American Journal of Biostatistics 2 (1): 11-19, 2011
ISSN 1948-9889

Measures of Explained Variation and the Base-Rate Problem for Logistic Regression

[1]Dinesh Sharma, [2]Dan McGee and [3]B.M. Golam Kibria
[1]Department of Mathematics and Statistics,
James Madison University, Harrisonburg, VA 22801
[2]Department of Statistics, Florida State University, Tallahassee, FL 32306
[3]Department of Mathematics and Statistics, Florida International University, Miami, FL 33199

Abstract: Problem statement: Logistic regression, perhaps the most frequently used regression model after the General Linear Model (GLM), is extensively used in the field of medical science to analyze prognostic factors in studies of dichotomous outcomes. Unlike the GLM, many different proposals have been made to measure the explained variation in logistic regression analysis. One of the limitations of these measures is their dependency on the incidence of the event of interest in the population. This has clear disadvantage, especially when one seeks to compare the predictive ability of a set of prognostic factors in two subgroups of a population. **Approach:** The purpose of this article is to study the base-rate sensitivity of several R^2 measures that have been proposed for use in logistic regression. We compared the base-rate sensitivity of thirteen R^2 type parametric and nonparametric statistics. Since a theoretical comparison was not possible, a simulation study was conducted for this purpose. We used results from an existing dataset to simulate populations with different base-rates. Logistic models are generated using the covariate values from the dataset. **Results:** We found nonparametric R^2 measures to be less sensitive to the base-rate as compared to their parametric counterpart. Logistic regression is a parametric tool and use of the nonparametric R^2 may result inconsistent results. Among the parametric R^2 measures, the likelihood ratio R^2 appears to be least dependent on the base-rate and has relatively superior interpretability as a measure of explained variation. **Conclusion/Recommendations:** Some potential measures of explained variation are identified which tolerate fluctuations in base-rate reasonably well and at the same time provide a good estimate of the explained variation on an underlying continuous variable. It would be, however, misleading to draw strong conclusions based only on the conclusions of this research only.

Key words: Base-rate sensitivity, coefficient of determinant, latent scale linear model, R^2 statistic

INTRODUCTION

The Search for R^2 analogs in logistic regression: Prediction of future outcomes based on a given set of covariates is a key component of regression analysis. In Ordinary Least Squares (OLS) regression analysis, the predictive accuracy of a linear model is often judged using the R^2 statistic. This statistic has several mathematically equivalent definitions and multiple interpretations such as the proportion of variation in the dependent variable explained by the regressors, a measure of the strength of relationship between the covariate(s) and the response and the squared correlation between the observed and the predicted response. This statistic is usually not used as a measure of goodness-of-fit as other tools are better suited to that purpose (Hosmer *et al*., 2011). When the outcome variable is dichotomous, logistic regression model is the most popular choice. In most instances, interest lies in determining how well the model predicts the probability of group membership with respect to the dependent variable. Unlike the OLS regression, more than a dozen of R^2 measures have been suggested for the logistic regression model (Mittlbock and Schemper, 1996; Menard, 2000; DeMaris, 2002; Liao and McGee, 2003). But the best form of R^2 is not clear yet. Mittlbock and Schemper (1996) reviewed 12 measures of explained variation for logistic regression, Menard (2000) six and DeMaris (2002) seven, with some overlap. Other authors have proposed adjusted R^2 analogs (see, Liao and McGee, 2003; Mittlbock and Schemper, 2002), for exmaple). Recommendations based on various researches were different as different criteria were used to evaluate the R^2 analogs.

Corresponding Author: Dinesh Sharma, Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA 22801

Kvalseth's sixth criteria for a "good" R^2 statistic for the linear model (Kvalseth, 1985) requires an R^2 measure to be comparable across different models fitted to the same data. Menard (2000) extended this criteria requiring an R^2 measure to be comparable not only across different predictors but also across different dependent variables and different subsets of the dataset. With the help of an empirical example Menard (2000) demonstrated that R^2 measures in logistic regression are sensitive to the incidence of the event of interest in the population. Even if the coefficients associating particular variables to the outcome are the same in different populations, the value of the R^2 for populations with different incidence rates tend to be different. This phenomena is sometimes referred as the "base-rate" problem (Menard, 2000).

Having an R^2 measure that depends on the incidence of the response has disadvantage if one is seeking to compare the predictive ability of two different sets of prognostic factors or to compare the same set of factors in two subgroups of a population or in two different populations. If a R^2measure depends on the underlying incidence of the disease under study then the R^2 values for these two cases could differ because of the difference in the underlying incidence and not because of different predictive abilities. This phenomena is illustrated with the help of an empirical study in the following subsection.

An empirical example: The data used in this example are a subset of the Framingham Heart Study with known values of the covariates (age, systolic blood pressure, serum cholesterol, current cigarette smoking status and diabetic status). A logistic model with the ten-year incidence of Coronary Heart Disease (CHD) was estimated and thirteen different R^2 measures were calculated. Table 1 presents estimated R^2 s for each of the thirteen R^2 measures. The measures are larger for the female group than the male group, with only a few exceptions. If we had performed OLS regression, we would claim that we are able to predict CHD better in women than in men. However, women developed CHD at only half the rate that men did and if our measures are affected by the underlying rate of disease, then it would be misleading to make such a claim.

For a more detailed examination of the effect of the base-rate on potential measures of explained variance, we conducted a simulation study.

The purpose of this article is to study the base-rate sensitivity of several R^2 measures in logistic regression. We use an actual dataset to simulate populations with different base-rate. Logistic models are generated using actually occurring covariate values. The organization of the study is as follows: We introduce the R^2 measures to be examined. Simulation methods and simulation results are discussed. Summary and concluding remarks are given.

R^2-measures in logistic regression: We present some of the R^2 measures which have been proposed in the literature to estimate explained variation in logistic regression. Consider n observations (y_i, x_i) on a binary response variable y and a covariate vector $x = (x_1 \ldots x_p)$. The relationship between y and x is modeled by a logistic model Eq. 1:

$$\Pr(y_i = 1 | x_i) \equiv \pi_i(x_i) = \frac{e^{\beta_o + \beta' x}}{1 + e^{\beta_o + \beta' x}} \quad (1)$$

where β is a (p+1)-dimensional parameter vector. We denote the estimates from a logistic regression by $\hat{\Pr}(y_i = 1 | x_i) = \hat{\pi}(x_i)$ and $\hat{\Pr}(y_i = 1) \equiv \bar{y} = \sum_{i=1}^{n}(y_i / n)$. For logistic model with binary y it can be shown that $\bar{y} = \bar{\pi}$, the mean of conditional probability of success for all possible combinations of the covariate values.

Ordinary Least Squares R^2 (R^2_{OLS}): It is a natural extension of the coefficient of determination in OLS regression to the case of a binary y and is given by Eq. 2:

$$R^2_{OLS} = 1 - \sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (2)$$

Gini's Concentration R^2 (R^2_G): Gini's concentration measure $C(\pi) = 1 - \sum_{j=1}^{s}\pi_j^2$ is proposed as a measure of dispersion of a nominal random variable γ that assumes the integral values j, $1 \le j \le s$, with probability π_j (Haberman, 1982). If the outcome variable is binary, $C(\pi)$ reduces to $2\pi(1-\pi)$, where π is the probability that Y=1. The Gini's Concentration R^2 is the given by Eq. 3:

$$R^2_G = 1 - \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i) / [n\bar{y}(1-\bar{y})] \quad (3)$$

The Likelihood Ratio R^2 (R^2_L): Let L_0 be the likelihood of the model containing only the intercept and L_M be the likelihood of the model containing all of the predictors. The quantity $D_M = -2\log L_M$ represents the SSE for the full model and $D_0 = -2\log L_0$ represents the SSE of the model with only the intercept included, analogs to the total sum of squares (SST) in OLS. Thus the likelihood ratio R^2 for a logistic model becomes Eq. 4:

$$R^2_L = 1 - \log(L_M) / \log(L_o) \quad (4)$$

R^2 Based Upon Geometric Mean Squared Improvement (R^2_M)**:** In the linear regression model with normally distributed errors with zero mean and constant variance it can be shown that $R^2 = 1 - (L_o/L_M)^{2/n}$ (DeMaris, 2002).

Table 1: Male and female R^2 from a single cohort study

	R^2_p	R^2_{OLS}	R^2_G	R^2_L	R^2_M	R^2_N	R^2_C	R^2_{CS}	ACU	τ^2_a	τ^2_b	R^2_D	R^2_s
Females ($\bar{\pi}$ = 0.058)	0.060	0.060	0.062	0.110	0.048	0.133	0.047	0.151	0.756	0.003	0.029	0.512	0.043
Males ($\bar{\pi}$ = 0.119)	0.040	0.040	0.048	0.059	0.042	0.081	0.041	0.097	0.686	0.006	0.029	0.372	0.043

Since the method of maximum likelihood is the primary method of parameter estimation in the logistic regression, it seems quite natural to extend this concept of explained variation to the logistic regression setting. Maddala (1983) and Magee (1990) proposed the following R^2 analog Eq. 5:

$$R^2_M = 1 - e^{-\frac{2}{n}[\ln(L_M)-\ln(L_0)]} = 1-(L_0/L_M)^{2/n} \quad (5)$$

Since $L_0 \leq L_M$, R^2_M must be less than one. The maximum attainable value for R^2_M in Eq. 5 is max(R^2_M) =1-$(L_0)^{2/n}$. Nagelkerke (1991) proposed adjusting R^2_M by its maximum, $1-L_0^{2/n}$, to produce Eq. 6:

$$R^2_N = \frac{1-(L_0/L_M)^{2/n}}{1-L_0^{2/n}} \quad (6)$$

Contingency Coefficient R^2 (R^2_C): Aldrich and Nelson (1984) proposed an R^2 analog based on the model *Chi-squared* statistics G_M = -2log (L_0/L_M). It is a variant of the contingency coefficient and is given by Eq. 7:

$$R^2_C = G_M/(G_M+n). \quad (7)$$

R^2_C has the same mathematical form of the squared contingency coefficient and as such cannot equal one, even for a model that fits the data perfectly, because of the addition of the sample size in the denominator. Because of this limitation, Hagle and Mitchell (1992) proposed to adjust R^2_C by its maximum to produce Eq. 8:

$$R^2_{CS} = R^2_G / \max(R^2_G) \quad (8)$$

where, max $(R^2_C) = \frac{-2[\bar{y}\log\bar{y}+(1-\bar{y})\log(1-\bar{y})]}{1-2[\bar{y}\log\bar{y}+(1-\bar{y})\log(1-\bar{y})]}$ and $\bar{y} = \sum_i^n y_i / n$ is the sample proportion of cases for which y = 1.

Squared Pearson correlation (R^2_p): In linear regression R^2 is mathematically equivalent to the squared correlation between y and $\hat{y}$, its sample fitted value according to the model. The same idea is extended to the case of logistic regression and the R^2 analog is obtained by squaring the correlation coefficient between y and $\hat{\pi}$ as Eq. 9 (Maddala, 1983):

$$R^2_p = [\mathrm{corr}(y,\hat{\pi})]^2 = \frac{[\sum_{i=1}^n y_i\hat{\pi}_i - n\bar{y}^2]^2}{n\bar{y}(1-\bar{y})\sum_{i=1}^n(\hat{\pi}_i-\bar{y})^2} \quad (9)$$

Squared Spearman's Rho (ρ^2_S): Spearman's Rho is simply the Pearson's product moment correlation between ranks of y and $\hat{\pi}$. If we denote the rank of z by R(z) and mean of the ranks of both variables by $\bar{R} \equiv (n+1)/2$ then Spearman's ρ is given by Eq. 10:

$$r_s = \frac{\sum_{i=1}^n (R(y_i)-\bar{R})(R(\hat{\pi}_i)-\bar{R})}{\sum_{i=1}^n (R(y_i)-\bar{R})^2 \sum_{i=1}^n (R(\hat{\pi}_i)-\bar{R})^2} \quad (10)$$

Spearman's Rho is very close to Pearson's product moment correlation in normally distributed samples. For notational consistency, we will use R^2_S to denote squared r_S hereafter.

Squared Kendall's *τs* (τ^2_a and τ^2_b): Kendall (1990) suggested three possible coefficients, which he designated as τ_a, τ_b and τ_c. Only the first two of these coefficients are considered for our simulation study. Kendall's τ^2_a and τ^2_b are defined respectively as Eq. 11 and 12:

$$\tau_a = \frac{\sum_{i<j}\mathrm{sign}(y_j-y_i)\,\mathrm{sign}(\hat{\pi}_j-\hat{\pi}_i)}{n(n-1)/2} \quad (11)$$

And:

$$\tau_b = \frac{\sum_{i<j}\mathrm{sign}(y_j-y_i)\mathrm{sign}(\hat{\pi}_j-\hat{\pi}_i)}{\sqrt{\sum_{i<j}\mathrm{sign}^2(y_j-y_i)\sum_{i<j}\mathrm{sign}^2(\hat{\pi}_j-\hat{\pi}_i)}} \quad (12)$$

where, sign(z) is defined as sign (z) = $\begin{cases} 1 & \text{if } z>0 \\ 0 & \text{if } z=0 \\ -1 & \text{if } z>0. \end{cases}$

Squared Somers'd: Under the hypothesis that y causes or predicts $\hat{\pi}$, Somers (1962) proposed to use $d_{\hat{\pi}y}$ and for the hypothesis that $\hat{\pi}$ causes of predicts y, the proposed coefficient is $d_{y\hat{\pi}}$. The coefficients are defined respectively as Eq. 13 and 14:

$$d_{\hat{\pi}y} = \frac{\sum_{i<j} sign(y_j - y_i) sign(\hat{\pi}_j - \hat{\pi}_i)}{\sum_{i<j} sign^2(y_j - y_i)} \quad (13)$$

$$d_{y\hat{\pi}} = \frac{\sum_{i<j} sign(y_j - y_i) sign(\hat{\pi}_j - \hat{\pi}_i)}{\sum_{i<j} sign^2(\hat{\pi}_j - \hat{\pi}_i)} \quad (14)$$

with sign (z) defined as above. Somers' d's penalize for pairs tied on y only, in directional (asymmetric) hypotheses in which y causes or predicts $\hat{\pi}$; and to penalize for pairs tied on $\hat{\pi}$ only, in hypotheses in which $\hat{\pi}$ causes or predicts y. Kendall's τ_b is the geometric average of both asymmetric Somers' d, i.e., $\tau_b = \sqrt{d_{\hat{\pi}y} d_{y\hat{\pi}}}$. Because of this relationship, which is the same as the relationship between the classical regression coefficients and the product moment correlation ($r^2 = b_{xy} b_{yx}$), it is often viewed as an analog of a regression rather than a correlation coefficient. For notational consistency, we will use R_D^2 to denote squared $d_{y\hat{\pi}}$, hereafter.

Area Under ROC Curve (AUC): Suppose that the population under study can be divided into two sub-populations based on the status of the outcome variable Y: D (diseased) if Y = 1 and $\bar{D}$ (not diseased) if Y = 0. Let $F_1(.)$ and $F_0(.)$ be the CDFs of $\pi(x)$, the conditional probability of the outcome of interest, in D and $\bar{D}$, respectively. Let $c \in \mathbb{R}$ be such that:

$$Y = \begin{cases} 1 \text{ if } \pi(x) \geq c, \\ 0 \text{ otherwise} \end{cases}$$

For a given value of c, the sensitivity and specificity of a classification model are defined as sensitivity = Pr ($\pi(x) \geq c | Y=1) = 1-F_1(c)$ and specificity = Pr ($\pi(x) < c | Y = 0 = F_0(c)$ respectively. The ROC curve is then obtained by plotting $1-F_1(c)$ against $1-F_0(c)$ for all possible values of c. The area under the *ROC* curve is then given by Eq. 15:

$$\begin{aligned} AUC &= \int_{+\infty}^{-\infty} (1 - F_1(c)) d(1 - F_0(c)) \\ &= \int_{-\infty}^{+\infty} P[\pi_1(x) > c, \pi_0 = c] dc \\ &= P[\pi_1(x) > \pi_0(x)] \end{aligned} \quad (15)$$

where, $\pi_1(x)$ denotes conditional probability of disease in the diseased. The last equality follows because of the independence of the conditional probabilities in the two groups. Thus AUC represents the probability that a randomly chosen diseased subject is correctly rated or ranked with greater suspicion than a randomly chosen non-diseased subject.

MATERIALS AND METHODS

Simulation study: Consider a response variable Y and a covairate vector X = (X_1, X_2,… X_p)'. Let us further consider m different populations or m subsets of the same population and assume that each of the covariates $X_1, X_2, \ldots X_p$ has the same effect on the outcome variable Y in all populations (i.e. fixed effect across the populations) but each population has different proportions of successes (Y = 1). Using the logistic model, the odds of success in r^{th} (r =1, 2,…, m) population is given by Eq. 16:

$$\Phi^{(r)}(x) = e^{\beta_o^{(r)} + \beta' x}, r = 1, 2, ..., m \quad (16)$$

The odds ratio of j^{th} population relative to the k^{th} population is then given by Eq. 17:

$$OR = e^{\beta_o^{(j)}} - \beta_o^{(k)} \equiv t > 0 \quad (17)$$

This gives $\beta_0^{(j)} = \beta_0^{(k)} + \log(t)$. It follows that for a given t>1, $\bar{\pi}^{(j)}(x) > \bar{\pi}^{(k)}(x)$. Where $\bar{\pi}^{(r)} = p_r^{(r)}$ (Y = 1) is the base-rate in the r^{th} population. Therefore, by fixing the odds ratio to some constant t > 0, it is possible to find a β_0^* which can be used to generate new Y* with odds of success t times the odds success in the original data.

To design our simulation study, we elected to take advantage of naturally occurring covariate values by employing existing dataset to generate true logistic regression models. The data was a subset of the Framingham Heart Study data and consisted of 4,123 Men and women examined at a baseline examination and followed for 10 years. During the next 10 years, 370 (about 9%) developed Coronary Heart Disease (CHD). Males were twice as likely to develop CHD as females (6.0% for females, 12.7% for males). We simulated the logistic models as below.

Simulation algorithm: (i) Fit a logistic model to the original data that specifies: $\Phi(x) = e^{\beta o + \sum_{1=0}^{6} x_i \beta_i}$, where x_1 = age in years x_2= systolic blood pressure (mmHg), x_3 serum cholesterol (mg/dL), x_4= male gender (0 = female, 1 = male), x_5 = Cigarette smoker (0 = no, 1 = yes) and x_6 = diabetic (0 = no, 1 = yes). Compute R^2 measures and obtain the estimates $\hat{\beta}_o$ and $\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_6)$.

(ii) Let $\hat{\beta}_0^*$ be the estimate of β_0 from a data set with odds of success two times the odds of success in the original data. Substiture t = 2 in $\widehat{OR} = \frac{\hat{\Phi}^*(x)}{\hat{\Phi}(x)} = e^{\hat{\beta}_0^* - \hat{\beta}_0} = t$ and solve for $\hat{\beta}_0^*$. Compute $\hat{\pi}^*(x) = \frac{e^{\hat{\beta}_0^* + \hat{\beta}' x}}{1 + e^{\hat{\beta}_0^* + \hat{\beta}' x}}$, where $\hat{\beta}'$ is obtained in step i).

(iii) Generate y* such that:

$$y^* = \begin{cases} 1 & \text{if } \hat{\pi}^*(x) \geq U, \text{ where } U \sim Uni(0,1), \\ 0 & \text{otherwise} \end{cases}$$

(iv) Select a random sample of size n from the new data, fit the regression model $\Phi(x) = e^{\beta_0 + \sum_{i=0}^{6} x_i \beta_i}$ and compute R^2 measures.

(V) Repeat steps ii-iv for t = 3, 4,…, k. We used k =14 in our simulation. This yielded datasets with base rates ranging from 8.6-49.6%.

(Vi) Repeat steps ii-v 10,000 times for each of the sample sizes 500, 1000, 2000 and 4000. However, sample size did not affect the average value of any of the R^2 measures. Therefore, we present only the results for the sample size 4,000.

RESULTS

Simulation results: Intercorrelations of different R^2 measures and their correlations with the base-rate are presented in Table 2. Squared correlation of the R^2 measures with the base-rate are presented in the last row of the same table. Only two of the 13 R^2 measures, AUC and R_D^2, have very low (0.011) squared correlations with the base-rate. R_L^2 has some advantage over the other R^2 measures in the sense of having a low squared correlation with base-rate, but it is still substantial.

Means of the parametric and nonparametric measures are plotted against the base-rate in Fig. 1 and 2, respectively. All the 9 parametric measures exhibit a monotonically increasing tendency with the base-rate (Fig. 1). R_{CS}^2 is uniformly dominant over all other parametric measures, followed by R_N^2, across the levels of π. For small $\bar{\pi}$ (less than 0.2), R_L^2 appearers to be the third largest measure, but as $\bar{\pi}$ approached to 0.5 other measures come to the fore forcing R_L^2 to be the smallest R^2 measure for π >0.25. The remaining six parametric measures have almost identical means across the levels of $\bar{\pi}$.

Among the nonparametric measures, the AUC statistic consistently resulted in very large mean values irrespective of the base-rate followed by the R_D^2 (Fig. 2).

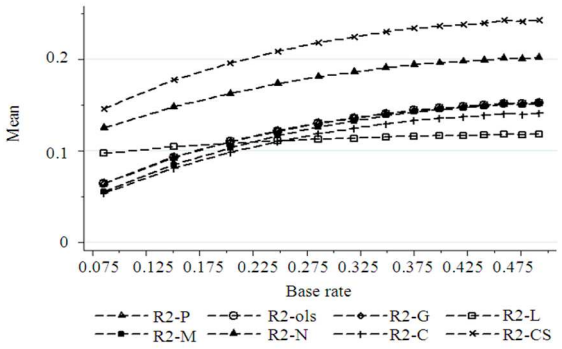

Fig. 1: Mean of the Parametric R^2 Measures by Base-rate

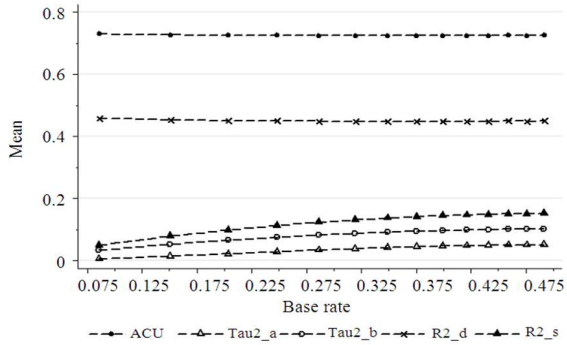

Fig. 2: Mean of the Nonparametric R^2 Measures by Base-rate

These two measures, unlike the rest of the nonparametric R^2 measures, exhibit a negative correlation with $\bar{\pi}$, which appears to be arising from the decreasing values of AUC for very low values of base-rate, particularly in the range 0%-2%. Otherwise, and R_D^2 appear to be mostly invariant with respect to the base-rate. All of these measure had very small standard deviations. We did not find any noticeable difference in their standard deviations (Table 3).

We evaluated the base-rate sensitivity of R^2 measures by examining the rate of change in their means associated with the small changes in the base-rate in the neighborhood of a given level of $\bar{\pi}$. In doing so, we numerically computed derivatives of the R^2 measures with respect to the base-rate using the "dydx" function available in stata @ 9.1 software (Stata Base Reference Manual, 2005). We did not consider the sign of the derivatives as we were particularly interested in the magnitude, rather than the direction of base-rate sensitivity of these R^2 measures. The results are presented in Fig. 3 for the parametric and in Fig. 4 for the nonparametric R^2 measures. The marked points represent absolute values of the numeric derivatives of the R^2 measures evaluated at each level of $\bar{\pi}$ employed.

Table 2: Correlation between base-rate and various R^2 measures. The last row gives the squared correlation of $\bar{\pi}$ with each of the R^2 measures.

R^2	$\bar{\pi}$	R_p^2	R_s^2	R^2_{OLS}	R^2_G	R^2_L	R^2_M	R^2_N	R^2_C	R^2_{CS}	ACU	τ_a^2	τ_b^2	R^2_D
$\bar{\pi}$	1.000													
R_p^2	0.899	1.000												
R_s^2	0.927	0.994	1.000											
R^2_{OLS}	0.899	1.000	0.994	1.000										
R^2_G	0.954	0.942	0.961	0.942	1.000									
R^2_L	0.549	0.797	0.754	0.796	0.570	1.000								
R^2_M	0.919	0.996	0.998	0.997	0.958	0.767	1.000							
R^2_N	0.848	0.985	0.972	0.985	0.880	0.882	0.977	1.000						
R^2_C	0.917	0.996	0.998	0.996	0.960	0.764	1.000	0.976	1.000					
R^2_{CS}	0.857	0.990	0.979	0.990	0.897	0.863	0.984	0.998	0.984	1.000				
ACU	-0.105	0.165	0.114	0.164	-0.133	0.683	0.120	0.306	0.114	0.269	1.000			
τ_a^2	0.970	0.969	0.986	0.969	0.968	0.680	0.980	0.935	0.979	0.942	0.028	1.000		
τ_b^2	0.927	0.994	1.000	0.994	0.961	0.754	0.998	0.972	0.998	0.979	0.114	0.986	1.000	
R^2_D	-0.105	0.165	0.114	0.164	-0.133	0.683	0.120	0.306	0.114	0.269	1.000	0.028	0.114	1.000
r^2	With $\bar{\pi}$	0.809	0.859	0.809	0.911	0.302	0.844	0.718	0.841	0.734	0.011	0.940	0.859	0.011

Table 3: Mean and standard deviations of various R^2 measures at different base-rates (standard deviations are given in parenthesis).

	$\bar{\pi}$													
R^2	0.0857	0.1508	0.2036	0.2479	0.2860	0.3193	0.3489	0.3753	0.3992	0.4209	0.4408	0.4591	0.4760	0.4916
R_p^2	0.0650	0.0940	0.1100	0.1220	0.1310	0.1360	0.1410	0.1450	0.1470	0.1490	0.1500	0.1520	0.1520	0.1530
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
R_s^2	0.0500	0.0790	0.0980	0.1130	0.1230	0.1310	0.1370	0.1410	0.1440	0.1470	0.1480	0.1510	0.1500	0.1520
	(0.006)	(0.007)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
R^2_{OLS}	0.0640	0.0930	0.1100	0.1220	0.1300	0.1360	0.1410	0.1440	0.1470	0.1480	0.1500	0.1520	0.1520	0.1530
	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
R^2_G	0.0640	0.0930	0.1100	0.1220	0.1300	0.1360	0.1410	0.1440	0.1470	0.1490	0.1500	0.1510	0.1520	0.1530
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
R^2_L	0.0970	0.1050	0.1080	0.1110	0.1130	0.1140	0.1150	0.1160	0.1170	0.1170	0.1170	0.1190	0.1180	0.1190
	(0.012)	(0.010)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
R^2_M	0.0550	0.0850	0.104 0	0.1170	0.1270	0.1330	0.1390	0.1430	0.1450	0.1470	0.1490	0.1510	0.1500	0.1520
	(0.007)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
R^2_N	0.1250	0.1490	0.1630	0.1730	0.1810	0.1860	0.1910	0.1940	0.1960	0.1980	0.1990	0.2020	0.2010	0.2020
	(0.015)	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)	(0.012)	(0.013)	(0.013)	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)
R^2_C	0.0540	0.0810	0.0990	0.1100	0.1190	0.1250	0.1300	0.1330	0.1360	0.1370	0.1390	0.1410	0.1400	0.1410
	(0.006)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
R^2_{CS}	0.1460	0.1780	0.1960	0.2090	0.2190	0.2250	0.230	0.2340	0.2360	0.2380	0.2400	0.2430	0.2410	0.2430
	(0.017)	(0.016)	(0.015)	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
AUC	0.7290	0.7260	0.7250	0.7250	0.7240	0.7240	0.724	0.7240	0.7240	0.7240	0.7240	0.7250	0.7240	0.7250
	(0.013)	(0.010)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
τ_a^2	0.0050	0.0130	0.0210	0.0280	0.0340	0.0380	0.0420	0.0440	0.0460	0.0480	0.0490	0.0500	0.0500	0.0510
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
τ_b^2	0.0330	0.0530	0.0660	0.0750	0.0820	0.0870	0.0910	0.0940	0.0960	0.0980	0.0990	0.1010	0.1000	0.1010
	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
R^2_D	0.4580	0.4530	0.4490	0.4490	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4500	0.4480	0.4500
	(0.027)	(0.021)	(0.019)	(0.017)	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)	(0.015)	(0.015)	(0.015)

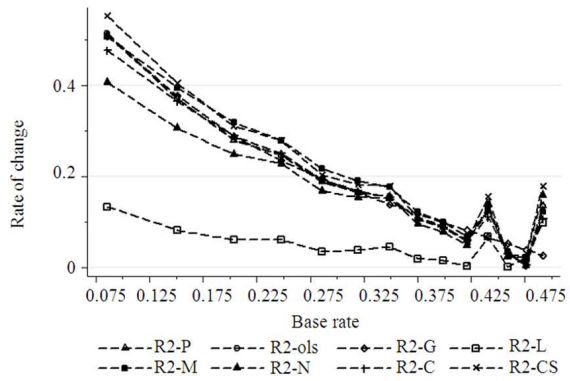

Fig. 3: Base-rate sensitivity of parametric R2 measures by base-rate

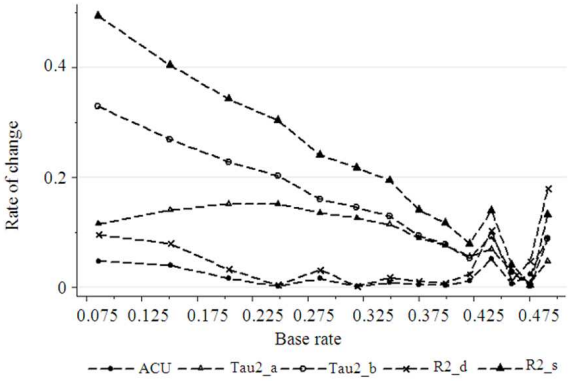

Fig. 4: Base-rate sensitivity of Nonparametric R^2 Measures by Base-rate

The large fluctuation observed at the higher end of $\bar{\pi}$ is attributed mainly to the error in estimating the derivatives at the end points.

It is evident from Fig. 3 that R_L^2 has a clear advantage over the rest of the parametric measures in the sense of having relatively small base-rate sensitivity. Like other parametric measures, it exhibits a steady decrease in base-rate sensitivity with increasing $\bar{\pi}$, but with a considerably slower rate as compared to the other parametric measures. The rest of the measures are in fairly good agreement with each other in terms of their sensitivity to the base-rate, at all levels of $\bar{\pi}$ employed. They exhibit very high levels of base-rate sensitive at small values of $\bar{\pi}$(<0.25). With increasing $\bar{\pi}$, their base-rate sensitivity rapidly decreases resulting quite low base-rate sensitivity when $\bar{\pi}$ is close to 0.5.

Among the nonparametric measures, the AUC and the R_D^2 appear to be the least base-rate sensitive at all levels of $\bar{\pi}$ (Fig. 4). In addition, these two measures demonstrate almost no fluctuation (except the fluctuation observed at the higher end of $\bar{\pi}$, which, as mentioned earlier, is primarily due to the estimation error) in their base-rate sensitivity for $\bar{\pi}$>0.2. τ_a^2, unlike other R^2 measures, exhibits a convex relationship with $\bar{\pi}$. Its base-rate sensitivity remains in between that of τ_b^2, the second worst measure in terms of the base-rate sensitivity, and R_D^2.

DISCUSSION

Summary and concluding remarks: The very existence of a plethora of R^2 measures for logistic regression sometime creates confusion about which measure to use in conjunction with a logistic regression analysis. Researchers have suggested various criteria for making judgment on these measures (for example see (Mittlbock and Schemper, 1996; Kvalseth, 1985; Sharma, 2006). Although the base-rate sensitivity of these R^2 measures has been documented (Menard, 2000; Gordon *et al*., 1979; Ash and Shwartz, 1999), the issue of whether this relationship to $\bar{\pi}$ is always a weakness of the R^2 measures is debated. Ash and Shwartz (1999) used a simple parametric model, applicable to a very specific situation, to clarify the effect of base-rate on R_{OLS}^2 and argued that it was in fact a strength rather than the weakness of R_{OLS}^2. Because in real-world situations the value of a diagnostic test does, in fact, depend on the prevalence of the problem in the population being tested (Ash and Shwartz, 1999; Hilden, 1991). This idea was further augmented by Mittlbock and Schemper (2002). They argued that if the base-rate is either close to 0 or 1, then the outcome is already pretty much determined and there is not much uncertainty left to be explained. However, on the other hand if the base-rate is large (i.e. if $\bar{\pi}$ is close to 0.5) the total variability in the dependent variable is high and the covariates may explain more of the uncertainty.

However, having an R^2 measure that depends on the incidence of the response has clear practical disadvantages if one is seeking to compare the predictive ability of a set of predictors in two sub-groups of a population. As illustrated in the empirical example presented, the analysis could lead to misleading conclusions. If a model, based on a particular R^2 value, shows better predictability in one population than the other, it may be simply because of the difference in the underlying incidence rate and not because of different predictive abilities of the set of predictors used. In this study we have examined the

base-rate sensitivity of thirteen R^2 type measures that are reported to have potential to be used as measures of explained variation in logistic regression analysis. Eight of these measures are parametric and the rest are nonparametric in nature. All of the R^2 measures are sensitive to the fluctuations in the base-rate. The magnitude of the base-rate sensitivity varies greatly from one measure to another. Results show that nonparametric measures tend to be less base-rate sensitive than the parametric measures. Four of these, τ_a, τ_b, R^2_D and R^2_S, are measures of ordinal association. Use of measures of ordinal association with a logistic regression model may result inconsistent behavior. For example, if a weak continuous covariate is added to a model with a strong binary covariate, the proportion of a explained variance, as measured by a parametric R^2, will increase slightly. But as a consequence of adding a continuous covariate in the model, ranks that were tied in the single covariate model are forced to slightly different values of the predictor. This may result a noticeable decrease in the proportion of explained variance, as measured by squared rank correlation, for example.

Among the parametric measures, R^2_L is the most base-rate invariant. In addition, its base-rate sensitivity fluctuates less as compared to other parametric measures, across the levels of $\bar{\pi}$. The closest competitors are the R^2_N and the R^2_{CS}. The observed difference between the base-rate sensitivity of these measures and that of the R^2_L is only marginal.

CONCLUSION

Use of R^2 in logistic regression has become a standard practice and many researchers have recommended it: Stata@ reports R^2_L as the part of its logistic regression analysis; Menard (2000) also preferred R^2_L over other R^2 measures because of its interpretability and independence from the base-rate; and Liao and McGee (2003) recommended routine use of R^2_L for logistic regression analysis. In spite of its interpretability and relatively superior ability to withstand fluctuations in base-rate, it is often criticized as having small values (Hosmer *et al*., 2011). If we consider y to be a binary proxy for a latent continuous variable y* that follows a multiple linear regression model, then the R^2 analogs can be viewed as the estimates of the ρ^2, the R^2 of the latent scale y*. Sharma and McGee (2008) found R^2_{CS} to be numerically most consistent with the underlying ρ^2 with R^2_N its nearest competitor. R^2_{CS} is based on the model chi-squared statistics and therefore has the advantages of being based on the quantity the model tries to maximize. Therefore, these two measures deserve serious consideration, especially when it is reasonable to believe that a underlying latent variable exists. They provide valuable information that R^2_L fails to provide, regarding the strength of relationship between the covariates and the underlying latent variable.

There are other potential factors whose effect on the base-rate sensitivity of R^2 measure is not studied in the current research. It would be dangerous to draw strong conclusions based only on the conclusions of this research. Some potential measures of explained variation are identified which tolerate fluctuations in base-rate reasonably well and at the same time provide a good estimate of the explained variation on underlying continuous variable.

ACKNOWLEDGEMENT

The data from the Framingham Heart Study was obtained from the National Heart, Lung and Blood Institute. The view expressed in this articals are those of the authors and do not necessarily reflect those of this agency.

REFERENCES

Aldrich, J.H. and F.D. Nelson, 1984. Linear Probability, Logit and Probit Models. 6th Edn., Sage Publications, Beverly Hills, ISBN-10: 0803921330 pp: 95.

Ash, A. and M. Shwartz, 1999. R^2: A useful measure of model performance when predicting a dichotomous outcome. Stat. Med., 18: 375-384. DOI: 10.1002/(SICI)1097-0258(19990228)18:4<375::AID-SIM20>3.0.CO;2-J

DeMaris, A., 2002. Explained variance in logistic regression a monte carlo study of proposed measures. Sociol. Methods Res., 31: 27-74. DOI: 10.1177/0049124102031001002

Gordon, T., W.B. Kannel and M. Halperin, 1979. Predictability of coronary heart disease. J. Chronic Disease, 32: 427-440. DOI: 10.1016/0021-9681(79)90103-6

Haberman, S.J., 1982. Analysis of dispersion of multinomial responses. J. Am. Stat. Assoc., 77: 568-580.

Hagle, T.M. and G.E. Mitchell, 1992. Goodness-of-Fit measures for probit and logit. Am. J. Polit. Sci., 36: 762-784.

Hilden, J., 1991. The area under the ROC curve and its competitors. Med. Decision Making, 11: 95-101. DOI: 10.1177/0272989X9101100204

Hosmer, D.W., S. Lemeshow and S. May, 2011. Applied Survival Analysis. John Wiley and Sons, ISBN-10: 1118211588, pp: 416.

Kendall, M.G. and J.D. Gibbons, 1990. Rank Correlation Methods. 5th Edn., Oxford University Press, New York, ISBN-10: 0195208374, pp: 260.

Kvalseth, T.O., 1985. Cautionary Note About R^2. Am. Stat., 39: 279-285.

Liao, J.G. and D. McGee, 2003. Adjusted coefficient of determination for logistic regression. Am. Stat., 57: 161-165.

Maddala, G.S., 1983. Limited-Dependent and Qualitative Variables in Econometrics. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521338255 pp: 401.

Magee, L., 1990. R^2 Measures based on wald and likelihood ratio joint significance tests. Am. Stat., 44: 250-253.

Menard, S., 2000. Coefficient of determination for multiple logistic regression analysis. Am. Statistician, 54: 17-24.

Mittlbock, M. and M. Schemper, 1996. Explained Variation for logistic regression. Stat. Med., 15: 1987-1997.

Mittlbock, M. and M. Schemper, 2002. Explained variation for logistic regression – small sample adjustments, confidence intervals and predictive precision. Biomet. J., 44: 263-272.

Nagelkerke, M.J.D., 1991. A note on a general definition of the coefficient of determination. Biometrika, 78: 691-692. DOI: 10.1093/biomet/78.3.691

Sharma, D. and D. McGee, 2008. Estimating proportion of explained variation for an underlying linear model using logistic regression analysis. J. Stat. Res., 42: 59-69.

Sharma, D.R., 2006. Logistic Regression, Measures of Explained Variation and the Base Rate Problem. Ph.D. Thesis, Florida State University, USA.

Somers, R.H., 1962. A new asymmetric measure of association for ordinal variables. Am. Sociolog. Rev., 27: 799-811.