

Acute effects of ambient exposures

Time series and case-crossover studies

Brooke Anderson, Colorado State University

✉: brooke.anderson@colostate.edu

🐦: [@gbwanderson](https://twitter.com/gbwanderson)

👤: github.com/geanders

Overview

Example data: Chicago NMMAPS

Concept: Time series studies

Implementation: Time series studies

Concept: Case-crossover studies

Implementation: Case-crossover studies

If you would like a copy of these slides, you can download a copy by clicking on this link:

https://github.com/geanders/GuestLectures/raw/master/CaseCrossover/feb_13_2018.pdf

If you would like to try out the example code, you can get a copy of the .Rmd file used to make the lecture:

https://github.com/geanders/GuestLectures/blob/master/CaseCrossover/feb_13_2018.Rmd

Overview

Community-wide effects of ambient exposures

Today, I'll talk about two popular study designs for studying the **community-wide effects** of **ambient environmental exposures** on human health outcomes.

The two study designs are:

- **Time series**
- **Case-crossover**

1952 “London Fog”



Source: *The Crown*, Season 1, Episode 4, Netflix

1952 “London Fog”

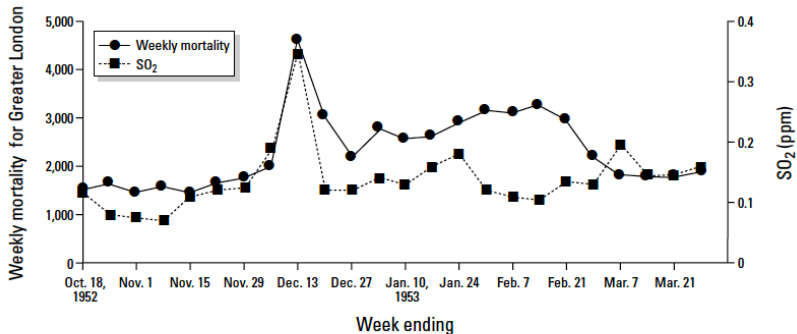
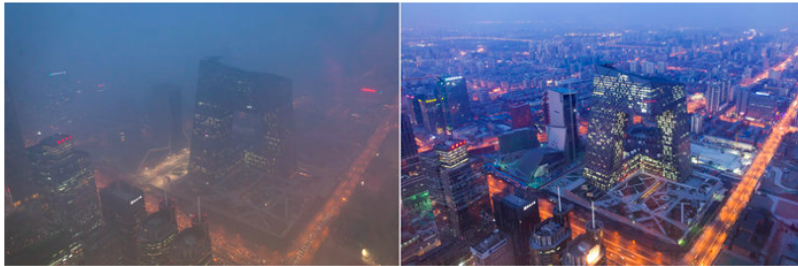


Figure 1. Approximate weekly mortality and SO₂ concentrations for Greater London, 1952–1953.

Source: Bell and Davis, 2001

2013 Beijing “Airpocalypse”

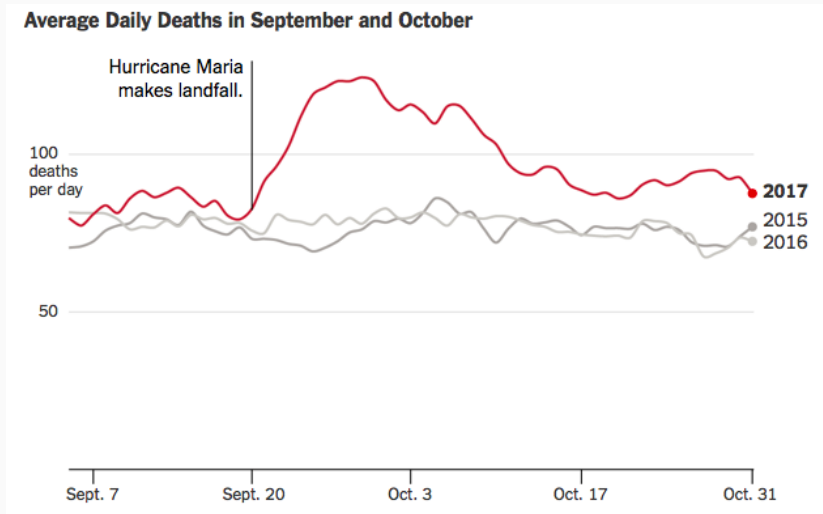


This combination of photos shows (left) the Beijing skyline during severe pollution Monday, and the same view (right) taken during clear weather on Feb. 4, 2012.

Ed Jones/AFP/Getty Images

Source: <https://www.npr.org/>

2017 Hurricane Maria



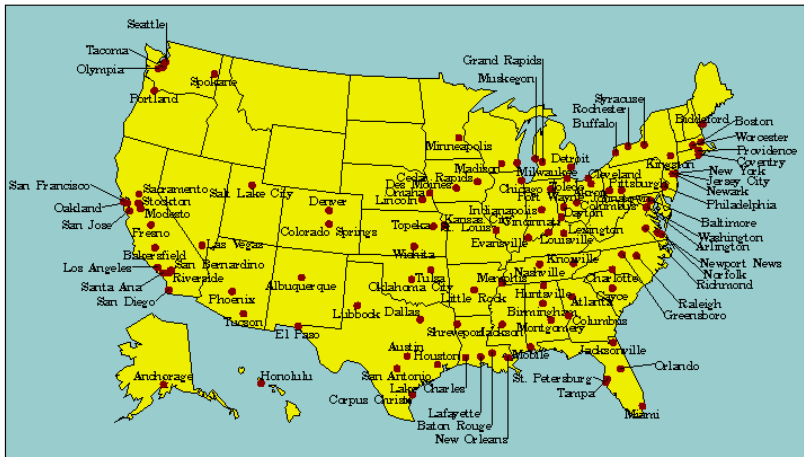
Source: The New York Times

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



Source: www.ihapss.jhsph.edu

“NMMAPs 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, 2006

“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.”

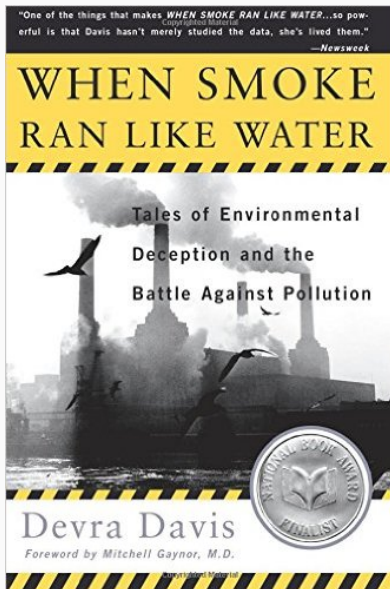
Source: Peng et al, 2006

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, 2012

Community-wide effects of ambient exposures



Example data: Chicago NMMAPS

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")]
```

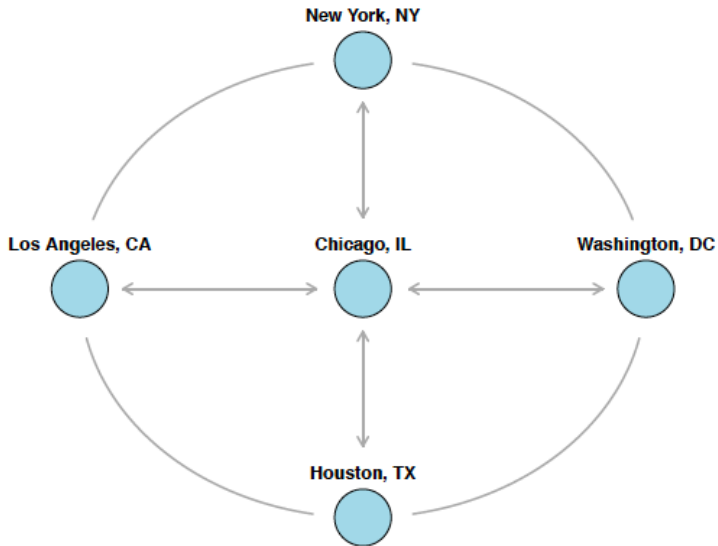
##		date	cvd	temp	dptp	pm10
## 1	1987-01-01	65	-0.2777778	31.500	26.95607	
## 2	1987-01-02	73	0.5555556	29.875		NA
## 3	1987-01-03	43	0.5555556	27.375	32.83869	

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

```
?chicagoNMMAPS
```

Concept: Time series studies

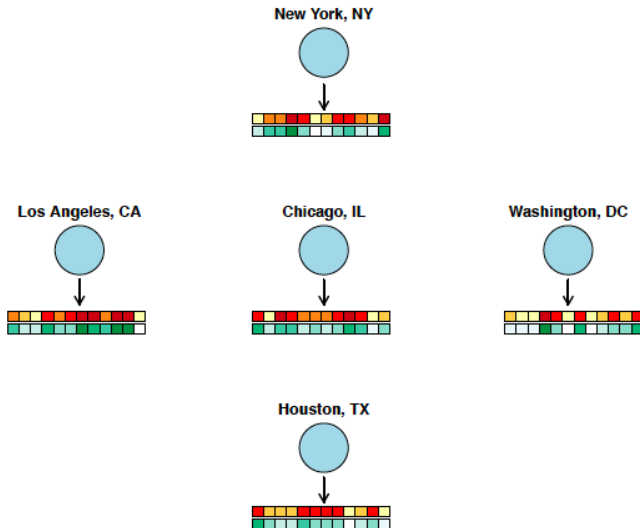
Model design



Model design



Model design



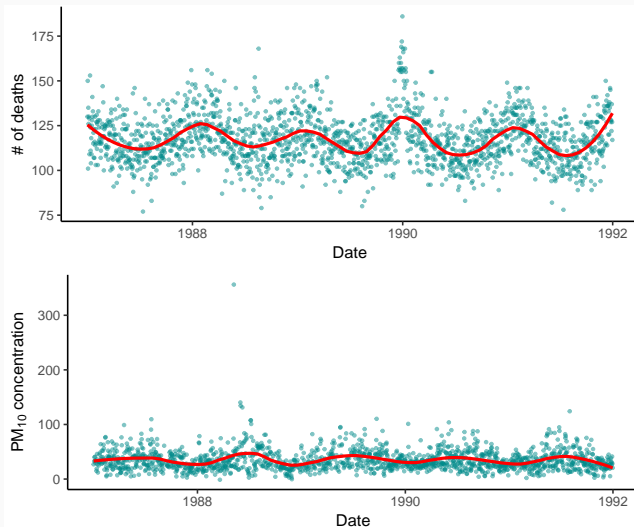
The model we fit is:

$$Y(t) \sim \text{Quasipoisson}(\mu_t, \sigma^2)$$
$$\log(\mu_t) = \beta_0 + \beta_1 PM_t + f(t) + Z(t)$$

where:

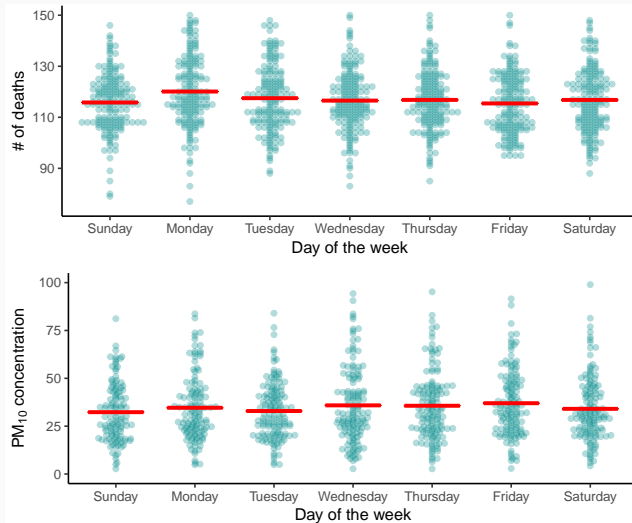
- $Y(t)$: Daily mortality count in community t
- PM_t : Daily PM_{10} count
- $f(t)$: Smooth function of time
- $Z(t)$: Other confounders

Confounders



Temporal trends in daily mortality and particulate matter, Chicago, IL

Confounders

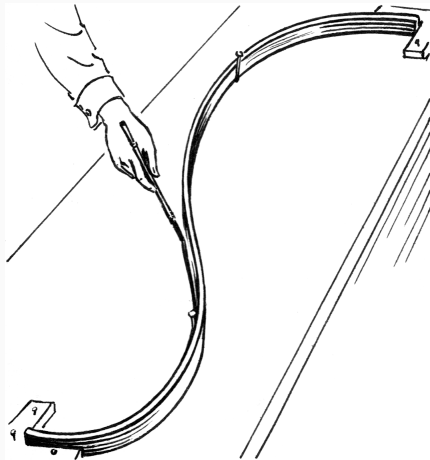


Day-of-week trends in daily mortality and particulate matter, Chicago, IL

Confounders

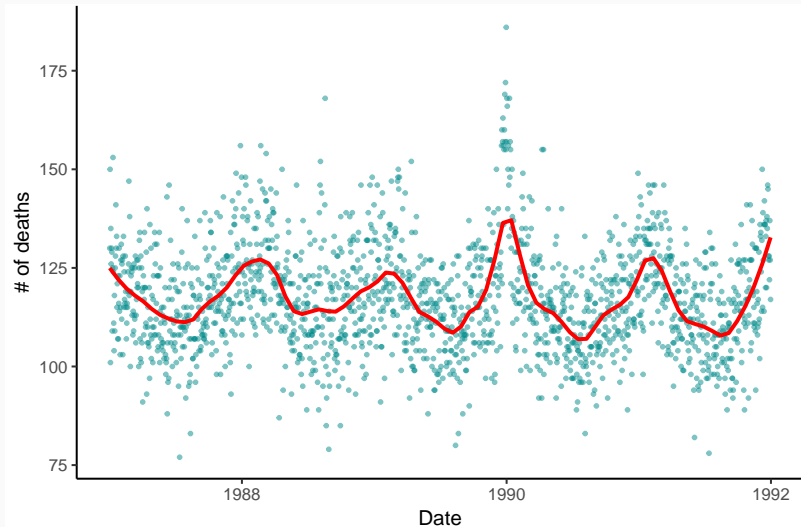
- Measured confounders
 - Temperature
 - Dew point temperature
 - Day of the week
- Unmeasured confounders
 - Long-term time trends
 - Changing population size
 - Changing population demographics
 - Seasonal time trends
 - Respiratory infections
 - Influenza

Splines



Source: Wikipedia

Confounders



Conceptual example of fitting a spline to control for long-term and seasonal trends

Convergence: “GAM-gate”



Volume 156

Number 3

August 1, 2002

American Journal of EPIDEMIOLOGY

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 Dominici¹, Aidan McDermott¹, Scott L. Zeger¹, and Jonathan M. Samet²

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 PM_{10} concentration.

Controlling for confounders

We usually want to control for other confounders. For example, when we look at the association between PM_{10} 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)] %>% head(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

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, 9:10), 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

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:

```
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
```

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)]
```


##	Estimate	Std. Error
## (Intercept)	4.162023561	0.0583344638
## pm10	0.000202813	0.0001540345
## ns(time, 7 * 14)1	-0.059709991	0.0583191728
## ns(time, 7 * 14)2	-0.181691561	0.0770088168
## ns(time, 7 * 14)3	-0.240738856	0.0700453542

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

Alternatively, you can use the `tidy` function from the `broom` package:

```
library(broom)
mod_e %>%
  tidy() %>%
  filter(term == "pm10")
```

```
##   term      estimate    std.error statistic    p.value
## 1 pm10 0.000202813 0.0001540345   1.316673 0.1880119
```

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 $\mu\text{g}/\text{m}^3$ PM_{10} .

Interpreting model coefficients

Often, epidemiology studies will present relative risk for a 10-unit, rather than 1-unit, increase in exposure (e.g., per $10 \mu\text{g}/\text{m}^3$ PM_{10}). 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 $10 \mu\text{g}/\text{m}^3$ PM_{10} .

Interpreting model coefficients

Sometimes, epidemiology studies will present results as % increase in mortality instead of relative risk. You can calculate this as:

$$\% \text{ increase} = 100 * (RR - 1)$$

For our example model, you could calculate:

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

```
## Estimate
```

```
## 0.2030188
```

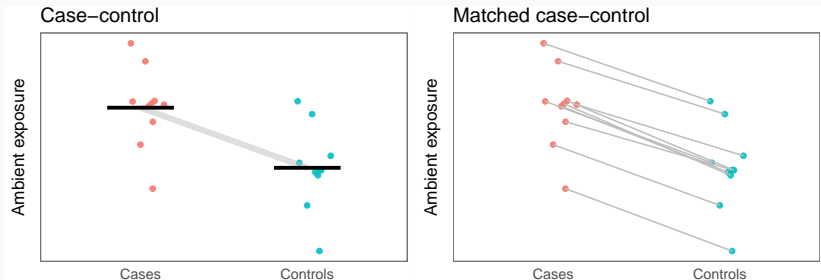
Therefore, there is a 0.203% increase in mortality for an increase of $10 \mu\text{g}/\text{m}^3$ PM_{10} .

►► [Go To Case-Crossover Implementation](#)

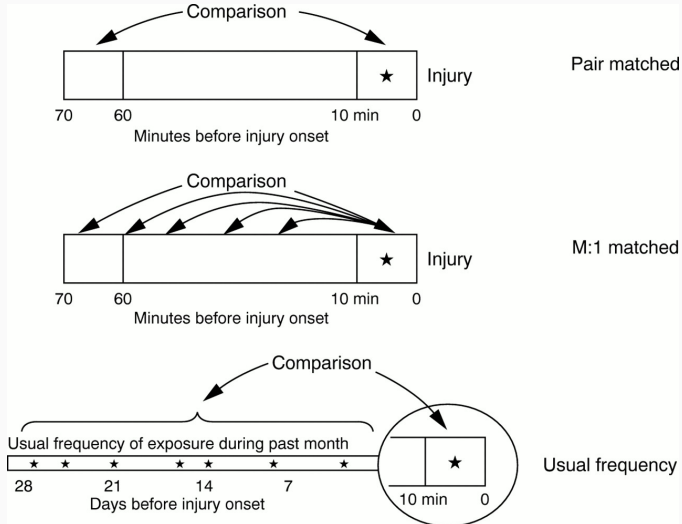
Concept: Case-crossover studies

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).

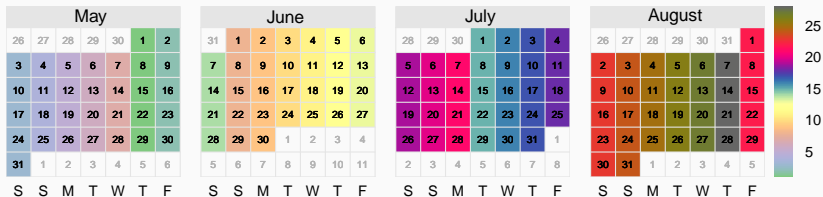


Types of case-crossover designs



Strata for case-crossover

Strata for a case-crossover: Year, month, day of week



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? Is the level of exposure associated with that probability?

$$Pr(\textit{Death}|\textit{Stratum}, \textit{Exposure})$$

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*

*"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:

$$Y(t) \sim \text{Quasipoisson}(\mu_t, \sigma^2)$$

$$\log(\mu_t) = \beta_0 + \beta_1 PM_t + \beta_2 \text{Stratum}_t + Z(t)$$

where:

- $Y(t)$: Daily mortality count in community t
- PM_t : Daily PM_{10} count
- Stratum_t : The stratum to which the day belongs
- $Z(t)$: Other confounders

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

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 1987-01-Tue 1987-01-Wed 1987-01-Thu 1987-01-Fri
```

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, 7:9), 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

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 $10 \mu\text{g}/\text{m}^3$ PM_{10} .

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:

```
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
```