Causal Paths and Exogeneity Tests in generalCorr Package for Air Pollution and Monetary Policy

Hrishikesh D. Vinod *

June 6, 2017

Abstract

Since causal paths are important for all sciences, my package 'generalCorr' provides sophisticated R functions using four orders of stochastic dominance and generalized partial correlation coefficients. A new test (in Version 1.0.3) replaces Hausman-Wu medieval-style diagnosis of endogeneity relying on showing that a dubious cure (instrumental variables) works. An updated weighted index summarizes causal path results from three criteria: (Cr1) lower absolute gradients, (Cr2) lower

*address: H. D. Vinod, Professor of Economics, Fordham University, Bronx, New York, USA 10458. E-mail: vinod@fordham.edu. Tel. 201-568-5976, Fax 718-817-3518, JEL codes C30, C51. I thank Prof. J. Francis for suggesting the 'excess bond premium' application.

absolute residuals, both quantified by stochastic dominance of four orders, and (Cr3) from goodness of fit. We illustrate with air-pollution data and causal strength of six variables driving 'excess bond premium,' a good predictor of US recessions.

1 Introduction

Econometrics still relies on the medieval diagnosis of a disease (endogeneity) because a remedy of instrumental variables (IV) estimator appears to "work." Actually, the IV remedy has been long known to be seriously flawed as shown by Bound et al. (1995) with a provocative title "the cure can be worse than the disease". This paper illustrates the use of an R package "generalCorr" to develop a new test which does not use any IV estimator. We indicate the very few lines of code needed to assess the preponderance of evidence in support of a causal path using macroeconomic examples which can serve as a template in many areas of research.

Review of Hausman-Wu test

Consider a possibly non-linear nonparametric regression:

$$Y = f(X_1, X_2, \ldots X_p) + \epsilon_1, \tag{1}$$

where the researcher wants to make sure that $E(X_i \epsilon_1) \neq 0$ holds. Assuming linear regressions, Wu (1973) provided a formal test of exogeneity of X_i often called Hausman-Wu test. It defines a vector of contrasts, $d = b_{OLS} - b_{IV}$, between ordinary least squares (OLS), an efficient but potentially inconsistent (due to endogeneity) estimator and inefficient but consistent (by assumption)

IV estimator. The covariance matrix of d can be shown to be $V_d = V(b_{IV}) - V(b_{OLS})$, and a quadratic form, $d'(V_d)^{-1}d$, is asymptotically a $\chi^2(p)$, with p degrees of freedom. The Hausman-Wu test amounts to medieval diagnosing of a disease (endogeneity) by showing that a cure (b_{IV}) works.

Koopmans (1950) test checks whether exogenous variables "approximately cause" the endogenous variables, i.e., whether the causal path $X_i \to Y$ holds. The underlying concept is same as in modern texts such as (Davidson and MacKinnon, 2004, p. 89) stating that the data generating process (DGP) generating X_i should be independent of Y manifest through the randomness of ϵ_1.

New Test Compares Flipped Models

Now consider a model obtained by flipping Y and X_i

$$X_i = f(Y, X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots X_p) + \epsilon_2. \tag{2}$$

which assumes the approximate path $Y \to X_i$. Engle et al. (1983) assume $p = 1$ and that f is a linear function to prove that both flipped models have identical $R^2 = r^2_{xy}$ values, where r_{xy} is the correlation coefficient. Therefore, these authors argued that Koopmans' approximate causality criterion is "ambiguous" without offering a practical alternative. This paper demonstrates that the alleged ambiguity is due to linearity and readily avoided in modern computing environment by extending Vinod (2015b).

Urgency of Replacing the Hausman-Wu test

Many authors including Bound et al. (1993) and Kiviet and Niemczyk (2007), have warned that in finite samples IV estimators "have systematic estima-

tion errors too, and may even have no finite moments." Moreover they can be very inefficient (even in large samples) and unnecessarily change the original specification. This paper is motivated by the following disadvantages of Hausman-Wu tests:

1. One must replace X_i with *ad hoc*, potentially weak and/or irrelevant instrumental variable before testing for its exogeneity.

2. The test needs to be repeated for each choice of IV to replace X_i.

3. Davidson and MacKinnon (2004) show that degrees of freedom p for the $\chi^2(p)$ test can be inappropriate when all X_i are not endogenous.

4. The Chi-square sampling distribution is subject to unverified assumptions of linearity and normality, especially unrealistic in finite samples.

Retaining $p = 1$ and relaxing linearity, consider a general nonlinear nonparametric kernel regression Model 1:

$$Y_t = G_1(X_t) + \epsilon_{1t}, \quad t = 1, \ldots, T, \tag{3}$$

where errors are no longer Normal and independent. Our nonparametric estimate $g_1(x)$ of the population conditional mean function $G_1(x)$ is:

$$g_1(x) = \frac{\sum_{t=1}^{T} Y_t K(\frac{X_t - x}{h})}{\sum_{t=1}^{T} K(\frac{X_t - x}{h})}, \tag{4}$$

where $K(.)$ is the well known Gaussian kernel function and h is the bandwidth parameter often chosen by leave-one-out cross validation, Li and Racine (2007) and (Vinod, 2008, Sec. 8.4).

Kernel Regressions in generalCorr package

It is well known that kernel regression fits are superior to OLS. The flipped kernel regression Model 2, obtained by interchanging X and Y in eq. (3), is:

$$X_t = G_2(Y_t) + \epsilon_{2t}, \quad t = 1, \ldots, T. \tag{5}$$

The generalized measures of correlation defined by eq. (2) in Zheng et al. (2012) are:

$$[GMC(Y|X), GMC[X|Y] = \left[[1 - \frac{E(Y - E(Y|X))^2}{var(Y)}], [1 - \frac{E(X - E(X|Y))^2}{var(X)}] \right], \tag{6}$$

which are computed simply as the R^2 values of flipped Models 1 and 2. Since they generally do differ from each other, the ambiguity in Koopmans' method mentioned above is removed.

As measures of correlation the non-negative GMC's in the range [0,1] provide no information regarding the up or down overall direction of the relation between Y and X, revealed by the sign of r_{xy}, the Pearson coefficient. Since a true generalization of of r_{xy} should not provide less information, Vinod (2014) and Vinod (2015a) propose the following modification. A general asymmetric correlation coefficient from the $GMC(Y|X)$ is:

$$r^*_{y|x} = \text{sign}(r_{xy})\sqrt{GMC(Y|X)}, \tag{7}$$

where $-1 \leq r^*_{y|x} \leq 1$. A matrix of generalized correlation coefficients denoted by R^* is asymmetric: $r^*_{x|y} \neq r^*_{y|x}$, as desired. A function in the generalCorr package, `gmcmtx0`, provides the R^* matrix from a matrix of data.

Our new test of exogeneity uses the "preponderance of evidence" standard quantified by a comprehensive index, which is a weighted sum of causal direction signs using three criteria Cr1 to Cr3. Our Cr3 which compares R^2 of flipped models is from Vinod (2014). Since elementary statistics teaches us not to rely on R^2 alone, an additional criterion (Cr1) considers evidence from probability distributions of the absolute values of gradients by using stochastic dominance (SD). Similarly our second criterion Cr2 compares absolute residuals.

An outline of the remaining paper is as follows. Section 2 provides an operational definition of kernel causality including our assumptions, definitions, a description of our 'sum' criterion incorporating Cr1 to Cr3, and decision rules explained with a simulation. Section 3 considers statistical inference using the bootstrap. Section 4 considers examples with a subection 4.1 for the famous Klein I model and 4.2 considers what macroeconomic variables drive (cause) excess bond premium known to be a good predictor of recessions. All examples include bootstrap inference for the new test. Section 5 contains a summary and final remarks.

2 Kernel Causality Explained

Assessing philosophically true causality from non-experimental data is nontrivial, Pearl (2009). Instead, we define a modified causality, called kernel causality which holds only under certain assumptions, and where the name kernel causality acknowledges that all our criteria rely on nonlinear nonparametric kernel regressions. We emphasize that Kernel causality has almost

nothing to do with Granger causality typically involving *linear* time series regressions.

Kernel Causality Assumptions:

Our assumptions are:

(A1) Assume that a DGP consists of (X, Y, Z), three sets of variables with main focus on dependence (causal) links between X and Y with Z representing additional (confounding or control) variable(s), if any.

(A2) There exists a conditional expectation function $E(Y|X, Z)$ for Model 1 and analogous function $E(X|Y, Z)$ for Model 2 obtained by flipping X and Y.

(A3) Model 1 DGP is such that X is independently generated (or exogenous) and the dependence of Y on X can be *nonlinear* and subject to *nonnormal* random noise. Model 2 data generation is identical, except for flipped X and Y.

(A4) It is possible to compare whether Model 1 or Model 2 is better supported by the data by using quantifiable empirical criteria.

Note that we are assuming away functional relations such as Boyle's law (pressure *volume = a constant) because it fails A1 and A3: (i) It fails A1 because one does not typically focus on knowing whether pressure causes volume or vice versa. (ii) It fails A3 because both pressure and volume can be independently generated in a typical laboratory.

If a majority of Cr1 to Cr3 support the causal path ($X \to Y$), assumptions A1 to A4 guarantee that X is exogenous (independently generated) and *kernel* causes Y. We begin with two digressions: (i) stochastic dominance, needed for Cr1 and Cr2, and (ii) partial correlations needed for Cr3.

Digression 1: Stochastic Dominance Notation

Let us describe stochastic dominance (SD) concepts surveyed in Levy (1992) without attempting to summarize the vast and growing published and unpublished literature motivated by financial economists' portfolio choice problem. We say that one density $f(x)$ dominates another density $f(y)$ in the first order (SD1) if their respective empirical cumulative distribution functions (ecdf) satisfy: $F(x) \leq F(y)$. It is well known that SD1 provides a comprehensive picture of the ranking between two probability distributions with a focus on locally defined first moment (mean).

The underlying computation requires bringing the two densities on a common 'support,' requiring ecdf's to have up to $2T$ possible jumps or steps. Hence there are $2T$ estimates of $F(x) - F(y)$ denoted by a $2T \times 1$ vector (sd1). Anderson (1996) shows how a simple pre-multiplication by a large patterned matrix implements computation of (sd1). Let us use a simple average Av(sd1) whose sign $(+1, 0, -1)$ helps summarize the first order stochastic dominance into only one number.

Second order dominance (SD2) of $f(x)$ over $f(y)$ requires further integrals of ecdf's to satisfy: $\int F(x) \leq \int F(y)$. One computes the numerical integral by using the trapezoidal rule described in terms of a large patterned matrix whose details are given in (Vinod, 2008, ch.4) and in Anderson (1996). The

$2T$ estimates of SD2 denoted by (sd2) are locally defined variances. Their simple average is denoted as Av(sd2), whose sign $(+1, 0, -1)$ summarizes the information regarding second order dominance.

Similarly, SD of order 3 is estimated by a vector (sd3) of $2T$ locally defined skewness values defined from $\int\int F(x) \leq \int\int F(y)$. The sd3 is further summarized by the sign of Av(sd3). Analogous SD of order 4 for kurtosis requires $\int\int\int F(x) \leq \int\int\int F(y)$ and measures investor 'prudence' according to Vinod (2004). Average of pointwise kurtosis estimates of SD4 are Av(sd4), whose sign $(+1, 0, -1)$ summarizes the SD4 dominance information.

Remark 1: By analogy with two streams of investment returns, stochastic dominance allows us to study realistic but fuzzy inequalities (may not hold for subsets of points) of the type $(x_t < y_t)$ for $t = 1, \ldots, T$. Stochastic dominance of four orders associated with the four moments yield $2T$ estimates of sd1 to sd4. The signs of their averages, Av(sd1) to Av(sd4), indicate whether the inequality holds true in an overall sense.

Digression 2: Partial Correlations

Consider a general matrix A with elements a_{ij}. The minor A_{ij} of A is obtained by deleting i-th row and j-th column. The cofactor of A is a signed determinant, (Vinod, 2011, Sec. 6.1), defined as: $(-1)^{(i+j)}|A_{ij}|$.

Having defined cofactors, we are ready to use the matrix of generalized correlation coefficients R^* to define generalized *partial* correlation between (X_1, X_2) after removing the effect of control or confounding variables

$(X_3, \ldots, X_p)$ as:

$$r^*_{12;3\ldots p} = \frac{R^*_{21}}{\sqrt{R^*_{11}R^*_{22}}}, \tag{8}$$

where R^*_{ij} is the cofactor of R^*.

Since the minor obtained by deleting first row and second column is distinct from one obtained by deleting second row and first column, $R^*_{21} \neq R^*_{21}$. Hence the numerator cofactor for R^*_{21} will differ from the numerator cofactor for R^*_{21}. Thus, $r^*_{12;3\ldots p} \neq r^*_{21;3\ldots p}$, our partial correlations are asymmetric.

In particular, when $p = 3$ we have a new formula:

$$r^*_{12;3} = \frac{r^*_{12} - r^*_{13}r^*_{32}}{\sqrt{(1 - r^*_{13}r^*_{31})}\sqrt{(1 - r^*_{23}r^*_{32})}}, \tag{9}$$

which is similar to but not the same as the well known partial correlation coefficient formula from textbooks.

2.1 Kernel Causality from Flipped Model Choice

We determine whether X drives Y, or vice versa by considering the evidence from the majority of three criteria. They are:

(Cr1) The path $X \to Y$ should be more successful than $Y \to X$ in minimizing absolute values of local kernel regression gradients evaluated at X_t, Y_t, respectively, for $t = 1, 2, \ldots, T$:

$$|\partial g_1(Y_t|X_t, Z_t)/\partial X_t|_{(X_t)} < |\partial g_2(X_t|Y_t, Z_t)/\partial Y_t|_{(Y_t)}. \tag{10}$$

(Cr2) The path $X \to Y$ should have "smaller" absolute residuals than those of the flipped model, that is, for $t = 1, 2, \ldots, T$:

$$|Y_t - g_1(X_t, Z_t)| = (|\hat{\epsilon_{1t}}|) < |X_t - g_2(Y_t, Z_t)| = (|\hat{\epsilon_{2t}}|). \tag{11}$$

(Cr3) The fit (and forecasts) implied by the path $X \to Y$ should have a larger $R^2 = GMC(Y|X, Z)$ than those of reversed path:

$$|r^*_{(y|x;\, z)}| > |r^*_{(x|y;\, z)}|, \tag{12}$$

where generalized partial correlation coefficients defined in eq. (8) remove the effect of control variable(s), if any.

The inequalities of equations (10) and (11) are fuzzy, requiring stochastic dominance tools summarized in Remark 1 above. Let us begin with some definitions.

Definition 1: According to Legal Information Institute (2017) the preponderance of evidence means a burden to show that greater than 50% of evidence points to something.

Definition 2: Assuming A1 to A4, we say that X is the kernel cause of Y (causal path: $X \to Y$), if at least two of Cr1 to Cr3 criteria satisfying the preponderance of evidence standard support the path.

Definition 3: Bidirectional causality ($X \leftrightarrow Y$) or causality marred by the presence of confounding variable(s) occurs if the evidence does not support either ($X \to Y$) or ($Y \to X$).

Remark 2: If relations are strictly linear and/or the errors are precisely normally distributed, flipped R^2 are almost identical creating an ambiguity of Koopmans' criterion criticized by Engle et al. (1983). Since we are using kernel regressions, not OLS, this problem obviously disappears, especially in light of assumption A3 and stochastic dominance for Cr1 and Cr2 unrelated to normality or linearity.

2.2 Weighted sum index from Cr1, Cr2 and Cr3

Applying Remark 1 to the inequality (10) for Cr1, we compute Av(sdℓ) for $\ell = 1, \ldots 4$, magnitudes from absolute gradients of two flipped models. Define a tolerance constant $\tau = 0.01$, say. If $|Av(sd\ell)| < \tau$, we say that the sign is ambiguous, denoted as zero for the ℓ-th SD. When $|Av(sd\ell)| > \tau$, only the signs of Av(sdℓ) not their magnitudes matter. These signs (sg) from the set $(+1, 0, -1)$, are denoted as $sg_{1\ell}$, where the first subscript 1 refers to Cr1. In practice, the signs sg_{11} to sg_{14} are rarely distinct.

Since it is cumbersome to track four signs, we propose a weighted sum, using the signs, $(+1, 0, -1)$, not magnitudes of Av(sd1) to Av(sd4). Statistical theory suggests that weights on magnitudes should be inversely proportional to the increasing sampling variances of the first four central moments. We choose the following weakly declining weights: (1.2/4, 1.1/4, 1.05/4, 1/4), with an option to change them in the R functions `silentPairs` and `causeSummary` of the 'generalCorr package.

Denote a summary sign index based on Cr1 as sC_1. It is computed as:

$$sC_1 = [1.2 * sg_{11} + 1.1 * sg_{12} + 1.05 * sg_{13} + sg_{14}]/4. \tag{13}$$

When all four (Av(sd1) to Av(sd4)) suggest the same sign, ie, all are (± 1), the largest magnitude of our weighted index of sign by Cr1 is $sC_1 = \pm 1.0875$.

Analogous signs $(+1, 0, -1)$ of Av(sd1) to Av(sd4) representing absolute residuals help define their weighted sum for Cr2 is

$$sC_2 = [1.2 * sg_{21} + 1.1 * sg_{22} + 1.05 * sg_{23} + sg_{24}]/4. \tag{14}$$

As before, if all four dominance measures suggest the same sign, the largest

magnitude of sC_2 is 1.0875. Hence, the sign index based on Cr2 lies in the closed interval: $sC_2 \in [-1.0875, 1.0875]$.

The computation of a Cr3 from the inequality test of (12) states that $X \to Y$ if the sign defined as: $sg_3 = (+1, 0, -1)$ of the absolute difference between flipped partial correlations equals (-1). We denote the sign index based on Cr3 as:

$$sC_3 = sign(|r^*_{(x|y;\,z)}| - |r^*_{(y|x;\,z)}|) \tag{15}$$

where the largest score, $\max(sg_3)= 1$. When $sg_3 < 0$, the causal path by Cr3 is $X \to Y$. Note that index always lies in the closed interval: $sC_3 \in [-1, 1]$.

So far, we have three sign indexes (sC_1, sC_2, sC_3) for the three criteria, summarizing the evidence supporting the causal path: $X \to Y$. Since our definition of kernel causality requires us to consider all three criteria, we compute their 'sum' defined as:

$$sum = sC_1 + sC_2 + sC_3, \tag{16}$$

from the observed sample data. Let us denote the corresponding true unknown population value with upper case letters as 'SUM'. When $(SUM < 0)$ holds, the causal path is $X \to Y$. Based on the preponderance of evidence, the sign of *sum* suggests the direction of the path, while its magnitude approximates the strength of sample evidence in support of that causal path.

Combining the three largest possible scores verify that: $\max(sum) = 3.175$, and $sum \in [-3.175, 3.175]$, a closed interval. A summary index is defined as $100(sum/3.175)$ in the range [–100,100]. Since the 'sum' and 'summary index' measure the extent of agreement among the three criteria,

its magnitude is a reasonable indicator of the strength (or unanimity) of evidence for a particular causal path.

Single number summarizing Cr1 to Cr3

The R command `causeSummary(mtx,ctrl=Z, nam=colnames(mtx))` requires a data matrix with p columns called 'mtx' with the first column for the dependent variable and remaining column(s) for regressors. The order of columns is very important. For example, `mtx=cbind(x1,x2,x3)`, where the matrix 'mtx' has three columns, denoted as $p = 3$. Our flipped models fix the first column x1 and pair it with either x2 or x3 for flipping. We do not pair x2 with x3. Thus we always have $p - 1$ possible flipped pairs. The code indicates an error if $p < 2$ or if it is not a matrix. Sometimes one needs to use `as.matrix(mtx)`. Note that control variables are a separate argument (not within mtx), as in: `causeSummary(mtx, ctrl=0)`, where the default value zero means absence of control variables.

The output of 'causeSummary' is self-explanatory based on 'preponderance of evidence' from a weighted combination of Cr1 to Cr3. Since we have exactly $(p-1)$ possible causal path pairs, the summary reports each printed to the screen. For each pair it reports the name of the causal variable, then the name of the response variable, the strength index in terms of unanimity of the sign of the reported causal path. It also reports Pearson correlation coefficient and its p-value for testing the null hypothesis: $\rho = 0$. If the strength is close to zero, in the range $[-5, 5]$, one should conclude that $X \leftrightarrow Y$, even though the computer output wrongly picks one of the two paths.

The code `su=causeSummary(mtx);xtable(su)` may be used to create a

Latex table of results from the output of the function. It is a matrix of $(p-1)$ rows and 5 columns providing summary of pair-wise causal path results. The first column entitled 'cause' names the causal variable, while the second column entitled 'response' names the response. The third column entitled 'strength' has absolute value of summary strength index, printed above but now in the positive range [0,100], summarizing preponderance of evidence from Cr1 to Cr3 from four orders of stochastic dominance and generalized partial correlations. The fourth column entitled 'corr' has Pearson correlation coefficient while the fifth column entitled 'p-value' is for testing the null of zero Pearson correlation coefficient.

Our notion of causality is not the true philosophical causality, but an approximation where a 'kernel cause' is simply the variable which is generated independently. That is, its innovations are self-generated. The dependent variable or the response variable responds to the innovations of the other variable in the flipped pair. This notion of causality allows us to create the causal and dependent variable pairs for the purpose of a simulation. After considering such a simulation in the next subsection, we discuss bootstrap statistical inference using the bootstrap proportion $P^*(\pm 1)$ of occurrences of positive or negative signs in Section 3, further illustrated in our examples later.

2.3 Simulation for checking decision rules

The simulation generates the X variable independently and then define Y to depend on X after adding a noise term, $\epsilon \sim N(0,1)$, a the standard normal deviate. Here the causal path is known to be $X \to Y$, by construction. Our

sample size is $n = 100$ and our `mtx=cbind(X,Y)` enters X as the first column implying that the correct signs are positive.

Let m denote the count for indeterminate signs when we repeat the experiments $N = 1000$ times. Define the success probability for each experiment as:

$$(succ.prob) = \frac{(count\ of\ correct\ signs)}{N - m}. \tag{17}$$

1. Time regressor:

   $X = \{1, 2, 3, \ldots, n\}$

   $Y = 3 + 4X + \epsilon$

2. Uniform Quadratic:

   X has n uniform random numbers

   $Y = 3 + 4X - 3X^2 + \epsilon$

3. Two Uniforms:

   $X1, X2$ each have n uniform random numbers

   $Y = 3 + 4X1 + 3X2 + \epsilon$

4. Three Uniforms:

   $X1, X2, X3$ each have n uniform random numbers

   $Y = 3 + 4X1 + 5X2 - 6X3 + \epsilon$

The large success rate reported on the last row of Table 1 for the experiments shows that our decision rules using a 'sum' from Cr1 to Cr3 work well. Thus, our procedure using flipped models to identify independently generated (causal) variables is supported by the simulation.

Table 1: Summary statistics for results of using the 'sum' measure for correct identification of causal path indicated by its positive sign using N=1000 repetitions, n=100 sample size.

	Expm=1	Expm=2	Expm=3	Expm=4
Min.	1.000	-3.175	-3.175	-3.175
1st Qu.	1.500	1.000	1.000	1.175
Median	2.575	1.000	1.175	3.175
Mean	2.373	1.084	1.658	2.110
3rd Qu.	3.175	1.175	3.175	3.175
Max.	3.175	3.175	3.175	3.175
succ.prob	1.000	0.905	0.882	0.970

3 A Bootstrap Exogeneity Test

Statistical inference regarding causal paths and exogeneity uses the 'sum' statistic defined in equation (16) for estimating the parameter 'SUM' mentioned before.

What is the sampling distribution of the 'sum' test statistic? We use the maximum entropy bootstrap (meboot) described in Vinod and López-de-Lacalle (2009) because it retains the dependence structure (e.g. rankings of countries) in the data recently supported by simulations in Yalta (2016), Vinod (2015b) and elsewhere. Here we use meboot to compute a large number ($J = 999$) of resamples of (X, Y, Z) data. These are an approximation to what the data might look like due to random variation in the population, or the ensemble. The observed (X, Y, Z) data represent only one realization from the ensemble. One can, of course, use other bootstrap algorithms.

Recall that sC_1 to sC_3 is a weighted sum of only three numbers (–1, 0, +1), implying an ordered categorical random variable. Since their *sum* defined in equation (16) can have only a finite set of values, the sampling distribution of the *sum* statistic has nonzero mass only at those set of points in the closed interval:

$$sum \in [-3.175, 3.175]. \tag{18}$$

Since computing the *sum* automatically cancels positive numbers with negative numbers, its magnitude measures a weighted vote count, as it were, in favor of the most enduring (empirically supported) sign of the *sum*. If, for example, $sum = -3.175$, reaching the lower limit of the range, Cr1 to Cr3 are unanimity supporting the causal path $X \to Y$.

Let sum_j denote the j-th bootstrap *sum* where $j = 1, \ldots J$, for each flipped pair. A direct study of the properties of the sampling distribution looks at the summary statistics of the J replicates sum_j, such as: (mean, median, quartiles), etc. The signs of these summary statistics reveal the most preponderant sign in the bootstrap approximation to their population, illustrated later in Table 2 below. The sign of the mode (most frequently observed sum_j) is also of interest.

A further summary of the sampling distribution can be obtained by computing bootstrap proportion of positive or negative values:

$$P^*(+1) = \#(sum_j > 0.5)/J, \quad \text{and} \quad P^*(-1) = \#(sum_j < -0.5)/J, \tag{19}$$

where $\#(sum_j > 0)$ denotes the number of occurrences of positive signs out of J computations while ignoring the magnitudes. Thus $P^*(\pm 1)$ is a

bootstrap approximation to the probability of a positive or negative sign in determining the causal path direction.

In the context of simultaneous equation models, consider the null hypothesis that X_j of eq. (1) is exogenous. Then the path implied by eq. (1) should have greater support than (2). We expect the preponderance of evidence supporting a negative 'SUM'.

Define the null and alternative hypotheses for exogeneity as:

$$H_0 : SUM \leq 0, \quad \text{against} \quad H_1 : SUM > 0, \tag{20}$$

Negative values of SUM are desirable, if we want to assure ourselves that the regressor is exogenous. A simple rule for statistical inference is to *reject* the hypothesized exogeneity whenever the bootstrap proportion $P^*(+1)$ sufficiently exceeds $P^*(-1)$ for the problem at hand. The Definition 1 suggests preponderance of evidence or $> 50\%$ standard. In our experience and illustrations below a much larger percentage is often attainable.

4 Application Examples

Let us begin with an example where the cause is known to illustrate our statistical inference using the *sum* statistic. Vinod (2015a) describes a cross section data example where Y denotes the number of police officers per 1000 population, and X denotes the number of crimes per 1000 population in T =29 European countries in 2008.

```
require(generalCorr);require(Hmisc)
attach(EuroCrime)#bring package data into memory
```

```
causeSummary(cbind(crim,off))
pcause(crim,off,n999=29)
```

The output of above code below shows that crime causes officer deployment with strength 100, while bootstrap resampling success proportion is about 0.59.

```
causeSummary(cbind(crim,off))
[1] crim      causes    off       strength= 100
[1] corr=  0.99   p-val= 0
     cause  response strength corr.  p-value
[1,] "crim" "off"    "100"    "0.99" "0"
pcause(crim,off,n999=29) #illustrative bootstrap
[1] 0.5862069
```

A single bootstrap computation for these data when $J = 999$ on a home PC requires about 20 minutes of CPU time. An approximate sampling distribution of 'sum' statistic for these data is depicted in Figure 1. We are using a histogram because the sampling distribution is categorical with nonzero frequency counts only at a finite set of points. The mode is clearly seen at –3.175 in the histogram. suggesting that the path (crime→officer deployment) is not due to random noise, but likely to be present in the population. The descriptive statistics for the set of J values of (sum_j) are: (first quartile=–3.175, median =–1.175, third quartile=1), and proportion of negatives, $P^*(-1) = 0.641$.

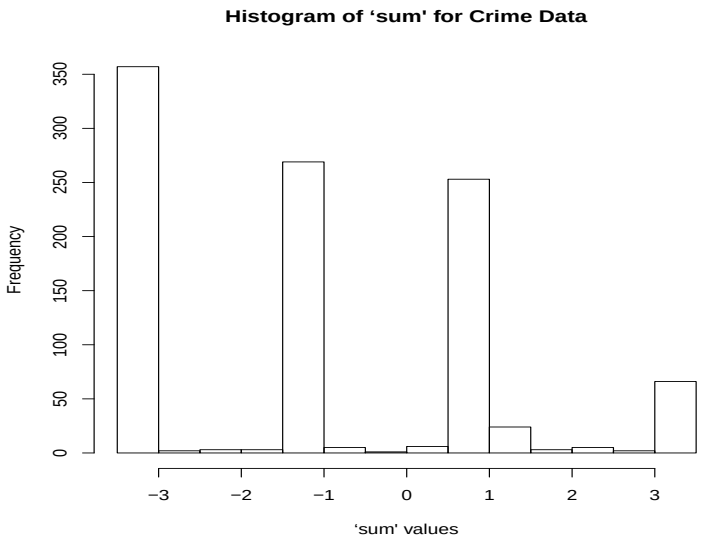

Figure 1: European Crime Data Approximate Sampling Distribution of the *sum* statistic

4.1 Klein I simultaneous equations model

This section reports the results for our three criteria regarding exogeneity of each of the regressors of the three equations of the famous Klein I model. Let us use the following four-character abbreviations using the upper case trailing L for lagged version of a variable: cons=consumption, coPr=corporate profits, coPL= corporate profits with a lag, wage=wages, inve=investment, capL=capital with a lag, prWg=private sector wages, gnpL=GNP with a lag, and finally, tren=time trend.

Klein's specification of the expected consumption equation (stated in terms of fitted coefficients) is:

$$E(\text{cons}) = a_{10} + a_{11}\,\text{coPr} + a_{12}\,\text{coPL} + a_{13}\,\text{wage}. \tag{21}$$

The second (investment) equation of the Klein I model is given by:

$$E(\text{inve}) = a_{20} + a_{21}\,\text{coPr} + a_{22}\,\text{coPL} + a_{23}\,\text{capL}. \tag{22}$$

The third (wage) equation of the Klein I model is given by:

$$E(\text{prWg}) = a_{30} + a_{31}\,\text{gnp} + a_{32}\,\text{gnpL} + a_{33}\,\text{tren}. \tag{23}$$

We report summary statistics for all three criteria combined into the $sum_j, j = 1, \ldots J$ defined in eq. (16) leading to a $J = 999 \times 1$ vector of summary signs, for brevity.

Three columns of Table 2 are for the three equations of the Klein I model. The rows report descriptive statistics: the minimum, maximum, quartiles Q1 and Q3, mean and median based on $J = 999$ bootstrap realizations. The bottom row of the Table reports the bootstrap probability of a positive

Table 2: Klein I model: Bootstrap summary statistics for 'sum' of eq. (16) using 999 resamples to represent the population. A positive mean and median with a large $P^*(+1)$ imply the relevant regressor might not be exogenous.

	cons	inve	prWg
Minimum	-3.1750	-3.1750	-3.1750
1st Quartile, Q1	-1.1750	-1.1750	-1.1750
Median	1.0000	-0.9250	0.0875
Mean	0.4443	-0.1892	0.1874
3rd Quartile, Q3	1.1750	1.1750	1.1750
Maximum	3.1750	3.1750	3.1750
$P^*(+1)$	0.597	0.481	0.504

result, $P^*(+1)$ defined in eq. (19), which are all close to 0.5. The fact that all equations have the same minimum, maximum, Q1 and Q3 show that the bootstrap variability is considerable in both tails making the causal path subject to sampling variability, implying considerable uncertainty in the estimated 'sum.'

The signs of means and medians are both positive in columns 1 and 3 for consumption and private wage equations, implying that wage appears to be endogenous in the consumption equation (21), while gnp may be endogenous in the private wage equation (23). The $P^*(+1) = 0.481 < 0.5$, along with the negative sign of the mean and the median in the second column entitled 'inve' suggests that coPr appears to be exogenous in the investment equation (22).

4.2 Macro Risk Factors for Excess Bond Premium

US Macroeconomists and Federal Reserve researchers have developed new awareness of their failure to forecast the great recession of 2007-2008. Some have developed new data series. For example, Gilchrist and ZakrajŽek (2012) have developed excess bond premium (EBP) and shown that it predicts risk of a recession. It is interesting to find what causes the EBP itself, possibly allowing us to understand why EBP predicts recession risk.

Potential causes are: unemployment rate (UnemR), credit creation (CrCrea, not seasonally adjusted), credit destruction (CrDstr, not seasonally adjusted), yield on 10-year treasury bonds (Yld10, not seasonally adjusted), effective federal funds rate (EffFFR), and money stock (M2, seasonally adjusted billions of dollars). Arguments for using separate variables for CrCrea and CrDstr are found in Contessi and l. Francis (2013) with additional references. We use Federal Reserve's quarterly data from 1973Q1 to 2012Q4, with some data missing. Our software tools can efficiently handle missing data.

We study endogeneity of variables in the following regression model:

$$\text{EBP} = \text{f}(\text{UnemR}, \text{CrCrea}, \text{CrDstr}, \text{Yld10}, \text{EffFFR}, \text{M2}) \tag{24}$$

After getting the data and relevant packages into R memory, we can use the following commands:

```
mtx=cbind(EBP,UnemR,CrCrea, CrDstr,Yld10,EffFFR,M2)
p=NCOL(ntx);print(colnames(mtx)[2:p])
silentPairs(mtx)
```

The output of this shows that only CrCrea, CrDstr and M2 are negative implying that they are exogenous.

```
[1] "UnemR"  "CrCrea" "CrDstr" "Yld10"  "EffFFR" "M2"
[1]  1.000   -1.000   -1.000    3.175     1.000 -1.000
```

The above output of 'sum' index is in the range: $[-3.175, 3.175]$. The results in more intuitive translated range: $[-100, 100]$ plus Pearson correlation and its p-values require simple code:

```
su=causeSummary(mtx);require(xtable)
xtable(su)
```

The results are printed in the following Table 3. Note that only CrCrea, CrDstr and M2 are likely to be independently generated (exogenous) causing the excess bond premium, while the other variables seem to be caused by EBP (endogenous). None of the magnitudes in the column entitled 'strength' is less than 5, implying that we do not have bidirectional paths.

Table 3: Excess Bond Premium and possible causes

	cause	response	strength	corr.	p-value
1	EBP	UnemR	31.496	0.1443	0.0688
2	CrCrea	EBP	31.496	-0.087	0.2739
3	CrDstr	EBP	31.496	0.1998	0.0113
4	EBP	Yld10	100	0.064	0.4216
5	EBP	EffFFR	31.496	0.0657	0.4091
6	M2	EBP	31.496	-0.0103	0.8976

What about sampling variability of strength index? The bootstrap inference is computer time intensive. It requires the function `pcause` as illustrated in the following code.

```
p=NCOL(mtx)
ou2=matrix(NA,nrow=p-1,ncol=2)
for (i in 2:p){
pp=pcause(mtx[,1],mtx[,i],n999=999)
ou2[i-1,1]=colnames(mtx)[i]
ou2[i-1,2]=round(pp,6)  }
print(ou2)
colnames(ou2)=c("variable", "P(-1,0,1)")
xtable(ou2)
```

The printed output of the above code is suppressed for brevity. Instead, our Table 4 shows that sampling distribution results provide a distinct piece of information not covered by the results about the strength or p-value in Table 3.

Graphics on Pair-wise Relations

Pretty scatterplots with locally best fitting lines for each pair of data have now become possible with a nice R package called 'PerformanceAnalytics' by Carl and Peterson (2010) with the function `chart.Correlation` modified for our purposes in the following code.

```
require(PerformanceAnalytics)
chartCorr2=function(mtx,temp="temp",nam=colnames(mtx)){
```

Table 4: Bootstrap success rates for causal direction using 999 resamples

	variable	P(±1)
1	UnemR	0.801802
2	CrCrea	0.927928
3	CrDstr	0.626627
4	Yld10	0.947948
5	EffFFR	0.600601
6	M2	1

```
p=NCOL(mtx)
#print(c("colnames=",nam))
if (p<2) stop("chartCorr2 has input mtx with <2 columns")
nameoplot=nam[2:p]
print(nameoplot)
for (i in 2:p) {
mypath<-file.path("C:",temp,paste(nameoplot[i-1],".pdf",sep=""))
pdf(file=mypath,width=9,height=7)
chart.Correlation(mtx[,c(1,i)])
dev.off()
}# end i loop
}#end function
chartCorr2(mtx)
```

All figures are analogous. Histograms of the two variables is seen in the diagonal panels. The South West panel has a scatter diagram and locally best fitting free hand curve. The number in the North East panel is the

ordinary correlation coefficient whose font size suggests its statistical significance, with stars increasing with 10%, 5% and 1% level. Figures provide visual impressions while the exact correlation coefficients and their p-values are also found in Table 3 with more decimal points.

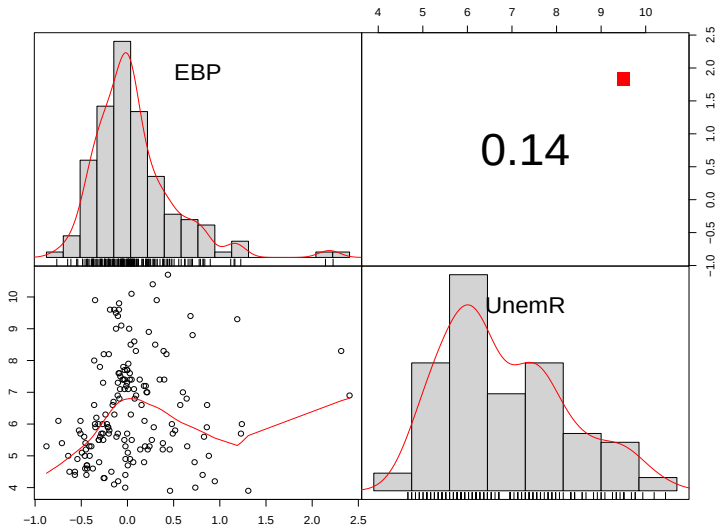

Figure 2: Scatterplot with nonlinear curve: EBP-UnemR

Our evidence including Figure 2 suggests that the variation in UnemR is endogenous, caused by EBP with a scatterplot having a mildly up-down-up pattern.

Our evidence including Figure 3 suggests that the variation in credit creation is exogenous. Its scatterplot is mostly flat and lots of noise.

Our evidence including Figure 4 suggests that the variation in credit destruction is exogenous. This scatterplot is also mostly flat with lots of noise, similar to credit creation.

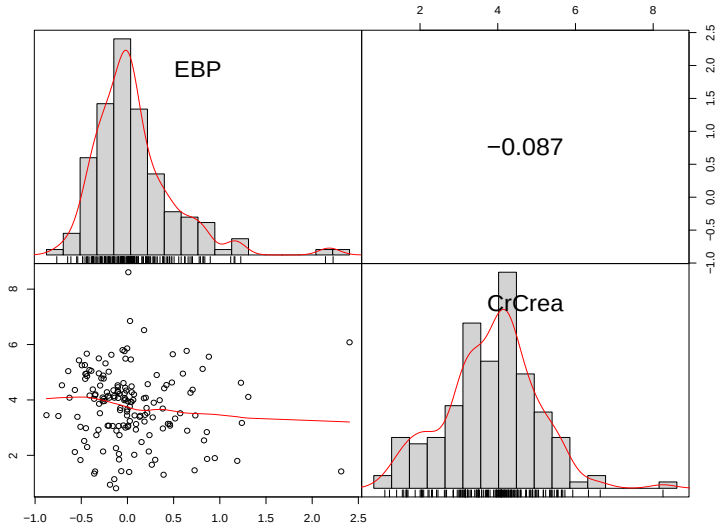

Figure 3: Scatterplot with nonlinear curve: EBP-CrCrea

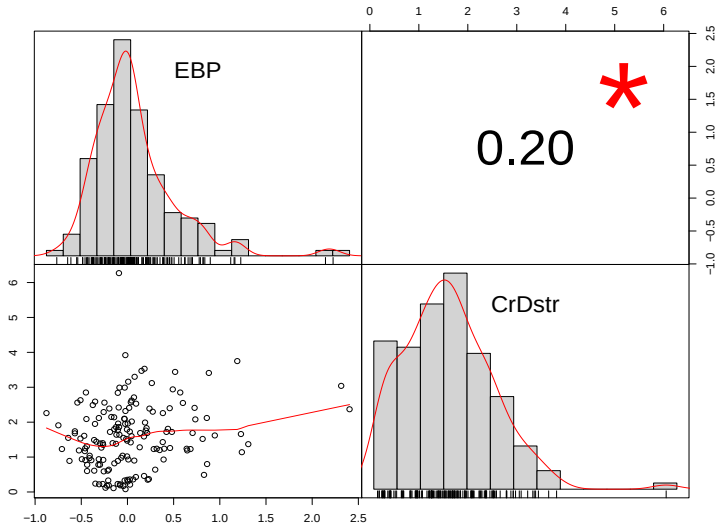

Figure 4: Scatterplot with nonlinear curve: EBP-CrDstr

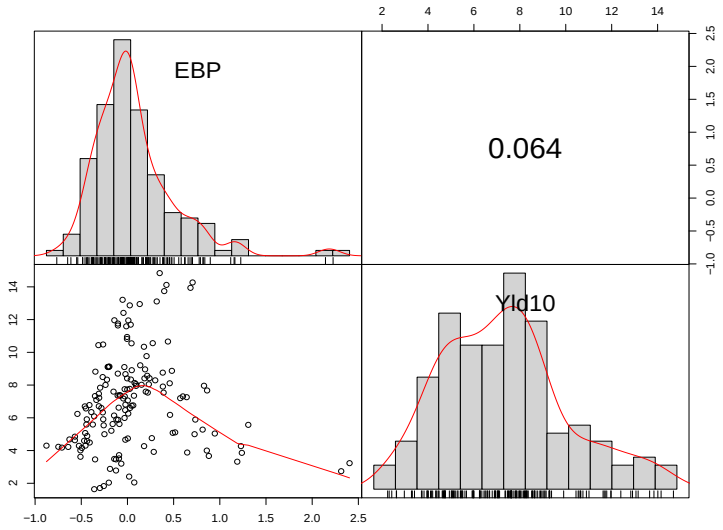

Figure 5: Scatterplot with nonlinear curve: EBP-Yld10

Our evidence including Figure 5 suggests that the variation in the yield on 10-year notes is endogenous, caused by EBP with a scatterplot having a mildly up-down pattern.

Our evidence including Figure 6 suggests that the variation in the effective federal funds rate is endogenous, caused by EBP with a scatterplot having a mildly up-down pattern.

Our evidence including Figure 7 suggests that the variation in money stock M2 is exogenous with a scatterplot having a mildly down-up pattern.

4.3 Airquality data

This example shows how the `causeSummary` function of the package provides reasonable results showing that all meteorological variables are exogenous

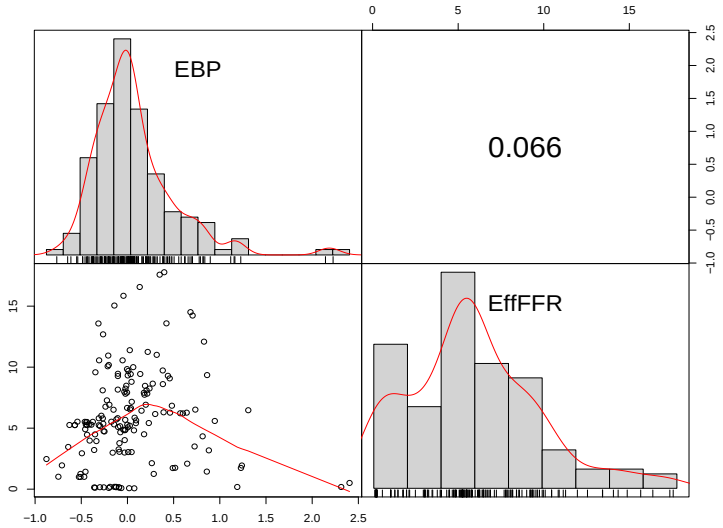

Figure 6: Scatterplot with nonlinear curve: EBP-EffFFR

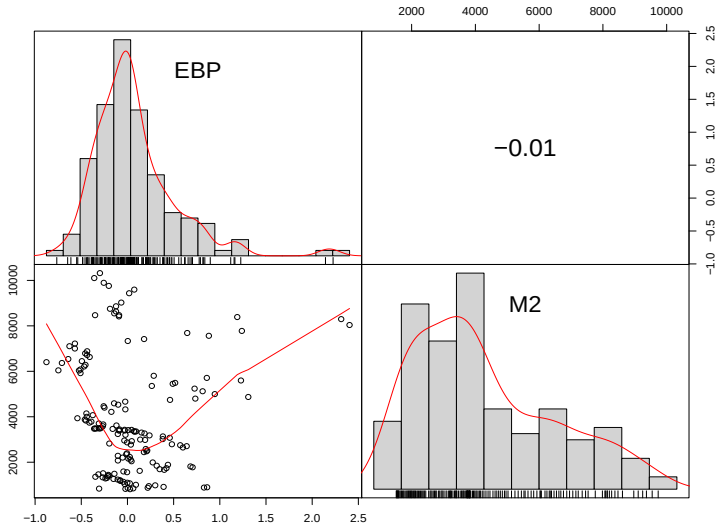

Figure 7: Scatterplot with nonlinear curve: EBP-M2

for Ozone (ppb) air pollution in New York in 1973, using some famous data always available in R.

```
library(generalCorr)
c1=causeSummary(as.matrix(airquality))
library(xtable)
xtable(c1)
```

Table 5: Ozone pollution and its various known causes

	cause	response	strength	corr.	p-value
1	Solar.R	Ozone	100	0.3483	2e-04
2	Wind	Ozone	31.496	-0.6015	0
3	Temp	Ozone	100	0.6984	0
4	Month	Ozone	31.496	0.1645	0.0776
5	Day	Ozone	31.496	-0.0132	0.8879

The results in Table 5 show that solar radiation (lang) and temperature (degrees F) have strongly independent variation, influencing Ozone pollution levels with high strength of 100 for both, suggesting unanimity of Cr1 and Cr2 criteria at all four stochastic dominance levels and further confirmed by Cr3.

Other variables: Wind (mph), month number (1:12) and Day number (1:31) also affect Ozone, but the causal direction is not unanimous. Hence the strength index is only 31.496 for them. Not surprisingly, high wind reduces Ozone pollution is indicated by the significantly negative (–0.6015) Pearson correlation coefficient with a near zero p-value. Additional comments about Table 5 are omitted for brevity.

We use following code to generate a table of bootstrap results.

```
options(np.messages=FALSE)
bb=bootPairs(airquality, n999=999)
ap=apply(bb$out,2,summary)
ap2=rbind(ap,bb$probSign)#P* at the bottom of summary table
xtable(ap2,digits=3)
```

The results are summarized in Table 6, where the 'sum' index is in the range $[-3.175, 3.175]$. We can focus of the means to obtain the overall effect. The bottom row of Table 6 reports the relative frequency of negative values according to the definition (17) implying a success probability in obtaining a negative sign after removing from the denominator all bootstrap estimates m lying in the bidirectional range $[-0.05, 0.05]$. For our example, $m = 0$ for all columns. The bottom line shows that the negative signs in all columns are very reliably estimated. It may be convenient to simply set $m = 0$ in the denominator $(N - m)$, leading to conservative estimates of success rates.

5 Summary and Final Remarks

Medicine has long rejected medieval-style diagnoses of diseases by simply showing that a cure works. Hausman-Wu tests are shown to be similarly flawed as they use IV estimators which can "do more harm than good" (Bound et al., 1995, p. 449), and are criticized as being "very inefficient" by Kiviet and Niemczyk (2007), Dufour, and others. This paper suggests an alternative

Koopmans (1950) suggested that exogenous variables X_i should "approximately cause" the dependent variables Y, but not vice versa. Engle et al.

Table 6: Variability of 'sum' over 999 bootstrap resamples using airquality data

	Solar.R	Wind	Temp	Month	Day
Min.	-3.175	-3.175	-3.175	-3.175	-3.175
1st Qu.	-3.175	-2.575	-1.500	-1.600	-1.000
Median	-3.175	-1.000	-1.175	-1.000	-1.000
Mean	-2.347	-1.539	-1.520	-1.531	-0.957
3rd Qu.	-1.175	-1.000	-1.175	-1.000	-1.000
Max.	1.975	1.175	1.000	-0.500	2.025
$P^*(-1)$	0.9459	0.9299	0.9710	1.0000	0.9760

(1983) correctly show that Koopmans' methods cannot unambiguously identify the causal variable, since two flipped linear regressions (Y on X_i) and (X_i on Y) have the same R^2. We show that modern computing tools and concepts including Zheng et al. (2012) allow us to remove the linearity assumption and focus on Koopmans' valuable insight that exogenous variables should have an independently generated DGP.

We suggest that the endogeneity problem is present in an equation if the left-hand-side variable 'kernel causes' the right-hand-side variable in terms of preponderance of evidence. Hence, we define kernel causality as requiring satisfaction of at least two out of three criteria Cr1 to Cr3. The Cr3 uses 'goodness of fit' when it compares generalized correlation coefficients, suggested in Vinod (2014), such that $|r^*_{y|x}| > |r^*_{x|y}|$ implies that X is the kernel cause of Y. Vinod (2015a) reports favorable simulations using Cr3 alone. Section 2.3 here shows how independently generated (exogenous) variables

are mostly correctly identified by using flipped model performance comparisons based on a summary of Cr1 to Cr3.

Since it is not safe to rely on goodness of fit alone, the other two criteria (Cr1, Cr2) here yield two fuzzy inequalities. The Cr1 involves absolute values of the gradients of kernel regressions and the Cr2 involves absolute values of residuals. Financial economics has long ago developed tools for a comprehensive study of fuzzy inequalities between stock market returns of two competing investment opportunities (e.g., mutual funds) called stochastic dominance of orders 1 to 4 (or SD1 to SD4). See a survey in Levy (1992) and discussion of SD4 in Vinod (2004).

Our sample statistics measuring SD1 to SD4 are called Av(sd1) to Av(sd4) which are further aggregated by using weights inversely related to their sampling variances. Weighted sums quantify the Cr1 and Cr2. Our decision rules based on the 'sum' statistic incorporating all three criteria are simulated in section 2.3 with high success rate.

Our new bootstrap test for exogeneity in section 3 can do statistical inference for the 'sum' statistic, using about a thousand estimates. Descriptive statistics of these estimates, illustrated in Table 2, provide a view of their sampling distribution to assess the preponderant sign and hence the causal direction. If significant endogeneity problem persists, econometricians will, of course, use simultaneous equation models. Koopmans' "departmental principle" gives practitioners some flexibility in designating certain non-economic variables as exogenous, without any need for statistical testing.

We illustrate the new test using the Klein I simultaneous equations model. Our Section 4.2 considers a novel model explaining the 'excess bond premium'

(EBP) known to be a good predictor of US recessions. We study detailed relation between EBP and six variables including various criteria and graphics, providing software tools for implementation based on the R package 'generalCorr.' Our evidence suggests that the variation in three variables: credit creation (CrCrea), credit destruction (CrDstr) and money stock (M2), is exogenous (independenly generated) causing changes in EBP.

Clearly, practitioners can use our tools readily implemented with very few lines of code. It is straightforward to extend and modify our tools, if indicated by future research, since they are open source.

References

Anderson, G., 1996. Nonparametric tests of stochastic dominance in income distributions. Econometrica 64(5), 1183–1193.

Bound, J., Jaeger, D. A., Baker, R., 1993. The Cure Can Be Worse than the Disease: A Cautionary Tale Regarding Instrumental Variables. NBER Working Paper No. 137.
URL `http://ssrn.com/paper=240089`

Bound, J., Jaeger, D. A., Baker, R., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. Journal of the American Statistical Association 90, 443–450.

Carl, P., Peterson, B. G., 2010. PerformanceAnalytics: Econometric tools for

performance and risk analysis.
URL http://CRAN.R-project.org/package=PerformanceAnalytics

Contessi, S., l. Francis, J., 2013. u.s. commercial bank lending through 2008:q4: new evidence from gross credit flows. Economic Inquiry 51(1), 428–444.

Davidson, R., MacKinnon, J. G., 2004. Econometric Theory and Methods. New York: Oxford Univ. Press.

Engle, R. F., Hendry, D. F., Richard, J.-F., 1983. Exogeneity. Econometrica 51, 277–304.

Gilchrist, S., ZakrajŽek, E., 2012. Credit spreads and business cycle fluctuations. American Economic Review 102(4), 1692–1720.

Kiviet, J. F., Niemczyk, J., 2007. The asymptotic and finite-sample distributions of OLS and simple IV in simultaneous equations. Computational Statistics & Data Analysis 51, 3296–3318.

Koopmans, T. C., 1950. When is an equation system complete for statistical purposes. Tech. rep., Yale University.
URL http://cowles.econ.yale.edu/P/cm/m10/m10-17.pdf

Legal Information Institute, 2017. Wex Legal Dictionary. Cornell Law School, Ithaka, NY.
URL https://www.law.cornell.edu/wex/preponderance_of_the_evidence

Levy, H., 1992. Stochastic dominance and expected utility: Survey and analysis. Management Science 38(4), 555–593.

Li, Q., Racine, J. S., 2007. Nonparametric Econometrics. Princeton University Press.

Pearl, J., 2009. Causality: Models, Reasoning and Inference. New York: Wiley.

Vinod, H. D., 2004. Ranking mutual funds using unconventional utility theory and stochastic dominance. Journal of Empirical Finance 11(3), 353–377.

Vinod, H. D., 2008. Hands-on Intermediate Econometrics Using R: Templates for Extending Dozens of Practical Examples. World Scientific, Hackensack, NJ, ISBN 10-981-281-885-5.
URL http://www.worldscibooks.com/economics/6895.html

Vinod, H. D., 2011. Hands-on Matrix Algebra Using R: Active and Motivated Learning with Applications. World Scientific, Hackensack, NJ, ISBN 978-981-4313-68-1.
URL http://www.worldscibooks.com/mathematics/7814.html

Vinod, H. D., 2014. Matrix algebra topics in statistics and economics using R. In: Rao, M. B., Rao, C. R. (Eds.), Handbook of Statistics: Computational Statistics with R. Vol. 34. North Holland, Elsevier Science, New York, Ch. 4, pp. 143–176.

Vinod, H. D., 2015a. Generalized correlation and kernel causality with applications in development economics. Communications in Statistics - Sim-

ulation and ComputationAccepted Nov. 10, 2015.
URL http://dx.doi.org/10.1080/03610918.2015.1122048

Vinod, H. D., 2015b. New bootstrap inference for spurious regression problems. Journal of Applied Statistics.
URL http://www.tandfonline.com/doi/full/10.1080/02664763.2015.1049939

Vinod, H. D., López-de-Lacalle, J., 2009. Maximum entropy bootstrap for time series: The meboot R package. Journal of Statistical Software 29 (5), 1–19.
URL http://www.jstatsoft.org/v29/i05/

Wu, D.-M., 1973. Alternative tests of independence between stochastic regressors and disturbances. Econometrica 77(5), 733–750.

Yalta, A. T., 2016. Bootstrap inference of level relationships in the presence of serially correlated errors: A large scale simulation study and an application in energy demand. Computational Economics 48, 339–366.

Zheng, S., Shi, N.-Z., Zhang, Z., 2012. Generalized measures of correlation for asymmetry, nonlinearity, and beyond. Journal of the American Statistical Association 107, 1239–1252.