

ÜBERPRÜFUNG STOCHASTISCHER MODELLE MIT PSEUDO-RESIDUEN

Dissertation

zur Erlangung des wirtschaftswissenschaftlichen Doktorgrades

an der Wirtschaftswissenschaftlichen Fakultät

der Universität Göttingen

vorgelegt von

Andreas Stadie

aus Langenhagen

Göttingen, 2002

Erstgutachter: Prof. Dr. Walter Zucchini
Zweitgutachter: Prof. Dr. Fred Böker
Tag der mündlichen Prüfung: 05.02.2003

Inhaltsverzeichnis

1	Einleitung	1
1.1	Problemstellung und Ziele	1
1.2	Aufbau der Arbeit	3
2	Theoretische Grundlagen	7
2.1	Grundlegende Möglichkeiten bei der Definition von Residuen	7
2.2	Definition eines stochastischen Modells	15
2.3	Pseudo-Residuen bei der Modellierung stetiger Merkmale	17
2.4	Pseudo-Residuen bei der Modellierung diskreter Merkmale	25
2.5	Einführende Beispiele	28
2.5.1	Überprüfung einer gemeinsamen Verteilung für den Brusthöhendurchmesser und die Höhe von Bäumen	29
2.5.2	Überprüfung einer Poissonverteilung für den radioak- tiven Zerfall von Polonium	33
2.6	Abschließende Bemerkungen zum Kapitel	37
3	Überprüfung verallgemeinerter linearer Modelle	39
3.1	Einführung	39
3.2	Grundform des verallgemeinerten linearen Modells	41
3.3	Graphische Analyse der Residuen	45
3.3.1	Analyse von QQ-Plots mit Pseudo-Residuen	46
3.3.2	Graphische Residuenanalyse bei diskreten Zielvaria- blen	55

3.3.2.1	QQ-Plots für NPR-Intervalle	56
3.3.2.2	Histogramme für UPR-Intervalle	58
3.3.2.3	Dreieck-Plots für UPR-Intervalle	66
3.4	Berücksichtigung der Parameterschätzung	75
3.5	Überprüfung von Modellen für spezielle Problemstellungen	88
3.5.1	Wiederholte Messung binärer Variablen	89
3.5.2	Matrizen für Markenwechsel	101
3.5.3	Modellierung zyklischer Daten	114
3.6	Abschließende Bemerkungen zum Kapitel	119
4	Pseudo-Residuen bei der Modellierung von Aktienrenditen	123
4.1	Einführung	123
4.2	Ausgewählte Zeitreihenmodelle zur Modellierung von Aktienrenditen	124
4.2.1	GARCH-Modelle	124
4.2.2	Hidden Markov-Modelle	126
4.2.3	Modelle mit gleitenden Dichten	129
4.3	Daten und Modellanpassung	130
4.4	Überprüfung der Voraussageverteilungen	132
4.4.1	Grundlegende Gestalt der Voraussageverteilungen	132
4.4.2	Voraussage des VaR	133
4.4.3	Abbildung extremer Beobachtungen	134
4.4.4	Erfassung zeitlicher Abhängigkeiten	136
4.4.5	Voraussage der Volatilität im Zeitverlauf	138
4.4.6	Perioden unter- bzw. überschätzter Volatilität	141
4.4.6.1	Algorithmus zur Identifikation von Änderungszeitpunkten der Varianz	141
4.4.6.2	Simulationsstudie zum Vergleich der Algorithmen	145
4.4.6.3	Anwendung des Algorithmus auf die Zeitreihe der NPR	150

4.5	Kalibrierung der Voraussageverteilungen	154
4.5.1	Mängel der Modellanpassung und Grundgedanke der Kalibrierung	154
4.5.2	Vorgehensweise zur Kalibrierung der Voraussagen	155
4.5.3	Anwendung der Kalibrierung	157
4.6	Abschließende Bemerkungen zum Kapitel	168
5	Überprüfung multivariater Verteilungen	169
5.1	Einführung	169
5.2	Definition des stochastischen Modells für multivariate Verteilungen	170
5.3	Multivariate Pseudo-Residuen	171
5.4	Überprüfung multivariater Verteilungen	175
5.4.1	Grundlegende Gestalt der gemeinsamen Verteilung	175
5.4.2	Erfassung von extremen Beobachtungen	182
5.4.2.1	Identifikation extremer Beobachtungen in multivariaten Datensätzen	182
5.4.2.2	Distanzbasierte Analyse der mNPR	184
5.4.2.3	Winkelbasierte Analyse der multivariaten Pseudo-Residuen	191
5.4.3	Erfassung der Abhängigkeiten zwischen den Variablen	205
5.4.3.1	Abbildung bedingter Erwartungswerte und Varianzen	205
5.4.3.2	χ -Plots zur Identifikation nicht spezifizierter Abhängigkeiten	210
5.5	Abschließende Bemerkungen zum Kapitel	217
6	Zusammenfassung und Ausblick	219
A	Programme	225
	uhistbounds	226
	psglm	227

psqq	229
bands	230
qqnorm.dis	232
set.trisys	233
tri.bands	234
tri.xy	235
hist.dis1	236
hist.dis2	237
tableplot	238
rowrpr	239
colrpr	240
cusumsq	241
varbreaks	243
varperiods	245
hist3d	247
reshist	249
multnormres	251
resangles	253
chiplot	255
Literaturverzeichnis	257

Abbildungsverzeichnis

1.1	Aufbau der Arbeit.	4
2.1	Grundlegende Ideen bei der Definition von Residuen.	8
2.2	Dichtefunktion einer Gammaverteilung mit den Parametern $\nu = \lambda = 2$.	13
2.3	Erwarteter Verlauf des QQ-Plots für vier verschiedene Residuen bei Gültigkeit der in Abbildung 2.2 dargestellten Gammaverteilung.	14
2.4	Pseudo-Residuen (UPR und NPR) einer stetigen Beobachtung.	18
2.5	Dichtefunktion der NPR für unterschiedliche Werte der erklärenden Variablen.	23
2.6	Pseudo-Residuen (UPR-Intervall und NPR-Intervall) einer diskreten Beobachtung.	26
2.7	Beobachtungen sowie Höhenlinien der angepassten gemeinsamen Dichtefunktion.	30
2.8	Histogramme für den Brusthöhendurchmesser und die Höhe der Bäume sowie angepasste Randverteilungen.	31
2.9	Histogramme der UPR für die Randverteilungen mit simultanen Konfidenzintervallen.	32
2.10	QQ-Plots der NPR für die Randverteilungen.	33
2.11	Relative Häufigkeiten für die Anzahl emittierter α -Partikel und Wahrscheinlichkeitsfunktion der angepassten Poissonverteilung.	34
2.12	Indexplot der NPR-Intervalle für die beobachteten Häufigkeiten.	36

3.1	Indexplot der NPR-Intervalle.	44
3.2	QQ-Plot der NPR.	49
3.3	NPR des QQ-Plots.	50
3.4	Angepasste Regressionsmodelle für den Zusammenhang zwischen n und p^- .	53
3.5	QQ-Plot der NPR mit simultanen Konfidenzbändern.	55
3.6	QQ-Plot der NPR-Intervalle.	57
3.7	Erster Schritt bei der Konstruktion von Histogrammen für UPR-Intervalle: Die Umhüllende.	59
3.8	Zweiter Schritt bei der Konstruktion von Histogrammen für UPR-Intervalle: Gleiche Klassenbreite.	61
3.9	Plot der Anzahl von Arztbesuchen in 1997 gegen das Alter (links) und gegen die Anzahl der Besuche in 1995 (rechts).	62
3.10	Umhüllende für die betrachteten Modelle sowie Umhüllende bei Gültigkeit des jeweiligen Modells.	63
3.11	Histogramme der UPR-Intervalle für die betrachteten Modelle bei poissonverteilter Zielvariable.	64
3.12	Histogramme der UPR-Intervalle für die betrachteten Modelle bei negativ binomialverteilter Zielvariable.	65
3.13	Darstellung des Punktes $(0.1, 0.2, 0.7)$ im dreieckigen Koordinatensystem.	67
3.14	Dreieck-Plot der UPR-Intervalle (links) und QQ-Plot der NPR-Intervalle (rechts).	70
3.15	Wahrscheinlichkeitsfunktionen für Y_5 und Y_{11} unter dem angepassten Modell.	71
3.16	Dreieck-Plots der UPR-Intervalle für die Modelle 1 und 3 (links) sowie für Modell 2 (rechts).	73
3.17	Dreieck-Plot mit Referenzverteilung von UPR-Intervallen.	75
3.18	Dichtefunktion der geschätzten UPR bei Verwendung des internen (links) und externen (rechts) Varianzschätzers für $x = 1$ (oben) und $x = 5$ (unten).	80

3.19	Dekomposition der Verteilungsfunktion der geschätzten UPR bei Verwendung des internen (oben) und externen (unten) Varianzschätzers.	81
3.20	<i>Bootstrap</i> -Schätzer der Verteilungsfunktionen der geschätzten UPR. (Abszisse und Ordinate sind jeweils durch null und eins begrenzt.)	83
3.21	Plot der ursprünglichen NPR gegen die kalibrierten NPR (links) und QQ-Plot der kalibrierten NPR (rechts).	84
3.22	<i>Bootstrap</i> -Schätzer der Verteilungen der UPR-Intervallmittelpunkte.	86
3.23	Scatterplot für die Beobachtungen der beiden erklärenden Variablen.	87
3.24	Plot der Mittelpunkte der UPR-Intervalle der <i>Bootstrap</i> -Verteilung gegen die entsprechenden Intervallgrenzen.	88
3.25	Plot der ursprünglichen NPR gegen die kalibrierten NPR (links) und QQ-Plot der kalibrierten NPR (rechts).	89
3.26	Relative Häufigkeiten der Stromstöße in den einzelnen Versuchen und angepasstes Modell 1.	90
3.27	Verteilung der Summe der Stromstöße eines einzelnen Hundes unter Modell 1.	92
3.28	QQ-Plot der NPR-Intervalle für die Summe der Stromstöße der einzelnen Hunde unter Modell 1.	93
3.29	Individuelle Lernkurven der Hunde nach Modell 2.	94
3.30	Verteilung der P-Werte für die Hypothese, dass die Hunde nicht lernen.	95
3.31	Lernkurven nach den Modellen 3 (links) und 4 (rechts).	97
3.32	QQ-Plot der NPR-Intervalle für die Summe der Stromstöße der einzelnen Hunde unter Modell 3 bzw. Modell 4.	98
3.33	Lernkurven des Modells 2 mit Markierung der Ausreißer.	99
3.34	Durchschnittliche Residuen für die einzelnen Versuche und ihre Standardabweichung unter Modell 3 bzw. Modell 4.	100
3.35	NPR-Intervalle der einzelnen Versuche unter Modell 3 bzw. Modell 4.	101

3.36	QQ-Plot der NPR-Intervalle für die Zelhäufigkeiten aus dem gesättigten Modell.	104
3.37	Indexplot der NPR-Intervalle aus dem Modell für Unabhängigkeit.	106
3.38	QQ-Plots für Unabhängigkeit und Quasi-Unabhängigkeit.	107
3.39	Tableplots für Unabhängigkeit und Quasi-Unabhängigkeit.	108
3.40	QQ-Plots für Symmetrie und Quasi-Symmetrie.	112
3.41	Tableplots für Symmetrie und Quasi-Symmetrie.	113
3.42	Anteil von Niederschlagstagen für jeden Tag des Jahres während der Jahre von 1950 bis 1994. (Die Bedeutung der grauen Kurve wird im Text erläutert.)	114
3.43	Geschätzte Niederschlagswahrscheinlichkeiten für die Jahre 1971 (links) und 1989 (rechts).	116
3.44	QQ-Plot und Indexplot der NPR-Intervalle für die Summe von Niederschlagstagen an den einzelnen Tagen des Jahres.	118
3.45	QQ-Plot und Indexplot der NPR-Intervalle für die Summe von Niederschlagstagen in den einzelnen Jahren.	119
4.1	Zeitreihe für die kontinuierlichen Renditen der Aktie der Deutschen Bank.	131
4.2	Histogramme der UPR der drei Modelle.	133
4.3	Vorgegebenes und tatsächliches Niveau bei Schätzung des <i>VaR</i> .	134
4.4	QQ-Plots der NPR mit simultanen Konfidenzbändern.	135
4.5	NPR der QQ-Plots.	136
4.6	Autokorrelationsfunktionen der NPR (links) und der quadrierten NPR (rechts).	137
4.7	Zentrierte CUSUMSQ der NPR.	139
4.8	Folge betragsmäßig größerer NPR bei Prognose geringerer Volatilität (oben) und Folge betragsmäßig kleinerer NPR bei Prognose größerer Volatilität (unten).	140
4.9	Erste Stufe des Algorithmus zur Identifikation von Änderungszeitpunkten der Varianz.	143

4.10	Zweite Stufe des Algorithmus zur Identifikation von Änderungszeitpunkten der Varianz.	144
4.11	Geschätzte Abweichung bei der Reproduktion des Varianzverlaufs ohne Anwendung der zweiten Stufe (links) und mit Anwendung der zweiten Stufe (rechts).	147
4.12	Geschätzte Anzahl von Änderungszeitpunkten für Zeitreihen mit drei tatsächlichen Änderungszeitpunkten ohne Anwendung der zweiten Stufe (links) und mit Anwendung der zweiten Stufe (rechts).	149
4.13	Mögliche Änderungszeitpunkte der Residualvarianz für das Hidden Markov-Modell.	151
4.14	Endgültige Änderungszeitpunkte der Residualvarianz für das Hidden Markov-Modell.	152
4.15	Zeitreihe und geschätzte Residualvarianz.	153
4.16	Schätzung der Transformation für $t = 2501$ und Transformation des entsprechenden UPR.	159
4.17	Korrelation der Residuen vor und nach der Transformation.	160
4.18	Empirische Verteilungsfunktion der transformierten Residuen für $t = 2501$.	161
4.19	Verteilungsfunktionen der Voraussage für $t = 2501$ vor Kalibrierung (links) und nach Kalibrierung (rechts).	162
4.20	Histogramme der UPR vor Kalibrierung (oben) und nach Kalibrierung (unten).	163
4.21	NPR der QQ-Plots vor Kalibrierung (oben) und nach Kalibrierung (unten).	164
4.22	Autokorrelation der UPR vor Kalibrierung (links) und nach Kalibrierung (rechts).	165
4.23	Autokorrelation der zentrierten und quadrierten UPR vor Kalibrierung (links) und nach Kalibrierung (rechts).	166
4.24	Zentrierte CUSUMSQ der NPR vor Kalibrierung (oben) und nach Kalibrierung (unten).	167
5.1	Beobachtungen und Höhenlinien der Dichtefunktion der S_{BB} -Verteilung.	177

5.2	Dreidimensionale Histogramme der bivariaten UPR.	179
5.3	Durch Graustufen dargestellte Mittelpunkte der NPR–Intervalle der dreidimensionalen Histogramme.	180
5.4	Übertragung der Graustufen auf die ursprünglichen Beobachtungen.	181
5.5	Mischung von zwei bivariaten Normalverteilungen und extreme Beobachtung.	183
5.6	QQ–Plots der quadrierten Mahalanobisdistanzen (links) und der NPR der Distanzen (rechts).	187
5.7	QQ–Plot der quadrierten Mahalanobisdistanzen mit simultanem 95%–Konfidenzband.	188
5.8	QQ–Plots der NPR der Distanzen bei alternativer Definition der bivariaten NPR.	189
5.9	Extreme Beobachtungen bei Verwendung der NPR der Distanzen.	190
5.10	Extreme Beobachtungen bei Verwendung der Mahalanobisdistanz.	191
5.11	Gruppe konzentrierter bivariater NPR.	192
5.12	Scatterplot standardisierter bivariater NPR (links) und Projektionen auf den Einheitskreis (rechts).	194
5.13	Verteilungsfunktion der Winkel für unterschiedliche Werte von p .	195
5.14	Scatterplot bivariater NPR mit einer Konzentration nördlich des Ursprungs (oben links), Projektionen der standardisierten Residuen (oben rechts) und Indexplots der NPR der Winkel (unten).	196
5.15	Äquidistante Projektionen und Konzentration von Projektionen (oben links) sowie Verlauf der Funktionen f (oben rechts), g (unten links) und h (unten rechts) in Abhängigkeit von der Referenzrichtung.	199
5.16	Identifikation der konzentrierten mNPR nach winkelbasiertem Vorgehen (oben) und mit Hilfe der Mahalanobisdistanzen (unten).	204

5.17	Scatterplot der Beobachtungen für U_1 gegen die von U_2 sowie nicht-parametrische Schätzung von $E(U_2 U_1)$.	208
5.18	Analyse des dreidimensionalen Histogramms zur Überprüfung der Schätzung von $E(Y_H Y_D)$.	209
5.19	Scatterplot der Beobachtungen für U_1 gegen die von T sowie nicht-parametrische Schätzung von $E(T U_1)$.	210
5.20	Scatterplot für 10 Beobachtungen einer bivariaten $U(0,1)$ -Verteilung und zugehöriger χ -Plot.	211
5.21	Scatterplot für jeweils 500 bivariate Beobachtungen mit verschiedenen Formen stochastischer Abhängigkeiten und zugehörige χ -Plots.	214
5.22	χ -Plots für die Elemente von $\mathbf{U}^{(1)}$ und $\mathbf{U}^{(2)}$ bei Verwendung der S_{BB} -Verteilung.	215
5.23	χ -Plots für die Elemente von $\mathbf{U}^{(1)}$ und $\mathbf{U}^{(2)}$ bei Verwendung einer Mischung von zwei bivariaten Normalverteilungen.	216

Tabellenverzeichnis

2.1	ML-Schätzer für die Mischverteilung der bivariaten Normalverteilungen.	29
2.2	Häufigkeiten für die Anzahl emittierter Partikel in Zeitintervallen mit konstanter Länge.	34
2.3	Häufigkeiten für die Anzahl emittierter Partikel und Beiträge zur χ^2 -Prüfgröße.	35
3.1	Anzahl der Maschinenausfälle, Betriebszeiten in zwei verschiedenen Betriebsarten sowie Schätzer der Erwartungswerte.	44
3.2	Beobachtungen und Schätzer für den Widerstand des Halbleiters sowie Ausprägungen der Faktoren.	48
3.3	Parameterschätzer des verallgemeinerten linearen Modells.	48
3.4	Schätzer für p^- in Abhängigkeit von Stichprobenumfang und Signifikanzniveau.	52
3.5	Schätzer der Regressionsmodelle für den Zusammenhang zwischen n und p^- .	53
3.6	Beobachtungen für die Zeit, die Anzahl benötigter Getränkekisten und den zurückzulegenden Fußweg.	54
3.7	Parameterschätzer und AIC der vier Modelle zur Beschreibung der Anzahl von Arztbesuchen.	62
3.8	Anzahl von Personen mit Anzeichen von Senilität aus n_i Personen mit Punktzahl x_i im Intelligenztest.	69
3.9	Beobachtete Anzahl entkeimter Staubbeutel aus n_{ij} Staubbeuteln unter verschiedenen Bedingungen.	72
3.10	Parameterschätzer der drei Modelle zur Beschreibung der Anzahl entkeimter Staubbeutel.	72

3.11	Beobachtungen für das Hundeexperiment mit 25 Versuchen (Spalten) für jeden der 30 Hunde (Zeilen).	91
3.12	Beobachtete Markenwechsel auf dem Automarkt und geschätzte Zellhäufigkeiten unter Unabhängigkeit und Quasi-Unabhängigkeit.	102
3.13	Beobachtete Markenwechsel auf dem Automarkt und geschätzte Zellhäufigkeiten unter Symmetrie und Quasi-Symmetrie.	110
3.14	Anzahl überlagernder Sinus- und Kosinusfrequenzen und AIC der entsprechenden Modelle.	115
3.15	Parameterschätzer des Modells für $K = 4$.	115
4.1	Zustände und Übergangswahrscheinlichkeiten eines Hidden Markov-Modells.	127
4.2	ML-Schätzer der drei Modelle für die Renditezeitreihe.	131
4.3	Faktoren und Faktorstufen des Simulationsexperiments.	146
4.4	Mögliche Änderungszeitpunkte der Residualvarianz für das Hidden Markov-Modell.	153
4.5	Geschätzte Varianz der NPR für die identifizierten Zeitabschnitte.	154
4.6	Geschätzte Varianz der Residuen des Modells mit gleitenden Dichten nach Kalibrierung.	167
5.1	ML-Schätzer der S_{BB} -Verteilung.	178
5.2	Vier verschiedene Messungen bezüglich der Festigkeit von 30 Brettern und quadrierte Mahalanobisdistanzen.	186
5.3	Künstlich erzeugte und in den Datensatz eingefügte Beobachtungen.	203
5.4	Relative Häufigkeiten der dichotomisierten Variablen.	212

Kapitel 1

Einleitung

1.1 Problemstellung und Ziele

Die Analyse von Residuen stellt ein zentrales Instrument zur Überprüfung stochastischer Modelle dar. Durch eine gezielte Untersuchung der Residuen erhält man verschiedene Informationen über die Güte eines verwendeten Modells, beispielsweise, ob ein Modell in der Lage ist, extreme Beobachtungen abzubilden oder ob die Annahme bestimmter Verteilungen gerechtfertigt ist.

Die Überprüfung stochastischer Modelle besitzt sowohl in der statistischen Methodologie als auch bei anwendungsorientierten Problemstellungen eine große Bedeutung. Aus Anwendersicht spielt die Frage nach der Modellgüte vielleicht die entscheidende Rolle. Beispielsweise ist die Schätzung des *Value at Risk* von Finanztiteln ein derzeit häufig diskutiertes Problem aus dem Bereich des Risikomanagements. Hier möchte der Nutzer statistischer Methoden wissen, wie gut ein zur Verfügung stehendes Modell die Risiken eines Finanztitels vorher sagt. Fragen bezüglich der Parameterschätzung oder der Modellauswahl treten für ihn in den Hintergrund, wenn das Modell die gewünschten Eigenschaften erfüllt.

Aus statistisch methodologischer Sicht können diese Fragen jedoch nicht getrennt betrachtet werden. Dies spiegelt sich auch in der Literatur wider, in der Verfahren zur Modellüberprüfung häufig in Abhängigkeit vom betrachteten Modell vorgeschlagen werden. Typische Publikationstitel sind beispielsweise „*Regression Diagnostics*“ (BELSLEY, KUH AND WELSCH, 1980), „*Residuals in Generalized Linear Models*“ (PIERCE AND SCHAFER, 1996) oder (etwas spezieller) „*Residuals for Proportional Hazard Models with Interval-Censored*

Survival Data“ (FARRINGTON, 2000). In diesen und ähnlichen Veröffentlichungen werden Residuen definiert bzw. vergleichend untersucht, die für die jeweilige Modellsituation geeignet sind, deren Verwendung zur Überprüfung anderer Modelle jedoch nicht sinnvoll erscheint. Eine allgemeinere Definition von Residuen, die viel Beachtung gefunden hat, stammt von COX AND SNELL (1968). Sie definieren Residuen für ein Modell, das einige klassische Modellsituationen als Spezialfall enthält.

Mit Blick auf die Überprüfung stochastischer Modelle ist es wünschenswert, Residuen zu definieren, die für eine Vielzahl verschiedener Modellsituationen verwendet werden können und unabhängig vom betrachteten Modell bekannte Verteilungseigenschaften besitzen. Darüber hinaus ist es für die Analyse der Residuen von Vorteil, wenn diese bei Gültigkeit des zu überprüfenden Modells unabhängig und identisch standardnormalverteilt oder $U(0, 1)$ -verteilt sind, da diese Verteilungen (wie im Verlauf der Arbeit deutlich wird) gute Voraussetzungen für eine Analyse der Residuen schaffen. Ein Residuenkonzept, das diese Anforderungen annähernd erfüllt, ist das Konzept der Pseudo-Residuen. (Eine Beschreibung weiterer Konzepte erfolgt zu Beginn des nachfolgenden Kapitels.) Auch diese Residuen sind vereinzelt in der Literatur zu finden. Sie wurden zum Beispiel zur Überprüfung verallgemeinerter linearer Modelle oder für die Beurteilung von Zeitreihenmodellen verwendet.

Übergeordnetes Ziel dieser Arbeit ist es, die breiten Anwendungsmöglichkeiten der Pseudo-Residuen bei der Überprüfung stochastischer Modelle aufzuzeigen. Es soll für ausgewählte Modellsituationen (beispielsweise für verallgemeinerte lineare Modelle) anhand von Beispielen die Bestimmung und Analyse der Pseudo-Residuen dargestellt und erläutert werden. Das Ziel schließt auch die Betrachtung von Pseudo-Residuen bei der Modellierung diskreter Merkmale ein. In solchen Fällen ist eine geeignete Definition von Residuen oftmals problematisch, insbesondere dann, wenn sich die im Modell verwendeten Verteilungen nur ungenau durch stetige Verteilungen approximieren lassen.

Bei der Analyse von Residuen besitzen graphische Methoden eine große Bedeutung. Beispielsweise QQ-Plots (in der englischen Sprache unter der Begriff *‘quantile-quantile plot’* bekannt) oder Plots der Residuen gegen einen Zeitindex können als Standardanalysen bezeichnet werden. Derartige Standardanalysen basieren auf der Idee von Punktresiduen. Wie später erläutert wird, sind Pseudo-Residuen, die aus der Modellierung diskreter Merkmale resultieren, jedoch als Intervalle definiert. Zu ihrer Analyse sind die Standardmethoden daher nicht immer geeignet bzw. nur in einer verallgemeinerten Form anwendbar. Dieser Schritt, d.h. die Erweiterung von Standardmethoden zur Analyse

von Punktresiduen bzw. die Entwicklung von graphischen Methoden zur Analyse von Intervallresiduen, stellt ein weiteres Ziel dieser Arbeit dar.

Ferner werden die in dieser Arbeit erläuterten Methoden in **R** (vgl. IHAKA AND GENTLEMAN, 1996) implementiert. Im Anhang wird eine kurze Beschreibung der Funktionen und ein jeweils kleines Beispiel zur Anwendung der Funktionen gegeben. Alle beschriebenen Funktionen sind in der **R**-*library* `psresiduals` zusammengefasst, die als Anlage zu dieser Arbeit zu verstehen ist.

1.2 Aufbau der Arbeit

Der Aufbau der Arbeit ist in Abbildung 1.1 charakterisiert.

Das nachfolgende Kapitel ‚Theoretische Grundlagen‘ umfasst zunächst die Darstellung grundlegender Ideen bei der Definition von Residuen sowie eine komprimierte Übersicht zu häufig bzw. traditionell verwendeten Residuen. Danach wird ein stochastisches Modell definiert, auf dessen Grundlage sich die Definition der Pseudo-Residuen in allgemeiner Form anschließt. Diese erfolgt getrennt für Modelle zur Modellierung stetiger Merkmale einerseits sowie für Modelle zur Modellierung diskreter Merkmale andererseits. Jeweils werden einige Eigenschaften der Pseudo-Residuen abgeleitet und erläutert.

Die drei anschließenden Kapitel befinden sich aus Gliederungssicht auf einer Ebene und beinhalten jeweils Problemstellungen der Modellüberprüfung in ausgewählten Modellsituationen.

In Kapitel 3 wird die Überprüfung von verallgemeinerten linearen Modellen betrachtet. Im Anschluss an eine kurze Einführung sowie der allgemeinen Formulierung des verallgemeinerten linearen Modells werden bestimmte Problemstellungen bei der Überprüfung von Modellen aus dieser Modellklasse behandelt. Dazu zählt als erstes die Erweiterung bzw. Entwicklung von graphischen Methoden zur Analyse der Residuen. Im Abschnitt ‚Graphische Analyse von Pseudo-Residuen‘ wird gezeigt, wie mit Hilfe von Pseudo-Residuen die Gestalt eines QQ-Plots analysiert werden kann und welche Möglichkeiten bei der graphischen Analyse von Intervallresiduen gewählt werden können. Im darauf folgenden Abschnitt ‚Berücksichtigung der Parameterschätzung‘ wird die Auswirkung der Parameterschätzung auf die Verteilung der Residuen betrachtet. Während die Verteilungseigenschaften bei Kenntnis der Modellparameter bekannt sind, ist dies bei einer Schätzung der Parameter nicht der Fall und es werden hier Möglichkeiten zur Korrektur bezüglich des Einflusses der Parame-

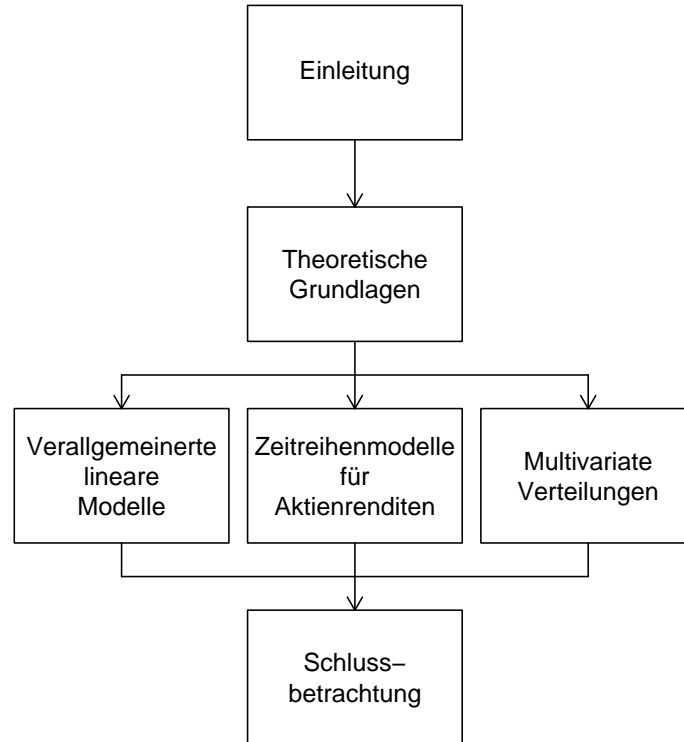


Abbildung 1.1: Aufbau der Arbeit.

terschätzung dargestellt. Die beiden Teilabschnitte ‚Graphische Analyse von Pseudo-Residuen‘ und ‚Berücksichtigung der Parameterschätzung‘ besitzen allgemeinen Charakter, d.h. die dargestellten Methoden beschränken sich nicht auf den Bereich der verallgemeinerten linearen Modelle. Das Kapitel schließt mit drei umfangreicheren Beispielen, in denen verallgemeinerte lineare Modelle zur Modellierung speziellerer Problemstellungen verwendet werden. Jeweils werden die besonderen Anforderungen an die Modellüberprüfung erläutert und geeignete Methoden zur Analyse der Pseudo-Residuen dargestellt.

Das Kapitel 4 ist der Überprüfung von Modellen für Aktienrenditen gewidmet. Überprüft wird die Güte der Voraussagen von drei ausgewählten Zeitreihenmodellen bei der Prognose der Renditen für die Aktie der Deutschen Bank. Dabei werden verschiedene Anforderungen an die Voraussagen (beispielsweise die korrekte Voraussage der Volatilität) unterschieden und entsprechende Methoden dargestellt, mit denen überprüft werden kann, ob die Modelle die-

se Anforderungen erfüllen. In einem weiteren Abschnitt wird gezeigt, wie die Eigenschaften der aus der Modellanpassung berechneten Residuen genutzt werden können, um die Güte der Voraussagen zu verbessern. Die Ausführungen gehen also in diesem Teil der Arbeit über den Bereich der Modellüberprüfung hinaus.

Das Kapitel ‚Überprüfung multivariater Verteilungen‘ beschreibt Methoden zur Überprüfung von Modellen zur gemeinsamen Beschreibung mehrerer Merkmale. Hier ist es in einem ersten Schritt erforderlich, die Definition des allgemeinen stochastischen Modells sowie die Definition der Pseudo-Residuen mit Blick auf die multivariate Problemstellung zu erweitern. Anschließend werden Methoden zur Überprüfung verschiedener Aspekte der Anpassungsgüte der multivariaten Verteilungen mit Hilfe von Beispielen illustriert.

Ein grundlegendes Merkmal dieser Arbeit ist die Veranschaulichung der Methoden und theoretischen Aspekte anhand von Beispielen. Um zu vermeiden, dass dadurch allgemeine Gesichtspunkte der betrachteten Problemstellungen nicht ausreichend zum Ausdruck kommen, werden am Ende eines jeden Kapitels wesentlich erscheinende Aspekte kurz zusammengefasst.

Das Kapitel 6 fasst die Ergebnisse der Arbeit zusammen und gibt einen Ausblick auf weitere mögliche Arbeiten zu Pseudo-Residuen.

Kapitel 2

Theoretische Grundlagen

2.1 Grundlegende Möglichkeiten bei der Definition von Residuen

Als Einstieg in das Thema werden in diesem Abschnitt drei grundsätzliche Ideen bei der Definition von Residuen dargestellt. Dabei wird das Konzept der Pseudo-Residuen gegenüber anderen Konzepten abgegrenzt und es wird ein kurzer Überblick zu Residuen gegeben, die sehr häufig in der Literatur vorgeschlagen werden. Dieser Überblick erhebt nicht den Anspruch auf Vollständigkeit. Vielmehr umfasst er einige ausgewählte Residuen, die zur Veranschaulichung der grundlegenden Definitionsideen verwendet werden.

Eine allgemeine Definition von Residuen, die alle hier erläuterten Konzepte beinhaltet, stammt von NARDI AND SCHEMPER (1999):

Residuen sind Statistiken, die die Eigenschaft besitzen, im Betrag größer zu werden, je weiter eine Beobachtung von einem unter dem Modell erwarteten Wert abweicht.

Auf der Grundlage dieser Definition lassen sich drei grundlegende Ideen zur Definition von Statistiken mit der genannten Eigenschaft unterscheiden. Abbildung 2.1 dient zur graphischen Veranschaulichung dieser Ideen. Sie zeigt eine Dichtefunktion (durchgezogene Kurve), die in diesem Fall das zu überprüfende Modell beschreibt. (Beispielhaft wurde hier die Dichtefunktion einer Normalverteilung gewählt.) Ferner ist eine Beobachtung y auf der Abszisse eingezeichnet, für die ein Residual zu definieren ist.

Die einfachste Möglichkeit zur Definition des Residuals (in der Abbildung mit

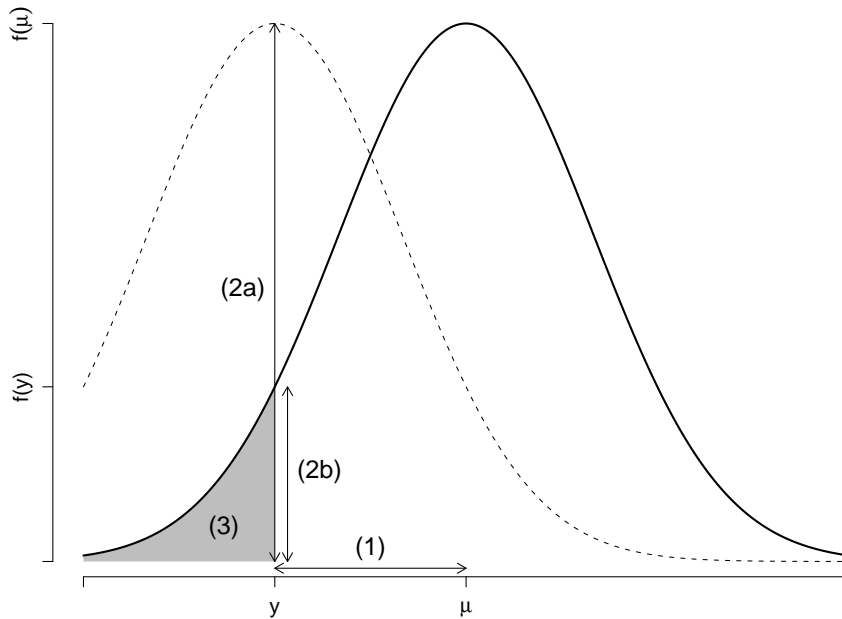


Abbildung 2.1: Grundlegende Ideen bei der Definition von Residuen.

(1) gekennzeichnet) basiert auf einem Vergleich von der Beobachtung und dem Erwartungswert der Beobachtung unter dem Modell. Quantifiziert man den Vergleich durch die Differenz zwischen der Beobachtung und dem Erwartungswert

$$r^{\text{ord}} = y - \mu,$$

erhält man die sogenannten *ordinary residuals*, deren Verwendung in linearen Regressionsmodellen eine lange Tradition besitzt. Sie stellen die Basis zahlreicher Transformationen mit unterschiedlichen Zielen dar. Analysiert man beispielsweise mehrere Residuen simultan und wurden die Residuen aus unterschiedlichen Verteilungen berechnet, ist es für die Vergleichbarkeit der Residuen notwendig, dass alle Residuen die gleiche Varianz besitzen. Ein mögliches Ziel der Transformation ist demnach die Standardisierung der Residuen bezüglich ihrer Varianz, die erreicht wird, indem man die *ordinary residuals* durch ihre Standardabweichung dividiert:

$$r^{\text{P}} = \frac{r^{\text{ord}}}{\sqrt{\text{Var}(r^{\text{ord}})}}.$$

Die Varianz dieser sogenannten *Pearson residuals* ist eins. Da die Varianz in der Regel geschätzt werden muss, ergeben sich bei der Präzisierung der Definition je nach verwendetem Varianzschätzer unterschiedliche Möglichkeiten.

Einige dieser Möglichkeiten werden im Zusammenhang mit der Überprüfung klassischer linearer Regressionsmodelle bei CHATTERJEE AND HADI (1988, S. 72–80) diskutiert. Sie geben Gründe sowie Literaturverweise an, die für die Verwendung der sogenannten *externally Studentized residuals* sprechen. (Eine genauere Betrachtung dieser Residuen erfolgt in Abschnitt 3.4.) Einen umfassenden Überblick zu Residuen für klassische lineare Modelle geben COOK AND WEISBERG (1982). Für verallgemeinerte lineare Modelle ist

$$r^P = \frac{y - \mu}{\sqrt{\phi V(\mu)}}$$

die varianzstabilisierende Transformation der *ordinary residuals*. Hier wird die Varianz einer Beobachtung durch das Produkt der Varianzfunktion $V(\mu)$ und dem sogenannten Dispersionsparameter ϕ bestimmt (vgl. beispielsweise MCCULLAGH AND NELDER, 1989, S. 37). In verallgemeinerten linearen Modellen ist die unterstellte Verteilung der zu erklärenden Variablen jedoch häufig schief, was dazu führt, dass neben einer Standardisierung bezüglich der Varianz auch die Beseitigung der Schiefe in der Verteilung der Residuen notwendig ist, wenn mehrere Residuen simultan untersucht werden sollen. Diese Korrektur wird durch die *Anscombe residuals*

$$r^A = \frac{t(y) - E(t(Y))}{\sqrt{\text{Var}(t(Y))}}$$

berücksichtigt (vgl. beispielsweise PIERCE AND SCHAFER, 1986). Hier wird das Residual nicht für die Beobachtung selbst, sondern für eine Transformation der Beobachtung definiert. Wählt man

$$t(x) = \int \frac{1}{V^{1/3}(x)} dx,$$

sind die transformierten Residuen approximativ normalverteilt (vgl. MCCULLAGH AND NELDER, 1989, S. 38). Die Transformation ist bei BARNDORFF-NIELSEN (1978, S. 177–182) für einige Verteilungen aus der Exponentialfamilie explizit spezifiziert; hier werden die *Anscombe residuals* im Rahmen des nächsten Beispiels für eine Gammaverteilung betrachtet.

Die bisher dargestellten Residuen basieren auf der Differenz zwischen der Beobachtung und ihrem Erwartungswert unter dem Modell. Eine weitere grundlegende Möglichkeit, das Residual einer Beobachtung zu definieren, ergibt sich aus dem Vergleich der Wahrscheinlichkeit der Beobachtung (bzw. des Funktionswertes der Dichtefunktion an der Stelle der Beobachtung) unter dem zu

überprüfenden Modell mit der Wahrscheinlichkeit der Beobachtung unter dem gesättigten Modell. Diese zweite Idee bei der Definition von Residuen ist in Abbildung 2.1 mit (2a) und (2b) gekennzeichnet. Dabei charakterisiert die gestrichelte Kurve (eine Normalverteilung, deren Erwartungswert der Beobachtung entspricht,) das gesättigte Modell. Je größer das (transformierte) Verhältnis der Wahrscheinlichkeiten (bzw. der *Likelihoods*)

$$d = 2 \log \frac{L_{\text{ges}}(y)}{L_{\text{üb}}(y)}$$

betragsmäßig ist, desto größer ist auch das Residual. ($L_{\text{ges}}(y)$ bzw. $L_{\text{üb}}(y)$ beschreiben die *Likelihood* der Beobachtung im gesättigten bzw. im zu überprüfenden Modell.) Umgesetzt ist diese Idee bei PREGIBON (1981), der die *deviance residuals*

$$r^d = \text{sign}(y - \mu) \sqrt{d}$$

für die Überprüfung logistischer Regressionsmodelle vorschlägt. PIERCE AND SCHAFER (1986) untersuchen die Eigenschaften verschiedener Residuen für verallgemeinerte lineare Modelle. Unter anderem betrachten sie die *adjusted deviance residuals*

$$r^{\text{da}} = r^d + \frac{1}{6} \cdot E \left(\frac{(y - \mu)^3}{\text{Var}^{3/2}(Y)} \right),$$

die von ihnen als „*surprisingly good in terms of normality*“ (auch für Verteilungen, die nicht der Normalverteilung ähneln,) beurteilt werden, und bemerken, dass sie auch bei der Überprüfung von Modellen für diskrete Merkmale problemlos eingesetzt werden können, wenn entsprechende Stetigkeitskorrekturen durchgeführt werden (siehe dazu auch DAVISON AND GIGLI, 1989). Bemerkenswert ist, dass die berechneten Werte der *deviance residuals* und der *Anscombe residuals* sehr ähnlich sind (vgl. MCCULLAGH AND NELDER, 1989, S. 39), obwohl ihre Definition, wie hier dargestellt wurde, auf unterschiedlichen Ideen beruht.

Die dritte Alternative (grau markierte Fläche in Abbildung 2.1) bei der Definition von Residuen ist an das Konzept der P-Werte angelehnt. Hier wird beurteilt, wie extrem eine Beobachtung unter dem zu überprüfenden Modell ist. Dazu wird die Wahrscheinlichkeit bestimmt, unter dem Modell eine kleinere Beobachtung als y zu erhalten. Ist diese Wahrscheinlichkeit sehr klein oder sehr groß, erscheint die Beobachtung nicht plausibel. Die direkte Anwendung dieser Idee, d.h. der Verzicht auf weitere Transformationen führt zu einer

ersten Variante von Pseudo-Residuen:

$$u = P\{Y \leq y\} = F_Y(y). \quad (2.1)$$

Streng genommen erfüllt u nicht die Eigenschaften der allgemeinen, zu Beginn des Abschnitts genannten Definition von Residuen, weil u nicht im Betrag größer wird, je weiter die Beobachtung von einem unter dem Modell erwarteten Wert abweicht. (Es ist allerdings offensichtlich, dass die einfache Transformation $u^* = u - 0.5$ die allgemeine Definition erfüllt.) Dennoch erweist sich die nach Gleichung 2.1 definierte Variante der Pseudo-Residuen bei der Überprüfung bestimmter Aspekte der Modellanpassung als sehr hilfreich, weil $U = F_Y(Y)$ bei Gültigkeit des zu überprüfenden Modells rechteckverteilt ist über dem Intervall $[0, 1]$.

Die Transformation

$$r = \Phi^{-1}(u) \quad (2.2)$$

beschreibt die zweite Variante der Pseudo-Residuen. Diese Variante erfüllt die allgemeine Definition der Residuen, da r im Betrag größer wird, je weiter eine Beobachtung vom Median der zu überprüfenden Verteilung entfernt liegt. Ferner ist $R = \Phi^{-1}(U)$ bei Gültigkeit des Modells standardnormalverteilt, was die Identifikation von extremen Beobachtungen auf einer gut geeigneten und häufig verwendeten Skala ermöglicht.

Die Verwendung von Pseudo-Residuen zur Überprüfung von stochastischen Modellen erfolgt beispielsweise bei ZUCCHINI AND MACDONALD (1999). Sie illustrieren die Analyse der Pseudo-Residuen bei der Modellierung stetiger, diskreter und gemischt stetig-diskreter Merkmale anhand verschiedener Beispiele. Sie verwenden für die Residuen aus Gleichung 2.2 den Begriff *pseudo-residuals*, der für diese Arbeit in seiner deutschen Übersetzung übernommen wird. DUNN AND SMYTH (1996) verwenden für diese Variante der Pseudo-Residuen die Bezeichnung *quantile residuals* und nutzen diese zur Überprüfung verallgemeinerter linearer Modelle. Ferner findet man im Zusammenhang mit der Überprüfung von Zeitreihenmodellen für die in Gleichung 2.1 definierten Residuen häufig die Bezeichnung *uniform residuals* (vgl. beispielsweise Diebold Gunther and Tay, 1998).

In Abschnitt 2.3 wird gezeigt, dass beide Varianten der Pseudo-Residuen als Spezialfall der *Cox-Snell residuals* (vgl. COX AND SNELL, 1968) darstellbar sind. Letztere werden häufig bei der Überprüfung von Modellen für Überlebenszeiten verwendet (vgl. beispielsweise KLEIN AND MOESCHBERGER, 1997) und können als Ursprung der Pseudo-Residuen verstanden werden.

Die Darstellung der drei grundlegenden Ideen bei der Definition von Residuen erfolgte anhand *einer* Beobachtung. In praktischen Situationen sind in der Regel allerdings *mehrere* Residuen, die aus unterschiedlichen Verteilungen berechnet wurden, gleichzeitig zu untersuchen. Auch und gerade in solchen Situationen ist es wünschenswert, dass alle Residuen (unter dem Modell) die gleichen Verteilungseigenschaften besitzen, vor deren Hintergrund beispielsweise QQ-Plots oder Histogramme beurteilt werden können. Dies ist, wie bereits erwähnt, für die Pseudo-Residuen der Fall, und zwar unabhängig davon, unter welchen Verteilungen sie berechnet werden. Anhand des nachfolgenden Beispiels wird deutlich, dass die Verteilungsannahmen für einige der übrigen Residuen verletzt ist, wenn als Modell eine (schiefe) Gamma-Verteilung betrachtet wird. Ihre Verwendung erscheint daher in gewisser Weise problematisch, wenn ein Modell mit verschiedenen (möglicherweise schiefen und/oder diskreten) Verteilungen überprüft werden soll.

Beispiel 2.1 Um die bisherigen Ausführungen zu verdeutlichen, werden im Folgenden vier der zuvor erwähnten Residuen für eine Gamma-Verteilung mit der Dichtefunktion

$$f_Y(y) = \begin{cases} \frac{\lambda^\nu y^{\nu-1} e^{-\lambda y}}{\Gamma(\nu)} & y \geq 0 \\ 0 & \text{sonst} \end{cases}$$

explizit bestimmt und vergleichend untersucht. Der Erwartungswert und die Varianz der Gamma-Verteilung sind durch

$$E(Y) = \mu = \nu/\lambda \quad \text{und} \quad \text{Var}(Y) = \nu/\lambda^2$$

gegeben. Man bestimmt demnach für die *Pearson residuals* der Gamma-Verteilung

$$r^P = \frac{y - \mu}{\sqrt{\mu/\lambda}}.$$

Die *Anscombe residuals* berechnen sich durch (vgl. MCCULLAGH AND NELDER, 1989, S. 38)

$$r^A = \frac{3(y^{1/3} - \mu^{1/3})}{\mu^{1/3}}.$$

Ferner sind die *deviance residuals* durch

$$r^d = \text{sign}(y - \mu) \sqrt{2((y - \mu)/\mu - \log(y/\mu))}$$

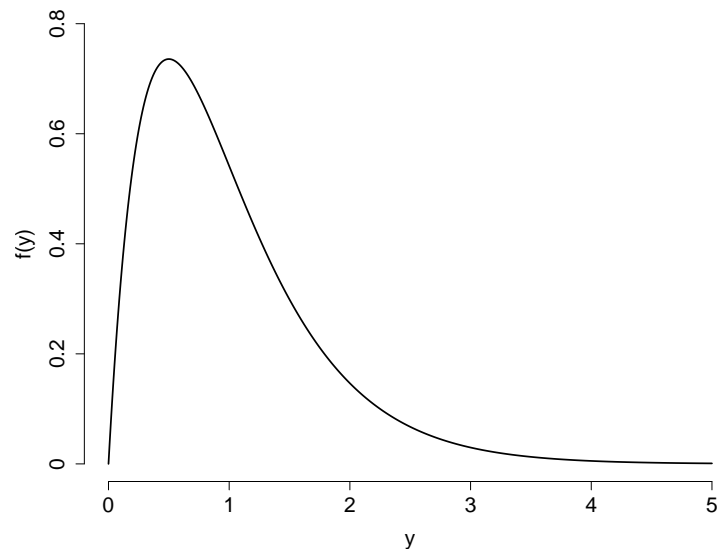


Abbildung 2.2: Dichtefunktion einer Gammaverteilung mit den Parametern $\nu = \lambda = 2$.

bestimmt (vgl. MCCULLAGH AND NELDER, 1989, S. 34) und die hier betrachtete Variante der Pseudo-Residuen sei

$$r = \Phi^{-1}(F_Y(y)),$$

wobei $F_Y(y)$ in diesem Fall die Verteilungsfunktion der Gammaverteilung beschreibt.

Eine graphische Analyse der Residuen erfolgt häufig durch deren Darstellung in einem QQ-Plot, d.h. die der Größe nach sortierten Residuen werden gegen die Quantile der Verteilung der Residuen bei Gültigkeit des Modells geplottet. Diese Referenzverteilung ist für die in diesem Beispiel betrachteten Residuen eine Standardnormalverteilung. Liegen die Punkte des QQ-Plots in etwa auf der Winkelhalbierenden, beschreibt die zu überprüfende Verteilung die Daten korrekt. Abweichungen von der Winkelhalbierenden hingegen werden in der Regel als Hinweis auf eine schlechte Modellanpassung interpretiert. Diese treten jedoch ebenfalls auf, wenn die Referenzverteilung und die tatsächliche Verteilung der Residuen bei Gültigkeit des Modells voneinander abweichen. Es ist also möglich, dass eine zu überprüfende Verteilung die Beobachtungen eines Datensatzes korrekt beschreibt, die Punkte des QQ-Plots aber dennoch von der Winkelhalbierenden abweichen. Der erwartete Verlauf eines QQ-Plots bei

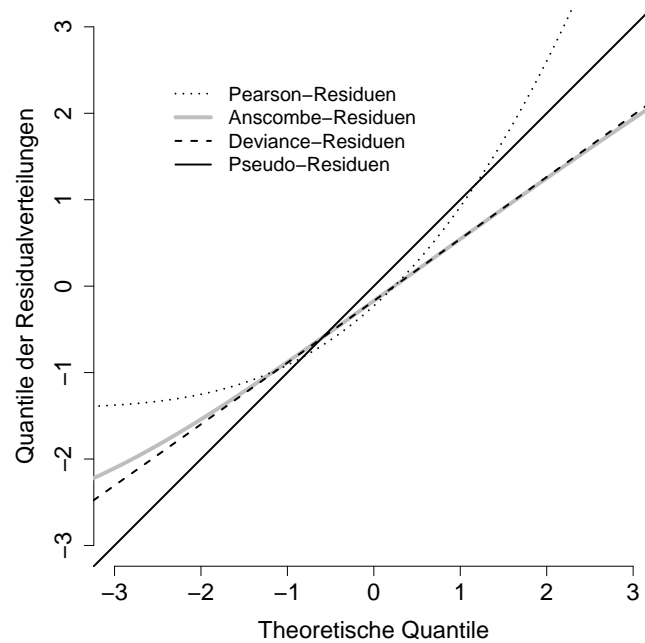


Abbildung 2.3: Erwarteter Verlauf des QQ-Plots für vier verschiedene Residuen bei Gültigkeit der in Abbildung 2.2 dargestellten Gammaverteilung.

Gültigkeit des Modells kann daher als Gütekriterium der Residuen interpretiert werden (vgl. dazu auch ZUCCHINI, 2002).

Abbildung 2.3 zeigt den erwarteten Verlauf der QQ-Plots bei Gültigkeit des Modells für die in diesem Beispiel betrachteten Residuen. Für die zu überprüfende Verteilung wurde hier eine Gammaverteilung mit den Parametern $\nu = \lambda = 2$ gewählt (siehe Abbildung 2.2). Es ist zu erkennen, dass bei Verwendung der *Pearson residuals* deutliche Abweichungen von der Winkelhalbierenden zu erwarten sind, obwohl das Modell korrekt ist. Die Ursache hierfür ist die ausgeprägte Schiefe der Gammaverteilung, die bei der Definition der *Pearson residuals* nicht berücksichtigt ist, so dass die Approximation ihrer Verteilung durch die Standardnormalverteilung sehr ungenau ist. Die *Anscombe residuals*, bei denen die Schiefe der zu überprüfenden Verteilung berücksichtigt ist, und die *deviance residuals* schneiden im Vergleich zu den *Pearson residuals* besser ab. Dennoch weicht auch bei ihnen der erwartete Verlauf des QQ-Plots von der Winkelhalbierenden ab. Ferner ist anhand der Graphik die zuvor erwähnte

Ähnlichkeit der *Anscombe residuals* und der *deviance residuals* zu erkennen. Nur die Pseudo-Residuen sind (per Definition) exakt standardnormalverteilt, so dass der erwartete Verlauf und die Winkelhalbierende übereinstimmen.



2.2 Definition eines stochastischen Modells

Im letzten Abschnitt wurde bereits deutlich, dass das Residual einer Beobachtung durch zwei Komponenten bestimmt wird: dem Modell und der Beobachtung selbst. Während im letzten Abschnitt eine Normalverteilung sowie eine Gamma-Verteilung beispielhaft als Modell verwendet wurden, wird der Begriff des stochastischen Modells in diesem Abschnitt allgemeiner formuliert und erläutert: Alle in dieser Arbeit betrachteten Beispiele lassen sich auf dieses Modell zurückführen. Dies gilt allerdings auch für Problemstellungen aus anderen Bereichen der statistischen Modellbildung. Es wird damit deutlich, dass die weiteren Kapitel der Arbeit „lediglich“ eine Auswahl möglicher Situationen darstellt, in denen die Pseudo-Residuen für die Überprüfung stochastischer Modelle verwendet werden können.

Alle stochastischen Modelle, die in den nachfolgenden Kapiteln betrachtet werden, können als Menge von unabhängigen Zufallsvariablen $\{Y_1, \dots, Y_n\}$ interpretiert werden, deren Verteilung in der Form

$$Y_i \sim G_i(\boldsymbol{\theta}_i) \quad \text{für } i = 1, \dots, n \quad (2.3)$$

spezifiziert ist. Dabei symbolisiert G_i eine beliebige Verteilung, die vom Parameter $\boldsymbol{\theta}_i$ abhängt und n beschreibt die Anzahl der Beobachtungen. Die Verwendung der Indizes am Parametervektor und am Verteilungssymbol impliziert, dass sowohl die Parameter der Verteilungen als auch die Verteilungstypen für die einzelnen Beobachtungen unterschiedlich sein können. Je nachdem, ob der Verteilungstyp und/oder die Parameter der Zufallsvariablen unterschiedlich oder identisch sind, können die drei folgenden Situationen abgegrenzt werden:

- (1) $Y_i \sim G(\boldsymbol{\theta})$
- (2) $Y_i \sim G(\boldsymbol{\theta}_i)$
- (3) $Y_i \sim G_i(\boldsymbol{\theta}_i)$

In der ersten Situation handelt es sich um ein Modell für n unabhängig und identisch verteilte Zufallsvariablen. Aus Sicht der Modellüberprüfung geht es

hier um die Frage, ob die Beobachtungen einer Zufallsstichprobe korrekt durch die Verteilung G mit dem Parameter θ beschrieben werden. Diese Situation wird beispielsweise in Abschnitt 2.5 betrachtet. Dort wird überprüft, ob eine Poissonverteilung geeignet ist, die Anzahl emittierter α -Partikel eines radioaktiven Materials zu beschreiben.

Die zweite Situation charakterisiert in erster Linie verschiedene Formen der Regression. Es handelt sich hier um Modelle, bei denen nicht davon ausgegangen wird, dass alle Zufallsvariablen des Modells identisch verteilt sind. Zwar besitzen sie den gleichen Verteilungstyp, ihre Parameter aber hängen von erklärenden Variablen ab. Die verallgemeinerten linearen Modelle, die in Kapitel 3 betrachtet werden, sind in diese Situation einzuordnen. Ferner sind die GARCH-Modelle zur Modellierung der Aktienrenditen (siehe Kapitel 4), bei denen die Voraussagen durch Normalverteilungen mit zeitabhängiger Varianz spezifiziert sind, ein Beispiel für diese Situation.

Weniger häufig ist der dritte Fall, in dem sowohl unterschiedliche Verteilungstypen als auch unterschiedliche Parameter für die Zufallsvariablen des Modells unterstellt werden. Ein in dieser Arbeit betrachtetes Beispiel ist das Modell mit gleitenden Dichten, das in Kapitel 4 zur Prognose von Aktienrenditen verwendet wird. Bei diesem Modell kann die Voraussageverteilung als Mischverteilung von Normalverteilungen dargestellt werden, wobei die Anzahl der Mischungskomponenten im Zeitverlauf steigt.

Beispiel 2.2 Viele Beispiele dieser Arbeit basieren auf der zweiten Situation, d.h. das zu überprüfende Modell impliziert den gleichen Verteilungstyp mit jeweils unterschiedlichen Parametern für die Zufallsvariablen des Modells. Es sei daher für das Modell der klassischen linearen Regression dargestellt, wie dessen Einbettung in das allgemeine stochastische Modell zu verstehen ist. Das Modell der klassischen linearen Regression kann durch

$$Y_i \sim N(\mu_i; \sigma^2) \quad \text{für } i = 1, \dots, n \text{ mit } \text{Cov}(Y_i, Y_j) = 0 \text{ für } i \neq j$$

dargestellt werden (vgl. beispielsweise FAHRMEIR, KAUFMANN UND KREDLER, 1984, S. 85). Es handelt sich im Sinn der Definition 2.3 um ein Modell mit n unabhängigen Zufallsvariablen, die allgemeine Verteilung G_i ist für alle i durch eine Normalverteilung spezifiziert und für die Parameter der Verteilungen gilt $\theta_i = (\mu_i, \sigma^2)'$.

◀

Das allgemeine stochastische Modell aus Gleichung 2.3 kann alternativ durch

$$Y_i \text{ besitzt die Verteilungsfunktion } F_i(y; \theta_i) \quad \text{für } i = 1, \dots, n \quad (2.4)$$

beschrieben werden. Bei dieser Notation rücken die Verteilungsfunktionen der Zufallsvariablen in den Vordergrund. Auf deren Grundlage basiert die allgemeine Definition der Pseudo-Residuen, die in den beiden nachfolgenden Abschnitten erläutert wird.

2.3 Pseudo-Residuen bei der Modellierung stetiger Merkmale

Die Definition der Pseudo-Residuen bei der Modellierung stetiger Merkmale wurde bereits im Rahmen der Abgrenzung der grundlegenden Ideen bei der Definition von Residuen (Abschnitt 2.1) skizziert. Diese Definition wird hier auf der Grundlage des stochastischen Modells aus dem letzten Abschnitt präzisiert. Ferner wird der Zusammenhang zu den *Cox-Snell residuals* beschrieben und die Verteilung der Pseudo-Residuen hergeleitet.

Definition stetiger Pseudo-Residuen Seien y_1, \dots, y_n Beobachtungen der stetigen Zufallsvariablen Y_1, \dots, Y_n . Zu überprüfen sei die Plausibilität der Beobachtungen unter dem Modell 2.4. Dann ist die erste Variante der Pseudo-Residuen für die i -te Beobachtung durch den Wert der Verteilungsfunktion F_i an der Stelle der Beobachtung definiert:

$$u_i = F_i(y_i; \boldsymbol{\theta}_i) \quad \text{für } i = 1, \dots, n. \quad (2.5)$$

Transformiert man diese Residuen mit der Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung, erhält man die Definition der zweiten Variante der Pseudo-Residuen:

$$r_i = \Phi^{-1}(F_i(y_i; \boldsymbol{\theta}_i)) = \Phi^{-1}(u_i) \quad \text{für } i = 1, \dots, n. \quad (2.6)$$

Um die beiden Definitionen begrifflich zu unterscheiden, werden die Residuen der zweiten Variante im Folgenden als Normal-Pseudo-Residuen (abgekürzt durch NPR) und die der ersten Variante als *Uniform*-Pseudo-Residuen (abgekürzt durch UPR) bezeichnet. Die Wahl dieser Begriffe erklärt sich aus der Tatsache, dass bei Gültigkeit des Modells die NPR standardnormalverteilt und die UPR $U(0, 1)$ -verteilt sind (siehe unten).

Beispiel 2.3 Betrachtet sei für gegebenes i die Beobachtung $y_i = 1.8$, deren Pseudo-Residuen (UPR und NPR) unter dem Modell einer Exponentialverteilung mit Parameter $\lambda_i = 1$ zu bestimmen sind. (Es sei bemerkt, dass in dieser

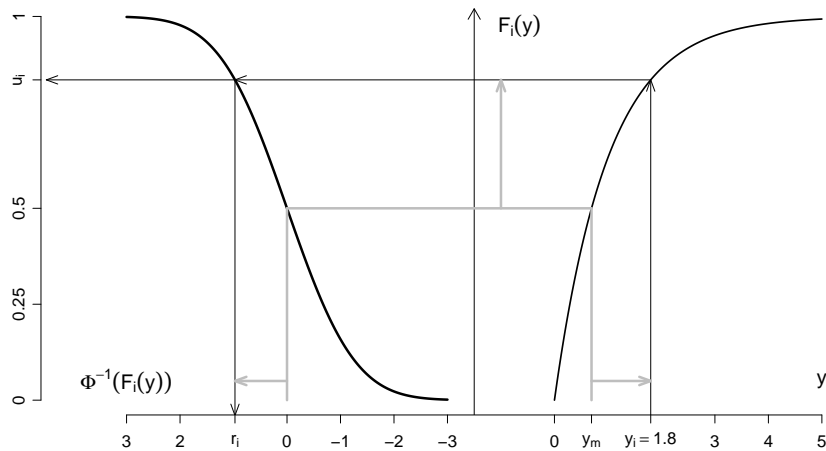


Abbildung 2.4: Pseudo-Residuen (UPR und NPR) einer stetigen Beobachtung.

Arbeit ein Punkt statt eines Kommas verwendet wird, um die Dezimalstellen einer Zahl kenntlich zu machen. Dies erscheint sinnvoll, da die verwendete und programmierte Software ebenso verfährt.) Man erhält in diesem Fall

$$u_i = 1 - e^{-1.8} \approx 0.84$$

$$\text{bzw. } r_i = \Phi^{-1}(1 - e^{-1.8}) \approx 0.97.$$

Abbildung 2.4 zeigt eine graphische Veranschaulichung der Berechnungen. Der Wert der Verteilungsfunktion (rechte Seite der Graphik) an der Stelle $y_i = 1.8$ entspricht dem UPR und dessen Transformation mit Φ^{-1} (linke Seite) liefert das NPR.

Das UPR der Beobachtung (definiert als Wert der Verteilungsfunktion an der Stelle der Beobachtung) kann als Wahrscheinlichkeitsaussage interpretiert werden: Unter dem Modell einer Exponentialverteilung mit Parameter $\lambda_i = 1$ ist die Wahrscheinlichkeit, eine Beobachtung zu erhalten, die kleiner ist als $y_i = 1.8$, in etwa 0.84.

Das NPR kann als das Quantil einer Standardnormalverteilung interpretiert werden, das dem beobachteten Quantil (also der Beobachtung selbst) unter dem Modell entspricht. Entspricht eine Beobachtung beispielsweise dem 50%-Quantil (also dem Median) der unterstellten Verteilung, so ist das entsprechende Quantil der Standardnormalverteilung und somit das NPR null. Dieser Spezialfall ist in der Abbildung durch die grau eingezeichneten Linien dargestellt. Er macht deutlich, dass die NPR die Abweichung zwischen der Beobachtung und dem Median der unterstellten Verteilung quantifizieren: Im hier

betrachteten Beispiel ist die Beobachtung $y_i = 1.8$ größer als der Median der Exponentialverteilung ($y_m \approx 0.69$). Daher ist das entsprechende NPR größer als null.



Pseudo-Residuen und Cox-Snell residuals Wie bereits weiter oben erwähnt, sind die Pseudo-Residuen ein Spezialfall der *Cox-Snell residuals*. Diese wurden 1968 von COX AND SNELL für ein Modell der Form

$$Y_i = g_i(\boldsymbol{\beta}, \epsilon_i) \quad \text{für } i = 1, \dots, n \quad (2.7)$$

definiert, wobei die g_i streng monotone Funktionen sind und $\boldsymbol{\beta}$ ein unbekannter Parametervektor ist. Ferner beschreiben die ϵ_i unabhängig und identisch verteilte Störvariablen, deren Verteilung bekannt bzw. durch die Modellformulierung impliziert ist (vgl. DAVISON AND GIGLI, 1989). Das Residual einer Beobachtung y_i definieren COX AND SNELL durch

$$e_i = g_i^{-1}(y_i, \hat{\boldsymbol{\beta}}).$$

Der Zusammenhang mit den Pseudo-Residuen ist wie folgt zu verstehen: Seien H_i die Verteilungsfunktionen der Störvariablen ϵ_i und sei $\boldsymbol{\beta}$ bekannt, dann gilt:

$$\begin{aligned} e_i = g_i^{-1}(y_i, \boldsymbol{\beta}) &= H_i^{-1}(H_i(g_i^{-1}(y_i, \boldsymbol{\beta}))) \\ &= H_i^{-1}(P\{\epsilon_i \leq g_i^{-1}(y_i, \boldsymbol{\beta})\}) \\ &= H_i^{-1}(P\{g_i(\epsilon_i, \boldsymbol{\beta}) \leq y_i\}) \\ &= H_i^{-1}(P\{Y_i \leq y_i\}) \\ &= H_i^{-1}(F_i(y_i)). \end{aligned}$$

Sind die Störvariablen des Modells 2.7 rechteckverteilt über dem Intervall $[0, 1]$, erhält man

$$e_i = H_i^{-1}(F_i(y_i)) = F_i(y_i) = u_i,$$

d.h. in diesem Fall entsprechen die *Cox-Snell residuals* den UPR. Analog ergeben sich die NPR als Spezialfall, wenn die Störvariablen standardnormalverteilt sind:

$$e_i = H_i^{-1}(F_i(y_i)) = \Phi^{-1}(F_i(y_i)) = r_i.$$

Der Unterschied zwischen den *Cox-Snell residuals* und den Pseudo-Residuen kann anhand der einfachen linearen Regression verdeutlicht werden. Wird dieses Modell in der Form

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

notiert, wird eine Annahme über die Verteilung der zu erklärenden Variable getroffen. Auf dieser Verteilung basiert die Definition der Pseudo-Residuen. Alternativ kann das Modell durch

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{mit} \quad \epsilon_i \sim N(0, \sigma^2)$$

formuliert werden. Hier wird eine Annahme über die Verteilung der Störvariablen getroffen, die bei bekanntem Parameter den *Cox-Snell residuals* entsprechen. Je nach betrachtetem Modell können verschiedene Annahmen über die Verteilung der Störvariablen sinnvoll sein, d.h. auch die Verteilung der *Cox-Snell residuals* ist nicht einheitlich definiert bzw. abhängig vom betrachteten Modell (vgl. WALKER, 1968). Im Gegensatz dazu ist die Verteilung der Pseudo-Residuen unabhängig vom betrachteten Modell.

Verteilung der Pseudo-Residuen Es wurde bereits mehrfach darauf hingewiesen, dass die UPR bei Gültigkeit des Modells $U(0, 1)$ -verteilt und die NPR standardnormalverteilt sind. Im Folgenden wird die Verteilung der Pseudo-Residuen allgemein, d.h. unabhängig von der Gültigkeit des Modells, hergeleitet.

Sei das zu überprüfende Modell weiterhin durch (2.4) beschrieben. Den Verteilungsfunktionen aus dem Modell stehen die Verteilungsfunktionen gegenüber, die das betrachtete Phänomen tatsächlich erzeugen. Letztere werden im Folgenden als wahres oder korrektes Modell bezeichnet und mit G_i bezeichnet. Außerdem seien Q_i die Verteilungsfunktionen der UPR. Mit diesen Definitionen lässt sich das wahre Modell wie folgt darstellen:

$$\begin{aligned} G_i(y) &= P\{Y_i \leq y\} \\ &= P\{F_i(Y_i) \leq F_i(y)\} \\ &= Q_i(F_i(y)) \quad \text{für } i = 1, \dots, n. \end{aligned}$$

Differenziert man die Gleichung und formt das Ergebnis um, erhält man für die Dichtefunktionen der UPR:

$$\begin{aligned} g_i(y) &= q_i(F_i(y)) \cdot f_i(y) \\ \Leftrightarrow q_i(u) &= \frac{g_i(F_i^{-1}(u))}{f_i(F_i^{-1}(u))} \quad \text{für } i = 1, \dots, n. \end{aligned} \quad (2.8)$$

Wenn das zu überprüfende Modell und das wahre Modell übereinstimmen ($f_i = g_i$ für alle i), ergibt sich $q_i(u) = 1$, d.h. die UPR sind $U(0, 1)$ -verteilt.

Unterscheidet sich das zu überprüfende Modell vom wahren Modell, kann die Verteilung der UPR nicht allgemein bestimmt werden, weil die Dichtefunktionen g_i des wahren Modells unbekannt sind. Es ist in diesem Fall lediglich möglich, die Dichtefunktionen der UPR für ausgewählte Fehlspezifikationen zu bestimmen. Dies kann hilfreich sein, wenn bei der Analyse der Residuen eine Fehlspezifikation identifiziert wird und somit eine Modifikation des Modells notwendig ist. Mit anderen Worten: Es ist hilfreich, die Verteilung der Residuen bei bestimmten Fehlspezifikationen zu kennen, weil dann aus der beobachteten Verteilung der Residuen eventuell auf die Art der notwendigen Modifikationen des Modells geschlossen werden kann.

Beispiel 2.4 Sei das zu überprüfende Modell durch n unabhängig exponentialverteilte Zufallsvariablen mit identischen Parametern λ charakterisiert. Das wahre Modell sei ebenfalls eine Menge von unabhängig exponentialverteilten Zufallsvariablen, deren Parameter μ sich allerdings von den Parametern des zu überprüfenden Modells unterscheiden. In diesem Fall sind die Dichtefunktionen des zu überprüfenden bzw. des wahren Modells durch

$$f_i(y) = \lambda e^{-\lambda y} \quad \text{bzw.} \quad g_i(y) = \mu e^{-\mu y}$$

bestimmt. Unter Verwendung der Umkehrfunktionen der Verteilungsfunktionen aus dem zu überprüfenden Modell

$$F_i^{-1}(u) = -\frac{1}{\lambda} \log(1 - u)$$

und durch Anwendung von 2.8 ergibt sich für die Dichtefunktion der UPR:

$$\begin{aligned} q_i(u) &= \frac{\mu \cdot \exp\left\{\frac{\mu}{\lambda} \log(1 - u)\right\}}{\lambda \cdot \exp\left\{\frac{\lambda}{\lambda} \log(1 - u)\right\}} \\ &= \frac{\mu \cdot (\exp\{\log(1 - u)\})^{\mu/\lambda}}{\lambda(1 - u)} \\ &= \mu/\lambda \cdot (1 - u)^{\mu/\lambda - 1} \quad \text{für } i = 1, \dots, n. \end{aligned}$$

Es handelt sich hier um die Dichtefunktion einer Beta-Verteilung mit den Parametern $\alpha = 1$ und $\beta = \mu/\lambda$. Für $\mu = \lambda$, also bei Übereinstimmung des wahren und des zu überprüfenden Modells ergibt sich mit $\beta = 1$ die $U(0, 1)$ -Verteilung. Für $\lambda > \mu$ bzw. $\beta < 1$ ist die Verteilung der UPR nach rechts verschoben, d.h. bei einer Unterschätzung des wahren Erwartungswerts $1/\mu$ sind die UPR im Durchschnitt größer als dies bei Verwendung des korrekten Modells zu erwarten wäre. Umgekehrt sind die UPR im Mittel kleiner, wenn der Erwartungswert des wahren Modells überschätzt wird.



Analog zum obigen Vorgehen erhält man die Dichtefunktion der NPR. Auch die Verteilungsfunktionen der NPR (hier mit K_i bezeichnet) können verwendet werden, um die Verteilungsfunktionen des wahren Modells darzustellen:

$$\begin{aligned} G_i(y) &= P\{Y_i \leq y\} \\ &= P\{\Phi^{-1}(F_i(Y_i)) \leq \Phi^{-1}(F_i(y))\} \\ &= K_i(\Phi^{-1}(F_i(y))) \quad \text{für } i = 1, \dots, n. \end{aligned}$$

Und auch hier ist die Gleichung zu differenzieren und umzuformen, um die Dichtefunktionen der Residuen zu bestimmen. Man erhält

$$\begin{aligned} g_i(y) &= k_i(\Phi^{-1}(F_i(y))) \cdot \frac{\partial \Phi^{-1}(F_i(y))}{\partial F_i(y)} \cdot f_i(y) \\ \Leftrightarrow g_i(y) &= \frac{k_i(\Phi^{-1}(F_i(y))) \cdot f_i(y)}{\varphi(\Phi^{-1}(F_i(y)))} \\ \Leftrightarrow k_i(r) &= \frac{g_i(F_i^{-1}(\Phi(r)))}{f_i(F_i^{-1}(\Phi(r)))} \cdot \varphi(r) \quad \text{für } i = 1, \dots, n, \end{aligned} \quad (2.9)$$

wobei φ die Dichtefunktion der Standardnormalverteilung darstellt. Wenn das wahre Modell und das zu überprüfende Modell übereinstimmen, gilt $k_i(r) = \varphi(r)$, d.h. die NPR sind bei Gültigkeit des Modells standardnormalverteilt. Wie im Fall der UPR sind auch die Dichtefunktionen der NPR nicht allgemein bzw. nur für ausgewählte Fehlspezifikationen zu bestimmen, wenn das zu überprüfende Modell vom wahren Modell abweicht.

Beispiel 2.5 Zu überprüfen sei ein einfaches verallgemeinertes lineares Modell mit exponentialverteilter zu erklärender Variable, Identitätslink (eine allgemeine Erläuterung der Linkfunktion eines verallgemeinerten linearen Modells wird in Abschnitt 3.2 gegeben) und einer erklärenden Variablen:

$$F_i(y) = 1 - e^{-\lambda_i y} \quad \text{mit } E(Y_i) = 1/\lambda_i = \beta_0 x_i \quad \text{für } i = 1, \dots, n.$$

Angenommen, das wahre Modell

$$G_i(y) = 1 - e^{-\mu_i y} \quad \text{mit } E(Y_i) = 1/\mu_i = \alpha_1 + \beta_1 x_i \quad \text{für } i = 1, \dots, n$$

besitzt zusätzlich einen Achsenabschnitt, dann ergibt sich für die Dichtefunktion des i -ten NPR (die Berechnung erfolgt analog zur Berechnung der Dichte der UPR im vorangegangenen Beispiel):

$$k_i(r) = \frac{\mu_i}{\lambda_i} (1 - \Phi(r))^{\mu_i/\lambda_i - 1} \cdot \varphi(r) \quad \text{für } i = 1, \dots, n.$$

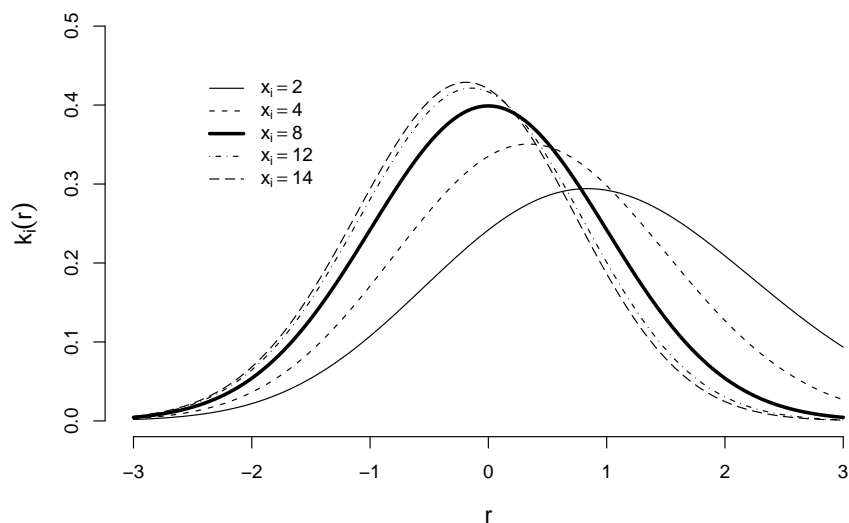


Abbildung 2.5: Dichtefunktion der NPR für unterschiedliche Werte der erklärenden Variablen.

Schreibt man die Dichte als Funktion der Parameter der linearen Erklärungsfunktionen (β_0 , α_1 und β_1) und der Beobachtungen für die erklärende Variable, erhält man

$$k_i(r) = \frac{\beta_0 x_i}{\alpha_1 + \beta_1 x_i} (1 - \Phi(r))^{\frac{\beta_0 x_i}{\alpha_1 + \beta_1 x_i} - 1} \cdot \varphi(r) \quad \text{für } i = 1, \dots, n.$$

Während die Verteilung der Residuen für die unterschiedlichen Beobachtungen im letzten Beispiel identisch waren, ist dies hier nicht der Fall. Beispielsweise ergibt sich am Schnittpunkt der linearen Erklärungsfunktionen, also für

$$\begin{aligned} \beta_0 x_c &= \alpha_1 + \beta_1 x_c \\ \Leftrightarrow x_c &= \frac{\alpha_1}{\beta_0 - \beta_1} \end{aligned}$$

die Dichtefunktion der Standardnormalverteilung, obwohl das zu überprüfende Modell und das wahre Modell nicht übereinstimmen. Für $x_i \neq x_c$ ist die Dichtefunktion schief sowie nach links oder rechts verschoben. Ist beispielsweise der Steigungsparameter β_0 des zu überprüfenden Modells größer als der des wahren Modells (β_1), wird der Erwartungswert der zu erklärenden Variable für $x_i < x_c$ unterschätzt und für $x_i > x_c$ überschätzt. Demnach ist die Dichtefunktion der NPR für $x_i < x_c$ nach rechts und für $x_i > x_c$ nach links verschoben. Beispielhaft ist dies in Abbildung 2.5 verdeutlicht. Sie zeigt die Dichtefunktion der NPR für $\beta_0 = 1$, $\alpha_1 = 4$ und $\beta_1 = 0.5$ bei unterschiedlichen Werten

der erklärenden Variablen. Am Schnittpunkt der linearen Erklärungsfunktionen ($x_c = 8$) sind die NPR standardnormalverteilt. Links vom Schnittpunkt (für $x_i = 2$ und $x_i = 4$) bzw. rechts davon (für $x_i = 12$ und $x_i = 14$) ist die Dichtefunktion der NPR nach rechts bzw. links verschoben.



Bislang wurde bei der Bestimmung der Verteilung der Pseudo-Residuen davon ausgegangen, dass die Parameter des zu überprüfenden Modells bekannt sind. Werden diese aus einer Stichprobe geschätzt, sind die Pseudo-Residuen auch bei Gültigkeit des Modells weder unabhängig noch $U(0,1)$ - bzw. standardnormalverteilt. Oft ist die Anzahl der zu schätzenden Parameter im Vergleich zur Anzahl der Beobachtungen gering, so dass sich die Eigenschaften der Pseudo-Residuen bei bekanntem und bei unbekanntem Parameter nur geringfügig unterscheiden (letztere werden im Folgenden als geschätzte Pseudo-Residuen bezeichnet). Ist dies jedoch nicht der Fall, muss der Einfluss der Parameterschätzung berücksichtigt werden, um Fehlinterpretationen bei der Residuenanalyse zu vermeiden.

Ein erster Ansatz zur Berücksichtigung der Parameterschätzung findet sich bei COX AND SNELL (1968), die den Erwartungswert, die Varianz und die Kovarianz der *Cox-Snell residuals* für ihr Modell (siehe Gleichung 2.7) bestimmen, wenn der Parameter β geschätzt wird. Wie zuvor gezeigt wurde, sind die Pseudo-Residuen ein Spezialfall der *Cox-Snell residuals*. Die Ergebnisse können demnach auf die Pseudo-Residuen übertragen werden, beispielsweise um eine Biaskorrektur für die geschätzten Pseudo-Residuen vorzunehmen.

Im Vergleich zu der Bestimmung der Momente der geschätzten Pseudo-Residuen ist die Bestimmung ihrer (gemeinsamen) Verteilung ein vollständigerer Ansatz zur Berücksichtigung der Parameterschätzung. Bereits 1948 betrachten DAVID AND JOHNSON (1948) die sogenannte *probability integral transform*

$$u_i = \int_{-\infty}^{y_i} f(y; \hat{\theta}) dy,$$

die als äquivalente Darstellung der geschätzten UPR zu verstehen ist. (Es sei bemerkt, dass hier von identischen Verteilungen der Y_i ausgegangen wird.) Sie leiten die gemeinsame Verteilung der U_i (und der Parameterschätzer) in allgemeiner Form her, weisen jedoch gleichzeitig darauf hin, dass die explizite Bestimmung der Verteilung in einer konkreten Anwendungssituation kompliziert sein kann.

Bei Kenntnis der Verteilung der geschätzten Pseudo-Residuen kann diese (anstatt der $U(0, 1)$ -Verteilung bzw. der Standardnormalverteilung) als Referenzverteilung bei der Residuenanalyse verwendet und somit der Einfluss der Parameterschätzung berücksichtigt werden. Eleganter ist es allerdings, die geschätzten Residuen so zu korrigieren, dass sie auch bei Schätzung der Parameter (und bei Gültigkeit des Modells) $U(0, 1)$ -verteilt bzw. standardnormalverteilt sind. Eine solche Korrektur ist bei LOYNES (1969) erläutert. Er stellt eine Transformation der *Cox-Snell residuals* vor, deren Verteilung gegen die Verteilung der Störvariablen ϵ_i konvergiert.

In dieser Arbeit wird die Berücksichtigung der Parameterschätzung bzw. die Korrektur der Pseudo-Residuen in Abschnitt 3.4 ausführlich am Beispiel der verallgemeinerten linearen Modelle illustriert. In allen übrigen Abschnitten wird davon ausgegangen, dass der Einfluss der Parameterschätzung zu vernachlässigen ist.

2.4 Pseudo-Residuen bei der Modellierung diskreter Merkmale

Definition diskreter Pseudo-Residuen Seien y_1, \dots, y_n Beobachtungen der diskreten Zufallsvariablen Y_1, \dots, Y_n . Zu überprüfen sei die Plausibilität der Beobachtungen unter dem Modell 2.4. Dann ist das UPR für die i -te Beobachtung durch

$$u_i = [F_i^-(y_i; \boldsymbol{\theta}_i), F_i^+(y_i; \boldsymbol{\theta}_i)] \equiv [u_i^-, u_i^+] \quad \text{für } i = 1, \dots, n \quad (2.10)$$

mit $F_i^-(y_i; \boldsymbol{\theta}_i) = P\{Y_i < y_i\}$ und $F_i^+(y_i; \boldsymbol{\theta}_i) = P\{Y_i \leq y_i\}$ definiert. Transformiert man die Intervallgrenzen des Intervalls mit der Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung, erhält man die Definition der NPR bei der Modellierung diskreter Merkmale:

$$\begin{aligned} r_i &= [\Phi^{-1}(F_i^-(y_i; \boldsymbol{\theta}_i)), \Phi^{-1}(F_i^+(y_i; \boldsymbol{\theta}_i))] \\ &\equiv [r_i^-, r_i^+] \quad \text{für } i = 1, \dots, n. \end{aligned} \quad (2.11)$$

Die grundlegende Idee, die Residuen bei der Modellierung diskreter Merkmale in Form von Intervallen zu definieren, ist bei MACDONALD AND ZUCCHINI (1997, S. 98) zu finden. ZUCCHINI AND MACDONALD (1999) bezeichnen die Definition nach Gleichung 2.11 als *pseudo-residual segment*. Daran angelehnt werden die UPR bzw. NPR für diskrete Merkmale im Folgenden als UPR-Intervalle bzw. NPR-Intervalle bezeichnet.

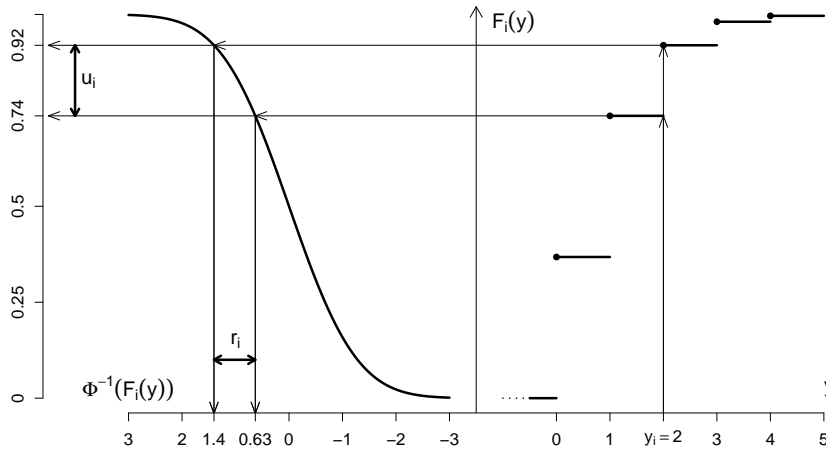


Abbildung 2.6: Pseudo-Residuen (UPR-Intervall und NPR-Intervall) einer diskreten Beobachtung.

Beispiel 2.6 Betrachtet sei für gegebenes i die Beobachtung $y_i = 2$, deren Pseudo-Residuen (UPR-Intervall und NPR-Intervall) unter dem Modell einer Poissonverteilung mit Parameter $\lambda_i = 1$ zu bestimmen sind. Es ergeben sich hier

$$[u_i^-, u_i^+] = \left[\sum_{y=0}^1 \frac{1^y e^{-1}}{y!}, \sum_{y=0}^2 \frac{1^y e^{-1}}{y!} \right] \approx [0.74, 0.92]$$

bzw. $[r_i^-, r_i^+] = [\Phi^{-1}(0.74), \Phi^{-1}(0.92)] \approx [0.63, 1.40]$.

Die Berechnung der Intervalle ist in Abbildung 2.6 verdeutlicht. Auf der rechten Seite ist die Verteilungsfunktion der hier betrachteten Poissonverteilung dargestellt, deren Wert an der Stelle $y_i = 2$ der rechten Grenze des UPR-Intervalls entspricht. Die Transformation mit Φ^{-1} (linke Seite der Graphik) liefert die rechte Grenze des NPR-Intervalls. Die linken Grenzen der Pseudo-Residual-Intervalle erhält man analog bei Betrachtung der Verteilungsfunktion an der Stelle $y = 1$, da $F_i^-(2) = P\{Y_i < 2\} = P\{Y_i \leq 1\}$ gilt.

Das UPR-Intervall enthält (wie in der stetigen Variante) Aussagen zur Wahrscheinlichkeit der Beobachtung unter dem Modell. Die linke Grenze u_i^- gibt die Wahrscheinlichkeit für eine kleinere Beobachtung als y_i unter dem Modell an, $1 - u_i^+$ für eine größere Beobachtung als y_i . Ferner entspricht die Intervallbreite der Wahrscheinlichkeit der Beobachtung unter dem Modell. Unter einer Poissonverteilung mit Parameter $\lambda_i = 1$ erhält man demnach mit Wahrscheinlichkeit 0.74 bzw. 0.08 eine kleinere bzw. größere Beobachtung als $y_i = 2$, und

die Wahrscheinlichkeit, unter der Verteilung genau diesen Wert zu realisieren, beträgt $0.92 - 0.74 = 0.18$.

Das NPR-Intervall ist der Bereich in der Standardnormalverteilung, der mit der Beobachtung y_i in der zu überprüfenden Verteilung korrespondiert. Während im stetigen Fall der Beobachtung ein eindeutiges Quantil in der Standardnormalverteilung zugeordnet werden kann, ist dies bei der diskreten Variante nicht möglich. Hier ist die Wahrscheinlichkeit der Beobachtung größer als null, d.h. sie besitzt Masse. Diese bleibt bei der Definition des NPR-Intervalls erhalten: Die Wahrscheinlichkeit einer Beobachtung im Intervall $[r_i^-, r_i^+]$ für eine Standardnormalverteilung entspricht der Wahrscheinlichkeit der Beobachtung y_i unter dem zu überprüfenden Modell.



Verteilung der Pseudo-Residuen Die UPR-Intervalle bzw. die NPR-Intervalle sind offensichtlich nicht $U(0,1)$ -verteilt bzw. standardnormalverteilt. Um die Verteilung der Intervalle für gegebenes i formal zu beschreiben, müssen ihnen entsprechende Wahrscheinlichkeiten zugeordnet werden, wobei für jede Zufallsvariable Y_i eine Menge von möglichen UPR-Intervallen bzw. NPR-Intervallen existiert, die sich aus der Struktur der Wahrscheinlichkeitsfunktion von Y_i ergibt. Die Wahrscheinlichkeiten dieser möglichen Intervalle stimmen mit den Wahrscheinlichkeiten der entsprechenden Beobachtungen von Y_i überein. Es wird an dieser Stelle deutlich, dass die formale Beschreibung der Verteilung wenig hilfreich ist: Es handelt sich um diskrete Verteilungen, deren mögliche Ausprägungen Intervalle sind. Darüber hinaus unterscheiden sich die möglichen Ausprägungen für unterschiedliche i , wenn die Zufallsvariablen Y_i keine identische Verteilung besitzen. An die Stelle einer formalen Beschreibung der Verteilung tritt daher die folgende Interpretation, die sich aus der Verteilung der UPR bzw. NPR bei der Modellierung stetiger Merkmale ableitet:

Im stetigen Fall sind die UPR bzw. NPR bei Gültigkeit des Modells $U(0,1)$ -verteilt bzw. standardnormalverteilt. In Anlehnung daran können die UPR-Intervalle bzw. die NPR-Intervalle bei Gültigkeit des Modells als intervallzensierte Beobachtungen einer $U(0,1)$ -Verteilung bzw. einer Standardnormalverteilung interpretiert werden.

Intervallzensierte Beobachtungen einer $U(0,1)$ -Verteilung sind Intervalle $[u_l, u_r] \subseteq [0, 1]$, deren Wahrscheinlichkeit gerade der Intervallbreite $u_r - u_l$

entspricht. Für das UPR-Intervall der i -ten Beobachtung gilt:

$$P\{[U_i^-, U_i^+] = [F_i^-(y_i; \boldsymbol{\theta}_i), F_i^+(y_i; \boldsymbol{\theta}_i)]\} = P\{Y_i = y_i\} \quad \text{für alle } y_i \in \Omega_i,$$

wobei Ω_i die Menge der möglichen Ausprägungen von Y_i kennzeichnet. Die Wahrscheinlichkeiten $P\{Y_i = y_i\} = G_i^+(y_i; \boldsymbol{\theta}_i) - G_i^-(y_i; \boldsymbol{\theta}_i)$ auf der rechten Seite der Gleichung sind durch das wahre Modell determiniert. Stimmen die Verteilungsfunktionen F_i des zu überprüfenden Modells mit denen des wahren Modells (repräsentiert durch die Verteilungsfunktionen G_i) überein, so entsprechen die Wahrscheinlichkeiten der UPR-Intervalle deren Breite. In diesem Fall verhalten sich die UPR-Intervalle demnach wie intervallzensierte Beobachtungen einer $U(0, 1)$ -Verteilung.

Die Wahrscheinlichkeit einer intervallzensierten Beobachtung $[r_l, r_r]$ mit $r_l, r_r \in \mathbb{R}$ aus einer Standardnormalverteilung beträgt $\Phi(r_r) - \Phi(r_l)$. Diese Eigenschaft ist bei Gültigkeit des Modells für die NPR-Intervalle erfüllt, da in diesem Fall gilt:

$$\begin{aligned} & P\{[R_i^-, R_i^+] = [\Phi^{-1}(F_i^-(y_i; \boldsymbol{\theta}_i)), \Phi^{-1}(F_i^+(y_i; \boldsymbol{\theta}_i))]\} \\ &= G^+(y_i; \boldsymbol{\theta}_i) - G^-(y_i; \boldsymbol{\theta}_i) \\ &= F^+(y_i; \boldsymbol{\theta}_i) - F^-(y_i; \boldsymbol{\theta}_i) \\ &= \Phi(\Phi^{-1}(F_i^+(y_i; \boldsymbol{\theta}_i))) - \Phi(\Phi^{-1}(F_i^-(y_i; \boldsymbol{\theta}_i))). \end{aligned}$$

Analog zum stetigen Fall gelten die zuvor beschriebenen Ergebnisse nur approximativ, wenn die Parameter des Modells geschätzt werden. Auch bei der Modellierung diskreter Merkmale ist jedoch eine Korrektur der Intervalle möglich, die in Abschnitt 3.4 erläutert wird.

2.5 Einführende Beispiele

Um die bisherigen Darstellungen zu verdeutlichen, werden im Folgenden zwei Anwendungsbeispiele betrachtet. Anhand der Beispiele werden zunächst einige der grundlegenden Ideen bei der Analyse von Residuen verdeutlicht, bevor in den nachfolgenden Kapiteln speziellere Problemstellungen betrachtet werden. Im Rahmen des ersten Beispiels wird der Brusthöhendurchmesser und die Höhe von Bäumen modelliert, d.h. es handelt sich um die Modellierung stetiger Merkmale. Im zweiten Beispiel wird die Beschreibung eines diskreten Merkmals (die Anzahl emittierter α -Partikel) durch eine Poissonverteilung betrachtet.

2.5.1 Überprüfung einer gemeinsamen Verteilung für den Brusthöhendurchmesser und die Höhe von Bäumen

In diesem Abschnitt wird die Überprüfung der Anpassungsgüte einer Mischverteilung betrachtet, deren Mischungskomponenten zwei bivariate Normalverteilungen sind. Dieses Modell wird zur Beschreibung der gemeinsamen Verteilung des Brusthöhendurchmessers und der Höhe von 1 242 Bäumen eines Buchenwalds im Solling verwendet. Die Modellanpassung wurde (mit Hilfe der NPR) bereits von ZUCCHINI ET AL. (2000) untersucht und mit der Güte anderer Modelle verglichen (siehe dazu auch ZUCCHINI AND MACDONALD, 1999).

Seien φ_1 und φ_2 die Dichtefunktionen der Mischungskomponenten (bivariate Normalverteilungen) mit den Parametern $\mu_{j,D}$, $\mu_{j,H}$, $\sigma_{j,D}^2$, $\sigma_{j,H}^2$ und ρ_j , wobei der Index $j = 1, 2$ die erste bzw. zweite Mischungskomponente anzeigt und der Index D bzw. H den (Brusthöhen-)Durchmesser bzw. die Höhe. Mit diesen Definitionen ist die gemeinsame Dichtefunktion von Brusthöhendurchmesser und Höhe durch

$$f(y_D, y_H) = \alpha \cdot \varphi_1(y_D, y_H) + (1 - \alpha) \cdot \varphi_2(y_D, y_H)$$

gegeben, wobei $0 < \alpha < 1$ den Mischungsparameter darstellt. Die Schätzer der bivariaten Normalverteilungen sind in Tabelle 2.1 gegeben. Der Schätzer des Mischungsparameters ist $\hat{\alpha} = 0.19$ (vgl. ZUCCHINI ET AL., 2000).

Abbildung 2.7 zeigt die Beobachtungen und die Höhenlinien der angepassten gemeinsamen Dichtefunktion. Es ist zu sehen, dass die grundlegenden Eigenschaften der Daten durch das Modell angemessen beschrieben werden. Detailspekte der Anpassungsgüte sind jedoch anhand dieser Graphik nicht zu beurteilen.

Die weiter oben definierten Pseudo-Residuen basieren auf univariaten Verteilungen. Daher werden in diesem Beispiel die Randverteilungen für den Brusthöhendurchmesser und die Höhe der Bäume als Grundlage der Modellüber-

j	$\hat{\mu}_{j,D}$	$\hat{\mu}_{j,H}$	$\hat{\sigma}_{j,D}^2$	$\hat{\sigma}_{j,H}^2$	$\hat{\rho}_j$
$j = 1$	18.00	17.91	5.16	4.85	0.88
$j = 2$	39.06	27.27	9.69	2.90	0.63

Tabelle 2.1: ML-Schätzer für die Mischverteilung der bivariaten Normalverteilungen.

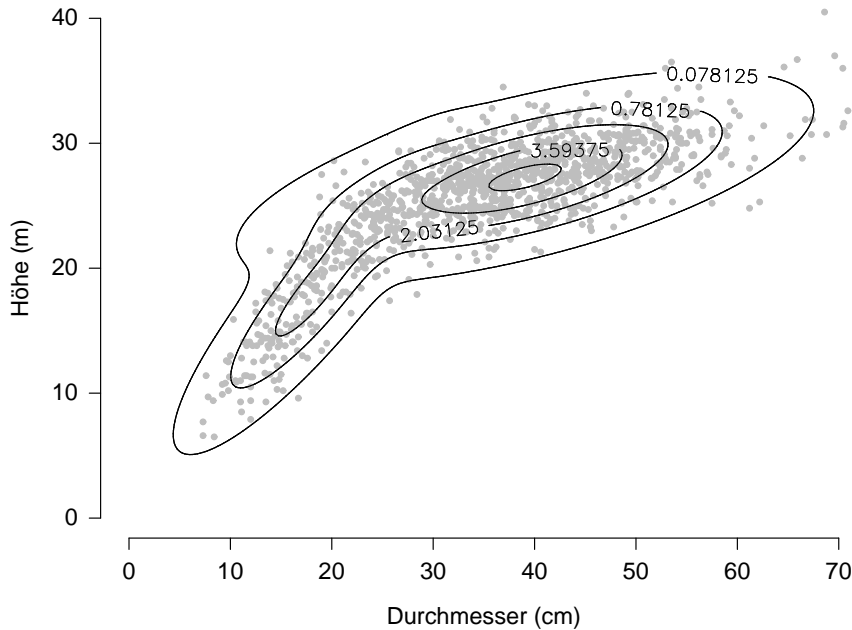


Abbildung 2.7: Beobachtungen sowie Höhenlinien der angepassten gemeinsamen Dichtefunktion.

prüfung verwendet. Die Dichtefunktionen der Randverteilungen

$$f_D(y_D) = \alpha \frac{\exp\left\{-\frac{(y_D - \mu_{1,D})^2}{2\sigma_{1,D}^2}\right\}}{\sqrt{2\pi\sigma_{1,D}^2}} + (1 - \alpha) \frac{\exp\left\{-\frac{(y_D - \mu_{2,D})^2}{2\sigma_{2,D}^2}\right\}}{\sqrt{2\pi\sigma_{2,D}^2}}$$

$$f_H(y_H) = \alpha \frac{\exp\left\{-\frac{(y_H - \mu_{1,H})^2}{2\sigma_{1,H}^2}\right\}}{\sqrt{2\pi\sigma_{1,H}^2}} + (1 - \alpha) \frac{\exp\left\{-\frac{(y_H - \mu_{2,H})^2}{2\sigma_{2,H}^2}\right\}}{\sqrt{2\pi\sigma_{2,H}^2}}$$

beschreiben Mischverteilungen univariater Normalverteilungen, deren Parameter sich aus den beiden Mischungskomponenten der bivariaten Mischverteilung ableiten.

Abbildung 2.8 zeigt die Histogramme für den Brusthöhendurchmesser und die Höhe der Bäume sowie die angepassten Randdichten. Auch aus diesem Blickwinkel beschreibt das Modell die Daten recht gut. Allerdings ist die Dichtefunktion für den Brusthöhendurchmesser bimodal. Diese Eigenschaft ist in den Daten (also im Histogramm) nicht zu erkennen. Im Gegensatz dazu erscheint die Beschreibung der beobachteten Baumhöhen (rechte Graphik) durch die entsprechende Dichtefunktion sehr gut. Die Beurteilung der Modellanpassung

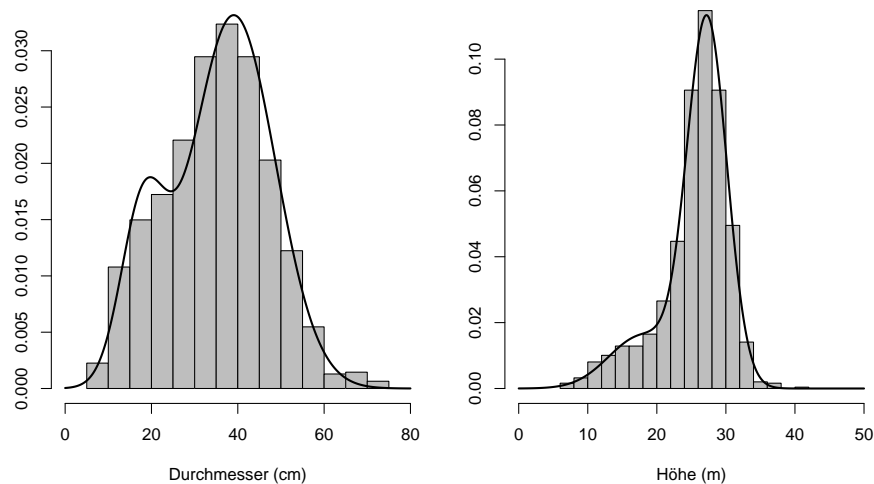


Abbildung 2.8: Histogramme für den Brusthöhendurchmesser und die Höhe der Bäume sowie angepasste Randverteilungen.

anhand dieser Graphiken ist allerdings weniger geeignet, weil die Dichtefunktionen der Randverteilungen keine einfache Struktur besitzen. Sinnvoller ist hier die Betrachtung der UPR für die Randverteilungen

$$\begin{aligned}
 u_{D i} &= \alpha \cdot \Phi(y_{D i}; \mu_{1,D}, \sigma_{1,D}^2) + (1 - \alpha) \cdot \Phi(y_{D i}; \mu_{2,D}, \sigma_{2,D}^2) \\
 u_{H i} &= \alpha \cdot \Phi(y_{H i}; \mu_{1,H}, \sigma_{1,H}^2) + (1 - \alpha) \cdot \Phi(y_{H i}; \mu_{2,H}, \sigma_{2,H}^2),
 \end{aligned}$$

die sich bei Gültigkeit des Modells wie Realisationen einer $U(0, 1)$ -Verteilung verhalten.

Abbildung 2.9 zeigt die Histogramme der UPR der Randverteilungen. Im linken Histogramm liegen mehr Beobachtungen links von 0.5 (657 Beobachtungen) als rechts davon (585 Beobachtungen). Die UPR sind also tendenziell kleiner als es unter dem Modell zu erwarten wäre. Dies ist ein Indiz dafür, dass der Brusthöhendurchmesser tendenziell überschätzt wird. Allerdings sind die Abweichungen des Histogramms von der $U(0, 1)$ -Verteilung nicht signifikant: Die vertikalen Pfeile in den Histogramm begrenzen einen aus den Daten geschätzten simultanen 99%-Konfidenzbereich für die erwarteten Höhen der Säulen. Da die erwartete Höhe der Säulen unter Gültigkeit des Modells eins beträgt und alle Pfeile diesen Wert abdecken, kann die Hypothese, dass die UPR $U(0, 1)$ -verteilt sind, nicht verworfen werden. Der hier verwendete Test basiert auf simultanen Konfidenzintervallen für die Erfolgswahrscheinlichkeiten einer Multinomialverteilung, die bei SISON AND GLAZ (1995) beschrieben sind.

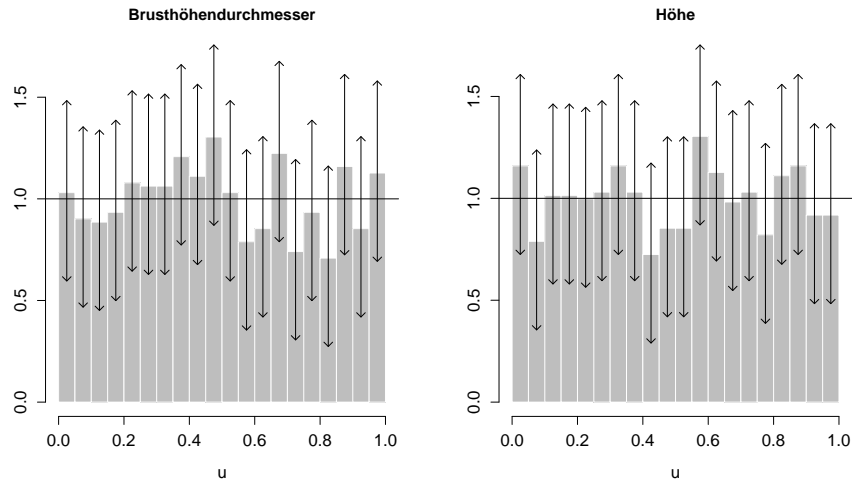


Abbildung 2.9: Histogramme der UPR für die Randverteilungen mit simultanen Konfidenzintervallen.

Um die Anpassung an den Verteilungsenden zu beurteilen, ist es sinnvoll die NPR zu betrachten, die bei Gültigkeit des Modells standardnormalverteilt sind. Diese Skalierung betont die Enden der Verteilung der Residuen und vereinfacht somit eine graphische Identifikation von Ausreißern, d.h. von Beobachtungen, die unter dem Modell zu weit am Rand liegen (zur Ausreißerproblematik vgl. beispielsweise BARNETT AND LEWIS, 1994, S. 11–16). Darüber hinaus ist man im Umgang mit der Standardnormalverteilung sehr vertraut, was bei der graphischen Analyse der NPR natürlich von Vorteil ist.

Eine Standardmethode zur graphischen Analyse der NPR ist ein QQ-Plot. Abbildung 2.10 zeigt die QQ-Plots der NPR für den Brusthöhendurchmesser und die Höhe der Bäume bezüglich der entsprechenden Randverteilungen. Bei Gültigkeit des Modells liegen die Punkte des Plots in etwa auf der Winkelhalbierenden. In den mittleren Bereichen der QQ-Plots ist diese Eigenschaft erfüllt, an den Enden ergeben sich jedoch Abweichungen von der Winkelhalbierenden. Für die NPR des Brusthöhendurchmessers (linke Graphik) liegen die Punkte an beiden Enden oberhalb der Winkelhalbierenden, d.h. an beiden Enden sind die Residuen (bzw. die Beobachtungen) größer als unter dem Modell und dem gegebenen Stichprobenumfang erwartet. Es sei an dieser Stelle erwähnt, dass die gemeinsame Analyse der Residuen immer vor dem Hintergrund des Stichprobenumfangs erfolgen muss. So zeigt der QQ-Plot auf der linken Seite Residuen, die kleiner sind als minus zwei. Separat beurteilt sind diese Residuen extrem. Bei dem hier betrachteten Stichprobenumfang allerdings wären bei Gültigkeit des Modells noch extremere Residuen zu erwarten,

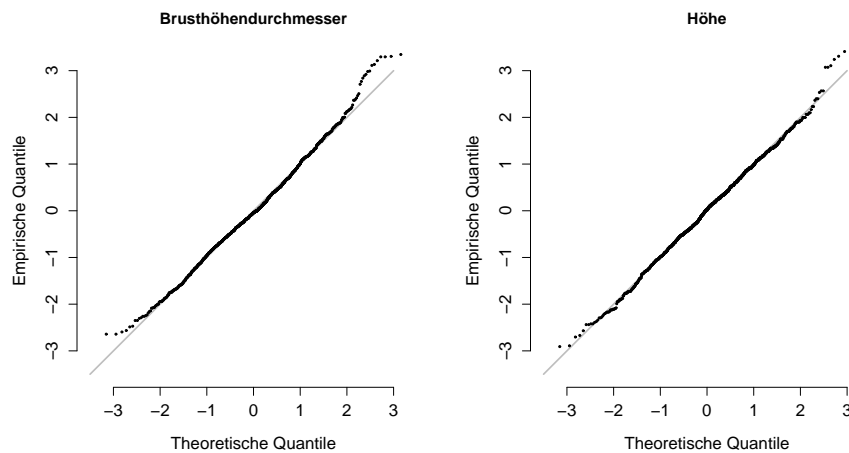


Abbildung 2.10: QQ-Plots der NPR für die Randverteilungen.

da die entsprechenden Punkte im QQ-Plot oberhalb der Winkelhalbierenden liegen. Dieser Aspekt ist auch bei vielen anderen Beispielen dieser Arbeit von Bedeutung, wobei auf einen expliziten Hinweis verzichtet wird, wenn offensichtlich ist, dass eine Berücksichtigung des Stichprobenumfangs erforderlich ist.

Im QQ-Plot der NPR für die Höhe der Bäume (rechte Graphik) sind am rechten Ende einige Punkte zu erkennen, die deutlich von der Struktur der übrigen Punkte abweichen. Auch die Beobachtungen, die zu diesen Residuen gehören, sind größer als unter dem Modell erwartet. Es handelt sich demnach um extrem große Beobachtungen, die nicht angemessen durch das Modell beschrieben werden, d.h. um Ausreißer unter dem Modell.

Im Rahmen dieses Beispiels wurden die Randverteilungen für den Brusthöhendurchmesser und die Höhe der Bäume zur Überprüfung des Modells verwendet. Die Aussagen zur Güte des Modells sind daher auf die Randverteilungen beschränkt. In Kapitel 5 wird dieses Beispiel erneut aufgegriffen, um Methoden zur Überprüfung der gemeinsamen Verteilung zu erläutern.

2.5.2 Überprüfung einer Poissonverteilung für den radioaktiven Zerfall von Polonium

In einer Untersuchung aus dem Jahr 1910 betrachteten RUTHERFORD AND GEIGER den radioaktiven Zerfall von Polonium. Sie zählten die Anzahl emit-

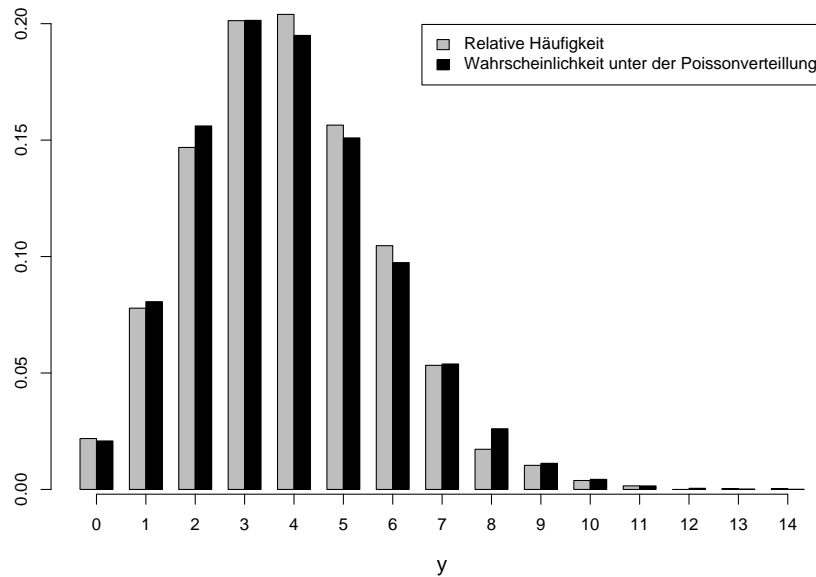


Abbildung 2.11: Relative Häufigkeiten für die Anzahl emittierter α -Partikel und Wahrscheinlichkeitsfunktion der angepassten Poissonverteilung.

tierter α -Partikel in $n = 2\,608$ Zeitintervallen mit einer Länge von jeweils 72 Sekunden. Die Daten der Untersuchung (hier zitiert nach HAND ET AL., 1994, S. 223) sind in Tabelle 2.2 wiedergegeben.

Unter der Annahme, dass der radioaktive Zerfall einem Poissonprozess folgt, ist die Anzahl der emittierten Partikel in den Zeitintervallen poissonverteilt, d.h. die beobachteten relativen Häufigkeiten sollten in diesem Fall mit den Werten einer angepassten Poissonverteilung in etwa übereinstimmen. Für den Parameterschätzer der Verteilung erhält man $\hat{\lambda} = \bar{y} \approx 3.87$. Abbildung 2.11 zeigt die beobachteten relativen Häufigkeiten und zum Vergleich die angepasste Poissonverteilung. Die Anpassung erscheint akzeptabel. HAND ET AL. (1994, S. 223) sprechen sogar von einer insgesamt exzellenten Übereinstimmung.

Anzahl emittierter Partikel	0	1	2	3	4	5	6	7
Beobachtete Häufigkeit	57	203	383	525	532	408	273	139
Anzahl emittierter Partikel	8	9	10	11	12	13	14	
Beobachtete Häufigkeit	45	27	10	4	0	1	1	

Tabelle 2.2: Häufigkeiten für die Anzahl emittierter Partikel in Zeitintervallen mit konstanter Länge.

i	beobachte Häufigkeit	erwartete Häufigkeit	Beitrag zu χ^2	i	beobachte Häufigkeit	erwartete Häufigkeit	Beitrag zu χ^2
0	57	54.31	0.13	8	45	67.99	7.78
1	203	210.28	0.25	9	27	29.25	0.17
2	383	407.06	1.42	10	10	11.32	0.15
3	525	525.31	0.00	11	4	3.99	0.00
4	532	508.44	1.09	12	0	1.29	1.29
5	408	393.69	0.52	13	1	0.38	0.99
6	273	254.03	1.42	14	1	0.11	7.55
7	139	140.50	0.02				

Tabelle 2.3: Häufigkeiten für die Anzahl emittierter Partikel und Beiträge zur χ^2 -Prüfgröße.

Der Datensatz wird häufig zur Veranschaulichung des χ^2 -Anpassungstests verwendet. Der Wert der Prüfgröße beträgt in etwa 22.78, was einem P-Wert von ca. 0.06 entspricht. Untersucht man die einzelnen Beiträge zur Prüfgröße (siehe Tabelle 2.3), zeigt sich, dass die Differenz zwischen den erwarteten und den beobachteten Häufigkeiten für $i = 8$ und $i = 14$ zu relativ großen Beiträgen führt. Die Anwendung des χ^2 -Anpassungstests ist in diesem Beispiel allerdings problematisch, weil die erwarteten Häufigkeiten zum Teil kleiner sind als fünf. Insbesondere für $i = 14$ stellt sich die Frage, ob der hohe Beitrag zur Prüfgröße die Folge einer schlechten Modellanpassung ist oder aus einer Verletzung der Annahmen des Signifikanztests resultiert.

Diese Frage kann durch eine Analyse der Pseudo-Residuen beantwortet werden: Bei Gültigkeit des Modells können die beobachteten Häufigkeiten y_i für $i = 0, \dots, 14$ als Realisationen von binomialverteilten Zufallsvariablen interpretiert werden. Dabei ergibt sich der Parameter $n = 2608$ aus der Gesamtzahl der Zeitintervalle und die Erfolgswahrscheinlichkeiten aus den Werten der Wahrscheinlichkeitsfunktion der Poissonverteilung. Bei Gültigkeit des Modells gilt demnach:

$$Y_i \sim b(n, \pi_i) \quad \text{für } i = 0, \dots, 14$$

mit $\pi_i = \frac{3.87^i e^{-3.87}}{i!}$.

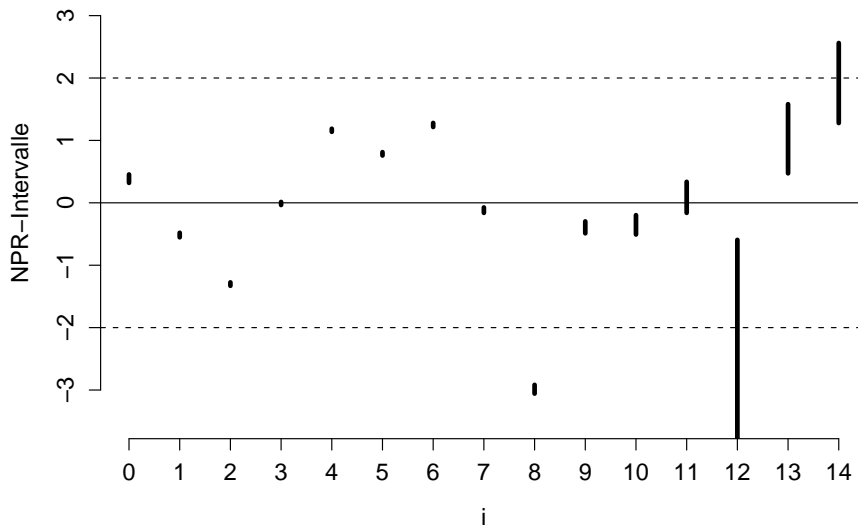


Abbildung 2.12: Indexplot der NPR-Intervalle für die beobachteten Häufigkeiten.

Daraus ergeben sich für jede beobachtete Häufigkeit die Intervallgrenzen

$$u_i^- = \sum_{y=0}^{y_i-1} \binom{2608}{y} \pi_i^y (1 - \pi_i)^{2608-y}$$

bzw.

$$u_i^+ = \sum_{y=0}^{y_i} \binom{2608}{y} \pi_i^y (1 - \pi_i)^{2608-y}$$

der UPR-Intervalle. Für $i = 14$ erhält man $u_{14} \approx [0.90, 0.99]$. Die Wahrscheinlichkeit, unter dem Modell bei 2608 beobachteten Zeitintervallen genau einmal den Wert 14 zu beobachten ($y_{14} = 1$), beträgt also in etwa 0.09. Vor diesem Hintergrund erscheint der relativ große Beitrag zur χ^2 -Prüfgröße eher ungerechtfertigt, weil die Beobachtung unter dem Modell nicht besonders ungewöhnlich ist. Im Gegensatz dazu ist die Wahrscheinlichkeit höchstens 45 mal die Acht zu beobachten ($y_8 = 45$) kleiner als 0.002. Die beobachtete Häufigkeit ist also wesentlich geringer als unter dem Modell erwartet.

Graphisch sind die Residuen in Abbildung 2.12 veranschaulicht. Sie zeigt einen Indexplot der NPR-Intervalle für die beobachteten Häufigkeiten. Die exponierte Stellung von $r_8 = [-3.05, -2.92]$ ist sofort zu erkennen, weil die extremen Residuen auch bei der Modellierung diskreter Merkmale durch die Transformation mit Φ^{-1} betont werden. Die übrigen NPR-Intervalle sind nicht auffällig. Insbesondere erscheint r_{14} auch bei der graphischen Analyse nicht extrem.

2.6 Abschließende Bemerkungen zum Kapitel

Im ersten Teil dieser Arbeit wurden allgemeine Fragestellungen erörtert, die mit dem Konzept der Pseudo-Residuen verbunden sind. Die grundlegende Idee, Residuen für die Modellüberprüfung einzusetzen, wurde ursprünglich für einfache Situationen (beispielsweise klassische lineare Regressionsmodelle) entwickelt. In solchen Fällen war die Definition der Residuen wenig problematisch, weil beispielsweise die *Pearson residuals* befriedigende Eigenschaften besaßen. Mit der zunehmenden Vielfalt stochastischer Modelle allerdings sind höhere Anforderungen an ein Residuenkonzept verknüpft: Alle Residuen sollten (bei Gültigkeit des Modells) auch für kompliziertere Modelle identische Verteilungseigenschaften besitzen. Nur dann ist es möglich, die Residuen gemeinsam zu analysieren (beispielsweise mit Hilfe von QQ-Plots) bzw. sie miteinander zu vergleichen. Die Pseudo-Residuen erfüllen diese Anforderung zumindest approximativ, wobei der Begriff ‚approximativ‘ an dieser Stelle verwendet wird, um anzudeuten, dass der Einfluss der Parameterschätzung noch keine Berücksichtigung fand. (Dies wird in Abschnitt 3.4 nachgeholt.)

Im Rahmen der Beschreibung der drei grundlegenden Ideen zur Definition von Residuen wurde deutlich, dass sich die Pseudo-Residuen grundsätzlich von den üblicher Weise verwendeten Residuen unterscheiden: Sie basieren nicht auf der Differenz zwischen einer Beobachtung und dem Erwartungswert der Beobachtung unter dem Modell. Vielmehr werden sie auf der Grundlage von Wahrscheinlichkeiten der Beobachtungen unter dem Modell berechnet. Daher sind mit den UPR explizite Wahrscheinlichkeitsaussagen verbunden, und die NPR messen eine Abweichung vom Median unter Berücksichtigung der Eigenschaften der Verteilung, unter der sie berechnet wurden. Beide Varianten der Pseudo-Residuen (UPR und NPR) sind bei der Modellüberprüfung hilfreich. Beispielsweise lässt sich die grundlegende Gestalt der zu überprüfenden Verteilungen gut mit einem Histogramm der UPR beurteilen. Die Anpassung an den Enden kann hingegen besser mit Hilfe der NPR beurteilt werden, da bei dieser Skalierung extreme Beobachtungen betont werden.

Die Verwendung von Residuen in Form von Intervallen bei der Modellierung diskreter Merkmale orientiert sich an der „Natur“ der Daten. Während bislang in der Literatur auch für den diskreten Fall in erster Linie Punktresiduen vorgeschlagen wurden, vermeidet das hier beschriebene Konzept den Verlust einer wichtigen Information: die Wahrscheinlichkeit der Beobachtung unter dem Modell. Es wird im Verlauf der Arbeit deutlich werden, dass dieser Gedanke in einigen Situationen hilfreich ist, um Fehlinterpretationen bei der Residuen-

analyse zu vermeiden, und dass auch für die Intervalle geeignete Methoden zur graphischen Analyse zur Verfügung stehen bzw. entwickelt werden können.

Kapitel 3

Überprüfung verallgemeinerter linearer Modelle

3.1 Einführung

In diesem Kapitel wird mit den verallgemeinerten linearen Modellen eine Modellklasse betrachtet, die eine Reihe von linearen Modellen für bestimmte Problemstellungen als Spezialfälle enthält. Verallgemeinerte lineare Modelle umfassen beispielsweise einfache lineare Regressionsmodelle, logistische Regressionsmodelle für binäre Zielvariablen oder loglineare Modelle für Zähldaten. Der erste Schritt in Richtung einer Zusammenfassung dieser und weiterer Modelle erfolgte durch NELDER AND WEDDERBURN (1972), die zeigten, dass die meisten der gewöhnlich verwendeten Regressionsmodelle zu einer Modellfamilie gehören und dass die Maximum-Likelihood-Schätzer für alle Modelle mit demselben Algorithmus (mit der Methode der gewichteten kleinsten Quadrate) bestimmt werden können (vgl. LINDSEY, 1997, S. 5).

Inzwischen existiert eine Vielzahl von Arbeiten, in denen verallgemeinerte lineare Modelle betrachtet werden. Die grundlegende Monographie wurde 1983 von MCCULLAGH AND NELDER publiziert und sechs Jahre später (1989) in einer zweiten Auflage wesentlich erweitert. DOBSON (1990) veröffentlichte eine Einführung in das Thema, die im Jahr 2002 ebenfalls in der zweiten Auflage erschienen ist. Bis in die heutige Zeit erscheinen weitere Monographien, die sich mit speziellen Aspekten der verallgemeinerten linearen Modelle beschäftigen. Erwähnt sei hier die anwendungsorientierte Darstellung der verallgemeinerten linearen Modelle von MYERS, MONTEGOMERY AND VINING (2002). Die hier genannten Monographien stellen lediglich eine Auswahl dar. Eine umfassenden-

dere Liste der relevanten Bücher stellt SMYTH im Internet unter der Adresse <http://www.isd.sdu.dk/~gks/glm/books.html> zur Verfügung.

Die Verallgemeinerung der verschiedenen Modelle zu einer Modellklasse führte auch zu neuen Anforderungen an die Überprüfung der Modelle. Für die *klassischen* linearen Regressionsmodelle wurde bereits Anfang der 1960-er Jahre die Basis der Residuenanalyse geschaffen (vgl. beispielsweise ANSCOMBE, 1961; ANSCOMBE AND TUKEY, 1963), die für diese Modelle spätestens seit der umfassenden Bearbeitung des Themas durch COOK AND WEISBERG (1982) ein etablierter Bestandteil der Modellbildung ist.

Die im Rahmen der klassischen linearen Regressionsmodelle verwendeten Residuen (die *ordinary residuals* sowie verschiedene Standardisierungen derselben; vgl. COOK AND WEISBERG, 1982, Kap. 2) sind jedoch nur bedingt zur Überprüfung der übrigen Modelle aus der Klasse der verallgemeinerten linearen Modelle geeignet: „*For generalized linear models we require an extended definition of residuals, applicable to all the distributions that may replace the Normal*“ (MCCULLAGH AND NELDER, 1989, S. 37). Die ‚erweiterten Definitionen‘, die bei MCCULLAGH AND NELDER (1989, S. 37–40) beschrieben werden, sind die *Pearson residuals*, die *Anscombe residuals* und die *deviance residuals*. Dass zumindest die *Pearson residuals* wenig geeignet sind, wenn an die Stelle der Normalverteilung eine Gammaverteilung tritt, wurde bereits in Beispiel 2.1 gezeigt. Die *Anscombe residuals*, deren Idee lange vor Einführung der verallgemeinerten linearen Modelle existierte (vgl. ANSCOMBE, 1953), und die *deviance residuals*, die zielgerichtet zur Überprüfung verallgemeinerter linearer Modelle definiert wurden (vgl. PREGIBON, 1981), unterscheiden sich (wie bereits erwähnt) nur unwesentlich. Nach Ansicht von PIERCE AND SCHAFER (1986), die diese Residuen vergleichen, ist es „*a matter of taste and computational convenience*“, welcher Definition der Vorzug gegeben wird.

Neben den von MCCULLAGH AND NELDER genannten Residuen, werden in der Literatur auch die NPR für die Überprüfung verallgemeinerter linearer Modelle vorgeschlagen. DUNN AND SMYTH (1996) bezeichnen die NPR als *quantile residuals* und empfehlen bei der Modellierung diskreter Merkmale die Verwendung der *randomized quantile residuals* $r^{\text{qu}} = \Phi^{-1}(U^*)$, wobei U^* eine Zufallszahl aus dem Intervall $[u^-, u^+]$ darstellt. Die Anwendung der *normal deviate residuals* im Rahmen der Cox-Regression (vgl. NARDI AND SCHEMPER, 1999) ist ein weiteres Beispiel für den Gebrauch der NPR im Zusammenhang mit verallgemeinerten linearen Modellen.

Für die hier erwähnten Residuen existiert eine Reihe von Transformationen

für spezielle Anwendungen, beispielsweise in Form von Stetigkeitskorrekturen bei diskreten Zielvariablen oder Varianzstabilisierende Transformationen zur Berücksichtigung des Einflusses einzelner Beobachtungen auf die Parameterschätzung. Auf die Erläuterung weiterer Details wird an dieser Stelle verzichtet, und es sei verwiesen auf MCCULLAGH AND NELDER (1989, Kap. 12), LINDSEY (1997, Anhang B) sowie auf DAVISON AND TSAI (1992).

Der Rest des Kapitels ist wie folgt gegliedert. Im folgenden Abschnitt wird die Grundform des verallgemeinerten linearen Modells definiert und die Berechnung der entsprechenden Pseudo-Residuen erläutert. Die weiteren Abschnitte behandeln spezielle Problemstellungen der Modellüberprüfung. Zunächst werden einige graphische Methoden zur Analyse der Residuen erläutert (Abschnitt 3.3). In Abschnitt 3.4 wird die Verteilung der Pseudo-Residuen unter Berücksichtigung der Parameterschätzung betrachtet. Im letzten Teil des Kapitels wird die Residuenanalyse für speziellere Modelle aus der Klasse der verallgemeinerten linearen Modelle erörtert und an Beispielen verdeutlicht.

3.2 Grundform des verallgemeinerten linearen Modells

Die Grundform des verallgemeinerten linearen Modells

$$g(E(\mathbf{Y})) = \eta \tag{3.1}$$

wird in der Regel als Komposition von drei Komponenten verstanden: Den Zielvariablen des Modells, der linearen Erklärungsfunktion (incl. der Modellparameter) und der Linkfunktion (vgl. LINDSEY, 1997, S. 18).

Die Zielvariablen $\mathbf{Y} = (Y_1, \dots, Y_n)'$ sind unabhängig verteilt und besitzen eine Verteilung aus der Exponentialfamilie mit Erwartungswert $E(\mathbf{Y}) = (E(Y_1), \dots, E(Y_n))'$. (Die Erwartungswerte der einzelnen Variablen werden im Folgenden mit μ_i bezeichnet.) Die Verteilung einer Zufallsvariablen Y , die von einem Parameter θ abhängt, gehört zur Exponentialfamilie, wenn ihre Dichtefunktion durch

$$f_Y(y) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}$$

darstellbar ist, wobei a , b , c und d bekannte Funktionen darstellen. Der Parameter θ ist der sogenannte ‚interessierende Parameter‘. Hängt die Verteilung von weiteren Parametern ab, so werden diese als ‚Störparameter‘, d.h.

als bekannter Teil der Funktionen a , b , c und/oder d aufgefasst. (Ein solcher Störparameter ist beispielsweise der Parameter σ^2 einer Normalverteilung.)

Die lineare Erklärungsfunktion η setzt sich aus der Designmatrix \mathbf{X} und dem Parametervektor $\boldsymbol{\beta}$ zusammen:

$$\eta = \mathbf{X}\boldsymbol{\beta},$$

wobei die $(n \times p)$ -Designmatrix

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

die Beobachtungen der erklärenden Variablen enthält (x_{ij} ist die i -te Beobachtung der Variablen X_j) und $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})'$ den Vektor der unbekannt Parameter beschreibt. Besitzt das Modell einen Achsenabschnitt, so besteht die erste Spalte der Designmatrix aus Einsen und der Parameter β_0 repräsentiert den Achsenabschnitt.

Die Linkfunktion g (monoton und differenzierbar) spezifiziert den Zusammenhang zwischen der linearen Erklärungsfunktion und dem Erwartungswert der Zielvariablen. Durch die Verwendung einer geeigneten Linkfunktion wird gewährleistet, dass die Schätzer $\hat{\mu}_i$ in einem zulässigen Bereich liegen, der vom Verteilungstyp der Zielvariablen abhängt. Besitzt die Zielvariable beispielsweise eine Poissonverteilung, so muss $\mu_i > 0$ für alle i gelten. Die Verwendung des Log-Links

$$g(\mu_i) = \log(\mu_i)$$

gewährleistet in diesem Fall, dass

$$\hat{\mu}_i = g^{-1}(\hat{\eta}_i) = \exp\{\hat{\eta}_i\}$$

für beliebige Werte der linearen Erklärungsfunktion größer ist als null. In anderen Situationen erhöht die Linkfunktion die Flexibilität bei der Modellbildung. Soll beispielsweise ein einfacher Wachstumsprozess durch ein exponentielles Wachstum $\mu_i = \alpha e^{\beta i}$ beschrieben werden (vgl. LINDSEY, 1997, S. 70), so ist die Erklärungsfunktion nicht linear. Hier führt die Verwendung des Log-Links zu $\log(\mu_i) = \log(\alpha) + \beta i$ und somit zu einem Modell mit linearer Erklärungsfunktion.

Die Parameterschätzung im verallgemeinerten linearen Modell erfolgt mit der Methode der iterativ gewichteten kleinsten Quadrate. Die Methode basiert auf

der iterativen Lösung des Gleichungssystems

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{z}, \quad (3.2)$$

wobei \mathbf{W} eine Diagonalmatrix mit

$$w_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

und $z_i = \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}$

die Elemente des Vektors \mathbf{z} bilden. Die Iteration beginnt mit der Berechnung von \mathbf{W} und \mathbf{z} unter Verwendung eines geeigneten Startwerts für $\hat{\boldsymbol{\beta}}^{(1)}$ (beispielsweise dem gewöhnlichen Kleinsten-Quadrate-Schätzer). Anschließend wird nach Gleichung 3.2 ein neuer Wert des Schätzers $\hat{\boldsymbol{\beta}}^{(2)}$ bestimmt, der für eine zweite Berechnung von \mathbf{W} und \mathbf{z} verwendet wird. Mit den neuen Werten für \mathbf{W} und \mathbf{z} kann $\hat{\boldsymbol{\beta}}^{(3)}$ berechnet werden usw. Die Iteration wird beendet, wenn die Differenz zwischen den sukzessive berechneten Werten von $\hat{\boldsymbol{\beta}}$ hinreichend klein ist. CHARNES ET AL. (1976) zeigten, dass die im Verlauf der Iterationen berechneten Schätzer gegen den Maximum-Likelihood-Schätzer konvergieren. Für eine umfassendere Beschreibung der Parameterschätzung sei auf DOBSON (2002, Kap. 4) verwiesen. Im Rahmen dieser Arbeit wird die Parameterschätzung mit der **R**-Funktion `glm` durchgeführt.

Zur Berechnung der Pseudo-Residuen werden die Verteilungsfunktionen der Zufallsvariablen Y_i benötigt. Der Verteilungstyp ist bei den verallgemeinerten linearen Modellen ein Teil der Modellformulierung und die (geschätzten) Parameter der Verteilungen ergeben sich aus dem angepassten verallgemeinerten linearen Modell.

Beispiel 3.1 Zur Verdeutlichung der Modelldefinition und der Berechnung der Pseudo-Residuen wird ein einfaches Beispiel betrachtet, in dem die Anzahl der Ausfälle eines Maschinenbauteils je Woche (Y) durch die Betriebszeiten der Maschine in zwei verschiedenen Betriebsarten (X_1 und X_2) erklärt wird (vgl. JORGENSON, 1969). Die Beobachtungen, die für einen Zeitraum von neun Wochen vorliegen, sind in Tabelle 3.1 dargestellt.

Im hier betrachteten Modell wird von einer Poissonverteilung der Zielvariablen mit

$$\log \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_9 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{91} & x_{92} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

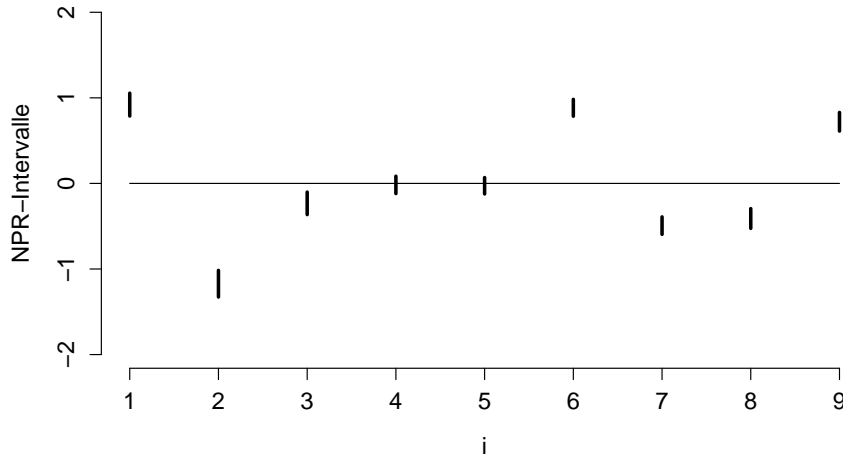


Abbildung 3.1: Indexplot der NPR-Intervalle.

ausgegangen, d.h. die Erwartungswerte sind über einen Log-Link mit der linearen Erklärungsfunktion verknüpft.

Der Parameterschätzer $\hat{\beta} = (2.18, 7.02 \cdot 10^{-3}, 2.55 \cdot 10^{-3})'$ des verallgemeinerten linearen Modells führt zu den in Tabelle 3.1 angegebenen Schätzern $\hat{\mu}_i$. Durch deren Verwendung zur Spezifikation der Verteilungsfunktionen der Poissonverteilungen erhält man schließlich die UPR-Intervalle

$$u_i = [u_i^-, u_i^+] = \left[\sum_{y=0}^{y_i-1} \frac{\hat{\mu}_i^y e^{-\hat{\mu}_i}}{y!}, \sum_{y=0}^{y_i} \frac{\hat{\mu}_i^y e^{-\hat{\mu}_i}}{y!} \right] \quad \text{für } i = 1, \dots, 9$$

Anzahl der Ausfälle	15	9	14	24	27
Betriebsart 1	33.3	52.2	64.7	137.0	125.9
Betriebsart 2	25.3	14.4	32.5	20.5	97.6
$\hat{\mu}_i$	11.86	13.17	15.06	24.25	27.31
Anzahl der Ausfälle	27	23	18	22	
Betriebsart 1	116.3	131.7	85.0	91.9	
Betriebsart 2	53.6	56.6	87.3	47.8	
$\hat{\mu}_i$	22.82	25.62	19.97	18.95	

Tabelle 3.1: Anzahl der Maschinenausfälle, Betriebszeiten in zwei verschiedenen Betriebsarten sowie Schätzer der Erwartungswerte.

und die NPR-Intervalle

$$r_i = [\Phi^{-1}(u_i^-), \Phi^{-1}(u_i^+)] \quad \text{für } i = 1, \dots, 9.$$

Ein Indexplot der NPR-Intervalle ist in Abbildung 3.1 dargestellt. Der Index i beschreibt hier aufeinanderfolgenden Wochen, d.h. er besitzt einen Zeitbezug. Systematische Strukturen im Indexplot wären also ein Hinweis darauf, dass die Anzahl der Ausfälle zeitlichen Veränderungen unterliegt. Im hier betrachteten Beispiel scheint dies allerdings nicht der Fall zu sein.



3.3 Graphische Analyse der Residuen

Anhand der einführenden Beispiele aus dem letzten Kapitel wurde bereits deutlich, dass die graphische Analyse der Residuen ein hilfreiches Instrument der Modellüberprüfung ist. CHAMBERS ET AL. (1983, S. 1) bringen die Ideen bzw. die Vorteile der graphischen Datenanalyse auf den Punkt:

„Our eye-brain system is the most sophisticated information processor ever developed, and through graphical displays we can put this system to good use to obtain deep insight into the structure of data.“

In der Literatur ist eine Vielzahl graphischer Methoden zur Überprüfung von verallgemeinerten linearen Modellen mit stetigen Zielvariablen zu finden. WANG (1987) beschreibt mit dem *constructed-variable plot* eine Möglichkeit, die Linearität für eine Teilmenge der erklärenden Variablen zu überprüfen. Seine Darstellung basiert auf dem *added-variable plot* (WANG, 1985), der zur Beurteilung der Aufnahme weiterer erklärender Variablen in das Modell verwendet wird. Diese Problematik wird auch von COOK AND WEISBERG (1994) behandelt. Sie betrachten einen „dynamischen“ Plot, in dem die geschätzten Erwartungswerte der Zielvariablen $\hat{\mu}_i(\lambda)$ gegen die Residuen $r_i(\lambda)$ geplottet werden, wobei $\lambda \in [0, 1]$ ein „control parameter“ ist: Ausgehend von $\lambda = 0$ werden zunächst die Residuen eines gegebenen Modells dargestellt und durch schrittweise Erhöhung von λ (bis $\lambda=1$) schließlich die Residuen des Modells mit einer weiteren erklärenden Variable. Ihr Vorgehen zeigt das Verhalten der Residuen bei einer „weichen“ Aufnahme einer erklärenden Variable, was die Bezeichnung *Adding REgressors Smoothly (ARES)*-Plots erklärt. In einer jüngeren Arbeit diskutieren COOK AND CROOS-DABRERA (1998) die Verwendung

von *partial residual plots*, die Hinweise auf notwendige Transformationen erklärender Variablen geben. Für weitere Möglichkeiten der graphischen Analyse von Residuen sowie entsprechende Literaturhinweise sei auf MCCULLAGH AND NELDER (1989, Kap. 12) verwiesen.

Die Art der verwendeten Residuen ist bei den genannten Darstellungen austauschbar. Für die *ARES*-Plots zum Beispiel wird die Verwendung von G^2 -Residuen (vgl. WILLIAMS, 1987) vorgeschlagen. Eine analoge Analyse könnte aber auch mit Hilfe der NPR durchgeführt werden. Es ist nicht die primäre Absicht, in diesem Abschnitt existierende graphische Darstellungen auf die Pseudo-Residuen zu übertragen. Vielmehr liegt der Schwerpunkt in der Analyse von Graphiken *mit Hilfe* von Pseudo-Residuen und der Entwicklung geeigneter Methoden zur Analyse der Pseudo-Residuen bei der Modellierung diskreter Zielvariablen: Im folgenden Abschnitt wird dargestellt, wie Pseudo-Residuen bei der graphischen Analyse von QQ-Plots verwendet werden können, um die Struktur der Plots objektiv zu beurteilen. Anschließend (Abschnitt 3.3.2) werden drei Möglichkeiten zur graphischen Analyse der diskreten Pseudo-Residuen erläutert.

3.3.1 Analyse von QQ-Plots mit Pseudo-Residuen

QQ-Plots sind ein sehr nützliches Instrument für den Vergleich zweier Verteilungen. Im Besonderen kann mit ihnen visuell beurteilt werden, ob eine Stichprobe durch eine bestimmte Verteilung (also beispielsweise durch eine Standardnormalverteilung) gut beschrieben wird. Dazu werden die Quantile der unterstellten Verteilung mit Verteilungsfunktion F

$$Q(p_k) = F^{-1}(p_k) \quad \text{für } k = 1, \dots, n$$

gegen die entsprechenden Quantile der Stichprobe

$$\hat{Q}(p_k) = x_{(k)} \quad \text{für } k = 1, \dots, n$$

geplottet, wobei im Rahmen dieser Arbeit $p_k = k/(n + 1)$ gewählt wird und $x_{(1)} \leq \dots \leq x_{(k)} \leq \dots \leq x_{(n)}$ die Orderstatistiken der Stichprobe (im Folgenden als empirische Quantile bezeichnet) darstellen. (Zu einer Diskussion alternativer Definitionen von p_k siehe KIMBALL, 1960.)

Liegen die Punkte $(Q(p_k), \hat{Q}(p_k))$ für $k = 1, \dots, n$ in etwa auf der Winkelhalbierenden, ist dies ein Hinweis darauf, dass die Stichprobe gut von der unterstellten Verteilung beschrieben wird. Abweichungen hingegen deuten auf

Unterschiede zwischen der empirischen Verteilung und der unterstellten Verteilung hin. In diesem Fall können mit Hilfe des QQ-Plots unterschiedliche Lage und/oder Streuung, unterschiedliche Verteilungsenden, Ausreißer sowie Regionen schlechter Anpassung identifiziert werden (vgl. SIEVERS, 1986).

Es ist offensichtlich, dass die QQ-Plots im Rahmen dieser Arbeit eingesetzt werden, um zu überprüfen, ob die NPR aus einer Standardnormalverteilung stammen. So wurde für das Beispiel zu den Bäumen aus in Abschnitt 2.5.1 bereits ein QQ-Plot für die NPR, basierend auf den Punkten

$$(\Phi^{-1}(p_k), r_{(k)}) \quad \text{für } k = 1, \dots, n,$$

untersucht, um auf diesem Weg die Güte der Modellanpassung zu beurteilen.

Obwohl die QQ-Plots in erster Linie eine explorative Methode darstellen, ist eine objektive Beurteilung der Abweichungen von der Winkelhalbierenden wünschenswert. Ob diese systematisch oder zufällig sind, kann mit Hilfe der Verteilungen der Orderstatistiken der UPR beurteilt werden: Seien $u_{(k)}$ die empirischen Quantile einer Stichprobe vom Umfang n aus einer $U(0, 1)$ -Verteilung. Es ist bekannt, dass die $U_{(k)}$ in diesem Fall eine Betaverteilung mit den Parametern $\alpha = k$ und $\beta = n - k + 1$ besitzen (vgl. beispielsweise KENDALL AND STUART, 1977, S. 347). Wenn ein zu überprüfendes Modell eine *korrekte* Beschreibung der Daten liefert, gilt daher für die Orderstatistiken der UPR

$$U_{(k)} \sim B(k, n - k + 1) \quad \text{für } k = 1, \dots, n.$$

Auf der Basis dieser Verteilungen können die Pseudo-Residuen

$$\begin{aligned} U_{(k)}^{\text{qq}} &= F_{\text{Beta}}(U_{(k)}; k, n - k + 1) \\ \text{bzw. } R_{(k)}^{\text{qq}} &= \Phi^{-1}(U_{(k)}^{\text{qq}}) \end{aligned}$$

der empirischen Orderstatistiken der UPR definiert werden. $U_{(k)}^{\text{qq}}$ und $R_{(k)}^{\text{qq}}$ sind als Pseudo-Residuen eines QQ-Plots zu interpretieren, da auf der Ordinate des QQ-Plots gerade die empirischen Quantile der Pseudo-Residuen abgetragen werden. Sie können somit verwendet werden, um die Abweichung eines Punktes $(\Phi^{-1}(p_k), r_{(k)})$ von der Winkelhalbierenden zu quantifizieren.

Beispiel 3.2 Um die Analyse der QQ-Plots mit Pseudo-Residuen zu veranschaulichen, wird ein Experiment zur Messung des Widerstandes von Halbleitern betrachtet (vgl. MYERS AND MONTEGOMERY, 1997). Tabelle 3.2 enthält neben den Beobachtungen für die Zielvariable (dem Widerstand) die Ausprägungen der berücksichtigten binären Faktoren. (Letztere werden bei MYERS AND MONTEGOMERY (1997) nicht näher spezifiziert.)

i	x_1	x_2	x_3	x_4	y_i	$\hat{\mu}_i$	i	x_1	x_2	x_3	x_4	y_i	$\hat{\mu}_i$
1	-	-	-	-	193.4	202.75	9	-	-	-	+	220.0	210.05
2	+	-	-	-	247.6	247.71	10	+	-	-	+	256.4	256.62
3	-	+	-	-	168.2	164.18	11	-	+	-	+	165.7	170.09
4	+	+	-	-	205.0	200.58	12	+	+	-	+	203.5	207.81
5	-	-	+	-	303.4	313.78	13	-	-	+	+	285.0	271.00
6	+	-	+	-	339.9	328.15	14	+	-	+	+	268.0	283.41
7	-	+	+	-	226.3	213.03	15	-	+	+	+	169.1	183.98
8	+	+	+	-	208.3	222.79	16	+	+	+	+	208.5	192.41

Tabelle 3.2: Beobachtungen und Schätzer für den Widerstand des Halbleiters sowie Ausprägungen der Faktoren.

Zur Beschreibung der Daten verwenden MYERS, MONTEGOMERY AND VINING (2002, S. 176–183) ein verallgemeinertes lineares Modell mit gammaverteilten Zielvariablen und Log-Link, das neben den Haupteffekten ausgewählte Wechselwirkungen berücksichtigt:

$$\begin{aligned} \log(\mu_i) = & \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} \\ & + \beta_5 x_{1i} x_{3i} + \beta_6 x_{2i} x_{3i} + \beta_7 x_{3i} x_{4i} \quad \text{für } i = 1, \dots, 16. \end{aligned}$$

Codiert man die Faktoren in der Form $x_{ij} = 1$ für ‚+‘ bzw. $x_{ij} = -1$ für ‚-‘, erhält man die in Tabelle 3.3 angegebenen Parameterschätzer des Modells. Diese liefern schließlich die Schätzer der Erwartungswerte $\hat{\mu}_i$ (siehe Tabelle 3.2). Die Schätzer für die Varianz der Y_i ergeben sich im verallgemeinerten linearen Modell aus dem Produkt der Varianzfunktion $V(\mu_i)$ und dem Dispersionsparameter ϕ . Bei gammaverteilten Zielvariablen gilt $V(\mu_i) = \mu_i^2$ und $\phi = \nu^{-1}$ (vgl. McCullagh and Nelder, 1989, S. 30). Man erhält also in diesem Fall $Var(Y_i) = \mu_i^2 \nu^{-1}$. Mit diesem Ergebnis bestimmen sich die Parameter ν und λ einer Gammaverteilung, deren Dichtefunktion in der Form

$$\begin{aligned} f_Y(y) &= \frac{\lambda^\nu y^{\nu-1} e^{-\lambda y}}{\Gamma(\nu)} \\ \text{mit } E(Y) &= \nu/\lambda \quad \text{und} \quad Var(Y) = \nu/\lambda^2 \end{aligned}$$

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
5.41	0.0613	-0.150	0.0899	-0.0278	-0.0389	-0.0441	-0.0455

Tabelle 3.3: Parameterschätzer des verallgemeinerten linearen Modells.

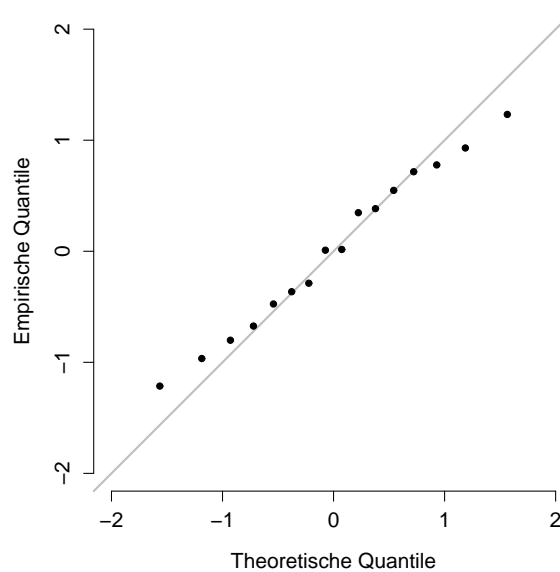


Abbildung 3.2: QQ-Plot der NPR.

parametrisiert ist durch

$$\nu = \phi^{-1} \quad \text{und} \quad \lambda_i = (\mu\phi)^{-1}.$$

Im hier betrachteten Beispiel ist der Schätzer des Dispersionsparameters $\hat{\phi} = 4.52 \cdot 10^{-3}$, der zusammen mit den Schätzern $\hat{\mu}_i$ verwendet wird, um die Verteilungen der Zielvariablen vollständig zu spezifizieren.

Der QQ-Plot der NPR ist in Abbildung 3.2 dargestellt. Die Struktur der Punkte scheint (relativ zur Winkelhalbierenden) leicht im Uhrzeigersinn gedreht. Dies bedeutet, dass die Residuen eine zu geringe Streuung besitzen. Mit anderen Worten wird bei Gültigkeit des Modells eine größere Streuung der Beobachtungen erwartet. Besonders deutlich wird die Struktur des QQ-Plots bei einer Betrachtung der NPR des QQ-Plots (Abbildung 3.3). Der klar zu erkennende negative Trend nährt die Vermutung, dass das Modell nicht adäquat ist. Es muss hier allerdings erwähnt werden, dass die Nichtberücksichtigung der Parameterschätzung in diesem Beispiel eine Rolle spielt, weil die Anzahl der geschätzten Parameter (acht) relativ groß ist im Vergleich zu der Anzahl der Beobachtungen (sechzehn). Es sei an dieser Stelle nur erwähnt, dass die Verteilung der Residuen bei einer Korrektur um den Einfluss der Parameterschätzung breiter ist, so dass die hier dargestellten Ergebnisse mit Vorsicht

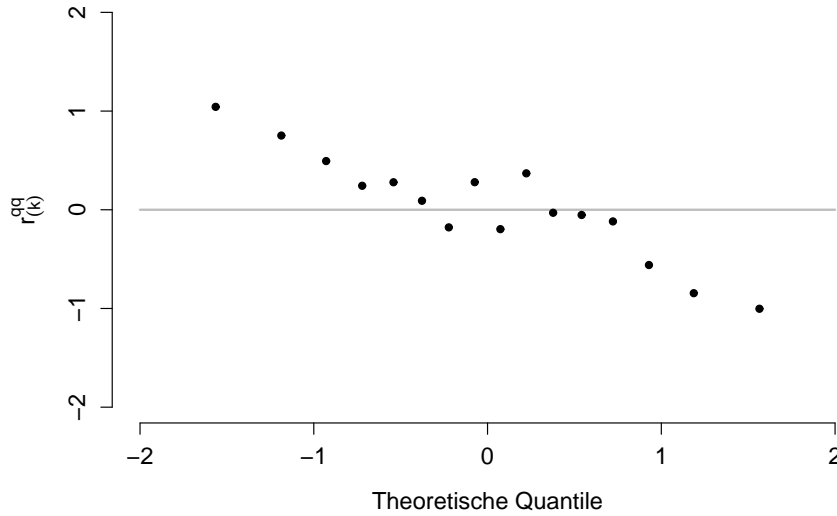


Abbildung 3.3: NPR des QQ-Plots.

zu interpretieren sind. Genauer wird diese Problemstellung (auch für dieses Beispiel) in Abschnitt 3.4 betrachtet.



Neben einer Verwendung der Pseudo-Residuen des QQ-Plots zur graphischen Analyse können diese auch verwendet werden, um simultane Konfidenzbänder für einen QQ-Plot zu konstruieren. Diese Konfidenzbänder grenzen den Bereich ein, der mit vorgegebener Wahrscheinlichkeit $1 - \alpha$ nicht von den Punkten des QQ-Plots verlassen wird. Seien

$$Z_{\min} = \min \left\{ U_{(1)}^{\text{qq}}, \dots, U_{(n)}^{\text{qq}} \right\}$$

bzw.

$$Z_{\max} = \max \left\{ U_{(1)}^{\text{qq}}, \dots, U_{(n)}^{\text{qq}} \right\}$$

das Minimum bzw. Maximum der UPR eines QQ-Plots. Bei Gültigkeit des zu überprüfenden Modells besitzen Z_{\min} bzw. Z_{\max} die Verteilung des Minimums bzw. Maximums von *nicht unabhängig* $U(0, 1)$ -verteilten Zufallsvariablen. Für gegebenes n können diese Verteilungen simuliert werden, indem hinreichend oft n unabhängig und identisch $U(0, 1)$ -verteilte Zufallszahlen erzeugt werden und aus den so erzeugten Daten die Verteilungen von Z_{\min} und Z_{\max} approximiert werden. Mit diesen Verteilungen können in einem nächsten Schritt die Werte

p^- und p^+ geschätzt werden, für die

$$\begin{aligned} P\{Z_{\min} < p^-\} &= \alpha/2 \\ \text{bzw. } P\{Z_{\max} > p^+\} &= \alpha/2 \end{aligned}$$

gilt. Da Z_{\min} bzw. Z_{\max} das Minimum bzw. Maximum der UPR des QQ-Plots beschreiben, ist die Wahrscheinlichkeit, dass alle UPR eines QQ-Plots im Intervall $[p^-, p^+]$ liegen annähernd $1 - \alpha$:

$$\begin{aligned} &P\left\{U_{(1)}^{\text{qq}}, \dots, U_{(n)}^{\text{qq}} \in [p^-, p^+]\right\} \\ &= 1 - (P\{Z_{\min} < p^- \text{ oder } Z_{\max} > p^+\}) \\ &= 1 - (P\{Z_{\min} < p^-\} + P\{Z_{\max} > p^+\} - P\{Z_{\min} < p^- \text{ und } Z_{\max} > p^+\}) \\ &\approx 1 - (P\{Z_{\min} < p^-\} + P\{Z_{\max} > p^+\}) \\ &= 1 - \alpha. \end{aligned}$$

Die Bestimmung der Intervallgrenzen p^- und p^+ basiert auf den simulierten Verteilungen von Z_{\min} und Z_{\max} . Um den Rechenaufwand zu verringern, kann $p^+ = 1 - p^-$ verwendet werden, d.h. auf die Simulation der Verteilung von Z_{\max} kann verzichtet werden. Dies ergibt sich aus

$$\begin{aligned} P\{Z_{\min} < p^-\} &= P\left\{\min\{U_{(1)}^{\text{qq}}, \dots, U_{(n)}^{\text{qq}}\} < p^-\right\} \\ &= P\left\{1 - \min\{U_{(1)}^{\text{qq}}, \dots, U_{(n)}^{\text{qq}}\} > 1 - p^-\right\} \\ &= P\left\{\max\{U_{(1)}^{\text{qq}}, \dots, U_{(n)}^{\text{qq}}\} > 1 - p^-\right\} \\ &= P\{Z_{\max} > 1 - p^-\}. \end{aligned}$$

Nach der Bestimmung von p^- und p^+ können die simultanen Konfidenzbänder für einen QQ-Plot der NPR wie folgt konstruiert werden: In dem QQ-Plot sind auf der Ordinate die Werte

$$\begin{aligned} r_{(k)} &= \Phi^{-1}(u_{(k)}) \\ &= \Phi^{-1}\left(F_{\text{Beta}}^{-1}(u_{(k)}^{\text{qq}}; k, n - k + 1)\right) \quad \text{für } k = 1, \dots, n \end{aligned}$$

abgetragen. Da es sich hierbei um monotone Transformationen der UPR des QQ-Plots handelt, formen die Punkte

$$\begin{aligned} &(\Phi^{-1}(p_k); \Phi^{-1}(F_{\text{Beta}}^{-1}(p^-; k, n - k + 1))) \quad \text{für } k = 1, \dots, n \\ \text{bzw. } &(\Phi^{-1}(p_k); \Phi^{-1}(F_{\text{Beta}}^{-1}(p^+; k, n - k + 1))) \quad \text{für } k = 1, \dots, n \end{aligned}$$

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5	0.01303	0.005769	$1.05 \cdot 10^{-3}$
10	0.00813	0.003702	$6.54 \cdot 10^{-4}$
20	0.00536	0.002447	$3.72 \cdot 10^{-4}$
50	0.00336	0.001480	$2.49 \cdot 10^{-4}$
100	0.00245	0.001090	$1.61 \cdot 10^{-4}$
200	0.00200	0.000857	$1.32 \cdot 10^{-4}$
500	0.00142	0.000599	$9.36 \cdot 10^{-5}$
1000	0.00123	0.000531	$1.09 \cdot 10^{-4}$
2000	0.00102	0.000465	$7.30 \cdot 10^{-5}$

Tabelle 3.4: Schätzer für p^- in Abhängigkeit von Stichprobenumfang und Signifikanzniveau.

die untere bzw. obere Grenze des gewünschten simultanen Konfidenzbereichs für den QQ-Plot.

Tabelle 3.4 zeigt einige Werte für p^- bei unterschiedlichen Stichprobengrößen und Signifikanzniveaus. Hier wurden jeweils 10 000 Simulationsläufe durchgeführt, um die Verteilung von Z_{\min} bzw. Z_{\max} zu approximieren. Das $\alpha/2$ -Quantil der angenäherten Verteilung von Z_{\min} bzw. das $1 - \alpha/2$ -Quantil der angenäherten Verteilung von Z_{\max} wurden dann verwendet, um p^- bzw. p^+ zu schätzen. Wie bereits erwähnt, gilt $p^- = 1 - p^+$. Daher sind die Ergebnisse der Schätzungen in Tabelle 3.4 zusammengefasst, indem für jede Kombination von n und α das arithmetische Mittel von \hat{p}^- und $1 - \hat{p}^+$ angegeben ist.

Um die Ergebnisse auf andere als die angegebenen Stichprobengrößen zu übertragen, kann die Regressionsgleichung

$$\log(p^-) = \beta_0 + \beta_1 \log(n) + \beta_2 (\log(n))^2$$

verwendet werden. Die Schätzer (bestimmt nach der Methode der kleinsten Quadrate) sind für die verschiedenen Werte von α in Tabelle 3.5 angegeben. Abbildung 3.4 zeigt, dass die angepassten Regressionsmodelle eine gute Beschreibung des Zusammenhangs zwischen n und p^- liefern.

Beispiel 3.3 Um die Verwendung der simultanen Konfidenzbänder zu illustrieren, werden die in Tabelle 3.6 wiedergegebenen Daten eines Beispiels von MONTGOMERY AND PECK (1992, S. 115–116) betrachtet. Zu erklären ist die Zeit Y , die ein Arbeiter eines Getränkeshändlers benötigt, um die bei einem

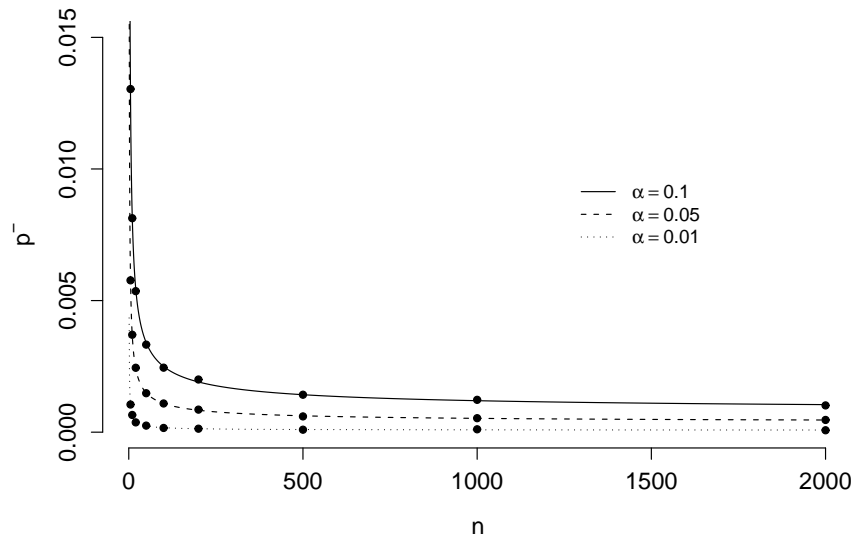


Abbildung 3.4: Angepasste Regressionsmodelle für den Zusammenhang zwischen n und p^- .

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
$\alpha = 0.10$	-3.17	-0.809	0.0425
$\alpha = 0.05$	-3.91	-0.839	0.0452
$\alpha = 0.01$	-5.44	-0.978	0.0600

Tabelle 3.5: Schätzer der Regressionsmodelle für den Zusammenhang zwischen n und p^- .

Kunden aufgestellten Getränkeautomaten zu füllen. Es wird davon ausgegangen, dass die Zeit durch die Anzahl der benötigten Getränkeboxen (X_1) und durch die Länge des zurückzulegenden Fußweges (X_2) erklärt werden kann.

In dem hier verwendeten Modell wird eine Gammaverteilung für die Zielvariable unterstellt und als Linkfunktion wird der Inverslink verwendet. Für das angepasste Modell (ohne Wechselwirkung und mit Achsenabschnitt) erhält man:

$$\frac{1}{E(\mathbf{Y})} = \mathbf{X} \begin{pmatrix} 7.13 \cdot 10^{-2} \\ -1.73 \cdot 10^{-3} \\ -6.43 \cdot 10^{-6} \end{pmatrix}.$$

Der Schätzer des Dispersionsparameters ist $\hat{\phi} = 0.068$.

Abbildung 3.5 zeigt den QQ-Plot der NPR und die simultanen Konfidenz-

i	y	x_1	x_2	i	y	x_1	x_2
1	16.68	7	560	14	19.75	6	462
2	11.50	3	220	15	24.00	9	448
3	12.03	3	340	16	29.00	10	776
4	14.88	4	80	17	15.35	6	200
5	13.75	6	150	18	19.00	7	132
6	18.11	7	330	19	9.50	3	36
7	8.00	2	110	20	35.10	17	770
8	17.83	7	210	21	17.90	10	140
9	79.24	30	1460	22	52.32	26	810
10	21.50	5	605	23	18.75	9	450
11	40.33	16	688	24	19.83	8	635
12	21.00	10	215	25	10.75	4	150
13	13.50	4	255				

Tabelle 3.6: Beobachtungen für die Zeit, die Anzahl benötigter Getränkeboxen und den zurückzulegenden Fußweg.

bänder für verschiedene Signifikanzniveaus. Es ist zu sehen, dass der QQ-Plot innerhalb der Konfidenzbänder verläuft. Versteht man die Darstellung als Signifikanztest, kann daher die Hypothese, dass die NPR standardnormalverteilt sind, nicht verworfen werden.



Die hier erläuterten Konfidenzbänder sind nur dann (asymptotisch) exakt, wenn die Parameter des Modells bekannt sind. Mit anderen Worten bleibt auch hier die Streuung der Schätzer unberücksichtigt. ROSENKRANTZ (2001) beschreibt ein Vorgehen zur Erzeugung simultaner Konfidenzbänder, das gerade an diesem Punkt ansetzt, d.h. die Konfidenzbänder basieren auf Konfidenzbereichen für die Parameter. Letztere können allerdings auf verschiedene Arten definiert werden. Wird beispielsweise ein simultaner Konfidenzbereich für die Parameter μ und σ^2 einer Normalverteilung definiert, so führen unterschiedliche Konfidenzbereiche für die Parameter zu unterschiedlichen Konfidenzbändern für den QQ-Plot. Darüber hinaus ist seine Methode auf die Überprüfung weniger Verteilungstypen beschränkt. Es sei an dieser Stelle erwähnt, dass in Abschnitt 3.4 eine Kalibrierung der Pseudo-Residuen erläutert wird, die auch bei unbekanntem (d.h. zu schätzenden) Modellparametern zu $N(0, 1)$ -verteilten NPR führt, so dass die hier vorgestellten Konfidenzbänder bei ge-

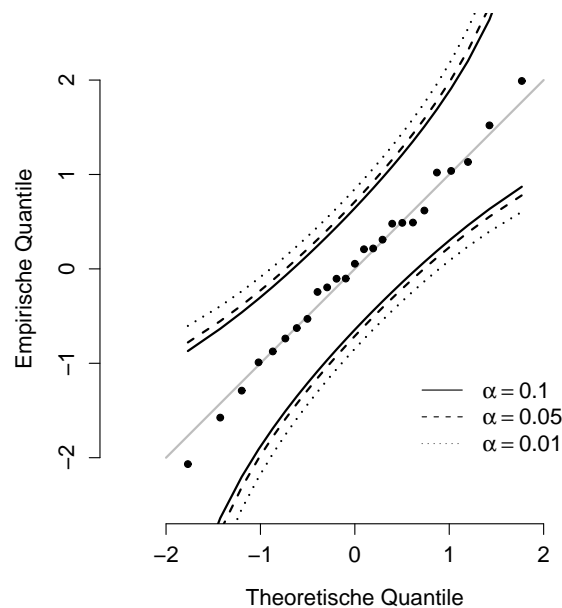


Abbildung 3.5: QQ-Plot der NPR mit simultanen Konfidenzbändern.

ringem Stichprobenumfang auf die kalibrierten Residuen angewendet werden sollten.

Eine auf der *Bootstrap*-Methode basierende Möglichkeit zur Konstruktion von simultanen Konfidenzbändern wird von DAVISON AND HINKLEY (1997, S. 150–154) beschrieben. Sie betrachten die Überprüfung von Daten auf eine $N(0, 1)$ -Verteilung und bestimmen simultane Konfidenzbänder für den entsprechenden QQ-Plot. Die hier vorgestellten Konfidenzbänder sind allgemeiner, da sie für beliebige QQ-Plots leicht zu bestimmen sind: Soll mit einem QQ-Plot überprüft werden, ob eine Stichprobe durch eine Verteilung mit Verteilungsfunktion F gut beschrieben wird, so besteht die untere bzw. obere Grenze des Konfidenzbands aus den Punkten

$$\begin{aligned} & (F^{-1}(p_k); F^{-1}(F_{\text{Beta}}^{-1}(p^-; k, n - k + 1))) \quad \text{für } k = 1, \dots, n \\ \text{bzw.} & (F^{-1}(p_k); F^{-1}(F_{\text{Beta}}^{-1}(p^+; k, n - k + 1))) \quad \text{für } k = 1, \dots, n. \end{aligned}$$

3.3.2 Graphische Residuenanalyse bei diskreten Zielvariablen

Die in der Literatur vorgeschlagenen Methoden zur graphischen Analyse von Residuen basieren in der Regel auf Punktresiduen. Da die Pseudo-Residuen bei

der Modellierung diskreter Merkmale als Intervalle definiert sind, erscheint es daher sinnvoll, die vorgeschlagenen Methoden zu erweitern bzw. entsprechende Methoden zu entwickeln. Mit dem Indexplot der NPR-Intervalle aus dem einführenden Beispiel dieses Abschnitts wurde bereits ein Standardverfahren bei der Analyse von Punktresiduen auf die NPR-Intervalle übertragen. Im Folgenden werden drei weitere graphische Darstellungen vorgestellt. Zum einen werden QQ-Plots für NPR-Intervalle und Histogramme für UPR-Intervalle beschrieben. Ferner wird mit den Dreieck-Plots (so die Bezeichnung in dieser Arbeit) eine Darstellung der UPR-Intervalle erläutert, die kein Gegenstück bei der Überprüfung von Modellen für stetige Zielvariablen besitzt.

3.3.2.1 QQ-Plots für NPR-Intervalle

QQ-Plots für stetige NPR werden konstruiert, indem die der Größe nach sortierten Beobachtungen gegen die entsprechenden Quantile der Standardnormalverteilung geplottet werden (eine genauere Beschreibung erfolgte bereits im vorangegangenen Abschnitt). Es ist naheliegend, den QQ-Plot für NPR-Intervalle analog zu konstruieren, indem die Intervalle sortiert werden und in Form von vertikalen Strecken gegen die Quantile der Standardnormalverteilung geplottet werden.

Seien $r_{(k)}^-$ und $r_{(k)}^+$ die Grenzen des k -ten Intervalls nach Sortierung der Intervalle. Dann werden in einem QQ-Plot für NPR-Intervalle die Strecken

$$S_k = \left(\Phi^{-1}(p_k), \overline{r_{(k)}^- r_{(k)}^+} \right) \quad \text{für } k = 1, \dots, n$$

dargestellt, wobei $p_k = k/(n+1)$ wie im stetigen Fall gewählt werden kann.

Bei der Konstruktion des Plots ist es erforderlich, ein Vorgehen zur Sortierung der Intervalle festzulegen. Während im stetigen Fall die Sortierung eindeutig ist, hat man bei der Sortierung von Intervallen mehrere Möglichkeiten. ZUCCHINI AND MACDONALD (1999), die eine Analyse der QQ-Plots für NPR-Intervalle bei loglinearen Modellen beschreiben, sortieren die Intervalle nach ihrem Mittelpunkt, der für die i -te Beobachtung durch

$$r_i^m = \Phi^{-1} \left(\frac{u_i^- + u_i^+}{2} \right)$$

auf Basis der UPR-Intervalle definiert ist. Dieses Vorgehen wird auch in dieser Arbeit verwendet.

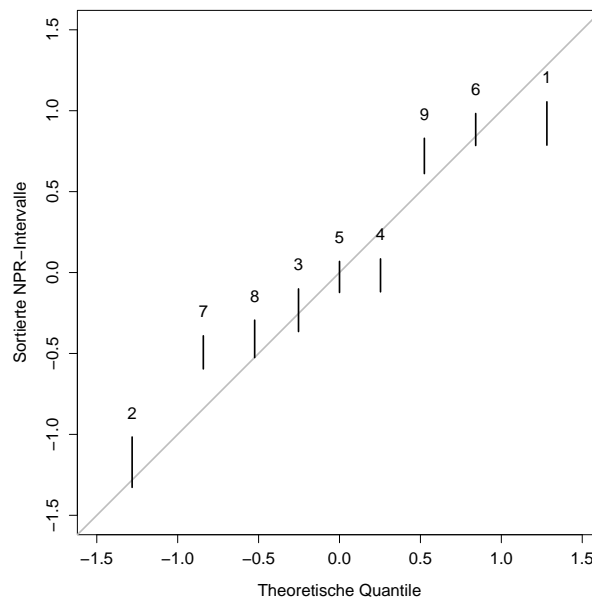


Abbildung 3.6: QQ-Plot der NPR-Intervalle.

Beispiel 3.4 Zur Illustration wird das einführende Beispiel zu den Maschinenausfällen erneut betrachtet. Die NPR-Intervalle des Beispiels sind bereits in einem Indexplot (Abbildung 3.2) dargestellt worden. Den entsprechenden QQ-Plot zeigt Abbildung 3.6, wobei die Segmente mit dem Index der entsprechenden Beobachtung gekennzeichnet sind.

Bei Gültigkeit des Modells sollten die Intervalle in etwa auf der Winkelhalbierenden liegen. Abweichungen von der Winkelhalbierenden hingegen sind als Hinweis auf ein fehlerhaft spezifiziertes Modell zu interpretieren. Im letzten Abschnitt wurde gezeigt, wie im stetigen Fall das Ausmaß der Abweichungen mit Hilfe der Pseudo-Residuen eines QQ-Plots beurteilt werden kann. Dieses Vorgehen ist nicht auf den diskreten Fall übertragbar, da die Verteilungen der Orderstatistiken (hier gleichbedeutend mit den sortierten Intervallen) nicht bekannt ist bzw. nicht im herkömmlichen Sinn in Form einer Dichte- oder Wahrscheinlichkeitsfunktion darstellbar ist. In Abbildung 3.6 weichen die Eintragungen für die vierte und siebte Beobachtung relativ deutlich von der Winkelhalbierenden ab. Es wird in einem späteren Beispiel dargestellt, wie das Ausmaß dieser Abweichungen mit Hilfe von Dreieck-Plots graphisch beurteilt werden kann.



3.3.2.2 Histogramme für UPR-Intervalle

Während sich die Konstruktion von QQ-Plots für NPR-Intervalle unmittelbar aus der stetigen Variante des QQ-Plots ergibt, erfordert die Konstruktion der Histogramme für UPR-Intervalle einige Modifikationen. Ziel ist es hier, eine graphische Darstellung für die UPR-Intervalle zu entwickeln, die bei Gültigkeit des zu überprüfenden Modells eine histogrammähnliche Darstellung einer $U(0,1)$ -Verteilung liefert. Abbildung 3.7 verdeutlicht die Umsetzung dieser Idee anhand von drei beispielhaft ausgewählten UPR-Intervallen:

$$u_1 = [0.20, 0.70], u_2 = [0.50, 0.80] \quad \text{und} \quad u_3 = [0.85, 0.95].$$

In der Darstellung ist jedem der drei Intervalle eine Graustufe zugeordnet. Das dunkelgraue Rechteck repräsentiert u_1 . Es ist horizontal durch die Intervallgrenzen von u_1 begrenzt und besitzt die Höhe

$$H_1 = \frac{1}{(u_1^+ - u_1^-)n} = \frac{1}{(0.7 - 0.2)3},$$

so dass die Fläche des Rechtecks $1/3$ beträgt. Analog ist das mittelgraue Rechteck im rechten Teil der Abbildung konstruiert. Es ist horizontal durch 0.85 und 0.95 begrenzt und seine Höhe beträgt $10/3$, d.h. auch hier ergibt sich eine Fläche von $1/3$. Die Darstellung von u_2 (hellgrauer Bereich) ist zwar horizontal ebenfalls durch die entsprechenden Intervallgrenzen bestimmt, jedoch existiert hier ein Bereich, der bereits von u_1 beansprucht wird. Für diesen Bereich zwischen 0.5 und 0.7 wird die Höhe von u_3 ($H_3 = (0.8 - 0.5)^{-1}/3$) zu der Höhe H_1 addiert. Für den Bereich zwischen 0.7 und 0.8, der noch nicht durch ein anderes UPR-Intervall beansprucht wird, beträgt die Höhe einfach $H_3 = (0.8 - 0.5)^{-1}/3$. Der hellgraue Bereich besitzt ebenfalls die Fläche $1/3$. Daher beträgt die Gesamtfläche der grau dargestellten Bereiche eins.

Durch diese Schichtung der UPR-Intervalle entsteht eine Umhüllende

$$H(u) = \frac{1}{n} \sum_{i=1}^n \frac{1}{u_i^+ - u_i^-} I\{u \in [u_i^-, u_i^+]\},$$

wobei $I\{A\}$ die Indikatorfunktion des Ereignisses A darstellt. Das erwartete Niveau der Umhüllenden (in der Abbildung fett eingezeichnet) beträgt bei Gültigkeit des Modells eins. Um dies zu verdeutlichen, sei angenommen, dass

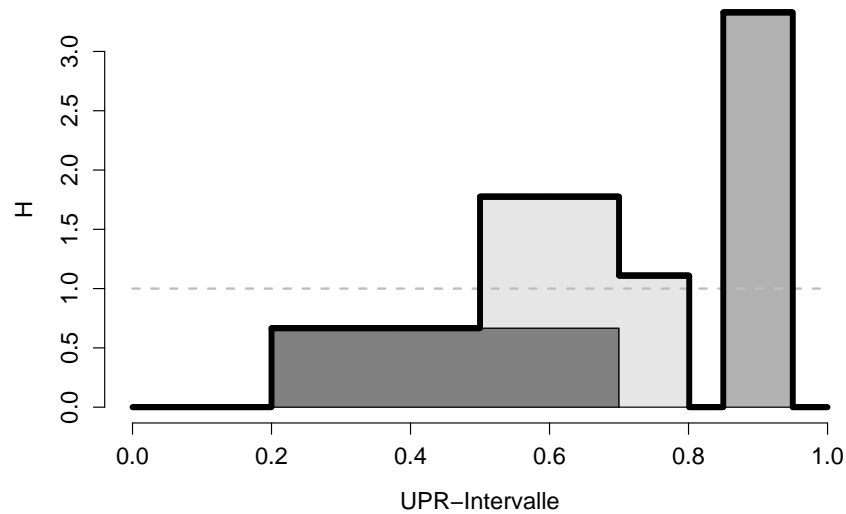


Abbildung 3.7: Erster Schritt bei der Konstruktion von Histogrammen für UPR-Intervalle: Die Umhüllende.

u_1 eine Realisation aus der Wahrscheinlichkeitsfunktion

$$P\{Y_1 = y\} = \begin{cases} 0.2 & \text{für } y = 1 \\ 0.5 & \text{für } y = 2 \\ 0.3 & \text{für } y = 3 \\ 0 & \text{sonst} \end{cases}$$

ist. Betrachtet sei das Niveau $H_1(u)$ der Umhüllenden, die durch die erste Beobachtung erzeugt wird (u_2 und u_3 bleiben also für den Moment außer Betracht). $H_1(u)$ besitzt im Bereich $u \in [0, 0.2[$ den Wert $1/0.2n$, wenn $Y_1 = 1$ beobachtet wird, andernfalls den Wert null. Da $P\{Y_1 = 1\} = 0.2$ gilt, ist der Erwartungswert des Niveaus in diesem Bereich $0.2/0.2n + 0.8 \cdot 0 = 1/n$. Das identische Ergebnis erhält man durch analoges Vorgehen für die Bereiche $u \in [0.2, 0.7[$ und $u \in [0.7, 1.0[$, d.h. für beliebiges u erhält man

$$E(H_1(u)) = 1/n.$$

Auf die gleiche Art lässt sich die Varianz des Niveaus bestimmen. Für $u \in [0.0, 0.2[$ erhält man

$$\text{Var}(H_1(u)) = \frac{0.2}{(0.2n)^2} - \frac{1}{n^2} = 4/n^2.$$

Berechnet man analog die Varianz des Niveaus für die Bereiche $u \in [0.2, 0.7[$ bzw. $u \in [0.7, 1.0[$ und fasst die Ergebnisse zusammen, ergibt sich:

$$\text{Var}(H_1(u)) = \begin{cases} 4/n^2 & \text{für } u \in [0.0, 0.2[\\ 1/n^2 & \text{für } u \in [0.2, 0.7[\\ 7/3n^2 & \text{für } u \in [0.7, 1.0[. \end{cases}$$

Im Gegensatz zum Erwartungswert des Niveaus ist dessen Varianz also von der Struktur der Wahrscheinlichkeitsfunktion von Y_1 abhängig.

Betrachtet man nun wieder die Umhüllende, die durch alle der drei oben genannten UPR-Intervalle erzeugt wird, ergibt sich das erwartete Niveau für $u \in [0, 1[$ durch

$$E(H(u)) = \sum_{i=1}^n E(H_i(u)) = \sum_{i=1}^3 \frac{1}{3} = 1.$$

Dieses Ergebnis gilt für beliebiges n und es basiert auf der Verwendung der korrekten Wahrscheinlichkeiten $P\{Y_i = y\}$. Bei einer Analyse der UPR-Intervalle kann die Umhüllende daher mit der Dichtefunktion der $U(0, 1)$ -Verteilung verglichen werden. Bei Gültigkeit des Modells sollte sie in etwa den gleichen Verlauf besitzen.

Abweichungen von der Umhüllenden können grob anhand der Varianz des Niveaus

$$\text{Var}(H(u)) = \sum_{i=1}^n \text{Var}(H_i(u))$$

beurteilt werden. Diese lässt sich für das zuvor betrachtete Beispiel allerdings nur bei Kenntnis der Wahrscheinlichkeitsfunktionen von Y_1 , Y_2 und Y_3 explizit bestimmen. Allgemein hängt sie von der Struktur sämtlicher Wahrscheinlichkeitsfunktionen des zu überprüfenden Modells ab.

Bei der Analyse von Histogrammen verwendet man in der Regel identische Klassenbreiten. Eine solche Darstellung erhält man, indem die Umhüllende in Abschnitte gleicher Länge unterteilt wird und als Niveau das durchschnittliche Niveau des jeweiligen Abschnitts verwendet. Zur Konstruktion von Abbildung 3.8 wurde die Umhüllende aus Abbildung 3.7 in Abschnitte der Länge 0.2 unterteilt und für diese jeweils das durchschnittliche Niveau der Umhüllenden berechnet. Die so berechneten durchschnittlichen Niveaus entsprechen den Höhen der Balken in Abbildung 3.8, die bei Gültigkeit des Modells ebenfalls den Erwartungswert eins besitzen.

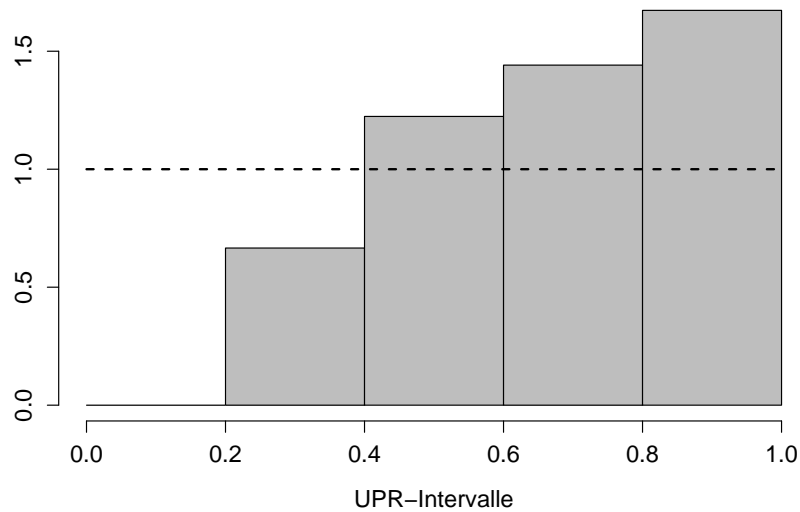


Abbildung 3.8: Zweiter Schritt bei der Konstruktion von Histogrammen für UPR-Intervalle: Gleiche Klassenbreite.

Beispiel 3.5 In diesem Beispiel wird die Anzahl der Arztbesuche (Y) von Patienten aus einer Stichprobe des Jahres 1997 betrachtet. Als erklärende Variablen werden das Alter der Patienten (X_1) und die Anzahl der Arztbesuche im Jahr 1995 (X_2) herangezogen. Die Daten wurden von ZUCCHINI (2000) verwendet, um Methoden der Modellauswahl zu veranschaulichen.

Eine Besonderheit der Stichprobe (insgesamt stehen $n = 23\,618$ Beobachtungen zur Verfügung) ist, dass sie keine Patienten enthält, die weder 1995 noch 1997 beim Arzt waren, d.h. die bedingte Verteilung von Y gegeben $X_2 = 0$ kann nicht korrekt geschätzt werden. Daher wurden aus den Daten 2483 Beobachtungen gestrichen, für die $X_2 = 0$ gilt. Ferner wurden Beobachtungen entfernt, bei denen die Anzahl der Arztbesuche in 1997 den Wert 100 übersteigt (5 Beobachtungen) und Beobachtungen für Personen, die älter sind als 100 Jahre (15 Beobachtungen). Zumindest im ersten Fall ist davon auszugehen, dass die extremen Beobachtungen auf die Datenerhebung zurückzuführen sind (Missbrauch der Versicherungskarten) und somit die Modellbildung nicht beeinflussen sollten.

Nach den Streichungen verbleiben 21 116 Beobachtungen, aus denen für das hier beschriebene Beispiel 500 zufällig ausgewählt wurden. Diese sind in Abbildung 3.9 dargestellt. In den beiden Graphiken ist die Anzahl der Arztbesuche in 1995 bzw. das Alter der Patienten gegen die Anzahl der Arztbesuche in 1997

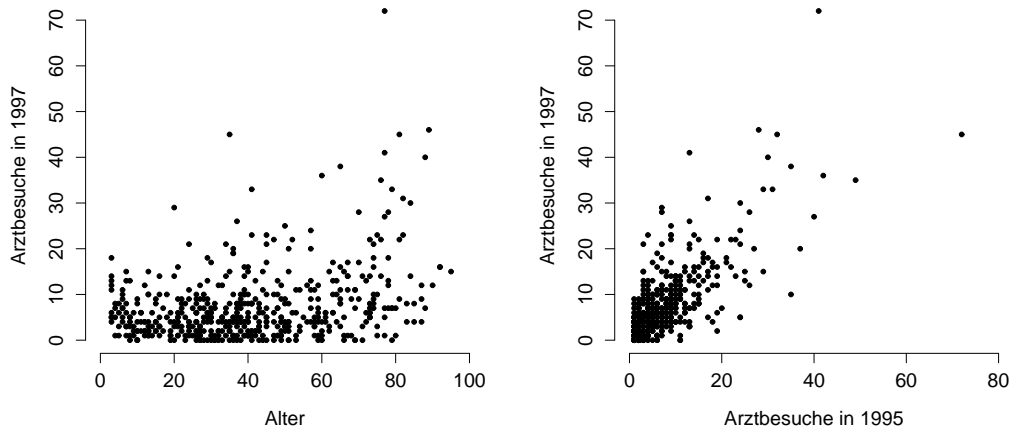


Abbildung 3.9: Plot der Anzahl von Arztbesuchen in 1997 gegen das Alter (links) und gegen die Anzahl der Besuche in 1995 (rechts).

geplottet. Untersucht wird die Beschreibung der Daten durch vier unterschiedliche Modelle:

$$\text{Modell 1: } \log(\lambda_i) = \beta_0 + \beta_1 x_{1i}$$

$$\text{Modell 2: } \log(\lambda_i) = \beta_0 + \beta_1 x_{2i}$$

$$\text{Modell 3: } \log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

$$\text{Modell 4: } \log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

Jeweils wird von einer Poissonverteilung der Zielvariablen ausgegangen. Alle in Tabelle 3.7 angegebenen Parameterschätzer sind signifikant von null verschieden. Würde man das Akaike-Informationskriterium zur Beurteilung der Modelle heranziehen, wäre das vierte Modell zu wählen.

Die schwarz gezeichneten Kurven in Abbildung 3.10 zeigen die Umhüllenden

	Modell 1	Modell 2	Modell 3	Modell 4
$\hat{\beta}_0$	1.335	1.556	1.309	0.897
$\hat{\beta}_1$	0.015	0.048	0.007	0.013
$\hat{\beta}_2$			0.043	0.090
$\hat{\beta}_3$				-0.001
AIC	4531.9	3644.7	3572.5	3462.9

Tabelle 3.7: Parameterschätzer und AIC der vier Modelle zur Beschreibung der Anzahl von Arztbesuchen.

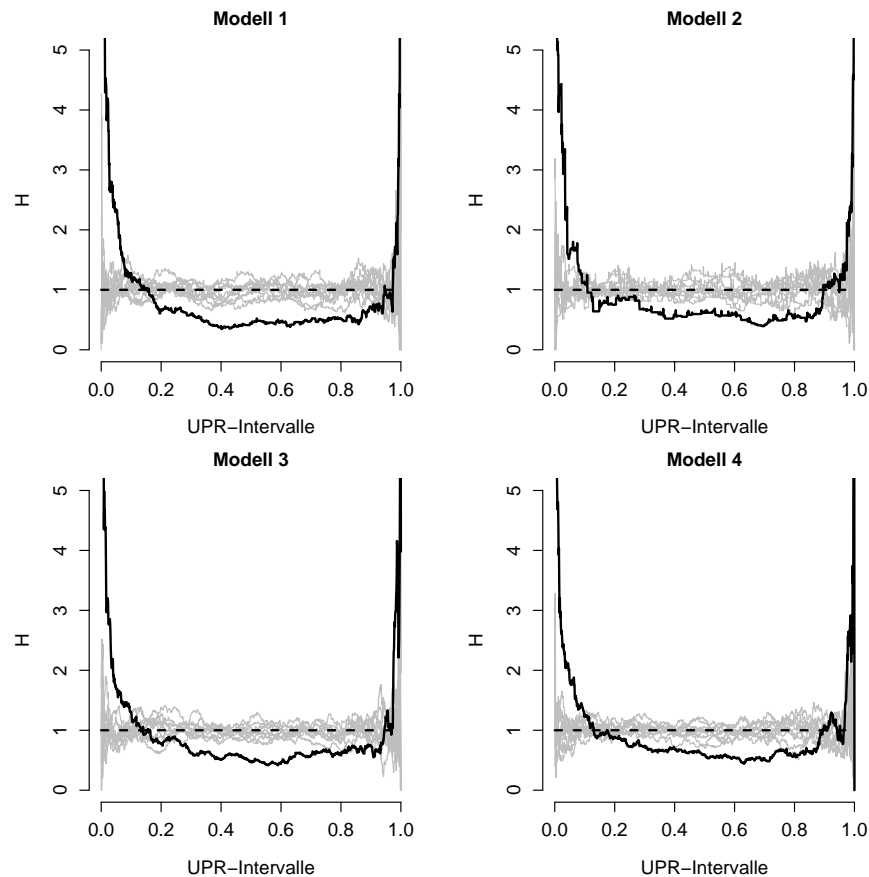


Abbildung 3.10: Umhüllende für die betrachteten Modelle sowie Umhüllende bei Gültigkeit des jeweiligen Modells.

der UPR-Intervalle. Es ist zu sehen, dass deren Niveau an den Rändern deutlich höher ist als im mittleren Bereich. Die grau eingezeichneten Linien zeigen jeweils zehn Umhüllende für Beobachtungen der UPR-Intervalle unter Gültigkeit der entsprechenden Modelle. Diese Beobachtungen wurden mit der parametrischen *Bootstrap*-Methode erzeugt: Die angepassten Modelle wurden verwendet, um entsprechende Zufallszahlen zu erzeugen, an die erneut die Modelle angepasst wurden und für die wiederum die UPR-Intervalle bestimmt wurden. (Eine genauere Beschreibung des Vorgehens erfolgt im nächsten Abschnitt.) Die grauen Linien lassen erkennen, dass die Varianz des Niveaus an den Rändern größer ist als im mittleren Bereich. Im gesamten Bereich streut das Niveau in etwa um den Wert eins, was aus dem oben beschriebenen theoretischen Resultat bezüglich des erwarteten Niveaus folgt.

Für alle Modelle weichen die beobachteten Umhüllenden deutlich von den un-

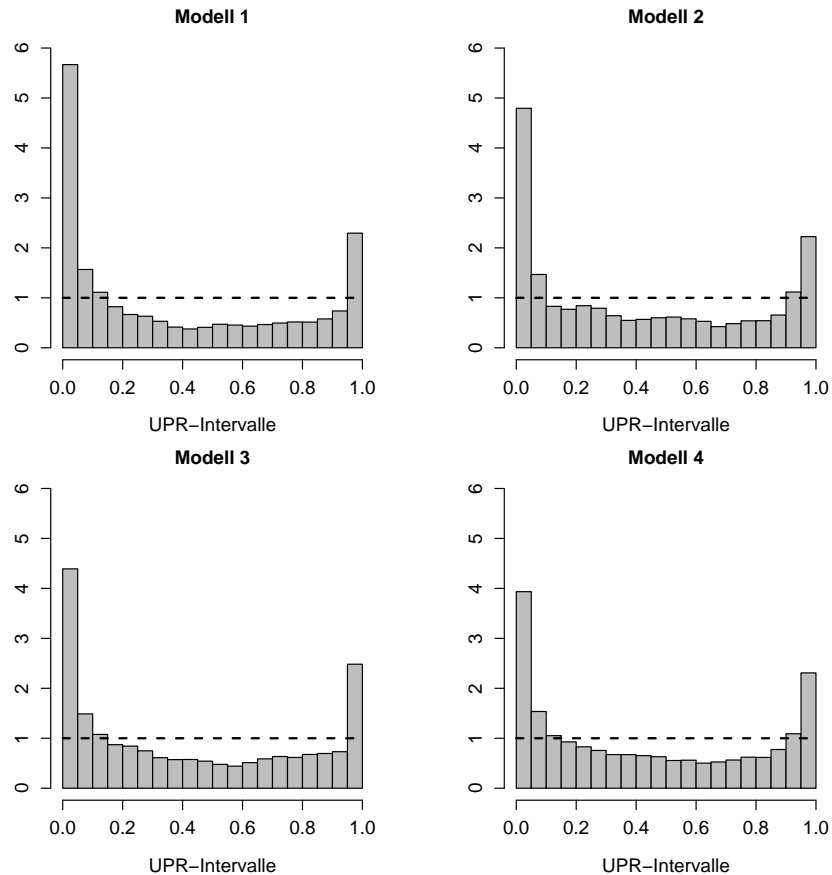


Abbildung 3.11: Histogramme der UPR-Intervalle für die betrachteten Modelle bei poissonverteilter Zielvariable.

ter dem Modell zu erwartenden Umhüllenden ab. Demnach liefert keines der betrachteten Modelle eine adäquate Beschreibung der Daten. Präziser formuliert werden für alle Modelle tendenziell zu wenig UPR-Intervalle im mittleren Bereich bzw. tendenziell zu viele UPR-Intervalle an den Randbereichen beobachtet. Mit Blick auf die Modellanpassung bedeutet dies, dass die Varianz der Poissonverteilungen (die den Erwartungswerten entspricht) zu gering ist. Auf der anderen Seite sind die Umhüllenden in etwa achsensymmetrisch um die vertikale Achse bei $u = 0.5$, d.h. der Erwartungswert der Poissonverteilungen wird in etwa korrekt geschätzt. Zu den gleichen Ergebnissen und Interpretationen führen die in Abbildung 3.11 dargestellten Histogramme der UPR-Intervalle. Diese Histogramme können als Glättung der Umhüllenden verstanden werden, die eine bessere Veranschaulichung der grundlegenden Struktur der Verteilung der UPR-Intervalle ermöglicht.

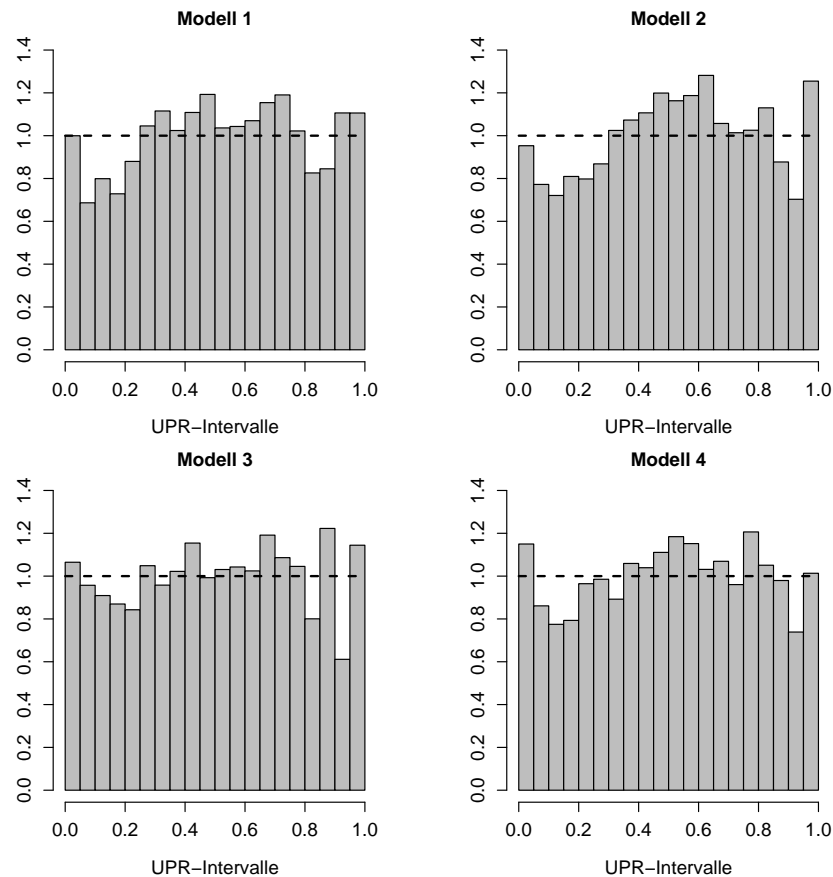


Abbildung 3.12: Histogramme der UPR-Intervalle für die betrachteten Modelle bei negativ binomialverteilter Zielvariable.

Das allgemeine Problem im hier betrachteten Beispiel ist, dass die Varianz der Y_i größer ist als ihr Erwartungswert und die verwendete Poissonverteilung nur einen Parameter besitzt, der sowohl dem Erwartungswert als auch der Varianz entspricht. Dieses Problem ist unter dem englischen Begriff *overdispersion* bekannt und kann bei der Modellierung berücksichtigt werden, indem zusätzlich ein Dispersionsparameter ϕ in das Modell aufgenommen wird, der den Zusammenhang $Var(Y) = \phi E(Y)$ zwischen Erwartungswert und Varianz liefert (siehe beispielsweise DOBSON, 2002, S. 167). In **R** erfolgt die Schätzung des Dispersionsparameters durch die Verwendung des Arguments `family='quasipoisson'` innerhalb der Funktion `glm`. Für die vier oben beschriebenen Modelle ergibt sich $\hat{\phi}_1 = 6.61$, $\hat{\phi}_2 = 4.04$, $\hat{\phi}_3 = 3.86$ und $\hat{\phi}_4 = 3.73$ (die übrigen Schätzer entsprechen denen aus Tabelle 3.7).

Für diese Modelle können die Pseudo-Residuen natürlich nicht mehr mit Hil-

fe der Poissonverteilung berechnet werden, da die Modelle unterschiedliche Schätzer für die Varianzen und die Erwartungswerte der Y_i liefern. Stattdessen werden die Schätzer zur Spezifikation von negativen Binomialverteilungen verwendet. Dazu ist die Parametrisierung der negativen Binomialverteilungen mit den Parametern $\mu_i > 0$ und $\tau_i > 0$ durch

$$p_i(y) = \begin{cases} \frac{\Gamma(y+\tau_i^{-1})}{\Gamma(\tau_i^{-1})\Gamma(y+1)} \left(\frac{1}{1+\tau_i\mu_i}\right)^{1/\tau_i} \left(\frac{\tau_i\mu_i}{1+\tau_i\mu_i}\right)^y & \text{für } y = 0, 1, 2, \dots \\ 0 & \text{sonst} \end{cases}$$

mit

$$E(Y) = \mu_i \quad \text{und} \quad \text{Var}(Y) = \mu_i(1 + \tau_i\mu_i).$$

sinnvoll. Konvergiert τ_i gegen null, ergibt sich eine Poissonverteilung. Wird der Parameter größer, erhöht sich die Varianz bei konstantem Erwartungswert (vgl. ZUCCHINI AND ADAMSON, 1989). Die Verteilung kann als Poissonverteilung mit Varianzflexibilität interpretiert werden, d.h. es handelt sich um eine Verteilung, die dem oben beschriebenen Problem gerecht wird. Schätzt man die Parameter der negativen Binomialverteilungen für das Modell j durch

$$\begin{aligned} \hat{\mu}_i &= \hat{\lambda}_i \\ \text{bzw. } \hat{\tau}_i &= \frac{\hat{\phi}_j - 1}{\hat{\lambda}_i}, \end{aligned}$$

so ergeben sich die in Abbildung 3.12 dargestellten Histogramme der UPR-Intervalle. Die Anpassung ist für alle Modelle deutlich besser als bei Verwendung der Poissonverteilung. Dieses Beispiel macht somit deutlich, wie die Informationen aus den Histogrammen der UPR-Intervalle zu einem modifizierten Modell führen, das eine geeignetere Beschreibung der Daten liefert.

◀

3.3.2.3 Dreieck-Plots für UPR-Intervalle

Eine weitere Möglichkeit, die UPR graphisch zu untersuchen, bieten Dreieck-Plots. Sie basieren auf den drei Wahrscheinlichkeiten

$$\begin{aligned} P^- &= P\{Y_i < y_i\} = u_i^-, \\ P &= P\{Y_i = y_i\} = u_i^+ - u_i^- \quad \text{und} \\ P^+ &= P\{Y_i > y_i\} = 1 - u_i^+, \end{aligned}$$

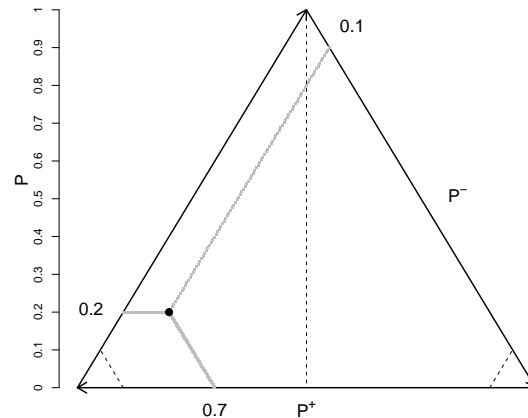


Abbildung 3.13: Darstellung des Punktes $(0.1, 0.2, 0.7)$ im dreieckigen Koordinatensystem.

die sich aus dem UPR-Intervall ableiten. Die Summe dieser Wahrscheinlichkeiten beträgt für beliebiges i eins, so dass eine Darstellung der drei Wahrscheinlichkeiten in zwei Dimensionen möglich ist. Geometrisch interpretiert, liegen die drei Wahrscheinlichkeiten in der Fläche, die durch die Punkte $(1, 0, 0)$, $(0, 1, 0)$ und $(0, 0, 1)$ in einem dreidimensionalen Koordinatensystem gebildet wird. Die Seiten dieser Fläche können als Achsen eines zweidimensionalen Koordinatensystems verwendet werden, das im Englischen häufig als *triangular coordinate system* bezeichnet wird. Für eine detaillierte Beschreibung der Konstruktion des Koordinatensystems sowie der Berechnung der Koordinaten der im System darzustellenden Punkte sei auf GREENACRE (1993, S. 11–16) verwiesen.

Hier sei beispielhaft die Darstellung des UPR-Intervalls $[0.1, 0.3]$ erläutert sowie auf einige Informationen hingewiesen, die aus einem Dreieck-Plot mit Blick auf die Modellprüfung abgeleitet werden können. Aus dem UPR-Intervall $[0.1, 0.3]$ ergeben sich die Wahrscheinlichkeiten $P^- = 0.1$, $P = 0.2$ und $P^+ = 0.7$. Diese sind als Punkt in einem dreieckigen Koordinatensystem in Abbildung 3.13 dargestellt. Die rechte Seite des Dreiecks repräsentiert die P^- -Achse, die untere Seite die P^+ -Achse und die linke Seite die P -Achse. Die grauen Linien in dem Plot zeigen, wie der Punkt auf die Achsen zu projizieren ist, um dessen Koordinaten graphisch zu bestimmen.

Eine grundlegende Idee bei der Definition der UPR in Form von Intervallen ist der Erhalt der Information über die Wahrscheinlichkeit einer Beobachtung un-

ter dem Modell. Diese Wahrscheinlichkeit, die im Dreieck-Plot mit P bezeichnet ist und direkt an der vertikalen Achse abgelesen wird, kann beispielsweise als zusätzliche Information bei der Beurteilung extremer Residuen herangezogen werden: Sie zeigt, ob eine Beobachtung aufgrund ihrer Wahrscheinlichkeit ungewöhnlich ist oder nicht.

Um zu beurteilen, ob ein UPR-Intervall extrem ist, müssen dessen Intervallgrenzen betrachtet werden. Ist die rechte Intervallgrenze sehr klein bzw. die linke Intervallgrenze sehr groß, ist die Wahrscheinlichkeit, unter dem Modell eine kleinere bzw. größere Beobachtung zu erhalten, sehr gering. Solche Intervalle, die extreme Beobachtungen unter dem Modell charakterisieren, liegen im Dreieck-Plot in der linken bzw. rechten unteren Ecke. In diese Ecken sind in Abbildung 3.13 gestrichelte Linien eingezeichnet. Sie begrenzen den Bereich, in dem Residuen liegen, deren rechte Intervallgrenze kleiner ist als 0.1 bzw. deren linke Intervallgrenze größer ist als 0.9.

Eine Beobachtung, deren UPR-Intervall den Mittelpunkt

$$u^m = \frac{u^- + u^+}{2} = 0.5$$

besitzt, ist unter dem Modell unabhängig von ihrer Wahrscheinlichkeit nicht extrem (zumindest für unimodale Verteilungen). Für UPR-Intervalle mit dem Mittelpunkt 0.5 gilt $P^- = P^+$ und die entsprechenden Punkte liegen im Dreieck-Plot auf dem Lot aus der oberen Ecke des Dreiecks (senkrechte gestrichelte Linie). Diese Aussage lässt sich dahingehend verallgemeinern, dass zwei Punkte, die auf einer vertikalen Linie liegen, UPR-Intervalle mit identischen Mittelpunkten besitzen. Die Sortierung der Punkte von links nach rechts entspricht damit der Sortierung im zuvor erläuterten QQ-Plot für NPR-Intervalle.

Fasst man die wesentlichen Eigenschaften des Dreieck-Plots zusammen, handelt es sich um einen Plot der Mittelpunkte der UPR-Intervalle gegen die Wahrscheinlichkeiten der Beobachtungen unter dem Modell, wobei auch die Wahrscheinlichkeiten für eine größere bzw. kleinere Beobachtung unter dem Modell (durch Projektion auf die entsprechenden Achsen) abgelesen werden können.

Beispiel 3.6 Die Daten in Tabelle 3.8 zeigen die Ergebnisse eines Versuchs, der bei DOBSON (2002, S. 129–131) betrachtet wird. Für 54 Personen wurde festgestellt, ob sie Symptome von Senilität zeigen ($Y = 1$) oder ob dies nicht der Fall ist ($Y = 0$). Gleichzeitig wurde für die Personen die erreichte Punkt-

i	x_i	n_i	y_i	i	x_i	n_i	y_i
1	4	2	1	10	13	6	1
2	5	1	1	11	14	7	2
3	6	2	1	12	15	3	0
4	7	3	2	13	16	4	0
5	8	2	2	14	17	1	0
6	9	6	2	15	18	1	0
7	10	6	1	16	19	1	0
8	11	6	1	17	20	1	0
9	12	2	0				

Tabelle 3.8: Anzahl von Personen mit Anzeichen von Senilität aus n_i Personen mit Punktzahl x_i im Intelligenztest.

zahl X bei einem Intelligenztests erhoben. Insgesamt wurden 17 unterschiedliche Punktzahlen beobachtet. In der Tabelle beschreibt y_i die beobachtete Anzahl der Personen mit Senilitäterscheinungen aus n_i Personen mit erreichter Punktzahl x_i im Intelligenztest.

Das angepasste Modell

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = \text{logit}(\hat{\pi}_i) = 2.40 - 0.32x_i \quad \text{für } i = 1, \dots, 17$$

unterstellt eine Binomialverteilung der Y_i , wobei der Zusammenhang zwischen den Erfolgswahrscheinlichkeiten und der linearen Erklärungsfunktion über den Logit-Link hergestellt wird.

Abbildung 3.14 zeigt den Dreieck-Plot der UPR-Intervalle und den QQ-Plot der NPR-Intervalle des Beispiels. Der QQ-Plot zeigt keine Auffälligkeiten, die auf ein fehlerhaft spezifiziertes Modell schließen lassen. Im Dreieck-Plot liegen die Punkte der UPR-Intervalle für die Beobachtungen $y_5 = 2$ und $y_{11} = 2$ (isoliert von den übrigen Punkten) am rechten unteren Rand. Abbildung 3.15 zeigt die Wahrscheinlichkeitsfunktionen der angepassten Verteilungen

$$Y_5 \sim b(2, 0.45) \quad \text{und} \quad Y_{11} \sim b(7, 0.11).$$

In beiden Fällen liegen die entsprechenden Beobachtungen am rechten Rand der angepassten Verteilungen. Da die Verteilungen eine unterschiedliche Gestalt besitzen, fällt es schwer, allgemein zu beurteilen, welche der Beobachtungen extremer ist. Optisch erscheinen beide Beobachtungen in etwa gleich

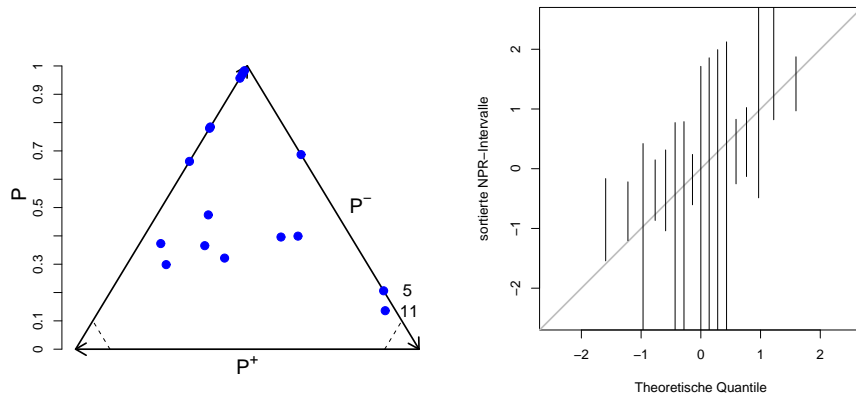


Abbildung 3.14: Dreieck-Plot der UPR-Intervalle (links) und QQ-Plot der NPR-Intervalle (rechts).

extrem. Berechnet man für beide Beobachtungen allerdings die *deviance residuals*, die häufig für logistische Regressionsmodelle verwendet werden, erhält man

$$r_5^d = 1.78 \quad \text{und} \quad r_{11}^d = 1.31.$$

Das Ergebnis der optischen Eindrücke spiegelt sich in diesen Werten kaum (wenn nicht sogar fehlerhaft) wider. Besser geeignet erscheint die Wiedergabe der Situation durch die Mittelpunkte

$$u_5^m = 0.897 \quad \text{und} \quad u_{11}^m = 0.902$$

der UPR-Intervalle, die für beide Beobachtungen nahezu identisch sind. Dies ist auch unmittelbar anhand des Dreieck-Plots zu erkennen, da die Punkte in etwa auf einer vertikalen Linie liegen.

Ferner erhält man aus dem Plot zusätzlich die Information über die Wahrscheinlichkeit der Beobachtungen. Die Wahrscheinlichkeit für $y_5 = 2$ beträgt in etwa 0.2. Die Beobachtung besitzt in diesem Sinn eine gewisse Plausibilität unter dem Modell, d.h. sie ist nicht ungewöhnlich. Die Idee, die Residuen bei der Modellierung von diskreten Merkmalen als Intervalle zu definieren, d.h. Informationen über die Wahrscheinlichkeiten der Beobachtungen unter dem Modell zu erhalten, ist somit in der graphischen Analyse des Dreieck-Plots umgesetzt.

◀

Beispiel 3.7 Die Daten eines weiteren Beispiels von DOBSON (2002, S. 123–125) sind in Tabelle 3.9 wiedergegeben. Die ursprünglichen Daten stammen

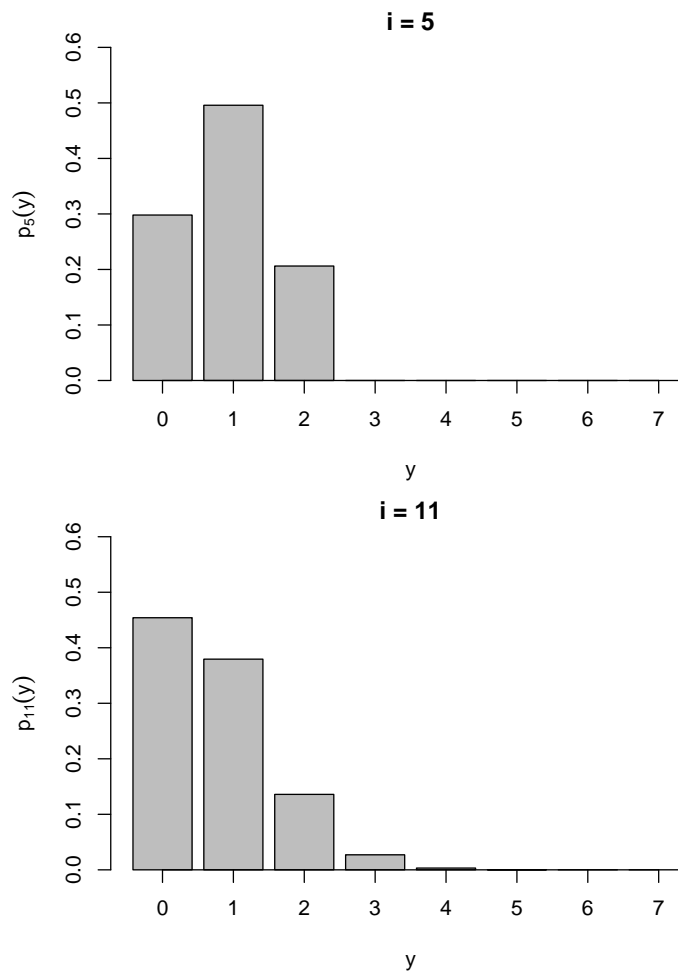


Abbildung 3.15: Wahrscheinlichkeitsfunktionen für Y_5 und Y_{11} unter dem angepassten Modell.

von SANGWAN–NORRELL (1977) und sie wurden ferner von WOOD (1978) untersucht. Modelliert wird die Anzahl entkeimter Staubbeutel Y_{ij} einer Pflanze mit insgesamt n_{ij} Staubbeuteln. Die Pflanzen wurden verschiedenen Bedingungen ausgesetzt: Zum einen wird ein Faktor berücksichtigt, der die Lagerung der Pflanzen beschreibt. Die Behandlung der Pflanzen erfolgte durch eine Lagerung bei 3 Grad Celsius für eine Zeit von 48 Stunden ($i = 1$). Pflanzen der Kontrollgruppe wurden dieser Behandlung nicht ausgesetzt ($i = 2$). Zum anderen beinhaltet der Versuch (bzw. das Modell) eine quantitative erklärende Variable, die die Kraft X_j einer Zentrifuge misst, mit der die Pflanzen behandelt wurden.

Lagerung	Kraft der Zentrifuge			
		$x_1 = \log(40)$	$x_2 = \log(150)$	$x_3 = \log(350)$
Behandlung	y_{1j}	55	50	50
	n_{1j}	76	81	90
Kontrollgruppe	y_{2j}	55	52	57
	n_{2j}	102	99	108

Tabelle 3.9: Beobachtete Anzahl entkeimter Staubbeutel aus n_{ij} Staubbeuteln unter verschiedenen Bedingungen.

Modell 1	$\hat{\alpha} = 1.021$	$\hat{\beta} = -0.148$		
Modell 2	$\hat{\alpha}_1 = 0.877$	$\hat{\alpha}_2 = 1.285$	$\hat{\beta} = -0.155$	
Modell 3	$\hat{\alpha}_1 = 0.234$	$\hat{\alpha}_2 = 2.211$	$\hat{\beta}_1 = -0.023$	$\hat{\beta}_2 = -0.341$

Tabelle 3.10: Parameterschätzer der drei Modelle zur Beschreibung der Anzahl entkeimter Staubbeutel.

Zur Modellierung der Daten seien die drei Modelle

$$\text{Modell 1: } \text{logit}(\pi_{ij}) = \alpha + \beta x_j$$

$$\text{Modell 2: } \text{logit}(\pi_{ij}) = \alpha_i + \beta x_j$$

$$\text{Modell 3: } \text{logit}(\pi_{ij}) = \alpha_i + \beta_i x_j$$

betrachtet. Modell 1 unterstellt, dass die Lagerung der Pflanzen keinen Einfluss auf die Anzahl entkeimter Staubbeutel besitzt (gleicher Achsenabschnitt und gleiche Steigung in beiden Gruppen). Bei der Verwendung von Modell 2 ist die Lagerungsart zwar berücksichtigt, jedoch ist der Einfluss der Zentrifuge in beiden Gruppen gleich (unterschiedlicher Achsenabschnitt und gleiche Steigung). Die Erweiterung des dritten Modells besteht darin, dass auch der Einfluss der Zentrifuge je nach Lagerung unterschiedlich sein kann (unterschiedlicher Achsenabschnitt und unterschiedliche Steigung). Die Parameterschätzer der angepassten Modelle sind in Tabelle 3.10 angegeben.

Abbildung 3.16 zeigt die Dreieck-Plots der UPR-Intervalle für die drei Modelle (die UPR-Intervalle der Modelle 1 und 3 sind gemeinsam im linken Plot eingezeichnet; betrachtet wird nur der untere Teil der Dreiecke, weil die Wahrscheinlichkeiten der Beobachtungen unter dem Modell (also die Werte von P) aufgrund der relativ großen Werte für n_{ij} gering sind). Jeweils sind die Residuen für die Kontrollgruppe als Punkte und die der Behandlungsgruppe als Dreiecke dargestellt.

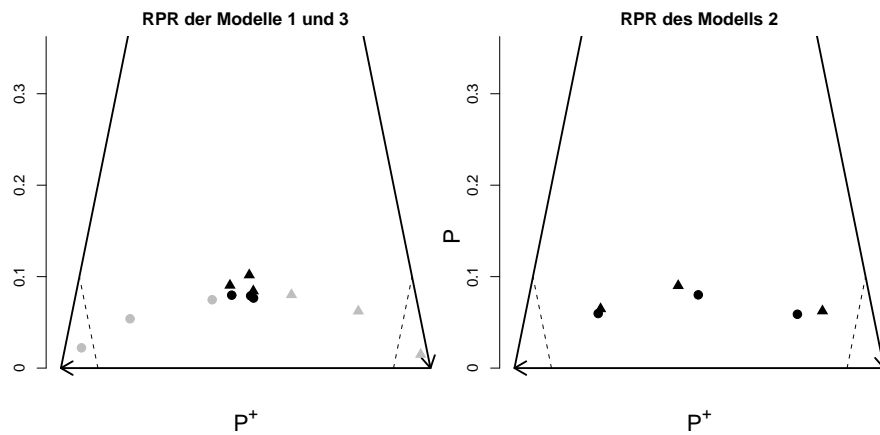


Abbildung 3.16: Dreieck-Plots der UPR-Intervalle für die Modelle 1 und 3 (links) sowie für Modell 2 (rechts).

Die grauen Punkte bzw. Dreiecke in der linken Graphik repräsentieren die UPR-Intervalle des Modells 1. Es ist zu sehen, dass der Einfluss der Lagerung nicht vernachlässigt werden kann: Alle Residuen der Kontrollgruppe befinden sich auf der linken Seite, alle Residuen der Behandlungsgruppe auf der rechten Seite des Plots. Die Anzahl entkeimter Staubbeutel wird durch dieses Modell in der Kontrollgruppe also überschätzt und in der Behandlungsgruppe unterschätzt.

Alle schwarz dargestellten UPR-Intervalle (Modell 3) liegen in der Mitte des Plots, d.h. die Beschreibung der Daten ist sehr gut (alle Beobachtungen liegen im Zentrum der angepassten Verteilungen). Dies liegt natürlich daran, dass hier vier Parameter zur Beschreibung von sechs Beobachtungen geschätzt wurden. Das *angepasste* Modell beschreibt daher zwar die Daten, von einer Beschreibung des datenerzeugenden Phänomens ist allerdings nicht auszugehen. Ein solches ‚*over-fitting*‘ ist nicht erwünscht (vgl. DOBSON, 2002, S. 124).

Die rechte Graphik zeigt die Residuen des Modells 2. Jeweils zwei Residuen liegen dicht beieinander, was darauf schließen lässt, dass die Residuen nicht unabhängig sind. Dies ist allerdings auch nicht zu erwarten, da auch hier die Anzahl der Parameter relativ groß ist im Vergleich zur Anzahl der Beobachtungen. Allerdings liegen hier jeweils ein Residual der Kontrollgruppe und eines der Behandlungsgruppe beieinander. Für das linke Paar ist die Kraft der Zentrifuge in der Kontrollgruppe gering ($x_1 = \log(40)$) und in der Behandlungsgruppe groß ($x_3 = \log(350)$), für das rechte Paar gilt die umgekehrte Kombination und für das mittlere Paar ist der Einfluss in beiden Gruppen

identisch ($x_2 = \log(150)$). Dies lässt auf eine Wechselwirkung schließen, d.h. auf einen unterschiedlichen Einfluss der Zentrifuge in den beiden Gruppen. Die gruppierte Anordnung der Punkte im Dreieck-Plot zeigt hier also, dass ein korrektes Modell einen unterschiedlichen Einfluss der Zentrifuge in den beiden Gruppen berücksichtigen muss.



Beispiel 3.8 Wie bereits im Beispiel zu den Maschinenausfällen (Beispiel 3.4) angedeutet wurde, sind die Dreieck-Plots geeignet, um Abweichungen von der Winkelhalbierenden in QQ-Plots für NPR-Intervalle zu beurteilen. In Abbildung 3.6 war auffällig, dass die Segmente der Beobachtungen vier und sieben nicht auf der Winkelhalbierenden liegen und es stellt sich die Frage, ob diese Abweichungen unter dem Modell zu erwarten sind oder nicht.

Soll beispielsweise beurteilt werden, ob das NPR-Intervall mit dem kleinsten Intervallmittelpunkt (also das erste der sortierten Intervalle) im QQ-Plot auf der Winkelhalbierenden liegt, so ist die Referenz eine Verteilung von Intervallen mit unterschiedlicher Länge und unterschiedlicher Lage. Die Beurteilung ist daher problematisch und ein analytisches Vorgehen wie im stetigen Fall (siehe Abschnitt ‚Pseudo-Residuen für QQ-Plots‘) ergibt sich nicht unmittelbar. Der Dreieck-Plot bietet den Vorteil, dass hier eine Verteilung von UPR-Intervallen als Punktwolke dargestellt werden kann. Ferner wurde bereits erwähnt, dass die Sortierung der Punkte im Dreieck-Plot von links nach rechts und die Sortierung der NPR-Intervalle bei der Konstruktion des QQ-Plots identisch sind. Daher ist es hier die Idee, eine Punktwolke der Referenzverteilung gemeinsam mit dem beobachteten Intervall im Dreieck-Plot darzustellen und dessen Lage in der Referenzverteilung graphisch zu beurteilen.

In Abbildung 3.17 sind die UPR-Intervalle der Beobachtungen sieben und vier als fettgedruckte Punkte dargestellt (der linke Punkt gehört zur siebten Beobachtung, der rechte zur vierten). Die Beobachtung sieben besitzt das zweite und die Beobachtung vier das sechste der sortierten NPR-Intervalle. Die hellgrauen bzw. dunkelgrauen Punkte repräsentieren jeweils 200 Beobachtungen aus der Verteilung des zweiten bzw. sechsten Intervalls nach Sortierung unter Gültigkeit des Modells. Beide Verteilungen sind auch hier mit der parametrischen *Bootstrap*-Methode erzeugt worden (eine genauere Beschreibung des Vorgehens erfolgt im nächsten Abschnitt). Der Dreieck-Plot zeigt, dass der Punkt der siebten Beobachtung am rechten Rand und der Punkt der vierten Beobachtung am linken Rand der jeweiligen Referenzverteilung liegt. Dies ist zu erwarten, da die Segmente im QQ-Plot über bzw. unter der Winkelhalbieren-

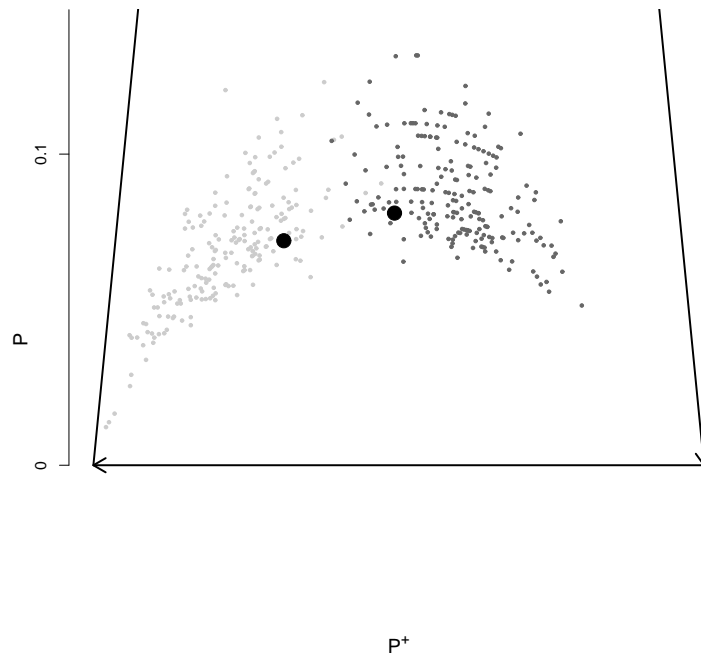


Abbildung 3.17: Dreieck-Plot mit Referenzverteilung von UPR-Intervallen.

den liegen. Die Lage der Punkte in der Punktwolke und somit die Abweichung der Segmente von der Winkelhalbierenden erscheint jedoch in beiden Fällen nicht extrem.



3.4 Berücksichtigung der Parameterschätzung

In allen bislang dargestellten Analysen wurde davon ausgegangen, dass die Schätzung und die damit verbundene Streuung der Parameterschätzer keinen Einfluss auf die Verteilung der Residuen ausübt bzw. zu vernachlässigen ist. Auf der anderen Seite ist klar, dass die mit dem angepassten Modell berechneten Residuen eine andere Verteilung besitzen als die Residuen des Modells, das die Daten erzeugt hat. Der Unterschied zwischen den Verteilungen ist von geringerer Bedeutung, wenn die Anzahl der Beobachtungen groß ist im Vergleich zu der Anzahl der Parameter. Dies ist beispielsweise für die in Kapitel 4 untersuchten Zeitreihenmodelle für Aktienrenditen der Fall. Ist die Anzahl der Beobachtungen hingegen relativ gering im Vergleich zur Anzahl der Para-

meter, kann eine Nichtberücksichtigung der Schätzung zu Fehlinterpretationen bei der Residuenanalyse führen. Im Folgenden wird ein Vorgehen zur Kalibrierung der Residuen vorgestellt: Durch eine Transformation der Residuen wird erreicht, dass diese auch bei Schätzung der Modellparameter $U(0, 1)$ -verteilt bzw. standardnormalverteilt sind. Die Ausführungen dieses Abschnitts orientieren sich an ZUCCHINI (2002) sowie STADIE, ZUCCHINI AND MACDONALD (2002).

Formal kann die Problematik wie folgt beschrieben werden: Es ist bekannt, dass die $U_i = F_i(Y_i)$ bzw. $R_i = \Phi^{-1}(U_i)$ unabhängig und identisch $U(0, 1)$ -verteilt bzw. standardnormalverteilt sind. Bei einer Schätzung der Parameter werden allerdings die Verteilungen der geschätzten UPR bzw. NPR

$$\hat{U}_i = \hat{F}_i(Y_i) \quad \text{bzw.} \quad \hat{R}_i = \Phi^{-1}(\hat{U}_i)$$

benötigt, wobei $\hat{F}_i(Y_i) = F_i(Y_i; \hat{\theta})$ die Schätzer der Verteilungsfunktionen darstellen. Die Verteilung der \hat{U}_i ist zwar im Intervall $[0, 1]$ definiert, jedoch handelt es sich nicht um eine Rechteckverteilung. Seien Q_i die Verteilungsfunktionen der geschätzten UPR, dann ist

$$\begin{aligned} F_i(y) &= P\{Y_i \leq y\} \\ &= P\{\hat{F}_i(Y_i) \leq \hat{F}_i(y)\} \\ &= Q_i(\hat{F}_i(y)). \end{aligned}$$

Daraus folgt, dass die kalibrierten UPR

$$U_i^k = Q_i(\hat{F}_i(Y_i)) \quad \text{für } i = 1, \dots, n$$

bei Gültigkeit des Modells und Schätzung der Parameter $U(0, 1)$ -verteilt sind. An dieser Stelle ergibt sich allerdings ein neues Problem: Die Verteilungsfunktionen Q_i sind in der Regel nicht bekannt. Im allgemeinen hängen sie von den (unbekannten) Parametern, den verwendeten Verteilungen und der Schätzmethode ab. Es gibt jedoch einige wenige Fälle, in denen die Q_i bekannt und unabhängig von den Parametern des Modells sind. Diese werden nachfolgend beschrieben, bevor anschließend gezeigt wird, wie die Verteilungsfunktionen in den anderen Fällen mit Hilfe der *Bootstrap*-Methode geschätzt werden können.

Betrachtet sei ein klassisches lineares Regressionsmodell

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}),$$

wobei \mathbf{I} eine $n \times n$ Einheitsmatrix repräsentiert. In der Terminologie der verallgemeinerten linearen Modelle handelt es sich hier um ein Modell mit normalverteilten Zielvariablen und Identitätslink. Die UPR und die geschätzten UPR sind durch

$$U_i = \Phi\left(\frac{Y_i - X_i\boldsymbol{\beta}}{\sigma}\right) \quad \text{bzw.} \quad \hat{U}_i = \Phi\left(\frac{Y_i - X_i\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)$$

gegeben, wobei X_i die i -te Zeile der Designmatrix und $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ den Schätzer für $\boldsymbol{\beta}$ nach der Methode der kleinsten Quadrate darstellt. Neben den Regressionskoeffizienten muss auch σ^2 geschätzt werden. Diesbezüglich beschreiben COOK AND WEISBERG (1982, S. 18–20) zwei Alternativen: den internen Schätzer

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2$$

und den externen Schätzer

$$\hat{\sigma}_{(i)}^2 = \frac{(n-p)\hat{\sigma}^2 - (y_i - \hat{\mu}_i)^2/(1-h_{ii})}{n-p-1},$$

wobei h_{ii} das i -te Diagonalelement der so genannten *Hat*-Matrix ist. Diese besitzt für klassische lineare Modelle die Form

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

Betrachtet sei zunächst die Verteilung der geschätzten UPR bei Verwendung des internen Varianzschätzers. Hier erhält man aus den Verteilungsfunktionen der geschätzten UPR:

$$\begin{aligned} Q_i^{\text{int}}(u) &= P\left\{\hat{U}_i \leq u\right\} \\ &= P\left\{\Phi\left(\frac{Y_i - X_i\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) \leq u\right\} \\ &= P\left\{\frac{Y_i - X_i\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{1-h_{ii}}} \leq \frac{\Phi^{-1}(u)}{\sqrt{1-h_{ii}}}\right\}. \end{aligned}$$

Sei

$$W = \frac{Y_i - X_i\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

dann ist das kalibrierte UPR durch

$$U_i^{k \text{ int}} = F_W \left(\frac{\Phi^{-1}(\hat{U}_i)}{\sqrt{1 - h_{ii}}} \right)$$

gegeben. Um die $U_i^{k \text{ int}}$ explizit berechnen zu können, ist F_W zu bestimmen. Den Ansatz dazu liefert die Tatsache, dass $W^2/(n-p)$ betaverteilt ist mit den Parametern $\alpha = 1/2$ und $\beta = (n-p-1)/2$ (siehe beispielsweise COOK AND WEISBERG, 1982, S. 19). Man erhält daher

$$F_{\text{Beta}} \left(\frac{w^2}{n-p} \right) = P \left\{ \frac{W^2}{n-p} \leq \frac{w^2}{n-p} \right\} = P \{W^2 \leq w^2\}$$

Daraus folgt für $w \geq 0$

$$\begin{aligned} P \{-w \leq W \leq w\} &= F_{\text{Beta}} \left(\frac{w^2}{n-p} \right) \\ \Rightarrow F_W(w) - F_W(-w) &= F_{\text{Beta}} \left(\frac{w^2}{n-p} \right) \\ \Rightarrow F_W(w) - (1 - F_W(w)) &= F_{\text{Beta}} \left(\frac{w^2}{n-p} \right) \\ \Rightarrow F_W(w) &= \frac{1}{2} + \frac{1}{2} F_{\text{Beta}} \left(\frac{w^2}{n-p} \right). \end{aligned}$$

In der dritten Zeile wurde $F_W(-w) = 1 - F_W(w)$ verwendet. Diese Identität resultiert aus der Achsensymmetrie der Dichtefunktion von W um die vertikale Achse bei $w = 0$ (vgl. COOK AND WEISBERG, S. 19). Weiterhin folgt aus $F_W(-w) = 1 - F_W(w)$ für $w < 0$

$$F_W(w) = \frac{1}{2} - \frac{1}{2} F_{\text{Beta}} \left(\frac{w^2}{n-p} \right).$$

Fasst man die Ergebnisse zusammen, erhält man für die gesuchte Verteilungsfunktion von W

$$F_W(w) = \begin{cases} \frac{1}{2} - \frac{1}{2} F_{\text{Beta}} \left(\frac{w^2}{n-p} \right) & -\sqrt{n-p} \leq w < 0 \\ \frac{1}{2} + \frac{1}{2} F_{\text{Beta}} \left(\frac{w^2}{n-p} \right) & 0 \leq w \leq \sqrt{n-p}, \end{cases}$$

wobei der Definitionsbereich $[-\sqrt{n-p}, \sqrt{n-p}]$ von W aus dem Definitionsbereich $[0, 1]$ der Betaverteilung von $W^2/(n-p)$ abgeleitet ist.

Für den externen Varianzschätzer ergibt sich aus der Verteilungsfunktion der geschätzten UPR

$$Q_i^{\text{ext}}(u) = P \left\{ \frac{Y_i - X_i \hat{\beta}}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}} \leq \frac{\Phi^{-1}(u)}{\sqrt{1 - h_{ii}}} \right\}.$$

Hier ist bekannt, dass (COOK AND WEISBERG, 1982, S. 20)

$$\frac{Y_i - X_i \hat{\beta}}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}}$$

t -verteilt ist mit $n - p - 1$ Freiheitsgraden, d.h. man berechnet die kalibrierten UPR bei Verwendung des externen Varianzschätzers durch

$$U_i^{k \text{ ext}} = F_{t, n-p-1} \left(\frac{\Phi^{-1}(\hat{U}_i)}{\sqrt{1 - h_{ii}}} \right).$$

Bemerkenswert ist, dass in beiden Fällen (für den internen und den externen Schätzer der Varianz) die Verteilungsfunktion der geschätzten UPR nicht von den Modellparametern β und σ^2 abhängt. Daher sind die kalibrierten UPR, die natürlich auch zur Berechnung der kalibrierten NPR verwendet werden können, auch bei Schätzung der Parameter exakt $U(0, 1)$ -verteilt. Es ist also im Fall der *klassischen* linearen Regression möglich, auch bei unbekanntem Parameter UPR bzw. NPR zu bestimmen, die bei Gültigkeit des Modells exakt $U(0, 1)$ - bzw. $N(0, 1)$ -verteilt sind.

Beispiel 3.9 Um die Bedeutung der Kalibrierung bzw. den Einfluss der Parameterschätzung an einem numerischen Beispiel zu veranschaulichen, sei ein klassisches Regressionsmodell mit $p = 2$ Parametern für $n = 4$ Beobachtungen betrachtet. Die Beobachtungen für die erklärende Variable seien $x_1 = 0, x_2 = 1, x_3 = 2$ und $x_4 = 5$. Abbildung 3.18 zeigt die Dichtefunktionen der nicht kalibrierten UPR für die zweite und die vierte Beobachtung. Es ist zu sehen, dass bei Verwendung des internen Varianzschätzers (linke Graphiken) die Dichtefunktion für bestimmte Bereiche des Intervalls $[0, 1]$ null ist. Allgemein gilt bei Verwendung des internen Schätzers

$$\hat{U}_i^{\text{int}} \in \left[\Phi \left(-\sqrt{(n-p)(1-h_{ii})} \right), \Phi \left(\sqrt{(n-p)(1-h_{ii})} \right) \right].$$

Ferner ist zu bemerken, dass, sowohl bei der Verwendung des internen als auch bei Verwendung des externen Varianzschätzers, die Abweichung von der $U(0, 1)$ -Verteilung für $x_4 = 5$ größer ist als für $x_2 = 1$. Dies liegt daran, dass x_4 im Vergleich zu den übrigen Beobachtungen der erklärenden Variablen relativ isoliert (bzw. extrem in der Verteilung der erklärenden Variablen) ist. Allgemein führt eine extreme Beobachtung x_i für die erklärende Variable dazu, dass $\hat{\mu}_i$ relativ dicht bei y_i liegt und einen relativ großen Einfluss auf die Parameterschätzung besitzt, so dass die Verteilung von \hat{U}_4 im hier betrachteten Beispiel vergleichsweise schmal ist. (Für diese Hebelwirkung sind in der englischen Sprache die Begriffe ‚leverage‘ oder auch ‚influence‘ etabliert.)

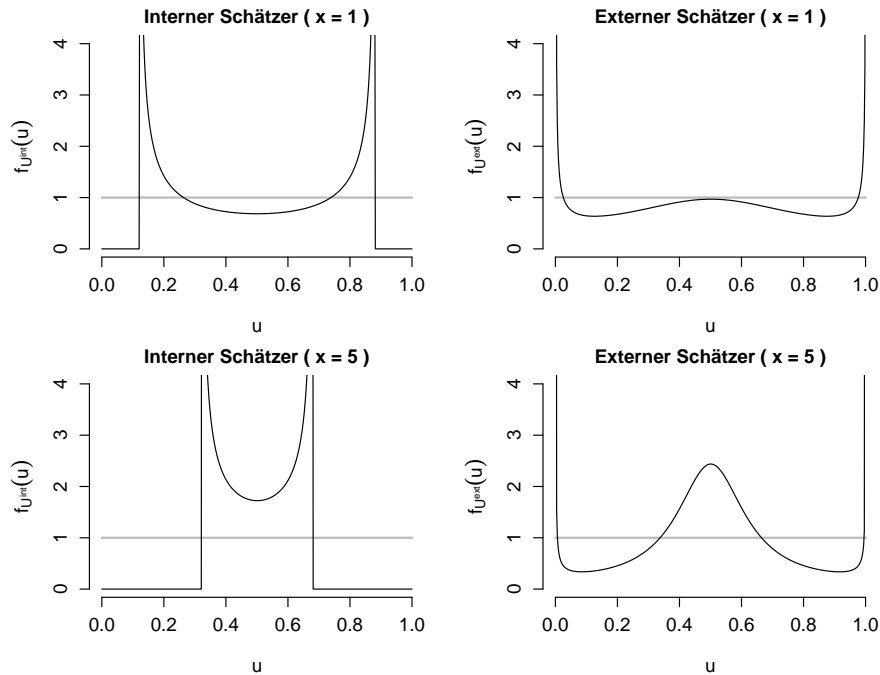


Abbildung 3.18: Dichtefunktion der geschätzten UPR bei Verwendung des internen (links) und externen (rechts) Varianzschätzers für $x = 1$ (oben) und $x = 5$ (unten).

Für spätere Zwecke sei im Rahmen dieses Beispiels gezeigt, dass die oben hergeleiteten Verteilungsfunktionen der geschätzten UPR als Komposition

$$Q_i(u) = Q_2 \circ Q_{1i}(u)$$

verstanden werden können mit

$$Q_{1i}(u) = \Phi \left(\frac{\Phi^{-1}(u)}{\sqrt{1 - h_{ii}}} \right)$$

und

$$Q_2(u) = F_W(\Phi^{-1}(u)) \quad \text{bzw.} \quad Q_2(u) = F_{t, n-p-1}(\Phi^{-1}(u))$$

bei Verwendung des internen bzw. externen Varianzschätzers. Durch die Transformation Q_{1i} erfolgt die Korrektur der UPR um den Einfluss der i -ten Beobachtung auf die Parameterschätzung. Wäre es möglich, dass eine Beobachtung keinen Einfluss auf die Parameterschätzung besitzt ($h_{ii} = 0$), wäre die erste Transformation überflüssig. Die zweite Transformation (Q_2) kann als Kalibrierung der Verteilung interpretiert werden. Sie ist einzig durch den verwendeten

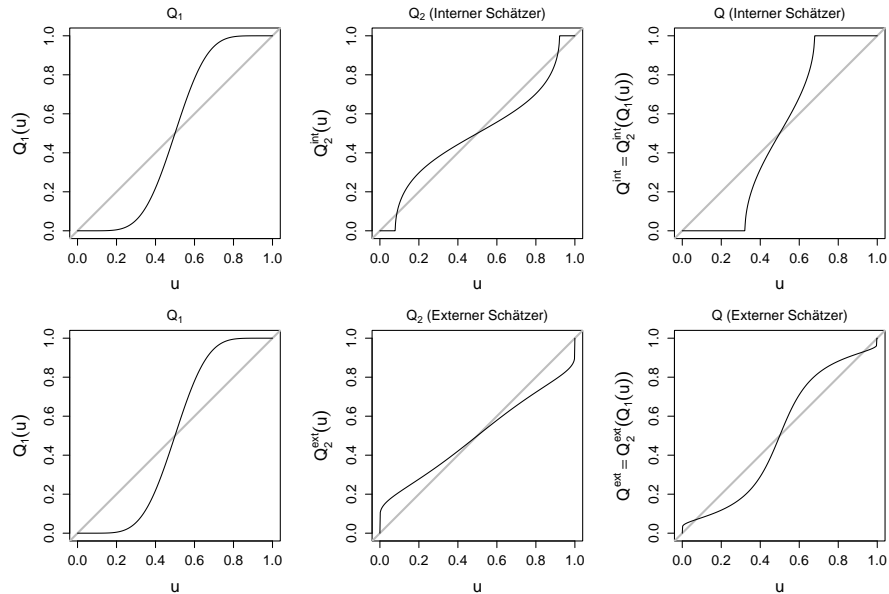


Abbildung 3.19: Dekomposition der Verteilungsfunktion der geschätzten UPR bei Verwendung des internen (oben) und externen (unten) Varianzschätzers.

Varianzschätzer bestimmt und somit unabhängig von i (bzw. vom Einfluss der Beobachtung auf die Schätzung).

Abbildung 3.19 zeigt die Bedeutung der beiden Transformationen für die extreme Beobachtung der erklärenden Variablen $x_4 = 5$. Es wird deutlich, dass, sowohl bei Verwendung des internen als auch bei Verwendung des externen Schätzers, die Korrektur um den Einfluss der Beobachtung eine größere Bedeutung besitzt als die Verteilungskorrektur. Für die anderen Beobachtungen (hier allerdings nicht dargestellt) gilt dies nicht, was auf den geringeren Einfluss dieser Beobachtungen auf die Parameterschätzung zurückzuführen ist.



Für klassische lineare Regressionsmodelle sind die Verteilungsfunktionen der geschätzten UPR nicht von den Parametern des Modells abhängig. Allgemein ist dies für verallgemeinerte lineare Modelle mit der *Hat*-Matrix

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{\frac{1}{2}}$$

jedoch nicht der Fall, weil die Diagonalmatrix \mathbf{W} mit

$$w_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

in der Regel von den Modellparametern abhängt. In diesen Fällen sind die Verteilungsfunktionen Q_i zu schätzen, bevor die UPR kalibriert werden können.

Eine Möglichkeit zur Schätzung der Verteilungsfunktionen basiert auf der Dekomposition von Q_i , die im vorangegangenen Beispiel erläutert wurde. So könnte bei der Kalibrierung auf die Verteilungskorrektur verzichtet werden, d.h.

$$Q_i(u) = Q_{1i}(u) = \Phi\left(\frac{\Phi^{-1}(u)}{\sqrt{1-h_{ii}}}\right) \quad (3.3)$$

verwendet werden. Dieses Vorgehen erscheint weniger geeignet, wenn die Verteilung der Zielvariablen schief ist, da die Verteilungskorrektur mit zunehmender Schiefe eine größere Bedeutung erlangt. Ein ähnliches Vorgehen wurde auch für die *Pearson residuals* und die *deviance residuals* vorgeschlagen (siehe beispielsweise MCCULLAGH AND NELDER, 1989, S. 397–398).

Eine zweite Möglichkeit zur Schätzung der Q_i basiert auf der Anwendung der parametrischen *Bootstrap*-Methode. Die Erzeugung des *Bootstrap*-Schätzers von Q_i kann bei stetigen Zielvariablen wie folgt durchgeführt werden:

- (1) Aus den angepassten Verteilungsfunktionen $\hat{F}_i(y)$ ($i = 1, \dots, n$) werden B Vektoren von Zufallszahlen der Zielvariablen

$$\mathbf{y}^{*b} = (y_1^{*b}, \dots, y_n^{*b})' \quad \text{für } b = 1, \dots, B$$

erzeugt, wobei B (die Anzahl der *Bootstrap*-Wiederholungen) hinreichend groß zu wählen ist.

- (2) An die Zufallszahlen der Zielvariablen wird das Modell unter Verwendung der ursprünglichen Designmatrix für jede *Bootstrap*-Wiederholung erneut angepasst. Man erhält in diesem Schritt:

$$\hat{F}_i^{*b}(y) \quad \text{für } i = 1, \dots, n \text{ und } b = 1, \dots, B.$$

- (3) Die Verteilungsfunktionen $\hat{F}_i^{*b}(y)$ werden verwendet, um für jede *Bootstrap*-Wiederholung den Vektor der UPR zu bestimmen:

$$\begin{aligned} \mathbf{u}^{*b} &= (\hat{F}_1^{*b}(y_1^{*b}), \dots, \hat{F}_n^{*b}(y_n^{*b}))' \\ &= (u_1^{*b}, \dots, u_n^{*b})' \quad \text{für } b = 1, \dots, B. \end{aligned}$$

- (4) Der *Bootstrap*-Schätzer der Verteilung von \hat{U}_i ist durch die Werte

$$u_i^{*b} \quad \text{für } b = 1, \dots, B$$

bestimmt und als Schätzer der Verteilungsfunktion kann beispielsweise die empirische Verteilungsfunktion verwendet werden.

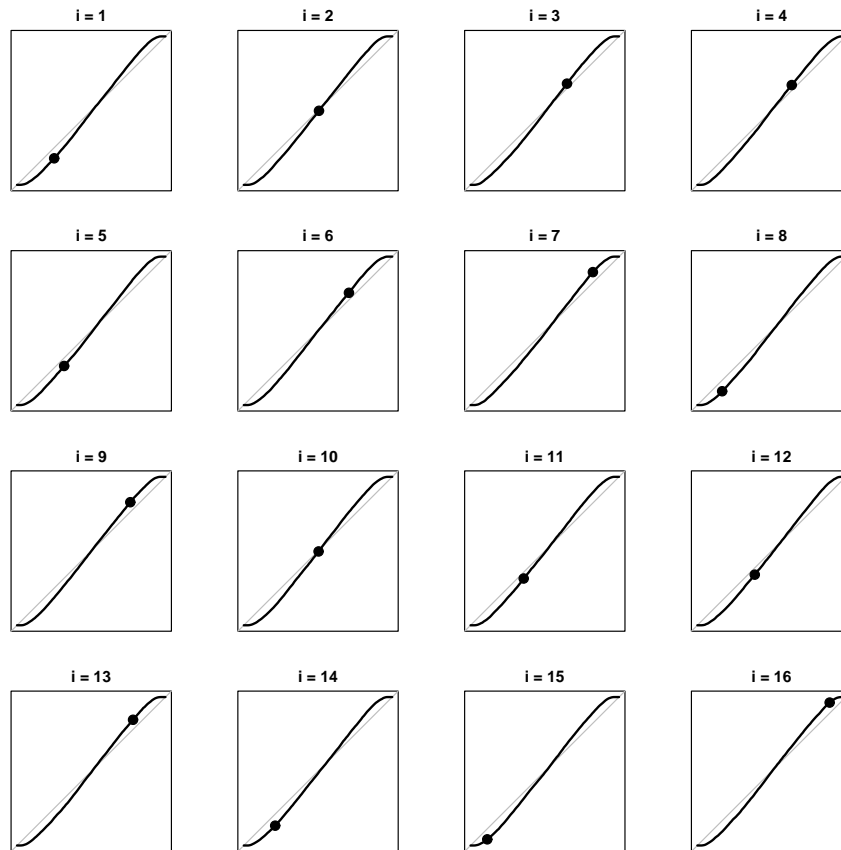


Abbildung 3.20: *Bootstrap*-Schätzer der Verteilungsfunktionen der geschätzten UPR. (Abszisse und Ordinate sind jeweils durch null und eins begrenzt.)

Die Idee, die parametrischen *Bootstrap*-Methode zur Schätzung der Verteilung der Residuen zu verwenden, wurde im Zusammenhang mit der Überprüfung logistischer Regressionsmodelle bei LANDWEHR, PREGIBON AND SHOEMAKER (1984) vorgestellt.

Beispiel 3.10 Um die Verwendung der *Bootstrap*-Methode zur Kalibrierung der Pseudo-Residuen zu verdeutlichen, wird auf das Beispiel 3.2 zurückgegriffen, in dem der Widerstand von Halbleitern durch vier Faktoren erklärt wurde. Bei der Modellanpassung wurden aus sechzehn Beobachtungen acht Parameter geschätzt. Die Betrachtung des QQ-Plots der NPR sowie der NPR des QQ-Plots zeigte, dass die Streuung der Residuen etwas geringer ist, als es unter dem Modell zu erwarten wäre. Es wurde allerdings darauf hingewiesen, dass dies an der Nichtberücksichtigung der Parameterschätzung liegen kann. Daher werden nun mit Hilfe der *Bootstrap*-Methode die Verteilungsfunktionen

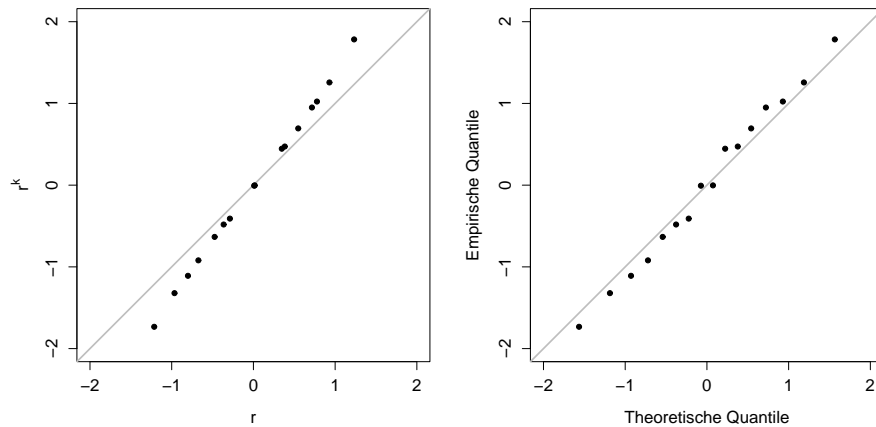


Abbildung 3.21: Plot der ursprünglichen NPR gegen die kalibrierten NPR (links) und QQ-Plot der kalibrierten NPR (rechts).

der geschätzten UPR geschätzt, um diese anschließend zur Kalibrierung der Residuen zu verwenden.

Abbildung 3.20 zeigt die Schätzer der Verteilungsfunktionen, wobei für dieses Beispiel 10 000 *Bootstrap*-Wiederholungen durchgeführt wurden. Es ist zu sehen, dass sich die Transformationen für die unterschiedlichen Beobachtungen nur unwesentlich unterscheiden. Dies liegt daran, dass es im hier betrachteten Design keine extremen Beobachtungen der erklärenden Variablen gibt und der Einfluss der Beobachtungen auf die Parameterschätzung somit für alle Beobachtungen in etwa gleich ist. Ferner ist an der Gestalt der Verteilungsfunktionen zu erkennen, dass die Verteilung der geschätzten Residuen schmäler ist als eine $U(0, 1)$ -Verteilung. Daher sind die kalibrierten Residuen extremer als die ursprünglichen Residuen. Dies ist in der linken Graphik der Abbildung 3.21 dargestellt, in der ursprüngliche und kalibrierte NPR gegeneinander geplottet sind. Die rechte Graphik zeigt einen QQ-Plot der kalibrierten NPR. Dieser verläuft in etwa entlang der Winkelhalbierenden, d.h. die ursprünglich beobachteten Abweichungen von der Winkelhalbierenden sind vermutlich auf die Nichtberücksichtigung der Parameterschätzung zurückzuführen und nicht auf ein fehlerhaftes Modell.

◀

Auch bei diskreten Zielvariablen können die geschätzten Residuen kalibriert werden. Da die Pseudo-Residuen im diskreten Fall in Form von Intervallen definiert sind, kann das oben beschriebene Vorgehen allerdings nicht di-

rekt verwendet werden. Stattdessen wird hier ein Vorgehen erläutert, das auf zwei Schritten basiert. In einem ersten Schritt werden die Mittelpunkte der geschätzten UPR-Intervalle

$$\hat{U}_i^m = \frac{\hat{U}_i^+ + \hat{U}_i^-}{2}$$

kalibriert, wobei die Kalibrierung auf die gleiche Weise erfolgt wie im stetigen Fall für die \hat{U}_i . Sind die Mittelpunkte kalibriert, müssen ihnen in einem zweiten Schritt geeignete Intervallgrenzen zugeordnet werden, wobei der Begriff ‚geeignet‘ im Rahmen des nachfolgenden Beispiels erläutert wird.

Beispiel 3.11 Zur Verdeutlichung der Kalibrierung bei diskreten Zielvariablen wird das einführende Beispiel (Beispiel 3.1) zu den Maschinenausfällen erneut betrachtet. Dort wurden aus neun Beobachtungen drei Parameter geschätzt: Der Achsenabschnitt und zwei Koeffizienten für die Laufzeiten der Maschine in unterschiedlichen Betriebsarten.

Die simulierte Verteilung der Mittelpunkte der UPR-Intervalle ist in Abbildung 3.22 dargestellt. Die Histogramme für die neun Beobachtungen basieren jeweils auf 5 000 *Bootstrap*-Wiederholungen. Sie geben einen ersten Eindruck von der Streuung der geschätzten UPR-Intervalle. Beispielsweise ist zu sehen, dass die Mittelpunkte für $i = 4$ eine vergleichsweise schmale Verteilung besitzen und für $i = 9$ in etwa eine $U(0, 1)$ -Verteilung vorliegt. Die Ursache für die deutlichen Unterschiede zwischen den Verteilungen ist der unterschiedliche Einfluss der erklärenden Variablen auf die Parameterschätzung. Abbildung 3.23 zeigt die Beobachtungen der erklärenden Variablen in einem Scatterplot, wobei die Punkte des Plots durch den Index der Beobachtung ergänzt sind. Hier ist zu sehen, dass die Beobachtung der erklärenden Variablen für $i = 4$ am äußeren Rand und für $i = 9$ im Zentrum der gemeinsamen Verteilung von X_1 und X_2 liegt. Dies führt zu der oben beschriebenen Hebelwirkung der vierten Beobachtung: Durch den ausgeprägten Einfluss dieser Beobachtung auf die Parameterschätzung ist der Unterschied zwischen $\hat{\mu}_4^*$ und y_4^* bei den *Bootstrap*-Wiederholungen tendenziell klein und die Folge ist eine relativ schmale Verteilung der Mittelpunkte der UPR-Intervalle.

Die Punkte in den Histogrammen der Abbildung 3.22 sind die aus der Originalstichprobe bestimmten Mittelpunkte der UPR-Intervalle. Deren Lage in der *Bootstrap*-Verteilung bestimmt die kalibrierten Mittelpunkte. Hier wurden (wie im stetigen Fall) die empirischen Verteilungsfunktionen der *Bootstrap*-Verteilungen verwendet, um im ersten Schritt die Mittelpunkte der UPR-Intervalle zu kalibrieren.

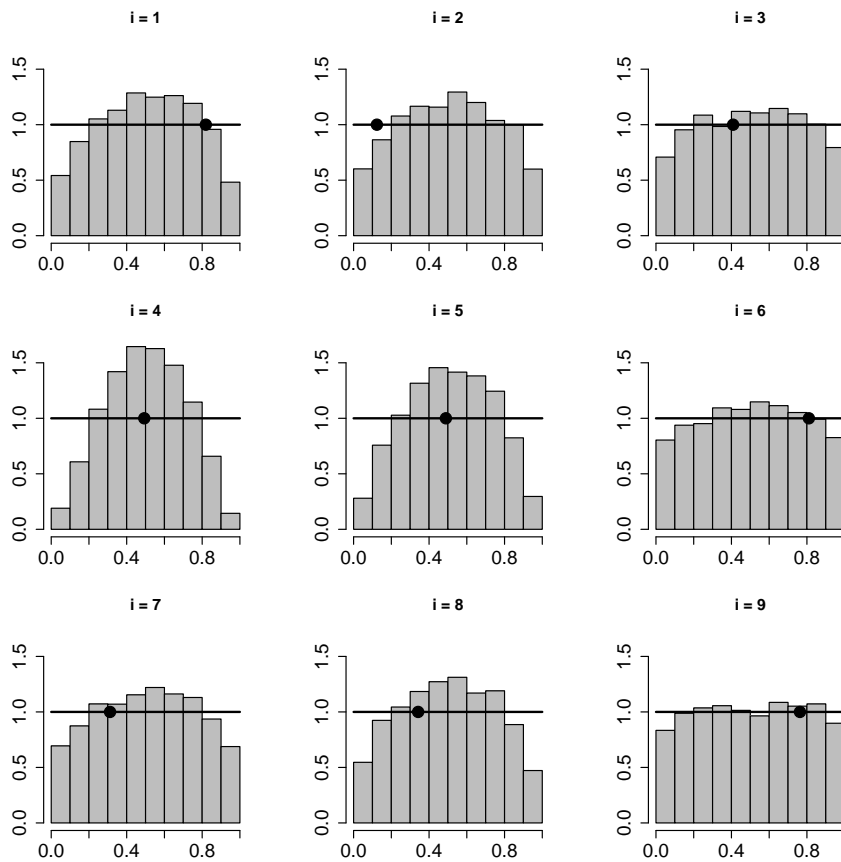


Abbildung 3.22: *Bootstrap*-Schätzer der Verteilungen der UPR-Intervallmittelpunkte.

Im zweiten Schritt müssen jedem kalibrierten Mittelpunkt geeignete Intervallgrenzen zugeordnet werden. Die hier erläuterte Zuordnung basiert auf einem Zusammenhang zwischen den Mittelpunkten und den Intervallgrenzen, der aus der *Bootstrap*-Verteilung abgeleitet wird. Abbildung 3.24 zeigt diesen Zusammenhang für die neun Beobachtungen des Beispiels. Jeweils wurden die Intervallmittelpunkte der *Bootstrap*-Verteilung gegen die untere und die obere Intervallgrenze geplottet. Um für gegebenes i die Intervallgrenzen zu ermitteln, wird zunächst der Mittelpunkt aus der *Bootstrap*-Verteilung bestimmt, der dem kalibrierten Mittelpunkt am nächsten ist. Ist dieser Mittelpunkt der *Bootstrap*-Verteilung identifiziert, werden dessen Intervallgrenzen dem kalibrierten Mittelpunkt zugeordnet. Alternativ hätten die Intervallgrenzen um den Betrag verschoben werden können, um den auch der Mittelpunkt bei der Kalibrierung verschoben wurde. Das hier vorgeschlagene Vorgehen besitzt je-

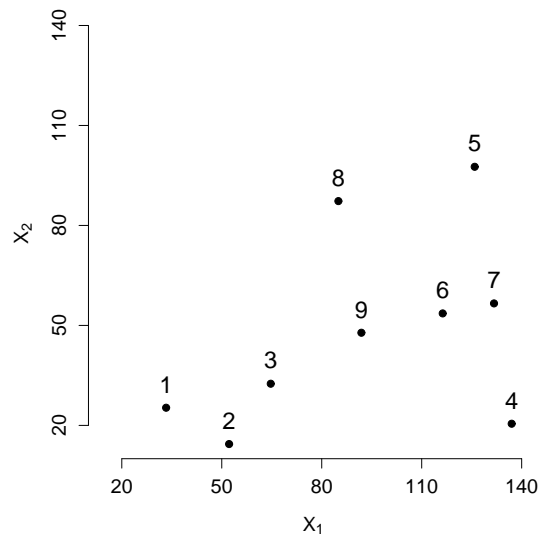


Abbildung 3.23: Scatterplot für die Beobachtungen der beiden erklärenden Variablen.

doch den Vorteil, dass die kalibrierten Intervallgrenzen nicht kleiner als null bzw. größer als eins werden können.

Abbildung 3.25 verdeutlicht die Ergebnisse der Kalibrierung. Sie zeigt in der linken Graphik einen Plot der geschätzten NPR-Intervalle $[\hat{r}_i^-, \hat{r}_i^+]$ gegen die kalibrierten NPR-Intervalle $[r_i^{k-}, r_i^{k+}]$ (die Schnittpunkte sind jeweils die entsprechenden Mittelpunkte der Intervalle). Analog zum stetigen Fall ist zu erkennen, dass die kalibrierten Residuen tendenziell extremer sind als die ursprünglich berechneten Residuen. In diesem Zusammenhang ist zu bemerken, dass eine relativ schmale Verteilung der geschätzten Residuen nicht zwangsläufig eine große Veränderung durch die Kalibrierung impliziert. So bleibt das NPR-Intervall der vierten Beobachtung von der Kalibrierung nahezu unberührt, weil der Mittelpunkt des ursprünglichen UPR-Intervalls in etwa 0.5 ist. Auch ist klar, dass die Veränderung nur geringfügig ist, wenn die Verteilung der Mittelpunkte in etwa einer $U(0, 1)$ -Verteilung entspricht, wie dies für $i = 9$ der Fall ist. Die Veränderung durch die Kalibrierung ist nur dann von größerem Ausmaß, wenn sowohl der ursprüngliche Mittelpunkt des UPR-Intervalls relativ nah bei null oder eins liegt als auch die Verteilung der geschätzten Residuen relativ schmal ist. Dies ist beispielsweise für die zweite Beobachtung der Fall.

Die rechte Graphik in Abbildung 3.25 zeigt einen QQ-Plot der kalibrierten

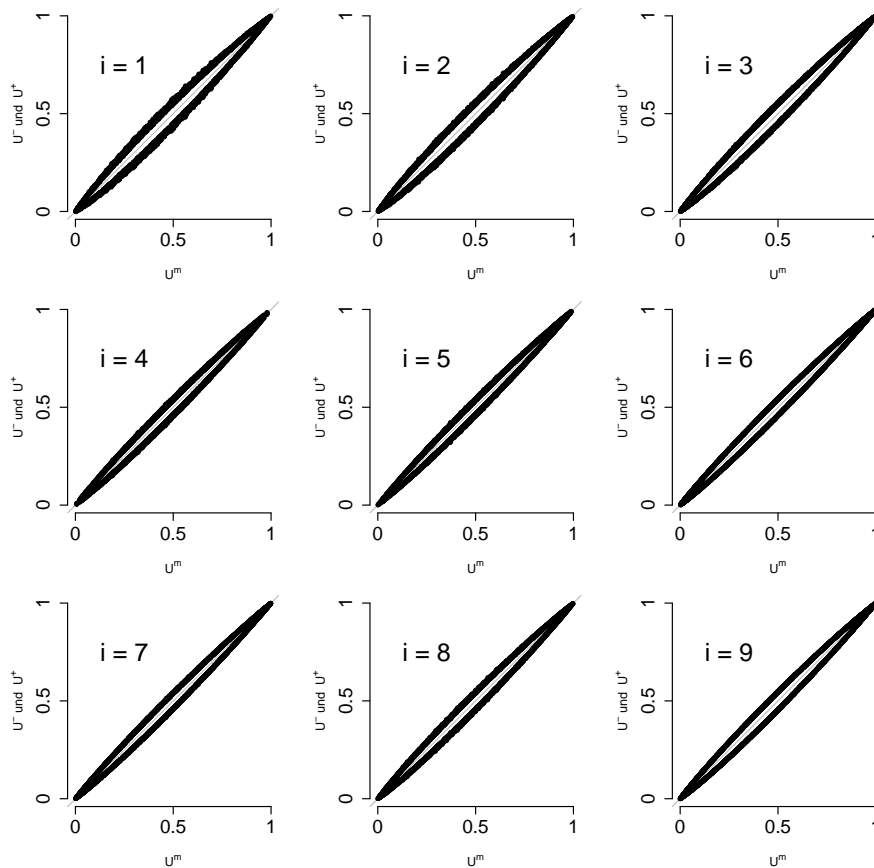


Abbildung 3.24: Plot der Mittelpunkte der UPR-Intervalle der *Bootstrap*-Verteilung gegen die entsprechenden Intervallgrenzen.

NPR-Intervalle. Hier sind keine Hinweise auf eine schlechte Modellanpassung zu erkennen, d.h. die NPR-Intervalle verlaufen in etwa auf der Winkelhalbierenden. Insbesondere schneidet das NPR-Intervall der ersten Beobachtung nach der Kalibrierung die Winkelhalbierende, während das nicht kalibrierte Intervall unterhalb der Winkelhalbierenden liegt (siehe Abbildung 3.6).



3.5 Überprüfung von Modellen für spezielle Problemstellungen

In diesem Abschnitt werden drei Problemstellungen betrachtet, in denen sich aus der Struktur der Daten besondere Anforderungen an die Modellüber-

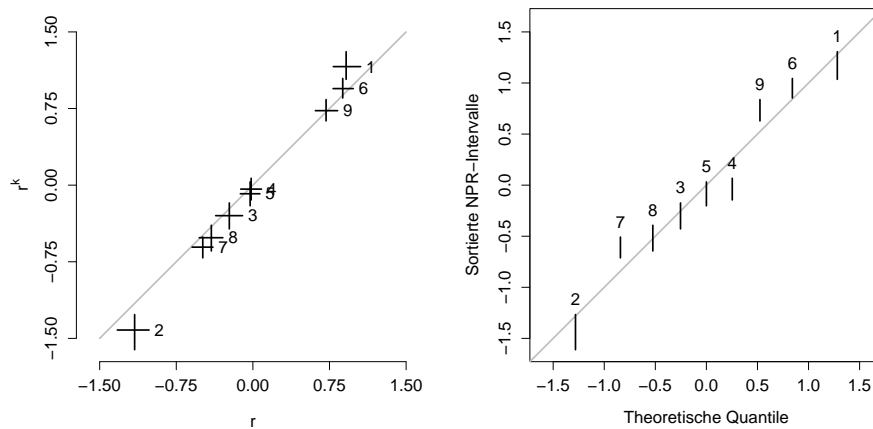


Abbildung 3.25: Plot der ursprünglichen NPR gegen die kalibrierten NPR (links) und QQ-Plot der kalibrierten NPR (rechts).

prüfung ergeben. Im Einzelnen werden Modelle für wiederholte Messungen, loglineare Modelle und Modelle zur Beschreibung von zyklischen Daten betrachtet. Jeweils wird anhand eines Beispiels beschrieben, welche Besonderheiten bei der Modellierung zu beachten sind und gezeigt, welche Ansätze zur Überprüfung der Modelle geeignet sind.

3.5.1 Wiederholte Messung binärer Variablen

Die Daten, die in diesem Abschnitt betrachtet werden, unterscheiden sich in ihrer grundlegenden Struktur von den bislang betrachteten Datensätzen. Es werden hier wiederholte Messungen an identischen Untersuchungseinheiten betrachtet, d.h. die Beobachtungen sind nicht unabhängig. Ferner werden simultan mehrere Untersuchungseinheiten betrachtet, d.h. die Daten haben nicht die Form einer einfachen Zeitreihe. In der englischen Sprache ist für diese Datenstruktur der Begriff ‚*Repeated Measurements*‘ etabliert (vgl. LINDSEY, 1993, S. 3).

Die Daten der Tabelle 3.11 stammen von einem Experiment, das zur Untersuchung eines Lernprozesses von Hunden diente und bei BUSH AND MOSTELLER (1955, S. 238–239) genauer beschrieben ist. In dem Experiment wurden mit 30 Hunden jeweils 25 aufeinanderfolgende Versuche durchgeführt: Die Hunde befanden sich in einem Käfig und der Versuch bestand darin, dass ein Licht ausgeschaltet wurde sowie zehn Sekunden später ein elektrischer Schlag durch

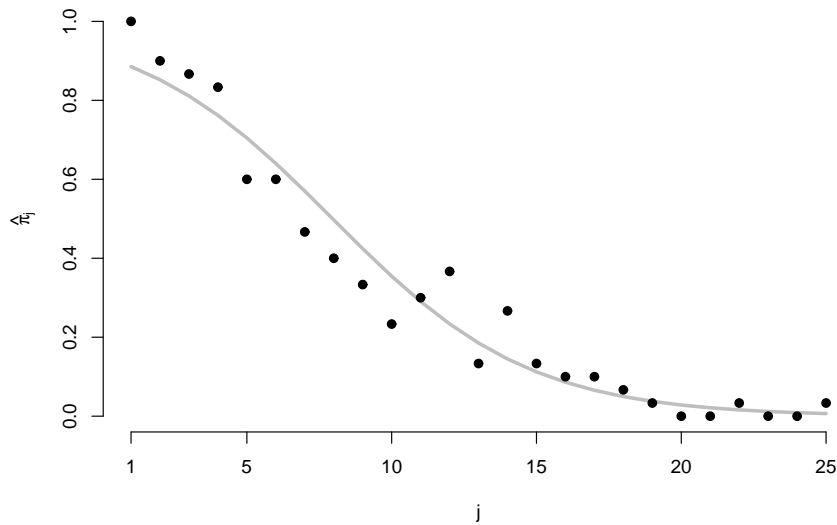


Abbildung 3.26: Relative Häufigkeiten der Stromstöße in den einzelnen Versuchen und angepasstes Modell 1.

den Käfig geschickt wurde. Die Hunde hatten in jedem Versuch die Möglichkeit, in einen benachbarten Käfig zu springen, um den Stromstoß zu vermeiden. Für die Beobachtung y_{ij} (die Beobachtung für den i -ten Hund in Versuch j) wurde der Wert $Y_{ij} = 1$ notiert, wenn der Stromstoß den Hund erreichte und $Y_{ij} = 0$, wenn der Stromstoß vermieden werden konnte. Für eine bessere Übersicht sind die Zeilen der Tabelle 3.11 nach dem Zeitpunkt des letzten Stromstoßes sortiert.

Im Folgenden werden vier Modelle zur Beschreibung des Experiments untersucht. Dabei sind die beiden ersten Modelle allerdings nicht als ernsthafte Modelle zu verstehen. Vielmehr verdeutlichen sie, welche Fehler bei der Modellierung von wiederholten Messungen gemacht werden können und wie diese mit Hilfe von Pseudo-Residuen zu identifizieren sind. Anschließend werden die beschriebenen Methoden verwendet, um zwei in der Literatur betrachtete Modelle zu überprüfen.

Ein naives Modell zur Beschreibung der Daten ergibt sich bei Betrachtung der Spaltensummen, d.h. bei Betrachtung der Gesamtzahl der Stromstöße in den einzelnen Versuchen. Unterstellt man für jeden Hund in Versuch j die gleiche Erfolgswahrscheinlichkeit π_j und setzt man ferner voraus, dass sich die Hunde unabhängig verhalten, ist die Summe der Stromstöße in Versuch j binomialverteilt mit den Parametern $n = 30$ und π_j . Wählt man als erklärende

i	j				
	1-5	6-10	11-15	16-20	21-25
1	1 1 1 1 0	1 0 0 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
2	1 1 0 1 0	1 0 0 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
3	1 1 1 1 1	1 1 0 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
4	1 1 1 1 1	1 1 0 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
5	1 1 1 0 1	0 0 1 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
6	1 1 1 1 0	1 0 1 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
7	1 1 1 1 1	1 1 1 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
8	1 1 1 0 1	0 0 1 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
9	1 1 1 0 0	1 0 1 1 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
10	1 1 1 0 0	0 0 0 1 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
11	1 0 1 1 1	1 0 1 1 1	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
12	1 1 1 1 1	1 1 0 0 1	1 0 0 0 0	0 0 0 0 0	0 0 0 0 0
13	1 1 0 1 0	0 0 1 0 0	1 0 0 0 0	0 0 0 0 0	0 0 0 0 0
14	1 1 1 1 0	1 1 0 0 1	0 1 0 0 0	0 0 0 0 0	0 0 0 0 0
15	1 1 1 1 1	0 0 0 0 0	1 1 0 0 0	0 0 0 0 0	0 0 0 0 0
16	1 1 1 1 1	0 1 1 1 1	1 1 0 0 0	0 0 0 0 0	0 0 0 0 0
17	1 1 1 0 1	0 1 0 0 0	1 0 1 0 0	0 0 0 0 0	0 0 0 0 0
18	1 1 1 1 0	0 0 0 0 0	1 0 1 0 0	0 0 0 0 0	0 0 0 0 0
19	1 1 1 1 1	1 0 0 0 0	1 1 0 1 0	0 0 0 0 0	0 0 0 0 0
20	1 0 1 0 1	1 1 0 1 0	0 0 0 1 0	0 0 0 0 0	0 0 0 0 0
21	1 0 0 1 1	0 0 0 0 1	0 1 0 1 0	0 0 0 0 0	0 0 0 0 0
22	1 1 1 1 0	0 1 0 1 0	0 1 0 1 0	0 0 0 0 0	0 0 0 0 0
23	1 1 1 1 1	1 1 0 1 1	1 1 1 1 0	0 0 0 0 0	0 0 0 0 0
24	1 1 1 1 1	1 0 1 0 0	0 0 1 0 1	0 0 0 0 0	0 0 0 0 0
25	1 1 1 1 1	1 1 1 1 1	0 0 0 0 0	0 1 0 0 0	0 0 0 0 0
26	1 1 1 1 1	0 0 1 0 0	1 1 0 0 1	0 1 0 0 0	0 0 0 0 0
27	1 1 0 1 0	1 0 0 0 0	0 0 0 0 0	0 1 1 0 0	0 0 0 0 0
28	1 1 1 1 0	0 1 1 0 0	0 1 0 1 0	1 0 1 0 0	0 0 0 0 0
29	1 1 1 1 1	1 1 1 0 0	0 1 0 1 1	1 0 0 1 0	0 0 0 0 0
30	1 1 1 1 1	0 1 0 1 0	0 1 0 1 1	1 0 0 0 0	0 1 0 0 1

Tabelle 3.11: Beobachtungen für das Hundexperiment mit 25 Versuchen (Spalten) für jeden der 30 Hunde (Zeilen).

Variable die Zeit, ist

$$\begin{aligned} \text{logit}(\pi_j) &= \beta_0 + \beta_1 j \\ &\text{für } j = 1, \dots, 25 \end{aligned}$$

ein mögliches Modell (im Folgenden als Modell 1 bezeichnet) zur Beschreibung des Lernprozesses. Abbildung 3.26 zeigt die relativen Häufigkeiten der Hunde, die einen Stromstoß erhalten haben, für die einzelnen Versuche sowie das angepasste Modell (die Parameterschätzer sind $\hat{\beta}_0 = 2.339$ und $\hat{\beta}_1 = -0.294$).

Das Modell 1 berücksichtigt keine Unterschiede zwischen den Hunden, d.h. die Modellierung unterschiedlicher Lernprozesse bzw. Lernerfolge bei den Hunden ist mit dem Modell nicht möglich. Wenn solche Unterschiede existieren, liegt es nah, dass die geschätzten Wahrscheinlichkeiten $\hat{\pi}_j$ für einige Hunde tendenziell zu groß und für andere tendenziell zu klein sind. Basierend auf diesem

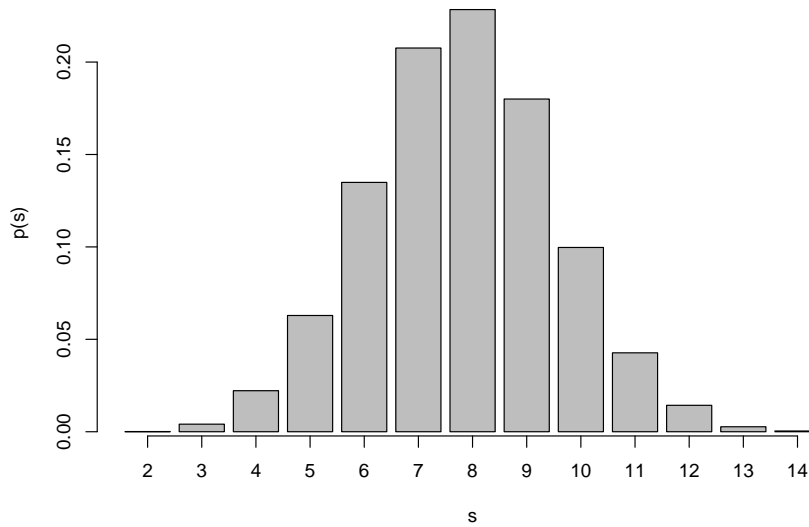


Abbildung 3.27: Verteilung der Summe der Stromstöße eines einzelnen Hundes unter Modell 1.

Gedanken kann für den einzelnen Hund ein Pseudo-Residual definiert werden, dass dessen Lernleistung über alle Versuche widerspiegelt: Betrachtet sei die Verteilung der Gesamtzahl der Stromstöße, die ein einzelner Hund in den 25 Versuchen erleidet, d.h. die Verteilung der Zeilensummen

$$S_i = \sum_{j=1}^{25} Y_{ij} \quad \text{für } i = 1, \dots, 30.$$

Da die Erfolgswahrscheinlichkeit in einem bestimmten Versuch für alle Hunde identisch ist, ist die Verteilung der S_i ebenfalls für alle Hunde identisch. Auf abstrakter Ebene handelt es sich hier um die Verteilung der Summe von 25 unabhängigen Bernoulli-Experimenten mit unterschiedlichen Erfolgswahrscheinlichkeiten, d.h. um eine sogenannte Poisson Binomial-Verteilung (vgl. WANG, 1993). Mit Hilfe dieser Verteilungen kann für jeden Hund ein NPR-Intervall

$$R_i = [\Phi^{-1}(F_{S_i}(S_i - 1)), \Phi^{-1}(F_{S_i}(S_i))] \quad \text{für } i = 1, \dots, 30$$

definiert werden. Es beschreibt, wie extrem die Summe der Stromstöße eines Hundes unter dem Modell ist. Bei Gültigkeit des Modells verhalten sich die NPR-Intervalle bekanntlich wie intervallzensierte Beobachtungen einer Standardnormalverteilung. Wenn sich die Hunde allerdings unterscheiden, sind tendenziell extremere Residuen zu erwarten, weil die geschätzten Erfolgswahr-

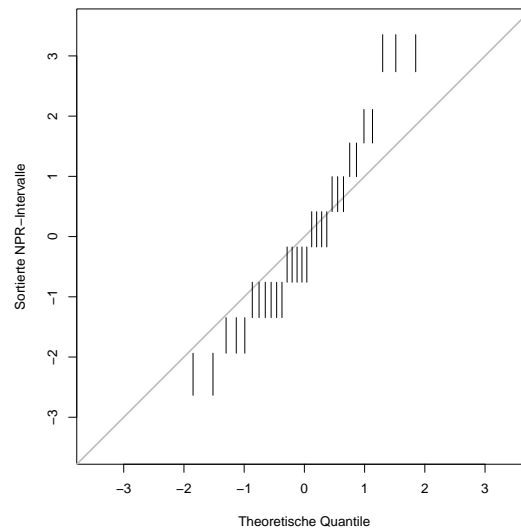


Abbildung 3.28: QQ-Plot der NPR-Intervalle für die Summe der Stromstöße der einzelnen Hunde unter Modell 1.

scheinlichkeiten für einige Hunde tendenziell zu klein und für andere tendenziell zu groß sind.

Da die explizite Bestimmung der Wahrscheinlichkeiten für eine Poisson Binomial-Verteilung sehr aufwendig ist, wurde die Verteilung der S_i simuliert. Insgesamt wurden 10 000 Realisationen von S_i nach Maßgabe des angepassten Modells erzeugt. Die so approximierte Verteilung (Abbildung 3.27 zeigt die entsprechende Wahrscheinlichkeitsfunktion) wurde anschließend zur Berechnung der NPR-Intervalle verwendet. Abbildung 3.28 zeigt den QQ-Plot der NPR-Intervalle für die 30 Hunde. Die Verteilung ist breiter als unter dem Modell erwartet, d.h. die Residuen sind tendenziell extremer. Die vermuteten Unterschiede zwischen den Hunden existieren also tatsächlich und können bei der Modellierung nicht vernachlässigt werden.

Die Fehlerquelle des Modells 1 ist die unzureichende Berücksichtigung von hundespezifischen Informationen bei der Schätzung der Erfolgswahrscheinlichkeiten: Die Information, ob ein Hund im Versuch j einen Stromstoß erhalten hat oder nicht, besitzt zwar einen Einfluss auf die Zielvariable (also die Summe der Stromstöße über alle Hunde), jedoch wird die individuelle Information nicht genutzt, um individuelle Erfolgswahrscheinlichkeiten zu schätzen. Das andere Extrem, d.h. die Schätzung der Erfolgswahrscheinlichkeiten ausschließ-

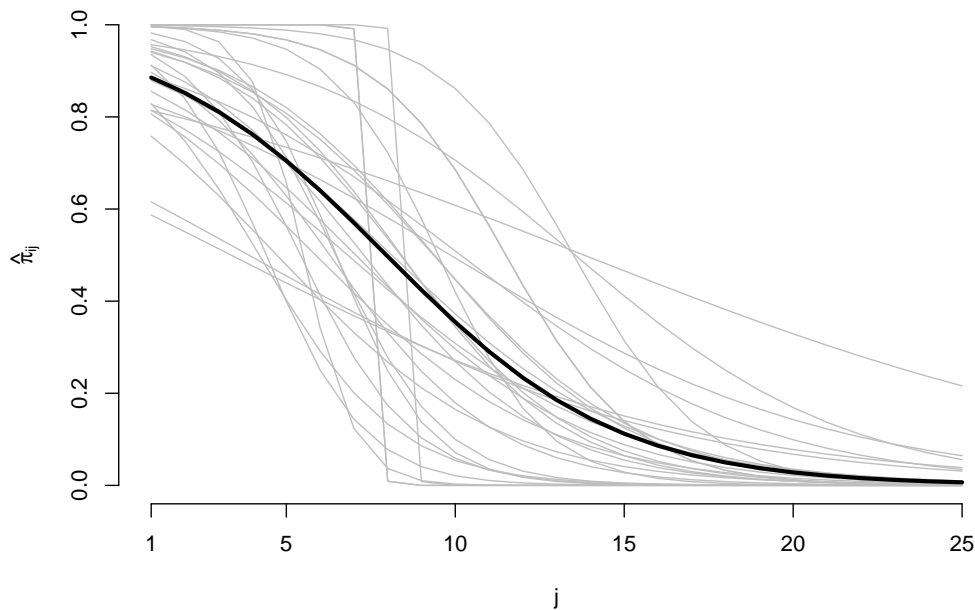


Abbildung 3.29: Individuelle Lernkurven der Hunde nach Modell 2.

lich unter Verwendung von hundespezifischen Informationen, ist in Modell 2

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \beta_{0i} + \beta_{1i}j \\ &\text{für } i = 1, \dots, 30 \text{ und } j = 1, \dots, 25 \end{aligned}$$

verkörpert. Dieses Modell besitzt zwei Parameter für jeden Hund (also insgesamt 60 Parameter), die jeweils nur aus Informationen über den entsprechenden Hund geschätzt werden. Die grauen Kurven in Abbildung 3.29 zeigen die individuellen Lernkurven der Hunde nach Modell 2. (Ergänzend ist die gemeinsame Lernkurve nach Modell 1 eingezeichnet.) Die Unterschiede zwischen individuellen Lernkurven bestätigen, dass die individuellen Eigenschaften der Hunde bei der Modellierung berücksichtigt werden sollten.

Auf der anderen Seite müssen aber auch die gemeinsamen Charakteristika der einzelnen Zeitreihen in die Modellbildung einbezogen werden. Schließlich ist es das Ziel, den Lernprozess allgemein zu beschreiben und nicht die Lernprozesse der einzelnen Hunde. Die Folge einer Vernachlässigung der gemeinsamen Eigenschaften kann anhand der P-Werte der Signifikanztests bezüglich der Koeffizienten β_{1i} quantifiziert werden. Betrachtet man die Nullhypothesen, dass die Hunde im Zeitverlauf nicht lernen ($H_i : \beta_{1i} = 0$), ergibt sich die in Abbildung 3.30 dargestellte Verteilung der P-Werte. (Drei Werte am rechten Rand der

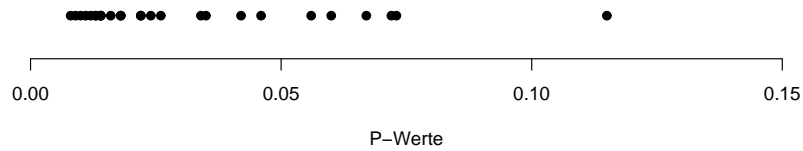


Abbildung 3.30: Verteilung der P–Werte für die Hypothese, dass die Hunde nicht lernen.

Verteilung sind nicht in die Graphik aufgenommen.) Bei gemeinsamer Betrachtung der P–Werte ist die Nullhypothese wenig plausibel, da die Verteilung der P–Werte deutlich nach links verschoben ist. Andererseits sieht man bei separater Betrachtung, dass die P–Werte für neun Hunde größer sind als 0.05, d.h. für diese Hunde erscheint die Nullhypothese nicht unplausibel. Ein Test des Koeffizienten β_1 aus Modell 1 kann als Zusammenfassung der Tests bezüglich der individuellen Koeffizienten des Modells 2 verstanden werden. Hier erhält man als P–Wert $4.8 \cdot 10^{-43}$, d.h. die Nullhypothese, dass die Hunde im Zeitverlauf nicht lernen, ist auszuschließen. Die Berücksichtigung der gemeinsamen Charakteristika ist insofern als Erhöhung des Stichprobenumfangs zu verstehen, die mit einer Erhöhung der Güte des Tests (besser bekannt unter der im Englischen verwendeten Bezeichnung *Power*) verbunden ist.

Die Ausführungen zu den bisher betrachteten Modellen verdeutlichen, dass es bei der Modellierung von wiederholten Beobachtungen an mehreren Untersuchungseinheiten zum einen darauf ankommt, die gemeinsamen Charakteristika der einzelnen Zeitreihen zu beschreiben und zum anderen dabei die individuellen Eigenschaften der Untersuchungseinheiten zu berücksichtigen. Während das Modell 2 die gemeinsamen Charakteristika vernachlässigt, unterstellt man bei Verwendung des Modells 1, dass die Hunde keine individuellen Eigenschaften besitzen, die sich im Lernprozess widerspiegeln. Zwei Modelle, die sowohl individuelle als auch gemeinsame Informationen bei der Schätzung der hundespezifischen Erfolgswahrscheinlichkeiten berücksichtigen, sind die Modelle 3 und 4:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \\ \log(\pi_{ij}) &= \gamma_1 x_{1ij} + \gamma_2 x_{2ij} \\ &\text{jeweils für } i = 1, \dots, 30 \text{ und } j = 1, \dots, 25. \end{aligned}$$

In beiden Modellen beschreiben

$$x_{1ij} = \sum_{t=1}^{j-1} (1 - y_{it}) \quad \text{bzw.} \quad x_{2ij} = \sum_{t=1}^{j-1} (y_{it})$$

die Summe der von Hund i vor dem Versuch j vermiedenen bzw. erhaltenen Stromstöße. Die Modelle basieren demnach auf der Idee, dass die Hunde sowohl durch den Erhalt als auch durch die Vermeidung eines Stromstoßes lernen. Die Unterschiede zwischen den Modellen liegen zum einen in der verwendeten Linkfunktion und zum anderen in der Aufnahme bzw. Vernachlässigung eines Achsenabschnitts. Die Verwendung des Logitlinks (Modell 3) erscheint bei der Modellierung von Erfolgswahrscheinlichkeiten naheliegend, da er gewährleistet, dass die Schätzer der π_{ij} zwischen null und eins liegen. Auf der anderen Seite weisen GELMAN ET AL. (2000) darauf hin, dass die Erfolgswahrscheinlichkeiten des Modells 3 für keine der beiden möglichen Realisationen (also $Y_{ij} = 0$ und $Y_{ij} = 1$) eins werden kann. Dies ist bei dem hier betrachteten Experiment nachteilig, da dessen Aufbau einen Stromstoß im ersten Versuch impliziert. Eine einfache Möglichkeit, das Problem zu umgehen, ist die Anpassung des Modells an die Daten ohne Berücksichtigung der Beobachtungen im ersten Versuch. Eine andere Möglichkeit ist die Verwendung des Modells 4, aus dem man für jeden Hund $\hat{\pi}_{i1} = 1$ berechnet. Dieser bereits bei BUSH AND MOSTELLER (1955, S. 240–241) beschriebene Modellierungsansatz hat weiterhin den Vorteil, dass die Modellparameter eine im Hinblick auf die Problemstellung aussagekräftige Interpretation besitzen. Es gilt

$$\pi_{ij} = \begin{cases} \pi_{i(j-1)} \cdot e^{\gamma_1} & \text{für } y_{i(j-1)} = 0 \\ \pi_{i(j-1)} \cdot e^{\gamma_2} & \text{für } y_{i(j-1)} = 1. \end{cases}$$

Die Wahrscheinlichkeit für einen Stromstoß verringert sich also im Versuch j um den Faktor e^{γ_1} , wenn im Versuch $j-1$ der Stromstoß vermieden wurde bzw. um den Faktor e^{γ_2} , wenn ein Hund den Stromstoß erhalten hat (vgl. LINDSEY, 1993, S. 197).

Die Lernkurven der angepassten Modelle sind in Abbildung 3.31 dargestellt. (Die Parameterschätzer des Modells 3 sind $\hat{\beta}_0 = 1.790$, $\hat{\beta}_1 = -0.355$ und $\hat{\beta}_2 = -0.209$, die des Modells 2 sind $\hat{\gamma}_1 = -0.236$ und $\hat{\gamma}_2 = -0.0793$; letztere sind übernommen von LINDSEY, 1997, S. 197). Beide Modelle können als Kompromiss zwischen den Modellen 1 und 2 verstanden werden:

- Beide Modelle berücksichtigen hundespezifische Eigenschaften, da die erklärenden Variablen vom individuellen Verhalten der Hunde abhängen. Dies führt (anders als bei Modell 1) zu unterschiedlichen Lernkurven.

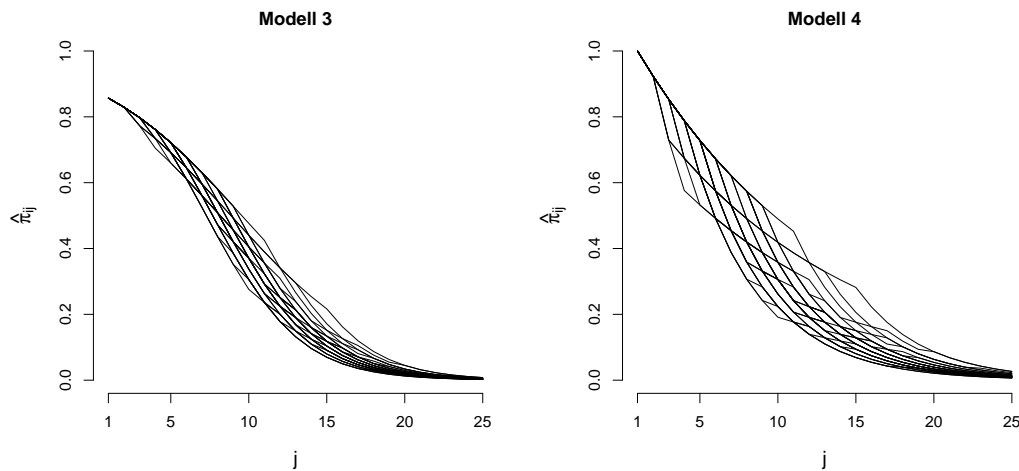


Abbildung 3.31: Lernkurven nach den Modellen 3 (links) und 4 (rechts).

- Beide Modelle berücksichtigen gemeinsame Charakteristika, da die Summe der erklärenden Variablen in den einzelnen Versuchen für alle Hunde identisch ist. Was die Beobachtungen also verbindet, ist die gemeinsame Eigenschaft ‚Zeitpunkt der Beobachtung‘. Dies führt dazu, dass sich die Lernkurven weniger stark unterscheiden als bei Modell 2.

Ein Vergleich der individuellen Lernkurven der Modelle 3 und 4 zeigt, dass sich die Lernerfolge der Hunde nach Modell 4 stärker unterscheiden als nach Modell 3 (siehe Abbildung 3.31). Dies könnte als Hinweis darauf verstanden werden, dass die individuellen Lernfähigkeiten in Modell 3 nicht ausreichend unterschieden werden, es also Hunde gibt, die tendenziell geringere bzw. höhere Erfolgswahrscheinlichkeiten besitzen als unter dem Modell erwartet. Anhand des Modells 1 wurde bereits gezeigt, wie mit Hilfe eines QQ-Plots der NPR-Intervalle für die beobachteten Zeilensummen beurteilt werden kann, ob dies der Fall ist. Die entsprechenden QQ-Plots für die Modelle 3 und 4 sind in Abbildung 3.32 zu sehen. In beiden Fällen scheinen die NPR-Intervalle etwas stärker zu streuen als unter den Modellen erwartet. Es ist allerdings kaum zu beurteilen, ob es sich hierbei um einen systematischen Fehler handelt. Es sei an dieser Stelle auf GELMAN ET AL. (2000) verwiesen: Sie untersuchen die Güte des Modells 4 im Hinblick auf die hinreichende Berücksichtigung der hundespezifischen Lerneigenschaften und kommen zu dem Schluss, dass die beobachteten Unterschiede nicht größer sind, als es unter dem Modell zu erwarten wäre.

Wesentlich auffälliger als die leicht erhöhte Streuung der NPR-Intervalle sind

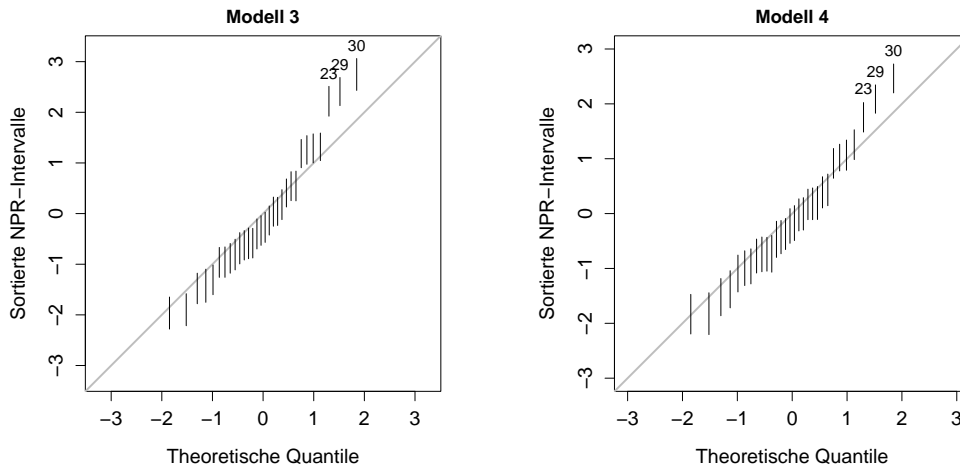


Abbildung 3.32: QQ-Plot der NPR-Intervalle für die Summe der Stromstöße der einzelnen Hunde unter Modell 3 bzw. Modell 4.

die jeweils drei extremen NPR-Intervalle auf der rechten Seite der QQ-Plots. Sie liegen in beiden Graphiken über der Winkelhalbierenden, im QQ-Plot für das Modell 3 deutlich darüber. Die entsprechenden NPR-Intervalle sind demnach größer als unter den beiden Modellen (und bei dem gegebenen Stichprobenumfang) erwartet. Da es sich hier um NPR-Intervalle für die Summe der Stromstöße eines Hundes handelt, markieren die extremen NPR-Intervalle solche Hunde, die schlechter lernen als durch die Modelle prognostiziert. Die Indizes dieser Hunde ($i = 23$, $i = 29$ und $i = 30$; eingefügt in die QQ-Plots) sind in beiden Fällen identisch. Abbildung 3.33 zeigt noch einmal die individuell geschätzten Lernkurven des Modells 2, wobei die Lernkurven der im QQ-Plot auffälligen Hunde schwarz gezeichnet sind. Es ist zu erkennen, dass diese im Vergleich zu den übrigen Lernkurven deutlich nach rechts verschoben sind. Diese Hunde, deren Wahrscheinlichkeit für einen Stromstoß bei individueller Modellierung in nahezu allen Versuchen höher ist als für die übrigen Hunde, können daher als extrem betrachtet werden: Sie lernen deutlich schlechter als die übrigen Hunde. Die Auffälligkeit der NPR-Intervalle in den QQ-Plots zeigt somit, dass diese extremen Beobachtungen nicht adäquat durch die Modelle beschrieben werden.

Ein weiterer, bis zu diesem Zeitpunkt nicht betrachteter Aspekt der Anpassungsgüte ist die Beschreibung des Lernprozesses im Zeitverlauf. Bei der Überprüfung dieses Kriteriums rücken die Spalten der Tabelle 3.11 in den Vordergrund, d.h. es werden nun die einzelnen Versuche betrachtet (und nicht mehr die einzelnen Hunde). Eine Möglichkeit, die Erfassung der zeitlichen Verände-

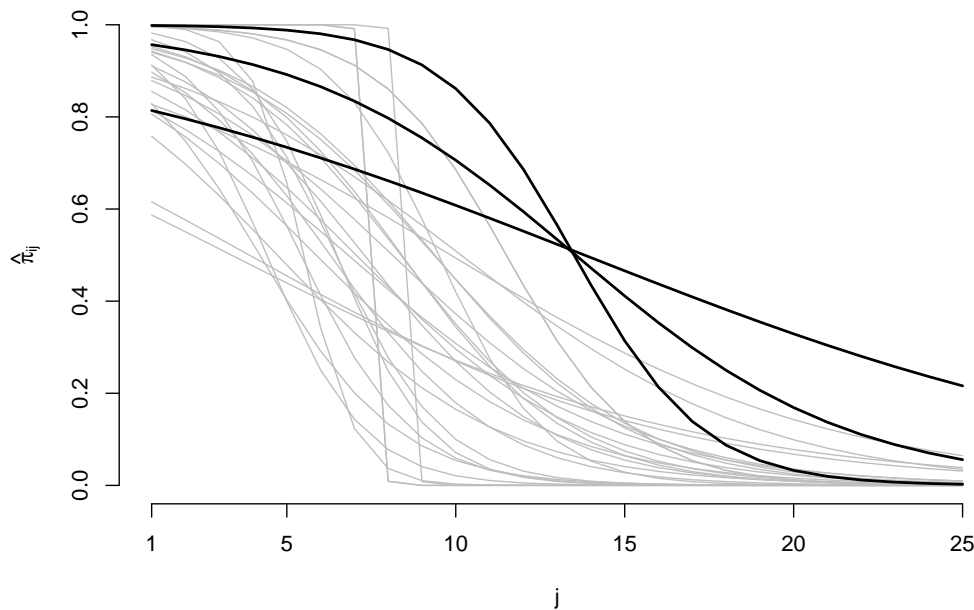


Abbildung 3.33: Lernkurven des Modells 2 mit Markierung der Ausreißer.

rungen durch die Modelle zu überprüfen, ist bei GELMAN ET AL. (2000) erläutert. Sie untersuchen die durchschnittlichen Residuen in den einzelnen Versuchen

$$\bar{e}_j = \frac{1}{30} \sum_{i=1}^{30} y_{ij} - \hat{\pi}_{ij}.$$

Der Indexplot der durchschnittlichen Residuen (Abbildung 3.34) zeigt, dass bei Verwendung des Modells 3 die Erfolgswahrscheinlichkeiten in den ersten vier Versuchen des Experiments unterschätzt (positive Residuen) und in den darauffolgenden sechs Versuchen tendenziell überschätzt (negative Residuen) werden. Diese systematische Struktur der Residuen im Zeitverlauf ist für das Modell 4 nicht zu erkennen. Bei der graphischen Analyse ist allerdings zu beachten, dass die Varianz der Residuen

$$\begin{aligned} \text{Var}(\bar{e}_j) &= \frac{1}{j^2} \sum_{i=1}^{30} \text{Var}(Y_{ij}) \\ &= \frac{1}{j^2} \sum_{i=1}^{30} \pi_{ij}(1 - \pi_{ij}) \end{aligned}$$

zu den verschiedenen Zeitpunkten unterschiedlich ist. Die geschätzten Standardabweichungen der Residuen sind mit positivem und negativem Vorzei-

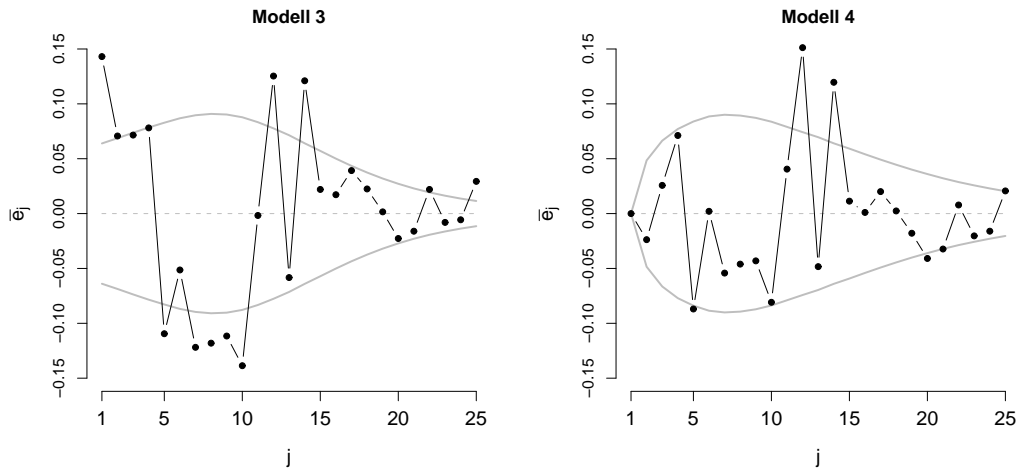


Abbildung 3.34: Durchschnittliche Residuen für die einzelnen Versuche und ihre Standardabweichung unter Modell 3 bzw. Modell 4.

chen (die Schätzung erfolgte durch Einsetzen der $\hat{\pi}_{ij}$ in die obige Gleichung für die Varianz) in Abbildung 3.34 grau dargestellt. Es wird an dieser Stelle klar, dass die unterschiedliche Streuung der Residuen die Interpretation der Indexplore-Plots verzerrt. Dieses Problem kann vermieden werden, wenn man an Stelle der durchschnittlichen Residuen die NPR-Intervalle der beobachteten Spaltensummen analysiert. Analog zum obigen Vorgehen, bei dem die NPR-Intervalle für die beobachteten Zeilensummen (also für die Hunde) bestimmt wurden, sind hier die Verteilungen für die Spaltensummen unter den Modellen zu simulieren und anschließend die entsprechenden NPR-Intervalle zu berechnen. Letztere sind in den Indexplore-Plots der Abbildung 3.35 zusammen mit einer Glättung der Intervallmittelpunkte (verwendet wurde der **R**-Befehl `lowess`; siehe CLEVELAND, 1979 und 1981) dargestellt. Das systematische Verhalten der geglätteten Intervallmittelpunkte bei Modell 3 impliziert eine sehr ähnliche Interpretation wie bei der Analyse der durchschnittlichen Residuen, d.h. die Erfolgswahrscheinlichkeiten werden zunächst unter- und später überschätzt. Die Untersuchung der NPR-Intervalle besitzt allerdings den Vorteil, dass die Amplitude (also das Ausmaß der Schwankungen) nicht durch eine unterschiedliche Varianz der Residuen beeinflusst wird. Abschließend sollte hier bemerkt werden, dass bei der Interpretation der Strukturen natürlich eine Berücksichtigung des (im hier betrachteten Beispiel relativ geringen) Stichprobenumfangs erforderlich ist.

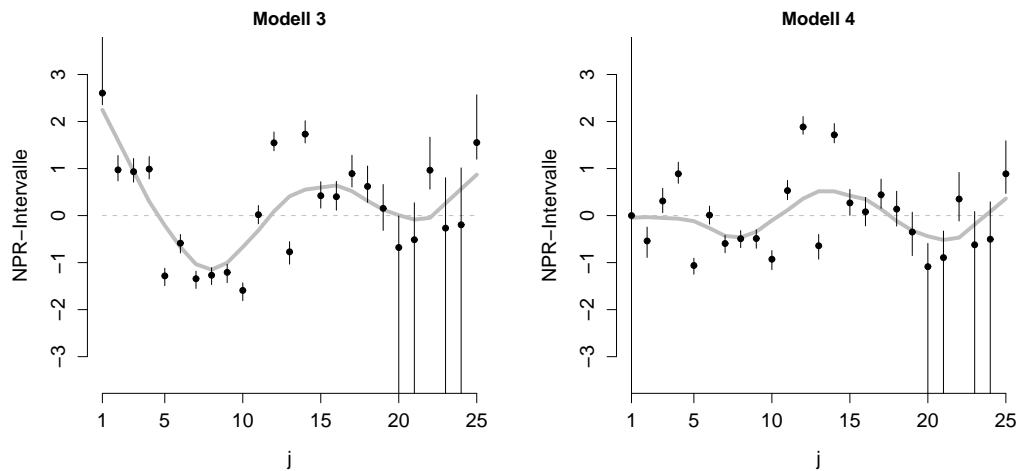


Abbildung 3.35: NPR-Intervalle der einzelnen Versuche unter Modell 3 bzw. Modell 4.

3.5.2 Matrizen für Markenwechsel

Die im folgenden Abschnitt betrachteten Daten sind Zähldaten, die in einer Kontingenztafel (bzw. Matrix) mit gleicher Zeilen- und Spaltenzahl dargestellt werden können, wobei die Zeilen und Spalten dieselbe (kategoriale) Variable zu einem unterschiedlichen Zeitpunkt repräsentieren. Die Einträge y_{ij} (für $i = 1, \dots, I, j = 1, \dots, J$ und $I = J$) in den Zellen (i, j) entsprechen der Anzahl von Beobachtungseinheiten, die in der Periode t zur Kategorie j gehören und zuvor zur Kategorie i der Variablen gehörten.

Tabelle 3.12 enthält die Daten des in diesem Abschnitt betrachteten Beispiels. Es handelt sich um die Darstellung von Beobachtungen, die auf dem französischen Automarkt im Jahr 1989 gemacht wurden. Die Daten stammen von COLOMBO, EHRENBERG AND SABVALA (1994) und sind hier zitiert nach IACOBUCCI AND HENDERSON (1997). Die Spalten repräsentieren die Automarken, die während der Erhebung von den befragten Personen gefahren wurden, die Zeilen geben an, welche Automarke sie zuvor besaßen. Die Beobachtungen stehen jeweils in der mittleren Zeile einer Kategorie. (Darüber und darunter sind mit verschiedenen Modellen geschätzte Häufigkeiten angegeben, die später betrachtet werden.) Beispielsweise liest man aus der Zelle $(1, 3)$ ab, dass 19 Personen, die zuvor einen Alpha gefahren haben, im Jahr 1989 einen Citroen besitzen, d.h. sie haben einen Markenwechsel (daher die Überschrift dieses Abschnitts) von Alpha zu Citroen vollzogen.

Nachfolgend wird zunächst dargestellt, wie die hier betrachteten Daten durch

	AI	BMW	Ci	Fi	Fo	GM	La	Me	Pe	Re	Ro	Se	VW	Vo	Σ
Alpha	Unabhängigkeit	2	3.6	32	16.2	18.3	12.8	2.3	3	58.2	79.6	4.2	3.9	19.8	2
	Beobachtungen	97	5	19	19	14	7	0	3	27	35	5	6	17	4
	Quasi-Unabhgk.	—	2.1	15.4	13.1	13.2	10.8	1.9	1.8	37.8	41.4	3.6	4.3	14.1	1.3
BMW	Unabhängigkeit	2.8	5	44.3	22.4	25.3	17.7	3.2	4.2	80.5	110.1	5.9	5.4	27.4	2.8
	Beobachtungen	4	163	20	9	14	6	0	19	39	40	6	4	29	4
	Quasi-Unabhgk.	1.3	—	18.7	15.9	16	13.1	2.3	2.2	45.8	50.3	4.4	5.3	17.1	1.6
Citroen	Unabhängigkeit	25.7	46	405.5	204.6	231.3	161.5	29.6	38.5	736.1	1006.9	53.7	49.4	250.9	25.4
	Beobachtungen	6	13	1811	136	98	67	21	15	464	477	29	28	90	10
	Quasi-Unabhgk.	11	21.3	—	129.7	130.7	107.3	18.9	17.7	374.7	410.8	36.1	43	139.5	13.4
Fiat	Unabhängigkeit	8.9	16	141.3	71.3	80.6	56.3	10.3	13.4	256.6	351	18.7	17.2	87.4	8.8
	Beobachtungen	11	5	69	526	50	49	12	3	134	148	27	17	82	5
	Quasi-Unabhgk.	4.6	8.8	63.3	—	54.1	44.4	7.8	7.3	155.3	170.2	14.9	17.8	57.8	5.5
Ford	Unabhängigkeit	10	18	158.6	80	90.5	63.2	11.6	15.1	287.9	393.8	21	19.3	98.1	9.9
	Beobachtungen	4	10	45	53	696	76	8	8	121	146	20	23	60	7
	Quasi-Unabhgk.	4.3	8.4	60.2	51.1	—	42.2	7.5	7	147.5	161.7	14.2	16.9	54.9	5.3
GM	Unabhängigkeit	5.4	9.6	85.1	42.9	48.5	33.9	6.2	8.1	154.4	211.3	11.3	10.4	52.6	5.3
	Beobachtungen	2	7	20	23	50	362	7	4	68	80	7	12	39	4
	Quasi-Unabhgk.	2.4	4.6	32.9	27.9	28.1	—	4.1	3.8	80.7	88.5	7.8	9.3	30	2.9
Lada	Unabhängigkeit	1.3	2.4	21.2	10.7	12.1	8.5	1.5	2	38.6	52.7	2.8	2.6	13.1	1.3
	Beobachtungen	0	0	12	13	11	12	68	0	25	16	4	3	7	0
	Quasi-Unabhgk.	0.7	1.4	9.9	8.4	8.5	7	—	1.1	24.3	26.6	2.3	2.8	9	0.9
Mercedes	Unabhängigkeit	1.5	2.7	24.1	12.2	13.7	9.6	1.8	2.3	43.7	59.8	3.2	2.9	14.9	1.5
	Beobachtungen	2	17	3	3	5	2	0	136	4	10	3	1	5	3
	Quasi-Unabhgk.	0.4	0.8	5.6	4.7	4.8	3.9	0.7	—	13.7	15	1.3	1.6	5.1	0.5
Peugeot	Unabhängigkeit	38.3	68.6	605.4	305.5	345.4	241.1	44.1	57.5	1099	1503.5	80.2	73.8	374.6	37.9
	Beobachtungen	13	23	273	164	195	168	34	23	2928	728	47	60	199	20
	Quasi-Unabhgk.	17.4	33.7	241.5	205	206.5	169.5	29.9	28	—	649.1	57	67.9	220.4	21.1
Renault	Unabhängigkeit	60.2	107.8	950.8	479.8	542.4	378.7	69.3	90.3	1726	2361.2	126	115.8	588.3	59.5
	Beobachtungen	22	37	353	334	308	254	40	33	896	4861	78	106	310	24
	Quasi-Unabhgk.	25.7	49.8	357.1	303.1	305.4	250.6	44.3	41.4	875.7	—	84.3	100.4	325.9	31.3
Rover	Unabhängigkeit	2.6	4.6	40.6	20.5	23.2	16.2	3	3.9	73.7	100.8	5.4	4.9	25.1	2.5
	Beobachtungen	2	4	18	25	30	3	6	1	30	45	11.5	6	40	2
	Quasi-Unabhgk.	1.5	2.9	20.6	17.5	17.6	14.5	2.6	2.4	50.6	55.4	—	5.8	18.8	1.8
Seat	Unabhängigkeit	1.1	1.9	17	8.6	9.7	6.8	1.2	1.6	30.9	42.3	2.3	2.1	10.5	1.1
	Beobachtungen	2	0	13	10	10	11	0	0	10	29	3	36	13	0
	Quasi-Unabhgk.	0.7	1.4	9.9	8.4	8.4	6.9	1.2	1.1	24.2	26.5	2.3	—	9	0.9
VW Audi	Unabhängigkeit	10.8	19.4	170.8	86.2	97.4	68	12.4	16.2	310	424.1	22.6	20.8	105.7	10.7
	Beobachtungen	6	21	50	51	65	59	2	10	177	115	13	25	772	9
	Quasi-Unabhgk.	4.5	8.8	62.8	53.3	53.7	44.1	7.8	7.3	154	168.8	14.8	17.7	—	5.5
Volvo	Unabhängigkeit	1.3	2.3	20.1	10.2	11.5	8	1.5	1.9	36.5	50	2.7	2.5	12.4	1.3
	Beobachtungen	1	3	11	5	4	6	0	3	9	17	3	4	18	78
	Quasi-Unabhgk.	0.6	1.1	8.1	6.8	6.9	5.7	1	0.9	19.8	21.7	1.9	2.3	7.4	—
Σ	Unabhängigkeit	172	308	2717	1371	1550	1082	198	258	4932	6747	360	331	1681	170
	Beobachtungen	172	308	2717	1371	1550	1082	198	258	4932	6747	360	331	1681	170
	Quasi-Unabhgk.	75	145	906	845	854	720	130	122	2004	1886	245	295	909	92

Tabelle 3.12: Beobachtete Markenwechsel auf dem Automarkt und geschätzte Zellhäufigkeiten unter Unabhängigkeit und Quasi-Unabhängigkeit.

verallgemeinerte lineare Modelle beschrieben werden können, kurz auf die Parameterschätzung dieser Modelle eingegangen und erläutert, wie geeignete Pseudo-Residuen zu bestimmen sind. Anschließend werden vier Modelle betrachtet, die für das Beispiel dieses Abschnitts besonders geeignet erscheinen. Es wird anhand dieser Modelle gezeigt, wie mit einer Analyse von Tableplots die in der Kontingenztafel beobachtete Marktsituation detailliert untersucht werden kann.

Im Kontext der verallgemeinerten linearen Modelle erfolgt die Modellierung von Kontingenztafeln durch sogenannte loglineare Modelle. Eine vollständige Beschreibung der Daten liefert die Verwendung des gesättigten loglinearen Modells, das in Analogie zur Varianzanalyse durch

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad \text{für alle } i, j$$

mit

$$\begin{aligned} \lambda_{ij} &= E(Y_{ij}), \\ \mu &= \frac{1}{IJ} \sum_{i,j} \log \lambda_{ij}, \\ \alpha_i &= \frac{1}{J} \sum_j \log \lambda_{ij} - \mu, \\ \beta_j &= \frac{1}{I} \sum_i \log \lambda_{ij} - \mu \quad \text{und} \\ (\alpha\beta)_{ij} &= \log \lambda_{ij} - \left(\frac{1}{J} \sum_j \log \lambda_{ij} + \frac{1}{I} \sum_i \log \lambda_{ij} \right) + \mu \end{aligned}$$

dargestellt werden kann (vgl. BISHOP FIENBERG AND HOLLAND, 1975, S. 24). Die Nebenbedingungen

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

des Modells reduzieren die Anzahl der unabhängigen Parameter auf die Anzahl der Zellen.

Zur Schätzung der Modellparameter wird bei den verallgemeinerten linearen Modellen eine Verteilung der Zielvariable aus der Exponentialfamilie vorausgesetzt. Im hier betrachteten Beispiel ist die gemeinsame Verteilung der Y_{ij} aufgrund des Versuchsaufbaus durch eine Multinomialverteilung mit der ge-

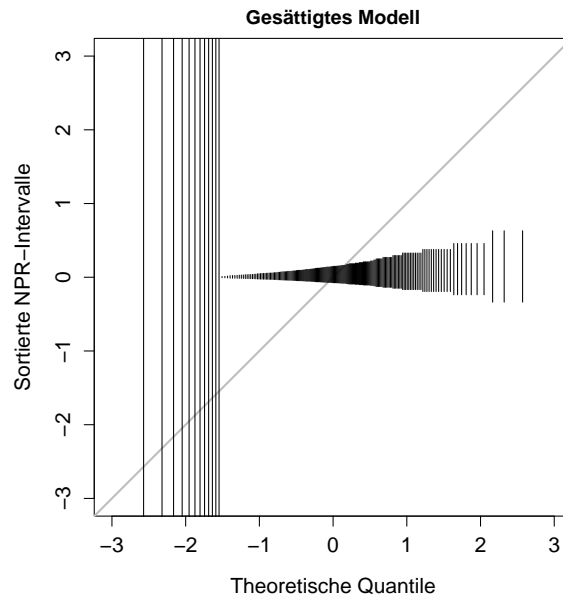


Abbildung 3.36: QQ-Plot der NPR-Intervalle für die Zelhäufigkeiten aus dem gesättigten Modell.

meinsamen Wahrscheinlichkeitsfunktion

$$f(\mathbf{y}) = \frac{n!}{\prod_{ij} y_{ij}!} \prod_{i=1}^I \prod_{j=1}^J \pi_{ij}^{y_{ij}}$$

gegeben, wobei \mathbf{y} den Vektor aller Beobachtungen y_{ij} , π_{ij} die entsprechenden Erfolgswahrscheinlichkeiten und n die Anzahl der Beobachtungseinheiten darstellt. DOBSON (2002, S. 135–136) zeigt, dass die gemeinsame Verteilung von unabhängig poissonverteilten Zufallsvariablen Y_{ij} , gegeben die Summe der Beobachtungen, durch eine Multinomialverteilung mit den Parametern $\pi_{ij} = \lambda_{ij} / \sum_{ij} \lambda_{ij}$ dargestellt werden kann. Zum Zwecke der Parameterschätzung kann daher von einer Poissonverteilung ausgegangen werden, wenn die Gesamtzahl der Beobachtungen nach dem Modell erhalten bleibt (vgl. DOBSON, 2002, S. 164). Dieses Resultat ist ein Spezialfall der allgemeineren Darstellungen von BIRCH (1963), nach denen für die Parameterschätzung auch für andere Formen des Versuchsaufbaus von einer Poissonverteilung ausgegangen werden kann, wenn bestimmte Bedingungen beachtet werden. Andere Formen des Versuchsaufbaus und die implizierten gemeinsamen Verteilungen der Zelhäufigkeiten sind anschaulich bei TUTZ (2000, S. 265–268) beschrieben. Um die Pseudo-Residuen für die einzelnen Zellen zu bestimmen, kann nicht von einer Poissonverteilung der Zelhäufigkeiten ausgegangen werden. Diese

sind binomialverteilt mit den Parametern n und $\pi_{ij} = \lambda_{ij} / \sum_{ij} \lambda_{ij}$ und besitzen die Kovarianz $Cov(Y_{ij}, Y_{kl}) = -n\pi_{ij}\pi_{kl}$ (vgl. AGRESTI, 1990, S. 44). Verwendet man die Schätzer aus dem gesättigten Modell (also die Beobachtungen selbst) zur Spezifikation dieser Verteilungen und berechnet die entsprechenden Pseudo-Residuen, so erhält man den in Abbildung 3.36 dargestellten QQ-Plot der NPR-Intervalle. Der QQ-Plot verläuft horizontal, was die exakte Wiedergabe der Beobachtungen durch das gesättigte Modell widerspiegelt (siehe auch ZUCCHINI AND MACDONALD, 1999).

Natürlich ist auch bei der Beschreibung von Kontingenztafeln die exakte Wiedergabe der Daten nicht das Ziel der Modellierung. Vielmehr muss nach allgemeinen, grundlegenden Strukturen gesucht werden, die in den Daten enthalten sind und in Form eines vereinfachten Modells formuliert werden können. Für die hier betrachteten Matrizen für Markenwechsel sind Modelle für Unabhängigkeit, Quasi-Unabhängigkeit, Symmetrie und Quasi-Symmetrie besonders geeignet, weil die Überprüfung dieser Modelle konkrete Hinweise zur Beurteilung des betrachteten Marktes liefert: Jedes dieser Modelle basiert auf bestimmten Annahmen, deren Verletzung mit Hilfe einer Analyse der Pseudo-Residuen aufgedeckt werden kann und gezielte Interpretationen bezüglich des Verhaltens der Käufer ermöglicht. Im Folgenden werden die Annahmen der genannten Modelle in Hinsicht auf ihre Bedeutung für den Automarkt erläutert. Außerdem wird die Anpassung der Modelle überprüft und interpretiert.

Das Modell für Unabhängigkeit

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j \quad \text{für alle } i, j$$

liefert eine adäquate Beschreibung der Daten, wenn die Zeilenvariable und die Spaltenvariable unabhängig sind. Mit Blick auf den Automarkt impliziert das Modell die Unabhängigkeit der Markenwahl zum Zeitpunkt t von der zuvor besessenen Automarke. Bereits ein Blick auf die Daten verrät, dass diese Art der Unabhängigkeit hier nicht erfüllt ist: Die sehr großen Werte auf der Diagonalen der Kontingenztafel zeigen eine ausgeprägte Loyalität der Käufer an, d.h. bei einer erneuten Kaufentscheidung wird sehr häufig die zuvor besessene Marke gewählt.

Die angepassten Zellhäufigkeiten des Modells für Unabhängigkeit sind in Tabelle 3.12 jeweils über der Beobachtung angegeben. Stellt man die NPR-Intervalle der Zellhäufigkeiten in einem Indexplot (Abbildung 3.37) dar, wird das Ausmaß der Fehlspezifikation deutlich. Die grau dargestellten Punkte zeigen die Mittelpunkte der NPR-Intervalle, die spaltenweise angeordnet sind, d.h. der erste Punkt gehört zur Zelle (1, 1), der zweite zur Zelle (1, 2) usw. Al-

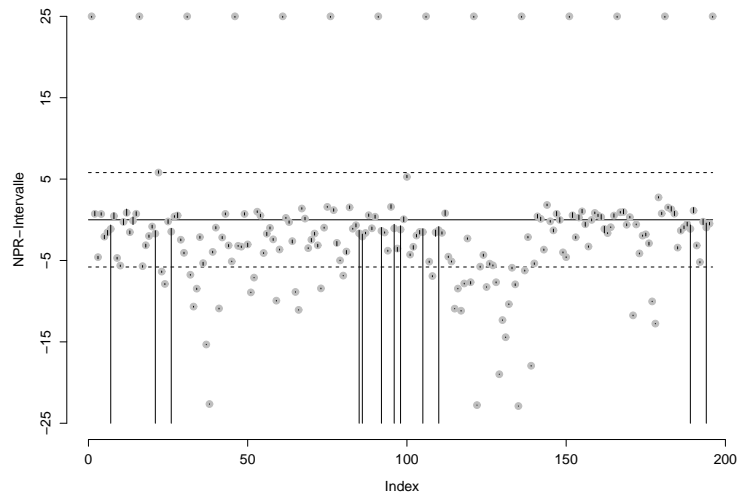


Abbildung 3.37: Indexplot der NPR-Intervalle aus dem Modell für Unabhängigkeit.

len Mittelpunkten der NPR-Intervalle, die einem Diagonalelement der Kontingenztafel entsprechen, ist in der Graphik der Wert 25 zugeordnet. Tatsächlich sind diese Residuen so groß, dass \mathbf{R} den Wert ∞ berechnet. Sie zeigen an, dass die Anzahl der Personen, die bei einer erneuten Kaufentscheidung die zuvor besessene Marke erneut wählen, drastisch unterschätzt wird. Alle übrigen Residuen sind tendenziell zu klein, d.h. die Häufigkeiten für Markenwechsler werden im Durchschnitt überschätzt. Neben den Mittelpunkten der NPR-Intervalle sind in Abbildung 3.37 auch die Intervalle selbst als schwarze Linien eingezeichnet. Diese sind allerdings aufgrund der Skalierung der Ordinate nur für wenige Residuen zu erkennen. Eine Ausnahme bilden die NPR-Intervalle für Beobachtungen, die in der Kontingenztafel den Wert null annehmen und deren linke Intervallgrenze daher $-\infty$ beträgt. Beispielsweise gilt dies für das siebte Residual, das für den Übergang von der Marke Lada zur Marke Alpha berechnet wird.

Weitere Möglichkeiten zur Überprüfung und Interpretation der Residuen werden im Folgenden zusammen mit dem Modell für Quasi-Unabhängigkeit

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j \quad \text{für alle } i \neq j$$

beschrieben. Das Modell für Quasi-Unabhängigkeit basiert auf der Annahme, dass unter der Bedingung, dass ein Markenwechsel vollzogen wird, die gewählte Marke zum Zeitpunkt t unabhängig ist von der zuvor besessenen Marke. Die Schätzer für die Zellhäufigkeiten erhält man, indem man das Modell für

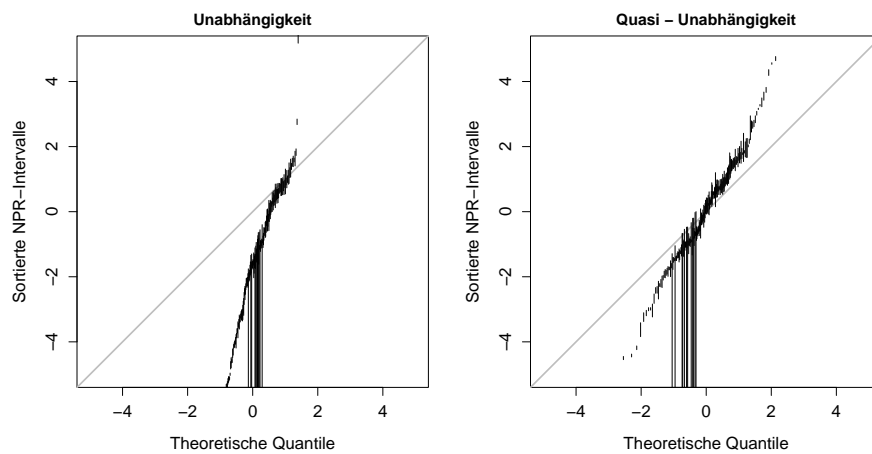
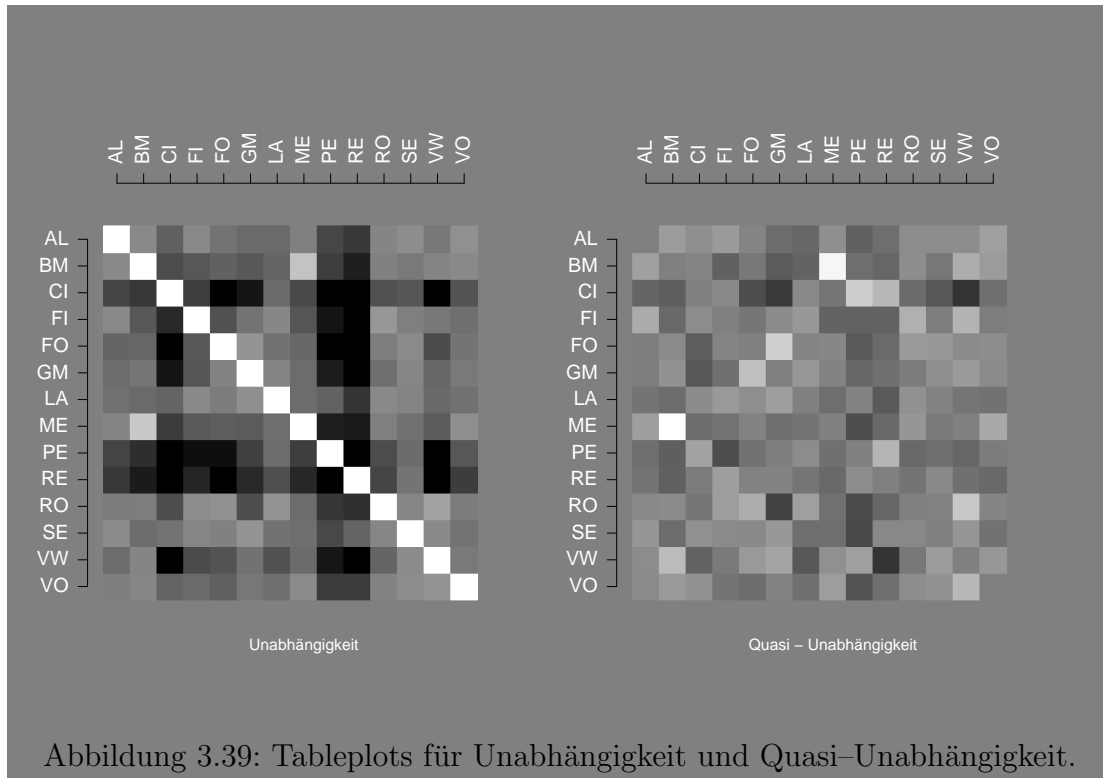


Abbildung 3.38: QQ-Plots für Unabhängigkeit und Quasi-Unabhängigkeit.

Unabhängigkeit an alle Beobachtungen, die nicht auf der Diagonalen liegen, anpasst (vgl. LINDSEY, 1997, S. 32). Dies ist im hier betrachteten Beispiel für 9 228 Personen der Fall. Die nach dem Modell für Quasi-Unabhängigkeit geschätzten Häufigkeiten sind in Tabelle 3.12 jeweils unterhalb der entsprechenden Beobachtung angegeben.

Abbildung 3.38 zeigt die QQ-Plots der NPR-Intervalle die sich nach dem Modell für Unabhängigkeit bzw. Quasi-Unabhängigkeit ergeben. Die linke Graphik (Unabhängigkeit) zeigt (wie bereits der Indexplot in Abbildung 3.37), dass die Markenwahl zum Zeitpunkt t keinesfalls unabhängig von der zuvor gewählten Marke ist. Aus einem Vergleich mit dem Indexplot wird deutlich, dass viele extreme NPR-Intervalle nicht in dem QQ-Plot dargestellt sind. Dies gilt für alle Intervalle, die im Indexplot außerhalb der gestrichelten Linien liegen. Der rechte QQ-Plot zeigt, dass die Anpassungsgüte des Modells für Quasi-Unabhängigkeit besser ist. Allerdings variieren auch hier die Residuen stärker, als es bei Gültigkeit des Modells zu erwarten wäre, d.h. ein Teil der Variation in den Daten wird nicht durch das Modell erklärt.

Während die QQ-Plots geeignet sind, um die Anpassung des Modells insgesamt zu beurteilen, bieten die Tableplots (vgl. Abbildung 3.39) eine differenziertere Betrachtungsweise. Die Tableplots besitzen die gleiche Struktur wie die betrachtete Kontingenztafel, d.h. sie zeigen die gleichen Zeilen und Spalten wie die Kontingenztafel. Anstelle der Zelhäufigkeiten sind allerdings die Mittelpunkte der NPR-Intervalle für die beobachteten Zelhäufigkeiten unter dem Modell in Form von Graustufen dargestellt. Eine relativ hell (dunkel) ausgefüll-



te Zelle repräsentiert einen relativ großen (kleinen) Wert des entsprechenden Residuals. Für helle bzw. dunkle Zellen der Tabelle werden die Häufigkeiten demnach unterschätzt bzw. überschätzt. Beträgt der Wert des Mittelpunktes eines NPR-Intervalls in etwa null, so ist die entsprechende Zelle in dem Grau gezeichnet, das dem Hintergrund entspricht. Ein ähnliches Vorgehen zur Analyse der Modelle bzw. des Marktes findet sich bei IACOBUCCI AND HENDERSON (1997). Sie definieren bestimmte Grenzen für standardisierte Residuen und teilen diese anhand der Grenzen in Klassen ein. In einer Tabelle verwenden sie dann verschiedene Symbole für die Klassen, um die Beschreibung der Zelhäufigkeiten durch das Modell zu beurteilen.

Die Tableplots der Modelle für Unabhängigkeit und Quasi-Unabhängigkeit sind in Abbildung 3.39 dargestellt. Die deutlich helle Diagonale der linken Graphik (Modell für Unabhängigkeit) spiegelt die bereits erwähnte Loyalität der Kunden wider, d.h. deutlich mehr Käufer bleiben ihrer Automarke treu und entscheiden sich in diesem Sinn nicht unabhängig für eine der zur Auswahl stehenden Marken. Die relativ dunklen Zeilen und Spalten der französischen Automarken Peugeot, Renault und Citroen bilden eine weitere systematische Struktur des Tableplots. Entlang der Zeilen gelesen, bedeutet dies, dass die Besitzer von französischen Automarken tendenziell weniger häufig zu anderen Marken wechseln, als dies bei Unabhängigkeit der Fall wäre. Sie sind nach dieser Interpretation also besonders treue Kunden. Die dunklen Spalten zeigen an, dass auch ein Wechsel von den nicht französischen zu den französischen Au-

tomarken tendenziell weniger häufig stattfindet, als es unter Unabhängigkeit der Kaufentscheidungen zu erwarten wäre. Die gemeinsame Interpretation der dunklen Zeilen und Spalten führt zu der Vermutung, dass die französischen Automarken und die übrigen Automarken zwei separate Märkte bilden, zwischen denen weniger häufig gewechselt wird als innerhalb der beiden Märkte.

Der Tableplot des Modells für Quasi-Unabhängigkeit (rechte Graphik) ist ähnlich zu interpretieren. Allerdings sind die Aussagen zum Käuferverhalten in diesem Fall auf die Käufer zu begrenzen, die einen Wechsel der Marke vollzogen haben (die Diagonale ist in der Hintergrundfarbe gezeichnet, da das Modell keine sinnvoll zu interpretierenden Schätzer für die entsprechenden Häufigkeiten liefert). Beispielsweise ist der Wechsel von Mercedes zu BMW sowie der Wechsel von BMW zu Mercedes in der Graphik durch relativ helle Zellen markiert. Markenwechsler, die zuvor einen Mercedes besessen haben, wechseln demnach häufiger zu BMW (und umgekehrt), als es unter Unabhängigkeit zu erwarten wäre. Zwischen den beiden Marken existiert demnach eine ausgeprägte Konkurrenz. Einseitig ist der Verlust von Kunden der Marke Rover an die Marke VW, d.h. Rover verliert mehr Kunden an VW als unter dem Modell für Quasi-Unabhängigkeit geschätzt und kann den Verlust im Gegenzug nicht entsprechend kompensieren. Eine dunkle Zelle repräsentiert weniger Kundenbewegungen als bei Quasi-Unabhängigkeit erwartet: So wäre es beispielsweise für VW interessant, die Ursache dafür zu erforschen, dass es nur selten gelingt, Markenwechsler von Citroen zu akquirieren. Allgemein helfen die aus den Tableplots abgelesenen Informationen den Unternehmen, ihre Marktposition zu beurteilen. Sie erkennen ihre Stärken und Schwächen gegenüber der Konkurrenz und können diese durch entsprechende Maßnahmen ausbauen bzw. abbauen.

Bei der Untersuchung der Tableplots ist unter anderem aufgefallen, dass die Beziehung zwischen den Marken Mercedes und BMW in etwa symmetrisch ist. Es wurden 17 Käufer beobachtet, die Mercedes an BMW abgeben musste und 19 Käufer, die BMW an Mercedes verloren hat (in beiden Fällen deutlich mehr Wechsler als unter den Modellen für Unabhängigkeit und Quasi-Unabhängigkeit erwartet). Die beiden nachfolgend dargestellten Modelle basieren auf der Annahme von Symmetrien dieser Art. Das Modell für Symmetrie

$$\log(\lambda_{ijk}) = \mu + \gamma_k \quad \text{für alle } i \neq j \text{ und } k = 1, \dots, (I^2 - I)/2$$

enthält den Faktor γ , der für die Beobachtungen eines Zellenpaars — gemeint sind hiermit zwei gegenüberliegende Zellen (i, j) und (j, i) — die gleiche Stufe besitzt (vgl. LINDSEY, 1997, S. 34). Das Modell basiert auf der Annahme,

	Al	BMW	Ci	Fi	Fo	GM	La	Me	Pe	Re	Ro	Se	VW	Vo	Σ
Alpha	Symmetrie	4.5	12.5	15	9	4.5	0	2.5	20	28.5	3.5	4	11.5	2.5	118
	Beobachtungen	97	5	19	14	7	0	3	27	35	5	6	17	4	258
	Quasi-Sym.	5.6	14.1	22	13.6	7.4	0	4	26.8	34.5	5	6.9	17.5	3.5	161
BMW	Symmetrie	4.5	16.5	7	12	6.5	0	18	31	38.5	5	2	25	3.5	169.5
	Beobachtungen	4	163	20	9	14	6	0	19	39	40	6	4	29	357
	Quasi-Sym.	3.4	14.4	8.7	15.6	9.6	0	25.6	34	36.8	6	3.1	32.7	4.1	194
Citroen	Symmetrie	12.5	16.5	102.5	71.5	43.5	16.5	9	368.5	415	23.5	20.5	70	10.5	1180
	Beobachtungen	6	13	1811	136	98	21	15	464	477	29	28	90	10	3265
	Quasi-Sym.	10.9	18.6	139.7	101.2	68.3	22.3	13.7	450.5	450.8	30.9	33.9	99.6	13.6	1454
Fiat	Symmetrie	15	7	102.5	51.5	36	12.5	3	149	241	26	13.5	66.5	5	728.5
	Beobachtungen	11	5	69	526	50	49	3	134	148	27	17	82	5	1138
	Quasi-Sym.	8	5.3	65.3	54.6	45.4	12.4	3.6	126.2	172.2	24.6	18.6	71.2	4.6	612
Ford	Symmetrie	9	12	71.5	51.5	63	9.5	6.5	158	227	25	16.5	62.5	5.5	717.5
	Beobachtungen	4	10	45	53	696	8	8	121	146	20	23	60	7	1277
	Quasi-Sym.	4.4	8.4	41.8	48.4	75.7	8.8	7.4	149.7	149.7	22.1	21.9	63.1	4.8	581
GM	Symmetrie	4.5	6.5	43.5	36	63	9.5	3	118	167	5	11.5	49	5	521.5
	Beobachtungen	2	7	20	23	50	362	7	68	80	7	12	39	4	685
	Quasi-Sym.	1.6	3.4	18.7	26.6	50.3	6.9	2.8	71.1	82.2	3.5	1.5	39.5	3.4	323
Lada	Symmetrie	0	0	16.5	12.5	9.5	9.5	0	29.5	28	5	1.5	4.5	0	116.5
	Beobachtungen	0	0	12	13	11	12	68	0	25	16	4	3	7	171
	Quasi-Sym.	0	0	10.7	12.6	10.2	12.1	0	25.3	20.3	4.8	2.1	4.9	0	103
Mercedes	Symmetrie	2.5	18	9	3	6.5	3	0	13.5	21.5	2	0.5	7.5	3	90
	Beobachtungen	2	17	3	3	5	2	0	136	4	10	3	1	5	194
	Quasi-Sym.	1	10.4	4.3	2.4	5.6	3.2	0	8.9	11.6	1.5	0.6	6.5	2.2	58
Peugeot	Symmetrie	20	31	368.5	149	158	118	29.5	13.5	812	38.5	35	188	14.5	1975.5
	Beobachtungen	13	23	273	164	195	168	34	23	2928	47	60	199	20	4875
	Quasi-Sym.	13.2	28	286.5	171.8	191.5	164.9	33.7	18.1	699.3	42.3	52.6	229.4	15.7	1947
Renault	Symmetrie	28.5	38.5	415	241	227	167	28	21.5	812	61.5	67.5	212.5	20.5	2340.5
	Beobachtungen	22	37	353	334	308	254	40	33	896	78	106	310	24	7656
	Quasi-Sym.	22.5	40.2	379.2	309.8	304.3	251.8	35.7	31.4	924.7	76	108	286.6	25	2795
Rover	Symmetrie	3.5	5	23.5	26	25	5	2	38.5	61.5	—	4.5	26.5	2.5	228.5
	Beobachtungen	2	4	18	25	30	3	6	1	30	45	11.5	6	40	327
	Quasi-Sym.	2	4	16.1	27.4	27.9	6.5	5.2	34.7	47	6.4	29.8	2.5	212	
Seat	Symmetrie	4	2	20.5	13.5	16.5	11.5	1.5	35	67.5	4.5	—	19	2	198
	Beobachtungen	2	0	13	10	10	11	0	10	29	3	36	13	0	137
	Quasi-Sym.	1.1	0.9	7.1	8.4	11.1	10	0.9	17.4	27	2.6	—	13	1.1	101
VW Audi	Symmetrie	11.5	25	70	66.5	62.5	49	4.5	7.5	188	212.5	26.5	19	—	13.5
	Beobachtungen	6	21	50	51	65	59	2	10	177	115	13	25	772	9
	Quasi-Sym.	5.5	17.3	40.4	61.8	61.9	58.5	4.1	8.5	146.6	138.4	23.2	25	—	11.6
Volvo	Symmetrie	2.5	3.5	10.5	5	5.5	5	0	3	14.5	20.5	2.5	2	13.5	88
	Beobachtungen	1	3	11	5	4	6	0	3	9	17	3	4	18	162
	Quasi-Sym.	1.5	2.9	7.4	5.4	6.2	6.6	0	3.8	13.3	16	2.5	2.9	15.4	84
Σ	Symmetrie	118	169.5	1180	728.5	717.5	521.5	116.5	90	1975.5	228.5	198	756	88	9228
	Beobachtungen	172	308	2717	1371	1550	1082	198	258	4932	6747	360	331	1681	170
	Quasi-Sym.	75	145	906	845	854	720	130	122	2004	1886	245	295	909	92

Tabelle 3.13: Beobachtete Markenwechsel auf dem Automarkt und geschätzte Zellhäufigkeiten unter Symmetrie und Quasi-Symmetrie.

dass die erwartete Anzahl von Kunden, die von einer Marke A zur Marke B wechseln und die erwartete Anzahl der Wechsler von der Marke B zur Marke A identisch sind. Stellt man diese Annahme wie bei BISHOP, FIENBERG AND HOLLAND (1975, S. 282) durch $\lambda_{ij} = \lambda_{ji}$ für alle $i \neq j$ formal dar, überrascht es nicht, dass die Maximum-Likelihood-Schätzer der Zellhäufigkeiten (in Tabelle 3.13 jeweils über den Beobachtungen angegeben) aus dem verallgemeinerten linearen Modell mit den Mittelwerten

$$\hat{\lambda}_{ij} = \frac{y_{ij} + y_{ji}}{2} \quad \text{für alle } i \neq j$$

der Beobachtungen in den entsprechenden Zellenpaaren übereinstimmen.

Das Modell für Symmetrie impliziert die Homogenität der Randverteilungen $\sum_i \lambda_{ij} = \sum_j \lambda_{ij}$ für $i = j$. Da das Modell ebenfalls nur für die Beobachtungen angepasst wird, die nicht auf der Diagonalen liegen, bedeutet dies im hier betrachteten Beispiel, dass die Summe der verlorenen Kunden und die Summe der akquirierten Kunden für die Marken jeweils identisch ist. Die Erweiterung des Modells um die Faktoren der Zeilen- und Spaltenvariable führt zu dem Modell für Quasi-Symmetrie

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k \quad \text{für alle } i \neq j \text{ und } k = 1, \dots, (I^2 - I)/2.$$

Das Modell für Quasi-Symmetrie „... , *implies (more or less) equal transition probabilities both ways between pairs of response categories, within the constraints of marginal probabilities being those values observed, ...*“ (LINDSEY, 1997, S. 34). Betrachtet man beispielsweise das Paar der Marken Alpha und BMW, so tauschen diese Marken neun Kunden aus. (Fünf wechseln von Alpha zu BMW und vier von BMW zu Alpha.) Während unter dem Modell für Symmetrie davon ausgegangen wird, dass jeweils die Hälfte dieser neun Kunden von Alpha zu BMW bzw. von BMW zu Alpha wechseln, sind im Modell für Quasi-Symmetrie die Randverteilungen berücksichtigt. Diese zeigen, dass BMW insgesamt 194 Kunden verloren und 145 Kunden gewonnen hat. Alpha hat 161 Kunden verloren und 75 gewonnen. Beide Marken verlieren netto also Kunden, wobei der Verlust für Alpha mit 86 Kunden größer ist als für BMW mit 49 Kunden. Die relativ schlechte Situation für Alpha (bzw. die relativ gute Situation für BMW) kommt im Modell für Quasi-Symmetrie darin zum Ausdruck, dass nur von 3.4 der neun Wechsler erwartet wird, von BMW zu Alpha zu wechseln, während der Wechsel in die umgekehrte Richtung von 5.6 Kunden erwartet wird. (Die entsprechenden Schätzer der Zellhäufigkeiten sind in Tabelle 3.13 jeweils unter den Beobachtungen angegeben.)

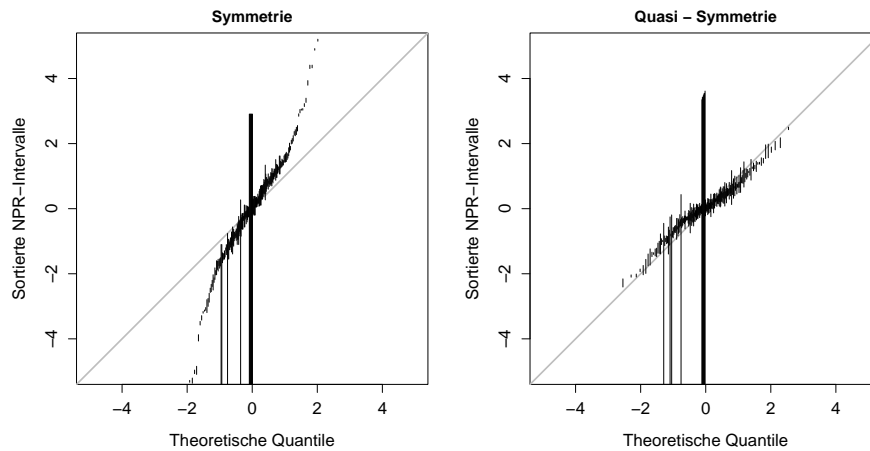


Abbildung 3.40: QQ-Plots für Symmetrie und Quasi-Symmetrie.

Abbildung 3.40 zeigt die QQ-Plots der NPR-Intervalle, die sich nach den Modellen für Symmetrie bzw. Quasi-Symmetrie ergeben. Das Modell für Symmetrie (linke Graphik) beschreibt die Daten nicht korrekt. (Einige extreme NPR-Intervalle sind nicht in der Graphik enthalten.) Im Gegensatz dazu verläuft der QQ-Plot des Modells für Quasi-Symmetrie in etwa entlang der Winkelhalbierenden. Von den bislang betrachteten Modellen beschreibt das Modell für Quasi-Symmetrie die Daten am besten.

Auch hier können differenziertere Aussagen aus der Analyse der entsprechenden Tableplots (Abbildung 3.41) abgeleitet werden. Betrachtet man im Tableplot des Modells für Symmetrie (linke Graphik) beispielsweise die zu der Marke Renault gehörige Zeile und Spalte, ist zu erkennen, dass die Zeile tendenziell helle und die Spalte tendenziell dunkle Zellen enthält. Dies deutet darauf hin, dass Renault tendenziell mehr Kunden verloren als hinzugewonnen hat. Die relativ vielen weißen bzw. schwarzen Zellen zeigen, dass die Kundenverluste häufig extrem sind. Tatsächlich lässt sich aus den Randverteilungen des Modells für Quasi-Symmetrie, die den Randverteilungen der Markenwechsler entsprechen, ablesen, dass Renault insgesamt 2 795 Kunden an andere Marken verloren hat und nur 1 886 Kunden von den anderen Anbietern abschöpfen konnte. Lediglich gegen die Marke Citroen konnte sich Renault deutlich behaupten, was an der entsprechenden schwarzen (weißen) Zelle in der zu Renault gehörenden Zeile (Spalte) zu erkennen ist. Hier steht ein Verlust von 353 Kunden einem Zugewinn von 477 Kunden gegenüber.

Im Gegensatz zu Renault konnte Seat die Anzahl der Kunden insgesamt er-

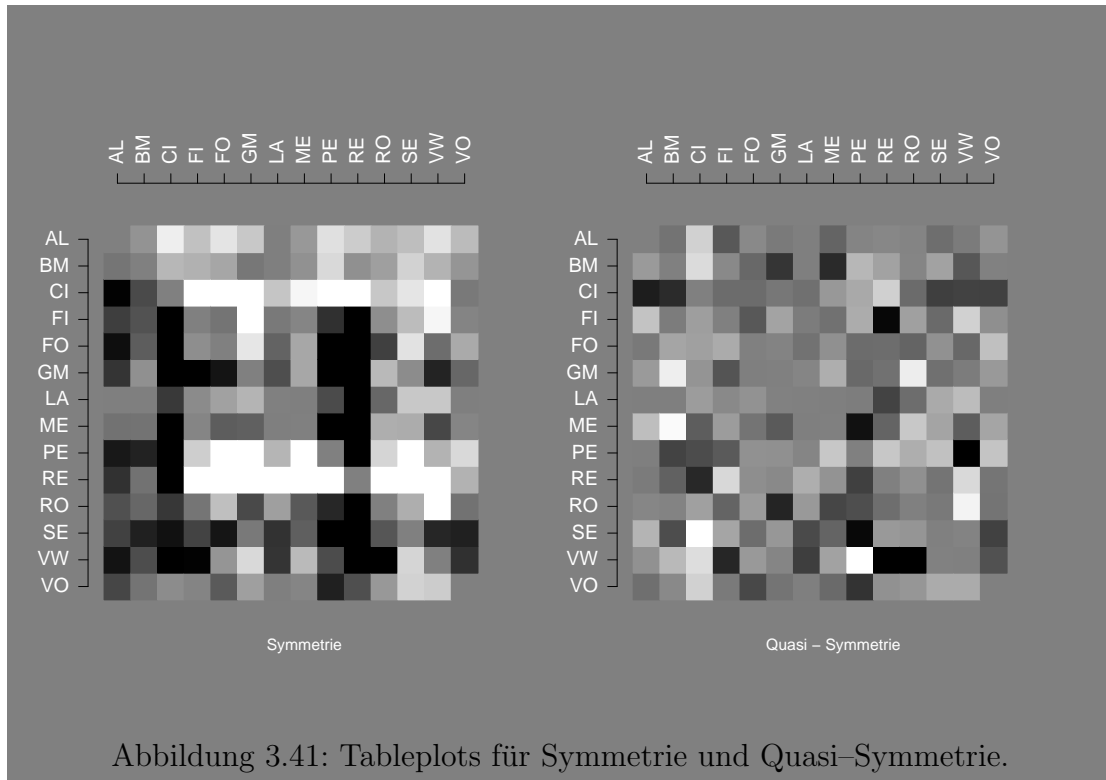


Abbildung 3.41: Tableplots für Symmetrie und Quasi-Symmetrie.

höhen. Alle Zellen der entsprechenden Spalte sind heller als der Hintergrund, d.h. Seat war in der Lage von allen Konkurrenten mehr Kunden zu akquirieren als an diese zu verlieren. Die Färbung der Zellen ist allerdings weniger extrem als bei Renault. Dies zeigt, dass der Nettogewinn mit insgesamt 194 Kunden relativ gering ist.

Bei der Analyse des Modells für Symmetrie steht die Interpretation von absoluten Zahlen im Mittelpunkt. Der große Marktanteil von Renault (zum Zeitpunkt der Erhebung knapp 31%) ist ein Grund für die hohe Anzahl der verlorenen Kunden, d.h. die Konkurrenten können aus einer großen Masse von Renaultkunden schöpfen. Das Modell für Quasi-Symmetrie ist um diesen Effekt bereinigt. Für die Annahme identischer Übergangswahrscheinlichkeiten zwischen den Paaren bei gegebenen Randverteilungen ist weder die zu Renault gehörende Zeile noch die Spalte besonders auffällig (rechter Tableplot in Abbildung 3.41). Diese Situation ist wie folgt zu interpretieren: Gegeben, dass die Situation für Renault schlecht ist (insgesamt 909 netto verlorene Kunden) und bei zusätzlicher Berücksichtigung von Zugewinnen bzw. Verlusten der anderen Marktteilnehmer, gibt es keine auffällig hohen oder niedrigen Abwanderungen zu anderen Marken.

Seat hat (wie zuvor erläutert) von allen Konkurrenten jeweils mehr Kunden abgeschöpft als an diese abgegeben. Betrachtet man beispielsweise die Austauschbeziehung mit Citroen, so sind absolut 13 Kunden zu Citroen abgewan-

dert und 28 Kunden von Citroen akquiriert worden. Im linken Tableplot für Symmetrie ist dies an der dunklen Zelle (*Seat, Citroen*) und an der hellen Zelle (*Citroen, Seat*) zu erkennen. Im Modell für Quasi-Symmetrie hingegen ist Zelle (*Citroen, Seat*) dunkler als die Zelle (*Seat, Citroen*), d.h. bei Berücksichtigung der schlechten Situation von Citroen (1454 Kunden verloren und 906 gewonnen) und der guten Situation von Seat (295 Kunden gewonnen und 101 Kunden verloren) wären weniger Abwanderungen von Seat zu Citroen (7.1 statt 13) und mehr Abwanderungen von Citroen nach Seat (33.9 statt 28) erwartet worden.

3.5.3 Modellierung zyklischer Daten

In diesem Abschnitt wird die Verwendung von verallgemeinerten linearen Modellen zur Beschreibung von zyklischen Daten betrachtet. Konkret handelt es sich um Beobachtungen für den täglichen Niederschlag über mehrere Jahre. Es ist offensichtlich, dass es sich um zyklische Daten handelt, da mit der Wiederkehr der Jahreszeiten niederschlagsbestimmende Einflüsse regelmäßig wiederkehren.

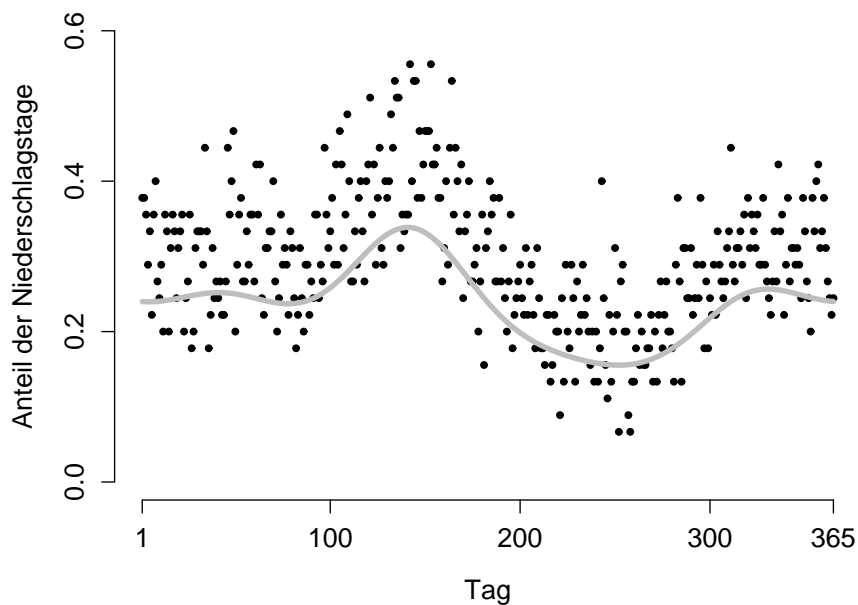


Abbildung 3.42: Anteil von Niederschlagstagen für jeden Tag des Jahres während der Jahre von 1950 bis 1994. (Die Bedeutung der grauen Kurve wird im Text erläutert.)

K	0	1	2	3	4	5	6
AIC	18 800.23	18 687.21	18 586.55	18 585.24	18 581.83	18 584.18	18 585.32

Tabelle 3.14: Anzahl überlagernder Sinus- und Kosinusfrequenzen und AIC der entsprechenden Modelle.

Die Punkte in Abbildung 3.42 zeigen die durchschnittliche Anzahl von Niederschlagstagen für jeden Tag des Jahres, wobei ein Tag als Niederschlagstag definiert ist, wenn mindestens 0.1 mm Niederschlag beobachtet wurde. (Die Bedeutung der eingezeichneten Kurve wird weiter unten erläutert.) Jeder Durchschnittswert basiert auf 45 Messungen, die während der Jahre von 1950 bis 1994 in Ivailo (Bulgarien) durchgeführt wurden. Eine Bereinigung um Schaltjahre erfolgte, indem Niederschlagsmengen am 29. Februar dem 01. März zugerechnet wurden. Die Daten stammen aus der *Hydro meteorological network data base of the National hydro meteorological service* und stellen einen Ausschnitt der bei ZUCCHINI, NEYKOV AND NEYTCHEV (2001a) verwendeten Daten zur Entwicklung eines Modells für den täglichen Niederschlag in Süd-West Bulgarien dar.

Die Beschreibung von täglichem Niederschlag erfolgt in der Regel durch zwei Modelle. Dabei beschreibt eines der Modelle, ob ein Niederschlagstag vorliegt oder nicht, und ein weiteres Modell beschreibt die Menge des Niederschlags unter der Bedingung, dass ein Niederschlagstag vorliegt (vgl. WALDEN AND GUTTORP, 1992, S. 72). Die hier erläuterten Methoden der Modellüberprüfung beziehen sich auf Modelle der ersten Art. Anhand eines beispielhaft ausgewählten Modells wird im Folgenden gezeigt, dass die Analyse der Residuen für die hier betrachtete Problemstellung auf sehr ähnliche Weise erfolgen kann wie im Fall der wiederholten Messungen aus dem vorletzten Abschnitt.

Insgesamt wurden die Niederschlagsmengen an $T = 16\,425$ Tagen (jeweils 365 Tage in 45 Jahren) gemessen. Da hier ein Modell für das Auftreten von Niederschlägen betrachtet wird, sind die Variablen binär codiert: $Y_t = 0$ bzw. $Y_t = 1$ für ‚kein Niederschlag‘ bzw. ‚Niederschlag‘ am Tag t . Die Wahrscheinlichkeit π_t für Niederschlag am Tag t wird hier (wie auch bei ZUCCHINI, NEYKOV AND

$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\delta}_4$
-1.189	0.160	0.284	-0.236	0.050	-0.006	0.020	0.105	-0.020	-0.069

Tabelle 3.15: Parameterschätzer des Modells für $K = 4$.

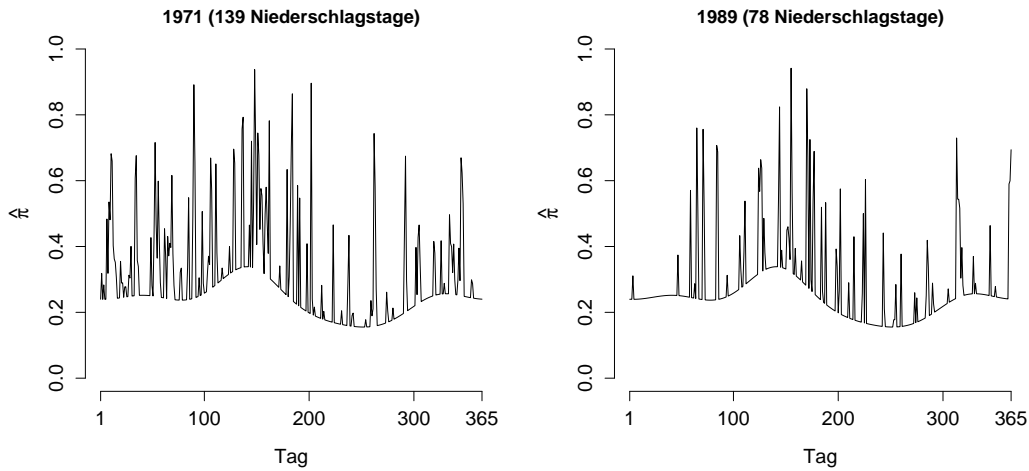


Abbildung 3.43: Geschätzte Niederschlagswahrscheinlichkeiten für die Jahre 1971 (links) und 1989 (rechts).

NEYTCHEV, 2001b) durch das Modell

$$\text{logit}(\pi_t) = \alpha + \beta\sqrt{x_{t-1}} + \sum_{k=1}^K \gamma_k \sin\left(\frac{2\pi tk}{365}\right) + \delta_k \cos\left(\frac{2\pi tk}{365}\right)$$

beschrieben. (Für eine Diskussion weiterer Modelle siehe GRUNWALD AND JONES, 2000.) Die rechte Seite der Modellgleichung kann als Komposition von zwei Komponenten verstanden werden. Eine Komponente beschreibt das jährlich wiederkehrende, zyklische Verhalten des Niederschlags. Im hier betrachteten Modell ist diese Komponente durch (jeweils K) überlagernde Sinus- und Kosinuskurven unterschiedlicher Frequenz sowie den Achsenabschnitt beschrieben. Diese zyklische Struktur ist in Abbildung 3.42 als graue Kurve eingezeichnet. Dabei beträgt die Anzahl der berücksichtigten Frequenzen vier. Dieser Wert ergibt sich im Rahmen der Modellauswahl, wenn das AIC für alternative Werte von K berechnet wird (siehe Tabelle 3.14). Die geschätzten Koeffizienten der Sinus- und Kosinuskurven und der geschätzte Achsenabschnitt sind zusammen mit dem Schätzer für β in Tabelle 3.15 angegeben.

Neben der Komponente zur Beschreibung des zyklischen Verhaltens besitzt das Modell eine Komponente zur Beschreibung der jahresspezifischen Eigenschaften. Konkret erfolgt diese im Modell durch den Term $\beta\sqrt{x_{t-1}}$, wobei x_{t-1} die beobachtete Niederschlagsmenge am Tag $t - 1$ darstellt. Abbildung 3.43 verdeutlicht die Wirkung des Terms anhand der geschätzten Niederschlagswahrscheinlichkeiten für die Jahre 1971 und 1989. Im Jahr 1971 wurden insgesamt 139 Niederschlagstage, im Jahr 1989 nur 78 Niederschlagstage beob-

achtet. Während die zyklische Komponente der überlagernden Sinus- und Kosinuskurven für beide Jahre identisch ist, besitzt die Kurve der geschätzten Erfolgswahrscheinlichkeiten für das Jahr 1971 mehr Sprünge als die des Jahres 1989. Diese Sprünge, d.h. diese jahresspezifischen Besonderheiten, resultieren aus beobachteten Niederschlägen, die jeweils am Vortag gemacht wurden, und zeigen daher, dass die geschätzte Niederschlagswahrscheinlichkeit am Tag t größer ist, wenn am Vortag Niederschlag beobachtet wurde. Weil im Jahr 1971 im Vergleich zum Jahr 1989 relativ viele Niederschlagstage beobachtet wurden, sind demnach die geschätzten Niederschlagswahrscheinlichkeiten im Jahr 1971 tendenziell größer als im Jahr 1989.

Da hier zum einen gemeinsame Charakteristika eines Jahres unter Berücksichtigung jahresspezifischer Eigenschaften modelliert werden und zum anderen binäre Beobachtungen vorliegen, kann die Struktur des Modells auf die gleiche Weise interpretiert werden wie die des Modells für den Lernprozess der Hunde aus Abschnitt 3.5.1. So können die Beobachtungen y_t für $t = 1, \dots, 16425$ auch durch y_{ij} für $i = 1, \dots, 45$ und $j = 1, \dots, 365$ notiert werden und demnach als Beobachtung für den Tag j im Jahr i aufgefasst werden. Während für die Hunde die Summe der Stromstöße in den einzelnen Versuchen bzw. die Summe der Stromstöße der einzelnen Hunde zur Überprüfung des Modells verwendet wurde, kann hier analog die Summe der Niederschlagstage an den einzelnen Tagen des Jahres

$$S_{.j} = \sum_{i=1}^{45} Y_{ij} \quad \text{für } j = 1, \dots, 365$$

bzw. die Summe der Niederschlagstage in den einzelnen Jahren

$$S_{i.} = \sum_{j=1}^{365} Y_{ij} \quad \text{für } i = 1, \dots, 45$$

zur Überprüfung des Modells verwendet werden. Auch hier können mit Hilfe der Simulation die resultierenden Poisson Binomial-Verteilungen unter dem Modell (also unter Verwendung der Schätzer $\hat{\pi}_{ij}$) approximiert und anschließend die entsprechenden NPR-Intervalle berechnet werden.

Abbildung 3.44 zeigt die NPR-Intervalle für die beobachteten Summen der Niederschlagstage an den einzelnen Kalendertagen. Der QQ-Plot in der linken Graphik verläuft recht genau entlang der Winkelhalbierenden. Dies zeigt, dass die Erfolgswahrscheinlichkeiten für Niederschläge an den einzelnen Tagen adäquat durch das Modell geschätzt werden. Lediglich eine Beobachtung

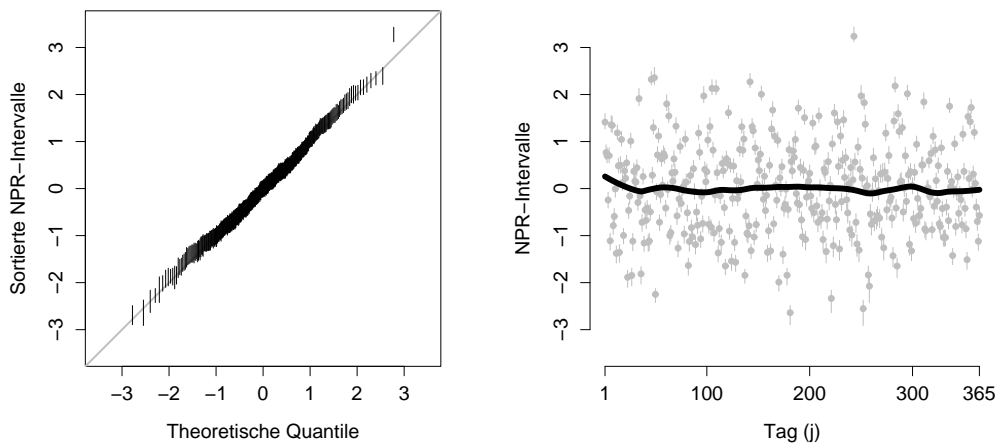


Abbildung 3.44: QQ-Plot und Indexplot der NPR-Intervalle für die Summe von Niederschlagstagen an den einzelnen Tagen des Jahres.

(rechts oben im QQ-Plot) könnte als Ausreißer unter dem Modell interpretiert werden. Es handelt sich um das Residual für den 31. August (bzw. für $j = 243$). Für diesen Tag wird die Niederschlagswahrscheinlichkeit im Durchschnitt (betrachtet wird hier nur die Summe) deutlicher unterschätzt, als es unter dem Modell (und der gegebenen Größe der Stichprobe) zu erwarten wäre.

Der Indexplot (rechte Graphik) untersucht das Verhalten der Residuen im Jahresverlauf. Wie bereits in Abbildung 3.35 wurde auch hier eine Glättung der Mittelpunkte der NPR-Intervalle (schwarze Kurve) verwendet, um die grobe Struktur der zeitlichen Entwicklung zu erfassen. Diese zeigt keine Systematik (beispielsweise in Form jahreszeitlicher Schwankungen), d.h. die zyklische Struktur der Daten ist gut durch das Modell beschrieben.

Im Gegensatz zur zyklischen Struktur sind die jahresspezifischen Eigenschaften nicht korrekt durch das Modell erfasst. Der QQ-Plot in Abbildung 3.45 ist im Vergleich zur Winkelhalbierenden deutlich gegen den Uhrzeigersinn gedreht. Die Summe der Niederschlagstage in den einzelnen Jahren ist demnach in einigen Jahren kleiner und in anderen größer, als es unter dem Modell (und der gegebenen Stichprobengröße) zu erwarten wäre. Mit anderen Worten sind die Unterschiede zwischen den Jahren stärker als die durch das Modell beschriebenen (und in Abbildung 3.43 angedeuteten) Unterschiede.

Da die einzelnen Jahre eine logische Reihenfolge besitzen (dies war für die Hunde aus Abschnitt 3.5.1 nicht der Fall) kann auch hier der Indexplot der NPR-Intervalle (rechte Graphik) zeitlich interpretiert werden. Die Glättung

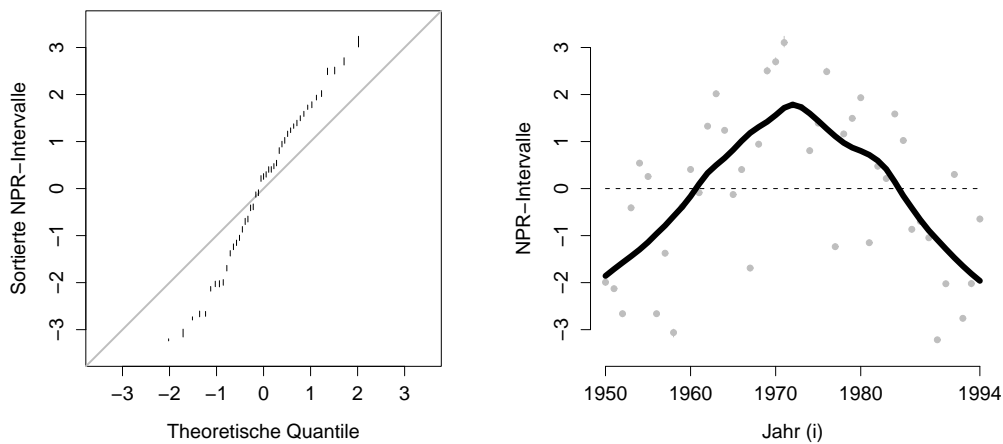


Abbildung 3.45: QQ-Plot und Indexplot der NPR-Intervalle für die Summe von Niederschlagstagen in den einzelnen Jahren.

der NPR-Intervallmittelpunkte zeigt deutlich, dass die Residuen im mittleren Bereich des Beobachtungszeitraums zu groß und an den Enden zu gering sind. Während also in den ersten und letzten Jahren die Niederschlagswahrscheinlichkeit tendenziell überschätzt wird, wird sie im mittleren Bereich tendenziell unterschätzt. Betrachtet man die *beobachtete* Summe von Niederschlagstagen, so ist diese für Jahre aus dem mittleren Bereich des Beobachtungszeitraums relativ groß und für Jahre an den Randbereichen vergleichsweise gering. (In Abbildung 3.43 wurde dies exemplarisch anhand der Jahre 1971 und 1989 mit 139 bzw. 78 Niederschlagstagen gezeigt.) Diese systematische Struktur in den *Beobachtungen* kann allerdings nicht als Rechtfertigung für die Struktur in den Residuen interpretiert werden. Letztere zeigt vielmehr, dass die jahresspezifischen Eigenschaften nicht korrekt durch den Term $\beta\sqrt{x_{t-1}}$ erfasst werden.

3.6 Abschließende Bemerkungen zum Kapitel

Mit den Darstellungen in den ersten beiden Abschnitten dieses Kapitels wurden allgemein verwendbare Vorgehensweisen am Beispiel der verallgemeinerten linearen Modelle illustriert. Auf der einen Seite standen hier graphische Methoden im Vordergrund. Dabei sind die NPR des QQ-Plots streng genommen keine graphische Methode zur Analyse der Residuen. Vielmehr stellen sie eine spezielle Definition der NPR für eine bestimmte Problemstellung dar, die häufig in der Statistik auftritt: die Beurteilung eines QQ-Plots. Ferner wurde

mit den verschiedenen Methoden zur Analyse der URR- bzw. NPR-Intervalle gezeigt, dass auch für diese Art von Residuen geeignete Graphiken für die Modellüberprüfung zu konstruieren sind. Auch wurden hier die Vorteile der Verwendung von Intervallresiduen deutlich (vgl. insbesondere Beispiel 3.6).

Auf der anderen Seite wurde ein Vorgehen dargestellt, die Residuen um den Einfluss der Parameterschätzung zu korrigieren. Während für klassische lineare Regressionsmodelle explizite Resultate abgeleitet wurden, ist für andere Situationen der Einsatz der parametrischen *Bootstrap*-Methode beschrieben worden. Hier sind der Nutzen der Kalibrierung (die Unterschiede zwischen den Residuen vor und nach Kalibrierung waren verhältnismäßig gering) und der nicht unerhebliche Rechenaufwand in der konkreten Anwendungssituation gegeneinander abzuwägen. Häufig wird es so sein, dass eine Korrektur nicht notwendig erscheint, weil die Anzahl der Parameter gering ist im Vergleich zur Anzahl der Beobachtungen. Dabei ist die Korrektur bezüglich des Einflusses der einzelnen Beobachtungen (siehe Gleichung 3.3) natürlich eine generelle, mit weniger Aufwand verbundene Alternative zur Berücksichtigung der Parameterschätzung.

Mit den drei umfangreicher bearbeiteten Beispielen am Ende des Kapitels wurde gezeigt, wie spezielle Problemstellungen bei der Modellüberprüfung bearbeitet werden können. Für das Beispiel mit den Hunden und das Beispiel mit dem täglichen Niederschlag wurde beschrieben, wie aus ganz bestimmten Anforderungen an das Modell (resultierend aus einer bestimmten Struktur der Daten), geeignete Residuen definiert werden können. Die Analyse dieser Residuen ermöglichte es, spezielle, problembezogene Modellfehler aufzudecken und somit letztlich konkrete Hinweise auf notwendige Modifikationen des Modells zu geben. An dieser Stelle wurde deutlich, dass die Analyse der Pseudo-Residuen einen wesentlichen Beitrag zum Prozess der statistischen Modellbildung leisten kann: Mit Hilfe dieser Hinweise ist es möglich, in einem nächsten Schritt ein geeigneteres Modell zu spezifizieren. (In Beispiel 3.5 zu den Arztbesuchen wurde dieser Gedanke explizit umgesetzt.) In dem Beispiel zu den Markenwechslern auf dem französischen Automarkt kam es in erster Linie auf die Identifikation von Verletzungen bestimmter Modellannahmen und den damit verbundenen Interpretationen auf die betrachtete Problemstellung an. Hier war es mit Hilfe der Pseudo-Residuen (dargestellt in den Tableplots) möglich, ganz konkrete Hilfestellungen bei der Beurteilung der Marktsituation zu leisten.

Insgesamt verdeutlicht bereits die Vielzahl der in diesem Kapitel betrachteten

Beispiele das breite Anwendungsspektrum der Pseudo-Residuen. Weitere Problemstellungen, in denen die Verwendung der Pseudo-Residuen hilfreich ist, werden in den nachfolgenden Kapiteln betrachtet.

Kapitel 4

Pseudo–Residuen bei der Modellierung von Aktienrenditen

4.1 Einführung

Bei der Modellierung von Aktienrenditen gilt das praktische Interesse in der Regel der Voraussage zukünftiger Werte. In der heutigen Zeit ist es üblich, diese Voraussage in Form einer Verteilung anzugeben (also nicht als Punkt– oder Intervallvoraussage). Verteilungsprognosen besitzen den Vorteil, dass verschiedene Fragen der Anwender bezüglich zukünftiger Beobachtungen beantwortet werden können, beispielsweise wie groß die Varianz oder der *Value at Risk* eines Finanztitels (unter dem Modell) sein wird. (Für eine allgemeinere Darstellung der Vorteile siehe DIEBOLD, GUNTHER AND TAY, 1998.) Zeitreihenmodelle mit Verteilungsprognosen liefern, basierend auf den Informationen bis zur Periode $t - 1$ (im Folgenden durch $\Omega_{t-1} = \{Y_1 = y_1, \dots, Y_{t-1} = y_{t-1}\}$ gekennzeichnet), eine Voraussageverteilung für die Folgeperiode. Beschreibt man die Voraussageverteilung durch ihre Verteilungsfunktion, entspricht das Modell

$Y_t|\Omega_{t-1}$ besitzt die Verteilungsfunktion $F_{Y_t|\Omega_{t-1}}(y; \boldsymbol{\theta}_t)$ für $t = 2, \dots, T$

der allgemeinen Darstellung des stochastischen Modells aus Abschnitt 2.2, wobei $\boldsymbol{\theta}_t$ den entsprechenden Parametervektor der Verteilung darstellt. Pseudo–Residuen, definiert durch

$$\begin{aligned} U_t &= F_{Y_t|\Omega_{t-1}}(Y_t|\Omega_{t-1}; \boldsymbol{\theta}_t) \\ \text{bzw. } R_t &= \Phi^{-1}(U_t) \quad \text{für } t = 2, \dots, T, \end{aligned}$$

bieten demnach die Möglichkeit, die Güte der Voraussagen eines Zeitreihenmodells zu beurteilen. Diese Art von Pseudo-Residuen wird von MACDONALD AND ZUCCHINI (1997, S. 99–101) als *forecast pseudo-residuals* bezeichnet. Die UPR werden seit den 1990-er Jahren bei der Analyse von Zeitreihenmodellen angewendet (vgl. beispielsweise ELERIAN, CHIB AND SHEPARD, 1998; PEDERSEN, 1994 sowie SEILLIER-MOISEWITSCH, 1993). Eine systematischere Betrachtung verschiedener Eigenschaften der UPR erfolgt bei DIEBOLD, GUNTHER AND TAY (1998), die auch auf die Möglichkeit hinweisen, dass die Residuen zur Verbesserung der Voraussagen verwendet werden können. Das entsprechende Vorgehen wird in einer späteren Publikation von DIEBOLD, HAHN AND TAY (1999) beschrieben. Auch die NPR werden bei der Analyse der Voraussagen verwendet (vgl. beispielsweise KIM, SHEPARD AND CHIB, 1998). Ihre Anwendung ist jedoch weniger verbreitet als die der UPR. Dass die Überprüfung von Zeitreihenmodellen für Finanztitel ein aktuell diskutiertes Thema ist, verdeutlicht auch die jüngere Arbeit von BERKOWITZ (2001), der (basierend auf den UPR) einen Test für die Gültigkeit der Voraussageverteilungen beschreibt. Für einen Überblick und weitere Literaturhinweise sei auf TAY AND WALLIS (2000) verwiesen.

Der Rest des Kapitels ist wie folgt gegliedert: Im folgenden Abschnitt werden drei ausgewählte Zeitreihenmodelle zur Voraussage von Aktienrenditen beschrieben. Anschließend erfolgt eine kurze Beschreibung der in diesem Kapitel modellierten Renditezeitreihe der Deutsche Bank-Aktie und der Modellanpassung. In Abschnitt 4.4 erfolgt die Überprüfung verschiedener Kriterien der Voraussageverteilungen. Es wird sich zeigen, dass keines der Modelle eine insgesamt befriedigende Anpassung liefert. Daher wird im letzten Teil des Kapitels ein Vorgehen zur Kalibrierung der Voraussagen vorgestellt, die zu einer Verbesserung der Verteilungsprognosen führt.

4.2 Ausgewählte Zeitreihenmodelle zur Modellierung von Aktienrenditen

4.2.1 GARCH-Modelle

Die grundlegende Idee der GARCH-Modelle (GARCH steht für *Generalized AutoRegressive Conditional Heteroskedastic*) liegt in der Abhängigkeit der Varianz der Voraussageverteilungen von den vorangegangenen Beobachtungen und den Varianzen zurückliegender Voraussageverteilungen (vgl. BOLLERS-

LEV, 1986). Unterstellt man normalverteilte Renditen, kann ein GARCH(p,q)-Modell allgemein durch

$$Y_t | \Omega_{t-1} \sim N(0, h_t)$$

$$\text{mit } h_t = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i}^2 + \sum_{j=1}^q \beta_j h_{t-j}$$

beschrieben werden. Die bedingte Verteilung von Y_t , gegeben alle vorherigen Beobachtungen (also die Voraussageverteilung), ist eine Normalverteilung mit Erwartungswert null und einer Varianz h_t , die eine lineare Funktion der p vorangegangenen quadrierten Beobachtungen und der q vorangegangenen bedingten Varianzen ist.

Die Modelle stellen eine Erweiterung der ARCH-Modelle (vgl. ENGLE, 1982) dar, bei denen die bedingte Varianz lediglich von den vorausgegangenen Beobachtungen der Zeitreihe abhängt. Die Entwicklung beider Modelle ist aus der Beobachtung ökonomischer Zeitreihen motiviert, deren Varianz häufig nicht konstant ist und somit bei der Modellierung berücksichtigt werden muss. Beide Arbeiten sind also im Kontext der wachsenden Erkenntnis zu sehen, dass die Volatilität im Bereich der Finanzwissenschaft eine bedeutende Rolle spielt. SHEPHARD (1996) bezeichnet Modelle dieser Art als *volatility models* und gibt einen Überblick zu ARCH- und GARCH-Modellen sowie eng verwandten stochastischen Prozessen. Die große Bedeutung der Modellierungsidee (die dynamische Varianz) spiegelt sich in zahlreichen Arbeiten zu spezielleren Problemstellungen wider, die relativ kurz nach der Publikation von ENGLE (1982) veröffentlicht wurden. (Für einen Überblick siehe BOLLERSLEV, CHOU AND KRONER, 1992.)

Im Rahmen dieser Arbeit wird ein GARCH(1,1)-Modell betrachtet. Für dieses Modell reduziert sich die bedingte Varianz auf

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}.$$

Als Verteilungstyp der Voraussageverteilung wird hier eine Normalverteilung unterstellt. Mit dieser Annahme wird den Ausführungen von BOLLERSLEV (1986) gefolgt. Es sei allerdings darauf hingewiesen, dass auch die Verwendung anderer Verteilungen denkbar ist. Weil die Modellformulierung der GARCH-Modelle auf den bedingten Verteilungen von Y_t basiert, ergibt sich für die Voraussagen des GARCH(1,1)-Modells mit Normalverteilungsannahme unmittelbar

$$Y_t | \Omega_{t-1} \sim N(0, \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}).$$

Da nach dem Modell der Erwartungswert von Y_t für alle Perioden den Wert null besitzt, erfolgt die Anpassung an die um den Stichprobenmittelwert \bar{y} bereinigten Daten. Somit ergibt sich für die Pseudo-Residuen des GARCH(1,1)-Modells mit Normalverteilungsannahme

$$U_t = \Phi \left(\frac{Y_t - \bar{y}}{\sqrt{\alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}}} \right)$$

bzw. $R_t = \frac{Y_t - \bar{y}}{\sqrt{\alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}}}.$

Um die Residuen ab der zweiten Periode berechnen zu können, wird die Varianz ab der ersten Periode benötigt. Weil diese sich nicht direkt aus dem Modell ergibt, wird sie hier durch den Schätzer der unbedingten Varianz des Prozesses ersetzt. Letztere ist durch $\alpha_0/(1 - \alpha_1 - \beta_1)$ gegeben.

4.2.2 Hidden Markov-Modelle

Bei Hidden Markov-Modellen ist die Verteilung von Y_t durch den Zustand einer Markov-Kette zum Zeitpunkt t gegeben. Insofern besteht das Modell aus zwei Zufallsmechanismen:

- (1) Einem Zustandsprozess C_t , der durch eine Markov-Kette beschrieben wird.
- (2) Den zustandsabhängigen Verteilungen von Y_t , aus denen die Beobachtungen y_t realisiert werden.

Dabei wird angenommen, dass die bedingte Verteilung von Y_t , gegeben Ω_{t-1} und alle c_j für $j \leq t$, ausschließlich von C_t abhängig ist. Diese Annahme wird als bedingte Unabhängigkeit der Y_t bezeichnet. Ferner ist zu bemerken, dass nur die Realisationen y_t beobachtet werden und die Zustände c_t der Markov-Kette im Verborgenen (*Hidden*) bleiben.

Für zwei Zustände verdeutlicht Tabelle 4.1 die Zusammenhänge zwischen der interessierenden Variable Y_t und der Markov-Kette. Die Zeilen der Tabelle repräsentieren den Zustand der Markov-Kette in der Periode $t - 1$. Im Hidden Markov-Modell implizieren diese Zustände zustandsabhängige Verteilungen, die in der Tabelle mit f_1 bzw. f_2 bezeichnet sind. Die Wahrscheinlichkeit, dass sich der Zustand der Markov-Kette beim Übergang zur Periode t (repräsentiert in den Spalten) ändert, beträgt γ_1 , wenn sich die Markov-Kette zum Zeitpunkt

		t	
		$C_t = 1$ bzw. f_1	$C_t = 2$ bzw. f_2
$t - 1$	$C_{t-1} = 1$ bzw. f_1	$1 - \gamma_1$	γ_1
	$C_{t-1} = 2$ bzw. f_2	γ_2	$1 - \gamma_2$

Tabelle 4.1: Zustände und Übergangswahrscheinlichkeiten eines Hidden Markov-Modells.

$t - 1$ in Zustand eins befindet. Analog findet ein Übergang von Zustand zwei zu Zustand eins mit einer Wahrscheinlichkeit von γ_2 statt. Die Wahrscheinlichkeiten, dass sich der Zustand nicht ändert, betragen daher $1 - \gamma_1$ bzw. $1 - \gamma_2$. Verdeutlicht man sich das Geschehen für eine beliebige Anzahl von Perioden, wird die Idee der Hidden Markov-Modelle ersichtlich: Die Realisationen von Y_t stammen in jeder Periode entweder aus f_1 oder aus f_2 , wobei die Wahrscheinlichkeit, dass eine Beobachtung aus f_1 bzw. f_2 realisiert wird, ausschließlich vom Zustand der Markov-Kette in der Vorperiode abhängt. Diese Prozesslogik gilt analog für mehr als zwei Zustände, d.h. die Anzahl der Zustände ist allgemein nicht auf zwei begrenzt.

Die Anwendungsbereiche der Hidden Markov-Modelle sind breit gefächert. Eine intensive Nutzung erfolgt seit einigen Jahrzehnten im Bereich der Spracherkennung (einen Überblick geben JUANG AND RABINER, 1991). Auch bei der Modellierung von Aktienrenditen sind Hidden Markov-Modelle sehr hilfreich (vgl. JUANG, 1985 sowie RYDÉN, TERSÄSVIRTA AND ÅSBRINK, 1998). Allerdings ist ihre Anwendung in diesem Bereich derzeit noch nicht sehr weit verbreitet. Für einen umfassenden Überblick zu theoretischen Aspekten und Anwendungsmöglichkeiten der Modelle sei auf MACDONALD AND ZUCCHINI (1997) verwiesen.

Im Rahmen dieser Arbeit wird ein Hidden Markov-Modell mit zwei Zuständen betrachtet, wobei für die zustandsabhängigen Verteilungen eine Normalverteilung und eine symmetrische Exponentialverteilung verwendet werden. Letztere ist auch unter der Bezeichnung doppelte Exponentialverteilung oder Laplace-Verteilung bekannt (vgl. ORD, 1983) und ihre Dichtefunktion ist

$$f_2 = \lambda_2/2 \cdot e^{-\lambda_2|y-\mu_2|} \quad y \in R.$$

Um die Voraussageverteilung eines Hidden Markov-Modells ableiten zu können, werden zwei weitere Definitionen benötigt. Zum einen wird die Matrix der

Übergangswahrscheinlichkeiten oder kurz Übergangsmatrix

$$\Gamma = \begin{pmatrix} 1 - \gamma_1 & \gamma_1 \\ \gamma_2 & 1 - \gamma_2 \end{pmatrix}$$

definiert, deren Interpretation aus den Erläuterungen zu Tabelle 4.1 hervorgeht. Zum anderen sind für die Voraussageverteilung die Gleichgewichtswahrscheinlichkeiten $\delta = (\delta_1, \delta_2)$, die als Randwahrscheinlichkeiten der Markov-Kette interpretiert werden können, von Bedeutung. Sie sind für irreduzible und homogene Markov-Ketten eindeutig aus der Übergangsmatrix zu bestimmen und für zwei Zustände gilt

$$\delta = \frac{1}{\gamma_1 + \gamma_2}(\gamma_2, \gamma_1).$$

Unter Verwendung dieser Definitionen erhält man für die bedingte Dichte von Y_t , gegeben alle Beobachtungen bis zum Zeitpunkt $t - 1$, d.h. für die Dichtefunktion der Voraussageverteilung (vgl. MACDONALD AND ZUCCHINI, 1997, S. 78 und S. 82):

$$\begin{aligned} f_{Y_t|\Omega_{t-1}}(y) &= \frac{f_{Y_1 Y_2 \dots Y_t}(y_1, y_2, \dots, y)}{f_{Y_1 Y_2 \dots Y_{t-1}}(y_1, y_2, \dots, y_{t-1})} \\ &= \frac{\delta \lambda(y_1) \Gamma \lambda(y_2) \Gamma \dots \lambda(y_{t-1}) \Gamma \lambda(y) (1, 1)'}{\delta \lambda(y_1) \Gamma \lambda(y_2) \Gamma \dots \lambda(y_{t-2}) \Gamma \lambda(y_{t-1}) (1, 1)'} \end{aligned}$$

mit

$$\lambda(y_t) = \begin{pmatrix} f_1(y_t) & 0 \\ 0 & f_2(y_t) \end{pmatrix}.$$

Definiert man zur Vereinfachung der Dichtefunktion

$$(c_t, d_t) = \delta \lambda(y_1) \Gamma \lambda(y_2) \Gamma \dots \lambda(y_{t-2}) \Gamma \lambda(y_{t-1}),$$

wird nach Einsetzen von (c_t, d_t) in die Dichtefunktion der Voraussageverteilung deutlich, dass diese für eine beliebige Periode t eine Mischverteilung mit der Dichtefunktion

$$\begin{aligned} f_{Y_t|\Omega_{t-1}}(y) &= \frac{c_t(1 - \gamma_1) + d_t \gamma_2}{c_t + d_t} \cdot f_1(y) \\ &+ \frac{c_t \gamma_1 + d_t(1 - \gamma_2)}{c_t + d_t} \cdot f_2(y) \end{aligned}$$

ist. Aus dieser Darstellung ergeben sich unmittelbar die zur Berechnung der Pseudo-Residuen benötigten Verteilungsfunktionen

$$F_{Y_t|\Omega_{t-1}}(y) = \frac{c_t(1-\gamma_1) + d_t\gamma_2}{c_t + d_t} \cdot \Phi\left(\frac{y - \mu_1}{\sigma_1}\right) + \frac{c_t\gamma_1 + d_t(1-\gamma_2)}{c_t + d_t} \cdot F_2(y), \quad (4.1)$$

wobei $F_2(y)$ (die Verteilungsfunktion der doppelten Exponentialverteilung) durch

$$F_2(y) = \begin{cases} 1/2 \cdot e^{-\lambda_2(\mu_2 - y)} & y \leq \mu_2 \\ 1 - 1/2 \cdot e^{-\lambda_2(y - \mu_2)} & y > \mu_2 \end{cases}$$

gegeben ist.

4.2.3 Modelle mit gleitenden Dichten

Modelle mit gleitenden Dichten unterstellen, dass die Verteilung zukünftiger Werte durch die Dichtefunktion

$$f_{Y_t|\Omega_{t-1}}(y) = \alpha\varphi(y; \mu_{t-1}, \sigma_{t-1}^2) + (1 - \alpha)f_{Y_{t-1}|\Omega_{t-2}}(y)$$

beschrieben werden kann, wobei φ die Dichtefunktion einer Normalverteilung mit den Parametern

$$\begin{aligned} \mu_{t-1} &= y_{t-1} \\ \sigma_{t-1}^2 &= (\beta + \gamma | y_{t-1} |)^2 \end{aligned}$$

darstellt (vgl. NEUMANN, 2000, S. 145–146). Das Modell mit gleitenden Dichten basiert auf der Idee des exponentiellen Glättens. Letzteres schätzt das aktuelle Niveau einer Zeitreihe durch

$$N_t = \alpha y_{t-1} + (1 - \alpha)N_{t-1} \quad \text{mit } N_1 = y_1,$$

d.h. durch eine Funktion der vorangegangenen Beobachtung und des Niveaus aus der Vorperiode. In Analogie setzt sich die Voraussageverteilung bei Modellen mit gleitenden Dichten zusammen aus: (1) Einer Verteilung, deren Parameter von der vorangegangenen Beobachtung abhängt und die mit Gewicht α berücksichtigt wird, und (2) der Voraussageverteilung aus der Vorperiode, die mit Gewicht $1 - \alpha$ berücksichtigt wird. Es sei bemerkt, dass die hier verwendete Normalverteilung nur eine Alternative bei der Spezifikation eines Modells mit

gleitenden Dichten darstellt. Je nach Problemstellung kann die Dichtefunktion der Normalverteilung durch die Dichtefunktion anderer Verteilungen (mit geeigneten Parametern) ersetzt werden.

Modelle mit gleitenden Dichten wurden erstmals bei ZUCCHINI AND NEUMANN (1999) und NEUMANN (2000) untersucht. In beiden Arbeiten wird die Güte des Modells bei der Prognose des *Value at Risk* von Aktien mit anderen Modellen verglichen. Untersucht wurden (neben weiteren Modellen) auch die in den beiden vorangegangenen Abschnitten dargestellten Modelle.

Zur Berechnung der Pseudo-Residuen ist es sinnvoll, die Dichtefunktionen der Voraussageverteilungen umzuschreiben und in der Form

$$f_{Y_i|\Omega_{t-1}}(y) = \sum_{i=0}^{t-2} \alpha(1-\alpha)^i \varphi(y; \mu_{t-1-i}, \sigma_{t-1-i}^2) + (1-\alpha)^{t-1} f_{Y_1}(y)$$

darzustellen. Unter der Annahme, dass Y_1 ebenfalls normalverteilt ist, kann die Voraussageverteilung für die Periode t demnach als Mischung von t Normalverteilungen ausgedrückt werden, wobei das Gewicht für weiter in der Vergangenheit liegende Verteilungen exponentiell abnimmt. Diese Darstellung hat den Vorteil, dass die Werte der Verteilungsfunktionen der Voraussagen

$$F_{Y_i|\Omega_{t-1}}(y) = \sum_{i=0}^{t-2} \alpha(1-\alpha)^i \Phi\left(\frac{y - y_{t-1-i}}{\beta + \gamma |y_{t-1-i}|}\right) + (1-\alpha)^{t-1} \Phi\left(\frac{y - \mu}{\sigma}\right)$$

auf einfache Weise mit \mathbf{R} berechnet werden können, wobei μ und σ hier den Erwartungswert und die Standardabweichung in der ersten Periode beschreiben. Mit anderen Worten wird der Rechenaufwand bei der Bestimmung der Pseudo-Residuen erheblich reduziert, da auf die Verwendung numerischer Methoden bei der Integration der Dichtefunktionen verzichtet werden kann.

4.3 Daten und Modellanpassung

Die Überprüfung (und die Kalibrierung) der zuvor beschriebenen Modelle wird anhand der Renditezeitreihe für die Aktie der Deutschen Bank illustriert. Modelliert werden die kontinuierlichen Eintagesrenditen

$$Y_t = \log \frac{P_t}{P_{t-1}},$$

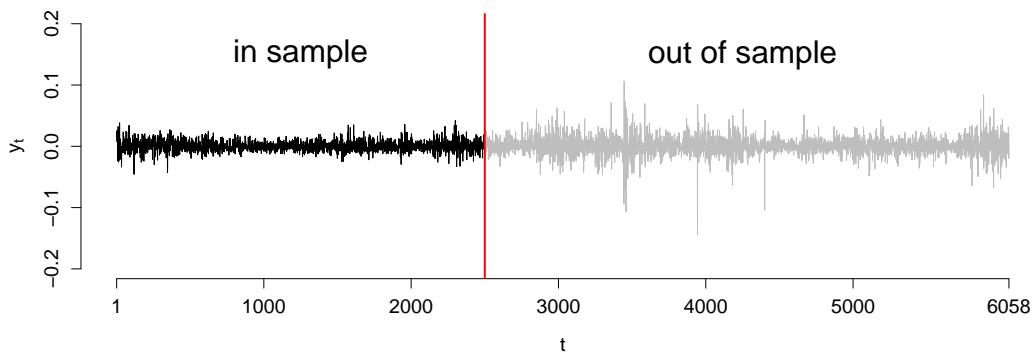


Abbildung 4.1: Zeitreihe für die kontinuierlichen Renditen der Aktie der Deutschen Bank.

GARCH(1,1)-Modell	Hidden Markov-Modell	Modell mit gleitenden Dichten
$\hat{\alpha}_0 = 3.02 \cdot 10^{-7}$	$\hat{\mu}_1 = 4.63 \cdot 10^{-5}$	$\hat{\alpha} = 3.94 \cdot 10^{-3}$
$\hat{\alpha}_1 = 4.56 \cdot 10^{-2}$	$\hat{\sigma}_1 = 6.74 \cdot 10^{-3}$	$\hat{\beta} = 1.00 \cdot 10^{-7}$
$\hat{\beta}_1 = 9.53 \cdot 10^{-2}$	$\hat{\mu}_2 = -3.71 \cdot 10^{-12}$	$\hat{\gamma} = 6.69 \cdot 10^{-2}$
	$\hat{\lambda}_2 = 119.50$	
	$\hat{\gamma}_1 = 4.45 \cdot 10^{-3}$	
	$\hat{\gamma}_2 = 4.75 \cdot 10^{-3}$	

Tabelle 4.2: ML-Schätzer der drei Modelle für die Renditezeitreihe.

wobei P_t den Kurs der Aktie zum Zeitpunkt t darstellt. Weitere Möglichkeiten, die Preisänderungen der Aktie zu definieren, sind beispielsweise bei NEUMANN (2000, S. 72–75) beschrieben. Insgesamt stehen 6058 Beobachtungen für die Zeit vom 03.01.1974 bis zum 31.03.1998 zur Verfügung (siehe Abbildung 4.1). Die Daten sind in zwei Perioden unterteilt: Die ersten 2500 Beobachtungen (*in sample*) werden verwendet, um die Parameter der Modelle zu schätzen. Die angepassten Modelle werden dann verwendet, um die Voraussageverteilungen für den übrigen Zeitraum (*out of sample*) zu erzeugen. Überprüft (und später kalibriert) werden die Voraussageverteilungen für die *out of sample*-Periode.

Die Ergebnisse der zum Teil sehr aufwendigen Parameterschätzungen werden aus der Arbeit von NEUMANN (2000) übernommen (Tabelle 4.2 zeigt die entsprechenden ML-Schätzer). Für das Modell mit gleitenden Dichten muss erwähnt werden, dass die Parameter μ und σ^2 (dies sind die Parameter der Normalverteilung aus der ersten Periode) durch den Stichprobenmittelwert

bzw. die Stichprobenvarianz der *in sample*-Periode ersetzt werden.

Die hier betrachteten Daten und Modelle stellen einen Ausschnitt aus der umfangreichen Untersuchung von NEUMANN (2000) dar. Sie untersuchte die Modelle im Hinblick auf ihre Güte bei der Beschreibung der Autokorrelation der quadrierten Renditen und der Kurtosis der Renditen. Ferner betrachtete sie die Güte der Modelle bei der Voraussage des *Value at risk*, was als Überprüfung der Anpassung an den Enden der Voraussageverteilungen verstanden werden kann. Durch eine Analyse der Pseudo-Residuen, die im folgenden Abschnitt durchgeführt wird, erhält man ebenfalls Informationen zu diesen und weiteren Gütekriterien.

4.4 Überprüfung der Voraussageverteilungen

4.4.1 Grundlegende Gestalt der Voraussageverteilungen

In einem ersten Schritt wird hier die grundlegende Gestalt der Voraussageverteilungen überprüft. Genauer formuliert geht es um die Frage, ob die beobachteten Renditen als Realisationen der entsprechenden Voraussageverteilungen aufgefasst werden können, wobei spezielle Aspekte (wie beispielsweise die Erfassung extremer Beobachtungen) in den Hintergrund rücken. Wenn die grundlegende Gestalt der Voraussageverteilungen korrekt ist, verhalten sich die UPR wie zufällige Realisationen einer $U(0, 1)$ -Verteilung. Letzteres lässt sich, wie bereits im einführenden Beispiel zu den Bäumen gezeigt, gut durch eine Analyse der Histogramme der UPR überprüfen.

Abbildung 4.2 zeigt die Histogramme der UPR der *out of sample*-Periode für die drei betrachteten Zeitreihenmodelle. Neben der Dichtefunktion der $U(0, 1)$ -Verteilung sind in den Graphiken 95%-Konfidenzbänder (gestrichelte Linien) für die einzelnen Säulen des Histogramms eingezeichnet, d.h. die Wahrscheinlichkeit, dass die Säulen des Histogramms den markierten Bereich verlassen, beträgt bei Gültigkeit des Modells jeweils 5%. Die Konfidenzbänder basieren auf einer Binomialverteilung mit den Parametern $n = 3558$ (die Anzahl von Residuen in der *out of sample*-Periode) und $\pi = 1/20$ (die Anzahl der Klassen). Dies ist die Verteilung der Häufigkeiten in den einzelnen Klassen, wenn die UPR eine $U(0, 1)$ -Verteilung besitzen. Für das Modell mit gleitenden Dichten (rechte Graphik) scheint die grundlegende Gestalt der Voraussageverteilungen korrekt zu sein. Für das GARCH(1,1)- und das Hidden Markov-Modell hingegen weicht die beobachtete Verteilung der UPR deutlich von der

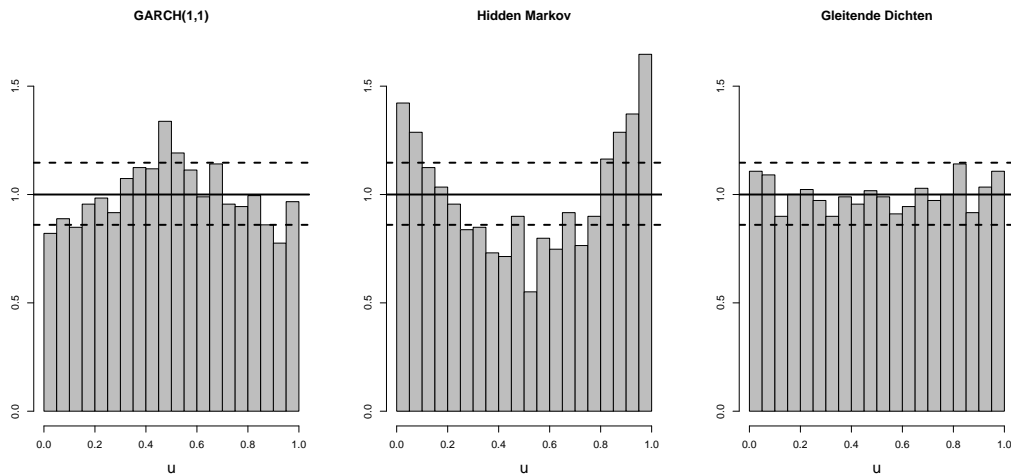


Abbildung 4.2: Histogramme der UPR der drei Modelle.

$U(0, 1)$ -Verteilung ab: Für das GARCH-Modell liegen mehr Beobachtungen um den Wert 0.5 sowie weniger Beobachtungen nahe null und eins vor, als es bei Gültigkeit des Modells zu erwarten wäre. Dies bedeutet, dass die Voraussageverteilungen des Modells tendenziell zu breit sind. Der umgekehrte Fall ist für das Hidden Markov-Modell zu beobachten, dessen Voraussageverteilungen im Durchschnitt zu schmal sind. Hier und in den folgenden Darstellungen ist allerdings zu berücksichtigen, dass in praktischen Situationen die Parameter der Modelle in sinnvollen Zeitabständen zu aktualisieren sind, was eine verbesserte Anpassung erwarten lässt.

4.4.2 Voraussage des VaR

Eines der Untersuchungsziele bei NEUMANN (2000) war die Beurteilung der Modelle in Hinblick auf die Voraussage des 99%-*Value at Risk*. Der *Value at Risk* (im Folgenden mit *VaR* abgekürzt) ist ein Risikomaß zur Messung des Verlusts eines Finanztitels innerhalb einer bestimmten Haltedauer (hier ein Tag), der mit gegebener Wahrscheinlichkeit (hier 0.99) nicht überschritten wird (vgl. JORION, 1997, S. 17). Aus statistischer Sicht handelt es sich um die Betrachtung der 1%-Quantile der Voraussageverteilungen. Je besser die Anpassung der Voraussageverteilungen in diesem Bereich ist, desto genauer wird der 99%-*VaR* vorausgesagt. Zur Überprüfung dieser Eigenschaft wird bei NEUMANN (2000) der Anteil der Beobachtungen, die kleiner sind als das 1%-Quantil der entsprechenden Voraussageverteilungen, bestimmt. Je weiter dieser Anteil von 0.01 entfernt lag, desto schlechter wurde das entsprechende

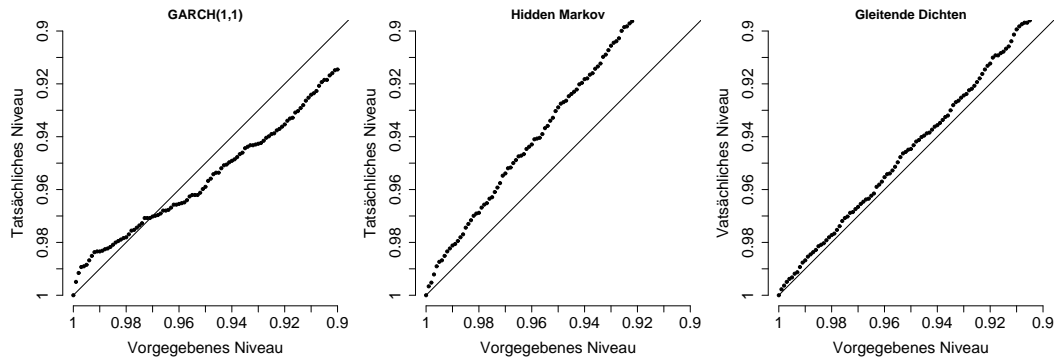


Abbildung 4.3: Vorgegebenes und tatsächliches Niveau bei Schätzung des VaR .

Modell bezüglich der Voraussage des VaR beurteilt. Mit Hilfe der Pseudo-Residuen ist es möglich, diese Analyse für ein beliebiges vorgegebenes Niveau $(1 - \alpha)$ durchzuführen, weil der Anteil der UPR, die kleiner sind als α , und der Anteil der Voraussagen, bei dem der $(1 - \alpha)$ - VaR unterschätzt wird, identisch sind. Ersterer kann daher als Schätzer für das tatsächliche Niveau verstanden und mit dem vorgegebenen Niveau verglichen werden.

In Abbildung 4.3 sind für die drei Modelle vorgegebenes und tatsächliches VaR -Niveau für $(1 - \alpha) \in [0.9, 1]$ gegeneinander geplottet. Es ist zu sehen, dass keines der Modelle das vorgegebene Niveau exakt einhält. Beispielsweise liegen die Punkte für das GARCH(1,1)-Modell im linken Bereich der Graphik über der Winkelhalbierenden. In diesem Bereich wird der VaR unterschätzt, d.h. extrem kleine Beobachtungen der Zeitreihe werden häufiger realisiert als unter dem Modell erwartet. Am besten erscheint die Einhaltung des Niveaus für das Modell mit gleitenden Dichten. Dieses Ergebnis deckt sich mit den Resultaten von NEUMANN (2000, S. 185), die berichtet, dass die Einhaltung eines angestrebten VaR -Niveaus für das Modell mit gleitenden Dichten besser ist als bei allen übrigen untersuchten Modellen.

4.4.3 Abbildung extremer Beobachtungen

In diesem Abschnitt wird überprüft, ob die Modelle die extremen Renditen korrekt erfassen. Ähnlich wie im letzten Abschnitt wird die Anpassung an den Enden der Voraussageverteilungen betrachtet. Es werden hier jedoch beide Enden betrachtet und die extremen Beobachtungen betont, indem die QQ-Plots der NPR analysiert werden (siehe Abbildung 4.4). Anhand der QQ-Plots (mit

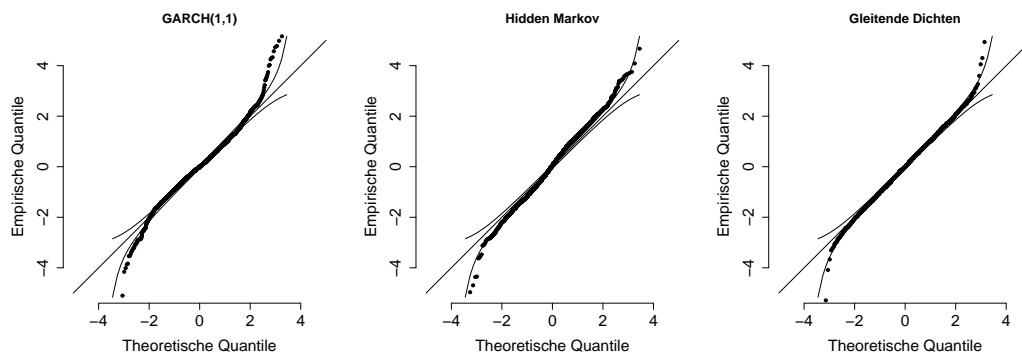


Abbildung 4.4: QQ-Plots der NPR mit simultanen Konfidenzbändern.

den simultanen 95%-Konfidenzbändern aus Abschnitt 3.3.1) ist zu erkennen, dass die extremen Quantile der beobachteten NPR für alle Modelle am linken Ende kleiner und am rechten Ende größer sind, als es bei Gültigkeit der Modelle zu erwarten ist. (Für das GARCH-Modell und für das Modell mit gleitenden Dichten sind jeweils vier, für das Hidden Markov-Modell ein Ausreißer nicht in den QQ-Plots abgebildet.) Die Abweichungen an Randbereichen des QQ-Plots resultieren aus zu kurzen Enden der Voraussageverteilungen, d.h. keines der Modelle ist in der Lage, die extremen Beobachtungen korrekt abzubilden. Insbesondere gilt dies für das GARCH(1,1)-Modell (linke Graphik).

Dass die QQ-Plots des GARCH- und des Hidden Markov-Modells auch im mittleren Bereich von der Winkelhalbierenden abweichen, ist in Abbildung 4.4 nur schwer zu erkennen. Daher sind in Abbildung 4.5 die NPR der QQ-Plots (vgl. Abschnitt 3.3.1) separat dargestellt. Auch wenn diese hoch korreliert sind und eine entsprechende Vorsicht bei der Interpretation der Ergebnisse berücksichtigt wird, ist es offensichtlich, dass die QQ-Plots für das GARCH(1,1)- und für das Hidden Markov-Modell in nahezu allen Bereichen deutlich von der Winkelhalbierenden abweichen. Für das GARCH- und das Hidden Markov-Modell existieren Bereiche, in denen die Abweichung von der Winkelhalbierenden so extrem ist, dass die NPR des QQ-Plots in diesen Bereichen größer sind als zehn (in der Abbildung gleich zehn gesetzt). Für das Hidden Markov-Modell gehören zu diesem Bereich 826 Beobachtungen, die im mittleren Bereich des QQ-Plots liegen.

Die Ergebnisse in diesem Abschnitt ergänzen die Informationen aus dem vorletzten Abschnitt, in dem bei der Analyse der Histogramme der UPR bereits darauf hingewiesen wurde, dass die grundlegende Gestalt der Voraussagever-

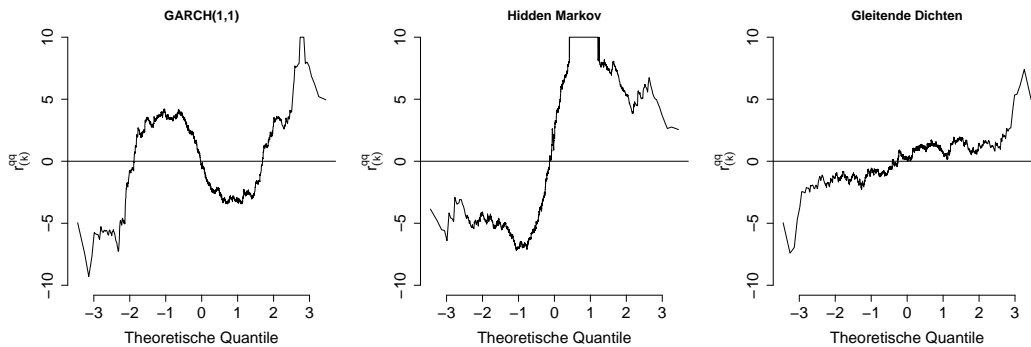


Abbildung 4.5: NPR der QQ-Plots.

teilungen des GARCH(1,1)- und des Hidden Markov-Modells nicht korrekt ist. Bemerkenswert ist, dass die Voraussagen des GARCH(1,1)-Modells zwar tendenziell zu breit sind aber dennoch zu kurze Enden besitzen. Dies kann auch an der wellenförmigen Struktur in Abbildung 4.5 abgelesen werden.

4.4.4 Erfassung zeitlicher Abhängigkeiten

Der Grundgedanke von Zeitreihenmodellen mit Verteilungsprognosen ist, als Voraussageverteilung für die Periode t die bedingte Verteilung von Y_t , bei gegebenen Informationen über die zuvor gemachten Beobachtungen, zu verwenden. Fließen sämtliche Informationen in die bedingte Verteilung ein, so sind die Voraussageverteilungen und letztlich auch die Pseudo-Residuen unabhängig. (Ein entsprechender Beweis ist bei PEDERSEN, 1994 dargestellt.) Eine weitere Möglichkeit, die NPR zu analysieren, ist daher die Betrachtung ihrer Autokorrelation. Auch eine Untersuchung der Autokorrelation der quadrierten Residuen sowie höherer Potenzen der Residuen gibt Aufschluss darüber, inwieweit die Informationen zurückliegender Beobachtungen in die Voraussageverteilung einfließen.

Abbildung 4.6 zeigt die Autokorrelationsfunktionen der NPR (links) und der quadrierten NPR (rechts). Für alle Modelle ist die Autokorrelation der NPR bei $lag = 1$ signifikant von null verschieden. Ferner ist die Autokorrelation der quadrierten Residuen für das Hidden Markov-Modell und das Modell mit gleitenden Dichten zu groß, d.h. die Werte der entsprechenden Autokorrelationsfunktionen sind auch für große $lags$ noch signifikant. Allgemein zeigt eine Autokorrelation der quadrierten NPR, dass nach der Beobachtung eines extremen NPR in der Periode t tendenziell ein extremes NPR in der Periode $t + 1$

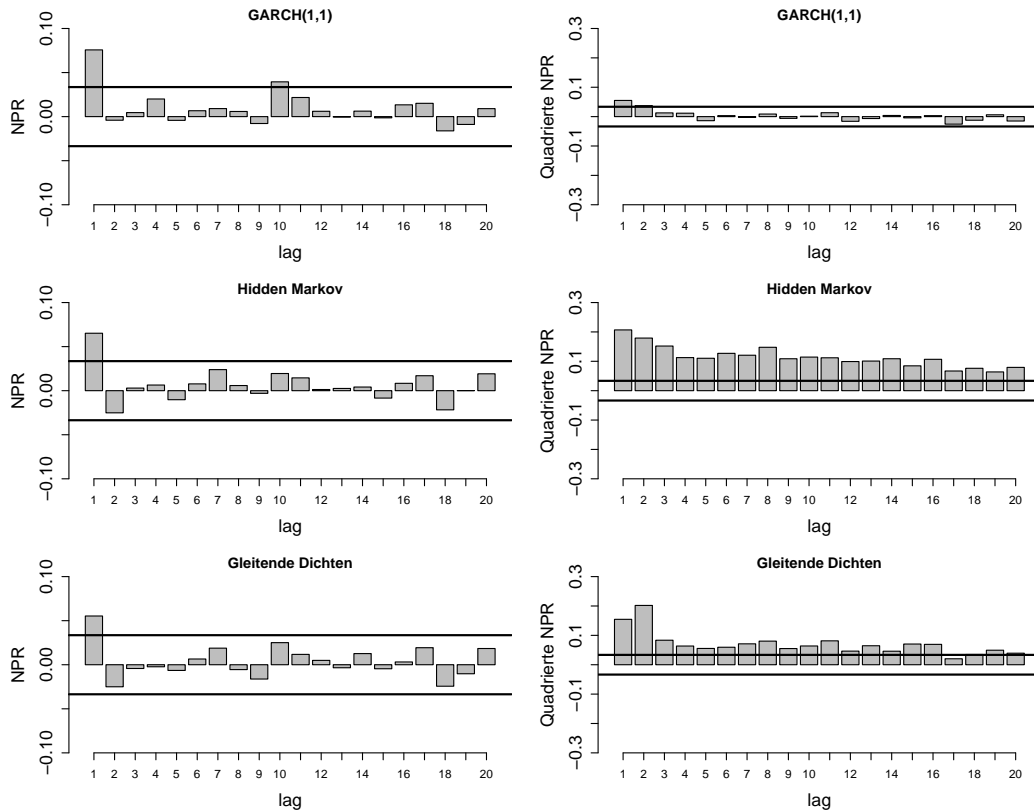


Abbildung 4.6: Autokorrelationsfunktionen der NPR (links) und der quadrierten NPR (rechts).

folgt (gleiches gilt für NPR nahe null). Mit Blick auf die Güte eines Modells bedeutet dies, dass Informationen über die Volatilität aus der Vorperiode nicht hinreichend in der Voraussageverteilung der Folgeperiode berücksichtigt sind: Die Modelle „... *fail to capture the volatility dynamics operative in the process*“ (DIEBOLD GUNTHER AND TAY, 1998). Die Modellierungsidee des GARCH-Modells basiert gerade auf derartigen Abhängigkeiten der Beobachtungen: Eine betragsmäßig relativ große Beobachtung führt zu einem relativ großen Wert des Schätzers für die Varianz der Voraussageverteilung in der Folgeperiode. Analog führt eine betragsmäßig kleine Beobachtung zu einem kleinen Wert des Schätzers. Daher ist die Autokorrelationsfunktion der quadrierten NPR des GARCH(1,1)-Modells relativ unauffällig.

4.4.5 Voraussage der Volatilität im Zeitverlauf

Die bislang dargestellten Methoden besitzen einen statischen Charakter, d.h. bei der Überprüfung der Voraussageverteilungen sind zeitliche Aspekte nicht berücksichtigt. In der Regel ändert sich jedoch das Verhalten der Zeitreihe im Beobachtungszeitraum, was dazu führt, dass auch die Güte der Voraussagen zeitlichen Veränderungen unterliegt. Unter der Annahme, dass zumindest der Verteilungstyp eines Modells korrekt ist und sich im Zeitverlauf nicht verändert, kann eine Veränderung der Parameter durch eine Aktualisierung der Parameterschätzer berücksichtigt werden. Erfolgt diese jedoch nicht, führen Veränderungen der Zeitreihe (auch bei korrektem Verteilungstyp) zu Veränderungen in der Struktur der Residuen im Zeitverlauf.

Wie bereits deutlich wurde, besitzt bei der Modellierung von Aktienrenditen die Voraussage der Varianz (bzw. Volatilität) zukünftiger Renditen eine besondere Bedeutung. Diese begründet sich auf dem gewachsenen Bedürfnis, die Risiken von Finanztiteln, die durch das Ausmaß ihrer Volatilität charakterisiert werden können, zu quantifizieren (vgl. NEUMANN, 2000, S. 1). Der im letzten Absatz verwendete Begriff ‚Struktur‘ wird daher im Folgenden durch die Varianz der Residuen spezifiziert: Es wird gezeigt, wie man mit Hilfe der kumulierten quadrierten NPR (nachfolgend durch CUSUMSQ abgekürzt; abgeleitet aus dem Englischen Begriff *Cumulative Sums of Squares*) untersuchen kann, ob die Varianz der Residuen und somit die Güte der vorausgesagten Volatilität zeitlichen Schwankungen unterliegt. Sei

$$s_t = \frac{\sum_{i=k}^t r_i^2}{\sum_{i=k}^K r_i^2} \quad \text{für } t = k, \dots, K$$

die beobachtete CUSUMSQ der NPR in der Periode t bei Betrachtung einer Teilzeitreihe $t = k, \dots, K$. Ferner seien alle NPR unabhängig und identisch standardnormalverteilt (das Modell also gültig im gesamten Betrachtungszeitraum), dann besitzt S_t für eine beliebige Teilzeitreihe eine Betaverteilung mit den Parametern (vgl. TANIZAKI, 1995):

$$\alpha = \frac{t - k + 1}{2} \quad \text{und} \quad \beta = \frac{K - t}{2}.$$

Die Analyse der CUSUMSQ ist bei der hier betrachteten Problemstellung von Interesse, weil sie gut geeignet ist, um Veränderungen der Varianz einer Zeitreihe zu identifizieren (vgl. TANIZAKI, 1995): Verändert sich die Varianz NPR im Zeitverlauf, so spiegelt sich dies im Verlauf der CUSUMSQ der NPR wider.

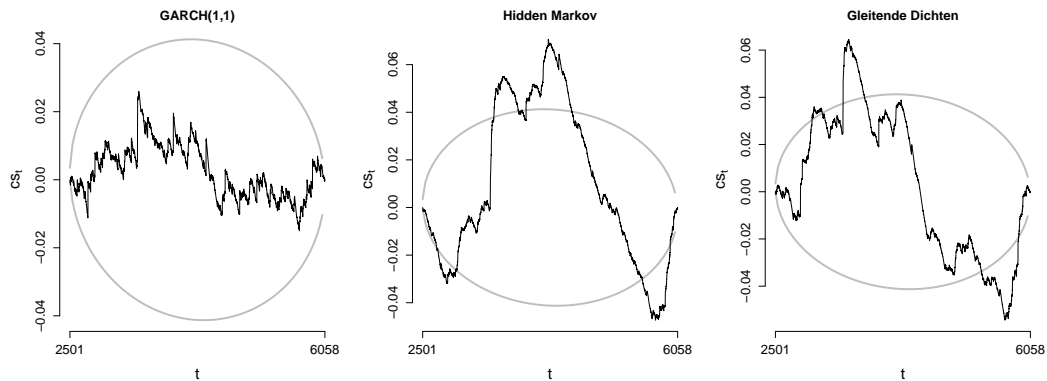


Abbildung 4.7: Zentrierte CUSUMSQ der NPR.

Die zentrierte und daher graphisch besser zu beurteilende Variante der CUSUMSQ

$$cs_t = s_t - \frac{t - k + 1}{K - k + 1}$$

ist für die beobachteten NPR der drei Modelle in Abbildung 4.7 zu sehen. Untersucht werden auch hier die Residuen der *out of sample*-Periode, d.h. es gilt $k = 2501$ und $K = 6058$. Die grau eingezeichneten Kurven markieren simultane 95%-Konfidenzbänder (siehe dazu TANIZAKI, 1995 sowie BOS, DING AND FETHERSTON, 1996). Sie grenzen den Bereich ein, der bei unabhängig standardnormalverteilten NPR (d.h. bei Gültigkeit des Modells) mit einer Wahrscheinlichkeit von 0.95 nicht verlassen wird. Es ist zu erkennen, dass die CUSUMSQ des Hidden Markov-Modells und die des Modells mit gleitenden Dichten zum Teil außerhalb des Konfidenzbereichs verlaufen. Dies zeigt, dass die Varianz der NPR für diese Modelle nicht konstant ist. Betrachtet man beispielsweise das Hidden Markov-Modell (mittlere Graphik), so besitzt der Verlauf zu Beginn des betrachteten Zeitraums einen negativen Trend. In diesem Bereich ist die Varianz der NPR tendenziell zu klein. Anschließend folgt ein Bereich, in dem die Varianz der NPR tendenziell zu groß ist (positiver Trend) usw.

In einem nächsten Schritt muss geklärt werden, welche Bedeutung die schwankende Varianz der NPR mit Blick auf die Güte eines Modells besitzt. TANIZAKI (1995) sowie BOS, DING AND FETHERSTON (1996) verwenden die hier beschriebene Methode, um Veränderungen der Varianz der ursprünglichen Zeitreihe zu identifizieren. Hier wird die Methode allerdings zur Untersuchung der Zeitreihe der NPR eingesetzt, was die Interpretation ihrer Ergebnisse verändert: Ist die Varianz der NPR in einem bestimmten Bereich tendenziell zu

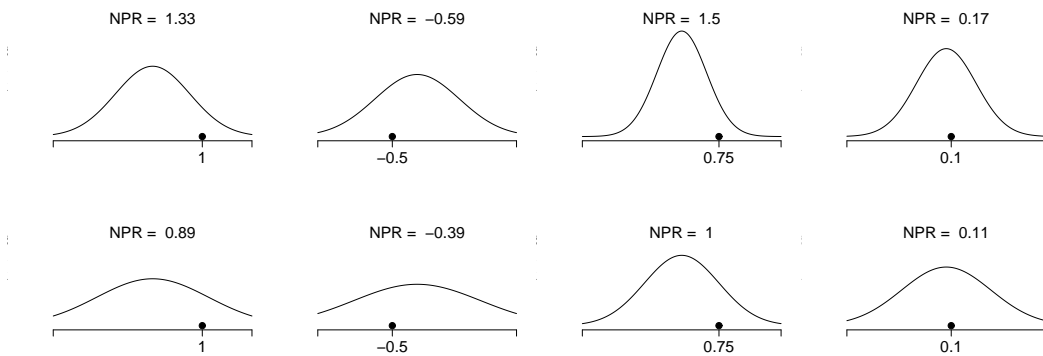


Abbildung 4.8: Folge betragsmäßig größerer NPR bei Prognose geringerer Volatilität (oben) und Folge betragsmäßig kleinerer NPR bei Prognose größerer Volatilität (unten).

klein, wird die Volatilität der Zeitreihe in diesem Bereich überschätzt und umgekehrt. Zur Verdeutlichung dieser Interpretation sei nachfolgend ein kleines Beispiel betrachtet.

Beispiel 4.1 In Abbildung 4.8 sind vier Voraussageverteilungen zweier unterschiedlicher Modelle dargestellt. Die Voraussageverteilungen des ersten Modells (oben) sind schmal im Vergleich zu denen des zweiten Modells (untere Voraussageverteilungen). Dies führt bei identischen Beobachtungen (also identischer Volatilität) zu betragsmäßig größeren Residuen für das obere Modell. Daher kann beispielsweise von einer Folge relativ großer quadrierter NPR (oberes Modell) nicht auf eine geringe Volatilität der Zeitreihe geschlossen werden. Eine solche Folge bedeutet vielmehr, dass die prognostizierte Volatilität relativ gering ist.



Verwendet man diese Interpretation zur Beurteilung der Voraussagen des Hidden Markov-Modells (Abbildung 4.7, mittlere Graphik), so zeigt die systematische Schwankung der CUSUMSQ zunächst, dass die Varianz der Residuen nicht konstant ist und es somit Perioden geben muss, in denen die Volatilität der Zeitreihe nicht korrekt vorausgesagt wird. Es ist allerdings nicht möglich, vom Verlauf der Kurve auf den konkreten Wert der Residualvarianz zu schließen. Da man aber wissen muss, ob die Residualvarianz größer bzw. kleiner ist als eins, um eine Unter- bzw. Überschätzung der Volatilität zu identifizieren, muss die Residualvarianz explizit für bestimmte Zeitabschnitte geschätzt werden, wenn man wissen möchte, in welchen Abschnitten die Volatilität korrekt

bzw. nicht korrekt vorausgesagt wird (siehe nächster Abschnitt). Abbildung 4.7 deutet auch für das Modell mit gleitenden Dichten auf Perioden hin, in denen die Varianz der Zeitreihe nicht korrekt geschätzt wird. Nur die CUSUMSQ der NPR des GARCH-Modells lässt nicht auf einen derartigen Fehler des Modells schließen.

4.4.6 Perioden unter- bzw. überschätzter Volatilität

Eine Erweiterung der im letzten Abschnitt beschriebenen Methode ist die Abgrenzung von Perioden unterschiedlicher Residualvarianz: Nachfolgend wird die graphische Analyse ergänzt, indem anhand der CUSUMSQ Zeitpunkte identifiziert werden, zu denen sich die Varianz der NPR ändert. Diese Zeitpunkte grenzen dann Perioden mit unterschiedlicher Varianz der NPR ab. Durch eine separate Schätzung der Varianz für die einzelnen Perioden kann letztlich beurteilt werden, in welchen Zeiträumen das Modell die Varianz der Zeitreihe tendenziell überschätzt bzw. unterschätzt oder korrekt prognostiziert.

4.4.6.1 Algorithmus zur Identifikation von Änderungszeitpunkten der Varianz

Zur Umsetzung der Idee wird ein Algorithmus benötigt, mit dem die Änderungszeitpunkte der Varianz einer Zeitreihe identifiziert werden können. Daher wird zunächst erläutert, wie ein solcher Algorithmus aussehen kann, bevor er auf die NPR der hier betrachteten Daten und Modelle angewendet wird. Der hier verwendete Algorithmus besteht aus zwei Stufen: Auf der ersten Stufe werden mögliche Änderungszeitpunkte der Residualvarianz bestimmt, auf der zweiten Stufe werden die möglichen Änderungszeitpunkte überprüft und korrigiert.

Erste Stufe

1. Start mit der *out of sample*-Periode: Setze $k = k_1$, wobei k_1 der Zeitpunkt ist, für den das erste Residual der *out of sample*-Periode berechnet wurde und setze $K = T$.

2. Bestimme den extremsten Wert der beobachteten CUSUMSQ:

$$s_t^* = \min_t \{ \min \{ P\{S_t \leq s_t\}, P\{S_t \geq s_t\} \} \} \quad \text{für } t = k, \dots, K.$$

Zur Lösung des Minimierungsproblems sind die im letzten Abschnitt genannten Betaverteilungen zu verwenden.

3. Liegt s_t^* außerhalb des simultanen $(1 - \alpha)$ -Konfidenzbands?
- (a) Nein: Es werden keine Perioden unterschiedlicher Varianz für die Teilzeitreihe $t = k, \dots, K$ identifiziert.
- (b) Ja:
- Speichere den Zeitpunkt t^* des Extremwerts s_t^* als möglichen Änderungszeitpunkt der Varianz.
 - Definiere durch eine Splittung des betrachteten Zeitraums an der Stelle t^* zwei neue Teilzeitreihen, wobei t^* zur ersten der beiden neuen Teilzeitreihen gehört.
4. Führe die Schritte 2 und 3 für alle in 3(b) definierten Teilzeitreihen durch, bis keine möglichen Änderungszeitpunkte mehr gefunden werden.

Beispiel 4.2 Abbildung 4.9 verdeutlicht die erste Stufe des Algorithmus. In der Wurzel des Baums wird die *out of sample*-Periode betrachtet und als möglicher Änderungszeitpunkt wird t_1^* identifiziert. Es ergeben sich die Teilzeitreihen $t = k_1, \dots, t_1^*$ und $t = t_1^* + 1, \dots, T$. In der zweiten Teilzeitreihe wird kein weiterer Änderungszeitpunkt festgestellt. In der ersten Teilzeitreihe hingegen wird t_2^* als möglicher Änderungszeitpunkt identifiziert, so dass erneut eine Splittung durchgeführt wird, die zu den Teilzeitreihen $t = k_1, \dots, t_2^*$ und $t = t_2^* + 1, \dots, t_1^*$ führt. Der letzte mögliche Änderungszeitpunkt t_3^* wird in der zweiten dieser Teilzeitreihen identifiziert. Als Ergebnis der ersten Stufe liegen somit drei mögliche Änderungszeitpunkte vor und die zugehörige Zerlegung der *out of sample*-Periode lässt sich anhand der grau markierten Blätter des Baums ablesen.



Zweite Stufe

1. Wurden $J \geq 2$ mögliche Änderungszeitpunkte identifiziert, so sortiere diese der Größe nach: $t_{(1)}^*, \dots, t_{(J)}^*$. Definiere außerdem $t_{(0)}^* = k_1 - 1$ und $t_{(J+1)}^* = T$.

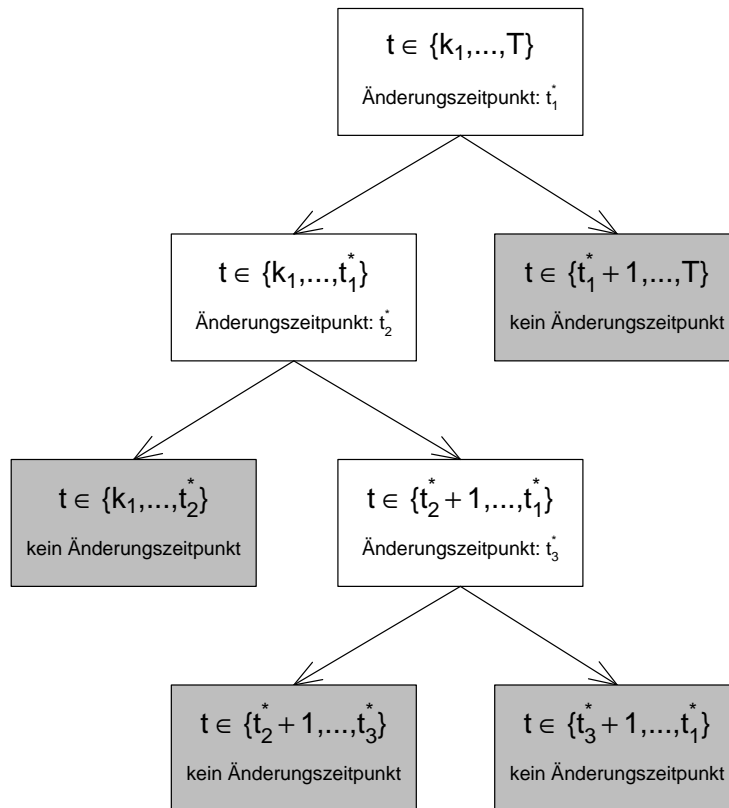


Abbildung 4.9: Erste Stufe des Algorithmus zur Identifikation von Änderungszeitpunkten der Varianz.

2. Berechne $s_{t_j}^*$ für alle $k = t_{(j-1)}^* + 1$ und $K = t_{(j+1)}^*$ mit $j = 1, \dots, J$ und überprüfe für alle $s_{t_j}^*$, ob sie außerhalb des simultanen $(1 - \alpha)$ -Konfidenzbandes liegen.
 - (a) Nein: Streiche die entsprechenden Änderungszeitpunkte $t_{(j)}^*$.
 - (b) Ja: Ersetze die entsprechenden Änderungszeitpunkte $t_{(j)}^*$ durch den zu $s_{t_j}^*$ gehörenden Änderungszeitpunkt $t_{(j)}^{**}$.
3. Wiederhole Schritt 2, bis sich die Anzahl der Änderungszeitpunkte nicht mehr ändert und für alle ersetzten Änderungszeitpunkte $t_{(j)}^* = t_{(j)}^{**}$ gilt.

Beispiel 4.3 Zur Verdeutlichung der zweiten Stufe wird das vorangegangene Beispiel erneut aufgegriffen. Als Ergebnis der ersten Stufe stehen die möglichen

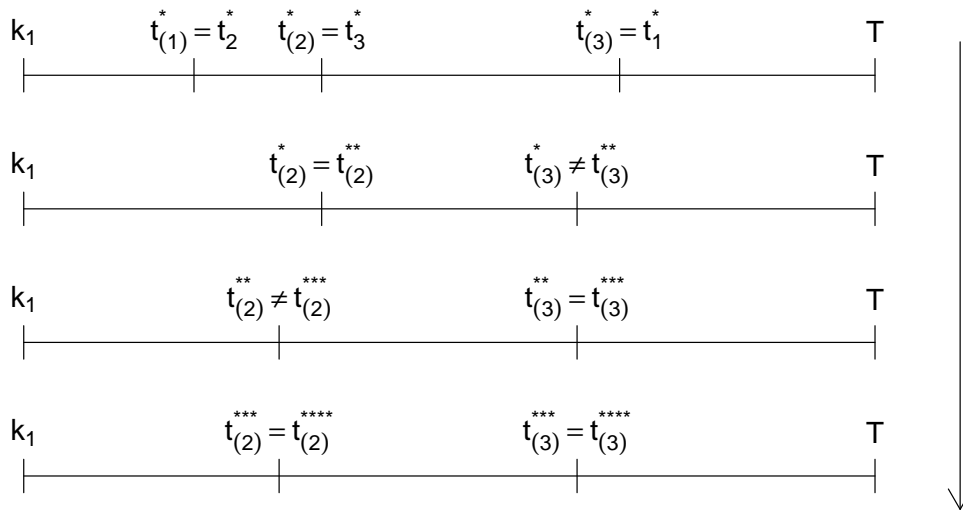


Abbildung 4.10: Zweite Stufe des Algorithmus zur Identifikation von Änderungszeitpunkten der Varianz.

Änderungszeitpunkte t_1^* , t_2^* und t_3^* zur Verfügung. Ordnet man diese der Größe nach, ergibt sich $t_{(1)}^* = t_2^*$, $t_{(2)}^* = t_3^*$ und $t_{(3)}^* = t_1^*$. Graphisch ist dies in Abbildung 4.10 dargestellt. Die obere Zeitachse zeigt die *out of sample*-Periode und eine beispielhafte Anordnung der möglichen Änderungszeitpunkte, die im Folgenden auf der zweiten Stufe zu überprüfen sind. Zunächst wird untersucht, ob die CUSUMSQ im Abschnitt $t = k_1, \dots, t_{(2)}^*$ das Konfidenzband verlässt. Im Beispiel sei dies nicht der Fall. Daher ist $t_{(1)}^*$ nicht mehr auf der darunter liegenden Zeitachse markiert. Betrachtet man den nächsten möglichen Änderungszeitpunkt, muss $t_{(2)}^* = t_{(2)}^{**}$ gelten, weil er auf der ersten Stufe des Algorithmus im identischen Zeitabschnitt $t = t_{(1)}^* + 1, \dots, t_{(3)}^*$ identifiziert wurde. Ferner sei hier angenommen, dass das Konfidenzband im Bereich $t = t_{(2)}^* + 1, \dots, T$ verlassen wird, sich aber $t_{(3)}^*$ verschiebt und durch $t_{(3)}^{**}$ ersetzt wird. Die erste Iteration ist an dieser Stelle beendet. Es folgt eine weitere, da sich sowohl die Anzahl der Änderungszeitpunkte als auch deren Lage verändert hat. Im Rahmen der nächsten Iteration verschiebt sich $t_{(2)}^{**}$ und wird durch $t_{(2)}^{***}$ ersetzt. Da $t_{(3)}^{**}$ bei der vorherigen Iteration im Zeitraum $t = t_{(2)}^{**} + 1, \dots, T$ identifiziert wurde, bleibt dieser Änderungszeitpunkt unverändert. Analog gilt dies für $t_{(2)}^{***}$ in der letzten Iteration, die durchgeführt werden muss, da sich die Lage der Änderungszeitpunkte zuvor noch verändert hat. $t_{(3)}^{***}$ könnte sich noch einmal

verändern, soll dies aber in diesem Beispiel nicht tun, so dass als endgültige Änderungszeitpunkte der Residualvarianz $t_{(2)}^{***}$ und $t_{(3)}^{***}$ vorliegen.



Der hier verwendete Algorithmus ist eine Kombination von zwei in der Literatur vorgeschlagenen Algorithmen. Dabei handelt es sich zum einen um den ICSS (*iterated cumulative sum of squares*)-Algorithmus von INCLAN AND TIAO (1994) und zum anderen um einen bei BOS, DING AND FETHERSTON (1998) beschriebenen Algorithmus (im Folgenden als BDF-Algorithmus bezeichnet). Ersterer basiert auf einer Analyse der Werte

$$s_t^{\text{IT}} = cs_t \sqrt{\frac{K - k + 1}{2}} \quad \text{für } t = k, \dots, K,$$

die sich bei unabhängig standardnormalverteilten Werten der betrachteten Zeitreihe wie eine *Brownian Bridge* verhalten. Die Struktur des ICSS-Algorithmus und des hier verwendeten Algorithmus ist identisch, d.h. in beiden Fällen erfolgt die Identifikation der Änderungszeitpunkte in zwei Stufen. Allerdings werden beim ICSS-Algorithmus lineare Konfidenzbänder eingesetzt, um auf Varianzänderungen zu testen. Dieses Vorgehen besitzt den Nachteil, dass die *Power* des Tests für Änderungszeitpunkte im mittleren Bereich der betrachteten Zeitreihe größer ist als für Änderungszeitpunkte an den Randbereichen. Der BDF-Algorithmus hingegen verwendet die hier beschriebenen Konfidenzbänder. Für diese ist die Wahrscheinlichkeit, dass die CUSUMSQ die Bänder verlässt, für alle Zeitpunkte identisch. Allerdings verzichtet der BDF-Algorithmus auf die zweite Stufe, d.h. auf eine Überprüfung der möglichen Änderungszeitpunkte. Diese erscheint aber nach theoretischen Überlegungen sinnvoll: Betrachtet man beispielsweise eine Zeitreihe für $t = 1, \dots, 1000$ und identifiziert auf der ersten Stufe $t_1^* = 250$ und $t_2^* = 500$ als mögliche Änderungszeitpunkte, so stellt t_1^* nicht den Änderungszeitpunkt der Varianz bezüglich der Periode $t = 1, \dots, 500$ dar, weil er in der gesamten Zeitreihe identifiziert wurde. Eine Überprüfung des Punktes für den korrespondierenden Zeitraum ist daher angebracht.

4.4.6.2 Simulationsstudie zum Vergleich der Algorithmen

Um das Vorgehen mit und ohne Durchführung der zweiten Stufe (also den hier verwendeten Algorithmus und den BDF-Algorithmus) zu vergleichen, wurde ein Simulationsexperiment durchgeführt. Als Aufbau wurde ein $6 \times 6 \times 6$ faktorielles Design gewählt, in dem die Länge einer Zeitreihe (T), die Anzahl der

Länge der Zeitreihe: T	Anzahl der Änderungszeitpunkte: m	Faktor der Varianzänderung: c
100	2	1
500	3	2
1000	6	4
2000	7	8
4000	10	16
6000	11	32

Tabelle 4.3: Faktoren und Faktorstufen des Simulationsexperiments.

Änderungszeitpunkte (m) und die Stärke der Varianzänderungen (c) als Faktoren berücksichtigt wurden: Eine einzelne erzeugte Zeitreihe setzt sich aus T unabhängig $N(0, \sigma_t^2)$ -verteilten Zufallszahlen y_t zusammen, wobei die Varianz zu m zufällig in der Zeitreihe verteilten Zeitpunkten mit dem Änderungsfaktor c multipliziert oder durch diesen dividiert wurde (jeweils mit Wahrscheinlichkeit 0.5). Für jede Kombination der in Tabelle 4.3 gegebenen Faktorstufen wurden 40 Zeitreihen erzeugt, d.h. insgesamt wurden 8 640 Zeitreihen erzeugt.

Die Beurteilung der Algorithmen basiert auf einem Vergleich der wahren Änderungszeitpunkte (die aus der Simulation bekannt sind) mit den Schätzern der Änderungszeitpunkte (bestimmt mit den Algorithmen). Es werden hier zwei Kriterien betrachtet, die von den Algorithmen erfüllt werden sollten: Zum einen wird untersucht, ob die Anzahl der Änderungszeitpunkte korrekt geschätzt wird und zum anderen, wie gut der Varianzverlauf der erzeugten Zeitreihen mit Hilfe der Algorithmen reproduziert werden kann.

Betrachtet sei zunächst die Reproduktion der Varianz. Die zur Erzeugung der Zeitreihen verwendete Varianz σ_t^2 ist eine Funktion der Zeit, die an den Änderungszeitpunkten der Varianz Sprungstellen besitzt und zwischen diesen Sprungstellen jeweils horizontal verläuft. Diese Treppenfunktion kann mit Hilfe der Algorithmen geschätzt werden, indem nach der Schätzung der Änderungszeitpunkte die Varianz der Zeitreihe separat für die entsprechenden Teilzeitreihen geschätzt wird. Ein Vergleich zwischen der wahren Varianzfunktion und der durch einen Algorithmus geschätzten Varianzfunktion zeigt dann, wie gut der Varianzverlauf der Population reproduziert wird. Für diesen Vergleich wird hier der Logarithmus der durchschnittlichen prozentualen Abweichung

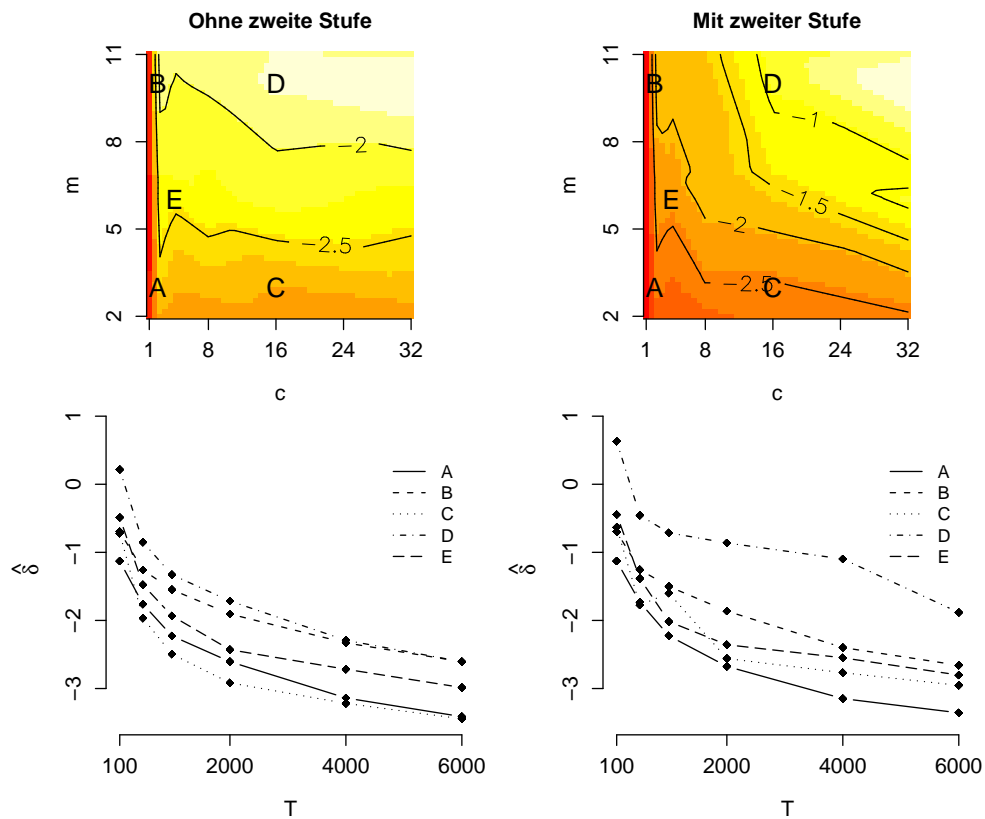


Abbildung 4.11: Geschätzte Abweichung bei der Reproduktion des Varianzverlaufs ohne Anwendung der zweiten Stufe (links) und mit Anwendung der zweiten Stufe (rechts).

zwischen geschätzter und tatsächlicher Varianz

$$\delta = \log \frac{1}{T} \sum_{t=1}^T \frac{|s_t^2 - \sigma_t^2|}{\sigma_t^2} \tag{4.2}$$

verwendet, wobei

$$s_t^2 = \frac{1}{K - k} \sum_{i=k}^K y_i^2 \quad \text{für } t = k, \dots, K$$

den Schätzer der Varianz in der Teilzeitreihe $t = k, \dots, K$ darstellt, der für alle identifizierten Perioden mit unterschiedlicher Varianz separat bestimmt wird.

Die Berücksichtigung der Faktoren erfolgt durch die Anpassung eines gesättigt-

tes Modells der Varianzanalyse:

$$\begin{aligned} \delta_{ijkl} &= \mu + T_i + m_j + c_k + (Tm)_{ij} + (Tc)_{ik} + (mc)_{jk} + (Tmc)_{ijk} + e_{ijkl} \\ &\text{mit } i, j, k = 1, \dots, 6, l = 1, \dots, 40 \\ &\text{und } e_{ijkl} \stackrel{\text{iid}}{\sim} N(0, \sigma^2). \end{aligned}$$

Durch die Anpassung des Modells unter den üblichen Nebenbedingungen erhält man die Schätzer $\hat{\delta}_{ijk}$, die als Schätzer der erwarteten Abweichung (definiert nach Gleichung 4.2) bei Zeitreihen der Länge T_i mit m_j Änderungen der Varianz um den Faktor c_k zu verstehen sind. Abbildung 4.11 enthält eine geglättete Darstellung dieser Schätzer (zur Glättung wurde die **R**-library `akima` verwendet; siehe AKIMA, 1978 und 1996). Die beiden Imageplots zeigen die Ergebnisse bei unterschiedlicher Anzahl der Änderungszeitpunkte und unterschiedlichen Änderungsfaktoren sowie konstanter Länge der Zeitreihen (hier $T_4 = 2000$). In den helleren Bereichen sind die Abweichungen zwischen der tatsächlichen und der geschätzten Varianzfunktion relativ groß, in den dunkleren Bereichen relativ gering. Es ist zu sehen, dass die Abweichungen für beide Algorithmen größer werden, wenn die Anzahl der Änderungszeitpunkte zunimmt. Wird auf die zweite Stufe verzichtet (linker Plot, BDF-Algorithmus), bleibt die Abweichung bei steigendem Änderungsfaktor in etwa konstant, während die Abweichungen größer werden, wenn die Änderungszeitpunkte auf der zweiten Stufe überprüft werden (rechter Plot, hier verwendeter Algorithmus). Anhand der eingezeichneten Höhenlinien ist zu erkennen, dass insbesondere für größere Änderungsfaktoren der Verzicht auf die zweite Stufe vorteilhaft ist.

Die dunklen Bereiche am linken Rand der beiden Imageplots zeigen, dass die Unterschiede zwischen geschätzter und tatsächlicher Varianz für kleine Varianzänderungen relativ gering ist. In diesem Bereich ist die tatsächliche Varianz der Zeitreihen nahezu konstant, so dass in der Regel keine Änderungszeitpunkte identifiziert werden und die Abweichungen in erster Linie auf die Streuung des Varianzschätzers und nicht auf eine fehlerhafte Schätzung der Änderungszeitpunkte zurückzuführen ist.

In den beiden unteren Graphiken ist das Verhalten von δ in Abhängigkeit von der Länge der Zeitreihen zu sehen. Jede der Kurven gehört zu einem Punkt aus den darüber liegenden Imageplots, d.h. die Anzahl der Änderungszeitpunkte und der Änderungsfaktor der Varianz sind jeweils konstant. Da der Standardfehler des Varianzschätzers bei zunehmender Länge der Zeitreihe kleiner wird, werden die Abweichungen bei der Reproduktion des Varianzverlaufs ebenfalls kleiner. Die jeweils zum Punkt D gehörenden Kurven unterscheiden sich am deutlichsten. Dies bestätigt, dass der Verzicht auf die zweite Stufe bei relativ

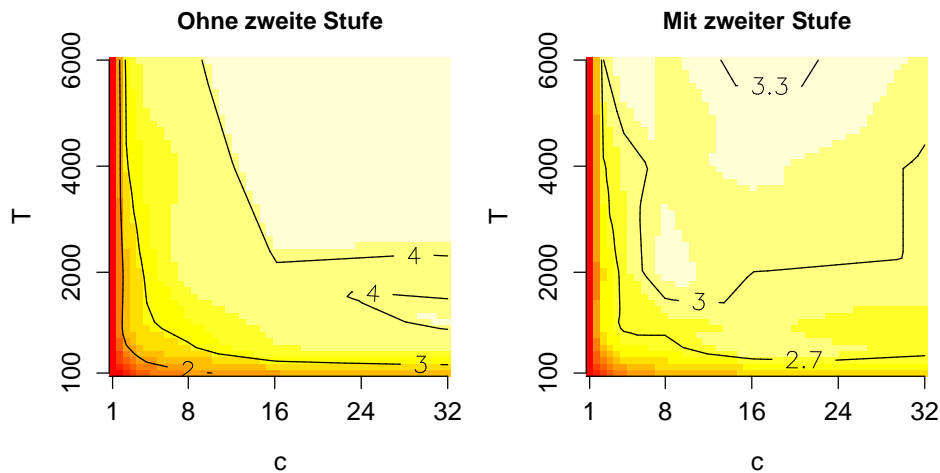


Abbildung 4.12: Geschätzte Anzahl von Änderungszeitpunkten für Zeitreihen mit drei tatsächlichen Änderungszeitpunkten ohne Anwendung der zweiten Stufe (links) und mit Anwendung der zweiten Stufe (rechts).

großem Änderungsfaktor mit besseren Ergebnissen bei der Reproduktion des Varianzverlaufs verbunden ist.

Wie bereits zuvor erwähnt, ist der Schätzer für die Anzahl der Änderungszeitpunkte (im Folgenden mit \hat{m} bezeichnet) das zweite Kriterium, nach dem die Algorithmen beurteilt werden. Einige Pilotversuche zeigten, dass der Erwartungswert des Schätzers (neben der tatsächlichen Anzahl der Änderungszeitpunkte) von der Länge der Zeitreihe und von dem Änderungsfaktor der Varianz abhängt. Diese Faktoren werden daher in einem verallgemeinerten linearen Modell (mit den üblichen Nebenbedingungen)

$$\begin{aligned} \log(\lambda_{ikl}) &= \mu + T_i + c_k + (Tc)_{ik} \\ &\text{mit } i, k = 1, \dots, 6, l = 1, \dots, 40 \\ &\text{und } \hat{m} \sim Po(\lambda_{ik}) \end{aligned}$$

als erklärende Variablen verwendet, um den Erwartungswert von \hat{m} zu schätzen. Die Modellanpassung liefert die Schätzer $\hat{\lambda}_{ij}$, die als Schätzer des Erwartungswerts von \hat{m} für Zeitreihen der Länge T_i mit Änderungsfaktor c_k zu verstehen sind. Hält man die Anzahl der tatsächlichen Änderungszeitpunkte konstant, können die angepassten Werte des verallgemeinerten linearen Modells verwendet werden, um zu überprüfen, ob die Anzahl der Änderungszeitpunkte korrekt von den Algorithmen geschätzt wird: Entsprechen die $\hat{\lambda}_{ij}$ der tatsächlichen An-

zahl der Änderungszeitpunkte, ist die Schätzung erwartungstreu.

Abbildung 4.12 zeigt eine Glättung der $\hat{\lambda}_{ij}$ für Zeitreihen mit drei tatsächlichen Änderungszeitpunkten. Der dunkle Bereich am jeweils linken Rand der Imageplots zeigt, dass die Anzahl der Änderungszeitpunkte bei geringem Änderungsfaktor von beiden Algorithmen unterschätzt wird. Hierbei handelt es sich um einen Spezialfall eines allgemeinen Problems beim Testen von Hypothesen: Bei geringem Änderungsfaktor unterscheidet sich die Alternativhypothese nur geringfügig von der Nullhypothese, so dass die *Power* der zugrunde liegenden Tests klein ist.

Die eingezeichneten Höhenlinien zeigen, dass die Anzahl der Änderungszeitpunkte vom BDF-Algorithmus tendenziell überschätzt wird, wenn relativ lange Zeitreihen mit relativ großem Änderungsfaktor betrachtet werden. Werden die Änderungszeitpunkte hingegen im Rahmen der zweiten Stufe überprüft, führt dies zu einer in etwa erwartungstreuen Schätzung. Das Problem der wiederholten Signifikanztests, das in beiden Algorithmen existiert, wird durch die zweite Stufe also entschärft. Sehr ähnliche Ergebnisse (hier allerdings nicht explizit dargestellt) ergaben sich, wenn die Anzahl der tatsächlichen Änderungszeitpunkte einem der übrigen in Tabelle 4.3 gegebenen Werten entsprach.

4.4.6.3 Anwendung des Algorithmus auf die Zeitreihe der NPR

In Abschnitt 4.4.5 wurde gezeigt, dass die CUSUMSQ der NPR für das GARCH(1,1)-Modell innerhalb der simultanen Konfidenzbänder verläuft. Es ist klar, dass daher bei Verwendung der hier betrachteten Algorithmen keine Perioden unterschiedlicher Varianz für das GARCH(1,1)-Modell identifiziert werden. Für das Hidden Markov-Modell und das Modell mit gleitenden Dichten liegen die CUSUMSQ hingegen teilweise außerhalb des Konfidenzbands, so dass eine Abgrenzung von Perioden unterschiedlicher Residualvarianz möglich ist. Im Folgenden wird das entsprechende Vorgehen zunächst anhand des Hidden Markov-Modells verdeutlicht, bevor anschließend eine Interpretation der Ergebnisse für beide Modelle vorgenommen wird.

Die erste Stufe des Algorithmus ist in Abbildung 4.13 veranschaulicht. Die Graphik oben links zeigt erneut die CUSUMSQ der NPR der *out of sample*-Periode. Man erhält hier $t_1^* = 5865$, d.h. den ersten möglichen Änderungszeitpunkt der Residualvarianz. Die beiden Teilzeitreihen, die sich aus der entsprechenden Splittung ergeben, sind in den beiden folgenden Graphiken (rechts neben der ersten Graphik) betrachtet. In der ersten Teilzeitreihe wird der zweite mögliche Änderungszeitpunkt $t_2^* = 4254$ identifiziert, für die zweite Teilzeitrei-

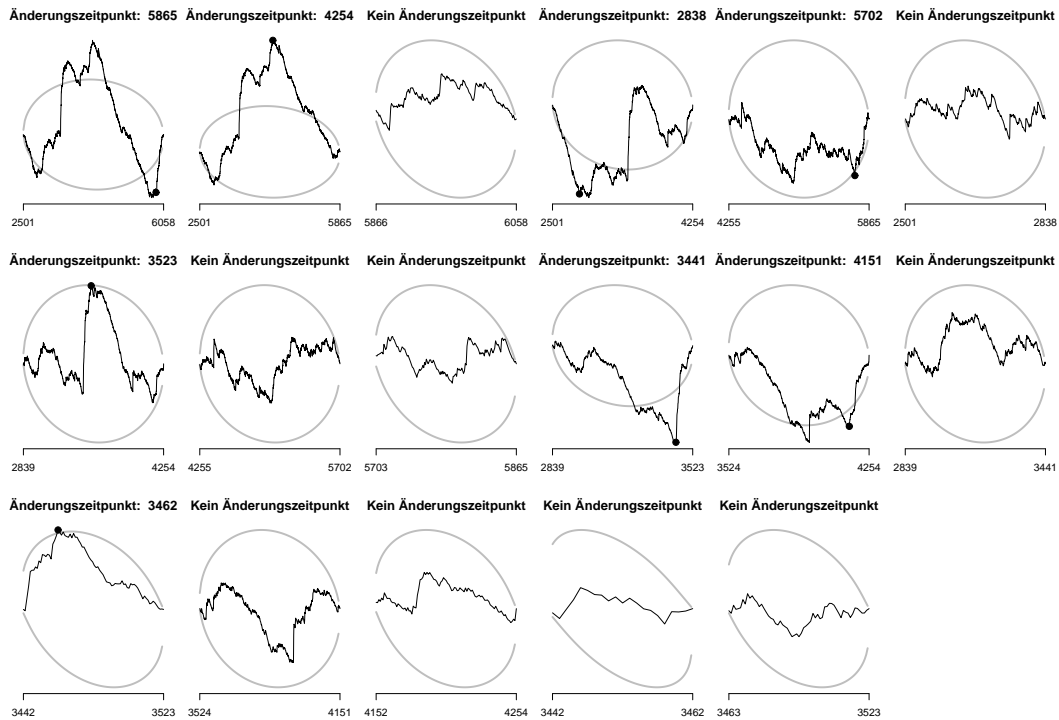


Abbildung 4.13: Mögliche Änderungszeitpunkte der Residualvarianz für das Hidden Markov-Modell.

he verläuft die CUSUMSQ der NPR innerhalb der Konfidenzbänder. Alle weiteren Graphiken sind analog zu interpretieren, d.h. sie zeigen Teilzeitreihen, die sich aus zuvor identifizierten möglichen Änderungszeitpunkten ergeben. Jeweils ist im Titel der Graphik angegeben, ob für den entsprechenden Zeitraum eine Varianzänderung identifiziert wird. Eine Zusammenfassung der acht (zeitlich sortierten) Änderungszeitpunkte, die auf der ersten Stufe identifiziert wurden, ist in Tabelle 4.4 gegeben.

Die Überprüfung der möglichen Änderungszeitpunkte (zweite Stufe des Algorithmus) ist in Abbildung 4.14 veranschaulicht. Es ist zu sehen, dass bei der ersten Iteration (dargestellt in den acht oberen Graphiken) die ersten sieben Änderungszeitpunkte unverändert bleiben. Lediglich der achte Änderungszeitpunkt ändert seine Position, so dass $t_{(8)}^* = 5865$ durch $t_{(8)}^{**} = 5884$ ersetzt wird. Bei der zweiten Iteration ergeben sich keine weiteren Veränderungen und die endgültigen Änderungszeitpunkte der Residualvarianz sind gefunden.

Im Anschluss an die Bestimmung der Änderungszeitpunkte der Residualvarianz ist die Varianz der Residuen in den resultierenden Teilzeitreihen separat

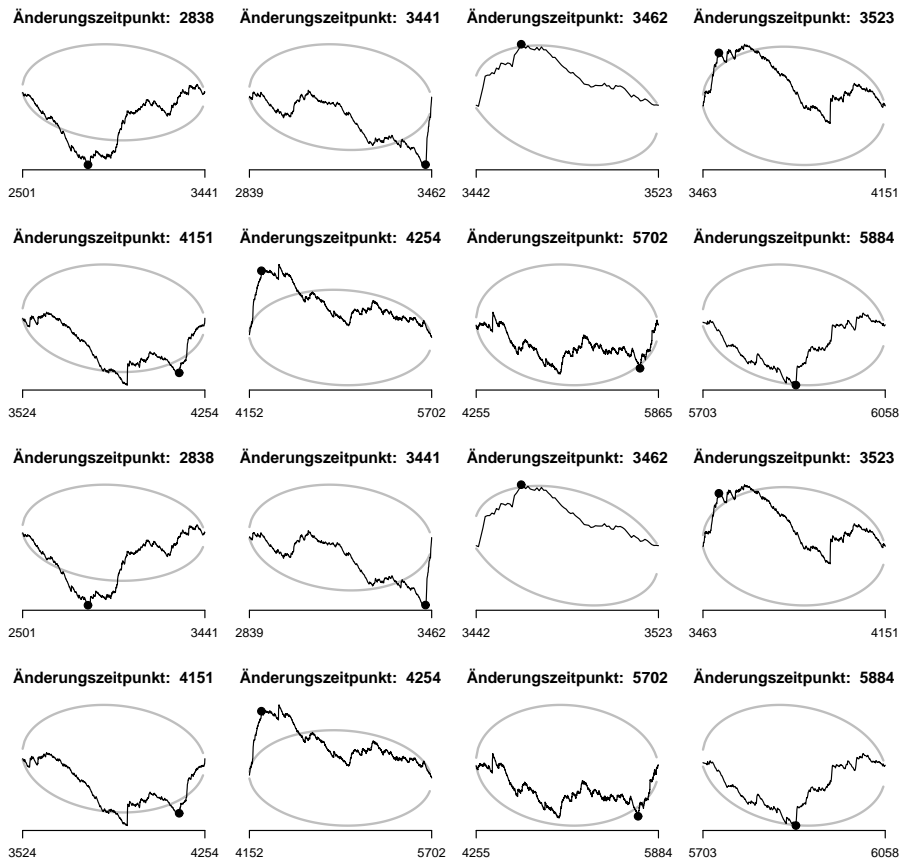


Abbildung 4.14: Endgültige Änderungszeitpunkte der Residualvarianz für das Hidden Markov-Modell.

zu schätzen, wobei als Schätzer die Varianz der in den jeweiligen Teilzeitreihen beobachteten NPR

$$\widehat{Var}(R_t) = \frac{1}{K-k} \sum_{t=k}^K \left(r_t - \frac{1}{K-k+1} \sum_k^K r_t \right)^2$$

verwendet werden kann. Bei korrekter Vorhersage der Volatilität besitzt die Varianz der Residuen den Wert eins. Größere Werte der geschätzten Residualvarianz deuten auf eine Unterschätzung der Volatilität in dem korrespondierenden Zeitabschnitt hin, kleinere Werte auf eine Überschätzung (siehe auch Abbildung 4.8). Die entsprechenden Schätzergebnisse für die Residuen des Hidden Markov-Modells und für das Modell mit gleitenden Dichten zeigt Tabelle 4.5.

Graphisch ist die geschätzte Varianz der NPR zusammen mit der Rendite-

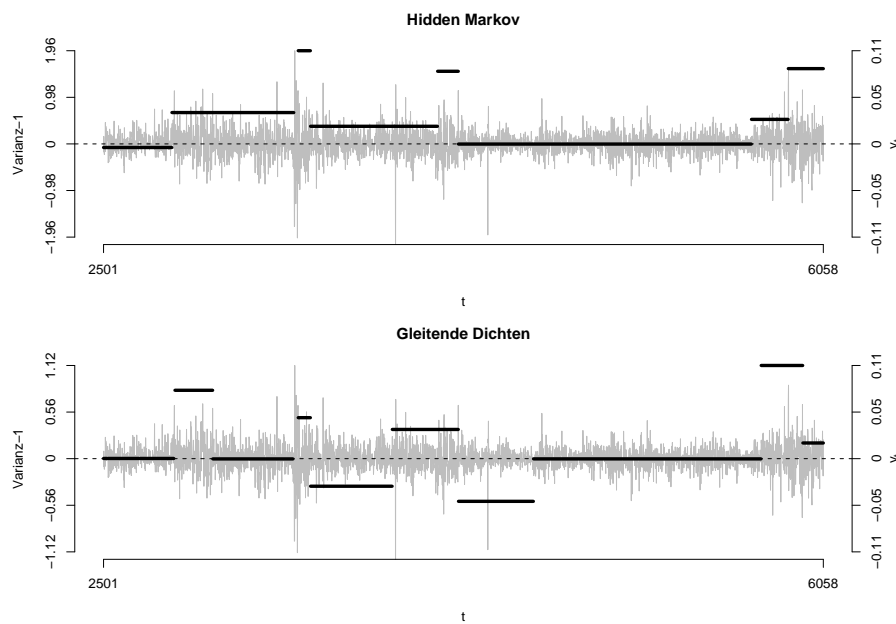


Abbildung 4.15: Zeitreihe und geschätzte Residualvarianz.

zeitreihe in Abbildung 4.15 dargestellt, wobei auf der linken Achse die Differenz $\widehat{Var}(R_t) - 1$ abgetragen ist, um eine zentrierte Darstellung zu erhalten. Verlaufen die horizontalen Linien unterhalb bzw. oberhalb von null, wird die Volatilität in der entsprechenden Periode tendenziell überschätzt bzw. unterschätzt. Die obere Graphik zeigt, dass die Volatilität der Zeitreihe vom Hidden Markov-Modell im größten Teil der *out of sample*-Periode unterschätzt wird. Die Ursache hierfür ist vermutlich, dass die Modellierung der Varianz nicht explizit in dem Modell berücksichtigt ist und auf eine Aktualisierung der Parameter verzichtet wird. Dies führt einerseits zu Zeitabschnitten unterschiedlicher Residualvarianz und andererseits aufgrund der relativ geringen Volatilität in der *in sample*-Periode (siehe Abbildung 4.1) zur tendenziellen Unterschätzung der Volatilität in der *out of sample*-Periode. Ersteres gilt analog für das Modell mit gleitenden Dichten (untere Graphik). Beim GARCH-Modell hingegen ist

$t_{(1)}^*$	$t_{(2)}^*$	$t_{(3)}^*$	$t_{(4)}^*$	$t_{(5)}^*$	$t_{(6)}^*$	$t_{(7)}^*$	$t_{(8)}^*$
2838	3441	3462	3523	4151	4254	5702	5865

Tabelle 4.4: Mögliche Änderungszeitpunkte der Residualvarianz für das Hidden Markov-Modell.

Hidden Markov-Modell									
Periodenbeginn	2501	2839	3442	3463	3524	4152	4255	5703	5885
Periodenende	2838	3441	3462	3523	4151	4254	5702	5884	6058
$\widehat{Var}(R_t)$	0.92	1.66	7.98	2.96	1.37	2.53	1.00	1.52	2.59
Modell mit gleitenden Dichten									
Periodenbeginn	2501	2844	2853	3040	3441	3446	3523	3848	4254
Periodenende	2843	2852	3039	3440	3445	3522	3847	4253	4624
$\widehat{Var}(R_t)$	1.01	8.83	1.85	1.00	8.06	1.49	0.67	1.35	0.49
Periodenbeginn	4625	5750	5956						
Periodenende	5749	5955	6058						
$\widehat{Var}(R_t)$	1.00	2.12	1.19						

Tabelle 4.5: Geschätzte Varianz der NPR für die identifizierten Zeitabschnitte.

die Modellierung einer dynamischen Volatilität der zentrale Gedanke, was dazu führt, dass keine Perioden unterschiedlicher Residualvarianz identifiziert werden und die Varianz der Residuen in der *out of sample*-Periode mit einem Wert von 1.07 nur geringfügig zu groß ist.

4.5 Kalibrierung der Voraussageverteilungen

4.5.1 Mängel der Modellanpassung und Grundgedanke der Kalibrierung

Im letzten Abschnitt wurde gezeigt, dass mit keinem der hier betrachteten Modelle die korrekte Verteilung zukünftiger Werte prognostiziert werden kann. Im Einzelnen ist

- (1) die Gestalt der Voraussageverteilungen (grundlegende Gestalt, Voraussage des VaR und Abbildung von Ausreißern) des GARCH-Modells und des Hidden Markov-Modells nicht zufriedenstellend,
- (2) die Erfassung zeitlicher Abhängigkeiten durch das Hidden Markov-Modell und das Modell mit gleitenden Dichten unzureichend und
- (3) die Güte der prognostizierten Volatilität bei dem Hidden Markov-Modell und dem Modell mit gleitenden Dichten durch zeitliche Schwankungen gekennzeichnet.

Da keines der Modelle korrekte Voraussagen liefert, ist es ein naheliegender Gedanke, die Modelle zu verwerfen und alternative Modelle zu untersuchen, bis ein Modell gefunden ist, dessen Anpassung keine Mängel aufweist. Es gibt jedoch eine weitere Möglichkeit, die in diesem Abschnitt erläutert und auf das betrachtete Beispiel angewendet wird. Dabei ist es die Idee, die fehlerhaften Voraussagen als erste Approximation zu verwenden und diese zu kalibrieren, wobei zur Kalibrierung die Eigenschaften der Pseudo-Residuen, die für vergangene Perioden berechnet wurden, genutzt werden.

Im Folgenden wird zunächst das allgemeine Vorgehen zur Kalibrierung der Voraussageverteilungen erläutert. Dabei wird explizit auf die drei oben genannten Mängel der Modelle eingegangen, an denen sich die Kalibrierung orientiert. Anschließend werden die Voraussageverteilungen der hier betrachteten Modelle kalibriert und die Modellanpassung nach Kalibrierung mit der Modellanpassung der ursprünglichen Modelle verglichen.

4.5.2 Vorgehensweise zur Kalibrierung der Voraussagen

Zunächst sei es das Ziel, die Gestalt der Voraussageverteilungen zu kalibrieren. Dazu kann eine Methode verwendet werden, die 1999 von DIEBOLD, HAHN AND TAY vorgestellt wurde. Sei Q die Verteilungsfunktion der UPR. Ferner seien die UPR unabhängig und identisch verteilt und $G_{Y_t|\Omega_{t-1}}$ sei die korrekte (wahre) Verteilungsfunktion der Voraussageverteilung für die Periode t , dann kann letztere wie folgt dargestellt werden:

$$\begin{aligned} G_{Y_t|\Omega_{t-1}}(y) &= P\{Y_t \leq y \mid \Omega_{t-1}\} \\ &= P\{F_{Y_t|\Omega_{t-1}}(Y_t) \leq F_{Y_t|\Omega_{t-1}}(y)\} \\ &= Q(F_{Y_t|\Omega_{t-1}}(y)). \end{aligned}$$

Man kann also die (zu schätzende) Verteilungsfunktion der UPR verwenden, um die Verteilungsfunktion der Voraussage unter dem Modell zu transformieren (bzw. zu kalibrieren) und erhält als Ergebnis die korrekte Verteilungsfunktion der Voraussageverteilung. Dieses Resultat wurde bereits an zwei Stellen dieser Arbeit betrachtet. In Kapitel 2 wurde es benutzt, um die Verteilung der UPR herzuleiten. Ferner wurde es in Abschnitt 3.4 verwendet, um die Schätzung der Parameter zu berücksichtigen. Während dort allerdings die Residuen (rechte Seite der Gleichung) im Mittelpunkt standen, werden hier die Verteilungsfunktionen des wahren Modells (linke Seite der Gleichung) fokussiert. Alternativ kann auch die Verteilungsfunktion der NPR (hier mit K be-

zeichnet) verwendet werden, um die Voraussageverteilungen zu kalibrieren:

$$\begin{aligned} G_{Y_t|\Omega_{t-1}}(y) &= P \{ \Phi^{-1}(F_{Y_t|\Omega_{t-1}}(Y_t)) \leq \Phi^{-1}(F_{Y_t|\Omega_{t-1}}(y)) \} \\ &= K(\Phi^{-1}(F_{Y_t|\Omega_{t-1}}(y))). \end{aligned}$$

Es gibt jedoch zwei Probleme, die bei der Anwendung der Methode zu berücksichtigen sind. Zum einen müssen die Pseudo-Residuen unabhängig sein. Wie in den Beispielen des vorangegangenen Abschnitts gezeigt wurde, muss diese Eigenschaft jedoch nicht immer erfüllt sein. Zum anderen wurde dargestellt, dass die Varianz der Residuen nicht konstant ist. Dies bedeutet, dass auch die Gestalt der Verteilungsfunktionen Q bzw. K zeitlichen Veränderungen unterliegt. Anwendbar ist die Methode jedoch nur, wenn Verteilungen der Residuen im Betrachtungszeitraum unverändert sind. Es wird im Folgenden erläutert, wie diese Probleme bei der Kalibrierung der Voraussageverteilungen berücksichtigt werden können.

Um die Autokorrelation der Residuen zu berücksichtigen, wird für diese ein weiteres, separates Zeitreihenmodell benötigt. Es handelt sich der Idee nach um eine Modellierung der zeitlichen Abhängigkeit in zwei Schritten. Das ursprüngliche Modell sollte die zeitlichen Abhängigkeiten der Daten bereits im Wesentlichen beschreiben, so dass im zweiten Schritt ein einfaches Modell ausreicht, um die Autokorrelation der Residuen zu erfassen. Letzteres Modell liefert erneut Residuen, die im Idealfall unabhängig sind, deren Abhängigkeit aber zumindest geringer ist als für die Residuen des ursprünglichen Modells. Seien U_t^* bzw. R_t^* die Residuen des Modells für die UPR bzw. NPR und sei die zugehörige Transformation $H_U : U_t \mapsto U_t^*$ bzw. $H_R : R_t \mapsto R_t^*$ streng monoton, dann kann die korrekte Verteilungsfunktion zukünftiger Werte durch

$$G_{Y_t|\Omega_{t-1}}(y) = Q^T(H_U(F_{Y_t|\Omega_{t-1}}(y))) \quad (4.3)$$

$$\text{bzw. } G_{Y_t|\Omega_{t-1}}(y) = K^T(H_R(\Phi^{-1}(F_{Y_t|\Omega_{t-1}}(y)))) \quad (4.4)$$

beschrieben werden, wobei Q^T bzw. K^T die Verteilungsfunktionen der transformierten UPR bzw. NPR darstellen. Zusammengefasst ist also eine streng monotone Transformation zu finden und auf die ursprünglichen Residuen anzuwenden, so dass die transformierten Residuen unabhängig sind bzw. deren Abhängigkeit reduziert ist, bevor die oben beschriebene Methode zur Kalibrierung der Gestalt angewendet wird.

Beispiel 4.4 Angenommen, die Gestalt der Voraussageverteilungen soll mit Hilfe der NPR kalibriert werden und die Korrelationsfunktion der NPR zeigt

typische Eigenschaften eines autoregressiven Prozesses erster Ordnung, dann ist die direkte Anwendung wegen der korrelierten Residuen nicht möglich. Passt man an die Residuen zuvor jedoch ein AR(1)-Modell

$$R_t = \phi R_{t-1} + R_t^*$$

an und geht davon aus, dass die Residuen R_t^* des AR(1)-Modells unabhängig sind, so erhält man durch die Transformation

$$H : R_t \mapsto R_t^* \quad \text{mit} \quad H(R_t) = R_t - \phi R_{t-1}$$

die erwünschten unabhängigen Transformationen der NPR und die Kalibrierung kann durch

$$G_{Y_t|\Omega_{t-1}}(y) = K^T \left(\Phi^{-1}(\hat{F}_{Y_t|\Omega_{t-1}}(y)) - \hat{\phi}r_{t-1} \right)$$

erfolgen.



Das zweite Problem ist das veränderte Verhalten der Zeitreihe im Beobachtungszeitraum, welches sich in der Struktur der Residuen widerspiegelt. Um dies zu berücksichtigen, können die Parameter der Zeitreihe in regelmäßigen Zeitabständen oder bei Bedarf, d.h. wenn man Informationen über ein verändertes Verhalten der Zeitreihe besitzt, erneut geschätzt werden. Überträgt man dieses Vorgehen auf die hier vorgeschlagene Kalibrierung, so sind die Transformation H und deren Verteilungsfunktion Q^T bzw. K^T zu aktualisieren. (Die entsprechenden Symbole erhalten daher im Folgenden einen Zeitindex.) Der Aufwand dieser Schätzungen ist im Vergleich zu einer Aktualisierung der Parameterschätzer der ursprünglichen Modelle geringer. Dies gilt insbesondere für das Modell mit gleitenden Dichten, bei dem die für die Parameterschätzung benötigte Rechenzeit sehr groß ist.

4.5.3 Anwendung der Kalibrierung

In diesem Abschnitt werden die Daten und Modelle aus Abschnitt 4.4 erneut aufgegriffen. Die entsprechenden Voraussageverteilungen werden zunächst kalibriert und anschließend überprüft. Die Kalibrierung erfolgt hier mit Hilfe der UPR: In einem ersten Schritt werden die UPR der Modelle mit der Verteilungsfunktion einer Betaverteilung transformiert. Diese Transformation ist als

Zeitreihenmodell für die UPR zu interpretieren und dient dazu, die Autokorrelation der Residuen zu reduzieren, damit in einem zweiten Schritt die Gestalt der Voraussage kalibriert werden kann.

Zur Reduktion der Autokorrelation der UPR wird die streng monotone Transformation

$$H_t : U_t \mapsto U_t^* \quad \text{mit} \quad H_t(U_t) = F_{\text{Beta}}(U_t; \mu_t, \sigma_t^2)$$

verwendet, wobei die Modellierung der zeitlichen Abhängigkeit über die Parameter der Betaverteilung erfolgt, welche vom Residual der Vorperiode abhängen:

$$\mu_t = \delta_t(U_{t-1} - 0.5) + 0.5 \quad \text{und} \quad \sigma_t^2 = \frac{\mu_t(1 - \mu_t)}{3}.$$

Die Transformation besitzt einen Parameter δ_t , der, wie im letzten Abschnitt angedeutet, nicht aus den Daten der *in sample*-Periode zu schätzen ist, sondern wiederholt zu aktualisieren ist, um das veränderte Verhalten der Zeitreihe zu berücksichtigen. Hier erfolgt die Schätzung von δ_t jeweils aus den 100 vorangegangenen Beobachtungen mit der Methode der kleinsten Quadrate. Die Varianz der Betaverteilung ist so gewählt, dass sich eine $U(0, 1)$ -Verteilung ergibt, wenn ein Erwartungswert von 0.5 berechnet wird.

Dass diese Transformation den Charakter eines Zeitreihenmodells besitzt, wird deutlich, wenn man U_t in Abhängigkeit von U_{t-1} darstellt:

$$U_t = F_{\text{Beta}}^{-1}(U_t^*; \mu_t(U_{t-1}), \sigma_t(U_{t-1})).$$

An dieser Stelle ist ebenfalls zu erkennen, dass die transformierten Residuen U_t^* als *Cox-Snell residuals* (siehe Kapitel 2) des Modells für die UPR interpretiert werden können. Wenn das Modell die zeitliche Abhängigkeit der UPR korrekt beschreibt, sind die transformierten Residuen unabhängig, so dass in einem späteren Schritt die Gestalt der Voraussageverteilungen kalibriert werden kann.

Zur Verdeutlichung des Vorgehens wird im Folgenden das Vorgehen für den Zeitpunkt $t = 2501$ anhand des Hidden Markov-Modells explizit dargestellt. Die linke Graphik in Abbildung 4.16 basiert auf den UPR für den Zeitraum $t = 2401, \dots, 2500$. In ihr sind die Werte von $(u_{t-1} - 0.5)$ und die von $(u_t - 0.5)$ gegeneinander geplottet. Die eingezeichnete Regressionsgerade, die mit der Methode der kleinsten Quadrate berechnet wurde, repräsentiert die Schätzung von δ_{2501} . Sie schätzt die Abhängigkeit von U_t vom vorangegangenen Residual U_{t-1} . Für den hier betrachteten Zeitpunkt ergibt sich mit $\hat{\delta}_{2501} \approx 0.12$

$$\begin{aligned} (U_t - 0.5) &= 0.12 \cdot (U_{t-1} - 0.5) \\ \text{bzw.} \quad U_t &= 0.12 \cdot (U_{t-1} - 0.5) + 0.5. \end{aligned}$$

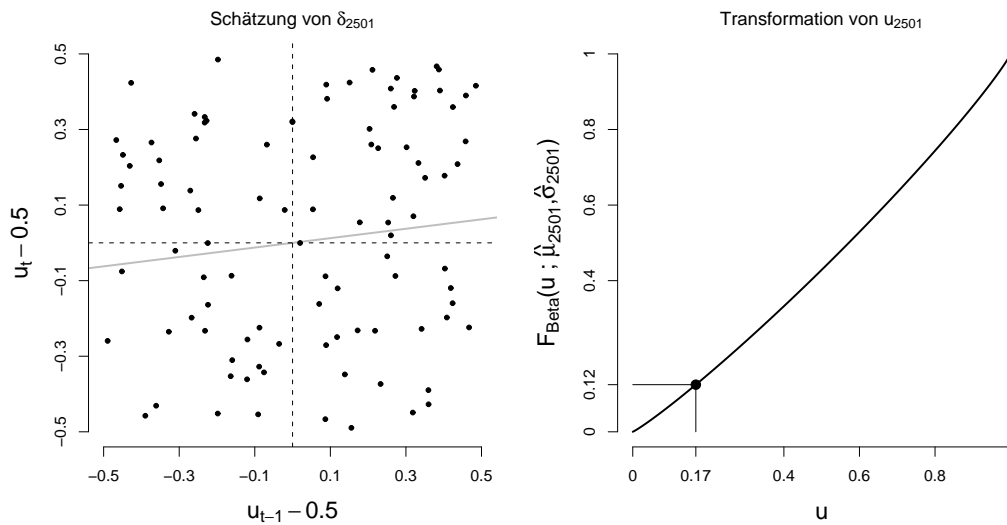


Abbildung 4.16: Schätzung der Transformation für $t = 2501$ und Transformation des entsprechenden UPR.

Das UPR für den Zeitpunkt $t = 2500$ besitzt in etwa den Wert 0.92, so dass die Schätzer für den Erwartungswert und die Varianz der Betaverteilung

$$\begin{aligned}\hat{\mu}_{2501} &= 0.12 \cdot (0.92 - 0.5) + 0.5 = 0.550 \\ \text{und } \hat{\sigma}_{2501}^2 &= \frac{0.55(1 - 0.55)}{3} = 0.0825\end{aligned}$$

sind. Die entsprechende Verteilungsfunktion ist im rechten Teil der Abbildung 4.16 dargestellt. Diese dient zur Transformation von $u_{2501} \approx 0.17$ und man erhält für den transformierten Wert

$$u_{2501}^* = F_{\text{Beta}}(0.17; \hat{\mu}_{2501}, \hat{\sigma}_{2501}^2) \approx 0.12.$$

In Abbildung 4.17 wird die Autokorrelation der ursprünglichen UPR mit der Autokorrelation der transformierten UPR verglichen. Auf der linken Seite der Abbildung sind die Autokorrelationsfunktionen von u_t und $(u_t - 0.5)^2$ dargestellt. (Diese sind vergleichbar mit den Autokorrelationsfunktionen der NPR und der quadrierten NPR aus Abbildung 4.6.) Die rechte Seite zeigt die entsprechenden Autokorrelationsfunktionen nach der Transformation. Durch den Vergleich der Graphiken (A) und (C) wird klar, dass die Autokorrelation der UPR für $lag = 1$, die vor der Transformation signifikant von null verschieden war, nach der Transformation deutlich geringer ist. Für $lag = 2$ hat sich

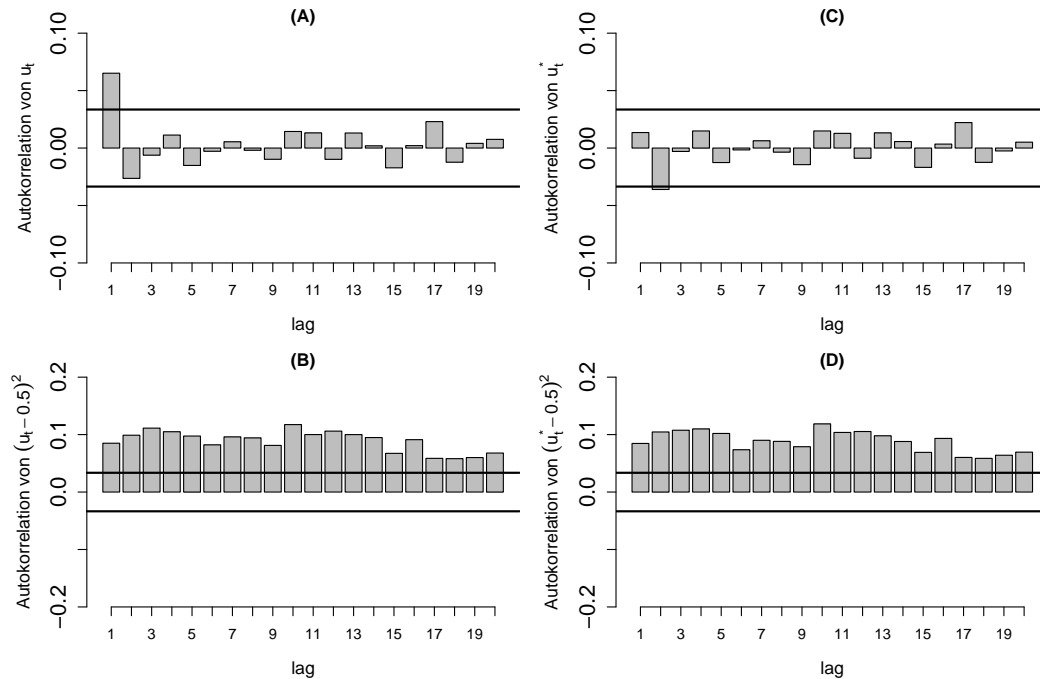


Abbildung 4.17: Korrelation der Residuen vor und nach der Transformation.

die Autokorrelation betragsmäßig leicht erhöht, was vermutlich darauf zurückzuführen ist, dass der Parameter der Transformation lediglich vom letzten beobachteten Residual abhängt. Ferner ist die Transformation nicht in der Lage, die Volatilitätsdynamik der Residuen zu erfassen, was bei einem Vergleich der Graphiken (B) und (D) deutlich wird: Die Autokorrelation der zentrierten und quadrierten UPR bleibt nahezu unverändert. Ein besseres Ergebnis wird erreicht, wenn neben dem Erwartungswert auch die Varianz der Beta-Verteilung aus den vergangenen Beobachtungen geschätzt wird. Eine weitere Möglichkeit, die Ergebnisse zu verbessern, wäre die Verwendung eines komplexeren Modells für die Residuen, beispielsweise eines GARCH-Modells. Diese Alternativen widersprechen jedoch der Idee, ein möglichst einfaches Modell für die Residuen zu verwenden, so dass die nachfolgende Kalibrierung der Gestalt der Voraussageverteilungen auf der oben beschriebenen Transformation basiert.

Um die Gestalt der Voraussageverteilungen zu kalibrieren, muss die Verteilungsfunktion der transformierten Residuen geschätzt werden. Die entsprechende Verteilung ist im Intervall $[0, 1]$ definiert, welches dem Bildbereich der Transformation H_t entspricht. Ein mögliches parametrisches Modell stellt daher die Beta-Verteilung dar. Eine Alternative ist die Verwendung eines Kernschätzers. DIEBOLD, HAHN AND TAY (1999) berichten jedoch, dass die

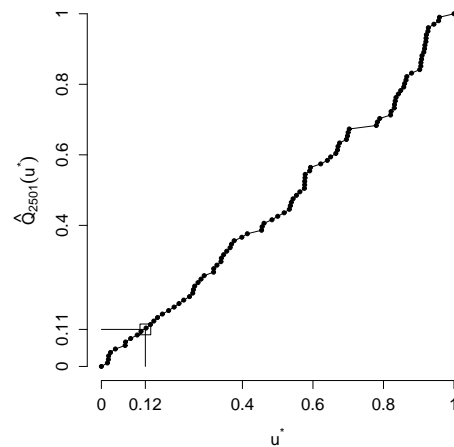


Abbildung 4.18: Empirische Verteilungsfunktion der transformierten Residuen für $t = 2501$.

Betaverteilung nicht flexibel genug ist, um die häufig multimodale Gestalt der Verteilung abzubilden, und dass ein Kernschätzer weniger geeignet ist, weil der Definitionsbereich der Verteilung begrenzt ist. Sie bevorzugen die Verwendung der empirischen Verteilungsfunktion, die auch hier verwendet wird, um die Gestalt der Voraussagen zu kalibrieren. Abbildung 4.18 zeigt die empirische Verteilungsfunktion der transformierten UPR des Hidden Markov-Modells aus dem Zeitraum $t = 2401, \dots, 2500$. Auf der Abszisse sind die sortierten Werte $u_{(t)}^*$ für den entsprechenden Zeitraum abgetragen, denen die Funktionswerte $i/(100+1)$ mit $i = 1, \dots, 100$ zugeordnet sind. Ergänzt sind diese Beobachtungen um die Punkte $(0, 0)$ und $(1, 1)$, die sich aus dem bekannten Definitionsbereich von U_t^* ergeben. Ferner sind alle Punkte linear miteinander verbunden. Diese Funktion wird als Schätzer für Q_{2501}^T verwendet und dient somit zur Kalibrierung der entsprechenden Voraussageverteilung.

Abbildung 4.19 zeigt die ursprünglichen und die kalibrierten Verteilungsfunktionen der Voraussage für den Zeitpunkt $t = 2501$. Die kalibrierten Verteilungsfunktionen (rechte Graphik) basieren auf einer Reihe vorangegangener Berechnungen. Ihre allgemeine Form ist in Gleichung 4.3 beschrieben. Für das Hidden Markov-Modell sind die zur Anwendung der Gleichung benötigten Funktionen zuvor berechnet worden. So ist die ursprüngliche Verteilungsfunktion $F_{Y_i|\Omega_{t-1}}(y)$ in Gleichung 4.1 gegeben, die Transformation H_{2501} ist die Verteilungsfunktion einer Betaverteilung mit Erwartungswert 0.55 und Varianz 0.0825 (siehe Abbildung 4.16, rechte Graphik) und Q_{2501}^T ist durch die in Abbildung 4.18 dargestellte Funktion geschätzt. Die kalibrierten Verteilungs-

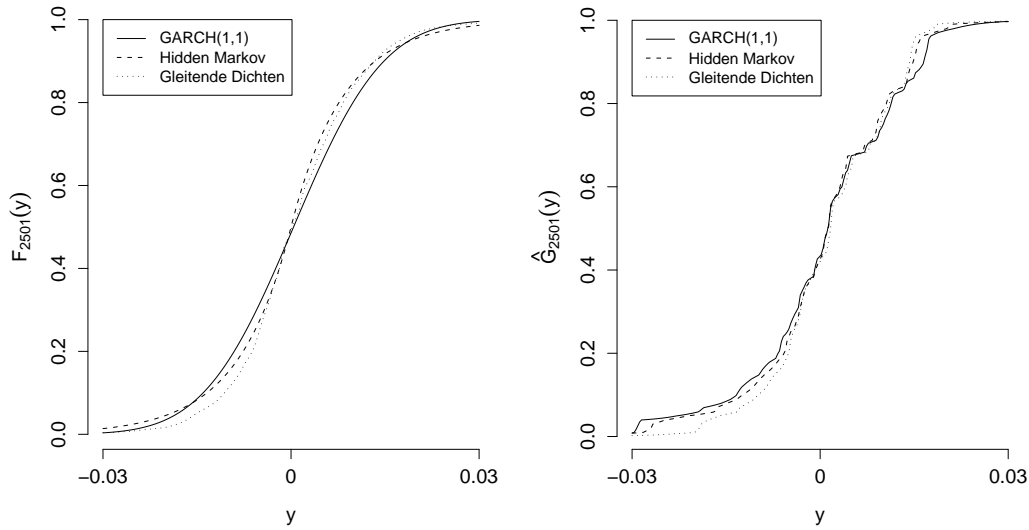


Abbildung 4.19: Verteilungsfunktionen der Voraussage für $t = 2501$ vor Kalibrierung (links) und nach Kalibrierung (rechts).

funktionen der Voraussage für das GARCH-Modell und für das Modell mit gleitenden Dichten erhält man durch analoge Anwendung der für das Hidden Markov-Modell beschriebenen Schritte. Abbildung 4.19 zeigt, dass die kalibrierten Verteilungsfunktionen nicht glatt sind. Dies liegt daran, dass zur Schätzung von Q^T die empirische Verteilungsfunktion der transformierten Residuen verwendet wird. Würde man stattdessen ein parametrisches Modell verwenden oder die empirische Verteilungsfunktion glätten, wären auch die kalibrierten Verteilungsfunktionen glatt.

Um die Güte der kalibrierten Voraussagen zu überprüfen, können dieselben Methoden verwendet werden, die in Abschnitt 4.4 zur Überprüfung der ursprünglichen Voraussagen verwendet wurden. Dabei ist es zur Berechnung der entsprechenden Pseudo-Residuen nicht notwendig, die kalibrierten Verteilungsfunktionen der Voraussage explizit zu bestimmen. Vielmehr reicht es aus, die bereits berechneten Pseudo-Residuen der ursprünglichen Modelle auf die gleiche Art zu transformieren wie die Verteilungsfunktionen der Voraussage:

$$u_t^{**} = \hat{G}_{Y_t|\Omega_{t-1}}(y_t) = \hat{Q}_t^T(H_t(F_{Y_t|\Omega_{t-1}}(y_t))) = \hat{Q}_t^T(H_t(u_t)).$$

Für das Hidden Markov-Modell ist die Berechnung des UPR der kalibrierten Voraussage für den Zeitpunkt $t = 2501$ in Abbildung 4.18 graphisch dargestellt. Nachdem zuvor bereits $H_{2501}(u_{2501}) = 0.12$ berechnet wurde, zeigt die Abbildung, dass man hier $u_t^{**} = \hat{Q}_{2501}^T(0.12) = 0.11$ erhält. Analog sind

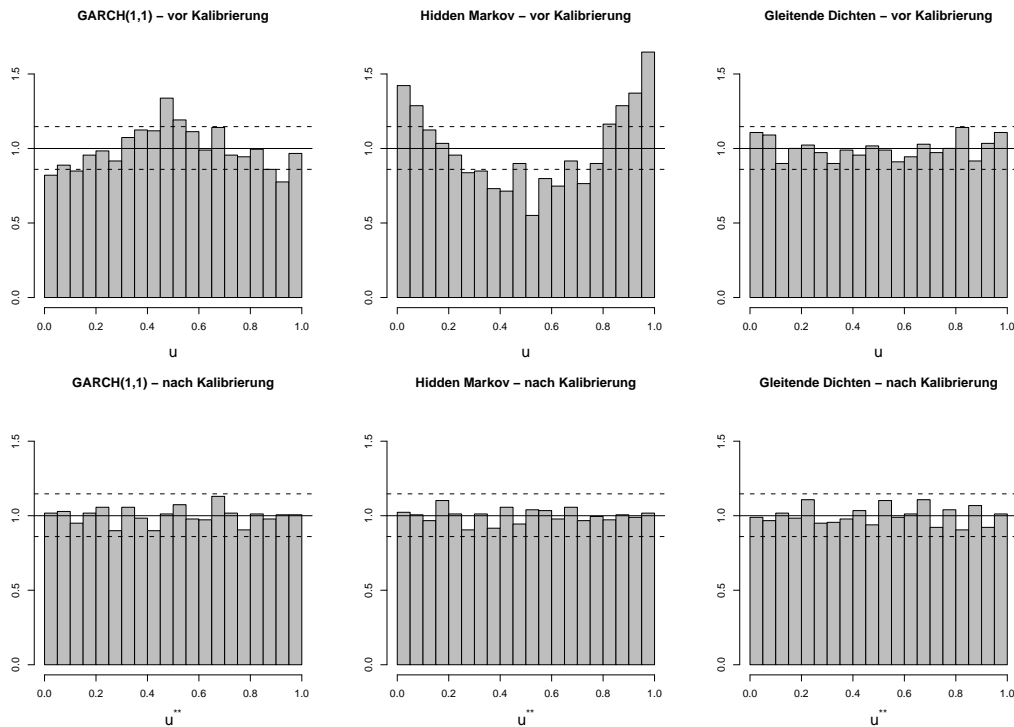


Abbildung 4.20: Histogramme der UPR vor Kalibrierung (oben) und nach Kalibrierung (unten).

sämtliche Pseudo-Residuen der *out of sample*-Periode der drei ursprünglichen Zeitreihenmodelle transformiert worden, um die Pseudo-Residuen der kalibrierten Voraussagen zu erhalten. In Abbildung 4.20 werden die Histogramme der UPR vor und nach der Kalibrierung verglichen. Die Histogramme besitzen nach der Kalibrierung in etwa die Gestalt einer $U(0, 1)$ -Verteilung. Die grundlegende Gestalt der kalibrierten Voraussageverteilungen ist also korrekt. Ferner gelingt die Abbildung extremer Beobachtungen mit den kalibrierten Voraussagen deutlich besser, was bei Betrachtung von Abbildung 4.21 zu erkennen ist. Die Abbildung zeigt die NPR der QQ-Plots vor und nach Kalibrierung der Voraussagen. Nach der Kalibrierung verlaufen die QQ-Plots deutlich näher entlang der Winkelhalbierenden. Insbesondere an den Enden der QQ-Plots gibt es weniger Ausreißer, was eine bessere Modellierung der extremen Renditen impliziert. Lediglich für die kalibrierte Version des Modells mit gleitenden Dichten sind zwei Beobachtungen zu registrieren, die vor der Kalibrierung besser beschrieben wurden (rechter Teil der Graphik unten rechts).

Abbildung 4.22 zeigt die Autokorrelation von $(u_t - 0.5)$ und von $(u_t^{**} - 0.5)$. Für alle Modelle konnte die signifikante Autokorrelation für $lag = 1$ deut-

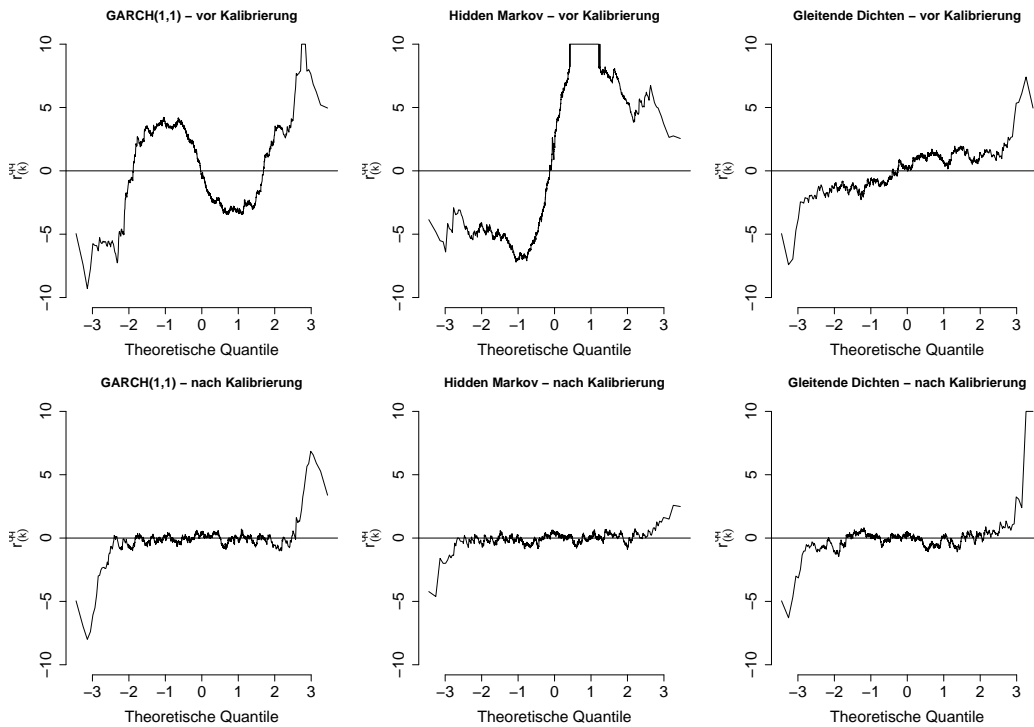


Abbildung 4.21: NPR der QQ-Plots vor Kalibrierung (oben) und nach Kalibrierung (unten).

lich verringert werden. Ferner wird durch einen Vergleich mit Abbildung 4.17 deutlich, dass die Transformation Q_t^T , die zur Kalibrierung der Gestalt genutzt wird, keinen wesentlichen Einfluss auf die Autokorrelation besitzt.

Die Autokorrelation der quadrierten Residuen (siehe Abbildung 4.23) konnte nur geringfügig verringert werden. Für das GARCH-Modell gab es vor der Kalibrierung keine signifikanten Werte, was sich auch durch die Kalibrierung nicht ändert. Für das Hidden Markov-Modell zeigt sich eine leichte Verbesserung, d.h. die Autokorrelation der zentrierten und quadrierten UPR ist nach der Kalibrierung geringer. Diese Verbesserung kann als Nebenprodukt der Kalibrierung der Gestalt verstanden werden, da die Transformation H_t alleine noch nicht zu einer Verringerung der entsprechenden Autokorrelation führte (siehe Abbildung 4.17). Ein identisch zu interpretierendes Ergebnis ergibt sich für das Modell mit gleitenden Dichten.

Die Voraussage der Volatilität im Zeitverlauf vor und nach der Kalibrierung wird in Abbildung 4.24 verglichen. Für das GARCH-Modell verläuft die CUSUMSQ der NPR sowohl vor als auch nach der Kalibrierung innerhalb der Kon-

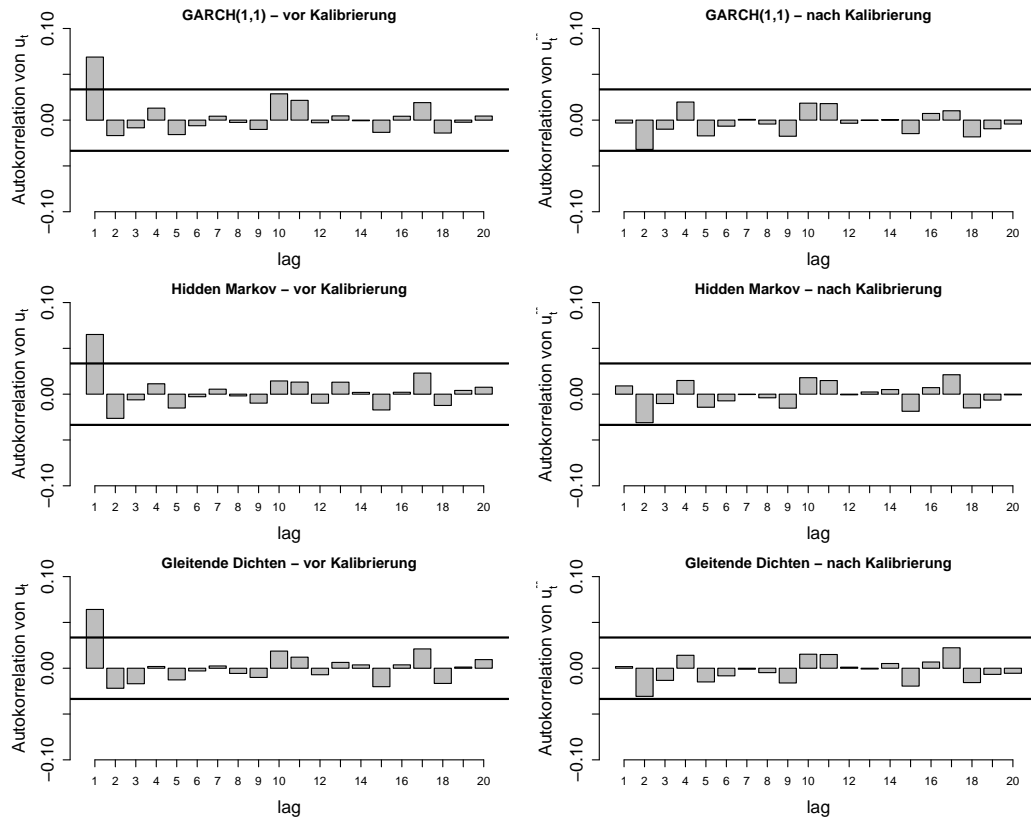


Abbildung 4.22: Autokorrelation der UPR vor Kalibrierung (links) und nach Kalibrierung (rechts).

fidenzbänder. Für das Hidden Markov-Modell wird die systematische Struktur der Kurve durch die Kalibrierung der Voraussagen beseitigt, d.h. nach der Kalibrierung verlässt die entsprechenden CUSUMSQ den kritischen Bereich nicht. Die Varianz der NPR in der gesamten *out of sample*-Periode beträgt nach der Kalibrierung ca. 1.02. Die Voraussage der Volatilität kann durch die Kalibrierung demnach wesentlich verbessert werden. Dies gilt auch für das Modell mit gleitenden Dichten. Allerdings verlässt die CUSUMSQ auch nach der Kalibrierung die Konfidenzbänder und bei einer Bestimmung der Änderungszeitpunkte der Residualvarianz ergeben sich die in Tabelle 4.6 dargestellten Ergebnisse. Die Anzahl der Änderungszeitpunkte kann durch die Kalibrierung von elf auf drei reduziert werden. Diese Ergebnisse lassen erkennen, dass die Kalibrierung mit den jeweils aktualisierten Transformationsparametern auch in der Lage ist, die zeitlichen Veränderungen der Zeitreihe zu berücksichtigen.

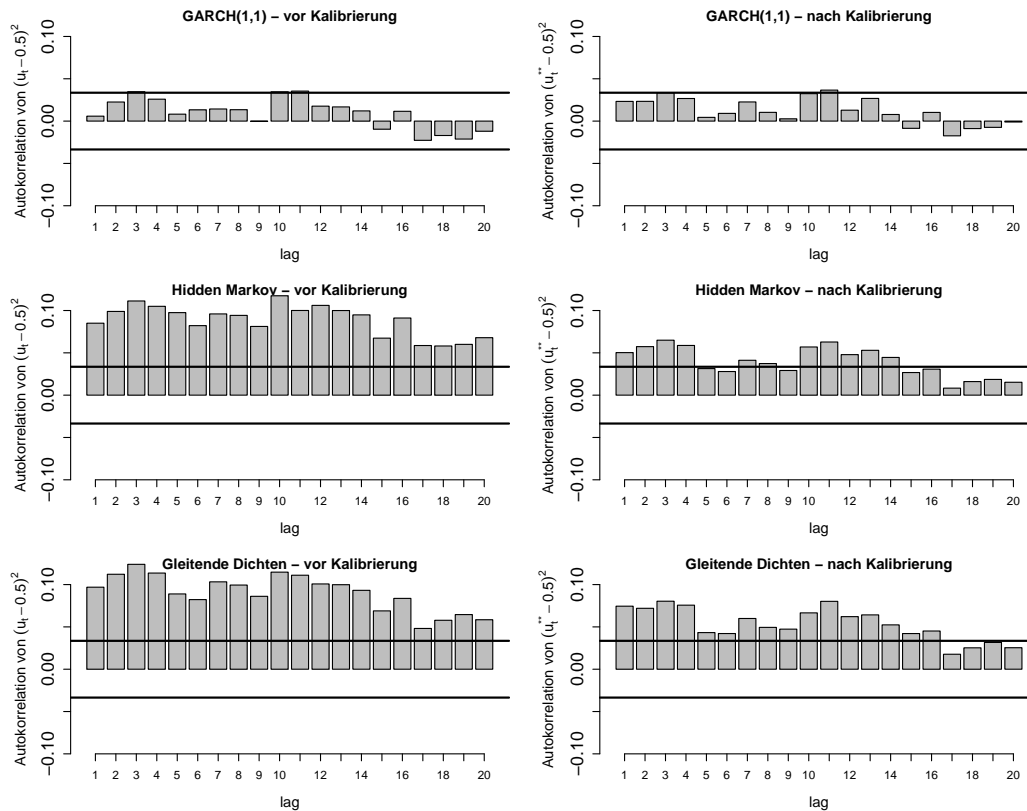


Abbildung 4.23: Autokorrelation der zentrierten und quadrierten UPR vor Kalibrierung (links) und nach Kalibrierung (rechts).

Insgesamt wurde die Voraussage zukünftiger Werte durch die Kalibrierung der Voraussagen deutlich verbessert. Weitere (hier nicht explizit dargestellte) Berechnungen zeigten, dass es für die in dieser Arbeit betrachtete Zeitreihe nicht ausreicht, die Parameter der Transformationen aus den Beobachtungen der *in sample*-Periode zu schätzen und Ergebnisse für die gesamte *out of sample*-Periode zu verwenden. Auch die Anzahl der zurückliegenden Beobachtungen, die zur Schätzung von H_t und Q_t^T verwendet wird, beeinflusst die Ergebnisse der Kalibrierung. Je größer diese Anzahl gewählt wird, desto weniger flexibel reagieren die Transformationen auf ein verändertes Verhalten der Zeitreihe. Werden nicht mehr als ca. 500 vorangegangene Beobachtungen zur Schätzung der Transformationsparameter verwendet, unterscheiden sich die Ergebnisse der Kalibrierung nur unwesentlich von den oben dargestellten Ergebnissen. Verwendet man jedoch eine größere Anzahl, tendieren die Modelle zu fehlerhaften Voraussagen. Wenn andererseits nur wenige Beobachtungen verwendet werden, um die Transformationen zu aktualisieren, reagieren diese flexibler auf

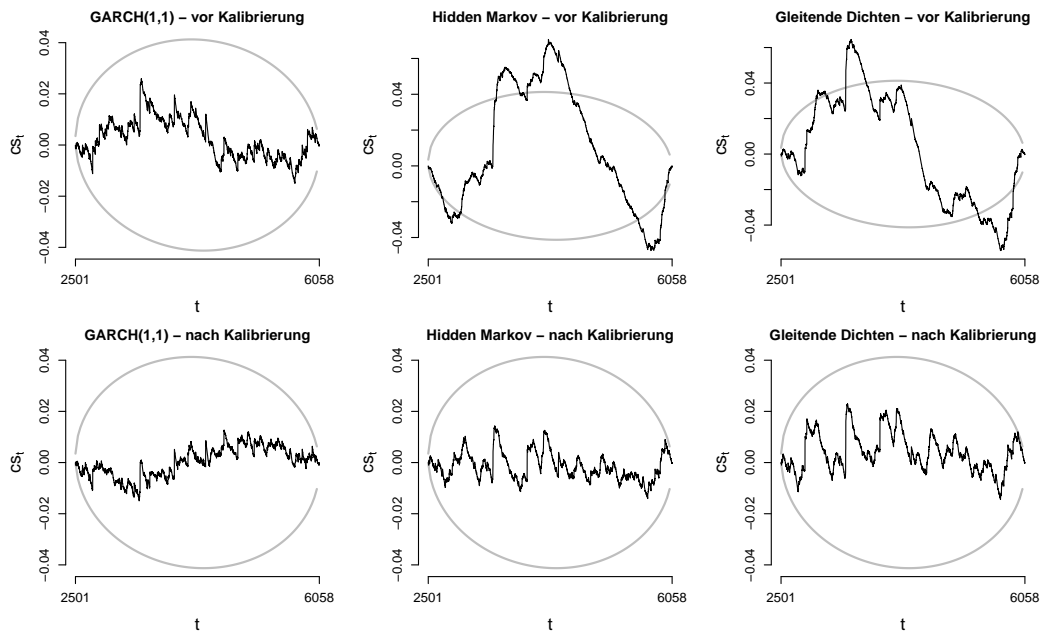


Abbildung 4.24: Zentrierte CUSUMSQ der NPR vor Kalibrierung (oben) und nach Kalibrierung (unten).

Veränderungen der Zeitreihe. Die relativ große Variation von $\hat{\delta}_t$, (dem Parameterschätzer der Transformation H_t , siehe Abbildung 4.16), kann allerdings dazu führen, dass der geschätzte Erwartungswert der Beta-Verteilung für einige Zeitpunkte größer als eins bzw. kleiner als null ist. Um dies zu vermeiden, müssen beispielsweise für das GARCH-Modell mindestens zwölf vorangegangene Beobachtungen verwendet werden, um die δ_t zu schätzen. Wie bereits zuvor erwähnt, werden in dieser Arbeit die jeweils letzten 100 berechneten UPR verwendet. Diese Anzahl ist heuristisch gewählt. Sie gewährleistet eine relativ gute Modellierung der zeitlichen Veränderungen in der Zeitreihe und es gab bei keinem der betrachteten Modelle Probleme mit den geschätzten Erwartungswerten der Beta-Verteilung.

Periodenbeginn	2501	5636	5703	5956
Periodenende	5635	5702	5955	6058
$\widehat{Var}(R_t)$	1.04	0.50	1.43	0.61

Tabelle 4.6: Geschätzte Varianz der Residuen des Modells mit gleitenden Dichten nach Kalibrierung.

Die hier betrachteten Modelle wurden auch an die Renditezeitreihen anderer Aktien angepasst. Insgesamt wurden die Aktien von acht weiteren Unternehmen aus unterschiedlichen Branchen untersucht. Dabei wurde wie in dem hier dargestellten Beispiel die Güte der Voraussagen vor und nach der Kalibrierung beurteilt. In allen Fällen führte die Kalibrierung zu einer deutlichen Verbesserung der Voraussagen.

4.6 Abschließende Bemerkungen zum Kapitel

Die Analyse von Pseudo-Residuen zur Überprüfung von Zeitreihenmodellen für Aktienrenditen ist (zumindest für die UPR) ein etabliertes und hilfreiches Instrument zur Identifikation von Anpassungsfehlern. Hier liegt der Schwerpunkt auf dynamischen Modellen, d.h. auf Modellen mit zeitabhängiger Varianz (GARCH-Modell) oder auch zeitabhängigen Verteilungstypen (Modell mit gleitenden Dichten). Es ist offensichtlich, dass einfache Residuenkonzepte wie beispielsweise die *ordinary residuals* in solchen Situationen nicht zur Modellüberprüfung geeignet sind. Mit Hilfe der Pseudo-Residuen hingegen kann eine Vielzahl unterschiedlicher Anpassungskriterien untersucht werden, was beispielsweise für die Erfassung der zeitlichen Abhängigkeiten oder die Abbildung extremer Beobachtungen dargestellt wurde. Insbesondere die korrekte Voraussage der Volatilität stellt ein wesentliches Kriterium dar, das von einem Modell erfüllt werden sollte, und es wurde hier gezeigt, wie (basierend auf einer Analyse der NPR) Perioden identifiziert werden können, in denen ein Modell dieses Kriterium nicht erreicht.

Ferner wurde beschrieben, wie die Pseudo-Residuen genutzt werden können, um die Voraussagen eines Modells zu verbessern. Hier war es die Idee, ein vorhandenes Modell als erste Approximation zu verwenden und dieses unter Verwendung eines zweiten Modells für die Pseudo-Residuen zu kalibrieren. Dieses „*fine-tuning*“ hat den Vorteil, dass die entsprechenden Schätzer (auch bei regelmäßiger Aktualisierung) ohne großen Aufwand bestimmt werden können. Auf der anderen Seite muss allerdings darauf hingewiesen werden, dass ein gewisser Interpretationsverlust mit diesem Vorgehen verbunden ist: Während die ursprünglichen Modelle jeweils auf ganz bestimmten Vorstellungen vom Verhalten der Zeitreihe basieren, ist es schwer, solche Vorstellungen für das „geschachtelte“ Modell zu entwickeln.

Kapitel 5

Überprüfung multivariater Verteilungen

5.1 Einführung

Alle bis zu dieser Stelle der Arbeit beschriebenen Methoden zur Analyse von Pseudo-Residuen eignen sich zur Überprüfung univariater Verteilungen. Auch wenn bereits in der Einleitung die Güte einer multivariaten Verteilung (eine Mischverteilung zweier bivariater Normalverteilungen) untersucht wurde, so bezog sich die Analyse auf univariate Pseudo-Residuen, d.h. die Anpassung der Randverteilungen wurde separat untersucht. Die so erhaltenen Aussagen zur Güte des Modells waren somit auf die Randverteilungen beschränkt. Sicherlich ist es möglich, die Analyse auf die bedingten Verteilungen auszuweiten und somit weitere Informationen über die Anpassungsgüte zu gewinnen. Aber auch hier können durch eine separate Analyse der Residuen nur Teilaspekte der Modellanpassung untersucht werden.

In diesem Abschnitt werden multivariate Pseudo-Residuen definiert, die bei der Überprüfung von multivariaten Verteilungen verwendet werden können und detaillierte Aussagen zur Anpassungsgüte der gemeinsamen Verteilung zulassen. Der wesentliche Unterschied zu den bisherigen Betrachtungen liegt darin, dass mehrere univariate Residuen simultan untersucht werden: Ihre gemeinsame Verteilung wird mit verschiedenen Methoden analysiert, um Abweichungen von der bei Gültigkeit des betrachteten Modells vorliegenden gemeinsamen Verteilung zu identifizieren.

Im Folgenden wird zunächst die Definition des stochastischen Modells aus Ab-

schnitt 2.2 mit Blick auf die in diesem Abschnitt betrachtete Problemstellung erweitert. Anschließend werden die multivariaten Pseudo-Residuen definiert und es wird die gemeinsame Verteilung der Residuen bestimmt. In Abschnitt 5.4 werden Methoden zur Analyse der Residuen erläutert, die sich auf drei verschiedene Aspekte der Modellanpassung beziehen. Dabei handelt es sich um die grundlegende Gestalt der gemeinsamen Verteilung, die Erfassung extremer Beobachtungen sowie die Fähigkeit der gemeinsamen Verteilung, die Abhängigkeit zwischen den zu modellierenden Variablen korrekt zu beschreiben.

5.2 Definition des stochastischen Modells für multivariate Verteilungen

Die allgemeine Beschreibung der in diesem Abschnitt betrachteten Modelle, d.h. der multivariaten Verteilungen, erfolgt durch die Formulierung:

Die Elemente von \mathbf{Y}_i besitzen die
gemeinsame Verteilungsfunktion $F_i(\mathbf{y}; \boldsymbol{\theta}_i)$ für $i = 1, \dots, n$.

Dabei sind die $\mathbf{Y}_i = (Y_{1i}, \dots, Y_{ki}, \dots, Y_{pi})'$ für $i = 1, \dots, n$ Vektoren von jeweils p Zufallsvariablen und die Parametervektoren der gemeinsamen Verteilung sind weiterhin mit $\boldsymbol{\theta}_i$ bezeichnet.

Analog zum univariaten Fall können auch hier verschiedene Modellsituationen unterschieden werden. Gilt beispielsweise $F_i = F$ und $\boldsymbol{\theta}_i = \boldsymbol{\theta}$ für alle i , handelt es sich um ein Modell, das alle (multivariaten) Beobachtungen durch eine identische gemeinsame Verteilung beschreibt, d.h. es wird davon ausgegangen, dass der Zufallsprozess, der die Beobachtungen erzeugt, für alle Beobachtungen der gleiche ist. Hier ist unter anderem das einführende Beispiel einzuordnen, bei dem der Brusthöhendurchmesser und die Höhe der Bäume für alle Beobachtungen durch eine Mischung von zwei bivariaten Normalverteilungen mit identischen Parametern erfolgte.

Komplizierter wird das Modell für unterschiedliche Parametervektoren. Wie auch im univariaten Fall sei in diesem Zusammenhang ein Regressionsmodell als Beispiel erwähnt. Die Erweiterung gegenüber dem klassischen linearen Regressionsmodell besteht darin, dass die zu erklärende Variable multivariat ist und die Parameter der gemeinsamen Verteilung von erklärenden Variablen abhängen, sich also in der Regel unterscheiden. Ein solches Modell wird zum

Beispiel bei OAKES AND RITZ (2000) betrachtet. Im Rahmen dieser Arbeit wird allerdings nicht näher auf diesen Fall eingegangen, da die grundsätzlichen Ideen dieses Abschnitts auch für identische Parametervektoren und somit anhand einfacherer Modelle dargestellt werden können. Aus dem gleichen Grund wird auch auf die dritte Modellsituation, d.h. auf Modelle mit unterschiedlichen Parametervektoren und unterschiedlichen gemeinsamen Verteilungstypen, nicht eingegangen. Durch diese Einschränkung kann und wird im Folgenden auf den Index i an der Verteilungsfunktion und am Parametervektor verzichtet. Ferner wird der Index — wenn möglich und sinnvoll — bei beobachtungsbezogenen Ausdrücken vernachlässigt, also beispielsweise \mathbf{u} statt \mathbf{u}_i oder Y_k statt Y_{ki} geschrieben.

5.3 Multivariate Pseudo-Residuen

Definition stetiger multivariater Pseudo-Residuen Das multivariate *Uniform-Pseudo-Residual* (mUPR)

$$\mathbf{u} = (u_1, \dots, u_p)'$$

einer Beobachtung $\mathbf{y} = (y_1, \dots, y_p)'$ ist (wie die Beobachtung selbst) ein Vektor mit p Elementen. Die Berechnung der einzelnen Elemente erfolgt mit Hilfe

- (1) der univariaten Verteilungsfunktion *einer* Randverteilung und
- (2) den univariaten Verteilungsfunktionen von $p - 1$ bedingten Verteilungen.

Konkret wird ein $k_1 \in \{1, \dots, p\}$ ausgewählt und mit Hilfe der Randverteilung von Y_{k_1} wird

$$u_{k_1} = F_{Y_{k_1}}(y_{k_1})$$

bestimmt, d.h. u_{k_1} ist das univariate UPR für die Beobachtung y_{k_1} bezüglich der Randverteilung von Y_{k_1} . Anschließend wird mit Hilfe der bedingten Verteilungsfunktion einer weiteren Zufallsvariablen gegeben $Y_{k_1} = y_{k_1}$ ein zweites Element von \mathbf{u} bestimmt:

$$u_{k_2} = F_{Y_{k_2}|Y_{k_1}}(y_{k_2}) \quad \text{mit } k_2 \in \{1, \dots, p\} \text{ und } k_2 \neq k_1.$$

Das dritte Element von \mathbf{u} berechnet sich dann durch $u_{k_3} = F_{Y_{k_3}|Y_{k_1}Y_{k_2}}(y_{k_3})$, d.h. die Zufallsvariablen in der Bedingung sind diejenigen, für die zuvor ein

entsprechendes Element von \mathbf{u} bestimmt wurde. Gleiches gilt auch bei der Berechnung aller weiteren Elemente des mUPR.

Da die Reihenfolge, in der die Elemente bestimmt werden, grundsätzlich frei gewählt werden kann, gibt es $p!$ verschiedene Möglichkeiten, die Residuen zu berechnen. Wählt man beispielsweise $k_1 = 1, k_2 = 2, \dots, k_p = p$ ist das mUPR allgemein durch

$$\begin{aligned} \mathbf{U} &= (U_1, \dots, U_p)' \\ &= (F_{Y_1}(Y_1), F_{Y_2|Y_1}(Y_2), F_{Y_3|Y_1Y_2}(Y_3), \dots, F_{Y_p|Y_1\dots Y_{p-1}}(Y_p))' \\ &\equiv (U_{1|}, U_{2|1}, U_{3|12}, \dots, U_{p|1\dots(p-1)})' \end{aligned}$$

definiert. (Bezüglich der Notation sei bemerkt, dass ein einfacher Index an den Elementen des Residuals (erste Zeile) die Elemente in ihrer allgemeinen Form kennzeichnet und eine konkrete Reihenfolge bei der Definition wie in der dritten Zeile indiziert wird.) Je nach Problemstellung kann es sinnvoll sein, eine bestimmte Reihenfolge bei der Berechnung der Elemente von \mathbf{u} zu wählen. Dies ist dann der Fall, wenn die Güte bestimmter bedingter Verteilungen von besonderem Interesse ist. In Abschnitt 5.4.3.1 wird beispielsweise die gemeinsame Modellierung der Höhe und des Durchmessers von Bäumen erneut aufgegriffen. Dort ist die bedingte Verteilung der Baumhöhen bei gegebenem Brusthöhendurchmesser von besonderem Interesse. Grundsätzlich gilt allerdings, dass eine vollständige Analyse der Residuen die Betrachtung aller möglichen Definitionen (Reihenfolgen) umfasst.

Analog zum univariaten Fall wird auch bei der Modellierung multivariater Beobachtungen durch

$$\begin{aligned} \mathbf{R} &= (\Phi^{-1}(U_1), \dots, \Phi^{-1}(U_p))' \\ &= (R_1, \dots, R_p)' \end{aligned}$$

ein multivariates Normal-Pseudo-Residual (mNPR) definiert, das für bestimmte Analysen der Modellgüte besser geeignet ist als das mUPR.

Beispiel 5.1 Zur Verdeutlichung der Definition der multivariaten Pseudo-Residuen sei als Modell eine bivariate Normalverteilung

$$\mathbf{Y} = (Y_1, Y_2)' \sim BN(\mu_{Y_1} = 0, \mu_{Y_2} = 1, \sigma_{Y_1}^2 = 1, \sigma_{Y_2}^2 = 4, \rho = 0.5)$$

betrachtet. Ferner sei $\mathbf{y} = (1, 0.5)'$ die Beobachtung, für die das mUPR und das mNPR zu bestimmen ist. Es gibt in diesem Fall zwei Möglichkeiten, die

multivariaten Pseudo-Residuen (mUPR und mNPR) der Beobachtung zu berechnen. Eine Möglichkeit ist die Verwendung von

$$\mathbf{u}^{(1)} = (u_{1|\cdot}, u_{2|1})' = (F_{Y_1}(1), F_{Y_2|Y_1}(0.5))'$$

Die Randverteilung von Y_1 ist hier eine (univariate) Normalverteilung mit Erwartungswert $E(Y_1) = \mu_{Y_1} = 0$ und Varianz $Var(Y_1) = \sigma_{Y_1}^2 = 1$. Die bedingte Verteilung von $Y_2|Y_1$ ist ebenfalls eine univariate Normalverteilung. Diese besitzt den Erwartungswert $E(Y_2|Y_1) = \mu_{Y_2} + \rho\sigma_{Y_2}(y_1 - \mu_{Y_1})/\sigma_{Y_1} = 2$ und die Varianz $Var(Y_2|Y_1) = \sigma_{Y_2}^2(1 - \rho^2) = 3$. Damit ergibt sich für das mUPR der Beobachtung

$$\mathbf{u}^{(1)} = \left(\Phi\left(\frac{1-0}{\sqrt{1}}\right), \Phi\left(\frac{0.5-2}{\sqrt{3}}\right) \right)' = (0.84, 0.19)'$$

und das mNPR ist

$$\mathbf{r}^{(1)} = \left(\frac{1-0}{\sqrt{1}}, \frac{0.5-2}{\sqrt{3}} \right)' = (1, -0.87)'$$

Die zweite Möglichkeit zur Berechnung des mUPR ist die Verwendung von

$$\mathbf{u}^{(2)} = (u_{1|2}, u_{2|\cdot})' = (F_{Y_1|Y_2}(1), F_{Y_2}(0.5))'$$

Durch eine analoge Bestimmung der Parameter der Verteilungen von $Y_1|Y_2$ und Y_2 erhält man hier

$$\mathbf{u}^{(2)} = \left(\Phi\left(\frac{1 - (-0.125)}{\sqrt{0.75}}\right), \Phi\left(\frac{0.5 - 1}{\sqrt{4}}\right) \right)' = (0.90, 0.40)'$$

für das mUPR und das entsprechende mNPR ist

$$\mathbf{r}^{(2)} = \left(\frac{1 - (-0.125)}{\sqrt{0.75}}, \frac{0.5 - 1}{\sqrt{4}} \right)' = (1.30, -0.25)'$$

◀

Gemeinsame Verteilung der Elemente Zur Analyse der Residuen wird die gemeinsame Verteilung der Elemente von \mathbf{U} benötigt. Beschreibt die multivariate Verteilung die Daten korrekt, so ist die gemeinsame Dichtefunktion der Elemente von \mathbf{U} (unabhängig von der gewählten Reihenfolge ihrer Berechnung) durch

$$f_{\mathbf{U}}(u_1, \dots, u_p) = \begin{cases} 1 & \text{für } 0 \leq u_1, \dots, u_p \leq 1 \\ 0 & \text{sonst} \end{cases} \quad (5.1)$$

gegeben, d.h. die Elemente von \mathbf{U} sind bei Gültigkeit des Modells unabhängig und identisch $U(0, 1)$ -verteilt (vgl. ROSENBLATT, 1952). Um dies zu zeigen, wird eine bivariate Problemstellung betrachtet, in der das mUPR durch

$$\mathbf{U} = (U_1, U_2)' = (F_{Y_1}(Y_1), F_{Y_2|Y_1}(Y_2))'$$

definiert sei. Es handelt sich hier um eine Transformation $(Y_1, Y_2) \mapsto (U_1, U_2)$ der Zufallsvariablen Y_1 und Y_2 mit

$$u_1 = F_{Y_1}(y_1) \quad \text{und} \quad u_2 = F_{Y_2|Y_1}(y_2).$$

Sei $g_{Y_1 Y_2}(y_1, y_2)$ die gemeinsame Dichtefunktion von Y_1 und Y_2 , dann ist die Dichtefunktion der gemeinsamen Verteilung von U_1 und U_2 durch

$$f_{U_1 U_2}(u_1, u_2) = \begin{cases} g_{Y_1 Y_2} \left(F_{Y_1}^{-1}(u_1), F_{Y_2|Y_1}^{-1}(u_2) \right) \cdot |J| & (u_1, u_2) \in A \\ 0 & \text{sonst} \end{cases} \quad (5.2)$$

gegeben, wobei $A = \{(u_1, u_2) : 0 \leq u_1, u_2 \leq 1\}$ den Bildbereich der Transformationen darstellt und die Jacobideterminante $|J|$ für die betrachtete Transformation durch

$$\begin{aligned} |J| &= \begin{vmatrix} \frac{\partial y_1}{\partial u_1} & \frac{\partial y_1}{\partial u_2} \\ \frac{\partial y_2}{\partial u_1} & \frac{\partial y_2}{\partial u_2} \end{vmatrix} \\ &= \frac{1}{f_{Y_1}(F_{Y_1}^{-1}(u_1))} \cdot \frac{1}{f_{Y_2|Y_1}(F_{Y_2|Y_1}^{-1}(u_2))} \\ &= \frac{1}{f_{Y_1 Y_2} \left(F_{Y_1}^{-1}(u_1), F_{Y_2|Y_1}^{-1}(u_2) \right)} \end{aligned}$$

gegeben ist. Durch Einsetzen der Jacobideterminante in (5.2) erhält man

$$f_{U_1 U_2}(u_1, u_2) = \begin{cases} \frac{g_{Y_1 Y_2} \left(F_{Y_1}^{-1}(u_1), F_{Y_2|Y_1}^{-1}(u_2) \right)}{f_{Y_1 Y_2} \left(F_{Y_1}^{-1}(u_1), F_{Y_2|Y_1}^{-1}(u_2) \right)} & (u_1, u_2) \in A \\ 0 & \text{sonst.} \end{cases} \quad (5.3)$$

An dieser Stelle wird deutlich, dass man bei Verwendung des korrekten Modells ($f_{Y_1 Y_2} = g_{Y_1 Y_2}$) den bivariaten Spezialfall der allgemeinen Dichtefunktion aus Gleichung 5.1 erhält. Für höherdimensionale Fälle und/oder für eine andere Reihenfolge bei der Definition der Elemente von \mathbf{U} ist der Beweis analog zu führen (siehe dazu auch DIEBOLD, HAHN AND TAY, 1999). Weiterhin ergibt sich aus der gemeinsamen Verteilung der Elemente des mUPR unmittelbar, dass die Elemente des mNPR unabhängig und identisch standardnormalverteilt sind.

5.4 Überprüfung multivariater Verteilungen

Wie auch im univariaten Fall basiert die Überprüfung multivariater Verteilungen auf der Analyse der beobachteten Verteilung der Residuen. Diese wird auf Abweichungen von der bei Gültigkeit des Modells geltenden Verteilung untersucht und entsprechende Abweichungen sind mit Blick auf die Modellanpassung zu interpretieren. Allerdings ist die Analyse der Residuen bei der in diesem Kapitel betrachteten Problemstellung komplizierter, weil die entsprechenden Methoden zur Untersuchung von multivariaten Verteilungen geeignet sein müssen. Einige Methoden, die dieser Anforderung genügen und sich auf verschiedene Aspekte der Modellanpassung beziehen, werden in den folgenden Abschnitten beschrieben.

5.4.1 Grundlegende Gestalt der gemeinsamen Verteilung

Eine erste Möglichkeit die gemeinsame Verteilung der Residuen zu untersuchen, ist die Verwendung von Signifikanztests. Das Ergebnis eines Signifikanztests ist jedoch sehr allgemein, d.h. die Hypothese, dass die multivariaten Pseudo-Residuen gemeinsam rechteckverteilt bzw. standardnormalverteilt sind, wird entweder abgelehnt oder nicht abgelehnt. Damit erscheinen sie für die hier betrachtete Problemstellung weniger geeignet, weil sie keine Auskunft darüber geben, welche Aspekte der Daten nicht hinreichend durch das angepasste Modell beschrieben werden. Es seien in diesem Zusammenhang dennoch zwei Arbeiten erwähnt, weil entsprechende Testverfahren (im Gegensatz zu Anpassungstest für univariate Beobachtungen) weniger verbreitet sind und sie trotz der begrenzten Aussagekraft in einem engen Zusammenhang zu der hier betrachteten Problemstellung stehen: Jeweils einen kurzen Überblick für Testverfahren zur Prüfung auf multivariate Standardnormalverteilung bzw. multivariate $U(0, 1)$ -Verteilung geben BOGDAN (1999) bzw. LIANG ET AL. (2000).

In diesem Abschnitt wird dargestellt, wie man einen ersten Eindruck von der grundlegenden Gestalt des angepassten Modells erhalten kann. Dazu wird die Darstellung der mUPR in einem dreidimensionalen Histogramm erläutert, mit dem die gemeinsame Verteilung bivariater UPR bzw. die gemeinsame Verteilung von Paaren der Elemente eines mUPR analysiert werden kann. Neben der graphischen Beurteilung des Histogramms, wird die Verwendung von diskre-

ten Pseudo-Residuen zur Untersuchung der beobachteten Häufigkeiten in den einzelnen Klassen beschrieben.

Für univariate Verteilungen wurde im Verlauf der Arbeit bereits veranschaulicht, dass aus einem Vergleich des Histogramms beobachteter UPR mit der Dichtefunktion der $U(0, 1)$ -Verteilung Informationen über die Güte des angepassten Modells abgeleitet werden können. Bei der hier betrachteten multivariaten Problemstellung sind bei Gültigkeit des angepassten Modells alle Paare von Elementen eines mUPR bivariat und unabhängig $U(0, 1)$ -verteilt, d.h. die gemeinsame Dichtefunktion ist (geometrisch interpretiert) ein Würfel mit der Seitenlänge eins. Die graphische Darstellung (ein dreidimensionales Histogramm), die für den Vergleich mit dieser bivariaten Dichte verwendet wird, basiert auf der Konstruktionsidee von Histogrammen für univariate Beobachtungen. Während letztere die beobachteten Häufigkeiten in Form von Rechtecken abbilden, werden hier allerdings Quader benötigt, um die Verteilung der bivariaten UPR graphisch darzustellen. Um diese zu bestimmen, wird der Definitionsbereich $A = \{(u_1, u_2) : 0 \leq u_1, u_2 \leq 1\}$ der bivariaten UPR zunächst in Teilflächen A_t zerlegt, d.h. die Teilflächen sind disjunkt und die Vereinigung aller Teilflächen entspricht dem Definitionsbereich A . Analog zu den herkömmlichen Histogrammen werden diese Teilflächen im Folgenden als Klassen bezeichnet. Sie entsprechen den Grundflächen der Quader. Sei N_t die Anzahl der beobachteten UPR in der Klasse A_t und F_t der Flächeninhalt von A_t , dann ist die Höhe der Quader durch

$$H_t = \frac{N_t}{F_t n}$$

definiert, wobei n auch hier die Gesamtzahl der Beobachtungen darstellt. Das Volumen der einzelnen Quader entspricht demnach der relativen Häufigkeit der in der entsprechenden Klasse beobachteten UPR und das Gesamtvolumen aller Quader beträgt eins.

Um die Gestalt des dreidimensionalen Histogramms objektiv zu beurteilen, ist es sinnvoll, die Verteilungen der H_t zu betrachten. Bei Gültigkeit des angepassten Modells sind die N_t binomialverteilt mit n Wiederholungen und den Erfolgswahrscheinlichkeiten $\pi_t = F_t$. Daraus ergibt sich für die Erwartungswerte und Varianzen der Quaderhöhen

$$\begin{aligned} E(H_t) &= \frac{1}{F_t n} E(N_t) = 1 \\ \text{und} \quad \text{Var}(H_t) &= \frac{1}{F_t^2 n^2} \text{Var}(N_t) = \frac{(1 - F_t)}{F_t n}. \end{aligned}$$

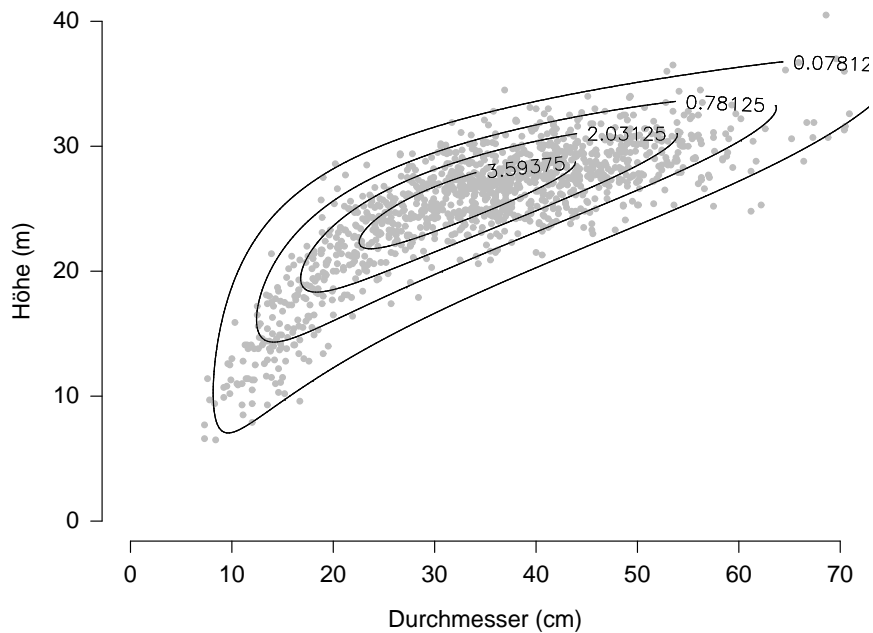


Abbildung 5.1: Beobachtungen und Höhenlinien der Dichtefunktion der S_{BB} -Verteilung.

Die erwartete Höhe eines beliebigen Quaders entspricht damit dem Funktionswert der bivariaten $U(0, 1)$ -Verteilung. Die Varianzen können verwendet werden, um das Ausmaß von Abweichungen zu beurteilen. Beispielsweise könnten

$$1 - 2\sqrt{\text{Var}(H_t)} \quad \text{und} \quad 1 + 2\sqrt{\text{Var}(H_t)}$$

als grobe Approximation der Grenzen eines kritischen Bereichs verwendet werden, innerhalb dessen die Höhe von ca. 95% der Quader liegen sollte, wenn das betrachtete Modell die Daten korrekt beschreibt. Eine andere Möglichkeit zur Beurteilung der Abweichungen ist die Verwendung der simultanen Konfidenzintervalle für die Erfolgswahrscheinlichkeiten $\pi_t = F_t$, die in Kapitel 2 beschrieben wurden und leicht auf die hier betrachtete Problemstellung übertragen werden können. Eine dritte Alternative ist die Bestimmung von diskreten Pseudo-Residuen für die beobachteten Häufigkeiten in den einzelnen Klassen. Diese Möglichkeit zur Beurteilung der dreidimensionalen Histogramme wird im Anschluss an das nachfolgende Beispiel genauer betrachtet.

Beispiel 5.2 Zur Veranschaulichung werden erneut die Daten für den Brusthöhendurchmesser und die Höhe der Buchen betrachtet, die bereits in Ab-

schnitt 2.5.1 modelliert wurden. In diesem Abschnitt wird mit Johnsons S_{BB} -Verteilung allerdings ein konkurrierendes Modells betrachtet (vgl. JOHNSON, 1949; bezüglich der Verwendung zur Modellierung des Durchmessers und der Höhe von Bäumen siehe HAFLEY, 1977). Sie basiert auf einer Transformation der bivariaten Normalverteilung und besitzt die gemeinsame Dichtefunktion

$$f(y_D, y_H) = \frac{J}{2\pi\sqrt{1-\rho^2}} \exp \left\{ \frac{z_D^2 - 2\rho z_D z_H + z_H^2}{2(1-\rho^2)} \right\}$$

mit

$$\begin{aligned} z_D &= \gamma_D + \delta_D \log \left(\frac{y_D - \xi_D}{\lambda_D + \xi_D - y_D} \right), \\ z_H &= \gamma_H + \delta_H \log \left(\frac{y_H - \xi_H}{\lambda_H + \xi_H - y_H} \right) \quad \text{und} \\ J &= \frac{\delta_D \lambda_D \delta_H \lambda_H}{(y_D - \xi_D)(\lambda_D + \xi_D - y_D)(y_H - \xi_H)(\lambda_H + \xi_H - y_H)}. \end{aligned}$$

Die Maximum-Likelihoodschätzer der Modellparameter wurden bereits von ZUCCHINI, SCHMIDT AND V. GADOW (2000) bestimmt. Ihre Ergebnisse sind in Tabelle 5.1 wiedergegeben und werden für diese Arbeit übernommen. Abbildung 5.1 zeigt die Daten und die Höhenlinien der angepassten Dichtefunktion.

Abbildung 5.2 zeigt die dreidimensionalen Histogramme der bivariaten UPR für die S_{BB} -Verteilung. Das linke Histogramm basiert auf 1.) den univariaten UPR für die Randverteilung des Brusthöhendurchmessers und 2.) den univariaten UPR der bedingten Verteilung der Baumhöhe bei gegebenem Brusthöhendurchmesser:

$$\mathbf{U}^{(1)} = (F_{Y_D}(Y_D), F_{Y_H|Y_D}(Y_H))'.$$

Das rechte Histogramm basiert auf der alternativen Definition der bivariaten UPR, d.h. auf

$$\mathbf{U}^{(2)} = (F_{Y_D|Y_H}(Y_D), F_{Y_H}(Y_H))'.$$

Beide Histogramme zeigen deutliche Strukturen, d.h. es existieren Bereiche, in denen weniger bzw. mehr Residuen beobachtet wurden, als es bei Gültigkeit der S_{BB} -Verteilung der Fall wäre. Die zweifache Standardabweichung der

$\hat{\xi}_D$	$\hat{\lambda}_D$	$\hat{\delta}_D$	$\hat{\gamma}_D$	$\hat{\xi}_H$	$\hat{\lambda}_H$	$\hat{\delta}_H$	$\hat{\gamma}_D$	$\hat{\rho}$
3.767	86.213	1.408	0.901	-974.514	1027.804	5.723	-20.580	0.843

Tabelle 5.1: ML-Schätzer der S_{BB} -Verteilung.

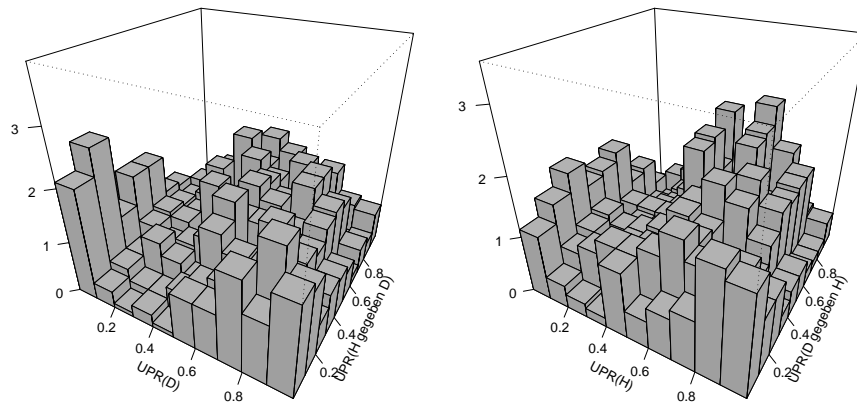


Abbildung 5.2: Dreidimensionale Histogramme der bivariaten UPR.

Quaderhöhen beträgt im hier betrachteten Beispiel in etwa 0.56. Grob approximiert sollte also bei Gültigkeit des Modells die Höhe für ca. fünf Quader außerhalb des Intervalls $[0.44, 1.56]$ liegen. Es ist hier allerdings für beide Histogramme zu erkennen, dass die Anzahl extremer Bereiche (Quader) deutlich höher ist.



Die Struktur der dreidimensionalen Histogramme liefert erste Eindrücke von der Güte der Modellanpassung. Eine detailliertere Analyse der Histogramme kann mit Hilfe von Pseudo-Residuen für die einzelnen Quader erfolgen. Wie bereits zuvor erwähnt, ist die Anzahl der Beobachtungen in den Klassen jeweils binomialverteilt mit den oben genannten Parametern. Daher lässt sich durch

$$U_t = \left[\sum_{j=0}^{N_t-1} \binom{n}{j} F_t^j (1 - F_t)^{n-j}, \sum_{j=0}^{N_t} \binom{n}{j} F_t^j (1 - F_t)^{n-j} \right] = [U_t^-, U_t^+]$$

für jeden Quader ein UPR-Intervall sowie ein entsprechendes NPR-Intervall

$$R_t = [\Phi^{-1}(U_t^-), \Phi^{-1}(U_t^+)] = [R_t^-, R_t^+]$$

definieren. Eine graphische Analyse dieser Intervalle sollte in engem Bezug zu den dreidimensionalen Histogrammen stehen, d.h. es sollte unmittelbar zu erkennen sein, welche Quader extrem sind. Dazu wird hier eine Darstellung gewählt, die den Betrachter von oben auf das Histogramm blicken lässt, und es wird durch Graustufen angedeutet, wie extrem die Höhen der Quader sind,

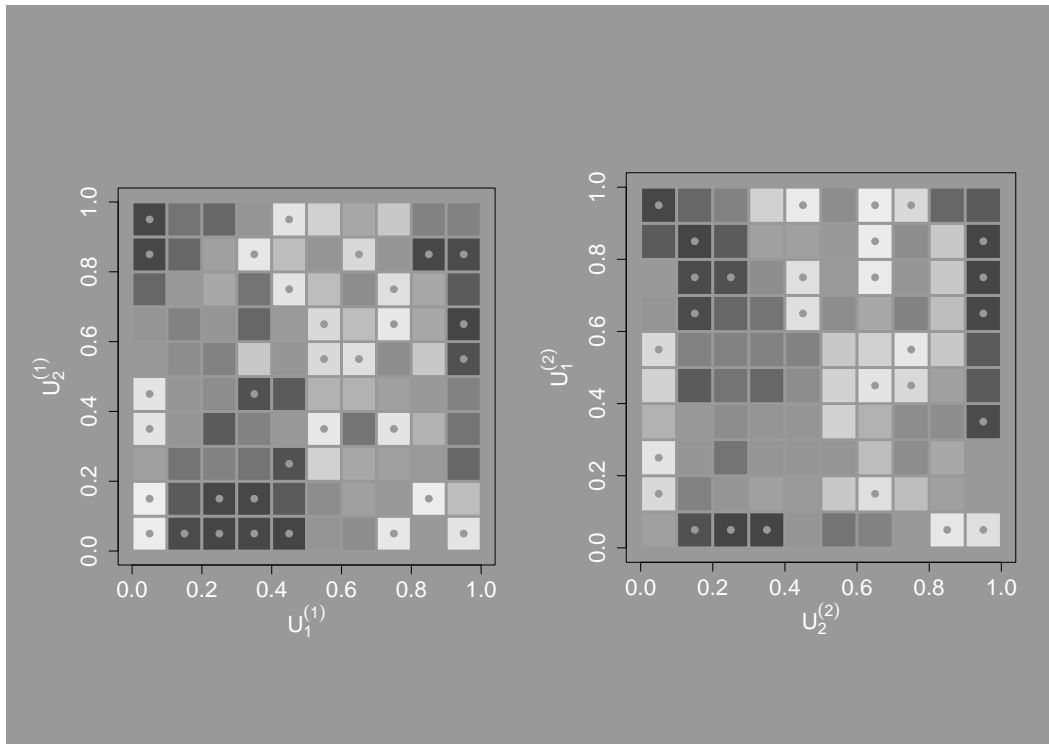


Abbildung 5.3: Durch Graustufen dargestellte Mittelpunkte der NPR-Intervalle der dreidimensionalen Histogramme.

wobei die „Quantifizierung“ der Graustufen durch die Mittelpunkte der entsprechenden NPR-Intervalle

$$R_t^m = \Phi^{-1} \left(\frac{U_t^- + U_t^+}{2} \right)$$

erfolgt. Liegt ein beobachteter Mittelpunkt nahe bei null, so entspricht die Graustufe in etwa der des Hintergrunds. Für größere bzw. kleinere Werte werden hellere bzw. dunklere Graustufen vergeben. Somit kennzeichnen helle Flächen Bereiche, in denen mehr UPR beobachtet wurden als unter Gültigkeit des Modells erwartet, während dunkle Flächen Bereiche markieren, in denen tendenziell zu wenig UPR beobachtet wurden.

Beispiel 5.3 Die Anwendung des Vorgehens auf die dreidimensionalen Histogramme des vorangegangenen Beispiels führt zu den Graphiken in Abbildung 5.3. Bei einer korrekten Beschreibung der Daten durch die S_{BB} -Verteilung sollten in den Graphiken keine systematischen Strukturen zu erkennen sein. Dies scheint hier allerdings nicht der Fall zu sein: Sowohl in der linken als auch in der rechten Graphik sind Bereiche zu erkennen, in denen mehrere aneinandergrenzende Klassen dunkler bzw. heller sind als der Hintergrund, also Bereiche, in

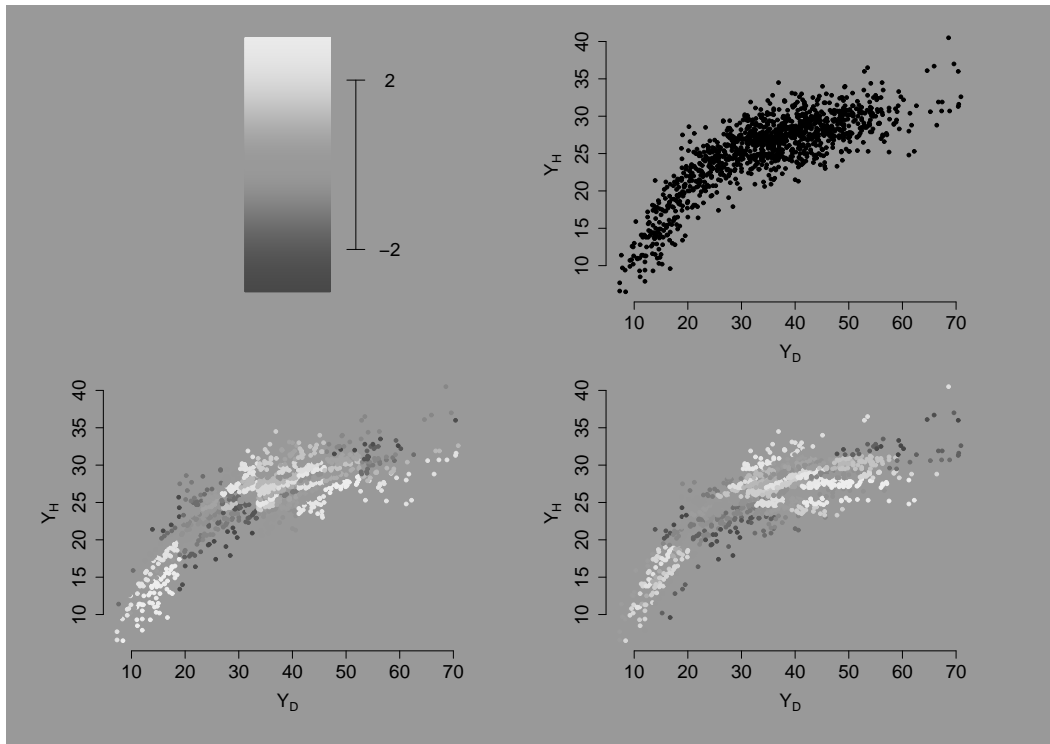


Abbildung 5.4: Übertragung der Graustufen auf die ursprünglichen Beobachtungen.

denen weniger bzw. mehr Residuen beobachtet wurden, als es unter Gültigkeit des Modells zu erwarten wäre.

Darüber hinaus sind in beiden Graphiken die Klassen mit einem Punkt versehen, für die der Mittelpunkt des NPR-Intervalls betragsmäßig größer ist als $\Phi^{-1}(0.975) \approx 1.96$. In den so gekennzeichneten Bereichen wurden demnach extrem wenige bzw. extrem viele mUPR beobachtet. Dies gilt in beiden Graphiken für etwa 30% der Klassen, was ein deutliches Zeichen für die schlechte Beschreibung der Daten durch die S_{BB} -Verteilung ist.

Aus Abbildung 5.3 ist nicht ersichtlich, welche der ursprünglichen Beobachtungen nicht adäquat durch das Modell beschrieben werden, mit anderen Worten: Es wurde bislang lediglich gezeigt, in welchen Bereichen die Verteilung der beobachteten mUPR von der bivariaten $U(0,1)$ -Verteilung abweicht. Daher ist es in einem nächsten Schritt sinnvoll, die Ergebnisse der Analyse auf die S_{BB} -Verteilung zu übertragen und somit zu zeigen, in welchen Bereichen sie keine korrekte Beschreibung der Daten liefert. Dazu kann der ursprüngliche Datensatz erneut geplottet und den einzelnen Beobachtungen die Graustufe zugeordnet werden, die der Graustufe der Klasse entspricht, in der das zur

Beobachtung gehörende bivariate UPR liegt. In den beiden unteren Graphiken der Abbildung 5.4 ist diese Darstellung für das hier betrachtete Beispiel gezeigt. Ferner sind zum Vergleich alle Beobachtungen in schwarz (oben rechts) sowie die verwendete Graustufenskala (oben links) dargestellt. Die beiden unteren Graphiken zeigen ein ähnliches Muster: Zwei größere helle Bereiche, in denen mehr Beobachtungen liegen, als es bei Gültigkeit des Modells zu erwarten wäre, werden durch einen dunkleren Bereich, in dem sich tendenziell zu wenig Beobachtungen befinden, getrennt. Betrachtet man erneut die Daten und die Höhenlinien der Dichtefunktion (Abbildung 5.1), erscheinen diese Ergebnisse sehr plausibel.

◀

5.4.2 Erfassung von extremen Beobachtungen

5.4.2.1 Identifikation extremer Beobachtungen in multivariaten Datensätzen

Wenn untersucht werden soll, ob eine multivariate Verteilung in der Lage ist, extreme Beobachtungen adäquat zu beschreiben, stellt sich zwangsläufig als erstes die Frage, was unter einer extremen Beobachtung zu verstehen ist. Für eine univariate (und unimodale) Verteilung ist es üblich, eine Beobachtung als extrem zu betrachten, wenn sie sich am äußeren Rand der Verteilung befindet, und die Überprüfung, ob eine solche Beobachtung adäquat durch das Modell beschrieben wird, kann beispielsweise mit einem QQ-Plot der NPR erfolgen.

Auch für multivariate Verteilungen kann eine Beobachtung als extrem interpretiert werden, wenn sie am Rand der Verteilung liegt. Für bivariate Beobachtungen kann die graphische Identifikation einer extremen Beobachtung erfolgen, indem die Lage der Beobachtung anhand der Höhenlinien der gemeinsamen Dichtefunktion beurteilt wird. Abbildung 5.5 zeigt die Höhenlinien der gemeinsamen Dichtefunktion einer Mischung von zwei bivariaten Normalverteilungen. Der eingezeichnete Punkt ist mit Blick auf die gemeinsame Verteilung sicherlich extrem. Bemerkenswert ist allerdings, dass er bezüglich der Randverteilungen von X und Y nicht extrem ist. Mit Blick auf die bimodale bedingte Verteilung von Y gegeben X ist der Punkt zwar extrem (er ist dem äußeren Rand einer der beiden Mischungskomponenten zuzuordnen), jedoch ist das entsprechende NPR nicht auffällig, weil sich die Beobachtung in der Nähe des Medians der bedingten Verteilung befindet. Lediglich für die bedingte Verteilung von X gegeben Y liegt der Punkt am linken Rand der Verteilung und

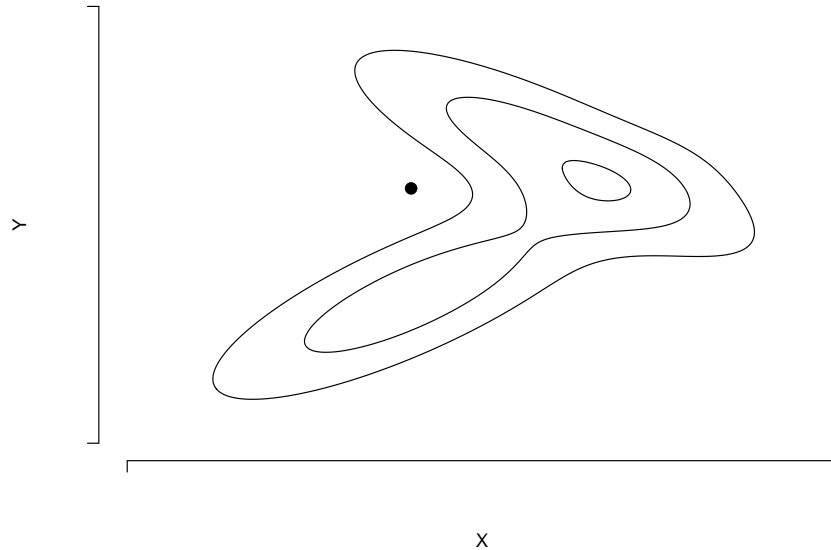


Abbildung 5.5: Mischung von zwei bivariaten Normalverteilungen und extreme Beobachtung.

liefert ein extrem kleines NPR. Dieses Beispiel macht deutlich, dass die Identifikation einer extremen Beobachtung bei multivariaten Beobachtungen eine differenzierte Betrachtung sämtlicher Randverteilungen und bedingter Verteilungen erfordert. Mit anderen Worten ist die Identifikation nur sichergestellt, wenn beide Definitionen der bivariaten NPR, d.h.

$$\mathbf{R} = (R_{1\cdot}, R_{2\cdot 1})' \quad \text{und} \quad \mathbf{R} = (R_{1\cdot 2}, R_{2\cdot})',$$

betrachtet werden.

Während für bivariate (und eventuell auch für dreidimensionale) Beobachtungen eine graphische Identifikation in der oben dargestellten Form möglich ist, sind für höherdimensionale Datensätze andere Methoden erforderlich. In direktem Zusammenhang dazu steht die Identifikation von Ausreißern in multivariaten Datensätzen, die relativ häufig in der Literatur betrachtet wird (für einen Überblick siehe BARNETT AND LEWIS, 1994, Kap. 7). Viele der beschriebenen Methoden basieren auf der Mahalanobisdistanz (siehe beispielsweise ROCKE AND WOODRUFF, 1996; ROUSSEEUW AND VAN ZOMEREN, 1990 sowie HADI, 1992), die für eine Beobachtung \mathbf{y}_i durch

$$MD_i = \sqrt{(\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}})}, \quad (5.4)$$

definiert ist, wobei $\bar{\mathbf{y}}$ bzw. \mathbf{S} den Vektor der Mittelwerte bzw. die Kovarianzmatrix der Stichprobe darstellen. MD_i ist eine standardisierte Distanz der Beobachtung \mathbf{y}_i zum Zentrum $\bar{\mathbf{y}}$ der Daten und insofern in bestimmten Situationen (wenn die Beobachtungen gemeinsam normalverteilt sind) geeignet, um zu messen, wie extrem eine multivariate Beobachtung ist.

Es gibt jedoch auch Situationen, in denen die Verwendung der Mahalanobisdistanz weniger hilfreich ist, um extreme Beobachtungen zu identifizieren. Betrachtet man beispielsweise noch einmal den Punkt aus Abbildung 5.5, so erhält man den Eindruck, dass dessen Distanz zum Zentrum der Verteilung und somit die Mahalanobisdistanz nicht besonders auffällig ist. In solchen Fällen kann anstatt der Distanz der Beobachtung selbst die Distanz des entsprechenden mNPR verwendet werden: Da die eingezeichnete Beobachtung bezüglich der bedingten Verteilung von X gegeben Y offensichtlich extrem ist, wird die Distanz von $\mathbf{R} = (R_{1|2}, R_{2|1})'$ zum Punkt $(0, 0)'$ groß. Diese Idee wird in Abschnitt 5.4.2.2 vertieft.

Die Analyse der Mahalanobisdistanzen der ursprünglichen Beobachtungen liefert ebenfalls unbefriedigende Ergebnisse, wenn die extremen Beobachtungen dicht beieinander liegen. Dieser Fall wird bei JUAN AND PRIETO (2001) diskutiert. Sie schlagen für derartige Situationen eine alternative Methode vor, die auf der Berechnung bestimmter Winkel basiert. Ihr Verfahren wird in Abschnitt 5.4.2.3 beschrieben und in leicht modifizierter Form für die Analyse der mNPR genutzt.

5.4.2.2 Distanzbasierte Analyse der mNPR

Die in diesem Abschnitt beschriebenen Methoden basieren auf der euklidischen Distanz der mNPR zum Ursprung (dem Punkt $\mathbf{0} \in \mathbb{R}^p$). Für die i -te Beobachtung ist diese durch

$$D_{\mathbf{R}i} = \sqrt{\sum_{k=1}^p R_{ki}^2} \quad \text{für } i = 1, \dots, n$$

gegeben, wobei R_{ki} das k -te Element des mNPR der Beobachtung \mathbf{y}_i darstellt. Bei Gültigkeit des Modells sind die R_{ki} für $k = 1, \dots, p$ und $i = 1, \dots, n$ unabhängig und identisch standardnormalverteilt, woraus für alle $D_{\mathbf{R}i}^2$ eine χ^2 -Verteilung mit p Freiheitsgraden folgt. Somit können, basierend auf den Distanzen, für die multivariaten Beobachtungen univariate NPR

$$R_{\mathbf{R}i} = \Phi^{-1} \left(F_{\chi_p^2}(D_{\mathbf{R}i}^2) \right) \quad \text{für } i = 1, \dots, n$$

definiert werden, die im Folgenden als NPR der Distanzen bezeichnet werden und bei Gültigkeit des Modells unabhängig und identisch $N(0, 1)$ -verteilt sind. Durch eine Analyse dieser Residuen können verschiedene Aspekte der Anpassungsgüte beurteilt werden. Beispielhaft seien drei mögliche Situationen beschrieben:

- (1) Ist die Verteilung der Residuen nach rechts verschoben, sind tendenziell zu wenig kleine und zu viele große Distanzen beobachtet worden, d.h. unter dem Modell wären mehr Beobachtungen im Zentrum und weniger Beobachtungen an den Randbereichen erwartet worden: Die gemeinsame Verteilung besitzt demnach zu viel Masse im Zentrum der Daten.
- (2) Besitzen die Residuen eine größere Streuung als unter dem Modell erwartet, so sind tendenziell zu viele kleine und zu viele große Distanzen beobachtet worden. Es liegen also einerseits mehr Beobachtungen im Zentrum der multivariaten Verteilung und andererseits mehr Beobachtungen in ihren Randbereichen als bei Gültigkeit der gemeinsamen Verteilung erwartet. Gleichbedeutend kann mit Blick auf das Modell festgestellt werden, dass die gemeinsame Verteilung zu wenig Masse im Zentrum und an den Randbereichen besitzt.
- (3) Sind einige Residuen wesentlich größer, als es unter dem Modell (und bei gegebenem Stichprobenumfang) erwartet wird, handelt es sich um extreme Beobachtungen, die nicht adäquat durch die gemeinsame Verteilung beschrieben werden.

Die Punkte (1) und (2) beziehen sich auf Aussagen zur grundlegenden Gestalt der Verteilungen und können daher als Ergänzung der in Abschnitt 5.4.1 dargestellten Methoden verstanden werden. Der Punkt (3) hingegen bezieht sich auf die hier untersuchte Fähigkeit einer multivariaten Verteilung, extreme Beobachtungen adäquat zu beschreiben. Die Betrachtung der NPR der Distanzen erfolgt an dieser Stelle, weil sie einen engen Bezug zur Mahalanobisdistanz besitzen, die (wie zuvor erläutert) häufig zur Identifikation von extremen Beobachtungen in multivariaten Datensätzen verwendet wird. Genauer wird dieser Zusammenhang im Rahmen des nachfolgenden Beispiels beschrieben.

Beispiel 5.4 Die in Tabelle 5.2 dargestellten Daten beschreiben vier verschiedenen Messungen für die Festigkeit von 30 Brettern. Bei der Messung von Y_1 wurde ein Schlag auf die Bretter ausgeübt, zur Messung von Y_2 wurden die Bretter vibriert und die Beobachtungen für Y_3 sowie Y_4 stammen von statischen

i	y_1	y_2	y_3	y_4	MD^2	i	y_1	y_2	y_3	y_4	MD^2
1	1889	1651	1561	1778	0.60	16	1954	2149	1180	1281	16.85
2	2403	2048	2087	2197	5.48	17	1325	1170	1002	1176	3.50
3	2119	1700	1815	2222	7.62	18	1419	1371	1252	1308	3.99
4	1645	1627	1110	1533	5.21	19	1828	1634	1602	1755	1.36
5	1976	1916	1614	1883	1.40	20	1725	1594	1313	1646	1.46
6	1712	1712	1439	1546	2.22	21	2276	2189	1547	2111	9.90
7	1943	1685	1271	1671	4.99	22	1899	1614	1422	1477	5.06
8	2104	1820	1717	1874	1.49	23	1633	1513	1290	1516	0.80
9	2983	2794	2412	2581	12.26	24	2061	1867	1646	2037	2.54
10	1745	1600	1384	1508	0.77	25	1856	1493	1356	1533	4.58
11	1710	1591	1518	1667	1.93	26	1727	1412	1238	1469	3.40
12	2046	1907	1627	1898	0.46	27	2168	1896	1701	1834	2.38
13	1840	1841	1595	1741	2.70	28	1655	1675	1414	1597	3.00
14	1867	1685	1493	1678	0.13	29	2326	2301	2065	2234	6.28
15	1859	1649	1389	1714	1.08	30	1490	1382	1214	1284	2.58

Tabelle 5.2: Vier verschiedene Messungen bezüglich der Festigkeit von 30 Brettern und quadrierte Mahalanobisdistanzen.

Messungen (vgl. JOHNSON AND WICHERN, 1998, S. 198). Ferner sind in der Tabelle die quadrierten Mahalanobisdistanzen der Beobachtungen angegeben. Zur Beschreibung der Daten wird eine vierdimensionale Normalverteilung mit der gemeinsamen Dichtefunktion

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -(\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) / 2 \right\} \quad -\infty < y_1, \dots, y_4 < \infty$$

verwendet. Die aus den Beobachtungen berechneten Schätzer für den Erwartungswertvektor und die Kovarianzmatrix sind

$$\hat{\boldsymbol{\mu}} = (1906.10, 1749.53, 1509.13, 1724.97)'$$

$$\text{und } \hat{\Sigma} = \begin{pmatrix} 105616.30 & 94613.53 & 87289.71 & 94230.73 \\ 94613.53 & 101510.12 & 76137.10 & 81064.36 \\ 87289.71 & 76137.10 & 91917.09 & 90352.38 \\ 94230.73 & 81064.36 & 90352.38 & 104227.96 \end{pmatrix}.$$

Um Ausreißer unter dem Modell zu identifizieren, untersuchen JOHNSON AND WICHERN (1998, S. 202–204) die Mahalanobisdistanzen, die bei Gültigkeit des

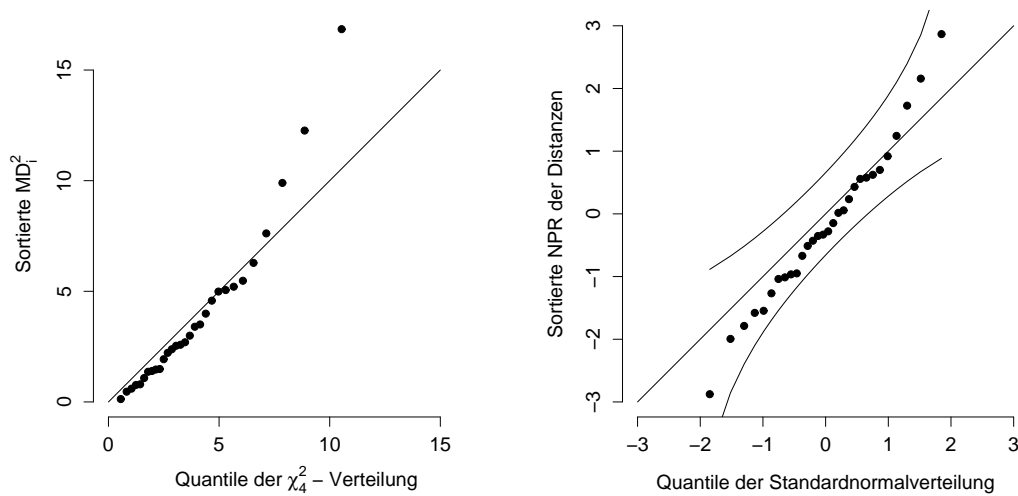


Abbildung 5.6: QQ-Plots der quadrierten Mahalanobisdistanzen (links) und der NPR der Distanzen (rechts).

Modells χ^2 -verteilt sind mit vier Freiheitsgraden. Sie verwenden dazu einen QQ-Plot, in dem die beobachteten Distanzen gegen die entsprechenden Quantile der χ^2 -Verteilung geplottet sind (siehe Abbildung 5.6, linke Graphik) und kommen zu dem Ergebnis: „*The two specimens (9 and 16) with large squared distances stand out as clearly different from the rest of the pattern . . .*“.

Bevor diese Interpretation genauer betrachtet wird, sei an dieser Stelle bemerkt, dass die Mahalanobisdistanz der Beobachtungen und die euklidische Distanz der mNPR zum Ursprung identisch sind, wenn das zu überprüfende Modell wie im hier betrachteten Beispiel eine multivariate Normalverteilung ist, und zwar unabhängig von der Reihenfolge bei der Berechnung der einzelnen Elemente des mNPR. Damit sollte eine Analyse der NPR der Distanzen zu den gleichen Ergebnissen führen wie eine Analyse der Mahalanobisdistanzen.

Dass dies nicht der Fall ist, wird bei der Betrachtung des entsprechenden QQ-Plots deutlich (siehe Abbildung 5.6, rechte Graphik). Es ist zu sehen, dass die Punkte die simultanen 95%-Konfidenzbänder nicht verlassen und insbesondere erscheinen die NPR der Distanzen für die neunte und sechzehnte Beobachtung nicht auffällig extrem. Vielmehr gliedern sie sich in die gegen den Uhrzeigersinn gedrehte Struktur des QQ-Plots ein, d.h. die Streuung der NPR der Distanzen ist größer als unter dem Modell erwartet, bzw. die Verteilung besitzt sowohl im Zentrum als auch an den Randbereichen zu wenig Masse.

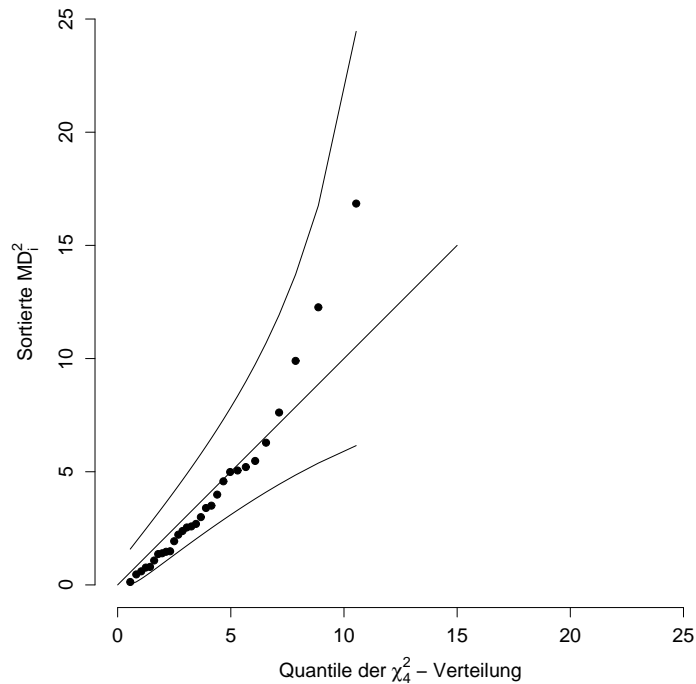


Abbildung 5.7: QQ-Plot der quadrierten Mahalanobisdistanzen mit simultanem 95%-Konfidenzband.

Neben dieser Aussage bezüglich der Modellanpassung wird anhand des hier betrachteten Beispiels ein allgemeiner Vorteil der NPR besonders deutlich: ihre Skalierung auf Basis der Standardnormalverteilung. Weil diese symmetrisch ist und man in der Statistik mit ihrem Umgang sehr vertraut ist, lässt sich ein entsprechender QQ-Plot gut interpretieren. Die χ^2 -Verteilung mit vier Freiheitsgraden hingegen ist schief, so dass die Analyse der Abweichungen von der Winkelhalbierenden leicht zu Fehlinterpretationen führen kann. Abbildung 5.7 zeigt noch einmal den QQ-Plot der quadrierten Mahalanobisdistanzen. Zusätzlich sind hier die simultanen 95%-Konfidenzbänder eingezeichnet, die, wie in Abschnitt 3.3.1 beschrieben, für eine beliebige Referenzverteilung erzeugt werden können. Verdeutlicht man sich, dass die Wahrscheinlichkeit für das Verlassen der Bänder für jeden Punkt identisch ist, wird auch hier ersichtlich, dass die Mahalanobisdistanzen der neunten und sechzehnten Beobachtung nicht extrem groß sind.

◀

Weil die Mahalanobisdistanz und die Distanz der mNPR zum Ursprung bei

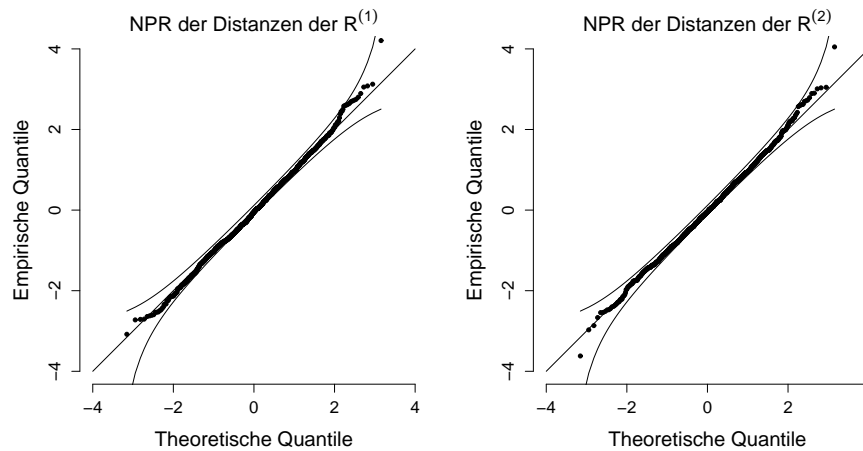


Abbildung 5.8: QQ-Plots der NPR der Distanzen bei alternativer Definition der bivariaten NPR.

der Analyse multivariater Normalverteilungen identisch sind, können die quadrierten Elemente des mNPR als Beiträge zur Mahalanobisdistanz verstanden werden: Jedes Element R_{ki}^2 misst, wie extrem die Beobachtung \mathbf{y}_i bezüglich der k -ten Zufallsvariable ist, wobei letztere je nach Definition des mNPR die Randverteilung von Y_k oder eine bedingte Verteilung von Y_k bei einer gegebenen Menge anderer Zufallsvariablen ist. Diese Interpretation der Elemente eines mNPR gilt auch wenn das betrachtete Modell *keine* multivariate Normalverteilung ist, so dass die Distanzen der mNPR auch in diesen Fällen geeignet sind, um extreme Beobachtungen in multivariaten Datensätzen zu identifizieren. Dies ist bemerkenswert, weil die Mahalanobisdistanz in solchen Situationen zu unplausiblen Ergebnissen bei der Identifikation extremer Beobachtungen führen kann. Es sei in diesem Zusammenhang an das einführende Beispiel erinnert (siehe Abbildung 5.5).

Beispiel 5.5 Zur gemeinsamen Modellierung des Brusthöhendurchmessers und der Höhe der 1242 Buchen wurde in Kapitel 2 die Mischung von zwei bivariaten Normalverteilungen verwendet. Für dieses Modell sind beide möglichen Definitionen der mNPR, d.h.

$$\begin{aligned} \mathbf{R}^{(1)} &= ((\Phi^{-1}F_{Y_D}(Y_D), \Phi^{-1}F_{Y_H|Y_D}(Y_H)))' \\ \text{und } \mathbf{R}^{(2)} &= ((\Phi^{-1}F_{Y_D|Y_H}(Y_D), \Phi^{-1}F_{Y_H}(Y_H)))' \end{aligned}$$

zu betrachten, weil hier keine multivariate Normalverteilung untersucht wird

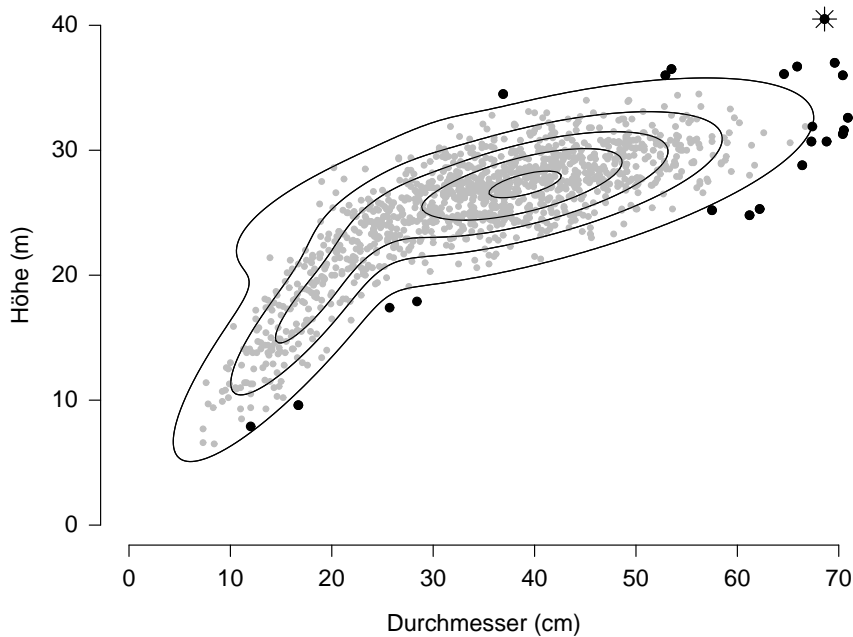


Abbildung 5.9: Extreme Beobachtungen bei Verwendung der NPR der Distanzen.

und somit die Distanzen der bivariaten mNPR zum Ursprung je nach Definition unterschiedlich sind.

Berechnet man nach beiden Definitionen die NPR der Distanzen und stellt diese jeweils in einem QQ-Plot dar (Abbildung 5.8), ist zu erkennen, dass das hier betrachtete Modell eine gute Beschreibung der Daten liefert. Für beide Definitionen liegen die Punkte dicht an der Winkelhalbierenden und verlassen die simultanen 95%-Konfidenzbänder nicht. Allerdings ist in beiden QQ-Plots am jeweils rechten Rand ein Punkt zu sehen, der sich von der grundlegenden Struktur der Plots abhebt. Es handelt sich hierbei in beiden Fällen um die 308-te Beobachtung, den höchsten der 1242 Bäume. Dieser Baum repräsentiert eine extreme Beobachtung, die nicht adäquat durch das Modell beschrieben wird. Dass dies tatsächlich der Fall ist, zeigt Abbildung 5.9, in der erneut die Höhenlinien der gemeinsamen Dichtefunktion zu sehen sind und in der die Beobachtung durch einen Stern markiert ist.

Zuvor wurde erwähnt, dass die Distanzen der mNPR im Gegensatz zur Mahalanobisdistanz auch bei der Betrachtung von Modellen, die keine multivariaten Normalverteilungen sind, zur Identifikation von extremen Beobachtungen verwendet werden kann. Dazu wurden für beide Definitionen der bivariaten NPR

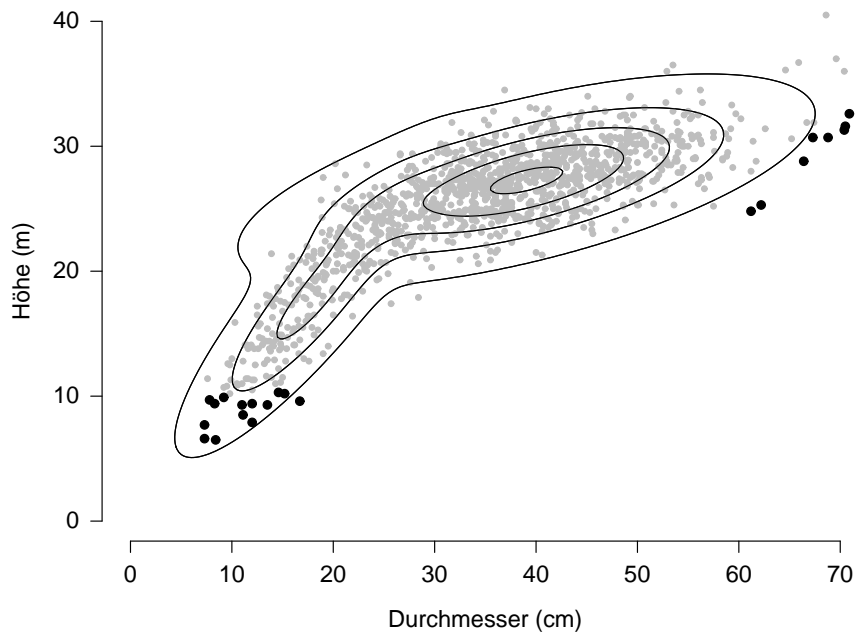


Abbildung 5.10: Extreme Beobachtungen bei Verwendung der Mahalanobisdistanz.

($\mathbf{R}^{(1)}$ und $\mathbf{R}^{(2)}$) jeweils die Beobachtungen bestimmt, die zu den 20 größten NPR der Distanzen führen. Da diese für beide Definitionen in den meisten Fällen identisch sind, wurden auf diese Weise insgesamt 22 Beobachtungen identifiziert und in Abbildung 5.9 schwarz markiert. Eine Beurteilung dieser Beobachtungen anhand der Höhenlinien zeigt, dass dieses Vorgehen sehr erfolgreich ist, um die extremen Beobachtungen unter dem Modell zu bestimmen. Zum Vergleich sind in Abbildung 5.10 die Beobachtungen mit den 22 größten Mahalanobisdistanzen markiert. Letztere sind im hier betrachteten Beispiel nicht geeignet, die extremen Beobachtungen unter dem Modell zu identifizieren, weil sie nur vor dem Hintergrund einer multivariaten Normalverteilung sinnvoll zu interpretieren sind.



5.4.2.3 Winkelbasierte Analyse der multivariaten Pseudo-Residuen

Die in diesem Abschnitt dargestellte Methode wurde von JUAN AND PRIETO (2001) verwendet, um konzentrierte Ausreißer in multivariaten Beobachtungen sogenannter isotropischer Verteilungen (zum Beispiel einer multivariaten

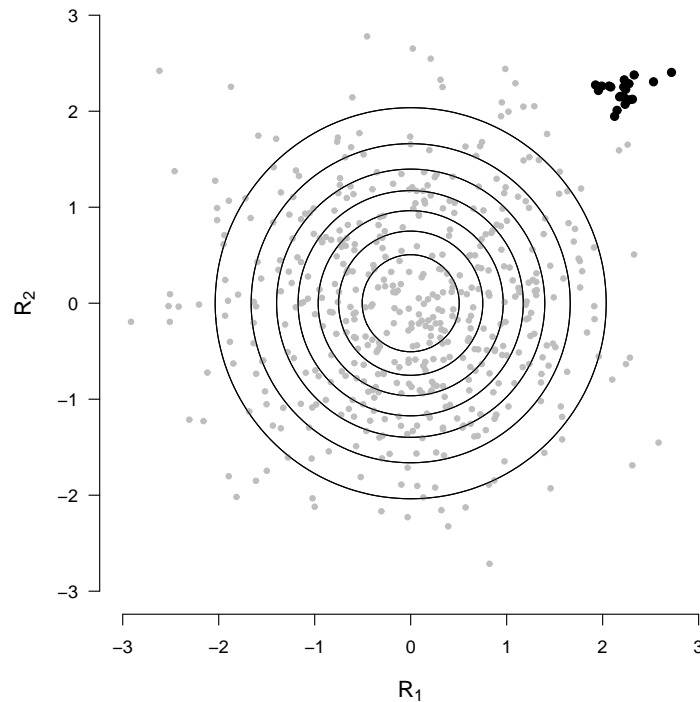


Abbildung 5.11: Gruppe konzentrierter bivariater NPR.

Normalverteilung) zu identifizieren. (Zu den isotropischen Verteilungen siehe EATON, 1983.) Unter dem Begriff ‚konzentrierte Ausreißer‘ wird dabei eine Gruppe von Beobachtungen verstanden, die sich von den übrigen Beobachtungen (als Gruppe) zu unterscheiden scheint. Wird die Methode zur Analyse der mNPR eingesetzt, wird demnach nach einer Konzentration von beobachteten mNPR aus \mathbb{R}^p gesucht, die bei unabhängig und identisch standardnormalverteilten Residuen (im Folgenden als p -dimensionale Standardnormalverteilung bezeichnet) nicht erwartet wird. Für den bivariaten Fall ist die Situation in Abbildung 5.11 verdeutlicht. Sie zeigt die Höhenlinien einer bivariaten Standardnormalverteilung sowie 500 zufällige Beobachtungen (graue Punkte) aus dieser. Ferner sind 20 schwarze Punkte eingezeichnet, die eine Gruppe konzentrierter bivariater NPR bilden. Findet man eine solche Situation bei der Analyse der Residuen vor, gibt es einen Bereich, in dem das zu überprüfende Modell zu wenig Masse besitzt: Mit der Konzentration der Residuen in der bivariaten Standardnormalverteilung ist eine Konzentration von ursprünglichen Beobachtungen verbunden, die unter dem zu überprüfenden Modell nicht zu erwarten ist.

Die Mahalanobisdistanz ist zur Identifikation von konzentrierten Ausreißern wenig geeignet, weil die entsprechenden Beobachtungen die Schätzer des Erwartungswertvektors und der Kovarianzmatrix stark beeinflussen. Aus diesem Grund schlagen JUAN AND PRIETO (2001) eine Methode zur Identifikation der Beobachtungen vor, die auf der Berechnung bestimmter Winkel basiert. Ihr Vorgehen wird im Folgenden zunächst allgemein erläutert und anschließend in leicht modifizierter Form anhand eines Beispiels illustriert.

Die von JUAN AND PRIETO (2001) beschriebene Methode kann in drei Schritte unterteilt werden:

- (1) Standardisierung der Beobachtungen
- (2) Projektion der standardisierten Beobachtungen
- (3) Wahl einer Referenzrichtung und Berechnung der Winkel

Standardisierung der Beobachtungen Der erste Schritt ist die Standardisierung der beobachteten mNPR \mathbf{r}_i durch

$$\mathbf{r}_i^* = \mathbf{S}^{-1/2}(\mathbf{r}_i - \bar{\mathbf{r}}) \quad \text{für } i = 1, \dots, n,$$

wobei $\bar{\mathbf{r}}$ den Mittelwertvektor und \mathbf{S} die empirische Kovarianzmatrix der mNPR darstellen.

Handelt es sich bei der zu überprüfenden Verteilung um eine multivariate Normalverteilung, sind die Elemente von $\bar{\mathbf{r}}$ null, und $\mathbf{S}^{-1/2}$ ist eine Einheitsmatrix. In diesem Fall kann auf die Standardisierung verzichtet werden, da $\mathbf{r}^* = \mathbf{r}$ ist. Ist das zu überprüfende Modell keine multivariate Normalverteilung, so ist die Standardisierung durchzuführen, weil dies bei Vorliegen einer Gruppe konzentrierter mNPR zu einer zuverlässigeren Identifikation dieser Gruppe beiträgt.

Projektion der standardisierten Beobachtungen Die (zur Berechnung der Winkel benötigten) Projektionen erhält man, indem die standardisierten Residuen durch ihre euklidische Distanz zum Ursprung (durch ihre Länge) dividiert werden:

$$\mathbf{z}_i = \frac{\mathbf{r}_i^*}{\|\mathbf{r}_i^*\|}.$$

Für zwei Dimensionen ist die Transformation in Abbildung 5.12 dargestellt. Die linke Seite zeigt einen Scatterplot von 200 zufälligen Beobachtungen standardisierter bivariater NPR. Werden diese durch ihre Länge dividiert, erhält

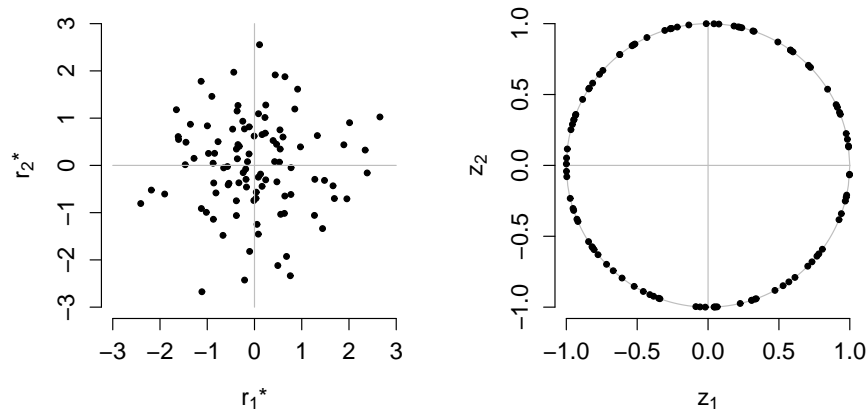


Abbildung 5.12: Scatterplot standardisierter bivariater NPR (links) und Projektionen auf den Einheitskreis (rechts).

man Projektionen, die (bei Gültigkeit des zu überprüfenden Modells) auf dem Einheitskreis gleichverteilt sind. Allgemein (auch für $p > 2$) liegt \mathbf{z} in der Sphäre $S_{p-1} = \{\mathbf{z} \in \mathbb{R}^p : \|\mathbf{z}\| = 1\}$, und bei Gültigkeit des zu überprüfenden Modells ist \mathbf{Z} gleichverteilt in S_{p-1} (EATON, 1983).

Wahl einer Referenzrichtung und Berechnung der Winkel Die projizierten Beobachtungen werden in einem nächsten Schritt verwendet, um die Winkel zwischen den Strecken vom Ursprung zu den Projektionen \mathbf{z}_i und einer Referenzrichtung \mathbf{z}_0 zu berechnen. Für die i -te Beobachtung erhält man diesen Winkel durch

$$w_i = \arccos(\mathbf{z}_0^T \mathbf{z}_i) \quad \text{mit } \|\mathbf{z}_0\| = 1.$$

Bei einer korrekten Beschreibung der Daten durch die zu überprüfende multivariate Verteilung, d.h. bei einer Gleichverteilung von \mathbf{Z} in S_{p-1} , ist die Verteilungsfunktion dieser Winkel für eine beliebige Referenzrichtung \mathbf{z}_0 durch

$$F_W(w) = \begin{cases} 0.5 \cdot F_{\text{Beta}}(\sin^2 w; (p-1)/2, 1/2) & 0 \leq w \leq \pi/2 \\ 1 - 0.5 \cdot F_{\text{Beta}}(\sin^2 w; (p-1)/2, 1/2) & \pi/2 < w \leq \pi \end{cases}$$

gegeben, wobei F_{Beta} die Verteilungsfunktion einer Beta-Verteilung mit den angegebenen Parametern darstellt. Es sei dazu bemerkt, dass bei der hier betrachteten multivariaten Problemstellung, der Winkel durch die Orientierung der Strecken festgelegt ist und daher im Intervall $[0, \pi]$ liegt (vgl. beispielsweise OBERLÄNDER, 1979, S. 588). Abbildung 5.13 zeigt die Verteilungsfunktion für

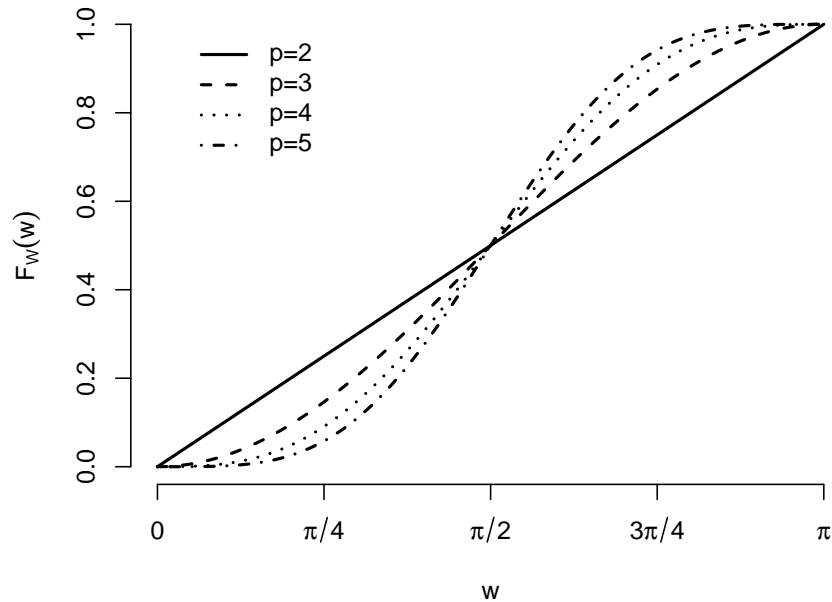


Abbildung 5.13: Verteilungsfunktion der Winkel für unterschiedliche Werte von p .

verschiedene Werte von p . Für zwei Dimensionen erhält man die Verteilungsfunktion einer Rechteckverteilung über dem Intervall $[0, \pi]$. Für $p > 2$ ist die Verteilung symmetrisch um $\pi/2$, wobei die Wahrscheinlichkeit für Winkel im mittleren Bereich der Verteilung größer ist als für Winkel an den Randbereichen.

Bevor die Winkel explizit berechnet werden können, ist die Referenzrichtung \mathbf{z}_0 zu bestimmen. Diese sollte in die Richtung der Gruppe der konzentrierten Projektionen zeigen, weil dadurch eine zuverlässigere Identifikation der konzentrierten mNPR ermöglicht wird. Verdeutlicht ist dies in Abbildung 5.14. Oben links ist ein Scatterplot von 35 Beobachtungen $(r_{1i}, r_{2i})'$ bivariater NPR zu sehen. Die ersten 30 Beobachtungen ($i = 1, \dots, 30$) stammen aus einer bivariaten Standardnormalverteilung, weitere 5 Beobachtungen ($i = 31, \dots, 35$) aus einer bivariaten Normalverteilung mit den Parametern $\mu_{R_1} = 0$, $\mu_{R_2} = 2$, $\sigma_{R_1}^2 = 0.01$, $\sigma_{R_2}^2 = 0.01$ und $\rho = 0$. Letztere bilden eine Gruppe konzentrierter bivariater NPR, die sich nördlich des Ursprungs befindet. Die Projektionen der standardisierten Residuen auf den Einheitskreis sind oben rechts dargestellt.

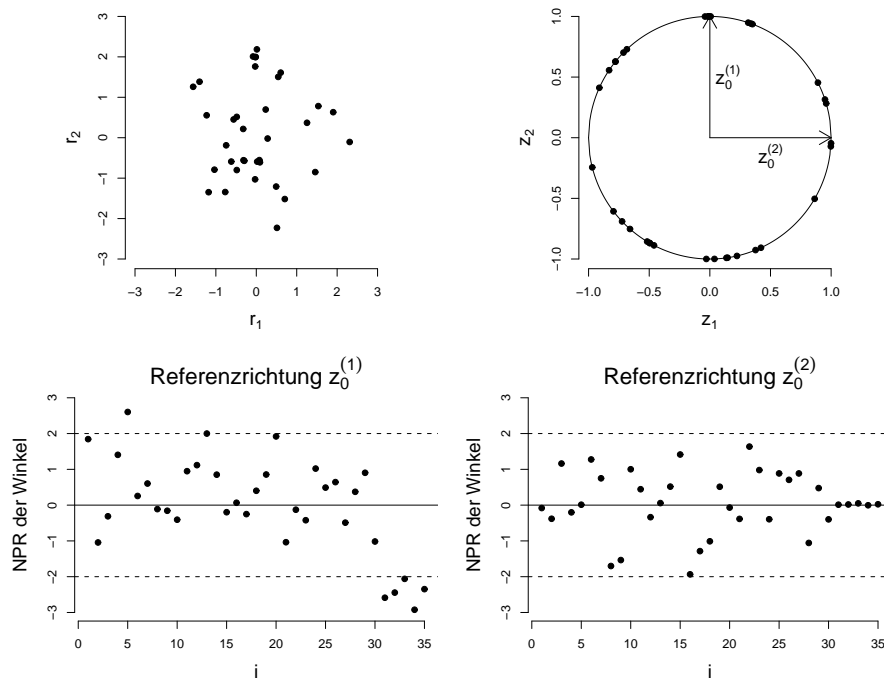


Abbildung 5.14: Scatterplot bivariater NPR mit einer Konzentration nördlich des Ursprungs (oben links), Projektionen der standardisierten Residuen (oben rechts) und Indexplots der NPR der Winkel (unten).

Ferner sind in diese Graphik zwei potenzielle Referenzrichtungen $\mathbf{z}_0^{(1)}$ und $\mathbf{z}_0^{(2)}$ eingezeichnet: Die Referenzrichtung $\mathbf{z}_0^{(1)}$ zeigt in die Richtung der konzentrierten Projektionen, die alternative Referenzrichtung zeigt nach Osten (ist also orthogonal zu $\mathbf{z}_0^{(1)}$). Die Winkel für die Beobachtungen aus der Gruppe der konzentrierten Projektionen betragen daher bei Verwendung von $\mathbf{z}_0^{(1)}$ in etwa null und bei Verwendung von $\mathbf{z}_0^{(2)}$ in etwa $\pi/2$. Dies kommt in den (univariaten) NPR der Winkel

$$r_i^W = \Phi^{-1}(F_W(w_i)) \quad \text{für } i = 1, \dots, 35$$

zum Ausdruck, deren Indexplots in den beiden unteren Graphiken zu sehen sind. Für den linken Plot wurde $\mathbf{z}_0^{(1)}$ als Referenzrichtung verwendet. Hier ist zu erkennen, dass die NPR der Winkel für $i = 31, \dots, 35$ sehr klein sind und sich als Gruppe deutlich von den übrigen NPR unterscheiden. Zeigt \mathbf{z}_0 also in Richtung der konzentrierten Projektionen, so ist anhand der NPR der Winkel die Gruppe der konzentrierten Beobachtungen gut zu identifizieren. Auch bei Verwendung von $\mathbf{z}_0^{(2)}$ bilden die NPR der Winkel für $i = 31, \dots, 35$ eine Gruppe. Sie sind in diesem Fall in etwa null (siehe Indexplot unten rechts).

Die Unterschiede zu den NPR der Winkel der übrigen Beobachtungen werden hier allerdings nicht betont. Daher kann die Gruppe der konzentrierten Beobachtungen nicht von den übrigen Beobachtungen abgegrenzt werden.

In praktischen Situationen ist die Richtung, in der sich die konzentrierten Projektionen befinden, nicht bekannt und muss daher aus den Daten geschätzt werden. Diese Schätzung kann durch die Verwendung einer Funktion erfolgen, die in Abhängigkeit von der Referenzrichtung ein Kriterium misst, das die Abweichung der empirischen Verteilung der Winkel von der Verteilung der Winkel unter dem Modell beschreibt. Diese Abweichung wird groß, wenn die Referenzrichtung auf die Gruppe der konzentrierten Projektionen zeigt, d.h. durch die Maximierung der Funktion erhält man einen Schätzer für die geeignete Referenzrichtung. Weil die Wahl der Funktion bzw. des Abweichungskriteriums einen Einfluss auf den Erfolg der Suche nach den konzentrierten Projektionen besitzt, werden im Folgenden drei mögliche Funktionen dargestellt und ihre Eigenschaften an einem Beispiel illustriert.

JUAN AND PRIETO (2001) verwenden die Funktion

$$f(\mathbf{z}) = \sum_{i=1}^n (v_{(i)} - \cos(q_i))^2$$

mit $v_i = \mathbf{z}^T \mathbf{z}_i$ für $i = 1, \dots, n$
 und $q_i = F_W^{-1} \left(\frac{n - i + 0.5}{n} \right)$

zur Schätzung der Referenzrichtung, wobei die $v_{(i)}$ den i -ten Wert der sortierten v_i darstellt. Sie messen auf diese Weise die Abweichung der empirischen Verteilung der Kosinus von der bei Gültigkeit des zu überprüfenden Modells vorliegenden Verteilung der Kosinus (letztere ergibt sich aus der Verteilung der Winkel unter dem Modell). Durch die Lösung des kontinuierlichen Optimierungsproblems

$$f(\mathbf{z}) \rightarrow \max_{\mathbf{z}} \quad \text{mit Nebenbedingung} \quad \|\mathbf{z}\| = 1$$

erhält man die Referenzrichtung, d.h. die Richtung \mathbf{z}_0 , die $f(\mathbf{z})$ maximiert, wird als Referenzrichtung verwendet. Da die Funktion viele lokale Extrema besitzt, ist ein geeigneter Startwert für den Maximierungsalgorithmus zu bestimmen. Dazu werden die Funktionswerte für alle beobachteten Projektionen

$$f(\mathbf{z}_i) \quad \text{für } i = 1, \dots, n$$

bestimmt, und die Projektion, die den größten Funktionswert liefert, wird als Startwert verwendet.

Eine weitere Funktion, die weniger lokale Maxima besitzt und ebenfalls zur Schätzung der Referenzrichtung verwendet werden kann, ist

$$g(\mathbf{z}) = \sum_{i=1}^n (u_i^W(\mathbf{z}) - 0.5)^2,$$

wobei $u_i^W(\mathbf{z})$ das UPR für den Winkel der i -ten Beobachtung bei Verwendung der Referenzrichtung \mathbf{z} darstellt. Der Ausdruck $(u_i(\mathbf{z}) - 0.5)^2$ ist ein Maß für die Abweichung des Winkels der i -ten Beobachtung vom Median der bei Gültigkeit des Modells vorliegenden Verteilung der Winkel. Die Funktion g misst demnach die Abweichung der empirischen Verteilung der Winkel von der Verteilung der Winkel unter dem Modell direkt (und nicht über die entsprechenden Kosinus).

Eine kompliziertere Funktion, die für das nachfolgende Beispiel allerdings bessere Ergebnisse bei der Identifikation der Referenzrichtung erzielt, ist

$$h(\mathbf{z}) = \sum_{i=2}^{n-1} \left(\Phi^{-1} \left(F_{\text{Beta}} \left(u_{(k)}^W(\mathbf{z}) \right) \right) \right)^2 = \sum_{i=2}^{n-1} \left(r_{(k)}^{\text{qq}}(\mathbf{w}) \right)^2.$$

Die Abweichung der empirischen Verteilung der Winkel von der bei Gültigkeit des Modells vorliegenden Verteilung wird hier durch die Summe der quadrierten NPR des QQ-Plots der NPR der Winkel quantifiziert (vgl. Abschnitt 3.3.1). Die Idee, nicht alle Residuen des QQ-Plots zu verwenden, d.h. den kleinsten und den größten Winkel zu vernachlässigen, basiert auf der Tatsache, dass die lokalen Maxima der Funktion entstehen, wenn die Richtung \mathbf{z} und die Richtung eines *beobachteten* Punkts \mathbf{z}_i annähernd übereinstimmen: Für \mathbf{z}_i wird bei gleicher bzw. entgegengesetzter Orientierung der Referenzrichtung ein sehr kleiner bzw. sehr großer Winkel berechnet, ohne dass dies zwangsläufig eine Konzentration impliziert. Dies ist nur dann der Fall, wenn mindestens zwei sehr kleine bzw. große Winkel identifiziert werden. Es erscheint daher plausibel, den größten und kleinsten Winkel zu vernachlässigen und somit das Problem der lokalen Maxima zu entschärfen.

Ein Vergleich der drei Funktionen für $p = 2$ ist in Abbildung 5.15 dargestellt. Links oben sind äquidistante Projektionen zu sehen, die eine ideale Gleichverteilung auf dem Einheitskreis charakterisieren. Zusätzlich sind vier eng beieinander liegende Projektionen im Bereich $(z_1, z_2)' = (0, 1)'$ eingezeichnet, die eine Konzentration von Projektionen in diesem Bereich repräsentiert. Die geeignete Referenzrichtung $\mathbf{z}_0 = (0, 1)'$ ist in diesem Fall bekannt. Sie maximiert alle hier betrachteten Funktionen, deren Verlauf bei Abweichungen von \mathbf{z}_0 in den drei weiteren Graphiken gezeigt ist. Es ist deutlich zu erkennen, dass die Funktionen g und h für diesen Datensatz bessere Maximierungseigenschaften besitzen,

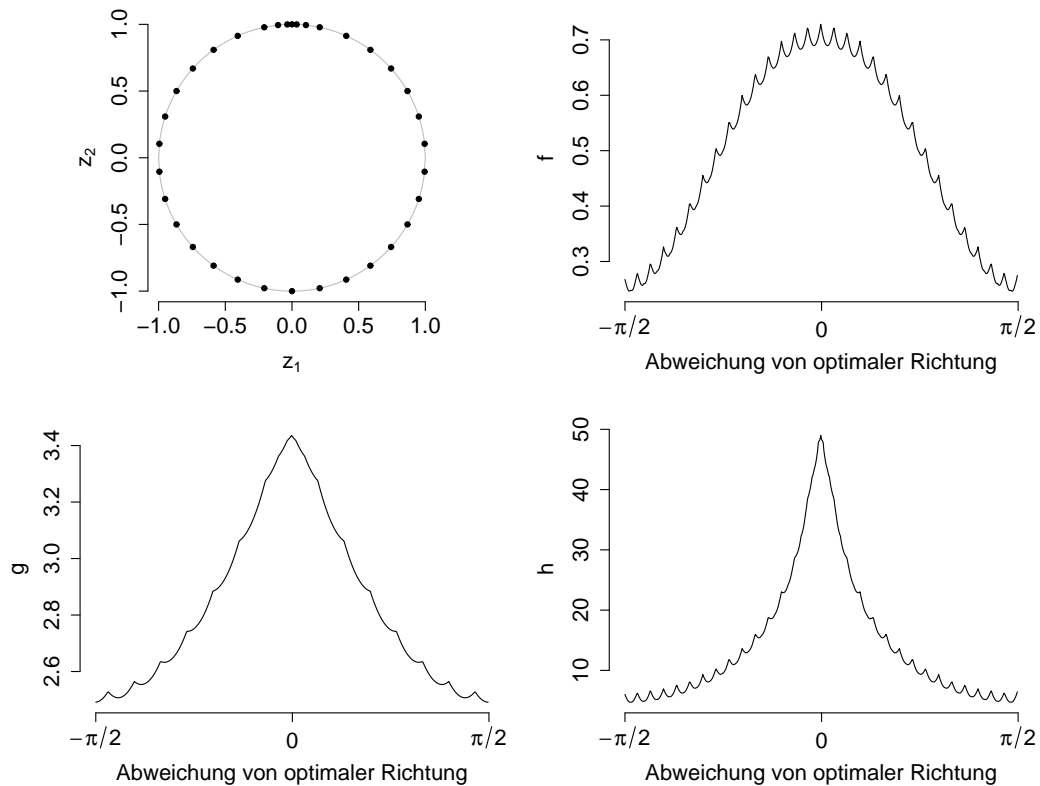


Abbildung 5.15: Äquidistante Projektionen und Konzentration von Projektionen (oben links) sowie Verlauf der Funktionen f (oben rechts), g (unten links) und h (unten rechts) in Abhängigkeit von der Referenzrichtung.

da sie zum einen weniger lokale Maxima zeigen und zum anderen einen steileren Verlauf im Bereich des globalen Maximums aufweisen. Allerdings besitzen auch diese Funktionen lokale Maxima, d.h. es ist auch hier notwendig, einen geeigneten Startwert für die Maximierung zu bestimmen. Dieser kann analog zum oben beschriebenen Vorgehen ermittelt werden.

Beispiel 5.6 Zur Verdeutlichung des Vorgehens werden erneut die Daten zur Festigkeit der Bretter (siehe Tabelle 5.2) betrachtet, wobei auch hier eine vierdimensionale Normalverteilung das zu überprüfende Modell sei.

Die Definition der mNPR kann in diesem Beispiel je nach gewählter Reihenfolge bei der Berechnung der Elemente der mNPR auf $4! = 24$ verschiedene Arten erfolgen. Die Reihenfolge hat allerdings keinen Einfluss auf die empirische Verteilung der Winkel, wenn die zu überprüfende Verteilung eine multivariate Normalverteilung ist, so dass die Überprüfung für eine beliebig ausgewählte

Definition ausreichend ist. Um dies zu zeigen, seien die alternativen Definitionen

$$\begin{aligned} \mathbf{R}^{(1)} &= (R_{1|\cdot}, R_{2|1})' = (\Phi^{-1}(F_{Y_1}(Y_1)), \Phi^{-1}(F_{Y_2|Y_1}(Y_2)))' \\ \text{und } \mathbf{R}^{(2)} &= (R_{1|2}, R_{2|\cdot})' = (\Phi^{-1}(F_{Y_1|Y_2}(Y_1)), \Phi^{-1}(F_{Y_2}(Y_2)))' \end{aligned}$$

der mNPR im Fall einer bivariaten Problemstellung betrachtet. In diesem Fall erhält man für eine Beobachtung $(Y_1, Y_2) = (y_1, y_2)$ die alternativen Residuen

$$\begin{aligned} \mathbf{r}^{(1)} &= (r_{1|\cdot}, r_{2|1})' = \left(\frac{y_1 - \hat{\mu}_1}{\hat{\sigma}_1}, \frac{y_2 - \hat{\mu}_{2|1}}{\hat{\sigma}_{2|1}} \right)' \\ \text{und } \mathbf{r}^{(2)} &= (r_{1|2}, r_{2|\cdot})' = \left(\frac{y_1 - \hat{\mu}_{1|2}}{\hat{\sigma}_{1|2}}, \frac{y_2 - \hat{\mu}_2}{\hat{\sigma}_2} \right)', \end{aligned}$$

wobei $\hat{\mu}_{i|j}$ bzw. $\hat{\sigma}_{i|j}^2$ die Schätzer für den bedingten Erwartungswert bzw. für die bedingte Varianz von Y_i gegeben Y_j darstellen. Die entsprechenden Projektionen auf den Einheitskreis sind durch

$$\begin{aligned} \mathbf{z}^{(1)} &= \left(\frac{r_{1|\cdot}}{\sqrt{r_{1|\cdot}^2 + r_{2|1}^2}}, \frac{r_{2|1}}{\sqrt{r_{1|\cdot}^2 + r_{2|1}^2}} \right)' \\ \text{und } \mathbf{z}^{(2)} &= \left(\frac{r_{1|2}}{\sqrt{r_{1|2}^2 + r_{2|\cdot}^2}}, \frac{r_{2|\cdot}}{\sqrt{r_{1|2}^2 + r_{2|\cdot}^2}} \right)' \end{aligned}$$

bestimmt. Sei α der Winkel zwischen den Strecken vom Mittelpunkt des Einheitskreises zu den Projektionen $\mathbf{z}^{(1)}$ und $\mathbf{z}^{(2)}$, dann berechnet sich der Kosinus des Winkels α durch

$$\cos(\alpha) = \mathbf{z}^{(1)T} \mathbf{z}^{(2)} = \frac{r_{1|\cdot} r_{1|2} + r_{2|1} r_{2|\cdot}}{\sqrt{r_{1|\cdot}^2 + r_{2|1}^2} \sqrt{r_{1|2}^2 + r_{2|\cdot}^2}}. \quad (5.5)$$

Zur Vereinfachung der Gleichung seien zunächst die Wurzeln

$$d_1 = \sqrt{r_{1|\cdot}^2 + r_{2|1}^2} \quad \text{und} \quad d_2 = \sqrt{r_{1|2}^2 + r_{2|\cdot}^2}$$

im Nenner betrachtet. Diese stellen die euklidischen Distanzen von $\mathbf{r}^{(1)}$ und $\mathbf{r}^{(2)}$ zum Ursprung dar. Unter Verwendung von

$$r_{2|1} = \frac{y_2 - (\hat{\mu}_2 + \hat{\sigma}_2 \hat{\rho} r_{1|\cdot})}{\hat{\sigma}_2 \sqrt{1 - \hat{\rho}^2}} = \frac{r_{2|\cdot} - \hat{\rho} r_{1|\cdot}}{\sqrt{1 - \hat{\rho}^2}} \quad (5.6)$$

erhält man

$$\begin{aligned}
 d_1 &= \sqrt{r_{1\cdot}^2 + \frac{(r_{2\cdot} - \hat{\rho}r_{1\cdot})^2}{1 - \hat{\rho}^2}} \\
 &= \sqrt{\frac{r_{1\cdot}^2 - r_{1\cdot}^2\hat{\rho}^2 + r_{2\cdot}^2 - 2r_{2\cdot}\hat{\rho}r_{1\cdot} + (\hat{\rho}r_{1\cdot})^2}{1 - \hat{\rho}^2}} \\
 &= \sqrt{\frac{r_{1\cdot}^2 + r_{2\cdot}^2 - 2r_{2\cdot}\hat{\rho}r_{1\cdot}}{1 - \hat{\rho}^2}}.
 \end{aligned}$$

Durch analoges Vorgehen berechnet man das identische Ergebnis für d_2 , d.h. die Distanzen von $\mathbf{r}^{(1)}$ bzw. $\mathbf{r}^{(2)}$ zum Ursprung sind für die alternativen Definitionen gleich. Der Nenner aus Gleichung 5.5 ist daher

$$\text{Nenner} = \frac{r_{1\cdot}^2 - 2\hat{\rho}r_{1\cdot}r_{2\cdot} + r_{2\cdot}^2}{1 - \hat{\rho}^2}.$$

Für den Zähler der Gleichung erhält man (ebenfalls durch die Verwendung des Ergebnisses aus Gleichung 5.6):

$$\begin{aligned}
 \text{Zähler} &= r_{1\cdot} \frac{r_{1\cdot} - \hat{\rho}r_{2\cdot}}{\sqrt{1 - \hat{\rho}^2}} + \frac{r_{2\cdot} - \hat{\rho}r_{1\cdot}}{\sqrt{1 - \hat{\rho}^2}} r_{2\cdot} \\
 &= \frac{r_{1\cdot}^2 - r_{1\cdot}\hat{\rho}r_{2\cdot} + r_{2\cdot}^2 - \hat{\rho}r_{1\cdot}r_{2\cdot}}{\sqrt{1 - \hat{\rho}^2}} \\
 &= \frac{r_{1\cdot}^2 - 2\hat{\rho}r_{1\cdot}r_{2\cdot} + r_{2\cdot}^2}{\sqrt{1 - \hat{\rho}^2}},
 \end{aligned}$$

so dass sich der Gesamtausdruck für den Kosinus des Winkels α zu

$$\frac{\text{Zähler}}{\text{Nenner}} = \sqrt{1 - \hat{\rho}^2}$$

reduziert. Dieses Ergebnis besagt, dass der Kosinus des Winkels zwischen $\mathbf{r}^{(1)}$ und $\mathbf{r}^{(2)}$ nur von $\hat{\rho}$ abhängt. Insbesondere besitzt die Beobachtung $(y_1, y_2)'$ selbst keinen Einfluss auf den Winkel, d.h. für eine beliebige Beobachtung ist die Veränderung des Winkels identisch, wenn eine andere Reihenfolge bei der Definition des mNPR verwendet wird. Daher bleibt die empirische Verteilung der Winkel unverändert, wenn die Referenzrichtung (im Zuge der oben erläuterten Optimierung) ebenfalls um den Winkel $\alpha = \arccos(\sqrt{1 - \hat{\rho}^2})$ korrigiert wird. Entsprechende Berechnungen für die Daten des hier betrachteten Beispiels zeigten, dass dies auch für $p = 4$ gilt.

Da die empirische Verteilung der Winkel für unterschiedliche Definitionen der mNPR identisch ist, wird im Folgenden die naheliegende Definition

$$\mathbf{r} = (\Phi^{-1}(F_{Y_1}(y_1)), \Phi^{-1}(F_{Y_2|Y_1}(y_2)), \Phi^{-1}(F_{Y_3|Y_1Y_2}(y_3)), \Phi^{-1}(F_{Y_4|Y_1Y_2Y_3}(y_4)))'$$

verwendet. Um die mNPR explizit berechnen zu können, werden die Verteilungsfunktionen verschiedener bedingter Verteilungen benötigt, die aus der gemeinsamen Verteilung von Teilmengen aus $\mathbf{Y} = (Y_1, \dots, Y_4)'$ zu bestimmen sind. Jede Teilmenge \mathbf{Y}_A von gemeinsam normalverteilten Zufallsvariablen ist ebenfalls gemeinsam normalverteilt und besitzt die entsprechenden Parameter der gemeinsamen Verteilung von \mathbf{Y} . Beispielsweise wird zur Berechnung von R_3 die gemeinsame Verteilung von $\mathbf{Y}_A = (Y_1, Y_2, Y_3)'$ benötigt, und es ergibt sich eine dreidimensionale Normalverteilung mit Erwartungswertvektor $\boldsymbol{\mu}_A = (\mu_1, \mu_2, \mu_3)'$ und Kovarianzmatrix

$$\Sigma_A = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix},$$

wobei μ_k den Erwartungswert von Y_k und σ_{kj} die Kovarianz von Y_k und Y_j bezeichnen.

Die (univariaten) bedingten Zufallsvariablen Y_k gegeben $\mathbf{Y}_A = \mathbf{y}_A$ sind ebenfalls normalverteilt und ihre Parameter sind (vgl. JOHNSON AND WICHERN, 1998, S. 170)

$$\begin{aligned} \mu_{k|\mathbf{Y}_A} &= \mu_k + \Sigma_{kA}^T \Sigma_A^{-1} (\mathbf{y}_A - \boldsymbol{\mu}_A) \\ \text{und} \quad \sigma_{k|\mathbf{Y}_A}^2 &= \sigma_k^2 - \Sigma_{kA}^T \Sigma_A^{-1} \Sigma_{kA}, \end{aligned}$$

wobei Σ_{kA} den Vektor der Kovarianzen von Y_k und den Elementen von \mathbf{Y}_A beschreibt. Beispielsweise ergibt sich für die Parameter der bedingten Verteilung von Y_3 gegeben $(Y_1, Y_2)' = (y_1, y_2)'$

$$\begin{aligned} \mu_{3|Y_1Y_2} &= \mu_3 + (\sigma_{31}, \sigma_{32}) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}^{-1} ((y_1, y_2)' - (\mu_1, \mu_2)') \\ \text{und} \quad \sigma_{3|Y_1Y_2}^2 &= \sigma_3^2 - (\sigma_{31}, \sigma_{32}) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}^{-1} (\sigma_{31}, \sigma_{32})'. \end{aligned}$$

Bestimmt man auf diesem Weg die Parameterschätzer sämtlicher zur Berechnung der mNPR benötigten Verteilungen und anschließend die mNPR, können

im nächsten Schritt die entsprechenden Winkel bestimmt werden. Ist die Beschreibung der Daten durch vierdimensionale Normalverteilung korrekt, besitzen die Winkel die Verteilungsfunktion

$$F_W(w) = \begin{cases} 0.5 \cdot F_{\text{Beta}}(\sin^2 w; 3/2, 1/2) & 0 \leq w \leq \pi/2 \\ 1 - 0.5 \cdot F_{\text{Beta}}(\sin^2 w; 3/2, 1/2) & \pi/2 < w \leq \pi. \end{cases}$$

Um zu zeigen, welche Ergebnisse die Methode liefert, wenn das Modell die Daten aufgrund einer Konzentration extremer Beobachtungen nicht adäquat beschreibt, wird der Datensatz künstlich kontaminiert, d.h. um eine Konzentration extremer Beobachtungen erweitert. Dazu werden vier weitere Beobachtungen in den Datensatz eingefügt, die in der Nähe des Punktes \mathbf{y}_{16} liegen. (Die quadrierte Mahalanobisdistanz dieser Beobachtung besitzt den größten Wert und ist in diesem Sinn als extreme Beobachtung zu verstehen.) Die künstlichen Beobachtungen werden durch die Addition normalverteilter Zufallszahlen zu den Elementen von $\mathbf{y}_{16} = (1954, 2149, 1180, 1281)'$ erzeugt: Zu jedem Element der Beobachtung wird eine $N(0; 0.05 \cdot \hat{\sigma}_k^2)$ -verteilte Zufallszahl addiert, wobei $\hat{\sigma}_k^2$ die geschätzten Varianzen der Y_k darstellen. Die auf diese Weise erhaltenen Beobachtungen (gekennzeichnet mit den Indizes $i = 31, \dots, 34$) sind (auf ganze Zahlen gerundet) in Tabelle 5.3 gegeben.

Abbildung 5.16 zeigt in der oberen Graphik den Indexplot der NPR der Winkel für den kontaminierten Datensatz. Für die Beobachtung \mathbf{y}_{16} sowie für die hinzugefügten Beobachtungen erhält man Residuen, die sich als Gruppe deutlich von den Residuen der übrigen Beobachtungen unterscheiden, d.h. die vierdimensionale Normalverteilung ist (wie zu erwarten war) nicht in der Lage, die Konzentration der extremen Beobachtungen korrekt abzubilden. Die untere Graphik zeigt einen Indexplot der quadrierten Mahalanobisdistanzen, welche (wie bereits weiter oben erwähnt) nicht geeignet sind, dicht beieinander liegende extreme Beobachtungen zu identifizieren. Dies wird hier bestätigt, da die

i	y_1	y_2	y_3	y_4
31	1922	2131	1114	1284
32	1986	2135	1179	1284
33	1942	2166	1121	1250
34	1904	2177	1243	1300

Tabelle 5.3: Künstlich erzeugte und in den Datensatz eingefügte Beobachtungen.

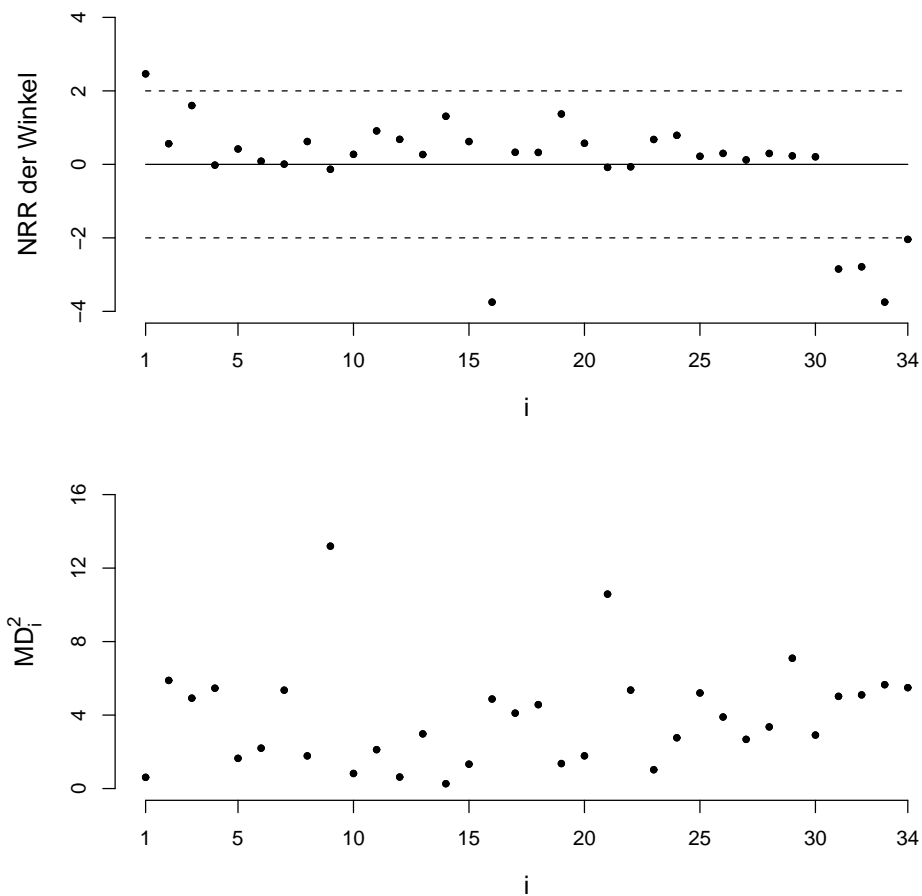


Abbildung 5.16: Identifikation der konzentrierten mNPR nach winkelbasiertem Vorgehen (oben) und mit Hilfe der Mahalanobisdistanzen (unten).

quadratierten Mahalanobisdistanzen der entsprechenden Beobachtungen nicht auffällig sind.



Abschließend sei festgehalten, dass in diesem Abschnitt mit der winkelbasierten Analyse der mNPR das von JUAN AND PRIETO (2001) vorgestellte Vorgehen verallgemeinert wurde: Werden die Winkel der mNPR analysiert, ist die Anwendung der Methode nicht auf die Untersuchung isotropischer Verteilungen begrenzt.

5.4.3 Erfassung der Abhängigkeiten zwischen den Variablen

Eine weitere Eigenschaft, die eine multivariate Verteilung erfüllen sollte, ist die korrekte Beschreibung der stochastischen Abhängigkeit zwischen den zu modellierenden Variablen. Das wohl am häufigsten verwendete Maß zur Beschreibung der Abhängigkeit ist der Korrelationskoeffizient, der die lineare Abhängigkeit zwischen zwei Variablen misst. Wird beispielsweise eine bivariate Normalverteilung für die Modellierung zweier Variablen verwendet, so ist die Korrelation integraler Bestandteil des Modells. Die stochastische Abhängigkeit kann jedoch auch eine Vielzahl anderer Formen besitzen, die nicht durch den Korrelationskoeffizienten beschrieben werden können (für einen Überblick siehe MARI AND KOTZ, 2001). Zum Beispiel ist bekannt, dass die Renditen eines Marktes dazu tendieren, zwar unkorreliert, aber abhängig zu sein (vgl. EMBRECHTS ET AL., 2001). In solchen Fällen muss die Abhängigkeit mit differenzierteren Methoden modelliert werden, d.h. die Modellierung mit einer einzelnen Statistik wie dem Korrelationskoeffizient ist wenig hilfreich, um ein solch komplexes Konstrukt wie die stochastische Abhängigkeit zu beschreiben (vgl. FISHER, 1985).

Beschreibt eine multivariate Verteilung die Abhängigkeit zwischen den Variablen korrekt, so sind die Elemente der multivariaten Pseudo-Residuen unabhängig. Für den Fall, dass die Abhängigkeit nicht korrekt durch das Modell erfasst wird, spiegelt sich dies in der Verteilung der Residuen wider, d.h. die Elemente der multivariaten Pseudo-Residuen sind stochastisch abhängig.

In diesem Abschnitt wird erläutert, wie Abhängigkeiten zwischen den Residuen zu identifizieren sind und welche Bedeutung sie mit Blick auf die Anpassung der multivariaten Verteilung besitzen: Im Folgenden wird zunächst gezeigt, wie die bedingten Erwartungswerte und Varianzen von Elementen der Residuen analysiert werden können. Anschließend wird eine Methode beschrieben, die verwendet werden kann, um explorativ nach unbekanntem Formen der Abhängigkeit (d.h. die Form der Abhängigkeit braucht zur Anwendung der Methode nicht näher spezifiziert werden) zwischen den Elementen der multivariaten Pseudo-Residuen zu suchen.

5.4.3.1 Abbildung bedingter Erwartungswerte und Varianzen

Wenn die Abbildung bedingter Erwartungswerte und Varianzen untersucht wird, rückt die Interpretation der multivariaten Verteilung als Regressionsmo-

dell in den Vordergrund. Werden beispielsweise die Variablen Y_1 und Y_2 durch eine bivariate Verteilung beschrieben, so impliziert dies auch eine Schätzung von $E(Y_1|Y_2)$, $Var(Y_1|Y_2)$, $E(Y_2|Y_1)$ und $Var(Y_2|Y_1)$, d.h. in Analogie zu einem Regressionsmodell wird der Erwartungswert und die Varianz einer Zielvariablen, gegeben eine erklärende Variable, implizit geschätzt.

Die Güte der Schätzung der bedingten Erwartungswerte und Varianzen kann untersucht werden, indem die bedingten Erwartungswerte und Varianzen der Elemente der multivariaten Pseudo-Residuen betrachtet werden. Angenommen es sei für ein bivariates Modell die Beschreibung des bedingten Erwartungswerts von Y_2 gegeben Y_1 zu untersuchen, dann ist es sinnvoll, das mUPR durch

$$(U_1, U_2)' = (F_{Y_1}(Y_1), F_{Y_2|Y_1}(Y_2))'$$

zu definieren. Beschreibt das Modell die Daten korrekt, so besitzt U_2 eine $U(0, 1)$ -Verteilung, und da U_1 und U_2 unabhängig sind, erhält man

$$E(U_2) = E(U_2|U_1) = 0.5.$$

Plottet man also die Beobachtungen für U_1 gegen die von U_2 und schätzt mit einer geeigneten Methode den Erwartungswert von U_2 gegeben U_1 , so sollte daraus in etwa eine horizontale Linie auf dem Niveau 0.5 resultieren. Wird der Wert überschritten, so deutet dies auf tendenziell größere Werte von U_2 in dem entsprechenden Bereich hin, d.h. der Erwartungswert von Y_2 gegeben Y_1 wird unterschätzt. Dabei bezieht sich der Begriff ‚Bereich‘ auf bestimmte Werte von U_1 : Überschreitet die Linie den Wert 0.5 beispielsweise im ‚Bereich‘ $0.25 < U_1 < 0.5$, so bedeutet dies, dass der Erwartungswert von Y_2 gegeben Y_1 unterschätzt wird, wenn der beobachtete Wert von Y_1 zwischen dem ersten Quartil und dem Median der Randverteilung von Y_1 liegt. Dieses Vorgehen stellt eine graphische Methode zur Überprüfung der Abbildung des bedingten Erwartungswertes von Y_2 gegeben Y_1 dar. Eine weitere Methode mit diesem Ziel wird im Rahmen des Beispiels 5.7 erläutert.

Neben der Betrachtung des bedingten Erwartungswerts kann auf ähnliche Weise auch überprüft werden, ob die bedingte Varianz korrekt durch das Modell beschrieben wird. Zur Verdeutlichung sei auch hier die zuvor erläuterte bivariate Problemstellung betrachtet. Das mNPR ist in diesem Fall durch

$$(R_1, R_2)' = (\Phi^{-1}(F_{Y_1}(Y_1)), \Phi^{-1}(F_{Y_2|Y_1}(Y_2)))'$$

definiert. Beschreibt das Modell die Daten korrekt, so besitzt R_2 eine $N(0, 1)$ -Verteilung, und da R_2 und U_1 unabhängig sind, erhält man

$$Var(R_2) = Var(R_2|U_1) = 1.$$

Ferner ist bei Gültigkeit des Modells $E(R_2) = 0$ und somit $Var(R_2) = E(R_2^2)$. Analog zum zuvor dargestellten Vorgehen könnte hier also ein Plot der Beobachtungen für U_1 gegen die von R_2^2 mit einer entsprechenden Schätzung von R_2^2 gegeben U_1 verwendet werden, um zu überprüfen, ob $Var(R_2|U_1)$ korrekt abgebildet wird. Es erscheint hier allerdings sinnvoller, die Beobachtungen für R_2^2 zunächst durch

$$T = F_{\chi_1^2}(R_2^2)$$

monoton zu transformieren ($F_{\chi_1^2}$ sei die Verteilungsfunktion einer χ^2 -Verteilung mit einem Freiheitsgrad): Weil T bei Gültigkeit des Modells ebenfalls eine $U(0, 1)$ -Verteilung besitzt, wird die Schätzung von $E(T|U_1)$ — im Gegensatz zur Schätzung von $E(R_2^2|U_1)$ — nicht durch Ausreißer beeinflusst. Der Erwartungswert von T bei Gültigkeit des Modells beträgt 0.5. Größere Werte indizieren eine größere Streuung der Residuen als unter dem Modell erwartet, d.h. die Varianz von Y_2 gegeben Y_1 wird unterschätzt. Schätzt man also aus einem Scatterplot $E(T|U_1)$, wird ersichtlich, für welche Bereiche von U_1 die bedingte Varianz von Y_2 gegeben Y_1 korrekt bzw. nicht korrekt geschätzt wird.

Beispiel 5.7 Um die Überprüfung der Abbildung bedingter Erwartungswerte und Varianzen zu verdeutlichen, werden erneut die Beobachtungen für den Brusthöhendurchmesser und die Höhe der Buchen betrachtet. Dabei resultieren die hier betrachteten Residuen aus einer Modellierung der Daten durch die S_{BB} -Verteilung (siehe auch Beispiel 5.2).

Interpretiert man die gemeinsame Verteilung als Regressionsmodell, ist es bei den hier betrachteten Daten sinnvoll, den Brusthöhendurchmesser als erklärende Variable und die Baumhöhe als Zielvariable zu definieren, weil der Brusthöhendurchmesser einfach zu messen ist und somit in praktischen Situationen zur Schätzung der Baumhöhen verwendet werden kann. Daher beziehen sich die Darstellungen in diesem Beispiel auf eine Analyse der Residuen

$$\begin{aligned} (U_1, U_2)' &= (F_{Y_D}(Y_D), F_{Y_H|Y_D}(Y_H))' \\ \text{bzw. } (R_1, R_2)' &= (\Phi^{-1}(F_{Y_D}(Y_D)), \Phi^{-1}(F_{Y_H|Y_D}(Y_H)))'. \end{aligned}$$

Betrachtet sei zunächst die Beschreibung des Erwartungswerts von Y_H gegeben Y_D . Ein Scatterplot der Beobachtungen von U_1 gegen die von U_2 ist in Abbildung 5.17 dargestellt. Ferner ist eine nicht-parametrische Glättung der Punkte dargestellt (verwendet wurde der **R**-Befehl `lowess`; siehe CLEVELAND, 1979 und 1981), die als Schätzer des Erwartungswerts von U_2 gegeben U_1 zu interpretieren ist. Dieser beträgt bei einer korrekten Beschreibung der Daten durch

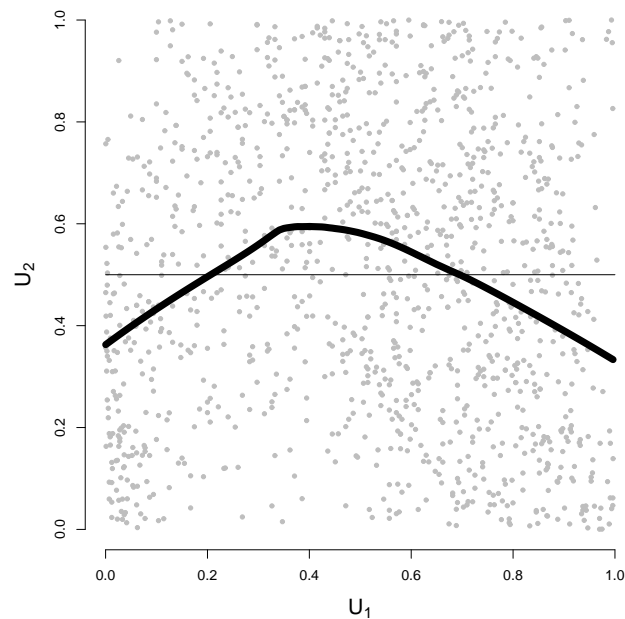


Abbildung 5.17: Scatterplot der Beobachtungen für U_1 gegen die von U_2 sowie nicht-parametrische Schätzung von $E(U_2|U_1)$.

die S_{BB} -Verteilung 0.5. In der Abbildung ist allerdings zu sehen, dass er in den Randbereichen von U_1 kleiner und im mittleren Bereich größer ist als 0.5, d.h. an den Rändern bzw. im mittleren Bereich gibt es tendenziell zu viele kleine bzw. große Werte von U_2 . Mit Blick auf die S_{BB} -Verteilung und den Regressionsgedanken bedeutet dies, dass der bedingte Erwartungswert der Baumhöhe für Bäume mit mittlerem Brusthöhendurchmesser tendenziell unterschätzt und für Bäume mit eher geringem oder großem Brusthöhendurchmesser tendenziell überschätzt wird.

Eine alternative Vorgehensweise, die zuvor noch nicht abstrakt erläutert wurde, zur Überprüfung der Schätzung des bedingten Erwartungswertes der Baumhöhe besteht in einer (speziellen) Analyse der in Abschnitt 5.4.1 beschriebenen dreidimensionalen Histogramme. Wählt man die Klassen des Histogramms so, dass die in Abbildung 5.18 (links) dargestellte Sicht von oben auf das Histogramm entsteht, kann durch einen Vergleich der Graustufen von jeweils zwei vertikal angeordneten Klassen beurteilt werden, ob sich in der oberen und unteren Klasse die Anzahl beobachteter Residuen unterscheidet: Ist die obere Klasse dunkler, befinden sich dort weniger Residuen als in der unteren und umgekehrt. Bei Gültigkeit des Modells ist die erwartete Anzahl von Residu-

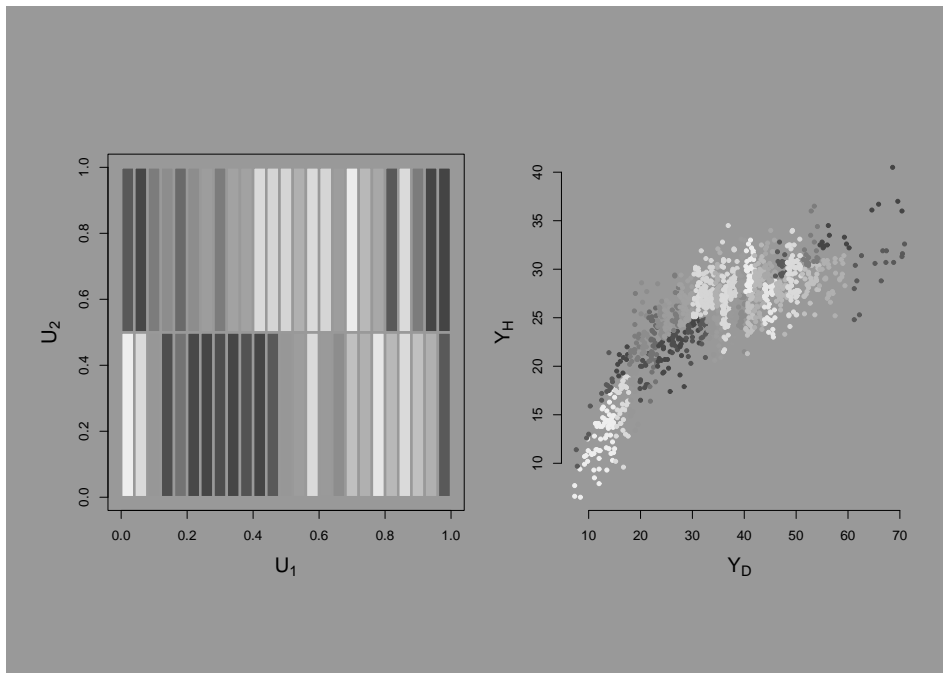


Abbildung 5.18: Analyse des dreidimensionalen Histogramms zur Überprüfung der Schätzung von $E(Y_H|Y_D)$.

en in der oberen und unteren Klasse identisch, so dass durch einen Vergleich der Graustufen beurteilt werden kann, ob sich in dem betrachteten Bereich (gemeint ist wieder ein Bereich von U_1) tendenziell zu viele große oder kleine Werte von U_2 befinden. In Abbildung 5.18 ist für die ersten drei Vergleiche (linker Bereich der linken Graphik) jeweils die untere Klasse heller, d.h. in diesem Bereich sind die Beobachtungen für U_2 tendenziell zu klein. Übertragen auf die Modellanpassung bedeutet dies, dass die Höhe der Bäume bei gegebenem Brusthöhendurchmesser in diesem Bereich überschätzt wird. Anhand der rechten Graphik, in der die Graustufen auf die ursprünglichen Daten übertragen sind (siehe auch Abschnitt 5.4.1), ist zu erkennen, dass dies für Bäume gilt, deren Brusthöhendurchmesser kleiner ist als ca. 20 cm. Auf eine Erläuterung der übrigen Vergleiche wird an dieser Stelle verzichtet, und es sei abschließend auf die Analogie der hier erhaltenen Ergebnisse zu den Ergebnissen aus der Glättung des Scatterplots (Abbildung 5.17) hingewiesen: Die geglättete Kurve verläuft in etwa in dem Bereich über der horizontalen Hilfslinie, in dem die jeweils oberen Klassen heller sind und umgekehrt.

Nach der Abbildung des bedingten Erwartungswerts wird jetzt die Schätzung der bedingten Varianz von Y_H gegeben Y_D beurteilt. Abbildung 5.19 zeigt einen Scatterplot der Beobachtungen von U_1 gegen die von $T = F_{\chi^2_1}(R_2^2)$.

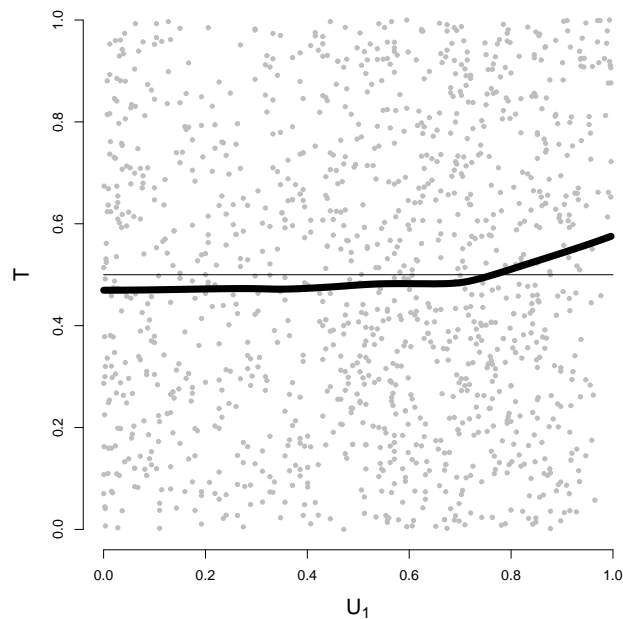


Abbildung 5.19: Scatterplot der Beobachtungen für U_1 gegen die von T sowie nicht-parametrische Schätzung von $E(T|U_1)$.

Die nicht-parametrische Schätzung von $E(T|U_1)$ — auch hier wurde der **R**-Befehl `lowess` verwendet — lässt erkennen, dass die bedingte Varianz von Y_H gegeben Y_D für $U_1 \in [0.0, 0.7]$ geringfügig überschätzt wird. Lediglich im rechten Teil der Abbildung steigt die geglättete Kurve an und indiziert für den entsprechenden Bereich eine Unterschätzung der bedingten Varianz der Baumhöhe bei gegebenem Brusthöhendurchmesser.

Zuvor wurde durch die Analyse eines dreidimensionalen Histogramms eine alternative Vorgehensweise zur Überprüfung der Abbildung des bedingten Erwartungswertes erläutert. Diese Alternative kann analog bei der Überprüfung der bedingten Varianz angewendet werden, worauf hier jedoch verzichtet wird.

◀

5.4.3.2 χ -Plots zur Identifikation nicht spezifizierter Abhängigkeiten

Während im vorangegangenen Abschnitt mit dem bedingten Erwartungswert und der bedingten Varianz die Form der untersuchten Abhängigkeit vor der Analyse der Residuen konkret spezifiziert wurde, dienen die hier betrachte-

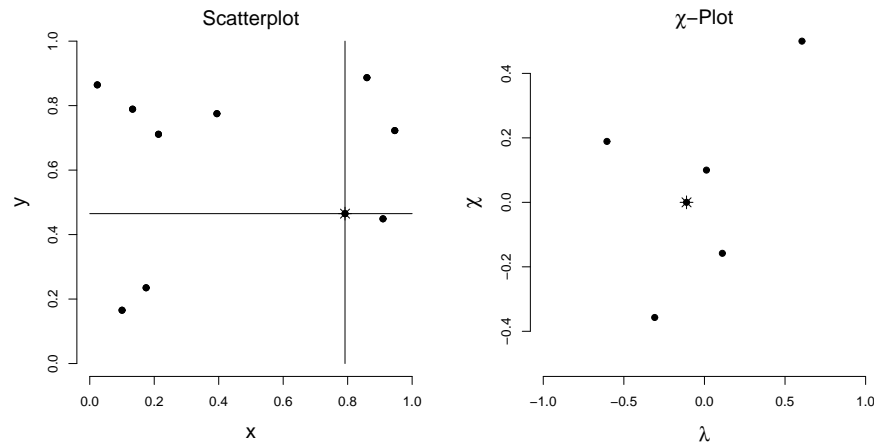


Abbildung 5.20: Scatterplot für 10 Beobachtungen einer bivariaten $U(0,1)$ -Verteilung und zugehöriger χ -Plot.

ten χ -Plots zur allgemeinen Überprüfung der Elemente multivariater Pseudo-Residuen auf verschiedene Formen der stochastischen Abhängigkeit. Die Verwendung der χ -Plots wurde von FISHER AND SWITZER (1985 und 2001) beschrieben. Ferner illustriert ZUCCHINI (2002) ihre Anwendung bei der Analyse multivariater Pseudo-Residuen.

Im Folgenden wird zunächst anhand eines kleinen Beispiels die Konstruktion der χ -Plots erläutert. Im Rahmen von zwei weiteren Beispielen wird anschließend die Gestalt der χ -Plots für verschiedene Formen der Abhängigkeit gezeigt und ihre Verwendung bei der Untersuchung der Residuen beschrieben.

Beispiel 5.8 Die Konstruktion der χ -Plots lässt sich am einfachsten aus einem Scatterplot der Beobachtungen zweier Variablen ableiten. Abbildung 5.20 zeigt in der linken Graphik einen Scatterplot für zehn Beobachtungen (x_i, y_i) einer bivariaten $U(0,1)$ -Verteilung und auf der rechten Seite den zugehörigen χ -Plot. Um den Wert χ_i (abgetragen auf der Ordinate des χ -Plots) eines Punktes (x_i, y_i) zu berechnen, müssen zunächst die Anteile

$$H_i = \sum_{j \neq i} \frac{I\{x_j \leq x_i \text{ und } y_j \leq y_i\}}{n-1},$$

$$F_i = \sum_{j \neq i} \frac{I\{x_j \leq x_i\}}{n-1} \quad \text{sowie}$$

$$G_i = \sum_{j \neq i} \frac{I\{y_j \leq y_i\}}{n-1}$$

		X		
		0	1	
Y	0	2/9	1/9	3/9
	1	4/9	2/9	6/9
		6/9	3/9	

Tabelle 5.4: Relative Häufigkeiten der dichotomisierten Variablen.

bestimmt werden, wobei $I\{A\}$ die Indikatorfunktion für das Ereignis A darstellt. Anschließend erhält man durch die Berechnung von

$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i(1 - F_i)G_i(1 - G_i)}} \quad (5.7)$$

den Wert der Ordinate für den Punkt (x_i, y_i) im χ -Plot. Die Berechnung von χ_i ist in Abbildung 5.20 für den Punkt $(0.79, 0.46)$ angedeutet. Da sechs Punkte links von dem markierten Punkt und drei Punkte darunter beobachtet wurden, erhält man hier $F_i = 6/9$ und $G_i = 3/9$. H_i , der Anteil der Punkte links unterhalb des Punktes $(0.79, 0.46)$, ist in diesem Beispiel $2/9$. Mit diesen Werten erhält man $\chi_i = 0$, den Wert der Ordinate des in der rechten Graphik markierten Punktes. Für die übrigen Punkte berechnet sich der Wert analog, es sei denn, der Nenner aus Gleichung 5.7 ist null. In diesem Fall ist χ_i nicht definiert und der entsprechende Punkt (x_i, y_i) wird nicht in den χ -Plot übertragen.

Der Wert χ entspricht dem Korrelationskoeffizient der dichotomisierten Variablen und misst somit eine besondere Form des linearen Zusammenhangs zwischen den betrachteten Merkmalen. Tabelle 5.4 verdeutlicht diese Interpretation. Sie zeigt die relativen Häufigkeiten, wenn die Dichotomisierung der Variablen mit dem oben betrachteten Punkt erfolgt. Werden die Daten in dieser Form dargestellt, entspricht der Schätzer des Korrelationskoeffizienten

$$\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}}$$

dem gesuchten Wert von χ . Der Schätzer für ρ und die rechte Seite der Gleichung 5.7 sind also identisch.

Auf der Abszisse eines χ -Plots werden die Werte

$$\begin{aligned} \lambda_i &= 4S_i \max\{(F_i - 0.5)^2; (G_i - 0.5)^2\} \\ \text{mit } S_i &= \text{sign}\{(F_i - 0.5)(G_i - 0.5)\} \end{aligned}$$

abgetragen. Beispielsweise erhält man für den in Abbildung 5.20 markierten Punkt

$$\lambda_i = -4 \max \left\{ \frac{1}{36}; \frac{1}{36} \right\} = -\frac{1}{9}.$$

Die λ_i sind als Abstände der Beobachtungen vom bivariaten Median mit Berücksichtigung eines Vorzeichens zu interpretieren: Ist der Abstand einer Beobachtung zum bivariaten Median relativ groß, so erhält man einen relativ großen Wert für $\max\{(F_i - 0.5)^2; (G_i - 0.5)^2\}$ und das Vorzeichen ist positiv, wenn x_i und y_i größer oder kleiner sind als die Mediane der entsprechenden Randverteilungen. Andernfalls ergibt sich ein negatives Vorzeichen. Sind die Beobachtungen unabhängig, verhalten sich die λ_i wie Realisationen einer Rechteckverteilung über dem Intervall $[-1, 1]$ (vgl. FISHER AND SCHWITZER, 1985). Liegen hingegen abhängige Beobachtungen vor, spiegelt sich dies in den Werten der λ_i wider: Bei einer positiven bzw. negativen Abhängigkeit sind auch die Werte der λ_i tendenziell positiv bzw. negativ. Auch die auf der Abszisse abgetragenen Werte besitzen somit Informationen über die Abhängigkeit der betrachteten Variablen.

Anhand dieses Beispiels wird auch klar, dass die χ -Plots auf dem Rang der Beobachtungen basieren. Bei einer Analyse der Residuen spielt es daher keine Rolle, ob die mUPR oder die mNPR untersucht werden, d.h. die resultierenden χ -Plots sind identisch.



Beispiel 5.9 Um einen Eindruck von der Gestalt der χ -Plots bei verschiedenen Formen der Abhängigkeit zu vermitteln, sind in Abbildung 5.21 beispielhaft drei Scatterplots (oben) und die zugehörigen χ -Plots (unten) dargestellt, die jeweils auf 500 Beobachtungen basieren. Die χ -Plots enthalten zusätzlich zwei horizontale Linien bei $\chi = 1.78/\sqrt{n}$ und $\chi = -1.78/\sqrt{n}$. Diese grenzen den Bereich ein, in dem bei Unabhängigkeit in etwa 95% der Punkte liegen (vgl. FISHER AND SWITZER, 2001).

Links oben ist ein Scatterplot von unabhängig und identisch standardnormalverteilten Zufallszahlen dargestellt. Da die betrachteten Variablen nicht assoziiert sind, zeigt der zugehörige χ -Plot (links unten) kaum interpretierbare Strukturen.

Die beiden mittleren Graphiken basieren auf Zufallszahlen einer bivariaten Normalverteilung mit Korrelationskoeffizient $\rho = 0.3$. Die Erwartungswerte

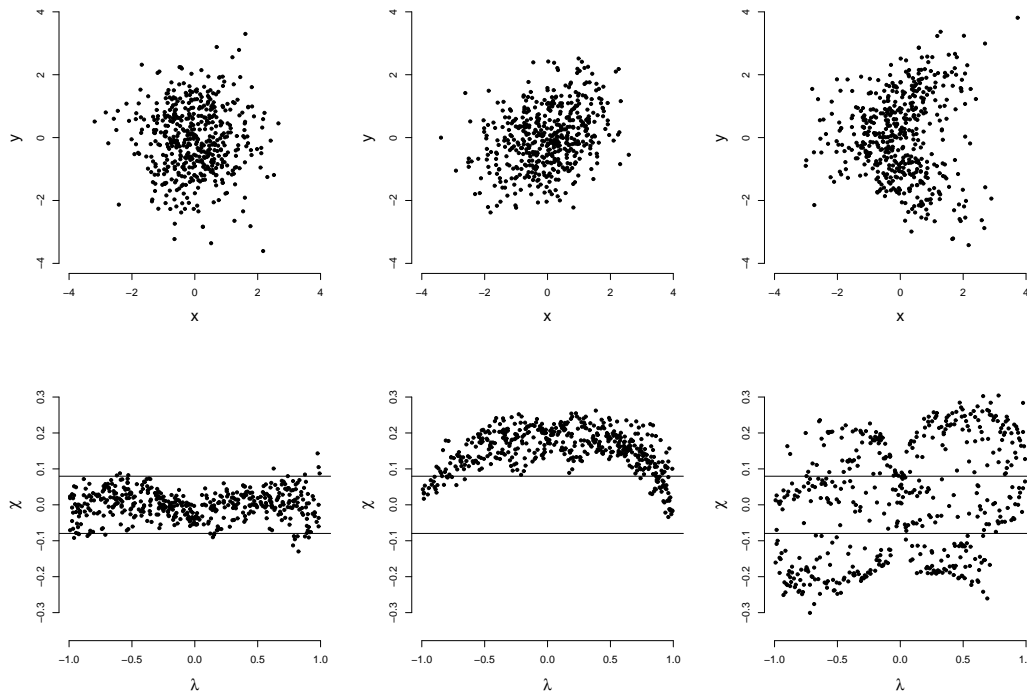


Abbildung 5.21: Scatterplot für jeweils 500 bivariate Beobachtungen mit verschiedenen Formen stochastischer Abhängigkeiten und zugehörige χ -Plots.

bzw. Varianzen der Randverteilungen sind auch hier null bzw. eins. Der positive Zusammenhang zeigt sich im χ -Plot zum einen durch die Wölbung nach oben, d.h. die Korrelationskoeffizienten der dichotomisierten Variablen sind tendenziell positiv. Zum anderen befinden sich auf der horizontalen Achse relativ viele Punkte rechts von null. Tendenziell sind die Werte von x_i und y_i demnach beide größer oder beide kleiner als die Mediane der entsprechenden Randverteilungen.

Im rechten Teil der Abbildung ist die Gestalt des χ -Plots für Zufallszahlen aus einer Mischung von zwei bivariaten Normalverteilungen veranschaulicht. Im Scatterplot sind jeweils 250 Zufallszahlen aus einer $BN(0, -1, 1, 1, -0.5)$ -Verteilung und aus einer $BN(0, 1, 1, 1, 0.5)$ -Verteilung zu sehen. Der zugehörige χ -Plot (rechts unten) besitzt mit Blick auf die horizontale Achse keine auffälligen Strukturen, d.h. die Beobachtungen für λ_i sind in etwa $U(-1, 1)$ -verteilt. Die Korrelation der dichotomisierten Variablen führt allerdings zu einer deutlichen Struktur, die von FISHER AND SWITZER (2001) durch die

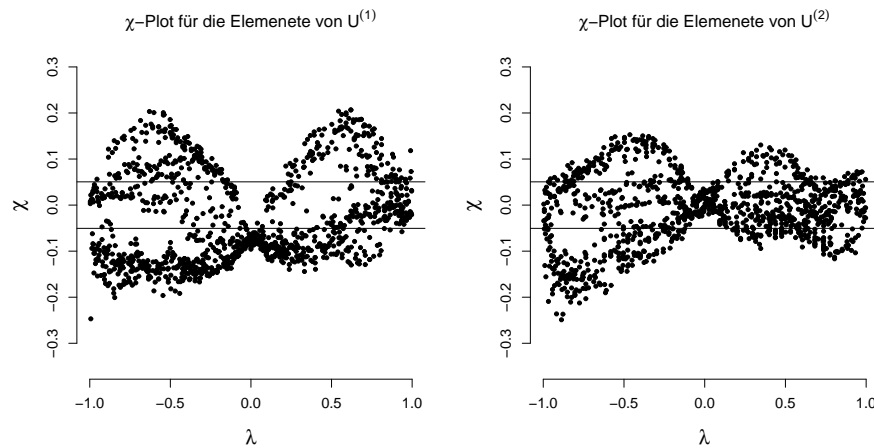


Abbildung 5.22: χ -Plots für die Elemente von $\mathbf{U}^{(1)}$ und $\mathbf{U}^{(2)}$ bei Verwendung der S_{BB} -Verteilung.

Formulierung „grouped into two lobes“ beschrieben und ebenfalls bei der Betrachtung einer Mischung von zwei bivariaten Verteilungen beobachtet wird.



Beispiel 5.10 Die χ -Plots werden nun verwendet, um die bivariaten Residuen zu analysieren, die bei der Modellierung des Brusthöhendurchmessers und der Baumhöhe durch die S_{BB} -Verteilung einerseits (siehe Beispiel 5.2) und durch die Mischung von zwei bivariaten Normalverteilungen andererseits (siehe Abschnitt 2.5.1) beobachtet wurden.

Abbildung 5.22 zeigt die χ -Plots für die beiden alternativen Definitionen der Residuen

$$\begin{aligned} \mathbf{U}^{(1)} &= (F_{Y_D}(Y_D), F_{Y_H|Y_D}(Y_H|Y_D))' \\ \text{bzw. } \mathbf{U}^{(2)} &= (F_{Y_D|Y_H}(Y_D|Y_H), F_{Y_H}(Y_H))' \end{aligned}$$

bei einer Modellierung der Daten mit der S_{BB} -Verteilung. In beiden Plots sind die im vorangegangenen Beispiel beschriebenen ‚lobes‘ zu sehen, was zu der Vermutung führt, dass die Elemente der Residuen aus einer Mischung von zwei bivariaten Verteilungen stammen. Dies ist der Fall, wenn die zu modellierenden Beobachtungen aus einer Mischung von zwei bivariaten Zufallsvariablen stammen, zur Beschreibung jedoch eine Verteilung verwendet wird, die keine Mischung darstellt: Besitzt die (wahre) gemeinsame Verteilung von Y_1 und Y_2

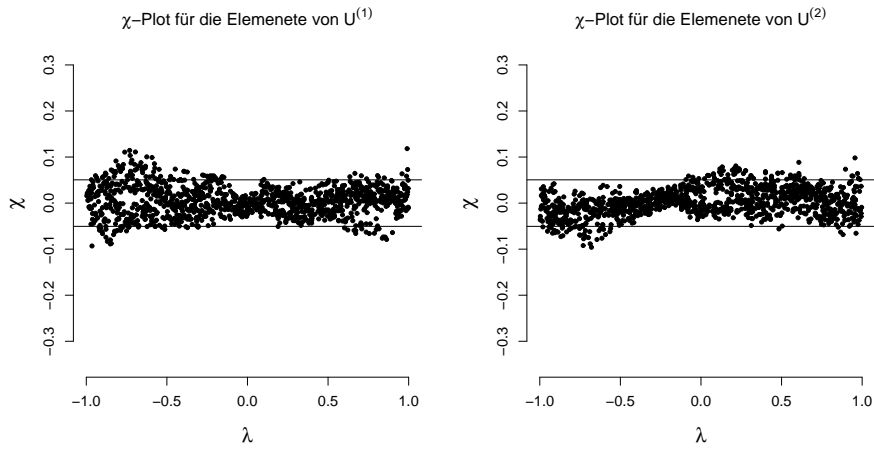


Abbildung 5.23: χ -Plots für die Elemente von $\mathbf{U}^{(1)}$ und $\mathbf{U}^{(2)}$ bei Verwendung einer Mischung von zwei bivariaten Normalverteilungen.

die Dichtefunktion

$$g_{Y_1 Y_2}(y_1, y_2) = \alpha g_{Y_1 Y_2}^{(1)}(y_1, y_2) + (1 - \alpha) g_{Y_1 Y_2}^{(2)}(y_1, y_2) \quad (5.8)$$

einer Mischverteilung, lässt sich die gemeinsame Dichtefunktion von U_1 und U_2 durch

$$q_{U_1 U_2} = \alpha \frac{g_{Y_1 Y_2}^{(1)}}{f_{Y_1 Y_2}} + (1 - \alpha) \frac{g_{Y_1 Y_2}^{(2)}}{f_{Y_1 Y_2}}$$

darstellen. (Das Ergebnis ergibt sich durch Einsetzen von 5.8 in die allgemeine Form der gemeinsamen Dichtefunktion von U_1 und U_2 aus Gleichung 5.3, wobei hier auf eine Angabe der Funktionsargumente verzichtet wird.) Die gemeinsame Dichtefunktion von U_1 und U_2 gehört demnach zu einer Mischverteilung, deren Mischungskomponenten die gemeinsamen Dichtefunktionen

$$\frac{g_{Y_1 Y_2}^{(1)}}{f_{Y_1 Y_2}} \quad \text{und} \quad \frac{g_{Y_1 Y_2}^{(2)}}{f_{Y_1 Y_2}}$$

besitzen. Vor dem Hintergrund dieses Resultats ist die Struktur der χ -Plots in Abbildung 5.21 als Hinweis auf eine Modellmodifikation zu verstehen: Zur Modellierung der Daten sollte eine Mischverteilung verwendet werden. Ein solches Modell (eine Mischung zweier bivariater Normalverteilungen) wurde bereits im Rahmen des einführenden Beispiels in Abschnitt 2.5.1 beschrieben. Analysiert man die entsprechenden Residuen des Modells mit Hilfe der χ -Plots (siehe Abbildung 5.23), so zeigen diese kaum systematische Strukturen.

Die stochastische Abhängigkeit der Beobachtungen erscheint also durch die Mischverteilung wesentlich besser modelliert als durch die S_{BB} -Verteilung.



5.5 Abschließende Bemerkungen zum Kapitel

Die Überprüfung multivariater Verteilungen impliziert bestimmte Problembereiche der Modellüberprüfung, die im univariaten Fall keine Bedeutung besitzen. Zunächst ist hier festzuhalten, dass es keine einheitliche Definition der Residuen gibt, weil $p!$ verschiedene Alternativen bei der Definition der Residuen zur Verfügung stehen: Eine vollständige Analyse umfasst die Untersuchung aller möglichen Definitionen. Auf der anderen Seite gibt es aber auch Situationen, in denen eine bestimmte Reihenfolge aus der Struktur des betrachteten Problems resultiert (siehe Beispiel 5.7) oder in denen die Reihenfolge keine Rolle spielt (siehe Beispiel 5.6). Ein weiteres Kriterium der Anpassungsgüte, das im univariaten Fall keine Rolle spielt, ist die Beschreibung der Abhängigkeiten zwischen den modellierten Merkmalen. Hier wurde mit den χ -Plots eine Methode angewendet, um die Abhängigkeiten zwischen den Elementen der multivariaten Residuen und somit die Erreichung des zuvor genannten Kriteriums zu beurteilen. Ein anderes spezielles Problem stellte die Identifikation der konzentrierten Ausreißer dar. Hier wurde gezeigt, dass durch eine winkelbasierte Analyse der Residuen bessere Ergebnisse erzielt werden als bei Verwendung der Mahalanobisdistanzen.

Bei der winkelbasierten Analyse ist das multivariate Problem auf eine Dimension reduziert worden, indem für eine multivariate Beobachtung ein univariates Residual (das NPR des Winkels) definiert wurde. Die gleiche Idee steht auch hinter den NPR der Distanzen (vgl. Abschnitt 5.4.2.2) und den NPR-Intervallen für die dreidimensionalen Histogramme (vgl. Beispiel 5.3). In diesen Fällen können, unter Berücksichtigung der entsprechenden Interpretation, die Analysemethoden eingesetzt werden, die auch bei der Betrachtung univariater Problemstellungen verwendet wurden.

Kapitel 6

Zusammenfassung und Ausblick

Mit den Pseudo-Residuen wurde in dieser Arbeit ein Residuenkonzept beschrieben, das in einer Vielzahl unterschiedlicher Situationen zur Überprüfung stochastischer Modelle verwendet werden kann. Der grundlegende Vorteil bei der Verwendung von Pseudo-Residuen resultiert aus der Tatsache, dass sie bei Gültigkeit des zu überprüfenden Modells, unabhängig von den im Modell betrachteten Verteilungen, gleiche Verteilungseigenschaften besitzen: Sie können aus diesem Grund sowohl einzeln als auch gemeinsam analysiert werden, und insbesondere die Möglichkeit einer gemeinsamen Analyse macht sie zu einem attraktiven Instrument der Modellüberprüfung. Als solches können sie einen wesentlichen Beitrag zur statistischen Modellbildung leisten: Versteht man die Modellbildung als iterativen Lernprozess (vgl. BOX, HUNTER AND HUNTER, 1978, S. 1–3), können aus einer Analyse der Pseudo-Residuen häufig konkrete Informationen zur Verbesserung des Modells abgeleitet werden. Die Einbettung des Themas in diesen Lernprozess macht klar, dass die Residuenanalyse von der Aussicht auf ein geeigneteres Modell getrieben wird und nicht bei der Identifikation von Anpassungsfehlern endet.

Das übergeordnete Ziel dieser Arbeit war die Veranschaulichung der breiten Anwendungsmöglichkeiten von Pseudo-Residuen bei der Überprüfung stochastischer Modelle. Zu diesem Zweck erfolgte in einem ersten Schritt die Darstellung der theoretischen Grundlagen (vgl. Kapitel 2). Nach einer Beschreibung der grundlegenden Ideen bei der Definition von Residuen wurden auf der Grundlage eines allgemeineren stochastischen Modells die Pseudo-Residuen beschrieben, wobei zwei Varianten (die UPR und die NPR) unterschieden wurden. Für die Pseudo-Residuen bei der Modellierung stetiger Merkmale wurde der Zusammenhang zu den *Cox-Snell residuals* dargestellt und die Verteilung

der Pseudo-Residuen (bei Kenntnis der Modellparameter) hergeleitet. Bei der Modellierung diskreter Merkmale stand die Idee der Definition der Residuen in Form von Intervallen im Vordergrund und es wurde gezeigt, dass sich diese Intervalle bei Gültigkeit des zu überprüfenden Modells wie intervallzensierte Beobachtungen einer $U(0, 1)$ - bzw. $N(0, 1)$ -Verteilung verhalten. Die Verwendung von Intervallen basiert auf der Idee, die Information über die Wahrscheinlichkeit einer Beobachtung unter dem Modell im Residual zu erhalten. Dies hat gegenüber der Verwendung von Punktresiduen den Vorteil, dass vermeintlich extreme Beobachtungen anhand eines zweiten Kriteriums (ihrer Wahrscheinlichkeit unter dem Modell) beurteilt werden können. Abschließend wurden in diesem Kapitel die grundlegenden Ideen bei der Analyse der Residuen im Rahmen zweier einführender Beispiele illustriert.

In einem nächsten Schritt wurden drei ausgewählte Modellsituationen betrachtet, für die anhand einer Reihe von Beispielen besondere Problemstellungen bei der Modellüberprüfung beschrieben wurden.

In Kapitel 3 wurden verallgemeinerte lineare Modelle betrachtet. Nach einer Beschreibung der Modelle wurde zunächst auf die graphische Analyse der Residuen eingegangen. Für stetige Zielvariablen wurde gezeigt, wie der Verlauf von QQ-Plots mit Hilfe entsprechender Pseudo-Residuen objektiv zu beurteilen ist und wie unter Verwendung dieser Residuen simultane Konfidenzbänder für einen QQ-Plot konstruiert werden können. Anschließend wurden mit den QQ-Plots für NPR-Intervalle und den Histogrammen für UPR-Intervalle graphische Methoden zur Analyse bei diskreten Zielvariablen beschrieben. Ferner wurden die Dreieck-Plots erläutert, die auf der Idee basieren, bei der Modellierung diskreter Merkmale die Wahrscheinlichkeit der Beobachtungen unter dem Modell zu erhalten. Nach der graphischen Analyse der Residuen wurde der Einfluss der Parameterschätzung auf die Residuen untersucht. In diesem Zusammenhang wurde ein Vorgehen zur Korrektur der Residuen beschrieben, das auch bei einer Schätzung der Parameter zu exakt $U(0, 1)$ - bzw. $N(0, 1)$ -verteilten Residuen führt. Die entsprechende Transformation konnte für klassische lineare Modelle explizit angegeben werden und für andere Fälle (auch für diskrete Zielvariablen) wurde gezeigt, wie die Transformation unter Verwendung der parametrischen *Bootstrap*-Methode erfolgen kann.

Während die bis zu dieser Stelle der Arbeit betrachteten Problemstellungen allgemeinen Charakter besitzen und auf andere Modellsituationen übertragen werden können, wurden zum Abschluss des Kapitels speziellere Modelle aus der Klasse der verallgemeinerten linearen Modelle betrachtet. Zum einen wur-

den (für das Beispiel mit den Hunden) Residuen definiert, die bei wiederholten Messungen einer binären Variablen geeignet sind, um einerseits zu überprüfen, ob die spezifischen Eigenschaften der einzelnen Untersuchungseinheiten im Modell hinreichend berücksichtigt sind, und andererseits, ob die zeitlichen Eigenschaften der Beobachtungen korrekt durch das Modell beschrieben werden. Diese Art von Pseudo-Residuen konnte auch zur Überprüfung des Modells für den täglichen Niederschlag verwendet werden, das sich aus einem zyklischen, jährlich wiederkehrenden Teil und einem Teil für die jahresspezifischen Eigenschaften des Niederschlags zusammensetzt. Zum anderen wurden Residuen für loglineare Modelle untersucht, die bei der Modellierung von Kontingenztafeln verwendet werden. Hier wurde mit den Tableplots eine graphische Methode zur Analyse der Residuen erläutert und anhand eines Beispiels illustriert, in dem der Markenwechsel auf dem französischen Automarkt modelliert wurde. Die Analyse der Tableplots für verschiedene (auf unterschiedlichen Annahmen basierende) loglineare Modelle erwies sich bei der Beurteilung der Marktsituation als sehr hilfreich.

Im anschließenden Kapitel 4 wurden Anwendungsmöglichkeiten der Pseudo-Residuen bei der Modellierung von Renditezeitreihen dargestellt. Nach einer kurzen Beschreibung von drei ausgewählten Modellen wurden einige graphische Methoden zur Analyse der Residuen beschrieben, die zur Überprüfung verschiedener Aspekte der Voraussageverteilungen geeignet sind. Beispielsweise wurde gezeigt, wie die Vorhersage des VaR oder die Erfassung der zeitlichen Abhängigkeiten mit Hilfe der Pseudo-Residuen beurteilt werden kann. Eine besondere Betonung lag hier auf der Beurteilung der prognostizierten Volatilität, die zur Messung des mit einem Finanztitel verbundenen Risikos herangezogen werden kann. Diesbezüglich wurde ein Vorgehen zur Bestimmung von Zeitabschnitten erörtert, in denen ein Modell keine korrekte Vorhersage der Volatilität liefert. Das Vorgehen basiert auf einer Identifikation von Zeitpunkten, zu denen sich die Varianz der NPR ändert. Dazu wurde hier ein Algorithmus verwendet, der zwei in der Literatur vorgeschlagene Algorithmen kombiniert, und mit Hilfe eines Simulationsexperiments wurde gezeigt, dass der kombinierte Algorithmus eine zuverlässigere Schätzung der Anzahl der Änderungszeitpunkte liefert, jedoch leichte Nachteile bei der Reproduktion der Populationsvarianz besitzt.

Im Rahmen der Modellüberprüfung zeigte sich, dass keines der Modelle alle untersuchten und wünschenswerten Kriterien erfüllte. Es wurde daher eine Methode zur Kalibrierung der Voraussagen vorgestellt, um die Anpassung der

Modelle zu verbessern. Die Idee war hier, ein „zweites“ Modell an die beobachteten Residuen anzupassen, um deren Autokorrelation zu eliminieren und in einem weiteren Schritt die Gestalt der Voraussageverteilungen zu korrigieren. Es zeigte sich, dass die Kalibrierung geeignet ist, um die Autokorrelation der Residuen, nicht aber die der quadrierten Residuen zu beseitigen, wenn ein einfaches Modell für die beobachteten Residuen verwendet wird. Die Kalibrierung der Gestalt wurde in ähnlicher Weise bereits in der Literatur vorgeschlagen und erwies sich als erfolgreich. Insbesondere lieferten die Modelle nach der Kalibrierung wesentlich bessere Voraussagen für die Volatilität der Renditezeitreihen.

Methoden zur Überprüfung der Anpassung multivariater Verteilungen wurden in Kapitel 5 beschrieben. Nach einer entsprechenden Erweiterung des stochastischen Modells wurden Pseudo-Residuen für multivariate Beobachtungen definiert und deren gemeinsame Verteilung in allgemeiner Form beschrieben. Anschließend wurden die dreidimensionalen Histogramme der mUPR zur Überprüfung der grundlegenden Gestalt einer bivariaten Verteilung vorgeschlagen. Insbesondere wurde hier gezeigt, wie die Form der Histogramme mit Hilfe von NPR-Intervallen für die einzelnen Klassen beurteilt werden kann.

Nachfolgend wurden zwei Vorgehensweisen beschrieben, mit denen überprüft werden kann, ob eine multivariate Verteilung extreme Beobachtungen angemessen erfasst: Im Anschluss an eine kurze Einführung in die Problematik wurde zunächst eine Methode beschrieben, die auf einer Analyse der Distanzen der mNPR zum Ursprung basiert. Sie erwies sich als sehr erfolgreich bei der Identifikation von unter dem Modell extremen Beobachtungen, einer Problemstellung, die nicht trivial ist, wenn das betrachtete Modell keine multivariate Normalverteilung ist. Die zweite Methode basiert auf der Berechnung bestimmter Winkel. Sie wurde verwendet, um Konzentrationen extremer Beobachtungen zu identifizieren, die nicht angemessen durch die multivariate Verteilung beschrieben werden. (In diesem Fall liefern übliche, auf der Mahalanobisdistanz basierende Vorgehensweisen keine befriedigenden Ergebnisse.) Bei der Überprüfung multivariater Normalverteilungen ist die hier beschriebene Methode dem Ansatz von JUAN AND PRIETO (2001) sehr ähnlich. Allerdings wurde hier eine geeignetere Identifikation der Referenzrichtung erläutert und darüber hinaus ist das hier dargestellte Vorgehen nicht auf die Analyse isotropischer Verteilungen begrenzt, weil hier die mNPR und nicht die ursprünglichen Beobachtungen analysiert werden.

Abschließend wurde gezeigt, wie überprüft werden kann, ob eine multivariate Verteilung die Abhängigkeitsstrukturen der zu modellierenden Variablen kor-

rekt beschreibt. Zum einen wurde erläutert, wie eine Glättung von Scatterplots der Residuen verwendet werden kann, um zu überprüfen, ob die implizit modellierten bedingten Erwartungswerte und Varianzen korrekt beschrieben werden. Zum anderen wurde mit den χ -Plots eine graphische Methode zur Identifikation verschiedener Formen der stochastischen Abhängigkeit bei der Analyse der multivariaten Residuen eingesetzt. Mit Hilfe der Plots konnten Abhängigkeiten der Residuen identifiziert werden, die auf eine fehlerhaft beschriebene Abhängigkeit der zu modellierenden Merkmale zurückzuführen sind.

Die vorliegende Arbeit ist eine Komposition ausgewählter Beispiele aus verschiedenen Bereichen der statistischen Modellbildung mit jeweils unterschiedlichen Anforderungen an die Modellüberprüfung. Es ist klar, dass aufgrund des begrenzten Umfangs dieser Arbeit für zukünftige Arbeiten eine Vielzahl von Anwendungssituationen existieren, auf die hier nicht eingegangen wurde. Einige der hier beschriebenen Methoden können mehr oder weniger direkt, insbesondere mit der bereitgestellten Software, auf andere Problembereiche übertragen werden. Beispielsweise können die Pseudo-Residuen des QQ-Plots in allen Situationen eingesetzt werden, in denen die Plots zur Überprüfung von Verteilungsannahmen genutzt werden. Für andere Anwendungsbereiche sind die hier beschriebenen Vorgehensweisen als kreativitätsfördernde Ideen zu verstehen. Beispielsweise wurde bei der Modellierung des Lernprozesses der Hunde aus einer konkreten Anforderung an das Modell eine spezielle, zielgerichtete Form von Residuen definiert. Es ist zu erwarten, dass für besondere Anforderungen an Modelle in anderen Situationen (beispielsweise die Bestimmung des Glättungsparameters in verallgemeinerten additiven Modellen) ähnliche Wege zur Definition geeigneter Residuen gangbar sind. Weiterhin gibt es Anwendungen (beispielsweise Tests bezüglich der Residuen), in denen es wünschenswert ist, Residuen zu definieren, die auch bei einer Schätzung der Modellparameter unabhängig sind. Dieses prinzipielle Problem wurde im Rahmen dieser Arbeit nur am Rande erwähnt. Eine mögliche Lösung ist die Verwendung rekursiver Residuen (vgl. BROWN, DURBIN AND EVANS, 1975), die eine gewisse Ähnlichkeit mit den Pseudo-Residuen der Voraussageverteilungen (vgl. Kapitel 4) besitzen. Auch erscheint bei Inkaufnahme einiger Nachteile (beispielsweise des Verlusts beobachtungsspezifischer Interpretationen der Residuen) eine Transformation der Residuen möglich, die zur Unabhängigkeit führt (vgl. COOK AND WEISBERG, 1982, S. 34–35).

Einige konkrete Anregungen für zukünftige Tätigkeiten ergeben sich auch aus Verknüpfungen von einzelnen Teilen dieser Arbeit, die hier nicht im Zusam-

menhang betrachtet wurden. So wurden in Kapitel 5 einige Methoden zur Überprüfung multivariater Verteilungen beschrieben. Diese Methoden besitzen das Potenzial zur Überprüfung von multivariaten Zeitreihenmodellen für Aktienportfolios. Entsprechende Anforderungen an solche Modelle können aus Kapitel 4 abgeleitet werden, in dem Modelle für univariate Renditezeitreihen überprüft wurden. Aus der Verknüpfung dieser Aspekte resultieren beispielsweise die Fragen, ob Modelle für Aktienportfolios den entsprechenden VaR korrekt prognostizieren oder ob extreme multivariate Renditen adäquat durch das Modell erfasst werden. Ein anderer Schwerpunkt weiterer Arbeiten zu Pseudo-Residuen könnte in der Überprüfung von multivariaten Verteilungen für diskrete Merkmale oder auch diskrete und stetige Merkmale liegen. Anstöße zur Entwicklung entsprechender Methoden sind im Rahmen dieser Arbeit bei der graphischen Analyse der UPR- bzw. NPR-Intervalle aufgezeigt worden.

Man könnte abschließend festhalten, dass eine Vielzahl von Fragen offen bleibt. Dieses Ergebnis ist allerdings keinesfalls deprimierend. Vielmehr ist es im Sinn von Sokrates „Ich weiß, dass ich nichts weiß“ zu deuten: Die offenen Fragen resultieren aus dem breiten Spektrum potenzieller Anwendungsmöglichkeiten, die im begrenzten Raum dieser Arbeit nur beispielhaft veranschaulicht werden konnten.

Anhang A

Programme

Eine Vielzahl der in dieser Arbeit vorgestellten graphischen Methoden sowie einige Funktionen zur Berechnung von Pseudo-Residuen sind in der **R**-*library* `psresiduals` zusammengestellt. Aufgrund des großen Umfangs der Programme wird an dieser Stelle eine Kurzbeschreibung der einzelnen Funktionen angegeben, so dass die wesentlichen Inhalte der Programme nachvollzogen werden können. Die Software wird auf Anfrage gerne vom Autor bereitgestellt (astadie@uni-goettingen.de). Entsprechend der **R**-Standards sind die folgenden Beschreibungen in englischer Sprache verfasst.

uhistbounds *Histograms with simultaneous confidence intervals*

Description

This function computes the simultaneous confidence intervals for the bars of a histogram of values between zero and one.

Usage

```
uhistbounds(x,k=10,clevel=0.95,pl=F)
```

Arguments

<code>x</code>	A vector of values for which the histogram is required.
<code>k</code>	The number of classes of the histogram.
<code>clevel</code>	The required (simultaneous) confidence level.
<code>pl</code>	Logical value. If 'pl=FALSE' (default), a histogram is not plotted.

See Also

`bands`

Examples

```
# Under the model
x<-runif(1000)
uhistbounds(x,k=10,clevel=0.95,pl=F)

# Under an alternative
x<-rbeta(1000,0.5,0.5)
uhistbounds(x,k=10,clevel=0.95,pl=F)
```

psglm

Pseudo residuals for a generalized linear model

Description

This function computes the UPRs and NPRs from an object of class ‘glm’.

Usage

```
psglm(mod, m=F)
```

Arguments

<code>mod</code>	Object of class ‘glm’.
<code>m</code>	Vector which contains the number of trials for each observation if family=“binomial”.

Value

<code>rpr</code>	The UPRs from the model.If the independent variable is discrete the residual intervals are returned as matrix with 3 columns: the first and third column are the interval limits, the second column is the midpoint.
<code>npr</code>	The NPRs from the model.If the independent variable is discrete the residual intervals are returned as matrix with 3 columns: the first and third column are the interval limits, the second column is the midpoint.

Examples

```
# discrete case
x<-1:10
s<-rbinom(10,20,x/11)
y<-matrix(c(s,20-s),10,2)
mod<-glm(y~x,family="binomial")
```

```
psglm(mod,rep(20,10))

# continous case
x<-1:10
y<-rgamma(10,shape=x)
mod<-glm(y~x,family="Gamma")
psglm(mod)
```

psqq

Residuals for a qq-plot

Description

This function computes the UPRs and NPRs of a quantile-quantile plot.

Usage

```
psqq(rpr)
```

Arguments

rpr The (continuous) UPRs from the model to be assessed.

Examples

```
rpr<-runif(200)
rofqq<-psqq(rpr)
par(mfrow=c(1,3))
# qq-plot
qqnorm(qnorm(rpr))
abline(0,1)
# Indexplot of its NPRs
plot(1:200,rofqq$nprqq)
# Histogram of its UPRs
hist(rofqq$rprqq)
```

bands	<i>Simultaneous confidence bands for a normal qq-plot</i>
-------	---

Description

This function adds simultaneous confidence bands to a normal qq-plot.

Usage

```
bands(npr, alpha=0.05, ...)
```

Arguments

npr	The NPRs from the model to be assessed.
alpha	The required confidence level (0.01,0.05 (default) and 0.1 are available).
...	Further graphical parameters to 'lwd', 'col' and 'lty'.

Details

The confidence bands can be used as a test for the NPRs of a model. If the points wander outside the boundaries 'cross=T' is returned, otherwise 'cross=F'.

Value

cross	Logical value. If 'cross=F' the qq-plot lies entirely between the bands otherwise outside.
-------	--

See Also

uhistbounds

Examples

```
# Under the model
x<-rnorm(200)
qqnorm(x)
bands(x,col=2)
```

```
# Under an alternativ
x<-rt(200,5)
qqnorm(x)
bands(x,col=2)
```

qqnorm.dis *qq-plot for NPR-intervals*

Description

This function produces a qq-plot for the NPR-intervals from a model for discrete observations.

Usage

```
qqnorm.dis(mat,l=4,...)
```

Arguments

<code>mat</code>	Matrix with three columns that contain the NPR-intervals. Such a matrix is for example returned by the function ‘psglm’.
<code>l</code>	Absolute value for the x-limits and y-limits of the plot.
<code>...</code>	Further graphical parameters to ‘lwd’, ‘cex.axis’ and ‘cex.lab’.

Examples

```
x<-1:10
s<-rbinom(10,20,x/11)
y<-matrix(c(s,20-s),10,2)
mod<-glm(y~x,family="binomial")
res<-psglm(mod,rep(20,10))
qqnorm.dis(res$npr,l=2.5)
```

set.trisys	<i>Triangular coordinate system</i>
------------	-------------------------------------

Description

This function plots a triangular system to which points can be added in a subsequent step.

Usage

```
set.trisys(hl=T,...)
```

Arguments

hl	Logical value. If 'hl=TRUE' (default) horizontal lines are added to the system.
...	Further graphical parameters to 'cex.axis', 'cex.lab' and 'main'.

See Also

```
tri.bands,tri.xy
```

Examples

```
set.trisys()
```

<code>tri.bands</code>	<i>Critical regions for a triangular system</i>
------------------------	---

Description

This function displays some critical regions in a triangular system.

Usage

```
tri.bands(alphavec=c(0.05,0.1),...)
```

Arguments

<code>alphavec</code>	Vector giving the significance levels.
<code>...</code>	Further graphical parameters to 'lwd', 'col' and 'lty'.

Details

The probability than an extreme UPR-interval falls in the critical region is less than 'alphavec'.

See Also

```
set.trisys,tri.xy
```

Examples

```
set.trisys()  
tri.bands(col=2)
```

tri.xy	<i>Coordinates for UPR-intervals in a triangular system</i>
--------	---

Description

This function computes the coordinates for UPR-intervals in a triangular system produced by 'set.trisys'.

Usage

```
tri.xy(rpr)
```

Arguments

rpr	The matrix 'rpr' is computed from a matrix of UPRs (such as returned by the function 'psglm'). The first, second and third columns give the probabilities that an observation is less, equal or bigger than that observed, respectively, under the model.
-----	---

See Also

```
set.trisys, tri.bands
```

Examples

```
x<-1:10
s<-rbinom(10,20,x/11)
y<-matrix(c(s,20-s),10,2)
mod<-glm(y~x,family="binomial")
rpr<-psglm(mod,rep(20,10))$rpr
rpr[,2]<-rpr[,3]-rpr[,1]
rpr[,3]<-1-rpr[,3]
xy<-tri.xy(rpr)
set.trisys()
points(xy$x,xy$y,pch=20)
```

hist.dis1	<i>Envelope of a histogram for UPR-intervals</i>
-----------	--

Description

This function computes the envelope of a histogram for UPR-intervals.

Usage

```
hist.dis(rpr,plot=T,...)
```

Arguments

rpr	Matrix with three columns that contain the UPR-intervals (such as returned by the function 'psglm').
plot	Logical value. If 'plot=T' (default) the envelop is plotted.
...	Further graphical parameters to 'lwd'.

Value

x	the x-values of the envelope.
y	the y-values of the envelope.

See Also

hist.dis2

Examples

```
rpr<-matrix(c(0.2,0.45,0.7,  
0.5,0.65,0.8,0.85,0.9,0.95),ncol=3,byrow=T)  
hist.dis1(rpr,lwd=5,plot=T)
```

hist.dis2 *Histogram for UPR-intervals*

Description

This function produces a histogram for UPR-intervals.

Usage

```
hist.dis2(y,nclass,ymax=F)
```

Arguments

y	The y-values of the envelope produced by 'hist.dis1'.
nclass	The number of classes of the histogram.
ymax	The y-limit of the plot. If not specified the height of the highest bar is used.
...	Further graphical parameters to 'xlab', 'ylab', 'cex.axis' and 'cex.lab'.

See Also

hist.dis1

Examples

```
rpr<-matrix(c(0.2,0.45,0.7,  
0.5,0.65,0.8,0.85,0.9,0.95),ncol=3,byrow=T)  
env<-hist.dis1(rpr,plot=F)  
hist.dis2(env$y,5)
```

`tableplot`*Table-plots for loglinear models*

Description

This function produces a table-plot based on the NPR-intervals from a loglinear model.

Usage

```
tableplot(nprmid, rowlab=I:1, collab=1:J, lim=4, ...)
```

Arguments

<code>nprmid</code>	Matrix that contains the midpoints of the NPR-intervals from a loglinear model for each cell of the contingency table.
<code>rowlab</code>	Names of the row categories.
<code>collab</code>	Names of the column categories.
<code>lim</code>	Midpoints greater than ‘lim’ are painted white on the plot; those smaller than ‘-lim’ are painted black.
<code>...</code>	Further graphical parameters to ‘cex.axis’ and ‘cex.lab’.

Examples

```
mids<-matrix(rnorm(50),ncol=5,nrow=10)
tableplot(mids,lim=2.5)
```

rowrpr	<i>Residuals for repeated measurements on binary outcomes</i>
--------	---

Description

This function computes the UPR-interval for each individual (each row) from a model for repeated measurements on binary outcomes.

Usage

```
rowrpr(X,fit,B=1000)
```

Arguments

X	Matrix that contains the observed outcomes. The individuals are represented by rows of the matrix.
fit	Matrix that contains the probabilities of success for each outcome. These are usually estimated under a given model.
B	Number of replications used to approximate the Poisson Binomial distribution under the model.

See Also

colrpr

Examples

```
obs<-matrix(sample(c(0,1),100,replace=T),10,10)
fit<-matrix(mean(obs),10,10)
rpr<-rowrpr(obs,fit,100)
qqnorm.dis(qnorm(rpr),l=2.5)
```

colrpr	<i>Residuals for repeated measurements on binary outcomes</i>
--------	---

Description

This function computes the UPR-interval for each trial (each column) from a model for repeated measurements on binary outcomes.

Usage

```
colrpr(X,fit,B=1000)
```

Arguments

<code>X</code>	Matrix that contains the observed outcomes. The trials are represented by columns of the matrix.
<code>fit</code>	Matrix that contains the probabilities of success for each outcome. These are usually estimated under a given model.
<code>B</code>	Number of replications used to approximate the Poisson Binomial distribution under the model.

See Also

`rowrpr`

Examples

```
obs<-matrix(sample(c(0,1),100,replace=T),10,10)
fit<-matrix(mean(obs),10,10)
rpr<-colrpr(obs,fit,100)
qqnorm.dis(qnorm(rpr),l=2.5)
```

`cusumsq`*CUSUMSQ for time series*

Description

This function computes the simultaneous critical bands for the CUSUMSQ of a time series.

Usage

```
cusumsq(y, von=1, bis=length(y), pl=1, test=T, achse2=F, ...)
```

Arguments

<code>y</code>	A vector which contains a time series.
<code>von</code>	The time point for the start of the series.
<code>bis</code>	The time point for the end of the series.
<code>...</code>	Further graphical parameters to 'main', 'xlab', 'ylab', 'cex.axis' and 'cex.lab'.

Details

The critical bands are produced under the assumption that the observations are independently distributed with constant variance (see section 4.4.5).

Value

<code>k.stern</code>	is zero if there are no variance change points; otherwise it is the estimated position of the change point relative to 'von'.
----------------------	---

See Also

`varbreaks`

Examples

```
# Plots the centered CUSUMSQ with simultaneous confidence bands
p1<-rnorm(100,0,sqrt(2))
p2<-rnorm(100,0,1)
Y<-c(p1,p2)
cusumsq(Y,test=F)

# Approximately 5 change points are identified
test<-rep(0,100)
B<-100
for (i in 1:B)
  {
    y<-rnorm(100)
    test[i]<-cusumsq(y,p1=F)
  }
length(test[test!=0])
```

varbreaks *Variance change points for time series*

Description

This function estimates variance change points of a time series.

Usage

```
varbreaks(y, von=1, bis=length(y), p1=1, p2=1)
```

Arguments

<code>y</code>	A vector which contains a time series.
<code>von</code>	The time point for the start of the series.
<code>bis</code>	The time point for the end of the series.
<code>p1</code>	Numeric. If 'p1=1' (default), the tested CUSUMSQs of the first step are plotted.
<code>p2</code>	Numeric. If 'p2=1' (default), the tested CUSUMSQs of the refinement are plotted.

Details

The function produces two estimates for the variance change points: The first estimate 'stufel' is computed by the BDF-algorithm; the second ('grenzen.new') by the algorithm used in this work (see section 4.4.6).

Value

<code>stufel</code>	The estimated positions of the change points after the first step. These are the estimates produced by the BDF-algorithm.
<code>grenzen.new</code>	The estimated positions of the change points after the refinement step.
<code>it</code>	Number of iterations in the refinement step.

See Also

cusumsq, varperiods

Examples

```
# Shows the first step and the refinement step of the algorithm
set.seed(4)
p1<-rnorm(100,0,sqrt(4))
p2<-rnorm(100,0,sqrt(0.5))
p3<-rnorm(100,0,1)
par(mfrow=c(2,3))
Y<-c(p1,p2,p3)
varbreaks(Y,p12=F)
par(mfrow=c(2,3))
varbreaks(Y,p11=F)
```

varperiods	<i>The variance of the NPRs in separate periods of a time series</i>
------------	--

Description

This function plots a time series and the estimated variance of the corresponding NPRs minus one.

Usage

```
varperiods(y,ps,breaks,von=1,main="",omit=0,pl=T)
```

Arguments

y	A vector which contains a time series.
ps	A vector which contains the NPRs of the series.
breaks	The estimated variance change points NPRs.
omit	Numeric. The number of extreme variance values omitted in the plot.
pl	Logical. If 'pl=T' (default), the plot is made.

Value

varper	The estimated variance of the NPRs in the periods of different variances.
--------	---

See Also

varbreaks

Examples

```
set.seed(4)
p1<-rnorm(100,0,sqrt(4))
p2<-rnorm(100,0,sqrt(0.5))
p3<-rnorm(100,0,1)
Y<-c(p1,p2,p3)
s<-sd(Y)
ps<-qnorm(pnorm(Y,0,s),0,1)
br<-varbreaks(Y,p11=F,p12=F)$grenzen.new
varperiods(Y,ps,br)
```

hist3d *Three-dimensional histograms*

Description

This function produces a histogram for bivariate observations.

Usage

```
hist3d<-function (x,y,nclassx=10,nclassy=10,colortype="heat",  
th=30,ph=20,lth=50,lph=50,prob=T,...)
```

Arguments

x	Vector which contains the observations of the first variable.
y	Vector which contains the observations of the second variable.
nclassx	Number of classes for the first variable.
nclassy	Number of classes for the second variable.
colortype	Color palette used for the histogram.
th,ph	Angles defining the viewing direction. 'th' gives the azimuthal direction and 'ph' the colatitude.
lth,lph	If non-zero values are specified for 'lth' and 'lph', the surface is shaded as though it was being illuminated from the direction specified by azimuth 'lth' and colatitude 'lph'.
prob	Logical. If 'prob=T' (default) the volume under the histogram is one.
...	Further graphical parameters to 'xlab' and 'ylab'.

See Also

reshist

Examples

```
x<-rnorm(1000)
y<-rnorm(1000)
hist3d(x,y,prob=T)
```

reshist	<i>Residuals for a three-dimensional histogram</i>
---------	--

Description

This function computes the NPR-intervals for a histogram of bivariate UPRs.

Usage

```
reshist(rprmat,nclassx=10,nclassy=10,plot=T,point=T,...)
```

Arguments

<code>rprmat</code>	Matrix containing the bivariate UPRs in two columns.
<code>nclassx</code>	Number of classes for the first variable.
<code>nclassy</code>	Number of classes for the second variable.
<code>plot</code>	Logical. If ‘plot=T’ (default) the midpoints of the histogram residuals are plotted as gray scales in a table-plot.
<code>point</code>	Logical. If ‘point=T’ (default) points are used to indicate extreme areas of the histogram.
<code>...</code>	Further graphical parameters to ‘xlab’, ‘ylab’, ‘cex.lab’ and ‘cex.axis’.

Value

<code>xbreaks</code>	The breaks for the first variable which correspond to ‘nclassx’.
<code>ybreaks</code>	The breaks for the second variable which correspond to ‘nclassy’.
<code>counts</code>	A matrix that contains the counts for each class of the histogram.

<code>npr</code>	An array that contains the NPR-intervals of the histogram. The first two dimensions define the class. The third dimension includes the left and the right interval limit (<code>'npr[,1]'</code> and <code>'npr[,3]'</code>) and the midpoint of the NPR-interval (<code>'npr[,2]'</code>).
<code>rprex</code>	A matrix that contains the midpoints of the NPR-intervals for each class of the histogram (same as <code>'npr[,2]'</code>).
<code>rprexcol</code>	A matrix that contains the gray scales used the NPR-intervals for each class of the histogram.
<code>rprmat</code>	See arguments.
<code>indmat</code>	Matrix with two columns. The first column contains the index of all original observations that fall in extreme classes of the histogram. The second column contains the color of the corresponding class on a gray scale.
<code>back</code>	Background gray scale.

See Also

`hist3d`

Examples

```
x<-rbeta(1000,1.4,1.4)
y<-rbeta(1000,1.4,1.4)
rprmat<-matrix(c(x,y),1000,2)
par(mfrow=c(1,2),bg=gray(0.6))
hist3d(x,y)
reshist(rprmat,nclassx=10,nclassy=10,plot=T,xlab="",ylab="",
point=T,cex.lab=1,cex.axis=1)
```

multnormres *Residuals for a multivariate normal distribution.*

Description

This function computes the mUPRs for multivariate observations under a multivariate normal distribution.

Usage

```
multnormres(obs,mu,Ssquare,ord=1:ncol(obs))
```

Arguments

obs	Matrix which contains the observations. A row represents an observation, a column a variable.
mu	Mean vector of the model (typically the mean over the columns of 'obs').
Ssquare	Covariance matrix of the model (typically 'cov(obs)').
ord	The order in which the elements of the mUPRs are defined.

Value

umat	Matrix which contains the mUPRs. A row represents an observation, a column a variable.
------	--

Examples

```
# Under the model
set.seed(1)
x<-rnorm(100,0,1)
y<-rnorm(100,x,1)
obs<-matrix(c(x,y),ncol=2)
mu<-apply(obs,2,mean)
Ssquare<-cov(obs)
```

```
a<-multnormres(obs,mu,Ssquare)
par(mfrow=c(1,2))
qqnorm(qnorm(a$umat[,1]))
abline(0,1)
qqnorm(qnorm(a$umat[,2]))
abline(0,1)

# Under an alternativ
set.seed(1)
x<-rnorm(100,0,1)
y<-rnorm(100,x,1)
obs<-matrix(c(x,y),ncol=2)
mu<-c(0,0)
Ssquare<-matrix(c(1,-0.5,-0.5,1),2,2)
a<-multnormres(obs,mu,Ssquare)
par(mfrow=c(1,2))
qqnorm(qnorm(a$umat[,1]))
abline(0,1)
qqnorm(qnorm(a$umat[,2]))
abline(0,1)
```

`resangles` *Residuals for the angles of mNPRs*

Description

This function computes residuals of the angles for mNPRs.

Usage

```
resangles(mat)
```

Arguments

`mat` Matrix which contains the mNPRs. A row represents an observation, a column a variable.

Value

`u` Matrix which contains the coordinates of the projections of the residuals on the unit sphere.

`u0` Direction of the residual which maximizes the discrepancy, namely the lack of fit. Used as the starting value for the ‘nlm’.

`umax` Direction which maximizes the discrepancy.

`indmax` Index of the observation with the biggest absolute NPR of the angle.

`angles` Angles of the observations to the reference direction ‘umax’.

`rpr` UPRs of the angles for each observation.

`npr` NPRs of the angles for each observation.

`maxi` Object returned by ‘nlm’ which is used to maximize the discrepancy.

`dsquare` Squared Mahalanobis distances of the mNPRs given in ‘mat’.

`sortobs` Order of the observations with respect to their angles.

Examples

```
# Under the model
set.seed(10)
x<-rnorm(100,0,1)
y<-rnorm(100,0,1)
z<-rnorm(100,0,1)
mat<-matrix(c(x,y,z),ncol=3)
res<-resangles(mat)
qqnorm(res$npr)
abline(0,1)
```

```
# For contaminated mNPRs
set.seed(1)
x<-c(rnorm(50,0,1),rnorm(5,2,0.1))
y<-c(rnorm(50,0,1),rnorm(5,2,0.1))
z<-c(rnorm(50,0,1),rnorm(5,2,0.1))
mat<-matrix(c(x,y,z),ncol=3)
res<-resangles(mat)
plot(res$npr)
```

`chiptot`*Chiptot for bivariate observations.*

Description

This function computes the positions in a chiptot for bivariate observations.

Usage

```
chiptot(x,y)
```

Arguments

<code>x</code>	Vector that contains the observations for the first variable.
<code>y</code>	Vector that contains the observations for the second variable.

Value

<code>x</code>	See ‘arguments’.
<code>y</code>	See ‘arguments’.
<code>H</code>	Values of the marginal empirical distribution function evaluated at the values of <code>x</code> .
<code>G</code>	Values of the marginal empirical distribution function evaluated at the values of <code>y</code> .
<code>F</code>	Values of the joint empirical distribution function evaluated at the values of <code>x,y</code> .
<code>Lam</code>	Vector that contains the x-coordinates in the chiptot.
<code>Chi</code>	Vector that contains the y-coordinates in the chiptot.
<code>cp</code>	Vector with three elements from which linear boundaries can be drawn. The boundaries determine the regions in which (under independence) approximately 0.9, 0.95 or 0.99 of the points should lie.

Examples

```
# For independent observations
set.seed(10)
x<-rnorm(100,0,1)
y<-rnorm(100,0,1)
pos<-chiplot(x,y)
plot(pos$Lam,pos$Chi,ylim=c(-0.25,0.25))
abline(h=c(pos$cp[2],-pos$cp[2]),col=2)

# For dependent observations
set.seed(1)
x<-rnorm(100,0,1)
y<-rnorm(100,x*0.5,1)
pos<-chiplot(x,y)
plot(pos$Lam,pos$Chi,ylim=c(-0.5,0.5))
abline(h=c(pos$cp[2],-pos$cp[2]),col=2)
```

Literaturverzeichnis

- AGRESTI A. (1990): *Categorical Data Analysis*. Wiley, New York et al.
- ANSCOMBE F.J. (1953): Contribution to the discussion on Professor Hotelling's Paper. *Journal of the Royal Statistical Society, Series B*, **16**, 229–230.
- ANSCOMBE F.J. (1961): Examination of Residuals. *Proceedings of the Fourth Berkeley Symposium*, **1**, 1–36.
- ANSCOMBE F.J. AND TUKEY W. (1963): The Examination and Analysis of Residuals. *Technometrics*, **5**, 141–160.
- BARNDORFF-NIELSEN O. (1978): *Information and Exponential Families*. Wiley & Sons, Chichester et al.
- BARNETT V. AND LEWIS T. (1994): *Outliers in Statistical Data*. Wiley & Sons, Chichester et al.
- BELSLEY D.A., KUH E. AND WELSCH R.E. (1980): *Regression Diagnostics*. Wiley & Sons, New York et al.
- BERKOWITZ J. (2001): Testing Density Forecasts, With Applications to Risk Management. *Journal of Business & Economic Statistics*, **19**, 465–474.
- BIRCH M.W. (1963): Maximum-Likelihood in Three-Way Contingency Tables. *Journal of the Royal Statistical Society, Series B*, **25**, 220–233.
- BISHOP Y.M.M., FIENBERG S.E. AND HOLLAND P.W. (1975): *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge et al.
- BODGAN M. (1999): Data driven smooth tests for bivariate normality. *Journal of Multivariate Analysis*, **68**, 26–53.

- BOLLERSLEV T. (1986): Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.
- BOS T., DING D. AND FETHERSTON T.A. (1998): Searching for Periods of Volatility: A Study of the Behaviour of Volatility in Thai Stocks. *Pacific-Basin Finance Journal*, **6**, 295–306.
- BOX G.E.P., HUNTER W.G. AND HUNTER J.S. (1978): *Statistics for Experimenters*, Wiley & Sons, New York et al.
- BRILLINGER D.R. AND PREISLER H.K. (1983): Maximum Likelihood Estimation in a Latent Variable Problem. KARLIN S., AMEMIYA T. AND GOODMAN L.A. (eds.): *Studies in Econometrics, Time Series, and Multivariate Statistics*. Academic Press, New York et al., 31–65.
- BROWN R.L., DURBIN J. AND EVANS J.M. (1975): Techniques for Testing the Constancy of Regression Relationships (with Discussion). *Journal of the Royal Statistical Society, Series B*, **37**, 149–163.
- BUSH R.R. AND MOSTELLER F. (1955): *Stochastic models for Learning*. Wiley, New York.
- CHAMBERS J.M., CLEVELAND W.S., KLEINER B. AND TUKEY P.A. (1983): *Graphical Methods for Data Analysis*. Duxbury Press, Boston.
- CHARNES A., FROME E.L. AND YU P.L., (1976): The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family. *Journal of the American Statistical Association*, **71**, 169–171.
- CLEVELAND W.S. (1979): Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.
- CLEVELAND W.S. (1981): LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. *The American Statistician*, **35**, 54.
- COLOMBO R., EHRENBERG A. AND SABVALA D. (1994): The Car Challenge: Diversity in Analysing Brand Switching Tables. *Unpublished manuscript*. (Wie bei IACOBUCCI AND HENDERSON (1997) zitiert.)

- COOK R.D. AND CROOS–DABRERA R. (1998): Partial Residual Plots in Generalized Linear Models. *Journal of the American Statistical Association*, **93**, 730–739.
- COOK R.D. AND WEISBERG S. (1982): *Residuals and Influence in Regression*. Chapman & Hall, New York, London.
- COOK R.D. AND WEISBERG S. (1994): ARES plots for generalized linear models. *Computational Statistics and Data Analysis*, **17**, 303–315.
- COX D.R. AND SNELL E.J. (1968): A General Definition of Residuals. *Journal of the Royal Statistical Society, Series B*, **30**, 248–265.
- DAVIS F.N. AND JOHNSON N.L. (1948): The Probability Integral Transform when Parameters are Estimated from the Sample. *Biometrika*, **35**, 183–190.
- DAVISON A.C. AND GIGLI A. (1989): Deviance residuals and normal scores plots. *Biometrika*, **76**, 211–221.
- DAVISON A.C. AND HINKLEY D.V. (1997): *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge et al.
- DAVISON A.C. AND TSAI C.–L. (1992): Regression Model Diagnostics. *International Statistical Review*, **60**, 337–353.
- DIEBOLD F.X., GUNTHER T.A. AND TAY A.S. (1998): Evaluating Density Forecasts. *International Economic Review*, **39**, 863–883.
- DIEBOLD F.X., HAHN J. AND TAY A.S. (1999): Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange. *Review of Economics and Statistics*, **81**, 661–673.
- DOBSON A.J. (1990): *An Introduction to Generalized Linear Models*. First edition, Chapman & Hall, London et al.
- DOBSON A.J. (2002): *An Introduction to Generalized Linear Models*. Second edition, Chapman & Hall, Boca Raton et al.
- DUNN P.K. AND SMYTH G.K. (1996): Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

- EATON M.L. (1983): Isotropic Distributions. KOTZ S. AND JOHNSON N. L. (eds.): *Encyclopedia of Statistical Sciences*, Wiley, New York, **4**, 265–267.
- ELERIAN O., CHIB S. AND SHEPARD N. (2001): Likelihood Inference for Discretely Observed Non-Linear Diffusions. *Econometrica*, **69**, 959–993.
- EMBRECHTS P., LINDSKOG F. AND MCNEIL A. (2001): Modelling Dependence with Copulas and Applications to Risk Management. *Working Paper*, Department of Mathematics ETHZ, CH-8092 Zürich, Schweiz, erhältlich unter <http://www.math.ethz.ch/~baltes/ftp/papers.html>.
- ENGLE R.F. (1982): Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, **50**, 987–1007.
- FAHRMEIR L., KAUFMANN H. UND KREDLER C. (1984): Regressionsanalyse. FAHRMEIR L. UND HAMERLE A. (Hrsg.): *Multivariate statistische Verfahren*, de Gruyter, Berlin und New York, 83–153.
- FARRINGTON C.P. (2000): Residuals for Proportional Hazard Models with Interval-Censored Survival Data. *Biometrics*, **56**, 473–482.
- FISHER N.I. AND SWITZER P. (1985): Chi-Plots for Assessing Dependence. *Biometrika*, **72**, 253–265.
- FISHER N.I. AND SWITZER P. (2001): Graphical assessment of Dependence: Is a Picture Worth than 100 Tests? *The American Statistician*, **55**, 233–239.
- GELMAN A., GOEGEBEUR Y., TUERLINCKX F. AND MECHELEN I.V. (2000): Diagnostic checks for discrete data regression models using posterior predictive simulations. *Applied Statistics*, **49**, 247–268.
- GREENACRE M.J. (1993): *Correspondence Analysis in Practice*. Academic Press, London et al.
- GRUNWALD G.K. AND JONES R.J. (2000): Markov Models for Time Series with Mixed Distribution. *Envirometrics*, **11**, 327–339.
- HADI A.S. (1992): Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society, Series B*, **54**, 761–771.

- HAFLEY W.L. AND SCHREUDER H.T. (1977): Statistical distributions for fitting diameter and height data in even-aged stands. *Canadian Journal of Forest Research*, **7**, 481–487.
- HAND D.J., DALY F., LUNN A.D., MCCONWAY K.J. AND OSTROWSKI E. (1994): *A Handbook of Small Data Sets*. Chapman & Hall, London et al.
- IACOBUCCI D. AND HENDERSON G. (1997): Log Linear Models for Consumer Brand Switching Behavior: What a Manager Can Learn from Studying Standardized Residuals. *Advances in Consumer Research*, **24**, 375–380.
- IHAKA R. AND GENTLEMAN R. (1996): R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.
- INCLAN C. AND TIAO G.C. (1994): Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance. *Journal of the American Statistical Association*, **89**, 913–923.
- JOHNSON N.L. (1949): Bivariate Distributions Based on Simple Translation Systems. *Biometrika*, **36**, 297–304.
- JOHNSON R.A. AND WICHERN D.W. (1998): *Applied Multivariate Statistical Analysis*. Fourth edition, Prentice–Hall, Upper Saddle River.
- JORION P. (1997): *Value at Risk*. McGraw–Hill, New York et al.
- JORGENSON D.W. (1961): Multiple Regression Analysis of a Poisson Process. *Journal of the American Statistical Association*, **56**, 235–245.
- JUANG B.H. (1985): Maximum–Likelihood estimation for mixture multivariate stochastic observations of Markov chains. *AT&T Technology Journal*, **64**, 1235–1249.
- JUANG B.H. AND RABINER L.R. (1991): Hidden Markov Models for Speech Recognition. *Technometrics*, **33**, 251–272.
- JUAN J. AND PRIETO F.J. (2001): Using Angles to Identify Concentrated Multivariate Outliers. *Technometrics*, **43**, 311–322.
- KENDALL M. AND STUART A. (1977): *The Advanced Theory of Statistics*. Vol. 1, Distribution Theory, Charles Griffin, London.

- KIM S., SHEPARD N. AND CHIB S. (1998): Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies*, **65**, 361–393.
- KIMBALL B.F. (1960): On the Choice of Plotting Positions on Probability Paper. *Journal of the American Statistical Association*, **55**, 546–560.
- KLEIN J.P. AND MOESCHBERGER M.L. (1997): *Survival Analysis*. Springer, New York et al.
- LANDWEHR J.M., PREGIBON D. AND SCHOEMAKER A.C. (1984): Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, **79**, 61–71. (Comments and Rejoinder, 72–83.)
- LIANG J.J. ET AL. (2000): Testing Multivariate Uniformity and its Applications. *Mathematics of Computation*, **70**, 337–355.
- LINDSEY J.K. (1993): *Models for Repeated Measurements*. Clarendon Press, Oxford.
- LINDSEY J.K. (1997): *Applying Generalized Linear Models*. Springer–Verlag, New York et al.
- LOYNES R.M. (1969): On Cox and Snell’s General Definition of Residuals. *Journal of the Royal Statistical Society, Series B*, **31**, 103–106.
- MACDONALD I.L. AND ZUCCHINI W. (1997): *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall, London et al.
- MARI D.D. AND KOTZ S. (2001): *Correlation and Dependence*. Imperial College Press, London.
- MCCULLAGH P. AND NELDER J.A. (1983): *Generalized Linear Models*. First edition, Chapman & Hall, London and New York.
- MCCULLAGH P. AND NELDER J.A. (1989): *Generalized Linear Models*. Second edition, Chapman & Hall, London et al.
- MONTGOMERY D.C. AND PECK E.A. (1992): *Introduction to Linear Regression Analysis*. Wiley & Sons, New York et al.
- MYERS R.H. AND MONTGOMERY D.C. (1997): A Tutorial on Generalized Linear Models. *Journal of Quality Technology*, **29**, 274–291.

- MYERS R.H., MONTEGOMERY D.C. AND VINING G.G. (2002): *Generalized Linear Models With Applications in Engeneering and the Sciences*. Wiley & Sons, New York.
- NARDI A. AND SCHEMPER M. (1999): New Residuals for Cox Regression and Their Application to Outlier Screening. *Biometrics*, **55**, 523–529.
- NELDER J.A. AND WEDDERBURN R.W.M. (1972): Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.
- NEUMANN K. (2000): *Zeitreihenmodelle zur Schätzung des Value at Risk von Aktien*. In: BOMS E., KÖSTERS W. UND MATTHES W. (Hrsg.): *Quantitative Ökonomie*, Band 104, Josef Eul, Lohmar und Köln.
- OAKES D. AND RITZ J. (2000): Regression in a bivariate copula model. *Biometrika*, **87**, 345–352.
- OBERLÄNDER S. (1979): Analytische Geometrie des Raumes. GELLERT W., KÜSTNER H., HELLWICH M. UND KÄSTNER H. (Hrsg.): *Kleine Enzyklopädie Mathematik*, Leipzig, 579–598.
- ORD J.K. (1983): Laplace Distribution. KOTZ S. AND JOHNSON N.L. (eds.): *Encyclopedia of Statistical Science*, Wiley & Sons, New York, **4**, 473–475.
- PEDERSEN A.R. (1994): Uniform Residuals for Discretely Observed Diffusion Processes. *Research Reports*, No. 292, August 1994, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.
- PIERCE D.A. AND SCHAFFER D.W. (1986): Residuals in Generalized Linear Models. *Journal of the American Statistical Association*, **81**, 977–986.
- PREGIBON D. (1981): Logistic Regression Diagnostics. *The Annals of Statistics*, **9**, 705–724.
- ROCKE D.M. AND WOODRUFF D.L. (1996): Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association*, **91**, 1047–1061.
- ROSENBLATT M. (1952): Remarks on a Multivariate Transformation. *Annals of Mathematical Statistics*, **23**, 470–472.

- ROSENKRANTZ W.A. (2001): Confidence Bands for Quantile Functions: A Parametric and Graphic Alternative for Testing Goodness of Fit. *The American Statistician*, **54**, 185–190.
- ROUSSEEUW P.J. AND VAN ZOMEREN B.C. (1990): Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, **85**, 633–639.
- RUTHERFORD E. AND GEIGER M. (1910): The Probability Variations in the Distribution of Alpha-Particles. *Philosophical Magazine*, Series 6, **20**, 698–704.
- RYDÉN T., TERSÄSVIRTA T. AND ÅSBRINK S. (1998): Stylized Facts of Daily Return Series and the Hidden Markov Model. *Journal of Applied Econometrics*, **13**, 217–244.
- SANGWAN-NORRELL B.S. (1977): Androgenic stimulating factor in the anther and isolated pollen grain culture of *Datura innoxia* Mill. *Journal of Experimental Biology*, **28**, 843–852.
- SEILER-MOISEWITSCH F. (1993): Sequential Probability Forecasts and the Probability Integral Transform. *International Statistical Review*, **61**, 395–408.
- Shepard N. (1996): Statistical aspects of ARCH an stochastic volatility. COX D.R., HINKLEY D.V. AND BARNDORFF-NIELSEN O.E. (eds.): *Time Series Models — In econometrics, finance and other fields*, Chapman & Hall, London et al., 1–67.
- SIEVERS G.L. (1986): Probability Plotting. KOTZ S. AND JOHNSON N.L. (eds.): *Encyclopedia of Statistical Sciences*. Wiley, New York, **7**, 232–237.
- SISON C.P. AND GLAZ J. (1995): Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions. *Journal of the American Statistical Association*, **90**, 366–369.
- STADIE A., ZUCCHINI W. AND MACDONALD I.L. (2002): Calibrating Pseudo-Residuals. *Internal Report*, 2002/05–01, Institut für Statistik und Ökonometrie, Universität Göttingen.

- TANIZAKI H. (1995): Asymptotically Exact Confidence Intervals of CUSUM and CUSUMSQ Tests: A Numerical Derivation Using Simulation Technique. *Communications in Statistics — Simulation and Computation*, **24**, 1019–1036.
- TAY A.S. AND WALLIS K.F. (2000): Density Forecasting: A Survey. *Journal of Forecasting*, **19**, 235–254.
- TUTZ G. (2000): *Die Analyse kategorialer Daten*. Oldenbourg Verlag, München und Wien.
- WALDEN A.T. AND GUTTORP P. (1992): *Statistics in the environmental & earth sciences*. Edward Arnold, London et al.
- WALKER A.M. (1968): Discussion on the paper by Professor Cox and Mrs Snell. *Journal of the Royal Statistical Society, Series B*, **30**, 272.
- WANG P.C. (1985): Adding a Variable in Generalized Linear Models. *Technometrics*, **27**, 273–276.
- WANG P.C. (1987): Residual Plots for Detecting Nonlinearity in Generalized Linear Models. *Technometrics*, **29**, 435–438.
- WANG Y.H. (1993): On the Number of Successes in Independent Trials. *Statistica Sinica*, **3**, 295–312.
- WILLIAMS D.A. (1987): Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. *Applied Statistics*, **36**, 181–191.
- WOOD C.L. (1978): Comparison of Linear Trends in Binomial Proportions. *Biometrics*, **34**, 496–504.
- ZUCCHINI W. (2000): An Introduction to Model Selection. *Journal of Mathematical Psychology*, **44**, 41–61.
- ZUCCHINI W. (2002): Generalized Residuals and their Applications. *Proceedings of the 21. International Biometric Conference*, 403–423.
- ZUCCHINI W. AND ADAMSON P.T. (1989): Bootstrap confidence intervals for design storms from exceedance series. *Hydrological Sciences—Journal—des Sciences Hydrologiques*, **34**, 41–48.

- ZUCCHINI W. AND MACDONALD I.L. (1999): Illustrations of the use of pseudo-residuals in assessing the fit of a model. FRIEDL H., BERGHOLD A. AND KAUERMANN D. (eds.): *Proceedings of the 14th International Workshop on Statistical Modelling*, Graz (Austria).
- ZUCCHINI W., SCHMIDT M. AND VON GADOW, K. (2000): A Model for the Diameter–Height Distribution in an Uneven–Aged Beech Forest and a Method to Assess the Fit of Such Models. *Silva Fennica*, **35**, 169–183.
- ZUCCHINI W. AND NEUMAN K. (2001): A comparison of several time series models for assessing the value at risk of shares. *Applied Stochastic Models in Business and Industry*, **17**, 135–148.
- ZUCCHINI W., NEYKOV N. AND NEYTCHEV P. (2001a): *Report on the Project: Development of a Daily Precipitation Model for South–West Bulgaria*. Institut für Statistik und Ökonometrie (University of Göttingen, Germany) and National Institute of Meteorology and Hydrology (Bulgarian Academy of Sciences, Bulgaria).
- ZUCCHINI W., NEYKOV N. AND NEYTCHEV P. (2001b): *Report on the Project: An Assessment of the Effects of Climate Change on Precipitation in Bulgaria*. Institut für Statistik und Ökonometrie (University of Göttingen, Germany) and National Institute of Meteorology and Hydrology (Bulgarian Academy of Sciences, Bulgaria).