3-13-2006

ON THE EQUIVALENCE OF CASE-CROSSOVER AND TIME SERIES METHODS IN ENVIRONMENTAL EPIDEMIOLOGY

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

Scott L. Zeger
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, szeger@jhsph.edu

On the equivalence of case-crossover and time series methods in environmental epidemiology

YUN LU*

Department of Biostatistics, Johns Hopkins Bloomberg School of Public health, 615 N. Wolfe Street,Baltimore, MD 21205-2179, USA

ylu@jhsph.edu, 1-410-614-5086(phone), 1-410-955-0958 (fax)

SCOTT L. ZEGER

Department of Biostatistics, Johns Hopkins Bloomberg School of Public health, 615 N. Wolfe Street,Baltimore, MD 21205-2179, USA

ABSTRACT

Time series and case-crossover methods are often viewed as competing alternatives in environmental epidemiologic studies. Several recent studies have compared the time series and case-crossover methods. In this paper, we show that case-crossover using conditional logistic regression is a special case of time series analysis when there is a common exposure such as in air pollution studies. This equivalence provides computational convenience for case-crossover analyses and a better understanding of time series models. Time series log-linear regression accounts for over-dispersion of the Poisson variance, while case-crossover analyses typically do not. This equivalence

also permits model checking for case-crossover data using standard log-linear model diagnostics.

Keywords: air pollution; case-crossover design; environmental epidemiology; log-linear model; Poisson regression; time series

1 INTRODUCTION

The case-crossover design was introduced in epidemiology fifteen years ago as a method for studying the effects of a risk factor on a health event using only cases (Maclure, 1991). The idea is to compare a case's exposure immediately prior or during the case-defining event with that same person's exposure at otherwise similar "reference" times. Each person's exposure values comprise a matched set with a single case exposure during the event interval. Conditional logistic regression is typically used to estimate an odds ratio as the measure of association (e.g. Bateson and Schwartz, 1999). The case-crossover design is attractive because it only involves cases and each case is compared to himself, thereby controlling for time-invariant personal factors.

Maclure originally proposed that only the intervals before the one in which the event occurred can be used for reference (Maclure, 1991). Greenland (1996) and Navidi (1998) pointed out that this choice produces a biased odds ratio estimate in the presence of a secular trend. As an alternative, Navidi (1998) proposed the "full-stratum" design such that all intervals other than the event interval can be used for reference. Bateson and Schwartz (1999) suggested a "symmetrical bidirectional" reference window that uses control intervals equidistant shortly before and after the event to control for bias induced by long-term and seasonal trends. Lumley and Levy (2000) and Janes *et al.* (2005b) have shown that in the bidirectional design, conditional logistic regression gives "overlap" biased estimates of the odds ratio because the reference windows are not chosen independently of the event time. They favor the use of pre-specified reference windows or "time-stratified designs".

The substantial statistical interest in case cross-over designs reflects its common application in many subspecialties of epidemiology including cardiovascular disease

(e.g. Koton *et al.*, 2004), HIV (e.g. Schneider *et al.*, 2005), accidents (Hagel *et al.*, 2005) and health service quality assessment (e.g. Polevoi *et al.*, 2005). The number of papers per year that include "case-crossover" in their title or keywords as identified in a Science Citation Index (SCI) search has grown more than ten-fold since 1993.

This work is motivated by our group's research on the effects of air pollution on morbidity or mortality where the case-crossover method is especially popular. (e.g. Dominici *et al.*, 2004, 2006; Wellenius *et al.*, 2005; Zanobetti and Schwartz, 2005). Case-crossover methods are used to estimate the relative rate of events per unit increase in exposure, controlling for potential confounding variables through matching. For example, Zanobetti and Schwartz (2005) applied conditional logistic regression to data from each of 21 regions to study the relative risk of emergency room admission for myocardial infarction associated with PM10 exposure (particulate matter 10 microns or smaller in aerodynamic diameter). This application and many others like it are characterized by the fact that the exposure for a given day is assumed to be the same for all persons.

An alternative approach to the analysis of daily exposure and case only data is time series analysis (e.g. Diggle, 1990). Here, log-linear regression models express the expected total number of events on each day as a function of the exposure level and potential confounding variables. In time series analyses of air pollution, smooth functions of time and weather are the main confounders. The smooth function of time is typically modelled using a flexible parametric or non-parametric curve to represent longer-term trends in the outcome due to changes in the population, its health behaviors and services and to represent seasonality. Zeger *et al.* (2006) and Bell *et al.* (2004) present overviews of time series methods in general and with application to air pollution epidemiology specifically.

The current understanding is that case-crossover methods control for potential confounding "by design" while time series methods control by modeling (Bateson and Schwartz, 2001; Janes *et al.*, 2005b; Mittleman, 2005; Zanobetti and Schwartz, 2005). The case-crossover idea is to control for personal variation in baseline risk by matching each case with himself and to control for time-varying confounders by matching each event period with reference periods that have the same value of time-varying confounders. In this way, case-crossover analysis apparently avoids the need to control through statistical modelling.

The relative merits of time series and case-crossover studies have been discussed by several recent papers in the environmental epidemiology literature. For example, Checkoway *et al.* (2000) selected the case-crossover approach as an alternative to time series methods in order to make causal inferences about air pollution effects. Bateson and Schwartz (1999; 2001) demonstrated that strong confounding by seasonality could be controlled by design in the case-control approach.

In this paper, we demonstrate that when exposure is common to the cohort at each time as in air pollution studies, the case-crossover approach is an application of log-linear time series analysis rather than an alternative approach. This equivalence has previously been noted in special cases by Levy *et al.* (2001) and by Janes *et al.* (2005a). We show how the choice of reference intervals in the case-crossover design is synonymous with the choice of estimator for the confounding function of time in the time series analysis. Given this correspondence, we offer an alternate perspective on bias of inferences from case-crossover designs. We show that inferences from case-crossover designs based upon conditional logistic regression do not account for over-dispersion as is routinely done in time series analyses. The connection of case-crossover and time series analyses also sheds some new light on the time series

applications.

Section 2 reviews the general framework for case-crossover and time series methods and demonstrates their equivalence and connects choice of reference window with an estimator of the smooth function of time in the time series case. Section 3 illustrates the bias in case-crossover designs and time series estimators when the smooth function of time is misspecified. Section 4 performs a simple data analysis with model checking and is followed by discussion.

2 GENERAL FRAMEWORK

Let X_{it} be the exposure for person i in interval t, $t = 1, \ldots, T$ and let Y_{it} indicate whether subject i has the event in interval t (1 - event; 0 - not). Assume that the outcome $Y_{it} = 1$ is rare and that the probability that subject i fails in interval t is given by the relative risk model:

$$\lambda_i(t, X_{it}) = \lambda_{0it} \exp(\beta X_{it}) = \lambda_{0i} \exp(\beta X_{it} + \gamma_{it}). \tag{2.1}$$

Each subject is assumed to have his own baseline risk λ_{0it} at time t consisting of two parts; λ_{0i} is a constant frailty for person i and $\exp(\gamma_{it})$ is the effect of unmeasured time-varying factors on his risk. The exposure $X_{it}$is assumed to have a common effect on each individual as quantified by the log relative risk β.

For air pollution and other similar studies, the population is assumed to have common exposure during each interval so that $X_{it} = X_t$.

2.1 Time series analysis

Denote the population from which cases arise by $\mathcal{I}$, the observed number of events Y_t in interval t is just $Y_t = \sum_{i \in \mathcal{I}} Y_{it}$. The expected number of events is given by the sum over the entire population of the individual risks:

$$\mu_t = \sum_{i \in \mathcal{I}} \lambda_i(t, X_t) = \exp(\beta X_t) \sum_{i \in \mathcal{I}} \lambda_{0it} = \exp(\beta X_t + S_t), \tag{2.2}$$

where $\exp(S_t) = \sum_{i \in \mathcal{I}} \lambda_{0it}$. The target of inference is the regression coefficient β, the common log relative rate of the event per unit change in the exposure. S_t is the log of the total population risk on day t. The total risk integrates across the population the individual baseline risks and behaviors such as exercise, smoking and health care seeking. It also represents factors that affect the population as a whole such as influenza epidemics or improved medical services. In time series analysis, S_t is assumed to be a smooth function of time and is modeled with parametric or non-parametric curves such as regression or smoothing splines (e.g. Kelsall *et al.*, 1997). Because S_t is not the scientific focus, most time series investigators examine the sensitivity of inferences about the exposure relative risk β to the choice of model for S_t (e.g. Dominici *et al.*, 2004).

To estimate the log relative risk parameter β, we assume Y_t follows a log linear model with mean $E(Y_t) = \mu_t$ and $Var(Y_t) = \phi\mu_t$. The following Poisson estimating equation is solved for the estimate $\hat{\beta}$

$$U(\beta) = \sum_{t=1}^{T} X_t(Y_t - e^{\beta X_t} \exp(\hat{S}_t(\beta))) = \sum_{t=1}^{T} X_t(Y_t - \hat{\mu}_t(\beta)), \tag{2.3}$$

where $\hat{\mu}_t(\beta)$ will depend on the estimate of the nuisance function $\hat{S}_t(\beta)$. We choose the

estimate of β that makes the observed number of events Y_t on each day t on average equal to the model-based predicted value $\hat{\mu}_t(\beta)$. Inferences about β are made robust to the Poisson assumption by allowing the variance of the data to exceed its mean using the method of "quasi-likelihood" or by using a robust variance estimator(Liang and Zeger, 1986; McCullagh and Nelder, 1989; White, 1982; Zeger, 1988).

2.2 Case-crossover design

In the case-crossover approach, the exposure of cases in interval t_i is compared to the exposures from a set of reference periods. We denote the event interval by t_i and the set of reference periods by $W(t_i)$. For example for day 10, $W(10)$ might be $\{8, 9, 10, 11, 12\}$ indicating we use the two days before and after the event day as the reference intervals. The key assumption of a case-crossover design is that the time-varying effect γ_{ij} is constant for all j within the reference window $W(t_i)$.

Conditional on an individual being a case within a pre-specified reference window $W(t_i)$, the probability p_{it_i} that subject i fails at time t_i is

$$\begin{aligned} p_{it_i} &= P(T_i = t_i | \mathbf{X}, W(t_i), \sum_{m=1}^{T} Y_{im} = 1) = \frac{P(T_i = t_i, \sum_{m=1}^{T} Y_{im} = 1 | \mathbf{X}, W(t_i))}{\sum_{j \in W(t_i)} P(T_i = j, \sum_{m=1}^{T} Y_{im} = 1 | \mathbf{X}, W(t_i))} \\ &= \frac{\lambda_i \exp(\beta X_{it_i} + \gamma_{it_i})}{\sum_{j \in W(t_i)} \lambda_i \exp(\beta X_{ij} + \gamma_{ij})} = \frac{\exp(\beta X_{it_i})}{\sum_{j \in W(t_i)} \exp(\beta X_{ij})}, \end{aligned} \tag{2.4}$$

which is free of the time-constant effect λ_i and time-varying effects γ_{ij}.

As Janes *et al.* (2005b) have pointed out, this probability is not correct if the reference window depends on t, e.g. in the symmetric bidirectional design. However, equation (2.4) can still be used to construct an estimating equation for β.

If we assume subjects are independent, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} p_{it_i} = \prod_{i=1}^{n} \left[\frac{\exp(\beta X_{it_i})}{\sum_{j \in W(t_i)} \exp(\beta X_{ij})} \right]. \tag{2.5}$$

The estimating equation for β is

$$U(\beta) = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} \left[X_{it_i} - \sum_{m \in W(t_i)} X_{im} \frac{\exp(\beta X_{im})}{\sum_{j \in W(t_i)} \exp(\beta X_{ij})} \right]. \tag{2.6}$$

This estimating equation is the sum over subjects of the difference between each subject's exposure at the index time t_i and a weighted average of exposures at all times in the reference window $W(t_i)$ (Janes *et al.*, 2005a). By solving (2.6), we estimate β by the value that on average makes the relative-risk weighted average of exposures on reference days equal to the exposure on the event days.

If we assume common exposure, $X_{it} = X_t$, (2.6) can be rewritten as (see appendix I)

$$\begin{aligned} U(\beta) &= \sum_{t=1}^{T} X_t \left[Y_t - e^{\beta X_t} \sum_{m \in \mathcal{R}(t)} \frac{Y_m}{\sum_{j \in W(m)} \exp(\beta X_j)} \right] \\ &= \sum_{t=1}^{T} X_t \left[Y_t - e^{\beta X_t + \hat{S}_t} \right] = \sum_{t=1}^{T} X_t \left(Y_t - \hat{\mu}_t^{(cc)} \right). \end{aligned} \tag{2.7}$$

Here, $\mathcal{R}(t)$ is the set of days containing day t in their reference window. For the symmetric bidirectional and time-stratified designs, this set is identical to the reference set for day t itself, that is $\mathcal{R}(t) = W(t)$. But this is not true for other designs so the distinction between $\mathcal{R}(t)$ and $W(t)$ is essential.

In equation (2.7), $\hat{S}_t$ is the weighted average of number of events across days that have day t in their reference window. The weight for Y_m is the probability of having event on day t given the reference window $W(m)$, where day m has day t in its reference window $W(m)$.

The case-crossover equation (2.7) is a special case of the time series equation (2.3) in which S_t is estimated by a weighted average of the observed numbers of events for those intervals m that include interval t in their reference windows. The weights are determined by the conditional probabilities that an event occurs in t given it occurs within the window.

Two special cases are worth considering further: time-stratified design (TSD) and symmetric bidirectional design (SBD). For TSD, time is a priori divided into strata $s(t) = 1, \ldots, S$. The reference window for day t is the set of days in its stratum (Lumley and Levy, 2000). Levy *et al.* (2001) previously pointed out that the time-stratified case-crossover design leads to the same estimate as obtained from a Poisson regression with dummy variables indicating the strata. The score equation can be written as

$$\sum_{i=1}^{n} U_i(\beta) = \sum_{t=1}^{T} X_t \left[Y_t - e^{\beta X_t} \frac{\sum_{m \in s(t)} Y_m}{\sum_{j \in s(t)} \exp(\beta X_j)} \right] = \sum_{t=1}^{T} X_t \left(Y_t - \hat{\mu}_t^{(a)} \right), \qquad (2.8)$$

where $\hat{\mu}_t^{(a)} = e^{\beta X_t} \frac{\sum_{t \in g(t)} Y_t}{\sum_{t \in g(t)} \exp(\beta X_t)}$ is the expected number of events on day t. Note that $\exp(\hat{S}_{s(t)}) = \frac{\sum_{t \in g(t)} Y_t}{\sum_{t \in g(t)} \exp(\beta X_t)}$ is the maximum likelihood estimator (MLE) of $\exp(S_{s(t)})$. The smooth function of time is assumed to be a step function with a separate level of population baseline risk for each pre-specified stratum. Whether to expect the total baseline risk to change abruptly at each stratum boundary as assumed in this design is a question specific to each application. However, if it does not, assuming S_t is a

step function may introduce bias in the estimator or the pollution log relative risk β.

In the symmetric bidirectional design, symmetric control days close to the event time are used. As the simplest example, define the controls as the days immediately before and after the event day. Then the score equation can be written as

$$\sum_{i=1}^{n} U_i(\beta) = \sum_{t=1}^{T} X_t \left[Y_t - e^{\beta X_t} \sum_{m=t-1,t,t+1} \frac{Y_m}{e^{\beta X_{m-1}} + e^{\beta X_m} + e^{\beta X_{m+1}}} \right] = \sum_{t=1}^{T} X_t \left(Y_t - \hat{\mu}_t^{(b)} \right),$$

This is equivalent to using a locally weighted running-mean smoother to estimate S_t in time series analysis.

3 BIAS IN CASE-CROSSOVER OR TIME SERIES METHODS

There are several papers that investigate the bias associated with case-crossover designs using conditional logistic regression (Lumley and Levy, 2000; Janes *et al.*, 2005a; 2005b). However, each one assumes that the time-varying confounders γ_{it} do not vary within the reference window. In this section, we give a standard expression for the asymptotic bias from the perspective that a true S_t exists and it can be mismodelled with either a case-crossover or time series approach. We then evaluate the actual bias in an air pollution mortality example via simulation.

Assume the true model is $\log \mu_t = X_t\beta + S_t$. The expectation of the estimating

equation for the case-crossover design can be written as

$$\begin{aligned} E[U(\beta)] &= \sum_{t=1}^{T} X_t e^{\beta X_t} \left[\exp(S_t) - \sum_{m \in \mathcal{R}(t)} \frac{\exp(\beta X_m + S_m)}{\sum_{j \in W(m)} \exp(\beta X_j)} \right] \\ &= \sum_{t=1}^{T} X_t e^{\beta X_t} \left[\exp(S_t) - E[\exp(\hat{S}_t(\beta))] \right] \\ &= \sum_{t=1}^{T} X_t e^{\beta X_t} \Delta e^{S_t}(\beta), \end{aligned} \tag{3.1}$$

where $\Delta e^{S_t}(\beta)$ is the difference between true $\exp(S_t)$ and the expectation of the estimated $\exp(\hat{S}_t(\beta))$. The estimating equation will be unbiased if $E[\exp(\hat{S}_t(\beta))] = \exp(S_t)$ for all t, i.e. $\Delta e^{S_t}(\beta) = 0$ for all t, and its solution $\hat{\beta}$ will be asymptotically unbiased (given regularity conditions).

For case-crossover design, we know that

$$\frac{\partial U(\beta)}{\partial \beta} = -\sum_{t=1}^{T} X_t e^{\beta X_t} \sum_{m \in \mathcal{R}(t)} Y_m \left[\frac{X_t}{\sum_{j \in W(m)} \exp(\beta X_j)} - \frac{\sum_{j \in W(m)} X_j \exp(\beta X_j)}{\left[\sum_{j \in W(m)} \exp(\beta X_j) \right]^2} \right]. \tag{3.2}$$

We can obtain

$$E\left[\frac{\partial U(\beta)}{\partial \beta} \right] = -\sum_{t=1}^{T} X_t e^{\beta X_t} \sum_{m \in \mathcal{R}(t)} e^{\beta X_m + S_m} \left[\frac{X_t}{\sum_{j \in W(m)} \exp(\beta X_j)} - \frac{\sum_{j \in W(m)} X_j \exp(\beta X_j)}{\left[\sum_{j \in W(m)} \exp(\beta X_j) \right]^2} \right]. \tag{3.3}$$

The asymptotic bias in $\hat{\beta}$ is given by

$$-\left\{E\left[\frac{\partial U(\beta)}{\partial \beta}\right]\right\}_{\beta}^{-1} E\left[U(\beta)\right] = -\left\{E\left[\frac{\partial U(\beta)}{\partial \beta}\right]\right\}_{\beta}^{-1} \sum_{t=1}^{T} X_t e^{\beta X_t} \Delta e^{S_t}(\beta), \qquad (3.4)$$

which is a function of $\Delta e^{S_t}(\beta)$. The finite sample bias can be evaluated using simulation. The finite sample bias and the asymptotic bias will be very similar when the number of days in the data set are large enough.

3.1 Simulation study

We report a brief simulation study to quantify the bias in $\hat{\beta}$ for an air-pollution-mortality example. We use the daily average PM10 in Chicago from January 1st, 1995 to December 31st, 1996 as the exposure X_t. We simulated Y_t, daily mortality, from a Poisson(μ_t), where $\log \mu_t = \beta X_t + S_t$. We let β be 0, 1, 2, and 5 percent change in daily mortality per 10 $\mu g/m^3$ increase in PM10 and use the following true functions S_t. Data and software for this simulation are available at http://www.ihapss.jhsph.edu/data/data.htm.

Scenario A: S_t is constant over time, S_t=4.10, which equals to the natural logarithm of mean daily mortality for people 75 years and older in Chicago for our two years of exposure.

Scenario B: S_t is a combination of a linear function of time and a cosine function with period equaling one year ($S_t = 4.10 + 0.0001(t - 365) + 0.1\cos(2\pi t/365)$). Here we assume that the base line risk increases over time, and there exists a seasonal trend that peaks in winter.

Scenario C: Scenario C is Scenario B plus a day of the week effect $S_{\text{dow},t}$ for Monday through Sunday: 0, 0.002, 0.01, 0.008, 0.005, 0.008, and -0.005, respectively.

Four models were estimated for each scenario.

Model 1: the symmetric bidirectional case-crossover design (SBD) using 14 and 7 days before and after event day as control days.

Model 2: a time-stratified case-crossover design (TSD) using the days with the same day of the week in the same month and year of the event day as control days.

Model 3: time series with $\log \mu_t = \beta X_t + dow_t \times ns(t, df = 12 \text{ per year})$, where ns is natural spline, and dow_t is the 7 level factor for day of the week. Model 3 has similar degrees of freedom as Model 2 but allows the day of week effect to vary as a smooth function over time.

Model 4: time series with $\log \mu_t = \beta X_t + dow_t \times ns(t, df = 4 \text{ per year})$. Model 4 has the same form as Model 3 with less degrees of freedom in smoothing than the other three models.

In this simulation study, the exposures are considered fixed. We used 1000 replicates. The estimated bias and ratio of mean squared error (MSE) relative to Model 4 were reported.

The standard deviations for the estimates are roughly constant across the 3 assumed scenarios and 4 values of β for the same model. The order of the standard deviations for $\hat{\beta}$ are: Model 4 $<$ Model 1$<$ Model 2 $<$ Model 3. The biases for $\hat{\beta}$ vary across scenarios and β values, but the magnitude of the bias is usually much smaller than the standard deviation, hence the standard deviation dominates the MSE. Model

4 has the smallest MSE (Table 1).

The bias results are shown in Table 2. For Scenario A, the symmetric bidirectional case-crossover method (SBD) performs the worst, while the other three models have similar bias. This demonstrates the previous finding that SBD suffers overlap bias (Janes *et al.*, 2005a, 2005b; Levy *et al.*, 2001; Lumley and Levy, 2000). TSD seems to have slightly smaller bias than the two time series models when the true S_t is constant, since the assumption that S_t is a step function is correct. For Scenario B and C, the time series method have much smaller bias than the case-crossover methods. When the true S_t has a trend, seasonality, and day of the week effects, the case-crossover methods can not capture the smoothness of S_t, producing some bias.

4 DATA ANALYSIS

We illustrate the connection of the case-crossover and time series methods with an analysis of mortality for persons over 75 and PM10 data from Chicago for the two years 1995-6. These data are typical of the air pollution time series problems that motivated the use of case-crossover designs. The data were analyzed using Models 2 and Model 4 as defined in the simulation study. Temperature would be included in an actual time series analysis. It is omitted here to demonstrate the value of a time series formulation of the case-crossover analyses.

The estimates of β are 1.09 (SE 0.35) and 0.69 (SE 0.45) for Model 2 (case-crossover)and Model 4 (time series), respectively. Note that for Model 2, the conditional logistic regression does not take into account of the over-dispersion in the Poisson variance.

To empirically demonstrate the equivalence of case-crossover and time series method proven here, we have fit Model 2 using both the conditional logistic regression and the Poisson regression programmed in R. The estimated β coefficients agree to 6 significant digits. The variance estimates are different if we allow over-dispersion in the Poisson regression. The estimated over-dispersion parameter of 1.92 for the case-crossover Model 2 indicates that there is greater variation in the numbers of deaths within matching strata than can be explained by the model. Much of this could be explained by including a non-linear function of temperature.

An important idea is to use the time series formulation of the case-crossover analysis and perform model checking using standard log-linear model diagnostics. Figure 1 illustrates the predicted daily mortality for Mondays, where Model 2 assumes a step function for S_t while Model 4 uses a natural spline.

Figure 2 shows the Dffits of Model 2 and Model 4. The Dffits statistic is a scaled measure of the change in the predicted value for the ith observation when it is omitted from the regression. Large absolute values of Dffits indicate influential observations. The top two graphs in Figure 2 indicate that the Dffits are mostly in the range of $(-1.5, 1.5)$ for Model 2 and $(-1, 1)$ for Model 4. There are several points with high Dffits for both models. We can set aside the influential points and refit both models, then check the Dffits again. After two rounds of checking, the influential points are identified as July 14th through 18th in 1995, when high daily mortality occurred due to the unusually high temperatures. After deleting these points, the estimates of β are 0.87 (SE 0.35) and 0.79 (SE 0.33) for Model 2 and Model 4, respectively. The estimate for Model 2 changes greater because the original $\hat{S}_t$ for Model 4 had a wider band width than did the Model 2 estimate, and was less influenced by the local pertubation in mortality. After removing highly influential points, the estimated

over-dispersion parameter for Model 2 decreases to 1.08. The bottom two graphs in Figure 2 are the Dffits after refitting the models.

Figure 3 shows the Q-Q plots of the standardized residuals for Models 2 and 4 before and after removing influential points. The standardized residuals $(Y_t - \hat{\mu}_t)/\sqrt{\hat{\mu}_t}$ are approximately Gaussian (0, 1) if the model is correctly specified because of the large Poisson mean. Figure 3 suggests that the standardized residuals are quite skewed to the right before removing influential points, indicating a of violation of the Poisson assumptions. After removing the influential points, the standardized residuals are very close to the Gaussian distribution for both models.

5 SUBJECT-SPECIFIC EXPOSURES

In the previous sections, we emphasized intervals with common exposures across individuals. We often have subject-specific exposures. In this section, we consider the connection between case-crossover and time series methods for subject-specific exposure data.

Let X_{it} be the exposure for person i in interval t, $t = 1, \ldots, T$ and let Y_{it} indicate whether subject i has the event in interval t (1 - event; 0 - not) as described above in the general framework.

We can rewrite equation (2.6) as (see appendix II)

$$\begin{aligned}
U(\beta) &= \sum_{i=1}^{n} U_i(\beta) \\
&= \sum_{i=1}^{n} \left[X_{it_i} - \sum_{m \in W(t_i)} X_{im} \frac{\exp(\beta X_{im})}{\sum_{j \in W(t_i)} \exp(\beta X_{ij})} \right] \\
&= \sum_{t=1}^{T} \left[\left(\sum_{i=1}^{n} Y_{it} \right) \frac{\sum_{i=1}^{n} X_{it} Y_{it}}{\sum_{i=1}^{n} Y_{it}} \right] - \sum_{t=1}^{T} \sum_{i=1}^{n} \left[Y_{it} \sum_{m \in W(t)} X_{im} \frac{\exp(\beta X_{im})}{\sum_{j \in W(t)} \exp(\beta X_{ij})} \right] \\
&= \sum_{t=1}^{T} Y_t \bar{X}_t - \sum_{t=1}^{T} \bar{X}_t \left\{ \sum_{i=1}^{n} \left[Y_{it} \frac{\sum_{m \in W(t)} X_{im} \exp(\beta X_{im})}{\sum_{j \in W(t)} \exp(\beta X_{ij})} / \bar{X}_t \right] \right\} \\
&= \sum_{t=1}^{T} \bar{X}_t \left[Y_t - \sum_{i=1}^{n} \left[Y_{it} \bar{X}_i^{(W(t))}(\beta) / \bar{X}_t \right] \right] \\
&= \sum_{t=1}^{T} \bar{X}_t \left(Y_t - \hat{\mu}_t^{(ccs)} \right), \qquad (5.1)
\end{aligned}$$

where $Y_t = \sum_{i=1}^{n} Y_{it}$, $\bar{X}_t = \frac{\sum_{i=1}^{n} X_{it} Y_{it}}{\sum_{i=1}^{n} Y_{it}}$, $\bar{X}_i^{(W(t))}(\beta) = \frac{\sum_{m \in W(t)} X_{im} \exp(\beta X_{im})}{\sum_{j \in W(t)} \exp(\beta X_{ij})}$. Here, $\bar{X}_t$ is the average exposure for day t, averaged across all subjects whose event is on day t. Here, $\bar{X}_i^{(W(t))}(\beta)$ is the weighted average exposure for subject i within the reference window of day t. We can consider $\hat{\mu}_t^{(ccs)}$ to be the expected number of events on day t given the exposures on days in its reference window and β.

This time series formulation of the case-crossover design with subject-specific exposures can be used to perform model checking similar to what we described in the previous section.

If the Poisson assumption is valid, the standardized residuals $\hat{r}_t = (Y_t - \hat{\mu}_t)/\sqrt{\hat{\mu}_t}$ should have mean 0 and variance 1. The plot of $\hat{r}_t$ vs. $\hat{\mu}_t$ can detect violations from this assumption. If the mean for the Poisson variable Y_t is large, $\hat{r}_t$ will approximately follow a standardized Gaussian distribution that can be checked using a Q-Q plot.

We can calculate the over-dispersion parameter ϕ using $\hat{\phi} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{Y_t - \hat{\mu}_t}{\sqrt{\hat{\mu}_t}}\right)$ and rely upon a robust variance estimator when $\hat{\phi} > 1$. We can check influential data points using Dffits as described before.

6 DISCUSSION

This paper has shown that the conditional logistic regression estimating equation used to obtain the case-crossover estimate of relative risk is a special case of the time series log-linear model estimating equation when exposure is common across subjects in each interval. Time series and case-crossover analyses simply offer different parameterizations for S_t.

The time-stratified case-crossover design is equivalent to Poisson regression with indicator variables for strata (Levy *et al.*, 2001). The estimated smooth function of time $\hat{S}_t$ is assumed to be a step function with different levels of total population baseline risk for each stratum. The symmetric bidirectional case-crossover design is equivalent to Poisson regression using a locally weighted running-mean smoother for S_t.

The equivalence of the case-crossover and time series methods improves our understanding of both methods and provides computational convenience. Most case-crossover analyses use conditional logistic regression (CLR) for estimation. When the number of time intervals and the number of controls for each case are large (e.g. full-stratum design), standard CLR is computationally inefficient and Poisson regression software is computationally less expensive.

Each case-crossover design corresponds to a model (or estimator) for S_t. The

equivalence of case-crossover and time series methods permits model checking for case-crossover data using standard log-linear model diagnostics tools (McCullagh and Nelder, 1989).

Despite the equivalence of estimates from time series and case-crossover analyses, they can give different standard errors. This is because time series analysis allows for over-dispersion of the Poisson variance, while case-crossover design uses the exact Poisson variance to calculate the standard error. In many applications, the Poisson assumption is not valid.

This connection also informs our interpretation of time series analysis. For example, in Dominici *et al.* (2004), time series models are used to estimated a PM effect on daily mortality. The degrees of freedom to estimate S_t with a regression spline are allowed to vary nine folds from 2.3 to 21 degrees of freedom per year, yet the standard error of the pollution effect changes little. For matched case-control studies, there is little gain in efficiency when the number of controls per case is beyond roughly four (McCullagh and Nelder, 1989). In a case-crossover design, this corresponds to four control days per event day or equivalently 90 degrees of freedom per year, which is much greater than the entire range included by Dominici *et al.* (2004).

In this paper, we focused on exposures common to all subjects. In many applications of the case-crossover design, exposures vary among subjects. We have shown how our approach extends to the case with subject-specific exposures. Further work on this topic is of interest.

ACKNOWLEDGEMENTS

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.

APPENDIX I

If we again assume common exposure, i.e., $X_{it} = X_t$, the score equation (2.6) can be written as

$$\begin{aligned}
\sum_{i=1}^{n} U_i(\beta) &= \sum_{i=1}^{n} \left[X_{it_i} - \sum_{m \in W(t_i)} X_{im} \frac{\exp(\beta X_{im})}{\sum_{j \in W(t_i)} \exp(\beta X_{ij})} \right] \\
&= \sum_{i=1}^{n} \left[X_{t_i} - \sum_{m \in W(t_i)} X_m \frac{\exp(\beta X_m)}{\sum_{j \in W(t_i)} \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} Y_t \left[X_t - \sum_{m \in W(t)} X_m \frac{\exp(\beta X_m)}{\sum_{j \in W(t)} \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} Y_t X_t - \sum_{t=1}^{T} \left[Y_t \sum_{m \in W(t)} X_m \frac{\exp(\beta X_m)}{\sum_{j \in W(t)} \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} Y_t X_t - \sum_{t=1}^{T} \sum_{m=1}^{T} Y_t X_m \frac{I[m \in W(t)] \exp(\beta X_m)}{\sum_{j=1}^{T} I[j \in W(t)] \exp(\beta X_j)} \\
&= \sum_{t=1}^{T} Y_t X_t - \sum_{m=1}^{T} \sum_{t=1}^{T} Y_t X_m \frac{I[m \in W(t)] \exp(\beta X_m)}{\sum_{j=1}^{T} I[j \in W(t)] \exp(\beta X_j)} \\
&= \sum_{t=1}^{T} Y_t X_t - \sum_{m=1}^{T} \left[X_m \sum_{t=1}^{T} Y_t \frac{I[m \in W(t)] \exp(\beta X_m)}{\sum_{j=1}^{T} I[j \in W(t)] \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} Y_t X_t - \sum_{t=1}^{T} \left[X_t \sum_{m=1}^{T} Y_m \frac{I[t \in W(m)] \exp(\beta X_t)}{\sum_{j=1}^{T} I[j \in W(m)] \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} X_t \left[Y_t - \sum_{m=1}^{T} Y_m \frac{I[t \in W(m)] \exp(\beta X_t)}{\sum_{j=1}^{T} I[j \in W(m)] \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} X_t \left[Y_t - \sum_{m \in \mathcal{R}(t)} Y_m \frac{\exp(\beta X_t)}{\sum_{j \in W(m)} \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} X_t \left[Y_t - e^{\beta X_t} \sum_{m \in \mathcal{R}(t)} \frac{Y_m}{\sum_{j \in W(m)} \exp(\beta X_j)} \right] \\
&= \sum_{t=1}^{T} X_t \left(Y_t - \hat{\mu}_t^{(cc)} \right),
\end{aligned}$$

where $\mathcal{R}(t)$ is the set of days containing day t in their reference window. For each day $m \in \mathcal{R}(t)$, the reference window for day m is $W(m)$, which contains day t. The probability of having event on day t given the reference window $W(m)$ is $\frac{\exp(\beta X_t)}{\sum_{j \in W(m)} \exp(\beta X_j)}$. Given the number of events Y_m on day m, the expected number of events on day t is $Y_m \frac{\exp(\beta X_t)}{\sum_{j \in W(m)} \exp(\beta X_j)}$. Only the days $m \in \mathcal{R}(t)$ will have contribution to $\hat{\mu}_t^{(cc)}$, because we assume m will have similar baseline risk as t. Hence we can consider $\hat{\mu}_t^{(cc)} = \sum_{m \in \mathcal{R}(t)} Y_m \frac{\exp(\beta X_t)}{\sum_{j \in W(m)} \exp(\beta X_j)}$ as the expected number of events on day t, given the number of events Y_m for all $m \in \mathcal{R}(t)$.

APPENDIX II

If we don't assume common exposure, then the score equation can be written as

$$\begin{aligned}
\sum_{i=1}^{n} U_i(\beta) &= \sum_{i=1}^{n}\left[X_{it_i} - \sum_{m\in W(t_i)} X_{im}\frac{\exp(\beta X_{im})}{\sum_{j\in W(t_i)}\exp(\beta X_{ij})}\right] \\
&= \sum_{t=1}^{T}\sum_{i \text{ s.t.} t_i=t}\left[X_{it} - \sum_{m\in W(t)} X_{im}\frac{\exp(\beta X_{im})}{\sum_{j\in W(t)}\exp(\beta X_{ij})}\right] \\
&= \sum_{t=1}^{T}\sum_{i=1}^{n}\left\{Y_{it}\left[X_{it} - \sum_{m\in W(t)} X_{im}\frac{\exp(\beta X_{im})}{\sum_{j\in W(t)}\exp(\beta X_{ij})}\right]\right\} \\
&= \sum_{t=1}^{T}\left[\left(\sum_{i=1}^{n} Y_{it}\right)\frac{\sum_{i=1}^{n} X_{it}Y_{it}}{\sum_{i=1}^{n} Y_{it}}\right] - \sum_{t=1}^{T}\sum_{i=1}^{n}\left[Y_{it}\sum_{m\in W(t)} X_{im}\frac{\exp(\beta X_{im})}{\sum_{j\in W(t)}\exp(\beta X_{ij})}\right] \\
&= \sum_{t=1}^{T} Y_t\bar{X}_t - \sum_{t=1}^{T}\bar{X}_t\left\{\sum_{i=1}^{n}\left[Y_{it}\frac{\sum_{m\in W(t)} X_{im}\exp(\beta X_{im})}{\sum_{j\in W(t)}\exp(\beta X_{ij})}/\bar{X}_t\right]\right\} \\
&= \sum_{t=1}^{T}\bar{X}_t\left[Y_t - \sum_{i=1}^{n}\left[Y_{it}\bar{X}_i^{(W(t))}/\bar{X}_t\right]\right] \\
&= \sum_{t=1}^{T}\bar{X}_t\left(Y_t - \hat{\mu}_t^{(ccs)}\right),
\end{aligned}$$

where $Y_t = \sum_{i=1}^{n} Y_{it}$, $\bar{X}_t = \frac{\sum_{i=1}^{n} X_{it}Y_{it}}{\sum_{i=1}^{n} Y_{it}}$, $\bar{X}_i^{(W(t))} = \frac{\sum_{m\in W(t)} X_{im}\exp(\beta X_{im})}{\sum_{j\in W(t)}\exp(\beta X_{ij})}$.

REFERENCES

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

BATESON, T.F. AND SCHWARTZ, J. (2001). Selection bias and confounding in case-crossover of environmental time-series data. *Epidemiology* **12**, 654-661.

BELL, M.L., SAMET, J.M. AND DOMINICI F. (2004). Time-series studies of particulate matter. *Annual Review of Public Health* **25**, 247-280.

CHECKOWAY, H., LEVY, D., SHEPPARD, L., KAUFMAN, J., KOENIG, J. AND SISCOVICK, D. (2000). A case-crossover analysis of fine particulate matter air pollution and out-of-hospital sudden cardiac arrest. *Health Effects Institute Research Report* **99**.

DIGGLE, P. (1990). *Time Series*. Oxford:Oxford University Press.

DOMINICI, F., MCDERMOTT, A. AND HASTIE, T.J. (2004). Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association* **99**, 938-948.

DOMINICI, F., PENG, R.D, , BELL, M.L., PHAM, L., MCDERMOTT, A., ZEGER, S.L. AND SAMET J.M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Journal of the American Medical Association* **295**,1127-1134.

GREENLAND, S. (1996). Confounding and exposure trends in case-crossover and case-time-control designs. *Epidemiology*, **7**, 231-239.

HAGEL, B.E., PLESS, I.B., GOULET, C., PLATT, R.W. AND ROBITAILLE, Y. (2005). Effectiveness of helmets in skiers and snowboarders: case-control and case crossover study. *British Medical Journal* **330**, 281-283.

JANES, H., SHEPPARD, L. AND LUMLEY, T. (2005a). Case-crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias. *Epidemiology* **16**, 717-726.

JANES, H., SHEPPARD, L. AND LUMLEY, T. (2005b). Overlap bias in the case-crossover design, with application to air pollution exposures. *Statistics in Medicine* **24**, 285-300.

KELSALL, J.E., SAMET, J.M., ZEGER, S.L. AND XU, J. (1997). Air pollution and mortality in Philadelphia, 1974-1988. *American Journal of Epidemiology* **146**, 750-762.

KOTON, S., TANNE, D., BORNSTEIN, N.M. AND GREEN, M.S. (2004). Triggering risk factors for ischemic stroke - A case-crossover study. *Neurology* **63**, 2006-2010.

LEVY, D., LUMLEY, T., SHEPPARD, L., KAUFMAN, J. AND CHECKOWAY, H. (2001). Referent selection in case-crossover analyses of acute health effects of air pollution. *Epidemiology* **12**, 186-192.

LIANG, K.Y. AND ZEGER, S.L. (1986). Longitudinal data-analysis using generalized linear-models. *Biometrika* **73**, 13-22.

LUMLEY, T. AND LEVY, D. (2000). Bias in the case-crossover design: implicaitons for studies of air pollution. *Environmetrics* **11**, 689-704.

MACLURE, M. (1991). The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* **133**, 144-153.

MCCULLAGH, P. AND NELDER, J.A. (1989). *Generalized Linear Models*, 2nd Edition. London:Chapman & Hall/CRC.

MITTLEMAN, M.A. (2005). Optimal referent selection strategies in case-crossover studies-A settled issue. *Epidemiology* **16**, 715-716.

NAVIDI, W. (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics* **54**, 596-605.

POLEVOI, S.K., QUINN, J.V., AND KRAMER, N.R. (2005). Factors associated with patients who leave without being seen. *Academic Emergency Medicine* **12**, 232-236.

SCHNEIDER, M.F., GANGE, S.J., MARGOLICK, J.B., DETELS, R., CHMIEL, J.S., RINALDO, C. AND ARMENIAN H.K. (2005). Application of case-crossover and case-time-control study designs in analyses of time-varying predictors of T-cell homeostasis failure. *Annals of Epidemiology* **15**, 137-144.

WELLENIUS, G.A., BATESON, T.F., MITTLEMAN, M.A. AND SCHWARTZ, J. (2005). Particulate air pollution and the rate of hospitalization for congestive heart failure among Medicare beneficiaries in Pittsburgh, Pennsylvania. *American Journal of Epidemiology* **161**, 1030-1036.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrika* **50**, 1-26.

ZANOBETTI, A. AND SCHWARTZ, J. (2005). The effect of particulate air

pollution on emergency admissions for myocardial infarction: A multicity case-crossover analysis. *Environmental Health Perspectives* **113**, 978-982.

ZEGER, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

ZEGER, S.L., IRIZARRY, R.A. AND PENG, R.D. (2006). On time series analysis of public health and biomedical data. *Annual Review of Public Health* doi:10.1146/annurev.publhealth.26.021304.144517.

Table and Figure Captions

Table 1. Ratio of MSE vs. Model 4 from the simulation study.

Table 2. Bias of the estimate $\times 1000$ from the simulation study.

Figure 1. The predicted Monday mortality vs. calender day in Chicago for persons 75 years and older. Model 2 (denoted as solid line) is a time-stratified case-crossover design (TSD) using the days with the same day of the week in the same month and year of the event day as control days. Model 4 (denoted as dashed line) is a time series method using a natural spline with 8 degrees of freedom.

Figure 2. The Dffits statistics of Model 2 and Model 4 before (top)and after (bottom) removing influential points. Model 2 is a time-stratified case-crossover design (TSD) using the days with the same day of the week in the same month and year of the event day as control days. Model 4 is a time series method using a natural spline with 8 degrees of freedom.

Figure 3. The Q-Q plot of standardized residuals for Model 2 and Model 4 before (top) and after (bottom) removing influential points. Model 2 is a time-stratified case-crossover design (TSD) using the days with the same day of the week in the same month and year of the event day as control days. Model 4 is a time series method using a natural spline with 8 degrees of freedom.

Tables and Figures

Table 1: Ratio of MSE vs. Model 4 from the simulation study.

MSE ratio		Model 1	Model 2	Model 3
Scenario A	β=0	1.1805	1.2520	1.3077
	β=1	1.1502	1.2252	1.2736
	β=2	1.1298	1.2055	1.2370
	β=5	1.2962	1.2106	1.2491
Scenario B	β=0	1.1801	1.3093	1.3564
	β=1	1.2173	1.2457	1.2885
	β=2	1.2155	1.2642	1.3473
	β=5	1.5108	1.3570	1.4450
Scenario C	β=0	1.2480	1.3357	1.3540
	β=1	1.2012	1.3274	1.3223
	β=2	1.2339	1.2792	1.2545
	β=5	1.3389	1.2342	1.2400

Table 2: Bias of the estimate $\times$1000 from the simulation study.

Bias		Model 1	Model 2	Model 3	Model 4
Scenario A	β=0	−12.0	−1.2	3.7	−7.8
	β=1	−38.8	−8.0	−5.0	−11.7
	β=2	−43.6	9.6	11.7	9.7
	β=5	−122.9	−0.4	−3.5	−3.0
Scenario B	β=0	−40.7	−32.1	−2.6	−4.6
	β=1	−85.4	−60.2	−24.3	−24.9
	β=2	−71.3	−18.3	7.5	10.7
	β=5	−147.3	−32.4	5.1	0.5
Scenario C	β=0	−53.6	−40.7	−13.9	−12.6
	β=1	−57.1	−27.5	6.4	4.8
	β=2	−91.3	−40.6	−4.5	−8.2
	β=5	−144.8	−30.0	3.1	−2.9

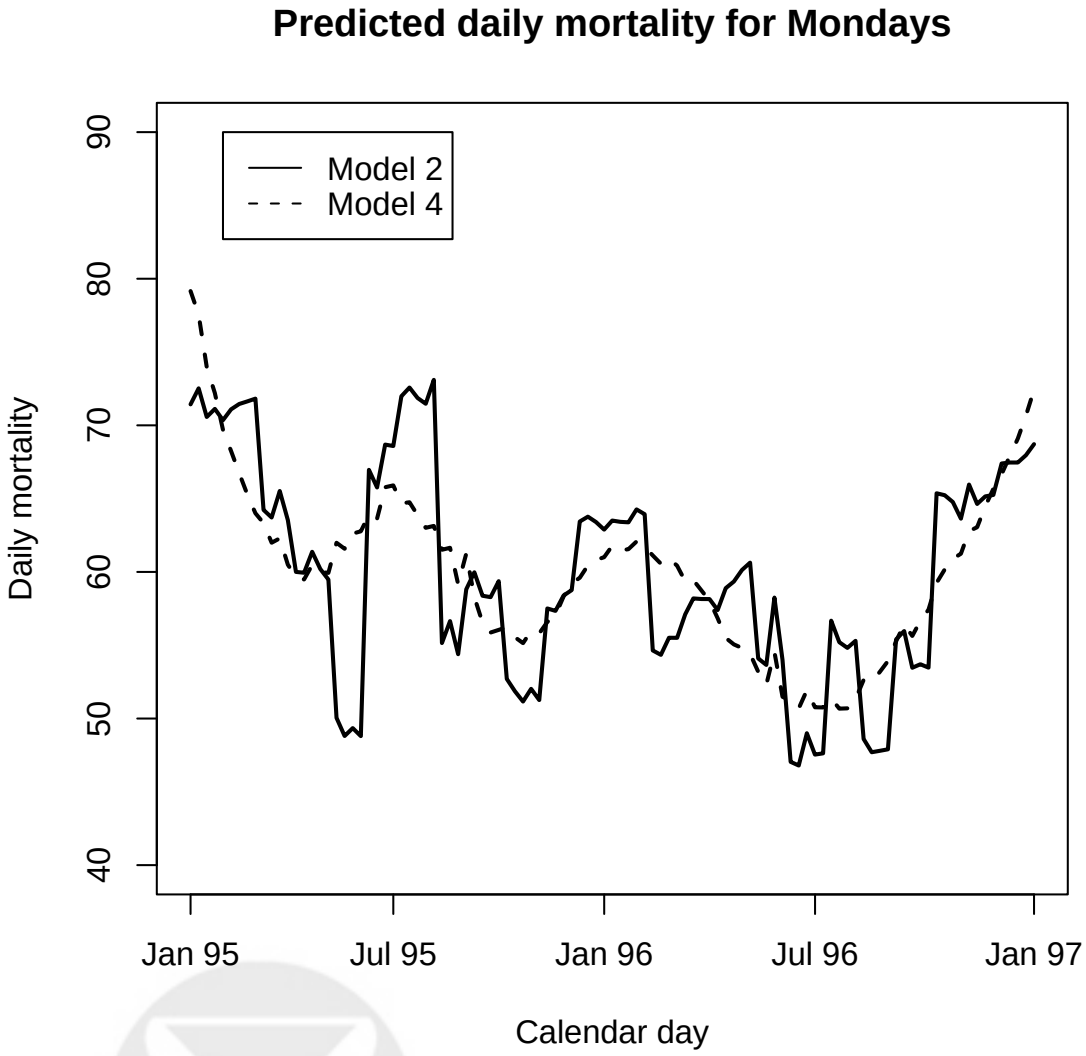

Figure 1:

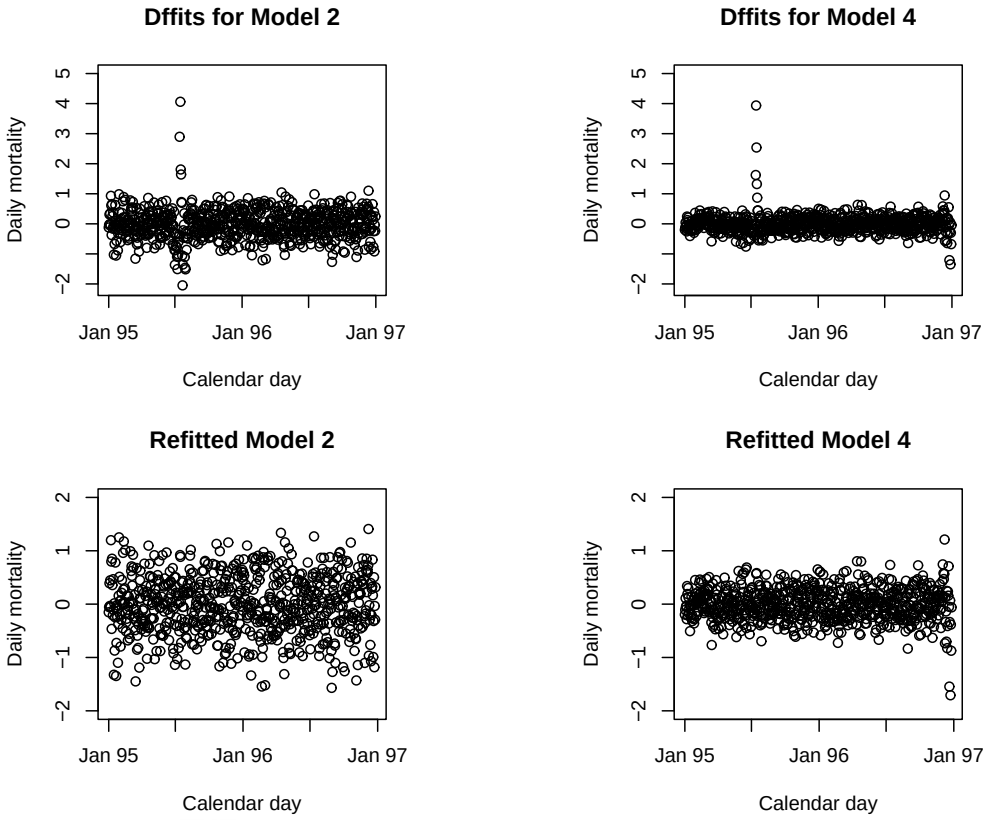

Figure 2:

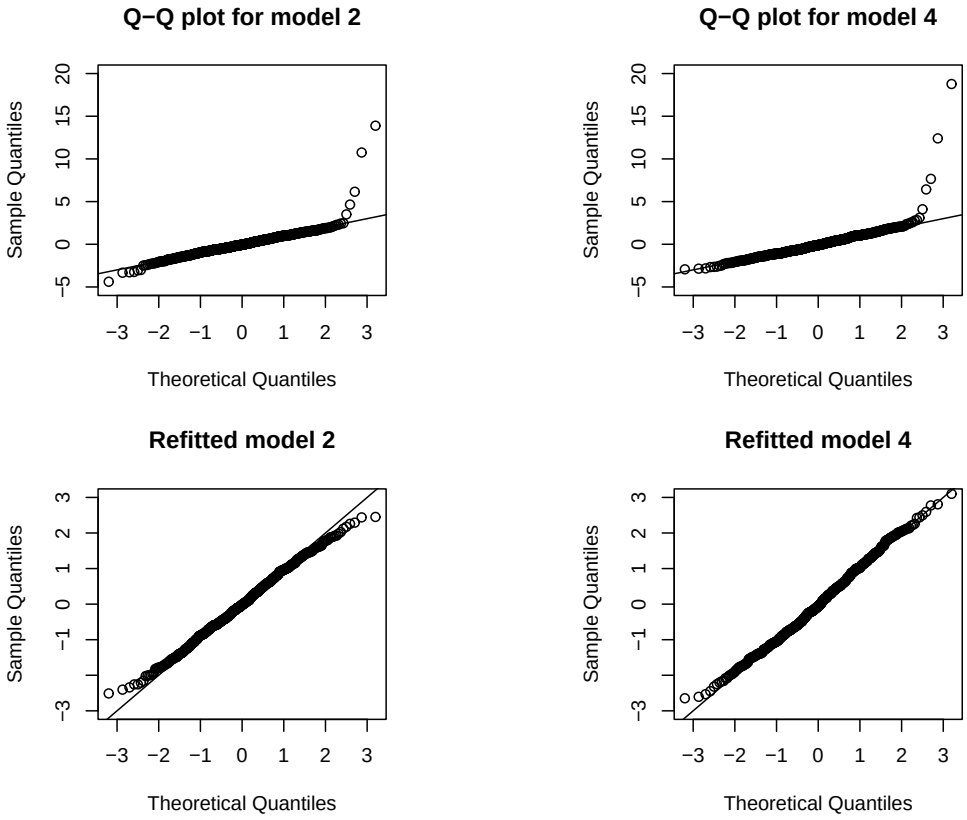

Figure 3: