**Abstract**

Crashes each year result in thousands of lives lost and injured victims, and billions of dollars in property damage in the USA. Accurate data are required to support the development, implementation, and assessment of highway safety programs aimed at reducing this toll. The analysis of highway-crash data has long been used as a basis for determining vehicle and highway designs. It also has great influences on directing and implementing a wide variety of regulatory policies aimed at improving safety. With the improvement in statistical methodologies, researchers are able to extract more information from crash databases to guide a wide array of safety design and policy improvements. In this project, we mainly use Crash Report Sampling System (CRSS) of 2018 which contains all types of police-reported crashes, ranging from property-damage-only crashes to those that result in fatalities. Six data files in CRSS will be mainly analyzed, involving ACCIDENT Data File, ACC\_AUX Data File, VEHICLE Data File, VEH\_AUX Data File, PERSON Data File and PER\_AUX Data File to find the correlation between injury severity and different parameters like alcohol involvement, collision type, ejection, speed, road surface and so on. In addition, we will use linear regression and logistic regression to build models predicting the injury and fatal rate in certain conditions.

**Introduction and Background**

More than 1.2 million people died every year in highway-related crashes and almost 50 million people are injured all over the world. It is reported that the highway-related crash is the 5th leading causes of the death worldwide. Apart from that, highway-related crashes result in many billions of dollars in property damage. To provide vital information on motor vehicle traffic crashes, the National Highway Traffic Safety Administration (NHTSA) annually publishes nationally representative estimates of police-reported motor vehicle traffic crashes and their characteristics. From 1988 to 2015 NHTSA created national estimates using data from the National Automotive Sampling System General Estimates System (NASS GES), which sampled police crash reports from police jurisdictions across the United States. In 2016 NHTSA replaced NASS GES with the Crash Report Sampling System (CRSS), which is a sample of police-reported crashes involving all types of motor vehicles, pedestrians, and cyclists, ranging from property-damage-only crashes to those that result in fatalities. CRSS is used to estimate the overall crash picture, identify highway safety problem areas, measure trends, drive consumer information initiatives, and form the basis for cost and benefit analyses of highway safety initiatives and regulations. The CRSS obtains its data from a nationally representative probability sample selected from the more than six million police-reported crashes which occur annually. Although various sources suggest that there are many more crashes that are not reported to the police, the majority of these unreported crashes involve only minor property damage and no significant personal injury. By restricting attention to police-reported crashes, the CRSS concentrates on those crashes of greatest concern to the highway safety community and the general public. The datasets we are going to use is from CRSS 2018.

Several popular injury severity model structures have been studied in the extant literature. Sequential binary logit models1, ordered-response probit models2 and multinomial or nested logit models3 are typical examples. The sequential binary logit and ordered-response probit models represent the ordinality in the discrete categories of the injury severity. Sequential binary logit models assume that the factors determining the level of the severity change according the level of the severity itself, while ordered-response probit models assume that the same factors are correlated with all levels of injury severity. In multinomial and nested logit models, ordinality is not theoretically implemented, thus information relating to ordering of severities is not inherently captured in those structures. Compared with the previous two models, multinomial and nested logit models are structurally flexible in the sense that independent variables are not forced to be the same across all severities.

**Predictive Problem and Approach**

The maximum injury severity is of great importance to transportation analysis but difficult to predict. We select some reprehensive variables which have strong correlation with injury severity

by using heatmap, and then linear regression and logistic regression are applied to generate models to predict injury severity.

**Data Collection**

In CRSS 2018, totally 48443 representative crashes are selected from over six million police-reported crashes in 2018, involving 86105 vehicles and 120230 people. ACCIDENT contains information of the crash, date, location, injuries, environmental condition, light condition and alcohol etc. VEHICLE includes vehicles information, like make, model, body type, speed and so on, while for PERSON includes people involved, age, sex, seating position, alcohol and drug involvement information. ACC\_AUX, PER\_AUX and VEH\_AUX is the auxiliary file of the ACCIDENT, PERSON and VEHICLE file respectively. Weight should be involved when analyzing frequency to ensure unbiased and robust estimate.

**Data Analysis**

**Linear Regression**is a machine learning algorithm based on**supervised learning,** which models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. It performs the task to predict a dependent variable value (injury severity or death) based on a given independent variable (collision type, alcohol involvement and so on).

Logistic regression is a classification algorithm which assigns observations to a discrete set of variables. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to more discrete variables.

~~~~~~~~~~~to be continued~~~~~~~~~~

**Reference**

1. Saccomanno, F. F., Nassar, S. A., & Shortreed, J. H. (1996). Reliability of statistical road accident injury severity models. *Transportation Research Record*, *1542*(1), 14-23.
2. O'donnell, C. J., & Connor, D. H. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention*, *28*(6), 739-753.
3. Shankar, V., & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of safety research*, *27*(3), 183-194.