AS.280.347
CLASS 2.1

- Review your work!
- Interpreting log-linear models
- Natural splines

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)

Assignment 2.1 – your work!

- 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.

- 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)

- Make a tabular display that compares the estimated log relative mortality rate for PM10 for these three models

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?

1. What is a reasonable causal graph (DAG) for this question?

2. What are the key confounders (that cause death and are related to, but not caused by, air pollution)?

3. How can the data be displayed for one city to “see” the air pollution effect controlling for season and/or year?

Module 2: Particulate air pollution and mortality

Question 2.2 (Q2.2): Is the estimate of the pollution effect sensitive to assumptions about seasonal or weather effects?

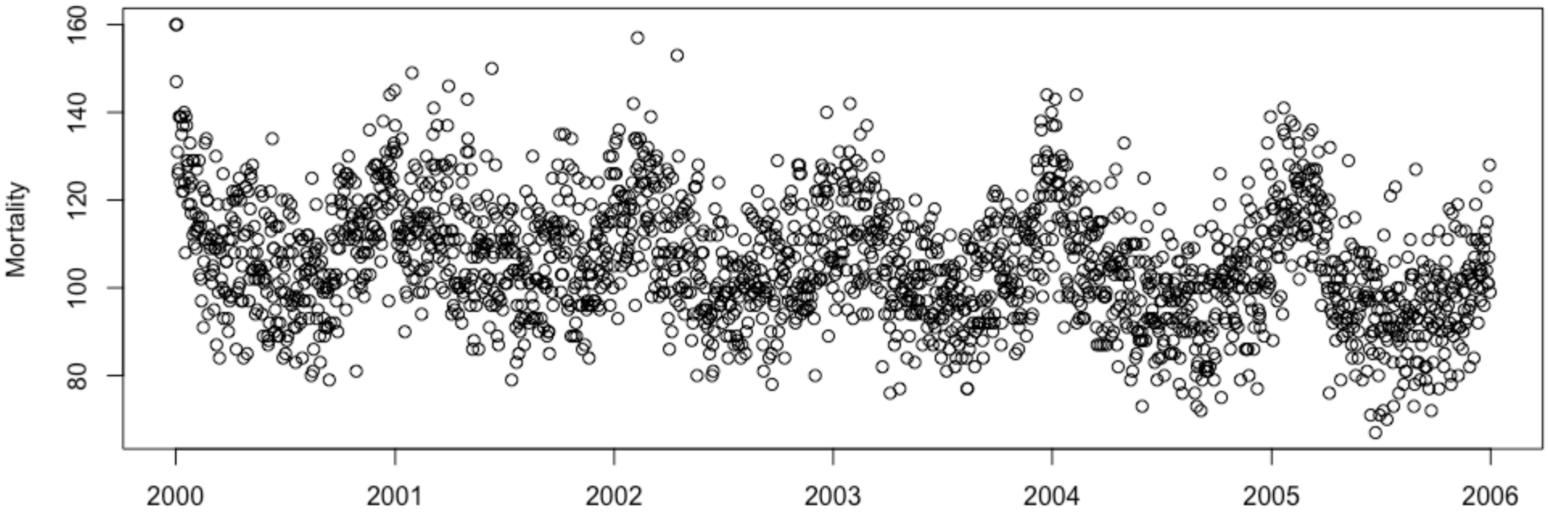
1. What statistical analysis will help us answer this question?

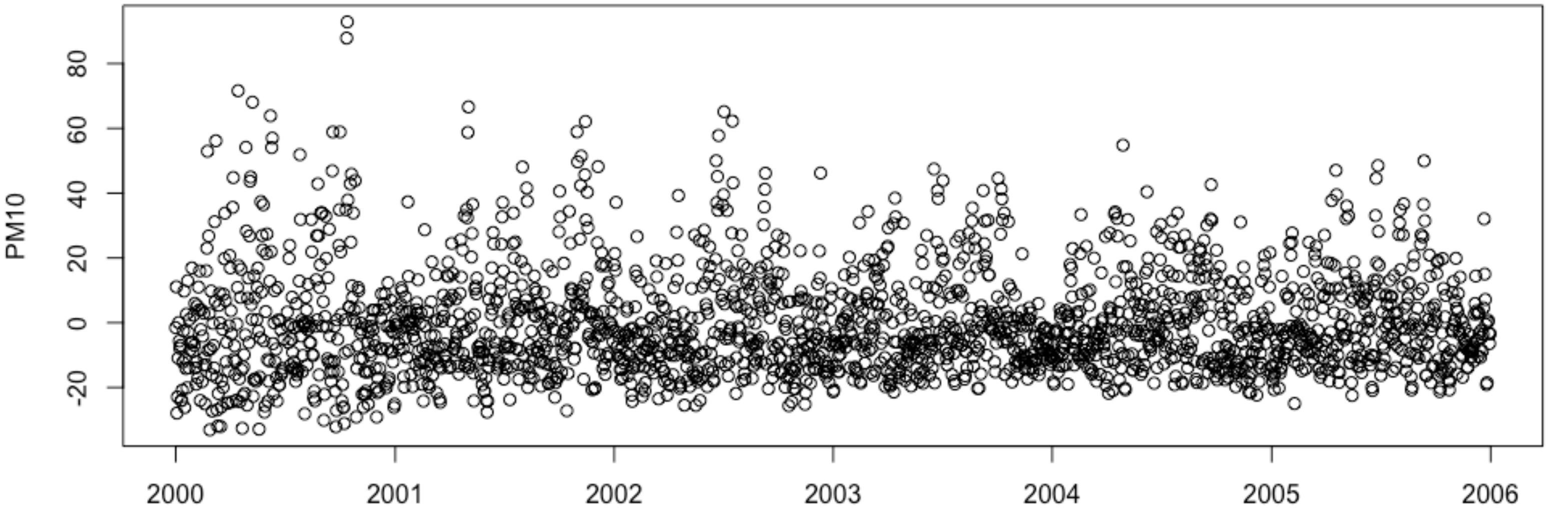
2. What displays might be use to answer this question?

Design of statistical analysis

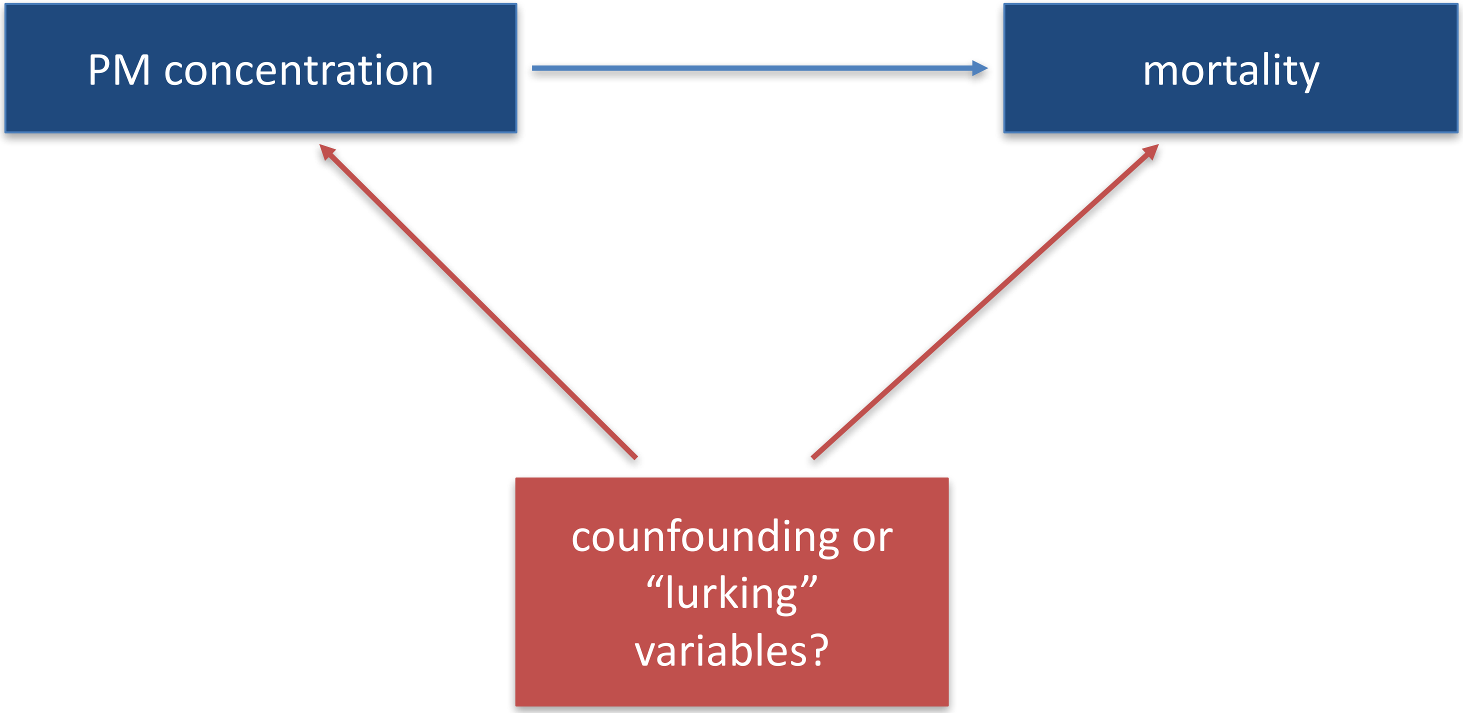
- Question
- Outcome
- Design
- Data analysis
  - Primary analysis plan
  - Communicating results in tables, figures
  - Model checking for robustness of findings to key assumption, data perturbations
- Interpretation
- Dissemination

NMMAPS data -- Chicago





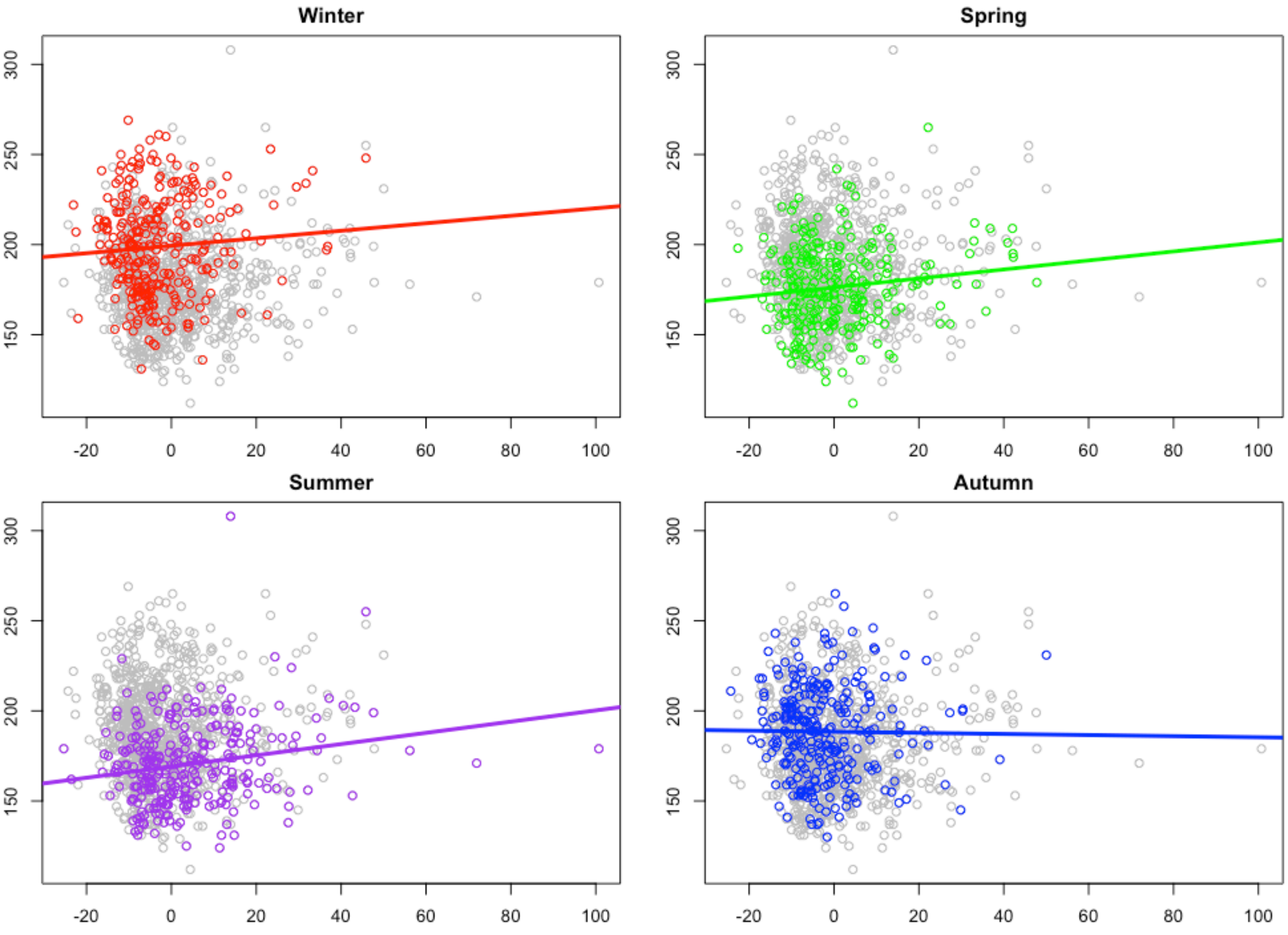
DAG of relationship



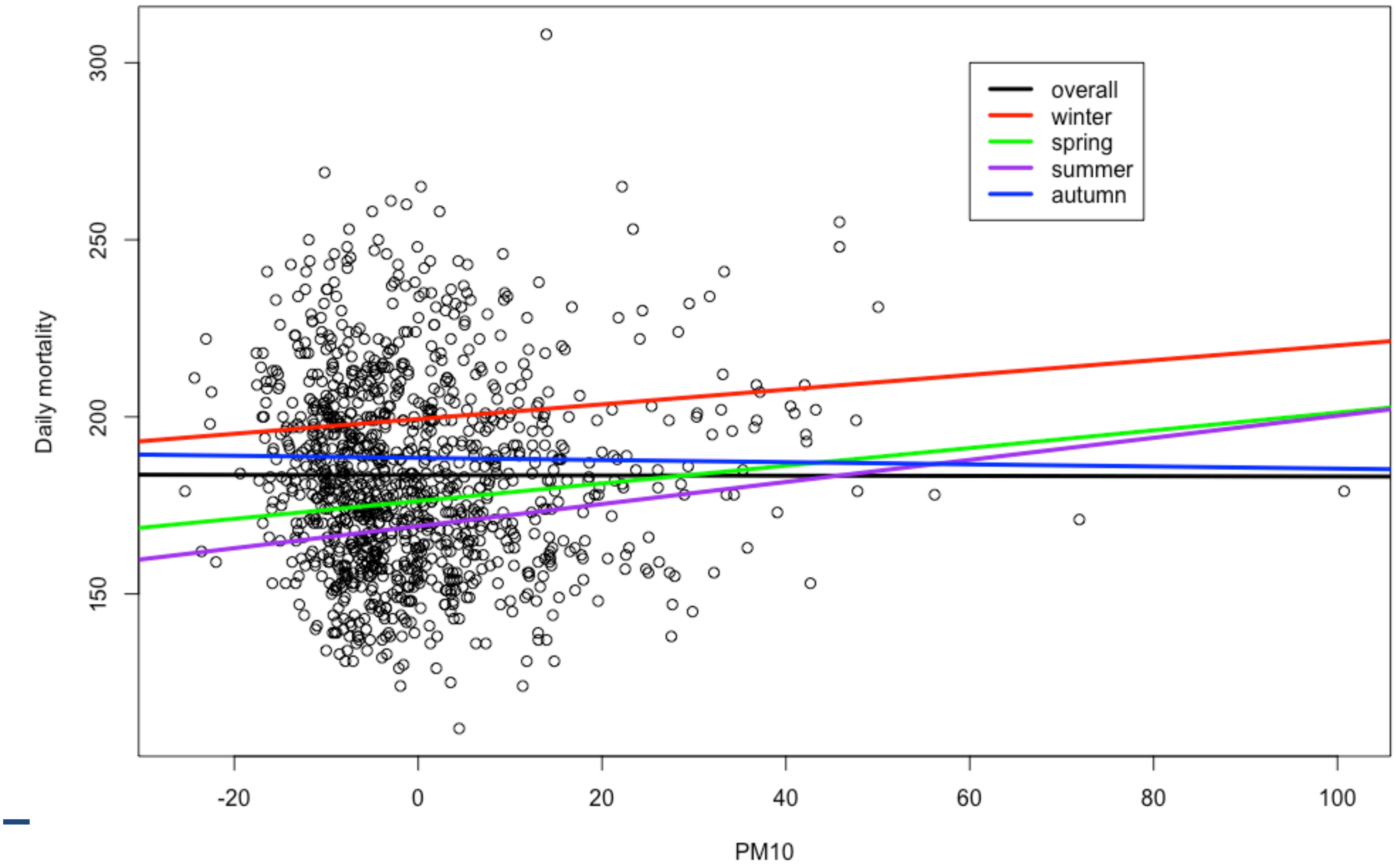
PM concentration

mortality

counfounding or “lurking” variables?


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:

$$\log(\mu|\, X_1 = 6) - \log(\mu|\, X_1 = 5)$$

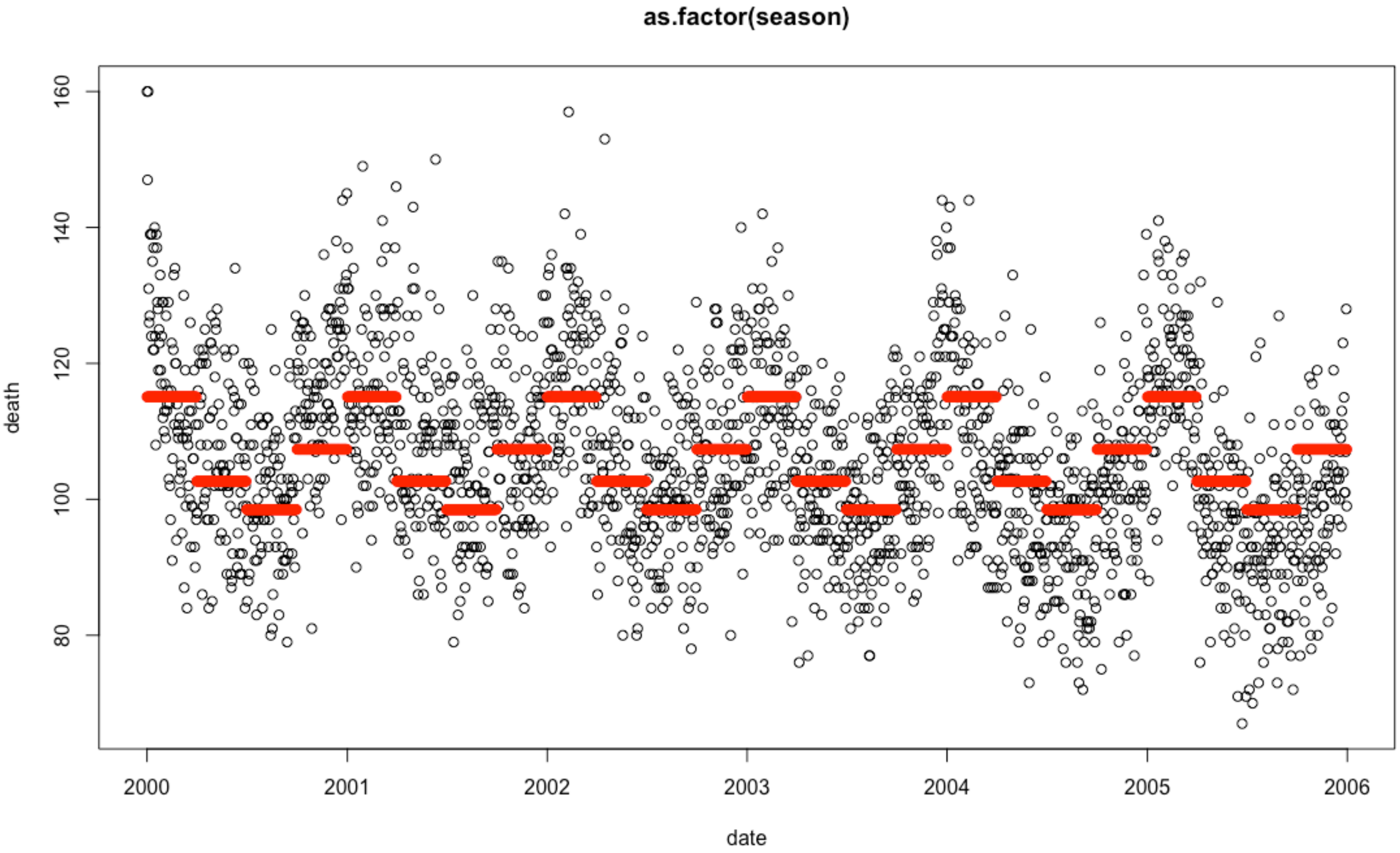
$$= (\beta_0 + \mathbf{6}\boldsymbol{\beta_1} + \beta_2 X_2 + \cdots + \beta_p X_p) - (\beta_0 + \mathbf{5}\boldsymbol{\beta_1} + \beta_2 X_2 + \cdots + \beta_p X_p) = \boldsymbol{\beta_1}$$

- Log relative rate (log rate ratio):

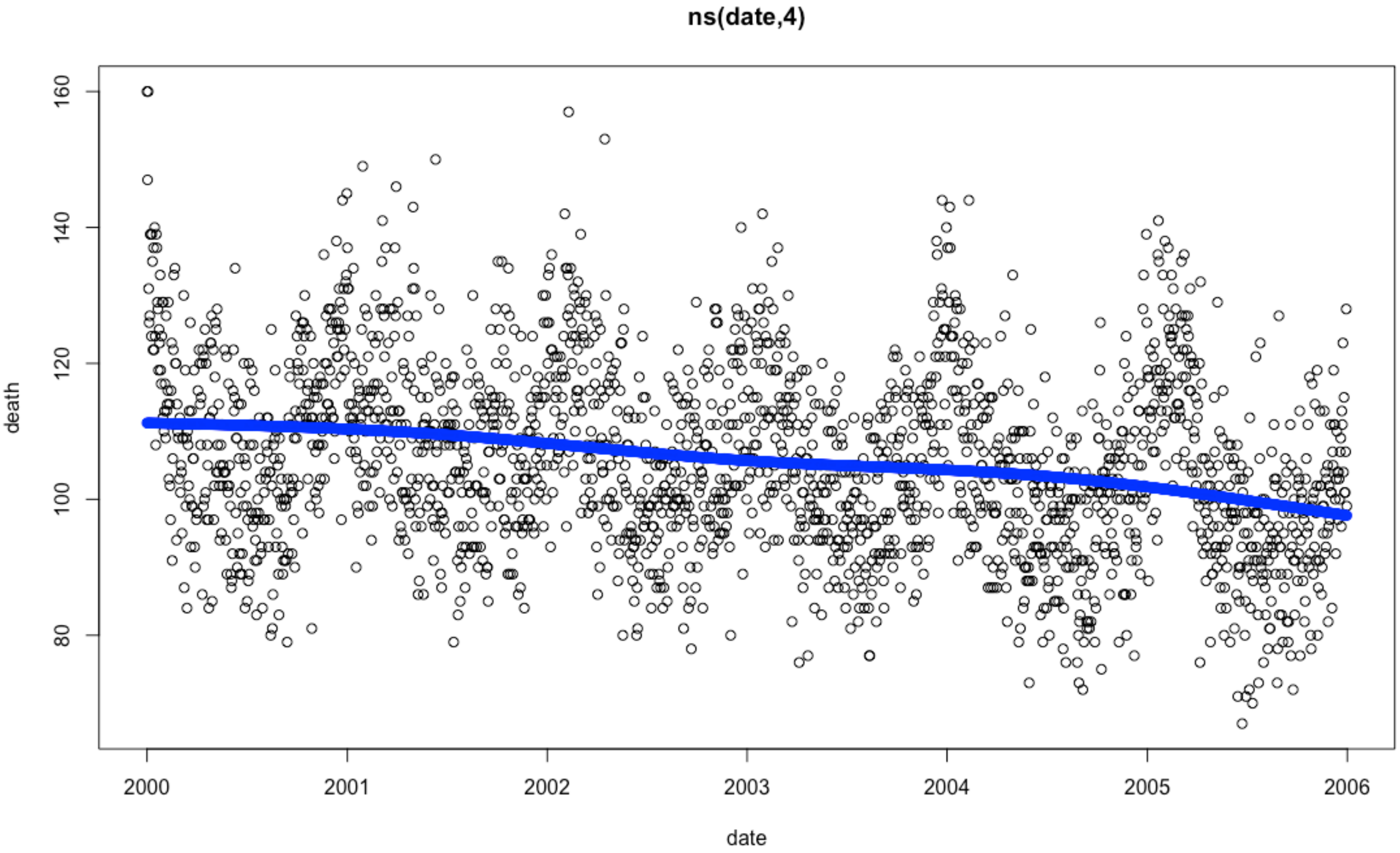
$$\log(\text{relative rate}) = \log\left(\frac{\mu|X_1=6}{\mu|X_1=5}\right)$$
$$= \log(\mu|\, X_1 = 6) - \log(\mu|\, X_1 = 5) = \boldsymbol{\beta_1}$$

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

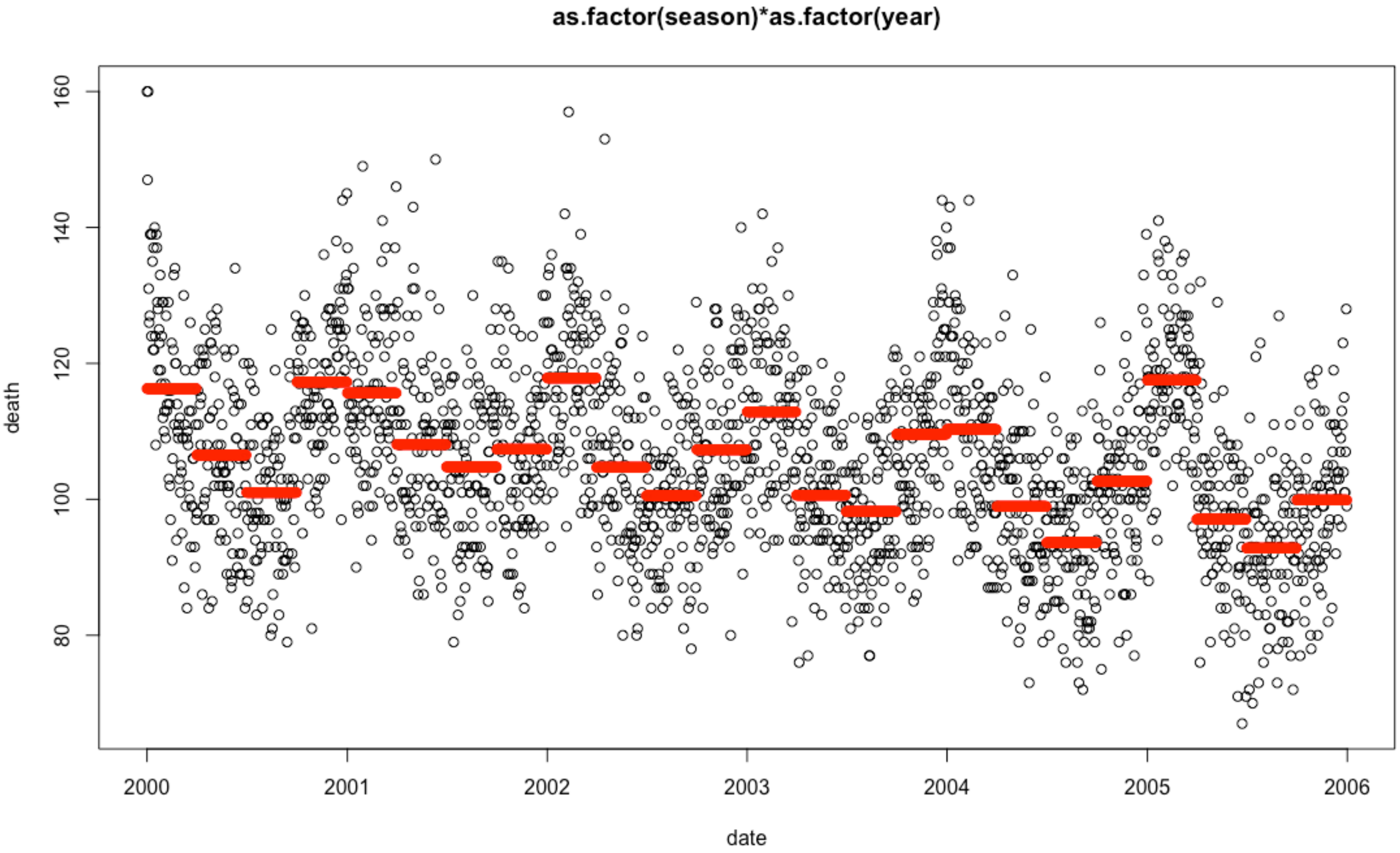
Modeling seasonal effects



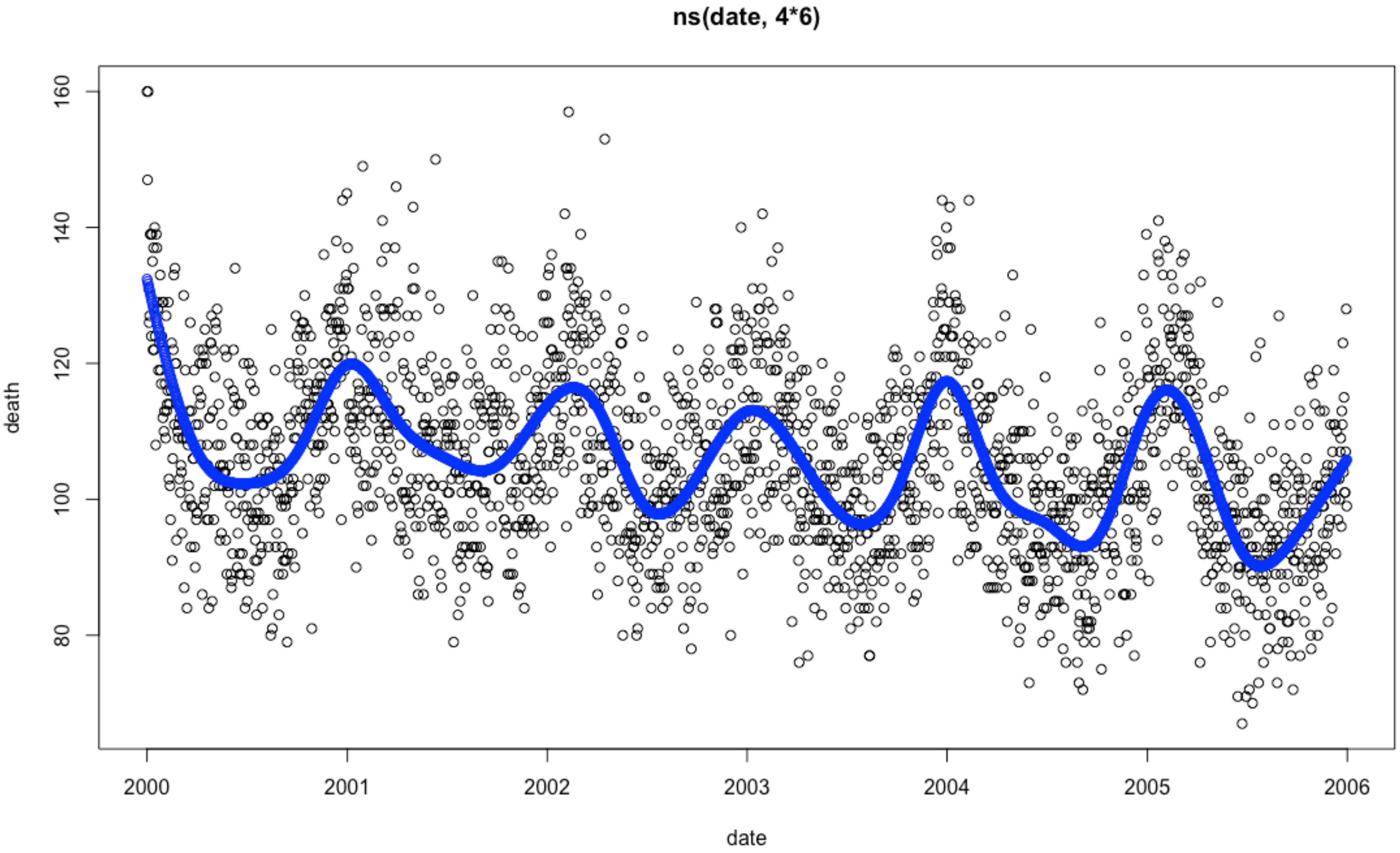
Modeling seasonal effects



Modeling seasonal effects



Modeling seasonal effects



Assignment 2.2

1. Update your time series display of PM10, temperature, and total mortality versus date.

2. For each city, regress mortality on PM10 using a log-linear (Poisson) model with different indicator variables for time:
   - A: death ~ pm10 0 degrees of freedom (df)
   - B: death ~ pm10 + as.factor(season) 4-1 = 3 df
   - C: death ~ pm10 + as.factor(month) 12-1 = 11 df
   - D: death ~ pm10 + as.factor(season)*as.factor(year) 4*19-1 = 75 df
   - E: death ~ pm10 + as.factor(month)*as.factor(year) 12*19-1 = 227 df

3. Plot mortality against time and add a continuous line for each predicted model.

4. Display the five PM10 coefficients (A-E) with confidence intervals in a table or graph to see the effect of the method of control for seasonality.

5. Repeat the regression models (A-E) with natural splines to give a smooth relationship between mortality and time using ns(time, df) for df = 0, 3, 11, 75, 227. Plot the data and predicted curves once again.

- Work together in groups!
- Submit your assignment in R markdown through Blackboard by Sunday @ midnight.