GLOBAL GOODNESS-OF-FIT TESTS IN LOGISTIC REGRESSION WITH SPARSE DATA

Oliver Kuss

Institute of Medical Epidemiology, Biostatistics, and Informatics,
University of Halle-Wittenberg,
Magdeburger Str. 27, 06097 Halle/Saale, Germany
Phone: +49-345-5573582, E-mail: Oliver.Kuss@medizin.uni-halle.de

1. The Logistic Regression Model

The logistic regression regression model is the standard tool for regression analysis with binary responses. This has many reasons of which the most prominent are that parameters can be interpreted as odds ratios, that prognoses for the event of interest are possible, and the availability of standard software.

Notation:
Consider N independent **grouped** observations (y_i, x_i), i=1, ..., N which are **grouped by pattern of covariate values** with
x_i : vector of p+1 covariates,
y_i : number of successes, realisation of $Y_i \sim B(m_i, \pi_i)$,
m_i: number of trials in the i-th covariate pattern,

M= $\sum_{i=1}^{N} m_i$ number of **individual** observations

Model equation:

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

2. Assessing Goodness-of-Fit (GOF)

Classical goodness-of-fit measures are the Pearson statistic

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

and the deviance

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

which compare observed and expected values from the model and indicate lack-of-fit by large values. A statistical test can be calculated by comparing these statistics to a χ^2_{N-p-1}-distribution.

However, the validity of the χ^2-distribution relies on the assumption of large m_i!!!

This is unrealistic with a large number of covariates or continuous covariates, more the rule than the exception in today´s data sets.
In the extreme case of m_i=1 (M=N), the deviance becomes[1]

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

is independent of the y_i, and thus contains no information on the model fit.
The Pearson statistic equals $X^2 \approx N$ in this case[1], the sample size being also a nonsense measure of fit.

3. Alternative Tests

Hosmer-Lemeshow-Test[2]
This test introduces a new grouping of the individual observations to avoid sparseness where the grouping depends on the estimated probabilities from the model. It became a new quasi-standard, although it has some deficiencies. For example, it might depend heavily on the number of new groups and the calculating algorithm[3], and even on the ordering of observations[4]. Hosmer et al.[3] reported results from fitting the same data set in six major statistical software packages, obtaining six different p-values for the Hosmer-Lemeshow test ranging from 0.02 to 0.16!

Modification of distribution
Osius/Rojek[5] (X^2_O) and McCullagh[6] (X^2_{McC}) showed that X^2 and D are normally distributed under sparseness assumptions (N, $m_i \to \infty$).

Use other test statistics
Farrington[7] calculated approximate moments for

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

and showed how to calculate a statistical test. Unfortunately, for m_i=1 there is $X^2_F \equiv N$.

The information matrix test[8] compares two different estimators of the information matrix which should yield comparable results under a satisfactory model fit. Evaluating the difference between the diagonal elements of these two estimators results in the ((p+1)x1)-vector

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

with $z_i = (1, x_{i1}^2, \ldots, x_{ip}^2)^t$, where the summation is now over the individual observations. Standardization with an appropriate variance estimator leads to the test statistic IM_{DIAG} which can be compared to a χ^2_{p+1}-distribution.

Hosmer et al.[3] extended the RSS-test of Copas[9]

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

to the logistic regression case and showed how to calculate asymptotic moments and a statistical test. Note that this test uses the numerator terms of X^2, also summing over the individual observations.

4. Which Test Works???

Up to now, there has only been one single systematic investigation of GOF tests in logistic regression[3]. In an own simulation study we considered in more depth the behaviour of the tests under varying degrees of sparseness and added the described tests which have not yet been fully investigated. Below some results of this study are given.

Table 1. Empirical level (in %) under the null hypothesis for X^2 and D for various m_i, various model specifications, M=500, α=5%, 1000 replications.

	Constellation of m_i			
	1	2	5	10
	Model: logit(π_i) = 0			
X^2	0.00	1.15	3.53	3.93
D	100.00	99.69	30.64	9.41
	Model: logit(π_i) = 0.693x_i, $x_i \sim$ N(0,1)			
X^2	0.00	1.02	3.43	4.57
D	100.00	97.96	33.91	11.42
	Model: logit(π_i) = 0.223x_{1i}+0.405x_{2i}+0.693x_3+, x_{ji} iid N(0,1)			
X^2	0.00	1.41	4.01	4.31
D	100.00	95.89	32.86	11.79

This shows that X^2 and D are no valid goodness-of-fit tests in logistic regression with sparse data.

Table 2. Empirical level (in %) under the alternative hypothesis of a misspecified model for the Hosmer-Lemeshow test (HL), X^2_O, X^2_{McC}, X^2_F, IM_{DIAG} and RSS for various m_i, various model specifications, M=500, α=5%, 1000 replications. The fitted model in all cases is a logistic regression model with logit(π_i)=β_0+$\beta_1 x_{1i}$.

	Constellation of m_i			
	1	2	5	10
Missing covariate, Model: logit(π_i)=0.405x_{1i}+0.223x_3, x_{ji} iid U(-6,6)				
HL	5.6	8.0	18.9	38.7
X^2_O	3.8	37.6	80.5	94.6
X^2_{McC}	4.2	40.2	83.8	95.9
X^2_F	0.0	41.7	85.0	95.9
IM_{DIAG}	5.8	6.6	9.7	17.3
RSS	4.8	5.8	8.0	13.5
Overdispersion,				
Model: logit(π_i)=β_0+0.405x_{1i}, $x_{1i} \sim$ U(-6,6), E(β_0)=0, Var(β_0)=0.323				
HL	4.6	5.2	11.1	23.1
X^2_O	4.5	21.1	46.9	64.5
X^2_{McC}	4.7	23.0	52.4	69.4
X^2_F	0.0	23.2	52.1	69.9
IM_{DIAG}	4.3	4.0	7.9	12.3
RSS	4.5	5.3	5.7	10.7
Misspecified link function, Model: log[log(1- π_i)] = 0.405x_{1i}, $x_{1i} \sim$ U(-6,6)				
HL	20.0	19.7	20.1	19.5
X^2_O	0.0	0.1	1.3	2.5
X^2_{McC}	0.0	0.1	1.8	3.7
X^2_F	0.0	0.5	10.6	12.6
IM_{DIAG}	54.1	54.5	55.0	51.7
RSS	27.5	27.7	28.1	26.7

The power increases with increasing m_i where X^2_O, X^2_{McC} and X^2_F on one hand, and IM_{DIAG} and RSS on the other hand behave very similarly. But all tests have low power for detecting lack-of-fit with small m_i.

5. An Example

574 (=M) hairdressers were followed for the occurence of occupational hand eczema. 340 events were observed, 6 covariates (endogenous and exogeneous risk factors) defined 334 (=N) different patterns of covariates where 205 (61%) were occupied once, and 68 (20%) twice ⇒ VERY SPARSE DATA!!!
Table 3 gives the results of assessing goodness-of-fit for the full data (p) and after removal of two outliers with large Pearson residuals (p*).

Table 3. GOF (p-values) for hairdresser data

	p	p*
X^2	0.053	0.391
D	0.012	0.033
X^2_O	0.044	0.511
X^2_{McC}	0.031	0.458
HL	0.451	0.299
X^2_F	0.408	0.421
IM_{DIAG}	0.365	0.873
RSS	0.062	0.734

Note the strange behaviour of the HL test which indicates worse fit after removal of two definite outliers!!!

6. Conclusions

- The classical goodness-of-fit tests in logistic regression (X^2 and D) are not valid with sparse data.
- Valid alternative tests exist but lack power in extremely sparse data.
- The fundamenal dilemma remains: A non-significant GOF test doesn´t mean that your model is correct.
- A SAS/IML® macro that calculates the described tests is available from the author[10].

7. References

1. McCullagh, P. and Nelder, J.A. (1986), *Generalized Linear Models*, Chapman & Hall.
2. Hosmer, D.W. and Lemeshow, S. (1980), "Goodness of fit tests for the multiple logistic regression model," *Communications in Statistics – Theory and Methods*, 9, 1043-1069.
3. Hosmer, D.W., Hosmer, T., Le Cessie, S. and Lemeshow, S. (1997), "A comparison of goodness-of-fit tests for the logistic regression model," *Statistics in Medicine*, 16, 965-980.
4. Bertolini, G., D'Amico, R., Nardi, D., Tinazzi, A., Apolone, G. (2000), "One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model," *Journal of Epidemiology and Biostatistics*, *5*, 251-253.
5. Osius, G. and Rojek, D. (1992), "Normal Goodness-of-Fit Tests for Multinomial Models With Large Degrees of Freedom," *Journal of the American Statistical Association*, 87, 1145-1152.
6. McCullagh, P. (1985), "On the Asymptotic Distribution of Pearson´s Statistic in Linear Exponential-Family Models," *International Statistical Review*, 53, 61-67.
7. Farrington, C.P. (1996), "On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data," *Journal of the Royal Statistical Society, B*, 58, 349-360.
8. Orme, C. (1988), "The calculation of the information matrix test for binary data models," *The Manchester School*, 54, 370-376.
9. Copas, J.B. (1989), "Unweighted Sum of Squares Test for Proportions," *Applied Statistics*, 38, 71-80.
10. Kuss, O. (2001), "A SAS/IML® Macro for Goodness-of-Fit Testing in Logistic Regression Models with Sparse Data," Proceedings of the 26th Annual SAS® Users Group International Conference, Paper 265-26.