SAS GLOBAL FORUM

Measures of Fit for Logistic Regression

Paul D. Allison, Ph.D.
Statistical Horizons LLC

Paper 1485-2014

Introduction

"How do I know if my model is a good model?"

Translation: "How can I convince my boss/reviewer/regulator that this model is OK?"

What statistic can I show them that will justify what I've done?

The ideal would be a single number that indicates that the model is OK if it the number is above or below a certain value.

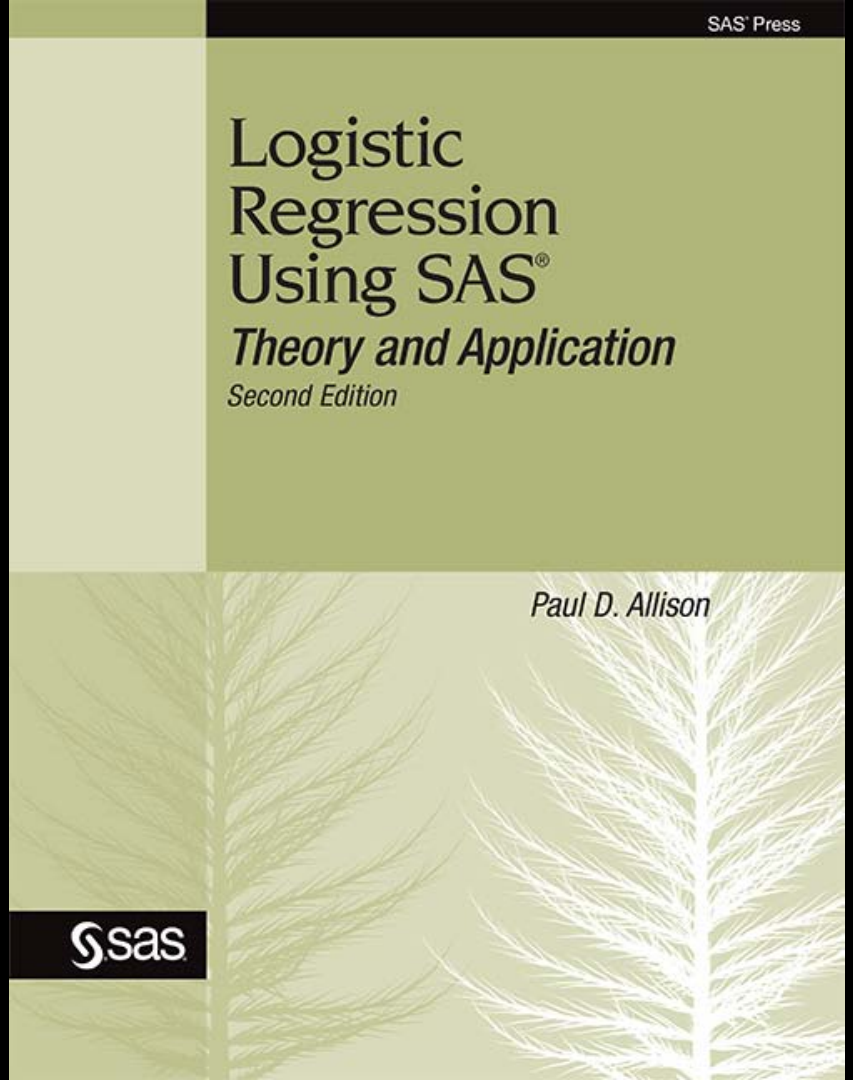
May be asking too much. Usually, you need at least two numbers.

Two classes of fit statistics

1. Measures of predictive power—How well can we explain/predict the dependent variable based on the independent variables.
   - R-square measures
   - Rank-order correlations
   - Area under the ROC curve
2. Goodness-of-fit (GOF) tests
   - Deviance
   - Pearson chi-square
   - Hosmer-Lemeshow.

Predictive power and GOF are very different things

- A model can have very high R-square, yet GOF is terrible.
- Similarly, GOF might be great but R-square is low.


SAS Press
Logistic Regression Using SAS®
Theory and Application
Second Edition
Paul D. Allison
sas

R-square for logistic regression

Many different measures

PROC LOGISTIC: Cox-Snell (regular and “max-rescaled)

PROC QLIM: Cox-Snell, McFadden, 6 others.

Stata: McFadden

SPSS: Cox-Snell for binary, McFadden for multinomial.

I’ve recommended Cox-Snell over McFadden for many years, but recently changed my mind.

Let L_0 be the value of the maximized likelihood for a model with no predictors, and let L_M be the likelihood for the model being estimated.

Cox-Snell: $R^2_{C\&S} = 1 - (L_0 / L_M)^{2/n}$

Rationale: For linear regression, this formula is a identity. A “generalized” R-square.

McFadden vs. Cox-Snell

McFadden: $R^2_{McF} = 1 - \log(L_M)/\log(L_0)$

Rationale: the log-likelihood plays a role similar to residual sum of squares in regression. A “pseudo” R-square.

Problem with Cox-Snell: An upper bound less than 1.

$$Upper\ Bound = 1 - \left[p^p (1-p)^{(1-p)}\right]^2$$

where *p* is the overall proportion of events. The maximum upper bound is .75 when *p*=.5. When *p*=.9 or .1, the upper bound is only .48.

Simple solution: divide Cox-Snell by its upper bound yielding “max-rescaled R-square” (Nagelkerke). But no longer has same appealing rationale. Tends to be higher than most other R-squares.

So, I give the nod to McFadden.

Tjur R^2 (*American Statistician* 2009)

For each category of the response variable, compute the mean of the predicted values. Then take the absolute value of the difference between the two means.

Intuitive appeal, upper bound is 1.0, and closely related to R^2 for linear models.

Example: Mroz (1987) data

```
PROC LOGISTIC DATA = my.mroz DESC;
  MODEL inlf = kidslt6 age educ huswage city exper;
  OUTPUT OUT = a PRED = yhat;
PROC TTEST DATA = a;
  CLASS inlf; VAR yhat; RUN;
```

Output for Tjur R^2

The TTEST Procedure
Variable: yhat (Estimated Probability)

INLF	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	325	0.4212	0.2238	0.0124	0.0160	0.9592
1	426	0.6787	0.2119	0.0103	0.1103	0.9620
Diff (1-2)		-0.2575	0.2171	0.0160		

Compare: Cox-Snell = .25, max re-scaled = .33, McFadden = .21, squared correlation between observed and predicted = .26.

Classic goodness of fit statistics

Classic GOF statistics can be used when cases can be aggregated into "profiles". A profile is a set of cases that have exactly the same values of all predictor variables.

Aggregation is most often possible when predictors are categorical.

Example: In MROZ data, CITY has two values (0,1) and NKIDSLT6 has integer values 0 through 3.

```
PROC LOGISTIC DATA = my.mroz DESC;
  MODEL inlf = kidslt6 city / AGGREGATE SCALE=NONE;
 RUN;
```

AGGREGATE says to group the data into profiles, and SCALE=NONE requests the Pearson and deviance GOF tests.

GOF Output

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	4.1109	5	0.8222	0.5336
Pearson	3.9665	5	0.7933	0.5543

Number of unique profiles: 8

High *p*-values indicate that the model fits well.

Formulas

For each cell in the 8 x 2 contingency table, Let O_j be the observed frequency and let E_j be the expected frequency. Then the deviance is

$$G^2 = 2\sum_j O_j \log\left(\frac{O_j}{E_j}\right)$$

The Pearson chi-square is

$$X^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}$$

If the fitted model is correct, both statistics have approximately a chi-square distribution. DF is number of profiles minus number of estimated parameters.

What are they testing?

Deviance is a likelihood ratio chi-square comparing the fitted model with a "saturated" model, which can be obtained by allowing all possible interactions and non-linearities:

```
PROC LOGISTIC DATA = my.mroz DESC;
  CLASS kidslt6;
  MODEL inlf = kidslt6 city kidslt6*city / AGGREGATE
      SCALE=NONE;
```

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.0000	0	.	.
Pearson	0.0000	0	.	.

What are they NOT testing?

- How well you can predict the dependent variable.
- Whether other predictor variables could improve the model.
- Whether there is unobserved heterogeneity at the individual level.
- If the profiles represent naturally occurring groups (e.g., hospitals, companies, litters), GOF tests can be affected by unobserved heterogeneity produced by *group*-level characteristics.

What if aggregation isn't possible?

Nowadays, most logistic regression models have one more continuous predictors and cannot be aggregated.

Expected values in each cell are too small (between 0 and 1) and the GOF tests don't have a chi-square distribution.

Hosmer & Lemeshow (1980): Group data into 10 approximately equal sized groups, based on *predicted values* from the model. Calculate observed and expected frequencies in the 10 x 2 table, and compare them with Pearson's chi-square (with 8 df).

```
PROC LOGISTIC DATA = my.mroz DESC;
  MODEL inlf = kidslt6 age educ huswage city exper /
    LACKFIT;
```

H-L output

Partition for the Hosmer and Lemeshow Test

Group	Total	INLF = 1		INLF = 0	
		Observed	Expected	Observed	Expected
1	75	14	10.05	61	64.95
2	75	19	19.58	56	55.42
3	75	26	26.77	49	48.23
4	75	24	34.16	51	40.84
5	75	48	41.42	27	33.58
6	75	53	47.32	22	27.68
7	75	49	52.83	26	22.17
8	75	54	58.87	21	16.13
9	75	68	65.05	7	9.95
10	76	71	69.94	5	6.06

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
15.6061	8	0.0484

Problems with Hosmer-Lemeshow

1. Can be highly sensitive to number of groups, which is arbitrary. For the model just fitted we get

   Stata: 10 groups p=.05
   9 groups p=.11
   11 groups p=.64

2. Very common that adding a highly significant interaction or non-linearity to a model makes the HL fit worse. Or adding a non-significant interaction or non-linearity makes the fit better.

3. Some simulation studies show low power.

Many alternative GOF statistics have been proposed (some by Hosmer and Lemeshow).

New GOF tests

New tests fall into two groups

- Those that use alternative methods of grouping. Once the data are grouped, apply Pearson's chi-square.
- Those that do not require grouping.

Focus on ungrouped tests here. Four seem especially promising:

- Standardized Pearson tests
- Unweighted sum of squares
- Information matrix test
- Stukel test

For ungrouped data, you can't create a test based on the deviance–it depends *only* on the fitted values, not the observed values.

Standardized Pearson

When applied to ungrouped data, the Pearson GOF can be written as

$$X^2 = \sum_i \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

where the sum is taken over all individuals, *y* is the observed value of the dependent variable (0 or 1) and π-hat is the predicted value.

This doesn't have a chi-square distribution but it does have a large-sample normal distribution. Use its mean and standard deviation to create a *z*-statistic. At least two ways to get the means and SD:

McCullagh (1985)
Osius and Rojek (1992) These two are usually almost identical.

Unweighted sum of squares

Copas (1989) proposed using

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

This also has a normal distribution in large samples under the null hypothesis that the fitted model is correct.

Hosmer et al. (1997) showed how to get its mean and standard deviation, which can be used to construct a *z*-test.

Information matrix test

White (1982) proposed comparing two different estimates of the covariance matrix of the parameter estimates (the negative inverse of the information matrix), one based on first derivatives of the log-likelihood, the other based on second derivatives.

In this context, we get the following formula

$$IM = \sum_{i=1}^{n} \sum_{j=0}^{p} (y_i - \hat{\pi}_i)(1 - 2\hat{\pi}_i) x_{ij}^2$$

where the *x*'s are the *p* predictors in the model. After standardization with an estimated variance, this has a chi-square distribution with *p*+1 DF.

Stukel test

Stukel (1988) proposed a generalization of the logistic regression model with two additional parameters. These allow for departures from the logit link function at each end of the curve.

The logit model can be tested against this more general model as follows: Let $g_i = \mathbf{x}_i'\mathbf{b}$ where $\mathbf{x}_i$ is the vector of covariate values for individual i and $\mathbf{b}$ is the vector of estimated coefficients. Create two new variables:

$z_a = g^2$ if $g \geq 0$, otherwise $z_a = 0$
$z_b = g^2$ if $g < 0$, otherwise $z_b = 0$.

Add these two variables to the model and test the null hypothesis that both coefficients are equal to 0.

Implementing the Stukel test

```
PROC LOGISTIC DATA=my.mroz DESC;
  MODEL inlf=kidslt6 age educ huswage city exper;
  OUTPUT OUT=a XBETA=xb;
DATA b;
  SET a;
  za=(xb>=0)*xb**2;
  zb=(xb <0)*xb**2;
  num=1; /* for use later */
PROC LOGISTIC DATA=b DESC;
  MODEL inlf = kidslt6 age educ huswage city exper za zb;
  TEST za=0,zb=0;
```

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
Test 1	0.1195	2	0.9420

GOFLOGIT macro for other tests

Macro developed by Oliver Kuss, presented at SUGI 25 (2001)

```
%GOFLOGIT(DATA=b, Y=inlf, XLIST=kidslt6 age educ
   huswage city exper, TRIALS=num)
```

NUM is a “variable” that is always equal to 1, indicating that each data line corresponds to only 1 observation.

Problem with the macro: Gives one-sided *p*-values for standardized Pearson statistics. But theory and simulation evidence indicate that two-sided tests are needed.

Change: posius = 1-probnorm(tosius);

to posius = 2*(1-probnorm(abs(tosius));

Output from GOFLOGIT

TEST	Value	p-Value
Standard Pearson Test	751.049	0.421
Standard Deviance	813.773	0.038
Osius-Test	0.003	0.499
McCullagh-Test	0.029	0.489
Farrington-Test	0.000	1.000
IM-Test	11.338	0.125
RSS-Test	136.935	0.876

Note: Direct testing finds a highly significant effect of experience squared.

Simulation evidence

Several studies report that all these tests have the right “size”: when a correct model is fit with α=.05, they reject the null about 5% of the time.

So the important question is how powerful are these tests at detecting various kinds of departures from the model.

Not satisfied with the available simulation studies so I did my own.

Quadratic:

True model: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x + \beta_2 x^2$

Fitted model: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x$

Quadratic results

Quadratic Effect	Very Low		Low		Medium		High	
N	100	500	100	500	100	500	100	500
Osius	0.068	0.106	0.328	0.840	0.604	0.990	0.832	1.000
McCullagh	0.072	0.108	0.344	0.844	0.616	0.990	0.842	1.000
USS	0.064	0.066	0.348	0.890	0.654	0.994	0.858	1.000
IM	0.048	0.070	0.292	0.872	0.584	0.994	0.826	1.000
Stukel	0.030	0.064	0.192	0.866	0.436	0.992	0.708	1.000
Wald X^2	0.044	0.104	0.380	0.912	0.636	1.000	0.980	1.000

Interaction

Correct model:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd$$

where *d* is a dichotomous variable.

Fitted model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x + \beta_2 d$$

Interaction results

Interaction	Very Low		Low		Medium		High		Very High	
N	100	500	100	500	100	500	100	500	100	500
Osius	0.083	0.086	0.130	0.262	0.211	0.414	0.338	0.570	0.497	0.639
McCullagh	0.093	0.086	0.138	0.264	0.215	0.416	0.348	0.574	0.501	0.639
USS	0.079	0.086	0.130	0.254	0.211	0.406	0.340	0.566	0.499	0.631
IM	0.032	0.054	0.077	0.310	0.168	0.518	0.320	0.658	0.545	0.745
Stukel	0.059	0.214	0.114	0.664	0.274	0.906	0.421	0.952	0.634	0.964
Wald X^2	0.120	0.426	0.342	0.950	0.666	1.000	0.864	1.000	0.966	1.000

Incorrect link function

Correct model:

$$\log(-\log(1-\pi_i)) = \beta_0 + \beta_1 x$$

i.e., complementary log-log model.

Fitted model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x$$

Link function results

N	100 (β_1 = .81)	500 (β_1 = .81)	1000 (β_1 = .81)	1000 (β_1 =.405)
Osius	0	0.112	0.574	0.428
McCullagh	0	0.092	0.548	0.428
USS	0.054	0.290	0.586	0.430
IM	0.076	0.552	0.884	0.350
Stukel	0.036	0.478	0.878	0.352

Why the reversal in the last column? When β_1 = .81, the predicted probabilities cover a wide range. The two standardized Pearson tests put more weight on extreme values.

Summary

- All of the new GOF tests with ungrouped data are potentially useful in detecting misspecification.
- For detecting interaction, the Stukel test was markedly better than the others. But it was somewhat weaker for detecting quadratic effects.
- None of the tests was great at distinguishing a logistic model from a complementary log-log model, but IM and Stukel were best.
- Tests for specific kinds of misspecification may be much more powerful than global GOF tests. This was particularly evident for interactions. For many applications a targeted approach may be the way to go.

Summary

- I recommend using all these GOF tests. If your model passes all of them, you can feel relieved. If any one of them is significant, it's probably worth doing targeted tests.
- As with any GOF tests, when the sample size is quite large, it may not be possible to find any reasonably parsimonious model with a *p*-value greater than .05.
- If you use the GOFLOGIT macro, modify it to calculate two-sided *p*-values for the Osius and McCullagh versions of the standardized Pearson statistic.