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 various different survival models in addition to two machine learning methods. 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 among elderly population. Specifically, this paper disentangles heterogeneity in age. 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.

The contribution of this paper is suggested to be twofold. Firstly, this paper spends a sizable amount of defining a suitable estimate for life years left, partially by 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 two machine learning methods namely survival random forest and a cox-lasso 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.

Provide evidence that there is significant variation within the population. Identify small subset of population most vulnerable.

2 Background

Use PM 2.5, allowing to capture domestic pollution and pollution that is transported via the wind from other areas. For reference, PM 2.5 is a common pollution measure in environmental studies, capturing Authors suggest that PM 2.5 has been linked to adverse health outcomes in prior research.

3 Literature Review

4 Data

4.1 Air Pollution

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

4.2 Athmospheric Conditions

-

4.3 Mortality, Morbidity and Medical Costs

- Looking at population of 65 and above
-

5 Method used in the Paper

3 pillars:

- Mortality and health care use

- Life-Years Lost
- Treatment Effect Heterogeneity

5.1 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

5.2 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.
- Using the lagged variables to ensure that there is no reverse causality.

The problem with using lags is determining the lag structure of our data. Here, we simple use a single year lag which is somewhat naive, because we don't know whether the medicare data would actually translate in that manner. This is perhaps also a source of mis-specification.

- Unsupervised learning (cluster analysis) to group pollution effect by county across the US.

5.3 Treatment Effect Heterogeneity

- CDDF by Chernozhukov
- Best linear predictor of conditional average treatment effect under general conditions
- Computational constraints, making it unable to include 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

5.3.1 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
- Only heterogeneity in age but not in class or socioeconomic standing or geography
- Reverse causality lag does not look at optimal lag structure
- Some health conditions given in the might only translate at a later stage

5.3.2 Model evaluation for ML

Traditional model evaluation is omitted in this paper, justified by the argument that if there is no variation in the results, it does not matter how precise our estimates are. In my opinion, this argument is very weak and to ensure scientific rigor, one should have at least reported some sort of model performance for the machine learning methods. Irrespective,model evaluation would have been a substantial improvement to this paper because it would have allowed to assess the credibility of these different estimates allowing the reader to make his own judgment about the contribution of using these estimates apart from statistical significance of the results. Given the similarity of the estimates, it is not unlikely that these estimates would have hinted towards an only very marginal improvement, if any.

Model evaluation for classical classification tasks and survival tasks differentiate. To accommodate the right censored nature of our data (50 % of the individuals in our sample don't die within the time frame), we would use methods employing some sort of estimation for our right censored observations e.g.via inverse weighted propensity estimates (IWPC). Not going into too much detail, we could have used the concordance statistics (or in short c-statistics) or a brier score here to evaluate the performance. For more robustness, usally one compares

the different thresholds of the data (25, 50, 75 quantile percentage).

5.3.3 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

5.3.4 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.

The difference between the cox-ph model and the survival random forest or the cox-lasso is marginal (5.33 to 5.23). The difference between the cox-ph model and the cox-lasso model is slighter bigger(5.33 to 4.8) but still not really a jump as opposed to the jump between medicare FFS average and the cox model.

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.

Further the argument that you don't need a concise estimate for identification is somewhat confusing. The entire point of building the different survival models is to get a more precise survival estimate, accounting for further conditions. If we disregard precision and don't report any information on model evaluation, why do we build a machine learning model that is so marginally different from standard measures in the first place? If it is not for the more accurate prediction, there is no justification for using the survival random forest or the cox-lasso model. Further the argument that it only matters if the results play out differently is unscientific. Generally speaking, this does not imply that machine learning tools are not applicable here, solely that the contribution made here is marginal as compared to what it could have been

5.3.5 Computational Shortcomings

Due to the vast number of observations and data, it is computationally unfeasible to apply advanced tools from atmospheric science models here. Henceforth one can only speculate how these models would have performed.

Unable to include country, state-by month, and month by year FE. All replaced by month, year, and division fixed effects.

Only use random 5 percent of FFS medicare sample, still amounting to 1.2 million individuals. This subset is justified in the paper as a tool for computational ease. While they do run the regression multiple times, it might have been nice to see other medicare beneficiaries. Little to nothing is known about this random sample and maybe by looking other specific estimates, accounting for factors like geography would have unfolded further variation in our treatment or at least influence our life expectancy estimates. Especially given how sparse information on performance of these estimates is in this paper, this might have significantly altered the estimates.

6 Conclusion

In conclusion, this paper does spark an interesting discourse about pollution exposure, shifting the narrative from focusing on improving the most polluted areas to focusing on the most affected areas. Essentially, this study does provide systemic evidence for heterogeneity in treatment, suggesting that the most vulnerable groups are more affected by pollution. Irrespective, there are various things that could have been done to further improve this study and enhanced their scientific rigor.

The biggest shortcoming was the implementation of machine learning method. While it did not harm the results, it contributed near to nothing to this study. There was no significant improvement or variation by using these tools as compared to the extensive cox-model. What is even worse is the disregard for actual model evaluation, reducing their credibility and justification even further. The argument that precision is only important for identification when results differ is not very scientific and misplaced.

A second shortcoming of this paper is the lack of scope in the treatment heterogeneity. We clearly see strong variation in age groups when it comes to treatment heterogeneity, but this is just one of many important factors to disentangle. Especially from a policy perspective, this study should have explored

6.0.1 Unanswered Questions

As mentioned, this paper truely addresses a broader field of research. One of the many interesting questions that remain unanswered is the effect of pollution in other shares of the populations. One interesting addition would be looking at infant mortality and the impact of pollution on younger age groups. Further, there are more convoluted aspects that remain hard to disentangle. It remains uncertain whether we will be able to causally disentangle the impact of pollution in the long run and perhaps accounting for heterogeneity over exposure time to pollution. Ultimately, the question remains whether these insights will really be causal or merely a mere approximation of an causal effect, capturing some other sort of variation due to omitted factors.

7 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.

Deryugina, T., Miller, N., Molitor, D., & Reif, J. (2021). Geographic and socioeconomic heterogeneity in the benefits of reducing air pollution in the United States. Environmental and energy policy and the economy, 2(1), 157-189.

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.