Analysis of association between economic conditions and traffic fatality in Maryland state

Boyang Zhang

Introduction

Traffic safety is a socio-economic aspect and therefore agencies have been focusing on enacting preventive measures, such as speed limits, DUI laws and seatbelt enforcement. However, the annual number of traffic accidents and fatalities have not yet significantly decreased. Apart from regulatory safety policies, macroeconomic factors, such as unemployment rate, could also exert great influence on traffic fatalities.

Understanding the relationship between economic conditions and traffic fatality can help us identify counties that might be at a higher risk for fatal crashes. The purpose of this study was to estimate association between economic conditions and traffic fatalities for Maryland state and counties in MD. Using exploratory analysis and regression models, including count-data regression methods, we showed that positive associations between favorable economic condition (low unemployment rate) and high traffic safety (low fatality rate), after adjusting for drivers' age, seasonality, weather conditions, light conditions and state/county's population. Therefore, state public policy makers should take this factor into consideration when allocating traffic safety related resources.

Methods

Data collection

The data used in this research is obtained from the Fatality Analysis Reporting System of the National Highway Traffic Safety Administration from 2005 to 2015. Population estimates of each county in Maryland State was obtained from U.S. Census Bureau. Maryland and county-specific monthly unemployment rate were obtained from Local Area Unemployment Statistics. These data were downloaded using the R programming language on 2016-10-21 except post 2010 population estimates needed to be downloaded by hand.

Key variable definitions

All variables were aggregated on a monthly basis from Janurary 2005 to December 2015, either summarizing as a state or by county.

- age: average age for accidents-related drivers
- drunk_prop: proportion of drunk drivers involved in the accidents
- ym: date defined by year and month
- weather: proportion of severe weather conditions in a month
- lgt_cond: average light conditions when accidents happened: 0-night;1-obscure(night light,dawn,dusk);2-daylight
- population: county-level annual population estimates
- unemp_rate: county-level monthly unemployment rate

Exploratory analysis

We aggregated raw data on months by concatenating data from different sources based on shared variables and extracting both state-level and county-level information for Maryland. Also, we assumed that population estimates would not change significantly throughout a year.

To explore association between traffic fatality and economic condition, we first used Maryland state as an example to illustrate. Exploratory plots of MD from 2005 to 2015 show seasonal pattern for traffic fatalities, indicating a non-linear dependence on date. Hence, we applied a natural spline of date with 10 degrees of freedom to describe the general trend. On the other hand, Maryland has a decreasing trend from 2005 to 2008 and from 2010 to 2015 but experienced a soaring unemployment rate from 2008 to 2010 when overall economy underwent great recession. Based on our prior knowledge and exploratory analysis, we include person-level information (i.e. age, proportion of drunk drivers and population), climate information (i.e. weather and light conditions) and economic conditions (i.e. unemployment rate) as our main effects.

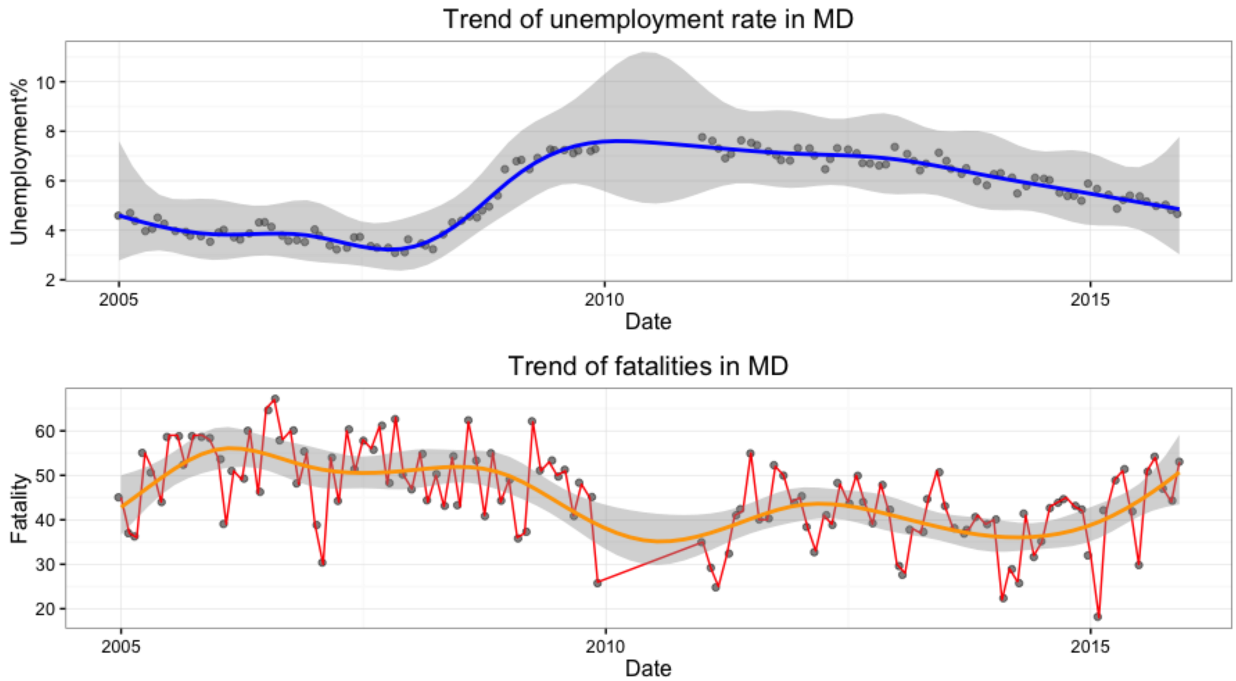

Figure 1: Exploratory analysis of Maryland state

Statistical Modeling

To analyze relation between unemployment rate and traffic fatality in MD, we performed a zero-truncated Poisson regression model to account for non-zero fatality in our dataset. For variable selection, we first identified possible main effects by our prior knowledge and exploratory analysis. Then, we used likelihood ratio test as a criteria for inclusion of variables.

Reproducibility

All data analyses performed in this report can be reproduced in the R markdown file: trafficfatality.Rmd. To reproduce exact results, data obtained online should be the same version of data we downloaded on 2016-10-21.

Results

The raw FARS dataset contained person-level, vehicle-level and accident-level data In order to simplify the problem, we decided to focus on accident-level data by borrowing information from person-level data and other data resources online.

For person-level files, we organized data in the same format by concatenating them based on shared variables, extracted all counties in MD, matched their FIPS code (geo_id). Then we imputed some missingness in the indicator of drunk drivers by setting threshold for blood-alcohol content (BAC) at 0.08, according to Maryland's DUI laws. By aggregating on monthly basis, we summarized average drivers'age involved in accidents (age) and proportion of drunk drivers (drunk_prop). Similarly, we made summaries of total fatalities in a specific month (fatals), proportion of severe conditions in a month (weather), average light condition when accidents happened (lgt_cond) from accident-level data. As for population estimates, we merged pre-2010 and post-2010 population estimates by county's FIPS code (geo_id). Finally, we merged all available preprocessed data by date (ym) and county FIPS code (geo_id).

Since non-negative integers of crash-frequency data were recorded, it was necessary to perform count-data regression methods that can properly account for integer nature of data structure. The most common technique employed to model count data was poisson regression and negative binomial regression. But both methods may predict zero counts even though there's no zero values in the dataset. Therefore, we employed zero-truncated poisson regression to solve this issue.

We divided our data into training set (60%) and test set (40%), trained our model on training set and assessed prediction accuracy in test set. We first fitted a regression model by including main effects that were statistically significant (threshold: 0.1). Then we applied likelihood ratio to check statistically significant interaction terms but none was discovered. Thus, our final zero-truncated poisson regression model was: