die Log-Likelihood erhält man bis auf eine additive Konstante

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_{i1} \log \pi_{i1} + \ldots + y_{iq} \log \pi_{iq} + y_{ic} \log \pi_{ic}),$$

$$\text{mit} \quad y_{ic} = n_i - y_{i1} - \ldots - y_{iq}, \quad \pi_{ic} = 1 - \pi_{i1} - \ldots - \pi_{iq}.$$

Daraus erhält man die Score-Funktion $s(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ als Vektor der 1. Ableitung und den ML-Schätzer durch iterative Lösung der ML-Gleichungen

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}.$$

Mit der erwarteten Informationsmatrix

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(s(\boldsymbol{\beta}) s'(\boldsymbol{\beta}))$$

gilt unter ähnlichen Regularitätsannahmen wie im univariaten Fall, dass der ML-Schätzer $\hat{\boldsymbol{\beta}}$ konsistent und asymptotisch bzw. approximativ normalverteilt ist

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \hat{\boldsymbol{A}}),$$

mit der geschätzten Kovarianzmatrix $\hat{\boldsymbol{A}} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$, also der inversen Informationsmatrix ausgewertet an der Stelle $\hat{\boldsymbol{\beta}}$.

Die Score-Funktion und die Informationsmatrix besitzen in multivariater Form die gleiche Struktur wie bei binären Regressionsmodellen und für univariate GLM in Kapitel 4. Zur expliziten Darstellung führen wir eine einheitliche Notation ein. In jedem kategorialen Regressionsmodell sind die Wahrscheinlichkeiten $\mathrm{P}(Y_i = r \,|\, \boldsymbol{x}_i) = \mathrm{P}(y_{ir} = 1 \,|\, \boldsymbol{x}_i) = \pi_{ir}$ für das Auftreten der Kategorie $r = 1, \ldots, q$ über Responsefunktionen mit linearen Prädiktoren η_{ir}, $r = 1, \ldots, q$ verknüpft, d.h. es gilt

$$\pi_{ir} = h_r(\eta_{i1}, \ldots, \eta_{iq}), \qquad r = 1, \ldots, q.$$

Die Responsefunktionen und die linearen Prädiktoren hängen dabei vom speziellen kategorialen Regressionsmodell ab.

Für das mehrkategoriale Logit-Modell (5.5) ist

$$\pi_{ir} = \frac{\exp(\eta_{ir})}{1 + \sum_{s=1}^{q} \exp(\eta_{is})} = h_r(\eta_{i1}, \ldots, \eta_{iq}) \tag{5.20}$$

mit $\eta_{ir} = \boldsymbol{x}_i'\boldsymbol{\beta}_r = \beta_{r0} + x_{i1}\beta_{r1} + \ldots + x_{ik}\beta_{rk}$. In der erweiterten Form mit kategorienspezifischen Einflussgrößen $\boldsymbol{w}_{ir} - \boldsymbol{w}_{ic}$ und globalen Parametern $\boldsymbol{\gamma}$ ist

$$\eta_{ir} = \boldsymbol{x}_i'\boldsymbol{\beta}_r + (\boldsymbol{w}_{ir} - \boldsymbol{w}_{ic})'\boldsymbol{\gamma}. \tag{5.21}$$

Für das ordinale kumulative Modell (5.11) ist

$$\pi_{ir} = F(\eta_{ir}) - F(\eta_{i,r-1}) = h_r(\eta_{i1}, \ldots, \eta_{iq}), \qquad r = 1, \ldots, q,$$

mit

$$\eta_{ir} = \theta_r + x_{i1}\beta_1 + \ldots + x_{ik}\beta_k. \tag{5.22}$$

Fasst man die Prädiktoren zum Prädiktorvektor

$$\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iq})$$

und alle Parameter zu einem Parametervektor $\boldsymbol{\beta}$ zusammen, so erhält man mit geeigneten Designmatrizen $\boldsymbol{X}_i$ immer die lineare Form

$$\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}.$$

Für das mehrkategoriale Logit-Modell mit dem erweiterten Prädiktor (5.21) ist $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q, \boldsymbol{\gamma})'$ und

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_i' & & & \boldsymbol{w}_{i1}' - \boldsymbol{w}_{ic}' \\ & \boldsymbol{x}_i' & & \boldsymbol{w}_{i2}' - \boldsymbol{w}_{ic}' \\ & & \ddots & \vdots \\ & & \boldsymbol{x}_i' & \boldsymbol{w}_{iq}' - \boldsymbol{w}_{ic}' \end{pmatrix}.$$

Für ordinale kumulative Modelle ist $\boldsymbol{\beta} = (\theta_1, \ldots, \theta_q, \beta_1, \ldots, \beta_k)'$ und

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & & & \boldsymbol{x}_i' \\ & 1 & & \boldsymbol{x}_i' \\ & & \ddots & \vdots \\ & & 1 & \boldsymbol{x}_i' \end{pmatrix}.$$

Führt man noch die q-dimensionalen Responsefunktionen $\boldsymbol{h}(\boldsymbol{\eta}_i) = (h_1(\boldsymbol{\eta}_i), \ldots, h_q(\boldsymbol{\eta}_i))'$ ein, so können alle kategorialen Regressionsmodelle für $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iq})'$ in kompakter Matrixnotation geschrieben werden als

$$\boldsymbol{\pi}_i = \boldsymbol{h}(\boldsymbol{\eta}_i), \qquad \boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}.$$

Die Score-Funktion $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ hat dann in q-dimensionaler Erweiterung die gleiche Struktur wie für univariate GLM in Kapitel 4. Es gilt

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{D}_i\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i),$$

wobei $\boldsymbol{D}_i = \partial\boldsymbol{h}(\boldsymbol{\eta}_i)/\partial\boldsymbol{\eta}$ die Matrix der partiellen Ableitungen an der Stelle $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$ bezeichnet und $\boldsymbol{\Sigma}_i$ die Kovarianzmatrix

$$\boldsymbol{\Sigma}_i = \frac{1}{n_i}\begin{pmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \cdots & -\pi_{i1}\pi_{iq} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & & \\ \vdots & & \ddots & \vdots \\ -\pi_{iq}\pi_{i1} & \cdots & & \pi_{iq}(1-\pi_{iq}) \end{pmatrix}$$

der (skalierten) Multinomialverteilung. Mit der (von $\boldsymbol{\beta}$ abhängigen) Gewichtsmatrix

$$\boldsymbol{W}_i = \boldsymbol{D}_i \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i'$$

lässt sich die Fisher-Matrix in der q-dimensionalen Erweiterung

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}$$

schreiben. Fasst man die Beobachtungen und Wahrscheinlichkeiten zusammen zu $\boldsymbol{y}' = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')$ und $\boldsymbol{\pi}' = (\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_n')$ und bildet die Gesamt-Designmatrix

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix}$$

sämtlicher Beobachtungen und die Blockdiagonalmatrizen

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_i), \quad \boldsymbol{D} = \operatorname{diag}(\boldsymbol{D}_i), \quad \boldsymbol{W} = \operatorname{diag}(\boldsymbol{W}_i),$$

so erhält man noch

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \quad \boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}.$$

Numerische Bestimmung des ML-Schätzers

Zur iterativen Bestimmung lässt sich wiederum das in Abschnitt 4.1.2 bereits betrachtete Fisher-Scoring-Verfahren einsetzen, für das die $(k+1)$-te Iteration gegeben ist durch

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

In der Darstellung als iterativ gewichtete KQ-Schätzung erhält man

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})\tilde{\boldsymbol{y}}^{(k)},$$

wobei der „Arbeitsvektor" $\tilde{\boldsymbol{y}}^{(k)} = (\tilde{\boldsymbol{y}}_1(\hat{\boldsymbol{\beta}}^{(k)}), \ldots, \tilde{\boldsymbol{y}}_n(\hat{\boldsymbol{\beta}}^{(k)}))$ bestimmt ist durch die Komponenten

$$\tilde{\boldsymbol{y}}_i(\hat{\boldsymbol{\beta}}^{(k)}) = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(k)} + (\boldsymbol{D}_i^{-1}(\hat{\boldsymbol{\beta}}^{(k)}))'(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}^{(k)})).$$

Asymptotische Eigenschaften und Tests linearer Hypothesen

Die asymptotischen Eigenschaften des ML-Schätzers sind analog zum univariaten Fall. Unter relativ schwachen Regularitätsbedingungen betrachtet man wiederum für wachsenden Gesamtstichprobenumfang $n \to \infty$ den zugehörigen ML-Schätzer $\hat{\boldsymbol{\beta}}$ und erhält asymptotische Existenz, Konsistenz und die asymptotische Normalverteilung

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})).$$

Tests linearer Hypothesen der Form $H_0 : \boldsymbol{C\beta} = \boldsymbol{d}$ lassen sich völlig analog zum univariaten Fall (Kapitel 4) durchführen. Es sind nur die Score-Funktion und die Informationsmatrix durch die multivariaten Versionen zu ersetzen.

Beispiel 5.5 Infektion nach Kaiserschnitt-Geburt

Wir testen nun, ob sich die Effekte von *NPLAN* und *RISK* für beide Typen von Infektionen überhaupt signifikant unterscheiden und bilden somit die Hypothesen

$$H_0 : \beta_{1N} = \beta_{2N} \text{ und } \beta_{1R} = \beta_{2R}, \qquad H_1 : \beta_{1N} \neq \beta_{2N} \text{ oder } \beta_{1R} \neq \beta_{2R}.$$

Dies lässt sich als spezielle lineare Hypothese der Form $\boldsymbol{C\beta} = \boldsymbol{0}$ schreiben, wobei $\boldsymbol{C}$ den Rang 2 besitzt. Als Likelihood-Quotienten-Statistik ergibt sich $lq = 0.8467$. Zum Signifikanzniveau $\alpha = 5\%$ ist $\chi^2_{0.95}(2) = 5.99$, d.h. H_0 kann nicht abgelehnt werden. Somit unterscheiden sich die Effekte in den beiden Kategorien nicht signifikant.

△

Beispiel 5.6 Zustand des Waldes

Wir modellieren die Abhängigkeit der dreikategorialen Zielvariable Schädigung durch ein kumulatives Logit-Modell. Der Einfluss des Jahres (*jahr*) und des Bestandalters (*alter*) wird durch kubische Polynome approximiert, während alle übrigen metrischen Kovariablen aus Tabelle 1.5 des Beispiels 1.4 mit linearen Effekten eingehen. Die in Tabelle 1.5 als metrisch präsentierte Variable Dichte der Humusschicht in cm (*humus*) wird dabei als kategoriale Variable mit den fünf geordneten Kategorien *humus0* ($0 \leq humus \leq 1$), *humus1* ($1 < humus \leq 2$), *humus2* ($2 < humus \leq 3$), *humus2* ($3 < humus \leq 4$) und *humus4* ($humus > 4$). Die binären bzw. kategorialen Kovariablen sind effektkodiert, wobei als Referenzkategorie (mit -1 kodiert) jeweils die häufigste Kategorie gewählt wurde. Die räumliche Information, die durch den Standort des Baumes gegeben ist, wird hier vernachlässigt. Die Ergebnisse in Tabelle 5.5 und in Abbildung 5.2 sind daher eher illustrativ oder als vorläufig anzusehen. In Kapitel 8.4 werden die Daten dann durch ein flexibleres, geoadditives Modell analysiert werden. Die in Abbildung 5.2 wiedergegebenen nichtlinearen Effekte des Bestandsalters und des Jahres geben – wie sich in Kapitel 8.4 zeigen wird – auch nur grobe Hinweise auf die Kovariableneffekte. Es scheint so zu sein, dass jüngere und ältere Bäume generell eine geringere Schädigungstendenz aufweisen, während Bäume mittleren Alters (ca. 60 – 130 Jahre) besonders gefährdet sind. Der Kalendereffekt *jahr* spiegelt ebenfalls den zeitlichen Trend grob wider. Ein steiler Anstieg des Risikos für Schädigung beginnt 1983, flacht sich gegen Ende der 80er Jahre ab, während das Risiko dann ab ca. 1990 – 2000 kontinuierlich sinkt. In jüngster Zeit ist wieder ein ansteigender Trend zu beobachten.

Wir kommentieren nur einige ausgewählte lineare Effekte: Ein hoher Beschirmungsgrad (*schirm*) vermindert signifikant das Risiko von Schädigungen, ebenso wie Düngung und ein mäßig trockener Boden (*frische1*), wobei der letzte Effekt für Buche charakteristisch ist. Dagegen erhöhen feuchte Böden und reiner Laubwald die Wahrscheinlichkeit von Schädigungen signifikant. In Beispiel 8.6 (Seite 416) reanalysieren wir die Daten mit einem flexiblen geoadditiven Modell. Dabei zeigt sich, dass einige Effekte relativ sensitiv in Bezug auf die Modellwahl reagieren, insbesondere wenn räumliche Effekte miteinbezogen werden.

△

Variable	Koeffizient	Standard-abweichung	p-Wert	95% Konfidenz-intervall	
β_{01}	-0.601	1.431	0.674	-3.407	2.205
β_{02}	3.456	1.434	0.016	0.645	6.267
hang	<0.001	0.007	0.959	-0.013	0.014
schirm	-3.189	0.305	<0.001	-3.786	-2.591
hoehe	>-0.001	0.001	0.795	-0.003	0.002
grund	-0.021	0.007	0.003	-0.034	-0.007
ph	-0.860	0.251	0.001	-1.351	-0.369
artkat	0.396	0.074	<0.001	0.251	0.542
dueng	-0.445	0.125	0.001	-0.691	-0.200
humus0	-0.126	0.132	0.343	-0.385	0.134
humus2	0.310	0.116	0.008	0.083	0.537
humus3	0.169	0.135	0.212	-0.096	0.434
humus4	-0.274	0.179	0.125	-0.625	0.077
frische1	-0.627	0.151	<0.001	-0.923	-0.331
frische3	0.493	0.108	<0.001	0.280	0.705
alkali1	0.365	0.159	0.022	0.053	0.676
alkali3	-0.545	0.156	0.001	-0.851	-0.239
alkali4	0.923	0.222	<0.001	0.488	1.358

Tabelle 5.5. *Waldschäden: Schätzungen der Kovariableneffekte.*

Abb. 5.2. *Waldschäden: Polynomiale Schätzungen des Alterseffekts und des Trends.*

Beispiel 5.7 Lungenfunktionstest

Der Einfluss von Alter (A) und Rauchverhalten (R) auf das Testergebnis wird durch ein sequentielles Modell spezifiziert. Die Einflussvariablen liegen in Effektkodierung vor. Die Kovariable x_A (mit $x_A = 1$ für Alter < 40, $x_A = -1$ für Alter ≥ 40) kodiert das Alter und x_R^1 (mit $x_R^1 = 1$ für Nichtraucher, $x_R^1 = 0$ für frühere Raucher, $x_R^1 = -1$ für derzeitiger Raucher) sowie x_R^2 (mit $x_R^2 = 1$ für frühere Raucher, $x_R^1 = 0$ für Nichtraucher, $x_R^1 = -1$ für derzeitiger Raucher) kodieren die Kovariable Rauchverhalten. Zusätzlich zu den Haupteffekten ist es notwendig, Interaktionseffekte zwischen Alter und Rauchverhalten zu berücksichtigen.

In Tabelle 5.6 sind Schätzungen für zwei Modellvarianten gegeben. In beiden Modellen geht das Alter mit globalen Koeffizienten in der Form $x_A\beta_A$ ein. Für das Rauchverhalten wird sowohl global modelliert mit $x_R^1\beta_{R(1)} + x_R^2\beta_{R(2)}$ (Modell 1) als auch kategorienspezifisch mit $x_R^1\beta_{R(1),1} + x_R^2\beta_{R(2),1}$ für den ersten Übergang von normal zu grenzwertig

Modell 1			Modell 2		
β_{01}	2.379	<0.001	β_{01}	2.379	<0.001
β_{02}	1.510	<0.001	β_{02}	1.516	<0.001
β_A	0.092	0.385	β_A	0.094	<0.000
$\beta_{R(1)1}$	0.915	<0.001			
$\beta_{R(1)2}$	0.675	0.108	$\beta_{R(1)}$	0.882	<0.001
$\beta_{R(2)1}$	-0.375	0.008			
$\beta_{R(2)2}$	-0.163	0.609	$\beta_{R(2)}$	-0.356	0.008
$\beta_{A*R(1),1}$	-0.561	0.003			
$\beta_{A*R(1),2}$	-0.894	0.047	$\beta_{A*R(1)}$	-0.601	0.001
$\beta_{A*R(2),1}$	0.015	0.912			
$\beta_{A*R(2),2}$	0.532	0.161	$\beta_{A*R(2)}$	0.092	0.492

Tabelle 5.6. *Lungenfunktionstest: Schätzungen für sequentielle Modelle.*

und mit $x_R^1\beta_{R(1),2} + x_R^2\beta_{R(2),2}$ für den zweiten Übergang von grenzwertig zu abnormal (Modell 2). Entsprechend gehen die Interaktionen global oder kategorienspezifisch ein. Wie man deutlich sieht, setzt sich bei der Vereinfachung zu globalen Effekten jeweils der Effekt für den ersten Übergang stärker durch. Das ist dadurch begründet, dass für den ersten Übergang wesentlich mehr Daten zur Verfügung stehen als für den zweiten Übergang. Der Interaktionseffekt zeigt deutlich, dass der starke Einfluss des Rauchens (0.882 für Nichtraucher) vor allem in Verbindung mit dem Alter wirksam wird ($\beta_{A*R(1)} = -0.601$ relativiert den Nichtrauchereffekt von 0.882).

△

5.5 Bemerkungen und Ergänzungen

Die Konzepte der Bayes-Inferenz für GLM, insbesondere für binäre Regressionsmodelle, lassen sich in kompletter Analogie zu Kapitel 4.6 auf kategoriale Regressionsmodelle erweitern. Für die *volle Bayes-Inferenz* kann man also wieder mit dem dort skizzierten MH-Algorithmus und der Normalverteilung $\mathrm{N}(\boldsymbol{\mu}^{(t)}, (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1})$ als Vorschlagsdichte arbeiten. Dabei sind $\boldsymbol{\mu}^{(t)}$ und die Fisher-Matrix $\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X}$ durch die entsprechenden mehrkategorialen Versionen aus Abschnitt 5.4 zu ersetzen.

Für Probit-Modelle ergibt sich eine alternative Möglichkeit zur vollen Bayes-Inferenz, indem man die Darstellung über latente Nutzenmodelle (Abschnitt 5.2) oder über Schwellenwert-Modelle (Abschnitt 5.3) nutzt. Dabei werden die latenten Variablen $U_i = \eta_i + \varepsilon_i$ als unbekannte Parameter angesehen und in einem zusätzlichen Ziehungsschema ebenfalls aus ihren (vollständig) bedingten Dichten gezogen. Für aktuelle Werte $U_i^{(t)}$ kann dann $\hat{\boldsymbol{\beta}}^{(t)}$ direkt aus Normalverteilungsdichten, ohne MH-Schritt, gezogen werden, siehe Albert & Chib (1993), Chen & Dey (2000) sowie Abschnitt 8.5.3 und Fahrmeir & Lang (2001b) für die Erweiterung auf strukturiert-additive Regressionsmodelle.

Kategoriale Regressionsmodelle sind in den Büchern von Tutz (2000) und Agresti (2002) ausführlich beschrieben. Knappere Darstellungen sind in Fahrmeir et al. (1996, Kap. 6) und Fahrmeir & Tutz (2001, Kap. 3) enthalten. Die Herleitung von kategorialen Regressionsmodellen und allgemeineren diskreten Wahlmodellen (Discrete Choice Models) geht auf McFadden zurück, der dafür mit dem Nobelpreis für Wirtschaftswissenschaften ausgezeichnet wurde, siehe z.B. McFadden (1984) oder Train (2003).

6 Gemischte Modelle

Gemischte Modelle (mixed models) beziehen in den Prädiktor $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ linearer, generalisierter linearer und kategorialer Modelle neben den bislang betrachteten festen Effekten auch zufällige Effekte oder Koeffizienten ein. Man spricht deshalb auch von Modellen mit zufälligen Effekten (random effects models). Ein wichtiger Anwendungsbereich dieser Modelle ist die Analyse von Longitudinaldaten, die sich durch zeitlich wiederholte Beobachtungen von Individuen bzw. Objekten im Rahmen von Längsschnittstudien ergeben. Eng verwandt damit ist die Analyse von sogenannten Clusterdaten, wenn aus Primäreinheiten (Cluster, Klumpen) mehrere Objekte ausgewählt und zugehörige Variablen erhoben werden. In Querschnittsstudien können die Cluster oder Primäreinheiten zum Beispiel Kliniken, Schulen oder Firmen sein, in denen jeweils an einer – unter Umständen auch kleinen – Teilstichprobe von Patienten, Schülern oder Kunden Daten zu interessierenden Variablen erhoben werden. In beiden Fällen liegen für jedes Individuum bzw. für jeden Cluster $i = 1, \ldots, m$ wiederholte Beobachtungen

$$(y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ij}, \ldots, \boldsymbol{x}_{in_i})$$

für Ziel- und Kovariablen vor. Bei Longitudinaldaten bezeichnet y_{ij} eine Beobachtung an Individuum i zum Zeitpunkt t_{ij}, während bei Clusterdaten y_{ij} der Wert der Zielvariablen für das j-te Objekt aus dem Cluster i ist, und n_i ist jeweils die Anzahl wiederholter Beobachtungen. Gemischte Modelle ermöglichen es, individuen- oder clusterspezifische Effekte zu modellieren und zu schätzen, und zwar insbesondere für den in Anwendungen oft auftretenden Fall relativ kurzer Längen n_i der individuellen Zeitreihen bzw. kleiner Umfänge n_i der Teilstichproben aus den Clustern $i = 1, \ldots, m$. Dazu wird der lineare Prädiktor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ für die Beobachtung y_{ij} zu $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i$ erweitert. Dabei ist $\boldsymbol{u}_{ij}$ oft ein Teilvektor der Kovariablen und $\boldsymbol{\gamma}_i$ ein Vektor individuen- bzw. clusterspezifischer zufälliger Effekte. Lineare und generalisierte lineare Modelle mit festen Effekten sind dazu in der Regel unzureichend geeignet, da eine im Vergleich zum Stichprobenumfang zu große Zahl freier Parameter zu schätzen ist. Ein zusätzlicher Vorteil gemischter Modelle ist, dass Korrelationen, die infolge der wiederholten Beobachtungen pro Individuum bzw. Cluster entstehen, berücksichtigt werden können.

Wir führen im Folgenden zunächst lineare gemischte Modelle (LMM, linear mixed models) für Longitudinal- und Clusterdaten ein und beschränken uns dabei auf die klassische, von sogenannten Varianzkomponentenmodellen herrührende Annahme, dass die zufälligen Effekte unabhängig und identisch gemäß einer gemeinsamen Mischverteilung verteilt sind. Anschließend erweitern wir die Modellklasse, indem auch korrelierte zufällige Effekte zugelassen werden. Damit erhält man eine äußerst breite Familie von Modellen, die sich durch geeignete Spezifikation von Korrelationsstrukturen unter anderem auch zur Analyse von räumlich und zeitlich-räumlich korrelierten Regressionsdaten eignet.

Der zweite Teil dieses Kapitels erweitert in analoger Weise generalisierte lineare Modelle für nicht-normale Zielvariablen zu generalisierten linearen gemischten Modellen (GLMM, generalized linear mixed models). Im Vergleich zu linearen gemischten Modellen beruht

die statistische Inferenz auf den gleichen Konzepten, wird aber in der Durchführung schwieriger und algorithmisch aufwändiger.

Von Bedeutung ist darüber hinaus die Möglichkeit, semiparametrische Regressionsmodelle, die auf Glättungs- oder Penalisierungsansätzen beruhen, durch gemischte Modelle darzustellen. Somit werden existierende Inferenztechniken gemischter Modelle für moderne semiparametrische Regressionsansätze in den Kapiteln 7 und 8 nutzbar gemacht, vergleiche insbesondere die Abschnitte 7.1.9 und 8.5.

6.1 Lineare gemischte Modelle für Longitudinal- und Clusterdaten

Wir illustrieren die Idee linearer gemischter Modelle zunächst an zwei Beispielen. Zunächst greifen wir nochmal Beispiel 2.9 (Seite 35) auf.

Beispiel 6.1 Hormontherapie bei Ratten

Um die Wirkung von Testosteron auf das Wachstum von Ratten zu untersuchen, wurden insgesamt 50 Ratten zufällig einer Kontrollgruppe oder einer von zwei Therapiegruppen zugewiesen. Als Therapie wurde dabei eine niedrige oder eine hohe Dosis des Mittels Decapeptyl gegeben, mit dem die Testosteronproduktion gehemmt wird. Die Behandlung begann im Alter von 45 Tagen. Beginnend mit dem 50. Tag, wurde alle 10 Tage das Wachstum des Kopfes mittels Röntgenuntersuchung gemessen. Zielvariable war der Abstand (gemessen in Pixeln) zwischen zwei wohldefinierten Punkten des Kopfes, welche die Höhe des Kopfes charakterisieren. Die Anzahl n_i von wiederholten Messungen y_{ij}, $j = 1, \ldots, n_i$ dieser Zielvariable war für die Ratten unterschiedlich. Die nach den drei Gruppen getrennten, individuellen Zeitreihen $\{y_{ij}, j = 1, \ldots, n_i\}$ für die Ratten $i = 1, \ldots, 50$ sind in Abbildung 2.10 (Seite 36) zu finden.

Mit dem logarithmisch transformierten Alter

$$t = \log(1 + (alter - 45)/10)$$

als Zeitskala ($t = 0$ entspricht dem Behandlungsbeginn), den Werten

$$t_{ij} = \log(1 + (alter_{ij} - 45)/10)$$

für die j-te Messung an der Ratte i, und den Indikatorvariablen

$$C_i = \begin{cases} 1, & \text{Ratte } i \text{ in Kontrollgruppe} \\ 0, & \text{sonst,} \end{cases}$$

$$N_i = \begin{cases} 1, & \text{Ratte } i \text{ in Niedrigdosisgruppe} \\ 0, & \text{sonst,} \end{cases}$$

$$H_i = \begin{cases} 1, & \text{Ratte } i \text{ in Hochdosisgruppe} \\ 0, & \text{sonst.} \end{cases}$$

wurde in Beispiel 2.9 das lineare gemischte Modell

$$y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 N_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} + \gamma_{1i} \cdot t_{ij} + \varepsilon_{ij} \tag{6.1}$$

mit individuenspezifischen Abweichungen γ_{0i} von der Populationskonstanten β_0 und individuenspezifischen Abweichungen γ_{1i} von den Populationssteigungen β_1, β_2 und β_3 formuliert.

Im Gegensatz zu den „fixen" Effekten $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ werden die individuenspezifischen Effekte $\boldsymbol{\gamma}_i = (\gamma_{0i}, \gamma_{1i})'$ als zufällige Größen angesehen, da die Ratten eine Zufallsauswahl aus der Population sind. Wir treffen dazu die spezifische Annahme, dass die zufälligen Effekte unabhängig und identisch normalverteilt sind mit

$$\boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{0i} \\ \gamma_{1i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{pmatrix} \right). \tag{6.2}$$

mit Erwartungswert $\mathbf{0} = (0, 0)'$ und Kovarianzmatrix

$$\boldsymbol{D} = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{pmatrix}.$$

Der Erwartungswert kann dabei ohne Einschränkung $\mathbf{0}$ gesetzt werden, da die Populationsmittelwerte bereits in den fixen Effekten $\boldsymbol{\beta}$ enthalten sind. Die Varianzen τ_0^2 und τ_1^2 geben jeweils an, wie stark die individuenspezifischen Effekte um die globale Konstante und die globale Steigung streuen. Durch die Kovarianz $\tau_{01} = \tau_{10}$ ist es möglich, Abhängigkeiten zwischen den individuenspezifischen Effekten zu berücksichtigen. Diese treten beispielsweise dann auf, wenn Individuen mit großer Steigung tendenziell eine kleinere Konstante aufweisen. In diesem Fall sind Konstante und Steigung negativ korreliert. Die Normalverteilung in (6.2) wird auch als *Mischverteilung* bezeichnet.

Für die Messfehler ε_{ij} nehmen wir in diesem Beispiel an, dass sie unabhängig und identisch normalverteilt sind, d.h.

$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2). \tag{6.3}$$

Die Abbildung 2.10 legt bei genauerer Betrachtung nahe, dass sich die individuenspezifischen Steigungen nicht stark unterscheiden und somit möglicherweise $\gamma_{1i} = 0$ gesetzt werden kann. Dies führt auf das vereinfachte Modell

$$y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 N_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} + \varepsilon_{ij}, \tag{6.4}$$

bei dem nur die konstanten Terme individuenspezifisch variieren. Durch geeignete statistische Inferenz möchte man untersuchen, ob (6.4) zur Modellierung ausreicht. Dabei gilt nun die Annahme, dass die γ_{0i} unabhängig und identisch $\mathrm{N}(0, \tau_0^2)$-verteilt sind.

△

Beispiel 6.2 Wortschatztest

Intelligenztests enthalten in der Regel sowohl verbale als auch sprachfreie Aufgaben. Der Wortschatztest zählt zum Verbalteil und besteht aus 42 Testaufgaben zum Erkennen von Wörtern. Jede einzelne Testaufgabe entspricht einer Zeile im Fragebogen. Diese enthält jeweils das zu erkennende Zielwort (z.B. evident) und fünf Distraktoren (z.B. elisent, devital, evelent, senisant, osalant). Richtig gelöst ist eine Aufgabe, falls der Proband das Zielwort erkannt und keinen Distraktor markiert hat. Die 42 Aufgaben bzw. Zeilen sind nach ansteigendem Schwierigkeitsgrad geordnet. Aus den gelösten und ungelösten Aufgaben wird dann ein Z-Wert errechnet, wobei höhere Z-Werte einer höheren Leistung beim Wortschatztest entsprechen. Die betrachteten Daten stammen von Patienten eines Krankenhauses, die chronisch an einer affektiven oder schizophrenen Psychose litten, sowie von deren Familienangehörigen. Da die Testergebnisse von Mitgliedern einer Familie vermutlich korreliert sind, handelt es sich um Clusterdaten: Die jeweiligen Familien bilden die Cluster $i = 1, \ldots, m$, und der Teilstichprobenumfang n_i entspricht der Anzahl der jeweils rekrutierten Familienmitglieder. Abbildung 6.1 zeigt die Häufigkeitsverteilung der Z-Werte für die insgesamt $n = 424$ Probanden aus $m = 199$ Familien. Ziel der Untersuchung war es unter anderem, herauszufinden, ob

Abb. 6.1. *Häufigkeitsverteilung der Z-Werte: Kerndichteschätzer (durchgezogene Linie) und Normalverteilungsapproximation (gestrichelt).*

die gestellte Diagnose sowie andere Kovariablen einen Einfluss auf den Wortschatztest haben. Tabelle 6.1 enthält die Merkmale, kategoriale Variablen sind dummy-kodiert.

Für die Analyse wählen wir ein gemischtes lineares Modell

$$y_{ij} = \beta_0 + \gamma_{0i} + \beta_{D1}\, D1 + \beta_{D2}\, D2 + \beta_{D3}\, D3 + \ldots + \beta_{M2}\, M2 + \beta_{M3}\, M3 + \varepsilon_{ij},$$

mit familienspezifischen Abweichungen γ_{0i} vom Populationswert β_0.

An dieser Stelle wollen wir noch einmal kurz die bereits in Kapitel 2.4 skizzierten Vorteile gemischter Modelle diskutieren. Da sich die Mitglieder einer Familie in Hinsicht auf die für das Testergebnis relevanten Eigenschaften ähnlicher sein dürften als beliebig ausgewählte Personen, sind die Testergebnisse der Familienmitglieder vermutlich positiv korreliert, so dass die Annahme unabhängiger Beobachtungen des klassischen linearen Modells nicht erfüllt sind. Theoretisch könnte man versuchen, die Ähnlichkeit der Familienmitglieder auf Kovariableneinflüsse zurückzuführen, beispielsweise genetische Veranlagungen. Häufig sind solche Kovariablen aber nicht für die Analyse erhältlich. In diesem Fall kann der familienspezifische Effekt γ_{0i} als Surrogat für diese (unbeobachteten) Kovariableneinflüsse angesehen werden. Darüberhinaus werden durch die stochastische Modellierung von γ_{0i} Korrelationen zwischen den Familienmitgliedern bewirkt, während Beobachtungen aus verschiedenen Familien weiterhin unabhängig bleiben. Berücksichtigt man die Korrelation der Daten nicht, so kann dies unter Umständen deutliche Auswirkungen auf die Schätzungen und insbesondere die geschätzten Standardabweichungen haben. Im Fall positiv korrelierter Zielvariablen führt der Einbezug zufälliger Effekte meist zu größeren Standardabweichungen, da der Tatsache Rechnung getragen wird, dass eine Stichprobe mit Gruppierungsstruktur typischerweise weniger Information enthält als eine Stichprobe unabhängiger Individuen. Dementsprechend ist die Berücksichtigung zufälliger Effekte auch notwendig, um gültige Konfidenzintervalle und Tests zu erhalten.

△

In allgemeiner Notation lassen sich gemischte lineare Modelle für Longitudinaldaten

$$(y_{ij}, \boldsymbol{x}_{ij}), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i$$

mit Beobachtungen zu den Zeitpunkten $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$ wie in Beispiel 6.1 in der Form

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i + \varepsilon_{ij} \tag{6.5}$$

Merkmal	Ausprägungen	Kodierung
Diagnose	- schizophren	$D_1 = 1$
	- affektiv erkrankt	$D_2 = 1$
	- minder schwer psychiatrisch erkrankt	$D_3 = 1$
	- gesund	Referenzkategorie
Schulbildung	- kein Abschluss	$S_1 = 1$
	- Hauptschulabschluss	Referenzkategorie
	- Realschulabschluss	$S_2 = 1$
	- Gymnasium abgebrochen	$S_3 = 1$
	- Abitur/Fachabitur	$S_4 = 1$
	- abgeschlossenes Studium	$S_5 = 1$
Geschlecht	- männlich	G=0
	- weiblich	G=1
Alter	Alter bei Erhebung	A metrisch
CGI	Clinical Global Impression: allgemeiner Gesundheitszustand, auf einer Skala von 0-8 bewertet (je höher der Wert, desto schwerer der Erkrankungsgrad)	C metrisch
Medikation	- keine	Referenzkategorie
	- Standardneuroleptika/Neuroleptika in atypischer Kombination	$M_1 = 1$
	- atypische Neuroleptika	$M_2 = 1$
	- sonstige	$M_3 = 1$

Tabelle 6.1. *Wortschatztest: Zusätzlich zur Testleistung erhobene Merkmale.*

schreiben. Dabei ist $\boldsymbol{x}_{ij}$ ein Vektor von Kovariablen, die zeitlich variieren dürfen, wie das transformierte Alter t_{ij} im Beispiel (6.1), oder zeitkonstant sein können, wie die Indikatorvariablen N_i, H_i und C_i. In der Regel enthält $\boldsymbol{x}_{ij}$ die 1 als eine Komponente, um eine feste Konstante β_0 in das Modell aufzunehmen. Dies gilt auch für den Vektor $\boldsymbol{u}_{ij}$, der neben der 1 auch weitere Komponenten umfassen kann, die oft aus bereits in $\boldsymbol{x}_{ij}$ enthaltenen Kovariablen bestehen, wie t_{ij} im Modell (6.1).

Für eine kompakte Matrixnotation fasst man die Zielvariablen y_{ij}, Kovariablenvektoren $\boldsymbol{x}_{ij}$, $\boldsymbol{u}_{ij}$ und die Fehlervariablen ε_{ij}, $j = 1, \ldots, n_i$, für jedes Individuum $i = 1, \ldots, m$ zu Vektoren bzw. Designmatrizen

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \vdots \\ \boldsymbol{x}'_{ij} \\ \vdots \\ \boldsymbol{x}'_{in_i} \end{pmatrix}, \quad \boldsymbol{U}_i = \begin{pmatrix} \boldsymbol{u}'_{i1} \\ \vdots \\ \boldsymbol{u}'_{ij} \\ \vdots \\ \boldsymbol{u}'_{in_i} \end{pmatrix}, \quad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ij} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix} \tag{6.6}$$

zusammen und erhält

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, \qquad \mathrm{E}(\boldsymbol{\varepsilon}_i) = \boldsymbol{0} \tag{6.7}$$

als Beobachtungsmodell für Individuum $i = 1, \ldots, m$. Durch zusätzliche Annahmen über die zufälligen Effekte $\boldsymbol{\gamma}_i$ und die Fehler $\boldsymbol{\varepsilon}_i$ ergeben sich Modelle verschiedener Komplexität. Für die zufälligen Effekte nehmen wir in diesem Abschnitt an, dass sie wie in

Beispiel 6.1 unabhängig und identisch normalverteilt sind,

$$\boldsymbol{\gamma}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D}), \tag{6.8}$$

mit der $(q \times q)$-Matrix $\boldsymbol{D}$ als Kovarianzmatrix. Diese kann auch von spezieller Struktur sein. Insbesondere entspricht der Spezialfall unabhängiger Komponenten einer Diagonalmatrix mit Elementen $\tau_j^2 = \mathrm{Var}(\gamma_{ij})$. Man spricht dann auch von einem *Varianzkomponentenmodell*. Dieser Spezialfall wird sehr häufig in praktischen Anwendungen verwendet, da sich insbesondere die Modellkomplexität deutlich verringert.

Die strengste Annahme für die Fehler ist, dass sie wie im Beispiel 6.1 unabhängig und identisch $\mathrm{N}(0, \sigma^2)$-verteilt sind. Daraus folgt

$$\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_i}), \tag{6.9}$$

und die Unabhängigkeit von $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m$. Wegen der Annahme (6.9) wird die Korrelation zwischen den wiederholten Beobachtungen eines Individuums nur durch den für diese Beobachtungen gemeinsamen Vektor $\boldsymbol{\gamma}_i$ von zufälligen Effekten erzeugt. Um flexiblere Modelle zu erhalten, erweitern wir (6.9), indem wir

$$\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_i) \tag{6.10}$$

mit beliebiger $(n_i \times n_i)$-Kovarianzmatrix $\boldsymbol{\Sigma}_i$ zulassen. Die spezielle Struktur von $\boldsymbol{\Sigma}_i$ ergibt sich aus Annahmen über den korrelierten Fehler-Prozess $\{\varepsilon_{ij}\}$. So führt beispielsweise die Annahme autoregressiver Fehler zu Kovarianzmatrizen wie sie in Abschnitt 3.4.4 behandelt wurden. Man beachte, dass $\boldsymbol{\Sigma}_i$ von i nur über die Anzahl n_i der Beobachtungen für das Individuum bzw. Cluster i abhängt.

Für Clusterdaten ändert sich bei gleicher Notation nur die Interpretation: Zu jedem Cluster $i = 1, \ldots, m$ liegen n_i Beobachtungen vor, die durch den gemeinsamen clusterspezifischen Vektor $\boldsymbol{\gamma}_i$ von zufälligen Effekten und durch die Kovarianzmatrix $\boldsymbol{\Sigma}_i$ korreliert sind.

Wie im linearen Regressionsmodell kann man für einige inferentielle Aussagen auf die Normalverteilungsannahme verzichten. Da aber die übliche Inferenzmethodik zum Schätzen unbekannter Elemente oder Parameter in $\boldsymbol{D}$ und $\boldsymbol{\Sigma}_i$ auf dem Likelihood-Prinzip basiert, setzen wir die Normalverteilung als Standardannahme voraus und weisen auf einige Aussagen, die ohne diese Annahme gültig sind, in Bemerkungen hin.

Beispiel 6.3 Hormontherapie bei Ratten

Die Messungen erfolgen für alle Ratten, solange sie noch an der Studie teilnehmen, im gleichen Alter von $a = 50, 60, \ldots, 110$ Tagen. Für das logarithmisch transformierte Alter $t = \log(1 + (a - 45)/10)$ entspricht dies den Werten

$$t_1 = \log(1.5), \quad t_2 = \log(2.5), \quad \ldots, \quad t_7 = \log(7.5)\,.$$

Ratte 17 aus der Niedrigdosisgruppe $(N_{17} = 1, H_{17} = 0, C_{17} = 0)$ wurde fünf mal gemessen. Somit erhalten wir aus Modell (6.1)

$$\boldsymbol{y}_{17} = \begin{pmatrix} y_{17,1} \\ \vdots \\ y_{17,5} \end{pmatrix}, \quad \boldsymbol{X}_{17} = \begin{pmatrix} 1 & \log(1.5) & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \log(5.5) & 0 & 0 \end{pmatrix}, \quad \boldsymbol{U}_{17} = \begin{pmatrix} 1 & \log(1.5) \\ \vdots & \vdots \\ 1 & \log(5.5) \end{pmatrix}.$$

Lineares gemischtes Modell für Longitudinal- und Clusterdaten

Ein lineares gemischtes Modell (LMM) für Longitudinal- oder Clusterdaten ist durch

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, m$$

gegeben. Dabei ist $\boldsymbol{y}_i$ der n_i-dimensionale Vektor von Zielvariablen für das Individuum bzw. Cluster i, m ist die Anzahl der Individuen bzw. Cluster, $\boldsymbol{X}_i$ und $\boldsymbol{U}_i$ sind $(n_i \times p)$- und $(n_i \times q)$- dimensionale Designmatrizen zu bekannten Kovariablen, $\boldsymbol{\beta}$ ist der p-dimensionale Vektor fester Effekte, $\boldsymbol{\gamma}_i$ ist ein q-dimensionaler Vektor individuen- bzw. clusterspezifischer Effekte und $\boldsymbol{\varepsilon}_i$ ein n_i-dimensionaler Fehlervektor. Für $\boldsymbol{\gamma}_i$ und $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, m$ gelten die Verteilungsannahmen

$$\begin{aligned}
&\boldsymbol{\gamma}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D}) \\
&\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i) \\
&\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m \quad \text{sind unabhängig.}
\end{aligned}$$

Für Ratte 12 aus der Hochdosisgruppe ($N_{12} = 0, H_{12} = 1, C_{12} = 0$) mit vier Messungen ergibt sich

$$\boldsymbol{y}_{12} = \begin{pmatrix} y_{12,1} \\ \vdots \\ y_{12,4} \end{pmatrix}, \quad \boldsymbol{X}_{12} = \begin{pmatrix} 1 & 0 & \log(1.5) & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \log(4.5) & 0 \end{pmatrix}, \quad \boldsymbol{U}_{12} = \begin{pmatrix} 1 & \log(1.5) \\ \vdots & \vdots \\ 1 & \log(4.5) \end{pmatrix}$$

und für Ratte 22 aus der Kontrollgruppe ($N_{22} = 0, H_{22} = 0, C_{22} = 1$)

$$\boldsymbol{y}_{22} = \begin{pmatrix} y_{22,1} \\ \vdots \\ y_{22,7} \end{pmatrix}, \quad \boldsymbol{X}_{22} = \begin{pmatrix} 1 & 0 & 0 & \log(1.5) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \log(7.5) \end{pmatrix}, \quad \boldsymbol{U}_{22} = \begin{pmatrix} 1 & \log(1.5) \\ \vdots & \vdots \\ 1 & \log(7.5) \end{pmatrix}.$$

Da wir unabhängig und identisch verteilte Messfehler annehmen, ist

$$\boldsymbol{\Sigma}_{12} = \sigma^2 \boldsymbol{I}_4, \quad \boldsymbol{\Sigma}_{17} = \sigma^2 \boldsymbol{I}_5, \quad \boldsymbol{\Sigma}_{22} = \sigma^2 \boldsymbol{I}_7.$$

Für die Kovarianzmatrix $\boldsymbol{D}$ der zufälligen Effekte werden wir später auch vereinfachend eine Diagonalmatrix

$$\boldsymbol{D} = \begin{pmatrix} \tau_0^2 & 0 \\ 0 & \tau_1^2 \end{pmatrix}$$

annehmen. Für das Modell (6.4) ohne individuenspezifische „Steigungen" $\gamma_{1i} \cdot t_{ij}$ entfällt für die Designmatrizen $\boldsymbol{U}_{12}$, $\boldsymbol{U}_{17}$, $\boldsymbol{U}_{22}$ die zweite Spalte.

△

6.2 Das allgemeine lineare gemischte Modell

Das Longitudinaldaten-Modell $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i$, $i = 1, \ldots, m$ lässt sich noch kompakter in der Form eines allgemeinen linearen gemischten Modells schreiben, wenn wir alle Beobachtungen zusammenfassen. Wir definieren dazu die Vektoren

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_i \\ \vdots \\ \boldsymbol{y}_m \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_i \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_i \\ \vdots \\ \boldsymbol{\gamma}_m \end{pmatrix} \tag{6.11}$$

von allen Ziel- und Fehlervariablen sowie zufälligen Effekten. Mit den Designmatrizen

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_i \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}, \quad \boldsymbol{U} = \operatorname{diag}(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_i, \ldots, \boldsymbol{U}_m) = \begin{pmatrix} \boldsymbol{U}_1 & & & & \boldsymbol{0} \\ & \ddots & & & \\ & & \boldsymbol{U}_i & & \\ & & & \ddots & \\ \boldsymbol{0} & & & & \boldsymbol{U}_m \end{pmatrix} \tag{6.12}$$

erhalten wir als Gesamtmodell

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mit $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{R})$, $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G})$ und den blockdiagonalen Kovarianzmatrizen

$$\boldsymbol{R} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_i, \ldots, \boldsymbol{\Sigma}_m), \qquad \boldsymbol{G} = \operatorname{diag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_i, \ldots, \boldsymbol{D}_m), \tag{6.13}$$

wobei $\boldsymbol{D}_i \equiv \boldsymbol{D}$. Damit sind lineare gemischte Modelle für Longitudinal- und Clusterdaten ein Spezialfall der folgenden allgemeinen Definition:

Lineares gemischtes Modell (LMM)

In allgemeiner Form ist ein lineares gemischtes Modell durch

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{6.14}$$

mit

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \right) \tag{6.15}$$

gegeben. Dabei sind $\boldsymbol{X}$ und $\boldsymbol{U}$ Designmatrizen, $\boldsymbol{\beta}$ ein Vektor von fixen Effekten und $\boldsymbol{\gamma}$ ein Vektor von zufälligen Effekten. Die Kovarianzmatrizen für $\boldsymbol{\gamma}$ und $\boldsymbol{\varepsilon}$ werden als nichtsingulär (und damit positiv definit) angenommen, $\boldsymbol{\gamma}$ und $\boldsymbol{\varepsilon}$ sind unabhängig.

Bemerkungen:

1. Die Normalverteilungsannahme in (6.15) ist nicht für alle Aussagen zur Inferenz in LMM notwendig. Wir weisen darauf in den nächsten Abschnitten hin. Die übliche Schätzung unbekannter Parameter in $\boldsymbol{G}$ und $\boldsymbol{R}$ basiert jedoch auf dem Likelihood-Prinzip, so dass wir die Normalverteilungsannahme in die Definition mit aufnehmen.

2. Das LMM (6.14), (6.15) lässt sich auch als *zweistufiges, hierarchisches Modell* schreiben: Die *bedingte Verteilung von* $\boldsymbol{y}$ *gegeben* $\boldsymbol{\gamma}$ ist gegeben durch

$$\boldsymbol{y} \,|\, \boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{X\beta} + \boldsymbol{U\gamma}, \boldsymbol{R}) \tag{6.16}$$

und wird durch die (Rand-) Verteilung

$$\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}) \tag{6.17}$$

für die zufälligen Effekte vervollständigt. Ein *marginales Modell*, d.h. die Randverteilung von $\boldsymbol{y}$, ergibt sich, indem man das LMM in der Form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}^*, \qquad \boldsymbol{\varepsilon}^* = \boldsymbol{U\gamma} + \boldsymbol{\varepsilon}$$

schreibt. Dies ergibt das allgemeine lineare Modell

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \boldsymbol{R} + \boldsymbol{UGU}') \tag{6.18}$$

bzw. $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\eta}, \boldsymbol{V})$ mit dem linearen Prädiktor $\boldsymbol{\eta} = \boldsymbol{X\beta}$ für die Erwartungswert-Struktur und der Kovarianzmatrix-Struktur

$$\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{UGU}'. \tag{6.19}$$

Beim Vergleich von zweistufiger und marginaler Darstellung ist folgendes wichtig: Aus der Definition bzw. der hierarchischen Darstellung (6.16), (6.17) lässt sich zwar das marginale Modell (6.18) ableiten, umgekehrt kann man aber aus der marginalen Verteilung von $\boldsymbol{y}$ alleine nicht die bedingte Verteilung für $\boldsymbol{y}$ gegeben $\boldsymbol{\gamma}$ und die Verteilung von $\boldsymbol{\gamma}$ rekonstruieren. Dies hat auch Konsequenzen für die Schätzung: Ist man nur an den fixen Effekten $\boldsymbol{\beta}$ interessiert, kann man das marginale Modell verwenden. Dabei können auch Kovarianzmatrizen $\boldsymbol{V}$ zum Einsatz kommen, die sich nicht in der strukturierten Form (6.19) fassen lassen. Will man jedoch auch die zufälligen Effekte, z.B. individuenspezifische Effekte in LMM für Longitudinaldaten, schätzen, so benötigt man die zweistufige Darstellung.

6.3 Likelihood-Inferenz für LMM

Dieser Abschnitt behandelt zunächst die auf (verallgemeinerten) KQ- oder ML-Ansätzen basierende Schätzung der fixen und zufälligen Effekte sowie die Schätzung unbekannter Parameter in $\boldsymbol{G}$ und $\boldsymbol{R}$. Bei den zufälligen Effekten spricht man auch von Vorhersage bzw. Prädiktion statt Schätzung. Das Testen von Hypothesen wird in Abschnitt 6.3.4 behandelt. Die Bayesianische Inferenz ist Inhalt von Abschnitt 6.5

6.3.1 Schätzung fixer und zufälliger Effekte bei bekannter Kovarianzstruktur

Wir setzen zunächst voraus, dass $\boldsymbol{R}$ und $\boldsymbol{G}$ und somit $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{UGU}'$ vollständig bekannt sind. Vom marginalen Modell $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \boldsymbol{V})$ ausgehend, bietet sich für $\boldsymbol{\beta}$ der ML- bzw. verallgemeinerte KQ-Schätzer

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{6.20}$$

an, der bei bekanntem $\boldsymbol{V}$ verschiedene Optimalitätseigenschaften besitzt, vgl. Abschnitt 3.4. Für die Herleitung eines Schätzers für den Vektor $\boldsymbol{\gamma}$ zufälliger Effekte gibt es verschiedene Zugänge. Unter der Normalverteilungsannahme ist der bedingte Erwartungswert $\mathrm{E}(\boldsymbol{\gamma}\,|\,\boldsymbol{y})$ von $\boldsymbol{\gamma}$, gegeben die Daten $\boldsymbol{y}$, der beste Schätzer. Zur Herleitung betrachten wir zunächst die gemeinsame Verteilung von $\boldsymbol{y}$ und $\boldsymbol{\gamma}$. Sie ergibt sich aus dem zweistufigen hierarchischen Modell für $\boldsymbol{y}\,|\,\boldsymbol{\gamma}$ und $\boldsymbol{\gamma}$ zu

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{\gamma}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{X\beta}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{V} & \boldsymbol{UG}\\ \boldsymbol{GU}' & \boldsymbol{G}\end{pmatrix}\right).$$

Daraus erhält man gemäß Satz B.4 (Seite 465) aus Anhang B

$$\mathrm{E}(\boldsymbol{\gamma}\,|\,\boldsymbol{y}) = \boldsymbol{GU}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{6.21}$$

Formel (6.21) kann auch ohne Normalverteilungsannahme als *bester linearer Prädiktor* (BLP) abgeleitet werden. Ersetzt man in (6.21) den unbekannten Vektor $\boldsymbol{\beta}$ durch den Schätzer $\tilde{\boldsymbol{\beta}}$, so ergibt dies den Schätzer

$$\tilde{\boldsymbol{\gamma}} = \boldsymbol{GU}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \tag{6.22}$$

für die zufälligen Effekte.

Die gleichen Schätzer erhält man auch nach anderen Konzepten, bei denen $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ simultan und symmetrisch betrachtet werden. Eine Möglichkeit ist die Herleitung von $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ als *beste lineare unverzerrte Prädiktion* (BLUP). Eine weitere Möglichkeit besteht darin, die gemeinsame (Log-) Likelihood von $\boldsymbol{y}, \boldsymbol{\gamma}$ simultan bezüglich $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ zu maximieren. Da dies in engem Zusammenhang zu (empirischer) Bayes-Schätzung und penalisierter Likelihood-Schätzung steht, gehen wir darauf näher ein.

Wegen $p(\boldsymbol{y}, \boldsymbol{\gamma}) = p(\boldsymbol{y}\,|\,\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ ergibt sich (bis auf additive Konstanten) für die Log-Likelihood

$$-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{U\gamma})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{U\gamma}) - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}^{-1}\boldsymbol{\gamma}\,.$$

Die Maximierung dieses Ausdrucks ist äquivalent zur Minimierung des *penalisierten KQ-Kriteriums*

$$\mathrm{KQ}_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{U\gamma})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{U\gamma}) + \boldsymbol{\gamma}'\boldsymbol{G}^{-1}\boldsymbol{\gamma}. \tag{6.23}$$

Der erste Term entspricht einem (verallgemeinerten) KQ-Kriterium. Ohne den zweiten Term würde damit auch $\boldsymbol{\gamma}$ wie ein fixer Effekt geschätzt werden. Der zweite Term berücksichtigt, dass $\boldsymbol{\gamma}$ aus einer Verteilung stammt. Da $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G})$ mit positiv definiter Kovarianzmatrix $\boldsymbol{G} > 0$ angenommen wird, bestraft der zweite Term $\boldsymbol{\gamma}'\boldsymbol{G}^{-1}\boldsymbol{\gamma}$ Abweichungen von $\boldsymbol{0}$, und zwar umso mehr bzw. weniger, je „kleiner" bzw. „größer" $\boldsymbol{G}$ ist. Man spricht deshalb auch von einem Straf- oder Penalisierungsterm. Im Grenzfall $\boldsymbol{G} \to \infty$ bzw. $\boldsymbol{G}^{-1} \to 0$ geht der Penalisierungsterm gegen 0 und $\boldsymbol{\gamma}$ wird wie ein fixer Effekt ohne Verteilungsrestriktionen geschätzt.

Differenzieren von $KQ_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ und Nullsetzen der Ableitungen ergibt die *Schätzgleichungen (mixed model equations)*

$$\begin{pmatrix}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{U}\\ \boldsymbol{U}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{U}'\boldsymbol{R}^{-1}\boldsymbol{U} + \boldsymbol{G}^{-1}\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\gamma}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y}\\ \boldsymbol{U}'\boldsymbol{R}^{-1}\boldsymbol{y}\end{pmatrix}. \tag{6.24}$$

Matrixumformungen zeigen, dass die Lösung $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ zu (6.20) und (6.22) äquivalent ist. Definiert man $\boldsymbol{C} = (\boldsymbol{X}, \boldsymbol{U})$ und die partitionierte Matrix

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} \end{pmatrix},$$

so lässt sich die Lösung der Schätzgleichung (6.24) auch in der Form

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\gamma}} \end{pmatrix} = (\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{B})^{-1}\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{y} \tag{6.25}$$

schreiben. In dieser Form erkennt man die enge Beziehung zur Ridge-Schätzung (vergleiche Abschnitt 3.6.4).

6.3.2 Schätzung der Kovarianzstruktur

Wir beschreiben in diesem Abschnitt den Maximum-Likelihood (ML)- und restringierten Maximum-Likelihood (REML, restricted ML)-Ansatz zur Schätzung unbekannter Parameter $\boldsymbol{\vartheta}$ in $\boldsymbol{R}, \boldsymbol{G}$ und $\boldsymbol{V}$. Diese beiden Schätzansätze sind derzeit am gebräuchlichsten und unter anderem in SAS, S-PLUS und R implementiert.
Um die Abhängigkeit von $\boldsymbol{\vartheta}$ zu betonen, schreiben wir

$$\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\vartheta}) = \boldsymbol{U}\boldsymbol{G}(\boldsymbol{\vartheta})\boldsymbol{U}' + \boldsymbol{R}(\boldsymbol{\vartheta}).$$

Die ML-Schätzung von $\boldsymbol{\vartheta}$ geht meist von der Likelihood des marginalen Modells

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}(\boldsymbol{\vartheta}))$$

aus. Die Log-Likelihood für $\boldsymbol{\beta}$ und $\boldsymbol{\vartheta}$ ist, von additiven Konstanten abgesehen,

$$l(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = -\frac{1}{2}\{\log|\boldsymbol{V}(\boldsymbol{\vartheta})| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}. \tag{6.26}$$

Maximiert man $l(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ für festes $\boldsymbol{\vartheta}$ bezüglich $\boldsymbol{\beta}$, so erhält man

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

Einsetzen in $l(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ liefert die *Profil-Log-Likelihood*

$$l_P(\boldsymbol{\vartheta}) = -\frac{1}{2}\{\log|\boldsymbol{V}(\boldsymbol{\vartheta})| + (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\vartheta}))'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\vartheta}))\}. \tag{6.27}$$

Maximieren von $l_P(\boldsymbol{\vartheta})$ bezüglich $\boldsymbol{\vartheta}$ liefert den ML-Schätzer $\hat{\boldsymbol{\vartheta}}_{ML}$.
Statt $l_P(\boldsymbol{\vartheta})$ verwendet man zur Schätzung von $\boldsymbol{\vartheta}$ jedoch vorzugsweise die *marginale Log-Likelihood*

$$l_R(\boldsymbol{\vartheta}) = \log\left(\int L(\boldsymbol{\beta}, \boldsymbol{\vartheta})d\boldsymbol{\beta}\right),$$

die man erhält, indem $\boldsymbol{\beta}$ aus der Likelihood herausintegriert wird. Die marginale Log-Likelihood lässt sich im Kontext linearer Modelle auch alternativ als *restringierte Log-Likelihood* herleiten, vergleiche Abschnitt 3.2.2 und die folgende Diskussion. Aus beiden Herleitungen ergibt sich

$$l_R(\boldsymbol{\vartheta}) = l_P(\boldsymbol{\vartheta}) - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{X}|. \tag{6.28}$$

Maximieren von $l_R(\boldsymbol{\vartheta})$ liefert den restringierten ML-Schätzer $\hat{\boldsymbol{\vartheta}}_{REML}$.
Wie in Abschnitt 3.2.2 besprochen, beseitigt der REML-Schätzer

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2$$

die Verzerrung des ML-Schätzers für die Fehlervarianz σ^2 im linearen Modell. Ebenso ist auch bei LMM die Reduktion der Verzerrung von $\hat{\boldsymbol{\vartheta}}_{ML}$ der Hauptgrund für die Verwendung von $\hat{\boldsymbol{\vartheta}}_{REML}$ als Schätzer von $\boldsymbol{\vartheta}$. Es ist jedoch allgemein nicht gesichert, ob auch der MSE geringer wird. Die Herleitung der restringierten Log-Likelihood kann zum Beispiel mit Hilfe einer „linearen Kontrastmatrix" $\boldsymbol{A} \neq \boldsymbol{0}$ erfolgen, die so konstruiert ist, dass das lineare Modell für den transformierten Vektor $\tilde{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$ nicht mehr von $\boldsymbol{\beta}$ abhängt. Eine alternative Herleitung aus Bayesianischer Sichtweise zeigt, dass $\hat{\boldsymbol{\vartheta}}_{REML}$ als Posteriori-Modus-Schätzer bei Vorgabe einer diffusen Priori-Verteilung für $\boldsymbol{\vartheta}$ aufgefasst werden kann, siehe Abschnitt 6.5.
Die konkrete numerische Berechnung von $\hat{\boldsymbol{\vartheta}}_{REML}$ oder $\hat{\boldsymbol{\vartheta}}_{ML}$ erfolgt iterativ, z.B. durch Maximierung von $l_R(\boldsymbol{\vartheta})$ oder $l_P(\boldsymbol{\vartheta})$ bzw. Minimierung von $-2l_R(\boldsymbol{\vartheta})$ oder $-2l_P(\boldsymbol{\vartheta})$ mit einem Newton-Raphson- oder Fisher-Scoring-Algorithmus. Da sich $l_R(\boldsymbol{\vartheta})$ und $l_P(\boldsymbol{\vartheta})$ nur durch den von $\boldsymbol{\beta}$ unabhängigen Term $-0.5\log|\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{X}|$ in (6.28) unterscheiden, kann man $\boldsymbol{\beta}$ und $\boldsymbol{\vartheta}$ auch simultan durch Maximierung von

$$l(\boldsymbol{\beta}, \boldsymbol{\vartheta}) - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{X}|$$

erhalten. Auch ein Schaukelalgorithmus, der abwechselnd Iterationswerte $\hat{\boldsymbol{\vartheta}}^{(k)}$ und $\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\vartheta}}^{(k)})$, $\hat{\boldsymbol{\gamma}}^{(k+1)} = \hat{\boldsymbol{\gamma}}(\hat{\boldsymbol{\vartheta}}^{(k)})$ aus den Schätzgleichungen (6.24) bestimmt, ist möglich. Einsetzen von $\hat{\boldsymbol{\vartheta}}$ nach Konvergenz ergibt die geschätzten Kovarianzmatrizen

$$\hat{\boldsymbol{R}} = \boldsymbol{R}(\hat{\boldsymbol{\vartheta}}),\ \hat{\boldsymbol{G}} = \boldsymbol{G}(\hat{\boldsymbol{\vartheta}}) \quad \text{und damit} \quad \hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\vartheta}}).$$

6.3.3 Schätzung fixer und zufälliger Effekte

Die Verwendung der geschätzten Kovarianzmatrizen $\hat{\boldsymbol{R}}, \hat{\boldsymbol{G}}$ und $\hat{\boldsymbol{V}}$ in den Schätzformeln für die fixen Effekte $\boldsymbol{\beta}$ und die zufälligen Effekte $\boldsymbol{\gamma}$ aus Abschnitt 6.3.1 für die dort als bekannt vorausgesetzten Matrizen $\boldsymbol{R}, \boldsymbol{G}$ und $\boldsymbol{V}$ liefert schließlich die Schätzer $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\gamma}}$. Durch das Einsetzen von $\hat{\boldsymbol{\vartheta}}$ gelten jedoch die Optimalitätseigenschaften nicht mehr exakt. Ebenso sind die Varianzen bzw. Kovarianzmatrizen der Schätzer analytisch nicht exakt zugänglich.

Bemerkungen:

1. Bei bekannter Kovarianzstruktur lässt sich aus (6.29)

$$\operatorname{Cov}\begin{pmatrix}\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\end{pmatrix} = (\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{B})^{-1} \tag{6.30}$$

Likelihood-Inferenz für LMM

Schätzung der Kovarianzstruktur

Sei $\boldsymbol{\vartheta}$ der unbekannte Parametervektor in $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\vartheta}), \boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\vartheta}), \boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\vartheta})$. ML-Schätzer für $\boldsymbol{\vartheta}$ erhält man durch Maximieren der (Profil-) Log-Likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = -\frac{1}{2}\{\log|\boldsymbol{V}(\boldsymbol{\vartheta})| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}$$

mit

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

REML- (restringierte ML)-Schätzer erhält man durch Maximieren der restringierten Log-Likelihood

$$l_R(\boldsymbol{\vartheta}) = l_P(\boldsymbol{\vartheta}) - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{X}|.$$

Schätzer für feste und zufällige Effekte

Schätzer $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\gamma}}$ für die fixen bzw. zufälligen Effekte ergeben sich als

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}\\ \hat{\boldsymbol{\gamma}} &= \hat{\boldsymbol{G}}\boldsymbol{U}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).\end{aligned}$$

bzw. äquivalent dazu als Lösung von

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\gamma}}\end{pmatrix} = (\boldsymbol{C}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{C} + \hat{\boldsymbol{B}})^{-1}\boldsymbol{C}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{y} \tag{6.29}$$

mit $\boldsymbol{C} = (\boldsymbol{X}, \boldsymbol{U})$ und $\hat{\boldsymbol{B}} = \begin{pmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \hat{\boldsymbol{G}}^{-1}\end{pmatrix}$. Diese Schätzer werden auch als *empirische beste lineare Schätzer* bezeichnet, kurz EBLP oder EBLUP.

als Posteriori-Kovarianzmatrix herleiten, siehe Abschnitt 6.5. Einsetzen von $\hat{\boldsymbol{R}}$ und $\hat{\boldsymbol{B}}$ ergibt die *geschätzte* bzw. *approximative Kovarianzmatrix*

$$\widehat{\mathrm{Cov}}\begin{pmatrix}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\\ \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\end{pmatrix} = \boldsymbol{A} = (\boldsymbol{C}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{C} + \hat{\boldsymbol{B}})^{-1}. \tag{6.31}$$

Man kann zeigen, dass $\boldsymbol{A}$ in partitionierter Form

$$\boldsymbol{A} = \begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix} \tag{6.32}$$

mit

$$\begin{aligned}
\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{A}_{11} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1} \\
\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= \boldsymbol{A}_{21} = -\hat{\boldsymbol{G}}\boldsymbol{U}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\boldsymbol{A}_{11} \\
\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\gamma}}) &= \boldsymbol{A}_{22} = (\boldsymbol{U}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{U} + \hat{\boldsymbol{G}}^{-1})^{-1} - \boldsymbol{A}_{21}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{U}\hat{\boldsymbol{G}}
\end{aligned}$$

geschrieben werden kann. Geschätzte Varianzen bzw. Standardabweichungen für die Komponenten von $\hat{\boldsymbol{\beta}}$ bzw. $\hat{\boldsymbol{\gamma}}$ erhält man dann aus entsprechenden Diagonalelementen von $\boldsymbol{A}$ bzw. $\boldsymbol{A}_{11}$ und $\boldsymbol{A}_{22}$.

2. Da beim Einsetzen der Schätzer $\hat{\boldsymbol{R}}, \hat{\boldsymbol{G}}$ und $\hat{\boldsymbol{V}}$ für $\boldsymbol{R}, \boldsymbol{G}$ und $\boldsymbol{V}$ ihre Variabilität nicht berücksichtigt wird, unterschätzt $\boldsymbol{A}$ tendenziell die tatsächliche Kovarianzmatrix von $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ und damit auch Varianzen und Standardabweichungen der Komponenten. Deshalb wurden verschiedene Alternativen vorgeschlagen, insbesondere Bayesianische Ansätze (vgl. Abschnitt 6.5).
3. Neben der approximativen Kovarianzmatrix $\boldsymbol{A}$ in (6.31) wird auch eine Alternative vorgeschlagen, bei der $\boldsymbol{\gamma}$ wie ein fixer Effekt behandelt, bzw. auf $\boldsymbol{\gamma}$ bedingt wird. Dann erhält man aus (6.29) nach den Regeln für Lineartransformationen wegen $\mathrm{Cov}(\boldsymbol{y} \,|\, \boldsymbol{\gamma}) = \boldsymbol{R}$

$$\mathrm{Cov}\left(\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\gamma}} \end{pmatrix} \,|\, \boldsymbol{\gamma}\right) = (\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{B})^{-1}\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{B})^{-1}.$$

   Diese Kovarianzmatrix besitzt die Form einer sogenannten „Sandwich“-Matrix, da sie wie ein Sandwich aus den Bestandteilen $(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{B})^{-1}$ und $\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}$ zusammengesetzt ist (vergleiche auch die Abschnitte 3.4.3, Seite 135, und 7.1.9, Seite 357). Nach Einsetzen der Schätzer $\hat{\boldsymbol{R}}, \hat{\boldsymbol{B}}$ erhält man als geschätzte Kovarianzmatrix

$$\boldsymbol{A}\hat{\boldsymbol{C}}'\hat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{C}}\boldsymbol{A}, \tag{6.33}$$

   wobei $\boldsymbol{A}$ durch (6.31) gegeben ist. Da im Vergleich zu $\boldsymbol{A}$ tendenziell eine zusätzliche Unterschätzung entsteht, präferieren wir (6.31). Aufgrund der Herleitung wird $\boldsymbol{A}$ manchmal als Bayesianische und (6.33) als frequentistische Version der Kovarianzmatrix bezeichnet.
4. Auch für die Schätzer $\hat{\boldsymbol{\vartheta}}_{ML}$ und $\hat{\boldsymbol{\vartheta}}_{REML}$ lassen sich mit Hilfe der Newton-Raphson- bzw. Fisher-Scoring-Iterationen Schätzungen für Informationsmatrizen bzw. Approximationen für $\mathrm{Cov}(\hat{\boldsymbol{\vartheta}})$ berechnen. Da die Schätzer, insbesondere von unbekannten Varianzen, schief verteilt sind, werden diese Approximationen relativ grob. Sie sind daher auch zur Konstruktion von Konfidenzintervallen oder Tests wenig geeignet.

Zur Konstruktion von Konfidenzintervallen bzw. Konfidenzbändern nimmt man an, dass die Verteilung von $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ asymptotisch, d.h. für große Stichproben, durch eine Normalverteilung approximiert werden kann. Für spezielle LMM, zum Beispiel Varianzkomponentenmodelle, oder LMM für Longitudinaldaten, lässt sich dies unter geeigneten Voraussetzungen auch nachweisen. Eine allgemein gültige asymptotische Normalitätsaussage ist jedoch nicht streng bewiesen. Mit Hilfe der Normalverteilungsapproximation lassen sich dann Quantile bzw. Konfidenzintervalle für (die Komponenten von) $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\gamma}}$ angeben.

6.3.4 Hypothesentests

In der Regel sind Hypothesen über fixe Effekte von primärem Interesse. Im einfachsten Fall möchte man für eine Komponente β_j von $\boldsymbol{\beta}$

$$H_0 : \beta_j = d_j \;\text{ gegen }\; H_1 : \beta_j \neq d_j$$

testen. Für $d_j = 0$ testet man damit wie im linearen Modell die Signifikanz des entsprechenden j-ten Effekts. Unter der Annahme einer approximativen Normalverteilung für $\hat{\boldsymbol{\beta}}$ kann der Test mit Hilfe eines (approximativen) Konfidenzintervalls für $\hat{\beta}_j$ durchgeführt werden. Liegt d_j außerhalb dieses Konfidenzintervalls, lehnt man H_0 zu einem vorgegebenen Signifikanzniveau α ab. Äquivalent dazu kann man die Teststatistik

$$t_j = \frac{\hat{\beta}_j - d_j}{\hat{\sigma}_j}$$

verwenden, wobei $\hat{\sigma}_j$ die Wurzel des j-ten Diagonalelements der approximativen Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ ist. Selbst unter Normalverteilungsannahme für $\boldsymbol{\gamma}$ und $\boldsymbol{\varepsilon}$ ist t_j unter H_0 nur in Spezialfällen exakt t-verteilt. Im Allgemeinen hofft man, dass t_j für große Stichproben approximativ standardnormalverteilt ist und lehnt H_0 ab, wenn $|t_j| > z_{1-\alpha/2}$ ist.

Allgemeinere lineare Hypothesen der Form

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \;\text{ gegen }\; H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d}$$

lassen sich durch die Wald-Statistik

$$W = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'(\boldsymbol{C}'\boldsymbol{A}_{11}\boldsymbol{C})^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$$

überprüfen.

Alternativ kann ein Likelihood-Quotienten-Test basierend auf der Log-Likelihood (6.26) für das unrestringierte Modell (unter H_1) und das restringierte Modell (unter H_0) durchgeführt werden. Wiederum hängt die Güte der Approximation der Verteilung der Teststatistik von der speziellen Struktur des zugrundeliegenden LMM ab.

Konzeptionell eignen sich Likelihood-Quotienten-Tests auch für Hypothesentests über unbekannte Parameter $\boldsymbol{\vartheta}$ in den Kovarianzmatrizen. Beispielsweise ist im folgenden, einfachen LMM für Longitudinaldaten

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \varepsilon_{ij}\,, \quad i = 1, \ldots, m\,,\; j = 1, \ldots, n\,,$$

mit $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$, $\gamma_{0i} \sim \mathrm{N}(0, \tau_0^2)$, die Hypothese $H_0 : \tau_0^2 = 0$ äquivalent zur Gültigkeit des einfachen linearen Modells

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

ohne die zufälligen, individuenspezifischen Effekte γ_{0i}.

Für den Test des entsprechenden Hypothesenpaares

$$H_0 : \tau_0^2 = 0 \;\text{ gegen }\; H_1 : \tau_0^2 > 0$$

kann die Likelihood-Quotienten (LQ)-Statistik basierend auf der Profil-Log-Likelihood (6.27) verwendet werden. Da jedoch $\tau_0^2 = 0$ auf dem Rand des Parameterraums $\{\tau_0^2 \geq 0\}$ liegt, ist die klassische asymptotische Likelihood-Test-Theorie nicht mehr gültig. In einfachen Spezialfällen lässt sich die Verteilung der LQ-Statistik zwar prinzipiell noch bestimmen (vergleiche etwa Crainiceanu & Ruppert, 2004), die zugehörigen Tests sind aber in den gängigen Statistik-Programmen noch nicht enthalten und zudem noch Gegenstand aktueller Forschung.

6.4 Likelihood-Inferenz für Longitudinal- und Clusterdaten-Modelle

Wir fassen die wesentlichen Ergebnisse des vorherigen Abschnitts für den Spezialfall von linearen gemischten Modellen für Longitudinal- und Clusterdaten zusammen und illustrieren anschließend die Inferenzmethodik anhand der Ratten-Studie und des Wortschatztests.

Für Longitudinaldaten-Modelle

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i\,,$$

mit $\boldsymbol{\gamma}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$ und $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ ergeben sich die in (6.12) und (6.13) beschriebenen speziellen Matrizen $\boldsymbol{X}$, $\boldsymbol{U}$, $\boldsymbol{R}$ und $\boldsymbol{G}$ für das LMM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,.$$

Dem marginalen Modell (6.18) $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{V} = \boldsymbol{R}+\boldsymbol{U}\boldsymbol{G}\boldsymbol{U}')$ entsprechen dann marginale Modelle

$$\boldsymbol{y}_i \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta},\, \boldsymbol{V}_i = \boldsymbol{\Sigma}_i + \boldsymbol{U}_i\boldsymbol{D}\boldsymbol{U}_i') \quad i = 1, \ldots, m$$

für die unabhängigen Beobachtungen $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$. Für den Spezialfall unkorrelierter Messfehler ist

$$\boldsymbol{V}_i = \sigma^2\boldsymbol{I} + \boldsymbol{U}_i\boldsymbol{D}\boldsymbol{U}_i'.$$

Ist auch noch $\boldsymbol{D}$ diagonal mit Elementen $\tau_j^2 = \mathrm{Var}(\gamma_{ij})$, $j = 1, \ldots, q$, so ist der Vektor $\boldsymbol{\vartheta}$ unbekannter Parameter in $\boldsymbol{\Sigma}_i$ und $\boldsymbol{D}$, und somit in $\boldsymbol{V}$, $\boldsymbol{R}$ und $\boldsymbol{G}$, gegeben durch

$$\boldsymbol{\vartheta} = (\sigma^2, \tau_1^2, \ldots, \tau_q^2)'\,.$$

Allgemein kann $\boldsymbol{\vartheta}$ weitere unbekannte Parameter enthalten, zum Beispiel Nebendiagonalelemente von $\boldsymbol{D}$.

Für die Kovarianzmatrix $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{U}\boldsymbol{G}\boldsymbol{U}'$ ergibt sich die Blockdiagonal-Matrix

$$\boldsymbol{V} = \mathrm{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_i, \ldots, \boldsymbol{V}_m).$$

Somit erhält man für die Schätzer $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\gamma}}$ die im Kasten angegebenen Ausdrücke.

Beispiel 6.4 Hormontherapie bei Ratten

Wir schätzen zunächst das Modell (6.1), das sowohl individuenspezifische Abweichungen γ_{0i} von der Populationskonstanten β_0 als auch individuenspezifische Steigungsparameter γ_{1i} enthält. Basierend auf den Resultaten von Verbeke & Molenberghs (2000) nehmen wir von vornherein an, dass γ_{0i} und γ_{1i} unkorreliert sind, die Kovarianzmatrix also diagonal ist:

$$\boldsymbol{D} = \begin{pmatrix} \tau_0^2 & 0 \\ 0 & \tau_1^2 \end{pmatrix}.$$

Wir schätzen die fixen Effekte, die Varianzparameter σ^2, τ_0^2, τ_1^2 und auch die zufälligen Effekte. Tabelle 6.2 enthält die Schätzwerte und Standardabweichungen für die fixen Effekte und die Schätzwerte für die Varianzparameter. Da der Schätzwert $\hat{\tau}_1^2$ für $\mathrm{Var}(\gamma_{1i})$ sehr klein ist, schätzen wir auch das vereinfachte Modell (6.4), das keine individuenspezifischen Terme $\gamma_{1i}t_{ij}$ enthält. Die Ergebnisse sind ebenfalls in Tabelle 6.2 zu finden. Es zeigt sich, dass die Schätzungen sehr ähnlich sind.

Schätzung für Longitudinal- und Clusterdaten

Schätzer

Für Longitudinaldaten vereinfachen sich die Schätzer für $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\gamma}}$ zu

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{y}_i$$

$$\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{D}} \boldsymbol{U}_i' \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}), \quad i = 1, \ldots, m.$$

Dabei sind $\hat{\boldsymbol{D}}$ und $\hat{\boldsymbol{V}}_i$ die geschätzten Kovarianzmatrizen nach Einsetzen des REML-Schätzers $\hat{\boldsymbol{\vartheta}}_{REML}$. Dieser ergibt sich durch Maximieren von

$$l_R(\boldsymbol{\vartheta}) = -\frac{1}{2} \sum_{i=1}^{m} \Big\{ \log |\boldsymbol{V}_i(\boldsymbol{\vartheta})| + \log |\boldsymbol{X}_i' \boldsymbol{V}_i(\boldsymbol{\vartheta})^{-1} \boldsymbol{X}_i| + (\boldsymbol{y}_i - \boldsymbol{X}_i' \hat{\boldsymbol{\beta}})' \boldsymbol{V}_i(\boldsymbol{\vartheta})^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}) \Big\}.$$

Kovarianzmatrizen und Tests

Die approximativen Kovarianzmatrizen sind gegeben durch

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{A}_{11} = \left\{ \sum_{i=1}^{m} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right\}^{-1}$$

und eine entsprechende spezielle Form von $\boldsymbol{A}_{22}$ in (6.31) bzw. (6.32). Approximative Varianzen bzw. Standardabweichungen $\hat{\sigma}_j$ der Komponenten von $\hat{\boldsymbol{\beta}}$ sind durch die entsprechenden Diagonalelemente bzw. deren Wurzeln gegeben. Daraus können approximative t-Werte und p-Werte berechnet werden.

Effekt	Parameter	Modell (6.1)		Modell (6.4)		Parametrisch	
		Schätzwert	s.e.	Schätzwert	s.e.	Schätzwert	s.e.
Konstante	β_0	68.607	0.338	68.607	0.331	68.687	0.348
Niedrigdosis	β_1	7.505	0.228	7.507	0.225	7.677	0.286
Hochdosis	β_2	6.874	0.231	6.871	0.228	6.529	0.284
Kontrolle	β_3	7.313	0.285	7.314	0.281	7.212	0.326
$\mathrm{Var}(\gamma_{0i})$	τ_0^2	3.739		3.565			
$\mathrm{Var}(\gamma_{1i})$	τ_1^2	<0.001					
$\mathrm{Var}(\varepsilon_{ij})$	σ^2	1.481		1.445		4.730	

Tabelle 6.2. *Hormontherapie bei Ratten: Likelihood-Schätzung.*

Um den Einfluss des zufälligen Effekts γ_{0i} abschätzen zu können, zeigt Tabelle 6.2 darüberhinaus Schätzungen eines parametrischen Modells ohne zufällige Effekte. Wäh-

Abb. 6.2. *Hormontherapie bei Ratten: Links: Kerndichteschätzer (durchgezogene Linie) und Normalverteilungsapproximation (gestrichelt), rechts: Normal-Quantil-Plot.*

rend sich die Schätzergebnisse für die Konstante kaum unterscheiden, zeigen sich Unterschiede für die Behandlungsgruppen. Die Schätzwerte selbst variieren leicht, insbesondere vergrößern sich aber die zugehörigen Standardfehler. Dies liegt in der deutlich vergrößerten Fehlervarianz σ^2 begründet, die im parametrischen Modell ungefähr dreimal so groß ist im Vergleich zu den Modellen mit zufälligen Effekten. Offenbar wird in diesem Beispiel individuenspezifische Variation in den Fehlerterm „verschoben“, so dass auch die Standardfehler der Schätzer vergrößert werden. Man beachte jedoch, dass in Anwendungen auch das umgekehrte Phänomen auftreten kann: Durch Einbezug zufälliger Effekte werden die Standardfehler im Vergleich zu rein parametrischen Modellen vergrößert.

Für das vereinfachte Modell zeigt Abbildung (6.2) den für die Schätzwerte $\hat{\gamma}_{0i}$, $i = 1, \ldots, 50$, berechneten Kerndichteschätzer und den Normal-Quantil-Plot. Die Abweichungen von der angenommenen Normalverteilung sind nicht gravierend.

△

Beispiel 6.5 Wortschatztest

Tabelle 6.3 enthält Schätzwerte für die fixen Effekte des linearen gemischten Modells aus Beispiel 6.2 sowie die dazugehörigen Standardabweichungen, p-Werte und 95%-Konfidenzintervalle. Als Schätzung für die Varianzen ergab sich $\hat{\sigma}^2 = 34.183$ und $\hat{\tau}^2 = \widehat{\mathrm{Var}}(\gamma_{0i}) = 4.835$. Während Diagnose (D), Medikation (M) und Geschlecht (G) keinen signifikanten Einfluss besitzen, sind Schulbildung (S) und Alter (A) hochsignifikant. Die Bildungsstufen ,,Realschule”, ,,Abitur bzw. Fachabitur”, ,,Gymnasium abgebrochen” und ,,abgeschlossenes Studium” wirken sich verglichen mit der Referenzkategorie ,,Hauptschulabschluss” positiv auf das Ergebnis des Wortschatztests aus. Hierbei fällt auf, dass die Parameterschätzungen für ,,Realschule” und ,,Gymnasium abgebrochen” deutlich unter denen für ,,Abitur bzw. Fachabitur” und ,,abgeschlossenes Studium” liegen. ,,Kein Abschluss” hingegen hat auf die erwartete Testleistung einen negativen Einfluss. Für die verschiedenen Kategorien der Variable Schulbildung wurde zusätzlich getestet, ob zwischen den einzelnen Parameterschätzungen signifikante Unterschiede bestehen. Hierbei konnten keine Abweichungen zwischen ,,Realschulabschluss” und ,,Gymnasium abgebrochen” einerseits und zwischen ,,Abitur bzw. Fachabitur” und ,,abgeschlossenes Studium” andererseits nachgewiesen werden. Für alle anderen möglichen Kombinationen ergaben sich signifikante Unterschiede.

Variable	Effekt	Standard-abweichung	p-Wert	95% Konfidenz-intervall	
Konstante	92.807	1.531	<0.001	89.805	95.808
D1	2.088	1.400	0.134	-0.547	4.823
D2	1.875	1.278	0.142	-0.631	4.381
D3	1.289	0.991	0.194	-0.653	3.231
S1	-4.177	1.970	0.034	-8.039	-0.315
S2	5.087	0.841	<0.001	3.437	6.736
S3	12.340	1.003	<0.001	10.374	14.306
S4	6.601	1.687	<0.001	3.295	9.908
S5	12.025	0.891	<0.001	10.277	13.774
G	-0.401	0.624	0.520	-1.226	0.822
A	0.116	0.024	<0.001	0.069	0.163
C	-0.538	0.351	0.125	-1.226	0.150
M1	-0.372	1.381	0.787	-3.080	2.335
M2	-1.396	1.326	0.293	-3.996	1.204
M3	-0.357	1.188	0.764	-2.686	1.973

Tabelle 6.3. *Wortschatztest: Likelihood-Schätzung.*

Abb. 6.3. *Wortschatztest: Links: Kerndichteschätzer für die zufälligen Effekte (durchgezogene Linie) und Normalverteilungsapproximation (gestrichelt). Rechts: Normal-Quantil-Plot für die zufälligen Effekte.*

Die Normal-Quantil-Plots und Kerndichteschätzungen der Abbildungen 6.3 und 6.4 für die geschätzten zufälligen Effekte und Residuen zeigen keine auffälligen Abweichungen von einer Normalverteilung.

△

6.5 Bayesianische gemischte lineare Modelle

Aus Bayesianischer Sicht werden in einem LMM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

Abb. 6.4. *Wortschatztest: Links: Kerndichteschätzer für die Residuen (durchgezogene Linie) und Normalverteilungsapproximation (gestrichelt). Rechts: Normal-Quantil-Plot für die Residuen.*

nicht nur die zufälligen Effekte $\boldsymbol{\gamma}$, sondern auch die fixen Effekte $\boldsymbol{\beta}$ als Zufallsgrößen angesehen. Sie unterscheiden sich jedoch durch die gewählte Priori-Verteilung. Wenn für $\boldsymbol{\beta}$ in Analogie zum „frequentistischen" Ansatz kein Vorwissen bekannt ist oder benutzt werden soll, setzt man eine *nichtinformative* Priori

$$p(\boldsymbol{\beta}) \propto const$$

an. Vorwissen kann z.B. durch eine Priori-Normalverteilung

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M})$$

formuliert werden, wobei $\boldsymbol{m}$ und $\boldsymbol{M}$ als bekannt vorausgesetzt werden. Die nichtinformative Priori $p(\boldsymbol{\beta}) \propto const$ erhält man daraus als Grenzfall, wenn für die Präzisionsmatrix $\boldsymbol{M}^{-1} \to \mathbf{0}$ gilt. Die Wahl einer Normalverteilung ist auch dadurch begründet, dass die Posteriori-Inferenz wie in linearen Modellen vergleichsweise einfach wird. Die im linearen gemischten Modell in Abschnitt 6.2 angenommene Verteilung

$$\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G})$$

für die zufälligen Effekte $\boldsymbol{\gamma}$ erscheint im Bayesianischen Modell ebenfalls als Priori-Verteilung. Jedoch enthält die Kovarianzmatrix $\boldsymbol{G}$ wie auch die Kovarianzmatrix $\boldsymbol{R} = \mathrm{Cov}(\boldsymbol{\varepsilon})$ im Allgemeinen unbekannte Parameter, die wir wie bisher in einem Vektor $\boldsymbol{\vartheta}$ zusammenfassen und als „Hyperparameter" bezeichnen. Es gilt also genauer

$$\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G}(\boldsymbol{\vartheta})), \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{R}(\boldsymbol{\vartheta})).$$

Zusätzlich nimmt man an, dass $\boldsymbol{\beta}, \boldsymbol{\gamma}$ und $\boldsymbol{\varepsilon}$ a priori unabhängig sind.

In Bezug auf den Hyperparameter $\boldsymbol{\vartheta}$ gibt es zwei Sichtweisen und daraus resultieren zwei Möglichkeiten der Bayes-Inferenz: In einem genuinen, *vollen Bayes-Ansatz* wird auch $\boldsymbol{\vartheta}$ als Zufallsvariable aufgefasst, für die eine weitere (Hyper-) Priori $p(\boldsymbol{\vartheta})$ angenommen wird. Dagegen wird $\boldsymbol{\vartheta}$ im sogenannten *empirischen Bayes-Ansatz* als unbekannter, aber fester Parameter aufgefasst, den man zum Beispiel durch Maximieren der (marginalen) Likelihood schätzt. Der resultierende Schätzer $\hat{\boldsymbol{\vartheta}}$ wird in $\boldsymbol{G}(\boldsymbol{\vartheta})$ bzw. $\boldsymbol{R}(\boldsymbol{\vartheta})$ eingesetzt und die Bayes-Inferenz für $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ wird mit als gegeben aufgefassten Kovarianzmatrizen $\hat{\boldsymbol{G}} = \boldsymbol{G}(\hat{\boldsymbol{\vartheta}})$ und $\hat{\boldsymbol{R}} = \boldsymbol{R}(\hat{\boldsymbol{\vartheta}})$ durchgeführt.

Posteriori-Verteilung bei bekannter Kovarianzstruktur

Wir betrachten zunächst die *Posteriori-Verteilung* von $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ bei *bekannten* bzw. *gegebenen* $\boldsymbol{G}$ und $\boldsymbol{R}$, da diese für beide Ansätze benötigt wird. Für die *gemeinsame* Posteriori-Verteilung gilt nach dem Satz von Bayes

$$\begin{aligned} p(\boldsymbol{\beta},\boldsymbol{\gamma}\,|\,\boldsymbol{y},\boldsymbol{G},\boldsymbol{R}) &\propto p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{\gamma})\,p(\boldsymbol{\beta})\,p(\boldsymbol{\gamma}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})\right) \\ &\quad\cdot\exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right)\cdot\exp\left(-\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}^{-1}\boldsymbol{\gamma}\right). \end{aligned} \tag{6.34}$$

Nach einigen algebraischen Umformungen erhält man mit

$$\boldsymbol{C}=(\boldsymbol{X},\boldsymbol{U}),\quad \boldsymbol{B}=\begin{pmatrix}\boldsymbol{M}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{G}^{-1}\end{pmatrix},\quad \tilde{\boldsymbol{m}}=\begin{pmatrix}\boldsymbol{M}^{-1}\boldsymbol{m}\\ \boldsymbol{0}\end{pmatrix}$$

eine Posteriori-Normalverteilung mit Erwartungswert

$$\begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\gamma}}\end{pmatrix}=\mathrm{E}[(\boldsymbol{\beta}',\boldsymbol{\gamma}')\,|\,\boldsymbol{y},\boldsymbol{G},\boldsymbol{R}]=(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{B})^{-1}(\tilde{\boldsymbol{m}}+\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{y})$$

und Kovarianzmatrix

$$\mathrm{Cov}[(\boldsymbol{\beta}',\boldsymbol{\gamma}')\,|\,\boldsymbol{y},\boldsymbol{G},\boldsymbol{R}]=(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{B})^{-1}.$$

Für eine nichtinformative Priori-Verteilung $p(\boldsymbol{\beta})\propto const$ können wir $\boldsymbol{M}^{-1}=\boldsymbol{0}$ setzen und erhalten

$$\boldsymbol{B}=\begin{pmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{G}^{-1}\end{pmatrix},\quad \tilde{\boldsymbol{m}}=\begin{pmatrix}0\\ 0\end{pmatrix}.$$

Dann ist der Posteriori-Erwartungswert gleich dem BLUP-Schätzer (6.29), und die Posteriori-Kovarianzmatrix ist die „Bayesianische" Kovarianzmatrix in (6.30).

Empirische Bayes-Schätzung

Die eben gezeigte Äquivalenz muss gelten: Für eine nichtinformative Priori-Verteilung $p(\boldsymbol{\beta})\propto const$ (bzw. $\boldsymbol{M}^{-1}=\boldsymbol{0}$ in (6.34)) ist $p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ nach Logarithmieren bis auf den Faktor $-\frac{1}{2}$ identisch mit dem penalisierten KQ-Kriterium $\mathrm{KQ}_{pen}(\boldsymbol{\beta},\boldsymbol{\gamma})$ in (6.23). Daher sind die daraus abgeleiteten BLUP-Schätzer $(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\gamma}})$ in (6.29) identisch mit dem Posteriori-Modus-Schätzer als Maximierer von

$$\log p(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{\gamma})+\log p(\boldsymbol{\gamma})=-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})-\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}^{-1}\boldsymbol{\gamma}. \tag{6.35}$$

Da die Posteriori-Verteilung von $(\boldsymbol{\beta},\boldsymbol{\gamma})$ eine Normalverteilung ist, sind Posteriori-Erwartungswert und Posteriori-Modus und somit die BLUP-Schätzer (6.29) als Punktschätzer identisch.

Den *empirischen Bayes-Schätzer* für $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ erhält man, indem man für $\boldsymbol{G}$ und $\boldsymbol{R}$ geschätzte Kovarianzmatrizen $\hat{\boldsymbol{G}}$ und $\hat{\boldsymbol{R}}$ einsetzt. Damit ist der empirische Bayes-Schätzer

Empirische Bayes-Schätzung

Mit einer nichtinformativen Priori $p(\boldsymbol{\beta}) \propto const$ gilt: Maximieren der marginalen Likelihood $p(\boldsymbol{y} \,|\, \boldsymbol{\vartheta})$ ergibt den REML-Schätzer $\hat{\boldsymbol{\vartheta}}_{REML}$. Die empirischen Bayes-Schätzer für $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ sind mit den Likelihood-Schätzern $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\gamma}}$ aus Abschnitt 6.3 identisch.

bei nichtinformativer Priori für $\boldsymbol{\beta}$ mit dem EBLP $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ in (6.29) und dessen geschätzter Kovarianzmatrix in (6.31) identisch.

Die Bayesianische Betrachtungsweise von LMM ermöglicht auch eine schon in 6.3.2 erwähnte Herleitung des REML-Schätzers $\hat{\boldsymbol{\vartheta}}_{REML}$ als Maximierer der restringierten Log-Likelihood $l_R(\boldsymbol{\vartheta})$ in (6.28). Es lässt sich nämlich zeigen, dass $\hat{\boldsymbol{\vartheta}}_{REML}$ bei Annahme einer nichtinformativen Priori $p(\boldsymbol{\beta}) \propto const$ die marginale Likelihood

$$p(\boldsymbol{y} \,|\, \boldsymbol{\vartheta}) = \int p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta})\, p(\boldsymbol{\gamma})\, d\boldsymbol{\beta}\, d\boldsymbol{\gamma} \tag{6.36}$$

maximiert bzw. dass bis auf additive Konstanten $\log p(\boldsymbol{y} \,|\, \boldsymbol{\vartheta}) = l_R(\boldsymbol{\vartheta})$ gilt.

Volle Bayes-Schätzung

In einem *vollen Bayes-Ansatz* ist auch für die unbekannten Parameter $\boldsymbol{\vartheta}$ in $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\vartheta}), \boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\vartheta})$ eine Priori-Verteilung $p(\boldsymbol{\vartheta})$ zu spezifizieren. Die Inferenz basiert dann auf der Posteriori

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma})\, p(\boldsymbol{\beta})\, p(\boldsymbol{\gamma})\, p(\boldsymbol{\vartheta}), \tag{6.37}$$

wobei $\boldsymbol{\beta}, \boldsymbol{\gamma}$ und $\boldsymbol{\vartheta}$ als unabhängig angenommen werden. Damit $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta} \,|\, \boldsymbol{y})$ eine echte Posteriori-Dichte ist, muss

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma})\, p(\boldsymbol{\beta})\, p(\boldsymbol{\gamma})\, p(\boldsymbol{\vartheta})\, d\boldsymbol{\beta}\, d\boldsymbol{\gamma}\, d\boldsymbol{\vartheta} < \infty \tag{6.38}$$

für den Nenner im Satz von Bayes gelten. Wenn $p(\boldsymbol{\vartheta})$ eine echte, *informative* Priori mit $\int p(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta} = 1$ ist, so existiert auch $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta} \,|\, \boldsymbol{y})$. Für eine *nichtinformative* Priori mit $\int p(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta} = \infty$ ist die Existenz der Posteriori jedoch nicht allgemein gesichert.

Nimmt man eine nichtinformative Priori $p(\boldsymbol{\vartheta}) \propto const$ an und setzt die Existenz der gemeinsamen Posteriori $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta} \,|\, \boldsymbol{y})$ voraus, ergeben sich folgende Beziehungen zum ML-Schätzer $\hat{\boldsymbol{\vartheta}}_{ML}$ und zum REML-Schätzer $\hat{\boldsymbol{\vartheta}}_{REML}$: Wegen

$$p(\boldsymbol{\vartheta} \,|\, \boldsymbol{y}) = p(\boldsymbol{y} \,|\, \boldsymbol{\vartheta}) \frac{p(\boldsymbol{\vartheta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \,|\, \boldsymbol{\vartheta})$$

ist $\hat{\boldsymbol{\vartheta}}_{REML}$ auch der Posteriori-Modus der *marginalen* Posteriori $p(\boldsymbol{\vartheta} \,|\, \boldsymbol{y})$, während $\hat{\boldsymbol{\vartheta}}_{ML}$ die $\boldsymbol{\vartheta}$-Komponente des Posteriori-Modus der gemeinsamen Posteriori $p(\boldsymbol{\beta}, \boldsymbol{\vartheta} \,|\, \boldsymbol{y})$ ist.

Da das Integral in (6.38) im Allgemeinen nicht analytisch zugänglich ist, ist auch die Posteriori $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta} \,|\, \boldsymbol{y})$ nicht in geschlossener Form darstellbar. Die volle Bayes-Inferenz

wird daher mit MCMC-Simulation durchgeführt. Dazu teilt man den gesamten Vektor $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta})$ in Teilvektoren zusammengehöriger Parameter auf. Solange keine spezifischen Kovarianzstrukturen vorliegen, sind dies $\boldsymbol{\beta}, \boldsymbol{\gamma}$ und $\boldsymbol{\vartheta}$. Dazu werden jeweils die *vollständig bedingten Dichten* (full conditionals), gegeben die restlichen Parameter und die Daten $\boldsymbol{y}$, gebildet und daraus werden sequentiell Zufallszahlen gezogen, vergleiche auch Anhang B.5.

Ohne spezifische Annahme für $p(\boldsymbol{\vartheta})$ erhalten wir für die vollständig bedingten Dichten

$$
\begin{aligned}
&p(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}, \boldsymbol{\vartheta}, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta}) p(\boldsymbol{\beta}) \\
&\quad \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \boldsymbol{m}) \right\}, \\
&p(\boldsymbol{\gamma} \,|\, \boldsymbol{\beta}, \boldsymbol{\vartheta}, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta}) p(\boldsymbol{\gamma}) \\
&\quad \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}' \boldsymbol{G}(\vartheta)^{-1} \boldsymbol{\gamma} \right\}, \\
&p(\boldsymbol{\vartheta} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\vartheta}) p(\boldsymbol{\gamma}) p(\boldsymbol{\vartheta}) \propto |\boldsymbol{R}(\vartheta)|^{-1/2} |\boldsymbol{G}(\vartheta)|^{-1/2} \\
&\quad \cdot \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}' \boldsymbol{G}(\boldsymbol{\vartheta})^{-1} \boldsymbol{\gamma} \right\} p(\boldsymbol{\vartheta}).
\end{aligned}
$$

Für die beiden bedingten Dichten $p(\boldsymbol{\beta} \,|\, \cdot)$ und $p(\boldsymbol{\gamma} \,|\, \cdot)$, wobei „·" für die bedingenden Variablen steht, erhält man nach algebraischen Umformungen

$$
p(\boldsymbol{\beta} \,|\, \cdot) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) \right\},
$$

mit

$$
\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\beta}} &= (\boldsymbol{X}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} \boldsymbol{X} + \boldsymbol{M}^{-1})^{-1} (\boldsymbol{M}^{-1} \boldsymbol{m} + \boldsymbol{X}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma})) \\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= (\boldsymbol{X}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} \boldsymbol{X} + \boldsymbol{M}^{-1})^{-1},
\end{aligned}
$$

und

$$
p(\boldsymbol{\gamma} \,|\, \cdot) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{\gamma}})' \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{\gamma}}) \right\}
$$

mit

$$
\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\gamma}} &= (\boldsymbol{U}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} \boldsymbol{U} + \boldsymbol{G}(\boldsymbol{\vartheta})^{-1})^{-1} (\boldsymbol{U}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})) \\
\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} &= (\boldsymbol{U}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} \boldsymbol{U} + \boldsymbol{G}(\boldsymbol{\vartheta})^{-1})^{-1}.
\end{aligned}
$$

Die vollständig bedingten Dichten für $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ sind also Normalverteilungen,

$$
\boldsymbol{\beta} \,|\, \cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \quad \boldsymbol{\gamma} \,|\, \cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}).
$$

Für eine nichtinformative Priori $p(\boldsymbol{\beta}) \propto \mathit{const}$ bzw. $\boldsymbol{M}^{-1} = \boldsymbol{0}$ ist der bedingte Erwartungswert

$$
\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\boldsymbol{X}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{R}(\boldsymbol{\vartheta})^{-1} (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma})
$$

ein gewichteter KQ-Schätzer, angewandt auf $\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma}$, die um die zufälligen Effekte $\boldsymbol{U}\gamma$ bereinigten Daten $\boldsymbol{y}$. Für eine informative Priori $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M})$ ist $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ ein Matrix-gewichtetes Mittel aus KQ-Schätzer und Priori-Erwartungswert. Analog ist $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$ ein Bayes-Schätzer mit Priori $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\vartheta}))$, angewandt auf $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

Die vollständig bedingte Dichte $p(\boldsymbol{\vartheta}\,|\,\cdot)$ lässt sich für speziell strukturierte LMM bei geeignet gewählter Priori $p(\boldsymbol{\vartheta})$ vereinfachen. Wir betrachten dazu Bayesianische LMM für Longitudinaldaten, also

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1,\ldots,m \\ \text{bzw.}\quad y_{ij} &= \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_i + \varepsilon_{ij}, \qquad j = 1,\ldots,n_i \end{aligned}$$

mit unabhängigen, identisch verteilten Fehlern $\boldsymbol{\varepsilon}_{ij}$, $j = 1,\ldots,n_i$. Dann gilt

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\varepsilon}_i) &= \sigma^2\boldsymbol{I}_{n_i}, \quad \boldsymbol{R} = \sigma^2\boldsymbol{I}, \\ \mathrm{Cov}(\boldsymbol{\gamma}_i) &= \boldsymbol{D}, \quad \boldsymbol{G} = \mathrm{diag}(\boldsymbol{D},\ldots,\boldsymbol{D}) \end{aligned}$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_i \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} \qquad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_1 & & & & \boldsymbol{0} \\ & \ddots & & & \\ & & \boldsymbol{U}_i & & \\ & & & \ddots & \\ \boldsymbol{0} & & & & \boldsymbol{U}_m \end{pmatrix},$$

vergleiche Abschnitt 6.2. Für den „fixen" Effekt nehmen wir eine nichtinformative Priori

$$p(\boldsymbol{\beta}) \propto const$$

an.

Der Vektor $\boldsymbol{\vartheta}$ besteht aus der unbekannten Varianz σ^2 und, soweit keine spezielleren Annahmen getroffen werden, aus den Elementen von $\boldsymbol{D}$. Die Wahl der nichtinformativen (aber uneigentlichen) Priori-Verteilungen von Jeffreys

$$p(\sigma^2) \propto \sigma^{-2}, \quad p(\boldsymbol{D}) \propto |\boldsymbol{D}|^{-\frac{r+1}{2}}$$

mit $r = \dim(\boldsymbol{\gamma}_i)$, führt im Allgemeinen zu uneigentlichen (nicht normierbaren) Posteriori-Verteilungen, so dass auch MCMC-Techniken versagen. Deshalb wird für die Varianz eine schwach informative inverse Gammaverteilung $\sigma^2 \sim \mathrm{IG}(a_0, b_0)$ vorgeschlagen, wobei a_0 und b_0 kleine Werte sind, etwas $a_0 = b_0 = 0.001$. Für $\boldsymbol{D}$ wird als mehrdimensionale Form der inversen Gammaverteilung oft eine inverse Wishart-Verteilung verwendet. Wir nehmen zur Vereinfachung an, dass $\boldsymbol{D}$ eine Diagonalmatrix mit Elementen τ_j^2, $j = 1,\ldots,q$, ist. Damit zerfällt die inverse Wishart-Verteilung in das Produkt unabhängiger inverser Gammaverteilungen

$$\tau_j^2 \sim \mathrm{IG}(a_j, b_j), \quad j = 1,\ldots,q.$$

Implizit unterstellen wir also, dass die Komponenten γ_{ij} in $\boldsymbol{\gamma}_i$ a priori unabhängig sind. Die vollständig bedingten Dichten eines LMM für Longitudinaldaten ergeben sich als:

$$\begin{aligned} &\boldsymbol{\beta}\,|\,\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\ &\quad \text{mit } \boldsymbol{\mu}_{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \\ &\boldsymbol{\gamma}_i\,|\,\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_i}), \quad i = 1,\ldots,m \\ &\quad \text{mit } \boldsymbol{\mu}_{\boldsymbol{\gamma}_i} = (\boldsymbol{U}_i'\boldsymbol{U}_i + \sigma^2\boldsymbol{D}^{-1})^{-1}\boldsymbol{U}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_i} = \sigma^2(\boldsymbol{U}_i'\boldsymbol{U}_i + \sigma^2\boldsymbol{D}^{-1})^{-1}, \\ &\sigma^2\,|\,\cdot \sim \mathrm{IG}(\tilde{a}_0, \tilde{b}_0) \\ &\quad \text{mit } \tilde{a}_0 = a_0 + \frac{n}{2}, \quad \tilde{b}_0 = b_0 + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma}) \end{aligned}$$

Variable	**Koeffizient**	**Standard-abweichung**	**2.5%-Quantil**	**97.5%-Quantil**
Konstante	68.560	0.343	67.965	69.259
Niedrigdosis	7.450	0.229	7.061	7.975
Hochdosis	6.873	0.231	6.415	7.314
Kontrolle	7.317	0.285	6.751	7.880

Tabelle 6.4. *Hormon-Therapie bei Ratten: Bayes-Schätzung.*

$$\tau_j^2 \mid \cdot \sim \mathrm{IG}(\tilde{a}_j, \tilde{b}_j),\ j = 1, \ldots, q,$$
$$\text{mit } \tilde{a}_j = a_j + \frac{m}{2}, \quad \tilde{b}_j = b_j + \frac{1}{2}\sum_{i=1}^{m} \gamma_{ij}^2 . \tilde{b}_j = b_j + \frac{1}{2}\sum_{i=1}^{m} \gamma_{ij}^2 .$$

Da alle vollständig bedingten Dichten eine bekannte Form besitzen, ergibt sich der im Kasten zusammengefasste Algorithmus.

Gibbs-Sampler für Longitudinaldaten-LMM

Initialisierung: Gebe Startwerte $\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}_i^{(0)}, \quad i = 1, \ldots, m$, und $\sigma^{2(0)}, D^{(0)}$ vor.
Ziehe für $t = 1, 2, \ldots$ Zufallszahlen

1. $\boldsymbol{\beta}^{(t)}$ aus $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ mit $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t-1)}, \sigma^2 = \sigma^{2(t-1)}$
2. $\boldsymbol{\gamma}_i^{(t)}$ aus $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_i})$ mit $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}, \sigma^2 = \sigma^{2(t-1)}, \boldsymbol{D} = \boldsymbol{D}^{(t-1)}$
3. $(\sigma^2)^{(t)}$ aus $\mathrm{IG}(\tilde{a}_0, \tilde{b}_0)$ mit $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}$
4. $(\tau_j^2)^{(t)}$ aus $\mathrm{IG}(\tilde{a}_j, \tilde{b}_j)$ mit $\gamma_{ij} = \gamma_{ij}^{(t)}$.

Beispiel 6.6 Hormontherapie bei Ratten

Wir reanalysieren das in Beispiel 6.1 angegebene, vereinfachte Modell (6.4), das nur eine zufällige Konstante γ_{0i} enthält. Als Prioris wählen wir für σ^2 und $\tau_0^2 = \mathrm{Var}(\gamma_{0i})$ jeweils inverse Gammaverteilungen mit Hyperparametern $a_0 = a = b_0 = b = 0.001$. Tabelle 6.4 zeigt die der Tabelle 6.2 entsprechenden Schätzwerte und zusätzlich die 2.5 %- und 97.5 %-Quantile. Man erkennt, dass die Schätzer sehr gut mit den ML-Schätzern übereinstimmen.

△

Beispiel 6.7 Wortschatztest

Wir reanalysieren das in Beispiel 6.2 angegebene Modell mit einer Bayes-Schätzung. Dabei werden für σ^2 und τ_0^2 ebenfalls Gamma-Priori-Verteilungen mit den gleichen Hyperparametern wie in Beispiel 6.6 gewählt. Tabelle 6.5 enthält die dazugehörigen Ergebnisse. Die Punktschätzer stimmen sehr gut mit den ML-Schätzern überein, obwohl hier der Posteriori-Erwartungswert anstelle des Posteriori-Modus angegeben ist. Die Vertrauensintervalle sind bei der Bayes-Inferenz tendenziell etwas breiter, da die Unsicherheit der Punktschätzung mit berücksichtigt wird. Das führt dazu, dass die Kategorie

Variable	Koeffizient	Standard-abweichung	2.5%-Quantil	97.5%-Quantil
Konstante	92.628	1.621	89.596	95.856
D1	2.099	1.375	-0.664	4.667
D2	1.860	1.282	-0.719	4.344
D3	1.318	0.989	-0.541	3.296
S1	-4.009	1.980	-7.969	0.040
S2	5.172	0.887	3.523	6.899
S3	12.489	1.046	10.368	14.658
S4	6.613	1.644	3.348	9.852
S5	12.154	0.932	10.342	13.949
G	-0.354	0.654	-1.669	0.958
A	0.116	0.025	0.068	0.164
C	-0.518	0.352	-1.202	0.136
M1	-0.369	1.366	-3.133	2.329
M2	-1.432	1.328	-3.949	1.164
M3	-0.436	1.189	-2.889	1.934

Tabelle 6.5. *Wortschatztest: Bayes-Schätzung.*

S1 (kein Abschluss) an der Grenze zur Nichtsignifikanz im Vergleich zur Referenzkategorie (Hauptschulabschluss) liegt. Zu berücksichtigen ist bei der Interpretation, dass die Fallzahl für die Kategorie S1 vergleichsweise gering ist.

△

6.6 Generalisierte lineare gemischte Modelle

Wie bei linearen Modellen kann es aus den gleichen Gründen sinnvoll sein auch bei generalisierten linearen Modellen (GLM) zufällige Effekte im Prädiktor zu berücksichtigen. Damit erhält man generalisierte lineare gemischte Modelle (GLMM, generalized linear mixed models) als Erweiterung für Zielvariablen, die nicht normalverteilt sind, sondern zum Beispiel binär, diskret oder schief verteilt sein können. Ähnlich wie bei der Erweiterung von LM zu GLM wird dabei die Inferenz technisch komplizierter.

6.6.1 Definition und Eigenschaften von GLMM

In Abschnitt 4.4.1 wurde ein GLM folgendermaßen definiert: Die (bedingte) Dichte von y_i, gegeben der lineare Prädiktor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, gehört zur Klasse der Exponentialfamilien (z.B. Binomial-, Poisson- oder Gammaverteilung) und der lineare Prädiktor η_i ist mit dem (bedingten) Erwartungswert $\mu_i = \mathrm{E}(y_i)$ durch die Beziehung

$$\mu_i = h(\eta_i) \quad \text{bzw.} \quad \eta_i = g(\mu_i)$$

verknüpft. Für die Definition von GLMM erweitern wir in Analogie zum Abschnitt 6.2 den Prädiktorvektor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ eines GLM durch Hinzunahme zufälliger Effekte.

GLMM für Longitudinal- und Clusterdaten

Die Zielvariablen y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, mit n_i Messwiederholungen pro Individuum oder Cluster i, können nun zum Beispiel binäre Variablen oder Zählvariablen sein. Mit individuen- bzw. clusterspezifischen zufälligen Effekten $\boldsymbol{\gamma}_i$ ist dann $y_{ij} \,|\, \boldsymbol{\gamma}_i$ zum Beispiel binär, Binomial- oder Poisson-verteilt. Der bedingte Erwartungswert $\mu_{ij} = \mathrm{E}(y_{ij} \,|\, \boldsymbol{\gamma}_i)$ ist mit einem linearen (gemischten) Prädiktor der gleichen Form wie in (6.5) für LMM,

$$\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_i\,, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n_i,$$

durch $\mu_{ij} = h(\eta_{ij})$ mit einer geeigneten Responsefunktion h verbunden.
Für binäre Zielvariablen ergeben sich somit Logit-Modelle mit zufälligen Effekten durch

$$\log \frac{\mathrm{P}(y_{ij} = 1 \,|\, \boldsymbol{\gamma}_i)}{\mathrm{P}(y_{ij} = 0 \,|\, \boldsymbol{\gamma}_i)} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_i \tag{6.39}$$

und für Zählvariablen $y_{ij} \sim \mathrm{Po}(\lambda_{ij}|\boldsymbol{\gamma}_i)$ log-lineare Modelle durch

$$\log(\lambda_{ij}) = x_{ij}'\boldsymbol{\beta} + u_{ij}'\boldsymbol{\gamma}_i.$$

Wie in LMM besteht $\boldsymbol{u}_{ij}$ oft aus Komponenten von $\boldsymbol{x}_{ij}$. Speziell für $\boldsymbol{u}_{ij} \equiv 1$ erhält man den Prädiktor

$$\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \gamma_{0i} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_k x_{ijk} + \gamma_{0i}\,, \tag{6.40}$$

bei dem γ_{0i} von Kovariablen unabhängige, zufällige Abweichungen von der Populationskonstanten β_0 repräsentieren.

GLMM in allgemeiner Form

Wie für LMM kann man durch geeignete Wahl der Designmatrizen $\boldsymbol{X}$ und $\boldsymbol{U}$ den Prädiktorvektor $\boldsymbol{\eta} = (\eta_{11}, \ldots, \eta_{1n_1}, \eta_{21}, \ldots, \eta_{m1}, \ldots, \eta_{mn_m})'$ in der Form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma}$$

schreiben, vergleiche Abschnitt 6.3. Ebenso wie für LMM nimmt man an, dass $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m$ unabhängig und identisch

$$\boldsymbol{\gamma}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$$

verteilt sind und somit $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G})$ mit $\boldsymbol{G} = \mathrm{diag}(\boldsymbol{D}, \ldots, \boldsymbol{D})$ gilt.
Für die folgende *allgemeine Definition* bezeichnen wir nun mit y_i die i-te (skalare) Komponente des gesamten Vektors $\boldsymbol{y} = (y_1, \ldots, y_i, \ldots, y_n)'$ der Zielvariablen und mit η_i die entsprechende Komponente

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma}$$

von $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma}$, d.h. $\boldsymbol{x}_i'$ bzw. $\boldsymbol{u}_i'$ sind die i-te Zeile der geeignet spezifizierten Designmatrizen $\boldsymbol{X}$ bzw. $\boldsymbol{U}$. Man beachte: Für wiederholte Beobachtungen wie bei Longitudinaldaten entfällt in dieser Durchnummerierung aller individuellen Beobachtungen die Doppelindizierung bei $\eta_{11}, \ldots, \eta_{1n_1}$ etc., d.h. der Index i ist die Nummer der i-ten einzelnen Beobachtung bei einfacher Durchnummerierung sämtlicher Beobachtungen.

Generalisierte lineare gemischte Modelle (GLMM)

1. Verteilungsannahme:
   Gegeben die zufälligen Effekte $\boldsymbol{\gamma}$ und die Kovariablen $\boldsymbol{x}_i$, sind die Zielvariablen y_i bedingt unabhängig und die bedingte Dichte $f(y_i \,|\, \boldsymbol{\gamma})$ gehört zu einer Exponentialfamilie wie in Abschnitt 4.4.1 für GLM.
2. Strukturannahme:
   Der bedingte Erwartungswert $\mu_i = \mathrm{E}(y_i \,|\, \boldsymbol{\gamma})$ ist mit dem linearen Prädiktor
   $$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma}$$
   durch
   $$\mu_i = h(\eta_i) \quad \text{bzw.} \quad \eta_i = g(\mu_i)$$
   verknüpft, wobei h die Responsefunktion und $g = h^{-1}$ die Linkfunktion ist.
3. Verteilungsannahme für die zufälligen Effekte $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$:
   $$\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}),$$
   mit positiv definiter Kovarianzmatrix $\boldsymbol{G}$.

Bemerkungen:

1. Ähnlich wie im LMM sind für binäre gemischte Modelle mit rein additiven zufälligen Effekten $\boldsymbol{\gamma}_i$ wie in (6.40) Messwiederholungen ($n_i > 1$) notwendig. Für gemischte Poisson-Modelle ist dies jedoch nicht notwendig. So ist es möglich, in einem log-linearen Poissonmodell die (zufällige) Rate auch ohne Messwiederholungen zu spezifizieren, wenn also
   $$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma}_i)$$
   oder speziell
   $$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \gamma_{0i}), \tag{6.41}$$
   mit $\gamma_{0i} \sim \mathrm{N}(0, \tau_0^2)$, gilt. Ein log-lineares Poisson-Modell mit dem Prädiktor (6.40) heißt auch *Poisson-Normal- oder Poisson-Lognormal-Modell*, da (6.41) äquivalent ist zu
   $$\lambda_i = \nu_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \tag{6.42}$$
   wobei $\nu_i = \exp(\gamma_{0i})$ dann Lognormalverteilt ist. Poisson-(Log) Normal-Modelle eignen sich dazu, Zähldaten mit Überdispersion zu analysieren, siehe dazu auch die Bemerkung 3 und Beispiel 6.8.
2. Die Forderung der bedingten Unabhängigkeit der $\boldsymbol{y}_i$, gegeben $\boldsymbol{\gamma}$, entspricht der Forderung nach unabhängigen Messfehlern $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ in LMM, d.h. $\boldsymbol{R} = \sigma^2\boldsymbol{I}$. Man kann diese Forderung prinzipiell abschwächen, indem man zum Beispiel bei binären Zielvariablen $\boldsymbol{y}_i$ für Longitudinaldaten geeignete Abhängigkeitsstrukturen, z.B. durch „odds-ratios", definiert. Diese sind jedoch deutlich komplizierter als im Fall von Normalverteilungen, so dass man üblicherweise die Abhängigkeit über zufällige Effekte

im Prädiktor modelliert. Die 4. Bemerkung zeigt, dass selbst im einfachsten Poisson-Normal-Modell (6.41) die Zielvariablen y_i und y_j, $i \neq j$, *marginal korreliert* sind.

3. *Marginale und konditionale Verteilungen:*
Die marginale Dichte $f(y_i)$ kann prinzipiell aus der (durch die spezifische Exponentialfamilie bestimmten) konditionalen Dichte $f(y_i \,|\, \boldsymbol{\gamma})$ durch Integration

$$f(y_i) = \int f(y_i \,|\, \boldsymbol{\gamma}) f(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}$$

bestimmt werden. Dabei ist $f(\boldsymbol{\gamma})$ die Dichte des Vektors $\boldsymbol{\gamma}$ der zufälligen Effekte, hier also die Dichte einer $\mathrm{N}(\mathbf{0}, \boldsymbol{G})$-Verteilung. Bis auf den Fall, dass auch die bedingte Dichte $f(y_i \,|\, \boldsymbol{\gamma})$ eine Normalverteilung ist und somit ein LMM vorliegt, lässt sich die Integration, von sehr speziellen Modellen abgesehen, nicht analytisch durchführen. Dies ist auch ein wesentlicher Grund dafür, dass die Inferenz für GLMM schwieriger ist als für LMM. Auch marginale Erwartungswerte, Varianzen und Kovarianzen lassen sich nur in Spezialfällen analytisch bestimmen. Es gilt

$$\mathrm{E}(y_i) = \mathrm{E}(\mathrm{E}(y_i \,|\, \boldsymbol{\gamma})) = \mathrm{E}(\mu_i) = \mathrm{E}(h(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma}))\,,$$

wobei der Erwartungswert bezüglich $\boldsymbol{\gamma}$ zu bilden ist. Da die Funktion h nichtlinear ist, kann dieser Erwartungswert im Allgemeinen nicht analytisch bestimmt werden.

Für die marginale Varianz gilt allgemein

$$\begin{aligned}\mathrm{Var}(y_i) &= \mathrm{Var}(\mathrm{E}(y_i \,|\, \boldsymbol{\gamma})) + \mathrm{E}(\mathrm{Var}(y_i \,|\, \boldsymbol{\gamma})) \\ &= \mathrm{Var}(\mu_i) + \mathrm{E}(\phi\, v(\mu_i)) \\ &= \mathrm{Var}(h(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma})) + \phi\, \mathrm{E}(v(h(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma})))\,,\end{aligned}$$

wobei ϕ der Skalenparameter der Exponentialfamilie ist (mit $\phi = 1$ für binäre und Poisson-Modelle) und $v(\mu_i)$ die Varianzfunktion, mit $v(\mu_i) = v(\pi_i) = \pi_i(1 - \pi_i)$ für binäre Modelle und $v(\mu_i) = \lambda_i$ für Poisson-Modelle. Somit lässt sich auch die marginale Varianz ohne spezielle Annahmen über h und die bedingte Verteilung von y_i gegeben $\boldsymbol{\gamma}$ nicht analytisch angeben.

Für die bedingte Kovarianz gilt

$$\mathrm{Cov}(y_i, y_j \,|\, \boldsymbol{\gamma}) = 0, \quad i \neq j,$$

da y_i und y_j als bedingt unabhängig angenommen werden. Damit folgt für die marginale Kovarianz

$$\begin{aligned}\mathrm{Cov}(y_i, y_j) &= \mathrm{Cov}(\mathrm{E}(y_i \,|\, \boldsymbol{\gamma})\,,\, \mathrm{E}(y_j \,|\, \boldsymbol{\gamma})) + \mathrm{E}(\mathrm{Cov}(y_i, y_j \,|\, \boldsymbol{\gamma})) \\ &= \mathrm{Cov}(h(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma})\,,\, h(\boldsymbol{x}_j'\boldsymbol{\beta} + \boldsymbol{u}_j'\boldsymbol{\gamma}))\,.\end{aligned}$$

4. Für das *log-lineare Poisson-Normal-Modell* (6.41) lassen sich die marginalen Erwartungswerte, Varianzen und Kovarianzen analytisch bestimmen. Dies liegt daran, dass

$$\mu_i \,|\, \gamma_{0i} = \nu_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

gilt und ν_i lognormal verteilt ist. Deshalb gilt

$$\mathrm{E}(y_i) = \mathrm{E}(\nu_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) = \exp(\tau_0^2/2)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\tau_0^2/2 + \boldsymbol{x}_i'\boldsymbol{\beta})\,.$$

Von der additiven Konstanten abgesehen, sind also die Populationseffekte $\boldsymbol{\beta}$ im marginalen und konditionalen Modell gleich. Weiterhin lässt sich

$$\mathrm{Var}(y_i) = \mathrm{E}(y_i)\{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})((\exp(3\tau_0^2/2) - \exp(\tau_0^2/2)) + 1)\}$$

zeigen. Da der zweite Faktor in der geschweiften Klammer größer als 1 ist, ist die marginale Varianz größer als der marginale Erwartungswert. Somit kann die marginale Verteilung keine Poisson-Verteilung sein, sondern ist eine Verteilung mit Überdispersion, für die jedoch keine analytische Formel angegeben werden kann.

Nach einiger Rechnung lässt sich auch

$$\mathrm{Cov}(y_i, y_j) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{x}_j'\boldsymbol{\beta})\{\exp(\tau_0^2)(\exp(\tau_0^2) - 1)\}$$

zeigen. Damit sind y_i und y_j marginal positiv korreliert, jedoch mit einer sehr speziellen Kovarianzstruktur.

5. Während im log-linearen Poisson-Normal-Modell die fixen Kovariableneffekte, von der Konstanten abgesehen, im konditionalen und marginalen Modell identisch sind, gilt dies allgemein nicht. Für ein *Probit-Normal-Modell*

$$\mathrm{E}(y_i \,|\, \gamma_{0i}) = \mathrm{P}(y_i = 1 \,|\, \gamma_{0i}) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta} + \gamma_{0i}), \quad \gamma_{0i} \sim \mathrm{N}(0, \sigma_\gamma^2),$$

lässt sich

$$\mathrm{P}(y_i = 1) = \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sqrt{\tau_0^2 + 1}}\right) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}^*)$$

mit $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\sqrt{\tau_0^2 + 1}$ zeigen, d.h. die Absolutwerte der Populationseffekte werden im marginalen Modell um den Faktor $1/\sqrt{\tau_0^2 + 1}$ verkleinert. Für ein analoges Logit-Normal-Modell lässt sich nicht einmal eine exakte Formel für die Linkfunktion des marginalen Modells angeben, sondern nur eine Approximation durch Taylorentwicklung, siehe z.B. Fahrmeir & Tutz (2001, Kap. 7.7).

Beispiel 6.8 Anzahl der Zitate von Patenten

In Beispiel 4.7 (Seite 214) wurden log-lineare Poisson-Modelle mit und ohne Überdispersionsparameter für Regressionsanalysen mit der Zielvariable *azit* (Anzahl der Zitate) eingesetzt. Es zeigte sich, dass deutliche Überdispersion vorliegt. Log-lineare Poisson-Normal-Modelle der Form (6.41) mit einem additiven zufälligen Effekt $\gamma_{0i} \sim \mathrm{N}\left(0, \tau_0^2\right)$ sind eine alternative Möglichkeit vorhandene Überdispersion zu modellieren. Zur Illustration beschränken wir uns auf das einfache log-lineare Modell mit rein linearen Effekten und erweitern es zum log-linearen Poisson-Normal-Modell

$$\begin{aligned}\log(\lambda_i) = \eta_i = \beta_0 &+ \beta_1 jahrc_i + \beta_2 alandc_i + \beta_3 anspc_i + \beta_4 biopharm_i + \\ &+ \beta_5 uszw_i + \beta_6 patus_i + \beta_7 patdsg_i + \beta_8 einspruch_i + \gamma_{0i}\,.\end{aligned}$$

△

Kategoriale gemischte Regressionsmodelle

Gemischte Modelle für kategoriale Zielvariablen können durch geeignete Erweiterungen der Prädiktoren von kategorialen Regressionsmodellen (Kapitel 5) definiert werden.

Multinomiale gemischte Logit-Modelle für eine Zielvariable Y mit ungeordneten Kategorien $1, \ldots, c$ und globalen Kovariablen, aber kategorienspezifischen fixen und zufälligen Effekten $\boldsymbol{\gamma}_r$ sind durch

$$\mathrm{P}(Y_i = r \,|\, \boldsymbol{\gamma}_r) = \frac{\exp(\eta_{ir})}{1 + \sum_{s=1}^{q} \exp(\eta_{is})} \qquad r = 1, \ldots, q, \quad q = c - 1,$$

mit kategorienspezifischen Prädiktoren

$$\eta_{ir} = \boldsymbol{x}_i' \boldsymbol{\beta}_r + \boldsymbol{u}_i' \boldsymbol{\gamma}_r$$

definiert. Die zufälligen Effekte $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ werden im einfachsten Fall als unabhängig und normalverteilt, $\boldsymbol{\gamma}_r \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G}_r)$, vorausgesetzt. Allgemeiner kann man auch

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \ldots \boldsymbol{\gamma}_q') \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G})$$

mit nicht-diagonaler Kovarianzmatrix $\boldsymbol{G}$ zulassen.

Bei Modellen für geordnete Kategorien ist darauf zu achten, dass notwendige Ordnungsrestriktionen für Schwellenwerte in kumulativen Modellen einzuhalten sind. Eine Erweiterung von kumulativen Logit-Modellen auf gemischte Prädiktoren der Form

$$\eta_{ir} = \theta_r + \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{u}_i' \boldsymbol{\gamma}\,,$$

mit $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G})$ ist unproblematisch, solange die Ordnungsrestriktion

$$-\infty = \theta_0 < \theta_1 < \ldots < \theta_q < \theta_c = \infty$$

für fixe Schwellenwerte $\theta_1, \ldots, \theta_q$ erfüllt ist. Problematischer wird die Erweiterung, wenn die Schwellenwerte als zufällig modelliert werden, also

$$\eta_{ir} = \theta_{ir} + \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{u}_i' \boldsymbol{\gamma}$$

angenommen wird. Die übliche Normalverteilungsannahme

$$\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iq})' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G}_\theta)$$

verletzt die Ordnungsrestriktion, so dass Reparametrisierungen notwendig werden, siehe Fahrmeir & Tutz (2001, Kap. 7.2). Wie für LMM und GLMM bilden kategoriale gemischte Regressionsmodelle für Longitudinaldaten einen wichtigen Spezialfall. Für ordinale Zielvariablen Y_{ij}, $i = 1, \ldots, m$, $j = 1 \ldots, n_i$, nimmt man für den Prädiktor eines kumulativen gemischten Logit-Modells zum Beispiel

$$\eta_{ij} = \theta_r + \boldsymbol{x}_{ij}' \boldsymbol{\beta} + \boldsymbol{u}_{ij}' \boldsymbol{\gamma}_i$$

mit unabhängigen individuenspezifischen zufälligen Effekten $\boldsymbol{\gamma}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$, $i = 1, \ldots, m$, an. In analoger Weise lassen sich Longitudinaldaten-Modelle für ungeordnete Zielvariablen formulieren.

6.7 Likelihood- und Bayes-Inferenz in GLMM

Konzeptionell beruht die Inferenz in GLMM auf den gleichen Likelihood- oder Bayes-Ansätzen wie für LMM. Da jedoch die bedingte Likelihood im Allgemeinen nicht normalverteilt ist und zudem die Verknüpfung zwischen Prädiktor und Erwartungswert für GLMM nichtlinear ist, sind die entscheidenden Teile dieser Konzepte nicht mehr analytisch durchführbar, sondern müssen durch numerische Verfahren oder geeignete Approximationen ersetzt werden. Wir konzentrieren uns dabei auf einen (approximativen) penalisierten Likelihood-Ansatz und auf MCMC-Techniken für die volle Bayes-Inferenz, die sich beide im Spezialfall von LMM auf Methoden der Abschnitte 6.3 bis 6.5 reduzieren. Da der penalisierte Likelihood-Schätzer auch als empirischer Bayes-Schätzer angesehen werden kann, ergeben sich enge Beziehungen zwischen beiden Konzepten. Zur Vereinfachung beschränken wir uns notationell auf Modelle mit Skalenparameter $\phi = 1$. Der allgemeinere Fall, etwa bei der Gammaverteilung, kann wie in GLM behandelt werden.

6.7.1 Penalisierte Likelihood- und empirische Bayes-Schätzung

Im Likelihood-Ansatz werden die fixen Effekte $\boldsymbol{\beta}$ und die Parameter $\boldsymbol{\vartheta}$ in $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\vartheta})$ als feste, unbekannte Größen betrachtet. Die der Log-Likelihood $l(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ in (6.26) entsprechende *marginale* Likelihood $L(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ ergibt sich durch Integration der *bedingten Likelihood*

$$L(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = \int p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma})\, p(\boldsymbol{\gamma} \,|\, \boldsymbol{\vartheta})\, d\boldsymbol{\gamma}. \tag{6.43}$$

Dabei ist wegen der bedingten Unabhängigkeit der y_i

$$p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} p(y_i \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}),$$

und $p(y_i \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma})$ ist die Dichte der gewählten Exponentialfamilie, also zum Beispiel eine Binomial- oder Poisson-Wahrscheinlichkeitsfunktion bzw. -dichte, mit (bedingtem) Erwartungswert $\mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma})$. Wegen $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\vartheta}))$ ist $p(\boldsymbol{\gamma} \,|\, \boldsymbol{\vartheta})$ eine (möglicherweise hochdimensionale) multivariate Normalverteilung.

Da das Integral in (6.43) analytisch nur in Spezialfällen lösbar ist, kann man versuchen, *numerische Integrationsmethoden*, zum Beispiel die Gauß-Hermite-Integration, oder *Monte-Carlo-Methoden* einzusetzen, siehe z.B. Fahrmeir & Tutz (2001, Kap. 7.3). Da diese Integrationstechniken nur für vergleichsweise niedrige Dimensionen durchführbar sind, ist die sogenannte *Laplace-Approximation* des Integrals die von uns bevorzugte Alternative. Sie führt auf ein penalisiertes Likelihood-Kriterium hin, das eine Verallgemeinerung des penalisierten KQ-Kriteriums (6.23) darstellt und sich wie bei LMM von einem empirischen Bayes-Ansatz ausgehend herleiten lässt. Dazu betrachten wir wie in Abschnitt 6.3.1 für zunächst festgehaltenes $\boldsymbol{\vartheta}$ bzw. $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\vartheta})$ die gemeinsame Likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta}) = p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma})\, p(\boldsymbol{\gamma} \,|\, \boldsymbol{\vartheta})$$

und maximieren sie simultan bezüglich $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$. Wegen

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma})\, p(\boldsymbol{\gamma} \,|\, \boldsymbol{\vartheta})$$

ist dies äquivalent zur Bestimmung des *Posteriori-Modus-Schätzers* $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ bzw. nach Einsetzen eines Schätzers $\hat{\boldsymbol{\vartheta}}$ für $\boldsymbol{\vartheta}$ zu einem *empirischen Bayes-Schätzer*. Logarithmieren von $L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta})$ liefert die *penalisierte Log-Likelihood*

$$l_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}(\boldsymbol{\vartheta})^{-1}\boldsymbol{\gamma}. \tag{6.44}$$

Dabei ist die Log-Likelihood $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ wie für ein GLM definiert, nur ist der Prädiktor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ durch den erweiterten Prädiktor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma}$ zu ersetzen. Für ein loglineares Poisson-Modell ist also zum Beispiel

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n}(y_i \log(\lambda_i) - \lambda_i) = \sum_{i=1}^{n}(y_i\eta_i - \exp(\eta_i))$$

mit $\log(\lambda_i) = \eta_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma})$. Für ein LMM ergibt sich speziell

$$l_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta}) = -\frac{1}{2}KQ_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

d.h. das penalisierte Log-Likelihood-Kriterium (6.44) stellt eine Verallgemeinerung des penalisierten KQ-Kriteriums (6.23) für nicht normalverteilte Zielvariablen dar. Bis auf additive, nur von $\boldsymbol{\vartheta}$ abhängige Konstanten ist $l_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta})$ auch die Laplace-Approximation für (6.44).

Zur Maximierung von $l_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta})$ bildet man in Analogie zu GLM die Score-Funktionen bezüglich $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ und erhält in Matrixnotation

$$\begin{aligned} s_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \frac{\partial\, l_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta})}{\partial\, \boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \\ s_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \frac{\partial\, l_{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta})}{\partial\, \boldsymbol{\gamma}} = \boldsymbol{U}'\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{G}(\boldsymbol{\vartheta})^{-1}\boldsymbol{\gamma}. \end{aligned}$$

Dabei hängen $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y}\,|\,\boldsymbol{\gamma})$, $\boldsymbol{D} = \mathrm{diag}\left(\frac{\partial h(\eta_i)}{\partial \eta}\right)$ und $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$ über den Prädiktor von $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ (und eventuell vom Skalenparameter ϕ) ab.

Für die entsprechend partitionierte Fisher-Informationsmatrix erhält man

$$F(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\boldsymbol{U} \\ \boldsymbol{U}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{U}'\boldsymbol{W}\boldsymbol{U} + \boldsymbol{G}(\boldsymbol{\vartheta})^{-1} \end{pmatrix},$$

wobei $\boldsymbol{W} = \boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}$ von $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ abhängt.

Der penalisierte ML- bzw. Posteriori-Modus-Schätzer wird simultan durch die Schätzgleichungen

$$s_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}) = \boldsymbol{0}, \qquad s_{\boldsymbol{\gamma}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}) = \boldsymbol{0}$$

bestimmt. Die numerische Berechnung erfolgt iterativ unter Verwendung des Fisher-Scoring-Verfahrens, das sich mit Hilfe der *Pseudo-Beobachtungen*

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{U}\tilde{\boldsymbol{\gamma}} + \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{6.45}$$

auch in der Form einer iterativen gewichteten penalisierten KQ-Schätzung

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}^{(k)}\boldsymbol{U} \\ \boldsymbol{U}'\boldsymbol{W}^{(k)}\boldsymbol{X} & \boldsymbol{U}'\boldsymbol{W}^{(k)}\boldsymbol{U} + \boldsymbol{G}^{-1}(\boldsymbol{\vartheta}) \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}}^{(k+1)} \\ \tilde{\boldsymbol{\gamma}}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{W}^{(k)}\tilde{\boldsymbol{y}}^{(k)} \\ \boldsymbol{U}'\boldsymbol{W}^{(k)}\tilde{\boldsymbol{y}}^{(k)} \end{pmatrix} \tag{6.46}$$

$k = 1, 2, \ldots$, schreiben lässt. Dabei bedeutet die Notation „$\cdot^{(k)}$“, dass $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\gamma}}$, $\boldsymbol{\mu}$, $\boldsymbol{D}$ und $\boldsymbol{W}$ für die aktuellen Iterationswerte $\tilde{\boldsymbol{\beta}}^{(k)}$, $\tilde{\boldsymbol{\gamma}}^{(k)}$ ausgewertet werden. Im Falle eines unbekannten Skalenparameters ϕ hängt $\boldsymbol{W}$ auch von diesem ab. Jeder Iterationsschritt erfordert also die Lösung des Gleichungssystems (6.46), das genau dem Gleichungssystem (6.24) für LMM entspricht, wenn dort $\boldsymbol{R}^{-1} = \boldsymbol{W}^{(k)}$ gesetzt wird.

Die *Schätzung der Varianzparameter* $\boldsymbol{\vartheta}$ (und gegebenfalls des Skalenparameters ϕ) kann ebenfalls mit Hilfe einer approximativen marginalen Log-Likelihood durchgeführt werden. Es lässt sich mit der Laplace-Approximation zeigen (Breslow & Clayton 1993), dass für die bedingte Log-Likelihood die Normalverteilungsapproximation

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta}) = \log\, p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}) \approx (\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

gilt, wobei $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ wieder von $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ abhängen. Aus der Definition (6.45) der Pseudobeobachtungen $\tilde{\boldsymbol{y}}$ ergibt sich

$$\boldsymbol{y} - \boldsymbol{\mu} = \boldsymbol{D}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}}).$$

Somit folgt

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\vartheta}) \approx (\tilde{\boldsymbol{y}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}})'\boldsymbol{W}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}}).$$

Damit kann die bedingte Log-Likelihood eines GLMM für festes $\boldsymbol{\vartheta}$ durch die Log-Likelihood eines LMM

$$\tilde{\boldsymbol{y}} \,|\, \boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma}, \boldsymbol{W}^{-1})$$

approximiert werden. Mit

$$\boldsymbol{V} = \boldsymbol{W} + \boldsymbol{U}'\boldsymbol{G}(\boldsymbol{\vartheta})^{-1}\boldsymbol{U}$$

erhält man als (approximative) *marginale restringierte Log-Likelihood*

$$l_R(\boldsymbol{\vartheta}) = -\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\vartheta})| - \frac{1}{2}\log(|\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{X}|) - \frac{1}{2}(\tilde{\boldsymbol{y}}(\boldsymbol{\vartheta}) - \boldsymbol{X}\tilde{\boldsymbol{\beta}})'\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}(\tilde{\boldsymbol{y}}(\boldsymbol{\vartheta}) - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).$$

Dabei hängen $\boldsymbol{V}(\boldsymbol{\vartheta})$ und $\tilde{\boldsymbol{y}}(\boldsymbol{\vartheta})$ auch noch von $\tilde{\boldsymbol{\beta}}$ und $\tilde{\boldsymbol{\gamma}}$ ab.

Dies legt den im Kasten beschriebenen Schaukel-Algorithmus zwischen der Schätzung von $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ und der Schätzung von $\boldsymbol{\vartheta}$ nahe.

Als geschätzte Kovarianzmatrix verwendet man die für $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\vartheta}}$ ausgewertete Inverse der Fisher-Informationsmatrix $\hat{\boldsymbol{F}} = \boldsymbol{F}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\vartheta}})$,

$$\widehat{\mathrm{Cov}}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} \approx \hat{\boldsymbol{F}}^{-1}.$$

Unter der Annahme einer approximativen Normalverteilung

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, \hat{\boldsymbol{F}}^{-1}\right)$$

lassen sich Konfidenzintervalle und Tests für lineare Hypothesen in Analogie zu LMM konstruieren. Wenn die Anzahl der Parameter im Verhältnis zum Stichprobenumfang

Penalisierte Likelihood- / empirische Bayes-Schätzung

0. Wähle Startwerte $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\gamma}}^{(0)}, \hat{\boldsymbol{\vartheta}}^{(0)}$.
1. Bestimme für $k = 1, 2, \ldots$ die Pseudodaten

$$\tilde{\boldsymbol{y}}^{(k)} = \tilde{\boldsymbol{y}}(\hat{\boldsymbol{\vartheta}}^{(k)}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{U}\hat{\boldsymbol{\gamma}}^{(k)} + \boldsymbol{D}(\hat{\boldsymbol{\vartheta}}^{(k)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\vartheta}}^{(k)}))$$

   und die Gewichtsmatrix $\boldsymbol{W}^{(k)} = \boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)}, \hat{\boldsymbol{\vartheta}}^{(k)})$. Bestimme $\hat{\boldsymbol{\beta}}^{(k+1)}$, $\hat{\boldsymbol{\gamma}}^{(k+1)}$ durch Lösen des Gleichungssystems (6.46).
2. Bestimme für $k = 1, 2, \ldots$ den (approximativen) ML-Schätzer $\hat{\boldsymbol{\vartheta}}^{(k+1)}$ durch einen Iterationsschritt eines numerischen Verfahrens zur Maximierung von $l_R(\boldsymbol{\vartheta}; \hat{\boldsymbol{\beta}}^{(k+1)}, \hat{\boldsymbol{\gamma}}^{(k+1)})$, z.B. durch einen Newton-Raphson-Schritt

$$\hat{\boldsymbol{\vartheta}}^{(k+1)} = \hat{\boldsymbol{\vartheta}}^{(k)} + \boldsymbol{F}_R(\hat{\boldsymbol{\vartheta}}^{(k)})^{-1}\boldsymbol{s}_R(\hat{\boldsymbol{\vartheta}}^{(k)}),$$

   wobei $\boldsymbol{s}_R = \partial\, l_R / \partial\, \boldsymbol{\vartheta}$ die Score-Funktion und $\boldsymbol{F}_R(\boldsymbol{\vartheta})$ die Fisher-Informationsmatrix von $l_R(\boldsymbol{\vartheta})$, ausgewertet für $\hat{\boldsymbol{\vartheta}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k+1)}, \hat{\boldsymbol{\gamma}}^{(k+1)}$ sind.
3. Nach Iteration von 1. und 2. bis zur Konvergenz (gemäß einem Abbruchkriterium) erhält man $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\vartheta}}$.

groß ist, muss jedoch mit einer deutlichen Verzerrung und einer Verschlechterung der Normalverteilungsapproximation gerechnet werden. Dies gilt insbesondere für Zielvariablen mit geringer Information, wie etwa binären Zielvariablen. Als Alternative oder Ergänzung ist dann eine Analyse mit der vollen Bayes-Inferenz des folgenden Abschnitts zu empfehlen.

6.7.2 Volle Bayes-Inferenz mit MCMC

Wie für LMM kann die penalisierte Likelihood-Inferenz des vorigen Abschnitts auch als *empirische Bayes-Schätzung*, basierend auf dem Posteriori-Modus, aufgefasst werden. Für eine *volle Bayes-Inferenz* wählen wir für $\boldsymbol{\beta}$ und $\boldsymbol{\vartheta}$ die gleichen Priori-Verteilungen wie für LMM, d.h. eine Normalverteilung

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M})$$

mit dem Grenzfall einer nichtinformativen Priori

$$p(\boldsymbol{\beta}) \propto const \quad \text{für} \quad \boldsymbol{M}^{-1} \to \boldsymbol{0}.$$

Bezüglich $p(\boldsymbol{\vartheta})$ beschränken wir uns wie bei LMM im Weiteren auf *GLMM für Longitudinal- oder Clusterdaten* mit

$$\mathrm{E}(y_{ij} \,|\, \boldsymbol{\gamma}_i) = \mu_{ij} = h(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{u}_{ij}'\boldsymbol{\gamma}_i), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i,$$

und zufälligen i.i.d. Effekten

$$\boldsymbol{\gamma}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D}).$$

Dabei nehmen wir wieder zur Vereinfachung $\boldsymbol{D} = \mathrm{diag}(\tau_1^2, \ldots, \tau_q^2)$ und $p(\boldsymbol{D}) = p(\tau_1^2) \cdot \ldots \cdot p(\tau_q^2)$ mit

$$\tau_j^2 \sim \mathrm{IG}(a_j, b_j),\ j = 1, \ldots, q.$$

an und betrachten nur GLMM mit bekanntem Skalenparameter $\phi = 1$. Andernfalls, wie zum Beispiel bei gammaverteilten Zielvariablen, muss wie bei LMM mit normalverteilten Zielvariablen eine zusätzliche Priori für ϕ gewählt werden.

Für die *Posteriori-Verteilung* gilt

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m, \boldsymbol{D} \,|\, \boldsymbol{y}) \propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(\boldsymbol{y}_{ij} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}_i) p(\boldsymbol{\beta}) \prod_{i=1}^{m} p(\boldsymbol{\gamma}_i \,|\, \boldsymbol{D}) p(\boldsymbol{D}).$$

Daraus folgt für die *vollständig bedingten Dichten*:

$$\begin{aligned}
p(\boldsymbol{\beta} \,|\, \cdot) &\propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(\boldsymbol{y}_{ij} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}_i) p(\boldsymbol{\beta}) \\
p(\boldsymbol{\gamma}_i \,|\, \cdot) &\propto \prod_{j=1}^{n_i} p(\boldsymbol{y}_{ij} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}_i) p(\boldsymbol{\gamma}_i \,|\, \boldsymbol{D}), \quad i = 1, \ldots, m \\
p(\tau_j^2 \,|\, \cdot) &\propto \prod_{i=1}^{m} p(\boldsymbol{\gamma}_{ij} \,|\, \tau_j^2) p(\tau_j^2), \quad j = 1, \ldots, q.
\end{aligned}$$

Im Gegensatz zu LMM mit Normalverteilungsannahme sind nun für GLMM mit nicht normalverteilten Zielvariablen, wie für Binomial- oder Poisson-Modelle, die vollständig bedingten Dichten für $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}_i$, $i = 1, \ldots, m$, nicht mehr durch eine bekannte Verteilung in geschlossener Form darstellbar. Somit können Zufallszahlen aus $p(\boldsymbol{\beta} \,|\, \cdot)$ und $p(\boldsymbol{\gamma}_i \,|\, \cdot)$ nicht direkt gezogen werden, und statt des Gibbs-Samplers für LMM in Abschnitt 6.5 werden Metropolis-Hastings (MH)-Algorithmen verwendet. Dazu gibt es verschiedene Vorschläge. Wir bevorzugen den automatisch ablaufenden Algorithmus von Gamerman (1997) mit sogenannten IWLS (iteratively weighted least squares)-Vorschlagsdichten.

Die vollständig bedingten Dichten $p(\tau_j^2 \,|\, \cdot)$ sind weiterhin invers gammaverteilt

$$\tau_j^2 \,|\, \cdot \sim \mathrm{IG}(\tilde{a}_j, \tilde{b}_j)$$

mit den Parametern

$$\tilde{a}_j = a_j + \frac{m}{2}, \qquad \tilde{b}_j = b_j + \frac{1}{2} \sum_{i=1}^{m} \gamma_{ij}^2.$$

Beispiel 6.9 Anzahl der Zitate von Patenten

Wir analysieren das log-lineare Poisson-Normal-Modell aus Beispiel 6.8 mit voller Bayes-MCMC-Inferenz. Für die fixen Effekte wählen wir nichtinformative Priori-Verteilungen und für die Varianz der γ_{0i} eine inverse Gammaverteilung $\tau_0^2 \sim \mathrm{IG}(a, b)$ mit $a = b = 0.001$. Tabelle 6.6 zeigt für die festen Effekte der Kovariablen eine weitgehend gute Übereinstimmung mit den Ergebnissen aus Beispiel 4.7 (Tabelle 4.10, Seite 214), sowohl für die geschätzten Effekte selbst als auch für die zugehörigen geschätzten Standardabweichungen. Die Effekte von *uszw* und *patus* sind wiederum nichtsignifikant. Es ist

Volle Bayes-Schätzung mit MCMC

Wähle Startwerte $\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}_i^{(0)}, i = 1, \ldots, m, \boldsymbol{D}^{(0)}$ bzw. $\sigma_\gamma^{2(0)}$. Ziehe für $k = 1, 2, \ldots$ Zufallszahlen aus den vollständig bedingten Dichten.

1. $\boldsymbol{\beta}^{(k)} \mid \cdot$ mit IWLS-MH-Algorithmus.
2. $\boldsymbol{\gamma}_i^{(k)} \mid \cdot$, $i = 1, \ldots, m$, mit IWLS-MH-Algorithmus.
3. $(\tau_j^2)^{(k)} \mid \cdot$ aus inversen Gammaverteilungen.

Variable	Koeffizient	Standard-abweichung	2.5%-Quantil	97.5%-Quantil
Konstante	-0.356	0.064	-0.479	-0.230
jahrc	-0.091	0.006	-0.104	-0.079
alandc	-0.035	0.008	-0.051	-0.021
anspc	0.019	0.003	0.014	0.024
biopharm	0.198	0.064	0.074	0.329
uszw	0.004	0.049	-0.094	0.101
patus	-0.126	0.052	-0.224	-0.018
patdsg	-0.306	0.059	-0.423	-0.192
einspruch	0.417	0.047	0.325	0.508

Tabelle 6.6. *Anzahl der Zitate von Patenten: Poisson-Bayes-Modell.*

deshalb nicht verwunderlich, dass sich die Schätzwerte in den beiden Tabellen dafür (scheinbar) deutlicher unterscheiden. Die geschätzte Konstante -0.356 kann nicht direkt mit der entsprechenden Konstante 0.176 aus Tabelle 4.10 verglichen werden: Es ist ja eine Korrektur um $\tau_0^2/2$ vorzunehmen, vgl. die Bemerkung 4 zum Poisson-Normal-Modell.

Als Schätzwert für τ_0^2 ergibt sich $\hat{\tau}_0^2 = 1.265$. Nach der Korrektur sind auch die geschätzten Konstanten vergleichbar. Wie in Beispiel 4.9 zeigen wir in Abbildung 6.5 geschätzte Posteriori-Dichten der (Populations-) Effekte der Kovariablen *uszw*, *patus* und *patdsg*. Die Posteriori-Dichten erscheinen wieder in guter Näherung als normalverteilt. Zusätzlich kann auch die Posteriori-Dichte der zufälligen Effekte γ_{0i} geschätzt werden. (Abbildung 6.5 rechts unten). Diese Posteriori-Dichte weicht nun deutlich von einer Normalverteilungsdichte ab.

△

6.8 Bemerkungen und Ergänzungen

LMM für Longitudinaldaten werden ausführlich von Verbeke & Molenberghs (2000) beschrieben. Der Schwerpunkt liegt methodisch in der Likelihood-Inferenz und inhaltlich bei biostatistischen Anwendungen. Dies gilt in ähnlicher Weise auch für entsprechende Kapitel über LMM und GLMM im Standardwerk von Diggle et al. (2002) zur Analyse

Abb. 6.5. *Anzahl der Zitate von Patenten: Posteriori-Dichten (durchgezogen) und Normalverteilungsapproximationen (gestrichelt) für die Effekte von uszw, patus und patdsg sowie die Posteriori-Verteilung der zufälligen Effekte (rechts unten).*

von Longitudinaldaten. Stärker methodisch orientiert ist die Darstellung von McCulloch & Searle (2001).

Zur Analyse von Longitudinaldaten existieren neben gemischten Modellen zwei weitere konzeptionelle Ansätze, die wir nicht behandeln: Sogenannte *konditionale* oder *autoregressive Modelle* erweitern lineare und generalisierte lineare Modelle für Querschnittsdaten dahingehend, dass in den linearen Prädiktor $\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$ zusätzlich Werte der Zielvariablen für vorangegangene Zeitpunkte $j-1, j-2, \ldots$ einbezogen werden. Die einfachste Möglichkeit ist

$$\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + y_{i,j-1}\alpha,$$

also ein autoregressiver Prädiktor 1. Ordnung. Für die statistische Inferenz wird $y_{i,j-1}$ formal wie eine zusätzliche Kovariable behandelt.

Marginale Modelle, auch *GEE* (generalized estimating equations)-*Ansätze* genannt, gehen von einer multivariaten Version von Score-Funktionen für die Beobachtungsvektoren $\boldsymbol{y}_i, i = 1, \ldots, m$, aus. Diese Score-Funktionen haben genau die gleiche Struktur wie die Score-Funktion für kategoriale Zielvariablen in Abschnitt 5.4, jedoch wird die Kovarianzmatrix $\boldsymbol{\Sigma}_i$ der Multinomialverteilung durch eine ,,Arbeitskovarianzmatrix" $\boldsymbol{S}_i$ für die zeitliche Korrelation ersetzt. Dies führt zu einer multivariaten Erweiterung des Quasi-Likelihood-Ansatzes aus Abschnitt 4.5. Aus der Schätzgleichung $\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = 0$ erhält man durch (Quasi-) Fisher-Scoring den *Quasi-Likelihood-* oder *GEE-Schätzer* $\hat{\boldsymbol{\beta}}$ in marginalen Modellen.

Beide Modellklassen werden in Diggle et al. (2002) und, etwas knapper, in Fahrmeir & Tutz (2001) dargestellt.

7 Nichtparametrische Regression

Wesentliches Ziel der nichtparametrischen Regression ist die flexible Modellierung des Einflusses metrischer Kovariablen auf die abhängige Variable. Dass eine rein lineare Modellierung in vielen praktischen Anwendungen nicht ausreichend ist, haben wir bereits an zahlreichen Beispielen und Anwendungen gesehen. Grundlage waren hierfür einerseits inhaltliche Überlegungen, zum anderen aber auch die Unsicherheit über die genaue Form des Einflusses der Kovariablen. In Abschnitt 3.1.4 haben wir zwei mögliche Vorgehensweisen zur Modellierung des Einflusses metrischer Kovariablen kennengelernt: einfache Transformationen von Kovariablen und Polynome. Als diagnostisches Hilfsmittel wurden dabei Streudiagramme der Residuen, insbesondere partielle Residuenplots verwendet. Obwohl einfache Problemstellungen mit Hilfe dieser Verfahren häufig noch zufriedenstellend bearbeitet werden können, ist die praktische Umsetzung bei einer größeren Zahl von Einflussgrößen kaum noch handhabbar. Darüberhinaus ist durch die vorgegebene Menge zugelassener Transformationen die funktionale Form des Einflusses erneut relativ eingeschränkt und wenig flexibel. Ziel dieses Kapitels ist es daher, flexiblere Modellierungsmöglichkeiten einzuführen, die darüberhinaus möglichst automatisiert anwendbar sein sollen. Dabei werden wir uns zunächst auf die Fragestellung konzentrieren, wie der Einfluss *einer* einzelnen metrischen Kovariablen auf eine (approximativ) normalverteilte abhängige Variable modelliert werden kann. Die dabei betrachteten Ansätze bilden dann die Grundlage für die in Abschnitt 7.2 betrachteten bivariaten Glättungsverfahren und Kapitel 8 behandelten komplexeren Modelle.

Betrachten wir als Motivation für die folgenden Überlegungen zwei Beispiele: Im ersten Beispiel beschäftigen wir uns, ähnlich wie in Beispiel 1.2 (Seite 5), mit der Analyse des Unterernährungsrisikos von Kindern in Entwicklungsländern, in diesem Fall in Tansania. Wie in Beispiel 1.2 verwenden wir den Z-Score für chronische Unterernährung als Zielvariable, dessen Abhängigkeit vom Alter des Kindes im Folgenden untersucht werden soll. Abbildung 7.1 zeigt das zugehörige Streudiagramm für einen der Distrikte Tansanias (Ruvuma), auf dessen Betrachtung wir uns in diesem Kapitel beschränken werden, um durch den geringeren Stichprobenumfang übersichtlichere Grafiken zu erhalten. Die damit erzielten Ergebnisse sind jedoch im Wesentlichen auf den gesamten Datensatz übertragbar. Aus Abbildung 7.1 lässt sich ein nichtlinearer Zusammenhang vermuten, bei dem junge Kinder offenbar einen höheren Z-Score, d.h. ein geringeres Risiko für Unterernährung besitzen als ältere Kinder. Auf den ersten Blick erscheint zur Modellierung des Alterseinflusses ein einfaches Polynom oder die inverse Transformation des Alters geeignet zu sein, wir werden jedoch im Laufe von Abschnitt 7.1 erkennen, dass eine flexiblere Modellierung zusätzliche Erkenntnisse erbringt.

Als zweites Beispiel verwenden wir einen simulierten Datensatz, der in Abbildung 7.2 visualisiert ist. Die linke Grafik zeigt ein Streudiagramm der simulierten Daten, in der rechten Abbildung ist zusätzlich der zugrunde liegende, wahre Effekt der Kovariablen eingezeichnet. Offenbar ist es schwierig, diese Funktion durch einfache Transformationen oder Polynome zu approximieren, obwohl aus dem Streudiagramm die Struktur der Funktion bereits relativ deutlich zu erkennen ist.

Abb. 7.1. *Streudiagramm des Z-Score für chronische Unterernährung gegen das Alter des Kindes für einen der Distrikte Tansanias (Ruvuma).*

Abb. 7.2. *Streudiagramm eines simulierten Datensatzes mit nichtlinearem Einfluss einer Kovariablen. In der rechten Abbildung ist zusätzlich die wahre funktionale Form des Einflusses eingezeichnet. Die Daten wurden gemäß dem Modell* $y = f(x) + \varepsilon$ *mit* $f(x) = \sin(2(4x-2)) + 2\exp(-(16^2)(x-0.5)^2)$ *und* $\varepsilon \sim N(0, 0.3^2)$ *erzeugt.*

7.1 Univariate Glättung

Thema dieses ersten, grundlegenden Abschnitts zur nichtparametrischen Regression ist die Behandlung von Verfahren, die eine flexible Modellierung des Einflusses genau einer metrischen Kovariablen auf eine ebenfalls als metrisch angenommene abhängige Variable erlauben. Solche Verfahren werden auch als *Streudiagramm-Glätter* (Scatterplot-Smoother) bezeichnet, da sich die zugrunde liegende Datenlage am Besten in einem Streudiagramm visualisieren lässt und das Ziel die Bestimmung einer glatten Funktion der Kovariablen ist. Genauer gehen wir im Folgenden davon aus, dass Daten der Form (y_i, z_i), $i = 1, \ldots, n$, gegeben sind, wobei y_i Beobachtungen der Zielvariablen und z_i die zugehörigen Werte einer metrischen Kovariablen sind. Im Standardmodell der univariaten nichtparametrischen Regression geht man davon aus, dass sich die Zielvariable erklären lässt durch eine deterministische Funktion der Kovariablen und einen additiven Messfehler:

$$y_i = f(z_i) + \varepsilon_i.$$

Unterschiedliche Annahmen über die Funktion f führen dann zu verschiedenen Modellierungsmöglichkeiten. Häufig werden dabei qualitative Annahmen über die Glatt-

Univariate Glättung

Daten

Gegeben sind Messungen (y_i, z_i), $i = 1, \ldots, n$, zu einer metrischen Zielvariable y und einer metrischen Kovariable z.

Modell

$$y_i = f(z_i) + \varepsilon_i$$

mit unabhängig und identisch verteilten Fehlern für die

$$\mathrm{E}(\varepsilon_i) = 0 \quad \text{und} \quad \mathrm{Var}(\varepsilon_i) = \sigma^2$$

gilt. In einigen Fällen wird zusätzlich angenommen, dass die Fehler normalverteilt sind, also

$$\varepsilon_i \text{ i.i.d. } \mathrm{N}(0, \sigma^2).$$

heit der Funktion, also über ihre Stetigkeit oder Differenzierbarkeit zugrunde gelegt, um zu einem einfacheren Schätzproblem zu gelangen. Dabei sei bereits hier darauf hingewiesen, dass einige unter der Bezeichnung nichtparametrische Regression behandelte Glättungsverfahren tatsächlich rein parametrisch sind. Im Unterschied zu den bisher behandelten Regressionsmodellen werden diese Modelle jedoch in der Regel von einer großen Zahl von Parametern bestimmt, die einzeln nicht mehr sinnvoll interpretierbar sind. Wir werden im Folgenden dennoch weiter die gebräuchliche Bezeichnung nichtparametrische Regression verwenden, häufig jedoch auch von univariater Glättung sprechen.

Über den Fehler werden in nichtparametrischen Regressionsmodellen die gleichen Annahmen getroffen wie in Kapitel 3 zum linearen Modell: Die Fehler sind unabhängig und identisch verteilt mit

$$\mathrm{E}(\varepsilon_i) = 0 \quad \text{und} \quad \mathrm{Var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n.$$

Daraus folgt, ähnlich wie im linearen Modell,

$$\mathrm{E}(y_i) = f(z_i) \quad \text{und} \quad \mathrm{Var}(y_i) = \sigma^2, \quad i = 1, \ldots, n,$$

so dass durch die Funktion f der Erwartungswert der Zielvariablen modelliert wird. Für einige Überlegungen, insbesondere die Konstruktion von Konfidenzintervallen, werden wir darüberhinaus annehmen, dass die Fehler normalverteilt sind.

7.1.1 Polynom-Splines

Als erste Modellierungsmöglichkeit zur flexiblen Beschreibung des Einflusses einer stetigen Kovariablen wollen wir nun die sogenannten *Polynom-Splines* oder auch *Regressions-Splines* behandeln, die an der Modellierung über Polynome orientiert sind. Betrachten wir zur Motivation zunächst noch einmal das polynomiale Modell

$$f(z) = \gamma_0 + \gamma_1 z + \ldots + \gamma_l z^l$$

Abb. 7.3. *Polynomiale Regression für den simulierten Datensatz.*

in dem der Einfluss von z auf y durch ein Polynom vom Grad l modelliert wird. Damit befinden wir uns wieder im Rahmen der in Kapitel 3 behandelten Modelle, so dass die Regressionskoeffizienten der Polynome und damit auch f durch den üblichen KQ-Schätzer bestimmt werden können. Im Gegensatz zu Kapitel 3 bezeichnen wir die Regressionskoeffizienten mit γ_j, um die im Weiteren Verlauf dieses Kapitels behandelten flexibleren Modellierungsmöglichkeiten von linearen Modellen zu unterscheiden.

Warum eine rein polynomiale Modellierung zur Beschreibung von $f(z)$ häufig nicht ausreichend ist, illustriert Abbildung 7.3 anhand des simulierten Datenbeispiels. Hier wurde die Schätzung von f mit Hilfe von Polynomen bis zu einem Grad von $l = 15$ durchgeführt. Offenbar sind Polynome selbst für einen (bereits extrem hohen) Grad von $l = 11$ nicht in der Lage, das lokale Maximum bei $z = 0.5$ adäquat zu erfassen. Für $l = 15$ wird zwar dieses Maximum annähernd erreicht, dafür ist der geschätzte Zusammenhang aber in anderen Bereichen sehr rau, so dass auch diese Schätzung nicht als besonders geeignet erscheint.

Um die polynomiale Modellierung flexibler zu gestalten, zerlegen wir nun den Wertebereich der Kovariablen in Intervalle und schätzen jeweils separate Polynome. Anstelle eines globalen Modells verwendet man also mehrere, lokal definierte Polynome, die aber weiterhin mit den in Kapitel 3 behandelten Verfahren für lineare Modelle geschätzt werden können. Die linke Grafik in Abbildung 7.4 illustriert dieses Vorgehen. Hier haben wir den Definitionsbereich von z in 10 Intervalle der Breite 0.1 eingeteilt und jeweils ein kubisches Polynom angepasst. Damit wird tatsächlich eine flexiblere Funktionsschätzung erreicht, die den Verlauf der wahren Funktion gut widerspiegelt. Allerdings wird aus der geschätzten Funktion auch unmittelbar der wesentliche Nachteil stückweiser Poly-

Abb. 7.4. *Stückweise polynomiale Regression und Polynom-Splines.*

nome deutlich: Die auf den einzelnen Intervallen geschätzten Polynome ergeben keine glatte Funktion, da sie an den Intervallgrenzen verschiedene Funktionswerte aufweisen und damit zu einer unstetigen Funktion führen. Wünschenswert wäre es also, an den Intervallgrenzen zusätzliche Glattheitsanforderungen an die zu schätzende Funktion zu stellen und somit die rechte Grafik in Abbildung 7.4 zu erhalten. Die Einführung solcher zusätzlicher Glattheitsbedingungen führt nun genau zur Klasse der Polynom-Splines. Die zugrunde liegende Idee ist weiterhin die lokale Definition von Polynomen auf Intervallen des Definitionsbereichs der Kovariablen. Um für ein Polynom vom Grad l aber auch global eine ausreichende Glattheit zu gewährleisten, fordern wir zusätzlich, dass die resultierende Funktion an den Intervallgrenzen $(l-1)$-mal stetig differenzierbar sein soll. Damit kommen wir zu folgender Definition:

Polynom-Splines

Eine Funktion $f : [a, b] \to \mathbb{R}$ heißt Polynom-Spline vom Grad $l \geq 0$ zu den Knoten $a = \kappa_1 < \ldots < \kappa_m = b$, falls sie die folgenden Bedingungen erfüllt:

1. $f(z)$ ist $(l-1)$-mal stetig differenzierbar. Für $l = 1$ entspricht dies der Forderung, dass $f(z)$ stetig ist, für $l = 0$ werden keine Glattheitsanforderungen an $f(z)$ gestellt.
2. $f(z)$ ist auf den durch die Knoten gebildeten Intervallen $[\kappa_j, \kappa_{j+1})$ ein Polynom vom Grad l.

Damit werden unsere obigen Überlegungen präzisiert und man erhält eine stückweise polynomiale Funktion, wobei die Zerlegung des Definitionsbereichs über die sogenannten *Knoten* $\kappa_1 < \ldots < \kappa_m$ erfolgt. Darüberhinaus soll die Funktion insgesamt $(l-1)$-mal stetig differenzierbar sein, so dass sich insbesondere die gewünschten Glattheitsrestriktionen an den Knoten ergeben. Abbildung 7.5 zeigt zu einer kleinen Knotenmenge einige einfache Polynom-Splines, die den Einfluss des angenommenen Splinegrads veranschaulichen. Offenbar bestimmt der Splinegrad l die globale Glattheit des Polynom-Splines, während die Vielfältigkeit darstellbarer Funktionen im Wesentlichen durch die Knotenzahl festgelegt ist. Je mehr Knoten verwendet werden, desto größer ist die Zahl an stückweisen Polynomen, aus denen der Polynom-Spline zusammengesetzt ist. Wir werden später den Einfluss des Splinegrads und der Knotenzahl noch ausführlicher diskutieren.

Abb. 7.5. *Beispiele für Splines vom Grad 0, 1, 2 und 3 zu den Knoten* $\kappa_1 = 0$, $\kappa_2 = 0.25$, $\kappa_3 = 0.5$, $\kappa_4 = 0.75$ *und* $\kappa_5 = 1$.

Bevor wir Polynom-Splines in der nichtparametrischen Regression einsetzen können, benötigen wir zunächst eine konstruktive Darstellung, die es uns erlaubt, die Menge der Polynom-Splines zu charakterisieren. Dazu existieren verschiedene, äquivalente Möglichkeiten. In den beiden folgenden Abschnitten werden wir die zwei populärsten Varianten besprechen: Trunkierte Potenzen und B-Splines.

Polynom-Splines und trunkierte Potenzen

Gehen wir zunächst von dem Regressionsmodell

$$y_i = \gamma_1 + \gamma_2 z_i + \ldots + \gamma_{l+1} z_i^l + \gamma_{l+2}(z_i - \kappa_2)_+^l + \ldots + \gamma_{l+m-1}(z_i - \kappa_{m-1})_+^l + \varepsilon_i$$

mit

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

aus. Dieses Modell besteht zunächst aus einem globalen Polynom vom Grad l, wie wir es zu Beginn dieses Kapitels betrachtet haben. Zusätzlich ändert sich der Koeffizient des höchsten Polynoms aber an jedem der Knoten $\kappa_2, \ldots, \kappa_{m-1}$. Damit wird einerseits erlaubt, dass man auf jedem der durch die Knoten gebildeten Intervalle lokale Polynome verwenden kann, andererseits wird aber auch die globale Glattheitsanforderung erfüllt. Abbildung 7.6 verdeutlicht dies am Beispiel eines Polynom-Splines vom Grad $l = 1$. Abbildung 7.6 a) zeigt die Funktionen, die unser Modell definieren, also die Funktionen eines globalen Polynoms vom Grad $l = 1$ (als gestrichelte Linien) und die weiteren, sogenannten trunkierten Geraden (als durchgezogene Linien). Diese Funktionen werden dann

Abb. 7.6. *Konstruktion von Basisfunktionen für lineare Polynom-Splines.*

durch die Regressionskoeffizienten passend skaliert, so dass sich Abbildung 7.6 b) ergibt. Dabei wurden wieder äquidistante Knoten mit Abstand 0.1 zugrunde gelegt. Die horizontale Linie bei $y \approx 0.8$ entspricht der globalen Konstante γ_1. Auf dem ersten Intervall $[0, 0.1)$ erhält man eine von diesem globalen Niveau fallende Funktion, deren Steigung durch γ_2 gegeben ist. Ab dem Knoten $\kappa_3 = 0.2$ wird diese Steigung überlagert durch γ_3. In unserem Beispiel ist auch γ_3 negativ, so dass sich insgesamt ein noch etwas steiler fallender Funktionsverlauf ergibt. Durch den positiven Koeffizienten γ_4 wird die negative Steigung dann ab $\kappa_4 = 0.3$ wieder verringert, allerdings ergibt sich insgesamt immer noch eine negative Steigung. Betrachtet man so den weiteren Verlauf der Funktion, so geben die Koeffizienten γ_j jeweils die Änderung der Steigung an, die an einem Knoten vorgenommen wird. Durch die spezielle Konstruktion der Funktionen $(z - \kappa_j)_+^l$ wird dabei sichergestellt, dass diese Steigungsänderungen glatt genug erfolgen, so dass die Anforderungen an einen Polynom-Spline erfüllt sind. Summiert man schließlich die skalierten Funktionen, so ergibt sich die in Abbildung 7.6 c) wiedergegebene Darstellung von $f(z)$.

Formaler kann man zeigen, dass jeder Polynom-Spline zum Grad l und mit Knoten $\kappa_1 < \ldots < \kappa_m$ eindeutig als Linearkombination der $d = m + l - 1$ Funktionen

$$B_1(z) = 1, \quad B_2(z) = z, \quad \ldots, \quad B_{l+1}(z) = z^l,$$
$$B_{l+2}(z) = (z - \kappa_2)_+^l, \quad \ldots, \quad B_d(z) = (z - \kappa_{m-1})_+^l,$$

dargestellt werden kann (vergleiche die Abbildungen 7.7 und 7.8, in denen diese Funktionen für $l = 0$ und $l = 2$ für eine kleine Knotenmenge dargestellt sind). Damit ergibt sich folgende Repräsentation des nichtparametrischen Regressionsproblems:

Abb. 7.7. *TP-Basis für Splines vom Grad 0 zur Knotenmenge {0,0.25,0.5,0.75,1}.*

$$y_i = f(z_i) + \varepsilon_i = \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i.$$

Die Funktionen $B_1, \ldots, B_d$ werden als *Basisfunktionen* bezeichnet, da die Menge aller Polynom-Splines eindeutig durch die Angabe dieser Funktionen charakterisiert werden kann. Genauer spricht man hier von der *Basis der trunkierten Potenzen* (Truncated Power Series Basis, TP-Basis), um zwischen verschiedenen Darstellungen des Spline-Raums unterscheiden zu können. Tatsächlich kann man zeigen, dass es sich bei dem Spline-Raum um einen d-dimensionalen Vektorraum handelt, so dass die Bezeichnung als Basis gerechtfertigt ist (vergleiche Hämmerlin & Hoffmann (1994) für einen mathematisch exakten Nachweis).

Der Vorteil der Modellierung von $f(z)$ als Polynom-Spline liegt darin, dass sich das nichtparametrische Regressionsmodell immer noch als lineares Modell auffassen lässt, allerdings mit einer möglicherweise großen Zahl von Parametern. Definiert man den Vektor der beobachteten Zielvariablen $\boldsymbol{y}$ und der Fehlerterme $\boldsymbol{\varepsilon}$ sowie zusätzlich die Designmatrix

$$\boldsymbol{Z} = \begin{pmatrix} B_1(z_1) & \ldots & B_d(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \ldots & B_d(z_n) \end{pmatrix} = \begin{pmatrix} 1 & z_1 & \ldots & z_1^l & (z_1-\kappa_2)_+^l & \ldots & (z_1-\kappa_{m-1})_+^l \\ \vdots & & & & & & \vdots \\ 1 & z_n & \ldots & z_n^l & (z_n-\kappa_2)_+^l & \ldots & (z_n-\kappa_{m-1})_+^l \end{pmatrix},$$

so erhält man die Modellgleichung

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mit dem Koeffizientenvektor $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$. Offensichtlich handelt es sich dabei um ein lineares Modell in dem Parametervektor $\boldsymbol{\gamma}$. Als üblichen Schätzer erhält man den

Abb. 7.8. *TP-Basis für Splines vom Grad 2 zur Knotenmenge* $\{0{,}0.25{,}0.5{,}0.75{,}1\}$.

KQ-Schätzer

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

Allerdings ist in diesem Modell im Gegensatz zu den linearen Modellen in Kapitel 3 eine Interpretation der einzelnen Parameterschätzer $\hat{\gamma}_j$ kaum sinnvoll. Stattdessen interessiert man sich für die Form der geschätzten Funktion, die sich aus den geschätzten Koeffizienten ergibt, d.h. man betrachtet

$$\hat{f}(z) = \boldsymbol{z}'\hat{\boldsymbol{\gamma}}$$

mit $\boldsymbol{z} = (B_1(z), \ldots, B_d(z))'$ in Abhängigkeit von Kovariablenausprägungen z. Die Güte der Modellanpassung lässt sich beispielsweise in einem Streudiagramm der Daten in Kombination mit der geschätzten Kurve beurteilen. Wir werden später noch auf geeignete Maßzahlen zur Beurteilung der Modellanpassung zurückkommen.

Abb. 7.9. *Nichtparametrische Schätzung des Alterseffekts basierend auf Polynom-Splines.*

Beispiel 7.1 Unterernährung in Tansania: Modellierung durch Polynom-Splines

Als erste Demonstration zur nichtparametrischen Regression vergleichen wir verschiedene Spezifikationen für den Splinegrad und die Knotenzahl anhand der Tansania-Daten. Abbildung 7.9 gibt eine Reihe solcher Schätzungen wieder. Offenbar bestimmt der Splinegrad im Wesentlichen die Glattheit der geschätzten Funktion, während die Knotenzahl die Flexibilität der Schätzung beeinflusst. Wie unmittelbar aus der Definition von Polynom-Splines ersichtlich ist, erhält man für den Splinegrad $l = 0$ eine unstetige, stückweise konstante Funktion. Dies widerspricht jedoch den inhaltlichen Überlegungen, die einen glatteren Verlauf des Z-Score bei variierendem Alter nahe legen. Erhöht man

den Splinegrad entsprechend, erhält man zunächst eine stetige ($l = 1$), dann eine auch stetig differenzierbare Funktion ($l = 2$).

Ein Vergleich der beiden Knotenzahlen zeigt, dass eine größere Zahl von Knoten zu deutlich raueren Funktionsschätzungen führt, unabhängig vom verwendeten Splinegrad. Offenbar entspricht die Funktionsschätzung mit 11 Knoten besser den theoretischen Vorstellungen über den Zusammenhang zwischen Alter und Unterernährung, während für 21 Knoten viele zusätzliche Schwankungen im Funktionsverlauf auftreten, die nicht sinnvoll interpretierbar sind. Interessanterweise zeichnet sich neben dem niedrigeren Unterernährungsrisiko für junge Kinder auch ein unerwarteter Anstieg der Funktionsschätzung nach ca. 24 Monaten ab. Dieser Anstieg lässt sich jedoch gut interpretieren, da zu diesem Zeitpunkt die Referenzpopulation wechselt, zu der die Größe der tansanischen Kinder in Beziehung gesetzt wird. Genauer werden bis zu einem Alter von 24 Monaten US-amerikanische Kinder weißer Eltern mit hohem sozio-ökonomischem Stand betrachtet, während für ältere Kinder eine repräsentative Stichprobe aller US-amerikanischen Kinder zugrunde gelegt wird. Durch diesen Wechsel ergibt sich künstlich ein höherer Z-Score, d.h. ein verringertes Unterernährungsrisiko in einem Alter von 24 Monaten. Dieser Einfluss der Referenzpopulation wäre mit Hilfe einfacher Transformationen oder Polynome geringen Grads vermutlich nicht zu identifizieren gewesen.

△

Einfluss der Knoten auf die Schätzung

Anhand von Beispiel 7.1 erkennt man, dass die Wahl der Zahl und Position der Knoten sowie des Splinegrads einen entscheidenden Einfluss auf die resultierende Funktionsschätzung besitzt. In der Regel werden kubische Splines als Standard verwendet, da sich so eine glatte, zweimal stetig differenzierbare Funktion ergibt. Für die Knotenzahl ist es dagegen wesentlich schwieriger, Faustregeln anzugeben. Generell gilt wie in Beispiel 7.1, dass eine größere Knotenzahl zu einer flexibleren Funktionsschätzung führt, während eine kleine Knotenzahl tendenziell in einer glatteren Schätzung resultiert. Dies ist für einen Polynom-Spline vom Grad 3 exemplarisch noch einmal in Abbildung 7.10 dargestellt, in der für verschiedene Knotenzahlen Schätzungen für den simulierten Datensatz eingezeichnet sind. Da wir für die simulierten Daten die wahre, zugrunde liegende Funktion kennen, ist es hier einfacher, die Güte der Schätzung in Abhängigkeit von der Knotenzahl zu beurteilen.

Es ergibt sich ein ähnliches Bild wie in Beispiel 7.1: Für eine geringe Zahl von Knoten ist die resultierende Funktionsschätzung sehr glatt und in unserem Beispiel nicht besonders gut an die wahre Funktion angepasst. Insbesondere wird das Maximum bei $z = 0.5$ nicht erkannt. Für 20 Knoten erhält man ein zufriedenstellendes Ergebnis, während eine größere Knotenzahl dann wieder zusätzliche Artefakte produziert und in einer zu rauen, nicht mehr sinnvoll interpretierbaren Schätzung resultiert.

Zusätzlich zur Knotenzahl ist auch noch deren Verteilung bzw. Position auf der Kovariablenachse festzulegen. In der Praxis sind die folgenden drei Ansätze verbreitet:

- *Äquidistante Knoten*: Der Wertebereich $[a, b]$ von z wird in $m - 1$ Intervalle der Breite

$$h = \frac{b - a}{m - 1}$$

  eingeteilt, so dass sich die Knoten

Abb. 7.10. *Einfluss der Knotenzahl auf die Schätzung von Polynom-Splines. Die geschätzte Funktion ist jeweils als durchgezogene Linie und die wahre Funktion als gestrichelte Linie dargestellt.*

$$\kappa_j = a + (j-1) \cdot h, \quad j = 1, \ldots, m,$$

ergeben. In den bisherigen Beispielen sind wir stillschweigend von äquidistanten Knoten ausgegangen.

- *Quantilbasierte Knoten*: Verwende die $(j-1)/(m-1)$-Quantile ($j = 1, \ldots, m$) der beobachteten Kovariablenausprägungen $z_1, \ldots, z_n$ als Knoten. Damit werden viele Knoten in Bereiche gelegt, in denen auch eine große Zahl von Beobachtungen vorliegt, so dass eine bessere Anpassung an die Verteilung der unabhängigen Variablen erreicht wird.
- *Visuelle Knotenwahl anhand des Streudiagramms*: Durch Betrachtung des Streudiagramms können die Knoten auch gezielt gesetzt werden, um die Knotendichte an die Variabilität der Daten anzupassen oder um inhaltliche Überlegungen zu berücksichtigen.

Keine der Strategien gibt jedoch eine Antwort auf die wesentliche Frage nach der Anzahl der zu verwendenden Knoten. Hierzu existieren zwei wesentliche Möglichkeiten: Die Regularisierung des Schätzproblems durch Einführung eines Strafterms ähnlich wie in der Ridge-Regression (vergleiche Abschnitt 3.6.4) sowie die adaptive (d.h. automatisiert an die Daten angepasste) Wahl der Knoten mit Hilfe von Modellwahlstrategien. Penalisierungsansätze werden in diesem Buch den Schwerpunkt bilden und sind Thema von Abschnitt 7.1.2. Adaptive Verfahren werden später in Abschnitt 7.1.10 behandelt. Zuvor betrachten wir jedoch noch eine alternative Darstellung von Polynom-Splines.

Abb. 7.11. *Einzelne B-Spline-Basisfunktionen vom Grad* $l = 0, 1, 2, 3$ *zu äquidistanten Knoten.*

B-Splines

Obwohl die TP-Basis einfach verständlich ist, existieren alternative Darstellungen von Polynom-Splines, die insbesondere aus numerischer Sicht vorzuziehen sind, aber auch in den im Folgenden behandelten Penalisierungsansätzen von Nutzen sein werden. Aufgrund ihrer Konstruktion über abgeschnittene Polynome kann die Berechnung von TP-Basen für betragsmäßig große Werte der Kovariablen zum Gleitkommaüberlauf bei der numerischen Auswertung führen. Darüberhinaus sind die Basisfunktionen der TP-Basis annähernd kollinear (also linear abhängig), besonders wenn zwei Knoten nur relativ geringen Abstand aufweisen. Wir werden deshalb als Alternative die Darstellung von Polynom-Splines mit Hilfe von *Basic-Spline*- oder einfach *B-Spline*-Basisfunktionen behandeln.

Bevor wir eine mathematisch exakte Definition von B-Splines geben, wollen wir zunächst die Konstruktion der Basis-Funktionen motivieren. Ausgangspunkt hierfür sind die Betrachtungen, die wir zu Beginn des Kapitels zur Approximation von $f(z)$ durch stückweise definierte Polynome angestellt haben. Dabei wurde offensichtlich, dass an den Knoten an die Funktion $f(z)$ zusätzliche Glattheitsanforderungen zu stellen sind. B-Spline-Basisfunktionen werden nun genau so konstruiert, dass Polynomstücke des gewünschten Grads an den Knoten ausreichend glatt zusammengesetzt werden. Genauer besteht eine B-Spline-Basisfunktion aus $(l + 1)$ Polynomstücken vom Grad l, die $(l - 1)$-mal stetig differenzierbar zusammengesetzt werden. Abbildung 7.11 zeigt die sich aus diesen Überlegungen ergebenden Basisfunktionen für die Splinegrade $l = 0, \ldots, 3$. Zur Darstellung von Polynom-Splines werden alle B-Spline-Basisfunktionen basierend auf der zugrunde liegenden Knotenmenge gebildet. Solche vollständigen B-Spline-Basen sind in Abbildung 7.12 für eine äquidistante und eine ungleichmäßig verteilte Knotenmenge dargestellt. Mit Hilfe dieser vollständigen Basen ergibt sich dann wieder die Darstellung von $f(z)$ durch eine Linearkombination von $d = m + l - 1$ Basisfunktionen, d.h.

Abb. 7.12. *B-Spline-Basen vom Grad $l = 1, 2, 3$ zu äquidistanten Knoten (links) und ungleichmäßig verteilten Knoten (rechts).*

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z).$$

Der wesentliche Vorteil der B-Spline-Basis liegt in ihrer lokalen Definition. Im Gegensatz zu den trunkierten Potenzen der TP-Basis, die ab einem bestimmten Knoten über den gesamten weiteren Definitionsbereich positiv sind, sind die B-Spline-Basisfunktionen nur über einem durch $l + 2$ benachbarte Knoten gebildeten Intervall positiv. Darüberhinaus sind die B-Spline-Basisfunktionen nach oben beschränkt, so dass die zuvor beschriebenen numerischen Probleme der TP-Basis nicht auftreten.

Für B-Splines vom Grad $l = 0$ lässt sich aus Abbildung 7.11 unmittelbar die Definition

$$B_j^0(z) = \mathbb{1}_{[\kappa_j, \kappa_{j+1})}(z) = \begin{cases} 1 & \kappa_j \leq z < \kappa_{j+1}, \\ 0 & \text{sonst}, \end{cases} \qquad j = 1, \ldots, d-1,$$

ablesen. In diesem Fall lässt sich auch die Äquivalenz zur TP-Basis besonders einfach zeigen. Bildet man nämlich sukzessive Differenzen der TP-Basisfunktionen vom Grad $l = 0$, so ergeben sich Funktionen, die über dem durch zwei benachbarte Knoten gebildeten Intervall konstant sind, also die B-Spline-Basis. Für B-Splines höheren Grads kann

man analoge Darstellungen aus stückweise zusammengesetzten Polynomen vom Grad l konstruieren. So ergeben sich die Basisfunktionen vom Grad $l = 1$ zu

$$B_j^1(z) = \frac{z - \kappa_j}{\kappa_{j+1} - \kappa_j} \mathbb{1}_{[\kappa_j, \kappa_{j+1})}(z) + \frac{\kappa_{j+2} - z}{\kappa_{j+2} - \kappa_{j+1}} \mathbb{1}_{[\kappa_{j+1}, \kappa_{j+2})}(z),$$

d.h. die Basisfunktion ist aus zwei linearen Teilstücken auf den Intervallen $[\kappa_j, \kappa_{j+1})$ und $[\kappa_{j+1}, \kappa_{j+2})$ definiert, die am Knoten κ_{j+1} stetig zusammengesetzt sind. Allgemein können B-Spline-Basisfunktionen höheren Grads rekursiv definiert werden durch die Vorschrift:

$$B_j^l(z) = \frac{z - \kappa_j}{\kappa_{j+l} - \kappa_j} B_j^{l-1}(z) + \frac{\kappa_{j+l+1} - z}{\kappa_{j+l+1} - \kappa_{j+1}} B_{j+1}^{l-1}(z).$$

Wendet man diese Formel beispielsweise auf B-Splines vom Grad $l = 1$ an, so ergibt sich genau die obige explizite Formel. Obwohl B-Splines durch ihre rekursive Definition komplizierter erscheinen als TP-Splines, wird dieser Nachteil durch die besseren numerischen Eigenschaften aufgewogen. Darüberhinaus stellen die meisten statistischen Programmpakete (etwa R oder STATA) Implementationen für B-Splines zur Verfügung, so dass der Anwender sich nicht selbst unmittelbar mit der Implementation beschäftigen muss.

Im Folgenden fassen wir einige grundlegende Eigenschaften der B-Spline-Basisfunktionen zusammen, die sich meist unmittelbar aus den Abbildungen 7.11 bzw. 7.12 ablesen lassen:

1. *B-Splines bilden eine lokale Basis.* Jede Basisfunktion ist nur in einem von $l+2$ Knoten gebildeten Bereich positiv. Bei äquidistanten Knoten besitzen zudem alle Basisfunktionen genau die gleiche Gestalt und sind lediglich entlang der z-Achse verschoben. An einer beliebigen Stelle $z \in [a, b]$ sind genau $l + 1$ Basisfunktionen positiv.
2. *Zerlegung der Einheit.* An jeder Stelle $z \in [a, b]$ gilt

$$\sum_{j=1}^{d} B_j(z) = 1.$$

3. *Überlappung mit $2l$ benachbarten Basisfunktionen.* Jede Basisfunktion (innerhalb des Definitionsbereichs $[a, b]$) überlappt sich mit genau $2l$ benachbarten Basisfunktionen.
4. *Beschränkter Wertebereich.* Der Wertebereich der einzelnen Basisfunktionen ist nach oben beschränkt.
5. *Ableitungen.* Da die einzelnen Basisfunktionen aus Polynomstücken zusammengesetzt sind, lassen sich Ableitungsformeln einfach bestimmen. Für jede einzelne Basisfunktion ergibt sich:

$$\frac{\partial}{\partial z} B_j^l(z) = l \cdot \left(\frac{1}{\kappa_{j+l} - \kappa_j} B_j^{l-1}(z) - \frac{1}{\kappa_{j+l+1} - \kappa_{j+1}} B_{j+1}^{l-1}(z) \right).$$

Daraus erhält man als Ableitung für den gesamten Polynom-Spline

$$\frac{\partial}{\partial z} \sum_j \gamma_j B_j^l(z) = l \cdot \sum_j \frac{\gamma_j - \gamma_{j-1}}{\kappa_{j+l} - \kappa_j} B_j^{l-1}(z). \tag{7.1}$$

Demzufolge lässt sich die Ableitung eines Polynom-Splines ausdrücken durch die Differenzen der entsprechenden Basiskoeffizienten und B-Spline-Basisfunktionen von um

eins verringertem Grad. Somit erhält man durch Schätzung der Koeffizienten γ_j nicht nur eine Schätzung für die Funktion selbst, sondern auch für ihre Ableitungen. Analog lassen sich Ableitungen höherer Ordnung in Abhängigkeit von höheren Differenzen der Koeffizienten und Basisfunktionen geringeren Grads schreiben (vergleiche De Boor (2001), S. 115ff)

Bevor wir die Schätzung nichtparametrischer Funktionen mit Hilfe von B-Splines behandeln, müssen wir uns zunächst noch mit einem eher technischen Problem befassen. Um die rekursive Definition von B-Splines zur Berechnung der Basisfunktionen verwenden zu können, benötigen wir zusätzlich zu den *inneren Knoten* $\kappa_1, \ldots, \kappa_m$ weitere $2l$ Knoten außerhalb des Definitionsbereichs $[a, b]$. Dies führt zur *erweiterten Knotenmenge* $\kappa_{1-l}, \kappa_{1-l+1}, \ldots, \kappa_{m+l-1}, \kappa_{m+l}$. Für äquidistante Knoten ist die Definition dieser erweiterten Knotenmenge unproblematisch, da auch außerhalb des Definitionsbereichs $[a, b]$ der gleiche Abstand h zur Definition der Knoten herangezogen werden kann. Für ungleichmäßig verteilte Knoten ist dies dagegen nicht möglich. Üblich ist es hier, den Abstand zwischen den beiden kleinsten bzw. den beiden größten Knoten innerhalb des Definitionsbereichs zu verwenden.

Wie für die TP-Basis lässt sich die Schätzung eines Polynom-Splines in B-Spline-Darstellung auf die Schätzung eines linearen Modells mit großer Parameterzahl zurückführen. Als Designmatrix erhält man

$$\boldsymbol{Z} = \begin{pmatrix} B_1^l(z_1) & \ldots & B_d^l(z_1) \\ \vdots & & \vdots \\ B_1^l(z_n) & \ldots & B_d^l(z_n) \end{pmatrix}.$$

Aufgrund der zuvor diskutierten Eigenschaften der B-Spline-Basis besitzt diese Designmatrix einige spezielle Merkmale. Zunächst fällt auf, dass die Matrix keine Konstante enthält. Diese ist durch die Zerlegung der Einheit aber bereits implizit in den übrigen Basisfunktionen enthalten und eine zusätzlich spezifizierte Konstante würde zu einem nicht identifizierbaren Modell führen. Da die B-Spline-Basis darüber hinaus lokal definiert ist, besitzt $\boldsymbol{Z}$ eine spezielle Struktur und besteht zum größten Teil aus Nullen. Für $\boldsymbol{Z}'\boldsymbol{Z}$ ergibt sich damit eine Bandmatrix der Bandweite l. Diese Eigenschaft lässt sich zum numerisch effizienten Lösen der Normalgleichungen $\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma} = \boldsymbol{Z}'\boldsymbol{y}$ ausnutzen.

In Abbildung 7.13 ist die Schätzung eines B-Splines anhand des simulierten Datenbeispiels schematisch dargestellt. Zunächst wird eine vollständige B-Spline-Basis (hier vom Grad 3) zu einer vorgegebenen Knotenzahl berechnet (Abbildung a)). Der KQ-Schätzer $\hat{\boldsymbol{\gamma}}$ liefert dann für jede Basisfunktion die zugehörige Amplitude $\hat{\gamma}_j$, mit der die Basisfunktion skaliert wird (Abbildung b)). Addiert man die skalierten Basisfunktionen, erhält man schließlich die in Abbildung c) gezeigte Schätzung.

7.1.2 Penalisierte Splines (P-Splines)

Wie wir im vorigen Abschnitt gesehen haben, hängt die Güte einer auf Polynom-Splines basierenden nichtparametrischen Funktionsschätzung stark von der Zahl der verwendeten Knoten ab. Um diesem Problem zu begegnen, existieren im Wesentlichen zwei Strategien:

- Die adaptive Wahl der Knoten basierend auf Modellwahlstrategien und
- die Regularisierung des Schätzproblems durch Penalisierungsansätze.

Abb. 7.13. *Schematische Darstellung der Schätzung eines nichtparametrischen Effekts mit B-Splines.*

Wir werden uns im Wesentlichen auf die Betrachtung von Penalisierungsansätzen beschränken und in Abschnitt 7.1.10 lediglich einen kurzen Überblick zu adaptiven Verfahren geben.

Die grundlegende Idee von *penalisierten Splines* (*P-Splines*) lässt sich folgendermaßen zusammenfassen:

- Approximiere die zu schätzende Funktion $f(z)$ durch einen Polynom-Spline mit einer *großen* Zahl von Knoten (üblicherweise ca. 20 bis 40). Damit wird gewährleistet, dass $f(z)$ prinzipiell flexibel genug ist, auch stark variierende Funktionen darzustellen.
- Führe zusätzlich einen Strafterm ein, der eine zu große Variabilität der Schätzung bestraft, und minimiere anstelle des üblichen KQ-Kriteriums ein um diesen Strafterm erweitertes *penalisiertes KQ-Kriterium*.

P-Splines basierend auf der TP-Basis

Betrachten wir dazu zunächst wieder die TP-Basis, d.h.

$$f(z) = \gamma_1 + \gamma_2 z + \ldots + \gamma_{l+1} z^l + \gamma_{l+2}(z - \kappa_2)_+^l + \ldots + \gamma_d (z - \kappa_{m-1})_+^l.$$

Offenbar besteht diese Basis aus zwei verschiedenen Teilen, die eine globale polynomiale Form durch die ersten $l + 1$ Basisfunktionen und Abweichungen von diesem Polynom

durch die abgeschnittenen Potenzen modellieren. Raue Funktionsschätzungen entstehen im Wesentlichen durch eine zu große Variabilität des zweiten Modellteils, so dass eine naheliegende Regularisierungsidee darin besteht, einen *Strafterm* für die Koeffizienten der entsprechenden Basisfunktionen einzuführen. Ein solcher Strafterm lässt sich beispielsweise durch die quadrierten Koeffizienten

$$\sum_{j=l+2}^{d} \gamma_j^2$$

konstruieren, so dass betragsmäßig große Koeffizienten der abgeschnittenen Potenzen bestraft werden. Anstelle der üblichen Residuenquadratsumme

$$\mathrm{KQ} = \sum_{i=1}^{n} (y_i - f(z_i))^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2$$

die den Basisfunktionenansätzen des letzten Abschnitts zugrunde lag, ist dann die penalisierte Residuenquadratsumme

$$\mathrm{PKQ}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2$$

zu minimieren.

Aufgabe des Strafterms ist es, zu rauen Funktionsschätzungen entgegen zu wirken und damit eine zu starke Anpassung an die Daten zu verhindern. Gesteuert wird die Stärke der Bestrafung durch den *Glättungsparameter* $\lambda \geq 0$. Für $\lambda \to 0$ bekommt der Strafterm nur ein sehr geringes Gewicht, so dass die penalisierte Residuenquadratsumme annähernd der üblichen Residuenquadratsumme entspricht und sich damit eine nahe am KQ-Schätzer gelegene Schätzung für γ ergibt. Gilt umgekehrt $\lambda \to \infty$, so wird die Schätzung vollständig durch den Strafterm dominiert und man erhält ein Polynom vom Grad l als Schätzung für $f(z)$. Durch Variation des Glättungsparameters kann man kontinuierlich zwischen diesen beiden Extremen auswählen.

Der große Vorteil des Penalisierungsansatzes liegt nun darin, dass die Glattheit der Schätzung nicht mehr über die Zahl und die Position der Knoten, sondern im Wesentlichen durch einen einzelnen Parameter, nämlich den Glättungsparameter gesteuert wird. Die Frage nach der Wahl geeigneter Werte für den Glättungsparameter werden wir noch genauer in Abschnitt 7.1.9 behandeln. Da eine große Zahl von Knoten verwendet wird, spielt die genaue Positionierung dieser Knoten nur noch eine untergeordnete Rolle. Häufig benutzt man der Einfachheit halber äquidistante Knoten, aber auch quantilbasierte Knoten finden in der Praxis Verwendung. Dass die mit Hilfe von Penalisierungsansätzen erhaltenen Funktionsschätzungen tatsächlich nur noch in geringem Umfang von der Knotenzahl abhängen, illustriert Abbildung 7.14. Für eine geringe Zahl von Knoten ist die Schätzung noch nicht flexibel genug. Wird jedoch eine ausreichende Zahl von Knoten verwendet, so ändern sich die Funktionsschätzungen nur noch geringfügig, wenn die Knotenzahl erhöht und der Glättungsparameter im Verhältnis dazu geeignet gewählt wird.

Abb. 7.14. *Einfluss der Knotenzahl auf die Schätzung von P-Splines.*

P-Splines basierend auf B-Splines

Verwendet man anstelle der TP-Basis eine Darstellung von $f(z)$ basierend auf B-Spline-Basisfunktionen, so ist eine geeignete Penalisierung weniger offensichtlich, da keine Zerlegung in einen parametrischen Anteil und Abweichungen von diesem parametrischen Anteil erfolgt. Um einen Strafterm für B-Splines zu konstruieren, müssen wir also anders vorgehen. Möchte man allgemein die Glattheit einer Funktion beurteilen, so bietet sich hierzu die Verwendung von (quadrierten) Ableitungen an, da diese als Maß für die Variabilität der Funktion angesehen werden können. Besonders häufig verwendet werden auf der zweiten Ableitung basierende Strafterme der Form

$$\lambda \int (f''(z))^2 dz, \tag{7.2}$$

da die zweite Ableitung ein Maß für die Krümmung einer Funktion definiert. Diese Form der Bestrafung liegt etwa der Implementation von penalisierten Splines im R-Paket `mgcv` zugrunde, vergleiche Wood (2006). Darüber hinaus kann dieser Strafterm auch in allgemeineren Ansätzen angewendet werden, vergleiche Abschnitt 7.1.3. Für B-Splines existiert jedoch auch eine einfache Approximation an die Ableitungen, die zur Konstruktion von Straftermen verwendet werden kann. Formel (7.1) zeigt, dass sich die erste Ableitung eines B-Splines in Abhängigkeit von den ersten Differenzen der zugehörigen Parameter schreiben lässt. Um eine im Sinne der ersten Ableitung glatte Funktion zu erhalten und zu große Werte der ersten Ableitung zu verhindern, liegt es nahe, Strafterme basierend auf genau diesen Differenzen einzuführen. Wir beschränken uns im Folgenden

auf äquidistante Knoten, da sich in diesem Fall besonders einfache Ausdrücke für den Differenzen-Strafterm ergeben.

Analog zu Differenzen erster Ordnung lassen sich auch Differenzen höherer Ordnung verwenden, um eine im Sinne höherer Ableitungen glatte Funktion zu erhalten. Dies führt zur penalisierten Residuenquadratsumme

$$\mathrm{PKQ}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=k+1}^{d} (\Delta^k \gamma_j)^2,$$

wobei Δ^k Differenzen k-ter Ordnung bezeichnet, die rekursiv definiert sind durch

$$\begin{aligned}
\Delta^1 \gamma_j &= \gamma_j - \gamma_{j-1}, \\
\Delta^2 \gamma_j &= \Delta^1 \Delta^1 \gamma_j = \Delta^1 \gamma_j - \Delta^1 \gamma_{j-1} = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}, \\
&\vdots \\
\Delta^k \gamma_j &= \Delta^{k-1} \gamma_j - \Delta^{k-1} \gamma_{j-1}.
\end{aligned}$$

Um eine genauere Vorstellung von der Wirkung des Differenzen-Strafterms zu bekommen, visualisieren wir ähnlich wie in Abbildung 7.13 die Schätzung eines penalisierten Splines für verschiedene Werte des Glättungsparameters. (Wie diese Schätzung genau funktioniert, werden wir anschließend genauer beschreiben.) Für zweite Differenzen (und kubische B-Splines) sind die resultierenden Ergebnisse in Abbildung 7.15 dargestellt. Typischerweise werden in Anwendungen ebenfalls Differenzen zweiter Ordnung verwendet, seltener auch Differenzen dritter oder erster Ordnung.

Wie aus der Schätzung für große Glättungsparameter ersichtlich ist, ergibt sich für $\lambda \to \infty$ eine annähernd lineare Schätzung der Funktion $f(z)$. Allgemein erhält man für große Glättungsparameter $\lambda \to \infty$ ein Polynom vom Grad $k-1$, falls der Grad des Splines mindestens ebenso groß ist wie die betrachtete Ordnung der Differenzen, d.h. für $l \geq k$. Für erste Differenzen lässt sich dies aus der Formel für die erste Ableitung eines B-Splines in (7.1) begründen: Die erste Ableitung ist genau dann gleich Null, wenn alle ersten Differenzen benachbarter Koeffizienten Null sind. Genau dies wird erreicht, wenn der Glättungsparameter dem Strafterm ein großes Gewicht in der Schätzung zuweist. Für größere Differenzenordnungen ergibt sich die obige Aussage aus den zugehörigen Ableitungen höherer Ordnung.

Im Vergleich zu penalisierten Splines basierend auf der TP-Basis gewinnt man die zusätzliche Freiheit, den Grad des unpenalisierten Polynoms, das sich für $\lambda \to \infty$ ergibt, und den Grad des Splines, der zur Modellierung von $f(z)$ selbst eingesetzt wird, separat zu wählen. Für die TP-Basis ist das unpenalisierte Polynom stets vom Grad l, da nur die Koeffizienten der trunkierten Potenzen bestraft werden. Bei der Verwendung von B-Splines ist es dagegen möglich, den Grad des unpenalisierten Polynoms durch die zur Penalisierung verwendete Differenzenordnung k unabhängig vom Grad l des zugrunde liegenden Splines zu wählen. Damit gewinnt man zusätzliche Flexibilität in der Modellierung und erreicht außerdem eine saubere Trennung zweier verschiedener Modelleigenschaften.

Abb. 7.15. *Einfluss des Glättungsparameters auf die Schätzung eines P-Splines basierend auf zweiten Differenzen.*

Penalisierte KQ-Schätzung

Zur Herleitung des penalisierten KQ-Schätzers (PKQ-Schätzers) ist es zunächst günstig, die Strafterme in Matrix-Notation zu schreiben. Für die TP-Basis lässt sich dies einfach erreichen durch

$$\lambda \sum_{j=l+2}^{d} \gamma_j^2 = \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

mit dem Vektor $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ und der *Strafmatrix*

$$\boldsymbol{K} = \operatorname{diag}(\underbrace{0,\ldots,0}_{(l+1)},\underbrace{1,\ldots,1}_{(m-2)}).$$

Für den auf Differenzen basierenden Strafterm für B-Splines drücken wir den Vektor der ersten Differenzen zunächst mit Hilfe einer Differenzenmatrix

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

der Dimension $(d-1) \times d$ aus und erhalten

$$\boldsymbol{D}_1\boldsymbol{\gamma} = \begin{pmatrix} \gamma_2 - \gamma_1 \\ \vdots \\ \gamma_d - \gamma_{d-1} \end{pmatrix}.$$

Rekursiv lassen sich auch höhere Differenzen mit Hilfe von Differenzenmatrizen der Form

$$\boldsymbol{D}_k = \boldsymbol{D}_1\boldsymbol{D}_{k-1}$$

schreiben. Für $k = 2$ erhält man so beispielsweise die $(d-2) \times d$ Differenzenmatrix

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

Damit ergibt sich für den Strafterm die Darstellung

$$\lambda \sum_{j=k+1}^{d} (\Delta^k \gamma_j)^2 = \lambda \boldsymbol{\gamma}' \boldsymbol{D}_k' \boldsymbol{D}_k \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}' \boldsymbol{K}_k \boldsymbol{\gamma}$$

mit den Strafmatrizen

$$\boldsymbol{K}_1 = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 1 & \end{pmatrix}, \qquad \boldsymbol{K}_2 = \begin{pmatrix} 1 & -2 & 1 & & & & & \\ -2 & 5 & -4 & 1 & & & & \\ 1 & -4 & 6 & -4 & 1 & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & 1 & -4 & 6 & -4 & 1 & \\ & & & 1 & -4 & 5 & -2 & \\ & & & & 1 & -2 & 1 & \end{pmatrix}$$

für erste beziehungsweise zweite Differenzen.

Wird der Strafterm durch das Integral der quadrierten zweiten Ableitung wie in (7.2) definiert, so erhält man ebenfalls einen quadratischen Strafterm:

$$\begin{aligned}\int (f''(z))^2 dz &= \int \left(\sum_{j=1}^{d} \gamma_j B_j''(z) \right)^2 dz \\ &= \int \left(\sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_i \gamma_j B_i''(z) B_j''(z) \right) dz \\ &= \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_i \gamma_j \int B_i''(z) B_j''(z) dz \\ &= \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}.\end{aligned}$$

Die Einträge der Strafmatrix $\boldsymbol{K}$ ergeben sich dabei aus den zweiten Ableitungen der B-Spline-Basisfunktionen.

Allgemein lässt sich die penalisierte Residuenquadratsumme also mit Hilfe einer Strafmatrix $\boldsymbol{K}$ geeigneten Typs darstellen als

$$\begin{aligned}\mathrm{PKQ}(\lambda) &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{Z}\boldsymbol{\gamma} - \boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{y} + \boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma} + \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{y} + \boldsymbol{\gamma}'(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})\boldsymbol{\gamma}.\end{aligned}$$

Differentiation nach $\boldsymbol{\gamma}$ (vergleiche Satz A.32 in Anhang A) und Nullsetzen des sich ergebenden Ausdrucks liefert das Gleichungssystem

$$-2\boldsymbol{Z}'\boldsymbol{y} + 2(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})\boldsymbol{\gamma} = 0,$$

so dass man den PKQ-Schätzer

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1} \boldsymbol{Z}'\boldsymbol{y}$$

erhält. Der Vektor der beobachteten Funktionswerte $\hat{\boldsymbol{f}} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))'$ berechnet sich damit als

$$\hat{\boldsymbol{f}} = \boldsymbol{Z}\hat{\boldsymbol{\gamma}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1} \boldsymbol{Z}'\boldsymbol{y}.$$

Man beachte, dass der PKQ-Schätzer eine dem unpenalisierten KQ-Schätzer sehr ähnliche Gestalt besitzt und sich lediglich durch den zusätzlichen Term $\lambda \boldsymbol{K}$ von diesem unterscheidet. Dieser zusätzliche Term repräsentiert den Einfluss des Strafterms im PKQ-Kriterium. Für $\lambda = 0$ erhält man wieder die Lösung des unpenalisierten Optimierungsproblems. Als Kovarianzmatrix des PKQ-Schätzers ergibt sich durch ähnliche Überlegungen wie für den KQ-Schätzer in Abschnitt 3.2.3

$$\mathrm{Cov}(\hat{\boldsymbol{\gamma}}) = \sigma^2 (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1} \boldsymbol{Z}'\boldsymbol{Z} (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1}.$$

Diese Kovarianzmatrix besteht im Gegensatz zur Kovarianzmatrix des KQ-Schätzer aus drei Faktoren und besitzt die (aus Abschnitt 6.3.3 bekannte) Form einer „Sandwich“-Matrix. Die zusätzlichen Ausdrücke, die die Strafmatrix enthalten, sind dabei für die Stabilisierung des PKQ-Schätzers verantwortlich und führen im Vergleich mit dem unpenalisierten KQ-Schätzer zu geringeren Varianzen. Der Preis für diese gewonnene Effizienz ist allerdings, dass der PKQ-Schätzer im Allgemeinen nicht mehr erwartungstreu ist. In der Regel sollte der PKQ-Schätzer aber bei geeigneter Glättungsparameterwahl

Penalisierte Splines

Modell

Die Funktion f wird durch einen Polynom-Spline approximiert, so dass sich das nichtparametrische Regressionsmodell als lineares Modell

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

schreiben lässt.

Penalisiertes KQ-Kriterium

Anstelle der üblichen Residuenquadratsumme wird das penalisierte KQ-Kriterium

$$\mathrm{PKQ}(\lambda) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$$

zur Schätzung von $\boldsymbol{\gamma}$ minimiert. Der Glättungsparameter $\lambda \geq 0$ steuert den Kompromiss zwischen Datentreue und Glattheit der resultierenden Funktionsschätzung. Für Splines in TP-Basis-Darstellung werden die quadrierten Koeffizienten der abgeschnittenen Potenzen penalisiert. Für B-Splines wird der Strafterm durch Differenzen benachbarter Koeffizienten oder das Integral der quadrierten zweiten Ableitung der Funktion konstruiert.

Penalisierter KQ-Schätzer

Der penalisierte KQ-Schätzer ergibt sich in allen Fällen als

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

einen geringeren MSE aufweisen, so dass er aus statistischer Sicht insgesamt günstigere Eigenschaften besitzt.

Vergleicht man die Form des penalisierten KQ-Schätzers mit dem in Abschnitt 3.5 (Seite 147ff.) hergeleiteten Posteriori-Erwartungswert

$$\mathrm{E}(\boldsymbol{\gamma}|\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

eines linearen Modells mit Priori-Verteilung $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{M})$, so besitzen beide im Wesentlichen die gleiche Form. Dementsprechend lassen sich penalisierte Splines auch Bayesianisch interpretieren, wie wir im folgenden Abschnitt noch ausführlicher diskutieren werden. Andererseits besitzen penalisierte Splines auch eine enge Verbindung zu Modellen mit zufälligen Effekten, wie ein Vergleich mit der in (6.29) auf Seite 265 angegebenen Schätzgleichung zeigt. Diese beiden Zusammenhänge werden wir in Abschnitt 7.1.9 verwenden, um Schätzverfahren für den Glättungsparameter λ herzuleiten.

Beispiel 7.2 Unterernährung in Tansania: Modellierung durch P-Splines

Ähnlich wie in Beispiel 7.1 wollen wir anhand des Datensatzes zur Unterernährung in Tansania verschiedene Spezifikationen für penalisierte Splines vergleichen. Wir konzentrieren uns dabei auf P-Splines in B-Spline Darstellung und untersuchen verschiedene

Abb. 7.16. *P-Spline-Schätzung für den Tansania-Datensatz basierend auf Penalisierung mit ersten (links) und zweiten Differenzen (rechts) sowie verschiedenen Glättungsparametern.*

Glättungsparameterwahlen sowie Penalisierungsterme basierend auf ersten und zweiten Differenzen. Abbildung 7.16 zeigt einige der zugehörigen Ergebnisse. In der ersten Zeile sieht man die geschätzten Funktionen $\hat{f}(z)$, die sich mit einem kleinen Wert für den Glättungsparameter ergeben. Konsequenterweise resultieren relativ raue Schätzungen, wobei die Ergebnisse für zweite Differenzen tendenziell etwas glatter ausfallen. Wählt man einen großen Glättungsparameter (zweite Zeile in Abbildung 7.16), so ergeben sich sehr glatte Schätzungen. Insbesondere erhält man für erste Differenzen bereits nahezu den Grenzfall einer konstanten Schätzung, die sich für $\lambda \to \infty$ ergibt. Für zweite Differenzen ist dagegen noch eine deutliche Abweichung von der linearen Form zu erkennen. Die letzte Zeile der Grafik zeigt zusätzlich Schätzungen, die sich mit einem 'optimalen'

Glättungsparameter ergeben, der mit Hilfe eines der in Abschnitt 7.1.9 behandelten Kriterien gewählt wurde. Hier ergibt sich ein (zumindest visuell) plausibler Kompromiss zwischen Datentreue und Glattheit. Die zwei bereits zuvor angesprochenen wesentlichen Eigenschaften des Zusammenhangs zwischen Alter und Unterernährung sind für beide Strafterme gut zu erkennen: Ein höherer Z-Score, d.h. ein niedrigeres Risiko für Unterernährung junger Kinder, und ein durch den Wechsel der Referenzpopulation hervorgerufenes, lokales Maximum des Z-Scores in einem Alter von etwa 24 Monaten. Ohne automatische, datengesteuerte Wahl des Glättungsparameters wäre es insbesondere schwierig, den Effekt des Wechsels der Referenzpopulation zu entdecken.

△

Bayesianische P-Splines

Als Alternative zur Betrachtung penalisierter KQ-Kriterien lassen sich P-Splines auch Bayesianisch motivieren. Damit können grundsätzlich auch Bayesianische Verfahren zur Schätzung von P-Splines verwendet werden, vergleiche Abschnitt 7.1.9. Wir werden uns zunächst in unseren Betrachtungen auf penalisierte Splines basierend auf B-Splines konzentrieren und später kurz auf TP-Splines eingehen.

Der Bayesianischen Formulierung soll wieder das gleiche Beobachtungsmodell wie im vorigen Abschnitt zugrunde liegen, d.h. es gelte

$$y_i = \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i$$

mit B-Spline-Basisfunktionen B_j. Anstelle eines Penalisierungsansatzes soll nun eine geeignete Priori-Annahme für den Vektor $\boldsymbol{\gamma}$ konstruiert werden, die eine glatte Schätzung gewährleistet. Hierzu bietet sich das stochastische Analogon zur Differenzenbestrafung an, nämlich *Random Walks* (Irrfahrten) der Ordnung k (RWk). Ein Random Walk erster Ordnung (RW1) ist definiert durch

$$\gamma_j = \gamma_{j-1} + u_j, \quad u_j \sim \mathrm{N}(0, \tau^2), \quad j = 2, \ldots, d.$$

Äquivalent lässt sich diese Annahme auch als

$$\gamma_j - \gamma_{j-1} = u_j, \quad u_j \sim \mathrm{N}(0, \tau^2), \quad j = 2, \ldots, d,$$

schreiben, so dass bereits eine gewisse Nähe zu den zuvor konstruierten Differenzenstraftermen erkennbar ist. Über den Startwert γ_1 müssen zusätzliche Vereinbarungen getroffen werden. In der Regel verwendet man eine nichtinformative Priori-Verteilung $p(\gamma_1) \propto const.$

Betrachtet man die durch einen RW1 definierten bedingten Verteilungen, so erhält man

$$\gamma_j | \gamma_{j-1}, \ldots, \gamma_1 \sim \mathrm{N}(\gamma_{j-1}, \tau^2). \tag{7.3}$$

Offenbar unterstellt der RW1 also eine spezielle Abhängigkeitsstruktur, so dass die bedingte Verteilung von γ_j gegeben alle vorherigen Werte nur von dem um eins verzögerten Wert γ_{j-1} abhängt. Allgemein spricht man in diesem Zusammenhang auch davon, dass der RW1 die *Markov-Eigenschaft* besitzt.

Betrachten wir nun genauer die durch den Random Walk unterstellte Abhängigkeitsstruktur des Parametervektors $\boldsymbol{\gamma}$. Aus Gleichung (7.3) folgt, dass der bedingte Erwartungswert

Abb. 7.17. *Durch einen Random Walk erster Ordnung definierte bedingte Verteilung bei großer (links) und kleiner Varianz (rechts).*

von γ_j gerade der um eins verzögerte Wert γ_{j-1} ist, so dass sich für den Erwartungswert ein konstanter Trend ergibt. Dies ist in Abbildung 7.17 illustriert, die gleichzeitig den Einfluss der Varianz τ^2 des Fehlerterms u_j veranschaulicht. Je größer die Varianz ist, desto größer ist auch die zugelassene Streuung um den bedingten Erwartungswert. Wie wir aus der Diskussion von P-Splines im vorigen Abschnitt wissen, führt ein konstanter Wert aller B-Spline Koeffizienten zu einer konstanten Schätzung für die Funktion $f(z)$. Dies entspricht nun genau dem Fall, dass die Varianz des RW1 annähernd Null ist, da in diesem Fall nur sehr geringe Abweichungen zwischen γ_j und γ_{j-1} zugelassen sind und man einen nahezu konstanten Verlauf der Sequenz $\gamma_1, \ldots, \gamma_d$ erhält. Im umgekehrten Fall einer großen Varianz τ^2 können benachbarte Koeffizienten sehr stark voneinander abweichen, so dass sich eine raue Funktionsschätzung ergibt. Dementsprechend lässt sich der Varianzparameter τ^2 in gewisser Weise wie ein inverser Glättungsparameter interpretieren. Wie sich der Zusammenhang zwischen λ und τ^2 genau darstellt, werden wir im Folgenden noch genauer darstellen.

Durch die Formulierung eines Random Walks in (7.3) werden zwar zunächst nur die bedingten Verteilungen von $\gamma_j|\gamma_{j-1}, \ldots, \gamma_1$ festgelegt, aus diesen bedingten Verteilungen lässt sich aber die gemeinsame, multivariate Priori-Verteilung des vollständigen Vektors $\boldsymbol{\gamma}$ unter Annahme eines RW1 bestimmen:

$$
\begin{aligned}
p(\boldsymbol{\gamma}|\tau^2) &= \prod_{j=1}^{d} p(\gamma_j|\gamma_{j-1}, \ldots, \gamma_1) \\
&= p(\gamma_1) \prod_{j=2}^{d} p(\gamma_j|\gamma_{j-1}) \\
&\propto \prod_{j=2}^{d} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\gamma_j - \gamma_{j-1})^2\right) \\
&= \frac{1}{(2\pi\tau^2)^{(d-1)/2}} \exp\left(-\frac{1}{2\tau^2}\sum_{j=2}^{d}(\gamma_j - \gamma_{j-1})^2\right) \\
&= \frac{1}{(2\pi\tau^2)^{(d-1)/2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}_1\boldsymbol{\gamma}\right).
\end{aligned}
$$

Der erste Schritt basiert dabei auf der Anwendung des Gesetzes der totalen Wahrscheinlichkeit zur Faktorisierung der gemeinsamen Dichte. Aufgrund der Markov-Eigenschaft

ergeben sich dann vereinfachte Ausdrücke für die bedingten Verteilungen, so dass sich die Dichten der durch die RW1-Priori definierten Normalverteilungsannahmen einsetzen lassen. Das Proportionalitätszeichen ergibt sich aus der flachen Priori für den ersten Regressionskoeffizienten γ_1, die nicht exakt, sondern nur bis auf Proportionalität spezifiziert ist. Durch weitere Umformungen zeigt man dann, dass die gemeinsame Verteilung eine multivariate Normalverteilung mit Erwartungswert $\mathbf{0}$ und Präzisionsmatrix $1/\tau^2\boldsymbol{K}_1$ ist, wobei $\boldsymbol{K}_1 = \boldsymbol{D}_1'\boldsymbol{D}_1$ genau wie im vorigen Abschnitt durch die Differenzenmatrix erster Ordnung $\boldsymbol{D}_1$ definiert ist. Allerdings besitzt diese Präzisionsmatrix nicht vollen Rang, da sich sowohl Zeilen als auch Spalten zu Null addieren. Dementsprechend lässt sich nicht die Inverse bilden, die Kovarianzmatrix „$\tau^2\boldsymbol{K}_1^{-1}$“ der Priori-Verteilung existiert also nicht. Die Verteilung von $\boldsymbol{\gamma}$ ist damit nicht normierbar bzw. uneigentlich, d.h. das Integral über die Dichte divergiert und es handelt sich bei der Verteilung um eine singuläre Normalverteilung (vergleiche Anhang B.3.2). Genauer spricht man in diesem Fall von einer teilweise uneigentlichen Verteilung, da die Präzisionsmatrix zwar nicht vollen Rang besitzt, aber $\mathrm{rg}(\boldsymbol{K}_1) = d - 1 > 0$ gilt. Für einen Random Walk erster Ordnung entspricht der in Anhang B.3.2 definierte, deterministische Anteil der Priori-Verteilung einem konstanten Vektor, der das Niveau der Funktion f repräsentiert. Für dieses Niveau wird durch die teilweise uneigentliche Priori implizit eine flache Priori definiert. Der stochastische Anteil repräsentiert dann die Abweichungen von der Konstanten. Diese Aufteilung werden wir in Abschnitt 7.1.9 verwenden, um eine Repräsentation von Penalisierungsansätzen durch gemischte Modelle herzuleiten.

Aus der multivariaten Priori-Verteilung des Vektors $\boldsymbol{\gamma}$ lassen sich auch ungerichtete Formen der bedingten Verteilung in (7.3) ablesen. In diesem Fall bedingt man also auf alle übrigen Parameter und nicht nur auf die „Vergangenheitsinformation“ $\gamma_{j-1}, \ldots, \gamma_1$. Für den RW1 ergibt sich (mit Ausnahme der Randparameter γ_1 und γ_d)

$$\gamma_j \sim \mathrm{N}\left(\frac{1}{2}(\gamma_{j-1} + \gamma_{j+1}), \frac{\tau^2}{2}\right), \quad j = 2, \ldots, d-1.$$

Man erkennt wieder die Markov-Eigenschaft (diesmal allerdings in ungerichteter Form), nach der die Verteilung von γ_j nur von den beiden unmittelbaren Nachbarn abhängt. Genauer ergibt sich (mit Ausnahme der Ränder), dass der bedingte Erwartungswert der lokale Mittelwert der beiden benachbarten Regressionskoeffizienten ist. Eine ausführlichere Diskussion der Random Walk-Prioris, die auch Aussagen für die Randparameter beinhaltet, findet man in Lang (2004).

Obwohl die gemeinsame Priori-Verteilung teilweise uneigentlich ist, ergibt sich eine eigentliche Posteriori-Verteilung für $\boldsymbol{\gamma}$, deren Form sich aus den Überlegungen in Abschnitt 3.5 ergibt. Insbesondere erhält man als Posteriori-Erwartungswert und Posteriori-Kovarianzmatrix für $\boldsymbol{\gamma}$ die Ausdrücke

$$\mathrm{E}(\boldsymbol{\gamma}|\boldsymbol{y}, \sigma^2, \tau^2) = \left(\boldsymbol{Z}'\boldsymbol{Z} + \frac{\sigma^2}{\tau^2}\boldsymbol{K}_1\right)^{-1}\boldsymbol{Z}'\boldsymbol{y},$$

(also den penalisierten KQ-Schätzer mit Glättungsparameter $\lambda = \sigma^2/\tau^2$) und

$$\mathrm{Cov}(\boldsymbol{\gamma}|\boldsymbol{y}, \sigma^2, \tau^2) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K}_1)^{-1}.$$

Im Gegensatz zu den nicht-Bayesianischen Betrachtungen im letzten Abschnitt ergibt sich hier also eine einfachere Form für die Kovarianzmatrix. Der Grund hierfür liegt in der

Abb. 7.18. *Durch einen Random Walk zweiter Ordnung unterstellte bedingte Verteilung.*

stochastischen Formulierung des Bayesianischen Modells: Die Regressionskoeffizienten $\boldsymbol{\gamma}$ werden als zufällig betrachtet, während sie im frequentistischen Ansatz feste, unbekannte Parameter bezeichnen.

Der Glättungsparameter λ wird in der Bayesianischen Formulierung von P-Splines bestimmt durch das Verhältnis der Varianz des Fehlerterms und der Varianz des RW1. Dies konkretisiert unsere früheren Überlegungen zum Einfluss der Varianz τ^2 des Random Walks und führt gleichzeitig zu einer interessanten Interpretation des Glättungsparameters λ: Je größer die Varianz der Priori-Verteilung im Verhältnis zur Residualstreuung ist, desto weniger stark wird die Schätzung penalisiert. Dementsprechend ist der konkrete Wert von τ^2 immer relativ zur Varianz σ^2 des Messfehlers $\boldsymbol{\varepsilon}$ zu interpretieren und λ lässt sich als Rauschen/Signal-Verhältnis auffassen.

Für Random Walks höherer Ordnung lassen sich nun analoge Überlegungen anstellen. Wir werden diese am Beispiel des RW2 noch einmal kurz diskutieren. Eine RW2-Priori für $\boldsymbol{\gamma}$ wird definiert durch

$$\gamma_j = 2\gamma_{j-1} - \gamma_{j-2} + u_j, \quad u_j \sim \mathrm{N}(0, \tau^2), \quad j = 3, \ldots, d,$$

wobei man in der Regel wieder nichtinformative Priori-Verteilungen $p(\gamma_1) \propto \mathit{const}$ und $p(\gamma_2) \propto \mathit{const}$ für die Startwerte annimmt. Wieder erkennt man die Ähnlichkeit zu den auf Differenzen beruhenden Straftermen des vorigen Abschnitts, wenn man die RW2-Definition umschreibt zu

$$\gamma_j - 2\gamma_{j-1} + \gamma_{j-2} = u_j, \quad u_j \sim \mathrm{N}(0, \tau^2), \quad j = 3, \ldots, d.$$

Spezifiziert man die Annahme eines RW2 über die bedingten Verteilungen, so ergibt sich

$$\gamma_j | \gamma_{j-1}, \gamma_{j-2}, \ldots, \gamma_1 \sim \mathrm{N}(2\gamma_{j-1} - \gamma_{j-2}, \tau^2),$$

was einer lokal linearen Extrapolation für den bedingten Erwartungswert entspricht (vergleiche Abbildung 7.18). Für die gemeinsame Verteilung des gesamten Vektors $\boldsymbol{\gamma}$ erhält man

$$p(\boldsymbol{\gamma}|\tau^2) \propto \frac{1}{(2\pi\tau^2)^{(d-2)/2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}_2\boldsymbol{\gamma}\right).$$

Auch hier besitzt die Präzisionsmatrix nicht vollen Rang, so dass die Priori-Verteilung für einen RW2 teilweise uneigentlich ist mit $\mathrm{rg}(\boldsymbol{K}_2) = d - 2$.

Obwohl man für Bayesianische P-Splines bei gegebenem Glättungsparameter analytisch Punktschätzer und Kovarianzmatrizen erhalten kann, werden wir die Bayesianische Formulierung im Wesentlichen dazu verwenden, in Abschnitt 7.1.9 ein Bayesianisches Inferenzkonzept für P-Splines (und allgemeinere Penalisierungsansätze) herzuleiten. Insbesondere ergibt sich hieraus eine einfache Möglichkeit, den Glättungsparameter mitzubestimmen. Darüber hinaus lassen sich im Rahmen von Bayes-Ansätzen die in Abschnitt 8 behandelten Erweiterungen für mehrere Einflussgrößen leichter umsetzen.

P-Splines in TP-Basis Darstellung lassen sich ebenfalls Bayesianisch motivieren. Hier lässt sich die Priori-Verteilung sogar noch einfacher konstruieren, da sich als Äquivalenz zur Ridge-Bestrafung des letzten Abschnitts eine i.i.d. Priori für die Regressionskoeffizienten anbietet. Genauer nimmt man an, dass die Koeffizienten $\gamma_{l+2}, \ldots, \gamma_d$ der abgeschnittenen Polynome unabhängig und identisch $\mathrm{N}(0, \tau^2)$ verteilt sind. Für die Koeffizienten der globalen Polynome $\gamma_1, \ldots, \gamma_{l+1}$ wird dagegen eine nichtinformative Priori-Verteilung zugrunde gelegt, d.h. $p(\gamma_j) \propto const$, $j = 1, \ldots, l+1$. Die beiden verschiedenen Typen von Priori-Verteilungen spiegeln ebenso wie im frequentistischen Penalisierungsansatz des letzten Abschnitts eine Unterscheidung in unrestringiert und restringiert zu schätzende Parameter wieder. Die Bayesianische Formulierung von TP-Splines stellt auch bereits einen Zusammenhang zur Formulierung von nichtparametrischen Regressionsmodellen als gemischte Modelle her, da neben der Bayesianischen Interpretation als Priori-Verteilungen die Verteilungsannahmen für die Regressionskoeffizienten auch als Unterscheidung in feste und zufällige Effekte betrachtet werden können. Dieser Zusammenhang wird uns ebenfalls zur Konstruktion eines geeigneten Schätzverfahrens für den Glättungsparameter verhelfen.

Wie für P-Splines basierend auf B-Splines lassen sich nun völlig analog für TP-Splines Posteriori-Erwartungswert und Kovarianzmatrix herleiten. Insbesondere erhält man als Posteriori-Erwartungswert den penalisierten KQ-Schätzer.

7.1.3 Allgemeine Penalisierungsansätze

Bevor wir uns weiteren Modellen der nichtparametrischen Regression zuwenden, wollen wir an dieser Stelle eine Reihe von Eigenschaften penalisierter Splines zusammenstellen und diese auf allgemeine Penalisierungsansätze generalisieren. Wir werden im Folgenden feststellen, dass sich viele der in Abschnitt 7.1 und 7.2 behandelten univariaten und bivariaten Glättungsverfahren tatsächlich in diesen allgemeinen Rahmen einbetten lassen.

Die Darstellung polynomialer Splines mit Hilfe von Basisfunktionen

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z) \tag{7.4}$$

ergab ein großes lineares Modell der Form

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{7.5}$$

in dem die Designmatrix durch die Basisfunktionen definiert wurde. In (7.4) ist man aber natürlich keineswegs auf TP-Splines oder B-Splines beschränkt, sondern kann praktisch beliebige Funktionen zur Approximation von $f(z)$ heranziehen. Durch die Vorgabe der Basisfunktionen B_j wird festgelegt, welche Funktionenklasse zur Beschreibung von

$f(z)$ zur Verfügung steht. Beispielsweise ergibt sich durch die Vorgabe der TP- oder der B-Spline-Basis genau der Raum aller möglichen Polynom-Splines (zur zugehörigen Knotenmenge und zum Grad l). In den folgenden Abschnitten werden wir aber auch andere Modelle der Form (7.5) kennenlernen, die in der nichtparametrischen Regression von Nutzen sind.

Auch das zur Herleitung von P-Splines verwendete Konzept penalisierter Residuenquadratsummen lässt sich allgemeiner anwenden. Geht man von der Darstellung (7.4) aus, so lässt sich im Allgemeinen ein geeigneter Strafterm über das Integral der quadrierten zweiten Ableitung definieren:

$$\int (f''(z))^2 dz = \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_i \gamma_j \int B_i''(z) B_j''(z) dz = \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}.$$

Damit ergibt sich wieder ein in $\boldsymbol{\gamma}$ quadratischer Strafterm mit Strafmatrix $\boldsymbol{K}$, deren Einträge sich aus den Ableitungen der Basisfunktionen bestimmen lassen. In einigen Basisfunktionenansätzen existieren alternative Konstruktionsmechanismen für Strafterme, wie wir sie für P-Splines kennengelernt haben. Diese führen jedoch häufig auch zu quadratischen Straftermen $\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$.

Als Optimalitätskriterium ergibt sich jeweils ein penalisiertes Kleinste-Quadrate-Kriterium der Form

$$(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma},$$

das sich für verschiedene Penalisierungsansätze lediglich durch die Form der Designmatrix $\boldsymbol{Z}$ und der Strafmatrix $\boldsymbol{K}$ unterscheidet. Als üblichen Schätzer erhält man damit den penalisierten KQ-Schätzer

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1} \boldsymbol{Z}' \boldsymbol{y}.$$

Äquivalent dazu lassen sich Penalisierungsansätze mit quadratischem Strafterm auch Bayesianisch betrachten, da Penalisierungen der Form $\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$ der Annahme einer multivariaten Normalverteilung mit Dichte

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{2\pi\tau^2}\right)^{\mathrm{rg}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}\right)$$

als Priori-Verteilung für den Koeffizientenvektor entsprechen. Dabei ist zu beachten, dass die Strafmatrix $\boldsymbol{K}$ häufig nicht vollen Rang besitzt, die Dichte in der Regel also nicht normierbar und damit nur bis auf Proportionalität definierbar ist.

In Abschnitt 7.1.9 werden zwei der betrachteten Verfahren zur Bestimmung optimaler Glättungsparameter explizit auf Modelle der in diesem Abschnitt diskutierten Form allgemeiner Penalisierungsansätze zugeschnitten sein.

An dieser Stelle wollen wir außerdem kurz auf die Möglichkeit der Modellierung nicht normalverteilter Zielvariablen im Rahmen der nichtparametrischen Regression eingehen, die uns in Kapitel 8 noch ausführlicher beschäftigen wird. Überträgt man die Ideen der nichtparametrischen Regression auf Zielvariablen aus Exponentialfamilien wie in Kapitel 4, so ergeben sich Modelle der Form

$$\mathrm{E}(y_i|\eta_i) = h(\eta_i), \qquad \eta_i = f(z_i)$$

Allgemeine Penalisierungsansätze

Modell:

Die Funktion f wird dargestellt mit Hilfe eines großen linearen Modells, so dass sich die Modellgleichung

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

ergibt. Häufig geschieht dies durch die Repräsentation von $f(z)$ mit Hilfe von Basisfunktionen:

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z).$$

Penalisierung:

Zur Regularisierung wird ein quadratischer Strafterm der Form

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

bzw. die Normalverteilungs-Priori

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{2\pi\tau^2}\right)^{\mathrm{rg}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right)$$

für $\boldsymbol{\gamma}$ angenommen. Dies führt zum PKQ-Kriterium

$$(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

und dem PKQ-Schätzer

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1} \boldsymbol{Z}'\boldsymbol{y}.$$

für den Erwartungswert der Zielvariablen. Dabei bezeichnet h wieder eine bekannte Responsefunktion, etwa die logistische Verteilungsfunktion im Logit-Modell oder die Exponentialfunktion im log-linearen Poisson-Modell. Setzt man Basisfunktionen zur Modellierung der Funktion f ein, so ergibt sich für den Vektor der Prädiktoren das lineare Modell $\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\gamma}$. Da nun die gleichen Probleme wie in der nichtparametrischen Regression mit normalverteilter Zielvariable auftreten können, wird analog zur penalisierten Residuenquadratsumme das penalisierte Log-Likelihood-Kriterium

$$l_{pen}(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{\lambda}{2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma} \tag{7.6}$$

zur Schätzung der Regressionskoeffizienten verwendet. Dabei bezeichnet $l(\boldsymbol{\gamma})$ die übliche Log-Likelihood eines generalisierten linearen Modells, vergleiche Abschnitt 4.4.1. Tatsächlich ist (7.6) für normalverteilte Zielvariablen äquivalent zu den bisher betrachteten PKQ-Kriterien.

Um den penalisierten ML-Schätzer zu bestimmen, wird das in Abschnitt 4.4.2 behandelte Fisher-Scoring-Verfahren geeignet um den Strafterm ergänzt. Insbesondere kann $\hat{\gamma}$ durch iterativ gewichtete penalisierte KQ-Schätzung gewonnen werden. Wir wollen darauf jedoch an dieser Stelle nicht genauer eingehen und verweisen hierfür auf Kapitel 8. Um die Diskussion der weiteren Modellierungsmöglichkeiten der nichtparametrischen Regression in den Abschnitten 7.1 und 7.2 möglichst einfach zu halten, werden wir uns außerdem im Weiteren auf normalverteilte Zielvariablen beschränken und erst in Kapitel 8 auf den allgemeineren Fall zurückkommen.

7.1.4 Glättungssplines

Zur Herleitung von penalisierten Splines haben wir uns von vornherein auf den durch die Basisfunktionen gebildeten Funktionenraum eingeschränkt. Man kann jedoch auch mit allgemeineren Funktionenklassen arbeiten, wie wir nun am Beispiel von *Glättungssplines* (*Smoothing Splines*) demonstrieren wollen. Hier wird nur vorausgesetzt, dass die nichtparametrisch zu schätzende Funktion $f(z)$ zweimal stetig differenzierbar ist, so dass man das penalisierte Kleinste-Quadrate-Kriterium

$$\sum_{i=1}^{n}(y_i - f(z_i))^2 + \lambda \int (f''(z))^2 dz \tag{7.7}$$

als zu minimierendes Optimalitätskriterium verwenden kann. Der Penalisierungsterm wird hier (wie im vorigen Abschnitt zur Penalisierung in allgemeinen Basisfunktionenansätzen) durch die quadrierte zweite Ableitung als Maß für die Krümmung einer Funktion definiert. Überraschenderweise stellt sich heraus, dass sich trotz der Allgemeinheit des gewählten Ansatzes wieder eine sehr spezielle Funktionenklasse als Lösung des Optimierungsproblems ergibt. Diese Funktionenklasse wird gebildet durch die sogenannten *natürlichen kubischen Splines*, die eine Teilmenge spezieller kubischer Polynom-Splines darstellen.

Natürliche kubische Splines

Zu einer vorgegebenen Knotenmenge $a < \kappa_1 < \ldots < \kappa_m < b$ ist die Funktion $f(z)$ genau dann ein natürlicher kubischer Spline, wenn gilt

(i) $f(z)$ ist ein kubischer Polynom-Spline zur obigen Knotenmenge und
(ii) $f(z)$ erfüllt die Randbedingungen $f''(a) = f''(b) = 0$, d.h. $f(z)$ ist linear in den Intervallen $[a, \kappa_1]$ und $[\kappa_m, b]$.

Offensichtlich ist also jeder natürliche kubische Spline auch ein kubischer Polynom-Spline. Durch die Randbedingungen erhält man aber zusätzliche Restriktionen, so dass sich ein natürlicher kubischer Spline durch m Basisfunktionen darstellen lässt (im Vergleich zu $m + 2$ Basisfunktionen für gewöhnliche kubische Polynom-Splines). Die beiden Randbedingungen sind es auch, die zur Namensgebung „natürlicher kubischer Splines“ geführt haben. Das Wort Spline bezeichnet im Englischen ein Kurvenlineal, also ein biegsames Lineal, dass es erlaubt, eine glatte Funktion durch eine vorgegebene Menge von Punkten zu zeichnen. In unserem Beispiel wären dies die Funktionswerte an den Knoten der Knotenmenge. Außerhalb der Knotenmenge nimmt das Lineal dann wieder seine „natürliche“, ungekrümmte Lage ein.

Abb. 7.19. *Basis eines natürlichen kubischen Splines zu äquidistanten Knoten (links) und ungleichmäßig verteilten Knoten (rechts).*

Als Lösung des Optimierungsproblems in (7.7) ergibt sich nun gerade ein natürlicher kubischer Spline, dessen Knotenmenge durch die d verschiedenen, geordneten Kovariablenausprägungen $z_{(1)} < \ldots < z_{(d)}$ gebildet wird. Ein solcher Spline wird dann verkürzt auch als *Glättungsspline* oder *Smoothing Spline* bezeichnet. Häufig geht man bei der Herleitung von Glättungssplines direkt davon aus, dass die Kovariablenausprägungen alle voneinander verschieden und geordnet sind, so dass sich $z_{(1)} = z_1, \ldots, z_{(d)} = z_n$ ergibt. Bevor wir hierauf genauer eingehen, wollen wir aber zunächst noch natürliche kubische Splines zu einer allgemeinen Knotenmenge detaillierter betrachten.

Da natürliche kubische Splines spezielle Polynom-Splines sind, lassen sie sich dementsprechend auch wieder mit Hilfe der dort behandelten Basisfunktionen darstellen. Dabei ist jedoch zu beachten, dass zur Beschreibung eines natürlichen kubischen Splines d Basisfunktionen ausreichen, so dass die Darstellung durch die üblichen kubischen Polynom-Spline-Basisfunktionen zwei überflüssige Basisfunktionen enthält. Eine „echte“ Basis für natürliche kubische Splines erhält man durch geeignete Modifikationen der Polynom-Spline-Basisfunktionen am Rand des Definitionsbereichs. In der Darstellung

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z)$$

bleiben die Basisfunktionen $B_3, \ldots, B_{d-2}$ unverändert gegenüber der üblichen Darstellung, während die Basisfunktionen B_1, B_2, B_{d-1} und B_d so verändert werden müssen, dass der resultierende Spline die natürlichen Randbedingungen erfüllt. Wir verzichten an dieser Stelle auf eine explizite Herleitung der resultierenden Basisfunktionen und visualisieren diese stattdessen in Abbildung 7.19 für die gleichen Knotenmengen, die in Abbildung 7.12 (Seite 304) der Konstruktion von B-Splines zugrunde gelegt wurden. Offensichtlich bleiben die Basisfunktionen im Inneren des Wertebereichs tatsächlich unverändert, während die äußeren Basisfunktionen gerade so modifiziert werden, dass sie die Randbedingungen erfüllen, die an den natürlichen kubischen Spline gestellt werden.

Zur Minimierung des penalisierten KQ-Kriteriums (7.7) benötigen wir noch einen Ausdruck für den durch die quadrierte zweite Ableitung definierten Penalisierungsterm. Diesen erhält man aus der Darstellung des natürlichen kubischen Splines mit Hilfe von Basisfunktionen, wie im vorigen Abschnitt diskutiert, als

$$\int (f''(z))^2 dz = \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_i \gamma_j \int B_i''(z) B_j''(z) dz = \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}.$$

Glättungssplines

Gesucht wird die Funktion f, die das penalisierte KQ-Kriterium

$$\sum_{i=1}^{n}(y_i - f(z_i))^2 + \lambda \int (f''(z))^2 dz$$

unter allen zweimal stetig differenzierbaren Funktionen minimiert. Lösung des Minimierungsproblems ist ein natürlicher kubischer Spline mit Knoten an den beobachteten Kovariablenausprägungen. Natürliche kubische Splines lassen sich mit Hilfe einer modifizierten B-Spline-Basis darstellen, so dass der PKQ-Schätzer wie für P-Splines berechnet werden kann durch

$$\hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

Damit ergibt sich das zu minimierende penalisierte KQ-Kriterium zu

$$(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma},$$

wobei $\boldsymbol{Z}$ die an den beobachteten z-Werten ausgewerteten Basisfunktionen enthält. Die Form des zu minimierenden Ausdrucks entspricht also dem allgemeinen, im vorigen Abschnitt betrachteten PKQ-Kriterium, so dass sich der PKQ-Schätzer unmittelbar ergibt:

$$\hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{K})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

Alternativ kann man für Glättungssplines das Minimierungsproblem auch direkt für die Funktionsauswertungen $f(z_{(j)})$ schreiben. Sammelt man diese in dem Vektor $\boldsymbol{f} = (f(z_{(1)}), \ldots, f(z_{(d)}))'$ und definiert die zugehörige Designmatrix $\boldsymbol{Z}_f$ als eine Inzidenzmatrix mit Einträgen

$$\boldsymbol{Z}_f[i,j] = \begin{cases} 1 & \text{falls } z_i = z_{(j)}, \\ 0 & \text{sonst,} \end{cases}$$

so ergibt sich das penalisierte KQ-Kriterium

$$(\boldsymbol{y} - \boldsymbol{Z}_f\boldsymbol{f})'(\boldsymbol{y} - \boldsymbol{Z}_f\boldsymbol{f}) + \lambda \boldsymbol{f}'\boldsymbol{K}_f\boldsymbol{f}.$$

Die Strafmatrix kann hier basierend auf den Abständen $\delta_j = z_{(j+1)} - z_{(j)}$ konstruiert werden (vergleiche etwa Fahrmeir & Tutz (2001), S. 179/80). Die Minimierung des penalisierten KQ-Kriteriums für $\boldsymbol{f}$ ergibt nun einen Ausdruck für $\hat{\boldsymbol{f}}$, der in seiner Struktur dem des PKQ-Schätzers $\hat{\gamma}$ entspricht:

$$\hat{\boldsymbol{f}} = (\boldsymbol{Z}_f'\boldsymbol{Z}_f + \lambda \boldsymbol{K}_f)^{-1}\boldsymbol{Z}_f'\boldsymbol{y}.$$

Glättungssplines basieren in der Regel auf einer hochdimensionalen Basis, so dass eine sehr große Zahl von Parametern zu schätzen ist. Insbesondere wächst die Basis üblicherweise mit dem Stichprobenumfang an, so dass für große Datensätze die praktische Berechnung schnell aufwändig wird. Deshalb wird in vielen Implementationen von

Glättungssplines (etwa PROC GAM in SAS) direkt eine reduzierte Basis verwendet. Für diese reduzierte Basis gilt dann aber nicht mehr die Optimalitätsaussage, die zur Konstruktion von natürlichen kubischen Splines führte. Dementsprechend erscheint es naheliegender, direkt mit dem einfacheren P-Spline-Ansatz zu arbeiten, weshalb wir auch auf die Demonstration von Glättungssplines im Rahmen von Anwendungsbeispielen verzichten.

7.1.5 Random Walks

Random Walks lassen sich nicht nur zur Konstruktion von Priori-Verteilungen in Bayesianischen Regressionsmodellen verwenden, sondern können auch direkt zur Modellierung nichtlinearer Effekte eingesetzt werden. Um den im Rahmen der räumlichen Statistik in Abschnitt 7.2.4 behandelten Markov-Zufallsfeldern ein univariates Äquivalent entgegenzusetzen, werden wir deren Einsatzmöglichkeiten knapp skizzieren.

Random Walk-Modelle finden insbesondere in der Analyse von Zeitreihen Anwendung, so dass wir von einem zeitlichen Trend ausgehen, der wie zuvor die Funktion $f(z)$ nichtparametrisch modelliert werden soll:

$$y_t = f(t) + \varepsilon_t, \quad t = 1, \ldots, T.$$

Im Rahmen eines Random Walk-Ansatzes identifiziert man zunächst jede der Funktionsauswertungen $f(t)$ mit einem Parameter γ_t und modelliert dann diese Parameter als Random Walk, d.h.

$$f(t) = \gamma_t = \gamma_{t-1} + u_t, \quad u_t \sim \mathrm{N}(0, \tau^2),$$

beziehungsweise

$$f(t) = \gamma_t = 2\gamma_{t-1} - \gamma_{t-2} + u_t, \quad u_t \sim \mathrm{N}(0, \tau^2),$$

für Random Walks erster und zweiter Ordnung. Man beachte, dass sich Random Walks als spezielle P-Splines betrachten lassen, wenn man B-Splines vom Grad $l = 0$ zugrunde legt und die Knoten mit den beobachteten Zeitpunkten identifiziert. Wie für Bayesianische P-Splines ergibt sich aus dieser Annahme ein penalisiertes KQ-Kriterium, dass sich speziell für den RW2 schreiben lässt als

$$\mathrm{PKQ}(\lambda) = \sum_{t=1}^{T} (y_t - \gamma_t)^2 + \lambda \sum_{t=3}^{T} (\gamma_t - 2\gamma_{t-1} + \gamma_{t-2})^2.$$

Die Lösung des zugehörigen Minimierungsproblems ist in der Ökonometrie auch als Hodrick-Prescott-Filter bekannt (vergleiche Hodrick & Prescott (1997)), geht aber tatsächlich bereits auf Whittaker (1922/23) zurück. In Hodrick & Prescott (1997) werden Faustregeln zur Wahl des Glättungsparameters durch theoretische Überlegungen zum Signal-Rauschen-Verhältnis motiviert. Dagegen bietet die Einbettung in den allgemeineren Kontext von nichtparametrischen Regressionsmodellen die Möglichkeit, die in Abschnitt 7.1.9 behandelten Verfahren zur optimalen Wahl des Glättungsparameters zu verwenden.

Während für Zeitreihen die Annahme gleichabständiger Beobachtungen häufig erfüllt ist, gilt dies in der Regel nicht mehr, wenn der RW-Ansatz auf allgemeinere Kovariableneffekte $f(z)$ übertragen werden soll. In diesem Fall werden entsprechend der Abstände gewichtete Random Walk-Definitionen benötigt, auf die wir an dieser Stelle jedoch nicht detailliert eingehen wollen. Eine ausführliche Diskussion hierzu findet man in Rue & Held (2005) im Rahmen der allgemeineren Gauß-Markov-Zufallsfelder.

7.1.6 Kriging

In diesem Abschnitt beschäftigen wir uns mit einem Glättungsverfahren, das nicht unmittelbar zur univariaten Glättung entwickelt wurde, sondern dessen Ursprung in der räumlichen Statistik liegt. Dort wird der Ansatz als Kriging bezeichnet, nach dem südafrikanischen Bergbauingenieur D. G. Krige, der ihn zur Modellierung des Verlaufs von Gesteinsadern basierend auf Tiefenbohrungen verwendete. Obwohl Kriging zur univariaten Glättung nur selten verwendet wird, wollen wir bereits hier darauf eingehen, um die prinzipielle Vorgehensweise und die Nähe zu Basisfunktionenansätzen zu diskutieren.

Klassisches Kriging

Die wesentliche Idee des Kriging ist es, (im Originalvorschlag räumliche) Korrelationen der Daten durch eine parametrische Korrelationsfunktion auszudrücken, um sie damit bei der Schätzung eines Regressionsmodells berücksichtigen zu können. Dies geschieht durch die Einbeziehung eines stationären und multivariat normalverteilten stochastischen Prozesses, der analog zur Annahme zufälliger Effekte die gewünschten Korrelationen induziert. Daher wird Kriging allgemeiner auch als Annahme eines stationären Gauß-Prozesses (bzw. eines stationären Gauß-Feldes in der räumlichen Statistik) bezeichnet. Da wir uns momentan für univariate Glättungsverfahren interessieren, betrachten wir im Folgenden zunächst Zeitreihen anstelle von räumlich angeordneten Daten, die ausführlicher in Abschnitt 7.2.3 behandelt werden.

Gegeben sei also zunächst eine Zeitreihe

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \epsilon_t,$$

wobei der Zeitindex $t \in \mathbb{R}$ als stetig betrachtet wird. Im Gegensatz zum klassischen linearen Modell ist hier die Annahme unkorrelierter Fehler in der Regel nicht erfüllt, so dass es nahe liegt, zeitlich korrelierte Fehler zu betrachten. Dazu modellieren wir den Fehlerterm $\boldsymbol{\epsilon}$ als multivariat normalverteilt mit Erwartungswert $\mathbf{0}$ und Kovarianzmatrix $\sigma^2\boldsymbol{I} + \tau^2\boldsymbol{R}$. Dabei bezeichnet $\boldsymbol{R}$ eine Korrelationsmatrix, die durch eine parametrische Korrelationsfunktion ρ beschrieben wird, so dass sich die Einträge

$$\boldsymbol{R}[t,s] = \mathrm{Corr}(\epsilon_t, \epsilon_s) = \rho(t,s)$$

ergeben. Die Kovarianzmatrix setzt sich also aus einem unkorrelierten Teil $\sigma^2\boldsymbol{I}$ und einem korrelierten Teil $\tau^2\boldsymbol{R}$ zusammen. Während der erste Teil der üblichen Kovarianzmatrix im klassischen linearen Modell entspricht, modelliert der zweite Teil die zeitliche Abhängigkeit. Sind die Parameter σ^2 und τ^2 sowie die Korrelationsfunktion ρ vollständig bekannt, so erhält man ein allgemeines lineares Modell mit Kovarianzmatrix $\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{V} = \sigma^2\boldsymbol{I} + \tau^2\boldsymbol{R}$, das prinzipiell wie in Abschnitt 3.4 geschätzt werden kann. Insbesondere ergibt sich für $\boldsymbol{\beta}$ der gewichtete KQ-Schätzer

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

Die Varianzparameter σ^2 und τ^2 (und auch mögliche weitere Parameter der Korrelationsfunktion) können ähnlich wie in Kapitel 6 durch Maximieren der Profil-Likelihood oder der entsprechenden restringierten Likelihood bestimmt werden.

Üblicherweise geht man von einer *stationären* Fehlerverteilung aus, das heißt, $\rho(t,s)$ hängt nur von der Zeitdifferenz $|t-s|$ und nicht von der genauen Lokation der Zeitpunkte t und s auf der Zeitachse ab. Damit gelangt man zu Korrelationsfunktionen der Form $\rho(h)$ mit dem Argument $h = |t-s| > 0$.

Für jede gültige Korrelationsfunktion gilt $\rho(0) = 1$. Darüber hinaus setzt man häufig noch eine Reihe weiterer Eigenschaften voraus:

- Für $h \to \infty$ gilt $\rho(h) \to 0$, d.h. für Beobachtungen in einem großen zeitlichen Abstand verschwindet die Korrelation.
- Die Korrelationsfunktion $\rho(h)$ ist monoton fallend. Dies drückt die intuitive Idee aus, dass zeitlich nahe Beobachtungen auf jeden Fall eine größere Korrelation aufweisen sollten als zeitlich weiter voneinander entfernte Beobachtungen.

Einige häufig verwendete parametrische Korrelationsfunktionen ergeben sich aus den folgenden Klassen:

Die Familie der *sphärischen Korrelationsfunktionen* ist definiert durch

$$\rho(h;\phi) = \begin{cases} 1 - \frac{3}{2}|h/\phi| + \frac{1}{2}|h/\phi|^3 & 0 \le h \le \phi, \\ 0 & h > \phi. \end{cases}$$

Der einzige Parameter der sphärischen Korrelationsfunktion ist gegeben durch $\phi > 0$. Dieser steuert, bis zu welchem Wert noch positive Korrelationen auftreten bzw. ab welchem Abstand die Korrelation auf 0 fällt (vergleiche dazu auch Abbildung 7.20 a), die sphärische Korrelationsfunktionen für verschiedene Werte von ϕ zeigt). Aufgrund dieser Interpretation wird ϕ auch als *Range* der Korrelationsfunktion bezeichnet. Problematisch an der sphärischen Korrelationsfunktion ist ihre Nicht-Differenzierbarkeit im Punkt $h = \phi$. Offenbar erfolgt der Wechsel von positiven Korrelationen zu nicht mehr vorhandenen Korrelationen in einer zwar stetigen, aber eben nicht mehr differenzierbaren Form. Diese Eigenschaft führt bei der Maximum-Likelihood-Schätzung von ϕ zu Problemen.

Die *Potenz-Exponential-Familie* ist definiert durch

$$\rho(h;\phi,\kappa) = \exp\left(-|h/\phi|^{\kappa}\right)$$

mit den Parametern $\phi > 0$ und $0 < \kappa \le 2$. Speziell für $\kappa = 1$ ergibt sich die einfache Exponential-Korrelationsfunktion und für $\kappa = 2$ die sogenannte Gauß-Korrelationsfunktion, die an die Dichte einer Normalverteilung angelehnt ist. In Abbildung 7.20 b) ist die Potenz-Exponential-Familie für einige verschiedene Werte von κ (und $\phi = 1$) abgebildet. Im Gegensatz zur sphärischen Korrelationsfunktion werden durch Potenz-Exponential-Korrelationsfunktionen auch beliebig weit voneinander entfernte Punkte noch als korreliert angenommen. Der Range ist hier also formal unendlich. Zur Charakterisierung solcher Korrelationsfunktionen verwendet man dann häufig den *effektiven Range*, d.h. den Abstand, an dem die Korrelationsfunktion einen vorgegebenen, kleinen Wert (z.B. 0.01) unterschreitet. In Anwendungen werden Korrelationen für größere Abstände dann häufig auch tatsächlich als vernachlässigbar betrachtet und auf Null gesetzt, damit man eine dünner besetzte Korrelationsmatrix $\boldsymbol{R}$ erhält. Der effektive Range der Potenz-Exponential-Familie wird im Wesentlichen vom Parameter ϕ gesteuert, der die Skalierung der Zeitachse bestimmt, unterscheidet sich aber auch für verschiedene Werte von κ.

Für die einfache Exponential-Korrelationsfunktion ergibt sich auch eine interessante Verbindung zu linearen Modellen mit autokorrelierten Fehlern (vergleiche auch Ab-

Abb. 7.20. *Beispiele für Korrelationsfunktionen. In Abbildung a) sind sphärische Korrelationsfunktionen für $\phi = 1$ (—), $\phi = 1.75$ (- - -) und $\phi = 2.5$ ($\cdots$) visualisiert. Abbildung b) zeigt Exponential-Korrelationsfunktionen (für $\phi = 1$) mit $\kappa = 0.5$ (—), $\kappa = 1$ (- - -) und $\kappa = 2$ ($\cdots$). In Abbildung c) sind Matérn-Korrelationsfunktionen (für $\phi = 1$) mit $\kappa = 1.5$ (—), $\kappa = 2.5$ (- - -) und $\kappa = 3.5$ ($\cdots$) dargestellt.*

schnitt 3.4.4). In diesem Fall kann der Zeitindex t und damit auch die betrachtete Zeitdifferenz h nur diskrete Werte annehmen. Das einfachste und am häufigsten verwendete Modell zur Berücksichtigung von Korrelationen ist in dieser Situation ein AR1-Prozess, der folgende Struktur für die Residuen unterstellt:

$$\epsilon_t = \rho\epsilon_{t-1} + u_t, \quad u_t \sim \mathrm{N}(0, \tau^2). \tag{7.8}$$

Der Parameter ρ steuert den Einfluss des verzögerten Wertes ϵ_{t-1} auf ϵ_t, so dass sich für $\rho > 0$ positiv korrelierte Fehler ergeben mit Korrelationsfunktion

$$\mathrm{Corr}(\epsilon_t, \epsilon_{t+h}) = \rho^h, \quad h = 1, 2, \ldots$$

Damit erhält man eine in h exponentiell fallende Korrelation, die sich in Verbindung setzen lässt mit dem exponentiellen Abfall der Exponential-Korrelationsfunktion im zeitstetigen Modell. Definiert man den Parameter $\rho = \exp(-1/\phi)$, so lässt sich die Korrelationsfunktion auch schreiben als

$$\rho(h) = \exp(-h/\phi) = \exp(-1/\phi)^h = \rho^h.$$

Damit lässt sich der autoregressive Fehlerprozess (7.8) als eine diskretisierte Version eines zeitstetigen Prozesses mit Exponential-Korrelationsfunktion interpretieren.

Eine der flexibelsten Familien von Korrelationsfunktionen, die sich daher auch in der Praxis zum Standard der räumlichen Statistik entwickelt hat, ist die *Matérn-Familie* $\rho(h;\phi,\kappa)$ mit $\phi > 0$, $\kappa > 0$. Die allgemeine Darstellung von Korrelationsfunktionen aus dieser Familie ist jedoch nur noch mit Hilfe von modifizierten Bessel-Funktionen der Ordnung κ möglich, die sich nur numerisch auswerten, aber nicht explizit darstellen lassen. Darum beschränkt man sich häufig auf die Teilmenge der Matérn-Korrelationsfunktionen, die sich für $\kappa = 0.5, 1.5, 2.5, \ldots$ ergeben. In diesen Fällen lassen sich die Korrelationsfunktionen direkt angeben und man erhält beispielsweise

$$\begin{aligned}
\rho(h;\phi,\kappa=0.5) &= \exp(-|h/\phi|),\\
\rho(h;\phi,\kappa=1.5) &= \exp(-|h/\phi|)(1+|h/\phi|),\\
\rho(h;\phi,\kappa=2.5) &= \exp(-|h/\phi|)(1+|h/\phi|+\tfrac{1}{3}|h/\phi|^2),\\
\rho(h;\phi,\kappa=3.5) &= \exp(-|h/\phi|)(1+|h/\phi|+\tfrac{2}{5}|h/\phi|^2+\tfrac{1}{15}|h/\phi|^3).
\end{aligned}$$

Für $\kappa = 0.5$ entspricht die Matérn-Korrelationsfunktion also wieder der einfachen Exponential-Korrelationsfunktion. Die übrigen drei Korrelationsfunktionen sind in Abbildung 7.20 c) dargestellt. Offenbar führen größere Werte von κ zu Korrelationsfunktionen mit größerem effektivem Range, also größeren Korrelationen für weiter auseinander liegende Zeitpunkte. Lässt man κ gegen unendlich gehen, so erhält man als Grenzfall die Gauß-Korrelationsfunktion, also wieder eine spezielle Potenz-Exponential-Korrelationsfunktion.

Kriging als Glättungsverfahren für Zeitreihen

Um die Verwendung von Korrelationsfunktionen und stationären Gauß-Prozessen in der nichtparametrischen Funktionsschätzung zu motivieren, zerlegen wir den Fehler ϵ_t in einen zeitlich korrelierten Fehler γ_t und den unabhängig und identisch verteilten Restfehler ε_t:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \gamma_t + \varepsilon_t, \quad t \in \{t_{(1)}, \ldots, t_{(d)}\}.$$

Durch geeignete Spezifikation von γ_t und ε_t erhält man dann genau die zuvor spezifizierte Kovarianzstruktur für die Zielvariable $\boldsymbol{y}$.

In Matrixnotation lässt sich das Modell darstellen als

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

wobei $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ den Vektor der d verschiedenen, zeitlich geordneten Fehlerterme γ_t und $\boldsymbol{Z} = \boldsymbol{I}_d$ die d-dimensionale Einheitsmatrix bezeichnet. In einem zweiten Schritt reparametrisieren wir das Modell zu

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{R}\cdot\boldsymbol{R}^{-1}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}$$

mit $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{R}$ und $\tilde{\boldsymbol{\gamma}} = \boldsymbol{R}^{-1}\boldsymbol{\gamma}$. Offenbar führt diese Reparametrisierung zur gleichen Verteilung für $\boldsymbol{y}$ und damit zu einer äquivalenten Modellformulierung. Allerdings ändert sich die Interpretation der Designmatrix $\tilde{\boldsymbol{Z}}$. Aufgrund der speziellen Struktur von $\boldsymbol{Z}$ ergibt sich hier für die individuellen Einträge

$$\tilde{\boldsymbol{Z}}[i,j] = \rho(t_{(i)}, t_{(j)}).$$

Vergleicht man diese Definition mit der Konstruktion der Designmatrix für B-Splines oder TP-Splines, so erkennt man, dass die Korrelationsfunktion ρ wie eine Basisfunktion verwendet wird und die beobachteten Zeitpunkte $t_{(j)}$ die Rolle von Knoten übernehmen. Damit erhält man die Darstellung

$$f(t) = \gamma_t = \sum_{j=1}^{d} \tilde{\gamma}_j \rho(t, t_{(j)})$$

für den zeitlichen Trend. Die gemeinsame Verteilung der zeitlich korrelierten Effekte $\tilde{\boldsymbol{\gamma}}$ ergibt sich zu

$$\tilde{\boldsymbol{\gamma}} \sim \mathrm{N}(0, \tau^2 \boldsymbol{R}^{-1}),$$

besitzt also eine Dichte der Form

$$p(\tilde{\boldsymbol{\gamma}}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\tilde{\boldsymbol{\gamma}}'\boldsymbol{R}\tilde{\boldsymbol{\gamma}}\right).$$

Diese Dichte entspricht in ihrer Form genau den im Zusammenhang mit den bisher betrachteten Penalisierungsansätzen behandelten Glättungs-Prioris. Basierend auf der Dichte lässt sich auch wieder ein äquivalentes penalisiertes KQ-Kriterium definieren durch

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\gamma}})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\gamma}}) + \frac{1}{2\tau^2}\tilde{\boldsymbol{\gamma}}'\boldsymbol{R}\tilde{\boldsymbol{\gamma}}$$

Obwohl der Kriging-Ansatz also aus einer rein stochastischen Modellformulierung zur Beschreibung korrelierter Daten hergeleitet wurde, ist er formal äquivalent zu einem Basisfunktionenansatz mit Basisfunktionen ρ und Strafmatrix $\boldsymbol{R}$. Dies erklärt auch, warum die gleichzeitige Modellierung eines zeitlichen Trends durch einen nichtparametrischen Ansatz (oder einfache Polynome) und die Berücksichtigung eines zeitlich korrelierten Fehlers zu Identifizierbarkeitsproblemen führen kann. Durch den zeitlich korrelierten Fehler wird implizit eine nichtparametrische Trendfunktion mit geschätzt, so dass die zusätzliche Annahme eines weiteren Trends zu einem eventuell kaum identifizierbaren Modell führt. Aus der Darstellung des Krigingansatzes als Basisfunktionenansatz erkennt man auch, dass die Schätzung des zeitlichen Effekts die Glattheitseigenschaften der Korrelationsfunktion ρ erbt. Insbesondere ist die geschätzte Funktion genau dann stetig bzw. stetig differenzierbar, wenn dies für die Korrelationsfunktion zutrifft.

Kriging als Glättungsverfahren der nichtparametrische Regression

Um Kriging allgemein zur Schätzung nichtparametrischer Kovariableneffekte einsetzen zu können, gehen wir nun wieder von der Problemstellung

$$y_i = f(z_i) + \varepsilon_i$$

aus. Bezeichnen dann wie zuvor $z_{(1)} < \ldots < z_{(d)}$ die verschiedenen, geordneten Kovariablenausprägungen, so definiert man die Parameter $\gamma_j = f(z_{(j)})$ und nimmt für diese einen stationären Gauß-Prozess mit Erwartungswert 0, Varianz τ^2 und Korrelationsfunktion

$$\rho(\gamma_j, \gamma_k) = \rho(|z_{(j)} - z_{(k)}|)$$

Kriging als Glättungsverfahren der nichtparametrischen Regression

Die Funktion f wird in Basisfunktionen entwickelt, die sich aus parametrischen Korrelationsfunktionen $\rho(h)$ ergeben. Damit erhält man das penalisierte KQ-Kriterium

$$\mathrm{PKQ}(\lambda) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$$

mit $\boldsymbol{Z}[i,j] = \rho(|z_i - z_{(j)}|)$ und $\boldsymbol{K}[j,k] = \rho(|z_{(j)} - z_{(k)}|)$. Glattheits-Eigenschaften der gewählten Korrelationsfunktion übertragen sich auf die Schätzung $\hat{f}(z)$.

wie im vorherigen Zeitreihen-Beispiel an. Analog kann man die Korrelationsfunktion dann wieder als eine Basisfunktion betrachten und die Kovariablenausprägungen $x_{(j)}$ wie für Glättungssplines als Knoten identifizieren. Damit erhält man die Darstellung

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mit $\boldsymbol{Z}[i,j] = \rho(|z_i - z_{(j)}|)$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ und als zu optimierendes Kriterium die penalisierte Residuenquadratsumme

$$\mathrm{PKQ}(\lambda) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$$

mit $\boldsymbol{K}[j,k] = \rho(|z_{(j)} - z_{(k)}|)$ und Glättungsparameter $\lambda = \sigma^2/\tau^2$. Damit ist das übliche Instrumentarium zur penalisierten KQ-Schätzung wieder anwendbar. Insbesondere ergibt sich wie zuvor für P-Splines oder Glättungssplines die bekannte Form des PKQ-Schätzers.

Beispiel 7.3 Unterernährung in Tansania: Kriging

In Abbildung 7.21 sind verschiedene Schätzungen für den Einfluss des Alters auf den Unterernährungs-Score basierend auf Kriging-Ansätzen dargestellt. Dabei wurden die vier vorgestellten speziellen Matérn-Korrelationsfunktionen verwendet und der Skalierungsparameter mit Hilfe der Regel

$$\hat{\phi} = \max_{j,k} |z_{(j)} - z_{(k)}|/c \tag{7.9}$$

gewählt. Dabei ist $c > 0$ eine Konstante, die so bestimmt wird, dass $\rho(c)$ klein ist, das heißt, c bestimmt den effektiven Range der Korrelationsfunktion. Durch die obige Wahl wird bewirkt, dass sich die Korrelationsfunktionen günstig über den Definitionsbereich verteilen. Für Polynom-Splines wurde dies automatisch durch die Wahl der Knoten garantiert. Im Falle des Kriging könnte ein zu kleiner effektiver Range dazu führen, dass sich die Korrelationsfunktionen nicht mehr überlappen, so dass in den entsprechenden Bereichen der nonparametrische Effekt nicht mehr adäquat erfasst werden kann. Darüber hinaus wird durch (7.9) die Skalierungsinvarianz der Schätzung gesichert (d.h. bei einer Skalierung der Kovariablenausprägungen z_i wird diese Skalierung durch $\hat{\phi}$ wieder kompensiert). Die Varianz τ^2 (und damit der Glättungsparameter λ) wurde per REML-Schätzung (vergleiche Abschnitt 6.3.2) bestimmt. In Abschnitt 7.1.9 werden wir sehen, dass sich die REML-Schätzung tatsächlich auch zur Bestimmung von Glättungsparametern in penalisierten Basisfunktionenansätzen und weiteren nichtparametrischen Regressionsmodellen einsetzen lässt.

An den geschätzten Effekten des Alters erkennt man die bereits vorher diskutierte Auswirkung der Eigenschaften der verwendeten Korrelationsfunktion auf die Schätzung

Abb. 7.21. *Kriging-Schätzungen für den Tansania-Datensatz.*

$\hat{f}(z)$. Für $\kappa = 0.5$ erhält man eine eher raue, insbesondere nicht differenzierbare Funktionsschätzung. Für größere Werte von κ wird die Funktion dann entsprechend glatter geschätzt mit relativ geringen Unterschieden für die unterschiedlichen Korrelationsfunktionen.

△

7.1.7 Lokale Glättungsverfahren

Wir wollen uns nun einer anderen Klasse von Schätzern zuwenden, die sich in der Regel nicht mehr aus der globalen Formulierung eines Regressionsproblems herleiten lassen, sondern *lokal* definiert sind. Viele dieser lokalen Schätzer sind äußerst intuitiv und einfach zu verstehen und werden deshalb häufig in explorativen Analysen eingesetzt.

Nächste-Nachbarn-Schätzer

Ein zur (deskriptiven) Glättung von Zeitreihen häufig verwendetes Verfahren ist die Bildung sogenannter *gleitender Durchschnitte* (Running Means). Beispielsweise ist für eine Zeitreihe y_t, $t = 1, \ldots, T$, ein gleitender Durchschnitt der Ordnung 3 definiert durch

$$\hat{y}_t = \frac{1}{3}(y_{t-1} + y_t + y_{t+1})$$

mit geeigneten Modifikationen an den Rändern. Offenbar werden durch die Bildung der Durchschnitte die zufälligen Schwankungen einer Zeitreihe geglättet, so dass man einen ersten Eindruck vom zugrunde liegenden Trend erhalten kann.

Nächste-Nachbarn-Schätzer erweitern das Konzept gleitender Durchschnitte in einem allgemeineren Rahmen und insbesondere auf die Anwendung in nichtparametrischen Regressionsmodellen. Allgemein wird der *Nächste-Nachbarn-Schätzer* definiert durch

$$\hat{f}(z) = \underset{j \in N(z)}{\mathrm{MWO}}\, y_j,$$

wobei MWO einen beliebigen Mittelwertsoperator und $N(z)$ eine Nachbarschaft von z bezeichnet. Im obigen Beispiel des gleitenden Durchschnitts der Ordnung 3 ist der Mittelwertsoperator das arithmetische Mittel und die Nachbarschaft wird durch die zwei (zeitlich) benachbarten Werte $t-1$ und $t+1$ (sowie t selbst) gebildet.

Die folgenden Mittelwertsoperatoren werden häufig zur Bestimmung von Nächste-Nachbarn-Schätzern verwendet:

1. *Arithmetisches Mittel* (Running Mean): Bestimme das arithmetische Mittel der abhängigen Variablen aus der Nachbarschaft von z_i, d.h.

$$\hat{f}(z) = \frac{1}{|N(z)|} \sum_{j \in N(z)} y_j,$$

   wobei $|N(z)|$ die Anzahl der Nachbarn von z angibt.
2. *Median* (Running Median): Bestimme den Median der abhängigen Variablen aus der Nachbarschaft von z, d.h.

$$\hat{f}(z) = \mathrm{Median}\{y_j, j \in N(z)\}.$$

3. *Lineare Einfachregression* (Running Line): Schätze eine lineare Einfachregression basierend auf den Nachbarn von z und verwende die Vorhersage aus diesem Modell als Schätzwert, d.h. verwende

$$\hat{f}(z) = \hat{\gamma}_{0,z} + \hat{\gamma}_{1,z} z$$

   mit den aus $\{(z_j, y_j), j \in N(z)\}$ bestimmten KQ-Schätzern $\hat{\gamma}_{0,z}$ und $\hat{\gamma}_{1,z}$.

Zur vollständigen Bestimmung eines Nächste-Nachbarn-Schätzers benötigen wir noch geeignete Definitionen von Nachbarschaften. Üblicherweise verwendet werden

1. Symmetrische Nachbarschaften der Ordnung k oder
2. Nachbarschaften bestehend aus den k nächsten Nachbarn.

Ein einfaches Beispiel für eine symmetrische Nachbarschaft haben wir bereits durch die gleitenden Durchschnitte der Ordnung 3 kennengelernt. Zur allgemeinen Definition symmetrischer Nachbarschaften (mit ungerader Ordnung k) geht man folgendermaßen vor: Verwende zur Schätzung von $f(z_i)$ (zusätzlich zu z_i selbst) die jeweils $(k-1)/2$ nächsten Beobachtungen rechts und links von z_i auf dem Definitionsbereich. Man teilt also die Ordnung symmetrisch auf und sucht rechts und links von z diejenigen $(k-1)/2$ Beobachtungen mit den am nächsten an z_i gelegenen Beobachtungen. Gehen wir der Einfachheit halber von geordneten Beobachtungen mit $z_1 \leq \ldots \leq z_n$ aus, so lässt sich die symmetrische Nachbarschaft formal definieren als

Nächste-Nachbarn-Schätzer

Allgemeine Form eines Nächste-Nachbarn-Schätzers:

$$\hat{f}(z) = \underset{j \in N(z)}{\text{MWO}}\, y_j,$$

mit geeignetem Mittelwertsoperator MWO und einer lokalen Nachbarschaft $N(z)$. Als Mittelwertsoperatoren werden häufig das arithmetische Mittel, der Median oder die lineare Einfachregression verwendet. Nachbarschaften können entweder symmetrisch um z oder basierend auf den k nächsten Nachbarn definiert werden.

$$N(z_i) = \{\max(1, i-(k-1)/2), \ldots, i-1, i, i+1, \ldots, \min(n, i+(k-1)/2)\}\,.$$

In dieser Definition sind auch bereits Korrekturen für den Rand des Definitionsbereichs enthalten, da dort nicht mehr genügend Beobachtungen zur Verfügung stehen, um die übliche Definition anzuwenden. Man beachte jedoch, dass die Nachbarschaft an diesen Rändern nicht mehr symmetrisch ist und nicht mehr k Beobachtungen enthält.

Alternativ zur Ordnung k lassen sich symmetrische Nachbarschaften auch durch den Anteil bzw. die *Bandbreite* $\omega \in (0,1)$ von Beobachtungen in der Nachbarschaft definieren. Die Ordnung k ergibt sich dann als ωn, wobei für nicht ganzzahliges ωn geeignet gerundet werden muss. Sowohl die Ordnung k als auch der Anteil ω lassen sich als Glättungsparameter des Nächste-Nachbarn-Schätzers interpretieren. In Abbildung 7.22 ist dies für den Tansania-Datensatz und das lokale arithmetische Mittel demonstriert. Je näher die Bandbreite ω bei 1 liegt, desto mehr Nachbarn werden in die Schätzung einbezogen, so dass sich eine glatte Schätzung ergibt. Wird ω dagegen sehr klein, also nahe bei 0 gewählt, so basiert die Schätzung für $\hat{f}(z_i)$ nur auf einer sehr kleinen Zahl von Beobachtungen und man erhält eine sehr raue Schätzung.

Die Definition einer Nachbarschaft basierend auf k nächsten Nachbarn führt im Allgemeinen zu asymmetrischen Nachbarschaften. Die Nachbarschaft $N(z)$ ist hier definiert als

$$N(z) = \{i : d_i \in \{d_{(1)}, \ldots, d_{(k)}\}\},$$

wobei $d_{(1)}, \ldots, d_{(n)}$ die geordneten Distanzen $d_i = |z_i - z|$ bezeichnet. Hier wird auch an den Rändern immer die gleiche Anzahl von Nachbarn verwendet, es sind also keine Randmodifikationen notwendig. Ähnlich wie für symmetrische Nachbarschaften lässt sich die Anzahl verwendeter Nachbarn k als Glättungsparameter interpretieren. Ist k klein, resultiert eine raue Schätzung, während ein nahe beim Stichprobenumfang n liegender Wert zu einer sehr glatten Schätzung führt, die der globalen Anwendung des Mittelwertoperators entspricht.

Lokal polynomiale Regression und Nadaraya-Watson-Schätzer

Obwohl wir bereits zu Beginn dieses Kapitels festgestellt haben, dass eine globale Darstellung der zu schätzenden Funktion $f(z)$ durch Polynome in der Regel nicht ausreichend

Abb. 7.22. *Gleitender Durchschnitt für den Tansania-Datensatz und verschiedene Bandbreiten.*

flexibel ist, kann man $f(z)$ häufig lokal durch ein Polynom approximieren (vergleiche hierzu auch Abbildung 2.12 auf Seite 42). Dabei verwenden wir als mathematisches Hilfsmittel die (lokale) Approximation einer l-mal stetig differenzierbaren Funktion $f(z_i)$ durch eine *Taylorreihe um den Entwicklungspunkt* z

$$f(z_i) \approx f(z) + (z_i - z)f'(z) + (z_i - z)^2 \frac{f''(z)}{2!} + \ldots + (z_i - z)^p \frac{f^{(l)}(z)}{l!}.$$

Die Funktion $f(z_i)$ lässt sich somit in einer Umgebung von z_i durch Polynome der Form $(z_i - z)^j$ approximieren, die gewichtet werden durch die Ableitungen $f^{(j)}(z)/j!$ ausgewertet am Entwicklungspunkt (wobei $f(z)$ als nullte Ableitung von f betrachtet wird).

Wendet man die Approximation durch eine Taylorreihe auf das Problem der nichtparametrischen Schätzung der Funktion $f(z)$ an einer vorgegebenen Stelle z an, so erhält man für jede Beobachtung (y_i, z_i) unter Verwendung von z als Entwicklungspunkt

$$\begin{aligned} y_i &= f(z_i) + \varepsilon_i \\ &\approx f(z) + (z_i - z)f'(z) + (z_i - z)^2 \frac{f''(z)}{2!} + \ldots + (z_i - z)^l \frac{f^{(l)}(z)}{l!} + \varepsilon_i \\ &= \gamma_0 + (z_i - z)\gamma_1 + (z_i - z)^2\gamma_2 + \ldots + (z_i - z)^l\gamma_l + \varepsilon_i. \end{aligned}$$

Es ergibt sich also ein polynomiales Regressionsmodell für y_i basierend auf Polynomen der Form $(z_i - z)^j$ und Regressionskoeffizienten $\gamma_j = f^{(j)}(z)/j!$. Bestimmt man nun aus diesem Regressionsmodell Schätzungen für die Regressionsparameter, so ergibt sich

Abb. 7.23. *Beispiele für Kernfunktionen.*

implizit eine Schätzung für den Funktionswert $f(z)$ durch $\hat{\gamma}_0 = \hat{f}(z)$ und allgemeiner sogar eine Schätzung für dessen Ableitungen durch $j!\hat{\gamma}_j = \hat{f}^{(j)}(z)$.

Prinzipiell lässt sich also $f(z)$ durch die Schätzung eines linearen Modells bestimmen. Da die der Taylorreihe zugrunde liegende Approximation jedoch nur lokal gilt, legt man der Schätzung eine gewichtete Version der üblichen Residuenquadratsumme zu Grunde. Dies ergibt das gewichtete KQ-Kriterium

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{l} \gamma_j (z_i - z)^j \right)^2 w_\lambda(z, z_i)$$

mit Gewichten $w_\lambda(z, z_i)$. Diese werden auf Grundlage der Abstände $|z_i - z|$ bestimmt, wobei das Gewicht umso größer sein soll, je kleiner dieser Abstand ist. Eine allgemeine Klasse solcher Gewichte ergibt sich aus der Verwendung von sogenannten *Kernfunktionen* K durch

$$w_\lambda(z, z_i) = K\left(\frac{z_i - z}{\lambda}\right). \tag{7.10}$$

Typische Beispiele hierfür sind

$$K(u) = \begin{cases} \frac{1}{2} & -1 \le u \le 1 \\ 0 & \text{sonst} \end{cases} \qquad \text{Gleichverteilungs-Kern}$$

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & -1 \le u \le 1 \\ 0 & \text{sonst} \end{cases} \qquad \text{Epanechnikov-Kern}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \qquad \text{Gauß-Kern}$$

(vergleiche auch Abbildung 7.23). Man beachte, dass sowohl der Epanechnikov-Kern als auch der Gleichverteilungskern alle Gewichte außerhalb des Intervalls $[-1, 1]$ auf Null setzen, während der Gauß-Kern auch noch für beliebig weit entfernte Beobachtungen positive Gewichte vergibt. Der zusätzliche Parameter λ (die *Bandweite* des Kerns) steuert dabei, wie schnell die Gewichte gegen Null gehen, und kann als Glättungsparameter der lokal polynomialen Regression betrachtet werden. Kernfunktionen finden auch in der nichtparametrischen Dichteschätzung Verwendung (vergleiche etwa Fahrmeir et al. (2007) oder Härdle (1990)) und sind daher selbst auch wieder Dichten. In der Regel werden Kernfunktionen als symmetrisch um Null vorausgesetzt.

Zur Bestimmung des gewichteten KQ-Schätzers im lokal polynomialen Regressionsmodell schreiben wir zunächst die gewichtete Residuenquadratsumme in Matrixnotation als

Lokal polynomiale Regression

Basierend auf einer Taylor-Entwicklung um den Punkt z erhält man lokal die Darstellung

$$y_i \approx \gamma_0 + (z_i - z)\gamma_1 + (z_i - z)^2\gamma_2 + \ldots + (z_i - z)^l\gamma_l + \varepsilon_i$$

mit Regressionskoeffizienten der Form $\gamma_j = f^{(j)}(z)/j!$. Zur Schätzung minimiert man das gewichtete KQ-Kriterium

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{l} \gamma_j (z_i - z)^j \right)^2 w_\lambda(z, z_i),$$

wobei die Gewichte $w_\lambda(z, z_i)$ durch eine Kernfunktion definiert werden. Als Schätzer erhält man den gewichteten KQ-Schätzer und insbesondere $\hat{f}(z) = \hat{\gamma}_0$. Für $l = 0$ ergibt sich der Spezialfall des Nadaraya-Watson-Schätzers.

$$(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})$$

mit der Designmatrix

$$\boldsymbol{Z} = \begin{pmatrix} 1 & (z_1 - z) & \ldots & (z_1 - z)^l \\ \vdots & \vdots & & \vdots \\ 1 & (z_n - z) & \ldots & (z_n - z)^l \end{pmatrix},$$

dem Vektor von Regressionskoeffizienten $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_l)'$ und der Gewichtsmatrix $\boldsymbol{W} = \operatorname{diag}(w_\lambda(z, z_1), \ldots, w_\lambda(z, z_n))$. Die Minimierung dieser Residuenquadratsumme entspricht der Schätzung eines verallgemeinerten linearen Modells, wie wir es in Abschnitt 3.4 kennengelernt haben, so dass sich der gewichtete KQ-Schätzer

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{y}$$

mit Kovarianzmatrix $\operatorname{Cov}(\hat{\boldsymbol{\gamma}}) = (\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}$ ergibt. Da sowohl die Designmatrix $\boldsymbol{Z}$ als auch die Gewichtsmatrix $\boldsymbol{W}$ von der Position z abhängen, für die man die Funktion f schätzen möchte, muss für jeden interessierenden Punkt eine separate Schätzung durchgeführt werden. In der Praxis gibt man sich hierfür üblicherweise ein regelmäßiges Gitter von Kovariablenwerten vor.

Ein interessanter Spezialfall ergibt sich, wenn man das lokal konstante polynomiale Modell betrachtet, also die gewichtete Residuenquadratsumme

$$\sum_{i=1}^{n} (y_i - \gamma_0)^2 \, w_\lambda(z, z_i).$$

Hier ergibt sich für den Schätzer $\hat{f}(z)$ die explizite Darstellung

$$\hat{f}(z) = \frac{\sum_i w_\lambda(z, z_i) y_i}{\sum_i w_\lambda(z, z_i)}. \tag{7.11}$$

Loess-Schätzung

1. Bestimme die Menge $N(z)$ der k nächsten Nachbarn von z.
2. Bestimme den größten Abstand zweier Datenpunkte in dieser Nachbarschaft
$$\Delta(z) = \max_{i,j \in N(z)} |z_i - z_j|.$$
3. Definiere Gewichte
$$w_{\Delta(z)}(z, z_i) = K\left(\frac{|z - z_i|}{\Delta(z)}\right)$$
mit der trikubischen Kernfunktion
$$K(u) = \begin{cases} (1 - |u|^3)^3 & -1 \leq u \leq 1, \\ 0 & \text{sonst.} \end{cases}$$
4. Bestimme $\hat{f}(z)$ durch gewichtete (lineare) Regression basierend auf den Datenpunkten in der Nachbarschaft $N(z)$.

Dieser Schätzer lässt sich alternativ auch aus Ideen zur nichtparametrischen Dichteschätzung herleiten und wird in diesem Zusammenhang als *Nadaraya-Watson-Schätzer* bezeichnet, vergleiche etwa Fahrmeir & Tutz (2001), Kap. 5 oder Härdle (1990).

Loess

Nächste-Nachbarn-Schätzer resultieren häufig auch dann in relativ rauen Schätzungen, wenn der Glättungsparameter gemäß einem geeigneten Optimalitätskriterium gewählt wird. Der Grund hierfür liegt in der Definition der lokalen Schätzung durch Nachbarschaften, da hier die Beobachtungen in der Nachbarschaft für die Schätzung volles Gewicht bekommen, während alle übrigen Beobachtungen vollständig ausgeschlossen werden. Kombiniert man die in der lokal polynomialen Regression eingeführte Gewichtung mit Nächste-Nachbarn-Schätzern, so lassen sich glattere Schätzungen erhoffen, da nun anstelle der abrupt auf Null fallenden Gewichte langsam auslaufende Gewichte verwendet werden. Dies führt zur *lokal gewichteten Regression* (Loess, Locally weighted regression), deren Originalvorschlag (Cleveland 1979) sich in dem im Kasten angegebenen Algorithmus zur Schätzung von $f(z)$ an einer vorgegebenen Stelle z zusammenfassen lässt.

Der Name Loess leitet sich sowohl aus einer Abkürzung für LOcal RegrESSion als auch von dem deutschen Wort Löß her. In der Geologie bezeichnet Löß ein durch den Wind angelagertes Sediment, das aus der Zerstörung anderer Gesteine hervorgegangen ist. Damit bilden Lößschichten eine glatte Oberfläche, was zur entsprechenden Namensgebung für die glatte Funktionsschätzung durch Loess führte. Häufig liest man auch die Bezeichnung *Lowess*, die für LOcally WEighted Scatterplot Smoothing steht und somit nochmal verstärkt, dass es sich bei Loess um einen gewichteten Schätzer zur nichtparametrischen Regression handelt. Die Verwendung der trikubischen Kernfunktion wird im Originalvorschlag anhand asymptotischer Überlegungen motiviert, in der Praxis können aber auch

andere Kernfunktionen zum Einsatz kommen. In statistischen Programmpaketen sind unter den Bezeichnungen Loess und Lowess Varianten von Loess erhältlich, die sich zwar im Detail unterscheiden, deren wesentliches Vorgehen aber durch den obigen Algorithmus beschrieben wird.

7.1.8 Allgemeine Streudiagramm-Glätter

Um verschiedene Glättungsverfahren der nichtparametrischen Regression vergleichen zu können, führen wir in diesem Abschnitt zunächst eine allgemeine Formulierung für Streudiagramm-Glätter ein und beschäftigen uns anschließend mit einer Reihe von Eigenschaften, die sich aus dieser Formulierung ableiten lassen.

Lineare Glättungsverfahren

Betrachtet man die in diesem Kapitel behandelten Glättungsverfahren, so stellt man fest, dass sich diese bis auf wenige Ausnahmen mit Hilfe von Gewichten $s(z, z_i)$ in der folgenden Form darstellen lassen:

$$\hat{f}(z) = \sum_{i=1}^{n} s(z, z_i) y_i. \tag{7.12}$$

Dementsprechend erhält man den Schätzwert $\hat{f}(z)$ als gewichtete Summe der Beobachtungen y_i, wobei die Gewichte sowohl von dem Wert der Kovariablen z abhängen, an dem die Funktion geschätzt werden soll, als auch von den beobachteten Kovariablenausprägungen $z_1, \ldots, z_n$. Da sich $\hat{f}(z)$ also als Linearkombination der beobachteten Zielvariablen darstellen lässt, bezeichnet man Glättungsverfahren, die Eigenschaft (7.12) erfüllen, auch als *lineare Glättungsverfahren*.

Besonders einfach lässt sich Eigenschaft (7.12) für Nächste-Nachbarn-Schätzer überprüfen (vergleiche Seite 333f). Hier wurde die Funktionsschätzung $\hat{f}(z)$ definiert als lokaler Mittelwert aller beobachteten abhängigen Variablen in einer Nachbarschaft von z. Verwendet man nun als Mittelwertsoperator einen linearen Operator, so ergibt sich ein lineares Glättungsverfahren. Konkret bedeutet dies, dass sowohl das lokale arithmetische Mittel als auch die lokale lineare Einfachregression lineare Glättungsverfahren sind, während der lokale Median ein nichtlinearer Glätter ist. Für das lokale arithmetische Mittel erhält man beispielsweise

$$s(z, z_i) = \begin{cases} \frac{1}{|N(z)|} & \text{falls } i \in N(z), \\ 0 & \text{sonst.} \end{cases}$$

Da der Median ein nichtlinearer Operator ist, lassen sich in diesem Fall keine solchen Gewichte spezifizieren und man erhält ein Beispiel für ein nichtlineares Glättungsverfahren.

Ein lineares Glättungsverfahren lässt sich in Matrixnotation schreiben als

$$\hat{f}(z) = \boldsymbol{s}(z)'\boldsymbol{y}$$

mit dem Vektor $\boldsymbol{s}(z) = (s(z, z_1), \ldots, s(z, z_n))'$. Insbesondere erhält man an den beobachteten Kovariablenausprägungen $\hat{f}(z_i) = \boldsymbol{s}(z_i)'\boldsymbol{y}$. Wendet man diese Gleichung nun auf alle n Kovariablenausprägungen an, so ergibt sich als linearer Glätter

$$\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{y},$$

mit $\hat{\boldsymbol{f}} = (\hat{f}(z_1), \ldots, \hat{f}(z_n))'$ und der aus den Vektoren $\boldsymbol{s}(z_i)$ zusammengesetzten $(n \times n)$ Matrix $\boldsymbol{S}$, die als *Glättungsmatrix* (*Smoother-Matrix*) bezeichnet wird. Sie bildet die Grundlage für einige der im Folgenden hergeleiteten Größen zur Beschreibung von Glättungsverfahren.

Das in Kapitel 3 behandelte lineare Modell, ebenso wie die in Abschnitt 3.4 behandelten Varianten zur allgemeinen linearen Regression, lassen sich als lineare Glätter auffassen. In diesen Fällen ist die Glättungsmatrix gegeben durch die Prädiktionsmatrix $\boldsymbol{H}$ (vergleiche Seite 93), d.h. durch

$$\boldsymbol{S} = \boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$$

im gewöhnlichen linearen Modell beziehungsweise durch

$$\boldsymbol{S} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{W}^{-1}$$

im allgemeinen linearen Modell. Damit ergibt sich unmittelbar, dass sich Polynom-Splines und auch beliebige andere Basisfunktionenansätze (ohne Penalisierung) unabhängig von der konkret verwendeten Basis ebenfalls als lineare Glätter auffassen lassen. Zur Schätzung an einem unbeobachteten Wert z einer Kovariablen erhält man hier (vergleiche Seite 299)

$$\boldsymbol{s}(z)' = \boldsymbol{z}'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}',$$

wobei $\boldsymbol{z}$ den Vektor der am Argument z ausgewerteten Basisfunktionen bezeichnet.

Für beliebige Penalisierungsansätze, die auf einer Modelldarstellung der Form

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mit Strafmatrix $\boldsymbol{K}$ und Glättungsparameter λ basieren, muss die Definition lediglich so modifiziert werden, dass für den Schätzer $\hat{\boldsymbol{\gamma}}$ der penalisierte KQ-Schätzer eingesetzt wird. Damit erhält man

$$\boldsymbol{S} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}'$$

und

$$\boldsymbol{s}(z)' = \boldsymbol{z}'(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}'.$$

Damit sind insbesondere alle aus Polynom-Splines abgeleiteten Penalisierungsansätze, aber auch Glättungssplines, Random Walks und Kriging-Ansätze lineare Glättungsverfahren.

Auch die beiden noch verbleibenden Ansätze (lokal polynomiale Regression und Loess) lassen sich als lineare Glättungsverfahren auffassen. Für die lokal polynomiale Regression ergab sich die Funktionsschätzung $\hat{f}(z)$ als Konstante eines gewichteten KQ-Schätzers (vergleiche Seite 337). Damit erhält man den entsprechenden Gewichtsvektor $\boldsymbol{s}(z)$ über

$$\hat{f}(z) = \hat{\gamma}_0 = \boldsymbol{e}_1'(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{y} = \boldsymbol{s}(z)'\boldsymbol{y}$$

mit $\boldsymbol{e}_1 = (1, 0, 0, \ldots)'$. Der Vektor $\boldsymbol{e}_1$ greift also lediglich die erste Komponente aus dem gewichteten KQ-Schätzer heraus. Für den Loess-Schätzer folgt die Linearität unmittelbar aus den Überlegungen zu Nächste-Nachbarn-Schätzern. Hier sind lediglich zusätzlich die Gewichte zu berücksichtigen, so dass die lokalen Gewichte entsprechend modifiziert werden müssen.

Als Fazit können wir festhalten, dass sich alle bisher behandelten Glättungsverfahren mit Ausnahme des lokalen Medians als lineare Glätter darstellen lassen. In den weiteren Betrachtungen werden wir uns daher im Wesentlichen auf lineare Glätter beschränken und die meisten Aussagen auch nicht mehr speziell für einzelne Glättungsverfahren herleiten, sondern direkt den allgemeinen Fall linearer Glätter zugrunde legen.

Konfidenzintervalle und -bänder

Neben der Bestimmung von Punktschätzern für die Funktion $f(z)$ sind für eine praktische Beurteilung der Schätzung $\hat{f}(z)$ auch Aussagen über die Variabilität bzw. entsprechende Konfidenzintervalle von Interesse. Betrachtet man zunächst nur die Funktionsschätzung an einem festen Punkt z, so ist die Herleitung einer Varianzformel für lineare Schätzer einfach möglich. Aus der Darstellung $\hat{f}(z) = \boldsymbol{s}(z)'\boldsymbol{y}$ ergibt sich mit $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$

$$\mathrm{Var}(\hat{f}(z)) = \sigma^2\boldsymbol{s}(z)'\boldsymbol{s}(z).$$

Setzt man zusätzlich voraus, dass die Fehler normalverteilt sind, so wird ein Konfidenzintervall zum Niveau α definiert durch

$$\hat{f}(z) \pm z_{1-\alpha/2}\sigma\sqrt{\boldsymbol{s}(z)'\boldsymbol{s}(z)}, \tag{7.13}$$

wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung bezeichnet. Genauer liegt dieser Konstruktion die Annahme

$$\hat{f}(z) - f(z) \overset{a}{\sim} \mathrm{N}(0, \sigma^2\boldsymbol{s}(z)'\boldsymbol{s}(z))$$

zugrunde. Entscheidend ist hierbei, dass zwar die Varianz des Schätzers $\hat{f}(z)$ exakt aus der Normalverteilung der Fehler bestimmt werden kann, zur Berechnung des Konfidenzintervalls aber zusätzlich vorausgesetzt wird, dass $\hat{f}(z)$ (approximativ) erwartungstreu ist. In der Regel gilt diese Aussage jedoch nur asymptotisch, so dass auch Aussagen zur Überdeckungswahrscheinlichkeit der so konstruierten Konfidenzintervalle nur asymptotische Gültigkeit besitzen.

Häufig wird das in (7.13) definierte Konfidenzintervall verwendet, um in Abbildungen der Funktionsschätzung $\hat{f}(z)$ einen Eindruck von der Variabilität der Schätzung zu vermitteln. Dazu werden neben der Funktionsschätzung selbst die obere und untere Grenze des Konfidenzintervalls gegen die Kovariable z abgetragen (vergleiche Abbildung 7.24). Bei der Interpretation ist jedoch zu beachten, dass die Überdeckungswahrscheinlichkeit bei punktweise konstruierten Konfidenzintervallen eben auch nur punktweise eingehalten wird. Genauer gilt für ein punktweises Konfidenzintervall der Form $[L(z), U(z)]$ zum Niveau α die Aussage

$$P\Big(L(z) \leq f(z) \leq U(z)\Big) \geq 1-\alpha \text{ für ein fest vorgegebenes } z,$$

d.h. die Wahrscheinlichkeit, dass das (zufällige) Intervall $[L(z), U(z)]$ den (festen) wahren Funktionswert überdeckt, beträgt mindestens $1-\alpha$. Entscheidend ist, dass die Aussage nur für ein fest vorgegebenes z gilt und nicht simultan für mehrere Kovariablenausprägungen. Häufig sind jedoch gerade solche simultanen Aussagen von Interesse, so dass die Betrachtung von (simultanen) *Konfidenzbändern* notwendig wird.

Gesucht ist also ein simultanes Konfidenzband $[L(z), U(z)]$ zum Niveau $1-\alpha$ mit

$$P\Big(L(z) \leq f(z) \leq U(z) \text{ für alle } z \in \{z_1, \ldots, z_r\}\Big) \geq 1-\alpha. \tag{7.14}$$

Hier gilt die Überdeckungswahrscheinlichkeit also nicht punktweise, sondern gleichzeitig für alle Punkte der Menge $\{z_1, \ldots, z_r\}$. Eine einfache Möglichkeit, eine solche simultane Wahrscheinlichkeitsaussage zu erreichen, ergibt sich aus der sogenannten *Bonferroni-Korrektur* des Niveaus. Diese Korrektur lässt sich aus der Bonferroni-Abschätzung motivieren, die für die simultane Wahrscheinlichkeit von Ereignissen $A_1, \ldots, A_r$ die Abschätzung

$$\mathrm{P}(A_1 \cap \ldots \cap A_r) \leq \sum_{j=1}^{r} \mathrm{P}(A_j)$$

liefert. Im vorliegenden Fall bezeichnet A_j das Ereignis „Die Funktionsschätzung $f(z_j)$ ist nicht im Intervall $[L(z_j), U(z_j)]$ enthalten“. Ziel ist es, die Wahrscheinlichkeit, dass (mindestens) eines dieser Ereignisse eintritt, durch α zu begrenzen. Mit Hilfe der Bonferroni-Abschätzung lässt sich dies erreichen, indem man punktweise Konfidenzintervalle zu Niveaus $\alpha_1, \ldots, \alpha_r$ verwendet, für die $\sum_j \alpha_j \leq \alpha$ gilt. Die naheliegendste Möglichkeit, solche α_j zu wählen, ist die gleichmäßige Aufteilung des Konfidenzniveaus durch $\alpha_j = \alpha/r$.

Mit Hilfe der Bonferroni-Korrektur erhält man also eine einfache Möglichkeit, simultane Konfidenzbänder zu definieren. Allerdings sind diese in der Regel relativ weit, so dass zwar das Niveau eingehalten wird, die Aussagen des Konfidenzbands aber relativ ungenau sind. Wir wollen deshalb noch eine zweite Möglichkeit zur Konstruktion von Konfidenzbändern betrachten, die zwar numerisch aufwändiger ist, aber deutlich genauere Aussagen im Sinne schmalerer Konfidenzbänder ermöglicht. Dazu bestimmen wir zunächst die Kovarianzmatrix des Vektors $\hat{\boldsymbol{f}}_r = (\hat{f}(z_1), \ldots, \hat{f}(z_r))'$, der sich für einen linearen Glätter schreiben lässt als $\hat{\boldsymbol{f}}_r = \boldsymbol{S}_r \boldsymbol{y}$, wobei $\boldsymbol{S}_r$ aus den Vektoren $\boldsymbol{s}(z_1), \ldots, \boldsymbol{s}(z_r)$ zusammengesetzt wird. Für die Kovarianzmatrix ergibt sich direkt der Ausdruck

$$\mathrm{Cov}(\hat{\boldsymbol{f}}_r) = \sigma^2 \boldsymbol{S}_r \boldsymbol{S}_r'$$

und damit

$$\hat{\boldsymbol{f}}_r - \boldsymbol{f}_r \overset{a}{\sim} \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{S}_r \boldsymbol{S}_r'). \tag{7.15}$$

Aus dieser gemeinsamen Verteilung erhält man ein simultanes Konfidenzband zum Niveau $1-\alpha$ durch

$$\hat{f}(z_j) \pm \sigma \cdot m_{1-\alpha} \sqrt{\boldsymbol{s}(z_j)' \boldsymbol{s}(z_j)},$$

wobei $m_{1-\alpha}$ das $(1-\alpha)$-Quantil der Verteilung von

$$\max_{1 \leq j \leq r} \left| \frac{\hat{f}(z_j) - f(z_j)}{\sigma \sqrt{\boldsymbol{s}(z_j)' \boldsymbol{s}(z_j)}} \right| \tag{7.16}$$

bezeichnet. In der Verteilung (7.16) wird die maximale standardisierte absolute Abweichung zwischen dem wahren Funktionswert und der Schätzung betrachtet. Da diese Verteilung und ihre Quantile analytisch nur schwer zugänglich sind, bietet sich eine Bestimmung per Simulation an. Dazu zieht man N Zufallsvektoren aus der asymptotischen Verteilung (7.15) und berechnet daraus die entsprechenden N Realisationen aus (7.16).

Konfidenzintervalle und -bänder für lineare Glätter

1. Punktweise Konfidenzintervalle zum Niveau α:

$$\hat{f}(z) \pm z_{1-\alpha/2}\sigma\sqrt{\boldsymbol{s}(z)'\boldsymbol{s}(z)}.$$

   Dabei bezeichnet $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung.
2. Simultanes Konfidenzband für $\{z_1, \ldots, z_j, \ldots, z_r\}$ mit Bonferroni-Korrektur:

$$\hat{f}(z_j) \pm z_{1-\alpha/(2r)}\sigma\sqrt{\boldsymbol{s}(z_j)'\boldsymbol{s}(z_j)}.$$

3. Simultanes Konfidenzband für $\{z_1, \ldots, z_j, \ldots, z_r\}$ basierend auf der gemeinsamen Verteilung von $(f(z_1), \ldots, f(z_r))'$:

$$\hat{f}(z_j) \pm \sigma \cdot m_{1-\alpha}\sqrt{\boldsymbol{s}(z_j)'\boldsymbol{s}(z_j)}.$$

   Das Quantil $m_{1-\alpha}$ von (7.16) wird per Simulation bestimmt.

Die gleichen Formeln können verwendet werden, wenn σ^2 durch einen konsistenten Schätzer $\hat{\sigma}^2$ ersetzt wird.

Ist N groß genug gewählt, so ergibt sich durch das entsprechende empirische $(1-\alpha)$-Quantil ein Schätzer für $m_{1-\alpha}$.

Zur tatsächlichen Bestimmung aller drei verschiedenen Typen von Konfidenzintervallen bzw. -bändern benötigen wir die Fehlervarianz σ^2. In praktischen Anwendungen ist diese in der Regel nicht bekannt, so dass die Konfidenzintervalle entsprechend angepasst werden müssen. Da die Überdeckungsaussagen jedoch ohnehin nur asymptotisch gültig sind, kann σ^2 durch einen konsistenten Schätzer $\hat{\sigma}^2$ ersetzt werden, ohne dass weitere Änderungen an der Definition der Konfidenzintervalle notwendig sind. Einen möglichen Schätzer für σ^2 werden wir im folgenden Abschnitt kennenlernen.

Beispiel 7.4 Konfidenzbänder für den Alterseffekt

Die Unterschiede zwischen punktweisen Konfidenzintervallen und simultanen Konfidenzbändern sollen am Beispiel des Alterseffekts für den Tansania-Datensatz verdeutlicht werden. Zur Schätzung haben wir einen kubischen P-Spline mit 20 inneren Knoten und einer auf zweiten Differenzen beruhenden Bestrafung verwendet. Der Glättungsparameter wurde auf $\lambda = 20$ gesetzt und ein Niveau von $\alpha = 0.05$ zugrunde gelegt. Ziel ist die Konstruktion eines Konfidenzbereiches für die gesamte Funktion, so dass wir der Berechnung der Konfidenzbänder alle 60 beobachteten Kovariablenausprägungen zugrunde legen.

Abbildung 7.24 zeigt die resultierende Funktionsschätzung gemeinsam mit den drei verschieden konstruierten Konfidenzbereichen. Die punktweisen Konfidenzintervalle bilden den engsten Konfidenzbereich, da sie nur die Unsicherheit der Schätzung an jeweils einem Punkt berücksichtigen. Das andere Extrem ergibt sich bei Verwendung der Bonferroni-Korrektur, die zum weitesten Konfidenzbereich führt. Hier haben wir das Niveau gleichmäßig auf alle 60 beobachteten Kovariablenausprägungen aufgeteilt und erhalten so jeweils ein Niveau von $\alpha_j = 0.0008333$. Damit basieren die Bonferroni-

Abb. 7.24. *Punktweises Konfidenzintervall (– – –), simultanes Konfidenzband basierend auf der gemeinsamen Verteilung (- - -) und simultanes Konfidenzband basierend auf der Bonferroni-Korrektur (· · ·) für den Alterseffekt im Tansania-Beispiel.*

Konfidenzbänder auf dem 0.9995833-Quantil der Standardnormalverteilung, das sich zu 3.34 ergibt. Im Vergleich zum punktweisen Konfidenzintervall, das auf dem 0.975-Quantil der Standardnormalverteilung $z_{0.975} \approx 1.96$ beruht, ergibt sich also ein circa 1.7 mal breiterer Konfidenzbereich.

Zur Konstruktion des Konfidenzbereichs basierend auf der gemeinsamen Verteilung der 60 verschiedenen geschätzten Funktionswerte müssen wir das entsprechende Quantil der Verteilung aus (7.16) bestimmen. Eine Simulation mit $N = 100.000$ Wiederholungen liefert $m_{0.95} = 2.66$, so dass sich (wie erwartet) ein zwischen dem punktweisen Konfidenzintervall und der Bonferroni-Korrektur liegendes Konfidenzband ergibt.

△

Äquivalente Freiheitsgrade (effektive Parameterzahl)

Eine Gemeinsamkeit aller bisher betrachteten Glättungsverfahren ist die Tatsache, dass die Glattheit der Funktionsschätzung durch (mindestens) einen Glättungsparameter gesteuert wird. Für Basisfunktionenansätze geschah dies im Wesentlichen durch die Zahl der Basisfunktionen, während für Penalisierungsansätze ein zusätzlicher Parameter eingeführt wurde, der den Einfluss des Strafterms steuert. Für Nächste-Nachbarn-Schätzer ergaben sich die Anzahl nächster Nachbarn bzw. die Ordnung der symmetrischen Nachbarschaft als die die Glattheit bestimmenden Größen. Offenbar sind die Glättungsparameter verschiedener Verfahren aber nicht unmittelbar vergleichbar, so dass eine allgemeine Größe zur Beurteilung der Glattheit von Funktionsschätzungen wünschenswert wäre.

Am einfachsten lässt sich eine solche Größe in Analogie zur linearen Regression herleiten. Dort ist die Komplexität eines Modells gegeben durch die Anzahl der im Modell vorhandenen Regressionsparameter. Diese Anzahl spiegelt sich unter anderem in der Spur der Prädiktionsmatrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ wieder, da $\mathrm{sp}(\boldsymbol{H}) = p$ gilt. Da die Prädiktionsmatrix des linearen Modells gerade der Glättungsmatrix $\boldsymbol{S}$ eines linearen Glätters entspricht, definiert man in Analogie die *äquivalenten Freiheitsgrade* (*equivalent degrees of freedom*) durch

Abb. 7.25. *Äquivalente Freiheitsgrade als Funktion des Glättungsparameters. Für P-Splines ist der logarithmierte Glättungsparameter aufgetragen. Die gestrichelten Linien geben die sich als Grenzfälle ergebenden äquivalenten Freiheitsgrade an.*

$$\mathrm{df}(\boldsymbol{S}) = \mathrm{sp}(\boldsymbol{S}). \tag{7.17}$$

Die äquivalenten Freiheitsgrade werden häufig auch als *effektive Parameterzahl* eines Glätters interpretiert. Tatsächlich erhält man für Basisfunktionenansätze ohne Penalisierung, wie im linearen Modell, die Anzahl der Parameter im Modell zurück, während für Penalisierungsansätze die effektive Parameterzahl durch den Strafterm heruntergewichtet wird.

Abbildung 7.25 zeigt für einige Glättungsverfahren den Zusammenhang zwischen dem natürlichen Glättungsparameter, mit dem die Verfahren eingeführt wurden, und den resultierenden effektiven Freiheitsgraden für das simulierte Datenbeispiel. Für den gleitenden Mittelwert (arithmetisches Mittel) führt, wie erwartet, eine größere Nachbarschaft zu einer glatteren Schätzung und damit zu einer geringeren Zahl effektiver Parameter. Als Grenzfall erhält man genau einen Parameter, wenn die Nachbarschaft alle Beobachtungen umfasst. Für Polynom-Splines entspricht die effektive Parameterzahl genau der Zahl verwendeter Basisfunktionen, so dass sich hier ein linearer Zusammenhang ergibt. Betrachtet man penalisierte Splines, so verringert sich die effektive Parameterzahl mit steigendem Glättungsparameter (der Glättungsparameter ist hier auf einer logarithmierten Skala wiedergegeben, um den Verlauf der Kurve zu verdeutlichen). Für $\lambda = 0$ erhält man den unpenalisierten Fall zurück, so dass sich die Anzahl verwendeter Basisfunktionen ergibt. Für einen sehr großen Glättungsparameter nähern sich die effektiven Freiheitsgrade offenbar einem Grenzwert, der von der verwendeten Differenzenordnung abhängt. In Abschnitt 7.1.2 haben wir gesehen, dass sich für $\lambda \to \infty$ und Penalisierung durch k-te Differenzen ein Polynom vom Grad $k-1$ ergibt. Ein solches Polynom wird gerade durch

k Parameter beschrieben, so dass man in Abbildung 7.25 als Grenzfall genau einen bzw. zwei Parameter erhält.

Will man umgekehrt zu vorgegebenen effektiven Freiheitsgraden die entsprechenden Glättungsparameter wählen, so lassen sich diese prinzipiell aus Abbildung 7.25 ablesen. Dabei ist allerdings zu beachten, dass der Glättungsparameter für Polynom-Splines und den lokalen Mittelwert nur eine endliche Menge von Werten annehmen kann. Somit ist nicht jeder Wert der effektiven Freiheitsgrade tatsächlich erreichbar. Für P-Splines kann man hingegen jeden Wert für die effektiven Freiheitsgrade wählen, der zwischen der Differenzenordnung k und der Anzahl der verwendeten Basisfunktionen liegt.

Die Definition äquivalenter Freiheitsgrade in (7.17) ist natürlich nicht die einzige Möglichkeit, die Komplexität nichtparametrischer Funktionsschätzungen zu charakterisieren. Es existieren noch zwei weitere Vorschläge, die sich ebenfalls aus Analogien zum linearen Modell ableiten lassen, aber seltener verwendet werden. Diese sind definiert durch

$$\mathrm{df}_{var}(\boldsymbol{S}) = \sum_{i=1}^{n} \mathrm{Var}(\hat{f}(z_i)) = \mathrm{sp}(\boldsymbol{S}\boldsymbol{S}')$$

und

$$\mathrm{df}_{res}(\boldsymbol{S}) = \mathrm{sp}(2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{S}')$$

und ergeben sich aus Überlegungen zur Summe der Varianzen der vorhergesagten Werte bzw. zum Erwartungswert des mittleren quadratischen Fehlers (vergleiche Hastie & Tibshirani (1990), Kap. 3.5 für ausführliche Herleitungen). Falls die Glättungsmatrix $\boldsymbol{S}$ symmetrisch und idempotent ist (also $\boldsymbol{S} = \boldsymbol{S}'$ und $\boldsymbol{S}^2 = \boldsymbol{S}$ gilt), fallen die Definitionen von df, df_{var} und df_{res} zusammen. Dies ist beispielsweise für die einfachen Basisfunktionenansätze der Fall, gilt aber bereits für Penalisierungsansätze nicht mehr.

In der Praxis wird die erste Definition effektiver Freiheitsgrade aufgrund der einfacheren Berechnung meist bevorzugt. Insbesondere kann man häufig die Berechnung der vollen $n \times n$ Matrix $\boldsymbol{S}$ vermeiden, da sich innerhalb der Spur Matrizen verschieben lassen (vergleiche Satz A.15 in Anhang A). Beispielsweise ergibt sich für P-Splines die Formel

$$\mathrm{df}(\boldsymbol{S}) = \mathrm{sp}(\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}') = \mathrm{sp}(\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}),$$

so dass nur noch die Spur eines Produkts aus zwei Matrizen berechnet werden muss, deren Dimension jeweils durch die Anzahl der verwendeten Basisfunktionen gegeben ist.

Schätzung der Fehlervarianz

Im linearen Modell konnte die Fehlervarianz σ^2 erwartungstreu geschätzt werden durch

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

In der nonparametrischen Regression liegt es nahe, die Anzahl der Parameter p durch die äquivalenten Freiheitsgrade aus dem vorherigen Abschnitt zu ersetzen. Tatsächlich kann man zeigen, dass sich unter der Annahme eines korrekt spezifizierten Modells der Erwartungswert der Residuenquadratsumme schreiben lässt als

$$\mathrm{E}\left(\sum_{i=1}^{n}(y_i - \hat{f}(z_i))^2\right) = (n - \mathrm{sp}(2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{S}'))\sigma^2 + \sum_{i=1}^{n} b_i^2,$$

wobei b_i die Verzerrung der Funktionsschätzung am Punkt z_i bezeichnet. Kann die Verzerrung vernachlässigt werden, ist der betrachtete Glätter also approximativ erwartungstreu, so erhält man den erwartungstreuen Schätzer

$$\hat{\sigma}^2 = \frac{1}{n - \mathrm{df}_{res}} \sum_{i=1}^{n}(y_i - \hat{f}(z_i))^2$$

für die Fehlervarianz. Obwohl die Erwartungstreue für $\hat{\sigma}^2$ streng genommen nur bei Verwendung von df_{res} gesichert ist, wird in der Praxis typischerweise auch die einfachere Definition df zur Schätzung der Fehlervarianz verwendet.

Bias-Varianz-Trade Off

In diesem Abschnitt untersuchen wir genauer, wie sich die Variation von Parametern des Glättungsverfahrens auf den mittleren quadratischen Fehler der Funktionsschätzung auswirkt. Der Einfachheit halber beschäftigen wir uns zunächst mit dem lokalen Mittelwert-Schätzer

$$\hat{f}(z_i) = \frac{1}{k} \sum_{j \in N(z_i)} y_j$$

basierend auf den k nächsten Nachbarn. Für diesen Schätzer lassen sich Erwartungswert und Varianz einfach berechnen und ergeben sich zu

$$\mathrm{E}(\hat{f}(z_i)) = \frac{1}{k} \sum_{j \in N(z_i)} f(z_j)$$

und

$$\mathrm{Var}(\hat{f}(z_i)) = \frac{\sigma^2}{k}.$$

Da sich der MSE additiv aus quadrierter Verzerrung und Varianz zusammensetzt, könnte man versuchen, den Glättungsparameter k so zu wählen, dass Verzerrung und Varianz gleichzeitig minimiert werden. Obigen Formeln zeigen aber, dass gerade dieses Vorgehen nicht möglich ist, da die Variation von k auf Verzerrung und Varianz gegenläufige Auswirkungen hat. Vergrößert man die Nachbarschaft, so verringert sich offensichtlich die Varianz. Andererseits ergibt sich tendenziell eine größere Verzerrung, da immer mehr Terme im Erwartungswert von $\hat{f}(z_i)$ zu berücksichtigen sind, die von $f(z_i)$ verschieden sind. Für eine Nachbarschaft der Größe $k = 1$ mit $N(z_i) = \{i\}$ ergibt sich ein unverzerrter Schätzer ohne Bias, der allerdings andererseits die größtmögliche Varianz besitzt. Ein ähnliches Dilemma haben wir bereits in Abschnitt 3.6 im Rahmen der Modellwahl in linearen Modellen kennengelernt und als klassisches Beispiel des Bias-Varianz-Trade Off identifiziert.

Üblicherweise führt die Wahl eines glatteren Schätzverfahrens mit einer kleineren Zahl effektiver Freiheitsgrade zu einer weniger variablen, aber stärker verzerrten Schätzung. Umgekehrt führt die Verwendung eines komplexen Modells mit einer großen Zahl effektiver Freiheitsgrade zu einer geringeren Verzerrung, aber großer Variabilität. Man

betrachte hierzu auch noch einmal die Grafiken, die wir im Verlauf dieses Kapitels zum Einfluss des Glättungsparameters verschiedener Glättungsverfahren gesehen haben und die genau dieses Phänomen widerspiegeln (etwa die Abbildungen 7.10 auf Seite 302, 7.16 auf Seite 315 oder 7.22 auf Seite 336). Im Gegensatz zum lokalen Mittelwertschätzer ist die explizite Herleitung von Erwartungswert und Varianz der meisten Streudiagramm-Glätter jedoch komplizierter, so dass wir an dieser Stelle als weiteres Beispiel nur noch lokal polynomiale Modelle betrachten wollen, für die zumindest asymptotische Ergebnisse vorliegen (vergleiche Fahrmeir & Tutz (2001), Kap. 5.1.3).

Die Glattheit wird hier im Wesentlichen durch die Bandweite λ des verwendeten Kerns gesteuert, so dass wir nun den Einfluss von λ auf Verzerrung und Varianz des Schätzers untersuchen wollen. Dabei werden wir uns auf asymptotische Betrachtungen beschränken, d.h. wir werden das Verhalten des Schätzers für großen Stichprobenumfang $n \to \infty$ betrachten. Hierfür ist zu beachten, dass für sinnvolle asymptotische Aussagen nicht nur Voraussetzungen an den Stichprobenumfang zu stellen sind, sondern auch der Glättungsparameter geeignet variiert werden muss, da mit größerem Stichprobenumfang automatisch auch eine größere Schätzgenauigkeit erzielt wird und damit eine geringere Bandweite notwendig ist. Genauer betrachten wir im Folgenden den Grenzübergang $\lambda \to 0$ und $n\lambda \to \infty$ d.h. wir lassen die Bandweite zwar gegen Null gehen, gleichzeitig wächst aber der Stichprobenumfang so schnell an, dass $n\lambda$ immer noch divergiert. Unter diesen Voraussetzungen ergibt sich für die asymptotische Verzerrung die Approximation

$$\mathrm{E}(\hat{f}(z) - f(z)) \approx \frac{\lambda^{l+1} f^{(l+1)}(z)}{(l+1)!} \mu_{l+1}(K)$$

bei ungeradem Polynomgrad l und

$$\mathrm{E}(\hat{f}(z) - f(z)) \approx \left(\frac{\lambda^{l+2} f^{(l+1)}(z) d_z'(z)}{(l+1)! d_z(z)} + \frac{\lambda^{l+2} f^{(l+2)}(z)}{(l+2)!} \right) \mu_{l+2}(K)$$

bei geradem Polynomgrad. Dabei bezeichnen $d_z(z)$ bzw. $d_z'(z)$ die Dichte der Verteilung von z und die zugehörige erste Ableitung, sowie $\mu_q(K) = \int u^q K(u) du$ das q-te zentrale Moment des verwendeten Kerns. Für die Varianz erhält man die asymptotische Approximation

$$\mathrm{Var}(\hat{f}(z)) \approx \frac{\sigma^2}{n\lambda d_z(z)} \int K^2(u) du.$$

Aus diesen beiden Formeln lassen sich eine Reihe interessanter Aussagen ablesen, die wieder den angesprochenen Trade Off zwischen Verzerrung und Variabilität nichtparametrischer Regressionsschätzer illustrieren. Zunächst ergibt sich für den Bias:

- Durch eine Verkleinerung der Bandweite λ verringert sich die Verzerrung von $\hat{f}(z)$.
- Für gerades l hängt die Verzerrung von der Verteilung von z (ausgedrückt durch die Dichte $d_z(z)$) ab. Es ergibt sich eine geringere Verzerrung wenn $d_z(z)$ einen großen Wert annimmt, also eine große Zahl von Beobachtungen in der Umgebung von z erwartet wird. Umgekehrt wird die Verzerrung in Bereichen mit geringer erwarteter Anzahl von Beobachtungen (also kleiner Dichte $d_z(z)$) vergrößert.
- Für ungerades l ist die Verzerrung unabhängig von der Verteilung der Kovariablen.
- Für $l = 0$ hängt die Verzerrung von der ersten und zweiten Ableitung der Funktion $f(z)$ ab. Offenbar ergibt sich ein größerer Bias in Bereichen mit großer Steigung $f'(z)$.

Da die zweite Ableitung ein Maß für die Krümmung der Funktion ist, erhält man darüber hinaus eine größere Verzerrung in Bereichen starker Krümmung. Genauer ergibt sich eine Unterschätzung (also eine negative Verzerrung) in lokalen Maxima und eine Überschätzung (also eine positive Verzerrung) in lokalen Minima.

- Für gerades l ergibt sich ein zusätzlicher Term in der Formel für die Verzerrung. Generell empfiehlt sich daher die Verwendung eines ungeraden Polynomgrads.

Für die Varianz erhalten wir die folgenden Aussagen:

- Durch die Vergrößerung des Glättungsparameters ergibt sich eine geringere Variabilität der Schätzung.
- Die Varianz hängt sowohl für gerades als auch für ungerades l von der Verteilung der Kovariablen ab. Analog zur Verzerrung ergibt sich für großes $d_z(z)$ eine Verringerung der Varianz und für kleines $d_z(z)$ eine Vergrößerung der Varianz (allerdings ist die Form des Einflusses für Bias und Varianz verschieden).

Zusammenfassend ist es also wieder nicht möglich, Varianz und Verzerrung gleichzeitig zu verkleinern. Einen geeigneten Kompromiss bietet daher die Verwendung des mittleren quadratischen Fehlers. Insbesondere könnte dieser zur Wahl eines optimalen Glättungsparameters herangezogen werden. Da der MSE aber im Allgemeinen von der unbekannten, wahren Funktion $f(z)$ abhängt, ist dieses Vorgehen nicht praktikabel. Es bildet jedoch die Grundlage für eine Strategie zur Wahl des Glättungsparameters, die auf einer Approximation des quadratischen Fehlers beruht.

7.1.9 Wahl des Glättungsparameters

Aus den bisherigen Betrachtungen zu nichtparametrischen Regressionsmodellen ergibt sich eine wesentliche verbleibende Fragestellung: Wie wählt man den Glättungsparameter, so dass sich eine optimale Beschreibung der Daten ergibt? Basierend auf den Überlegungen zum Bias-Varianz-Trade Off im letzten Abschnitt kann man diese Frage auch umformulieren in das Problem, einen geeigneten Kompromiss zwischen Verzerrung und Variabilität des Schätzers zu erhalten. Eine direkte Möglichkeit bietet die Betrachtung des mittleren quadratischen Fehlers, die im Mittelpunkt des ersten Abschnitts zur Glättungsparameterwahl stehen wird. Darüber hinaus werden wir hier kurz die Verwendung des Informationskriteriums AIC zur Glättungsparameterwahl diskutieren. Der zweite und dritte Abschnitt beschränken sich dann auf Penalisierungsansätze und beschreiben hier allgemeine Möglichkeiten, Glättungsparameter zu bestimmen. Dabei wird zum einen die Ähnlichkeit zwischen Penalisierungsansätzen und gemischten Modellen ausgenutzt, und zum anderen werden Bayesianische Markov-Chain-Monte-Carlo-Simulationsverfahren hergeleitet.

Glättungsparameterwahl basierend auf Optimalitätskriterien

Eine erste Möglichkeit zur Wahl des Glättungsparameters ergibt sich aus den Überlegungen zum Bias-Varianz-Trade Off. Dort hatten wir festgestellt, dass sowohl die Verzerrung als auch die Varianz von Glättungsverfahren vom Glättungsparameter abhängen, aber nicht simultan verringert werden können, so dass ein geeigneter Ausgleich gefunden werden muss. Einen solchen Kompromiss erhält man durch die Betrachtung des mittleren quadratischen Fehlers

$$\mathrm{MSE}(\hat{f}(z)) = \mathrm{E}\left((\hat{f}(z) - f(z))^2\right) = \left(\mathrm{E}(\hat{f}(z) - f(z))\right)^2 + \mathrm{Var}(\hat{f}(z)),$$

der sich additiv aus quadrierter Verzerrung und Varianz zusammensetzt. Um aus dem punktweisen MSE ein Maß für die Güte der gesamten Funktionsschätzung zu erhalten, betrachten wir den über die beobachteten Kovariablenausprägungen gemittelten MSE

$$\frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(\hat{f}(z_i) - f(z_i))^2\right).$$

Eine naive Approximationsmöglichkeit ergibt sich durch die Residuenquadratsumme

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(z_i))^2.$$

Diese ist allerdings nicht dazu geeignet, einen optimalen Glättungsparameter zu bestimmen, da sie sich durch $\hat{f}(z_i) = y_i$ minimieren lässt. Damit ergäbe sich durch Optimierung der Residuenquadratsumme immer der triviale, interpolierende Schätzer (vergleiche hierzu auch Abschnitt 3.6).

Daher geht man zur Glättungsparameterwahl zum quadratischen Fehler der Prognose für neue Beobachtungen y^* über. Da üblicherweise allerdings keine solchen neuen Beobachtungen erhältlich sind, verwendet man die sich durch Kreuzvalidierung ergebende Approximation. Dazu wird jeweils eine der Beobachtungen aus den Daten gelöscht, die Schätzung für den gegebenen Glättungsparameter basierend auf den verbleibenden $n-1$ Beobachtungen durchgeführt und dann der Funktionswert $f(z_i)$ für die gelöschte Beobachtung aus dieser Schätzung vorhergesagt. Bezeichnen wir mit $\hat{f}^{(-i)}(z)$ die Funktionsschätzung, die sich ohne die Beobachtung (z_i, y_i) ergibt, so erhält man daraus das Kreuzvalidierungskriterium

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

Die Minimierung des CV-Kriteriums ergibt dann einen im Sinne des Vorhersagefehlers optimalen Glättungsparameter. Die Verwendung des CV-Kriteriums lässt sich auch theoretisch rechtfertigen, da

$$\mathrm{E}(\mathrm{CV}) \approx \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{f}(z_i))^2$$

gilt, wobei y_i^* jeweils neue Beobachtungen an den Punkten z_i bezeichne. Im Erwartungswert ergibt sich über das CV-Kriterium also gerade der mittlere quadratische Vorhersagefehler.

Auf den ersten Blick sind zur Bestimmung des optimalen Glättungsparameters n separate Schätzungen nichtparametrischer Regressionsmodelle durchzuführen. Für Penalisierungsansätze kann man jedoch zeigen, dass sich wie im linearen Modell (vergleiche Abschnitt 3.6) $\hat{f}^{(-i)}(z_i)$ bestimmen lässt, ohne die zugehörige Schätzung durchzuführen. Dazu benötigen wir die Diagonalelemente s_{ii} der Glättungsmatrix $\boldsymbol{S}$, die der Prädiktionsmatrix in Abschnitt 3.6 entspricht. Damit erhält man

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(z_i)}{1 - s_{ii}}\right)^2,$$

so dass sich das Kreuzvalidierungskriterium bereits mit Hilfe der Schätzung $\hat{f}(\cdot)$ und der zugehörigen Glättungsmatrix bestimmen lässt. Diese Aussage gilt zwar nicht für alle möglichen Streudiagramm-Glätter, wird aber dennoch im Allgemeinen zur Approximation des CV-Kriteriums herangezogen. Die Berechnung der Glättungsmatrix bzw. ihrer Diagonalelemente ist allerdings (insbesondere für große Datensätze) immer noch numerisch aufwändig. Daher ersetzt man häufig die Diagonalelemente durch ihren Mittelwert, so dass sich das generalisierte Kreuzvalidierungs-Kriterium (GCV) ergibt:

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(z_i)}{1 - \mathrm{sp}(\boldsymbol{S})/n}\right)^2 .$$

Die Summe der Diagonalelemente entspricht genau der ersten Definition, die wir für äquivalente Freiheitsgrade kennengelernt haben, nämlich der Spur der Glättungsmatrix. Beispielsweise haben wir für P-Splines gesehen, dass sich diese Spur einfach berechnen lässt, da Matrizen innerhalb der Spur verschoben werden dürfen (vergleiche Seite 347). Neben der einfacheren Berechenbarkeit besitzt das GCV-Kriterium im Gegensatz zum CV-Kriterium aber auch den theoretischen Vorteil, invariant unter orthogonalen Transformationen der Daten zu sein (vergleiche Wood (2006), Kap. 4.5.2).

Alternativ zum generalisierten Kreuzvalidierungskriterium können auch andere aus der Modellwahl bekannte Kriterien zur Bestimmung des Glättungsparameters herangezogen werden. Ein typisches Beispiel ist Akaikes Informationskriterium AIC, das auch in Abschnitt 3.6 diskutiert wurde. Dort hing das AIC von der Zahl der Parameter im Modell ab, so dass die Likelihood geeignet bezüglich der Modellkomplexität korrigiert wurde (vergleiche auch Abschnitt B.4.5). In nichtparametrischen Regressionsmodellen ersetzt man die tatsächliche Parameterzahl durch die effektive Parameterzahl und erhält damit das Kriterium

$$\mathrm{AIC} = n\log(\hat{\sigma}^2) + 2(\mathrm{df} + 1),$$

das bezüglich des Glättungsparameters zu minimieren ist. Dabei bezeichnet $\hat{\sigma}^2 = \sum(y_i - \hat{f}(z_i))^2/n$ den ML-Schätzer der Fehlervarianz.

Abbildung 7.26 zeigt jeweils für einen kubischen P-Spline mit 20 inneren Knoten und auf zweiten Differenzen basierender Penalisierung den Verlauf des GCV und des AIC für den Tansania-Datensatz zusammen mit den resultierenden optimalen Schätzungen. Um eine einfachere Lesbarkeit der Verläufe zu Erreichen wurde dabei jeweils der minimale Werte des AIC bzw. des GCV auf Null gesetzt, so dass die Kurven den Unterschied zum optimalen Modell wiedergeben. Dabei fällt zunächst auf, dass GCV und AIC in Abhängigkeit von λ einen sehr ähnlichen Verlauf zeigen. Beide Größen sind tatsächlich asymptotisch äquivalent, in einzelnen Anwendungen können aber dennoch deutliche Unterschiede auftreten. Im vorliegenden Beispiel liefern beide Kriterien ähnliche, aber nicht vollständig identische, optimale Glättungsparameter und damit auch sehr ähnliche geschätzte Funktionsverläufe.

Wie lässt sich nun konkret der optimale Glättungsparameter für eines der zur Verfügung stehenden Kriterien bestimmen? Da wir uns bisher nur mit dem einfachsten Beispiel, der univariaten nichtparametrischen Regression beschäftigt haben, ist es möglich, die Optimierung direkt durch eine Gittersuche (oder effizientere Methoden zur numerischen Minimierung) durchzuführen. Dazu gibt man sich ein Gitter von möglichen Glättungsparametern vor, berechnet jeweils das verwendete Optimalitätskriterium und wählt dann aus dem Gitter den besten Wert aus. Dabei kann es sinnvoll sein, das Gitter

Abb. 7.26. *GCV und AIC im Tansania-Beispiel (links) und die sich aus dem jeweils optimalen Glättungsparameter ergebenden Schätzungen (rechts).*

anhand erster Versuche iterativ zu variieren, um eine möglichst genaue Identifikation des optimalen Wertes zu gewährleisten.

In komplexeren Problemen, die wir in Kapitel 8 behandeln werden, ist ein solches Vorgehen jedoch nicht mehr möglich oder zumindest äußerst aufwändig, da hier in der Regel mehrere Glättungsparameter simultan zu optimieren sind. Es existieren jedoch auch dann noch Algorithmen zur effizienten Minimierung des GCV-Kriteriums oder des AIC, die auf Newton-Verfahren, also der Ableitung der Kriterien nach den Glättungsparametern beruhen. Da diese Verfahren jedoch mathematisch und algorithmisch kompliziert sind, wollen wir sie an dieser Stelle nicht im Detail behandeln und verweisen hierzu auf die entsprechende Fachliteratur, insbesondere Wood (2000) und Wood (2006).

Repräsentation von Penalisierungsansätzen als gemischte Modelle

Die im vorigen Abschnitt zur Optimierung von Modellwahlkriterien beschriebenen Verfahren lassen sich allgemein zur Wahl des Glättungsparameters beliebiger Streudiagramm-Glätter verwenden. In den folgenden beiden Abschnitten werden wir uns auf eine spezielle Klasse von Glättungsverfahren einschränken und für diese weitere geeignete Möglichkeiten zur Bestimmung optimaler Glättungsparameter herleiten. Diese Klasse besteht aus Penalisierungsansätzen, deren Strafterm sich in der Form

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

schreiben lässt, also quadratisch ist in den Parametern und durch eine (symmetrische) Strafmatrix $\boldsymbol{K}$ definiert wird. Diese Einschränkung ist allerdings nicht besonders restrik-

tiv, da hierunter alle in den Abschnitten 7.1.2 bis 7.1.6 behandelten Verfahren fallen, also insbesondere penalisierte Splines, Glättungssplines, Random Walks und Kriging.

Gehen wir zunächst von einem P-Spline in TP-Darstellung aus. Dann ist das penalisierte KQ-Kriterium gegeben durch

$$\mathrm{PKQ}(\lambda) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda \sum_{j=l+2}^{d} \gamma_j^2.$$

(vergleiche Seite 308). Um eine Verbindung zu den gemischten Modellen aus Kapitel 6 herzustellen, zerlegen wir zunächst den Parametervektor $\boldsymbol{\gamma}$ in $\boldsymbol{\beta} = (\gamma_1, \ldots, \gamma_{l+1})'$ und $\tilde{\boldsymbol{\gamma}} = (\gamma_{l+2}, \ldots, \gamma_d)'$. Der Vektor $\boldsymbol{\beta}$ besteht also aus den Parametern des Polynoms, die nicht penalisiert werden, während $\tilde{\boldsymbol{\gamma}}$ die Parameter der abgeschnittenen Potenzen beinhaltet, die quadriert in den Strafterm eingehen. Definiert man die Teildesignmatrizen $\boldsymbol{X}$ und $\boldsymbol{U}$, so dass $\boldsymbol{Z} = [\boldsymbol{X}\ \boldsymbol{U}]$ gilt, so lässt sich das penalisierte KQ-Kriterium umformulieren zu

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}}) + \lambda\tilde{\boldsymbol{\gamma}}'\tilde{\boldsymbol{\gamma}}. \tag{7.18}$$

Die Schätzungen in einem gemischten Modell

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad \tilde{\boldsymbol{\gamma}} \sim \mathrm{N}(\boldsymbol{0}, \tau^2\boldsymbol{I}),$$

lassen sich, wie wir in Abschnitt 6.3.1 gesehen haben, ebenfalls durch die Minimierung eines penalisierten KQ-Kriteriums der Form

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}}) + \frac{\sigma^2}{\tau^2}\tilde{\boldsymbol{\gamma}}'\tilde{\boldsymbol{\gamma}}$$

bestimmen. Vergleicht man dieses Kriterium mit (7.18), so kann $\boldsymbol{\beta}$ als ein Vektor fester Effekte und $\tilde{\boldsymbol{\gamma}}$ als ein Vektor zufälliger Effekte in einem gemischten Modell aufgefasst werden. Die Varianz der zufälligen Effekte wird bestimmt durch $\tau^2 = \sigma^2/\lambda$, der Glättungsparameter λ lässt sich also, wie für Bayesianische P-Splines (Seite 316ff), als Verhältnis der Fehlervarianz zur Varianz der zufälligen Effekte interpretieren.

Der Vorteil der Einbettung von P-Splines in den Rahmen gemischter Modelle liegt darin, dass die in Kapitel 6 behandelten Schätzverfahren zur Bestimmung von σ^2 und τ^2 verwendet werden können, um einen optimalen Glättungsparameter zu erhalten. Insbesondere können die Schätzkonzepte aus Abschnitt 6.3.2 angewendet werden, also Maximum-Likelihood- und Restringierte Maximum-Likelihood-Schätzung. Basierend auf den Schätzern $\hat{\sigma}^2$ und $\hat{\tau}^2$ erhält man dann den optimalen Glättungsparameter $\hat{\lambda} = \hat{\sigma}^2/\hat{\tau}^2$.

Für P-Splines in TP-Darstellung ergibt sich also relativ unmittelbar ein Zusammenhang zu gemischten Modellen und damit eine alternative Darstellung. Nun wollen wir diesen Zusammenhang allgemeiner für beliebige Penalisierungsansätze untersuchen. Ausgangspunkt hierfür ist die penalisierte Residuenquadratsumme

$$\mathrm{PKQ}(\lambda) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma},$$

die allen von uns betrachteten Penalisierungsansätzen zugrunde liegt. Aufgrund der bisherigen Überlegungen könnte man nun versucht sein, ein gemischtes Modell der Form

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \tau^2\boldsymbol{K}^{-1}) \tag{7.19}$$

und $\lambda = \sigma^2/\tau^2$ zu definieren. Dabei entsteht das Problem, dass die Inverse $\boldsymbol{K}^{-1}$ im Allgemeinen nicht existiert, da die Strafmatrix $\boldsymbol{K}$ häufig (beispielsweise für auf B-Splines basierende P-Splines) nicht vollen Rang besitzt. Für den Kriging-Ansatz macht die Formulierung (7.19) direkt Sinn, da hier die Strafmatrix vollen Rang besitzt.

Um allgemeine Penalisierungsansätze als gemischte Modelle auffassen zu können, müssen wir also etwas anders vorgehen, als es für P-Splines in TP-Basis möglich war. Das Problem liegt darin begründet, dass sich die aus der Darstellung (7.19) ergebende Dichte der zufälligen Effekte

$$p(\boldsymbol{\gamma}) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right)$$

nicht normieren lässt. Man spricht in diesem Zusammenhang auch davon, dass die Dichte *uneigentlich* ist. Genauer erhält man eine teilweise uneigentliche Dichte, da zwar $\mathrm{rg}(\boldsymbol{K}) > 0$, aber nicht $\mathrm{rg}(\boldsymbol{K}) = \dim(\boldsymbol{\gamma})$ gilt. In einem gemischten Modell ist diese teilweise Uneigentlichkeit aufgelöst in eine nicht informative Verteilung für die festen Effekte und eine eigentliche Normalverteilung mit normierbarer Dichte für die zufälligen Effekte. Ziel ist es also, für beliebige Penalisierungsansätze eine solche Zerlegung herzuleiten (vergleiche hierzu auch Abschnitt B.3.2 in Anhang B).

Dazu müssen wir $\boldsymbol{\gamma}$ in Abhängigkeit vom Rangabfall von $\boldsymbol{K}$ in zwei Teilvektoren $\boldsymbol{\beta}$ und $\tilde{\boldsymbol{\gamma}}$ zerlegen. Bezeichne dazu im Folgenden $r = \mathrm{rg}(\boldsymbol{K})$ den Rang der Strafmatrix und $d = \dim(\boldsymbol{\gamma})$ die Dimension des Koeffizientenvektors. Dann betrachten wir Zerlegungen

$$\boldsymbol{\gamma} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}}$$

mit den $d \times (d-r)$ und $d \times r$ dimensionalen Designmatrizen $\tilde{\boldsymbol{X}}$ und $\tilde{\boldsymbol{U}}$ sowie den $(d-r)$ und r dimensionalen Parametervektoren $\boldsymbol{\beta}$ und $\tilde{\boldsymbol{\gamma}}$. Ziel ist es, die Designmatrizen $\tilde{\boldsymbol{X}}$ und $\tilde{\boldsymbol{U}}$ so zu wählen, dass sich der Strafterm $\lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$ umschreiben lässt zu $\lambda\tilde{\boldsymbol{\gamma}}'\tilde{\boldsymbol{\gamma}}$, so dass $\boldsymbol{\beta}$ als Vektor fester Effekte und $\tilde{\boldsymbol{\gamma}}$ als Vektor zufälliger Effekte aufgefasst werden kann. Genauer erfordert dies die folgenden Eigenschaften der Designmatrizen:

1. $\tilde{\boldsymbol{X}}'\boldsymbol{K} = 0$, so dass $\boldsymbol{\beta}$ nicht durch $\boldsymbol{K}$ penalisiert wird und
2. $\tilde{\boldsymbol{U}}'\boldsymbol{K}\tilde{\boldsymbol{U}} = \boldsymbol{I}$, so dass $\tilde{\boldsymbol{\gamma}}$ aus unabhängigen und identisch verteilten zufälligen Effekten besteht.

Damit ergibt sich nämlich für den Strafterm

$$\begin{aligned}
\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma} &= (\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}})'\boldsymbol{K}(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}}) \\
&= \boldsymbol{\beta}'\underbrace{\tilde{\boldsymbol{X}}'\boldsymbol{K}}_{=\boldsymbol{0}}\tilde{\boldsymbol{X}}\boldsymbol{\beta} + 2\boldsymbol{\beta}'\underbrace{\tilde{\boldsymbol{X}}'\boldsymbol{K}}_{=\boldsymbol{0}}\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\gamma}}'\underbrace{\tilde{\boldsymbol{U}}'\boldsymbol{K}\tilde{\boldsymbol{U}}}_{=\boldsymbol{I}}\tilde{\boldsymbol{\gamma}} \\
&= \tilde{\boldsymbol{\gamma}}'\tilde{\boldsymbol{\gamma}}.
\end{aligned}$$

Die Modellgleichung lässt sich dann ebenfalls zu einem gemischten Modell umformulieren:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = \boldsymbol{Z}(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}}) + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}$$

mit den Designmatrizen $\boldsymbol{X} = \boldsymbol{Z}\tilde{\boldsymbol{X}}$ und $\boldsymbol{U} = \boldsymbol{Z}\tilde{\boldsymbol{U}}$ sowie festen Effekten $\boldsymbol{\beta}$ und zufälligen Effekten $\tilde{\boldsymbol{\gamma}} \sim \mathrm{N}(\boldsymbol{0}, \tau^2\boldsymbol{I}_r)$. Damit lassen sich wie zuvor für P-Splines in TP-Darstellung Schätzverfahren für gemischte Modelle zur Bestimmung der Varianzparameter und damit der Glättungsparameter einsetzen.

Zu klären bleibt noch die Frage, wie man Designmatrizen $\tilde{\boldsymbol{X}}$ und $\tilde{\boldsymbol{U}}$ konstruieren kann, die die Anforderungen 1. und 2. erfüllen. Für die Designmatrix der festen Effekte kann man die gewünschte Orthogonalität zur Strafmatrix einfach erreichen, indem man als Spalten von $\tilde{\boldsymbol{X}}$ eine Basis des Nullraums von $\boldsymbol{K}$ verwendet (vergleiche (A.14) in Anhang A). Für P-Splines mit B-Spline-Basis lässt sich dieser Nullraum einfach charakterisieren, da durch den Differenzenstrafterm Polynome vom Grad $k-1$ nicht bestraft werden. Damit kann $\tilde{\boldsymbol{X}}$ definiert werden als

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} 1 & \kappa_1 & \dots & \kappa_1^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & \kappa_d & \dots & \kappa_d^{k-1} \end{pmatrix},$$

wobei $\kappa_1, \dots, \kappa_d$ die Knoten des Splines bezeichnen.

Die Designmatrix der zufälligen Effekte kann aus der Spektralzerlegung der Strafmatrix abgeleitet werden (vergleiche Satz A.22 in Anhang A). Zerlegt man diese zu $\boldsymbol{K} = \boldsymbol{\Gamma}\boldsymbol{\Omega}_+\boldsymbol{\Gamma}'$, wobei $\boldsymbol{\Omega}_+$ die Matrix der positiven Eigenwerte bezeichnet und $\boldsymbol{\Gamma}$ die orthonormale Matrix der zugehörigen Eigenvektoren, so lässt sich ein geeignetes $\tilde{\boldsymbol{U}}$ durch $\tilde{\boldsymbol{U}} = \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1}$ mit $\boldsymbol{L} = \boldsymbol{\Gamma}\boldsymbol{\Omega}^{1/2}$ definieren. Damit ergibt sich nämlich

$$\tilde{\boldsymbol{U}}'\boldsymbol{K}\tilde{\boldsymbol{U}} = (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'\boldsymbol{L}\boldsymbol{L}'\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1} = \boldsymbol{I}.$$

In bestimmten Situationen lässt sich die Berechnung über die Spektralzerlegung allerdings auch vermeiden. So kann beispielsweise für P-Splines auch $\boldsymbol{L} = \boldsymbol{D}'$ mit der Differenzenmatrix $\boldsymbol{D}$ gewählt werden. Daraus folgt insbesondere, dass die Zerlegung von $\boldsymbol{\gamma}$ nicht eindeutig ist.

Somit haben wir eine Möglichkeit kennengelernt, praktisch beliebige Penalisierungsansätze in gemischte Modelle umzuformulieren und damit ML- bzw. REML-Schätzer für den Glättungsparameter zu erhalten. Wir werden dies auch im folgenden Abschnitt 7.2 für weitere Penalisierungsansätze ausnutzen. Häufig wird als weiterer Vorteil der Darstellung als gemischtes Modell auch angeführt, dass Standard-Software für gemischte Modelle zur Schätzung verwendet werden kann. Dabei ist jedoch zu beachten, dass sich die aus Penalisierungsansätzen entstehenden gemischten Modelle in ihrer Struktur deutlich von den gemischten Modellen aus Kapitel 6 unterscheiden, obwohl sie formal genauso aufgeschrieben werden können. Insbesondere weisen die hier betrachteten gemischten Modelle keine Gruppierungsstruktur auf. Bei der Verwendung von Standardsoftware treten daher häufiger numerische Probleme auf, so dass im Allgemeinen die Schätzung über spezialisierte Software, die auf der Darstellung als gemischtes Modell beruht, vorzuziehen ist. Dies trifft insbesondere auf die in Kapitel 8 betrachteten komplexeren Modelle zu.

Zum Abschluss wollen wir uns noch einem methodischen Problem der Schätzung von Glättungsparametern über gemischte Modelle zuwenden. Aus einer frequentistischen Perspektive betrachtet, sind die Parameter $\boldsymbol{\gamma}$ in Penalisierungsansätzen feste, unbekannte Parameter, die es zu schätzen gilt. Durch die Umformulierung in ein gemischtes Modell wird ein Teil des Vektors $\boldsymbol{\gamma}$ in zufällige Effekte überführt und stellt damit formal keinen (festen) Parameter mehr dar, sondern eine Zufallsgröße. Gemäß der ursprünglichen Modellauffassung dürfte $\tilde{\boldsymbol{\gamma}}$ aber nicht als zufällig betrachtet werden, so dass streng genommen die Darstellung als gemischtes Modell nur als algorithmischer Trick, nicht aber als tatsächliche Umformulierung des Modells zu betrachten wäre. Bayesianisch betrachtet tritt dieses Problem nicht auf, da hier ohnehin alle Parameter als zufällig betrachtet

werden und damit die beiden Darstellungen nur äquivalente Formulierungen der gleichen Priori-Annahmen darstellen. Gemäß der Unterscheidung in Abschnitt 6.5 sind die sich aus der Bayesianischen Betrachtungsweise ergebenden Schätzer dann als empirische Bayes-Schätzer zu interpretieren, da zwar $\boldsymbol{\gamma}$ als zufällig betrachtet wird, die Varianzparameter aber frequentistisch, also likelihoodbasiert bestimmt werden.

Der Unterschied zwischen Bayesianischer und frequentistischer Auffassung von Penalisierungsansätzen zeigt sich auch deutlich in der Kovarianzmatrix der erhaltenen Schätzer (vergleiche hierzu auch die verschiedenen Kovarianzmatrix, die wir in Abschnitt 6.3.3 für gemischte Modelle erhalten haben). Grundlage für deren Konstruktion bildet die Koeffizientenmatrix

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{U} \\ \boldsymbol{U}'\boldsymbol{X} & \boldsymbol{U}'\boldsymbol{U} + \lambda \boldsymbol{I} \end{pmatrix},$$

die zur Schätzung der Regressionskoeffizienten benötigt wird, sowie die Kreuzproduktmatrix

$$\boldsymbol{H}_1 = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{U} \\ \boldsymbol{U}'\boldsymbol{X} & \boldsymbol{U}'\boldsymbol{U} \end{pmatrix}.$$

Allgemein ergibt sich für die Kovarianzmatrix dann zunächst die Form

$$\operatorname{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\tilde{\boldsymbol{\gamma}}} \end{pmatrix} = \boldsymbol{H}^{-1} \begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{U}' \end{pmatrix} \operatorname{Cov}(\boldsymbol{y})(\boldsymbol{X}\,\boldsymbol{U})\boldsymbol{H}^{-1}.$$

Betrachtet man sowohl $\boldsymbol{\beta}$ als auch $\tilde{\boldsymbol{\gamma}}$ als feste Parameter, so erhält man mit $\operatorname{Cov}(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}$

$$\operatorname{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\tilde{\boldsymbol{\gamma}}} \end{pmatrix} = \sigma^2 \boldsymbol{H}^{-1}\boldsymbol{H}_1\boldsymbol{H}^{-1}.$$

In Bayesianischer Betrachtungsweise ergäbe sich dagegen als Posteriori-Kovarianzmatrix

$$\sigma^2 \boldsymbol{H}^{-1}.$$

Im frequentistischen Ansatz erhält man also die aus Abschnitt 6.3.3 bekannte „Sandwich“-Form für die Kovarianzmatrix. Im Gegensatz dazu ergibt sich Bayesianisch eine einfachere Kovarianzmatrix, die zu etwas weiteren Konfidenzintervallen führt.

Bayesianische Glättungsparameterwahl basierend auf MCMC

Nach dem empirischen Bayes-Ansatz zur Bestimmung des Glättungsparameters beschäftigen wir uns in diesem Abschnitt mit einer voll Bayesianischen Alternative, die mit Hilfe von MCMC-Verfahren umgesetzt wird. Dabei beschränken wir uns wieder auf Penalisierungsansätze mit in den Parametern quadratischen Straftermen. Für Bayesianische P-Splines haben wir gesehen, dass sich ein solcher Strafterm auch aus Überlegungen zu geeigneten Glattheits-Prioris herleiten lässt. Allgemein ergibt sich die folgende Analogie: In einem Regressionsmodell der Form $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ entspricht der quadratische Strafterm

$$\frac{\sigma^2}{\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$$

aus der frequentistischen Betrachtungsweise einer multivariaten Normalverteilungs-Priori

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{2\pi\tau^2}\right)^{\mathrm{rg}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right).$$

Dabei ist zu beachten, dass sich diese Dichte in der Regel nicht normieren lässt, da $\boldsymbol{K}$ nicht vollen Rang besitzt. Damit ist $p(\boldsymbol{\gamma}|\tau^2)$ teilweise uneigentlich. Betrachtet man den Logarithmus der Dichte, so entspricht dieser (bis auf das Vorzeichen und eine additive Konstante) dem Strafterm aus der penalisierten KQ-Schätzung. Damit folgt (wie für Bayesianische P-Splines), dass der Posteriori-Modus $\hat{\boldsymbol{\gamma}}$ bei gegebenem Glättungsparameter mit dem penalisierten KQ-Schätzer übereinstimmt. In Analogie zur Penalisierung bestimmter Funktionstypen lässt sich die zugehörige Priori-Verteilung dann als Glattheits-Priori interpretieren, die a priori-Annahmen zur Glattheit der Funktionsschätzung ausdrückt.

Im Folgenden wollen wir uns aber nicht mit dem Posteriori-Modus-Schätzer beschäftigen, sondern mit Hilfe von MCMC-Verfahren eine Möglichkeit zur Bestimmung des Posteriori-Erwartungswerts kennenlernen. Grundlage hierfür ist eine vollständig Bayesianische Formulierung des nichtparametrischen Regressionsproblems. Dementsprechend sind alle unbekannten Parameter mit geeigneten Priori-Verteilungen zu versehen, auch die Varianzparameter, die im empirischen Bayes-Ansatz ML- bzw. REML-geschätzt wurden. Da die inverse Gammaverteilung $\mathrm{IG}(a, b)$ im Normalverteilungsfall die konjugierte Priori-Verteilung ist, bietet sich die Wahl von IG-Prioris für σ^2 und τ^2 an (vgl. hierzu auch Abschnitt 3.5 zu Bayesianischen linearen Modellen). Genauer gehen wir von den folgenden Annahmen aus:

$$\sigma^2 \sim \mathrm{IG}(a_0, b_0) \quad \text{und} \quad \tau^2 \sim \mathrm{IG}(a_1, b_1).$$

Unter der zusätzlichen Annahme, dass die Priori-Verteilungen (bedingt) unabhängig sind, können wir schließlich die gemeinsame Posteriori-Verteilung bestimmen:

$$\begin{aligned}
p(\boldsymbol{\gamma}, \sigma^2, \tau^2|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\gamma}, \sigma^2)p(\boldsymbol{\gamma}|\tau^2)p(\sigma^2)p(\tau^2) \\
&\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})\right) \\
&\quad \cdot (\tau^2)^{-\frac{\mathrm{rg}(\boldsymbol{K})}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right) \\
&\quad \cdot \frac{1}{(\sigma^2)^{a_0+1}} \exp\left(-\frac{b_0}{\sigma^2}\right) \frac{1}{(\tau^2)^{a_1+1}} \exp\left(-\frac{b_1}{\tau^2}\right)
\end{aligned}$$

Obwohl es sich dabei um die aus Abschnitt 3.5 bekannte Normal-inverse Gammaverteilung handelt, wollen wir im Folgenden die vollständig bedingten Dichten der einzelnen Parameterblöcke für einen MCMC-Algorithmus herleiten, da diese in analoger Form in Kapitel 8 auftreten werden. Aufgrund der konjugierten IG-Prioris ergeben sich für alle vollständig bedingten Dichten bekannte Verteilungen, so dass der Gibbs-Sampler zur Schätzung verwendet werden kann. Genauer erhält man für $\boldsymbol{\gamma}$ die Normalverteilung

$$\boldsymbol{\gamma}|\boldsymbol{y}, \sigma^2, \tau^2 \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$$

mit Erwartungswert und Kovarianzmatrix

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\gamma}} &= \mathrm{E}(\boldsymbol{\gamma}|\cdot) = \left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{K}\right)^{-1} \frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{y} \\
\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} &= \mathrm{Cov}(\boldsymbol{\gamma}|\cdot) = \left(\frac{1}{\sigma^2}\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{\tau^2}\boldsymbol{K}\right)^{-1}
\end{aligned}$$

und für die Varianzparameter inverse Gammaverteilungen

$$\tau^2|\boldsymbol{y},\boldsymbol{\gamma},\sigma^2 \sim \mathrm{IG}(a_1 + 0.5\mathrm{rg}(\boldsymbol{K}), b_1 + 0.5\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}),$$
$$\sigma^2|\boldsymbol{y},\boldsymbol{\gamma},\tau^2 \sim \mathrm{IG}(a_0 + 0.5n, b_0 + 0.5(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})).$$

Aus diesen Verteilungen können nun prinzipiell einfach Zufallszahlen gezogen werden, um eine MCMC-Stichprobe zu erzeugen und damit wie üblich die interessierenden Größen der Posteriori-Verteilung durch ihre empirischen Analoga abzuschätzen. Da $\boldsymbol{\gamma}$ in nichtparametrischen Regressionsmodellen jedoch in der Regel hochdimensional ist, können sich hier sehr lange Laufzeiten ergeben, weil zur Berechnung von $\boldsymbol{m}$ und $\boldsymbol{P}$ hochdimensionale Gleichungssysteme in jeder Iteration gelöst werden müssen. Um die (erheblich) Rechenzeiten zu verkürzen, müssen also effiziente Algorithmen zum Ziehen der Zufallszahlen verwendet werden. Grundlage für diese Algorithmen ist die Beobachtung, dass die Strafmatrix $\boldsymbol{K}$ und das Kreuzprodukt der Designmatrix $\boldsymbol{Z}'\boldsymbol{Z}$ in vielen Fällen eine spezielle, nur dünn besetzte Struktur aufweisen.

7.1.10 Adaptive Verfahren

In Abschnitt 7.1.1 haben wir Polynom-Splines zunächst ohne Penalisierung eingeführt und dann die Notwendigkeit einer geeigneten Regularisierung erkannt, die mit Hilfe verschiedener Penalisierungsansätze umgesetzt wurde. Alternativ lassen sich auch Algorithmen konstruieren, deren Ziel die optimale Bestimmung der Anzahl und der Position der Knoten eines Polynom-Splines *ohne* Penalisierung ist. Hier wird die Regularisierung also nicht indirekt über einen Strafterm, sondern direkt über geeignete Modellwahl-Strategien durchgeführt. Man spricht dann auch von *adaptiven Verfahren*, da die Konstruktion des Polynom-Splines von den Daten und insbesondere den beobachteten Werten der Zielvariablen adaptiv gesteuert wird, während in Penalisierungsansätzen die Position und Zahl der Knoten nur durch die Verteilung der Kovariablen bestimmt wird.

Mittlerweile existiert eine große Zahl verschiedener adaptiver Verfahren sowohl in Bayesianischer als auch in frequentistischer Formulierung, die wir hier nicht alle im Detail besprechen können. Stattdessen werden wir uns auf einen groben Überblick und die Vermittlung der grundlegenden Ideen beschränken und weiterführende Literatur in Abschnitt 7.4 angeben.

Multivariate adaptive Regressions-Splines (MARS)

Viele frequentistische Ansätze zur adaptiven nichtparametrischen Regression lassen sich in Analogie zu Ideen der Variablenselektion einführen, wie wir sie in Abschnitt 3.6 diskutiert haben. Als repräsentatives Beispiel hierfür werden wir im Folgenden eine Variante des populären MARS-Algorithmus betrachten. Grundlage hierfür ist die Basisfunktionendarstellung

$$y_i = f(z_i) + \varepsilon_i = \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i,$$

wobei jedoch die Basisfunktionen nicht mehr fest vorgegeben sind, sondern geeignet aus einer großen Menge von möglichen Kandidaten ausgewählt werden sollen. Zur Beurteilung der Modellgüte wird dabei eines der bereits besprochenen Kriterien verwendet, also

Abb. 7.27. *MARS-Schätzung für den simulierten Datensatz: Verlauf des GCV (links) und resultierende optimale Schätzung mit acht Basisfunktionen (rechts).*

beispielsweise GCV oder AIC. Eine mögliche Suchstrategie zur Bestimmung des optimalen Modells besteht aus dem folgenden Algorithmus:

1. Starte mit einem minimalen Modell, also der kleinsten zu betrachtenden Basis.
2. Erweitere das Modell schrittweise bis das volle Modell erreicht ist:
   - Berechne für alle noch nicht im Modell enthaltenen Basisfunktionen B_j die Score-Statistik zum Test auf $\gamma_j = 0$ und erweitere das Modell um die Basisfunktion mit der größten Score-Statistik.
   - Schätze das neue Modell und speichere das zugehörige Gütekriterium.
3. Verkleinere ausgehend vom vollen Modell die Menge der Basisfunktionen schrittweise bis zum minimalen Modell:
   - Berechne für alle noch im Modell enthaltenen Basisfunktionen B_j die Wald-Statistik zum Test auf $\gamma_j = 0$ und entferne die Basisfunktion mit der kleinsten Wald-Statistik aus dem Modell.
   - Schätze das neue Modell und speichere das zugehörige Gütekriterium.
4. Wähle das Modell mit dem optimalen Gütekriterium aus der entstandenen Sequenz als bestes Modell.

Der Algorithmus besteht also im Wesentlichen aus zwei Teilschritten, in denen eine Vorwärts- und eine Rückwärtssuche durch den Raum der möglichen Modelle vorgenommen werden. Dabei werden für die Aufnahme neuer Modellterme und die Streichung bereits enthaltener Terme unterschiedliche Kriterien zugrunde gelegt, so dass sich zwei verschiedene Sequenzen von Modellen ergeben. Die jeweils verwendeten Test-Kriterien sind so gewählt, dass ihre Berechnung mit möglichst geringem Aufwand möglich ist. Insbesondere kann die Score-Statistik zur Aufnahme einer neuen Basisfunktion aus der aktuellen Modellschätzung bestimmt werden, ohne dass das erweiterte Modell selbst geschätzt werden muss. Umgekehrt lässt sich die Wald-Statistik berechnen, ohne dass tatsächlich ein Submodell geschätzt werden muss. Damit ist die Durchführung des Algorithmus auch noch für eine größere Zahl von möglichen Basisfunktionen handhabbar.

Zur Demonstration analysieren wir noch das simulierte Datenbeispiel mit Hilfe einer MARS-Implementation, die unter der Bezeichnung `polymars` im Paket `polspline` des Programmpakets R erhältlich ist. Die Implementation verwendet die lineare TP-Basis

und startet mit dem Intercept als minimalem Modell. Als maximale Anzahl von Basisfunktionen wurde 17 gewählt, wobei alle beobachteten Kovariablenausprägungen als Positionen der Knoten in Frage kamen. Zur Beurteilung der Modellanpassung wurde das GCV-Kriterium verwendet. Abbildung 7.27 (links) zeigt den Verlauf des GCV für die Sequenz der berechneten Modelle. Offenbar besitzen die einfachen Modelle mit einer geringen Zahl von Basisfunktionen eine schlechte Anpassung an die Daten, so dass sich für die Modelle am Anfang und am Ende der Sequenz große Werte des GCV ergeben. Allerdings führt auch die Verwendung eines sehr großen Modells zu größeren Werten des GCV, da in diesem Fall die Funktionsschätzung zu stark an die Daten angepasst wird. Aus dem Verlauf des GCV erkennt man auch, dass die Vorwärts- und die Rückwärtssuche tatsächlich in verschiedenen Modellfolgen resultieren, die GCV-Kurve also nicht vollständig symmetrisch ist. In komplexeren Modellen mit einer größeren Zahl von Einflussgrößen sind sogar noch wesentlich deutlichere Abweichungen von der Symmetrie zu erwarten. Abbildung 7.27 (rechts) zeigt die sich aus dem MARS-Algorithmus ergebende optimale Schätzung mit acht Basisfunktionen. Im vorliegenden Beispiel ergibt sich bereits für lineare Splines eine zufriedenstellende Modellanpassung. Generell empfehlen sich zur Modellierung glatter Funktionen jedoch Basen höheren Grads, die in `polymars` allerdings nicht zur Verfügung stehen.

Regressionsbäume

Als zweites Beispiel zur adaptiven Modellierung betrachten wir nun ein Verfahren, dass die Funktion $f(z)$ durch eine stückweise konstante Funktion approximiert. Dabei soll der Definitionsbereich der z-Variablen so aufgeteilt werden, dass die entstehenden Gruppen (in Bezug auf die abhängige Variable) in sich möglichst homogen sind, während die Gruppen untereinander möglichst heterogen sein sollen. Im Rahmen der nichtparametrischen Regression lässt sich diese Problemstellung so auffassen: Die Funktion $f(z)$ soll dargestellt werden als

$$f(z) = \sum_{j=1}^{d} \gamma_j \mathbb{1}_{(\kappa_{j-1},\kappa_j]}(z),$$

d.h. der Definitionsbereich von z wird durch die Werte $\kappa_0 < \kappa_1 < \ldots < \kappa_d$ zerlegt in Intervalle $R_j = [\kappa_{j-1}, \kappa_j)$, auf denen die Funktion f den Wert γ_j annimmt. Sowohl die Zerlegung $\kappa_0, \ldots, \kappa_d$ als auch die Funktionswerte γ_j sollen nun so optimiert werden, dass die Residuenquadratsumme $\sum_i (y_i - f(z_i))^2$ minimiert wird. Ohne Restriktionen führt dies zur perfekten Datenanpassung, in der jede Beobachtung eine eigene Gruppe darstellt. In der Regel werden daher zusätzliche Nebenbedingungen an die Darstellung von $f(z)$ gestellt, die wir im Verlauf dieses Abschnitts noch genauer diskutieren werden. Für eine stückweise konstante Funktion ergibt sich die Residuenquadratsumme

$$\sum_{j=1}^{d} \sum_{i:z_i \in R_j} (y_i - \gamma_j)^2, \tag{7.20}$$

so dass sich für eine gegebene Zerlegung die Funktionswerte γ_j unmittelbar als lokale Mittelwerte der abhängigen Variablen mit $z_i \in R_j$ ergeben, also $\hat{\gamma}_j = \bar{y}(R_j)$. Die optimale Bestimmung der Zerlegung ist dagegen wesentlich schwieriger und im Allgemeinen nicht mehr direkt durchführbar. Daher verwendet man einen rekursiven Algorithmus,

Abb. 7.28. *Regressionsbaum für den Tansania-Datensatz. Die Zerlegung wurde gestoppt, wenn eine Gruppe weniger als 20 Beobachtungen enthielt.*

der den Definitionsbereich sukzessive in Teilintervalle zerlegt. Abbildung 7.28 illustriert dieses Vorgehen anhand des Unterernährungs-Beispiels (und veranschaulicht auch die Bezeichnung der sich ergebenden Zerlegung als *Regressionsbaum*). Zunächst wird der Definitionsbereich der Variablen Alter durch einen einzelnen Trennpunkt κ aufgeteilt, so dass die Variabilität innerhalb der entstehenden zwei Gruppen minimal ist. In unserem Beispiel ergibt sich der Trennpunkt $\kappa = 7.5$, wie aus Abbildung 7.28 an der obersten Verzweigung abzulesen ist. Anschließend wird das gleiche Prinzip wieder auf die beiden entstandenen Subgruppen angewendet und so schrittweise der Definitionsbereich zerlegt. Die entstehende Struktur lässt sich dann durch einen binären Baum darstellen, in dem jede Verzweigung einen Trennpunkt repräsentiert. In Abbildung 7.28 sind an den Verzweigungen jeweils die entsprechenden Trennpunkte angegeben. Zusätzlich kann am unteren Ende des Baums abgelesen werden, wie viele Beobachtungen jeweils in den resultierenden Gruppen verbleiben und welchen Wert $\hat{\gamma}_j = \bar{y}(R_j)$ für diese Gruppe annimmt.

Formaler geht man zur Definition eines Regressionsbaums also folgendermaßen vor: Initialisiert wird der Baum basierend auf der Zerlegung des Definitionsbereichs in die Intervalle

$$R_1 = \{z : z \leq \kappa\}, \qquad R_2 = \{z : z > \kappa\},$$

wobei κ gewählt wird durch Minimierung des Kriteriums

$$\sum_{\{i:z_i \in R_1\}} (y_i - \bar{y}(R_1))^2 + \sum_{\{i:z_i \in R_2\}} (y_i - \bar{y}(R_2))^2.$$

Da hier nur ein Trennpunkt bestimmt werden muss, ist die numerische Bestimmung von κ einfach. Zur Auswahl stehen alle beobachteten Kovariablenausprägungen z_i, so dass eine direkte Suche möglich ist. Anschließend wendet man das gleiche Prinzip auf die zwei

Abb. 7.29. *Kreuzvalidierungskriteriums in Abhängigkeit von der Anzahl der Verzweigungen.*

entstandenen Intervalle R_1 und R_2 an, so dass sich rekursiv eine Zerlegung der Form (7.20) ergibt.

Zur Erstellung eines Regressionsbaums stellt sich noch die Frage nach einem geeigneten Abbruchkriterium, das festlegt, bis zu welcher Ebene im Baum noch neue Verzweigungen aufgenommen werden sollen. Dazu gibt es im Wesentlichen zwei einfache Strategien:

1. Das aktuelle Intervall wird nur dann weiter zerlegt, wenn es mindestens eine bestimmte Anzahl von Beobachtungen enthält.
2. Das aktuelle Intervall wird nur weiter zerlegt, wenn dadurch eine bestimmte (absolute oder relative) Reduktion der Residuenquadratsumme erreicht werden kann.

Für das Beispiel in Abbildung 7.28 haben wir das erste Kriterium mit einer Mindestanzahl von 20 Beobachtungen verwendet.

In der Regel sind die mit diesen einfachen Abbruchkriterien berechneten Regressionsbäume zu komplex, so dass eine zusätzliche Strategie zur Komplexitätsreduktion angewendet wird. Dabei wird zunächst ein komplexer Baum mit einer großen Zahl von Verzweigungen bestimmt und anschließend mit Hilfe eines Komplexitätsmaßes überflüssige Verzweigungen wieder entfernt. Dieser Vorgang wird als *Stutzen* (*Pruning*) des Regressionsbaums bezeichnet. Dazu definiert man das komplexitätsadjustierte Optimalitätskriterium

$$\mathrm{PKQ}_\lambda(T) = \mathrm{KQ}(T) + \lambda|T|,$$

wobei $\mathrm{KQ}(T)$ die ursprüngliche Residuenquadratsumme für den Baum T bezeichnet und $|T|$ die Anzahl der durch den Baum definierten Verzweigungen angibt. Die Idee hinter diesem Kriterium entspricht unseren früheren Überlegungen zu Penalisierungsansätzen. Insbesondere lässt sich λ als Glättungs- bzw. Komplexitätsparameter interpretieren, durch dessen Variation man unterschiedlich komplexe Varianten des ursprünglichen Baums erhält. Genauer ergibt sich durch sukzessive Verringerung von λ eine Sequenz von Bäumen mit wachsender Komplexität, also wachsender Zahl von Verzweigungen. Der optimale Wert für λ lässt sich dann beispielsweise per Kreuzvalidierung bestimmen.

In Abbildung 7.29 ist das Kreuzvalidierungskriterium für den Tansania-Datensatz aufgetragen gegen die Anzahl von Verzweigungen im Baum. Die Anzahl von Verzweigungen wird dabei implizit bestimmt durch den Wert des Glättungsparameters λ. Man beachte, dass nicht alle Anzahlen von Verzweigungen tatsächlich erreicht werden, da man

Abb. 7.30. *Gestutzter Regressionsbaum und zugehörige, stückweise konstante Funktionsschätzung für den Tansania-Datensatz.*

durch Variation von λ unter Umständen unmittelbar zu einem Modell mit mehr als einer zusätzlichen Verzweigung gelangt. Offenbar führen sowohl zwei als auch vier Verzweigungen zu sehr ähnlichen Werten des Kreuzvalidierungskriteriums, mit einer leichten Tendenz zum komplexeren Modell. Damit ergibt sich die in Abbildung 7.30 visualisierte Schätzung mit vier Verzweigungen als optimales Modell. Man erkennt wieder die wesentlichen Eigenschaften der Daten, nämlich einen höheren Z-Score, d.h. ein niedrigeres Unterernährungsrisiko für jüngere Kinder und den durch den Wechsel der Referenzpopulation hervorgerufenen Anstieg in f nach dem 24. Monat.

Der Vorteil von Regressionsbäumen liegt in der einfachen Interpretierbarkeit der Ergebnisse. In komplexeren Modellen erlauben Regressionsbäume darüber hinaus die einfache Identifikation von Interaktionen. Problematisch ist in vielen Anwendungen die Unterstellung einer stückweise konstanten Funktion. Darüberhinaus sind Regressionsbäume instabil, d.h. bei einer geringfügigen Veränderung der Datengrundlage kann unter Umständen ein völlig anderer optimaler Baum entstehen. Dieses Problem hat auch einen Einfluss auf die Wahl eines optimalen Glättungsparameters durch Kreuzvalidierung. Je nach zugrunde gelegter Aufteilung des Datensatzes können hier deutliche Unterschiede entstehen. In modernen Erweiterungen wird diesem Problem begegnet, indem Resampling-Verfahren in Kombination mit gewichteter Mittelwertbildung über die entstehenden Bäume eingesetzt werden, vgl. die Hinweise in Abschnitt 7.4.

Bayesianische adaptive Verfahren I: Model Averaging

Zur Motivation Bayesianischer adaptiver Verfahren gehen wir wieder vom Basisfunktionenansatz

$$y_i = f(z_i) + \varepsilon_i = \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i$$

aus. Aus einer großen Zahl von möglichen Basisfunktionen soll eine geeignete Teilmenge ausgewählt werden. Anders formuliert soll die Frage beantwortet werden, welche der Koeffizienten γ_j von Null verschieden sind, so dass die entsprechenden Basisfunktionen die Schätzung beeinflussen. In einem Bayesianischen Ansatz kann diese Fragestellung direkt angegangen werden, indem Indikatorvariablen $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_d)'$ eingeführt werden, für die gilt

$$\rho_j = \begin{cases} 1 & \text{falls } \gamma_j \neq 0, \\ 0 & \text{falls } \gamma_j = 0. \end{cases}$$

Damit lässt sich das ursprüngliche Modell umschreiben als

$$f(z) = \sum_{j=1}^{d} \rho_j \gamma_j B_j(z_i).$$

Für einen gegebenen Vektor von Indikatoren bezeichne nun $\boldsymbol{\gamma}_{\boldsymbol{\rho}}$ den Vektor der Regressionskoeffizienten, der sich durch Streichung der auf Null gesetzten Komponenten ergibt. Ebenso sei $\boldsymbol{Z}_{\boldsymbol{\rho}}$ die sich aus den Basisfunktionen ergebende Designmatrix, in der nur Spalten mit $\rho_j = 1$ berücksichtigt sind. In Matrixnotation erhält man damit den Vektor der Funktionsauswertungen basierend auf dem durch $\boldsymbol{\rho}$ bestimmten Modell als $\boldsymbol{f}_{\boldsymbol{\rho}} = \boldsymbol{Z}_{\boldsymbol{\rho}} \boldsymbol{\gamma}_{\boldsymbol{\rho}}$. Vernachlässigen wir in der Notation die Fehlervarianz σ^2, so lässt sich die marginale Posteriori-Verteilung des Vektors $\boldsymbol{f}$ (ohne Bedingen auf ein konkretes Modell) schreiben als

$$p(\boldsymbol{f}|\boldsymbol{y}) = \sum_{\boldsymbol{\rho} \in \{0,1\}^d} p(\boldsymbol{f}_{\boldsymbol{\rho}}|\boldsymbol{\rho}, \boldsymbol{y}) p(\boldsymbol{\rho}|\boldsymbol{y}), \tag{7.21}$$

d.h. man erhält eine Mischung über alle möglichen Ausprägungen des Indikatorvektors $\boldsymbol{\rho}$, wobei die Posteriori-Verteilungen $p(\boldsymbol{f}_{\boldsymbol{\rho}}|\boldsymbol{\rho}, \boldsymbol{y})$ durch die Posteriori-Wahrscheinlichkeiten $p(\boldsymbol{\rho}|\boldsymbol{y})$ gewichtet werden. Insbesondere ergibt sich für den marginalen Posteriori-Erwartungswert von $\boldsymbol{f}$ auch eine Mischung über die bedingten Posteriori-Erwartungswerte:

$$\mathrm{E}(\boldsymbol{f}|\boldsymbol{y}) = \sum_{\boldsymbol{\rho} \in \{0,1\}^d} \mathrm{E}(\boldsymbol{f}_{\boldsymbol{\rho}}|\boldsymbol{\rho}, \boldsymbol{y}) p(\boldsymbol{\rho}|\boldsymbol{y}). \tag{7.22}$$

Dies ist ein spezielles Beispiel für das Konzept des *Model Averaging*, das wir noch in einem etwas allgemeineren Rahmen formulieren wollen. Gegeben sei eine Menge möglicher Modelle $M_1, \ldots, M_p$, beispielsweise die sich aus den 2^d möglichen Indikatorvektoren $\boldsymbol{\rho}$ in unserem Modellwahlproblem ergebenden Modelle. Liegt keine Prior-Information über die Wahrscheinlichkeit der Modelle vor, so kann als Priori-Verteilung eine Gleichverteilung angenommen werden. In unserem Beispiel erhielte damit jeder mögliche Indikatorvektor $\boldsymbol{\rho}$ die Priori-Wahrscheinlichkeit $1/2^d$. Durch Model Averaging fasst man nun alle möglichen Modelle zusammen, gewichtet dabei aber entsprechend der Posteriori-Wahrscheinlichkeiten $p(M_j|\boldsymbol{y})$. Ist das eigentliche Ziel die Schätzung eines allen Modellen gemeinsamen Parametervektors $\boldsymbol{\theta}$ (in unserem Beispiel $\boldsymbol{f}$), so ist die Posteriori-Verteilung von $\boldsymbol{\theta}$ gegeben durch die Mischverteilung

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{j=1}^{p} p(\boldsymbol{\theta}|M_j, \boldsymbol{y}) p(M_j|\boldsymbol{y}).$$

Während sich für konjugierte Priori-Verteilungen die bedingten Verteilungen $p(\boldsymbol{\theta}|M_j, \boldsymbol{y})$ noch relativ einfach berechnen lassen, ergeben sich bei der Bestimmung der Posteriori-Wahrscheinlichkeiten $p(M_j|\boldsymbol{y})$ häufig Schwierigkeiten.

In unserem Modellwahlproblem erhält man unter Verwendung von Normalverteilungs-Prioris für die Parameter $\boldsymbol{\gamma}_{\boldsymbol{\rho}}$ auch a posteriori Normalverteilungen für $\boldsymbol{\gamma}_{\boldsymbol{\rho}}$ und die Funktionsauswertungen $\boldsymbol{f}_{\boldsymbol{\rho}}$. Hier lassen sich theoretisch sogar die Posteriori-Wahrscheinlichkeiten explizit angeben. Allerdings wird der direkte Zugang über (7.21) durch eine andere Schwierigkeit praktisch limitiert. Für eine größere Zahl von Basisfunktionen müssen

$p = 2^d$ Modelle berücksichtigt werden, so dass in der Praxis die Bestimmung aller Posteriori-Wahrscheinlichkeiten bereits für eine relativ geringe Zahl von Basisfunktionen nicht mehr möglich ist. Daher verwendet man in der Regel nicht die expliziten Formeln zum Model Averaging, sondern eine auf MCMC-Simulationen beruhende Approximation. Dabei werden in jeder Iteration zunächst die Indikatoren $\boldsymbol{\rho}$ aktualisiert, so dass man eine aktuelles Modell erhält. Bedingt auf dieses aktuelle Modell kann dann $\boldsymbol{f}_{\boldsymbol{\rho}}$ simuliert werden. Genauer erhält man den folgenden Algorithmus, der auch die Schätzung der Fehlervarianz beinhaltet:

1. Wähle geeignete Startwerte $\boldsymbol{\rho}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ und $\sigma^{2(0)}$ sowie eine maximale Iterationszahl T und setze $t = 1$.
2. Für $j = 1, \ldots, d$ generiere $\rho_j^{(t)}$ aus der marginalen vollständig bedingten Dichte $p(\rho_j | \boldsymbol{\rho}_{-j}^{(t-1)}, \boldsymbol{y})$.
3. Simuliere $\boldsymbol{f}_{\boldsymbol{\rho}}^{(t)}$ gemäß der vollständig bedingten Dichte $p(\boldsymbol{f}_{\boldsymbol{\rho}} | \boldsymbol{\rho}^{(t)}, \sigma^{2(t-1)}, \boldsymbol{y})$.
4. Simuliere $\sigma^{2(t)}$ aus der vollständig bedingten Dichte $p(\sigma^2 | \boldsymbol{\rho}^{(t)}, \boldsymbol{f}_{\boldsymbol{\rho}}^{(t)}, \boldsymbol{y})$.
5. Falls $t < T$ setze $t = t + 1$ und gehe zurück zu 2.

Die Verwendung der marginalen vollständig bedingten Dichte zur Simulation von ρ_j (also einer nicht von $\boldsymbol{\gamma}_{\boldsymbol{\rho}}$ und σ^2 abhängenden Verteilung) führt in der Regel zu einem günstigeren Verhalten der erzeugten Markov-Kette. Verwendet man konjugierte Priori-Verteilungen, also Normalverteilungs-Prioris für die Regressionsparameter und eine inverse Gammaverteilung für σ^2, so können die Schritte 3. und 4. per Gibbs-Sampling durchgeführt werden. Eine ausführlichere Diskussion geeigneter Priori-Verteilungen und eine detailliertere Beschreibung des sich ergebenden MCMC-Algorithmus findet man in Smith & Kohn (1996).

Aus den Realisierungen des MCMC-Algorithmus lässt sich nun der Posteriori-Erwartungswert (7.22) approximieren durch

$$\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{f}_{\boldsymbol{\rho}}^{(t)}.$$

Da Modelle mit einer hohen Posteriori-Wahrscheinlichkeit im Laufe des MCMC-Algorithmus häufiger berücksichtigt werden, ergibt sich in diesem Mittelwert eine Gewichtung, die (zumindest approximativ) der theoretischen Gewichtung in (7.22) entspricht. Durch

$$\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\rho}^{(t)}$$

erhält man darüber hinaus Schätzungen für die Posteriori-Wahrscheinlichkeiten $\mathrm{P}(\rho_j = 1 | \boldsymbol{y})$, also Posteriori-Inklusionswahrscheinlichkeiten für die einzelnen Basisfunktionen. Wendet man einen Schwellenwert auf diese Wahrscheinlichkeiten an, so ergibt sich auch eine einfache Möglichkeit, basierend auf dem MCMC-Output ein einzelnes geeignetes Modell auszuwählen, das nur die im Sinne der Posteriori-Wahrscheinlichkeiten relevanten Basisfunktionen enthält.

Bayesianische adaptive Verfahren II: Reversible Jump MCMC

Eine zweite Bayesianische Möglichkeit zur adaptiven nichtparametrischen Regression besteht darin, die Anzahl und die Lokation der Knoten als zu schätzende Parameter in die

Modellformulierung aufzunehmen. Dazu müssen zunächst geeignete Priori-Verteilungen für diese neuen Parameter gefunden werden. Die Anzahl der Knoten m kann etwa durch eine Poisson-Verteilung beschrieben werden, deren Parameter λ die a priori-Annahmen über die Modellgröße repräsentiert. Für einen großen Parameter λ werden a priori komplexere Modelle bevorzugt, während kleine Parameter simplere Modelle mit einer geringeren Zahl von Knoten favorisieren. In der Praxis wird die Poisson-Verteilung häufig an einer Obergrenze m_{max} trunkiert, d.h. die maximale Anzahl von Basisfunktionen wird nach oben beschränkt, so dass stets $m \leq m_{max}$ gilt. Als Alternative zur Poisson-Verteilung kommt auch eine Gleichverteilung auf $\{0, \ldots, m_{max}\}$ in Frage. Zur Definition der Priori-Verteilung für die Positionen der Knoten konstruiert man üblicherweise zunächst eine große Kandidatenmenge von Knoten, aus denen die Positionen der aktuellen Knoten gewählt werden können. Unter Verwendung dieser Kandidatenmenge können dann alle Knotenpositionen a priori als gleich wahrscheinlich betrachtet werden. Eine alternative Priori-Verteilung ergibt sich aus der Verteilung der Ordnungsstatistik (Order Statistic) einer Gleichverteilung über den Definitionsbereich von z.

Die Schätzung des um die zusätzlichen Parameter erweiterten Modells soll nun wieder in einem MCMC-Algorithmus umgesetzt werden. Die Konstruktion eines solchen Algorithmus wird jedoch durch die Tatsache erschwert, dass der zu schätzende Parametervektor von Iteration zu Iteration in der Dimension variiert. Daher muss die übliche MCMC-Methodik dahingehend erweitert werden, was zum sogenannten *Reversible Jump MCMC* (RJMCMC) führt. Zusätzlich zu den bereits bekannten Schritten, die zur Simulation der Regressionskoeffizienten und der Fehlervarianz bei einer gegebenen Knotenkonfiguration anzuwenden sind, müssen wir also neue Simulationsschritte für die Anzahl und die Position der Knoten definieren. In einem RJMCMC-Algorithmus wird diese Simulation indirekt vorgenommen, indem nicht neue Zustände für diese Parameter vorgeschlagen, sondern die aktuellen Werte geeignet modifiziert werden. Dazu stehen die drei folgenden Möglichkeiten zur Verfügung:

1. *Erzeugen eines neuen Knotens*: Wähle zufällig einen bisher nicht im Modell enthaltenen Knoten aus und füge ihn der Knotenmenge hinzu.
2. *Löschen eines bestehenden Knotens*: Entferne zufällig einen Knoten aus der aktuellen Knotenmenge.
3. *Verschieben eines aktuellen Knotens*: Wähle zufällig einen Knoten aus und verschiebe diesen innerhalb des durch die beiden benachbarten Knoten gebildeten Bereichs.

Die ersten beiden Schritte modifizieren die Anzahl der Knoten, während der dritte Schritt eine Variation der Knotenposition bei festgehaltener Knotenzahl ermöglicht. Innerhalb des RJMCMC-Algorithmus wird nun in jeder Iteration zufällig entschieden, ob aus dem aktuellen Modell simuliert oder die Knotenkonfiguration durch einen der drei Schritte variiert wird. Bei der Modifikation der Knotenkonfiguration muss dann zusätzlich die Akzeptanzwahrscheinlichkeit angepasst werden, worauf wir hier allerdings nicht im Detail eingehen wollen, sondern auf die entsprechende Literatur verweisen (Green (1995) für theoretische Grundlagen zu RJMCMC sowie Biller (2000) und Denison, Mallick & Smith (1998) für Anwendungen im Rahmen Bayesianischer adaptiver Verfahren).

Durch die RJMCMC-Simulation ergibt sich wie beim Model Averaging eine Stichprobe möglicher Modelle, die entsprechend ihrer Posteriori-Wahrscheinlichkeiten gewichtet werden. Zusätzlich erhält man die Posteriori-Verteilung für die Anzahl der Knoten und deren Positionen, so dass auch diese mit den üblichen Verfahren analysiert werden können.

Abb. 7.31. *Mietspiegel München: Separate Schätzungen des Effekts der Wohnfläche bei Gruppierung nach Baujahr (links) sowie separate Schätzungen des Effekts des Baujahrs bei Gruppierung nach Wohnfläche (rechts). Die Gruppierung wurde dabei folgendermaßen gebildet: bjahr* ≤ 1938 *(—),* $1938 <$ *bjahr* ≤ 1958 *(- - -),* $1958 <$ *bjahr* ≤ 1978 *(· · ·), bjahr* > 1978 *(- · - ·) bzw. flaeche* ≤ 55 *(—),* $55 <$ *flaeche* ≤ 90 *(- - -),* $90 <$ *flaeche* ≤ 135 *(· · ·), flaeche* > 135 *(- · - ·).*

7.2 Bivariate Glättung und räumliche Effekte

Bisher haben wir uns auf den Fall *einer* metrischen Einflussgröße beschränkt. In diesem Abschnitt werden wir Ansätze zur bivariaten Glättung (also für Modelle mit zwei metrischen Einflussgrößen) und zur Modellierung räumlicher Effekte behandeln. Um einen Eindruck der unterschiedlichen denkbaren Fragestellungen und Datensituationen zu gewinnen, behandeln wir zunächst eine Reihe von Beispielen, deren Modellierung dann in den folgenden Abschnitten thematisiert wird.

Beispiel 7.5 Mietspiegel München – Interaktion zwischen Wohnfläche und Baujahr

In Kapitel 3 (Beispiel 3.5) haben wir gesehen, dass sowohl das Baujahr als auch die Wohnfläche einen möglicherweise nichtlinearen Einfluss auf die Nettomiete pro Quadratmeter ausüben. Mit den bisher in Abschnitt 7.1 behandelten Verfahren wäre es nur möglich, jeweils den Einfluss einer der beiden Variablen separat nichtparametrisch zu untersuchen. Dabei wird jedoch unterstellt, dass keine Interaktion zwischen den beiden Effekten vorliegt. Um diese Annahme zu überprüfen, haben wir die Daten sowohl nach Baujahr als auch nach Wohnfläche gruppiert und jeweils den Effekt der anderen Variablen nichtparametrisch geschätzt. Die in Abbildung 7.31 wiedergegebenen Ergebnisse legen nahe, dass zwischen den beiden Kovariablen Interaktionseffekte auftreten, da sich die Effekte für die einzelnen Gruppen jeweils deutlich unterscheiden. Ein flexibles Modell zur Beschreibung solcher Interaktionseffekte besitzt die Form

$$mieteqm_i = f(flaeche_i, bjahr_i) + \varepsilon_i,$$

wobei f wie in Abschnitt 7.1 eine nicht näher spezifizierte, glatte Funktion sein soll, die nun allerdings in Abhängigkeit von zwei Variablen variiert. Da die Interaktion zwischen Wohnfläche und Baujahr also durch eine Oberfläche modelliert wird, spricht man in diesem Zusammenhang auch von der Schätzung von Interaktionsoberflächen.

△

Abb. 7.32. *Waldzustandsanalyse in Baden-Württemberg. Links: An den durch offene Kreise markierten Lokationen ist die Kalziumkonzentration bekannt, an den durch gefüllte Kreise markierten Lokationen werden Vorhersagen benötigt. Rechts: Lokale Mittelwerte basierend auf der Einteilung der Beobachtungsregion in ein reguläres Gitter.*

Beispiel 7.6 Analyse des Waldzustands in Baden-Württemberg

Die diesem Beispiel zugrunde liegende Anwendung beschäftigt sich mit der Modellierung von Waldschäden in Baden-Württemberg. Dabei soll unter Anderem die Schädigung von Bäumen zur Konzentration verschiedener Chemikalien im Boden in Beziehung gesetzt werden. Dabei tritt folgendes Problem auf: Nicht an allen Lokationen, an denen Erhebungen zum Waldzustand durchgeführt wurden, sind auch die Chemikalienkonzentrationen gemessen worden. Diese Situation wird in Abbildung 7.32 am Beispiel Kalzium verdeutlicht. In der linken Grafik bezeichnet jeder Punkt eine Lokation, an der der Waldzustand erhoben wurde. Durch offene Kreise sind diejenigen Lokationen gekennzeichnet, an denen gleichzeitig auch die Kalziumkonzentration bestimmt wurde. An den durch gefüllte Kreise gekennzeichneten Lokationen ist dagegen keine Messung zur Kalziumkonzentration vorhanden. Um alle Beobachtungen in der Analyse berücksichtigen zu können, soll ein Modell der Form

$$Ca_i = f(x_i, y_i) + \varepsilon_i$$

verwendet werden, so dass an Lokationen ohne Messungen die Kalziumkonzentration vorhergesagt werden kann. Dabei bezeichnet Ca die gemessene Kalziumkonzentration und x bzw. y die Lokation der Messung in Längen- und Breitengrad. Prinzipiell befindet man sich also wieder in der gleichen Situation wie in Beispiel 7.5, d.h. eine metrische Zielvariable soll flexibel in Abhängigkeit zweier metrischer Einflussgrößen modelliert werden. Hier treten jedoch zwei zusätzliche Besonderheiten auf: Zum einen repräsentieren die Einflussgrößen räumliche Information, so dass spezielle Verfahren der räumlichen Statistik angewendet werden können, und zum anderen benötigen wir eine Darstellung von f, die an beliebigen Lokationen auswertbar ist. Letzteres kann beispielsweise durch Einteilung der Daten in ein regelmäßiges Gitter und lokale Mittelwertbildung nur schwer erreicht werden, da relativ große Lücken in den Daten vorhanden sind (vergleiche Abbildung 7.32 rechts). Man erkennt anhand dieser deskriptiven Abbildung aber bereits, dass sich die Kalziumkonzentration tatsächlich kontinuierlich verändert, so dass die Modellierung durch eine glatte Oberfläche sinnvoll erscheint.

△

Abb. 7.33. *Gehirnkartierung: Pixelweise Aktivierung bei visuellem Reiz. Die linke Grafik zeigt die Aktivierung zum Zeitpunkt $t = 18$, die rechte Grafik zum Zeitpunkt $t = 38$.*

Beispiel 7.7 Gehirnkartierung

Ziel der Gehirnkartierung ist die Identifikation von Arealen im Gehirn, die beispielsweise durch einen visuellen Stimulus aktiviert werden. In diesem Experiment wurden Probanden mit Hilfe der funktionellen Magnet-Resonanztomografie (fMRT) untersucht, wobei sich Ruhephasen mit Phasen visueller Stimulierung abwechselten. Zur Illustration zeigt Abbildung 7.33 eine Schicht des Gehirns zu zwei verschiedenen Zeitpunkten, an denen der visuelle Stimulus vorhanden war. Regionen mit erhöhter Aktivierung sind dabei durch hell gefärbte Pixel zu erkennen. Solche hellen Pixel befinden sich überwiegend im hinteren Teil des Gehirns, in dem der visuelle Cortex liegt, der für die Verarbeitung visueller Reize zuständig ist. Ziel ist es nun, mit Hilfe von räumlichen Glättungsverfahren das durch das Aufnahmeverfahren entstehende Rauschen aus den Daten herauszufiltern, um dadurch deutlicher zwischen aktivierten und nicht aktivierten Arealen unterscheiden zu können. Das Glättungsproblem lässt sich demnach wieder als Schätzung einer bivariaten (oder räumlichen) Funktion auffassen. Dabei liegt die räumliche Information streng genommen nur diskret vor. Da die Anzahl der Pixel aber sehr groß ist, kann man diese Diskretisierung vernachlässigen und jedes Pixel mit den Koordinaten seines Mittelpunkts identifizieren.

△

In den bisherigen Beispielen waren die Einflussgrößen bzw. die räumlichen Koordinaten stets auf einer kontinuierlichen Skala gemessen oder zumindest als quasi-stetig interpretierbar. Im Rahmen der räumlichen Statistik werden wir in diesem Fall auch von *stetigen Lokationsvariablen* sprechen. In einer anderen typischen Datensituation der räumlichen Statistik ist jedoch tatsächlich nur diskrete räumliche Information gegeben. In diesem Fall bezeichnen wir die Einflussgröße als *diskrete Lokationsvariable*. Diese kann beispielsweise aus der Information bestehen, in welchem der Landkreise Deutschlands eine beobachtete Person ihren Wohnsitz hat. Ziel ist es dann, die räumliche Anordnung der Landkreise auszunutzen, um eine räumlich glatte Funktion zu schätzen. Verschiedene Situationen dieser Art haben wir bereits in Abschnitt 2.8 kennengelernt. Ein weiteres Beispiel ergibt sich für den Münchner Mietspiegel:

Beispiel 7.8 Mietspiegel München – Untersuchung räumlicher Abhängigkeiten

In den Erhebungen zum Münchner Mietspiegel wird auch registriert, in welchem der Bezirksviertel sich die jeweilige Wohnung befindet. Abbildung 7.34 zeigt die durch-

Abb. 7.34. *Mietspiegel München: Durchschnittliche Nettomiete pro Quadratmeter in den Bezirksvierteln. Die schraffierten Bereiche kennzeichnen Bezierksviertel, in denen keine Mietwohnungen vorhanden sind (beispielsweise Parks oder Industriegebiete).*

schnittliche Nettomiete pro Quadratmeter, jeweils separat pro Bezirksviertel. Aus der Abbildung erkennt man beispielsweise, dass im Norden Münchens die Mieten tendenziell geringer sind als im Durchschnitt. Ziel einer räumlichen Analyse ist es, solche räumlich strukturierten Effekte deutlicher aufzudecken, indem die räumliche Nähe der Bezirksviertel geeignet zur Glättung verwendet wird.

△

Generell lassen sich also zwei unterschiedliche Fragestellungen charakterisieren, die wir im weiteren Verlauf dieses Kapitels untersuchen wollen:

- die Schätzung zweidimensionaler Oberflächen zur Modellierung von Interaktionen oder räumlichen Effekten bei stetig gemessenen Einflussgrößen bzw. Lokationsvariablen, und
- die Schätzung räumlicher Effekte basierend auf diskreter räumlicher Information in Form von Zugehörigkeit zu Regionen oder der räumlichen Anordnung auf einem diskreten Gitter.

Die erste Problemstellung wird in den Abschnitten 7.2.1 bis 7.2.3 behandelt. Dabei erweitern wir die in Abschnitt 7.1 besprochenen Penalisierungsansätze, insbesondere penalisierte Splines und den Kriging-Ansatz, vom univariaten auf den bivariaten Fall. Ansätze zur Modellierung diskreter räumlicher Information, die in gewisser Weise als räumliche Erweiterungen der in Abschnitt 7.1.5 behandelten Random Walk-Modelle betrachtet werden können, werden wir dann in Abschnitt 7.2.4 besprechen. In Abschnitt 7.2.6 diskutieren wir kurz einige bivariate Erweiterungen lokaler und adaptiver Verfahren.

7.2.1 Tensorprodukt-P-Splines

Tensorprodukt-Basen

In Abschnitt 3.1.4 wurden Interaktionen zwischen Kovariablen durch Produkte der entsprechenden Designvektoren modelliert. Diese Idee lässt sich auch auf die Modellierung

Abb. 7.35. *Tensorprodukt-Basis basierend auf linearen univariaten TP-Basen.*

von Oberflächen mit Hilfe von Basisfunktionen übertragen und führt zu den sogenannten Tensorprodukt-Basen. Dabei werden alle paarweisen Produkte univariater Basisfunktionen gebildet, so dass sich eine zweidimensionale Basis ergibt.

Betrachten wir dazu genauer die folgende Situation: Die Zielvariable y soll durch eine zweidimensionale Oberfläche $f(z_1, z_2)$ beschrieben werden, wobei z_1 und z_2 sowohl metrische Kovariablen als auch Koordinaten im Fall eines räumlichen Modells sein können. Dann bilden wir zunächst die univariaten Basen für z_1 und z_2 und erhalten so die Basisfunktionen $B_j^{(1)}(z_1)$, $j = 1, \ldots, d_1$, bzw. $B_k^{(2)}(z_2)$, $k = 1, \ldots, d_2$. Die Tensorprodukt-Basis besteht dann aus allen Basisfunktionen der Form

$$B_{jk}(z_1, z_2) = B_j^{(1)}(z_1) \cdot B_k^{(2)}(z_2), \quad j = 1, \ldots, d_1, \quad k = 1, \ldots, d_2,$$

so dass sich für $f(z_1, z_2)$ die folgende Darstellung ergibt:

$$f(z_1, z_2) = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \gamma_{jk} B_{jk}(z_1, z_2).$$

Speziell für Polynom-Splines erhält man die sogenannten *Tensorprodukt-Splines*, die auch als bivariate Polynom-Splines bezeichnet werden. Als Illustration zur Konstruktion von Tensorprodukt-Basen zeigt Abbildung 7.35 lineare Tensorprodukt-Splines basierend auf den univariaten TP-Basisfunktionen

Abb. 7.36. *Tensorprodukt-Basisfunktionen basierend auf univariaten B-Splines vom Grad* $l = 0, 1, 2$ *und* 3.

$$B_1^{(1)}(z_1) = 1, \quad B^{(1)}(z_1) = z_1, \quad B_3^{(1)}(z_1) = (z_1 - \kappa_1)_+$$

und

$$B_1^{(2)}(z_2) = 1, \quad B_2^{(2)}(z_2) = z_2, \quad B_3^{(2)}(z_2) = (z_2 - \kappa_2)_+.$$

Die links oben abgebildete, konstante Funktion ergibt sich aus dem Produkt der beiden univariaten konstanten Basisfunktionen $B_1^{(1)}$ und $B_1^{(2)}$. Die erste Zeile bzw. die erste Spalte erhält man, indem man die konstante Basisfunktion in z_1-Richtung mit den Basisfunktionen in z_2-Richtung multipliziert bzw. umgekehrt. Die übrigen vier Basisfunktionen entsprechen den verbleibenden Produkten der univariaten Basisfunktionen.

Zur Regularisierung von Tensorprodukt-TP-Splines werden wieder Strafterme in Form quadrierter Koeffizienten konstruiert, vergleiche Abschnitt 7.1.3. Die dort beschriebenen numerischen Probleme treten in zweidimensionalen Fragestellungen jedoch im Vergleich zum univariaten Fall noch verstärkt auf. Wir betrachten deshalb im Folgenden Tensorprodukte der numerisch stabileren B-Spline-Basis. Einzelne Tensorprodukt-B-Splines sind in Abbildung 7.36 für die Splinegrade $l = 0, 1, 2$ und 3 abgebildet. Dabei erkennt man wieder die größere Glattheit, die man für einen höheren Splinegrad erhält. Insbesondere sind Tensorprodukt-Splines vom Grad $l = 0$ nicht stetig und Tensorprodukt-Splines vom Grad $l = 1$ stetig, aber nicht differenzierbar (vergleiche die Definition bivariater Polynom-Splines in Dierckx (1993) für eine genauere Beschreibung der Stetigkeits- und Differenzierbarkeitseigenschaften von Tensorprodukt-Splines). Abbildung 7.37 zeigt zusätzlich eine größere Menge von kubischen B-Spline-Basisfunktionen. Um eine anschaulichere Abbildung zu erhalten, sind jedoch nicht alle Basisfunktionen einer vollständigen

Abb. 7.37. *Teile einer Tensorprodukt-Basis basierend auf kubischen univariaten B-Splines.*

Abb. 7.38. *Konturplots zu Tensorprodukt-B-Spline-Basisfunktionen vom Grad $l = 1, 2$ und 3.*

Basis abgebildet, da hier eine wesentlich stärkere Überlappung stattfinden würde, ähnlich wie wir sie für univariate B-Splines in Abschnitt 7.1.1 gesehen haben.

Betrachtet man die Konturplots (also die Höhenlinien) der Tensorprodukt-B-Splines (Abbildung 7.38), so erkennt man, dass sich die Höhenlinien insbesondere für einen geringen Splinegrad deutlich von Kreisen unterscheiden. Man spricht daher davon, dass Tensorprodukt-B-Splines nicht radial sind. Verschiedene radiale Basisfunktionen lernen wir noch in Abschnitt 7.2.2 und 7.2.3 kennen.

Obwohl Tensorprodukt-Ansätze auf den ersten Blick wesentlich komplexer erscheinen als univariate Basisfunktionenansätze, können auch sie in Form großer linearer Modelle dargestellt werden. Dazu definiert man die Designmatrix $\boldsymbol{Z}$ mit Zeilen

$$\boldsymbol{z}_i' = (B_{11}(z_{i1}, z_{i2}), \ldots, B_{d_1 1}(z_{i1}, z_{i2}), \ldots, B_{1d_2}(z_{i1}, z_{i2}), \ldots, B_{d_1 d_2}(z_{i1}, z_{i2}))$$

sowie den Vektor der zugehörigen Regressionskoeffizienten

$$\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{d_1 1}, \ldots, \gamma_{1d_2}, \ldots, \gamma_{d_1 d_2})'$$

und erhält damit die bekannte Regressionsgleichung

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{7.23}$$

Prinzipiell lassen sich also auch bivariate Glättungsansätze wieder im Rahmen von linearen Modellen schätzen. Die Zahl der zu schätzenden Parameter ist im Vergleich zum

Abb. 7.39. *Räumliche Nachbarschaften auf einem regulären Gitter: Die Nachbarn des Koeffizienten γ_{jk} sind jeweils durch schwarze Punkte gekennzeichnet.*

univariaten Fall üblicherweise wesentlich größer, so dass die numerisch effiziente Bestimmung von $\hat{\boldsymbol{\gamma}}$ eine noch größere Rolle spielt. Dabei kann man für Tensorprodukt-B-Splines beispielsweise wieder die dünn besetzte Struktur der Designmatrix ausnutzen, vergleiche die entsprechenden Anmerkungen auf Seite 306.

Wie für univariate Polynom-Splines stellt sich wieder die Frage nach der optimalen Wahl der Knoten, die zur Konstruktion des Polynom-Splines herangezogen werden sollen. Darüber hinaus ergibt sich bei der Modellierung von Oberflächen häufig das Problem, dass in bestimmten Datenbereichen keine Beobachtungen vorliegen, so dass die Koeffizienten der in diesem Bereich liegenden Basisfunktionen nicht geschätzt werden können. Dieses Phänomen lässt sich etwa anhand des Beispiels zur Gehirnkartierung erläutern. Durch Tensorprodukt-Ansätze werden über dem gesamten Bereich $[\min(z_1), \max(z_1)] \times [\min(z_2), \max(z_2)]$ Basisfunktionen konstruiert. Durch die Form des Gehirns sind jedoch insbesondere in den Ecken des entstehenden Rechtecks keine Beobachtungen möglich, so dass für die hier liegenden Basisfunktionen keine Daten vorhanden sind und damit auch keine Regressionskoeffizienten geschätzt werden können. Die gleichen Probleme können prinzipiell auch für univariate B-Splines auftreten, wenn die Daten große Lücken aufweisen. Sie sind dort jedoch seltener als im zweidimensionalen Fall. Sowohl die Frage der Knotenwahl als auch die durch Datenlücken entstehenden Probleme lassen sich durch die Regularisierung mit Straftermen beheben.

2D-Penalisierungsansätze

Dazu müssen geeignete Strafterme für bivariate Basisfunktionen eingeführt werden. Während man für die Tensorprodukt-TP-Basis wieder eine Form der Ridge-Bestrafung wählen kann, bietet es sich für Tensorprodukt-B-Splines an, die räumliche Anordnung der Basisfunktionen und damit der Regressionskoeffizienten auszunutzen. Für univariate B-Splines wurden Strafterme durch quadrierte Differenzen zu Koeffizienten benachbarter Basisfunktionen eingeführt. Um dieses Konzept auf den zweidimensionalen Fall übertragen zu können, müssen wir also zunächst geeignete räumliche Nachbarschaften definieren. Abbildung 7.39 zeigt mögliche Nachbarschaftsdefinitionen für vier, acht und zwölf Nachbarn. Im Folgenden werden wir verschiedene Penalisierungsansätze basierend auf diesen Nachbarschaften einführen.

Gehen wir zunächst von der einfachsten Nachbarschaft mit vier nächsten Nachbarn aus. Dann liegt es nahe, den Strafterm basierend auf quadrierten Differenzen zwischen γ_{jk} und

diesen Nachbarn zu definieren. Seien dazu $\boldsymbol{D}_1$ und $\boldsymbol{D}_2$ univariate Differenzenmatrizen erster Ordnung in z_1- und z_2-Richtung. Dann lassen sich die zeilenweisen ersten Differenzen durch Anwendung der aufgeblähten Differenzenmatrix $\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1$ auf den Vektor $\boldsymbol{\gamma}$ bestimmen, wobei $\boldsymbol{I}_d$ die d-dimensionale Einheitsmatrix und $\otimes$ das Kroneckerprodukt bezeichnen (vergleiche Definition A.9 in Anhang A). Tatsächlich erhält man durch Anwendung auf den Vektor der Regressionskoeffizienten

$$\boldsymbol{\gamma}'(\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1)'(\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1)\boldsymbol{\gamma} = \sum_{k=1}^{d_2} \sum_{j=2}^{d_1} (\gamma_{jk} - \gamma_{j-1,k})^2,$$

also die Summe aller zeilenweise gebildeten quadrierten Differenzen. Analog ergeben sich die spaltenweisen quadrierten Differenzen durch

$$\boldsymbol{\gamma}'(\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1})'(\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1})\boldsymbol{\gamma} = \sum_{j=1}^{d_1} \sum_{k=2}^{d_2} (\gamma_{jk} - \gamma_{j,k-1})^2.$$

Durch Quadrieren und Aufsummieren sowohl der zeilenweisen als auch der spaltenweisen Differenzen ergibt sich dann der Strafterm

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}' \left[(\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1)'(\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1) + (\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1})'(\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1}) \right] \boldsymbol{\gamma}. \tag{7.24}$$

Mit Hilfe von Rechenregeln für Kroneckerprodukte (vergleiche Satz A.4 in Anhang A) kann man zeigen, dass $\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$ sich ebenfalls berechnen lässt als

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}' \left[\boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1 + \boldsymbol{K}_2 \otimes \boldsymbol{I}_{d_1} \right] \boldsymbol{\gamma},$$

mit den univariaten Strafmatrizen $\boldsymbol{K}_1 = \boldsymbol{D}_1' \boldsymbol{D}_1$ und $\boldsymbol{K}_2 = \boldsymbol{D}_2' \boldsymbol{D}_2$.

Die Bayesianische Variante der Regularisierung ergibt sich wie folgt: Da sich die zeilen- und spaltenweisen Differenzenbestrafungen wie im univariaten Fall über Random Walks erster Ordnung definieren lassen, kann man $\boldsymbol{K}$ als Präzisionsmatrix der gemeinsamen Verteilung des Vektors $\boldsymbol{\gamma}$ unter Annahme eines zweidimensionalen Random Walks erster Ordnung interpretieren. Genauer ergibt sich die Priori-Verteilung für $\boldsymbol{\gamma}$ als

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{2\pi\tau^2} \right)^{\mathrm{rg}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} \right). \tag{7.25}$$

Berechnet man daraus die bedingte Verteilung von γ_{jk} gegeben alle übrigen Koeffizienten, so erhält man (außerhalb der Randbereiche)

$$\gamma_{jk}|\cdot \sim \mathrm{N}\left(\frac{1}{4}(\gamma_{j-1,k} + \gamma_{j+1,k} + \gamma_{j,k-1} + \gamma_{j,k+1}), \frac{\tau^2}{4} \right).$$

Für den Vektor $\boldsymbol{\gamma}$ gilt also eine räumliche Form der Markov-Eigenschaft, da die bedingte Verteilung von γ_{jk} nur von den vier nächsten (räumlichen) Nachbarn abhängt. Insbesondere ergibt sich als Erwartungswert der bedingten Verteilung der lokale Mittelwert, gebildet aus den vier nächsten Nachbarn. Damit verallgemeinert der Strafterm für zweidimensionale Oberflächen eine Eigenschaft, die wir in Abschnitt 7.1.2 für univariate P-Splines kennengelernt haben.

Tensorprodukt-P-Splines

Ziel ist die Schätzung der bivariaten Funktion f im Modell

$$y = f(z_1, z_2) + \varepsilon$$

für metrische Kovariablen z_1 und z_2. Tensorprodukt-Splines werden konstruiert durch Bilden aller paarweisen Produkte univariater Polynom-Splines für z_1 und z_2, also

$$B_{jk}(z_1, z_2) = B_j^{(1)}(z_1) \cdot B_k^{(2)}(z_2), \quad j = 1, \ldots, d_1, \quad k = 1, \ldots, d_2.$$

Damit ergibt sich die Darstellung

$$f(z_1, z_2) = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \gamma_{jk} B_{jk}(z_1, z_2)$$

für $f(z_1, z_2)$ und man erhält die Modellgleichung

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

wobei $\boldsymbol{Z}$ die ausgewerteten Basisfunktionen und $\boldsymbol{\gamma}$ die entsprechenden Regressionskoeffizienten enthält.
Strafterme lassen sich durch Kroneckerprodukte der univariaten Strafmatrizen konstruieren:

$$\boldsymbol{K} = \boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1 + \boldsymbol{K}_2 \otimes \boldsymbol{I}_{d_1}.$$

Es ergibt sich ein quadratischer Strafterm

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma},$$

so dass die in Abschnitt 7.1.9 behandelten Verfahren zur Bestimmung von λ angewendet werden können.

Allgemeiner kann man das auf zeilen- und spaltenweisen Differenzen basierende Prinzip zur Konstruktion zweidimensionaler Strafterme natürlich auch auf Differenzenmatrizen höherer Ordnung anwenden und erhält so Strafterme der Form

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}' \left[\boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1^{(k_1)} + \boldsymbol{K}_2^{(k_2)} \otimes \boldsymbol{I}_{d_1} \right] \boldsymbol{\gamma}$$

mit univariaten Strafmatrizen $\boldsymbol{K}_1^{(k_1)}$ und $\boldsymbol{K}_2^{(k_2)}$ der Ordnungen k_1 und k_2. Beispielsweise ergibt sich für $k_1 = k_2 = 2$ ein Strafterm basierend auf zweiten Differenzen dessen Nachbarschaftsstruktur aus den acht nächsten Nachbarn entlang der Koordinatenachsen besteht (vergleiche Abbildung 7.39).

Zusammenfassend lassen sich die in diesem Abschnitt behandelten Ansätze für 2D-Strafterme (und auch eine Reihe von Erweiterungen) in Kombination mit Tensorprodukt-B-Splines wieder in die allgemeine Form bringen, die wir bereits in Abschnitt 7.1 ken-

nengelernt haben: Der Vektor der Funktionsauswertungen lässt sich als großes lineares Modell der Form $\boldsymbol{Z}\boldsymbol{\gamma}$ mit einem quadratischen Strafterm $\lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$ bzw. einer entsprechenden Gauß-Priori für $\boldsymbol{\gamma}$ mit Dichte (7.25) schreiben. Damit sind auch die in Abschnitt 7.1.9 besprochenen Schätzverfahren prinzipiell wieder anwendbar. Da die Dimension des betrachteten Modells jedoch wesentlich größer ist als im univariaten Fall, ist die Verwendung numerisch effizienter Verfahren insbesondere bei der Implementation von MCMC-Algorithmen noch entscheidender.

Für die Schätzung mit Hilfe von gemischten Modellen benötigen wir darüber hinaus noch die die Zerlegung

$$\boldsymbol{\gamma} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}}$$

definierenden Designmatrizen. Wie wir in Abschnitt 7.1.9 gesehen haben, ergibt sich $\tilde{\boldsymbol{X}}$ aus einer Basis des Nullraums der Matrix $\boldsymbol{K}$. Um diesen Nullraum zu bestimmen, beschäftigen wir uns mit der Frage, welche Effekte durch die behandelten Strafmatrizen unpenalisiert bleiben. Verwendet man den Kroneckerprodukt-Ansatz mit ersten Differenzen, so lässt die Addition einer Konstanten sowohl den spaltenweisen als auch den zeilenweisen Strafterm unverändert. Tatsächlich kann man zeigen, dass $\mathrm{rg}(\boldsymbol{K}) = d_1 d_2 - 1$ gilt, so dass tatsächlich nur die Konstante unpenalisiert bleibt und der Nullraum der Matrix $\boldsymbol{K}$ durch einen $d_1 d_2$-dimensionalen Vektor von Einsen gebildet wird. Im Allgemeinen ist der Nullraum von $\boldsymbol{K}$ für Kroneckerprodukt-Strafmatrizen gegeben durch das Tensorprodukt der Nullräume der univariaten Strafmatrizen. Durch zweite Differenzen ergibt sich also ein vierdimensionaler Nullraum, der neben der Konstanten lineare Effekte in z_1- und z_2-Richtung sowie die Interaktion $z_1 \cdot z_2$ enthält. Die Basis dieses Nullraums lässt sich mit Hilfe der univariaten Knotenmengen $\kappa_1^{(1)}, \ldots, \kappa_{d_1}^{(1)}$ und $\kappa_1^{(2)}, \ldots, \kappa_{d_2}^{(2)}$ darstellen:

$$\begin{pmatrix}
1 & \kappa_1^{(1)} & \kappa_1^{(2)} & \kappa_1^{(1)}\kappa_1^{(2)} \\
1 & \kappa_2^{(1)} & \kappa_1^{(2)} & \kappa_2^{(1)}\kappa_1^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_{d_1}^{(1)} & \kappa_1^{(2)} & \kappa_{d_1}^{(1)}\kappa_1^{(2)} \\
1 & \kappa_1^{(1)} & \kappa_2^{(2)} & \kappa_1^{(1)}\kappa_2^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_{d_1}^{(1)} & \kappa_2^{(2)} & \kappa_{d_1}^{(1)}\kappa_2^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_1^{(1)} & \kappa_{d_2}^{(2)} & \kappa_1^{(1)}\kappa_{d_2}^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_{d_1}^{(1)} & \kappa_{d_2}^{(2)} & \kappa_{d_1}^{(1)}\kappa_{d_2}^{(2)}
\end{pmatrix}.$$

Zur Konstruktion von $\tilde{\boldsymbol{U}}$ können wir wie in Abschnitt 7.1.9 vorgehen, das heißt $\tilde{\boldsymbol{U}}$ wird mit Hilfe der Eigenwertzerlegung von $\boldsymbol{K}$ bestimmt. Man beachte dabei, dass hier tatsächlich die Eigenwertzerlegung verwendet werden muss, da sich $\boldsymbol{K}$ nicht in der Form $\boldsymbol{K} = \boldsymbol{D}'\boldsymbol{D}$ darstellen lässt.

Beispiel 7.9 Mietspiegel München – Interaktion zwischen Baujahr und Wohnfläche

Wir verwenden Tensorprodukt-Splines mit unterschiedlichen Straftermen, um den Interaktionseffekt zwischen Baujahr und Wohnfläche im Mietspiegelbeispiel zu untersuchen

Abb. 7.40. *Mietspiegel München: Geschätzte Interaktionsoberflächen zwischen Wohnfläche in Quadratmetern und Baujahr basierend auf kubischen B-Splines mit jeweils 20 inneren Knoten sowie ersten und zweiten Differenzen als Strafterm.*

(vergleiche Beispiel 7.5 (Seite 368)). Abbildung 7.40 zeigt Schätzungen dieses Effekts basierend auf kubischen B-Splines mit jeweils 22 Knoten sowie ersten und zweiten Differenzen.

Beide Schätzungen geben im Wesentlichen die gleiche Struktur wieder, mit höheren Quadratmetermieten für neue und kleine Wohnungen. Man erkennt deutlich, dass tatsächlich ein Interaktionseffekt vorliegt und die Schätzung univariater Funktionen von Baujahr und Wohnfläche vermutlich nicht ausreicht. Vergleicht man die Glattheit der beiden Schätzungen, so führt, wie bereits im univariaten Fall, die Verwendung höherer Differenzen zur Konstruktion des Strafterms tendenziell zu glatteren Funktionen.

△

Beispiel 7.10 Analyse des Waldzustands in Baden-Württemberg

Als zweites Beispiel analysieren wir die räumliche Verteilung der Kalziumkonzentration in Baden-Württemberg (vergleiche Beispiel 7.6). Abbildung 7.41 zeigt eine Schätzung für kubische Tensorprodukt-B-Splines mit jeweils 17 Knoten für Längen- und Breitengrad sowie einem auf ersten Differenzen basierenden Strafterm. Die geschätzte Oberfläche reproduziert den in der deskriptiven Analyse erhaltenen räumlichen Trend (Abbildung 7.32 auf Seite 369), besitzt aber den Vorteil, dass sie in einer wesentlich feineren Auflösung visualisiert und auch in Bereiche ohne Beobachtungen extrapoliert werden kann. Augustin, Lang, Musio & von Wilpert (2007) verwenden diese extrapolierten Werte, um darauf aufbauend den Waldzustand in Baden-Württemberg zu analysieren.

△

7.2.2 Radiale Basisfunktionen

Eine Alternative zur Konstruktion von Tensorprodukt-Basen ist die Verwendung einer anderen Klasse bivariater Basisfunktionen, der sogenannten radialen Basen. Allgemein lassen sich radiale Basisfunktionen definieren als Funktionen der euklidischen Distanz zwischen einem Knoten $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)$ und dem Beobachtungspunkt $\boldsymbol{z} = (z_1, z_2)$, d.h. als

$$B_{\boldsymbol{\kappa}}(\boldsymbol{z}) = B(||\boldsymbol{z} - \boldsymbol{\kappa}||) = B(r)$$

Abb. 7.41. *Waldzustandsanalyse in Baden-Württemberg: Geschätzter räumlicher Effekt der Kalziumkonzentration basierend auf kubischen Tensorprodukt-B-Splines mit 15 inneren Knoten und auf ersten Differenzen beruhendem Strafterm.*

für eine geeignete Funktion B und $r = ||\boldsymbol{z} - \boldsymbol{\kappa}|| = ((z_1 - \kappa_1)^2 + (z_2 - \kappa_2)^2)^{0.5}$. Die Bezeichnung radiale Basisfunktion ergibt sich aus der Tatsache, dass die Konturplots radialer Basisfunktionen gemäß ihrer Konstruktion aus kreisförmigen Höhenlinien bestehen.

Alle radialen Basisfunktionen besitzen die gleiche funktionale Form und sind, im Gegensatz zu Tensorprodukt-B-Splines, genau einem Knoten zugeordnet. Üblicherweise sind die verwendeten Knoten eine Teilmenge der Beobachtungspunkte, also $\{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_d\} \subset \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$. Damit passt sich die Verteilung der Basisfunktionen automatisch an die Verteilung der Daten an, während für Tensorprodukt-Basen eine große Zahl von Basisfunktionen in Bereichen liegen kann, in denen keine Beobachtungen gemacht wurden. Dies ist etwa im Beispiel zur Analyse von Waldschäden der Fall. Hier werden durch den Tensorprodukt-Ansatz auch außerhalb von Baden-Württemberg eine große Zahl von Knoten und damit Basisfunktionen gesetzt, da die Knoten gleichmäßig über $[\min(z_1), \max(z_1)] \times [\min(z_2), \max(z_2)]$ verteilt werden. Andererseits erhält man so die Möglichkeit, auch in diese Bereiche hinein zu extrapolieren, während der Extrapolationsbereich für radiale Basisfunktionen eingeschränkter ist.

Das bekannteste Beispiel radialer Basisfunktionen erhält man durch die Optimierung des Kriteriums

$$\sum_{i=1}^{n} (y_i - f(\boldsymbol{z}_i))^2 + \lambda \int\int \left[\left(\frac{\partial^2}{\partial^2 z_1} + 2 \frac{\partial^2}{\partial z_1 \partial z_2} + \frac{\partial^2}{\partial^2 z_2} \right) f(z_1, z_2) \right]^2 dz_1 dz_2 \to \max_f$$

über die Klasse aller zweimal stetig differenzierbaren Funktionen$f(\boldsymbol{z})$. Dabei entspricht

$$\int\int \left[\left(\frac{\partial^2}{\partial^2 z_1} + 2 \frac{\partial^2}{\partial z_1 \partial z_2} + \frac{\partial^2}{\partial^2 z_2} \right) f(z_1, z_2) \right]^2 dz_1 dz_2$$

dem bivariaten Analogon der quadrierten zweiten Ableitung, die wir im Zusammenhang mit Glättungssplines betrachtet haben. Ziel ist also die Erweiterung des Glättungsspline-Ansatzes auf die Schätzung von Oberflächen. Als Lösung ergibt sich der sogenannte *Thin-Plate-Spline*, der wie natürliche kubische Splines die natürlichen Randbedingungen

erfüllt und damit außerhalb des Beobachtungsbereichs linear ausläuft. Ein Thin-Plate-Spline lässt sich mit Hilfe von radialen Basisfunktionen darstellen als

$$f(z_1, z_2) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \sum_{j=1}^{n} \gamma_j B_j(z_1, z_2)$$

mit

$$B_j(z_1, z_2) = B(||\boldsymbol{z} - \boldsymbol{z}_j||) = ||\boldsymbol{z} - \boldsymbol{z}_j||^2 \log(||\boldsymbol{z} - \boldsymbol{z}_j||).$$

Der Thin-Plate-Spline basiert also auf radialen Basisfunktionen

$$B(r) = r^2 \log(r).$$

Andere typischerweise verwendete radiale Basisfunktionen sind etwa

$$B(r) = r^l, \quad l \text{ ungerade},$$

oder

$$B(r) = \sqrt{r^2 + c^2} \quad \text{für eine Konstante } c > 0.$$

Im folgenden Abschnitt werden wir sehen, dass sich auch stationäre Gauß-Felder, wie wir sie bereits in Abschnitt 7.1.6 kennengelernt haben, als radiale Basisfunktionenansätze interpretieren lassen.

Auch auf Modelle mit radialen Basisfunktionen lassen sich die in Abschnitt 7.1.9 behandelten Schätzverfahren für Penalisierungsansätze anwenden. Der Strafterm ergibt sich dabei aus den quadrierten zweiten Ableitungen, aus denen sich wie im univariaten Fall ein quadratischer Strafterm herleiten lässt.

7.2.3 Kriging: Räumliche Glättung bei stetiger Lokationsvariable

In Abschnitt 7.1.6 haben wir bereits die Möglichkeit kennengelernt, zeitliche Korrelationen mit Hilfe von stationären Gauß-Prozessen und parametrischen Korrelationsfunktionen zu beschreiben. Wir wollen nun diesen Ansatz übertragen auf die Modellierung räumlicher Korrelationen durch stationäre Gauß-Felder. Dabei handelt es sich im Gegensatz zu den bisher betrachteten Basisfunktionenansätzen zur Modellierung von Interaktionsoberflächen zunächst um einen probabilistischen Ansatz, d.h. der räumliche Effekt wird durch einen stochastischen Prozess beschrieben. Wie im univariaten Fall wird sich jedoch herausstellen, dass die probabilistische Modellformulierung äquivalent ist zu einem speziellen Basisfunktionenansatz.

Allgemein lässt sich ein *Gauß-Feld* $\{\gamma(s), s \in \mathbb{R}^2\}$ charakterisieren durch die Erwartungswertfunktion $\mu(s) = \mathrm{E}(\gamma(s))$, die Varianzfunktion $\tau^2(s) = \mathrm{Var}(\gamma(s))$ und die Korrelationsfunktion $\rho(s,t) = \mathrm{Corr}(\gamma(s), \gamma(t))$. Für *stationäre* Gauß-Felder sind Erwartungswert und Varianz räumlich konstant ($\mu(s) \equiv \mu$ und $\tau^2(s) \equiv \tau^2$) und die Korrelationsfunktion hängt nur noch von der Differenz $s - t$ ab, so dass man $\rho(s,t) = \rho(s-t) = \rho(h)$ mit $h = s - t$ erhält. Häufig beschränkt man sich bei der Analyse räumlicher Korrelationen sogar auf den Spezialfall *isotroper* Korrelationsfunktionen, für die

$$\rho(s,t) = \rho(||s - t||) = \rho(r)$$

Abb. 7.42. *Höhenlinien anisotroper Korrelationsfunktionen, die mit Hilfe des Abstandsmaßes (7.26) konstruiert wurden.*

mit $r = ||s - t||$ gilt. Die Korrelation zwischen zwei Punkten s und t hängt also nur noch von ihrem euklidischen Abstand ab, aber nicht mehr von der genauen Position der Punkte und auch nicht von der Richtung des Vektors zwischen den beiden Punkten. Für die Korrelationsfunktion ρ lassen sich unter der Annahme der Isotropie dann wieder alle in Abschnitt 7.1.6 (Seite 327) besprochenen parametrischen Klassen anwenden, da die eigentlich bivariate, räumliche Korrelationsfunktion in diesem Fall nur noch von dem Skalar r abhängt.

Eine einfache Möglichkeit Anisotropie zuzulassen erhält man, indem man den euklidischen Abstand

$$||s - t|| = \sqrt{(s-t)'(s-t)}$$

ersetzt durch

$$\sqrt{(s-t)'R(\psi)'D(\delta)R(\psi)(s-t)}, \tag{7.26}$$

wobei $R(\psi)$ eine Rotationsmatrix mit Anisotropie-Winkel $\psi \in [0, 2\pi]$ bezeichnet, also

$$R(\psi) = \begin{pmatrix} \cos(\psi) & \sin(\psi) \\ -\sin(\psi) & \cos(\psi) \end{pmatrix},$$

und $D(\delta)$ eine Dehnungsmatrix mit Anisotropie-Verhältnis $\delta \geq 1$, d.h.

$$D(\delta) = \begin{pmatrix} \delta^{-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Abbildung 7.42 zeigt die durch verschiedene Wahlen von ψ und δ entstehenden Höhenlinien des resultierenden, anisotropen Abstandsmaßes. Durch (7.26) lassen sich allerdings nur Korrelationsfunktionen mit elliptischen Höhenlinien definieren. Allgemeinere Ansätze zur Konstruktion anisotroper Korrelationsfunktionen lassen aber auch komplexer verlaufende Höhenlinien zu.

Klassische Geostatistik

Entwickelt wurden Kriging-Ansätze in der räumlichen Statistik zur Glättung oder Interpolation räumlicher Phänomene. Ausgangspunkt hierfür ist das Modell

$$y(s) = \boldsymbol{x}(s)'\boldsymbol{\beta} + \gamma(s) + \varepsilon(s),$$

das auch als klassisches geostatistisches Modell bezeichnet wird. Dabei sind

$\boldsymbol{x}(s)'\boldsymbol{\beta}$ der durch Kovariablen $\boldsymbol{x}$ parametrisierte räumliche Trend,

$\gamma(s)$ ein stationärer Gauß-Prozess mit Erwartungswert 0, Varianz τ^2 und Korrelationsfunktion $\rho(h)$, sowie

$\varepsilon(s)$ der übliche Fehlerterm, also $\varepsilon(s) \sim \mathrm{N}(0, \sigma^2)$ (unabhängig von $\gamma(s)$).

Falls der räumliche Trend konstant ist, also $\boldsymbol{x}(s)'\boldsymbol{\beta} \equiv \beta_0$, so spricht man auch von gewöhnlichem Kriging, andernfalls von universellem Kriging.
In Matrixnotation ergibt sich das Kriging-Modell zu

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

wobei $\boldsymbol{\gamma} = (\boldsymbol{\gamma}(s_{(1)}), \ldots, \boldsymbol{\gamma}(s_{(d)}))'$ die Werte des stationären Gauß-Prozesses an den d verschiedenen beobachteten räumlichen Lokationen $s_{(1)}, \ldots, s_{(d)}$ bezeichnet. Die Matrix $\boldsymbol{Z}$ entspricht einer Inzidenzmatrix mit

$$\boldsymbol{Z}[i,j] = \begin{cases} 1 & \text{falls } y_i \text{ am Punkt } s_{(j)} \text{ beobachtet wurde (d.h. es gilt } s_i = s_{(j)}) \text{ und} \\ 0 & \text{sonst.} \end{cases}$$

Für die Kovarianzmatrix der Zielvariablen $\boldsymbol{y}$ ergibt sich aus den Modellannahmen

$$\mathrm{Cov}(\boldsymbol{y}) = \tau^2 \boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}' + \sigma^2 \boldsymbol{I}_n$$

mit der Kovarianzmatrix der räumlichen Effekte

$$\boldsymbol{R} = (\mathrm{Corr}(\gamma(s_{(i)}), \gamma(s_{(j)})) = (\rho(s_{(i)} - s_{(j)})),$$

so dass sich die räumliche Korrelation des Vektors $\boldsymbol{\gamma}$ auf $\boldsymbol{y}$ überträgt.
Ziel der klassischen Geostatistik ist es, optimale Vorhersagen für $\gamma(s_0)$ oder $y(s_0) = \boldsymbol{x}(s_0)'\boldsymbol{\beta} + \gamma(s_0)$ an neuen Lokationen s_0 zu erhalten. Unter Verwendung der Bezeichnungen $\boldsymbol{r} = (\rho(s_1 - s_0), \ldots, \rho(s_n - s_0))'$, $y_0 = y(s_0)$ und $\mu_0 = \boldsymbol{x}(s_0)'\boldsymbol{\beta}$ lässt sich die optimale Prognose aus der gemeinsamen Normalverteilung

$$\begin{pmatrix} \boldsymbol{y} \\ y_0 \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \tau^2 \boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}' + \sigma^2 \boldsymbol{I}_n & \tau^2 \boldsymbol{r}' \\ \tau^2 \boldsymbol{r} & \tau^2 \end{bmatrix} \right)$$

herleiten. Aus den Eigenschaften der Normalverteilung (vergleiche Satz B.4 in Anhang B.3) ergibt sich die bezüglich der mittleren quadratischen Abweichung optimale Vorhersage dann gerade als der bedingte Erwartungswert von y_0 gegeben $\boldsymbol{y}$:

$$\hat{y}_0 = \mathrm{E}(y_0|\boldsymbol{y}) = \mu_0 + \tau^2 \boldsymbol{r}'(\tau^2 \boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}' + \sigma^2 \boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Die bedingte Varianz von $\hat{y}_0$ ergibt sich ebenfalls aus der gemeinsamen Normalverteilung:

$$\mathrm{Var}(y_0|\boldsymbol{y}) = \tau^2 - \tau^2 \boldsymbol{r}'(\tau^2 \boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}' + \sigma^2 \boldsymbol{I}_n)^{-1} \boldsymbol{r} \tau^2.$$

Auch ohne die Normalverteilungsannahme kann man zeigen, dass $\hat{y}_0$ immer noch gewisse Optimalitätseigenschaften erfüllt. Wie bei den Schätzungen zufälliger Effekte in Abschnitt 6.3.1 (Seite 261) ist $\hat{y}_0$ bester linearer unverzerrter Prädiktor.

Kriging als Basisfunktionenansatz

Mit Hilfe der klassischen Geostatistik erhält man also optimale Vorhersagen des räumlichen Effekts für beliebige Punkte s_0, aber im Gegensatz zu den Basisfunktionenansätzen, die wir bisher behandelt haben, keine kompakte Darstellung dieser Prognosen. Eine solche Möglichkeit ergibt sich aus unseren Betrachtungen in Abschnitt 7.1.6 (Seite 327). Dort haben wir gesehen, dass sich der stochastische Kriging-Ansatz tatsächlich umschreiben lässt in einen Basisfunktionenansatz mit aus der Korrelationsfunktion $\rho(h)$ gebildeten Basisfunktionen. Diese Reparametrisierung lässt sich auch für das räumliche Kriging-Modell anwenden, so dass sich das klassische geostatistische Modell darstellen lässt als

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon},$$

mit Designmatrix

$$\tilde{\boldsymbol{Z}}[i,j] = \rho(s_i, s_{(j)}).$$

Für die einzelnen Beobachtungen ergibt sich somit das Modell

$$y(s) = \boldsymbol{x}(s)'\boldsymbol{\beta} + f_{geo}(s) + \varepsilon(s)$$

mit dem räumlichen Effekt

$$f_{geo}(s) = \sum_{j=1}^{d} \tilde{\gamma}_j B_j(s)$$

und den auf der Korrelationsfunktion basierenden Basisfunktionen

$$B_j(s) = \rho(s, s_{(j)}).$$

Für isotrope Korrelationsfunktionen erhält man insbesondere radiale Basisfunktionen der Form $B_j(s) = \rho(||s - s_{(j)}||)$. Der Strafterm für $\tilde{\boldsymbol{\gamma}}$ ergibt sich wie in Abschnitt 7.1.6 aus der Korrelationsmatrix $\boldsymbol{R}$:

$$\lambda\tilde{\boldsymbol{\gamma}}'\boldsymbol{K}\tilde{\boldsymbol{\gamma}} = \frac{\sigma^2}{\tau^2}\tilde{\boldsymbol{\gamma}}'\boldsymbol{R}\tilde{\boldsymbol{\gamma}}.$$

Wählt man speziell Matérn-Korrelationsfunktionen, so ergibt sich die Klasse der sogenannten Matérn-Splines. Abbildung 7.43 zeigt zwei isotrope Matérn-Spline-Basisfunktionen, die sich für die Hyperparameter $\kappa = 0.5$ und $\kappa = 1.5$ ergeben. Zusätzlich sind Konturplots eingezeichnet, die noch einmal die Radialität der so entstehenden Basisfunktionen verdeutlichen.

In der Bayesianischen Betrachtungsweise entspricht der Kriging-Ansatz wieder einer bestimmten Glattheits-Priori, die in der Basisfunktionen-Repräsentation durch

$$\tilde{\boldsymbol{\gamma}} \sim \mathrm{N}(0, \tau^2\boldsymbol{R}^{-1}) \tag{7.27}$$

gegeben ist.

Im Unterschied zu den Tensorprodukt-Ansätzen, die wir für Polynom-Splines betrachtet haben, ergeben sich die Knoten eines Kriging-Ansatzes aus den im Datensatz beobachteten Lokationen. Üblicherweise entspricht die Anzahl der Knoten damit etwa dem Stichprobenumfang, so dass sich eine sehr große Zahl zu schätzender Koeffizienten ergibt. Um die damit verbundenen großen Rechenzeiten zu vermeiden, verwendet man in der Praxis häufig nur eine Teilmenge

Abb. 7.43. *Perspektiv- und Konturplots von Matérn-Spline-Basisfunktionen für* $\kappa = 0.5$ *und* $\kappa = 1.5$.

$$\mathcal{D} = \{\kappa_1, \ldots, \kappa_m\} \subset \mathcal{C} = \{s_{(1)}, \ldots, s_{(d)}\}$$

der beobachteten Lokationen als Knoten. Basierend auf geeigneten Überdeckungskriterien kann diese Teilmenge repräsentativ für die ursprüngliche Menge von Beobachtungspunkten ausgewählt werden, vergleiche Johnson, Moore & Ylvisaker (1990) und Nychka & Saltzman (1998) für Details. Damit erhält man eine Approximation der Kriging-Oberfläche durch $\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\gamma}}$, wobei nun die Designmatrix definiert ist durch

$$\tilde{\boldsymbol{Z}}[i,j] = \rho(s_i - \kappa_j) \tag{7.28}$$

und der m-dimensionale Koeffizientenvektor besitzt die Verteilung

$$\tilde{\boldsymbol{\gamma}} \sim \mathrm{N}(0, \tau^2 \boldsymbol{R}^{-1}) \quad \text{mit} \quad \boldsymbol{R}[i,j] = \rho(\kappa_i, \kappa_j). \tag{7.29}$$

Schätzung von Kriging-Modellen

Die Schätzung von Kriging-Ansätzen kann auf zwei Arten erfolgen, die sich aus den unterschiedlichen Sichtweisen auf das geostatistische Modell ergeben. Betrachtet man dieses aus der ursprünglichen, stochastischen Perspektive, so entspricht die Annahme eines Gauß-Feldes für den räumlichen Effekt im Wesentlichen der Annahme eines räumlich korrelierten zufälligen Effekts. Dementsprechend können die in Kapitel 6 behandelten Schätzverfahren eingesetzt werden. Insbesondere können zusätzliche Parameter der Korrelationsfunktion (wie beispielsweise der Range-Parameter) mit Hilfe von ML- bzw. REML-Verfahren simultan mitbestimmt werden.

Kriging

Klassisches geostatistisches Modell

$$y(s) = \mu(s) + \gamma(s) + \varepsilon(s), \quad s \in \mathbb{R}^2.$$

Dabei bezeichnet $\mu(s) = \boldsymbol{x}(s)'\boldsymbol{\beta}$ den räumlichen Trend und $\gamma(s)$ ein (stationäres) Gauß-Feld mit Erwartungswert 0, Varianz τ^2 und parametrischer Korrelationsfunktion $\rho(h)$.

Schätzung

Für gegebene Varianzen ergeben sich die besten linearen unverzerrten Vorhersagen für $y_0 = \boldsymbol{x}(s_0)'\boldsymbol{\beta} + \gamma(s_0)$ aus der gemeinsamen Normalverteilung für $\boldsymbol{y}$ und y_0. Allgemeiner kann die Schätzung mit Hilfe der Verfahren für gemischte Modelle aus Kapitel 6 erfolgen, da $\gamma(s)$ einem räumlich korrelierten zufälligen Effekt entspricht.

Darstellung als Basisfunktionenansatz

Wie für den univariaten Kriging-Ansatz ergibt sich eine Repräsentation als Basisfunktionenansatz mit Basisfunktionen

$$B_j(s) = \rho(s, s_{(j)}) \quad \text{bzw.} \quad B_j(s) = \rho(||s - s_{(j)}||).$$

Damit erhält man die Modellgleichung

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}$$

mit $\tilde{\boldsymbol{Z}}[i,j] = \rho(s_i, s_{(j)})$ bzw. $\tilde{\boldsymbol{Z}}[i,j] = \rho(||s_i - s_{(j)}||)$. Der Strafterm für $\tilde{\boldsymbol{\gamma}}$ ergibt sich aus der Korrelationsmatrix $\boldsymbol{R}$ mit Einträgen $\boldsymbol{R}[i,j] = \rho(s_{(i)}, s_{(j)})$ zu

$$\lambda\tilde{\boldsymbol{\gamma}}'\boldsymbol{K}\tilde{\boldsymbol{\gamma}} = \frac{\sigma^2}{\tau^2}\tilde{\boldsymbol{\gamma}}'\boldsymbol{R}\tilde{\boldsymbol{\gamma}}.$$

Betrachtet man den Kriging-Ansatz dagegen als einen Basisfunktionenansatz, so werden in der Regel alle Parameter der Korrelationsfunktion vorab ausgewählt oder wie in Beispiel 7.3 anhand einfacher Regeln vorab aus den Daten geschätzt. Lediglich die Varianz τ^2 verbleibt in diesem Fall als unbekannter Parameter, der als Glättungsparameter des Basisfunktionenansatzes fungiert. Die Strafmatrix $\boldsymbol{R}$ ist dagegen vollständig spezifiziert. Zur Bestimmung von τ^2 können dann die in Abschnitt 7.1.9 behandelten Verfahren, insbesondere die Repräsentation als gemischtes Modell und die MCMC-basierten Algorithmen eingesetzt werden.

Beispiel 7.11 Gehirnkartierung

Wir setzen nun den räumlichen Kriging-Ansatz zur Analyse der Gehirnaktivierung bei visuellem Reiz ein (vergleiche Beispiel 7.7 auf Seite 370). Abbildung 7.44 zeigt zwei Schätzungen für diese Aktivierung basierend auf Matérn-Splines mit Parameter $\kappa = 1.5$

Abb. 7.44. *Gehirnkartierung: Kriging-Schätzungen für die Aktivierung zum Zeitpunkt* $t = 38$ *basierend auf Matérn-Splines mit* $\kappa = 1.5$ *und 100 bzw. 200 Knoten.*

und 100 bzw. 200 Knoten. Der Rangeparameter der Korrelationsfunktion wurde dabei gemäß der Faustregel gewählt, die wir bereits in Beispiel 7.7 verwendet haben.

Für beide Knotenzahlen erkennt man eindeutig die starke Aktivierung im Bereich des visuellen Cortex. Durch den Glättungsansatz kann nicht nur pixelweise entschieden werden, ob Aktivierung vorliegt, sondern man erhält durch die Basisfunktionenrepräsentation eine kontinuierliche Aktivierungsoberfläche. Vergleicht man die Ergebnisse der unterschiedlichen Knotenzahlen, so führt die Verwendung einer größeren Knotenzahl zu einer etwas raueren Schätzung. Üblicherweise reagiert die Funktionsschätzung aber nicht besonders sensitiv auf die Knotenzahl.

△

7.2.4 Markov-Zufallsfelder: Räumliche Glättung bei diskreter Lokationsvariable

Bisher haben wir räumliche Effekte für räumliche Koordinaten, also für stetig vorliegende räumliche Information betrachtet. Wie wir in Beispiel 7.8 (Seite 370) gesehen haben, kann räumliche Information jedoch auch diskret, in Form von Regionenvariablen $s \in \{1, \ldots, d\}$ vorliegen. Ein anderes Beispiel diskreter räumlicher Information ergibt sich beim Vorliegen von Daten, die auf einem regulären Gitter angeordnet sind. Auch die Daten zur Gehirnkartierung sind von diesem Typ, können aber aufgrund der großen Zahl von Gitterpunkten durch die Identifikation jedes Pixels mit den zugehörigen Koordinaten auch mit Hilfe stetiger räumlicher Modelle analysiert werden.

Nachbarschaften und penalisiertes KQ-Kriterium

Während im Fall stetiger räumlicher Information der Abstand zwischen zwei Lokationen einfach, etwa mit Hilfe der euklidischen Distanz, bestimmt werden kann, ist dies im Fall diskreter räumlicher Information nicht mehr möglich. Wir werden deshalb ein anderes Konzept verwenden, um die räumliche Anordnung der Daten zu beschreiben. Dieses beruht auf der Definition geeigneter Formen von Nachbarschaften. Solche Nachbarschaften lassen sich dabei auf verschiedene Art und Weise konstruieren:

- Für Regionendaten, wenn die Variable s also die Zugehörigkeit einer Beobachtung zur Region s bezeichnet, werden Nachbarschaften üblicherweise über gemeinsame Grenzen definiert (vergleiche Abbildung 7.45). Modifikationen dieser Definition können notwendig werden, wenn einige der Regionen Inseln sind oder der Beobachtungsbereich in räumlich separierte Teilbereiche zerlegt ist.

Abb. 7.45. *Nachbarschaften erster Ordnung auf einem regulären Gitter und für irregulär angeordnete Regionendaten. Die Nachbarn der schwarz gekennzeichneten Region sind in grau wiedergegeben.*

- Auf regulären Gittern werden häufig die nächsten Nachbarn auf dem Gitter verwendet, also beispielsweise die vier oder acht nächsten Nachbarn. Theoretisch können auch Nachbarschaften höherer Ordnung betrachtet werden, wie wir sie im Zusammenhang mit bivariaten Polynom-Splines kennengelernt haben. Wir werden uns im Folgenden jedoch auf einfache Nachbarschaften beschränken.

Als Bezeichnungsweise für zwei benachbarte Regionen s und r verwenden wir im Folgenden die Notation $s \sim r$.

Ähnlich wie wir es im Fall von Random Walks in Abschnitt 7.1.5 für die verschiedenen Zeitpunkte einer Zeitreihe getan haben, weisen wir jeder Region s einen eigenen Regressionskoeffizienten $f_{geo}(s) = \gamma_s$, $s = 1, \ldots, d$, zu. Da sich damit üblicherweise eine große Zahl von Koeffizienten ergibt, benötigen wir dann wieder eine geeignete Möglichkeit, glatte räumliche Effekte zu erzielen und damit die effektive Parameterzahl zu reduzieren. Für einen glatten räumlichen Effekt sollten sich die Koeffizienten benachbarter Regionen nicht zu stark unterscheiden. Wir konstruieren daher einen Strafterm basierend auf den quadrierten Differenzen zwischen Parametern benachbarter Regionen. Genauer betrachten wir das penalisierte KQ-Kriterium

$$\mathrm{PKQ}(\lambda) = \sum_{i=1}^{n} \left(y_i - f_{geo}(s_i)\right)^2 + \lambda \sum_{s=2}^{d} \sum_{r \in N(s), r<s} (\gamma_r - \gamma_s)^2,$$

wobei $N(s)$ die Menge aller Nachbarn von Region s bezeichne. Damit besteht der Strafterm aus quadrierten Differenzen aller möglichen Kombinationen benachbarter Regionen, wobei jede Kombination genau einmal im Strafterm berücksichtigt wird. Damit werden wie gewünscht zu große Abweichungen zwischen Effekten benachbarter Regionen bestraft.

Um diesen Ansatz in den Rahmen allgemeiner Penalisierungsansätze einzubetten, stellen wir das penalisierte KQ-Kriterium in Matrixnotation dar. Dazu definieren wir zunächst die Designmatrix $\boldsymbol{Z}$ mit Einträgen

$$\boldsymbol{Z}[i,s] = \begin{cases} 1 & \text{falls } y_i \text{ in Region } s \text{ beobachtet wurde (d.h. es gilt } s_i = s) \text{ und} \\ 0 & \text{sonst.} \end{cases}$$

Damit lässt sich der Vektor der Funktionsauswertungen $\boldsymbol{f}_{geo} = (f_{geo}(s_1), \ldots, f_{geo}(s_n))'$ als lineares Modell $\boldsymbol{Z\gamma}$ schreiben. Der Strafterm wiederum besteht aus quadrierten Dif-

Abb. 7.46. *Nachbarschaftsmatrix für die deutschen Kreise in zwei verschiedenen Reihenfolgen.*

ferenzen und lässt sich kompakt als quadratische Form $\lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$ darstellen mit

$$\boldsymbol{K}[s,r] = \begin{cases} -1 & s \neq r, s \sim r, \\ 0 & s \neq r, s \nsim r, \\ |N(s)| & s = r. \end{cases}$$

Damit ergibt sich durch Minimierung des penalisierten KQ-Kriteriums wieder der bekannte penalisierte KQ-Schätzer $\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}'\boldsymbol{y}$.

Die Matrix $\boldsymbol{K}$ besitzt nach ihrer Definition die Struktur einer Nachbarschaftsmatrix, da die Einträge $\boldsymbol{K}[s,r]$ nur dann von Null verschieden sind, wenn s und r Nachbarn sind. Weil $\boldsymbol{K}$ also nur sehr dünn besetzt ist, können effiziente numerische Methoden zur Speicherung und Bearbeitung eingesetzt werden. Dabei ist zu beachten, dass die Reihenfolge der Regionen im Vektor $\boldsymbol{\gamma}$ eine wesentliche Rolle für die Struktur von $\boldsymbol{K}$ spielt. Abbildung 7.46 verdeutlicht dies am Beispiel der deutschen Kreise in zwei verschiedenen Reihenfolgen. Während in der linken Grafik die Kreise gemäß ihrer Kreiskennziffer angeordnet sind, ergibt sich die Anordnung in der rechten Grafik aus dem Versuch, die von Null verschiedenen Einträge möglichst dicht an der Diagonalen zu sammeln, um $\boldsymbol{K}$ zu einer Bandmatrix mit möglichst geringer Bandweite zu machen. Die Umordnung erfolgte dabei mit Hilfe des sogenannten Cuthill-McKee-Algorithmus, vergleiche etwa George & Liu (1981). Dort findet man auch weitere Details zu numerisch effizienten Algorithmen für dünn besetzte Matrizen. Aus statistischer Perspektive werden einige dieser Algorithmen auch in Rue & Held (2005) besprochen.

Bayesianische Modellformulierung

Wie für die übrigen, bisher betrachteten Penalisierungsansätze, lässt sich auch die obige Modellformulierung Bayesianisch interpretieren. Dies führt zu den sogenannten *Markov-Zufallsfeldern*, die wir nun zunächst etwas allgemeiner einführen werden. In der Literatur wird der beschriebene Penalisierungsansatz für diskrete räumliche Lokationen typischerweise in dieser Form Bayesianisch motiviert.

Zunächst definieren wir allgemein, was unter Markov-Zufallsfeldern (MZF) zu verstehen ist. Diese verallgemeinern die zeitliche Markov-Eigenschaft, die wir für Random Walks in Abschnitt 7.1.5 (Seite 326) kennengelernt haben, auf den allgemeineren, räumlichen

Fall. Bezeichnet $D = \{1, \ldots, s, \ldots, d\}$ die Menge aller Regionen bzw. Gitterpunkte, so heißt $\boldsymbol{\gamma} = \{\gamma_s, s \in D\}$ MZF, wenn die bedingte Verteilung von γ_s gegeben alle übrigen Effekte γ_r, $r \neq s$, nur von den Nachbarn abhängt. Die zugehörige (bedingte) Dichte lässt sich also schreiben als

$$p(\gamma_s|\gamma_r, r \neq s) = p(\gamma_s|\gamma_r, r \in N(s)).$$

Betrachten wir wieder die Modellspezifikation

$$y_i = f_{geo}(s_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2),$$

in der die Verteilung von $f_{geo}(s) = \gamma_s$ nun gemäß eines MZF bestimmt sei. Eine geeignete Klasse von latenten MZF zur Beschreibung räumlicher Effekte ergibt sich aus Gauß-Markov-Zufallsfeldern (GMZF), also MZF, für die die bedingten Verteilungen (und auch die gemeinsame Verteilung) Normalverteilungen sind. Das übliche Vorgehen zur Modellspezifikation ist es dann, die bedingten Verteilungen $\gamma_s|\gamma_r$, $r \in N(s)$ vorzugeben, aus denen dann die gemeinsame Verteilung des gesamten Vektors $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ herzuleiten ist. Dabei ist zu beachten, dass nicht jede Spezifikation zulässig ist, um daraus eine gültige gemeinsame Verteilung zu erhalten. Wir wollen hier jedoch nicht auf die, zur Formulierung geeigneter Klassen bedingter Verteilungen notwendigen, theoretischen Überlegungen eingehen, da sich aus dem penalisierten KQ-Kriterium sofort eine gültige Spezifikation ableiten lässt.

Diese spezielle Klasse besitzt die Form

$$\gamma_s|\gamma_r, r \in N(s) \sim \mathrm{N}\left(\sum_{r:r\sim s} \frac{1}{|N(s)|}\gamma_r, \frac{\tau^2}{|N(s)|}\right), \tag{7.30}$$

wobei $|N(s)|$ die Anzahl der Nachbarn von Region s bezeichnet. Gemäß dieser Annahme ist der bedingte Erwartungswert des räumlichen Effekts in Region s gegeben durch das arithmetische Mittel der räumlichen Effekte in den benachbarten Regionen. Durch die Varianz τ^2 wird gesteuert, wie stark der räumliche Effekt γ_s von diesem Mittelwert abweichen kann. Wir erhalten damit eine stochastische Formulierung der gewünschten Eigenschaft, dass benachbarte Regionen einen ähnlichen räumlichen Effekt aufweisen sollen. Die gemeinsame Verteilung aller räumlichen Effekte lässt sich dann aus den bedingten Verteilungen herleiten (die in diesem Fall tatsächlich ein System konsistenter Verteilungen definieren) und besitzt die Dichte

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{2\pi\tau^2}\right)^{(d-1)/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right). \tag{7.31}$$

Diese ist nur bis auf Proportionalität definiert. Die gemeinsame Verteilung ist also teilweise uneigentlich. Die Präzisionsmatrix $\boldsymbol{K}$ entspricht genau der zuvor aus dem PKQ-Kriterium hergeleiteten Strafmatrix, so dass der Bayesianische Ansatz (7.30) tatsächlich zu einer äquivalenten Modellformulierung führt. Man beachte, dass GMZFs mit teilweise uneigentlicher Verteilung auch als *intrinsische* GMZF (IGMZF) bezeichnet werden. Es lassen sich auch MZF mit normierbarer gemeinsamer Verteilung definieren, vergleiche hierzu etwa Banerjee, Carlin & Gelfand (2003, Kap. 3).

Bisher haben wir lediglich die räumliche Nachbarschaft zur Konstruktion von MZF ausgenutzt. Es kann jedoch sinnvoll sein, diese Definition etwas zu verallgemeinern und den Einfluss der Nachbarn geeignet zu gewichten. Dazu erweitern wir (7.30) zu

Abb. 7.47. *Mietspiegel München: Räumlicher Effekt basierend auf der voll-Bayesianischen, MCMC-basierten Schätzung eines IGMZF ohne Gewichtung.*

$$\gamma_s|\gamma_r, r \in N(s) \sim \mathrm{N}\left(\frac{w_{sr}}{w_{s+}} \sum_{r:r\sim s} \gamma_r, \frac{\tau^2}{w_{s+}}\right),$$

mit symmetrischen Gewichten $w_{sr} = w_{rs}$ und $w_{s+} = \sum_{r:r\sim s} w_{sr}$. Damit ist der bedingte Erwartungswert für γ_s nun gegeben durch das *gewichtete* Mittel der Koeffizienten der Nachbarn. Zur Definition der Gewichte w_{sr} werden häufig die folgenden Strategien verwendet:

- Gleiches Gewicht für alle Nachbarn, also $w_{sr} = 1$. Damit erhält man unsere ursprüngliche MZF-Definition zurück.
- Gewichte invers proportional zum Abstand der Zentroide, z.B. $w_{sr} \propto \exp(-d(s,r))$, wobei $d(s,r)$ die euklidische Distanz zwischen den Zentroiden der Regionen s und r bezeichnet.
- Gewichte proportional zur Länge der gemeinsamen Grenze der Regionen s und r.

Sollen Gewichte berücksichtigt werden, so muss die Straf- bzw. Präzisionsmatrix $\boldsymbol{K}$ ebenfalls modifiziert werden und ergibt sich nun als

$$\boldsymbol{K}[s,r] = \begin{cases} -w_{rs} & s \neq r, s \sim r, \\ 0 & s \neq r, s \nsim r, \\ w_{s+} & s = r. \end{cases}$$

Beispiel 7.12 Mietspiegel München – Untersuchung räumlicher Abhängigkeiten

Abbildung 7.47 zeigt die MCMC-basierte Schätzung eines ungewichteten IGMZF für die Münchner Mietspiegeldaten. Vergleicht man den geschätzten Effekt mit den separaten Mittelwerten, die wir in Beispiel 7.34 (Seite 371) erhalten haben, so erkennt man deutlich den Glättungseffekt, der durch die Annahme eines IGMZF eingeführt wird. Dies

Gauß-Markov-Zufallsfelder

Modell

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mit

$$\boldsymbol{Z}[i,s] = \begin{cases} 1 & \text{falls } y_i \text{ in Region } s \text{ beobachtet wurde (d.h. es gilt } s_i = s) \text{ und} \\ 0 & \text{sonst.} \end{cases}$$

und $\boldsymbol{\gamma} = (f_{geo}(s_1), \ldots, f_{geo}(s_n))'$.

IGMZF für den räumlichen Effekt

$$\gamma_s | \gamma_r, r \in N(s) \sim \mathrm{N}\left(\sum_{r:r\sim s} \frac{w_{sr}}{w_{s+}} \gamma_r, \frac{\tau^2}{w_{s+}} \right)$$

mit $\gamma_s = f_{geo}(s)$ und $w_{sr} = w_{rs}$. Die (teilweise uneigentliche) gemeinsame Verteilung ergibt sich daraus zu

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{2\pi\tau^2} \right)^{(d-1)/2} \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} \right)$$

mit

$$\boldsymbol{K}[s,r] = \begin{cases} -w_{rs} & s \neq r, s \sim r, \\ 0 & s \neq r, s \nsim r, \\ w_{s+} & s = r. \end{cases}$$

hat insbesondere den Vorteil, dass in Bezirksvierteln, in denen nur eine geringe Zahl von Wohnungen erhoben wurde, die Schätzung durch die Einbeziehung der Nachbarschaftsinformation stabilisiert und damit weniger variabel wird. Darüber hinaus ist die Einteilung in größere Gebiete mit über- oder unterdurchschnittlicher Miete basierend auf dem glatteren Effekt sinnvoller möglich.

Überraschenderweise erkennt man in Abbildung 7.47, dass Wohnungen in der Stadtmitte eine geringere Nettomiete pro Quadratmeter aufzuweisen scheinen, während Wohnungen am Stadtrand teurer sind. Hier zeigt es sich, dass die alleinige Verwendung eines räumlichen Effekts zu schwer interpretierbaren Ergebnissen führen kann, da andere wichtige erklärende Variablen nicht berücksichtigt werden. In unserem Fall liegt die Erklärung für die überraschenden Ergebnisse darin, dass die Wohnungen in der Stadtmitte typischerweise älter sind als am Stadtrand. Berücksichtigt man diesen Effekt, wie wir es in Kapitel 8 in einer Reihe von Beispielen tun werden, so kehrt sich der Effekt um und Wohnungen in der Stadtmitte sind wie erwartet teurer als Wohnungen am Stadtrand. Es zeigt sich also, dass die *simultane* Berücksichtigung von Kovariableneffekten und räumlichem Effekt zur Erzielung interpretierbarer Schätzungen notwendig ist.

△

Räumlich autoregressive Prozesse

Speziell in der Ökonometrie werden Markov-Zufallsfelder auch zur Definition einer anderen Klasse von Modellen zur Analyse räumlicher Effekte verwendet. Hier wird in Verallgemeinerung zeitlicher autoregressiver Prozesse ein Markov-Zufallsfeld für die Zielvariable y selbst angenommen. Dies führt zu räumlich autoregressiven Prozessen (Spatially Autoregressive, SAR), deren einfachster Spezialfall gegeben ist durch die Spezifikation

$$y_s = \alpha \sum_{r \in N(s)} w_{sr} y_r + \varepsilon_s, \quad \varepsilon_s \sim \mathrm{N}(0, \sigma^2)$$

mit Gewichten w_{sr} mit $w_{ss} = 0$. Daraus erhält man die bedingten Verteilungen

$$y_s | y_r, r \neq s \sim \mathrm{N}\left(\alpha \sum_{r \in N(s)} w_{sr} y_r, \sigma^2\right),$$

so dass der Vektor der Zielvariablen $\boldsymbol{y} = \{y_s, s = 1, \ldots, d\}$ offenbar ein Markov-Zufallsfeld bildet. Soll die Annahme eines SAR-Modells um Kovariableneffekte erweitert werden, so geschieht dies häufig durch die Spezifikation eines SAR-Modells für die Residuen $y_s - \boldsymbol{x}_s'\boldsymbol{\beta}$. Alternativ können SAR-Modelle auch direkt um Kovariableneffekte erweitert werden zu

$$y_s = \boldsymbol{x}_s'\boldsymbol{\beta} + \alpha \sum_{r \in N(s)} w_{sr} y_r + \varepsilon_s, \quad \varepsilon_s \sim \mathrm{N}(0, \sigma^2).$$

Die Interpretation der Kovariableneffekte unterscheidet sich in den beiden Modellspezifikationen jedoch deutlich, da diese entweder auf die um räumliche Effekte bereinigten Zielvariablen oder direkt auf die Zielvariablen wirken.

Da sich SAR-Modelle nicht in den von uns behandelten allgemeinen Rahmen von Penalisierungsansätzen fassen lassen, gehen wir im Folgenden nicht detaillierter auf diese Modellklasse ein (vgl. hierzu etwa Anselin (1988) oder Banerjee, Carlin & Gelfand (2003), Kap. 3).

7.2.5 Fazit

Bevor wir uns noch kurz mit einigen lokalen und adaptiven Verfahren der bivariaten Glättung befassen, wollen wir die Gemeinsamkeiten und Eigenschaften der bisher behandelten Ansätze zusammenfassen. Wie sich herausgestellt hat, lassen sich bivariate Basisfunktionenansätze, stationäre Gauß-Felder und Gauß-Markov-Zufallsfelder in den in Abschnitt 7.1.3 eingeführten allgemeinen Rahmen einbetten. Als Modellgleichung ergab sich jeweils ein lineares Modell

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mit einer großen Zahl von Regressionskoeffizienten, die entweder die Koeffizienten der Basisfunktionen oder die räumlichen Effekte repräsentieren. Um die Schätzung von $\boldsymbol{\gamma}$ zu regularisieren, wurden jeweils quadratische Strafterme

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

eingeführt, die je nach Ansatz basierend auf Differenzen, Ableitungsoperatoren, Korrelationen oder Nachbarschaftsstrukturen konstruiert wurden. Aufgrund der gemeinsamen Struktur lassen sich die in Abschnitt 7.1.9 behandelten Verfahren zur Schätzung des Glättungsparameters einsetzen.

Zwischen den Ansätzen bestehen im bivariaten Fall jedoch deutlichere Unterschiede als im univariaten Fall. So unterscheidet man üblicherweise zunächst zwischen stetiger und diskreter räumlicher Information, für die jeweils spezifische Ansätze entwickelt wurden. In unserem Fall waren dies Basisfunktionenansätze und stationäre Gauß-Felder einerseits und Markov-Zufallsfelder andererseits. Diese Aufteilung ist jedoch keineswegs so eindeutig, wie es auf den ersten Blick den Anschein haben mag. So können Verfahren, die stetige räumliche Information voraussetzen, beispielsweise auch für Regionendaten eingesetzt werden, indem man die Zentroide der Regionen zur Grundlage der Analyse macht. Umgekehrt können auf diskreter räumlicher Information beruhende Modelle auch im stetigen Fall eingesetzt werden, indem etwa der Beobachtungsraum durch ein reguläres Gitter diskretisiert wird.

Vergleicht man die unterschiedlichen Ansätze bezüglich ihrer numerischen Eigenschaften, so ergeben sich unterschiedliche Aussagen, je nachdem, welches Inferenzkonzept zugrunde gelegt wird. Werden MCMC-Verfahren verwendet, so muss in jeder Iteration ein hochdimensionales Gleichungssystem zur Simulation von $\boldsymbol{\gamma}$ gelöst werden. Mit Hilfe effizienter Verfahren für dünn besetzte Matrizen lassen sich diese Berechnungen erheblich beschleunigen. Dünn besetzte Strafmatrizen $\boldsymbol{K}$ ergeben sich in den von uns behandelten Verfahren insbesondere für Gauß-Markov-Zufallsfelder und bivariate P-Splines. Sollen hingegen stationäre Gauß-Felder oder radiale Basen eingesetzt werden, so ist $\boldsymbol{K}$ in der Regel voll besetzt. In diesem Fall kann die Dimension des Schätzproblems lediglich dadurch reduziert werden, dass der Parametervektor $\boldsymbol{\gamma}$ in Teilvektoren zerlegt und diese getrennt im MCMC-Algorithmus aktualisiert werden. Da die Einträge in $\boldsymbol{\gamma}$ jedoch in der Regel hoch korreliert sind, birgt dieses Vorgehen die Gefahr, die Mischungs- und Konvergenzeigenschaften des MCMC-Algorithmus deutlich zu verschlechtern.

Beruht die Inferenz auf der Darstellung als gemischtes Modell, so spielt die Struktur der Strafmatrix $\boldsymbol{K}$ eine wesentlich geringere Rolle, da diese im Zuge der Reparametrisierung nur einmal bearbeitet werden muss. Hier ist die Dimension des Parametervektors $\boldsymbol{\gamma}$ deutlich entscheidender. Dementsprechend sind radiale Basisfunktionenansätze und stationäre Gauß-Felder ohne eine geeignete Knotenwahlstrategie für diesen Ansatz weniger geeignet als bivariate P-Splines und Gauß-Markov-Zufallsfelder. Wird jedoch eine Knotenwahlstrategie angewendet, so lässt sich die Dimension von $\boldsymbol{\gamma}$ häufig so stark reduzieren, dass alle Schätzverfahren eine ähnliche Komplexität aufweisen.

7.2.6 Lokale und adaptive Glättungsverfahren

Auch zur Analyse von Interaktionen und räumlichen Effekten werden häufig lokale Glätter eingesetzt, die auf Erweiterungen der in Abschnitt 7.1.7 betrachteten Ansätze beruhen. Für Nächste-Nachbarn-Schätzer ist diese Erweiterung besonders offensichtlich, da sich etwa die k nächsten Nachbarn auch in bivariaten Problemen mit Hilfe der euklidischen Distanz bestimmen lassen. Damit können auch der Loess-Schätzer, ebenso wie der auf Kerndichte-Schätzern beruhende Nadaraya-Watson-Schätzer, zu bivariaten Glättungsverfahren verallgemeinert werden. Die Optimierung der Glättungsparameter kann entweder über eine Gittersuche oder mit Hilfe geeigneter numerischer Verfahren

durchgeführt werden (vergleiche hierzu beispielsweise Kauermann & Opsomer (2004)). Allerdings ist die Verwendung lokaler Glättungsverfahren als Baustein in komplexeren Modellen schwierig, wenn etwa nichtparametrische Effekte und zufällige Effekte kombiniert werden sollen. Wir werden daher nicht genauer auf lokale Verfahren eingehen, sondern verweisen auf die in Abschnitt 7.4 angegebene Literatur.

Adaptive Verfahren können auch für die bivariate Glättung und komplexere Modelle effizient eingesetzt werden. Hier liegt der Vorteil insbesondere darin, dass die Verfahren in der Regel in der Lage sind, nicht nur festzustellen, wie der gemeinsame Einfluss von z_1 und z_2 zu modellieren ist, sondern auch, ob überhaupt Interaktionseffekte vorliegen und diese ebenfalls adaptiv mitzubestimmen. Dies ist insbesondere bei den Erweiterungen in Kapitel 8 von Vorteil, wenn nicht nur der Einfluss zweier, sondern einer größeren Zahl von Einflussgrößen zu untersuchen ist.

Für den MARS-Algorithmus ergibt sich die bivariate Erweiterung, indem in jeder Iteration nicht nur die Basisfunktionen einer Kovariablen, sondern sowohl die univariaten Basisfunktionen für z_1 und z_2 sowie die zugehörigen Interaktionen auf ihr Verbesserungspotential in Bezug auf das gewählte Modellwahlkriterium untersucht werden. Damit ergibt sich eine Sequenz von Modellen, die sowohl komplexe Interaktionsmodelle als auch einfache Modelle mit nur univariaten Effekten der Kovariablen beinhalten kann. Über das gewählte Modellwahlkriterium wird dann adaptiv entschieden, welches der Modelle zur Modellierung des Einflusses der Kovariablen am geeignetsten ist.

Für Regressionsbäume ergeben sich Erweiterungen für komplexere Modelle ebenfalls, indem in jeder Iteration nicht nur die Trennpunkte einer Variablen, sondern die aller Variablen berücksichtigt werden. Dementsprechend wird in jedem Schritt entschieden, nach welcher Variablen und wo genau der Datensatz als nächstes geteilt werden soll. Damit ergibt sich adaptiv eine Zerlegung der Daten in homogene Bereiche ähnlicher Zielvariablen, die einer stückweise konstanten Funktionsschätzung $f(z_1, z_2)$ entsprechen.

Auf ähnliche Art und Weise lassen sich auch die Bayesianischen adaptiven Verfahren verallgemeinern. Da wir uns in Kapitel 8 jedoch auf Penalisierungsansätze konzentrieren werden, gehen wir hier nicht genauer auf adaptive Verfahren ein. Weitere Hinweise zu adaptiven Verfahren geben wir ebenfalls am Ende dieses Kapitels in Abschnitt 7.4.

7.3 Höherdimensionale Glättung

Prinzipiell lassen sich die in Abschnitt 7.2 behandelten Ideen auch zur Modellierung höherdimensionaler Oberflächen, also zur Analyse von Modellen der Form

$$y = f(z_1, \ldots, z_q) + \varepsilon$$

einsetzen. Beispielsweise ergeben sich höherdimensionale Tensorprodukt-Splines, indem man alle möglichen Interaktionen der univariaten Splines für $z_1, \ldots, z_q$ betrachtet. Die zugehörigen Strafmatrizen lassen sich dann mit Hilfe von Kroneckerprodukten wie in Abschnitt 7.2.1 konstruieren. So erhält man beispielsweise für eine dreidimensionale Funktionsschätzung die Strafmatrix

$$\boldsymbol{I}_{d_3} \otimes \boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1 + \boldsymbol{I}_{d_3} \otimes \boldsymbol{K}_2 \otimes \boldsymbol{I}_{d_1} + \boldsymbol{K}_3 \otimes \boldsymbol{I}_{d_2} \otimes \boldsymbol{I}_{d_1},$$

wobei $\boldsymbol{K}_1$, $\boldsymbol{K}_2$ und $\boldsymbol{K}_3$ die univariaten Strafmatrizen und $\boldsymbol{I}_d$ Einheitsmatrizen geeigneter Dimension bezeichnen. Für radiale Basen und auf isotropen Korrelationsfunktionen

basierende Kriging-Terme ist die Konstruktion höherdimensionaler Oberflächen sogar noch einfacher, da hier die Basisfunktionen nur auf der euklidischen Distanz zwischen zwei Punkten beruhen, die im $\mathbb{R}^q$ ebenso wie im Spezialfall des $\mathbb{R}^2$ definiert ist. Auch die Konstruktion von Markov-Zufallsfeldern lässt sich auf höherdimensionale Probleme ausdehnen, indem man das Nachbarschaftskonzept entsprechend erweitert.

Allerdings ergeben sich unabhängig vom gewählten Ansatz Probleme bei der nonparametrischen Schätzung hochdimensionaler Funktionen, da diese in der Regel auf einer sehr großen Zahl von Parametern beruhen und die entsprechenden Schätzverfahren sehr aufwändig werden. Beispielsweise besitzt ein Tensorprodukt-Spline, der in jeder Dimension durch eine 20-dimensionale Basis beschrieben wird, im $\mathbb{R}^3$ bereits $20^3 = 8000$ Parameter. Obwohl durch die Bestrafung die effektive Parameterzahl wieder reduziert wird, muss zur Bestimmung dennoch ein 8000-dimensionales Gleichungssystem (in der Regel sogar mehrfach) gelöst werden.

Darüber hinaus ergeben sich Probleme höherdimensionaler Ansätze aus dem sogenannten „Fluch der Dimension“ (curse of dimensionality). Betrachten wir dazu die folgende Fragestellung: Gegeben sei ein q-dimensionaler Würfel mit Kantenlänge 1, in dem Daten gleichmäßig verteilt seien. Welche Kantenlänge l muss dann ein Teilwürfel haben, der einen Anteil p der Daten enthält? Für $q = 1$ ergibt sich direkt $l = p$, während für $q = 2$ bereits $l = \sqrt{p}$ gelten muss. Als allgemeine Vorschrift erhält man

$$l = p^{\frac{1}{q}},$$

also beispielsweise für $p = 0.1$

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

Bereits für eine dreidimensionale Oberfläche muss der Teilwürfel also eine Kantenlänge von fast 0.5 besitzen. Offenbar stimmt unsere intuitive Vorstellung von räumlicher Lokalität hier nicht mit den tatsächlichen Begebenheiten überein. Darüber hinaus legt der Fluch der Dimension nahe, dass in höherdimensionalen Glättungsproblemen die Anzahl von Parametern in jeder einzelnen Dimension größer gewählt werden sollte als bei univariaten Glättungsproblemen, um eine gute Überdeckung des Beobachtungsbereichs zu erhalten. Damit tritt aber noch einmal verstärkt das Problem auf, dass eine sehr große Zahl von Parametern zu schätzen ist.

Daher werden in der Praxis nur selten Oberflächen höherer Dimension als $q = 2$ betrachtet und stattdessen eine besondere, nämlich additive Struktur an die Funktion $f(z_1, \ldots, z_q)$ unterstellt, also

$$f(z_1, \ldots, z_q) = f_1(z_1) + \ldots + f_q(z_q).$$

Dies führt zur Klasse der additiven Modelle und deren Erweiterungen, die wir im folgenden Kapitel 8 behandeln werden.

7.4 Bemerkungen und Ergänzungen

Im Folgenden geben wir weiterführende Literatur an, die einzelne Aspekte aus Kapitel 7 vertieft oder allgemeiner behandelt. Ein Großteil der angegebenen Referenzen ist gleichermaßen für die in Kapitel 8 behandelten Erweiterungen von Interesse.

Univariate Glättungsverfahren wurden von Hastie & Tibshirani (1990) im Rahmen von additiven Modellen systematisch eingeführt und untersucht. Insbesondere werden hier allgemeine Fragen zu linearen Glättungsverfahren ausführlich behandelt. Einen guten Überblick, auch zu moderneren Erweiterungen, bieten Fahrmeir & Tutz (2001, Kap. 5). Der Überblicksartikel von Wand (2000) vergleicht einige univariate Glättungsverfahren per Simulation, insbesondere auch adaptive Verfahren. Wood (2006) beschreibt auf Penalisierungsansätzen und Splines beruhende nichtparametrische Regressionsmodelle und gibt darüber hinaus eine Einführung in die Verwendung des R-Pakets `mgcv`. Mit Schwerpunkt auf ihre mathematischen Eigenschaften werden Splines in Hämmerlin & Hoffmann (1994), De Boor (2001) sowie Dierckx (1993) behandelt. Detaillierte Beschreibungen zu lokalen Regressionsmodellen findet man in Härdle (1990), Fan & Gijbels (1996), Loader (1999) und Härdle, Müller, Sperlich & Werwatz (2004).

Ausführliche Darstellungen der Verbindung zwischen penalisierten Splines und gemischten Modellen, wie wir sie in Abschnitt 7.1.9 verwendet haben, findet man in Ruppert et al. (2003), Wand (2003) oder Kauermann (2006). Fahrmeir, Kneib & Lang (2004) beschreiben diesen Zusammenhang für allgemeinere Penalisierungsansätze und insbesondere auch für Glättungsverfahren der räumlichen Statistik. Erweiterungen für Überlebenszeitmodelle werden in Kauermann & Khomski (2007) oder Kneib & Fahrmeir (2007) betrachtet. Auf MCMC basierende Bayesianische Verfahren zur Glättungsparameterwahl werden in Fahrmeir & Lang (2001a), Lang & Brezger (2004) sowie Brezger & Lang (2006) genauer diskutiert.

Friedman (1991), Stone, Hansen, Kooperberg & Truong (1997) und Hansen & Kooperberg (2002) diskutieren adaptive Verfahren basierend auf frequentistischen Modellwahlstrategien. Entsprechende Bayesianische Ansätze beschreiben die Artikel von Biller & Fahrmeir (2001), Biller (2000), Yau, Kohn & Wood (2003), Denison et al. (1998) oder DiMatteo, Genovese & Kass (2001). Eine einführende Darstellung zu Regressionsbäumen bietet Tutz (2000, Kap. 9), die klassische Referenz im Bereich Regressions- und Klassifikationsbäume ist Breiman, Friedman, Stone & Olshen (1984). Wie bereits in Abschnitt 7.1.10 angesprochen, sind Regressionsbäume instabil, da bereits kleine Änderungen an der Datenstruktur in deutlich veränderten Baumstrukturen resultieren können. Resampling-Verfahren, die nicht nur einen, sondern ein ganzes Ensemble von Regressionsbäumen erzeugen, erlauben es, diese Instabilität zu beseitigen. Beispiele hierfür sind Bagging (Breiman 1996) oder Random Forests (Breiman 2001). Tatsächlich lassen sich ähnliche Resampling-Verfahren auch in Regressions- und Glättungsverfahren einsetzen, vergleiche etwa Bühlmann & Yu (2003), Bühlmann & Yu (2006) oder Tutz & Binder (2006).

Abschließend sei noch auf einige Ansätze verwiesen, die wir in diesem Kapitel nicht näher behandeln konnten. Wavelets sind ein aus der Fourier-Analyse motiviertes Glättungsverfahren, das insbesondere zur Darstellung von funktionalen Zusammenhängen mit stark ausgeprägter (lokaler) Variabilität angewendet werden kann. Prinzipiell lassen sich Wavelets ebenfalls als Basisfunktionenansatz betrachten. Die Berechnung erfolgt jedoch über effiziente Algorithmen, die spezielle Eigenschaften der Wavelet-Basisfunktionen ausnut-

zen und deutlich von den hier behandelten Verfahren abweichen. Daher sind Wavelets insbesondere aus numerischen Gesichtspunkten interessant, haben aber den Nachteil, dass zu ihrem Verständnis mathematisch anspruchsvollere Werkzeuge notwendig sind. Eine knappe Einführung zu Wavelets findet man in Hastie, Tibshirani & Friedman (2003). Ausführlichere, aber auch anspruchsvollere Beschreibungen findet man etwa in Ogden (1997) oder Gençay, Selçuk & Whitcher (2002).

Eine andere Klasse nichtparametrischer Regressionsmodelle wird in der funktionalen Datenanalyse betrachtet. Hier werden sowohl die Einflussgröße als auch die Zielvariable als Funktionen aufgefasst und speziell auf diese Situation angepasste Verfahren angewendet. Eine ausführliche Diskussion hierzu findet man in Ramsay & Silverman (2005). Ramsay & Silverman (2002) beschreiben die funktionale Datenanalyse im Rahmen einiger Fallstudien.

Zu den in Abschnitt 7.2 betrachteten Verfahren der räumlichen Statistik existieren mehrere empfehlenswerte Einführungsbücher, beispielsweise Banerjee et al. (2003) oder Schabenberger & Gotway (2005). Mit einzelnen Aspekten beschäftigen sich Rue & Held (2005) (Gauß-Markov-Zufallsfelder) sowie Chiles & Delfiner (1999) und Stein (1999) (Kriging). Möglichkeiten, Anisotropie in geostatistischen Modellen zu berücksichtigen, werden in Ecker & Gelfand (2003) und Zimmermann (1993) diskutiert. Nychka (2000) gibt einen mathematisch rigorosen Zugang zum Zusammenhang zwischen Kriging und Glättungsverfahren basierend auf radialen Basisfunktionen. Eilers & Marx (2003) und Dierckx (1993) schließlich behandeln bivariate Polynom-Splines.

8 Strukturiert-additive Regression

Im vorhergehenden Kapitel wurde gezeigt, wie sich im Rahmen der nichtparametrischen Regression der Einfluss einer metrischen Kovariablen z auf die Zielvariable y ohne Vorgabe einer restriktiven funktionalen Form des Effekts $f(z)$ flexibel modellieren und schätzen lässt, und wie man die Konzepte für zwei metrische Kovariablen (z_1, z_2) oder eine geografisch-räumliche Lokationsvariable s erweitern kann. Wie wir in Kapitel 3 und vielen Beispielen, etwa zum Mietspiegel, zur Unterernährung in Afrika, zum Kredit-Scoring und zur Patenterteilung gesehen haben, liegt in Anwendungen häufig folgende Situation vor: Für einen Teil der Kovariablen lässt sich der Einfluss auf die Zielvariable adäquat durch einen linearen Prädiktor beschreiben. Zusätzlich gibt es jedoch eine oder mehrere metrische Kovariablen $z_1, \ldots, z_q$, deren Effekte sich nicht – zumindest a priori – durch eine einfache funktionale Form beschreiben lassen. Prinzipiell ist man also an der flexiblen Modellierung des Einflusses dieser Kovariablen in Form einer Funktion $f(z_1, \ldots, z_q)$ interessiert. Wie wir jedoch in Abschnitt 7.3 gesehen haben, ist die Schätzung solcher hochdimensionalen Funktionen schwierig und erfordert (um dem Fluch der Dimension entgegenzuwirken) einen sehr großen Stichprobenumfang. Daher unterstellt man in der Regel eine stärker eingeschränkte, additive Struktur für den Einfluss der Kovariablen, also beispielsweise

$$f(z_1, \ldots, z_q) = f_1(z_1) + \ldots + f_q(z_q).$$

Darüber hinaus sind auch Modelle mit Interaktionseffekten von Interesse, oder geoadditive Modelle, die räumliche Information in Form einer Lokationsvariablen berücksichtigen. Die Abschnitte 8.1 bis 8.3 behandeln solche Modellklassen für Regressionsprobleme, bei denen die Zielvariable metrisch und – wie im linearen Modell – auch approximativ normalverteilt ist. Die Modellklassen lassen sich in einem einheitlichen Ansatz als strukturiert-additive Regressionsmodelle formulieren und für nicht normalverteilte, insbesondere diskrete Zielvariablen verallgemeinern. Dabei konzentrieren wir uns auf die in den Abschnitten 7.1 und 7.2 diskutierten Penalisierungsansätze, da diese die Schätzung aller Modellparameter mit Hilfe einheitlicher Schätzverfahren erlauben.

8.1 Additive Modelle

Wir gehen von folgender Datensituation aus: Wie in der linearen Regression liegen Beobachtungen $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, zu einer metrischen Zielgröße y und zu Kovariablen $x_1, \ldots, x_k$ vor, deren Einfluss auf y durch einen linearen Prädiktor modelliert werden kann. Zusätzlich liegen Beobachtungen $(z_{i1}, \ldots, z_{iq})$, $i = 1, \ldots, n$, zu weiteren *metrischen* Kovariablen $z_1, \ldots, z_q$ vor, deren Einfluss nichtparametrisch modelliert und analysiert werden soll. Im Mietspiegelbeispiel ist y die Nettomiete pro Quadratmeter, z_1 ist die Wohnfläche, z_2 das Baujahr, und die Kovariablen $x_1, \ldots, x_k$ umfassen binär kodierte Ausstattungsmerkmale sowie die Beurteilung der Wohnlage durch Experten in den drei Kategorien normale, gute und beste Lage.

Additive Modelle erweitern das lineare Modell zu

$$\begin{aligned} y_i &= f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \\ &= f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin} + \varepsilon_i \\ &= \eta_i^{add} + \varepsilon_i \end{aligned} \tag{8.1}$$

mit

$$\eta_i^{lin} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \qquad \eta_i^{add} = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin}.$$

Die Funktionen $f_1(z_1), \ldots, f_q(z_q)$ sind nichtlineare, glatte Effekte der Kovariablen $z_1, \ldots, z_q$. Sie werden nichtparametrisch modelliert und geschätzt (siehe Abschnitt 8.5), wobei prinzipiell alle Streudiagramm-Glätter aus Kapitel 7.1 als Bausteine eingesetzt werden können. Wir beschränken uns hier allerdings auf Penalisierungsansätze.

Bei additiven Modellen tritt ein *Identifikationsproblem* auf, das bereits in Kapitel 2.6 erwähnt wurde: Addiert man etwa zur Funktion $f_1(z_1)$ eine Konstante $c \neq 0$ und subtrahiert c gleichzeitig von einer zweiten Funktion $f_2(z_2)$, so bleibt die Summe

$$f_1(z_1) + f_2(z_2) = f_1(z_1) + c + f_2(z_2) - c$$

unverändert, d.h. durch den Übergang von $f_1(z_1)$ zu $\tilde{f}_1(z_1) = f_1(z_1) + c$ und von $f_2(z_2)$ zu $\tilde{f}_2(z_2) = f_2(z_2) - c$ ändert sich der Wert des Prädiktors nicht. Zwar ist die Form von $f_1(z_1)$ und $\tilde{f}_1(z_1)$ bzw. $f_2(z_2)$ und $\tilde{f}_2(z_2)$ die gleiche, aber durch die willkürliche Veränderung der Höhe um c bzw. $-c$ ist diese ohne weitere Restriktionen nicht identifizierbar. Man muss deshalb die Höhe oder das Niveau der Funktionen festlegen. Eine häufig eingesetzte Möglichkeit besteht darin, die Funktionen „um Null zu zentrieren“, so dass

$$\sum_{i=1}^{n} f_1(z_{i1}) = \ldots = \sum_{i=1}^{n} f_q(z_{iq}) = 0$$

gilt. Wir verwenden im Weiteren diese Zentrierung zur Behebung des Identifikationsproblems.

Für die Fehlervariablen ε_i werden im Standardmodell der additiven Regression die gleichen Annahmen wie im klassischen linearen Modell getroffen, d.h. die ε_i sind i.i.d. mit $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ und gegebenenfalls normalverteilt. Allgemeinere Annahmen sind in Analogie zu linearen Modellen möglich, werden aber in diesem Kapitel nicht betrachtet.

Wie im linearen Modell übertragen sich die Annahmen über die Fehlervariablen ε_i auf die Zielgrößen y_i, d.h. die y_i sind für gegebene Kovariablenwerte (bedingt) unabhängig mit

$$\begin{aligned} \mathrm{E}(y_i) &= \mu_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \\ \mathrm{Var}(y_i) &= \sigma^2 \end{aligned}$$

und gegebenenfalls normalverteilt mit

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2).$$

In der Literatur wird häufig der Spezialfall

Standardmodell der additiven Regression

Daten

$(y_i, z_{i1}, \ldots, z_{iq}, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, zu y und $x_1, \ldots, x_k$ wie in der linearen Regression und zu stetigen Kovariablen $z_1, \ldots, z_q$.

Modell

Es gilt

$$y_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i.$$

Die Funktionen $f_1(z_1)$, ..., $f_q(z_q)$ werden als glatt vorausgesetzt und beschreiben nichtlineare Effekte der metrischen Kovariablen $z_1, \ldots, z_q$, die additiv zusammenwirken. Für die Fehlervariablen ε_i werden die gleichen Annahmen getroffen wie im Standardmodell der multiplen linearen Regression.

$$y_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \varepsilon_i, \tag{8.2}$$

also ohne den zusätzlichen linearen Prädiktor η_i^{lin}, als additives Modell bezeichnet. Das Modell (8.1) heißt dann *partiell lineares Modell* oder auch *semiparametrisches Modell*.

In jedem Fall enthalten additive Modelle (8.1) oder (8.2) keine Terme für Interaktionen zwischen zwei stetigen Kovariablen, etwa z_1, z_2 oder auch z_1, x mit einer stetigen Kovariable x, und auch keine Interaktion zwischen einer stetigen Kovariable z und einer kategorialen oder binären Kovariable x. Modelle mit solchen Interaktionen betrachten wir in Abschnitt 8.3.

Die Interpretation der nichtlinearen und linearen Effekte erfolgt wie bereits in Kapitel 2.6 beschrieben. Wir illustrieren dies erneut anhand der Mietspiegeldaten, vergleiche Beispiel 8.1 etwas weiter unten.

Wir werden in diesem Kapitel zur nichtparametrischen Modellierung und Schätzung der Funktionen vorzugsweise die in Abschnitt 7.1.3 allgemein beschriebenen Penalisierungsansätze für Basisfunktionen einsetzen. Dann wird jede Funktion f_j, $j = 1, \ldots, q$, durch

$$f_j(z_j) = \sum_{l=1}^{d_j} \gamma_{jl} B_l(z_j),$$

approximiert, wobei die Basisfunktionen B_l Polynomsplines in TP- oder B-Spline-Basis, Glättungssplines oder auch Kriging-Ansätze repräsentieren. Prinzipiell können für die verschiedenen Funktionen f_j auch verschiedene Typen von Basisfunktionen gewählt werden, etwa Polynomsplines von verschiedenem Grad. Wir werden dies im Folgenden jedoch notationell unterdrücken. Der Vektor $\boldsymbol{f}_j = (f_j(z_{1j}), \ldots, f_j(z_{nj}))'$ der Funktionswerte an den beobachteten Ausprägungen $z_{1j}, \ldots, z_{nj}$ der Kovariable lässt sich dann immer durch

$$\boldsymbol{f}_j = \boldsymbol{Z}_j \boldsymbol{\gamma}_j$$

Additive Regression in Basisfunktionenansätzen

Die Funktionen f_j werden durch Basisfunktionen in der Form

$$f_j(z_j) = \sum_{l=1}^{d_j} \gamma_{jl} B_l(z_j)$$

approximiert. Unabhängig von der gewählten Basis ergibt sich das additive Modell

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

wobei die Designmatrizen $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q$ aus den an den beobachteten Kovariablenausprägungen ausgewerteten Basisfunktionen bestehen und die Designmatrix $\boldsymbol{X}$ wie im linearen Modell gebildet wird.

ausdrücken. Dabei ist $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jd_j})'$ der Koeffizientenvektor, und die Designmatrix $\boldsymbol{Z}_j$ besteht aus den an den Kovariablenausprägungen ausgewerteten Basisfunktionen, d.h. es gilt $\boldsymbol{Z}_j[i,l] = B_l(z_{ij})$.

Bezeichnet noch wie in linearen Modellen $\boldsymbol{y} = (y_1, \ldots, y_n)'$ den Vektor der Zielgrößen, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ den Vektor der Fehlergrößen, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ den Koeffizientenvektor des linearen Teils des Prädiktors und $\boldsymbol{X}$ die entsprechende Designmatrix, dann lässt sich das additive Modell in Matrixnotation in der Form

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8.3}$$

schreiben. Auf den ersten Blick sieht (8.3) aus wie ein großes lineares Modell. Falls die Koeffizientenvektoren $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ der Basisfunktionen unrestringiert nach der KQ-Methode geschätzt würden, ist dies auch tatsächlich der Fall. In der Regel wird jedoch die Schätzung durch die gleichen Straffunktionen $\lambda_j \boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j$ regularisiert, wie dies in 7.1.3 für eine Funktion f beschrieben ist. Die Strafmatrizen haben je nach Wahl des Basisfunktionenansatzes unterschiedliche Gestalt. Die genauere Durchführung der Schätzung wird in Abschnitt 8.5 beschrieben.

Wir verwenden wie im folgenden Beispiel meist penalisierte Splines, vergleiche Abschnitt 7.1.2 oder die entsprechende Bayesianische Version, bei der die Straffunktion durch die Gauß-Priori $p(\boldsymbol{\gamma}_j \mid \tau_j^2) \propto \exp(-0.5\boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j / \tau_j^2)$ ersetzt wird.

Beispiel 8.1 Mietspiegel – Additives Modell

Als Zielvariable y verwenden wir die Nettomiete pro Quadratmeter. Als Kovariablen nehmen wir die Wohnfläche (*flaeche*), das Baujahr (*bjahr*) und die Wohnlage (*lage*) auf. Dies führt zu dem additiven Modell

$$mieteqm = f_1(flaeche) + f_2(bjahr) + \beta_0 + \beta_1 glage + \beta_2 blage + \varepsilon.$$

Die Funktionen f_1 und f_2 werden durch P-Splines mit 20 inneren Knoten und auf der Summe der quadrierten zweiten Differenzen basierenden Straftermen modelliert. Für die Wohnlage werden die Dummy-Variablen *glage* für gute Wohnlagen und *blage* für beste Wohnlagen ins Modell aufgenommen. Als Referenzkategorie fungiert die normale

Abb. 8.1. *Mietspiegel München: Geschätzte nichtlineare Effekte der Wohnfläche und des Baujahrs.*

Wohnlage. Man beachte, dass es sich bis jetzt um ein reines Haupteffekt-Modell handelt, d.h. es werden mögliche Interaktionen, etwa zwischen Wohnfläche und Lage außer Acht gelassen. Modelle mit solchen Interaktionen werden in Abschnitt 8.3 vorgestellt.

Abbildungen 8.1 zeigt die geschätzten nichtlinearen Effekte von Wohnfläche und Baujahr. In die jeweils linken Grafiken sind zusätzlich partielle Residuen eingezeichnet. Diese werden völlig analog wie in Kapitel 3.2.3 im Rahmen der linearen Modelle beschrieben berechnet. Die partiellen Residuen bezüglich der Wohnlage sind also beispielsweise definiert durch

$$\hat{\varepsilon}_{flaeche,i} = \hat{\varepsilon}_i + \hat{f}_1(flaeche_i).$$

Zu partiellen Residuen in additiven Modellen vergleiche man auch noch Abschnitt 8.5.1.

Der geschätzte Effekt der Wohnfläche zeigt – wie zu erwarten – einen mit wachsender Wohnfläche nichtlinear fallenden Effekt. Ab ca. 110 qm weiten sich die punktweisen Konfidenzintervalle, da für größere Wohnflächen weniger Daten vorliegen. Deshalb sollte auch der leicht ansteigende Verlauf am rechten Rand nicht überinterpretiert werden: Ein ab ca. 110 qm horizontaler Verlauf würde problemlos in die Konfidenzbänder passen! Der geschätzte Baujahrseffekt ist zunächst annähernd konstant und ab circa 1945 linear ansteigend. Insgesamt sind die geschätzten nichtlinearen Effekte relativ ähnlich zu dem in Kapitel 3 rein parametrisch geschätzten Modell aus Beispiel 3.5 (Seite 76). Trotzdem ist die Anpassung an die Daten mit der hier verwendeten Modellierung noch besser. Verwendet man wie in Beispiel 3.5 die Transformation $1/flaeche$ für die Wohnfläche und ein kubisches Polynom zur Modellierung des Effekts des Baujahrs, so erhält man als AIC den Wert 11496.8. Mit der hier benutzten nonparametrischen Modellierung

Variable	Koeffizient	Standard-abweichung	p-Wert	95% Konfidenz-intervall	
Konstante	13.811	0.233	< 0.001	13.354	14.269
glage	1.235	0.149	< 0.001	0.944	1.526
blage	2.931	0.455	< 0.001	2.040	3.823

Tabelle 8.1. *Mietspiegel München: Geschätzte Effekte der Wohnlage.*

erhalten wir stattdessen den deutlich niedrigeren AIC-Wert 11479.1. Zur Berechnung des AIC und anderer Modellgütemaße in additiven Modellen vergleiche man Abschnitt 8.5.1. Ein weiterer Vorteil im Vergleich zur parametrischen Modellierung in Beispiel 3.5 ist, dass die Schätzungen hier automatisiert erzielt wurden, d.h. es war nicht nötig sich Gedanken zu machen über Variablentransformationen oder den Grad des Polynoms bei polynomialer Modellierung.

Die Zuschläge für gute und beste Lage sind mit ca. 1.23 DM bzw. 2.93 DM deutlich und hochsignifikant (Tabelle 8.1).

△

8.2 Geoadditive Regression

Die Daten für den Mietspiegel enthalten neben der drei-kategorialen Experteneinschätzung der Wohnlage zusätzlich die Information in welchem der 336 Bezirksviertel in München die betreffende Wohnung liegt. Will man diese kleinräumige, feinere Information der Lokationsvariable *bezv* nutzen und gleichzeitig auf die Experteneinschätzung verzichten, so führt dies zum geoadditiven Modell

$$mieteqm = \beta_0 + f_1(flaeche) + f_2(bjahr) + f_{geo}(bezv) + \varepsilon.$$

Man kann aber durchaus zusätzlich die Experteneinschätzung mit in das Modell einbeziehen. Dann erhält man das geoadditive Modell

$$mieteqm = f_1(flaeche) + f_2(bjahr) + f_{geo}(bezv) + \beta_0 + \beta_1\, glage + \beta_2\, blage + \varepsilon.$$

Im Allgemeinen sind die Daten einer geoadditiven Regression durch Beobachtungen $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, $i = 1, \ldots, n$ für die Zielgröße y sowie Kovariablen $\boldsymbol{x}, \boldsymbol{z}$ wie im additiven Modell und durch die zusätzlichen Werte s_i, $i = 1, \ldots, n$, einer räumlich-geografischen Lokationsvariable s gegeben. Wie bereits in Kapitel 7.2 beschrieben, kann die Variable s diskret sein und Werte $s \in \{1, \ldots, d\}$ annehmen (vergleiche Abschnitt 7.2.4), oder sie kann stetig in einem Teilbereich des $\mathbb{R}^2$ variieren (vergleiche die Abschnitte 7.2.1–7.2.3). Im diskreten Fall bezeichnet s_i meist eine spezielle Region, in der das Individuum oder die Einheit i beobachtet wurde, z.B. Bezirksviertel in München, Distrikte in Sambia, Landkreise oder Postleitzahlbezirke in Deutschland, usw. Im stetigen Fall ist s meist durch die Koordinaten des $\mathbb{R}^2$ gegeben, d.h. die Lokation s_i der Einheit i ist (fast) exakt bekannt. So ist in der Original-Stichprobe des Münchner Mietspiegels die exakte Adresse der Wohnung gegeben. Aus Datenschutzgründen ist uns aber nur das Bezirksviertel, in der die Wohnung liegt, bekannt. Ähnlich verhält es sich mit Versicherungsdaten von individuellen Versicherungsnehmern. Auch hier sind der Versicherung die exakten Adressen

bekannt, werden aber aus Datenschutzgründen zu einer diskreten Lokationsvariable z.B. für Landkreise oder Zulassungsbezirke aggregiert.

In geoadditiven Modellen wird der Prädiktor additiver Modelle durch einen räumlichen Effekt $f_{geo}(s)$ der Lokationsvariable s ergänzt, d.h. es gilt

$$\begin{aligned} y_i &= \eta_i^{add} + f_{geo}(s_i) + \varepsilon_i \\ &= f_1(z_{i1}) + \ldots + f_q(z_{iq}) + f_{geo}(s_i) + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \end{aligned} \tag{8.4}$$

$i = 1, \ldots, n$, mit $\boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$. Für die Kovariablen $\boldsymbol{x}_i, \boldsymbol{z}_i$ und die Fehlervariablen ε_i gelten die gleichen Annahmen wie im additiven Modell.

Der räumliche Effekt $f_{geo}(s)$ kann als Surrogat für unbeobachtete, nicht durch die Daten erfasste Variablen mit räumlicher Information aufgefasst werden. Als Bausteine zur Modellierung und Schätzung von $f_{geo}(s)$ dienen die in Kapitel 7.2 vorgestellten Ansätze. Für diskretes $s \in \{1, \ldots, d\}$ sind Markov-Zufallsfelder (siehe Abschnitt 7.2.4) besonders geeignet. Man kann aber auch die zunächst für stetiges s konzipierten räumlichen Glätter, insbesondere Kriging (siehe Abschnitt 7.2.3), einsetzen: Dazu verwendet man als Koordinaten z.B. die Schwerpunkte (Zentroide) der Regionen.

Für diskretes s, wenn also zum Beispiel $s_i = s$ die Zugehörigkeit von i zur Region s aus der Menge $\{1, \ldots, d\}$ von insgesamt d Regionen bezeichnet, ist $f_{geo}(s_i)$ der räumliche Effekt der Region $s = s_i$. Der Vektor $\boldsymbol{f}_{geo} = (f_{geo}(s_1), \ldots, f_{geo}(s_n))'$ des räumlichen Effekts für die beobachteten Einheiten $i = 1, \ldots, n$ lässt sich wie bei den Basisfunktionenansätzen des vorangehenden Abschnitts durch

$$\boldsymbol{f}_{geo} = \boldsymbol{Z}_{geo}\boldsymbol{\gamma}_{geo} \tag{8.5}$$

darstellen. Dabei ist $\boldsymbol{\gamma}_{geo} = (\gamma_{geo,1}, \ldots, \gamma_{geo,d})'$ der Vektor der räumlichen Effekte, und die $(n \times d)$-Designmatrix $\boldsymbol{Z}_{geo}$ ist eine sogenannte Inzidenzmatrix, mit $\boldsymbol{Z}_{geo}[i, s] = 1$ für das s-te Element der i-ten Zeile wenn $s_i = s$ gilt und 0 sonst. Für stetiges s kann $\boldsymbol{f}_{geo}$ ebenfalls in der Form (8.5) geschrieben werden. Die speziellen Formen von $\boldsymbol{Z}_{geo}$ und $\boldsymbol{\gamma}_{geo}$ finden sich in den Abschnitten 7.2.1–7.2.3. Bei Kriging-Ansätzen wird dabei meist mit einer repräsentativen Knotenmenge gearbeitet (Seite 384).

Damit können geoadditive Modelle in Matrixnotation in der Form

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{Z}_{geo}\boldsymbol{\gamma}_{geo} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8.6}$$

geschrieben werden, d.h. das additive Modell 8.3 wird um den räumlichen Effekt $\boldsymbol{Z}_{geo}\boldsymbol{\gamma}_{geo}$ erweitert. Wie bei der Schätzung der Koeffizientenvektoren $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ wird auch die Schätzung von $\boldsymbol{\gamma}_{geo}$ durch Straffunktionen der Form $\lambda_{geo}\boldsymbol{\gamma}_{geo}'\boldsymbol{K}_{geo}\boldsymbol{\gamma}_{geo}$ bzw. durch entsprechende Gauß-Prioris wie in Abschnitt 7.2 beschrieben regularisiert, vergleiche Abschnitt 8.5.

Beispiel 8.2 Mietspiegel

Wir betrachten zwei Modelle, bei denen der Effekt $f_{geo}(bezv)$ der Bezirksviertel von München berücksichtigt wird:

$$mieteqm = f_1(flaeche) + f_2(bjahr) + f_{geo}(bezv) + \beta_0 + \varepsilon$$

und

$$mieteqm = f_1(flaeche) + f_2(bjahr) + f_{geo}(bezv) + \beta_0 + \beta_1\, glage + \beta_2\, blage + \varepsilon.$$

Abb. 8.2. *Mietspiegel München: Räumlicher Effekt basierend auf Kriging. Links: Ohne Verwendung der Expertenbeurteilung der Wohnlage. Rechts: Mit Verwendung der Expertenbeurteilung der Wohnlage.*

Abb. 8.3. *Mietspiegel München: Kerndichteschätzer für den räumlichen Effekt. Gestrichelte Linie: Modell ohne Expertenurteil, durchgezogene Linie: Modell mit Expertenurteil.*

Im ersten Modell wird die Experteneinschätzung durch die Wohnlagevariable nicht berücksichtigt, während sie im zweiten Modell zusätzlich in den Prädiktor aufgenommen wird. Der räumliche Effekt $f_{geo}(bezv)$ wird in beiden Modellen durch einen Kriging-Ansatz mit den Zentroiden der Bezirksviertel als fiktive Messpunkte modelliert. Durch den Vergleich der Schätzungen $\hat{f}_{geo}$ für die beiden Modelle können wir die Güte der Experteneinschätzung überprüfen. Ist die Expertenbeurteilung der Wohnlage brauchbar, so müssten im zweiten Modell die Lage-Dummies *glage* und *blage* einen Großteil der räumlichen Heterogenität auffangen und der geschätzte räumliche Effekt $\hat{f}_{geo}$ sollte deutlich schwächer sein als im ersten Modell.

Abbildung 8.2 zeigt die geschätzten räumlichen Effekte. Die linke Abbildung spiegelt den kleinräumigen Einfluss der Lage in München, adjustiert auf die Effekte von Wohnfläche und Baujahr, sehr gut wieder. Teile im Süden, wie etwa Grünwald oder Menterschwaige, oder in der Nähe des Nymphenburger Parks und am Englischen Garten liegen bei den Miethöhen deutlich über dem Durchschnitt, während Gebiete im Norden, in der Nähe und nördlich des Frankfurter Rings deutlich unter dem Durchschnitt liegen.

Werden zusätzlich die Effekte der Wohnlagevariablen (gemäß Experteneinschätzung) in den Prädiktor aufgenommen, schwächt sich dieses Muster deutlich ab. Um einen Eindruck von der geringeren Variabilität zu gewinnen, zeigt Abbildung 8.3 Kerndichteschätzer der geschätzten räumlichen Effekte. Mit Berücksichtigung der Experten-

Geoadditive Modelle

Daten

Neben der stetigen Zielvariablen y, den stetigen Kovariablen $z_1, \ldots, z_q$ und den üblichen Kovariablen $x_1, \ldots, x_k$ liegen noch Beobachtungen zur geografischen Lokation s vor. Dabei kann $s \in \{1, \ldots, d\}$ diskret sein und z.B. Regionen bezeichnen, oder s ist stetig und bezeichnet die Koordinaten in der Ebene.

Modell

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + f_{geo}(s_i) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

wobei für die Fehlervariablen ε_i die gleichen Annahmen wie für das klassische lineare Regressionsmodell getroffen werden. Die Modellierung von f_{geo} basiert für diskretes s auf Markov-Zufallsfeldern, für stetiges s werden Kriging oder zweidimensionale Basisfunktionenansätze verwendet. In Matrixnotation ergibt sich das geoadditive Modell

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{Z}_{geo}\boldsymbol{\gamma}_{geo} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

einschätzung ist die Verteilung deutlich stärker um Null konzentriert, weist jedoch noch eine sichtbare Restvariabilität auf. Somit erfasst die Experteneinschätzung tatsächlich einen großen Teil der räumlichen Heterogenität. Trotzdem scheint noch Raum für eine Verbesserung der Expertenbeurteilung vorhanden.

Die Güte der Expertenbeurteilung der Wohnlage kann auch an der Modellgüte gemessen durch das AIC abgelesen werden. Wir erhalten AIC = 11476.2 im ersten Modell und AIC = 11450 im zweiten Modell. Der Wert des AIC im ersten Modell ist nur geringfügig um 2.9 Einheiten kleiner als im Modell von Beispiel 8.1, in dem der Effekt der Wohnlage rein parametrisch durch die beiden Lage-Dummies modelliert ist. Nur durch die Kombination von Expertenurteil und datengesteuerter nichtparametrischen Modellierung der räumlichen Heterogenität verringert sich der Wert des AIC deutlich um fast 30 Einheiten.

△

8.3 Modelle mit Interaktionen

In Kapitel 3.1.4 (Seite 83) wurde bereits dargestellt, wie zusätzlich zu den Haupteffekten von Kovariablen auch (paarweise) Interaktionen zwischen Kovariablen in linearen Modellen berücksichtigt werden können. Die in den vorangehenden Abschnitten betrachteten additiven und geoadditiven Modelle sind bezüglich der nichtparametrischen Komponenten $f_1, \ldots, f_q$ und der räumlichen Komponente f_{geo} reine Haupteffekt-Modelle. In diesem Abschnitt zeigen wir, wie sie durch Einbeziehen von (paarweisen) nichtparametrischen Interaktionstermen erweitert werden können, um Wechselwirkungen zwischen Kovariablen zu berücksichtigen. Wir betrachten zunächst den Fall einer metrischen Kovariable

z und einer binären oder kategorialen Kovariablen x. Wegen der besonderen Interpretation spricht man auch von Modellen mit variierenden Koeffizienten. Modelle dieser Art lassen sich auch dahingehend modifizieren, dass Interaktionen zwischen einer Lokationsvariablen s und x erlaubt werden. Anschließend gehen wir auf paarweise Interaktionen zwischen zwei metrischen Kovariablen z_1 und z_2 ein. Zur Illustration dient wieder das Mietspiegelbeispiel.

8.3.1 Modelle mit variierenden Koeffizienten

Eine Interaktion zwischen einer metrischen Variablen, etwa z_1, und einer binären Variablen, etwa x_1, kann in additiven Modellen durch zusätzliches Einbeziehen eines Interaktionsterms $f_{z_1|x_1}(z_1) \cdot x_1$ berücksichtigt werden. Dabei ist $f_{z_1|x_1}(z_1)$ eine nichtparametrische, aus den Daten wie der Haupteffekt $f_1(z_1)$ zu schätzende glatte Funktion. Dies ergibt das Modell

$$y_1 = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + f_{z_1|x_1}(z_{i1})\, x_{i1} + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{8.7}$$

wobei ansonsten die gleichen Annahmen wie für additive Modelle (8.1) gelten. Da das Modell (8.7) auch die Haupteffekte $f_1(z_1)$ und $\beta_1 x_1$ enthält, muss die Funktion $f_{z_1|x_1}$ wie die anderen Funktionen $f_1, \ldots, f_q$ um Null zentriert werden. Ansonsten würde der Übergang zu $\tilde{f}_{z_1|x_1} = f_{z_1|x_1}(z_1) + c$ und zu $\tilde{\beta}_1 = \beta_1 - c$ den gleichen Prädiktor ergeben, d.h. das Modell wäre nicht identifizierbar. Falls keine Zentrierung um Null vorgenommen wird, muss der Haupteffekt $\beta_1 x_1$ aus dem Modell entfernt werden.

Das Modell lässt sich folgendermaßen interpretieren:

- $f_1(z_1)$ ist der nichtlineare Effekt von z_1 falls $x_1 = 0$;
- $f_1(z_1) + f_{z_1|x_1}(z_1) + \beta_1$ ist der nichtlineare Effekt von z_1 falls $x_1 = 1$;
- $f_{z_1|x_1}(z_1) + \beta_1$ ist ein in Abhängigkeit von z_1 *variierender Effekt* für $x_1 = 1$ (im Vergleich zu $x_1 = 0$). Die Variable z_1 heißt auch *Effektmodifizierer* von x_1 und x_1 *Interaktionsvariable*.

Statt des üblichen festen Koeffizienten β_1 von x_1 erhält man also einen mit z_1 variierenden Koeffizienten $f_{z_1|x_1}(z_1)+\beta_1$. Zur Veranschaulichung betrachtet man am besten die beiden Effekte $f_1(z_1)$ und $f_1(z_1) + f_{z_1|x_1}(z_1) + \beta_1$ in einer gemeinsamen Grafik. Zusätzlich kann der variierende Effekt $f_{z_1|x_1}(z_1) + \beta_1$ in Abhängigkeit von z_1 in einer separaten Grafik dargestellt werden.

In analoger Weise kann man eine Interaktion zwischen z_1 und einer drei-kategorialen Variable $x \in \{1, 2, 3\}$ einführen, wobei x durch die Dummy-Variablen x_1 und x_2 mit 3 als Referenzkategorie kodiert ist. Dies führt zum Modell

$$y_i = f_1(z_{i1}) + \ldots + f_{z_1|x_1}(z_{i1})\, x_{i1} + f_{z_1|x_2}(z_{i1}) x_{i2} + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \varepsilon_i.$$

Interpretation und Visualisierung erfolgen analog: $f_1(z_1) + f_{z_1|x_1}(z_1) + \beta_1$ ist der Effekt von z_1 falls $x = 1$, und $f_1(z_1)+f_{z_1|x_2}(z_1)+\beta_2$ der Effekt von z_1 falls $x = 2$, während $f_1(z)$ der nichtlineare Effekt von z ist, falls die Referenzkategorie 3 vorliegt. Zur Visualisierung zeichnet man wieder die (addierten) Effekte in Grafiken ein.

Abb. 8.4. *Links: Effekte der Wohnfläche in normaler (durchgezogene Linie) und bester Lage (gestrichelte Linie). Die Punkte im unteren Teil der Grafik zeigen die Verteilung der Beobachtungen in normaler (dunkle Punkte) und bester Lage (hellere Punkte). Rechts: Variierender Effekt der besten Wohnlage.*

Abb. 8.5. *Geschätzte Funktionen f_1 und f_2 inklusive punktweise Konfidenzintervalle.*

Beispiel 8.3 Mietspiegel – Interaktion zwischen Wohnfläche und Lage

Dieses Beispiel kann als nichtparametrisches Analogon zur rein parametrischen Version in Beispiel 3.8 (Seite 88) angesehen werden. Wie in Beispiel 3.8 verwenden wir zur Illustration nur Daten für Wohnungen in normaler (Referenzkategorie) und bester Lage und schätzen das Modell

$$mieteqm = f_1(flaeche) + f_2(flaeche)\, blage + f_3(bjahr) + \beta_0 + \beta_1\, blage + \varepsilon.$$

Die Funktion $f_1(flaeche)$ kann als nichtlinearer Effekt der Wohnfläche für normale Lage ($blage = 0$) und der Term $f_1(flaeche) + f_2(flaeche) + \beta_1$ als nichtlinearer Effekt der Wohnfläche für beste Lage ($blage = 1$) interpretiert werden. Bei dem Term $f_2(flaeche) + \beta_1$ handelt es sich um den mit der Wohnfläche variierenden Effekt der besten Lagen. Die Abbildungen 8.4 und 8.5 zeigen die geschätzten Effekte bzw. die Funktionen f_1, f_2.

Die geschätzten Effekte der Wohnfläche für normale und beste Lagen in der linken Grafik von Abbildung 8.4 zeigen visuell das Ausmaß der Interaktion. Sind die Kurven nur parallel verschoben, so liegt keine Interaktion vor. Die Funktion $f_2(flaeche)$ (Abbildung 8.5 rechts) müsste dann (annähernd) Null sein. Im vorliegenden Fall erkennen wir jedoch eine verhältnismäßig stark ausgeprägte Interaktion. Für kleine Wohnflächen ist

der Effekt in normalen und besten Lagen fast gleich. Je größer die Wohnfläche, desto weiter driften die Effekte auseinander, d.h. desto größer ist der Unterschied in der Miete pro qm zwischen Wohnungen in normaler und bester Lage.

Die rechte Grafik der Abbildung 8.4 zeigt, dass der Effekt der besten Lage zunächst klein ist, dann stark zunimmt, und ab ca. 70 qm Wohnfläche auf einem konstanten Niveau mit einem Zuschlag von etwa 3.5 DM verbleibt. Man erhält somit ein differenzierteres Bild des Effekts der Wohnlage, den man auch als Interaktionseffekt zwischen Wohnfläche und Wohnlagekategorie sehen kann.

Das Modell wurde bereits mit rein parametrischer Modellierung in Beispiel 3.8 (Seite 88) geschätzt. Dort wurde der Interaktionseffekt linear modelliert. Im Vergleich dazu ergeben sich für den variierenden Effekt der besten Lage Unterschiede für Wohnflächen ab 70 qm. Bei nichtparametrischer Modellierung ergibt sich ab einer Wohnfläche von 70 qm eher ein konstanter Effekt. Der Wert des AIC ist im parametrischen Modell 7156.75 und im nonparametrischen Modell 7158.6. Das parametrische Modell ist hier also gleichwertig bzw. sogar geringfügig besser als das nichtparametrische Modell.

△

Falls zur Modellierung der nichtparametrischen Funktionen $f_1, \ldots, f_q$ sowie $f_{z_1|x_1}$ (und $f_{z_2|x_2}$) Basisfunktionenansätze verwendet werden, folgt

$$f_{z_1|x_1}(z_1) = \sum_{l=1}^{d} \gamma_{int,l} B_l(z_1) \quad \text{bzw.} \quad f_{z_1|x_1}(z_1) \cdot x_1 = \sum_{l=1}^{d} \gamma_{int,l} B_l(z_1)\, x_1.$$

Der Vektor der Interaktionsterme $\boldsymbol{f}_{int} = (f_{z_1|x_1}(z_{11})\, x_{11}, \ldots, f_{z_1|x_1}(z_{n1})\, x_{n1})'$ lässt sich mit dem Vektor $\boldsymbol{\gamma}_{int} = (\gamma_{int,1}, \ldots, \gamma_{int,d})'$ der Regressionskoeffizienten durch

$$\boldsymbol{f}_{int} = \boldsymbol{Z}_{int} \boldsymbol{\gamma}_{int}$$

darstellen. Dabei enthält nun die Designmatrix die Werte der Basisfunktionen multipliziert mit den entsprechenden x_1-Werten, d.h. $\boldsymbol{Z}_{int}[i,l] = B_l(z_{i1})\, x_{i1}$. Damit lässt sich das Modell (8.7) in Matrixnotation in der Form

$$\boldsymbol{y} = \boldsymbol{Z}_1 \boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q \boldsymbol{\gamma}_q + \boldsymbol{Z}_{int} \boldsymbol{\gamma}_{int} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

eines großen linearen Modells schreiben. Dies gilt in analoger Weise, wenn mehrere Interaktionsterme bzw. variierende Koeffizienten in das Modell einbezogen werden.

Das Konzept bleibt auch unverändert bestehen, wenn an Stelle von z_1 in geoadditiven Modellen eine Interaktion $f_{geo|x}(s) \cdot x$ zwischen der Lokationsvariablen s und einer binären (oder mehrkategorialen) Variablen x einbezogen wird. Man spricht dann von *Modellen mit räumlich variierenden Koeffizienten.*

8.3.2 Interaktion zwischen zwei metrischen Kovariablen

Für zwei metrische Kovariablen z_1 und z_2 werden Interaktionen in nichtparametrischer Form durch additives Einbeziehen einer glatten zweidimensionalen Funktion $f_{1|2}(z_1, z_2)$ berücksichtigt. Dabei kommen prinzipiell zwei Alternativen in Frage:

- *Zweidimensionale Funktion ohne Haupteffekte:* Der Effekt zwischen z_1 und z_2 wird allein durch die zweidimensionale Funktion $f_{1|2}(z_1, z_2)$ modelliert. Eindimensionale

Haupteffekte $f_1(z_1)$ und $f_2(z_2)$ werden nicht zusätzlich ins Modell aufgenommen. Damit erhalten wir das Modell

$$y_i = f_{1|2}(z_1, z_2) + f_3(z_{i3}) + \ldots + f_q(z_{iq}) + \eta_i^{lin} + \varepsilon_i.$$

Wie bisher ist das Niveau der Funktion $f_{1|2}(z_1, z_2)$ nicht identifizierbar, so dass die Funktion zusätzlich um Null zentriert wird, d.h. es muss gelten

$$\sum_{i=1}^{n} f_{1|2}(z_{i1}, z_{i2}) = 0.$$

Die Interpretation der Ergebnisse ist insgesamt schwieriger und aufwändiger als in reinen Haupteffektmodellen. Im nachfolgenden Beispiel zeigen wir auf wie durch geeignete Darstellung der Ergebnisse die Interpretation erleichtert wird.

Bevor die Schätzung einer zweidimensionalen Interaktion in Angriff genommen wird, sollte man sich unbedingt vergewissern, dass für alle Wertekombinationen von z_1 und z_2 ausreichend Daten zur Verfügung stehen. Dies kann durch ein einfaches Streudiagramm zwischen z_1 und z_2 geschehen. Liegen in Teilbereichen Lücken vor, so sollten zumindest die Ergebnisse in diesen Bereichen mit Vorsicht interpretiert werden. Unter Umständen reicht die vorhandene Datenbasis nicht aus um eine zweidimensionale Oberfläche zwischen z_1 und z_2 zu schätzen, so dass die Schätzung eines reinen Haupteffektmodells vorzuziehen ist.

- *Zweidimensionale Funktion mit Haupteffekten:* Hier wird die zweidimensionale Funktion $f_{1|2}(z_1, z_2)$ zusätzlich zu den eindimensionalen Haupteffekten $f_1(z_1)$ und $f_2(z_2)$ ins Modell aufgenommen. Wir erhalten dann

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + f_{1|2}(z_1, z_2) + f_3(z_{i3}) + \ldots + f_q(z_{iq}) + \eta_i^{lin} + \varepsilon_i.$$

Durch die Hinzunahme der Haupteffekte wird die Frage der Identifizierbarkeit komplizierter. Welche Zusatzbedingungen an $f_{1|2}$ gestellt werden müssen, hängt hier von den für die Haupteffekte und die Interaktion verwendeten Basisfunktionen und auch von der Penalisierung ab. In der Regel wird neben der Zentrierung sämtlicher Funktionen um Null noch gefordert, dass sämtliche Schnitte der Funktion $f_{1|2}(z_1, z_2)$, d.h. alle eindimensionalen Kurven bei festgehaltenem Wert von z_1 oder z_2, ebenfalls um Null zentriert sind. Dadurch wird die Identifizierbarkeit in jedem Fall erreicht und darüber hinaus die Interpretierbarkeit erleichtert. Die Funktion $f_{1|2}(z_1, z_2)$ kann dann als Abweichung von den Haupteffekten interpretiert werden. Diese Modellierungsvariante benötigt allerdings einen noch größeren Stichprobenumfang um zu verlässlichen Ergebnissen zu gelangen. Darüber hinaus erfordert die Interpretation der Interaktionsoberfläche relativ viel Erfahrung.

Für die Modellierung von $f_{1|2}$ werden in der Regel die in den Abschnitten 7.2.1–7.2.3 beschriebenen Tensorprodukt-P-Splines, radiale Basisfunktionen und Kriging-Ansätze verwendet. Wie dort angegeben, lässt sich dann der Vektor

$$\boldsymbol{f}_{1|2} = (f_{1|2}(z_{11}, z_{12}), \ldots, f_{1|2}(z_{n1}, z_{n2}))'$$

wieder durch

$$\boldsymbol{f}_{1|2} = \boldsymbol{Z}_{1|2}\boldsymbol{\gamma}_{1|2}$$

mit einer geeignet definierten Designmatrix $\boldsymbol{Z}_{1|2}$ und einem Koeffizientenvektor $\boldsymbol{\gamma}_{1|2}$ darstellen. Somit erhält man wiederum ein großes lineares Modell

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{Z}_{1|2}\boldsymbol{\gamma}_{1|2} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

wobei die Koeffizienten $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{1|2}$ unter Berücksichtigung von Bestrafungsrestriktionen zu schätzen sind, vergleiche Abschnitt 8.5.

Beispiel 8.4 Mietspiegel München – Interaktion zwischen Wohnfläche und Baujahr

Wir ersetzen im geoadditiven Modell von Beispiel 8.2 die beiden Haupteffekte der Wohnfläche und des Baujahrs durch eine zweidimensionale Oberfläche $f_{1|2}(\mathit{flaeche}, \mathit{bjahr})$ und erhalten das Modell

$$\mathit{mieteqm} = f_{1|2}(\mathit{flaeche}, \mathit{bjahr}) + f_{geo}(\mathit{bezv}) + \beta_0 + \beta_1\, \mathit{glage} + \beta_2\, \mathit{blage} + \varepsilon.$$

Die zweidimensionale Funktion $f_{1|2}$ wurde wie in Beispiel 7.5 (Seite 368) durch einen Tensorprodukt P-Spline modelliert. Für den Strafterm wurden sowohl erste als auch zweite Differenzen getestet. Für das AIC erhält man bei Verwendung erster Differenzen den Wert 11433.7 und bei Verwendung zweiter Differenzen den Wert 11420.6. In beiden Fällen ist das AIC deutlich niedriger als im geoadditiven Modell aus Beispiel 8.2 (AIC = 11450). Im Folgenden beschränken wir uns auf die Beschreibung des besten Modells basierend auf Straftermen mit zweiten Differenzen.

Bevor wir die Schätzergebnisse interpretieren, betrachten wir noch die Verteilung der beobachteten Werte für die Wohnfläche und das Baujahr, vergleiche Abbildung 8.6 links. Die Abbildung führt die Schwierigkeiten bei der Schätzung zweidimensionaler Kovariableneffekte deutlich vor Augen. Offenbar sind in den Bereichen mit Wohnflächen kleiner als 30 und größer als 110 Quadratmeter Wohnfläche zu wenige Beobachtungen vorhanden um verlässliche Aussagen zu treffen. Hier wären schlecht interpretierbare Effekte nicht überraschend. Für Wohnungen zwischen 30 und 110 Quadratmeter Wohnfläche scheint die Datenlage aber ausreichend (Abbildung 8.6 rechts). Wir beschränken daher unsere Interpretation auf diesen Bereich. In der Praxis hat man im Wesentlichen die Wahl zwischen vier Alternativen: Die zusätzliche Erhebung neuer Daten, die zusätzliche Berücksichtigung von Daten aus früheren Mietspiegeln, die Einschränkung der Gültigkeit des Mietspiegel auf Wohnungen zwischen 30 und 110 Quadratmetern Wohnfläche oder die Verwendung von Modellen mit geringerer Komplexität (insbesondere reine Haupteffektmodelle). Am besten umsetzbar erscheint die letztgenannte Alternative.

Bei der Darstellung der Schätzergebnisse beschränken wir uns auf die Interaktion zwischen Wohnfläche und Baujahr. Die räumlichen Effekte bleiben im Vergleich zu Beispiel 8.2 nahezu unverändert. Abbildung 8.7 (links) zeigt die geschätzte zweidimensionale Funktion der Wohnfläche und des Baujahrs. Häufig fällt die Interpretation zweidimensionaler Oberflächen nicht leicht, so dass es sinnvoll sein kann, zusätzlich einige Schnitte der zweidimensionalen Funktion zu zeichnen. Die rechte Grafik von Abbildung 8.7 zeigt daher den geschätzten Effekt der Wohnfläche für die Baujahre 1918, 1957, 1972 und 1991. Die Schnittfunktionen zeigen visuell, ob die Modellierung einer Interaktion überhaupt notwendig ist. Verlaufen die Schnittfunktionen nahezu parallel liegt keine Interaktion vor und ein Modell mit den beiden Haupteffekten ist ausreichend. Im vorliegenden Fall verlaufen die Schnitte keineswegs parallel, so dass die Einbeziehung einer Interaktion gerechtfertigt erscheint. Diese Beobachtung wird durch die deutlich niedrigeren AIC-Werte für die Interaktionsmodelle im Vergleich zu den Haupteffektmodellen zusätzlich gestützt. Insbesondere anhand der Schnittfunktionen erkennen wir, dass im Bereich kleiner Wohnungen der größte Interaktionseffekt vorliegt. Bei älteren und moderneren Wohnungen ist der Abfall der durchschnittlichen Miete pro qm deutlich flacher

Abb. 8.6. *Verteilung der Wohnfläche und des Baujahrs. Links: Alle Beobachtungen. Rechts: Nur Beobachtungen mit Wohnfläche zwischen 30 und 110 Quadratmeter Wohnfläche.*

Abb. 8.7. *Geschätzte zweidimensionale Funktion* $\hat{f}_{1|2}(flaeche, bjahr)$ *(links) und Schnitte für verschiedene Baujahre (rechts).*

als bei Wohnungen mit mittlerem Baujahr. Für sehr alte Wohnungen verläuft der Effekt der Wohnfläche sehr flach und sogar annähernd linear.

△

8.4 Strukturiert-additive Regression

Die Modellterme der bisher betrachteten Modellklassen lassen sich prinzipiell beliebig kombinieren. Beispielsweise könnte ein Modell neben rein additiven Effekten zusätzlich einen räumlichen Effekt und ein oder mehrere Terme mit variierenden Koeffizienten zur Modellierung von Interaktionen enthalten. Darüber hinaus können auch zufällige Effekte wie in gemischten Modellen einbezogen werden. Die so entstehenden Modelle lassen sich in einheitlicher und allgemeiner Form als *strukturiert-additive Regressions-(STAR)-Modelle*

$$y = f_1(v_1) + \ldots + f_q(v_q) + \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \tag{8.8}$$

beschreiben. Dabei sind $v_1, \ldots, v_q$ ein- oder auch mehrdimensionale Variablen verschiedenen Typs, die aus den eigentlichen Kovariablen gebildet werden. Ebenso sind auch

die Funktionen von unterschiedlicher Struktur. Für folgende Typen von Variablen und Funktionen erhält man die bisherigen Modelle als Spezialfälle, wobei auch nicht im Detail betrachtete Kombinationsmöglichkeiten zulässig sind:

$f_1(v_1) = f_1(z_1),$	$v_1 = z_1,$	nichtlinearer Effekt von z_1.
$f_2(v_2) = f_{geo}(s),$	$v_2 = s,$	räumlicher Effekt der Lokationsvariablen s.
$f_3(v_3) = f(z)\,x,$	$v_3 = (z, x),$	mit z variierender Effekt von x.
$f_4(v_4) = f_{1\|2}(z_1, z_2),$	$v_4 = (z_1, z_2),$	nichtlineare Interaktion zwischen z_1 und z_2.
$f_5(v_5) = \gamma_i u,$	$v_5 = (u, i),$	individuenspezifischer zufälliger Effekt von u.

Formal könnte auch der lineare Teil $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = \boldsymbol{x}'\boldsymbol{\beta}$ des Prädiktors als Funktion $f_0(\boldsymbol{v}_0) = \boldsymbol{x}'\boldsymbol{\beta}$, $\boldsymbol{v}_0 = (1, x_1, \ldots, x_k)'$, geschrieben werden. Da der lineare Teil, zumindest die Konstante β_0, immer im Modell ist, führen wir ihn in (8.8) explizit auf.

Falls die Funktionen $f_1, \ldots, f_q$ durch Basisfunktionenansätze modelliert und geschätzt werden, erhält man mit den Vektoren $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_q$ der Funktionswerte an den Variablenausprägungen von $v_1, \ldots, v_q$ ein großes lineares Modell

$$\boldsymbol{y} = \boldsymbol{f}_1 + \ldots + \boldsymbol{f}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{V}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

mit geeignet definierten Designmatrizen $\boldsymbol{V}_j$ und Koeffizientenvektoren $\boldsymbol{\gamma}_j$. Die Rauheit der Funktionen wird durch Strafterme der Form $\lambda_j \boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j$ mit Glättungsparameter λ_j und Strafmatrix $\boldsymbol{K}_j$ bestraft. In einem vollen Bayes-Ansatz korrespondieren die Strafterme mit Glattheitsprioris der Form

$$p(\boldsymbol{\gamma}_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j\right).$$

Jeder Modellterm in STAR-Modellen wird also durch die Gestalt der Designmatrix $\boldsymbol{V}_j$ und der Strafmatrix $\boldsymbol{K}_j$ charakterisiert. Einen Überblick über häufig verwendete Terme gibt Tabelle 8.2. Auf Details zur Schätzung der Parameter gehen wir im nächsten Abschnitt ein.

Das einheitliche Konzept von STAR-Modellen lässt sich auch auf Regressionssituationen mit nicht-normalverteilten, insbesondere binären, diskreten und kategorialen Zielgrößen erweitern. Dazu nimmt man wie in generalisierten linearen Modellen an, dass die Zielvariablen y_i bei gegebenen Prädiktoren η_i (bedingt) unabhängig sind und der strukturiert-additive Prädiktor

$$\eta_i^{strukt} = f(v_{i1}) + \ldots + f(v_{iq}) + \boldsymbol{x}_i'\boldsymbol{\beta}$$

durch die Verknüpfung

$$\mathrm{E}(y_i) = \mu_i = h(\eta_i^{strukt})$$

mittels einer Funktion h mit dem (bedingten) Erwartungswert $\mathrm{E}(y_i) = \mu_i$ verbunden ist. Wir sprechen dann von einem *generalisierten STAR-Modell*.

Diese allgemeine Modellklasse enthält die bisher behandelten Modellklassen als Spezialfälle. Zum Beispiel erhält man *generalisierte additive Modelle* (GAM), wenn

$$\eta_i^{strukt} = \eta_i^{add} = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \boldsymbol{x}_i'\boldsymbol{\beta}$$

gewählt wird. Mit

Termtyp	Designmatrix $\boldsymbol{V}$	Strafmatrix $\boldsymbol{K}$	Kapitel
P-Spline	Basisfunktionen ausgewertet an den Beobachtungen.	$\boldsymbol{K} = \boldsymbol{D}_k'\boldsymbol{D}_k$, mit $\boldsymbol{D}_k$ Differenzenmatrix k-ter Ordnung.	7.1.2
Allgemeine Basisfunktionenansätze	Basisfunktionen ausgewertet an den Beobachtungen.	Basierend auf dem Integral der quadrierten zweiten Ableitung.	7.1.3
Variierender Koeffizient	Basisfunktionen multipliziert mit der Interaktionsvariablen.	$\boldsymbol{K} = \boldsymbol{D}_k'\boldsymbol{D}_k$ wie für P-Splines.	8.3.1
2D-P-Spline	2D-Basisfunktionen ausgewertet an den Beobachtungen.	$\boldsymbol{K} = \boldsymbol{I} \otimes \boldsymbol{K}_1 + \boldsymbol{K}_2 \otimes \boldsymbol{I}$ mit Einheitsmatrix $\boldsymbol{I}$ und Strafmatrizen $\boldsymbol{K}_1$ und $\boldsymbol{K}_2$ wie für univariate P-Splines.	7.2.1
Kriging	Auf der Korrelationsfunktion basierende Basisfunktionen.	$\boldsymbol{K} = \boldsymbol{R}$, mit Korrelationsmatrix $\boldsymbol{R}$.	7.1.6, 7.2.3
Markov-Zufallsfeld	0/1-Inzidenzmatrix, die Beobachtungen und Regionen verknüpft.	$\boldsymbol{K}$ = Nachbarschaftsmatrix.	7.2.4
Zufällige Konstante	0/1-Inzidenzmatrix, die Beobachtungen und Cluster verknüpft.	$\boldsymbol{K} = \boldsymbol{I}$, mit Einheitsmatrix $\boldsymbol{I}$.	6

Tabelle 8.2. *Übersicht über verschiedene, häufig verwendete Modellterme in STAR-Modellen.*

$$\eta_i^{strukt} = \eta_i^{add} + f_{geo}(s_i)$$

erhält man *generalisierte geoadditive Modelle*, usw.

Beispiel 8.5 Schadenshäufigkeiten bei Kfz-Versicherung

Bereits in Kapitel 2 in den Beispielen 2.12, 2.13 und 2.15 haben wir Beispiele für generalisierte STAR-Modelle kennengelernt. In Beispiel 2.15 (Seite 51) wurde die erwartete Schadenshäufigkeit einer Kfz-Versicherung durch ein *strukturiert-additives Poisson-Modell* mit Prädiktor

$$\eta^{strukt} = f_1(alterv) + f_2(alterkfz) + f_3(ps) + f_4(bm) + f_5(district) + \beta_0 + \beta_1 geschl_i + \ldots$$

modelliert. In einer Fallstudie von Denuit & Lang (2005) wurde zusätzlich eine Interaktion zwischen dem Geschlecht und dem Alter des Versicherungsnehmers in Form eines mit dem Alter variierenden Geschlechtseffekts modelliert. Damit erhält man dann den erweiterten strukturiert additiven Prädiktor

$$\begin{aligned}\eta^{strukt} &= f_1(alterv) + f_2(alterkfz) + f_3(ps) + f_4(bm) + f_5(district) + \\ &\quad + f_6(alterv, geschl) + \beta_0 + \beta_1 geschl + \ldots,\end{aligned}$$

wobei die Interaktion f_6 die Form $f_6(alterv, geschl) = f(alterv)geschl$ eines variierenden Koeffizienten Terms mit nichtparametrisch modellierter Funktion f besitzt. Abbildung 8.8 (links) zeigt den Effekt $\hat{f}_1$ des Alters des Versicherungsnehmers für Frauen und den Alterseffekt $\hat{f}_1 + \hat{f}_6 + \hat{\beta}_1$ für Männer ($geschl = 1$). Die rechte Grafik von Abbildung 8.8 zeigt den in Abhängigkeit des Alters variierenden Effekt $\hat{f}_6 + \hat{\beta}_1$ der Männer im Vergleich zu den Frauen. Die Schätzungen wurden jeweils mit Hilfe der Darstellung von STAR-Modellen als gemischte Modelle erzielt, vergleiche Abschnitt 8.5.2.

Abb. 8.8. *Links: Effekt des Alters des Versicherungsnehmers für Frauen (durchgezogene Linie) und Männer (gestrichelte Linie). Rechts: In Abhängigkeit des Alters variierender Effekt der Männer im Vergleich zu den Frauen.*

Für Frauen nimmt die durchschnittliche Schadenshäufigkeit mit zunehmendem Alter linear ab. Für Männer ergibt sich die bereits in den Beispielen von Kapitel 2 festgestellte nichtlineare Form. Jüngere und sehr alte Männer haben deutlich erhöhte Schadenshäufigkeiten. Männer im Alter zwischen ca. 30 und 80 verursachen sogar etwas weniger Schäden als Frauen im gleichen Alter. Insgesamt besteht ein stark ausgeprägter Interaktionseffekt. Auf die Darstellung der übrigen Schätzergebnisse verzichten wir, da sich kaum Unterschiede zu den in Kapitel 2 dargestellten Effekten ergeben.

△

Im nächsten Beispiel wird gezeigt, wie der Ansatz für strukturiert-additive *kategoriale* Regressionsmodelle erweitert werden kann. Außerdem sehen wir, dass bei der Modellierung räumlicher Effekte Vorsicht angebracht ist, da (beispielsweise bei ökologischen Anwendungen) einige Kovariablen *lokationsspezifisch* sind und daher der räumliche Effekt den Einfluss dieser Variablen überdecken kann.

Beispiel 8.6 Zustand des Waldes

Wir führen die in den Kapiteln 1 und 5 betrachteten Beispiele 1.4 (Seite 9) und 5.6 (Seite 250) fort und erweitern den linearen Prädiktor $\eta_{it}^{(r)} = \beta_{0r} + \boldsymbol{x}_{it}'\boldsymbol{\beta}$, $r = 1, 2$ des dreikategorialen ordinalen Logit-Modells $P(Y_{it} = r)$ für Baum i im Jahr t zum (zunächst rein) additiven Prädiktor

$$\eta_{it}^{(r)} = \beta_{0r} + f_1(alter_{it}) + f_2(hang_i) + f_3(schirm_{it}) + f_4(t) + f_5(hoehe_i) + \boldsymbol{x}_{it}'\boldsymbol{\beta}.$$

Im Vergleich zu Beispiel 5.6 werden die parametrischen Modellierungen der meisten metrischen Kovariablen durch nichtparametrische Effekte ersetzt. Lediglich der Effekt der Gründigkeit (*grund*) und des Ph-Werts (*ph*) werden weiterhin linear modelliert, da hier frühere nichtparametrische Modellierungen jeweils lineare Effekte ergaben. Die Modellierung der nichtlinearen Effekte f_1 bis f_5 erfolgt durch kubische P-Splines mit 20 inneren Knoten und Straftermen basierend auf zweiten Differenzen. Der Kovariablenvektor $\boldsymbol{x}_{it}$ enthält die restlichen Regressoren mit linearen Effekten $\boldsymbol{\beta}$ wie in Tabelle 8.3 angegeben. Abbildung 8.9 enthält die (mit Hilfe der Methodik für gemischte Modelle, vgl. Abschnitt 8.5.2) geschätzten nichtlinearen Effekte f_1 bis f_5.

Für das AIC wird der Wert 1682.8 erreicht, was eine deutliche Verringerung im Vergleich zum rein parametrischen Modell in Beispiel 5.6 (Seite 250) bedeutet. Hier wird lediglich ein AIC-Wert von 2029.16 erreicht.

Abb. 8.9. *Waldschäden: Schätzungen der nonparametrischen Effekte im Modell ohne räumlichen Effekt inklusive punktweisen 80% und 95% Konfidenzintervallen.*

Ein Blick auf die Kovariableneffekte zeigt jedoch einige überraschende Ergebnisse:

- Wie bereits im parametrischen Modell scheinen ältere Bäume ähnlich wie jüngere Bäume weniger geschädigt. Zu erwarten wäre hier eher ein monoton steigender Zusammenhang.
- Der Effekt der Hangneigung ist extrem rau und praktisch nicht interpretierbar.
- Bäume in kleiner und großer Höhe scheinen mehr geschädigt als Bäume in mittleren Höhenlagen. Auch dieser Effekt ist überraschend. Eher würde man einen monoton steigenden Effekt erwarten.
- Die Effekte der linearen Kovariablen sind im Vergleich mit den Ergebnissen aus Beispiel 5.6 relativ instabil, vergleiche etwa den Effekt des Alkali-Anteils.

	Koeffizient	p-Wert	Koeffizient	p-Wert
β_{01}	-4.000	0.009	-1.515	0.453
β_{02}	1.107	0.470	4.846	0.019
grund	-0.048	<0.001	-0.006	0.838
ph	-0.718	0.018	-0.249	0.491
artkat	0.644	<0.001	0.202	0.373
dueng	-0.571	0.004	-0.570	0.253
humus0	-0.419	0.008	-0.453	0.014
humus2	0.413	0.002	0.279	0.057
humus3	0.087	0.573	0.218	0.223
humus4	-0.032	0.879	0.192	0.443
frische1	-0.801	<0.001	-0.762	0.167
frische3	0.483	0.001	0.812	0.031
alkali1	1.165	<0.001	0.920	0.144
alkali3	-0.966	<0.001	-0.359	0.523
alkali4	0.195	0.507	-0.201	0.798

Tabelle 8.3. *Waldschäden: Schätzungen der fixen Effekte im Modell ohne räumlichen Effekt (links) und im Modell mit räumlichem Effekt (rechts).*

Wie lassen sich diese eigenartigen Ergebnisse erklären? Als ersten Schritt erweitern wir das Regressionsmodell um einen zusätzlichen räumlichen Effekt und verwenden zur Modellierung einen zweidimensionalen P-Spline. Tabelle 8.3 zeigt wieder die resultierenden linearen Effekte, während Abbildung 8.10 diejenigen nichtparametrisch modellierten Effekte darstellt, die sich bei Einbezug des räumlichen Effekts deutlich ändern. Abbildung 8.11 zeigt den räumlichen Effekt $f_{geo}(s)$. Jetzt können wir folgendes beobachten:

- Der Abfall des Alterseffekts für sehr alte Bäume verschwindet.
- Die Effekte der Gründigkeit (*grund*), des Ph-Wertes (*ph*), der Art des Waldes (*artkat*), des Alkali-Anteils (*alkali*), der Hangneigung (*hang*) und der Höhenlage (*hoehe*) verschwinden oder sind zumindest stark verringert. Bei all diesen Variablen handelt es sich um *standortspezifische Kovariablen*, deren Werte sich über die Zeit nicht oder kaum verändern.

Insbesondere die letzte Beobachtung liefert den entscheidenden Hinweis für die Erklärung der überraschenden Schätzergebnisse. Im ersten Modell ohne räumlichen Effekt wurden die standortspezifischen Effekte zum Teil durch die Kovariablen modelliert. Jedoch scheinen die standortspezifischen Einflüsse insgesamt sehr kompliziert, so dass die Hangneigung und die Höhenlage offenbar den Einfluss weiterer, unbeobachteter bzw. nicht berücksichtigter Kovariablen auffangen und damit deren Effekte schlecht interpretierbar sind. Im zweiten Modell werden durch die sehr flexible Modellierung des räumlichen Effekts die Einflüsse der standortspezifischen Kovariablen überdeckt und größtenteils in diesem Effekt vereint. Dies ist natürlich nicht wünschenswert, da ja der räumliche Effekt lediglich die Einflüsse unbeobachteter Kovariablen auffangen und nicht die Effekte beobachtbarer Kovariablen überdecken soll.

Welches Fazit können wir ziehen? Offenbar ist in dieser Anwendung eine Trennung der verschiedenen Einflüsse nicht möglich oder zumindest erschwert. Ein möglicher Weg wäre die Schätzung eines Modells mit räumlichem Effekt aber ohne standortspezifische Kovariablen. Anschließend könnte man durch sorgfältige Untersuchung der Eigenschaften der Standorte mit besonders großem (bzw- kleinem) räumlichen Effekt versuchen die relevanten Einflussgrößen zu extrahieren. Eventuell könnte hier auch die Definition neuer Kovariablen, etwa geeignet gewählter Linearkombinationen, von Nutzen sein.

△

Abb. 8.10. *Waldschäden: Schätzungen ausgewählter nichtparametrischen Effekte im Modell mit räumlichem Effekt inklusive punktweisen 80% und 95% Konfidenzintervallen.*

Abb. 8.11. *Waldschäden: Schätzung des räumlichen Effekts.*

8.5 Inferenz

Wir beschränken uns auf die Schätzung von STAR-Modellen basierend auf Basisfunktionenansätzen. Ähnlich wie in Kapitel 7 können wir drei unterschiedliche Inferenzkonzepte unterscheiden:

- Penalisierte KQ- bzw. Likelihood-Schätzung in Verbindung mit der Minimierung eines Modellwahlkriteriums zur Schätzung der Glättungsparameter (Abschnitt 8.5.1).

- Inferenz basierend auf der Repräsentation als gemischtes Modell (Abschnitt 8.5.2).
- Volle Bayes-Inferenz basierend auf MCMC-Simulationstechniken (Abschnitt 8.5.3).

8.5.1 Penalisierte KQ- bzw- Likelihood-Schätzung

Wir gehen zunächst von STAR-Modellen

$$\boldsymbol{y} = \boldsymbol{f}_1 + \ldots + \boldsymbol{f}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{V}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8.9}$$

mit metrischer Zielgröße $\boldsymbol{y}$ aus. Zur Schätzung wird in Erweiterung der penalisierten KQ-Schätzung aus Kapitel 7 das penalisierte KQ-Kriterium

$$\mathrm{KQ}_{pen}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = \mathrm{KQ}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) + \sum_{j=1}^{q} \lambda_j \boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j \tag{8.10}$$

bezüglich $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}$ minimiert. Dabei ist

$$\mathrm{KQ}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{V}_1\boldsymbol{\gamma}_1 - \ldots - \boldsymbol{V}_q\boldsymbol{\gamma}_q - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{V}_1\boldsymbol{\gamma}_1 - \ldots - \boldsymbol{V}_q\boldsymbol{\gamma}_q - \boldsymbol{X}\boldsymbol{\beta})$$

das gewöhnliche KQ-Kriterium ohne Berücksichtigung der quadratischen Straffunktionen $\lambda_j \boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j$. Die spezifische Form der Strafmatrix $\boldsymbol{K}_j$ hängt wie in Kapitel 7 vom speziellen Typ der Funktion f_j und der gewählten Penalisierung ab. Einen Überblick gibt Tabelle 8.2. Der Glättungsparameter λ_j steuert das Ausmaß der quadratischen Bestrafung $\boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j$. Wir setzen zunächst λ_j als bekannt oder gegeben voraus.
Bei der Minimierung haben wir die Wahl zwischen zwei Alternativen: Zum Einen eine iterative Minimierung mit Hilfe des sogenannten Backfitting-Algorithmus und zum Anderen die direkte Minimierung des penalisierten KQ-Kriteriums. Der Vorteil des Backfitting-Algorithmus beruht darin, dass die Schätzung auf die ein- oder zweidimensionale Glättung, wie in Kapitel 7 beschrieben, zurückgeführt werden kann. Damit erlaubt der Backfitting-Algorithmus im Prinzip die Kombination beliebiger Glättungsverfahren als Bausteine für die Schätzung von STAR-Modellen. D.h. es können nicht nur Basisfunktionenansätze, sondern auch lokale Glätter als Bausteine verwendet werden. Auch eine Mischung verschiedener Ansätze ist möglich.

Backfitting

Die Idee des Backfitting ist es, die Schätzer $\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_q, \hat{\boldsymbol{\beta}}$ iterativ durch Glättung von Residuen zu berechnen. Unter Vernachlässigung des Fehlers $\boldsymbol{\varepsilon}$ in (8.9) gilt approximativ

$$\boldsymbol{f}_j \approx \boldsymbol{y} - \boldsymbol{f}_1 - \ldots - \boldsymbol{f}_{j-1} - \boldsymbol{f}_{j+1} - \ldots - \boldsymbol{f}_q - \boldsymbol{X}\boldsymbol{\beta}.$$

Falls bereits Schätzungen für $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{f}}_k$, $k \neq j$ vorliegen, lässt sich $\boldsymbol{y} - \sum_{k \neq j} \hat{\boldsymbol{f}}_k - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ als (partieller) Residuenvektor ohne $\boldsymbol{f}_j$ auffassen. Für die Schätzung von $\boldsymbol{f}_j$ wendet man den zu $\boldsymbol{f}_j$ gehörigen Streudiagramm- oder auch bivariaten Glätter $\boldsymbol{S}_j$ auf diesen Residuenvektor an und erhält

$$\hat{\boldsymbol{f}}_j = \boldsymbol{S}_j(\boldsymbol{y} - \sum_{k \neq j} \hat{\boldsymbol{f}}_k - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Sämtliche Schätzer basierend auf Basisfunktionenansätzen haben dabei die Form

$$\hat{\boldsymbol{f}}_j = (\boldsymbol{V}_j'\boldsymbol{V}_j + \lambda_j \boldsymbol{K}_j)^{-1}\boldsymbol{V}_j'(\boldsymbol{y} - \sum_{k \neq j} \hat{\boldsymbol{f}}_k - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Die spezielle Gestalt von $\boldsymbol{V}_j$ und $\boldsymbol{K}_j$ hängt vom Typ des verwendeten Glätters ab, vergleiche Tabelle 8.2.

Ausgehend von einer einfachen Startlösung wird dieses Konzept iterativ für $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_q$ durchgeführt. Dies ergibt den *Backfitting-Algorithmus*:

Backfitting-Algorithmus

1. Initialisierung: Setze z.B. $\hat{\boldsymbol{f}}_1 = \ldots = \hat{\boldsymbol{f}}_q = \boldsymbol{0}, \hat{\boldsymbol{\beta}} = \boldsymbol{0}$.
2. Erhalte verbesserte Schätzungen $\hat{\boldsymbol{f}}_j, j = 1, \ldots, q$, durch
$$\hat{\boldsymbol{f}}_j = \boldsymbol{S}_j(\boldsymbol{y} - \sum_{k \neq j} \hat{\boldsymbol{f}}_k - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$
3. Erhalte eine verbesserte Schätzung $\hat{\boldsymbol{\beta}}$ durch
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \hat{\boldsymbol{f}}_1 - \ldots - \hat{\boldsymbol{f}}_q).$$
4. Iteriere die Schritte 2–3 so lange bis sich die geschätzten Funktionen bis auf eine kleine, vorgegebene Abweichung nicht mehr ändern.

In dieser vereinfachten Darstellung wird nicht sichtbar, dass die Glätter $\boldsymbol{S}_j$ in der Regel noch unbekannte Glättungsparameter enthalten. Wie diese ebenfalls datengesteuert gewählt werden können, besprechen wir weiter unten.

Direkte Minimierung des penalisierten KQ-Kriteriums

Der penalisierte KQ-Schätzer kann in vielen Fällen auch in einem nicht iterativen Verfahren durch direkte Minimierung von (8.10) bestimmt werden. Die Minimierung erfolgt durch Nullsetzen der ersten Ableitung und Auflösen des resultierenden Gleichungssystems nach den unbekannten Parametern.

Nullsetzen der ersten Ableitungen führt auf das Gleichungssystem

$$\begin{pmatrix} \boldsymbol{V}_1'\boldsymbol{V}_1 + \lambda_1\boldsymbol{K}_1 & \cdots & \boldsymbol{V}_1'\boldsymbol{V}_q & \boldsymbol{V}_1'\boldsymbol{X} \\ \vdots & \ddots & & \\ \boldsymbol{V}_q'\boldsymbol{V}_1 & & \boldsymbol{V}_q'\boldsymbol{V}_q + \lambda_q\boldsymbol{K}_q & \boldsymbol{V}_q'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{V}_1 & \cdots & \boldsymbol{X}'\boldsymbol{V}_q & \boldsymbol{X}'\boldsymbol{X} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_1'\boldsymbol{y} \\ \vdots \\ \boldsymbol{V}_q'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} \end{pmatrix}. \tag{8.11}$$

Die direkte (nicht iterative) Lösung des Gleichungssystem funktioniert gut für eine nicht zu große Zahl von Regressionskoeffizienten, etwa wenn für kleines q die Funktionen $f_1, \ldots, f_q$ durch P-Splines mit einer nicht zu großen Anzahl von Knoten spezifiziert werden. Für hohe Dimensionen von $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ bietet die numerische Analysis den iterativen *Gauss-Seidel-Algorithmus* zur Lösung des Gleichungssystems (8.11) an. Es zeigt sich,

dass dieser mit dem oben beschriebenen Backfitting-Algorithmus identisch ist, vergleiche Buja, Hastie & Tibshirani (1989). Unter Verwendung der geschätzten Regressionskoeffizienten lassen sich dann ähnlich wie im linearen Modell Schätzungen $\hat{\boldsymbol{y}}$ für den (bedingten) Erwartungswert $\mathrm{E}(\boldsymbol{y})$ der Zielgrößen bestimmen als $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, wobei man die Prädiktionsmatrix bzw. Gesamt-Glättungsmatrix $\boldsymbol{H}$ erhält als

$$\boldsymbol{H} = (\boldsymbol{V}_1\,\boldsymbol{V}_2 \ldots \boldsymbol{V}_q\,\boldsymbol{X}) \begin{pmatrix} \boldsymbol{V}_1'\boldsymbol{V}_1 + \lambda_1\boldsymbol{K}_1 & \cdots & \boldsymbol{V}_1'\boldsymbol{V}_q & \boldsymbol{V}_1'\boldsymbol{X} \\ \vdots & \ddots & & \\ \boldsymbol{V}_q'\boldsymbol{V}_1 & & \boldsymbol{V}_q'\boldsymbol{V}_q + \lambda_q\boldsymbol{K}_q & \boldsymbol{V}_q'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{V}_1 & \cdots & \boldsymbol{X}'\boldsymbol{V}_q & \boldsymbol{X}'\boldsymbol{X} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{V}_1' \\ \vdots \\ \boldsymbol{V}_q' \\ \boldsymbol{X}' \end{pmatrix}.$$

Generalisierte STAR-Modelle

Für generalisierte strukturiert-additive Modelle mit nicht-normalverteilten Zielvariablen und Prädiktor $\boldsymbol{\eta}^{strukt} = \boldsymbol{V}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta}$ wird zur Schätzung das penalisierte Log-Likelihood-Kriterium

$$l_{pen}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = l(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{q} \lambda_j \boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j$$

bezüglich $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}$ maximiert. Dabei ist $l(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta})$ die übliche Log-Likelihood von generalisierten linearen Modellen für den Prädiktor $\boldsymbol{\eta}^{strukt}$. Nullsetzen der ersten Ableitung liefert das hochdimensionale Gleichungssystem

$$\begin{aligned} \boldsymbol{s}_{pen,0}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) &= \boldsymbol{s}_0(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = \boldsymbol{0} \\ \boldsymbol{s}_{pen,j}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) &= \boldsymbol{s}_j(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) - \lambda_j \boldsymbol{K}_j \boldsymbol{\gamma}_j = \boldsymbol{0}. \end{aligned}$$

Die Funktionen $\boldsymbol{s}_0(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = \partial\, l / \partial\, \boldsymbol{\beta}$, $\boldsymbol{s}_j(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}) = \partial\, l / \partial\, \boldsymbol{\gamma}_j$ besitzen dabei die Form der Score-Funktionen in generalisierten linearen Modellen. Die Struktur des Gleichungssystems ähnelt (8.11), wobei aber zusätzlich wie in generalisierten linearen Modellen Arbeitsgewichte und Arbeitsbeobachtungen enthalten sind. Das Gleichungssystem kann durch einen Fisher-Scoring-Algorithmus ähnlich wie in Kapitel 4 oder in Kombination mit Backfitting gelöst werden, siehe z.B. Fahrmeir & Tutz (2001, Kap. 5) und Tutz (2000, Kap. 8).

Schätzung der Glättungsparameter

Zur Schätzung der Glättungsparameter kann man geeignete Modellwahlkriterien wie AIC oder GCV bezüglich der Glättungsparameter optimieren. Dabei muss jeweils die Anzahl effektiver Parameter des Modells berücksichtigt werden, die sich in Verallgemeinerung von Abschnitt 7.1.8 (Seite 345) als Spur der Gesamt-Glättungsmatrix $\boldsymbol{H}$ eines STAR-Modells definieren lassen. Diese Glättungsmatrix bildet für normalverteilte Zielgrößen den Vektor $\boldsymbol{y}$ auf $\hat{\boldsymbol{y}}$ ab. Für nicht-normalverteilte Zielvariablen wird die entsprechende Matrix aus dem Arbeitsmodell verwendet. Für STAR-Modelle mit normalverteilten Fehlern ist das AIC Kriterium damit gegeben durch

$$\mathrm{AIC} = n \log(\hat{\sigma}^2) + 2(\mathrm{sp}(\boldsymbol{H}) + 1),$$

wobei $\hat{\sigma}^2 = \sum(y_i - \hat{\eta}_i^{strukt})^2/n$ der ML-Schätzer der Fehlervarianz ist. Als generalisierte Kreuzvalidierungsfunktion GCV erhält man

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\eta}_i^{strukt}}{1 - \mathrm{sp}(\boldsymbol{H})/n}\right)^2 .$$

Bis vor wenigen Jahren wurde überwiegend eine einfache aber wenig effektive Gittersuche bei der Minimierung der Gütemaße eingesetzt, so dass Probleme mit mehr als 4 – 5 Glättungsparametern kaum zu bewältigen waren. Mittlerweile existieren effiziente Algorithmen zur Optimierung der Modellwahlkriterien, die aber technisch aufwändig sind, so dass wir an dieser Stelle auf eine genauere Beschreibung verzichten. Eine detaillierte Beschreibung sowohl zum methodischen Hintergrund als auch zur Benutzung der im R-Paket `mgcv` implementierten Verfahren findet man für generalisierte additive Modelle in Wood (2006).

Wird der iterative Backfitting-Algorithmus zur Schätzung der Regressionskoeffizienten verwendet, ist die Gesamt-Glättungsmatrix nicht erhältlich und damit die Bestimmung der Gütemaße so nicht möglich. In der Praxis approximiert man daher die Spur der Gesamt-Glättungsmatrix durch die Summe der Spuren der einzelnen Glättungsmatrizen. Solange die einzelnen Terme nicht zu stark korreliert sind funktioniert diese Approximation verhältnismäßig gut.

Modellwahl und Diagnose

Zur Modellwahl und Diagnose in nicht- und semiparametrischen Modellen existiert relativ wenig Literatur und noch weniger einfach benutzbare Software. Einige der in Kapitel 3.6 zur Modellwahl in linearen Modellen behandelten Konzepte lassen sich aber auf STAR-Modelle leicht übertragen. So können konkurrierende Modelle beispielsweise durch Gütemaße wie AIC und GCV verglichen werden.

Zur Modelldiagnose können vor allem die verschiedenen Arten von Residuen verwendet werden. Einfache, standardisierte und partielle Residuen lassen sich völlig analog zum linearen Modell definieren, vergleiche für den Fall metrischer Zielgrößen den nachfolgenden Kasten. Ein Analogon zu studentisierten Residuen ist allerdings nicht verfügbar. Im Fall generalisierter STAR-Modelle findet man Hinweise zur Definition und Anwendung von Residuen in den bereits in Kapitel 4.7 genannten Büchern McCullagh & Nelder (1989), Tutz (2000) sowie Collett (1991). Maße zur Einfluss- und Ausreißeranalyse wie die Cook-Distanz sind prinzipiell übertragbar, bisher jedoch nicht oder kaum untersucht und vor allem nicht in den gängigen Software-Paketen implementiert.

8.5.2 Inferenz basierend auf der Repräsentation als gemischtes Modell

Ein zweiter Ansatz zur Schätzung von STAR-Modellen beruht auf der Repräsentation der Modelle durch gemischte Modelle. In Abschnitt 7.1.9 wurde bereits die Repräsentation von Penalisierungsansätzen als gemischte Modelle behandelt. Hier gehen wir kurz auf die Erweiterung der Vorgehensweise für STAR-Modelle ein.

Dazu wird für jede Modellkomponente $\boldsymbol{f}_j = \boldsymbol{V}_j\boldsymbol{\gamma}_j$ der Vektor $\boldsymbol{\gamma}_j$ zerlegt in

$$\boldsymbol{\gamma}_j = \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j + \tilde{\boldsymbol{U}}_j\tilde{\boldsymbol{\gamma}}_j .$$

Residuen in STAR-Modellen mit metrischer Zielgröße

Gewöhnliche Residuen

$$\hat{\varepsilon}_i = y_i - \hat{\eta}_i^{strukt} = y_i - \hat{f}_1(v_{i1}) - \ldots - \hat{f}_q(v_{iq}) - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}.$$

Standardisierte Residuen

Das i-te standardisierte Residuum ist gegeben durch

$$r_i = \frac{y_i - \hat{\eta}_i^{strukt}}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

wobei h_{ii} das i-te Diagonalelement der Gesamt-Glättungsmatrix $\boldsymbol{H}$ ist.

Partielle Residuen

Die partiellen Residuen bezüglich v_j sind definiert als

$$\begin{aligned}\hat{\varepsilon}_{v_j,i} &= y_i - \hat{f}_1(v_{i1}) - \ldots - \hat{f}_{j-1}(v_{i,j-1}) - \hat{f}_{j+1}(v_{i,j+1}) - \ldots - \hat{f}_q(v_{iq}) - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} \\ &= y_i - \hat{\eta}_i^{strukt} + \hat{f}_j(v_{ij}).\end{aligned}$$

Durch geeignete Wahl der Designmatrizen $\tilde{\boldsymbol{X}}$ und $\tilde{\boldsymbol{U}}$ lässt es sich dann immer erreichen, dass $\boldsymbol{\beta}_j$ einem Vektor fester Effekte und $\tilde{\boldsymbol{\gamma}}_j \sim \mathrm{N}(\boldsymbol{0}, \tau_j^2\boldsymbol{I})$ einem Vektor von unabhängigen und identisch verteilten zufälligen Effekten entspricht. Analog lässt sich der Vektor der Funktionsauswertungen darstellen als

$$\boldsymbol{f}_j = \boldsymbol{V}_j(\tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j + \tilde{\boldsymbol{U}}_j\tilde{\boldsymbol{\gamma}}_j) = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{U}\tilde{\boldsymbol{\gamma}}_j.$$

Wendet man diese Zerlegung auf alle Modellkomponenten sukzessive an, so ergibt sich schließlich ein großes gemischtes Modell, dessen feste Effekte aus den ursprünglichen parametrischen Effekten $\boldsymbol{\beta}$ und den aus der Reparametrisierung entstehenden festen Effekten $\boldsymbol{\beta}_j$, $j = 1, \ldots, q$, besteht. Entsprechend ergeben sich die gesamten zufälligen Effekte aus den Vektoren $\tilde{\boldsymbol{\gamma}}_j$, $j = 1, \ldots, q$ mit gemeinsamer Verteilung $\mathrm{N}(\boldsymbol{0}, \mathrm{diag}(\tau_1^2\boldsymbol{I}, \ldots, \tau_q^2\boldsymbol{I}))$. Es handelt sich bei dem entstehenden gemischten Modell also speziell um ein sogenanntes Varianzkomponentenmodell. Zur Schätzung dieses Varianzkomponentenmodells (und insbesondere der zugehörigen Varianzen bzw. Glättungsparameter) kann nun wieder die in Kapitel 6 behandelte Methodik sowohl im Normalverteilungsfall als auch für generalisierte STAR-Modelle herangezogen werden. Dabei ist zu beachten, dass durch die Reparametrisierung in der Regel redundante Spalten in der Designmatrix der festen Effekte entstehen. Der Grund hierfür liegt in dem Identifizierbarkeitsproblem additiver Modelle: Jeder der nichtparametrischen Modellterme besitzt ein eigenes Niveau, dass in der Reparametrisierung einer eigenen Konstante entspricht. Nach der Reparametrisierung besitzt das Modell also zusätzlich zu der bereits in $\boldsymbol{\beta}$ enthaltenen Konstante weitere Konstanten, so dass die gemeinsame Designmatrix nicht mehr vollen Rang besitzt. Entsprechend müssen redundante Spalten aus dieser Designmatrix entfernt werden, um zu einem iden-

tifizierbaren Modell zu gelangen. Dies entspricht für die Konstanten im Wesentlichen der früher diskutierten Zentrierungsbedingung an die Funktionen f_j.

Für additive Modelle wird die auf gemischten Modellen beruhende Inferenz ausführlich in Ruppert et al. (2003) besprochen. Auf generalisierte STAR-Modelle gehen Fahrmeir et al. (2004) genauer ein, Kneib & Fahrmeir (2006) beschreiben die entsprechenden Erweiterungen für kategoriale Zielgrößen. Eine sehr ausführliche Darstellung findet man auch in Kneib (2005).

Modellwahl und Diagnose

Die oben gemachten Anmerkungen zur Modellwahl und Diagnose behalten auch bei Inferenz basierend auf gemischten Modellen ihre Gültigkeit.

Möchte man überprüfen, ob die flexible, nichtparametrische Modellierung der Effekte einzelner Kovariablen notwendig ist, so kann dies prinzipiell anhand statistischer Tests für den Varianzparameter geschehen. Soll beispielsweise ein P-Spline mit quadrierten zweiten Differenzen als Penalisierung gegen eine Gerade getestet werden, so entspricht dies dem Hypothesenpaar

$$H_0 : \tau^2 = 0 \quad \text{gegen} \quad H_1 : \tau^2 > 0.$$

Zur Durchführung des Tests bietet sich die Likelihood-Quotienten-Statistik an, die die Likelihoods unter H_0 und H_1 vergleicht. Es ergibt sich allerdings das Problem, dass der zu testende Parameter unter der Null-Hypothese auf dem Rand des Parameterraums liegt, was den Regularitätsbedingungen widerspricht, die bei der Herleitung der asymptotischen χ^2-Verteilung der LQ-Teststatistik verwendet werden. Dementsprechend ergibt sich für obiges Testproblem auch keine χ^2-Verteilung als Verteilung der Teststatistik. Die exakte Verteilung lässt sich bisher nur in relativ einfachen Spezialfällen mit normalverteilten Zielvariablen herleiten und hängt insbesondere von den betrachteten Daten ab. Weitere Informationen hierzu findet man etwa in Crainiceanu & Ruppert (2004) oder Crainiceanu, Ruppert, Claeskens & Wand (2005).

8.5.3 Bayesianische Inferenz mit MCMC

Normalverteilte Zielgrößen

Im Folgenden gehen wir zunächst vom einfachsten Fall normalverteilter Zielgrößen aus, d.h.

$$\boldsymbol{y} \,|\, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{\eta}^{strukt}, \sigma^2 \boldsymbol{I}).$$

Damit ist die Dichte des Beobachtungsmodells proportional zur Likelihood und gegeben durch

$$p(\boldsymbol{y} \,|\, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \left(\boldsymbol{y} - \boldsymbol{\eta}^{strukt}\right)' \left(\boldsymbol{y} - \boldsymbol{\eta}^{strukt}\right) \right).$$

Bayesianische Inferenz basiert auf der in Abschnitt 7.1.9 auf Seite 357 eingeführten Glattheits-Priori

$$p(\boldsymbol{\gamma}_j \,|\, \tau_j^2) \propto \left(\frac{1}{\tau_j^2}\right)^{\mathrm{rg}(\boldsymbol{K}_j)/2} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right) \tag{8.12}$$

für die Regressionskoeffizienten $\boldsymbol{\gamma}_j$ des j-ten Terms des strukturiert additiven Prädiktors. Für die Varianzparameter τ_j^2 wird eine inverse Gammaverteilung mit Hyperparametern a_j und b_j, d.h.

$$\tau_j^2 \sim \mathrm{IG}(a_j, b_j),$$

als Priori-Verteilung spezifiziert. Analog wird für die Varianz σ^2 des Störterms eine inverse Gammaverteilung mit Hyperparametern a_0 und b_0 definiert. In der Praxis werden häufig kleine Werte für a_j und b_j verwendet, etwa $a_j = 0.001$ und $b_j = 0.001$ oder $a_j = 0.0001, b_j = 0.0001$.

Damit kann die Dichte der Priori-Verteilung aller Parameter als Produkt der Dichten in (8.12) sowie der Dichten der inversen Gammaverteilungen der Varianzparameter geschrieben werden. Die Posteriori-Verteilung ist proportional zum Produkt aus Likelihood und Priori-Verteilung und ist damit gegeben durch

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \propto\; & \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{\eta}^{strukt}\right)'\left(\boldsymbol{y} - \boldsymbol{\eta}^{strukt}\right)\right) \\ & \prod_{j=1}^{q} \frac{1}{(\tau_j^2)^{\mathrm{rg}(\boldsymbol{K}_j)/2}} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right) \prod_{j=1}^{q} (\tau_j^2)^{-a_j-1} \exp\left(-\frac{b_j}{\tau_j^2}\right) \\ & (\sigma^2)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right), \end{aligned}$$

wobei $\boldsymbol{\theta}$ hier der Vektor aller im Modell vorkommenden Parameter (inklusive der Varianzparameter) ist. Im Folgenden beschreiben wir einen Gibbs-Sampler (vergleiche in Anhang B Abschnitt B.5.3) zur Ziehung von Zufallszahlen aus der Posteriori-Verteilung. Hierzu zerlegen wir den gesamten Parametervektor in die Blöcke $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_q^2, \sigma^2$. Zur Anwendung eines Gibbs-Samplers müssen die vollständig bedingten Dichten der Blöcke bekannte Verteilungen darstellen, aus denen leicht Zufallszahlen gezogen werden können. Ähnlich wie bereits in Abschnitt 7.1.9 ausgeführt, handelt es sich bei den bedingten Dichten der Regressionsparameter um multivariate Normalverteilungen und bei den bedingten Dichten der Varianzparameter um inverse Gammaverteilungen.

Genauer gilt

$$\boldsymbol{\gamma}_j \,|\, \cdot \;\sim \mathrm{N}(\boldsymbol{m}_j, \boldsymbol{\Sigma}_j)$$

mit Erwartungswert und Kovarianzmatrix

$$\begin{aligned} \boldsymbol{m}_j &= \mathrm{E}(\boldsymbol{\gamma}_j \,|\, \cdot) = \left(\frac{1}{\sigma^2}\boldsymbol{V}_j'\boldsymbol{V}_j + \frac{1}{\tau_j^2}\boldsymbol{K}_j\right)^{-1} \frac{1}{\sigma^2}\boldsymbol{V}_j'(\boldsymbol{y} - \boldsymbol{\eta}_{-j}^{strukt}) \\ \boldsymbol{\Sigma}_j &= \mathrm{Cov}(\boldsymbol{\gamma}_j \,|\, \cdot) = \left(\frac{1}{\sigma^2}\boldsymbol{V}_j'\boldsymbol{V}_j + \frac{1}{\tau_j^2}\boldsymbol{K}_j\right)^{-1}. \end{aligned}$$

Bei dem Vektor $\boldsymbol{\eta}_{-j}^{strukt} = \boldsymbol{\eta}^{strukt} - \boldsymbol{V}_j\boldsymbol{\gamma}_j$ handelt es sich um den Prädiktor $\boldsymbol{\eta}^{strukt}$ ohne den j-ten Term. Analog erhalten wir für den Vektor $\boldsymbol{\beta}$ eine multivariate Normalverteilung mit Erwartungswert und Kovarianzmatrix

$$\boldsymbol{m}_{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta} \,|\, \cdot) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'(\boldsymbol{y} - \tilde{\boldsymbol{\eta}}^{strukt})$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathrm{Cov}(\boldsymbol{\beta} \,|\, \cdot) = \frac{1}{\sigma^2}(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

wobei $\tilde{\boldsymbol{\eta}}^{strukt} = \boldsymbol{\eta}^{strukt} - \boldsymbol{X}\boldsymbol{\beta}$ jetzt der Prädiktor $\boldsymbol{\eta}^{strukt}$ ohne die linearen Effekte ist. Für die Varianzparameter erhalten wir

$$\tau_j^2 \,|\, \cdot \sim \mathrm{IG}(a_j + 0.5\mathrm{rg}(\boldsymbol{K}_j), b_j + 0.5\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j),$$
$$\sigma^2 \,|\, \cdot \sim \mathrm{IG}(a_0 + 0.5n, b_0 + 0.5(\boldsymbol{y} - \boldsymbol{\eta}^{strukt})'(\boldsymbol{y} - \boldsymbol{\eta}^{strukt})).$$

Damit lässt sich ohne weiteres ein Gibbs-Sampler initiieren, wobei darauf zu achten ist, dass insbesondere aus den zum Teil hochdimensionalen Normalverteilungen effizient Zufallszahlen gezogen werden. Hier kommen insbesondere numerische Verfahren für Bandmatrizen bzw. dünn besetzte Matrizen zum Einsatz. Für Details verweisen wir auf die Originalliteratur, insbesondere Rue & Held (2005), Lang & Brezger (2004), Rue (2001) und George & Liu (1981).

Latente normalverteilte Zielgrößen

Für eine Reihe von nicht normalverteilten Zielgrößen, insbesondere für kategoriale Variablen, können Repräsentationen der Modelle durch dahinter stehende latente Variablen zur Inferenz herangezogen werden. Wir illustrieren die Vorgehensweise für binäre Zielgrößen, speziell für binäre Probitmodelle. In Probitmodellen sind die binären Beobachtungen y_i bei gegebenen Kovariablen binomialverteilt, d.h. $y_i \sim \mathrm{B}(1, \pi_i)$ mit Wahrscheinlichkeit $\pi_i = \Phi(\eta_i^{strukt}) = P(y_i = 1)$. In Abschnitt 4.1.1 wurde ein Probitmodell äquivalent über die latenten Variablen

$$\tilde{y}_i = \eta_i^{strukt} + \varepsilon_i$$

mit normalverteilten Fehlern $\varepsilon_i \sim \mathrm{N}(0, 1)$ definiert. Der Zusammenhang mit den Beobachtungen ist gegeben durch $y_i = 1$ falls $\tilde{y}_i \geq 0$ und $y_i = 0$ falls $\tilde{y}_i < 0$, vergleiche auch nochmal Abschnitt 4.1.1 (Seite 193). Dieser Schwellenwertmechanismus besitzt nicht nur eine interessante Interpretation, sondern kann im Rahmen von MCMC-Inferenzmethoden auch zum Schätzen der Parameter des Modells benutzt werden. Dieser Ansatz wurde zum ersten Mal in einem Artikel von Albert & Chib (1993) beschrieben. Die grundlegende Idee ist, die latenten Variablen als weitere Parameter im Modell anzusehen und die Posteriori-Verteilung bezüglich des erweiterten Parameterraums zu betrachten. Zur Inferenz basierend auf MCMC müssen dann in jeder Iteration Zufallszahlen aus den vollständig bedingten Dichten der latenten Variablen $\tilde{y}_i$, $i = 1, \ldots, n$ gezogen werden. Die gezogenen latenten Variablen $\tilde{y}_i$ erfüllen dann die Rolle von metrischen Zielgrößen. Zufallszahlen für die Regressions- bzw. Varianzparameter können dann im Wesentlichen wie im letzten Abschnitt beschrieben gezogen werden. Im dortigen Gibbs-Sampler müssen lediglich in jeder Iteration des Algorithmus die Beobachtungen y_i durch die latenten Variablen $\tilde{y}_i$ ersetzt werden. Auf den ersten Blick scheint die Idee einen entscheidenden Haken zu besitzen, schließlich müssen in jeder Iteration des Algorithmus zusätzlich n latente Variablen $\tilde{y}_i$ gezogen werden. Der Zusatzaufwand ist aber selbst für großen Stichprobenumfang praktisch vernachlässigbar. Bei den vollständig bedingten Verteilungen der latenten Variablen $\tilde{y}_i$ handelt es sich nämlich um *trunkierte Normalverteilungen* aus denen leicht und schnell Zufallszahlen gezogen werden können. Konkret ist $\tilde{y}_i \,|\, \cdot \sim \mathrm{TN}_{0,\infty}(\eta_i^{strukt}, 1)$ falls $y_i = 1$

und $\tilde{y}_i \,|\, \cdot \sim \mathrm{TN}_{-\infty,0}(\eta_i^{strukt}, 1)$ falls $y_i = 0$, vergleiche zur trunkierten Normalverteilung auch Definition B.1 (Seite 459) in Anhang B. Die einfachste Variante wie beispielsweise aus der $\mathrm{TN}_{0,\infty}(\eta_i^{strukt}, 1)$ Verteilung gezogen werden kann, ist solange Zufallszahlen aus einer $\mathrm{N}(\eta_i^{strukt}, 1)$ Verteilung zu ziehen bis die gezogene Zufallszahl tatsächlich positiv ist. Diese einfache Vorgehensweise kann allerdings durch effizientere Algorithmen noch deutlich verbessert werden, vergleiche hierzu die Artikel von Geweke (1991) und Robert (1995).

Die für binäre Probit-Modelle skizzierte Vorgehensweise lässt sich auch für mehrkategoriale Probit- und auch Logit-Modelle anwenden, vergleiche hierzu Holmes & Held (2006).

Nicht-normalverteilte Zielgrößen

Bei Poisson- oder gammaverteilten Zielvariablen kann die im letzten Abschnitt beschriebene Repräsentation durch latente Variablen nicht mehr zur Inferenz herangezogen werden. In diesem Fall basiert die MCMC-Inferenz auf MH-Algorithmen mit den in Kapitel 4.6 beschriebenen Normalverteilungsdichten als Vorschlagsdichten, wobei geeignete Adaptionen für die nichtparametrisch modellierten Terme vorzunehmen sind. Eine genaue Darstellung der Vorgehensweise findet man in Brezger & Lang (2006) und Rue & Held (2005)

Modellwahl und Diagnose

Zum Vergleich konkurrierender Modelle kann bei MCMC-Inferenz das DIC (vergleiche Abschnitt B.5.4) in Anhang B verwendet werden. Der Nachteil des DIC ist allerdings, dass es sich hierbei um eine zufällige Größe handelt, die in jedem MCMC-Lauf einen veränderten Wert aufweist. Sind bei zwei Modellen die DIC sehr unterschiedlich sollten keine Probleme beim Modellvergleich auftreten. Bei ähnlichen DIC-Werten ist hingegen Vorsicht geboten. In diesem Fall kann ein Vergleich nur anhand mehrerer MCMC-Läufe (mindestens 20) und anschließender Ermittlung der durchschnittlichen DIC-Werte erfolgen. Diese Vorgehensweise ist dann gegebenenfalls sehr zeitintensiv.

Zur Modelldiagnose können wieder, wie oben bereits beschrieben, die verschiedenen Residuenarten verwendet werden. Bei Bayes-Inferenz basierend auf MCMC-Methoden sind zusätzlich einige Besonderheiten zu beachten. Insbesondere sollte bei jeder Analyse die Konvergenz der verwendeten MCMC-Algorithmen überprüft werden. Dazu werden in der Literatur eine Fülle von Diagnosewerkzeugen vorgeschlagen, deren Verwendbarkeit aber auf verhältnismäßig einfache Modelle mit wenigen Parametern beschränkt ist. Einen Überblick geben die Artikel Cowles & Carlin (1996) und Mengersen, Robert & Guihenneuc-Jouyaux (1999). In jedem Fall, d.h. auch bei komplexen Modellen mit vielen Parametern, können die Samplingpfade und Autokorrelationsfunktionen der gezogenen Zufallszahlen untersucht werden. Hier werden Auffälligkeiten, insbesondere Nichtkonvergenz, niedrige Akzeptanzraten und hohe Autokorrelationen deutlich sichtbar, vergleiche hierzu auch Beispiel B.8 in Anhang B. Bei Verwendung der in diesem Kapitel beschriebenen MCMC-Algorithmen sollten allerdings Probleme nur dann auftreten, wenn angesichts der vorhandenen Daten zu komplexe Modelle geschätzt werden.

Als weiterer Modelldiagnoseschritt empfiehlt sich bei Bayes-Modellen die Durchführung einer sogenannten *Sensitivitätsanalyse*. Darunter versteht man die Untersuchung der Auswirkungen von kleinen Änderungen in den Priori-Verteilungen auf die Schätzergebnisse.

Abb. 8.12. *Links: Geschätzter nichtparametrischer Effekt der Anzahl der Länder für die Patentschutz gelten soll. Rechts: Geschätzter Effekt von aland basierend auf vier unterschiedlichen Kombinationen der Hyperparameter a und b der Priori-Verteilung für den Varianzparameter.*

Bei den in diesem Kapitel diskutierten STAR-Modellen sollte vor allem die Sensitivität der Wahl der Hyperparameter a_j und b_j der inversen Gamma Priori-Verteilungen für die Varianzparameter τ_j^2 untersucht werden. Je nach Wahl der Hyperparameter können sich die Schätzergebnisse unter Umständen erheblich unterscheiden. Das ist häufig dann der Fall, wenn zumindest in Teilbereichen des Wertebereichs wenig Information, d.h. vor allem wenig Beobachtungen, vorliegt. In den meisten Fällen sind die Ergebnisse allerdings sehr stabil. Die mögliche Abhängigkeit von der Wahl der Hyperparameter ist kein Nachteil der vollen Bayes-Inferenz. Vielmehr ermöglicht es eine sorgfältige Sensitivitätsanalyse, die Stabilität des gewählten Modells zu untersuchen.

Zuletzt sei noch angemerkt, dass zu den in Abschnitt 7.1.10 beschriebenen, Bayesianischen adaptiven Verfahren einige Erweiterungen existieren, die eine simultane Schätzung und Modellwahl erlauben. So können beispielsweise mit Hilfe der beim Model Averaging betrachteten Dummy-Variablen einzelne Effekte adaptiv aus dem Modell ausgeschlossen bzw. mit aufgenommen werden. Entsprechende Verfahren sind in den bereits in Abschnitt 7.1.10 genannten Referenzen detaillierter beschrieben. Da jedoch momentan keine Standardsoftware existiert, in der die entsprechende Methodik umgesetzt ist, verzichten wir auf eine genauere Beschreibung.

Beispiel 8.7 Patenterteilung

In den Beispielen 2.8 (Seite 33), 4.1 (Seite 194) und 4.4 (Seite 206) wurde der Einfluss der Kovariable *aland* auf die binäre Zielvariable Einspruch durch ein kubisches Polynom modelliert. Zum Vergleich schätzen wir nun den Effekt $f(aland)$ nichtparametrisch. Dies führt zum additiven Logit-Modell

$$P(einspruch_i = 1) = \frac{\exp(\eta_i^{add})}{1 + \exp(\eta_i^{add})}$$

mit dem additiven Prädiktor

$$\eta_i^{add} = f(aland_i) + \beta_0 + \beta_1\, jahr_i + \beta_2\, azit_i + \beta_3\, ansp_i + \beta_4\, uszw_i + \beta_5\, patus_i + \beta_6 patdsg_i.$$

Die Funktion $f(aland)$ wird durch einen kubischen Bayesianischen P-Spline modelliert und mit MCMC geschätzt. Abbildung 8.12 (links) zeigt den geschätzten Posteriori-Erwartungswert zusammen mit (punktweisen) 80%- und 95% Konfidenzbändern. Offensichtlich weicht die Funktion nicht deutlich von einer Geraden ab. Für das DIC

erhalten wir den Wert DIC = 2993.3. Bei linearer bzw. polynomialer Modellierung des Effekts von *aland* erhalten wir DIC = 2992.9 bzw. DIC = 2994.3, so dass eine lineare Modellierung offenbar ausreichend ist. Damit wird das Testergebnis aus Beispiel 4.4 bestätigt.

Die rechte Grafik von Abbildung 8.12 zeigt noch das Ergebnis einer Sensitivitätsanalyse, wobei die Hyperparameter a und b der Priori für den Varianzparameter τ^2 variiert wurden. Verwendet wurden die Kombinationen $a = 0.001, b = 0.001$, $a = 0.0001, b = 0.0001$, $a = 0.00001, b = 0.00001$ und $a = 1, b = 0.001$. Offenbar unterscheiden sich die Schätzergebnisse kaum. Auch zeigt sich nochmal deutlich, dass die Modellierung eines linearen Effekts für *aland* ausreichend ist.

△

8.5.4 Software-Hinweise

Während die in den ersten sechs Kapiteln besprochenen parametrischen Regressionsmodelle in praktisch allen gängigen Statistikprogrammen implementiert sind, findet sich gut benutzbare Software für semiparametrische Modelle nicht in allen Programmpaketen. Wir geben daher an dieser Stelle einige Hinweise zur vorhandenen Software, vergleiche auch die Homepage zum Buch, die detailliertere Hinweise enthält.

Generalisierte additive Modelle inklusive Erweiterungen für zweidimensionale Glättung, mit variierenden Koeffizienten und zufälligen Effekten können mit dem R Paket `mgcv` geschätzt werden. Die Schätzung räumlicher Effekte ist unter Verwendung zweidimensionaler Glätter möglich, wird aber nicht explizit unterstützt. Modelle mit (mehr-)kategorialen Zielvariablen sind nicht implementiert. Eine ausführliche Beschreibung der Software (und Methodik) findet man in Wood (2006). Die Schätzung erfolgt im Rahmen eines penalisierten KQ- bzw. Likelihood-Ansatzes. Die Glättungsparameter werden durch Minimierung des GCV-Kriteriums bestimmt.

Sämtliche Varianten von STAR-Modellen sind im Programm BayesX enthalten, wobei zwei alternative Schätzansätze implementiert sind. Die Objektklasse `remlreg` erlaubt die Schätzung basierend auf gemischten Modellen. Alternativ kann mit Hilfe der Objektklasse `bayesreg` ein voller Bayes-Ansatz basierend auf MCMC-Methoden durchgeführt werden. Beide Varianten liefern in der Regel sehr ähnliche Ergebnisse, die Schätzungen basierend auf Methoden für gemischte Modelle sind erfahrungsgemäß jedoch etwas glatter. Hinsichtlich der benötigten Rechenzeit könnte man vermuten, dass die simulationsbasierten Schätzungen deutlich länger dauern. Dies ist aber nicht generell der Fall. Vor allem in Modellen mit sehr vielen Parametern, etwa wenn räumliche Effekte geschätzt werden, ist der simulationsbasierte Ansatz häufig schneller. Der Grund hierfür ist die durch die Zerlegung des Gesamt-Parameterraums in kleinere Blöcke erzielte Komplexitätsreduktion. Das Programm BayesX ist im Internet frei verfügbar unter:

`http://www.stat.uni-muenchen.de/ bayesx`

Die Homepage enthält auch einige Tutorien, die vor allem bei Verwendung des vollen Bayes-Ansatzes und bei geringer Erfahrung mit MCMC-Methoden zu empfehlen sind.

Schließlich sind additive Modelle und einige Erweiterungen basierend auf lokalen Glättern in den (kommerziellen) Statistikprogrammen S-Plus und Xplore implementiert.

8.6 Fallstudie: Unterernährung in Sambia

Dieser letzte Abschnitt illustriert anhand der Daten zur Unterernährung in Sambia wie die in diesem Buch vorgestellten Methoden in der Praxis verwendet werden können. Im nächsten Unterabschnitt geben wir zunächst allgemeine Hinweise wie Regressionsfragestellungen bearbeitet werden können. Anschließend wenden wir diese allgemeine Vorgehensweise auf die Sambia-Daten an.

8.6.1 Hinweise zur grundsätzlichen Vorgehensweise

Die Bearbeitung von Regressionsfragestellungen lässt sich im Wesentlichen in die folgenden Schritte untergliedern:

Deskriptive Analyse der Rohdaten

Der erste wichtige Arbeitsschritt bei *jeder* statistischen Auswertung ist die sorgfältige Analyse der Rohdaten. Dazu verschafft man sich zunächst einen Überblick über die (univariaten) Verteilungen der Variablen im Datensatz. Hilfsmittel sind die grafischen Darstellungsmöglichkeiten und statistischen Kennzahlen, die in Kapitel 1.2.1 skizziert wurden. Mit dieser ersten Analyse werden folgende Ziele verfolgt:

- *Beschreibung der Verteilungen der Variablen.*
- *Auffinden von Datenanomalien.* Hierzu gehören Fehlkodierungen, unplausible und extreme Werte. Bei Regressionsanalysen kommt der Analyse des Wertebereichs der Kovariablen eine besondere Bedeutung zu. Bei einigen metrischen Kovariablen sind die Daten an den Rändern des Wertebereichs sehr dünn, so dass aufgrund der geringen Information für diese Bereiche nur sehr unsichere Aussagen über den Einfluss der Kovariable auf die Zielgröße getroffen werden können. Unter Umständen ist es sogar ratsam einige extreme Beobachtungen von der weiteren Analyse auszuschließen. Ein typisches Beispiel haben wir bei der Analyse der Patentdaten in Beispiel 2.8 (Seite 33) gesehen. Dort wurde für die Kovariable *ansp* eine extrem linkssteile Verteilung beobachtet. Der überwiegende Teil der Beobachtungen ist auf den Bereich 0–60 konzentriert, einige wenige Beobachtungen reichen aber bis zum maximalen Wert 355. Da für den Bereich 61–355 praktisch keine Information vorhanden ist, wurden die Beobachtungen mit Werten $ansp > 60$ von der weiteren Analyse ausgeschlossen.

  Das Problem ungleichmäßig verteilter Werte kann auch bei kategorialen Kovariablen auftreten. Häufig sind einige Kategorien sehr dünn besetzt. Im Extremfall kann es passieren, dass eine Variable als Kovariable bei der Regressionsanalyse nicht sinnvoll verwendet werden kann. Dies ist beispielsweise bei binären Kovariablen dann der Fall, wenn eine der beiden Kategorien nicht oder nur sehr dünn besetzt ist. Dann kann aufgrund der fehlenden Variation der Einfluss auf die Zielvariable nicht untersucht werden. In vielen Fällen kann das Problem dünn besetzter Kategorien durch Zusammenfassung einzelner Kategorien behoben werden, sofern dies sinnvoll möglich ist.

Datenaufbereitung

Nach der ersten Analyse der Rohdaten sollten die Daten für die spätere Regressionsanalyse aufbereitet werden. Das bedeutet insbesondere, dass einige extreme Beobachtungen, möglicherweise sogar einige Kovariablen, von der Analyse ausgeschlossen werden.

Darüber hinaus müssen eventuell vorhandene Datenfehler korrigiert und für die Analysen benötigte neue Kovariablen erzeugt werden (beispielsweise orthogonale Polynome oder andere Transformationen).

Schließlich ist zu beachten, dass die Rohdaten nicht im Zuge der Datenaufbereitung (versehentlich) überschrieben werden. Stattdessen sollte ein neuer aufbereiteter Datensatz erzeugt und die zur Aufbereitung nötigen Schritte sorgfältig dokumentiert werden.

Grafische zweidimensionale Zusammenhangsanalyse

Bei metrischen Zielgrößen geben zweidimensionale grafische Zusammenhangsanalysen zwischen Zielgröße und jeweils einer Kovariable wichtige Hinweise über die Art und Stärke des Zusammenhangs. Bei der Interpretation ist jedoch stets zu beachten, dass dabei alle übrigen Kovariablen nicht beachtet werden, so dass sich die beobachteten Effekte in simultanen Analysen durchaus stärker von den rein bivariaten Zusammenhängen unterscheiden können. Als Hilfsmittel können die in Kapitel 1.2.2 beschriebenen grafischen Darstellungsmöglichkeiten (Streudiagramme, Boxplots) verwendet werden. Ein besonderes Augenmerk sollte auf mögliche Ausreißer (vergleiche hierzu Kapitel 3.6.4) gelegt werden. Häufig – jedoch nicht immer – können solche Beobachtungen bereits anhand der Streudiagramme identifiziert werden.

Im Fall diskreter Zielgrößen, insbesondere kategorialer Variablen, können grafische Zusammenhangsanalysen jedoch nicht oder nur sehr eingeschränkt eingesetzt werden.

Schätzung erster Arbeitsmodelle

Dazu sind vor allem die in diesem und im letzten Kapitel besprochenen semiparametrischen Regressionsmodelle nützlich. Nichtlineare Kovariableneffekte, räumliche und clusterspezifische Heterogenität können mit diesen Modellen schnell und automatisiert entdeckt werden. In der Regel wird man sich zunächst auf reine Haupteffekt-Modelle beschränken und Interaktionen erst in einem zweiten Schritt berücksichtigen. Bei der Formulierung der Modelle sollten, neben den Ergebnissen der uni- und bivariaten Analysen, vorhandenes Fachwissen und die Erkenntnisse früherer oder ähnlicher Untersuchungen einfließen. In keinem Fall ist es ratsam, sämtliche zur Verfügung stehenden Variablen unreflektiert in die Regressionsmodelle zu integrieren.

In der Regel stehen mehrere konkurrierende Modellspezifikationen zur Auswahl. Beispielsweise können Modelle mit und ohne räumliche Komponente in Frage kommen, oder der Einfluss einer Variablen könnte unsicher sein, so dass Modelle mit und ohne die betroffene Kovariable geschätzt werden. Die konkurrierenden Modelle können dann mit einem Gütekriterium wie beispielsweise dem AIC verglichen werden. Offensichtlich weniger taugliche Spezifikationen können von der weiteren Analyse ausgeschlossen werden.

Das Resultat dieses Arbeitsschrittes sind in der Regel einige wenige, hinsichtlich des verwendeten Gütekriteriums nahezu gleichwertige Modellspezifikationen.

Modelldiagnose und Verfeinerung der Arbeitsmodelle

Hauptziel dieses Arbeitsschrittes ist die Verbesserung bzw. Verfeinerung der im vorangegangenen Abschnitt gefundenen Modelle. Dazu können folgende Hilfsmittel verwendet werden:

Überführung in ein einfaches parametrisches Modell

In einigen Fällen handelt es sich bei den durch die semiparametrische Analyse gefundenen Kovariableneffekten um annähernd lineare oder zumindest einfache nichtlineare Funktionen. Hier bietet sich zumindest teilweise eine Überführung in ein parametrisches Modell an.

Analyse der räumlichen bzw. clusterspezifischen Effekte

Diese Effekte können als Surrogat für nicht beobachtete bzw. nicht im Modell enthaltene Kovariableneffekte interpretiert werden. In einigen Fällen liefert eine sorgfältige Analyse der Effekte Hinweise auf Kovariablen bzw. neue Aspekte, die bisher nicht in Erwägung gezogen wurden. Beispielsweise haben wir bei der Analyse der Schadensfälle bei Kfz-Versicherungen in Beispiel 2.12 (Seite 47) festgestellt, dass die Schadenshäufigkeit vor allem in den Ballungsgebieten im Vergleich zu den eher ländlichen Gegenden erhöht ist. Hier könnte man also unter Umständen den nichtparametrischen räumlichen Effekt durch eine binäre Dummy-Variable ersetzen, die Ballungsräume identifiziert. Das ideale Endziel ist in jedem Fall die Ersetzung oder zumindest deutliche Verringerung der räumlichen bzw. clusterspezifischen Effekte durch Hinzunahme gut interpretierbarer Kovariablen, die durch die anfänglich geschätzten Effekte motiviert sind.

Wie Beispiel 8.6 (Seite 416) zeigt, kann die Hinzunahme räumlicher oder clusterspezifischer Effekte dazu führen, dass Identifikationsprobleme entstehen. Insbesondere können die Effekte anderer Kovariablen im räumlichen Effekt aufgehen und somit verdeckt werden. Es ist daher immer ratsam, den Unterschied in den Schätzergebnissen mit und ohne räumliche bzw. clusterspezifische Komponente sorgfältig zu untersuchen.

Modelldiagnose

Instrumente zur Modelldiagnose dienen dazu, mögliche alternative Modellformulierungen zu finden. Im Fall von parametrischen linearen Modellen existiert ein umfangreiches Instrumentarium zur Modelldiagnose, vergleiche hierzu Abschnitt 3.6.4. Im Fall von semiparametrischen Modellen können die Instrumente, insbesondere die Analyse der Residuen bzw. partiellen Residuen, zum Teil einfach übertragen werden. Einige Hilfsmittel zur Ausreißeranalyse sind jedoch nicht oder nur eingeschränkt methodisch entwickelt oder in Softwarepaketen umgesetzt. In Modellen mit diskreten Zielgrößen sind auch Residuenplots nur eingeschränkt verwendbar.

Bei metrischen Zielgrößen geben Residuenplots Hinweise auf unberücksichtigte bzw. nicht korrekt spezifizierte Kovariableneinflüsse, Ausreißer und heteroskedastische bzw. korrelierte Fehler. Je nach Befund kommen verschiedene alternative Modelle in Frage:

- *Fehlspezifizierte Kovariableneinflüsse.* Bei Hinweisen auf nicht korrekt spezifizierte Kovariableneinflüsse kommen prinzipiell drei Varianten in Frage:
  - *Nichtlinearer Einfluss.* Ein zunächst linear modellierter Einfluss einer metrischen Kovariablen erweist sich als nichtlinear. In diesem Fall muss eine flexiblere Modellierung gewählt werden.
  - *Nicht berücksichtigte Kovariablen.* Zunächst nicht berücksichtigte aber vorhandene Kovariablen können jederzeit ins Modell aufgenommen werden. Schwieriger ist

die Situation bei nicht beobachtbaren Einflussgrößen. Hier können räumliche oder clusterspezifische Effekte zur Modellierung der unbeobachteten Heterogenität herangezogen werden.

  – *Nicht berücksichtigte Interaktionen.* Komplexe Interaktionen können grundsätzlich mit den in Abschnitt 3.1.4 diskutierten parametrischen Modellierungen und insbesondere mit den in Abschnitt 8.3 beschriebenen nichtparametrischen Alternativen geschätzt werden. Interaktionen sollten aber möglichst sparsam verwendet werden, da sie die Modellkomplexität deutlich erhöhen. Hier sollte man sich vor allem an der Fragestellung und dem theoretischen Vorwissen orientieren.

- *Heteroskedastische Fehler.* Hier kommen grundsätzlich die in Abschnitt 3.4.3 (Seite 132) diskutierten Alternativen in Frage, d.h. die Transformation der Zielgröße und eine zweistufige Schätzung. In parametrischen linearen Modellen kann auch die von White vorgeschlagene korrigierte Kovarianzmatrix verwendet werden.
- *Korrelierte Fehler.* Eine in den Residuen vorhandene Autokorrelation deutet wie in den Abschnitten 3.1.3 und 3.4.4 diskutiert auf eine vorhandene Fehlspezifikation hin. Ziel sollte also sein, die Fehlspezifikation soweit wie möglich zu beseitigen. Die Verwendung von semiparametrischen Modellen mit gleichzeitig korrelierten Störtermen befindet sich erst in der Entwicklung. Darüber hinaus sind Identifikationsprobleme zu erwarten, etwa wenn gleichzeitig ein nichtparametrischer Zeittrend und korrelierte Fehler modelliert werden.
- *Ausreißer.* In diesem Fall gelten die im Zusammenhang mit dem linearen Regressionsmodell diskutierten Hinweise, vergleiche in Abschnitt 3.6.4 Seite 175. Von den dort beschriebenen Maßnahmen ist die Verwendung robuster Verfahren nicht oder nur eingeschränkt möglich, da hierzu in semiparametrischen Modellen noch wenig Methodik und vor allem keine gut benutzbare Software existiert. Die wichtigste Maßnahme dürfte neben der Suche nach Datenfehlern und Erklärungen für Ausreißer die alternative Schätzung ohne Berücksichtigung der auffälligen Punkte und die anschließende Beschreibung der Unterschiede sein.

Darstellung der Ergebnisse

Das Ergebnis einer Regressionsanalyse ist in den seltensten Fällen ein alle anderen überragendes „bestes“ Modell. In der Regel sind einige wenige Modelle nahezu gleichwertig. Diese werden sich jeweils nur in Teilaspekten unterscheiden, die oft für die Fragestellung unwesentlich sind. Bei der Darstellung und Beschreibung der Ergebnisse sollte zum Einen der gemeinsame Kern aller Modelle hervorgehoben werden und zum Anderen sollten die Unterschiede zwischen den konkurrierenden Modellen dargestellt werden. Hinsichtlich dieser Unterschiede besteht dann Unsicherheit was die Schlussfolgerungen der Analyse anbelangt.

Zuletzt weisen wir darauf hin, dass die Analysen auch für nicht an der Untersuchung beteiligte Personen gut nachvollziehbar und reproduzierbar sein sollten. Hierzu sind ausreichend kommentierte Fassungen der verwendeten Auswertungsprogramme inklusive der verwendeten Befehle notwendig.

Abb. 8.13. *Streudiagramm zwischen dem Alter des Kindes und der Stilldauer. Beobachtungen unterhalb der Winkelhalbierenden sind unplausibel.*

8.6.2 Deskriptive Analysen

Im Zuge der Vorstellung des Datensatzes zur Unterernährung in Kapitel 1 wurden bereits ausführliche deskriptive und grafische Analysen für die Variablen im Datensatz durchgeführt, vergleiche Tabelle 1.3 (Seite 7) sowie Abbildung 1.4 (Seite 14). Die Ergebnisse können wie folgt interpretiert werden:

- *Zielgröße Z-Score:* Der Z-Score besitzt eine annähernd symmetrische Verteilung mit Werten zwischen -600 und 503, wobei der Großteil der Beobachtungen im Bereich -600 bis 300 liegt. Der durchschnittliche Z-Score liegt bei -171.19, d.h. die Kinder in Sambia sind im Durchschnitt deutlich kleiner als Kinder der selben Altersklasse in der Referenzpopulation. Unplausible oder extreme Werte sind nicht zu erkennen.
- *Alter des Kindes und Stilldauer:* Die beiden Variablen Alter des Kindes und Stilldauer hängen zusammen und müssen gemeinsam betrachtet werden, da die Stilldauer nicht länger als das Alter des Kindes sein kann.

  Eine zunächst univariate Betrachtung zeigt, dass das Alter nahezu gleichverteilt ist, ältere Kinder sind im Datensatz nur geringfügig seltener vertreten als jüngere. Die ungewöhnlich raue Gestalt des Histogramms scheint eher ein Artefakt zu sein, wie eine genaue Betrachtung der Häufigkeitstabelle bzw. wiederholte Berechnungen mit modifizierten Klassenbreiten zeigen.

  Die Stilldauer schwankt im Datensatz zwischen 0 und 46 Monaten. Die Verteilung ist deutlich bimodal mit Modi bei 0 und circa 18 Monaten. Fast 25 Prozent der Kinder wurden nie gestillt. Eine Stilldauer größer als 30 Monate ist sehr selten (nur 24 Beobachtungen).

  Zur Plausibilitätsprüfung zeichnen wir in Abbildung 8.13 zusätzlich ein Streudiagramm zwischen dem Alter des Kindes und der Stilldauer. Da die Stilldauer nicht länger als das Alter sein darf, sind alle Werte unterhalb der zusätzlich eingezeichneten Winkelhalbierenden nicht plausibel und werden aus diesem Grund von der weiteren Analyse ausgeschlossen. Da eine Stilldauer länger als 30 Monate extrem selten ist und außerdem im folgenden zweidimensionale Glätter bei der Modellierung des Einflusses von Alter und Stilldauer verwendet werden, schließen wir auch Beobachtungen mit $k_still > 30$ aus.

- *Alter der Mutter bei der Geburt:* Das Alter der Mutter schwankt zwischen 13 und 48 Jahren mit einem Mittelwert von 26.40 Jahren. Die Verteilung ist deutlich linkssteil. Einige sehr kleine und große Werte sind ungewöhnlich, jedoch nicht unmöglich.
- *Größe der Mutter:* Die Mütter haben eine Körpergröße zwischen 134 und 185 cm mit einem Mittelwert von 158.06 cm. Die Verteilung ist nahezu symmetrisch. Auffällige Werte liegen nicht vor.
- *BMI der Mutter:* Der Body-Mass-Index ist ein bekanntes (wenn auch teilweise umstrittenes) Maß zur Beurteilung der Ernährungssituation bei Erwachsenen. Bei einem BMI zwischen 18.5 und 25 geht man von Normalgewicht aus. Ein BMI kleiner als 18.5 bedeutet Untergewicht. Von (leichtem) Übergewicht spricht man ab einem BMI von mehr als 25. Ein BMI größer als 30 bedeutet starkes Übergewicht bzw. Adipositas. Im Datensatz schwankt der BMI zwischen 13.15 und 39.29 mit einem Mittelwert von 21.99. Der Großteil der Mütter ist normalgewichtig oder leicht übergewichtig. Ein nicht zu vernachlässigender Teil ist jedoch untergewichtig (knapp 10 Prozent), ein kleinerer Teil auch stark übergewichtig (circa 2.6 Prozent). Unplausible oder besonders auffällige Werte lassen sich nicht finden.
- *Bildung und Erwerbsstatus der Mutter:* Die überwiegende Mehrheit der Mütter (circa 80 Prozent) besitzt keine abgeschlossene Schulausbildung. Circa 17 Prozent besitzen einen Volksschulabschluss und lediglich 1.75 Prozent einen darüber hinaus gehenden Abschluss. Etwas mehr als die Hälfte der Mütter geht einer Erwerbsarbeit nach, circa 45 Prozent der Mütter sind nicht erwerbstätig. In beiden Variablen sind alle Kategorien ausreichend besetzt, so dass keine Kategorien zusammengefasst werden müssen oder gar eine der Variablen ausgeschlossen werden muss.

Ebenfalls in Kapitel 1 wurden erste grafische Zusammenhangsanalysen durchgeführt, vergleiche die Abbildungen 1.8 (Seite 17) und 1.9 (Seite 18). Wir stellen fest:

- *Alter des Kindes:* Der durchschnittliche Z-Score nimmt zunächst annähernd linear ab und bleibt dann ab einem Alter von circa 15-20 Monaten annähernd konstant. Die beobachtete Form des Zusammenhangs wird ähnlich in der entsprechenden Literatur zur Unterernährung postuliert. Man nimmt an, dass der Ernährungszustand in den ersten 6 Monaten unmittelbar nach der Geburt am besten ist, da das Kind durch die Ernährung im Mutterleib bzw. nach der Geburt durch die Muttermilch gut versorgt ist. Anschließend erwartet man eine Verschlechterung des Ernährungszustands, da das Kind zunehmend nicht ausreichender, qualitativ minderwertiger und unhygienischer Nahrung ausgesetzt ist. Um das zweite Lebensjahr herum wird dann eine Stabilisierung auf allerdings sehr niedrigem Niveau erwartet.
- *Stilldauer:* Die WHO empfiehlt neugeborene Kinder mindestens ein Jahr voll (d.h. ausschließlich) zu stillen, da Muttermilch in diesem Alter die beste Nahrung darstellt. Die Literatur zur Unterernährung geht daher generell von einem positiven Effekt des Stillens auf den Ernährungszustand aus. Werden Kinder deutlich länger als ein Jahr voll gestillt, nimmt man jedoch an, dass sich der Effekt verringert oder gar negativ wird. Die lange Stilldauer gilt dann als Indikator für eine unzureichende Versorgungsmöglichkeit mit fester Nahrung.

  Im Streudiagramm ist ein Zusammenhang mit der Stilldauer kaum zu erkennen, eventuell nimmt der durchschnittliche Z-Score mit wachsender Stilldauer geringfügig linear ab, was im Widerspruch zur Literatur stehen würde. Bei der Beurteilung des Einflusses der Stilldauer ist jedoch Vorsicht geboten, da hier eine Interaktion mit dem Alter des

Kindes zu erwarten ist. Der Effekt einer Stilldauer von einem Monat etwa dürfte bei einem ein Monat alten Kind ein anderer sein als bei einem bereits drei Jahre alten Kind. Im ersten Fall ist die einmonatige Stilldauer völlig normal. Im zweiten Fall wurde das Kind sehr früh abgestillt. Wir werden später den Effekt des Alters und der Stilldauer gemeinsam durch eine zweidimensionale Funktion modellieren.

- *BMI der Mutter:* Die Literatur geht von einem invers U-förmigen Einfluss des Ernährungszustands der Mutter aus. Kinder untergewichtiger und stark übergewichtiger Mütter sind schlechter ernährt als Kinder mit normalgewichtigen Müttern. Im Streudiagramm ist ein weitgehend linearer Zusammenhang mit dem Z-Score zu erkennen, d.h. der durchschnittliche Z-Score steigt mit wachsendem BMI linear an. Ab einem BMI von 30–35, d.h. bei stark übergewichtigen Müttern, scheint der durchschnittliche Effekt tatsächlich wieder abzusinken, so dass die postulierte inverse U-Form zumindest näherungsweise gegeben scheint.
- *Alter bei der Geburt:* Das Alter der Mutter bei der Geburt scheint einen schwach linearen Einfluss zu besitzen, d.h. je älter die Mutter bei der Geburt, desto besser ist im Durchschnitt der Ernährungszustand. Tatsächlich geht die Literatur davon aus, dass ältere Mütter besser für ihr Kind sorgen können. Diese Frauen haben eine höhere Chance auf eine abgeschlossene Berufsausbildung und besitzen in der Regel mehr Berufserfahrung, so dass ein höheres Einkommen zu erwarten ist.
- *Größe der Mutter:* Es ist bekannt, dass die Größe der Eltern einen Einfluss auf die Größe der Kinder besitzt. Da die altersstandardisierte Körpergröße als Maß für den Ernährungszustand verwendet wird, ist es nicht verwunderlich, dass im Streudiagramm ein linear steigender Effekt der Größe der Mutter zu sehen ist.
- *Bildungsstand und Erwerbsstatus:* Wie nicht anders zu erwarten scheint ein höherer Bildungsstand der Mutter mit besser ernährten Kindern einher zu gehen. Ein Zusammenhang mit dem Erwerbsstatus ist nicht erkennbar (Grafik nicht abgedruckt).
- *Geschlecht:* Geschlechtsspezifische Unterschiede sind nicht erkennbar (Grafik nicht abgedruckt).

8.6.3 Modellierungsvarianten

Basierend auf den im letzten Abschnitt beschriebenen, in der Literatur zur Unterernährung formulierten Hypothesen zum Einfluss der einzelnen erklärenden Variablen und den Ergebnissen der deskriptiven Analysen wurden für die einzelnen Kovariablen die nachfolgend skizzierten Modellierungsvarianten untersucht. Um später die Ergebnisse kompakter darstellen zu können, unterscheiden wir verschiedene Aspekte der Modellierung durch entsprechende Abkürzungen.

- *Räumlicher Effekt:* Unter Umständen vorhandene räumliche Heterogenität wurde entweder durch ein Markov-Zufallsfeld (Modellierungsvariante M), unabhängige zufällige Effekte (Z) oder parametrisch unter Verwendung von Dummy-Variablen für die Variable *region* (P) modelliert.
- *Alters des Kindes und Stilldauer:* Wie bereits erwähnt, hängt der Effekt der Stilldauer vom Alter des Kindes ab. Daher wurde der Effekt beider Variablen durch eine zweidimensionale Funktion basierend auf dem Krigingansatz geschätzt (Modellierungsvariante I). Der Krigingansatz wurde anderen zweidimensionalen Modellierungsvarianten, etwa zweidimensionalen P-Splines, vorgezogen, da Kriging am besten geeignet scheint,

den deutlich nicht rechteckigen Wertebereich der beiden Variablen (siehe Abbildung 8.13) zu berücksichtigen.

Neben der zweidimensionalen Modellierung wurde auch eine nichtparametrische Modellierung des Alterseffekts unter Vernachlässigung des Effekts der Stilldauer getestet (Modellierungsvariante N). Verwendet wurde ein P-Spline mit 20 inneren Knoten basierend auf quadrierten zweiten Differenzen für den Strafterm.

- *Alter der Mutter bei der Geburt, Größe und BMI der Mutter*: Der Einfluss der restlichen metrischen Variablen wurde auf vier unterschiedliche Arten modelliert:
  - Alle Effekte nichtparametrisch mit P-Splines (Modellierungsvariante N),
  - Nichtparametrische Modellierung für den Body-Mass-Index der Mutter und eine lineare Modellierung für die beiden anderen Variablen (B),
  - Modellierung des Body-Mass-Index durch ein quadratisches Polynom und eine lineare Modellierung für die beiden anderen Variablen (Q),
  - Alle Effekte linear (P).
- *Kategoriale Kovariablen*: In jedem Modell wurden die kategorialen Variablen *k_geschl*, *m_bildung* und *m_arbeit* unter Verwendung von entsprechenden Dummy-Variablen mit aufgenommen.

Jedes untersuchte Modell lässt sich eindeutig kennzeichnen durch eine Kombination aus drei Buchstaben. Beispielsweise kennzeichnet die Kombination (MIN) das Modell mit einem Markov Zufallsfeld für den räumlichen Effekt, einem Kriginganstz für den zweidimensionalen Interaktionseffekt des Alters und der Stilldauer und jeweils nichtparametrisch modellierten Effekten der übrigen metrischen Variablen. Das Modell (ZNQ) besteht aus unabhängigen zufälligen Effekten für den räumlichen Effekt, einem nichtparametrischen Effekt des Alters des Kindes, einem quadratischen Effekt für den Body Mass Index der Mutter und linearen Effekten für die übrigen Variablen.

8.6.4 Schätzergebnisse und Modellevaluation

Sämtliche untersuchten Modelle wurden unter Verwendung von Techniken für gemischte Modelle geschätzt, vergleiche die Abschnitte 7.1.9 und 8.5.2. Zur Beurteilung der Modellgüte verwenden wir als Gütemaß das AIC. Zum Vergleich wurden die Modelle auch in einem vollen Bayes-Ansatz unter Verwendung von MCMC-Techniken geschätzt. Die Unterschiede bei den Schätzergebnissen sind relativ gering, so dass die nachfolgend dargestellten Ergebnisse in der Regel auf der Repräsentation durch gemischte Modelle beruhen. Wir weisen jedoch auf vorhandene Unterschiede in den Ergebnissen an entsprechender Stelle hin.

Von allen untersuchten Modellen besitzt das Modell (MIB) basierend auf einem Markov-Zufallsfeld für den räumlichen Effekt, einem zweidimensionalen Kriginganstz für den Effekt des Alters und der Stilldauer, einem nichtlinearen Effekt des BMI und ansonsten linearen Effekten den kleinsten AIC-Wert. Abbildung 8.14 zeigt für sämtliche Modelle die AIC-Differenzen zu diesem, im Sinne des AIC besten Modell. Die Modelle sind aufsteigend nach AIC-Werten geordnet, so dass die gemäß AIC besten Modelle weiter links stehen. Zum Vergleich wurde darüberhinaus noch ein rein parametrisches Modell bestimmt, in dem der Alterseffekt über einen linearen Spline mit einem Knoten bei $k_alter = 23.5$ modelliert wurde. Für dieses Modell ergibt sich eine AIC-Differenz von 308, d.h. es handelt sich um das mit großem Abstand schlechteste Modell.

Abb. 8.14. *AIC-Differenzen zum besten Modell für die verschiedenen Modellspezifikationen.*

Abbildung 8.14 lässt folgende Schlussfolgerungen zu:

- Die Modelle mit zweidimensionaler Modellierung des Effekts des Alters und der Stilldauer sind den Modellen mit lediglich nichtparametrischer Modellierung des Altersseffekts deutlich überlegen. Hier sind die Unterschiede im AIC sehr groß.
- Bezüglich der Modellierung des räumlichen Effekts werden die besten Ergebnisse bei Verwendung von Markov-Zufallsfeldern erzielt. Die Verwendung von zufälligen Effekten zur Modellierung der räumlichen Heterogenität liefert um 2.5–2.7 Punkte höhere AIC Werte. Hier sind die Unterschiede verhältnismäßig gering. Die rein parametrische Modellierung basierend auf Regionen-Dummies schneidet hingegen deutlich schlechter ab.
- Wie bereits Abbildung 2.15 (Seite 45) zeigt, sind die Effekte der übrigen drei metrischen Variablen sehr glatt und jeweils nicht weit von einer Geraden entfernt. Dementsprechend liefern die oben angesprochenen vier Modellierungsvarianten für diese Variablen nahezu identische AIC-Werte. In jedem Fall sind die geschätzten Effekte kaum unterscheidbar. Lediglich beim Einfluss des BMI der Mutter herrscht eine gewisse Unsicherheit, ob ein linearer oder ein schwach nichtlinearer Effekt vorliegt.

Die Abbildungen 8.15 und 8.16 zeigen die geschätzten Effekte der metrischen Variablen und den räumlichen Effekt für das Modell (MIN), bei dem die Effekte sämtlicher metrischer Variablen nichtparametrisch modelliert wurden. Es handelt sich um das Modell mit dem zweit niedrigsten AIC-Wert. Tabelle 8.4 zeigt die Ergebnisse für die kategorialen Variablen. Schließlich sei darauf hingewiesen, dass Ergebnisse für das Modell (MNN) in Abbildung 2.15 (Seite 45) zu finden sind. Das Modell (MNN) unterscheidet sich vom Modell (MIN) dadurch, dass die Stilldauer unberücksichtigt bleibt.

Die Ergebnisse lassen sich wie folgt interpretieren:

- *Alter und Stilldauer:* Der durchschnittliche Z-Score nimmt mit wachsendem Alter zunächst annähernd linear ab. Anschließend stabilisiert sich der durchschnittliche Ernährungszustand für Kinder, die älter als 18 Monate sind. Ab 3 Jahren könnte sogar eine leichte Verbesserung eintreten. Die Kurve von nie gestillten Kindern liegt deutlich unterhalb der Kurven für Kinder mit Stillzeiten von 8, 16 und 24 Monaten. Es gibt also einen positiven Effekt des Stillens. Unterschiede für die verschiedenen Stillzeiten sind jedoch kaum auszumachen. Die im letzten Abschnitt formulierten Hypothesen zum Alters- und Stilldauereffekt bestätigen sich also weitgehend.

Abb. 8.15. *Interaktionseffekt des Alters und der Stilldauer im Modell (MIB) (links). Die rechte Grafik zeigt den Alterseffekt für verschiedene Stilldauern.*

Abb. 8.16. *Nichtparametrische Effekte und räumlicher Effekt für das Modell (MIB).*

- *BMI der Mutter:* Der dargestellte geschätzte Effekt zeigt, dass Kinder untergewichtiger Mütter im Durchschnitt schlechter ernährt sind als Kinder normal gewichtiger oder leicht übergewichtiger Mütter. Für Kinder mit stark übergewichtigen Müttern ist eine Abschwächung des ansonsten nahezu linearen Effekts zu erkennen. Basierend auf diesen Ergebnissen sinkt also der durchschnittliche Ernährungszustand bei stark

Abb. 8.17. *Schätzungen des Effekts des BMI der Mutter für alternative Modelle. Die linke Grafik zeigt eine Schätzung im Modell (MIB) basierend auf einem quadratischen Polynom. Die rechte Grafik zeigt eine nonparametrische Schätzung basierend auf einem vollen Bayes-Ansatz mit MCMC-Techniken.*

übergewichtigen Kindern nicht wie eigentlich mit einem invers U-förmigen Effekt postuliert.

Im Bereich stark übergewichtiger Frauen beobachten wir jedoch deutlich breiter werdende Konfidenzintervalle, hervorgerufen durch die insgesamt geringe Beobachtungszahl. Hier empfiehlt es sich zusätzlich die Ergebnisse alternativer Schätzungen heranzuziehen. Im praktisch gleichwertigen Modell (MIQ) mit quadratischer Modellierung für den BMI ist die postulierte inverse U-Form zu erkennen, wenn auch schwach ausgeprägt. Hier sinkt der durchschnittliche Z-Score für stark übergewichtige Frauen tatsächlich leicht ab, vergleiche die linke Grafik in Abbildung 8.17. Beim Effekt des BMI zeigen sich auch die größten Unterschiede zur alternativen Schätzung basierend auf MCMC-Methoden, vergleiche die rechte Grafik in Abbildung 8.17. Basierend auf MCMC-Schätztechniken erhalten wir eine noch etwas ausgeprägtere U-Form als durch die quadratische Modellierung.

Welches Fazit lässt sich also ziehen? Eine in der Praxis nicht seltene Vorgehensweise ist, das den vorher formulierten Hypothesen am nächsten kommende Ergebnis zu präsentieren (in unserem Fall die Bayes-Schätzungen) und die anderen Schätzungen zu unterschlagen. Diese Vorgehensweise ist natürlich nicht korrekt. Vielmehr empfiehlt es sich den gemeinsamen Kern der erzielten Ergebnisse herauszuarbeiten und auf Unterschiede hinzuweisen. Im vorliegenden Fall beobachten wir in jedem Fall einen annähernd linearen Effekt für unter-, normal- und leicht übergewichtige Mütter. Bei stark übergewichtigen Müttern ist der Effekt des BMI unsicher. Der vorher lineare Effekt schwächt sich ab oder sinkt sogar leicht. Für den in der Literatur postulierten invers U-förmigen Effekt gibt es also Hinweise, jedoch keinesfalls eindeutige Beweise. Um mehr Informationen zu erhalten, könnte man beispielsweise in einer der nachfolgenden Untersuchungen stark übergewichtige Mütter überrepräsentieren um genauere Aussagen zu erhalten. Sollte dies nicht möglich sein, könnte man die Daten anderer Länder zusätzlich untersuchen.

- *Größe der Mutter:* Hier zeigt sich ein linearer Zusammenhang. Je größer die Mutter ist, desto größer ist im Durchschnitt das Kind und damit der Z-Score. Rückschlüsse auf den Ernährungszustand liefert dieser Effekt jedoch nicht. Im Gegenteil, dieser sehr

Variable	Koeffizient	Standard-abweichung	95% Konfidenz-intervall	
Konstante	-186.926	22.685	-231.398	-142.454
k_geschl	-15.144	3.606	-22.213	-8.075
m_bildung1	-11.043	5.042	-20.927	-1.158
m_bildung3	21.774	5.055	11.863	31.684
m_bildung4	80.814	14.184	53.008	108.620
m_arbeit	-4.420	3.807	-11.883	3.044

Tabelle 8.4. *Schätzungen der parametrischen Effekte im Modell (MIB).*

große, rein durch Vererbung verursachte Effekt stellt eher den Z-Score als (alleiniges) Maß für chronische Unterernährung in Frage.

- *Alter der Mutter:* Auch dieser Effekt ist praktisch linear. Ältere Mütter haben also im Durchschnitt besser ernährte Kinder. Hier finden sich wieder leichte Unterschiede zu den Ergebnissen des alternativen Bayes-Ansatzes. Basierend auf diesem Ansatz ist ab einem Alter von circa 40 Jahren ein steilerer Anstieg des durchschnittlichen Z-Score zu beobachten. Der positive Effekt würde sich bei besonders alten Müttern noch einmal verstärken. Die im letzten Abschnitt formulierten Hypothesen lassen sich somit bestätigen.
- *Räumlicher Effekt:* Wir erkennen ein deutliches Nord-Süd Gefälle. Im Norden des Landes ist die Ernährungssituation deutlich besser als im Süden. Einer der Gründe für diesen starken Unterschied könnten schlechtere klimatische Bedingungen im Süden des Landes sein, da diese Landesteile eine deutlich niedrigere Höhenlage aufweisen als die nördlichen.
- *Kategoriale Kovariablen* Die geschätzten Parameter für die verschiedenen Ausbildungskategorien zeigen einen deutlichen Bildungseffekt. Je besser die Ausbildung der Mutter, desto besser die Ernährungssituation der eigenen Kinder. Der Unterschied zwischen Kindern, deren Mütter keine Ausbildung und lediglich eine Grundschulausbildung besitzen, ist mit durchschnittlich 11 Punkten noch relativ klein. Die Unterschiede für Kinder, deren Mütter eine darüber hinausgehende Ausbildung besitzen sind allerdings deutlich ausgeprägter. Im Durchschnitt besitzen Kinder von Müttern mit höherem Bildungsabschluss einen um über 90 Punkte verbesserten Z-Score im Vergleich zu Kindern, deren Mütter über keinerlei Ausbildung verfügen.

  Der Erwerbsstatus der Mutter scheint keine Rolle zu spielen, der Effekt ist deutlich nicht signifikant.

  Wie häufig in der Literatur festgestellt, scheinen Mädchen etwas robuster zu sein. Im Durchschnitt haben Jungen einen um 15.14 Punkte niedrigeren Z-Score.

Ein Blick auf die standardisierten Residuen in Abhängigkeit von den geschätzten Werten (Abbildung 8.18 links) und den Kovariablen (Grafiken nicht abgedruckt) liefert Hinweise auf vorhandene Heteroskedastizität. Eine erneute zweistufige Schätzung zur Berücksichtigung der Heteroskedastizität liefert aber keine nennenswerten Unterschiede in den Ergebnissen der Analyse. Der QQ-Plot der Residuen zeigt deutliche Hinweise auf Abweichungen von der Normalverteilung im Bereich besonders hoher Residuen mit Werten größer als 2. Es handelt sich überwiegend um Kinder mit positivem Z-Score, d.h. sehr gutem Ernährungszustand. Im großen und ganzen ist eine geringfügige Verletzung

Abb. 8.18. *Standardisierte Residuen des Basismodells gegen die vorhergesagten Werte (links) und QQ-Plot der standardisierten Residuen (rechts).*

der Normalverteilungsannahme nicht sehr problematisch, da die Verfahren in der Regel robust sind gegenüber kleineren Abweichungen von der Normalverteilung und bei dem vorliegenden Stichprobenumfang die asymptotische Gültigkeit von Tests und Konfidenzintervallen greifen sollte. Eine denkbare Alternative für weitergehende Analysen wären Ansätze, die zusätzlich die Modellierung der Schiefe einer Verteilung erlauben. Derartige Modelle sind in Rigby & Stasinopoulos (2005) beschrieben und stehen im R-Paket `GAMLSS` zur Verfügung.

8.7 Bemerkungen und Ergänzungen

Die in Abschnitt 7.4 von Kapitel 7 gegebenen Literaturhinweise zur nichtparametrischen Regression behalten auch für dieses Kapitel ihre Gültigkeit. Wir fassen hier lediglich die speziell zu Additiven Modellen und deren Erweiterungen verfassten Monographien kurz zusammen.

Der (nach wie vor lesenswerte) Klassiker unter den Monographien zur generalisierten additiven Regression ist sicherlich Hastie & Tibshirani (1990). Wood (2006) beschreibt generalisierte additive Modelle basierend auf P-Splines und zeigt die Anwendung der Verfahren mit dem Statistikprogramm R, insbesondere dem Paket `mgcv`. Ruppert et al. (2003) stellen generalisierte additive Modelle und Erweiterungen in einem einheitlichen Rahmen als gemischte Modelle dar und beschreiben die Schätzung basierend auf Methoden und Software für diese Modellklasse. Fan & Gijbels (1996), Loader (1999) und Härdle et al. (2004) befassen sich mit Ansätzen basierend auf lokalen Glättern. Die Monographie von Fotheringham, Brunsdon & Charlton (2002) beschreibt Modelle mit räumlich variierenden Koeffizienten ebenfalls basierend auf lokalen Glättungsverfahren. Das Buch von Yatchew (2003) beschreibt die Anwendung additiver Modelle in der Ökonometrie. Die mathematisch anspruchsvolle Monographie von Gu (2002) bettet Additive Modelle ein in die Theorie reproduzierender Hilberträume. Schließlich beschreiben Denison, Holmes, Mallick & Smith (2002) Bayesianische adaptive Regressionssplines als Bausteine für Generalisierte Additive Modelle und Erweiterungen.

A Matrix-Algebra

In diesem Anhang geben wir eine kompakte Einführung in die Matrizenrechnung bzw. Matrix-Algebra. Die Darstellung ist dabei an die Bedürfnisse des Buches angelehnt und beschränkt sich auf die Darstellung der wesentlichen Definitionen und Sätze ohne jedoch genauer auf Beispiele und Beweise einzugehen. Eine detailliertere Behandlung, die auch Beispiele beinhaltet, ist über die Internetseite des Buchs zu erreichen. Eine leicht lesbare Einführung mit sehr vielen Beispielen bietet die „Einführung in die Moderne Matrix-Algebra“ von Schmidt & Trenkler (2006).

A.1 Definition und elementare Operationen

Definition A.1 Reelle Matrix

Ein nach n Zeilen und p Spalten geordnetes Schema $\boldsymbol{A}$ von $n \cdot p$ Elementen $a_{ij} \in \mathbb{R}$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

heißt reelle Matrix der Ordnung $n \times p$, der Dimension $n \times p$ oder kurz $n \times p$ Matrix. Kurzschreibweise: $\boldsymbol{A} = (a_{ij})$, $i = 1, \ldots, n, j = 1, \ldots, p$.
Die Zeilen von $\boldsymbol{A}$ können als Vektoren des $\mathbb{R}^p$ (sog. Zeilenvektoren) und die Spalten als Vektoren des $\mathbb{R}^n$ (sog. Spaltenvektoren) angesehen werden.

Definition A.2 Transponierte Matrix

Sei $\boldsymbol{A} = (a_{ij})$ eine $n \times p$ Matrix. Dann ist die transponierte Matrix $\boldsymbol{A}'$ definiert als diejenige Matrix, die man durch das Vertauschen der Zeilen und Spalten von $\boldsymbol{A}$ erhält:

$$\boldsymbol{A}' = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{np} \end{pmatrix}$$

Die Matrix $\boldsymbol{A}'$ ist von der Ordnung $p \times n$.

Definition A.3 Quadratische Matrix

Eine Matrix $\boldsymbol{A}$ heißt quadratisch, falls sie von der Ordnung $n \times n$ ist. Die Diagonale, welche aus den Elementen $a_{11}, \ldots, a_{nn}$ besteht, heißt Hauptdiagonale.

Definition A.4 Diagonalmatrix

Eine quadratische Matrix $\boldsymbol{D}$ heißt Diagonalmatrix, wenn ihre Einträge unter- und oberhalb der Hauptdiagonalen Null sind, d.h. $\boldsymbol{D}$ besitzt folgende Gestalt:

$$\boldsymbol{D} = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & d_n \end{pmatrix}$$

Schreibweise: $\boldsymbol{D} = \operatorname{diag}(d_1, \dots, d_n)$.

Definition A.5 Einheitsmatrix

Die Diagonalmatrix

$$\boldsymbol{I}_n = \operatorname{diag}(1, \dots, 1) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & 1 \end{pmatrix}$$

heißt Einheitsmatrix.

Definition A.6 Symmetrische Matrix

Eine quadratische Matrix $\boldsymbol{A}$ heißt symmetrisch, wenn $\boldsymbol{A} = \boldsymbol{A}'$ gilt.

Definition A.7 Summe und skalare Multiplikation von Matrizen

Die Summe $\boldsymbol{A} + \boldsymbol{B}$ zweier $n \times p$ Matrizen $\boldsymbol{A} = (a_{ij})$ und $\boldsymbol{B} = (b_{ij})$ ist definiert als:

$$\boldsymbol{A} + \boldsymbol{B} = (a_{ij} + b_{ij}).$$

Die Multiplikation von $\boldsymbol{A}$ mit einem Skalar $\lambda \in \mathbb{R}$ ist definiert als

$$\lambda \boldsymbol{A} = (\lambda\, a_{ij}).$$

Satz A.1 Rechenregeln

Für beliebige $n \times p$ Matrizen $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ und beliebige Skalare $r, k \in \mathbb{R}$ gilt:

1. Assoziativgesetz für die Addition: $\boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C}$.
2. Kommutativgesetz: $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$.
3. Distributivgesetze für die skalare Multiplikation: $(k + r)\boldsymbol{A} = k\boldsymbol{A} + r\boldsymbol{A}$ bzw. $k(\boldsymbol{A} + \boldsymbol{B}) = k\boldsymbol{A} + k\boldsymbol{B}$.
4. Assoziativgesetz für die skalare Multiplikation: $(kr)\boldsymbol{A} = k(r\boldsymbol{A})$.
5. $(k\boldsymbol{A})' = k\boldsymbol{A}'$.
6. $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$.

Definition A.8 Matrixmultiplikation

Das Produkt der $n \times p$ Matrix $\boldsymbol{A} = (a_{ij})$ mit der $p \times m$ Matrix $\boldsymbol{B} = (b_{ij})$ ist die $n \times m$ Matrix

$$\boldsymbol{AB} = \boldsymbol{C} = (c_{ik}) \qquad \text{mit} \qquad c_{ik} = \sum_{j=1}^{p} a_{ij} b_{jk}.$$

Ausführlich erhalten wir demnach

$$\boldsymbol{A} \cdot \boldsymbol{B} = \begin{pmatrix} \sum_j a_{1j} b_{j1} & \cdots & \sum_j a_{1j} b_{jm} \\ \vdots & \ddots & \vdots \\ \sum_j a_{nj} b_{j1} & \cdots & \sum_j a_{nj} b_{jm} \end{pmatrix}.$$

Man beachte, dass zwei Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ nur dann multiplizierbar sind, wenn die Anzahl der Spalten von $\boldsymbol{A}$ gleich der Anzahl der Zeilen von $\boldsymbol{B}$ ist. Im Allgemeinen ist die Matrixmultiplikation darüberhinaus nicht kommutativ, d.h. es gilt $\boldsymbol{B} \cdot \boldsymbol{A} \neq \boldsymbol{A} \cdot \boldsymbol{B}$.

Satz A.2 Darstellung von Summen als Matrixprodukte

Seien $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ und $\mathbf{1}$ der $n \times 1$ Vektor, dessen Einträge sämtlich aus Einsen bestehen. Dann gilt:

1. $\displaystyle\sum_{i=1}^{n} x_i = \mathbf{1}'\boldsymbol{x} = \boldsymbol{x}'\mathbf{1}.$
2. $\displaystyle\sum_{i=1}^{n} x_i y_i = \boldsymbol{x}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{x}.$
3. $\displaystyle\sum_{i=1}^{n} x_i^2 = \boldsymbol{x}'\boldsymbol{x}.$

Satz A.3 Rechenregeln für die Matrixmultiplikation

Für Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$ und $\boldsymbol{C}$ passender Ordnungen gilt:

1. $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$.
2. $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$.
3. $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$.
4. $\boldsymbol{AI}_n = \boldsymbol{A}$ bzw. $\boldsymbol{I}_n\boldsymbol{A} = \boldsymbol{A}$.

Definition A.9 Kroneckerprodukt

Seien $\boldsymbol{A}$ und $\boldsymbol{B}$ Matrizen der Ordnungen $n \times p$ und $r \times q$. Dann ist das Kroneckerprodukt von $\boldsymbol{A}$ und $\boldsymbol{B}$ definiert als diejenige Matrix $\boldsymbol{C}$ der Ordnung $nr \times pq$ mit

$$\boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1p}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{n1}\boldsymbol{B} & a_{n2}\boldsymbol{B} & \cdots & a_{np}\boldsymbol{B} \end{pmatrix}.$$

Satz A.4 Rechenregeln für das Kroneckerprodukt

Seien $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ und $\boldsymbol{D}$ Matrizen passender Ordnungen sowie k ein Skalar. Dann gelten die folgenden Rechenregeln:

1. $k(\boldsymbol{A} \otimes \boldsymbol{B}) = (k\boldsymbol{A}) \otimes \boldsymbol{B} = \boldsymbol{A} \otimes (k\boldsymbol{B})$.
2. $\boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes \boldsymbol{C}$.
3. $\boldsymbol{A} \otimes (\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) + (\boldsymbol{A} \otimes \boldsymbol{C})$.
4. $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$.
5. $(\boldsymbol{AB}) \otimes (\boldsymbol{CD}) = (\boldsymbol{A} \otimes \boldsymbol{C})(\boldsymbol{B} \otimes \boldsymbol{D})$.

Definition A.10 Orthogonale Matrix

Eine quadratische Matrix $\boldsymbol{A}$ heißt orthogonal, wenn $\boldsymbol{AA}' = \boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$ gilt.

Satz A.5 Eigenschaften orthogonaler Matrizen

Sei $\boldsymbol{A}$ eine orthogonale Matrix. Dann gilt:

1. Die Zeilenvektoren bzw. die Spaltenvektoren bilden ein Orthonormalsystem, d.h. die Vektoren besitzen Länge Eins und sind paarweise orthogonal.
2. $\boldsymbol{AB}$ ist orthogonal, wenn $\boldsymbol{A}$ und $\boldsymbol{B}$ orthogonal sind.

Definition A.11 Idempotente Matrix

Eine quadratische Matrix $\boldsymbol{A}$ heißt idempotent, wenn gilt: $\boldsymbol{AA} = \boldsymbol{A}^2 = \boldsymbol{A}$.
Eine spezielle, in der Statistik wichtige idempotente Matrix ist die $n \times n$ Matrix

$$\boldsymbol{C} := \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}'.$$

Es gelten die folgenden Aussagen:

1. Multiplikation von $\boldsymbol{C}$ mit einem beliebigen $n \times 1$ Vektor $\boldsymbol{a}$ ergibt

$$\boldsymbol{Ca} = \begin{pmatrix} a_1 - \bar{a} \\ \vdots \\ a_n - \bar{a} \end{pmatrix},$$

   d.h. man erhält den mittelwertszentrierten Vektor.
2. Multiplikation von $\boldsymbol{C}$ mit einer $n \times m$ Matrix $\boldsymbol{A}$ liefert

$$\boldsymbol{CA} = \begin{pmatrix} a_{11} - \bar{a}_1 & \cdots & a_{1m} - \bar{a}_m \\ \vdots & & \vdots \\ a_{n1} - \bar{a}_1 & \cdots & a_{nm} - \bar{a}_m \end{pmatrix},$$

   wobei $\bar{a}_1, \ldots, \bar{a}_m$ die Mittelwerte der Spalten von A sind.
3. $\boldsymbol{C}\mathbf{1} = \mathbf{0}$.

4. $\mathbf{1}'\boldsymbol{C} = \mathbf{0}'$.
5. $\mathbf{1}\mathbf{1}'\boldsymbol{C} = \boldsymbol{C}\mathbf{1}\mathbf{1}' = \mathbf{0}$.
6. $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \boldsymbol{x}'\boldsymbol{C}\boldsymbol{x}$ wobei $\boldsymbol{x} = (x_1, \ldots, x_n)'$.

Satz A.6 Eigenschaften idempotenter Matrizen

Für idempotente Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ gilt:

1. $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$, also $\boldsymbol{A}\boldsymbol{B}$ idempotent.
2. $\boldsymbol{I} - \boldsymbol{A}$ ist idempotent.
3. $\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{A}) = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{A} = \mathbf{0}$.

A.2 Der Rang einer Matrix

Definition A.12 Zeilenrang, Spaltenrang

Sei $\boldsymbol{A}$ eine $n \times p$ Matrix. Die Maximalzahl linear unabhängiger Spaltenvektoren des $\mathbb{R}^n$ heißt Spaltenrang von $\boldsymbol{A}$, geschrieben $\mathrm{rgs}(\boldsymbol{A})$. Entsprechend erhält man den Zeilenrang $\mathrm{rgz}(\boldsymbol{A})$ von $\boldsymbol{A}$ als die Maximalzahl linear unabhängiger Zeilen von $\boldsymbol{A}$.

Satz A.7 Spaltenrang = Zeilenrang

Spaltenrang und Zeilenrang einer $n \times p$ Matrix $\boldsymbol{A}$ sind gleich, d.h.

$$\mathrm{rgs}(\boldsymbol{A}) = \mathrm{rgz}(\boldsymbol{A}).$$

Definition A.13 Rang einer Matrix

Der Rang $\mathrm{rg}(\boldsymbol{A})$ einer $n \times p$ Matrix $\boldsymbol{A}$ ist definiert als

$$\mathrm{rg}(\boldsymbol{A}) := \mathrm{rgs}(\boldsymbol{A}) = \mathrm{rgz}(\boldsymbol{A}) \leq \min\{n, p\}$$

Gilt $\mathrm{rg}(\boldsymbol{A}) = \min\{n, p\}$, so besitzt $\boldsymbol{A}$ vollen Rang und wird als regulär bezeichnet. Für $\mathrm{rg}(\boldsymbol{A}) = n$ ($\mathrm{rg}(\boldsymbol{A}) = p$) heißt $\boldsymbol{A}$ zeilenregulär (spaltenregulär).

Satz A.8 Allgemeine Rangbeziehungen

Für Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$ und $\boldsymbol{C}$ passender Ordnungen gilt:

1. $\mathrm{rg}(\boldsymbol{A}) = \mathrm{rg}(-\boldsymbol{A})$.
2. $\mathrm{rg}(\boldsymbol{A}') = \mathrm{rg}(\boldsymbol{A})$.
3. $\mathrm{rg}(\boldsymbol{A} + \boldsymbol{B}) \leq \mathrm{rg}(\boldsymbol{A}) + \mathrm{rg}(\boldsymbol{B})$.
4. $\mathrm{rg}(\boldsymbol{A}\boldsymbol{B}) \leq \min\{\mathrm{rg}(\boldsymbol{A}), \mathrm{rg}(\boldsymbol{B})\}$.
5. $\mathrm{rg}(\boldsymbol{I}_n) = n$.

Definition A.14 Nullraum

Der Nullraum $N(\boldsymbol{A})$ einer $n \times p$ Matrix $\boldsymbol{A}$ ist definiert als die Menge

$$N(\boldsymbol{A}) := \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

Definition A.15 Zeilenraum, Spaltenraum

Der Zeilenraum $Z(\boldsymbol{A})$ einer $n \times p$ Matrix $\boldsymbol{A}$ ist der durch die Zeilen von $\boldsymbol{A}$ aufgespannte Unterraum des $\mathbb{R}^n$:

$$Z(\boldsymbol{A}) := \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} \text{ für ein } \boldsymbol{y} \in \mathbb{R}^p\}.$$

Analog lässt sich der Spaltenraum (als Teilraum des $\mathbb{R}^p$) definieren.

Satz A.9 Eigenschaften des Nullraums

Sei $\boldsymbol{A}$ eine $n \times p$ Matrix. Dann gilt:

1. Der Nullraum ist ein Unterraum des $\mathbb{R}^p$.
2. $\mathrm{rg}(\boldsymbol{A}) + \dim(N(\boldsymbol{A})) = p$ bzw. $\dim(N(\boldsymbol{A})) = p - \mathrm{rg}(\boldsymbol{A})$. Die Dimension des Nullraums $N(\boldsymbol{A})$ wird als Defekt von $\boldsymbol{A}$ bezeichnet.
3. Der Nullraum $N(\boldsymbol{A})$ ist das orthogonale Komplement des Zeilenraums $Z(\boldsymbol{A})$ von $\boldsymbol{A}$.
4. $N(\boldsymbol{A}'\boldsymbol{A}) = N(\boldsymbol{A})$.

Definition A.16 Inverse einer Matrix

Sei $\boldsymbol{A}$ eine quadratische Matrix. Eine Matrix $\boldsymbol{A}^{-1}$ heißt Inverse zur Matrix $\boldsymbol{A}$, falls gilt:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = I$$

Satz A.10 Existenz und Eindeutigkeit der Inversen

Die Inverse einer quadratischen $n \times n$ Matrix $\boldsymbol{A}$ existiert genau dann, wenn $\mathrm{rg}(\boldsymbol{A}) = n$ gilt, also wenn $\boldsymbol{A}$ regulär ist. Die Inverse ist dann eindeutig bestimmt und die Matrix $\boldsymbol{A}$ heißt invertierbar.

Satz A.11 Rechenregeln für Inverse

Seien $\boldsymbol{A}$, $\boldsymbol{B}$ und $\boldsymbol{C}$ invertierbare Matrizen gleicher Ordnung und $k \neq 0$ ein Skalar. Dann gilt

1. $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.
2. $(k\boldsymbol{A})^{-1} = k^{-1}\boldsymbol{A}^{-1} = \dfrac{1}{k}\boldsymbol{A}^{-1}$.
3. $(\boldsymbol{A}')^{-1} = (\boldsymbol{A}^{-1})'$.
4. $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
5. $(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C})^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

6. $\boldsymbol{A}$ symmetrisch $\Rightarrow \boldsymbol{A}^{-1}$ symmetrisch.
7. Für eine Diagonalmatrix $\boldsymbol{A} = \operatorname{diag}(a_1, \ldots, a_n)$ gilt
$$\boldsymbol{A}^{-1} = \operatorname{diag}(a_1^{-1}, \ldots, a_n^{-1}).$$
8. Falls $\boldsymbol{A}$ orthogonal ist, gilt $\boldsymbol{A}^{-1} = \boldsymbol{A}'$.
9. Sei $\boldsymbol{A}$ partitioniert in
$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$
und seien die Submatrizen $\boldsymbol{A}_{11}$ und $\boldsymbol{A}_{22}$ quadratisch und invertierbar. Dann gilt
$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{B}^{-1} & -\boldsymbol{B}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}^{-1} & \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \end{pmatrix}$$
mit $\quad \boldsymbol{B} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21},$

und $\quad \boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}\boldsymbol{A}_{12}\boldsymbol{C}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{C}^{-1} \\ -\boldsymbol{C}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{C}^{-1} \end{pmatrix}$

mit $\quad \boldsymbol{C} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}.$

A.3 Determinante und Spur einer Matrix

Definition A.17 Determinante

Die Determinante einer quadratischen Matrix $\boldsymbol{A}$ der Ordnung $n \times n$ ist definiert als
$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |\boldsymbol{A}_{-ij}|,$$
wobei $\boldsymbol{A}_{-ij}$ die $n-1 \times n-1$ dimensionale Matrix bezeichnet, die durch Streichung der i-ten Zeile und der j-ten Spalte aus $\boldsymbol{A}$ entsteht. Für skalare Matrizen $\boldsymbol{A} = (a_{11})$ der Ordnung 1×1 gilt $|\boldsymbol{A}| = a_{11}$. Für eine 2×2 Matrix erhält man $|\boldsymbol{A}| = a_{11}a_{22} - a_{12}a_{21}$. In diesem Fall lässt sich die Determinante auch geometrisch interpretieren. Die Determinante von $\boldsymbol{A}$ entspricht dem Flächeninhalt des von den beiden Spaltenvektoren gebildeten Parallelogramms. Für 3×3 Matrizen entspricht die Determinante von $\boldsymbol{A}$ dem Volumen des von den drei Spaltenvektoren aufgespannten Körpers. Für $n > 3$ ergeben sich analoge Interpretationen.

Satz A.12 Determinante der Transponierten

Für eine quadratische Matrix $\boldsymbol{A}$ gilt $|\boldsymbol{A}'| = |\boldsymbol{A}|$.

Satz A.13 Determinanten einiger bestimmter Matrizen

Sei $\boldsymbol{A}$ eine quadratische Matrix. Dann gilt:

1. Wenn eine Zeile (Spalte) von $\boldsymbol{A}$ aus Nullen besteht, dann gilt $|\boldsymbol{A}| = 0$.
2. Wenn $\boldsymbol{A}$ zwei identische Zeilen (Spalten) besitzt, dann gilt $|\boldsymbol{A}| = 0$.
3. Die Determinante einer Matrix in Dreiecksform ist das Produkt der Diagonalelemente. Eine Matrix besitzt Dreiecksform, wenn alle Elemente ober bzw. unterhalb der Hauptdiagonalen gleich Null sind.
4. $|\boldsymbol{I}| = 1$

Satz A.14 Eigenschaften von Determinanten

Für die Determinante einer $n \times n$ Matrix $\boldsymbol{A}$ gilt:

1. $|k\boldsymbol{A}| = k^n|\boldsymbol{A}|$.
2. $|\boldsymbol{A}| \neq 0 \Longleftrightarrow \mathrm{rg}(\boldsymbol{A}) = n$.
3. $|\boldsymbol{AB}| = |\boldsymbol{A}| \cdot |\boldsymbol{B}|$.
4. $|\boldsymbol{A}^{-1}| = \dfrac{1}{|\boldsymbol{A}|}$.
5. $\boldsymbol{A}$ orthogonal $\Rightarrow |\boldsymbol{A}| = \pm 1$.

Definition A.18 Spur einer Matrix

Sei $\boldsymbol{A} = (a_{ij})$ eine quadratische $n \times n$ Matrix. Dann heißt die Summe der Diagonalelemente Spur von $\boldsymbol{A}$, in Zeichen

$$\mathrm{sp}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}.$$

Satz A.15 Eigenschaften der Spur

Für die Spur der $n \times n$ Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ gilt:

1. $\mathrm{sp}(\boldsymbol{A} + \boldsymbol{B}) = \mathrm{sp}(\boldsymbol{A}) + \mathrm{sp}(\boldsymbol{B})$.
2. $\mathrm{sp}(\boldsymbol{A}) = \mathrm{sp}(\boldsymbol{A}')$.
3. $\mathrm{sp}(k\boldsymbol{A}) = k \cdot \mathrm{sp}(\boldsymbol{A})$.
4. $\mathrm{sp}(\boldsymbol{AB}) = \mathrm{sp}(\boldsymbol{BA})$. Dies bleibt auch für den Fall gültig, dass $\boldsymbol{A}$ eine $n \times p$ und $\boldsymbol{B}$ eine $p \times n$ Matrix ist.
5. Seien $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Dann gilt $\mathrm{sp}(\boldsymbol{xy}') = \mathrm{sp}(\boldsymbol{yx}') = \mathrm{sp}(\boldsymbol{x}'\boldsymbol{y}) = \boldsymbol{x}'\boldsymbol{y}$.

A.4 Verallgemeinerte Inverse

Definition A.19 Verallgemeinerte Inverse

Sei $\boldsymbol{A}$ eine beliebige $n \times p$ Matrix mit $n \leq p$. Dann heißt die $p \times n$ Matrix $\boldsymbol{A}^-$ verallgemeinerte Inverse oder g-Inverse (generalized Inverse) von $\boldsymbol{A}$ falls gilt

$$\boldsymbol{AA}^-\boldsymbol{A} = \boldsymbol{A}.$$

Satz A.16 Existenz der verallgemeinerten Inversen

Zu jeder Matrix $\boldsymbol{A}$ existiert eine verallgemeinerte Inverse, die aber im Allgemeinen nicht eindeutig ist.

Satz A.17 Eigenschaften der verallgemeinerten Inversen

Sei $\boldsymbol{A}^-$ eine verallgemeinerte Inverse der Matrix $\boldsymbol{A}$. Dann gilt:

1. $\mathrm{rg}(\boldsymbol{A}) = \mathrm{rg}(\boldsymbol{A}\boldsymbol{A}^-) = \mathrm{rg}(\boldsymbol{A}^-\boldsymbol{A})$.
2. $\mathrm{rg}(\boldsymbol{A}) \leq \mathrm{rg}(\boldsymbol{A}^-)$.
3. $\boldsymbol{A}$ regulär $\Rightarrow \boldsymbol{A}^- = \boldsymbol{A}^{-1}$. Insbesondere ist in diesem Fall die verallgemeinerte Inverse eindeutig.
4. $\boldsymbol{A}^-\boldsymbol{A}$ und $\boldsymbol{A}\boldsymbol{A}^-$ sind idempotent.

A.5 Eigenwerte und Eigenvektoren

Definition A.20 Eigenwert und Eigenvektor

Sei $\boldsymbol{A}$ eine quadratische $n \times n$ Matrix. Dann heißt (die im Allgemeinen komplexe Zahl) $\lambda \in \mathbb{C}$ Eigenwert von $\boldsymbol{A}$, wenn ein (im Allgemeinen komplexer) Vektor $\boldsymbol{x} \in \mathbb{C}^n$ mit $\boldsymbol{x} \neq \boldsymbol{0}$ existiert, so dass gilt:

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \qquad \text{bzw.} \qquad (\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}.$$

Der Vektor $\boldsymbol{x}$ heißt dann Eigenvektor zum Eigenwert λ.

Definition A.21 Charakteristisches Polynom

Sei $\boldsymbol{A}$ eine quadratische $n \times n$ Matrix. Dann heißt

$$q(\lambda) := |\boldsymbol{A} - \lambda\boldsymbol{I}|$$

charakteristisches Polynom von $\boldsymbol{A}$.

Aus der Definition der Determinante (siehe Definition A.17), ergibt sich, dass $q(\lambda)$ tatsächlich ein Polynom vom Grad n ist, so dass sich $q(\lambda)$ äquivalent darstellen lässt als

$$q(\lambda) = (-\lambda)^n + \alpha_{m-1}(-\lambda)^{m-1} + \cdots + \alpha_1(-\lambda) + \alpha_0, \tag{A.1}$$

wobei die Skalare $\alpha_0, \ldots, \alpha_{m-1}$ zunächst unspezifiziert sind.

Das Polynom $q(\lambda)$ lässt sich stets auch in die Gestalt

$$q(\lambda) = |\boldsymbol{A} - \lambda\boldsymbol{I}| = \prod_{i=1}^{n}(\lambda_i - \lambda) \tag{A.2}$$

bringen, wobei $\lambda_1, \ldots, \lambda_n$ die Nullstellen des Polynoms sind. Nach dem Fundamentalsatz der Algebra hat dieses Polynom genau n nicht notwendig verschiedene und auch nicht notwendig reellwertige Nullstellen.

Satz A.18 Berechnung der Eigenwerte über das charakteristische Polynom

Die Eigenwerte einer quadratischen Matrix $\boldsymbol{A}$ sind die Nullstellen des charakteristischen Polynoms, also die Lösungen von

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0.$$

Satz A.19 Eigenschaften von Eigenwerten

Für die Eigenwerte λ_i einer $n \times n$ Matrix gelten folgende Eigenschaften:

1. $|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i$.
2. $\mathrm{sp}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$.
3. $\boldsymbol{A}$ ist genau dann regulär, wenn alle Eigenwerte ungleich Null sind.
4. Die Matrizen $\boldsymbol{A}$ und $\boldsymbol{A}'$ besitzen dasselbe charakteristische Polynom und damit dieselben Eigenwerte.
5. Ist λ ein Eigenwert einer regulären Matrix $\boldsymbol{A}$, dann ist $\frac{1}{\lambda}$ ein Eigenwert von $\boldsymbol{A}^{-1}$.
6. Die Eigenwerte einer Diagonalmatrix $\boldsymbol{D}$ sind gerade die Hauptdiagonalelemente.
7. Für die Eigenwerte λ_i einer orthogonalen Matrix $\boldsymbol{A}$ gilt $\lambda_i = \pm 1$.
8. Die Eigenwerte einer idempotenten Matrix $\boldsymbol{A}$ sind 1 oder 0.

Definition A.22 Eigenraum

Sei $\boldsymbol{A}$ eine quadratische Matrix und λ ein Eigenwert von $\boldsymbol{A}$. Die Menge

$$\boldsymbol{A}_\lambda := \{\boldsymbol{x} \in \mathbb{C}^n | \boldsymbol{x} \text{ Eigenvektor zu } \lambda\} \cup \{0\}$$

heißt Eigenraum zum Eigenwert λ.

Definition A.23 Ähnliche Matrizen

Zwei Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ heißen ähnlich (in Zeichen $\boldsymbol{A} \sim \boldsymbol{B}$), wenn eine reguläre Matrix $\boldsymbol{C}$ existiert, so dass $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^{-1}$ gilt.

Satz A.20 Eigenwerte ähnlicher Matrizen

Für ähnliche Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ gilt:

1. $\boldsymbol{A}$ und $\boldsymbol{B}$ haben dasselbe charakteristische Polynom und damit dieselben Eigenwerte.
2. Ist $\boldsymbol{x}$ Eigenvektor zum Eigenwert λ, so ist $\boldsymbol{C}\boldsymbol{x}$ Eigenvektor der Matrix $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^{-1}$.

Satz A.21 Eigenwerte und Eigenvektoren symmetrischer Matrizen

Sei $\boldsymbol{A}$ eine symmetrische $n \times n$ Matrix. Dann gilt:

1. Alle Eigenwerte sind reell.
2. Die zu verschiedenen Eigenwerten gehörenden Eigenvektoren sind paarweise orthogonal.

Satz A.22 Spektralzerlegung

Sei $\boldsymbol{A}$ eine symmetrische $n \times n$ Matrix mit $\mathrm{rg}(\boldsymbol{A}) = r$. Dann existiert eine $n \times r$ Matrix $\boldsymbol{P}$, so dass gilt:

$$\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r) \qquad \text{bzw.} \qquad \boldsymbol{A} = \boldsymbol{P}\mathrm{diag}(\lambda_1, \ldots, \lambda_r)\boldsymbol{P}'.$$

Dabei sind die λ_i die von Null verschiedenen Eigenwerte von $\boldsymbol{A}$ (Insbesondere entspricht der Rang von $\boldsymbol{A}$ der Anzahl der von Null verschiedenen Eigenwerte). Die Spaltenvektoren von $\boldsymbol{P}$ entsprechen den (paarweise orthonormalen) zugehörigen Eigenvektoren.

Satz A.23 Spektralzerlegung einer idempotenten Matrix

Sei $\boldsymbol{A}$ eine symmetrische und idempotente $n \times n$ Matrix mit $\mathrm{rg}(\boldsymbol{A}) = r$. Dann existiert eine orthogonale Matrix $\boldsymbol{A}$ so dass gilt

$$\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P} = \boldsymbol{I}_r$$

Außerdem ergibt sich

$$\mathrm{rg}(\boldsymbol{A}) = \mathrm{sp}(\boldsymbol{A}).$$

A.6 Quadratische Formen

Definition A.24 Quadratische Form

Sei $\boldsymbol{A}$ eine symmetrische $n \times n$ Matrix. Eine quadratische Form in einem Vektor $\boldsymbol{x} \in \mathbb{R}^n$ ist definiert durch:

$$Q(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} a_{ij}x_ix_j.$$

Definition A.25 Definite Matrizen

Die quadratische Form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ und die Matrix $\boldsymbol{A}$ heißen

1. positiv definit, falls $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ für alle $\boldsymbol{x} \neq 0$. Schreibweise: $\boldsymbol{A} > 0$.
2. positiv semidefinit, falls $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$ und $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = 0$ für mindestens ein $\boldsymbol{x} \neq 0$.
3. nichtnegativ definit, falls $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ bzw. $\boldsymbol{A}$ entweder positiv oder positiv semidefinit ist. Schreibweise: $\boldsymbol{A} \geq 0$.
4. negativ definit, wenn $-\boldsymbol{A}$ positiv definit ist.
5. negativ semidefinit, wenn $-\boldsymbol{A}$ positiv semidefinit ist.
6. indefinit in allen anderen Fällen.

Satz A.24 Kriterium für die Definitheit einer Matrix

Sei $\boldsymbol{A}$ eine symmetrische Matrix mit den (reellen) Eigenwerten $\lambda_1, \ldots, \lambda_n$. Dann ist $\boldsymbol{A}$ genau dann

1. positiv definit, wenn $\lambda_i > 0$ für $i = 1, \ldots, n$,
2. positiv semidefinit, wenn $\lambda_i \geq 0$ für $i = 1, \ldots, n$ und mindestens ein $\lambda_i = 0$,
3. negativ definit, wenn $\lambda_i < 0$ für alle $i = 1 \ldots, n$,
4. negativ semidefinit, wenn $\lambda_i \leq 0$ für $i = 1, \ldots, n$ und mindestens ein $\lambda_i = 0$,
5. indefinit, wenn $\boldsymbol{A}$ mindestens einen positiven und einen negativen Eigenwert besitzt.

Satz A.25 Eigenschaften positiv definiter Matrizen

Sei $\boldsymbol{A}$ positiv definit. Dann gilt:

1. $\boldsymbol{A}$ ist regulär (und damit invertierbar).
2. $\boldsymbol{A}^{-1}$ ist positiv definit.
3. Für die Diagonalelemente a_{ii}, $i = 1, \ldots, n$ gilt: $a_{ii} > 0$.
4. $\mathrm{sp}(\boldsymbol{A}) > 0$.
5. Sei $\boldsymbol{B}$ positiv semidefinit. Dann ist $\boldsymbol{A} + \boldsymbol{B}$ positiv definit.

Satz A.26

Seien $\boldsymbol{A}$ eine $n \times n$ und $\boldsymbol{Q}$ eine $n \times m$ Matrix. Dann gilt:

1. Ist $\boldsymbol{A}$ nichtnegativ definit, so ist auch $\boldsymbol{Q}'\boldsymbol{A}\boldsymbol{Q}$ nichtnegativ definit.
2. Ist $\boldsymbol{A}$ positiv definit und $\boldsymbol{Q}$ spaltenregulär, so ist auch $\boldsymbol{Q}'\boldsymbol{A}\boldsymbol{Q}$ positiv definit.

Satz A.27

Sei $\boldsymbol{B}$ eine $n \times p$ Matrix. Dann ist die Matrix $\boldsymbol{B}'\boldsymbol{B}$ symmetrisch und nicht negativ definit. Sie ist positiv definit, wenn $\boldsymbol{B}$ spaltenregulär ist. Neben $\boldsymbol{B}'\boldsymbol{B}$ ist dann auch $\boldsymbol{B}\boldsymbol{B}'$ nichtnegativ definit.

Satz A.28 Eigenwerte von $B'B$ und BB'

Sei $\boldsymbol{B}$ eine $n \times p$ Matrix mit $\mathrm{rg}(\boldsymbol{B}) = r$. Dann gilt:

1. Sowohl $\boldsymbol{B}\boldsymbol{B}'$ als auch $\boldsymbol{B}'\boldsymbol{B}$ besitzen r von Null verschiedene Eigenwerte λ_j, $j = 1, \ldots, r$. Diese sind positiv und identisch für $\boldsymbol{B}\boldsymbol{B}'$ und $\boldsymbol{B}'\boldsymbol{B}$.
2. Falls $\boldsymbol{v}$ ein Eigenvektor von $\boldsymbol{B}'\boldsymbol{B}$ zum Eigenwert λ ist, dann ist

$$\boldsymbol{u} := \frac{1}{\sqrt{\lambda}}\boldsymbol{B}\boldsymbol{v}$$

ein Eigenvektor von $\boldsymbol{B}\boldsymbol{B}'$ zum Eigenwert λ.

Satz A.29 Cholesky-Zerlegung

Jede symmetrische und positiv definite $n \times n$ Matrix $\boldsymbol{A}$ lässt sich eindeutig darstellen als

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}',$$

wobei $\boldsymbol{L}$ die Gestalt einer unteren Dreiecksmatrix mit positiven Diagonalelementen besitzt. $\boldsymbol{L}$ heißt Cholesky-Faktor von $\boldsymbol{A}$.

A.7 Differentiation von Matrixfunktionen

Definition A.26 Differentiation nach einem Skalar

Sei $\boldsymbol{A} = (a_{ij})$ eine $n \times p$ Matrix, deren Elemente differenzierbare Funktionen der reellen Variablen t seien. Dann heißt die Matrix

$$\frac{\partial \boldsymbol{A}}{\partial t} = \left(\frac{\partial a_{ij}}{\partial t}\right)$$

Ableitung von $\boldsymbol{A}$ nach t.

Satz A.30 Rechenregeln

Sei $\boldsymbol{A}$ und $\boldsymbol{B}$ Matrizen passender Ordnungen. Dann gilt:

1. $\dfrac{\partial \boldsymbol{A}}{\partial a_{ij}} = e_i e_j'$, wobei $e_i = (0, \ldots, \underbrace{1}_{i}, \ldots, 0)'$.
2. $\dfrac{\partial \boldsymbol{A}'}{\partial a_{ij}} = e_j e_i'$.
3. $\dfrac{\partial \boldsymbol{A}\boldsymbol{B}}{\partial t} = \dfrac{\partial \boldsymbol{A}}{\partial t}\boldsymbol{B} + \boldsymbol{A}\dfrac{\partial \boldsymbol{B}}{\partial t}$ (Produktregel).

Satz A.31 Differentiation von Funktionalen einer Matrix

Sei $\boldsymbol{A}$ eine quadratische Matrix, deren Elemente differenzierbare Funktionen der reellen Variablen t seien. Dann gilt:

1. Die Ableitung der Spur ist die Spur der Ableitung:
$$\frac{\partial \mathrm{sp}(\boldsymbol{A})}{\partial t} = \mathrm{sp}\left(\frac{\partial \boldsymbol{A}}{\partial t}\right).$$
2. Ist $\boldsymbol{A}$ invertierbar, so ergibt sich die Ableitung der Inversen als
$$\frac{\partial \boldsymbol{A}^{-1}}{\partial t} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial t}\boldsymbol{A}^{-1}.$$
3. Ist $\boldsymbol{A}$ invertierbar, so ergibt sich die Ableitung der logarithmierten Determinante als
$$\frac{\partial \log(|\boldsymbol{A}|)}{\partial t} = \mathrm{sp}\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial t}\right).$$

Definition A.27 Differentiation nach einer Matrix

Sei $\boldsymbol{A} = (a_{ij})$ eine $n \times p$ Matrix und $f(\boldsymbol{A})$ eine differenzierbare reellwertige Funktion der np Elemente a_{ij}. Dann heißt die $n \times p$ Matrix

$$\frac{\partial f}{\partial \boldsymbol{A}} = \left(\frac{\partial f}{\partial a_{ij}}\right)$$

Ableitung von f nach $\boldsymbol{A}$.

Satz A.32 Rechenregeln

Seien $\boldsymbol{A}$ und $\boldsymbol{B}$ Matrizen, f und g Funktionen von Matrizen sowie $\boldsymbol{x}$ und $\boldsymbol{y}$ Vektoren. Bei den folgenden Größen wird angenommen, dass sie existieren und von passender Ordnung sind. Dann gelten folgende Rechenregeln:

1. $\dfrac{\partial fg}{\partial \boldsymbol{A}} = \dfrac{\partial f}{\partial \boldsymbol{A}} g + f \dfrac{\partial g}{\partial \boldsymbol{A}}$.
2. $\dfrac{\partial \mathrm{sp}(\boldsymbol{A})}{\partial \boldsymbol{A}} = I$.
3. $\dfrac{\partial \mathrm{sp}(\boldsymbol{B}\boldsymbol{A})}{\partial \boldsymbol{A}} = \boldsymbol{B}'$.
4. $\dfrac{\partial \mathrm{sp}(\boldsymbol{A}'\boldsymbol{B}\boldsymbol{A})}{\partial \boldsymbol{A}} = (\boldsymbol{B} + \boldsymbol{B}')\boldsymbol{A}$.
5. $\dfrac{\partial \mathrm{sp}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}')}{\partial \boldsymbol{A}} = \boldsymbol{A}'(\boldsymbol{B} + \boldsymbol{B}')$.
6. $\dfrac{\partial \mathrm{sp}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A})}{\partial \boldsymbol{A}} = \boldsymbol{A}'\boldsymbol{B}' + \boldsymbol{B}'\boldsymbol{A}'$.
7. $\dfrac{\partial \boldsymbol{y}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{y}$.
8. $\dfrac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{y}}{\partial \boldsymbol{A}} = \boldsymbol{x}\boldsymbol{y}'$.
9. $\dfrac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}$.
10. Für symmetrisches $\boldsymbol{A}$ gilt

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x} = 2\boldsymbol{A}'\boldsymbol{x}.$$

B Wahrscheinlichkeitsrechnung und induktive Statistik

Dieser Anhang enthält in knapper Form Teile der Stochastik und der induktiven Statistik, die im Rahmen des Buchs verwendet werden, jedoch teilweise über Grundkenntnisse, wie sie in Einführungen zur Statistik vermittelt werden, hinaus gehen. Dies betrifft neben einigen eindimensionalen Verteilungen insbesondere mehrdimensionale Zufallsvariablen sowie die Likelihood- und Bayes-Inferenz bei mehrdimensionalen Parametern. Einführungen zur Statistik findet man etwa in den Lehrbüchern Fahrmeir et al. (2007), Mosler & Schmid (2005), Schira (2005) oder Bortz (2004). Mehrdimensionale Zufallsvariablen und multivariate Likelihood-Inferenz sind ausführlich in Fahrmeir et al. (1996) beschrieben. Die Homepage des Buches enthält eine ausführlichere Version dieses Anhangs.

B.1 Einige eindimensionale Verteilungen

Definition B.1 Normalverteilung und trunkierte Normalverteilung

Eine stetige Zufallsvariable X heißt normalverteilt, in Zeichen $X \sim \mathrm{N}(\mu, \sigma^2)$, wenn sie die Dichte

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

besitzt. Für den Erwartungswert und die Varianz gilt $\mathrm{E}(X) = \mu$ und $\mathrm{Var}(X) = \sigma^2$. Die spezielle Verteilung mit $\mu = 0$ und $\sigma^2 = 1$ heißt Standardnormalverteilung.

Die Verteilung von X unter der Bedingung $a \leq X \leq b$ heißt trunkierte Normalverteilung, in Zeichen $X \sim \mathrm{TN}_{a,b}(\mu, \sigma^2)$. Die Dichte ist gegeben durch

$$g(x) = \begin{cases} \dfrac{f(x)}{\mathrm{P}(a \leq X \leq b)}, & a \leq x \leq b \\ 0 & \text{sonst.} \end{cases}$$

Durch die Bedingung wird X also auf den Wertebereich $[a, b]$ eingeschränkt und die Dichte entsprechend renormiert. Für $a = -\infty$ bzw. $b = \infty$ spricht auch von der links bzw. rechts trunkierten Normalverteilung.

Definition B.2 Lognormalverteilung

Eine stetige, nicht-negative Zufallsvariable X heißt logarithmisch normalverteilt, in Zeichen $X \sim \mathrm{LN}(\mu, \sigma^2)$, falls die transformierte Zufallsvariable $Y = \log(X)$ $\mathrm{N}(\mu, \sigma^2)$-verteilt ist. Die Dichte von X ist gegeben durch

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x} \exp\left(-(\log(x) - \mu)^2/2\sigma^2\right), \qquad x > 0.$$

Für Erwartungswert und Varianz gilt

$$\begin{aligned} \mathrm{E}(X) &= \exp(\mu + \sigma^2/2), \\ \mathrm{Var}(X) &= \exp(2\mu + \sigma^2) \cdot (\exp(\sigma^2) - 1). \end{aligned}$$

Definition B.3 Gammaverteilung

Eine stetige, nicht-negative Zufallsvariable X heißt gammaverteilt mit Parametern $a > 0$ und $b > 0$, in Zeichen $X \sim \mathrm{G}(a, b)$, wenn sie die Dichte

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \qquad x > 0,$$

besitzt. Erwartungswert und Varianz sind gegeben durch $\mathrm{E}(X) = a/b$ und $\mathrm{Var}(X) = a/b^2$. Der Modus liegt bei $(a-1)/b$ (falls $a > 1$).

Gelegentlich wird die Gammaverteilung in einer alternativen Parametrisierung definiert. In Abhängigkeit vom Erwartungswert μ und dem Skalenparameter ν ist die Dichte gegeben durch

$$f(x) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} \exp\left(-\frac{\nu}{\mu} x\right), \qquad x > 0.$$

Eine Verwendung dieser alternativen Parametrisierung findet man in Kapitel 4 im Zusammenhang mit der Gamma-Regression, vergleiche Seite 217.

Definition B.4 Exponentialverteilung

Eine stetige, nicht-negative Zufallsvariable X heißt exponentialverteilt mit Parameter $\lambda > 0$, in Zeichen $X \sim \mathrm{Expo}(\lambda)$, wenn sie folgende Dichte besitzt:

$$f(x) = \lambda \exp(-\lambda x), \qquad x > 0.$$

Für den Erwartungswert und die Varianz erhält man $\mathrm{E}(X) = 1/\lambda$ und $\mathrm{Var}(X) = 1/\lambda^2$. Die Exponentialverteilung ist eine spezielle Gammaverteilung mit $a = 1$ und $b = \lambda$.

Definition B.5 χ^2-Verteilung

Eine stetige, nicht-negative Zufallsvariable X mit Dichte

$$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp\left(-\frac{1}{2} x\right), \qquad x > 0,$$

heißt χ^2-verteilt mit n Freiheitsgraden, in Zeichen $X \sim \chi^2_n$. Der Erwartungswert ist gegeben durch $\mathrm{E}(X) = n$ und die Varianz durch $\mathrm{Var}(X) = 2n$. Die χ^2-Verteilung ist eine spezielle Gammaverteilung mit $a = n/2$ und $b = 1/2$.

Sind $X_1, \ldots, X_n$ unabhängig und identisch standardnormalverteilt, so ist

$$Y_n = \sum_{i=1}^{n} X_i^2$$

χ^2-verteilt mit n Freiheitsgraden. Aus dieser Darstellung folgt auch (mit Hilfe des Gesetzes der großen Zahlen), die (fast sichere) Konvergenz $Y_n/n \to 1$ für $n \to \infty$.

Definition B.6 Inverse Gammaverteilung

Gilt $Y \sim \mathrm{G}(a,b)$, so heißt $X = 1/Y$ invers gammaverteilt, in Zeichen $X \sim \mathrm{IG}(a,b)$. Die Dichte von X besitzt folgende Gestalt:

$$f(x) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp\left(-b/x\right), \qquad x > 0.$$

Man erhält $\mathrm{E}(X) = b/(a-1)$ und $\mathrm{Var}(X) = b/((a-1)^2(a-2))$.

Definition B.7 t-Verteilung

Eine stetige Zufallsvariable X heißt t-verteilt mit n Freiheitsgraden, in Zeichen $X \sim t_n$, wenn Sie folgende Dichte besitzt:

$$f(x) = \frac{\Gamma(n+1)/2}{\sqrt{n\pi}\Gamma(n/2)(1+x^2/n)^{(n+1)/2}}.$$

Erwartungswert und Varianz sind gegeben durch $\mathrm{E}(X) = 0$ (für $n > 1$) und $\mathrm{Var}(X) = n/(n-2)$ (für $n > 2$). Die t_1-Verteilung wird auch als Cauchy-Verteilung bezeichnet.
Sind X und Y unabhängig standardnormal- bzw. χ^2_n-verteilt, so gilt

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n.$$

Sind $X_1, \ldots, X_n$ unabhängig $\mathrm{N}(\mu, \sigma^2)$-verteilt, so gilt

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \sim t_{n-1}$$

mit

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{und} \quad \bar{X} = \sum_{i=1}^{n} X_i.$$

Definition B.8 F-Verteilung

Sind X_1 und X_2 unabhängig χ^2_n- bzw. χ^2_m-verteilt, so heißt

$$F = \frac{X_1/n}{X_2/m}$$

F-verteilt mit n und m Freiheitsgraden, in Zeichen $F \sim F_{n,m}$.

B.2 Zufallsvektoren

Definition B.9 Zufallsvektor

Der p-dimensionale Vektor $\boldsymbol{X} = (X_1, \ldots, X_p)'$ heißt Zufallsvektor oder p-dimensionale Zufallsvariable, wenn die Komponenten $X_1, \ldots, X_p$ eindimensionale Zufallsvariablen

sind. Der Zufallsvektor $\boldsymbol{X}$ heißt stetig, wenn es eine Funktion $f(\boldsymbol{x}) = f(x_1, \ldots, x_p) \geq 0$ gibt, so dass gilt

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_p \leq X_p \leq b_p) = \int_{a_p}^{b_p} \ldots \int_{a_1}^{b_1} f(x_1, \ldots, x_p)\, dx_1 \ldots dx_p.$$

Die Funktion f heißt (gemeinsame) Dichte von $\boldsymbol{X}$.
Der Zufallsvektor $\boldsymbol{X}$ heißt diskret, wenn $\boldsymbol{X}$ nur Werte aus einer höchstens abzählbaren Menge $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\} \subset \mathbb{R}^p$ annimmt. Die Funktion f mit

$$f(\boldsymbol{x}) = \begin{cases} P(\boldsymbol{X} = \boldsymbol{x}) & \boldsymbol{x} \in \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\} \\ 0 & \text{sonst} \end{cases}$$

heißt Wahrscheinlichkeitsfunktion oder diskrete Dichte von $\boldsymbol{X}$.

Definition B.10 Randverteilungen und bedingte Verteilungen

Gegeben sei der p-dimensionale Zufallsvektor $\boldsymbol{X} = (X_1, \ldots, X_p)'$ und eine Partition von $\boldsymbol{X}$ in den p_1-dimensionalen Vektor $\boldsymbol{X}_1$ und den p_2-dimensionalen Vektor $\boldsymbol{X}_2$, d.h. $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$. Dann heißt die p_1-dimensionale Dichte bzw. Wahrscheinlichkeitsfunktion $f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)$ von $\boldsymbol{X}_1$ Randdichte bzw. Randwahrscheinlichkeitsfunktion von $\boldsymbol{X}$. Die Randdichte bzw. Randwahrscheinlichkeitsfunktion ist gegeben durch

$$f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(\boldsymbol{x}_1, \boldsymbol{x}_2)\, dx_{p_1+1} \ldots dx_p$$

bei stetigen Zufallsvektoren bzw.

$$f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = \textstyle\sum_{\boldsymbol{x}_2} f(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

bei diskreten Zufallsvektoren. Die bedingte Dichte bzw. Wahrscheinlichkeitsfunktion von $\boldsymbol{X}_1$ bei fest vorgegebenem Wert $\boldsymbol{x}_2$ von $\boldsymbol{X}_2$ ist gegeben durch

$$f(\boldsymbol{x}_1|\boldsymbol{x}_2) = \begin{cases} \dfrac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)} & \text{für } f_{\boldsymbol{X}_2}(\boldsymbol{x}_2) > 0 \\ 0 & \text{sonst} \end{cases}$$

und einen analogen Ausdruck für die bedingte Dichte bzw. Wahrscheinlichkeitsfunktion von $\boldsymbol{X}_2$ gegeben $\boldsymbol{x}_1$.

Definition B.11 Erwartungswertvektor

Sei $\boldsymbol{X} = (X_1, \ldots, X_p)'$ ein p-dimensionaler Zufallsvektor. Dann heißt

$$\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)' = (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_p))'$$

Erwartungswertvektor von $\boldsymbol{X}$.

Definition B.12 Kovarianzmatrix, Korrelationsmatrix und Präzisionsmatrix

Die Kovarianzmatrix $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$ eines p-dimensionalen Zufallsvektors $\boldsymbol{X}$ ist definiert durch

$$\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{p1} & \dots & \sigma_{pp} \end{pmatrix},$$

wobei $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, $i \neq j$, die Kovarianz zwischen X_i und X_j ist und $\sigma_{ii} = \sigma_i^2 = \mathrm{Var}(X_i)$ die Varianz von X_i.
Ausgehend von der Kovarianzmatrix $\boldsymbol{\Sigma}$ ergibt sich die Korrelationsmatrix $\boldsymbol{R}$ als:

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \vdots & & & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix}$$

mit

$$\rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i) \cdot \mathrm{Var}(X_j)}}.$$

Sowohl $\boldsymbol{\Sigma}$ als auch $\boldsymbol{R}$ sind symmetrisch und positiv semidefinit. Ist $\boldsymbol{\Sigma}$ sogar positiv definit, so heißt die Inverse $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}$ Präzisionsmatrix.

Satz B.1 Rechenregeln für Erwartungswertvektoren und Kovarianzmatrizen

Seien $\boldsymbol{X}$ und $\boldsymbol{Y}$ Zufallsvektoren, $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{a}, \boldsymbol{b}$ geeignet dimensionierte Matrizen bzw. Vektoren sowie $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ und $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Dann gilt:

1. $\mathrm{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{E}(\boldsymbol{X}) + \mathrm{E}(\boldsymbol{Y})$.
2. $\mathrm{E}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A} \cdot \mathrm{E}(\boldsymbol{X}) + \boldsymbol{b}$.
3. $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}') - \mathrm{E}(\boldsymbol{X})\mathrm{E}(\boldsymbol{X})'$.
4. $\mathrm{Var}(\boldsymbol{a}'\boldsymbol{X}) = \boldsymbol{a}'\mathrm{Cov}(\boldsymbol{X})\boldsymbol{a} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \sigma_{ij}$.
5. $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}'$.
6. $\mathrm{E}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \mathrm{sp}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$.

Definition B.13 Empirische Mittelwerte, Kovarianzmatrix und Korrelationsmatrix

Sei $\boldsymbol{x}_1, \dots \boldsymbol{x}_n$ eine i.i.d. Stichprobe aus der Verteilung von $\boldsymbol{X}$. Dann lässt sich der Erwartungswertvektor durch die empirischen Mittelwerte schätzen:

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_p)' = (\bar{x}_1, \dots, \bar{x}_p)' = \bar{\boldsymbol{x}}$$

mit

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \bar{x}_j.$$

Für die Kovarianzmatrix erhält man als Schätzung die empirische Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}$ durch die empirischen Varianzen

$$\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

und die empirischen Kovarianzen

$$\hat{\sigma}_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

Die empirische Korrelationsmatrix $\hat{\boldsymbol{R}}$ ergibt sich analog durch die empirischen Korrelationskoeffizienten.

B.3 Die multivariate Normalverteilung

B.3.1 Definition und Eigenschaften

Definition B.14 Multivariate Normalverteilung

Ein p-dimensionaler Zufallsvektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ heißt multivariat normalverteilt, wenn $\boldsymbol{X}$ die Dichte

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right] \tag{B.1}$$

mit $\boldsymbol{\mu} \in \mathbb{R}^p$ und positiv semidefiniter $p \times p$ Matrix $\boldsymbol{\Sigma}$ besitzt.
Für die Herleitung von Posteriori-Verteilungen in Bayesianischen Modellen erweist es sich als zweckmäßig die Dichte der multivariaten Normalverteilung in etwas anderer Form darzustellen. Unter Vernachlässigung aller Faktoren in (B.1), die nicht von $\boldsymbol{x}$ abhängen, erhält man die Darstellung

$$\begin{aligned} f(\boldsymbol{x}) &\propto \exp\left(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \\ &= \exp\left(-\tfrac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\ &\propto \exp\left(-\tfrac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right). \end{aligned} \tag{B.2}$$

Für multivariat normalverteiltes $\boldsymbol{X}$ ist die Dichte also stets proportional zu (B.2).

Satz B.2 Erwartungswert und Kovarianzmatrix

Für den Erwartungswert und die Kovarianzmatrix von $\boldsymbol{X}$ gilt $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ und $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Wir verwenden daher die Bezeichnung

$$\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

die bereits für die eindimensionale Normalverteilung geläufig ist. Häufig wird der Index p unterdrückt, wenn sich die Dimension aus dem Zusammenhang erschließen lässt. Für $\boldsymbol{\mu} = \boldsymbol{0}$ und $\boldsymbol{\Sigma} = \boldsymbol{I}$ spricht man von der (multivariaten) Standardnormalverteilung.

Satz B.3 Lineare Transformationen

Sei $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{d} \in \mathbb{R}^q$ und $\boldsymbol{D}$ eine $q \times p$ Matrix. Dann gilt für $\boldsymbol{Y} = \boldsymbol{d} + \boldsymbol{D}\boldsymbol{X}$:

$$\boldsymbol{Y} \sim \mathrm{N}_q(\boldsymbol{d} + \boldsymbol{D}\boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}').$$

Im Folgenden stellen wir wichtige Eigenschaften der multivariaten Normalverteilung zusammen. Dabei werden wir die zwei Fälle $\mathrm{rg}(\boldsymbol{\Sigma}) = p$ und $\mathrm{rg}(\boldsymbol{\Sigma}) < p$ unterscheiden. Im Fall $\mathrm{rg}(\boldsymbol{\Sigma}) < p$ spricht man von einer uneigentlichen, singulären oder entarteten Normalverteilung. Diesen Fall wollen wir zunächst außer Acht lassen und nehmen an, dass $\mathrm{rg}(\boldsymbol{\Sigma}) = p$, d.h. die Kovarianzmatrix ist positiv definit und invertierbar. Auf den singulären Fall gehen wir dann im Anschluss ein.

Satz B.4 Randverteilungen und bedingte Verteilungen

Sei $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ multivariat normalverteilt. Betrachte die Partition von $\boldsymbol{X}$ in die beiden Subvektoren $\boldsymbol{Y} = (X_1, \ldots, X_r)'$ und $\boldsymbol{Z} = (X_{r+1}, \ldots, X_p)'$, d.h.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{Z} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_Z \end{pmatrix}.$$

Dann ist der Subvektor $\boldsymbol{Y}$ wieder r-dimensional normalverteilt mit $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$.
Die bedingte Verteilung von $\boldsymbol{Y}$ gegeben $\boldsymbol{Z}$ ist ebenfalls eine multivariate Normalverteilung mit Erwartungswert

$$\boldsymbol{\mu}_{Y|Z} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YZ} \cdot \boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_Z)$$

und Kovarianzmatrix

$$\boldsymbol{\Sigma}_{Y|Z} = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\Sigma}_{ZY}.$$

Darüberhinaus gilt bei normalverteilten Zufallsvariablen die Äquivalenz zwischen Unabhängigkeit und Unkorreliertheit. $\boldsymbol{Y}$ und $\boldsymbol{Z}$ sind genau dann unabhängig, wenn $\boldsymbol{Y}$ und $\boldsymbol{Z}$ unkorreliert sind, d.h. $\boldsymbol{\Sigma}_{YZ} = \boldsymbol{\Sigma}_{ZY} = \boldsymbol{0}$ gilt. Für nicht normalverteilte Zufallsvektoren gilt die Äquivalenz zwischen Unabhängigkeit und Unkorreliertheit im Allgemeinen nicht. Hier folgt lediglich aus der Unabhängigkeit die Unkorreliertheit.

B.3.2 Die singuläre Normalverteilung

Definition B.15 Singuläre Normalverteilung

Sei $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Die Verteilung von $\boldsymbol{X}$ heißt singulär, wenn $\mathrm{rg}(\boldsymbol{\Sigma}) < p$ gilt. In diesem Fall wird die Verteilung häufig auch durch die Präzisionsmatrix $\boldsymbol{P}$ (mit $\mathrm{rg}(\boldsymbol{P}) < p$) anstelle der Kovarianzmatrix charakterisiert. Der Zufallsvektor $\boldsymbol{X}$ besitzt dann die Dichte

$$f(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{P}(\boldsymbol{x} - \boldsymbol{\mu})\right]$$

die nur bis auf Proportionalität definiert ist. Diese Form der singulären Normalverteilung verwenden wir insbesondere in den Kapiteln 7 und 8.

Satz B.5 Charakterisierung der singulären Normalverteilung

Sei der p-dimensionale Zufallsvektor $\boldsymbol{X}$ singulär normalverteilt, d.h. $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, mit $\mathrm{rg}(\boldsymbol{\Sigma}) = r < p$. Sei $(\boldsymbol{G}\ \boldsymbol{H})$ eine orthogonale Matrix, wobei die Spalten der $p \times r$ Matrix $\boldsymbol{G}$ eine Basis des Spaltenraums von $\boldsymbol{\Sigma}$ und die Spalten von $\boldsymbol{H}$ eine Basis des Nullraums von $\boldsymbol{\Sigma}$ bilden. Betrachte die Transformation

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{G} & \boldsymbol{H} \end{pmatrix}' \boldsymbol{X} = \begin{pmatrix} \boldsymbol{G}'\boldsymbol{X} \\ \boldsymbol{H}'\boldsymbol{X} \end{pmatrix}$$

Dann ist $\boldsymbol{Y}_1$ der stochastische Anteil von $\boldsymbol{X}$ und nicht singulär mit

$$\boldsymbol{Y}_1 \sim \mathrm{N}(\boldsymbol{G}'\boldsymbol{\mu}, \boldsymbol{G}'\boldsymbol{\Sigma}\boldsymbol{G}).$$

$\boldsymbol{Y}_2$ ist der deterministische Anteil von $\boldsymbol{X}$ mit

$$\mathrm{E}(\boldsymbol{Y}_2) = \boldsymbol{H}'\boldsymbol{\mu} \qquad \text{und} \qquad \mathrm{Var}(\boldsymbol{Y}_2) = 0$$

Die Dichte f des stochastischen Anteils $\boldsymbol{Y}_1 = \boldsymbol{G}'\boldsymbol{X}$ hat die Gestalt

$$f(\boldsymbol{y}_1) = \frac{1}{(2\pi)^{\frac{r}{2}} (\prod_{i=1}^{r} \lambda_i)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{G}'\boldsymbol{\mu})'(\boldsymbol{G}'\boldsymbol{\Sigma}\boldsymbol{G})^{-1}(\boldsymbol{y}_1 - \boldsymbol{G}'\boldsymbol{\mu})\right] \tag{B.3}$$

Dabei sind λ_i die r von Null verschiedenen Eigenwerte und $\boldsymbol{\Sigma}^-$ eine g-Inverse von $\boldsymbol{\Sigma}$.

B.3.3 Verteilungen quadratischer Formen

Verteilungen quadratischer Formen normalverteilter Zufallsvektoren spielen beim Testen linearer Hypothesen eine Rolle, vergleiche insbesondere Abschnitt 3.3 in Kapitel 3.

Satz B.6 Verteilungen quadratischer Formen

1. Sei $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mit $\boldsymbol{\Sigma} > 0$. Dann gilt:
$$Y = (\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_p.$$
2. Sei $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{B}$ eine $n \times p$ $(n \leq p)$ Matrix und $\boldsymbol{R}$ eine symmetrische, idempotente $p \times p$ Matrix mit $\mathrm{rg}(\boldsymbol{R}) = r$. Dann gilt:
   - $\boldsymbol{X}'\boldsymbol{R}\boldsymbol{X} \sim \chi^2_r$.
   - Aus $\boldsymbol{B}\boldsymbol{R} = \boldsymbol{0}$ folgt die Unabhängigkeit der quadratischen Form $\boldsymbol{X}'\boldsymbol{R}\boldsymbol{X}$ von der Linearform $\boldsymbol{B}\boldsymbol{X}$.
3. Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit $X_i \sim \mathrm{N}(\mu, \sigma^2)$ und
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$
   Dann gilt:
   - $\dfrac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$.
   - S^2 und $\bar{X}$ sind unabhängig.
4. Seien $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{R}$ und $\boldsymbol{S}$ symmetrische und idempotente $n \times n$ Matrizen mit $\mathrm{rg}(\boldsymbol{R}) = r$ und $\mathrm{rg}(\boldsymbol{S}) = s$ und $\boldsymbol{R}\boldsymbol{S} = \boldsymbol{0}$. Dann gilt:
   - $\boldsymbol{X}'\boldsymbol{R}\boldsymbol{X}$ und $\boldsymbol{X}'\boldsymbol{S}\boldsymbol{X}$ sind unabhängig.
   - $\dfrac{s}{r} \dfrac{\boldsymbol{X}'\boldsymbol{R}\boldsymbol{X}}{\boldsymbol{X}'\boldsymbol{S}\boldsymbol{X}} \sim F_{r,s}$.

B.3.4 Multivariate t-Verteilung

Ein p-dimensionaler Zufallsvektor $\boldsymbol{X} = (X_1, \ldots, X_p)'$ heißt multivariat t-verteilt mit ν Freiheitgraden, Lokationsparameter $\boldsymbol{\mu}$ und (positiv definiter) Dispersionsmatrix $\boldsymbol{\Sigma}$, wenn er die Dichte

$$f(x) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} (\nu\pi)^{-\frac{p}{2}} \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)} \left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-(\nu+p)/2}$$

besitzt. Jeder Teilvektor von $\boldsymbol{X}$ ist wieder t-verteilt mit ν Freiheitsgraden sowie den entsprechenden Subvektoren bzw. -matrizen von $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ als Lokations- bzw. Dispersionsparameter.

In Definition B.7 wurde eine konstruktive Definition der t-Verteilung basierend auf einer standardnormalverteilten und einer χ^2-verteilten Zufallsvariablen angegeben. Analog lässt sich die multivariate t-Verteilung basierend auf einer multivariat normalverteilten und einer χ^2-verteilten Zufallsvariablen definieren.

B.4 Likelihood-Inferenz

In diesem Abschnitt beschreiben wir die Maximum-Likelihood (ML)-Methode zur Schätzung unbekannter Parameter in statistischen Modellen und darauf beruhende Tests zu Hypothesen über diese Parameter.

B.4.1 Maximum-Likelihood-Schätzung

Gegeben sei eine Stichprobe $Y_1, \ldots, Y_n$ mit realisierten Werten $y_1, \ldots, y_n$. Die gemeinsame Wahrscheinlichkeit (für diskrete $Y_1, \ldots, Y_n$)

$$P(Y_1 = y_1, \ldots, Y_n = y_n \,|\, \boldsymbol{\theta})$$

bzw. der Wert der gemeinsamen Dichte (für stetige $Y_1, \ldots, Y_n$)

$$f(Y_1 = y_1, \ldots, Y_n = y_n \,|\, \boldsymbol{\theta})$$

der realisierten Werte hänge von einem unbekannten, zu schätzenden Parametervektor $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \in \boldsymbol{\Theta}$ ab. Diese Wahrscheinlichkeit bzw. dieser Wert der Dichte aufgefasst als Funktion des unbekannten Parameters $\boldsymbol{\theta}$ heißt Likelihood und wird mit $L(\boldsymbol{\theta})$ bezeichnet:

$$L(\boldsymbol{\theta}) = P(Y_1 = y_1, \ldots, Y_n = y_n \,|\, \boldsymbol{\theta})$$

bzw.

$$L(\boldsymbol{\theta}) = f(Y_1 = y_1, \ldots, Y_n = y_n \,|\, \boldsymbol{\theta}).$$

Gemäß dem *Maximum-Likelihood-Prinzip* wird der *Maximum-Likelihood-(ML-) Schätzer* $\hat{\boldsymbol{\theta}}$ für $\boldsymbol{\theta}$ so bestimmt, dass $\hat{\boldsymbol{\theta}}$ die Likelihood maximiert.

Im diskreten Fall besagt das ML-Prinzip also: Bestimme den ML-Schätzer $\hat{\boldsymbol{\theta}}$ so, dass die Wahrscheinlichkeit für das Auftreten der beobachteten Stichprobe $y_1, \ldots, y_n$ maximal

wird und somit diese Stichprobe möglichst plausibel (likely) ist. Im stetigen Fall soll in analoger Weise die Dichte maximal und damit die Stichprobe möglichst plausibel werden.

In der Regel, insbesondere für alle Modelle dieses Buches, ist die Likelihood $L(\boldsymbol{\theta})$ differenzierbar bzgl. $\boldsymbol{\theta}$ und das Maximum kann durch differenzieren, Null setzen und anschließendes Auflösen des resultierenden Gleichungssystems bezüglich $\boldsymbol{\theta}$ bestimmt werden. Aus technischen Gründen wird üblicherweise nicht die Likelihood maximiert, sondern die sogenannte Log-Likelihood $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$, die man durch Logarithmieren der Likelihood erhält. Da der Logarithmus eine streng monotone Funktion ist, besitzt $l(\boldsymbol{\theta})$ das selbe Maximum wie $L(\boldsymbol{\theta})$.

Die einfachste Situation der ML-Schätzung liegt vor, wenn Y eine Zufallsvariable mit Wahrscheinlichkeitsfunktion bzw. Dichte $f(y \mid \theta)$ ist, die einen unbekannten Parametervektor $\boldsymbol{\theta}$ enthält. Zur ML-Schätzung wird eine i.i.d. Stichprobe $Y_1, \ldots, Y_n$ aus $f(y \mid \boldsymbol{\theta})$ gezogen. Dann ist die gemeinsame Dichte wegen der Unabhängigkeit der $Y_1, \ldots, Y_n$ das Produkt der einzelnen Dichten $f(y_i \mid \theta)$. Somit ist die Likelihood durch

$$L(\boldsymbol{\theta}) = f(y_1 \mid \boldsymbol{\theta}) \cdot \ldots \cdot f(y_n \mid \boldsymbol{\theta})$$

und die Log-Likelihood durch die Summe

$$l(\boldsymbol{\theta}) = \log f(y_1 \mid \boldsymbol{\theta}) + \ldots + \log f(y_n \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta})$$

mit den Log-Likelihood-Beiträgen $l_i(\boldsymbol{\theta}) = \log f(y_i \mid \boldsymbol{\theta})$ gegeben. Diese Situation liegt in einfacher Form im folgenden Beispiel vor.

Beispiel B.1 Poisson-Verteilung – ML-Schätzer

Gegeben sei eine Poisson-verteilte Zufallsvariable $Y \sim \mathrm{Po}(\lambda)$ mit unbekanntem Parameter λ. Zur Schätzung von λ ziehen wir eine i.i.d. Stichprobe $Y_1, \ldots, Y_n$ mit $Y_i \sim \mathrm{Po}(\lambda)$ und realisierten Werten $y_1, \ldots, y_n$. Den ML-Schätzer für λ erhält man in vier Schritten:

1. Schritt: Bestimmung der Likelihood. Da es sich um eine i.i.d. Stichprobe handelt, ist die gemeinsame Wahrscheinlichkeit für die realisierte Stichprobe das Produkt der einzelnen Wahrscheinlichkeiten und wir erhalten für die Likelihood

$$\begin{aligned} L(\lambda) &= P(Y_1 = y_1, \ldots, Y_n = y_n \mid \lambda) \\ &= P(Y_1 = y_1 \mid \lambda) \cdot \ldots \cdot P(Y_n = y_n \mid \lambda) \\ &= \frac{\lambda^{y_1}}{y_1!} \exp(-\lambda) \cdot \ldots \cdot \frac{\lambda^{y_n}}{y_n!} \exp(-\lambda) \end{aligned}$$

Da die Faktoren $1/y_1!, \ldots, 1/y_n!$ nicht von λ abhängen, können wir sie bei der Maximierung vernachlässigen und erhalten

$$L(\lambda) \propto \exp(-n\lambda) \cdot \lambda^{y_1} \cdot \ldots \cdot \lambda^{y_n}.$$

Oft bezeichnet man auch bereits die rechte Seite unter Vernachlässigung konstanter Faktoren als Likelihood und schreibt einfach

$$L(\lambda) = \exp(-n\lambda) \cdot \lambda^{y_1} \cdot \ldots \cdot \lambda^{y_n}.$$

2. Schritt: Bestimmung der Log-Likelihood. Durch Logarithmieren erhält man

$$l(\lambda) = -n\lambda + \sum_{i=1}^{n} y_i \log(\lambda).$$

3. Schritt: Differenzieren und Nullsetzen.

$$\frac{\partial\, l(\lambda)}{\partial \lambda} = -n + \sum_{i=1}^{n} y_i \frac{1}{\lambda} = 0$$

Die erste Ableitung $s(\lambda) = \partial\, l(\lambda)/\partial\, \lambda$ nennt man *Score-Funktion*.

4. Schritt: Auflösen nach dem unbekannten Parameter liefert den ML-Schätzer

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}.$$

5. Schritt: Zweite Ableitung zur Überprüfung des Maximums. Die zweite Ableitung der Log-Likelihood ist gegeben durch

$$\frac{\partial^2\, l(\lambda)}{\partial \lambda^2} = -\sum_{i=1}^{n} y_i \frac{1}{\lambda^2} < 0,$$

d.h. bei dem im 4. Schritt gefundenen Schätzer handelt es sich tatsächlich um ein Maximum der Log-Likelihood bzw. der Likelihood. Die *negative* zweite Ableitung $-\partial^2\, l(\lambda)/\partial\, \lambda^2$ heißt *beobachtete Fisher-Information*.

△

Die im vorausgegangenen Beispiel erwähnte Score-Funktion und Fisher-Information spielen in der Likelihood-Theorie eine wichtige Rolle. Die Score-Funktion ist allgemein definiert als der Vektor der partiellen Ableitungen der Log-Likelihood

$$\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \right)' .$$

Die beobachtete Fisher-Information ist definiert durch

$$\boldsymbol{H}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_p} \end{pmatrix} .$$

Die beobachtete Fisher-Informationsmatrix kann, wie der Name bereits andeutet, als (lokales) Maß für die Information, welche die Likelihood über den unbekannten Parameter liefert, betrachtet werden. Die zweite Ableitung einer Funktion ist ein Maß für die Krümmung an der Stelle $\boldsymbol{\theta}$. Je stärker die Krümmung der Log-Likelihood in der Nähe des Maximums, desto mehr Information über den unbekannten Parameter enthält die Log-Likelihood. Da die Krümmung in der Regel negativ ist, wird die Fisher-Information als die negative Krümmung definiert.

Da die Likelihood bzw. die Log-Likelihood von den Werten der Stichprobenvariablen $Y_1, \ldots, Y_n$ abhängt, wird die Likelihoodfunktion bei wiederholter Stichprobenziehung jeweils eine andere Gestalt annehmen. Likelihood, Log-Likelihood und deren Ableitungen können also auch als Zufallsvariablen aufgefasst werden, die insbesondere auch Erwartungswert und Varianz besitzen. Der Erwartungswert der Score-Funktion ist stets gleich $\mathbf{0}$, d.h.

$$\mathrm{E}(\boldsymbol{s}(\boldsymbol{\theta})) = \mathbf{0}.$$

Der Erwartungswert wird dabei bezüglich der Stichprobenvariablen $Y_1, \ldots, Y_n$ gebildet. Von besonderem Interesse ist der Erwartungswert der Fisher-Information:

$$\boldsymbol{F}(\boldsymbol{\theta}) = \mathrm{E}\left(-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = \mathrm{Cov}(\boldsymbol{s}(\boldsymbol{\theta})) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{\theta})').$$

Die Matrix $\boldsymbol{F}(\boldsymbol{\theta})$ wird als erwartete Fisher-Information bezeichnet und kann als globales Informationsmaß angesehen werden, da sie unabhängig von der Stichprobe ist.

Für unabhängige Stichproben ist die Information additiv: Seien $\boldsymbol{H}_{Y_i}(\boldsymbol{\theta})$ und $\boldsymbol{F}_{Y_i}(\boldsymbol{\theta})$ beobachtete und erwartete Fisherinformation bezüglich der i-ten Stichprobenvariablen, d.h. die Information, wenn nur die i-te Ziehung berücksichtigt wird. Dann ist die Information bezüglich der gesamten Stichprobe $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ gegeben durch

$$H_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} H_{Y_i}(\boldsymbol{\theta}) \quad \text{bzw.} \quad F_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} F_{Y_i}(\boldsymbol{\theta}).$$

Beispiel B.2 Poisson-Verteilung – Score-Funktion und Fisher-Information

Gegeben sei eine i.i.d. Stichprobe $Y_1, \ldots, Y_n$ einer Poisson-verteilten Zufallsvariablen $Y \sim \mathrm{Po}(\lambda)$ mit unbekannten Parameter λ. Log-Likelihood und Score-Funktion der i-ten Beobachtung sind gegeben durch (vergleiche Beispiel B.1)

$$l_i(\lambda) = y_i \log(\lambda) - \log(y_i!) - \lambda \qquad \text{und} \qquad s_i(\lambda) = \frac{y_i}{\lambda} - 1.$$

Anhand des Beispiels verifizieren wir, dass $\mathrm{E}(s(\lambda)) = 0$ gilt. Wir erhalten

$$\mathrm{E}(s(\lambda)) = \mathrm{E}\left(\sum_{i=1}^{n}\left(\frac{Y_i}{\lambda} - 1\right)\right) = \frac{1}{\lambda}\sum_{i=1}^{n}\mathrm{E}(Y_i) - n = \frac{1}{\lambda}n\lambda - n = 0.$$

Als beobachtete Fisherinformation für die i-te Beobachtung berechnen wir

$$H_{y_i}(\lambda) = -\frac{\partial^2 l_i(\lambda)}{\partial^2\lambda} = -\frac{\partial s_i(\lambda)}{\partial\lambda} = -\left(-\frac{y_i}{\lambda^2}\right) = \frac{y_i}{\lambda^2}.$$

Aufgrund der Additivität der Information erhalten wir dann für die gesamte Stichprobe

$$H_{\boldsymbol{Y}}(\lambda) = \frac{1}{\lambda^2}\sum_{i=1}^{n} y_i.$$

Für die i-te Stichprobenvariable ergibt sich mit $\mathrm{E}(Y_i) = \lambda$

$$F_{Y_i}(\lambda) = \mathrm{E}\left(\frac{Y_i}{\lambda^2}\right) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

und wiederum aufgrund der Additivität der Information

$$F_{\boldsymbol{Y}}(\lambda) = \frac{n}{\lambda}.$$

Die erwartete Information wächst also linear mit dem Stichprobenumfang n. Darüber hinaus ist die erwartete Information umso kleiner, je größer der unbekannte Parameter λ ist. Intuitiv ist das klar, denn die Varianz des ML-Schätzers $\overline{Y}$ für λ ist $\mathrm{Var}(\overline{Y}) = \frac{1}{n}\lambda$, d.h. die Variabilität der Schätzung nimmt mit wachsendem λ zu und damit die Information ab.

△

In diesem Buch liegt meistens eine Regressionssituation mit Zielvariablen $Y_1, \ldots, Y_n$ vor, zu denen Werte $y_1, \ldots, y_n$ und zusätzliche Kovariablen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ als Stichprobe gegeben sind. Der Parametervektor $\boldsymbol{\theta}$ ist dann oft der Vektor $\boldsymbol{\beta}$ der Kovariableneffekte. Die $Y_1, \ldots, Y_n$ sind (bei gegebenen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$) nun nicht mehr identisch verteilt, werden aber weiterhin als unabhängig angenommen. Die Wahrscheinlichkeitsfunktion bzw. Dichten der Y_i hängen über $\boldsymbol{x}_i$ von i ab, d.h. es ist $Y_i \sim f_i(y_i \,|\, \boldsymbol{\theta}) = f(y_i \,|\, \boldsymbol{x}_i; \boldsymbol{\beta})$ mit $\boldsymbol{\theta} = \boldsymbol{\beta}$. Die Likelihood ist dann weiter das Produkt

$$L(\boldsymbol{\theta}) = f_1(y_1 \,|\, \boldsymbol{\theta}) \cdot \cdots \cdot \cdots \cdot f_n(y_n \,|\, \boldsymbol{\theta})$$

der einzelnen Dichten und die Log-Likelihood

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(y_i \,|\, \boldsymbol{\theta})$$

die Summe der Log-Likelihood-Beiträge $l_i(\boldsymbol{\theta}) = \log f_i(y_i \,|\, \boldsymbol{\theta})$.

Beispiel B.3 Poisson-Regression

Wir betrachten das folgende Poisson-Regressionsmodell:

$$\begin{aligned} y_i &\sim \mathrm{Po}(\lambda_i) \qquad i = 1, \ldots, n, \\ \lambda_i &= \exp(\beta_0 + \beta_1 x_i) = \exp(\eta_i), \\ \eta_i &= \beta_0 + \beta_1 x_i. \end{aligned}$$

Unser Ziel ist die Bestimmung des ML-Schätzers für β_0 und β_1. Dabei wird sich herausstellen, dass eine analytische Lösung, wie im vorausgegangenen Beispiel B.1, nicht existiert. Vielmehr sind wir auf numerische Verfahren zur Optimierung angewiesen, die wir im nächsten Abschnitt besprechen. Die Bestimmung des Schätzers erfolgt durch folgende Schritte:

1. Schritt: Bestimmung der Likelihood. Die Likelihood der i-ten Beobachtung ist unter Vernachlässigung von Faktoren, die nicht von β_0 und β_1 abhängen, gegeben durch

   $$L_i(\beta_0, \beta_1) = \lambda_i^{y_i} \exp(-\lambda_i).$$

   Die Likelihood für die Stichprobe ist das Produkt der individuellen Likelihoods:

   $$L(\beta_0, \beta_1) = \prod_{i=1}^{n} L_i(\beta_0, \beta_1).$$

2. Schritt: Bestimmung der Log-Likelihood. Durch Logarithmieren der individuellen Likelihood erhalten wir

   $$l_i(\beta_0, \beta_1) = y_i \log(\lambda_i) - \lambda_i = y_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i)$$

   und damit

   $$l(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i)).$$

3. Schritt: Bestimmung der Score-Funktion. Die partiellen Ableitungen der individuellen Log-Likelihood bezüglich β_0 und β_1 sind gegeben durch

   $$\frac{\partial\, l_i(\beta_0, \beta_1)}{\partial \beta_0} = y_i - \exp(\beta_0 + \beta_1 x_i) = y_i - \lambda_i,$$

$$\frac{\partial\, l_i(\beta_0,\beta_1)}{\partial \beta_1} = y_i x_i - \exp(\beta_0 + \beta_1 x_i)\cdot x_i = x_i(y_i - \lambda_i).$$

Unter Verwendung der Vektoren $\boldsymbol{y} = (y_1,\ldots,y_n)'$, $\boldsymbol{x} = (x_1,\ldots,x_n)'$, $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_n)'$ und $\mathbf{1} = (1,\ldots,1)'$ erhalten wir damit für die Score-Funktion

$$\boldsymbol{s}(\beta_0,\beta_1) = \begin{pmatrix} \sum_{i=1}^n (y_i - \lambda_i) \\ \sum_{i=1}^n x_i(y_i - \lambda_i) \end{pmatrix} = \begin{pmatrix} \mathbf{1}'(\boldsymbol{y} - \boldsymbol{\lambda}) \\ \boldsymbol{x}'(\boldsymbol{y} - \boldsymbol{\lambda}) \end{pmatrix}.$$

Schließlich ergibt sich mit der Designmatrix $\boldsymbol{X} = (\mathbf{1}\ \boldsymbol{x})$ für die Score-Funktion

$$\boldsymbol{s}(\beta_0,\beta_1) = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\lambda}).$$

Wenn wir die Score-Funktion Null setzen erhalten wir ein nichtlineares Gleichungssystem, das mit numerischen Verfahren gelöst werden muss, vergleiche hierzu den nächsten Abschnitt. Ein Baustein des numerischen Verfahrens ist die Fisher-Information, die deshalb im nächsten Schritt berechnet wird.

4. Schritt: Bestimmung der (beobachteten) Fisher-Informationsmatrix. Die zweiten partiellen Ableitungen der individuellen Log-Likelihood sind gegeben durch

$$\begin{aligned} \frac{\partial^2\, l_i(\beta_0,\beta_1)}{\partial \beta_0^2} &= -\exp(\beta_0 + \beta_1 x_i) = -\lambda_i, \\ \frac{\partial^2\, l_i(\beta_0,\beta_1)}{\partial \beta_1^2} &= -\exp(\beta_0 + \beta_1 x_i) x_i^2 = -\lambda_i x_i^2, \\ \frac{\partial^2\, l_i(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} &= -\exp(\beta_0 + \beta_1 x_i) x_i = -\lambda_i x_i. \end{aligned}$$

Damit erhalten wir die beobachtete Informationsmatrix

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\frac{\partial^2\, l(\beta_0,\beta_1)}{\partial\, \boldsymbol{\beta} \partial\, \boldsymbol{\beta}'} = -\sum_{i=1}^n \frac{\partial^2\, l_i(\beta_0,\beta_1)}{\partial\, \boldsymbol{\beta} \partial\, \boldsymbol{\beta}'} = \begin{pmatrix} \sum_{i=1}^n \lambda_i & \sum_{i=1}^n \lambda_i x_i \\ \sum_{i=1}^n \lambda_i x_i & \sum_{i=1}^n \lambda_i x_i^2 \end{pmatrix}$$

als die mit -1 multiplizierte Matrix der zweiten partiellen Ableitungen. Mit der Matrix $\boldsymbol{W} = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$ erhalten wir kompakter

$$\boldsymbol{H}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}.$$

Da die beobachtete Informationsmatrix $\boldsymbol{H}(\boldsymbol{\beta})$ nicht mehr von $\boldsymbol{y}$ abhängt ist sie in diesem Beispiel gleich der erwarteten Informationsmatrix

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{y}}(\boldsymbol{H}(\boldsymbol{\beta})) = \boldsymbol{H}(\boldsymbol{\beta}).$$

Allgemein gilt jedoch $\boldsymbol{F}(\boldsymbol{\beta}) \neq \boldsymbol{H}(\boldsymbol{\beta})$.

5. Schritt: Iterative numerische Bestimmung des ML-Schätzers $\hat{\boldsymbol{\beta}}$ als Nullstelle der Score-Funktion $\boldsymbol{s}(\boldsymbol{\beta})$, d.h. als Lösung des nichtlinearen Gleichungssystems

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) \stackrel{!}{=} \mathbf{0}.$$

Die iterative Bestimmung erfolgt mit dem im nächsten Abschnitt skizzierten Newton- oder Fisher-Scoring-Algorithmus. △

Abschließend bemerken wir, dass das Likelihood-Prinzip auch allgemein für abhängige Variablen $Y_1, \ldots, Y_n$ anwendbar ist. Eine solche Situation liegt beispielsweise in Kapitel 6.1 (Seite 254) vor. Dort ist $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ multivariat normalverteilt

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\Sigma}(\boldsymbol{\alpha}))$$

mit einem von $\boldsymbol{\beta}$ abhängigen Erwartungswertvektor $\boldsymbol{\mu}(\boldsymbol{\beta})$ und einer von $\boldsymbol{\alpha}$ abhängigen, nicht diagonalen Kovarianzmatrix $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$. Die Likelihood $L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = L(\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ ist dann durch die Dichte der multivariaten Normalverteilung für $\boldsymbol{Y}$ gegeben. Diese zerfällt nun nicht mehr in das Produkt separater Dichten für die einzelnen Variablen $Y_1, \ldots, Y_n$.

B.4.2 Numerische Berechnung des ML-Schätzers

In den meisten Anwendungsfällen kann der Maximum-Likelihood-Schätzer nicht analytisch berechnet werden, da das nach dem Differenzieren der Score-Funktion resultierende nichtlineare Gleichungssystem nicht nach den unbekannten Parametern aufgelöst werden kann. Vergleiche hierzu etwa das vorangegangene Beispiel B.3. In diesen Fällen ist man auf numerische Verfahren zur Bestimmung der Nullstellen der Score-Funktion angewiesen. In der Literatur existieren eine Vielzahl von Verfahren, vergleiche z.B. Lange (2000) für einen Überblick. Wir wollen hier lediglich die beiden nach wie vor gängigsten Verfahren, das Newton- (bzw. Newton-Raphson-) Verfahren und das sogenannte Fisher-Scoring, behandeln.

Wir illustrieren das Newton-Raphson-Verfahren zunächst für eindimensionale Parameter. Ziel ist die (numerische) Bestimmung der Nullstelle der Score-Funktion, d.h. die Lösung der im Allgemeinen nichtlinearen Gleichung

$$s(\theta) = 0.$$

Das Newton-Raphson-Verfahren bestimmt ausgehend von einer Startlösung $\theta^{(0)}$ iterativ die Nullstelle der Score-Funktion. Dabei wird wie folgt vorgegangen (vergleiche auch Abbildung B.1): An der Stelle $\theta^{(0)}$ wird eine Tangente an die Score-Funktion gelegt und man erhält eine verbesserte Lösung $\theta^{(1)}$ als Nullstelle der angelegten Tangente. Die Tangente ist gegeben durch

$$g(\theta) = s\left(\theta^{(0)}\right) + s'\left(\theta^{(0)}\right) \cdot \left(\theta - \theta^{(0)}\right).$$

Es handelt sich dabei um die Taylorreihenentwicklung erster Ordnung von $s(\theta)$ im Entwicklungspunkt $\theta^{(0)}$. Die Nullstelle der Tangente liefert dazu die verbesserte Lösung

$$\theta^{(1)} = \theta^{(0)} - \frac{1}{s'\left(\theta^{(0)}\right)} \cdot s\left(\theta^{(0)}\right).$$

Da $-s'(\theta)$ die beobachtete Fisher-Information $H(\theta)$ ist, können wir auch

$$\theta^{(1)} = \theta^{(0)} + H\left(\theta^{(0)}\right)^{-1} s\left(\theta^{(0)}\right)$$

schreiben.

Ausgehend von $\theta^{(1)}$ erhalten wir dazu eine weiter verbesserte Lösung $\theta^{(2)}$, indem bei $\theta^{(1)}$ erneut eine Tangente an $s(\theta)$ angelegt wird und die Nullstelle bestimmt wird. Das Verfahren wird so lange iterativ fortgesetzt, bis sich die Lösungen nicht mehr ändern.

Abb. B.1. *Illustration des Newton-Verfahrens.*

Der beschriebene Algorithmus lässt sich auch auf mehrdimensionale Parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ übertragen. Ausgehend von einer aktuellen Lösung $\boldsymbol{\theta}^{(k)}$ erhält man eine verbesserte Lösung für $\boldsymbol{s}(\boldsymbol{\theta}) = \mathbf{0}$ durch

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \left(\frac{\partial \boldsymbol{s}\left(\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}'} \right)^{-1} \boldsymbol{s}\left(\boldsymbol{\theta}^{(k)}\right) = \boldsymbol{\theta}^{(k)} + \boldsymbol{H}\left(\boldsymbol{\theta}^{(k)}\right)^{-1} \boldsymbol{s}\left(\boldsymbol{\theta}^{(k)}\right).$$

Ersetzt man die beobachtete Informationsmatrix durch die erwartete Informationsmatrix $\boldsymbol{F}(\boldsymbol{\theta}^{(k)})$, so erhält man das *Fisher-Scoring-Verfahren*. Der Vorteil des modifizierten Algorithmus besteht darin, dass in vielen Fällen die Formel für die erwartete Informationsmatrix einfacher sind und damit weniger Rechenzeit zur Bestimmung des ML-Schätzers benötigt wird.

Beispiel B.4 Poisson-Regression

Wir setzen unser Beispiel B.3 zur Poisson-Regression fort. Gemäß dem Newton-Raphson-Verfahren lassen sich die Parameterschätzungen $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ durch folgendes Verfahren bestimmen:

1. Setze Startwerte $\hat{\boldsymbol{\beta}}^{(0)}$, z.B. $\hat{\boldsymbol{\beta}}^{(0)} = (0, 0)'$. Setze $k = 1$.
2. Erhalte verbesserte Schätzungen $\hat{\boldsymbol{\beta}}^{(k)}$ durch

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} + \boldsymbol{H}\left(\hat{\boldsymbol{\beta}}^{k-1}\right)^{-1} \boldsymbol{s}\left(\hat{\boldsymbol{\beta}}^{k-1}\right) = \hat{\boldsymbol{\beta}}^{(k-1)} + (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\lambda})$$

wobei die Gewichtsmatrix $\boldsymbol{W}$ und der Erwartungswertvektor $\boldsymbol{\lambda}$ von den aktuellen Parameterschätzungen $\hat{\boldsymbol{\beta}}^{(k-1)}$ abhängen, d.h. $\boldsymbol{W} = \boldsymbol{W}\left(\hat{\boldsymbol{\beta}}^{(k-1)}\right)$ und $\boldsymbol{\lambda} = \boldsymbol{\lambda}\left(\hat{\boldsymbol{\beta}}^{(k-1)}\right)$.

3. Falls für einen sehr kleinen Wert $\varepsilon > 0$

$$\frac{||\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}||}{||\hat{\boldsymbol{\beta}}^{(k-1)}||} \leq \varepsilon$$

beende das Verfahren. Andernfalls setze $k = k + 1$ und fahre fort mit 2.

△

B.4.3 Asymptotische Eigenschaften des ML-Schätzers

Unter schwachen Regularitätsvoraussetzungen (die im Detail vom zugrunde liegenden Modell $f(y_1, \ldots, y_n \,|\, \boldsymbol{\theta})$ abhängen), gelten folgende asymptotischen Eigenschaften: Für $n \to \infty$ bzw. für großen Stichprobenumfang ist der ML-Schätzer konsistent, asymptotisch (bzw. approximativ) erwartungstreu und asymptotisch bzw. approximativ multivariat normalverteilt

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\theta}, \hat{\boldsymbol{V}})$$

mit (geschätzter) Kovarianzmatrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{V}}$. Diese erhält man als Inverse der beobachteten oder erwarteten Informationsmatrix, d.h.

$$\hat{\boldsymbol{V}} = \boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}}) \quad \text{bzw.} \quad \hat{\boldsymbol{V}} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\theta}}),$$

ausgewertet an der Stelle $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ des ML-Schätzers. Beim Newton- bzw. Fisher-Scoring-Algorithmus werden diese Inversen im Schlussschritt berechnet. Die Diagonalelemente $\hat{v}_{jj}$ von $\hat{\boldsymbol{V}}$ sind dann die geschätzten Varianzen $\widehat{\mathrm{Var}}(\hat{\theta}_j)$ der j-ten Komponente von $\hat{\boldsymbol{\theta}}$, und $\sqrt{\hat{v}_{jj}}$ ist der geschätzte Standardfehler: $se_j = \sqrt{\hat{v}_{jj}}$.

Schließlich sind ML-Schätzer auch noch asymptotisch effizient, d.h. zumindest für sehr großen Stichprobenumfang besitzen sie kleinere Varianz als alle alternativen Schätzer.

Die Frage nach dem Stichprobenumfang n ab dem diese asymptotischen Eigenschaften hinreichend gut erfüllt sind, lässt sich nicht allgemein beantworten. Dieser hängt davon ab, wie gut die gemeinsame Dichte bzw. die Log-Likelihood an der Stelle $\hat{\boldsymbol{\theta}}$ des ML-Schätzers durch eine Normalverteilung bzw. durch eine quadratische Funktion approximiert werden kann. Der Stichprobenumfang n muss in jeden Fall ein Vielfaches (etwa 10 – 20-faches) der Anzahl p sein. Im Zweifelsfall sind Simulationsstudien notwendig.

B.4.4 Likelihood-basierte Tests für lineare Hypothesen

Wir beschränken uns auf das Testen linearer Hypothesen über $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \in \boldsymbol{\Theta}$

$$H_0 : \boldsymbol{C\theta} = \boldsymbol{d} \quad \text{gegen} \quad H_1 : \boldsymbol{C\theta} \neq \boldsymbol{d},$$

wobei die $(r \times p)$-Matrix $\boldsymbol{C}$ vollen Zeilenrang $r \leq p$ besitzt. In diesem Buch ist $\boldsymbol{\theta}$ meist der Parametervektor $\boldsymbol{\beta}$ linearer Effekte in Regressionsmodellen. Ein wichtiger Spezialfall sind Hypothesen über einen r-dimensionalen Teilvektor $\boldsymbol{\theta}_r$ von $\boldsymbol{\theta}$ von der Form

$$H_0 : \boldsymbol{\theta}_r = \boldsymbol{d} \quad \text{gegen} \quad H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{d}.$$

Wenn $\boldsymbol{\theta}_r = \boldsymbol{\beta}_r$ ein Teilvektor von linearen Effekten in einem Regressionsmodell und $\boldsymbol{d} = \mathbf{0}$ ist, dann testet man damit auf Signifikanz der zugehörigen Kovariablen. Zum Testen können die Likelihood-Quotienten-, Wald- und die Score-Statistik verwendet werden.

Die Idee für die *Likelihood-Quotienten-Statistik* beruht auf dem Likelihood-Quotienten

$$LQ = \frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})},$$

wobei $\hat{\boldsymbol{\theta}}$ der ML-Schätzer unter H_1, also ohne die lineare Restriktion der Nullhypothese, und $\tilde{\boldsymbol{\theta}}$ der ML-Schätzer unter der Gleichungsrestriktion $\boldsymbol{C\theta} = \boldsymbol{d}$ von H_0 ist. Der Quotient LQ vergleicht somit das Maximum $L(\tilde{\boldsymbol{\theta}})$ der Likelihood unter H_0 mit dem Maximum $L(\hat{\boldsymbol{\theta}})$ ohne diese Einschränkung auf H_0. Damit gilt in jedem Fall $LQ \geq 1$ und für Werte nahe bei 1 ist die Einschränkung von $\boldsymbol{\theta}$ auf H_0 nahezu ohne Auswirkungen. Somit wird man intuitiv H_0 nicht ablehnen. Größere Werte von LQ sprechen dagegen gegen die Nullhypothese.

Ähnlich wie beim ML-Schätzer geht man noch durch Logarithmieren über zur *Log-Likelihood-Quotienten-Statistik*

$$lq = 2\{l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})\} = -2\{l(\tilde{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}})\}$$

wodurch die Log-Likelihoodwerte $l(\hat{\boldsymbol{\theta}})$ und $l(\tilde{\boldsymbol{\theta}})$ verglichen werden. Falls das unrestringierte Maximum $l(\hat{\boldsymbol{\theta}})$ deutlich größer als $l(\tilde{\boldsymbol{\theta}})$ ist, wird lq groß und H_0 zugunsten von H_1 abgelehnt.

Zur Berechnung von $\tilde{\boldsymbol{\theta}}$ muss die Maximierung von $l(\boldsymbol{\theta})$ unter Beachtung der linearen Nebenbedingung $\boldsymbol{C\theta} = \boldsymbol{d}$ durchgeführt werden. Dieser rechentechnische Nachteil lässt sich mit der *Wald-Statistik*

$$w = (\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{d})'(\boldsymbol{C}\hat{\boldsymbol{V}}\boldsymbol{C}')^{-1}(\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{d})$$

vermeiden. Dabei ist $\hat{\boldsymbol{V}}$ die (geschätzte) Kovarianzmatrix von $\hat{\boldsymbol{\theta}}$ und somit $\boldsymbol{C}\hat{\boldsymbol{V}}\boldsymbol{C}'$ die (geschätzte) Kovarianzmatrix von $\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{d}$. Die Wald-Statistik misst die durch die inverse Kovarianzmatrix gewichtete Differenz zwischen $\boldsymbol{C}\hat{\boldsymbol{\theta}}$ und $\boldsymbol{d}$. Falls H_0 zutrifft, sollte dieser Wert klein sein. Große Werte von w sprechen für $H_1 : \boldsymbol{C\theta} \neq \boldsymbol{d}$. Für die spezielle Hypothese $H_0 : \boldsymbol{\theta}_r = \boldsymbol{d}$ ist

$$w = (\hat{\boldsymbol{\theta}}_r - \boldsymbol{d})'\hat{\boldsymbol{V}}_r^{-1}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{d})$$

mit $\hat{\boldsymbol{V}}_r = \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}_r)$ als (geschätzter) Kovarianzmatrix von $\hat{\boldsymbol{\theta}}_r$.

Die *Score-Statistik*

$$u = \boldsymbol{s}'(\tilde{\boldsymbol{\theta}})\boldsymbol{F}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{s}(\tilde{\boldsymbol{\theta}})$$

misst die gewichtete Distanz zwischen dem Wert $\mathbf{0} = \boldsymbol{s}(\hat{\boldsymbol{\theta}})$ der Score-Funktion an der Stelle $\hat{\boldsymbol{\theta}}$ und dem Wert $\boldsymbol{s}(\tilde{\boldsymbol{\theta}})$ ausgewertet am restringierten ML-Schätzer $\tilde{\boldsymbol{\theta}}$.

Wald-Tests sind dann rechnerisch vorteilhaft, wenn zu einem bereits geschätzten Modell ein Teilmodell getestet werden soll. Dann braucht das Teilmodell nicht zusätzlich geschätzt werden. Umgekehrt sind Score-Tests günstig, wenn ein geschätztes Modell gegen ein Obermodell getestet werden soll. Mit Wald- und Score-Tests können deshalb Rückwärts- bzw. Vorwärtsselektionen von Variablen numerisch effizient durchgeführt werden.

Unter ähnlich schwachen Voraussetzungen wie für die asymptotische Normalität des ML-Schätzers sind die drei Teststatistiken unter H_0 asymptotisch äquivalent und asymptotisch bzw. approximativ χ^2-verteilt mit r Freiheitsgraden:

$$lq, w, u \overset{a}{\sim} \chi^2_r.$$

Kritische Werte oder p-Werte werden über diese asymptotische Verteilung berechnet. Für mittleren Stichprobenumfang ist die Approximation durch die χ^2-Grenzverteilung in der Regel ausreichend. Für kleinere Stichprobenumfänge können sich die Werte der Teststatistiken jedoch deutlich unterscheiden.

B.4.5 Modellwahl

Will man zwischen mehreren konkurrierenden statistischen Modellen mit verschiedenen Prädiktoren und Parametern auswählen, muss ein Kompromiss zwischen möglichst guter Datenanpassung und zu großer Modellkomplexität, d.h. einer hohen effektiven Anzahl von Parametern getroffen werden. So wird etwa in linearen Regressionsmodellen das Bestimmtheitsmaß R^2 durch Einbeziehen zusätzlicher Kovariablen, Interaktionen etc. immer weiter erhöht, jedoch ist dies meist mit einer Überanpassung (overfitting) an den vorliegenden (Lern-) Datensatz verbunden und geht mit einem Verlust an Prognosefähigkeit und der Generalisierbarkeit für neue Daten einher. Deshalb müssen Kriterien zur reinen Modellanpassung, wie R^2 bzw. in der Likelihood-Inferenz die Devianz oder die Log-Likelihood, so modifiziert werden, dass eine Überanpassung an den Datensatz durch Bestrafung zu hoher Komplexität, d.h. einer zu hohen (effektiven) Anzahl von Parametern, vermieden wird. Im Rahmen der parametrischen Likelihood-Inferenz mit einem p-dimensionalen Parametervektor $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ ist das AIC (Akaike's Information Criterion) gegeben durch

$$\text{AIC} = -2\, l(\hat{\boldsymbol{\theta}}) + 2\, p,$$

wobei der Term $2\,p$ ein überparametrisiertes Modell bestraft. Bei der Wahl zwischen verschiedenen Modellen wird das Modell mit dem kleinsten AIC-Wert (und nicht mit der größten Log-Likelihood $l(\hat{\boldsymbol{\theta}})$!) bevorzugt. Statt des AIC wird oft auch der mit dem Stichprobenumfang n dividierte Wert AIC$/n$ verwendet.

Für non- und semiparametrische Modelle, wie für Glättungsverfahren in Kapitel 7, wird die fixe Dimension p durch die „effektive" Dimension

$$df = \text{sp}(\boldsymbol{S}),$$

(also die Spur der Glättungsmatrix) ersetzt (vgl. Abschnitt 7.1.8, Seite 345). Zudem existieren alternative Strafterme, wie zum Beispiel beim korrigierten AIC$_{corr}$

$$\text{AIC}_{corr} = -2\, l(\hat{\boldsymbol{\theta}}) + \frac{2n(df+1)}{n-df-2}.$$

Das AIC besitzt eine informationstheoretische Rechtfertigung, die sehr schön in dem Buch von Burnham & Anderson (2002) beschrieben ist.

B.5 Bayes-Inferenz

Mit der Entwicklung von computerintensiven, simulationsbasierten Markov-Chain-Monte-Carlo (MCMC) Methoden seit Anfang der 1990er Jahre kann die Bayes-Inferenz für statistische Modelle von realitätsnaher Komplexität eingesetzt werden. Wir stellen zunächst das Konzept der Bayes-Inferenz und anschließend MCMC-Verfahren vor.

B.5.1 Grundlagen der Bayes-Inferenz

Der fundamentale Unterschied zur Likelihood-basierten Inferenz besteht darin, dass die unbekannten Parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ keine festen deterministischen Größen sind, sondern als zufällig angenommen werden und eine *Priori-Verteilung* besitzen. Ein Bayesianisches Modell besteht daher aus zwei Teilen:

- *Priori-Verteilung:* Das (subjektive) Vorwissen über die unbekannten Parameter wird durch die Spezifikation einer Wahrscheinlichkeitsverteilung für den unbekannten Parameter $\boldsymbol{\theta}$ ausgedrückt. Diese Verteilung wird als Priori-Verteilung von $\boldsymbol{\theta}$ bezeichnet. Die Spezifikation einer Priori-Verteilung für $\boldsymbol{\theta}$ bedeutet nicht notwendigerweise, dass die unbekannten Parameter $\boldsymbol{\theta}$ tatsächlich stochastisch sind. Vielmehr drücken wir durch die Priori-Verteilung den Grad der Unsicherheit über den unbekannten Parameter vor Durchführung der statistischen Analyse aus. Die Dichte bzw. Wahrscheinlichkeitsfunktion der Priori-Verteilung bezeichnen wir im Folgenden mit $p(\boldsymbol{\theta})$.
- *Beobachtungsmodell:* Im sogenannten Beobachtungsmodell wird die bedingte Verteilung der beobachtbaren Größen, also der Stichprobenvariablen $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, bei gegebenem unbekanntem Parameter spezifiziert. Die Dichte dieser Verteilung ist proportional zur Likelihood $L(\boldsymbol{\theta})$ und wird im Folgenden mit $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ bezeichnet.

Basierend auf der Priori-Verteilung und dem Beobachtungsmodell können wir unter Zuhilfenahme des Satzes von Bayes die Verteilung von $\boldsymbol{\theta}$ nach Durchführung des Experiments, also die bedingte Verteilung von $\boldsymbol{\theta}$ bei gegebenen Beobachtungen $\boldsymbol{y} = (y_1, \ldots, y_n)'$, bestimmen. Wir erhalten

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = c \cdot p(\boldsymbol{y} \mid \boldsymbol{\theta})\; p(\boldsymbol{\theta}),$$

mit der Normierungskonstanten $c = [\int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}]^{-1}$. Diese Verteilung wird als *Posteriori-Verteilung* bezeichnet.

Beispiel B.5 Poisson-Verteilung

Wir betrachten eine Poisson-verteilte i.i.d. Stichprobe $Y_1, \ldots, Y_n$, d.h. $Y_i \sim \mathrm{Po}(\lambda)$. Der Parameter λ ist unbekannt und soll mit Bayesianischer Inferenz geschätzt werden. Die gemeinsame Wahrscheinlichkeit der beobachteten Stichprobe $\boldsymbol{y} = (y_1, \ldots, y_n)'$ ist wie in Beispiel B.1 gegeben durch

$$p(\boldsymbol{y} \mid \lambda) = \frac{1}{y_1! \cdots y_n!} \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda).$$

Als Priori-Verteilung für λ verwenden wir eine Gammaverteilung mit Parametern a und b, also $\lambda \sim \mathrm{G}(a, b)$. Eine Begründung für die spezielle Wahl der Priori-Verteilung geben wir weiter unten. Damit hat λ die Dichte

$$p(\lambda) = k\,\lambda^{a-1}\exp(-b\lambda)$$

mit $k = \frac{b^a}{\Gamma(a)}$. Als Posteriori-Verteilung erhält man

$$p(\lambda\,|\,\boldsymbol{y}) = \frac{p(\boldsymbol{y}\,|\,\lambda)\,p(\lambda)}{\int p(\boldsymbol{y}\,|\,\lambda)\,p(\lambda)\,d\lambda} = c\,\frac{1}{y_1!\cdots y_n!}\lambda^{\sum_{i=1}^n y_i}\exp(-n\lambda)\,k\lambda^{a-1}\exp(-b\lambda).$$

Bei der Bestimmung des Verteilungstyps können wir sämtliche Faktoren in der Posteriori-Dichte, die nicht von λ abhängen, ignorieren. Wir erhalten

$$p(\lambda\,|\,\boldsymbol{y}) \propto \lambda^{\sum_{i=1}^n y_i}\exp(-n\lambda)\,\lambda^{a-1}\exp(-b\lambda) = \lambda^{a+\sum_{i=1}^n y_i - 1}\exp(-(b+n)\lambda).$$

Damit besitzt die Posteriori-Dichte die Gestalt einer Gammaverteilung mit Parametern $a' = a + \sum_{i=1}^n y_i$ und $b' = b + n$, d.h.

$$\lambda\,|\,\boldsymbol{y} \sim G\left(a + \sum_{i=1}^n y_i, b + n\right).$$

Als Posteriori-Verteilung erhalten wir also denselben Verteilungstyp wie die Priori-Verteilung. Dieser wünschenswerte Fall kann in vielen (einfacheren) Bayesianischen Modellen erreicht werden. Wir sprechen in diesem Zusammenhang von der Priori-Verteilung als der zum Poisson-Stichprobenmodell konjugierten Verteilung, weil die Posteriori denselben Verteilungstyp besitzt wie die Priori-Verteilung.

△

Beispiel B.6 Bayesianisches Logit-Modell

Wir betrachten ein Logit-Modell mit einer Kovariablen x:

$$y_i = B(1, \pi_i), \quad \pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}, \quad \eta_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

Mit diesen Spezifikationen ist das Beobachtungsmodell gegeben durch

$$p(\boldsymbol{y}|\beta) \propto L(\boldsymbol{\beta}) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

wobei $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ der Vektor der Regressionskoeffizienten ist. Zur vollständigen Spezifikation eines Bayesianischen Modells benötigen wir eine Priori-Verteilung für β_0 und β_1. Da die geschätzten Regressionskoeffizienten in Regressionsmodellen häufig zumindest asymptotisch normalverteilt sind, liegt es nahe eine zweidimensionale Normalverteilung als Priori anzunehmen. Wir spezifizieren also für $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ die Priori-Verteilung

$$p(\boldsymbol{\beta}) \sim N(\boldsymbol{m}, \boldsymbol{M})$$

mit Priori-Erwartungswert $\boldsymbol{m}$ und Priori-Kovarianzmatrix $\boldsymbol{M}$. Wenn beispielsweise Ergebnisse einer vorangegangenen Untersuchung vorliegen, kann der dort erhaltene Punktschätzer für $\boldsymbol{m}$ und die geschätzte Kovarianzmatrix für $\boldsymbol{M}$ eingesetzt werden. Je nachdem wie weit die Ergebnisse zurückliegen, könnte man auch die Kovarianzmatrix mit einem Faktor $c > 1$ multiplizieren um die Varianzen in $\boldsymbol{M}$ zu vergrößern und damit die Priori-Information abzuschwächen.

In vielen Fällen stehen wir aber vor der Situation, dass keinerlei Vorwissen über die Regressionskoeffizienten vorliegt. Wenn wir die Varianzen in $\boldsymbol{M}$ immer mehr aufblähen, d.h. c immer weiter vergrößern, wird die Priori-Normalverteilungsdichte immer flacher

und nähert sich einer Gleichverteilung an. Für $c \to \infty$ ist die Priori-Dichte proportional zu einer Konstanten, d.h.

$$p(\boldsymbol{\beta}) \propto \text{const.}$$

Das Integral dieser Verteilung über $\mathbb{R}$ ist unendlich, so dass es sich bei $p(\boldsymbol{\beta})$ nicht um die Dichte einer Verteilung im eigentlichen Sinne handelt. Wir sprechen auch von einer uneigentlichen oder diffusen Verteilung. Trotzdem sind diffuse Verteilungen als Priori-Verteilungen in vielen Fällen zulässig, und zwar immer dann, wenn die gemäß dem Satz von Bayes resultierende Posteriori-Verteilung eine im Sinne der Wahrscheinlichkeitsrechnung eigentliche Verteilung ist. In einem Bayesianischen Logit-Modell ist dies der Fall falls der ML-Schätzer für $\boldsymbol{\beta}$ existiert. Als Posteriori Verteilung erhalten wir mit einer diffusen Priori-Verteilung die Dichte

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})}{\int p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})d\,\boldsymbol{\beta}} \propto p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

Obwohl die Posteriori-Verteilung eine eigentliche Verteilung ist, kann über den Verteilungstyp keine Aussage getroffen werden. Diese Beobachtung hat erhebliche Auswirkungen auf die statistische Inferenz, da Charakteristika der Posteriori-Verteilung analytisch nur schwer zugänglich sind. Ein auf der Ziehung von Zufallszahlen basierender Zugang über MCMC-Methoden ist hingegen möglich, siehe Abschnitt B.5.3 und insbesondere Beispiel B.8

△

Von der Posteriori-Verteilung gehen im Bayes-Ansatz sämtliche Inferenzschlüsse bezüglich der unbekannten Parameter $\boldsymbol{\theta}$ aus. Im folgenden Abschnitt beschreiben wir sowohl Bayesianische Punkt- als auch Intervallschätzer.

B.5.2 Punkt- und Intervallschätzer

Punktschätzer

Die üblichen Punktschätzer der Bayesianischen Inferenz sind Posteriori-Erwartungswert, Posteriori-Median und Posteriori-Modus. Alle drei Punktschätzer können unter entscheidungstheoretischen Gesichtspunkten diskutiert und gerechtfertigt werden, wir gehen darauf jedoch nicht genauer ein und stellen lediglich die Schätzer vor.

Der Posteriori-Erwartungswert als Punktschätzer ist gegeben durch

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta} \,|\, \boldsymbol{y}) = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \,|\, \boldsymbol{y})\, d\boldsymbol{\theta} = c \cdot \int \boldsymbol{\theta}\, p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Bei der analytischen oder numerischen Bestimmung des Posteriori-Erwartungswerts sind somit (unter Umständen hochdimensionale) Integrale zu berechnen. Die damit verbundenen Schwierigkeiten waren lange Zeit ein Hauptgrund für die geringe Akzeptanz Bayesianischer Methoden in der Praxis. Mit Hilfe simulationsbasierter Methoden insbesondere MCMC-Verfahren, halten sich die Schwierigkeiten aber mittlerweile in Grenzen. Hierzu mehr in Abschnitt B.5.3.

Als weiterer Punktschätzer $\hat{\boldsymbol{\theta}}$ für $\boldsymbol{\theta}$ kommt der Posteriori-Modus, also das globale Maximum der Posteriori-Verteilung in Frage:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Der Vorteil des Posteriori-Modus im Vergleich zum Posteriori-Erwartungswert besteht darin, dass die Normierungskonstante zur Berechnung nicht benötigt wird.

Da der Median robust gegenüber Ausreißern ist, wird er häufig gegenüber dem Erwartungswert als Punktschätzer für $\boldsymbol{\theta}$ bevorzugt.

Beispiel B.7 Poisson-Verteilung

In Beispiel B.5 haben wir als Posteriori-Verteilung für λ eine Gammaverteilung mit Parametern $a' = a + \sum_{i=1}^n x_i$ und $b' = b + n$ erhalten. Als Posteriori-Erwartungswert ergibt sich

$$\mathrm{E}(\lambda \,|\, \boldsymbol{x}) = \frac{a + \sum_{i=1}^n x_i}{b + n}.$$

Je kleiner a (im Vergleich zu $\sum x_i$) und b (im Vergleich zu n), desto mehr stimmt der Posteriori-Erwartungswert als Punktschätzer mit dem üblichen Maximum-Likelihood-Schätzer $\hat{\lambda} = \overline{x}$ für λ überein. Je stärker die Priori-Information, d.h. je größer a und b, desto mehr weichen Posteriori-Erwartungswert und der ML-Schätzer voneinander ab. △

Intervallschätzung

Punktschätzer reduzieren die Information in der Posteriori-Verteilung auf eine einzige Größe. Insbesondere die Genauigkeit der Schätzungen wird damit aber nicht erfasst. Natürliche Maße für die Genauigkeit sind die üblichen Kennzahlen für die Variabilität von Zufallsvariablen. Wird der Posteriori-Erwartungswert als Punktschätzer verwendet, so bietet sich die Posteriori-Varianz an. Im Falle des Medians als Punktschätzer kann der Interquartilsabstand als Maß für die Variabilität der Schätzung verwendet werden. Schließlich kann bei Verwendung des Posteriori-Modus die Krümmung der Posteriori-Dichte beim Modus, also die beobachtete Fisher-Information, herangezogen werden.

Neben den genannten Maßzahlen finden vor allem sogenannte Vertrauensintervalle oder Bayesianische Konfidenzintervalle als Analoga zu den frequentistischen Konfidenzintervallen zur Beurteilung der Variabilität Anwendung. Vertrauensintervalle bzw. allgemeiner Vertrauensbereiche sind wie folgt definiert:

Eine Teilmenge $C \subset \boldsymbol{\Theta}$ des Parameterraums heißt $(1-\alpha)$-Vertrauensbereich oder Bayesianisches Konfidenzintervall für $\boldsymbol{\theta}$, wenn

$$P(\boldsymbol{\theta} \in C \,|\, \boldsymbol{y}) \geq 1 - \alpha$$

ist. Ein Vertrauensbereich für $\boldsymbol{\theta}$ ist also so definiert, dass er (mindestens) die vorgegebene Wahrscheinlichkeitsmasse $1-\alpha$ der Posteriori-Verteilung enthält. Man beachte, dass Vertrauensbereiche eine völlig andere Interpretation als klassische Konfidenzbereiche besitzen. Beim klassischen Konfidenzbereich wird keine Wahrscheinlichkeitsaussage über den unbekannten Parameter $\boldsymbol{\theta}$ getroffen. Dieser ist ja im klassischen Inferenz-Konzept keine zufällige Größe. Vielmehr wird eine Wahrscheinlichkeitsaussage über die Stichprobe $\boldsymbol{y} = (y_1, \ldots, y_n)'$ getroffen. Der Konfidenzbereich $C(\boldsymbol{y})$ enthält den unbekannten, wahren Parameter $\boldsymbol{\theta}$ mit Wahrscheinlichkeit $1-\alpha$. Wird also das Schätzverfahren häufig wiederholt, so enthält der konstruierte Konfidenzbereich den unbekannten Parameter in ungefähr $(1-\alpha)\cdot 100$ Prozent der Fälle. Beim Bayesianischen Vertrauensbereich hingegen wird tatsächlich eine Wahrscheinlichkeitsaussage über den als zufällig angesehenen Parameter $\boldsymbol{\theta}$ getroffen. Der Vertrauensbereich umfasst $(1-\boldsymbol{\alpha})\cdot 100$ Prozent der Wahrscheinlichkeitsmasse der Posteriori-Verteilung, d.h. eine aus der Posteriori gezogene Zufallszahl $\boldsymbol{\theta}$ ist mit Wahrscheinlichkeit $1-\alpha$ im Vertrauensbereich enthalten.

Bayesianische Vertrauensbereiche sind mit analytischen Methoden oft schwer zu berechnen. Sie lassen sich aber leicht mit den im folgenden Abschnitt beschriebenen MCMC-Methoden zur Ziehung von Zufallszahlen aus der Posteriori-Verteilung bestimmen.

B.5.3 MCMC-Methoden

Die Hauptschwierigkeit bei der Anwendung Bayesianischer Methoden bestand lange Zeit darin, dass die Posteriori-Verteilung in vielen Fällen analytisch und numerisch unzugänglich war. Mit dem Aufkommen moderner sogenannter Markov-Chain-Monte-Carlo- (MCMC-) Methoden Anfang der 1990er Jahre konnten die Schwierigkeiten größtenteils gelöst werden. Heute sind nun selbst solche Probleme zugänglich, die mit Standardverfahren der Statistik bisher nicht gelöst werden konnten. Obwohl MCMC-Methoden bereits in den 1950er Jahren (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller 1953) vor allem unter Physikern bekannt waren, dauerte es bis Anfang der 1990er Jahre (Gelfand & Smith 1990), bis die Nützlichkeit und enorme Vielseitigkeit dieser Methoden insbesondere im Zusammenhang mit Bayesianischer Inferenz unter Statistikern Beachtung fand. Eine gute Einführung in die Methodik findet man etwa in Gilks, Richardson & Spiegelhalter (1996) und Green (2001). Hier beschränken wir uns lediglich auf die Darstellung der Grundidee und die Beschreibung der wichtigsten Algorithmen.

MCMC-Methoden erlauben das Ziehen von Zufallszahlen aus einer in der Regel analytisch unzugänglichen Posteriori-Verteilung (und prinzipiell auch aus jeder anderen Verteilung). Auf diesem Wege kann eine beliebig genaue Schätzung von Charakteristika der Posteriori-Verteilung wie Erwartungswert, Varianz oder Quantile bis hin zur Schätzung der Dichte der Verteilung vorgenommen werden. Der wichtigste Vorteil der MCMC-Methoden im Vergleich zu alternativen Methoden der Zufallszahlenziehung, gedacht sei hier z.B. an das Importance- oder Rejection-Sampling, besteht darin, dass man nicht auf Posteriori-Dichten niedriger Dimension beschränkt ist, sondern vielmehr hochkomplexe Dichten mit Dimension 1000 oder höher bewältigen kann. Ein weiterer Vorteil von MCMC-Methoden beruht auf der Tatsache, dass die Normierungskonstante der Posteriori-Verteilung, ein in der Regel hochdimensionales und damit auch numerisch unlösbares Integral, nicht bekannt sein muss.

Die Grundidee bei MCMC-Verfahren ist relativ einfach. Sei $\boldsymbol{\theta}$ der unbekannte Parametervektor in einem Bayesianischen Modell und $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ die Dichte die Posteriori-Verteilung. Wir beschränken uns in diesem Kapitel auf stetige Verteilungen. Anstelle aus der Dichte $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ direkt eine unabhängige Stichprobe zu ziehen, wird statt dessen eine Markovkette erzeugt, deren Iterationen des Übergangskerns gegen die interessierende Posteriori-Verteilung konvergieren. Auf diese Weise erhält man dann eine Stichprobe, die nach einer gewissen Konvergenzzeit, auch „*Burn In*“-Phase genannt, als (allerdings abhängige) Stichprobe aus der Posteriori-Verteilung angesehen werden kann. Bevor wir die dazu benötigten Algorithmen genauer besprechen, wollen wir noch einmal darauf hinweisen, dass MCMC-Verfahren nicht nur zur Simulation aus Posteriori-Verteilungen in der Bayes-Inferenz, sondern allgemein zur Simulation von Zufallszahlen aus komplexen Verteilungen verwendet werden können. Dazu ist in den folgenden Abschnitten lediglich die Posteriori-Verteilung $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ durch die entsprechende Dichte zu ersetzen.

Metropolis-Hastings-Algorithmus

Der Basis-Algorithmus, von dem dann alle weiteren Algorithmen abgeleitet sind, ist wie folgt konstruiert: Zunächst wird ein Startwert $\boldsymbol{\theta}^{(0)}$ festgelegt. Anstatt aus der interessierenden Posteriori-Dichte $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ direkt zu ziehen, wird in jeder Iteration des Algorithmus eine neue Zufallszahl $\boldsymbol{\theta}^*$ aus einer sogenannten *Vorschlagsdichte* q gezogen. In der Regel hängt die Vorschlagsdichte vom aktuellen Zustand $\boldsymbol{\theta}^{(t-1)}$ ab, d.h. $q = q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$. Die Wahl der Vorschlagsdichte ist dabei prinzipiell beliebig, jedoch sollte es sich um eine Verteilung handeln, aus der relativ leicht Zufallszahlen gezogen werden können. Da die Vorschlagsdichte nicht mit der Posteriori-Verteilung übereinstimmt, können nicht alle vorgeschlagenen Zufallszahlen $\boldsymbol{\theta}^*$ als neuer aktueller Zustand $\boldsymbol{\theta}^{(t)}$ akzeptiert werden, sondern nur mit einer gewissen Wahrscheinlichkeit α. Dabei handelt es sich im Wesentlichen um den Quotienten der Posteriori-Dichte und der Vorschlagsdichte ausgewertet am aktuellen Zustand $\boldsymbol{\theta}^{(t-1)}$ und dem vorgeschlagenen Wert $\boldsymbol{\theta}^*$. Genauer gilt

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) = \min\left\{\frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{y})\, q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{y})\, q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})}, 1\right\}.$$

Wird der vorgeschlagene Parametervektor $\boldsymbol{\theta}^*$ nicht akzeptiert, so setzt man $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$. Die Dichte $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ geht in $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$ lediglich im Verhältnis $p(\boldsymbol{\theta}^* \mid \boldsymbol{y})/p(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{y})$ ein, so dass alle konstanten Ausdrücke in $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ nicht berücksichtigt werden müssen. Insbesondere bedeutet dies, dass die Normierungskonstante der Posteriori-Verteilung nicht bekannt sein muss. Dies ist einer großen Vorteile von MCMC Methoden im Vergleich zu herkömmlichen Verfahren zur Zufallszahlenziehung.

Der Metropolis-Hastings-Algorithmus vereinfacht sich nochmals, wenn eine *symmetrische* Vorschlagsdichte mit $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)$ verwendet wird. In diesem Fall erhält man als Akzeptanzwahrscheinlichkeit

$$\alpha(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(t-1)}) = \min\left\{\frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{y})}, 1\right\}.$$

Es handelt sich hier um den 1953 veröffentlichten, ursprünglichen *Metropolis-Algorithmus*, der dann von Hastings (1970) auf den allgemeineren Fall nicht symmetrischer Dichten erweitert wurde.

Die Kunst bei der Verwendung des Metropolis-Hastings-Algorithmus besteht darin eine geeignete Vorschlagsdichte zu wählen. Die Vorschlagsdichte muss garantieren, dass die Akzeptanzwahrscheinlichkeiten groß genug sind und die hintereinander gezogenen Zufallszahlen eine möglichst geringe Abhängigkeit aufweisen. Je geringer die Abhängigkeit, desto geringer ist der erforderliche Stichprobenumfang an Zufallszahlen zur Schätzung von Charakteristika der Posteriori-Verteilung. Wir illustrieren die Konstruktion eines Metropolis-Hastings-Algorithmus am Beispiel eines Bayesianischen Logit-Modells:

Beispiel B.8 Bayesianisches Logit-Modell

Wir betrachten das folgende simulierte Logit-Modell mit zwei Kovariablen x_1 und x_2:

$$\begin{aligned} y_i &= B(1, \pi_i) \qquad i = 1, \ldots, 500, \\ \pi_i &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \\ \eta_i &= -0.5 + 0.6\, x_{i1} - 0.3\, x_{i2}. \end{aligned}$$

Metropolis-Hastings-Algorithmus

Sei $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ die Dichte der Verteilung, aus der Zufallszahlen gezogen werden sollen. Gemäß dem Metropolis-Hastings-Algorithmus werden Zufallszahlen wie folgt gezogen:

1. Wähle einen Startwert $\boldsymbol{\theta}^{(0)}$ und die Anzahl der Iterationen T. Setze $t = 1$.
2. Ziehe eine Zufallszahl $\boldsymbol{\theta}^*$ aus der Vorschlagsdichte $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$ und akzeptiere diese als neuen Zustand $\boldsymbol{\theta}^{(t)}$ mit Wahrscheinlichkeit $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$, andernfalls setze $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
3. Falls $t = T$ beende den Algorithmus, ansonsten setze $t = t + 1$ und fahre fort mit 2.

Nach einer gewissen Konvergenzphase t_0 können die Zufallszahlen $\boldsymbol{\theta}^{(t_0+1)}, \ldots, \boldsymbol{\theta}^{(T)}$ als (abhängige) Stichprobe aus der Dichte $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ angesehen werden.

Die Kovariablen x_1 und x_2 wurden jeweils gemäß einer Standardnormalverteilung simuliert. Wir stellen uns die Aufgabe für die simulierten Daten einen Metropolis-Hastings-Algorithmus zur Schätzung der Parameter $\boldsymbol{\beta} = (-0.5, 0.6, -0.3)'$ zu konstruieren.

Als Priori-Verteilung verwenden wir unabhängige diffuse Verteilungen $p(\beta_j) \propto \text{const}$. Mit diesen Spezifikationen ist die Posteriori-Verteilung proportional zur Likelihood:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{500} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

Zur Implementation eines Metropolis-Hastings-Algorithmus verwenden wir als Vorschlagsdichte eine dreidimensionale Normalverteilung mit dem aktuellen Zustand $\boldsymbol{\beta}^{t-1}$ der Kette als Erwartungswert. Als Kovarianzmatrix fungiert zunächst die Diagonalmatrix $\Sigma = \text{diag}(0.4^2, 0.4^2, 0.4^2)$. Abbildung B.2 (erste Zeile) zeigt die ersten 2000 der mit dieser Vorschlagsdichte gezogenen Zufallszahlen für die Parameter β_0 und β_1. Da wir diffuse Verteilungen als Priori verwendet haben, müssen die Bayes-Schätzer für die Regressionskoeffizienten ungefähr mit den Maximum-Likelihood Schätzern übereinstimmen. In den Grafiken sind daher die ML-Schätzer in Form von horizontalen Geraden eingezeichnet. Offenbar werden mit diesem ersten Algorithmus nur wenige neu vorgeschlagene Zufallszahlen akzeptiert, teilweise bleibt der Zustand für mehr als 100 Iterationen unverändert. Die Akzeptanzwahrscheinlichkeiten sind also viel zu klein. Höhere Akzeptanzwahrscheinlichkeiten erhalten wir, indem die Varianzen in der Vorschlagsdichte verkleinert werden. Wir initialisieren also einen weiteren MH-Algorithmus mit Kovarianzmatrix $\Sigma = \text{diag}(0.1^2, 0.1^2, 0.1^2)$ für die Vorschlagsdichte. Die zweite Reihe in Abbildung B.2 zeigt die resultierenden ersten 2000 Zufallszahlen für β_0 und β_1. Wir können deutlich die kurze Konvergenzzeit von circa 50 Iterationen erkennen. Anschließend erhalten wir einen zufrieden stellenden Algorithmus mit relativ hohen Akzeptanzraten. Wenn wir die Varianzen in Kovarianzmatrix weiter verkleinern mit $\Sigma = \text{diag}(0.02^2, 0.02^2, 0.02^2)$ (dritte Zeile in Abbildung B.2) erhöhen sich zwar die Akzeptanzraten weiter, jedoch verharren hintereinander gezogene Parameter fast im gleichen Zustand.

Ein nützliches und wichtiges Hilfsmittel bei der Beurteilung von MCMC-Algorithmen ist die Autokorrelationsfunktion der gezogenen Parameter, vergleiche zur Autokorrelationsfunktion auch Abschnitt 3.4.4 (Seite 137). Idealerweise sollten die Autokorrelationen möglichst rasch gegen Null abfallen. Je geringer die Abhängigkeiten hintereinander

Abb. B.2. *Samplingpfade für β_0 und β_1 für die verschiedenen MH-Algorithmen. Die rechte Spalte zeigt die jeweiligen Autokorrelationsfunktionen für β_1.*

gezogener Parameter, desto genauer können Charakteristika der Posteriori-Verteilung abgeschätzt werden bei gleicher Stichprobenlänge T der Kette. Die rechte Spalte von Abbildung B.2 zeigt exemplarisch für die drei bisher konstruierten MH-Algorithmen die jeweiligen Autokorrelationsfunktionen für β_1. Offenbar sind die Autokorrelationen im ersten und dritten Algorithmus auch bei einer Verzögerung von 40 noch extrem hoch, während beim zweiten Algorithmus die Autokorrelationen für Verzögerungen ab circa 20 nahezu verschwinden. In der Praxis werden häufig die gezogenen Zufallszahlen „ausgedünnt“, d.h. es wird nur jede k-te gezogene Zufallszahl gespeichert, so dass die verbleibenden Zufallszahlen annähernd unkorreliert sind. Auf diese Art und Weise wird angesichts oft hochdimensionaler Modelle Speicherplatz gespart ohne dass die Schätzgenauigkeit wesentlich beeinflusst wird. Wollte man im vorliegenden Beispiel eine annähernd unabhängige Stichprobe vom Umfang 1000 speichern, so müssten beim zweiten Algorithmus nach der kurzen Burn-in Phase insgesamt circa 20000 Zufallszahlen gezogen und anschließend jede 20. gespeichert werden.

Wir können also festhalten: Kleine Varianzen der Vorschlagsdichte bewirken große Akzeptanzraten. Große Varianzen bewirken den gegenteiligen Effekt, die Akzeptanzraten sind eher klein. In beiden Extremfällen sehr großer bzw. kleiner Varianzen sind die Autokorrelationen hintereinander gezogener Parameter hoch. Die Kunst bei der Konstruktion von MH-Algorithmen besteht also darin, durch geeignete Wahl der Vorschlagsdichten einen Algorithmus zu konstruieren, der hohe Akzeptanzraten verbunden mit niedrigen Autokorrelationen aufweist. Darüber hinaus ist ein möglichst automatisiertes Verfahren wünschenswert, bei dem die Feineinstellung der Parameter der Vorschlagsdichte nicht von Hand vorgenommen werden muss.

Ein Verfahren, das die genannten wünschenswerten Eigenschaften besitzt ist der in Kapitel 4.6 (Seite 228) ausführlich beschriebene MH-Algorithmus. Dieser benötigt keine Adjustierung per Hand und liefert hohe Akzeptanzraten und geringe Autokorrelationen. Die letzte Reihe von Abbildung B.2 zeigt die zu den Daten des Beispiels passenden Samplingpfade für die Parameter β_0 und β_1 und die Autokorrelationsfunktion für β_1. Offenbar besitzt der Algorithmus die besten Eigenschaften von allen bisher getesteten Algorithmen. Die Autokorrelationen verschwinden bereits ab einer Verzögerung von circa 13. Neben der automatisierten optimalen Wahl der Varianzen in der Vorschlagsdichte berücksichtigt dieser Algorithmus auch Korrelationen zwischen den Parametern.

Basierend auf dem automatisierten Algorithmus aus Kapitel 4.6 wurde eine Markov-Kette erzeugt und nach der Burn-In-Phase 20000 Zufallszahlen gezogen, wovon jede 20. gespeichert wurde. Mit den gespeicherten Zufallszahlen können wir sämtliche Charakteristika der Posteriori-Verteilung abschätzen. Als Schätzung für den Posteriori-Erwartungswert verwenden wir die jeweiligen arithmetischen Mittel der gezogenen Zufallszahlen und erhalten $\hat{\boldsymbol{\beta}} = (-0.64, 0.65, -0.38)'$. Diese Schätzung stimmt praktisch mit den ML-Schätzungen überein. Schätzungen für Bayesianische Vertrauensintervalle können wir auf den Quantilen der gezogenen Zufallszahlen aufbauen. Beispielsweise erhalten wir 95% Vertrauensintervalle durch die 2.5% Quantile als untere Grenze und die 97.5% Quantile als obere Grenze. Konkret ergeben sich aus der gezogenen Stichprobe die Intervalle $[-0.87, -0.42]$, $[0.52 - 0.78]$ und $[-0.52, -0.26]$ für die Regressionskoeffizienten. Auch diese stimmen gut mit den aus der Likelihood-Theorie gewonnenen Intervallen überein (müssen aber anders interpretiert werden).

Der große Vorteil von Inferenzvefahren basierend auf Zufallszahlen besteht darin, dass sehr leicht Inferenzschlüsse auch für komplizierte (nichtlineare) Transformationen der Parameter bestimmt werden können. Sind wir beispielsweise an einem Vertrauensbereich für $\exp(\beta_0 + \beta_1 + \beta_2)$ interessiert, so können wir einfach für jede Ziehung die transformierte Größe berechnen und anschließend auf Basis der gewonnen Zufallszahlen den Vertrauensbereich bestehend aus dem 2.5% und 97.5% Quantil bestimmen. Damit ist keinesfalls ausgeschlossen, dass mit anderen Inferenzverfahren ähnliche Inferenzschlüsse möglich sind. In der Regel ist jedoch die dahinter stehende Methodik deutlich komplexer.

△

Gibbs-Sampler und Hybrid-Algorithmen

In vielen praktischen Anwendungen ist der Parametervektor hochdimensional. So sind Probleme mit 1000 und mehr Parametern keine Seltenheit. In diesen Fällen sind die Akzeptanzraten auch bei sorgfältig konstruierten MH-Algorithmen zu klein, weil gleichzeitig eine hochdimensionale Zufallszahl akzeptiert oder verworfen werden muss. Hier schaffen auf dem MH-Algorithmus aufbauende sogenannte Hybrid-Algorithmen Abhilfe. Grundlage ist eine Art „Teile und Herrsche“-Strategie. Der hochdimensionale Parametervektor

$\boldsymbol{\theta}$ wird zunächst in kleinere Blöcke $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_S$ zerlegt. Anschließend werden separate MH-Schritte für die entstandenen Teilvektoren konstruiert.

Den einfachsten Spezialfall dieser Strategie stellt der sogenannte *Gibbs-Sampler* dar. Meistens existieren zwar keine (einfachen) Verfahren zur direkten Ziehung von Zufallszahlen aus der Dichte $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ des gesamten Parametervektors. Häufig können jedoch aus den bedingten Dichten $p(\boldsymbol{\theta}_1 \mid \cdot), p(\boldsymbol{\theta}_2 \mid \cdot), \ldots, p(\boldsymbol{\theta}_S \mid \cdot)$ *direkt* Zufallszahlen gezogen werden. Die Notation $p(\boldsymbol{\theta}_s \mid \cdot)$ ist dabei eine Abkürzung für die bedingte Dichte von $\boldsymbol{\theta}_j$ gegeben alle anderen Blöcke $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{s-1}, \boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_S$ und die Daten $\boldsymbol{y}$. Diese Dichten werden auch als vollständig bedingte Dichten (full conditionals) bezeichnet. Beim Gibbs-Sampler zieht man in jeder Iteration des Samplers hintereinander Zufallszahlen aus den vollständig bedingten Dichten und akzeptiert diese in jedem Fall, d.h. mit Wahrscheinlichkeit Eins, als aktuellen Zustand der Markovkette. Nach einer gewissen Konvergenzzeit können die gezogenen Zufallszahlen als Realisationen aus den Marginalverteilungen $p(\boldsymbol{\theta}_1 \mid \boldsymbol{y}), p(\boldsymbol{\theta}_2 \mid \boldsymbol{y}), \ldots, p(\boldsymbol{\theta}_S \mid \boldsymbol{y})$ angesehen werden. Details zum Gibbs-Sampler entnimmt man dem nachfolgenden Kasten.

Gibbs-Sampler

Sei $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ die Dichte der Posteriori-Verteilung, aus der Zufallszahlen gezogen werden sollen. Sei weiterhin der Parametervektor $\boldsymbol{\theta}$ in S Blöcke $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S$ aufgeteilt. Gemäß dem Gibbs-Sampler werden Zufallszahlen wie folgt gezogen:

1. Wähle Startwerte $\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_S^{(0)}$ und die Anzahl der Iterationen T. Setze $t = 1$.
2. Für $s = 1, \ldots, S$: Ziehe Zufallszahlen $\boldsymbol{\theta}_s^{(t)}$ aus der vollständig bedingten Dichte
$$p(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_{s+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)}, \boldsymbol{y}).$$
Man beachte dass in der Bedingung jeweils die momentan aktuellen Zustände verwendet werden.
3. Falls $t = T$ beende den Algorithmus, ansonsten setze $t = t + 1$ und fahre fort mit 2.

Nach einer gewissen Konvergenzphase t_0 können die Zufallszahlen $\boldsymbol{\theta}_s^{(t_0+1)}, \ldots, \boldsymbol{\theta}_s^{(T)}$ als Ziehungen aus den Marginalverteilungen von $\boldsymbol{\theta}_s \mid \boldsymbol{y}$ angesehen werden.

Sollte aus Teilen der vollständig bedingten Dichten keine direkte Zufallszahlenziehung möglich sein, so bedient man sich wieder der im vorangegangenen Abschnitt beschriebenen MH-Algorithmen zur Ziehung. Dabei wird für den entsprechenden Teilvektor $\boldsymbol{\theta}_s$ eine Vorschlagsdichte
$$q_s(\boldsymbol{\theta}_s^* | \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_s^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)})$$
verwendet, aus der Zufallszahlen $\boldsymbol{\theta}_s^*$ als Vorschläge gezogen werden. Diese werden als neue Zustände der Markov-Kette mit Wahrscheinlichkeit
$$\alpha(\boldsymbol{\theta}_s^* | \boldsymbol{\theta}_s^{(t-1)}) = \min\left\{ \frac{p(\boldsymbol{\theta}_s^* \mid \boldsymbol{\theta}_{-s}^{(t-1)}) q_s(\boldsymbol{\theta}_s^{(t-1)} | \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_s^*, \ldots, \boldsymbol{\theta}_S^{(t-1)})}{p(\boldsymbol{\theta}_s^{(t-1)} \mid \boldsymbol{\theta}_{-s}^{(t-1)}) q_s(\boldsymbol{\theta}_s^* | \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_s^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)})}, 1 \right\}$$
akzeptiert, wobei $p(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{-s}^{(t-1)}) = p(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_{s+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)}, \boldsymbol{y})$ die vollständig bedingte Dichte von $\boldsymbol{\theta}_s$ bezeichnet. Andernfalls setzt man wie im MH-Algorithmus

$\boldsymbol{\theta}_s^{(t)} = \boldsymbol{\theta}_s^{(t-1)}$. Zusammengefasst wird Schritt 2 aus dem Gibbs-Sampler-Algorithmus also ersetzt durch

2.* Für $s = 1, \ldots, S$: Ziehe Zufallszahlen $\boldsymbol{\theta}_s^*$ aus den Vorschlagsdichten $q_s(\boldsymbol{\theta}_s^*|\cdot)$ und akzeptiere diese mit Wahrscheinlichkeit $\alpha(\boldsymbol{\theta}_s^*|\boldsymbol{\theta}_s^{(t-1)})$ als neue Zustände $\boldsymbol{\theta}_s^{(t)}$, andernfalls setze $\boldsymbol{\theta}_s^{(t)} = \boldsymbol{\theta}_s^{(t-1)}$.

Man beachte jedoch, dass durchaus MH- und Gibbs-Sampling-Schritte in 2.* kombiniert eingesetzt werden können. Dazu wird formal die entsprechende Vorschlagsdichte direkt mit der vollständig bedingten Dichte gleichgesetzt. In diesem Fall ergibt sich die Akzeptanzwahrscheinlichkeit unmittelbar als $\alpha(\boldsymbol{\theta}_s^*|\boldsymbol{\theta}_s^{(t-1)}) = 1$, so dass jeder vorgeschlagene Wert akzeptiert wird.

B.5.4 Modellwahl

Die klassische Herangehensweise zur Modellwahl in Bayes-Ansätzen ist der Vergleich konkurrierender Modelle anhand der *Posteriori-Wahrscheinlichkeiten* der Modelle. Gegeben seien zunächst zwei konkurrierende Modelle M_1 und M_2 mit Parametern $\boldsymbol{\theta}_1$ und $\boldsymbol{\theta}_2$. Dann sind die Posteriori-Wahrscheinlichkeiten für die beiden Modelle gegeben durch

$$p(M_j|\boldsymbol{y}) \propto p(M_j)\, p(\boldsymbol{y}|M_j) \propto p(M_j) \int p(\boldsymbol{y}|\boldsymbol{\theta}_j, M_j)\, p(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j \quad j = 1, 2, \tag{B.4}$$

wobei $p(M_j)$ die Priori-Wahrscheinlichkeit für Modell M_j ist, $p(\boldsymbol{\theta}_j|M_j)$ die Priori-Verteilung für den Parameter $\boldsymbol{\theta}_j$ in Modell M_j und $p(\boldsymbol{y}|\boldsymbol{\theta}_j, M_j)$ das jeweilige Beobachtungsmodell darstellt. Die beiden Modelle können verglichen werden, indem das Posteriori-Verhältnis

$$\frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} = \frac{p(M_1)}{p(M_2)} \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}$$

gebildet wird. Ist dieses Verhältnis größer als Eins entscheidet man sich für Modell 1, andernfalls für Modell 2. Bei a priori gleichen Wahrscheinlichkeiten für Modell 1 und 2, d.h. $p(M_1) = p(M_2) = 1/2$ vereinfacht sich das Posteriori-Verhältnis der Wahrscheinlichkeiten zum sogenannten *Bayes-Faktor*

$$BF(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}.$$

Da wir in den meisten Fällen a priori keines der zur Auswahl stehenden Modelle bevorzugen, d.h. alle Modelle a priori gleich wahrscheinlich sind, werden konkurrierende Modelle in Bayes-Ansätzen häufig anhand der Bayes-Faktoren verglichen.

In vielen praktisch relevanten Fragestellungen ist die exakte Berechnung des Bayes-Faktors nicht leicht möglich, da $p(\boldsymbol{y}|M_j)$ schwer berechenbar ist. Eine Approximation für die mit -2 multiplizierte Wahrscheinlichkeit $p(\boldsymbol{y}|M_j)$ ist gegeben durch

$$-2p(\boldsymbol{y}|M_j) = -2 \cdot \log(p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_j, M_j)) + \log(n)\, p_j,$$

wobei p_j die Dimension des Parametervektors $\boldsymbol{\theta}_j$ ist und $\hat{\boldsymbol{\theta}}_j$ der Posteriori-Modus-Schätzer. Diese Formel beruht auf einer Laplace-Approximation des Integrals in (B.4).

Insgesamt führt dies zum sogenannten Bayesian Information Criterion (BIC), das allgemein für ein Modell mit Parametervektor $\boldsymbol{\theta}$, Log-Likelihood $l(\boldsymbol{\theta})$ und ML-Schätzer $\hat{\boldsymbol{\theta}}$ definiert ist durch

$$BIC = -2l(\hat{\boldsymbol{\theta}}) + \log(n)\, p.$$

Wir können damit unter einer Menge konkurrierender Modelle das Modell mit dem kleinsten BIC auswählen. Die Minimierung von BIC bedeutet dann gleichzeitig, dass wir (approximativ) das Modell mit der größten Posteriori-Wahrscheinlichkeit auswählen. Formal besitzt das BIC eine große Ähnlichkeit mit dem AIC. Lediglich der Faktor 2, mit dem beim AIC die Parameterzahl multipliziert wird, wird durch $\log(n)$ ersetzt. Man beachte jedoch, dass beide Informationskriterien völlig unterschiedlich motiviert sind. In der Praxis werden bei Verwendung des BIC in der Regel weniger komplexe Modelle (d.h. Modelle mit weniger Parametern) selektiert als mit dem AIC, da die Bestrafung beim BIC deutlich stärker ist.

Tatsächlich wird das BIC in Bayesianischen Analysen relativ selten eingesetzt. Dies hat im Wesentlichen die folgenden Gründe: Zum Einen sind die der approximativen Herleitung des BIC zu Grunde liegenden Voraussetzungen insbesondere in komplexeren Modellen nicht oder nur schlecht erfüllt. In diesem Zusammenhang ist auch zu beachten, dass in dem Faktor $\log(n)$ die Variable n nicht in jedem Fall den Gesamt-Stichprobenumfang des Datensatzes zu bezeichnen hat. In gemischten Modellen für Longitudinaldaten wäre n gegeben durch die Anzahl der Individuen. Zum Anderen werden komplexere Bayesianische Modelle typischerweise mit Hilfe von MCMC-Verfahren geschätzt, so dass das BIC ohnehin nicht direkt anwendbar ist.

Im Zusammenhang mit MCMC-Methoden existiert seit kurzem ein weiteres Kriterium zur Beurteilung der Güte von Modellen, das sogenannte Deviance Information Criterion (DIC), vergleiche Spiegelhalter, Best, Carlin & van der Linde (2002). Das DIC ist im Rahmen eines MCMC-Algorithmus besonders einfach und schnell, quasi nebenbei berechenbar und findet daher breite Anwendung. Gegeben sei eine durch MCMC gewonnene Stichprobe $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ von Zufallszahlen aus der Posteriori-Verteilung des Modells. Die Berechnung des DIC beruht dann auf zwei Größen. Zum Einen ist dies die sogenannte unstandardisierte Devianz

$$D(\boldsymbol{\theta}) = -2\log(p(\boldsymbol{y}|\boldsymbol{\theta}))$$

und zum Anderen die sogenannte effektive Zahl an Parametern p_D im Modell. Die effektive Anzahl an Parametern kann geschätzt werden durch

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}),$$

wobei

$$\overline{D(\boldsymbol{\theta})} = \frac{1}{T}\sum_{t=1}^{T} D(\boldsymbol{\theta}^{(t)})$$

die durchschnittliche Posteriori-Devianz ist und $D(\bar{\boldsymbol{\theta}})$ die Devianz ausgewertet am Mittelwert $\bar{\boldsymbol{\theta}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\theta}^{(t)}$ der gezogenen Zufallszahlen. Das DIC lässt sich dann berechnen durch

$$\mathrm{DIC} = \overline{D(\boldsymbol{\theta})} + p_D = 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}).$$

Ein Nachteil des DIC ist, dass der Wert in verschiedenen MCMC-Durchläufen schwankt, da das DIC basierend auf einer Stichprobe aus der Posteriori berechnet wird. Es kann also

passieren, dass der Vergleich zweier Modelle anhand des DIC in verschiedenen MCMC-Läufen unterschiedlich ausfällt. Dies wird aber in der Regel nur dann der Fall sein, wenn die beiden DIC-Werte nahe beieinander liegen.

Beispiel B.9 Bayesianisches Logit-Modell – DIC

Wir illustrieren die Verwendung des DIC mit den simulierten Daten aus Beispiel B.8. Wenn wir x_2 fälschlicherweise nicht berücksichtigen und ein Logit-Modell mit nur einer Kovariablen x_1 schätzen, erhalten wir als effektive Anzahl an Parametern den Wert $p_D = 1.99$ und als DIC den Wert 571.6. Die effektive Anzahl an Parametern von ungefähr zwei ist plausibel, da wir ja genau zwei Parameter β_0 und β_1 geschätzt haben. Bei korrekt spezifiziertem Modell erhalten wir stattdessen $p_D = 2.93$ und DIC $= 540.3$. Die effektive Anzahl an Parametern erhöht sich also, wie zu erwarten, ungefähr um Eins. Das DIC ist deutlich kleiner als im fehlspezifizierten Modell, so dass das komplexere Modell deutlich zu bevorzugen ist.

Zur Illustration starten wir fünf weitere MCMC-Läufe, wobei wir uns auf das korrekt spezifizierte Modell beschränken. Wir erhalten für p_D die Werte 3.05, 2.99, 3.15, 2.87 und 3.23. Für das DIC ergeben sich die Werte 540.56, 540.42, 540.73, 540.19 und 540.91. Damit wird nochmal deutlich, dass das DIC in unterschiedlichen MCMC-Läufen schwankt. Jedoch ist die Variabilität in aller Regel sehr gering.

△

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis (2. Auflage)*, Wiley, New York.

Albert, J. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**: 669–679.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Boston.

Augustin, N. H., Lang, S., Musio, M. & von Wilpert, K. (2007). A spatial model for the needle losses of pine-trees in the forests of Baden-Württemberg: an application of Bayesian structured additive regression, *Applied Statistics* **56**: 29–50.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2003). *Hierarchical Modelling and Analysis for Spatial Data*, Chapman & Hall / CRC, Boca Raton, FL.

Belsley, D. A., Kuh, E. & Welsch, R. E. (2003). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, Hoboken, NJ.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models, *Journal of Computational and Graphical Statistics* **9**: 122–140.

Biller, C. & Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines, *Statistical Modelling* **1**: 195–211.

Blossfeld, H.-P., Hamerle, A. & Mayer, K. U. (1986). *Ereignisanalyse*, Campus Verlag, Frankfurt am Main.

Blossfeld, H.-P. & Rohwer, G. (2001). *Techniques of Event History Modeling (2. Auflage)*, Erlbaum, Mahwah, NJ.

Bortz, J. (2004). *Statistik (6. Auflage)*, Springer, Berlin.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

Breiman, L. (2001). Random forests, *Machine Learning* **45**: 5–32.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984). *Classification and Regression Trees*, Chapman & Hall/CRC, Boca Raton, FL.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**: 9–25.

Brezger, A. & Lang, S. (2006). Generalized additive regression based on Bayesian P-splines, *Computational Statistics and Data Analysis* **50**: 967–991.

Brockwell, P. J. & Davis, R. A. (2002). *Introduction to Time Series and Forecasting (2. Auflage)*, Springer, New York.

Bühlmann, P. (2006). Boosting for high-dimensional linear models, *Annals of Statistics* **34**: 559–583.

Bühlmann, P. & Yu, B. (2003). Boosting with the L2 loss: regression and classification, *Journal of the American Statistical Association* **98**: 324–339.

Bühlmann, P. & Yu, B. (2006). Sparse boosting, *Journal of Machine Learning Research* **7**: 1001–1024.

Buja, A., Hastie, T. & Tibshirani, R. (1989). Linear smoothers and additive models, *Annals of Statistics* **17**: 453–510.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodal Inference (2. Auflage)*, Springer, New York.

Cameron, A. C. & Trivedi, P. K. (1998). *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models (2. Auflage)*, Chapman & Hall / CRC, Boca Raton, FL.

Chen, M. H. & Dey, D. K. (2000). Bayesian analysis for correlated ordinal data models, *in* D. K. Dey, S. K. Ghosh & B. K. Mallick (eds), *Generalized linear models: A Bayesian perspective*, Marcel Dekker, New York, pp. 133–159.

Chiles, J.-P. & Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*, Wiley, New York.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**: 829–836.

Collett, D. (1991). *Modelling Binary Data*, Chapman & Hall, London.

Collett, D. (2003). *Modelling Survival Data in Medical Research (2. Auflage)*, Chapman & Hall / CRC, Boca Raton, FL.

Cowles, M. & Carlin, B. (1996). Markov chain monte carlo convergence diagnostics: A comparative review, *Journal of the American Statistical Association* **91**: 883–904.

Crainiceanu, C. M. & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component, *Journal of the Royal Statistical Society B* **66**: 165–185.

Crainiceanu, C., Ruppert, D., Claeskens, G. & Wand, M. (2005). Exact likelihood ratio tests for penalised splines, *Biomtrika* **92**: 91–103.

De Boor, C. (2001). *A practical Guide to Splines*, Springer, New York.

Dellaportas, P. & Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via gibbs sampling, *Applied Statistics* **42**: 443–459.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). Automatic Bayesian curve fitting, *Journal of the Royal Statistical Society B* **60**: 333–350.

Denison, D., Holmes, C., Mallick, B. & Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*, Wiley, Chichester.

Denuit, M. & Lang, S. (2005). Nonlife ratemaking with bayesian gam's., *Insurance: Mathematics and Economics* **35**: 627–647.

Dey, D., Gosh, S. K. & Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective*, Dekker, New York.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*, Clarendon Press, Oxford.

Diggle, P. J., Heagerty, P., Liang, K.-L. & Zeger, S. L. (2002). *Analysis of Longitudinal Data (2. Auflage)*, Oxford University Press, Oxford.

DiMatteo, I., Genovese, C. R. & Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines, *Biometrika* **88**: 1055–1071.

Ecker, M. D. & Gelfand, A. E. (2003). Spatial modelling and prediction under stationary nongeometric range anisotropy, *Environmental and Ecological Statistics* **10**: 165–178.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression, *Annals of Statistics* **32**: 407–451.

Eilers, P. H. C. & Marx, B. D. (2003). Multidimensional calibration with temperature interaction using two-dimensional penalized signal regression, *Chemometrics and Intelligent Laboratory Systems* **66**: 159–174.

Fahrmeir, L., Hamerle, A. & Tutz, G. (1996). *Multivariate Statistische Verfahren (2. Auflage)*, De Gruyter, Berlin.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *The Annals of Statistics* **13**: 342–368.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective, *Statistica Sinica* **14**: 731–761.

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2007). *Statistik: Der Weg zur Datenanalyse (6. Auflage)*, Springer, Berlin.

Fahrmeir, L. & Lang, S. (2001a). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Applied Statistics* **50**: 201–220.

Fahrmeir, L. & Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time-space data, *Annals of the Institute of Statistical Mathematics* **53**: 11–30.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2. Auflage)*, Springer, New York.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

Faraway, J. J. (2004). *Linear Models with R*, Chapman & Hall / CRC, Boca Raton, FL.

Forthofer, R. N. & Lehnen, R. G. (1981). *Public Program Analysis: A New Categorical Data Approach*, Lifetime Learning Publications, Belmont, Calif.

Fotheringham, A., Brunsdon, C. & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley, Chichester.

Friedman, J. H. (1991). Multivariate adaptive regression splines, *The Annals of Statistics* **19**: 1–141.

Furnival, G. M. & Wilson, R. W. (1974). Regression by leaps and bounds, *Technometrics* **16**: 499–511.

Galton, F. (1889). *Natural Inheritance*, Macmillan, London.

Gamerman, D. (1997). Efficient sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**: 57–68.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**: 398–409.

Gençay, R., Selçuk, F. & Whitcher, B. (2002). *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*, Academic Press, San Diego.

George, A. & Liu, J. W. (1981). *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Geweke, J. (1991). Efficient Simulation From the Multivariate Normal and Student-t Distribution Subject to Linear Constraints, *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, Alexandria, pp. 571–578.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**: 711–732.

Green, P. J. (2001). A primer on Markov chain Monte Carlo, *in* O. Barndorff-Nielsen, D. Cox & C. Klüppelberg (eds), *Complex Stochastic Systems*, Chapman & Hall / CRC, Boca Raton, FL, pp. 1–51.

Greene, W. H. (2000). *Econometric Analysis (4. Auflage)*, Prentice Hall, Upper Saddle River, NJ.

Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer, New York.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press.

Hämmerlin, G. & Hoffmann, K.-H. (1994). *Numerische Mathematik (4. Auflage)*, Springer, Berlin.

Hansen, M. H. & Kooperberg, C. (2002). Spline adaptation in extended linear models, *Statistical Science* **17**: 2–51.

Härdle, W. (1990). *Smoothing Techniques*, Springer, New York.

Härdle, W., Müller, M., Sperlich, S. & Werwatz, A. (2004). *Nonparametric and Semiparametric Models*, Springer, Berlin.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall / CRC, London.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. (2003). *The Elements of Statistical Learning*, Springer, New York.

Hastings, W. K. (1970). Monte-carlo sampling methods using Markov chains and their applications, *Biometrika* **57**: 97–109.

Hodrick, R. J. & Prescott, E. C. (1997). Postwar U.S. business cycles: An empirical investigation, *Journal of Money, Credit, and Banking* **29**: 1–16.

Holmes, C. C. & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Analysis* **1**: 145–168.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Johnson, M. E., Moore, L. M. & Ylvisaker, D. (1990). Minimax and maximin distance designs, *Journal of Statistical Planning and Inference* **26**: 131–148.

Jorgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*, Springer, New York.

Judge, G. G., Griffith, W. E., Hill, R. C., Lütkepohl, H. & Lee, T.-C. (1980). *The Theory and Practice of Econometrics*, Wiley, New York.

Kauermann, G. (2006). Nonparametric models and their estimation, *Allgemeines Statistisches Archiv* **90**: 135–150.

Kauermann, G. & Khomski, P. (2007). Additive two way hazards model with varying coefficients, *Erscheint in Computational Statistics and Data Analysis* .

Kauermann, G. & Opsomer, J. (2004). Generalized cross-validation for bandwidth selection of backfitting estimates in generalized additive models, *Journal of Computational and Graphical Statistics* **13**: 66–89.

Klein, J. P. & Moeschberger, M. L. (2005). *Survival Analysis (2. Auflage)*, Springer, New York.

Kneib, T. (2005). *Mixed model based inference in structured additive regression*, Dr. Hut-Verlag, München, erhältlich unter `http://edoc.ub.uni-muenchen.de/archive/00005011/`.

Kneib, T. & Fahrmeir, L. (2006). Structured additive regression for multicategorical space-time data: A mixed model approach, *Biometrics* **62**: 109–118.

Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression, *Erscheint im Scandinavian Journal of Statistics* .

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, New York.

Lang, S. (2004). *Vorlesungsskript Computerintensive Verfahren.* Erhätlich auf der Internetseite des Buchs.

Lang, S. & Brezger, A. (2004). Bayesian P-splines, *Journal of Computational and Graphical Statistics* **13**: 183–212.

Lange, K. (2000). *Numerical Analysis for Statisticians*, Springer, New York.

Leeflang, P. S. H., Wittink, D. R., Wedel, M. & Naert, P. A. (2000). *Building Models for Marketing Decisions*, Kluwer, Boston.

Loader, C. (1999). *Local Regression and Likelihood*, Springer, New York.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models (2. Auflage)*, Chapman & Hall / CRC, London.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*, John Wiley.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour, *in* P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York.

McFadden, D. (1984). Econometric analysis of qualitative response models, *in* Z. Griliches & M. Intriligator (eds), *Handbook of Econometrics*, North Holland, Amsterdam, pp. 1395–1457.

Mengersen, K. L., Robert, C. & Guihenneuc-Jouyaux, C. (1999). Mcmc convergence diagnostics: a „reviewww“, *in* B. et al. (ed.), *Bayesian Statistics 6*, Oxford University Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**: 1087–1091.

Miller, A. (2002). *Subset Selection in Regression*, Chapman & Hall / CRC, Boca Raton, FL.

Mosler, K. & Schmid, F. (2005). *Wahrscheinlichkeitsrechnung und schließende Statistik (2. Auflage)*, Springer, Berlin.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society A* **135**: 370–384.

Nychka, D. (2000). Spatial-process estimates as smoothers, *in* M. Schimek (ed.), *Smoothing and Regression: Approaches, Computation and Application*, Wiley, New York, pp. 393–424.

Nychka, D. & Saltzman, N. (1998). Design of air quality monitoring networks, *in* D. Nychka, W. W. Piegorsch & L. H. Cox (eds), *Case Studies in Environmental Statistics*, Springer, New York, pp. 51–76.

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhäuser, Boston.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference*, Arnold, London.

Ramsay, J. O. & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis (2. Auflage)*, Springer, New York.

Rawlings, J. O., Pantula, S. G. & Dickey, A. D. (2001). *Applied Regression Analysis (2. Auflage)*, Springer, New York.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, *Applied Statistics* **54**: 507–554.

Robert, C. P. (1995). Simulation of truncated normal variables, *Statistics and Computing* **5**: 121–125.

Rousseeuw, P. J. & Leroy, A. M. (2003). *Robust Regression and Outlier Detection*, Wiley, Hoboken, NJ.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields with applications, *Journal of the Royal Statistical Society B* **63**: 325–338.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields*, Chapman & Hall / CRC, Boca Raton, FL.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

Schabenberger, O. & Gotway, C. (2005). *Statistical Methods for Spatial Data Analysis*, Chapman & Hall / CRC, Boca Raton, FL.

Schira, J. (2005). *Statistische Methoden der VWL und BWL. Theorie und Praxis (2. Auflage)*, Pearson Studium, München.

Schlittgen, R. & Streitberg, B. (2001). *Zeitreihenanalyse (9. Auflage)*, Oldenbourg Verlag, München.

Schmidt, K. & Trenkler, G. (2006). *Einführung in die Moderne Matrix-Algebra (2. Auflage)*, Springer, Berlin.

Schneeweiß, H. (1990). *Ökonometrie*, Physika Verlag, Heidelberg.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling*, Chapman & Hall / CRC, Boca Raton, FL.

Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics* **75**: 317–343.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit., *Journal of the Royal Statistical Society B* **65**: 583–639.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

Stone, C. J., Hansen, M. H., Kooperberg, C. & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling, *Annals of Statistics* **25**: 1371–1470.

Toutenburg, H. (2003). *Lineare Modelle*, Physica-Verlag, Heidelberg.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*, University Press, Cambridge.

Tutz, G. (2000). *Die Analyse kategorialer Daten*, Oldenbourg, München.

Tutz, G. & Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting, *Biometrics* **62**: 961–971.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer, New York.

von Auer, L. (2005). *Ökonometrie (3. Auflage)*, Springer, Berlin.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures, *Computational Statistics* **15**: 443–462.

Wand, M. P. (2003). Smoothing and mixed models, *Computational Statistics* **18**: 223–249.

Weisberg, S. (2005). *Applied Linear Regression (3. Auflage)*, Wiley, New York.

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity, *Econometrica* **48**: 817–838.

Whittaker, E. T. (1922/23). On a new method of graduation, *Proceedings of the Edinburgh Mathematical Society* **41**: 63–75.

Winkelmann, R. (1997). *Econometric Analysis of Count Data (2. Auflage)*, Springer, Berlin.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society B* **62**: 413–428.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall / CRC, Boca Raton, FL.

Wooldridge, J. M. (2006). *Introductory Econometrics (3. Auflage)*, Thomson, Mason, Ohio.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*, Cambridge University Press.

Yau, P., Kohn, R. & Wood, S. (2003). Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression, *Journal of Computational and Graphical Statistics* **12**: 23–54.

Zimmermann, D. L. (1993). Another look at anisotropy in geostatistics, *Mathematical Geology* **25**: 453–470.

Index

χ^2-Verteilung, 460

Abhängige Variable, 1, 19
Adaptive Verfahren, 359, 394
Ähnliche Matrizen, 454
AIC, 477
– Korrigiertes, 477
– Lineare Modelle, 161
– Nichtparametrische Regression, 352
– Strukturiert additive Regression, 422
Aitken-Schätzer, 126
All-Subset-Selection, 164
Allgemeine Streudiagramm-Glätter, 340
Allgemeines lineares Modell, 125
Anisotropie, 382
Äquivalente Freiheitsgrade, 345
Autokorrelation, 67, 136
– Erster Ordnung, 67, 137
– Partielle, 138
Autokorrelationsfunktion, 137
– Partielle, 138
Autoregressive Modelle, 290

B-Splines, 303
Backfitting, 420
Backward-Selection, 164
Basic-Splines, 303
Basisfunktionen, 298
Bayes-Faktor, 488
Bayes-Inferenz, 478
– Intervallschätzer, 481
– Punktschätzer, 480
Bayesianische P-Splines, 316
Bayesianischer Vertrauensbereich, 481
Bayesianisches Konfidenzintervall, 481
Bayesianisches lineares Modell, 151
Bedingte Verteilung, 462
Beobachtete Informationsmatrix, 470
Beobachtungsmodell, 478
Bester linearer Prädiktor, 262
Bester linearer unverzerrter Prädiktor, 262, 383
Bestimmtheitsmaß, 99
Bestimmtheitsmaß, 98
– Korrigiertes, 160
Bias-Varianz-Trade Off, 348
BIC, 162, 488
Binäre Regressionsmodelle, 192
Bivariate Polynom-Splines, 372
BLUP, 262, 383
Bonferroni-Korrektur, 175, 343
Boxplot, 11
Breusch-Pagan-Test, 131

Charakteristisches Polynom, 453, 454
Cholesky-Zerlegung, 457
Clusterspezifische Effekte, 35, 38, 254
Cook-Distanz, 178
CV, 351

Defekt einer Matrix, 450
Degrees of freedom, 345
Dehnungsmatrix, 382
Designmatrix, 61
Determinante, 451
Devianz, 205
Diagonalmatrix, 446
Differentiation von Matrixfunktionen, 457
Differenzen k-ter Ordnung, 310
Differenzenmatrix, 312
diffuse Verteilung, 480
Diskreter Zufallsvektor, 461
Dummy-Kodierung, 80
Dummy-Variablen, 81
Durbin-Watson-Test, 141

Effekt-Kodierung, 82
Effektmodifizierer, 408
Eigenraum, 454
Eigenvektor, 453
Eigenwert, 453
Eigenwerte
– Symmetrischer Matrizen, 454
Einflussanalyse, 177
Einheitsmatrix, 446
Empirisch beste lineare Schätzer, 265
Empirische Bayes-Schätzung, 273, 284, 357
Empirische Korrelationsmatrix, 463
Empirische Kovarianzmatrix, 463
Epanechnikov-Kern, 337

Equivalent degrees of freedom, 345
Erklärende Variable, 1, 19
Erwartete Informationsmatrix, 469
Erwarteter quadrierter Prognosefehler, 158
Erwartungswertvektor, 462
Exponential-Korrelationsfunktion, 328
Exponentialfamilie, 218
Exponentialverteilung, 460

F-Test, 113
F-Verteilung, 461
Fehlerterm, 19
Fisher-Information, 469
Fisher-Informationsmatrix, 469
Fisher-Matrix, 469
Fisher-Scoring, 202, 249, 264, 473
Fluch der Dimension, 396
Forward-Selection, 164
Freiheitsgrade, 345

G-Inverse, 452
Gammaverteilung, 460
Gauß-Kern, 337
Gauß-Markov-Theorem, 104
Gauß-Felder, 381
Gauß-Korrelationsfunktion, 328
Gauß-Markov-Theorem, 103
Gauß-Markov-Zufallsfelder, 390
GCV, 352, 423
Generalisierte Schätzfunktion, 226
Generalisierte Schätzgleichung, 226
Generalisiertes Kreuzvalidierungskriterium, 352, 423
Generalized Estimating Equations, 226, 290
Geoadditive Modelle, 51, 404
Geometrische Eigenschaften des KQ-Schätzers, 97
Geostatistik, 382
Gewichtete KQ-Schätzung, 126
Gewichteter KQ-Schätzer, 127
Gibbs-Sampler, 487
Glättungsmatrix, 341, 422
Glättungsparameter, 308, 319, 345, 350, 422
Glättungsparameterwahl, 350, 422
Glättungsverfahren, 291
Glättunssplines, 323
Gleichverteilungs-Kern, 337
Gleitende Durchschnitte, 333
GMZF, 390
Gram-Schmidt-Orthogonalisierung, 97
Gruppierte Daten, 127, 195, 237
Gruppiertes Cox-Modell, 244

Hat-Matrix, 93
Hauptdiagonale, 445
Hauptkomponentenregression, 172
Hazardrate, 53
Hazardratenmodelle, 53
Hebelwerte, 177
Heteroskedastische Fehler, 61, 64, 124
– Diagnose, 129
– Variablentransformation, 132
– White-Schätzer, 135
– Zweistufige Schätzung, 133
Histogramm, 11
Hodrick-Prescott-Filter, 326
Homoskedastische Fehler, 61, 64

Idempotente Matrix, 448
Identifikationsproblem additiver Modelle, 400
IGMZF, 390
Indefinit, 455
Individuenspezifische Effekte, 35, 38, 254
Informationsmatrix, 469
– Beobachtete, 470
– Erwartete, 469
Interaktionen zwischen Kovariablen, 83, 407
Interaktionsvariable, 408
Intrinsische Gauß-Markov-Zufallsfelder, 390
Inverse Gammaverteilung, 461
Inverse Matrix, 450
Irreduzibler Prognosefehler, 159
Irrfahrten, 326
– k-ter Ordnung, 316
Isotrop, 381
Iterativ Gewichtete KQ-Schätzung, 222, 249

Kanonische Linkfunktion, 220
Kanonischer Parameter, 218
Kerndichteschätzer, 11
Kernfunktionen, 337
Klassische Normalregression, 62
Klassisches lineares Regressionsmodell, 62
Knoten, 295
– Äquidistante Knoten, 301
– Quantilbasierte Knoten, 302
Kodierung
– Effekt-, 82
Kollinearität, 170
Kollinearitätsanalyse, 170
Komplementäres log-log-Modell, 192
Konditionale Modelle, 290
Konfidenzbänder, 119, 342
Konfidenzintervalle, 119, 342
Korrelationsfunktion, 327

– Exponential-, 328
– Gauß-, 328
– Isotrope, 381
– Matérn, 330
– Potenz-, 328
– Range, 328
– Sphärische, 328
– Stationäre, 328
Korrelationsmatrix, 462
– Empirische, 463
Korrigiertes Bestimmtheitsmaß, 160
Kovariable, 1, 19
Kovarianzmatrix, 462
– Empirische, 463
KQ-Methode, 90
KQ-Schätzer
– Asymptotische Eigenschaften, 105
– Eigenschaften, 104
– Erwartungswert, 101
– Gewichteter, 127
– Kovarianzmatrix, 101, 104
KQ-Schätzung
– Geometrische Eigenschaften, 97
– Gewichtete, 126
– Penalisierte, 262, 311
Kreuzvalidierungskriterium, 161, 351, 423
Kriging, 327, 381
– Gewöhnliches, 383
– Universelles, 383
Kroneckerprodukt, 447
Kumulative Modelle, 244
Kumulatives Extremwertmodell, 244

Laplace-Approximation, 284
Lebensdaueranalyse, 52
Likelihood-Quotienten-Test, 476
Lineare gemischte Modelle für Longitudinal- und Clusterdaten, 38, 259
Lineare Glättungsverfahren, 340
Lineare Schätzer, 102
Linearer Prädiktor, 190
Lineares Wahrscheinlichkeitsmodell, 190
Linkfunktion, 190, 217
– Kanonische, 220
– Natürliche, 220
Locally weighted Regression, 339
Loess, 339, 394
Log-lineares Poisson-Modell, 210
Log-lineares Poisson-Normal-Modell, 281
Logistisches sequentielles Modell, 246
Logit-Modell, 32, 192
Lognormalverteilung, 72, 459
Lokal polynomiale Regression, 335
Lokale Glättungsverfahren, 333, 394

Mallows' C_p, 161
Markov-Chain-Monte-Carlo-Methoden, 482
Markov-Eigenschaft, 316, 376
Markov-Zufallsfelder, 387
MARS, 359, 395
Matérn Korrelationsfunktionen, 330
Matérn-Splines, 384
Matrix, 445
– Ähnlich, 454
– Definite, 455
– Diagonal-, 446
– Idempotente, 448
– Inverse, 450
– Orthogonale, 448
– Quadratische, 445
– Spur, 452
– Symmetrische, 446
– Transponiert, 445
Matrizenmultiplikation, 447
Maximum-Likelihood Schätzung, 467
– Asymptotische Eigenschaften, 475
MCMC, 482
Median-Regression, 90
Mehrkategoriales Logit-Modell, 239
Metropolis-Hastings-Algorithmus, 483
Mietspiegel, 5
Mischverteilung, 255
Mixed Model Equations, 262
Mixed Models, 35
ML-Gleichung, 199, 222, 247
ML-Schätzung, 467
– Asymptotische Eigenschaften, 475
Model Averaging, 364
Modell der proportionalen kumulativen Chancen, 244
Modelle mit zufälligen Effekten, 35
Modellwahlkriterien, 159, 477, 488
Multikategoriales Logit-Modell, 239
Multinomialverteilung, 236
Multivariate adaptive Regressions-Splines, 359, 395
Multivariate Normalverteilung, 464
– Bedingte Verteilungen, 465
– Erwartungswert, 464
– Kovarianzmatrix, 464
– Lineare Transformationen, 464
– Marginalverteilungen, 465
Multivariate t-Verteilung, 467
MZF, 387

Nächste-Nachbarn-Schätzer, 41, 333, 394
Nachbarschaften, 387
Nachbarschaftsmatrix, 389
Nadaraya-Watson-Schätzer, 335, 394
Natürliche kubische Splines, 323
Natürliche Linkfunktion, 220
Natürlicher Parameter, 218
Negativ definit, 455
Newton-Raphson-Algorithmus, 264, 473
Newton-Verfahren, 473
Nichtlineare Zusammenhänge durch Polynome, 80
Nichtlineare Zusammenhänge durch Variablentransformation, 80
Nichtparametrische Regression, 291
Nominal skalierte Zielvariable, 235
Nonparametrische Regression, 291
Normal-inverse Gammaverteilung, 147
Normalverteilung
– Multivariate, 464
– Singuläre, 465
– Trunkierte, 459
– Univariate, 459
Nullraum einer Matrix, 450

Ordinal skalierte Zielvariable, 235
Ordnung einer Matrix, 445
Orthogonale Matrix, 448
Orthogonale Polynome, 78
Overdispersion, 197, 210

P-Splines, 42, 306
– Bayesianische, 316
Partiell lineares Modell, 401
Partielle Autokorrelationsfunktion, 138
Partielle Residuen, 63, 110, 424
Penalisierte KQ-Schätzung, 308, 311
Penalisierte Log-Likelihood, 229, 285, 322
Penalisierte Residuenquadratsumme, 308
Penalisierte Splines, 306
– Bayesianische, 316
Penalisierter KQ-Schätzer, 311
Penalisierter ML-Schätzer, 229
Penalisiertes KQ-Kriterium, 262
Penalisierungsansätze, 306
PKQ-Schätzung, 311
Polynom-Splines, 42, 293, 295
– Bivariate, 372
Polynomiale Regression, 75
Populationseffekte, 35, 38, 254
Positiv definit, 455
Posteriori-Modus-Schätzer, 480
Posteriori-Verteilung, 478
Potenz-Exponential Korrelationsfunktionen, 328
Prädiktionsmatrix, 93, 341, 345
Präzisionsmatrix, 462
Priori-Verteilung, 478
Probit-Modell, 192
Probit-Normal-Modell, 282
Prognosefehler
– Erwarteter quadrierter, 158
– Irreduzibler, 159
Prognoseintervalle, 121
Proportional Odds Model, 244

Quadratische Form, 455
Quadratische Matrix, 445
Quantil-Regression, 91
Quasi-Likelihood-Modelle, 226, 290
Quasi-Score-Funktion, 226

Räumlich Autoregressive Prozesse, 393
Rückwärts-Selektion, 164
Radiale Basisfunktionen, 379, 384, 395
Random Walk, 326
– k-ter Ordnung, 316
Randverteilung, 462
Rang einer Matrix, 449
Range, 328
Regressions-Splines, 293
Regressionsbäume, 361, 395
Regressor, 1, 19
Regulär, 449
Reguläre Inverse, 450
REML, 94
Residualplots, 129
Residuen, 63
– Übersicht, 110
– Matrixnotation, 93
– Partielle, 63, 110, 424
– Standardisierte, 108, 110, 424
– Studentisierte, 109, 110
Responsefunktion, 190, 217, 220
Restringierte Log-Likelihood, 263
Restringierte Maximum-Likelihood-Schätzung, 94
Reversible Jump MCMC, 366
Ridge-Regression, 171
Rotationsmatrix, 382
Running Line, 334
Running Mean, 334
Running Median, 334

Sandwich-Matrix, 136, 266, 313, 357

SAR, 393
Scatterplot-Smoother, 292
Schrittweise-Selektion, 164
Schwellenwertmechanismus, 193, 242, 427
Score-Funktion, 469
Score-Test, 476
Semidefinit, 455
Semiparametrisches Modell, 401
Sequentielles Modell, 246
– Extremwertmodell, 246
– Logistisches, 246
Singuläre Normalverteilung, 465
Smoother-Matrix, 341, 422
Smoothing Splines, 323
SMSE, 158
Spaltenraum einer Matrix, 450
Spaltenregulär, 449
Spaltenvektor, 445
Spatially Autoregressive, 393
Spektralzerlegung, 455
Sphärische Korrelationsfunktionen, 328
Splines, 293, 295
Spur einer Matrix, 452
Störgröße, 2, 19, 60
Standardisierte Residuen, 108, 110, 424
Standardnormalverteilung, 459
Stationäre Gauß-Felder, 381
Stationärer Gauß-Prozess, 327
Stepwise-Selection, 164
Stetiger Zufallsvektor, 461
Stochastische Komponente, 19
Strafterm, 308
Streudiagramm-Glätter, 292, 340
Streuungszerlegung, 98
Strukturiert-additive Regression, 413
Studentisierte Residuen, 109, 110
Symmetrische Matrix, 446
Systematische Komponente, 19, 60

t-Test, 116
t-Verteilung, 461
– Multivariate, 467
Tensorprodukt-Basen, 371
Tensorprodukt-P-Splines, 377
Tensorprodukt-Splines, 371
Tests linearer Hypothesen, 112, 120, 249
Teststichprobe, 159
Thin-Plate-Spline, 380
TP-Splines, 296
Transponierte Matrix, 445
Truncated Power Series, 296
Trunkierte Normalverteilung, 459
Trunkierte Potenzen, 296

Überdispersion, 197, 210
Unbeobachtete Heterogenität, 197
Univariate Normalverteilung, 459
Unterernährung in Sambia, 5

Validierungsstichprobe, 159, 161
Variablentransformation, 73
Varianzinflationsfaktor, 171
Varianzkomponentenmodell, 253, 258, 266, 424
Variierende Koeffizienten, 408
Verallgemeinerte Inverse, 452
Vollständig bedingte Dichten, 487
Vorwärts-Selektion, 164

Wald-Test, 476
White-Schätzer, 135

Zeilenraum einer Matrix, 450
Zeilenregulär, 449
Zeilenvektor, 445
Zentrierung, 78, 400, 411, 425
Zielvariable, 1, 19
Zufallsvariable, 459
– Mehrdimensionale, 461
Zufallsvektor, 461
– Bedingte Verteilung, 462
– Diskret, 461
– Randverteilung, 462
– Stetig, 461

Die Autoren

Ludwig Fahrmeir war wissenschaftlicher Assistent an der Technischen Universität München (1970 – 1977), danach Professor für Statistik an der Universität Regensburg (1978 – 1991). Seit 1991 ist er Professor für Statistik an der LMU München und war von 1995 – 2006 Sprecher des Sonderforschungsbereichs "Statistische Analyse diskreter Strukturen (mit Anwendungen in Biometrie und Ökonometrie)". Derzeit koordiniert er das LMUinnovativ Projekt "Analysis and Modelling of Complex Systems in Biology and Medicine (BioMed-S)". Seine Forschungsinteressen sind die statistische Analyse und Modellierung in Wirtschafts-, Sozial und Lebenswissenschaften, Bayesianische semiparametrische Regression, die Analyse von zeitlich-räumlich strukturierten Daten und BioImaging.

Thomas Kneib studierte von 1998 bis 2003 Statistik in München und promovierte dort 2006. Im Sommersemester 2007 war er Gastprofessor für Angewandte Statistik an der Universität Ulm. Im Wintersemester 2008/09 vertrat er den Lehrstuhl für Statistik an der Universität Göttingen. 2009 habilitierte er sich und ist seit April 2009 Professor für Angewandte Statistik an der Universität Oldenburg. Seine aktuellen Forschungsschwerpunkte sind semiparametrische geoadditive Regressionsmodelle, Bayesianische Regularisierung und Boosting-Verfahren.

Stefan Lang studierte von 1991 bis 1997 Statistik in München. Von 1997 bis 2005 war er wissenschaftlicher Mitarbeiter bzw. Assistent am Institut für Statistik in München. Seine Promotion schloss er 2001 ab, 2004 folgte die Habilitation. Von April 2005 bis September 2006 war er Professor für Statistik in Leipzig, seit Oktober 2006 hat er die Professor für Angewandte Statistik an der Universität Innsbruck inne. Seine Forschungsinteressen sind Bayesianische semiparametrische Regression, räumliche Statistik, hierarchische Modelle und Multilevel-Modelle, sowie simulationsbasierte Inferenz.