Oliver Kuß
Dr. sc. hum.

Globale Anpassungstests im logistischen Regressionsmodell bei fehlenden Messwiederholungen

Geboren am 16.07.1969 in Crailsheim
Reifeprüfung am 06.05.1988 in Crailsheim
Zwischenprüfung am 08.10.1991 (Mathematik) und 22.10.93 (Geograhie) an der Universität Heidelberg
Staatsexamen am 30.11.1996 an der Universität Heidelberg

Promotionsfach: Arbeits- und Sozialmedizin
Doktorvater: Prof. Dr. med. T.L. Diepgen

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 v.a. 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 und 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 spezifizierten 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 der Informationsmatrix-Test IM_{DIAG}, der zwei unter korrekter Modellspezifikation äquivalente Schätzer der Informationsmatrix vergleicht und der dritte schließlich 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, v.a. 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.