ONE MORE TIME ABOUT R^2 MEASURES OF FIT IN LOGISTIC REGRESSION

Ernest S. Shtatland, Ken Kleinman, Emily M. Cain
Harvard Medical School, Harvard Pilgrim Health Care, Boston, MA

ABSTRACT

In logistic regression, the demand for pseudo R^2 measures of fit is undeniable. There are at least a half dozen such measures, with little consensus on which is preferable. Two of them, both based on the maximum likelihood, are used in almost all statistical software systems. The first, R^2_1, has been implemented in SAS and SPSS. The second, R^2_2, (also known as McFadden's R^2, R^2_{MF}, the deviance R^2_{DEV} and the entropy R^2_E) is implemented in STATA and SUDAAN as well as SPSS.

Until recently these two measures have been considered independent. We will show in our presentation, which is a sequel to our SUGI 25 paper, that there exists a one-to-one correspondence between R^2_1 and R^2_2. If we know one of them, we know the other. The relationship between these measures of fit is required to understand which of them is preferred on a theoretical basis. To make this choice we consider our ability to interpret the measure in a reasonable way, the measure's dependence on the base rate as well as its degree of susceptibility to over-dispersion. We conclude that R^2_2 should be regarded as the standard R^2 measure.

INTRODUCTION

R^2 is probably the most popular measure of fit in statistical modeling. The measure provides a simple and clear interpretation, takes values between 0 and 1, and becomes larger as the model "fits better", in particular when we add more predictors. R^2 dominates in the SAS REG and GLM procedures. Researchers like to use the R^2 of the linear regression model and would like to have something similar to report in logistic regression.

According to Hosmer and Lemeshow (2000, p. 167):

> "Unfortunately, low R^2 values in logistic regression are the norm and this presents a problem when reporting their values to an audience accustomed to seeing linear regression values… Thus we do not recommend routine publishing of R^2 values from fitted logistic regression models."

We disagree with this opinion. It seems to us that the question of interpretation of R^2 is more important than the range of its values. It is much more important to know *what we measure* rather than to have the range of R^2 values similar to those of linear regression. We should make our choice of R^2 in logistic regression based on the intuitively meaningful interpretation. The problem of making this choice is nontrivial.

R^2 AVAILABLE IN SAS PROC LOGISTIC: R^2_{SAS}

R^2_{SAS} can be defined by the following equation:

$$R^2_{SAS} = 1 - \exp\{2[\log L(M) - \log L(0)] / n\} \quad (1)$$

where logL(M) and logL(0) are the maximized log likelihood for the fitted (current) model and the "null" model containing only the intercept term, and n is

the sample size. This definition is equivalent to that used in *SAS/STAT Software Changes and Enhancements Through Release 6.11*. (See also Maddala (1983), Cox and Snell (1989), Nagelkerke (1991), and Mittlbock and Schemper (1996)). The defined R^2_{SAS} cannot attain the value of 1 even if the model fits perfectly and residuals are zero (Mittlbock and Schemper (1996)). Nagelkerke (1991) proposed the following adjustment:

$$\text{Adj-}R^2_{SAS} = R^2_{SAS} / [1 - \exp(2 \log L(0) / n)] \quad (2)$$

In SAS this value is labeled "Max-rescaled RSquare". Although Adj-R^2_{SAS} can reach the maximum value of 1, the correction appears cosmetic and does not guarantee that intermediate values of Adj-R^2_{SAS} are adequate (see Mittlbock and Schemper (1996), p.1991).

An even more serious disadvantage is the lack of a reasonable interpretation. Unlike the linear model, R^2_{SAS} cannot be interpreted as a proportion of variation in the dependent variable that is explained by the predictors. According to Mittlbock and Schemper (1999) the values of R^2_{SAS} and Adj-R^2_{SAS} cannot be interpreted in any useful way. We cannot but agree with these authors.

THE PROPOSED R^2 MEASURES: R^2_{DEV}, R^2_E AND R^2_{MF}

The deviance R^2 can be defined as follows:

$$R^2_{DEV} = [\log L(M) - \log L(0)] / [\log L(S) - \log L(0)] \quad (3)$$

where logL(M), logL(0), and logL(S) are the maximized log likelihoods for the currently fitted, "null", and saturated models correspondingly (Hosmer and Lemeshow (2000), Agresti (1990), Menard (1995), Mittlbock and Schemper (1999) and Menard (2000)). If we work with *single-trial* syntax or *individual-level data*, then the saturated model has a dummy variable for each observation. Thus logL(S) = 0, and R^2_{DEV} simplifies to McFadden R^2 (R^2_{MF}). In case of *events / trials* syntax or *grouped-level data*, these two measures are different. For simplicity, we will consider mostly the individual-level data case, in which $R^2_{DEV} = R^2_E = R^2_{MF}$.

R^2_{MF} can be interpreted in two ways: first as proportional reduction in the - 2log-likelihood statistic (Menard (2000)). This interpretation is intuitively meaningful and fits the spirit of the maximum likelihood principle, the statistical basis for logistic regression. In addition, the formula

$$R^2_{MF} = 1 - \log L(M)/\log L(0) \quad (4)$$

is parallel to the formula for the ordinary least squares R^2 (R^2_{OLS}) as shown in (Menard (2000)).

Another interpretation of R^2_{MF} and also R^2_{DEV} and R^2_E is in terms of the information: they can be interpreted as the ratio of the estimated information gain when using the current model M in comparison with the null model to the estimate of the information potentially recoverable by including *all possible* explanatory variables (see Kent (1983) and Hastie (1987)). Both interpretations of R^2_{MF} are intuitively reasonable.

R^2_{SAS} AND R^2_{DEV}: FUNCTIONAL RELATIONSHIP

As far as we know, the researchers working with logistic regression treat R^2_{SAS} and R^2_{DEV} as independent R^2 measures. However, there exists a simple functional relationship between them. From (1) and (3), it is not difficult to show that

$$R^2_{SAS} = 1 - \exp\{-R^2_{DEV}\, 2[\log L(S) - \log L(0)]/n\} \quad (5)$$

If we use the notation

$$T = 2[\log L(S) - \log L(0)]/n,$$

formula (5) becomes

$$R^2_{SAS} = 1 - \exp(-R^2_{DEV} * T) \quad (5a)$$

If logL(S) = 0 (a single-trial case), then

$$R^2_{SAS} = 1 - \exp(-R^2_{MF} * T) \quad (5b)$$

and T = - 2logL(0)/*n*. We will consider only the single-trial case with logL(S) = 0 in the rest of the paper.

The maximized log-likelihood for the null model can be written as follows

$$\text{LogL}(0) = n[Y\log Y + (1 - Y)\log(1 - Y)] \quad (6)$$

and as a result

$$T = -2[Y\log Y + (1 - Y)\log(1 - Y)] \quad (7)$$

where $Y = (\sum y_i) / n$ and y_i denotes the binary outcome (see, for example, Agresti (1990), p. 110). Y is known as the *base rate* (Menard (2000)). From (5b) and (7) we can arrive at a number of important conclusions.

1) We see *how* and *why* R^2_{SAS} so strongly depends on the base rate Y. As noted in Menard (2000), R^2_{MF} and Y are almost un-correlated ($\text{corr}(R^2_{MF}, Y) = 0.002$). At the same time, Menard (2000) reported an extremely high empirical correlation between R^2_{SAS} and Y ($\text{corr}(R^2_{SAS}, Y) = 0.982$). (5b) and (7) provide us with a theoretical explanation of this obvious difference.

2) It is interesting that using (5b) we can calculate theoretically the values of R^2_{SAS} if R^2_{MF} and logL(0) (or the base rate) are known. And vise versa, if we know R^2_{SAS} and logL(0), we can calculate the values of R^2_{MF}. Applying formula (5b) to Table 1 in Menard (2000) with R^2_{MF} and the base rate as known quantities and R^2_{SAS} to be calculated, we find an excellent agreement between calculated (theoretical) and empirical values of R^2_{SAS} from Table 1.

3) From (5b) it can be seen that T is a key parameter in the relationship between R^2_{SAS} and R^2_{MF}. For sufficiently small values of R^2_{MF}, which are typical according to Hosmer and Lemeshow (2000), p. 167, formula (5b) can be linearized as follows:

$$R^2_{SAS} \cong T * R^2_{MF} \quad (8)$$

We would like to find the range of possible values of T, the key parameter in (5b) and (8). A priori, we can think that $0 < T < \infty$. Actually, this interval is much smaller. Note that parameter T as a function of base rate Y on the interval [0,1] is symmetrical with respect to Y = .5. It is equal to 0 at the ends Y = 0 and Y = 1, increases from 0 to 2ln2 on [0,1/2] and then decreases from 2ln2 to 0 on [1/2,1]. Thus, $0 \le T \le 2\ln 2 \cong 1.3863$. If $1 < T \le 2\ln 2$ which corresponds to $.2 < Y < .8$ then it can be shown that for small (enough) values of both R^2 measures we have $R^2_{SAS} > R^2_{MF}$. Otherwise, R^2_{SAS} and R^2_{MF} "switch" positions: $R^2_{SAS} < R^2_{MF}$. Note that as the upper limit for T is a rather small number of $2\ln 2 \cong 1.3863$, R^2_{SAS} and R^2_{MF} are of similar magnitude even if $R^2_{SAS} > R^2_{MF}$. This theoretical conclusion can be confirmed by empirical data from Table 1 in Menard (2000), for example:

$Y = 0.4$, $R^2_{SAS} = 0.291$, $R^2_{MF} = 0.255$
$Y = 0.495$, $R^2_{SAS} = 0.254$, $R^2_{MF} = 0.211$

and by the data from Mittlbock and Schemper (1999):

$Y = 0.4286$, $R^2_{SAS} = 0.4611$, $R^2_{MF} =$

0.4527
$Y = 0.5483$, $R^2_{SAS} = 0.1051$, $R^2_{MF} = 0.0806$

On the other hand, if $T \le 1$ then $R^2_{SAS} < R^2_{MF}$ and if T is small (which is equivalent to small base rate Y) then formulas (5b) and (8) show that R^2_{MF} can be *substantially* greater than R^2_{SAS}. The data from Table 1 (Menard (2000)) confirm our theoretical conclusion

$Y = 0.010$, $R^2_{SAS} = 0.020$, $R^2_{MF} = 0.187$
$Y = 0.021$, $R^2_{SAS} = 0.061$, $R^2_{MF} = 0.312$.

Thus for $0 < Y < .2$ or $.8 < Y < 1$, R^2_{SAS} gives smaller estimates of the performance of the model.

DOES R^2_{MF} ASSESS GOODNESS-OF-FIT?

Hosmer and Lemeshow (2000, p. 164) note that R^2 measures in logistic regression are based on comparisons of the current fitted model M to the null model and " as a result they do not assess goodness-of-fit". This remark is true for R^2_{SAS}. But R^2_{MF} is different because it compares the current model M with the saturated one and addresses the question "Does there exist a model which is substantially better than M?" Thus, R^2_{MF} could be treated as a goodness-of-fit measure.

SUMMARY OF COMPARISONS OF R^2_{SAS} AND R^2_{MF}

Since we know that there exists a one-to-one functional relationship between R^2_{SAS} and R^2_{MF}, it is natural to use *only one* of the measures.
To make the right choice we perform the following comparisons.

Interpretation. R^2_{SAS} has no satisfactory interpretation. At the same time, R^2_{MF} (or R^2_{DEV} and R^2_E) has at least two useful interpretations.

Base rate. R^2_{SAS} is strongly dependent on the base rate Y. Empirical results in Menard (2000) show that $corr(R^2_{SAS},Y) = .910$. At the same time, the correlation between R^2_{MF} (or R^2_{DEV} or R^2_E) and Y is almost negligible (.002).

Is the value of 1 attainable? R^2_{SAS} can never attain the value of 1. That is why adj-R^2_{SAS} is introduced, a measure that is even less interpretable than R^2_{SAS}. R^2_{MF} (or R^2_{DEV} or R^2_E) attains the value of 1 for a saturated model.

Overdispersion. It is easy to see that R^2_{SAS} is susceptible to overdispersion and R^2_{MF} is not.

Summarizing these results, we conclude that R^2_{MF} is undoubtedly superior to R^2_{SAS} and should be used instead of R^2_{SAS} in the SAS LOGISTIC procedure.

ADJUSTING FOR THE NUMBER OF PREDICTORS

Even after considering all the advantages R^2_{MF} enjoys over its competitor R^2_{SAS}, that measure is of rather limited use in model selection. It can be used only for comparison of models with the *same* number of explanatory variables. The reason for this is that R^2_{MF} always increases with any additional predictor. This is a common feature of all well-behaved R^2 measures. To make R^2_{MF} more useful in model selection (in particular, in comparing nested models), we have to adjust it by penalizing for model complexity -- the number of covariates in the model. We can do this in several ways: by using the adjustment as in R^2 for linear regression, by using the ideas behind information criteria (AIC and BIC), or by combining both previous approaches (Shtatland et al (2000)). We will mention here only the Akaike's type adjustment

$$\text{Adj-}R^2_{MF} = 1-(\log L(M) - k - 1) /(\log L(0) - 1) \quad (9)$$

where k is the number of covariates without an intercept. About this adjustment see (Mittlbock and Schemper (1996) and Menard (1995, p. 22)). About Harrell's adjustment see (Shtatland et al (2000)). Adjustment (9) works *exactly* as Akaike's Information Criterion (AIC), the most popular criterion in model selection. Nevertheless, we prefer to work with Adj-R^2_{MF} rather than AIC. The reason for this is that Akaike Information Criterion, being the estimate of the expected log-likelihood, takes rather arbitrary values: from very large positive to very large negative, which are hard to interpret. At the same time, in most cases adjustment (9) takes values between 0 and 1, which are far easier to interpret. This is why we suggest using Adj-R^2_{MF} at least as a supplement to AIC, if not instead of AIC.

CONCLUSION

In this paper, we have shown that two popular R^2 measures, R^2_{SAS} and R^2_{MF}, are not independent statistics. Instead, there exists a one-to-one correspondence between them. To avoid this redundancy, we should work either with R^2_{MF} or R^2_{SAS}. Thorough comparative analysis shows that R^2_{MF} has a number of important advantages over R^2_{SAS} and undoubtedly should be chosen as the standard R^2 measure in PROC LOGISTIC. To facilitate using R^2_{MF} in model selection, we propose to adjust it for the number of parameters. We suggest that the adj-R^2_{MF} has some important advantages over the popular information criterion AIC in terms of interpretability.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Cox, D. R. & Snell, E. J. (1989). *The Analysis of Binary Data,* Second Edition, London: Chapman and Hall.

Hastie, T. (1987). A Closer Look at the Deviance. *The American Statistician*, **41**, 16 – 20.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression (2nd Edition),* New York: John Wiley & Sons, Inc.

Kent J. T. (1983). Information gain and a general measure of correlation. *Biometrika,* **70**, 163 - 173.

Maddala, G. S. (1983). *Limited-Dependent and Quantitative Variables in Econometrics,* Cambridge: University Press.

Menard, S. (1995). *Applied Logistic Regression Analysis*, Sage University Paper series Quantitative Applications in the Social Sciences, series no. 07-106, Thousand Oaks, CA: Sage Publications, Inc.

Menard, S. (2000). Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, **54**, 17 – 24.

Mittlbock, M & Schemper, M. (1996). Explained Variation for Logistic Regression. *Statistics in Medicine*, **15**, 1987-1997.

Mittlbock, M & Schemper, M. (1999). Computing measures of explained variation for logistic regression models. *Computer Methods and Programs in Biomedicine*, **58**, 17 - 24.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika,* **78**, 691-692.

SAS Institute Inc. (1996). *SAS/STAT Software Changes and Enhancements Through 6.11*,

Cary, NC: SAS Institute Inc.

Shtatland, E. S., Moore, S. & Barton, M. B. (2000). Why We Need R^2 measure of fit (and not only one) in PROC LOGISTIC and PROC GENMOD. *SUGI 2000 Proceedings*, 1338 – 1343, Cary, SAS Institute Inc.

CONTACT INFORMATION:

Ernest S. Shtatland
Department of Ambulatory Care and Prevention
Harvard Pilgrim Health Care & Harvard Medical School
133 Brookline Avenue, 6th floor
Boston, MA 02115
tel: (617) 509-9936
email: ernest_shtatland@hphc.org