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

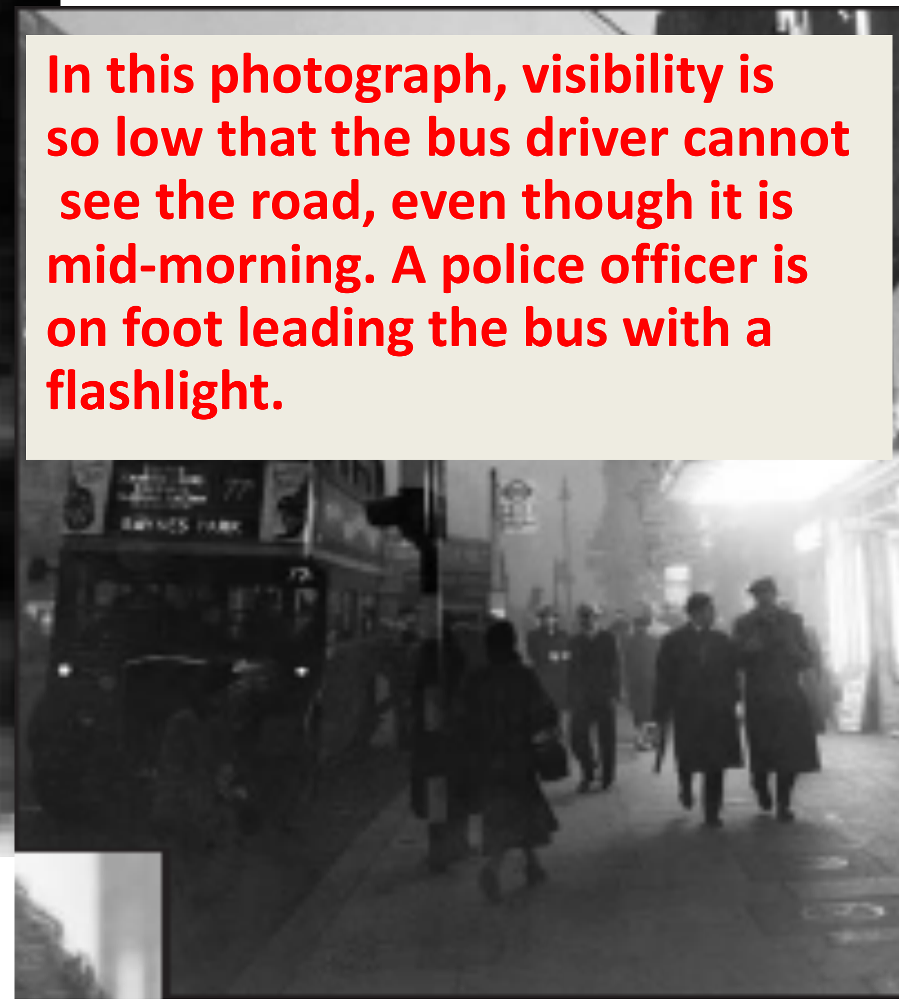
Module 2: Particulate air pollution and mortality

- **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

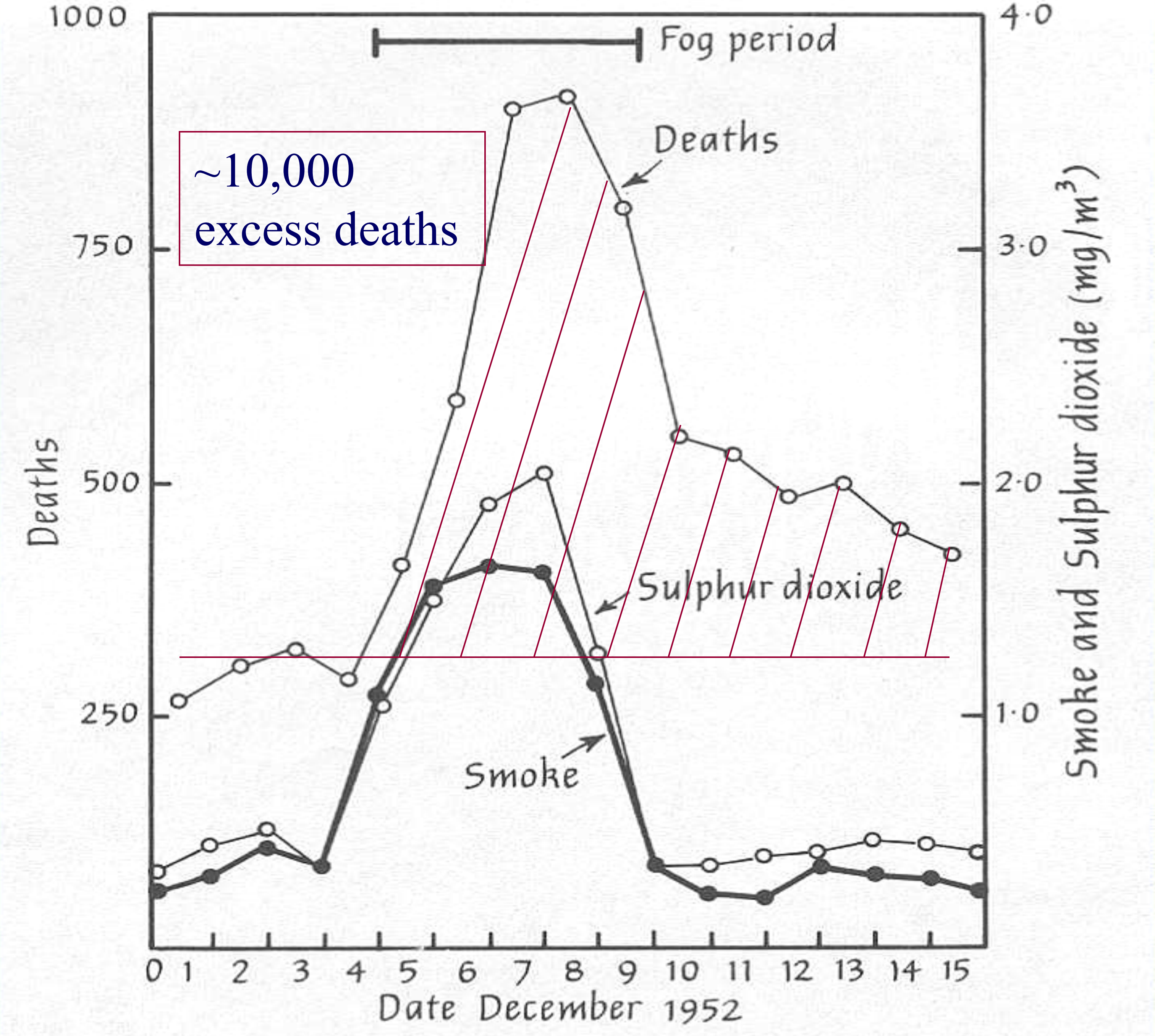
London, 12 noon, December, 1952; Particulate levels – 3,000 µg/m^3

Particulate air pollution and mortality



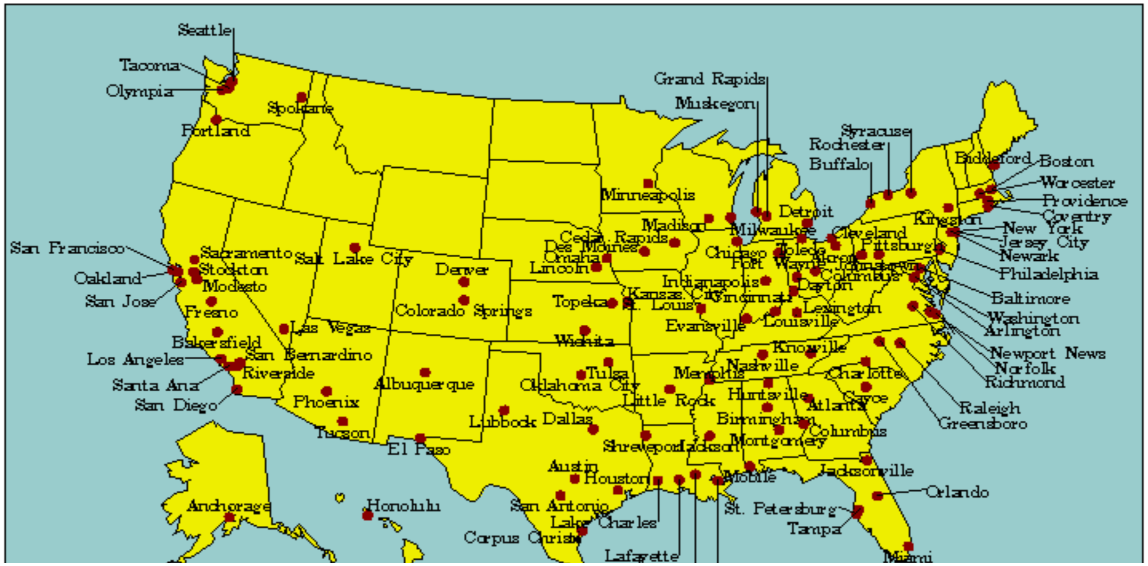
In this photograph, visibility is so low that the bus driver cannot see the road, even though it is mid-morning. A police officer is on foot leading the bus with a flashlight.

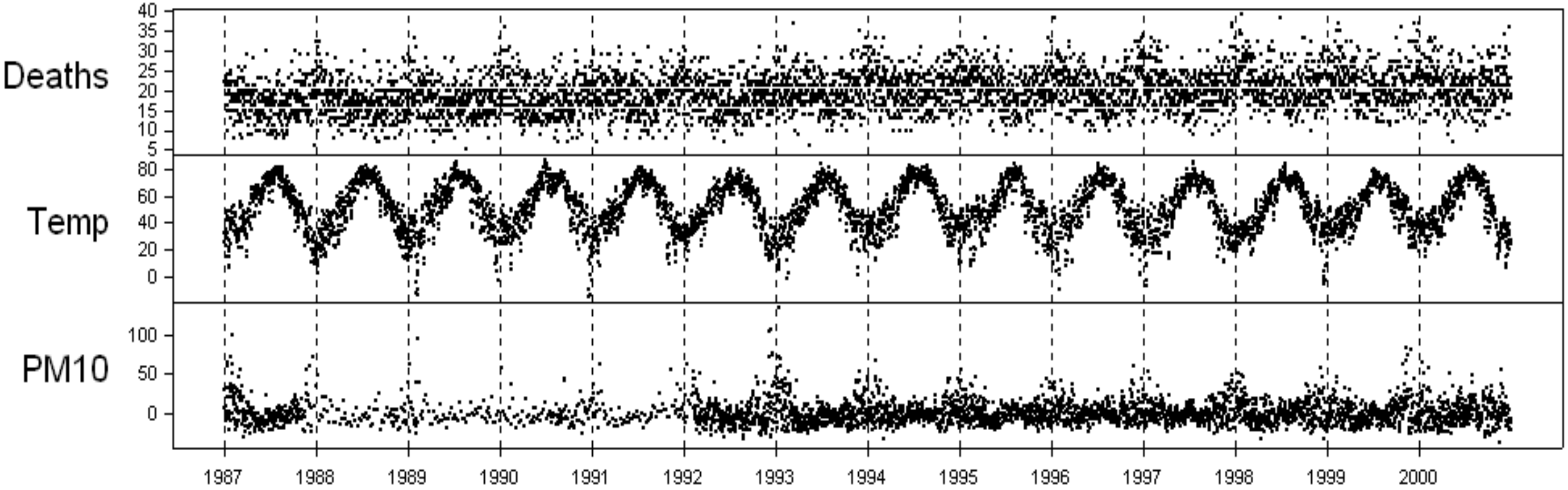
London, 12 noon, December, 1952; Particulate levels – 3,000 μg/m^3

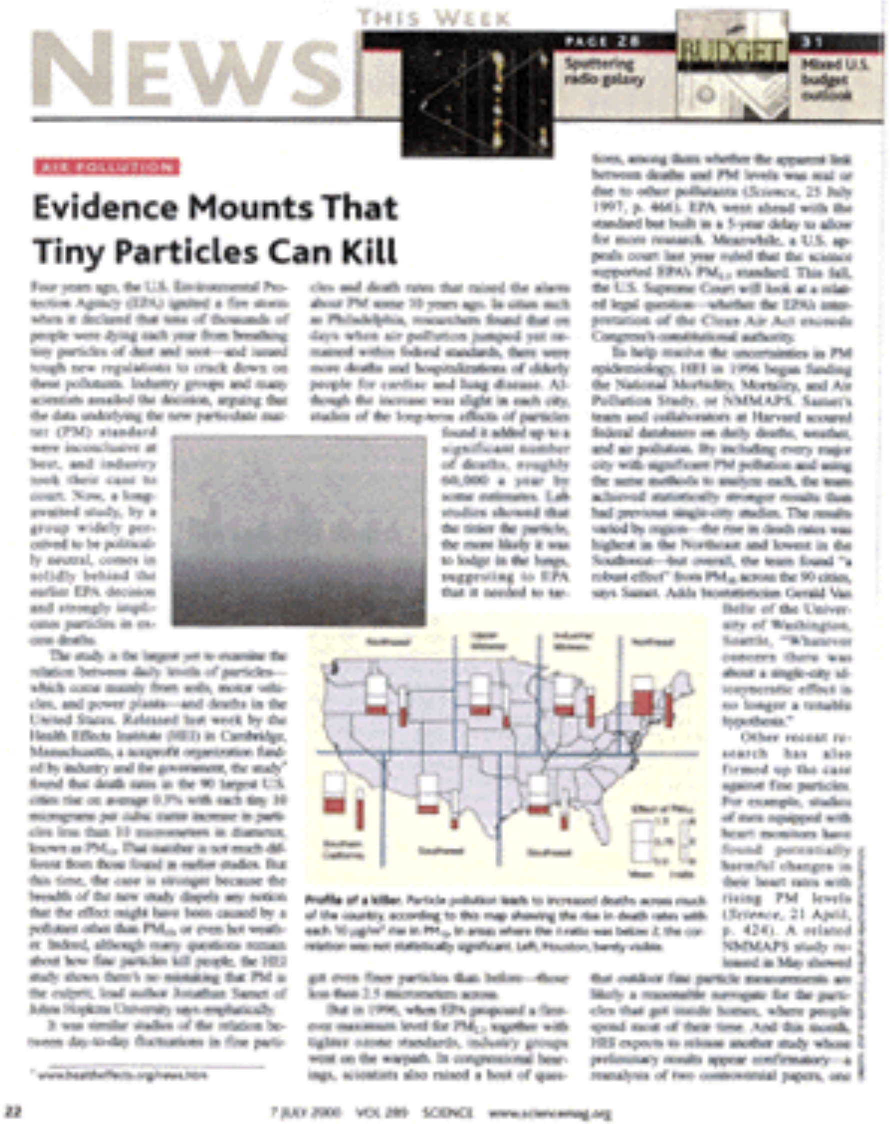


~10,000
excess deaths
Fog period
Deaths
Sulphur dioxide
Smoke
1000
750
500
250
4·0
3·0
2·0
1·0
Smoke and Sulphur dioxide (mg/m³)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Date December 1952

NMMAPS data


Seattle
Tacoma
Olympia
Spokane
Portland
San Francisco
Oakland
San Jose
Sacramento
Stockton
Modesto
Fresno
Bakersfield
Los Angeles
Santa Ana
San Diego
San Bernardino
Riverside
Las Vegas
Salt Lake City
Phoenix
Tucson
Denver
Colorado Springs
Albuquerque
El Paso
Anchorage
Honolulu
Minneapolis
Grand Rapids
Muskegon
Madison
Milwaukee
Cedar Rapids
Des Moines
Omaha
Lincoln
Topeka
Kansas City
St. Louis
Wichita
Tulsa
Oklahoma City
Lubbock
Dallas
Austin
Houston
San Antonio
Lake Charles
Corpus Christi
Lafayette
Little Rock
Shreveport
Jackson
Memphis
Nashville
Knoxville
Huntsville
Birmingham
Montgomery
Mobile
Atlanta
Columbus
Jacksonville
Orlando
St. Petersburg
Tampa
Miami
Chicago
Detroit
Toledo
Fort Wayne
Indianapolis
Cincinnati
Dayton
Evansville
Louisville
Lexington
Cleveland
Akron
Columbus
Pittsburgh
Charlotte
Sayce
Raleigh
Greensboro
Syracuse
Rochester
Buffalo
Biddeford
Boston
Worcester
Providence
Coventry
Kingston
New York
Jersey City
Newark
Philadelphia
Baltimore
Washington
Arlington
Newport News
Norfolk
Richmond

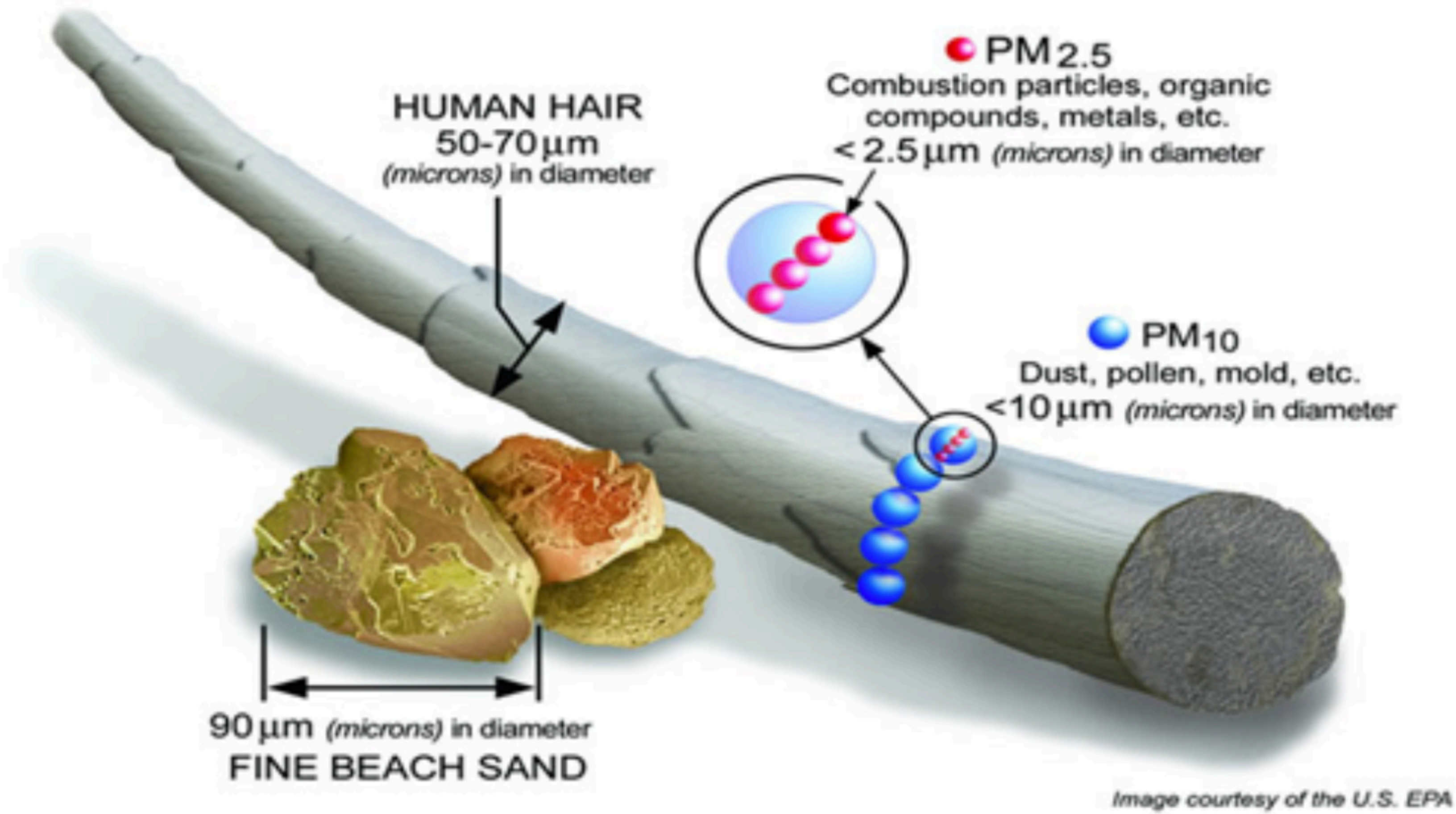

Deaths
Temp
PM10
40
35
30
25
20
15
10
5
80
60
40
20
0
100
50
0
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000



"(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?

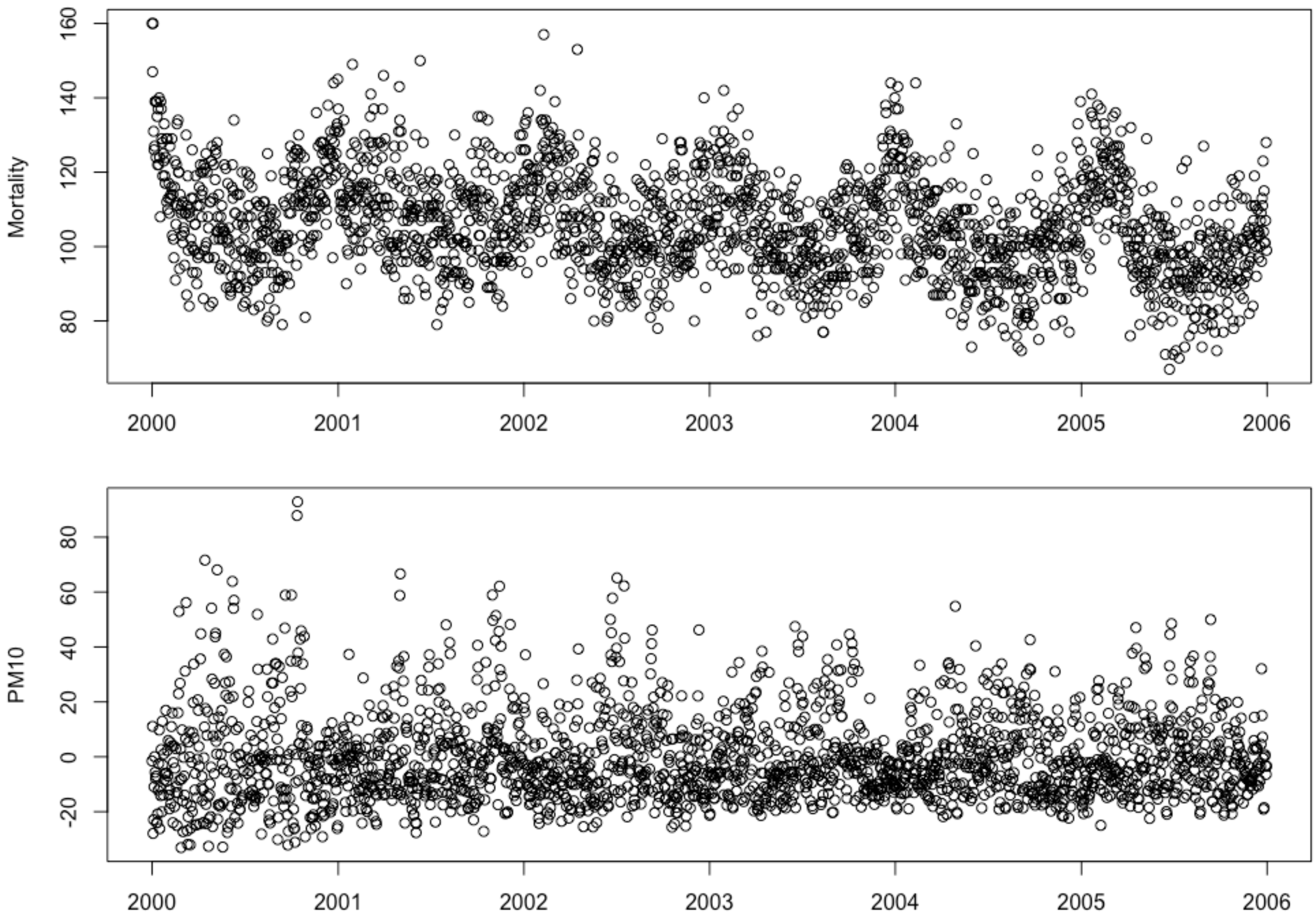


http://www.irceline.be/en/documentation/faq/what-is-pm10-and-pm2.5

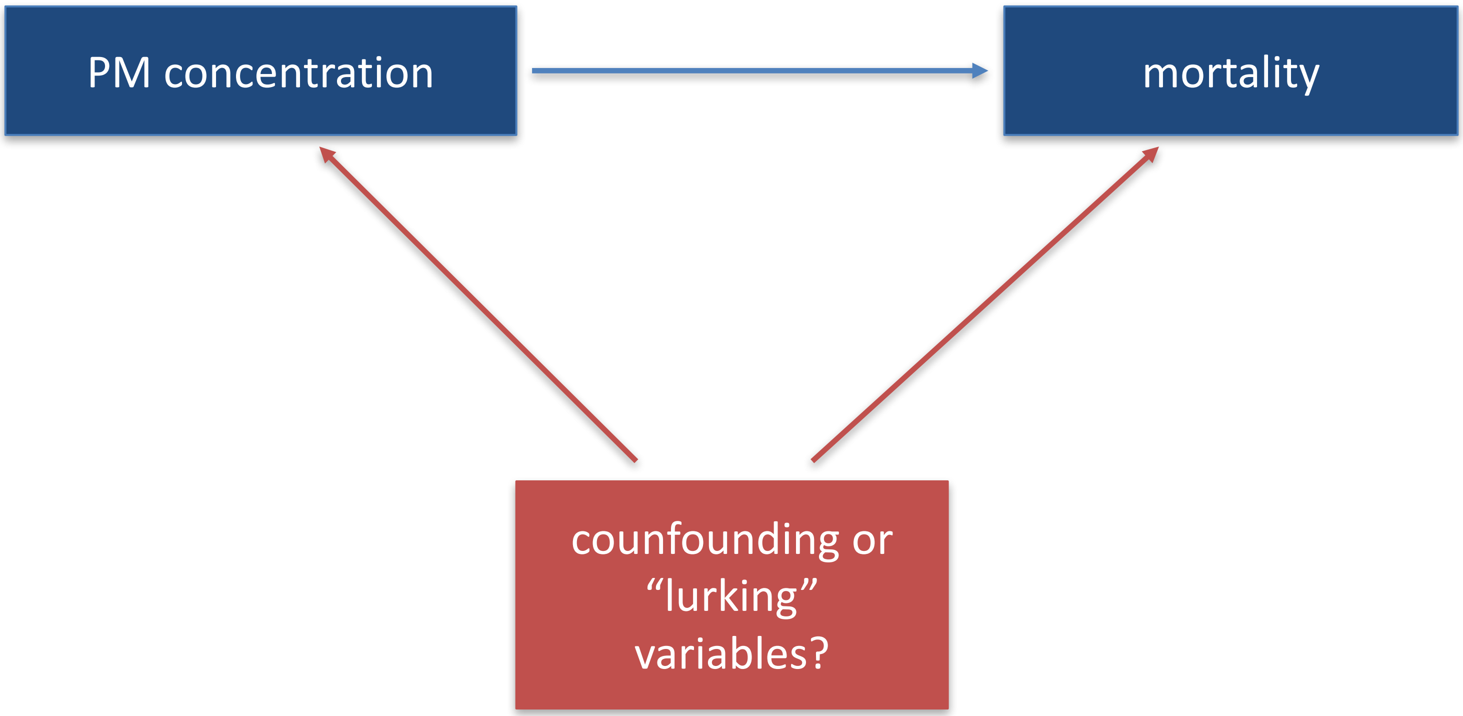
NMMAPS data -- Chicago

```
> 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
```

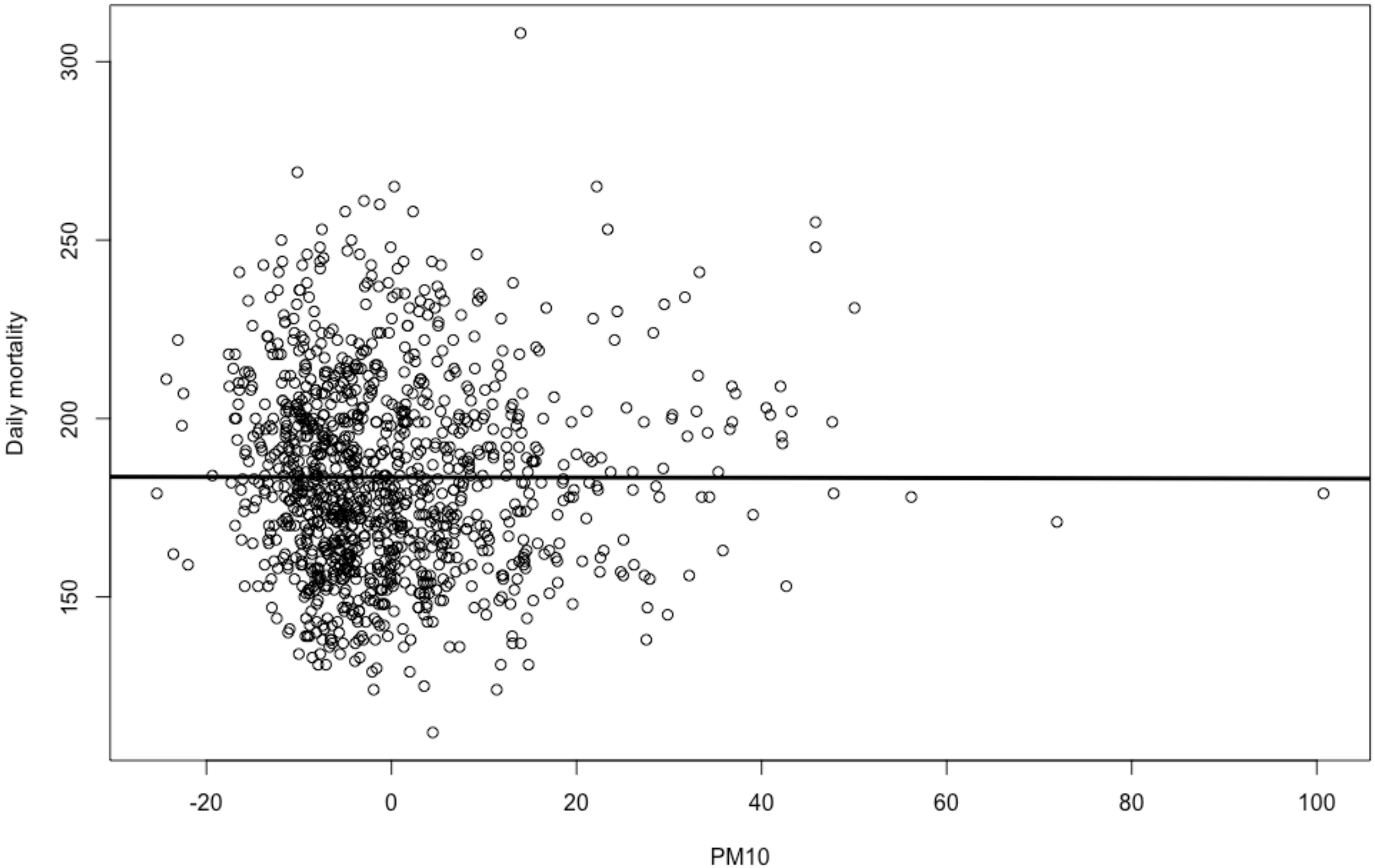
NMMAPS data -- Chicago



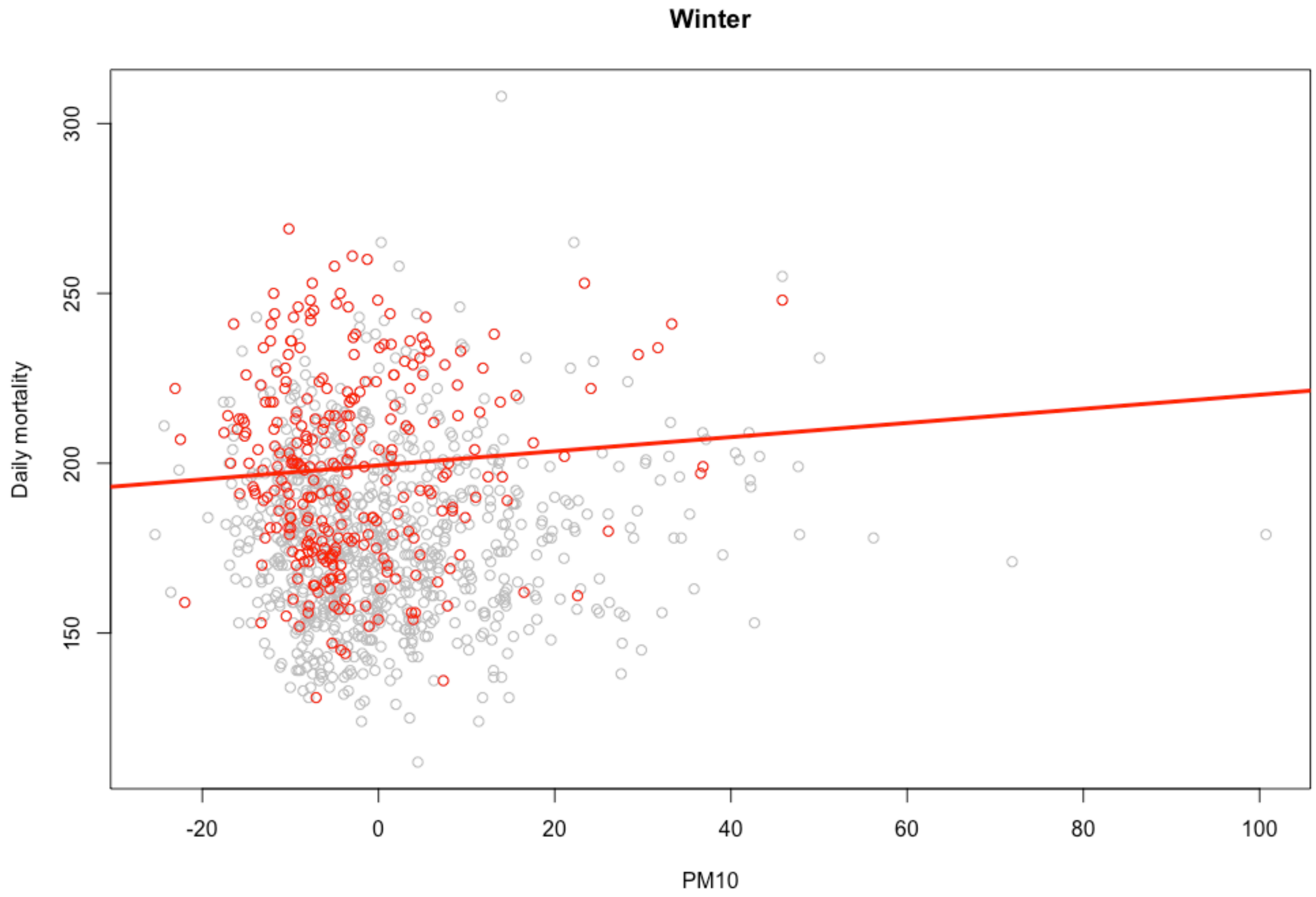
DAG of relationship

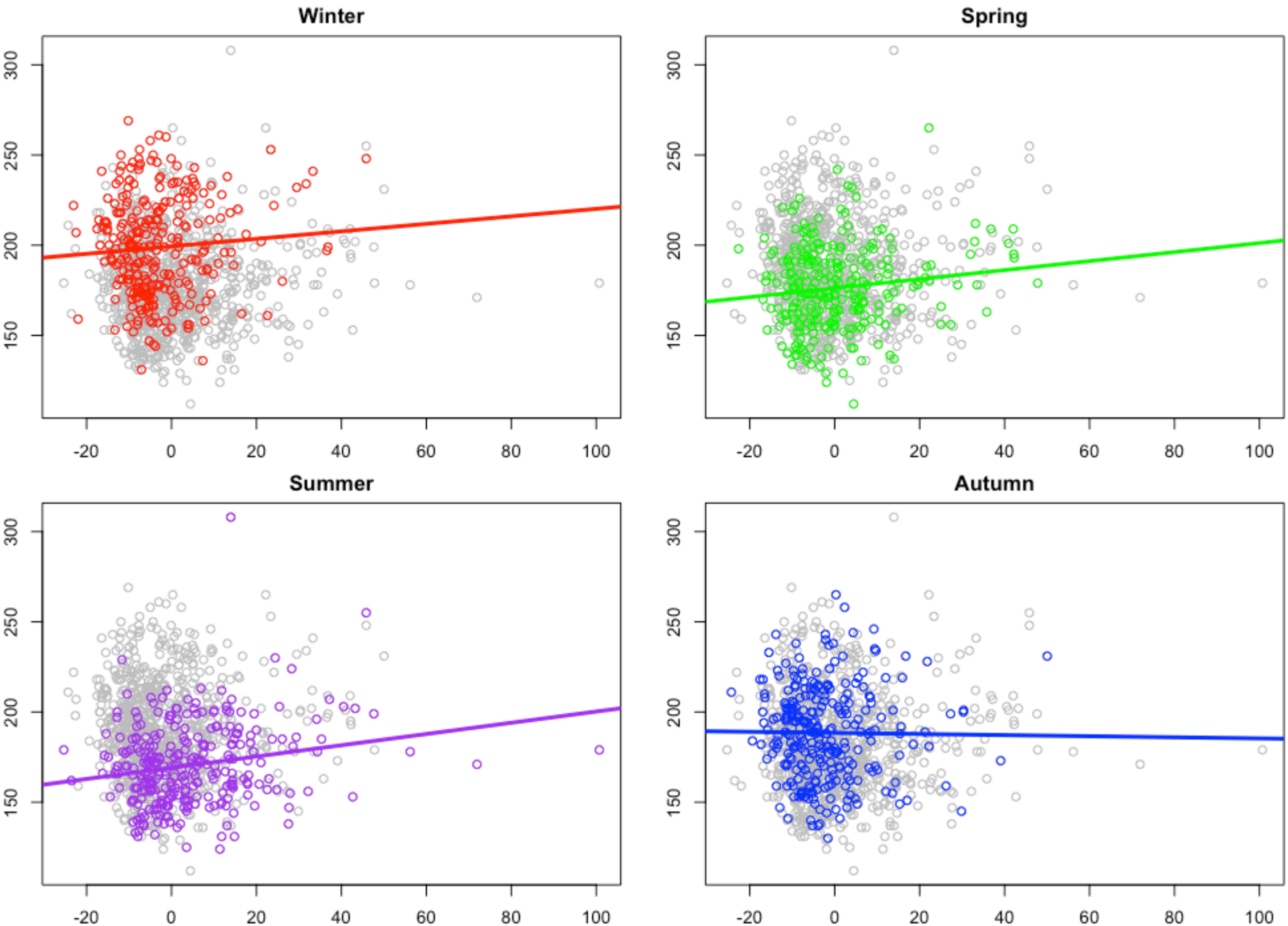


NMMAPS data – New York



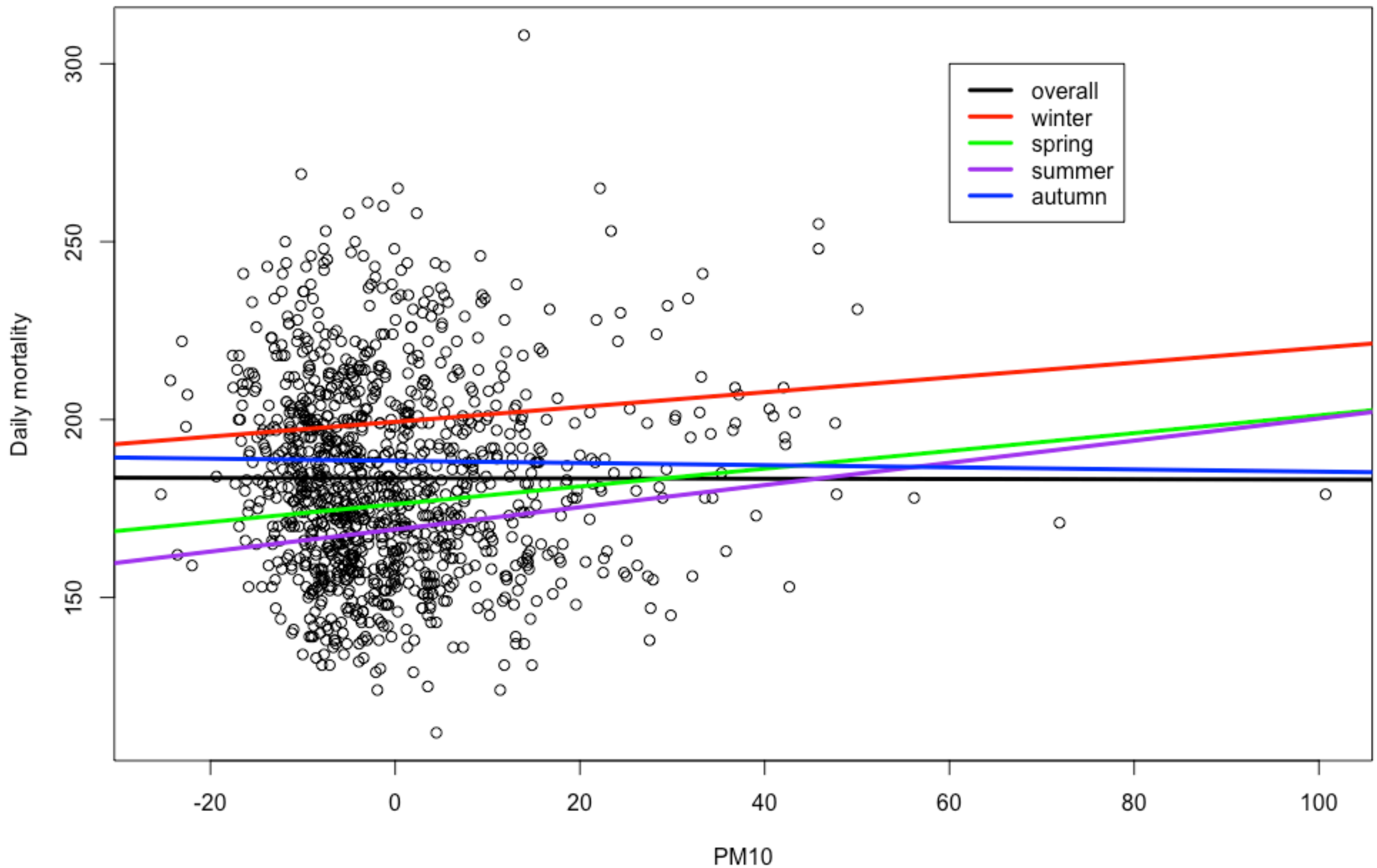
NMMAPS data – New York (winter)





Winter
Spring
Summer
Autumn
300
250
200
150
-20
0
20
40
60
80
100

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: $\boldsymbol{log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$
- $\log(\mu|\, X_1 = 5) = \beta_0 + \boldsymbol{\beta_1} \cdot (\mathbf{5}) + \beta_2 X_2 + \cdots + \beta_p X_p$
- $\log(\mu|\, X_1 = 6) = \beta_0 + \boldsymbol{\beta_1} \cdot (\mathbf{6}) + \beta_2 X_2 + \cdots + \beta_p X_p$

- Difference in mean rates for $X_1 = 6$ compared to $X_1 = 5$, holding other variables fixed:

$$\log(\mu|\, X_1 = 6) - \log(\mu|\, X_1 = 5)$$
$$= (\beta_0 + \mathbf{6}\boldsymbol{\beta_1} + \beta_2 X_2 + \cdots + \beta_p X_p) - (\beta_0 + \mathbf{5}\boldsymbol{\beta_1} + \beta_2 X_2 + \cdots + \beta_p X_p) = \boldsymbol{\beta_1}$$

- Log relative rate (log rate ratio):

$$\log(\text{relative rate}) = \log\left(\frac{\mu|X_1=6}{\mu|X_1=5}\right)$$
$$= \log(\mu|\, X_1 = 6) - \log(\mu|\, X_1 = 5) = \boldsymbol{\beta_1}$$

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

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: death ~ pm10
   - Model B: death ~ pm10 + as.factor(season)
   - Model C: death ~ pm10 + as.factor(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.