EXPLAINING CONTRADICTIONS FROM ECONOMETRIC MODELS

Frank G. Landram, West Texas A & M University
Amjad Abdullat, West Texas A & M University

ABSTRACT

Contradictions may exist when applying significance tests to multiple, partial, and simple determination coefficients. These contradictions are usually the result of multicollinearity among the independent variables and the manner in which it affects these tests. By understanding the manner in which these contradictions occur and under what conditions multicollinearity is desirable, users are able to give a more informed interpretation of significance tests. A clear understanding will ensure the proper treatment and will enhance the effectiveness of econometric applications.

INTRODUCTION

It is common knowledge that testing the coefficient of determination (R^2) is equivalent to the analysis of variance test in multiple regression (Ramanathan, 1998). Similarly, testing partial determination coefficient $r^2_{y(j)}$ is equivalent to testing partial regression coefficients (b_j.) (Gustafson, 1961). However, it is not common knowledge among users that contradictions exist when applying various significance tests. Using obvious notation, this article explains these contradictions thereby giving analysts additional insight into the subject matter. An understanding of these contradictions will also enhance analytical and research skills in econometric modeling and in the proper interpretation of the model's output.

RESULTS OF SIGNIFICANCE TESTS

When testing multiple R^2, simple r^2, and the equation's partial regression coefficients, the following results may occur:

1. R^2 and some, if not all, b_j are significant
2. Neither R^2 nor any b_j are significant
3. R^2 but no b_j are significant
4. R^2 is not significant; some, if not all, b_j are significant
5. Partial $r^2_{y(j)}$ is significant; its simple $r^2_{y.j}$ is insignificant

Result 1 is the expected result. Researchers and analysts do not engage in model building using unrelated variables.

Result 2 should never happen. Given a knowledge of the subject matter, it is highly unlikely to collect data with no variable or combination of variables being significantly related to the dependent variable Y.

Result 3 is caused by multicollinearity among the X_j variables. If one or more of these variables are deleted, many of the remaining variables may become highly significant. An explanation is that no variable is needed (significant) as long as other duplicate variables are in the model. When some of these duplicate variables are deleted, the remaining variables become significant. This is found in the "pork" study published by SAS (1991) — R^2 is significant and all partial $r^2_{y(j)}$ coefficients are insignificant. When multicollinear (duplicate) variables are deleted, the other variables become significant. This usually happens when the independent variables possess significant simple $r^2_{y.j}$ coefficients but their relationships with Y are highly duplicated by other independent variables. Thus, multicollinearity can render the independent variables' b_j or partial $r^2_{y(j)}$ coefficients insignificant even though their simple $r^2_{y.j}$ coefficients are are significant.

Result 4 is extremely rare. It can occur when the X variables are barely significant but negatively correlated with other variables. This can make the combination of X variables which are reflected in R^2 insignificant. It can also occur when synergism among the independent variables causes the partial $r^2_{y(j)}$ coefficients to become significant faster than R^2. A solution is to obtain a further knowledge of the subject matter and investigate the inclusion of additional variables in the regression model.

Result 5. A Significant partial $r^2_{y(j)}$ and an insignificant simple $r^2_{y.j}$ is the definition of a synergistic variable (Landram and Abdullat, 2002). Variable X_j is only significantly related with Y when used in a combination with other variables. Used alone, X_j is insignificant. Also, synergistic variables produce a multiple R^2 that is greater than the sum of the simple determination coefficients;

$$R^2_{y.jk} > r^2_{y.j} + r^2_{y.k}$$

With respect to the frequency of synergism in statistical models, the literature has evolved from Freund's (1988) opinion that "it rarely happens" to Shieh's (2001) research that "it is more common than one thinks". Actually, synergism is fairly common. The probability is surprisingly high that synergism may be found in an economic study with empirical data. By testing both partial and simple determination coefficients, synergistic variables are identified. This enables analysts and researchers to obtain additional insight into the subject matter. It may also generate more questions than answers. This insight is illustrated in the following example.

APPLICATIONS

The following data are taken from *Introductory Econometrics: With Applications* (Ramanathan,1998). Death rates in the US due to coronary heart disease for 1947-1986 are analyzed using multiple regression.

Y_i is death rate per 100,000

X_1 is daily per capital consumption of calcium.

X_2 is percentage of labor force (over 15 years) per 1000 persons.

X_3 is per capita of cigarettes in pounds of tobacco.

X_4 is per capita intake of edible fat and oil in pounds.

X_5 is per capita intake of meat in pounds (beef, veal, pork, lamb, mutton)

X_6 is per capita consumption of distilled spirits in taxed gallons.
X_7 is per capita consumption of malted liquor in taxed gallons.
X_8 is per capita consumption of wine in taxed gallons.

Table 1
Comparing Variable Section Criteria

Model	R_j^2	$\bar{R}_j^2$	C_p	$\overline{PC}_j$	P_j^2	Variables in Model
8	0.736	0.642	9.00	0.545	0.471	X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8
7	0.735	0.664	7.03	0.572	0.506	X_1 X_2 X_3 X_4 X_5 X_6 X_7
6	0.733	0.674	5.22	0.595	0.239	X_2 X_3 X_4 X_5 X_6 X_7
5	0.730	0.682	3.51	0.615	0.567	X_3 X_4 X_5 X_6 X_7
4	0.712	0.672	3.21	0.613	0.554	X_3 X_4 X_6 X_7
3	0.658	0.624	5.82	0.567	0.521	X_4 X_6 X_7
2	0.585	0.558	11.26	0.504	0.480	X_6 X_7

Table 1 illustrates the results from employing the backward selection algorithm. Observe, although X_8 possesses a significant simple r^2 value, it is deleted early in the selection process because of its multicollinearity with X_3, X_4, X_6, and X_7. The models with four and five variables possess the lowest Mallows C_p statistic. The other variable selection statistics--the adjusted determination coefficient $\bar{R}_j^2$, Amemiya's $\overline{PC}_j$, and relative PRESS P_j^2 statistics --are also favorable for these models (Maddala 1981). Mallows C_p (1973), Amemiya's $\overline{PC}_j$ (1980), and P_j^2 (Myres, 1990) statistics measure goodness of prediction while R^2 and adjusted $\bar{R}_j^2$ measure goodness of fit.

Table 2
Computer Output For The Reduced Model

VARIABLE	X_0	X_3	X_4	X_6	X_7
REG COEF	139.68	10.706	3.380	26.75	-4.132
S(b)	**	4.590	0.967	7.037	0.863
$F=[b/S(b)]^2$	**	5.439	12.227	14.450	22.925
PROB>F	**	0.025	0.002	0.001	0.000

Correlation Coefficient Matrix

	Y	X_3	X_4	X_6	X_7
Y	1.0000	0.4548	-0.2047	-0.2283	-0.4502
X_3	*********	1.0000	-0.9141	-0.8840	-0.9364
X_4	*********	*********	1.0000	0.9192	0.9126
X_6	*******	*********	************	1.0000	0.9472
X_7	*********	*********	*********	*********	1.0000

The least significant value for r is LSV r = $|0.2887|$, $\alpha = 0.10$

Table 2 reveals that X_4 and X_6 are synergistic variables. They possess significant partial F values and insignificant simple r^2 coefficients; also

$$R^2_{y.3467} > r^2_{y.3} + r^2_{y.4} + r^2_{y.6} + r^2_{y.7}.$$

Variables X_4 and X_6 must be combined with X_3 and/or X_7 before they are significant related with Y. Does this mean they are only harmful when associated with smoking X_3 and malted liquor X_7? Or, are other variables, such as age and obesity, needed in the model? Furthermore, observe that the simple r^2 coefficients for X_4 and X_6 are negative while their regression coefficients are positive. The death rate Y increases when they are combined with X_3 and X_7. Significance tests on simple correlation coefficients as well as on partial F values provide additional insight into the subject matter. It may also generate more questions than answers. However, these significant tests may direct the analyst toward further investigations and the final solution.

CONCLUDING REMARKS

In testing R^2, researchers initially determine whether or not a relationship exists between Y_i and at least one X variable or a combination of X variables. Therefore, result 1 is the expected result. Preliminary procedures such as identifying influential variables and properly collecting the data somewhat guarantee that result 2 will not happen. Result 3 is usually caused by multicollinearity. Deleting irrelevant variables, increasing the sample size, and combining logically compatible variables are some of the procedures used in treating multicollinearity (Feldstein, 1973). A knowledge of the subject matter should assist users in deleting variables from the statistical equation. Many times the inclusion of additional variables will change the value of an insignificant b_j. Thus, analysts and researchers should make certain that all influential variables are included in the model.

Result 5 identifies synergistic variables thereby adding another dimension to the analysis. An important part of econometric modeling is identifying and explaining the logic underlying a synergistic combination of X variables. Variables included in synergistic combinations should be identified and retained in the model if (a) their partial F values are significant or (b) their presence enables other variables to possess significant partial F values. Truly, identifying synergistic combinations directs one to areas where a further knowledge of the subject matter is needed. Indeed, understanding significance tests and their contradictions is imperative for obtaining a properly specified econometric model and in correctly interpreting its output. Hence, this understanding greatly enhances the analytical and research skills of analysts and researchers.

REFERENCES

Amemiya, T. (1980), "Selection of Regressors," *International Economic Review*, 21, 331-354.

Feldstein, M. (1973), Multicollinearity and the Mean Square Error of Alternative Estimators, *Econometrica 41, 337-346*.

Freund, R.J. (1988), "When is $R^2 > r_{y.1}^2 + r^2_{y.2}$? (Revisited)," *The American Statistician*, 42, 89-90.

Gustafson, R.(1961), "Partial Correlations in Regression Computations, "*Journal of the American Statistical Association*, 56, 363-366.

Landram, F.G., and Abdullat, A. (2002), "Lurking Variables In Econometric Models", *Southwestern Economic Review*, 29, 1.

Maddala, G. S. (1981), "Model Selection," *Journal of Econometrics, Vol. 16.*

Mallows, C.L. (1973), "Some Comments on C_p," *Technometric, 15.*

Myres, R.H. (1990), *Classical and Modern Regression With Application,*2nd edition, Boston, Massachusetts: Duxbury.

Ramanathan, R.(1998), *Introductory Econometrics: With Applications*, 4th edition, Fort Worth: The Dryden Press.

SAS Institute (1991), Freund, R.J. and R.C. Littell (1991), *SAS System for Regression*, 2nd ed., Cary, NC.

Shieh, Gwowne,(2001), "The Inequality Between the Coefficient of Determination and the Sum of Squared Simple Correlation Coefficients", *The American Statistician*, 42, 121-124.

ACKNOWLEDGMENTS

This research was supported in part by a Killgore Research Grant at West Texas A & M University.