

## Measures of Fit for Logistic Regression

Paul D. Allison, Statistical Horizons LLC and the University of Pennsylvania

### ABSTRACT

One of the most common questions about logistic regression is “How do I know if my model fits the data?” There are many approaches to answering this question, but they generally fall into two categories: measures of predictive power (like R-square) and goodness of fit tests (like the Pearson chi-square). This presentation looks first at R-square measures, arguing that the optional R-squares reported by PROC LOGISTIC might not be optimal. Measures proposed by McFadden and Tjur appear to be more attractive. As for goodness of fit, the popular Hosmer and Lemeshow test is shown to have some serious problems. Several alternatives are considered.

### INTRODUCTION

One of the most frequent questions I get about logistic regression is “How can I tell if my model fits the data?” Often the questioner is expressing a genuine interest in knowing whether a model is a good model or a not-so-good model. But a more common motivation is to convince someone else--a boss, an editor, or a regulator--that the model is OK.

There are two very different approaches to answering this question. One is to get a statistic that measures how well you can predict the dependent variable based on the independent variables. I'll refer to these kinds of statistics as *measures of predictive power*. Typically, they vary between 0 and 1, with 0 meaning no predictive power whatsoever and 1 meaning perfect predictions. Predictive power statistics available in PROC LOGISTIC include R-square, the area under the ROC curve, and several rank-order correlations. Obviously, the higher the better, but there is rarely a fixed cut-off that distinguishes an acceptable model from one that is not acceptable.

The other approach to evaluating model fit is to compute a goodness-of-fit statistic. With PROC LOGISTIC, you can get the deviance, the Pearson chi-square, or the Hosmer-Lemeshow test. These are formal tests of the null hypothesis that the fitted model is correct, and their output is a *p*-value--again a number between 0 and 1 with higher values indicating a better fit. In this case, however, a *p*-value below some specified  $\alpha$  level (say, .05) would indicate that the model is not acceptable.

What many researchers fail to realize is that measures of predictive power and goodness-of-fit statistics are testing very different things. It's not at all uncommon for models with very high R-squares to produce unacceptable goodness-of-fit statistics. And conversely, models with very low R-squares, can fit the data very well according to goodness-of-fit tests. As I'll explain in more detail later, what goodness-of-fit statistics are testing is *not* how well you can predict the dependent variable, but whether you could do even *better* by making the model more complicated, specifically, adding non-linearities, adding interactions, or changing the link function.

The goal of this paper is to discuss several issues that I've been grappling with over the years regarding both predictive power statistics and goodness-of-fit statistics. With one exception, I make no claim to originality here. Rather, I'm simply trying to make sense out of a rather complicated literature, and to distill it into some practical recommendations.

I begin with measures of predictive power, and I'm going to focus exclusively on R-square measures. I don't mean to imply that these are either better or worse than alternative measures like the area under the ROC curve. But I personally happen to like R-square statistics just because they are so familiar from the context of ordinary linear regression.

### $R^2$ STATISTICS FOR LOGISTIC REGRESSION

There are many different ways to calculate  $R^2$  for logistic regression and, unfortunately, no consensus on which one is best. Mittlbock and Schemper (1996) reviewed 12 different measures; Menard (2000) considered several others. The two methods that are most often reported in statistical software appear to be one proposed by McFadden (1974) and another that is usually attributed to Cox and Snell (1989) along with its “corrected” version.

The Cox-Snell  $R^2$  (both corrected and uncorrected) was actually discussed earlier by Maddala (1983) and by Cragg and Uhler (1970). Cox-Snell is the optional  $R^2$  reported by PROC LOGISTIC. PROC QLIM reports eight different  $R^2$  measures including both Cox-Snell and McFadden. Among other statistical packages that I'm familiar with, Statistica reports the Cox-Snell measures. JMP reports both McFadden and Cox-Snell. SPSS reports the Cox-Snell measures for binary logistic regression but McFadden's measure for multinomial and ordered logit.

For years, I've been recommending the Cox-Snell  $R^2$  over the McFadden  $R^2$ , but I've recently concluded that that was a mistake. I now believe that McFadden's  $R^2$  is a better choice. However, I've also learned about another  $R^2$  that has good properties, a lot of intuitive appeal, and is easily calculated. At the moment, I like it better than the McFadden  $R^2$ , but I'm not prepared to make a definitive recommendation at this point.

Here are some details. Logistic regression is, of course, estimated by maximizing the likelihood function. Let  $L_0$  be the value of the likelihood function for a model with no predictors, and let  $L_M$  be the likelihood for the model being estimated. McFadden's  $R^2$  is defined as

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

where  $\ln(\cdot)$  is the natural logarithm. The rationale for this formula is that  $\ln(L_0)$  plays a role analogous to the residual sum of squares in linear regression. Consequently, this formula corresponds to a proportional reduction in "error variance". It's sometimes referred to as a "pseudo"  $R^2$ .

The Cox and Snell  $R^2$  is

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

where  $n$  is the sample size. The rationale for this formula is that, for normal-theory linear regression, it's an identity. In other words, the usual  $R^2$  for linear regression depends on the likelihoods for the models with and without predictors by precisely this formula. It's appropriate, then, to describe this as a "generalized"  $R^2$  rather than a pseudo  $R^2$ . By contrast, the McFadden  $R^2$  does *not* have the OLS  $R^2$  as a special case. I've always found this property of the Cox-Snell  $R^2$  to be very attractive, especially because the formula can be naturally extended to other kinds of regression estimated by maximum likelihood, like negative binomial regression for count data or Weibull regression for survival data.

It's well known, however, that the big problem with the Cox-Snell  $R^2$  is that it has an upper bound that is less than 1.0. Specifically, the upper bound is  $1 - L_0^{2/n}$ . This can be a lot less than 1.0, and it depends only on  $p$ , the marginal proportion of cases with events:

$$\text{upper bound} = 1 - [p^p(1-p)^{(1-p)}]^2$$

I have not seen this formula anywhere else, so it may be the only original thing in this paper. The upper bound reaches a maximum of .75 when  $p=.5$ . By contrast, when  $p=.9$  (or .1), the upper bound is only .48.

For those who want an  $R^2$  that behaves like a linear-model  $R^2$ , this is deeply unsettling. There is a simple correction, and that is to divide  $R^2_{C\&S}$  by its upper bound, which produces the  $R^2$  attributed to Nagelkerke (1991) and which is labeled in SAS output as the max-rescaled  $R^2$ . But this correction is purely ad hoc, and it greatly reduces the theoretical appeal of the original  $R^2_{C\&S}$ . I also think that the values it typically produces are misleadingly high, especially compared with what you get from just doing OLS with the binary dependent variable. (Some might view this as a feature, however).

So, with some reluctance, I've decided to cross over to the McFadden camp. As Menard (2000) argued, it satisfies almost all of Kvalseth's (1985) eight criteria for a good  $R^2$ . When the marginal proportion is around .5, the McFadden  $R^2$  tends to be a little smaller than the uncorrected Cox-Snell  $R^2$ . When the marginal proportion is nearer to 0 or 1, the McFadden  $R^2$  tends to be larger.

But there's another  $R^2$ , recently proposed by Tjur (2009), that I'm inclined to prefer over McFadden's. It has a lot of intuitive appeal, its upper bound is 1.0, and it's closely related to  $R^2$  definitions for linear models. It's also easy to calculate.

The definition is very simple. For each of the two categories of the dependent variable, calculate the mean of the predicted probabilities of an event. Then, take the absolute value of the difference between those two means. That's it!

The motivation should be clear. If a model makes good predictions, the cases with events should have high predicted values and the cases without events should have low predicted values. Tjur also showed that his  $R^2$  (which he called the coefficient of discrimination) is equal to the arithmetic mean of two  $R^2$  formulas based on squared residuals, and equal to the geometric mean of two other  $R^2$ 's based on squared residuals.

Here's an example of how to calculate Tjur's statistic in SAS. I used a well-known data set on labor force participation of 751 married women (Mroz 1987). The dependent variable INLF is coded 1 if a woman was in the labor force, otherwise 0. A logistic regression model was fit with six predictors.

Here's the code:

```
proc logistic data=my.mroz;
  model inlf(desc) = kids1t6 age educ huswage city exper;
```

```

output out=a pred=yhat;
proc ttest data=a;
class inlf; var yhat; run;

```

The OUTPUT statement produces a new data set called A with predicted probabilities stored in a new variable called YHAT. PROC TTEST is a convenient way to compute the mean of the predicted probabilities for each category of the dependent variable, and to take their difference. The output is shown in Table 1. Ignoring the sign of the difference, the Tjur  $R^2$  is .2575. By contrast, the Cox-Snell  $R^2$  is .2477, and the max-rescaled  $R^2$  is .3322. McFadden  $R^2$  is .208. The squared correlation between the observed and predicted values is .2572.

**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
<b>Diff (1-2)</b>		-0.2575	0.2171	0.0160		

**Table 1. PROC TTEST Output to Compute Tjur's  $R^2$ .**

One possible objection to the Tjur  $R^2$  is that, unlike Cox-Snell and McFadden, it's not based on the quantity being maximized, namely, the likelihood function. As a result, it's possible that adding a variable to the model could reduce the Tjur  $R^2$ . But Kvalseth (1985) argued that it's actually preferable that  $R^2$  not be based on a particular estimation method. In that way, it can legitimately be used to compare predictive power for models that generate their predictions using very different methods. For example, one might want to compare predictions based on logistic regression with those based on a linear model or on a classification tree method.

Another potential complaint is that the Tjur  $R^2$  cannot be easily generalized to ordinal or nominal logistic regression. For McFadden and Cox-Snell, the generalization is trivial.

## CLASSIC GOODNESS-OF-FIT STATISTICS

I now turn to goodness-of-fit (GOF) tests, which can help you decide whether your model is correctly specified. GOF tests produce a  $p$ -value. If it's low (say, below .05), you reject the model. If it's high, then your model passes the test.

Classic GOF tests are readily available for logistic regression when the data can be aggregated or grouped into unique "profiles". Profiles are groups of cases that have exactly the same values on the predictors. For example, suppose we fit a model to the Mroz data with just two predictor variables, CITY (1=urban, 0=nonurban) and NKIDSLT6 which has integer values ranging from 0 to 3. There are then eight profiles, corresponding to the eight cells in the cross-classification of CITY by NKIDSLT6.

After fitting the model, we can get an observed number of events and an expected number of events for each profile. There are two well-known statistics for comparing the observed number with the expected number: the deviance and Pearson's chi-square. Here's how to get them with PROC LOGISTIC:

```

proc logistic data=my.mroz;
model inlf(desc) = kidslt6 city / aggregate scale=none; run;

```

The AGGREGATE option says to aggregate the data into profiles based on the values of the predictor variables. The SCALE=NONE option requests the deviance and the Pearson chi-square, based on those profiles. Here are the results.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
<b>Deviance</b>	4.1109	5	0.8222	0.5336
<b>Pearson</b>	3.9665	5	0.7933	0.5543

Number of unique profiles: 8

**Table 2. PROC LOGISTIC Output of GOF Statistics**

For both statistics, the chi-squares are low relative to the degrees of freedom, and the  $p$ -values are high. This is exactly what we want to see. There is no evidence to reject the null hypothesis, which is that the fitted model is correct. Now let's take a closer look at these two statistics.

The formula for the deviance is

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

where each  $j$  is a cell in the 2-way contingency table with each row being a profile and each column being one of the two categories of the dependent variable.  $O_j$  is the observed frequency and  $E_j$  is the expected frequency based on the fitted model. If  $O_j=0$ , the entire term in the summation is set to 0. The degrees of freedom is the number of profiles minus the number of estimated parameters.

The Pearson chi-square is calculated as

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

If the fitted model is correct, both statistics have approximately a chi-square distribution, with the approximation improving as the sample gets larger.

But what exactly are these statistics testing? This is easiest to see for the deviance, which is a likelihood ratio test comparing the fitted model to a "saturated" model that perfectly fits the data. In our example, a saturated model would treat KIDSLT6 as a CLASS variable, and would also include the interaction of KIDSLT6 and CITY. Here's the code for that model, with the GOF output in Table 3.

```
proc logistic data=my.mroz;
  class kidslt6;
  model inlf(desc) = kidslt6 city kidslt6*city / aggregate scale=none; run;
```

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
<b>Deviance</b>	0.0000	0	.	.
<b>Pearson</b>	0.0000	0	.	.

**Table 3. PROC LOGISTIC Output for a Saturated model**

So the answer to the question "What are GOF tests testing?" is simply this: they are testing whether there are any non-linearities or interactions. You can always produce a satisfactory fit by adding enough interactions and nonlinearities. But do you really need them to properly represent the data? GOF tests are designed to answer that question.

A related issue is whether the "link" function is correct. Is it logit, probit, complementary log-log, or something else entirely? Note that in a saturated model, the link function is irrelevant. It's only when you suppress interactions or non-linearities that the link function becomes an issue. For example, it's possible (although unusual) that interactions

that are needed for a logit model could disappear when you fit a complementary log-log model.

Both the deviance and the Pearson chi-square have good properties when the expected number of events and the expected number of non-events for each profile is at least 5. But most contemporary applications of logistic regression use data that do not allow for aggregation into profiles because the model includes one or more continuous (or nearly continuous) predictors. That's certainly true for the Mroz data when you include age, education, husband's wage, and years of experience in the model. When there is only one case per profile, both the deviance and Pearson chi-square have distributions that depart markedly from a true chi-square distribution, yielding  $p$ -values that may be wildly inaccurate. In fact, with only one case per profile, the deviance does not depend on the observed values at all, making it utterly useless as a GOF test (McCullagh 1985).

What can we do? Hosmer and Lemeshow (1980) proposed grouping cases together according to their predicted values from the logistic regression model. Specifically, the predicted values are arrayed from lowest to highest, and then separated into several groups of approximately equal size. Ten groups is the standard recommendation.

For each group, we calculate the observed number of events and non-events, as well as the expected number of events and non-events. The expected number of events is just the sum of the predicted probabilities for all the individuals in the group. And the expected number of non-events is the group size minus the expected number of events.

Pearson's chi-square is then applied to compare observed counts with expected counts. The degrees of freedom is the number of groups minus 2. As with the classic GOF tests, low  $p$ -values suggest rejection of the model.

For the Mroz data, here is the code for a model with five predictors:

```
proc logistic data=my.mroz;
  model inlf(desc) = kidslt6 age educ huswage city exper / lackfit;
run;
```

The LACKFIT option requests the Hosmer-Lemeshow (HL) test. Results are in Table 4.

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

**Table 4. Hosmer-Lemeshow Results from PROC LOGISTIC.**

The  $p$ -value is just below .05, suggesting that we may need some interactions or non-linearities in the model.

The HL test seems like a clever solution, and it has become the de facto standard for almost all software packages. But it turns out to have serious problems. The most troubling problem is that results can depend markedly on the number of groups, and there's no theory to guide the choice of that number. This problem did not become apparent until some software packages (but not SAS) started allowing you to specify the number of groups, rather than just using 10.

When I estimated this model in Stata, for example, with the default number of 10 groups, I got a HL chi-square of 15.52 with 8 df, yielding a  $p$ -value of .0499—almost the same as what we just got in SAS. But if we specify 9 groups, the  $p$ -value rises to .11. With 11 groups, the  $p$ -value is .64. Clearly, it's not acceptable for the results to depend so greatly on minor changes that are completely arbitrary. Examples like this are easy to come by.

But wait, there's more. One would hope that adding a statistically significant interaction or non-linearity to a model would improve its fit, as judged by the HL test. But often that doesn't happen. Suppose, for example, that we add the square of EXPER (labor force experience) to the model, allowing for non-linearity in the effect of experience. The squared term is highly significant ( $p=.002$ ). But with 9 groups, the HL chi-square increases from 11.65 ( $p=.11$ ) in the simpler model to 13.34 ( $p=.06$ ) in the more complex model. Thus, the HL test suggests that we'd be better off with the model that excludes the squared term.

The reverse can also happen. Quite frequently, adding a *non*-significant interaction or non-linearity to a model will substantially improve the HL fit. For example, I added the interaction of EDUC and EXPER to the basic model above. The product term had a  $p$ -value of .68, clearly not statistically significant. But the HL chi-square (based on 10 groups) declined from 15.52 ( $p=.05$ ) to 9.19 ( $p=.33$ ). Again, unacceptable behavior.

I am certainly not the first person to point out these problems. In fact, in a 1997 paper, Hosmer, Lemeshow and others acknowledged that the HL test had several drawbacks, although that hasn't stopped other people from using it. But if the HL test is not good, then how can we assess the fit of the model? It turns out that there's been quite a lot of work on this topic, and many alternative tests have been proposed—so many that it's rather difficult to figure out which ones are useful. In the remainder of this paper, I will review some of the literature on these tests, and I will recommend four of them that I think are worthy of consideration.

## NEW GOODNESS-OF-FIT TESTS

Many of the proposed tests are based on alternative ways of grouping the data (Tsiatis 1980, Pigeon and Heyse 1991, Pulkstenis and Robinson 2002, Xie et al. 2008, Liu et al. 2012). Once the data have been grouped, a standard Pearson chi-square is calculated to evaluate the discrepancy between predicted and observed counts within the groups. The main problem with these kinds of tests is that the grouping process usually requires significant effort and attention by the data analyst, and there is a certain degree of arbitrariness in how it is done. What most analysts want is a test that can be easily and routinely implemented. And since there are several tests that fulfill that requirement, I shall restrict my attention to tests that can be calculated when there is only one case per profile and no grouping of observations.

Based on my reading of the literature, I am prepared to recommend four statistics for widespread use:

1. **Standardized Pearson Test.** With ungrouped data, the formula for the classic Pearson chi-square test is:

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

where  $y_i$  is the dependent variable with values of 0 or 1, and  $\hat{\pi}_i$  is the predicted probability that  $y_i=1$ , based on the fitted model.

As we've just discussed, the problem with the classic Pearson GOF test is that it does not have a chi-square distribution when the data are not grouped. But Osius and Rojek (1992) showed that  $X^2$  has an asymptotic normal distribution with a mean and standard deviation that they derived. Subtracting the mean and dividing by the standard deviation yields a test statistic that has approximately a standard normal distribution under the null hypothesis. McCullagh (1985) derived a different mean and standard deviation after conditioning on the vector of estimated regression coefficients. In practice, these two versions of the standardized Pearson are nearly identical, especially in larger samples. Farrington (1996) also proposed a modified  $X^2$  test, but his test does not work when there is only one case per profile. For the remainder of this paper, I shall refer to the standardized Pearson test as simply the Pearson test.

2. **Unweighted Sum of Squares.** Copas (1989) proposed the test statistic

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

This statistic also has an asymptotic normal distribution under the null hypothesis, and Hosmer et al. (1997) showed how to get its mean and standard deviation. As with the Pearson test, subtracting the mean and dividing by the standard deviation yields a standard normal test statistic.

3. **Information Matrix Test.** White (1982) proposed a general approach to testing for model misspecification by 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 function and the other based on second derivatives. If the fitted model is correct, the expected values of these two estimators should be the same. Orme (1988, 1990) showed how to apply this method to test models for binary data. The test statistic is

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

where the  $x_{ij}$ 's are the  $p$  predictor variables in the model and  $x_{i0}=1$ . After standardization with an appropriate variance, this statistic has approximately a chi-square distribution with  $p+1$  degrees of freedom under the null hypothesis.

4. **Stukel Test.** Stukel (1988) proposed a generalization of the logistic regression model that has two additional parameters, thereby allowing either for asymmetry in the curve or for a different rate of approach to the (0,1) bounds. Special cases of the model include (approximately) the complementary log-log model and the probit model. The logistic model can be tested against this more general model by a simple procedure. Let  $g_i$  be the linear predictor from the fitted model, that is,  $g_i = \mathbf{x}_i' \mathbf{b}$  where  $\mathbf{x}_i$  is the vector of covariate values for individual  $i$  and  $\mathbf{b}$  is the vector of estimated coefficients. Then create two new variables:

$$z_a = g^2 \text{ if } g \geq 0, \text{ otherwise } z_a = 0$$

$$z_b = g^2 \text{ if } g < 0, \text{ otherwise } z_b = 0.$$

Add these two variables to the logistic regression model and test the null hypothesis that both of their coefficients are equal to 0. Stukel proposed a score test, but there's no obvious reason to prefer that to a Wald test or a likelihood ratio test. Note that in many data sets,  $g$  is either never greater than 0 or never less than 0. In such cases, only one  $z$  variable will be needed.

## IMPLEMENTING THE TESTS

As we'll see, Stukel's test is easily performed in SAS without much difficulty. The others are not quite so easy. Fortunately, Oliver Kuss has written a SAS macro that will calculate these and other tests. In fact, he presented a paper on that macro at SUGI 25 in 2001. Currently, the macro can be downloaded at <https://github.com/friendly/SAS-macros/blob/master/goflogit.sas>. Unfortunately, there is a major problem with this macro that I will explain later.

Let's apply these tests to the Mroz data used earlier. Recall that the HL test with ten groups yielded a  $p$ -value of .048, suggesting a need for interactions or non-linearities in the model. Here's the code for doing the Stukel test:

```
proc logistic data=my.mroz;
  model inlf(desc) = kidslt6 age educ huswage city exper;
  output out=a xbeta=xb;
data b;
  set a;
  za=xb**2*(xb>=0);
  zb=xb**2*(xb<0);
  num=1;
proc logistic data=b;
  model inlf(desc) = kidslt6 age educ huswage city exper za zb ;
  test za=0,zb=0;
run;
```

We first fit the model of interest using PROC LOGISTIC. The OUTPUT statement produces a new data set A that contains all the variables in the model plus the new variable XB, which is the linear predictor based on the fitted

model. In the DATA step that follows, the two new variables needed for the Stukel test are created. In addition, NUM=1 creates a new “variable” that will be needed for the GOFLOGIT macro. The second PROC LOGISTIC step estimates the extended model with the two new variables, and tests the null hypothesis that both ZA and ZB have coefficients of 0. This produced a Wald chi-square of .12 (2 df), yielding a  $p$ -value of .94. A likelihood ratio chi-square (the difference in the -2logL for the two models) produced almost identical results. Clearly there is no evidence against the model.

To calculate the other GOF statistics, we call the GOFLOGIT macro with the following statement:

```
%goflogit(data=b, y=inlf, xlist=kidslt6 age educ huswage city exper, trials=num)
```

The macro fits the logistic regression model with the dependent variable specified in Y= and the independent variables specified in XLIST=. TRIALS=NUM is necessary because the macro is designed to calculate GOF statistics for either grouped or ungrouped data. For ungrouped data, the number of trials must be set to 1, which is why I created the NUM=1 variable in the earlier DATA step. The output is shown in Table 5.

Results from the Goodness-of-Fit Tests	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

**Table 5. Output from GOFLOGIT Macro**

The first two tests are the classic Pearson and Deviance statistics, with  $p$ -values that can't be trusted with ungrouped data. Osius and McCullagh are two different versions of the standardized Pearson. As noted earlier, the Farrington test is not appropriate for ungrouped data—it's always equal to 0. What the macro labels as the RSS test is what I'm calling the USS test.

The only test yielding a  $p$ -value less than .05 is the “standard deviance” but, as I said earlier, this test is useless for ungrouped data because it doesn't depend on the observed values of  $y$ . The Farrington test is also useless because, with ungrouped data, it's always equal to 0. Notice that the Osius and McCullagh tests are very close, which has been the case with every data set that I've looked at. As reported here, the IM test is a chi-square statistic with  $df=7$  (the number of covariates in the model). The RSS value is just the sum of the squared residuals. Calculation of its  $p$ -value requires subtracting its approximate mean and dividing by its approximate standard deviation, and referring the result to a standard normal distribution.

## PROPERTIES OF THE TESTS

Simulation results show that all these tests have about the right “size”. That is, if a correct model is fitted, the proportion of times that the model is rejected is about the same as the chosen alpha level, say, .05. So, in that sense, all the tests are properly testing the same null hypothesis. But then we must ask two related questions: what sorts of departures from the model are these tests sensitive to, and how much power do they have to detect various alternatives? We can learn a little from theory and a little from simulation results.

**Theory.** When the data are naturally grouped, the classic Pearson and deviance tests are truly omnibus tests. That is, they respond to any non-linearity, interaction or deviation from the specified link function. The newer tests appear to be more specific. For example, by its very design, the Stukel test should do well in detecting departures from the logit link function. Similarly, Osius and Rojek (1992) showed that the Pearson test can be derived as a score test for a parameter in different generalization of the logit model. They describe this test as “a powerful test against particular alternatives concerning the link [function]”.

For the IM test, Chesher (1984) demonstrated that it is equivalent to a score test for the alternative hypothesis that the regression coefficients vary across individuals, rather than being the same for everyone. Similarly, Copas (1989) showed that the USS test can be derived as a score test for the alternative hypothesis that that the  $\pi_i$ 's are

independent random draws from a distribution with constant variance and means determined by the  $x$ 's. These results suggest that both the IM and USS tests should be particularly sensitive to unobserved heterogeneity.

**Simulation.** There have been three major simulation studies designed to assess the power of goodness-of-fit tests for logistic regression: Hosmer et al. (1997), Hosmer and Hjort (2002) and Kuss (2002). For convenience, I'll refer to them as H+, HH, and K. Of the four statistics under consideration here, H+ includes the Pearson, USS and Stukel. HH only considers the Pearson and USS. K includes Pearson, USS and IM.

All three studies use only sample sizes of 100 and 500. Here's a summary of their results for various kinds of departure from the standard logistic model:

*Quadratic vs. linear effect of a covariate.* H+ report that Pearson and USS have moderate power for  $N=100$  and very good power (above 90% under most conditions) for  $N=500$ . Power for Stukel is similar but somewhat lower. HH get similar results for Pearson and USS. K, on the other hand, found *no* power for Pearson, and moderate to good power for USS and IM, with USS noticeably better than IM for  $N=100$ .

*Interaction vs. linear effects of two covariates.* H+ found virtually no power for all tests under all conditions. But they set up the simulation incorrectly, in my judgment. HH reported power of about 40% for both Pearson and USS at  $N=100$  for a very strong interaction. At  $N=500$ , the power was over 90%. For weaker interactions, power ranged between 5% and 70%. K did not examine interactions. None of the simulations examined the power of Stukel or IM for testing interactions.

*Alternative link functions.* H+ found that Pearson, USS and Stukel generally had very low power at  $N=100$  and only small to moderate power at  $N=500$ . Stukel was the best of the three. HH report similar results for USS and Pearson. Comparing logistic with complementary log-log, K found no power for Pearson and moderate power for IM and USS, with IM somewhat better.

*Missing covariate and overdispersion.* K found that neither Pearson, USS or IM had any appreciable power to detect these kinds of departures.

*Discussion.* The most alarming thing about these simulation studies is the inconsistency between Kuss and the other two studies regarding the performance of the Pearson test. H+ and HH found that Pearson had reasonably good power to test several different kinds of specifications. Kuss, on the other hand, found that the Pearson test had virtually no power to test either a quadratic model or a misspecified link function.

Unfortunately, I believe that Kuss's simulations, which were based on his GOFLOGIT macro, have a major flaw. For the standardized Pearson statistics, he used a one-sided test rather than the two-sided test recommended by Osius and Rojek, and also by Hosmer and Lemeshow in the 2013 edition of their classic text, *Applied Logistic Regression*. When I replicated Kuss's simulations using a two-sided test, the results (shown below) were consistent with those of H+ and HH. There is also a flaw in the H+ simulations. For their interaction models, the fitted models removed the interaction *and* the main effect of one of the two variables. This does not yield a valid test of the interaction.

## NEW SIMULATIONS

To address problems with previous simulations, I ran new simulations that included all the GOF tests considered here. Whenever appropriate, I tried to replicate the basic structure of the simulations used in previous studies. As reported in those studies, all the tests rejected the null hypothesis at about the nominal level when the fitted model was correct. So I shall only report the estimated power of the tests to reject the null hypothesis when the fitted model is incorrect. For each condition, 500 samples were drawn.

*Linear vs. quadratic.* The correct model was  $\text{logit}(\pi_i) = \beta_0 + \beta_1 x + \beta_2 x^2$ . Values of the coefficients were the same as those used by Hosmer and Hjort. Coefficients were varied to emphasize or deemphasize the quadratic component, in four configurations: very low, low, medium and high. The variable  $x$  was uniformly distributed between -3 and 3. The fitted model deleted  $x^2$ . Sample sizes of 100 and 500 were examined. The linear model was "rejected" if the  $p$ -value for the GOF test fell below .05. In addition to the new GOF tests, I checked the power of the standard Wald chi-square test for  $\beta_2$ , the coefficient for  $x^2$ . Table 6 shows the proportion of times that the linear model was rejected, i.e., estimates of the power of each test.

When the quadratic effect is "very low", none of the tests had any appreciable power. For the "low" quadratic effect, we see pretty good power at  $N=500$ , and some power at  $N=100$ . For the "medium" quadratic effect,  $N=500$  gives near perfect power for all tests, but just moderate power  $N=100$ . In this condition, the Stukel test seems noticeably weaker than the others. For the "high" quadratic effect, we see very good power at  $N=100$ , and 100 percent rejection at  $N=500$ . The Wald chi-square for the interaction generally does better than the GOF tests, especially at  $N=500$ .

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 $\chi^2$		.0440	0.104	0.380	0.912	0.636	1.000	0.980	1.000

**Table 6. Power Estimates for Detecting a Quadratic Effect**

*Linear vs. interaction.* The correct model was  $\text{logit}(\pi_i) = \beta_0 + \beta_1x + \beta_2d + \beta_3xd$ . For the predictor variables,  $x$  was uniformly distributed between -3 and +3,  $d$  was dichotomous with values of -1 and +1, and the two variables were independent. Coefficients were chosen to represent varying levels of interaction. The fitted model deleted the product term  $xd$ . Sample sizes of 100 and 500 were examined. The linear model was “rejected” if the  $p$ -value for the GOF test fell below .05. Table 7 shows the proportion of times that the linear model was rejected.

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 $\chi^2$		0.120	0.426	0.342	0.950	0.666	1.000	0.864	1.000	0.966	1.000

**Table 7. Power Estimates for Detecting an Interaction**

In Table 7, we see that power to detect interaction with GOF tests is generally on the low side. Of the five new tests, Stukel clearly outperforms the others, especially at  $N=500$ . IM generally comes in second. But by comparison, the standard Wald chi-square test for the interaction is far superior to any of these tests. This illustrates the general principle that, while GOF tests may be useful in detecting unanticipated departures from the model, tests that target specific departures from the model are often much more powerful.

*Incorrect Link Function.* Most software packages for binary regression offer only three link functions: logit, probit and complementary log-log. So the practical issue is whether GOF tests can discriminate among these three. Logit and probit curves are both symmetrical so it's very hard to distinguish them. Instead, I'll focus on logit vs. complementary log-log (which is asymmetrical). The true model was linear in the complementary log-log:

$$\log(-\log(1 - \pi_i)) = \beta_0 + \beta_1x,$$

with  $x$  uniformly distributed between -3 and 3,  $\beta_0 = 0$  and  $\beta_1 = .81$ . The fitted model was a standard logistic model. Results for the GOF tests are shown in Table 8.

For  $N=100$ , none of the tests is any good. For  $N=500$ , the standardized Pearson tests are awful, USS is marginal, and IM and Stukel are half decent.

Things look a little different in the last two columns, however, where I increased the sample size to 1,000. When the coefficient stays the same, IM and Stukel are still the best, although the others are much improved. But when I reduced the coefficient of  $x$  by half, the Pearson statistics look better than IM and Stukel. Why the reversal? As

others have noted, the Pearson statistic may be particularly sensitive to cases where the predicted value is near 1 or 0 and the observed value is in the opposite direction. That's because each residual gets weighted by

$$\frac{1}{\hat{\pi}_i(1 - \hat{\pi}_i)}$$

which will be large when the predicted values are near 0 or 1. When  $\beta_1 = .81$ , many of the predicted values are near 0 or 1. But when  $\beta_1 = .405$ , a much smaller fraction of the predicted values are near 0 or 1. This suggests that the earlier simulations should also explore variation in the range of predicted values.

N	100 ( $\beta_1 = .81$ )	500 ( $\beta_1 = .81$ )	1000 ( $\beta_1 = .81$ )	1000 ( $\beta_1 = .405$ )
Osious	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

**Table 8. Power Estimates for Detecting an Incorrect Link Function**

## CLOSING POINTS

- 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. The Pearson tests were much worse than the others when many predicted probabilities were close to 1 or 0, and better than the others when predicted probabilities were concentrated in the midrange. This suggests that more elaborate simulations are needed for a comparative evaluation of these statistics.
- 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.
- 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 Osious and McCullagh versions of the standardized Pearson statistic.

## REFERENCES

- Chesher A. (1984) "Testing for neglected heterogeneity." *Econometrica* 52:865–872.
- Cragg, J.G. and R.S. Uhler (1970) "The demand for automobiles." *The Canadian Journal of Economics* 3: 386-406.
- Copas, J.B. (1989) "Unweighted sum of squares test for proportions." *Applied Statistics* 38:71 –80.
- Cox, D.R. and E.J. Snell (1989) *Analysis of Binary Data*. Second Edition. Chapman & Hall.
- Farrington, C. P. (1996) "On assessing goodness of fit of generalized linear models to sparse data." *Journal of the Royal Statistical Society, Series B* 58: 344–366.
- Hosmer, D.W. and N.L. Hjort (2002) "Goodness-of-fit processes for logistic regression: Simulation results." *Statistics in Medicine* 21:2723–2738.
- Hosmer, D.W., T. Hosmer, S. Le Cessie and S. Lemeshow (1997). "A comparison of goodness-of-fit tests for the logistic regression model." *Statistics in Medicine* 16: 965–980.

- Hosmer D.W. and S. Lemeshow (1980) "A goodness-of-fit test for the multiple logistic regression model." *Communications in Statistics* A10:1043-1069.
- Hosmer D.W. and S. Lemeshow (2013) *Applied Logistic Regression*, 3<sup>rd</sup> Edition. New York: Wiley.
- Kvalseth, T.O. (1985) "Cautionary note about  $R^2$ ." *The American Statistician*: 39: 279-285.
- Kuss, O. (2001) "A SAS/IML macro for goodness-of-fit testing in logistic regression models with sparse data." Paper 265-26 presented at the SAS User's Group International 26.
- Kuss, O. (2002) "Global goodness-of-fit tests in logistic regression with sparse data." *Statistics in Medicine* 21:3789–3801.
- Liu, Y., P.I. Nelson and S.S. Yang (2012) "An omnibus lack of fit test in logistic regression with sparse data." *Statistical Methods & Applications* 21:437–452.
- McFadden, D. (1974) "Conditional logit analysis of qualitative choice behavior." Pp. 105-142 in P. Zarembka (ed.), *Frontiers in Econometrics*. Academic Press.
- Maddala, G.S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
- McCullagh, P. (1985). "On the asymptotic distribution of Pearson's statistics in linear exponential family models." *International Statistical Review* 53: 61–67.
- Menard, S. (2000) "Coefficients of determination for multiple logistic regression analysis." *The American Statistician* 54: 17-24.
- Mittlbock, M. and M. Schemper (1996) "Explained variation in logistic regression." *Statistics in Medicine* 15: 1987-1997.
- Mroz, T.A. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions." *Econometrica* 55: 765-799.
- Orme, C. (1988) "The calculation of the information matrix test for binary data models." *The Manchester School* 54(4):370–376.
- Orme, C. (1990) "The small-sample performance of the information-matrix test." *Journal of Econometrics* 46: 309-331.
- 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.
- Nagelkerke, N.J.D. (1991) "A note on a general definition of the coefficient of determination." *Biometrika* 78: 691-692.
- Pigeon, J. G., and Heyse, J. F. (1999) "An improved goodness of fit test for probability prediction models. *Biometrical Journal* 41: 71–82.
- Press, S.J. and S. Wilson (1978) "Choosing between logistic regression and discriminant analysis." *Journal of the American Statistical Association* 73: 699-705.
- Pulkstenis, E., and T. J. Robinson (2002) "Two goodness-of-fit tests for logistic regression models with continuous covariates." *Statistics in Medicine* 21: 79–93.
- Stukel, T. A. (1988) "Generalized logistic models." *Journal of the American Statistical Association* 83: 426–431.
- Tjur, T. (2009) "Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination." *The American Statistician* 63: 366-372.
- Tsiatis, A. A. (1980) "A note on a goodness-of-fit test for the logistic regression model." *Biometrika*, 67: 250–251.
- White H. (1982) "Maximum likelihood estimation of misspecified models." *Econometrica* 50:1 –25.
- Xie, X.J., J. Pendergast and W. Clarke (2008) "Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors." *Computational Statistics & Data Analysis* 52: 2703 – 2713.

## **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at:

Name: Paul D. Allison  
Organization: University of Pennsylvania and Statistical Horizons LLC  
Address: 3718 Locust Walk  
City, State ZIP: Philadelphia, PA 19104-6299  
Work Phone: 215-898-6717  
Email: [allison@statisticalhorizons.com](mailto:allison@statisticalhorizons.com)  
Web: [www.pauldallison.com](http://www.pauldallison.com)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.