The Journal of Economic Education

ISSN: 0022-0485 (Print) 2152-4068 (Online) Journal homepage: http://www.tandfonline.com/loi/vece20

Oh No! I Got the Wrong Sign! What Should I Do?

Peter E. Kennedy

To cite this article: Peter E. Kennedy (2005) Oh No! I Got the Wrong Sign! What Should I Do?, The Journal of Economic Education, 36:1, 77-92, DOI: 10.3200/JECE.36.1.77-92

To link to this article: http://dx.doi.org/10.3200/JECE.36.1.77-92

Content Articles in Economics

In this section, the *Journal of Economic Education* publishes articles concerned with substantive issues, new ideas, and research findings in economics that may influence or can be incorporated into the teaching of economics.

HIRSCHEL KASPER, Section Editor

Oh No! I Got the Wrong Sign! What Should I Do?

Peter E. Kennedy

Abstract: Getting a "wrong" sign in empirical work is a common phenomenon. Remarkably, econometrics textbooks provide very little information to practitioners on how this problem can arise. The author exposits a long list of ways in which a wrong sign can occur and how it might be corrected.

Key words: data mining, false significance, misspecification
JEL codes: A2, C00, C50

All researchers have experienced, far too frequently, the frustration caused by finding that the estimated sign on their favorite variable is the opposite of what they anticipated it would be. This is probably the most alarming thing "that gives rise to that almost inevitable disappointment one feels when confronted with a straightforward estimation of one's preferred structural model" (Smith and Brainard 1976, 1299). To address this problem, researchers might naturally seek help from applied econometrics textbooks, looking for a section entitled "How to deal with the wrong sign." It is remarkable that a perusal of existing texts does not supply sections devoted to this common problem.[1] A possible reason for this, in the words of a referee of this article, is that when asked about a wrong sign, textbook authors would advise as follows: "If you get a wrong sign, read this book because you have probably done the econometrics incorrectly."

This response is unfortunate; an exposition of examples of how a wrong sign can arise, and what to do about it, can be an eye-opener for students, a useful

Peter E. Kennedy *is a professor of economics at Simon Fraser University (e-mail: kennedy@sfu.ca).* *The author thanks Badi Baltagi, Bill Becker, John Fountain, Ed Maberly, Angelo Melino, Dorian Owen, Ron Smith, and anonymous referees for suggestions, some of which he has incorporated.*

resource for instructors, and a great help to practitioners struggling with this problem. In short, gathering together a variety of ways in which wrong signs can occur should enhance student and practitioner understanding of econometrics. In this spirit, I attempt to fill this void in our textbook literature by expositing several possible reasons for obtaining a wrong sign and suggesting how corrections might be undertaken. Because many such corrections smack of data mining, before moving to these examples, I present a discussion of the worrisome relationship between wrong signs and data mining.

DO WRONG SIGNS BREED DATA MINING?[2]

Finding a wrong sign will surely cause researchers to alter their empirical analysis on the basis of what they have learned from the data, a form of data mining. *Data mining* refers to "a broad class of activities that have in common a search over different ways to process or package data statistically or econometrically with the purpose of making the final presentation meet certain design criteria" (Hoover and Perez 2000, 196). It is typically denigrated by econometricians; for example, Mukherjee, White, and Wuyts (1998, 30) claim that " . . . any attempt to allow data to play a role in model specification . . . amounted to data mining, which was the greatest sin any researcher could commit." On the other hand, Hoover (1995, 243) maintains that " . . . data mining is misunderstood, and once it is properly understood, it is seen to be no sin at all."

Who is right here—is data mining a sin, or not? Both sides are right—some variants of data mining can be classified as the greatest of econometric sins, but other variants of data mining can be viewed as important ingredients in data analysis. Unfortunately, these two variants usually are not mutually exclusive and so frequently conflict in the sense that to gain the benefits of the latter, one runs the risk of incurring the costs of the former.

Two markedly different views of data mining lie within the scope of the general definition given above. One view is that it refers to experimenting with the data to produce a specification. The problem with this, and why it is viewed as a sin, is that such a procedure is almost guaranteed to produce a specification tailored to the peculiarities of that particular data set and, consequently, will be misleading in terms of what it says about the underlying process generating the data. Furthermore, traditional testing procedures used to "sanctify" the specification are no longer legitimate; these data, because they have been used to generate the specification, cannot be judged impartial if used to test that specification.

An alternative view of data mining is that it refers to experimenting with the data to discover empirical regularities that can inform economic theory. This approach to data mining has been welcomed into the mainstream of statistical analysis by the recent launching of the journal, *Data Mining and Knowledge Recovery*. Data mining's greatest virtue is that it can uncover empirical regularities that point to errors and omissions in theoretical specifications, an example of which was described by Kennedy (1998, 87). The spirit of this approach is captured by Thaler's (2000, 139) remark that "Some economists seem to feel that data-driven theory is, somehow, unscientific. Of course, just the opposite is true."

The art of the applied econometrician is to allow for data-driven theory while avoiding the considerable dangers inherent in data mining. In crude terms, researchers should listen to the data but know when to tell the data to shut up!

In summary, this second type of data mining identifies regularities in or characteristics of the data that should be accounted for and understood in the context of the underlying theory. A wrong sign, for example, may suggest a need to rethink the theory behind one's model, resulting in a new specification founded on a more broad-based understanding. This is to be distinguished from a new specification created by mechanically remolding the old specification to fit the data; this would risk incurring the costs described earlier when discussing the first variant of data mining.[3] This positive view of data mining is not new to the economics literature. Backhouse and Morgan (2000, 176) summarized a symposium on data mining, published in the June 2000 issue of the *Journal of Economic Methodology*, by noting that some of the papers in this symposium advocated an increase in data mining but "hedge this with strong warnings about the need for such data mining to be undertaken in a suitable manner."

In light of this discussion of data mining, a wrong sign could be considered a blessing, not a disaster. Getting a wrong sign is a traumatic but nonetheless friendly message that some detective work needs to be done—there is undoubtedly some shortcoming in the researcher's theory, interpretation, data, or estimation procedure. The empirical results are sending a message; a researcher should follow up by rethinking the analysis but throughout this process, remain sensitive to data mining dangers.

A more formal way of articulating this view is to note that the sign of a key variable can be used as a test statistic for a general specification test of the model being estimated: If the sign is significantly wrong, the null of this specification is rejected. The power of this test is not easily determined because (as this article makes abundantly clear) the alternative hypothesis is in general not clearly defined. Indeed, much of this article can be viewed as an effort to lay out possible alternative hypotheses (i.e., specifications) that may be relevant. Rejecting the null hypothesis that the coefficient is zero using a one-sided test reveals that the model in hand is misspecified, that an error has been made in interpreting the coefficients, that the data are deficient in some way, that an inappropriate estimation procedure has been used, or that a mistake has been made in conducting the test. This should initiate a major investigation by the researcher to determine the reason for this rejection. The examples presented below are illustrations of the kinds of things a researcher should be looking for during this investigation.

WRONG SIGN EXAMPLES

Faulty Economic Theory

A first step in dealing with a wrong sign should be to review the economic theory that gave rise to the prior belief concerning the sign. The following examples illustrate how a theoretical misspecification can give rise to a wrong sign.

Example 1: Inappropriate substitute. In a regression of the demand for Ceylonese tea on income, the price of Ceylonese tea, and the price of Brazilian coffee, the researchers obtained a positive sign on the price of Ceylonese tea (Rao and Miller 1971, 38–39). An explanation for this is that it is the price of other tea, such as Indian tea, that is the relevant substitute here.

Example 2: Real versus nominal. Estimates of consumption functions have often produced positive signs on the interest rate. Gylfason (1981) cited several such studies, explaining that the researchers obtained this wrong sign because they used the nominal rather than the real interest rate.

Example 3: Defining learning. In early studies in economic education, researchers regressed learning, measured as the difference between posttest and pretest scores, on the pretest score (as a measure of student ability) and other explanatory variables, obtaining a negative sign on pretest. Becker and Salemi (1977) spelled out several ways in which faulty theory could explain this wrong sign. One example is that the true specification may be that the posttest score depends on the pretest score with a coefficient less than unity. Subtracting pretest from both sides of this relationship produces a negative coefficient on pretest in the relationship connecting the score difference to the pretest score.

Example 4: Reaction function. Macroeconomic researchers were puzzled by empirical results indicating that contractionary monetary policy increased inflation; as noted by Sims (1992), this could happen if monetary authorities were forward-looking, contracting monetary policy in anticipation of higher inflation.

Interpretation Errors

A second step in checking reasons for a wrong sign is to ensure that the sign does not reflect an error in interpreting the empirical results. Such errors could arise from confusing the *ceteris paribus* interpretation of regression results, an algebraic error in interpreting the parameterization, neglecting interaction terms, using a nonlinear approximation, or failing to sort out dynamics.

Example 5: Ceteris paribus *confusion*. In a regression of yearling (racehorse) auction prices on various characteristics of the yearling, plus information on its sire (father) and dam (mother), Robbins and Kennedy (2001) found that although the estimated coefficient on dam dollar winnings was positive, the coefficient on number of dam wins was negative, suggesting that yearlings from dams with more race wins are worth less. This wrong sign problem is resolved by recognizing that the negative sign means that holding dam dollar winnings constant, a yearling is worth less if its dam required more wins to earn those dollars. Although proper interpretation solves the sign dilemma, in this case, an adjustment to the specification is attractive: replace the two dam variables with a new variable, earnings per win.

Example 6: Ceteris paribus *confusion*. In a regression of house price on square feet, number of bathrooms, number of bedrooms, and a dummy for a family room, the researcher obtains a negative sign on family room, suggesting that adding a family room onto a house will decrease its value. The coefficient on the family room dummy tells us the change in the house price if a family room is added, holding constant the other regressor values, in particular holding constant square feet. So adding a family room under this constraint must entail a reduction in square footage elsewhere, such as smaller bedrooms or loss of a dining room, which will entail a loss in house value. In this case, the net effect on price is negative. To estimate the impact on house price of adding a 600-square-foot family room, for example, one must account for contributions from both the square feet regressor and the family room dummy.

Example 7: Algebraic mixup. Regressing growth on male education levels and the gap between male and female education levels (GAP) should, when sorted out algebraically, give the same results as when the regressors are female education levels and GAP. As noted by Knowles, Lorgelly, and Owen (2002), however, some opposite signs appearing in the literature have resulted from a failure to do this algebraic sorting.

Example 8: Neglecting interaction terms. Economics exam scores are regressed on grade point average (GPA) and an interaction term that is the product of GPA and ATTEND, percentage of classes attended, as reported in Wooldridge (2003, 194–96). The interaction term is included to capture the belief that attendance benefits better students more than poorer students. Although the interaction term has a positive sign, GPA has a negative sign, suggesting that students with higher ability, as measured by GPA, have lower exam scores. This dilemma is easily explained—the partial derivative of exam scores with respect to GPA is the coefficient on GPA plus the coefficient on the interaction term times ATTEND. The second term outweighs the first for the vast majority of the observations, so the overall influence of GPA on exam scores is positive, as expected.

Example 9: Functional form approximation. In a regression of house prices on several characteristics of houses, including number of rooms and the square of the number of rooms, the researcher obtained a negative sign on number of rooms, as reported in Wooldridge (2003, 192). Although there is a positive coefficient on the square of number of rooms, this nonetheless suggests that, for a small number of rooms, having more rooms decreases price. This could happen because in the data there are no (or few) observations with a small number of rooms, so the quadratic term dominates the linear term throughout the range of the data. The negative sign on the linear term comes about because it provides the best approximation to the data and is relevant only to this range of the number of rooms.

Example 10: Dynamic confusion. In a regression of income on lagged income and investment spending, the investment coefficient, interpreted as the multiplier, is less than unity, a type of wrong sign, as reported in Rao and Miller (1971, 44–45).

Calculating the implied long-run impact on income (by setting lagged income equal to income and solving for income as a function of investment) resolves this dilemma.

Example 11: Dynamic confusion. Panel data on U.S. states are used to estimate the impact of public capital stock (in addition to private capital stock and labor input) on state output. Fixed-effects estimation produces a negative sign on the public capital stock coefficient estimate. Baltagi and Pinnoi (1995) noted that this could be because fixed effects estimates the short-run reaction; pooled OLS (ordinary least squares), the "between" estimator, and random effects all produce the expected positive sign, suggesting that the long-run impact is positive.

A similar point is made by Durlauf and Quah (1999, 286) in the context of estimating the determinants of economic growth. In cross-country panel studies, any estimation method that removes individual effects also removes long-run variation (i.e., across countries) in growth rates, leaving higher-frequency variation (such as business cycles) for which the magnitudes and possibly the signs are different.

In both these examples, the sign difference should alert a researcher to a difference between short- and long-run responses and thereby offer more information about the data-generating process. If data permit, one response might be to add lags to focus more sharply on the dynamics.

Example 12: Dynamic confusion. Suppose x affects y positively with both immediate and lagged effects. Regressing y_t on x_t and x_{t-1} may produce a negative coefficient on x_{t-1}. The explanation for this is that the long-run impact of x is smaller than its short-run impact—the short-run impact is measured by the coefficient on x_t, whereas the long-run impact is measured by the sum of the two coefficients.

Data Problems

A third way in which a wrong sign could arise is related to a variety of data problems. These examples illustrate several ways in which this could occur: bad data, inappropriate data definitions, measurement errors, influential observations, poor instruments, and reversed measures.

Example 13: Bad data. A prominent myth about the American economy is that small businesses are responsible for the majority of new jobs created. Davis, Haltiwanger, and Schuh (1996, 70–72) noted that one reason for this false conclusion is that many researchers were employing the Dun and Bradstreet Market Identifier (DMI) database. They document enormous discrepancies between these data and other databases. For example, in 1986 in the DMI database, employment was 9 million people greater than was employment in other databases, 81 percent of mass layoffs identified in the DMI database were not layoffs at all but were recorded as such because of some other event such as a change in ownership structure, and 96 percent of new firms recorded by other databases were not identified as such in the DMI database.

Example 14: Data definitions. Davis, Haltiwanger, and Schuh (1996, 62–66) noted another reason for the wrong sign on the main source of job creation. The way in which the dependent variable is defined by some researchers allows firms to migrate between size categories from one year to the next, causing misleading measures of job creation by firm size. Suppose a large firm, firm A, loses a lot of jobs and as a result in the second period is classified as a small firm. The number of jobs in small firms in the second period jumps dramatically, not because small firms created more jobs but rather because firm A's jobs are now counted in the pool of small firm jobs.

Example 15: Data definitions. Theory may suggest that bad weather depresses stock market traders, causing them to sell, so that regressing stock price changes on a dummy for bad weather should produce a negative sign on the weather dummy. Kramer and Runde (1997) reported a positive sign when bad weather is defined as 100 percent cloud cover plus relative humidity above 70 percent. By changing the definition of bad weather to cloud cover more than 80 percent or relative humidity outside the range 25 to 75 percent, the sign magically changes. This example illustrates more than the role of variable definitions or measurement in affecting coefficient signs—it illustrates that data mining can produce whatever results one wants, highlighting the importance of sensitivity analyses that report results from a range of specifications.

Example 16: Data definitions. Regressing growth on female education produces a negative sign, as reported by Barro and Lee (1994). Knowles, Lorgelly, and Owen (2002) concluded that this wrong sign resulted from Barro and Lee's use of a base-period measure of female education, rather than a measure whose timing matches that of the output-per-worker measure.

Example 17: Measurement errors. It is not uncommon to regress the crime rate on the per capita number of police and obtain a positive coefficient, suggesting that more police engender more crime. One possible reason for this is that having extra police causes more crime to be reported—actual crime may not be affected, or could be decreased.

Example 18: Measurement errors. If measurement errors are correlated with the true value of the variable being measured (contrary to the usual econometric assumption), bias sufficient to change a coefficient's sign can result. Bound, Brown, and Mathiowetz (2001) documented that this often is the case.

Example 19: Influential observations. In a regression of infant mortality on doctors per thousand population, using data on the 50 U.S. states plus the District of Columbia, the sign on doctors was positive, as reported by Wooldridge (2003, 314–15). This happened because the District of Columbia is an unrepresentative observation—relative to other observations it has pockets of extreme poverty and, for extraneous reasons (because the District of Columbia is the nation's capital), a large number of doctors. Removing this influential observation resolved the

sign dilemma. Dealing with unusual and influential observations must be done with care. As Zellner (1981) emphasized, these observations could be the most important observations in the data because they provide variation that allows an equivocal result to be unequivocal. If other considerations in the context of the problem at hand argue for deletion, as is the case for the District of Columbia, omission of the observation is appropriate. If deletion cannot be justified on other grounds, this influential observation should be noted when presenting results, following Kennedy's (2002) 10th commandment of applied econometrics: Report a sensitivity analysis.

Example 20: Inappropriate instruments. Instrumental variable (IV) estimation is usually employed to alleviate bias caused by correlation between an explanatory variable and the equation error. Consider a regression of incidence of violent crime on percentage of population owning guns, using data on U.S. cities. Because gun ownership may be endogenous (i.e., higher crime causes people to obtain guns), gun magazine subscriptions is used as an IV for gun ownership, as was done, for example, in Kleck and Patterson (1993). The appropriate IV estimator, two-stage least squares, produced a negative sign, the reverse of the sign obtained using OLS.[4] This probably happened because an unsuitable IV was employed. The IV gun subscriptions was representing gun ownership that is culturally patterned, linked with a rural hunting subculture, and so did not represent gun ownership by individuals residing in urban areas, who own guns primarily for self-protection. IV estimation bases its estimate on that part of the explanatory variable that moves in concert with the IV; this variation in the explanatory variable may correspond to dependent variable reactions that are different from dependent variable reactions associated with other variation in the explanatory variable.

Example 21: Weak instruments. When an IV is only weakly correlated with the regressor for which it is serving as an instrument, most of the bias of OLS persists in IV estimation, even for large sample sizes, and standard errors are large and, most worrisome, underestimated. This could give rise to questionable empirical results, in particular estimates with wrong signs or testing significantly different from zero when they should not do so. Shea (1997) suggested a partial R^2 measure that should be checked before proceeding with IV estimation. Zivot, Startz, and Nelson (1998) contains a good summary of this literature, with examples.

Example 22: Reversed measure. A regression of consumption on a consumer confidence measure, among other variables, unexpectedly produces a negative sign on the consumer confidence measure. This could happen if a researcher does not realize that small numbers for the consumer confidence measure correspond to high consumer confidence. A similar problem occurs if one of several time series is reversed chronologically, perhaps because it comes from a different data source. It has been known[5] for economists to present an entire seminar trying to explain a wrong sign only to discover afterwards that it resulted from their software reversing the coding on the dependent variable in their logit analysis.

Classic Econometric Problems

This category includes several econometric phenomena, well documented in econometrics textbooks, that can give rise to wrong signs: an omitted explanatory variable, nonstationarity problems, high variances, selection bias, and lack of identification.

Example 23: Omitted explanatory variable. Using data on a cross section of countries, Barro (1991) regressed growth in per capita GDP (gross domestic product) on initial per capita GDP, obtaining a positive sign (as shown in his Figure 1), a wrong sign for convergence. Adding a set of control variables to allow for the determinants of the steady state reverses this sign (as shown in his Figure 2). In general, an omitted explanatory variable with a positive (negative) coefficient in the regression (which is negatively [positively] correlated with initial per capita GDP) could be causing this wrong sign.

Example 24: Omitted explanatory variable. A sample of women was asked whether they smoke and then were resampled 20 years later. As reported by Appleton, French, and Vanderpump (1996), a probit (using the smoking dummy as the explanatory variable) was run on whether the women were still alive after 20 years, and the smokers were found more likely to be alive. This could happen if the nonsmokers in the sample were mostly older and the smokers mostly younger. Adding age as an explanatory variable resolved this problem.

Example 25: Omitted trend. In time series data, a common positive (negative) trend could swamp what would otherwise be a negative (positive) relationship between two variables; omitting the common trend would give rise to the wrong sign. Wooldridge (2003, 348–49) provided an example in which housing investment was regressed on a housing price index, producing a positive sign on price. This sign is reversed when a time trend is added as a regressor. Interpreting this result is problematic because lack of identification suggests that both signs are wrong. In general, however, common trends can give rise to spurious relationships, concealing fundamental relationships.

Example 26: Mixing orders of integration. The Puerto Rico employment rate was regressed on several variables, including U.S. GNP (gross national product), producing a negative (but insignificant) sign on GNP, as reported by Wooldridge (2003, 351). By including a time trend among the regressors, the sign on GNP became positive (and significant). Although this example could be interpreted as an omitted explanatory variable, it illustrates a much more fundamental issue. In this example, a stationary variable (employment rate) was regressed on a nonstationary variable (GNP), without a cointegrating relationship (among the regressors) to avoid conflict between orders of integration. A nonsense result follows. In this case, GNP is trendstationary,[6] so that after removing the time trend (by including the time trend as a regressor), it becomes stationary, permitting this regression to make sense; the positive sign on GNP can now be interpreted as reflecting the influence of GNP departing from its trend.

Example 27: Ignoring nonstationarity. The preceding two examples could be viewed as cases in which removing a time trend renders nonstationary variables stationary and so avoids misleading results. But removing a time trend may not eliminate the nonstationarity; it may be necessary to first difference the data because the data contain a unit root—the trend is stochastic, not deterministic. Wooldridge (2003, 379–80) presented an example in which the log of hourly wage was regressed on the log of output per hour and a time trend, obtaining an elasticity estimate greater than unity, a type of wrong sign. When first differences are employed to remove a unit root, the elasticity estimate becomes less than unity.[7]

Example 28: High variances. When estimated coefficients have high variances, their sampling distributions are widely spread and may straddle zero, implying that it is quite possible that a draw from this distribution will produce a wrong sign. Estimating a demand curve by regressing quantity of coffee on the price of coffee and the price of tea, using time series data, could produce a positive sign on the price of coffee. This could happen because over time the prices of coffee and tea are highly collinear, resulting in estimated coefficients with high variances. (Indeed, one of the casual indicators of multicollinearity is the presence of wrong signs.) In this example, an attractive solution to this problem is to use the ratio of the two prices as the explanatory variable, rather than their levels. In an ideal world, of course, all information would be included before estimation, to guard against the undesirable form of data mining.

Example 29: High variances. The preceding example is one in which the wrong sign problem is solved by incorporating additional information to reduce high variances. However, multicollinearity is not the only source of high variances; they could result from a small sample size or minimal variation in the explanatory variables. Leamer (1978, 8) presented another example of how additional information can solve a wrong sign problem. Regressing household demand for oranges on total expenditure E, the price p_o of oranges and the price p_g of grapefruit (all variables logged) could produce wrong signs on the two price variables. Imposing homogeneity (if prices and expenditure double, the quantity of oranges purchased should not change) implies that the sum of the coefficients of E, p_o, and p_g is zero. This extra information reverses the price signs.

Example 30: Sample selection. A nonrandom sample because of sample selection could create a wrong sign. Regressing academic performance, as measured by SAT scores (the scholastic aptitude test is taken by many students to enhance their chances of admission to the college of their choice) on per student expenditures on education, using aggregate data on states, produced a negative sign on per student expenditures on education, as reported by Guber (1999). This wrong sign may result from sample selection bias. In states with high education expenditures, a larger fraction of students may take the test. A consequence of this is that the overall ability of the students taking the test may not be as high as in states with lower education expenditures and a lower fraction of students taking the test. Some

kind of correction for this selection bias is necessary. In this example, putting in the fraction of students taking the test as an extra explanatory variable resolved this dilemma.

Example 31: Sample selection. Regressing the birthweight of children on several family and background characteristics, including a dummy for participation in AFDC (aid for families with dependent children), produced a negative sign on the AFDC dummy, as reported by Currie and Cole (1993). This could happen because mothers self-selected themselves into this program—mothers believing they were at risk for delivering a low birthweight child may have been more likely to participate in AFDC. One way of dealing with this problem is to use an instrumental variable for AFDC participation (Currie and Cole used measures of the generosity of state welfare programs). An alternative is to use the Heckman two-stage correction for selection bias or an appropriate maximum likelihood procedure. Another alternative is to confine the sample to mothers with two children, for only one of which the mother participated in the AFDC program. A panel data method such as fixed effects (or differences) could then be used to control for the unobservables that are causing the problem.

Example 32: Lack of identification. Historically, regressions of an agricultural product on price produced negative coefficients and were interpreted as demand curves—the exogenous variable "weather" affected supply but not demand, rendering this regression an identified demand curve. Estimating an unidentified equation would produce estimates of an arbitrary combination of the supply and demand equation coefficients, and so could be of arbitrary sign. The lesson here is check for identification. A classic example is Moore (1914) who regressed quantity of pig iron on price, obtained a positive coefficient and announced a new economic discovery—an upward-sloping demand curve. He was quickly rebuked for confusing supply and demand curves. Morgan (1990, ch. 5) discussed historical confusion on this issue.

Example 33: Simultaneity. Even if an equation in a simultaneous system is identified, estimation by OLS can create bias, perhaps sufficient to produce the wrong sign. More policemen may serve to reduce crime, for example, but higher crime will cause municipalities to increase their police force, so when crime is regressed on police, it is possible to get a positive coefficient estimate. An appropriate estimation procedure, such as two-stage least squares, should be used; the crucial thing here, as illustrated in example 20, is to ensure that a suitable instrumental variable is employed.

Either Sign Is a Wrong Sign

In many situations, the "right" sign is no sign, in the sense that the result should be insignificantly different from zero. In such cases, a "significant" sign could result from regression to the mean, nonstationarity, or underestimated variances.

Example 34: Regression to the mean. Regressing average annual growth for several countries over the period 1950–1979 on GDP per work hour in 1950 produced a negative coefficient on GDP per work hour, a result interpreted as supporting the convergence hypothesis. Friedman (1992) noted that this could arise from the regression to the mean phenomenon. Suppose there is substantive measurement error in GDP. Large underestimates of GDP in 1950 will result in low GDP per work hour and, at the same time, likely produce a higher annual growth rate over the subsequent period (because the 1979 GDP measure will likely not have a similar large underestimate). Large overestimates will have an opposite effect. As a consequence, this regression is likely to find convergence, even when none exists.

A similar example was identified by Hotelling (1933). A set of firms with high business-to-sales ratios had this measure regressed against time, finding a negative relationship, that is, over time the average ratio declined. In this case, the chosen firms probably had high ratios by chance; the negative sign came about because in subsequent years they reverted to a more normal ratio.

Davis, Haltiwanger, and Schuh (1996, 66–70) presented another example, in the context of the myth of small firms creating more jobs than large firms. On average, firms classified as large in the base year are more likely to have experienced a recent transitory increase in employment and so are more likely to contract in the following year. The opposite occurs on average to the firms classified as small.

Example 35: Nonstationarity. Regressing a random walk on an independent random walk should produce a slope coefficient insignificantly different from zero but far too frequently does not, as is now well-known. This spurious correlation is a very old problem, identified by Yule (1926) in an article entitled, "Why do we sometimes get nonsense correlations between time series?" A similar problem occurs when variables are nonstationary because they contain a trend. A classic example provided by Hendry (1980) is the ability of cumulative rainfall to "explain" the price level.

Nonstationarity causes most test statistics to mislead. For example, Mankiw and Shapiro (1985) showed that nonstationarity caused researchers to conclude that consumption is excessively sensitive to income, and Kleidon (1986) showed that nonstationarity caused researchers to conclude that variance bounds were violated, implying market inefficiency.

Example 36: Systematic measurement error. A group of students was given free milk at lunch, and a control group was not. After six months there was a significant difference in their weight gains. As Kadane and Seidenfeld (1996) reported, students were not chosen completely at random for the two groups; there was a tendency for poorer students who "needed" the milk to be assigned by teachers to the treatment group. Furthermore, they were weighed in winter clothing at the beginning of the experiment and then in spring clothing at the end of the experiment. Because wealthier students tended to wear heavier winter clothing, a systematic measurement error was introduced. This measurement error, in

conjunction with the sample selection bias (more poor students in the treatment group), was responsible for the significant results.

Example 37: Influential observations. A regression of economic growth on several explanatory variables, including male and female education, produces a significant negative sign on female education. Lorgelly and Owen (1999) reported that low base-period levels of female education in the Asian Tigers are influential in producing this result; dropping these observations produced an insignificant coefficient on female education. Rowthorn (1975) pointed out that a regression confirming Kaldor's law (that productivity growth is a positive function of employment growth) resulted from a random scatter of points and an outlier, Japan. Dyl and Maberly (1986) showed that a significant "weekend effect" was due to a major measurement error in one of the observations in the data; correcting this rendered the result without significance, as the efficient markets hypothesis predicts. As stressed in an earlier example, researchers need to report a sensitivity analysis, warning readers about the role of influential observations.

Example 38: Underestimated variance. It is common in empirical work to estimate variances using asymptotic formulas, which in small samples can produce marked underestimates of true variances, causing irrelevant variables to become statistically significant. A classic example appears in Laitinen (1978), who showed that failure to use small-sample adjustments explained why demand homogeneity had been rejected so frequently in the literature.

Example 39: Underestimated variance. The Poisson model assumes that the variance of the counts is equal to its expected value. This extra "information" causes Poisson estimation to produce marked underestimates of variances in the typical case in which there is overdispersion (the count variance is larger than its expected value). This, in turn, can cause irrelevant coefficients in Poisson regressions to be statistically significant. For example, List (2000) reported results much more significant than one would expect from his data. Researchers should always check for overdispersion and, if present, use either a negative binomial regression model, if the overdispersion is thought to be caused by heterogeneity, or a zero-inflated or hurdle model, if the dispersion is thought to result from excess zeros (See Greene 2003, 740–52). Wooldridge (2003, 576) suggested dealing with this problem by adjusting the standard errors, assuming that the variance is proportional to (i.e., as opposed to equal to) the mean. Wooldridge (2002, ch. 19) extended this suggestion to an unrestricted form of variance.

Example 40: Honest mistakes. Researchers sometimes make mistakes that can cause results to be opposite to what should have been found. Levitt (1997) found that the impact of police on crime was significantly negative. McCrary (2002) noted that Levitt transformed his data to correct for heteroskedasticity by multiplying rather than dividing; when the correct transformation was used, there was no significant impact of police on crime. Mistakes can be produced by software as well as researchers. McCullough and Vinod (2003) explained how software

used for estimation of complicated nonlinear specifications can easily produce incorrect answers; they suggest several ways in which researchers can guard against this problem.

CONCLUSION

The examples presented in this article catalog a wide range of reasons why a wrong sign might arise. Some are clearly more important than others, namely faulty economic theory, omitted explanatory variables, ignoring nonstationarity, sample selection bias, high variances, lack of identification, influential observations, and *ceteris paribus* confusion, but all are worthy of note. Although I suggest solutions, clearly there is no easily identified route to finding the reason for a wrong sign: in general, solutions are context specific. Beyond looking at the specific context, researchers should seek to understand their result by undertaking a selection of investigative actions, such as viewing the data with imaginative graphs, checking regressions on subsets of the data, forecasting extra-sample data, and looking for collaborative and falsifying evidence. By providing a range of examples, this article should help researchers searching for a solution to a wrong-sign problem.

What should be done if a researcher's detective work can turn up no reasonable explanation for the wrong sign? Try and get it published. Wrong sign puzzles, such as the Leontief paradox, are a major stimulus to the development of the economic discipline. A prominent example is the wrong sign result of Card and Krueger (1994), who found that an increase in the minimum wage leads to an increase in employment, not a decrease as standard economic theory suggests. As a second example, recent evidence suggests that there is a positive relationship between import tariffs and growth across countries in the late 19th century, a wrong sign in many economists' view. Irwin (2002) extended the relevant economic theory to offer an explanation for this. As a third example, consider the forward discount anomaly: The change in the future exchange rate is typically found to be negatively related to the forward premium, as surveyed, for example, in Engle (1996).

There is no definitive list of ways in which wrong signs can be generated. In general, any theoretical oversight, interpretation error, data problem, or inappropriate estimating technique could give rise to a wrong sign. Observant readers might have noted that many, perhaps all, could be classified under a single heading: Researcher foolishness. This underlines the importance of the first of Kennedy's (2002) 10 commandments of applied econometrics: Use common sense.

NOTES

1. Most textbooks mention this phenomenon but provide few examples of different ways in which it could arise. Wooldridge (2003) is an exception; several examples of wrong signs are scattered throughout his textbook.
2. In this section I borrowed heavily from Kennedy's (2002) discussion of his seventh commandment of applied econometrics: Understand the costs and benefits of data mining.
3. Recent research has developed variants of this first type of data mining that markedly decrease its drawbacks. See, for example, Krolzig and Hendry (2001).

4. I am indebted to Tomislav Kovandzic for this example.
5. I am indebted to Marie Rekkas for this anecdote.
6. Although GNP in this example is trend stationary, GNP itself is nonstationary because its mean is not constant. Regressing a stationary variable on a trend stationary variable does not make sense unless the trend is removed.
7. A drawback to first differencing is that it removes the long-run relationship, causing estimation to reflect the short-run relationship, which may be different. An error correction model may be a more appropriate alternative here if the log of hourly wage and the log of output per hour are cointegrated.

REFERENCES

Appleton, D. R., J. M. French, and M. P. J. Vanderpump. 1996. Ignoring a covariate: An example of Simpson's paradox. *American Statistician* 50 (4): 340–41.

Backhouse, R. E., and M. S. Morgan. 2000. Introduction: Is data mining a methodological problem? *Journal of Economic Methodology* 7 (2): 171–81.

Baltagi, B. H., and N. Pinnoi. 1995. Public capital stock and state productivity growth: Further evidence from an error components model. *Empirical Economics* 20 (2): 351–59.

Barro, R. J. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics* 106 (2): 407–43.

Barro, R. J., and J-W. Lee. 1994. Sources of economic growth. *Carnegie-Rochester Conference Series on Public Policy* 40 (1): 1–46.

Becker, W. E., and M. K. Salemi. 1977. The learning and cost effectiveness of AVT supplemented instruction: Specification of learning models. *Journal of Economic Education* 8 (2): 77–92.

Bound, J., C. Brown, and N. Mathiowetz. 2001. Measurement error in survey data. In J. J. Heckman and E. E. Leamer, eds., *Handbook of econometrics*, vol. v. 3705–843. Amsterdam: North Holland.

Card, D. E., and A. B. Krueger. 1994. Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review* 84 (4): 772–93.

Currie, J., and N. Cole. 1993. Welfare and child health: The link between AFDC participation and birth weight. *American Economic Review* 83 (4): 971–83.

Davis, S. J., J. C. Haltiwanger, and S. Schuh. 1996. *Job creation and destruction*. Cambridge, MA: MIT Press.

Durlauf, S. N., and D. T. Quah. 1999. The new empirics of economic growth. In J. B. Taylor and M. Woodford, eds., *Handbook of macroeconomics*, vol. IA. 235–308. Amsterdam: North Holland.

Dyl, E. A., and E. D. Maberly. 1986. The weekly pattern in stock index futures: A further note. *Journal of Finance* 41 (5): 1149–52.

Engle, C. 1996. The forward discount anomaly and the risk premium: A survey of recent evidence. *Journal of Empirical Finance* 3 (2): 123–92.

Friedman, M. 1992. Do old fallacies ever die? *Journal of Economic Literature* 30 (4): 2129–32.

Greene, W. H. 2003. *Econometric analysis*. 5th ed. Upper Saddle River, NJ: Prentice-Hall.

Guber, D. C. 1999. Getting what you pay for: The debate over equity in public school expenditures. *Journal of Statistics Education* 7 (2). [Online] Available at: www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm.

Gylfason, H. 1981. Interest rates, inflation, and the aggregate consumption function. *Review of Economics and Statistics* 63 (2): 233–45.

Hendry, D. F. 1980. Econometrics—alchemy or science? *Economica* 47 (Nov.): 387–406.

Hoover, K. D. 1995. In defense of data mining: Some preliminary thoughts. In K. D. Hoover and S. M. Sheffrin, eds., *Monetarism and the methodology of economics: Essays in honor of Thomas Mayer.* 242–57. Aldershot, England: Edward Elgar.

Hoover, K. D., and S. J. Perez. 2000. Three attitudes towards data mining. *Journal of Economic Methodology* 7 (2): 195–210.

Hotelling, H. 1933. Review of *The triumph of mediocrity in business*, by Horace Secrist. *Journal of the American Statistical Association* 28:463–65.

Irwin, D. A. 2002. Interpreting the tariff-growth correlation of the late 19th century. *American Economic Review, Papers and Proceedings* 92 (2): 165–69.

Kadane, J. D., and T. Seidenfeld. 1996. Statistical issues in the analysis of data gathered in the new designs. In J. D. Kadane, ed., *Bayesian methods and ethics in a clinical trial design*. 115–25. New York: Wiley.

Kennedy, P. E. 1998. *A guide to econometrics*. 4th ed. Cambridge, MA: MIT Press.

———. 2002. Sinning in the basement: What are the rules? The ten commandments of applied econometrics. *Journal of Economic Surveys* 16 (4): 569–89.

Kleck, G., and E. B. Patterson. 1993. The impact of gun control and gun ownership levels on violence rates. *Journal of Quantitative Criminology* 9 (2): 249–87.
Kleidon, A. W. 1986. Variance bounds tests and stock price valuation methods. *Journal of Political Economy* 94 (5): 953–1001.
Knowles, S., P. K. Lorgelly, and P. D. Owen. 2002. Are educational gender gaps a brake on human development? Some cross-country empirical evidence. *Oxford Economic Papers* 54 (1): 118–48.
Kramer, W., and R. Runde. 1997. Stocks and the weather: An exercise in data mining or yet another capital market anomaly? *Empirical Economics* 22 (4): 637–41.
Krolzig, H.-M., and D. F. Hendry. 2001. Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control* 25 (7): 831–66.
Laitinen, K. (1978). Why is demand homogeneity so often rejected? *Economics Letters* 1 (2): 187–91.
Leamer, E. E. 1978. *Specification searches: Ad hoc inference with nonexperimental data.* New York: John Wiley.
Levitt, S. 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review* 87 (3): 270–90.
List, J. A. 2000. Interview scheduling strategies of new Ph.D. economists. *Journal of Economic Education* 31 (2): 191–201.
Lorgelly, P. K., and P. D. Owen. 1999. The effect of female and male schooling on economic growth in the Barro-Lee model. *Empirical Economics* 24 (3): 537–57.
Mankiw, N. G., and M. D. Shapiro. 1985. Trends, random walks and tests of the permanent income hypothesis. *Journal of Monetary Economics* 16 (2): 165–74.
McCrary, J. 2002. Using electoral cycles in police hiring to estimate the effect of police on crime: Correction. *American Economic Review* 92 (4): 1236–43.
McCullough, B. D., and H. D. Vinod. 2003. Verifying the solution from a nonlinear solver: A case study. *American Economic Review* 93 (3): 873–92.
Moore, H. L. 1914. *Economic cycles—Their law and cause*. New York: Macmillan.
Morgan, M. S. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.
Mukherjee, C., H. White, and M. Wuyts. 1998. *Econometrics and data analysis for developing countries*. London: Routledge.
Rao, P., and R. Miller. 1971. *Applied econometrics*. Belmont, CA: Wadsworth.
Robbins, M., and P. E. Kennedy. 2001. Buyer behaviour in a regional thoroughbred yearling market. *Applied Economics* 33 (8): 969–77.
Rowthorn, R. E. 1975. What remains of Kaldor's law? *Economic Journal* 85 (1): 10–19.
Shea, J. 1997. Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics* 79 (2): 348–52.
Sims, C. A. 1992. Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review* 36 (5): 975–1011.
Smith, G., and W. Brainard. 1976. The value of *a priori* information in estimating a financial model. *Journal of Finance* 31 (5): 1299–322.
Thaler, R. H. 2000. From homo economicus to homo sapiens. *Journal of Economic Perspectives* 14 (1): 133–41.
Wooldridge, J. M. 2002. *Econometric analysis of cross-section and panel data*. Cambridge, MA: MIT Press.
———. 2003. *Introductory econometrics*. 2nd ed. Cincinnati: South-Western.
Yule, G. U. 1926. Why do we sometimes get nonsense correlations between time-series? *Journal of the Royal Statistical Society* 89:1–64.
Zellner, A. 1981. Philosophy and objectives of econometrics. In D. Currie, R. Nobay, and D. Peel, eds., *Macroeconomic analysis: Essays in macroeconomics and econometrics*. 24–34. London: Croom Helm.
Zivot, E., R. Startz, and C. R. Nelson. 1998. Valid confidence intervals and inference in the presence of weak instruments. *International Economic Review* 39 (4): 1119–46.