Predicting Drug-induced Deaths Using Educational Attainment and Household Data in Ohio Counties

by Gezhi Cheng, Ziyi Liu, Yang Lu

Figure 1. Forecast Drug-Induced Deaths in Comparison with Real Deaths Data

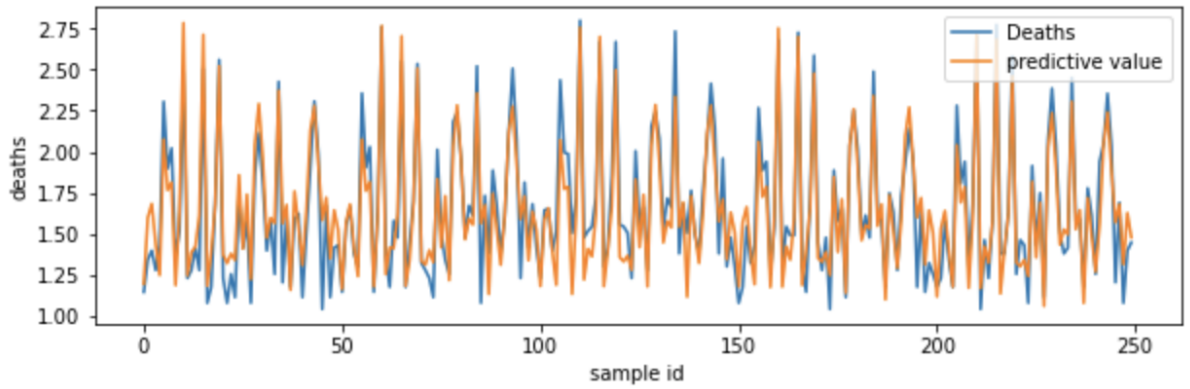
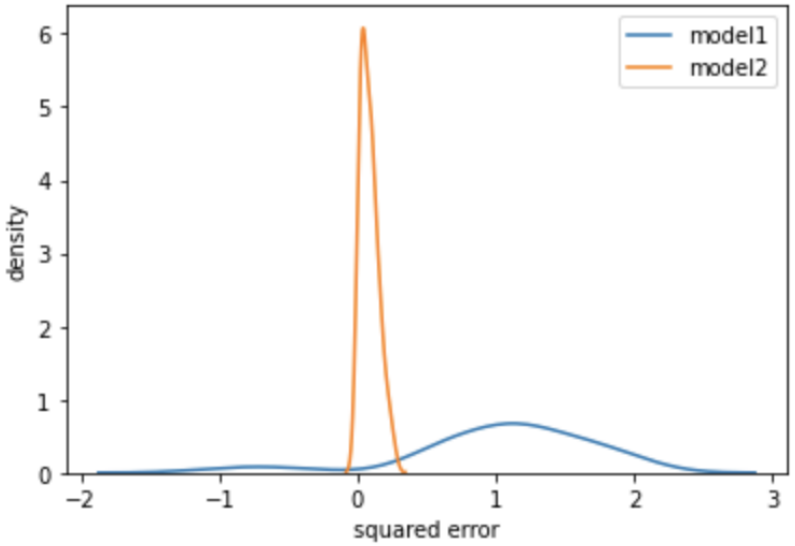


Figure 2. Squared Error Distribution Density Plot of Two Models



Background:

This one-month long research project mainly focused on predicting risk factors that may have contributed to drug-induced death cases in Ohio. We chose Ohio as the target location of our project since it has one of the highest rates of drug overdose deaths involving opioids in the United States. Due to Covid-19 in recent years, physical and mental health has become a topic of great concern, which reminded us of the effects of drug abuse on people's health. We primarily obtain and analyze data from the CDC WONDER database [1], Ohio Department

of Health [2] and the U.S. Census Bureau [3]. The methods used for data cleaning and analysis were Python, multiple linear regression model, and $\log_{10}(N+1)$ regression model.

Question:

How many deaths in Ohio will be caused by drugs with the educational attainment and household data known?

Conclusions:

The R^2 of our original multiple linear prediction model (model 1) is 0.937 with RMSE of 36.29 ($\log_{10}(N+1)=1.572$); after we implemented $\log_{10}(N+1)$ to independent and dependent variables (model 2), our model's R^2 was elevated to 0.947 with RMSE of 0.105. Since educational attainment showed great importance in our model, we ran the linear regression with solely educational attainment as an independent variable and discovered that the prediction model's R^2 is 0.84 with RMSE of 0.16.

With the help of the (log-)multiple linear regression model, we found that educational attainment showed significant importance in the study. That is to say, if more people receive better education, the drug abuse-induced deaths in Ohio may be eased. Besides, all the other predicting factors also showed relevance in this study, including population, divorce count, nonrelatives in nonfamily households, people living alone, number of foster children and people living in the group quarters.

References:

1. https://wonder.cdc.gov/controller/datarequest/D77

2. https://odh.ohio.gov/wps/portal/gov/odh/explore-data-and-stats/published-reports/marriagedivorcedatasourcetable

3. https://data.census.gov/cedsci/all

Code and Datasets:

https://github.com/ggzzcc/Introduction-to-Data-Analytics-in-Business