THEORY AND METHODS

A bootstrap method to avoid the effect of concurvity in generalised additive models in time series studies of air pollution

Adolfo Figueiras, Javier Roca-Pardiñas, Carmen Cadarso-Suárez

J Epidemiol Community Health 2005;**59**:881–884. doi: 10.1136/jech.2004.026740

See end of article for authors' affiliations

Correspondence to: Dr A Figueiras-Guzmán, Dto de Medicina Preventiva y Salud Pública, Facultad de Medicina, c/ San Francisco s/n, 15705 Santiago de Compostela (A Coruña), Spain; aldolfo.figueiras@usc.es

Accepted for publication 3 March 2005

Background: In recent years a great number of studies have applied generalised additive models (GAMs) to time series data to estimate the short term health effects of air pollution. Lately, however, it has been found that concurvity—the non-parametric analogue of multicollinearity—might lead to underestimation of standard errors of the effects of independent variables. Underestimation of standard errors means that for concurvity levels commonly present in the data, the risk of committing type I error rises by over threefold.
Methods: This study developed a conditional bootstrap methology that consists of assuming that the outcome in any observation is conditional upon the values of the set of independent variables used. It then tested this procedure by means of a simulation study using a Poisson additive model. The response variable of this model is a function of an unobserved confounding variable (that introduces trend and seasonality), real black smoke data, and temperature. Scenarios were created with different coefficients and degrees of concurvity.
Results: Conditional bootstrap provides confidence intervals with coverages close to nominal (95%), irrespective of the degree of concurvity, number of variables in the model or magnitude of the coefficient to be estimated (for example, for a concurvity of 0.85, bootstrap confidence interval coverage is 95% compared with 71% in the case of the asymptotic interval obtained directly with S-plus gam function).
Conclusions: The bootstrap method avoids the problem of concurvity in time series studies of air pollution, and is easily generalised to non-linear dose-risk effects. All bootstrap calculations described in this paper can be performed using S-Plus gam.boot software.

In recent years, time series studies with Poisson regression using generalised additive models (GAMs)[1–4] have been the reference method for analysing the short term health effects of air pollution. Lately, however, these models have been shown to suffer from an important limitation, namely, that there is underestimation of the standard error (SE) of the estimated effect of any given pollutant in those cases where concurvity is present in the data.[5–8]

Briefly, concurvity is the non-parametric analogue of multicollinearity, in which a function of one of the independent variables can be approximated by a linear combination of functions of the remaining variables, with these functions being estimated in the same way as the corresponding functions in the original model.[6] It has been seen that, in the presence of concurvity in GAMs, confidence intervals (CI) are too narrow, p values are understated, and type I error rate is greater than that established a priori[6 8] (for example, for a concurvity of 0.6—a value based on real data—type I error has shown a rise of almost threefold[6]). One way of eliminating the problem of underestimation of SE in the presence of concurvity is to have recourse to bootstrap resampling techniques. Accordingly, in this paper we propose the use of a conditional bootstrap method in GAMs to calculate valid CI for the estimated effect of any given pollutant. For the purpose of performing such calculations, we have developed a routine, termed gam.boot.

METHODS

Conditional bootstrap

In this type of bootstrap,[9–13] B bootstrap replicates are generated. In each of these, the values of the independent variables are the same as those of the observed data, with only the values of the response variable being varied from replicate to replicate. The value assumed by the outcome in each observation is conditional (hence the technique's name) upon the values of the set of independent variables in said observation. To this end, a Poisson model is first applied to the data for the sample, and the coefficient and predictions of the model are then obtained for each observation (also known as pilot estimates, as these are taken as reference).

To illustrate this procedure, it is as well to start with a simple model (see expression 1) having only two covariates $X = (X_1, X2)$, one of which, X_1, is linearly related to the response via parameter β_1 (the parameter of interest), and the other, X_2, is related to the response via an unknown smooth function, f. Based on n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, with $X_i = (X_{i1}, X_{i2})$, the following model is then fitted

$$\begin{aligned} Y_i &\sim Poisson(\mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 X_{i1} + f(X_{i2}) \end{aligned} \qquad (1)$$

and thereafter the estimated coefficient $\hat{\beta}_1$ and predictions $\hat{\mu}_1, \ldots \hat{\mu}_n$ are obtained for each of the observations.

In a second step, the B conditional bootstrap replicates from b = 1 to B (for example, B = 1000) are generated, so that the values of the dependent variable, Y_i, in each observation follow a random Poisson distribution with a mean of $\hat{\mu}_i$, that is to say, $Y_i^{*b} \sim Poisson\ (\hat{\mu}_i)$. In each of these B replicates, an estimate of the coefficient $\hat{\beta}_1^{*b}$ is obtained, using a GAM with the same number of degrees of freedom as the original model.

Finally, the $100percent \times (1-\alpha)$ limits for the confidence interval of β_1 are given by $(2\hat{\beta}_1 - \hat{\beta}_1^{1-\alpha/2}, 2\hat{\beta}_1 - \hat{\beta}_1^{\alpha/2})$ where $\hat{\beta}^p$

Abbreviations: GAM, generalised additive models; BS, black smoke

represents the percentile p of the bootstrapped estimates $\hat{\beta}_1^{*1},\ldots,\hat{\beta}_1^{*B}$.

Bootstrap validation: a simulation study

Scenario

A three year (1096 day) series was constructed in which the number of events per day followed a Poisson distribution, in line with model (2) below,

$$Y_1 \sim \text{Poisson}(\mu_i)$$
$$\log(\mu_i) = \log(22) + \beta_1 BS + \beta_2 trend + f_{temp}(temp_t) \quad (2)$$

where Y_t denotes the number of events per day, BS_t is the black smoke, and $temp_t$ is the temperature recorded on day t. The coefficient β_1 denotes the log relative rate of Y associated with increase in black smoke (BS). The BS is genuine and was drawn from data recorded for the city of Vigo (north west Spain) over the period 1996–1998 (see fig 1A). By default, the true β_1 had a value of 0.001, although other values for this parameter (from 0.001 to 0.008) were also considered to ascertain the effect of coefficient magnitude on the 95% CI coverage and bias in the estimated coefficient, $\hat{\beta}_1$.

In model (2), *trend* is a function that introduces trend and seasonality into the simulated data to simulate an unobserved confounding variable in the data. In ecological time series studies, smooth functions (smoothing splines or LOWESS) of the time variable [1–4] are used to control for the confounding effect of these unobserved confounding variables. The *trend* function is a sinusoidal function (see fig 1B) that had already been used in other simulation studies.[14–16] In (2), as in such studies, β_2 assumes the value of 0.1. The function f_{temp} represents the functional form of the effect of temperature on mortality in Vigo (see fig 1C).

Concurvity generation

In our simulation study, the effect of interest was BS. The concurvity measure proposed by Ramsay *et al*[6] was based on the correlation between the fitted values obtained from the GAM with pollution, BS_t, as the response, and time and temperature as the predictors. Specifically, to assess the degree of concurvity in our data, one should: (a) fit the GAM

$$BS_t = g_1(t) + g_2(temp_t) + \varepsilon_t \quad (3)$$

where the partial functions g_1 and g_2 are estimated using smoothing splines with 7 and 4 degrees of freedom respectively, and ϵ_t is a zero mean error variable; and (b) then compute the squared correlation (R^2) between BS_t and, using the fitted values from (3), namely:

$$Concurvity = \left(\text{correlation}\left(BS_t, \widehat{BS}_t = \hat{g}_1(t) + \hat{g}_2(temp_t)\right)\right)$$

In our real data, concurvity was 0.29. To vary the concurvity and thereby assess its influence on CI coverage and bias in the parameter estimate, $\hat{\beta}_1$, in model (2), the original BS was replaced by a new variable

$$BSs_t = k_1 BS_t + k_2 \widehat{BS}_t + (1 - k_1 - k_2)\hat{\varepsilon}_t$$

where $\hat{\epsilon}_t = BS_t - \widehat{BS}_t$, k_1 and k_2 were constants, such that $0 \leqslant k_1, k_2 \leqslant 1$ and $0 \leqslant k_1 + k_2 \leqslant 1$. Note that BSs_t is a combination of the original BS_t of the estimate $\widehat{BS}_t$ and the errors $\hat{\epsilon}_t = BS_t - \widehat{BS}_t$ obtained from the model (3). Thus, by varying the constants k_1 and k_2, the degree of concurvity between BSs and the remaining independent variables may easily be modified, from 0 ($k_1 = k_2 = 0$) through 1 ($k_1 = 0$, $k_2 = 1$).

Data analysis

In our study, a number of scenarios were defined using different values for the coefficient β_1 (from 0.001 to 0.009) and different values for k_1 and k_2, so that the degree of concurvity varied from 0.05 to 0.65. In each of the scenarios, 1000 samples $\{t, temp_t, Y_t\}$ were generated with 1096 points in time (days) each ($t = 1,\ldots,1096$) based on the model (2). In each sample, $\hat{\beta}_1$ was obtained on the basis of the estimated model (4)

$$\log(\hat{\mu}_t) = \hat{\beta}_0 + \hat{\beta}_1 BS + \hat{f}_{trend}(t) + \hat{f}_{temp}(temp_t) \quad (4)$$

Lastly, the corresponding bootstrap and asymptotic 95% CI were calculated for β_1. The CI coverages were calculated as the proportion of samples in which these included the true β_1.

The estimates in (4) were obtained by using smoothing splines with 7 degrees of freedom per year for the estimated trend effect, $\hat{f}_{trend}$ (following Dominici *et al*),[5] and 4 degrees of freedom for estimated temperature effect, $\hat{f}_{temp}$.

RESULTS

Figure 1A shows the BS sequence for the city of Vigo, which was used as the independent variable to generate the replicates; fig 1B shows the unobserved confounding variable having trend and seasonality; fig 1C shows the functional form of the relation between temperature and mortality, which was obtained from previous estimates in a relation observed for Vigo.

In table 1, the CI coverages are compared for different degrees of concurvity using asymptotic GAM and GAM bootstrap. It will be seen that, with the asymptotic approach, increases in concurvity were accompanied by a progressive decline in coverage (see fig 2A), which decreased to as low as 80% when concurvity was 0.6. With the bootstrap method, however, coverage remained above 94% throughout, irrespective of the degree of concurvity. The coverage was also evaluated for different values of the true coefficient. With regard to the coefficient's influence on coverage (see fig 2B), coverage proved seemingly independent of the magnitude of the coefficient, for the asymptotic and bootstrap methods alike.

DISCUSSION

The results of our study show that application of conditional bootstrap techniques could enable the CI for the effect of pollutants in time series studies to be calculated with optimal coverage, regardless of the degree of concurvity, number of covariates in the model or magnitude of the coefficient to be estimated.

The publication of two scientific papers a few months ago, the first by Dominici *et al*[5] and the second by Ramsay *et al*,[6] showed that standard errors are considerably underestimated in the presence of concurvity.[7 8] To eliminate the problem of concurvity, flexible regression methods can be used, such as parametric cubic splines[6] or penalised splines, which do not require the use of backfitting to estimate model components. However, parametric cubic splines might not provide sufficient flexibility to capture the trend or seasonality of the series, as the knots have to be established a priori. In this respect, penalised splines[17 18] might afford greater flexibility, although, as far as we are aware, no assessment has yet been made of the performance of such smoothers in the context of time series studies.

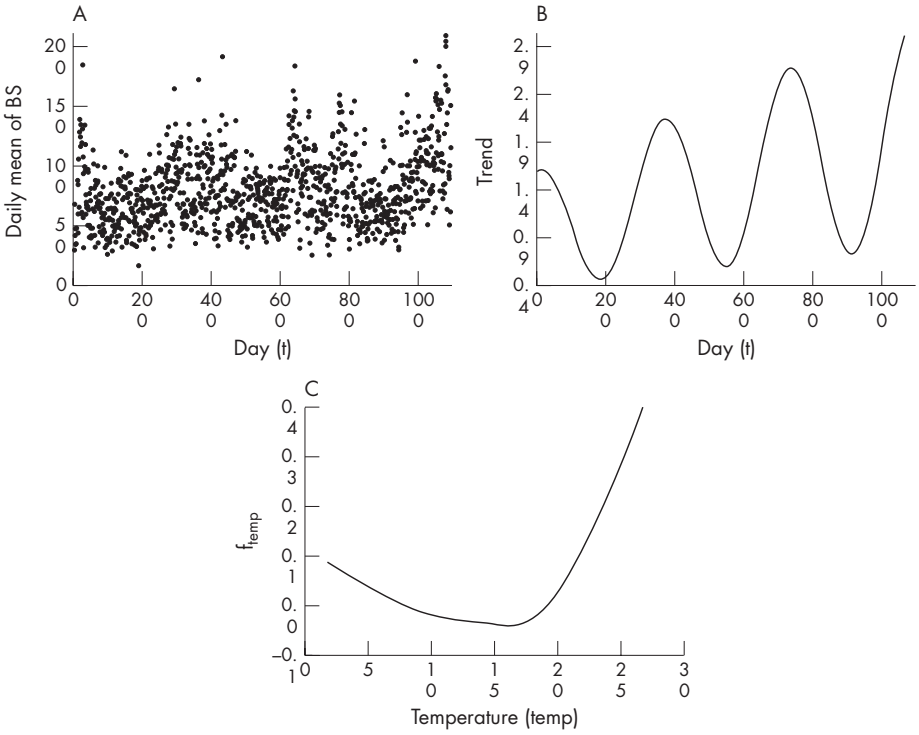

Figure 1 Basis of the simulation. (A) Daily mean black smoke (BS) levels in Vigo for 1096 days, in the period 1996–1998. (B) Effect of unobserved confounding variable that has a seasonal and trend component. (C) Smoothed temperature effect (f_{temp}).

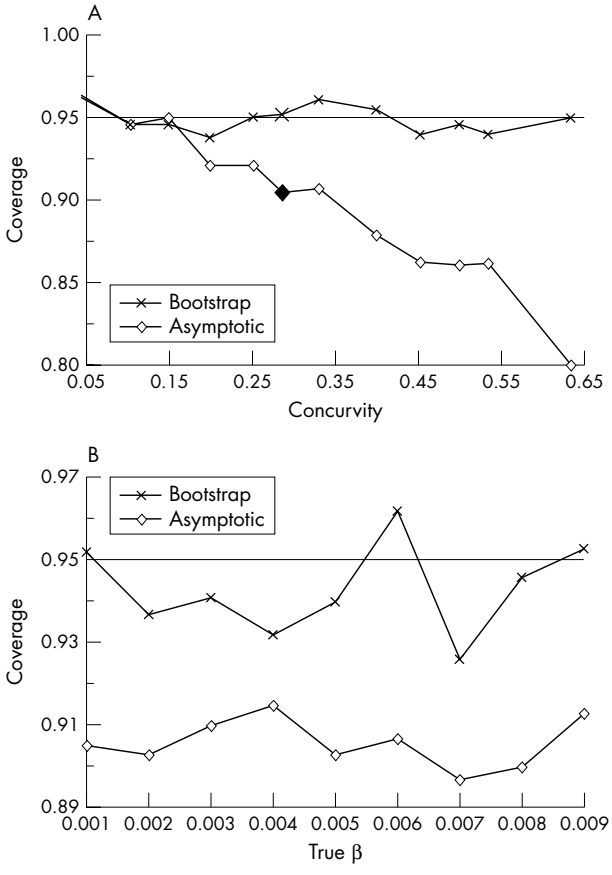

Figure 2 Effect of concurvity (A) and magnitude of coefficient associated with black smoke (β_1) on 95% CI coverage* (B). *Coverage = % of replicates that include the true value of the coefficient within the 95% CI of the true coefficient 0.001. Nominal coverage is 0.95.

To eradicate the problem of concurvity, consideration has been given to the application of alternative designs, such as that of case-crossover, which control trend and seasonality by means of design and thereby eliminate the effect of concurvity on the estimation of standard errors.[16] Dominici *et al*[19] have introduced a closed form estimate of the asymptotically exact covariance matrix of the linear component of a GAM, and offer a development of the gam.exact package, which is an extended version of gam.[20] However, it has also been observed that, in given conditions (with many smoothing functions included in the model), the routine gam.exact can give rise to standard errors very much greater than those that would result from application of parametric techniques.

In this paper, we propose an alternative method based on conditional bootstrap resampling techniques, which provides

Table 1 Effect of concurvity on confidence interval coverage*. Results of the simulation study using asymptotic (standard method in gam software) and conditional bootstrap. True coefficient = 0.001

	Coverage*	
Concurvity	**Asymptotic**	**Bootstrap**
0.045	0.962	0.964
0.103	0.946	0.946
0.149	0.950	0.946
0.199	0.921	0.938
0.251	0.921	0.951
0.286†	**0.905**	**0.952**
0.329	0.907	0.961
0.399	0.879	0.955
0.451	0.863	0.940
0.498	0.861	0.946
0.533	0.862	0.940
0.632	0.801	0.950

*Percentage of replicates that include the true value of the coefficient within the 95% CI of the true coefficient 0.001. Nominal coverage is 0.95. †Degree of concurvity for our real data.

What this paper adds

Concurvity—the non-parametric analogue of multicollinearity—leads to an important underestimation of standard errors, which means that the risk of committing type I error rises by over threefold for concurvity levels commonly present in the data. We propose a method based on conditional bootstrap resampling techniques, which provides optimal coverages of the confidence intervals of the estimated effect in generalised additive models (GAMs) applied to time series data. This methodology affords the following additional advantages over a previous method: (1) our method is applicable to models with any type of smoother (smoothing splines, LOWESS, natural splines, penalised splines) and any combination of same; (2) this approach is easily generalised to non-linear dose-risk relations; and (3) our method could also be extended to the calculation of CI for possible interactions between pollutants and/or climatological variables.

Policy implications

- Up until 2002, most studies that assessed the effects of pollution on health were time series studies that used generalised additive models.
- In 2002, however, it was shown that, because of concurvity (the analogue of collinearity), false statistically significant associations could be found at a frequency more than three times higher than that established a priori (for example, type I error of 18% compared with 5%).
- This could imply that some of the findings reported to date on the health related effects of pollution might be erroneous. Accordingly, in time series studies that use generalised additive models, it would be advisable for such data to be assessed for the presence of concurvity and for methods such as that developed in this paper to be applied, to prevent false associations being found between pollution and health related effects.

optimal coverages and affords the following additional advantages: (1) unlike the gam.exact method, which requires the smooth functions to be of the smoothing-spline type, gam.boot is general, in that it is applicable to models with any type of smoother, such as smoothing splines, LOWESS, natural splines, and any combination of same; (2) although our bootstrap method was initially developed to construct CI for coefficients, this approach is easily generalised to non-linear dose-risk relations. Thus, non-linear dose-response relations can be evaluated for pollutants or climatological variables; and lastly, (3) our method could also be extended to the calculation of CI for possible interactions between pollutants and/or climatological variables.

Furthermore, a recent study seems to show that, in cases where the results for a number of cities are combined using hierarchical models, underestimation of standard errors in each of the cities seems to have little effect on the overall results under most conditions.[21] In this regard, our method could be generalised to a hierarchical bootstrap method that would start on the basis of standard errors correctly calculated at the first level (city), and then proceed to estimate the overall effects at a multicity level.

While the main limitation of our method might possibly lie in the computational cost entailed, we do not regard this as too high. Furthermore, there is no need to apply the bootstrap method at each stage of building the model: suffice to apply it once the basal model has been constructed and the final effect of each pollutant is to be ascertained.

Indeed, the conditional bootstrap method proposed in this paper could enable the community of air pollution researchers to make a satisfactory calculation of CI in dose-response relations, whether linear or non-linear. To this end, the S-plus gam.boot software, a user friendly application, will be made freely available to readers on request (roca@uvigo.es).

ACKNOWLEDGEMENTS

The authors are grateful to Dr Marc Saez for his advice and helpful comments, and to Michael Benedict for his help with the translation of the manuscript.

Authors' affiliations

A Figueiras, Department of Preventive Medicine, University of Santiago de Compostela, Spain
J Roca-Pardiñas, Department of Statistics and Operations Research, University of Vigo, Spain
C Cadarso-Suárez, Unit of Biostatistics, Department of Statistics and Operations Research, University of Santiago de Compostela

Funding: Dr Adolfo Figueiras' work on this project was funded by Health Research Fund (Fondo de Investigación Sanitaria) grants 00/0010-05 and 99/1189 from the Spanish Ministry of Health, Javier Roca-Pardiñas' work was funded by a grant from the University of Vigo (Vigo, Spain), and Dr Carmen Cadarso-Suárez's work was funded by grant BMF2002-03213 from the Spanish Ministry of Science and Technology.

Conflicts of interest: none.

REFERENCES

1 **Hastie TJ**, Tibshirani RJ. *Generalized additive models*. London: Chapman and Hall, 1990.
2 **Saez M**, Ballester F, Barcelo MA, *et al*. A combined analysis of the short-term effects of photochemical air pollutants on mortality within the EMECAM project. *Environ Health Perspect* 2002;**110**:221–8.
3 **Le Tertre A**, Medina S, Samoli E, *et al*. Short term effects of particulate air pollution on cardiovascular diseases in eight European cities. *J Epidemiol Community Health* 2002;**56**:773–9.
4 **Samet JM**, Dominici F, Curriero FC, *et al*. Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *N Engl J Med* 2000;**343**:1742–9.
5 **Dominici F**, McDermott A, Zeger SL, *et al*. On generalized additive models in time series studies of air pollution and health. *Am J Epidemiol* 2002;**156**:193–203.
6 **Ramsay TO**, Burnett RT, Krewski D. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology* 2003;**14**:18–23.
7 **Samet JM**, Dominici F, McDermott A, *et al*. New problems for an old design: time series analyses of air pollution and health. *Epidemiology* 2003;**14**:11–12.
8 **Lumley T**, Sheppard L. Time series analyses of air pollution and health: straining at gnats and swallowing camels? *Epidemiology* 2003;**14**:13–14.
9 **Efron E**, Tibshirani RJ. *An introduction to the bootstrap*. London: Chapman and Hall, 1993.
10 **Kauermann G**, Opsomer JD. Local likelihood estimation in generalized additive models. *Scand J Stat* 2003;**30**:317–37.
11 **Roca-Pardiñas J**, González-Manteiga W, Febrero-Bande M, *et al*. Predicting binary time series of SO2 using generalized additive models with unknown link function. *Environmetrics* 2004;**15**:729–42.
12 **Härdle W**, Mammen E. Comparing nonparametric versus parametric regression fits. *Ann Statist* 1993;**21**:1926–47.
13 **Roca-Pardiñas J**, Cadarso-Suárez C, González-Manteiga W. Testing for interactions in generalized additive models: applications to SO2 pollution data. *Stat Comput*, (in press).
14 **Bateson TF**, Schwartz J. Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures. *Epidemiology* 1999;**10**:539–44.
15 **Navidi W**, Weinhandl E. Risk set sampling for case-crossover designs. *Epidemiology* 2002;**13**:100–5.
16 **Figueiras A**, Carracedo-Martínez E, Saez M, *et al*. Analysis of case-crossover designs using longitudinal approaches: a simulation study. *Epidemiology* 2005;**16**:239–46.
17 **Eilers PHC**, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996;**11**:89–121.
18 **Ruppert D**, Wand MP, Carroll RJ. *Semiparametric regression*. Cambridge: Cambridge University Press, 2003.
19 **Dominici F**, McDermott A, Hastie TJ. *Issues in semiparametric regression: a case study of time series models in air pollution and mortality*. Baltimore, MD: Department of Biostatistics, Johns Hopkins University, 2003.
20 **Dominici F**, McDermott A, Hastie TJ. Software for computing the asymptotically exact standard errors in GAM. http://www.ihapss.jhsph.edu/software/gam.exact/gam.exact.htm.
21 **Daniels MJ**, Dominici F, Zeger S. Underestimation of standard errors in multi-site time series studies. *Epidemiology* 2004;**15**:57–62.