# Acute effects of ambient exposures

## Time series and case-crossover studies

Brooke Anderson

February 3, 2016

# Overview

# Air pollution studies

- Inform policy choices
- Evaluate effectiveness of interventions or policy changes
- Gives clues to biological mechanism

# Air pollution studies

> *"NMMAPs [a large study of the acute effects of air pollution] played a central role in the Environmental Protection Agency's development of national ambient air quality standards for the six 'criteria' pollutants'."*

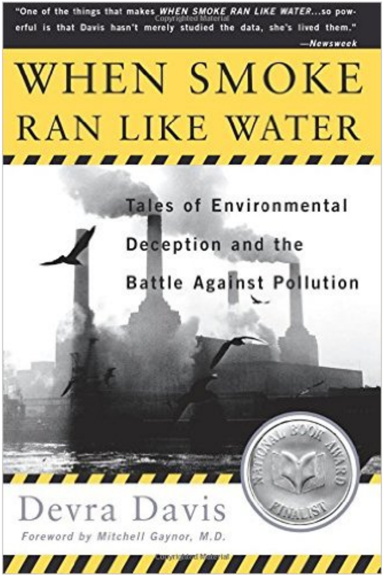
*Source: Peng et al. JRSS-A 2006*

# Air pollution studies

> *"The critical role of the NMMAPs in the development of the air quality standards attracted intense scrutiny from the scientific community and industrial groups regarding the statistical models that are used and the methods that are employed for adjusting for potential confounding."*

*Source: Peng et al. JRSS-A 2006*

# Air pollution studies



Example data: Chicago NMMAPS

# chicagoNMMAPS data

For the examples in this lecture, I'll use some data from Chicago on mortality, temperature, and air pollution. These data are available as part of the `dlnm` package. You can load them in R using the following code:

```
library(dlnm)
data("chicagoNMMAPS")
```

# chicagoNMMAPS data

To make the data a little easier to use, I'll rename the data frame as `chic`:

```
chic <- chicagoNMMAPS
chic[1:3, c("date", "cvd", "temp", "dptp", "pm10")]
```

```
##         date cvd       temp   dptp     pm10
## 1 1987-01-01  65 -0.2777778 31.500 26.95607
## 2 1987-01-02  73  0.5555556 29.875       NA
## 3 1987-01-03  43  0.5555556 27.375 32.83869
```

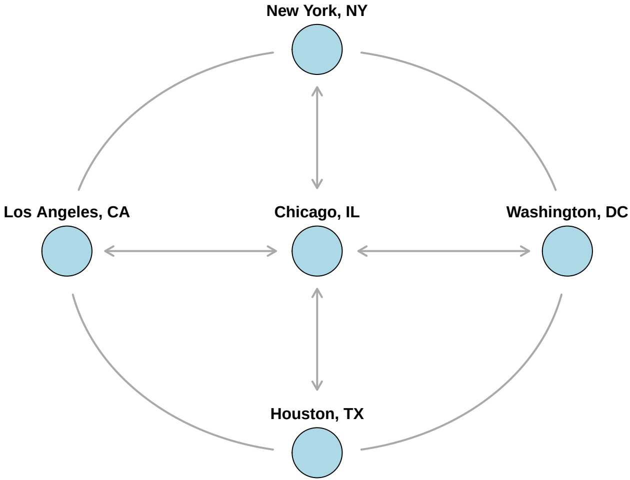
# chicagoNMMAPS data

To find out more about this data, you can look at its help file:

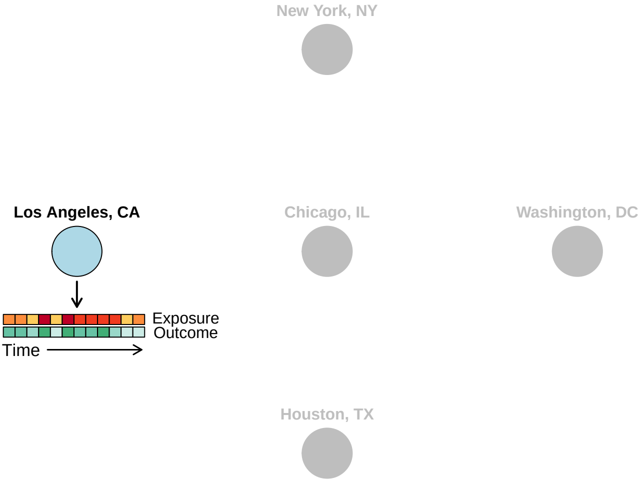
```
?chicagoNMMAPS
```

# Concept: Time series studies

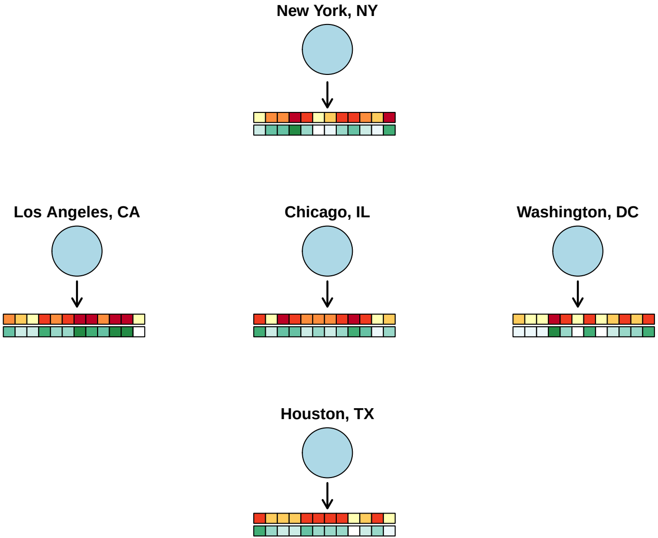
# Model design



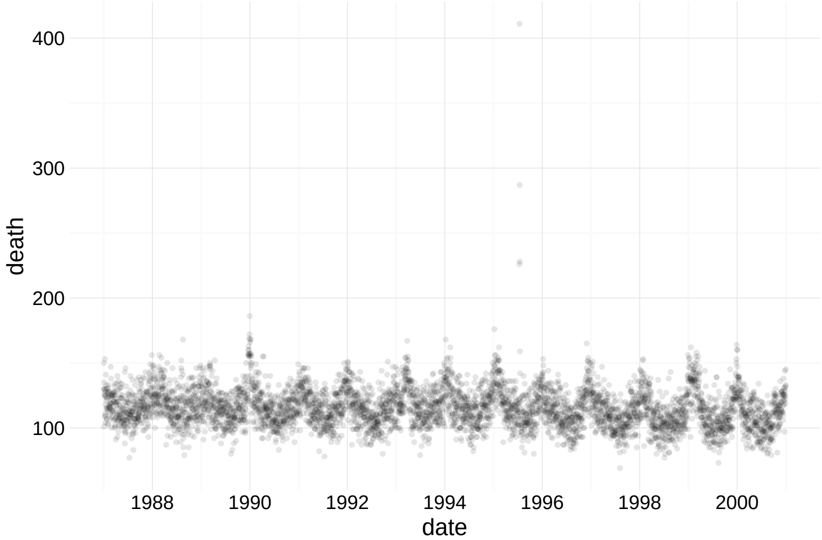
# Model design



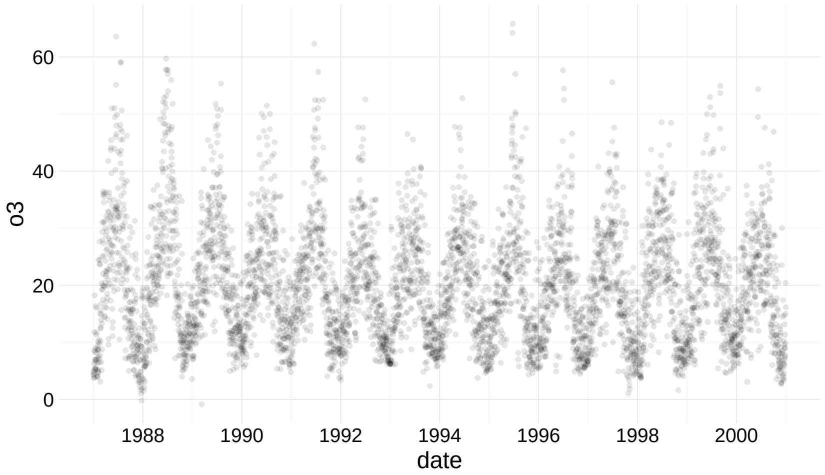
# Model design



# Confounders



# Confounders



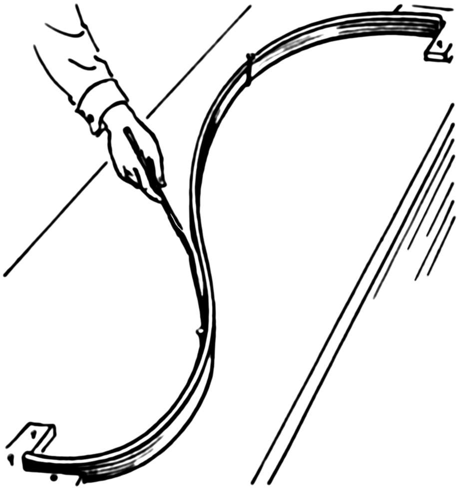
# Confounders

- Measured confounders
  - Temperature
  - Dew point temperature
  - Day of the week
- Unmeasured confounders
  - Long-term time trends
    - Changing population size
    - Changing population demographics
  - Seasonal time trends
    - Respiratory infections
    - Influenza

# Controlling for confounders

Some cofounders you might want to fit using a more complex form. For example, the relationship between temperature and mortality is often non-linear, with the lowest risk at mild temperatures and increasing risk as temperature gets colder or hotter.
Because of that, we often are interested in including temperature in the model using a natural cubic spline.

# Splines



*Source: Wikipedia*

# Choosing degrees of freedom

- Data-driven: Minimize a goodness-of-fit metric
- *A priori*: Use a reasonable value based on prior knowledge

# Choosing degrees of freedom

> *"An alternate approach is to use a fixed degrees of freedom, perhaps based on biological knowledge or previous work. For multisite studies, this approach leads to fitting the same model to data from each location. One can explore the sensitivity of $\hat{\beta}$ by varying the df used in the model(s) and examining the associated changes in $\hat{\beta}$."*

*Source: Peng and Dominici, "Statistical Methods for Environmental Epidemiology with R"*

# Convergence: "GAM-gate"

American Journal of
EPIDEMIOLOGY

Volume 156

Number 3

August 1, 2002

Copyright © 2002 by The Johns Hopkins
Bloomberg School of Public Health
Sponsored by the Society for Epidemiologic Research
Published by Oxford University Press

**COMMENTARY**

## On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health

**Francesca Dominici[1], Aidan McDermott[1], Scott L. Zeger[1], and Jonathan M. Samet[2]**

[1] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD.
[2] Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD.

# Implementation: Time series studies

# Overdispersed Poisson example

For example, say we wanted to fit an overdispersed Poisson regression for the `chic` data of whether cardiovascular mortality is associated with particulate matter (note: in this simplified example, I'm not controlling for many things we normally would, like season and temperature).

```
mod_d <- glm(cvd ~ pm10, data = chic,
             family = quasipoisson())
summary(mod_d)$coef[ , c(1, 2, 4)]
```

```
##                  Estimate   Std. Error  Pr(>|t|)
## (Intercept)  3.9327868499 0.0059351054 0.0000000
## pm10        -0.0001760049 0.0001530262 0.2501337
```

# Overdispersed Poisson example

Here, the model coefficient gives the **log relative risk** of cardiovascular mortality associated with a unit increase in PM10 concentration.

# Controlling for confounders

We usually want to control for other confounders. For example, when we look at the association between PM10 and cardiovascular mortality, we probably want to control for things like day of the week, seasonal and long-term mortality trends, and temperature. We can control for these potential confounders by adding them in to the right-hand side of the formula:

```
# Note: This is pseudocode
[health outcome] ~ [exposure of interest] +
                   [confounder1] + [confounder 2] ...
```

# Controlling for confounders

For example, we usually want to control for day of the week as a factor. To do that, first make sure that day of the week has the class `factor`:

```
class(chic$dow)
```

```
## [1] "factor"
```

# Controlling for confounders

If so, you can include it in your model:

```
mod_e <- glm(cvd ~ pm10 + dow, data = chic,
             family = quasipoisson())
summary(mod_e)$coef[ , c(1, 2, 4)]
```

```
##                  Estimate    Std. Error     Pr(>|t|)
## (Intercept)   3.914136910 0.0089931211 0.000000e+00
## pm10         -0.000211628 0.0001550096 1.722355e-01
## dowMonday     0.048178938 0.0111293675 1.528087e-05
## dowTuesday    0.030462708 0.0111124226 6.141692e-03
## dowWednesday  0.011251065 0.0111850000 3.145106e-01
## dowThursday   0.012227882 0.0111654287 2.735028e-01
## dowFriday     0.019722730 0.0111817715 7.782369e-02
## dowSaturday   0.016837469 0.0111119395 1.297719e-01
```

# Controlling for confounders

You can use `ns()` from the `splines` package to fit temperature using a spline. Here, I am fitting a spline with four degrees of freedom:

```
library(splines)
mod_e <- glm(cvd ~ pm10 + dow + ns(temp, 4),
             data = chic,
             family = quasipoisson())
summary(mod_e)$coef[c(1:2, 9:12), c(1, 2, 4)]
```

```
##                   Estimate   Std. Error     Pr(>|t|)
## (Intercept)   4.0533073621 0.0357364704 0.000000e+00
## pm10          0.0008714696 0.0001579088 3.590381e-08
## ns(temp, 4)1 -0.1767349276 0.0318373157 2.988366e-08
## ns(temp, 4)2 -0.3278324170 0.0254813081 2.842988e-37
## ns(temp, 4)3 -0.1992809283 0.0743260391 7.361296e-03
## ns(temp, 4)4 -0.0279239953 0.0272149287 3.049171e-01
```

# Controlling for confounders

Controlling for seasonal and long-term trends is similar. Often, we will use a spline with around 7 degrees of freedom per year. To fit this, first find out how many years are in your data:

```r
length(unique(chic$date)) / 365
```

```
## [1] 14.01096
```

# Controlling for confounders

Then add a column for `time`:

```
chic$time <- scale(chic$date, scale = FALSE,
                   center = TRUE)
chic$time[1:3]
```

```
## [1] -2556.5 -2555.5 -2554.5
```

```
summary(chic$time)
```

```
##        V1
##  Min.   :-2556
##  1st Qu.:-1278
##  Median :    0
##  Mean   :    0
##  3rd Qu.: 1278
##  Max.   : 2556
```

# Controlling for confounders

Now you can fit the model:

```
mod_e <- glm(cvd ~ pm10 + dow + ns(temp, 4) +
             ns(time, 7 * 14),
             data = chic,
             family = quasipoisson())
summary(mod_e)$coef[c(1:2, 13:15), c(1, 2, 4)]
```

```
##                       Estimate   Std. Error     Pr(>|t|
## (Intercept)        4.162023561 0.0583344638 0.000000000
## pm10               0.000202813 0.0001540345 0.188011886
## ns(time, 7 * 14)1 -0.059709991 0.0583191728 0.305958995
## ns(time, 7 * 14)2 -0.181691561 0.0770088168 0.018346681
## ns(time, 7 * 14)3 -0.240738856 0.0700453542 0.000593458
```

# Controlling for convergence problems

One way to account for "GAM-gate" is to change the convergence default threshold using the `control` option in `glm`:

```
mod_e <- glm(cvd ~ pm10 + dow + ns(temp, 4) +
                   ns(time, 7 * 14),
             data = chic,
             family = quasipoisson(),
             control = glm.control(epsilon=10E-8,
                                   maxit = 10000))
```

Generally, it is good practice to include this when modeling air pollution-health relationships.

# Interpreting model coefficients

You can pull the model coefficient you're interested in from the model summary using this code:

```
pm_coef <- summary(mod_e)$coefficients["pm10", ]
pm_coef
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
## 0.0002028130 0.0001540345 1.3166725629 0.1880118867
```

# Interpreting model coefficients

Remember that this model coefficient is the **log** relative risk, since we fit a quasi-Poisson model. To get a relative risk estimate, you'll need to take the exponent:

```
exp(pm_coef[1])
```

```
## Estimate
## 1.000203
```

Therefore, there is a relative risk of 1.0002028 for each increase of 1 $\mu g/m^3$ PM10.

# Interpreting model coefficients

Often, epidemiology studies will present relative risk for a 10-unit, rather than 1-unit, increase in exposure (e.g., per 10 $\mu g/m^3$ PM10). To estimate this, you need to multiple the coefficient by 10 *before* taking the exponential:

```
exp(10 * pm_coef[1])
```

```
## Estimate
##  1.00203
```

Therefore, there is a relative risk of 1.0020302 for an increase of 10 $\mu g/m^3$ PM10.

# Interpreting model coefficients

Sometimes, epidemiology studies will present results as % increase in mortality instead of relative risk. You can calculate this as:
% increase = 100 * (RR - 1)
For our example model, you could calculate:

```
100 * (exp(10 * pm_coef[1]) - 1)
```

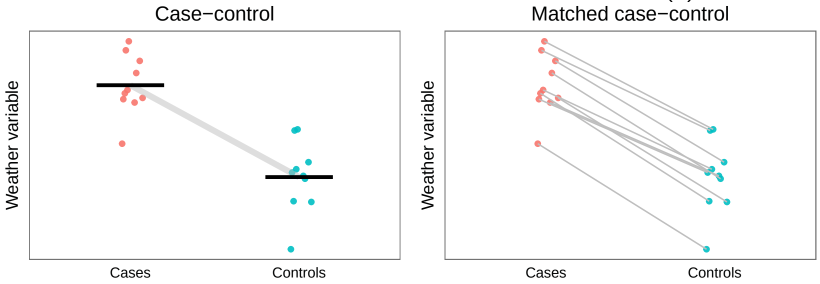
```
##  Estimate
## 0.2030188
```

Therefore, there is a 0.203% increase in mortality for an increase of 10 $\mu g/m^3$ PM10.

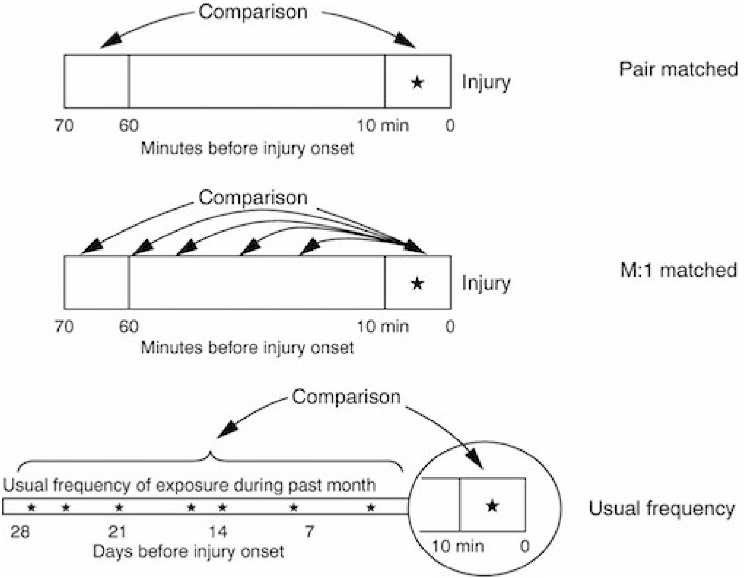
# Concept: Case-crossover studies

# Case-crossover models

Case-crossover model designs are based on the idea of matched case-control studies. For these, instead of comparing averages of exposure for cases versus controls, you compare the average difference across each matched set of case and control(s).

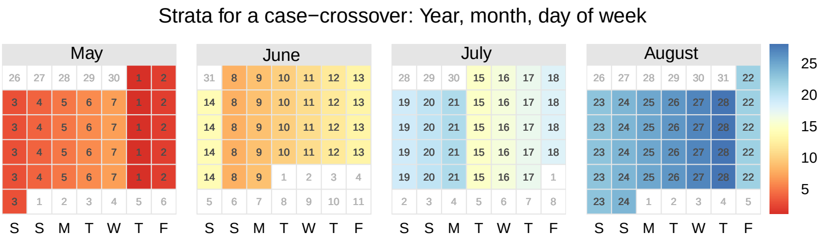


# Types of case-crossover designs




*Source: Sorock et al. 2001, Injury Prevention*

# Strata for case-crossover



# Concept of case-crossover

For each death in the dataset: Given that the death happened on one of the days in its strata, what is the probability that it happened on the day it did?

$$Pr(Death|Stratum, Exposure)$$

# Conditional logistic vs. GLM

**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 North Wolfe Street, Baltimore, MD 21205-2179, USA*
ylu@jhsph.edu

SCOTT L. ZEGER

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

# Conditional logistic vs. GLM

> *"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."*

*Source: Lu and Zeger Biostatistics 2007*

# Conditional logistic vs. GLM

Case-crossover fit using a GLM:

$$E(log(Y_t)) \sim \beta_0 + \beta_1 PM_t + \beta_2 Stratum_t$$

# Conditional logistic vs. GLM

**Table 2 Excerpt from example daily data in original format**

| Stratum | Date | Ozone | Temp-erature | n. of deaths |
|---|---|---|---|---|
| 2002 1 Sun | 06 jan 2002 | 2.4 | 7.1 | 198 |
| 2002 1 Sun | 13 jan 2002 | 17.6 | 8.2 | 204 |
| 2002 1 Sun | 20 jan 2002 | 49.9 | 8.9 | 167 |
| 2002 1 Sun | 27 jan 2002 | 42.5 | 10.5 | 169 |
| 2002 1 Mon | 07 jan 2002 | 4.1 | 5.2 | 180 |
| . . . . | | | | |

*Source: Armstrong et al. BMC Medical Research Methodology 2014*

# Conditional logistic vs. GLM

**Table 3 Excerpt from example data in semi-expanded format for case crossover conditional logistic analysis**

| Stratum | Case-con set | Date | Ozone | Temp-erature | Case day | Weight |
|---|---|---|---|---|---|---|
| 2002 1 Sun | 2002 1 Sun 1 | 06 jan 2002 | 2.4 | 7.1 | 1 | 198 |
| 2002 1 Sun | 2002 1 Sun 1 | 13 jan 2002 | 17.6 | 8.2 | 0 | 198 |
| 2002 1 Sun | 2002 1 Sun 1 | 20 jan 2002 | 49.9 | 8.9 | 0 | 198 |
| 2002 1 Sun | 2002 1 Sun 1 | 27 jan 2002 | 42.5 | 10.5 | 0 | 198 |
| 2002 1 Sun | 2002 1 Sun 2 | 06 jan 2002 | 2.4 | 7.1 | 0 | 204 |
| 2002 1 Sun | 2002 1 Sun 2 | 13 jan 2002 | 17.6 | 8.2 | 1 | 204 |
| 2002 1 Sun | 2002 1 Sun 2 | 20 jan 2002 | 49.9 | 8.9 | 0 | 204 |

*Source: Armstrong et al. BMC Medical Research Methodology 2014*

# Implementation: Case-crossover studies

# GLM method

To code using a GLM, first you need to create a column with the stratum. In R, you can use `format` with the date to do this easily, and then convert the formatted date for a `factor` class:

```
chic$casecross_stratum <- format(chic$date, "%Y-%m-%a")
chic$casecross_stratum <- factor(chic$casecross_stratum)
head(chic$casecross_stratum, 3)
```

```
## [1] 1987-01-Thu 1987-01-Fri 1987-01-Sat
## 1176 Levels: 1987-01-Fri 1987-01-Mon 1987-01-Sat 1987-0
```

# Case-crossover

Now you can include this factor in your model (note: this takes the place of model control for time trends and day of week in a typical time series model):

```
mod_f <- glm(cvd ~ pm10 + ns(temp, 4) + casecross_stratum,
             data = chic,
             family = quasipoisson())
summary(mod_f)$coef[c(1:2, 7:10), c(1, 2)]
```

```
##                                 Estimate   Std. Error
## (Intercept)                 4.0482946294 0.0855215089
## pm10                        0.0001909843 0.0001680322
## casecross_stratum1987-01-Mon 0.1907876393 0.1137590495
## casecross_stratum1987-01-Sat 0.0855529446 0.1168756412
## casecross_stratum1987-01-Sun 0.3300835895 0.1099033832
## casecross_stratum1987-01-Thu 0.0462517003 0.1043859066
```

# Case-crossover

You can interpret the coefficients now in the same way as with the time series model:

```
pm_coef <- summary(mod_f)$coefficients["pm10", ]
100 * (exp(10 * pm_coef[1]) - 1)
```

```
##  Estimate
## 0.1911668
```

Therefore, for this model, there is a 0.191% increase in mortality for an increase of 10 $\mu g/m^3$ PM10.

# Case-crossover

There are also other methods for fitting case-crossover models:

- Armstrong et al. (Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis) suggest using a conditional Poisson regression model (`gnm()`) to speed up computational time.
- The `casecross` function in the `season` package by Adrian Barnett uses 28-day strata (rather than by month) and a Cox proportional hazards regression model to fit the model.

If you are using this method for a paper, it is worthwhile testing the different methods to see if you get similar results.

# Case-crossover

Using a conditional Poisson model:

```r
library(gnm)
mod_g <- gnm(cvd ~ pm10 + ns(temp, 4),
             eliminate = casecross_stratum,
             data = chic,
             family = quasipoisson())
pm_coef <- summary(mod_g)$coefficients["pm10", ]
100 * (exp(10 * pm_coef[1]) - 1)
```

```
##  Estimate
## 0.1911668
```

# Case-crossover

Using a Cox proportional hazards regression model:

```
library(season)
mod_h <- casecross(cvd ~ pm10 + temp,
                   matchdow = TRUE,
                   data = chic)
```

```
## Note, irregularly spaced data...
## ...check your data for missing days
```

```
pm_coef <- mod_h$c.model$coefficients[1]
100 * (exp(10 * pm_coef[1]) - 1)
```

```
##      pm10
## 0.4320228
```

# Multi-city studies

# Samet editorial



EPIDEMIOLOGY March 2002, Vol. 13 No. 2

## Air Pollution and EPIDEMIOLOGY: "Déjà Vu All Over Again?"

For centuries, air pollution has been a public health and aesthetic concern, managed by governments (with varying degrees of success) to protect the public. Although there is still uncertainty about many aspects of air pollution and health, there are now evidence-based regulations in many countries to protect the public from air pollution by motor vehicles and by

Methods developed for air pollution research have also been creatively applied to other areas of epidemiology, such as infectious disease.[2,3] EPIDEMIOLOGY has provided a forum for discussion of these new methodologies and for divergent views on the findings and their interpretation.[4–7]

Not surprisingly, many of our recently submitted manuscripts on air pollution follow in the footsteps of

# NMMAPS


*Source: www.ihapss.jhsph.edu*

# NMMAPS package

NMMAPSdata package

**Data**

- akr
- albu
- Anch

*and 105 other US cities*

- *Meta-data on cities (population, location, counties, Census variables)*

**Functions**

- readCity
- getMetaData

*and various other functions for different versions of the package*

**Documentation**

- PDF users' manual
- Instructions for each function within R
- Examples for each function within R
- Website

# Impact of NMMAPS

Research impacts of NMMAPS package

- As of November 2011, 67 publications had been published using this data, with 1,781 citations to these papers
- Research using NMMAPS has been used by the US EPA in creating regulatory impact statements for air pollution (particulates and ozone)
- "Thanks to NMMAPS, there is probably no other country in the world with a greater understanding of the health effects of air pollution and heat waves in its population.""

*Source: Barnett, Huang, and Turner, "Benefits of Publicly Available Data", Epidemiology 2012*