Paper 265-26

A SAS/IML® Macro for Goodness-of-Fit Testing in Logistic Regression Models with Sparse Data

Oliver Kuss, Institute of Medical Epidemiology, Biometry and Informatics,
Halle/Saale, Germany

ABSTRACT

The logistic regression model has become the standard analyzing tool for binary responses in a variety of disciplines. Methods for assessing goodness-of-fit, however, are less developed and this is especially pronounced in calculating goodness-of-fit tests with sparse data, when the standard tests (deviance and Pearson test) behave unsatisfactorily.

In our paper we show two solutions to the problem that are implemented in the LOGISTIC procedure in SAS® software, and introduce five additional testing procedures from the statistical literature. By means of a simulation study we show that these additional tests are valid instruments for assessing goodness-of-fit in logistic regression models, even with sparse data. Finally, we present the SAS/IML macro %GOFLOGIT which allows calculation of the introduced tests and illustrate the macro with an example from occupational epidemiology on hand eczema in hairdressers.

INTRODUCTION

The logistic regression model has become the standard analyzing tool for binary responses in a variety of disciplines. This has many reasons: ease of interpretation of parameters as adjusted odds ratios, possibility of calculating prognoses for the event of interest, and availability of standard software. The LOGISTIC procedure is the standard tool in SAS software for fitting logistic regression models, but solutions with the GENMOD, the PROBIT or the CATMOD procedure are also possible.

Methods for assessing goodness-of-fit, however, are less developed, which may be due to the relative youth and the enhanced mathematical complexity of the logistic regression model, compared to, for example, the linear regression model.

In principle, there are two different approaches to assessing goodness-of-fit in logistic regression models. The first, known as residual analysis, investigates the model on the level of individual observations and looks for those observations which are not adequately described by the model or which are highly influential on the model fit. Among the different SAS procedures for logistic regression PROC LOGISTIC offers the most extensive possibilities for residual analysis: the INFLUENCE option in the MODEL statement supplies a number of influence and outlier diagnostics, and the IPLOTS option provides the corresponding plots.

The second approach to goodness-of-fit on which we will focus seeks to combine the information on the amount of lack-of-fit in a single number. Statistical tests, so called goodness-of-fit-tests, are then performed to judge if the observed lack-of-fit is statistically significant or due to random chance. There are two standard procedures, the deviance and the Pearson test, and these are routinely provided by PROC GENMOD and PROC PROBIT and optionally by PROC LOGISTIC (use the SCALE=none option in the MODEL statement).

These tests, however, have a serious problem with sparse data, where “sparse data” means, that for every pattern of covariate values we have only a small number of observations. In general, this would be the case if there are continuous or many covariates. In extreme cases each individual observation has its own risk profile or pattern of covariates. In this case of sparseness, which in our view is more the rule than the exception in today’s data sets, the deviance and the Pearson test no longer have a chi-square distribution under the null hypothesis and so no longer are valid measures of model fit. Note that this is just an extension of the familiar problem of small cell counts in contingency tables.

In the following, we state the problem with some more mathematical rigor, show two possibilities to circumvent the problem with PROC LOGISTIC, give five additional testing procedures from the statistical literature and present some results from a simulation study which demonstrate that these additional tests are valid goodness-of-fit tests for logistic regression models, even with sparse data. Finally, we present the SAS/IML macro %GOFLOGIT that allows the calculation of this new procedures and illustrate the macro with an example from occupational epidemiology on hand eczema in hairdressers.

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

THE MODEL

Let y_i be the response with $y_i \sim \text{binomial}(m_i, \pi_i)$. The model equation is $\text{logit}(\pi_i)=x_i\beta$, $i=1,\ldots,N$, where $\beta=(\beta_0,\ldots,\beta_p)'$ is a vector of regression parameters corresponding to a vector of p+1 covariates $x_i=(1, x_{i1},\ldots,x_{ip})$. Estimates of the β_j are usually calculated by maximum likelihood and we get estimates of the π_i by plugging the $\hat{\beta}_j$ into the model equation.

Note that we consider *grouped* observations where two individual observations with the same covariate pattern belong to the same group. Translated into PROC LOGISTIC language this means that we consider the model to be specified in the `events/trials` syntax where `events` counts the number of events (y_i) and `trials` counts the number of individual observations (m_i) in a specific covariate pattern.

STANDARD GOODNESS-OF-FIT TESTS

To assess goodness-of-fit in logistic regression one in general calculates the Pearson statistic

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

or the deviance

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

Both rely on the principle of comparing observed (y_i) to predicted ($m_i\hat{\pi}_i$) values and should be large if the model does not fit the data well. To judge statistical significance they are usually compared to a χ^2_{N-p-1}-distribution. The validity of this distribution, however, relies on the assumption of large m_i, and both tests show unsatisfactory behaviour with sparse data, that is, small m_i. It can be shown (McCullagh and Nelder, 1986) that D degenerates to

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

in the case of extreme sparseness ($m_i \equiv 1$) when it is completely independent from the observations and contains no information about the model fit. The Pearson statistic does not perform that much better in this situation, it can be shown (McCullagh and Nelder, 1986) that $X^2 \approx N$, the sample size also not being a reasonable measure of model fit.

SOLUTIONS IN PROC LOGISTIC

PROC LOGISTIC offers two solutions for a reliable assessment of model fit even in the situation of sparse data.

First, there is an alternative goodness-of-fit test, the so called Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980) which was introduced in Version 6.07 and is invoked by the LACKFIT option in the MODEL statement. It relies on a new grouping of the individual observations in approximately ten groups with roughly the same size where the grouping depends on the percentiles of the estimated probabilities ($\hat{\pi}_i$) from the model. Observed and expected numbers of events are determined for each of the new groups, and their discrepancies are summed. Lack-of-Fit is judged by comparing this sum, which is, after standardization, a Pearson statistic from a 2 x g-table with g being the number of new groups, to a χ^2_{g-2}-distribution. Normally, we would expect a χ^2_{g-1}-distribution to judge statistical significance, but this loss of one degree of freedom accounts for the fact that the new grouping depends on estimated parameters and was not fixed in advance.

This test has some deficiencies, however. It was shown that the value of the test statistic might depend on the number of new groups and on the calculating algorithm. Hosmer et al. (1997) reported results from fitting the same data set in six major statistical software packages (including SAS software), obtaining identical values for the estimated parameters but six different values for the p-value of the Hosmer-Lemeshow test ranging from 0.02 to 0.16! A further disadvantage is that observations belonging to same new group might differ considerably in their covariate values and so we get no information where there is lack-of-fit in the space of covariates. Finally, people being more mathematically rigorous might be disturbed by the fact that the distribution for calculating p-values for the Hosmer-Lemeshow test is simulation-based and not derived exactly.

The second possibility for a valid assessment of goodness-of-fit in PROC LOGISTIC is offered through the AGGREGATE= (*variable-list*) option. It enables specifying the groups for which the Pearson Test and the deviance are calculated by yourself where observations with identical values in the given list of variables are regarded as coming from the same group. As Pearson Test and deviance are calculated only when the SCALE= option is specified, the AGGREGATE-option has no effect if the SCALE= option is not specified. By the aid of these options it is in your own responsibility to give cell counts large enough to make the usual chi-square distribution valid, but we might feel a bit uneasy with the arbitraryness of this method which is open for manipulation.

ALTERNATIVE TESTS

However, these two described solutions are not the only alternatives, and the statistical literature offers a variety of additional testing procedures which do not rely on the assumption of large cell counts.

Osius and Rojek (1992) proposed to use the usual X^2 as the test statistic and derived asymptotic moments for it under the assumption of fixed m_i, which need not be large for the result to hold. A statistical test can be calculated by standardizing X^2 with these moments and comparing the resulting test statistic to the standard normal distribution.

McCullagh (1985) made a similar proposal and relaxed the assumption of large m_i, but he argued to use *conditional* asymptotic moments for X^2 given the parameter estimates $\hat{\beta}$. His test statistic is also compared, after standardization with the conditional asymptotic moments, to a standard normal distribution.

Farrington (1996) investigated a family of generalized Pearson statistics which extend X^2 by an additive constant. He showed that the member

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

has optimal properties regarding variance and local orthogonality. Approximate moments for the Farrington statistic can be calculated in closed form, and the standardized statistic can be compared to the standard normal distribution. Unfortunately, the Farrington statistic has a structural deficiency: in the case of extreme sparseness ($m_i \equiv 1$) there is $X_F^2 = N$ and the test will never reject the null hypothesis of a good fit.

The information matrix test (IM-test), originally proposed by White (1982), relies on the principle of comparing two different estimators of the information matrix (which is the inverse of the covariance matrix of the parameter estimates) that should give comparable results under a satisfactory model fit.

Hosmer/Lemeshow (1989) termed the IM-test as being “elegant, but difficult to compute in praxis”, but Orme (1988) showed how to calculate this test for logistic regression models. Evaluating the difference of the diagonal elements of the two estimators results in the ((p+1) x 1)-vector

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

with $M = \Sigma m_i$, the number of individual observations, and $z_j = (1, x_{i1}^2, \ldots, x_{ip}^2)'$. In the case of a good model fit the components of d sum up to 0. Standardization of this vector with an appropriate variance estimator provides the test statistic (IM_{DIAG}) as a single number which should be compared to a χ^2_{p+1}-distribution. Note that the IM-test is calculated for the individual (summation over M) and not for the grouped observations and so we do not expect problems in sparse data.

The RSS-(Residual Sum of Squares)-test by Copas (1989) only considers the numerator of the X^2, where the summation is again over the individual obervations:

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

Hosmer et al. (1997) show how to calculate asymptotical moments of RSS and to perform a statistical test.

COMPARISON OF THE VARIOUS GOODNESS-OF-FIT TESTS

Up to now, there has only been one single systematic investigation of goodness-of-fit tests in logistic regression which is due to Hosmer et al. (1997). Our work can be seen as a kind of supplement to this study where we considered in more depth the behaviour of the tests under varying degree of sparseness and added some tests which have not been investigated by Hosmer et al.

In the following we give some results of our simulation study. We report the empirical levels of the tests under various situations concerning null and alternative hypothesis and under differing m_i. Under the null hypothesis empirical levels are only shown for the standard tests X^2 and D (see table 1) because the other tests kept to the prespecified level of 5% in most of the cases.

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%, 10000 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~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_{3i}+, 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

We find a dramatic anti-conservatism of D for all m_i<5 where the situation gets worse when the m_i get smaller. The Pearson test is too conservative in situations with m_i<2 but behaves satisfactorily with m_i>5. These simulation results show (and thus confirm existing knowledge) that X^2 and D are no valid goodness-of-fit tests in logistic regression with sparse data where the deviance D shows by far the more erratic behaviour.

In Table 2 we report some results for the power of the various goodness-of-fit tests from our simulation under three different alternative hypotheses (missing covariate (a), overdispersion (b), and misspecified link function (c)). This time we consider the newly introduced alternative tests and the Hosmer-Lemeshow test because they kept the prespecified level under the null hypothesis, as opposed to the standard tests X^2 and D.

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 models is a logistic regression model with an intercept and one continuous covariate.

	Constellation of m_i			
	1	2	5	10
	Model (a): logit(π_i)=0.405x_{1i}+0.223x_{2i}, 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
	Model (b): logit(π_i)=β_0+0.405x_{1i}, x_{1i} ~ 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
	Model (c): log[log(1-π_i)]=0.405x_{1i}, x_{1i} ~ 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	6.7	10.6	12.6
IM_{DIAG}	54.1	54.5	55.0	51.7
RSS	27.5	27.7	28.1	26.7

Several points can be made:

- X^2_O, X^2_{McC} and X^2_F behave very similarly where X^2_{McC} performs better than X^2_O but is outperformed by X^2_F.
- IM_{DIAG} and RSS also behave similarly where IM_{DIAG} outperforms RSS.
- In every situation there is a competing GOF-test that outperforms HL, which is the standard procedure for assessing goodness-of-fit in PROC LOGISTIC.
- In general, all tests gain power with increasing m_i.
- All in all, there is low power for detecting lack-of-fit with small m_i.

%GOFLOGIT: A SAS/IML MACRO FOR ASSESSING GOODNESS-OF-FIT IN LOGISTIC REGRESSION MODELS WITH SPARSE DATA

As the simulation programm which produced the previously reported results was written in SAS/IML, it was not difficult to write a SAS/IML macro that computes the five alternative goodness-of-fit tests which proved to be valid in our simulation. The macro, named %GOFLOGIT, is a kind of stand-alone-application since it also estimates the parameter estimates, their standard errors and estimated probabilities where the IML module from the SAS/IML User's Guide (1988), p. 171-173 is utilized. PROC LOGISTIC can be invoked on demand to get the Hosmer-Lemeshow test which is not supplied by the macro, but this is not necessary for calculating the implemented GOF-tests.

The following command, suitably changed to describe your data, invokes the %GOFLOGIT-macro:

```
%goflogit(data=,y=,m=,xlist=,logistic=,
work=2000,syms=200)
```

where

`data=`	specifies the data set you are using,
`y=`	specifies the variable that contains the number of observed events (y_i) for each covariate pattern,
`m=`	specifies the variable that contains the number of observed individual observations (m_i) for each covariate pattern,
`xlist=`	specifies the list of covariates in the model,
`logistic=`	controls the optional running of PROC LOGISTIC (default: `logistic=on`),
`work=`	specifies the worksize for SAS/IML (default: `work=2000`) and
`syms=`	specifies the size of symbol space for SAS/IML (default: `SYMS=200`).

Note the following remarks:

- The macro does not check for syntactical errors, however, it checks for separation, that is non-existence of the maximum likelihoood estimates (see So, 1993), and issues a warning message if separation is suspected.
- %GOFLOGIT has no CLASS statement, that is, if you have nominal or ordinal covariates you have to recode them as dummies in a previous DATA step.
- %GOFLOGIT expects observations to be grouped by covariate patterns. If your data show extreme sparseness (m_i≡1) and every covariate pattern consists of a single individual observation you should specify a variable which constantly equals 1 in a previous DATA step (for example: `numobs=1;`) and set this variable as the `m=`-variable (`m=numobs`).

AN ILLUSTRATIVE EXAMPLE

We illustrate the %GOFLOGIT macro with an example from a prospective cohort study of 574 hairdresser apprentices conducted in two different German cities. The main aim of the study was to assess exogeneous and endogeneous risk factors for developing hand eczema during the course of the hairdressers' apprenticeship. After one year of follow-up a hand eczema was diagnosed in 340 hairdressers.

Six known or suspected risk factors were included in a logistic regression model to evaluate their influence on developing hand eczema: wet work (</≥ 4h/day), working with permanent wave (</≥ 1h/day), atopic disposition (square root of the continuous atopy-score by Diepgen et al., 1996), diagnosed dyshidrosis (yes/no), center (Erlangen/Dortmund) and change in skin protection behaviour (continuous score ranging from 0 to 5), because it was known from previous studies that this is an important confounder.

Depending on these six covariates, the 574 hairdressers can be divided into 334 different covariate groups with identical values of the covariates within each group. The distribution of the m_i was the following:

m_i	frequency
1	205
2	68
3	35
>3	26

We notice a certain degree of sparseness and decide not to rely on the standard tests X^2 and D in assessing goodness-of-fit. The statement

```
%goflogit(data=hairdresser,y=y,m=m,
xlist=wet_work permanent_wave atopy
dyshidrosis center skin_protection,
logistic=ON);
```

invokes the %GOFLOGIT macro and produces (besides other) the following output where the upper part is from the PROC LOGISTIC run:

```
Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square       DF     Pr > ChiSq

    7.5735        8         0.4762


                          Value      p-Value
Standard Pearson Test    369.246       0.054
Standard Deviance        387.080       0.012
Osius-Test                 1.702       0.044
McCullagh-Test             1.861       0.031
Farrington-Test            0.232       0.408
IM-Test                    6.426       0.491
RSS-Test                 108.862       0.063
```

The goodness-of-fit of the model does not seem to be very high, even the standard Pearson test (X^2) which in the simulation study showed a conservative behaviour indicates some lack-of-fit. A second look at the results reveals that most of the tests which are based on a summation of residuals (X^2, X^2_O, X^2_{McC} and RSS) indicate lack-of-fit, as opposed to the tests that rely on different computing principles (HL and IM_{DIAG}). This arouses the suspicion that some outlying observations are responsible for the bad fit of the model and indeed, a residual analysis identifies two observations which had an estimated probability of 0.96 for developing a hand eczema, but in both cases *no* hand eczema was observed. A reanalysis of the data after removal of these two outliers gave the following results regarding goodness-of-fit:

```
Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square       DF     Pr > ChiSq

    9.2703        8         0.3200


                          Value      p-Value
Standard Pearson Test    331.402       0.391
Standard Deviance        373.570       0.033
Osius-Test                -0.027       0.511
McCullagh-Test             0.106       0.458
Farrington-Test            0.185       0.427
IM-Test                    3.097       0.876
RSS-Test                 106.923       0.734
```

and there no longer is any indication of a bad model fit.

Note, however, the rather bizarre behaviour of the Hosmer-Lemeshow test which adds another point to the list of its peculiarities. The removal of two obviously outlying observations leads to a *worse* model fit, at least according to the Hosmer-Lemeshow test.

CONCLUSION

Assessing goodness-of-fit in logistic regression is an important, but non-trivial task, especially with sparse data. We saw that in certain cases the standard methods offered by SAS software may not suffice and should be completed by some more advanced statistical tests which are not yet implemented in SAS software but can be calculated easily with the %GOFLOGIT macro.

REFERENCES

Copas, J.B. (1989), "Unweighted Sum of Squares Test for Proportions," *Applied Statistics,* 38, 71-80.

Diepgen, T.L., Sauerbrei, W., Fartasch, M. (1996), "Development and validation of diagnostic scores for atopic dermatitis incorporating criteria of data quality and practical usefulness," *Journal of Clinical Epidemiology,* 49, 1031-1038.

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.

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.

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.

Hosmer, D.W., Lemeshow, S. (1989), *Applied logistic regression,* John Wiley & Sons, Inc.

McCullagh, P. (1985), "On the Asymptotic Distribution of Pearson´s Statistic in Linear Exponential-Family Models," *International Statistical Review*, 53, 61-67.

McCullagh, P. and Nelder, J.A. (1986), *Generalized Linear Models*, Chapman & Hall.

Orme, C. (1988), "The calculation of the information matrix test for binary data models," *The Manchester School,* 54, 370-376.

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.
SAS Institute Inc. (1988), *SAS/IML™ Users's Guide, Release 6.03 Edition,* Cary, NC: SAS Institute Inc.
So, Y. (1993), "A Tutorial on Logistic Regression," *Proceedings of the Eighteenth Annual SAS Users Group International Conference,* 18, 1290-1295.
White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica,* 50, 1-25.

CONTACT INFORMATION

The %GOFLOGIT macro is available on request from the author. Contact him at:

Oliver Kuss
Institute of Medical Epidemiology, Biometry and Informatics
06097 Halle/Saale, Germany
Phone: +49-345-5573582
Fax: +49-345-5573580
Email:Oliver.Kuss@medizin.uni-halle.de
WWW: http://imebmi.medizin.uni-halle.de/