AN INFORMATION-GAIN MEASURE OF FIT IN PROC LOGISTIC

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

ABSTRACT

This paper is a continuation of the paper [1] presented at NESUG, 1997. In [1] it was proposed to use information-gain or information-difference statistics as the most natural and most meaningful criteria of goodness of fit in such popular statistical modeling procedures as PROC REG, LOGISTIC, GENMOD, PHREG and others. The information-difference statistics have a very convenient property of additivity that allows SAS users to evaluate the contribution of an individual explanatory variable or a group of explanatory variables in terms of information gain in bits. It is interesting that these statistics are directly related to *F*- or Chi-Square statistics used in testing statistical significance. This is especially important when we are interested in understanding the relationship between statistical significance and substantive significance or importance. In [1] we focused mostly on PROC REG. Here the case of PROC LOGISTIC, the most popular regression procedure in health care research, is paid more attention. Much of the material given below is equally related to both PROC REG and PROC LOGISTIC. Moreover, we will find a striking analogy between these two types of regression analysis regarding the interplay of statistical significance and substantive importance if both are measured by using the same 'currency' - information.

.

SAS USERS AND SAS STATISTICAL MODELING PROCEDURES

One of the most difficult points in using SAS statistical modeling procedures (after the type of the model is chosen) is to understand whether the model we have built is good enough in terms of some goodness-of-fit criterion. This means that at a minimum, we have to understand and interpret printouts. The difficulty of this task is compounded because, for example, PROC REG has at least eleven statistics of fit ([2], p. 1369), PROC GENMOD - nine statistics ([3], p. 273), PROC LOGISTIC - four ([2], p. 1088, and [3], pp. 413-414), PROC PHREG - three ([3], p. 832), PROC MIXED - six ([3], pp. 547, 606-607), and so on. The expertise required to understand and interpret properly each of these statistics is not available to all users. Therefore it would be desirable to have a universal criterion with an easy to understand meaning that would be common for many applications.

POSSIBLE CANDIDATES FOR THE ROLE OF A UNIVERSAL CRITERION OF FIT

There are two clear candidates for the role of a universal criterion of fit: *R*-Square and a statistic based on likelihood.

1. *R*-SQUARE

R-Square is probably the most popular statistic of fit. Its popularity is so widespread that it can be found in Vanguard or Fidelity mutual funds booklets. *R*-Square dominates in PROC REG and PROC GLM (in the latter it is the sole criterion). *R*-Square has been added to the original list of goodness-of-fit criteria for PROC LOGISTIC. Also, *R*-Square is sometimes used in Survival Analysis. Unfortunately, *R*-Square is very often misused and misinterpreted. Perhaps it is the reason why in spite of all the popularity of *R*-Square, it ' is a measure many statisticians love to hate.' ([4], p. 113).

2. LIKELIHOOD

This is certainly a very fundamental characteristic, even more fundamental than *R*-Square. It appears in the form of the logarithm of the likelihood (log-likelihood) in printouts of almost all nonlinear regression procedures as itself or as a part of *AIC*, *BIC*, *SC* or *DEVIANCE*

(see [1], pp. 1088-1089, 1369; [2], p. 273). But sometimes it is used with the '+' sign, sometimes with the '-' sign; sometimes you can see the log-likelihood with the factor 2, sometimes without it. This is confusing, especially if one does not know whether the log-likelihood statisticis meaningful in itself or it is used instead of likelihood because of some mathematical convenience (by the way, the latter is the most typical explanation).

_R_-SQUARE AND INFORMATION - WHAT THEY HAVE IN COMMON

It has been noted ([5]-[7]) that there exists a simple formula relating *R*-Square and log-likelihood statistics

$$- \log(1- R^2) = \frac{2}{N}[\log L(b) - \log L(0)] \quad (1)$$

where N is a sample size, $\log L(b)$ and $\log L(0)$ denote the log-likelihoods of the fitted and "null" (with intercept only) models respectively. Here *log* is either the natural or binary logarithm. The natural logarithm is used more often with the Gaussian or normal distribution (in linear regression) and the base 2 logarithm with discrete distributions (binomial, Poisson etc.).

As shown in [3], we can think of the right side of (1) as a sample estimate of the difference between the information characteristics of the 'null' model (b=0) and the model under consideration. The most natural interpretation of this difference is the information gain (*IG*) we have when moving from the simplest 'null' model to the fitted model ([8]-[10]). Thus, we can write

$$IG = - \log(1- R^2) \quad (2)$$

Taking the first derivative of both sides of this equation, we have

$$\frac{dIG}{dR^2} = 1/(1-R^2) \quad (3)$$

From equation (3), we can see that

$$\frac{dIG}{dR^2} = 1 \quad if \ R^2 = 0, \quad (4)$$

$$\frac{dIG}{dR^2} = 2 \quad if \ R^2 = 0.5, \quad (5)$$

$$\frac{dIG}{dR^2} >> 1 \quad if \ R^2 \sim 1 \quad (6)$$

From (4), (5), and (6) we arrive at three important conclusions:

A) for small values of *R*-Square, the increase in *R*-Square is equivalent to the increase in information;

B) for *R*-Square~0.5, the information gain is twice as large as the increase in *R*-Square;

C) for *R*-Square close to 1, the change of information is much faster than that of *R*-Square.

Actually, it is almost obvious that the improvement in *R*-Square is much more 'valuable' if our initial value of *R*-Square is already high. But *how much* more valuable is an open question. Equation (2) allows us to quantify this intuitive idea in terms of information.

INFORMATION AND STATISTICAL SIGNIFICANCE *vs.* SUBSTANTIVE SIGNIFICANCE

A rather serious disadvantage of *R*-Square as a typical criterion of fit is the lack of a direct relationship between statistical significance of a model and its substantive significance. It is possible for the model to be statistically significant (say, $p < 0.0001$) with R-Square being not substantively significant (e.g., R-Square less than 0.005), which is not infrequent for a large sample. It is also possible for a model to be substantively significant (e.g., R-Square =0.4) but not statistically significant (which is rather typical for a small sample). The same is true for

differences in *R*-Square (the so-called *uniqueness indexes*, see [11], p. 407). It is a confusing moment in model building, particularly for the layperson.

Below we demonstrate that by using the concept of information we can build a bridge between statistical significance and substantive significance. It is well-known that the most powerful test of significance for overall regression is performed by using an *F*-statistic which in terms of *R*-Square is expressed by the formula (see, for example [12], p. 126):

$$F = \frac{N-K-1}{K} [R^2 / (1-R^2)] \quad (7)$$

where K is the number of explanatory or independent variables in the model and N is the number of observations. The test of significance is performed by comparing the statistic F with the critical level $F(K, N-K-1)$ of the *F*-distribution with K and $N-K-1$ degrees of freedom for the numerator and denominator correspondingly.

It is easy to see that (7) is equivalent to

$$-\log(1-R^2) = \log(1+F\frac{K}{N-K-1}) \quad (8)$$

From equations (2) and (8) we can conclude that for sufficiently small values of *R*-Square we have the following approximate *linear* relationships between the statistics *F* and *IG*:

$$IG \sim F\frac{K}{N-K-1} \quad (9)$$

or

$$F \sim IG\frac{N-K-1}{K} \quad (10)$$

The requirement for *R*-Square not to take comparatively large values is not a limitation at all in such fields, for example, as health care research, psychology, and behavioral sciences. J. Cohen in his well-known book on statistical power calculations, [13], proposed the following classification: he defined a *small* effect as the effect with *R*-Square of 0.02, a *moderate* effect - with *R*-Square of 0.13, and a *large* effect - with *R*-Square of 0.30. See also [14], p. 214.

If we can work with the approximation (10), we have this bridge between statistical significance and substantive importance and both are measured by using the same 'currency' - information. Indeed, working with (10), we can use, on one hand, IG as an indicator of substantive importance which shows how much information about our dependent variable is contained in the explanatory variables used in the model. On the other hand, the same information gain, IG multiplied by $N-K-1$ and divided by K (see the right side of (10)) is the statistic to be compared with $F(K, N-K-1)$, the corresponding critical level of *F*-distribution when testing statistical significance.

It is worth noting that the right side of (10) also has the meaning of information - it can be treated as the quantity of information in our data (per parameter) left *after* estimating the model parameters. This information can be used for prediction purposes. It is natural to call it the *information for prediction*. The decrease in the originally available information comes from the fact that $K+1$ degrees of freedom are used for estimating $K+1$ parameters.

Note also that this double role of information gain, *IG*, as the indicator of substantive significance and the statistic for testing statistical significance, is possible due to the linearity of (9) and (10).

INFORMATION APPROACH AND THE LAW OF PARSIMONY

From (2) and (8) it is easy to come to the following inequality

$$F > IG\frac{N-K-1}{K} \quad (11)$$

This inequality, which is always true, becomes the asymptotic equality (10) if *R*-Square is small enough. Inequality (11) shows also that working with the

information for prediction

$$IG\,\frac{N-K-1}{K} \quad (12)$$

instead of exact values of *F*-statistic we are more *conservative* and follow the law of parsimony (otherwise known as Occam's Razor). In the setting of statistical modeling, this law states that we should explain our data with as few parameters (explanatory variables) as possible. Our conservative information approach discussed above ideally meets this requirement.

Below we will see that such a conservative approach can be useful in logistic regression analysis.

INFORMATION STATISTICS AS CRITERIA OF FIT IN PROC LOGISTIC

Frequently applications of PROC LOGISTIC belong to medicine and health and behavioral sciences ([2], [3], [15]-[20]). For a typical application of logistic regression and PROC LOGISTIC in health care research see also [21].

For PROC LOGISTIC, equation (1) was used to introduce the *generalized R*-Square which was added (see [3], p. 414) to the original list of model fitting criteria (-2Log*L*, *AIC*, and *SC*). Equations (2) - (5) remain valid for PROC LOGISTIC. Equation (6) however, is not valid because *R*-Square in logistic regression achieves a maximum of less than 1 for discrete models ([3], [7], [8]).

The most popular criterion for assessing model fit is (-2log*L*) . It is also used for testing the statistical significance of the model, based on the fact that the difference

$$2(\log L(b) - \log L(0)) \qquad (13)$$

has *approximately* a Chi-Square distribution with *K* degrees of freedom as the number of observations, *N* increases. However, the two cautions are worth mentioning.

The first caution is about the interpretability of Chi-Square statistics. It was noticed (see [22] and references therein) that 'in spite of its practical utility the chi-square statistic is not an estimate of anything meaningful'. It is easy to cope with this disadvantage in our case: dividing (13) by *N*, we obtain the quantity which has the meaning of information.

The second caution is much more serious. The point is that the validity of a Chi-Square approximation in (13) could be guaranteed for *large* values of N only. But in many studies (especially of the exploratory, not confirmatory type) in such fields as health care research, behavioral sciences, psychology etc., sample sizes are either moderate or even *small*. In this case Chi-Square approximation in (13) can be rather poor, and Chi-Square statistics consistently tend to choose the more complex model, i.e., to over- parameterize the model (see [23]). Gelfand and Dey in [24] also show that likelihood ratio tests inherently tend to favor full models in contrast to reduced ones. In this light, using the Chi-Square approximations is at least questionable or even misleading when the sample size is not large enough.

It should be added that even for large values of *N* there is an upward bias, sometimes rather substantial, when treating the log-likelihood ratio (13) as a Chi-Square statistic ([8], [22]). Of course, the problem is expanded with the smaller sample size.

Thus, we have no reliable, theoretically based test for statistical significance of logistic regression in the abundance of cases with small-to-moderate sample sizes which are potentially most interesting. A realistic approach, adopted by many practitioners, is to treat the statistic (13), divided by its degrees of freedom *K* , as an *F*- statistic with *K* and *N - K - 1* degrees of freedom (see, for example, [25], p. 89). It makes the test more conservative. By using this 'practitioners' approach, equations (1) and (2), and the same notation for the *F*-statistic as above, we arrive at the following sequence of equations:

$$N(-\log(1-R^2)) = 2[\log L(b) - \log L(0)] \quad (14)$$

$$IG\,\frac{N}{K} = \frac{2}{K}[\log L(b) - \log L(0)] \quad (15)$$

If we change N for $N - K - 1$ in (16), we achieve two goals:
A) we are getting an even more conservative test which is important in exploratory, small-size studies;
B) we arrive at the equation

$$F \sim IG\,\frac{N-K-1}{K} \qquad (17)$$

$$IG\,\frac{N}{K} = F \qquad (16)$$

which is absolutely analogous to equation (10) obtained in the PROC REG setting. It is important to notice that our conclusion following equation (10), about bridging statistical significance and substantive importance in terms of information, remains valid in logistic regression.

This striking analogy between equation (10) in multiple linear regression and equation (17) in logistic regression could be treated as an indirect indication of the fact that our information approach that allows us to measure statistical significance and substantive importance of the model by using the same 'currency' - information, is a universal approach.

It is important to emphasize that we propose to use the *F*- approximation (16) - (17) to the Chi-Square test of significance in logistic regression not as a substitute, but rather as a supplement to the conventional test. It is similar to the case of overdispersion in PROC GENMOD in which it is also recommended to use both a Chi-Square and the corresponding *F*-test ([3], pp. 266-269). However, there is no a direct relationship between overdispersion and the sample size.

Only simulation studies can show how much superior to the conventional Chi-Square test is our proposed information-gain *F*-approximation.

CONCLUSIONS

1. The facts discussed above support the suggestion that SAS users in their statistical analyses and especially interpretations should think in terms of the information gain or information differences rather than *R*-Square and uniqueness indices. The information gain, not the uniqueness index, gives us a real value (in universal information units) of the improvements in our model when we add some new explanatory variables.
2. Information statistics provide us with a bridge between statistical significance and substantive importance of the models for both linear and logistic regression.
3. As we have seen, there is a striking similarity between equations (10) and (17) expressing linear relationship between *F*-statistics and information for prediction for both linear and logistic regressions.
4. We propose to use 'information' *F*-approximations in logistic regression (16) - (17) not as a substitute, but rather as a supplement to the conventional Chi-Square test of significance. We believe that there is an urgent need to supplement the already existing logistic regression software in order to have available some statistics discussed above that are useful in evaluating the logistic regression model.
5. Extensive simulation studies will measure the degree of improvements in the model building process when using our proposed 'information' *F*-approximations *vs*. conventional Chi-Square statistics.

REFERENCES

1. Shtatland, E. S. & Barton, M. B. (1997). Information as a unifying measure of fit in SAS statistical modeling procedures. *NESUG* '97 *Proceedings*, Baltimore, Me, 875-880.
2. SAS Institute Inc. (1990). *SAS/STAT User's Guide*, Version 6, Fourth Edition, Volume 2, Cary, NC: SAS Institute Inc.
3. SAS Institute Inc. (1996). *SAS/STAT Software Changes and Enhancements Through Release* 6.11, Cary, NC: SAS Institute Inc.
4. Anderson-Sprecher, R. (1994). Model comparisons and R^2 . *The American Statistician*, **48**, 113-117.
5. Cox, D. R. & Snell, E. J. (1989). *The Analysis of Binary Data*, Second Edition, London: Chapman and Hall.
6. Maddala, G. S. (1983). *Limited-Dependent and Quantitative Variables in Econometrics*, Cambridge: University Press.
7. Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691-692.
8. Kent J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, **70**, 163-173.
9. Kent J. T. & O'Quigley, J. (1988). Measures of dependence for censored survival data. *Biometrika*,**75**,

525-534.
10. El Hasnaoui A. & Jais, J. P. (1992). Study and applications of an informational measure of dependence in survival models. *Methods of Information in Medicine*, **31**, 275-283.
11. Hatcher, L. & Stepanski, E. J. (1994). *A Step-by -Step Approach to Using the SAS System for Univariate and Multivariate Statistics*, Cary, NC: SAS Institute Inc.
12. Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods*, Second Edition, Boston: PWS-Kent.
13. Cohen, J. (1987). *Statistical Power Analysis for the Behavioral Sciences* (Rev. ed), Hillside, NJ: Lawrence Erlbaum Associates.
14. Munro, B. H. & Page, E. B. (1993). *Statistical Methods for Health Care Research*, Philadelphia, Pennsylvania: J. B. Lippincott Company.
15. Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.
16. 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.
17. McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, Second Edition, New York: Chapman and Hall.
18. Dobson, A. J. (1983). *An Introduction to Statistical Modelling*, London: Chapman and Hall.
19. Dobson, A. J. (1990). *An Introduction to Generalized Linear Models*, London: Chapman and Hall.
20. Firth, D. (1991). Generalized Linear Models. In *Statistical Theory and Modelling*. Ed. Hinkley, D.V., Reid, N., and Snell, E. J., London: Chapman and Hall.
21. Barton, M. B., Fletcher, S. W. (1995) Preventive Services for Medicaid Enrollees in an HMO. *J Gen Int Med*, 10 (Suppl): 61.
22. Akaike, H. (1977). On Entropy Maximization Principle. In *Applications of Statistics*. Ed. Krishnaiah, P. R., New York: North-Holland Publishing Company.
23. Lin, T. H. & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, **22**, 249-264.
24. Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 501-504.
25. Aldrich, J. H. & Nelson, F. D. (1984). *Linear Probability, Logit, And Probit Models*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-045, Beverly Hills and London: Sage Publications, Inc.

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

Mary B. Barton, MD, MPP
Department of Ambulatory Care and Prevention
Harvard Medical School & Harvard Pilgrim Health Care
126 Brookline Avenue, Suite 200
Boston, MA 02115
tel: (617) 421-6011
email: mbarton@warren.med.harvard.edu