Paper 256-25

WHY WE NEED AN R^2 MEASURE OF FIT (AND NOT ONLY ONE) IN PROC LOGISTIC AND PROC GENMOD

Ernest S. Shtatland, PhD
Sara Moore, MPH
Mary B. Barton, MD, MPP
Harvard Medical School, Harvard Pilgrim Health Care, Boston, MA

ABSTRACT

We propose to use two seemingly different R^2 measures of fit in PROC LOGISTIC and PROC GENMOD (SAS/STAT), and we show that they are closely related to each other in terms of the amount of information gained when including predictors, in comparison with the "null" model. We suggest working with these R^2 measures *simultaneously* rather than separately because they can be seen as two faces of the same coin – information, and using them in parallel is similar to the binocular vision.

The intended audience: SAS users of all levels who work with SAS/STAT.

INTRODUCTION

R^2 is probably the most popular measure of fit in statistical modeling. R^2 dominates in PROC REG and PROC GLM. There is a natural appeal for a measure that can be computed for a fitted model, takes values between 0 and 1, becomes larger as the model "fits better", and provides a simple and clear interpretation. Researchers like to use the R^2 of the linear regression model and would like to have something similar to report for other models.

In this paper, we propose to use two seemingly different R^2 measures of fit in SAS PROC LOGISTIC and PROC GENMOD, and we show that they are closely related to each other in terms of the amount of information gained when including predictors, in comparison to the model without covariates.

One of the measures is already available in PROC LOGISTIC since Release 6.10. We will call it R$^2_{SAS}$. In Shtatland and Barton (1998), we discussed a very strong relationship between R$^2_{SAS}$ and the information gain statistic (see also Kent (1983)). R$^2_{SAS}$ has not gained much popularity and has been underused.

The R^2 measure that we propose to use is well known in statistical literature. It is used under different names: the likelihood-ratio-index, *aka* McFadden's pseudo R^2, *aka* the deviance R^2, etc. It is available in such statistical packages as STATA, GAUSS and SPSS, but not in SAS.

The use of both measures of fit simultaneously has been rarely reported because they can often demonstrate erratic behavior regarding each other, which may be confusing. However, we will show that working with them *simultaneously* is very instructive and advantageous, and that these measures are closely related to each other if we interpret them in terms of the information content of the data. Actually, they can be seen as two faces of the same coin – information. When we work with R$^2_{SAS}$, we compare the fitted model M with the "null" model as a reference. When we use the alternative R^2, we compare our fitted model with another baseline - the saturated model (though the intercept-only model also participates in this comparison).

In PROC GENMOD, there are only two basic measures of fit: the Deviance and Pearson's statistic. McCullagh and Nelder (1989) caution against the use of the deviance (and Pearson's statistic) alone to assess model fit. Thus, in PROC GENMOD it is even more urgent to have R^2 measures of fit.

R^2 MEASURES IN PROC LOGISTIC

a) The R^2 available in SAS: 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 model and the "null" model containing only an 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* (1996). See also Maddala (1983), Cox and Snell (1989), Nagelkerke (1991), and Mittlbock and Schemper (1996). Formula (1) can be rewritten as follows

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

As shown in Shtatland and Barton (1998), the right side of (2) can be interpreted as the amount of information gained when including the predictors into model M in comparison with the "null" model (see also Kent (1983)). By using notation IG(M) for the information gain, we have

$$- \log(1 - R^2_{SAS}) = IG(M), \quad (3)$$

$$R^2_{SAS} = 1 - \exp(- IG(M)) \quad (4).$$

The defined R^2_{SAS} cannot attain a value of 1 which is an obvious disadvantage of the measure. For example, it is possible that the model fits perfectly and residuals are zero but R^2_{SAS}= 0.75 (Mittlbock and Schemper (1996)). Nagelkerke (1991) proposed the following adjustment:

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

In SAS it is labeled as "Max-rescaled RSquare". Although Adj-R^2_{SAS} can reach a maximum value of 1, the correction appears cosmetic as it can only force the maximum of Adj-R^2_{SAS} to 100 % and there is no indication why the scaling of the intermediate values of Adj-R^2_{SAS} should be adequate (see Mittlbock and Schemper (1996), p.1991). Thus, a situation may arise in which the value of R^2_{SAS} is too small, and the value of Adj-R^2_{SAS} is too large. This is one potential explanation for the lack of popularity of R^2_{SAS}.

b) The proposed R^2 : R^2_{DEV}

The deviance R^2 can be defined as follows

$$R^2_{DEV}=[\log L(M)-\log L(0)]/[\log L(S)-\log L(0)] \quad (6)$$

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 (1989), Agresti (1990) and Menard (1995)). If we work with *single-trial* syntax, then the saturated model has a dummy variable for each observation. Thus logL(S) = 0, and R^2_{DEV} simplifies to McFadden's pseudo R^2. In case of *events* / *trials* syntax, these two measures are different.

R^2_{DEV} has been criticized by Hosmer and Lemeshow (1989), p.149 on the following grounds: the denominator in (6) is constant and the numerator is one-half the likelihood ratio test for significance; thus, the quantity R^2_{DEV} is "nothing more than an expression of the likelihood ratio test and, as such, is not a measure of goodness of fit." We disagree with this opinion. From our point of view, the basic idea behind R^2_{DEV} is to compare the log-likelihood gain achieved by the fitted model (the numerator in (6)) with the maximum *potential* log-likelihood gain (the denominator in (6)). Being a measure of comparison of the two log-likelihood gains (current *vs*. potential), R^2_{DEV} can be treated as an indicator of goodness-of-fit. Also R^2_{DEV} has the additional advantage of interpretation in terms of proportionate reduction in recoverable information.

Also, it has been mentioned in Agresti (1990), p. 110-112, that R^2_{DEV} can be large even when the strength of association is weak. But similar behavior occurs for R^2 in linear regression and it does not prevent R^2 from being the most popular measure of fit. Moreover, the fact that R^2_{SAS} and R^2_{DEV} work sometimes in opposite directions: R^2_{SAS} can be too low and R^2_{DEV} can be too high in some cases, creates an additional incentive to use them in combination with each other. Thus, the question is not whether to use R^2_{DEV} but rather *how* to use it in combination with R^2_{SAS}. Of course, of paramount importance is the relationship between R^2_{SAS} and R^2_{DEV}.

c) R^2_{SAS} and R^2_{DEV}: functional relationship

From (1) and (6), it is not difficult to see that

$$R^2_{SAS}=1 - \exp\{-R^2_{DEV}\ 2[\log L(S) - \log L(0)]/n\} \quad (7)$$

We will use the following notation

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

As a result, formula (7) reduces to

$$R^2_{SAS}=1 - \exp(-R^2_{DEV} * T\ ) \quad (7a)$$

We can interpret term T in formula (7a) in three different ways:

1) as an indicator of overdispersion of the "null" model (Dev[0] = 2[logL(S) - logL(0)]);
2) as an indicator of the *potentially* recoverable information (IG(S) = 2[logL(S) - logL(0)]/*n*);
3) as a characteristic of the *total variation* in the data: TOT_VAR = [logL(S) - logL(0)] .

The interpretation in terms of overdispersion can be useful, though, it is recommended to use the maximal available (but not saturated) model in estimating overdispersion (see McCullagh and Nelder (1989)).The information interpretation of T is important because in this case our basic formula relating R^2_{SAS} and R^2_{DEV} looks natural and transparent

$$R^2_{SAS}=1 - \exp(-R^2_{DEV} * IG(S)) \quad (7b)$$

Comparing (4) and (7b), we arrive at the formula

$$IG(M) = R^2_{DEV} * IG(S) \qquad (8)$$

This formula is equivalent to the original definition of R^2_{DEV} given by (6) and shows that R^2_{DEV} is nothing but the ratio of the estimated information gain when using fitted model M to the estimate of the information *potentially recoverable* by inclusion of all possible explanatory variables.

Summarizing, we can describe the R^2_{SAS} *vs*. R^2_{DEV} relationship as follows:

1) R^2_{SAS} compares the fitted model M with the smallest model of interest (the "null" model) as a reference;
2) when we use R^2_{DEV}, we compare our fitted model M with saturated model S, and the "null" model participates in this comparison.

Also, it is important to note that in formula (7a), term T does not depend on the fitted model, it depends only on our (unfitted) data and characterizes the total variability or potentially recoverable information in the data. Thus, term T can be considered constant in the process of selecting the best fitted model (within the given model specification).

d) R^2_{SAS} and R^2_{DEV}: quantitative relationship

First, let us discuss extreme cases. If our fitted model M is the "null" model then obviously R^2_{SAS} = Adj-R^2_{SAS} = R^2_{DEV} = 0. Also it is easy to see that if R^2_{DEV} = 1 then Adj-R^2_{SAS} = 1 and

$$R^2_{SAS} = 1 - \exp(2 \log L(0) / n) \quad (9)$$

which is its maximum attainable value (it is smaller than 1). In an intermediate case (R^2_{DEV} is between zero and one) quantitative R^2_{SAS} *vs.* R^2_{DEV} relations are more complex. If $T <= 1$ (no overdispersion for the "null" model or the information content per observation is less than or equal to 1), then it can be shown that $R^2_{SAS} < R^2_{DEV}$. At the same time, it has been shown in Cameron and Windmeijer (1997) that $R^2_{DEV} <$ Adj-R^2_{SAS}. Combining two previous unequalities we arrive at the conclusion that under our assumption of no overdispersion for the "null" model

$$R^2_{SAS} < R^2_{DEV} < \text{Adj-}R^2_{SAS} \quad (10)$$

Thus, we can think that R^2_{DEV} keeps balance between "too low" R^2_{SAS} and "too high" Adj-R^2_{SAS}. If $T > 1$ (there is some overdispersion for the "null" model or our data are "rich" in terms of information), then it can be shown that for small values of both R^2's we have $R^2_{SAS} > R^2_{DEV}$ then after some "critical" value they "switch" positions: $R^2_{SAS} < R^2_{DEV}$. Thus, for large enough values of both R^2's we always have $R^2_{SAS} < R^2_{DEV}$ which means that R^2_{SAS} can "soften" an unreasonably high value of R^2_{DEV} (compare this to the criticism in Agresti (1990), p. 111).

It is worth re-emphasizing the crucial role of the overdispersion parameter T in R^2_{SAS} *vs.* R^2_{DEV} relations: the same value $R^2_{DEV} = 0.05$ corresponds to $R^2_{SAS} = 0.049$ with $T = 1$, to $R^2_{SAS} = 0.095$ with $T = 2$, to $R^2_{SAS} = 0.221$ with $T = 5$, and to $R^2_{SAS} = 0.394$ with $T = 10$. This illustrates that R^2_{SAS} is sensitive to overdispersion, unlike R^2_{DEV}. In general, any correction for overdispersion affects R^2_{SAS}, but does not change R^2_{DEV}. Also, we have to remember that R^2_{SAS} compares the current model with the "null" model and R^2_{DEV} – with the saturated one. Thus, R^2_{SAS} and R^2_{DEV} mutually complement each other. We have to note again that both measures can be interpreted in terms of information content of the data. Then, R^2_{SAS} and R^2_{DEV} are likelihood-based, i.e. they are in agreement with the maximum-likelihood fitting procedure which is the basis of logistic and Poisson regression. And last but not least, it is simpler and more natural to adjust R^2_{SAS} and R^2_{DEV} than other measures for the number of parameters. All these facts support the use of R^2_{SAS} and R^2_{DEV} rather than many other competing R^2 measures. Among these competitors, we mention only the squared Pearson correlation between observed responses and their fitted values, and also the R^2 based on the residual sum of squares. The latter R^2 stems directly from the classic coefficient of determination in linear regression. There is some evidence of a trend in favor of log-likelihood / information based R^2 measures in generalized linear models in major statistical packages.

R^2 MEASURES AND INFORMATION CRITERIA IN MODEL SELECTION

With all the advantages over their competitors, R^2_{SAS} and R^2_{DEV} are of rather limited use in model selection. They can be used only for comparison of models with the same number of covariates. The reason is that both measures always increase with any additional covariate. This is a common feature of all well-behaved R^2 measures. To make our R^2 measures more useful in model selection, we have to adjust them by penalizing for the model complexity: the number of covariates in the model. We will demonstrate adjusting of R^2_{DEV} in the case of logistic regression with single-trial syntax. In this case R^2_{DEV} can be written as McFadden's pseudo R^2

$$R^2_{DEV} = 1 - \log L(M)/\log L(0) \qquad (11)$$

We can adjust R^2_{DEV} in a way the R^2 in linear regression is usually adjusted

$$\text{Adj1-}R^2_{DEV} = 1\text{-}\log L(M)/\log L(0))\ ((n-1)/(n\text{-}k\text{-}1)) \qquad (12)$$

where n is the sample size, and k is the number of covariates (without an intercept). This adjustment is supposed to work as well as it works in linear regression. If n is large and k is small to moderate, then the adjustment provided by (12) can be negligible. To have the adjustment more "k-oriented", let us define

$$\text{Adj2-}R^2_{DEV} = 1\text{-}(\log L(M) - k - 1)/(\log L(0) - 1) \qquad (13)$$

This is an " Akaike's type correction" (see Mittlbock & Schemper (1996) and also Menard (1995), p. 22 about Harrell's R^2). It works exactly as Akaike's Information Criterion (AIC) in model selection. In particular, it can be too "liberal" and tend to select too many covariates and overfit the model. To avoid overfitting, we can either use a "Schwarz' type correction", penalizing overfitting more strictly, or we can combine adjustments (12) and (13) into one adjustment

$$\text{Adj3-}R^2_{DEV} = 1\text{-}(\log L(M)\text{-}(k+1)(n\text{-}1)/(n\text{-}k\text{-}1))/(\log L(0)\text{-}1) \qquad (14)$$

Adjustment (14) is similar to the corrected versions of the AIC that were introduced to avoid possible bias in small sample cases (see, for example, Hurvich & Tsai (1995) and references therein). Though Adj2-R^2_{DEV} and Adj3-R^2_{DEV} are based on AIC-type corrections we prefer to work with (13) and (14) rather than AIC or its corrections. The reason is that Akaike Information Criteria and their corrections (being the estimates of the expected log-likelihood) take rather arbitrary values: from very large positive to very large negative, which are hard to interpret. At the same time, in most cases adjustments (13) and (14) take values between 0 and 1. And, more important, an R^2 measure is always meaningful: it is interpretable as a re-scaled measure of variation, as a comparison between two models: the current model and the reference one. This is why we suggest to use Adj2-R^2_{DEV} and Adj3-R^2_{DEV} at least as a supplement to the AIC and its family.

With the enhanced capabilities in Version 8 to output the resulting statistics for many SAS statistical procedures, it is not difficult to write SAS macros for calculating the proposed R^2 measures.

R^2 MEASURES IN PROC GENMOD

All of the above is equally related to PROC GENMOD. We can add that in PROC GENMOD it is even more necessary to have R^2 measures of goodness of fit. The reason is that now we have only two basic measures of fit: the Deviance and Pearson's statistic. McCullagh and Nelder (1989) caution against the use of the deviance (and Pearson's statistic) alone to assess model fit. A good discussion of R^2 measures in generalized linear models can be found in Kent (1983) and

Cameron and Windmeijer (1996, 1997).

CONCLUSIONS

In this paper, we show that two seemingly different R^2 measures of fit, R^2_{SAS} and R^2_{DEV}, can and should be used simultaneously in SAS PROC LOGISTIC and PROC GENMOD because they mutually complement each other. R^2_{SAS} and R^2_{DEV} can be interpreted in terms of information content of the data. They are likelihood-based, and as such they are in agreement with the maximum-likelihood method which is the basic fitting method in PROC LOGISTIC and PROC GENMOD. This is why R^2_{SAS} and R^2_{DEV} should be preferred to other competing R^2 measures. To facilitate using R^2_{SAS} and R^2_{DEV} in model selection, we propose to adjust them for the number of parameters and the sample size. We suggest that the adjusted R^2_{SAS} and R^2_{DEV} have some advantages over the popular information criteria (AIC, SIC, etc) in terms of interpretability.

REFERENCES

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

Cameron, A. C. & Windmeijer, F. A. (1996). R-Squared measures for count data regression models with applications to health care utilization. *Journal of Business & Economic Statistics*, **14**, 209-220.

Cameron, A. C. & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, **77**, 329-342.

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

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression,* New York: John Wiley & Sons, Inc.

Hurvich, C. M. & Tsai, C. L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, 51, 1077-1084.

Kent J. T. (1983). Information gain and a general measure of correlation. *Biometrika,* **70**, 163-73.

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 on Quantitative Applications in the Social Sciences, series no. 07-106, Thousand Oaks, CA: Sage Publications, Inc.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models,* Second Edition, New York: Chapman and Hall.

Mittlbock, M & Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, **15**, 1987-1997.

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 Release 6.11*, Cary, NC: SAS Institute Inc.

Shtatland, E. S. & Barton, M. B. (1998). An information-gain measure of fit in PROC LOGISTIC. *SUGI'98 Proceedings*, Cary, SAS Institute Inc., 1194-1199

CONTACT INFORMATION:

Ernest S. Shtatland
Department of Ambulatory Care and Prevention
Harvard Pilgrim Health Care & Harvard Medical School
126 Brookline Avenue, Suite 200
Boston, MA 02115
tel: (617) 421-2671
email: ernest_shtatland@hphc.org