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.

Ensure no mortality displacement by using other time windows (5,14,28 days).

Using 40 billion people observations, using generic machine learning approach by Chernozhukov et al. (2018). This approach allows us to examine heterogeneity in treatment amongst elderly population. We can see htat there is a near 0 effect among 75 percent of population, but allows us to see that in top 5 percent has an substantial effect.

Contribute because are using 2 new approaches. Introducing machine learning for time to event studies for a economics study, looking beyond age and sex as determinants for life expectancy. Specifically, this paper uses survival random forest and a survival cox-proportional hazard model. Secondly, this paper uses this proposed generic machine learning for inference on heterogeneous treatment effects to a quasi-experimental study. Few papers have implemented this method.

2 Literature Review

3 Data

3.1 Air Pollution

- Data on various different pollution types
- PM 2.5 is most robust measure

3.2 Athmospheric Conditions

-

3.3 Mortality, Morbidity and Medical Costs

- Looking at population of 65 and above
-

4 Method used in the Paper

3 pillars:

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

4.1 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

4.2 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)
- There are various way to define statistical value of life.
- Examples include: Excess death, hazard fraction, relative risk, premature deaths, attributable deaths.
- These various notions of statistical value in itself offer variety to create an estimate for exposure.
- Here, we are only looking at very few comparisons, without actually even comparing actual performance of these measures.
- Henceforth it is somewhat speculative to argue one way or another.
- Irrespective, it is not unlikely that there is also variation in these different estimates.
- Solely based on the research presented in this paper, one cannot exclude that the chosen measure was cherry picked to ensure significance.
- Inevitably, absence of proof and proof of absence are not the same.

- One cannot know by merely looking at the values whether it would have changed anything but it surely would have added considerable robustness, allowing for different values of statistical life measures rather than focusing entirely on estimates prediction scores without ever evaluating their performance.
- Unsupervised learning (cluster analysis) to group pollution effect by county across the US.

4.3 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

4.3.1 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
- Contribution ML algorithm

4.3.2 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

4.4 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

4.4.1 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

4.4.2 Importance ML method

- Here generic machine learning inference is used for subsequent HTE analysis.
- Inevitably, one should consider the actual contribution of this method.
- Novel machine learning method here lead to lower expected prediction of life expectancy.
- Unfortunately, it is largely unclear whether these lower predictions are actually a more accurate representation of actual survival, or whether these score are simply lower.
- As mentioned, to actually get a more comprehensive performance evaluation
- Would have been more interesting using causal RF or other machine learning methods compared to OLS as opposed to generic ml inference for a subsequent
- Instead this paper relies very heavily on traditionally econometric tools for statistical inference, adding little by introducing machine learning into their study.
- This does not imply that machine learning are not applicable here, solely that the contribution made here is marginal as compared to what it could have been

4.4.3 Discourse Double Machine Learning vs Generic Machine Learning

- Performance Random Forest
-

4.4.4 Unanswered Questions

- Only disentangle effect in old population
- Long-run effect or immediate effect?
- Are we really capturing immediate repercussions caused by pollution
- Causality questions always challenging

5 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.

6 References

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., & Reif, J. (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. American Economic Review, 109(12), 4178-4219.

Hammitt, J. K., Morfeld, P., Tuomisto, J. T., & Erren, T. C. (2020). Premature deaths, statistical lives, and years of life lost: identification, quantification, and valuation of mortality risks. Risk Analysis, 40(4), 674-695.