The Mortalility and Medical Costs of Air Pollution- Review

Author: Daniel Saggau — daniel.saggau@campus.lmu.de
Supervisor: Nadzeya Laurentsyeva

15/03/2021

1 Introduction

Paper uses heterogeneous treatment effect and generic machine learning inference. Mortality in elderly population 25% are effected. Conduct first large scale quasi experimental investigation on health care use and medical case. Wind direction as an IV identification strategy. Mortality displacement as problem with identification.

Use cox-lasso model as further model.

2 Literature Review

3 Method used in the Paper

3 pillars:

- Mortality and health care use
- Life-Years Lost
- Treatment Effect Heterogeneity

3.1 Mortality and health care use

- Dependent Variables: health care use, health care spending and net of any confounding factors
- Include FE: country level, state by month level and month by year level
- Cluster all standard errors at country level and weight all estimates by population for per captia dependent variables.
- Robustness to different clustering choices.
- IV- Strategy: Daily wind direction varying by geography

3.2 Life-Years Lost

- Statistical value of Life
- Counterfactual life expectancy is unobserved
- All previous studies use either counterfactuual life exepectancy via population life tables
- Or: Using estimation change in cause and age specific mortaility over time (here argue all prior studies only use age and sex but here we are using preconditions)
- Problem: Disentangle affect by pollution
- Solution: using rich set of different health and non-health characteristics
- left censored data (use cox ph model to estimate)
- Unsupervised learning (cluster analysis) to group pollution effect by county across the US.
-

3.3 Treatment Effect Heterogeneity

- CDDF by Chernozhukov
- Best linear predictor of conditional average treatment effect under general conditions

- Computational constraints, making it unable to unclude country, state-by-month and ,month-by-year fixed effects in our regression
- Instead use month, year and division fixed effects
- Using subgroups to estimate equations and average (using 250 subsets)
- Gradient boosted decision tree

3.3.1 Criticism:

- Shortcoming: External validity
- Model evaluation for ML method not SOTA
- Bayesian Estimation (probabilistic modelling)
- Added value ML here: Prediction of pollution
- monotonicity assumption
- Robustness not really complete and more transparency needed
- Potentially

3.3.2 Model evaluation for ML

Traditionally, the way to assess prediction quality is not simply comparing the expectancies. For more robustness, usally one compares the different thresholds of the data (25, 50, 75 quantile percentage).

- c-index
- brier-score

3.4 Modification or difference it may have made

- Choice of survival random forest unclear ()
- Random forest objective is prediction while lasso is more statistically grounded and there for explanation
- Only comparing non-parametric and parametric models

3.4.1 Usage of OLS regression

- Usage of regularized method for machine learning method but not for main regression
- PLS for regular regression for transparent feature selection rather than p-value hacking
- Also possible to model with bayesian approach, allowing for more insightsful information rather than point estimate

4 Conclusion

Additional Variation via estimation of life expectancy. Using heterogeneous treatment effect on quasi-experiment. Suggest that life expectancy models vulnerability to air pollution exposure.

5 References