The College Of
WILLIAM & MARY

A Note on the Use of R^2 in Model Selection

Alfredo A. Romero*
College of William and Mary

College of William and Mary
Department of Economics
Working Paper Number 62

October 2007

* Alfredo A. Romero is a Visiting Assistant Professor at the College of William and Mary and a Ph.D. candidate in economics at Virginia Polytechnic Institute and State University

COLLEGE OF WILLIAM AND MARY
DEPARTMENT OF ECONOMICS
WORKING PAPER #62
October 2007

A Note on the Use of $\bar{R}^2$ in Model Selection

Abstract

The use of $\bar{R}^2$ in Model Selection is a common practice in econometrics. The rationale is that the statistic produces a consistent estimator of the true coefficient of determination for the underlying data while taking into consideration the number of variables involved in the model. This pursuit of parsimony comes with a cost: The researcher has no control over the error probabilities of the statistic. Alternative measures of goodness of fit, such as the Schwarz Information Criterion, provide only a marginal improvement to the problem. The F-Test under the Neyman-Pearson testing framework will provide the best alternative for model selection criteria.

JEL Codes: C12, C52

Keywords: Adjusted R squared, Schwarz Information Criterion BIC, Neyman-Pearson Testing, Nonsense Correlations

Alfredo A. Romero
College of William and Mary
Williamsburg, VA 23187-8795
aromero@wm.edu
aromeroa@vt.edu

1 Introduction

The use of $\bar{R}^2$ in model selection is a common practice in econometrics. Nearly every econometric software would provide the researcher with this measure of goodness of fit along with some other measures, such as the unadjusted R squared, the Akaike Information Criterion, and the Schwarz Information Criterion. These statistics are then used in model selection following a pre-specified rule. For the $\bar{R}^2$, the model with the highest statistic would represent the best-fit model. In this document, we will show that such a mechanism would lead the researcher to unwanted conclusions. In section two, we will show that the $\bar{R}^2$ leaves the researcher with no control over the error probabilities of the model selection. In section three, we extend these results to the Schwarz Information Criterion. In section four, we will put both statistics to the test. Finally, section five presents the conclusions.

2 The use of $\bar{R}^2$ in model selection

In econometric modeling, the addition of explanatory variables to a normal linear regression model can never decrease the value of the unadjusted R squared, R^2, even if the additional variables have no explanatory power. Because of this and other reasons, the R^2 is not the preferred model selection criterion. For instance, as Montgomery and Morrison (1973) have shown, the unadjusted R^2 is a (positively) biased estimator of the true coefficient of determination for the underlying population. Additionally, the coefficient does not penalize the likelihood function for having additional variables.

A somehow less biased estimator of the coefficient of determination is the adjusted coefficient of determination, Adjusted R Squared, $\bar{R}^2$. Barton (1962) has shown that, although positively biased, the estimator is consistent since it converges to the true coefficient of determination when the sample size increases[1].

Thus, although a consistent estimator of the true coefficient of determination, there is an inherent danger in using $\bar{R}^2$ in model selection. It is widely believed that, since the statistic takes explicit account of the number of regressor used in the equation, it is useful for comparing the fit of specifications that differ in the addition or deletion of explanatory variables (Johnston and Dinardo, 1997). In practice, a model with a larger number of explanatory variables and thus a higher unadjusted R^2 would be preferred if and only if the $\bar{R}^2$ is higher too. We will attempt to show that such a rational suffers from problems in statistical grounds by using the criterion to select between two competing models.

For instance, consider the following polynomial regression model,

$$y_t = \beta_0 + \Sigma_{k=1}^{K-1} \beta_k x_t^k + u_t,$$

where $u_t \sim N(0, \sigma^2)$, $t \in T$,

to be fitted to a data set $\{(x_t, y_t), t = 1, 2, ..., T\}$.

[1]For $T > 10$, the biased is 0.01 whereas for $T > 100$, the biased is 0.001, hence rendering a relatively small bias even for small sample sizes.

The $\bar{R}^2$ statistic, necessary for model selection, is defined as,

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k},$$

where $R^2 = \left(1 - \frac{RSS}{TSS}\right) = \left(1 - \frac{\hat{\sigma}^2 \cdot T}{TSS}\right)$, RSS is the Residual Sum of Squares and TSS is the Total Sum of Squares. A little algebra shows that,

$$\bar{R}^2 = 1 - \frac{T-1}{T-k}\frac{\hat{\sigma}^2 \cdot T}{TSS}.$$

Thus, it is possible to relate the $\bar{R}^2$ to the Maximum Likelihood Estimator of σ^2. Recall that the log-likelihood function takes the form,

$$\ln L(\boldsymbol{\theta}) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\Sigma_{t=1}^{T} u_t^2,$$

which takes its minimum value at

$$\ln L(\hat{\boldsymbol{\theta}}) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\Sigma_{t=1}^{T} \hat{u}_t^2,$$

where $\hat{\sigma}^2 = \frac{1}{T}\Sigma_{t=1}^{T}\hat{u}_t^2$, $\hat{u}_t = (y_t - b_0 - \Sigma_{k=1}^{K-1} b_k x_t^k)$, and $\boldsymbol{b}$ denote the Maximum Likelihood Estimators of $\boldsymbol{\beta}$.

Now suppose the existence of two competing models,

M_2: $y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + u_t$

M_1: $y_t = \beta_0 + \beta_1 x_t + u_t$

and suppose that M_1 was chosen over M_2 based on the $\bar{R}^2$ criterion. That is, the addition of the explanatory variables does not contribute to the explanatory power of the model. This implies that,

$$1 - \frac{T-1}{T-k_1}\frac{\hat{\sigma}_1^2 \cdot T}{TSS} > 1 - \frac{T-1}{T-k_2}\frac{\hat{\sigma}_2^2 \cdot T}{TSS},$$

that is

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} < \frac{T-k_1}{T-k_2}.$$

Equation 1

Note that it is possible to relate the previous condition to a Neyman-Pearson testing framework via an F-test. In an F-test, we can relate the decision of accepting M_1 instead of M_2 with that of a test for,

$$H_0\text{: } \beta_2 = \beta_3 = 0, \text{ vs. } H_1\text{: } \beta_2 \neq 0, \text{ or } \beta_3 \neq 0.$$

The F-test for these hypotheses takes the form:

$$F(y) = \left\{ \left(\frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\hat{\sigma}_2^2} \right) \left(\frac{T-k_2}{k_2-k_1} \right), \;\; C_1 = F(y) > c_\alpha \right\},$$

where C_1 denotes the rejection region and c_α the critical value associated with $F(k_2 - k_1, T - k_2)$. This suggests that the $\bar{R}^2$ of Equation 1 amounts to accepting H_0 when,

$$F(y) = \left(\frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\hat{\sigma}_2^2} \right) \left(\frac{T-k_2}{k_2-k_1} \right) < 1.$$

This is an extremely dangerous result! The criterion leads the researcher with absolutely no control over the significance level of the test. For instance, for $(k_2 - k_1) = 2$ and $T = 100$, the implicit significance level is around $\alpha = 0.37$. That is, the probability to making a Type-I error is 37 percent, way far from the standard levels of significance of 0.05, 0.01 and 0.001, commonly used in the reporting of econometric research.

We can extend these results to compute different probabilities. For instance, we can compute the probability of a model being overfitted when using this criterion. The fact that a model is overfitted implies that $\bar{R}^2_{K+L} > \bar{R}^2_K$. The probability that the $\bar{R}^2$ criterion selects the overfitted model is given by,

$$
\begin{aligned}
& P\left(1 - \frac{T-1}{T-K-L}\frac{\hat{\sigma}^2_{K+L}\cdot T}{TSS} > 1 - \frac{T-1}{T-K}\frac{\hat{\sigma}^2_{K}\cdot T}{TSS}\right) \\
& = P\left(-\frac{T-1}{T-K-L}\frac{\hat{\sigma}^2_{K+L}\cdot T}{TSS} > -\frac{T-1}{T-K}\frac{\hat{\sigma}^2_{K}\cdot T}{TSS}\right) \\
& = P\left(\frac{T-1}{T-K-L}\frac{\hat{\sigma}^2_{K+L}\cdot T}{TSS} < \frac{T-1}{T-K}\frac{\hat{\sigma}^2_{K}\cdot T}{TSS}\right) \\
& = P\left(\frac{\hat{\sigma}^2_{K+L}}{T-K-L} < \frac{\hat{\sigma}^2_{K}}{T-K}\right) \\
& = P\left(\frac{\hat{\sigma}_{K}}{\hat{\sigma}^2_{K+L}} > \frac{T-K}{T-K-L}\right) \\
& = P\left(\left(\frac{\hat{\sigma}^2_K - \hat{\sigma}^2_{K+L}}{\hat{\sigma}^2_{K+L}}\right)\left(\frac{T-K-L}{L}\right) > 1\right) \\
& = P(F(y) > 1) \\
& = 1 - CDF\,(1,\ \ L,\ \ (T-K-L)).
\end{aligned}
$$

where $CDF()$ is the cumulative distribution function of the F distribution.

Asymptotically, the situation is not alleviated. Given that asymptotically the F test converges in distribution to a chi-squared distribution, that is,

$$F(y) \Rightarrow^{D} \frac{1}{L}\chi^2(L),$$

then

$$\lim_{T\to\infty} P(\bar{R}^2_{K+L} > \bar{R}^2_K) = P(\chi^2(L) > L) > 0.$$

For instance, in our previous example, the probability that the criterion asymptotically overfits the model by one is 31.73 percent.

3 A (conditionally) better alternative

Akaike (1974) started the pursuit of a model selection criterion based on different penalizations of the likelihood function with his Akaike Information Criterion (AIC_K). Although still highly used in econometric reporting, early testing on the AIC showed that the statistic suffered from two main drawbacks,

1) In small samples the criterion led to overfitting.

2) Asymptotically the chosen K was not a consistent estimator of the true $K*$.

To deal with the small sample overfitting, several modified criterions have been suggested. Amongst them, the Modified-AIC_K, proposed by Hurvich and Tsai (1989) and the Final

Prediction Error, proposed by Akaike. Spanos (2007) has shown that, in small samples, only the M-AIC performs better but maintains a positive probability of overfitting even asymptotically.

To deal with the problem of inconsistency the preferred alternatives have been the Schwarz Information Criterion (BIC)and the Hannan-Quinn Information Criterion[2]. These two criteria perform better relative to AIC in small samples, with lower probabilities of overfitting, and are asymptotically consistent, that is, the probability of asymptotically overfitting the true model using these criteria is zero.

Following the previous discussion, we will derive the same probabilities described in the $\bar{R}^2$ criterion. The BIC is defined as,

$$BIC_K = \ln \hat{\sigma}^2 + \frac{K \ln(T)}{T}$$

Thus, selecting M_1 over M_2 implies that

$$\ln \hat{\sigma}_1^2 + \frac{K_1 \ln(T)}{T} < \ln \hat{\sigma}_2^2 + \frac{K_2 \ln(T)}{T},$$

or,

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} < T^{\frac{k_2 - k_1}{T}}.$$

[2]Of these, BIC has become the standard in econometric reporting.

The previous equation implies the following F-statistic,

$$F(y) = \left(\frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\hat{\sigma}_2^2}\right)\left(\frac{T-k_2}{k_2-k_1}\right) < \left(\frac{T-k_2}{k_2-k_1}\right)\left[T^{\frac{k_2-k_1}{T}} - 1\right].$$

Thus, for instance, for $(k_2 - k_1) = 2$ and $T = 10$, the implicit level of significance given by the BIC is 0.10; for $T = 25$, $\alpha = 0.03$; and for $T = 100$, $\alpha = 0.01$; clearly decreasing with the sample size.

Similarly, the probability of overfitting a model by L variables using this criterion is

$$P(F(y) > \frac{T-K-L}{L}[T^{\frac{L}{T}} - 1]),$$

which has been shown to be equal to zero asymptotically.

Does this mean that BIC will lead the researcher to the true model when the sample size is large? The answer is a big 'IT DEPENDS'. If the true model lies within the pre-specified family of model the answer is yes. However, if the true model does not lie within the pre-specified family of models the answer is no (see next section). Even worse, in the latter case, the researcher would have no way of assessing how off the proposed model is from the true model.

4 Empirical Exercise

To illustrate the danger of utilizing $\bar{R}^2$ and BIC criteria in model selection, we created two series, y_t and x_t, that seem to have a 'strong' relation.

Our base model is $y_t = \beta_0 + \beta_1 x_t + u_t$. Ordinary Least Squares estimation produces the following results,

$Variable$	$Coefficient$	$Std.Error$	$t-Statistic$	$Prob.$
b_0	20.5874	1.8512	11.1207	0.0000
b_1	0.8186	0.0228	35.8330	0.0000
$\boldsymbol{\bar{R}^2}$	**0.8657**	$\boldsymbol{BIC_k}$	**7.5382**	

Then, we proceeded to increase the number of variables by sequentially adding lags in both, x_t and y_t, until the selected criterion indicated the 'best' model. Using the $\bar{R}^2$ criterion, the selected model is $y_t = d + a_1 y_{t-1} + \ldots + a_5 y_{t-5} + b_0 x_t + b_1 x_{t-1} + \ldots + b_6 x_{t-6} + u_t$, with an $\bar{R}^2 = 0.9642$; an improvement with respect to the base model of 11 percent in the fit. The BIC tells a different story. The best-fit model is $y_t = d + a_1 y_{t-1} + b_0 x_t b_1 x_{t-1} + b_2 x_{t-2} + u_t$, producing a BIC of 6.6262, the lowest possible amongst the class.

Clearly, both criteria would produce an answer with respect to what model would produce the best fit while taking into consideration the number of variables. The endless look for parsimony. The problem is that neither approach looks into the statistical adequacy of the models. This creates the problem of sacrificing statistical significance for explanatory power. In fact, the true model is neither of the previous two but,

$$y_t = d + a \cdot y_{t-2} + b \cdot t + e_t$$

where no statistical relationship between x_t and y_t exists (see Appendix).

5 Conclusion

Although both, the $\bar{R}^2$ and the BIC would lead the researcher to deciding what model represents the best fit to the data, neither of them warrants the existence of a statistically meaningful model. By construction, they are unable to provide control over the level of significance in their model selection procedure, unlike a Neyman-Pearson testing framework. Goodness of fit should be considered a highly unreliable measure in model selection and used only in conjunction with relevant F-tests.

6 Appendix

Failure to correctly capture the effect of trends in the expected value of the marginal distributions of x_t and y_t is the source of nonsense correlations between them. As we know, the maximum likelihood estimator for b_1 in a bivariate linear regression model is $b_1 = \frac{Cov(x,y)}{Var(x)}$. By definition, $Cov(y,x) = \frac{1}{n}\Sigma x_i y_i - E(x)E(y)$. If $E(x) = a + bT_i$ and $E(y) = c + dT_i$, then, $Cov(y,x) = \frac{1}{n}\Sigma x_i y_i - (a + bT_i)(c + dT_i)$. Notice that $Cov(x,y) = 0$ if $b = \frac{\Sigma x_i y_i}{ncT_i + dT_i^2} - \frac{a}{T_i}$. This is precisely how we created the series (y_t, x_t). Ignoring the heterogeneity in both means, that is, letting $b = d = 0$, produces $b_1 = \frac{\Sigma(y_i - c)(x_i - a)}{\Sigma(x_i - a)^2} \neq 0$. Thus, creating the impression of a statistical relation between x_t and y_t.

7 References

1. Akaike, Hirotugu (1974). "A new look at the statistical model identification." IEEE Transactions on Automatic Control 19 (6): 716723.

2. Barten, A.P. (1962). "Note on unbiased estimation of the squares multiple correlation coefficient." Statistica Neerlandica 16 (2): 151-63.

3. Hurvich, Clifford M. and Chih-Ling Ttsai (1989). "Regression and time series model selection in small samples." Biometrika 1989 76(2):297-307

4. Montgomery, David and Donald Morrison (1973). "A note on adjusting R^2." Journal of Finance 28: 1009-13.

5. Johnston, Jack and John Dinardo (1997). "Econometric Methods," McGraw-Hill/Irwin.

6. Schwarz, Gideon, 1978. "Estimating the dimension of a model." Annals of Statistics 6(2):461-464.

7. Spanos, Aris (2007). "Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach," Working Paper. VPI.