ON NEGATIVE - VALUED R^2

By

ANASTASSIOS GAGALES
Bank of Greece

One of the first things to learn in statistics is that the coefficient of determination, R^2, by definition takes values always in the closed internal [0, 1]. Occasionally, however, one is puzzled by a computer output reporting negative R^2. How should this be interpreted? Is it an indication that the model fits data poorly? If so, in which direction should the model be modified? Or is it due to rounding» errors and the near singularity of the observation matrix? Moreover, can this reasoning sufficiently justify values of R^2 equal to —19.7?

These and other more wild guesses can create, at least, great discomfort and uncertaintly. Initially, the whole affair is viewed as a mere curiocity and is assertively attributed to rounding errors and the ill-conditionality of the observation matrix. As, however, the frequency of negative R^2 increases, so does the mistrust and skepticism (both exhibit a very strong positive correlation with the frequency of negative R^2 appearances); more fundamental but as yet unintelligible deficiencies of the statistical package are thought to be the villains. Gradually, one reaches a state of absolute mistrust for the package and adopts a nihilistic attitude, coupled with a scornful tone rowards the estimates. Lacking, however, any alternative, investigators willy-nilly accept whatever the computer grinds out. The cynics among them either report bluntly the negative valued coefficients of determination, or they opt not to report this statistic at all. Others, in an obscure footnote aknowledge their incomprehention and desparation while they appeal apologetically to some higher authority. Neither however can come to grips with the invisible forces that are operating behind the scenes. The ensuing paragraphs offer a resolution to this unfortunate state of affairs and a boost of confidence for the estimates.

The most widely used econometric package is the Time Series Processor (TSP). It was developed in the mid-sixties at the University of Chicago, but since then it has undergone substantial revisions pua extensions, keeping pace with the evolution of econometric thinking.

TSP calculates the coefficient of determination using the formula $R^2_{TSP} =$

$$= 1 - \frac{e'e}{Y'AY}$$ where $Y'AY/N$ denotes the sample variance[1,2] of the dependent variable Y and $e = Y - X\hat{b}$. (Cooper 1973, Hall and Hall 1980). As it will be shown shortly, this formula is responsible for the «perverse» behavior of R^2_{TSP} since it is inappropriate when a constant is not included among the regressors. In this case R^2_{TSP} underestimates the proportion of the variance of Y explained by the model. Furthermore, the closer the coefficient of variation of Y (defined as $C_Y = \frac{S_Y}{\bar{Y}}$) is to zero, the greater are the chances that R^2_{TSP} takes on a negative value. It is proved that R^2_{TSP} takes values in the half-closed interval $(-\infty, 1]$, i.e. it does not possess a lower bound.

Let that the linear model $Y = ib_0 + Xb + e$ (1) is fitted to the N. (K+1)—dimensional observational matrix (Y, i, X). e, the residual vector, is by construction orthogonal to all regressors; $(i' \; x')e = 0$. This orthogonality property of e facilitates the decomposition of the variance of Y into two parts, one «explained»

1. $A = I - ii/N$ stands for the idempotent linear operator that transforms the original observation matrix into deviations from its sample mean (see Theil, pp. 12 - 14).

2. The reason for using R^2_{TSP} instead of $R^2 = b'x'Axb/y'AY$ is that the value of e'e is needed anyway in the computation of the covariance matrix, F - statistics etc., whereas the value of b'x'Axb is not used elsewhere in the calculations. To use R^2_{TSP} instead of R^2 is therefore more economical in computer money.

by the model and the residual variance, i.e. $Y'AY = b'x'Axb + e'e$. Hence,

$$R^2_{TSP} - R^2 = 1 - \frac{e'e}{Y'AY}.$$

If instead, we fit a linear model differing from (1) only in that it does not contaiu a constant, i.e. $Y = Xb_0 + e'_0$ (2) the variance of Y will be decomposed as follows:

$Y'AY = b_0'X'AXb_0 + e'_0 - N\bar{e}_0^2$ (3) where $\bar{e}_0$ denotes the sample mean of the residual vector [3]. From (3) an explicit relation between R^2 and R^2_{STP} can be derived;

namely, $R^2 = R^2_{TSP} + \frac{N\bar{e}_0^2}{Y'AY}$ (4) We see, therefore, that the omission of the constant from the regression gives rise to the emergence of a wedge between the sample value of the coefficient of determination and R^2_{TSP}; consequently R^2 is underestimated by an amount that varies inversely with C_y.

A diagram might help clarify the preceding arguments. Let that models (1) and (2) one fitted to the scatter $\{(X_i, Y_i)\}$. Evidently, the second model fits the data poorly and is characterized by a relatively low R^2. Let us now make the following thought experiment: from each Y_i subtract a constant Y_e, to obtain a new scatter denoted $\{(x_i\, y_i - y_e)\}$. Choose Y_e in such a fashion that models(1)and(2)both fit the transformed sample equally well, i.e. if model(1) is employed it will yield a zero estimate for the coefficient of the constant. (It is trivial to show that such a value of Y_e always exist, and is equal to $Y_e = b_0$). The estimates of model (1) are invariant under this affine transformation of the sample.

A cursory inspection of the diagram reveals that as Y_e increases, the fit of model (2) deteriorates. A measure of how poorly model (2) fits the data is the absolute

3. In model (3) it is not usually the case that $i'e = 0$. This implies that except for a set of measure zero, $Ae_c = e_o$, so that model (3) cannot be brought into the form $AY = AX + e_o$.

value of the sample mean of the residuals, i.e. the systematic influences on Y that are incorporated in the random variable. As $|\bar{e}_0|$ increases, the wedge between R^2 and R^2_{TSP} widens (see eq. 4). The two samples $[(X_i, Y_i)]$ and $[(X_i, Y_i - Y_e)]$ however, differ only in the mean and the coefficient of variation of the dependent variable, Y. An inverse association between C_Y and $(R^2 - R^2_{TSP})$ is thus evident.

To express this relation analytically note first that $\bar{e}_0 - \bar{Y} - \bar{X}'b_0 =$

$= [1 - \bar{x}'(x'x)^{-1}x'i]Y_e$, and

$$\bar{e}_0^2 = [1 - \bar{x}'(x'x)^{-1}x'i]^2 (Y_e/\bar{Y})^2 \bar{Y}^2 \equiv \kappa^2 \lambda^2 \bar{Y}^2.$$

Therefore, $N \bar{e}_0^2 / Y'AY = \kappa^2\lambda^2\bar{Y}^2/S^2_Y = \kappa^2\lambda^2/C^2_Y.$

This expression can be exploited in the construction of rather sharp and effective bounds for R^2_{TSP}, namely: $-\kappa^2\lambda^2/C^2_Y \leqq R^2_{TSP} \leqq 1 - \kappa^2\lambda^2/C^2_Y$. Since C_Y can take on any arbitrarily small positive value, R^2_{TSP} cannot be bounded below. Furthermore, the smaller the value of C_Y, the greater the possibility, ceteris paribus, that R^2_{TSP} becomes negative.

A negative valued R^2_{TSP} means that a much better fit and superior explanatory power could have been obtained if a constant were used as the sole regressor. In other words, the sample mean of Y contains valuable information about $\{Y_i\}$, since the bulk of the observations is clustered around it. Consequently, the modification that the model is asking for in the presence of a negative valued R^2_{TSP} is obvious: Relaxation of the constraint $b_0 = 0$.

When does the problem arise?[4] Usually, when theoretical and a priori considerations suggest that a constant should not be used; more often, however, when the regressors are transformed in the presence of heteroscedastic terms. In this case the constant term of the original model disappears. What should be done then; Ideally, the procedure used to calculate R^2 should be appropriately modi fied. If this is not feasible (and usually it is not, or it is a rather costly endeavor) R^2_{TSP} should be interpreted as a lower bound of R^2. Provided that R^2_{TSP} is relatively high and positive, this is a satisfactori procedure. If, however, R^2_{TSP} negative, it should be ignored, with a concomitant loss of some valuable information.

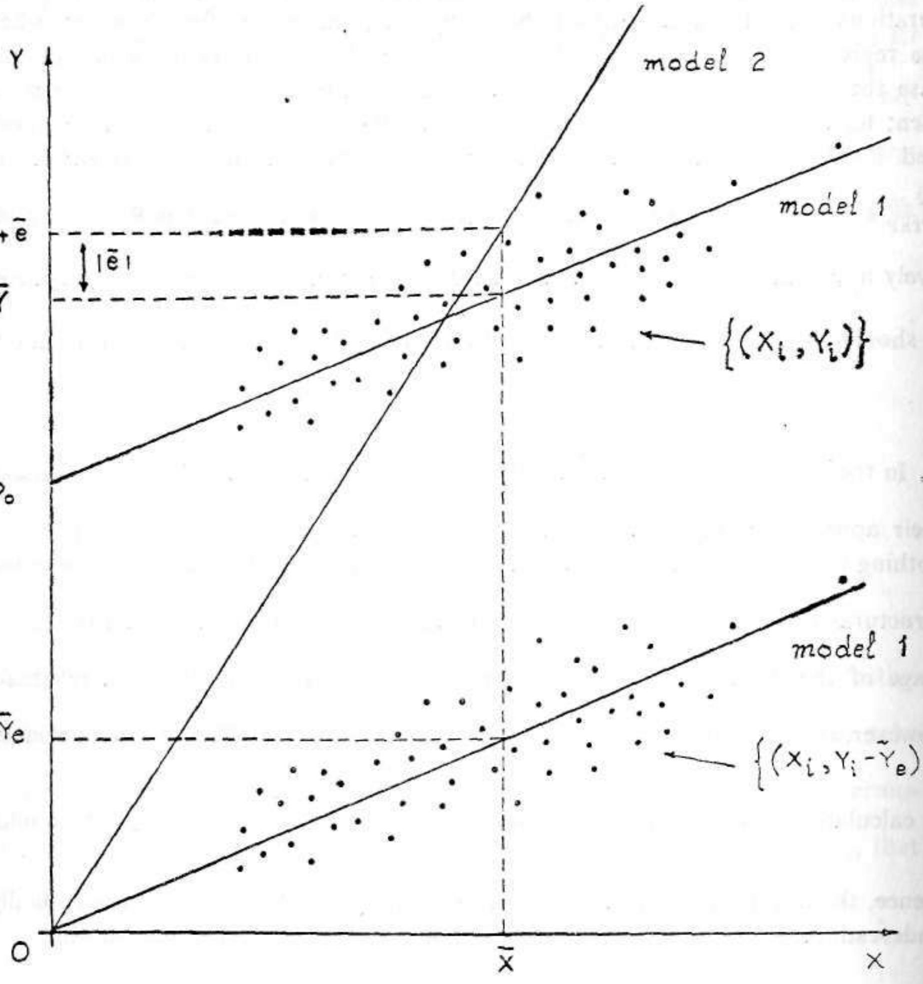
In the case of instrumental variable estimation negative valued R^2_{TSP} can make their appearance even if a constant is included among the regressors. There is nothing perverse in that. TSP calculates the coefficient of determination using the structural residuals, $e = Y - x(\hat{x}'\hat{x})^{-1}\hat{x}'Y$, and not the ones obtained at the second stage of the iteration, $\hat{e} = Y - \hat{x}(\hat{x}'\hat{x})^{-1'}\hat{x}'y$. The mean of the structural residuals however, is not necessarily zero (i.e. in general $i'e = 0 - i'\hat{e}$). The residual variance is calculated in TSP from $\frac{e'e}{N}$ and not from $e'Ae = e'e - (i'e)^2/N$, as it should. Hence, the unexplained variance is overestimated while R^2_{TSP} is disproportionally underestimated. Using the preceding methodology one can establish that

$-m^2q^2/c^2_Y < R^2_{TSP} < 1 - m^2q^2/c^2_Y$, i.e. in instrumental variable estimation the

4. Rounding errors can very well reduce Y'AY disproportionately and give rise to negative valued R^2_{TSP}. In this note however, we concentrated on only two, but certainly the most frequent source of high (in absocute terms) negative values for R^2_{TSP}.

5. As in the case of ordinary least square, statistical inference should be carried out using the structural (e) and not the second stage (e) residuals.

FIGURE 1



coefficient of determination does not admit a finite lower bound. What actio: should be taken in the presence of negative-valued coefficients of determination The answer remains unchanged: this statistic should be ignored. After all, the coeffi cient of determination, adjusted or not, is not the decisive criterion in performin specification analysis.

LITTERATURE

Cooper, J. P. Econometric Software Package, User's manual,May 1973.

Hall, B. and R. Hall : Time Series Processor, Version 3, 5, User's Manual Nov. 1980.

Theil, H. : Principles of Econometrics, John Wiley, 1971.