Journal of Statistical Research *ISSN 0256 - 422 X*
2008, Vol. 42, No. 1, pp. 59-69
Bangladesh

ESTIMATING PROPORTION OF EXPLAINED VARIATION FOR AN UNDERLYING LINEAR MODEL USING LOGISTIC REGRESSION ANALYSIS

Dinesh Sharma
Department of Statistics
Florida International University, Miami, Florida 33199
Email: dinesh.sharma@fiu.edu

Daniel McGee
Department of Statistics
Florida State University, Tallahassee, Florida 32306
Email: dan@stat.fsu.edu

summary

Eight R^2 type statistics proposed to use in logistics regression analysis are evaluated based upon their ability to predict the proportion of explained variation (ρ^2) for an underlying linear model with latent scale continuous dependent variable. Functional relationships between these statistics are also studied. Predictive quality of these statistics depends mainly upon the proportion of success in the sample and ρ^2, the quantity to be predicted. We found R^2_{CS} (Hagle and Mitchell (1992)) to be numerically closest to the underlying ρ^2. There is a one-to-one correspondence between the likelihood based R^2 statistics, some of which have been considered independent until recently.

Keywords and phrases: Measure of explained variation, coefficient of determinant, R^2 statistic, logistic regression, latent scale linear model.

AMS Classification: 62J05

1 Introduction

Logistic regression analysis is popularly used to model the relationship between a binary response variable Y and a set of covariates $\mathbf{X} = (X_1, X_2 \ldots X_p)'$. Unlike the ordinary least squares (OLS) regression analysis, where R^2 is widely accepted as a measure of explained variation, more than a dozen summary measures have been suggested for logistic regression models (Mittlböck and Schemper (1996); Menard (2000); DeMaris (2002); Liao and McGee (2003)). Mittlböck and Schemper (1996) review 12 coefficients of determination for logistic regression, Menard (2000) six and DeMaris (2002) seven, with some overlap. Other authors have proposed adjusted R^2 analogs (for example, see Mittlböck and Schemper (2002) and Liao and McGee (2003)). But there is no consensus on the "best" R^2 statistic for use with logistic models. Recommendations based on various research are different as different criteria are used to evaluate the R^2 analogs. Nevertheless, these statistics are reported to have two major drawbacks. Firstly, all of the statistics are sensitive to the proportion of success (i.e. $P(Y = 1)$) in the data and secondly, they usually have small values. The second drawback is of particular importance when the binary random variable Y represents the categorization of an underlying continuous random variable. In many practical situations Y is used as a proxy for an underlying continuous random variable U, such that $Y = 1$ if $U \geq c$ for some $c \in \mathbb{R}$, and $Y = 0$ otherwise. If U is linearly related to $\mathbf{X}$, the usual coefficient of determinant R^2 can be used as a measure of the extent to which U is explained by $\mathbf{X}$. But measurements on U are usually not available and a researcher using a logistic model and seeking answer to the question "How well the covariates explain U?" has to rely on some measure of explained variation based on the logistic regression analysis of Y on $\mathbf{X}$. The tendency of some of the R^2 statistics to have very small values in logistic regression analysis poses a problem while estimating the extent of variation in U explained by $\mathbf{X}$ based on the logistic regression analysis of Y.

The focus of this paper is to compare and evaluate some commonly used R^2 statistics for logistic regression according to their ability to estimate the explained variation in an underlying continuous outcome variable. We consider eight R^2 statistics for our study. These statistics are described in Section 2. In Section 3 we present functional relationships between these statistics and how they compare with each other. A model for the unobservable underlying continuous outcome variable is developed and numerical results are presented in Section 4. Finally the conclusions of our study are presented in Section 1.2.

2 Measures of Explained Variation

Consider n observations $(y_i, \mathbf{x}_i)$, where $y_i = 0$ or 1 and is the outcome of the i^{th} subject, and the relationship between y_i and $\mathbf{x}_i$ is modeled by

$$P(y_i = 1|\mathbf{x}_i) \equiv \pi_i = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}},$$

where β is a *(p+1)*-dimensional parameter vector. We denote the estimates from a logistic regression by $\hat{P}(y_i = 1|\mathbf{x}_i) = \hat{\pi}_i$ and $\hat{P}(y_i = 1) \equiv \bar{\pi} = \sum_{i=1}^{n}(y_i/n)$. Furthermore, let $D(y_i)$ denote a measure of dispersion for the i^{th} observation relative to the marginal distribution of y and $D(y_i|\mathbf{x}_i)$ represents the measure computed conditional on the model and covariate vector $\mathbf{x}_i$. Then the reduction in variation of the outcome variable due to the covariates can be expressed by the difference $\sum\limits_{i=1}^{n} D(y_i) - \sum\limits_{i=1}^{n} D(y_i|\mathbf{x}_i)$ and the ratio $PEV = \left[\sum_{i=1}^{n} D(y_i) - \sum_{i=1}^{n} D(y_i|\mathbf{x}_i)\right] / \sum_{i=1}^{n} D(y_i)$ is the proportion of explained variation (Efron (1978); Agresti (1986)).

A number of variation functions have been proposed for a binary response. Some examples of these functions are the squared error, prediction error, entropy and linear error (Efron (1978)). Three of the eight measures of explained variation discussed in this section differ in their specification of $D(y_i)$ and $D(y_i|x_i)$ and have a PEV interpretation. Though the remaining five statistics do not have intuitive PEV interpretation, they are either derived as direct extensions of the different forms of R^2 statistics used in OLS regression analysis or are scaled versions of these R^2 analogs.

Ordinary Least Squares R^2: In OLS regression the coefficient of determination is defined as $R^2 = 1-\text{SSE/SST}$ and uses $SST = \sum\limits_i D(y_i)$ and $SSE = \sum\limits_i D(y_i|x_i)$. A natural extension of this idea to the case of a binary y would be to use $D(y_i) = (y_i - \bar{y})^2$ and $D(y_i|x_i) = (y_i - \hat{\pi}_i)^2$. For binary dependent variable, this R^2 statistic becomes

$$R^2_{OLS} = 1 - \sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{2.1}$$

Gini's Concentration R^2: 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 Y 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 $\pi = P(Y = 1)$. Thus using $D(y_i) = 2\bar{y}(1-\bar{y})$ and $D(y_i|x_i) = 2\hat{\pi}_i(1-\hat{\pi}_i)$ this R^2 statistic reduces to

$$R^2_G = 1 - \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i) / \sum_{i=1}^{n}\bar{y}(1-\bar{y}). \tag{2.2}$$

The Likelihood Ratio R^2: 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 using $\sum_i D(y_i) = -2\log(L_0)$ and $\sum\limits_{i=1}^{n} D(y_i|\mathbf{x}_i) = -2\log L_M$ the likelihood ratio R^2 for a logistic model becomes

$$R^2_L = 1 - \log(L_M)/\log(L_0). \tag{2.3}$$

R^2 Based Upon Geometric Mean Squared Improvement: In the linear regression model with normally distributed errors with zero mean and constant variance it can be shown that $R^2 = 1-(L_0/L_M)^{2/n}$ (DeMaris (2002)). 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 (1900) proposed using the following R^2 analog:

$$R^2_M = 1 - e^{-\frac{2}{n}[ln(L_M)-ln(L_0)]} = 1 - (L_0/L_M)^{2/n}. \tag{2.4}$$

Since $L_0 \leq L_M$, R^2_M must be less than one. The maximum attainable value for R^2_M in Equation (2.4) 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

$$R^2_N = \frac{1 - (L_0/L_M)^{2/n}}{1 - L_0^{2/n}}. \tag{2.5}$$

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

$$R^2_C = G_M/(G_M + n). \tag{2.6}$$

R^2_C has the same mathematical form of the squared contingency coefficient and as such can not 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:

$$R^2_{CS} = R^2_C/max(R^2_C), \tag{2.7}$$

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: 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 $\hat{\pi}$ (Maddala (1983)). This R^2 statistic becomes

$$R^2_P = [corr(y, \hat{\pi})]^2 = \frac{\left[\sum_{i=1}^n y_i\hat{\pi}_i - n\bar{y}^2\right]^2}{n\bar{y}(1-\bar{y})\sum_{i=1}^n(\hat{\pi}_i - \bar{y})^2}. \tag{2.8}$$

3 Comparison of R^2 Analogs

Comparison of Likelihood Based R^2 Measures: In Section 2 we have presented five R^2 statistics based on the likelihood function. There is a close relationship between these statistics and all of them can be expressed as a function of any of the other likelihood based R^2 statistic. With simple algebra R^2_M, R^2_N, R^2_C and R^2_{CS} can be written as a function of

R^2_L as follow:

$$R^2_M = 1 - (L_0/L_M)^{2/n} = 1 - e^{-\frac{2}{n}[\log\, L_M - \log\, L_0]} = 1 - e^{-\delta\, R^2_L} \tag{3.1a}$$

$$R^2_N = \frac{1 - e^{-\delta R^2_L}}{1 - L_0^{2/n}} = \frac{1 - e^{-\delta R^2_L}}{1 - e^{-\delta}} \tag{3.1b}$$

$$R^2_C = \frac{-2(\log\, L_0 - \log\, L_M)}{-2(\log\, L_0 - \log\, L_M) + n} = \frac{\delta R^2_L}{\delta R^2_L + 1} \tag{3.1c}$$

$$R^2_{SS} = \frac{R^2_C}{max(R^2_C)} = \left[\frac{\delta R^2_L}{\delta R^2_L + 1}\right] / \left[\frac{\delta}{\delta + 1}\right] = \frac{(1+\delta) R^2_L}{\delta R^2_L + 1} \tag{3.1d}$$

where, $\delta \equiv -2\, \log\, L_0/n = -2\, [\bar{\pi} \log \bar{\pi} + (1 - \bar{\pi}) \log{(1 - \bar{\pi})}]$.

Typically R^2_L is small in value. Therefore, for sufficiently small R^2_L the right hand side in (3.1a) can be linearized as $R^2_M \approx \delta\, R^2_L$. Similarly, using (3.1c) we may write $R^2_C/R^2_L = \delta R^2_L / \left(\delta (R^2_L)^2 + R^2_L\right) \approx \delta$ for sufficiently small R^2_L. Now δ, the key parameter in defining the above functional relationships is a parabolic function of $\bar{\pi}$ in the range $[0, 1]$ and attains its maximum at $\bar{\pi} = 0.5$. Furthermore, $\delta > 1$ if $0.2 < \bar{\pi} < 0.8$, and $\delta < 1$ otherwise. This suggests that, for sufficiently small values of R^2_L, $R^2_L < R^2_M$ and R^2_C when $0.2 < \bar{\pi} < 0.8$, and $R^2_L > R^2_M$ and R^2_C, otherwise. Using simple algebra it can be easily shown that that $R^2_C < R^2_M < R^2_N < R^2_{CS}$ and $R^2_L < R^2_N < R^2_{CS}$.

Comparison of R^2_{OLS}, R^2_P and R^2_G: R^2_G differs from the R^2_{OLS} and R^2_P in that it is based on the predicted values rather than the error vector. However there is a close relationship between these three R^2 statistics (Hu et al. (2006))

$$\left(R^2_G + R^2_{OLS}\right)^2 = 4 R^2_P R^2_G \tag{3.2}$$

Applying the Cauchy-Schwartz inequality to the right hand side of Equation (3.2) we have the following inequality

$$R^2_P \geq R^2_{OLS} \tag{3.3}$$

Similarly, we can write

$$\left(R^2_P + R^2_{OLS}\right)^2 \geq 4 R^2_P R^2_{OLS}, \text{ and}$$
$$\left(R^2_P + R^2_G\right)^2 \geq 4 R^2_P R^2_G.$$

Subtracting the first inequality from the second and after simplification we have the following

$$R^2_G \geq R^2_P \tag{3.4}$$

This leads to the following relationship between R^2_{OLS}, R^2_P and R^2_G

$$R^2_G \geq R^2_P \geq R^2_{OLS} \tag{3.5}$$

4 Simulation Study

Let the underlying (continuous) random variable U be related with a set of covariates $\mathbf{X}$ through a linear model $U = a + \mathbf{b}'\mathbf{X} + \varepsilon$, where ε is an error term assumed to be *iid* $N(0,1)$. The usual coefficient of determinant $\rho^2 = 1 - E\left[\text{Var}(U|\mathbf{X})\right]/\text{Var}(U)$ measures the extent to which the covariates of interest explain the underlying outcome variable. U is unobservable in practice but quite often is represented by it's binary proxy Y such that $Y = 1$ if $U \geq c$ for some $c \in \mathbb{R}$, and $Y = 0$ otherwise. The relationship between Y and $\mathbf{X}$ is quite often modeled by the logistic model

$$\pi \equiv P(Y = 1|\mathbf{X}) = \exp(\beta_0 + \beta'\mathbf{X})/\left[1 + \exp(\beta_0 + \beta'\mathbf{X})\right].$$

If the estimation of the extent of variation in U due to $\mathbf{X}$ is of interest, it is desirable to compute a coefficient of determinant from the logistic model and use it as an estimate of ρ^2, the theoretical coefficient of determinant of the underlying linear model. Obviously the objective is then to select an R^2 analog that, along with other desirable criteria, is closest to the ρ^2 of the underlying model.

In our simulation study we considered linear regression models with a single covariate, which was generated from $N(0, \sigma_x^2)$. The regression parameters b and σ_x were selected in such a way that the underlying model would produce a given value of ρ^2. A continuous random variable U was generated for five levels of ρ^2 ranging from 0.1 to 0.9. Each U was then transformed to a binary dependent variable Y with probability of success $\bar{\pi}$, also termed as the base-rate, ranging from 0.05 to 0.5; in total 10 levels of $\bar{\pi}$ were considered. Pseudo R^2 statistics were then computed for each combination of ρ^2 and $\bar{\pi}$. The simulation experiment was replicated 10,000 times each for five sample sizes ranging from 250 to 2000.

The desired criteria of agreement of the Pseudo R^2 statistics with the ρ^2 of the underlying linear model was evaluated by comparing the means of the sampling distribution of the various R^2 statistics with the underlying ρ^2. To the extent that the means of the R^2 statistics depart from the ρ^2 they exhibit bias. The results are presented in Figure 1, for experimental conditions with $n = 250$ and $\bar{\pi} = 0.05, 0.2, 0.35$ and 0.5. Any deviation from the 45° line indicates bias.

All the statistics responded to the underlying ρ^2 to some degree. With the exception of R^2_{CS}, which was upwardly biased for estimating small ρ^2 at moderate to large $\bar{\pi}$, all other statistics underestimated the underlying ρ^2. The hierarchical order among R^2_{CS}, R^2_N, R^2_M and R^2_C derived in Section 3, is also evident from Figure 1 with R^2_C and R^2_M most severely underestimating the ρ^2. This is obviously due to the fact that the upperbounds of these two statistics are functions of the base-rate and cannot exceed 0.75 and 0.5802 for R^2_M and R^2_C, respectively. However, when scaled by their respective maximum this upperbound restriction was eliminated and the resulting statistics: R^2_{CS} and R^2_N provided much improved estimates of the underlying ρ^2. As noted in Section 3, R^2_{CS} provided the best estimates for the underlying ρ^2 followed by the R^2_N in all simulation conditions. Although R^2_L uniformly provided poorer estimates of ρ^2 as compared to R^2_N and R^2_{CS} at all experimental conditions, it was interesting to note that it performed better than the rest of the R^2 statistics at low

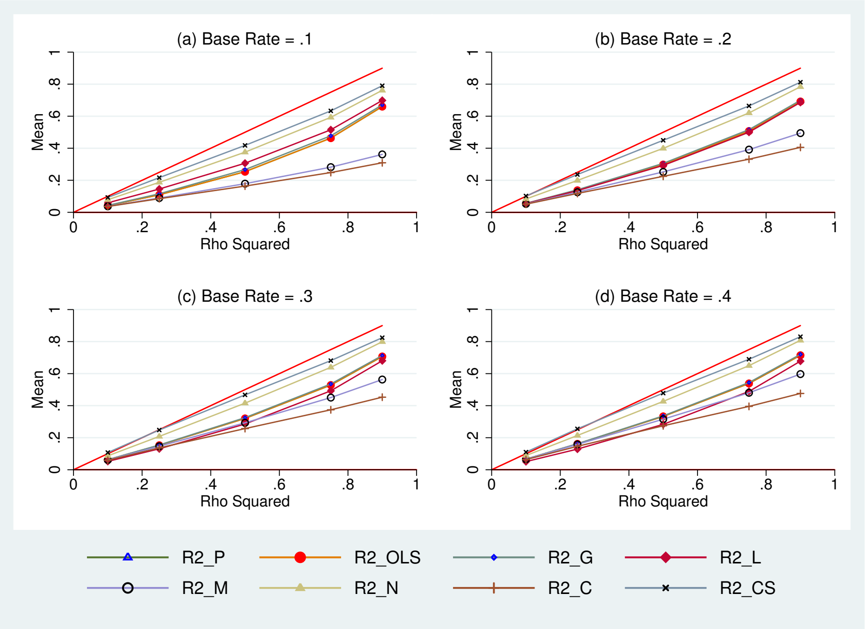

Figure 1: Means of R^2 Statistics by ρ^2 at $\bar{\pi} = 0.05, 0.2, 0.35, 0.5$; and $n = 250$.

base-rate conditions. When the proportion of successes are less than 20%, R^2_L performed better than the R^2_P, R^2_{OLS} and R^2_G at all level of ρ^2 , and than the R^2_M and R^2_C while estimating small ρ^2.

In general, we observed a greater degree of discrepancy between the actual and estimated value of moderately large ρ^2 (around 0.75) at small base-rate conditions, which tended to become narrower as the $\bar{\pi}$ approached to 0.5.

The ability of these statistics to predict underlying ρ^2 is evaluated in Table 1, which presents the *Mean Squared Errors* (MSE) for estimating the underlying ρ^2 of various R^2 statistics under selected experimental conditions. The R^2_{CS} appeared to have clear advantages over other R^2 statistics at most experimental conditions except when estimating small ρ^2 (e.g. in the neighborhood of 0.1); the R^2_N provided better estimates of ρ^2 in such situations irrespective to the base-rate. It is also noted that the R^2_L was ranked at third place (average rank 2.6 to 3.0) when $\bar{\pi} \leq 0.15$. However the relative performance of this R^2 statistic deteriorated noticeably when the base-rate was greater than 0.2 (average rank 6.6 to 7.2). In general, the MSE tends to decrease with increasing base-rate for all R^2 statistics, except the R^2_L, which shows a small but increasing tendency with the base-rate.

The predictive quality of these statistics is further investigated by comparing the MSE of the corresponding R^2 statistic to that of the R^2_{CS}, which produced the smallest MSE in

Table 1: Mean Squared Error Evaluation of R^2 Statistics by Selected Levels of ρ^2 and $\bar{\pi}$ at $n = 250$.

$\bar{\pi}$	ρ^2	R_P^2	R_{OLS}^2	R_G^2	R_L^2	R_M^2	R_N^2	R_C^2	R_{CS}^2
0.10	0.10	0.0043	0.0043	0.0041	0.0030	0.0044	**0.0027**	0.0045	0.0030
	0.25	0.0224	0.0226	0.0206	0.0139	0.0270	0.0085	0.0282	**0.0068**
	0.50	0.0663	0.0667	0.0596	0.0422	0.1049	0.0215	0.1142	**0.0133**
	0.75	0.0897	0.0900	0.0807	0.0605	0.2210	0.0299	0.2529	**0.0185**
	0.90	0.0640	0.0640	0.0582	0.0449	0.2918	0.0226	0.3500	**0.0149**
0.25	0.10	0.0026	0.0026	0.0026	0.0030	0.0027	**0.0019**	0.0027	0.0024
	0.25	0.0128	0.0128	0.0122	0.0153	0.0141	0.0053	0.0159	**0.0040**
	0.50	0.0387	0.0388	0.0363	0.0470	0.0521	0.0121	0.0669	**0.0055**
	0.75	0.0569	0.0569	0.0531	0.0681	0.1071	0.0172	0.1560	**0.0083**
	0.90	0.0427	0.0428	0.0397	0.0498	0.1348	0.0132	0.2186	**0.0078**
0.45	0.10	0.0020	0.0020	0.0020	0.0030	0.0020	**0.0017**	0.0021	0.0023
	0.25	0.0093	0.0093	0.0091	0.0160	0.0095	0.0041	0.0114	**0.0036**
	0.50	0.0298	0.0298	0.0288	0.0497	0.0343	0.0085	0.0505	**0.0037**
	0.75	0.0459	0.0460	0.0436	0.0710	0.0701	0.0117	0.1227	**0.0052**
	0.90	0.0363	0.0363	0.0340	0.0525	0.0882	0.0094	0.1759	**0.0056**

Smallest *MSE*s in each experimental condition are given in bold type.

most of the experimental conditions, except when $\rho^2 \leq 0.1$. This error ratio is termed as the *Relative Mean Squared Error* (*RMSE*), and can be interpreted as the average squared error loss of a R^2 statistic relative to that of the R_{CS}^2. *RMSE* of various R^2 statistics are plotted against $\bar{\pi}$ at selected levels of ρ^2 in Figure 2.

The relative losses in comparison with the R_{CS}^2 appeared to increase as the underlying ρ^2 increased and could be noticeably large. For example the average squared loss of R_C^2 was about 25 time larger than R_{CS}^2 and that of the R_L^2, the most commonly used R^2 statistic, is about 15 times larger when the $\rho^2 = 0.75$ and $\bar{\pi} = 0.5$. While the The $RMSE$ of R_L^2 tended to increase with $\bar{\pi}$, relative losses could rise or fall for other statistics with $\bar{\pi}$ depending upon the ρ^2. However, at $\rho^2 \geq 0.5$ the relative losses appeared to increase with $\bar{\pi}$ in general.

5 Summary

Many real life events, including many diseases are progressive in nature. Based on some predefined criteria scientists determine whether a particular event has occurred or not.

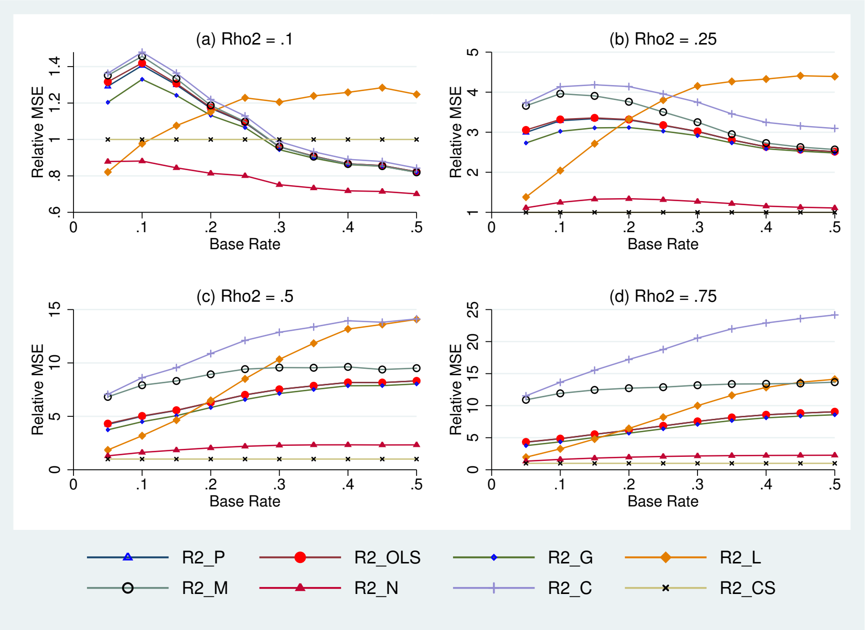

Figure 2: Relative Mean Squared Error of R^2 Statistics by $\bar{\pi}$ at $\rho^2 = 0.1, 0.25, 0.5, 0.75$; and $n = 250$.

Therefore, existence of a latent variable underlying a dichotomous variable Y, representing whether the event has occurred or not, is often a reasonable assumption to make. For example, suppose that a clinician conducts a Fasting Plasma Glucose (FPG) test to determine whether a subject is diabetic or not. Using the National Institute of Health classification for diabetes, he/she classifies the subjects as diabetic if the FPG test reveals a fasting blood glucose level equal or greater than 126 mg/dL. Further suppose that an applied researcher is interested in modeling the determinants of diabetes in the study population but he/she has only access to data on whether the subjects are diabetic or not. Assuming that data on covariates of interest are also recorded for these subjects, the researcher might want to model diabetes using logistic regression. However, his/her interest is to study blood glucose level *per se*, and not just whether someone is diabetic. Here the interest is on blood glucose level as a latent scale and the binary classification is just a crude measure of it. If the goal is to estimate the extent of variation in blood glucose level due to the set of covariates then it is desirable to use an R^2 statistic in logistic regression analysis that is numerically consistent with the ρ^2 of the underlying linear model. In other words, if the underlying ρ^2 exists, a "good" R^2 statistic should be able to estimate it reasonably well. Knowing the value of an R^2 statistic that is numerically close to the underlying ρ^2 helps to explain the strength of the relationship between the covariates and the latent variable underlying the observed binary dependent variable.

Using the Monte Carlo Simulations we find that while all the eight R^2 statistics are strongly correlated with the ρ^2, some of them severely underestimate it. R^2_{CS} provides the best estimate of the underlying ρ^2 : it has the smallest bias at all experimental conditions and the smallest *MSE* for estimating $\rho^2 > 0.1$. We also showed that the underestimation depends on the proportion of successes in the sample as well as ρ^2 of the underlying model. In particular, underestimation of the R^2_L increases as the base-rate increases. While the relative mean squared loss of R^2_L is quite small for estimating small ρ^2, it increases noticeably as the ρ^2 increases. We found R^2_N to be the next best predictor of ρ^2.

In this paper we also showed that the likelihood based R^2 statistics, some of which have been considered independent until recently, are not independent statistics. In fact there is a one-to-one correspondence between them and each of the likelihood based R^2 statistic can be presented as a function of R^2_L (or any other likelihood based R^2 statistic for that matter) and a parameter δ, which is a parabolic function of $\bar{\pi}$ in the range $[0, 1]$. There is, however, a clear hierarchical order among these R^2 statistics with R^2_{CS} being the largest in value followed by R^2_N. This hierarchical order can be summarized as below:

1. $R^2_C < R^2_M < R^2_N < R^2_{CS}$ irrespective of ρ^2 and $\bar{\pi}$;
2. For sufficiently small ρ^2,

$$\begin{array}{ll} R^2_L < R^2_M \text{ and } R^2_C & \text{when } .2 < \bar{\pi} < .8, \text{ and} \\ R^2_L > R^2_M \text{ and } R^2_C & \text{otherwise;} \end{array}$$

3. $R^2_L < R^2_N < R^2_{CS}$ irrespective of ρ^2 and $\bar{\pi}$;
4. And finally, $R^2_G \geq R^2_P \geq R^2_{OLS}$.

These results are also verified from simulation results.

The very existence of a plethora of R^2 statistics for logistic regression sometime creates confusion about which statistic to use to evaluate the worth of a logistic regression model. Use of R^2_L in logistic regression has become a standard practice and many researchers have recommended it (for example, Menard (2000); Liao and McGee (2003)). In this paper we have showed that R^2_{CS} deserve a serious consideration specially when it is reasonable to believe that a underlying latent variable exists and estimation of explained variation in the underlying dependent variable is of interest. This statistic provides valuable information regarding the strength of relationship between the covariates and the underlying latent variable, which R^2_L fails to provide.

References

[1] Agresti, A. (1986). Applying R^2-type measures to ordered categorical data. *Technometrics*, **28**, 133–138.

[2] Aldrich, J. and Nelson, F. (1984). *Linear Probability, Logit, and Probit Models*. Beverly Hills, CA: Sage.

[3] DeMaris, A. (2002). Explained variance in logistic regression. A Monte Carlo study of proposed measures. *Sociological Methods and Research*, **31**, 27–74.

[4] Efron, B. (1978). Regression and anova with zero-one data: Measures of residual variation. *Journal of the American Statistical Association*, **73**, 113–121.

[5] Haberman, S. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, **77**, 568–580.

[6] Hagle, T. and Mitchell, G. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, **36**, 762–784.

[7] Hu, B., Palta, M., and Shao, J. (2006). Properties of r^2 statistics for logistic regression. *Statistics in Medicine*, **25**, 1383–1395.

[8] Liao, J. and McGee, D. (2003). Adjusted coefficient of determination for logistic regression. *The American Statistician*, **73**, 161–165.

[9] Maddala, G. (1983). Limited-Dependent and Qualitative Variables in Economics. Cambridge, UK: Cambridge University Press.

[10] Magee, L. (1990). R^2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, **44**, 250–253.

[11] Menard, S. (2000). Coefficient of determination for multiple logistic regression analysis. *The American Statistician*, **54**, 17–24.

[12] Mittlböck, M. and Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, **15**, 1987–1997.

[13] Mittlböck, M. and Schemper, M. (2002). Explained variation for logistic regression - small sample adjustments, confidence intervals and predictive precision. *Biometrical Journal*, **44**, 263–272.

[14] Nagelkerke, N. (1991). A note on a general dentition of the coefficient of determination. *Biometrika*, **78**, 691–692.