Car Crash Case Analysis

Easton Huch and Jacob Merrell

STAT 536

No Institute Given

Abstract. Thousands of people die each year in car crashes. This study explores the relationships of various explanatory variables with the severity of injury in the car crashes. We show which factors increase and decrease the odds of suffering a severe injury. Policy makers are encouraged to help enact laws to help make roads safe.S

1 Introduction

In 2015, more than 34,000 people died in motor vehicle. The federal highway administration (FHWA) is charged with keeping highways as safe as possible. The FHWA created General Estimates System (GES); a system that collects data on freeway crashes. The goal of this study is to use GES data to see the relationship between independent variables like speed limits, driving habits, seat belt types, etc. and the probability of a serious injury. Understanding these relationships will, ideally, lead to policies that will save lives and create safer driving conditions. The Crash dataset comes from the GES database. The SEVERITY variable is an indicator if at least one person in the accident sustained a serious (including fatal) injury.

2 Exploratory Data Analysis

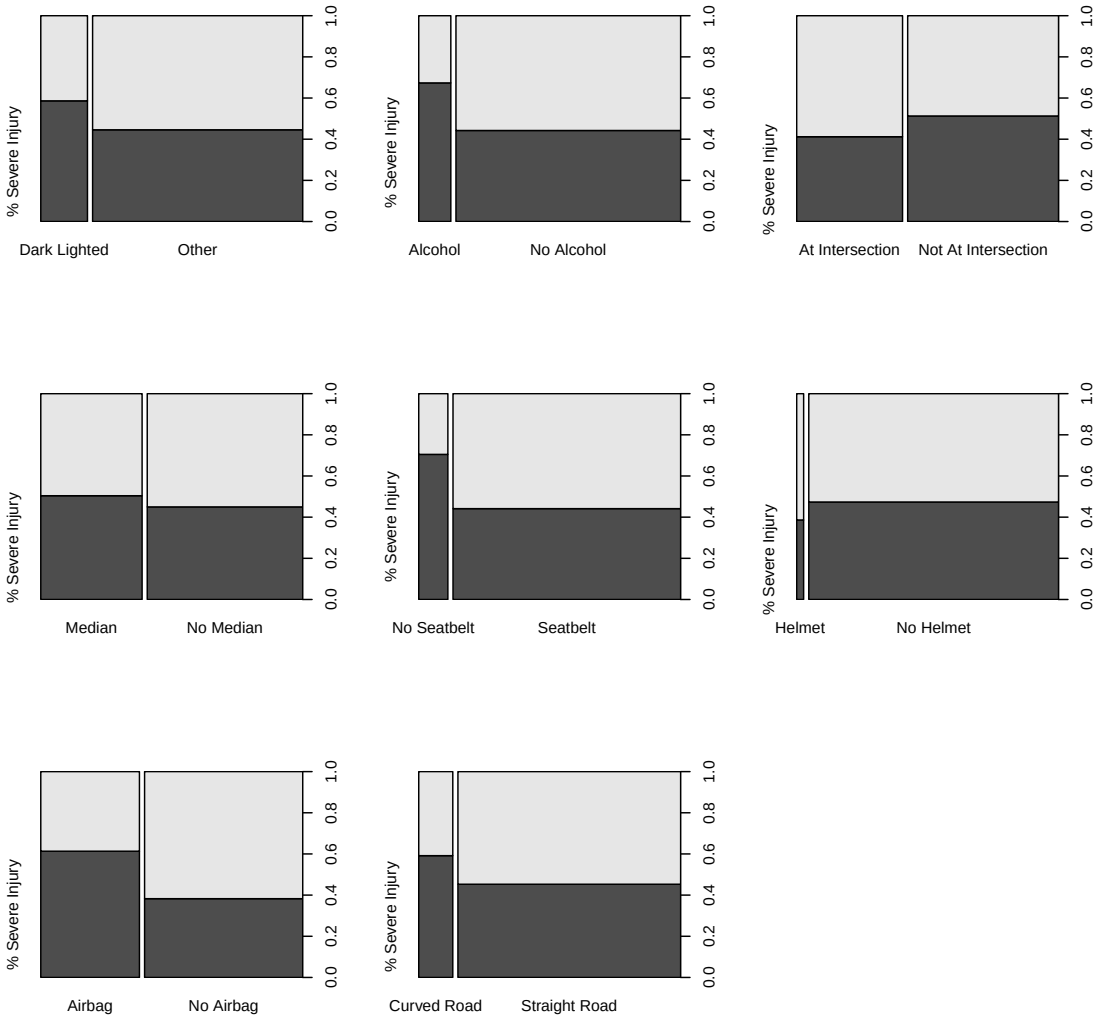
The data included many potential explanatory variables, most of which contained multiple categories. The large number of categories posed three problems:

1. Many categories represented a small percentage of the data (often less than 1%), creating the potential for high standard errors and serious overfitting
2. Many categories represented similar driving conditions (e.g., sleet vs. snow), unnecessarily increasing model complexity
3. Many categories, especially those suffering from the first two problems, had similar probabilities of being a severe crash, suggesting they are not strong predictors of collision severity

Because of these issues, our first task was to combine some categories to alleviate these problems. For example, we had separate categories for a variety of intersection types. However, we noticed that many intersection types (e.g., T-intersections, L-intersections, traffic circles, and roundabouts) were uncommon and had similar probabilities of collision severity, so we combined all intersection types into a new variable "At Intersection" that was 1 when the collision occurred at an intersection and 0 otherwise. After combining categories in this fashion, we were left with 25 explanatory variables, of which 23 were categorical. The following variables are a subset of the 25 explanatory variables:

- Hour: the hour of the day, ranging from 0 to 23
- Dark Lighted: whether the collision occurred after sunset in an artificially lighted area
- Alcohol: whether alcohol was involved in the collision
- At Intersection: whether the collision occurred at an intersection
- Median: whether the road had a median
- Seatbelt: whether the occupants were wearing any type of seatbelt
- Helmet: whether the occupants were wearing a helmet
- Airbag: whether any type of airbag deployed during the collision
- Curved Road: whether the road was curved

We will later show the subset selection methodology used to choose the variables above.



As expected, drivers who drank alcohol, those that didn't use restraints, and crashes when air bag(s) displayed higher rates of severe injury. Other significant explanatory variables are described further in the study.

3 Model Specification

We will use the logistic regression model to analyze the car crash data. It can be written as follows:

$$Y_i \overset{ind}{\sim} \text{Bernoulli}(p_i), \;\; \log\Big(\frac{p_i}{1-p_i}\Big) = \mathbf{x}_i'\boldsymbol{\beta} \tag{1}$$

In Equation (1), the symbols have the following meaning:

- Y_i: The response variable for crash i; it is equal to 1 if the crash was severe and 0 otherwise
- p_i: The probability that crash i is severe
- $\mathbf{x}_i$: A p-dimensional vector of explanatory variables for crash i
- $\boldsymbol{\beta}$: A p-dimensional vector containing the coefficients for each of the covariates; each coefficient represents the average effect of a one-unit increase in the covariate on $\log\left(\frac{p_i}{1-p_i}\right)$, the log odds ratio

The logistic regression model involves two key assumptions. First, the outcome variable is **independent** conditional on the covariates. This means knowing whether a given crash was severe does not impact the probability that any other crash was severe; this probability only depends on the covariates in the manner specified in (1). Second, the probability (p_i) is a **monotonic** function of the covariates. In other words, as a given covariate increases, p_i either increases, decreases, or stays the same; it cannot increase and then decrease (or vice versa).

The logistic regression model will allow us to accomplish our goals because it allows us to perform inference on the vector $\boldsymbol{\beta}$. These coefficients will allow us to determine the partial effect of the covariates on the probability of a car crash being severe. In a general linear model, these coefficients have a simple interpretation. Each coefficient represents the average effect on the outcome variable of increasing its covariate by unit, holding the other covariates constant. However, the interpretation is more nuanced in the case of logistic regression because we are modeling the log odds ratio, not the outcome variable itself. We will demonstrate how to interpret the coefficients in in our Results section.

In addition to a point estimate, our model also allows us to calculate the asymptotic distribution of $\boldsymbol{\beta}$ because we are using maximum likelihood estimation. The asymptotic distribution will allow us to create a confidence interval for each coefficient, quantifying the uncertainty surrounding each one.

The binary nature of the outcome variable means that the general linear model may not be the best choice for this data. Our data is not conditionally normally distributed, so maxmimum likelihood estimation would not be possible. We could still use the least squares solution, but then we could no longer use traditional inference procedures and our model could produce impermissiable probabilities. Model comparison would also be difficult because we would not be able to calculate AIC or BIC without a likelihood function. We would also have to account for the heteroscedasticity of our data using robust standard errors.

The logistic regression model, on the other hand, overcomes all of these potential issues. It assumes that the data are distributed as a Bernoulli random variable, which fits perfectly with the binary nature of our outcome variable. This means that it naturally accounts for the heteroscedasticity of our data and allows us to use traditional methods based on the likelihood function, including model comparison and asymptotic inference. These features of the logistic regression model make it a good choice for our data.

4 Model Justification

In this section, we explain how we performed variable subset selection and then justify the assumptions for our chosen model.

4.1 Variable Subset Selection

We show the two quantitative variables—hour of the day and speed limit—in Figure 1. The hour of the day (shown in plot (a)) appears to exhibit a cubic relationship with p_i, so we included it in our model as an orthogonal cubic polynomial. The speed limit (in plot(b)), on the other hand exhibits a fairly linear relationship, so we incorporated it into our model as a linear term. The points at 5 and 75 miles per hour fall far below the line, but each accounts for a tiny proportion of the data (less than 0.5% combined).

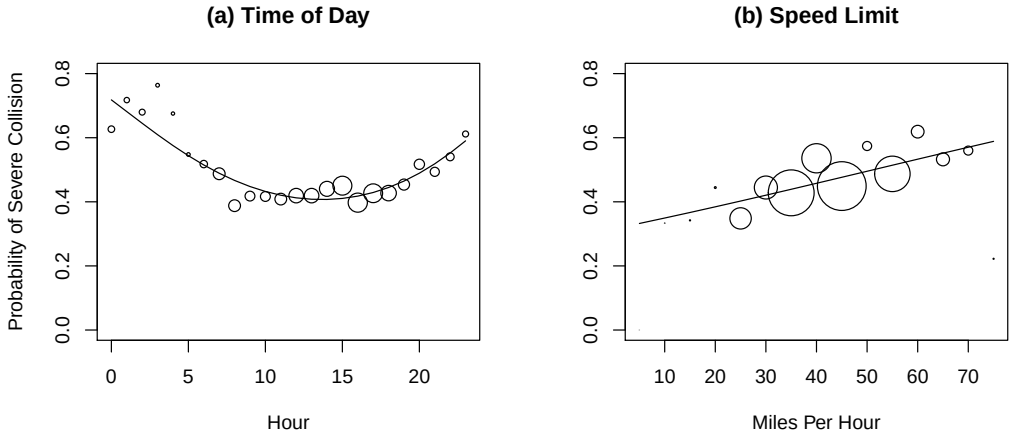


Fig. 1: Plot (a) shows that a cubic fit is reasonable for time of day while plot (b) shows that a linear fit works well for speed limit. The size of the points is proportional to their frequency in our data.

After assessing the quantitative variables in this manner, we began performing variable subset selection with backward BIC selection. We chose BIC as our selection criterion because we are primarily interested in inference. With all of the explanatory variables in the model, we saw that the model BIC would improve significantly if we removed any of the variables indicating the number of lanes. None of these indicator variables were statistically or practially significant, and the data did not show any clear trend as the number of lanes increased. So we removed all of these indicator variables and continued our selection process. We specified that the hour of the day should remain in the model because Figure 1 showed a clear pattern and we did not want to impair our model's interpretability by removing only some of the polynomial terms.

4.2 Assumption Verification

The first assumption is that of independence. We do not have any diagnostics or plots to check this assumption, but we would not expect the severity of different car crashes to be related to one another. So this assumption seems reasonable.

The second assumption is that the probability of a car crash being severe is a monotone function of the covariates. We know that the probability will show a monotone relationship with each of the categorical variables, so the only variable we have to check is "Hour" because it is quantitative. In Figure 2, we display added-variable plots for each of the "Hour" polynomial terms included in our model. In each plot, the relationship between the residuals is monotone, so the monotonicity assumption appears reasonable.

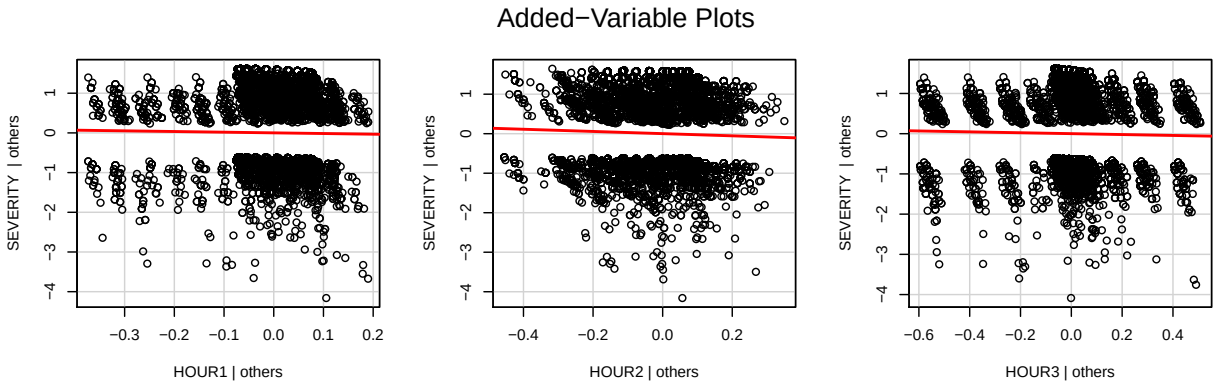


Fig. 2: The relationship between the plotted residuals is monotone for each term.

5 Performance Evaluation

The Reciever Operator Characteristic (ROC) curves for our model are plotted in Figure 3. Plot (a) shows the ROC curve for in-sample predictions while plot (b) shows the ROC curve for out-of-sample predictions. The curves are well above the coin flip rate (the dotted line), indicating that our model fits well and predicts well out-of-sample. They are also nearly identical to each other, suggesting that our model generalizes well (i.e., it is not overfitting). The Area Under the Curve (AUC) for in-sample predictions is 0.677, and the corresponding value for out-of-sample predictions is 0.676.

In Tables 1 and 2, we plot the confusion matrices from the in-sample and out-of-sample predictions respectively. For both tables, we chose the threshold value that minimized the misclassification rate. For Table 1, this value was 0.496, while for Table 2, it was 0.484. The number of crashes in each category is almost identical between the two tables, leading to similar classification metrics. This fact serves as more

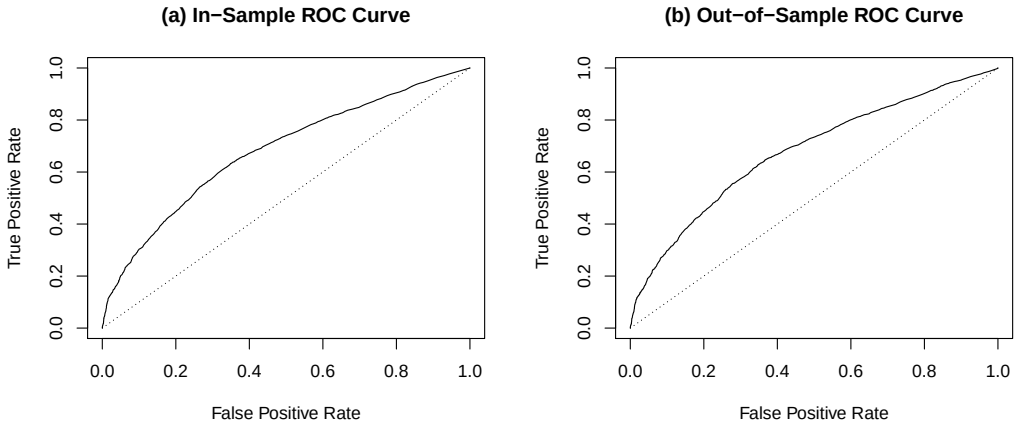


Fig. 3: This figure shows both the in-sample and out-of-sample ROC curves.

evidence that our model generalizes well. Based on the metrics in these tables, we know the follwing about our model:

- About 64% of its predictions (either positive or negative) are correct
- It correctly identifies about 55% of severe accidents
- It correctly identifies about 73% of non-severe accidents

	Predicted Severe	**Predicted Not Severe**	
Severe	2,196	1,858	**Sensitivity:** 54.2%
Not Severe	1,201	3,348	**Specificity:** 73.6%
	PPV: 64.6%	**NPV:** 64.3%	**Misclassification:** 35.6%

Table 1: This table shows the confusion matrix and various **in-sample** prediction measures. PPV stands for Positive Predicted Value and NPV stands for Negative Predictive Value.

	Predicted Severe	**Predicted Not Severe**	
Severe	2,250	1,804	**Sensitivity:** 55.5%
Not Severe	1,263	3,286	**Specificity:** 72.2%
	PPV: 64.0%	**NPV:** 64.6%	**Misclassification:** 35.7%

Table 2: This table shows the confusion matrix and various **out-of-sample** prediction measures.

6 Results

	Coefficients	Lower Bound	Upper Bound
(Intercept)	0.96	0.65	1.27
Hour 1	-0.34	-0.90	0.22
Hour 2	-0.59	-0.97	-0.21
Hour 3	-0.22	-0.53	0.08
Dark Lighted	0.24	0.10	0.39
Alcohol	0.46	0.30	0.61
At Intersection	-0.19	-0.29	-0.10
Median	0.22	0.13	0.32
Seatbelt	-1.31	-1.50	-1.11
Helmet	-1.29	-1.62	-0.96
Airbag	0.81	0.72	0.91
curved Road	0.29	0.15	0.43

Table 3: This table shows the model coefficients for the intercept and the upper and lower bounds for their 95% confidence intervals

As shown before, the model fits the data well, and predicts well out of sample. Also, the assumptions for this study hold. Therefore, we can use the model coefficients and their uncertainties to understand the relationship of the explanatory variables and the probability of suffering a severe injury in a crash. For example, if someone drives drunk instead of driving sober, the log odds of suffering a severe injury increase by 0.46 on average. The 95% confidence interval for the effect of alcohol consumption on injury severity is (0.30,0.61). This means that if the study were repeated several times using different crash data, we would expect 95% of the confidence intervals to contain the true coefficient for the effect of alcohol on the log odds of severe injury. The lower and upper bounds for all coeficients listed in Table 3 are the 95% confidence intervals for each coefficient, and are interpreted in the same way as the example given for alcohol consumption. Note that two of the intervals (Hour 1 and Hour 3) contain 0. If 0 is included in the interval that means the effect of that variable is statistically insignificant at 95% confidence level. In general, if the coefficient is positive that means being part of that category increases the chances of severe injury on average. Therefore we can see air bags, alcohol, curved roads, medians, and lighted roads at night all increase the risk of severe injury on average. Negative coefficients show relationships that are expected to decrease the risk of severe injury. These include wearing a seatbelt, wearing a helmet, and being in an intersection.

7 Conclusion

The model does a good job in identifying the explanatory variables that have the largest effect on the the severity of injuries during car crashes. Air bags, alcohol, curved roads, medians, and lighted roads at night

all increase the risk of severe injury, while wearing a seatbelt, wearing a helmet, and being in an intersection decrease the risk of injury. Even though the model does well identifying relationships between the variables and the severity of injury, perhaps interactions may be explored in future research. For example, perhaps consuming alcohol could have a different effect on severity of injury at different times of the day. Also, only crash data from 2013 was used in the study. Using more years of data would allow us to explore more relationships while still not overfitting. The next step for this research is to inform policy makers on what affects the severity of injuries in crashes to be able to enact policies to help create safer roads.

References

1. R: A Language and Environment for Statistical Computing.R Core Team. R Foundation for Statistical Computing.Vienna, Austria. (2017).url = https://www.R-project.org/