Machine Learning for Causal Inference: An Application to Air Quality Impacts on House Prices

Jenny Ho

University of Washington

Abstract

Hedonic models are commonly used to recover the implicit prices of house attributes and local non-market public goods such as environmental quality. Yet they are plagued by omitted variable bias when variables that are correlated with the attribute in question are unobservable. Typically, researchers have relied on fixed effects, instrumental variables, or quasi-randomness to control for this. However, these methods require strong underlying assumptions that are often a priori implausible. The increase in availability of big data and unstructured data in the form of text and images allow for a more extensive set of variables that are relevant to consumers to be included in hedonic methods. Unstructured data are high-dimensional and require machine learning methods that are robust to multicollinearity and irrelevant variables. I collect a rich and comprehensive dataset of property listings from Zillow.com and extract features from house descriptions and curbside view images using natural language and computer vision tools. I apply machine learning techniques to estimate the effects of air pollution on house prices in Pennsylvania. Coupled with the inclusion of more data, this approach nests previous methods to further reduce bias. My results show that omitting important variables can understate the negative effects of air pollution on house prices by more than half.