A semi-bayesian Goodness-of-Fit Tests in Logistic Regression with Sparse Data

Oliver Kuss
Institute of Medical Epidemiology, Biostatistics and Informatics, Medical Faculty University of Halle-Wittenberg,
Oliver.Kuss@medizin.uni-halle.de

Programme:
1. The Logistic Regression Model
2. Checking Goodness-of-Fit
3. The Problem of Sparse Data
4. Some Solutions
5. A semi-bayesian solution
6. Discussion
7. Literature
1. The Logistic Regression Model

Standard method for the regression of binary responses

**Reasons:**
- Easy interpretation of parameters as odds-ratios
- One can predict response probabilities
- Software is available
- Valid in prospective and retrospective designs

**Notation:**
N independent observations grouped by covariate patterns \((y_i, x_i), i=1, ..., N\)

- \(x_i\) : vector of \(p+1\) covariates,
- \(y_i\) : number of successes, \(Y_i \sim B(m_i, \pi_i)\),
- \(m_i\) : number of trials,

\[M=\sum_{i=1}^{N} m_i\] : Number of individual observations
Data:

<table>
<thead>
<tr>
<th>Response</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1$</td>
<td>$m_1-Y_1$</td>
</tr>
<tr>
<td>Covariate 2</td>
<td>$Y_2$</td>
<td>$m_2-Y_2$</td>
</tr>
<tr>
<td>Pattern</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>N</td>
<td>$Y_N$</td>
<td>$m_N-Y_N$</td>
</tr>
</tbody>
</table>

Example:

Continuous covariate(s): \( N=M \) (\( m_i \equiv 1 \))

<table>
<thead>
<tr>
<th>Response</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Covariate 2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pattern</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Model equation:

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

with $\beta_j = (\beta_0, \ldots, \beta_p)$ vector of regression parameters.

Estimate parameters $\beta_j$ via ML.
2. Checking Goodness-of-Fit

Statistical modelling consists of two steps (Hosmer et al., 1991):

**Model building** and **Model checking**

**Model building**: How can I explain the variation in response values in terms of the covariates (*systematical component*)

**Model checking**: Assess all variation that is not explained by the systematical component by comparing observations and prognoses from the model (*error component*)

**Two different kinds of model checking:**

1) Consider deviations from prognoses for each single observations (numerically and graphically): ⇒ **Residual analysis**
2) Calculate goodness-of-fit measures and assess model fitting by a single number and a statistical test: ⇒ **Goodness-of-fit tests**
A classical Goodness-of-Fit test:

**Pearson statistic:**

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

Large values of \( X^2 \) indicate lack-of-fit

**Statistical test:** Compare \( X^2 \) to a \( \chi^2 \)-distribution with \( N-p-1 \) df
3. The Problem of Sparse Data

The $\chi^2$-limiting distribution is only valid for large $m_i$ (N fixed, $m_i \to \infty$ for all i)

Unrealistic with a large number of covariates or with continuous covariates

A disaster:
In the case of $m_i \equiv 1$ there is: $X^2 \approx N$
4. Some Solutions

4.1 Modify limiting distribution
- $X^2$, $D$ are asymptotically normal under $N$, $m_i \to \infty$ (Osius/Rojek, 1992; McCullagh, 1986)

4.2 Grouping observations
- Hosmer-Lemeshow test (Hosmer/Lemeshow, 1980)
  Maybe the standard test with sparse data nowadays, but it has some deficiencies (Hosmer et al, 1997, Bertolini et al., 2000)

4.3 Use other tests statistics
- $X^2_F$ (Farrington, 1996)
  \[
  X^2_F = X^2 + \sum_{i=1}^{N} \frac{1}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)} (y_i - m_i \hat{\pi}_i)
  \]
5. A semi-bayesian solution

Example:
Occupational hand eczema in hairdresser apprentices,
M=574 (340 „successes“),
Several covariates (p=6): genetical disposition, work hazards, confounders,
N=334,
Distribution of the m_i:

<table>
<thead>
<tr>
<th>m_i</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>205 (61%)</td>
</tr>
<tr>
<td>2</td>
<td>68 (20%)</td>
</tr>
<tr>
<td>3</td>
<td>35 (11%)</td>
</tr>
<tr>
<td>&gt;3</td>
<td>26 (8%)</td>
</tr>
</tbody>
</table>
Step 1:

Bayesian estimation of the model gives posterior distribution for all parameters, for all functionals of them, and thus also for the Pearson statistic (see Jackman, 2000)

Thus: Estimate model by MCMC (Gibbs-Sampling, WinBUGS) und get "exact" posterior distribution of $X^2$
Step 2:

The mode of the posterior gives a new value of the Pearson statistic ($X^2_{MCMC}$) and by the 2.5%- and 97.5%-percentile a corresponding 95%-credibility interval.
Step 3:

Interpret the 95%-credibility interval as a frequentist 95%-confidence interval (Mossman/Berger, 2001; Carlin/Louis, 2000)
Step 4:

Identify the 95%-confidence interval as acceptance region of a frequentist statistical test
95% - Annahmebereich
Step 5:

Which is the value that indicates the null hypotheses ("The model fits good")?

Use asymptotic expectations from standard theory:

\[
E(X^2_O) = N \quad (=334; \text{Osius/Rojek, 1992})
\]
\[
E(X^2_M) = N-p-1 \quad (=327; \text{McCullagh, 1986})
\]
Summary:

Value of test statistics:

\[ X^2 = 369.25 \]
\[ X^2_{\text{MCMC}} \text{ (Posterior Mode)} = 371.30 \]
\[ 2.5\%-\text{percentile} = 335.90 \]
\[ 97.5\%-\text{percentile} = 493.00 \]

p-values:

\[ X^2 \quad 0.053 \]
\[ X^2_O \quad 0.044 \]
\[ X^2_M \quad 0.031 \]
\[ X^2_{\text{MCMCO}} \quad 0.038 \]
\[ X^2_{\text{MCMCM}} \quad 0.012 \]
6. Discussion

- The results of the semi-bayesian test seem sensible, maybe better than the standard Pearson test (Kuss, 2002)

- To do:
  Check the MCMC-algorithm (convergence, autocorrelation)
  Check the principle→ simulation

- The semi-bayesian idea seems to be applicable to other measures in logistic regression ($R^2$, c, Somer's D etc.) or even to other statistical models
7. Literature