

AS.280.347

CLASS 2.1

- Your projects!
 - Air particulate pollution and mortality
 - NMMAPS data
 - Log-linear models
-

Thinking ahead: your project!

- Question:
- Data set and design
 - Outcome:
 - Predictor variables of primary interest:
 - Effect modifiers:
 - Confounders:
- Directed Acyclic Graph (DAG):
- Primary analysis to address question:
- Communicating results in tables and figures:

Thinking ahead: your project!

Before you leave for Spring Break, you should have a rough idea of:

- A research question of interest in public health
- A data source that you can use to answer this question

Framing a research question in public health:

- Start with a general area of public health in which you have interest, and then narrow to a specific question you'd like to answer.
- It can be helpful to frame your question in terms of investigating a relationship between a specific outcome variable (like “disease status” for our Module 1) and one or more primary predictor variables (“smoking status” for our Module 1.)
- Later you will need to think about the possibility of effect modifiers and possible confounders, but for now just think about that primary relationship of interest!

Locating data to answer this question:

- If you have a specific area of interest in mind, you can Google for data in that area
- Or explore the links below to see what type of data is available:

<https://www.healthdata.gov/>

<http://guides.lib.berkeley.edu/publichealth/healthstatistics/rawdata>

<http://www.datasciencecentral.com/profiles/blogs/10-great-healthcare-data-sets>

https://www.cdc.gov/nchs/data_access/ftp_data.htm

https://catalog.data.gov/dataset?_organization_limit=0&organization=hhs-gov#topic=health_navigation

Module 2: Particulate air pollution and mortality

- **Question 2.1 (Q2.1): How does the daily risk of death depend upon air pollution level in American cities?**
- Question 2.2 (Q2.2): Is the estimate of the pollution effect sensitive to assumptions about seasonal or weather effects?
- Question 2.3 (Q2.3): How do you pool PM effect (log relative rate) estimates from multiple cities taking account of both natural geographic variability in the true effects and statistical errors that might differ among cities?
- We will answer these questions using data from the National Morbidity and Mortality Air Pollution Study (NMMAPS)

Particulate air pollution and mortality



London, 12 noon, December, 1952; Particulate levels – $3,000 \mu\text{g}/\text{m}^3$

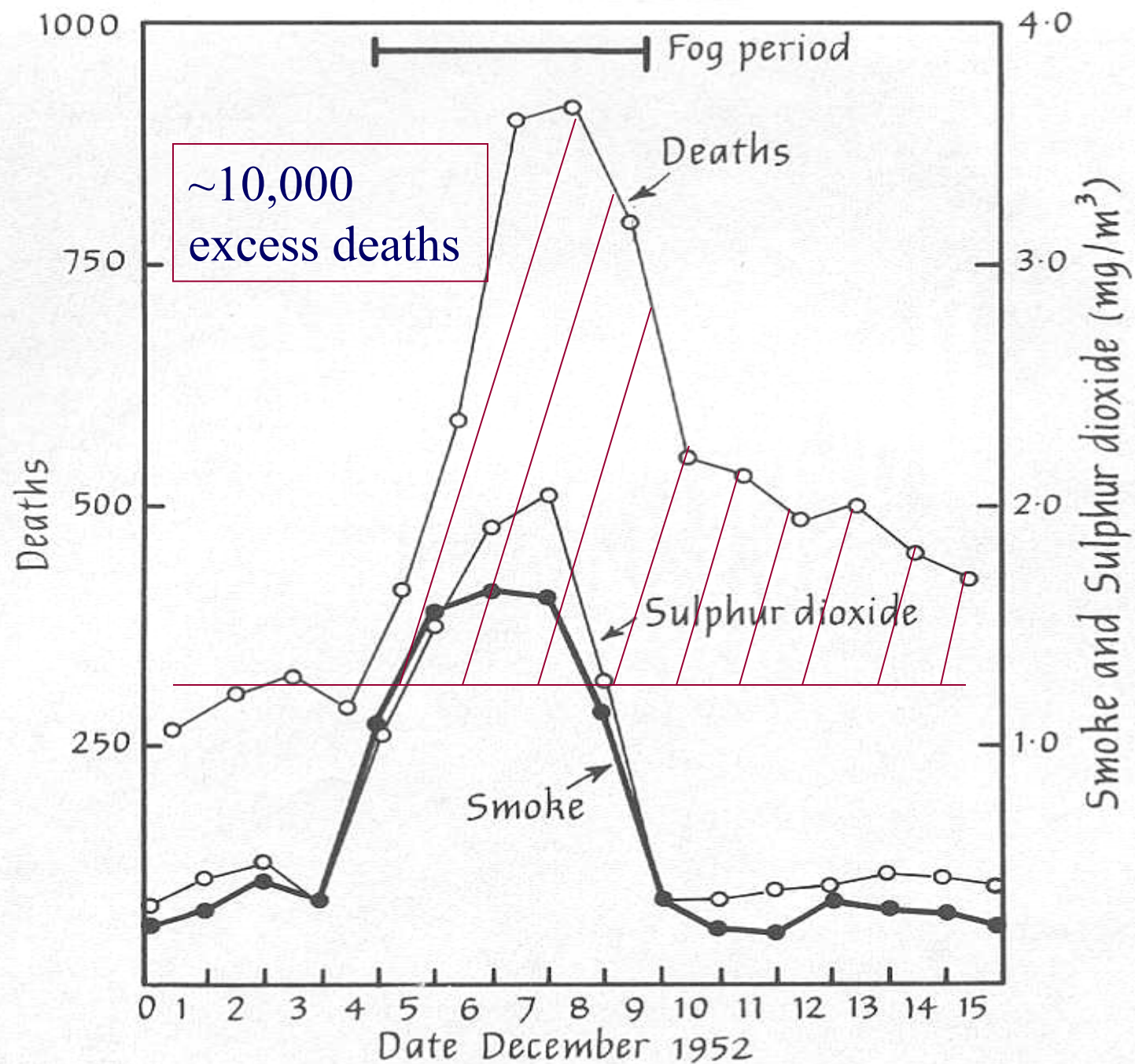
Particulate air pollution and mortality



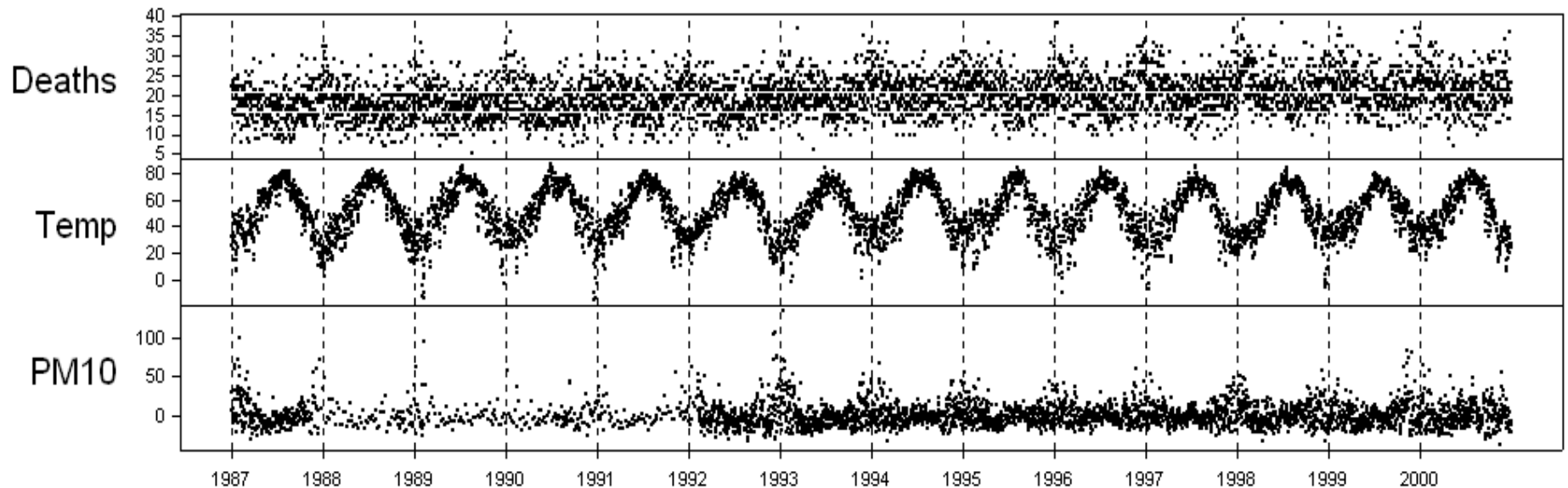
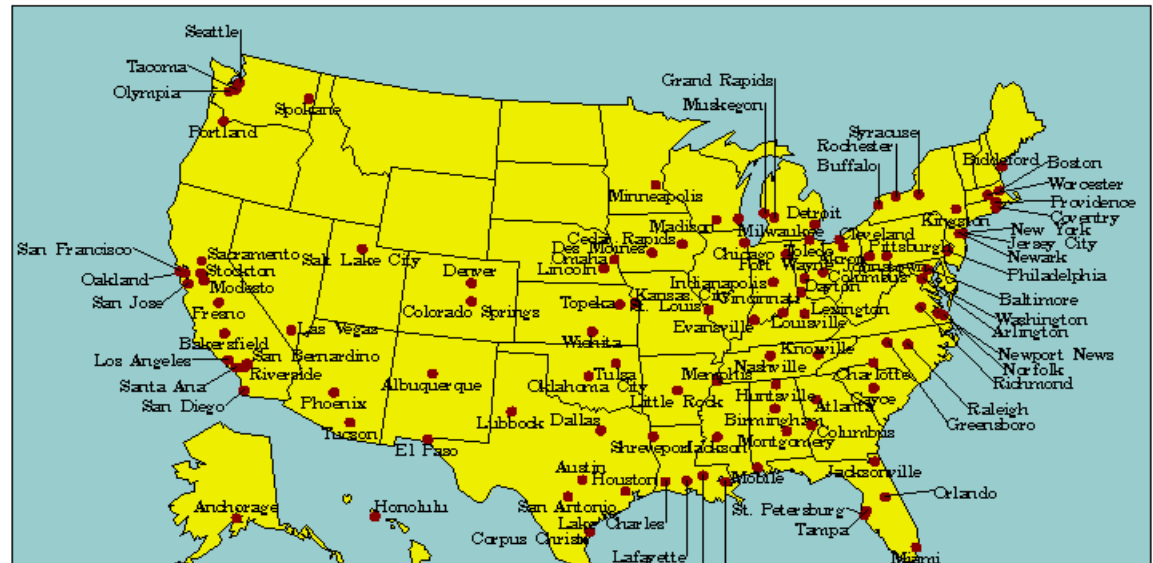
In this photograph, visibility is so low that the bus driver cannot see the road, even though it is mid-morning. A police officer is on foot leading the bus with a flashlight.



London, 12 noon, December, 1952; Particulate levels – $3,000 \mu\text{g}/\text{m}^3$



NMMAPS data



AIR POLLUTION

Evidence Mounts That Tiny Particles Can Kill

Four years ago, the U.S. Environmental Protection Agency (EPA) ignited a fire storm when it declared that tens of thousands of people were dying each year from breathing tiny particles of dust and soot—and issued tough new regulations to crack down on these pollutants. Industry groups and many scientists assailed the decision, arguing that the data underlying the new particulate matter (PM) standard were inconclusive at best, and industry took their case to court. Now, a long-awaited study, by a group widely perceived to be politically neutral, comes in solidly behind the earlier EPA decision and strongly implicates particles in excess deaths.

The study is the largest yet to examine the relation between daily levels of particles—which come mainly from roads, motor vehicles, and power plants—and deaths in the United States. Released last week by the Health Effects Institute (HEI) in Cambridge, Massachusetts, a nonprofit organization funded by industry and the government, the study found that death rates in the 90 largest U.S. cities rise on average 0.7% with each tiny 10 microgram per cubic meter increase in particles less than 10 micrometers in diameter, known as PM₁₀. That number is not much different from those found in earlier studies. But this time, the case is stronger because the breadth of the new study dispels any notion that the effect might have been caused by a pollutant other than PM₁₀, or even hot weather. Indeed, although many questions remain about how fine particles kill people, the HEI study shows there's no mistaking that PM is the culprit, lead author Jonathan Samet of Johns Hopkins University says emphatically.

It was similar studies of the relation between day-to-day fluctuations in fine particles and death rates that raised the alarm about PM some 10 years ago. In cities such as Philadelphia, researchers found that on days when air pollution jumped yet remained within federal standards, there were more deaths and hospitalizations of elderly people for cardiac and lung disease. Although the increase was slight in each city, studies of the long-term effects of particles found it added up to a significant number of deaths, roughly 60,000 a year* by some estimates. Lab studies showed that the finer the particles, the more likely it was to lodge in the lungs, suggesting to EPA that it needed to take

steps, among them whether the apparent link between deaths and PM levels was real or due to other pollutants (Science, 25 July 1997, p. 466). EPA went ahead with the standard but built in a 5-year delay to allow for more research. Meanwhile, a U.S. appeals court last year ruled that the science supported EPA's PM₁₀ standard. This fall, the U.S. Supreme Court will look at a related legal question—whether the EPA's interpretation of the Clean Air Act exceeds Congress's constitutional authority.

To help resolve the uncertainties in PM epidemiology, HEI in 1996 began funding the National Mortality, Mortality, and Air Pollution Study, or NMMAPS. Samet's team and collaborators at Harvard scored federal databases on daily deaths, weather, and air pollution. By including every major city with significant PM pollution and using the same methods to analyze each, the team achieved statistically stronger results than had previous single-city studies. The results varied by region—the rise in death rates was highest in the Northeast and lowest in the Southwest—but overall, the team found "a robust effect" from PM₁₀ across the 90 cities, says Samet. Adds biostatistician Gerald Van Belle of the University of Washington, Seattle, "Whatever concerns there was about a single-city idiosyncratic effect is no longer a testable hypothesis."



Profile of a killer. Particle pollution leads to increased deaths across much of the country, according to this map showing the rise in death rates with each 10 µg/m³ rise in PM₁₀. In areas where the ratio was below 2, the correlation was not statistically significant. Left, Houston, barely visible.

got even finer particles than before—those less than 2.5 micrometers across.

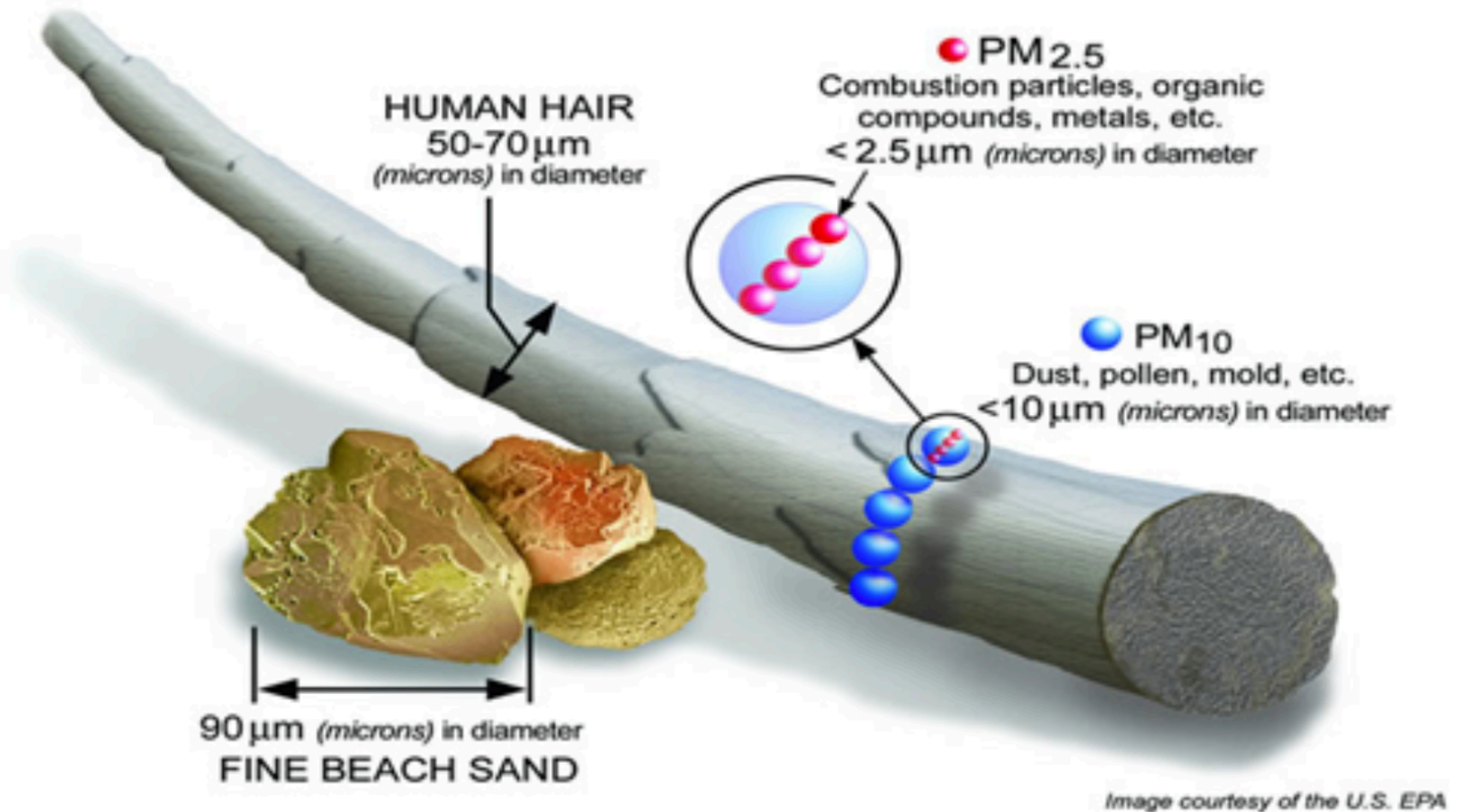
But in 1996, when EPA proposed a first-ever maximum level for PM_{2.5} together with tighter ozone standards, industry groups went on the warpath. In congressional hearings, scientists also raised a host of ques-

tion, among them whether the apparent link between deaths and PM levels was real or due to other pollutants (Science, 25 July 1997, p. 466). EPA went ahead with the standard but built in a 5-year delay to allow for more research. Meanwhile, a U.S. appeals court last year ruled that the science supported EPA's PM₁₀ standard. This fall, the U.S. Supreme Court will look at a related legal question—whether the EPA's interpretation of the Clean Air Act exceeds Congress's constitutional authority.

“(A)lthough many questions remain about how fine particles kill people, the NMMAPS study shows *there’s no mistaking that PM is the culprit*

Understatement of statistical uncertainty in the press

What is PM10?



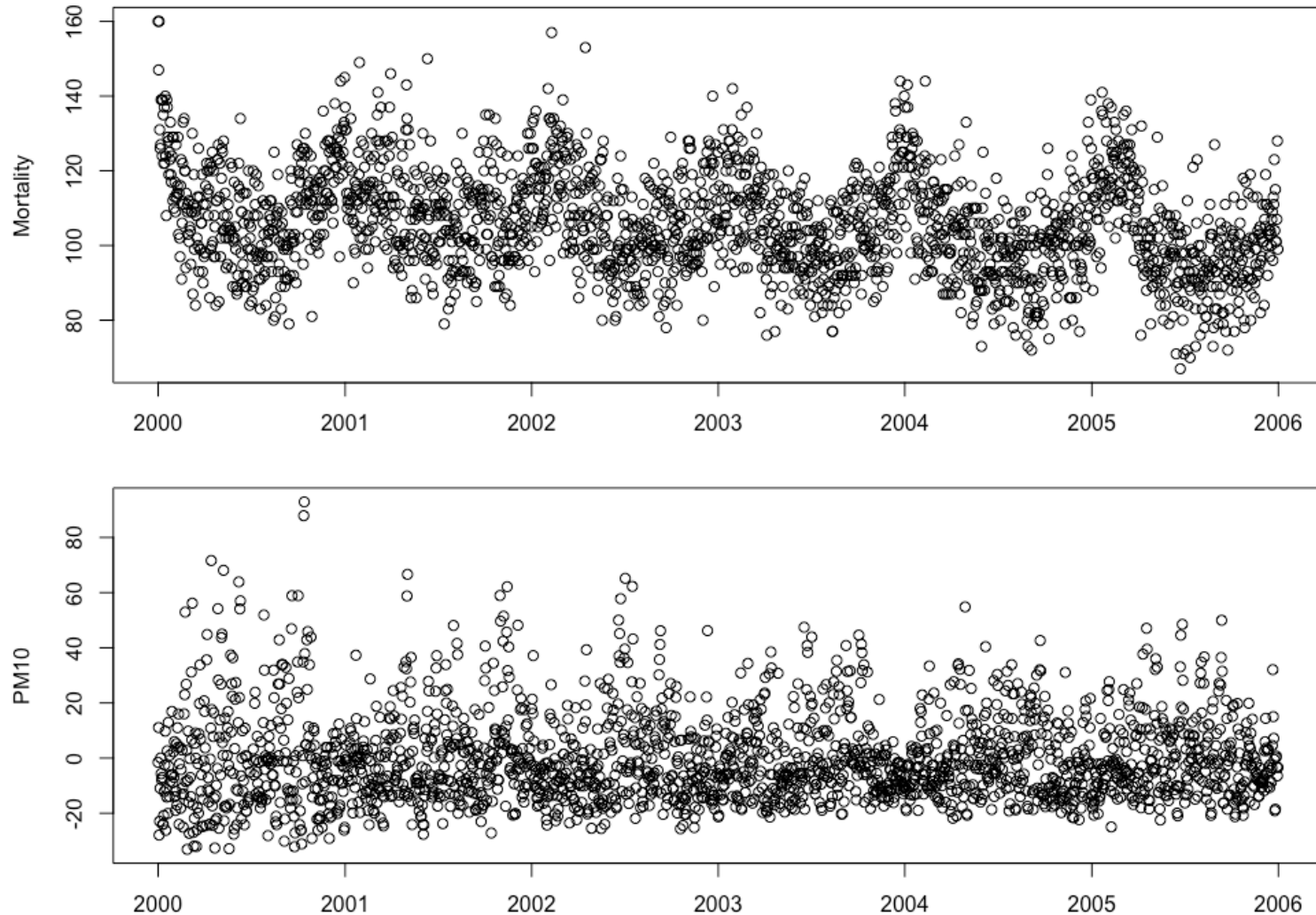
<http://www.irceline.be/en/documentation/faq/what-is-pm10-and-pm2.5>

NMMAPS data -- Chicago

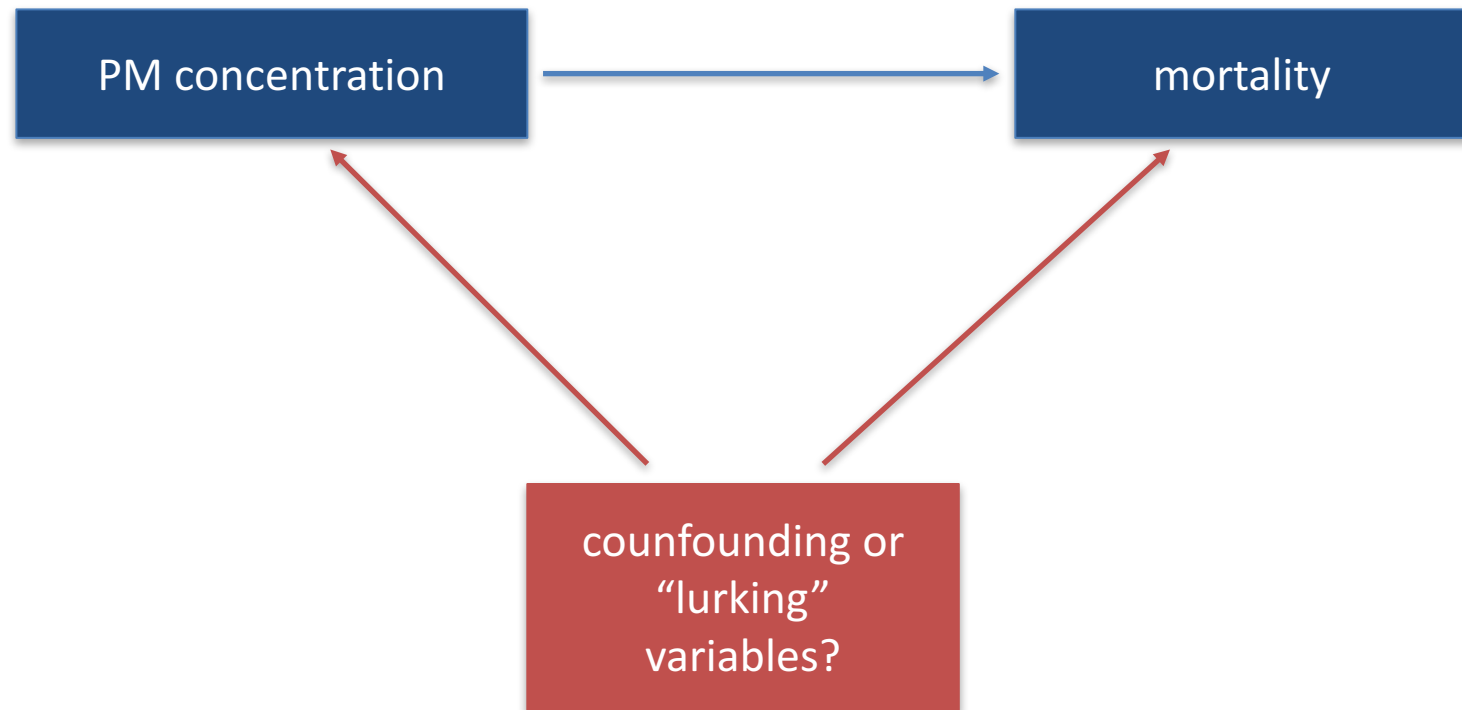
```
> head(chicago.data)
```

	date	city	death	pm10	tempF	o3	dow	season
1	1987-01-01	chic	130	-7.433544	31.5	-19.59234	Thursday	Winter
2	1987-01-02	chic	150	NA	33.0	-19.03861	Friday	Winter
3	1987-01-03	chic	101	-1.550923	33.0	-20.21734	Saturday	Winter
4	1987-01-04	chic	135	5.566456	29.0	-19.67567	Sunday	Winter
5	1987-01-05	chic	126	NA	32.0	-19.21734	Monday	Winter
6	1987-01-06	chic	130	6.566456	40.0	-17.63400	Tuesday	Winter

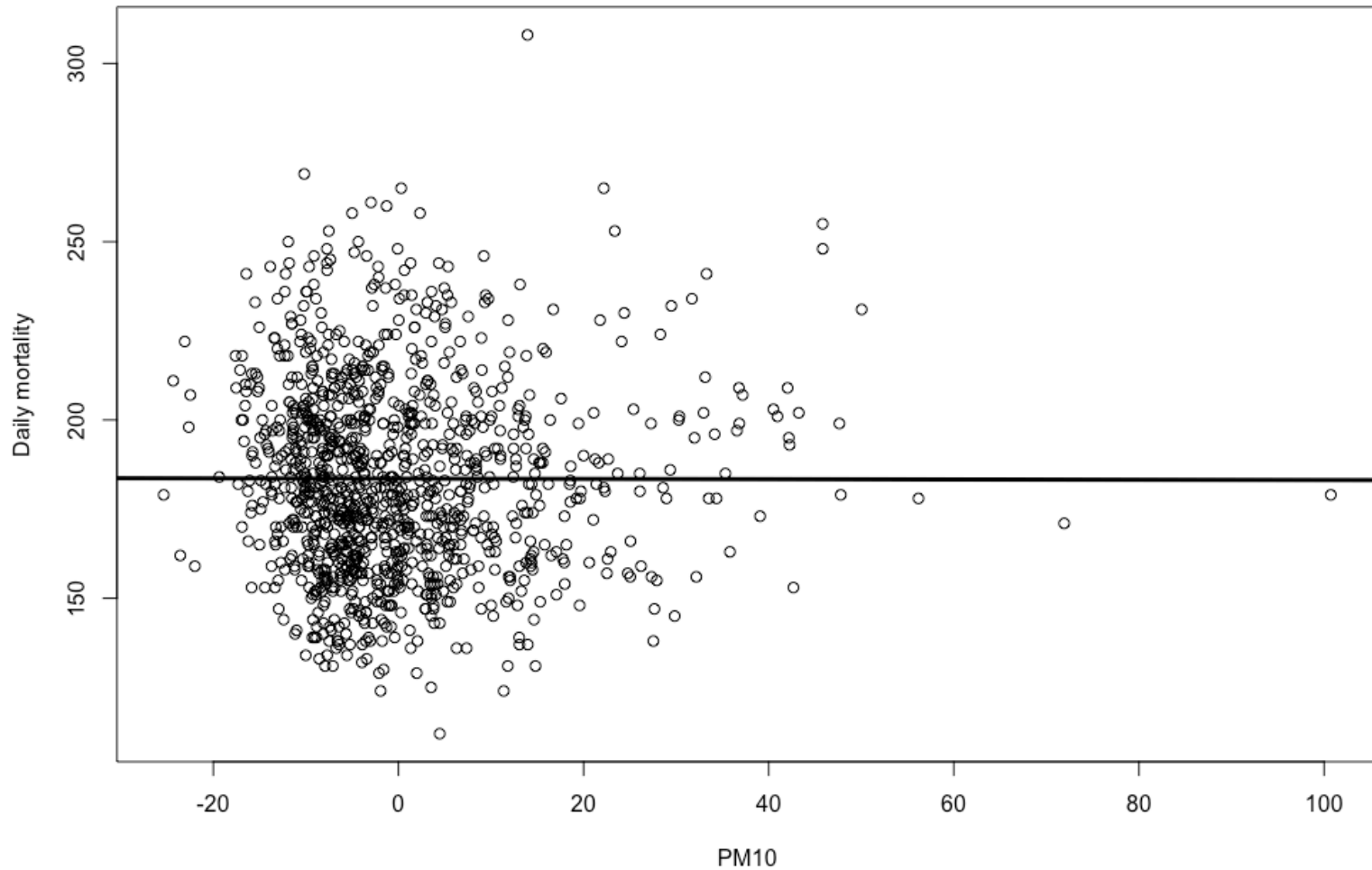
NMMAPS data -- Chicago



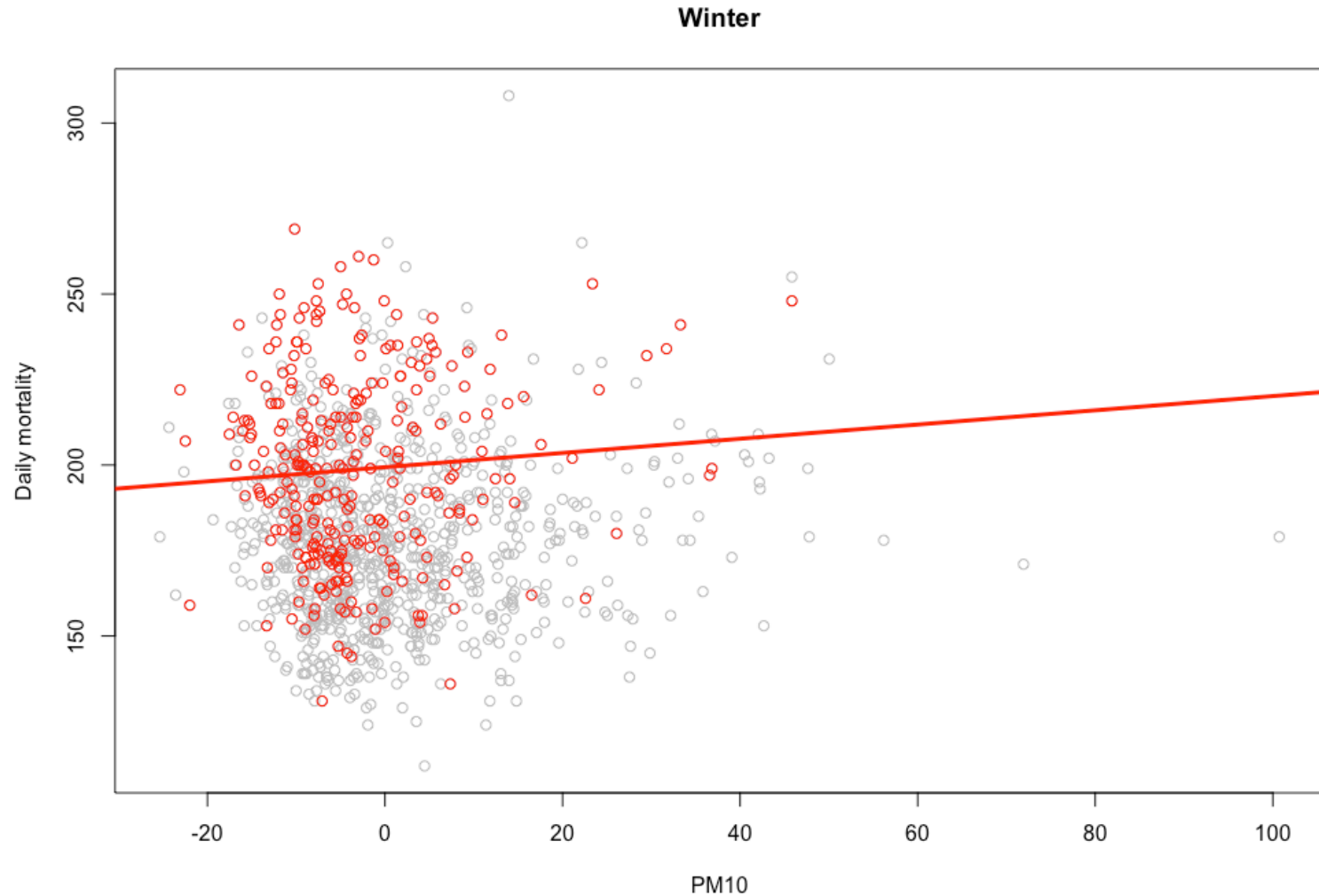
DAG of relationship



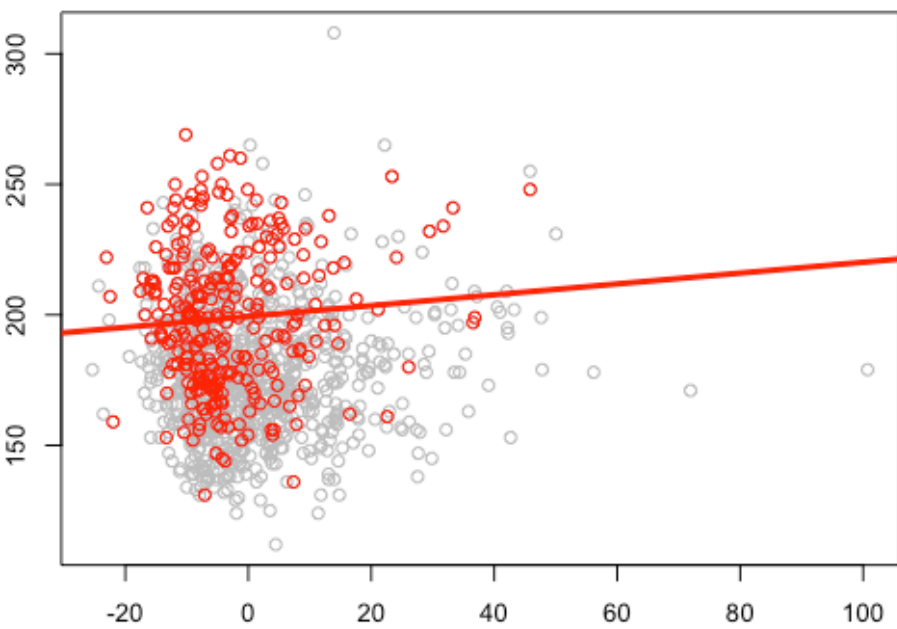
NMMAPS data – New York



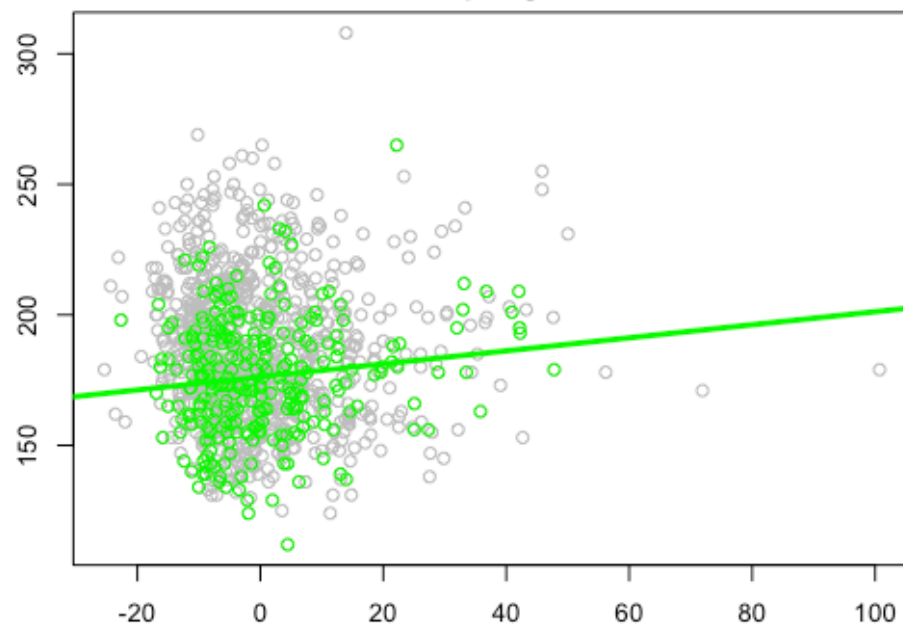
NMMAPS data – New York (winter)



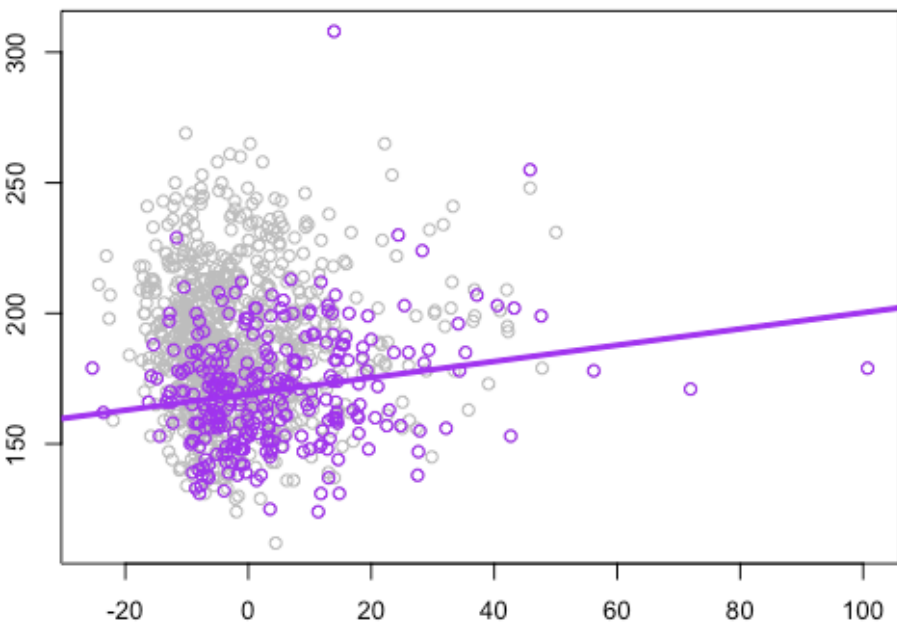
Winter



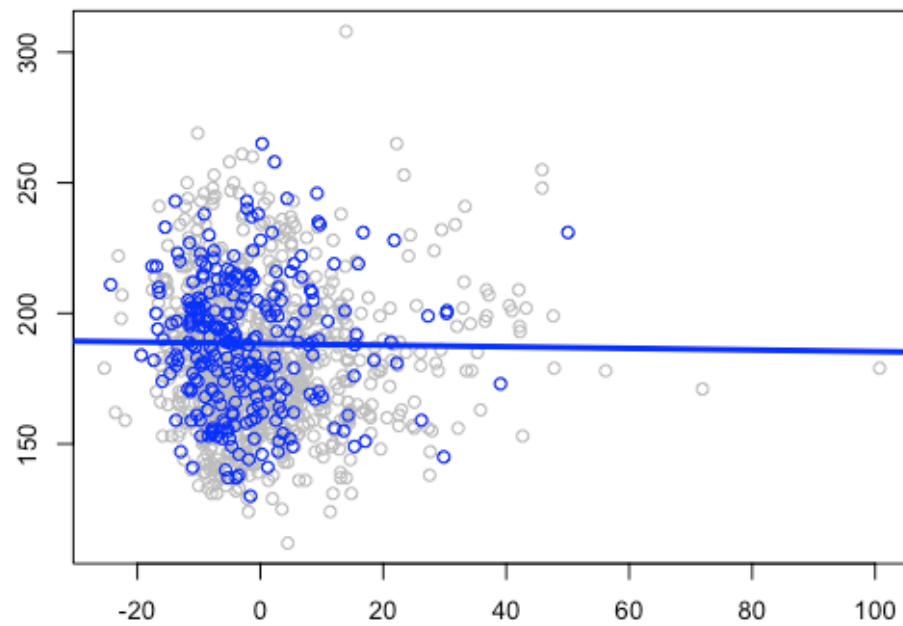
Spring



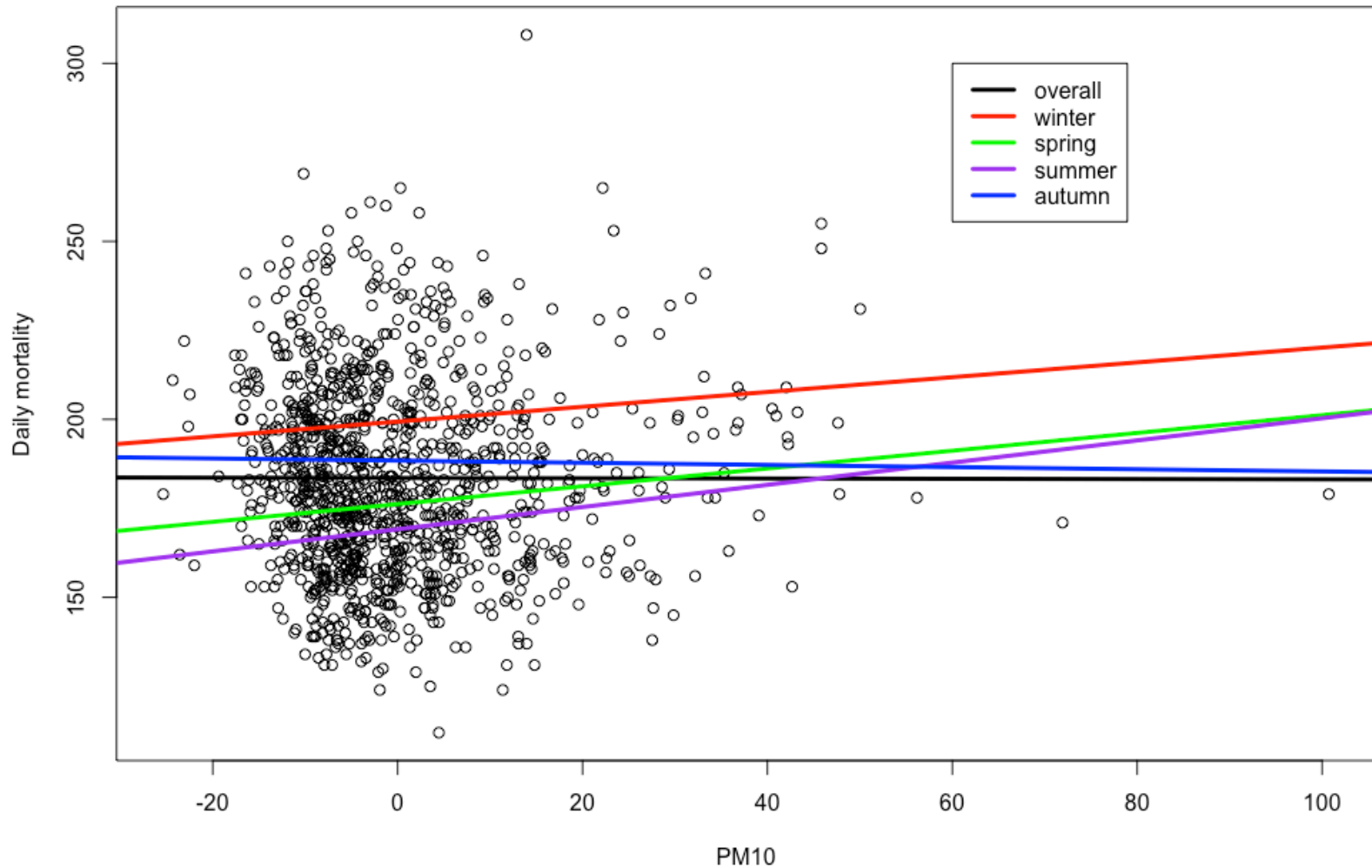
Summer



Autumn



NMMAPS data – New York (seasonal)



Log-linear (Poisson) regression

- Useful if the outcome Y counts the number of events in a fixed time period
- Let $\mu = E[Y|X]$ = expected or mean “rate” of events per day in the time period
- We can often model Y as a Poisson distribution with rate μ
- Model equation:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Interpreting coefficients

- Model equation: $\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$
- $\log(\mu | X_1 = 5) = \beta_0 + \beta_1 \cdot (5) + \beta_2 X_2 + \dots + \beta_p X_p$
- $\log(\mu | X_1 = 6) = \beta_0 + \beta_1 \cdot (6) + \beta_2 X_2 + \dots + \beta_p X_p$
- Difference in mean rates for $X_1 = 6$ compared to $X_1 = 5$, holding other variables fixed:

$$\begin{aligned} \log(\mu | X_1 = 6) - \log(\mu | X_1 = 5) \\ = (\beta_0 + 6\beta_1 + \beta_2 X_2 + \dots + \beta_p X_p) - (\beta_0 + 5\beta_1 + \beta_2 X_2 + \dots + \beta_p X_p) = \beta_1 \end{aligned}$$

- Log relative rate (log rate ratio):

$$\begin{aligned} \log(\text{relative rate}) &= \log\left(\frac{\mu|_{X_1=6}}{\mu|_{X_1=5}}\right) \\ &= \log(\mu | X_1 = 6) - \log(\mu | X_1 = 5) = \beta_1 \end{aligned}$$

- Relative rate: $e^{\log(\text{relative rate})} = e^{\beta_1}$

Assignment 2.1

1. Watch Dr. Roger Peng's guest lecture from Public Health Biostatistics:
(watch through time 35:35 at the following link)
<https://jh.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=33503a4e-7f30-4d4e-83cf-a9910149ad37>
 2. Access 2-4 of the 8 city data sets on Github to use in your analysis, one for each member of your group.
 3. For each of your chosen cities, make a time series display of PM10, temperature, and total mortality versus date. You want the display for each city to be a single page graphic, rather than separate graphics for each variable.
 4. For each city, fit the following three log-linear (Poisson) models:
 - Model A: $\text{death} \sim \text{pm10}$
 - Model B: $\text{death} \sim \text{pm10} + \text{as.factor}(\text{season})$
 - Model C: $\text{death} \sim \text{pm10} + \text{as.factor}(\text{month})$
 5. Make a tabular display that compares the estimated log relative mortality rate for PM10 for these three models
- Work together in groups!
 - Submit your assignment in R markdown through Blackboard by Sunday @ midnight.