Circling the Square: Experiments in Regression

R. D. Coleman [unaffiliated]

This document is excerpted from the research paper entitled “Critique of Asset Pricing Circularity” by Robert D. Coleman dated 15 February 2002.

Econometrics is inferential statistics applied to economics. Econometrics, an inductive method, is not mathematics, a deductive method. Time series analysis is not econometrics. Time series analysis of a sequence of variables generated by human behavior is not science, it is history. For example, time series analysis of a sequence of radioactive decay intervals is scientific. Time series analysis of a sequence of stock market prices is not scientific, but rather history. Likewise, the issue of equations in circular form is not econometrics, but rather logic and mathematics. Just as in grammar, a definition that includes the term to be defined is a circular definition and a fallacy, in math, an equation with the term to be solved on both sides is a circular equation and a fallacy.

In a theoretical algebraic equation, the relationship among the variables is known, but the values of the variables are unknown. In an applied algebraic equation, the value of one variable is unknown, and the values of the other variables are known. In contrast, in a statistical regression equation, the values of all variables are known in a sample of observations, but the relationships among the variables are unknown.

Conjoining variables renders an equation or variable relationships intractable due to statistical interaction and similar effects and thereby confounds interpretation. To avoid such inextricable intertwining, it is necessary in an algebraic equation to isolate the unknown variable on the left-hand side and the known variables on the right-hand side. For the same reason, it is necessary in a regression equation to isolate the explained or dependent variable on the left-hand side and the explanatory or independent variables on the right-hand side. This is accomplished by various operations. The basic operations are simplify, factor, reduce to lowest common denominator, substitute, transpose and

rearrange. Other operations are multiply, divide, add or subtract the same term or otherwise transform on both sides. Additional operations include multiplying or dividing one side with a combination of terms that is equivalent to one, and adding or subtracting on one side a combination of terms that is equivalent to zero.

If the equation contains practical variables rather than abstract variables, it is also necessary to insure that the values of the variables have consistent units of measurement. This is done by either converting variables or adjusting data values. Isolation and units analysis result in the proper reduced form of the equation. Only then is the equation ready for solution. An algebraic equation is solved in a specific case by replacing the variables on the right-hand side with their respective values. In a statistical regression, the equation is solved by estimating the coefficients of the right-hand side variables based on a sample of observations.

Any variable included on both sides of a least-squares regression equation may result in greater significance as measured by Student's *t* and in a better "fit" as measured by R-square. For such improvement in these and other test statistics, there is a price to pay in bias in the estimated coefficients. This bias is the result of the induced correlation due to logical circularity. The dependent explained variable may be isolated on the left-hand side and yet the equation may still be logically circular if there is any adjustment factor that appears on both sides. Such circular adjustment factors due to economic reasoning are sometimes necessary due to the theory represented by the equation. But any econometric technique that requires a circular adjustment factor is elective and not logically necessary.

An equation is logically circular in any variable that appears on both sides. Thus, an equation with the explained variable on both sides is logically circular in the explained variable. Such an equation lacks face validity and thus scientific interest, and is an unwarranted assertion at best. Moreover, a mere equation is not a theory. A regression equation represents the operationalization of a theory. Any author of a logically circular equation bears the burden of proof to justify the theory behind it. The specified equation must be motivated by a plausible story that explains the inclusion of each variable. This justification is of greater importance if application of the theory can result in greater damage.

A. Example Formula

The Table summarizes selected regressions of logically circular equations for unit rectangles. The terms of the equations are related by formulas, i.e., they are tautological. These ordinary least squares (OLS) regressions illustrate the kinds of bias in the estimated coefficients that result when the equation is not in proper reduced form. These estimated coefficients have greater precision because the observations are calculated from the formulas rather than measured on actual rectangles. The sample size includes all rectangles with lengths and widths from 2 to 9 integer units for a total of 64 observations.

First, we compare six regressions of length and width on perimeter, on area, and on compactness. Length (L) is defined as the vertical dimension, and width (W) as the horizontal dimension. Perimeter (P) is defined as length added to width and then multiplied by two. Area (A) is defined as length multiplied by width. Compactness (C) is defined as area divided by perimeter. Not surprisingly, length and width have high significance as measured by Student's *t* and high explanatory power as measured by R-

square. Also not surprisingly, significance and power decrease as we go from perimeter as the explained variable which is additive in length and width, to area which is simple multiplicative, to compactness which is compound multiplicative.

Next, we compare five regressions of area, perimeter, and inverse perimeter on compactness. The inverse of a variable is defined as one unit divided by that variable. It is not surprising that perimeter and area, each of which entails both length and width, are highly significant and that the specified variables have high explanatory power. It also is not surprising that in the equation with two explanatory variables, inverse Perimeter is more significant than perimeter because that is the form in which perimeter is entailed in compactness. A natural logarithm transform of both sides should result in a zero intercept, coefficients equal to 1.00, Student's *t* equivalent to zero probability, and 100% R-square.

Last, we compare 8 regressions of oddness, sides ratio, diagonal, and radius on compactness, and 4 regressions of these four variables on compactness with deflation and the intercept constrained to equal zero. Deflating the variables by oddness transforms the intercept to inverse oddness. Oddness (O) is defined as a nominally measured variable with four numerical categories: 1 if both length and width are even; 2 if length is even and width is odd, 3 if length is odd and width is even, and 4 if both length and width are odd. Sides ratio (S) is defined as the long side divided by the short side. Diagonal (D) is defined as the hypotenuse of the right triangle with length and width as sides. Radius (R) is defined as that of a circle with area equal to a rectangle with length and width as sides.

It is not surprising that all four of these variables contribute individually to explanatory power because they each entail length and width. It is not surprising that the

explanatory power of these variables increases from oddness to sides ratio to diagonal to radius due to increasing information and closer conformance to the form of the explained variable. For the opposite reasons, it is not surprising that oddness is not statistically significant at the 5% probability level in such a small sample.

It also is not surprising that the increase in explanatory power due to the step-wise addition of inverse oddness as a deflator to the regression equation is greater when the explanatory power at the prior step begins at a lower level. The increase for oddness alone is 54 percentage points from 4%. The increase is greater for oddness alone than in combination with sides ratio (up 39 percentage points from 29%), in combination with sides ratio than in combination with diagonal (up 12 percentage points from 78%), and in combination with diagonal than in combination with radius (up one percentage point from 98%).

Oddness entails length and width quite indirectly, in only four categories. In this sample, neither oddness nor its inverse is statistically significant when regressed on compactness, and each explains only 4% of the total variation in compactness. Yet, dividing all terms on both sides by oddness as a deflator in the regression of either radius, diagonal, or sides ratio on compactness, results in an increase of one to 39 percentage points in explanatory power. The regression of oddness on compactness illustrates the phenomenon of false negatives, i.e., logically circular variables do not necessarily result in high statistical significance or notable explanatory power. Values of test statistics depend on the form of the variable, the presence of other variables competing in the same equation, sample size, sample composition, and methodology.

This illustration of counter-intuitive false negatives is reinforced by comparison of the regressions of Oddness with the regressions of X, defined as a random variable equal to an integer ranging from one through 100 for each rectangle. One run each of two regressions of X is reported in the Table. Ten runs resulted in a Student's *t* ranging from 0.45 to 1.85 for X when regressed on compactness, but it was always equal to zero for inverse X when X was a deflator with the intercept constrained to equal zero. These ten runs resulted in a R-square ranging from near zero to 7% for each regression with different combinations of Student's *t* below or above the 5% probability level and R-square increasing or decreasing from the regression on compactness to the regression on compactness deflated by X.

Non-explanatory variables are sometimes used to modify operational variable definitions or to adjust sample data. In addition to "size" deflators and unit conversion factors, this includes heteroskedasticity correction and other data weightings. Such data adjustments are ad hoc and arbitrary. They are avoidable and not necessary. The specification of such adjustment factors on both sides of the equation may improve the goodness of fit at the price of bias in the estimated coefficients due to the spurious correlation that results from logical circularity. This bias in turn impacts both the levels of significance and the explanatory power. If a variable used as an adjustment factor is isolated on the left-hand side of an equation, then its role changes from adjustment factor to explained variable.

All regressions run on formulas of rectangles are reported in the Table. The most unexpected result is the high significance and explanatory power of oddness (t=10 and R-square=58%) when specified as a deflator in the explanation of sample variation in

compactness. Yet, it can be concluded there is no evidence that oddness belongs in the equation in any form. It appears in a significant form in the deflated equation only because the other variables were divided by it. Also unexpected is the high significance and explanatory power of diagonal (t=15 and R-square=78%) and of radius (t=56 and R-square=98%), even with each directly entailing both length and width.

B. Return Formula

To identify potential logical circularity in capital asset pricing models, we analyze the formula for return on investment. The definition of return on a common stock is:

$$R_t = \left(\frac{\frac{D_t}{S_t}}{\frac{P_{t-1}}{S_{t-1}}} \right) + \left(\frac{\frac{P_t}{S_t} - \frac{P_{t-1}}{S_{t-1}}}{\frac{P_{t-1}}{S_{t-1}}} \right) = \left(\frac{D_t}{(P_{t-1})(S_t/S_{t-1})} \right) + \left(\frac{P_t - (P_{t-1})(S_t/S_{t-1})}{(P_{t-1})(S_t/S_{t-1})} \right)$$

where R is return, D is dividend per share, P is share price at the end of a holding period, S is number of shares, and t indexes holding periods. It can be seen from this equation that R_t , D_t , P_t , P_{t-1}, S_t and S_{t-1} should not appear on the right-hand side of any least-squares regression equation to explain R_t on the left-hand side. None of the six variables should appear on the right-hand side in any form, whether directly, entailed, forming portfolios, or any other method.

The capital asset pricing model (CAPM) of modern portfolio theory was independently developed by Sharpe (1964) and Lintner (1965). Reconciliation of these two versions of the CAPM involves excluding the priced security from the market portfolio. To avoid logical circularity, the stock with beta risk factor, β_i, and return, R_i, on the left-hand side is not included in the market portfolio, represented by R_m, or its proxy, R_M, on the right-hand side of the equation. Thus,

$$\beta_{i,m-i} = \frac{\operatorname{cov}(\tilde{R}_i, \tilde{R}_{m-i})}{\sigma^2(\tilde{R}_{m-i})} \text{ and } \tilde{\bar{R}}_{it} = \tilde{a}_i + \tilde{b}_i R_{m-i,t}$$

where beta of the stock relative to the market is constant and return on the stock varies by holding period. For example, it would be logically circular to use the DJIA as a proxy for the stock market to estimate beta for one of the 30 component stocks in the DJIA.

Analysis of units of measurement may reveal changes in shares. Time-indexed shares represent different units when they are claims on different fractions of total equity and voting rights. For example, stock splits based on the exchange ratio, S_t:S_{t-1}, result in S_t "new" shares for each S_{t-1} "old" share. More generally, the shares adjustment factor equal to S_t/S_{t-1} is used for adjusting the old price, P_{t-1}. Before solving, old shares must be converted to new shares and the price of old shares adjusted. If there are no stock splits, stock dividends, spin-offs or recapitalizations that change the number of shares outstanding during the period from time *t* to time *t-1*, then S_t and S_{t-1} are effectively eliminated from the equation because both would be equal to one unit.

Net return can be calculated with the inclusion of variables for transaction costs and taxes. Real return can be calculated with the addition of a price-level inflation variable. Other-currency return can be calculated with the addition of a foreign exchange ratio variable. Adjustment factors for net, real, and foreign-currency return may be necessary in some cases, yet introduce a bias when present on both sides of an equation.

Autoregressions of an economic time sequence are not necessarily logically circular. Each term in the sequence is a different variable due to time indexing. The regression equation, $P_t = \beta * P_{t-1} + e_t$, is a first-order autoregression with two price variables. This autoregression is in proper reduced form and thus is not logically circular.

The regression equation, $R_t = \beta * R_{t-1} + e_t$, is also first-order autoregressive, but it is logically circular because P_{t-1} is entailed in R on both sides.

This analysis of the formula for return shows that high statistical significance and explanatory power are neither certain nor surprising if R_t, D_t, P_t, P_{t-1}, S_t or S_{t-1} is specified as an explanatory variable in a regression equation to explain either return or a variable entailing return. All five variables embedded in return are controlled by human behavior. The share prices are determined by stock market participants, and the shares outstanding and dividends per share are set by company directors. Thus, autoregressions of return or any of its entailed variables are not scientifically valid.

Table. Estimated OLS Regression of Logically Circular Equations for Unit Rectangles. Explained Variable (DV), Explanatory Variables (IV), estimated coefficients (B), Student's *t*, and R-square coefficient of multiple determination. Sample Size = 64.

DV	IV	B	t	IV	B	t	R-SQ
P	L	2.00	8				50%
P	L	2.00	2e15	W	2.00	2e15	100%
A	L	5.50	7				46%
A	L	5.50	19	W	5.50	19	92%
C	L	0.13	7				45%
C	L	0.13	16	W	0.13	16	90%
C	A	0.02	44				97%
C	P	0.06	23				90%
C	P	0.002	0.46	A	0.02	12	97%
C	1/P	-19.24	11				67%
C	1/P	-2.29	3	A	0.02	26	97%
C	O	0.08	1.66				4%
C	1/O	-0.30	1.55				4%
C/O *	1/O	1.16	10				58%
C	S	-0.26	5				27%
C	S	-0.25	5	O	0.06	1.39	29%
C/O *	S/O	-0.23	5	1/O	1.62	15	68%
C	D	0.17	15				78%
C	D	0.17	14	O	0.01	0.51	78%
C/O *	D/O	0.17	14	1/O	-0.14	1.46	90%
C	R	0.46	56				98%
C	R	0.46	55	O	-0.0001	0.11	98%
C/O *	R/O	0.46	55	1/O	-0.12	5	99%
C	X	-0.004	1.85				5%
C/X *	1/X	0.00	0.00				1%

* Intercept constrained to equal zero.
t (60 df, 2-tailed test): 2.00 = 5% probability and 2.66 = 1% probability.

LEGEND

A = area = L*W
C = compactness = A/P
D = diagonal = sqrt(L**2+W**2)
L = length (2, …, 9)
O = oddness (1, 2, 3 or 4)
P = perimeter = (L+W)*2
R = radius of circle = sqrt(A/π)
S = sides ratio = long side/short side
W = width (2, …, 9)
X = random integer (1, …, 100)

REFERENCES

Lintner, J., 1965, “The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets”, *Review of Economics and Statistics* 47 (February): 13-37.

Sharpe, W. F., 1964, “Capital asset prices: A theory of market equilibrium under conditions of risk”, *Journal of Finance* 19 (September): 425-442.