Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 43

5-1-2016

JMASM36: Nine Pseudo R^2 Indices for Binary Logistic Regression Models (SPSS)

David A. Walker
Northern Illinois University, dawalker@niu.edu

Thomas J. Smith
Northern Illinois University, tjsmith@niu.edu

DOI: 10.22237/jmasm/1462077720

JMASM36: Nine Pseudo R^2 Indices for Binary Logistic Regression Models (SPSS)

Erratum
This paper was originally published in JMASM Algorithms & Code without its enumeration, JMASM36.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 848-854.

ISSN 1538 − 9472

JMASM Algorithms and Code
Nine Pseudo R^2 Indices for Binary Logistic Regression Models

David A. Walker
Northern Illinois University
DeKalb, IL

Thomas J. Smith
Northern Illinois University
DeKalb, IL

This syntax program is an applied complement to Veall and Zimmermann (1994), Menard (2000), and Smith and McKenna (2013) and produces nine pseudo R^2 indices, not readily accessible in statistical software such as SPSS, which are used to describe the results from binary logistic regression analyses.

Keywords: Binary logistic regression, R^2 indices, SPSS, syntax

Introduction

The subsequent syntax-based software program (Walker & Smith, 2015) is intended to provide an application for users interested in various pseudo R^2 indices used to describe the results obtained from fitting binary logistic regression models, but not freely obtainable in the current SPSS operational format. In logistic regression, pseudo R^2 indices proffer an indication of model fit, and are similar to variance accounted for metrics affiliated with ordinary least-squares (OLS) regression models such as R^2, R^2 adjusted, or eta squared. Although values of pseudo R^2 indices typically range from zero to unity, values for some indices can exceed 1.0.

The majority of the indices in the current program have been applied previously under various research conditions by Veall and Zimmermann (1994), Menard (2000), and Smith and McKenna (2013). This program is intended to be a software-based complement to these studies. Of the indices affiliated with the nine pseudo R^2 measures, only two are produced in SPSS: Cox and Snell's (1989) and Nagelkerke's (1991).

The Cox and Snell index is represented as

Dr. Walker is a Professor of Educational Research and Assessment. Email him at: dawalker@niu.edu.

$$R^2_{\text{CS}} = 1 - \left(\frac{\text{L}(\text{Null})}{\text{L}(\text{Full})} \right)^{2/N} , \tag{1}$$

where L(Null) and L(Full) are the likelihood functions for the constant-only model and the model with the predictors, respectively, and *N* is the sample size. The Nagelkerke index, which is a "corrected" version of the Cox and Snell index in the sense that it constrains the index value so that it does not exceed 1.0, is expressed as

$$R^2_{\text{N}} = \frac{1 - \left(\text{L}(\text{Null}) \Big/ \text{L}(\text{Full}) \right)^{2/N}}{1 - \text{L}(\text{Null})^{2/N}} , \tag{2}$$

where the rescaling is accomplished by dividing Cox and Snell's index by its maximum possible value.

The remaining seven indices produced in the program are not produced in the default logistic regression output provided by SPSS. The McFadden (1974) and the McFadden adjusted metrics are stated as, respectively,

$$R^2_{\text{MF}} = 1 - \frac{\text{LL}(\text{Full})}{\text{LL}(\text{Null})} \tag{3}$$

and

$$R^2_{\text{MFA}} = 1 - \frac{\text{LL}(\text{Full}) - (K+1)}{\text{LL}(\text{Null})} , \tag{4}$$

where LL(Full) is the log-likelihood value for the model containing the predictors and LL(Null) is the log-likelihood value for the constant-only model. The latter index "adjusts" (penalizes) for the number of predictors (*K*) in the model. The Aldrich and Nelson (1984) index is expressed as

$$R^2_{\text{AN}} = \frac{2\left[\text{LL}(\text{Null}) - \text{LL}(\text{Full}) \right]}{2\left[\text{LL}(\text{Null}) - \text{LL}(\text{Full}) \right] + N} . \tag{5}$$

The Veall and Zimmerman (1994) index, which is a "corrected" version of the Aldrich and Nelson index, is formulated as

$$R^2_{\text{VZ}} = \frac{2\left[\text{LL}\left(\text{Null}\right) - \text{LL}\left(\text{Full}\right)\right]}{2\left[\text{LL}\left(\text{Null}\right) - \text{LL}\left(\text{Full}\right)\right] + N} \cdot \frac{2\,\text{LL}\left(\text{Null}\right) - N}{2\,\text{LL}\left(\text{Full}\right)}\ , \tag{6}$$

where LL(Full) and LL(Null) are as defined previously, and *N* is the sample size. The Sapra (2004) index is represented as

$$R^2_{\text{S}} = 1 - \left(\frac{\text{LL}\left(\text{Max}\right) - \text{LL}\left(\text{Full}\right) + K + 1/2}{\text{LL}\left(\text{Max}\right) - \text{LL}\left(\text{Null}\right) + 1/2} \right), \tag{7}$$

where LL(Null), LL(Full), and *K* are as defined previously, and LL(Max) indicates the maximum possible log-likelihood value for the saturated model (typically zero, for most data and models). Finally, the Estrella (1998) formulates both unadjusted and adjusted metrics, respectively,

$$R^2_{\text{E}} = 1 - \left(\frac{\text{LL}\left(\text{Full}\right)}{\text{LL}\left(\text{Null}\right)} \right)^{\left(-2/N\right)\text{LL}\left(\text{Null}\right)} \tag{8}$$

and

$$R^2_{\text{EA}} = 1 - \left(\frac{\text{LL}\left(\text{Full}\right) - K}{\text{LL}\left(\text{Null}\right)} \right)^{\left(-2/N\right)\text{LL}\left(\text{Null}\right)}\ , \tag{9}$$

where LL(Null), LL(Full), *N*, and *K* are as defined previously.

Program

The data set used in the current SPSS syntax program is a randomly sampled subset (*n* = 200) of the 1982 High School and Beyond data (Raudenbush & Bryk, 2002). Data were obtained from Acock (2008). To have the outcome variable (i.e., Write) fit the profile of a binary measure within a logistic regression model, the values were recoded, per the aforementioned data example, as a dichotomized variable

called "HonComp" with values ≥ 60 = 1 or "Yes" and values < 60 = 0 or "No." The full model also included three predictor variables: Female (0 = Male; 1 = Female), Read, and Science (note that the latter two variables were continuous in measurement).

The program is shown below. In the space between BEGIN DATA and END DATA, the user would insert the null model's -2 Log Likelihood (LL) value (LLNull), the full model's -2 LL value (LLFull), the sample size (*N*), and the number of predictors (*K*) in the model. The input values in the example from the program are, in this order, 231.289, 160.236, 200, 3, 0. Note that the last value place, LLMAX, has a suggested value that is frequently fixed at 0 and will be left as such in this example.

```
Pseudo R2 Program
**************************************************************************
Copyright David A. Walker and Thomas J. Smith, 2015
Contact dawalker@niu.edu or tjsmith@niu.edu
Northern Illinois University, 204 Gabel, DeKalb, IL 60115
            **APA 6th Edition Citation**
Walker, D. A., & Smith, T. J. (2015). Nine pseudo R2 indices for binary
logistic regression
models [Computer program]. DeKalb, IL: Authors.
**************************************************************************.
DATA LIST LIST / LLNull LLFull (2F9.3) N K (2F8.0) LLMAX.
**************************************************************************
Between BEGIN DATA and END DATA below, put the Null Model's -2 Log
Likelihood (LL)
value (LLNull), the Full Model's -2 LL value (LLFull), the sample size
(N), and the number of
predictors (K) in the model Note: LLMAX, the last value in the data, is
typically 0 rather than
always defaulted to 0
**************************************************************************.
BEGIN DATA
231.289 160.236 200 3 0
END DATA.
COMPUTE L0 = LLNull/-2.
COMPUTE L1 = LLFull/-2.
COMPUTE L2 = ((LLNull-LLFull)/2)+L0.
```

```
COMPUTE CoxSnell =1-EXP(-1*(LLNull-LLFull)/N).
COMPUTE Nagelkerke = CoxSnell/(1-EXP(-LLNull/N)).
COMPUTE McFadden = 1-(LLFull/LLNull).
COMPUTE McFaddenAdj = 1-((L1-(K+1))/L0).
COMPUTE Sapra = 1-((LLMax-LLFull+(K+1)/2)/(LLMax-LLNull+.5)).
COMPUTE AldrichNelson = ((LLNull-LLFull)/((LLNull-LLFull)+N)).
COMPUTE VeallZimmermann = AldrichNelson*((LLNull+N)/LLNull).
COMPUTE Estrella = 1-(L2/L0)**((-2/N)*L0).
COMPUTE EstrellaAdj = 1-((L2-K)/L0)**((-2/N)*L0).
EXECUTE.
FORMAT CoxSnell TO EstrellaAdj (F9.3).
VARIABLE LABELS CoxSnell 'Cox & Snell R2'/ Nagelkerke 'Nagelkerke R2'/
McFadden 'McFadden R2'/
VeallZimmermann 'Veall & Zimmermann R2'/Sapra 'Sapra R2'/McFaddenAdj
'McFadden Adjusted R2'/
AldrichNelson 'Aldrich & Nelson R2'/ Estrella 'Estrella R2'/ EstrellaAdj
'Estrella Adjusted R2'/.
REPORT FORMAT=LIST AUTOMATIC ALIGN (CENTER)
  /VARIABLES= CoxSnell Nagelkerke McFadden McFaddenAdj AldrichNelson
Sapra VeallZimmermann Estrella EstrellaAdj
  /TITLE "Pseudo R Squared Indices".
```

Results

As a simple, one-shot comparison, the values of pseudo R^2 obtained by applying the program to the High School and Beyond data, recorded in Table 1, indicated that seven of the nine indices were much lower in value than the R^2 (0.522) or the R^2 adjusted (0.515) values computed from an OLS model using the same predictors as the logistic regression model and with the non-dichotomized, continuous outcome variable (i.e., Write). The aforementioned pseudo R^2 values ranged from a minimum of 0.262 (Aldrich and Nelson) to a maximum of 0.489 (Veall and Zimmermann). It should be noted, though, that the lower values of these indices compared to OLS R^2 values may reflect, in part, less precision in the outcome due to the dichotomization of the continuous dependent variable for use in logistic regression. Cohen (1983) remarks on of the cost of engaging in such research practices where, "…the cost in the degradation of measurement due to dichotomization is a loss of one-fifth to two-thirds of the variance, and a concomitant loss of power…" (p. 253).

The values of the Nagelkerke and the Veall and Zimmermann indices, both of which are "corrected" indices, were noticeably similar (i.e., 0.436 and 0.489, respectively) to the OLS R^2 values. These indices' comparability in value to the OLS R^2 values was also found in Smith and McKenna (2013). Of interest is that the Smith and McKenna Monte Carlo simulation study, which included four continuous predictors, and the current findings indicate, potentially, that the Veall and Zimmermann pseudo R^2 index's highly favorable comparisons to OLS R^2 values may signify a robust nature within this index toward countering the full effect of dichotomizing an outcome variable in binary logistic regression modeling.

Table 1. Pseudo R^2 results

Cox-Snell	Nagelkerke	McFadden	McFadden Adj.	Aldritch-Nelson
0.299	0.436	0.307	0.273	0.626

Sapra	Veall-Zimmerman	Estrella	Estrella Adj.
0.314	.0489	0.346	0.317

References

Acock, A. C. (2008). *A gentle introduction to Stata* (2nd ed.). College Station, TX: Stata Press.

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Beverly Hills, CA: Sage.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253. doi: 10.1177/014662168300700301

Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). London, UK: Chapman and Hall.

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business & Economic Statistics, 16*(2), 198-205. doi: 10.1080/07350015.1998.10524753

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 104-142). New York, NY: Academic Press.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17-24. doi: 10.1080/00031305.2000.10474502

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692. doi: 10.2307/2337038

Raudenbush, S. W., & Bryk, A. S. (2002) *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Sapra, S. K. (2004). Comment by Sapra and reply. *The American Statistician, 58*(1), 90. doi: 10.1198/0003130042917

Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R^2 indices. *Multiple Linear Regression Viewpoints, 39*(2), 17-26. Retrieved from http://www.glmj.org/archives/articles/Smith_v39n2.pdf

Veall, M. R., & Zimmermann, K. F. (1994). Evaluating pseudo-R^2's for binary probit models. *Quality and Quantity, 28*(2), 151-164. doi: 10.1007/BF01102759

Walker, D. A., & Smith, T. J. (2015). Nine pseudo R^2 indices for binary logistic regression models [Computer program]. DeKalb, IL: Authors.