Model Selection and Health Effect Estimation in Environmental Epidemiology

Francesca Dominici, Chi Wang, Ciprian Crainiceanu, Giovanni Parmigiani

Correspondence to:
Francesca Dominici**,** PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street
Baltimore, MD 21205
Email: fdominic@jhsph.edu
Phone: 410-61451907
Fax: 410-9550958

Abstract word count: 243
Text word count: 1689

Abstract

In air pollution epidemiology, improvements in statistical analysis tools can translate into significant scientific advances, because of the unfavorable signal-to-noise ratios, and large correlations between exposures and confounders. Therefore, the use of a novel model selection approach in identifying time windows of exposure to pollutants that lead to adverse health effects is important and welcome. However, previous literature has raised concerns about approaches that select a model based on a given data set, and then estimate health effects in the same data assuming that the chosen model is correct. Problems can be particularly severe when: 1) the sample size is small for the magnitude of the true health effects to be detected; and 2) candidate predictors are highly correlated and likely to have a similar effect on the health outcome. Bayesian Model Averaging (BMA) has been advocated as a way of estimating health effects accounting for model uncertainty. However, BMA might not be as effective for effect estimation as it has proven to be for prediction. This is because posterior model probabilities might not reflect the ability of the model to provide an estimate of the health effect properly adjusted for confounding. In studies of air pollution and health, the focus should ideally be on estimating health effects, accounting for the uncertainty in the adjustment for confounding factors, especially when model choice and estimation are performed on the same data. However, the development of appropriate statistical tools remains an area of open investigation.

1. Introduction

In this issue of the *Journal,* Mortimer et al 2008 estimate the association between prenatal and lifetime exposures to air pollutants and pulmonary function measures in a cohort of children with asthma. They find large correlations between different pollutants as well as between different exposures time windows for the same pollutant. Therefore, rather than estimating the health effect of each exposure separately, they use a recently developed model selection procedure, the Deletion/Substitution/Addition algorithm (DSA) (Sinisi and van der Laan 2004) to identify the best predictive model, and base their conclusions about health effects on the model so selected. DSA is an iterative model-search algorithm, which optimizes global measures of prediction performance, here the mean squared error of residuals. Compared to stepwise model selection procedures, DSA has the advantages of being less sensitive to outliers, via the use of cross-validation during the search, and of allowing the search to move between statistical models that are not nested.

Air pollution epidemiology is an area where progress in statistical modeling strategies can translate into significant scientific advances, because of the unfavorable signal-to-noise ratios, and the number and correlation of both the exposures and the potential confounders exemplified in Mortimer et al. With this in mind, we welcome the thorough exploration of novel model selection approaches in a well conducted and statistically challenging study. However, we are also concerned about two potential issues: 1) is it safe to estimate health effects assuming that the best predictive model is correct? 2) How can one account for the uncertainty in the selection of population characteristics when estimating the association between an exposure and a health outcome?

2. Potential Pitfalls of Model Selection followed by Estimation

The authors' implementation of DSA is ambitious: it aims to identify which among: a) exposures X; b) population characteristics Z; c) functional form of the X; d) functional form of the Z; and e) interactions between the X and the Z lead to a model with the best predictive power. The chosen maximum model size is 10 terms, but the order of the interaction between covariates is up to 2, and the sum of the power of the polynomial function of each predictor is up to 3, leading to a large number of candidate terms for the algorithm to choose from. Health effect estimates and their variances are then obtained using GEE (see Table 3 of Mortimer et al) assuming that the selected model is correct.

Previous literature has shown that approaches that select a model based on a given data set, and then estimate health effects in the same data assuming that the chosen model is correct, can lead to misleading inferences. These approaches can identify exposure variables that appear to have strong predictive power even in randomly generated data, for which there are no true health effects (Raftery, Madigan, and Hoeting,1997 and Draper 1995). In addition, confidence intervals of regression coefficients calculated after model selection can have poor statistical properties (Benjamini and Yekutieli, 2005, Thomas et al 2005). These problems can be particularly severe when: 1) the sample size is small for the magnitude of the true health effects to be detected; and 2) the candidate predictors are highly correlated and likely to have a similar effect on the health outcome.

We illustrate these points using a simulation study modeled after Mortimer et al. For 232 subjects, we generate the 8 pollutant-specific metrics: prenatal and lifetime exposure to CO, NO2, O3, and PM10, from a multivariate normal distribution with mean zero, variance one, and correlation matrix as in Table 2 of Mortimer et al. This table does not include correlations between prenatal exposures and pollutants and thus we assumed that these correlations are the same as those estimated for lifetime exposures. We generate a continuous outcome, for example FVC, from a linear regression model with intercept 1.95 (from Table 1 of Mortimer et al). We include all 8 predictors, as linear terms, in the true model, and we assume that they all have the same regression coefficient of 0.1 (model 1). Finally we assume that errors are independent and identically distributed as normal with mean zero and variance 1. We then apply DSA using the inputs chosen by Mortimer et al (maxsize = 10, maxsumofpow = 3, maxorderint = 2, nsplits = 10). Table 1 shows the predictors that were identified by DSA and the estimates and 95% confidence intervals of the corresponding regression coefficients. Even though all predictors have the same predictive power, DSA selects only lifetime exposure to PM10 and prenatal exposure of NO2. The estimates of the health effects of these two pollutant metrics are severely biased upward, and their 95% confidence intervals do not include the true value of 0.1. In this example, the chosen model may be useful for prediction, but its interpretation in terms of etiology would be misleading in two ways: it would fail to indicate that the pollutants have similar effects, by singling out one pollutant as predictive, and it would overstate the effect of that pollutant on the outcome. We repeated the simulation study assuming that the true regression

model not only included all the linear terms, but also the quadratic and cubic terms of lifetime exposure to CO (model 2) with true regression coefficients of 0.01 and 0.001, respectively. In this scenario (see Table 2), DSA selects the linear terms of prenatal exposures to O3 and NO2 and the linear terms of lifetime exposure to NO2 and O3. Again, the estimated regression coefficients of these four predictors are biased and their 95% confidence intervals do not include the true values. Although we recognize that we may have simulated data from a somewhat unfavorable, though not unrealistic, situation, these simulations results suggest caution in the interpretation of the results of Mortimer et al. A critical consideration in this discussion is the relation of the sample size to the number of candidate predictors. If we generate data with a 100-fold sample size of 23200 under model 1, then DSA identifies all the correct predictors and provides unbiased estimates of the corresponding regression coefficients (Table 3).

3. Model Uncertainty versus Adjustment Uncertainty

The challenges highlighted in our simulations emphasize the importance of accounting for uncertainty arising from the variable selection stage when reporting health effects estimates. Bayesian Model Averaging (BMA) (George and McCulloch, 1993; Draper, 1995; George and Clyde, 2004) has been advocated as a way of estimating health effects accounting for model uncertainty. BMA treats the true model as an unknown random variable and estimates health effects by a weighted average of model-specific health effects estimates using the posterior model probabilities as weights (Clyde 2000, Koop and Tole 2004). In Crainiceanu et al. 2008 we have demonstrated that BMA might not be as effective for effect estimation as it has proven to be for prediction. In practice this can happen because the posterior model probabilities might not reflect the ability of the model to provide an estimate of the health effect properly adjusted for confounding. For example, large weights may be assigned to models that do not adequately adjust for confounders, leading to a biased estimate of the health effect. For example, consider an exposure X, an outcome Y, and two covariates Z1 and Z2. Assume that Z1 is independent from X, but a good predictor of Y, and that Z2 is highly correlated with X but a poor predictor of Y. Table 4 summarizes the three possible models and hypothetical weights that would be assigned based on both predictive ability and ability to estimate the health effects properly adjusting for confounding. BMA is likely to assign a high weight to models that do not include Z2 and therefore would provide a biased estimate of the health effect. Because the goal of inference is to obtain an estimate of the health

effect, the confounder Z2 needs to be included into the model. Standard BMA can over- or under-estimate health effects, depending on the correlations of the variables involved. Here, when BMA is applied to the simulated examples above, the health effect estimates are higher than the true ones for 6 out of 8 pollutants. The 95% posterior intervals are larger than DSA, with 3 of the 8 intervals including 0, but 2 of the 8 intervals are still failing to include the true value of 0.1. We used the implementation in the package in R with default settings, constraining exposures to be included in all models.

Our view is that in studies of air pollution and health, it is important to focus on estimating health effects that are properly adjusted for all the confounding factors (Dominici et al 2004), including exposure to other pollutants. Ideally, adjustment uncertainty should be fully incorporated into statistical inference whenever estimation is sensitive to model choice, especially when model choice and estimation are performed on the same data. However, the development of appropriate statistical tools to achieve this goal is still in progress. In the case of a single exposure, in Crainiceanu et al. 2008 and in Wang et al 2008, we discuss an approach to estimate a health effect accounting for uncertainty in the confounding adjustment. Similar methods for multiple exposures, as would have been required in Mortimer et al, are not available.

4. Concluding thoughts

Mortimer et al. present interesting results on the association between lifetime and prenatal exposure to air pollution and pulmonary functions in a cohort of asthmatic children. The authors used an innovative model search approach to select the exposure variables that lead to the best predictive model. In doing so they faced a challenge common in environmental epidemiology, where the health effects to be estimated are small and both the exposure variables and the potential confounders are correlated. In this context, using ambitious model selection methods that can search efficiently through a large number of possible models may illuminate on important novel directions for etiological research. However, inference that ignores the biases and uncertainty in the selection of predictors may lead to inflated effects and overly optimistic estimates of precision. Therefore, unless the sample size far exceeds the number of potential terms in the model, findings of this type of analysis require further validation.

Acknowledgements. Funding for Dr. Dominici was provided by the U.S. Environmental Protection Agency (RD-83241701). Funding for Drs Dominici and Crainiceanu was also provided by the National Institute for Environmental Health Sciences (ES012054-03) and by the NIEHS Center in Urban Environmental Health (P30 ES 03819). Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency through grant agreement # RD-83241701 to Johns Hopkins University, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

References

Mortimer K, Neugebauer R, Lurmann F, Alcorn S, Balmes J, Tager I. The Effect of Prenatal and Lifetime Exposure to Air Pollution on the Pulmonary Function of Asthmatic Children. *Epidemiology.* 2008*; ??*:??-??.

Sandra SE and van der Laan, Mark J. Deletion/Substitution/Addition Algorithm in Learning with Applications in Genomics. *Statistical Applications in Genetics and Molecular Biology*. 2004; Vol. 3 : Iss. 1, Article 18. Available at: http://www.bepress.com/sagmb/vol3/iss1/art18**.**

Raftery AE, Madigan D, and Hoeting J.A. Bayesian model averaging for linear regression models*. Journal of the American Statistical Association.* 1997; 92:179–191.

Draper D. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B.* 1995*;* 57**:** 45-97.

Benjamini Y,Yekutieli D. False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters. *Journal of the American Statistical Association*. 2005;11:71-81.

Thomas DC, Witte JS, and Greenland S. Dissecting Effects of Complex Mixtures: Who's Afraid of Informative Priors? *Epidemiology*. 2007; 18: 186-190.

George E. and Clyde M. Model Uncertainty. *Statistical Science*. (2004); 19: 81–94.

George E. and McCulloch R.Variable selection via Gibbs sampling. *Journal of the American Statistical Association*. 1993;88:881–889.

Clyde M. Model Uncertainty and health effects studies for particulate matter. *Environmetrics*. 2000; 11:745–763.

Koop G. and Tole, L. Measuring the health effects of air pollution: to what extent can we really say that people are dying of bad air. *Journal of Environmental Economics and Management. 2004:* 47: 30–54.

Crainiceanu C, Dominici F, and Parmigiani G. Adjustment Uncertainty in Effect Estimation. *Biometrika*. 2008 (to appear). Technical report available at http://www.bepress.com/jhubiostat/paper89/

Dominici F, McDermott A, and Hastie T. Improved Semi-parametric Time Series Models of Air Pollution and Mortality. *Journal of the American Statistical Association*. 2004; 468: 938–948.

Wang C, Parmigiani G, Crainiceanu CM, Dominici F. A Bayesian Approach to Effect Estimation Accounting for Adjustment Uncertainty. Technical report available at http://www.bepress.com/jhubiostat/paper157/

Table 1: Selected predictors and point estimates and 95% CI of the corresponding regression coefficients, data are generated from model 1, sample size =232.

Predictor	Point Estimate	95% C.I.
$L.PM_{10}$	0.320	(0.188, 0.453)
$P.NO_2$	0.274	(0.143, 0.405)

Table 2: Selected predictors and point estimates and 95% CI of the corresponding regression coefficients, data are generated from model 2, sample size =232.

Predictor	Point Estimate	95% C.I.
$P.O_3$	0.312	(0.159 0.465)
$P.NO_2$	0.239	(0.067 0.411)
$L.NO_2$	0.280	(0.105 0.456)
$L.O_3$	0.301	(0.145 0.456)

Table 3: Selected predictors and point estimates and 95% CI of the corresponding regression coefficients, data are generated from model 1, sample size =23200.

Predictor	Point Estimate	95% C.I.
$L.PM_{10}$	0.099	(0.080 0.119)
$L.NO_2$	0.103	(0.078 0.127)
$P.O_3$	0.102	(0.084 0.120)
$L.CO$	0.102	(0.081 0.124)
$P.CO$	0.120	(0.098 0.142)

$P.PM_{10}$	0.113	(0.095 0.132)
$L.O_3$	0.090	(0.072 0.109)
$P.NO_2$	0.088	(0.063 0.112)

Table 4: Toy example

Regression Models	Weights based on prediction	Weights based on estimation
$y = \beta x + \alpha_1 z_1$	0.9	0.0
$y = \beta x + \alpha_2 z_2$	0.0	0.9
$y = \beta x + \alpha_1 z_1 + \alpha_2 z_2$	0.1	0.1