Globale Anpassungstests im logistischen Regressionsmodell bei fehlenden Messwiederholungen

Oliver Kuß
Institut für Medizinsche Epidemiologie, Biometrie und Informatik,
Universität Halle-Wittenberg,
Magdeburger Str. 27, 06097 Halle/Saale
Oliver.Kuss@medizin.uni-halle.de

Programm:

1. Das logistische Regressionsmodell
2. Überprüfung des Modellanpassung
3. Das Problem der fehlenden Messwiederholungen
4. Lösungsvorschläge
5. Welche Lösung ist die beste???
6. Eigene Untersuchungen
7. Fazit
8. Literatur

1. Das logistische Regressionsmodell

Standardmethode zur Regressionsanalyse binärer Zielgrößen

Gründe:

- Leichte Interpretierbarkeit der Parameter als Odds-Ratios
- Prognosen für das Eintreten des Zielereignisses sind möglich
- Verfügbarkeit von geeigneter Software
- Analyse von prospektiven und retrospektiven Beobachtungsstudien möglich
- Ausgereifte Methodik (Loglineares Modell, GLIM, nichtlineares Regressionsmodell)

Notation:

N unabhängige **nach Kovariablenmustern gruppierte** Beobachtungen (y_i , x_i), i=1, ..., N

x_i : Vektor von p+1 Kovariablen,
y_i : Anzahl der Erfolge, Realisation von $Y_i \sim B(m_i, \pi_i)$,
m_i : Anzahl der Versuche,

$M=\sum_{i=1}^{N} m_i$: Anzahl der individuellen Beobachtungen

Daten:

		Zielgröße 1	Zielgröße 0	
	1	Y_1	m_1-Y_1	m_1
Kovariablen	2	Y_2	m_2-Y_2	m_2
Muster	:	:	:	:
	N	Y_N	m_N-Y_N	m_N

Beispiel:

Stetige Kovariable(n): N=M ($m_i \equiv 1$)

		Zielgröße 1	Zielgröße 0	
	1	1	0	1
Kovariablen	2	0	1	1
Muster	:	:	:	:
	N	1	0	1

Modellgleichung

$$\log\left(\pi_i / 1-\pi_i\right) = \sum_{j=0}^{p} x_{ij}\beta_j$$

mit $\beta_j = (\beta_0,..., \beta_p)$ als dem Vektor der Regressionsparameter.

Schätzung der Parameter β_j durch ML

2. Überprüfung der Modellanpassung

Statistische Modellbildung läuft in zwei Schritten ab (Hosmer et al., 1991):

Modellwahl und **Modellüberprüfung**

Modellwahl: Wie lässt sich die Variation in der Zielgröße durch andere an den Beobachtungen gemessene Kovariablen erklären (*systematische Komponente*)

Modellüberprüfung: Untersuchung der nicht durch die systematische Komponente erklärten Variation durch Vergleich von Beobachtungen und Prognose des Modells (*Fehlerkomponente*)

Systematische Komponente beschreibt den „mittleren" Wert der Zielgröße, die Fehlerkomponente beschreibt die Abweichung von diesem „mittleren" Wert

Modellüberprüfung geschieht auf zwei Ebenen

1) Betrachte individuelle Beiträge der einzelnen Beobachtungen zu diesen Statistiken (auch graphisch): Residuenanalyse
2) Berechne Anpassungsmaße und beurteile Anpassung anhand einer einzelnen Zahl: Anpassungstests

Zwei Klassiker unter den globalen Anpassungstests:

Pearson-Statistik:

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

Devianz:

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

Große Werte von X^2, D zeigen schlechte Anpassung an

Statistischer Test: Vergleiche X^2, D mit Quantil der χ^2-Verteilung mit N-p-1 Freiheitsgraden

3. Problem der fehlenden Messwiederholungen

Gültigkeit der Prüfverteilung von X^2, D hängt wesentlich von der Annahme von Meßwiederholungen ab (N fest, $m_i \to \infty$ für alle i)

Unrealistisch bei großer Anzahl von Kovariablen oder stetigen Kovariablen

Katastrophal:
Im Extremfall $m_i \equiv 1$ degeneriert D zu

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

und enthält keinerlei Information mehr über den Fit.

Für X^2 gilt in diesem Fall: $X^2 \approx N$

Das Problem ist altbekannt, ...

- 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, 1990)

- Neither X^2 nor D is appropriate in the many strata standard asymptotic model (*Anm.: p fest, N und* $m_i \rightarrow \infty$), because under this model there is no χ^2-limiting distribution. (Santner/Duffy, 1989)

- Thus, p-values calculated for X^2 and D when M ≈N, using the χ^2-distribution, are incorrect. (Hosmer/Lemeshow, 1989)

- The effect of sparseness is noticed mainly on D and X^2, which fail to have the properties required for goodness-of-fit statistics. (McCullagh/Nelder, 1989)

... aber was ist die Lösung???

- 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, 1999)

- 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, 1989)

- 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, 1989)

4. Lösungsvorschläge (Auswahl)

4.1 Modifizierte Prüfverteilung

- Unter n, $m_i \to \infty$ sind X^2, D asymptotisch normalverteilt (Osius/Rojek, 1992; McCullagh, 1986)

4.2 Gruppierung von Beobachtungen

- Hosmer-Lemeshow-Test (Hosmer/Lemeshow, 1980) Inzwischen Quasi-Standard, Anwendung aber nicht ohne Probleme (Hosmer et al, 1997, Bertolini et al., 2000)

4.3 Verwendung anderer Teststatistiken

- $\mathbf{X_F^2}$ (Farrington, 1996)

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

Approximative Momente:

$$E\left(X_F^2 \mid \hat{\beta}\right) = N - p - 1 + \sum_{i=1}^{N} \hat{\pi}_i(1-\hat{\pi}_i)\hat{Q}_{ii}$$

$$Var\left(X_F^2 \mid \hat{\beta}\right) = 2\left(1 - \frac{p+1}{N}\right)\sum_{i=1}^{N} \frac{m_i - 1}{m_i}$$

mit $\hat{Q} = X\left(X^t\hat{W}X\right)^{-1}X^t$, $\hat{W} = diag\left(m_i\hat{\pi}_i(1-\hat{\pi}_i)\right)$.

Teststatistik:

$$Z_F = \frac{X_F^2 - E\left(X_F^2 \mid \hat{\beta}\right)}{Var\left(X_F^2 \mid \hat{\beta}\right)^{1/2}}$$

ist unter der Nullhypothese N(0,1)-verteilt.

Problem: Für $m_i \equiv 1$ gilt $X_F^2 = N$

- **IM-Test** (White, 1982; Orme, 1988)
  Informationsmatrix-Gleichung:

$$-E\left(\frac{\partial^2 L}{\partial\beta\partial\beta'}\right)=E\left(\frac{\partial L}{\partial\beta}\frac{\partial L}{\partial\beta'}\right)$$

  Berechne ML-Schätzer der beiden Matrizen, Addition der Elemente auf der Hauptdiagonalen liefert ((p+1)×1)-Vektor

$$\hat{d}=\sum_{i=1}^{M}\left(y_i-\hat{\pi}_i\right)\left(1-2\hat{\pi}_i\right)z_i$$

  mit $z_i=(1,x_{i1}^2,\ldots,x_{ip}^2)^t$.

  Teststatistik:

$$IM=\frac{1}{M}\hat{d}^t\hat{V}^{-1}\hat{d}$$

  ist unter der Nullhypothese χ^2-verteilt mit (p+1) Freiheitsgraden

  mit

$$\hat{V}=\frac{1}{M}\left[Z^{*t}\left(I-X^*\left(X^{*t}X^*\right)^{-1}X^{*t}\right)Z^*\right],$$

$$X^*=\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)}X,$$

$$Z^*=\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)}(1-2\hat{\pi}_i)Z,$$

  Z die Matrix mit den z_i als Zeilen.

- $\mathbf{R_C}$ (Copas, 1986, Hosmer et al., 1997)

$$R_C = \sum_{i=1}^{M} \left(y_i - m_i \hat{\pi}_i\right)^2$$

Summation von rohen Pearson-Residuen

Asymptotische Momente:

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

$$Var\left(R_C - \sum_{i=1}^{M} \hat{\pi}_i \left(1 - \hat{\pi}_i\right) \right) = \left(1 - 2\hat{\pi}\right)^t \left(\hat{W} - \hat{W}\hat{Q}\hat{W}\right) \left(1 - 2\hat{\pi}\right)$$

mit $\hat{Q} = X\left(X^t \hat{W} X\right)^{-1} X^t$, $\hat{W} = diag\left(m_i \hat{\pi}_i (1 - \hat{\pi}_i)\right)$.

Teststatistik:

$$Z_C = \frac{R_C - \sum_{i=1}^{M} \hat{\pi}_i \left(1 - \hat{\pi}_i\right)}{Var\left(R_C - \sum_{i=1}^{M} \hat{\pi}_i \left(1 - \hat{\pi}_i\right) \right)^{1/2}}$$

ist unter der Nullhypothese N(0,1)-verteilt

Beispiel:
Berufsbedingte Handekzeme bei Auszubildenden im Friseurhandwerk
M=574 (340 „Erfolge“),
Mehrere Kovariablen (p=6): genetische Disposition, Arbeitsbelastungen, Confounder,
N=334,
Verteilung der m_i:

m_i	Häufigkeit
1	205 (61%)
2	68 (20%)
3	35 (11%)
>3	26 (8%)

Überprüfung der Modellanpassung:

	p-Wert
X^2	0,053
D	0,012
$X^2_{\boldsymbol{O}}$	0,044
$X^2_{\boldsymbol{M}}$	0,031
HL-Test	0,451
$X^2_{\boldsymbol{F}}$	0,408
IM-Test	0,365
$R_{\boldsymbol{C}}$	0,062

Wer hat recht???

5. Welche Lösung ist die beste???

Es liegt erst eine umfassendere Simulations-Studie zum Verhalten von Anpassungstests im logistischen Regressionsmodell vor (Hosmer et al., 1997)

Ergebnis:
R_C und X^2_M als „Sieger“

Ergänzungsbedarf:
- Aktualisierung, Vollständigkeit
- Verschiedene m_i

6. Eigene Untersuchungen

6.1. Nullhypothese

Eine stetige Kovariable x_1 mit $x_1 \sim N(0,1)$,
$\beta_0=0$, $\beta_1=0{,}693$,
M=500, 1000 Wiederholungen, $\alpha=0{,}05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1\text{-}10$	$M_i \equiv 10$
X^2	0,000	0,010	0,002	0,046
D	1,000	0,977	0,585	0,114
$X^2_{\boldsymbol{O}}$	0,061	0,043	0,040	0,041
$X^2_{\boldsymbol{M}}$	0,063	0,052	0,045	0,052
HL-Test	0,055	0,051	0,054	0,052
$X^2_{\boldsymbol{F}}$	0,000	0,051	0,055	0,062
IM-Test	0,057	0,049	0,045	0,049
$R_{\boldsymbol{C}}$	0,053	0,052	0,046	0,051

Drei stetige Kovariablen x_i mit x_i iid N(0,1),
$\beta_0=0$, $\beta_1=0.693$, $\beta_2=0.405$, $\beta_3=0.223$
M=500, 1000 Wiederholungen, $\alpha=0{,}05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1\text{-}10$	$M_i \equiv 10$
X^2	0,000	0,001	0,009	0,043
D	1,000	0,959	0,866	0,118
$X^2_{\boldsymbol{O}}$	0,074	0,039	0,042	0,026
$X^2_{\boldsymbol{M}}$	0,078	0,052	0,059	0,057
HL-Test	0,049	0,049	0,052	0,042
$X^2_{\boldsymbol{F}}$	0,000	0,052	0,058	0,058
IM-Test	0,051	0,049	0,058	0,051
$R_{\boldsymbol{C}}$	0,058	0,046	0,048	0,049

6.2. Alternative

Overdispersion

Stetige Kovariable x_1 mit $x_1 \sim U(-6,6)$, $\beta_0=0$, $\beta_1=0.405$,
Fehlspezifikation: β_0 zufällig, $E(\beta_0)=0, Var(\beta_0)=0.323$,
M=500, 1000 Wiederholungen, $\alpha=0,05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1\text{-}10$	$M_i \equiv 10$
$X^2_{\boldsymbol{O}}$	0,045	0,211	0,201	0,645
$X^2_{\boldsymbol{M}}$	0,047	0,230	0,230	0,694
HL-Test	0,046	0,052	0,121	0,231
$X^2_{\boldsymbol{F}}$	0,000	0,232	0,464	0,699
IM-Test	0,043	0,040	0,086	0,123
$R_{\boldsymbol{C}}$	0,045	0,053	0,061	0,107

Fehlspezifizierte Linkfunktion

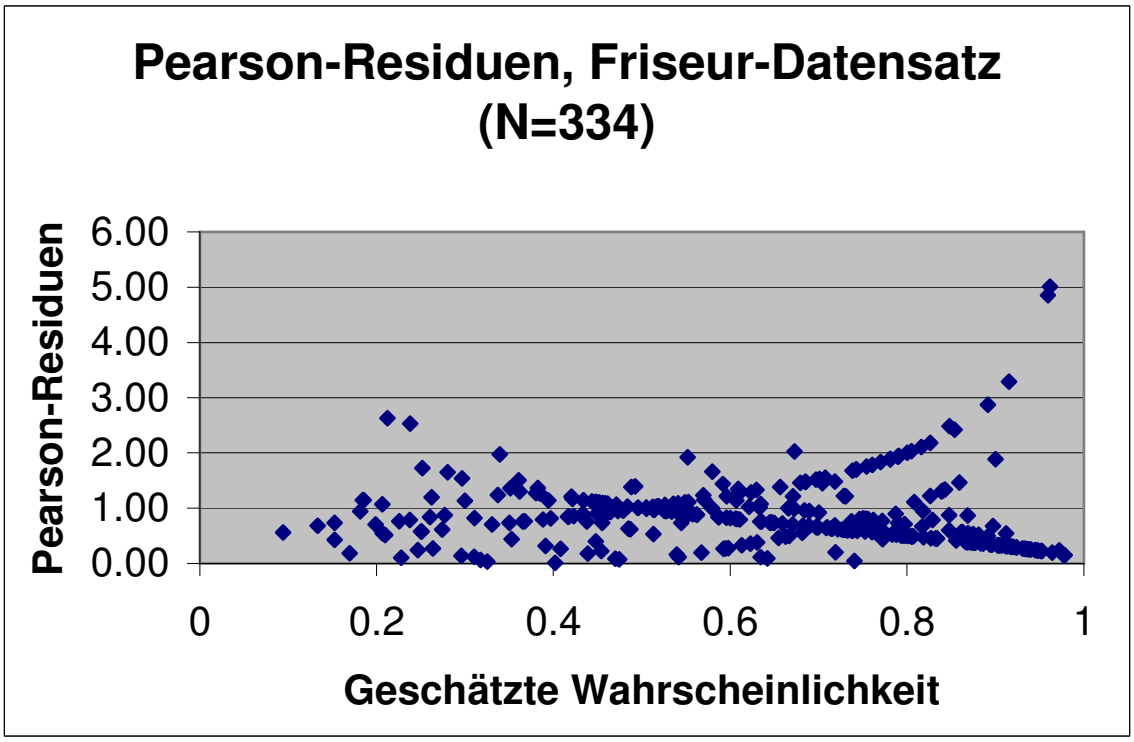
Stetige Kovariable x_1 mit $x_1 \sim U(-6,6)$,
Fehlspezifikation: $\log(-\log(1-\pi_i))=0.405x_1$
M=500, 1000 Wiederholungen, $\alpha=0,05$

	$m_i \equiv 1$	$m_i \equiv 2$	$m_i \equiv 1\text{-}10$	$M_i \equiv 10$
$X^2_{\boldsymbol{O}}$	0,000	0,001	0,000	0,025
$X^2_{\boldsymbol{M}}$	0,000	0,001	0,000	0,037
HL-Test	0,200	0,197	0,204	0,195
$X^2_{\boldsymbol{F}}$	0,000	0,067	0,059	0,126
IM-Test	0,541	0,545	0,527	0,517
$R_{\boldsymbol{C}}$	0,275	0,277	0,289	0,267

Beispiel Friseurstudie:

Vorsicht: Signifikantes Testergebnis bietet keine Hilfe bei der Reformulierung des Modelles

	p-Wert	p*-Wert
X^2	0,053	0,391
D	0,012	0,033
X^2_O	0,044	0,511
X^2_M	0,031	0,458
HL-Test	0,451	0,299
X^2_F	0,408	0,427
IM-Test	0,365	0,873
R_C	0,062	0,734



7.Fazit

- X^2 und D sind als Anpassungstests im logistischen Modell bei fehlenden Messwiederholungen ungeeignet.
- Es gibt Alternativen zu diesen Tests, auch der bisherige Standard (HL-Test) kann noch verbessert werden. Diese Tests sind mit vernünftigem Aufwand zu berechnen.
- Für fehlende Messwiederholungen ($m_i \equiv 1$) und kleine Fallzahlen haben aber auch diese Alternativtests eine niedrige Power

→ Globale Anpassungstests sind ein hilfreiches Werkzeug, aber kein Allheilmittel

Grundsätzliches Dilemma:
Ein Anpassungstest kann nur die Alternative prüfen, ein nicht-signifikanter Test sagt uns nicht, dass ein gutes Modell vorliegt.

Software:
SAS/IML-Makro %GOFLOGIT
Erhältlich unter Oliver.Kuss@medizin.uni-halle.de

8. Literatur

- Agresti A. *Categorical data analysis.* John Wiley & Sons, 1990.
- Bertolini G et al. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *J Epidem Biostat*, 5:251-253, 2000.
- Copas JB. Unweighted Sum of Squares Test for Proportions. *Appl Statist,* 38:71-80, 1989.
- Farrington CP. On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data. *J R Statist Soc B,* 58:349-360, 1996.
- Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Statist - Theor Meth*, 9:1043-1069, 1980.
- Hosmer DW, Lemeshow S. *Applied logistic regression.* John Wiley & Sons, 1989.
- Hosmer DW, Taber S, Lemeshow S. The Importance of Assessing the Fit of Logistic Regression Models: A Case Study. *Am J Public Health,* 81:1630-1635, 1991.
- Hosmer DW et al. A comparison of goodness-of-fit tests for the logistic regression model. *SiM,* 16:965-980, 1997.
- Lloyd CJ. *Statistical Analysis of Categorical Data.* John Wiley & Sons, 1999.
- McCullagh P. On the Asymptotic Distribution of Pearson´s Statistic in Linear Exponential-Family Models. *International Statistical Review*, 53:61-67, 1985.
- McCullagh P, Nelder JA. *Generalized Linear Models.* Chapman & Hall, 1989.
- Osius G, Rojek D. Normal Goodness-of-Fit Tests for Multinomial Models With Large Degrees of Freedom. *JASA,* 87:1145-1152, 1992.
- Orme C. The calculation of the information matrix test for binary data models. *The Manchester School,* 54:370-376, 1988.
- Santner TJ, Duffy DE. *The statistical analysis of discrete data.* Springer, 1989.
- White H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica,* 50:1-25, 1982.

Anpassungstests

- **Spezifisch:** Einbetten des logistischen Modells in eine umfassendere parametrische Familie und Testen des Parameters, der das Standardmodell beschreibt:
  z.B.: Pregibon, 1980

$$g(\pi_i, \lambda) = \log\left(\frac{(1/(1-\pi_i))^{\chi} - 1}{\chi}\right)$$

  Test auf $\lambda=1$ testet das Standardmodell.

- **Global:** Testen einer unspezifischen Nullhypothese „Das Modell ist gut angepasst“

Pregibon D. Goodness of link tests for generalized linear models. *Applied Statistics*, 29:15-24, 1980.