MEASURING GOODNESS OF FIT IN LINEAR AND NONLINEAR*

Walter Haessel
The Pennsylvania State University
August 1977

Agricultural experimentation

AAEA paper, July 31 - Aug 3, 1977, San Diego, CA

Forthcoming, Southern Economic Journal

I. INTRODUCTION

When a linear model which includes an intercept is estimated using ordinary least squares (OLS), a very useful summary statistic is the coefficient of determination, R^2. In that case, the R^2 measure is restricted to the interval [0,1], and has a simple interpretation as the proportion of the variation about the sample mean of the dependent variable that is associated with variation in the explanatory variables. Unfortunately, when the linear model is estimated by some method other than OLS or when the model does not contain an intercept, the conventional R^2 measure is no longer necessarily restricted to the [0,1] interval [2;7]. In this paper, two goodness-of-fit (GOF) measures are discussed which have a straight-forward, intuitive interpretation, which can be used with any estimation method, and can be applied to nonlinear models as well as linear models. The two measures differ only in that one measures GOF about the sample mean of the dependent variable, while the other meausres GOF about the origin.

II. GOF MEASURES

Consider a general model of the form

$$(1) \quad Y = f(X;\theta) + u,$$

where Y is an n x 1 vector of observations on the dependent variable, X is an n x k matrix of n observations on k explanatory variables, θ is a k x 1 vector of nonrandom, unknown coefficients to be estimated, and u is an n x 1 vector of unobservable error terms with mean 0 and covariance matrix $\sigma^2 V$. The k explanatory variables may include exogenous variables,

lagged values of the dependent variable Y in the case of time series data, or values of other endogenous variables if equation (1) is one equation in a system of simultaneous equations. Let θ^* be an estimator of θ and let $Y^* = f(X;\theta^*)$ be an estimator of Y. What is desired is a measure of how well the vector Y^* represents Y, or how well the variation in Y is accounted for by the variation in Y^*. Two types of variation can be differentiated: variation about the sample mean $\bar{y}$ of Y (the usual case in OLS with an intercept), and variation about the origin or total variation in Y. Since these are inherently different, separate GOF measures are discussed for the two cases. For a GOF measure to be useful as a descriptive statistic it should have a straight-forward and meaningful interpretation. This requirement almost necessitates that the measure be restricted to some definite range (such as the closed intervals [0,1] or [-1,1]). In addition, because of the widespread use of the conventional R^2, any proposed measure about the mean should be equivalent to the usual R^2 in the case of OLS with an intercept. The GOF measures discussed in this paper satisfy these conditions.

The dependent variable Y, and its estimator Y^*, are vectors in n-dimensional space. Let $P = \lambda Y^*$ denote an orthogonal projection of Y onto Y^*, where λ is a scalar to adjust the length of Y^* to P. Then Y, P, and (Y-P) form a right-angled triangle, and by the theorem of Pythogoras,

$$(2) \quad \Sigma y_i^2 = \Sigma p_i^2 + \Sigma (y_i - p_i)^2,$$

where $p_i = \lambda y_i^*$ are the elements of P. Dividing both sides of (2) by Σy_i^2 results in

$$(3) \quad 1 = \Sigma p_i^2 / \Sigma y_i^2 + \Sigma (y_i - p_i^2) / \Sigma y_i^2.$$

The term $\Sigma p_i^2/\Sigma y_i^2 = \lambda^2 \Sigma y^{*2}_i/\Sigma y_i^2$ can be interpreted as the proportion of the variation about the origin of the dependent variable that is associated with, or explained by, a vector of "optimal length" in the direction of Y*. Thus, λ is a scale-factor to adjust the length of Y* to the length of P, which is the optimal length of Y* to explain Y. This is because (Y-P) is the vector of shortest distance between the point Y and the vector Y*. Alternatively, the term $\Sigma p_i^2/\Sigma y_i^2$ may be viewed as $\text{Cos}^2\Psi$, where Ψ is the angle between the vectors Y and Y*. This can also be calculated as (see [8])

$$(4) \quad \text{Cos}^2\Psi = [\Sigma y_i y_i^*]^2/[\Sigma y_i^2 \Sigma y^{*2}].$$

This measure is restricted to the interval [0,1] and has a straight-forward, intuitive interpretation. $\text{Cos}^2\Psi = 1$ would indicate a perfect fit since the vectors Y and Y* are pointed in the same direction, and $\text{Cos}^2\Psi = 0$ would indicate that Y* provided no explanation of the direction of Y since the two vectors are orthogonal. In other words, if $\text{Cos}^2\Psi = 0$, the elements y_i^* of Y* provide no explanation of how the elements y_i of Y vary about the origin.

It is not always desired to explain the total variation in Y about the origin. In many cases the economic model is specified in terms of explaining variation in the deviations about the sample mean $(y_i - \bar{y})$, where $\bar{y} = \Sigma y_i/n$. This would be the case whenever the model in (1) contains an intercept. Define $\bar{Y}$ as an n x 1 vector with elements equal to $\bar{y} = \Sigma y_i/n$, and $\bar{Y}^*$ as an n x 1 vector of elements equal to $\bar{y}^* = \Sigma y_i^*/n$. The vectors $\bar{Y}$ and $\bar{Y}^*$ will have the same direction, but in general will not be of the same length. When the model is specified to include an intercept, a GOF measure equivalent to (4) would be

$$(5)\quad \cos^2\phi = \Sigma(y_i - \bar{y})(y_i^* - \bar{y}^*)^2 / \Sigma(y_i - \bar{y})^2 \Sigma(y_i^* - \bar{y}^*)^2,$$

where ϕ is the angle between the vectors $(Y - \bar{Y})$ and $(Y^* - \bar{Y}^*)$. Note that (5) is just the square of the correlation coefficient between y_i and y_i^*. Thus, Cos ϕ would have the usual interpretation given to correlation coefficients as measuring the degree of linear association between y_i and y_i^*. Since it is possible to make arguments analogous to those preceding equation (4), $\cos^2\phi$ can also be interpreted as the proportion of the variation in y_i about the sample mean that is explained by a vector of "optimal length" in the direction of $(Y^* - \bar{Y}^*)$.

The choice between the GOF measures defined in equations (4) and (5) should be based on the type of variation it is desired to explain. If the model to be estimated is specified with an intercept, then the measure in (5) is appropriate. If, on the other hand, the economic model is specified without an intercept, the appropriate measure would be given by equation (4). However, even in this case, the square of the correlation coefficient between y_i and y_i^* has a certain amount of intuitive appeal. A further attraction of this measure is its ease of computation. All that is required is to regress Y on Y* and an intercept using OLS and the resulting R^2 will be the GOF measure defined in 5.

III. COMPARISON WITH OTHER MEASURES

The conventional coefficient of determination, R^2, computed for linear models with intercepts estimated by ordinary least squares is well known. The measure defined in (5) to be used with models which contain intercepts is identical to the conventional R^2 if the model is estimated by ordinary least squares. If the linear model is estimated by ordinary least squares and does not contain an intercept, the conventional

interpretation of R^2 is no longer appropriate and is no longer restricted to the [0,1] interval. Theil [6] and Aigner [1] have recommended using measures equivalent to equation (4) in that case.

A common problem in empirical research is generalized least squares where the conventionally defined R^2 breaks down. Theil [5,221] proposed using $\text{Cos}^2\Psi$ as a GOF measure, where $\text{Cos}^2\Psi$ is calculated using the transformed data. Buse [2, 10] proposes using $\text{Cos}^2\phi$, where $\text{Cos}^2\phi$ is also calculated on the transformed variables. The foregoing discussion suggests this choice should be made on the basis of whether or not the transformed model contains an intercept. On the other hand, since it is desired to explain variation in Y, not the transformed data, an argument can be made to measure GOF on the original data using either (4) or (5).

Theil [5, Ch. 2] defined an inequality coefficient

$$u^2 = \Sigma(y_i - y_i^*)^2/\Sigma y_i^2$$

to be used in evaluating the accuracy of forecasts. This measure, unfortunately, is unbounded from above. Consequently, it is difficult to give an intuitive interpretation to values other than $u^2 = 0$ which indicates a perfect forecast.

Jobson [4] has defined an equality coefficient which can be related to Cos ϕ as

$$J = \frac{2M_y M_{y*} \text{ Cos } \phi}{M_y^2 + M_{y*}^2 + n(\bar{y} - \bar{y}^*)^2}$$

where $M_y^2 = \Sigma(y_i - \bar{y})^2$ and $M_{y*}^2 = \Sigma(y_i^* - \bar{y}^*)^2$. Since $2 M_y M_{y*} < (M_y^2 + M_{y*}^2)$ whenever $M_y \neq M_{y*}$, we have $|J| \leq |\text{Cos } \phi|$, with the strict equality holding only if either Cos $\phi = 0$ or $M_y = M_{y*}$ and $\bar{y} = \bar{y}^*$. The range of J is [-1,1], with 1 indicating a perfect fit and 0 indicating no fit. Note that M_y^2, M_{y*}^2 and $n(\bar{y} - \bar{y}^*)^2$ are the squared Euclidean lengths of Y, $\bar{Y}$, and $(\bar{Y} - \bar{Y}^*)$

respectively. Thus, J will be less than Cos $\emptyset$ whenever Y and $\bar{Y}$ or $\bar{Y}$ and Y* are not of equal length, and as a GOF measure J will be sensitive to these differences whereas Cos $\emptyset$ is not. It is, however, difficult to give any meaningful intuitive interpretations to values of other than J = 1 (perfect fit) or J = 0 [$(Y - \bar{Y})$ and $(Y^* - \bar{Y}^*)$ are orthogonal], and these interpretations are identical to the interpretations given to Cos $\emptyset$.

Carter and Nager [3] have recently proposed a GOF measure for single equations in a simultaneous system as well as for the entire system based on the reduced form equations. Thus, if it is desired to have a measure of how well an estimated structural equation fits the actual data, $\text{Cos}^2\emptyset$ or $\text{Cos}^2\Psi$ could be used as a summary statistic.

IV. SUMMARY AND CONCLUSIONS

In this paper it has been argued that useful summary measures of GOF for linear or nonlinear models are the square of the cosine of the angle between Y and its predictor Y*, or the square of the cosine between these vectors as deviations from their means, i.e., $(Y - \bar{Y})$ and $(Y^* - \bar{Y}^*)$. The choice between these two measures should depend on whether the model is intended to explain total variation in the dependent variable (i.e., about the origin), or whether it is intended that the model should explain variation about the mean of the dependent variable. The latter would be the case whenever the economic model is specified with an intercept. The two measures are restricted to the range [0,1], reduce to the conventional R^2 measure in the case of OLS with an intercept, and are equivalent to measures proposed in certain special cases. Furthermore, it is possible to give these measures a straight-forward and meaningful intuitive interpretation for any linear or nonlinear model and for any estimation method.

Walter Haessel

The Pennsylvania State University

References

1. Aigner, D. J. Basic Econometrics. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

2. Buse, A. "Goodness of Fit in Generalized Least Squares Estimation," The American Statistician, 27:106-108 (1973).

3. Carter, R. A. L. and A. L. Nager. "Coefficients of Correlation for Simultaneous Equation Systems," Journal of Econometrics, forthcoming.

4. Jobson, J. D. "A Coefficient of Equality for Questionnaire Items With Interval Scales," Journal of Educational and Psychological Measurement, forthcoming.

5. Theil, H. Economic Forecasts and Policy, (Amsterdam: North-Holland Publishing Co., 1958).

6. Theil, H. Principles of Econometrics, (New York: John Wiley and Sons, 1971).

7. Tomek, W. G. "R^2 in TSLS and GLS Estimation," American Journal of Agricultural Economics, 55:670 (1973).

8. Wonnacott, R. J. and T. H. Wonnacott. Econometrics. New York: John Wiley and Sons, 1970, Ch. 15.

Footnotes

*Paper number 5324 in the journal series of the Pennsylvania Agricultural Experiment Station. I am grateful to R. A. L. Carter and J. D. Jobson for numerous helpful discussions. Comments by M. C. Hallberg, R. H. Warland and an anonymous referee on earlier drafts were very helpful. The author is solely responsible for any remaining deficiencies.

[1]In generalized lease squares, if $E\ u\ u' = V$, and $T'T = V^{-1}$, then Theil proposed $\text{Cos}^2\Psi = \Sigma z_i z_i^* / \Sigma z_i^{*2}$, where $Z = TY$, $Z^* = TX\theta_a$, and $\theta_a = (X'V^{-1}X)^{-1}X'V^{-1}Y$ is the Aitken estimator.