An Examination of Ordinal Regression Goodness-of-Fit Indices Under Varied Sample Conditions and Link Functions

Thomas J. Smith
Northern Illinois University

Cornelius M. McKenna
Kishwaukee College

This study examined how the values of three distinct R^2 analogs used in ordinal regression vary as functions of number of predictors, choice of link function, and distribution type. Results indicated that Nagelkerke's (1991) index most closely approximated OLS R^2, and the use of a complementary log-log link function resulted in particularly variant values of the R^2 analogs.

Ordinal regression is a procedure used in situations where one or more predictor variables are used to predict the probability of a particular outcome, and where that outcome variable of interest is manifest as a set of J ordered categories. Specifically, the most commonly-used application of ordinal regression, the cumulative odds model using a logit link, fits an equation of the form:

$$\ln\left(\frac{\hat{\lambda}_j}{1-\hat{\lambda}_j}\right) = \beta_{0i} + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_K x_K \,, \quad (1)$$

where x_1 to x_K are the K (continuous or categorical) predictor variables, and $\hat{\lambda}_j$ is the predicted probability of an outcome occurring at or below category j. This equation is sometimes rephrased as

$$\ln\left(\frac{\hat{\lambda}_j}{1-\hat{\lambda}_j}\right) = \beta_{0i} - \left(\beta_1 x_1 + \beta_2 x_2 + \ldots \beta_K x_K\right), \quad (2)$$

so that the β_k (for k = 1, …, K) parameter estimates (i.e., the "location" parameter estimates) reflect the subsequent increase in the predicted odds associated with each unit increase of each predictor, and β_{0i} indicates the unique intercept ("threshold") associated with each of the J-1 categories.

The ordinal regression model is typically assessed for a given data set to determine whether it predicts significantly better than the threshold-only model. This is carried out by comparing the significance of the omnibus chi-square test of the model coefficients, which assesses the incremental decrease in the log-likelihood of the regression model containing the full set of predictor variables when it is compared to the model that contains only the intercept (threshold) terms, and determines whether the former significantly improves prediction over the latter. Although the omnibus chi-square statistic provides a formal test of whether the predictor variables add significant predictive capacity beyond the intercept-only model, it does not provide a true goodness-of-fit index, and spurious significance can result with large sample sizes. Goodness-of-fit of the ordinal regression model can be assessed using the Pearson statistic,

$$\chi^2 = \sum\sum \frac{\left(O_{jkl} - E_{jkl}\right)^2}{E_{jkl}} \,, \quad (3)$$

or the likelihood-ratio (i.e., "deviance") statistic,

$$\chi^2 = 2\sum\sum O_{jkl} \ln\left(\frac{O_{jkl}}{E_{jkl}}\right), \quad (4)$$

each of which is based on the discrepancies between the observed and expected frequencies at each combination of the ordinal outcome categories and predictor variable levels, and where non-statistically-significant results on the residual degrees of freedom indicate adequacy of model fit. When computing either the Pearson or deviance goodness-of-fit statistic, if zero frequencies occur in a substantial number of cells—as can occur, for example, with multiple predictors or the use of continuous predictors—the Pearson and deviance statistics do not provide stable estimates of goodness-of-fit (Agresti, 1984). Lipsitz, Fitzmaurice, and Molenberghs (1996) describe an alternative goodness-of-fit test, which might be considered an extension of the Hosmer-Lemeshow (2000) binary logistic regression goodness-of-fit procedure to ordinal models, whereby scores (s_j) are assigned to each of the J levels of the ordinal outcome (e.g., s_1 = 1, s_2 = 2, and s_3 = 3 for ordinal outcomes of "Poor," "Fair," and "Good."), and predicted mean scores, $\hat{\mu}_i = \sum_{j=1}^{J} s_j \hat{p}_{ij}$, computed for each case, where $\hat{p}_{ij}$ are the predicted probabilities from the ordinal regression for each of the J levels of the ordinal outcome. A set of G binary (0/1) indicators, I_{ig} are then constructed that reflect the deciles of these predicted mean scores. A second ordinal regression model,

$$\ln\left(\frac{\lambda_j}{1-\lambda_j}\right) = \beta_{0i} - \left(\beta_1 x_1 + \beta_2 x_2 + \ldots \beta_K x_K\right) + \sum_{g=1}^{G-1} I_{ig}\gamma_g \,, \quad (5)$$

is then fitted. If the original regression model (2) is appropriately specified, then $\gamma_1 = \gamma_2 = \ldots = \gamma_{G-1} = 0$, which can be tested via Wald statistics or likelihood ratio tests.

In addition, a number of "R^2 analog" goodness-of-fit indices have been developed for ordinal and logistic regression models that are intended as analogs of R^2 as used in ordinary least-squares (OLS) regression. McFadden (1974) proposed an index that is computed as one minus the ratio of the full-model log-likelihood to the intercept-only log-likelihood (1-[LL_{Full}/LL_{Null}]), and is considered perhaps the most direct analog to OLS R^2, at least in a conceptual sense. Cox and Snell (1989) propose computing an R^2 analog as (1-[L_{Null}/L_{Full}]$^{2/n}$), where L_{Null} is the value of the likelihood function for the intercepts-only (i.e., thresholds-only) model, L_{Full} is the corresponding value for the full model with all the predictors, and n is the sample size[†]. An issue with Cox and Snell's R^2 analog is that values can potentially exceed 1.0. Nagelkerke (1991), however, proposed a rescaling that divides this statistic by its maximum possible value (1-(L_{Null})$^{2/n}$), which results in a statistic that has an identical range to OLS R^2.

The goodness of fit for a specified regression model can be influenced by the specific link functions, and particular link functions may be preferred for specific distributional characteristics of the outcome variable. Norusis (2012) suggests that the choice of link function in an ordinal regression analysis should be driven by the distribution of the ordinal outcome. Specifically, the logit link is suggested for uniformly distributed outcomes, the complementary log-log link is suggested for distributions in which higher categories are more probable, the negative log-log link is suggested when smaller categories are more probable, the probit link is suggested when the underlying latent trait of the ordinal outcome is normally distributed, and the Cauchit link is suggested when many extreme values are present. Li and Duan (1989), however, suggest that the choice of link function has little effect on the estimated parameters. The IBM SPSS (Statistical Package for the Social Sciences) Statistics Information Center (2011) suggests that, in the absence of theoretical justification for a particular link function, model fit should be the guiding criterion. The purpose of the present study was to examine several goodness-of-fit indices used in ordinal regression under varying sample conditions and under the use of various, specified link functions.

Method

The present study used Monte Carlo techniques to examine and compare several goodness-of-fit indices used in ordinal regression. To accomplish the goals of the study, multiple samples of size of n = 200 were randomly drawn from a multivariate normal distribution, and where each sample consisted of five continuous variables with a specified correlational structure. One of these five variables was transformed to an ordinal outcome variable with four levels by binning the values using the procedures described below. We then carried out ordinal regression on each sample using this ordinal variable as the outcome measure and the remaining four continuous variables as predictors, and computed three distinct R^2 analogs (Cox & Snell's R^2, Nagelkerke's R^2, and McFadden's R^2). In addition, R^2 was computed using OLS regression, where the predictors were the same predictors used in the ordinal regression, and the outcome variable consisted of values for the original continuous variable generated in the simulation (prior to transformation to an ordinal outcome). We then examined how varying (1) the distribution of the ordinal outcome (i.e., uniform distribution, high categories more probable, or low categories more probable), as well as (2) the specified link function (i.e., logit, complementary log-log, negative log-log, probit, Cauchit) affected the resulting R^2 analogs. These simulations were then repeated using either two or three predictors in the regression model.

The simulated data values for five continuous variables (four "predictors" and one "outcome") were randomly generated from multivariate normal distribution with a specified correlational structure. Particularly, low correlation (r = .10) was specified among the predictors, and moderate correlation (r =

.50) was specified between each predictor and the outcome variable. An additional, ordinal outcome variable was then derived from fifth continuous "outcome" variable by binning this variable into four categories based on one of three experimental conditions: (1) specifying a uniform distribution (i.e., each category equally likely); (2) specifying a distribution in which lower categories were more probable (i.e., the categories, from low to high, contained 40%, 30%, 20%, and 10% of the data values, respectively); and (3) specifying a distribution in which higher categories were more probable (i.e., the categories, from low to high, contained 10%, 20%, 30%, and 40% of the data values, respectively). These data were then used as input for ordinal regression analyses with 1000 replications carried out for each condition.

Table 1. R^2 Analog Values for Simulated Data by Link Function: Ordinal Regression with Four Uniformly-Distributed Categorical Outcomes.

		Two predictors		Three predictors		Four predictors	
Statistic	Link Function	*M*	*SD*	*M*	*SD*	*M*	*SD*
Cox & Snell	Logit	.3942	.0494	.5490	.0421	.6803	.0317
Cox & Snell	Comp. Log-Log	.4043	.0910	.7557	.1350	.9335	.0148
Cox & Snell	Neg. Log-Log	.3749	.0525	.5296	.0483	.6650	.0334
Cox & Snell	Probit	.3957	.0492	.5490	.0424	.6860	.0410
Cox & Snell	Cauchit	.3563	.0522	.5060	.0471	.6509	.0368
McFadden	Logit	.1830	.0296	.2904	.0340	.4153	.0358
McFadden	Comp. Log-Log	.1934	.0684	.5739	.2244	.9881	.0522
McFadden	Neg. Log-Log	.1717	.0305	.2755	.0373	.3985	.0353
McFadden	Probit	.1839	.0296	.2904	.0345	.4236	.0518
McFadden	Cauchit	.1610	.0296	.2574	.0346	.3837	.0379
Nagelkerke	Logit	.4209	.0527	.5862	.0491	.7264	.0338
Nagelkerke	Comp. Log-Log	.4317	.0971	.8069	.1442	.9968	.0158
Nagelkerke	Neg. Log-Log	.4003	.0561	.5655	.0516	.7101	.0357
Nagelkerke	Probit	.4226	.0525	.5862	.0453	.7325	.0438
Nagelkerke	Cauchit	.3804	.0557	.5402	.0503	.6951	.0392
OLS R^2	(None)	.4563	.0522	.6288	.0414	.7725	.0277

Table 2. R^2 Analog Values for Simulated Data by Link Function: Ordinal Regression with Four Positively Skewed Categorical Outcomes.

		Two predictors		Three predictors		Four predictors	
Statistic	Link Function	*M*	*SD*	*M*	*SD*	*M*	*SD*
Cox & Snell	Logit	.3823	.0508	.5282	.0433	.6573	.0349
Cox & Snell	Comp. Log-Log	.4960	.1703	.8856	.0752	.9213	.0052
Cox & Snell	Neg. Log-Log	.3636	.0528	.5120	.0461	.6450	.0380
Cox & Snell	Probit	.3854	.0506	.5312	.0437	.6942	.0744
Cox & Snell	Cauchit	.3296	.0550	.4825	.0482	.6177	.0401
McFadden	Logit	.1904	.0317	.2966	.0347	.4228	.0377
McFadden	Comp. Log-Log	.3013	.1778	.9008	.1734	.9998	.0044
McFadden	Neg. Log-Log	.1788	.0322	.2834	.0360	.4901	.0401
McFadden	Probit	.1924	.0317	.2992	.0355	.4812	.1204
McFadden	Cauchit	.1582	.0316	.2603	.0354	.3801	.0393
Nagelkerke	Logit	.4148	.0547	.5732	.0461	.7133	.0366
Nagelkerke	Comp. Log-Log	.5382	.1849	.9611	.0816	.9999	.0011
Nagelkerke	Neg. Log-Log	.3945	.0569	.5556	.0493	.7000	.0402
Nagelkerke	Probit	.4182	.0544	.5764	.0466	.7535	.0807
Nagelkerke	Cauchit	.3576	.0592	.5235	.0515	.6704	.0424
OLS R^2	(None)	.4563	.0522	.6288	.0414	.7725	.0277

Table 3. R^2 Analog Values for Simulated Data by Link Function: Ordinal Regression with Four Negatively Skewed Categorical Outcomes.

		Two predictors		Three predictors		Four predictors	
Statistic	Link Function	*M*	*SD*	*M*	*SD*	*M*	*SD*
Cox & Snell	Logit	.3858	.0502	.5369	.0446	.6637	.0343
Cox & Snell	Comp. Log-Log	.3840	.0699	.6420	.1207	.8833	.0604
Cox & Snell	Neg. Log-Log	.3632	.0533	.5192	.0475	.6508	.0373
Cox & Snell	Probit	.3872	.0502	.5387	.0444	.6675	.0377
Cox & Snell	Cauchit	.3389	.0539	.4922	.0501	.6270	.0398
McFadden	Logit	.1929	.0321	.3043	.0368	.4303	.0384
McFadden	Comp. Log-Log	.1932	.0491	.4303	.1534	.8818	.1568
McFadden	Neg. Log-Log	.1787	.0330	.2897	.0381	.4156	.0406
McFadden	Probit	.1938	.0322	.3058	.0367	.4353	.0448
McFadden	Cauchit	.1639	.0319	.2681	.0379	.3898	.0404
Nagelkerke	Logit	.4188	.0542	.5827	.0478	.7203	.0364
Nagelkerke	Comp. Log-Log	.4167	.0757	.6967	.1301	.9587	.0643
Nagelkerke	Neg. Log-Log	.3941	.0577	.5635	.0509	.7063	.0397
Nagelkerke	Probit	.4203	.0543	.5846	.0475	.7244	.0401
Nagelkerke	Cauchit	.3677	.0582	.5341	.0538	.6805	.0423
OLS R^2	(None)	.4563	.0522	.6288	.0414	.7725	.0277

Note: Simulations based on 1000 sample replications, with $n = 200$ for each sample.
OLS R^2 based on continuous, normally-distributed outcome.

Results

Table 1 shows descriptive statistics for each of three R^2 analog statistics (Cox & Snell's R^2, Nagelkerke's R^2, and McFadden's R^2) based on ordinal regression models with two, three, and four predictors and a uniformly-distributed ordinal outcome, and using each of five distinct link functions (logit, complementary log-log, negative log-log, probit, and Cauchit).

Tables 2 and 3 show descriptive statistics for R^2 analog values resulting from ordinal regression models fitted to asymmetric distributions (either positively or negatively skewed).

Figure 1 shows variation in the mean R^2 analog values for each of the three indices by number of predictors and aggregated across link function and distribution type. As these tables and figure indicate, values of McFadden's index were generally lower in value than corresponding values of Cox & Snell's or Nagelkerke's indices, as has consistently been observed for other generalized linear models (Menard, 2000). This observation was consistent across variation in the number of predictor variables. Nagelkerke's index appeared to most closely approximate the OLS R^2 estimate, and this approximation became closer as the number of predictors in the model increased.

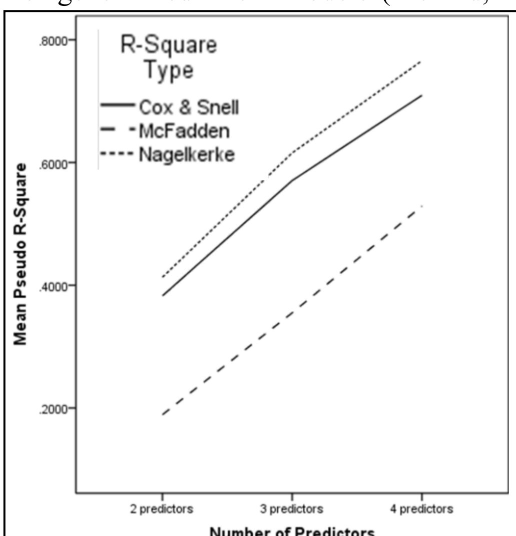

Figure 1. Mean R^2 analog values for simulated samples by index type and number of predictors.

Figures 2, 3, and 4 show how the analog values vary by distribution type and link function (aggregating across number of predictors) for each of the index types, respectively. Once again, values of McFadden's index were generally lower in value than corresponding values of the Cox & Snell's or Nagelkerke's indices, and this observation was consistent across distribution type and link function. However, as can be seen for each of the three R^2 analogs, the complementary log-log function consistently resulted in larger R^2 analog values than the other four link functions. Also, greater variability in the R^2 analog values by distribution type (i.e., uniform, positively skewed, or negatively skewed) was evident when the complementary log-log link function was employed than when the other link functions were used. Lastly, for two of the

R^2 analogs (Cox & Snell's and Nagelkerke's), the complementary log-log link function resulted in larger R^2 analog values for the positively skewed distribution than for uniformly or negatively distributed outcomes; whereas for the other link functions this effect was reversed (i.e., R^2 analog values were smallest under the positively skewed distribution).

We next examined the mean R^2 analog values by link function and distribution type, aggregating across index type. As Figure 5 indicates, mean R^2 analog values (as expected) for ordinal regression with a uniformly distributed outcome increased as the number of predictors increased. However, although the mean R^2 analog values resulting from each link function were relatively equal when two predictors were used, as the number of predictors increased, the value of R^2 analog increased at a much more rapid rate when the complementary log-log link function was used compared to when the other link functions (logit, negative log-log, probit, and Cauchit) were employed. Similar patterns of R^2 analog values were evident when the effect of link function and number of predictors was observed for each distinct distribution type (Figure 6).

McCullagh (1978, 1980) indicates that, with respect to estimated regression parameters and model fit, an ordinal regression model should preferably be "palindromic-invariant"—that is, invariant to a reversal of the outcome categories. One might infer, then, that this invariance should also be a desirable property of a R^2 analog. In the present study, when goodness of fit for the palindromic (i.e., positively vs. negatively skewed) distributions were compared, (1) Nagelkerke's R^2 analog resulted in the least variant values among the three R^2 analogs (see Figures 2, 3, and 4), and (2) estimation using any link function, other than the complementary log-log link function, resulted in relatively invariant values (see Figures 5 and 6 - Uniform and Positively Skewed panels).

Discussion

Although R^2 analog values for ordinal regression are widely reported and available in most statistical packages, few if any guidelines exist for their interpretation. The present study found that, compared to several other commonly-used indices, Nagelkerke's index consistently resulted in the highest R^2 analog values, and appeared to most closely approximate the OLS estimate of R^2. McFadden's index generally showed the smallest values. However, the OLS R^2 estimate cannot necessarily be held as a "gold standard" (any more than a particular R^2 analog might be held as a gold standard for OLS regression). It can validly be argued that particular R^2 analogs that result in lower values simply require their own set of interpretational guidelines. Nonetheless, most interpretations of R^2 analog values, for lack of explicit guidance, use OLS R^2-based guidelines (e.g., effect sizes such as f^2). The present study suggests that such interpretations may not be adequate.

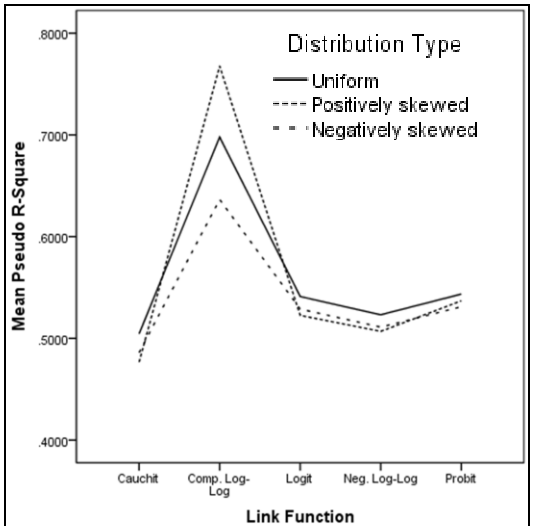

Figure 2. Mean Cox & Snell R^2 analog values by link function and distribution

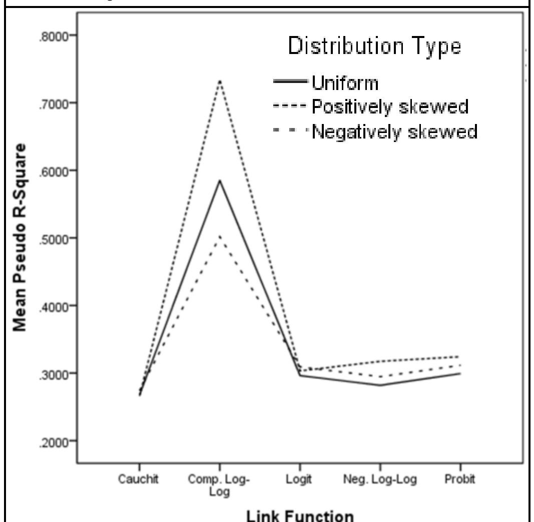

Figure 3. Mean McFadden R^2 analog values by link function and distribution

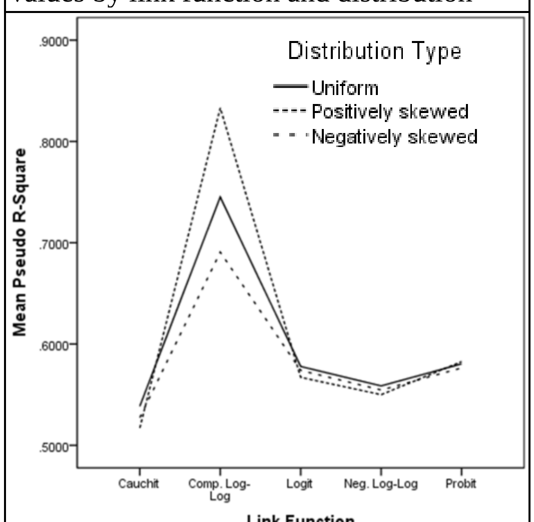

Figure 4. Mean Nagelkerke R^2 analog values by link function and distribution

Further, ordinal regression can be implemented using a variety of link functions and , although Norusis (2012) provides guidelines for the choice of link function, few studies investigate their comparative use in practice. Oftentimes, researchers will choose a link function based on the relative magnitude of the resulting R^2 analog values. As the present study indicates, use of the complementary log-log link function resulted in larger R^2 analog values than did the use of other link functions (and increased more markedly than other indices as the number of predictors was increased), and these values were also more variable. Moreover, the R^2 analog values resulting from the complementary log-log link function were so discrepant from those observed with other link functions that it appears a completely distinct set of interpretational guidelines might be necessary when this link function is used. Further, use of the complementary log-log link function appears to result in R^2 analog values that do not have the desirable attribute of palindromic invariance, and other link functions would be better choices in this regard. It also appears that, among the R^2 analogs, the Nagelkerke index appears to best exhibit palindromic invariance.

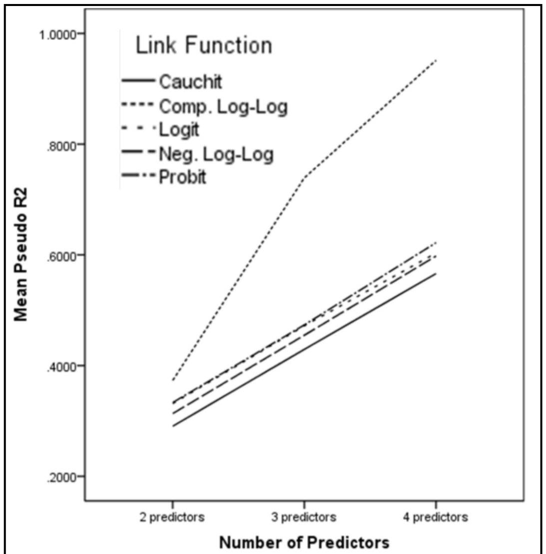

Figure 5 Mean R^2 analog by link function and number of predictors for all distributions.

Research that examines the effects of varying link functions and distributional variability in the ordinal outcome is important to an increased understanding of how these R^2 analogs should be used and interpreted. Continued research on this topic will help to guide interpretation in substantive studies.

Footnote

† A number of published sources mistakenly define this statistic as $1 - [LL_{\text{Null}}/LL_{\text{Full}}]^{2/n}$. The Cox and Snell (1989) statistic may be phrased using log-likelihoods as $1 - e^{[-2/n(LL(\text{Full}))-(LL(\text{Null}))]}$.

References

Agresti A. (1984). *Analysis of ordinal categorical data.* New York: Wiley.

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

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley & Sons.

IBM SPSS Statistics Information Center (2011). *Evaluating the choice of link function*. Retrieved from http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.cs%2Fplum_germcr_intro_02.htm

Li, K., & Duan, N. (1989). Regression analysis under link violation. *Annals of Statistics, 17*, 1009-1052.

Lipsitz, S. R., Fitzmaurice, G. M., & Molenberghs, G. (1996). Goodness-of-fit tests for ordinal response regression models. *Applied Statistics, 45*, 175-190.

McCullagh, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika, 65*, 413-415.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Series B, 42*, 109-142.

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

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*, 17-24.

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

Norusis, J. M. (2012). *IBM SPSS statistics 19.0 advanced statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.

Send correspondence to: Thomas J. Smith
Northern Illinois University
Email: tjsmith@niu.edu

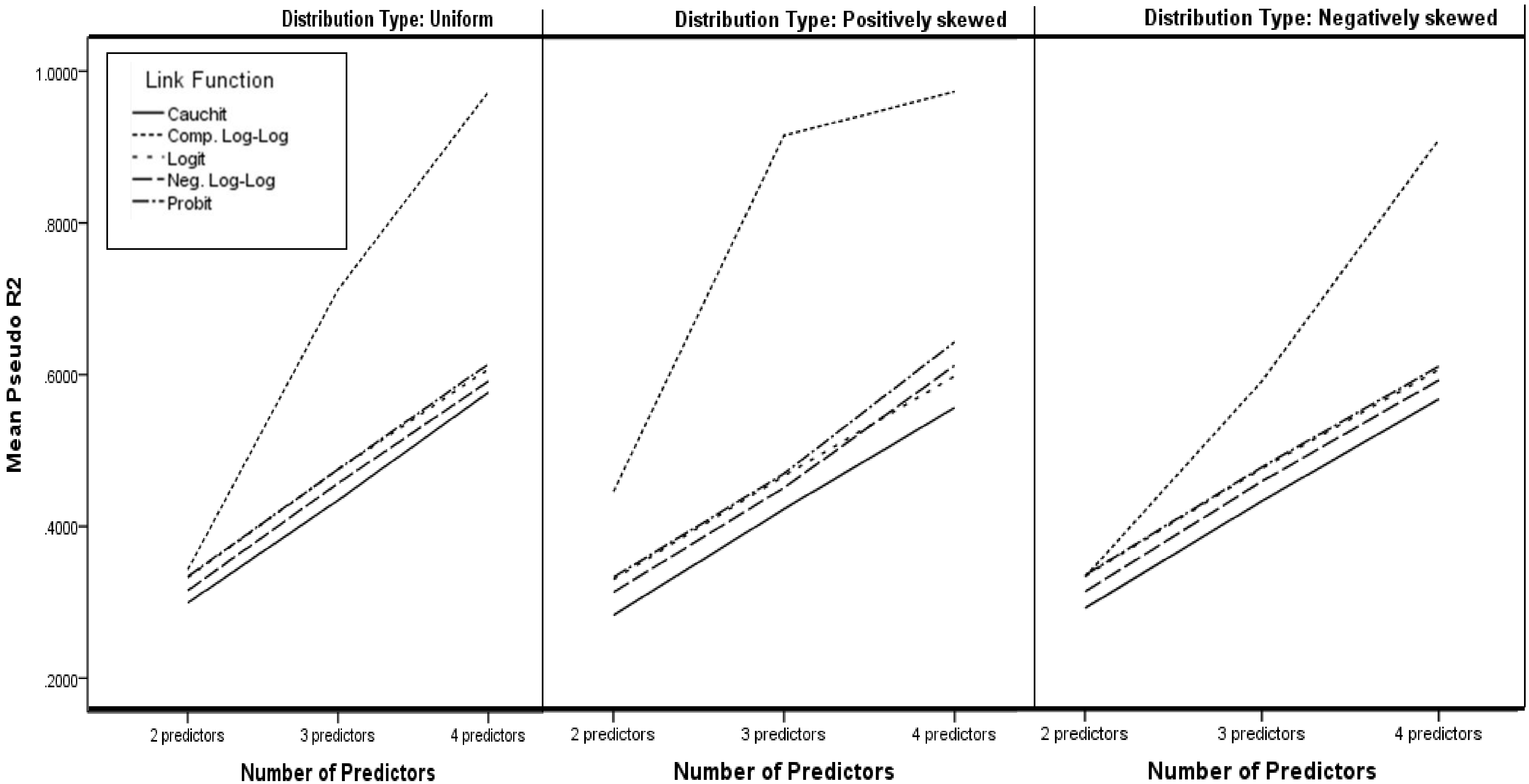

Figure 6. Mean R^2 analog values by link function and number of predictors for Uniform, Positvely-skewed, and Negatively-skewed outcomes. Simulations are based on 1000 sample replications, with n = 200 for each sample.