



UW Biostatistics Working Paper Series

10-19-2006

Statistical Analysis of Air Pollution Panel Studies: An Illustration

Holly Janes

Johns Hopkins University, hjanes@jhsph.edu

Lianne Sheppard

University of Washington, sheppard@u.washington.edu

Kristen Shepherd

University of Washington, shepherdk@gmail.com

Suggested Citation

Janes, Holly; Sheppard, Lianne; and Shepherd, Kristen, "Statistical Analysis of Air Pollution Panel Studies: An Illustration" (October 2006). *UW Biostatistics Working Paper Series*. Working Paper 300.
<http://biostats.bepress.com/uwbiostat/paper300>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

1. Introduction

The panel study design is a popular tool for studying the short-term effects of air pollution on human health. By observing individuals repeatedly over time, the design enables assessment of the health effects of within-subject changes in exposure over time. Standard statistical methods for analyzing longitudinal data can be applied. However, the literature reveals that the longitudinal models and statistical issues pertaining to the analysis of longitudinal data are not well understood by many practitioners. In this paper, we illustrate longitudinal data methods using a recent air pollution panel study, and clarify issues that are sources of confusion. In Section 2, we describe the Seattle, Washington panel study, which is used for illustration. This data is used in Section 3 to illustrate the different approaches to the analysis of longitudinal data. We contrast the approaches with respect to parameter interpretation, accounting for correlation, and dealing with missing data. In Section 4, we demonstrate various techniques for controlling for confounding, and advocate partitioning the exposure effect into between- and within-subject components. In Section 5, we illustrate methods for exploring and summarizing panel study data. Notes on software are included in the appendix.

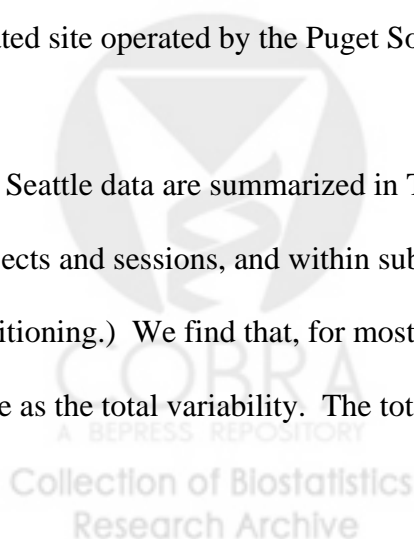
2. The Seattle Panel Study

The Seattle panel study was conducted between 1999 and 2002, and was designed to assess air pollution exposure and to evaluate the health effects of particulate matter (PM) and related pollutants among susceptible individuals (Koenig et al. 2003; Liu et al. 2003; Mar, Jansen, Shepherd, Lumley, Larson, and Koenig 2005; Allen et al. 2005). We restrict attention to the data collected on 19 children with asthma, aged 6-13 years. The children were recruited from a local allergy and asthma clinic, had physician-diagnosed asthma, and were taking asthma medications.

A full description of this dataset is published elsewhere (Koenig et al. 2003). The small study size means that assessing data structure and fitting complex models is difficult.

Each child participated in at least one 10-day monitoring session, with some sessions occurring in winter 2000-2001 and some in spring 2001. During the session, the children filled out daily questionnaires pertaining to asthma symptoms and daily activities. Exhaled nitric oxide (eNO), a measure of airway inflammation that is commonly elevated in asthmatics, was collected daily in the homes of the children, using a NO-inert and impermeable Mylar balloon. Measurements were taken in the afternoon or early evening, and children were asked to forego food intake for one hour before the measurement. The health outcomes we consider are eNO and what is called “overall well-being”, which is a binary variable based on whether the child reported feeling “better than average” on a given day. For simplicity, we refer to this adverse health event as “feeling worse”. $PM_{2.5}$, measured inside the subjects’ homes using single-stage inertial Harvard Impactors, serves as our exposure of interest. Indoor PM concentrations are thought to more closely represent personal exposures than do central-site monitoring data. The data also include daily measures of relative humidity and temperature, measured at Beacon Hill, a centrally located site operated by the Puget Sound Clean Air Agency.

The Seattle data are summarized in Table 1. Time-varying variables are summarized both across subjects and sessions, and within subject-sessions. (See Section 4.1 for the motivation for such partitioning.) We find that, for most variables, the variability within-subject-session is almost as large as the total variability. The total and within subject-session proportions of times subjects



reported feeling worse are also reported. Observe that for many subjects, there is little within-subject-session variability in this outcome.

3. Longitudinal Modeling

3.1 Outcome Variables

Longitudinal data models can be used with binary, count, and continuous outcomes. It is important that the outcomes vary over time within individuals, to allow for estimation of within-subject exposure effects.

3.2 Model specification and parameter interpretation

We consider three fundamentally different types of longitudinal models: marginal, conditional, and transition models. The Seattle panel study is used to illustrate each approach. For didactic purposes, we follow a somewhat non-traditional ordering: the analyses are presented in this section, and the data descriptives are saved for Section 5.

Denote the binary outcome, overall well-being, for subject i at time t by Y_{it} , and let X_{it} represent indoor $\text{PM}_{2.5}$ at time t . A marginal model specifies the form of $P(Y_{it} = 1 | X_{it})$, the probability of feeling worse as a function of the indoor $\text{PM}_{2.5}$ level. For example,

$$\text{logit } P(Y_{it} = 1 | X_{it}) = \beta_0 + \beta^M X_{it}. \quad (1)$$

This is the model for the mean of Y ; a model for the correlation among outcomes over time is specified separately, and will be discussed in Section 3.3.1. The parameters in (1) are estimated using generalized estimating equations (GEE) (Liang and Zeger 1986). The marginal

parameter, β^M , represents the difference in the log-odds of feeling worse between groups of children with a unit difference in indoor $\text{PM}_{2.5}$.

In contrast, the conditional model specifies a form for $P(Y_{it} = 1 | X_{it}, i)$, the probability of feeling worse for subject i (Diggle, Heagerty, Liang, and Zeger 2002). We use a random intercept model,

$$\text{logit } P(Y_{it} = 1 | X_{it}, i) = \beta_0 + b_i + \beta^C X_{it}, \quad (2)$$

and assume that the random intercept $b_i \sim N(0, \nu^2)$. A more complex conditional model would also allow the slope to be random. The parameters in (2) are estimated using maximum likelihood, since the distribution for the random effects induces a form for the likelihood of the data. The conditional parameter, β^C , represents a child's expected change in the log-odds of feeling worse due to a unit increase in indoor $\text{PM}_{2.5}$. This is the within-subject effect of changes in indoor $\text{PM}_{2.5}$. Hence, the conditional model facilitates making inferences about individuals, rather than groups of individuals. The random intercept b_i is the i^{th} child's baseline level of overall well-being, and so ν^2 describes the heterogeneity in baseline overall well-being across children.

A transition model specifies a form for $P(Y_{it} = 1 | X_{it}, Y_{it-1}, \dots, Y_{i1})$, the probability of feeling worse as a function of past overall well-being (Diggle, Heagerty, Liang, and Zeger 2002). We use a transition model that conditions on the previous day's overall well-being, though other previous outcomes could also be included. We write

$$\text{logit } P(Y_{it} = 1 | X_{it}, Y_{it-1}) = \beta_0 + \beta^T X_{it} + \gamma Y_{it-1}. \quad (3)$$

The parameters in (3) are estimated using GEE (likelihood methods can also sometimes be used for estimation (Diggle, Heagerty, Liang, and Zeger 2002)). The transition parameter, β^T , can be interpreted as the difference in the log-odds of feeling worse between groups of children which have one unit difference in indoor $\text{PM}_{2.5}$ today, but which had the same overall well-being yesterday. For this reason, a transition model for a binary outcome can be thought of as a model for incidence, while the marginal model describes the prevalence. The parameter γ represents the transition effect, or the effect of yesterday's overall well-being on today's overall well-being, holding today's indoor $\text{PM}_{2.5}$ level constant. In order for a transition model to be well defined, observations need to be equally spaced.

Models (1), (2), and (3) were fit to the Seattle data (see Table 2). In this data, the relationship between X and $\text{logit}(E(Y))$ is not linear, but piecewise linear (see Section 5). Hence, we model exposure, indoor $\text{PM}_{2.5}$, using a linear spline with one knot at the median (Greenland 1995). In general, it is important that the functional form of the model be explored; a linear model may not be adequate. We also include age, BMI, relative humidity, and temperature as covariates in all models, in order to increase precision.

We estimate different PM effects using the three types of models, especially above the PM median. In all three models, we see little evidence of a PM effect below the median. Above the median, the conditional PM effect is much larger than the marginal or transitional effects. The conditional model implies that a $10 \mu\text{g}/\text{m}^3$ increase in indoor $\text{PM}_{2.5}$ above the median increases the odds of feeling worse by 496% for a given child (95% CI: 71% to 1973%). According to the marginal model, a $10 \mu\text{g}/\text{m}^3$ difference in indoor $\text{PM}_{2.5}$ above the median is associated with

142% greater odds of feeling worse (95% CI: 33% less to 780% greater). Based on the transition model, there is an associated 153% higher odds of feeling worse in a population with $10 \mu\text{g}/\text{m}^3$ higher indoor $\text{PM}_{2.5}$ today (above the median), but the same overall well-being yesterday (95% CI: 7% to 495%). The different results are a consequence of the fact that the PM parameters represent different quantities. The parameter of interest should be chosen based on the scientific question.

All three types of longitudinal models can be used with other types of outcome variables. A more general form for the random intercept model, for example, is

$$g(E(Y_{it} | X_{it}, i)) = \beta_0 + b_i + \beta^C X_{it},$$

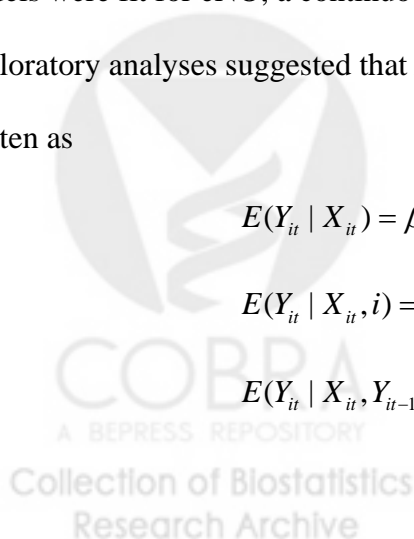
where g is called the link function. We have used the logit function for g , which is appropriate for a binary outcome. With a continuous outcome, it is common to let $g(w) = w$, and with a count outcome, $g(w) = \log(w)$.

Marginal, conditional (random intercept), and transitional (conditioning on yesterday's outcome) models were fit for eNO, a continuous outcome collected in the Seattle panel study (see Table 3). Exploratory analyses suggested that X be modeled linearly (see Section 5). The models are written as

$$E(Y_{it} | X_{it}) = \beta_0 + \beta^M X_{it}$$

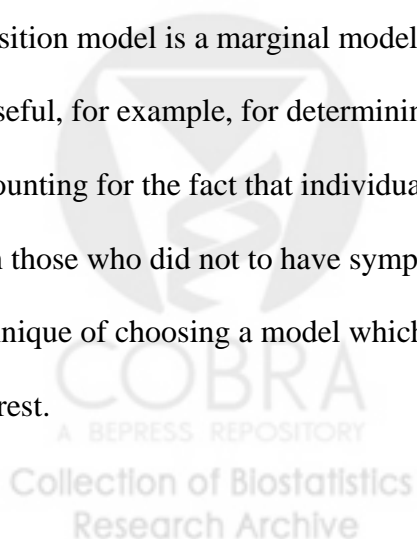
$$E(Y_{it} | X_{it}, i) = \beta_0 + b_i + \beta^C X_{it}$$

$$E(Y_{it} | X_{it}, Y_{i,t-1}) = \beta_0 + \beta^T X_{it} + \gamma Y_{i,t-1}.$$



The covariates age, gender, BMI, relative humidity, and temperature are also included as predictors. We find that the conditional and marginal parameter estimates are nearly identical, and differ slightly due to differences in estimation procedures. In fact, with an identity link function, the conditional and marginal parameters will always agree. With a certain parametrization of the transition model, this parameter will also be the same (Diggle, Heagerty, Liang, and Zeger 2002). According to the conditional model, we estimate that each $10 \mu\text{g}/\text{m}^3$ increase in indoor $\text{PM}_{2.5}$ is associated with a 4.10 ppb increase in the child's eNO (95% CI: 1.89 ppb to 6.32 ppb).

The scientific question of interest should dictate the type of longitudinal model that is used. A marginal model is used to estimate the effect of exposure on population average outcomes. Suppose, for example, that we want to assess the impact of increasing PM on rates of asthma-induced hospital visits. Since our interest is in a population mean, a marginal model should be used. If individual-level exposure effects are desired, a conditional model is more appropriate. Such a model can be used, for instance, to assess the effect of PM on individuals' immune system biomarkers, accounting for heterogeneity in baseline levels of the biomarkers. Finally, a transition model is a marginal model which controls for outcome history. This type of approach is useful, for example, for determining the effect of increasing PM on rates of asthma symptoms, accounting for the fact that individuals who experience symptoms on one day are more likely than those who did not to have symptoms on the following day. These examples illustrate the technique of choosing a model which corresponds most closely to the scientific question of interest.



An aggregated analysis has historically been a popular approach to analyzing panel data. All outcomes on the same day are collapsed into a total, called the panel average (or panel attack rate for binary outcomes) (Korn and Wittemore 1979). The panel average is then regressed on exposure. This approach is subject to many of the problems associated with ecological studies (Sheppard, Prentice, and Rossing 1996; Sheppard 2002). In addition, inference from a linear model which assumes that outcomes on successive days are independent and have constant variance may be incorrect. Finally, bias may be incurred due to the fact that subject-specific missing data patterns are not taken into account (Dominici, Sheppard, and Clyde 2003). For these reasons, aggregated analyses have been discouraged since 1979 (Korn and Wittemore 1979).

In our illustration, we specified exposure to be today's indoor $PM_{2.5}$. However, longitudinal models can incorporate other exposure measures, such as particular exposure lags, cumulative exposure, or different exposure components (see Section 4.1). Distributed lag exposure models can also be used (Schwartz 2000; Diggle, Heagerty, Liang, and Zeger 2002; Goodman, Dockery, and Clancy 2003). These models allow the health effects of air pollution to extend over time.

Exposure is modeled as a linear combination of exposure lags, say $\eta_0 X_{it} + \dots + \eta_{t-q} X_{it-q}$, where the coefficients, η_s , can be constrained to have a specific functional form. The sum of the

coefficients, $\sum_{s=0}^q \eta_s$, is interpreted as the net effect of exposure on the outcome.

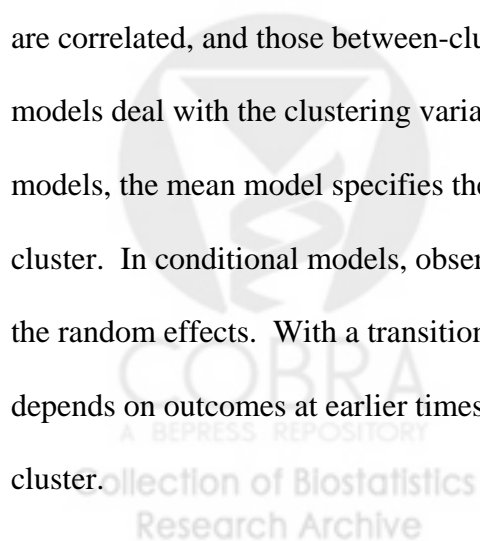
3.3 Accounting for correlation

3.3.1 Modeling the correlation structure

In the Seattle panel study, outcomes for each child are correlated, as measurements taken on the same child are likely to be more similar than those on different children. All longitudinal studies have correlation due to repeated measures. This is what makes using special longitudinal models necessary. But the Seattle panel study has additional correlation structure due to the fact that some children were observed for more than one session. Observations in the same session are likely to be more similar than those in different sessions. Multiple levels of correlation, due to observation times or measurement error, are common in longitudinal studies.

Correlation among outcomes is typically of two types: exchangeable and serial. An exchangeable structure implies that any pair of outcomes has the same correlation, while serial correlation decreases with the time elapsed between the two observations. There may be both serial and exchangeable correlation in an outcome if, for example, correlation decreases with time separation, but there is still long-term correlation between observations that are far apart.

In longitudinal models, correlation is accounted for by specifying the “clustering” variable, or the independent unit of observation. This specification indicates that observations within-cluster are correlated, and those between-clusters are independent. The three types of longitudinal models deal with the clustering variable in different ways. With conditional and transition models, the mean model specifies the correlation structure among observations in the same cluster. In conditional models, observations in the same cluster are correlated because they share the random effects. With a transition model, the mean model specifies that the outcome at time t depends on outcomes at earlier times, and this induces correlation among outcomes in the same cluster.



With a marginal model, the correlation structure is specified separately from the mean model, using a “working correlation matrix” (Diggle, Heagerty, Liang, and Zeger 2002). Common working correlation choices are independence (observations in the same cluster are assumed to be independent), autoregressive, and exchangeable structures. With a transition model, as well, additional correlation structure can be added using a working correlation model.

In the Seattle data, we specify subject-session as the clustering variable. This structure assumes that observations within the same subject-session are correlated, while observations in the same session but different subjects, or the same subject but in different sessions, are uncorrelated. We chose this model for simplicity and consistency with the literature. An alternative would be to cluster on both subject and session. This stipulates that observations in the same subject and the same session are correlated, as are those across subjects in the same session, and those across sessions for the same subject. But with many subjects having only one session, and others only two or three sessions, the data would not allow us to estimate the within-subject, between-session correlation precisely. Hence, in our notation, the i subscript refers to subject-session (i.e., X_{it} represents exposure for subject-session i at time t).

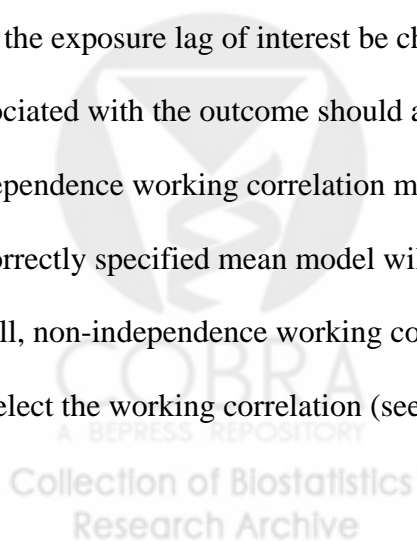
3.3.2 Modeling correlation structure in marginal models

With marginal and transition models (estimated using GEE), if the parameters in the mean model are the primary interest, correct specification of the mean model is essential. However, even if the correlation structure is incorrectly specified, the mean model parameter estimates and standard errors are still valid if so-called “robust” standard error estimates are used (White 1982;

Liang and Zeger 1986; Royall 2005). Robust standard errors are reported by most statistical packages, though the number of individuals must be large in order to ensure their validity. If, instead, the variance model is of interest, perhaps for studying the variation in within-subject outcomes, correctly specifying the correlation structure is also necessary.

Regardless of which parameters are of primary interest, correct specification of the correlation structure will lead to gains in efficiency (Diggle, Heagerty, Liang, and Zeger 2002). But Pepe and Anderson (1994) showed that, if the appropriate exposure lags are not included in the mean model, the parameter estimates will be biased unless a working independence correlation structure is used. Therefore, researchers face a dilemma: specify an appropriate working correlation, and risk bias in the parameter estimates, or specify working independence, report robust standard errors, and risk a loss of efficiency.

Schildcrout and Heagerty (2004) have provided some guidance in choosing between these two strategies. Their advice is based on the assumption that a marginal model with one exposure lag is the model of interest, since interpretation of a one-lag model is simplest. First, they suggest that the exposure lag of interest be chosen carefully. The degree to which other lags are associated with the outcome should also be determined. If there are other large lag effects, an independence working correlation matrix should be used, since the bias associated with the incorrectly specified mean model will be large. On the other hand, if the other lag effects are small, non-independence working correlation should be used. Exploratory analyses can be used to select the working correlation (see Section 5.2).



In the Seattle data, we use a working independence correlation structure and report robust standard errors to guard against bias in the parameter estimates.

3.4 Missing data

In the Seattle panel study, there is a substantial amount of missing data. The variables overall well-being, eNO, indoor PM_{2.5}, and relative humidity are intermittently missing. Of a total of 330 observations, 68 (21%) are missing. The sources of missingness, according to our best judgment, are tabulated in Table 4.

With longitudinal data, missing data is a common problem. Subjects may have observations missing intermittently, or they may drop out at some point. Missing observations are those that are missing unintentionally; measuring subjects at different pre-specified time points results in unbalanced, but not missing, data.

Fortunately, all of the longitudinal models we have discussed allow researchers to use the available observations for each subject. That is, subjects need not be dropped from the analysis if they have missing data. However, each of the models makes certain assumptions about the reasons for missingness.

Missingness is frequently partitioned into three categories, following the characterization proposed by Little and Rubin (1987). Data is said to be missing completely at random (MCAR) if missingness is independent of both observed and unobserved data. Hence, observations are missing simply due to random chance, as if on the basis of a coin toss. In contrast, if data is

missing at random (MAR), missingness depends only on the data that is collected. Finally, if data is informatively missing, the missingness mechanism depends on both observed and unobserved data.

Determining the type of missing data at hand is difficult. There are some ad-hoc procedures for using the data to distinguish between MCAR and MAR (see Diggle, Heagerty, Liang, and Zeger (2002)). However, it is impossible to determine from the observed data whether missingness is informative. In general, the type of missingness is determined by carefully considering the sources of missing data in the particular study. In the Seattle data, missingness is a consequence of equipment problems, and children not being home for the visit, eating within one hour of the eNO measurement, or leaving questions blank on the questionnaire (see Table 4). First, it is reasonable to assume that missingness due to equipment problems is true measurement error (and not measurements below the limit of detection). We assume that this data is MCAR, but this is certainly debatable, and should be mentioned as a limitation of our analysis. Data that is missing due to the behavior of the study subjects is more worrisome. There is a possibility that children who are not home for a visit, who neglect to answer the symptom questions on the questionnaire, or who eat within one hour of the eNO measurement, are not feeling well. Since this type of missingness may be associated with the outcome (an observed variable), we classify it as MAR. We assume that the missingness mechanism can be explained entirely by the observed data, and hence is not informative, but again this is debatable.

Likelihood based methods, such as conditional models, are valid so long as data is MAR (or MCAR). However, non-likelihood based analyses, such as marginal and transitional models (fit

using GEE), require that the data be MCAR, a stronger assumption. Hence, the type of missingness should be seriously considered when choosing a statistical analysis. A number of methods have been proposed for using GEE with MAR (Heyting, Tolboom, and Essers 1992; Robins et al. 1995; Scharfstein, Rotnitzsky, and Robins 1999), and for analyzing data with informative missingness (Wu and Bailey 1989; Little 1993; Diggle and Kenward 1994; Diggle, Heagerty, Liang, and Zeger 2002; Wu and Carroll 2005), but each depend on making additional assumptions, in particular about the missingness mechanism.

4. Controlling for Confounding

4.1 Between- and within-subject effects

Consider the marginal model for the Seattle eNO data:

$$E(Y_{it} | X_{it}) = \beta_0 + \beta^M X_{it}, \quad (4)$$

where Y_{it} represents eNO and X_{it} denotes indoor $PM_{2.5}$, for subject-session i at time t . (For simplicity, here we ignore the other covariates included in the model.) This model implies that the effect of increasing PM on eNO is the same, regardless of whether the difference is in the same subject-session, or comparing across subject-sessions. In fact, given the design of the study, the exposure can be partitioned into three components: the between-subject effect; the within-subject, between-session effect; and the within-subject, within-session effect. These three effects may be very different, due to confounding (Palta and Yao 1991; Diggle, Heagerty, Liang, and Zeger 2002). The between-subject effect is the difference in average eNO associated with a unit increase in indoor $PM_{2.5}$, comparing subjects. This is potentially confounded by time-independent confounders, such as the region in which the subject lives. The within-subject, between-session exposure effect is the difference in average eNO associated with a unit increase

in indoor $PM_{2.5}$, when comparing across sessions for the same subject. Since sessions are often in different seasons, and season is strongly associated with exposure and outcome, this exposure effect is confounded. The within-session, within-subject effect of exposure is the real parameter of interest; it represents the difference in average eNO associated with a unit increase in indoor $PM_{2.5}$, when comparing across days in the same session for the same subject. In (4), β^M represents a combination of these three exposure effects. In order to allow them to differ, we should partition exposure into the three components, and include each as an independent term in the model.

We rewrite the marginal model, allowing the three exposure effects to have different magnitudes.

In this section, we use new subscripting notation. Let i index subject, j session, and t time.

Denote the average of a variable Z by \bar{Z} . We partition the exposure, X_{ijt} , into three components:

\bar{X}_i : the average exposure for subject i (the between-subject component)

$\bar{X}_{ij} - \bar{X}_i$: for subject i , the average exposure in session j minus the overall average exposure (the within-subject, between-session component)

$X_{ijt} - \bar{X}_{ij}$: for subject i and session j , the exposure on day t minus the average exposure in the session (the within-subject, within-session component)

We write

$$E(Y_{it} | X_{it}) = \beta_0 + \beta_{bs}^M \bar{X}_i + \beta_{wsbs}^M (\bar{X}_{ij} - \bar{X}_i) + \beta_{wsws}^M (X_{ijt} - \bar{X}_{ij}). \quad (5)$$

The parameters β_{bs}^M , β_{wsbs}^M , and β_{wsws}^M are the between-subject; within-subject, between-session;

and within-subject, within-session effects of exposure. The parameter of interest is β_{wsws}^M , the

effect of day-to-day variation in indoor $PM_{2.5}$ on average eNO, for a given individual in a given

session. This parameter corresponds with the component of exposure variation that is least likely to be confounded.

If, by design, each subject has only one observation period, partitioning with respect to session is unnecessary. In this case, the two components of exposure are the between-subject and within-subject components. Alternatively, if the exposure is shared and all subjects are observed at the same times (as in a traditional panel study design), partitioning is impossible and unnecessary.

The within-subject, within-session exposure component is “mean balanced”, since it has mean zero for all subjects. Schildcrout and Heagerty (2004) suggest that mean balanced covariates behave more favorably in terms of the bias/efficiency tradeoff associated with marginal models (discussed in Section 3.3.2). That is, for such covariates, the use of independent working correlation is less inefficient, and use of non-independence working correlation is associated with less bias.

The parameter estimates for model (5) and the Seattle eNO data are displayed in Table 5. The covariates age, BMI, relative humidity, and temperature are also included as predictors. The between-subject and within-subject, between-session exposure effects are potentially confounded by season and are not of primary interest. Note that the within-subject, between-session exposure effect has a very large standard error, due to the fact that only 13 children participated in more than one session, and these children had only two or three sessions. We estimate that the within-subject, within-session effect of a $10 \mu\text{g}/\text{m}^3$ increase in indoor $\text{PM}_{2.5}$ is an associated increase of 3.90 ppb in eNO (95% CI: 0.74 ppb to 7.06 ppb). All three exposure effects are

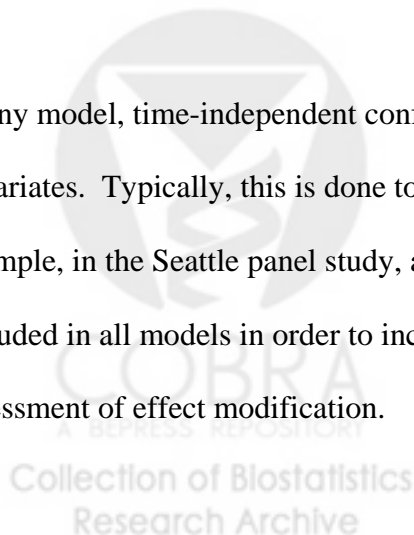
consistent with this value. As expected, the effect of unpartitioned indoor $PM_{2.5}$ in the marginal model (Table 3) is a weighted average of the coefficients of the three exposure components considered here.

4.2 Time-independent confounding

Factors such as the region in which a subject lives may be important time-independent confounders in panel studies. One of the uses of partitioning exposure is to control for such confounding. In particular, the inclusion of the between-subject exposure component controls for static differences between individuals. Such partitioning is useful in marginal, conditional, and transition models.

In a conditional model, the use of a random effect for each subject (or for each subject-session) will help control for differences between subjects. In a transition model, conditioning on previous outcomes controls for a large amount of time-independent and time-dependent confounding, since the exposure effect represents the association between exposure and outcome among observations with the same outcome history.

In any model, time-independent confounders can be further controlled by including them as covariates. Typically, this is done to gain efficiency in estimating the effect of exposure. For example, in the Seattle panel study, age and BMI, factors which are not confounders, are included in all models in order to increase efficiency. Such adjustment also allows for assessment of effect modification.



4.2 Time-dependent confounding

Air pollution is highly influenced by time-varying factors such as season and temperature, many of which are also associated with health outcomes of interest. The most common method of controlling for time-varying confounders is by including them in the regression model. (See, for example, Yu, Sheppard, Lumley, Koenig, and Shapiro (2000)) This approach requires selecting a functional form for the covariates, as well as choosing the appropriate lags to be included. Model selection concerns arise, as researchers select from among candidate models which control for different confounders in different ways. Lumley and Sheppard (2000) have shown that model selection bias can be of the same magnitude as the health effects themselves. In order to avoid this bias, the method of controlling for confounding should be chosen *a priori*.

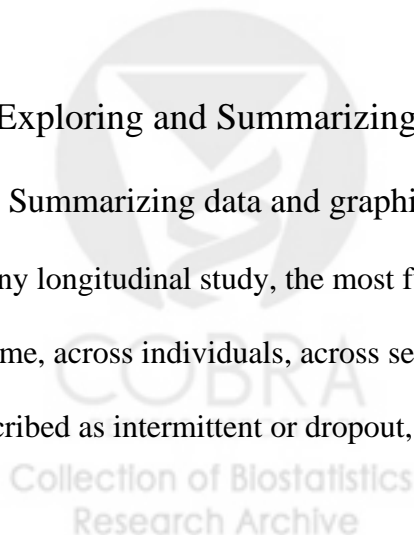
Partitioning exposure will control for a certain amount of time-dependent confounding.

However, confounders which vary within an individual's period of observation will not be controlled. For example, if the subject-session is long, there may be residual confounding due to season. Further partitioning could be employed to enable estimation of the within-subject, within-season effect of changes in exposure.

5. Exploring and Summarizing Longitudinal Data

5.1 Summarizing data and graphical displays

In any longitudinal study, the most fundamental data descriptives relate where the data lie (e.g., in time, across individuals, across sessions) (Kunzli and Schindler 2005). Missingness should be described as intermittent or dropout, and tabulated (see Tables 1 and 4 for examples).



A plot of the exposure and outcome trends over time is a simple and useful first descriptive (see Figure 1). This is especially informative if both exposure and outcome are continuous and measured over a long time period. Lines should connect observations within the same subject-session. Such a plot displays the within and between subject-session patterns and variability in exposure and outcome, and can be used to assess whether exposure and outcome series have peaks at similar times. In Figure 1, eNO and indoor $PM_{2.5}$ are plotted as a function of time. Observe the large amount of variability in both exposure and outcome series over time, and within subject-session. As is typical in the air pollution setting, where exposure effects are very small, it is difficult to see an association between exposure and outcome in this figure. The figure can also be misleading because any association observed may be attributable to confounding.

In order to observe the relationship between exposure and outcome apart from confounding, it is useful to plot exposure and outcome residuals from models that include the confounders. In Figure 2, we plot eNO residuals against indoor $PM_{2.5}$ residuals, each from linear models which include age, BMI, relative humidity, and temperature. A smooth curve fit to the residuals is overlaid. We see a modest, linearly increasing trend in eNO as a function of indoor $PM_{2.5}$, and note that much of the variation in eNO is not explained by the exposure.

The drawback of a plot such as Figure 2 is that it displays the association between overall (unpartitioned) exposure and outcome. In fact, the three exposure components should be modeled separately (see Section 4.1). Therefore, it is appropriate to examine three plots: within-subject/within-session outcome residuals versus within-subject/within session exposure

residuals; within-subject/between-session outcome residuals versus within-subject/between-session exposure residuals; and between-subject outcome residuals versus between-subject exposure residuals. These plots are not shown for the Seattle data because they look very similar to Figure 2.

In order to explore the transition model, it is useful to plot outcome residuals against previous lags of outcome residuals. This is shown in Figure 3 for the Seattle eNO data, where residuals come from linear models including the covariates: indoor $PM_{2.5}$, age, BMI, relative humidity, and temperature. We see that there is a strong association between eNO today and eNO yesterday, even after accounting for PM and other covariates.

With a binary outcome, it is useful to summarize the outcome cross-sectionally as well as within-individuals. See Table 1 for an example.

Exploring the association between a binary outcome and a continuous exposure is difficult, due to the fact that plotting the outcome is not very informative. A plot of the outcome residuals as a function of the exposure residuals can be useful. The residuals for overall well-being (from a logistic model) are plotted as a function of indoor $PM_{2.5}$ residuals (from a linear model) in Figure 4. Both models are adjusted for age, BMI, relative humidity, and temperature. The plot is still difficult to interpret, since the residuals cluster into two groups. We can, however, determine with a smoothed curve overlaid that the association between exposure and outcome does not appear to be linear. This figure suggests that indoor $PM_{2.5}$ be modeled piecewise-linearly in the overall well-being models. As with a continuous outcome, plots of partitioned exposures and

outcomes should also be examined. Plots of outcome residuals versus previous lags of outcome residuals can also be examined to determine the necessity of controlling for outcome history.

5.2 Exploring correlation structure

In order to correctly specify the correlation structure of the data, we must explore the dependence among outcomes in the same cluster. We are interested in the correlation that exists after the mean structure has been taken out. With a continuous outcome, the correlation structure can be summarized using a variogram (Diggle 1990; Diggle, Heagerty, Liang, and Zeger 2002). For residuals r_{ij} , the differences $v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2$ are plotted against the corresponding time differences $u_{ijk} = t_{ij} - t_{ik}$. Such a plot is shown in Figure 5 for the eNO residuals in the Seattle data. A smooth curve is overlaid to aid viewing of the trend (the y-axis is truncated at 300). A solid line is also added at the level of the total estimated variance in the eNO residuals. We find that the differences between observations tend to increase with time lag, corresponding to a serial correlation structure. The variogram increases all the way to the overall variance, indicating that the correlation decays to zero, and has no long-term component. At larger time lags, the variogram actually moves slightly above the overall variance, a consequence of the instability of the estimates at these large lags. We conclude that a serial correlation model seems to be adequate.

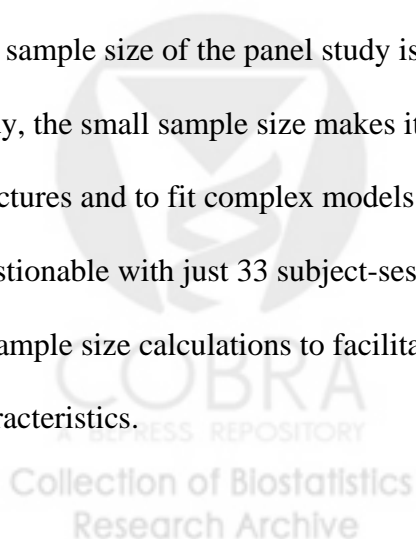
With a binary outcome, correlation is not a useful measure of association. A more meaningful summary of the relationships among binary observations over time is the odds ratio. Heagerty and Zeger (1998) suggest plotting a lorelogram, or the log-odds ratio for pairs of observations in the same cluster as a function of time separation. We show an example in Figure 6 for the

overall well-being outcome. The log odds ratios for pairs of observations in the same subject-session do not appear to decrease with time separation, and remain extremely high. Children who report feeling worse (better) than average on one day are likely to feel worse (better) than average on all other days in the same subject-session. This high degree of “correlation” persists when the data are stratified by various covariates. An exchangeable correlation structure seems a reasonable assumption.

6. Discussion

The panel study design is a powerful tool for assessing the short-term association between air pollution and health outcomes over time, within individuals. We reviewed here the existing tools for exploring and quantifying this association. The various modeling approaches differ primarily with respect to the scientific questions that they answer, but also with respect to how they model correlation, deal with missing data, and control for confounding. We have advocated partitioning of the exposure, in order to focus on the exposure effect that is least likely to be confounded.

The sample size of the panel study is an important attribute of the design. In the Seattle panel study, the small sample size makes it difficult to distinguish between different types of data structures and to fit complex models. The validity of our robust standard errors is also questionable with just 33 subject-sessions. An important future research topic is the derivation of sample size calculations to facilitate design of efficient studies with specified operating characteristics.



Appendix

GEE can be fit in STATA using *xtgee*, in Splus using *gee*, or in SAS using *proc genmod*. In Splus, robust standard errors are always reported, while STATA requires the *robust* option and SAS the *covb* option to request robust standard errors. Linear random effects models can be fit in STATA using *xtgee*, in Splus using *lme*, and in SAS using *proc mixed*. Logistic random intercept models can be fit in STATA using *xtlogit*, and general non-linear random effects models can be fit in SAS using *proc glimmix*.

References

- Diggle PJ. Time Series: A Biostatistical Introduction. Oxford: Oxford University Press, 1990.
- Diggle PJ, Heagerty P, Liang KY, Zeger SL. The Analysis of Longitudinal Data. 2 Ed. Oxford: Oxford University Press, 2002.
- Diggle PJ, Kenward MG. Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics* 1994; 43:49-73.
- Dominici F, Sheppard L, Clyde M. Health effects of air pollution: A statistical review. *International Statistical Review* 2003; 71:243-276.
- Goodman PG, Dockery DW, Clancy L. Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environmental Health Perspectives* 2003; 112:179-185.
- Greenland S. Dose-response and trend analysis in epidemiology: Alternatives to categorical analysis. *Epidemiology* 1995; 6(4):345-347.

- Heagerty P, Zeger SL. Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *Journal of the American Statistical Association* 1998; 93:150-162.
- Heyting A, Tolboom JTBM, Essers JGA. Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine* 1992; 11:2043-2062.
- Koenig JQ, Jansen K, Mar TF, Lumley T, Kaufman J, Trenga CA, Sullivan J, Liu L-JS, Shapiro GG, Larson TV. Measurement of offline exhaled nitric oxide in a study of community exposure to air pollution. *Environmental Health Perspectives* 2003; 111:1625-1629.
- Koenig JQ, Mar TF, Allen RW, Jansen K, Lumley T, Sullivan JH, Trenga CA. Pulmonary effects of indoor- and outdoor-generated particles in children with asthma. *Environmental Health Perspectives* 2003; 113: 449-503.
- Korn EL, Whittemore AS. Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* 1979; 35:795-802.
- Kunzli N, Schindler C. A call for reporting the relevant exposure tem in air pollution case-crossover studies. *Journal of Epidemiology and Community Health* 2005; 59:527-530.
- Liu LJ, Box M, Kalman D, Kaufman J, Koenig J, Larson T, Lumley T, Sheppard S, Wallace L. Exposure assessment of particulate matter for susceptible populations in Seattle. *Environmental Health Perspectives* 2003; 111:909-918.
- Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. New York: John Wiley, 1987.
- Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; 88:125-134.

- Lumley T, Sheppard L. Assessing seasonal confounding and model selection bias in air pollution epidemiology using positive and negative control analysis. *Epidemiology* 2000; 11:705-717.
- Mar TF, Jansen K, Shepherd K, Lumley T, Larson TV, Koenig JQ. Exhaled nitric oxide in children with asthma and short-term PM2.5 exposure in Seattle. *Environmental Health Perspectives* 2005; 113: 1791-4.
- Palta M, Yao T-J. Analysis of longitudinal data with unmeasured confounders. *Biometrics* 1991; 47:1355-1369.
- Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics, Part B- Simulation and Computation* 1994; 23:939-951.
- Robins JM, Rotnitzky A, Zhou LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; 90:106-121.
- Royall RM. Model robust inference using maximum likelihood estimators. *International Statistical Review* 2005; 54:221-226.
- Scharfstein D, Rotnitzky A, Robins JM. Adjusting for non-ignorable dropout using semiparametric non-response models (with discussion). *Journal of the American Statistical Association* 1999; 94:1096-1120.

- Schildcrout JS, Heagerty P. Regression analysis of longitudinal binary data with time-dependent environmental covariates: Bias and efficiency. *University of Washington Biostatistics Working Paper Series* 2004.
- Schwartz J. The distributed lag between air pollution and daily deaths. *Epidemiology* 2000; 11(3):320-326.
- Sheppard L. Ecologic study design. In: *Encyclopedia of Environmetrics*. New York: John Wiley and Sons: 673-705, 2002.
- Sheppard L, Prentice RL, Rossing MA. Design considerations for estimation of exposure effects on disease risk, Using aggregate data studies. *Statistics in Medicine* 1996; 15:1849-1858.
- White H. Maximum likelihood estimation of misspecified models. *Econometrics* 1982; 50:1-25.
- Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics* 1989; 45:939-955.
- Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of right censoring by modeling the censoring process. *Biometrics* 2005; 44:175-188.
- Yu OC, Sheppard L, Lumley T, Koenig JQ, Shapiro GG. Effects of ambient air pollution on symptoms of asthma in Seattle-area children enrolled in the CAMP study. *Environmental Health Perspectives* 2000; 108(12):1209-1214.

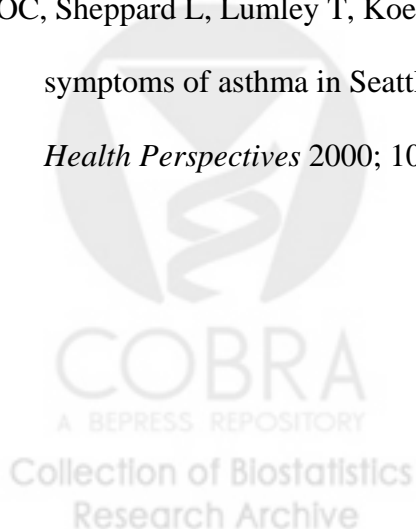


Table 1 The Seattle panel study. A total of 19 subjects were studied for one to three 10-day monitoring sessions each (mean 1.7), for a total of 330 observations.

(a) Subject Characteristics

	N ^a	Mean (SD)
Age (years)	19	9.01 (2.01)
BMI	19	19.78 (3.27)

^a out of 19 total subjects

(b) Time-varying variables

	N ^b	Overall mean (SD)	Within-subject, within-session SD ^c
Relative humidity (%)	313	79.03 (10.24)	7.79
Temperature (F)	313	44.32 (6.27)	3.63
Indoor PM _{2.5} (µg/m ³)	296	9.08 (5.87)	3.96
Overall well-being	314	0.59 ^c	0.60 (0.30, 1.00) ^d
eNO (ppb)	288	15.74 (9.98)	9.59

^b out of 330 total observations

^c proportion “positive”

^d mean and interquartile range of the within-subject, within-session proportions “positive”

^e summarizing the within-subject, within-session component of this variable, which has mean zero by definition; see Section 4.1 for motivation.

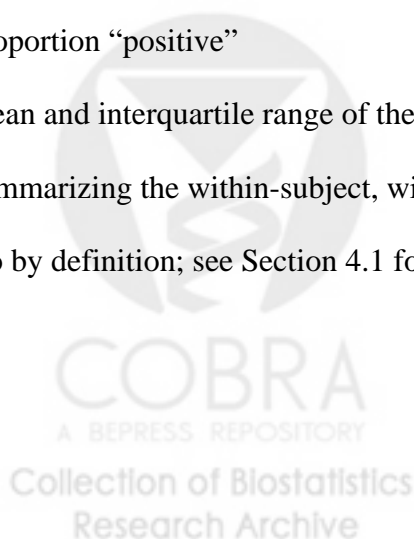


Table 2 The Seattle panel study: the association between indoor PM_{2.5} levels and overall well-being, for three different types of longitudinal models. Indoor PM_{2.5} is modeled using a spline with one knot at the median (7.46 µg/m³). The odds ratio (OR) relates to a 10 µg/m³ increase in indoor PM_{2.5}. All models include age, BMI, relative humidity, and temperature.

	OR below median	95% CI	p- value	OR above median	95% CI	p- value
Marginal model ^a	0.16	(0.01, 2.37)	0.180	2.42	(0.67, 8.80)	0.179
Random intercept model	0.22	(0.01, 6.89)	0.389	5.96	(1.71, 20.73)	0.005
Transitional model ^a	0.22	(0.02, 2.64)	0.233	2.53	(1.07, 5.95)	0.034

^aEstimated using GEE with independent working correlation, and robust standard errors are reported.



Table 3 The Seattle panel study: the association between indoor PM_{2.5} levels and eNO, for three different types of longitudinal models. The coefficient relates to a 10 µg/m³ increase in indoor PM_{2.5}, and all models include age, BMI, relative humidity, and temperature.

	Coefficient	95% CI	p- value
Marginal model ^a	4.15	(1.06, 7.24)	0.008
Random intercept model	4.10	(1.89, 6.32)	< 0.001
Transitional model ^a	3.28	(1.00, 5.57)	0.005

^aEstimated using GEE with independent working correlation, and robust standard errors are reported.



Table 4 Missing data in the Seattle panel study. The number of observations, variables missing, reason given, and type of missingness are listed.

No. of obs. missing	Variables missing	Reason given	Type of missingness
55	eNO, indoor PM _{2.5} , and/or relative humidity	Equipment problem	MCAR
5	eNO	Meal w/in 1 hour of measurement	MAR
3	Overall well-being	Subject left blank on questionnaire	MAR
5	All	Child not home	MAR



Table 5 The Seattle panel study: the marginal association between indoor PM_{2.5} levels and eNO, when exposure is partitioned into between-subject; within-subject, between-session; and within-subject, within-session components. The coefficient relates to a 10 µg/m³ increase in indoor PM_{2.5}, and the model also includes age, BMI, relative humidity, and temperature. Independent working correlation was used, and robust standard errors are reported.

	Coefficient	95% CI	p- value
Indoor PM _{2.5} , Between-subject	3.82	(0.59, 7.05)	0.021
Indoor PM _{2.5} , Within-subject, between-session	10.77	(-7.79, 29.33)	0.255
Indoor PM _{2.5} , Within-subject, within-session	3.90	(0.74, 7.06)	0.016



Titles and Legends to Figures

Figure 1 The Seattle panel study: indoor $PM_{2.5}$ and eNO as a function of time. Lines connect observations within the same subject-session.

Figure 2 The Seattle panel study: eNO residuals as a function of indoor $PM_{2.5}$ residuals. Residuals are generated from linear models with covariates age, BMI, relative humidity, and temperature. A smoothed curve is overlaid.

Figure 3 The Seattle panel study: eNO residuals plotted against eNO residuals, lagged one day. Residuals come from a linear model with covariates: indoor $PM_{2.5}$, age, BMI, relative humidity, and temperature. A smoothed curve is overlaid.

Figure 4 The Seattle panel study: overall well-being residuals as a function of indoor $PM_{2.5}$ residuals. Residuals are generated from a logistic model for overall well-being, and a linear model for indoor $PM_{2.5}$, each with covariates age, BMI, relative humidity, and temperature. A smoothed curve is overlaid.

Figure 5 Sample variogram for the eNO residuals in the Seattle data, as a function of time separation, in days. A smooth curve is overlaid (solid line). The dotted line represents the total sample variance for the eNO residuals. Residuals come from a linear model including the three indoor $PM_{2.5}$ exposure components, age, BMI, relative humidity, and temperature.

Figure 6 The lorelogram for overall well-being in the Seattle data: log-odds ratios for pairs of observations in the same subject-session, as a function of time separation, in days.

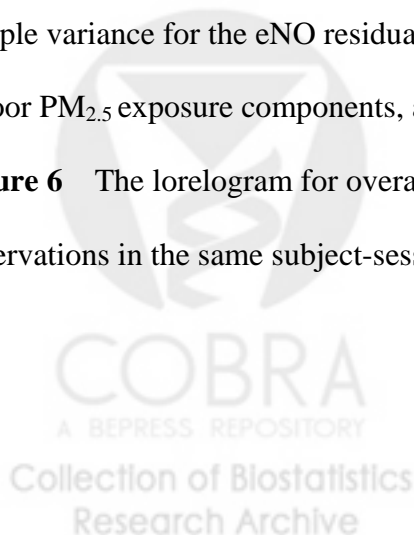


Figure 1

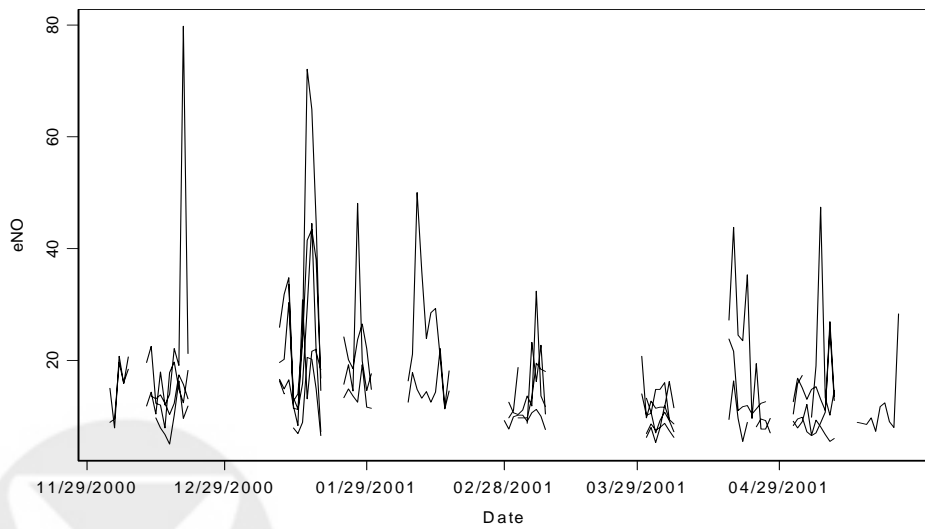
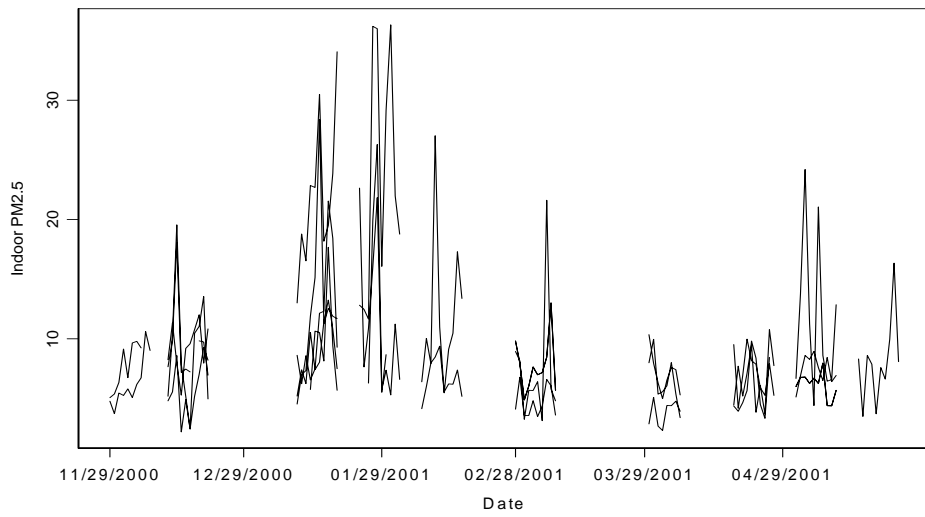


Figure 2

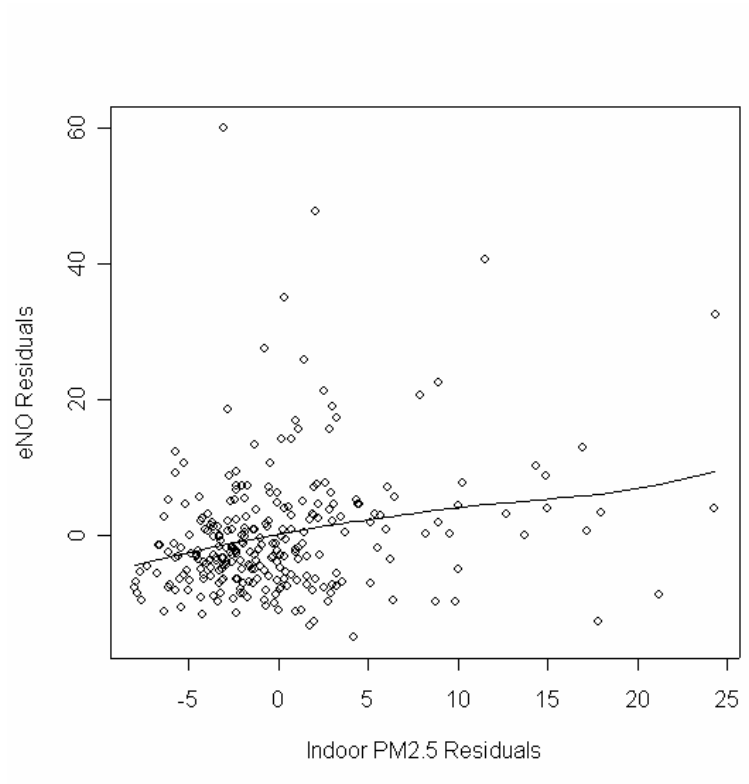


Figure 3

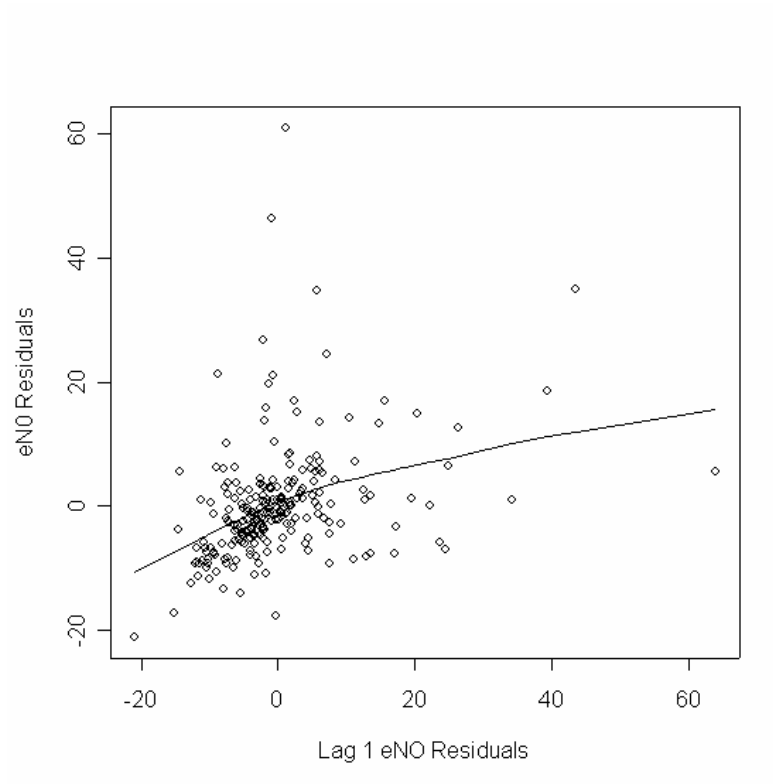


Figure 4

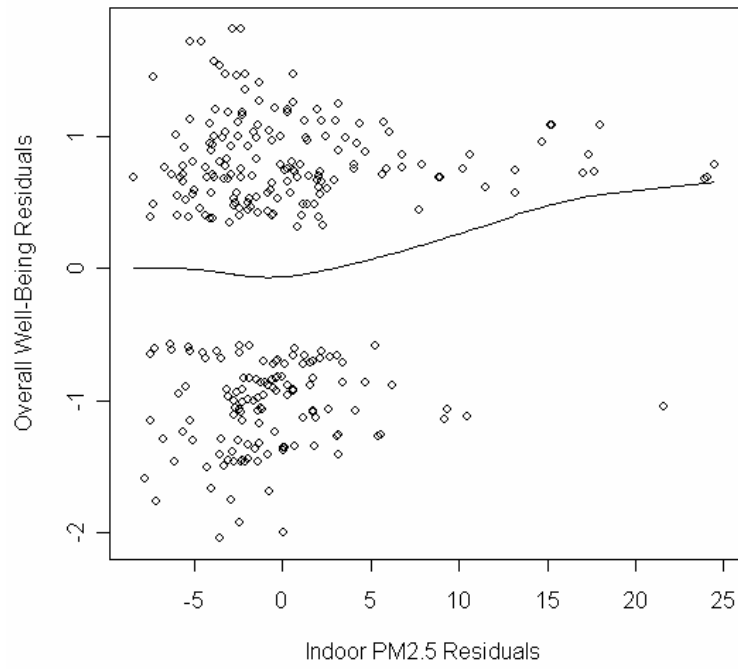


Figure 5

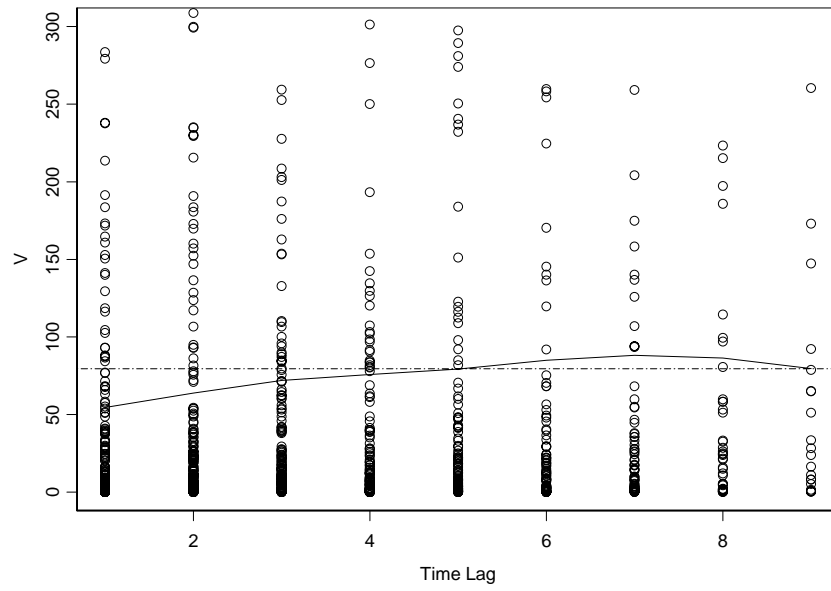


Figure 6

