Traffic Fatalities Regression Analysis

Vincent Kim

CSU East Bay

Note: All figures/tables referenced are shown in the Appendix.

Introduction.

It is no secret that traffic accidents are one of the leading causes of human fatalities in the U.S. In the past few years, the United States has seen a steady rise of fatality rates, reaching 400,000 total deaths for consecutive years in 2016 and 2017.[2] Preventable causes, such as traffic accidents, have also been identified as the prominent cause of death among adolescents, with total death rate among teenagers soaring 12% higher between 2014 to 2016 compared to 1999 to 2013.[3] Car accident fatalities have grown to be a worldwide predicament and it is this amplifying concern that induced the purpose of this study. This research examines deeply to uncover the main factors of traffic accidents fatalities in order to find and advocate potential methods to make the roads safer.

Data Description.

Data used for this study was amassed from the Fatality Analysis Reporting System (FARS), which was created by the National Highway Traffic Safety Administration in the United States. FARS records data information of any car accidents that involves a fatality within the 50 states, Puerto Rico, and Washington DC. The initial dataset accumulated over 34,000 accidents that occured in 2016, but through data cleaning in order to account for missing and unnecessary information for the study, it was reduced to approximately 1,500 data points. The response variable (Y) is injury severity, a binary variable that equals 0 if the driver who caused the accident was injured or unharmed and 1 if the driver was killed from the crash. FARS had over 30 columns of information, but most were not applied in the study since variables like longitude, latitude, date, county, street name, etc were not necessary for the analysis. Thus, the predictor variables chosen in the initial model consist of numeric variables such as travelling speed and age of the driver, and categorical ones like route (where the accident occured), manner of collision (where the cars collided), light condition (how bright the day of crash was), weather (grouped into light, moderate, severe), if drinking was involved (yes or no), body type of the car (small or large), and if the driver was licensed (yes or no). Table 1, shown in the Appendix, provides a more laid out visual description of the x variables. All in all, with injury severity acting as the response, nine predictors were selected to be tested in the statistical analysis.

Statistical Analysis and Results.

Because injury severity was a binary variable, logistic regression was selected to be the model of choice. The full model was initially tested in order to visualize how sufficient and significant the model would be with all predictor variables present; however, the model formulated a total AIC of 2117.1 with all predictor variables except age, drinking, and travelling speed being not significant. Because of this, the step function was used to find and optimize a more precise model that would help determine the predictor variables that should be left in and out.

The “forward” choice was selected to conduct the step procedure out. Comparing the null model to the final model that the step function outputted, the AIC did not decrease all that much, from 2193.70 to 2142.90, but it is deciphered that age, drinking, travelling speed indeed, as well as body type were the final predictor variables to be chosen for the model. However, what was interesting was that the body type variable was included in the final model despite being exceedingly non-significant. Therefore, the question was brought whether body type should be deleted from the final model or kept in. As a result, an ANOVA test was carried out in order to examine the significance of the body type variable. According to Table 3, the p-value was

0.00004.492 < 0.05; thus, the null hypothesis is rejected and can conclude that body type should be kept in the final model.

With body type being determined to be in the final model, the following results can be concluded. Table 2 shows the output of the final model, where INJ_SEV ~ factor(BODY_TYP) + AGE + factor(DRINKING) + TRAV_SP. Age, drinking, and travelling speed are significant, with age having a p-value of 0.00000145, drinking having a p-value of 0.0000315, and travelling speed with 0.0319. Thus, it can be concluded that age, drinking, and travelling speed have an effect on injury severity of a driver in a car accident. Interpreting these numbers into context, for example, a person who drinks has odds of dying in a car crash e^0.514875 = 1.67 times higher than those who don't drink; or in other words, a drunk driver increases their odds of getting killed by about 51%. As a result, with the x and y variables now confirmed from the final model, it was sought out to transition the study to examine deeper into the relationships between the predictor variables and response.

Box plots, histograms, and scatter plots were created in order to best facilitate the connection between the x variables and the y. The first relationship examined was drinking vs age as shown in Figure 1. Looking at the boxplot, it can be identified that there is more weight from younger people since the boxplots for both drinking (yes and no) is a bit skewed towards the bottom. Furthermore, the median for Drinking:Yes is lower than Drinking:No; thus, this is a sign that younger people tend to drink more than older folks. The second relationship studied was age vs travelling speed vs injury severity, using a scatter plot in Figure 2. From the plot, there is a plethora of points to the left side of the graph than the right, further showing that younger people tend to have more incident rates and death counts from car accidents. The scatter plot also indicates that a fair amount of crashes occur around a travel speed of 50. Furthermore, looking at the death rates (red points), it should be noted that there are far more red dots at the travel speed 50-100 region and age <50, implying that younger people tend to drive faster and lose their lives from car accidents. The final relationship investigated was drinking vs travelling speed and injury severity, as plotted out in Figure 3 and 4. Figure 3 clearly exhibits how people who drink tend to drive faster than those who do not drink, as marked by the higher distribution and median of the Drinking:Yes boxplot. Figure 4 expresses the fatality ratio for yes/no groups in drinking, where red indicates death and blue is survival of the driver in the crash. While there is an immense amount of non-drinking incidents, there is a higher fatality ratio for drunk driving. Looking, at the non-drinking bar graph, there is a little more than 50% chance of survival than death, but glancing over the drinking graph, drivers are more likely to get killed in a crash than surviving. With the usage of plots and graphs, one can see how the relationships between these predictor and response variables are exactly outlined.

Discussion.

By no means was this research the perfect study as there were numerous complications and missing points that could have been used for future work and extensions. First of all, the were a substantial amount of missing/unknown information that could have been useful in influencing the significance and accuracy of the statistical analysis if the material was given. To add on, while the FARS dataset assembled various amounts of variables convenient for the study, there were also several other factors that were left out that would have been of great use for the research. For instance, a blood alcohol content (BAC) component would have helped the study in examining how much alcohol impacts a person's ability to drive and potentially become involved in an accident.

There were also some possible extensions that could have been implemented if time permitted. For example, it would have been useful to conduct a confusion matrix that would have aided in examining the overall accuracy of the final model. Another extension would have been to further study the body type variable in order to divulge the relationship the variable might have with the other predictors and response. This might have helped reveal why the ANOVA test was significant and why body type was eventually kept in the final model.

References

2016 FARS / CRSS Coding and Validation Manual. *FARS.* Retrieved from https://ftp.nhtsa.dot.gov/fars/FARS-DOC/Coding%20and%20Validation%20Manual/2016%20FARS%20CRSS%20Coding%20and%20Validation%20Manual-812449.pdf

Bomey, N. (2018). U.S. vehicle deaths topped 40,000 in 2017, National Safety Council estimates. *USA Today.* Retrieved from https://www.usatoday.com/story/money/cars/2018/02/15/national-safety-council-traffic-deaths/340012002/

Howard, J. (2018). A 'wake-up call' about what's killing America's young people. *CNN.* Retrieved from https://www.cnn.com/2018/06/01/health/youth-injury-death-rate-cdc-study/index.html

Appendix.

Figure 1: Boxplot between age vs drinking

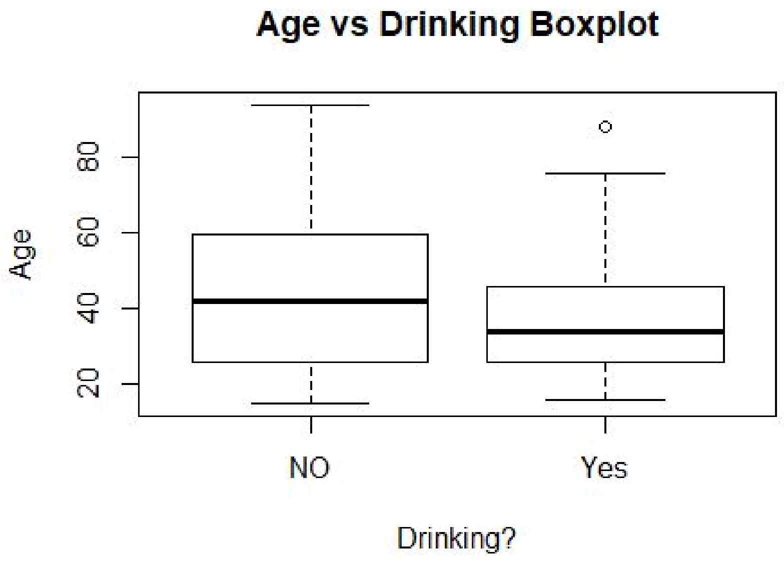


Figure 2: Scatterplot of Age*Travelling Speed grouped by Injury Severity

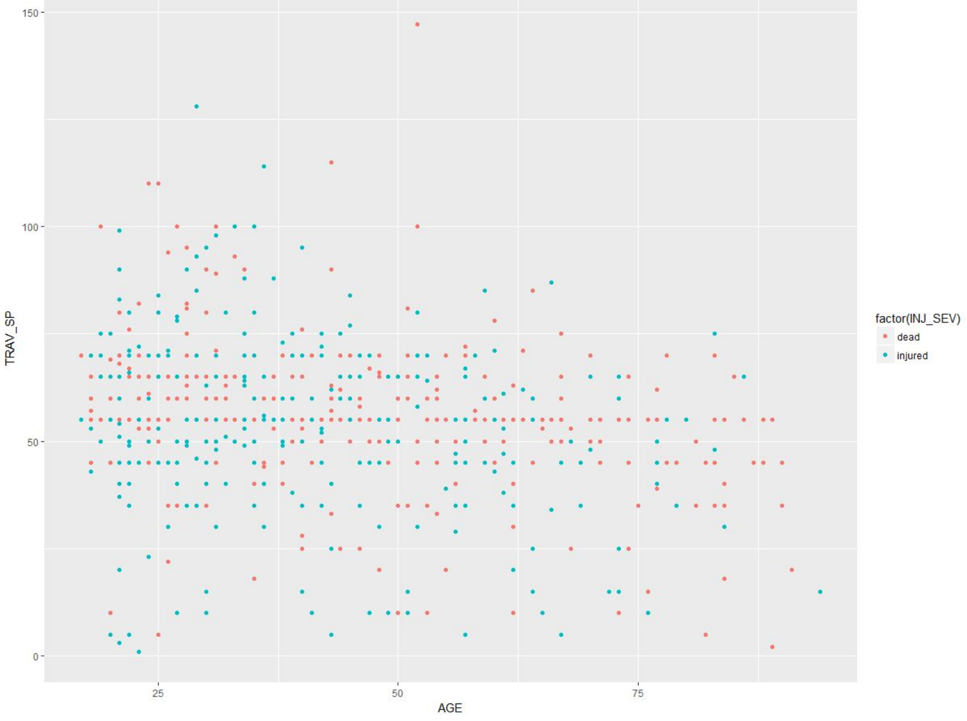


Figure 3: Boxplot of drinking vs travelling speed

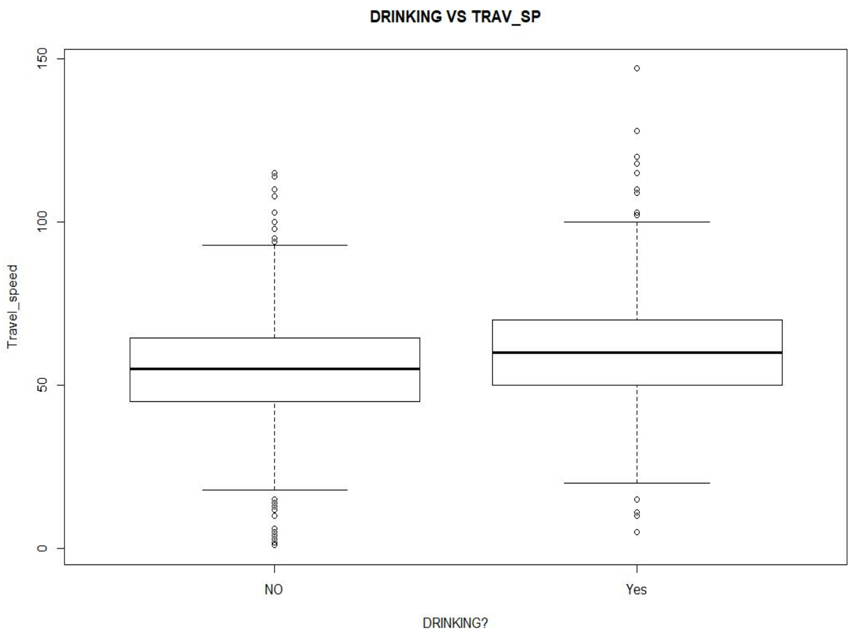
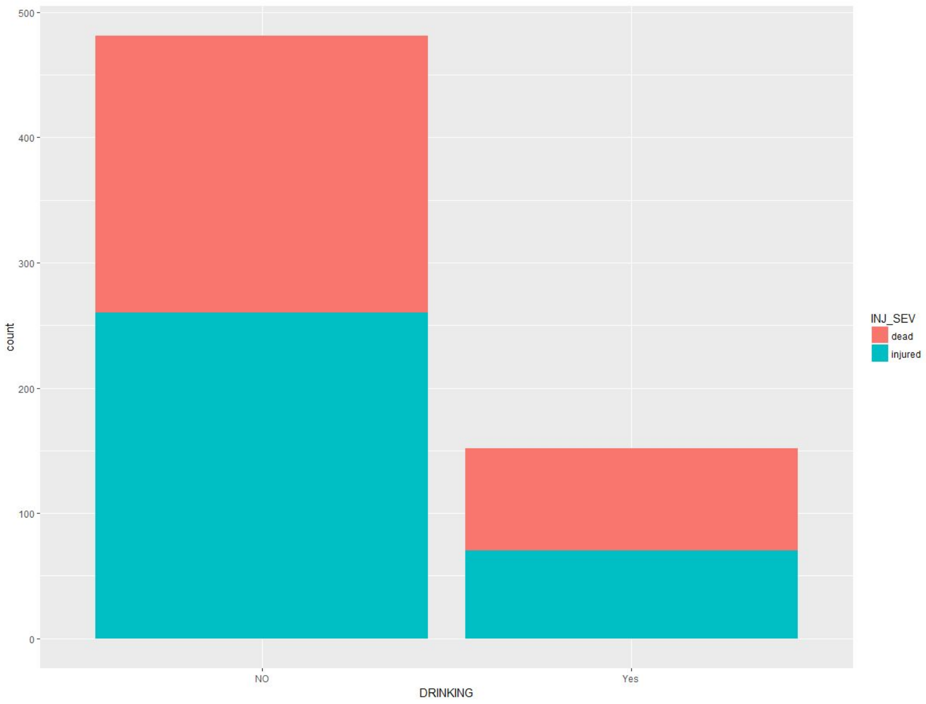
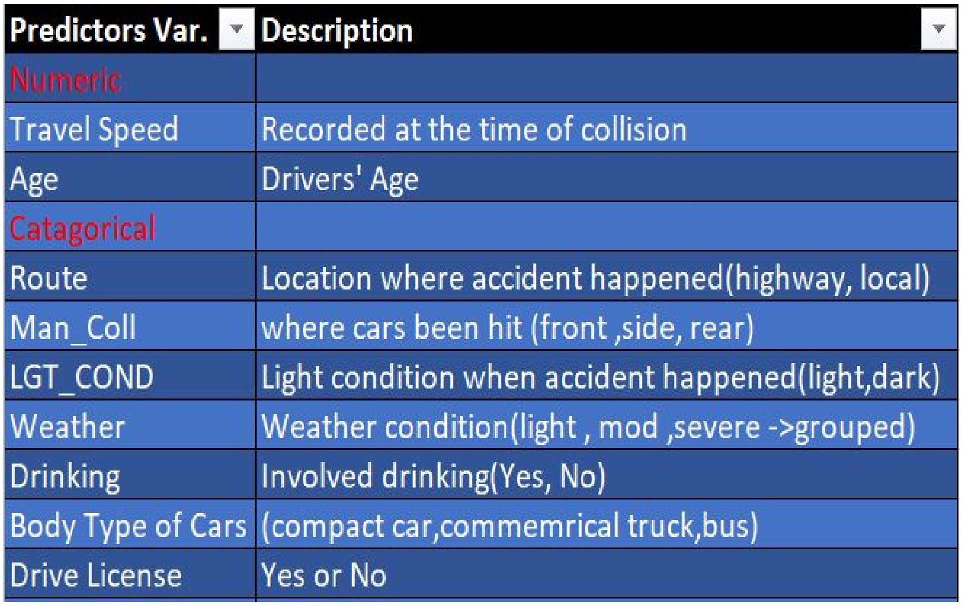


Figure 4: Bar graph of drinking vs injury severity



Note: All figures/tables referenced are shown in the Appendix.

Table 1: Initial predictor variables descriptions



Predictors Var.	Description
Numeric	
Travel Speed	Recorded at the time of collision
Age	Drivers' Age
Catagorical	
Route	Location where accident happened(highway, local)
Man_Coll	where cars been hit (front ,side, rear)
LGT_COND	Light condition when accident happened(light,dark)
Weather	Weather condition(light , mod ,severe ->grouped)
Drinking	Involved drinking(Yes, No)
Body Type of Cars	(compact car,commemrical truck,bus)
Drive License	Yes or No

Table 2: Final model table w/ coefficient statistics, values w/ stars indicate significance

Coefficients:	Estimate	Std. Error	Z value	P-value
Intercept	-15.376688	393.043232	-0.039	0.9688
Factor(BODY_TYP) LORRY	14.594409	393.043207	0.037	0.9704
Factor(BODY_TYP)small	14.093767	393.043195	0.036	0.9714
AGE	0.013683	0.002840	4.818	1.45e-06 ***
Factor(DRINKING)Yes	0.514875	0.123688	4.163	3.15e-05 ***
TRAV_SP	0.006003	0.002798	2.145	0.0319 *
AIC: 2142.90				

Note: All figures/tables referenced are shown in the Appendix.

Table 3: ANOVA comparison between Model 1 and Model 2
Model 1: INJ_SEV ~ factor(BODY_TYP) + AGE + factor(DRINKING) + TRAV_SP
Model 1: INJ_SEV ~ AGE + factor(DRINKING) + TRAV_SP

				Deviance	**P-value**
1	**1582**	**2130.90**			
2	**1584**	**2151.0**	**-2**	**-20.021**	**4.492e-05 *****

Final R Code:

```
null = glm(INJ_SEV ~ 1, family = binomial, data=fatal)
summary(null)
full = glm(INJ_SEV ~ factor(ROUTE) + factor(MAN_COLL) +
factor(LGT_COND)+factor(WEATHERlight)+factor(BODY_TYP)+AGE+factor(DRINKING)
+TRAV_SP+factor(L_TYPE), family=binomial, data=fatal)
summary(full)
step(null, scope=list(lower=null, upper=full),direction="forward")
final = glm(INJ_SEV ~ factor(BODY_TYP) + AGE + factor(DRINKING) + TRAV_SP, family
= binomial, data=fatal)
summary(final)
final2 = glm(INJ_SEV ~ AGE + factor(DRINKING) + TRAV_SP, family = binomial,
data=fatal)
summary(final2)
anova(final2, final, test="Chisq")
round(exp(cbind(Estimate=coef(final), confint(final))),2)

boxplot(AGE ~ DRINKING, data=fatal, main="Age vs Drinking Boxplot", xlab = "Drinking?",
ylab = "Age", ylim=c(0,100))
plot(fatal$TRAV_SP, fatal$AGE, groups=fatal$DRINKING, xlab = "Age", ylab = "TRAV_SP",
ylim=c(0,150), col=c("red", "blue") )
plot(fatal$TRAV_SP, fatal$AGE, groups=fatal$INJ_SEV, xlab = "Age", ylab = "TRAV_SP",
ylim=c(0,150), col=c("red", "blue") )
boxplot(TRAV_SP ~ DRINKING, data=fatal, main="Drinking vs Travelling Speed", xlab =
"Drinking?", ylab = "Travelling speed", ylim=c(0,150))
counts = table(fatal$INJ_SEV, fatal$DRINKING)
barplot(counts, col=c("blue", "red"))
```

Note: All figures/tables referenced are shown in the Appendix.