Interpreting Causal Relationships between Contributing Factors and Motor Vehicle Accidents in the New York Metropolitan Area

Jian Ruan, Zach Yu, Carrie Hashimoto, Sophia Lyu (Undergraduate), Patrick Yee (Graduate)

February 22, 2023

1 Motivation and Overview of Primary Data.

The aim of this project is to analyze contributing factors for motor vehicle accidents and related injuries in the New York Metropolitan area. A better statistical understanding of these factors could help guide more intelligent decisions in the interest of public safety by preemptively identifying conditions under which accidents are more likely. These analyses would be of interest to officials in both the public (legislative, infrastructual) and private (insurance, engineering) sectors.

The primary dataset for this project is a list of motor vehicle accidents in New York City between 1 Januray 2013 and 31 December 2022 (MVCC). The data are collected by the New York Police Department and maintained by the City of New York [1] (1.97m rows $\times$ 29 cols) and contains date/time, vehicle type, location, and injury data for each raw observation. Since the MVCC set is complete population data (eg, contains every known car accident over the specified 10-year interval and is not a sample), the main focus of this study will be on regression and causality, not statistical inference.

2 Number and severity of accidents as a function of location.

The dataset provided each individual car crash in New York from 2013-2022. For each car crash, location and borough was provided. In order to gain an understanding of car crashes within each borough of New York, data was organized to count the number of car crashes within each borough. A pie chart is plotted below to display the amount of car crashes in each borough relative to other boroughs.

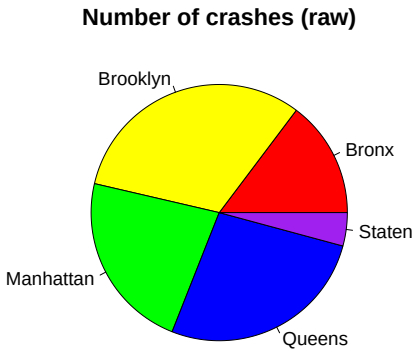


Figure 1: Number of accidents/borough, raw

Based on the pie chart, Brooklyn has the highest number of car crashes and Staten Island has the lowest number of car crashes in New York City from 2013-2022. Even though Brooklyn has the highest number

of car crashes, other boroughs follow closely behind such as Queens and Manhattan. However, even though it seems Brooklyn has the highest number of car crashes within New York from 2013-2022, the population within each borough must be accounted for to get an accurate representation. For example, Slaten Island may be as low as it seems when population of this borough is taken into account. Data was cross referenced from census data on population size in 2017 of each borough. A histogram is plotted below to illustrate the amount of car crashes per borough, per capita (population of each borough taken into account).

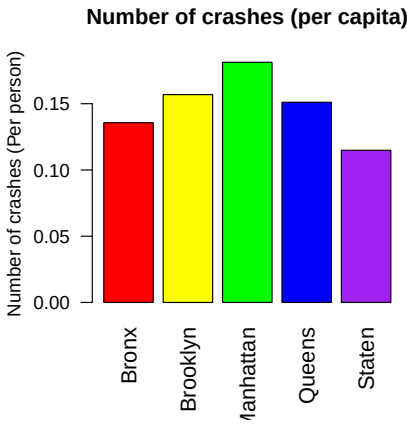


Figure 2: Number of accidents/borough, per capita

As shown, the highest number of car crashes in each borough is no longer Brooklyn when adjusting for population. [2] Clearly, the number of accidents per capita do not follow a uniform distribution, with Manhattan being the most dangerous and Staten Island being the least dangerous. These observations can be verified quantitatively:

```
        Chi-squared test for given probabilities

data:  observed.counts
X-squared = 16619, df = 4, p-value < 2.2e-16
```

3 Incorporating weather data.

It was hypothesized that weather influences the likelihood of accidents, as certain humidity levels can affect visibility and precipitation/high wind speeds can create unsafe road conditions. To investigate the correlation between accident risk and weather, the MVCC was cross referenced against weather data collected by JFK international airport from 2013 to 2022 (NCDC). [3] These data (3652 rows $\times$ 21 cols) contain precipitation, wind, temperature and other weather data for each day in the interval of interest. For each day, accident counts were aggregated from the primary data and regressed against rainfall for the corresponding entry in NCDC.

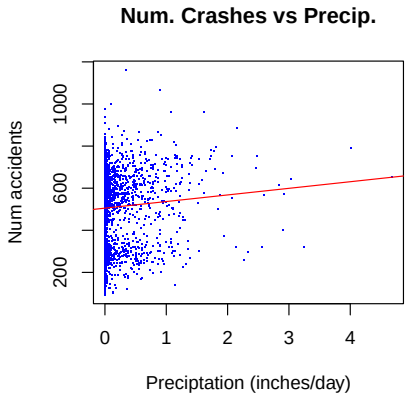


Figure 3: Precipitation (inches) versus number of crashes/day

The coefficient of this model is 31.4, which is significant ($p = .0002$). The associated residual plots are provided below.

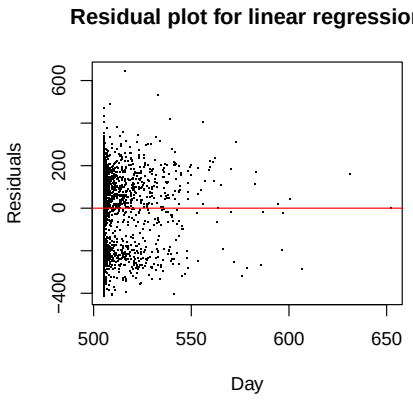


Figure 4: Residual plot for linear regression

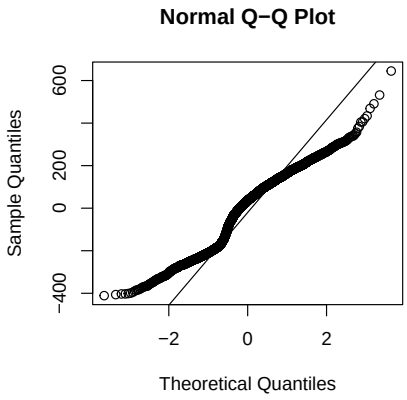


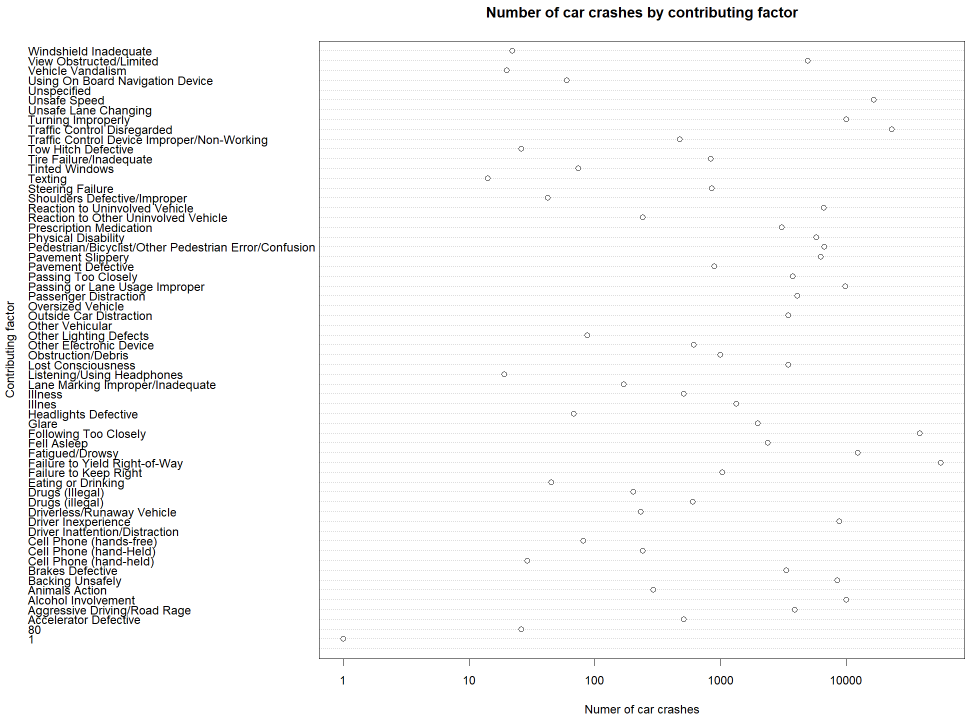
Figure 5: QQ plot for residuals

We see that the residuals have constant variance and the QQ plot is relatively symmetric about the main diagonal, so we conclude that linear regression is indeed apprioriate in this case.

4 Other contributing factors

To better understand the contributing factors on a granular level, a dot plot is used to map out the frequency of occurrence. Since data is grouped into categories in histograms, it is more difficult to read exact values and compare two data sets.

```
null device
          1
```



The dotplot does't show a significant pattern among different factors. The top four contributing factor includes: 1) driver inattention/distraction, 2) failure to yield right-of-way, 3) backing unsafely, and 4) fatigued/drowsy/sleep/unconsciousness.

5 Cumulative number of crashes over time

The trend of the number of car crashes was taken into interest to see if the rate of car crashes changed over time. To determine whether any patterns exist in the accumulation of crashes over time, accumulated crashes were plotted over time with a scattered line chart.

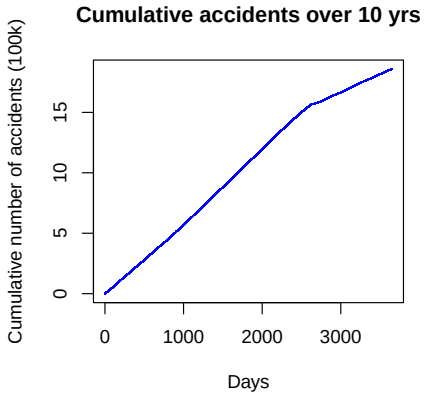


Figure 6: Precipitation (inches) versus number of crashes/day

Visually, crashes appear to increase linearly at a somewhat constant rate until 2020, at which point that

rate decreased. This trend suggests that the rate of car crashes per day may have decreased around this time. This change may reflect the decreased use of vehicles during the COVID-19 pandemic or other recent changes in driver behavior.

References

[1] (NYPD), Police Department. "Motor Vehicle Collisions - Crashes: NYC Open Data." Motor Vehicle Collisions - Crashes | NYC Open Data, 17 Feb. 2023, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

[2] "America." New York City Boroughs (USA): Boroughs - Population Statistics, Charts and Map, http://www.citypopulation.de/en/usa/newyorkcity/.

[3] National Centers for Environmental Information (NCEI). "Climate Data Online Search." Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC), https://www.ncdc.noaa.gov/cdo-web/search.