ECON316 Final Research Project 2019  
Jordyn Lipsey  
Foong Min Wong

**Factors Affecting Car Fatality Rates in the US**

Due to our interest in actuarial science, our initial project topic was characteristics of an individual that might affect the price of their car insurance premium depending on their risk of an accident. In our initial research into car premium prices, we investigated car accidents and what factors about the consumer could affect their risk. Our group looked at car premiums as a dependent variable, however, our group believed more data could be found assessing the risk rather than the premium. Our group narrowed our focused due to data being expensive and difficult to find in our position, and we switched our analysis to look specifically at car fatalities based on a number of independent variables. Our independent variables cover the individual in the accident and nationwide independent variables. To reach our final regression model and conclusion, we cover three sections: Literature Review, Data and Methodology, and Empirical Findings.

**Literature Review**

From the research that our group analyzed, many researchers broke down the independent variable into sections including fatal injury, disabling injury, serious injury, minor injury, other injury, and no injury resulting from an accident. Within the table below, researchers compared the independent variables to the breakdowns of the dependent variables. Regressors that researchers have included in their studies are weight of car, gender, type of vehicle, drunk drivers, age of driver, and characteristics of the other driver. Researchers also include conditional probabilities between gender and type of vehicle. In the research testing whether the weight of the car affected the type of injury received, the researchers wanted to control for alcohol use, age, and gender in order to test weight of the car effectively. The researchers used a Multinomial Logit Model (MLM) for the weights of the vehicles. Although these variables were in log, the rest of the independent variables used a linear model to estimate the dependent variable.

For the independent variable of alcohol use, the beta coefficient supports that the fatality rate will decrease when the driver is drunk, however, when the other driver is drunk, the fatality rate will increase. This is expected because drinking and driving usually results in an impaired driver harming another driver because they are more likely to drive at faster speeds. The beta coefficient of age indicate that as you age, the risk of a fatality accidents decrease, and the risk of a minor accidents increase because a driver will have more experience and can avoid the major or fatal accidents. A second study researched the effects of types of vehicles, and the conditional gender given types of vehicles. From this, researchers gathered the data that there is significance difference between male and female drivers. When female crashes are high in one area, the male crashes are low and vice versa. This is due to phycological differences that cause men and women to behave differently in situations. In the first study, researchers looked at how the weight of a car impacts the severity of accident and hold some variables constant in order to get unbiased results. From their coefficients, a driver with a heavier vehicle will decrease fatality rate, however, when the other driver has a heavier vehicle, it will increase fatality weights. This is expected because the driver with the heavier vehicle will be the safest yet cause the most damage.

The variables the demonstrated significant effects were the weights of the vehicles with a coefficient of –2.751 (driver’s car) and 1.966 (opposite driver’s car). For every one increase of weight of a driver’s car, fatality will decrease by 2.751 and for every one increase of weight of an opposite driver’s car, fatality will increase by 1.966. Some variables that could be possible indicators of risk are credit score, garaging locations, number of years since getting their license, accident history, and ticket history. All these are factors to estimate the likelihood of whether a driver will get into an accident. Lower credit scores, rural areas, new drivers, and high accident and ticket history will increase the risk of the driver getting into an accident. These variables could have been identified by the researchers, however, the research covered a significant number of variables.

**Data and Methodology**

Our group is researching the effects on car fatalities based on the independent variables of state driver is located, unemployment rate, income, beer tax, and miles. We chose to include the variable state because it will include the state’s specific laws about minimum sentencing, mandatory community service, preliminary breath test law, weather, and how this reflects in driving patterns. We included personal information about the average driver such as unemployment, income, and average miles driven per year. This information gives information about the state of mind of the drivers and the frequency an individual will be drinking. The beer tax independent variable was included to see how beer consumption affects drunk driving and alcohol related fatalities. Two variables not included in this set of data was the gender of the driver and the weights of the vehicles involved in the accidents. Fatalities would decrease for the driver when the driver’s car is heavier and vice versa. Information about the make and model of a vehicle could be beneficial to see how fatalities can be prevented given an accident occurs. Additionally, insurance charges young male drivers a higher insurance because they are considered riskier drivers compared to women.

After reviewing our independent variables, we decided to use a log functional form for income since income ranges from zero dollars to hundreds of thousands. By applying the log function, we observe the data on a more appropriate scale rather than the large range beforehand. We determined the rest of our variables follow a linear functional form after observing the plots between each independent and fatalities. Our coefficients were as expected once a regression model was summarized with the equation:

We predicted that increasing beer tax would cause people to drink less and therefore decrease alcohol related car fatalities. When we increase income, individuals can purchase more vehicles and increase the number of cars on the road. The more cars on the roads, the higher probability an accident will occur from oversaturation. However, the more miles driven will decrease the about of car fatalities due to more experienced drivers on the road. We predict that many of the states are positive or negative based on the laws they have like texting while drinking, minimum sentencing, mandatory community service, preliminary breath test law, weather, or road maintenance. The unemployment would have an inverse effect on car fatalities because less people are on the roads driving to or from work. For every one dollar increase in beer tax, car fatalities will decrease by 0.0234; for every 1 percent increase in income, car fatalities will increase by 0.03727; for every one increase in miles driven, car fatalities will decrease by 0.002239; car fatalities varies on each state, majority of the coefficients for states are negative (37 out of 47 states), and for every one increase in the unemployment rate, car fatalities will decrease by 0.1572.

Our regression has an adjusted R-squared value of .9921, which means our model predicts 99.21% of the variation in car fatalities. With a high R-squared and having expected signs, we have a quality model for predicting car fatalities.

The data we gathered was called “Fatalities” and inputted from the public library of RStudio data sets. This data was collected by the US Department of Transportation Fatal Accident Reporting System from 1982-1988 and contains 336 observations on 37 variables. Of the given variables, the independent variables we used for our model were state, unemployment, income, beer tax, and miles and we used fatalities as our dependent variable. The beer tax is used to measure the state’s overall alcohol taxes by using the tax on one case of beer in dollar value. The independent variable, miles, is the average miles traveled per driver and was obtained from the Department of Transportation, income per capita in 1987 dollars was obtained from the US Bureau of Economic Analysis, and the unemployment rate for the general US population was obtained from the US Bureau of Labor Statistics. The dependent variable, fatalities, is the number of vehicle fatalities and includes night, alcohol-related, and age-base fatalities.

**Empirical Findings**

In our model, we used the independent variables unemployment, log(income), beer tax, drinking age, miles, region, miles^2 to estimate our dependent variable, fatality rate. All our independent variables are statistically significance at 1% level, except for unemployment, beertax, and possibly regionnortheast. Our R squared value is 0.658 and our adjusted r squared value is 0.648. This means our model explains 65.8 percent of the variation in fatalities. We use linear regression to estimate our model with one log coefficient. The one log coefficient accounts for the large range of income. Income of a single individual can range from 20,000 to 300,000, and the log functional forms accounts for this. Based on further testing, we discovered our model has heteroskedasticity and serial correlation, which means our model is no longer BLUE and our Type I error has increased.

For every one unit increase in unemployment, fatality rate will slightly increase by 0.00201670582. This coefficient does not follows expectations because as less people are working, there are less commuters on the roads driving to work. This coefficient is also insignificant. For every one percent increase in income, fatality rate will decrease by 1.19966453733. This coefficient matches our expectation because as income increases, we expect life style changes that improving driving habits. These life style changes may include having children, more expensive vehicles, closer commutes, etc. For every one unit increase in the beer tax, fatalities will increase by 0.05195234744. This coefficient did not match our expectations because as beer tax increase, beer consumption will decrease. This implies that alcohol we believe that alcohol fatalities would increase because people would change their order to a different substance, possibly with a higher alcohol percent than beer. For every one unit increase in the drinking age, fatalities will decrease by 0.03585965773. This coefficient matches our expectations because with less young adults drinking and driving, fatalities will decrease. For every one mile increase in miles driven, fatalities will increase by 0.00037791681 by a decreasing rate of 0.00000001102miles2. This coefficient matches expectations because the more miles driven, the higher probability of an accident and fatality happening. When northeast has a value of 1, fatality rate will increase by 0.10200690137. When northeast has a value of 1, fatality rate will increase by 0.37379140100. When northeast has a value of 1, fatality rate will increase by 0.60658334319.

We are 90% confident that the interval [-0.016, 0.02] contains the true coefficient of unemployment. We are 90% confident that the interval [-1.5, -0.92] contains the true coefficient of log(income). We are 90% confident that the interval [-0.032, 0.14] contains the true coefficient of beer tax. We are 90% confident that the interval [-0.072, 0.00021] contains the true coefficient of drinking age. We are 90% confident that the interval [0.0003, 0.00045] contains the true coefficient of miles. We are 90% confident that the interval [-0.000000014, -0.0000000084] contains the true coefficient of miles.

Since our F statistic is high, we have sufficient evidence to say that the full linear model with our independent variables is useful in predicting fatalities. To fix the heteroskedasticity and serial correlation, we decided to add independent variables based on state. We divided the states into region dummy variables in order to decrease the effect on our degrees of freedom. This increases the standard error and lowers the t-values to avoid Type I error. One standardized coefficient we used was the fatality rate per people, rather than using fatality totals and population as an independent variable. Standardizing the coefficient helps scale the term and compares the relative importance of each coefficient in the regression model.

**Conclusion**

Compared to other research paper on fatalities from automobile accidents, our research contains very different independent variables. Many papers focused on characteristics of the individual driving and the car, but rarely measured national independent variables occurring at the time of collision. Their goal was to determine if a single coefficient was significant in determining fatalities while holding other variables constant. One research paper discovered that a 500 kg increase in the weight of the other car increases fatalities by 70% and another paper determined there was a significance n male and female drivers. During our analysis, we discovered significant coefficients that influence on our dependent variable and account for 68.4% of the variation in the rate of fatalities and this met our purposes of discovering several significant variables to predict fatality rates. Based on our finding, we found several significant results such as income, drinking age, miles, and the region. Our most influential coefficient was income with a value of -1.19966453733; having income decrease fatalities so drastically compared to our other variables means there is another reason for trying to increase income of the average American.

For future research to predict fatality rates, we encourage a combination of individual statistics along with national statistics. Personal characteristics of the car accident, car, and driver along with national data such as region and rates will increase the percentage of variability your model accounts for. Another alteration for future research would be to break down our dependent variable future into fatalities based on age of driver, time of day, alcohol related, falling asleep at the wheel, etc. Our group discovered insignificant results in unemployment and beer tax, but theory suggests these could be significant when analysing fatalities only based on alcohol involved. More accurate conclusions and suggestions for reducing fatalities could be introduced with broken down data.

**Table 1**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Fatality Rate | numeric. Vehicle fatalities per 10000 people. |
| Unemployment | numeric. Unemployment rate. |
| Income | numeric. Per capita personal income in 1987 dollars. |
| Beertax | numeric. Tax on case of beer. |
| Drinkage | numeric. Minimum legal drinking age. |
| Miles | numeric. Average miles per driver. |
| Regions | Factor indicating region of the US (East, Northeast, Midwest, South) |

**Table 2: Descriptive Statistics**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **fatal\_rate (%, per 10000 people)** | **unemp (%)** | **Income ($)** | **beertax**  **(%)** | **Drinkage (years old)** | **Miles (mi)** | **region** |
| median | 0.000000019559549565 | 7.00 | 13763.13 | 0.353 | 21.000 | 7796.22 | 3.00 |
| mean | 0.000000020404437844 | 7.35 | 13880.18 | 0.513 | 20.456 | 7890.75 | 2.54 |
| SE.mean | 0.000000000311066202 | 0.14 | 122.91 | 0.026 | 0.049 | 80.50 | 0.06 |
| CI.mean.0.95 | 0.000000000611889184 | 0.27 | 241.78 | 0.051 | 0.096 | 158.36 | 0.12 |
| var | 0.000000000000000033 | 6.42 | 5076217.59 | 0.228 | 0.808 | 2177568.69 | 1.21 |
| std.dev | 0.000000005701937675 | 2.53 | 2253.05 | 0.478 | 0.899 | 1475.66 | 1.10 |
| coef.var | 0.279445957712157800 | 0.34 | 0.16 | 0.931 | 0.044 | 0.19 | 0.43 |

**Table 3: Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Residuals** |  |  |  |  |
| Min | 1Q | Median | 3Q | Max |
| -0.7971 | -0.2127 | -0.0074 | 0.1684 | 1.4821 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Coefficients** |  |  | |  |  |
|  | **Estimate** | **Std. Error** | | **t value** | **Pr(>|t|)** |
| (Intercept) | 11.60527592492 | 1.73813270144 | | 6.68 | 0.0000000001052635 \*\*\* |
| unemp | 0.00201670582 | 0.01110448000 | | 0.18 | 0.86 |
| log(income) | -1.19966453733 | 0.16912015061 | | -7.09 | 0.0000000000081772 \*\*\* |
| beertax | 0.05195234744 | 0.05091293566 | | 1.02 | 0.31 |
| drinkage | -0.03585965773 | 0.02186735977 | | -1.64 | 0.10 |
| miles | 0.00037791681 | 0.00004571539 | | 8.27 | 0.0000000000000035 \*\*\* |
| regionnortheast | 0.10200690137 | 0.06423746097 | | 1.59 | 0.11 |
| regionsouth | 0.37379140100 | 0.05573835570 | | 6.71 | 0.0000000000882628 \*\*\* |
| regionwest | 0.60658334319 | 0.05730593236 | | 10.59 | < 0.0000000000000002 \*\*\* |
| miles2 | -0.00000001102 | 0.00000000161 | | -6.84 | 0.0000000000380655 \*\*\* |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | |
| Residual standard error: 0.34 on 326 degrees of freedom | | | | | |
| Multiple R-squared: 0.658 | | | Adjusted R-squared: 0.648 | | |
| F-statistic: 69.6 on 9 and 326 DF | | | p-value: <0.0000000000000002 | | |

**Table 4: Confidence Intervals (90%)**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | 5 % | 95 % | R codes |
| unemp | -0.016 | 0.02 | confint(lmfit4, 'unemp', level=0.9) |
| log(income) | -1.5 | -0.92 | confint(lmfit4, 'log(income)', level=0.9) |
| beertax | -0.032 | 0.14 | confint(lmfit4, 'beertax', level=0.9) |
| drinkage | -0.072 | 0.00021 | confint(lmfit4, 'drinkage', level=0.9) |
| miles | 0.0003 | 0.00045 | confint(lmfit4, 'miles', level=0.9) |
| miles2 | -0.000000014 | -0.0000000084 | confint(lmfit4, 'miles2', level=0.9) |

**Table 5: Heteroskedasticity**

|  |
| --- |
| R codes: par(mfrow=c(2,2)) # create 4 charts in 1 panel    The top-left is chart of residuals vs fitted values, while in the bottom-left one, it is standardized residuals on Y axis. If there is absolutely no heteroscedasticity, we should see a completely random, equal distribution of points throughout the range of X axis and a flat red line. In our case, from the top-left plot, we notice that the red line is slightly curved and the residuals increase as the fitted Y values increase. Thus, heteroscedasticity exists. |

**Table 6: Serial Correlation**

|  |  |
| --- | --- |
| **Durbin-Watson test** | **Rcodes** |
| data: lmfit4  DW = 0.7, p-value <0.0000000000000002  alternative hypothesis: true autocorrelation is greater than 0  n = 336, k = 6, dL = 1.707, dU=1.831 (According to Durbin-Watson Table)  Since dW = 0.7 (< dL), we reject null and conclude that there is a serial correlation. | > require(lmtest)  > dwtest(lmfit4) |

**Table 7: Multicollinearity**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Importance Score** | | | | **Rcodes** |
| **Variable** | **GVIF** | **Df** | **GVIF^(1/(2\*Df))** | > library(car)  > vif(lmfit4) |
| unemp | 2.3 | 1 | 1.5 |
| log(income) | 2.1 | 1 | 1.4 |
| beertax | 1.7 | 1 | 1.3 |
| drinkage | 1.1 | 1 | 1.1 |
| miles | 13.3 | 1 | 3.7 |
| region | 2.7 | 3 | 1.2 |
| Miles2 | 11.0 | 1 | 3.3 |

**References**

Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. (2003, March 15). Retrieved from <https://www.sciencedirect.com/science/article/pii/S0001457502001355>

Killing kilos in car accidents: Are external costs of car weight internalized? (2013, July 24). Retrieved from <https://www.sciencedirect.com/science/article/pii/S2212012213000142>

What determines the price of an auto insurance policy? (n.d.). Retrieved from <https://www.iii.org/article/what-determines-price-my-auto-insurance-policy>

Clarke, W. (2018, November 07). Car Insurance Rates: Factors That Can Cause Them to Increase. Retrieved from <https://www.aarp.org/auto/car-maintenance-safety/info-2018/car-insurance-rate-factors.html>

AER. (n.d.). Retrieved from <https://www.rdocumentation.org/packages/AER/versions/1.2-6/topics/Fatalities>

Ruhm, C. J. (1996). Alcohol Policies and Highway Vehicle Fatalities. *Journal of Health Economics*, **15**, 435--454.

Stock, J. H. and Watson, M. W. (2007). *Introduction to Econometrics*, 2nd ed. Boston: Addison Wesley.