Reply to Bezing's and Dunlevy's Comments on "A Scientific View of Economic Data Analysis"

Cornell A. Los

Cynthia Bezing and Kevin Dunlevy (B&D) provide an erroneous, but not uncommon, interpretation of the scientific problem of modeling from instance data and of its solution. To correct their error and clarify the concepts of exact and inexact modeling, I begin by asking the following: What model can be identified from data such as those in Figure 4? Specifically:

1. Are these data series exact or inexact, and are they stochastic or not?
2. How many linear relationships can these data be minimally described? As Einstein once said: "A scientific explanation must be as simple as possible, but not simpler."
3. What do these relationships look like?

In this sample case the answers to the three questions are given at once by plotting the data in their 3-dimensional data space. The black dots in the center of Figure 2 are the data points. The area dots in the three basis planes of $(x_1,x_2)$, $(x_2,x_3)$, and $(x_1,x_3)$, respectively, are the projections of these 32 observations. These projections are connected to the data points by drop lines. Note that:

1. The data are obviously exact: they lie along a ray, as is also shown by their 2-dim. projection on the basis planes.
2. The data can be minimally described by two independent linear equations.
3. These two linear model orders can be written in convenient matrix notation as $A = 0$, where $A$ is the coefficient matrix of order $(3 \times 2) = (2 \times 3)$ and the matrix of 32 observations on 3 variables, i.e. of order $(3 \times 3) = (3 \times 3)$.

There is no scientific justification for the statistician's distinction between "explained" and "explanatory" variables on the basis of only the data, since a linear model is nothing but the minimal linear combination of exact data series. Note from Figure 2 that $x_1$ is positively related to $x_2$ and simultaneously negatively to $x_3$. Consequently, $x_2$ is also negatively related to $x_3$.

The following example explains what is meant by exact data and exact linear modeling. Series $x_1$ happens to be an empirical data series, measured as deviations from their mean, while $x_2$ and $x_3$ are constructed from $x_1$ using the model. The model realizes all the data and the data are realized by the model. The model is thus an isomorphism of the exact data. Note that the concept of exactness does not say anything about stochasticity or probability. Series $x_1$, and thus $x_2$ and $x_3$, could be stochastic or deterministic. That question is irrelevant for identification.

The forecast example of exact data contains empirical economic data previously used in two technical articles [Los, 1968b, 1992]. Since one picture is worth a thousand words, it suffices to summarize that discussion by using pictures analogously to the ones above. (These pictures are not in the referred articles.)

The shaded box black empirical data dots are plotted in their 3-dim. data space in Figure 3. The area dots are their projections on the 2-dim. basis planes. This time only the drop lines for the projection on the $(x_2,x_3)$ plane are drawn to not clutter the picture too much.

The projections show positive correlation between $x_1$ and $x_2$, negative correlation between $x_1$ and $x_3$, and negative correlation between $x_2$ and $x_3$, as is evident in the data covariance matrix in the first North-West quadrant of Figure 4.

According to a very useful theorem of Kalman, each row of the inverse (or rather, the adjoint) of the data covariance matrix $A$ in the second, North-West quadrant of Figure 4, is an elementary, $q = 1$, single equation least squares $(1.5)$ regression. This can be demonstrated by normalizing on the diagonal elements a Toeplitz matrix in the South-East quadrant of Figure 4. For example, by dividing the first row by the first diagonal element we have the elementary regressions of $x_1$ on $x_2$ and $x_3$, and it can be repeated by running the conventional regression software on the data series in Los, 1968b, pp. 1301-1302. Similarly, the second row is the regression of $x_2$ on $x_1$ and $x_3$, and the third row of $x_3$ on $x_1$ and $x_2$. (Note that we use deviations from the means of the respective series. Inclusion of the means in the model is straightforward.)

To compare these regression results, we have normalized, arbitrarily, on the first element of the rows, by dividing each row by its first element. Other normalizations lead to equivalent results. Note that the system coefficients $b_1$ and $b_2$ of the system coefficient matrix $A = [1 \ b_1 b_2]$ are sign consistent with each other. For example, in normalized regressions $b_1$ and $b_2$ are negative, but in regression $c'_{1,2}$ is positive. The $b_0$ coefficients of these three regressions lie in different orthants. Ergo, such a one-equation model cannot resolve the required signs consistency with the data covariance matrix.

B&D's elementary problems and their solutions are nowhere explained in the statistical literature; also not by the alleged expert on "reverse regressions," who maintains the prejudice of a priori knowing the number of equations without understanding that precisely this prejudice leads to the observed inconsistencies with the complete data. Indeed, B&D state, "Exact regressions do not specify an exact system of equations to be used nor does it tell us the interrelationships of the variables in the system." Wrong. It does, as I will now demonstrate.

Let's increase the rank of the system coefficient matrix $A$ from 1 to 2. There are three ways of selecting two rows out of the three available in the inverted data covariance matrix. By arbitrary rotation, that is by pre-multiplication by a $(2 \times 2)$ positive definite matrix, I have created some ones and zeros, to reduce the number of valued coefficients and to facilitate comparison of the results.

Thus, analogously to the exact data case, I computed the three normalized least squares simultaneous two-equation systems $(H, E)$, $(F, G)$ and $(D, J)$, notice from the three South-East quadrants in Figure 5. That we now complete sign-consistency results and that, despite the dimension of the data, the range of values of the relevant A coefficients is very narrow.

Note that values of the coefficients in matrix $A$ remain arbitrary up to the multiplication by any positive definite matrix. That is why we expect that only invariant of a linear system is $q$, the minimal rank of $A$.

This equivalence result of parameterization is well-known in econometric "simultaneous equation estimation" and in factor analysis. The system coefficient values of $A$ have no intrinsic value. The "frame of reference" (i.e., the orthogas in Figures 2. and 3.) can always be retained to gain an interpretation that is meaningful to the economic researcher.

Figure 6 summarizes our 5-dim. results. The three computed LS two-equation models are represented by the arched rays through the black data points in the center of Figure 6. They capture all systematic variation consistent with the data, given the three possible projection choices of assumed exact variables. The corresponding three signless standard deviations show that the two-equation model explains between 64.9% and 86.5% of the variation of these empirical cross-sectional data. On the basis planes I have projected these three LS rays to show how well they represent the exact components of the covariances between $x_1$ and $x_2$, between $x_2$ and $x_3$, and between $x_1$ and $x_3$. The rays deviate slightly from each other because of the inexactness of the data or "lack of linear fits." If the data had been less inexact, the black dots would cluster closer along a ray from the cone formed by the three least squares regression results $L, J$, and $L$. (The origin of the data is located in the center of the picture.) Moreover, these three LS rays would cluster closer together. Of course, thanks to continuity, in the limit the black data points and the three LS rays coincide in the exact data case.

A single equation model would represent an indeterminate plane through the data ray in the center of Figure 2, which would satisfy only two of the three relevant correlation signs at a time, but not all three simultaneously. The coefficients of a single equation model are indeterminate when the data are are realized by a two-equation model. If the prejudice is $q = 1$, then $A = [1 \ b_1 b_2]$, $[1 \ b_1 b_2]$, and $[1 \ b_1 b_2]$, with $b_1$ the same. This is in case of exact data: the coefficients of the postulated one-equation model is indeterminate linear combinations of the coefficients of the true exact two-equation model. Most empirical
Example of Data
T = 32 Observations

3 - Dim Data Space
Bank Profitability Indicators

3 - Dim Data Space
Exact Model: \( (n, q) = (3, 2) \)
\[ x_1 + 0.0602x_3 = 0 \]
\[ x_1 + 0.0631x_2 = 0 \]

Covariance Analysis To Identify q

Data Covariance Matrix \( \Sigma \)

Inverted Data Covariance Matrix

Normalization on First Column

Normalization on Diagonal

\[ A' = [1 \ b_1 \ b_2] \]
economic researchers have observed such 'instability of coefficients' resulting from the misspecification of the rank of their model A', like for example in the case of the 'money demand equation,' the 'Phillips curve,' or the logarithmic 'Cobb-Douglas production function.' But none of these researchers understood that it was their own prejudice of q = 1 that caused such instability and that each of these three examples should have been modeled by q = 1 models.

Our aim is to transform economics from subjective (mathematized) philosophy into objective exact science. It is no longer a question if economic and financial data can be analyzed scientifically and economics can become a science. Economics can be subjected to the same scientific rigor as physics and the so-called 'hard' or 'exact' sciences, despite the criticisms of the data. But our critique is that this has not been accomplished by the conventional 'soft' or 'exact' probability methods originally proposed by Haavelmo in 1944. Our empirical method has not obtained the status of economics to deal with real data, as BAE does, since even BAE correctly acknowledge that we 'know no room for subjective analysis.'

Probability plays no material role in scientific identification. It is remarkable that economists and statisticians continue to mistake stochasticity for uncertainty. What is needed is not necessarily stochastic, only genuine randomness. Contrary to BAE's beliefs, in quantum physics it is still not established that electrons move probabilistically or that it is our lack of exact measurement that makes us assume they do. After 70 years of debate, it is likely that also conventional quantum theory is not science but (mathematized) philosophy, a point that Albert Einstein would have appreciated [Avanesov, Benteleini, Carriotti, Lahl, Namik, Von Frassen and van der Merwe, 1992].

The point is that Science is measurement. For example, in 1953 James Watson and Francis Crick directly observed the exact 3-dim. atomic model structure of the DNA molecule from the x-ray, or "DNA," 2-dim. X-ray diffraction pictures taken by Rosalind Franklin and Maurice Wilkins late in 1952. To achieve this model realization, Watson and Crick applied King's exact mathematical [Law 2, eq. (h)] = q. They directly measured the two-dimensional reflections in that picture, expressed by the wavelength of the x-ray radiation and the angle of incidence of the x-rays in the DNA molecule. This knowledge of the exact spacing and identified the atomic bonding and the precise double helix structure with a diameter of 2 nanometers and a periodicity of 3.4 nanometers.

In contrast, Haavelmo's 1944 approach to identifying models via their noise characteristics, is miss-directed after Kalman (1989) show first two theorems that a particular linear model A' can be fitted to the data in a consistent fashion with infinitely many different noise patterns or residuals. Thus noise is undeniable, as it should be. Second, the generalised LS scheme can never produce the residuals required to let any linear model A' of any rank fit the data. A characteristic of linear LS is that the covariance matrix of the impact components of the data, is the smallest in a certain sense of all the noise covariance matrices that can be considered for the linear scheme A' and the data. But surprisingly, this is not the defining characteristic of LS; the rank q of model A's.

The (only) generalised LS scheme remains prejudiced, but its prejudice of the rank q can be neutralized by computing the complete regression models, as Frisch proposed almost 60 years ago [Frisch, 1934], and Kalman proved [1990]. As I demonstrated, all columns of the data covariance matrix together reveal the true rank q of the model.

It cannot be emphasized enough that the given in scientific research are the data. Data covariances can always be computed. Our method identifies the allowable narrow range values of the coefficients of A', up to an equivalency relationship, from the invariant data and separates the exact model from the noise. It is even capable of distinguishing outliers and other extraneous aberrations. There are no "true" or "non-true" data. Just data. The degree of meaningfulness depends on how close the data cluster around the exact linear model. (Note that we use linearity in the coefficients only. One-to-one transformations of the data series to fashion the pictures, so to say, is always allowed.)

With our paradigm results are true or false. The hypotheses that the empirical-data of the three-variable example above, can be described by a single-equation model, similarly, have demonstrated that I am firmly theoretical hypotheses about positiveness or negativity of relationships.

Nowhere do I contend that "if enough data were available, economists could precisely predict our future." Our paradigm is about identification from the data, not prediction. Prediction implies extrapolation beyond the available data based on the belief in the continuing integrity of the identified model. Such a

**REFERENCES**


R. Frisch [1954], Statistical confidence analysis by means of complete regression systems. Publication No. 5, University of Oslo Economic Institute, Oslo, Norway.

