Confounding in Air Pollution Epidemiology: When Does Two-Stage Regression Identify the Problem?

Allan H. Marcus and Scott R. Kegler

National Center for Environmental Assessment - RTP, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

A two-stage approach has recently been proposed to assess confounding by copollutants or other variables in time–series epidemiology studies for airborne particulate matter (PM), using independent series from different cities. In the first stage of the proposed method, two regression models are fitted for each city in the analysis. The first relates the health effect to the putative causal variable such as PM without including any copollutant or confounder. The other first-stage model relates a putative confounding variable to PM. In the second stage of the analysis, the estimated city-specific regression slopes for the health-effect-versus-PM model are regressed against the estimated city-specific regression slopes for the confounder-versus-PM model. Under the proposed method, a nonzero intercept estimate in the second-stage regression would be interpreted as indicating a direct pathway from PM to the health effect, and a nonzero slope estimate would be interpreted as indicating at least partial confounding of PM with the putative confounder. A simple counterexample using an additional copollutant variable shows that inferences based on this method could be misleading. *Key words*: air pollution epidemiology, confounding, copollutants, model misspecification, multicollinearity, particulate matter epidemiology, two-stage regression. *Environ Health Perspect* **109:1193–1196 (2001). [Online 13 November 2001]**
http://ehpnet1.niehs.nih.gov/docs/2001/109p1193-1196marcus/abstract.html

In air pollution epidemiology studies, evaluating modeled associations between individual pollutants and the health outcome of concern is often complicated by multicollinearity among the measured pollutant concentrations. Statistical models that incorporate these copollutants simultaneously are often unstable, with the estimated regression coefficients possibly changing in both magnitude and direction depending on which copollutants are included. Further, coefficient standard errors will likely be inflated so that the estimated effects may not achieve statistical significance, despite actual relationships that may exist. Conversely, if single-pollutant models are used to evaluate potential associations, the problem of confounding arises: the single pollutant represents not only its own health effect but also the effects of the excluded pollutants with which it is associated.

Several authors (*1–3*) have recently applied two-stage regression techniques in an effort to identify the extent to which a pollutant is directly associated with human health effects, as opposed to acting indirectly through its association with other pollutants (confounding). Noting that relationships between copollutants differ by city, it has been proposed that such differences can facilitate the separation of direct and indirect effects through use of a second-stage meta-regression. When the first- and second-stage models have been correctly specified, the proposed approach may help clarify the role that confounding plays in observed associations between pollutants and health outcomes. In this paper we describe some general conditions under which the method may, however, be vulnerable to the effects of model misspecification.

In the applications described by Schwartz and colleagues (*1–3*), the response variable denoted by Y is most often a community-level health index, such as the number of deaths or number of hospital admissions per day (possibly transformed), rather than an individual-level health index. In reality, Y may have a Poisson or hyper-Poisson distribution, but the examples discussed below are not affected by such distributional properties. The variables W, X, and Z usually represent community-level indices of airborne particles or gaseous pollutants averaged over one or more stationary air monitoring stations. However, the two-stage method used in earlier papers (*1–3*) could potentially be applied over a much wider range of epidemiology studies, including those with individual-level health effects and exposure data.

Mathematical Model for the Two-Stage Method

We express the mathematical model for the two-stage approach to evaluating confounding using nonspecific variables to illustrate the generality of the problem. The simplest setting involves copollutant variables denoted by Z and X and a health outcome variable denoted by Y. The variables Z and X are known to be associated with each other, and one or both may be directly associated with Y. Following the approach of Schwartz (*3*), a model characterizing these relationships may be expressed as:

$$X = \gamma_0 + (\gamma_{1,\text{city}} \times Z) + \varepsilon_X \quad [1]$$

$$Y = \beta_0 + (\beta_1 \times Z) + (\beta_2 \times X) + \varepsilon_Y. \quad [2]$$

Thus, in our simple model, associations are linear and characterized parametrically. We assume that one "instance" of the system embodied in Equations 1 and 2 will be used to represent each city 1,...,K in a multicity analysis. The parameters β_1 and β_2 characterize the direct associations of Z with Y and X with Y, respectively. The parameter $\gamma_{1,\text{city}}$ characterizes the association between Z and X. While β_1 and β_2 are assumed constant across cities, $\gamma_{1,\text{city}}$ varies from city to city as the relationship between Z and X varies from city to city. The intercept terms γ_0 and β_0 may or may not depend on the city. The analyses below do not depend on city-specific values for these parameters, and for simplicity, they are assumed to be the same for all cities. Equation 1 is used to represent the association between Z and X for each city, one of many possible ways of expressing such associations but consistent with the approach described by Schwartz (*3*). In Equation 2 the variable Y is a response or health effect (such as the logarithm of mortality) that depends directly on Z as well as indirectly on Z, acting through X as a surrogate or proxy. By substituting Equation 1 into Equation 2, we have:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 \times Z) + \beta_2 \\ &\quad \times [\gamma_0 + (\gamma_{1,\text{city}} \times Z) + \varepsilon_X] + \varepsilon_Y \\ &= [\beta_0 + (\beta_2 \times \gamma_0)] + [\beta_1 + (\beta_2 \times \gamma_{1,\text{city}})] \times Z \\ &\quad + (\beta_2 \times \varepsilon_X) + \varepsilon_Y. \end{aligned} \quad [3]$$

The total Z effect is the regression coefficient of Y on Z, which from Equation 3 is

$$Z\,\text{effect}_{\text{city}} = \beta_1 + (\beta_2 \times \gamma_{1,\text{city}}) \quad [4]$$

for each city = 1,...,K. This total effect reflects both the direct and indirect associations between Z and Y.

Address correspondence to A.H. Marcus, Room 358 Catawba Building, 3210 Highway 54, Research Triangle Park, NC 27709 USA. Telephone: (919) 541-0636. Fax: (919) 541-1818. E-mail: marcus.allan@epa.gov

The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Received 25 October 2000; accepted 15 May 2001.

Equation 3 motivates the two-stage regression approach. Assuming that data on Z, X, and Y are available for city = 1,...,K, the two-stage regression process involves first estimating the marginal association between Y and Z using a first-stage regression model of Y against Z for each city, ignoring X. Consistent with the usual problems associated with model misspecification detailed in the regression literature (*4*) we anticipate that Z will also represent the effect of X and that we will recover estimates which in expectation should conform to

$$E[\hat{Z}\text{effect}_{\text{city}}] = \beta_1 + (\beta_2 \times \gamma_{1,\text{city}}) \qquad [5]$$

for city = 1,...,K. Thus, the expected Z effects are based on parameters reflecting both the direct effect of Z on Y (characterized by the parameter β_1) and the mediated effect of Z on Y acting through X (characterized by the parameters β_2 and $\gamma_{1,\text{city}}$). It is therefore of interest to further extract estimates of β_1 and β_2, and this is accomplished using a second-stage regression. To carry out the second-stage regression, estimates of $\gamma_{1,\text{city}}$ for city = 1,...,K are needed in addition to the estimated Z effects already obtained during the first stage. The estimates of $\gamma_{1,\text{city}}$ are obtained by fitting Equation 1 for each city. The second stage then consists of a regression of the estimated Z effects (i.e., $\hat{Z}\text{effect}_{\text{city}}$ for city = 1,...,K) against the estimated X with Z associations (i.e., $\hat{\gamma}_{1,\text{city}}$, for city = 1,...,$K$). Equivalently, the second stage consists of a regression of the estimates of the linear combinations $\beta_1 + (\beta_2 \times \gamma_{1,\text{city}})$ against estimates of the parameters $\gamma_{1,\text{city}}$ in an effort to estimate β_1 and β_2. The second-stage regression results in the fitted model:

$$\hat{\hat{Z}}\text{effect}_{\text{city}} = \hat{\delta}_1 + (\hat{\delta}_2 \times \hat{\gamma}_{1,\text{city}}). \qquad [6]$$

When there is sufficient variation in the parameters $\gamma_{1,\text{city}}$ across cities, the claim is that $\hat{\delta}_1$ provides an estimate of β_1 and that $\hat{\delta}_2$ provides an estimate of β_2. By way of these estimates it should be possible to separate the component of the Z effect due to direct association of Z with Y (by virtue of the estimated intercept $\hat{\delta}_1$ in the second-stage regression) from the component due to indirect association through an intermediate variable (by virtue of the estimated slope $\hat{\delta}_2$ in the second-stage regression). In particular, a near-zero estimate for the second-stage intercept suggests that there is no direct association between Z and Y, whereas a nonzero (generally positive) intercept estimate is consistent with a direct association between Z and Y. Similarly, a near-zero estimate for the second-stage slope suggests absence of any indirect association between Z and Y, whereas a nonzero (generally positive) slope estimate suggests an indirect association between Z and Y (confounding). If the estimated second-stage intercept is positive and the estimated slope is close to zero, then the two-stage approach described by Schwartz and colleagues (*1–3*) would imply an association of Z with Y primarily through a direct pathway (i.e., statistical associations between Z and Y would not be substantially attributed to confounding).

A Class of Counterexamples

The situation becomes more complicated with the introduction of another pollutant variable. As before, the variable Y may be regarded as the health outcome. The variables X and Z, and a new variable W, represent pollutants. Suppose now that Z is not directly associated with the outcome Y and is only indirectly associated through X and W. Moreover, both X and W are assumed to have associations with Z that may vary from city to city. A structural equation model analogous to the system of Equation 1 – Equation 2 that expresses these relationships is

$$X = \gamma_0 + (\gamma_{1,\text{city}} \times Z) + \varepsilon_X \qquad [7]$$

$$W = \tau_0 + (\tau_{1,\text{city}} \times Z) + \varepsilon_W \qquad [8]$$

$$Y = \beta_0 + (\beta_2 \times X) + (\beta_3 \times W) + \varepsilon_Y. \qquad [9]$$

As before, one instance of the system will be used to represent each city. The parameters β_2 and β_3, which characterize the direct relationships of Y with X and Y with W, are treated as constant across all instances of the system, whereas the parameters $\gamma_{1,\text{city}}$ and $\tau_{1,\text{city}}$, which characterize the relationships between Z and X and between Z and W, respectively, are allowed to vary across instances (cities) but are to be regarded as constant within any instance of the system.

Assuming that Z is the variable to be evaluated for a direct effect, the first stage of the analysis involves a regression of Y against Z. Again, due to model misspecification the effects picked up should be that of Z acting through both X and W. The first-stage estimates of the Z effects should thus be characterized by

$$E[\hat{Z}\,\text{effect}_{\text{city}}] = (\beta_2 \times \gamma_{1,\text{city}}) + (\beta_3 \times \tau_{1,\text{city}}) \qquad [10]$$

for city = 1,...,K. Although the variable W is present in the modified system, the second-stage regression uses only one set of estimated copollutant relationships, either X with Z associations or W with Z associations; for consistency, we assume the former. The results of the second-stage regression will thus be affected by the relationship (if any) between $\tau_{1,\text{city}}$ and $\gamma_{1,\text{city}}$. For example, suppose that $\tau_{1,\text{city}} = \tau_1$ for city = 1,...,K. Then, in the second-stage regression of estimated Z effects against estimated X with Z associations, we should anticipate an intercept approximating $\beta_3 \times \tau_1$ and a slope approximating β_2. If $\beta_3 \times \tau_1 > 0$ and $\beta_2 > 0$, then we are likely to correctly conclude that Z has an indirect effect on Y (confounding), but incorrectly conclude that Z also has a direct effect on Y. The reason that a zero intercept would not be expected is that the confounding mechanism involves two intermediaries, with Z acting through one of them in varying proportion over cities, and through the other in constant proportion (absent the error terms). More generally, suppose that

$$\tau_{1,\text{city}} = \eta_0 + (\eta_1 \times \gamma_{1,\text{city}}) \qquad [11]$$

for city = 1,...,K. By substitution of Equation 11 into Equation 10, it follows that the expected Z effects from the first-stage regression are given by

$$E[\hat{Z}\,\text{effect}_{\text{city}}] = (\beta_3 \times \eta_0) + [\beta_2 + (\beta_3 \times \eta_1)] \times \gamma_{1,\text{city}}. \qquad [10']$$

The second-stage regression can then be expected to yield an intercept that approximates $\beta_3 \times \eta_0$ and a slope that approximates $\beta_2 + (\beta_3 \times \eta_1)$. The two-stage method would likely lead to the incorrect conclusion that both indirect and direct Z effects exist and would also incorrectly estimate the magnitude of the Z-to-X-to-Y pathway.

Simulations

We developed a program using SAS software (SAS Institute Inc., Cary, NC) that simulates data for a system which fits into the class of models described by Equations 7–9. The simulation is carried out for $K = 10$ cities and $N = 100$ observations (days) per city under the following parameter settings:

$\gamma_0 = 100$; $\gamma_{1,\text{city}} = 0.05, 0.15, ..., 0.95$ successively for city = 1,...,K

$\tau_0 = 100$; $\tau_{1,\text{city}} = \eta_0 + (\eta_1 \times \gamma_{1,\text{city}})$ $= 0.5 + (0.05 \times \gamma_{1,\text{city}})$ for city = 1,...,K

$\beta_0 = 100$; $\beta_2 = 0.3$; $\beta_3 = 1.0$.

Positive values were assigned to the intercept terms to minimize generation of negative quantities during simulation; however, the intercepts do not play an essential role in the analysis. Note that the relationship between $\tau_{1,\text{city}}$ and $\gamma_{1,\text{city}}$ is relatively flat, so that $\tau_{1,\text{city}}$ is nearly constant with respect to $\gamma_{1,\text{city}}$. For simplicity, the variable Z is simulated according to a normal distribution:

$$Z \sim N(100, 25^2).$$

Each simulated value of Z is substituted into Equations 7 and 8 to obtain the systematic

components of X and W. The values for X and W are completed with the addition of error terms ε_X and ε_W, simulated according to normal distributions:

$$\varepsilon_X \sim N(0, 20^2)$$

$$\varepsilon_W \sim N(0, 20^2).$$

The values for X and W are then substituted into Equation 9 to obtain the systematic component of Y. The value for Y is completed by addition of the error term ε_Y, which is simulated according to a normal distribution:

$$\varepsilon_Y \sim N(0, 20^2).$$

The simulation structure for Z, ε_X, ε_W, and ε_Y does not incorporate any dependencies, so that effectively variables are independent and for simplicity the time-series values are free of serial correlation. The magnitude of the error terms was selected to introduce enough randomness to make the simulation and estimation process meaningful while retaining substantial collinearity between Z, X, and W.

After the data have been simulated, the first-stage and second-stage regressions are performed with the variable W omitted from the analysis. Referring to Equation 10′ we anticipate that the second-stage intercept $\hat{\delta}_1$ should approximate $\beta_3 \times \eta_0 = 0.5$ and that the slope estimate $\hat{\delta}_2$ should approximate $\beta_2 + (\beta_3 \times \eta_1) = 0.35$.

The complete simulation was replicated 1,000 times. Over all simulations, the second-stage intercept estimate $\hat{\delta}_1$ had a mean value of 0.498 (SD = 0.073) and the second-stage slope estimate $\hat{\delta}_2$ had a mean value of 0.351 (SD = 0.125). On average, therefore, we recovered parameter estimates that conform to the underlying model structure. Moreover, the reported standard deviations indicate that nonpositive estimates of either the second-stage intercept or the second-stage slope would be rare.

Figure 1 shows an example of the results obtained in a single simulation run. This particular replication was selected for display because the second-stage regression recovered intercept and slope estimates close to the underlying quantities $\beta_3 \times \eta_0$ and $\beta_2 + (\beta_3 \times \eta_1)$. Based on the two-stage analysis, the (incorrect) conclusion would be that Z has a direct pathway to Y and also a comparatively weak indirect pathway to Y. Hence, the complete confounding of Z is largely obscured.

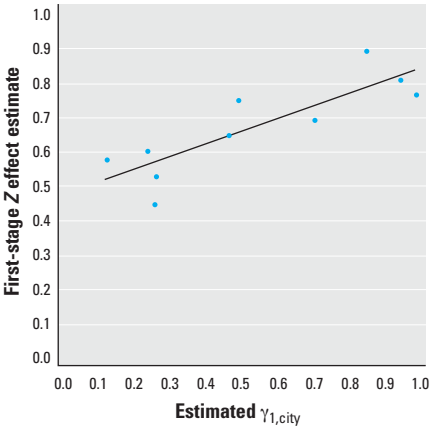

Figure 1. Results of one typical simulation run. The vertical axis represents the estimated Z effects and the horizontal axis represents the estimated values of $\gamma_{1,city}$ from the first-stage regressions. The fitted line represents the second-stage regression. For the second stage, the estimated intercept = 0.48 (SE = 0.050) and the estimated slope = 0.36 (SE = 0.081). Due to omission of an important variable (W), a positive intercept is estimated in the second stage and a direct effect is mistakenly attributed to Z.

Discussion

We have shown that two-stage regression approaches cannot necessarily be trusted in scenarios where a third factor, such as another air pollutant, may also play a role. Although the class of counterexamples presented is based on one particular model, many variations are possible. For example, introducing a direct association between Z and Y produces a new model that supports additional counterexamples. Depending on the parameters that characterize the relationships between variables in this revised model, different outcomes would be expected using the two-stage regression approach. In particular, if Z and the omitted variable W vary inversely, the direct effect of Z on Y could be partially or totally obscured. Hence the bias can go in either direction. More complicated examples involving multiple variables that are omitted from the estimation process can also be constructed, but such examples appear to offer little additional insight.

Samet et al.'s Figure 33 (*1*) and Schwartz's Figure 4 (*3*) are consistent with the hypothesis that the estimated PM_{10} (particulate matter < 10 μm in aerodynamic diameter) effects for hospital admissions and mortality, respectively, are not strongly confounded with sulfur dioxide and ozone. These figures are also not inconsistent with the hypothesis that the PM_{10} effect is due to another excluded air pollutant through mechanisms similar to those in the class of counterexamples presented above. Without necessarily subscribing to the logic of direct and indirect pathways implied by such models, the variable (PM_{10}) to be tested for confounding using the proposed methodology must nonetheless enter the analysis as the explanatory variable Z in the first-stage analysis. The two-stage models discussed in this paper, therefore, appropriately describe the type of models employed in the confounding analyses in previous reports (*1–3*).

In practice, multicollinearity is most effectively evaluated on the basis of entire correlation structures (e.g., as opposed to pairwise correlations), and factor analysis may be useful for this purpose. Published results providing adequate details on correlations between numerous pollutants in a multicity study, however, are not readily available. As a surrogate in this discussion, we consider results published by Schwartz (*2*) that show the correlation matrices for PM_{10}, carbon monoxide, temperature, and dew point for eight cities, but include no other air pollutants. Factor analyses of these four variables reveal considerable differences among the cities. The two smallest eigenvalues for each city are < 0.025, indicating strong multicollinearity among the four variables. In the four western cities (Colorado Springs, Colorado, and Seattle, Spokane, and Tacoma, Washington) and in New Haven, Connecticut, the first principal component explains 88–95% of the variation in these data, but includes both PM_{10} and CO, suggesting that these pollutants have similar sources and that their health effects may be difficult to separate. The second principal component in these cities explains < 13% of the variation, and reflects mainly the difference between PM_{10} and CO variations. By contrast, in three midwestern cities (Chicago, Illinois, and Minneapolis and St. Paul, Minnesota), the first principal component explains only 72–75% of the variation and puts much less weight on PM_{10} variations than on CO variations. The second principal component in the midwestern cities explains 26–27% of the variation, with PM_{10} the dominant variable, suggesting that it may be easier to separate the effects of the two pollutants in those cities than in western cities. These results suggest not only that multicollinearity is a significant problem in air pollution epidemiology but also that the nature of the problem does indeed vary from city to city. In short, the structural characteristics of the hypothetical models underlying the two-stage analyses reported by Schwartz and colleagues (*1–3*) and the more complicated counterexample we describe in this paper are quite plausible.

Factor analysis can also be useful for constructing alternative model inputs. Specifically, by obtaining factors that are often identifiable as "bundles" of closely related pollutants and that by construction are orthogonal, the multicollinearity problem can be largely eliminated. Although most recent efforts of this type have focused on factor analyses of concentrations of the elemental components of fine particles on

filters (*5–8*), they have also included gaseous copollutant concentrations as variables in the analyses. Mar et al. (*6*) used a factor analysis model that included SO_2, nitrogen dioxide, and CO along with particle elements for Phoenix, Arizona. Tsai et al. (*7*) used a model that included CO and sulfate concentrations, along with eight metals in respirable particles for three New Jersey cities. Özkaynak et al. (*8*) used a model with the coefficient of haze (an index of black particles) as well as NO_2 and CO and meteorologic variables for Toronto, Ontario, Canada. These models generally identified several important factors, or pollutant bundles, that suggested several significant sources contributed to urban air pollution, including motor vehicle emissions, coal combustion, fuel oil combustion, vegetative burning, resuspended road dust, soil and crustal material, local sources of SO_2, regional sulfate sources, nonferrous metal processing, and sea salt.

We suggest only that factor analysis is one useful tool for evaluating multicollinearity and is useful as a preprocessing step in regression modeling; we do not assert that it is the best or the only alternative to the proposed two-stage approach. A comprehensive treatment of the confounding/multicollinearity problem is beyond the scope of this paper, but clearly deserves additional study.

In this paper, we have purposely avoided the added complications associated with the "errors in variables" problem (*4,9*). In particular, if Z is measured with error, we may anticipate that attenuation bias will emerge in the first-stage regressions in situations properly characterized by Equations 1 and 2. A completely rigorous treatment would incorporate this aspect of modeling and estimation. However, dealing with such concerns in the present analysis would not appreciably alter the general conclusions and would introduce an unnecessary technical distraction.

Conclusion

We have given a brief background on the use of two-stage regression as it has been applied to the evaluation of direct and indirect associations of copollutants with human health outcomes. Selected counterexamples show, in simple idealized terms, mechanisms through which such analyses can lead to erroneous interpretations.

References and Notes

1. Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery DW, Schwartz J, Zanobetti A. The National Morbidity, Mortality, and Air Pollution Study Part II: Morbidity, Mortality, and Air Pollution in the United States. HEI Research Report no. 94, Part II. Cambridge, MA:Health Effects Institute, 2000.
2. Schwartz J. Air pollution and hospital admissions for heart disease in eight U.S. counties. Epidemiology 10:17–22 (1999).
3. Schwartz J. Assessing confounding, effect modification, and thresholds in the association between ambient particles and daily deaths. Environ Health Perspect 108:563–568 (2000).
4. Hanushek EA, Jackson JE. Statistical Methods for Social Scientists. New York:Academic Press, 1977.
5. Laden F, Neas LM, Dockery DW, Schwartz J. Association of fine particulate matter from different sources with daily mortality in six U.S. cities. Environ Health Perspect 108:941–947 (2000).
6. Mar TF, Norris GA, Koenig JQ, Larson TV. Associations between air pollution and mortality in Phoenix, 1995–1997. Environ Health Perspect 108:347–353 (2000).
7. Tsai FC, Apte MG, Daisey JM. An exploratory analysis of the relationship between mortality and the chemical composition of airborne particulate matter. Inhal Toxicol 12(suppl 2):121–135 (2000).
8. Özkaynak H, Xue J, Zhou H, Raizenne M. Associations between Daily Mortality and Motor Vehicle Pollution in Toronto, Canada. Boston:Department of Environmental Health, Harvard University School of Public Health, for Environmental Health Directorate, Health Canada, Ottawa, Ontario, Canada, 1996.
9. Fuller WA. Measurement Error Models. New York:John Wiley & Sons, 1987.