COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 381

FURTHER NOTES ON THE MISUSE OF R^2

Gary Smith

October 22, 1974

FURTHER NOTES ON THE MISUSE OF R^2

by

Gary Smith*

When estimating a linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \qquad i = 1, \ldots, n$$

one is often interested in the accuracy both of the estimated parameters and of in-sample forecasts of y_i based on these parameter estimates

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij} \qquad i = 1, \ldots, n .$$

In practice, the coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

is often explicitly used as a measure of goodness of fit or predictive precision in forecasting y_i and implicitly used as a decision tool for improving the accuracy of the estimated parameters through the deletion of variables

*Assistant Professor of Economics, Yale University. The research described in this paper was undertaken by grants from the National Science Foundation and from the Ford Foundation.

whose coefficients are statistically insignificant.

In a recent article in AmStat, James Barrett notes the usefulness of R^2 as a decision tool but argues that a confidence interval for the expected value of y is of more practical value than R^2 in assessing predictive precision, inasmuch as R^2 depends upon the true values of the β_j while a confidence interval does not.

In the first part of this paper, I note that Barrett's two alternatives are not strictly comparable as one is an absolute and the other a relative measure of goodness of fit. The limited usefulness of each in its own domain is more deeply explored and a broader view of the alternatives advocated. In the second part of the paper, I point out that R^2 is also of extremely limited usefulness as a decision tool of the type discussed by Barrett.

I. R^2 as a Measure of Goodness of Fit

Predictive accuracy can be measured in either absolute or relative terms. While an absolute measure is usually difficult to interpret or evaluate, a relative measure invokes a particular standard of comparison. Although absolute and relative measures are not comparable, they are obviously related since an absolute measure of goodness of fit is the basis of a relative measure.

Barrett advocates measuring predictive precision by a confidence interval for the expected value of y . This is an absolute measure which would provide specific point forecasts with a welcome touch of modesty. However, as a general measure of the predictive accuracy of an equation, confidence intervals require an assumed distribution for the ϵ_i and suffer from the obvious defect that they are dependent upon the selection of particular values for the explanatory variables. With $k > 2$, it would be quite inconvenient to graph or otherwise display these intervals for all possible values of the explanatory variables. In addition a comparison of particular confidence intervals for alternative sets of explanatory variables would require the establishment of either a deterministic or stochastic correspondence between the occurrence of particular values of one set of variables and the occurrence of particular values of the alternative set. Even with this, one would still have the awkward problem of comparing vectors of confidence intervals unless one were willing and able to specify multivariate likelihood and loss functions.

It would be considerably easier to focus on predictive accuracy for the actually observed sample values of the explanatory variables. One might plot these intervals on a graph with the units of y on one axis and i (the observation label) on the other axis, and calculate a (possibly weighted) average length of the confidence interval for the sample data. Although shortness is a desirable property, it is not the only consideration. It might also be of interest to know in which direction and by how far the interval errs, when it does err. While such an exercise might have some heuristic value, it turns out that two of the more obvious potential weightings give results which are closely related to the far more popular alternative of measuring predictive accuracy by a simple comparison of the model's forecasts with the actually observed values of y. (Proofs are given in the Appendix.)

The Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

is a commonly accepted (absolute) summary measure of a model's actual forecast errors, and is consistent with the minimization of squared errors in most estimation procedures. Interestingly, using the sample data for a comparison of the MSE's for alternative models is closely related to a comparison of Average Squared Confidence Intervals for the expected value of y, in that the latter can be shown to be equal (under the usual assumptions) to

$$ASCI = 4t^2_{\alpha/2}\left(\frac{k+1}{n-k-1}\right)MSE$$

where $t_{\alpha/2}$ is from student's t-distribution. Similarly, the squared confidence interval for the expected value of y when the explanatory variables are at their sample means is

$$SCI(\bar{X}) = 4t^2_{\alpha/2}\left(\frac{1}{n-k-1}\right)MSE \ .$$

For given n and k, these three alternative measures are equivalent. However, MSE ignores the size of k while ASCI and SCI($\bar{X}$) penalize the model with more explanatory variables. It is of course possible (and probably desirable) to modify the MSE criterion to include a degrees of freedom adjustment; however it is not obvious that one would want to use either of the specific adjustments implied by ASCI or SCI($\bar{X}$). One popular alternative is the Standard Error of Estimate

$$SEE = \left(\frac{n}{n-k-1}\right)MSE \ .$$

Whichever specific criteria is adopted, one should be careful not to compare predictive accuracy for different variables. A transformation is necessary for example if one model forecasts y while the alternative model forecasts $\text{Log}(y)$, y^2 , or y/z . The following discussion of relative measures assumes that any needed transformations have been made and that MSE is used as a summary measure of goodness of fit. A degrees of freedom adjustment would affect the detail but not the substance of the argument.

A goodness of fit comparison of alternative theoretical models may be ambiguous if different bodies of data are used or if the explanatory variables are themselves stochastic.[1] In addition, it is usually interesting (and often sobering) to examine not only the predictive accuracy of competing theoretical models but also to construct some benchmark level of performance, in order to determine if the race is between slow or fast horses. For these purposes, it would be desirable to examine the forecasting ability of a naive model such as might be used by someone who was given unlabeled data.

It is instructive in this context to view R^2 as a relative measure of goodness of fit where the comparison is between one's theoretical model and the naive forecaster who uses the sample mean as his forecast:

$$R^2 = 1 - \frac{\text{MSE}(\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij})}{\text{MSE}(\hat{y}_i = \bar{y})} .$$

Viewed in this way, the adequacy of R^2 as a measure of relative predictive

[1] For example, the stock market is a good forecaster of investment demand, but this relation may be of little predictive use if the stock market cannot itself be accurately forecast.

accuracy depends upon the adequacy of the sample mean as a benchmark standard. If the sample mean is a trivial or irrelevant competitor, then R^2 is uninteresting and need not be reported.

In practice it is not difficult to think of situations in which one will obtain an impressively high R^2 , not because of the success of one's model but rather because of the inadequacy of the benchmark model. Generally, these will be cases where y is very stable and predictable but has a large variance.

To be more specific, it is necessary to look more carefully at the

$$\text{MSE} = \sigma_\epsilon^2[1 - R_{\epsilon x}^2]$$

and the sample variance of y

$$\text{var}(y) = \sum_i \sum_j \beta_i\beta_j\sigma_i\sigma_j\rho_{ij} + 2 \sum_j \beta_j\sigma_\epsilon\sigma_j\rho_{\epsilon j} + \sigma_\epsilon^2$$

where σ_i^2 = sample variance of X_i
σ_ϵ^2 = sample variance of ϵ
ρ_{ij} = sample correlation coefficient for X_i and X_j
$\rho_{\epsilon j}$ = sample correlation coefficient for ϵ and X_j
$R_{\epsilon X}$ = sample multiple correlation coefficient for ϵ and all k X_j

Barrett argues that R^2 is misleading because it depends upon the true values of β_j while MSE (or alternative absolute measures) do not. It is clear from the explicit expressions that the variance of y also depends upon the variances of the X_j while MSE does not,[2] and that although

[2] This could have been deduced from the fact the units of β_j can be arbitrarily changed by altering the units in which X_j is measured.

both expressions have a number of terms in common, they enter in very different ways so that a change in a parameter such as ρ_{12} will affect R^2 in part because of the altered goodness of fit of one's theoretical model and in part because of the altered fit of the naive benchmark model.

Whether this is misleading or not depends upon one's expectations. Since R^2 is a relative measure of goodness of fit, it must necessarily reflect the success of the benchmark model as much as the theoretical model. R^2 will consequently be misleading whenever it is interpreted without reference to the appropriateness of the benchmark standard. The appropriateness of the benchmark model is a subjective question which depends upon the particular variable being forecast. It is not a simple question of the magnitudes of the β_j or the σ_j or any of the other parameters. The relevant question is instead whether or not the model $\hat{y} = \bar{y}$ gives a reasonable measure of how difficult y is to forecast.

These points are most clear in the simple regression case. Here

$$y_i = \alpha + \beta X + \epsilon_i \qquad i = 1, \ldots, n$$

and

$$\text{MSE} = (1 - \rho^2)\sigma_\epsilon^2$$

$$\text{var}(y) = \sigma_\epsilon^2 + \beta^2\sigma_X^2 + 2\beta\sigma_X\sigma_\epsilon\rho$$

$$R^2 = \frac{1}{1 + \dfrac{(1 - \rho^2)}{\left(\rho + \beta\dfrac{\sigma_X}{\sigma_\epsilon}\right)^2}}$$

where σ_X^2 and σ_ϵ^2 are the sample variances of X and ϵ, and ρ is the sample correlation coefficient between X and ϵ. Ceteris paribus, R^2 will approach 1 as $\sigma_\epsilon^2 \to 0$, $\rho^2 \to 1$, $\beta^2 \to \infty$, or $\sigma_X^2 \to \infty$; while MSE will approach zero only as $\rho^2 \to 1$ or $\sigma_\epsilon^2 \to 0$. Thus, if either the sample variance of X or the absolute value of β is very large, this will raise R^2 by worsening the fit of the naive model. And in the nonextreme situation, a change in either β, σ_X^2, σ_ϵ^2, or ρ has an ambiguous effect[3] on the fit of the naive model and consequently on R^2. Thus, even in this simple case the success of the naive model is not simple, and it is consequently difficult to say whether or not R^2 is appropriate. As illustrations, however, I will discuss two obvious situations[4] in which models that are only modestly successful will have impressive R^2.

<u>Trend Variables</u>

If y is a trend dominated or a highly autoregressive time series, then $\bar{y}$ will give particularly inaccurate forecasts. If x is similarly trend dominated or autoregressive, then one should expect to be able to find coefficients such that $y = \alpha + \beta x$ will dramatically outforecast the average level naive model, even if x and y are conditionally independent.

Consider for example two independently generated trend variables

[3] If $\rho = 0$, then matters are considerably simplified. The variance of y is then always increased and R^2 always reduced by an increase in β^2 or σ_X^2. In the multivariate case, if the explanatory variables are uncorrelated with one another as well as with ϵ, then an increase in any $\beta_j^2\sigma_j^2$ will unambiguously worsen the fit of the naive model and thereby improve R^2.

[4] Related discussions are contained, for example, in Bartlett, Yule, and Yule and Kendall.

$$y_i = a + bT + \epsilon_i \qquad i = 1, \ldots, n$$

$$x_i = c + dT + u_i$$

where $T = i$, and ϵ_i and u_i are independent disturbance terms. If we assume for convenience that the sample covariances among ϵ , u , and T are zero, then the simple coefficient of determination between x and y will be

$$R^2 = \frac{1}{\left[1 + \frac{\sigma_\epsilon^2}{b^2\sigma_T^2}\right]\left[1 + \frac{\sigma_u^2}{d^2\sigma_T^2}\right]} .$$

This will approach 1 as σ_ϵ^2 and σ_u^2 approach zero, as d^2 and b^2 become very large, and as n becomes very large. Thus, R^2 will be high if one considers two unrelated variables which have each changed rapidly and smoothly over a considerable period of time.

For the specific case where

$$\frac{\sigma_\epsilon^2}{b^2} = \frac{\sigma_u^2}{d^2} = \lambda$$

we have the following sample R^2 :

λ \ n	10	20	50	100
4	.454	.797	.963	.990
2	.648	.890	.981	.995
1	.795	.942	.990	.998
.5	.889	.971	.995	.999
.25	.942	.985	.998	.999

Distinct Subsets

If the observations on y and x are both taken from two widely disparate situations, then $\bar{y}$ will forecast poorly and the sample R^2 may be impressively large even though the only association that exists is the coincidence of the two disparate sample situations. For example, one might be using pre- and post-war data, or war and nonwar years, or cross-sectional data involving distinct subgroups. In this latter case, the sample might include distinct subgroups (in terms of region or nationality, race, religion, occupation, education, politics) which differ markedly in a number of ways (such as diet, family size, handgun ownership, movie attendance, annual rainfall, unemployment, church attendance); these characteristics will then appear to be highly correlated, even though they are unrelated within each subgroup. Rather than using specific variables, one could alternatively boost R^2 by using ad hoc dummy variables which coincide with the disparate situations or with the "abnormal values of y."

As an example of the general situation, consider this following simple model:

$$y_i = \alpha_1 + \epsilon_1 \qquad i = 1, \ldots, m$$

$$x_i = \beta_1 + \epsilon_2$$

$$y_i = \alpha_2 + \epsilon_1 \qquad i = m+1, \ldots, n$$

$$x_i = \beta_2 + \epsilon_2$$

where the α_i and β_i are constants, and ϵ_1 and ϵ_2 are independent disturbance terms. For convenience, assume that the sample means of ϵ_1

and ϵ_2 are zero and the sample covariance between ϵ_1 and ϵ_2 is zero. The simple correlation coefficient between y and x (for all n observations) can then be shown to be

$$R^2 = \frac{1}{\left[\dfrac{\sigma^2_{\epsilon_2}}{\frac{m}{n}\frac{n-m}{n}(\beta_1 - \beta_2)^2} + 1\right]\left[\dfrac{\sigma^2_{\epsilon_1}}{\frac{m}{n}\frac{n-m}{n}(\alpha_1 - \alpha_2)^2} + 1\right]} .$$

The smaller the variances of ϵ_i , the more nearly equal m and n-m , and the larger the difference between α_1 and α_2 and between β_1 and β_2 , the larger will be R^2 . In, for example, the quite plausible situation where $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = .25$, $m = n-m$, and $(\alpha_1 - \alpha_2) = (\beta_1 - \beta_2) = 10$, R^2 would be .98.

Alternative Goodness of Fit Criteria

I have argued here that since R^2 is a relative measure of goodness of fit, any interpretation of it must consider the stringency of the standard of comparison. If the sample mean is an uninteresting naive model, then R^2 is uninteresting and one should use a more challenging benchmark.

There are a large variety of alternative naive models to select from, and indeed one important question is in what ways the scope of the search should be limited. It seems clear that the naive model should not be theoretically motivated, which could be enforced by requiring that the identification of the data not affect the selection process. It is not as clear whether or not data on variables other than y_i should be used. Models in which right-hand side variables are selected randomly or mechanically according to purely statistical criteria are properly called naive, and yet at times are hard to distinguish from what pose in the journals as theoretical models.

For self-testing, a randomly selected benchmark model is unappealing; but there might be times when it is of interest to know how well frankly statistical models perform.[5] Simplicity is definitely a virtue, and for within-sample comparisons the naive model should probably be limited to having no more parameters than the theoretical model. Whatever the particular choice, it is important to acknowledge that a variety of options are available and to exercise some care in selecting a benchmark naive model.

For time series data, a variety of simple naive models are plausible, such as

$$\hat{y}_i = y_{i-1} \qquad \text{[no change]}$$

$$\hat{y}_i = y_{i-1} + (y_{i-1} - y_{i-2}) \qquad \text{[same change]}$$

$$\hat{y}_i = \bar{y} \qquad \text{[average level]}$$

$$\hat{y}_i = y_{i-1} + \overline{\Delta y} \qquad \text{[average change]}$$

$$\hat{y}_i = \alpha + \beta T \qquad [\text{trend,}\quad T = i\ ]\ .$$

Since these models may give very different forecasts and consequently provide markedly different benchmark standards, it might be preferable to estimate the parameters of a more general autoregressive model

[5] See Coen, et al., and the ensuing discussion, including the remark by Durbin that, "As far as short term economic forecasting is concerned, my feeling is that it is not clear at present whether one does better to fit economic variables based on postulated relationships between the variables, or to use statistical forecasting of a frankly ad hoc character."

$$\hat{y}_i = \sum_{\tau=1}^{q} \alpha_\tau y_{i-\tau}$$

or perhaps use a moving average or mixed autoregressive-moving average model (see Nelson, for example).

For cross section data, it is not as easy to think of reasonably simple alternatives to $\bar{y}$. One might look at the distribution of the y_i for evidence of outliers or distinct subsets, or perhaps directly regress y_i on available X_j data without resort to any theoretically inspired exclusion restrictions. It may of course be necessary (and fairer for within-sample tests) to exclude some X_j either randomly or on the basis of purely statistical criteria.

II. R^2 as an Hypothesis Test and Decision Rule

Barrett comments that, "Researchers find that R^2 is a handy index in searching for a useful regression equation including a subset of $x_1, x_2, \ldots, x_k$." It is well known that under the assumption that $\epsilon \sim N[0, \sigma^2 I_N]$, one can test the joint hypothesis that m of the β_j are zero by examining the increase in MSE , or equivalently the reduction in R^2 , when the m associated X_j are deleted from the regression equation. Alternatively, one can conduct this joint significance test at the $\alpha\%$ level by asking whether or not the origin is in an m-dimensional $(1-\alpha)\%$ confidence region for these m coefficients.

This provides another way of viewing our earlier results since, ceteris paribus, R^2 will be high and zero will be outside a confidence interval for β in a simple regression equation: (i) when β is far from zero;

(ii) when σ_X^2 is large or σ_ϵ^2 small, as this will make $\sigma_{\hat{\beta}}$ small; and (iii) when ϵ and X are highly correlated as X will account for most of the variation in ϵ and thus reduce $\hat{\sigma}_\epsilon$, which will reduce $\hat{\sigma}_{\hat{\beta}}$.

However, I will argue here that it is a dubious procedure at best to delete a variable whenever one can accept the null hypothesis that its coefficient is zero; i.e., to mechanically assume that a particular arbitrary value of a coefficient is true whenever this value is not contradicted by the data. The following discussion will ignore the question of ill-behaved disturbance terms and emphasize instead the inappropriateness of routinely using hypothesis tests for decision making.

One of the problems with this deletion procedure is that if we test each β_j in a search for coefficients for which zero cannot be rejected as a parameter value, then the probability of incorrectly rejecting at least one null hypothesis will be greater than the probability of rejecting a particular null hypothesis. Thus if we conduct each test at the 5% level, then the probability of incorrectly including at least one variable will be greater than 5%, though less than .05 multiplied by the number of tests. Similarly, the probability of committing at least one type II error by incorrectly excluding a variable will also be increased.

While it is clear that one should take into account the fact that more than one test will be conducted, it is not obvious in which direction the procedure should be modified; i.e., whether each test should be conducted at a higher or lower significance level.[6] To answer this would require a

[6] The usual modification seems to be to lower the probability of Type I Error for each test, which increases the probability of incorrectly excluding a variable.

resolution of the more basic question of why zero is given the presumptive weight of being the null hypothesis. Since any point in the confidence interval would be accepted if it were tested as a null hypothesis, one should wonder why the origin is to be preferred over all other unrejected hypotheses, and in particular why it should be preferrable to the unconstrained point estimate.

The usual answer is that all of the parameters often cannot be accurately estimated because of a limited number of observations on highly collinear variables; omitting some variables will reduce the variance and may improve the mean squared errors of the estimates of the coefficients of the remaining variables. In this situation, however, zero will often be in the confidence interval not because the point estimate is close to zero but rather because the confidence interval is very large; that is, zero is unrejected only because of the acknowledged imprecision of the estimate. And while it is true that the use of correct a priori information will improve efficiency, the automatic deletion of a variable is an ad hoc rather than an a priori restriction that the associated coefficient is zero, which may or may not improve matters and in any case should be inferior to an a priori inspired restriction.

It is useful here to write the model in matrix form as

$$\underset{nx1}{Y} = \underset{nxk-m}{X_1} \underset{k-mx1}{\beta_1} + \underset{nxm}{X_2} \underset{mx1}{\beta_2} + \underset{nx1}{\varepsilon} .$$

If the estimates of β_2 are constrained to equal b_2 , then

$$Y - X_2 b_2 = X_1 \beta_1 + [\varepsilon + X_2(\beta_2 - b_2)]$$

and the OLS constrained estimates will be

$$\hat{\hat{\beta}}_1 = (X_1'X_1)^{-1}X_1'(Y - X_2b_2)$$
$$\hat{\hat{\beta}}_2 = b_2$$

with mean squared errors displayed as the diagonal elements in the following format

$$\mathrm{MSE}(\hat{\hat{\beta}}_i) = \mathrm{BIAS}^2(\hat{\hat{\beta}}_i) + \mathrm{VAR}(\hat{\hat{\beta}}_i) \ .$$

Thus

$$\mathrm{MSE}(\hat{\hat{\beta}}_1) = G'(b_2 - \beta_2)(b_2 - \beta_2)'G + \sigma_\epsilon^2(X_1'X_1)^{-1}$$
$$\mathrm{MSE}(\hat{\hat{\beta}}_2) = \quad (b_2 - \beta_2)(b_2 - \beta_2)' \quad + \quad 0$$

where $G = X_2'X_1(X_1'X_1)^{-1}$.

In contrast, the unconstrained OLS estimates will have

$$\mathrm{MSE}(\hat{\beta}_1) = 0 + \sigma_\epsilon^2[(X_1'X_1)^{-1} + G'(X_2'\tilde{P}_1X_2)^{-1}G]$$
$$\mathrm{MSE}(\hat{\beta}_2) = 0 + \sigma_\epsilon^2[X_2'\tilde{P}_1X_2]^{-1}$$

where $\tilde{P}_1 = I - X_1(X_1'X_1)^{-1}X_1'$.

Since $\tilde{P}_1$ is idempotent, $\mathrm{VAR}(\hat{\beta}_1) - \mathrm{VAR}(\hat{\hat{\beta}}_1)$ is positive semi-definite and consequently has nonnegative diagonal elements. Thus, the imposition of exact parameter constraints unambiguously reduces (or leaves unchanged) the variance of the estimate of each coefficient. Notice though that this variance reduction is completely independent of the specific values of b_2 that are selected. In particular, deleting a variable from an equation does not reduce the variances of the estimates any more (or less) than does the selection of any other value for the associated coefficient.

The estimated variances will however depend upon b_2 since σ_ϵ^2 is unknown. One could consequently carry the usual two-step procedure to the absurd cosmetic extreme of using the unconstrained estimates of β_2 as the constraint b_2, thereby recording a maximum reduction of $\hat{\sigma}_\epsilon^2$ (through an unwarranted degrees of freedom increase) and thus achieving the maximum decrease in the reported variances of the parameter estimates.

The unreported problem with the two-step procedures is of course the biasing of the estimates, which depends critically but not unambiguously on the selected values of b_2. Looking at the mean squared errors, we have

$$\mathrm{MSE}(\hat{\beta}_1) - \mathrm{MSE}(\hat{\hat{\beta}}_1) = G'[\mathrm{MSE}(\hat{\beta}_2) - \mathrm{MSE}(\hat{\hat{\beta}}_2)]G$$

where

$$\mathrm{MSE}(\hat{\beta}_2) - \mathrm{MSE}(\hat{\hat{\beta}}_2) = \sigma_\epsilon^2[X_2'\tilde{P}_1X_2]^{-1} - (b_2 - \beta_2)(b_2 - \beta_2)' .$$

The MSE's in the constrained case are minimized by exactly correct constraints $(\beta_2 = b_2)$ and in this happy situation the constrained estimates dominate the unconstrained estimates.

In the case of a single restriction,[7] we have the further satisfying results that a decrease in $(b_2 - \beta_2)^2$ reduces (or leaves unchanged) the

[7]
$$\mathrm{MSE}(\hat{\beta}_1) = \sigma_\epsilon^2(X_1'X_1)^{-1} + \mathrm{MSE}(\hat{\beta}_2)G'G$$
$$\mathrm{MSE}(\hat{\hat{\beta}}_1) = \sigma_\epsilon^2(X_1'X_1)^{-1} + \mathrm{MSE}(\hat{\hat{\beta}}_2)G'G$$

where $\mathrm{MSE}(\hat{\beta}_2)$ and $\mathrm{MSE}(\hat{\hat{\beta}}_2)$ are scalars and $G'G$ is positive semi-definite.

MSE's of each of the remaining coefficients, and that the constrained estimates are either unambiguously superior or inferior to the unconstrained estimates, depending solely upon whether or not $(b_2 - \beta_2)^2$ is smaller than the MSE of the unconstrained estimate of β_2 . Notice that again $b_2 = 0$ has no special virtue, and will be thoroughly inferior to a more accurate restriction.

With more than one restriction, we have the possibility that some incorrect restrictions may offset others. Thus,

$$\mathrm{MSE}(\hat{\beta}_1) - \mathrm{MSE}(\hat{\hat{\beta}}_1) = G'[\mathrm{MSE}(\hat{\beta}_2) - \mathrm{MSE}(\hat{\hat{\beta}}_2)]G$$

will be semi-definite if $\mathrm{MSE}(\hat{\beta}_2) - \mathrm{MSE}(\hat{\hat{\beta}}_2)$ is semi-definite. However, the diagonal elements of this latter matrix may all be nonnegative or nonpositive without the matrix being semi-definite. It is consequently possible to have smaller mean squared errors for each of the elements of β_2 and yet have larger mean squared errors for each of the elements of β_1 . In other words, if more than 1 constraint is incorrect, then the effectiveness of the constraints in reducing the mean squared errors of the remaining estimates depends upon the entire risk matrix, $\mathrm{MSE}(\beta_2)$, and not solely upon the diagonal elements. This is true both of comparisons between alternative sets of constraints and of the choice between constrained and unconstrained estimates.

The ambiguity is not likely to be reduced without the specification of a loss function and information about how b_2 is likely to differ from β_2 .[8] This would lead naturally to a Bayesian or quasi-Bayesian approach

[8] Bock, Yancey, and Judge investigate the characteristics of risk functions over the parameter space for unconstrained, constrained, and preliminary test estimators.

(such as mixed estimates). Foregoing this, the decision to impose constraints must be based upon a vague weighing of the inadequacy of the estimates and one's confidence in the proposed constraints. This implies that mechanically imposing ad hoc constraints has little to recommend it. Unless truly a priori inspired, the benefits from routinely deleting variables are necessarily accidental.

APPENDIX

Let $\underset{1\mathrm{x}k+1}{X_i}$ be the i^{th} row of $\underset{n\mathrm{x}k+1}{X}$

$$k+1 = \mathrm{Tr}(I_{k+1}) = \mathrm{Tr}[X'X(X'X)^{-1}] = \mathrm{Tr}[X(X'X)^{-1}X']$$

$$= \mathrm{Tr}\left\{\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}(X'X)^{-1}[X_1' \vdots \cdots \vdots X_n']\right\}$$

$$= \sum_{i=1}^{n} X_i(X'X)^{-1}X_i' \ .$$

Now

$$\mathrm{ASCI} \equiv \frac{1}{n}\sum_{i=1}^{n} [2\tau_{\alpha/2}\hat{\sigma}_{\epsilon}\sqrt{X_i(X'X)^{-1}X_i'}]^2$$

$$= \frac{4\tau^2\hat{\sigma}_{\epsilon}^2}{n}\sum_{i=1}^{n} X_i(X'X)^{-1}X_i'$$

$$= 4\tau^2\left(\frac{k+1}{n-k-1}\right)\mathrm{MSE}$$

$$\begin{aligned} SCI(\bar{X}) &\equiv [2\tau_{\alpha/2}\hat{\sigma}_e\sqrt{\bar{X}(X'X)^{-1}\bar{X}'}]^2 \\ &= 4\tau^2\hat{\sigma}_e^2\bar{X}(X'X)^{-1}\bar{X}' \\ &= 4\tau^2 MSE/(n-k-1) \end{aligned}$$

since

$$I = (X'X)(X'X)^{-1} = (\underset{nx1}{1} \quad \underset{nxk}{X^*})'X(X'X)^{-1} = \begin{bmatrix} 1'x(X'X)^{-1} \\ X^{*\prime}x(X'X)^{-1} \end{bmatrix}$$

implies that

$$\bar{X}(X'X)^{-1}\bar{X}' = \frac{1}{n}[1'x(X'X)^{-1}]\bar{X}' = [\frac{1}{n} \quad 0 \quad 0 \; \ldots \; 0]\begin{bmatrix} 1 \\ \bar{X}_1 \\ \vdots \\ \bar{X}_k \end{bmatrix} = \frac{1}{n}$$

REFERENCES

Barrett, James.: "The Coefficient of Determination--Some Limitations," The American Statistician, 128 (1974), 19-20.

Bartlett, M.: "Some Aspects of the Time Correlation Problem in Regard to Tests of Significance," Journal of the Royal Statistical Society, 98 (1935), 536-43.

Bock, M., T. Yancey, and G. Judge: "The Statistical Consequences of Preliminary Test Estimators," Journal of the American Statistical Association, 68 (1973), 109-116.

Coen, P., E. Gomme, and M. Kendall: "Lagged Relationships in Economic Forecasting," Journal of the Royal Statistical Society, Series A, 132 (1969), 133-163.

Nelson, C.: Applied Time Series Analysis (1973).

Yule, G.: "Why Do We Sometimes Get Nonsense Correlations between Time Series?", Journal of the Royal Statistical Society, 89 (1926), 1-64.

Yule, G., and M. Kendall: An Introduction to the Theory of Statistics, 14th ed. (1950), esp. chs. 13 and 14.