Aus dem Institut und Poliklinik für Arbeits- und Sozialmedizin
Abteilung Klinische Sozialmedizin
(Ärztlicher Direktor: Prof. Dr. med. Thomas L. Diepgen)

Globale Anpassungstests im logistischen Regressionsmodell bei fehlenden Messwiederholungen

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr.sc.hum.)
der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Oliver Kuß

aus
Crailsheim
2000

Dekan: Prof. Dr. med. Sonntag
Referent: Prof. Dr. med. Thomas L. Diepgen

Inhaltsverzeichnis

1 Einleitung

1.1 Das logistische Regressionsmodell

„Das logistische Regressionsmodell hat sich seit seiner Einführung in den siebziger Jahren zu einer Standardmethode in der Biometrie und Epidemiologie entwickelt, wenn es um die Auswertung von binären Zielgrößen geht."
Mit einem Standardsatz wie diesem beginnt eine Reihe von Artikeln, die sich methodisch mit der logistischen Regression befassen. Dass diese Behauptung auch zu belegen ist, zeigt das Ergebnis einer MEDLINE-Recherche (vgl. Tabelle 1) nach Veröffentlichungen, die das Stichwort „Logistic Regression" im Abstrakt oder als Schlüsselwort enthalten. Es zeigt sich tatsächlich seit Anfang der siebziger Jahre ein nahezu exponentielles Anwachsen der Anzahl der Veröffentlichungen, die sich mit logistischer Regression beschäftigen oder in denen Daten mit Hilfe dieser Methode ausgewertet werden.

Tabelle 1: Anzahl medizinischer Fachartikel mit dem Stichwort „Logistic Regression" im Abstrakt oder als Schlüsselwort, adjustiert nach der gesamten Anzahl der Fachartikel (eigene Recherche)

Jahr	Anzahl Artikel (pro 100000)
1973	0
1978	4
1983	25
1988	110
1993	283
1998	622

Die Gründe für die Beliebtheit des logistischen Regressionsmodell sind vielfältig. Zum einen sind binäre Zielgrößen im medizinischen Bereich häufig (z.B. krank/gesund). Dazu kommt die leichte Interpretierbarkeit der geschätzten Parameter als Odds-Ratios, die Möglichkeit zu Prognosen über das Eintreten des Zielereignisses, die Verfügbarkeit von geeigneter Software und, für die Epidemiologie besonders wichtig, die Möglichkeit, das Modell zur Analyse sowohl von prospektiven als auch retrospektiven Beobachtungsstudien einzusetzen.
Aus methodischer Sicht ist noch die Tatsache zu nennen, dass das logistische

Regressionsmodell sowohl als log-lineares Modell als auch als generalisiertes lineares Modell aufgefasst werden kann, so dass mathematische Erkenntnisse aus diesen Bereichen auf das logistische Regressionsmodell übertragbar sind.

Notation

Das logistische Regressionsmodell beschreibt, ganz allgemein, den Zusammenhang zwischen einer binären Zielgröße und einer Menge von erklärenden Variablen, den so genannten *Kovariablen*. Um die Darstellung etwas zu präzisieren, wird im folgenden kurz die benötigte Notation eingeführt.
Etwas abweichend von der üblichen Notation betrachten wir die individuellen Beobachtungen bereits nach Kovariablen gruppiert. Das bedeutet, dass zwei individuelle Beobachtungen, die identische Ausprägungen bezüglich der Kovariablen haben, (wir sagen auch: das selbe Kovariablenmuster besitzen) zu einer Gruppe gehören. Unterscheiden können sich die individuellen Beobachtungen dann selbstverständlich bezüglich des Auftretens des Zielereignisses.
Gegeben sind also N Gruppen von individuellen Beobachtungen (y_i, x_i), $i = 1, \ldots, N$, wobei x_i ein Zeilenvektor von p+1 Kovariablen ist (mit einer 1 an der ersten Stelle zur Modellierung des konstanten Faktors) und y_i die beobachtete Anzahl der eingetretenen Zielereignisse für das i-te Kovariablenmuster.
Die y_i werden aufgefasst als Realisationen einer binomialverteilten Zufallsvariable Y_i mit Ereigniswahrscheinlichkeit π_i, $Y_i \sim \mathbf{B}(m_i, \pi_i)$, m_i bezeichnet somit die Anzahl der individuellen Beobachtungen, die das i-te Kovariablenmuster gemeinsam haben.
Die Gesamtanzahl der individuellen Beobachtungen bezeichnen wir mit M, d.h. $M = \sum_{i=1}^{N} m_i$, m steht für den $(N \times 1)$-Vektor der m_i. Analog stehen y, π für die Spaltenvektoren der y_i und π_i. X bezeichnet die $(N \times (p+1))$-Matrix der Kovariablen, die so genannte *Designmatrix*, die dadurch entsteht, dass alle x_i untereinander geschrieben werden.
Anschaulich kann man sich die Daten wie in Tabelle 2 dargestellt vorstellen. Liegt als Beispiel nur eine einzelne binäre Kovariable vor, so reduziert sich diese Datentabelle auf eine 2x2-Tafel. Im Falle von stetigen Kovariablen besitzt jede individuelle Beobachtung ein eigenes Kovariablenmuster , so dass alle $m_i \equiv 1$ sind und man im Innern der Tabelle ausschließlich Nullen und Einsen findet.
Um die Abhängigkeit der Ereigniswahrscheinlichkeiten von den $p + 1$ Kovariablen zu modellieren, verwenden wir die Modellgleichung

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{p} \beta_j x_{ji}\,. \tag{1}$$

Aufgelöst nach π_i erhält man

$$\pi_i = \frac{\exp(\sum_{j=0}^{p} \beta_j x_{ji})}{1+\exp(\sum_{j=0}^{p} \beta_j x_{ji})}\,. \tag{2}$$

Die Funktion, die die Ereigniswahrscheinlichkeit π_i mit den Kovariablen verknüpft (hier die Logit-Funktion $\log\left(\frac{\pi_i}{1-\pi_i}\right)$), ist die sogenannte Link-Funktion. Zur Schätzung der Parameter β_j, die die Stärke des Einflusses der jeweiligen Kovariable auf die Zielgröße beschreiben, wird in der Regel die Maximum-Likelihood-Methode verwendet. Die Likelihood-Funktion ist gegeben durch

$$L(\beta) = \prod_{i=1}^{N} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}\,. \tag{3}$$

Die Likelihood-Funktion hängt von den Ereigniswahrscheinlichkeiten π_i ab und damit aufgrund von (2) auch von β. Zur Schätzung der β bestimmt man die Werte $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, die die Likelihoodfunktion maximieren. Nach Übergang zum Logarithmus, Ableiten der logarithmierten Likelihoodfunktion nach den Parametern β_j und Nullsetzen der resultierenden p+1 Gleichungen erhält man ein nicht-lineares Gleichungssystem in den β_j, das im allgemeinen keine geschlossene Lösungsform hat, sondern numerisch gelöst werden muss. Dies geschieht iterativ mit einem IRLS-Algorithmus (iteratively reweighted least squares). Dabei wird in jedem Iterationsschritt eine Pseudo-Zielgröße $\hat{z}_i = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \frac{(y_i - m_i\hat{\pi}_i)}{(m_i\hat{\pi}_i(1-\hat{\pi}_i))}$ bestimmt und mit dieser eine gewichtete lineare Regression berechnet. Diese hat als unabhängige Variablen die

Tabelle 2: Daten als $N \times 2$-Tafel bzgl. der Zielgröße und der verschiedenen Kovariablenmuster

		Zielgröße 1	Zielgröße 0	
	1	Y_1	$m_1 - Y_1$	m_1
Kovariablen	2	Y_2	$m_2 - Y_2$	m_2
muster	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	N	Y_N	$m_N - Y_N$	m_N

Kovariablen aus dem logistischen Modell und die Gewichte $\hat{w}_i = m_i\hat{\pi}_i(1-\hat{\pi}_i)$, die bei jedem Iterationsschritt neu berechnet werden.
Nach dem Konvergieren des Algorithmus erhält man neben den geschätzten $\hat{\beta}_j$ durch Einsetzen auch Schätzer für die geschätzten Ereigniswahrscheinlichkeiten $\hat{\pi}_i$ durch

$$\hat{\pi}_i = \frac{\exp(\sum_{j=0}^{p} \hat{\beta}_j x_{ji})}{1 + \exp(\sum_{j=0}^{p} \hat{\beta}_j x_{ji})} \,. \tag{4}$$

Die Diagonalmatrix der geschätzten $\hat{w}_i$ aus dem letzten Iterationsschritt bezeichnen wir mit $\hat{W}$. Aus dieser ergibt sich ein Schätzer für die asymptotische Varianz-Kovarianz-Matrix der Parameterschätzer durch

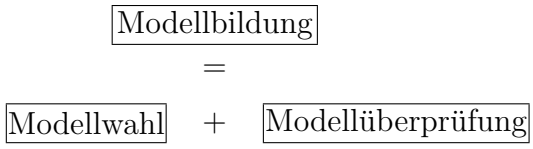
$$\mathrm{Cov}(\hat{\beta}) = (X^t \hat{W} X) \,. \tag{5}$$

1.2 Statistische Modellbildung

Die statistische Modellbildung ist nach Hosmer et al. ([27]) ein Prozess, der notwendigerweise in zwei Schritten abläuft.
Der erste Schritt betrifft die Modellwahl. Dabei wird die **systematische** Komponente des Modells festgelegt, worin beschrieben wird, wie sich die Variabilität in der Zielgröße über die Beobachtungen durch andere an diesen Beobachtungen gemessenen Kovariablen erklären lässt. Dabei gilt es, die für das vorliegende Problem (statistisch und inhaltlich) relevanten Kovariablen aus allen vorliegenden bzw. gemessenen auszuwählen und deren funktionale Form festzulegen.
Der zweite Schritt betrifft die Überprüfung der Modellgüte. Dabei wird die **Fehlerkomponente** des Modells im Hinblick auf verbleibende, nicht durch die systematische Komponente erklärte Variabilität in der Zielgröße untersucht. Hierbei wird das Ausmaß der Abweichung der **beobachteten** Werte der Zielgröße von den **erwarteten** Werten, die von der systematischen Komponente des Modells vorhergesagt wurden, beurteilt. In einem „guten“ Modell wird die systematische Komponente alle nicht-zufällige Variabilität in der Zielgröße beschreiben, so dass wir bei der Untersuchung der Fehlerkomponente nur noch unsystematische, zufällige Abweichungen zwischen beobachteten und erwarteten Werten der Zielgröße finden.
Zusammenfassend beschreibt die systematische Komponente eines Modells den „mittleren“ Wert der Zielgröße für eine gegebene Wertemenge der Kovariablen , die Fehlerkomponente beschreibt die Abweichung von diesem „mittleren“ Wert.

Modellbildung im Sinne von Hosmer et al. könnte man in folgendem Diagramm zusammenfassen:

Abbildung 1: Der Modellbildungsprozess nach Hosmer et al. ([27])

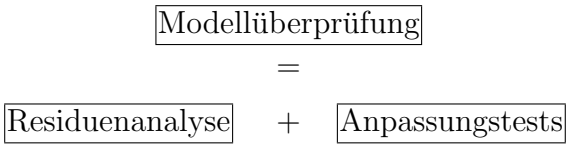


Wichtig dabei ist, dass Modellwahl und Modellbildung gleichberechtigt nebeneinander stehen und die Modellbildung kein lästiges Anhängsel der Modellwahl ist.

1.3 Überprüfung der Modellgüte

Der Vergleich zwischen beobachteten und erwarteten Werten der Zielgröße als zentrales Moment der Modellüberprüfung kann auf zwei grundsätzlich verschiedene Arten geschehen: einerseits bei Betrachtung der Abweichungen auf der Ebene der Beobachtungen (Residuenanalyse), andererseits durch Berechnung von Anpassungsmaßen, die die Modellgüte mit Hilfe einer einzelnen Zahl messen und das Ausmaß einer schlechten Modellanpassung mit Hilfe statistischer Tests beurteilen.

Abbildung 2: Der Prozess der Modellüberprüfung



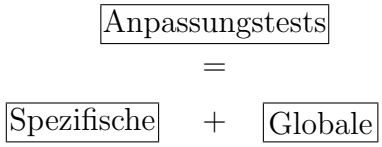
Die Entwicklung dieser Methoden für das logistische Regressionsmodell hinkt denen des linearen Regressionsmodells etwas hinterher, die Gründe dafür sind klar: Zum einen ist das logistische Modell mathematisch komplizierter, zum anderen blickt das lineare Regressionsmodell auf eine längere Entwicklungszeit zurück. Desweiteren erbringt eine direkte Anwendung bzw. Übert-

ragung von Techniken, die man aus der linearen Regression kennt, nicht notwendigerweise sinnvolle und nützliche Analysemethoden für das logistische Modell. Dies ist vor allem auf die Verteilung der Residuen im logistischen Modell zurückzuführen, die ungleich komplizierter ist als die der Residuen im linearen Modell [29].

1.4 Anpassungstests

Auch die Anpassungstests lassen sich, analog zur Modellbildung und Modellüberprüfung, in zwei große Komplexe unterteilen: in die spezifischen und die globalen Anpassungstests.

Abbildung 3: Typen von Anpassungstests



Die spezifischen Anpassungstests betten das zu überprüfende Modell in eine allgemeinere Familie von Modellen ein, wobei ersteres aus letzterer durch eine Reihe von Parameterrestriktionen hervorgeht. Ein Beispiel dafür ist eine Familie von Regressionsmodellen für binäre Zielgrößen mit einer sehr allgemeinen Linkfunktion (Pregibon, [47]):

$$g(\pi, \lambda) = \log \frac{(1/(1-\pi))^{\lambda} - 1}{\lambda}\,, \tag{6}$$

die für $\lambda = 1$ das logistische Regressionsmodell mit Logit-Link enthält. Somit liefert ein Test mit Nullhypothese: $\lambda = 1$ und Alternative: $\lambda \neq 1$ einen spezifischen Anpassungstest für das logistische Modell.
Demgegenüber sind globale Anpassungstests solche, die ganz allgemein auf Unzulänglichkeiten in der Anpassung des Modells hinweisen, ohne aber die Art der Abweichung genauer zu spezifizieren, d.h. sie prüfen die unspezifische Nullhypothese: „Das geschätzte Modell ist richtig“. Die Alternativhypothese bleibt ebenfalls unspezifisch, d.h. sie lautet einfach: „Die Nullhypothese ist falsch“, die Art der Abweichung von der Nullhypothese wird dabei also nicht näher spezifiziert.

Das Fehlen einer spezifischen Alternativhypothese bringt sowohl Vorteile als auch Nachteile für globale Anpassungstests mit sich. Es liegt in der Natur der globalen Anpassungstests, dass wir im Falle der Ablehnung der Nullhypothese wenig Hilfe zur Neuformulierung des Modells erhalten. Jedoch leisten dies die spezifischen Tests im allgemeinen auch nicht. Spezifische Anpassungstests haben sehr oft auch beträchtliche Power gegen Alternativen, für die sie gar nicht entwickelt wurden. Eine Ablehnung der Nullhypothese bei einem spezifischen Anpassungstest weist also nicht notwendig darauf hin, dass genau diejenige Modellverletzung vorliegt, die mit Hilfe der Alternative abgeprüft worden ist.
Ein weiteres, in eine ähnliche Richtung gehendes Argument gegen spezifische Anpassungstests liefern Alston/Chalfant([3]): Diese Tests sind alle unter der Annahme hergeleitet, dass *ausschließlich* die in der Alternative spezifizierte Modellverletzung vorliegt und deshalb auch nur gültig, wenn alle anderen Aspekte der Modellanpassung korrekt sind. Dies ist in der Praxis eine unwahrscheinliche Situation, da gänzlich isolierte Fehlspezifikationen selten vorkommen werden. Auch hier gilt also, dass eine abgelehnte Nullhypothese eines spezifischen Anpassungstests keine verlässliche Information über die konkrete Art der Modellverletzung liefert.
Andererseits können globale Anpassungstests auch sehr wohl Power gegen spezifische Alternativen haben. Der Informationsmatrix-Test, ein globaler Test, der zwei unter korrekter Modellspezifikation äquivalente Schätzer der Kovarianzmatrix vergleicht, besitzt die gleiche Teststatistik wie ein Score-Test auf Parameterhomogenität, d.h. auf konstante β über die Beobachtungen. Der augenscheinlich globale Anpassungstests mit unspezifischer Alternative hat also sehr wohl Power, eine ganz konkrete Modellverletzung (so sie isoliert vorliegt) zu erkennen.
Ein weiterer Nachteil spezifischer Anpassungstests ist der, dass eine Schätzung der notwendigerweise komplizierteren Modelle, die eine Verallgemeinerung des logistischen Modells darstellen, nötig wird (zumindest solange nicht nur der Score-Test berechnet wird). Dies wird nicht immer trivial sein, erst recht kann man nicht erwarten, dass Standard-Software dafür zur Verfügung steht. Schon die ML-Schätzer des einfachen logistischen Modells zu berechnen, verlangt bereits iterative Verfahren.

1.5 Klassische globale Anpassungstests im logistischen Regressionsmodell

Zwei Klassiker unter den globalen Anpassungsmaßen sind die Pearson-Statistik

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \ , \tag{7}$$

und die Devianz

$$D = 2\sum_{i=1}^{N} y_i \log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i(1-\hat{\pi}_i)}\right) . \tag{8}$$

Beide vergleichen, wie aus den Definitionen ersichtlich ist und wie es von Hosmer et al. ([27], vgl. 1.2) gefordert wird, beobachtete, y_i, und erwartete, $m_i\hat{\pi}_i$, Werte der Zielgröße und beurteilen so die Güte der Anpassung des Modells an die Daten.
Die Pearson-Statistik X^2 geht zurück auf Karl Pearson, der das Prinzip des Vergleichs zwischen beobachteten und geschätzten Häufigkeiten bereits 1900 ([44]) vorgeschlagen hatte. Der wissenschaftliche Wert dieser Veröffentlichung wird als extrem hoch eingeschätzt, sie wird z.B. als der Durchbruch zur modernen Statistik des 20. Jahrhunderts angesehen ([7]) oder in einer Aufzählung der 20 wichtigsten wissenschaftlichen Veröffentlichungen des 20. Jahrhunderts ([24]) neben Arbeiten zur Relativitätstheorie oder zur Quantenmechanik genannt. Es ist beachtenswert, dass ausgerechnet einer Arbeit zur Überprüfung der Anpassungsgüte eines Modells soviel Beachtung geschenkt wird.
Die Devianz D ist ebenfalls aus anderen Bereichen der Statistik wohl bekannt. Sie misst allgemein den Abstand zwischen zwei Modellen bzgl. deren Likelihoodfunktion. Im logistischen Regressionsmodell beschreibt sie dabei konkret den Abstand zwischen dem geschätzten Modell und dem saturierten Modell, also dem Modell, das genauso viele Parameter wie Beobachtungen hat und eine perfekte Anpassung an die Daten liefert. Ist dieser Abstand zu groß, dann beschreibt das vorliegende Modell die Daten nur unzureichend. Dieser Test ist von der Likelihood-Ratio-Statistik im logistischen Modell zu unterscheiden, die auch häufig als „Devianz“ bezeichnet wird, aber den Abstand zwischen einem Modell mit ausschließlich konstanten Term und dem geschätzten Modell misst.

Große Werte von X^2 und D weisen auf eine schlechte Modellanpassung hin. Um die statistische Signifikanz dieser Modellanpassung zu beurteilen, vergleicht man X^2 und D mit dem Quantil einer χ^2-Verteilung mit $N - p - 1$ Freiheitsgraden, d.h. Devianz und Pearson-Statistik sind asymptotisch äquivalent.

1.6 Das Problem fehlender Messwiederholungen

Ein großes und wohlbekanntes Problem dabei ist, dass die Gültigkeit der asymptotischen Prüfverteilung von X^2 und D wesentlich von der Annahme von Messwiederholungen abhängt (N fest, m_i groß für alle i). Das heisst, für jedes beobachtete Kovariablenmuster müssen hinreichend viele Beobachtungen vorliegen. Diese Annahme ist sicher unrealistisch bei einer großen Anzahl von Kovariablen oder bei stetigen Kovariablen, wo im Extremfall jedes Kovariablenmuster nur einmal besetzt ist ($m_i \equiv 1$). In diesem Fall degeneriert die Devianz zu

$$D = 2 \sum_{i=1}^{N} \hat{\pi}_i \log \left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + \log(1 - \hat{\pi}_i) \,, \tag{9}$$

so dass diese unabhängig von den beobachteten Zielereignissen y_i ist und also keinerlei Information mehr über die Anpassungsgüte enthält (vgl. [38], S. 121).
Auch die Pearson-Statistik X^2 hat im Fall $m_i \equiv 1$ nur noch begrenzten Wert (vgl. [38], S. 121). Angenommen, die Beobachtungen y_i wären identisch $\mathbf{B}(1, \pi)$ verteilt, dann wäre $\hat{\pi} = \bar{y}$ und

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{\bar{y}(1 - \bar{y})} = N \,, \tag{10}$$

also gleich der Anzahl der Beobachtungen, woraus sich ebenfalls keine sinnvolle Aussage zur Anpassungsgüte des Modells ergibt. Tatsächlich sind die y_i im logistischen Modell mit $m_i \equiv 1$ natürlich $\mathbf{B}(1, \pi_i)$ verteilt, so dass jede Beobachtung eine eigene Verteilung besitzt, aber es zeigt sich in den Simulationsuntersuchungen, dass sich in diesem Fall zwar nicht $X^2 = N$, aber doch $X^2 \approx N$ ergibt.
Das Problem der fehlenden Messwiederholungen und die Ungültigkeit der asymptotischen χ^2-Verteilung von X^2 und D in dieser Situation sind seit

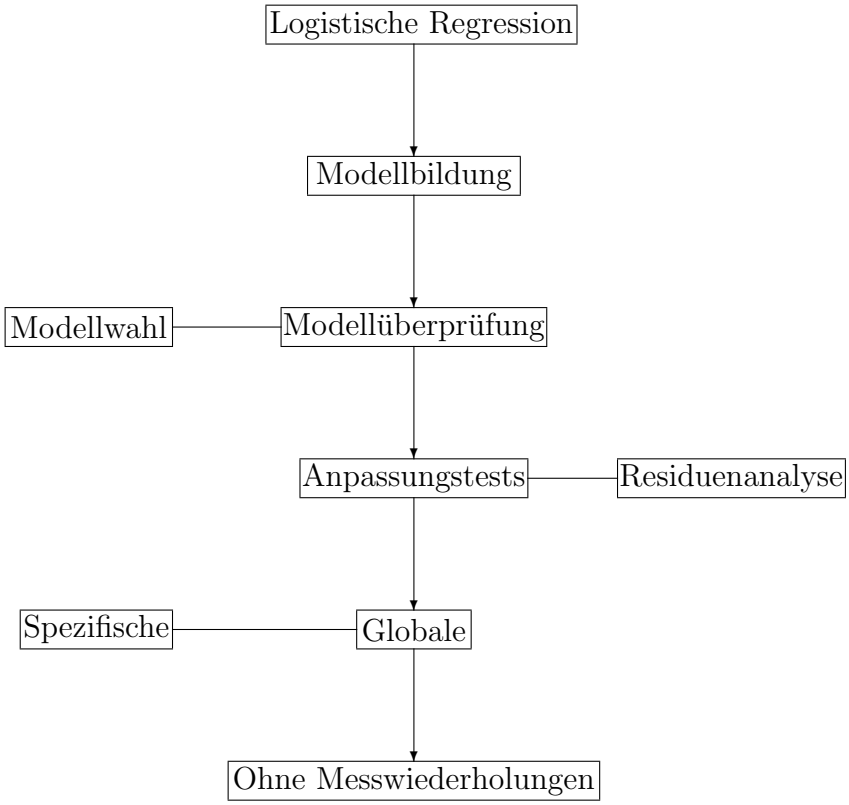
langem bekannt und haben auch ihren Weg in die einschlägigen Lehrbücher gefunden:

- The X^2 and D goodness-of-fit statistics do *not* have approximate chi-squared distributions when applied to logistic regression models with a continuous covariate, unless there are many observations at each level of the covariate. (Agresti, [1])
- Neither X^2 nor D is appropriate in the many strata standard asymptotic model (Anm.: p fest, N und $M \to \infty$), because under this model there is no χ^2_{N-p-1} limiting distribution. (Santner/Duffy, [49])
- Thus, p-values calculated for X^2 and D when $M \approx N$, using the χ^2_{N-p-1} distribution, are incorrect. (Hosmer/Lemeshow, [26])
- The effect of sparseness (Anm.: d.h. kleine m_i) is noticed mainly on D and X^2, which fail to have the properties required for goodness-of-fit statistics. (McCullagh/Nelder, [38])

Teilweise werden dort auch bereits Lösungsvorschläge für Datensätze mit fehlenden Messwiederholungen gemacht, diese sind aber uneinheitlich und widersprechen sich sogar in manchen Fällen:

- In principle it would seem preferable to accept the failure of the chi-square limit and to use a more accurate approximation to the null distribution without accumulating cells. (Lloyd, [33])
- Thus, to analyze lack of fit when explanatory variables are continuous, we apply goodness-of-fit statistics and related residual measures by grouping observed and fitted values for a partition of the space of explanatory variable values. (Agresti, [1])
- It is good statistical practice, however, not to rely on either D or X^2 as an absolute measure of goodness of fit in these circumstances. It is much better to look for specific deviations from the model of a type that is easily understood scientifically. (McCullagh/Nelder, [38])

Abbildung 4: Darstellung der Problemstruktur



Die oben stehende Abbildung macht noch einmal deutlich, wie das vorgestellte Problem in den Kontext der logistischen Regressionsmodelle einzuordnen ist.
Die vorliegende Arbeit soll einen Beitrag zur Modellüberprüfung in logistischen Regressionsmodellen liefern. Der Gegenstand der Betrachtung sind globale Anpassungstests im Falle von fehlenden Messwiederholungen. Nicht betrachtet werden Aspekte der Modellwahl (z.B. Variablenselektion) und der Modellüberprüfung auf Beobachtungsebene (Residuenanalyse).

2 Globale Anpassungstests im logistischen Regressionsmodell bei fehlenden Messwiederholungen

Die Vorschläge zur Lösung des Problems der globalen Anpassungstests im logistischen Modell bei fehlenden Messwiederholungen sind zahlreich und z.T. auch schon relativ alt, haben aber bisher nur in Ausnahmefällen den Weg aus der mathematischen Statistik in die angewandte medizinische Statistik gefunden.
Die tatsächliche Anwendung dieser Verfahren geht alles in allem nur selten über den Datensatz hinaus, der in der methodischen Originalveröffentlichung als Beispiel herangezogen wird. Dabei benutzen viele Autoren bereits vorliegende, von anderen veröffentlichte Datensätze, was verständlich ist, da diese Autoren als Mathematiker bzw. theoretische Statistiker nur selten mit realen Datensätzen in Berührung kommen. Das hat zwar den Vorteil, dass diese Datensätze allen, die auf diesem Gebiet arbeiten, bekannt sind, dadurch ist aber praktisch ausgeschlossen, dass diese Beispielanalysen irgendwelche medizinischen Entscheidungen beeinflussen.
In Anlehnung an McCullagh ([35]) kann man diese Lösungsversuche in drei Gruppen unterteilen. Diese Gruppierung deutet sich auch schon bei den auf S.13 beschriebenen Lösungsvorschlägen an.

1. **Modifikation der Prüfverteilung:** Dabei werden die beiden Teststatistiken X^2 und D als Prüfgrößen beibehalten und nur andere Prüfverteilungen benutzt, um die statistische Signifikanz zu beurteilen.

2. **Gruppierung von Beobachtungen:** Dabei gruppiert man die vorhandenen individuellen Beobachtungen, die bezüglich der ins Modell aufgenommenen Kovariablen nicht genügend Messwiederholungen pro Kovariablenmuster ergeben, neu.

   Dies kann geschehen, indem man die neuen Gruppen bzgl. der Kovariablen bildet, wodurch man die Möglichkeit erhält, diese inhaltlich sinnvoll und zielgerichtet zu definieren. Dies würde z.B. dann Sinn machen, wenn in einer epidemiologischen Studie ein spezieller Risikofaktor untersucht werden soll und die anderen Kovariablen nur zur Adjustierung ins Modell aufgenommen worden sind. Dann würde es nahe liegen, die

neuen Kovariablenmuster in Abhängigkeit von diesem speziellen Risikofaktor zu bilden.

Eine zweite Möglichkeit ist, die neuen Gruppen datenabhängig zu definieren. Dabei werden die Beobachtungen bezüglich ihrer geschätzten Wahrscheinlichkeiten $\hat{\pi}_i$ geordnet und entweder in Klassen mit a priori fest gelegten Klassengrenzen eingeteilt oder so gruppiert, dass sich gleich stark besetzte neue Gruppen ergeben.

3. **Verwendung anderer Teststatistiken:** Darunter fällt eine Reihe von Anpassungstests, die entweder Modifikationen von X^2 und D darstellen oder vollständig neue Prinzipien verwenden, um die Güte des Modells zu messen.

Illusorisch ist es zu versuchen, die exakte Verteilung der beiden Statistiken X^2 und D zu berechnen. Dies ist nur in simpelsten Spezialfällen möglich, z.B. für die ersten beiden Momente von X^2 im Modell für Unabhängigkeit in einer Vierfeldertafel, d.h. also für ein logistisches Modell mit einer einzigen binären Kovariable ([38]). Dann werden wir allerdings auch nicht mehr das Problem fehlender Messwiederholungen haben.
Im folgenden werden nun die gegenwärtig vorgeschlagenen Anpassungstests für das logistische Regressionsmodell, gruppiert nach den oben beschriebenen Testprinzipien, dargestellt.

2.1 Modifikation der Prüfverteilung

2.1.1 Asymptotische Normalverteilung

Intuitiv ist klar, dass die Prüfverteilung von Devianz und Pearson-Statistik für $N, M \to \infty$ gegen eine Normalverteilung konvergiert: für festes N ist die Prüfverteilung eine χ^2-Verteilung mit Freiheitsgraden proportional zu N und bei wachsender Anzahl der Freiheitsgrade strebt eine χ^2-Verteilung gegen eine Normalverteilung. Überraschend ist allerdings, dass in diesem Fall Devianz und Pearson-Statistik nicht mehr asymptotisch äquivalent sind, sondern gegen verschiedene Normalverteilungen konvergieren.
Unter leicht unterschiedlichen Voraussetzungen ist die asymptotische Normalverteilung als Prüfverteilung von mehreren Autoren hergeleitet worden ([42], [15], [57]), wobei die Darstellung hier von der Herleitung von Osius/Rojek ([42]) ausgeht. Zentraler Punkt ist dabei, dass die Anzahl der Kovariablenmuster nicht mehr länger fest bleiben muss, wenn die Anzahl der

individuellen Beobachtungen gegen unendlich geht (N fest für $m_i \to \infty$). Zur Erinnerung, unter dieser Bedingung war die herkömmliche χ^2-Verteilung für die Pearson-Statistik hergeleitet worden. Osius/Rojek (und auch die anderen Autoren) erlauben dagegen, dass die Anzahl der Kovariablenmuster mit der Anzahl der individuellen Beobachtungen wächst ($N, m_i \to \infty$).
Osius/Rojek leiten die asymptotische Prüfverteilung allgemeiner im Kontext der Power-Divergenz-Familie von Cressie/Read ([14]) her. Diese Familie von Anpassungstests lässt sich im logistischen Regressionsmodell in Abhängigkeit von einem Parameter λ schreiben als

$$CR_\lambda = \sum_{i=1}^{N} m_i \left[a_\lambda \left(\frac{Y_i}{m_i}, \hat{\pi}_i \right) + a_\lambda \left(1 - \frac{Y_i}{m_i}, 1 - \hat{\pi}_i \right) \right] , \tag{11}$$

mit

$$a_\lambda(\Pi, \pi) = \frac{2\Pi}{\lambda(\lambda+1)} \left[\left(\frac{\Pi}{\pi} \right)^\lambda - 1 \right] - \frac{2}{\lambda+1}(\Pi - \pi) , \tag{12}$$

wobei $-1 < \lambda < \infty$ ([42]). Der Fall $\lambda = 0$ wird dabei durch einen stetigen Übergang in λ definiert. Die Einschränkung $\lambda > -1$ kommt daher, dass $Y_i = 0$ bzw. $Y_i = m_i$ erlaubt sein soll, d.h. es sollen Kovariablenmuster vorkommen dürfen, bezüglich denen keine bzw. ausschließlich Zielereignisse beobachtet wurden.
Für $\lambda = 1$ ergibt sich die Pearson-Statistik und für $\lambda = 0$ die Devianz. Read/Cressie ([48]) schlagen aus mehreren Gründen die Verwendung der Statistik mit $\lambda = 2/3$ vor (allerdings nicht explizit für das logistische Modell).

In geschlossener Form ist die als neue Prüfverteilung zu verwendende asymptotische Normalverteilung bzw. deren Momente allerdings nur für die Pearson-Statistik herzuleiten, auch dann aber noch mit beträchtlichem Aufwand. Die ersten beiden Momente der asymptotischen Normalverteilung der Pearson-Statistik sind dann zu schätzen durch

$$\hat{\mathrm{E}}_O(X^2) = N \tag{13}$$

und

$$\begin{aligned} \widehat{\mathrm{Var}}_O(X^2) &= \left[\sum_{i=1}^{N} 2 + \frac{1}{m_i} \left(\frac{1}{\hat{\pi}_i(1-\hat{\pi}_i)} - 6 \right) \right] - \\ &\quad (X^t(\mathbf{1} - 2\hat{\pi}))^t (X^t \hat{W} X)^{-1} (X^t(\mathbf{1} - 2\hat{\pi})) . \end{aligned} \tag{14}$$

Ein formaler Signifikanztest kann durchgeführt werden, indem man die standardisierte Statistik

$$X_O^2 = \frac{X^2 - \hat{\mathrm{E}}_O(X^2)}{\widehat{\mathrm{Var}}_O(X^2)^{1/2}} \tag{15}$$

mit dem Quantil der Standardnormalverteilung vergleicht.
Zusätzlich zu den ersten beiden Momenten $\mathrm{E}_O(X^2)$und $\mathrm{Var}_O(X^2)$ können auch noch höhere Momente bzw. Kumulanten, d.h. Maße für Schiefe und Wölbung der Verteilung von X^2 geschätzt werden. Einen formalen Test erhält man dann durch die Edgeworth-Expansion [1]

$$\begin{aligned} P(X^2 \geq z) &= 1 - \Phi(z) + \phi(z)[\frac{1}{6}\rho_{3O} H_2(z) + \frac{1}{24}\rho_{4O} H_3(z) \\ &\quad + \frac{1}{72}\rho_{3O}^2 H_5(z)]\,. \end{aligned} \tag{16}$$

$\Phi(z)$ und $\phi(z)$ bezeichnen Verteilungsfunktion und Dichte der Standardnormalverteilung, $H_2(x) = x^2 - 1, H_3(x) = x^3 - 3x$ und $H_5(x) = x^5 - 10x^3 + 15x$ sind die Hermiteschen Polynome.
Zur Schätzung der Schiefe ρ_{3O} und der Wölbung ρ_{4O} bedarf es einiger Zwischenschritte. Zunächst werden Vektoren $\hat{\mu}_{xS}$ berechnet:

$$\hat{\mu}_{3S} = \frac{1 - 2/\hat{\pi}}{\sqrt{m\hat{\pi}(1-\hat{\pi})}}\,, \tag{17}$$

$$\hat{\mu}_{4S} = 3\left(1 - \frac{2}{m}\right) + \frac{1}{m\hat{\pi}(1-\hat{\pi})}\,, \tag{18}$$

$$\hat{\mu}_{5S} = \hat{\mu}_{3S}\left[2\left(5 - \frac{6}{m}\right) + \frac{1}{m\hat{\pi}(1-\hat{\pi})}\right]\,, \tag{19}$$

$$\hat{\mu}_{6S} = 5\left(3 - \frac{26}{m} + \frac{24}{m^2}\right) + \frac{5(5-6/m)}{m\hat{\pi}(1-\hat{\pi})} + \frac{1}{(m\hat{\pi}(1-\hat{\pi}))^2}\,, \tag{20}$$

$$\hat{\mu}_{7S} = \hat{\mu}_{3S}\Big[3\left(35 - \frac{154}{m} + \frac{120}{m^2}\right) + \frac{4(14-15/m)}{m\hat{\pi}(1-\hat{\pi})}$$

[1] Eine Edgeworth-Expansion ist eine Transformation, die eine Verteilung, von der man nur die Momente bzw. Kumulanten kennt, in Abhängigkeit von der Standardnormalverteilung darstellt. Die Verwendung einer solchen Edgeworth-Expansion im vorliegenden Fall ist umstritten ([15]), da diese die Diskretheit der asymptotischen Prüfverteilung nicht beachtet, sondern die zu approximierende Verteilung als stetig voraussetzt. Dies kann in manchen Fällen zu negativen p-Werten führen.

$$+\frac{1}{(m\hat{\pi}(1-\hat{\pi}))^2}\Big]\,, \tag{21}$$

$$\hat{\mu}_{8S} = 7\Big[\Big(15-\frac{340}{m}+\frac{1044}{m^2}-\frac{720}{m^3}\Big)+\frac{2(35-154/m+120/m^2)}{m\hat{\pi}(1-\hat{\pi})} +\frac{17-18/m}{(m\hat{\pi}(1-\hat{\pi}))^2}\Big]+\frac{1}{(m\hat{\pi}(1-\hat{\pi}))^3}\,, \tag{22}$$

wobei m, $\hat{\pi}$ Vektoren mit den Elementen m_i, $\hat{\pi}_i$ aus dem ursprünglichen logistischen Regressionsmodell sind. Multiplikation und Division mit diesen Vektoren ist hier elementweise durchzuführen. Man beachte auch den zweitletzten Term in der Darstellung von $\hat{\mu_{8S}}$, der in der Originalveröffentlichung ([43]) irrtümlich mit $(17-18/m)\times(m\hat{\pi}(1-\hat{\pi}))^2$ angegeben war und nach Rücksprache mit Prof. Osius verbessert wurde.
Mit dem Vektor $d_O=-(X^t(\mathbf{1}-2\hat{\pi}))^t(X^t\hat{W}X)^{-1}X^t$ findet man die Vektoren

$$\hat{\mu}_2 = \hat{\mu}_{4S}-1+2d_O\sqrt{m\hat{\pi}(1-\hat{\pi})}\hat{\mu}_{3S}+d_O^2m\hat{\pi}(1-\hat{\pi})\,, \tag{23}$$

$$\hat{\mu}_3 = (\hat{\mu}_{6S}-3\hat{\mu}_{4S}+2)+3d_O\sqrt{m\hat{\pi}(1-\hat{\pi})}(\hat{\mu}_{5S}-2\hat{\mu}_{3S}) +3d_O^2m\hat{\pi}(1-\hat{\pi})(\hat{\mu}_{4S}-1)+d_O^3(m\hat{\pi}(1-\hat{\pi}))^{3/2}\hat{\mu}_{3S}\,, \tag{24}$$

$$\hat{\mu}_4 = (\hat{\mu}_{8S}-4\hat{\mu}_{6S}+6\hat{\mu}_{4S}-3)+4d_O\sqrt{m\hat{\pi}(1-\hat{\pi})}(\hat{\mu}_{7S}-3\hat{\mu}_{5S}+3\hat{\mu}_{3S}) +6d_O^2m\hat{\pi}(1-\hat{\pi})(\hat{\mu}_{6S}-2\hat{\mu}_{4S}-2\hat{\mu}_{2S}+1) +4d_O^3(m\hat{\pi}(1-\hat{\pi}))^{3/2}(\hat{\mu}_{5S}-\hat{\mu}_{3S})+d_O^4(m\hat{\pi}(1-\hat{\pi}))^2\hat{\mu}_{4S}\,, \tag{25}$$

wobei wie oben das Potenzieren des Vektors d_O und die Multiplikation mit m und $\hat{\pi}$ elementweise durchzuführen ist.
Schließlich erhält man als Schätzer für ρ_{3O} und ρ_{4O} in (16) durch

$$\hat{\rho}_{3O}=\frac{\sum_{i=1}^N\hat{\mu}_3}{\Big(\sum_{i=1}^N\hat{\mu}_2\Big)^{3/2}} \tag{26}$$

und

$$\hat{\rho}_{4O}=\frac{\sum_{i=1}^N(\hat{\mu}_4-3\hat{\mu}_2^2)}{\Big(\sum_{i=1}^N\hat{\mu}_2\Big)^{2}}\,. \tag{27}$$

Es besteht noch eine dritte Möglichkeit, im vorliegenden Fall einen Signifikanztest für die Pearson-Statistik anzugeben. Dabei wird als Prüfverteilung eine skalierte χ^2-Verteilung verwendet, wodurch eine Art Kompromiss

zwischen der herkömmlichen χ^2-Verteilung und der asymptotischen Normalverteilung erreicht wird. Osius/Rojek schreiben diesen Vorschlag einem unbekannten Reviewer ihrer Arbeit zu, jedoch lässt sich dieser Vorschlag bis mindestens zu Cox/Hinkley ([13], S. 463) zurückverfolgen.
Die Idee dabei ist, eine Normalverteilung $N(\mu, \sigma^2)$ durch eine skalierte χ^2-Verteilung $\beta\chi^2_\nu$ mit gleichem Erwartungswert $\mu = \beta\nu$ und gleicher Varianz $\sigma^2 = 2\beta^2\nu$ darzustellen. Dann hat die skalierte Pearson-Statistik

$$X^2_{OSkal} = \frac{X^2}{\sigma^2/2\mu} \tag{28}$$

eine approximative χ^2-Verteilung mit $\nu = 2\mu/\sigma^2 - p - 1$ Freiheitsgraden (die in der Regel nicht mehr ganzzahlig sind). Für μ und σ^2 sind jeweils die Schätzer aus (13) und (14) einzusetzen.

2.1.2 Bedingte asymptotische Normalverteilung

McCullagh schlägt in einer Reihe von Arbeiten ([35],[36],[37]) vor, als Prüfverteilungen für Pearson-Statistik und Devianz die auf die Parameterschätzer $\hat{\beta}$ bedingten Verteilungen zu verwenden. Die Bedingung auf die Parameterschätzer eliminiert die Abhängigkeit der Teststatistiken von den unbekannten Parametern β. Dies hat den Vorteil, dass dadurch berücksichtigt wird, dass die $\hat{\pi}_i$, auf denen die Berechnung von X^2 und D basiert, geschätzt und nicht einfach als bekannt vorausgesetzt werden.
Eine Berechnung der Prüfverteilungen ist mit vernünftigem Aufwand allerdings auch hier nur für die Pearson-Statistik möglich, da X^2 und $\hat{\beta}$ asymptotisch (in N) unkorreliert sind. Dies gilt dagegen nicht für die Devianz, wo die Korrelation zwischen D und $\hat{\beta}$ auch asymptotisch nicht verschwindet. Somit erhalten wir eine weitere Erklärung für die Tatsache, die wir bereits im vorangegangenen Kapitel beobachtet hatten (vgl. (9)): Im Extremfall von ausschließlich einfach besetzten Kovariablenmustern hängt die Devianz nur von den Parametern ab und enthält keinerlei Information über die Anpassungsgüte des Modells mehr.
Tatsächlich ableitbar sind hier, ähnlich wie in 2.1.1, nicht die kompletten bedingten Verteilungen, sondern nur deren Momente. Zur Signifikanzprüfung bedient man sich dann wieder einer Normal-Approximation oder einer Edgeworth-Expansion.
Wir berechnen zunächst $\hat{s}^2_{McC}$ und $\hat{\gamma}_{McC}$ durch

$$\hat{s}^2_{McC} \;=\; U^t_{McC}(\hat{W} - \hat{W}X(X^t\hat{W}X)^{-1}X^t\hat{W})U_{McC} \tag{29}$$

$$\hat{\gamma}_{McC} = X(X^t\hat{W}X)^{-1}X^t\hat{W}U_{McC}\,, \tag{30}$$

wobei U_{McC} ein $(N \times 1)$-Vektor mit charakteristischem Element $U_{McCi} = \frac{1-2\hat{\pi}_i}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}$ ist.
Die ersten drei bedingten Kumulanten der Pearson-Statistik werden dann geschätzt durch

$$\hat{\mathrm{E}}(X^2|\hat{\beta}) = N - p - 1 - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\hat{\kappa}_{4i}}{\hat{\kappa}_{2i}} - \hat{\gamma}_{McCi}\hat{\kappa}_{3i}\right)\hat{Q}_{ii} \tag{31}$$

$$\widehat{\mathrm{Var}}(X^2|\hat{\beta}) = (1 - \frac{p+1}{N})\left[2\sum_{i=1}^{N}\frac{m_i - 1}{m_i} + \hat{s}^2_{McC}\right] \tag{32}$$

$$\begin{aligned}\hat{\kappa}_3(X^2|\hat{\beta}) = {} & (1 - \frac{p+1}{N})\Big(8\sum_{i=1}^{N}\frac{m_i - 3}{m_i} + 4\sum_{i=1}^{N}\frac{\hat{\kappa}^2_{3i}}{\hat{\kappa}^3_{2i}} + \sum_{i=1}^{N}\frac{\hat{\kappa}_{6i}}{\hat{\kappa}^3_{2i}} \\ & +18\hat{s}^2_{McC} - 3\sum_{i=1}^{N}\frac{\hat{\kappa}_{5i}}{\hat{\kappa}^2_{2i}}\hat{\gamma}_{McCi} + 3\sum_{i=1}^{N}\hat{\gamma}^2_{McCi}\frac{\hat{\kappa}_{4i}}{\hat{\kappa}_{2i}} \\ & -\sum_{i=1}^{N}\hat{\gamma}^3_{McCi}\kappa_{3i}\Big),\end{aligned} \tag{33}$$

mit $\hat{Q} = X(X^t\hat{W}X)^{-1}X^t$ und den $\hat{\kappa}_{xi}$ als geschätzten Beiträge des i-ten Kovariablenmusters zu den Kumulanten einer Binomialverteilung. Diese erhält man aus

$$\hat{\kappa}_{1i} = m_i\hat{\pi}_i\,, \tag{34}$$

$$\hat{\kappa}_{2i} = m_i\hat{\pi}_i(1 - \hat{\pi}_i)\,, \tag{35}$$

$$\hat{\kappa}_{3i} = m_i\hat{\pi}_i(1 - \hat{\pi}_i)(1 - 2\hat{\pi}_i)\,, \tag{36}$$

$$\hat{\kappa}_{4i} = m_i(\hat{\pi}_i - 7\hat{\pi}_i^2 + 12\hat{\pi}_i^3 - 6\hat{\pi}_i^4)\,, \tag{37}$$

$$\hat{\kappa}_{5i} = m_i(\hat{\pi}_i - 15\hat{\pi}_i^2 + 50\hat{\pi}_i^3 - 60\hat{\pi}_i^4 + 24\hat{\pi}_i^5)\,, \tag{38}$$

$$\hat{\kappa}_{6i} = m_i(\hat{\pi}_i - 31\hat{\pi}_i^2 + 180\hat{\pi}_i^3 - 390\hat{\pi}_i^4 + 360\hat{\pi}_i^5 - 120\hat{\pi}_i^6)\,. \tag{39}$$

Leider gibt McCullagh in seinen Veröffentlichungen verschiedene Formeln für die bedingten Kumulanten der Pearson-Statistik an. Zur Darstellung und Berechnung der Parameter wird in der vorliegenden Arbeit die Gleichung für $\hat{\mathrm{E}}(X^2|\hat{\beta})$ aus ([36]) und die Gleichungen für $\widehat{\mathrm{Var}}(X^2|\hat{\beta})$ und $\hat{\kappa}_3(X^2|\hat{\beta})$ aus [35] benutzt. In dieser Zusammensetzung gelingt es, die Ergebnisse für die Beispieldatensätze von McCullagh nachzuvollziehen.

Approximative Signifikanztests können berechnet werden, indem man die standardisierte Teststatistik

$$X^2_{McC} = \frac{(X^2 - \hat{\mathrm{E}}(X^2|\hat{\beta}))}{\widehat{\mathrm{Var}}(X^2|\hat{\beta})^{1/2}} \tag{40}$$

mit der Standardnormalverteilung vergleicht oder durch die Edgeworth-Expansion

$$P(X^2 \geq z) = 1 - \Phi(z) + \phi(z)(z^2 - 1)\rho_{3McC}/6\,, \tag{41}$$

wobei die standardisierte bedingte Schiefe ρ_{3McC} von Z_{McC} geschätzt wird durch

$$\hat{\rho}_{3McC} = \frac{\hat{\kappa}_3(X^2|\hat{\beta})}{\widehat{\mathrm{Var}}(X^2|\hat{\beta})^{3/2}}\,. \tag{42}$$

McCullagh benutzt in seiner Edgeworth-Expansion nur das erste Hermitesche Polynom, da die Herleitung des vierten standardisierten Kumulanten ρ_{4McC} zu kompliziert ist.
Einen Erweiterung dieses Tests geben Forster et al. ([22]) an. Sie begnügen sich nicht mit den asymptotischen Methoden der bedingten Verteilung, sondern leiten die exakte bedingte Verteilung her. Allerdings existiert in diesem Fall kein Anpassungstest in geschlossener Form, man benötigt vielmehr Monte Carlo-Methoden, um p-Werte zu berechnen.

2.1.3 F-Verteilung

Snapinn/Small ([50]) schlagen eine Ad-Hoc-Korrektur der Prüfverteilung der Devianz vor. Sie leiten diese eigentlich für ordinale logistische Regressionsmodelle her, aber da die logistische Regression ein Spezialfall dieser ist, ist diese Korrektur auch hier von Interesse. Ausgehend von der Beobachtung, dass der Signifikanztest für die Devianz bei Verwendung der χ^2-Verteilung zu liberal ist, empfehlen Snapinn/Small einfach die Verwendung einer F-Verteilung. Konkret sollte die Nullhypothese der guten Anpassung verworfen werden, falls

$$D_F = D/N > F_{N,M-p-1,1-\alpha}\,, \tag{43}$$

wobei $F_{N,M-p-1,1-\alpha}$ das $1-\alpha$-Quantil einer F-Verteilung mit N und $M-p-1$ Freiheitsgraden ist.
Dieser Test macht sich dabei zunutze, dass $NF_{N,M-p-1,1-\alpha} > \chi^2_{N,1-\alpha}$, so dass die Nullhypothese weniger oft verworfen wird und dadurch die Antikonservativität der Standard-Prozedur bekämpft wird. Daneben gilt

$$NF_{N,M-p-1,1-\alpha} - \chi^2_{N,1-\alpha} \to 0 \tag{44}$$

für $M \to \infty$, so dass der Unterschied zwischen Snapinn-Small-Test und der Standard-Prozedur verschwindet, falls M groß wird. Allerdings ist damit noch nichts über das Verhalten des Tests gesagt, wenn N ebenfalls groß wird.

2.2 Gruppierung von Beobachtungen

2.2.1 Hosmer-Lemeshow-Test

Die Grundlage des Hosmer-Lemeshow-Tests ([25]) ist eine 2×10 -Kontingenztafel (vgl. Abbildung 5), in der die individuellen Beobachtungen aus dem ursprünglichen logistischen Regressionsmodell neu gruppiert werden.

Abbildung 5: Neugruppierung der individuellen Beobachtungen als 2×10 -Kontingenztafel bzgl. der Zielgröße und der gruppierten Erfolgswahrscheinlichkeit zur Berechnung des Hosmer-Lemeshow-Tests

		Risikodezile			
		1	2	$\cdots$	10
Ziel-	1	O_{11}	O_{21}	$\cdots$	O_{101}
größe	0	$M_1 - O_{11}$	$M_2 - O_{21}$	$\cdots$	$M_{10} - O_{101}$
		M_1	M_2	$\cdots$	M_{10}

Die Zeilen dieser Kontingenztafel werden durch die Ausprägungen der Zielgröße definiert, die Spalten durch die Werte der geordneten geschätzten Ereigniswahrscheinlichkeiten $\hat{\pi}_i$. Das heisst, wir finden in der ersten Spalte die Anzahl der ersten $M/10$ Beobachtungen mit den kleinsten Ereigniswahrscheinlichkeiten und in der zehnten Spalte die Anzahl der letzten $M/10$ Beobachtungen mit den größten Ereigniswahrscheinlichkeiten. Dementsprechend finden wir in der ersten Zeile die Anzahl der Beobachtungen, bei denen das Zielereignis eingetreten ist, in der zweiten Zeile die Anzahl der Beobachtungen ohne Zielereignis. Hosmer und Lemeshow bezeichnen diese Art der Einteilung als die „Risiko-Dezil-Methode“. Die Neugruppierung der individuellen Beobachtungen erfolgt also quasi mit Hilfe einer neuen einzelnen „Meta-Kovariablen“, die die gesamte Information aus den ursprünglichen Kovariablen sammelt.

Bezeichnen wir dann die Anzahl der Beobachtungen in der g-ten Spalte mit M_g, $(g = 1, \ldots, 10)$, so ist die *erwartete* Anzahl von Zielereignissen in der g-ten Spalte gerade gleich der Summe über die geschätzten Erfolgswahrscheinlichkeiten $E_{g1} = \sum_{i=1}^{M_g} \hat{\pi}_i$, die erwartete Anzahl von Nichtereignissen in derselben Spalte ist gleich $E_{g2} = M_g - E_{g1}$, die *beobachtete* Anzahl von Zielereignissen in dieser Gruppe ist gleich $O_{g1} = \sum_{i=1}^{M_g} y_i$ und die beobachtete Anzahl von Nichtereignissen gleich $O_{g2} = M_g - O_{g1}$.
Die Teststatistik ist die herkömmliche Pearson-Statistik in einer Mehrfeldertafel, die die Diskrepanz zwischen beobachteten und erwarteten Häufigkeiten misst und ergibt sich zu

$$\hat{C} = \sum_{g=1}^{10} \sum_{k=1}^{2} \frac{(O_{gk} - E_{gk})^2}{E_{gk}} \,. \tag{45}$$

Als Prüfverteilung erwarten wir im vorliegenden Falle eine χ^2-Verteilung mit $9 = (10-1)(2-1)$ Freiheitsgraden. Hosmer/Lemeshow schlagen an verschiedenen Stellen ([25], [32], [26]), ausgehend von umfangreichen Simulationsuntersuchungen, jedoch vor, eine χ^2-Verteilung mit 8 Freiheitsgraden zur p-Wert-Berechnung zu verwenden. Dieser Verlust eines Freiheitsgrades kommt dadurch zustande, dass die Häufigkeiten in der Kontingenztafel von den geschätzten Modellparametern abhängen und also zufällig sind.
Hosmer/Lemeshow ([26], S. 141f.) weisen darauf hin, dass die Varianz der O_{gk} in der g-ten Spalte, die im Nenner von $\hat{C}$ steht und de facto durch $\sum_{i=1}^{M_g} \hat{\pi}_i(1-\hat{\pi}_i)$ geschätzt wird, systematisch unterschätzt wird. Korrekt wäre eigentlich die Verwendung von $M_g\bar{\pi}_g(1-\bar{\pi}_g)$ im Nenner von $\hat{C}$, wobei $\bar{\pi}_g = 1/M \sum_{i=1}^{M_g} \hat{\pi}_i$ die mittlere Erfolgswahrscheinlichkeit in der g-ten Gruppe ist. Diese Unterschätzung der Varianz resultiert in einer Überschätzung von $\hat{C}$.
Ausgehend von dieser Überlegung schlagen Pigeon/Heyse ([45], [46]) eine modifizierte Teststatistik $\hat{C}_{PH}$ vor mit

$$\hat{C}_{PH} = \sum_{g=1}^{10} \sum_{k=1}^{2} \frac{(O_{gk} - E_{gk}^2)}{\phi_g E_{gk}} \,, \tag{46}$$

und Korrekturfaktor

$$\phi_g = \frac{\sum_{i=1}^{M_g} \hat{\pi}_i(1-\hat{\pi}_i)}{M_g\bar{\pi}_g(1-\bar{\pi}_g)} \,, \tag{47}$$

die jetzt tatsächlich mit einer χ^2-Verteilung mit 9 Freiheitsgraden verglichen werden soll. Hosmer/Lemeshow bemerken dazu, dass $\hat{C}$ und $\hat{C}_{PH}$ sich in der

Regel kaum unterscheiden (was auch die Beispiele aus den Veröffentlichungen von Pigeon/Heyse nahelegen), so dass der auf $\hat{C}_{PH}$ basierende Test ein eher konservatives Verhalten zeigen wird.

Erweiterungen des Hosmer-Lemeshow-Tests In der Originalveröffentlichung von 1980 schlagen Hosmer/Lemeshow neben der Teststatistik $\hat{C}$ noch eine Reihe von weiteren Teststatistiken vor, die ebenfalls auf dem Prinzip der Neugruppierung der Beobachtungen in 2×10-Tafeln beruhen. Unterschiede zu $\hat{C}$ ergeben sich zum einen dadurch, dass die Partition der Erfolgswahrscheinlichkeiten vorher festgelegt wird, z.B. durch Unterteilung des Intervalls $[0, 1]$ in Subintervalle $[0, 0.1], [0.1, 0.2], \ldots$ und die Beobachtungen wieder entsprechend ihrer geschätzten Erfolgswahrscheinlichkeit $\hat{\pi}_i$ in diese Subintervalle eingeteilt werden. Dies hat den Vorteil, dass die Dezilgrenzen jetzt leichter interpretierbar sind, andererseits können auf diese Art schwach besetzte Zellen entstehen, so dass die Gültigkeit der χ^2-Verteilung fraglich wird.
Zum anderen leiten die beiden Autoren noch weitere Tests unter der verschärften Annahme der multivariaten Normalverteilung der Kovariablen her, die in der Standardtheorie der logistischen Regression nicht gebraucht wird, sondern aus der Diskriminanzanalyse stammt. Überraschenderweise zeigen sich diese Teststatistiken in Simulationsuntersuchungen als robust gegen eine Verletzung dieser Annahme der normalverteilten Kovariablen. Mehr noch: Sie sind in diesen Simulationen den anderen Tests bezüglich der Power überlegen. Der Grund, warum Hosmer/Lemeshow diese Tests nicht zur Anwendung empfehlen, ist die Tatsache, dass zu deren Berechnung numerische Integrationen erforderlich sind, was in den Frühzeiten der Computertechnik mit erheblichem Zeitaufwand verbunden war. Zum anderen zeigt eine Simulationsuntersuchung von Korn/Hosmer/Lemeshow ([31]), dass diese Teststatistik empfindlich auf diskrete Kovariablen reagiert.

Probleme mit dem Hosmer-Lemshow-Test Der Hosmer/Lemeshow-Test hat sich zu einem Standardverfahren entwickelt, wenn es um die Beurteilung der Anpassungsgüte von logistischen Regressionsmodellen geht. Dies kommt u.a. dadurch zum Ausdruck, dass er standardmäßig in allen großen statistischen Software-Paketen (BMDP, LOGXACT, SAS, SPSS, STATA, STATISTIX, SYSTAT) implementiert ist.
Seine Anwendung ist allerdings auch nicht ohne Probleme. Hosmer/Lemeshow ([28]) weisen selber daraufhin, dass die angeführten Programmpakete

unterschiedliche Algorithmen verwenden, um die Grenzen der Risikodezile zu berechnen. In manchen Fällen wird das Hauptaugenmerk daraufgelegt, dass die Dezile möglichst gleich groß sind, in anderen werden Beobachtungen mit der selben Ereigniswahrscheinlichkeit auf jeden Fall dem gleichen Dezil zugeteilt, was bei einer kleinen Zahl von verschiedenen Ereigniswahrscheinlichkeiten zu stark unterschiedlichen Dezilgrößen führen kann. Diese führt dazu, dass bei einem konkreten Datensatz die verschiedenen Programmpakete p-Werte des Hosmer-Lemeshow-Tests zwischen 0.02 und 0.16 liefern, obwohl der selbe Datensatz zugrunde liegt und die Parameterschätzer in allen Programmpaketen identisch sind.
Ein weiterer Nachteil des hier vorgestellten Testprizips ist der, dass in den Risikodezilen Beobachtungen nebeneinander liegen können, die sich bezüglich der ursprünglichen Kovariablen stark unterscheiden, d.h. der Test liefert nicht notwendig Information darüber, wo im Raum der Kovariablen das Modell eine schlechte Anpassung an die Daten liefert.
Daneben ist auch nicht unmittelbar einsichtig, warum ausgerechnet 10 neue Risikogruppen gebildet werden sollen. Es ist durchaus denkbar und wurde auch von Hosmer/Lemshow beobachtet, dass der Test bei einer festen Menge von Risikogruppen die Nullhypothese verwirft, bei einer anderen aber eine gute Anpassung des Modells anzeigt.
Eine umfassende Theorie für χ^2-Tests mit datenabhängigen Zellen legt Andrews ([4],[5]) vor. Diese basiert auf Ergebnissen über die schwache Konvergenz von empirischen Prozessen auf Mengen. Der Abstand zwischen beobachteten und erwarteten Zellhäufigkeiten wird dabei als bedingter empirischer Prozess aufgefasst, der (nach Standardisierung) bei korrekter Modellspezifikation asymptotisch gegen eine χ^2-Verteilung konvergiert. Die Anwendungsmöglichkeiten dieses Test sind sehr vielfältig und gehen weit über das logistische Modell hinaus.

2.2.2 Tsiatis-Test

Der Test von Tsiatis ([52]) beruht ebenfalls auf dem Testprinzip der Neugruppierung von Beobachtungen. Allerdings wird diese Gruppierung nicht datenabhängig durchgeführt, vielmehr wird a priori der Raum der Kovariablen partitioniert. Bezeichnet man die Q Gruppen dieser Partition mit $R_1, \ldots, R_Q$, mit $I[x \in R_l]$ die Indikatorfunktion für das Ereignis $\{x \in R_l\}$

und betrachtet dann das erweiterte logistische Modell

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{p} \beta_j x_{ji} + \sum_{l=0}^{Q} \gamma_l I[x \in R_l]\,, \tag{48}$$

so gilt unter korrekter Modellspezifikation $\gamma_1 = \ldots = \gamma_Q = 0$. Der Score-Test für diese Nullhypothese im Modell (48) liefert eine Teststatistik, die χ^2-verteilt ist mit $Q-1$ Freiheitsgraden.
Tsiatis zeigt, dass dieser Test auf dem Vergleich zwischen beobachteten und erwarteten Häufigkeiten in den Q Gruppen beruht, ohne dass diese dazu explizit berechnet werden müssten. Das hat den Vorteil, dass man sich diese Berechnung ersparen kann, aber den Nachteil, dass diese dann auch nicht zur inhaltlichen Interpretation zur Verfügung stehen. Der Tsiatis-Test kann ferner auch noch als Test auf konstanten Intercept β_0 in den Q Gruppen interpretiert werden.
Die bei diesem Test notwendige a priori Klasseneinteilung macht vor allem dann Sinn, wenn sie inhaltlich begründet ist. So könnte zum Beispiel in einer epidemiologischen Studie, in der ein einzelner Risikofaktor untersucht werden soll und die anderen Kovariablen im Modell nur der Adjustierung dienen, der Raum der Kovariablen nach den Ausprägungen des hauptsächlich interessierenden Risikofaktors partitioniert werden. Man erhält so einen Eindruck davon, wie gut die Anpassung des Modells bezüglich der Werte dieses Risikofaktors ist. Im allgemeinen ist diese Partitionierung aber inhaltlich nicht so leicht durchzuführen und lässt auch Platz für Manipulationen.

2.3 Verwendung anderer Teststatistiken

2.3.1 Bartlett-Korrektur von D

Dieser Vorschlag geht auf Bartlett zurück, der in mehreren Arbeiten multiplikative Korrekturfaktoren für Likelihood-Ratio-Statistiken ableitet, so dass der Erwartungswert der modifizierten Statistik, mit der dann Signifikanztests berechnet werden, näher am Erwartungswert der χ^2-Verteilung liegt. Cordeiro ([11]) berechnet einen solchen Korrekturterm für die Devianz in generalisierten linearen Modellen, von denen das logistische Regressionsmodell ein Spezialfall ist. Simulationsergebnisse ([12]) für Poisson- und Gamma-Regressionsmodelle zeigen, dass die korrigierte Devianz der unkorrigierten Devianz überlegen ist, und es ist zu erwarten, dass dies auch für logistische Regressionsmodelle gilt.

Die Bartlett-Korrektur der Devianz nach Cordeiro wird geschätzt durch

$$D_C = D/c \tag{49}$$

mit

$$c = \frac{1}{N-p-1}\left(N + \frac{1}{6}\sum_{i=1}^{N}\frac{1-\hat{\pi}_i(1-\hat{\pi}_i)}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} - (p+1) - \hat{\epsilon}_p\right). \tag{50}$$

Der Term $\hat{\epsilon}_p$ ergibt sich durch

$$\hat{\epsilon}_p = \frac{1}{4}\mathrm{tr}(\hat{H}_c\hat{Q}_d^2) + \frac{1}{12}\mathbf{1}^t\hat{F}_c(2\hat{Q}^{(3)} + 3\hat{Q}_d\hat{Q}\hat{Q}_d)\hat{F}_c\mathbf{1}\,, \tag{51}$$

mit $\hat{H}_c = \mathrm{diag}(2m_i\hat{\pi}_i(1-\hat{\pi}_i)^2 - m_i\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)^2)$, $\hat{Q}$ wie in 2.1.2, S.21, $\hat{Q}_d$ einer Diagonalmatrix der Diagonalelemente von $\hat{Q}$, $\hat{Q}^{(3)} = (\hat{q}_{ij}^3)$ als einer Matrix die die Elemente der Matrix $\hat{Q}$ in die dritte Potenz erhebt und $\hat{F}_c = \mathrm{diag}(m_i\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i))$. Die diag-Funktion bezeichnet dabei die Funktion, die aus einem N-dimensionalen Vektor eine $N \times N$-Matrix macht, in der die Elemente des Vektors auf der Hauptdiagonalen stehen und alle anderen Elemente der Matrix gleich Null sind.
Zur Überprüfung der Anpassungsgüte des Modells vergleicht man den beobachteten Wert von D_C mit dem Quantil der χ^2-Verteilung mit $N-p-1$ Freiheitsgraden, da der Erwartungswert von D_C besser durch den Erwartungswert einer χ^2-Verteilung approximiert wird als der von D. Damit wird jedoch nichts über die Approximation der höheren Momente von D_C ausgesagt, Bartlett-Korrekturen beziehen sich grundsätzlich nur auf Erwartungswerte.
Eine approximative Bartlett-Korrektur gibt Lugtenburg an ([34]). Diese ergibt sich aus der Cordeiro-Korrektur, wenn $\epsilon_p = 0$ gesetzt wird.

2.3.2 Farringtons modifizierter Pearson-Test

Farrington ([20],[21]) leitet im Kontext der verallgemeinerten linearen Modelle eine modifizierte Pearson-Statistik her. Ausgangspunkt dazu sind die Arbeiten von McCullagh ([35],[36],[37], vgl. 2.1.2, S. 20), der bedingte Momente der Pearson-Statistik und der Devianz für verallgemeinerte lineare Modelle mit kanonischer Linkfunktion, wie es das logistische Regressionsmodell ist, ableitet. Farrington verallgemeinert diese Ergebnisse für beliebige Linkfunktionen. Als Teilergebnis fällt dabei auch eine durch die Addition eines linearen Korrekturterms modifizierte Pearson-Statistik ab, die auch auf das logistische Regressionsmodell anwendbar ist.

Farrington schlägt dazu eine Familie von Pearson-Statistiken X_a^2 vor mit

$$X_a^2 = \sum_{i=1}^{N} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} + \sum_{i=1}^{N} a_i(y_i - m_i\hat{\pi}_i)\,. \tag{52}$$

Für $a_i \equiv 0$ ergibt sich die herkömmliche Pearson-Statistik X^2. Durch die Wahl $a_i = \frac{-(1-2\hat{\pi}_i)}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}$ ergibt sich die Statistik

$$X_F^2 = \sum_{i=1}^{N} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} + \sum_{i=1}^{N} \frac{-(1-2\hat{\pi}_i)}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}(y_i - m_i\hat{\pi}_i)\,. \tag{53}$$

Diese hat unter allen X_a^2 optimale Eigenschaften bezüglich Varianz und lokaler Orthogonalität. Desweiteren vereinfachen sich die approximativen bedingten Momente von X_F^2 im Vergleich zum allgemeinen Fall beträchtlich, so dass diese einer Berechnung zugänglich werden. Diese Momente werden dann geschätzt durch

$$\hat{\mathrm{E}}(X_F^2|\hat{\beta}) = N - p - 1 + \sum_{i=1}^{N}(\hat{\pi}_i(1-\hat{\pi}_i))\hat{Q}_{ii} \tag{54}$$

$$\widehat{\mathrm{Var}}(X_F^2|\hat{\beta}) = 2(1 - \frac{p+1}{N})\sum_{i=1}^{N}\frac{m_i - 1}{m_i} \tag{55}$$

$$\hat{\kappa}_3(X_F^2|\hat{\beta}) = 8(N-p-1)\Big(1 - \frac{1}{N}\sum_{i=1}^{N}\frac{5m_i - 4}{m_i^2} + \frac{1}{2N}\sum_{i=1}^{N}\frac{m_i - 1}{m_i^2\hat{\pi}_i(1-\hat{\pi}_i)}\Big)\,, \tag{56}$$

mit der Matrix $\hat{Q} = X(X^t\hat{W}X)^{-1}X^t$ wie in 2.1.2, S.21.
Für $m_i \to \infty$ streben diese Momente, wie zu erwarten, gegen die Momente einer χ^2-Verteilung mit $N-p-1$ Freiheitsgraden. Approximative Signifikanztests können berechnet werden, indem man die standardisierte Teststatistik

$$Z_F = \frac{(X_F^2 - \hat{\mathrm{E}}(X_F^2|\hat{\beta})}{\widehat{\mathrm{Var}}(X_F^2|\hat{\beta})^{1/2}} \tag{57}$$

mit der Standardnormalverteilung vergleicht oder mithilfe der Edgeworth-Expansion

$$P(Z_F \geq z) = 1 - \Phi(z) + \phi(z)(z^2-1)\rho_{3F}/6\,, \tag{58}$$

wobei die standardisierte bedingten Schiefe ρ_{3F} von Z_F zu schätzen ist mittels

$$\hat{\rho}_{3F} = \frac{\hat{\kappa}_3(X_F^2|\hat{\beta})}{\widehat{\mathrm{Var}}(X_F^2|\hat{\beta})^{3/2}}\,. \tag{59}$$

Im Extremfall von ausschließlich einfach besetzten Kovariablenmustern ($m_i \equiv 1$) enthält die Farrington-Statistik ähnlich der Devianz keinerlei Information über den Fit des Modells. Sie degeneriert in diesem Fall zu $X_F^2 \equiv N$.
Bei der Herleitung der Teststatistiken X_a^2 geht Farrington von einem Modell mit Overdispersion als Alternativmodell aus. Die Overdispersion wird dabei durch einen multiplikativen Korrekturterm ϕ der binomialen Varianz modelliert, d.h. die Varianz der Zielgröße ist nicht länger $m_i\pi_i(1-\pi_i)$, sondern $\phi m_i\pi_i(1-\pi_i)$. Es ist daher zu erwarten, dass X_F^2 hohe Power gegen Overdispersion hat. Eine kleine Simulationsuntersuchung von Farrington ([20]) bestätigt diese Vermutung.

2.3.3 Informationsmatrix-Test

Der Informationsmatrix-Test wird in der medizinischen Statistik bisher kaum angewandt, vorgeschlagen ([53]) und weiterentwickelt (z.B. [8], [16], [40]) wurde er vor allem von Ökonometrikern. Eine Ausnahme sind dabei Hosmer/Lemeshow, die von einem „eleganten“ Test sprechen, der aber „in der Praxis schwer zu berechnen ist und dessen Power noch nicht ausreichend untersucht wurde“ ([26], S. 169).
Der IM-Test wird dabei nur für ungruppierte Daten hergeleitet, d.h. $m_i \equiv 1$ bzw. $M \equiv N$ per Definition. Insofern ist bei der Berechnung darauf zu achten, dass die gruppierten Beobachtungen in individuelle Beobachtungen aufgelöst werden.
Dem Test zugrunde liegt die Informationsmatrix-Gleichung, die besagt, dass sich die Fisher-Information $I(\beta)$ (die gleich der Inversen der asymptotischen Varianz-Kovarianz-Matrix der Parameterschätzer, also eine $(p+1)\times(p+1)$-Matrix ist) bei korrekt spezifiziertem Modell auf zwei äquivalente Arten schreiben lässt, nämlich einmal in der Hesse'schen Form

$$I_1(\beta) = -E\left(\frac{\partial^2 \log L(\beta)}{\partial\beta_j\partial\beta_k}\right) \tag{60}$$

und einmal in der OPG(Outer Product Gradient)-Form

$$I_2(\beta) = E\left(\frac{\partial \log L(\beta)}{\partial\beta_j}\frac{\partial \log L(\beta)}{\partial\beta_k}\right) \tag{61}$$

mit jeweils $j, k = 0, \ldots, p$.
Die Idee des Tests ist es, die unbekannten β durch ihre Maximum-Likelihood-Schätzer zu ersetzen, die beiden Matrizen $I_1(\beta)$ und $I_2(\beta)$ zu schätzen, dann zu addieren und die Summe mit Null zu vergleichen. Da die Informationsmatrix symmetrisch ist, ist es nur notwendig, die $\frac{1}{2}(p+1)(p+2)$ Elemente im unteren Dreieck von $I_1(\beta)$ und $I_2(\beta)$ zu vergleichen.
Konkret ergibt sich durch Einsetzen der $(\frac{1}{2}(p+1)(p+2) \times 1)$-Vektor

$$\frac{1}{M}\sum_{i}^{M}\hat{l}_i = \frac{1}{M}\sum_{i}^{M}(y_i - \hat{\pi}_i)(1 - 2\hat{\pi}_i)z_i\,, \tag{62}$$

wobei $z_i = \text{vech}(x_i^t x_i)$ mit „vech“ als dem Operator, der die voneinander verschiedenen Elemente einer symmetrischen Matrix untereinander in eine Spalte schreibt. Das heisst anschaulich, aus jeder Spalte der transponierten Designmatrix wird in einem ersten Schritt durch Quadrierung eine $((p+1) \times (p+1))$-Matrix gewonnen, aus der dann in einem zweiten Schritt die Elemente im unteren Dreieck ausgelesen und in den Vektor z_i geschrieben werden.
Der Vektor $\frac{1}{M}\sum_i^M \hat{l}_i$ hat bei korrekter Spezifikation des Modells aufsummiert über die $\frac{1}{2}(p+1)(p+2)$ Komponenten den Wert Null. Ausgehend davon ergibt sich die Teststatistik

$$IM = \frac{1}{M}\left(\sum_{i=1}^{M}\hat{l}_i\right)^t \hat{V}_{IM}^{-1}\left(\sum_{i=1}^{M}\hat{l}_i\right)\,, \tag{63}$$

die unter der Nullhypothese eine asymptotische χ^2-Verteilung mit $\frac{1}{2}(p+1)(p+2)$ Freiheitsgraden hat.
Die Varianz-Kovarianzmatrix V_{IM} wird geschätzt mittels (vgl. [40])

$$\hat{V}_{IM} = \frac{1}{M}\left[(\hat{F}Z)^t(I - \hat{D}X((\hat{D}X)^t\hat{D}X)^{-1}(\hat{D}X)^t)\hat{F}Z\right]\,, \tag{64}$$

mit $\hat{D} = \text{diag}(\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)})$, $\hat{F} = \text{diag}(\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}(1-2\hat{\pi}_i))$ und Z als der Matrix mit den oben definierten z_i^t als Zeilen. Für ausschließlich binäre Kovariablen wird die Matrix V_{IM} singulär und nicht mehr invertierbar, was zur Nicht-Definiertheit von IM führt. Man muss dann eine verallgemeinerte Inverse von V_{IM} zur Berechnung der Teststatistik heranziehen.
Es müssen nicht notwendig genau alle verschiedenen Elemente aus $I(\beta)$ miteinander verglichen werden, um einen vernünftigen Anpassungstests herzuleiten. Genauso gut ist es denkbar, nur die Elemente auf der Hauptdiagonalen

der Informationsmatrix heranzuziehen, so dass $\frac{1}{M}\sum_i^M \hat{l}_i$ die Dimension $p+1$ hat, wodurch sich in analoger Weise die Teststatistik IM_{DIAG} ergibt, die dann mit einer χ^2-Verteilung mit $p+1$ Freiheitsgraden verglichen wird.
Das hier angewandte Testprinzip des Vergleichs von $I_1(\beta)$ und $I_2(\beta)$ führt auch in anderen Bereichen der Statistik außerhalb der logistischen Regression zu bekannten Anpassungstests (vgl. [39]). Beispiele sind ein Test auf Poisson-Verteilung einer Stichprobe oder ein Test auf Normalverteilung, der auf beobachteter Schiefe und Kurtosis beruht. Im Bereich der linearen Regression ergibt sich ein Test auf Unabhängigkeit der Residuen.
Die Berechnung der Teststatistik wird erleichtert durch die Tatsache, dass diese gleich der residualen Fehlerquadratsumme einer linearen Regression mit „Beobachtungen“ $r_i = \frac{(y_i - \hat{\pi}_i)}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$ und „Designmatrix“ $X^* = (\hat{D}X|\hat{F}Z)$ ist. Insofern kann die Teststatistik mit jedem herkömmlichen Statistikprogramm, das die Schätzung von linearen Regressionsmodellen erlaubt, berechnet werden. Es ist zu beachten, dass in dieser Regression kein konstanter Faktor mitgeschätzt werden darf.
An verschiedenen Stellen in der Literatur (z.B. [19], S. 146) wird der IM-Test heftig kritisiert, da er in verschiedenen Simulationsuntersuchungen dramatische Überschreitungen des empirischen Niveaus gezeigt hatte. Dieses antikonservative Verhalten des Tests geht zurück auf einen ungeeigneten Schätzer (den so genannten *OPG-Schätzer*) der Kovarianzmatrix V_{IM}. Wird diese Matrix, wie oben angegeben, durch ihren ML-Schätzer geschätzt, hält der Test das Niveau ein ([41]).
Chesher([8]) zeigt, dass der IM-Test auch als Score-Test auf Parameterhomogenität im logistischen Regressionsmodell interpretiert werden kann. Das heisst, ein Score-Test mit Nullhypothese „Die Parameter β sind fest und nicht zufällig“ gegen die Alternativhypothese „Die Parameter β folgen einer gemeinsamen zufälligen Verteilung“, hat als Prüfgröße gerade die IM-Teststatistik.
An verschiedenen Stellen in der ökonometrischen Literatur (s. z.B. [54]) wird betont, wie wichtig es ist, daß die Informationsmatrix-Gleichung erfüllt ist, da bei ihrer Verletzung die Kovarianzmatrix der Parameterschätzer $\mathrm{Cov}(\beta) = (X^t W X)$ nicht mehr konsistent geschätzt werden kann. In diesem Fall sind die herkömmlichen Parametertests (Wald-, Likelihood-Ratio- und Score--Tests) auch asymptotisch nicht mehr χ^2-verteilt und um zu gültigen Aussagen hinsichtlich der Parameter zu kommen, müssten verbesserte Varianzschätzer verwendet werden.

Es ist jedoch erfreulich, dass auch bei Verletzung der Informationsmatrix-Gleichung mit der herkömmlichen Methode ein quasi optimaler Schätzer für die β berechnet wird und zwar optimal dahingehend, dass dieser konsistent ein $\bar{\beta}$ schätzt, das im Sinne der Kullback-Leibler-Divergenz einen minimalen Abstand vom wahren β hat ([53]). Ausgehend von diesem so genannten *Quasi-Maximum-Likelihood-(QML)-Schätzer* können auch korrigierte Varianzschätzer berechnet werden, die dann robust gegen die Modellverletzung sind. Die korrigierten Parametertests halten dann auch bei Vorliegen der Fehlspezifikation das Niveau asymptotisch ein. Die Korrektur geht dabei von einer Kovarianzmatrix $\mathrm{Cov}_{QML}(\beta) = I_1(\beta)^{-1} I_2(\beta) I_1(\beta)^{-1}$ aus.

2.3.4 Residuen-Tests

Summen von standardisierten Residuen Die Herleitung dieser Gruppe von Tests beruht auf der Beobachtung, dass die beiden klassischen Anpassungstests D und X^2 Summen von quadrierten Residuen darstellen. Die Idee ist, diese einzelnen Residuen so zu standardisieren, dass diese „günstigere" Verteilungseigenschaften aufweisen, in der Hoffnung, dass diese Eigenschaften sich dann auf die aufsummierten Residuen übertragen, so dass gute globale Anpassungstests entstehen.
Ausgangspunkt sind die Pearson-Residuen

$$r_i = \frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}\,, \tag{65}$$

die quadriert und aufsummiert die Pearson-Statistik X^2 ergeben und die Devianz-Residuen

$$d_i = \mathrm{sgn}(y_i - m_i\hat{\pi}_i)\sqrt{2y_i \log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + 2(m_i - y_i)\log\left(\frac{m_i - y_i}{m_i(1-\hat{\pi}_i)}\right)} \tag{66}$$

mit $\mathrm{sgn}(y_i - m_i\hat{\pi}_i) = 1$, falls $y_i > m_i\hat{\pi}_i$ und $\mathrm{sgn}(y_i - m_i\hat{\pi}_i) = -1$ sonst. Diese ergeben quadriert und aufsummiert gerade die Devianz D.
Die Verteilung dieser Residuen ist schief und für kleine m_i weit von der Standardnormalverteilung entfernt. Eine Stabilisierung zumindest der Varianz dieser Verteilung liefert, genau wie bei der linearen Regression, eine Standardisierung der Residuen mit den Elementen der Hutmatrix $\hat{H} = \hat{W}^{1/2} X (X^t \hat{W} X)^{-1} X^t \hat{W}^{1/2}$.

Dadurch ergeben sich die studentisierten Pearson-Residuen

$$r_{Pi} = \frac{r_i}{\sqrt{1 - h_{ii}}}\,, \tag{67}$$

und die standardisierten Devianz-Residuen

$$r_{Di} = \frac{d_i}{\sqrt{1 - h_{ii}}}\,, \tag{68}$$

wobei die h_{ii} die Diagonalelemente der Hutmatrix sind. Durch die Studentisierung wird auch berücksichtigt, dass die $\hat{\pi}_i$ geschätzt wurden ([9], S.122), da die herkömmlichen Verteilungseigenschaften der Residuen in der Regel für feste π_i hergeleitet werden.
Einen Kompromiss zwischen studentisierten Pearson- und Devianz-Residuen liefern die Likelihood-Residuen

$$r_{Li} = \operatorname{sgn}(y_i - m_i\hat{\pi}_i)\sqrt{h_{ii}r_{Pi}^2 + (1 - h_{ii})r_{Di}^2}\,. \tag{69}$$

Diese haben ihren Namen davon, dass r_{Li} die (approximative) Veränderung der Devianz (also der Likelihood-Statistik) misst, wenn man die i-te Beobachtung aus dem Datensatz ausschließt. Da die h_{ii} im allgemeinen klein sind, werden die r_{Li} nahe bei r_{Di} liegen.
Desweiteren können die Likelihood-Residuen auch als Ausreißer-Residuen interpretiert werden, und zwar in dem Sinne, dass r_{Li} eine Prüfgröße dafür ist, ob die i-te Beobachtung ein Ausreißer ist. Konkret ist r_{Li} eine Approximation der LR-Statistik für $\gamma_i = 0$ im Outlier-Modell von Williams ([56])

$$\operatorname{logit}(\pi_i) = \sum_{j=0}^{p} \beta_j x_{ji} + u_i\gamma_i\,, \tag{70}$$

wobei u_i ein Einheitsvektor ist, der zwischen lauter Nullen nur in der i-ten Komponente eine 1 stehen hat. Die i-te Beobachtung ist also ein Ausreißer, wenn γ_i von Null verschieden ist.
Die bisher vorgestellten Residuen waren aus den rohen Summanden von X^2 und D durch eine multiplikative Adjustierung hervorgegangen. Erreicht wird dadurch eine Stabilisierung der Varianz der Residuen auf Eins. Eine andere Möglichkeit, die Verteilung der Residuen in Richtung der Normalverteilung zu transformieren, ist die Korrektur der Schiefe dieser Verteilung. Dadurch entstehen die Anscombe-Residuen in Anlehnung an die Arbeit von F.J.

Anscombe ([6]), der ein allgemeines Konstruktionsprinzip für Residuen vorschlägt, bei dem die Zielgröße transformiert wird, um günstige Eigenschaften der Verteilung der Residuen zu erreichen. Im Falle einer binomial-verteilten Zielgröße erhält man

$$r_{Ai} = \frac{\frac{(\Gamma(2/3))^2}{\Gamma(4/3)} \left[\mathcal{B}\left(\frac{y_i}{m_i}, \frac{2}{3}, \frac{2}{3}\right) - \mathcal{B}\left(\frac{m_i\hat{\pi}_i}{m_i}, \frac{2}{3}, \frac{2}{3}\right) \right]}{\left[\sqrt[6]{\frac{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})^3}{m_i^3}} \right]}\,, \tag{71}$$

wobei Γ für die Gamma-Funktion steht und $\mathcal{B}(p, a, b)$ für die Verteilungsfunktion der Beta-Verteilung.
Auch die Verteilung der Devianz-Residuen kann noch verbessert werden, wobei sich diese Adjustierung auf den Mittelwert dieser Verteilung bezieht. Wir erhalten bias-adjustierte Devianz-Residuen durch

$$r_{Bi} = d_i + \frac{1 - 2\hat{\pi}_i}{6\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}\,. \tag{72}$$

Insgesamt ergeben sich also durch Quadrieren und Aufsummieren über die Residuen $r_{Pi}, r_{Di}, r_{Li}, r_{Ai}$ und r_{Bi} die fünf globalen Anpassungstests R_P, R_D, R_L, R_A und R_B, die unter der Nullhypothese χ^2-verteilt sind mit N Freiheitsgraden (für R_P, R_D, R_L, da bei diesen die Schätzung der Parameter bereits explizit berücksichtigt ist) bzw. mit $N - p - 1$ Freiheitsgraden (für R_A, R_B).

Extrem-Residuum-Test Al-Sarraf und Young ([2]) schlagen vor, nur das dem Betrage nach größte studentisierte Pearson-Residuum $\max_i |r_{Pi}|$ zu betrachten. Da die r_{Pi} nach der Studentisierung (zumindest approximativ) standardnormalverteilt sind, weist ein großer Wert von $\max_i |r_{Pi}|$ auf einen schlechten Fit hin. Da alle N Residuen auf Abweichung von der Normalverteilung geprüft werden, ergibt sich der kritische Wert erst nach einer Bonferroni-Korrektur, d.h. um einen Test zum $(1-\alpha)$-Niveau zu erhalten, vergleichen wir $\max_i |r_{Pi}|$ mit dem $(1 - \frac{\alpha}{2N})$-Quantil der Standardnormalverteilung.

Copas' RSS(Residual Sum of Squares)-Test Hosmer et al. ([28]) schlagen in Anlehnung an Copas ([10]) vor, als globalen Anpassungstest die Summe der rohen quadrierten Residuen

$$R_C = \sum_{i=1}^{M} (y_i - \hat{\pi}_i)^2\,, \tag{73}$$

zu verwenden. Es ist zu beachten, dass die Summe über die Residuen hier über M läuft, d.h. es werden hier prinzipiell ungruppierte Residuen betrachtet. Durch die Nichtberücksichtigung des Nenners im Vergleich zu den Pearson-Residuen (vgl. (65)), wird Beobachtungen mit extremen $\hat{\pi}_i$ relativ mehr Gewicht gegeben. Die asymptotischen Momente der Teststatistik werden nach Standardisierung mit $\sum \hat{\pi}_i(1-\hat{\pi}_i)$ geschätzt durch

$$\hat{\mathrm{E}}\left(R_C - \sum_{i=1}^{M} \hat{\pi}_i(1-\hat{\pi}_i)\right) = 0 \tag{74}$$

$$\widehat{\mathrm{Var}}\left(R_C - \sum_{i=1}^{M} \hat{\pi}_i(1-\hat{\pi}_i)\right) = d_C^t(\hat{W} - \hat{W}X(X^t\hat{W}X)^{-1}X^t\hat{W})d_C\,, \tag{75}$$

$$\tag{76}$$

wobei d_C ein Vektor mit typischem Element $(1-2\hat{\pi}_i)$ ist. Um die Signifikanz zu beurteilen, wird die standardisierte Teststatistik $\frac{R_C - \sum \hat{\pi}_i(1-\hat{\pi}_i)}{\widehat{\mathrm{Var}}\left(R_C - \sum \hat{\pi}_i(1-\hat{\pi}_i)\right)^{1/2}}$ mit dem Quantil der Standardnormalverteilung verglichen.
Die Bezeichnung „RSS“ hat der Test von der Tatsache, dass die Varianz der standardisierten Teststatistik von R_C auch als residuale Quadratsumme (**R**esidual **S**um of **S**quares) einer gewichteten linearen Regression mit Beobachtungsvektor d_C, Designmatrix X und Gewichtsmatrix $\hat{W}$ berechnet werden kann.

3 Vergleich der vorgeschlagenen Anpassungstests

3.1 Bisheriger Kenntnisstand

Es gibt bisher kaum Evidenz darüber, wie sich die vorgeschlagenen globalen Anpassungstests in der Situation von stetigen Kovariablen bzw. fehlenden Messwiederholungen verhalten, weder analytisch noch durch Simulationsuntersuchnungen.
Eine Reihe von Autoren, die Tests vorgeschlagen haben, sichern ihre eigenen Vorschläge durch kleine Simulationen ab. Diese Untersuchungen sind aber limitiert durch kleine Anzahlen von überprüften Szenarios (häufig wird nur das Verhalten unter der Nullhypothese untersucht) und durch fehlende Vergleiche zu Referenztests (z.B zu X^2 oder zum Hosmer-Lemeshow-Test). Desweiteren ist eine gewisse Verzerrung dadurch zu befürchten, dass die Autoren ihre eigenen Vorschläge überprüfen und dadurch, ohne eine Absicht zu unterstellen, eventuell die Objektivität etwas leiden könnte.
Eine Ausnahme ist dabei die groß angelegte Simulationsuntersuchung von Hosmer et al.([28]), die eine ganze Reihe von (auch nicht-globalen) Tests unter verschiedenen Parameterkonstellationen (Anzahl der Kovariablen, Verteilung der Kovariablen, Anzahl der Beobachtungen, verschiedene Modellverletzungen) vergleichen. Dabei zeigt sich, dass unter den Tests, die Hosmer et al. vergleichen, alles in allem R_C (vgl. S.35) und X^2_{McC} (vgl. S.22) am besten abschneiden.
Doch leider weist auch diese Studie einige Mängel auf: Zum einen betrachten die Autoren nur den Fall ohne Messwiederholungen, d.h. $m_i \equiv 1$. Man erhält also keinerlei Information darüber, ob es, analog zur Faustregel der Gültigkeit des χ^2-Unabhängigkeitstests in Vierfeldertafeln, eine *untere* Grenze bzgl. der m_i gibt, *über* der die altbekannten Tests X^2 und D noch zu verlässlichen Ergebnissen führen. Man könnte sich, falls es eine solche Grenze gibt, die stellenweise sehr komplizierten Berechnungen der hier beschriebenen Tests ersparen. Dazu treten noch einige kleinere Mängel, z.B. die niedrige Anzahl von Replikationen (nur 500 Simulationsläufe pro Parameterkonstellation) und die Nichtberücksichtigung von D.

3.2 Eigene Untersuchungen

Angesichts der, eben beschriebenen, begrenzten Erkenntnis zum Verhalten der globalen Anpassungstests und der Tatsache, dass in dieser Arbeit Tests vorgestellt wurden, die noch nie in Simulationsuntersuchungen überprüft wurden, wurde eine eigene Simulationsstudie durchgeführt. Dabei werden die Tests sowohl unter der Nullhypothese eines korrekt spezifizierten Modells als auch unter der Alternative eines fehlspezifizierten Modells miteinander verglichen.
Unter der Nullhypothese wurden folgende Parameter variiert:

- Gesamtstichprobenumfang
- Anzahl der Messwiederholungen
- Anzahl der Kovariablen
- Verteilung der Kovariablen
- Stärke des Regressionseffekts

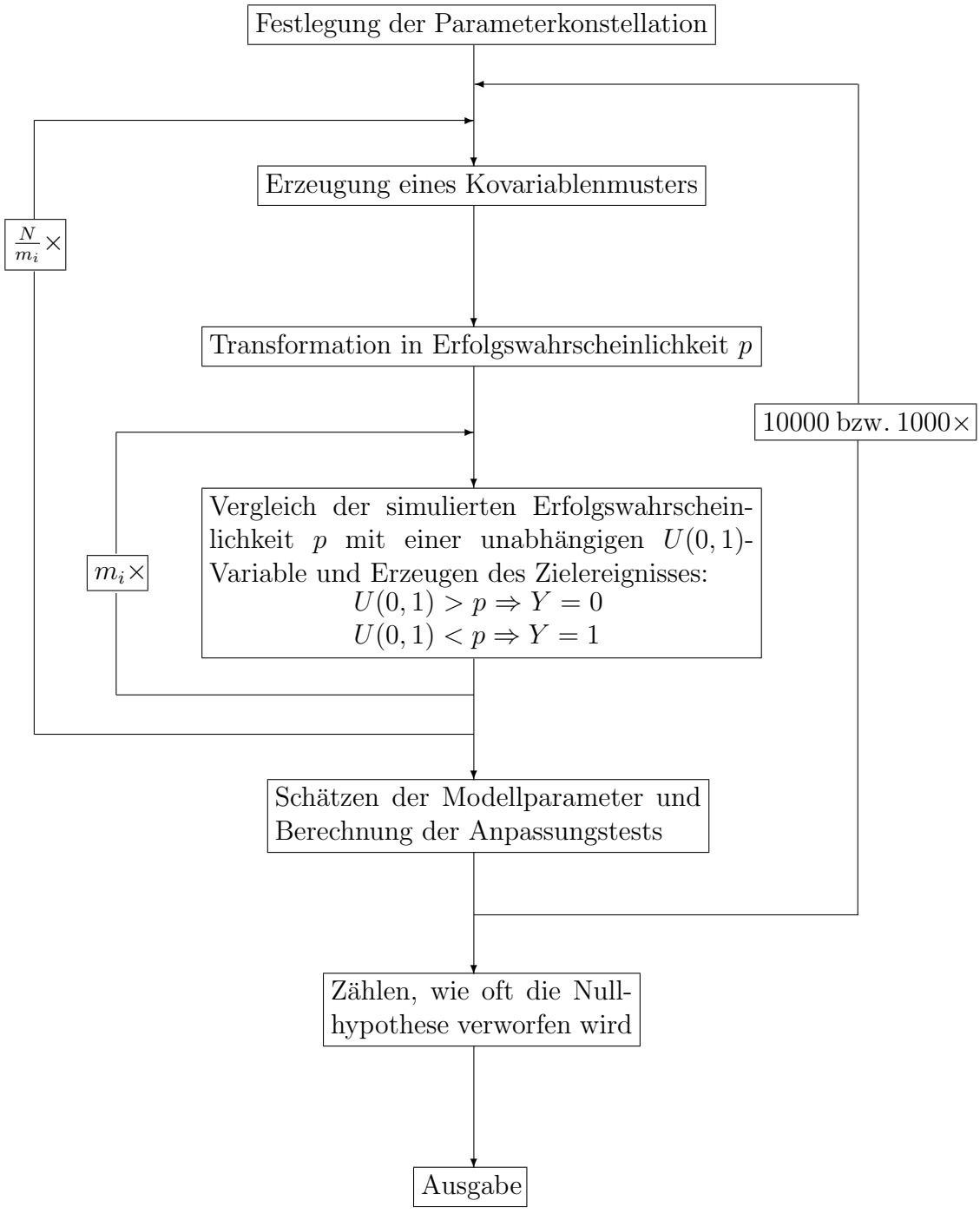
Bei der Festlegung dieser Parameter ist darauf zu achten, dass diese möglichst praxisnahe Werte annehmen, so dass eine Übertragung der erreichten Ergebnisse in die Realität von konkret vorliegenden Datensätzen möglich ist. Dies soll aber nicht soweit gehen, dass ein bestimmter Datensatz bezüglich seiner Eigenschaften konkret nachgebildet wird.
Unter der Alternative werden die Tests bzgl. folgender Modellverletzungen miteinander verglichen:

- Falsch spezifizierte funktionale Form der Kovariablen
- Nicht ins Modell aufgenommene Kovariablen
- Falsch spezifizierte Link-Funktion
- Overdispersion

Auch unter der Alternative ist darauf zu achten, dass möglichst realistische Szenarios überprüft werden, so dass ein Eindruck über die Power der Tests in alltäglichen Situationen entsteht. Es ist trivial, extreme Modellverletzungen zu überprüfen, die von allen Tests gefunden werden. Interessant sind vielmehr

Situationen, in denen sich echte Powerunterschiede zwischen den einzelnen Tests ergeben.
Von zentralem Interesse ist dabei das Verhalten der Tests, wenn die Anzahl der Messwiederholungen variiert. Es ist zu hoffen, dass eine untere Grenze bzgl. der m_i existiert , *über* der die altbekannten und routinemäßig abrufbaren Standard-Verfahren noch zu verlässlichen Ergebnissen führen, so dass die in einigen Fällen aufwendige Berechnung bzw. Programmierung der hier vorgestellten Tests nicht notwendig wird. Weiterhin wäre es wünschenswert, dass *unterhalb* dieser Grenze sich einer der vorgeschlagenen Alternativen für die Standardtests den anderen als überlegen erweist, so dass dieser in Zukunft zur allgemeinen Anwendung empfohlen werden kann. Dabei ist zu beachten, dass dieser Test nicht zu kompliziert ist, denn „the reality of model construction demands that diagnostic and specification tests be neither expensive nor cumbersome to construct. Once methods begin to cause trouble on either of these criteria, they are likely to be ignored.“ (A.R. Pagan, [23])
Die Simulationsuntersuchung wurde mit SAS/IML durchgeführt. Dabei wird (vgl. Abbildung 6) für eine vorgegebene Parameterkonstellation in einem ersten Schritt ein Kovariablenmuster zufällig erzeugt, das m_i individuelle Beobachtungen enthält. Aus den Werten der Kovariablen wird gemäß des spezifizierten Modells eine Erfolgswahrscheinlichkeit p für dieses Kovariablenmuster berechnet. Für jede der m_i individuellen Beobachtungen wird sodann unabhängig eine $U(0,1)$-verteilte Zufallszahl erzeugt, mit der die berechnete Erfolgswahrscheinlichkeit p verglichen wird: Ist p größer als diese Zufallszahl, so liegt ein Erfolg vor, das heisst die Zielgröße hat den Wert 1, dementsprechend den Wert 0, falls p kleiner als die erzeugte Zufallszahl ist. Dieser Schritt des Erzeugens der Zielgröße für eine individuelle Beobachtung wird m_i-mal innerhalb jedes Kovariablenmuster und für N/m_i verschiedene Kovariablenmuster wiederholt, so dass sich ein Datensatz mit N verschiedenen Kovariablenmustern und insgesamt M individuellen Beobachtungen ergibt, von denen jeweils m_i das selbe Kovariablenmuster besitzen.
Für diesen Datensatz werden die Modellparameter geschätzt und die beschriebenen Anpassungstests berechnet. Auf diese Art und Weise werden 10000 bzw. 1000 (10000 für Konstellationen mit $M = 100$, 1000 für Konstellationen mit $M = 500$) Datensätze erzeugt und gezählt, wie oft die Anpassungstests die Nullhypothese der korrekten Modellspezifikation verwerfen: Bei Vorliegen der Nullhypothese wird man erwarten, dass dies in 5% der Fälle geschieht, bei Verletzung der Nullhypothese wird man hoffen, dass dies möglichst oft geschieht.

Abbildung 6: Ablauf des Simulationsprogrammes



3.3 Situationen unter der Nullhypothese

3.3.1 Abgeprüfte Parameterkonstellationen

Die im letzten Abschnitt vorgestellten Parameter Gesamtstichprobenumfang, Anzahl der Messwiederholungen, Anzahl der Kovariablen, Verteilung der Kovariablen und die Stärke des Regressionseffekts wurden folgendermaßen variiert, um einen möglichst umfassenden Einblick in die Eigenschaften der vorgeschlagenen Anpassungstests zu erhalten.

- **Gesamtstichprobenumfang:** Dieser variiert zwischen zwei verschiedenen Einstellungen, und zwar zwischen $M = 100$ und $M = 500$. Dies spiegelt die durchschnittliche Fallzahl in den untersuchten Datensätzen wieder, die in der Abteilung Klinische Sozialmedizin analysiert wurden (vgl. dazu auch die Beispiele in Kapitel 4).

- **Anzahl der Messwiederholungen:** Hier wurden zum einen vier Einstellungen mit konstantem m_i ($m_i = 1, 2, 5, 10$) gewählt, desweiteren zwei Einstellungen mit variierendem m_i: Eine bildet einen Übergang zwischen den Situationen mit $m_i \equiv 1$ und $m_i \equiv 2$, indem das Verhältnis zwischen einfach und doppelt besetzten Kovariablenmustern auf $1:1$ gesetzt wird. Die andere ahmt die in den Beispielstudien typischerweise beobachtete Besetzung der Kovariablenmuster nach: Hier wird das Verhältnis zwischen einfach, doppelt, fünffach und zehnfach besetzten Kovariablenmuster auf $64:21:9:6$ eingestellt.

- **Anzahl der Kovariablen:** In diesem Fall werden zum einen Modelle mit einer einzelnen Kovariablen, zum anderen Modelle mit drei Kovariablen untersucht, diese werden in beiden Fällen als metrische bzw. stetige Variablen generiert und als solche geschätzt. Zusätzlich wird in allen Modellen ein konstanter Faktor mitgeschätzt, der in allen Simulationsläufen konstant auf 0 gesetzt wurde.

- **Verteilung der Kovariablen:** Hier werden insgesamt fünf verschiedene Szenarios abgeprüft. Im Falle von Modellen mit einer Kovariable wird diese entweder als $U(-6, 6)$-, $N(0, 1)$- oder als χ^2_4-verteilt eingestellt. Dies orientiert sich an den Vorgaben von Hosmer et al. ([28]), die ebenfalls auf diese Verteilung der Kovariablen zurückgreifen. Man beachte auch, dass in den ersten beiden Fällen eine symmetrische, im dritten Fall eine unsymmetrische Verteilung vorliegt.

Im Falle der Modelle mit drei Kovariablen werden diese in einem Fall als unabhängig und identisch $U(-6, 6)$-, im anderen Fall als unabhängig und identisch $N(0, 1)$-verteilt generiert.

- **Stärke des Regressionseffekts:** Dieser variiert nur in den Modellen mit einer Kovariable und zwar dergestalt, dass zwei verschiedene Stärken des Regressionseffekts eingestellt werden. Zusätzlich wurde noch eine Einstellung untersucht, in der überhaupt kein Einfluss einer Kovariable vorliegt.

3.3.2 Verglichene Anpassungstests

Bei der Darstellung der Ergebnisse wird aus Gründen der Übersichtlichkeit bereits eine gewisse Vorauswahl bezüglich der Darstellung getroffen. Nicht dargestellte Tests erbringen entweder keine neuen Erkenntnisse oder schneiden im Vergleich zu direkten Konkurrenten schlechter ab. Konkret dargestellt werden

- Die beiden klassischen Anpassungstests:
  - X^2, der Pearson-Test mit der herkömmlichen χ^2-Verteilung als Prüfverteilung, vgl. S.11, (7)
  - D, die Devianz mit der herkömmlichen χ^2-Verteilung als Prüfverteilung, vgl. S.11, (8)
- Aus der Gruppe der Tests mit modifizierter Prüfverteilung:
  - X^2_O, der Test von Osius/Rojek mit der asymptotischen Normalverteilung als Prüfverteilung von X^2, vgl. S.18, (15)
  - X^2_{OEd}, der Test von Osius/Rojek mit der asymptotischen Normalverteilung und Edgeworth-Expansion als Prüfverteilung von X^2, vgl. S.18, (16)
  - X^2_{OSkal}, der Test von Osius/Rojek, der einen Kompromiss zwischen der asymptotischen Normalverteilung und der herkömmlichen χ^2-Verteilung von X^2 macht, vgl. S.20, (28)
  - X^2_{McC}, der Test von McCullagh mit der bedingten asymptotischen Normalverteilung als Prüfverteilung von X^2, vgl. S.22, (40)

  - X^2_{McCEd}, der Test von McCullagh mit der bedingten asymptotischen Normalverteilung und Edgeworth-Expansion als Prüfverteilung von X^2, vgl. S.22, (41)
  - D_F, der Test von Snapinn und Small mit der F-Verteilung als Prüfverteilung von D, vgl. S.22, (43)

- Aus der Gruppe der Tests mit Gruppierung von Beobachtungen:
  - $\hat{C}$, der Test von Hosmer/Lemeshow, vgl. S.24, (45)

- Aus der Gruppe der Tests mit anderen Teststatistiken:
  - D_C, die Bartlett-Korrektur von D von Cordeiro, vgl. S.28, (49)
  - X^2_F, Farrington's Modifikation von X^2, vgl. S.29, (57)
  - X^2_{FEd}, Farrington's Modifikation von X^2 mit zusätzlicher Edgeworth-Expansion, vgl. S.29, (58)
  - IM, der Informationsmatrix-Test mit allen $\frac{1}{2}(p+1)(p+2)$ verschiedenen Elementen der Informationsmatrix, vgl. S.31, (63)
  - IM_{DIAG}, der Informationsmatrix-Test mit den $(p+1)$Elementen der Hauptdiagonalen der Informationsmatrix, vgl. S.32, (2.3.3)
  - R_P, der Anpassungstest, der durch Aufsummierung der studentisierten Pearson-Residuen entsteht, vgl. S.34, (67)
  - R_D, der Anpassungstest, der durch Aufsummierung der studentisierten Devianz-Residuen entsteht, vgl. S.34, (68)
  - R_L, der Anpassungstest, der durch Aufsummierung der Likelihood-Residuen entsteht, vgl. S.34, (69)
  - R_A, der Anpassungstest, der durch Aufsummierung der Anscombe-Residuen entsteht, vgl. S.35, (71)
  - R_B, der Anpassungstest, der durch Aufsummierung der bias-adjustierten Devianz-Residuen entsteht, vgl. S.35, (72)
  - $\max_i |r_{Pi}|$, der Extrem-Residuum-Test von AlSarraf/Young, vgl. S.35
  - R_C, der RSS-Test von Copas, der durch Aufsummierung der rohen Pearson-Residuen entsteht, vgl. S.35, (73)

3.3.3 Ergebnisse

Die Ergebnisse sind in Anhang A dargestellt. Dargestellt ist dabei das empirische Signifikanzniveau der oben beschriebenen Anpassungstests, d.h. die Anzahl (in %), in der der jeweilige Anpassungstest die Nullhypothese verworfen hat. Dieses sollte in diesem Fall (unter der Nullhypothese einer korrekten Modellanpassung und bei einem Signifikanzniveau der berechneten Anpassungstests von $\alpha = 0.05$) gleich 5 sein. Eine Zahl unter 5 weist auf ein zu konservatives, eine Zahl über 5 auf eine zu liberales Verhalten des Tests hin. In den Spalten sind die Belegungen der Kovariablenmuster abgetragen. Dabei stehen konstante m_i $(1, 2, 5, 10)$ für Belegungen, in denen alle Kovariablenmuster die selbe Belegung haben. Die Bezeichnung „1-2“ steht für eine Belegung, in dem eine Hälfte der Kovariablenmuster einfach, die andere Hälfte doppelt belegt sind. Die Bezeichnung „1-10“ steht für eine Belegung, in der die Kovariablenmuster einfach, doppelt, fünffach und zehnfach im Verhältnis $64:21:9:6$ belegt sind.
Es kann aus Gründen der Übersichtlichkeit nicht auf jede einzelne Zahl aus dieser Vielzahl von Ergebnistabellen eingegangen werden. Folgende Tendenzen sind aber sichtbar:
Die Pearson-Statistik X^2 mit der herkömmlichen χ^2-Verteilung liefert erst ab $m_i \geq 5$ verlässliche Ergebnisse. Bei kleinerer Besetzung der m_i ist der Test in der Regel zu konservativ. Dies beruht auf der Tatsache, dass im Falle von $m_i \equiv 1$ $X^2 \approx N$ (vgl. (10), S. 12) gilt und in diesem Falle die Nullhypothese nie verworfen wird. Die dadurch bedingte Konservativität zeigt sich bereits bei $m_i = 2$.
Die Devianz D ist als Anpassungstest im logistischen Regressionsmodell völlig unbrauchbar. Dies war nach den Vorbetrachtungen (vgl. (9), S. 12) bereits zu erahnen. Wir finden für D krasseste Niveauüberschreitungen auf bis zu 100% und in anderen Fällen deutliche Unterschreitungen bis zu 0%. Das Verhalten von D wird umso extremer, je mehr m_i gegen 1 geht.
X_O^2, der Test von Osius/Rojek mit der asymptotischen Normalverteilung als Prüfverteilung von X^2, hält das Niveau von 5% in der Mehrzahl der Fälle ein. Es ist eine leichte Tendenz zu einem konservativen Verhalten zu beobachten, wenn m_i größer wird, dies gilt aber nur für $M = 100$. Im Falle von drei Kovariablen zeigt sich für $m_i = 1$ ein leicht liberales und dann ein stetiger Abstieg zu einem leicht konservativen Verhalten bei $m_i = 10$.
Die Edgeworth-Korrektur von X_O^2, X_{OEd}^2, zeigt ein ähnliches Verhalten wie X_O^2 bezüglich des Abfallens des Niveaus von großen zu kleinen m_i. Jedoch

sind die Extreme hier stärker ausgebildet, so finden wir bei $m_i = 1$ doch erhebliche Niveauüberschreitungen, bis zu 28% im Modell mit einer χ^2_4-verteilten Kovariable. Diese Niveauverletzungen nivellieren sich bei großen Fallzahlen und bei symmetrisch verteilten Kovariablen.
X^2_{OSkal} als Übergang zwischen X^2 und X^2_O zeigt ein sehr gutes Verhalten und hält das Niveau in beinahe allen Fällen ein. Zwei Ausnahmen sind eine Unterschreitung bei einem fehlenden Regressionseffekt und $m_i = 1$, wo X^2_{OSkal} wohl zu X^2 und dessen konservativem Verhalten tendiert und ein zu liberales Verhalten (auch in diesem Fall analog zu X^2) bei einer $U(-6, 6)$-verteilten Kovariable und starkem Regressionseffekt.
Die Modifikation der Pearson-Statistik von McCullagh, X^2_{McC}, hält das Niveau sehr gut ein. Das größte beobachtete Niveau über alle Parameterkonstellationen ist 7.8, das kleinste bei 3.4, wobei aber kaum eine Systematik zu beobachten ist. Diese minimalen Abweichungen vom gewünschten empirischen Niveau von 5 sind wohl als zufällige Schwankungen anzusehen.
Analog zu X^2_{OEd} zeigt auch die Edgeworth-Korrektur des McCullagh-Tests, X^2_{McCEd}, eine Empfindlichkeit gegen schief verteilte Kovariablen, was sich durch eine Niveauüberschreitung bemerkbar macht. Aber auch X^2_{McCEd} hält in der Mehrzahl der Fälle das Niveau ein.
Die Snapinn/Small-Korrektur D_F, bei der die Devianz D mit einer F-Verteilung geprüft wird, zeigt wie D selber ein unbefriedigendes Verhalten. Auch hier finden wir empirische Niveaus zwischen 0% und 100%, wobei sich keine eindeutige Systematik zeigt, außer der, dass das Niveau von D_F definitonsbedingt immer unter dem von D liegt. Offenbar ist die F-Verteilung als Prüfverteilung für die Devianz ebenso ungeeignet wie die χ^2-Verteilung bzw. können auch mit der Hilfe der F-Verteilung die statistischen Probleme von D (vgl. (9)) nicht gelöst werden.
Der Hosmer/Lemeshow-Test $\hat{C}$ hält in der überwiegenden Anzahl von Parameterkonstellationen das vorgegebene Niveau ein. Zwei kleine Ausnahmen sind Niveauunterschreitungen bei den beiden Modellen mit drei Kovariablen, $M = 100$ und $m_i = 10$, wo das empirische Niveau unter 2% liegt.
Die Bartlett-Korrektur der Devianz von Cordeiro, D_C, zeigt in über 60% aller Fälle ein empirisches Niveau von 0% und ist damit bei weitem zu konservativ. Ein einigermaßen zufriedenstellendes Bild ergibt sich erst bei $m_i = 10$, aber auch dann nicht in allen Fällen.
Die Korrektur der Pearson-Statistik von Farrington, X^2_F, hält das vorgegebene Niveau in allen Fällen mit $m_i \neq 1$ ein. Es ist allenfalls eine ganz geringe Tendenz zur Überschätzung bei großen m_i zu beobachten. Zur Erinnerung,

im Falle von $m_i \equiv 1$ ist $X_F^2 \equiv N$, wodurch die Nullhypothese nie verworfen wird und das empirische Niveau notwendigerweise gleich Null ist. Dieses Handicap hat X_F^2 aber bereits bei der Parameterkonstellation „1-2“, wo zur Hälfte Kovariablenmuster mit $m_i \equiv 1$ und zur anderen Hälfte mit $m_i \equiv 2$ vorkommen, wieder wettgemacht und hält das Niveau zufrieden stellend ein.
Die Edgeworth-Korrektur der Farrington-Statistik, X_{FEd}^2, verhält sich ähnlich wie X_F^2. Auch hier ist das empirische Niveau gleich Null für $m_i = 1$, in den anderen Fällen wird das Niveau eingehalten.
Der Informationsmatrix-Test IM, der alle voneinander verschiedenen Elemente der Informationsmatrix in die Berechnung der Teststatistik mit einbezieht, hält das Niveau über alle m_i gut ein. Einzige Ausnahmen sind ein leicht konservatives Verhalten bei den Modellen ohne Regressionseinfluss und bei den Modellen mit drei Kovariablen und $M = 100$.
IM_{DIAG}, die Version des Informationsmatrix-Tests, die nur die Elemente auf der Hauptdiagonalen der Informationsmatrix mit einbezieht, schneidet über alle Parameterkonstellationen sehr gut ab. Sie zeigt auch keine Schwächen bei den Modellen, wo IM ein leicht konservatives Verhalten zeigt.
Der Anpassungstest R_P, der durch Aufsummierung der studentisierten Pearson-Residuen zustandekommt, zeigt ein ähnliches Verhalten wie X^2, ist also bei kleinen m_i zu konservativ und hält das Niveau erst ab $m_i \geq 5$ ein. Das empirische Niveau liegt dabei immer minimal über dem von X^2. Dies liegt an der Definition der studentisierten Person-Residuen, die die echten Pearson-Residuen um einen Faktor $\frac{1}{1-h_{ii}}$ mit kleinen, positiven h_{ii} vergrößern. Offensichtlich reicht diese Korrektur aber nicht aus, um die Konservativität von X^2 bei kleinen m_i zu korrigieren.
Ein analoges Verhalten zeigt der Anpassungstest aus den studentisierten Devianz-Residuen, R_D, dessen Niveau immer minimal größer als das von D ist. Dadurch ist dieser Test wie die Devianz D als Anpassungstests völlig unbrauchbar.
Die Teststatistik R_L wird aus Residuen gebildet, die einen Übergang zwischen den studentisierten Pearson- und Devianzresiduen bilden. Wie oben bereits ausgeführt, liegen diese aufgrund ihrer Definition näher bei den Devianz-Residuen als bei den Pearson-Residuen. Dementsprechend ist das Verhalten von R_L ähnlich schlecht wie das von D und R_D.
Das Verhalten von R_A, des Anpassungstests, der auf der Summation der Anscombe-Residuen basiert, ist ebenfalls völlig unbefriedigend. Für diesen finden wir in über 90% der untersuchten Parameterkonstellationen ein empirisches Niveau von über 10%, d.h. der Test lehnt die Modelle viel zu häufig

ab.
Der Anpassungstest R_B, der aus den bias-adjustierten Devianz-Residuen besteht, liefert in allen Fällen ein sehr ähnliches Ergebnis wie D und ist dadurch ebenfalls als Anpassungstest nicht geeignet.
Die Extremwert-Statistik $\max_i |r_{Pi}|$, die das dem Betrage nach größte studentisierte Pearson-Residuum zur Grundlage eines Anpassungstest macht, überschreitet in den meisten Fällen das nominale Niveau deutlich und ist als Anpassungstest ebenfalls ungeeignet. Ein einigermaßen brauchbares Verhalten zeigt die Teststatistik nur für $m_i = 10$. Dies liegt daran, dass die Pearson-Residuen für kleine m_i nicht normalverteilt sind ([29]).
Der Anpassungstest R_C, der auf den rohen Pearson-Residuen beruht, hält dagegen in praktisch allen Fällen das empirische Niveau ein. Diese Beobachtung wurde bereits von Hosmer et al. ([28]) gemacht. Leichte Unterschätzungen des Niveaus gibt es allenfalls bei den Modellen mit χ^2-verteilter Kovariable und $M = 100$.

Zusammenfassend lässt sich feststellen:
Die überprüften Anpassungstests spalten sich bezüglich des Verhaltens unter der Nullhypothese in zwei Gruppen auf:
Die Tests in der ersten Gruppe (X^2_O, X^2_{OEd}, X^2_{OSkal}, X^2_{McC}, X^2_{McCEd}, $\hat{C}$, X^2_F, X^2_{FEd}, IM, IM_{DIAG}, R_C) halten das vorgegebene Niveau über praktisch alle untersuchten Parameterkonstellationen ein. Für die Tests aus dieser Gruppe lohnt sich eine weitergehende Untersuchung auf das Verhalten unter der Alternativhypothese.
Diese Tests sind dabei robust gegen unterschiedliche Fallzahlen, gegen unterschiedliche Verteilungen der Kovariablen, gegen unterschiedliche Anzahlen der Kovariablen und gegen die Stärke des Regressionseffekts. Insbesondere sind die Tests robust gegen fehlende Messwiederholungen, was sie den Standardtests X^2 und D überlegen macht.
Die Tests der zweiten Gruppe (X^2, D, D_F, D_C, R_P, R_D, R_L, R_A, R_B, $\max_i |r_{Pi}|$) zeigen ein unbefriedigendes Verhalten bereits unter der Nullhypothese und sind als Anpassungstests für das logistische Modell ungeeignet.
Insbesondere von Interesse ist, dass die beiden Standardtests X^2 und D ebenfalls in diese Gruppe fallen. Der Pearson-Test X^2 zeigt dabei immerhin bis zu Situationen mit 5 Messwiederholungen pro Kovariablenmuster ein akzeptables Verhalten, ist bei weniger Messwiederholungen in der Regel aber zu konservativ. Die Devianz D ist als Anpassungstest im logistischen Regressionsmodell völlig ungeeignet.

3.4 Situationen unter der Alternative

3.4.1 Abgeprüfte Parameterkonstellationen

Der entscheidende Parameter für die Beurteilung der Anpassungstests unter der Alternative, d.h. unter einer Fehlspezifikation des Modells, ist selbstverständlich die *Art* der Fehlspezifikation, die in den Simulationen zur Erzeugung der Datensätze eingestellt wird. Auch hier sollte wieder eine Vielzahl von Konstellationen abgeprüft werden, um einen möglichst umfassenden Einblick in das Verhalten der Anpassungstests zu erhalten. Ausgewählt wurden letztendlich vier verschiedene Fehlspezifikationen, nämlich eine falsch spezifizierte funktionale Form einer Kovariablen, eine nicht ins Modell aufgenommene Kovariable, eine vorliegende Overdispersion und eine falsch spezifizierte Link-Funktion.
Umgesetzt werden diese Abweichungen dann dergestalt, dass die abzuprüfenden Datensätze unter diesen Fehlspezifikationen erzeugt werden, sodann aber Modelle geschätzt werden, die diese Abweichung nicht berücksichtigen. Es liegt dann eine „schlechte“ Anpassung des Modells vor, da die jeweilige „Verunreinigung“ des Datensatzes bei der Schätzung der Parameter nicht berücksichtigt wird. Diese schlechte Anpassung des Modells an die vorliegenden Daten sollte von den beteiligten Anpassungstests entdeckt werden. Wir erhoffen uns also in den Situationen unter der Alternative ein möglichst häufiges Verwerfen der Nullhypothese, da diese ja in der Realität nicht vorliegt.
Ausgehend von der Fragestellung nach einer eventuellen Grenze der m_i (vgl. 3.2, S. 39) über der die Standardverfahren zu immer noch verlässlichen Ergebnissen führen, wurde auch unter der Alternative die Anzahl der Messwiederholungen in der altbekannten Weise variiert. Desweiteren untersucht wurde auch wieder eine Änderung des Gesamtstichprobenumfangs.
Konkret wurden die oben angeführten Parameter wie folgt variiert:

- **Art der Fehlspezifikation:** Geschätzt wurde in allen Fällen ein Modell mit einem konstanten Faktor und einer weiteren Kovariable, d.h. die Modellschätzung ging von einem Modell $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1$ aus.

  Die Datensätze wurden aber erzeugt unter den folgenden Modellannahmen bzw. mit den folgenden Fehlspezifikation. Das heisst, dargestellt sind jeweils die „wahren“ Modelle. Dabei war der konstante Faktor β_0 immer auf Null gesetzt, die Kovariable x_1 war $U(-6, 6)$-verteilt.

- **Falsch spezifizierte funktionale Form der Kovariablen:** Die wahre Modellgleichung ist $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1^2$, d.h. die Kovariable geht quadratisch ins wahre Modell ein, wird aber linear geschätzt.
- **Nicht ins Modell aufgenommene Kovariablen:** Die wahre Modellgleichung ist $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, wobei $x_2 \sim U(-6, 6)$ und $\beta_1 = 2\beta_2$. Es wurde also eine zweite unabhängig $U(-6, 6)$-verteilten Kovariable erzeugt, die mit einem halb so grossen Regressionseffekt ins Modell einging.
- **Overdispersion:** Die wahre Modellgleichung ist $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1$, wobei β_0 jetzt nicht mehr als fest und nicht-zufällig eingestellt wird, sondern zufällig mit $E(\beta_0) = 0$ und $\text{Var}(\beta_0) = 0.323$. Dieses Modell für Overdispersion geht auf Williams zurück([55]). Durch die Aufnahme des zufälligen konstanten Faktors in die Modellgleichung wird die Varianz der Zielgröße aufgebläht zu

$$\text{Var}(Y) \simeq m_i \pi_i (1 - \pi_i)[1 + \text{Var}(\beta_0)(m_i - 1)\pi_i(1 - \pi_i)] \,. \qquad (77)$$

  Daraus wird ersichtlich, dass Overdispersion nur entstehen kann, falls $m_i > 1$ für mindestens ein i gilt. Wir erwarten also keine Power der Anpassungstests bei den Parameterkonstellationen mit $m_i \equiv 1$.
- **Falsch spezifizierte Link-Funktion:** Die wahre Modellgleichung ist $\log[-\log(1 - \pi_i)] = \beta_0 + \beta_1 x_1$. Hier wird die Erfolgswahrscheinlichkeit π_i mit der so genannten *komplementären LogLog-Linkfunktion* transformiert. Diese Linkfunktion bietet sich bei einer Reihe von Fragestellungen an (vgl. [9], S. 112), die für den medizinischen Bereich interessanteste ist dabei die der Analyse von gruppierten Überlebenszeiten.

- **Gesamtstichprobenumfang:** Dieser variiert wieder, analog zu den Parameterkonstellationen unter der Nullhypothese zwischen den beiden Einstellungen $M = 100$ und $M = 500$. Damit ist die Möglichkeit des Vergleichs zwischen dem Verhalten der Tests unter Nullhypothese und Alternative gegeben.

- **Anzahl der Messwiederholungen:** Auch in diesem Fall wurden die Einstellungen aus den Situationen unter der Nullhypothese übernom-

men. Das heisst, es wurden zum einen vier Einstellungen mit konstantem m_i ($m_i = 1, 2, 5, 10$) gewählt, desweiteren zwei Einstellungen mit variierendem m_i: Eine bildet einen Übergang zwischen den Situationen mit $m_i \equiv 1$ und $m_i \equiv 2$, indem das Verhältnis zwischen einfach und doppelt besetzten Kovariablenmustern auf $1:1$ gesetzt wird. Die andere ahmt die in den Beispielstudien typischerweise beobachtete Besetzung der Kovariablenmuster nach: Hier wird das Verhältnis zwischen einfach, doppelt, fünffach und zehnfach besetzten Kovariablenmuster auf $64:21:9:6$ eingestellt.

3.4.2 Verglichene Anpassungstests

Einbezogen werden nur die Tests, die unter der Nullhypothese ein zufriedenstellendes Verhalten gezeigt haben. Das sind aus der Gruppe der Tests mit modifizierter Prüfverteilung die Tests X^2_O, X^2_{OEd}, X^2_{OSkal}, X^2_{McC} und X^2_{McCEd}, der Hosmer-Lemeshow-Test $\hat{C}$ aus der Gruppe mit Tests mit gruppierten Beobachtungen und die Tests X^2_F, X^2_{FEd}, IM, IM_{DIAG} und R_C aus der dritten Gruppe mit neuer Teststatistik.

3.4.3 Ergebnisse

Die Ergebnisse sind in Anhang B dargestellt. Dargestellt ist dabei das empirische Signifikanzniveau der oben beschriebenen Anpassungstests, d.h. die Anzahl (in %), in der der jeweilige Anpassungstest die Nullhypothese verworfen hat. Dieses sollte in diesem Fall (unter der Alternative von vorliegenden Fehlspezifikationen) möglichst groß sein, wobei maximal die Zahl 100 erreicht werden kann. In diesem Fall hat ein Test die Fehlspezifikation in allen Fällen entdeckt. Zur Erinnerung, in den Situationen unter der Alternative wird jeweils ein Modell mit einem festen, nicht zufälligen Intercept und einer einzelnen Kovariable geschätzt.
In den Spalten sind die Belegungen der Kovariablenmuster abgetragen. Dabei stehen konstante m_i $(1, 2, 5, 10)$ für Belegungen, in denen alle Kovariablenmuster die selbe Belegung haben. Die Bezeichnung „1-2“ steht für eine Belegung, in dem eine Hälfte der Kovariablenmuster einfach, die andere Hälfte doppelt belegt sind. Die Bezeichnung „1-10“ steht für eine Belegung, in der die Kovariablenmuster einfach, doppelt, fünffach und zehnfach im Verhältnis $64:21:9:6$ belegt sind.
Auch hier soll aus Gründen der Übersichtlichkeit nicht auf jede einzelne Zahl

aus der Vielzahl von Ergebnistabellen eingegangen werden. Es zeigen sich jedoch folgende Tendenzen:
Die Tests aus der ersten Gruppe der Tests mit modifizierter Prüfverteilung X^2_O, X^2_{OEd}, X^2_{OSkal}, X^2_{McC} und X^2_{McCEd} zeigen ein sehr ähnliches Verhalten unter der Alternative. Sie sind vollkommen unsensibel gegen eine falsch spezifizierte funktionale Form der Kovariablen und eine fehlspezifizierte Linkfunktion, dies gilt für alle Formen der Verteilung der m_i. Gute Power haben die genannten Tests bei Modellen mit einer nicht ins Modell aufgenommenen Kovariablen und gegen Overdispersion. Leider versagen diese Tests komplett bei den Kovariablenmustern mit $m_i \equiv 1$, wo in keinem Fall das 5%-Niveau überschritten werden kann, d.h. die Tests also nur ihr Verhalten unter der Nullhypothese widerspiegeln.
Insgesamt sind die Varianten der Tests, die auf McCullagh zurückgehen, X^2_{McC} und X^2_{McCEd} den Osius-Tests, X^2_O, X^2_{OEd} und X^2_{OSkal}, etwas überlegen.
Es lohnt sich in keinem Fall, den beträchtlichen Aufwand der Berechnung der höheren Momente der Teststatistiken auf sich zu nehmen, um die vorgeschlagenen Edgeworth-Korrekturen zu berechnen. Beide Edgeworth-Varianten (X^2_{OEd} und X^2_{McCEd}) sind ihren unkorrigierten Partner-Statistiken X^2_O und X^2_{McC} unterlegen.
Die skalierte Version des Osius-Tests X^2_{OSkal} liegt von der empirischen Power zwischen den herkömmlichen Osius-Tests und den McCullagh-Tests.
Der Hosmer/Lemeshow-Test $\hat{C}$ zeigt in allen abgeprüften Situationen im Vergleich zu den anderen Tests ein durchschnittliches Verhalten. Sein Verhalten ist vom Gesamtstichprobenumfang abhängig und zwar insofern, als dass er höhere Power bei größeren Fallzahlen hat. Sein Verhalten über die verschiedenen Konfigurationen der m_i ist uneinheitlich. Bei einer falsch spezifizierten funktionalen Form der Kovariablen und bei einer fehlspezifizierten Linkfunktion finden wir über alle m_i ähnliche Power, in den beiden anderen Fällen steigt die beobachtete Power mit den m_i an.
Die beiden Tests von Farrington, X^2_F und X^2_{FEd} sind den Varianten des Pearson-Tests von Osius und McCullagh überlegen: In den Situationen, wo diese gut abgeschnitten hatten, sind die Farrington-Tests gleichwertig, in den Situationen, wo erstere schlecht abgeschnitten hatten, haben X^2_F und X^2_{FEd} doch noch einige Power, um die eingestellten Fehlspezifikationen zu entdecken. Vor allem bestätigt sich hier auch die Beobachtung von Farrington (vgl. 2.3.2, S. 30) der sehr guten Power von X^2_F und X^2_{FEd} gegen Overdispersion. Analog zu den Tests von Osius und McCullagh ist auch hier die

unkorrigierte Version, X^2_F, der mit Edgeworth-Korrektur, X^2_{FEd}, überlegen. Getrübt wird der Gesamteindruck der Farrington-Tests natürlich durch ihre Schwäche in den Situationen mit $m_i \equiv 1$, wo X^2_F und X^2_{FEd} definitionsbedingt nie das Modell verwerfen. Aber wie unter der Nullhypothese hat sich dieses Verhalten bereits bei der Konfiguration m_i =1-2 relativiert.
Die Tests IM und IM_{DIAG} zeigen im Vergleich zu den anderen Tests ein gutes Verhalten in den Modellen mit falsch spezifizierter funktionaler Form der Kovariablen und bei einer fehlspezifizierten Linkfunktion. Im letzterem Fall sind sie sogar allen anderen Test überlegen. Demgegenüber stehen Schwächen bei den Modellen mit nicht mit ins Modell aufgenommener Kovariable und bei denen mit Overdispersion, wo die IM-Tests den anderen Tests unterlegen sind. Wie nach der Definition der Tests nicht anders zu erwarten, zeigen diese ein konstantes Verhalten über die m_i.
Die Version IM_{DIAG}, die zur Berechnung der Teststatistik nur die Elemente auf der Hauptdiagonalen der Informationsmatrix benutzt, ist dabei der Vollversion IM überlegen. Insgesamt zeigen die beiden Tests auch eine beträchtliche Abhängigkeit vom Gesamtstichprobenumfang: In den Situationen, wo die beiden Tests den anderen Tests überlegen sind, wird der Unterschied erst bei den Modellen mit $M = 500$ richtig sichtbar.
Der Anpassungstest R_C, der auf den rohen Pearson-Residuen beruht, zeigt ein extrem schwankendes Verhalten, wobei er insgesamt ähnliche Tendenzen zeigt wie die IM-Tests. Diese sind: gutes Verhalten bei falsch spezifizierte Form der Kovariablen und bei fehlspezifizierter Linkfunktion, schlechteres Verhalten bei nicht ins Modell aufgenommener Kovariable und bei Overdispersion. Alles in allem ist R_C den IM-Tests jedoch leicht unterlegen.
Das Verhalten der untersuchten Tests unter der Alternative zusammenfassend lässt sich feststellen:
Auch wenn die Unterschiede zwischen den untersuchten Tests nicht dramatisch sind, kristallisieren sich drei Anpassungstests, der Farrington-Test X^2_F, IM_{DIAG} und R_C als den anderen leicht überlegen heraus.

4 Anwendungsbeispiele

Im folgenden werden drei real existierende Datensätze vorgestellt, anhand derer die Anwendung der vorgestellten Anpassungstests demonstriert werden soll. Weiterhin werden die Implikationen aus den Ergebnissen der Simulationsuntersuchungen diskutiert.

4.1 Berufsbedingte Handekzeme in der Automobilindustrie

Ziel dieser Studie war die Beurteilung von exogenen (vor allem Arbeitsbelastungen) und endogenen (genetische Disposition) Risikofaktoren bei der Entstehung von berufsbedingten Handekzemen in der Automobilindustrie. Dazu wurden zwischen 1990 und 1998 alle 2078 Auszubildenden der AUDI AG, die ihre Ausbildung in Ingolstadt 1990-1994 und Neckarsulm 1991-1994 begonnen hatten, im Rahmen einer prospektiven Kohortenstudie standardisiert untersucht. Untersuchungen erfolgten zu Beginn, nach dem ersten Ausbildungsjahr und am Ende der dreijährigen Ausbildung.
Durchgeführt wurde die Studie mit Unterstützung des Bundesministeriums für Bildung und Forschung in Zusammenarbeit von Dr. Ulrich Funke vom Bereich Gesundsheitswesen der AUDI AG und Prof. Dr. Thomas Diepgen und Prof. Dr. Manigé Fartasch von der Dermatologischen Universitätsklinik Erlangen, die statistische Auswertung erfolgte in der Abteilung Klinische Sozialmedizin der Universität Heidelberg.
Zur Beurteilung der Risikofaktoren mit Hilfe eines logistischen Regressionsmodells lagen nach Abschluss der Studie die Daten von 1910 Auszubildenden vollständig vor, in 270 Fällen war dabei ein berufsbedingtes Handekzem beobachtet worden. In Zusammenarbeit mit den Betriebsärzten der AUDI AG und den klinischen Verantwortlichen aus Erlangen/Heidelberg wurden folgende Kovariablen ins Modell aufgenommen:

Exogene Risikofaktoren: Schmutzarbeit (in Stunden), Feuchtarbeit (in Stunden) und die Interaktion zwischen beiden

Endogene Risikofaktoren: Atopie-Score ([18]), Vorliegen eines anamnestischen Beugenekzems, Vorliegen eines anamnestischen Handekzems, Vorliegen einer Dyshidrose

Confounder: Geschlecht, Berufsgruppe (Metall, Büroberufe, Sonstige)

Ausgehend von diesen Kovariablen ergeben sich 602 verschiedene Kovariablenmuster, auf die sich die 1910 Auszubildenden wie folgt verteilen:

m_i	Absolute Häufigkeit	Relative Häufigkeit
1	381	63.3
2	78	13.0
3-5	67	11.1
6-10	37	6.1
>10	39	6.5

Das heisst, mehr als 60% der Auszubildenden haben bezüglich der ins Modell aufgenommenen Kovariablen ein individuelles Risikoprofil, was im Hinblick auf die Überprüfung der Modellgüte mit den vorgestellten Anpassungstests zu berücksichtigen ist. In Tabelle 3 sind die Ergebnisse der Modellüberprüfung dargestellt.

Tabelle 3: Beurteilung der Anpassungsgüte des logistischen Regressionsmodells mit Hilfe der vorgestellten Anpassungstests in der AUDI-Studie

Anpassungstest	p-Wert	Anpassungstest	p-Wert
X^2	0.201	D_C	1.000
D	0.157	X^2_F	0.104
X^2_O	0.289	X^2_{FEd}	0.109
X^2_{OEd}	0.282	IM	0.008
X^2_{OSkal}	0.196	IM_{DIAG}	0.002
X^2_{McC}	0.231	R_P	0.184
X^2_{McCEd}	0.227	R_D	0.139
D_F	0.277	R_L	0.144
		R_A	0.000
$\hat{C}$	0.438	R_B	0.227
		$\max_i \lvert r_{Pi}\rvert$	0.002
		R_C	0.333

Die Mehrzahl der Tests weist nicht auf Mängel des Modells hin. Die signifikanten p-Werte von R_A und $\max_i |r_{Pi}|$ sind angesichts der Ergebnisse der durchgeführten Simulationsuntersuchung, wo diese beiden Anpassungstests viel zu liberal abgeschnitten haben, kein Grund zur Beunruhigung.
Auffällig sind dagegen die signifikanten p-Werte der beiden IM-Tests. Es läge jetzt nahe, angesichts der Ergebnisse der Simulationsuntersuchungen,

ein Modell mit einer anderen Linkfunktion (z.B. mit der komplementären LogLog-Linkfunktion, vgl. 3.4.1, S. 49) zu schätzen, hatten sich die IM-Tests doch gegen derartige Fehlspezifikationen den anderen Tests gegenüber als überlegen gezeigt. Ein solches Vorgehen ist jedoch nicht angebracht. Wie bereits dargestellt wurde (vgl. 1.4, S. 9), können wir uns von globalen (wie auch von spezifischen) Anpassungstests keine wirkliche Hilfe bei der Neuformulierung eines schlecht angepassten Modells erhoffen.
Ein Vergleich der AIC-Werte der beiden Modelle (einmal mit der kanonischen Logit-Linkfunktion und einmal mit der komplementären LogLog-Linkfunktion geschätzt), bestätigt diese Beobachtung: im Modell mit der Logit-Linkfunktion (AIC: 1429.1), also quasi im Standardmodell, erhalten wir eine bessere Anpassung als im Modell mit der komplementären LogLog-Linkfunktion (AIC: 1431.69).
Letztendlich zufriedenstellend ist diese Erkenntnis jedoch nicht: die schlechte Modellanpassung, wie sie zumindest von den IM-Tests angezeigt wird, bleibt. Da die IM-Tests auch als Tests gegen Parameterhomogenität (vgl. 2.3.3, S. 32) über die Beobachtungen interpretiert werden können, wurde die Modellschätzung, einer Empfehlung von Thomas[2] zufolge unter der Annahme von heteroskedastischen Fehlern (was in unserem Falle gleichbedeutend ist mit Parameterheterogenität) wiederholt.
Dazu wurde die robuste Methode von White [53] herangezogen (vgl. 2.3.3, S. 32). Diese Schätzmethode wird an verschiedenen Stellen (vgl. [17], [30]) in der ökonometrischen Literatur empfohlen, wenn die Informationsmatrixgleichung (deren Gültigkeit ja gerade von den IM-Tests geprüft wird) verletzt ist.
Es ergeben sich bei der Schätzung definitionsbedingt identische Parameterschätzer, aber unterschiedlich geschätzte Standardfehler und folglich auch zu unterschiedliche p-Werte für die Tests, die den Einfluss der einzelnen Risikofaktoren messen. Ein Vergleich dieser Standardfehler (Tabelle 4) zeigt, dass v.a. die exogenen Risikofaktoren Feuchtarbeit und Schmutzarbeit und die Interaktion zwischen beiden durch die Verwendung der QML-Methode „leiden“, d.h. dass die Schätzung dieser Faktoren mit Hilfe der QML-Methode mit größerer Unsicherheit behaftet ist und dass deren Einfluss auf die Entstehung von Handekzemen relativiert werden muss.

[2] „... the rejection of the model by the IM test and non-rejection by the others may suggest that the model should be estimated under the alternative of heteroscedastic errors...“, ([51])

Tabelle 4: Vergleich der QML- und der ML-Schätzung anhand der Standardfehler der Parameter in der AUDI-Studie

Risikofaktor	Standardfehler ML-Schätzung	Standardfehler QML-Schätzung	Quotient QML/ML
Konstante	0.1719	0.1964	1.143
Atopie-Score	0.0477	0.0462	0.969
Schmutzarbeit	0.0331	0.0381	1.151
Feuchtarbeit	0.0990	0.1184	1.196
Interaktion			
Feucht-/Schmutzarbeit	0.0145	0.0169	1.166
Anamn. Beugenekzem	0.3083	0.3244	1.052
Anamn. Handekzem	0.3172	0.3321	1.047
Dyshidrose	0.2283	0.2261	0.990
Geschlecht	0.1838	0.1834	0.998
Berufsgruppe (Metall-Sonst.)	0.1646	0.1724	1.047
Berufsgruppe (Metall-Büro)	0.3976	0.3755	0.944

4.2 Berufsbedingte Handekzeme im Friseurgewerbe

Auch in dieser prospektiven Kohortenstudie wurde versucht, exo- und endogene Risikofaktoren für die Entstehung von berufsbedingten Handekzemen zu finden und deren Stärke zu quantifizieren, in diesem Fall jedoch in einem anderen Berufsfeld, dem Friseurgewerbe. Die Studie wurde in den Jahren 1991-1994 mit Unterstützung des Hauptverbandes der gewerblichen Berufsgenossenschaften unter Federführung von Prof. Dr. Thomas Diepgen von der Dermatologischen Universitätsklinik Erlangen in Erlangen und Dortmund durchgeführt. Auch hier wurden wieder Auszubildende untersucht, da bei diesen ein relativ standardisiertes Belastungsprofil gegeben ist. Da die arbeitsplatzbezogenen Belastungen im Friseurgewerbe ungleich höher sind als in den meisten anderen Berufen, war bereits nach dem ersten Ausbildungsjahr in 340 von insgesamt 574 Fällen ein berufsbedingtes Handekzem beobachtet worden.
Zur Bewertung der Risikofaktoren wurde auch hier wieder ein logistisches Regressionsmodell berechnet, nach sorgfältigen Voruntersuchungen wurden folgende Kovariablen ins Modell aufgenommen:

Exogene Risikofaktoren: Feuchtarbeit ($</\geq$ 4 Stunden), Hantieren mit

saurer Dauerwelle ($</\geq$ 1 Stunde)

Endogene Risikofaktoren: Atopie-Score ([18]), Vorliegen einer Dyshidrose

Confounder: Zentrumseffekt, Veränderung des Hautschutzverhaltens (weil beobachtet worden war, dass solche Veränderungen erst unternommen werden, wenn bereits Hautprobleme an den Händen vorlagen)

Ausgehend von diesen Kovariablen ergeben sich 334 verschiedene Kovariablenmuster, auf die sich die 574 Auszubildenden wie folgt verteilen:

m_i	Absolute Häufigkeit	Relative Häufigkeit
1	205	61.4
2	68	20.4
3	35	10.5
>3	26	7.8

Auch hier finden wir nur relativ schwach besetzte Kovariablenmuster, was Auswirkungen auf die Überprüfung des Modells mit den vorgestellten Anpassungstests hat. Die Ergebnisse der verschiedenen vorgestellten Anpassungstests findet man in Tabelle 5 (Spalte „p-Wert“).
Die Anpassung des Modells scheint nicht optimal. Viele der Tests, die sich in der Simulationsuntersuchung als zuverlässig herausgestellt haben, zeigen eine signifikante Schwäche des Modells.
Auffällig ist, dass dabei sogar der Pearson-Test X^2 auf eine schlechte Anpassung des Modells hinweist, wo doch dieser Test in den Simulationsuntersuchungen ein konservatives Verhalten gezeigt hatte. Dies legt den Verdacht nahe, dass Ausreißer vorliegen, da die Abweichung von prognostiziertem und beobachtetem Wert im Zähler von X^2 quadriert eingeht. Diese Beobachtung wird beim Blick auf R_C bestätigt, in dessen Berechnung die rohen Pearson-Residuen eingehen. Demgegenüber sind die Tests, die nicht explizit auf der Summation von Residuen beruhen, wie z.B. der Hosmer-Lemeshow-Test $\hat{C}$ oder die beiden IM-Tests, unauffällig.
Eine Residuenanalyse (vgl. Abb. 7) zeigt, dass zwei Beobachtungen vorliegen, bei denen nahezu alle Risikofaktoren gemessen worden waren, so dass sich bei diesen eine geschätzte Wahrscheinlichkeit für das Eintreten des Zielereignis ($\hat{\pi}_i$) von 0.96 ergeben hatte. Tatsächlich war aber in beiden Fällen *kein* Handekzem beobachtet worden. Es ist zu überprüfen, ob hier Datenfehler vorliegen, so dass diese beiden Beobachtungen korrigiert werden könnten.

Tabelle 5: Beurteilung der Anpassungsgüte des logistischen Regressionsmodells mit Hilfe der vorgestellten Anpassungstests für die Friseur-Studie, wobei die p-Werte für den Original-Datensatz, die p*-Werte für den um zwei Ausreißer bereinigten Datensatz stehen

Anpassungstest	p-Wert	p*-Wert	Anpassungstest	p-Wert	p*-Wert
X^2	0.053	0.391	D_C	0.999	1.000
D	0.012	0.033	X^2_F	0.408	0.427
X^2_O	0.044	0.511	X^2_{FEd}	0.396	0.413
X^2_{OEd}	0.053	0.474	IM	0.308	0.142
X^2_{OSkal}	0.030	0.407	IM_{DIAG}	0.365	0.873
X^2_{McC}	0.031	0.458	R_P	0.054	0.387
X^2_{McCEd}	0.042	0.420	R_D	0.011	0.298
D_F	0.063	0.112	R_L	0.012	0.033
			R_A	0.000	0.000
$\hat{C}$	0.451	0.299	R_B	0.017	0.425
			$\max_i \lvert r_{Pi} \rvert$	0.000	0.188
			R_C	0.062	0.734

Nach dem Löschen dieser beiden Ausreißer ergeben sich keine Auffälligkeiten mehr (vgl. Tabelle 5, Spalte „p*-Wert“). All diejenigen Tests, die sich in den Simulationsuntersuchungen als zuverlässige Tests erwiesen hatten, zeigen nun eine zufrieden stellende Anpassung des Modelles an. Am augenscheinlichsten wird das bei R_C, wo der p-Wert von 0.062 auf 0.734 anwächst. Eine neuerliche Skurilität (vgl. 2.2.1, S. 25) zeigt sich beim Hosmer-Lemeshow-Test $\hat{C}$. Die Elimination der beiden offensichtlichen Ausreißer führt laut Hosmer-Lemeshow-Test zu einer *Verschlechterung* der Modellanpassung, der p-Wert sinkt von 0.451 auf 0.299.

4.3 Multizentrische Beobachtungsstudie zu Determinanten intratubarer Sterilität

Zugrunde liegt ein Kollektiv von 162 Frauen mit unerfülltem Kinderwunsch, die in einer internationalen Multicenter-Studie unter der Federführung von Herrn Dr. Rimbach von der Heidelberger Universitätsfrauenklinik untersucht worden waren. Das Motiv für die Durchführung dieser Studie war die Überprüfung einer neuen Methode (Falloposkopie) zur Diagnose von empfängnis-

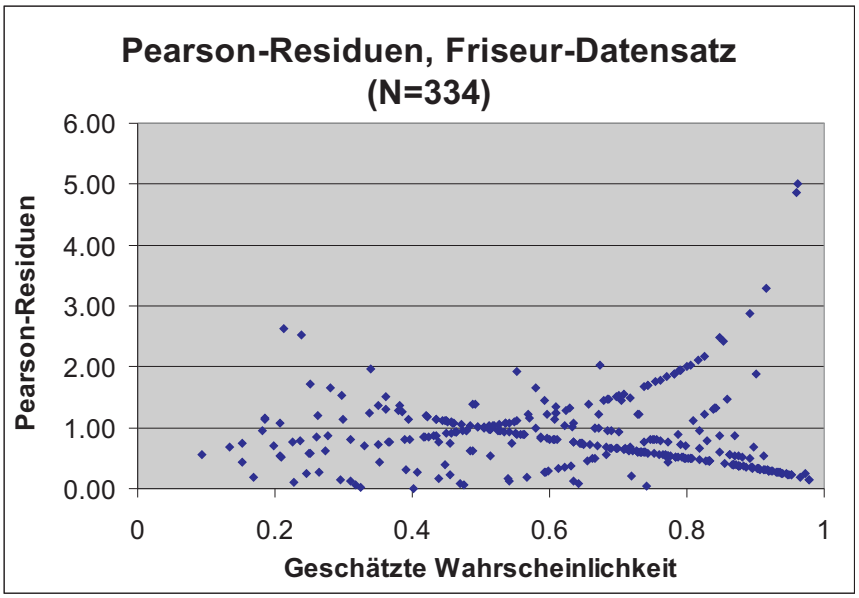


Abbildung 7: Pearson-Residuen aus dem Friseur-Datensatz

verhindernden Eileiterdefekten.
Die interessierende Zielgröße war dabei das Eintreten einer Schwangerschaft im Verlauf der Beobachtungszeit von drei Jahren. Neben der zentralen (binären) Kovariable „Vorliegen eines falloposkopisch diagnostizierten Eileiterdefekts“ wurde das Alter und die Dauer der Unfruchtbarkeit (jeweils in Jahren) als Confounder in ein logistisches Regressionsmodell aufgenommen. Das Zielereignis wurde insgesamt 47mal beobachtet, die 162 Frauen verteilen sich bezüglich der drei Kovariablen auf 108 Kovariablenmuster, und zwar wie folgt:

m_i	Absolute Häufigkeit	Relative Häufigkeit
1	76	70.4
2	18	16.7
3-5	14	12.9

In Tabelle 6 findet man die Ergebnisse der vorgestellten Anpassungstests. Diese weisen übereinstimmend auf eine gute Anpassung des Modells hin. Die einzige Ausnahme bildet R_A, wobei dieser Test aber in der Simulationsuntersuchung annähernd alle Modelle unter der Nullhypothese verworfen hat, so dass dieses Ergebnis keinen Anlass zur Sorge gibt.

Tabelle 6: Beurteilung der Anpassungsgüte mit Hilfe der vorgestellten Anpassungstests für die Falloposkopie-Studie

Anpassungstest	p-Wert	Anpassungstest	p-Wert
X^2	0.658	D_C	0.999
D	0.253	X_F^2	0.836
X_O^2	0.846	X_{FEd}^2	0.835
X_{OEd}^2	0.868	IM	0.866
X_{OSkal}^2	0.797	IM_{DIAG}	0.872
X_{McC}^2	0.798	R_P	0.664
X_{McCEd}^2	0.788	R_D	0.248
D_F	0.391	R_L	0.264
		R_A	0.006
$\hat{C}$	0.362	R_B	0.272
		$\max_i \lvert r_{Pi} \rvert$	0.568
		R_C	0.760

5 Diskussion

In der Literatur findet sich bis zum heutigen Tag erst eine einzige groß angelegte Untersuchung zum Verhalten von globalen Anpassungstests im logistischen Regressionsmodell bei fehlenden Messwiederholungen (Hosmer et al. [28]). Die hier vorliegende Arbeit bestätigt die Ergebnisse von Hosmer et al. dort, wo vergleichbare Dinge untersucht wurden, ergänzt sie aber in vielen Punkten und liefert einige neue Einblicke in das Problem der Überprüfung der Modellgüte im logistischen Modell.
Es zeigt sich, dass einige neue Tests, die zum Zeitpunkt der Untersuchung von Hosmer et al. noch nicht entwickelt bzw. den Autoren nicht bekannt waren, den von ihnen empfohlenen Tests überlegen sind. Hosmer et al. hatten zur allgemeinen Anwendung eine Variante von X^2_{McC} und R_C vorgeschlagen. Die hier durchgeführten Simulationsuntersuchungen bestätigen das gute Verhalten von R_C und X^2_{McC}. Es existieren aber zwei weitere Tests, die diesen beiden bzgl. der Power stellenweise überlegen sind, aber das Niveau unter der Nullhypothese ebenfalls einhalten.
Der erste ist der der Farrington-Test X^2_F, eine Variante des altbekannten Pearson-Tests X^2, der um eine additive Konstante erweitert wird. Dieser Test hat leider eine definitionsbedingte Schwäche in Situationen ohne Messwiederholungen ($m_i \equiv 1$), wo dieser nie das Modell verwirft. Der zweite, im medizinisch-statistischen Bereich bisher noch nicht angewandte Test, ist der Informationsmatrix-Test IM_{DIAG}. Dieser stammt aus dem Bereich der Ökonometrie und beruht auf dem Konzept des Vergleichs zweier Schätzer der Informationsmatrix, bei genauerer Betrachtung der Teststatistik werden aber auch hier letztendlich beobachtete und prognostizierte Werte der Zielgröße miteinander verglichen. Dieser Test behält seine guten Eigenschaften im Vergleich zu den anderen Tests auch in Situationen mit sehr wenigen Messwiederholungen.
Sowohl X^2_F, als auch IM_{DIAG} haben den Vorteil, dass sie mit vernünftigem Programmieraufwand zu berechnen sind , insofern ist auch die „Doktrin“ von Pagan (vgl. 3.2, S.39) erfüllt, die besagt, dass ein Anpassungstest nicht allzuviel Mühe bei der Berechnung machen sollte, weil ansonsten die Gefahr besteht, dass er überhaupt nicht angewendet wird.
Die beiden klassischen globalen Anpassungstests, der Pearson-Test X^2 und die Devianz D sind in logistischen Regressionsmodellen, außer in Fällen wo ausschließlich wenige binäre Kovariablen vorliegen, nicht zu gebrauchen. Auch hier bestätigt die vorliegende Arbeit bereits vorliegendes Wissen (vgl.

1.6, S.13), wobei die Devianz D stärker betroffen ist als der Pearson-Test. Da diese Tatsache bereits seit langem bekannt war, hatte sich ein neuer Quasi-Standard für die Modellüberprüfung im logistischen Regressionsmodell herausgebildet, der Hosmer-Lemeshow-Test $\hat{C}$. In der voliegenden Untersuchung zeigt sich, dass er diesen Status nicht zu unrecht erworben hat. Er hält unter der Nullhypothese das Niveau gut ein und schneidet auch bei den Untersuchungen zur Power zufriedenstellend ab. Angesichts einiger unerfreulichen Eigenschaften dieses Tests (vgl. 2.2.1, S.25) zu denen noch die neue Beobachtung tritt, dass bei einem real vorliegenden Datensatz das Löschen von zwei offensichtlichen Ausreißern zu einer postulierten Verschlechterung der Modellanpassung führen kann (vgl. 4.2, S.58), kann es aber nur von Vorteil sein, wenn es auch zum Hosmer-Lemeshow-Test noch Alternativen gibt.
Eine Reihe von Vorschlägen für globale Anpassungstests im logistischen Modell erweisen sich als völlig unbrauchbar. Dieses sind zum einen Korrekturen, die die Devianz D betreffen, wo die korrigierten Tests D_F und D_C die schlechten Eigenschaften der Devianz nicht korrigieren können. Zum anderen handelt es sich dabei um Tests, die auf der Weiterverarbeitung von Residuen (Summation nach Standardisierung oder Analyse des betragsmäßig größten Residuums) aufbauen. Hier bestätigt sich die Beobachtung von Jennings ([29]), dass die Techniken und Erkenntnisse zur Residuenanalyse im linearen Modell nicht ohne weiteres auf das logistische Modell übertragbar sind.

Bezüglich des notwendigen Stichprobenumfangs zeigt sich, dass die untersuchten Tests erst bei größeren Stichproben eine einigermaßen gute Power haben, Fehlspezifikation im Modell zu entdecken. Dies gilt nicht für die Situationen unter der Nullhypothese, wo das Verhalten der Tests weitestgehend unabhängig von der Stichprobengröße ist.

Schließlich muss natürlich auch auf die Mängel der vorliegenden Untersuchung hingewiesen werden. Trotz der Bemühungen, möglichst viele Parameter der untersuchten Modelle zu variieren und der dadurch „verbratenen“ Rechenzeit, stellen die hier abgeprüften Situationen nur einen winzigen Ausschnitt aller möglichen logistischen Modelle bezüglich Anzahl und Verteilung der Kovariablen und der Kovariablenmuster oder auch bezüglich der untersuchten Fehlspezifikationen dar. Insofern können die hier gewonnenen Erkenntnisse in vielerlei Hinsicht ergänzt werden.
Desweiteren liegt noch eine Reihe von Anpassungstests vor, die nicht in diese Untersuchung aufgenommen wurden, sei es, da sie sich hartnäckig einer

Programmierung entzogen oder dass keine Datensätze vorlagen, um die Programmierung zu überprüfen. Interessant wären in diesem Zusammenhang z.B. die vorliegenden Modifikationen des Hosmer-Lemeshow-Tests, wo zum einen Hosmer und Lemeshow selber eine Variante (vgl. 2.2.1, S.25) vorgeschlagen haben, die in deren Simulationsuntersuchungen gut abgeschnitten hatte, letztendlich aber verworfen wurde, weil sie in Situationen mit vielen Messwiederholungen unbefriedigende Ergebnisse zeigte. Zum anderen existiert ein Vorschlag von Andrews([4],[5], vgl. 2.2.1, S.26), der in allgemeinerem Kontext eine exakte Verteilung für eine Variante des Hosmer-Lemeshow-Tests herleitet.
Ein weiterer Kandidat wäre der Test von Cressie und Read (vgl. 2.1.1, S.17) mit $\lambda = 2/3$ für das logistische Regressionsmodell. Dieser Test wird von den Autoren für andere Modelle als bevorzugter Anpassungstest empfohlen([48]), Osius/Rojek([42]) zeigen, dass die Teststatistik unter der Nullhypothese ebenfalls normalverteilt ist.

Für die konkrete Anwendung zeigt sich, dass globale Anpassungstests sehr wohl einen wichtigen Beitrag zur Modellierung von Daten mit Hilfe logistischer Regressionsmodelle liefern können, so sie tatsächlich angewandt werden. Sie können aber ganz sicher nicht das einzige Mittel bei einer notwendigen sorgfältigen Überprüfung der Modellanpassung sein. Diese ist vor allem dann, wenn die Tests eine ungenügende Anpassung anzeigen, um eine Residuenanalyse zu erweitern, da die Tests keinerlei Hilfe bei der Neuformulierung des Modells geben können. Desweiteren ist die Power der Tests, vor allem in Situationen mit fehlenden Messwiederholungen und bei kleinen Fallzahlen insgesamt zu niedrig.
Es sei schließlich auch noch auf das Dilemma aller Anpassungstests hingewiesen, die ja letztendlich nicht die Güte des Modells überprüfen und statistisch absichern können, sondern immer nur die Schwäche des Modells: ein nichtsignifikanter Anpassungstest sagt uns nicht, dass ein gutes Modell vorliegt, er sagt uns nur, dass kein schlechtes Modell vorliegt.

Zusammenfassung

Das logistische Regressionsmodell hat sich seit seiner Einführung in den siebziger Jahren zu einer Standardmethode in der Biometrie und Epidemiologie entwickelt, wenn es um die Auswertung von binären Zielgrößen geht. Die Gründe dafür sind vielfältig. Exemplarisch seien genannt die leichte Interpretierbarkeit der geschätzten Parameter als Odds-Ratios, die Möglichkeit zu Prognosen über das Eintreten des Zielereignisses, die Verfügbarkeit von geeigneter Software und, für die Epidemiologie besonders wichtig, die Möglichkeit, das Modell zur Analyse sowohl von prospektiven als auch retrospektiven Beobachtungsstudien einzusetzen.

Methoden zur Überprüfung der Anpassungsgüte in logistischen Regressionsmodell haben diese stürmische Entwicklung nicht mitgemacht, was zum einen sicherlich an der höheren mathematischen Komplexität des logistischen Modells liegt, zum anderen an der relativen Jugend im Vergleich zu z.B. dem linearen Regressionsmodell.

Als globale Anpassungstests für das logistische Regressionsmodell werden vor allem die Devianz D oder die Pearson-Statistik X^2 empfohlen, die auf allgemein bekannten Testprinzipien basieren und auch in anderen Bereichen der Statistik Anwendung finden. Es ist jedoch bekannt, dass diese beiden Tests in Situationen mit fehlenden Messwiederholungen, also z.B. bei stetigen Kovariablen oder einer großen Anzahl von Kovariablen, eher die Regel als die Ausnahme in realen Datensätzen, nicht zu verlässlichen Ergebnissen führen, weil die Prüfgrößen auch asymptotisch nicht mehr χ^2-verteilt sind.

Die Lösungen dieses Problems sind im Prinzip seit langem bekannt, werden aber im biometrisch-epidemiologischen Bereich, mit der Ausnahme des Hosmer-Lemeshow-Tests, wenig eingesetzt. Die vorgeschlagenen Lösungen lassen sich in drei Gruppen einteilen: Zum ersten können D und X^2 als Prüfgrößen beibehalten werden, ihre statistische Signifikanz wird jedoch mit Hilfe anderer Prüfverteilungen beurteilt. Zum zweiten können die Beobachtungen zu Gruppen zusammengefasst werden, so dass ausreichend Messwiederholungen in diesen neuen Gruppen vorliegen. Zum dritten kann zu anderen Teststatistiken übergangen werden, die die altbekannten Tests modifizieren oder auf gänzlich neuen Testprinzipien beruhen.

Die vorgeschlagenen Tests werden dargestellt und im Rahmen einer Simulationsuntersuchung sowohl unter der Nullhypothese eines korrekt spezifierten Modells als auch unter der Alternative einer Fehlspezifikation des Modells miteinander verglichen. Es zeigt sich, dass die

Standardtests bereits in Modellen in denen die Anzahl der Messwiederholungen kleiner ist als fünf, nicht mehr zu verlässlichen Ergebnissen führen, die Devianz ist davon noch stärker betroffen als der Pearson-Test. Der Hosmer-Lemeshow-Test, der bekannteste unter allen Alternativen zu D und X^2, hält dagegen in allen Simulationen das vorgegebene Niveau ein und hat eine zufrieden stellende Power. Daneben treten drei weitere Tests, die noch zu etwas besseren Ergebnissen führen. Der erste Test ist der Farrington-Test X_F^2, der die herkömmliche Pearson-Statistik um eine additive Konstante erweitert. Der zweite Test ist der Informationsmatrix-Test IM_{DIAG}, der zwei unter korrekter Modellspezifikation äquivalente Schätzer der Informationsmatrix vergleicht. Der dritte Test schließlich ist der RSS-Test, R_C, der auf der Summation von unstandardisierten Residuen beruht. Der Farrington-Test ist den beiden anderen Tests leicht überlegen, hat aber den Nachteil, dass er in Situationen ohne Messwiederholungen definitionsbedingt nie eine schlechte Modellanpassung anzeigt. Alle drei Tests sind mit vernünftigem Aufwand zu berechnen.

Anhand dreier Anwendungsbeispiele aus der Praxis wird gezeigt, dass globale Anpassungstests sehr wohl einen wichtigen Beitrag zur Modellierung von Daten mit Hilfe logistischer Regressionsmodelle liefern können. Sie können aber ganz sicher nicht das einzige Mittel bei einer sorgfältigen Überprüfung der Modellanpassung sein, diese ist in der Regel, vor allem dann wenn die Tests eine ungenügende Anpassung anzeigen, um eine Residuenanalyse zu erweitern. Desweiteren ist die Power der Tests, vor allem in Situationen mit sehr wenigen Messwiederholungen und bei kleinen Fallzahlen insgesamt zu niedrig. Schließlich bleibt das Dilemma aller Anpassungstests bestehen, die ja letztendlich nicht die Güte des Modells überprüfen und statistisch absichern können, sondern immer nur die Schwäche des Modells: ein nicht-signifikanter Anpassungstest sagt uns nicht, dass ein gutes Modell vorliegt, er sagt uns nur, dass kein schlechtes Modell vorliegt.

Literatur

[1] A. Agresti. *Categorical data analysis.* John Wiley & Sons, 1990.

[2] Z. Al-Sarraf, D.H. Young. The Extreme Residuals in Logistic Regression. *J Stat Comput Sim,* 25:115-125, 1986.

[3] J.M. Alston, J.A. Chalfant. Unstable Models from Incorrect Forms. *Am J Agr Econ,* 73:1171-1181, 1991.

[4] D.W.K. Andrews. Chi-Square Diagnostic Tests for Econometric Models: Introduction and Application. *J Econometrics,* 37:135-156.

[5] D.W.K. Andrews. Chi-Square Diagnostic Tests for Econometric Models: Theory. *Econometrika,* 56:1419-1453.

[6] F.J. Anscombe. Contribution to the Discussion of a Paper by H. Hotelling. *J Roy Stat Soc B,* 15:229-230, 1953.

[7] G.A. Barnard. Introduction to Pearson (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Varaibles is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. *S. Kotz, N.L. Johnson.[Hrsg.] Breakthroughs in Statistics II. Springer-Verlag,* 1-10, 1992.

[8] A. Chesher. Testing for neglected heterogeneity. *Econometrica*, 52:865-872, 1984.

[9] D. Collett. *Modelling binary data.* Chapman & Hall, 1991.

[10] J.B. Copas. Unweighted Sum of Squares Test for Proportions. *Appl Stat,* 38:71-80, 1989.

[11] G.M. Cordeiro. Improved Likelihood Ratio Statistics for Generalized Linear Models. *J Roy Stat Soc B,* 45:404-413, 1983.

[12] G.M. Cordeiro. Performance of a Bartlett-type modification for the deviance *J Stat Comput Sim,* 51:385-403, 1995.

[13] D.R. Cox, D.V. Hinkley. *Theoretical Statistics.* Chapman & Hall, 1974.

[14] N.A.C. Cressie, T.R.C. Read. Multinomial Goodness-of-Fit tests. *J Roy Stat Soc B,* 46:440-464, 1984.

[15] J.R. Dale. Asymptotic Normality of Goodness-of-fit Statistics for Sparse Product Multinomials. *J Roy Stat Soc B,* 48:48-59, 1986.

[16] R. Davidson, J.G. MacKinnon. Convenient specification tests for logit and probit models. *J Econometrics,* 25:241-262, 1984.

[17] R. Davidson, J.G. MacKinnon. *Estimation and inference in econometrics.* Oxford University Press, 1993.

[18] T.L. Diepgen, W. Sauerbrei, M. Fartasch. Development and validation of diagnostic scores for atopic dermatitis incorporating criteria of data quality and practical usefulness. *J Clin Epidemiol,* 49:1031-1038, 1996.

[19] L. Fahrmeir, G. Tutz. *Multivariate statistical modelling based on generalized linear models.* Springer, 1996.

[20] C.P. Farrington. Pearson statistics, goodness of fit, and overdispersion in generalised linear models. *G.U.H Seeber et al.[Hrsg.] Statistical Modelling. Springer-Verlag,* 109-116, 1995.

[21] C.P. Farrington. On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data. *J Roy Stat Soc B,* 58:349-360, 1996.

[22] J.J. Forster, J.W. McDonald, P.W.F. Smith. Monte Carlo Exact Conditional Tests for Log-linear and Logistic Models. *J Roy Stat Soc B,* 58:445-453, 1996.

[23] L.G. Godfrey. *Misspecification tests in econometrics.* Cambridge University Press, 1988.

[24] I. Hacking. Trial by Number. *Science,* 84:69-70, 1984.

[25] D.W. Hosmer, S. Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Commun Stat - Theor M* 9:1043-1069, 1980.

[26] D.W. Hosmer, S. Lemeshow. *Applied logistic regression.* John Wiley & Sons, 1989.

[27] D.W. Hosmer, S.Taber, S. Lemeshow. The Importance of Assessing the Fit of Logistic Regression Models: A Case Study. *Am J Public Health* 81:1630-1635, 1991.

[28] D.W. Hosmer et al. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med,* 16:965-980, 1997.

[29] D.E. Jennings. Outliers and Residual Distributions in Logistic Regression. *J Am Stat Assoc,* 81:987-990, 1986.

[30] J. Johnston, J. DiNardo. *Econometric Methods.* The McGraw-Hill Companies, Inc., 1997.

[31] L.R. Korn, D.W. Hosmer, S. Lemeshow. The Performance of Goodness of Fit Tests for Logistic Regression with Discrete Covariates. *Biometrical J,* 28:697-708, 1986.

[32] S. Lemeshow, D.W. Hosmer. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 115:92-106, 1982.

[33] C.J. Lloyd. *Statistical Analysis of Categorical Data.* John Wiley & Sons, 1999.

[34] D. Lugtenburg. An Approximate Bartlett Adjustment for Testing the Goodness of Fit of Multinomial Regression Models. *J Roy Stat Soc B,* 53:393-398, 1991.

[35] P. McCullagh. Sparse data and conditional tests. *Bulletin of the International Statistics Institute, Proceedings of the 45th Session of ISI (Amsterdam), Invited Paper,* 28:1-10, 1985.

[36] P. McCullagh. On the Asymptotic Distribution of Pearsons Statistic in Linear Exponential-Family Models. *Int Stat Rev,* 53:61-67, 1985.

[37] P. McCullagh. The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data. *J Am Stat Assoc,* 81:104-107, 1986.

[38] P. McCullagh, J.A. Nelder. *Generalized Linear Models.* Chapman & Hall, 1989.

[39] C.E. Minder. Globale Anpassungstests für eine weite Klasse von statistischen Modellen. *G.U.H Seeber, C.E.Minder.[Hrsg.] Multivariate Modelle. Springer-Verlag,* 156-165, 1991.

[40] C. Orme. The calculation of the information matrix test for binary data models. *The Manchester School,* 54:370-376, 1988.

[41] C. Orme. The small-sample performance of the information-matrix test. *J Econometrics,* 46:309-331, 1990.

[42] G. Osius, D. Rojek. Normal Goodness-of-Fit Tests for Multinomial Models With Large Degrees of Freedom. *J Am Stat Assoc,* 87:1145-1152, 1992.

[43] G. Osius. Evaluating the Significance Level of Goodness-of-Fit Statistics for Large Discrete Data. *P. Dirschedl, R. Ostermann.[Hrsg.] Computational Statistics. Physica-Verlag,* 395-417, 1994.

[44] K. Pearson. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Varaibles is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. *Philosophical Transactions,* 50:157-175, 1900.

[45] J.G. Pigeon, J.F. Heyse. Methods for Assessing the Adequacy of Probability Prediction Models. *G.U.H Seeber et al. [Hrsg.] Statistical Modelling. Proceedings of the 10th International Workshop on Statistical Modelling Innsbruck, Austria, 10-14 July, Springer-Verlag,* 225-232, 1995.

[46] J.G. Pigeon, J.F. Heyse. An Improved Goodness of Fit Statistic for Probability Prediction Models. *Biometrical J,* 41:71-82, 1999.

[47] D. Pregibon. Goodness of link tests for generalized linear models. *Appl Stat* 29:15-24, 1980.

[48] T.R.C. Read, N.A.C Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer, 1989.

[49] T.J. Santner, D.E. Duffy. *The statistical analysis of discrete data.* Springer, 1989.

[50] S.M. Snapinn, R.D. Small. Tests of Significance Using Regression Models for Ordered Data. *Biometrics,* 42:583-592, 1986.

[51] J.M. Thomas. On Testing the Logistic Assumption in Binary Dependent Variable Models. *Emp Econom,* 18:381-392, 1993.

[52] A.A. Tsiatis. A note on a goodness-of-fit test for the logistic regression model. *Biometrika,* 67:250-251, 1980.

[53] H. White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50:1-25, 1982.

[54] H. White. *Estimation, Inference and Specification Analysis.* Cambridge University Press, 1994.

[55] D.A. Williams. Extra-binomial Variation in Logistic Linear Models. *Appl Stat,* 31:144-148, 1982.

[56] D.A. Williams. Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. *Appl Stat,* 36:181-191, 1987.

[57] F.A.G. Windmeijer. The asymptotic distribution of the sum of weighted squared residuals in binary choice models. *Stat Neerl,* 44:69-77, 1990.

Publikationsverzeichnis

Originalarbeiten

- **O. Kuß**, A. Koch. Meta-analysis macros for SAS. *The Statistical Software Newsletter* 22:325-333, 1996.

- S. Ziegler, **O. Kuß**, A. Koch. Meta-Analyse in einem Modell mit festen und mit zufälligen Effekten. *R. Muche, G. Büchele, D. Harder, W. Gaus (Hrsg). Medizinische Informatik, Biometrie und Epidemiologie - GMDS '97. MMV, München,* 361-365, 1997.

- **O. Kuß**, T.L. Diepgen. Proper statistical analysis of transepidermal water loss (TEWL) measurements in bioengineering studies. *Contact Dermatitis* 39:64-67, 1998.

- **O. Kuß**, T.L. Diepgen : Ergebnisse einer Kohortenstudie zu Risikofaktoren von Handekzemen bei Auszubildenden im Friseurhandwerk. *E. Greiser, M. Wischnewsky (Hrsg). Medizinische Informatik, Biometrie und Epidemiologie - GMDS '98. MMV, München,* C48 (Beitrag auf beigelegter CD-ROM), 1998.

- E. Schnetz, **O. Kuß**, J. Schmitt, T.L. Diepgen , M. Kuhn, M. Fartasch. Intra- and inter-individual variations in transepidermal water loss on the face: facial locations for bioengineering studies. *Contact Dermatitis* 40:243-247, 1999.

- **O. Kuß** : Logistische Regression in SAS. *C. Ortseifen (Hrsg.). Proceedings der 3. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE), 25./26. Februar 1999, Ruprecht-Karls-Universität Heidelberg,* 147-154, 1999.

- V. Mahler, T.L. Diepgen , **O. Kuß**, T. Leakakos, G. Schuler, D. Kraft, R. Valenta. Mutual boosting effects of sensitization with timothy grass pollen and latex glove extract on IgE antibody responses in a mouse model. *J Invest Dermatol* 114:1039-1043, 2000.

- E. Schnetz, T.L. Diepgen , P. Elsner P, P.J. Frosch, A.J. Klotz, J. Kresken, **O. Kuß**, H. Merk, H.J. Schwanitz, W. Wigger-Alberti, M. Fartasch. Multicentre study for the development of an in vivo model

to evaluate the influence of topical formulations on irritation. *Contact Dermatitis* 42:336-343, 2000.

- **O. Kuß**. Ein SAS-Makro zur Schätzung des Stereotype Regressionsmodells. *Bödeker, H (Ed.) Proceedings der 4. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE), Gießen, 9./10. 3. 2000,* 107-113, 2000.

- B. Feuerstein, T.G. Berger, C. Maczek, C. Röder, D. Schreiner, U. Hirsch, I. Haendle, W. Leisgang, A. Glaser, **O. Kuß**, T.L. Diepgen, Schuler G, B. Schuler-Thurner. A method for the production of cryopreserved aliquots of antigen-preloaded, mature Dendritic Cells ready for clinical use. *J Immunol Methods* 245:15-29, 2000.

- V. Mahler, S. Vrtala, **O. Kuß**, T.L. Diepgen , O. Cromwell, H. Fiebig, G. Schuler, D. Kraft, R. Valenta R. Immunological effects of immunization with genetically engineered hypoallergenic derivatives of the major birch pollen allergen Bet v 1 (rBet v 1 fragments and rBet v 1 trimer) in comparison with currently applied vaccines (allergoids) and unmodified rBet v 1 in a mouse model. *Eur Journal Immunol,* (zur Publikation eingereicht)

- J.W. Fluhr, **O. Kuß**, T.L. Diepgen, S. Lazzerini, A. Pelosi. Testing for irritation with a multifactorial approach: Comparison of eight noninvasive bioengineering parameters on five different irritation models. *Brit J Dermatol,* (zur Publikation angenommen)

- R.L. Bergmann, T.L. Diepgen , **O. Kuß**, K.E. Bergmann, J. Kujat, U. Wahn and the MAS-study group. Breastfeeding does not protect from atopic eczema. *Clin Exp Allergy,* (zur Publikation eingereicht)

Vorträge

- **O. Kuß**, A. Koch. SAS-Makros für Meta-Analyse. Statistical Computing '95: Arbeitstagung über Methoden und Werkzeuge der Informatik in der Statistik, Schloß Reisensburg (bei Günzburg), 18.-21.6. 1995.

- **O. Kuß**. Ordinal regression models in epidemiological research. The EDEN Congress: 2nd International Meeting on Epidemiology and Prevention of Skin Diseases, Bamberg, 2.-4.5. 1998.

- **O. Kuß**, T.L. Diepgen. How to analyse ordinal data without information loss. 4th Congress of the European Society of Contact Dermatitis (ESCD), Helsinki, Finland, 8.-11.7. 1998.

- **O. Kuß**, T.L. Diepgen. Ergebnisse einer Kohortenstudie zu Risikofaktoren von Handekzemen bei Auszubildenden im Friseurhandwerk. 43. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), Bremen, 14.-16.9. 1998.

- **O. Kuß**, T.L. Diepgen. Ordinale Regressionsmodelle in der Dermato-Epidemiologie. Zweite gemeinsame Tagung der AG „Epidemiologie, Biostatistik und Informatik“ der DDG und der AG „Dermato- und Allergoepidemiologische Methodik“ der Deutschen Arbeitsgemeinschaft für Epidemiologie, Mannheim, 25.-26.9. 1998.

- **O. Kuß**. Goodness-of-Fit Tests im logistischen Regressionsmodell ohne Meßwiederholungen. Herbstworkshop der Arbeitsgruppen „Generalisierte lineare Modelle“ und „Statistische Methoden in der Epidemiologie“ der Deutschen Region der Internationalen Biometrischen Gesellschaft und der Deutschen Arbeitsgemeinschaft für Epidemiologie, Heidelberg, 8.-9.10. 1998.

- **O. Kuß**. Logistische Regression in SAS. 3. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE), Heidelberg, 25.-26.2. 1999.

- **O. Kuß**, J. Hendrickx. Stereotype Regression - ein nahezu unbekanntes multinomiales logistisches Regressionsmodell. „Medical Decision Making - Methodische Aspekte“. Gemeinsamer Workshop der Arbeitsgruppen „Statistische Methoden in der Epidemiologie“, „Statistische Methoden in der Medizin“ der Deutschen Region der Internationalen Biometrischen Gesellschaft und „Statistische Methodik der Klinischen Forschung“, „Methoden der Prognose- und Entscheidungsfindung“ der Deutschen Gesellschaft für Medizinsche Informatik, Biometrie und Epidemiologie, Heidelberg, 18.-20.11. 1999.

- **O. Kuß**, J. Hendrickx. Ein SAS-Makro zur Schätzung des Stereotype Regressionsmodells. 4. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE), Gießen, 9.-10.3. 2000.

Publizierte Abstracts

- **O. Kuß**. Ordinal regression models in epidemiological research. *Allergologie,* 21:294, 1998.

- **O. Kuß**, T.L. Diepgen. How to analyse ordinal data without information loss. *L. Kanerva, A. Lauerma, F. Björksten, T. Estlander, R. Jolanki, M. Hannuksela (Ed.). Fourth Congress of the European Society of Contact Dermatitis. (People and Work, Research Reports 18, Finnish Institute of Occupartional Health),* 58, 1998.

- V. Mahler, T.L. Diepgen , **O. Kuß**, T. Leakakos, W. Truscott, G. Schuler, D. Kraft, R. Valenta. Wechselwirkungen zwischen Typ-I-Sensibilisierung gegen Pollen und Latex - Untersuchungen in einem Mausmodell. *Allergologie,* 23:263, 2000.

- J.W. Fluhr, **O. Kuß**, T. Diepgen, S. Lazzerini, A. Pelosi, E. Berardesca. Testing for irritation with a multifactorial approach: Comparison of eight non-invasive bioengineering parameters on five different irritation models. *Skin Research and Technology,* (zur Publikation angenommen)

Poster

- **O. Kuß**, T.L. Diepgen : How to analyse TEWL-measurements properly - A statistical approach. 3rd International Symposium on Irritant Contact Dermatitis (ISICD), Rome, Italy, 2.-4.10. 1997.

- E. Schnetz, F. Bahmer, T.L. Diepgen , G. Eichler, P. Elsner, P.J. Frosch, W. Gehring, A.J. Klotz, J. Kresken, A. Kurte, **O. Kuß**, M. Lange, H. Merck, H. Täuber, H.J. Schwanitz, S.W. Wassilew, W. Wigger-Alberti, M. Fartasch. Development and evaluation of an in vivo test model for cumulative irritation - first results of a multi center study. 3rd International Symposium on Irritant Contact Dermatitis (ISICD), Rome, Italy, 2.-4.10. 1997.

- E. Schnetz, **O. Kuß**, T.L. Diepgen , M. Fartasch. Evaluation of the influence of magistral formulas on irritation using a cumulative irritation model over three consecutive days. 4th Congress of the European Society of Contact Dermatitis (ESCD), Helsinki, Finland, 8.-11.7. 1998.

- E. Schnetz, F. Bahmer, M. Bock, T.L. Diepgen , G. Eichler, P. Elsner, P.J. Frosch, W. Gehring, F. Jugert, A.J. Klotz, J. Kresken, bf O. Kuss, M. Lange, H. Merck, H.J. Schwanitz, H. Täuber, S.W. Wassilew, W. Wigger-Alberti, M. Fartasch. Evaluation of the effect of creams in a repetitive irritation test - A multi center trial. 4th Congress of the European Society of Contact Dermatitis (ESCD), Helsinki, Finland, 8.-11.7. 1998.

- E. Schnetz E, **O. Kuß**, K.E. Andersen, T.L. Diepgen , M. Fartasch. A practical in vivo model to evaluate the efficacy of barrier creams. 4th Congress of the European Society of Contact Dermatitis (ESCD), Helsinki, Finland, 8.-11.7. 1998.

- R.L. Bergmann, K.E. Bergmann, T.L. Diepgen , **O. Kuß**, J. Kujat, U. Wahn. Does breastfeeding affect the risk of allergy. 9th International Conference of the International Society for Research in Human Milk and Lactation (ISRHML): Short and Long-Term Effects of Breastfeeding on Child Health, Kloster Irsee, 2.-6.6. 1999.

- H. Dickel, A. Schmidt, **O. Kuß**, T.L. Diepgen. Daten aus dem BKH-N zur Anwendung von Hautschutzmaßnahmen bei Versicherten mit berufsbedingten Hauterkrankungen. Kontaktallergie 2000, Pullach, 8.-9.9. 2000.

A Ergebnisse der Simulationsuntersuchung: Nullhypothese

Dargestellt ist das empirische Signifikanzniveau der verglichenen Anpassungstests, d.h. die Anzahl (in %), in der der jeweilige Anpassungstest die Nullhypothese verworfen hat. Dieses sollte in diesem Fall (unter der Nullhypothese einer korrekten Modellanpassung und bei einem Signifikanzniveau der berechneten Anpassungstests von $\alpha = 0.05$) gleich 5 sein. Eine Zahl unter 5 weist auf ein zu konservatives, eine Zahl über 5 auf eine zu liberales Verhalten des Tests hin.
In den Spalten sind die Belegungen der Kovariablenmuster abgetragen. Dabei stehen konstante m_i $(1, 2, 5, 10)$ für Belegungen, in denen alle Kovariablenmuster die selbe Belegung haben. Die Bezeichnung „1-2“ steht für eine Belegung, in dem eine Hälfte der Kovariablenmuster einfach, die andere Hälfte doppelt belegt sind. Die Bezeichnung „1-10“ steht für eine Belegung, in der die Kovariablenmuster einfach, doppelt, fünffach und zehnfach im Verhältnis $64:21:9:6$ belegt sind.

Tabelle 7: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0$, Kein Regressionseffekt

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.00	0.00	1.28	0.21	4.11	5.18
D	99.97	86.48	58.58	46.09	14.59	7.67
X^2_O	5.17	5.02	4.01	4.16	3.57	3.15
X^2_{OEd}	8.86	5.02	3.92	3.62	2.90	2.38
X^2_{OSkal}	0.26	5.18	4.96	4.60	4.79	5.49
X^2_{McC}	5.42	5.33	5.51	5.74	6.80	7.47
X^2_{McCEd}	6.95	5.34	5.49	4.90	4.88	5.06
D_F	0.00	25.02	24.22	8.02	5.66	2.60
$\hat{C}$	4.43	4.87	4.94	2.89	4.96	5.34
D_C	0.00	0.00	6.31	0.06	7.99	5.99
X^2_F	0.00	5.02	5.53	5.93	6.81	7.47
X^2_{FEd}	0.00	5.01	5.51	4.87	4.88	5.05
IM	2.93	3.05	2.58	2.77	2.97	2.80
IM_{DIAG}	4.78	4.97	4.55	4.82	4.77	5.01
R_P	0.00	0.00	1.36	0.33	4.79	7.04
R_D	99.97	87.85	59.93	47.39	16.15	9.94
R_L	99.97	85.12	57.50	44.33	14.70	9.38
R_A	100.00	100.00	94.31	97.69	32.45	18.76
R_B	99.96	86.20	58.37	45.69	14.48	7.69
$\max_i \lvert r_{Pi} \rvert$	0.00	0.00	0.00	0.13	0.67	3.87
R_C	5.25	4.56	4.89	5.09	4.96	5.42
M=500						
X^2	0.00	0.00	1.15	0.10	3.53	3.93
D	100.00	100.00	99.69	99.90	30.64	9.41
X^2_O	5.60	4.50	3.97	4.20	4.23	3.59
X^2_{OEd}	14.8	4.50	3.97	4.00	3.98	3.06
X^2_{OSkal}	0.00	4.50	4.56	4.70	4.98	4.63
X^2_{McC}	5.70	4.70	4.88	5.40	5.51	5.63
X^2_{McCEd}	11.30	4.70	4.88	4.90	5.12	4.62
D_F	100.00	100.00	98.25	99.20	20.25	5.34
$\hat{C}$	5.50	5.00	4.92	3.80	4.81	4.79
D_C	0.00	0.00	27.19	0.00	11.26	5.60
X^2_F	0.00	4.50	4.88	5.3	5.51	5.63
X^2_{FEd}	0.00	4.50	4.88	4.90	5.12	4.63
IM	2.10	2.20	2.06	1.40	2.11	1.94
IM_{DIAG}	4.90	4.80	4.44	3.70	4.59	4.29
R_P	0.00	0.00	1.16	0.10	3.72	4.31
R_D	100.00	100.00	99.70	99.90	31.37	9.96
R_L	100.00	100.00	99.68	99.90	30.41	9.68
R_A	100.00	100.00	100.00	100.00	57.35	17.22
R_B	100.00	100.00	99.68	99.90	30.58	9.41
$\max_i \lvert r_{Pi} \rvert$	0.00	0.00	0.00	0.00	0.00	1.51
R_C	5.60	5.50	5.15	5.00	5.05	4.40

Tabelle 8: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0.405x_{1i}$, $x_1 \sim U(-6, 6)$, schwacher Regressionseffekt (OR=1.5)

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.71	2.42	2.27	6.07	4.10	4.52
D	1.74	1.73	11.00	0.58	10.11	7.88
X^2_O	5.32	4.58	4.06	4.46	3.45	2.68
X^2_{OEd}	4.56	4.23	3.19	3.85	2.61	1.73
X^2_{OSkal}	5.64	5.21	4.69	5.84	4.81	4.81
X^2_{McC}	5.97	5.79	5.39	6.19	6.05	6.54
X^2_{McCEd}	4.89	5.36	4.18	5.24	4.21	3.82
D_F	0.00	0.00	1.33	0.02	2.80	2.23
$\hat{C}$	4.48	4.58	4.66	2.30	4.11	4.57
D_C	0.00	0.00	0.03	0.00	2.09	4.34
X^2_F	0.00	5.47	5.70	6.23	5.94	6.57
X^2_{FEd}	0.00	4.99	4.40	4.53	4.23	3.88
IM	4.86	4.89	5.09	5.15	4.91	4.98
IM_{DIAG}	4.70	4.75	5.15	4.79	5.04	5.22
R_P	0.72	2.45	2.48	6.20	4.78	6.16
R_D	1.83	1.87	11.60	0.73	11.95	10.95
R_L	1.60	1.75	10.99	0.70	10.85	9.58
R_A	79.03	45.93	53.81	14.91	31.20	25.18
R_B	1.37	1.51	9.83	0.44	9.23	7.36
$\max_i \lvert r_{Pi}\rvert$	31.81	34.51	19.78	42.94	9.25	5.77
R_C	4.83	4.79	4.96	5.13	4.96	5.31
M=500						
X^2	0.00	0.40	1.50	4.90	3.46	4.75
D	8.40	1.00	38.56	0.00	26.10	14.33
X^2_O	4.80	5.00	4.48	5.70	4.19	4.37
X^2_{OEd}	4.50	5.00	4.06	5.60	3.55	3.54
X^2_{OSkal}	4.80	5.20	4.92	6.30	4.77	5.27
X^2_{McC}	4.80	5.60	5.24	6.60	5.43	6.23
X^2_{McCEd}	4.80	5.50	4.68	6.10	4.59	5.07
D_F	0.50	0.00	17.27	0.00	15.25	8.59
$\hat{C}$	3.10	5.20	4.67	4.50	4.95	5.12
D_C	0.00	0.00	0.00	0.00	1.51	5.30
X^2_F	0.00	5.90	5.28	6.50	5.33	6.13
X^2_{FEd}	0.00	5.80	4.92	5.10	4.55	5.22
IM	4.50	4.40	5.01	3.70	4.92	5.01
IM_{DIAG}	4.90	4.20	5.17	4.60	5.20	5.06
R_P	0.00	0.40	1.53	4.80	3.63	5.07
R_D	8.40	1.10	38.73	0.00	26.57	15.11
R_L	8.40	1.00	38.33	0.00	25.89	14.48
R_A	100.00	91.50	97.40	18.60	62.26	31.16
R_B	5.80	0.60	31.74	0.00	22.47	13.13
$\max_i \lvert r_{Pi}\rvert$	11.70	14.60	39.54	20.90	15.71	9.20
R_C	3.50	5.00	4.87	4.80	5.53	5.15

Tabelle 9: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0.693x_{1i}$, $x_1 \sim U(-6, 6)$, starker Regressionseffekt (OR=2)

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	11.43	14.83	8.00	19.43	5.77	4.60
D	0.00	0.00	0.40	0.00	2.53	4.14
X^2_O	6.52	5.95	4.37	5.77	3.37	2.15
X^2_{OEd}	4.95	4.99	5.60	4.59	7.09	5.35
X^2_{OSkal}	7.93	9.37	6.18	13.18	5.44	4.94
X^2_{McC}	7.22	6.93	5.60	7.24	5.33	5.30
X^2_{McCEd}	4.59	4.93	2.76	4.68	2.71	2.44
D_F	0.00	0.00	0.02	0.00	0.53	0.88
$\hat{C}$	5.20	5.44	5.12	3.42	4.98	4.66
D_C	0.00	0.00	0.00	0.00	0.09	1.03
X^2_F	0.00	5.90	4.91	5.05	5.45	4.81
X^2_{FEd}	0.00	4.49	7.32	3.68	4.02	3.81
IM	5.67	5.34	5.45	5.27	5.03	5.03
IM_{DIAG}	4.97	4.68	4.44	4.79	4.61	4.61
R_P	11.40	14.87	8.11	19.40	6.21	5.81
R_D	0.00	0.00	0.42	0.00	3.42	6.09
R_L	0.00	0.00	0.42	0.00	3.19	5.55
R_A	2.57	1.31	7.53	0.44	13.47	17.22
R_B	0.00	0.00	0.33	0.00	2.04	3.63
$\max_i \lvert r_{Pi} \rvert$	85.01	84.20	54.49	70.87	21.36	9.37
R_C	4.62	4.76	4.37	4.39	4.40	4.34
M=500						
X^2	11.08	16.30	8.17	24.70	5.97	5.37
D	0.00	0.00	0.00	0.00	1.66	5.24
X^2_O	5.72	5.70	4.83	6.30	4.52	4.23
X^2_{OEd}	4.66	5.00	3.76	5.30	3.30	3.00
X^2_{OSkal}	6.06	6.60	5.31	8.20	5.11	4.99
X^2_{McC}	6.00	6.40	5.43	6.70	5.39	5.63
X^2_{McCEd}	4.75	5.30	3.96	5.60	3.81	4.01
D_F	0.00	0.00	0.00	0.00	0.65	2.38
$\hat{C}$	4.56	5.30	4.67	3.20	4.82	4.93
D_C	0.00	0.00	0.00	0.00	0.00	0.26
X^2_F	0.00	5.10	5.83	5.90	5.75	5.75
X^2_{FEd}	0.00	4.30	3.61	4.60	4.17	3.99
IM	4.68	5.00	4.82	5.90	4.75	4.68
IM_{DIAG}	4.48	5.50	4.75	4.50	5.18	4.63
R_P	11.08	16.30	8.19	24.70	6.07	5.63
R_D	0.00	0.00	0.00	0.00	1.78	5.57
R_L	0.00	0.00	0.00	0.00	1.75	5.41
R_A	1.05	0.00	5.41	0.00	13.30	16.30
R_B	0.00	0.00	0.00	0.00	1.12	3.90
$\max_i \lvert r_{Pi} \rvert$	99.91	99.90	93.08	99.70	44.30	23.22
R_C	4.84	5.40	4.99	5.90	4.92	4.80

Tabelle 10: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0.693x_{1i}$, $x_1 \sim N(0,1)$, schwacher Regressionseffekt (OR=2)

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.06	0.20	1.14	1.16	4.08	4.94
D	76.68	43.56	43.48	13.15	13.76	8.33
X^2_O	6.10	4.64	3.67	4.14	3.62	2.77
X^2_{OEd}	5.45	4.57	3.53	3.56	2.91	2.05
X^2_{OSkal}	5.87	5.02	4.45	4.62	4.77	5.18
X^2_{McC}	6.73	5.44	5.35	5.78	6.46	7.25
X^2_{McCEd}	5.40	5.32	4.96	4.95	4.75	4.75
D_F	0.00	4.60	13.68	1.47	5.18	2.60
$\hat{C}$	4.94	4.99	4.62	2.87	4.67	4.95
D_C	0.00	0.00	1.35	0.00	6.28	6.06
X^2_F	0.00	5.22	5.44	5.70	6.53	7.30
X^2_{FEd}	0.00	5.11	5.18	4.58	4.91	4.80
IM	5.72	5.53	5.14	6.24	4.77	5.08
IM_{DIAG}	4.78	4.76	4.65	5.26	4.81	4.86
R_P	0.06	0.21	1.28	1.31	4.77	6.80
R_D	77.30	44.61	44.52	14.25	15.61	11.10
R_L	75.98	42.73	42.35	13.19	14.11	9.92
R_A	99.97	97.24	87.82	71.04	34.62	21.92
R_B	74.67	41.93	42.19	12.41	13.44	8.18
$\max_i \lvert r_{Pi} \rvert$	3.09	3.80	3.10	5.25	3.30	3.98
R_C	4.67	4.68	4.40	4.90	4.74	5.24
M=500						
X^2	0.00	0.00	1.02	0.20	3.43	4.57
D	100.00	99.00	97.76	58.50	33.91	11.42
X^2_O	6.10	4.40	4.34	4.00	4.13	4.16
X^2_{OEd}	5.00	4.40	4.23	3.90	3.75	3.40
X^2_{OSkal}	6.10	4.50	4.57	4.40	4.66	5.21
X^2_{McC}	6.30	4.50	5.18	5.00	5.51	6.13
X^2_{McCEd}	4.90	4.50	4.96	4.50	4.73	5.15
D_F	100.00	92.30	92.07	32.40	22.60	6.89
$\hat{C}$	5.50	5.30	5.10	5.40	4.85	5.16
D_C	0.00	0.00	2.30	0.00	9.36	6.55
X^2_F	0.00	5.00	5.13	4.60	5.45	6.19
X^2_{FEd}	0.00	5.00	5.01	4.20	4.76	5.19
IM	6.20	5.70	5.18	6.10	4.92	4.99
IM_{DIAG}	5.70	4.60	4.86	4.60	4.90	4.88
R_P	0.00	0.00	1.07	0.20	3.56	4.83
R_D	100.00	99.00	97.81	58.80	34.55	12.28
R_L	100.00	99.00	97.69	58.10	33.69	11.75
R_A	100.00	100.00	100.00	99.80	64.84	22.28
R_B	100.00	98.80	97.32	53.20	32.48	11.11
$\max_i \lvert r_{Pi} \rvert$	0.80	1.40	2.09	1.90	3.37	3.61
R_C	5.30	4.00	5.18	4.60	4.95	5.12

Tabelle 11: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 1.386x_{1i}$, $x_1 \sim N(0,1)$, starker Regressionseffekt (OR=4)

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	2.27	3.94	3.22	7.42	4.35	4.55
D	4.12	3.19	14.90	0.56	10.45	7.72
X^2_O	5.10	4.22	3.90	4.22	3.32	2.57
X^2_{OEd}	11.02	8.36	7.53	8.04	4.12	2.30
X^2_{OSkal}	5.31	4.81	4.64	5.86	4.68	4.71
X^2_{McC}	5.45	5.02	5.34	5.50	5.85	6.29
X^2_{McCEd}	9.01	7.07	5.30	6.62	4.13	3.87
D_F	0.00	0.03	2.50	0.06	3.15	2.20
$\hat{C}$	4.70	4.31	4.50	2.38	4.24	4.57
D_C	0.00	0.00	0.02	0.00	2.39	4.47
X^2_F	0.00	5.57	5.56	6.09	5.78	6.45
X^2_{FEd}	0.00	5.28	4.19	4.70	4.01	3.93
IM	5.25	5.75	5.84	5.46	5.29	4.58
IM_{DIAG}	4.77	4.81	5.10	4.97	4.83	4.82
R_P	2.31	3.97	3.44	7.65	5.03	6.14
R_D	4.25	3.50	15.49	0.72	11.84	10.29
R_L	3.94	3.16	14.69	0.65	10.91	9.22
R_A	89.68	56.87	61.34	15.81	29.82	23.26
R_B	3.53	2.79	13.78	0.52	9.57	7.37
$\max_i \lvert r_{Pi}\rvert$	34.11	36.48	17.97	43.08	8.12	5.16
R_C	4.24	4.37	4.31	4.03	4.44	4.49
M=500						
X^2	1.30	4.61	2.94	7.34	4.09	4.76
D	35.81	4.21	53.73	0.00	23.82	12.17
X^2_O	4.91	5.62	4.87	4.52	4.34	4.26
X^2_{OEd}	6.12	6.52	4.95	5.43	3.81	3.52
X^2_{OSkal}	5.02	5.82	5.33	5.03	4.82	5.12
X^2_{McC}	5.12	6.12	5.67	5.03	5.63	6.04
X^2_{McCEd}	3.11	4.51	4.22	3.92	4.40	4.66
D_F	3.51	0.30	28.72	0.00	14.13	7.22
$\hat{C}$	4.61	3.21	4.44	3.72	4.57	4.80
D_C	0.00	0.00	0.00	0.00	1.27	4.10
X^2_F	0.00	4.21	5.34	5.53	5.62	6.02
X^2_{FEd}	0.00	4.21	4.78	4.82	4.70	4.98
IM	4.11	6.42	6.01	5.13	5.53	5.84
IM_{DIAG}	3.11	5.72	5.09	4.02	4.96	5.06
R_P	1.30	4.61	3.00	7.54	4.29	5.01
R_D	35.61	4.21	53.94	0.00	24.30	12.78
R_L	35.51	4.21	53.34	0.00	23.72	12.37
R_A	100.00	97.39	99.12	22.71	57.60	25.81
R_B	29.69	3.11	47.49	0.00	21.01	11.09
$\max_i \lvert r_{Pi}\rvert$	61.79	68.10	37.48	72.26	16.46	9.43
R_C	3.61	5.02	5.40	3.92	4.91	5.18

Tabelle 12: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0.223x_{1i}$, $x_1 \sim \chi^2_4$, schwacher Regressionseffekt (OR=1.25)

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.69	1.42	1.81	3.42	4.03	4.55
D	22.72	17.64	26.62	6.90	12.61	8.40
X^2_O	4.08	3.61	3.51	3.82	3.32	2.71
X^2_{OEd}	12.74	8.81	6.69	8.03	3.82	2.66
X^2_{OSkal}	4.00	4.02	4.05	4.61	4.77	4.82
X^2_{McC}	4.80	4.71	4.98	5.27	6.17	6.63
X^2_{McCEd}	11.66	9.22	5.94	8.71	4.87	4.28
D_F	0.00	0.72	5.79	0.69	4.23	2.66
$\hat{C}$	4.97	4.68	4.70	3.20	4.46	4.54
D_C	0.00	0.00	0.20	0.00	3.97	5.06
X^2_F	0.00	5.84	5.58	6.09	6.41	6.99
X^2_{FEd}	0.00	5.36	4.68	4.67	4.77	4.29
IM	4.54	4.93	4.92	4.88	4.35	4.54
IM_{DIAG}	4.39	4.61	4.59	4.47	4.77	4.88
R_P	0.74	1.48	2.14	3.68	4.79	6.16
R_D	23.20	18.30	27.48	7.77	14.35	10.72
R_L	21.85	17.27	26.08	7.13	13.10	9.59
R_A	98.76	88.87	76.15	55.30	33.24	23.39
R_B	20.68	16.11	24.84	6.35	11.87	8.06
$\max_i \lvert r_{Pi} \rvert$	14.44	16.00	8.82	20.69	5.94	4.63
R_C	3.59	3.73	3.98	3.99	4.34	4.75
M=500						
X^2	0.30	1.21	1.67	3.34	3.70	4.54
D	96.48	79.11	84.57	26.59	30.56	12.62
X^2_O	5.34	4.84	4.33	4.04	4.23	4.06
X^2_{OEd}	10.67	6.36	5.05	6.88	4.00	3.52
X^2_{OSkal}	5.44	4.94	4.65	4.45	4.83	5.14
X^2_{McC}	5.64	5.45	5.07	4.85	5.63	6.29
X^2_{McCEd}	7.65	5.45	4.62	5.56	4.67	4.94
D_F	64.25	44.90	64.41	10.11	19.46	7.49
$\hat{C}$	4.93	3.94	4.97	3.74	4.97	4.96
D_C	0.00	0.00	0.02	0.00	4.44	5.80
X^2_F	0.00	5.75	5.39	6.57	5.58	6.37
X^2_{FEd}	0.00	5.65	5.10	6.47	4.95	5.16
IM	5.74	5.55	5.81	4.75	5.93	5.37
IM_{DIAG}	5.34	5.65	5.38	4.25	5.44	5.18
R_P	0.30	1.21	1.68	3.34	3.85	4.88
R_D	96.48	79.11	84.65	26.79	31.02	13.30
R_L	96.48	79.01	84.19	26.49	30.23	12.81
R_A	100.00	100.00	99.93	97.67	63.47	25.63
R_B	94.96	75.68	81.08	24.17	27.93	11.92
$\max_i \lvert r_{Pi} \rvert$	30.51	29.47	17.69	36.80	9.61	6.84
R_C	4.23	5.25	4.77	3.44	4.74	4.71

Tabelle 13: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0.405x_{1i}$, $x_1 \sim \chi^2_4$, starker Regressionseffekt (OR=1.5)

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	2.73	4.94	3.12	9.43	3.88	4.44
D	0.03	0.51	3.31	0.72	6.36	6.75
X^2_O	2.88	2.89	2.47	2.99	2.54	2.10
X^2_{OEd}	28.47	20.28	17.57	18.36	8.67	5.02
X^2_{OSkal}	3.16	3.58	3.08	7.68	3.88	5.27
X^2_{McC}	3.48	3.78	3.47	4.02	4.71	5.83
X^2_{McCEd}	27.37	19.37	13.20	16.85	6.19	4.22
D_F	0.00	0.00	0.22	0.04	1.65	1.76
$\hat{C}$	4.04	4.46	4.37	3.60	4.25	4.45
D_C	0.00	0.00	0.00	0.00	0.57	2.61
X^2_F	0.00	6.88	5.37	5.78	5.62	6.72
X^2_{FEd}	0.00	5.21	3.04	4.02	3.42	3.51
IM	3.89	3.50	4.04	3.79	4.67	4.42
IM_{DIAG}	3.71	3.40	3.75	3.47	4.74	4.22
R_P	2.78	5.03	3.39	9.61	4.60	6.24
R_D	0.03	0.54	3.61	0.85	7.62	9.30
R_L	0.03	0.51	3.36	0.87	7.18	8.36
R_A	48.25	33.79	33.83	18.90	23.27	22.37
R_B	0.03	0.40	2.82	0.55	5.71	6.20
$\max_i \lvert r_{Pi} \rvert$	49.41	55.60	28.88	42.59	12.48	6.98
R_C	3.63	3.34	3.69	3.54	3.91	4.02
M=500						
X^2	7.96	11.44	5.26	11.72	5.30	5.06
D	0.00	0.24	5.04	0.00	11.19	8.98
X^2_O	5.47	4.87	3.97	4.31	3.85	3.60
X^2_{OEd}	21.14	17.76	14.07	12.44	8.83	5.67
X^2_{OSkal}	5.97	5.60	4.26	4.78	4.37	4.56
X^2_{McC}	5.97	5.84	4.38	4.55	4.85	5.39
X^2_{McCEd}	12.44	5.60	3.64	3.59	3.15	3.47
D_F	0.00	0.00	1.20	0.00	5.61	4.87
$\hat{C}$	4.48	4.62	4.91	4.78	5.08	4.97
D_C	0.00	0.00	0.00	0.00	0.01	1.11
X^2_F	0.00	6.08	5.45	7.42	5.19	5.85
X^2_{FEd}	0.00	5.84	3.83	5.98	3.78	4.18
IM	4.73	4.87	5.25	4.78	5.56	5.32
IM_{DIAG}	5.47	4.87	5.14	4.78	5.48	5.20
R_P	8.21	11.44	5.28	11.96	5.50	5.35
R_D	0.00	0.24	5.13	0.00	11.72	9.51
R_L	0.00	0.24	5.02	0.00	11.33	9.19
R_A	95.52	78.59	74.67	35.89	40.30	23.51
R_B	0.00	0.00	3.30	0.00	8.58	7.70
$\max_i \lvert r_{Pi} \rvert$	90.05	93.92	65.29	92.34	31.15	16.33
R_C	4.98	4.87	4.68	5.98	5.15	5.04

Tabelle 14: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0.118x_{1i} + 0.223x_{2i} + 0.405x_{3i}$, x_j iid $U(-6, 6)$

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	5.97	7.98	5.26	10.00	5.58	5.00
D	0.40	1.77	6.75	2.03	8.11	6.92
X^2_O	6.76	5.35	3.74	4.61	2.02	1.00
X^2_{OEd}	6.87	6.22	5.54	6.20	3.21	1.34
X^2_{OSkal}	8.58	8.85	6.19	11.08	6.41	5.69
X^2_{McC}	7.34	6.59	5.57	6.29	6.26	6.20
X^2_{McCEd}	5.33	5.43	4.03	6.06	4.14	3.83
D_F	0.00	0.00	0.22	0.03	1.05	0.66
$\hat{C}$	4.52	4.54	4.03	3.45	3.03	1.72
D_C	0.00	0.00	0.00	0.00	1.27	3.47
X^2_F	0.00	5.98	5.27	5.24	6.13	6.17
X^2_{FEd}	0.00	3.81	3.69	3.32	3.87	3.62
IM	6.88	6.29	6.30	6.92	5.65	0.66
IM_{DIAG}	5.34	5.35	5.22	5.31	5.31	4.76
R_P	5.92	7.98	5.61	9.97	7.44	9.40
R_D	0.50	2.28	8.33	3.15	12.20	13.43
R_L	0.35	1.87	7.23	2.62	9.90	11.27
R_A	69.37	56.55	51.54	42.22	39.40	39.47
R_B	0.32	1.42	5.88	1.68	7.45	6.48
$\max_i \lvert r_{Pi} \rvert$	57.87	49.07	29.92	38.03	11.98	5.62
R_C	5.07	5.01	5.07	5.04	4.87	4.74
M=500						
X^2	3.60	6.40	3.99	9.30	4.89	4.69
D	0.70	2.10	15.42	1.50	16.70	11.09
X^2_O	7.20	5.90	4.37	4.50	3.67	2.63
X^2_{OEd}	5.60	4.10	3.44	3.60	2.99	1.98
X^2_{OSkal}	8.00	6.60	5.60	5.60	5.34	4.78
X^2_{McC}	7.80	6.30	5.61	5.30	5.79	5.67
X^2_{McCEd}	5.20	4.30	4.16	3.80	4.53	4.20
D_F	0.00	0.00	3.79	0.20	6.59	4.06
$\hat{C}$	4.60	3.20	4.44	3.70	4.55	3.60
D_C	0.00	0.00	0.00	0.00	0.28	2.49
X^2_F	0.00	5.70	5.36	5.50	5.52	5.61
X^2_{FEd}	0.00	4.90	4.37	4.40	4.63	4.38
IM	4.80	4.00	5.80	5.70	5.32	5.38
IM_{DIAG}	2.70	4.50	5.11	3.90	5.14	4.96
R_P	3.60	6.40	4.02	9.00	5.08	5.23
R_D	0.70	2.50	16.26	1.60	17.74	12.65
R_L	0.70	2.00	15.33	1.50	16.66	11.70
R_A	99.50	95.60	90.62	75.40	55.86	34.49
R_B	0.30	1.20	11.67	1.40	13.36	9.58
$\max_i \lvert r_{Pi} \rvert$	87.50	85.50	55.22	76.80	22.66	12.25
R_C	5.50	4.80	5.05	5.80	4.85	5.07

Tabelle 15: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Nullhypothese eines korrekt spezifizierten Modells
Modell: $\text{logit}(\pi_i) = 0.223x_{1i} + 0.405x_{2i} + 0.693x_{3i}$, x_j iid $N(0,1)$

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.47	1.12	2.20	2.56	4.71	5.21
D	56.76	43.86	38.49	21.18	14.05	8.18
X^2_O	6.34	4.21	2.81	3.23	1.54	0.91
X^2_{OEd}	6.02	4.25	2.96	3.11	1.30	0.74
X^2_{OSkal}	6.70	5.29	4.44	4.99	5.01	4.93
X^2_{McC}	7.13	5.98	5.38	6.23	6.63	6.93
X^2_{McCEd}	5.57	5.53	4.89	5.19	4.96	4.42
D_F	0.00	1.83	6.16	1.21	2.25	0.89
$\hat{C}$	4.30	4.57	4.07	3.33	2.56	1.62
D_C	0.00	0.00	0.67	0.00	5.41	6.08
X^2_F	0.00	6.14	5.69	5.04	6.72	6.97
X^2_{FEd}	0.00	5.41	5.02	3.88	4.96	4.46
IM	6.57	6.43	5.32	6.57	5.02	0.59
IM_{DIAG}	6.15	5.67	5.30	5.86	4.88	5.50
R_P	0.48	1.25	2.79	3.06	6.86	10.16
R_D	59.57	46.85	41.95	24.53	18.33	13.66
R_L	55.30	42.51	37.32	21.09	15.46	12.16
R_A	99.90	98.57	89.00	88.19	46.83	36.69
R_B	53.89	41.59	36.68	19.93	13.52	8.00
$\max_i \lvert r_{Pi}\rvert$	9.99	9.46	6.85	8.21	4.53	4.19
R_C	4.54	4.88	4.66	4.50	4.63	5.23
M=500						
X^2	0.00	0.10	1.41	0.90	4.01	4.31
D	100.00	99.70	95.89	86.60	32.86	11.79
X^2_O	7.40	4.40	3.94	4.20	3.27	2.58
X^2_{OEd}	6.30	4.30	3.80	3.80	2.92	2.15
X^2_{OSkal}	7.50	4.60	4.89	5.40	5.22	4.69
X^2_{McC}	7.80	4.70	5.26	5.90	5.99	5.67
X^2_{McCEd}	6.20	4.50	5.07	5.50	5.30	4.67
D_F	99.10	95.00	83.69	59.90	17.56	4.66
$\hat{C}$	4.90	5.70	4.86	5.20	4.50	4.21
D_C	0.00	0.00	0.61	0.00	7.73	5.96
X^2_F	0.00	5.20	5.31	5.80	5.89	5.75
X^2_{FEd}	0.00	5.00	5.11	5.40	5.29	4.71
IM	7.50	5.80	5.42	6.80	5.34	4.84
IM_{DIAG}	5.10	5.00	4.88	5.80	5.21	5.05
R_P	0.00	0.10	1.51	1.10	4.31	5.07
R_D	100.00	99.70	96.03	86.60	34.14	13.13
R_L	100.00	99.70	95.62	86.10	32.45	12.16
R_A	100.00	100.00	100.00	100.00	68.67	29.70
R_B	100.00	99.60	94.91	85.30	30.98	11.22
$\max_i \lvert r_{Pi}\rvert$	6.30	6.60	6.88	6.40	5.33	4.64
R_C	5.80	5.00	4.63	4.80	4.74	4.85

B Ergebnisse der Simulationsuntersuchung: Alternative

Dargestellt ist das empirische Signifikanzniveau der verglichenen Anpassungstests, d.h. die Anzahl (in %), in der der jeweilige Anpassungstest die Nullhypothese verworfen hat. Dieses sollte in diesem Fall (unter der Alternative eines fehlspezifiziertem Modell) möglichst groß sein. Optimal wäre ein empirisches Signifikanzniveau von 100, in diesem Fall würde der Anpassungstest in allen Fällen die eingestellte Fehlspezifikation erkennen. Zur Erinnerung, in den Situationen unter der Alternative wird jeweils ein Modell mit einem festen, nicht zufälligen Intercept und einer einzelnen Kovariable geschätzt.
In den Spalten sind die Belegungen der Kovariablenmuster abgetragen. Dabei stehen konstante m_i $(1, 2, 5, 10)$ für Belegungen, in denen alle Kovariablenmuster die selbe Belegung haben. Die Bezeichnung „1-2“ steht für eine Belegung, in dem eine Hälfte der Kovariablenmuster einfach, die andere Hälfte doppelt belegt sind. Die Bezeichnung „1-10“ steht für eine Belegung, in der die Kovariablenmuster einfach, doppelt, fünffach und zehnfach im Verhältnis $64:21:9:6$ belegt sind.

Tabelle 16: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Alternative eines fehlspezifierten Modells
Fehlspezifikation: Falsch spezifizierte Form der Kovariablen
Modell: $\text{logit}(\pi_i) = 0.405x_{1i}^2$, $x_1 \sim U(-6, 6)$

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.43	0.65	0.43	0.92	2.14	5.21
D	0.00	0.00	0.41	0.00	3.71	7.59
X^2_O	0.06	0.05	0.16	0.13	1.22	3.02
X^2_{OEd}	0.08	0.10	0.46	0.10	2.89	5.08
X^2_{OSkal}	0.17	0.71	0.35	1.55	2.31	5.72
X^2_{McC}	0.08	0.08	0.20	0.17	2.50	6.45
X^2_{McCEd}	0.03	0.04	0.10	0.02	1.18	3.78
$\hat{C}$	5.06	5.02	5.25	2.44	4.87	5.24
X^2_F	0.00	10.29	6.67	10.29	9.86	13.42
X^2_{FEd}	0.00	8.41	6.23	5.68	7.09	9.40
IM	10.57	11.12	11.05	10.49	10.46	10.58
IM_{DIAG}	12.86	13.03	13.07	13.14	12.74	12.64
R_C	25.16	25.55	24.51	25.24	23.05	20.23
M=500						
X^2	0.00	0.00	0.00	0.00	0.80	5.50
D	0.00	0.00	0.00	0.00	2.30	12.70
X^2_O	0.00	0.00	0.00	0.00	0.70	4.40
X^2_{OEd}	0.00	0.00	0.00	0.00	0.40	3.30
X^2_{OSkal}	0.00	0.00	0.00	0.00	0.80	5.50
X^2_{McC}	0.00	0.00	0.00	0.00	0.80	6.50
X^2_{McCEd}	0.00	0.00	0.00	0.00	0.60	4.30
$\hat{C}$	59.90	57.50	57.50	60.50	55.40	54.90
X^2_F	0.00	13.10	9.80	21.40	18.80	33.50
X^2_{FEd}	0.00	12.40	5.70	16.80	14.50	27.30
IM	90.10	86.30	87.60	88.20	86.20	86.00
IM_{DIAG}	94.40	92.70	92.40	92.90	92.20	92.50
R_C	97.90	97.30	97.50	96.90	96.30	96.50

Tabelle 17: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Alternative eines fehlspezifierten Modells
Fehlspezifikation: Nicht ins Modell aufgenommene Kovariable
Modell: $\text{logit}(\pi_i) = 0.405x_{1i} + 0.223x_{2i}$, x_j iid $U(-6, 6)$

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.25	2.07	6.76	15.74	28.42	43.08
D	5.50	9.21	34.82	8.38	49.77	53.23
X^2_O	5.04	9.68	13.10	17.28	26.17	34.93
X^2_{OEd}	4.49	9.38	11.39	15.60	22.40	29.76
X^2_{OSkal}	5.27	10.50	14.80	19.95	31.56	43.88
X^2_{McC}	5.81	11.47	16.80	21.23	36.05	49.10
X^2_{McCEd}	4.75	11.14	14.27	19.11	29.84	40.93
$\hat{C}$	4.29	6.05	7.48	10.51	18.89	42.99
X^2_F	0.00	15.32	18.20	32.00	36.59	48.96
X^2_{FEd}	0.00	14.48	15.61	27.38	30.71	41.13
IM	4.73	5.96	6.68	12.99	12.12	24.08
IM_{DIAG}	4.75	5.70	6.45	11.60	11.21	20.53
R_C	5.34	5.72	6.08	10.08	9.48	15.12
M=500						
X^2	0.00	2.50	19.70	33.50	77.80	94.90
D	38.10	30.20	93.80	15.10	97.30	98.30
X^2_O	3.80	22.00	37.60	46.60	80.50	94.60
X^2_{OEd}	3.50	21.90	36.20	45.90	78.40	93.30
X^2_{OSkal}	3.90	22.30	38.60	48.20	82.40	95.50
X^2_{McC}	4.20	22.70	40.20	50.70	83.80	95.90
X^2_{McCEd}	3.70	22.70	38.60	48.90	82.20	95.30
$\hat{C}$	5.60	6.70	8.00	20.70	18.90	38.70
X^2_F	0.00	35.00	41.70	82.10	85.00	95.90
X^2_{FEd}	0.00	34.40	40.90	80.00	83.00	95.10
IM	6.60	5.40	6.70	11.50	13.10	19.30
IM_{DIAG}	5.80	6.30	6.60	11.90	9.70	17.30
R_C	4.80	6.40	5.80	9.50	8.00	13.50

Tabelle 18: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Alternative eines fehlspezifierten Modells
Fehlspezifikation: Overdispersion
Modell: $\text{logit}(\pi_i) = \beta_0 + 0.405x_{1i}$, $x_1 \sim U(-6, 6)$ mit $E(\beta_0) = 0$ und $\text{Var}(\beta_0) = 0.323$

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	0.36	2.07	4.51	11.35	15.27	24.20
D	3.28	4.63	22.21	3.08	30.19	33.62
X^2_O	5.66	6.79	8.67	10.93	13.46	17.76
X^2_{OEd}	4.82	6.53	7.26	9.70	11.21	13.95
X^2_{OSkal}	5.85	7.49	9.78	13.08	17.29	24.97
X^2_{McC}	6.26	8.22	11.09	13.87	20.42	29.74
X^2_{McCEd}	5.24	7.97	9.26	12.28	16.05	22.60
$\hat{C}$	4.59	5.47	5.86	5.87	11.04	24.15
X^2_F	0.00	10.20	12.08	18.71	20.71	29.58
X^2_{FEd}	0.00	9.80	9.94	14.74	16.51	22.98
IM	4.71	5.48	5.38	8.51	8.42	14.42
IM_{DIAG}	4.56	5.29	5.04	8.23	7.99	13.20
R_C	5.06	6.00	5.52	8.27	7.39	10.88
M=500						
X^2	0.00	0.70	8.70	14.10	43.10	65.80
D	23.90	10.50	77.10	2.30	85.40	83.50
X^2_O	4.50	11.50	21.10	20.10	46.90	64.50
X^2_{OEd}	4.20	11.50	19.70	19.70	43.40	61.10
X^2_{OSkal}	4.50	11.70	21.70	21.80	49.60	67.10
X^2_{McC}	4.70	11.90	23.00	23.00	52.40	69.40
X^2_{McCEd}	4.50	11.90	21.50	21.80	49.30	66.70
$\hat{C}$	4.60	6.00	5.20	12.10	11.10	23.10
X^2_F	0.00	21.30	23.20	46.40	52.10	69.90
X^2_{FEd}	0.00	20.80	22.20	44.20	49.90	67.30
IM	5.40	6.90	5.10	9.20	8.00	14.90
IM_{DIAG}	4.30	6.80	4.00	8.60	7.90	12.30
R_C	4.50	7.90	5.30	6.10	5.70	10.70

Tabelle 19: Empirisches Signifikanzniveau der verglichenen Anpassungstests unter der Alternative eines fehlspezifierten Modells
Fehlspezifikation: Falsch spezifizierte Linkfunktion
Modell: $\log[-\log(1-\pi_{\mathrm{i}})] = 0.405x_{1i}$, $x_1 \sim U(-6,6)$

Anpassungstest	Belegung m_i					
	1	1-2	2	1-10	5	10
M=100						
X^2	1.11	2.05	1.73	3.99	2.71	4.14
D	0.00	0.00	0.62	0.00	3.54	5.52
X^2_O	0.23	0.24	0.91	0.33	1.55	2.11
X^2_{OEd}	0.41	0.29	2.10	0.36	4.44	4.70
X^2_{OSkal}	0.45	1.33	1.41	4.35	2.80	4.69
X^2_{McC}	0.26	0.32	1.22	0.44	2.88	4.98
X^2_{McCEd}	0.19	0.14	0.68	0.25	1.55	2.89
$\hat{C}$	3.76	4.31	3.79	2.01	4.18	4.14
X^2_F	0.00	5.85	6.16	5.64	6.93	8.32
X^2_{FEd}	0.00	4.17	6.91	3.12	4.81	5.67
IM	6.32	6.76	6.47	6.18	6.34	6.10
IM_{DIAG}	8.70	9.32	9.29	8.89	8.94	8.05
R_C	3.68	3.69	3.86	3.71	4.38	4.83
M=500						
X^2	0.00	0.10	0.20	0.40	1.70	3.70
D	0.00	0.00	0.00	0.00	2.90	6.90
X^2_O	0.00	0.00	0.10	0.00	1.30	2.50
X^2_{OEd}	0.00	0.00	0.00	0.00	0.80	1.70
X^2_{OSkal}	0.00	0.00	0.10	0.00	1.40	3.30
X^2_{McC}	0.00	0.00	0.10	0.00	1.80	3.70
X^2_{McCEd}	0.00	0.00	0.00	0.00	1.20	2.30
$\hat{C}$	20.00	20.60	19.70	20.40	20.10	19.50
X^2_F	0.00	6.30	6.70	5.90	10.60	12.60
X^2_{FEd}	0.00	5.60	3.70	4.00	7.40	9.00
IM	42.10	40.40	41.10	39.80	41.70	38.50
IM_{DIAG}	54.10	53.00	54.50	52.70	55.00	51.70
R_C	27.50	26.40	27.70	28.90	28.10	26.70

Lebenslauf

PERSONALIEN	
Name und Vorname:	Oliver Kuß
Geburtsdatum:	16.07.1969
Geburtsort:	Crailsheim
Familienstand:	ledig
Vater:	Roland Kuß
Mutter:	Irmgard Kuß, geb. Hartnagel
SCHULISCHER WERDEGANG	
1975-1979	Grund- und Hauptschule Fichtenau
1979-1988	Albert-Schweitzer-Gymnasium Crailsheim
6.5.1988	Abitur
UNIVERSITÄRER WERDEGANG	
SS 1990 - SS 1991	Studium der Mathematik und der Politischen Wissenschaften (Staatsexamen) an der Universität Heidelberg
WS 1991/92	Studium der Mathematik und der Geographie (Staatsexamen) an der Universität Heidelberg
30.11.1996	Staatsexamen
1.4.1997-31.12.1999	Wissenschaftlicher Mitarbeiter an der Dermatologischen Universitätsklinik Erlangen
1.1.2000-31.12.2000	Wissenschaftlicher Mitarbeiter an der Abteilung Klinische Sozialmedizin, Universitätsklinik Heidelberg