The Unscientific Incompleteness and Bias of Unidirectional Projections (= Regressions): A Questionnaire

Cornelis A. Los, PhD, Associate Professor
Kent State University, Department of Finance, BSA416,
Kent, OH, 44242-0001, Email: clos500@cs.com

October 26, 2004

Abstract

Why do statisticians (econometricians, economists, financial analysts, etc.) continue to incompletely identify the algebraic/geometric structure of the multi-variate data series they profess to *analyze*, and instead continue to publish the results of incomplete, prejudiced and biased unidirectional projections (= "regressions") of such covariance structures? Such incomplete, prejudiced and biased representations cannot lead to scientific knowledge, as has been demonstrated already more than twenty years ago.

1 INTRODUCTION

Based on an extensive survey of published statistical articles and several data bases since my original *CAMWA* articles of 1989 (Los, 1989a and b), plus my rebuttal to Zellner and Jaynes (Los, 1992), and my "Galtons Error" article (Los, 1999), I would like to raise the following fundamental methodological question regarding (linear) modeling by statisticians (incl. econometricians, biometricians, psychometricians, financial analysts, etc.) and I would like to receive comments on it from mathematicians, statisticians and every other scientist who feels compelled to respond.

My question is as follows: Why do statisticians (econometricians, economists, financial analysts, etc.) continue to incompletely identify the algebraic/geometric structure of the multi-variate data series they profess to *analyze*, and instead continue to publish the results of incomplete, prejudiced and biased unidirectional projections (= "regressions") of such covariance structures? Such incomplete, prejudiced and biased representations cannot lead to scientific knowledge, as has been demonstrated aleady more than twenty years ago. Since 1983 they have leaned from the articles of Kalman and me that such unidirectional "regression analysis" is scientifically worthless.

I always put the following question to my students: "How can one assess the volume of a 3-dimensional object, by taking only one one-sided picture of it?" [Obvious answer: one can't! But statisticians continue to do so, despite the fact that all their "significance testing" is based on the same one one-sided picture and is thus not complete].

2 TWO SIMPLE QUESTIONS

This fundamental epistemological question can be further decomposed into two simpler questions:

[1] Why is it that statisticians always select one particular projection direction (unidirectional projection), by making an *a priori* distinction among the variables in a given data set, *i.e.*, between "left hand side variables" (e.g., "regressands," or "explained variables") and "right hand side variables" (e.g., "regressors," or "explanatory variables")? Neither the modeling algebra nor the correlation geometry of the data provides a scientific basis for such a presumption, because the models used are always equations and thus an *a priori* "distinction" between such variables is scientifically, logically and empirically, unwarranted.

Take, for example, the simple bivariate linear model:

$$a.x_1 + b.x_2 = 0$$

where the series x_1 and x_2 have $T > 2$ observations and a and b are the parametric coefficients to be identified from the uncertain data. Statisticians make first an *a priori* distinction between variables x_1 and x_2 and then write this model as

$$x_1 = c.x_2$$

(for the projection of x_1 on x_2), where the parameter $c = -b/a$, although the alternative choice

$$x_2 = d.x_1$$

(for the projection of x_2 on x_1), where the parameter $d = -a/b$, is just as valid an entity to be identified, solely based on the (linear) algebra and the two data series. (Undergraduate students recognize this as the "reverse regression," a somewhat misleading term, but after graduation everybody conveniently forgets about "reverse regressions").

If statisticians would attempt to realize the complete empirical data correlation structure by projection, they should present at least both extreme (orthogonal) projections, from which all other linear combinations can be derived:

$$\widehat{c} = \frac{\sigma_{12}}{\sigma_{22}} \textbf{ and } \widehat{d} = \frac{\sigma_{12}}{\sigma_{11}}$$

The *coefficient of determination*, which gives the percentage of the system identification is, after all:

$$\rho_{12}^2 = \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} = \widehat{c}.\widehat{d}$$

In other words, the coefficient of determination incorporates the measurements from both projections, and not from one unilateral projection.

But the published literature shows overwhelmingly that statisticians don't. Only one unique unilateral projection direction is published, $\widehat{c}$, and thus their analysis remains incomplete. If the coefficient of determination ρ_{12}^2 is co-published with the measured $\widehat{c}$ the second projection measurement $\widehat{d}$ can be easily established in the case of this simple bivariate system:

$$\widehat{d} = \rho_{12}^2 / \widehat{c}$$

For easy comparison one should set side by side the two measurement results on a normalized basis, either:

$$\begin{aligned} x_1 &= \widehat{c}.x_2 = \frac{\sigma_{12}}{\sigma_{22}}.x_2 \text{ and} \\ x_1 &= \frac{1}{\widehat{d}}.x_2 = \frac{\sigma_{11}}{\sigma_{12}}.x_2 \end{aligned}$$

or

$$\begin{aligned} x_2 &= \widehat{d}.x_1 = \frac{\sigma_{12}}{\sigma_{11}}.x_1 \text{ and} \\ x_2 &= \frac{1}{\widehat{c}}.x_1 = \frac{\sigma_{22}}{\sigma_{12}}.x_1 \end{aligned}$$

Statisticians should know that incomplete data analysis is prejudiced and biased, per definition, and that for an unprejudiced, unbiased scientific analysis a complete presentation of all the measurements is required. Of the two presentations only the first half of the first presentation is conventionally published.

Let's now extend our line of analysis towards the multivariate systems and ask the following question:

[2] Why is it that statisticians prefer single equation (linear) models (*i.e.*, "planes") above simultaneous independent equation models, when analyzing multi-dimensional data sets? This occurs despite the fact that, for example, lines in a 3-dimensional data space can only be described by two simultaneous independent equations.

Take, for example, the data set (x_1, x_2, x_3), where each of these three data series has $T > 3$ observations. Two simple linear (or linearized) model structures are possible:

(a) one single equation (= a plane):

$$a.x_1 + b.x_2 + c.x_3 = 0$$

(a, b and c are the parametric coefficients to be measured); and

(b) two simultaneous independent equations (= a line):

$$\begin{aligned} d.x_1 + e.x_2 &= 0 \text{ and} \\ f.x_1 + g.x_2 &= 0 \end{aligned}$$

and (d, e, f, and g are the parametric coefficients to be measured).

If statisticians would attempt to realize the empirical data correlation structure, they should present the complete set of identification results of both these models, in at least three extreme (orthogonal) projections for each model structure (*i.e.*, six projections in total). All other projection directions can then be found by linear combination.

The correlation structure of the given data is, hopefully, discriminatory enough to be able to discriminate between case (a), one plane, and case (b), two planes and thus a cross-line. If not, there is too much epistemic uncertainty for a linear model to be identified and not much scientific progress can be made.[1]

3 INCOMPLETENESS AND BIAS

We can quantify the percentage of analytic incompleteness and the under-reporting of the possible number of projections, using the following two measures.

First, the information matrix $\mathbf{\Sigma}^{-1}$ is the inverse of the covariance matrix $\mathbf{\Sigma}$ of all n variables in the data set.[2] Each row of the information matrix is an elementary regression or $(n, 1)$ unidirectional LS projection. Since only one of these elementary regressions is reported in each of the following articles, the

$$\text{Percentage of analytic incompleteness} = 100.(n-1)/n\%. \tag{1}$$

Second, the complete number of projections of the invariant number q of possible linear relations among n variables is given by

$$\text{Number of possible LS projections } = \sum_{q=1}^{n-1} \left(\begin{array}{c} n \\ q \end{array} \right) \tag{2}$$

The under-reporting is this number minus the one $(n, 1)$ unidirectional projection that is reported. The following Table 1. provides some examples of these measurements of published scientific incompleteness and bias to demonstrate the seriousness of the problems. These examples are discussed in greater detail in Los (2004).[3]

[1] A complete discussion of both the bivariate and trivariate cases can now be found in Los, 2001, Chapters 4 and 5, with corresponding solutions to their Exercises in Kassabov and Los (2004)..

[2] A lagged variable counts as a separate variable.

[3] LS = Least Squares projection (regression);
VAR = Vector Auto-Regression;
ADF = Augmented Dickey-Fuller test

TABLE 1. SCIENTIFIC INCOMPLETENESS			
Article (Type of Analysis)	**# of**	**Analytic**	**# of unreported**
	Variables	**Incompleteness %**	**LS projections**
Fama (1990) (LS)	5	80	29
Schwert (1990) (LS)	5	80	29
	9	89	509
	12	92	4,093
Bittlingmayer (1992) (LS)	8	88	253
Canova & De Nicolo (1995) (LS)	4	75	13
	6	83	61
	9	89	509
	11	91	2,045
Lee (1992) (VAR)	28	96	268,000,000
Gallinger (1994) (ADF)	13	92	8,189
	15	93	32,765
	17	94	131,060
	18	94	262,141
	19	95	542,285

4 CONCLUSION

The published academic and non-academic literature shows overwhelmingly that statisticians don't follow such a logical and complete analytic methodology. Based on multivariable data sets only one the measurement of one unilateral projection direction is published and that is usually the unilateral measurement of one single equation model, measured uniquely in one direction. This is the prevalent example of incomplete and thus prejudiced and biased data analysis that must be eradicated otherwise scientific progress remains stopped in its tracks, in particular in the so-called social sciences.

Notice that when the number of covariant empirical data series increases, the complexity of the system identification problem increases more than commensurately. But when one reads the statistical literature, one never senses such a dramatic increase of the complexity of the system identification problem. Statisticians blissfully tend to lump five, ten or more data series in a single linear equation (= a hyper-plane) and report the measurement results of only one unilateral projection direction.

We have been very surprised and saddened by the complete ignorance of this prejudiced practice among statisticians. Why do other scientists not protest more against such an incomplete prejudiced and biased statistical system identification? None of these incompletely measured statistical system identification results should be acceptable or deemed credible and valid, since incomplete scientific evidence is presented in each case. All the so-called "significance testing" is prejudiced and biased, because its testing statistics are based on one unilateral projection only and are incomplete.

All comments on this issue (except emotional "flames") are welcome. This is

a serious epistemological question and it is essential that all sciences (including the so-called social sciences) are honest and sincere about these observed prejudices and biases caused by incomplete measurements, for the sake of enlightened scientific progress. I only hope that my email box doesn't become overwhelmed by the avalanche of your responses, since we have been severely disappointed by the lack of response from the statistical community in the past two decades.

5 REFERENCES

Los, Cornelis A. (1989a), "The Prejudices of Least Squares, Principal Components and Common Factors," *Computers & Mathematics With Applications*, Vol. 17, No. 8/9, April, 1269 - 1283.

Los, Cornelis A. (1989b), "Identification of a Linear System from Inexact Data: A Three Variable Example," *Computers & Mathematics With Applications*, Vol. 17, No. 8/9, April 1989, 1285 - 1304.

Los, Cornelis A. (1992), "Reply to E. T. Jaynes' and A. Zellner's Comments on My Two Articles," *Computers & Mathematics With Applications*, Vol. 24, No. 8/9, August, 277 - 288.

Los, Cornelis A. (1999), "Galton's Error and Under-Representation of Systematic Risk," *Journal of Banking and Finance*, Vol. 23, No. 12, December, 1793 - 1829.

Los, Cornelis A. (2001), *Computational Finance: A Scientific Perspective*, World Scientific Publishing Co., Ltd, Singapore, 336 pages (ISBN: 9810244967, hard cover, January 2001; ISBN: 9810244975, paperback, July 2001. Currently in its 2nd print).

Los, Cornelis A. (2004), "System Identification in Noisy Data Environments: An Application to Six Asian Stock Markets," *Journal of Banking and Finance*, Vol. 28, 2004, 35 pages (on 10/31/2003 accepted for publication with small editorial changes).

Kassabov, Milen, and Cornelis A. Los (2004), *Solutions Manual to Accompany Computational Finance: A Scientific Perspective*, World Scientific Publishing Co., Ltd, Singapore, 2004, 113 pages. (ISBN: 981256036X, e-version).