**Fatality Analysis of Motor Vehicle Crashes in U.S. States [2010 – 2015]**



Team 4

Phimonkae Wilairat

Vidhya Lakshmi Vaithilingam Palanimurugan

ALY6015: Intermediate Analytics

Prof. Dr. Roy Wada

August 09, 2022

**Introduction**

The Fatality Analysis Reporting System (FARS) contains data on a census of fatal traffic crashes within the 50 states, the District of Columbia, and Puerto Rico. Two data sets combined contains information from 2010 to 2015 with a total of 185,426 unique cases of crash. Data is collected across 58 variables which include 34 quantitative variables and 24 categorical variables. The aim of this project is to perform preliminary analysis that show descriptive statistics of the data, charts, correlation, chi-square-test, ANOVA, and to predict fatality using linear regression models.

**Analysis**

This report conducts an analysis of the fatalities and the fatality rate (percentage of fatalities to the number of motorists in the crash by sub-categories from 2010 to 2015. The analysis is based on fatal crash data. Our group expects to see some categories showed significant change in fatalities and fatality rates for a given year. First, we start with observing the shift of fatality rate in five years, from 2010 to 2015. According to Figure 1, the graph shows the fatality rate decreased in 2011 Then, the rate gradually increased from 2012, and reach a peak at 48.8% in 2014. However, in 2015, the rate dropped to 47.82%. Although we can realize the change through years, the amount of change is not significant enough that most are less than 1% or equal to 1%. Therefore, year might not be a good subset for grouping information. Our group will try other subset such as state to see more significant difference in fatality number or fatality rate.



*Figure 1.* Fatality rate by year

*Table 1.* Categorical variables related to crash cases

Table

Description automatically generated

Five categorical variables are selected to study how variables affect the crash case through six years. We compared the figure in the form of percentages as the number of crash cases is eliminated from each variable unequally due to the existence of NA. We found that the selected variables not much affect the crash case that it is presented in a small percentage compared to the proportion that does not affect. When comparing the five variables, the result shows that any of the Drivers was Under the Influence of Alcohol, Drugs, or Medication at the time of the crash had the most impact on the crash cases, and any of the Drivers Had Any Other Physical Impairments at Time of Crash the least.

*Table 2.* Quantitative variable relate to crash cases

Table

Description automatically generated

We created a new data frame by collecting the sum of quantitative variables grouped by year in which the accident occurred. Table 2 illustrates various information likes number of fatalities, count of male and female drivers involved in the crash, their age range, the number of drivers with valid license and the number of drunk drivers.

**Chi-square test**

From the data on cases for five different years listed in table 2, can we formulate hypotheses and test them. We would like to test if the crashes are dependent on the gender of the driver involved in the crash at significance level of 95%.

**Hypotheses**

Null hypothesis, Ho: Crash cases are independent of driver’s gender

Alternate hypothesis, Ha: Crash cases are dependent of driver’s gender

Significance level: 95% (alpha = 0.05)

Text

Description automatically generated

*Figure 2.* Chi-square test on driver’s gender on fatality

**Inference**

We observed that the test value is greater than the critical value. Additionally, p-value is much small than the alpha value (α = 0.05). We can conclude with 95% confidence that there is no sufficient evidence to accept the null hypothesis. Therefore, we can say that the crashes are dependent on gender of the driver involved in the accidents.

**One-way ANOVA**

Is there sufficient evidence to conclude that there is a difference in mean of the number of drunk drivers, number of drivers with invalid licenses, and number of drivers under age 16? Assume the 0.05 level of significance.

**Hypotheses**

The claim: There is a difference in mean of the number of drunk drivers, number of drivers with invalid licenses, and number of drivers under age 16.

Null hypothesis, Ho: There's no difference in mean number of drivers among three types of drivers (µ1=µ2=µ3).

Alternative hypothesis, Ha: There's a difference in mean number of drivers among three types of drivers

Text

Description automatically generated

*Figure 3.* One-way ANOVA test

S**ummarize the results**

ANOVA’s result show that p-value is much smaller than the alpha value (α = 0.05); therefore, we reject the null hypothesis that the means are all equal. In other words, there's difference in mean number of drivers among drunk drivers, drivers with invalid licenses, and drivers under age 16.

By using Tukey's ‘Honest Significant Difference’ method, the result show that the p-value after adjustment for the multiple comparisons (p adj) of each pair of driver type is lower than significant (alpha level of 0.05). Therefore, there is sufficient evidence to prove that a significant difference in mean number of drivers exists among drunk drivers, drivers with invalid licenses, and drivers under age 16.

**Simple Linear Regression**

We split the entire data into training and testing dataset in a ratio of 80:20. Two Simple Linear Regression models were created. Simple Linear Regression model 1 is designed to predict the ‘fatals’ using all the variables in the dataset while the model 2 uses only 10 variables. The results of the models are illustrated in the table below. It is observed that the model 1 with all variables is preferred as is has low AIC and BIC values.

*Table 3.* Comparison of two simple linear regression models

| **rowname** | **AIC** | **BIC** | **RSME.Training.Data** | **RSME.Testing.Data** |
| --- | --- | --- | --- | --- |
| Simple Linear Model- 1 | 51708.94 | 52242.41 | 0.34 | 0.34 |
| Simple Linear Model- 2 | 58641.80 | 58752.17 | 0.36 | 0.36 |