Car Crashes: A Logistic Regression Analysis

Christian Davis and Jackson Curtis

April 1, 2018

1 Introduction

The federal highway administration (FHWA) is responsible for improving highway and road safety for states and local agencies across the United States. To help make roads safer, the FHWA implemented the General Estimates System (GES) to collect information such as weather conditions, air bag conditions, speed limit, the hour of the crash, and other factors that may contribute to the severity of a car crash. In 2015, there were approximately 93 deaths per day on average, which corresponds to over 34,000 total deaths due to motorized vehicle accidents in the U.S. These numbers are staggering. The goal of this analysis is to understand the relationship of various factors related to a vehicle-related accident gathered using the GES and the probability of a serious injury. Understanding these relationships will allow us to determine with factors contribute to safer roads and will allow us to predict the severity of accidents in various regions across United States.

2 Data

Table 1: Relationship of Alcohol, Air Bag, and Curvature of the Road and Severity

	Status	Not Severe	Severe
Alcohol	Yes	351	724
	No	4198	3330
Air Bag	Deployed	1278	2027
	Not Deployed	2909	1728
	Not Applicable	362	299
Road Type	Curved	466	675
	Straight	4083	3379

The data consist of 8,603 car crash observations each measuring 12 corresponding factors that may contribute to the seriousness of the vehicle-related accident in 2013. These factors are time of the accident, lighting conditions, weather, alcohol status, intersection type, seat belt status, airbag status, road type, number of lanes, speed limit, curvature of the road, and road conditions. Severity is a binary response variable that indicates if at least one involved in the accident sustained a serious injury. So a 0 suggests that no one involved in the accident sustained a serious injury while a 1 indicates that at least one person did sustain a serious injury.

We explored all variables to understand what categories were present, how big each sample size was, how predictive of severity it was, and whether the effect on severity was monotonic. The scatter smooth plot of hour in figure 1 suggest a more severe accident is more likely to occur late at night or early in the morning. Additionally, the scatter smooth plot of speed limit in figure 2 suggest that as speed limit increases, the probability of a more severe accident also increases. These findings are consistent with intuition. Table 1 contains a cross tabulation table to help understand the relationship of alcohol, airbag status, and curvature of the road and severity. Note that accidents involving alcohol are significantly more likely to be severe. We also learn that accidents where the air bag deployed are more likely to be severe than when the air bag was not deployed or when air bag status was not applicable. Finally, roads with any form of curvature are more likely to be severe than straight roads.

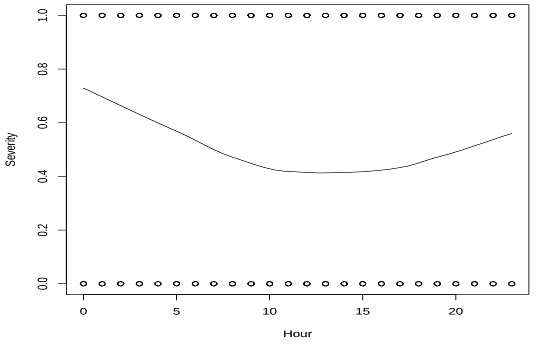

Figure 1: Hour with smoothed curve to show trend

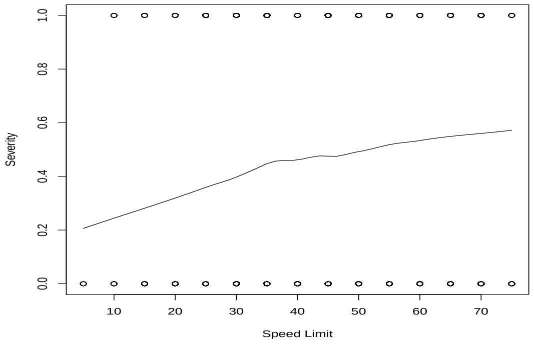

Figure 2: Speed limit vs. Severity

3 Choosing a Model

Because we are interested in being able to describe scenarios that likely resulted in a severe crash, we want a model with interpretable parameters. Each parameter should be able to describe how each explanatory variable changes the likelihood of a severe crash. One model that allows us to do this in logistic regression. In logistic regression:

$$Y_i \sim Bern(p_i) \tag{1}$$

where

$$log(\frac{p_i}{1-p_i}) = x_i'\beta$$

so our y_is are our 0/1 (not severe/severe) outcomes, and each of those has a unique probability, p_i, of happening. We will model the log odds as the sum of our explanatory variables multiplied by their coefficients, and then use maximum likelihood estimation to compute the coefficients. We can back-transform to get our predicted probability:

$$p_i = \frac{exp\{x_i'\beta\}}{1+exp\{x_i'\beta\}} \tag{2}$$

This model allows us to interpret an individual explanatory variable's effect on the odds of a severe crash. For example if a categorical variable (present/not present) had a coefficient of 1, switching from not present to present would make a severe outcome exp(1)=2.7 times more likely to occur.

This model assumes that each of our observed crashes are independent. Because crash data is gathered on an incident basis (and not a per car basis), this seems like a safe assumption. The only possible violation would be a case where a crash caused some kind of change to the road (debris, traffic back-up) that resulted in another crash far enough away to be classified as a separate incident. It is safe to assume events like this are rare and will not ruin our parameter estimates.

The other assumption of this model that each explanatory variable is monotone in its effect on the odds of a severe crash. For example, a continuous variable cannot increase the odds up to a certain point and then decrease them thereafter. This is not a problem for categorical variables because when they are coded to 0 or 1 they are forced to be monotonic. For continuous variables, we can check this assumption and introduce non-linear terms or other transformations when this appears to be violated. Figure 2 shows that speed limit appear to be monotonic. However, Figure 1 clearly is not monotonic. We applied a sine/cosine transformation to help with monotonicity and enforce that the effect at the end points is the same. Figure 3 shows that our two transformed hour variables both have a monotonic relationship with severity, so our assumption is met.

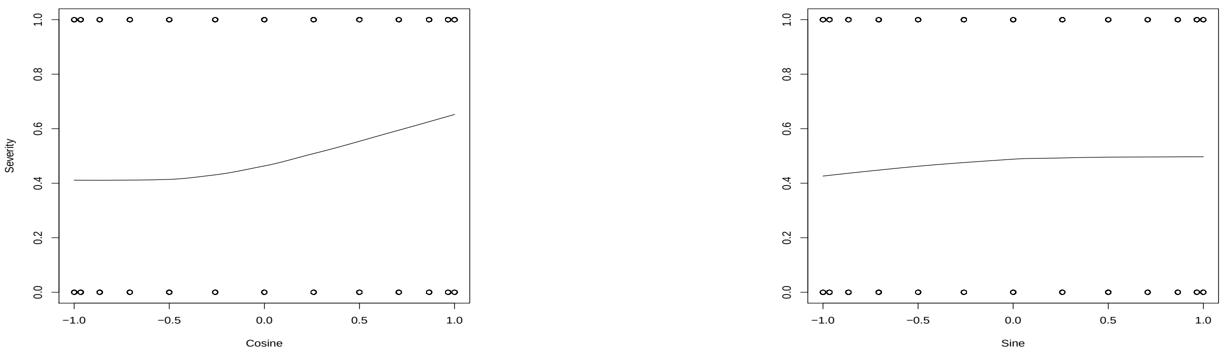

Figure 3: The cosine and sine transformations have a monotonic relation with severity.

4 Model Evaluation

We had many different options for modeling car crash severity. Our explanatory variables have many possible levels. Unfortunately, some of these levels have very few observations which would lead to unstable and hard to interpret parameter estimates. To solve this, we sought to create categories that shared similar interpretations but had a large sample size. For example, we collapsed several air bag categories (front, side, top, etc) into just three: air bag deployed, air bag did not deploy, and not applicable. Likewise, with weather we tried to group similar conditions (blowing sand and blowing snow) into one category. Additionally, we decided to treat the number of lanes as a continuous variable (because it had a monotone relationship with severity) and speed limit as continuous, although we could also have explored creating distinct groups from their levels.

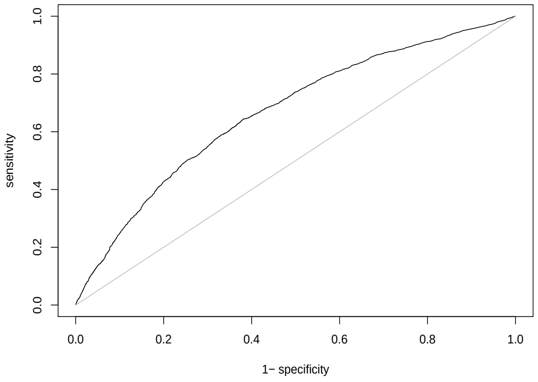

Figure 4: In-Sample ROC Curve

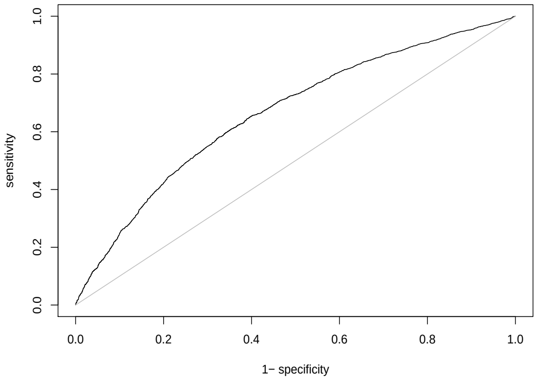

Figure 5: Out-of-Sample sample ROC Curve

From there, we used model selection to remove variables that did not make a significant contribution to the model. For model selection we used Bayesian Information Criterion to provide a simple, interpretable model. For large data sets like ours, the difference between AIC and BIC can be large because the penalty for BIC is much bigger (9.06 vs. 2 per parameter in our case), but BIC models tend to be simpler and more interpretable, which we believed would allow us to answer our question of interest: what are the main factors that distinguish a severe car crash?

Our model used bi-directional (forward and backward) step selection to minimize BIC. After we obtained a model with the best BIC, we did several experiments to compare this base BIC with various model modifications such as combining groups that that could reasonably be assumed to have a similar effect on severity and adding splines to see if they improved the fit of our continuous variables (they did not). Our final explanatory variables were: the time of day with a sine/cosine transformation to enforce circularity, whether alcohol was present, the use of air bags (not deployed, deployed, not applicable), the speed limit with a two degree spline, and whether the road was straight or curved.

R^2 can be used to measure model fit, however when using logistic regression good classification models may still have low R^2 values. Thus, we will use the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve to measure model fit. A ROC curve compares sensitivity to false positive rates for many thresholds. An in-sample ROC curve is displayed in figure 4. We can think of an ROC curve as a cost-benefit curve. The ROC curve indicates that we must sacrifice the false positive rate to increase the true positive rate. An in-sample AUC is restricted between 0.5 and 1. Generally, an larger AUC is better. The AUC we obtained for our model is 0.67.

Another method to measure model fit is to report in-sample sensitivity (percent of true positives), specificity (percent of true negatives), positive predictive value (percent of correctly predicted severe crashes), and negative predictive value (percent of correctly predicted non-severe crashes). These values are reported in table 2. In order obtain these estimates, we need determined a cutoff probability to classify a crash as severe or non-severe. As seen in the plot in figure 6, we choose a threshold where misclassifications are minimized. We obtained the threshold of 0.49. This means that we will slightly classify more crashes as not severe because our cutoff value is slightly less than 0.5. Our sensitivity value is fairly low. This means that we only predict about half of severe crashes as severe. However, our specificity is pretty high. We are happy with this because we predict the majority of non-severe crashes as non-severe and typically a car crash is not severe.

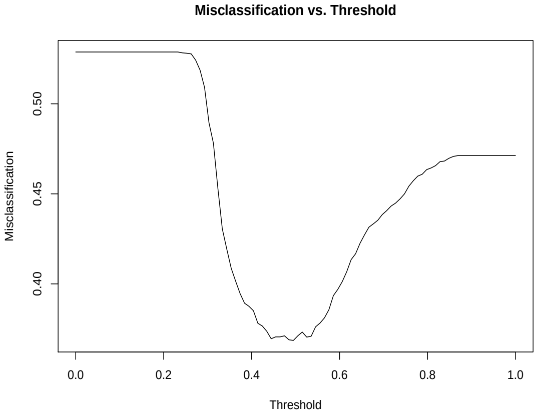

Figure 6: Misclassification Rate for Various Threshold Values

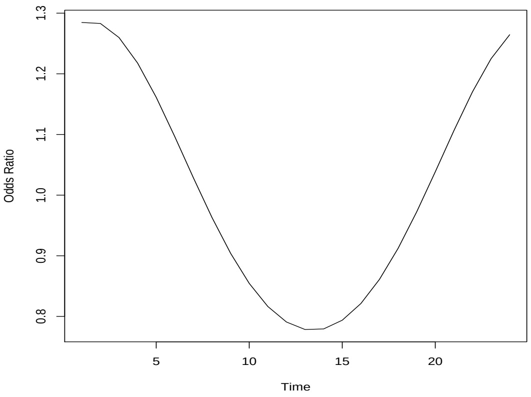

Figure 7: The effect of time of day on the odds ratio of a severe crash

One of the primary goals of this analysis is to predict severity of vehicle-related accidents. To assess how well the model predicts, we performed cross validation. Cross validation was performed by randomly splitting the entire dataset into 5 approximately equal segments to use as test data. The remaining data was used as training data. We then fit our model using the training data and used that model to predict all the observation in our test data. This process was repeated so that each of the 5 segments was used as test data so that we could obtain out-of-sample predictions for every observation in the dataset. To measure how well the model predicts, we will report out-of-sample AUC and out-of-sample sensitivity, specificity, positive predictive value, and negative predictive value. The out-of-sample ROC curve can been seen in figure 5. The out-of-sample AUC is 0.67, which is the same as the in-sample AUC. We expect the out-of-sample AUC to be slightly smaller, so we are pleased with this result. The out-of-sample sensitivity, specificity, positive predictive value, and negative predictive value are reported in table 2. These values are very similar to the in-sample results. Thus we can concluded that our model accurately predicts severity of vehicle-related accidents.

Table 2: Classification Performance

	In-Sample	Out-of-Sample
Sensitivity	0.577	0.569
Specificity	0.679	0.681
Positive Predictive Value	0.616	0.614
Negative Predictive Value	0.643	0.639
AUC	0.668	0.665

5 Results

Table 3 shows our parameter estimates for each term in our model. All variables were significant at a .05 level. Figure 7 shows the complex relationship between time and risk of a severe crash, demonstrating that crashes at night are more risky than daytime crashes. A crash that didn't involve alcohol is only 52% as likely to be severe. A crash where the air bags deployed is 2.4 times as likely to be severe as one that didn't, and a crash where air bags weren't applicable was 1.3 times more likely to be severe. Speed has a positive correlation with severe crashes. For example, a crash that happened on a road with a 70 mph speed limit is 1.2 times more likely to be severe than a crash on a 50 mph road. Finally a crash on a road that is straight is only 0.69 times as likely to be severe as a crash on a curve.

To put it all together, a crash on a bend in the highway, at night, involving air bags and alcohol would result in the absolute worse case scenario for crash severity. Indeed, our highest predicted crash had all five of these traits, so it received a predicted probability of 86% of being severe, and it was severe.

Our confidence intervals can be understood in the as the change in the log odds when the variable increases by unit, we can exponentiate these end points to get a reasonable range for the change in odds.

	$\hat{\beta}$	SE	95% CI
coshour	0.23	0.04	(0.161, 0.306)
sinhour	0.09	0.03	(0.027, 0.154)
alcohol-No	-0.64	0.08	(-0.783, -0.487)
air_bag-NA	0.27	0.09	(0.100, 0.434)
air_bag-Deployed	0.87	0.05	(0.778, 0.966)
Speed Limit	0.01	0.00	(0.008, 0.015)
Straight Road	-0.37	0.07	(-0.499, -0.233)

Table 3: Estimates and 95% confidence intervals based on the asymptotic distribution of our parameters

6 Conclusion

Our analysis was able to provide a simple description of the most important indicators of whether a crash would be severe or not. While the effects we describe have a very significant impact on the probability of being severe, our model does not provide a perfect description of crashes, and we are still only able to classify crashes as severe or not severe with about 63% accuracy. A lot of what happens in a crash is still left up to chance.

A criticism of our model could be that it ignores relevant variables for simplicity sake. While we appreciate this model as a simple description of the phenomena, we recognized that future researchers may be interested in answering specific questions from the data, such as how seat belts effect the severity of a crash. Our model would need to be adapted to provide a suitable analysis for that question. Additionally, other researchers may challenge our grouping, and may find better models by exploring different ways to group the variables.