LURKING VARIABLES IN ECONOMETRIC MODELS

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

ABSTRACT

Lurking variables are omitted variables that should be included in the regression model. If the lurking variable is part of a synergistic combination, the effects it has on a regression model are magnified. This paper illustrates the seriousness of omitting a variable that is part of a synergistic combination. When this happens, synergistic variables in the regression model act as if they are unrelated with the dependent variable. This drastically reduces the model's effectiveness and can lead to misleading results. An awareness of synergistic variables and their possible effect on underspecified regression models enable analysts to become more proficient in econometric modeling.

INTRODUCTION

Little research is available concerning the possible effects of excluding relevant variables which are part of a synergistic combination [4]. Although these effects are known in many circles by oral tradition and intuition, they have not been documented. Moreover, the results of excluding these types of variables are usually misleading.

Using obvious notations, this paper illustrates the seriousness of lurking variables in regression [10]. If the lurking variable is synergistic, the seriousness is magnified. Synergism in regression enables multiple R^2 to become greater than the sum of the simple r^2 coefficients; $R^2 > \Gamma r_{.j}$. These regressors which are seemingly unrelated with Y become significant when combined with other variable(s). Using empirical data, an example is given which illustrates the misleading results when synergistic variables are omitted from the regression model. Concluding remarks are given which will enhance ones ability in model building.

Benefits

This paper benefits readers by showing **(a)** the possible misleading results of lurking variables and underspecified models. This encourages analysts to become more diligent in their preliminary research procedures. **(b)** Readers are given a strong awareness that they should not always accept statistical tests as being definitive. If intuition and a knowledge of the subject matter indicate that an insignificant regressor is in fact related with Y, search for additional influential variables. The regressor in question may be a synergistic variable needing another regressor to complete the synergistic combination. **(c)** Variable selection algorithms can not replace a knowledge of the subject matter and must be used with caution. When the regression model is underspecified, variable selection algorithms usually magnify the problem. **(d)** A knowledge of synergism also enhances ones skills in regression analysis. Multicollinearity is desirable for synergistic variables but becomes a problem for other variables [3]. This concept is explained below.

SYNERGISTIC VARIABLES

A synergistic variable is defined as a variable whose partial r^2 value is greater than its simple r^2 value [4]. A variable is also considered synergistic if it possesses a significant partial F value and an insignificant simple r^2 value. Synergistic variables must be used in a combination with other variables. Several articles have illustrated synergism in regression [4][9]. Kendall and Stuart [5], show that given $r_{y.j}r_{y.k} > 0$, if variables X_j and X_k are inversely related ($r_{j.k} < 0$) or if

$$r_{j.k} > 2r_{y.j}r_{y.k} /(r_{.j} + r_{.k}), \qquad (1)$$

then the variables are synergistic; $R^2 > r_{.j} + r_{.k}$. They also give the conditions for identifying synergism when $r_{y.j}r_{y.k} < 0$.

Daniel and Wood [2] and Freund [3] show that synergism in regression is a function of multicollinearity. As multicollinearity increases among the regressors, the importance of the regressors may also increase; this increases the partial F values and decreases the residual mean square. This concept is illustrated in Figure 1 for the regression model

$$\hat{Y}= b_0 + b_1X_1 + b_2X_2. \qquad (2)$$

Multiple R^2 is measured on the vertical axis and the correlation or multicollinearity between X_1 and X_2 ($r_{1.2}$) is measured on the horizontal axis. Given specific values for $r_{y.1}$ and $r_{y.2}$ and starting at the extreme negative point for $r_{1.2}$, multiple R^2 decreases as $r_{1.2}$ increases between X_1 and X_2. This continues throughout the permissible range for $r_{1.2}$ until R^2 reaches its minimum after which it increases. Hence, multicollinearity is desirable if X_1 and X_2 are inversely related ($r_{1.2} < 0$) or if $r_{1.2}$ is greater than (1) given $r_{y.1}r_{y.1} > 0$. These values are where X_1 and X_2 become synergistic.

FINANCIAL ANALYSIS EXAMPLE

Stock prices (Y) are a function of annual return on investment and anticipated growth. The data in Table 1 was obtained from a sample of 35 companies in Dun's Review. The types of variables employed in describing stock prices are listed below.

X_1: yield = (Dividend + Price Change)/Current Price X_2: Dividends
X_3: Earnings per share X_4: Sales X_5: Income
X_6: Return on sales X_7: Return on equity (ROE) X_8: Exchange
traded

Figure 1
Synergism is a Function of Multicollinearity
Given $r_{y.1} = -0.442$ and $r_{y.2} = 0.191$
R^2

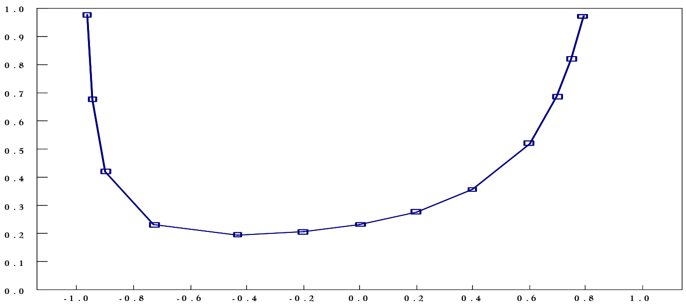

Table 1
Data for Selected Stocks

Company	Y	X1	X2	X3	X4	X5	X6	X7	X8
Cross(A.T.)	40.00	. 170	3.9	3.61	95.1	14.6	15.4	29.9	0
McDonough	41.25	.150	3.4	5.39	450.6	21.6	4.8	14.9	1
Brunswick	14.50	. 100	6.0	.39	1257.3	51.4	4.1	10.4	1
Leaseway Trn.	33.50	. 170	3.9	3.62	937.2	42.9	4.6	20.6	1
Com. Clr.House	31.38	. 130	3.2	2.11	215.8	19.6	9.1	65.7	0
Mallinckrodt	45.00	. 175	2.7	3.40	392.5	31.9	8.1	13.9	0
Empl. Casualty	39.50	. 210	3.0	5.55	178.8	17.2	9.6	23.6	0
Lib. Natl.Life	17.68	. 225	6.8	2.67	409.3	50.8	12.4	20.1	0
Ohio Casualty	36.75	. 330	4.8	7.11	809.2	82.5	10.2	23.9	0
Wstn Cas.	38.88	. 350	4.7	7.70	401.9	48.8	12.0	25.8	1
Crane	38.50	. 320	4.1	5.39	1573.2	55.0	3.5	14.7	1
Am. Bnkrs.Life	11.00	. 090	4.0	2.22	131.0	10.2	7.8	31.2	0
Toronto-Dom.Bk	30.50	. 340	4.5	.80 2	739.0	106.4	3.9	14.3	0
Kennametal	2.88	.190	2.3	3.06	325.9	36.7	11.3	19.7	1
Huyck Corp.	22.25	. 210	3.2	1.58	143.0	9.0	6.3	14.0	1
Std. Brands Pt	29.75	. 180	2.4	2.70	182.7	14.3	7.8	16.2	1
Nevada Power	19.78	. 610	11.7	3.37	175.1	17.1	9.8	13.1	1
Heinz (H.J.)	44.75	. 630	5.0	6.24	2924.8	142.9	4.9	16.4	1
Nashua Corp.	28.13	. 450	5.3	5.75	608.4	26.7	4.4	19.0	1
HB Fuller	12.38	. 120	3.2	1.75	258.7	7.9	3.1	14.0	0
Diebold Inc.	44.75	. 260	1.7	3.19	305.2	18.2	6.0	15.3	1
Kellogg	19.18	. 450	6.9	2.13	1846.6	162.6	8.8	24.6	1
Caterpillar	56.00	. 800	4.2	5.69	7613.2	491.	6.5	16.0	1
Ryl. Bank Can.	52.75	. 860	4.7	7.40	4215.5	270.7	6.4	20.9	0
Banco de Ponce	16.50	. 451	7.3	4.96	107.7	7.7	7.1	3.3	0
Fla. P&L	26.50	1.020	10.0	4.22	1933.9	204.7	10.6	11.1	1
Moore Products	29.00	. 310	2.7	2.84	48.3	5.7	1.8	19.8	0
Meyer (Fred)	38.50	. 220	1.4	3.29	1060.2	22.4	2.1	14.3	0
Eagle-Picher	18.88	. 380	4.7	3.03	590.0	30.7	5.2	15.3	1
Ga.-Pacific	26.38	. 520	4.6	3.12	5207.0	327.0	6.3	18.2	1
Ctl. Tel.& Ut.	23.25	. 880	8.6	3.34	750.5	83.1	11.1	15.1	1
Gnl. Shale	13.50	. 470	7.3	2.21	61.8	5.5	8.9	13.5	0
MT-Dak Util.	22.88	. 880	7.9	2.76	173.3	17.7	10.2	10.7	1
So. Union	45.25	980	4.2	4.78	724.0	34.1	4.7	9.6	1

Source: Dun's Review, Dun & Bradstreet Publications Corporation.

ANALYSIS

Table 2 gives the computer output for two scenarios. Assume the analyst excludes variable X and computes the initial least squares regression model with the following variables:

$$Y = f(X \ X_3 \ X_4 \ X_5 \ X_6 \ X_7 \ X_8) \qquad R^2 = 0.53 \tag{3}$$

Using stepwise regression with significant level at 0.10, the above model reduces to

$$Y = f(X_3 \ X_4) \qquad R^2 = 0.46 \tag{4}$$

By excluding X, both of the above models are underspecified.

Table 2
Partial F Measures In Regression Models

(a) $Y = f(X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_6 \ X_7 \ X)$ R = 0.787

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X
partial F	6.15	30.86	25.50	2.45	1.67	4.67	0.24	1.99
p-value	0.020	0.001	0.001	0.130	0.208	0.040	0.626	0.170

(b) $Y = f(X_1 \ X_2 \ X_3 \ X_4 \ X)$ R = 0.762

Variable	X_1	X_2	X_3	X_4	X_6
partial F	7.05	34.86	24.01	2.47	3.12
p-value	0.013	0.001	0.001	0.127	0.088

(c) $Y = f(X_1 \ X_2 \ X_3)$ R = 0.725

Variable	X_1	X_2	X_3
partial F	12.39	36.5	23.46
p-value	0.0014	0.0001	0.0001

(d) $Y = f(X_2 \ X_3 \ X_4 \ X_5 \ X_6 \ X_7 \ X)$ R = 0.535

Variable	X_2	X_3	X_4	X_5	X_6	X_7	X_8
partial F	0.26	17.49	2.88	1.81	1.20	1.33	1.24
p-value	0.617	0.001	0.101	0.190	0.283	0.259	0.275

When X is included, the stepwise regression algorithm reduces the initial model

$$Y = f(X\ X\ X\ X\ X\ X\ X\ X) \qquad R = 0.79 \qquad (5)$$

to

$$Y = f(X_1\ X_2\ X_3) \qquad R^2 = 0.73 \qquad (6)$$

The adjusted R^2 value for (5) and (6) are 0.722 and 0.70. Table 3 reveals that variables (X X X X) possess significant simple r-value. However, a variable is deleted from a regression model because it is either unrelated with Y or it is multicollinear with other X-variables (partial duplication of data) and not needed. The latter reason is why variables X and X are deleted from (5). Although variables X and X possess significant simple r-values, their multicollinearity causes them to become insignificant and deleted from the model.

Table 3
Correlation Coefficient Matrix (Bottom Diagonal)

Variable	Y	X1	X2	X3	X4	X5	X6	X7
Y	1.000	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
X1 -	0.442	1.000	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
X2	0.191	0.597	1.000	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
X3	0.633	0.033	0.348	1.000	xxxxx	xxxxx	xxxxx	xxxxx
X4	0.425	0.033	0.443	0.320	1.000	xxxxx	xxxxx	xxxxx
X5	0.383	0.132	0.512	0.309	0.967	1.000	xxxxx	xxxxx
X6	-0.096	0.256	0.048	-0.107	-0.270	-0.089	1.000	xxxxx
X7	0.073	-0.260	-0.289	-0.105	-0.118	-0.082	0.289	1.000
X8	0.034	0.184	0.311	-0.083	0.204	0.221	-0.192	-0.310

The Least Significant Absolute Value for simple r equals 0.254
(LSV-r = . 0.254 .). All r-values below . 0.254 . are insignificant.

By employing the stepwise regression algorithm, (6) is obtained from (5). By employing the all possible regressions algorithm on (5), the following model is obtained;

$$Y = f(X\ X\ X\ X\ X) \qquad R = 0.76 \qquad (7)$$

The adjusted R^2 value for (7) is 0.721. The partial F-value for X is marginal at 2.47; p-value = 0.13. However, X is a synergistic variable and only becomes significant when X_2 and X_4 are included in the model. Thus, the identification of synergistic combinations and a knowledge of the subject matter is used in selecting regression model (7) over model (6).

SYNERGISM

Observe, X_2 (dividends) only becomes significant when X_1 is included in the model. The absence of X_1 causes X_2 to be deleted thereby magnifying the misleading effects of the underspecified financial model. Table 3 reveals that $r_{y.2} = -0.442$ possesses a negative correlation coefficient. This is caused by the mathematics of the yield ratio. The larger the price, the larger is the base on the yield ratio. This causes high price stocks to possess lower yields than low price stocks. Observe that $r_{y.1}r_{y.2} < 0$. As stated above, when $r_{y.1}r_{y.2} < 0$, X_1 and X_2 are synergistic ($R^2 > r_{.1} + r_{.2}$) when $r_{1.2} > 0$. Hence, this property of synergism causes X_1 and X_2 to quickly become synergistic. In multiple regression,analysts should be alert in obtaining variables which are inversely related. These variables are potential synergistic variables.

The partial F and p-values in Table 2a along with $r_{y.2} = 0.191$ in Table 3 reveal that X_2 is synergistic. It possesses a significant partial F value (30.86) and an insignificant simple r-value. Observe from Table 2(d), X_2 acts as an unrelated variable until combined with X_1. The distinction between unrelated variables and synergistic variables without the proper combination is difficult to detect until an appropriate variable is included in the regression model.

DISCUSSION AND CONCLUDING REMARKS

The above example illustrates the concept of synergism in regression models. This example also illustrates the misleading results that may occur when a synergistic variable is excluded. It is important to keep the concept of synergism and the definition of synergistic variables separate. Synergism is when R is greater than the sum of the simple r values. A synergistic variable is a regressor with an insignificant simple determination coefficient and a significant partial determination coefficient.

Multicollinearity is helpful for synergistic variables but can become a problem for other variables possessing significant simple determination coefficients. For example, there is a considerable amount of multicollinearity between variables X and X ($r_{4.5} = 0.967$). However, since both variables possess significant simple r-values, the multicollinearity between them causes X to become unimportant and possess an insignificant partial F value. Thus, multicollinearity between variables with significant simple determination coefficients usually renders one or more of them unimportant. Multicollinearity between variables with insignificant simple determination coefficients is usually helpful in making then synergistic and their partial F values significant.

It is possible for all of the variables in a synergistic combination to possess insignificant simple r values but significant partial F values. It is also possible for a variable in a synergistic combination to possess an insignificant partial F value. A variable which is marginally related to Y but part of a synergistic combination should be included in the model. If it is not included in the model, the synergistic combination may disappear. In the above example, the deletion of X causes the synergistic variable X to become insignificant. When this happens, personal judgment and a knowledge of the subject matter must guide one into selecting the appropriate model.

In the 1970s, diagnostic test statistics, such as the Durbin-Watson statistic, were initially interpreted as suggesting estimation problems. These problems were dealt with by adopting more sophisticated estimation methods rather than dealing with misspecification of the chosen model [6]. Today, there is a realization that many times autocorrelation and violations such as normality and homoskedasticity may be correctly viewed as underspecification error. Kraner [7], also Maddala [8] have excellent expositions on misspecification errors. Belsley [1] argues for the use of prior information in specification analysis. Certainly, econometric models must be developed by people well grounded in economic theory and a firm knowledge of the subject matter.

Finally, in the preliminary research phase, always obtain a further knowledge of the subject matter. Take time to identify influential variables. In the analysis phase of the research, if intuition strongly indicates that a seemingly unrelated variable should be related to Y, return to the preliminary research phase. Search for additional variables so that the proper synergistic combination is included in the regression model.

REFERENCES

Besley, D.A. (1986), Model Selection in Regression Analysis, Regression Diagnostics and Prior Knowledge," *International Journal of Forecasting* 2, 41-6, and commentary 46-52.

Daniel, C., & Wood F.S. (2000). *Fitting Equations to Data*, 3nd.ed., New York: John Wiley.

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

Hamilton, David (1987), "Sometimes $R^2 > r_y{}^2{}_{.1} + r^2{}_{y.2}$: Correlated Variables Are Not Always Redundant," *The American Statistician*, 41, 129-132.

Kendall, M.G. and Stuart, A. (1973), *The Advanced Theory of Statistics*, (Vol. 2, 3rd. ed.). New York: Hafner Publishing.

Kennedy, Peter (1998), *A Guide To Econometrics,* 4th ed., Cambridge, Mass., The MIT Press

Kramer, w. (1985), "Diagnostic Checking in Practice", *Review of Economics and Statistics* 67, 118-23.

Maddala, G.S. (1995), "Specification Tests in Limited Dependent Variable Models", *Advances in Econometrics and Quantitative Economics*, Oxford, Blackwell, 1-49.

Mitra, S. (1988), "The relationship between the multiple and the zero- order correlation coefficients," *The American Statistician*, 42, 89.

Ryan, T.P. (1997), *Modern Regression Methods*, New York, Wiley.