Analysis of studies on the short term health effects of air pollution

Crab lab presentation

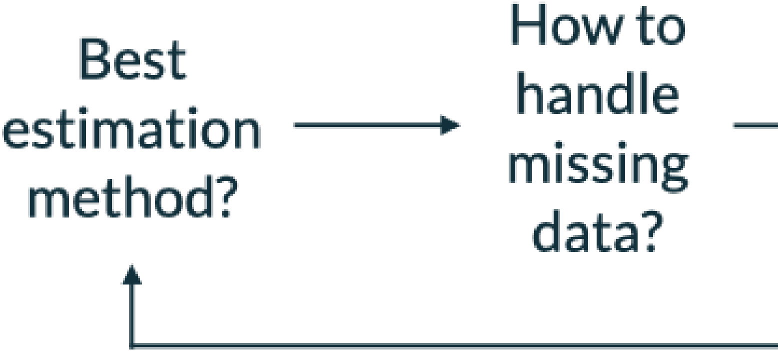
Vincent Bagilet - Leo Zabrocki

November 30, 2020

Analysis of studies on the short term health effects of air pollution

Vincent Bagilet - Leo Zabrocki

Genesis



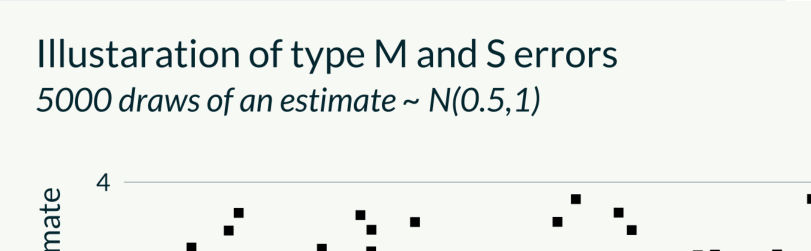
Main questions

- What is the usual **power** in studies of short-term health effects of air pollution ?
- How do **different identification strategies** perform to estimate these effects ?
- What is the impact of **missing data** on these estimates ?

USUAL POWER IN AIR POLLUTION STUDIES

Definitions

- **Power** : probability of finding an effect when there is actually one
- Power can be low when effects are small and/or variance of the estimates is large (**eg** small sample size)



Motivation

- Ioannidis et al (2017) showed that studies in economics are massively under-powered : median statistical power of 18%
- Is there a similar issue in studies of health effects of air pollution ?
- Health effects of air pollution are often tiny, making them difficult to detect
- Low power is associated with high rates of type M and S error

Method

- Follow Ioannidis and retrieve point estimates and s.e. of estimates
- Compute power, type M and type S errors
- Literature review of causal studies : yield a set of $\sim$ 30 studies
- Systematic literature review for other studies : for now, about 1000 estimates

Preliminary results

- Causal studies : some studies have high power, others have quite low power.

Performance of different estimation methods

Motivation

- Epidemiologists often use very simple models, with small sample size : is it enough to recover the true effects ?
- Are more “fancy” techniques necessary ?
- Some methods can perform better than others in some contexts but less well in others
- Want to look into the performance of Poisson generalized additive model, IV, DiD, event study, RD

? ? ?

- Fancy techniques also limit the number of situations in which we can compute the estimates of interest

Method

- For the sake of the example, let's focus on a simple Poisson generalized additive model :

$$h_{ct} = \alpha + \beta_c p_{ct} + \boldsymbol{W'_{ct}}\delta + \boldsymbol{C'_{ct}}\gamma + \epsilon_{ct}$$

- Use both actual and fake data (here focus on actual data)

1. Estimate the model on the existing data

–

1. Define a "fake", known, effect β_c

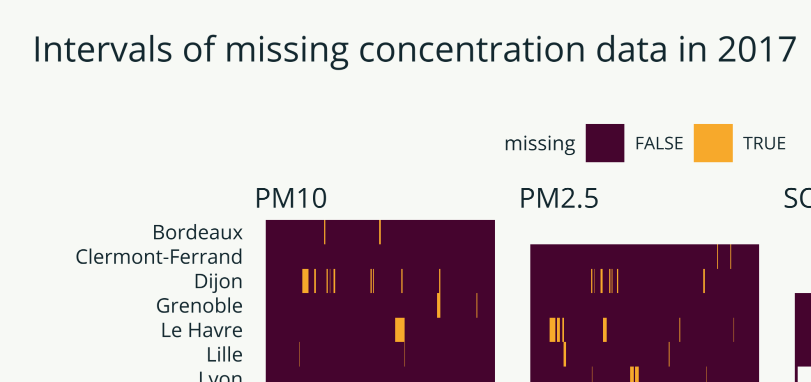
–

- Do this for each estimation method (with different DGPs of course)
- Look how our measures of interest vary with sample size and effect size.
- Where do papers in the literature lie ? $\rightarrow$ what problem could it be exposed to
- Reproduce the same analysis with fake data (**ie** generate all the data)

Impact of missing values

Motivation

- Air pollution data sets always display missing observations : not always clear how to handle them



Questions

- Does the literature discuss missing data issues ?
- To what extent do missing data affect estimates ?
- Are some estimation methods more robust to the missing data problems ?
- How does this vary with the type of missing data mechanism ?
- If it is actually a problem, which imputation method performs better ?

Method

1. Build a complete data set

—

1. Estimate the model and find the “true” effect

—

1. Delete data to create missing observations (create 1000 samples)

—

1. Estimate the model on the incomplete sets

—

Comparison of imputation methods

If missing data is actually a problem :

1. Redo the previous steps

–

1. On each incomplete set, impute missing observation

–

1. Estimate the model on the imputed sets

–

1. Compute bias, power, type I, type M, type S error

class : center, middle