

Ludwig Fahrmeir
Thomas Kneib
Stefan Lang

Regression

Modelle, Methoden und Anwendungen



Springer

Berlin Heidelberg New York
Barcelona Hong Kong
London Milan Paris
Tokyo

Vorwort

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 Bearbeitung 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*

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	Geoadditve 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	146
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	Auswirkung der Modellspezifikation auf Bias, Varianz und Prognosegüte	156
	Auswirkung der Modellspezifikation auf Bias und Varianz des KQ-Schätzers	156
	Auswirkung der Modellspezifikation auf die Prognosegüte	157
3.6.2	Modellwahlkriterien	159
	Das korrigierte Bestimmtheitsmaß	160
	Mallow's C_p	161
	Informationskriterium nach Akaike AIC	161
	Kreuzvalidierung	162
	Bayesianisches Informationskriterium BIC	162
3.6.3	Praktische Verwendung der Modellwahlkriterien	163
3.6.4	Modelldiagnose	167
	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 der ML- bzw. KQ-Schätzung im linearen Regressionsmodell	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 fixer und zufälliger Effekte 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 Penaliserungsansätze	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-Penalierungsansätze	375
7.2.2	Radiale Basisfunktionen	379
7.2.3	Kriging: Räumliche Glättung bei stetiger Lokationsvariable	381
	Klassische Geostatistik	382
	Kriging als Basisfunktionenansatz	384
	Schätzung von Kriging-Modellen	385
7.2.4	Markov-Zufallsfelder: Räumliche Glättung bei diskreter Lokationsvariable	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														Gesamt
	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	
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)

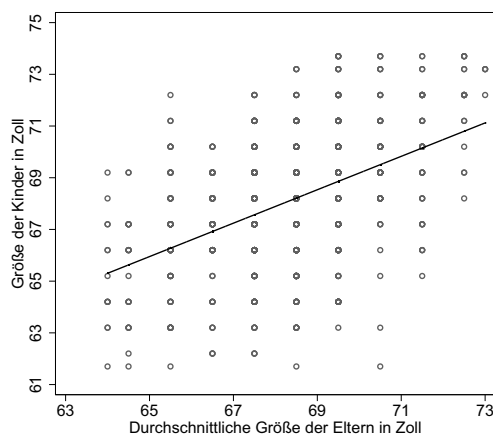


Abb. 1.1. Streudiagramm zwischen Größe der Kinder und Durchschnittsgröße der Eltern inklusive eingezeichneter Regressionsgerade.

den 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) , $i = 1, \dots, n$, 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 Gehirnnareale, 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 Mietspiegelschüre 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
<i>miete</i>	Nettomiete pro Monat (in DM)	895.90	381.53	79/3594.6
<i>mieteqm</i>	Nettomiete pro Monat und qm (in DM)	13.87	4.75	0.81/34.56
<i>flaeche</i>	Wohnfläche in qm	67.37	23.72	20/160
<i>bjahr</i>	Baujahr (in Jahren)	1956.31	22.31	1918/1997
<i>lage</i>	Lagekategorie gemäß Einschätzung durch Gutachter			
	1 = normale Lage	58.21		
	2 = gute Lage	39.26		
	3 = beste Lage	2.53		
<i>bad</i>	Ausstattung des Bades			
	0 = normal	93.80		
	1 = gehoben	6.20		
<i>kueche</i>	Ausstattung der Küche			
	0 = normal	95.75		
	1 = gehoben	4.25		
<i>zh</i>	Zentralheizung			
	0 = ohne Zentralheizung	10.42		
	1 = mit Zentralheizung	89.58		
<i>bez</i>	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
<i>zscore</i>	Z-Score des Kindes	-171.19	139.34	-600/503
<i>k_geschl</i>	Geschlecht des Kindes			
	1 = männlich	49.02		
	0 = weiblich	50.98		
<i>k_still</i>	Stilldauer in Monaten	11.11	9.42	0/46
<i>k_alter</i>	Alter des Kindes in Monaten	27.61	17.08	0/59
<i>m_alterg</i>	Alter der Mutter bei der Geburt in Jahren	26.40	6.87	13.16/48.66
<i>m_groesse</i>	Größe der Mutter in cm	158.06	5.99	134/185
<i>m_bmi</i>	Body-Mass-Index der Mutter	21.99	3.32	13.15/39.29
<i>m_bildung</i>	Ausbildung der Mutter			
	1 = keine Ausbildung	18.59		
	2 = Grundschule	62.34		
	3 = Volksschule	17.35		
	4 = höherer Abschluss	1.72		
<i>m_arbeit</i>	Erwerbsstatus der Mutter			
	1 = Mutter arbeitet	55.25		
	0 = Mutter arbeitet nicht	44.75		
<i>region</i>	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		
<i>district</i>	Wohnort in Zambia, 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
<i>einspruch</i>	Einspruch gegen das Patent			
	1 = Ja	41.49		
	0 = Nein	58.51		
<i>biopharm</i>	Patent aus der Biotechnologie- / Pharma-Branche			
	1 = Ja	44.31		
	0 = Nein	55.69		
<i>uszw</i>	US Zwillingspatent			
	1 = Ja	60.85		
	0 = Nein	39.15		
<i>patus</i>	Patentinhaber aus den USA			
	1 = Ja	33.74		
	0 = Nein	66.26		
<i>patdsg</i>	Patentinhaber aus Deutschland, der Schweiz oder Großbritannien			
	1 = Ja	23.49		
	0 = Nein	76.51		
<i>jahr</i>	Jahr der Patenterteilung			
	1980	0.18		
	⋮	⋮		
	1997	1.62		
<i>azit</i>	Anzahl der Zitationen für dieses Patent	1.64	2.74	0/40
<i>aland</i>	Anzahl der Länder, für die Patent- schutz gelten soll	7.8	4.12	1/17
<i>ansp</i>	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
<i>id</i>	Standort Identifikationsnummer			
<i>jahr</i>	Jahr der Erhebung	1993.59	6.34	1983/2004
<i>buche</i>	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		
<i>x</i>	x-Koordinate des Standorts			
<i>y</i>	y-Koordinate des Standorts			
<i>alter</i>	Bestandsalter, in Jahren	106.07	51.41	7/234
<i>schirm</i>	Beschirmungsgrad, d.h. Dichte der Laubdecke, in Prozent (0% – 100%)	77.29	23.70	0/100
<i>hang</i>	Hangneigung, in Prozent	15.45	11.27	0/46
<i>hoehe</i>	Höhe über dem Meeresspiegel, in Metern	386.99	58.88	250/480
<i>grund</i>	Gründigkeit, d.h. Bodentiefe, in der das Gestein beginnt, in cm	24.64	9.94	9/51
<i>ph</i>	pH-Wert in 0–2cm Tiefe	4.29	0.34	3.28/6.05
<i>frische</i>	Feuchtigkeitsstufe des Boden, in 3 Kategorien			
	1 = mäßig trocken	11.03		
	2 = mäßig frisch	55.12		
	3 = frisch oder mäßig wechselfeucht	33.85		
<i>alkali</i>	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		
<i>humus</i>	Dichte der Humusschicht in cm	1.57	1.38	0/9
<i>art</i>	Art des Waldes			
	0 = Mischwald	49.78		
	1 = Laubwald	50.22		
<i>dueng</i>	Düngung			
	0 = nicht gedüngt	80.90		
	1 = gedüngt	19.10		

Tabelle 1.5. Beschreibung der Variablen des Datensatzes zum Waldzustand.

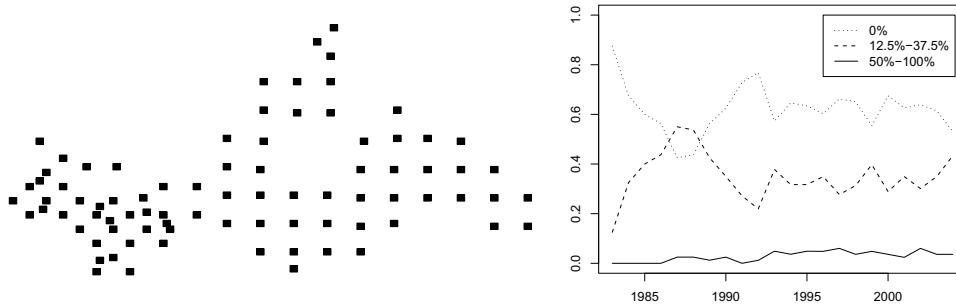


Abb. 1.2. Links: Beobachtungsstandorte. Im Zentrum befindet sich der Ort Rothenbuch. Rechts: Zeitlicher Trend der Schädigung.

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).

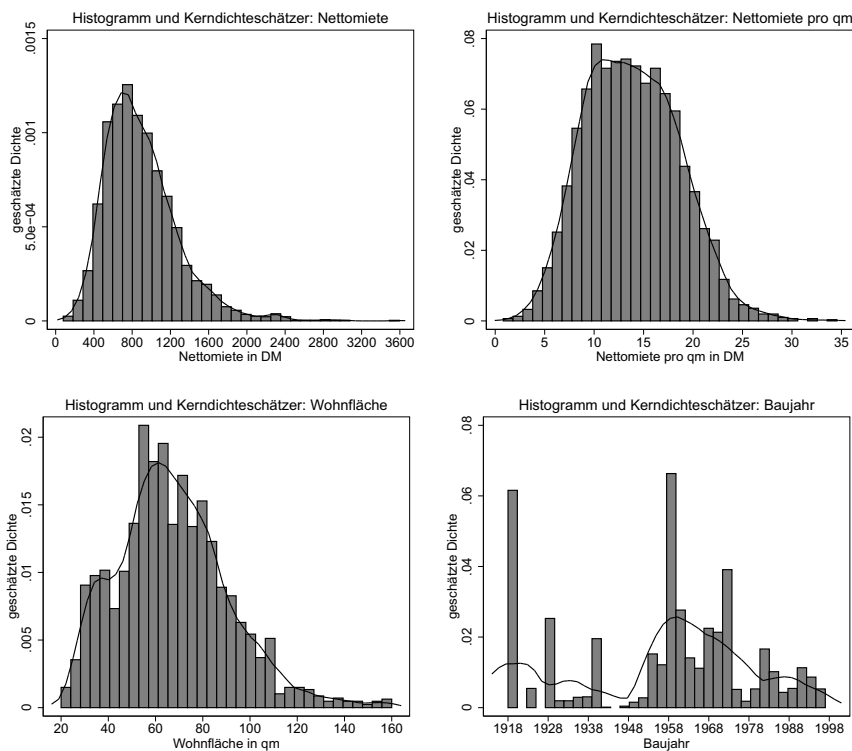


Abb. 1.3. *Mietspiegel: Histogramme und Kerndichteschätzer für die metrischen Variablen miete, mieteqm, flaeche und bjahr.*

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 teuren 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

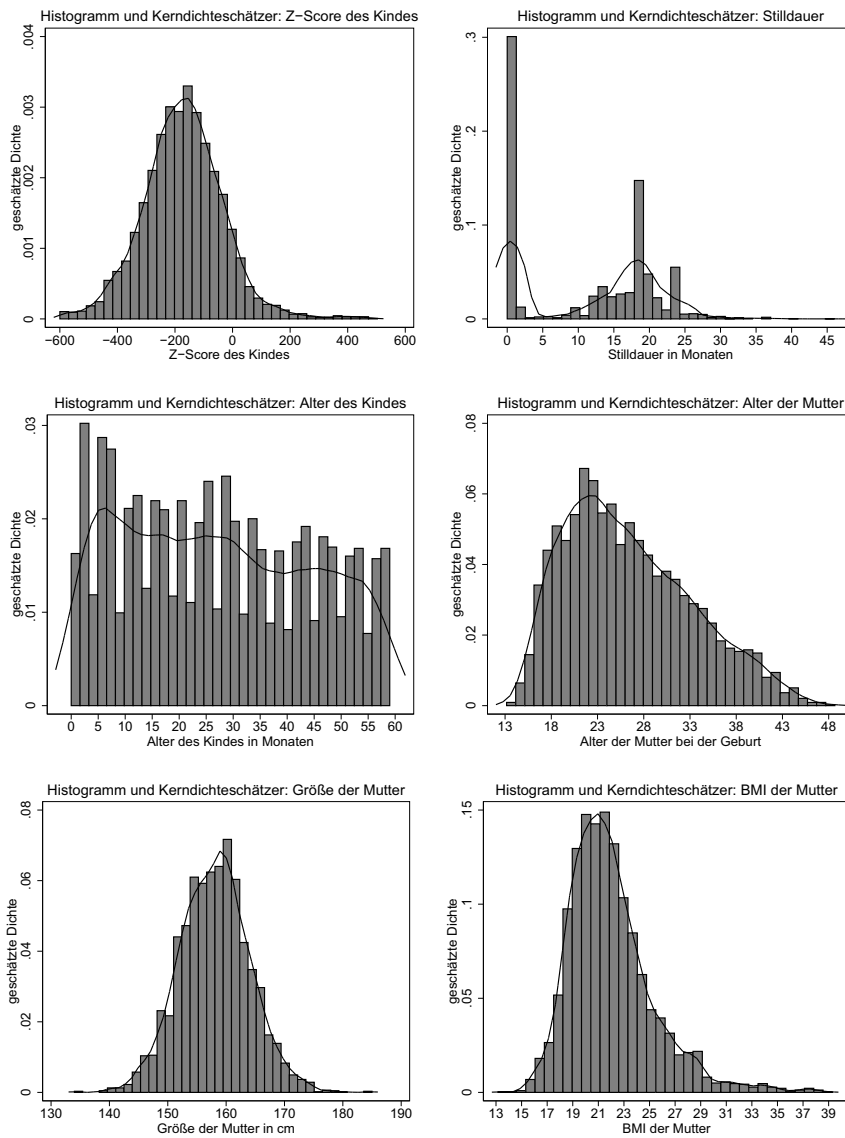


Abb. 1.4. *Unterernährung in Sambia: Verteilungen der metrischen Variablen.*

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-

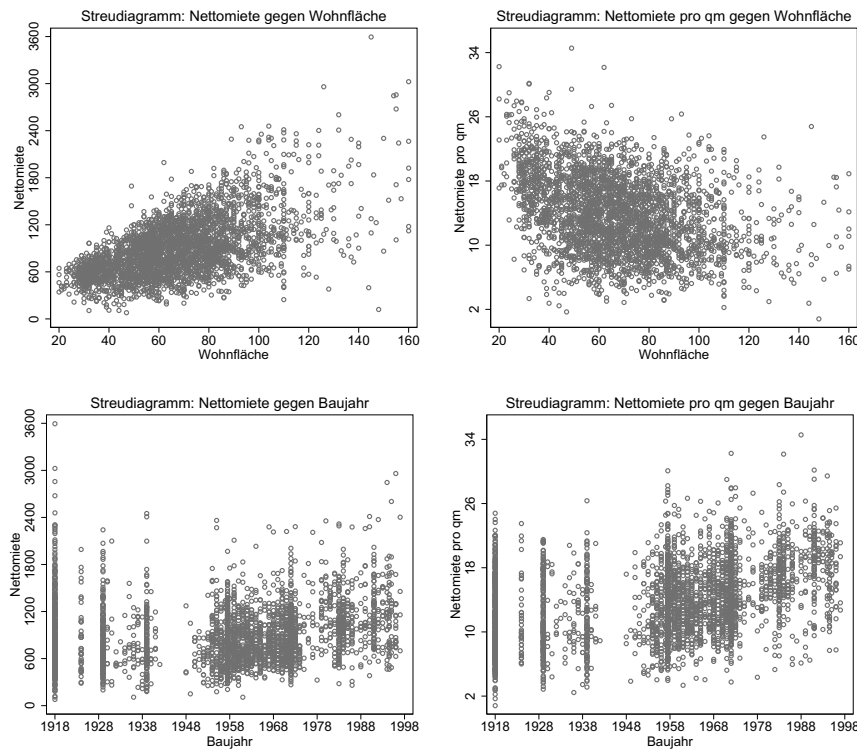


Abb. 1.5. *Mietspiegel: Streudiagramme zwischen Nettomiete bzw. Nettomiete pro qm und den erklärenden Variablen Wohnfläche und Baujahr.*

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

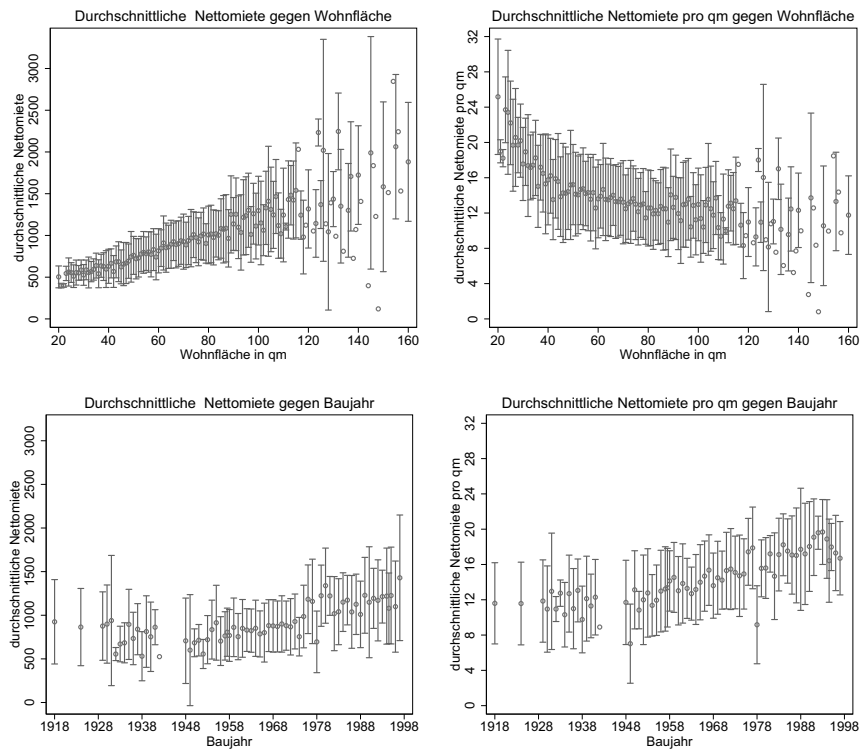


Abb. 1.6. Mittelwerte plus/minus eine Standardabweichung der Nettomiete bzw. Nettomiete pro qm versus Wohnfläche und Baujahr.

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.

△

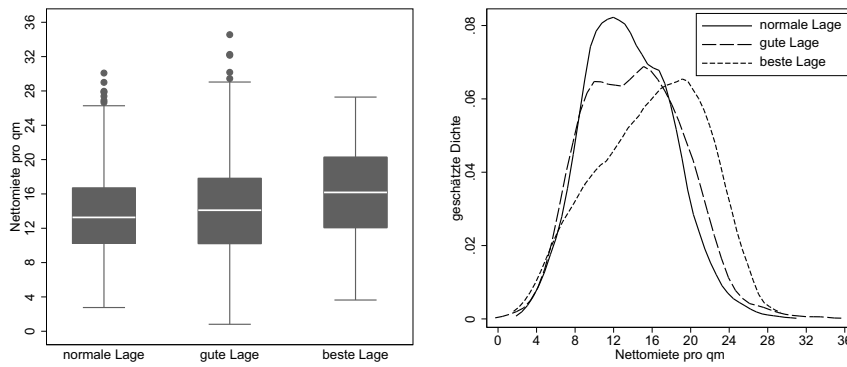


Abb. 1.7. Verteilung der Nettomiete pro qm in Abhängigkeit von der Wohnlage.

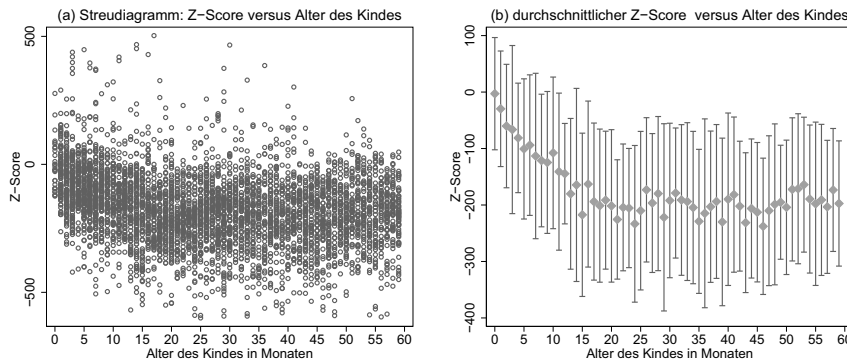


Abb. 1.8. Unterernährung in Sambia: Verschiedene grafische Darstellungen des Zusammenhangs zwischen Z-Score und Alter des Kindes.

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.

△

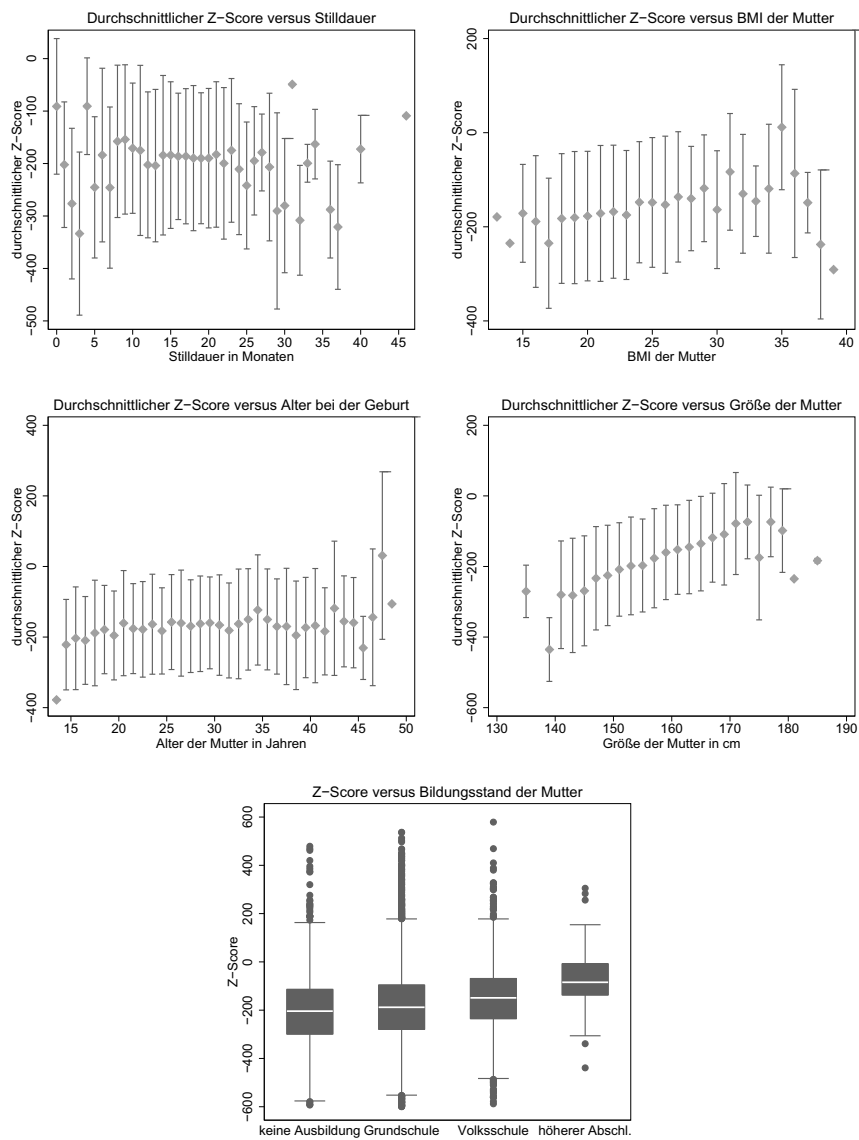


Abb. 1.9. *Unterernährung in Sambia: Grafische Darstellung des Zusammenhangs zwischen Z-Score und ausgewählten erklärenden Variablen.*

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, \dots, 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 $f(x_1, \dots, x_k)$ von x_1, \dots, 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 $E(y | x_1, \dots, x_k)$ von y in Abhängigkeit der Kovariablen. Der Erwartungswert ist also eine Funktion der Kovariablen:

$$E(y | x_1, \dots, x_k) = f(x_1, \dots, x_k)$$

Die Zielgröße lässt sich dann immer zerlegen in

$$y = E(y | x_1, \dots, x_k) + \varepsilon = f(x_1, \dots, x_k) + \varepsilon,$$

wobei ε die zufällige, nicht von den Kovariablen erklärte Abweichung vom Erwartungswert ist. Häufig bezeichnet man $f(x_1, \dots, x_k)$ 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 $y_i, x_{i1}, \dots, x_{ik}$, $i = 1, \dots, n$, zu schätzen und von der stochastischen Komponente ε zu trennen.

Am bekanntesten ist die Klasse der *linearen Regressionsmodelle*

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

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

$$E(y | x_1, \dots, x_k) = f(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

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

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

mit den unbekannten Parametern bzw. Regressionskoeffizienten β_0, \dots, β_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:

$$miete_i = \beta_0 + \beta_1 \cdot flaeche_i + \varepsilon_i. \quad (2.1)$$

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

$$E(miete | flaeche) = \beta_0 + \beta_1 \cdot flaeche,$$

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 $E(y | x)$ in der allgemeineren Beziehung

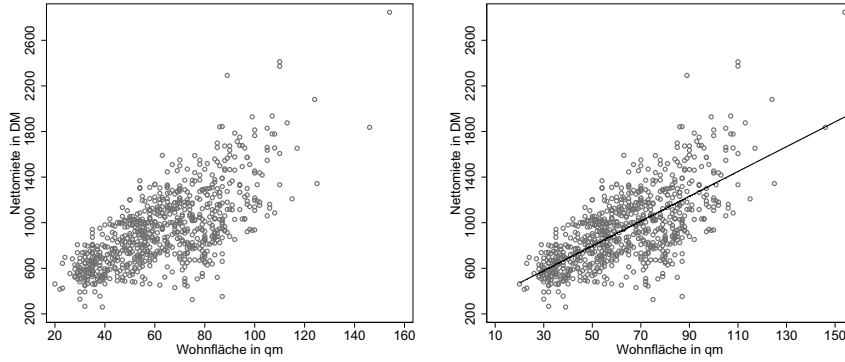


Abb. 2.1. Streudiagramm zwischen Nettomiete und Wohnfläche für nach 1966 gebaute Wohnungen in normaler Wohnlage (links). In der rechten Grafik ist zusätzlich die Regressionsgerade mit eingezeichnet.

$$y = f(x) + \varepsilon = E(y|x) + \varepsilon$$

als linear, d.h. $f(x) = E(y|x) = \beta_0 + \beta_1 x$ angenommen wird.

Genauer werden für das *Standardmodell der linearen Einfachregression* folgende Annahmen getroffen: Es gilt

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.2)$$

wobei die Fehlervariablen ε_i unabhängig und identisch mit

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

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*

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

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

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

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

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

bei gegebenen Daten (y_i, x_i) , $i = 1, \dots, n$, 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 $\widehat{E(y|x)}$ für den bedingten Erwartungswert von y bei gegebenem

Standardmodell der linearen Einfachregression

Daten

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

Modell

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

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

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

Die geschätzte Regressionsgerade $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ kann als Schätzung $\widehat{E(y|x)}$ 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 $\text{Var}(\varepsilon_i) = \text{Var}(y_i) = \sigma^2$ 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

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

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

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

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

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

definiert und eine Regression der Form

$$\text{mieteqm}_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \beta_0 + \beta_1 \frac{1}{\text{flaeche}_i} + \varepsilon_i \quad (2.3)$$

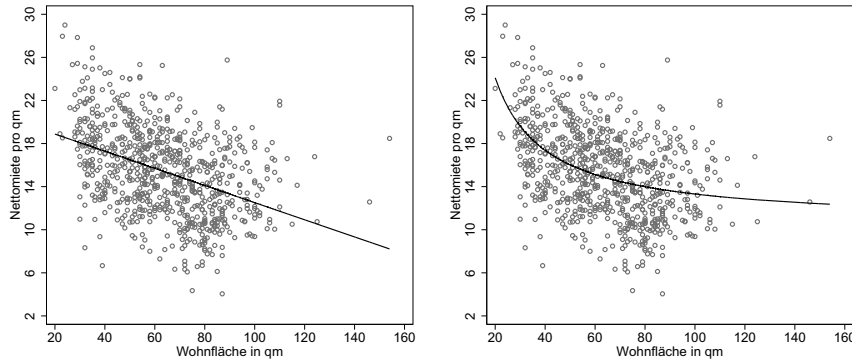


Abb. 2.2. Streudiagramm zwischen Nettomiete pro qm und Wohnfläche und geschätzte Funktionen \hat{f} bei Verwendung der Wohnfläche (links) und der inversen Wohnfläche (rechts) als erklärende Variable.

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

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

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

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

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. $\text{flaeche} = 30$ qm, ist

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

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

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

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:

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

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

$$\widehat{\text{mieteqm}}(60) - \widehat{\text{mieteqm}}(61) \approx 0.07 \text{ DM}.$$

△

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

$$miete_i = \beta_0 + \beta_1 flaeche_i + \beta_2 glage_i + \varepsilon_i \quad (2.4)$$

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

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

Mit der KQ-Methode erhält man als geschätzte Durchschnittsmiete

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

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

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

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, \dots, x_k :

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

Dabei ist x_{ij} der Wert der j -ten Kovariable für die i -te Beobachtung, $i = 1, \dots, n$. 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:

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

mit

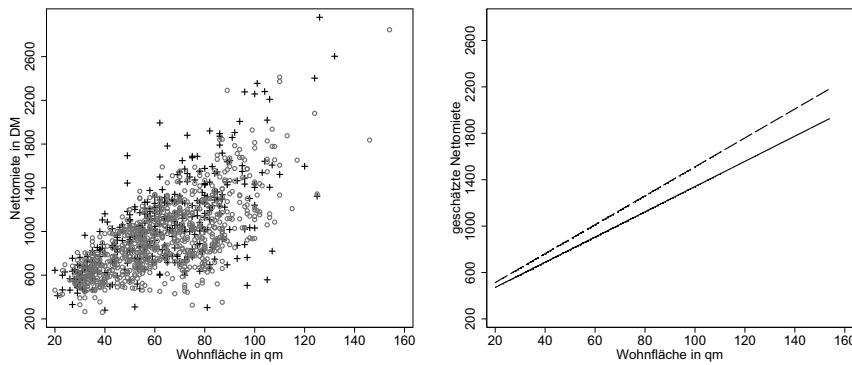


Abb. 2.3. Die linke Grafik zeigt das Streudiagramm zwischen Nettomiete und Wohnfläche für Wohnungen in normaler (Kreise) und guter Lage (Pluszeichen). Die rechte Grafik zeigt separat geschätzte Regressionsgeraden für Wohnungen in normaler (durchgezogene Linie) und guter Lage (gestrichelte Linie).

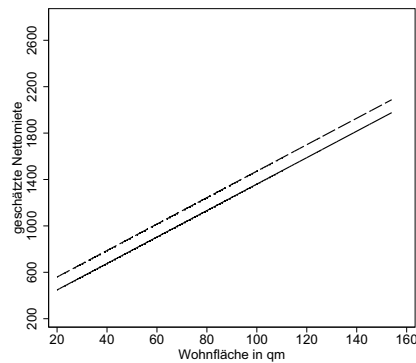


Abb. 2.4. Nach Modell (2.4) geschätzte Regressionsgeraden für Wohnungen in normaler (durchgezogene Linie) und guter Lage (gestrichelte Linie).

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

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 $x = \frac{1}{\text{flaeche}}$ und formulieren

$$\text{mieteqm}_i = \beta_0 + \beta_1 \cdot \frac{1}{\text{flaeche}_i} + \beta_2 \text{glage}_i + \varepsilon_i \quad (2.5)$$

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

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

Die beiden Kurven für die durchschnittlichen Quadratmetermieten

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

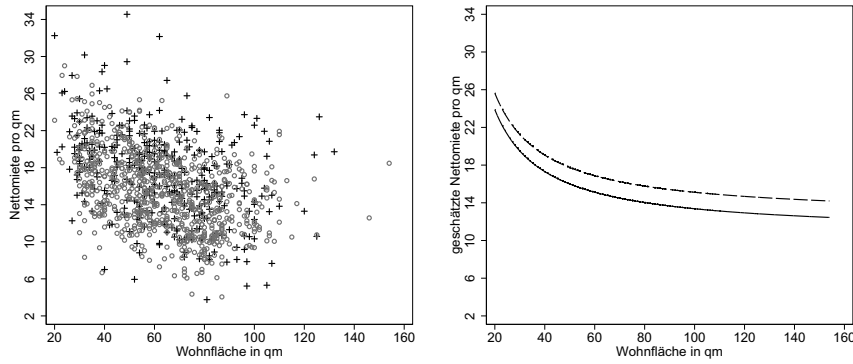


Abb. 2.5. Links: Streudiagramm zwischen Nettomiete pro qm und Wohnfläche für Wohnungen in normaler (Kreise) und guter Lage (Pluszeichen). Rechts: Geschätzte Regressionskurven für Wohnungen in normaler (durchgezogene Linie) und guter Lage (gestrichelte Linie).

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 2.4 111.66 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* $inter = flaeche \cdot glage$ einbeziehen, zu

$$miete_i = \beta_0 + \beta_1 flaeche_i + \beta_2 glage_i + \beta_3 inter_i + \varepsilon_i. \quad (2.6)$$

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

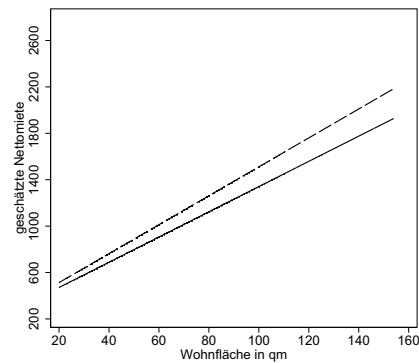


Abb. 2.6. Basierend auf dem Interaktionsmodell (2.6) geschätzte Regressionsgeraden für normale (durchgezogene Linie) und gute Wohnlagen (gestrichelte Linie).

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

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

geprüft werden, vergleiche Kapitel 3.3.

△

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

- 1 = normale Lage
- 2 = gute Lage
- 3 = beste Lage

eingeteilt. Da die Lagevariable kategorial und nicht metrisch ist, kann der Effekt der Lage nicht in der Form $\beta \cdot \text{lage}$, 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 $x \in \{1, \dots, c\}$, 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, \dots, x_{c-1} :

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

Für die Referenzkategorie c gilt dann

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

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

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

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 *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 Zentralheizungsindekatoren, 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

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

Modell

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

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

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

Die geschätzte lineare Funktion

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

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

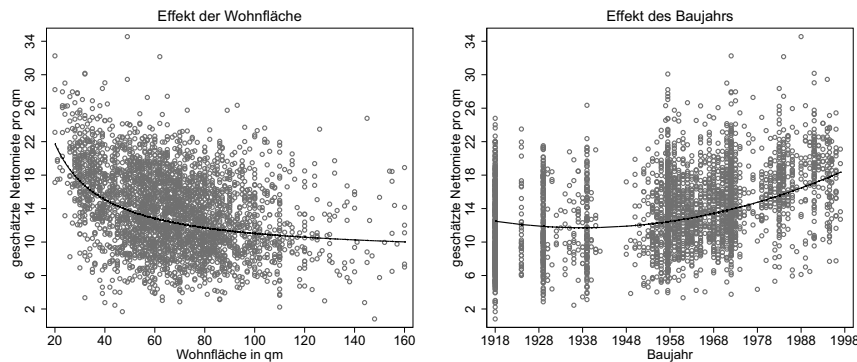


Abb. 2.7. Effekte der Wohnfläche (links) und des Baujahrs (rechts).

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

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

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

$$P(y = 1) = P(y = 1 \mid x_1, \dots, x_k) = \pi$$

in Abhängigkeit von den Kovariablen. Ein übliches lineares Regressionsmodell

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

mit $\varepsilon_i \sim N(0, \sigma^2)$ 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 $\text{Var}(\varepsilon_i) = \text{Var}(y_i | x_i)$ nicht homoskedastisch, d.h. gleich σ^2 sein. Da y_i Bernoulli-verteilt ist mit $\pi_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$, folgt, dass

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

ebenfalls von den Kovariablen und den Parametern β_0, \dots, β_k abhängt und somit nicht für alle i den gleichen Wert σ^2 besitzen kann.

- Das lineare Modell lässt für $P(y_i = 1)$ 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

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

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

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

mit dem *linearen Prädiktor*

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

Analog zum linearen Regressionsmodell wird angenommen, dass die binären Zielvariablen y_i bei gegebenen Kovariablenwerten $x_i = (x_{i1}, \dots, x_{ik})'$ (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

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

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

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

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

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

Modell

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

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

mit dem linearen Prädiktor

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

gewählt. Dazu äquivalent ist die Annahme

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

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

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

Ist x_1 speziell eine binäre Variable, so gilt

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

Für $\beta_1 > 0$ vergrößert sich also die Chance $P(y_i = 1)/P(y_i = 0)$, 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.

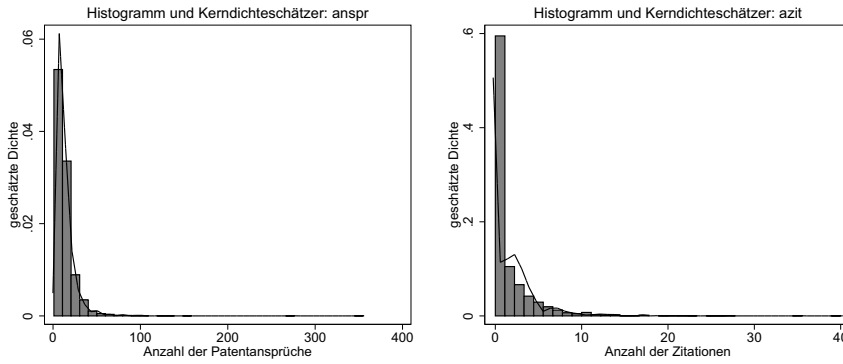


Abb. 2.8. Histogramme und Kerndichteschätzer für die metrischen Kovariablen *anspr* (links) und *azit* (rechts).

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 *anspr* und *azit* abgedruckt sind. Die Verteilungen beider Variablen sind extrem linkssteil. Der Großteil der Beobachtungen für *anspr* 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 *anspr* > 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

$$P(\text{einspruch}_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1 \text{jahr}_i + \beta_2 \text{azit}_i + \beta_3 \text{anspr}_i + \beta_4 \text{uszw}_i + \beta_5 \text{patus}_i + \beta_6 \text{patdsg}_i + \beta_7 \text{aland}_i$$

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

$$\frac{P(\text{Einspruch})}{P(\text{kein Einspruch})} = \exp(201.74) \cdot \exp(-0.102 \cdot \text{jahr}_i) \cdot \dots \cdot \exp(0.097 \cdot \text{aland}_i).$$

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 $P(\text{Einspruch}) / P(\text{kein Einspruch})$ mit Hilfe des Modells prognostizieren.

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

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

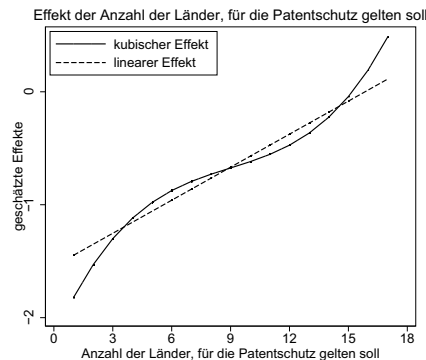


Abb. 2.9. *Einsprüche gegen Patente: Linearer und kubischer Effekt der Variable *aland*.*

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

$$\beta_7 \text{ aland} + \beta_8 \text{ aland}^2 + \beta_9 \text{ aland}^3.$$

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 $\{0, 1, 2, \dots\}$ 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 Querschnittstudien auftreten. Die Regressionskoeffizienten β_0, \dots, β_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 individuenpezifische 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 β_0, \dots, β_k zusätzlich individuen- bzw. clusterspezifische zufällige Effekte mit ein. Deren Modellierung und Schätzung ermöglicht weitergehende Analysen auf individuen spezifischer 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} , $j = 1, \dots, n_i$, dieser Zielvariable war für die Ratten $i = 1, \dots, 50$, 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 $\{y_{ij}, j = 1, \dots, n_i\}$ für die Ratten $i = 1, \dots, 50$.

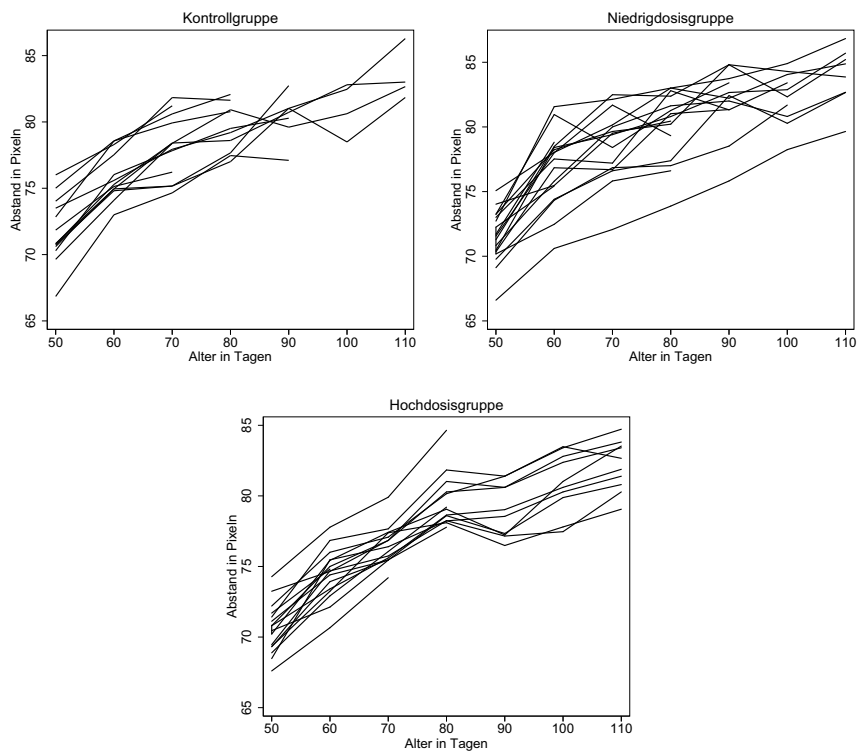


Abb. 2.10. Hormontherapie bei Ratten: Nach Dosierungsgruppe getrennte Zeitreihen.

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*

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

Der Wert $t = 0$ entspricht dann dem Behandlungsbeginn (Alter = 45 Tage).

Für die drei Gruppen definieren wir die Indikatorvariablen C , N , H

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

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 $E(y_{ij}) = \beta_0$, 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

$$y_{ij} = \beta_0 + \beta_1 N_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} + \varepsilon_{ij} \quad (2.11)$$

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 individuentpezifischer Information wirkt sich deshalb auch positiv auf die Qualität 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

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

mit individuentpezifischen Abweichungen γ_{0i} von der Populationskonstanten β_0 und individuentpezifischen Abweichungen γ_{1i} von den Populationssteigungen β_1 , β_2 und β_3 .

Im Gegensatz zu den „fixen“ Effekten $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ werden die individuentpezifischen Effekte $\gamma_i = (\gamma_{0i}, \gamma_{1i})'$ 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

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

Die Erwartungswerte können dabei ohne Einschränkung gleich Null gesetzt werden, da die Populationsmittelwerte bereits in den fixen Effekten β enthalten sind.

Lineare gemischte Modelle für Longitudinal- und Clusterdaten

Daten

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

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

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

Modell

Für ein lineares gemischtes Modell wird

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

$i = 1, \dots, m$, $j = 1, \dots, n_i$, angenommen. Dabei sind β_0, \dots, β_k feste *Populationseffekte* und $\gamma_{0i}, \gamma_{1i}, \dots, \gamma_{li}$ *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.

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

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

Die Kovariablen x_{ij1}, \dots, 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 $t_{i1} < \dots < t_{ij} < \dots < t_{in_i}$ für Individuum i bzw. für Objekte $j = 1, \dots, n_i$ im Cluster i in der Form

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

$i = 1, \dots, m$, $j = 1, \dots, n_i$, schreiben. Die festen Parameter β_0, \dots, β_k messen dabei Populationseffekte, während die zufälligen Parameter $\gamma_{0i}, \gamma_{1i}, \dots, \gamma_{li}$ individuen- bzw. clusterspezifische Effekte beschreiben. Die zusätzlichen Designvariablen u_{ij1}, \dots, u_{ijl} bestehen oft aus einem Teil der Kovariablen x_{ij1}, \dots, 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-)

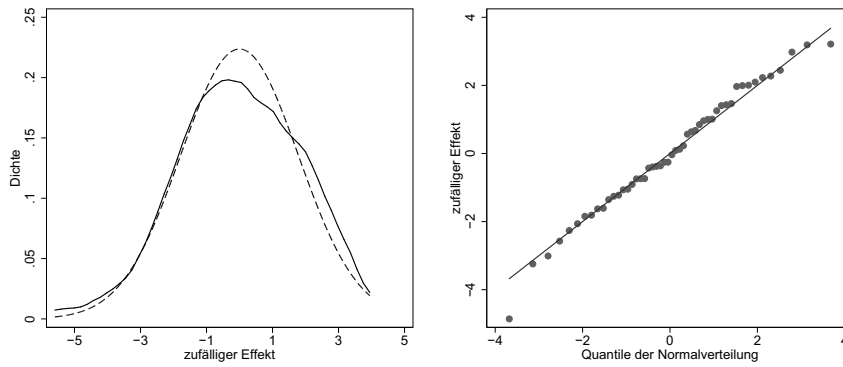


Abb. 2.11. Links: Kerndichteschätzer (durchgezogene Linie) und Normalverteilungsdichte (gestrichelt). Rechts: Normal-Quantil-Diagramm.

verteilt sind mit $E(\varepsilon_{ij}) = 0$ und $\text{Var}(\varepsilon_{ij}) = \sigma^2$. Es ist aber auch möglich, Korrelationen zwischen den Fehlern ε_{ij} , $j = 1, \dots, n_i$, 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 $\text{Var}(\gamma_{1i})$ 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 $\hat{\gamma}_{0i}$, $i = 1, \dots, 50$ 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
$\text{Var}(\gamma_{0i})$	τ_0^2	3.739	3.565
$\text{Var}(\gamma_{1i})$	τ_1^2	<0.001	
$\text{Var}(\varepsilon_{ij})$	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

$$f(\text{flaeche}) = \beta_0 + \beta_1/\text{flaeche}. \quad (2.15)$$

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. $f(\text{flaeche}) = \beta_0 + \beta_1 \log(\text{flaeche})$ oder $f(\text{flaeche}) = \beta_0 + \beta_1(\text{flaeche})^{\frac{1}{2}}$. 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

$$y_i = f(x_i) + \varepsilon_i \quad (2.16)$$

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

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

Modell

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n.$$

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

$$k_alter_1 < k_alter_2 < \dots < k_alter_m,$$

für die Funktionsschätzungen $\hat{f}(k_alter_j), j = 1, \dots, m$, 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 $\hat{f}(k_alter_1), \dots, \hat{f}(k_alter_m)$ 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

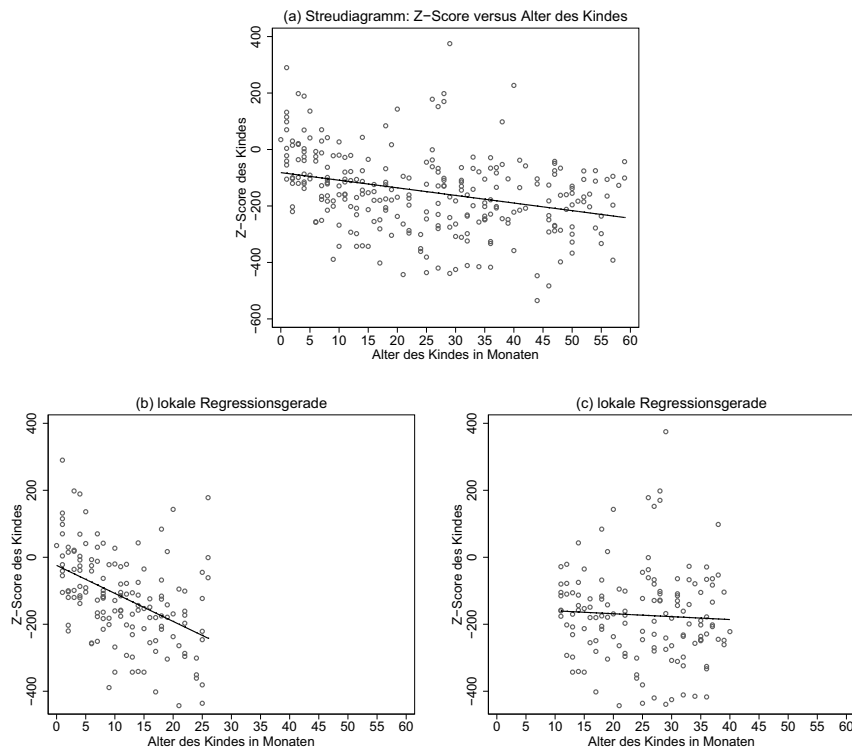


Abb. 2.12. Illustration globale und lokale Regression. Abbildung a) zeigt eine (globale) Regressionsgerade zur Schätzung des Zusammenhangs zwischen Z-Score und Alter des Kindes. Die Abbildung b) und c) zeigen lokale Regressionsgeraden basierend auf einem Ausschnitt der Daten.

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

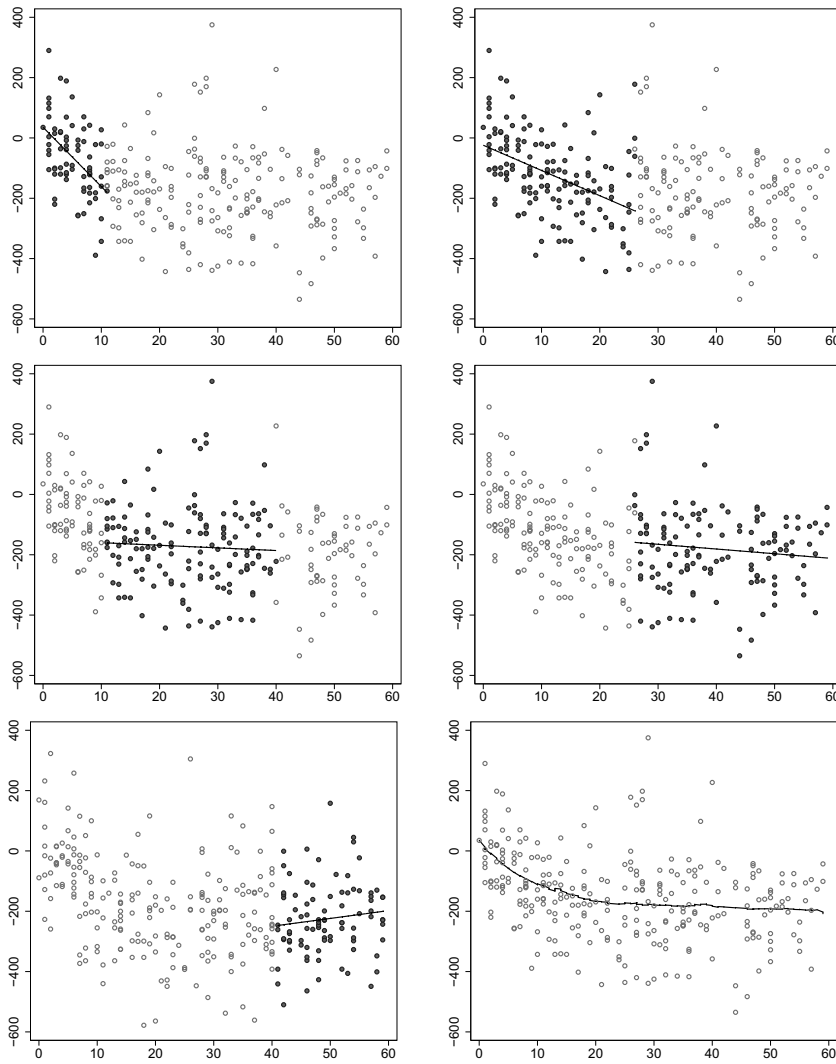


Abb. 2.13. Veranschaulichung einer nichtparametrischen Nächste-Nachbarn-Schätzung für den Zusammenhang zwischen Z-Score und Alter des Kindes.

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.

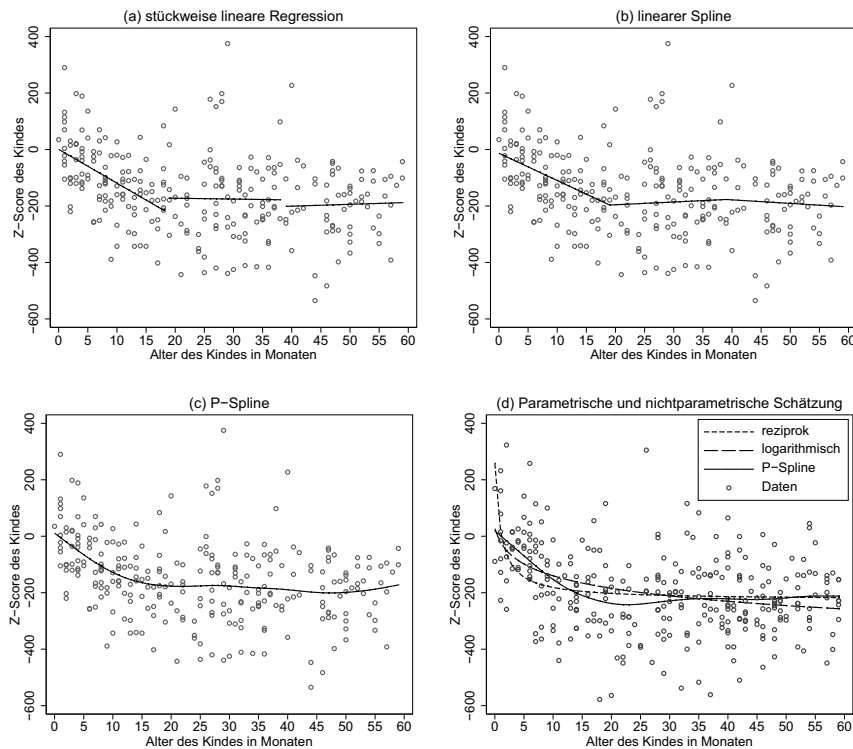


Abb. 2.14. Stückweise lineare Regression (Grafik a), linearer Spline (Grafik b) und kubischer P-Spline (Grafik c) zur Schätzung des Zusammenhangs zwischen Z-Score und Alter des Kindes, sowie Vergleich zwischen parametrischer und nichtparametrischer Regression (Grafik d).

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 nicht-linear ist, wird statt eines linearen Regressionsmodells ein *additives Modell*

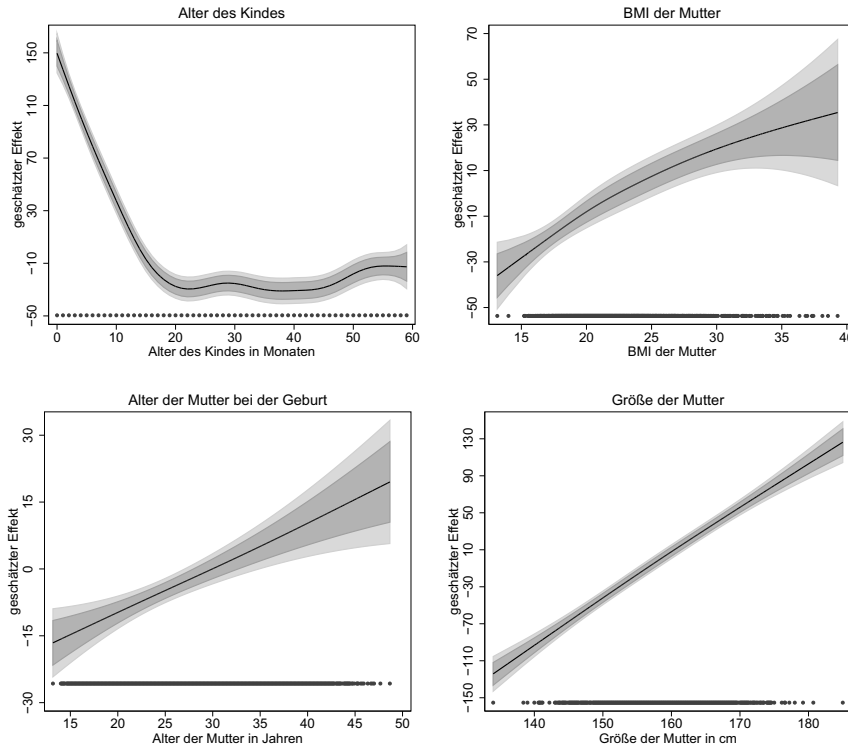


Abb. 2.15. Geschätzte nichtlineare Funktionen inklusive 80% und 95% punkweisen Konfidenzintervallen. Die Punkte im unteren Teil der Grafiken deuten die Verteilung der Kovariablenwerte an.

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

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 β_1, β_2, \dots 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 β_0, β_1, \dots der kategorialen Kovariablen geschätzt. Das Modell ist wegen der Effekte f_1, \dots, 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

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

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

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

wobei für die Fehlervariablen die gleichen Annahmen wie im linearen Regressionsmodell gelten. Die unspezifizierten glatten Funktionen f_1, \dots, f_q stellen die (Haupt-)Effekte der metrischen Kovariablen z_1, \dots, z_q dar und werden mit nichtparametrischen Techniken geschätzt, siehe Kapitel 8. Die Kovariablen x_1, \dots, 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

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

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

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

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 $\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$ die KQ-Methode und für die Schätzung der Funktionen f_1, \dots, f_q nichtparametrische Methoden, z.B. basierend auf Splines, verwendet werden können.

Standardmodell der additiven Regression**Daten**

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

Modell

Es gilt

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

Für die Fehlervariablen ε_i werden die gleichen Annahmen getroffen wie im klassischen linearen Regressionsmodell. Die Funktionen $f_1(z_1), \dots, f_q(z_q)$ werden als „glatt“ vorausgesetzt und beschreiben nichtlineare Effekte der stetigen oder metrischen Kovariablen z_1, \dots, 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 (*alterkfv*), 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 $0, 1, 2, \dots$ annehmen kann, sind Regressionsmodelle für metrische Zielgrößen nicht geeignet.

△

Eine häufig angemessene Verteilung für Zähldaten ist die Poisson-Verteilung. Wir können also annehmen, dass $y \sim \text{Po}(\lambda)$ gilt, mit erwarteter Häufigkeit $\lambda = E(y)$. 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 $y_i \sim \text{Po}(\lambda_i)$ durch

$$\lambda_i = E(y_i) = \exp(\eta_i)$$

mit einem *additiven Prädiktor*

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

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

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

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

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

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

$$\eta_i = f_1(\text{alter}_i) + f_2(\text{alterkfz}_i) + f_3(\text{ps}_i) + f_4(\text{bm}_i) + \beta_0 + \beta_1 \text{geschl}_i + \dots$$

mit möglicherweise nichtlinearen Funktionen f_1, \dots, f_4 der Variablen *alter*, *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 $\hat{f}_1, \dots, \hat{f}_4$ 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.

△

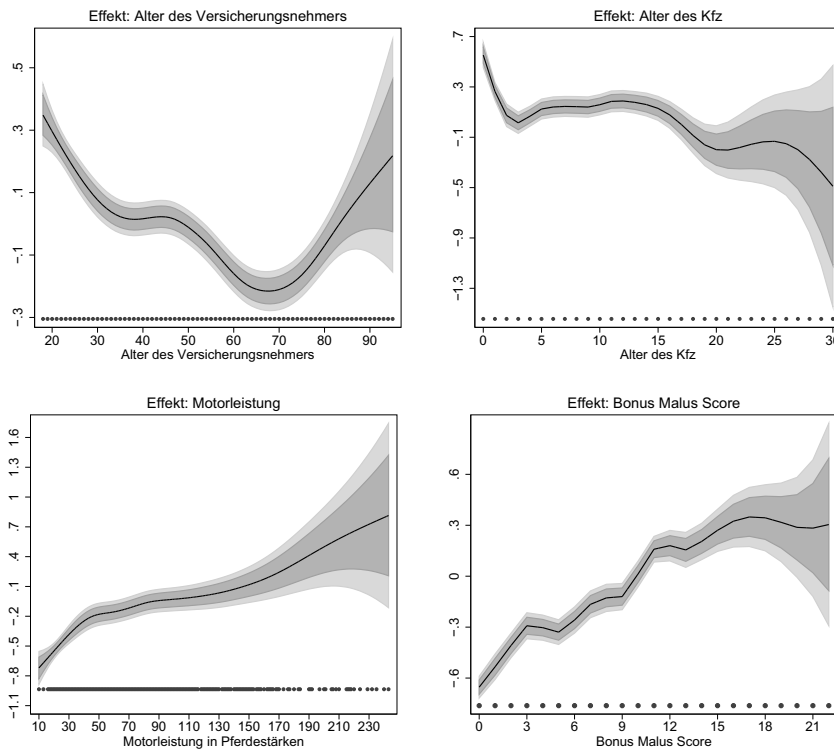


Abb. 2.16. Kfz-Versicherungsdaten: Geschätzte nichtlineare Funktionen inklusive 80% und 95% punktwisen Konfidenzintervallen. Die Punkte im unteren Teil der Grafik deuten die Verteilung der Kovariablenwerte an.

2.8 Geoadditive Regression

In vielen Anwendungen ist für die Individuen oder Untersuchungseinheiten $i = 1, \dots, n$ neben den Werten $(y_i, x_{i1}, \dots, x_{ik}, z_{i1}, \dots, z_{iq})$ 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 regionspezifischen 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

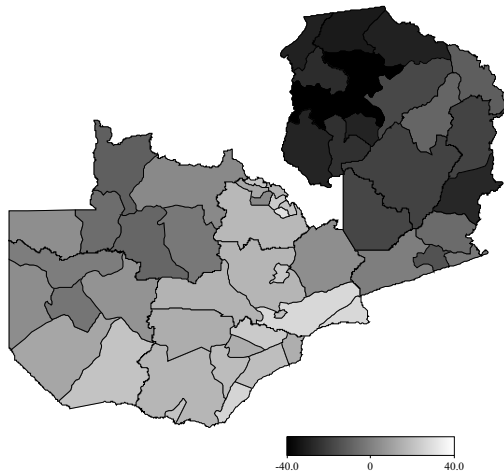


Abb. 2.17. Räumlicher Effekt für die Sambia-Daten.

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 $f_{geo}(\text{district})$ 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. $f(\text{alter})$. Wir reanalysieren somit die Daten mit folgendem *geoadditiven* Modell:

$$\text{zscore} = f_1(k_alter) + f_2(m_bmi) + f_3(m_alterg) + f_4(m_groesse) + f_{geo}(\text{district}) + \beta_0 + \beta_1 m_bildung1 + \dots + \beta_4 m_arbeit + \varepsilon$$

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 *geoadditiven* 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, \dots, z_q und den üblichen Kovariablen x_1, \dots, x_k liegen noch Beobachtungen zur räumlichen Lokation s vor.

Modell

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

wobei für die Fehlervariablen ε_i die gleichen Annahmen wie für das klassische lineare Regressionsmodell getroffen werden. Die unbekannten glatten Funktionen f_1, \dots, 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 $s \in \{1, \dots, S\}$ 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*

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

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

$$\eta_i = f_1(alterv_i) + f_2(alterkfi_i) + f_3(psi_i) + f_4(bmi_i) + f_{geo}(district_i) + \dots,$$

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,

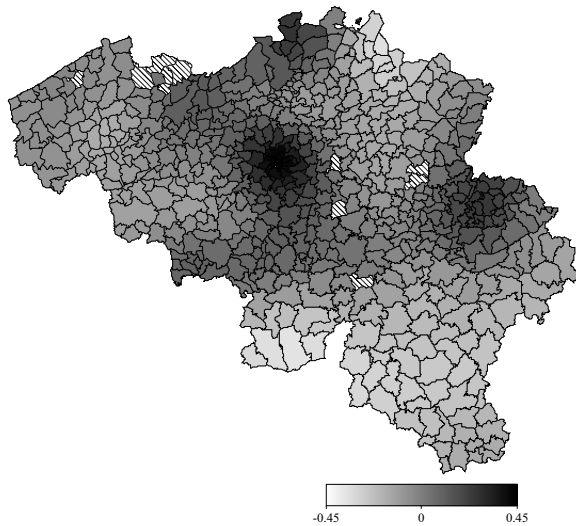


Abb. 2.18. Geschätzter räumlicher Effekt für die Kfz-Versicherungsdaten.

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

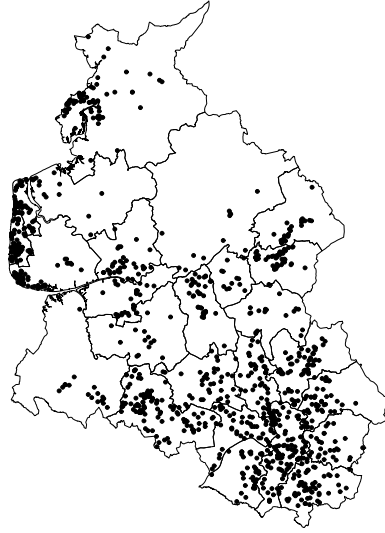


Abb. 2.19. Verteilung der Beobachtungen in Northwest-England. Jeder Punkt entspricht einer Beobachtung.

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

$$\lambda_i(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_i \leq t + \Delta t \mid T_i \geq t)}{\Delta t}.$$

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 *geoaditives 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 geoaditiven Prädiktor verknüpft:

$$\lambda_i(t) = \exp [g(t) + f_1(\text{alter}_i) + f_2(\text{ti}_i) + f_{\text{geo}}(s_i) + \beta_0 + \beta_1 \text{wb}_i + \beta_2 \text{sex}]$$

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*.

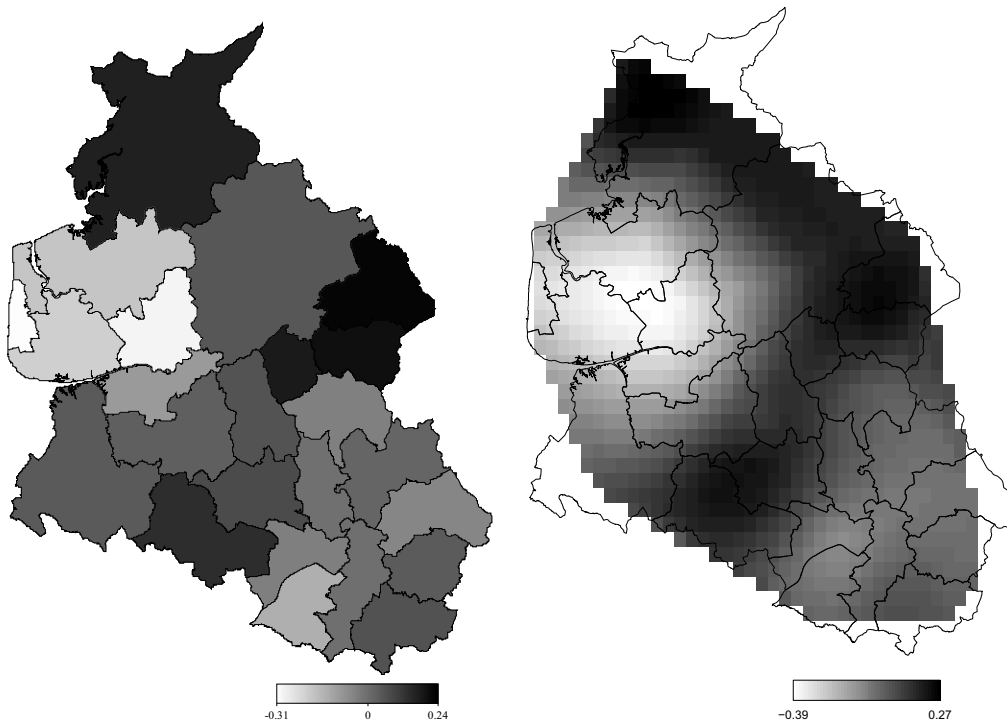


Abb. 2.20. Geschätzte räumliche Effekte basierend auf Distrikten (links) beziehungsweise den exakten Koordinaten der Beobachtungen (rechts).

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 $g(t)$, $f_1(\text{alter})$ 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.

△

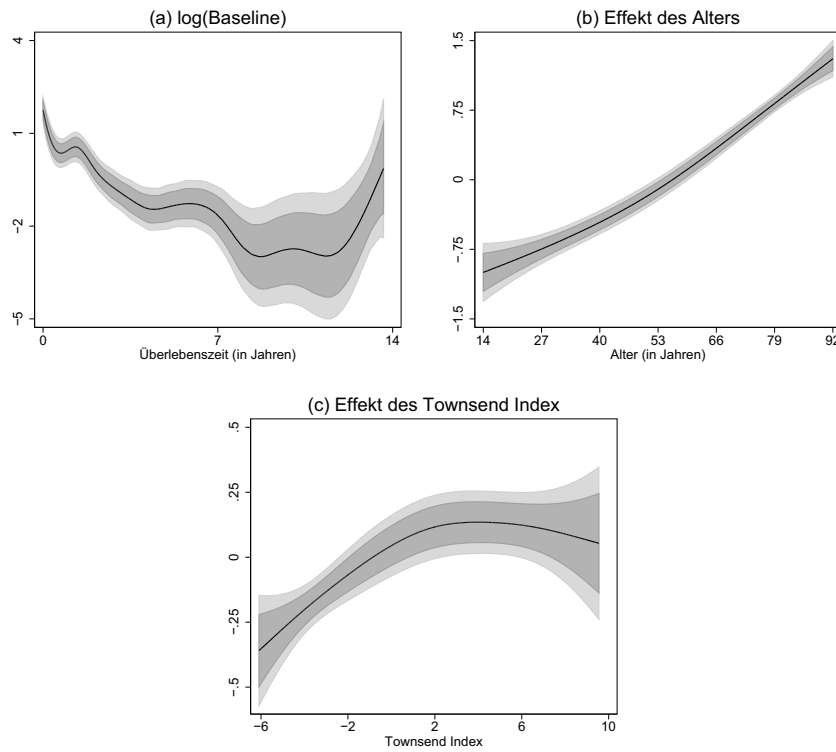


Abb. 2.21. Geschätzte nichtlineare Kovariableneffekte mit 80% und 95% Konfidenzintervallen.

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

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

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

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

- *Erwartungswert:*

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

- *Prädiktor:*

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

2.9.2 Logit-Modell (Kapitel 4)

- *Zielgröße:* Die Beobachtungen $y_i \in \{0, 1\}$ sind binär und $B(1, \pi_i)$ verteilt.
- *Erwartungswert:*

$$E(y_i) = P(y_i = 1) = \pi_i = \frac{\exp(\eta_i^{lin})}{1 + \exp(\eta_i^{lin})}.$$

- *Prädiktor:*

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

2.9.3 Poisson-Regression (Kapitel 4)

- *Zielgröße:* Die Beobachtungen $y_i \in \{0, 1, 2, \dots\}$ 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:*

$$E(y_i) = \lambda_i = \exp(\eta_i^{lin}).$$

- *Prädiktor:*

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

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:*

$$E(y_i) = \mu_i = h(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}) = h(\eta_i^{lin}),$$

wobei h eine (bekannte) Verknüpfungsfunktion ist, wie zum Beispiel $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$ im Logit-Modell.

- *Prädiktor:*

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

- *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} , $i = 1, \dots, m$, $j = 1, \dots, n_i$ 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:*

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

Die zufälligen Effekte $\gamma_{0i}, \gamma_{1i}, \dots$ werden als unabhängig und identisch verteilt betrachtet.

- *Prädiktor:*

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

- *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:*

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

- *Prädiktor:*

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

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

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

oder

$$\eta_i = \eta_i^{add} + f(z_1) x_1 + \dots$$

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

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

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:*

$$E(y_i) = \mu_i = h(\eta_i^{add})$$

mit der (bekannten) Verknüpfungsfunktion h .

- *Prädiktor:*

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

- *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:*

$$E(y_i) = \mu_i = h(\eta_i)$$

mit der Verknüpfungsfunktion h .

- *Prädiktor:*

$$\eta_i = f_1(v_{i1}) + \dots + f_q(v_{iq}) + \eta_i^{lin}$$

Dabei sind v_1, \dots, v_q ein- oder auch mehrdimensionale Variablen verschiedenen Typs, die aus den Kovariablen gebildet werden. Ebenso sind auch die Funktionen f_1, \dots, f_q von unterschiedlichem Typ. Beispiele sind:

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

- *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.