Causal Modeling in Environmental Epidemiology

Joel Schwartz

Harvard University

When I was Young…

What do I mean by Causal Modeling?

- What would have happened if the population had been exposed to a' instead of being exposed to a
- If that is different that what happened when they were exposed to a, there is a causal effect of changing exposure.

So the key issue is

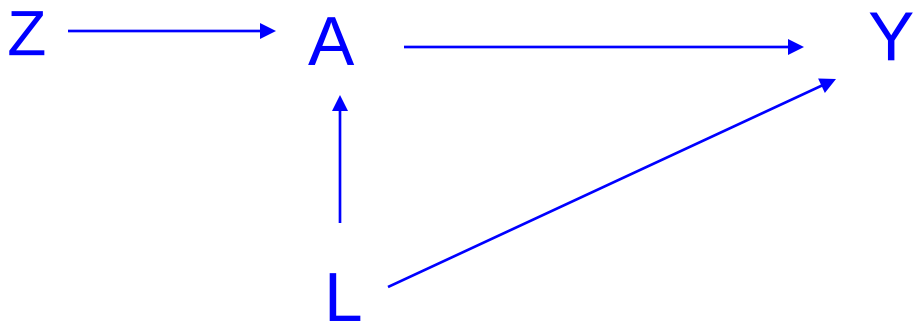
- The Average Treatment Effect is the difference between two unobserved outcomes—
  - What would have happened if everyone had exposure a, versus what would have happened if everyone had exposure a’
- Clearly, we need to estimate valid substitutes for these two potential outcomes

- Generally we have some people with exposure a and others with exposure a′
- We would like it to be the case that those who had exposure a and those who had exposure a′ are comparable (in their potential outcomes)
- If that were the case then the outcomes of those who had treatment a would be similar to the outcomes if the whole population had been given treatment a
- And the outcomes of those who had treatment a′ would be similar to the outcomes if the whole population had been given treatment a′
- How do we get there?

Notation

- Let $Y^{A=a}$ be the effect that **would have been observed** under treatment A=a (say Y under a=a)
- Let $Y^{A=a'}$ be the effect that **would have been observed** under treatment A=a'
- If these are different, there is a causal effect
- Causal Risk Ratio= $Y^{A=a} / Y^{A=a'}$
- Causal Difference= $Y^{A=a} - Y^{A=a'}$

Causal Diagram

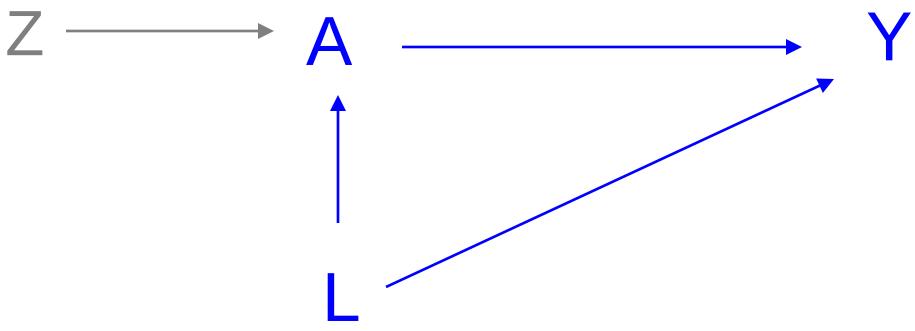


Common Cause (L) --->Confounding
Z (Random Assignment) Breaks Confounding
Z is not correlated with L

Randomized Experiments

- $Y^{a=1}$ for the people who happened to be treated is the same as $Y^{a=1}$ for the people who **were not treated**
- Pr(Y=1|A=1) is a consistent estimate of $Y^{a=1}$
- We believe this is true because randomization means that the distribution of confounders is the same in treated and untreated people
- But I am an observational epidemiologist

What about Observational Data?



Assignment to A may be different by levels of L

Idea of Causal Modeling

- Take Observational Data
- Make it look like a Randomized Control Trial
- Then we can believe it is causal
- Association is defined as a comparison of different subjects under different conditions
- A causal effect defines a comparison of the same subjects under different actions
  - How can we do this?
    - Make the distribution of covariates (other factors) among people exposed to a the same as among people exposed to a'.

Consider the NHANES Survey

- Suppose you want to look at the distribution of Blood Pressure in the Population
- You would like to know what it is like in different ethnic groups (Blacks, Hispanics, Asian)
- Blacks were 13% of the population at the time of NHANES II, Asians were 6%
- Will there be a large enough sample to understand the distribution?

Solution

- Oversample Blacks (etc)
- Suppose we sample Blacks at twice their frequency in the Population
- Good News- We can now look at the distribution of Blood Pressure in Blacks
- Bad News- How do we compute the Distribution in the Population with a Non-representative Sample?

We have a missing data problem

- We sampled twice as many Blacks, but not twice as many Whites (and others)
- We are missing those extra Whites
- However, if the Whites we did sample were a representative sample
- We can weight them twice as much as the Blacks and return to a Representative Sample of the Whole Population

That is

- We are assuming that CONDITIONAL on RACE
- People are Missing At Random
- We weight each subject by 1/probability of being selected

Observational Epidemiology

- Is like a Randomized Trial
- But with missing data
- If they were not missing, then the distribution of exposure would not be correlated with confounders
- For example, if exposure is correlated with age. We have confounding
- We are “Missing” people with low age and high exposure
- IF we can solve that, we have a causal model

There are two approaches

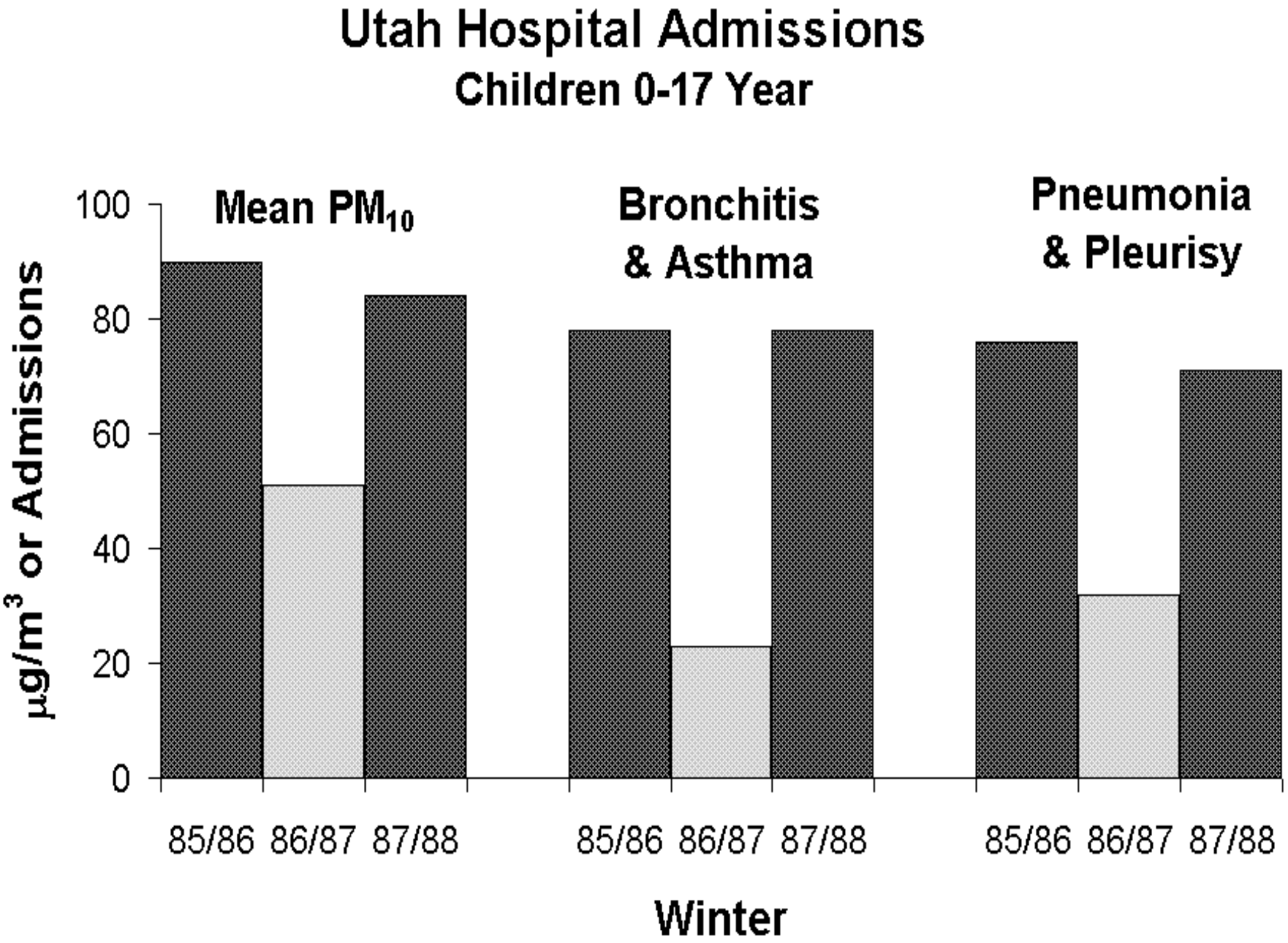
- Modeling
  - e.g. Inverse Probability Weighting, Propensity Scores
  - This makes the exposure independent of measured confounders, so $E(Y^{a=1}|A=1,C)=E(Y|A=1)$
  - Untestable Assumption: We measured all the confounders
- Quasi –Experimental
  - Find natural patterns that appear to randomize exposure, or part of the variation in exposure, with respect to confounders including unmeasured ones
  - Untestable Assumption: It is really random

The second approach is less common in Environmental Epidemiology

- So I will talk about it first
- I call it pseudo randomization because it involves directly or indirectly identifying things that manipulated exposure in ways that we believe are random with respect to other predictors of outcome.
- Three Examples
  - Natural Experiments
  - Differences in Differences
  - Instrumental Variables

Natural Experiments

- Lleras-Muney (2010) looked at all U.S. military families stationed in the United States over 11 years.
- About 1/3 of families are relocated each year to new bases, based on military need, presumably unrelated to air pollution levels or health.
- 114,612 children < age 2-5, 1.2% hospitalized each year for respiratory disease
- An additive risk model showed 10 ppb of annual ozone increased hospitalization by an additional 0.35% (95% CI 0.05, 0.65%), that is, from 1.2% to 1.55%



Pope, Amer J Public Health 1989; 79: 623

EZ Pass Study

- Toll booths slow traffic, increase air pollution
- Currie and Walker studied the introduction of electronic tolling at highway speeds to replace toll plazas
- Traffic goes under radio detectors at 100km/hr without slowing down
- The order of the installations was at the convenience of the contractors

Currie, 2008

Differences in Differences

- Walker and Currie considered two groups, those living within 1.5km of the toll plaza, and those living near the highway, but further (2-10 km) from the plaza
- They controlled for individual and small area covariates, but what about omitted confounders?

A pre/post estimate

- in Group 1 (near the toll booths) will control for unmeasured confounders that change slowly over time within the neighborhood
- If the time varying omitted confounders change similarly in Group 0, then their change can serve as a control for those omitted confounders
- The difference between post-pre in Group 1 and post-pre in Group 2 is this difference in differences estimate

Potential Outcome Version

- Let Y^0 and Y^1 be the potential outcomes under treatment or not
- G is the group, in our case location (close/far from toll)
- T is time period, 0 or 1

We can write

$E[Y^0 | T,G,X] = \alpha T + \beta G + X\theta$

Treatment only occurs in time 1, group 1

$$E[Y^1 | T = 1, G = 1, X] = \alpha + \beta + \tau + X\theta$$

And So

$$E[Y|T, G, X] = \alpha T + \beta G + \tau TG + X\theta$$

That is, if we regress outcome against a dummy variable for time period, a dummy variable for group, and an interaction between time period and group
We obtain a causal estimate of the effect τ

PM2.5 in New Jersey

- Wang (2016) geocoded all deaths for six years in New Jersey to Census Tract
- We used satellite remote sensing, land use terms, and meteorology to estimate PM2.5 and temperature on a 1km grid.
- We matched these estimates to each census tract.
- Performed a Difference in Differences Analysis controlling for census tract and time

Results

- HR of 1.19 (95%CI: 1.11–1.28) per 10 $\mu g/m^3$ increase in the annual $PM_{2.5}$ concentrations
- That is, a 1.9% increase in mortality rate per 1 $\mu g/m^3$

Kioumourtzoglou (2016) applied this method to a Survival Analysis

- Medicare Cohort: 35 million participants living in 207 US cities with PM2.5 data, 11 million deaths
- City Specific Survival Analysis→ No confounding by covariates varying across city
  - Stratified by age, sex, race
- City Specific time trend→No confounding by covariates varying over time

So we have controlled for

- Group (city), time trend (spline of time) and their interaction
- We only look at PM2.5 variations around it's city specific mean and city specific time trend
- This gives a causal estimate
- HR of 1.2; 95% confidence interval [CI]: 1.1, 1.3 per10 $\mu g/m^3$ increase in the annual $PM_{2.5}$ concentrations →
- 2% increase in mortality rate per 1 $\mu g/m^3$

Instrumental Variable Analysis

- Suppose the outcome depends on predictors in the following manner:
- $Y_t = A_t\theta + \eta_t$
- Where θ represents the effect of exposue A, and η_t represents the impact of all other variables on the outcome

- suppose further we can find a variable Z such that Z is associated with Y only through A.
- $A_t = Z_t\delta + \tau_t$

- Where τ_t represents the other sources of variations in *exposure*, and in particular, *all of the exposure variations that are associated with other predictors of outcome,* ***measured*** *or* ***unmeasured***.

- let Z1 be the Z such that E(A|Z1)=a, and similarly E(A|Z2)= a´

- $E(Y^{Z=Z1})=E(\theta A + \eta_t | Z=Z1)= \theta a + E(\eta_t)$

- And hence $E(Y^{Z=Z1} - Y^{Z=Z2})= \theta(a- a´)$
- And θ is the causal estimate of the effect of exposure independent of measured and unmeasured confounders
- This works because Z represents the part of the variation in A that is not confounded by other predictors of outcome because it is not associated with them

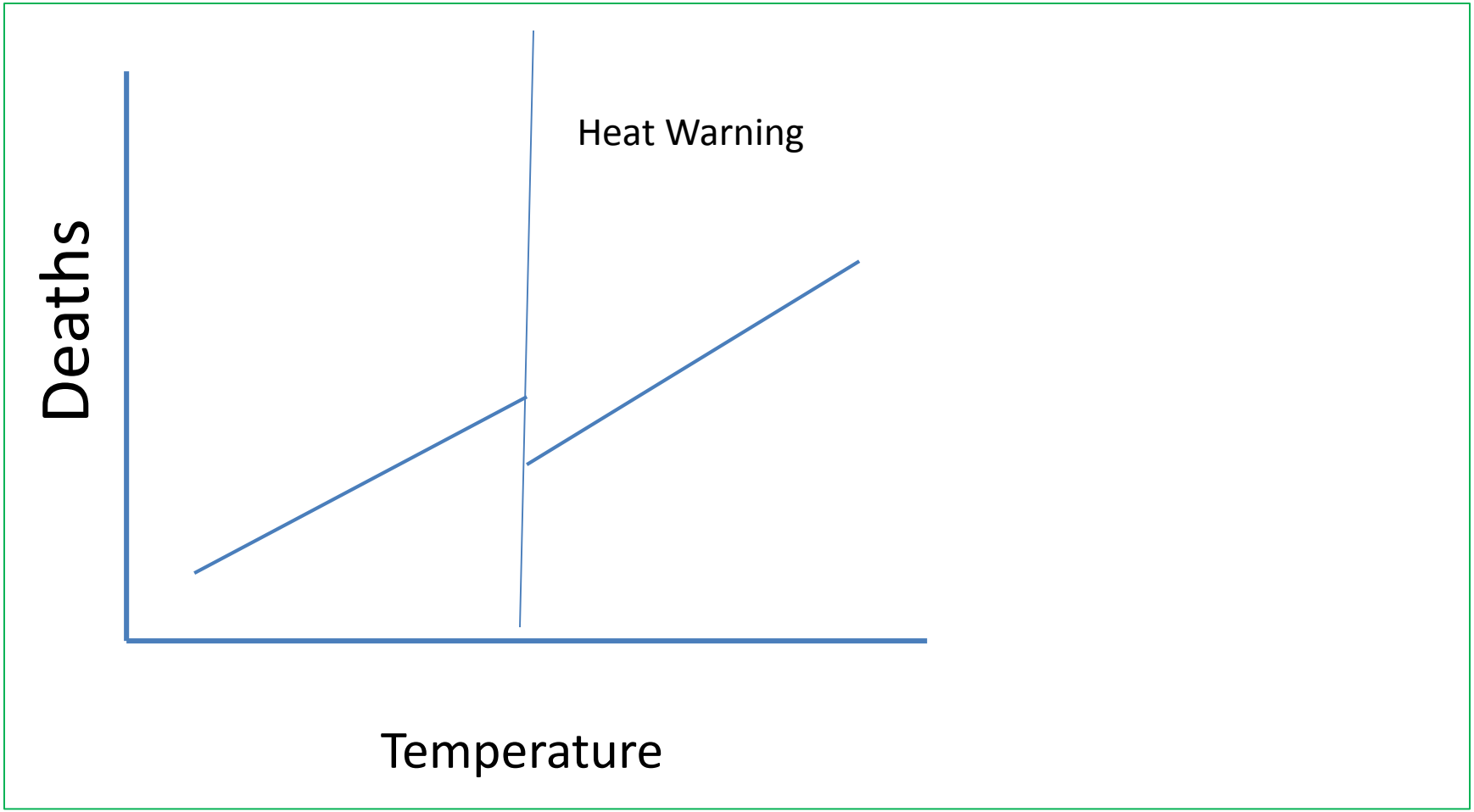
What is a Good Instrument?

- The mixing height is the height above ground below which air is well mixed. When it is higher, local emissions are diluted into more air.
- Likewise low wind speed results in local emissions piling up
- We used them (after control for season and temperature) as instruments for local PM
- We found a significant effect of the instrument on daily deaths (0.9% per IQR, 95% CI (0.25,1.56))

Regression Discontinuity

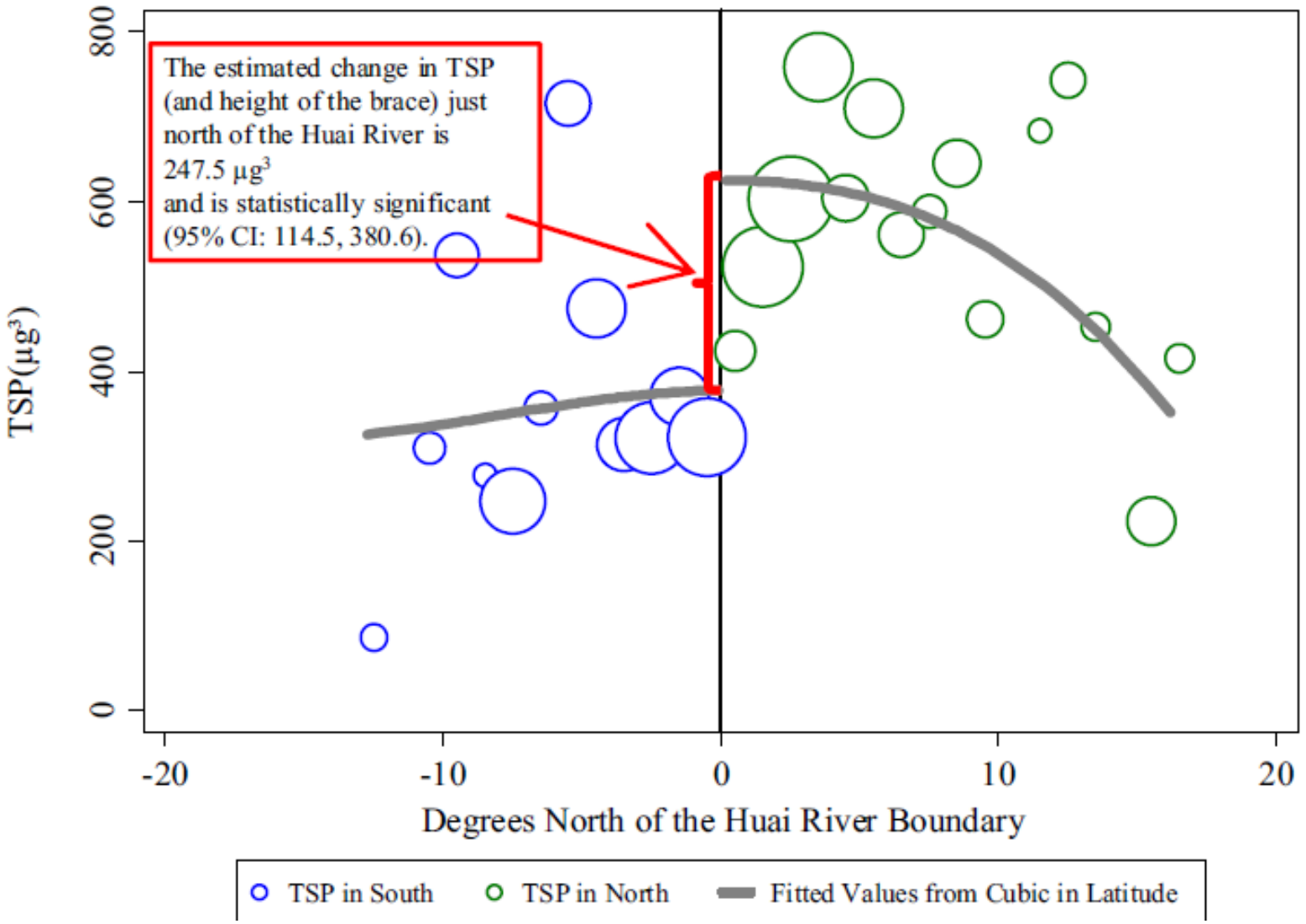
- Suppose something happens at a particular exposure level that modifies risk
- Than the dose response curve my be discontinuous there
- E.G. if at a specific threshold of temperature a heat warning is called, people may take defensive actions to reduce exposure
- In that case exposure may be lower just above the threshold than just below it
- Hence we may see fewer deaths just above the threshold

Effect of Temperature on Mortality



Greenstone Study of Coal Subsidy in China

- North of a boundary, subsidies were given for heating by coal
- Below they were not
- The temperatures were similar just above and below the line


The estimated change in TSP
(and height of the brace) just
north of the Huai River is
247.5 µg³
and is statistically significant
(95% CI: 114.5, 380.6).
TSP(µg³)
0
200
400
600
800
-20
-10
0
10
20
Degrees North of the Huai River Boundary
TSP in South
TSP in North
Fitted Values from Cubic in Latitude

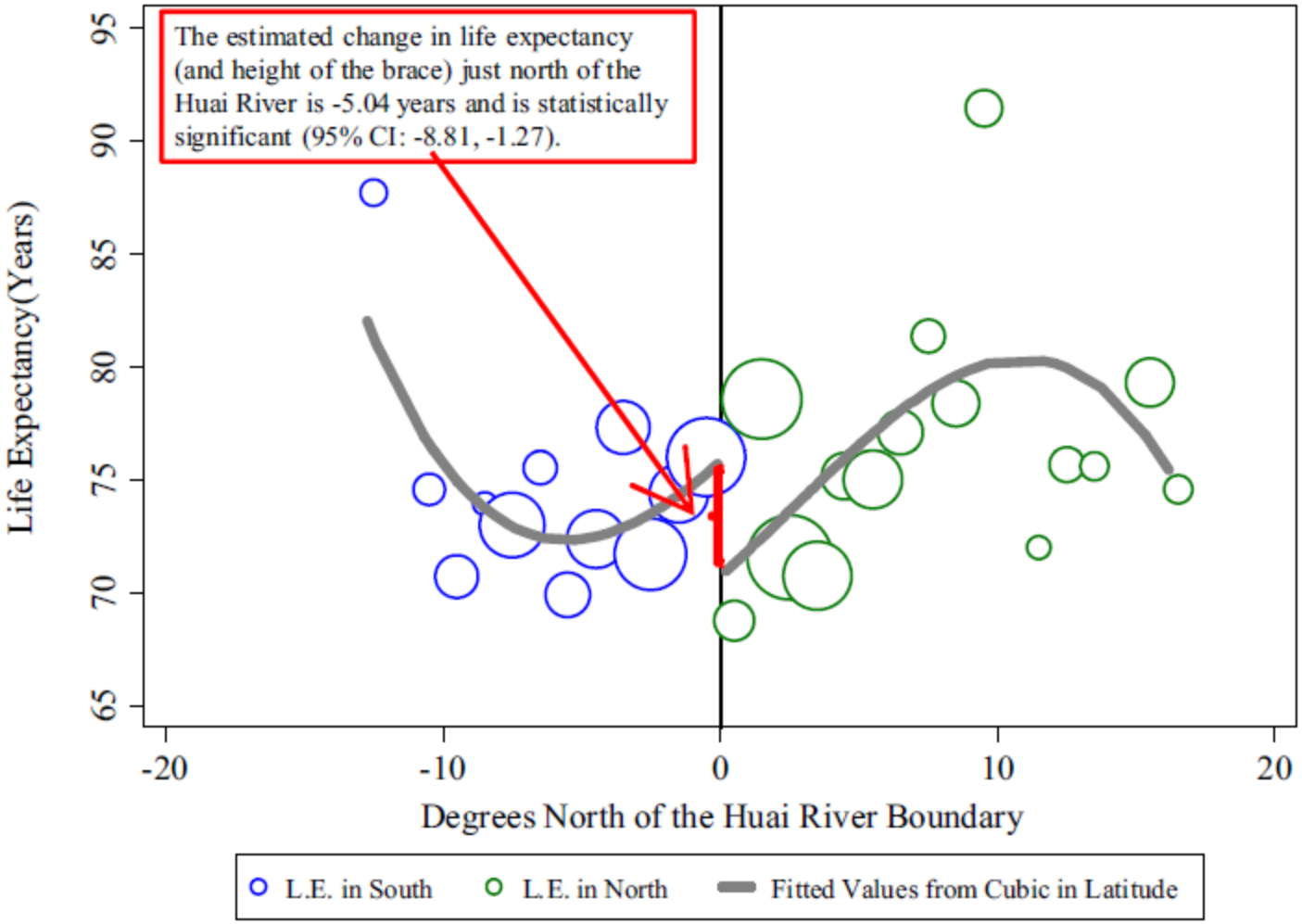


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

Propensity Score Analysis

- Remember the goal of causal modeling is to mimic as well as possible a randomized trial
- Trials balance the distribution of covariates between exposed and unexposed
- Propensity Scores are a way to mimic that because they are a balancing score
- Conditional on the score the distribution of observed covariates should be the same between exposed and unexposed

Propensity Scores (How)

- Suppose exposure is binary for simplicity
- If we predict the probability of high exposure as a function of all covariates
- Then we can weight each observation by the inverse of its probability given the covariates
- This makes the exposure in the weighted sample independent of the covariates
- Continuous exposures are accommodated by using probability densities

Or we could Match on the score

- Group subjects that have comparable chances of being assigned to the treatment (exposure) group vs the control (unexposed) based on their characteristics
- So, we compare exposed people who have a 10% chance of being exposed given the confounders with unexposed people who have a 10% chance of being exposed. Then there is no confounding by their characteristics, which are the same.

Other Attributes of Propensity Scores

- We can put more covariates in the model for the propensity score than we could in the model for the outcome, since we don't have to worry about collinearity
- So we can control for more covariates
- If we control for enough observed covariates we can hope that unobserved covariates have enough correlation with that large set that they are effectively controlled as well

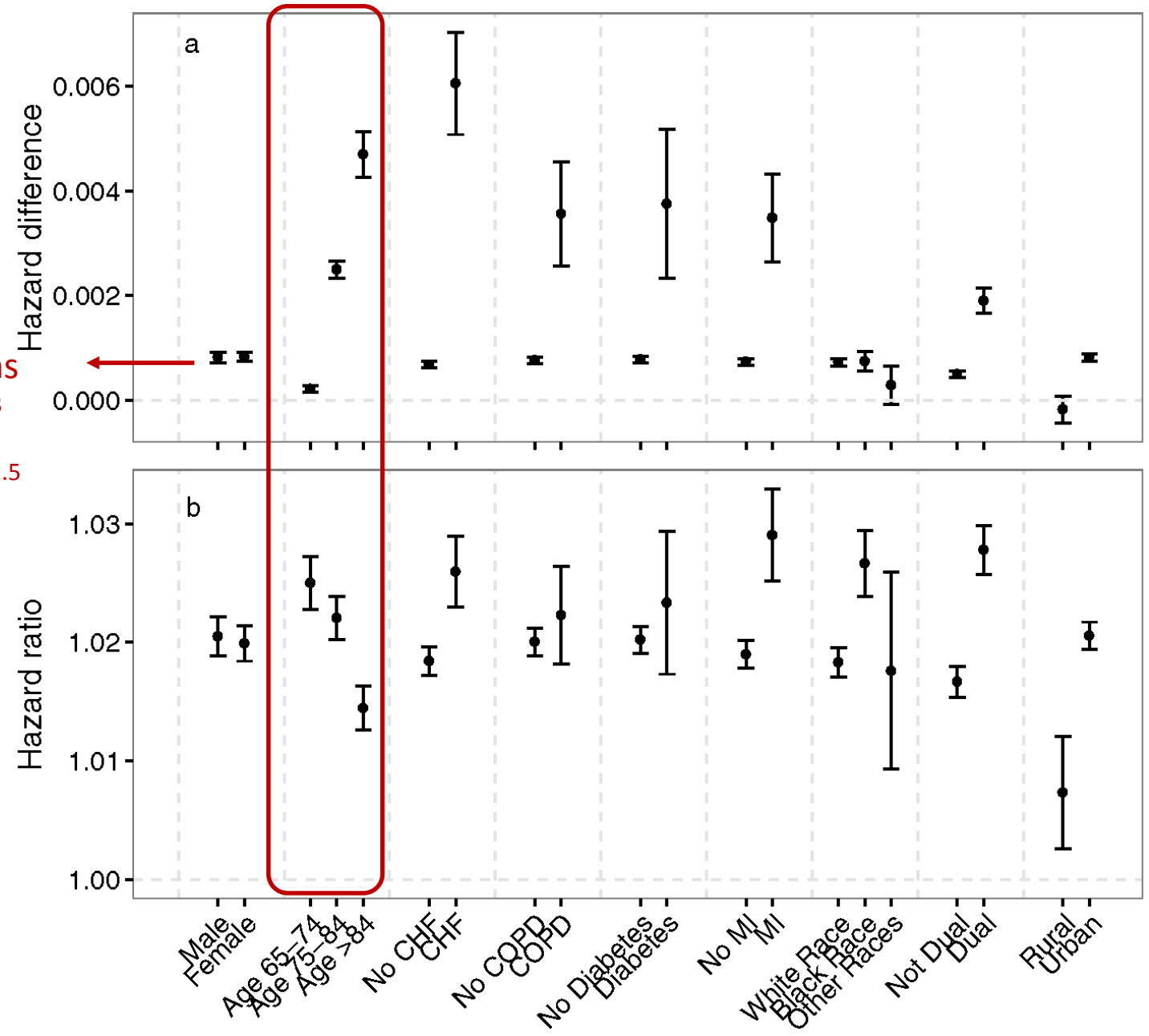
Application: $PM_{2.5}$ and Survival in Older Adults in the Southeast (Wang 2017)

- 13 million Medicare participants in the Southeastern US
- Followed for 12 years
- Matched to air pollution by Zip code
- We also demonstrated how to do this using an additive model for risk instead of a ratio model

Additive
Approach 1

Prevent **5000** premature deaths for each 1 µg m^{-3} reduction in $PM_{2.5}$

Multiplicative
Cox Model



These Models All

- Model the rate of death over time as a function of exposure
- No one ever asked their Physician “What will happen to my Instantaneous Rrisk of Dying if I Quit Smoking?”
- What the ask is “How much Longer will I Live if I Quit Smoking?”
- How do we answer that?

Causal Accelerated Failure Time Model

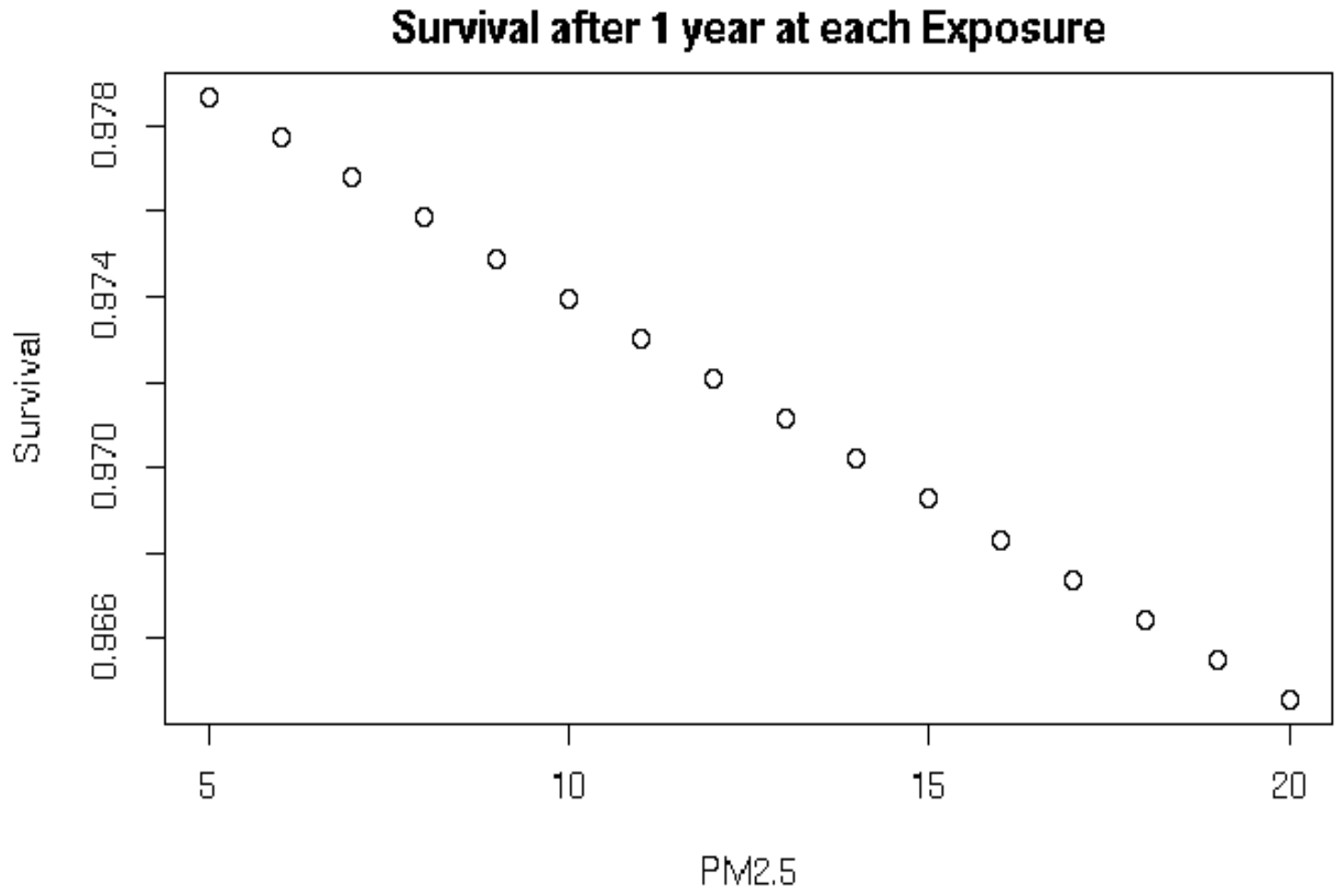
- $Log(T) = \beta x + Z\gamma$
- Requires an assumption about the distribution of T
- Gives us relative time, so we can say exposure a reduces life expectancy by d percent
- More meaningful to most people

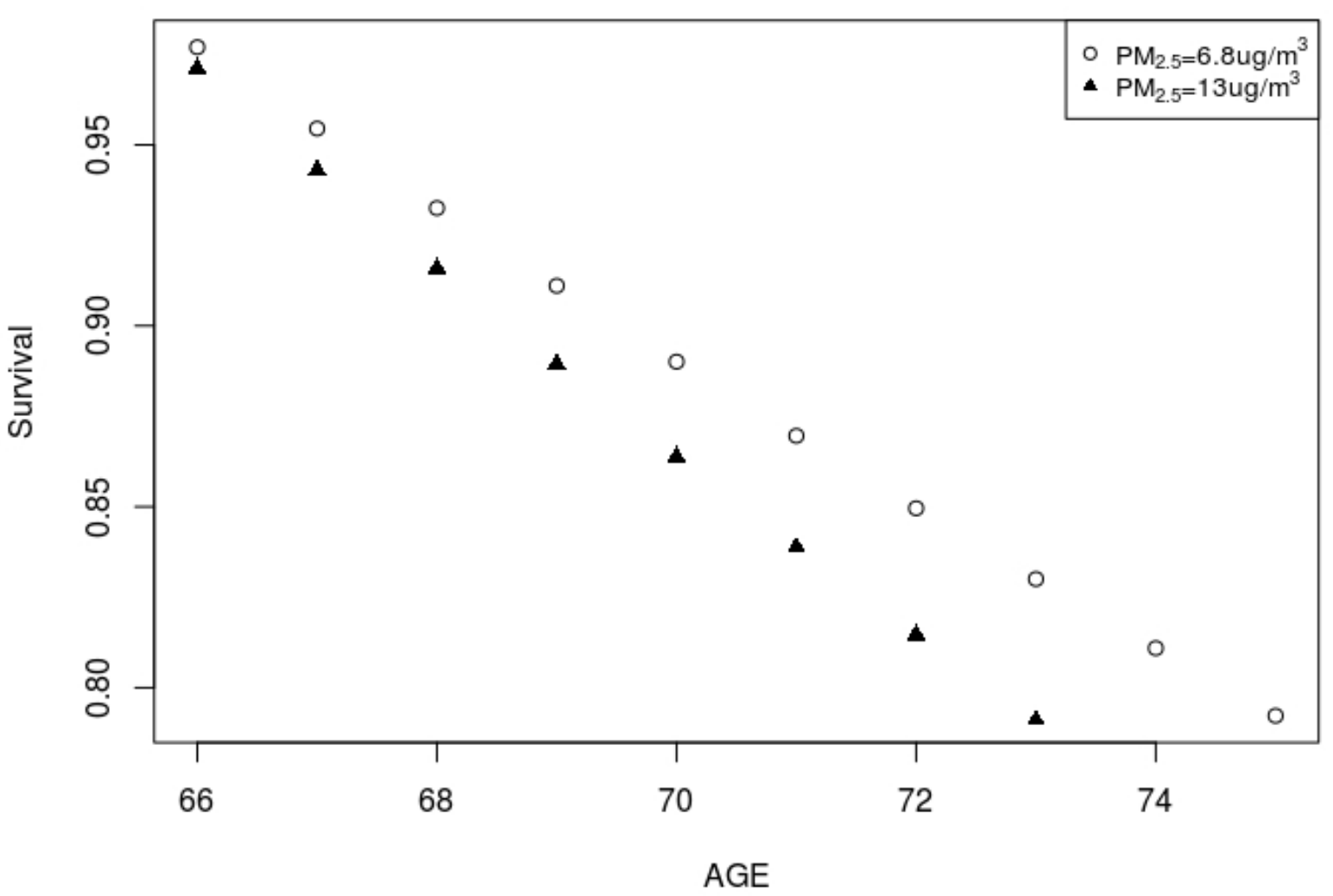
How do we make it causal?

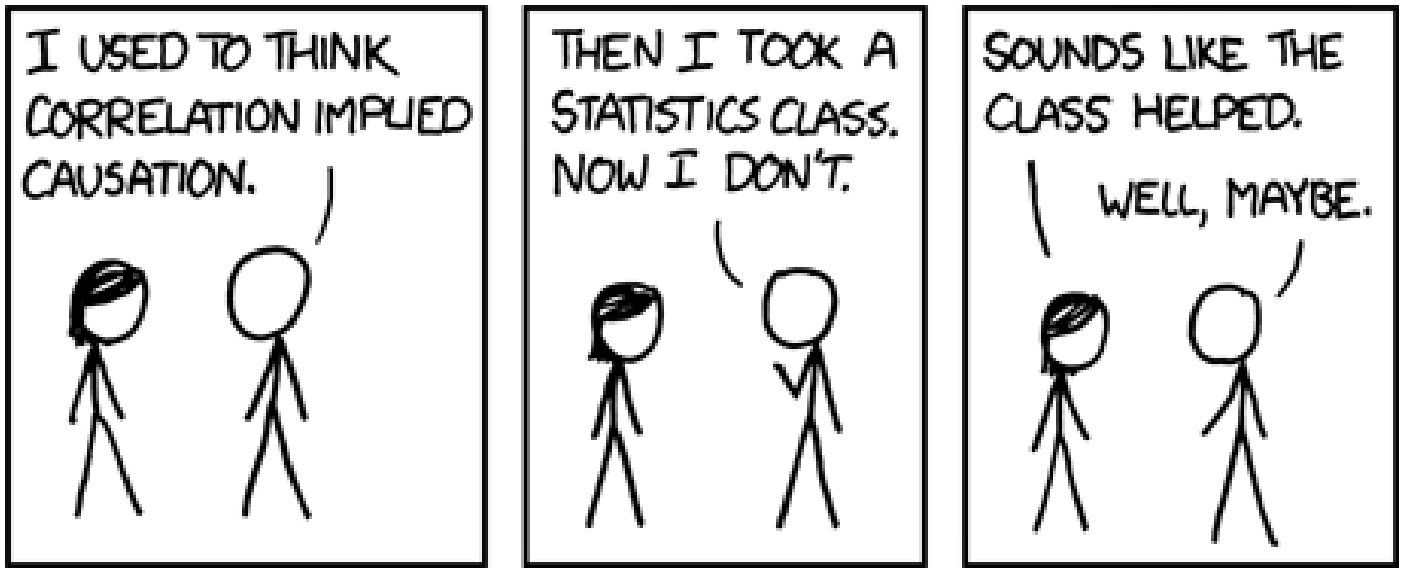
- This is a multiplicative model
- The probability of surviving for 2.5 years is the product of the probability of surviving for year 1, year 2, and half of year 3
- So w can fit the model using annual follow-up, and define the outcome (survival time) in each interval as 1, until the year of death, when it is a number between 0 and 1 depending on the date of death in that year
- We can incorporate time varying exposures
- We can then use IPW weights to make exposure independent of covariates and obtain a causal estimate

Population

- Everyone in New England covered by Medicare, the U.S. health insurance for people 65 and older (n=3,329,058 and 1,309,771 deaths) for the period 1999-2013.
- PM2.5 exposure estimate from Satellite Remote Sensing
- Low exposure (90^{th} percentile=11.7 $\mu g/m^3$)
- Covariates: Age, Race, Sex, poverty, Census Socio-economic and other data on the Zip Code level


Survival after 1 year at each Exposure
Survival
0.966
0.970
0.974
0.978
5
10
15
20
PM2.5


Fraction Surviving over 10 years at two Counterfactual Exposures
PM2.5=6.8ug/m3
PM2.5=13ug/m3
Survival
0.80
0.85
0.90
0.95
66
68
70
72
74
AGE


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.