Car Crash Analysis: Understanding Factors that Contribute to Severe Accidents

Matthew Oehler

March 30, 2018

Abstract

In 2015 alone, there were more than 34,000 deaths attributed to motor vehicle accidents. The federal highway administration (FHWA) is responsible for supporting state and local agencies to help increase highway and roadway safety. Using data from the General Estimates System (GES) database we use logistic regression to determine which independent variables have a relationship with the severity of a car accident. We fit a logistic regression model, verify that the necessary assumptions hold, interpret results, and assess predictive accuracy of the model. The conclusions reached can be helpful in creating and implementing policies that will improve road safety and ultimately save lives.

1 Introduction

In 2015 alone, there were more than 34,000 deaths attributed to motor vehicle accidents. The federal highway administration (FHWA) is responsible for supporting state and local agencies to help increase highway and roadway safety. One tool that the FHWA uses to collect information on motor vehicle accidents is the General Estimates System (GES)[1]. The GES database contains information about several factors that could potentially relate to road safety for a stratified sample of all vehicle crashes in the United States. Among the factors recorded in the GES is whether a particular accident is severe (the crash resulted in at least one fatality or serious injury). Leveraging the GES data, the FHWA, Congress, and the National Highway Traffic Safety Administration (NHTSA) can determine if there is a relationship between road condition/accident related factors and the severity of an accident. In this analysis we will assess which of the recorded factors in the GES dataset have a significant relationship with the severity of a car accident.

The primary purpose for this analysis is to use the discovered relationships to help institutions such as the FHWA, Congress, and the NHTSA be able to create and implement policies that will improve road safety and ultimately save lives. We will use logistic regression to perform this analysis since it will help us to estimate quantities that represent relationships between certain factors and car accident severity, and it will allow us to quantify the uncertainty associated with those elements. Logistic regression will also allow us to predict whether a car crash is severe or not given certain conditions, but the main focus of this analysis will simply be inference.

2 Data

As was previously mentioned, the data used for this analysis come from the GES data base. In particular we will be using variables to describe the light condition, road median, speed limit, road alignment, road, surface, type of intersection, number of lanes, airbag usage, seat belt usage, weather condition, time of day, alcohol involvement, and whether or not the car crash is severe. These variables can be seen in table 1. The details of how these variables are recorded can be found in the documentation provided by the US Department of Transportation [1].

Since some of the conditions had few occurrences, we manipulated the data by combining certain categories of variables to reduce the risk of model over fitting. To justify the regroupings for several of the variables, we looked at cross tabulations for each of the variables and the severity of the car crashes. The variables prefixed with 'my_' are the variables that were modified. The airbag variable was simplified to just be an indicator of whether or not any airbags were deployed in the accident. The surface condition variable was regrouped into dry, wet, ice, snow, or other to account

for the sparsely observed categories. The restraint variable was simplified to indicate whether or not a seat belt or other type of restraint was used (also to account for sparse observations). The intersection variables was simplified to indicate whether or not the accident occurred in an intersection. The number of lanes variable had a range of one to seven, but since there were very few occurrences of seven lanes, we grouped them in with the six lane observations, and left this variable as a quantitative variable since it seemed most intuitive. The weather variable was reduced to conditions of clear, rain, hail, snow, fog, windy, and cloudy. We grouped sparsely observed categories in with most similar condition. Time was simplified to day time (7am to 9pm) and night time (9pm to 7am). Alcohol was already measured as an indicator of whether or not alcohol was involed we just changed it to be more easily interpretable (1: alcohol was involved, 0: no alcohol was involved). As a side note, the speed limit variables only has observations at five mile per hour increments, but we left it as a quantitative variable to maintain its inherent ordinal structure.

Variable	Description
LGT_COND	Different light conditions (categorical)
VTRAFWAY	How the road was divided (categorical)
VSPD_LIM	Speed limit (quantitative)
VALIGN	Road alignment (categorical)
my_airbag	Whether or not the airbag was deployed (categorical)
my_surface	Surface condition of the road (categorical)
my_restraint	Were seat belts used (categorical)
my_intersection	Whether or not the crash occurred at an intersection (categorical)
my_numlane	Number of lanes (quantitative)
my_weather	Weather conditions (categorical)
my_time	Day or night (categorical)
my_alcohol	Whether or not alcohol was involved (categorical)
SEVERITY	Whether or not the crash was severe (categorical)

Table 1: Description of variables

3 Methods

Since our response variable (the severity of a car accident) is binary, we can't really use standard linear regression, which would results in response estimates that are not equal to zero or one (let alone the fact that the relationship isn't linear, and distributional assumptions don't hold.). Instead we can use logistic regression, which is a type of generalized linear model that constrains the response to be between zero and one. The setup for a logistic regression model is shown in equations 1, 2, and 3.

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

$$log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta} \qquad \text{(logit transform of odds ratio)} \tag{2}$$

$$p_i = \frac{exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1 + exp\{\mathbf{x}_i'\boldsymbol{\beta}\}} \tag{3}$$

$Y_i = i^{th}$ response $\quad (scalar)$

$\mathbf{x}_i =$ vector of covariates for the i^{th} observation $\quad (1 \times p)$

$\boldsymbol{\beta} =$ model coefficients $\quad (p \times 1)$

The response variables follows a Bernoulli distribution which, as desired, constrains the response to be binary. The p_i parameter associated with the Bernoulli distribution is constrained to be

between zero and one, but by doing a logit transform on it (see equation 2) the result is an unconstrained value which we can model with $\mathbf{x}_i'\boldsymbol{\beta}$ similar to ordinary linear regression. Back solving for p_i leaves us with the probability that an accident is severe vs. not severe.

When performing logistic regression there are a few assumptions that need to hold. First, we need to assume that the response variables are independent of each other. Second, we need to ensure that relationships are linear in terms of log-odds. This can be done using be checking to see if smoothed regression lines are monotonic in relation to the binary response variable. These assumptions will be verified in the following section.

4 Results

Using the manipulated data as described in above, we fit a logistic regression model and performed variable selection by comparing models with all different possible combinations of the variables. We picked the model using BIC, which yields a more parsimonious model. This is because we are more interested in inference than prediction (for which AIC would probably be better). Table 2 shows the coefficient estimates for the variables that were chosen for the best model. We then verified our model assumptions to ensure that we can proceed with our results and draw conclusions.

The only quantitative variable that we used in our model is speed limit, and as shown in figure 1 we can see that the smoothed regression line is indeed monotone. Also, since the severity of one car crash doesn't really provide information about the severity of another car crash, we conclude that the assumption of independence holds as well.

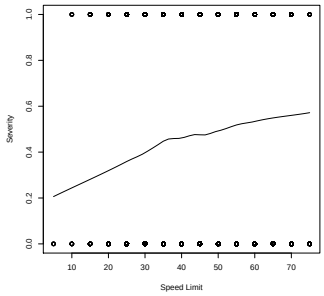

Figure 1: Verifying assumptions

The coefficient estimates are in terms of the log-odds ratio. Since that's not a very interpretable scale, we transformed the estimates, and included them with there transformed 95% confidence intervals in table 3. These can be interpreted in terms of the likeliness of an event. For example, we estimate that holding all else constant, an accident in which seat belts weren't used (my_restraint = 1) is 3.83 times more likely to be severe on average. We are 95% confident that the true value for $exp\{\beta_{my\_restraint}\}$ (the effect of not using a seat belt) is between 3.16 and 4.67. The other variables would be interpreted similarly. The results of the model seem to follow to follow for general intuition as well. The coefficients all have a positive relationship with severity. It makes sense that factors such as alcohol, night time, and not using a seat belt would increase the likelihood of a severe car accident. The airbag results can be misleading since airbags are a safety mechanism. However, since air bags don't deploy for minor crashes, it still makes sense that a deployed airbag is indicative of crash severity.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1687	0.0845	-13.84	0.00
VSPD_LIM	0.0109	0.0018	5.98	0.00
my_airbag1	0.7262	0.0458	15.85	0.00
my_restraint1	1.3436	0.0996	13.50	0.00
my_timeNight	0.3792	0.0610	6.22	0.00
my_alcohol1	0.5402	0.0760	7.11	0.00

Table 2: Table of model coefficients

	2.5 %	$exp\{\text{Estimate}\}$	97.5 %
(Intercept)	0.26	0.31	0.37
VSPD_LIM	1.01	1.01	1.01
my_airbag1	1.89	2.07	2.26
my_restraint1	3.16	3.83	4.67
my_timeNight	1.30	1.46	1.65
my_alcohol	1.48	1.72	1.99

Table 3: 95% confidence intervals for $exp\{\beta\}$

To assess the performance of the model we performed first had to select a threshold value to determine at what point the model classifies a crash as severe or not severe. To do this we iterated over many possible threshold values and picked the value that minimized the number of misclassified crashes. This can be seen in figure 2, and the resulting threshold value was 0.495. We then performed cross validation for 1000 iterations and calculated the mean accuracy, specificity, sensitivity, positive predicted value, and negative predicted value. The results of the cross validation are included in table 4. These statistics help us to see how well the model performs with out-of-sample data. We also created a receiver operating characteristic curve to assess out-of-sample performance of the model, which is displayed in figure 3. The area under the ROC curve (AUC) can be used to compare models to each other. The closer the value is to one, the better the value is. We calculated an AUC of 0.69, and had an overall prediction accuracy of 0.63. The model isn't amazing at prediction, but it does perform significantly better than random chance.

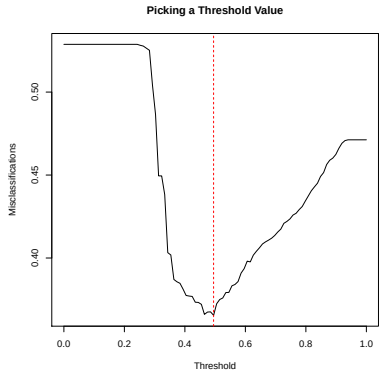

Figure 2: Selecting a threshold value

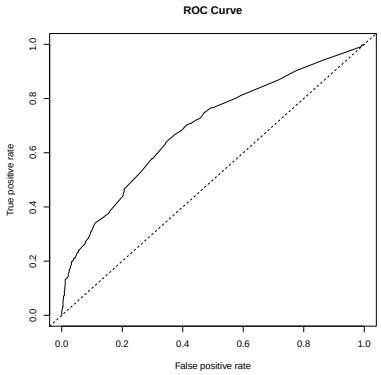

Figure 3: ROC Curve

	Model Performance
Sensitivity	0.57
Specificity	0.70
PPV	0.62
NPV	0.65
Accuracy	0.63
AUC	0.69

Table 4: Summary statistics for model performance

5 Conclusion

After performing this analysis we conclude that we were able to accomplish the goal of our analysis which was to determine which factors have a relationship with the severity of a car crash. The results calculated above should be able to help institutions such as the FHWA, Congress, and the NHTSA be able to create and implement policies that will improve road safety and ultimately

save lives. This study could be improved upon by looking at potential interactions between factors of accidents. This study could also potentially be expanded to look at roadways which have a particularly high proportion of severe crashes to see in further depth, what measures can be taken to reduce car crashes by as much as possible.

References

[1] US Department of Transportation. National automotive sampling system (nass) general estimates system (ges) analytical user's manual 1988-2015, August 2016.