Respiratory Deaths and Pollution: Principle Component Analysis

Rashad Dixon, Andy Wang, Chris Hyland
University of Texas at Austin

Abstract

Air pollution has long been associated with a variety of respiratory conditions such as emphysema, asthma and even lung cancer. In this paper we are trying to understand the persistence of the effects of pollution shocks on human death rates directly linked to respiratory and circulatory deaths. For this purpose, we have extracted yearly data for respiratory deaths and the concentration of the 3 main air pollutants CO2, O3, and SO2 in parts per million as well as PM 2.5 which is a fine particulate matter kicked into the atmosphere during many commercial processes. This analysis uses an ADL model as well as principal component analysis to predict the year over year changes in respiratory deaths based on the changes in the 3 main pollutants. The Pseudo-Out of Sample testing of the models resulted in insignificant estimators using the ADL process. PCA models may provide significant estimation and modeling of respiratory fatalities in the presence of pollution shocks, but additional estimation methods are required to determine persistence of the pollution-death relationship.

Introduction

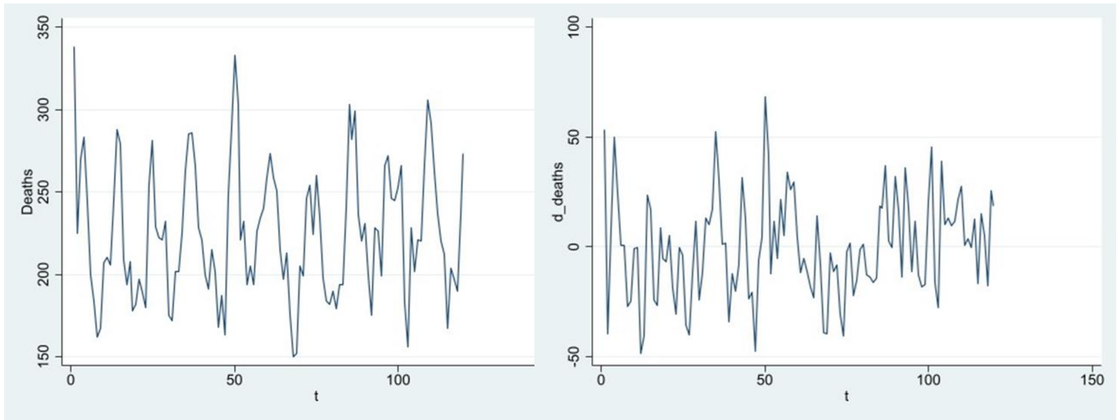
Respiratory pathways are susceptible to long-term damage when exposed to ambient air pollution. (Jiang et. al) Pollutant shocks occur as a result of weather-related incidents as well as being the by-product of industrial processes. Changes in energy prices and demand shocks are significantly correlated with variation in ambient air pollution (Ozaltun). Being able to accurately estimate the persistence of the relationship between pollution shocks and respiratory deaths allows policy-makers and health care providers to understand how to mitigate risk, and determine the course of regulatory action. Pollution risk is measured by moving averages of daily pollution. Ozaltun shows that shocks in production may have some impact on pollution levels. Production smoothing and demand estimation may reduce seasonal production shocks (Gorman & Brannon), and thus potentially reduce shocks in ambient air pollutants

Data

Data was collected from two sources: The CDC Wonder monthly reported deaths, and the EPA monthly pollutant reading levels. Both data sets cover the years from 2000 through 2010 and are organized by state. Respiratory are collected from the CDC Wonder database while weighted average measurements of pollutant levels (in parts per million) in each state are collected from the EPA website. We use the data to train an ADL model, while also using the data to test the accuracy of myriad popular alternative forecasting models.

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{Y}_{\boldsymbol{t}} - \boldsymbol{\beta M}_{\boldsymbol{t}}$$

The monthly death series is highly seasonal due to weather related illnesses and other seasonal factors. In order to effectively estimate the data we removed seasonality. Monthly deaths were estimated using a dummy variable for each month (estimates shown as BM_t), then the residuals were used as the new Y variable of interest. This transformed series estimated a p-value of 0.00 on the Dickey Fuller test, providing evidence of stationarity. The untransformed data (left) and transformed stationary data (right) are provided below:



Method

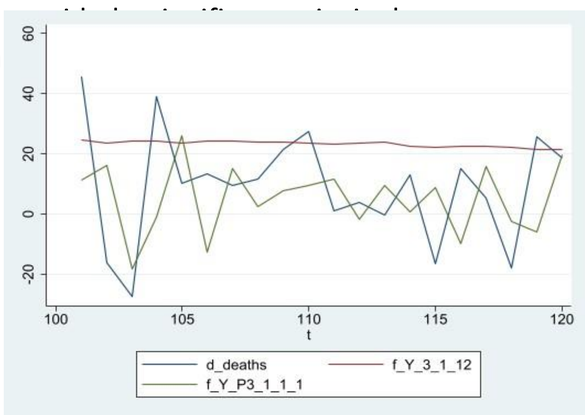
Autoregressive Distributed lag models were used the initial method of estimation. The example ADL model:

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{\beta}\Omega + \boldsymbol{\gamma\alpha} + \boldsymbol{\epsilon\theta}$$

with coefficient vectors β, γ, and ϵ . Ω is a n x 1 matrix of lagged values of $\widetilde{\boldsymbol{Y}}_{\boldsymbol{t}}$(transformed deaths), α is a m x 3 matrix of lagged pollution shocks, € is a i x I matrix of lagged PM 2.5 shocks. Model selection code was written to store the AIC, BIC, and RMSFE of models with up to 12 lags of deaths, 24 lags of (CO,SO2,O3), and up to 12 lags of PM2.5. The model with the lowest Bayesian Information Criterion and RMSFE is then selected as the representative model, and provides optimal values on m,n, and i. Principal component analysis (PCA) is then used to reduce the number of coefficients. Due to the contemporaneous nature of the pollution shocks the estimated coefficients are biased, and colinear. PCA allows for the reduction of the number of estimators in an attempt to reduce collinearity and bias from estimation.

Results

Pseudo-out of sample prediction of with ADL (top) and PCA (bottom). The BIC of the ADL model (1089) was minimized with 3 lags of the independent variable, 1 lag of each of the pollution shocks, and 12 lags of PM2.5 shocks. The BIC of the PCA model (1066) was minimized with 3 lags of the independent variable and 1 lags of the 3 estimated principal components. Though the ADL model resulted in insignificant estimators, the PCA model provided significant estimators for both deaths. The POOS forecast shown Provides evidence of the performance Of the principal component model. A limitation of these results are the Frequency of the data. Shorter lags in Pollutants may be an improvement as acute air pollution can have physical Impacts within hours of the shock.



Discussion

Though the models perform well in terms of their predictive ability, the estimates are insignificant. The model may benefit from the implementation of non-linear methods of estimation, or Poisson smoothing (Mokoena et. al). Additionally, there are confounding variables that are likely missing from estimation that limits the quality of the forecasts. Meteorological conditions may change the impact of air-borne pollutants. PM-10 pollution is also related to respiratory disease and fatalities, and may be an important factor to include in the model as well as regional demographics where these pollution shocks are occurring. Overall, the selected models fail to provide insight on the persistence of the mortality related impacts of air-borne pollutants.

Conclusions

Given the nature of pollutions impact on the human body, properly forecasting respiratory deaths as a function of shocks in pollutants is an important factor for decision makes in government regulation. Understanding how production, pollution, and human health are related can guide proper greenhouse gas and particulate interventions. The PCA model was successful in reducing the number of estimators required to optimize POOS forecasting of monthly deaths as a function of pollution shocks from 18 to 6. It was also more representative of the actual pollution data. Observations with higher frequency or the addition of demographic and meteorological data may increase the models predictive ability. Pollution smoothing may also be required to account for the gradual nature of pollution dispersal.

Contact

Rashad Dixon
Email: r.dixon@utexas.edu
Andy Wang
Email: andy.ct.wang@utexas.edu
Chris Hyland
Email: chris.f.hyland@gmail.com

References

1. Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., Luepker, R., Mittleman, M., Samet, J., Smith, S. C., and Tager, I. Air pollution and cardiovascular disease. Circulation 109 (6 2004), 2655–2671.
2. Mokoena, K. K., Ethan, C. J., Yu, Y., Shale, K., and Liu, F. Ambient air pollution and respiratory mortality in Xi'an, China: a time-series analysis. Respiratory Research 20 (12 2019), 139.
3. Jiang, Xu-Qin et al. "Air pollution and chronic airway diseases: what should people know and do?" Journal of thoracic disease vol. 8,1 (2016): E31-40. doi:10.3978/j.issn.2072-1439.2015.11.50
4. Ozaltun, Bora. "Learning from Supply Shocks in the Energy Market : Evidence from Local and Global Impacts of the Shale Revolution." Learning from Supply Shocks in the Energy Market : Evidence from Local and Global Impacts of the Shale Revolution, Massachusetts Institute of Technology, 1 Jan. 1970, https://dspace.mit.edu/handle/1721.1/127174.
5. Gorman, Michael F, and James I Brannon. "Seasonality and the Production-Smoothing Model." International Journal of Production Economics, vol. 65, no. 2, 2000, pp. 173–178., https://doi.org/10.1016/s0925-5273(99)00049-3.