

The University of Chicago  
Center for Integrating Statistical and Environmental Science  
[www.stat.uchicago.edu/~cises](http://www.stat.uchicago.edu/~cises)



Chicago, Illinois USA

TECHNICAL REPORT NO. 16

**Distributed Lag Model: Analysis of  
Air Pollution on Asthma Occurrence**

W. Gu, P. Rathouz

October 2004



\*Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency through STAR Cooperative Agreement #R-82940201-0 to The University of Chicago, 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.

# Distributed Lag Model: Analysis of Air Pollution on Asthma Occurrence

Wen Gu, Department of Statistics, University of Chicago  
Paul Rathouz, Department of Health Studies, University of Chicago

## Abstract

Many studies have reported associations between air pollution and daily asthma occurrence. Those studies have not consistently specified the lag between exposure and response, especially multiple lags, although most have found associations that persisted for more than 1 day. A systematic approach to specifying the lag association would allow better comparison and give insight into the nature of the relation. To examine this question, daily asthma claims (daily beta-agonist prescription counts) of about 20,000 adults were collected in metropolitan Chicago, aggregated at the ZIP code level. The study follows a case-crossover design wherein strata were defined as all days on the same weekday of 5 consecutive weeks in the same year. I fit distributed lag relation to the association between daily asthma claim with ozone,  $PM_{10}$ , relative humidity, pollen and temperature. We assume the weights  $w_l$  for lag  $l$  follow a binomial distribution function and let the total effect of these exposures be distributed over the past 14 days. We use conditional logistic regression for the distributed lag model and apply the Box-Tidwell method to estimate the parameter  $\mathbf{a}$  in the weight function. Models with and without the smooth time function are compared. Results show that exposure to ozone,  $PM_{10}$  and pollen have a positive effect on the response with effects most evident with lags on the order of 7 to 10 days. Exposures to temperature and relative humidity have a negative effect on daily asthma claim. The effect of temperature is strongest with small lags and relative humidity is the only one which doesn't have a lagged structure when jointly estimated with the other four covariates in the distributed lag model; only the exposure on the current day

is associated with the response. The effect of ozone is sensitive to which model is used. A valid test by fixing  $\alpha$  at 0.5 for the covariate we are interested is performed to check the significance of that pollutant.

## I. INTRODUCTION

Asthma is a chronic lung condition. It is characterized by difficulty in breathing (<http://www.lung.ca/asthma>). Its occurrence has become more prevalent during recent years. Many factors have been implicated in asthma exacerbations and a number of studies have implicated adverse effects of outdoor air quality.

In this project, we use a time series of asthma-related outcomes in the Chicago metropolitan area, aggregated at the ZIP code level, to identify linkages with measures of air quality at the spatial scale of individual neighborhoods, while allowing for neighborhood-specific effects.

The data has a longitudinal structure. The advantage of this structure is that it not only enhances the statistical efficiency but also controls confounding by relying on within-subject covariation of health outcomes with air pollution.

In this study, asthma data are collected over four summers 1995-1998, covering from July 1995 through June 1998 in Cook County, Illinois. Residence in the city of Chicago was approximated by including all patients with zip code 606xx. Health outcome data are drawn from billing records from the adult asthma patient Medicaid population in Illinois. Rather than counts of emergency department visits for hospital admissions, which were typically used in past analyses, we use the daily number of short acting beta-agonist prescription refills among Medicaid enrollees as a surrogate marker for exacerbated asthma symptoms. Beta-agonists are typically used to treat asthma and refills can be obtained on a refill-basis without necessarily requiring a physician consult. Because the daily number of beta agonist refills is much greater than either

ED visits or hospitalization, we thus have potentially much more information in this outcome variable. Another advantage is that we use daily Medicaid data clustered at the level of ZIP code of residence, of which we are using 52 in Cook County.

As our analysis is focused on ozone as a primary exposure of interest, we restricted analysis to the seven months of April through October when ozone was at relatively high levels. The summer of 1995 was represented from July 1 on, while data for summer 1998 ended on June 30. Summers of 1996 and 1997 have complete data for April 1 through October 31.

The data used in the analysis are the daily BA prescriptions for weekdays; weekends and the three summer holidays are excluded as number of prescriptions filled on those days were very much lower. The daily data are clustered at the ZIP code level, and are modeled with effects for temperature (F), relative humidity (%) and levels of ozone (ppb), particulate matter (PM<sub>10</sub>) ( $mg/m^3$ ) and pollen (in the log base). The units of temperature, relative humidity, PM<sub>10</sub> and ozone are 5F, 10%, 15  $mg/m^3$  and 20 ppb separately. Pollen was measured in units of grains per  $m^3$ . For the analysis here, we use log-pollen, which was standardized with mean zero and standard error 1. We next check the distributions of the exposures during the seven months from April to October. Ozone has a general trend of increasing then decreasing over time, with maximum level reached approximately during July or August. The range of its fluctuation is close to 6 units, which equals to about 120 ppb changes overall. No obvious trends were observed for relative humidity and PM<sub>10</sub> within the time period. The overall range of fluctuation is close to 70% for relative humidity and 120  $mg/m^3$  for PM<sub>10</sub>. Temperature has a very apparent trend of increasing and decreasing during the seven months, highest temperature was achieved in August or September. The overall change is close to 55F for the whole time period. There are

large fluctuations in log-pollen with a slight decreasing trend over time. There is a total of 6 units change in the standardized log-pollen.

The design of the study is case-crossover, in the sense that only cases are collected, each subject also serves as its own control. We choose the case interval to be the exposure time just before the failure and the control interval to be one or more prior or subsequent time periods. This concept will be explored in more detail in the following parts.

For inferences in the case-crossover design, we use the following model. Estimation will be performed via conditional logistic regression. The outcome of the data is  $Y_{ijt}$ , a binary variable indicating the occurrence of the asthma event for person  $i$  at ZIP code  $j$  and day  $t$ :

$$Y_{ijt} = \begin{cases} 1 & \text{asthma event occur} \\ 0 & \text{otherwise} \end{cases} .$$

We now set up the hazard model:

$$p_{ijt} = \Pr(Y_{ijt} = 1) = \exp(R(x_{ij}) + R_{ij} + R_{ijk} + \mathbf{b}' z_{ijt})$$

where  $z_{ijt}$  is air pollution and adjustment covariates for person  $(i, j)$  on day  $t$ .

$R_{ij}$  and  $R_{ijk}$  are person- and person-time-specific effects. We assume that  $R_{ijk} = R_{ijt} = \text{constant}$  for each  $t \in W_k$ . And  $R(x)$  is the “adjusted” log-asthma event risk at location  $x$ .

We partition  $\{1, \dots, t_{\max}\}$  into “time-windows”  $W_k$ .

$W_1$  = sequence of 5 Mondays;  $W_2$  = sequence of 5 Tuesdays; .....;  $W_5$  = sequence of 5 Fridays;  
 $W_6$  = sequence of next 5 Mondays; .....

Defining strata as all days on the same day of the week within the same month and year is a common choice of stratification in the air pollution epidemiology studies. For example, if the index day falls on the second Tuesday in August of 1996, the referents could be all other Tuesdays in August of 1996. Simulation studies have shown that shorter time periods for the

referents (e.g. 7 or 7&14 days before and after the index day) result in less confounding bias (Janes, 2003). Hence we choose the strata to be a sequence of 5 Mondays, Tuesdays, etc. People's refill behavior is more consistent on the sequence of days as defined above than, for instance, consecutive 5 days in a week. We can imagine that a person who has refilled the prescription on Wednesday is likely to refill the prescription again on a Wednesday. Confounding due to season and day of the week are controlled.

We assume  $R_{ijt} = R_{ijk} = \text{constant}$  for all  $t \in W_k$ , then  $\{R(x_{ij}) + R_{ij} + R_{ijk}\}$  is a stratum-level nuisance intercept and captures unobserved location-, person- and person-time-specific factors that could confound the association of  $Y_{ijt}$  to  $z_{ijt}$ . One advantage of conditional logistic regression is that the nuisance parameters won't be estimated, it only depends on the parameters of interest.

We perform conditional logistic regression for strata of response  $\{Y_{ijt}\}$  for  $t \in W_k$ .

In this study,  $z_{ijt}$  corresponds to  $\{\text{temp, pollen, relhum, pm10, ozone}\}$  on day  $t$ , zip code  $j$  currently and in past days up to lag  $L$ . Thus the model in its full form is:

$$\Pr(Y_{ijt} = 1) = \exp(R(x_{ij}) + R_{ij} + R_{ijk} + \mathbf{b}_1 \text{temp}_{ijt} + \mathbf{b}_2 \text{pollen}_{ijt} + \mathbf{b}_3 \text{relhum}_{ijt} + \mathbf{b}_4 \text{pm10}_{ijt} + \mathbf{b}_5 \text{ozone}_{ijt})$$

## II. CASE-CROSSOVER DESIGN

A case-control study is a design which is used to assess the relationship between the exposure to a risk factor and the development of a disease (Breslow, 1980). It compares the exposure distributions between groups of patients with (the cases) and without (the controls) the disease. It typically uses only a fraction of subjects in the non-disease group. A case-control study is a rapid and efficient way to evaluate a hypothesis about an exposure-disease relationship compared to other analytical designs.

Despite its practicality, the case-control study cannot be done well without considerable planning.

Indeed, a case-control study is perhaps the most challenging to design and conduct in such a way that bias is avoided (Maclure, 1991). This is mainly because of the non-straightforward selection of the control group. Healthy representatives of the general population are not always easy to obtain and reducing the selection bias in the control group is a big problem persisting in the case-control study. It's natural to ask who the best representatives of the population base that produced the cases would be. A simple answer is the cases themselves. This led to the development of the case-crossover design (Maclure, 1991).

The case-crossover design is an extension of the crossover design to observational studies. In the experimental setting, the term "crossover" is mainly used to describe experiments in which all subjects pass through both the treatment and placebo phases. The word "case" indicates that only cases are to be sampled and that each subject serves as his or her own control. The appeal of the basic design is that, within-subject comparisons make it possible to avoid confounding by subject-specific attributes that are constant over time. Case-crossover designs are useful for estimating effects that are acute and transient. If the effects can persist well beyond the cessation of a treatment, which are known as carry-over effects, they can result in severe bias under this study design.

The advantage of the case-crossover study is obvious: First, there is no need to select the control group, which can greatly reduce the expenses and efforts and eliminates sources of bias related to sampling of controls. Second, by making comparisons within subject, individual susceptibility factors can be controlled for under appropriate conditions.

Data analysis in case-crossover studies is done by standard case-control methods (Navidi, 1998) for matched studies. The basic principle is to estimate risk by comparing the exposure of the subject during a time interval just before failure (the case interval) with the exposure during one

or more other time periods (control intervals) in which no failure is observed. Conditional logistic regression is the general method for a case-crossover analysis.

In its original formulation, the case-crossover study was unidirectional retrospective, which means the control time was only selected to be prior to the event. The reason for this was that the original applications of the design involved outcomes that were likely to affect subsequent exposures. Thus, if control time is selected post failure, it could introduce reverse-causation bias. For example, failure may remove person for being at risk to fail, may change exposure status or even change what it means to “fail”. However, unidirectional retrospective control sampling can also be severely biased when time trends in exposure are strong, such as with air pollutants. For example, if the level of exposure is overall increasing or decreasing over time while allowing individual deviations from the general trend, and we only choose control intervals to be prior to the case interval and compare the exposures during the two intervals, bias due to the overall increasing or decreasing trend with the exposures will be resulted. Yet studying the effects of environmental rather than behavioral exposures has the advantage that subsequent levels of exposure are most likely unaffected by failures of subjects. Therefore, control time can be assessed either before or after the event. This method is referred to as bi-directional case-crossover method. This allows consistent estimators of risk to be computed regardless of time trends in exposure.

In a typical air pollution time series study, daily event counts, in this analysis the beta-agonist prescription counts, are regressed on the shared exposure series typically in log-linear regression models. Strong confounding effects of season and weather must be controlled for statistically in the model. The case-crossover design shares the same goal as the above regression analysis except that confounding is controlled for by design (by matching on person and time-window)

instead of in the regression model. Referents that are restricted to the same day of week and season as the index time control for these confounding effects by design. In addition, person-level confounding factors, e.g., ozone exposure variation across persons is also controlled because strata are defined within person.

Time-stratified referent sampling design divides time into disjoint strata. This strategy maintains control over confounders by design by ensuring the strata are matched on important confounders. In the air pollution context, if the time duration covered by a stratum is short and members are all on the same day of the week, confounding due to season and day of the week is controlled for. There is no bias due to time trend since there is no pattern in the placement of referents relative to the index time.

### III. CONDITIONAL LOGISTIC REGRESSION

Standard analysis of case-crossover design is derived by considering each subject crossed with each time-window to be a stratum, where the cases and controls are times. Failure times are the cases and the other times are the controls. We make inferences via conditional logistic regression (Breslow, 1980).

Conditional logistic regression is used to investigate the relationship between an outcome and a set of prognostic factors in matched case-control studies. The outcome is whether the subject at a given time is a case or a control. If there is only one case and one control, the matching is 1:1. The  $m:n$  matching refers to the situation in which there is a varying number of cases and controls in the matched sets.

Suppose for a given stratum composed of  $n_1$  cases and  $n_0$  controls,  $t_1, \dots, t_M$  are the times at which failures can occur and that we know the unordered values  $X_{ij}$  --the vector of exposure covariates for subject  $i$  at time  $t_j$ , but we don't know which values are associated with the case

times and which with the control times. Let  $\mathbf{b}$  represent the vector whose components represent the log odds of failure associated with one unit increase in the corresponding components of  $X_{ij}$ .

Thus we can write the log odds of failure for subject  $i$  on day  $t_j$  as

$$\log \frac{p_{ij}}{1-p_{ij}} = \mathbf{I}_{ik} + \mathbf{b}^T X_{ij} \quad , \quad (1)$$

where  $\mathbf{I}_{ik}$  is the baseline level specific to subject  $i$ , stratum  $k$ . Then the unconditional probability of failure for subject  $i$  on day  $j$ ,  $p_{ij}$  is

$$p_{ij} = \frac{\exp(\mathbf{I}_{ik} + \mathbf{b}^T X_{ij})}{1 + \exp(\mathbf{I}_{ik} + \mathbf{b}^T X_{ij})} \quad . \quad (2)$$

Let random variable  $n_i$  be the total number of failures for subject  $i$  within a stratum. We also define  $D_{n_i}$  to be the collection of all sets of  $n_i$  times, since the failures can occur at any time points within the stratum, there will be  $M!/[n_i!(M-n_i)!]$  possibilities, where  $M$  is the total number of time units at risk. Let  $A_i$  be a random subset of  $D_{n_i}$ . Then  $P(A_i)$ , the probability that subject  $i$  would experience failures precisely at the times in the set  $A_i$  and not experience failure in the complement of  $A_i$  can be written as

$$P(A_i) = \left( \prod_{t_j \in A_i} p_{ij} \right) \left( \prod_{t_j \notin A_i} (1-p_{ik}) \right) \quad (3)$$

If we condition on the total number of failures  $n_i$ , then the conditional probability that the failures occurred precisely at those times in set  $A_i$  is

$$P(A_i | n_i) = \frac{\left( \prod_{t_j \in A_i} p_{ij} \right) \left( \prod_{t_j \notin A_i} (1-p_{ik}) \right)}{\sum_{S \in D_{n_i}} \left( \prod_{t_j \in S} p_{ij} \right) \left( \prod_{t_k \notin S} (1-p_{ik}) \right)} \quad (4)$$

Combining the expression of  $p_{ij}$  with  $P(A_i | n_i)$  yields

$$P(A_i | n_i) = \frac{\exp(\mathbf{b}^T \sum_{t_j \in A_i} X_{ij})}{\sum_{S \in D_{n_i}} \exp(\mathbf{b}^T \sum_{t_k \in S} X_{ik})} \quad (5)$$

Let  $T_{iS} = \sum_{t_k \in S} X_{ik}$ , we can simplify the above conditional probability to be

$$P(A_i | n_i) = \frac{\exp(\mathbf{b}^T T_{iA_i})}{\sum_{S \in D_{n_i}} \exp(\mathbf{b}^T T_{iS})} \quad (6)$$

Summing the logarithm of the (6), we get the likelihood function as

$$L(\mathbf{b}) = \sum_i [\mathbf{b}^T T_{iA_i} - \log \sum_{S \in D_{n_i}} \exp(\mathbf{b}^T T_{iS})] \quad (7)$$

For each subject  $i$  and for each set  $S \in D_{n_i}$ , the quantity  $P(S | n_i)$ , computed by substituting  $S$  for  $A_i$  in (4), is the probability that  $S$  is the set of failure times for subject  $i$ , given the total number of failure  $n_i$ . If  $T_i$  is the sum of the covariates over the failure times for subject  $i$ , then

$$E(T_i) = \sum_{S \in D_{n_i}} T_{iS} P(S | n_i) \quad (8)$$

and

$$\text{cov}(T_i) = \sum_{S \in D_{n_i}} T_{iS} T_{iS}^T P(S | n_i) - E(T_i) E(T_i)^T \quad (9)$$

The score equation follows naturally

$$\frac{\partial L(\mathbf{b})}{\partial \mathbf{b}} = \sum_{i=1}^n [T_{iA_i} - E(T_i)] = 0 \quad (10)$$

and the asymptotic covariance of  $\hat{\mathbf{b}}$ , which maximizes the likelihood in (7) and solves (10) is

$$\text{cov}(\hat{\mathbf{b}}) = \left[ \sum_{i=1}^n \text{cov}(T_i) \right]^{-1} \quad (11)$$

The advantage of using conditional likelihood as a basis for statistical analysis of the results of a case-crossover study is that it depends only on the relative risk parameters of interest— $\mathbf{b}^T$  (see

equation (1)). Nuisance or baseline parameter  $I_{ik}$  will be eradicated, thus allowing for construction of tests and estimates without concern for the many  $I_{ik}$ 's.

#### IV. DISTRIBUTED LAG MODEL

In this study, we utilize a distributed lag model to model the effect of daily fluctuations of air pollution and weather on the occurrence of acute asthma event, in this case filling or refilling a beta-agonist prescription. Distributed lag models are not new, they have been used in social science and economics for decades. Pope and Schwartz recently described the use of this approach in epidemiology (Schwartz, 2000; Zeger, 2004).

It has been realized that different kinds of weather and pollution factors can affect not only asthma occurring on the same day, but on several subsequent days. Thus today's asthma occurrence will depend on the "same day" effect of weather and pollution factors, the "1-day lag" effect of yesterday's weather and pollution factors, the "2-day lag" effect of the day before yesterday's weather and pollution factors, etc. So the distributed lag models are actually time series models that allow an exposure to affect the response over an extended time period, and hence provide a natural framework for considering the effects of weather on asthma.

We use  $x_{i(j,t)}$  to denote a weather covariate  $x$  realized  $l$  days before the asthma event for person  $i$  at zip code  $j$ . The overall effect of a unit increase in the covariate on a single day is its impact on that day plus its impact on subsequent days. That is, it is the sum of  $\bar{b}_0 + \dots + \bar{b}_l$ . To see why this is true, note we can rewrite the equation (1) to be:

$$\text{logit}(p_{ijt}) = I_{ijk} + \sum_{l=0}^L \bar{b}_l x_{ij(t-l)} \quad (12)$$

as

$$\text{logit}(\mathbf{p}_{ijt}) = \mathbf{I}_{ijk} + \mathbf{b} \times \sum_{l=0}^L \mathbf{w}_l x_{i,j,t-l} \quad (13)$$

where  $\mathbf{w}_l$  are weights that sum to one, and  $\bar{\mathbf{b}}_l = \mathbf{b} \times \mathbf{w}_l$ . That is,  $\mathbf{b}$  is also interpretable as the effect on day  $t$  of a unit increase in all  $0, \dots, L$  lagged values of  $x_{ijt}$ .

Since weather and air pollution factors on days close together will have high correlation, the estimation of the individual  $\mathbf{b}$ 's will be very unstable because of the high collinearity. Nevertheless, the sum of the individual  $\mathbf{b}$ 's will be a consistent estimate of the overall effect of a unit increase in weather or air pollution variable. However, if we don't constrain the shape of the distribution of effect of weather or pollution over time, the estimates would be inefficient. Thus to gain more efficiency and more insight into the shape of the distributed effect of weather over time, it is useful to constrain the  $\mathbf{b}$ 's. If this is done flexibly, substantial gains in reducing the noise of the unconstrained distributed lag model can be obtained, with minimal bias. Also, this increase in efficiency will allow us to estimate multiple distributed lag effects, which has not been done very much before, especially for the analysis of air pollution.

In this problem, we assume the weight  $\mathbf{w}_l$  for lag  $l$  follow a binomial probability mass function, if the lag number  $l$  goes from 0 (the current day) to  $L$  days before, then  $l$  follow *Binomial*( $L, \mathbf{a}$ ), where the total number of "observations" is  $L$  and the probability is  $\mathbf{a}$ . More precisely, we can rewrite the weights  $\mathbf{w}_l$  as  $\mathbf{w}_l(\mathbf{a})$ , where

$$\mathbf{w}_l(\mathbf{a}) = \binom{L}{l} \mathbf{a}^l (1-\mathbf{a})^{L-l} \quad (14)$$

The reason why we choose binomial distribution to model the lags is because the mode of this distribution is flexible, which can shift from left to right with increase in the probably  $\mathbf{a}$ . Small

$\mathbf{a}$  indicates that the short lags have more important effect on the response, while large value of  $\mathbf{a}$  indicates the longer lags are more significant on predicting the response. See figure 1.

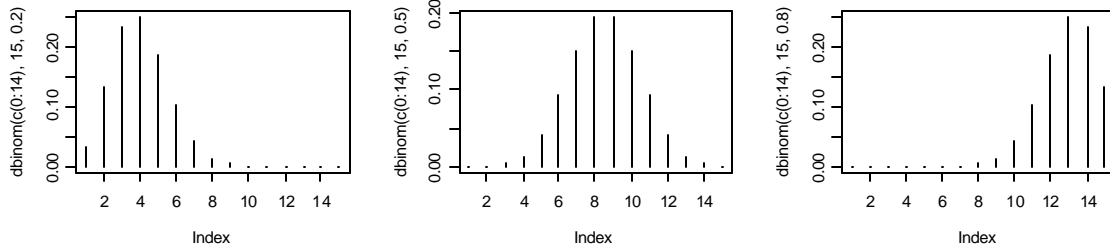


Figure 1. The shape of binomial distribution with  $L=15$ ,  $a=0.2, 0.5, 0.8$  from left to right. The mode shifted from left to right.

Thus we can rewrite the distributed lag model (13) as

$$\text{logit}(\mathbf{p}_{ijt}) = \mathbf{I}_{ijk} + \mathbf{b} \sum_{l=0}^L \Pr(T=l) x_{ij(t-l)} = \mathbf{I}_{ijk} + \mathbf{b} \sum_{l=0}^L \binom{L}{l} \mathbf{a}^l (1-\mathbf{a})^{L-l} x_{ij(t-l)} \quad (15)$$

I have chosen a maximum lag of 14 days before the asthma event for the weather variables, which seems to be appropriate and long enough to detect the variation of the effect of a specific weather and pollution variable. It's reasonable to assume that exposure levels on more than 14 days age have minimal effect on the response.

In this analysis, there are 5 air quality variables: temperature, pollen, relative humidity,  $\text{PM}_{10}$  and ozone. We decided to apply the distributed lags to all the 5 covariates, the lags goes from 0 to 14 days ago, they all follow binomial distribution with  $L=14$  but with different  $\mathbf{a}$ 's. To estimate the  $\mathbf{a}$ 's is the goal of this analysis since it will tell us within the 14 days preceding the event, the weather and pollution variables on which day will have the most important effect on the occurrence of the asthma event. The model is constructed as following:

$$\begin{aligned} \text{logit}(\mathbf{p}_{ijt}) = & R_{ijk} + \mathbf{b}_0 + \mathbf{b}_1 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_1) \times \text{temp}_{ij(t-l)} + \mathbf{b}_2 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_2) \times \text{pollen}_{ij(t-l)} \\ & + \mathbf{b}_3 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_3) \times \text{relhum}_{ij(t-l)} + \mathbf{b}_4 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_4) \times \text{pm10}_{ij(t-l)} + \mathbf{b}_5 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_5) \times \text{ozone}_{ij(t-l)} \end{aligned} \quad (16)$$

$\mathbf{b}_1, \dots, \mathbf{b}_5$  are the effects of a unit increase in the weighted average of the corresponding weather variables.  $\mathbf{a}_1, \dots, \mathbf{a}_5$  are the parameters of the weights function to be estimated.

## V. BOX-TIDWELL METHOD

From (16), we see that the model is not a generalized linear model; although  $\mathbf{b}_1, \dots, \mathbf{b}_5$  are the linear parameters,  $\mathbf{a}_1, \dots, \mathbf{a}_5$  are the non-linear parameters. To estimate  $\hat{\mathbf{a}}$ , we will need a fitting technique—Box-Tidwell appears to be appropriate in this case.

To deal with generalized non-linear models with non-linear parameter in the covariates, Box and Tidwell (1962) describe a fitting technique by linearization, which follow closely that for non-linear parameters in the link function. Model (16) contains terms  $\mathbf{b}g(x;\mathbf{q})$ , where  $x$  can be any observed covariate vector. We proceed as follows: Let  $g(x;\mathbf{q})$  be the covariate to be used, with  $\mathbf{q}$  unknown, we expand about an initial value  $\mathbf{q}^{(0)}$  to give the linear approximation (ignoring second order and higher terms)

$$g(x;\mathbf{q}) \approx g(x;\mathbf{q}^{(0)}) + (\mathbf{q} - \mathbf{q}^{(0)})[\partial g / \partial \mathbf{q}]_{\mathbf{q}=\mathbf{q}^{(0)}} \quad (17)$$

Thus if a non-linear term in the linear predictor is given by

$$\mathbf{b}g(x;\mathbf{q}),$$

we replace it by two linear terms

$$\mathbf{b}u + \mathbf{g}v$$

where

$$u = g(x;\mathbf{q}^{(0)}), \quad v = [\partial g / \partial \mathbf{q}]_{\mathbf{q}=\mathbf{q}^{(0)}} \quad \text{and} \quad \mathbf{g} = \mathbf{b}(\mathbf{q} - \mathbf{q}^{(0)})$$

An extra level of iteration is again required, and after fitting a model including  $u$  and  $v$  as covariates we obtain

$$\mathbf{q}^{(1)} = \mathbf{q}^{(0)} + \hat{\mathbf{g}}/\hat{\mathbf{b}} \quad (18)$$

(Because  $\mathbf{g} = \mathbf{b}(\mathbf{q}^{(1)} - \mathbf{q}^{(0)}) \Rightarrow \mathbf{q}^{(1)} = \mathbf{q}^{(0)} + \hat{\mathbf{g}}/\hat{\mathbf{b}}$ )

as the improved estimate, and iterate. Convergence is not guaranteed for starting values arbitrarily far from the solution. If the process does converge then the presence of the extra term  $\mathbf{g}v$  ensures that the asymptotic covariances produced for the remaining parameters are correctly adjusted for the fitting of  $\mathbf{q}$ . If we wish to obtain the asymptotic variance of  $\hat{\mathbf{q}}$  directly, we need a final iteration with  $\hat{\mathbf{b}}v$  in the place of  $v$ ; the components of  $(X^T W X)^{-1}$  corresponding to that covariate then give the approximate variance of  $\hat{\mathbf{q}}$  and its covariances with the other parameters.

To see why the variance of  $\hat{\mathbf{q}}$  and its covariances with the other parameters are correct after the final iteration, let's use  $l$  to indicate the log likelihood function,  $\mathbf{h}$  the predictor, either linear or non-linear. If  $\mathbf{h}$  is linear with

$$\mathbf{h} = \mathbf{a}_0 + \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 \quad (19)$$

then we can rewrite the score function and the second derivative with respect to  $\mathbf{a}_1$  and  $\mathbf{a}_2$  as

$$\frac{\partial l}{\partial \mathbf{a}_1} = \frac{\partial l}{\partial \mathbf{h}} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{a}_1}; \quad \frac{\partial^2 l}{\partial \mathbf{a}_1 \partial \mathbf{a}_2} = \frac{\partial l}{\partial \mathbf{h}} \cdot \frac{\partial^2 \mathbf{h}}{\partial \mathbf{a}_1 \partial \mathbf{a}_2} + \frac{\partial^2 l}{\partial \mathbf{h}^2} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{a}_1} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{a}_2} \quad (20)$$

$\partial l / \partial \mathbf{h}$  is the score function, and has expectation zero, and  $\partial^2 \mathbf{h} / \partial \mathbf{a}_1 \partial \mathbf{a}_2$  doesn't contain any data, we can approximate

$$\frac{\partial^2 l}{\partial \mathbf{a}_1 \partial \mathbf{a}_2} \approx \frac{\partial^2 l}{\partial \mathbf{h}^2} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{a}_1} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{a}_2} \quad (21)$$

This rule applies regardless whether the predictor  $\mathbf{h}$  is linear or non-linear.

For linear  $\mathbf{h}$ , we can derive the second derivative of the log likelihood with respect to  $\mathbf{a}_1$ ,  $\mathbf{a}_2$ , and  $\mathbf{a}_1$  and  $\mathbf{a}_2$  as the following

$$\begin{aligned}\frac{\partial^2 l}{\partial \mathbf{a}_1^2} &= \frac{\partial^2 l}{\partial \mathbf{h}^2} \left( \frac{\partial \mathbf{h}}{\partial \mathbf{a}_1} \right)^2 = \frac{\partial^2 l}{\partial \mathbf{h}^2} x_1^2; & \frac{\partial^2 l}{\partial \mathbf{a}_2^2} &= \frac{\partial^2 l}{\partial \mathbf{h}^2} \left( \frac{\partial \mathbf{h}}{\partial \mathbf{a}_2} \right)^2 = \frac{\partial^2 l}{\partial \mathbf{h}^2} x_2^2; \\ \frac{\partial^2 l}{\partial \mathbf{a}_1 \partial \mathbf{a}_2} &= \frac{\partial^2 l}{\partial \mathbf{h}^2} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{a}_1} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{a}_2} = \frac{\partial^2 l}{\partial \mathbf{h}^2} x_1 x_2\end{aligned}\quad (22)$$

the variance and covariance of the parameters can be obtained based on above results.

Now suppose we have non-linear predictor

$$\mathbf{h} = \mathbf{b}_0 + \mathbf{b}_1 g(x, \mathbf{q}) \quad (23)$$

Using the rules above, we can rewrite the second derivatives by

$$\begin{aligned}\frac{\partial^2 l}{\partial \mathbf{b}_1^2} &= \frac{\partial^2 l}{\partial \mathbf{h}^2} \left( \frac{\partial \mathbf{h}}{\partial \mathbf{b}_1} \right)^2 = \frac{\partial^2 l}{\partial \mathbf{h}^2} (g(x, \mathbf{q}))^2; & \frac{\partial^2 l}{\partial \mathbf{q}^2} &= \frac{\partial^2 l}{\partial \mathbf{h}^2} \left( \frac{\partial \mathbf{h}}{\partial \mathbf{q}} \right)^2 = \frac{\partial^2 l}{\partial \mathbf{h}^2} \left( \mathbf{b}_1 \frac{\partial g}{\partial \mathbf{q}} \right)^2 \\ \frac{\partial^2 l}{\partial \mathbf{b}_1 \partial \mathbf{q}} &= \frac{\partial^2 l}{\partial \mathbf{h}^2} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{b}_1} \cdot \frac{\partial \mathbf{h}}{\partial \mathbf{q}} = \frac{\partial^2 l}{\partial \mathbf{h}^2} \cdot (g(x, \mathbf{q})) \cdot \left( \mathbf{b}_1 \frac{\partial g}{\partial \mathbf{q}} \right)\end{aligned}\quad (24)$$

Remember during the last iteration of Box-Tidwell, convergence is reached and we are linearly regressing toward

$$u = g(x; \mathbf{q}) \quad v = \mathbf{b}[\partial g / \partial \mathbf{q}]$$

If we regard  $u$  as  $x_1$  and  $v$  as  $x_2$ , we can demonstrate the variance and covariance matrix obtained from the linear regression has exactly the same form as the one we derived starting from the non-linear predictor. Thus we have proved the correctness of the variance and covariance matrix by using Box-Tidwell fitting technique.

## VI. BOX-TIDWELL ITERATION AND GAUSS-SEIDEL ITERATION

Although the Box-Tidwell technique is undoubtedly useful, and indeed probably under-used, it is usually unwise to try to include more than a very few non-linear parameters in this way,

especially when the other covariates are themselves appreciably correlated in the data set. It will usually be found that estimates of the non-linear parameters have large sampling errors, and are highly correlated with the linear parameters and perhaps with each others.

In this problem, if we include the distributed lags for all of the five covariates, then there will be five non-linear  $\mathbf{a}$ 's to be estimated, recall model (16),

$$\begin{aligned} \text{logit}(\mathbf{p}_{ijt}) = & R_{ijk} + \mathbf{b}_0 + \mathbf{b}_1 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_1) \times \text{temp}_{ij(t-l)} + \mathbf{b}_2 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_2) \times \text{pollen}_{ij(t-l)} \\ & + \mathbf{b}_3 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_3) \times \text{relhum}_{ij(t-l)} + \mathbf{b}_4 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_4) \times \text{pm10}_{ij(t-l)} + \mathbf{b}_5 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_5) \times \text{ozone}_{ij(t-l)} \end{aligned}$$

Obviously it is unwise to use Box-Tidwell to estimate the five  $\mathbf{a}$ 's at the same time. Even if the starting values are appropriate, convergence is not guaranteed.

Gauss-Seidel iteration (Jeffrey and Jeffreys, 1988) is a good candidate to solve this problem. It is a technique for solving the  $n$  equations of the linear system of equations  $AX = b$  one at a time in sequence, and uses previously computed results as soon as they are available. This method is especially useful for estimating multiple  $\mathbf{a}$ 's. In this problem, for instance, instead of estimating the five  $\mathbf{a}$ 's at the same time, we estimate them one by one. Start from an arbitrary  $g_r(x_r; \mathbf{a}_r)$ ,  $r=1,2,3,4,5$ , fix the value of  $\mathbf{a}$  for the other four covariates and apply Box-Tidwell method only for the  $r^{\text{th}}$  term. Then, for each iteration, we estimate all  $\mathbf{b}$ 's and  $\mathbf{a}_r$ . When  $\mathbf{a}_r$  converge, go to the next  $r$  and again estimate all  $\mathbf{b}$ 's and  $\mathbf{a}_{r+1}, \dots$ , Iterate to convergence of all  $\mathbf{a}_r$ 's. For each iteration, always use the most recently updated values for the four  $\mathbf{a}$ 's not estimated at that step. This method is a special case of Gauss-Seidel iteration, also sometimes called the method of successive displacements.

A remarkable property about the Gauss-Seidel algorithm is that convergence is guaranteed whenever the matrix of known coefficients (Hessian, the second derivative matrix) is symmetric

positive definite (as it will generally be in regression computations, for instance) (Golub and Van Loan, 1983).

## VII. RESULTS

### 1. 5-day distributed lag model

We start by fitting the model which only includes the distributed lag for one covariate and 1-day lag effect for the other four covariates. Starting from ozone, fit model

$$\begin{aligned} \text{logit}(\mathbf{p}_{ijt}) = & R_{ijk} + \mathbf{b}_0 + \mathbf{b}_1 \text{temp}_{ij(t-1)} + \mathbf{b}_2 \text{pollen}_{ij(t-1)} + \mathbf{b}_3 \text{1relhum}_{ij(t-1)} \\ & + \mathbf{b}_4 \text{pm10}_{ij(t-1)} + \mathbf{b}_5 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}) \times \text{ozone}_{ij(t-l)} \end{aligned} \quad (17)$$

Before applying Box-Tidwell method for  $\mathbf{a}$ , we can maximize the likelihood for various values of  $\mathbf{a}$  and plot the log likelihood against  $\mathbf{a}$ , thereby obtaining a profile likelihood curve. The maximum, usually unique, gives  $\hat{\mathbf{a}}$ . This is a quick way to obtain  $\hat{\mathbf{a}}$  and provide a reasonable range for the starting value when using Box-Tidwell method.

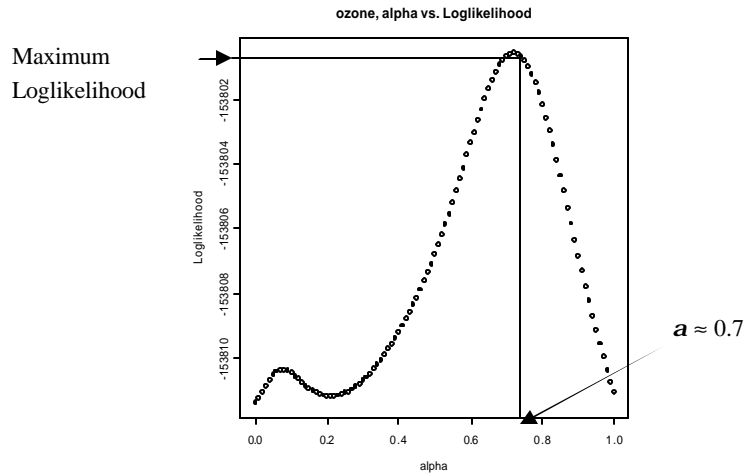


Figure 2. Likelihood profile for ozone

From figure 2, we see maximum likelihood is achieved when  $\mathbf{a}$  is around 0.7. Next, we apply

Box-Tidwell method to term  $\mathbf{b}_5 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}) \times \text{ozone}_{ij(t-l)}$ , estimate  $\hat{\mathbf{a}}$  and its standard error.

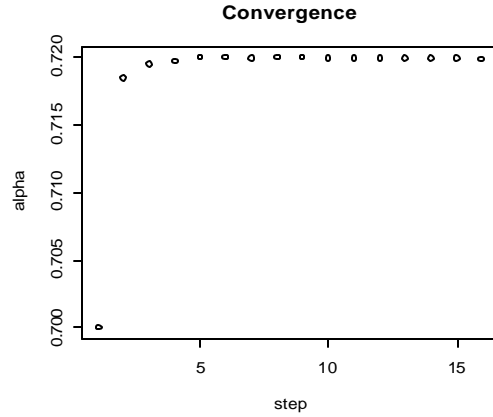


Figure 3. Convergence of  $\hat{\alpha}$  .

Let

$$g(\text{ozone}; \mathbf{a}) = \sum_{l=0}^{14} W_l(\mathbf{a}) \times \text{ozone}_{i(j+l)}, \quad \frac{\partial g}{\partial \mathbf{a}} \Big|_{\mathbf{a}=\mathbf{a}_0} = \sum_{l=0}^{14} \frac{l-14\mathbf{a}_0}{\mathbf{a}_0(1-\mathbf{a}_0)} W_l(\mathbf{a}_0) \times \text{ozone}_{ij(t-l)} .$$

Apply conditional logistic regression to the four 1-day lag effect of temperature, pollen, relative humidity, PM<sub>10</sub> as well as  $g$  and  $\partial g / \partial \mathbf{a}$  . Let 0.7 be the starting value for  $\mathbf{a}_{\text{ozone}}$  . Convergence is quickly reached, see figure 3.

$\hat{\alpha}$  is 0.72 with standard error 0.0017,  $\hat{\mathbf{b}}_{\text{ozone}}$  is 0.04—ozone with 10-day lag effect within the past 14 days exhibits the most strong positive effect on the asthma occurrence. There's no reason to think the other pollutants don't have such a pattern. Thus, we can do the same thing to estimate  $\hat{\alpha}$  for the other four covariates separately. As usual, we check the likelihood profile first to have a basic idea of where  $\hat{\alpha}$  should converge to (figure 4).

Except for temperature, whose  $\hat{\alpha}$  converges to about 0.1 (1- or 2-day lag effect is most strong), the maximum likelihood of pollen, relative humidity and PM<sub>10</sub> all correspond to  $\hat{\alpha}$  around 0.6 or 0.7 (8- or 9-day lag effect is most strong). I summarized the standard error and  $\hat{\alpha}$  for the five variables in table 1.

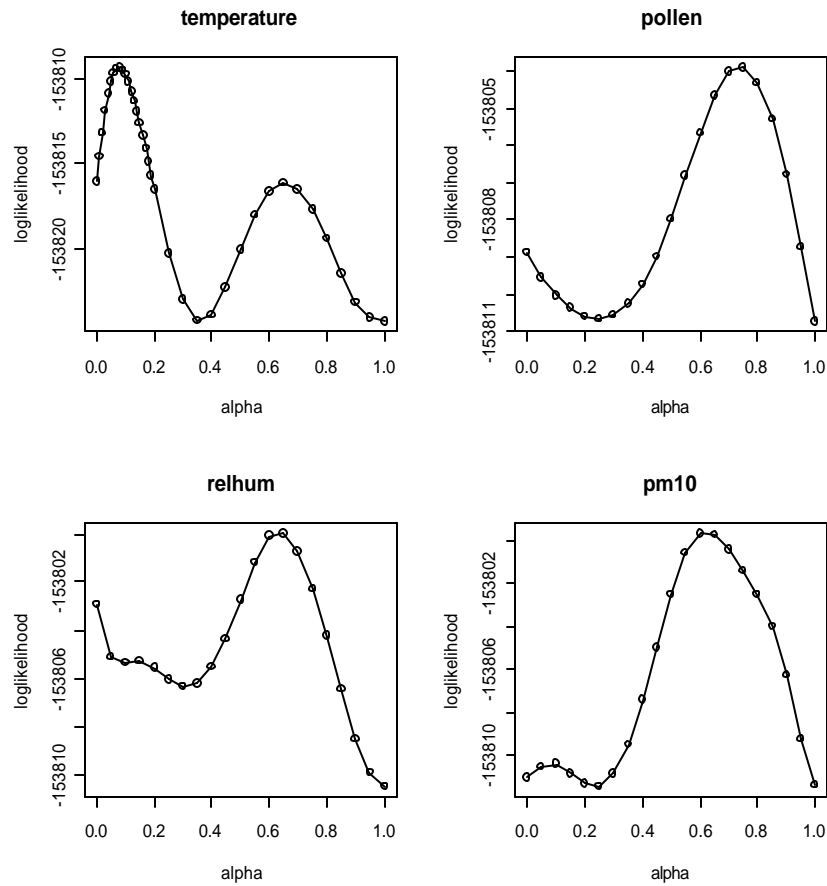


Figure 4. Likelihood profile for temperature, pollen, relative humidity and PM<sub>10</sub>.

	$\hat{\mathbf{a}}$	$sd(\hat{\mathbf{a}})$	log likelihood
temperature	0.077	0.0280	-153809.33
pollen	0.736	0.0992	-153803.87
relative humidity	0.634	0.0409	-153799.88
PM <sub>10</sub>	0.621	0.0374	-153799.59
ozone	0.720	0.0409	-153800.57

Table 1. Summarization of  $\hat{\mathbf{a}}$  and its standard error when estimated separately.

Remember these  $\hat{\mathbf{a}}$  's are estimated separately, we only include the distributed lag for the covariate we want to estimate and fix the 1-day lag effect for the other four. Our ultimate goal is to estimate  $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5$  for the model including the distributed lags for all five variables, model (16):

$$\begin{aligned} \text{logit}(\mathbf{p}_{ijt}) = & R_{ijk} + \mathbf{b}_0 + \mathbf{b}_1 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_1) \times \text{temp}_{ij(t-l)} + \mathbf{b}_2 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_2) \times \text{pollen}_{ij(t-l)} \\ & + \mathbf{b}_3 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_3) \times \text{relhum}_{ij(t-l)} + \mathbf{b}_4 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_4) \times \text{pm10}_{ij(t-l)} + \mathbf{b}_5 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_5) \times \text{ozone}_{ij(t-l)} \end{aligned}$$

Box-Tidwell is not suitable for estimating multiple non-linear parameters. However, we can use Gauss-Seidel algorithm for each outer cycling and within each cycling, apply Box-Tidwell separately for each covariate. Keep cycling around until all  $\mathbf{a}$ 's converge. We choose the starting values to be those  $\hat{\mathbf{a}}$  when estimated separately, see table 1. The results are summarized in table 2.

	$\hat{\mathbf{a}}$	$sd(\hat{\mathbf{a}})$	log likelihood
temperature	0.0658	0.0258	-153779.19
pollen	0.7460	0.1048	
relative humidity	0	/	
PM <sub>10</sub>	0.4989	0.0769	
ozone	0.7514	0.0542	

Table 2. Summarization of  $\hat{\mathbf{a}}$  and its standard error when jointly estimated.

We found that except for relative humidity,  $\hat{\mathbf{a}}$ 's and  $sd(\hat{\mathbf{a}})$ 's are not far from the values when estimated separately. Log likelihood for this distributed lag model is -153779.19, which is significantly smaller than the log likelihood's of the models when only one  $\mathbf{a}$  is estimated at a time.

But we should notice that, if we go back to the profile, we can find that though the maximum likelihood is unique, the likelihood curve is usually not unimodal. There's often a small mode around  $\mathbf{a} = 0.1$ , or for temperature, the small mode is near  $\mathbf{a} = 0.7$ . There's reason to be concerned that if we start from different values and perform Gauss-Seidel and Box-Tidwell algorithm,  $\hat{\mathbf{a}}$ 's may converge to different sets of values. To check the robustness of the result, it is suggested that we use different sets of starting value and see where  $\mathbf{a}$  converge to. Because of the position of the modes in the likelihood profile, we choose either 0.1 or 0.7 to be the starting

values. There are five  $\mathbf{a}$ 's to be estimated, a total of  $2^5 = 32$  sets of starting values need to be checked. Let “+” to denote that  $\mathbf{a}$  starts from high value (0.7), “-” indicates that  $\mathbf{a}$  starts from low value (0.1).

Model	temp	pollen	relhum	PM <sub>10</sub>	ozone	log likelihood
1.	-	+	-	+	-	<i>-153779.19</i>
	+	+	-	+	+	
	-	+	-	+	+	
<b>a</b> converge to	<i>0.06</i>	<i>0.75</i>	<i>0</i>	<i>0.50</i>	<i>0.75</i>	
2.	+	+	-	-	+	<i>-153781.98</i>
	-	+	-	-	+	
<b>a</b> converge to	<i>0.06</i>	<i>0.73</i>	<i>0</i>	<i>0.06</i>	<i>0.70</i>	
3.	-	-	-	+	-	<i>-153782.62</i>
	-	-	-	+	+	
	+	-	-	+	+	
<b>a</b> converge to	<i>0.08</i>	<i>0</i>	<i>0</i>	<i>0.50</i>	<i>0.77</i>	
4.	-	-	-	-	+	<i>-153786.59</i>
5.	+	+	+	+	+	<i>-153787</i>
	-	+	+	-	+	
	-	+	+	+	+	
6.	+	-	+	+	+	<i>-153787.91</i>
	-	-	+	+	+	
	-	-	+	-	+	
7.	-	+	+	+	-	<i>-153790</i>
8.	-	-	+	+	-	<i>-153791.51</i>
9.	+	+	+	-	-	<i>-153793</i>
	-	+	+	-	-	
	-	+	-	-	-	
	+	+	+	-	+	
10.	-	-	+	-	-	<i>-153796.56</i>
11.	+	+	-	-	-	<i>-153797</i>
12.	+	-	+	-	+	<i>-153797.46</i>
13.	+	-	+	-	-	<i>-153798.34</i>
14.	-	-	-	-	-	<i>-153798.43</i>
15.	+	-	-	-	+	<i>-153800.67</i>
16.	+	+	-	+	-	<i>-153803.4</i>
17.	+	+	+	+	-	<i>-153805</i>
18.	+	-	-	-	-	<i>-153805.25</i>
19.	+	-	+	+	-	<i>-153806.57</i>
20.	+	-	-	+	-	<i>-153807.36</i>

Table 3. Different sets of starting values and the corresponding log likelihood value when  $\mathbf{a}$ 's converge. Listed from maximum to minimum.

We found from table 3 that some of the different sets of starting values converge to the same  $\mathbf{a}$ 's and hence yield the same log likelihood. The maximum likelihood value is -153779.19, just the one we obtained before and  $\hat{\mathbf{a}}$ 's converge to exactly the same value as we listed in table 2. When listing the likelihood from high to low, we found that the second and third likelihood values are very close to the first one.  $\hat{\mathbf{a}}$  for the second model is only different from model 1 in  $\text{PM}_{10}$ , where in model 1,  $\hat{\mathbf{a}}_{pm10}$  converge to 0.5 and in model 2,  $\hat{\mathbf{a}}_{pm10}$  is 0.06.  $\hat{\mathbf{a}}$  for model 3 is different from model 1 in the  $\mathbf{a}$  value of pollen,  $\hat{\mathbf{a}}_{pollen}$  in model 1 is 0.75 and 0 in model 3. In order to see whether model 1 is indeed the only best model, we can check the likelihood profile again (figure 5, 6 and 7).

For the profiles below, I circled the position on the likelihood curve where  $\mathbf{a}$  converges to. For model 2 and model 1,  $\hat{\mathbf{a}}_{pm10}$  converges to 0.06 and 0.50 with difference in log likelihood  $\sim 2$ . For model 3 and model 1,  $\hat{\mathbf{a}}_{pollen}$  is 0 and 0.75 individually, with difference in log likelihood  $\sim 2$  also. Then  $2 \times \log\text{likelihood}$  is approximately 4, this difference is relatively large enough to distinguish model 2 or model 3 with model 1. Hence, we can be quite confident that model 1 is indeed the best model with  $\hat{\mathbf{a}}$  converge to the correct place.

Now we can explore model 1 more carefully.

In table 4, odds ratios for  $\mathbf{b}_1, \dots, \mathbf{b}_5$  when jointly estimated are compared with the corresponding odds ratios when separately estimated:

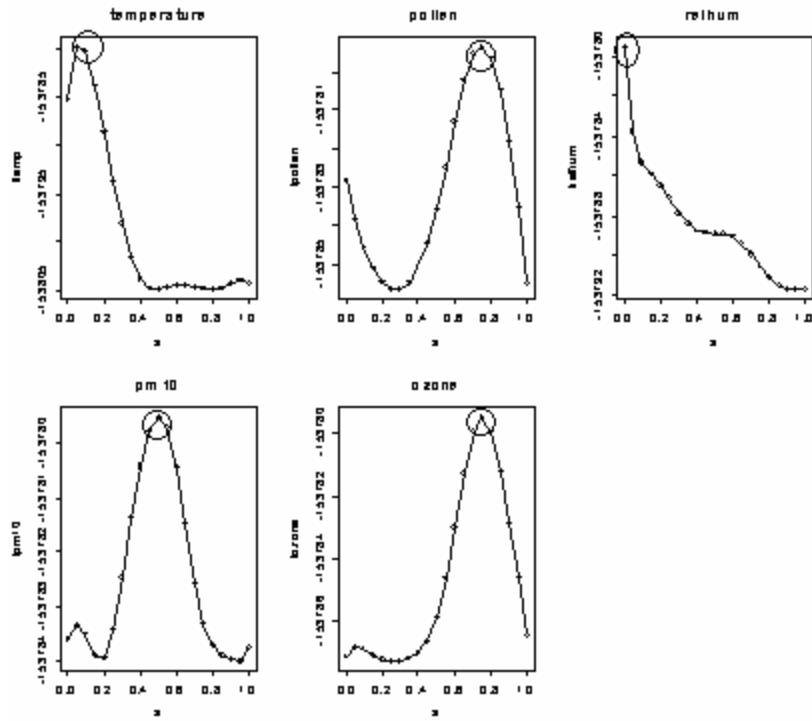


Figure 5. Model 1

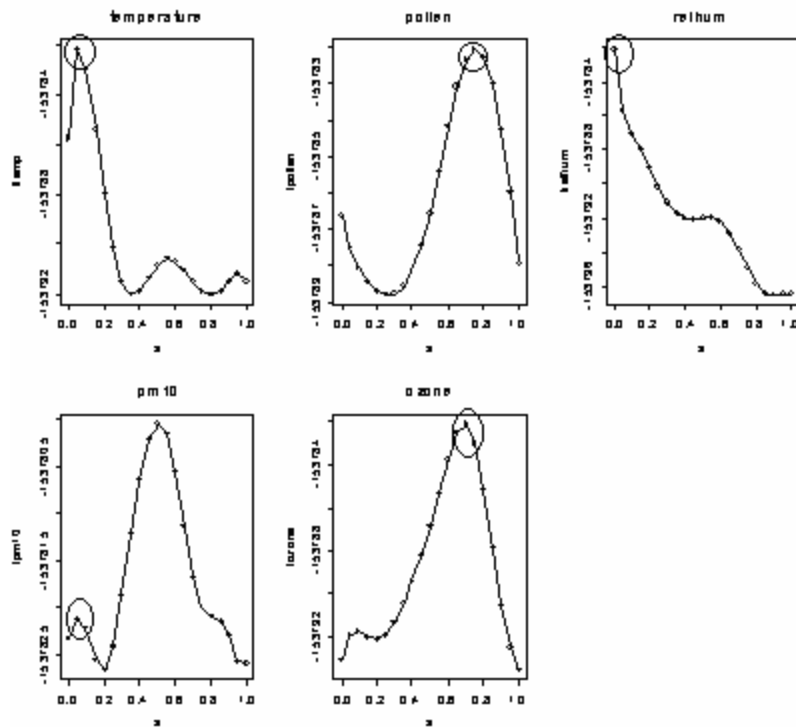


Figure 6. Model 2

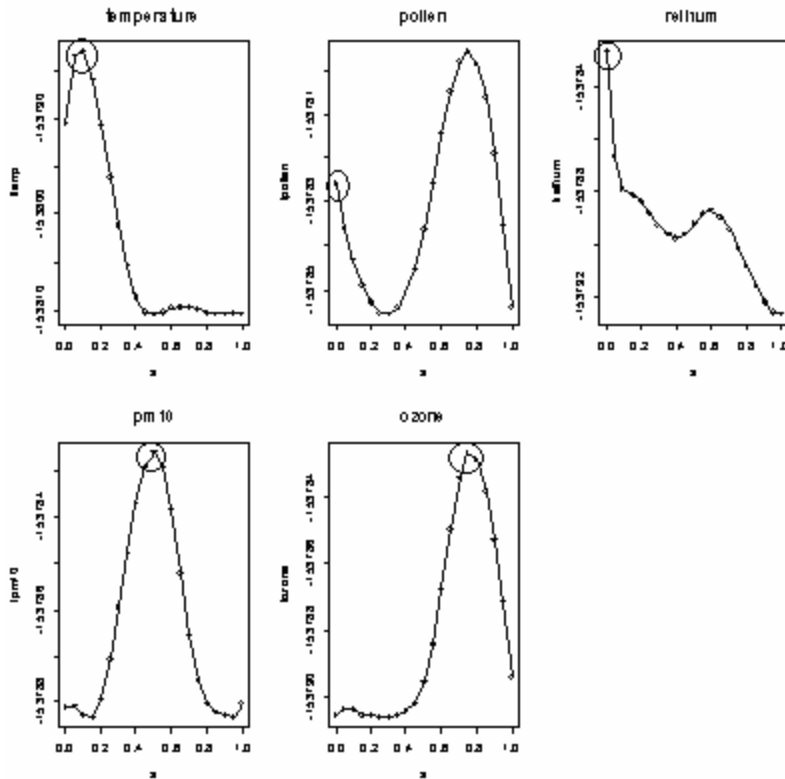


Figure 7. Model 3

	Odds Ratio for distributed lag models	
	Separately estimated	Jointly estimated
temperature (5F)	0.9810	0.9803
pollen (1 log (grains / m <sup>3</sup> ))	1.0183	1.0177
relative humidity (10%)	0.9746	0.9848
PM <sub>10</sub> (15 mg / m <sup>3</sup> )	1.0253	1.0156
ozone (20 ppb)	1.0433	1.0356

Table 4. Odds ratios for the five pollutants when jointly estimated and separately estimated

Fortunately, regardless whether estimated separately or jointly, the odds ratios for  $b_1, \dots, b_5$  always keep the same sign, but exhibit an overall trend of decreasing when jointly estimated, which means the effects are less strong, whether the effect is positive or negative, except for relative humidity. Now we can interpret the  $\hat{b}$ 's in the multiple distributed model in terms of

OR's and units of exposure variables. For a 5F increase in temperature, the OR is 0.9803. 1 unit increase in log-pollen indicates that the number of pollen grains/ $m^3$  is 2.718 times the previous number, this change results in OR of 1.0183. 10% increase in relative humidity will yield OR of 0.9848. For  $PM_{10}$ , we obtain OR of 1.0156 when there is a 15  $mg/m^3$  increase. 20 ppb increase in ozone will result in OR of 1.0356.

No strong collinearity is found after checking the correlation matrix.

	Temp	alpha1	pollen	alpha2	relhum	$PM_{10}$	alpha4	ozone	alpha5
temp	1								
alpha1	-0.21	1							
pollen	-0.07	-0.05	1						
alpha2	0.03	0.15	-0.02	1					
relhum	0	0.12	-0.07	-0.01	1				
$PM_{10}$	-0.06	0.22	0.01	-0.05	0.14	1			
alpha4	0.38	-0.23	-0.07	0.03	-0.01	0.13	1		
ozone	-0.08	0.02	-0.01	0.11	-0.12	-0.36	-0.44	1	
alpha5	0.02	-0.01	-0.04	-0.12	0.15	0.46	0.15	-0.32	1

Table 5. Correlation matrix

## 2. Test the significance of the covariates

To test the significance of the covariates in model (16)

$$H_0 : \mathbf{b} = 0 \quad \text{vs.} \quad H_1 : \mathbf{b} \neq 0$$

We cannot directly use the p-value and the standard error obtained after fitting the conditional logistic model. The reason why the p-values are wrong is because under the null hypothesis

when  $\mathbf{b} = 0$ , there's no information about  $\mathbf{a}$ , as the term  $\mathbf{b} \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}) \times \text{covariate}$  doesn't exist

anymore. While under the alternative, jointly estimating  $\mathbf{b}$  and  $\mathbf{a}$  with the data when the null hypothesis is correct will not yield a valid test.

To yield a valid test for  $\mathbf{b}$ , we can pick any value of  $\mathbf{a}$  for the covariate we want to test and after fixing this  $\mathbf{a}$ , use Gauss-Seidel and Box-Tidwell to estimate  $\hat{\mathbf{a}}$  for the other four covariates.

When convergence is reached, we can test the significance of the covariate whose value of  $\mathbf{a}$  is

fixed based on its p-value. Picking  $\mathbf{a}$  a-priori will yield a valid test though it's not most powerful because the result will change if we choose different priori values for  $\mathbf{a}$ . Notice  $\mathbf{a}$  can be any value between 0 and 1, under the null hypothesis they are all the same, though the “correct”  $\mathbf{a}$  will be most powerful under a given alternative hypothesis. In this analysis, we choose  $\mathbf{a}$  to be 0.5, which is just picked arbitrarily as if we knew nothing about the data set.

The result is summarized in the table 6:

	P-value
temperature	0.076
pollen	0.006
relative humidity	0.013
PM <sub>10</sub>	0.011
ozone	0.007

Table 6. Significance tests for the five covariates.  
For each test,  $\mathbf{a}$  is fixed at 0.5 for the covariate we are interested.

The results showed that except for temperature, all the other four pollutants are significant at 5% level. But we should bear in mind that this is only true when  $\mathbf{a} = 0.5$ .

### 3. Adding the smooth time function into the distributed lag model

Based on model (16), we can include a smooth function of time with 3 degrees of freedom and 1 extra linear term for each summer. The model is

$$\begin{aligned}
 \text{logit}(\mathbf{p}_{ijt}) = & R_{ijk} + \mathbf{b}_0 + \mathbf{b}_1 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_1) \times \text{temp}_{ij(t-l)} + \mathbf{b}_2 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_2) \times \text{pollen}_{ij(t-l)} \\
 & + \mathbf{b}_3 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_3) \times \text{relhum}_{ij(t-l)} + \mathbf{b}_4 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_4) \times \text{pm10}_{ij(t-l)} + \mathbf{b}_5 \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_5) \times \text{ozone}_{ij(t-l)} \quad (18) \\
 & + g(t) + I(\text{year} = 1996) \times t + I(\text{year} = 1997) \times t + I(\text{year} = 1998) \times t
 \end{aligned}$$

In our work,  $g(t)$  is modeled using a natural cubic spline.  $t$  is day of year, which spans from April 1<sup>st</sup> to October 31<sup>st</sup>. Regression splines can represent the fit as a piecewise polynomial, thus capture the local nature more exactly.

The time region (Apr. 1<sup>st</sup>—Oct. 31<sup>st</sup>) is separated by a sequence of knots or breakpoints,  $k_1, \dots, k_n$ . (see figure 8 for an example). In addition, it is customary to force the piecewise polynomials to join smoothly at these knots. Although many different configurations are possible, a popular choice consists of piecewise cubic polynomials constrained to be continuous and have continuous first and second derivatives at the knots.

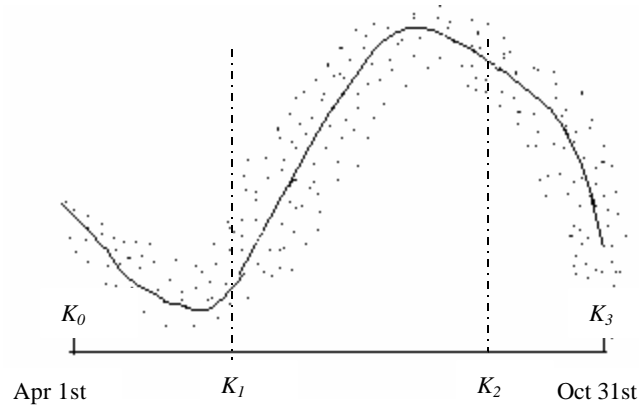


Figure 8. Piecewise cubic fits to time. The vertical lines indicate the location of the two knots.

We put one more constraint that the function is linear beyond the boundary knots. Because the spline is piecewise polynomial, if we don't put this constraint, the two ends will be very sensitive to the middle part. To enforce this condition we have to impose the constraint in each of the boundary regions:  $f'' = 0$ , which reduces the dimension of the space by 2. Then with  $K$  interior knots (and two boundary knots), the dimension of the space of fits is  $K+2$ .

In this analysis, we generate the basis matrix for the natural cubic spline with 3 degrees of freedom, thus only 1 interior knot.

Besides this, we also include

$$I(\text{year} = 1996) \times t, I(\text{year} = 1997) \times t, I(\text{year} = 1998) \times t$$

into the model, where

$$I(\text{year} = 1996), I(\text{year} = 1997), I(\text{year} = 1998)$$

are the indicator functions for year 1996, 1997 and 1998. So each year not only has its own level relative to the other years, but can also deviate in a linear fashion over time  $t$  from the other years. Comparing with the model without the smooth time function, this updated model can capture the confounding due to secular trends in exposure and asthma in the population. The 5-day time-window may be too long to capture the confounding across time within-person. This is another way to check the robustness of model (16).

We still choose the starting values to be the 32 different sets with  $\mathbf{a}$ 's start either from 0.1 or 0.7.

Model	temp	pollen	relhum	PM <sub>10</sub>	ozone	Log likelihood
1.	-	+	-	+	+	-153769
$\mathbf{a}$ converge to	0.07	0.76	0	0.52	0.77	
2.	-	+	-	+	-	-153770
$\mathbf{a}$ converge to	0.08	0.76	0	0.56	0.03	
3.	-	-	-	+	+	-153774
4.	-	-	-	-	+	-153774
5.	-	-	-	-	-	-153775
6.	-	+	-	-	+	-153776
7.	+	+	-	-	+	-153775
8.	-	-	-	+	-	-153776
9.	-	+	-	-	-	-153777
10.	-	+	+	+	+	-153779
11.	-	+	+	-	+	-153779
12.	-	-	+	+	+	-153779
13.	-	-	+	-	+	-153779
14.	+	-	+	+	+	-153779
15.	-	+	+	+	-	-153779
16.	-	-	+	+	-	-153780
17.	-	-	+	-	-	-153780
18.	+	+	-	-	-	-153781
19.	+	+	+	-	-	-153783
20.	-	+	+	-	-	-153784
21.	+	+	+	-	+	-153784
22.	+	-	-	-	+	-153788
23.	+	+	-	+	-	-153788
24.	+	-	+	-	+	-153790
25.	+	+	-	+	+	-153792

26.	+	-	+	-	-	-153792
27.	+	-	-	-	-	-153795
28.	+	-	-	+	-	-153795
29.	+	-	-	+	+	-153797
30.	+	+	+	+	-	-153797
31.	+	+	+	+	+	-153798
32.	+	-	+	+	-	-153800

Table 7. Different sets of starting values and the log likelihood value when  $\mathbf{a}$  converge. Listed from maximum to minimum. “+” means starting from 0.7, “-” means starting from 0.1.

We found that the likelihood values for the first and second model are very close and relatively larger than the rest. The  $\mathbf{a}$  values they converge to are only different in ozone. In model 1,  $\hat{\mathbf{a}}_{ozone}$  is 0.77 and in model 2 it is 0.03, and the values are very sensitive to the starting values. Again, we choose to check the likelihood profile for model 1 and 2 (see figure 9 and 10).

I circled the position on the likelihood curve where  $\mathbf{a}$  converge. For model 2 and model 1,  $\hat{\mathbf{a}}_{ozone}$  converges to 0.03 and 0.77 with difference in log likelihood less than 1. This difference is not large enough to distinguish model 2 with model 1. We conclude that model 1 and model 2 are equally good.

Next, we examine model 1 and 2 more carefully.

In table 8, I listed the value of  $\hat{\mathbf{a}}$  and its standard error together with the log likelihood values for the corresponding models. In table 9, I compared the odds ratios of the two models.

	Model 1		Model 2	
	$\hat{\mathbf{a}}$	$sd(\hat{\mathbf{a}})$	$\hat{\mathbf{a}}$	$sd(\hat{\mathbf{a}})$
temperature	0.073	0.022	0.078	0.022
pollen	0.760	0.104	0.756	0.100
relative humidity	0	/	0	/
PM <sub>10</sub>	0.521	0.060	0.565	0.045
ozone	0.769	0.104	0.024	0.089
log likelihood	-153769.73		-153770.58	

Table 8. Summarization of  $\hat{\mathbf{a}}$  and its standard error for model 1 and 2 which includes the smooth time function.

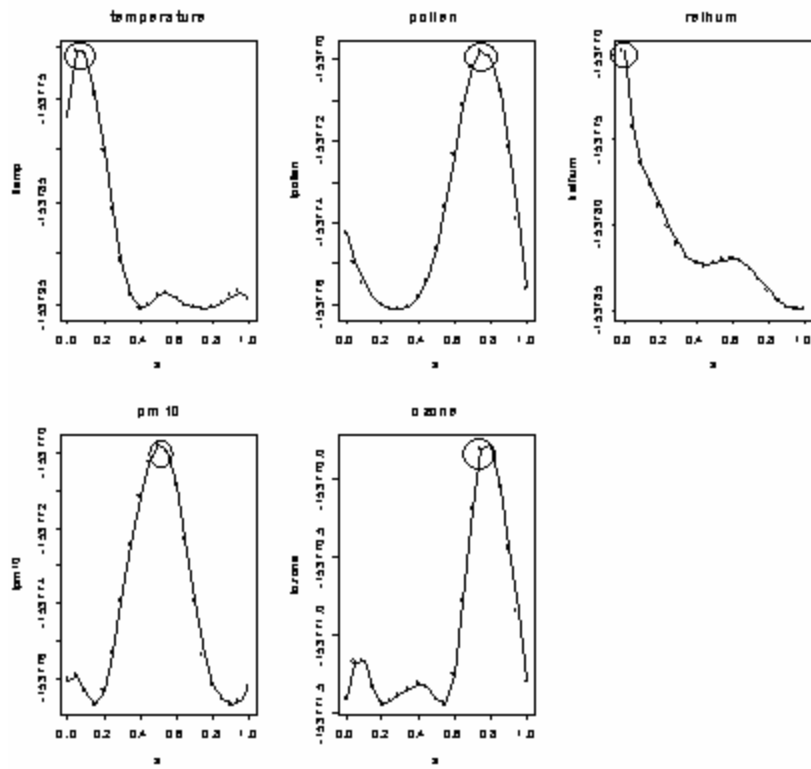


Figure 9. Model 1 (including smooth time function)

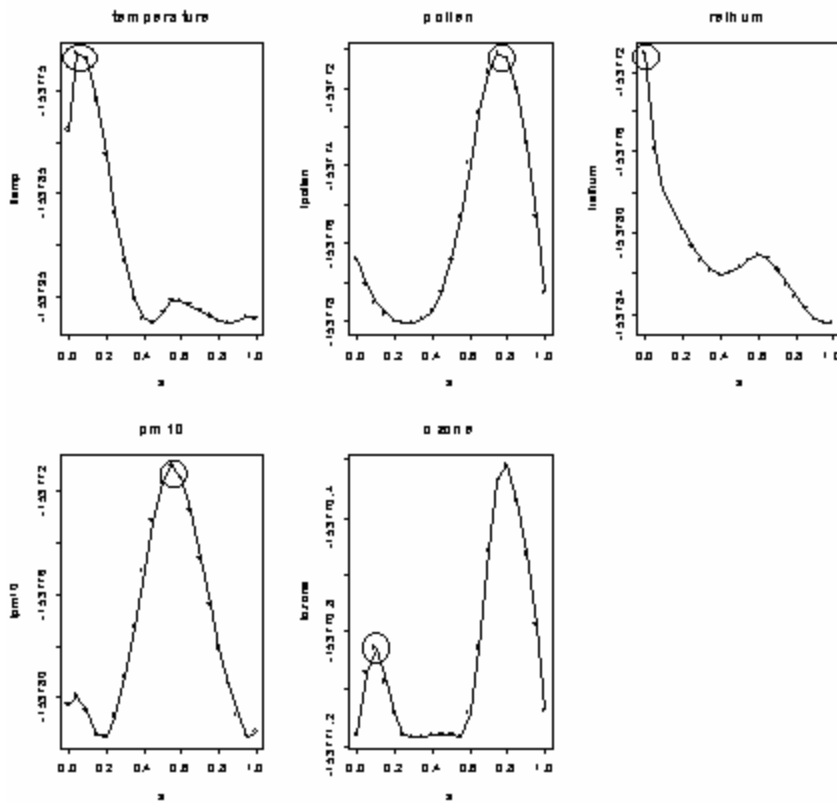


Figure 10. Model 2 (including smooth time function)

	Odds Ratio for distributed lag models with time function	
	Model 1	Model 2
temperature (5F)	0.976	0.973
pollen (1 log (grains / $m^3$ ))	1.018	1.019
relative humidity (10%)	0.982	0.983
PM <sub>10</sub> (15 $mg / m^3$ )	1.020	1.024
ozone (20 ppb)	1.018	1.008

Table 9. Compare the odds ratio for model 1 and model 2.

From table 8 and table 9, it's very obvious that  $\hat{\alpha}$ 's are only different in ozone, thus the corresponding odds ratios for temperature, pollen, relative humidity and PM<sub>10</sub> are almost the same. But the odds ratio for ozone is much smaller when  $\hat{\alpha}_{ozone}$  is 0.03 than when it is 0.77. Now a 5F increase in temperature will result in odds ratio of  $\sim 0.97$ . As explained before, when there's a one unit increase in log-pollen (number of pollen grains /  $m^3$  is 2.718 times the previous value), the odds ratio related to this change is about 1.02 for both models. A 10% increase in relative humidity gives odds ratio of 0.98 in both models as well. PM<sub>10</sub> has odds ratio around 1.02 for an increase of 15  $mg / m^3$ . Ozone, when  $\hat{\alpha}$  is 0.769, the odds ratio is 1.018; when  $\hat{\alpha}$  is 0.024, the odds ratio is 1.008, both are related to a 20 ppb increase in the ozone level.

We can also compare the results with the one obtained from the distributed lag model without the time function (table 2). It is not surprising to find that there's no big change in the odds ratios for the pollutants except ozone, while the estimated odds ratio for ozone is delicate in different models, in the original model it is 1.043 and 1.018 in model 1 and only 1.008 in model 2. Thus by adding the smooth time function, the estimated effect of ozone becomes much less strong.

## VIII. DISCUSSION

From the analysis above, we found that temperature and relative humidity have the most important effects on the asthma occurrence at small lags. And actually, relative humidity doesn't

have a lagged structure at all, we include and only include the observation of relative humidity on the current day to predict the response. Since  $\hat{\mathbf{a}}_{relhum}$  is 0, the first derivative of

$\mathbf{b}_{relhum} \sum_{l=0}^{14} \mathbf{w}_l(\mathbf{a}_{relhum}) \times relhum$  doesn't exist, so we can't estimate the standard error of  $\hat{\mathbf{a}}_{relhum}$ .

$\hat{\mathbf{a}}_{temp}$  is around 0.07, which means temperature with 1-day lag is most significantly related to the event. Temperature and relative humidity are the only two covariates with negative effects on asthma occurrence. With one unit increase in the weighted average of temperature or relative humidity, the odds is ~98% of the original odds, which means the probability of having asthma decreased. For temperature, one unit increase is actually an increase of 5 F for temperature centered at 60 F. One unit increase for relative humidity is a 10% increase at the true level centered at 70.

Pollen and PM<sub>10</sub> affect the response with large lags.  $\hat{\mathbf{a}}_{pollen}$  is ~-0.76 with odds ratio ~1.02 and  $\hat{\mathbf{a}}_{pm10}$  is ~-0.50 with odds ratio ~1.02. They all have positive effects. One unit increase in standardized log-pollen, or one unit increase in rescaled PM<sub>10</sub> – 15 mg/m<sup>3</sup> increase in actual level with baseline 30 mg/m<sup>3</sup>, the log odds will be 1.02 the original log odds. The probability of having asthma increases.

Above conclusion is applicable no matter for which model — the distributed lag model with or without time function. While for ozone, we need a further look since its estimated effect seems to be sensitive to the model we use. But its effect is always positive. In the no-time-function model,  $\hat{\mathbf{a}}_{ozone}$  is 0.75—effect is most obvious with large lags. One unit increase in the rescaled ozone variable results in 20 ppb increase in the actual ozone level centered at 30 ppb, the odds ratio is 1.04. In the time-function model,  $\hat{\mathbf{a}}_{ozone}$  can be either 0.77 or 0.02, which means exposure to

ozone at 1-day or 10/11-day lag both have a significant effect on daily asthma claim. The odds ratio is reduced from 1.04 to either 1.02 or 1.01. This change indicates that there might be some time trends with ozone in the 5-day time window, and the 5-day time window may not be small enough to control the confounders due to the time trends. We can capture the residual confounding by including the smooth time function, thus the effect of ozone becomes a little bit weaker.

How to test the significance of the covariates is another important issue in this analysis. Here, we choose  $\alpha$  to be fixed at a specific value. This test is valid but may not be very powerful. There is reason to believe that there must be some other more reasonable and powerful way to perform the significance test.

## **Acknowledgement**

Thanks for Dr. Vanja M. Dukic and Dr. Mei Wang for finalizing the paper and offering important advice.

## **Reference**

1. Holly Janes, Lianne Sheppard, Thomas Lumley. Referent selection in case-crossover analysis of the health effects of air pollution. 2003
2. J Schwartz, C Spix, G Touloumi, L Bacharova, T Barumamdzadeh, A le Tertre, T Piekarksi, A Ponce de Leon, A Ponka, G Rossi, M Saez, J P Schouten. Methodological issues in studies of air pollution and daily counts of deaths or hospital admission. J Epidemiol Community Health 1996;50 Suppl 1:S3-11
3. Leah J.Welty, Scott L. Zeger. Flexible distributed lag models: Are the acute effects of PM<sub>10</sub> on mortality the result of inadequate control for weather and season? Johns Hopkins University, Dept. of Biostatistics Working Papers 2004:38

4. Malcolm Maclure. The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* 1991; 133: 144-153
5. N.E. Breslow, N.E. Day. *Statistical methods in cancer research*. International Agency for Research on Cancer. 1980
6. Paul Rathouz. Longitudinal and spatial analysis of short-term respiratory health effects of air pollution.
7. P. McCullagh, J.A. Nelder. *Generalized linear models*. Chapman & Hall. 1999
8. Schwartz, Joel. The distributed lag between air pollution and daily deaths. *Epidemiology* 2000;11(3):320-326
9. Thomas Lumley, Drew Levy. Bias in the case-crossover design: implications for studies of air pollution. *Environmetrics* 2000;11:689-704
10. T.J. Hastie, R.J. Tibshirani. *Generalized additive models*. Chapman & Hall. 1994
11. Vanja Dukic, Paul Rathouz, Dana Draghicescu, Edward Naureckas, Gidon Eshel, Alexis Zubrow, John Frederick. Short-term respiratory health effects of air pollution in metropolitan Chicago. University of Chicago. Draft preliminary report.
12. William Navidi. Bidirectional case-crossover designs for exposures with time trends. *Biometrics* 1998;54:596-605