Published in final edited form as:

Int J Behav Dev. 2012 July 1; 36(4): 313–322. doi:10.1177/0165025412448944.

Determining Directional Dependency in Causal Associations

Sunthud Pornprasertmanit and Todd D. Little

University of Kansas

Abstract

Directional dependency is a method to determine the likely causal direction of effect between two variables. This article aims to critique and improve upon the use of directional dependency as a technique to infer causal associations. We comment on several issues raised by von Eye and DeShon (2012), including: encouraging the use of the signs of skewness and excessive kurtosis of both variables, discouraging the use of D'Agostino's K^2 , and encouraging the use of directional dependency to compare variables only within time points. We offer improved steps for determining directional dependency that fix the problems we note. Next, we discuss how to integrate directional dependency into longitudinal data analysis with two variables. We also examine the accuracy of directional dependency evaluations when several regression assumptions are violated. Directional dependency can suggest the direction of a relation if (a) the regression error in population is normal, (b) an unobserved explanatory variable correlates with any variables equal to or less than .2, (c) a curvilinear relation between both variables is not strong (standardized regression coefficient $\leq .2$), (d) there are no bivariate outliers, and (e) both variables are continuous.

Keywords

directional dependency; skewness; excessive kurtosis; unobserved explanatory variable

Determining Directional Dependency in Causal Associations

Directional dependency is a method to determine the likely causal direction of effect between two variables (Dodge & Rousson, 2000). Directional dependency can help to determine which of two correlation variables can be designated as the explanatory variable. Directional dependency is determined by comparing the shapes of the distributions of two continuous random variables. This article aims to provide a brief overview of directional dependency, critique and improve upon the use of the directional dependency tests presented by von Eye and DeShon (2012), and address the circumstances in which tests of directional dependency are appropriate.

This article is divided into several parts. First, we define skewness, kurtosis, and excessive kurtosis. We then show that skewness and *excessive* kurtosis (which is sometime confused with kurtosis) can be used to examine directional dependency (Dodge & Rousson, 2000,

2001; Dodge & Yadegari, 2010). Third, we comment on several issues raised by von Eye and DeShon (2012), including: encouraging the use of the signs of skewness and excessive kurtosis of both variables, discouraging the use of D'Agostino's K^2 , and showing that the test of directional dependency is only appropriate with concurrent data and the test cannot be used to determine the direction of effects over time. We then offer improved steps for determining directional dependency that fix the problems we note. In this regard, we discuss how to integrate directional dependency into longitudinal data analysis with two variables. Finally, we clarify the (in)accuracy of directional dependency tests when assumptions of regression analysis are violated.

Definitions

We begin with definitions that are needed below. *Skewness* (γ_1) measures the degree of asymmetry in the distribution of a random variable; skewness can be defined (Cramér, 1946) as

$$\gamma_1 = \frac{E[(X - \mu)^3]}{\sqrt{(E[(X - \mu)^2])^3}}. \quad (1)$$

where X is the target random variable, E is the expectation function or the average of the argument, μ is the population mean. The skewness of a normal distribution is 0.

Kurtosis (β_2) measures the degree of peakedness of a random variable's distribution; kurtosis can be defined (Cramér, 1946) as

$$\beta_2 = \frac{E[(X - \mu)^4]}{(E[(X - \mu)^2])^2}. \quad (2)$$

All notations are defined in Equation 1 (see Online Appendix A for definitions, derivations, and proofs; crmda.ku.edu/supplementals). The kurtosis of a normal distribution is 3.

Excessive kurtosis (γ_2), on the other hand, is defined as the degree of peakedness from the expected value of the normal distribution (i.e., 3; see Equation 2 and see Cramér, 1946):

$$\gamma_2 = \beta_2 - 3. \quad (3)$$

Therefore, the excessive kurtosis of a normal distribution is 0. Some articles use the term “kurtosis” when they mean “excessive kurtosis.” Differentiating between kurtosis and excessive kurtosis is essential to understand directional dependency and its calculation. Equations 1 and 3 are population-level estimates of skewness and excessive kurtosis.

The unbiased sample estimator of skewness and excessive kurtosis takes a different functional form from their population formulas. The sample estimates of skewness (g_1) and excessive kurtosis (g_2) are (Cramér, 1946)

$$g_1 = \frac{\sqrt{n(n-1)} E[(X-M)^3]}{(n-2) \sqrt{(E[(X-M)^2])^3}}, \quad (4)$$

and,

$$g_2 = \frac{(n+1)}{(n-2)(n-3)} \frac{E[(X-M)^4]}{(E[(X-M)^2])^2} - 3 \frac{(n-1)^2}{(n-2)(n-3)}. \quad (5)$$

where M is the sample mean, E is the expectation (average) function, and n is the sample size.

Directional Dependency Principle

Although a large part of this section is redundant to the proof from Dodge and Rousson (2000, 2001), Muddapur (2003), and Dodge and Yadegari (2010), we repeat this proof for the sake of clarity and because this proof is needed for additional proofs in the following sections.

Dodge and Rousson (2000, 2001) showed that directional dependency can be determined by the proportion of skewness of two variables, X and Y . They compared two models:

$$Y = \beta_{YX} X + e, \quad (6)$$

and,

$$X = \beta_{XY} Y + r, \quad (7)$$

where β_{YX} and β_{XY} are the regression coefficients in the models, and e and r are the respective regression errors. Note that intercepts are excluded to simplify the proof in this section, as well as the following sections; the simplification does not undermine the generality of the proof. Figures 1a and 1b illustrate the models in Equation 6 and 7, respectively. Dodge and Rousson examined which of these two models made more sense in representing the true directional relation in this simple two variable relation. As shown in Online Appendix A (crmda.ku.edu/supplementals), if Equation 6 is true (i.e., X predicting Y) and e is normally distributed, the third power of the correlation coefficient between both variables (ρ_{XY}) is equal to the ratio of the skewness of Y ($\gamma_1(Y)$) and the skewness of X ($\gamma_1(X)$):

$$\rho_{XY}^3 = \frac{\gamma_1(Y)}{\gamma_1(X)}, \quad (8)$$

and the fourth power of the correlation between both variables (Dodge & Yadegari, 2010) is equal to the ratio of the *excessive* kurtosis of Y ($\gamma_2(Y)$) and the *excessive* kurtosis of X ($\gamma_2(X)$):

$$\rho_{XY}^4 = \frac{\gamma_2(Y)}{\gamma_2(X)}. \quad (9)$$

Equations 8 and 9 will only hold if the regression error (e) is normally distributed.

Dodge and Rousson (2000, 2001) proposed that directional dependency can be determined based on the possible range of the cube of a correlation coefficient. Because the cube of a correlation will range between -1 and 1 (i.e., $-1 \leq \rho_{XY}^3 \leq 1$), then, based on Equation 8, the magnitude of the skewness of Y will not be greater than the magnitude of the skewness of X :

$$|\gamma_1(Y)| \leq |\gamma_1(X)|, \quad (10)$$

where $|\dots|$ is the absolute value function. In other words, If X is the predictor and Y is the outcome, then the absolute value of the skewness of Y would be less than the absolute value of the skewness of X . Because the fourth power of a correlation coefficient ranges from 0 to 1, then, based on Equation 9, the magnitude of the excessive kurtosis of Y will not be greater than the magnitude of the excessive kurtosis of X (Dodge & Yadegari, 2010):

$$|\gamma_2(Y)| \leq |\gamma_2(X)| \quad (11)$$

Comments on von Eye and DeShon (2012)

The proofs shown in the previous section lead to seven general comments on von Eye and DeShon (2012). Some comments are applicable to the methods proposed by Dodge and Rousson (2000, 2001) and Dodge and Yadegari (2010) as well.

Using Excessive Kurtosis instead of Kurtosis

The von Eye and DeShon's (2012) paper is not clear how they defined kurtosis. They use β_2 (Equation 2) for the directional dependency and argue against the use of γ_2 (Equation 3) in their discussion section. As shown in Equation 9, the ratio of γ_2 between two variables is equal to the fourth-order of the correlation coefficient instead of the ratio of β_2 . Therefore, γ_2 must be used for the directional dependency determination not β_2 . Although, in von Eye and DeShon, kurtosis is defined as β_2 , which cannot take on any negative values, their kurtosis of inattentive symptoms in Table 3 is negative. Caution, therefore, is warranted before using the kurtosis (or excessive kurtosis) values that they reported.

Accounting for the Sampling Error of the Difference

von Eye and DeShon (2012) acknowledged the sampling error of skewness and excessive kurtosis by proposing that the skewness (or excessive kurtosis) of at least one variable should be significant before determining directional dependency. The sampling error of the difference in skewness (or excessive kurtosis), however, was not account for. Because of the sampling error, the difference in skewness in the population is always different from the difference in skewness in the sample to some extent, especially when sample size is small.

In some situations, the differences in skewness in population and in sample have different signs. For example, the skewness of X can be greater than the skewness of Y in the sample when in fact the skewness of X is less than the skewness of Y in the population. To determine directional dependency, the difference in magnitude of skewness and excessive kurtosis must be significant. The significance of the difference in skewness (or excessive kurtosis) guarantees that the differences in population and sample levels have the same sign. Otherwise, the skewness (or excessive kurtosis) of X may be either greater or less than the skewness (or excessive kurtosis) of Y in the population and the decision is inconclusive.

Accounting for Other Possible Two-Variable Associations

Directional dependency is designed to choose whether the effect is from X to Y (Figure 1a) or from Y to X (Figure 1b). von Eye and DeShon (2012) does not discuss the possibility of a nonrecursive effect (Figure 1c) or an unobserved explanatory variable (Figure 1d). A nonrecursive effect is when X and Y affect each other simultaneously. An unobserved explanatory variable means an extraneous variable, Z , affects both X and Y but X and Y have no causal effect between each other, which would be a spurious effect. The directional dependency technique assumes one direction of effect between two variables with no extraneous variables.

Accounting for the Signs of Skewness and Excessive Kurtosis

Dodge and Rousson (2000, 2001), Dodge and Yadegari (2010), and von Eye and DeShon (2012) focused only on comparing the magnitudes of skewness and excessive kurtosis to determine directional dependency. The signs of skewness and excessive kurtosis are critical as well.

From Equation 8, the correlation coefficient, as well as the cube of the correlation coefficient, is positive if and only if the skewness of X and Y have the *same* sign. The correlation coefficient, as well as the cube of the correlation coefficient, is negative if and only if the skewness of X and Y have *different* signs. If the skewness signs are inconsistent with the sign of the correlation coefficient (e.g., the correlation is positive whereas the signs of skewness of both variables are different), Equation 8 is not satisfied. Thus, neither X (Eq. 6) or Y (Eq. 7) is an explanatory variable (the relation may be nonrecursive or spurious). The excessive kurtosis of X and Y must have the same sign because ρ_{XY}^4 can only be positive (Eq. 9). If the excessive kurtosis values are not the same sign then neither X or Y is an explanatory variable.

Ignoring the signs of skewness and excessive kurtosis can lead to a different conclusion. We illustrate this issue in the von Eye and DeShon's (2012) example of the relation between ADHD symptoms and the blood lead level. The skewness of blood lead level and inattentive symptoms are 2.156 and -0.815 , respectively. With a standard error of 0.198, the blood lead level variable is positively skewed and the inattentive symptom variable is negatively skewed in the population. If the von Eye and DeShon's (2012) guideline is used, blood lead level is suggested as the explanatory variable. Presumably, the correlation between ADHD symptoms and the blood lead level is positive. To support directional dependency, however,

these variables must have the same signs. Because they have different signs, neither variable is supported as the explanatory variable.

We assume that the kurtosis reported in von Eye and DeShon (2012) is excessive kurtosis (g_2), which can take a negative value. The excessive kurtosis of blood lead level and inattentive symptoms are 5.158 and -1.045 , respectively. With a standard error of 0.394, the lead level variable is leptokurtic but the inattentive symptoms variable is platykurtic in the population. If the von Eye and DeShon's (2012) guideline is used, blood lead level is suggested as the explanatory variable. Based on our argument, however, the excessive kurtosis of both variables must have the same sign to determine directional dependency. The signs of excessive kurtosis of both variables, in fact, are different; therefore, the direction dependency of ADHD and blood-lead content cannot be determined. Using our refined guidelines, therefore, a causal conclusion is not supported or warranted in either of these two examples.

Avoiding the Use of D'Agostino's K^2

Our next point is on the use of D'Agostino's K^2 (1971) statistic for directional dependency. Even though the use of this index for testing directional dependency was supported by von Eye and DeShon's (2008) simulation, the use of K^2 can be misleading because the directional dependency of both skewness and excessive kurtosis must each be supported individually. If the directional dependency of skewness indicates that one variable is an explanatory variable, the directional dependency of excessive kurtosis should also indicate that the variable is an explanatory variable or at least that the directional dependency status is undetermined. The directional dependency of excessive kurtosis should not contradict the directional dependency indicated by using skewness. By combining both skewness and excessive kurtosis into a single index, the information of individual directional dependency indices is lost. For example, in the directional dependency of lead level and hyperactive-impulsive symptoms in von Eye and Deshon's (2012), the skewness shows that lead level ($g_1 = 2.156$) is the explanatory variable and the hyperactive-impulsive symptoms ($g_1 = 0.223$) is the response variable (as the correlation between both variables is positive). The excessive kurtosis, however, shows that neither the lead level ($g_1 = 5.158$) nor the hyperactive-impulsive symptoms ($g_2 = -1.568$) are the explanatory variable because the excessive kurtosis of both variables have different signs. Because both conditions were not satisfied, a conclusion of directional dependency is not defensible. Combining skewness and kurtosis into a single index conceals the result from the use of excessive kurtosis that neither variable is the explanatory variable.

The Need to Establish the Two-Variable's Relation

Another important issue in determining directional dependency is that the relation between two variables must be established. von Eye and DeShon (2012) emphasized comparing the degree of nonnormality between both variables; however, the establishing a non-zero relation between two variables needs to be explicitly mentioned as well. One of the central conditions of causal inference is that a non-zero relation must exist between the variables (Cook & Campbell, 1979). We mention this issue because we are afraid that one may pick a pair of unrelated and nonnormal variables (i.e., neither Eq. 6 or 7 is true in the population)

and then compare skewness or excessive kurtosis directly to infer the direction of an effect. This practice will produce Type I errors—inferring a direction of effect when there is no effect. Therefore, a relation between both variables must exist before examining directional dependency.

Avoiding the Use of Directional Dependency in Longitudinal Data

von Eye and DeShon (2012) proposed two scenarios that applied directional dependency within a longitudinal data analysis context, each with multiple time points. In Scenario 1, X is observed at Time 2 only and Y is observed at both Time 1 and Time 2. They proposed that, if X is the explanatory variable that is active on Time 2 only, then (a) Y should have less skewness at Time 2 than at Time 1, and (b) X should have more skewness than Y at Time 2. In Scenario 2, X is observed in Time 1 only and Y is observed at both Time 1 and Time 2. They proposed that if X is the explanatory variable that is active at Time 1 only, then (a) Y should have more skewness at Time 2 than at Time 1, and (b) X should have more skewness than Y at Time 1.

We disagree with the first rule in determining directional dependency from both scenarios. The rule implies that the response variable measured when an explanatory variable is active will be closer to a normal distribution than one measured when the explanatory variable is not active. This argument is not necessarily true. Consider three propositions here:

1. From the directional dependency principle (i.e., Equations 8 and 9), a response variable is not normally distributed when the response variable has a predictor that is not normally distributed.
2. Many psychological variables deviate significantly from a normal distribution (Micceri, 1989).
3. As one of the assumptions of science, every effect (variable) has its cause (Maxwell & Delaney, 2004).

From proposition 3, a variable exists because it has a cause. From proposition 1, a psychological variable is not normally distributed if its cause is not normally distributed. Thus, a psychological variable must have an unobserved explanatory variable that makes them not normally distributed. In terms of mathematical notation, let Y be the psychological variable and Z be its unobserved cause. The regression model is

$$Y = \beta_{Y,Z} Z + e, \quad (12)$$

where β_{YZ} is the regression coefficient of Y on Z , and e is the normally distributed error. If X is active at Time 1 but not at Time 2, the two regression models will be:

$$Y_1 = \beta_{Y,X} X + \beta_{Y,Z,1} Z_1 + e_1, \quad (13)$$

and,

$$Y_2 = \beta_{Y,Z,2} Z_2 + e_2, \quad (14)$$

where Y_1 and Y_2 are the response variable at Time 1 and Time 2, respectively, Z_1 and Z_2 are the response variable's unobserved cause at Time 1 and Time 2, respectively, $\beta_{YZ,1}$ is the regression coefficients of Y on Z at Time 1 controlling for X , $\beta_{YZ,2}$ is the regression coefficient of Y on Z at Time 2, β_{YX} is the coefficient of Y on X controlling for Z_1 , and e_1 and e_2 are the errors.

Assuming that X and Z are independent and e_1 and e_2 are normally distributed, as shown in Online Appendix A (crmda.ku.edu/supplementals), the ratio of the skewness of Y when X is active to the skewness of Y when X is inactive is

$$\frac{\gamma_1(Y_1)}{\gamma_1(Y_2)} = \frac{\beta_{S-YX}^3 \gamma_1(X) + \beta_{S-YZ,2}^3 \gamma_1(Z_1)}{\beta_{S-YZ,2}^3 \gamma_1(Z_2)}. \quad (15)$$

Where $\gamma_1(Y_1)$ and $\gamma_1(Y_2)$ are the skewness of the response variable at Time 1 and Time 2, respectively, $\gamma_1(Z_1)$ and $\gamma_1(Z_2)$ are the skewness of the response variable's unobserved cause at Time 1 and Time 2, respectively, $\gamma_1(X)$ is the skewness of the explanatory variable, $\beta_{S,YZ,1}$ is the standardized regression coefficients of Y on Z at Time 1 controlling for X , $\beta_{S,YZ,2}$ is the standardized regression coefficient of Y on Z at Time 2, and $\beta_{S,YX}$ is the standardized regression coefficient of Y on X controlling for Z_1 .

The key implication of Equation 15 is that the skewness of X can make the ratio of the skewness of Y_1 and Y_2 either greater or less than 1. This possibility means that the skewness of Y at Time 1 (X is active) is not necessarily greater than the skewness of Y at Time 2 (X is inactive). Therefore, the skewness of the response variable should not be compared across time to determine directional dependency.

Improved Rules for Determining Directional Dependency

On the basis of the above, we offer improved rules for determining directional dependency. At the population level, the upper part of Table 1 shows a summary of using skewness and excessive kurtosis for directional dependency tests. The rules at the population and sample levels are different because of sampling error. The decision steps are:

1. If the correlation between X and Y is significant, go to the next step. If not, neither variable is the explanatory variable.
2. If $g_1(X)$ or $g_1(Y)$ is significantly different from 0, go to the next step. If not, both variables are likely normally distributed and directional dependency is undetermined.
3. If the correlation is positive and both $g_1(X)$ and $g_1(Y)$ are not significantly different from 0 in different directions, go to the next step. If the correlation is negative and both $g_1(X)$ and $g_1(Y)$ are not significantly different from 0 in the same directions, go to the next step. Otherwise, both variables' skewness values in the population are inconsistent with the sign of the correlation coefficient so neither variable is an explanatory variable.
4. If $|g_1(X)|$ is significantly different from $|g_1(Y)|$, go to the next step. Otherwise, the directional dependency is undetermined.

5. If $|g_1(X)|$ is significantly greater than $|g_1(Y)|$, then X is the explanatory variable and Y is the response variable. If $|g_1(X)|$ is significantly less than $|g_1(Y)|$, then Y is the explanatory variable and X is the response variable.

The decision steps for comparing excessive kurtosis at the sample level are:

1. If the correlation between X and Y is significant, go to the next step. If not, neither variable is the explanatory variable.
2. If $g_2(X)$ or $g_2(Y)$ is significantly different from 0, go to the next step. If not, both variables are likely normally distributed and directional dependency is undetermined.
3. If both $g_2(X)$ and $g_2(Y)$ are not significantly different from 0 in different directions (e.g., $g_2(X)$ is negative but $g_2(Y)$ is positive), go to the next step. Otherwise, both variables' excessive kurtosis values in the population are in different directions so neither variable is an explanatory variable.
4. If $|g_2(X)|$ is significantly different from $|g_2(Y)|$, go to the next step. Otherwise, the directional dependency is undetermined.
5. If $|g_2(X)|$ is significantly greater than $|g_2(Y)|$, then X is the explanatory variable and Y is the outcome variable. If $|g_2(X)|$ is significantly less than $|g_2(Y)|$, then Y is the explanatory variable and X is the outcome variable.

The lower part of Table 1 shows a summary of using skewness and excessive kurtosis for directional dependency tests at the sample level after a significant correlation between both variables is determined. All significance tests are easily done by using a simple bootstrap estimation method (Efron & Tibshirani, 1993). We provide an R script in Online Appendix B (crmda.ku.edu/supplementals) to run a bootstrap confidence interval for correlation, skewness, excessive kurtosis, and the difference in magnitudes of skewness and excessive kurtosis.

One prospective problem of a step approach is the inflation of Type I error. Because a step approach involves multiple significance testing, the chance that at least one test is a Type I error is greater than the specified alpha level for an individual test. The meaning of Type I error in this situation, however, is different from the multiple comparisons in the analysis of variance. Overall Type I error in this situation, in our opinion, is the probability that the directional dependency test suggests that X , Y , or neither is the explanatory variable when a decision of undetermined should be selected based on the information at the population level. We will discuss this issue in the following section after we show our Monte Carlo simulation results.

Guidelines of Using Directional Dependency in Longitudinal Data

As explained above, the proposed method for directional dependency determination is invalid in comparing between the same (or different) variables across time points. If longitudinal data can be obtained, causal inference is conducted more rigorously than in a cross-sectional analysis because temporal precedence can be supported (i.e., a cause

precedes an effect; Cook & Campbell, 1979). Figure 2 show the list of possible two-variable relationship in a longitudinal design: a) X is explanatory, b) Y is explanatory, c) X and Y have no relationship, d) X and Y have reciprocal relationship, and e) X and Y have spurious relationship. We assume that the causal structure does not change across time (i.e., stationary, Kenny, 1979). If X and Y are observed, we can test Model A, B, C, and D directly by nested model comparison using structural equation modeling (χ^2 difference test). If the observed data comes from Model E and only X and Y are observed, Model D would be selected by the nested model comparison.

The proposed directional-dependency method cannot replace the nested model comparison. This method can be used as additional information in causal inference by comparing different variables *at the same time point*. When Model A is selected by the nested model comparison, the directional dependency should suggest that X is explanatory or the directional dependency is undetermined at time point 2 and again at time point 3. When Model B is selected, the directional dependency should suggest that Y is explanatory or the directional dependency is undetermined. When Model C or D is selected (Model C, D, and E might be true in the population), the directional dependency should suggest that neither variable is explanatory or the directional dependency is undetermined. Thus, directional dependency tests can be used as an additional method for causal inference in a longitudinal design but directional dependency cannot be applied over time.

Robustness of Assumption Violations

Directional dependency is based on regression analysis which has several assumptions. Therefore, the (in)accuracy of directional dependency will be related to the violation of the regression assumptions to some degree. We highlight some important assumptions of basic regression analysis that would decrease the accuracy of any test of directional dependency. Those assumptions are (a) normality, (b) no unobserved explanatory variable exists, (c) the relation between X and Y is linear, (d) no outlier, and (e) the response variable is continuous.

Normality

Regression analysis and directional dependency assume that the error is normally distributed at the population level. If this assumption is not true, Equations 8-11 will not hold. The sample distribution, instead of the population distribution, is usually observed. Because of sampling error, the sample residual distribution need not be exactly normal, although it should be close to normal if it comes from a normal population error. When X or Y is suggested as the explanatory variable, the suggested (explanatory) variable and the other (response) variable can be used to find the sample regression residual. For example, if X is suggested as the explanatory variable, Y is regressed on X and the residual is saved. The residual is checked for normality (Maxwell & Delaney, 2004).

Unobserved Explanatory Variable

Directional dependency only focuses on the effect of one independent variable on one dependent variable. A dependent variable, however, may be caused by more than one variable. Unobserved explanatory variables (i.e., other predictor variables) might lead to

erroneous conclusions in directional dependency. For example, X and Z are explanatory variables; X is a normally distributed and observed while Z is a highly skewed but unobserved. In this situation, Y (the dependent variable) will be skewed because of the skewness of Z . If directional dependency between X and Y is of interest, a researcher would errantly pick Y as the explanatory variable.

We will further investigate the effect of an unobserved explanatory variable by a mathematical proof and a Monte Carlo simulation. The mathematical proof (as shown in Online Appendix A; crmda.ku.edu/supplementals) illustrates how an unobserved covariate affects directional dependency tests. The following Monte Carlo simulation examines the robustness of directional dependency on an unobserved explanatory variable. This simulation study result shows the robustness to violations of other regression assumptions described below as well.

In this simulation, X is an explanatory variable, Y is a response variable, and Z is a covariate. X and Z were independent ($\rho_{xz} = 0$). We only studied the directional dependency determined by skewness under four simulation conditions:

1. Skewness of explanatory variable: Strongly Negative ($\gamma_1 \approx -2$), Negative ($\gamma_1 \approx -1$), Normal ($\gamma_1 \approx -0$), Positive ($\gamma_1 \approx 1$), and Strongly Positive ($\gamma_1 \approx 2$).
2. Skewness of covariate: Strongly Negative, Negative, Normal, Positive, and Strongly Positive.
3. Correlation (standardized regression coefficient) between the response variable and the explanatory variable: .2 and .5.
4. Correlation (standardized regression coefficient) between the response variable and the covariate: 0 (baseline model), .2, .5, and .8.

To create different levels of skewness, we used a χ^2 distribution. The χ^2 distribution has a higher level of skewness if its degree of freedom is low. In the strongly negative and the strongly positive skewness condition we used the χ^2 distribution with 2 degrees of freedom. In the negative and the positive skewness condition we used the χ^2 distribution with 10 degrees of freedom. In the normal condition we used a standard normal distribution. We use a sample size of 1,000 and ran 1000 replications per condition. We used X as the explanatory variable and our proposed method of detection (Online Appendix B contains the R script used; crmda.ku.edu/supplementals). We used skewness (g_1) to determine directional dependency only. A simulation using von Eye and DeShon's (2012) method is available in Online Appendix C (crmda.ku.edu/supplementals).

The simulation study show the results when the correlations between X and Y are .2 and .5, respectively (Tables 2 and 3). The rows of both tables are divided into four panels that represent the correlation between Y and Z (0, .2, .5, and .8, respectively). There are four possible results from the directional dependency test: X is the explanatory variable, Y is the explanatory variable, neither is the explanatory variable, or the directional dependency is

undetermined. Each cell represents the proportion that the directional dependency showed each possible result.

The first panel of rows of both tables is the baseline model without the covariate effect. When the explanatory variable is skewed, the directional dependency can be determined almost 100% correctly. On the other hand, if the explanatory variable is normal, approximately 96% of replications were undetermined, 2% of replications suggested X as an explanatory variable, and 2% of replications suggested Y as an explanatory variable. The similar pattern of result is generalized to the second panel of rows of both tables when the effect of the covariate is small (i.e., the correlation between covariate and response variable is .2).

The third and fourth panels of rows of both tables represent the conditions where the effect of the covariate is medium ($\rho = .5$) and large ($\rho = .8$), respectively. In the normal explanatory variable and the skewed covariate, more than 45% of replications indicated that Y was the explanatory variable (Y was actually the outcome variable). If the effect of the covariate is large (i.e., the fourth panel), the explanatory variable and covariate are both skewed, the test is less likely to select X as explanatory variable and more likely to select “neither” and “undetermined”.

From this simple simulation, we can conclude that the test of directional dependency is robust to the small effect of an unobserved explanatory variable. When the covariate effect is medium or large, the result cannot be trusted because the skewness of a variable can result from either itself or the covariate. Thus, directional dependency can only be trusted if the existence of a nontrivial, unobserved explanatory variable can be ruled out. In other words, the correlations of the covariate with both variables used in directional dependency should not exceed .2.

From this simulation, Type I error can be examined when X is normal and the correlation between X and Z is 0 (the third row of the first panel). In this situation, the decision should be that the directional dependency is undetermined (see the top panel of Table 1). The proportion that the test suggested “undetermined” is .96. That is, the overall Type I error is .04, which is close to the nominal level of .05 even though this step approach involves multiple testing.

Another issue related to unobserved explanatory variables is that the correlation between two variables may be spurious (Figure 1d). Thus, the potential for a spurious correlation must also be ruled out before investigating the directional dependency. Here, if there is no potential effect of a covariate that is greater than a standardized regression coefficient on either variable of .2, the spurious contribution to a correlation will not be greater than .04. This magnitude of correlation is sufficiently trivial to ignore in a test of directional dependency. The existence or not of a trivial or not unobserved variable is similar to the elimination of plausible alternative causes in causal relation inference (Cook & Campbell, 1979). Instead of totally eliminating alternative causes, however, the alternative causes should have effects that are not too strong to compromise the accuracy of directional

dependency. Conducting tests of plausible third variables would be a recommended step to ensure that the directional dependency tests are accurate.

Curvilinear Relation

Dodge and Rousson (2001) indicated that the test of directional dependency is applicable for a linear relation. No effect between two variables, however, is exactly linear in the population (MacCallum & Austin, 2000). There will be curvatures in a relation to some degree. Consider the quadratic effect of a predictor. In multiple regression, a quadratic effect can be modeled as

$$Y = \beta_1 X + \beta_2 X^2 + e, \quad (16)$$

where β_1 is the regression coefficient of Y on X controlling X^2 , β_2 is the coefficient of Y on controlling X , and e is the regression error. Here, X^2 can be considered as a covariate. As implied in the simulation study, the effect size of the quadratic term (standardized regression coefficient) is the most important factor. If the relationship between the quadratic term and the response variable is small (i.e., the correlation between the curvature term and the response variable is not greater than .2), the directional dependency will still be reliable.

Outliers

Outliers can be considered as unobserved explanatory variables. If one observation is an outlier, the outlier can be modeled by a dummy variable (1 = Outlier and 0 = Others), D , as

$$Y = \beta_1 X + \beta_2 D + e, \quad (17)$$

where β_1 is the regression coefficient of Y on X controlling D , β_2 is the coefficient of Y on D controlling X , and e is the regression error, which is normally distributed after accounting for D .

The mathematical proof (as shown in Online Appendix A; crmda.ku.edu/supplementals) and the simulation study in the unobserved explanatory variable section also apply. Specifically, the effect size of the outlier (standardized regression coefficient) is the most important factor. If the correlation between a dummy variable indicating the outlier (D) and the dependent variable (Y) is 0, the effect size of the outlier is also 0 and the directional dependency will not be biased. That is, if the averages of the dependent variable between the outlier and other observations are not different, directional dependency determination is not biased. Therefore, the outlier influence is negligible when the outlier is univariate or in X only. If an outlier is bivariate, the outlier should be excluded (Cohen, Cohen, West, & Aiken, 2003).

Beyond the General Linear Model

The test for directional dependency is based on the general linear model assuming that the response variable is continuous and the relation is linear. The general linear model is not an exhaustive list of all two-variable relations. In fact, the general linear model is a submodel of the generalized linear model with a link function that is 1.0 (identity link). Directional dependency is not applicable for other submodels in the generalized linear model. For

example, Poisson regression is usually used to predict count variables (such as hours of exercise), which is the submodel of the generalized linear model with log link function and a Poisson distribution. In this submodel, a dependent variable may have a higher magnitude of skewness than an explanatory variable (i.e., a Poisson distribution is positively skewed when the rate is low).

In the example of the relation between ADHD and the number of inattentive symptoms, the general linear model (i.e., the traditional multiple regression) may not be appropriate in predicting the number of inattentive symptoms because the number of inattentive symptoms is a count variable ranging from 0 to 9. To predict this variable, Poisson regression (or a generalized linear model for an ordered categorical variable) is more appropriate. The error from either of these models is not normally distributed and comparing skewness or excessive kurtosis of the two variables is inappropriate.

As a result, before using directional dependency, the general linear model must be a legitimate choice for the nature of variables on both sides of the regression equation. If one of two variables is not continuous, the general linear model is appropriate in one direction but not appropriate in the other direction. The examples of inappropriate types of variables are count variables (e.g., number of children in family), unordered categorical variables (e.g., voting behavior), and censored variables (e.g., test score with floor or ceiling effect).

Conclusion

Directional dependency is a method to determine the likely causal direction of effect between two variables (Dodge & Rousson, 2000) whether one variable (e.g., X predicting Y) or the other variable is the explanatory variable (e.g., Y predicting X). In our critique, we noted several critical issues relating to the traditional method of directional dependency and offered critical improvements to von Eye and DeShon (2012), including how to integrate directional dependency tests with longitudinal data. There are several promising future directions. First, although we show that bootstrap approach for directional dependency determination can control Type I error to the nominal overall alpha level, this method still needs further examination of its accuracy and its efficiency. In addition, future research should examine the robustness of directional dependency with multiple unobserved explanatory variables because the multiple explanatory variables may have a cumulative effect on directional dependency tests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Author Note

Sunthud Pornprasertmanit and Todd D. Little, Department of Psychology and Center for Research Methods and Data Analysis, University of Kansas. Partial support for this project was provided by grant NSF 1053160 and by the Center for Research Methods and Data Analysis at the University of Kansas (Todd D. Little, director). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

We thank Paul E. Johnson, Pascal Deboeck, Aaron Boulton, Maneekwan Chandarasorn, and Katharina Jorgensen for their helpful comments on a previous draft of this paper.

References

- Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd. Routledge; New York: 2003.
- Cook, TD.; Campbell, DT. Quasi-experimentation: Design and analysis issues for field settings. Rand McNally; Chicago, IL: 1979.
- Cramér, H. Princeton University Press; Princeton, NJ: 1946.
- D'Agostino RB. An omnibus test of normality for moderate and large sample size. *Biometrika*. 1971; 58:341–348.
- Dodge Y, Rousson V. Direction dependence in a regression line. *Communication in Statistics: Theory and Methods*. 2000; 29:1957–1972.
- Dodge Y, Rousson V. On asymmetric properties of the correlation coefficient in the regression setting. *American Statistician*. 2001; 55:51–54.
- Dodge Y, Yadegari I. On direction of dependence. *Metrika*. 2010; 72:139–150.
- Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman & Hall/CRC; New York, NY: 1993.
- Gut, A. 2nd. Springer; New York, NY: 2009. An intermediate course in probability.
- Joanes DN, Gill CA. Comparing measures of sample skewness and kurtosis. *Statistician*. 1998; 47:183–189.
- Kenny, DA. Correlation and causality. Wiley; New York: 1979.
- Lance CE. Residual centering, explanatory and confirmatory moderator analysis, and decomposition of effects in path models containing interactions. *Applied Psychological Measurement*. 1988; 12:163, 175.
- Little TD, Bovaird JA, Widaman KF. On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling*. 2006; 13:497–519.
- MacCallum RC, Austin JT. Applications of structural equation modeling in psychological research. *Annual Review of Psychology*. 2000; 51:201–226.
- Maxwell, SE.; Delaney, HD. Designing experiments and analyzing data: A model comparison perspective. 2nd. Psychology Press; New York, NY: 2004.
- Micceri T. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*. 1989; 105:156–166.
- Muddapur MV. On directional dependence in a regression line. *Communication in Statistics: Theory and Methods*. 2003; 32:2053–2057.
- R Development Core Team. R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria: 2011. Available from <http://www.R-project.org/>
- von Eye, A.; DeShon, RP. Characteristics of measures of directional dependence: A Monte Carlo study. *InterStat*. Feb. 2008 Retrieved from <http://interstat.statjournals.net/>
- von Eye A, DeShon RP. Directional dependence in developmental research. *International Journal of Behavioral Development*. 2012

Figure 1.

The possible explanations of causal inference within a two-variable relation.

Figure 2.
The possible explanations of causal inference within a two-variable relation in a longitudinal design.

Table 1

The Summary of Using the Skewness or Excessive Kurtosis for Determining Directional Dependency between Two Variables at the Population and Sample Levels

	<i>X</i> is predictor	<i>Y</i> is predictor	Neither	Undetermined
Population Level				
Skewness	$\gamma_1(X) \times \gamma_2(X)$ is in the same direction with ρ_{XY} and $ \gamma_1(Y) \quad \gamma_1(X) $.	$\gamma_1(X) \times \gamma_2(X)$ is in the same direction with ρ_{XY} and $ \gamma_1(X) \quad \gamma_1(Y) $.	$\gamma_1(X) \times \gamma_2(X)$ is in the different direction with ρ_{XY} .	$\gamma_1(X) = \gamma_1(Y) = 0$
Excessive Kurtosis	$\gamma_2(X)$ and $\gamma_2(Y)$ are in the same directions and $ \gamma_2(Y) \quad \gamma_2(X) $.	$\gamma_2(X)$ and $\gamma_2(Y)$ are in the same directions and $ \gamma_2(X) \quad \gamma_2(Y) $.	$\gamma_2(X)$ and $\gamma_2(Y)$ are in different directions.	$\gamma_2(X) = \gamma_2(Y) = 0$
Sample Level				
Skewness	$ g_1(Y) $ is significantly less than $ g_1(X) $, and $g_1(X)$ or $g_1(Y)$ is significantly different from 0. and Both $g_1(X)$ and $g_1(Y)$ are not significantly different from 0 and $g_1(X) \times g_1(Y)$ is in the same direction with r_{XY} .	$ g_1(X) $ is significantly less than $ g_1(Y) $, and $g_1(X)$ or $g_1(Y)$ is significantly different from 0. and Both $g_1(X)$ and $g_1(Y)$ are not significantly different from 0 and $g_1(X) \times g_1(Y)$ is in the same direction with r_{XY} .	Both $g_1(X)$ and $g_1(Y)$ are significantly different from 0 and $g_1(X) \times g_1(Y)$ is in the different direction with r_{XY} .	$ g_1(X) $ and $ g_1(Y) $ are not significantly different, or Both $g_1(X)$ and $g_1(Y)$ are not significantly different from 0.
Excessive Kurtosis	$ g_2(Y) $ is significantly less than $ g_2(X) $, and $g_2(X)$ or $g_2(Y)$ is significantly different from 0, and Both $g_2(X)$ and $g_2(Y)$ are not significantly different from 0 and have different directions.	$ g_2(X) $ is significantly less than $ g_2(Y) $, and $g_2(X)$ or $g_2(Y)$ is significantly different from 0, and Both $g_2(X)$ and $g_2(Y)$ are not both significantly different from 0 and have different directions.	Both $g_2(X)$ and $g_2(Y)$ are significantly different from 0 and have different directions.	$ g_2(X) $ and $ g_2(Y) $ are not significantly different, or Both $g_2(X)$ and $g_2(Y)$ are not significantly different from 0.

Table 2

The Simulation Results of Directional Dependency if a Covariate Exists and the Correlation Between Explanatory and Response Variables Are .2 ($\rho_{XY} = .2$)

Explanatory Variable	Covariate Skewness											
	SNeg			Neg			Norm			Pos		
Skewness	X	Y	U	X	Y	U	X	Y	U	X	Y	U
Covariate Correlation = 0 (No Covariate Effect)												
SNeg	.98	.00	.00	.02	.98	.00	.02	.98	.00	.02	.98	.00
Neg	.98	.00	.00	.02	.98	.00	.02	.98	.00	.02	.98	.00
Norm	.02	.02	.96	.00	.02	.96	.00	.02	.96	.00	.01	.97
Pos	.97	.00	.00	.03	.98	.00	.02	.97	.00	.03	.98	.00
SPos	.98	.00	.00	.02	.99	.00	.01	.98	.00	.02	.98	.00
Covariate Correlation = .2												
SNeg	1.00	.00	.00	.01	.98	.00	.02	.98	.00	.03	.98	.00
Neg	.99	.00	.00	.01	.99	.00	.01	.98	.00	.02	.96	.00
Norm	.02	.02	.96	.00	.02	.96	.00	.02	.95	.00	.02	.97
Pos	.97	.00	.00	.03	.97	.00	.03	.99	.00	.01	.98	.00
SPos	.97	.00	.00	.03	.97	.00	.03	.98	.00	.02	.99	.00
Covariate Correlation = .5												
SNeg	1.00	.00	.00	.00	1.00	.00	.00	.98	.00	.02	.60	.00
Neg	1.00	.00	.00	.00	1.00	.00	.00	.98	.00	.02	.56	.00
Norm	.00	.48	.50	.02	.00	.24	.74	.02	.02	.96	.00	.23
Pos	.25	.00	.00	.75	.59	.00	.00	.42	.97	.00	.03	1.00
SPos	.26	.00	.00	.74	.56	.00	.00	.44	.98	.00	.02	1.00
Covariate Correlation = .8												
SNeg	.92	.00	.08	.00	1.00	.00	.00	.98	.00	.03	.00	.00
Neg	.26	.00	.74	.00	.92	.00	.09	.98	.00	.02	.00	.00
Norm	.00	.97	.00	.03	.00	.96	.00	.04	.02	.97	.00	.98
Pos	.00	.00	.00	1.00	.00	.00	.00	1.00	.98	.00	.02	.92

Note. X = Proportion of replications suggesting X as the explanatory variable, which is true. Y = Proportion suggesting Y as the explanatory variable, which is false. U = Proportion of undetermined directional dependency replications. N = Proportion suggesting neither is explanatory variable. $SNeg$ = Strongly negative skewed. Neg = Negative skewed. $Norm$ = Normal distribution. Pos = Positive skewed. $SPos$ = Strongly positive skewed. The bold numbers represent the conditions that are less than 95% accurate (i.e., when the proportion of replications suggesting X as explanatory variable is less than .95).

Table 3

The Simulation Results of Directional Dependency if a Covariate Exists and the Correlation between Explanatory and Response Variables Are .5 ($\rho_{XY} = .5$)

Explanatory Variable	Covariate Skewness											
	SNeg			Neg			Norm			Pos		
Skewness	X	Y	U	X	Y	U	X	Y	U	X	Y	U
Covariate Correlation = 0 (No Covariate Effect)												
SNeg	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
Neg	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
Norm	.02	.02	.96	.00	.02	.03	.95	.00	.01	.02	.96	.00
Pos	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
SPos	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
Covariate Correlation = .2												
SNeg	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
Neg	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
Norm	.02	.03	.96	.00	.02	.02	.96	.00	.01	.02	.97	.00
Pos	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
SPos	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
Covariate Correlation = .5												
SNeg	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.01
Neg	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	.98	.00	.14
Norm	.00	.47	.51	.02	.00	.22	.77	.01	.02	.01	.24	.73
Pos	.85	.00	.00	.15	.98	.00	.00	.02	1.00	.00	.00	.00
SPos	.99	.00	.00	.02	1.00	.00	.00	.00	1.00	.00	.00	.00
Covariate Correlation = .8												
SNeg	.84	.00	.16	.00	1.00	.00	.00	.00	1.00	.00	.00	.00
Neg	.10	.00	.90	.00	.80	.00	.21	.00	1.00	.00	.00	.00
Norm	.00	.97	.00	.03	.00	.96	.00	.04	.02	.02	.97	.00
Pos	.00	.00	.00	1.00	.01	.00	.00	1.00	.00	.00	.19	.00
SPos	.01	.00	.00	.99	.14	.00	.00	.87	1.00	.00	.00	.00

Note. X = Proportion of replications suggesting X as the explanatory variable, which is true. Y = Proportion suggesting Y as the explanatory variable, which is false. U = Proportion of undetermined directional dependency replications. N = Proportion suggesting neither is explanatory variable. $SNeg$ = Strongly negative skewed. Neg = Negative skewed. $Norm$ = Normal distribution. Pos = Positive skewed. $SPos$ = Strongly positive skewed. The bold numbers represent the conditions that are less than 95% accurate (i.e., when the proportion of replications suggesting X as explanatory variable is less than .95).