November 2007

DECOMPOSITION OF REGRESSION ESTIMATORS TO EXPLORE THE INFLUENCE OF "UNMEASURED" TIME-VARYING CONFOUNDERS

Yun Lu
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, ylu@jhsph.edu

Scott L. Zeger
Department of Biostatistics, The Johns Hopkins Bloomberg School of Public Health

Decomposition of Regression Estimators to Explore the Influence of "Unmeasured" Time-Varying Confounders

Yun LU, and Scott L. ZEGER

Yun Lu is Graduate Student (email: ylu@jhsph.edu), and Scott L. Zeger is Professor (email: szeger@jhsph.edu), Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205. The authors are grateful to partial support from the National Institute for Environmental Health Sciences grant ES012054-03 and the NIEHS Center in Urban Environmental Health grant P30 ES 03819.

ABSTRACT

In environmental epidemiology, exposure X and health outcome Y vary in space and time. We present a method to diagnose the possible influence of unmeasured confounders U on the estimated effect of X on Y and to propose several approaches to robust estimation. The idea is to use space and time as proxy measures for the unmeasured factors U. We start with the time series case where X_t and Y_t are continuous variables at equally-spaced times and assume a linear model. We define matching estimator $\tilde{\beta}(u)$s that correspond to pairs of observations with specific lag u. Controlling for a smooth function of time, S_t, using a kernel estimator is roughly equivalent to estimating β with a linear combination of the $\tilde{\beta}(u)$s with weights that involve two components: the assumptions about the smoothness of S_t and the normalized variogram of the X process. When an unmeasured confounder U exists, but the model otherwise correctly controls for measured confounders, the excess variation in $\tilde{\beta}(u)$s is evidence of confounding by U. We use the plot of $\tilde{\beta}(u)$s versus lag u, lagged-estimator-plot (LEP), to diagnose the influence of U on the effect of X on Y. We use appropriate linear combination of $\tilde{\beta}(u)$s or extrapolate to $\tilde{\beta}(0)$ to obtain novel estimators that are more robust to the influence of smooth U. The methods are extended to time series log-linear models and to spatial analyses. The LEP plot gives us a direct view of the magnitude of the estimators for each lag u and provides evidence when models did not adequately describe the data.

KEY WORDS: confounding, coefficient decomposition, time series, log-linear model, matching estimator, Laggard-Estimator-Plot

1 INTRODUCTION

Particulate matter (PM) air pollution is a public health problem around the world. In developing countries, with the rapid urbanization and industrialization, the PM air pollution has worsened, reaching concentrations at which serious adverse health outcomes are well documented (Aekplakorn et al., 2003; Chhabra et al., 2001; Ostro et al., 1999a, 1999b; Vichit-Vadakan et al., 2001). In developed countries, in spite of declining PM concentrations during the past 20 years due to stricter air quality standard, adverse health effects of particulate air pollution remain a regulatory and public health concern (Dominici et al., 2006, 2007; Jerrett et al., 2005a, 2005b; Ostro et al., 2006; Samet et al., 2000a, 2000b).

There are two major sources of evidence about the relationship of air pollution and health outcomes: cohort and time series studies. Prospective cohort studies provide important evidence on the long-term risks of particulate matter by comparing mortality rates adjusted for personal characteristic across cities with different long-term average levels of pollution. However, only a small number of cohort studies have been carried out due to the long follow up time, the complexity and costs of such studies. These studies can also be confounded by other unmeasured dissimilarities among city populations being compared.

Time series studies of adverse health effects (e.g. hospitalization and death) compare the same population on different days with varying pollution levels, thereby avoiding confounding by unmeasured population differences. However, this association can be confounded by time-varying factors (Schwartz et al. 1996; Katsouyanni et al. 1996; Peng et al. 2005). We can control the effect of measured confounders by including them in a regression model or by matching. Unmeasured confounders

U are variables which either cannot be measured directly or had not been controlled for in the study design. When U is associated with both outcome and PM, failure to take U into account will bias our estimation of the true association, either upward or downward.

One commonly used approach for time series of mortality or morbidity counts is to fit a Poisson log-linear model with linear terms of PM air pollution levels and smooth functions of time to adjust for the time-varying confounders (McCullagh and Nelder, 1989; Hastie and Tibshirani, 1990, 1995; Dominici et al., 2004). In such a model, we control for the effect of unmeasured confounders that vary smoothly in time by including functions of time in the model as proxies.

The case-crossover design has been increasingly applied to epidemiologic studies to investigate the association between short-term exposure to ambient air pollution and acute adverse health effects. We compare a case's exposure during the event interval with that same person's exposure at "otherwise similar reference" times and estimate an odds ratio as the measure of association using conditional logistic regression (Maclure 1991; Bateson and Schwartz, 1999). The case-crossover design is appealing because it involves cases only and it controls for the effect of unmeasured confounders by matching.

Lu and Zeger (2007) have shown that the case-crossover method is a special case of a time series log-linear model. Both methods control for confounding in their respective regression models, and it is equally important for both methods to evaluate key modeling assumptions about the nuisance function that represents the effect of potential temporal confounding.

In this paper, we propose a new model checking method: decomposition of re-

gression estimators by lag or distance to explore the possible influence of unmeasured confounders on the estimation of the true association. The first step is to decompose the data into pairs to obtain pairwise regression coefficients. The second step is to combine our pairwise coefficients with the same lag u to construct matching estimator $\tilde{\beta}(u)$. We proposed the LEP (Lagged-Estimator-Plot) as a new model-checking method, which is the graph of $\tilde{\beta}(u)$ versus lag u. When the model adequately captures the structure of unmeasured confounders, LEP will be a roughly horizontal line. In this article, we introduced the LEP plot for time series linear model, compared the performance of several proposed estimators using simulation study, and extended the results to the log-linear time series model and to log-linear spatial model.

2 METHODS AND RESULTS

2.1 Linear Time Series Model

2.1.1 Model

Before we proceed to the log-linear model motivated by our time series application, we start with the simple case of a Gaussian time series. Suppose we have time series data, generated from the true model $Y_t = \beta_0 + \beta X_t + S_t + \varepsilon_t = \mu_t + S_t + \varepsilon_t$, where Y_t is a health outcome, X_t is our risk factor of interest such as PM level, S_t is the effect of unmeasured confounders, and ε_t is an independent $N(0, \sigma^2)$ deviation.

If we ignore S_t and fit **Model I:** $E(Y_t) = \beta_0 + \beta X_t$, we obtain the least squares estimator $\hat{\beta}$. The conditional expectation of $\hat{\beta}$ given $\mathbf{X} = (X_1, \ldots, X_T)$ and $\mathbf{S} =$

$(S_1, \ldots, S_T)$ is

$$E\left(\hat{\beta}|\mathbf{X}, \mathbf{S},\right) \;=\; \beta - \frac{\sum_{i<j}(S_i - S_j)(X_i - X_j)}{\sum_{i<j}(X_i - X_j)^2}.$$

The bias depends on the covariance between the independent variable $\mathbf{X}$ and the time-varying $\mathbf{S}$. If $\mathbf{X}$ and $\mathbf{S}$ satisfy $\sum_t (X_t - \bar{X})(S_t - \bar{S}) = 0$, then the simple regression estimator will be unbiased. A simple special case is when $S_i = S_j$ for all i and j, ie. S_t is constant. However, there usually exists long-term or seasonal trends in both $\mathbf{X}$ and $\mathbf{S}$ for time series data, making the cross-product non-zero, in which case the estimator $\hat{\beta}$ from Model I is biased.

2.1.2 Coefficient decomposition

If we look at each data pair (X_i, Y_i) and (X_j, Y_j) individually, we can crudely estimate β using a pairwise regression estimator $\hat{\beta}_{i,j} = (Y_i - Y_j)/(X_i - X_j)$, which will be unbiased if $S_i = S_j$. If given $\mathbf{X}$ and $\mathbf{S}$, the ε_t are iid $N(0, \sigma^2)$, the variance of $\hat{\beta}_{i,j}$ is $Var(\hat{\beta}_{ij}) = 2\sigma^2/(X_i - X_j)^2$. The estimator $\hat{\beta}_{i,j}$ has been called "elemental regression" for simple linear regression (Mayo and Gray, 1995; Sheynin, 1973). Back in 1841, Jacobi first reported that the least squares estimator $\hat{\beta}$ can be written as the weighted average of $\hat{\beta}_{ij}$ for all i, j, with the weight proportional to $Var(\hat{\beta}_{ij})^{-1}$: $\hat{\beta} = \sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y}) / \sum_{t=1}^{T}(X_t - \bar{X})^2 = \sum_{i<j}(X_i - X_j)^2 \hat{\beta}_{ij}$ (Sheynin, 1973). The estimator $\hat{\beta}$ uses information from all possible data pairs. However, $\hat{\beta}$ is biased in the presence of unmeasured time-varying confounders.

Matching has been used to control for potential confounders. If we match day t with days $t + u$ and $t - u$, we can combine pairs of observations with lag u to obtain

matching estimator $\tilde{\beta}(u)$. Here

$$\begin{aligned}\tilde{\beta}(u) &= \frac{\sum_{t=1}^{T-u}(Y_t - Y_{t+u})(X_t - X_{t+u})}{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2} \\ &= \sum_{t=1}^{T-u}\left(\frac{Y_t - Y_{t+u}}{X_t - X_{t+u}}\right)\left(\frac{(X_t - X_{t+u})^2}{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2}\right) \\ &= \sum_{t=1}^{T-u}\hat{\beta}_{t,t+u}w_t(u), \end{aligned} \tag{2.1}$$

where $w_t(u) = (X_t - X_{t+u})^2 / \sum_{t=1}^{T-u}(X_t - X_{t+u})^2$.

We can also use matrix notation to get (See Appendix I for details)

$$\tilde{\beta}(u) = \frac{\mathbf{X}^t{\mathbf{D_u}}^t\mathbf{D_u}\mathbf{Y}}{\mathbf{X}^t{\mathbf{D_u}}^t\mathbf{D_u}\mathbf{X}} = \mathbf{H_u}\mathbf{Y}, \tag{2.2}$$

where $\mathbf{H_u} = \mathbf{X}^t{\mathbf{D_u}}^t\mathbf{D_u}/\mathbf{X}^t{\mathbf{D_u}}^t\mathbf{D_u}\mathbf{X}$ is a vector of length T. Let $\tilde{\beta} = (\tilde{\beta}(1), \ldots, \tilde{\beta}(T-1))^t$ and $\mathbf{H}$ be the $(T-1) \times T$ matrix with $\mathbf{H_u}$ as the uth row. Then we have $\tilde{\beta} = \mathbf{HY}$, which follows multivariate normal distribution and its covariance matrix is $\Sigma = \sigma^2\mathbf{H}\mathbf{H}^t$.

The bias for $\tilde{\beta}(u)$ will vary with lag u. If S_t is a smooth function of time, in particular, smoother than X_t, we would expect $\tilde{\beta}(u)$ to have little bias for small u. Note that the least squares estimator $\hat{\beta}$ can be expressed as a weighted average of

$\tilde{\beta}(u)$

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{u=1}^{T-1}\sum_{t=1}^{T-u}(Y_t - Y_{t+u})(X_t - X_{t+u})}{\sum_{u=1}^{T-1}\sum_{t=1}^{T-u}(X_t - X_{t+u})^2} \\
&= \sum_{u=1}^{T-1} \frac{\sum_{t=1}^{T-u}(Y_t - Y_{t+u})(X_t - X_{t+u})}{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2} \frac{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2}{\sum_{u=1}^{T-1}\sum_{t=1}^{T-u}(X_t - X_{t+u})^2} \\
&= \sum_{u=1}^{T-1} \tilde{\beta}(u) w(u), \qquad (2.3)
\end{aligned}$$

where $w(u) = \sum_{t=1}^{T-u}(X_t - X_{t+u})^2 / \sum_{u=1}^{T-1}\sum_{t=1}^{T-u}(X_t - X_{t+u})^2$. Here we can see that the least squares estimator $\hat{\beta}$ is a linear combination of matching estimators ($\tilde{\beta}(u)$s) that compare day t to days $t+u$ and $t-u$. How to optimally combine the $\tilde{\beta}(u)$ involves a trade-off of bias and variance as described below.

Another method to control for potential confounders is by modeling S_t. We can fit the following **Model II:** $E(Y_t) = \beta_0 + \beta X_t + S_t$, where S_t is estimated by $\hat{S}_t$, a smooth function of t with ν degrees of freedom. The estimator $\hat{\beta}_{\hat{\mathbf{S}}}$ is the same as regressing $Y_t^* = Y_t - \hat{S}_t$ on X_t, where $\hat{\mathbf{S}} = (\hat{S}_1, \ldots, \hat{S_{T-1}})$. We can again using the pairs with lag u to obtain

$$\tilde{\beta}_{\hat{\mathbf{S}}}(u) = \frac{\sum_{t=1}^{T-u}(Y_t - \hat{S}_t - Y_{t+u} + \hat{S}_{t+u})(X_t - X_{t+u})}{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2}. \qquad (2.4)$$

Similar to Equation 2.3, the estimator $\hat{\beta}_{\hat{\mathbf{S}}}$ can be expressed as the weighted average of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$, $\hat{\beta}_{\hat{\mathbf{S}}} = \sum_{u=1}^{T-1} \tilde{\beta}_{\hat{\mathbf{S}}}(u) w(u)$.

If $\hat{S}_t$ adequately captures the structure of S_t, we would expect $Y_t^* = Y_t - \hat{S}_t$ to be uncorrelated with the unmeasured time-varying confounders. The matching estimators $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ would be unbiased for all us, thus resulting in an unbiased $\hat{\beta}_{\hat{\mathbf{S}}}$.

As an obvious extension, we have measured confounders Z_t, we can fit **Model III:** $E(Y_t) = \beta_0 + \beta X_t + \gamma Z_t + S_t$, and obtain $\tilde{\beta}_{\hat{\mathbf{S}},\mathbf{Z},}(u)$ using $Y_t^* = Y_t - \hat{\gamma} Z_t - \hat{S}_t$ and X_t.

2.1.3 Connections between $\tilde{\beta}(u)$ and $\hat{\beta}_{\hat{\mathbf{S}}}$

As we mentioned before, unmeasured confounder can be controlled by matching or modeling. Lu and Zeger (2007) have established connections between case-crossover design and time series log-liner model for counts outcome, which are examples of controlling by matching and modeling. An obvious question is: what is the connection between matching and modeling for this linear model?

The estimator $\tilde{\beta}(u)$ can be obtained by matching day t with days $t+u$ and $t-u$, while $\hat{\beta}_{\hat{\mathbf{S}}}$ can be calculated by modeling $\mathbf{S}$ with $\hat{\mathbf{S}}$. In this section, we establish a connection between $\tilde{\beta}(u)$ and $\hat{\beta}_{\hat{\mathbf{S}}}$.

Suppose we fit a model $E(Y_t) = \beta_0 + \beta X_t + S_t$ for $t = 1, \ldots, T$. Here we use symmetric weighted running mean smoother to estimate S_t by defining $\hat{S}_t = \tilde{Y}_t = \sum_{u=-t+1}^{T-t} \lambda_u Y_{t+u}$ with $\sum_{u=-t+1}^{T-t} \lambda_u = 1$, where λ_u is symmetric, i.e. $\lambda_u = \lambda_{-u}$. We usually have $\lambda_u = 0$ for $u > k$, where k depends on the number of degrees of freedom in smoothing. For $t < k+1$ and $t > T-k$, there exists edge effect.

We can write $\hat{\beta}_{\hat{\mathbf{S}}}$ as the following expression (See APPENDIX for details)

$$\begin{aligned}
\hat{\beta}_{\hat{\mathbf{S}}} &= \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \hat{S}_t - \bar{Y})}{\sum_{t=1}^{T}(X_t - \bar{X})^2} \\
&= \frac{1}{\sum_{t=1}^{T}(X_t - \bar{X})^2}\left\{\sum_{t=1}^{T} Y_t\left[X_t - \sum_{u=-t+1}^{T-t}\lambda_u X_{t+u}\right]\right\} \\
&= \frac{1}{\sum_{t=1}^{T}(X_t - \bar{X})^2}\left\{\sum_{t=1}^{T} Y_t\left[X_t - \tilde{X}_t\right]\right\}, \qquad (2.5)
\end{aligned}$$

where $\tilde{X}_t = \sum_{u=-t+1}^{T-t}\lambda_u X_{t+u}$ is the symmetric weighted running mean smoother of X_t using the same smoothing method as estimating $\hat{S}_t = \tilde{Y}_t$. This result holds even when there exists edge effect, as far as we use the same smoothing method for both $\tilde{X}_t$ and $\tilde{Y}_t$.

The matching estimator $\tilde{\beta}(u)$ can be written as

$$\begin{aligned}
\tilde{\beta}(u) &= \frac{\sum_{t=1}^{T-u}(Y_t - Y_{t+u})(X_t - X_{t+u})}{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2} \\
&= \frac{1}{\sum_{t=1}^{T} X_t\left[X_t - \tilde{X}_t(u)\right]}\sum_{t=1}^{T} Y_t\left[X_t - \tilde{X}_t(u)\right], \qquad (2.6)
\end{aligned}$$

where $\tilde{X}_t(u) = (2X_t + X_{t+u})/3$ for $t < u+1$, $\tilde{X}_t(u) = (X_{t-u} + X_t + X_{t+u})/3$ for $u+1 \le t \le T-u$, and $\tilde{X}_t(u) = (2X_t + X_{t-u})/3$ for $t > T-u$. It can be shown that $\alpha(u)\tilde{\beta}(u)$ is exact the same as $\hat{\beta}_{\hat{\mathbf{S}}}$ using $\hat{S}_t = \tilde{Y}_t(u)$, where

$$\begin{aligned}
\alpha(u) &= \frac{\sum_{t=1}^{T} X_t(X_t - \tilde{X}_t(u))}{\sum_{t=1}^{T} X_t(X_t - \bar{X}_t)} \\
&= 1 - \frac{\sum_{t=1}^{T} X_t(\tilde{X}_t(u) - \bar{X}_t)}{\sum_{t=1}^{T} X_t(X_t - \bar{X}_t)}.
\end{aligned}$$

tional to estimating S_t using running average of Y_{t-u}, Y_t, and Y_{t+u}. The factor $\alpha(u)$ depends on the pattern of X_t and $\tilde{X}_t(u)$.

For the least square estimator $\hat{\beta}$, we have

$$\hat{\beta} = \frac{1}{\sum_{t=1}^{T}(X_t - \bar{X})^2}\left\{\sum_{t=1}^{T} Y_t(X_t - \bar{X})\right\},$$

hence $\hat{\beta}$ corresponds to using $\hat{S}_t = \bar{Y}$, with $\hat{\beta}_0 = \bar{Y} - \bar{S} = 0$.

We extend the data beyond [1, T] by $X_{t+u} = X_{t+u-T}$ for $t + u > T$ and $X_{t-u} = X_{t-u+T}$ for $t - u < 1$ to avoid edge effect. The data were analyzed using circular pattern, here $\lambda_u = \lambda_{-u} = \lambda_{T-u} = \lambda_{u-T}$. The matching estimator $\tilde{\beta}(u)$ can be written as

$$\begin{aligned}\tilde{\beta}(u) &= \frac{\sum_{t=1}^{T}(Y_t - Y_{t+u})(X_t - X_{t+u})}{\sum_{t=1}^{T}(X_t - X_{t+u})^2} \\ &= \frac{1}{\sum_{t=1}^{T} X_t\left[X_t - \tilde{X}_t(u)\right]}\sum_{t=1}^{T} Y_t\left[X_t - \tilde{X}_t(u)\right], \qquad (2.7)\end{aligned}$$

where $\tilde{X}_t(u) = (X_{t-u} + X_t + X_{t+u})/3$.

Denote $V(u) = \sum_{t=1}^{T}(X_t - X_{t+u})^2$ and $V_T = \sum_{u=1}^{T-1}\sum_{t=1}^{T}(X_t - X_{t+u})^2 = 2T\sum_{t=1}^{T}(X_t - \bar{X})^2$ (Note that $V_T = T\sum_{t=1}^{T}(X_t - \bar{X})^2$ if we don't use data circularly). For $\sum_{u=-t+1}^{T-t}\lambda_u = 1$, we have $\lambda_0 + \sum_{u=-T+1}^{T-1}\lambda_u = 2\sum_{u=-t+1}^{T-t}\lambda_u = 2$. We can obtain the linear combination of $\tilde{\beta}(u)$s

$$\begin{aligned}\hat{\beta}_{\mathbf{V}} &= \sum_{u=1}^{T-1}\frac{V(u)}{V_T/T}\lambda_u\tilde{\beta}(u) \\ &= \frac{1}{\sum_{t=1}^{T}(X_t - \bar{X})^2}\sum_{t=1}^{T}\left\{Y_t\left[X_t - \sum_{u=-t+1}^{T-t}\lambda_u X_{t+u}\right]\right\}, \qquad (2.8)\end{aligned}$$

which is exactly the same as estimator $\hat{\beta}_{\hat{\mathbf{S}}}$ using smoothing function $\hat{S}_t = \sum_{u=-t+1}^{T-t} \lambda_u Y_{t+u}$ (See APPENDIX for details). Hence the estimator $\hat{\beta}_{\hat{\mathbf{S}}}$ obtained by modeling S_t with $\hat{S}_t = \tilde{Y}_t = \sum_{u=-t+1}^{T-t} \lambda_u Y_{t+u}$ can be calculated as a linear combination of $\tilde{\beta}(u)$s, where the weight is proportional to λ_u and the variogram for each lag u. If we don't use data circularly, the result is approximate due to the edge effect.

2.1.4 Model-checking and Proposed Estimators

In the previous sections, we introduced a matching estimator $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ after modeling $\mathbf{S}$ using $\hat{\mathbf{S}}$. The estimator $\hat{\beta}_{\hat{\mathbf{S}}}$ can be written as a linear combination of $\tilde{\beta}_(u)$s, and the weight is related to how we smooth $\mathbf{Y}$ to obtain $\hat{\mathbf{S}}$. There are two questions we want to answer in this section: (1) How to assess whether $\hat{S}_t$ adequately captures the structure of S_t? (2) If $\hat{S}_t$ is not adequate, can we still obtain a less biased estimator?

As we mentioned above, when $\hat{S}_t$ adequately captures the structure of S_t, we would expect $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ to be unbiased for all u, hence the plot of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ versus lag u would on average be a horizontal line. We propose a model checking method by plotting $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ vs lag u, and we call it the Lagged-Estimator-Plot, LEP.

First we need to fit a model $E(Y_t) = \beta_0 + \beta X_t + S_t$ to obtain $\hat{S}_t$, then $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ can be calculated using Equation 2.4, and the variance of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ is $\sigma^2(u) = 2\sigma^2 / \sum_{t=1}^{T-u}(X_t - X_{t+u})^2$. The LEP plot will give us a visual display of how $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s change with u, which reflects the possible impact of unmeasured confounders on the true association or mis-specification of the model for measured confounders.

In order to test the equivalence of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s, we need to a test statistics. Under the assumption $Y_t = \beta_0 + \beta X_t + S_t + \varepsilon_t = \mu_t + S_t + \varepsilon_t$, if $\hat{S}_t$ captures the structure of

S_t, we would expect $Y_t^* = Y_t - \hat{S}_t$ to be approximately independent normal variates with mean μ_t and variance σ^2. The vector $\tilde{\beta}_{\hat{\mathbf{S}}} = \mathbf{HY}^*$ follows a multivariate normal distribution with estimated covariance matrix $\Sigma_{\hat{\mathbf{S}}} = \hat{\sigma^2}_{\hat{\mathbf{S}}}\mathbf{HH^t}$.

Let $\beta(u)$ denote the limit of $\tilde{\beta}(u)$ when the total number of days T goes to ∞. First we want to test the null hypothesis $\beta_{\hat{\mathbf{S}}}(1) = \ldots = \beta_{\hat{\mathbf{S}}}(u) = \cdots = \beta_{\hat{\mathbf{S}}}$. Denote the vector $(\tilde{\beta}_{\hat{\mathbf{S}}}(1) - \overline{\tilde{\beta}}_{\hat{\mathbf{S}}}, \ldots, \tilde{\beta}_{\hat{\mathbf{S}}}(T-2) - \overline{\tilde{\beta}}_{\hat{\mathbf{S}}})^t$ as $\Delta\tilde{\beta}_{\hat{\mathbf{S}}} = \mathbf{M}\tilde{\beta}_{\hat{\mathbf{S}}} = \mathbf{MHY}^*$, where $\overline{\tilde{\beta}}_{\hat{\mathbf{S}}}$ is the mean of all the $T-1$ $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s.

One obvious test statistics is $D_1 = \Delta\tilde{\beta}^t_{\hat{\mathbf{S}}}[\hat{Cov}(\Delta\tilde{\beta}_{\hat{\mathbf{S}}})]^{-1}\Delta\tilde{\beta}_{\hat{\mathbf{S}}}$, where $\hat{Cov}(\Delta\tilde{\beta}_{\hat{\mathbf{S}}})$ is the estimated covariance matrix for $\Delta\tilde{\beta}_{\hat{\mathbf{S}}}$. However, the test statistics D_1 equals to the residual degrees of freedom for the model. We can rewrite the test statistics as $(\mathbf{MHY}^*)^\mathbf{t}[\mathbf{MHH^tM^t}]^{-\mathbf{1}}\mathbf{MHY}^*/\hat{\sigma^2}_{\hat{\mathbf{S}}}$, where $(\mathbf{MHY}^*)^\mathbf{t}[\mathbf{MHH^tM^t}]^{-\mathbf{1}}\mathbf{MHY}^*$ equals sum squared error of the model, while $\hat{\sigma^2}_{\hat{\mathbf{S}}}$ is the mean squared error. Hence the test statistics is not a random variable, but a constant, which is the residual degrees of freedom.

We can construct another test statistics $D_2 = \Delta\tilde{\beta}^t_{\hat{\mathbf{S}}}[\hat{Var}(\Delta\tilde{\beta}_{\hat{\mathbf{S}}})]^{-1}\Delta\tilde{\beta}_{\hat{\mathbf{S}}}$, where $\hat{Var}(\Delta\tilde{\beta}_{\hat{\mathbf{S}}})$ is the estimated diagonal variance matrix. Note that D_2 equals to $\sum_{u=1}^{T-1}[(\tilde{\beta}_{\hat{\mathbf{S}}}(u) - \hat{\beta}_{\hat{\mathbf{S}}})^2/\hat{\sigma}^2(u)]$, where $\hat{\sigma}^2(u)$ is the estimated variance of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$. Because the $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s are dependent, D_2 does not have a $\chi^2(T-2)$ distribution and we will use time series bootstrap method to obtain a p-value for the test statistics as well as the 95% interval under the null hypothesis.

The bootstrap scheme introduced by Efron (Efron, 1979) was first geared toward independent data. When we apply the i.i.d. bootstrap to dependent data, the estimates of variances are typically inconsistent (Singh, 1981; Babu and Singh, 1983). Because the i.i.d. bootstrap "scrambles" the data, all information about dependence

will be lost. In order to preserve the possible autocorrelation in the time series data, we use block resampling instead of the i.i.d. bootstrap (Künsch, 1989).

The bootstrap procedure we used is the following. After fitting the model, we have $Y_t = \hat{Y}_t + e_t$. The time series of residuals e_t, $t = 1, \ldots, T$ is divided into disjoint m day strata to preserve the possible autocorrelation within blocks. The strata are sampled with replacement and we acquire a new sequence of Y_t^N using $Y_t^N = \hat{Y}_t + e_t^N$ for each bootstrap replicate. Bootstrap percentile p-value is then obtained. For each bootstrap step, we can calculate the matching estimator $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$. We then obtain the 95% interval of the $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ under the null hypothesis. We use different values of m to check the sensitivity of the result to the size of the strata.

Another question we want to answer is how to obtain a less biased estimator of β when $\hat{S}_t$ does not adequately describe S_t. If S_t changes smoothly with time, $\tilde{\beta}(u)$ tend to have little bias for small lag u. We can use the weighted average of $\tilde{\beta}(u)$ for small u as our estimator (eg, $\tilde{\beta}(u)$ for $u \leq k$).

Since our goal is to estimate "the true effect" of an exposure on human health, we want to know the relative risk of adverse health outcome with or without the exposure for the same population, keeping all the other factors constant. Our scientific interest is actually the counterfactual parameter $\beta(0)$. Another proposed estimator is obtained by extrapolating to $\tilde{\beta}(0)$ by fitting a regression model of $\tilde{\beta}(u)$ as a smooth function of u. We used a natural spline with 3 degrees of freedom.

2.1.5 Simulation study

A simulated data set is used to illustrate the LEP model-checking method. First we used the outcome Y_t, the daily morality for persons 75 years and older and the exposure X_t, the previous day daily average temperature in Chicago from March 1 to October 31, 1996. Data for this application are available at the Internet-based Health and Air Pollution Surveillance System (iHAPSS) website (URL: http://www.ihapss.jhsph.edu/data/data.htm
We fit a linear regression model $Y_t = \beta_0 + X_t\beta + S(t, df = 3) + \varepsilon_t$, using natural spline of t with three degrees of freedom as the smoothing function. The estimated coefficient and smoothing function were used as the true β and S_t in the simulation study. We used the same set of observed X_t in the simulation and simulated Y_t using $Y_t \sim N(\hat{Y}_t, \sigma^2)$, where σ^2 was the estimated variance from the linear regression. The true β is 0.1367.

The following four models were fit to the simulated data, and we obtained estimators $\hat{\beta}_{\hat{\mathbf{S}}}$ and $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s for each of them.

Model 1: $E(Y_t) = \beta_0 + X_t\beta$. Here we ignore S_t.

Models 2, 3 and **4:** $E(Y_t) = \beta_0 + X_t\beta + S(t, \nu)$, where $S(t, \nu)$ is a natural spline of t with ν degrees of freedom, for $\nu = 1, 3, 10$, respectively.

Simulation Study I: model-checking

A standardized residual plot is commonly used in model checking. The standardized residuals should have mean 0 and constant variance when the model is valid. Figure 1 is the residual plots for each of the four models using one realization of the simulated data, and the solid lines are the smooth spline curve of the standardized residuals with 3 degrees of freedom. However, from the residual plot itself, it is very

difficult to detect lack of fit and especially difficult to appreciate how any lack of fit affects the estimate of scientific interest β. The smooth spline lines suggest that Models 1 (ignore S_t) and 2 ($\nu = 1$) have a little bit of U shape in the mean of the standardized residuals, while Models 3 ($\nu = 3$) and 4 ($\nu = 10$) have standardized residuals with roughly mean 0 and constant variance.

Figure 2 is our proposed model checking method, the LEP plots using the same realization of the simulated data. The horizontal lines are the true β, and the dotted lines are the bootstrap 95% intervals under the null hypothesis using 5000 bootstrap iterations. We used block sizes 5, 7, and 10 days to investigate the sensitivity of the bootstrap on the block size, and the results turned out to be very similar. For models using less than enough degrees of freedom in estimating S_t (Models 1 and 2), we can clearly see the U shape curve of $\tilde{\beta}(u)$. The LEP plots give us more obvious evidence that Models 1 and 2 did not adequately describe the data in so far as our goal to estimate β. Had they done so, the $\tilde{\beta}(u)$ would be roughly independent of u. Moreover, the LEP plots give us a direct view of the magnitude of the estimators for each lag u. If we test the null hypotheses that $\beta(u) = \beta$ for all u, we reject with $p = 0.00$.

The LEP plots for both models 3 and 4 turn out to be roughly horizontal (p-value=0.47 and 0.16, respectively). Model 3 uses the correct degrees of freedom in estimating S_t, while Model 4 uses too many degrees of freedom. Both models should generate estimators with less bias. However, Model 4 will be more variable.

Simulation Study II: robust estimators

Using simulated data, we fit Models 1 through 4 as described in the model-checking session, and calculate $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ for each lag u. For each model, we obtain the following

estimators:

Method A: The estimator of β obtained using the model.

Method B: (a)Weighted average of the first 20 $\tilde{\beta}(u)$s. Weight $w(u) = \sum_{t=1}^{T-u}(X_t - X_{t+u})^2 / \sum_{u=1}^{T-1}\sum_{t=1}^{T-u}(X_t - X_{t+u})^2$.

(b)Weighted average of all the $\tilde{\beta}(u)$s. This estimator should be the same as in Method A.

Method C: (a)Regress the first 20 $\tilde{\beta}(u)$s on natural spline of u with 3 degrees of freedom, weighted by $w(u)$. The intercept $\tilde{\beta}(0)$ is our estimator.

(b) Similar to C(a), but using all the $\tilde{\beta}(u)$s. Table 5.2 is the mean and standard deviation (in parenthesis) of the estimates using 1000 simulations for the models and methods mentioned above.

The results confirm that Method A and Method B(b) generate the same estimator as shown in Equation 2.3, showing that $\hat{\beta}_{\hat{\mathbf{S}}}$ can be written as the weighted average of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s for linear models. When we use less than enough degrees of freedom for S_t (Models 1 and 2), extrapolation to lag 0 (Method C) gives a less biased estimator than the weighted average (Method B). Using the first 20 $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s produces less bias results than using all the pairs. The variance using all pairs is smaller reflecting a bias-variance trade-off.

When we used enough degrees of freedom for S_t (Models 3 and 4), we will have similar estimates for all the 5 methods, which suggests that when $\hat{S}_t$ captures the structure of S_t, the $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ will be roughly constant across u. Method C(a) is the least biased estimator for all models, but it has the biggest variance.

2.2 Log-Linear Time Series Model

For environmental time series data, we often have mortality or morbidity counts as the outcome. One commonly used approach is to fit a Poisson log-linear model with linear terms of PM air pollution levels and smooth functions of time to adjust for the time-varying confounders.

2.2.1 Model

We assume the true model is $\log \mu_t = \beta_0 + \beta X_t + S_t$, where Y_t is the number of events with $E(Y_t) = \mu_t$ and $Var(Y_t) = \phi\mu_t$, X_t is the exposure such as air pollution, and S_t is the value of a smooth function of time which represents the combined effect of unmeasured confounders. Time series log-linear models allow for over-dispersion relative to the Poisson variance, where ϕ is the over-dispersion parameter.

Suppose we ignore S_t and fit **Model I:** $\log(\mu_t) = \beta_0 + \beta X_t$ to obtain $\hat{\beta}$. For each data pair (X_i, Y_i) and (X_j, Y_j), we can obtain the estimator by solving estimating equation $X_i[Y_i - \exp(\beta_0 + \beta X_i)] + X_j[Y_j - \exp(\beta_0 + \beta X_j)] = 0$, however, there is no closed form solution. Similar to the linear time series linear model, for each data pair (X_i, Y_i) and (X_j, Y_j), we can crudely estimate β using $\hat{\beta}_{i,j} = (\log Y_i - \log Y_j)/(X_i - X_j)$. The estimator $\hat{\beta}$ obtained using Model I is roughly the weighted average of $\hat{\beta}_{i,j}$, and the weight is reciprocal to the variance of $\hat{\beta}_{i,j}$. The variance of $\hat{\beta}_{i,j}$ is $(1/\hat{\mu}_i + 1/\hat{\mu}_j)/(X_i - X_j)^2$, which is a function of X as well as μ. Note that for linear model, the least squares estimator is exactly the weighted average of $\hat{\beta}_{i,j}$, and the weight is proportional to $(X_i - X_j)^2$.

Using the results given by Lu et al. (2007), the expectation of the estimating

equation for the model can be written as

$$\begin{aligned} E[U(\beta)] &= \sum_{t=1}^{T} X_t e^{\beta X_t} \left[\exp(S_t) - E[\exp(\hat{S}_t(\beta))]\right] \\ &= \sum_{t=1}^{T} X_t e^{\beta X_t} \Delta e^{S_t}(\beta), \end{aligned} \tag{2.9}$$

where $\Delta e^{S_t}(\beta)$ is the difference between true $\exp(S_t)$ and the expectation of the estimated $\exp(\hat{S}_t(\beta))$.

We can combine pairs of observations with lag u to obtain a matching estimator

$$\begin{aligned} \tilde{\beta}(u) &= \sum_{t=1}^{T-u} \left(\frac{\log Y_i - \log Y_j}{X_t - X_{t+u}}\right) \left(\frac{(X_t - X_{t+u})^2/(1/\hat{\mu}_t + 1/\hat{\mu}_{t+u})}{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2/(1/\hat{\mu}_t + 1/\hat{\mu}_{t+u})}\right) \\ &= \sum_{t=1}^{T-u} \hat{\beta}_{t,t+u} w_t(u), \end{aligned} \tag{2.10}$$

where $w_t(u) = (X_t - X_{t+u})^2/(1/\hat{\mu}_t + 1/\hat{\mu}_{t+u})/\sum_{t=1}^{T-u}(X_t - X_{t+u})^2/(1/\hat{\mu}_t + 1/\hat{\mu}_{t+u})$. Here, $\tilde{\beta}(u)$ is approximately the estimator obtained using symmetric bidirectional case-crossover design using days $t-u$ and $t+u$ as the control days for event day u.

We can also fit **Model II:** $\log(\mu_t) = \beta_0 + \beta X_t + S_t$ to obtain $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$, and use the proposed model-checking method in the linear model section as well as obtain the proposed estimators.

2.2.2 Simulation study

Another simulation study was conducted. We used the same data set as in the linear model section and we fit a log-linear regression model $\log(\mu_t) = \beta_0 + \beta X_t + S(t, df = 3)$, using natural spline of t with three degrees of freedom as the smoothing function. The

estimated coefficient and smoothing function were used as the true β and S_t in the simulation study. We used the same set of observed X_t and simulated Y_t using $Y_t \sim$ Poisson(μ_t). The true β is 2.379, which is approximately the percentage increase of the mortality rate for every 10 degrees of increase in previous day temperature. We fit 4 models as before:

Model 1: $\log(\mu_t) = \beta_0 + \beta X_t$. Here we ignore S_t.

Models 2, 3 and **4** are $\log(\mu_t) = \beta_0 + \beta X_t + S(t, \nu)$, where $S(t, \nu)$ is a natural spline of t with ν degrees of freedom, for $\nu = 1, 3, 10$, respectively.

We again used one realization of the simulated data to perform model-checking. The standardized residual plots (Figure 3) and the LEP plots (Figure 4) show similar pattern as for linear time series model. It is very difficult to detect lack of fit from the standardized residual plot itself, even though the smooth spline lines suggest that models 1 (ignore S_t) and 2 ($\nu = 1$) have a little bit of U shape. Our proposed LEP plots clearly indicate the U shape curve of $\tilde{\beta}(u)$ for models 1 and 2, which suggest that those two models did not adequately describe the data. The dotted lines in the LEP plots are the bootstrap 95% intervals under the null hypothesis, $\beta(u) = \beta$ for all u, using block re-sampling bootstrap. We reject the null hypothesis with $p = 0.00$ for models 1 and 2. The LEP lots for both models 3 ($\nu = 3$) and 4 $\nu = 10$ turn out to be roughly horizontal (p=0.43 and 0.08, respectively). Using block sizes 5, 7, and 10 generates very similar result.

We used 1000 simulations to calculate mean and standard deviation of the proposed estimators (Table 5.2). The simulation results have shown that Method A and Method B(b) generate very similar results, which confirmed that $\hat{\beta}_{\hat{\mathbf{S}}}$ can be approximated using the weighted average of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s for log-linear models. Because there

is no closed-form solution for the log-linear model estimating equation, we can not compute the bias and variance directly. This approximation allows us to compute the estimator and variance without iterations. When we used less than enough degrees of freedom in S_t, extrapolation to lag 0 (Method C) gives less biased results comparing with weighted average (Method B), and using the first 20 $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s gives less biased results than using all the pairs, but again we have bias-variance trade-off. When we used enough degrees of freedom in S_t (Models 3 and 4), we will have similar estimates for all the 5 methods, which suggests that when $\hat{S}_t$ captures the structure of S_t, the $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$ will be roughly constant across u. Method C(a) is the most robust estimator, even though it has the biggest variance.

The simulation results have shown that matching estimators can be used to perform model-checking, compute bias and variance, and construct robust estimators for log-linear models.

2.3 Extension to Spatial Log-linear Model

In the previous sections, we considered time series models. In environmental epidemiology, we often have spatial data, which is more complicated because space is two-dimensional. We often have mortality or morbidity counts as the outcome, so need to fit log-linear model with linear terms of PM air pollution levels and smooth functions of space to adjust for the space-varying confounders. Usually we need to adjust for measured confounders such as SES and smoking due to differences between populations at different locations. The city studies can also be confounded by other unmeasured differences between city populations being compared, which makes spatial analyses more challenging.

Let's use the Medicare Cohort Air Pollution Study (MCAPS) data as a motivating example. We have average PM2.5 concentration and non-accidental mortality for 1055 zipcodes across the United states in the Medicare system. We divided the US into three geographical regions (Western Coast, Central US, and Eastern US). Suppose we are interested in the association between the numbers of deaths Y_s and average PM2.5 level X_s for each region. We also take into account of number of people at risk N_s, and control for measured confounders Z_s, SES status (proportion with high-school education, proportion with degree, proportion living in poverty, proportion unemployed, median income) and the chronic obstructive pulmonary disease (COPD) standardized mortality ratio as a surrogate for smoking. We assume a spatial log-linear model $\log(\mu_s) = \beta_0 + \beta X_s + \gamma Z_s + U_s + \log(N_s)$ for location s, where $E(Y_s) = \mu_s$, $Var(Y_s) = \phi\mu_s$, and U_s is the value of a smooth function of space at location s. We allow for over-dispersion relative to the Poisson variance, where ϕ is the over-dispersion parameter.

Suppose we ignore U_s and fit **Model I:** $\log(\mu_s) = \beta_0 + \beta X_s + \gamma Z_s$ to obtain $\hat{\beta}$ and $\hat{\gamma}$. Denote the $\hat{\beta}$ from Model I as **Estimator A** . Similar to time series log-linear model we described above, for each data pair (X_i, Y_i) and (X_j, Y_j), we can estimate β using $\hat{\beta}_{i,j} = [(\log(Y_i/N_i) - \hat{\gamma}Z_i) - (\log(Y_j/N_j) - \hat{\gamma}Z_j)]/(X_i - X_j)$, where $\hat{\gamma}$ is obtained from Model I mentioned above. Model I assumes U_s is constant across location s, which implies that $\hat{\beta}_{i,j}$ will be roughly constant for pairs at different distance d. Since there are so many $\hat{\beta}_{i,j}$s available, we need to construct new measures of $\beta(d)$ in order to perform model checking.

First, we fit Model I separately for each geographical region, then obtain $\hat{\beta}_{i,j}$ and $d_{i,j}$ for each data pair (X_i, Y_i) and (X_j, Y_j) within that region. We divide data pairs into 5 different categories based on their geographical distance $d_{i,j}$: $d \leq 10$ miles,

$10 < d \leq 100$, $100 < d \leq 250$, $250 < d \leq 500$, and $d > 500$ miles. We assume that U is roughly constant within each distance category. We then regress $\hat{\beta}_{i,j}$ on the dummy variables for each distance category, weighted by the reciprocal of the variance of $\hat{\beta}_{i,j}$. We will obtain an estimator $\tilde{\beta}(d)$ for each distance category. $\tilde{\beta}(d)$ can be considered as the weighted average of $\hat{\beta}_{i,j}$ within each distance category. If the assumption is valid, $\tilde{\beta}(d)$ will be roughly constant across distance category.

Due to the dependence of pairwise coefficients, we use block re-sampling bootstrap to calculate the standard errors of matching coefficient $\tilde{\beta}(d)$. First we fit model I to the data and obtain pairwise coefficients. We divide each geographical region into blocks based on the latitude and longitude of the sampling sites. We then re-sample the blocks with replacement and use the previously obtained pairwise coefficients within the blocks to calculate new $\tilde{\beta}(d)$s for each bootstrap replicate. The standard deviation of the 1000 bootstrap $\tilde{\beta}(d)$s is used as the standard error of matching coefficient $\tilde{\beta}(d)$.

Figures 5, 6 and 7 are the plots of $\tilde{\beta}(d)$ versus lag category with their 95% confidence intervals for three age groups and three geographical regions. Even though the 95% confidence intervals overlap each other, we can see that there exist trend of $\tilde{\beta}(d)$, especially for people 85 years and older, and for central US. Changing bootstrap block sizes generates similar results.

We define **Estimators B and C** to be the weighted average of $\hat{\beta}_{i,j}$ for all the pairs and pairs with distance $\leq$ 500 miles, respectively. The weight again is the reciprocal of the variance of $\hat{\beta}_{i,j}$. The **Estimator D** is obtained by regressing $\hat{\beta}_{i,j}$ on $log(d_{i,j})$ using i and j pairs with distance $\leq$ 500 miles, weighted by the reciprocal of the variance of $\hat{\beta}_{i,j}$.

The estimators obtained by extrapolation is robust in the simulation study for

time series model, however, it generates negative estimates for people in central US and for people 85 years and older even after we fit two-dimensional smoothing function of longitude and latitude with 3 degrees of freedom (results not shown). It suggests that the most commonly used spatial log-linear model doesn't adequately accounted for the unmeasured space-varying confounders. Due to the complexity of spatial data, a new measure of distance other than geographical distance d may need to be constructed to take into account the effect of unmeasured confounders which do not vary smoothly in space.

3 DISCUSSIONS

Confounding bias is an important issue in environmental epidemiology. The true association between health outcome and air pollution can be confounded by measured or unmeasured time-varying confounders. Matching and modeling have been used to control for these effects.

Time series analyses of air pollution data controlled for confounding bias by including smooth functions of time in the time series semi-parametric regression model (McCullagh and Nelder, 1989; Hastie and Tibshirani, 1990; Marx and Eilers, 1998; Dominici et al., 2004). The number of degrees of freedom in the smooth functions of time reflects the degree of adjustment for confounding factors and it can have a large impact on the magnitude and statistical uncertainty of the estimation of the true association.

In this paper, we introduced matching estimators $\tilde{\beta}(u)$, which can be considered as matching day t with days $t+u$ and $t-u$. The matching estimator $\tilde{\beta}(u)$ is a

weighted average of pairwise regression estimators $\hat{\beta}_{t,t+u}$. The pairwise regression estimators has been called "elemental regressions" for simple linear regression. The idea of elemental regressions have existed for centuries. Suppose we have k parameters in the model, only $p = k + 1$ observations are required to estimate the estimators $\beta_0, \beta_1, \ldots, \beta_k$ (Mayo and Gray, 1997). Back in 1981, Jacobi showed that the least squares estimator can be written as a weighted average of the elemental regressions in the linear model space(Sheynin, 1973). There are major differences between elemental regressions and our pairwise regressions. Assume the true model is $E(Y_t) = \beta_0 + \beta X_t + \gamma Z_t$. Elemental regressions would need 3 observations to estimate the estimators β_0, β and γ. For our pairwise regressions, we first use all the data to estimate $\hat{\beta}_0$ and $\hat{\gamma}$, then use $Y_t^* = Y_t - \hat{\gamma} Z_t$ as a new response variable to explore the association between Y_t^* and X_t. No matter how many covariates are included in the model, we can always use the data pair as the smallest element to perform pairwise regressions. For semi-parametric models such as $E(Y_t) = \beta_0 + \beta X_t + \gamma Z_t + S_t$, it may not be feasible to perform the traditional elemental regression.

Our pairwise regression estimator $\hat{\beta}_{i,j} = [(Y_i - \hat{\gamma} Z_i - \hat{S}_i) - (Y_j - \hat{\gamma} Z_j - \hat{S}_j)]/(X_i - X_j)$ would be approximately free of confounding effect if the $\hat{\mathbf{S}}$ is a close approximation to $\mathbf{S}$. If S_t changes smoothly with time, we can combine our pairwise regressions with the same lag u to construct a matching estimator $\tilde{\beta}(u)$.

The LEP (Lagged-Estimator-Plot) is the graph of $\tilde{\beta}(u)$ versus lag u. When the model for S_t adequately captures the influence of unmeasured confounders, the LEP will be a roughly horizontal line. The LEP plot shows the effect on the regression coefficient of interest, β, if not adequately describing S_t. Moreover, the LEP plots give us a direct view of the magnitude of the estimators for each lag u.

The least squares estimator $\hat{\beta}_{\mathbf{S}}$ can be written as the weighted average of $\tilde{\beta}_{\mathbf{S}}(u)$. Note that $\hat{\beta}_{\mathbf{S}}$ can be written as a linear combination of $\tilde{\beta}(u)$, where the weight depends on variogram for lag u as well as the how we smooth $\mathbf{Y}$ to obtain $\mathbf{S}$.

Based on $\tilde{\beta}_{\mathbf{S}}(u)$, we can construct new estimators using weighted average of $\tilde{\beta}_{\mathbf{S}}(u)$, or by extrapolation to lag 0. Simulation results suggest that estimators generated using extrapolation tend to have less bias, but the variance is bigger. The bias of $\tilde{\beta}_{\mathbf{S}}(u)$ is small for small enough u. When we calculated weighted average of $\tilde{\beta}_{\mathbf{S}}(u)$ for the first 20 u comparing with using all of the 244 us, the bias is much improved when the model uses less than enough degrees of freedom in smoothing, and the bias is bigger. When the model uses enough degrees freedom in smoothing, using the first 20 or all 244 $\tilde{\beta}_{\mathbf{S}}(u)$ would generate similar result. The result is somewhat surprising because we would expect the weighted average of the first 20 $\tilde{\beta}_{\mathbf{S}}(u)$ to have bigger variance since it throws away lots of data pairs.

Sentürk and Müller proposed covariate-adjusted regression (CAR) for situations where both predictors and outcome are contaminated by a multiplicative factor which is determined by a unknown function of a measured variable M (Sentürk and Müller, 2005). There is a varying coefficient model associated with CAR: $Y = \beta_0(M) + \beta(M)X + \psi(M)\varepsilon$, where $\psi(M)$ is the unknown multiplicative factor for the outcome. For time series data, time t is the measured variable. The $\beta(M)$ obtained in CAR is similar to our matching estimator when ignoring S_t, because both use the idea of matching. CAR divide data into disjoint strata by their M values, while our matching estimator uses running blocks. CAR uses weighted average of $\beta(M)$ to obtain an estimator for the true β. For our proposed estimators, we combine the idea of matching and modeling by including a smooth function of time in the model.

The results can be extended to time-series log-linear models. Simulations confirmed that $\hat{\beta}_{\hat{\mathbf{S}}}$ can be approximated using the weighted average of $\tilde{\beta}_{\hat{\mathbf{S}}}(u)$s for log-linear models. This approximation allows us to compute the estimator and variance without iterations. The matching estimators can be used to perform model-checking, compute bias and variance, and construct robust estimators for log-linear models. The mortality counts in our simulation study are relatively large (around 60), hence we can use $\log(Y_t/N_t)$ to construct matching coefficients. It would be interesting to extend our results to log-linear model with small counts for which $\log(Y_t/N_t)$ is not possible and to binary responses as well as other models in the generalized linear model families.

We applied our model-checking method to spatial data, which revealed that the most commonly used spatial log-linear model doesn't adequately accounted for the unmeasured space-varying confounders. We usually include smoothing function of location in the model, however, for spatial data, the confounders may not vary smoothly in space. For example, big cities and their surrounding suburbs can have big culture difference even though they are very close in distance, while big cities may be more similar in culture even though they are further apart. We want to extend our results to cases when β_{ij} depends on more than geographical distance d. Due to the complexity of spatial data, a new measure of distance other than d may need to be constructed to take into account the effect of unmeasured confounders which do not vary smoothly in space.

4 References

Aekplakorn, W.; Loomis, D.; Vichit-Vadakan, N.; Shy, C.; Wongtim, S.; and Vitayanon, P. Acute effect of sulphur dioxide from a power plant on pulmonary function of children, Thailand. *Int J Epidemiol*, **2003**, *32*, 854-861.

Babu, G. J. and Singh, K. Inference on means using the bootstrap. *Ann Statist*, **1983**, *11*, 999-1003.

Bateson, T. F. and Schwartz, J. Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures. *Epidemiology*, **1999**, *10*, 539-544.

Chhabra, S. K.; Chhabra, P.; Rajpal, S.; and Gupta, R. K. Ambient air pollution and chronic respiratory morbidity in Delhi. *Arch Environ Health*, **2001**, *56*, 58-64.

Dominici, F., Mcdermott, A.; and Hastie, T.J. Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association*, **2004**, *99*, 938-948.

Dominici, F.; Peng, R. D.; Bell, M. L.; Pham, L.; McDermott, A.; Zeger, S. L.; and Samet, J. M. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, **2006**, *295*, 1127-1134.

Dominici, F.; Peng, R. D.; Zeger, S. L.; White, R. H.; and Samet, J. M. Particulate air pollution and mortality in the United States: did the risks change from 1987 to 2000? *Am J Epidemiol*, **2007**, *166*, 880-888.

Efron, B. Bootstrap methods: Another look at the jackknife. *Ann Statist*, **1979**, *7*,

1-26.

Hastie, T. and Tibshirani, R. Generalized additive models. **1990**. London and New York: Chapman and Hall.

Hastie, T. and Tibshirani, R. Generalized additive models for medical research. *Stat Methods Med Res*, **1995**, *4*, 187-196.

Jerrett, M.; Burnett, R. T.; Ma, R.; Pope, C. A.; Krewski, D.; Newbold, K. B.; Thurston, G.; Shi, Y.; Finkelstein, N.; Calle, E. E.; and Thun, M. J. Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, **2005**, *16*, 727-736.

Jerrett, M.; Buzzelli, M.; Burnett, R. T.; and DeLuca, P. F. Particulate air pollution, social confounders, and mortality in small areas of an industrial city. *Soc Sci Med*, **2005**, *60*, 2845-2863.

Katsouyanni, K.; Schwartz, J.; Spix, C.; Touloumi, G.; Zmirou, D.; Zanobetti, A.; Wojtyniak, B.; Vonk, J. M.; Tobias, A.; Pönkä, A.; Medina, S.; Bachárová, L.; and Anderson, H. R. Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *J Epidemiol Community Health*, **1996**, *50* Suppl 1, S12-S18.

Künsch, H. R. The jacknife and the bootstrap for general stationary observations. *Anals of Statistics*, **1989**, *17*, 1217-1241.

Lu, Y.; Symons, J. M.; Geyh, A. S.; and Zeger, S. L. An approach to checking and improving upon case-crossover analyses based on equivalence with time series methods. *Epidemiology*. (to appear)

Lu, Y. and Zeger, S. L. On the equivalence of case-crossover and time series

methods in environmental epidemiology. *Biostatistics*, **2007**, *8*, 337-344.

Maclure, M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*, **1991**, *133*, 144-153.

Marx, B. D. and Eilers, P. H. C. Direct generalized additive modeling with penalized likelihood. **1998**, *Comp Stat Data Ana*, *28*, 193-209.

Mayo, M. S. and Gray, J. B. Elemental subsets: the building blocks of regression. *Am Statist*, **1997**, *51*, 122-129.

McCullagh, P. and Nelder, J.A. *Generalized Linear Models*, 2nd Edition. **1989**, London:Chapman & Hall/CRC.

Ostro, B.; Broadwin, R.; Green, S.; Feng, W.; and Lipsett, M. Fine particulate air pollution and mortality in nine California counties: results from CALFINE. *Environ Health Perspect*, **2006**, *114*, 29-33.

Ostro, B.; Chestnut, L.; Vichit-Vadakan, N.; and Laixuthai, A. The impact of particulate matter on daily mortality in Bangkok, Thailand. *J Air Waste Manag Assoc*, **1999**, *49*, 100-107.

Ostro, B. D.; Eskeland, G. S.; Sanchez, J. M.; and Feyzioglu, T. Air pollution and health effects: A study of medical visits among children in Santiago, Chile. *Environ Health Perspect*, **1999**, *107*, 69-73.

Peng, R. D.; Dominici, F.; Pastor-Barriuso, R.; Zeger, S. L.; and Samet, J. M. Seasonal analyses of air pollution and mortality in 100 US cities. *Am J Epidemiol*, **2005**, *161*, 585-594.

Samet, J. M.; Dominici, F.; Curriero, F. C.; Coursac, I.; and Zeger, S. L. Fine

particulate air pollution and mortality in 20 U.S. cities, 1987-1994. *N Engl J Med*, **2000**, *343*, 1742-1749.

Samet, J. M.; Dominici, F.; Zeger, S. L.; Schwartz, J.; and Dockery, D. W. The National Morbidity, Mortality, and Air Pollution Study. Part I: Methods and methodologic issues. *Res Rep Health Eff Inst*, **2000**, *5-14*; discussion 75-84.

Schwartz, J.; Dockery, D. W.; and Neas, L. M. Is daily mortality associated specifically with fine particles? *J Air Waste Manag Assoc*, **1996**, *46*, 927-939.

Scheynin, O. B. R.J. Boscovich's work on probability. *Arch Hist Exact Sci*, **1973**, *9*, 306-324.

Sentürk, D. and Müller, H.-G. Covariate-adjusted regression. *Biometrika*, **2005**, *92*, 75-89.

Singh, K. On the asymptotic accuracy of Efron's bootstrap. *Ann Statist*, **1981**, *9*, 1187-1195.

Vichit-Vadakan, N.; Ostro, B. D.; Chestnut, L. G.; Mills, D. M.; Aekplakorn, W.; Wangwongwatana, S.; and Panich, N. Air pollution and respiratory symptoms: results from three panel studies in Bangkok, Thailand. *Environ Health Perspect*, **2001**, *109* Suppl 3, 381-387.

5 APPENDIX

5.1 Appendix I

Let

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & & & & \dots & 0 \\ 0 & 1 & -1 & 0 & & & \dots & 0 \\ & & & \dots & & & & \\ 0 & & \dots & & & 0 & 1 & -1 \\ -- & -- & -- & -- & -- & -- & -- & -- \\ 1 & 0 & -1 & 0 & & & \dots & 0 \\ & & & \dots & & & & \\ 0 & & & \dots & 0 & 1 & 0 & -1 \\ -- & -- & -- & -- & -- & -- & -- & -- \\ & & & \dots & & & & \\ -- & -- & -- & -- & -- & -- & -- & -- \\ 1 & 0 & & \dots & & & 0 & -1 \end{pmatrix} \begin{pmatrix} T \\ 2 \end{pmatrix}_{\times T}$$

$$= \begin{pmatrix} \mathbf{D_1} \\ - - - \\ \mathbf{D_2} \\ - - - \\ \dots \\ - - - \\ \mathbf{D_{T-1}} \end{pmatrix} \begin{pmatrix} T \\ 2 \end{pmatrix}_{\times T} .$$

We have

$$\begin{aligned}\tilde{\beta}(u) &= \frac{\sum_{t=1}^{T-u}(Y_t - Y_{t+u})(X_t - X_{t+u})}{\sum_{t=1}^{T-u}(X_t - X_{t+u})^2} \\ &= \frac{(\mathbf{D_u X})^t \mathbf{D_u Y}}{(\mathbf{D_u X})^t \mathbf{D_u X}} = \frac{\mathbf{X}^t {\mathbf{D_u}}^t \mathbf{D_u Y}}{\mathbf{X}^t {\mathbf{D_u}}^t \mathbf{D_u X}} = \mathbf{H_u Y}, \end{aligned} \tag{5.1}$$

where $\mathbf{H_u} = \mathbf{X}^t {\mathbf{D_u}}^t \mathbf{D_u} / \mathbf{X}^t {\mathbf{D_u}}^t \mathbf{D_u X}$.

We have

$$\begin{aligned}\tilde{\beta} &= \begin{pmatrix} \tilde{\beta}(1) \\ \tilde{\beta}(2) \\ \cdots \\ \tilde{\beta}(T-1) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{H_1} \\ --- \\ \mathbf{H_2} \\ --- \\ \cdots \\ --- \\ \mathbf{H_{T-1}} \end{pmatrix} \mathbf{Y} \\ &= \mathbf{HY}. \end{aligned} \tag{5.2}$$

The covariance for $\tilde{\beta}$ is $\Sigma = \sigma^2 \mathbf{H}\mathbf{H}^t$.

5.2 Appendix II

We can estimate S_t with $\hat{S}_t$ and rewrite the model as $Y_t = \hat{\beta}_0 + \hat{\beta} X_t + \hat{S}_t + \varepsilon_t$.

Then we can get

$$\begin{aligned} \hat{\beta}_{0,\hat{\mathbf{S}}} &= \frac{1}{T}\sum_{t=1}^{T} Y_t - \frac{1}{T}\sum_{t=1}^{T} \hat{S}_t \\ &= \bar{Y} - \bar{S}, \end{aligned}$$

and

$$\hat{\beta}_{\hat{\mathbf{S}}} = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \hat{S}_t - \bar{Y})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}.$$

We can write the model as

$$Y_t = \bar{Y} - \bar{S} + \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \hat{S}_t - \bar{Y})}{\sum_{t=1}^{T}(X_t - \bar{X})^2} + \hat{S}_t + \varepsilon_t,$$

Let's suppose $\hat{S}_t = \sum_{u=-t+1}^{T-t} \lambda_u Y_{t+u}$ with $\sum_{u=-t+1}^{T-t} \lambda_u = 1$ where λ_u is symmetric i.e. $\lambda_u = \lambda_{-u}$.

Then we can get

$$
\begin{aligned}
\hat{\beta}_{\hat{\mathbf{S}}} &= \frac{\sum_{t=1}^{T}(X_t-\bar{X})(Y_t-\hat{S}_t-\bar{Y})}{\sum_{t=1}^{T}(X_t-\bar{X})^2} \\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\sum_{t=1}^{T}(X_t-\bar{X})(Y_t-\hat{S}_t-\bar{Y}) \\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left[\sum_{t=1}^{T}(X_t-\bar{X})(Y_t-\hat{S}_t)-\sum_{t=1}^{T}(X_t-\bar{X})\bar{Y}\right] \\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\sum_{t=1}^{T}(X_t-\bar{X})(Y_t-\hat{S}_t) \\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left[\sum_{t=1}^{T}(X_t-\bar{X})Y_t-\sum_{t=1}^{T}(X_t-\bar{X})\hat{S}_t\right] \\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left[\sum_{t=1}^{T}(X_t-\bar{X})Y_t-\sum_{t=1}^{T}\left\{(X_t-\bar{X})\sum_{u=-t+1}^{T-t}\lambda_u Y_{t+u}\right\}\right] \\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left[\sum_{t=1}^{T}(X_t-\bar{X})Y_t-\sum_{t=1}^{T}\sum_{u=-t+1}^{T-t}(X_t-\bar{X})\lambda_u Y_{t+u}\right] \\
(\text{let } l=t+u) &= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left[\sum_{t=1}^{T}(X_t-\bar{X})Y_t-\sum_{t=1}^{T}\sum_{l=1}^{T}(X_t-\bar{X})\lambda_{l-t}Y_l\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left[\sum_{t=1}^{T}(X_t-\bar{X})Y_t-\sum_{l=1}^{T}\left\{\sum_{t=1}^{T}(X_t-\bar{X})\lambda_{l-t}\right\}Y_l\right]\\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left[\sum_{t=1}^{T}(X_t-\bar{X})Y_t-\sum_{t=1}^{T}\left[\sum_{l=1}^{T}(X_l-\bar{X})\lambda_{-l+t}\right]Y_t\right]\\
(\text{by } \lambda_{-l+t}=\lambda_{l-t}) &= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left\{\sum_{t=1}^{T}Y_t\left[(X_t-\bar{X})-\sum_{l=1}^{T}(X_l-\bar{X})\lambda_{l-t}\right]\right\}\\
(\text{by } l=t+u) &= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left\{\sum_{t=1}^{T}Y_t\left[(X_t-\bar{X})-\sum_{u=-t+1}^{T-t}(X_{t+u}-\bar{X})\lambda_u\right]\right\}\\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left\{\sum_{t=1}^{T}Y_t\left[X_t-\sum_{u=-t+1}^{T-t}\lambda_u X_{t+u}\right]\right\}\\
&= \frac{1}{\sum_{t=1}^{T}(X_t-\bar{X})^2}\left\{\sum_{t=1}^{T}Y_t\left[X_t-\tilde{X}_t\right]\right\}.
\end{aligned}
$$

where $\tilde{X}_t=\sum_{u=-t+1}^{T-t}\lambda_u X_{t+u}$ is the symmetric weighted running mean smoother of X_t using the same smoothing method as estimating $\hat{S}_t=\tilde{Y}_t$. This result holds even when there exists edge effect, as far as we use the same smoothing method for both $\tilde{X}_t$ and $\tilde{Y}_t$.

We extend the data beyond [1, T] by $X_{t+u}=X_{t+u-T}$ for $t+u>T$ and $X_{t-u}=X_{t-u+T}$ for $t-u<1$ to avoid edge effect. The data were analyzed using circular pattern, hence $\lambda_u=\lambda_{-u}=\lambda_{T-u}=\lambda_{u-T}$. The matching estimator $\tilde{\beta}(u)$ can be

written as

$$\begin{aligned}
\tilde{\beta}(u) &= \frac{\sum_{t=1}^{T}(Y_t - Y_{t+u})(X_t - X_{t+u})}{\sum_{t=1}^{T}(X_t - X_{t+u})^2} \\
&= \frac{\sum_{t=1}^{T} Y_t \left[2X_t - (X_{t-u} + X_{t+u})\right]}{\sum_{t=1}^{T} X_t \left[2X_t - (X_{t-u} + X_{t+u})\right]} \\
&= \frac{\sum_{t=1}^{T} Y_t \left[X_t - (X_{t-u} + X_t + X_{t+u})/3\right]}{\sum_{t=1}^{T} X_t \left[X_t - (X_{t-u} + X_t + X_{t+u})/3\right]} \\
&= \frac{1}{\sum_{t=1}^{T} X_t \left[X_t - \tilde{X}_t(u)\right]} \sum_{t=1}^{T} Y_t \left[X_t - \tilde{X}_t(u)\right], \qquad (5.3)
\end{aligned}$$

where $\tilde{X}_t(u) = (X_{t-u}+X_t+X_{t+u})/3$. It can be shown that $[\sum_{t=1}^{T} X_t(X_t-\tilde{X}_t(u))]/[\sum_{t=1}^{T} X_t(X_t-\bar{X}_t)]\tilde{\beta}(u)$ is exact the same as $\hat{\beta}_{\hat{\mathbf{S}}}$ using $\hat{S}_t = \tilde{Y}_t(u) = (Y_{t-u} + Y_t + Y_{t+u})/3$, here $\hat{S}_t$ is the running mean of Y_{t-u}, Y_t, and Y_{t+u}. When the edge effect exists, the result still holds, but $\hat{S}_t = (2Y_t + Y_{t+u})/3$ for $t < u + 1$ and $\hat{S}_t = (2Y_t + Y_{t-u})/3$ for $t > T - u$.

For the least square estimator $\hat{\beta}$, we have

$$\hat{\beta} = \frac{1}{\sum_{t=1}^{T}(X_t - \bar{X})^2} \left\{ \sum_{t=1}^{T} Y_t(X_t - \bar{X}) \right\},$$

hence $\hat{\beta}$ corresponds to using $\hat{S}_t = \bar{Y}$.

Using linear combinations of $\tilde{\beta}(u)$ for different u, we will be able to construct the estimator corresponding to different type of $\hat{S}_t$.

Denote $V(u) = \sum_{t=1}^{T}(X_t - X_{t+u})^2$ and $V_T = \sum_{u=1}^{T-1}\sum_{t=1}^{T}(X_t - X_{t+u})^2 = 2T\sum_{t=1}^{T}(X_t - \bar{X})^2$. For $\sum_{u=-t+1}^{T-t} \lambda_u = 1$, we have $\lambda_0 + \sum_{u=-T+1}^{T-1} \lambda_u = 2\sum_{u=-t+1}^{T-t} \lambda_u = 2$. Let's con-

sider the following estimator

$$\begin{aligned}
\hat{\beta}_{\mathbf{V}} &= \sum_{u=1}^{T-1} \frac{V(u)}{V_T/T} \lambda_u \tilde{\beta}(u) \\
&= \frac{1}{V_T/T} \sum_{u=1}^{T-1} \sum_{t=1}^{T} \lambda_u Y_t \left(2X_t - X_{t+u} - X_{t-u}\right) \\
&= \frac{1}{V_T/T} \sum_{t=1}^{T} \left\{ Y_t \sum_{u=1}^{T-1} \lambda_u \left(2X_t - X_{t+u} - X_{t-u}\right) \right\} \\
&= \frac{1}{V_T/T} \sum_{t=1}^{T} \left\{ Y_t \left[X_t \sum_{u=1}^{T-1} (2\lambda_u + 2\lambda_0) - \sum_{u=-T+1}^{T-1} \lambda_u X_{t+u} - X_t \lambda_0 \right] \right\} \\
&= \frac{1}{V_T/T} \sum_{t=1}^{T} \left\{ Y_t \left[X_t (\lambda_0 + \sum_{u=-T+1}^{T-1} \lambda_u) - \sum_{u=-T+1}^{T-1} \lambda_u X_{t+u} - X_t \lambda_0 \right] \right\} \\
&= \frac{1}{V_T/T} \sum_{t=1}^{T} \left\{ Y_t \left[2X_t \sum_{u=-t+1}^{T-t} \lambda_u - 2 \sum_{u=-t+1}^{T-t} \lambda_u X_{t+u} \right] \right\} \\
&= \frac{1}{V_T/2T} \sum_{t=1}^{T} \left\{ Y_t \left[X_t - \sum_{u=-t+1}^{T-t} \lambda_u X_{t+u} \right] \right\} \\
&= \frac{1}{\sum_{t=1}^{T} (X_t - \bar{X})^2} \sum_{t=1}^{T} \left\{ Y_t \left[X_t - \sum_{u=-t+1}^{T-t} \lambda_u X_{t+u} \right] \right\},
\end{aligned}$$

which is exactly the same as the estimator $\hat{\beta}_{\hat{\mathbf{S}}}$ using smoothing function $\hat{S}_t = \sum_{u=-t+1}^{T-t} \lambda_u Y_{t+u}$.

Table 1: Mean and standard deviation (in parenthesis) of the estimates from the simulation study (1000 simulations) for linear model. Method A is the estimator obtained from the model. Method B(a) is the weighted average using the first 20 $\tilde{\beta}(u)$s, where weight $w(u) = \sum_{t=1}^{T-u}(X_t - X_{t+u})^2 / \sum_{u=1}^{T-1}\sum_{t=1}^{T-u}(X_t - X_{t+u})^2$. Method B(b) is the weighted average using all the $\tilde{\beta}(u)$s (it should be the same as Method A for linear model). Method C(a) is the intercept by regressing $\tilde{\beta}(u)$s on natural spline of u with 3 degrees of freedom, weighted . Method C(b) is the same as C(a) except it uses all the $\tilde{\beta}(u)$s. The true $\beta = 0.1367$.

df in S_t	A	B		C	
		(a)	(b)	(a)	(b)
0	-0.129 (0.032)	0.089 (0.061)	-0.129 (0.032)	0.131 (0.095)	0.089 (0.066)
1	-0.137 (0.039)	0.089 (0.062)	-0.137 (0.039)	0.131 (0.095)	0.089 (0.066)
3	0.134 (0.069)	0.135 (0.069)	0.134 (0.069)	0.131 (0.095)	0.134 (0.072)
10	0.134 (0.072)	0.134 (0.072)	0.134 (0.072)	0.131 (0.095)	0.134 (0.073)

Table 2: Mean and standard deviation (in parenthesis) of the estimates from the simulation study (1000 simulations) for log-linear model. Method A is the estimator obtained from the model. Method B(a) is the weighted average using the first 20 $\tilde{\beta}(u)$s, where weight $w_t(u)^{\log} = (X_t - X_{t+u})^2/(1/\hat{\mu}_t + 1/\hat{\mu}_{t+u}) / \sum_{t=1}^{T-u}(X_t - X_{t+u})^2/(1/\hat{\mu}_t + 1/\hat{\mu}_{t+u})$. Method B(b) is the weighted average using all the $\tilde{\beta}(u)$s (it should be similar to Method A for log-linear model). Method C(a) is the intercept by regressing $\tilde{\beta}(u)$s on natural spline of u with 3 degrees of freedom, weighted . Method C(b) is the same as C(a) except it uses all the $\tilde{\beta}(u)$s. The true $\beta = 1.043$.

df in S_t	A	B		C	
		(a)	(b)	(a)	(b)
0	-2.247 (0.521)	1.611 (1.063)	-2.285 (0.530)	2.453(1.649)	1.612 (1.123)
1	-2.392 (0.646)	1.601 (1.071)	-2.430 (0.654)	2.453(1.650)	1.607 (1.127)
3	2.415 (1.170)	2.424 (1.205)	2.390 (1.175)	2.444 (1.647)	2.424 (1.250)
10	2.415 (1.225)	2.426 (1.246)	2.389 (1.229)	2.443 (1.644)	2.431 (1.252)

Table 3: Mean and bootstrapping standard deviation (in parenthesis) of the estimate of the effect of PM2.5 on mortality while controlling for proportion with high-school education, proportion with degree, proportion living in poverty, proportion unemployed, median income, and the standardized mortality ratio for COPD for three different age groups and three geographical regions using MCAPS data. Estimator is obtained directly from the model. Estimators B and C are the weighted average using all the $\tilde{\beta}$s and $\tilde{\beta}$s within 500 miles, respectively. The weight $w_{ij} = (X_i - X_j)^2/(1/\hat{\mu}_i + 1/\hat{\mu}_j)$. The Estimator D is obtained by regressing $\hat{\beta}_{i,j}$ on $log(d_{i,j})$ using i and j pairs with distance ≤ 500 miles, weighted by the reciprocal of the variance of $\hat{\beta}_{i,j}$.

Age	Region	A	B	C	D
65-74	West	0.79 (2.66)	2.64 (2.53)	2.52 (3.46)	8.34 (11.42)
	Central	23.00 (4.60)	23.31 (4.52)	17.78 (5.69)	-3.29 (22.91)
	East	11.04 (2.64)	10.45 (2.33)	11.81 (3.68)	20.44 (20.67)
75-84	West	1.97 (1.81)	3.05 (2.38)	3.61 (2.89)	4.43 (9.86)
	Central	15.64 (2.97)	15.99 (2.08)	11.70 (4.56)	-20.75 (19.60)
	East	9.76 (1.72)	9.59 (1.58)	9.27 (2.08)	3.87 (11.34)
85+	West	1.92 (1.57)	3.08 (1.71)	2.85 (2.62)	-6.35 (8.20)
	Central	1.73 (2.28)	3.06 (1.90)	-0.23 (3.56)	-29.59 (9.53)
	East	4.57 (1.52)	4.79 (2.16)	0.32 (2.52)	-26.57 (11.39)

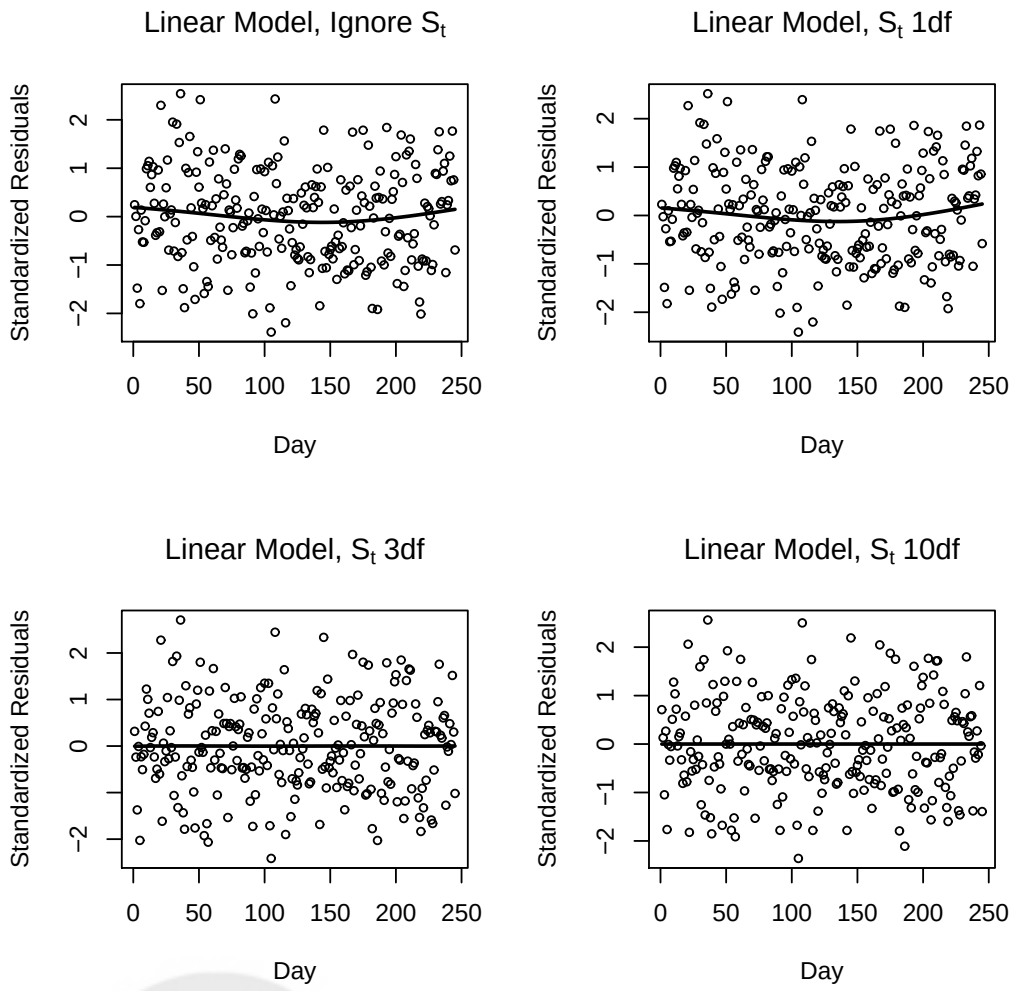

Figure 1: Standardized residual plots vs. time for four linear models using different degrees of freedom in estimating the effect of S_t. The solid lines are the smooth spline curve of the standardized residuals with 3 degrees of freedom.

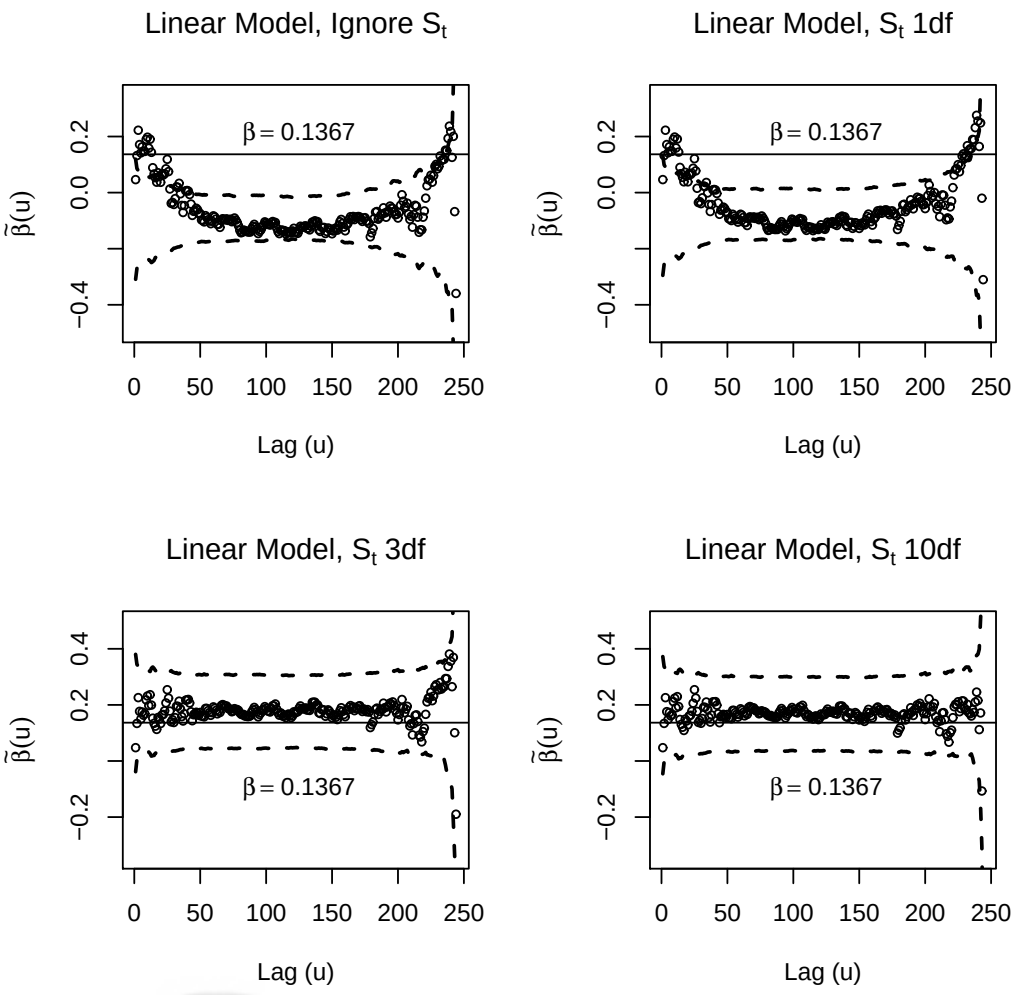

Figure 2: The LEP (Lagged-Estimator-Plot) $\tilde{\beta}(u)$ vs. lag u for four linear models using different degrees of freedom in estimating the effect of S_t. The horizontal line is the true β, and the dotted lines are the bootstrap 95% tolerance intervals under the null hypothesis.

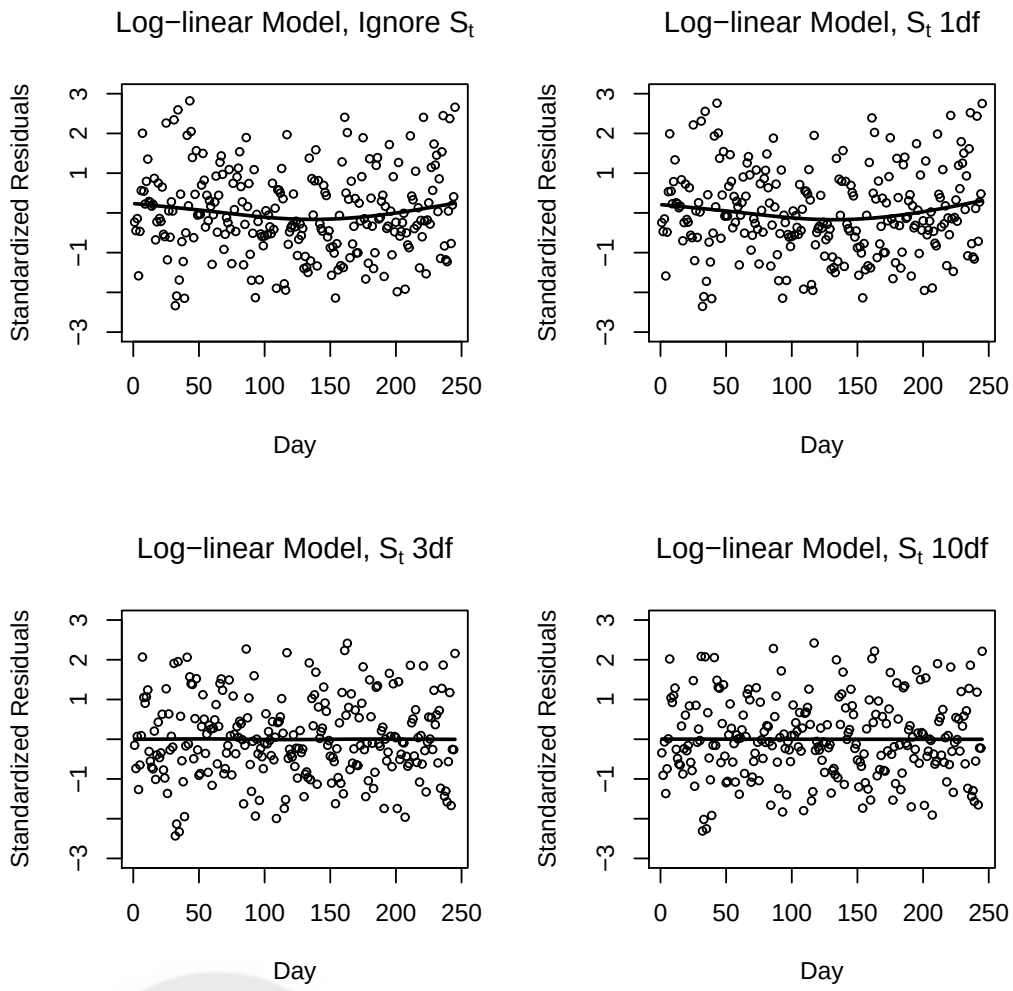

Figure 3: Standardized residual plots vs. time for four log-linear models using different degrees of freedom in estimating the effect of S_t. The solid lines are the smooth spline curve of the standardized residuals with 3 degrees of freedom.

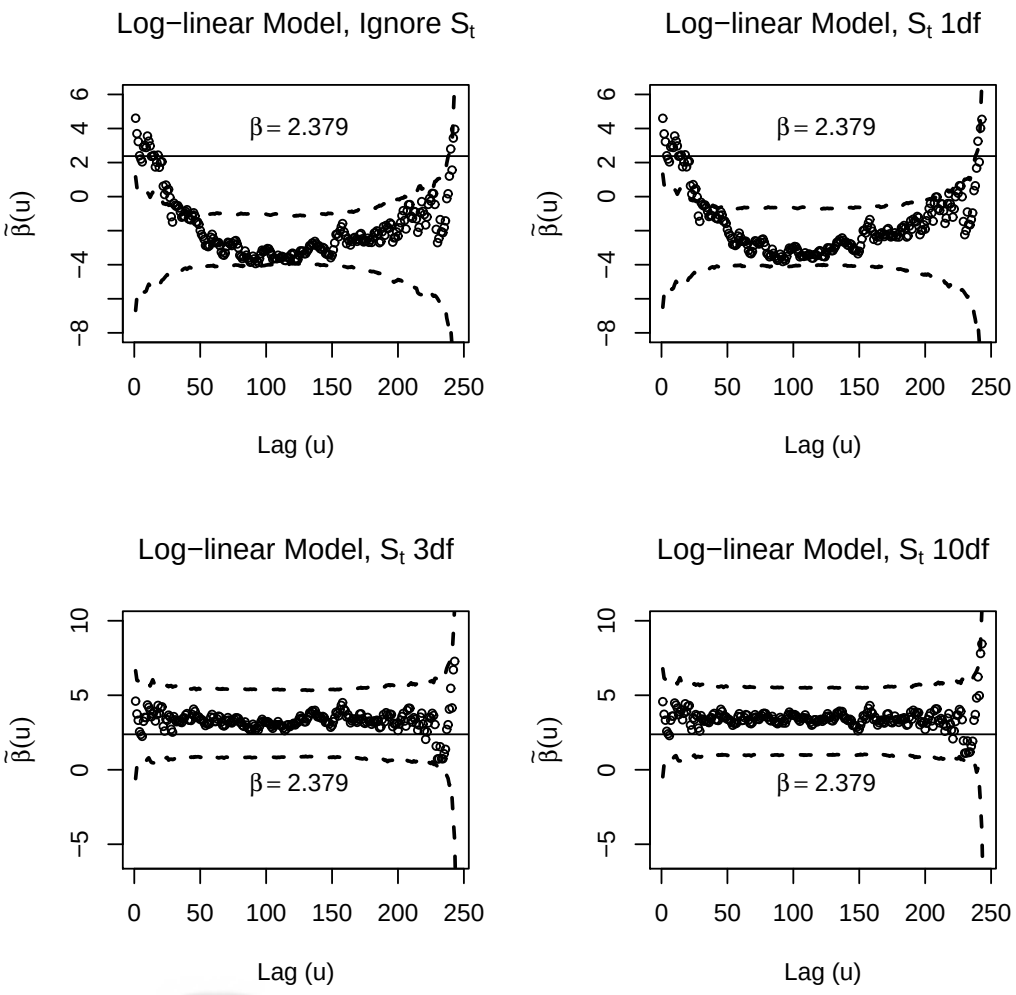

Figure 4: The LEP (Lagged-Estimator-Plot) $\tilde{\beta}(u)$ vs. lag u for four log-linear models using different degrees of freedom in estimating the effect of S_t. The horizontal line is the true β, and the dotted lines are the bootstrap 95% tolerance intervals under the null hypothesis.

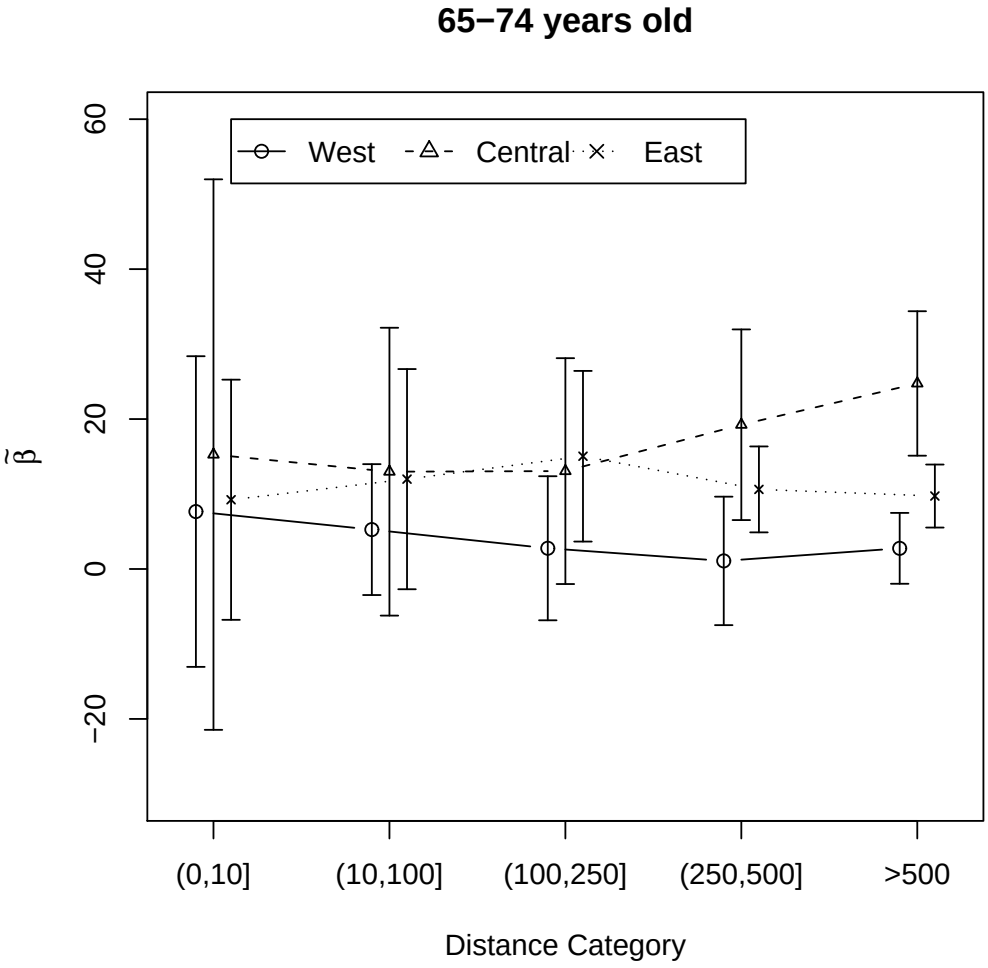

Figure 5: The LEP (Lagged-Estimator-Plot) $\tilde{\beta}$ vs. lag category with 95% confidence intervals using people 65-74 years old in three different geographical regions freedom while controlling for proportion with high-school education, proportion with degree, proportion living in poverty, proportion unemployed, median income, and the standardized mortality ratio for COPD.

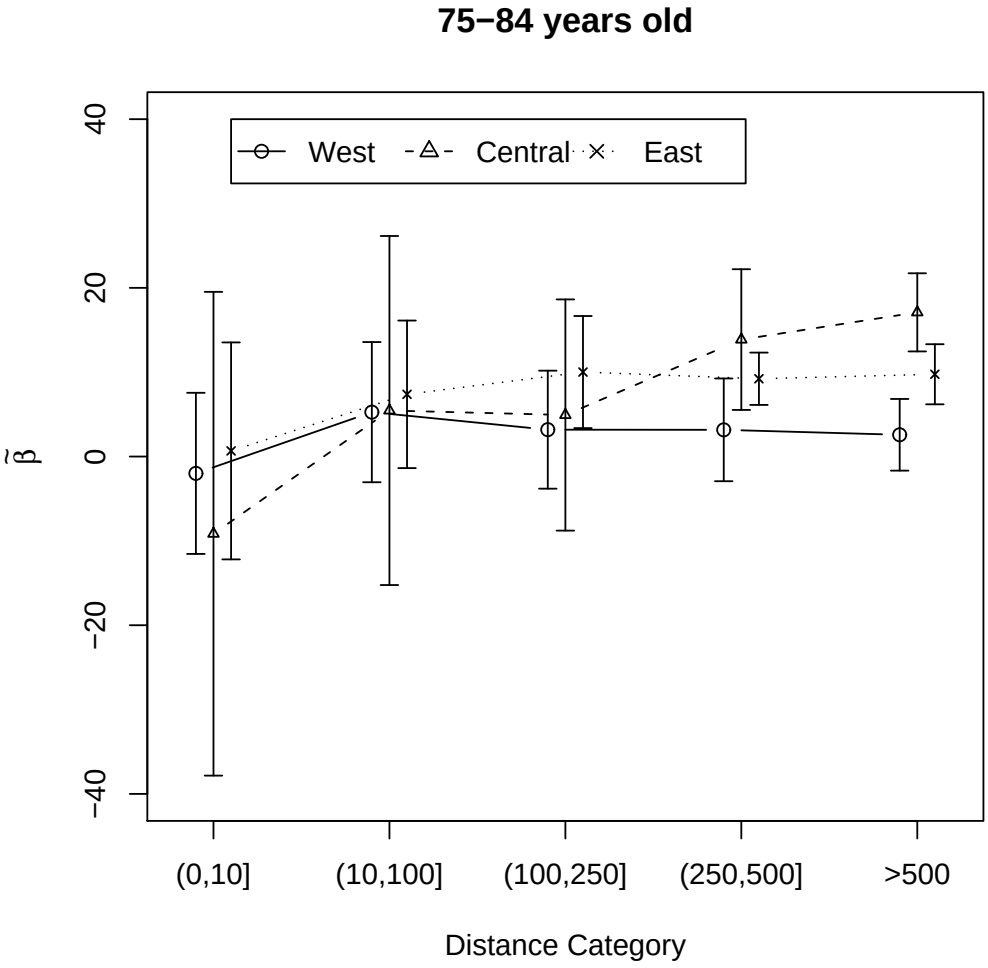

Figure 6: The LEP (Lagged-Estimator-Plot) $\tilde{\beta}$ vs. lag category with 95% confidence intervals using people 75-84 years old in three different geographical regions freedom while controlling for proportion with high-school education, proportion with degree, proportion living in poverty, proportion unemployed, median income, and the standardized mortality ratio for COPD.

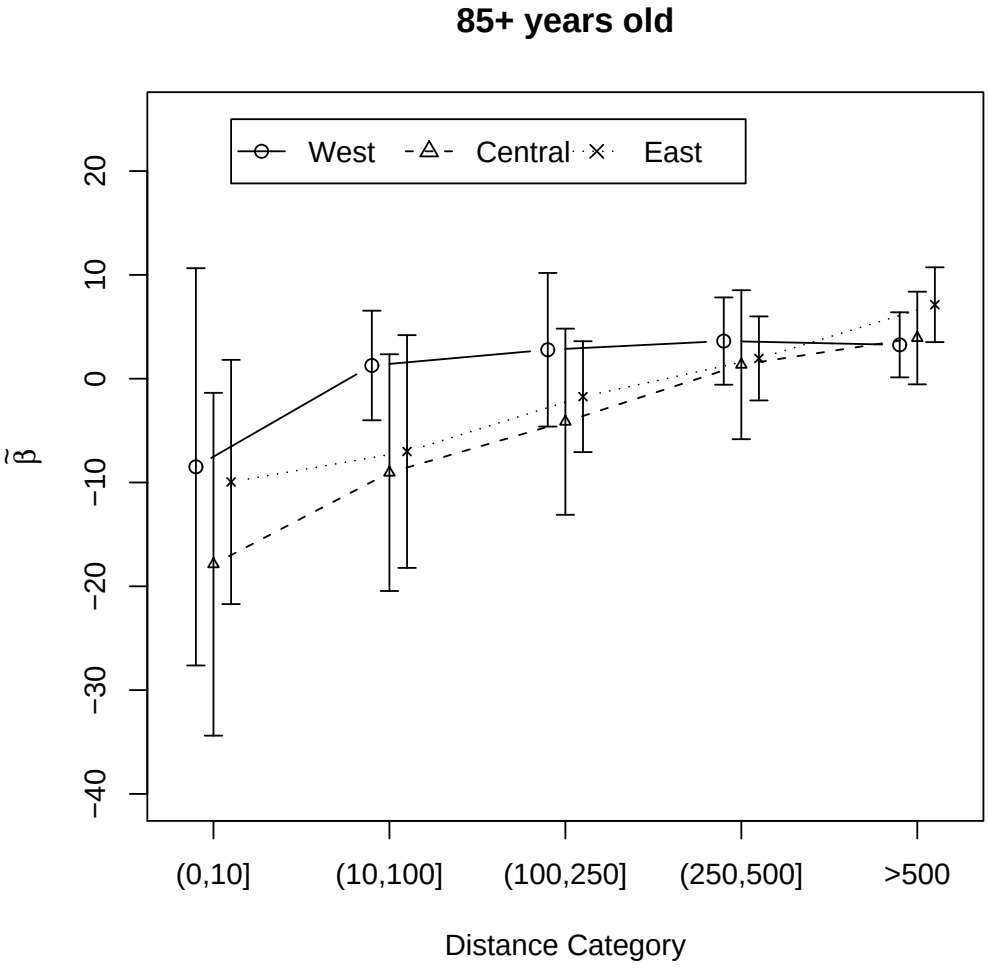

Figure 7: The LEP (Lagged-Estimator-Plot) $\tilde{\beta}$ vs. lag category with 95% confidence intervals using people 85 years and older in three different geographical regions while controlling for proportion with high-school education, proportion with degree, proportion living in poverty, proportion unemployed, median income, and the standardized mortality ratio for COPD.