Statistical Methodological Issues in Studies of Air Pollution and Respiratory Disease

Bircan Erbas[1] and Rob Hyndman[2]

[1] Department of Public Health, The University of Melbourne, VIC 3010, Australia
[2] Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia

Abstract: Epidemiological studies have consistently shown short term associations between levels of air pollution and respiratory disease in countries of diverse populations, geographical locations and varying levels of air pollution and climate. The aims of this paper are: (1) to assess the sensitivity of the observed pollution effects to model specification, with particular emphasis on the inclusion of seasonally adjusted covariates; and (2) to study the effect of air pollution on respiratory disease in Melbourne, Australia.

Keywords: air pollution, autocorrelation, generalized additive models, respiratory disease, seasonal adjustment.

1 Introduction

1.1 Background

The adverse effects of air pollution on respiratory disease have been widely documented in countries of diverse populations, geography and climate. Recently, there has been some effort to determine the replicability of these findings across a range of exposure outcomes. For example, the APHEA (Air Pollution and Health, a European Approach) produced a standard protocol designed to assess replicability across different countries (Katsouyanni et al. 1996).

We extend this work on replicability by examining the robustness of the estimated relationships between air pollution and respiratory disease under different statistical models. The work is motivated by the idea that applications of different statistical models with varying underlying methodological assumptions may lead to different conclusions regarding the air pollution and respiratory disease relation.

1.2 Data

COPD (Chronic Obstructive Pulmonary Disease) and asthma hospital admissions from all short-stay acute public hospitals in Melbourne, registered

on a daily basis by the Health Department of Victoria, were used as response variables for the period 1 July 1989 to 31 December 1992. International Classification of Disease (ICD) codes for COPD (490–492, 494, 496) and asthma (493) were used to define COPD and asthma.

Air pollution data were obtained from the Environment Protection Authority (EPA). Maximum hourly values were averaged each day across nine monitoring stations in Melbourne, for nitrogen dioxide, sulfur dioxide, and ozone, all measured in parts-per-hundred-million (pphm). Particulate matter was measured by a device which detects back-scattering (B_{scat}) of light by visibility-reducing particulates between 0.1 and 1μm in diameter. Air particles index (API) were derived from $B_{scat} \times 10^{-4}$. Meteorological data include three hourly maximum daily levels of relative humidity, dry bulbs temperature and dew point temperature. The measures were averaged across four monitoring stations in the Melbourne area.

1.3 Statistical Methodological Issues

A key issue which arises in studies of respiratory disease and pollution is controlling for seasonal variation. Several variables may be confounded with seasonality, leading to some possible spurious pollution effects.

To assess the strength and magnitude of seasonal variation in the pollutants and climatic variables, we utilise a method of seasonal adjustment called STL (Seasonal-Trend decomposition based on Loess smoothing) developed by Cleveland and Terpenning (1982). Covariates exhibiting strong seasonality were adjusted with the STL method and the resulting seasonally adjusted series were used in subsequent analysis.

We explore the robustness of the pollution-respiratory disease relation using a variety of regression type approaches, controlling for secular trends, seasonality, and confounding effects of climate. These models include: (1) Generalized Linear Models (GLM); (2) Generalized Additive Models (GAM); (3) Parameter Driven Poisson Regression Models (PDM); and (4) Transitional Regression Models (TRM). In each case, we consider models based on a Poisson distribution, incorporating over-dispersion and serial correlation where possible.

2 Statistical Models

2.1 Generalized Linear Models

For a Generalized Linear Model (GLM) with a log link function, we specify the expectation of a random variable Y_t as

$$\mathrm{E}(Y_t|\boldsymbol{X}_t) = \exp\left(\beta_0 + \sum_{i=1}^{r} \beta_i X_{t,i}\right). \tag{1}$$

Refer to McCullagh and Nelder (1989) for a detailed discussion of GLMs.

Here Y_t denotes daily counts of respiratory disease and air pollution and $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})'$ denotes the explanatory variables at time t. We assume an overdispersed Poisson model, estimated using a quasi-likelihood approach. Akaike's Information Criterion, AIC (Akaike, 1973) was used for variable selection.

2.2 Generalized Additive Models

A nonparametric alternative to the parametric GLM is the Generalized Additive Model (GAM). GAMs allow non-linear relations between the response variable and each explanatory variable (Hastie and Tibshirani, 1990). For a GAM, we assume

$$\mathrm{E}(Y_t|\boldsymbol{X}_t) = \exp\left(\beta_0 + \sum_{i=1}^{r} g_i(X_{t,i})\right) \tag{2}$$

where each g_i is a smooth, possibly non-linear, univariate function. Any of the g_i can be made linear to obtain a semi-parametric model. As with a GLM, we use quasi-likelihood estimation.

Cubic smoothing spline's were used to estimate the non-parametric functions g_i. We fix the smoothing parameter to be that value for which $\hat{g}_i$ has four "degrees of freedom" (see Hastie and Tibshirani, 1990).

A step-wise model selection procedure in S-PLUS (1999) was used to determine the optimal GAM. Both linear and non-linear terms were allowed for each covariate, and the step-wise procedure automatically selected whether each covariate should be included, and if so, whether it should be linear or non-linear. The AIC was used in this algorithm for variable selection.

2.3 Parameter Driven Models

In this section we include Zeger's (1988) extension of parameter driven models (PDM) for serially correlated time-ordered count data using results from quasi-likelihood. In a parameter driven model, serial correlation is set up through an unobservable latent process. A Poisson regression model has conditional mean

$$\mathrm{E}(Y_t|\epsilon_t, \boldsymbol{X}_t) = \exp(\boldsymbol{X}_t'\boldsymbol{\beta} + \epsilon_t), \tag{3}$$

where $\boldsymbol{\beta}$ denotes a vector of paramters, and ϵ_t is a latent process allowing both overdispersion and autocorrelation in Y_t. We allow ϵ_t to follow a first-order autoregressive process.

2.4 Transitional Regression Models

Transitional Regression Models were introduced by Brumback et al. (2000). To specify the general model, let $\boldsymbol{D_t}$ denote the present and past covariates and the past response at time t. That is, $\boldsymbol{D_t} = (\boldsymbol{Y}^{t-1}, \boldsymbol{X}^t)$, where $\boldsymbol{Y}^{t-1} = (Y_1, \ldots, Y_{t-1})$ and $\boldsymbol{X}^t = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$. Also, the conditional mean and variance of Y_t given past responses and the covariates are defined as $\mu_t = \mathrm{E}(Y_t|\boldsymbol{D_t})$ and $\nu_t = \mathrm{var}(Y_t|\boldsymbol{D_t})$.

The transitional model has conditional mean given by

$$h(\mu_t) = \boldsymbol{X}_t'\boldsymbol{\beta} + \sum_{i=1}^{s} \theta_i f_i(\boldsymbol{D_t}), \tag{4}$$

where h is a link function, $f_i's$ represent functions of the past outcomes, and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_r\}$ and $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_s\}$ are vectors of parameters.

In this paper we present a special case of a TRM, defined as GLM with time series errors. For a Poisson with AR(1) errors, $\mu_t = \exp(\boldsymbol{X}_t'\boldsymbol{\beta}) + \psi_1 e_{t-1}\sqrt{\upsilon_t}$, where $e_t = (Y_t - \upsilon_t)/\sqrt{\upsilon_t}$ and $\upsilon_t = \exp(\boldsymbol{X}_t'\boldsymbol{\beta})$. Here e_t is scaled to give constant variance. Note that $e_t = \psi_1 e_{t-1} + \delta_t$ where $\{\delta_t\}$ is an independent series with zero mean.

3 Asthma and COPD hospital admissions in Melbourne, Australia from 1 July 1989 to 31 December 1992

Each of the four models was fitted to the asthma and COPD hospital admissions data. To simplify the analysis of seasonality, we excluded the leap days of 29 February 1992 in each series. The following covariates were considered for each model.

- Fourier series functions $\sin(2\pi jt/365)$ and $\cos(2\pi jt/365)$ for $j = 1, 2, \ldots, J$. The value of J was chosen using the AIC. For COPD admissions, $J = 4$ and for asthma admissions, $J = 10$.
- Time trend (a quadratic time trend was considered for GLM, PDM and TRM).
- Day of week factor.
- Seasonally adjusted climatic variables: dry bulb temperature and humidity.
- Seasonally adjusted pollutants: nitrogen dioxide and ozone.
- Non-seasonally adjusted pollutants sulphur dioxide and the air particles index (API).

For sulphur dioxide and API, there was virtually no seasonality observed. Lagged values of each of the climatic and pollutant covariates were considered up to five days previously.

To allow comparison across different statistical models we use the following three measures:

- Mean square error (MSE) = mean $\left\{(Y_t - \hat{Y}_t)^2\right\}$, where $\hat{Y}_t$ are the (inverse link transformed) fitted values.
- Mean square proportional error (MSPE) = mean $\left\{(Y_t - \hat{Y}_t)^2/\hat{Y}_t\right\}$.
- AIC = $n\log(\sigma^2) + 2p$, where σ^2 is the variance of the raw residuals (response minus fitted values), and p is the number of degrees of freedom in each model.

Table 1 displays results from the analyses of COPD and asthma hospital admissions, using different statistical methods. Where a variable has been included in a linear function, the relative risk is shown. For the GAM, variables which were included using a smoothing spline are denoted by $g(\cdot)$.

TABLE 1. *Relative Risk and 95% CI of COPD and asthma admissions for an increase from the 10th to 90th percentile for levels of pollutants, generated using different statistical methods.*

COPD								
	GLM		GAM		PDM		TRM	
Pollutant	RR	95%CI	RR	95% CI	RR	95%CI	RR	95% CI
$NO_{2,t}$	1.06	1.00–1.12	1.06	1.01–1.11	1.05	1.00–1.11	1.05	1.00–1.10
$O_{3,t-2}$			1.06	1.00–1.11				
API_{t-2}			0.95	0.91–1.00				
$SO_{2,t-2}$			g( )					
MSE	13.23		12.76		12.88		12.29	
MSPE	1.24		1.19		1.13		1.16	
AIC	3340.42		3292.19		3252.84		3243.09	

Asthma								
	GLM		GAM		PDM		TRM	
Pollutant	RR	95%CI	RR	95% CI	RR	95%CI	RR	95% CI
$NO_{2,t}$	1.05	1.01–1.08	1.05	1.01–1.09	1.04	1.01–1.08	1.05	1.02–1.08
$NO_{2,t-1}$			0.96	0.92–0.99				
$O_{3,t}$			0.97	0.93–1.00				
$O_{3,t-1}$	0.96	0.93–0.99			0.97	0.94–1.09	0.97	0.95–0.99
$O_{3,t-2}$			g( )					
API_t								
$SO_{2,t}$								
MSE	57.81		53.68		56.05		55.8	
MSPE	1.75		1.61		1.70		1.69	
AIC	5244.03		5153.41		5207.56		5206.92	

Daily COPD hospital admissions increased significantly with increased ambient outdoor levels of same day nitrogen dioxide (NO_2). The observed

effects for ambient outdoor levels of ozone were highly sensitive to model specification and different statistical models. In this study the relationships between particulates (API) and both COPD and asthma hospital admissions were non-robust, because the estimated effects varied depending on the statistical methodology employed in the analysis. The relationship between sulfur dioxide and COPD was also non-robust, as was the relationship between ozone and both COPD and asthma admissions.

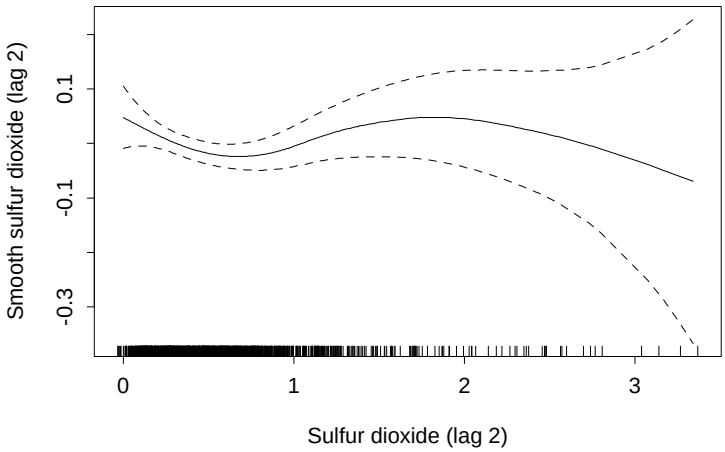

FIGURE 1. *The nonlinear function for sulfur dioxide (lagged 2 days). Dashed lines represent pointwise 95% confidence intervals.*

A GAM analysis showed a nonlinear relationship between sulfur dioxide and COPD hospital admissions in Melbourne, Australia (see Figure 1). This is similar to that found in London (Schwartz and Marcus, 1990) and Europe (Touloumi et al. 1994).

In this study, GLMs were inadequate because of the serial correlation remaining after controlling for trend, seasonality and climate. The inclusion of nonlinearities in a GAM analysis of COPD hospital admissions removed most of the correlation pattern in the residuals from a GLM analysis. However, a GAM was inadequate in representing the short-term association between asthma hospital admissions and air pollution. Significant residual correlation remained even after the inclusion of seasonally adjusted covariates, seasonal terms for the response variable, a time trend and confounding effects of climate.

A parameter driven Poisson regression model was adequate in representing the remaining correlation pattern in the residuals from a GLM analysis of COPD hospital admissions. Although, similar to the GAM anal-

ysis, a small but significant lag three correlation remained. A parameter driven Poisson regression model with an AR(1) process for the covariance structure, for daily asthma hospital admissions was inadequate.

A TGLM (transitional generalized linear model) was fitted to COPD hospital admissions and was adequate in representing the correlation pattern of the residuals with an AR(1) process. A TGLM of asthma hospital admissions and air pollution was inadequate due to the strong correlation pattern in the residuals.

4 Conclusion

This study extends recent epidemiological studies by focusing on the following question: How robust is the observed pollution-respiratory disease relation to different statistical models with various underlying methodological assumptions?

The statistical methodologies adopted in this study are all variations of regression methods. They range from popular nonnormal methods (generalized linear and additive models), to recently developed parameter and observation driven models (Poisson regression with autocorrelation and transitional regression models).

Table 2 displays the strengths and weaknesses of each model adopted in this study. A + indicates a strength and – indicates a weakness of the methodology.

TABLE 2. *Strengths and weaknesses of the statistical methods used in this study.*

	TF	**GLM**	**GAM**	**Par. Driven**	**TRM**
Methodological Issues					
Nonnormality	–	+	+	+	+
Overdispersion	–	+	+	+	+
Nonlinearity	–	–	+	–	–
Autocorrelation	+	–	–	+	+

The utility of GAM methodology for the analysis of respiratory disease and air pollution is demonstrated in this study. Although the GLM with time series errors adequately represented the correlation pattern in COPD admissions, the model was unable to capture the non-linear effect of sulfur dioxide. No model was adequate in representing the correlation pattern in asthma admissions.

The findings from this study show that the relation between ambient outdoor concentrations of nitrogen dioxide and COPD hospital admissions is consistent and robust to different statistical methodology. The findings for levels of ozone, sulfur dioxide and particulates are highly sensitive to model specification.

Acknowledgments: This study was funded by a Public Health Ph.D. Postgraduate Scholarship from the National Health and Medical Research Council.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory.* B.N. Petrov & F. Csaki (eds), Adademiai Kidao, Budapest; pp. 267–281.

Brumback, B.A., Ryan, L.M., Schwartz, J.D., Neas, L.M., Stark, P.C., & Burge, H.A. (2000). Transitional regression models, with application to environmental time series. *Journal of the American Statistical Association*, **95**, 16–27.

Cleveland, W.S., & Terpenning, I.J. (1982). Graphical methods for seasonal adjustment. *Journal of the American Statistical Association*, **77**, 52–62.

Hastie, T., & Tibshirani, R.J. (1990). *Generalized additive models.* London: Chapman and Hall.

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zmirou, D., Zanobetti, A., Wojtyniak, B., Vonk, J.M., Tobias, A., Ponka, A., Medina, S., Bacharova, L. & Anderson, H.R. (1996). Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *Journal of Epidemiology and Community Health, Supplement*, **50**, S12–S18.

McCullagh, P., & Nelder J.A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

Schwartz, J., & Marcus, A. (1990). Mortality and air pollution in London: a time series analysis. *American Journal of Epidemiology*, **131**, 185–194.

S-PLUS (1999). *Modern statistics and advanced graphics.* Seattle, Washington: MathSoft, Inc

Touloumi, G., Pocock, S.J., Katsouyanni, K., & Trichopoulos, D. (1994). Short-term effects of air pollution on daily mortality in Athens: a time series analysis. *International Journal of Epidemiology***23**, 957–967.

Zeger, S. (1988). A regression model for time series of counts. *Biometrika*, **75**, 621–629.