Finance, Economics and Monetary Policy
Discussion Papers

WHICH PSEUDO R-SQUARED?
CONCLUSIVE NEW EVIDENCE

Arturo Estrella

Discussion Paper No. 2202
June 2022

Which Pseudo R-Squared? Conclusive New Evidence

Arturo Estrella

Finance, Economics and Monetary Policy Discussion Paper No. 2202

June 2022

JEL Classification: C12, C25, C52

Abstract

What is the best way to assess the fit of an econometric model with a dichotomous dependent variable? Many applied researchers continue to employ pseudo R-squared measures that have serious mathematical, statistical and interpretational flaws. Drawing on mathematical analysis, hypothesis testing techniques and information theory, this paper presents conclusive evidence that clearly supports a single measure and reveals important shortcomings of the others.

Keywords: measure of fit, maximum likelihood, probit, logit, dichotomous dependent variable

Arturo Estrella
Professor of Economics, Emeritus
Rensselaer Polytechnic Institute
estrea@rpi.edu

1. Introduction

Too many researchers find themselves in a quandary when trying to assess the fit of an econometric model with a dichotomous dependent variable (DDV), such as probit or logit. The R-squared from the linear model is not available and some of the older readings on the topic point to a multitude of competing measures without providing clear evidence of their absolute or relative merits. Estrella (1998) proposed a new pseudo R-squared and showed that it consistently outperforms the other existing measures in a series of tests.

Since then, many applied researchers have taken advantage of the new measure and several major econometric and statistical software packages have incorporated it into their DDV routines and in their documentation.[1] Yet, some applied researchers continue to employ other measures that have serious shortcomings. The main purpose of the present article is to present new evidence that shows conclusively that the 1998 measure is the only reliable and interpretable choice to evaluate the fit of a DDV model.[2]

Researchers who already use the 1998 measure will find in these new results a deeper understanding of the measure as well as mathematical links to other DDV model statistics. Researchers who are still undecided about which measure to use should benefit from the clear comparative results that reveal important shortcomings of alternative measures. Finally, researchers who are unaware of the 1998 measure will have a chance to study the evidence and make an informed decision.

---

[1] A review of the documentation for a sample of statistical packages shows that some have incorporated the Estrella (1998) measure as the default R-squared for DDV models (RATS: Estima (2014), TSP: Hall and Cummins (2009)), some have included it in the output or documentation as one of several alternatives (LIMDEP: Greene (2012), SAS: SAS Institute (2021), SHAZAM: Whistler et al. (2011)) and some focus only on the ad hoc measures existing prior to 1998 (EViews, R, SPSS, Stata).

[2] As noted in Estrella (1998), the pseudo R-squared may also be applied to non-DDV models in which the average log likelihood ratio is bounded, such as multinomial logit and probit. For conciseness, the concrete analysis presented here is limited to the DDV case.

Admittedly, some readers may question the need for an R-squared at all. They may propose, for example, that testing for statistical significance is more important than assessing the fit of a model. But statistical significance by itself may not provide sufficient substantive information in a real-world application.[3]

Also, statistical significance by itself may be misleading. Anyone who consumes microeconomic research has run into cross-section or panel estimates with 10,000 observations in which everything is significant at the 1% level. However, if the R-squared is close to zero, the real-world information content of the model may be wanting.

The present article differs from the earlier literature that evaluates pseudo R-squared measures in several ways. First, rather than considering afresh the dozen or so measures that have been proposed historically, the field is narrowed to four candidates by eliminating the ones that previous analysis has shown to be clearly deficient in at least one important aspect. Section 2 identifies the four measures, all of which appear in some recent empirical work, and presents a brief review of previous tests.

Second, the analysis delves into foundational issues that make the construction of a measure of fit especially challenging in the DDV case. In particular, Section 3 examines the issue of boundedness of the average likelihood ratio statistic for a DDV model and applies the concepts of entropy and conditional entropy to compare the relative information content of dichotomous and continuous random variables. The components of the likelihood ratio test are seen to have clear interpretations in terms of entropy, which may be used in the evaluation of the pseudo R-squared measures.

[3] For example, Estrella and Mishkin (1998) found that a yield curve spread and the S&P 500 index were both statistically significant at the 1% level in separate probit equations to forecast U.S. recessions four quarters ahead. However, they also reported that the Estrella (1998) R-squared was .296 for the yield curve and .043 for the S&P. A policymaker or an investor would surely find that additional information useful.

Finally, Sections 4 and 5 compare the four candidates in terms of their general mathematical properties and in terms of their quantitative statistical performance, respectively. All the evidence points in the direction of the 1998 measure, but the quantitative tests in Section 5 show the differences most starkly. The main weakness of the other measures is a strong dependence on the mean of the dependent variable, which in the linear model does not influence R-squared as a measure of fit.

2. Brief review of earlier results

Estrella (1998) argued that for a pseudo R-squared to have an interpretation comparable to the linear model R-squared, it should satisfy the following three conditions.

1. The value of the measure should lie in the unit interval, with 0 representing "no fit" and 1 representing a "perfect fit."
2. The measure should be based on a valid test statistic of the hypothesis H_0 that all the coefficients of the model, except for the constant term, are zero.
3. The rate of growth of the measure with respect to the test statistic should be comparable to that of the linear model.

The derivation of the pseudo R-squared in Estrella (1998), reviewed later in this section, was based on solving a differential equation that gave precise mathematical meaning to those three goals.

The present analysis considers only those measures that have the potential to satisfy all three of the above criteria. Measures based on second moments are not included because the ordering of models they generate may differ from the ordering based on proper maximum

likelihood tests of H_0. Two likelihood-based measures from the 1998 article are also excluded because they are subject to upper bounds substantially less than 1.

That leaves four measures, denoted here as follows, defining Lc and Lu as the constrained (constant term only) and unconstrained log likelihood of the model and *n* as the number of observations.

$R2e = 1-\left(Lu/Lc\right)^{-2Lc/n}$ from Estrella (1998)

$R2m = 1-Lu/Lc$ from McFadden (1974)

$$R2cu = \frac{1-\exp\left(Lc-Lu\right)^{2/n}}{1-\exp\left(Lc\right)^{2/n}}$$ from Cragg and Uhler (1970)

$$R2vz = \frac{2\left(Lu-Lc\right)}{2\left(Lu-Lc\right)+n}\cdot\frac{2Lc-n}{2Lc}$$ from Veall and Zimmermann (1992)

Three of these measures may be interpreted as adopting their mathematical form from the relationship between R-squared and the average value of one of the classical statistics used to test H_0 in the linear model. To be precise, the test may be performed using the Lagrange multiplier statistic *LM* (estimating under the null hypothesis), the Wald statistic *W* (estimating under the alternative hypothesis) or the likelihood ratio statistic *LR* (estimating under both the null and the alternative). These tests are asymptotically equivalent in the linear model.

Define the average values of the three statistics as $ALM = LM/n$, $AW = W/n$, $ALR = LR/n$. Then the linear model R-squared has the following alternative exact expressions.

$$R^2 = ALM = 1-\exp\left(-ALR\right) = AW/\left(AW+1\right).$$

Each of the three average statistics has a value of 0 when the model does not fit at all ($Lu = Lc$), but as the fit improves, *ALM* is bounded above by 1 while *AW* and *ALR* are unbounded. These ranges are all consistent with R-squared values that go from 0 to 1.

In the DDV case, all four pseudo R-squared measures under consideration here are based explicitly or implicitly on the DDV average likelihood ratio statistic, which we denote as $A = 2\left(Lu - Lc\right)/n$. Values of this statistic range from zero to an upper bound $B = -2Lc/n$. Now $Lc/n = E\left(y\right)\log\left(E\left(y\right)\right) + \left(1 - E\left(y\right)\right)\log\left(1 - E\left(y\right)\right)$, where $E\left(y\right)$ is the mean value of the DDV, so this mean is the sole determinant of the upper bound *B*. Using this relationship, we see that the value of *B* lies in the range $0 \le B \le \log\left(4\right) = 1.386$.

The upper bound *B* is shown in Figure 1 as a function of $E\left(y\right)$. The function is symmetrical around $E\left(y\right) = \frac{1}{2}$, so that $E\left(y\right)$ and $1 - E\left(y\right)$ correspond to the same upper bound. In the subsequent analysis, the lower of these two proportions will generally be used, but the results apply symmetrically to the other value as well. The figure shows vertical lines at .05, .20 and .50, which will be used in various numerical illustrations as representative of the range of values encountered in practice. In the context of Figure 1, the pseudo R-squared measures have to deal with two basic issues: how high is *B* for a given $E\left(y\right)$ and how fast should the measure scale the slope from zero to *B*?

A simple mechanical way to construct a pseudo R-squared for the DDV case is to take one of the formulas that apply to the linear model and substitute into it the DDV likelihood ratio statistic *A*. Thus, one could use the value of *A* itself to correspond with *ALM,* $1 - \exp\left(-A\right)$ to correspond with *ALR* or $A/\left(A + 1\right)$ to correspond with *AW*. In fact, the last two forms have been

suggested in the literature, but they are unsatisfactory because their upper bounds are much lower than 1. The first alternative does not seem to have been proposed.

One way to make each of these alternatives fall in the desired range is to divide them by their respective upper bounds. This scaling adjustment is certainly not arbitrary, but it seems ad hoc in the absence of further information. The resulting measures are $R2m = A/B$, $R2cu = \left(1-\exp\left(-A\right)\right)/\left(1-\exp\left(-B\right)\right)$ and $R2vz = \left(A/\left(A+1\right)\right)\cdot\left(B/\left(B+1\right)\right)^{-1}$, as may be verified by substituting the definitions of *A* and *B* and comparing with the expressions given earlier.

Measure *R2e* is constructed differently, based on the argument that as the average likelihood ratio of a model grows, the measure of fit should increase at a rate comparable to that implied by the linear model. Rather than choose a particular functional form, the idea is to model marginal R-squared as inversely related to the proportional distance between *A* and its upper bound. Marginal R-squared is defined here as the derivative of the measure of fit in proportion to the difference between its level and the maximum value of 1, so the resulting differential equation is

$$\frac{1}{1-R2\left(A\right)}\frac{d}{dA}R2\left(A\right)=\frac{B}{B-A}.$$

Imposing the boundary condition $R2\left(0\right)=0$, the solution to the differential equation is the measure $R2e=1-\left(1-A/B\right)^{B}$. The solution also satisfies $R2\left(B\right)=1$ and $R2'\left(0\right)=1$. This last condition is satisfied by the three alternative measures in the linear model and is consistent with the objective of making the rate of growth consistent with the linear case.

Note that this technique also produces the correct R-squared formula in the linear model. In that case, there is no upper bound to *ALR* and, letting $B\to\infty$, the proportional distance on the

right-hand side of the differential equation becomes $(B-A)/B \to 1$. The solution to this alternative equation, subject to $R2(0)=0$, is $R2(A)=1-\exp(-A)$, which is the relationship between the average likelihood ratio statistic and R-squared in the linear model.

The four pseudo R-squared measures considered here were subjected in Estrella (1998) to a number of tests, the results of which are briefly summarized as follows.

Derivative at *A*=0: It may be verified that *E2e* is the only measure that has a unit derivative at *A*=0, as all the linear model measures do. The derivatives of *E2cu* and *E2vz* are always larger and for *E2m*, the derivative is 1 only if the expected value of the DDV is .1997 or .8003.

Odds ratio: In a model in which both the dependent (*y*) and independent (*x*) variables are dichotomous, the critical value of the odds ratio was computed for a constant p value and for various combinations of $E(x)$ and $E(y)$. Values of the four pseudo R-squared measures corresponding to the critical values were generated. Results for *R2e* were virtually constant across all cases, which is consistent with the constant p value, whereas *E2m*, *E2cu* and *E2vz* had values that were strongly inversely related to $E(y)$.

F statistic: In the linear model, R-squared may be transformed into an F statistic to test H_0 by taking $F = R^2/(1-R^2)\cdot(n-k-1)/k$, where *n* is the number of observations and *k* is the number of nonconstant regressors. For the two extreme DDV cases of $E(y)$ = .05 and .50, critical values of the likelihood ratio test for H_0 were computed at the .01, .05 and .10 levels of significance. Values of the four pseudo R-squared measures were generated from the critical values and transformed into *F* levels using the formula above. The level of significance was then

computed from the *F* distribution. Results for *R2e* were virtually the same as the DDV significance levels, but they were exceptionally different for the other three measures.

Overall, Estrella (1998) presented compelling evidence that favored measure *R2e*, both in terms of mathematical properties and statistical performance. Having narrowed the present field to four measures, Sections 4 and 5 present the results of new tests that reveal further sharp differences in the performance of the measures.

Before proceeding to the tests, however, it is helpful to explore further the boundedness of the average likelihood ratio statistic in the DDV case, which we have seen is a complicating factor in the construction of an R-squared measure.

3. Entropy, the upper bound *B* and residual uncertainty

3.1 Entropy and the upper bound

The asymptotic properties of the likelihood ratio test of H_0 are very similar whether the model is linear with a continuous dependent variable or nonlinear with a DDV. In both cases, the large-sample distribution of the test is chi square with *k* degrees of freedom, the latter representing the number of additional parameters included in the alternative hypothesis.[4] Why then is the test statistic bounded above only in the DDV case? The reason has to do with the amount of information that would have to be obtained to explain the dependent variable completely.

As an example, suppose that a DDV represents an event that happens 99% of the time, and suppose that I want to predict the outcome of the next observation. If I forecast that the event will occur, using no other information, I will be right 99% of the time. This prediction is not

[4] This result is known as the Wilks Theorem. See, for example, Wilks (1938).

difficult, but improving on it by incorporating additional information is hard because there is so little room for improvement. In contrast, a DDV corresponding to an event that happens 50% of the time is hard to predict but leaves much more room for improvement in the sense that additional information could increase predictive power substantially.

Entropy is a measure of how much information there is to be learned about a random variable. In the case of a DDV *y* that takes the values 0 and 1, entropy is defined as

$$\begin{aligned} H(y) &= -P(y=1)\log\big(P(y=1)\big) - \big(1-P(y=1)\big)\log\big(1-P(y=1)\big) \\ &= -E(y)\log\big(E(y)\big) - \big(1-E(y)\big)\log\big(1-E(y)\big). \end{aligned}$$

H is positive by construction and reaches its highest level when $P(y=1) = E(y) = .5$, at which $H = \log(2) = .693$.[5]

Comparing the expression for *H* with the definition of *B* in Section 2 shows that $H = -Lc/n$, hence that $B = 2H$. Thus, the reason that the average likelihood ratio statistic is bounded above by *B* is the limited information content of the DDV *y*, and this limit is a function solely of the mean of the variable. The multiplier 2 comes from the derivation of the likelihood ratio test statistic so that it follows asymptotically a chi-square distribution. Graphically, the function *H* has the same shape as the upper bound *B* seen earlier in Figure 1, but half the magnitude.

In the linear model, the dependent variable is usually a continuous random variable, which as far as entropy goes is very different from a DDV. Entropy for a continuous random variable may be defined by discretizing the probability density function over intervals of length

[5] In this definition, log is the natural logarithm and for that reason the function is sometimes called natural entropy. An alternative in information theory is to use the base 2 logarithm so that entropy is expressed in terms of bits rather than natural log units or nats. The natural scale fits better with the likelihood ratio application. The entropy concepts used in this section are explained in detail in textbooks on information theory such as Cover and Thomas (2006).

Δ and taking the limit of the sum of $-p_i \log(p_i)$ over these discrete terms as $\Delta \to 0$. In that case, the result includes a term that may be finite plus a component of the order of $-\log(\Delta) \to \infty,$ which is consistent with the unboundedness of entropy and the test statistic in the linear model case.

To compare the entropy of a continuous dependent variable with the DDV case, it is helpful to examine the finite component of the limit of the sum described above, which is known as differential entropy. It does not represent the absolute level of entropy, but it is commonly used to compare the difference in entropy for pairs of continuous random variables, since the infinite components corresponding to the two variables may be thought of as offsetting one another.

To illustrate the concept of differential entropy, suppose that $f(x)$ is the normal density function with mean μ and variance σ^2. Its differential entropy is defined as $h = -\int_{-\infty}^{\infty} f(x) \log(f(x))\, dx = \frac{1}{2} \log(\sigma^2 2\pi e)$. In this case, differential entropy is an increasing function of the variance σ^2 and, although *h* is finite for a given value of σ^2, it is unbounded as a function of σ^2.

Since the entropy of a continuous random variable is unbounded, the average likelihood ratio statistic for a linear model with a continuous dependent variable is not subject to an upper bound. However, only models that fit the data well are likely to have an *ALR* that exceeds the levels attainable in models with DDVs. We can be more precise about this point by employing the asymptotic distribution of the likelihood ratio statistic *LR*.

In a linear model with *k* nonconstant explanatory variables and *n* observations, $LR = n \cdot ALR$ follows asymptotically a chi-square distribution with *k* degrees of freedom. We

can use this result to estimate the probability that the linear model *ALR* exceeds any given value, say a level that falls within the bounds of standard DDV models.

Consider the value of *B* from a model in which the DDV has a mean of .05. In that case, $B = .397$. Applying the large-sample distribution, the estimated probability that the linear model *ALR* exceeds this level is very sensitive to the number of observations and to the number of variables in the model. For example, when $n = 100$ and $k = 1$, the probability that *ALR* exceeds .397 is very low, on the order of 10^{-10}. In contrast, with $n = 50$ and $k = 10$, the probability increases to .03.

Thus, in a linear model with a large data sample, it is possible that the average likelihood ratio statistic may surpass the upper bound for a realistic DDV model, but it is generally a low probability event. These higher levels are attained only by models that perform very well in terms of level of significance and fit. In the DDV case, applying the linear model R-squared formula directly would severely constrain the possible range of values and would result in underestimating the performance of well-fitting models.

3.2 Conditional entropy and residual uncertainty

The related concept of conditional entropy may be introduced to help with the interpretation of two of the pseudo R-squared measures under consideration. Suppose that *y* is a DDV and that *x* is a potential explanatory variable, not necessarily dichotomous. The conditional entropy of *y* with respect to *x*, denoted by $H(y \mid x)$, represents the amount of information in *y* that remains unexplained after the effects of *x* are taken into account. If *x* is a continuous random variable with density function $f(x)$ and support set *S*, then

$$H(y|x) = \int_S \left[-P(y=1|x)\log\left(P(y=1|x)\right) - \left(1-P(y=1|x)\right)\log\left(1-P(y=1|x)\right) \right] f(x)\,dx .$$

Heuristically, conditional entropy is the information contained in the residual of a regression equation. Since entropy $H(y)$ corresponds to the information content of the dependent variable, the ratio $U(y|x) = H(y|x)/H(y)$ represents the portion of the information in the dependent variable that is not explained by the regression. It may be interpreted as a proportional measure of residual uncertainty.

We saw earlier that in a DDV equation, $H(y) = -Lc/n$ and $B = 2H(y)$. Similarly, when the DDV model contains only *x* as the explanatory variable, $H(y|x) = -Lu/n$ and $A = 2\left(H(y) - H(y|x)\right)$. Combining these relationships, $A/B = 1 - U(y|x)$, which implies that two of the four pseudo R-squared measures may be defined in terms of the entropy ratio *U*.[6]

One case is $R2m = 1 - U(y|x)$. If the focus of interest is on the entropy ratio, measure *R2m* provides the most direct information for that purpose. In fact, McFadden (1974) suggests that *R2m* is analogous to "the 'mean squared error' explained or the 'variance' explained." Since *R2m* is the complement of the entropy ratio, which represents residual uncertainty, the measure may be interpreted as the proportion of the DDV entropy explained by the model.

However, this interpretation does not imply that *R2m* is a reliable measure of fit along the lines of the traditional R-squared. In the linear case, the interpretation in terms of "the proportion of variance explained" is associated with the Lagrange multiplier statistic, not with the likelihood ratio statistic on which *R2m* is constructed. Specifically, for the linear model, $ALM = 1 - URSS/RRSS$, where the terms on the right-hand side correspond to the unrestricted

[6] The other two measures depend on *A* and *B* separately rather than on their ratio.

and restricted residual sum of squares. When the restriction is H_0, that is, there is only a constant term in the linear equation, the ratio represents the variance of the residual in the full equation as a proportion of the variance of the dependent variable.

The problem is that *R2m* is not constructed from the Lagrange multiplier statistic, but as a linear rescaling of the average likelihood ratio statistic, which even in the linear model is a nonlinear function of both the sum of squares ratio and of R-squared. The mathematical relationships are given by $ALR = -\log\left(URSS/RRSS\right) = -\log\left(1-R^2\right)$. Thus, to convert the average likelihood ratio into a reliable measure of fit requires a nonlinear transformation, as will be clear in Section 4. In addition, the quantitative analysis in Section 5 points to serious problems with the use of *R2m* as a measure of fit, notwithstanding its close connection with the entropy ratio.

The second measure that is definable in terms of the entropy ratio is *R2e*, since the reciprocal of the entropy ratio is the driving function in the differential equation from which the measure is derived: $\left(1-R2e\left(A\right)\right)^{-1} dR2e\left(A\right)/dA = U\left(y \mid x\right)^{-1}$. The idea is to make marginal R-squared inversely related to the proportion of the information that remains to be explained. The solution to the differential equation, taking account of the implicit dependence of $U\left(y \mid x\right)$ on *A*, may be expressed in terms of entropy as $R2e = 1-U\left(y \mid x\right)^{2H(y)}$. Sections 4 and 5 will show that the adjustment in the exponent of *U*, as compared with the linear expression for *R2m*, makes the measure much more robust, particularly with respect to differences in the level of $E\left(y\right)$.

Returning to the linear model, the relationship between entropy and the likelihood ratio statistic is straightforward if the variables are normally distributed. Let $y \sim N\left(\mu,\sigma^2\right)$ be the

dependent variable and assume that all the regressors are also normally distributed. Then the variance of the residual may be expressed as $\sigma_u^2 = \left(1 - R^2\right)\sigma^2 = \exp\left(-ALR\right)\sigma^2$.

Using the expression given earlier for the differential entropy of a normal variable, the difference between the differential entropy of the dependent variable and the conditional differential entropy of the model is

$$h(y) - h(y \mid x) = \tfrac{1}{2}\log\left(\sigma^2 2\pi e\right) - \tfrac{1}{2}\log\left(\exp\left(-ALR\right)\sigma^2 2\pi e\right) = \tfrac{1}{2}ALR.$$

Much as in the DDV model, $ALR = 2\left(h(y) - h(y \mid x)\right)$. An important difference is that in this case *ALR* is unbounded and so is $h(y) - h(y \mid x)$.

4. Mathematical properties of the pseudo R-squared measures

This section compares the mathematical properties of the four pseudo R-squared measures as functions of the likelihood ratio statistic. Table 1 collects expressions for the functions themselves, their derivatives, and other transformations, all in terms of the average likelihood ratio statistic $A = 2\left(Lu - Lc\right)/n$ and its upper bound $B = -2Lc/n$. This notation facilitates comparisons across measures.

Consider first the functions themselves. It is clear from the first row of Table 1 that the four measures have distinct functional forms, which could be described as power, linear, exponential and rational. For this reason, they will tend to give different signals for most values of *A* and *B*, other than at the two extreme points for *A*. The only case in which two measures coincide is when $B = 1$ ($E(y) = .1997$ or $.8003$), for which $R2e = R2m$. Note that all the functions depend only on *A* and *B*. In particular, they do not depend on sample size directly, although sample size is used in the calculation of the averages *A* and *B*.

Inspection of the first row of Table 1 also shows that all four measures satisfy the first goal for a pseudo R-squared, as proposed at the beginning of Section 2. The range of each of the functions is the unit interval [0,1], they cover the full range as *A* goes from 0 to *B* and the function values at the endpoints correspond to "no fit" and "perfect fit," respectively.

For measures *R2m*, *R2cu* and *R2vz*, the unit value at $A = B$ is achieved by multiplicative rescaling, with the denominator in each case representing the value of the numerator when $A = B$. This rescaling is designed to adjust the value of the function at this single point, but its application affects the level of the function and its derivatives for all values of *A*.

The consequences of the rescaling are illustrated in the top two panels of Figure 2. The left panel shows the measures before rescaling, defined as $R2mu = A$, $R2cuu = 1 - \exp(-A)$ and $R2vzu = A/(A+1)$. The values of the unscaled measures are very similar to the linear model R-squared for low values of *A*. In fact, *R2cuu* is the exact linear model formula. Scaling problems for the unadjusted measures increase gradually as the value of *A* grows, especially when it gets closer to the upper bound *B*. Multiplicative rescaling is a perfect solution to the problem at the right endpoint, but at the cost of changing the levels for all values of *A* by the same proportion, including the low values of *A* where change was unnecessary and undesirable.

The top right panel of Figure 2 shows the actual adjusted measures for $E(y) = .05$ and compares them with the linear model. Constant proportional rescaling drives all three measures upwards far from the linear model, even for low values of *A* where scaling issues were minor.

In contrast, for *R2e*, the unit value at $A = B$ is obtained organically by solving a differential equation, as shown earlier. As in the linear model, the derivative of *R2e* is driven by the inverse of the proportional distance to the upper bound, so the extent of the rescaling adjusts smoothly as the value of *A* increases.

The results are illustrated in the bottom two panels of Figure 2, which compare *R2e* with the linear model for two values of the DDV mean. For low values of *A*, where rescaling is not much of an issue, the differential equation approach keeps the value of *R2e* close to the linear model. As *A* grows, the adaptive scaling drives the measure increasingly upwards, away from the linear model, to reflect the growing closeness to a perfect fit for the model.

The bottom right panel of Figure 2 shows the levels of *R2e* for $E(y) = .50$, which corresponds to the largest possible level of the upper bound *B*. As the upper bound increases, the *R2e* function gets closer to the linear model, and this panel shows the closest correspondence feasible given the limited information content of a DDV.

Turning to the derivatives in the second row of Table 1, consider first their values when $A = 0$. The derivative of *R2e* at this point is 1, which matches the derivative in the linear model. For measures *R2cu* and *R2vz*, the initial growth rate is above 1, suggesting that they overstate the fit of the model, at least at lower levels of *A*. For *R2m*, the relationship of the initial growth rate to unity depends inversely on *B*. The rate equals 1 in the special case when $B = 1$, but otherwise this measure starts growing faster or slower than in the case of the linear model.

Marginal R-squared appears in the third row of Table 1, defined as in Section 2 as $\left(1 - R2i(A)\right)^{-1} dR2i(A)/dA$, where i represents one of the four pseudo R-squared functions under consideration. Whereas the derivative is the rate of growth in absolute terms, marginal R-squared looks at the derivative in proportion to the distance from the current level of the function to the maximum value of 1. At $A = 0$, this distance is 1 so that marginal R-squared is the same as the derivative.

For the linear model, $R^2 = 1 - \exp(-A)$ and marginal R-squared as a function of *A* is always 1. Among the DDV measures, *R2e* is the only one for which marginal R-squared at $A = 0$ is 1 regardless of the DDV mean.

Table 1 also shows that marginal R-squared is predominantly an increasing function of *A* and that it goes to infinity as *A* approaches the upper bound *B*. Intuitively, the boundedness of *A* in the DDV case makes it necessary for each pseudo R-squared function to speed up as *A* approaches *B* so as to reach its upper bound when the fit is perfect. The only partial exception is *R2vz*, for which marginal R-squared has a small downward slope in the range $0 \le A < (B-1)/2$ if $B > 1$.

Among the DDV measures, *R2e* is the only one whose marginal R-squared may be expressed solely in terms of the entropy ratio $U = 1 - A/B$, as defined in Section 3. For each of the measures, substituting $A = B(1-U)$ into the expression for marginal R-squared converts it into a function of *U* and *B* only, but *R2e* is the only case in which dependence on *B*, and hence on $E(y)$, drops out completely. Section 5 will show quantitatively that this independence from $E(y)$ makes the measure more robust.

A generalization of the differential equation from which *R2e* is derived suggests that the condition that the derivative at $A = 0$ equals 1 is important to make the measure and its marginal R-squared less sensitive to differences in $E(y)$. Consider the generalized differential equation

$$\frac{1}{1 - R2(A)} \frac{d}{dA} R2(A) = \frac{C2}{C1 - A}.$$

The expression on the left is marginal R-squared and the right-hand side represents the reciprocal of the distance between *A* and a benchmark level *C*1, scaled by a constant *C*2. Three boundary

conditions are imposed, $R2(0)=0$, $R2(B)=1$ and $R2'(0)=1$, which together determine the values of $C1$, $C2$ and the constant of integration.

With the single boundary condition $R2(0)=0$, the solution to the general differential equation is $R2=1-(1-A/C1)^{C2}$. If in addition, we impose the condition that $R2(B)=1$, we have that $C1=B$ and $R2=1-(1-A/B)^{C2}$. Finally, the third condition $R2'(0)=1$ implies that $C2=B$, so that $R2=R2e=1-(1-A/B)^{B}$. Thus, the right-hand side of the differential equation reduces to $(1-A/B)^{-1}=U^{-1}$.

Measure *R2m* may be derived in a similar way by imposing the first two boundary conditions, but not the third, stipulating instead that $R2''(0)=0$ so that the function is linear. This condition implies that $C2=1$ and the result is $R2m=A/B$. In this case, the right-hand side of the differential equation becomes $(B-A)^{-1}=(BU)^{-1}$, which depends on *B* even after controlling for *U*.

One final point about the mathematical properties of the measures is based on counterfactual analysis of their functional forms, undertaken by allowing the upper bound *B* to approach infinity. We have seen that in the linear model, *ALR* is unbounded. So, what happens if we take the limit of any of the pseudo R-squared functions as *B* approaches infinity and apply the resulting function to the linear model *ALR*?

Despite their difference in functional form for finite *B*, the limit of both *R2e* and *R2cu* is $R^2=1-\exp(-A)$, which matches the linear case exactly. The other two measures produce very different results. In the case of *R2m*, the limit of the function is zero, which it is not useful for these purposes. In the case of *R2vz*, the result is $A/(A+1)$. As in the linear case, this expression

is increasing in *A* and converges to 1 in the limit as *A* approaches infinity. However, the level of this expression is uniformly lower than the linear model R-squared and would consistently understate the fit for any positive value of *ALR*.

In summary, the results of this counterfactual experiment clearly favor *R2e* and *R2cu*. The result for the latter is not unexpected, as the construction of *R2cu* may be interpreted as the direct application of the linear model relationship followed by multiplicative rescaling based on its upper bound. Taking the limit undoes the rescaling. The limiting function of *R2e* is perhaps less obvious but is nevertheless indicative of a fundamental correspondence in form with the linear case R-squared.

5. Evidence of statistical performance

This section turns to evidence based on the statistical properties of the four measures. The strategy consists primarily of performing numerical comparisons of the values of the different measures while controlling for the significance level of the likelihood ratio test and for the mean of the DDV.

The tests examine the internal consistency of each measure as $E(y)$ is allowed to vary for fixed *n* and *k*, while keeping constant the level of significance of the chi square test for H_0. Ideally, the value of each measure would remain constant across these changes in the DDV mean, as the level of significance is held fixed. That result clearly holds in the linear case, where R-squared is independent of the mean of the dependent variable. The values of the measures are also compared with the linear model R-squared for the same level of significance.

In the DDV case, the upper bound *B* creates complications and a dilemma. If the measure of fit is to have a value of 1 when the fit is perfect, as is the case with the four measures we are

considering, the function that defines each measure must depend on *B* and hence on $E(y)$, all else equal. Therefore, the goal of total independence of the value of the measure from $E(y)$ for a constant level of significance is not strictly attainable. A compromise solution admits a small degree of dependence on $E(y)$ for the range of parameter values most often encountered in practice. This compromise is possible with only one of the R-squared measures.

To set up the experiment, it is convenient to express each pseudo R-squared measure as a reduced-form function of four base parameters: the level of significance *p*, the number of explanatory variables *k*, the sample size *n* and the DDV mean $E(y)$. Table 1 gives an expression for each measure in terms of *A* and *B*. The fact that *B* is a function of the mean of the DDV has already been established. The statistic *A* may be expressed as a function of the other three parameters.

Specifically, the value of *A* satisfies the equation $F(nA;k)=1-p$, where $F(\cdot;k)$ is the chi square cumulative distribution function with *k* degrees of freedom. Since the cumulative distribution function is strictly increasing in *A*, *n* and *k*, the relationship may be inverted to obtain the implicit function $A(p,k,n)$, where *A* depends inversely on *p* and *n* and directly on *k*.

If *R2i* represents one of the four measures, we may substitute the foregoing relationships for *A* and *B* to write it in the form $R2i(p,k,n,E(y))$. Dependence on the parameters that enter through *A* has the same sign as for *A* so that $\partial R2i/\partial p<0$, $\partial R2i/\partial k>0$ and $\partial R2i/\partial n<0$.

In addition, each *R2i* depends inversely on the mean of the DDV for $E(y)\le\frac{1}{2}$.[7] In three of the four cases, this result is easy to show by taking the derivative with respect to *B* of the

[7] The results are symmetrical for *E*(*y*) above ½.

logarithm of the function, since we have seen that B and $E(y)$ are positively related in the range considered.

$$\partial \log(R2m)/\partial B = -1/B$$

$$\partial \log(R2cu)/\partial B = -\exp(-B)/(1-\exp(-B))$$

$$\partial \log(R2vz)/\partial B = -(B(B+1))^{-1}$$

In the case of *R2e*, the derivative is more complicated, but its form may be simplified by expressing it as a function of the entropy ratio $U = 1 - A/B$ defined in Section 3.

$$\partial R2e/\partial B = -U^{B-1}\left[U \log(U) + 1 - U\right]$$

Second-order series expansion of the term in square brackets at $U = 1$ shows that it is positive, so that the derivative is negative like the others.

Thus, as expected, none of the four measures is strictly insensitive to changes in the mean of the DDV when (p, k, n) are held fixed, which would be ideal. In fact, all four measures decline as the DDV mean increases in the range $E(y) \leq \frac{1}{2}$. The question is by how much. Since the values of the derivatives are difficult to compare analytically, it is helpful to plot the measures as functions of $E(y)$ for representative values of the other parameters and to compare the results graphically.

Figure 3 displays the values of the measures for $p = .01$, $k = 1$, $n = 100$ and with $E(y)$ running continuously on the horizontal axis from .05 to .50. Since every point in the graph represents the same level of statistical significance with the same number of explanatory variables and number of observations, there should be in principle close uniformity in the values of the all the measures of fit across the board.

The actual results show clearly that that is not the case, and the difference between *R2e* and the other measures is striking. Other than *R2e*, the levels of the measures start out very high, then fall sharply and eventually flatten out. *R2cu* and *R2vz* remain well above the others throughout, again suggesting that they tend to overstate the fit of the underlying model. *R2m* starts out almost as high as the other two, but then falls until it crosses below the level of the linear model just beyond the $E(y) = .20$ point, and it remains below that level as it flattens out.

Measure *R2e* is clearly much more stable than the others, which is consistent with the constant level of significance. There is a slight uptick as the DDV mean approaches .05, but the curve is very flat throughout the broad range of mean values. Moreover, the measure remains everywhere very close to the level of the linear model. A key reason for these results is that, as we have seen, marginal R-squared for *R2e* starts out at the same level as for the linear model and is everywhere independent of $E(y)$.

To verify the robustness of these results, the experiments were repeated for values of the significance level *p*, number of observations *n* and number of variables *k* representing substantial departures from the base case. Specifically, *p* was assigned values corresponding to an increase and a decrease by a factor of 10 from the base case, that is, .10 and .001. Similarly, the number of observations and number of variables were set at 10 times the base case, at 1000 and 10, respectively.

Qualitatively, the results are generally the same as in the base case. As in Figure 3, measure *R2e* remains stable as the other three measures fall from very high initial levels. *R2cu* and *R2vz* stay considerably above the linear model across the board and *R2m* starts high but falls below the linear case for values of $E(y)$ somewhat beyond the .20 level.

With the larger sample size, the pattern for *R2e* is even flatter and closer to the linear model than in the base case. With the larger number of variables, *R2e* shows a bit more separation from the linear model and a more noticeable uptick at .05, but the basic relative flatness remains when compared with the other measures. Overall, these experiments are highly supportive of the *R2e* measure.

Returning to the reduced-form expression $R2i\left(p,k,n,E\left(y\right)\right)$, Figure 3 shows that an important drawback of *R2m*, *R2cu* and *R2vz* is that they are overly sensitive to the last argument, the mean of the dependent variable. In fact, the sensitivities of these three measures to each of the remaining reduced-form arguments also present clear departures from the linear model.

Each panel of Figure 4 shows the values of the partial derivates of the pseudo R-squared measures with respect to one of the reduced-form parameters. Again, the base case is $p=.01, k=1, n=100$ and the value of $E\left(y\right)$ is allowed to run in the horizontal axis from .05 to .50. Since *k* is typically a small integer, sensitivity to *k* is calculated as a partial difference from $k=1$ to $k=2$. Each of the four panels also includes the corresponding partial derivative for the linear model, which is constant in each plot since R-squared is independent of the mean of the dependent variable. Their values are $-1.67, .0238, -.0006$ and 0, respectively.

We observe in Figure 4 that the partial derivatives of *R2e* are much more consistent across DDV mean levels that the derivatives of the other measures and that they are closer to the constant levels of the linear model. The quantitative divergences of the other three measures are substantial. Even in the case of the number of observations *n*, where all four measures exhibit very low sensitivity, as expected, *R2e* is much more stable than the others in relative terms.

The last panel of Figure 4 shows that *R2e* is less sensitive to $E(y)$ than all the other measures by two orders of magnitude for any starting value of $E(y)$. For the other measures, this considerable dependence on $E(y)$ compromises the consistency of their interpretability as a measure of fit across dependent variables with different means.[8]

We end this comparative analysis with a caveat. One danger implicit in the results of this section is that a strategically minded researcher might choose to report *R2cu* or *R2vz* in their work, particularly the latter, because all indications are that they tend to overstate the fit of the model and would thus portray it in the best possible light. Of course, it would then be up to the alert and informed reader or reviewer to insist on the use of *R2e* for accuracy of representation.

6. Conclusions

If the evidence presented in favor of the measure *R2e* in Estrella (1998) was compelling, the further evidence presented here is conclusive. Focusing on the four measures that earlier research found have the better properties, the present investigation provides greater depth of analysis and includes new tests that highlight the benefits and drawbacks of each of the four measures.

Each measure has some positive qualities. For instance, *R2m* is linearly related to the entropy ratio that captures the proportion of information in the dependent variable that is explained by the model. Measure *R2cu* incorporates the exponential form of the linear model and

[8] Hemmert et al. (2018, Table 4) examine the sensitivity of the four pseudo R-squared measures with respect to *k*, *n* and the DDV mean using meta-analysis of 274 published logistic regression models. They conclude that *R2m* performs best because it appears to be less sensitive in their calculations to changes in the reduced-form parameters. However, their procedure does not control for *p*, so the reported values of the measures could correspond to any level of significance and there should be no expectation of constancy or low sensitivity to the parameters.

is one of only two measures that converge to the linear model R-squared function if the boundedness afflicting the DDV statistic is removed.

Going back to the three criteria proposed in Section 2, all four measures considered clearly satisfy the first two criteria. It is the third criterion, regarding the rate of growth of the measure, that presents a severe challenge for all but one of the measures. Section 4 shows that the derivative and marginal R-squared of *R2e* are far more comparable to the linear model than the others. Moreover, in the quantitative analysis of Section 5, *R2e* has robust internal consistency and consistency with the linear model R-squared in assessing the fit of a DDV model. The differences are stark.

Why have some researchers and some statistical packages continued to use other measures in the last two decades? One reason could be that they are unaware of the literature. An internet search of reviews of pseudo R-squared measures during that period suggests that some of the reviewers are not familiar with the more recent work. Other reasons might be strategic, as in the example suggested at the end of the previous section regarding the use of measures that tend to overstate the fit of the model.

Developers of statistical software packages, as professional specialists in the field, should be aware of the literature and should provide more than an indiscriminate list of conceivable measures. If *R2e* is not included in a package or if no guidance is provided with respect to the choice of pseudo R-squared in the documentation for DDV model routines, the package has to be considered deficient in light of the evidence presented here.

The main issue is a practical one, to select the right tool to understand and interpret the statistical results of DDV models used in a broad range of empirical applications. Researchers,

reviewers, statistical software developers and anyone else who uses DDV models as producer or consumer of applied research should be aware of the evidence in this paper.

References

Cover, Thomas M. and Joy A.Thomas (2006) Elements of Information Theory, Second Edition, Hoboken, New Jersey: John Wiley and Sons.

Cragg, John G. and Russell S. Uhler (1970) “The Demand for Automobiles” Canadian Journal of Economics 3, 386-406.

Estima (2014) RATS Version 9.0 Reference Manual, Evanston, Illinois: Estima.

Estrella, Arturo (1998) “A New Measure of Fit for Equations with Dichotomous Dependent Variables” Journal of Business and Economic Statistics 16, 198-205.

Estrella, Arturo and Frederic S. Mishkin (1998) “Predicting U.S. Recessions: Financial Variables as Leading Indicators.” Review of Economics and Statistics 80: 45-61.

Greene, William H. (2012) LIMDEP Version 10 Econometric Modeling Guide, Plainview, New York: Econometric Software, Inc.

Hall, Bronwyn H. and Clint Cummins (2009) TSP 5.1 User’s Guide, Palo Alto, California: TSP International.

Hemmert, Giselmar A.J., Laura M. Schons, Jan Wieseke and Heiko Schimmelpfennig (2018) “Log-Likelihood-Based Pseudo-R^2 in Logistic Regression: Deriving Sample-Sensitive Benchmarks” Sociological Methods and Research 47, 507–531.

McFadden, Daniel (1974) “Conditional Logit Analysis of Qualitative Choice Behavior” In Frontiers of Economics, P. Zarembka, ed., New York: Academic Press.

SAS Institute (2021) SAS/ETS User’s Guide: The MDC Procedure, Cary, North Carolina: SAS Institute, Inc.

Veall, Michael R. and Klaus F. Zimmermann (1992) “Pseudo-R^2’s in the Ordinal Probit Model” Journal of Mathematical Sociology 16, 333-342.

Whistler, Diana, Kenneth J. White, David Bates, Madeleine Golding (2011) SHAZAM Reference Manual Version 11, Cambridge, England: SHAZAM Analytics, Ltd.

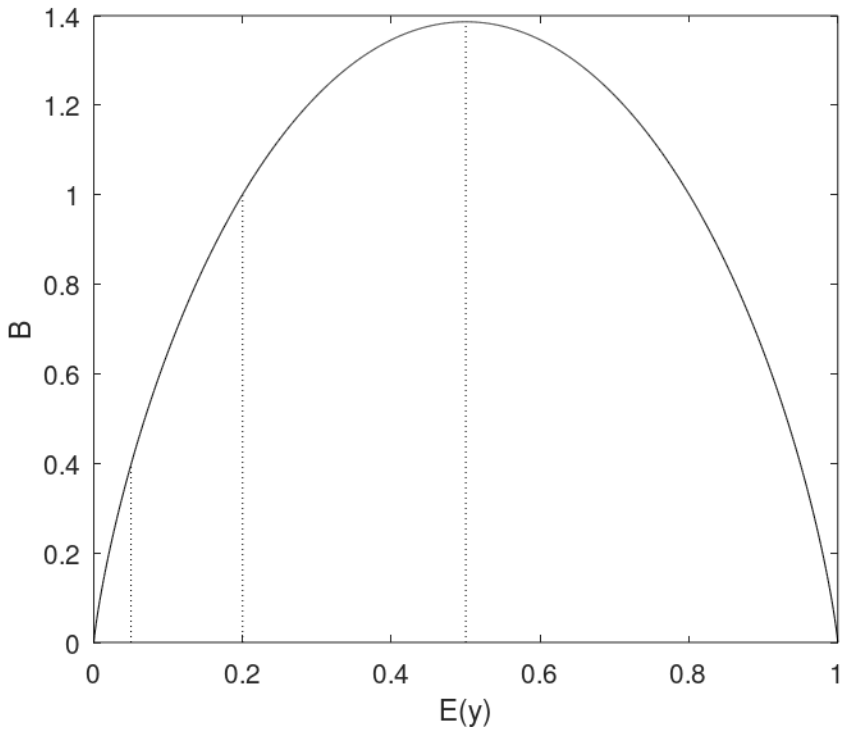
Wilks, Samuel S. (1938) “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses” The Annals of Mathematical Statistics 9, 60-62.

Table 1. Pseudo R-squared functions $R2i(A,B)$, derivatives and limits

	R2e	*R2m*	*R2cu*	*R2vz*
$R2i(A,B)$	$1-\left(1-\frac{A}{B}\right)^{B}$	$\frac{A}{B}$	$\frac{1-\exp(-A)}{1-\exp(-B)}$	$\frac{A/(A+1)}{B/(B+1)}$
$\partial R2i/\partial A$	$\left(1-\frac{A}{B}\right)^{B-1}$	$\frac{1}{B}$	$\frac{\exp(-A)}{1-\exp(-B)}$	$\frac{B+1}{(A+1)^2 B}$
Marginal $R2i$	$\frac{1}{1-A/B}$	$\frac{1}{B-A}$	$\frac{1}{1-\exp(A-B)}$	$\frac{B+1}{(B-A)(A+1)}$
$\lim_{B\to\infty} R2i$	$1-\exp(-A)$	0	$1-\exp(-A)$	$\frac{A}{A+1}$

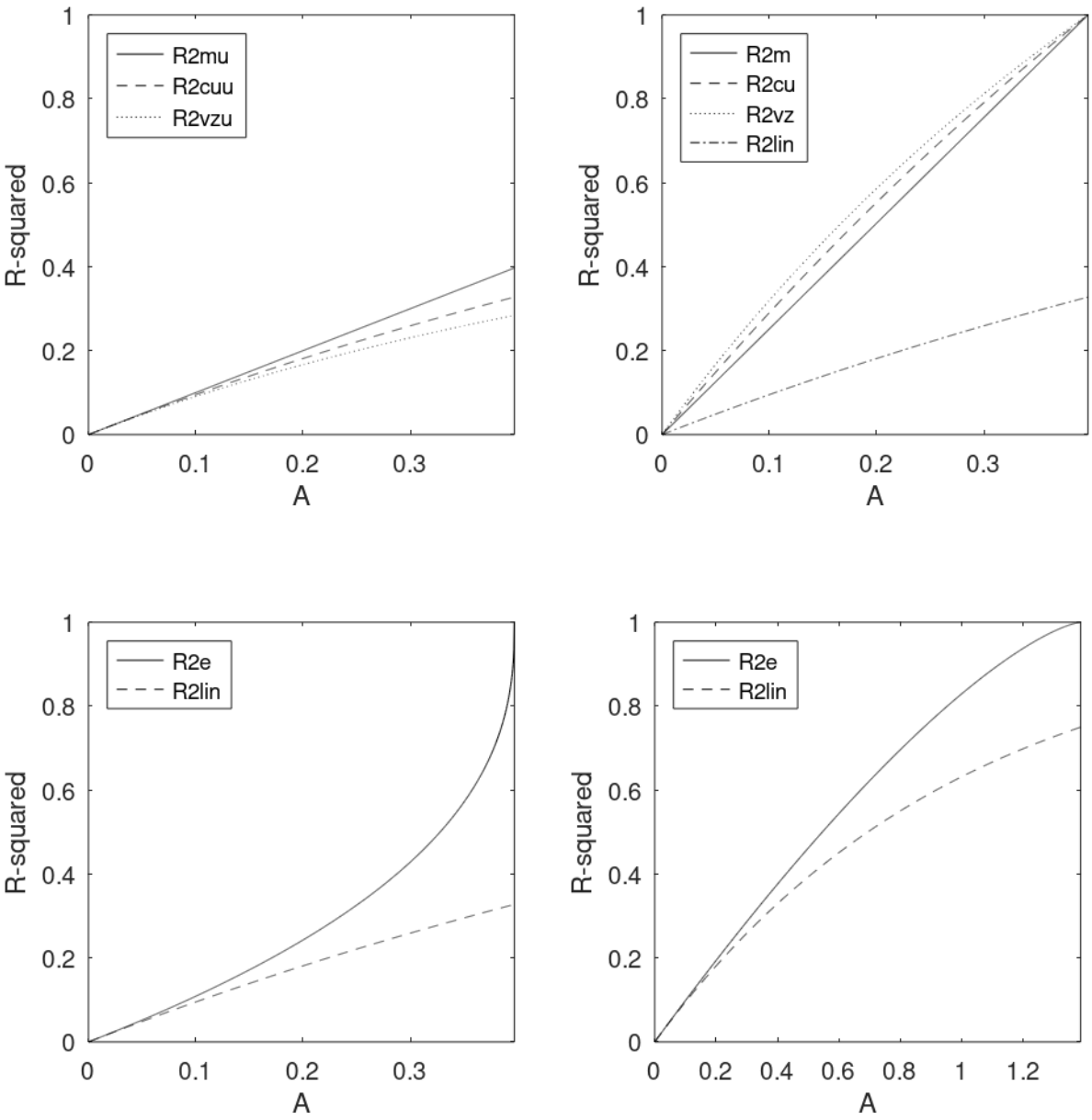
Notes: $i = e, m, cu, vz$. Marginal $R2i = (1-R2i)^{-1}\,\partial R2i/\partial A$.

Figure 1. Upper bound *B* as a function of *E*(*y*)



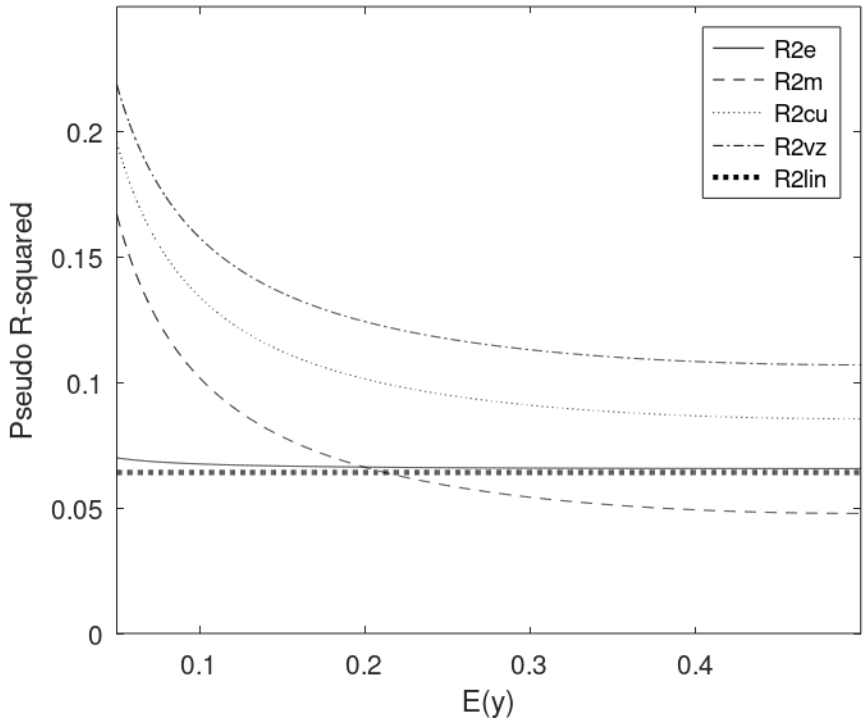
Note: The vertical lines at .05, .20, .50 correspond to illustrations in the text.

Figure 2. Scaling of pseudo R-squared measures



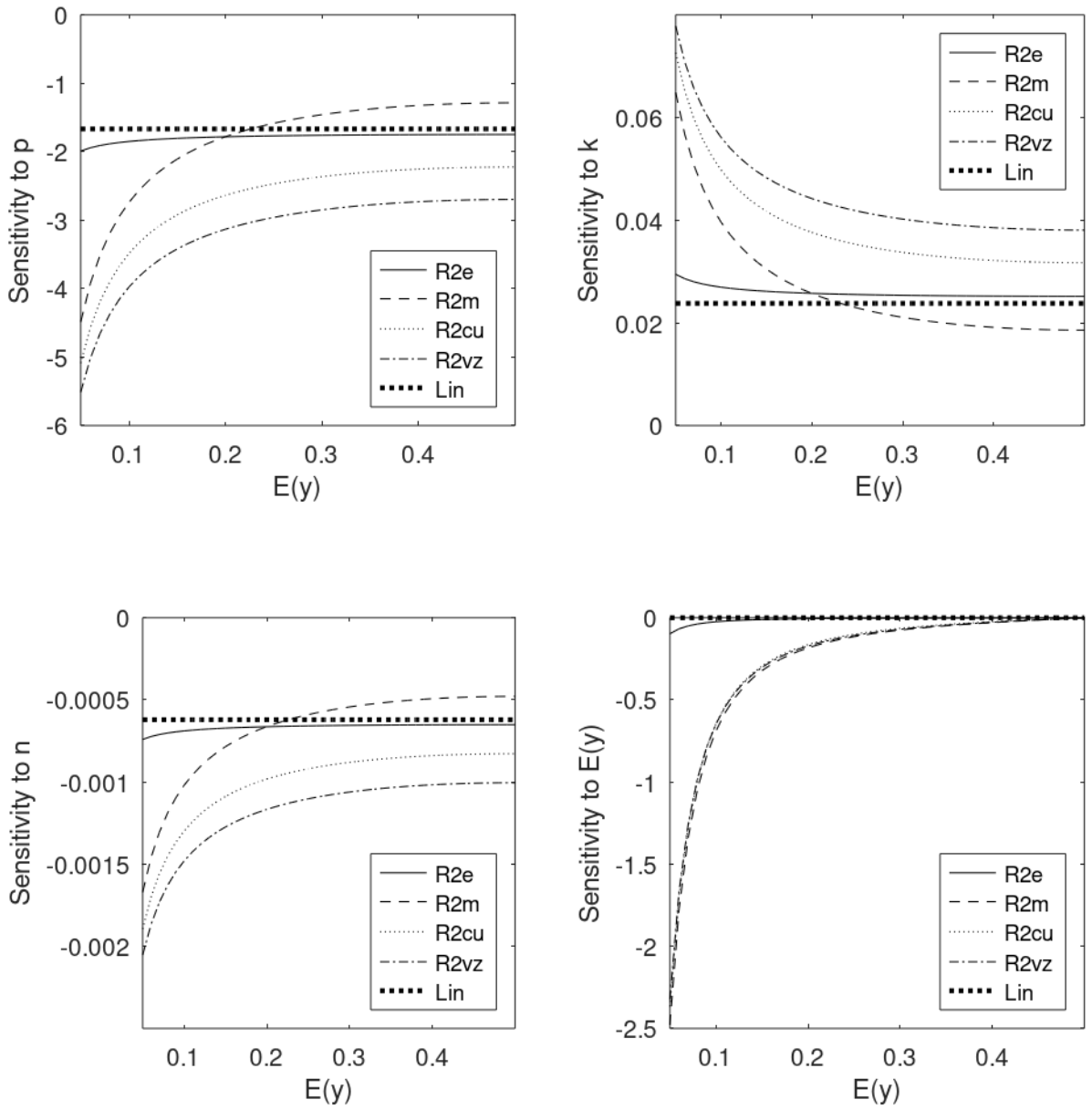
Notes: Measures in the first panel are shown before scaling by the respective multiplicative factors. In the first three panels, $E(y) = .05$ and $0 \le A \le B = .397$. In the bottom right-hand panel, $E(y) = .50$ and $0 \le A \le B = 1.386$.

Figure 3. Pseudo R-squared measures as functions of $E(y)$ $.05 \leq E(y) \leq .50$, chi square $p = .01$, $n = 100$ and $k = 1$



Note: Linear model R-squared (R2lin) is included for reference purposes.

Figure 4. Sensitivity of pseudo R-squared measures to each of the reduced-form arguments



Notes: The sensitivities are the partial derivatives (partial difference for *k*) of $R2i\left(p,k,n,E\left(y\right)\right)$ evaluated at $p=.01, n=100, k=1$ for the range of values $.05\le E\left(y\right)\le .50$ as indicated in the horizontal axis of each panel. Partial derivatives for the linear model (Lin), which do not depend on $E\left(y\right)$, are included for reference.