Statistica Sinica **16**(2006), 847-860

PSEUDO-R^2 IN LOGISTIC REGRESSION MODEL

Bo Hu, Jun Shao and Mari Palta

University of Wisconsin-Madison

Abstract: Logistic regression with binary and multinomial outcomes is commonly used, and researchers have long searched for an interpretable measure of the strength of a particular logistic model. This article describes the large sample properties of some pseudo-R^2 statistics for assessing the predictive strength of the logistic regression model. We present theoretical results regarding the convergence and asymptotic normality of pseudo-R^2s. Simulation results and an example are also presented. The behavior of the pseudo-R^2s is investigated numerically across a range of conditions to aid in practical interpretation.

Key words and phrases: Entropy, logistic regression, pseudo-R^2

1. Introduction

Logistic regression for binary and multinomial outcomes is commonly used in health research. Researchers often desire a statistic ranging from zero to one to summarize the overall strength of a given model, with zero indicating a model with no predictive value and one indicating a perfect fit. The coefficient of determination R^2 for the linear regression model serves as a standard for such measures (Draper and Smith (1998)). Statisticians have searched for a corresponding indicator for models with binary/multinomial outcome. Many different R^2 statistics have been proposed in the past three decades (see, e.g., McFadden (1973), McKelvey and Zavoina (1975), Maddala (1983), Agresti (1986), Nagelkerke (1991), Cox and Wermuch (1992), Ash and Shwartz (1999), Zheng and Agresti (2000)). These statistics, which are usually identical to the standard R^2 when applied to a linear model, generally fall into categories of entropy-based and variance-based (Mittlböck and Schemper (1996)). Entropy-based R^2 statistics, also called pseudo-R^2s, have gained some popularity in the social sciences (Maddala (1983), Laitila (1993) and Long (1997)). McKelvey and Zavoina (1975) proposed a pseudo-R^2 based on a latent model structure, where the binary/ multinomial outcome results from discretizing a continuous latent variable that is related to the predictors through a linear model. Their pseudo-R^2 is defined as the proportion of the variance of the latent variable that is explained by the covariate. McFadden (1973) suggested an alternative, known as "likelihood-ratio index", comparing a model without any predictor to a model including all predictors. It is defined as one minus the ratio of the log likelihood with intercepts only, and the log likelihood with all predictors. If the slope parameters are all 0, McFadden's R^2 is 0, but it is never 1. Maddala (1983) developed another pseudo-R^2 that can be applied to any model estimated by the maximum likelihood method. This popular and widely used measure is expressed as

$$R_M^2 = 1 - \left(\frac{L(\tilde{\theta})}{L(\hat{\theta})} \right)^{\frac{2}{n}}, \tag{1}$$

where $L(\tilde{\theta})$ is the maximized likelihood for the model without any predictor and $L(\hat{\theta})$ is the maximized likelihood for the model with all predictors. In terms of the likelihood ratio statistic $\lambda = -2\log(L(\tilde{\theta})/L(\hat{\theta}))$, $R_M^2 = 1 - e^{-\lambda/n}$. Maddala proved that R_M^2 has an upper bound of $1 - (L(\tilde{\theta}))^{2/n}$ and, thus, suggested a normed measure based on a general principle of Cragg and Uhler (1970):

$$R_N^2 = \frac{1 - \left(\frac{L(\tilde{\theta})}{L(\hat{\theta})} \right)^{\frac{2}{n}}}{1 - (L(\tilde{\theta}))^{\frac{2}{n}}}. \tag{2}$$

While the statistics in (1) and (2) are widely used, their statistical properties have not been fully investigated. Mittlböck and Schemper (1996) reviewed R_M^2 and R_N^2 along with other measures, but their results are mainly empirical and numerical. The R^2 for the linear model is interpreted as the proportion of the variation in the response that can explained by the regressors. However, there is no clear interpretation of the pseudo-R^2s in terms of variance of the outcome in logistic regression. Note that both R_M^2 and R_N^2 are statistics and thus random. In linear regression, the standard R^2 converges almost surely to the ratio of the variability due to the covariates over the total variability as the sample size increases to infinity. Once we know the limiting values of R_M^2 and R_N^2, these limits can be similarly used to understand how the pseudo-R^2s capture the predictive strength of the model. The pseudo-R^2s for a given data set are point estimators for the limiting values that are unknown. To account for the variability in estimation, it is desirable to study the asymptotic sampling distributions of R_M^2 and R_N^2, which can be used to obtain asymptotic confidence intervals for the limiting values of pseudo-R^2s. Helland (1987) studied the sampling distributions of R^2 statistics in linear regression.

In this article we study the behavior of R_M^2 and R_N^2 under the logistic regression model. In Section 2, we derive the limits of R_M^2 and R_N^2 and provide interpretations of them. We also present some graphs describing the behavior of R_N^2 across a range of practical situations. The asymptotic distributions of R_M^2 and R_N^2 are derived in Section 3 and some simulation results are presented. An example is given in Section 4.

2. What Does Pseudo-R^2 Measure

In this section we explore the issue of what R_M^2 in (1) and R_N^2 in (2) measure in the setting of binary or multinomial outcomes.

2.1. Limits of pseudo-R^2s

Consider a study of n subjects whose outcomes fall in one of m categories. Let $Y_i = (Y_{i1}, \ldots, Y_{im})'$ be the outcome vector associated with the ith subject, where $Y_{ij} = 1$ if the outcome falls in the jth category, and $Y_{ij} = 0$ otherwise. We assume that $Y_1, \ldots, Y_n$ are independent and that Y_i is associated with a p-dimensional vector X_i of predictors (covariates) through the multinomial logit model

$$P_{ij} = E(Y_{ij}|X_i) = \frac{\exp(\alpha_j + X_i'\beta_j)}{\sum_{k=1}^m \exp(\alpha_k + X_i'\beta_k)}, \qquad j = 1, \ldots, m, \tag{3}$$

where $\alpha_m = \beta_m = 0$, $\alpha_1, \ldots, \alpha_{m-1}$ are unknown scalar parameters, and $\beta_1, \ldots,$ β_{m-1} are unknown p-vectors of parameters. Let θ be the $(p+1)(m-1)$ dimensional parameter $(\alpha_1, \beta_1', \ldots, \alpha_{m-1}, \beta_{m-1}')$. Then the likelihood function under the multinomial logit model can be written as

$$L(\theta) = \prod_{i=1}^n P_{i1}^{Y_{i1}} P_{i2}^{Y_{i2}} \cdots P_{im}^{Y_{im}}. \tag{4}$$

Procedures for obtaining the maximum likelihood estimator $\hat\theta$ of θ are available in most statistical software packages. The following theorem provides the asymptotic limits of the pseudo-R^2s defined in (1) and (2). Its proof is given in the Appendix.

Theorem 1. *Assume that covariates X_i, $i = 1, \ldots, n$, are independent and identically distributed random p-vectors with finite second moment. If*

$$H_1 = -\sum_{j=1}^m E(P_{ij}) \log E(P_{ij}), \tag{5}$$

$$H_2 = -\sum_{j=1}^m E(P_{ij} \log P_{ij}), \tag{6}$$

then, as $n \to \infty$, $R_M^2 \to_p 1 - e^{2(H_2 - H_1)}$ and $R_N^2 \to_p (1 - e^{2(H_2 - H_1)})/(1 - e^{-2H_1})$, where $\to_p$ denotes convergence in probability.

2.2. Interpretation of the limits of pseudo-R^2s

It is useful to consider whether the limits of pseudo-R^2 can be interpreted much as R^2 can be for linear regression analysis.

Theorem 1 reveals that both R_M^2 and R_N^2 converge to limits that can be described in terms of entropy. If the covariates X_is are i.i.d., $Y_i = (Y_{i1}, \ldots, Y_{im})'$, $i=1, \ldots, n$, are also i.i.d. multinomial distributed with probability vector $(E(P_{i1}), \ldots, E(P_{im}))$ where the expectation is taken over X_i. Then H_1 given in (5) is exactly the entropy measuring the marginal variation of Y_i. Similarly, $-\sum_{j=1}^{m} P_{ij} \log P_{ij}$ corresponds to the conditional entropy measuring the variation of Y_i given X_i and H_2 can be considered as the average conditional entropy. Therefore $H_1 - H_2$ measures the difference in entropy explained by the covariate X, which is always greater than 0 by Jensen's inequality, and is 0 if and only if the covariates and outcomes are independent. For example, when (X_i, Y_i) is bivariate normal, $H_1 - H_2 = \log(\sqrt{1-\rho^2})^{-1}$ where ρ is the correlation coefficient, and the limit of R_M^2 is ρ^2.

The limit of R_M^2 is $1 - e^{-2(H_1 - H_2)}$ monotone in increasing $H_1 - H_2$. Then we can write the limit of R_N^2 as the limit of R_M^2 divided by its upper bound:

$$R_N^2 \to_p \frac{1 - e^{-2(H_1 - H_2)}}{1 - e^{-2H_1}} = \frac{e^{2H_1} - e^{2H_2}}{e^{2H_1} - 1}.$$

When both H_1 and H_2 are small, $1 - e^{-2(H_1 - H_2)} \approx 2(H_1 - H_2)$, $1 - e^{-2H_1} \approx 2H_1$ and the limit of R_N^2 is approximately $(H_1 - H_2)/H_1$, the entropy explained by the covariates relative to the marginal entropy H_1.

2.3. Limits of R_M^2 and R_N^2 relative to model parameters

For illustration, we examine the magnitude of the limits of R_M and R_N under different parameter settings when the X_is are i.i.d. standard normal and the outcome is binary. Figures 1 and 2 show the relationship between the limits of R_M^2 and R_N^2 and the parameters α and β. In these figures, profile lines of the limits are given for different levels of the response probability $e^\alpha/(1+e^\alpha)$ at the mean of X_i and odds ratio e^β per standard deviation of the covariate. The limits tend to increase as the absolute value of β increases with other parameters fixed, which is consistent with the behavior of the usual R^2 in linear regression models. However, we note that the limits tend to be low, even in models where the parameters indicate a rather strong association with the outcome. For example, a moderate size odds ratio of 2 per standard deviation of X_i is associated with the limit of R_N^2 at most 0.10. As the pseudo-R^2 measures do not correspond in magnitude to what is familiar from R^2 for ordinary regression, judgments about the strength of the logistic model should refer to profiles such as those provided in Figures 1 and 2. Knowing what odds ratio for a single predictor model produces the same pseudo-R^2 as a given multiple predictor model greatly facilitates subject matter relevance assessment.

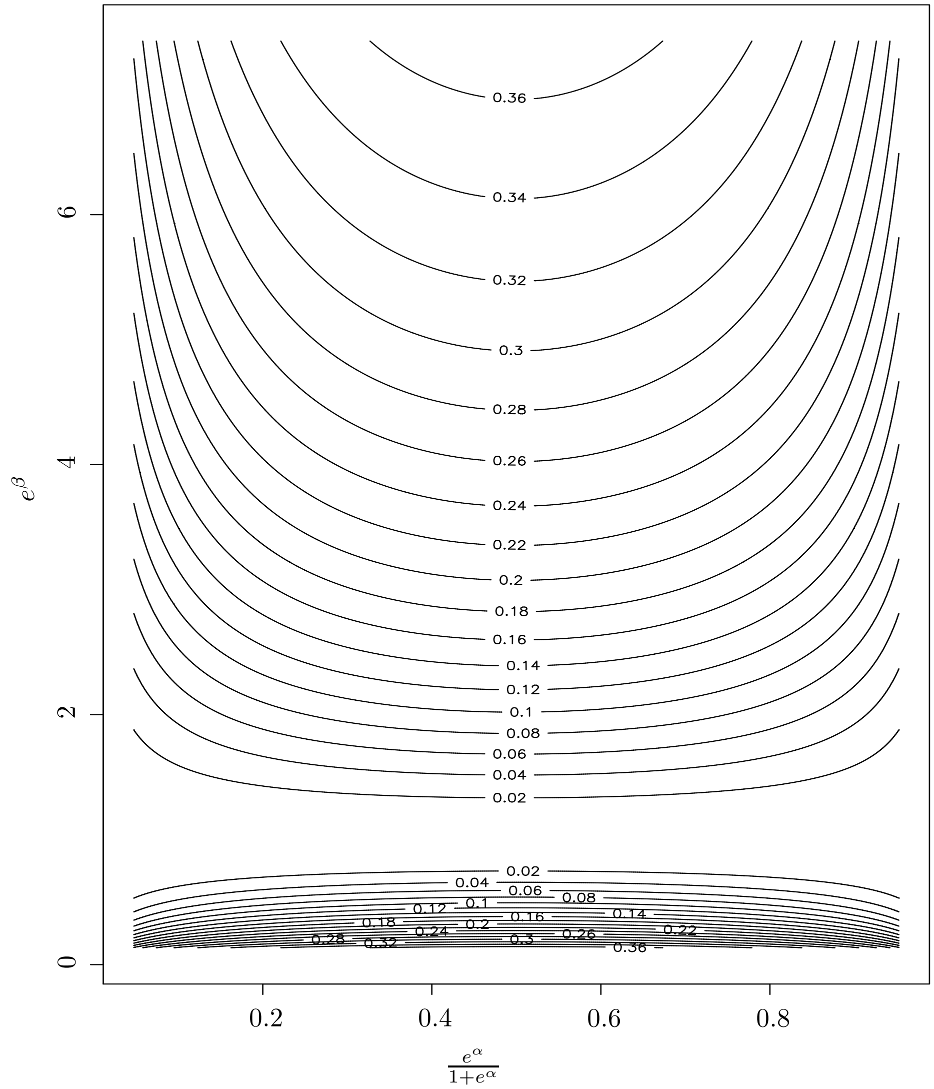

Figure 1. Contour plot of limits of R_M^2 against $e^\alpha/(1+e^\alpha)$ and odds ratio e^β.

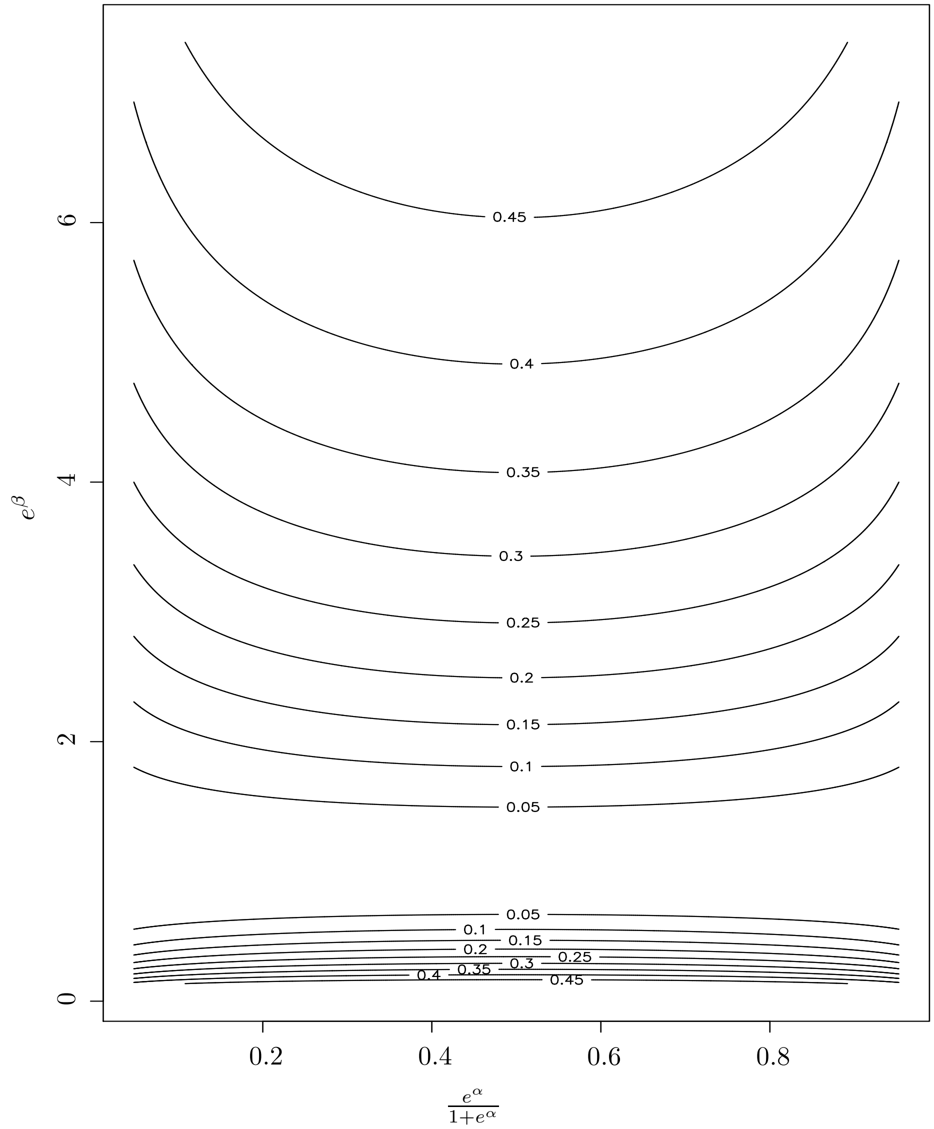

Figure 2. Contour plot of limits of R_N^2 against $e^\alpha/(1+e^\alpha)$ and odds ratio e^β.

It may be noted that neither R_N^2 nor R_M^2 can equal 1, except in degenerate models. This property is a logical consequence of the nature of binary outcomes. The denominator, $1-(L(\tilde{\theta}))^{2/n}$, equals the numerator when $L(\hat{\theta})$ equals 1, which occurs only for a degenerate outcome that is always 0 or 1. In fact, any perfectly fitting model for binary data would predict probabilities that are only 0 or 1. This constitutes a degenerate logistic model, which cannot be fit. In comparison to the R^2 for a linear model, R^2 of 1 implies residual variance of 0. As the variance and entropy of binomial and multinomial data depend on the mean, this again can occur only when the predicted probabilities are 0 and 1. The mean-entropy dependence influences the size of the pseudo-R^2s and tends to keep them away from 1 even when the mean probabilities are strongly dependent on the covariate.

For ease of model interpretation, investigators often categorize a continuous variable, which leads to a loss of information. Consider a standard normally distributed covariate. We calculate the limit of R_N^2 when cutting the normal covariate into two, three, five or six categories. The threshold points we choose are 0 for two categories, ± 1 for three categories, $\pm 0.5, \pm 1$ for five categories, and $0, \pm 0.5, \pm 1$ for six categories. In Figures 3 and 4, we plot the corresponding limits of R_N^2 against $e^\alpha/(1+e^\alpha)$ by fixing β at 1, and against e^β by setting $\alpha = 1$. The fewer the number of categories we use for the covariate, the more information we lose, i.e., the smaller the limit of R_N^2. In this example, we note that using five or six categories retains most of the information provided by the original continuous covariate.

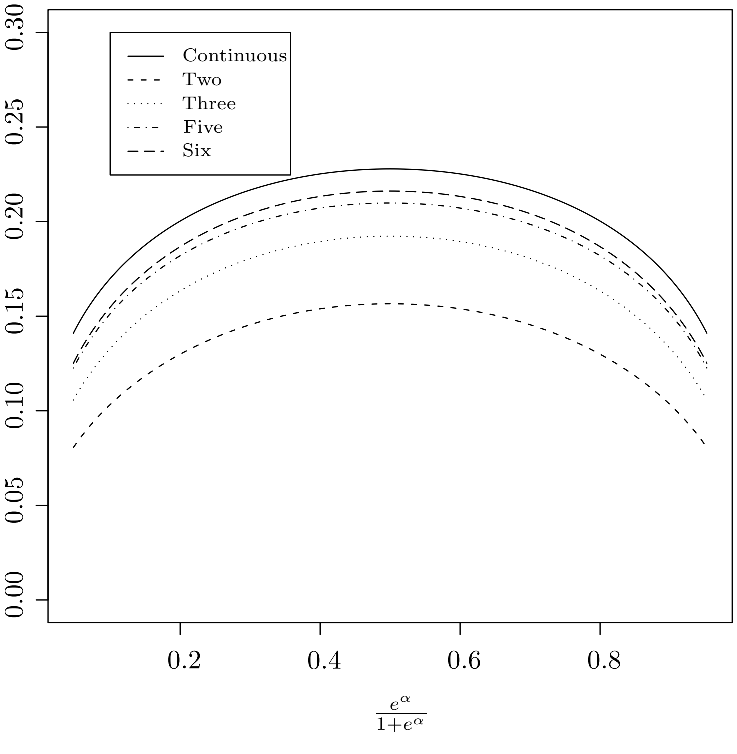

Figure 3. Limit of R_N^2 with covariate $N(0,1)$ dichotomized into K categories against $e^\alpha/(1+e^\alpha)$, $K = 2, 3, 5, 6$

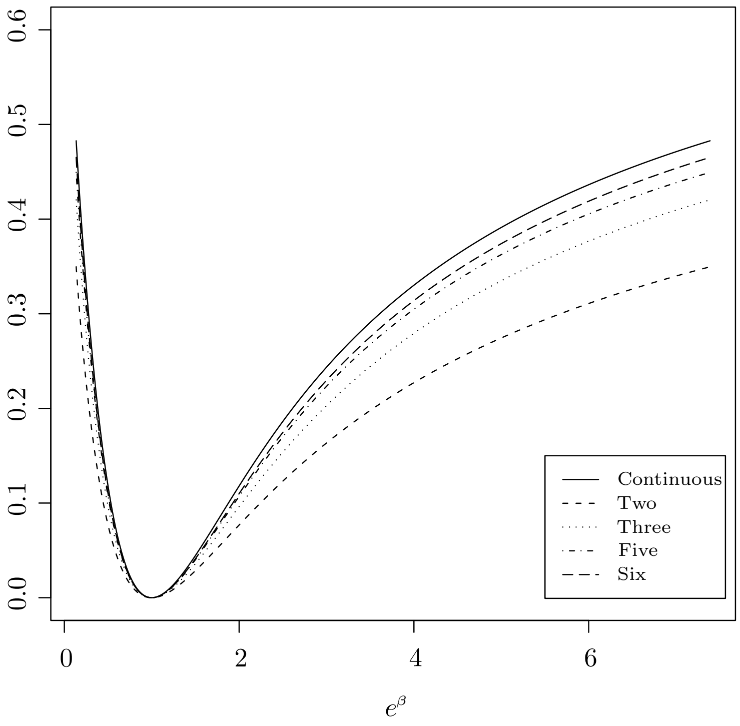

Figure 4. Limit of R_N^2 with covariate $N(0,1)$ dichotomized into K categories against odds ratio e^β, $K = 2,3,5,6$

3. Sampling Distributions of Pseudo-R^2s

The result in the previous section indicates that the limit of a pseudo-R^2 is a measure of the predictive strength of a model relating the logistic responses to some predictors (covariates). The quantities R_M^2 and R_N^2 are statistics and are random. They should be treated as estimators of their limiting values in assessing the model strength. In this section, we derive the asymptotic distributions of R_M^2 and R_N^2 that are useful for deriving large sample confidence intervals.

3.1. Asymptotic distributions of pseudo-R^2s

Theorem 2. *Under the conditions of Theorem* 1*,*

$$\sqrt{n}\left[R_M^2 - (1 - e^{2(H_2-H_1)})\right] \to_d N(0,\sigma_1^2) \tag{7}$$

$$\sqrt{n}\left[R_N^2 - \frac{1-e^{2(H_2-H_1)}}{1-e^{-2H_1}}\right] \to_d N(0,\sigma_2^2), \tag{8}$$

where H_1 and H_2 are given by (5) *and* (6), $\sigma_1^2 = g_1'\Sigma g_1$ *and* $\sigma_2^2 = g_2'\Sigma g_2$ *with*

$$g_1 = -2e^{2(H_2-H_1)}\left(1+\log\gamma_1,\ldots,1+\log\gamma_m,-1\right), \tag{9}$$

$$g_2 = \frac{e^{-2H_1}(1-e^{2H_2})}{(1-e^{-2H_1})^2}\left(1+\log\gamma_1,\ldots,1+\log\gamma_m,e^{2H_2}\frac{1-e^{2H_1}}{1-e^{2H_2}}\right), \tag{10}$$

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(Y_i) & \eta \\ \eta' & \epsilon \end{pmatrix}. \tag{11}$$

Here $\gamma_j = E_x(P_{ij})$, $j = 1,\ldots,m$, *is the expected probability that the outcome falls in jth category, the jth element of* η *is* $\eta_j = E_x(P_{ij}\log P_{ij}) + \gamma_j H_2$, *and* $\epsilon = \sum_{j=1}^m E_x\left(P_{ij}(\log P_{ij})^2\right) - H_2^2$.

When all the slope parameters β_j are 0 (i.e., X_i and Y_i are uncorrelated), both σ_1^2 and σ_2^2 are zero. g_1, g_2 and Σ can be estimated by replacing the unknown quantities, which are related to the covariate distribution, with consistent estimators. For example, γ can be estimated by $(\sum \hat{P}_{i1}/n,\ldots,\sum \hat{P}_{im}/n)'$.

Suppose g_k, $k = 1, 2$, and Σ are estimated by $\hat{g}_k$ and $\hat{\Sigma}$, respectively. Theorem 2 leads to the following asymptotic $100(1-\alpha)\%$ confidence interval for the limit of R_M^2:

$$\left(R_M^2 - Z_{\frac{\alpha}{2}}\hat{g}_1'\hat{\Sigma}\hat{g}_1, \quad R_M^2 + Z_{\frac{\alpha}{2}}\hat{g}_1'\hat{\Sigma}\hat{g}_1\right), \tag{12}$$

where Z_α is the $1-\alpha$ quantile of the standard normal distribution. A confidence interval for the limit of R_N^2 can be obtained by replacing R_M^2 and $\hat{g}_1$ in (12) with R_N^2 and $\hat{g}_2$, respectively. If the resulting lower limit of the confidence interval is below 0 or the upper limit is above 1, it is conventional to use the margin value of 0 or 1.

3.2. Simulation results

In this section, we examine by simulation the finite sample performance of the confidence intervals based on the asymptotic results derived in Section 3.1. Our simulation experiments consider the logistic regression model with binary outcome and a single normal covariate with mean 0 and standard deviation 1.

All the simulations were run with 3,000 replications of an artificially generated data set. In each replication, we simulated a sample of size 200 or 1,000 from the standard normal distribution as covariate vectors X, and simulated 200 or 1,000 binary outcomes according to success probability $\exp(\alpha+\beta X)/(1+\exp(\alpha+\beta X))$. Tables 1 and 2 show the results for different values of α and β.

In all the simulations with sample size 1,000, the estimated confidence intervals derived by Theorem 2 displayed coverage probability close to the expected level of 0.95. Coverage probability is less satisfactory with sample size 200 when the model is weak.

Table 1. Simulation average of pseudo-R^2s and 95% confidence intervals in the logit model with normal covariate (sample size=1,000).

α	β	R^2_M (limit)	CI (coverage*)	R^2_N (limit)	CI (coverage*)
2	0.5	0.028 (0.027)	(0.008,0.047) (0.930)	0.052 (0.050)	(0.015,0.088)(0.930)
	1	0.108 (0.103)	(0.069,0.137) (0.937)	0.178 (0.178)	(0.121,0.236) (0.940)
	2	0.298 (0.298)	(0.258,0.338) (0.929)	0.455 (0.454)	(0.396,0.513) (0.927)
1	0.5	0.047 (0.046)	(0.022,0.071) (0.937)	0.067 (0.066)	(0.032,0.103) (0.940)
	1	0.151 (0.150)	(0.113,0.189) (0.943)	0.213 (0.213)	(0.160,0.267) (0.945)
	2	0.351 (0.350)	(0.312,0.390) (0.922)	0.483 (0.483)	(0.430,0.536) (0.929)
0.5	0.5	0.054 (0.053)	(0.028,0.080) (0.940)	0.073 (0.072)	(0.038,0.109) (0.941)
	1	0.166 (0.166)	(0.127,0.204) (0.948)	0.224 (0.226)	(0.172,0.276) (0.948)
	2	0.367 (0.365)	(0.327,0.404) (0.931)	0.492 (0.491)	(0.443,0.546) (0.933)
0	0.5	0.056 (0.055)	(0.030,0.083) (0.938)	0.075 (0.074)	(0.039,0.111) (0.938)
	1	0.171 (0.171)	(0.133,0.210) (0.931)	0.229 (0.228)	(0.177,0.281) (0.933)
	2	0.371 (0.370)	(0.332,0.409) (0.925)	0.490 (0.494)	(0.443,0.546) (0.929)

*The relative frequency with which the intervals contain the true limit

Table 2. Simulation average of pseudo-R^2s and 95% confidence intervals in the logit model with normal covariate (sample size=200).

α	β	R^2_M (limit)	CI (coverage)	R^2_N (limit)	CI (coverage)
2	0.5	0.031 (0.027)	(0, 0.074) (0.912)	0.058 (0.050)	(0, 0.138) (0.922)
	1	0.107 (0.103)	(0.0310, 0.182) (0.918)	0.185 (0.178)	(0.0580, 0.312) (0.920)
	2	0.299 (0.298)	(0.2110, 0.387) (0.910)	0.458 (0.454)	(0.3290, 0.588) (0.913)
1	0.5	0.050 (0.046)	(0, 0.105) (0.915)	0.073 (0.066)	(0, 0.151) (0.914)
	1	0.152 (0.150)	(0.0680, 0.235) (0.928)	0.215 (0.213)	(0.0980, 0.332) (0.925)
	2	0.351 (0.350)	(0.2650, 0.437) (0.930)	0.484 (0.483)	(0.3660, 0.601) (0.932)
0.5	0.5	0.058 (0.053)	(0, 0.116) (0.919)	0.078 (0.072)	(0, 0.157) (0.920)
	1	0.168 (0.166)	(0.0820, 0.253) (0.927)	0.227 (0.226)	(0.1120, 0.343) (0.930)
	2	0.367 (0.365)	(0.2820, 0.452) (0.925)	0.494 (0.491)	(0.3790, 0.608) (0.925)
0	0.5	0.059 (0.055)	(0.0010, 0.118) (0.911)	0.079 (0.074)	(0.0010, 0.158) (0.912)
	1	0.174 (0.171)	(0.0880, 0.260) (0.928)	0.233 (0.228)	(0.1180, 0.347) (0.930)
	2	0.372 (0.370)	(0.2870, 0.457) (0.925)	0.496 (0.494)	(0.3830, 0.610) (0.922)

4. Example

We now turn to an example of logistic regression from Fox's (2001) text on fitting generalized linear models. This example draws on data from the 1976 U.S. Panel Study of Income Dynamics. There are 753 families in the data set with 8 variables. The variables are defined in Table 3. The logarithm of the wife's estimated wage rate is based on her actual earnings if she is in the labor force; otherwise this variable is imputed from other predictors. The definition of other variables is straightforward.

Table 3. Variables in the women labor force dataset.

Variable	Description	Remarks
lfp	wife's labor-force participation	factor: no,yes
k5	number of children ages 5 and younger	0-3, few 3's
k618	number of children ages 6 to 18	0-8, few > 5
age	wife's age in years	30-60, single years
wc	wife's college attendance	factor: no,yes
hc	husband's college attendance	factor: no,yes
lwg	log of wife's estimated wage rate	see text
inc	family income excluding wife's income	$\$1,000$s

We assume a binary logit model with no labor force participation as the baseline category. Other variables are treated as predictors in the model. The estimated model with all the predictors has the following form:

$$\log \frac{P}{1-P} = 3.18 - 1.47k5 - 0.07k618 - 0.06age + 0.81wc + 0.11hc + 0.61lwg - 0.03inc,$$

where P is the probability that the wife in the family is in the labor force. The variables $k618$ and hc are not statistically significant based on the likelihood-ratio test. Table 4 shows the values of R^2_M and R^2_N, as well as 95% confidence intervals of limits of R^2_M and R^2_N, for the model containg all the predictors, and models excluding certain predictors.

Table 4. R^2_M and R^2_N with 95% confidence intervals of models for women labor force data.

Model	R^2_M (95% CI.)		R^2_N (95% CI.)	
Use all predictors	0.152	(0.109, 0.195)	0.205	(0.147, 0.262)
Exclude $k5$	0.074	(0.040, 0.108)	0.100	(0.054, 0.145)
Exclude *age*	0.123	(0.083, 0.164)	0.165	(0.111, 0.219)
Exclude *wc*	0.138	(0.096, 0.180)	0.185	(0.129, 0.241)
Exclude *lwg*	0.133	(0.092, 0.174)	0.179	(0.123, 0.234)
Exclude *inc*	0.130	(0.087, 0.172)	0.175	(0.119, 0.230)
Exclude $k618$	0.151	(0.108, 0.194)	0.203	(0.145, 0.261)
Exclude *hc*	0.152	(0.109, 0.195)	0.204	(0.146, 0.262)
Use $k618$, *hc* only	0.003	(-0.005, 0.010)	0.004	(-0.006, 0.013)

For the model with all the covariates, R^2_M and R^2_N are around 0.15 and 0.20, respectively. The results imply a moderately strong model when referencing the odds ratio scale equivalents in Figure 1. Dropping a significant covariate results in a notable decrease in the values of pseudo-R^2s, while no significant change occurs if we drop the insignificant covariates. R^2_M and R^2_N are near zero when we exclude all the significant covariates. However, model selection procedures using pseudo-R^2 need further research.

Acknowledgements

The research work is supported by Grant CA-53786 from the National Cancer Institute. The authors thank the referees and an editor for helpful comments.

Appendix

For the proof of results in Section 3, we begin with a lemma and then sketch the main steps for Theorem 1 and 2.

Lemma 1. *Assume that covariates X_i, $i = 1, \ldots, n$, are i.i.d. random p-vectors with finite second moment, then $(\log L(\hat{\theta}) - \log L(\theta))/\sqrt{n} \rightarrow_p 0$, where $\hat{\theta}$ is the maximum likelihood estimator of θ.*

Proof of Lemma 1. We first prove that $\partial^2 \log L(\theta)/\partial\theta\partial\theta^{'} = O_p(n)$. The score function is

$$\frac{\partial \log L(\theta)}{\partial\theta} = \left(\sum_{i=1}^{n} (Y_{i1} - P_{i1}), \sum_{i=1}^{n} (Y_{i1} - P_{i1}) X_i^{\prime}, \ldots, \sum_{i=1}^{n} (Y_{im} - P_{im}) X_i^{\prime} \right)^{\prime}.$$

Let $\eta_k = (\alpha_k, \beta_k^{\prime})^{\prime} \in \mathcal{R}^{p+1}$ for $k = 1, \ldots, m$, and $U_i = (1, X_i^{\prime})^{\prime}$. Then

$$\frac{\partial^2 \log L(\theta)}{\partial\eta_k \partial\eta_k^{'}} = -\sum_{i=1}^{n} P_{ik}(1 - P_{ik}) U_i U_i^{\prime}, \qquad k = 1, 2, \ldots, m,$$
$$\frac{\partial^2 \log L(\theta)}{\partial\eta_k \partial\eta_l^{'}} = -\sum_{i=1}^{n} P_{ik} P_{il} U_i U_i^{\prime}, \qquad k \neq l.$$

Since $U_i U_i^{\prime} = \begin{pmatrix} 1 & X_i^{\prime} \\ X_i & X_i X_i^{\prime} \end{pmatrix}$, each element in the second derivative matrix $\partial^2 \log L(\theta)/\partial\theta\partial\theta^{'}$ is $O_p(n)$ by assumption. For simplicity, we write this as $\partial^2 \log L(\theta)/\partial\theta\partial\theta^{'} = O_p(n)$. Let $S_n(\theta) = \partial \log L(\theta)/\partial\theta$, $J_n(\theta) = -\partial^2 \log L(\theta)/\partial\theta\partial\theta^{'}$ and $I_n(\theta) = E(J_n(\theta))$, where the expectation is taken over covariates. It follows that the cumulative information matrix $I_n(\theta) = O_p(n)$. By a second-order Taylor expansion,

$$\begin{aligned} \frac{\log L(\hat{\theta}) - \log L(\theta)}{\sqrt{n}} &= \frac{S_n(\hat{\theta})^{'}}{\sqrt{n}} (\hat{\theta} - \theta) - \frac{1}{2\sqrt{n}} (\hat{\theta} - \theta)^{'} J_n(\theta^*) (\hat{\theta} - \theta) \\ &= (\hat{\theta} - \theta)^{'} I_n(\theta)^{\frac{1}{2}} I_n(\theta)^{-\frac{1}{2}} \sqrt{n} \frac{J_n(\theta^*)}{2n^{\frac{3}{2}}} \sqrt{n} I_n(\theta)^{-\frac{1}{2}} I_n(\theta)^{\frac{1}{2}} (\hat{\theta} - \theta), \end{aligned}$$

where θ^* is a vector between θ and $\hat{\theta}$. The asymptotic normality results of the MLE gives $I_n(\theta)^{1/2}(\hat{\theta}-\theta) \to N(0,\mathbf{1})$. The lemma then follows from the fact that $I_n(\theta)^{-1/2}\sqrt{n} = O_p(1)$ and $J_n(\theta^*)/n = O_p(1)$.

Proof of Theorem 1. Let $f(x) = \log(1-x)/2$, then

$$\begin{aligned} f(R^2) &= \frac{1}{n}\log L(\tilde{\theta}) - \frac{1}{n}\log L(\hat{\theta}) \\ &= \sum_{j=1}^{m} \frac{n_j}{n}\log(\frac{n_j}{n}) - \frac{1}{n}\log L(\theta) + \frac{1}{n}\Big(\log L(\theta) - \log L(\hat{\theta})\Big). \end{aligned}$$

The convergence of $\sum_{j=1}^{m}(n_j/n)\log(n_j/n)$ and $\log L(\theta)/n$ come from the Law of Large Numbers. The results of the theorem follow from the lemma and the Continuous Mapping Theorem.

Proof of Theorem 2. Let $S_M^2 = 1-(L(\tilde{\theta})/L(\theta))^{2/n}$ and $S_N^2 = (1-(L(\tilde{\theta})/L(\theta))^{2/n})/(1-(L(\tilde{\theta}))^{2/n})$. It follows from the lemma that S_M^2 and S_N^2 have the same asymptotic distribution as R_M^2 and R_N^2, respectively, in the sense that $\sqrt{n}(S_M^2 - R_M^2) \to_p 0$ and $\sqrt{n}(S_N^2 - R_N^2) \to_p 0$.

Define $Z_i = (Y_{i1},\ldots,Y_{im},W_i)$ where $W_i = \sum_{j=1}^{m} Y_{ij}\log P_{ij}$. Then Z_i's form a i.i.d. random sequence with $\mu = E(Z_i) = (\gamma', \sum_{j=1}^{m} E(P_{1j}\log P_{1j})) = (\gamma', -E_2)$, $\mathrm{Cov}\,(Z_i) = \Sigma$, γ and Σ as defined in Section 3. By the Multidimensional Central Limit Theorem,

$$\sqrt{n}\left(\bar{Z}-\mu\right) \to N(0,\Sigma). \tag{13}$$

Let $\phi_1(x_1,\ldots,x_m) = 1 - e^{2(\sum_{j=1}^{m} x_j\log x_j - x_{m+1})}$ and $\phi_2(x_1,\ldots,x_m) = (1-e^{2(\sum_{j=1}^{m} x_j\log x_j - x_{m+1})})/(1-e^{2\sum_{j=1}^{m} x_j \log x_j})$. Applying the delta-method with ϕ_1 and ϕ_2 to (13), respectively, leads to the asymptotic normality results in Theorem 2.

References

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

Ash, A. and Shwartz, M. (1999). R^2: a useful measure of model performance when predicting a dichotomous outcome. *Statist. Medicine* **18**, 375-384.

Cox, D. R. and Wermuch, N. (1992). A comment on the coefficient of determination for binary responses. *Amer. Statist.* **46**, 1-4.

Cragg, J. G. and Uhler, R. S. (1970). The demand for automobiles. *Canad. J. Economics* **3**, 386-406.

Draper, N. R. and Smith, H. (1998). *Applied Rregression Analysis.* 3rd edition. Wiley, New York.

Fox, J. (2001). *An R and S-Plus Companion to Applied Regression.* Sage Publications.

Helland, I. S. (1987). On the interpretation and use of R^2 in regression analysis, *Biometrics* **43**, 61-69.

Laitila, T. (1993). A pseudo-R^2 measure for limited and qualitative dependent variable models. *J. Econometrics* **56**, 341-356.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications.

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

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (Edited by P. Zarembka), 105-42. Academic Press, New York.

McKelvey, R. D. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *J. Math. Soc.* **4**, 103-120.

Mittlböck M. and Schemper, M. (1996). Explained variation for logistic regression. *Statist. Medicine* **15**, 1987-1997.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691-693.

Zheng B. Y. and Agresti, A. (2000). Summarizing the predictive power of a generalized liner model. *Statist. Medicine* **19**, 1771-1781.

Department of Statistics, University of Wisconsin-Madison, Madison, WI, 53706, U.S.A.

E-mail: bohu@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI, 53706, U.S.A.

E-mail: shao@stat.wise.edu

Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI, 53706, U.S.A.

E-mail: mpalta@wisc.edu

(Received August 2004; accepted July 2005)