Ein semi-bayes'scher Anpassungstest für das logistische Regressionsmodell mit schwach besetzten Zellen

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 schwach besetzten Zellen
4. Lösungsvorschläge
5. Ein semi-bayes'scher Lösungsvorschlag
6. Fazit
7. 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
Modellwahl und **Modellüberprüfung**

Modellwahl:
Wie beschreibe ich den mittleren Wert der Zielgröße??
Modellüberprüfung:
Inwiefern weicht dieser von den beobachteten Werten ab?

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

Ein Klassiker unter den globalen Anpassungstests:

Pearson-Statistik:

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

Große Werte von X^2 zeigen schlechte Anpassung an

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

3. Das Problem der schwach besetzten Zellen

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

Unrealistisch bei großer Anzahl von Kovariablen oder stetigen Kovariablen

Katastrophal:
Im Extremfall $m_i \equiv 1$ gilt für X^2: $X^2 \approx N$

4. Lösungsvorschläge (Auswahl)

4.1 Modifizierte Prüfverteilung

- Unter n, $m_i \rightarrow \propto$ ist X^2 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)

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

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

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

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

Summation von rohen Pearson-Residuen

5. Ein semi-bayes'scher Lösungsvorschlag

Problem 1:
Lösungen von McCullagh und Osius/Rojek berechnen asymptotische Momente von X^2 und verlassen sich auf die Normalverteilung

Idee 1:
Bayes'sche Schätzung des Modells liefert Posterior-Verteilung für alle Parameter und auch für alle Funktionen der Parameter.

Also: Schätze Model durch MCMC und erhalte dadurch komplette "exakte" Posterior-Verteilung von X^2.

Schön: MH-Algorithmus von Gamerman, 1997, verknüpft geschickt den IRLS-Algorithmus mit dem bayes'schen log. Modell, hat dadurch sehr effiziente Proposal-Dichte

Problem 2:
Was ist der "Nulleffekt" eines gut angepassten Modells??

Idee 2:
Nutze asymptotische Mittelwerte aus der frequentistischen Theorie ("semi-bayes'sch")

Beispiel:

Berufsbedingte Handekzeme bei Auszubildenden im Friseurhandwerk

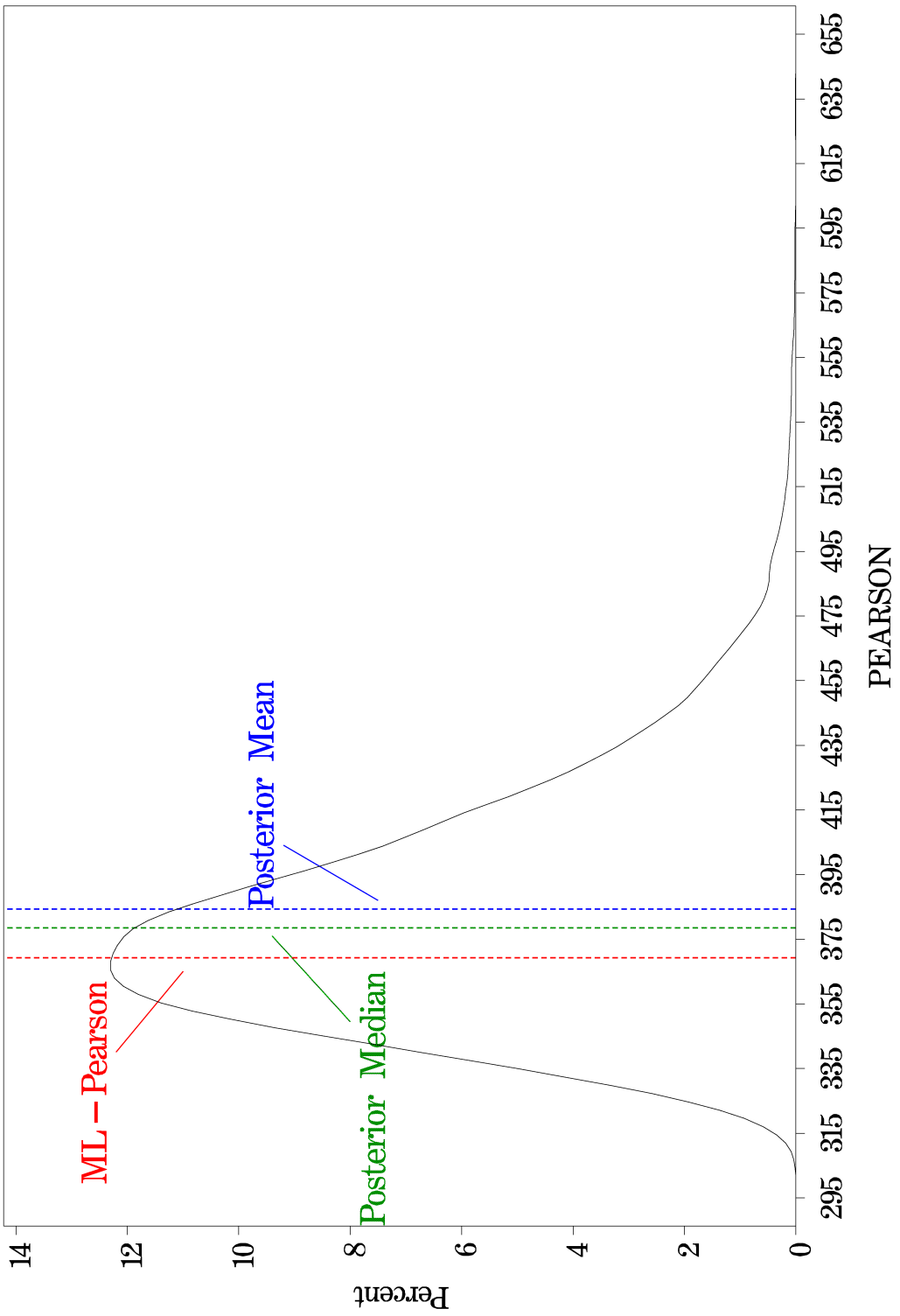
M=574 (340 „Erfolge“),

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%)


Percent
0
2
4
6
8
10
12
14
295
315
335
355
375
395
415
435
455
475
495
515
535
555
575
595
615
635
655
PEARSON
ML – Pearson
Posterior Median
Posterior Mean

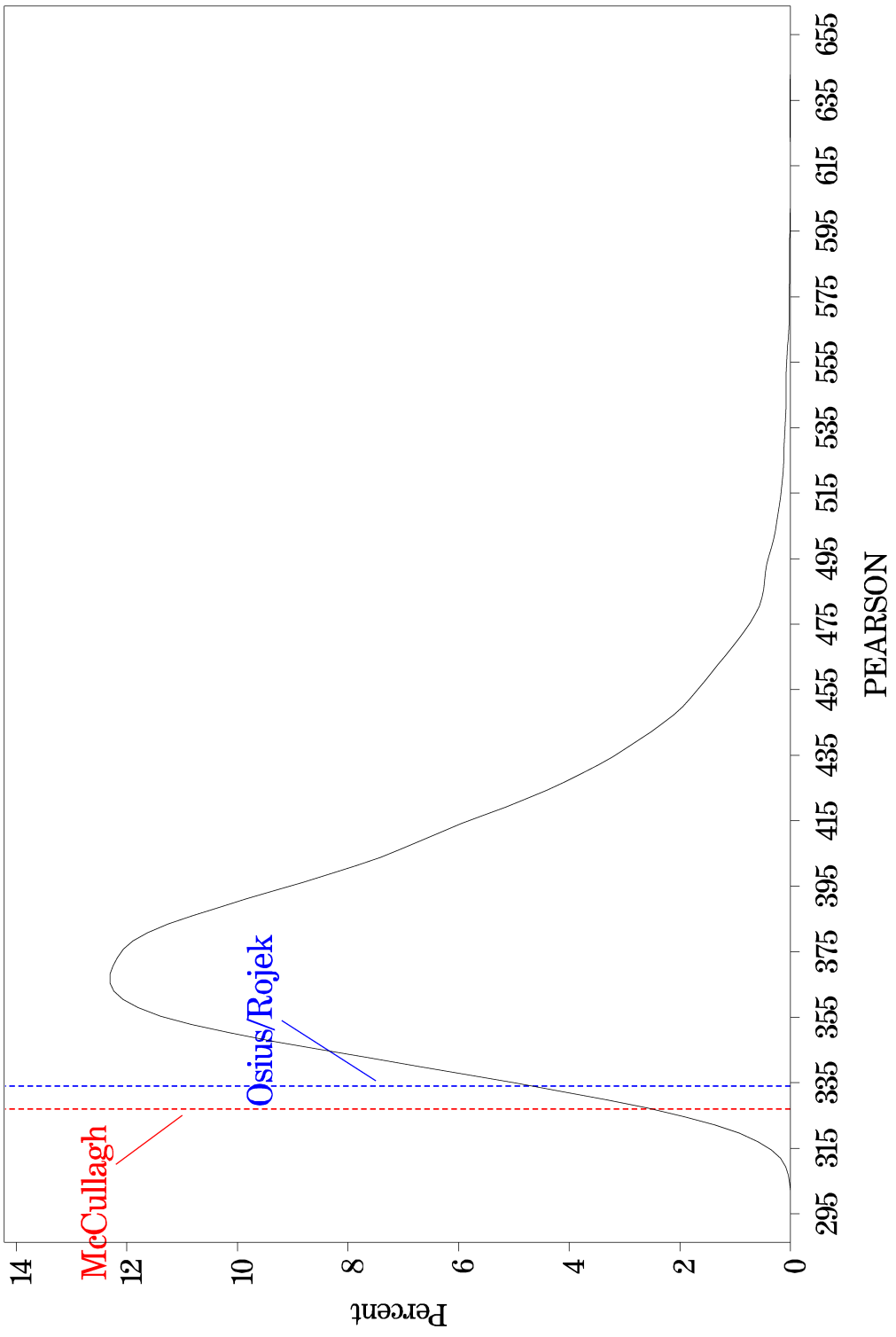
Überprüfung der Modellanpassung:

Werte der Teststatistiken:

$$X^2 = 369.25$$
$$X^2_{MH} \text{ (mean)} = 384.37$$
$$X^2_{MH} \text{ (median)} = 378.56$$

p-Werte:

Test	p-Wert
X^2	0,053
X^2_O	0,044
X^2_M	0,031
X^2_{MHO}	0,036
X^2_{MHM}	0,012


Percent
0
2
4
6
8
10
12
14
McCullagh
Osius/Rojek
295
315
335
355
375
395
415
435
455
475
495
515
535
555
575
595
615
635
655
PEARSON

6. Fazit

- Resultate des semi-bayes'schen Tests liegen in ähnlichem Bereich, evtl. sogar in besserer Richtung
- Es gibt noch viel zu tun:
  - Eigenschaften des MCMC-Algorithmus (Konvergenz, Autokorrelation)
  - ML-Pearson = Posterior Mode ?
  - Eigenschaften des Tests
    → Simulationsuntersuchungen

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

7. Literatur

- 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.
- Gamerman D. Efficient sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7: 57-68, 1997.
- Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Statist - Theor Meth*, 9:1043-1069, 1980.
- Hosmer DW et al. A comparison of goodness-of-fit tests for the logistic regression model. *SiM*, 16:965-980, 1997.
- McCullagh P. On the Asymptotic Distribution of Pearson´s Statistic in Linear Exponential-Family Models. *International Statistical Review*, 53:61-67, 1985.
- 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.
- White H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50:1-25, 1982.