Car Crash Analysis

Jeremy Meyer and Brittany Russell

April 2019

Outline

Motivation


[1]

- 37,461 US fatalities in 2016 (NHTSA department)[2]
- Understanding the relationships between road conditions and fatal injuries → safer roads

The Data

- FHWA collected data from 8603 accidents in 2013 nationwide
- Variable of interest: Crash severity (binary outcome)
  - 47% of data had severe crashes
- Variables collected describe various road/vehicle conditions.

Categorical Variables:

Hour*	Traffic-way
Lighting	Air Bag
Weather	Restraints
ALCOHOL	Road alignment
Intersection	Road surface
Severity	

Numeric Variables:

Number of Lanes	Speed Limit

The Data: Categorical Variables

- Most categorical variables contain several levels
- Problem: Factor levels need to be cleaned
  - Sparsity
  - Similar categories
  - Hour is a cyclic variable $\rightarrow$ categorize

Cleaning Example – Airbag Deployment

(a) Original contingency table.

	Not severe	Severe
None deployed	362	299
Front deployed	787	1259
Side deployed	61	53
Roof deployed	11	20
Other deployed	1	2
Combo deployed	128	248
Unknown deployed	290	445
Not applicable	2909	1728

(b) Original proportions by airbag.

	Not severe	Severe
None deployed	0.5477	0.4523
Front deployed	0.3847	0.6153
Side deployed	0.5351	0.4649
Roof deployed	0.3548	0.6452
Other deployed	0.3333	0.6667
Combo deployed	0.3404	0.6596
Unknown deployed	0.3946	0.6054
Not applicable	0.6273	0.3727

(a) Cleaned contingency table.

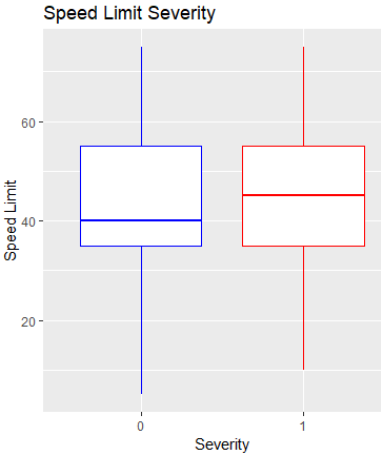
	Not severe	Severe
None deployed	362	299
1+ deployed	1278	2027
Not applicable	2909	1728

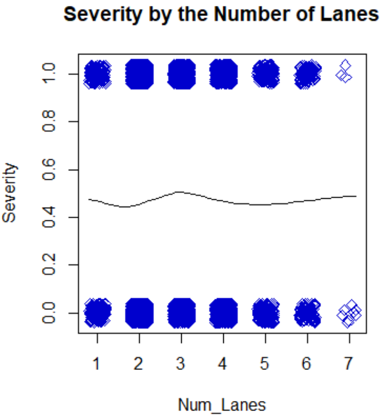
(b) Cleaned proportions by airbag.

	Not severe	Severe
None deployed	0.5477	0.4523
1+ deployed	0.3867	0.6133
Not applicable	0.6273	0.3727

The Data: Continuous variables

- Data also had numeric variables such as speed limit and number of lanes.





Goals of the Analysis

- Our goal is to understand the relationship between these variables and severe crashes.
- We will address the following:
  1. What makes a severe accident more likely?
  2. What is the probability of a severe crash for different groups?

Logistic Regression

- Response variables is binary – severe crash (1) or not severe crash (0)
- Cannot use linear regression (response does not follow a normal distribution)
- Use the Bernoulli distribution to model binary response
- Make the probability of “success” (severe crash) a function of the covariates

Model Statement

Model

$$Y_i \sim \text{Bernoulli}(p_i), \qquad \boldsymbol{x_i'}\beta = \log\left(\frac{p_i}{1-p_i}\right)$$

Y_i: the response for the ith crash
p_i: the probability of the ith crash being "severe"
$\boldsymbol{x_i'}$: the vector of covariates for the ith crash
β: coefficients of the covariates, the effect of the covariate on the log-odds

Model Continued

$$\boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1\text{airbag}_i + \beta_2(\text{no restraint})_i + \beta_3\text{unknown}_i + \beta_4\text{alcohol}_i + \beta_5\text{speed}_i + \beta_6\text{night}_i + \beta_7\text{left}_i$$

Covariates

airbag: binary, no airbags deployed or at least one deployed
no restraint: binary, known restraint used or no restraint used
unknown: binary, known restraint used or unknown restraint used
alcohol: binary, no alcohol involved or alcohol involved
speed: quantitative, speed limit, mph
night: binary, any other light condition or night with lights
left: binary, straight/other curvature or left curve

Model Assumptions

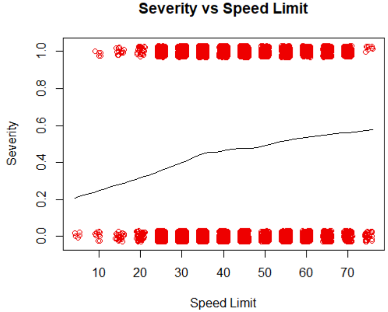
1. Independence
2. Monotonicity in the predictors
3. No Multicollinearity

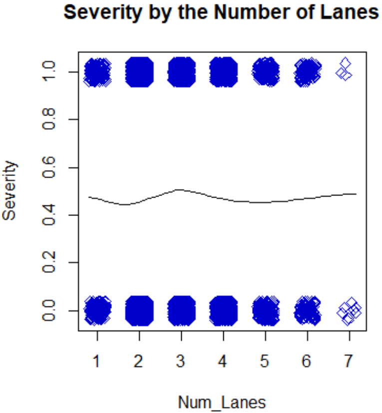
Independence

- Each individual car crash is independent of the others
- Reasonable since data has been pre-cleaned
  - Only 1 record per car accident

Monotonicity in the predictors

Only needs to be checked for continuous variables





Multicollinearity

- Potential Problems with:
  1. Hour (VIF 3.00) and Lighting (VIF 3.16)
  2. Weather (VIF 5.87) and surface conditions (VIF 5.56)

  (All other variables had a VIF less than 2)

Solution: We will allow step-wise regression to choose between the correlated predictors. We will then check collinearity with final model to verify it has dissipated.

Variable selection

- Problem: Even after data cleaning $\rightarrow$ 43 levels across 11 variables.
  - Multicollinearity / Overfitting
- We want the significant levels, not just significant variables.
- Stepwise Regression - treat each level as its own variable

Stepwise regression criterion

Criterion : BIC

- Lower standard errors
- Less chance for overfitting
- Better for our inference-type research questions

$$BIC = -2(loglikelihood) + Plog(N) \tag{1}$$

P = Number of predictors in the model
N = number of crashes in dataset

Stepwise regression will seek to minimize this quantity

Bidirectional stepwise selection algorithm

We used this to perform variable selection

1. Start with intercept only model
2. Consider all possible models after adding one of the variables not in the model or removing one variable currently in the model.
3. Find and choose the model with the lowest BIC.
4. Repeat 2 and 3 until doing nothing yields the lowest BIC

Which ones were eliminated?

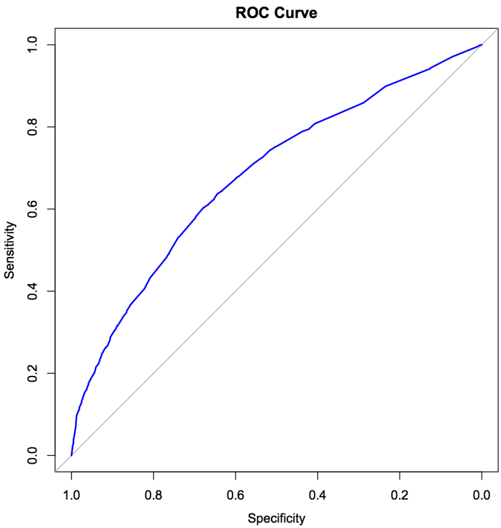
Trafficway (2w-Divided)	Weather (Clear)	Intersec type (None)	Restraint (Lap/Shoulder)
2w-Divided, unprotected	Rain	4 way	Shoulder Only
2w-Unprotected	Snow	5+ way	Lap Only
One way	Crosswinds	Y-int	Not Used
2w-mid lane	Cloudy	T-int	Motor Helmet
On/off ramp	Low Visibility	L-int	Other/Unknown
	Wintry Mix		None available
Surface Condition (None)	**Light (Day)**	**Hour (Day (10a-3p))**	**Alignment (Straight)**
Snow/Slush	Dark-not lit	Morning (6a-9a)	Curve Right
Ice	Dark-lit	Evening (4p-8p)	Curve Left
Water	Dawn	Night (9p-5a)	Unknown Curve
Dry	Dusk		
Wet			
Other Conditions			
Air Bag (Not deployed)	**Alcohol (None)**	**Speed Limit (0)**	**Number Lanes (0)**
Air Bag Deployed	Alcohol Used	Speed Limit	Number of Lanes

- The algorithm eliminated 35 variables and kept only 8
- In the final model, all VIFs were less than 2

Evaluating Fit

- R^2 does not work in this situation
- We need a measurement of classification accuracy
- ROC (Receiver Operating Characteristic) Curve
  - Uses many different cutoff values
  - Plots sensitivity (true positive rate) against specificity (true negative rate)
- AUC (Area Under the Curve) summarizes the ROC curve

ROC Curve


ROC Curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

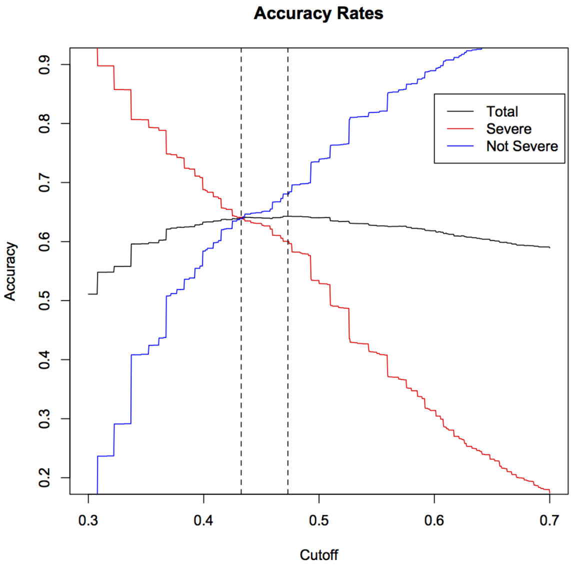
ROC Curve

- AUC $= 0.682$
- AUC $= 1$ (perfect classification), AUC $= 0.5$ (same as flipping a coin)
- Model classifies better than a coin flip
- Missing factors that distinguish severe and not severe crashes?

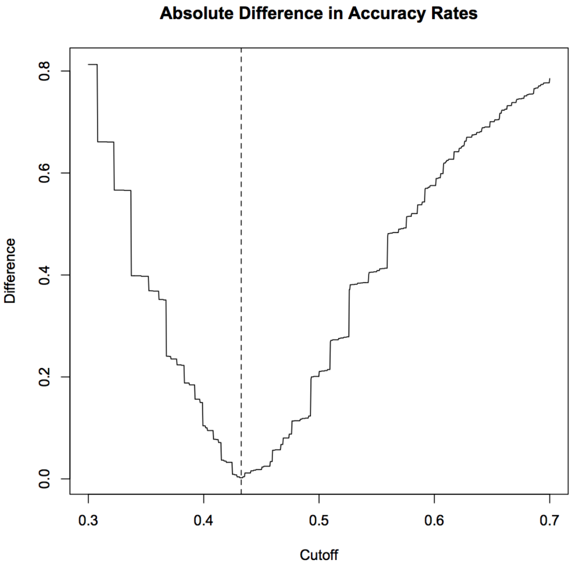
Choosing a Cutoff

- In order to classify, we must choose a cutoff value
- "Best" cutoff depends on the goals of the analysis
- Balance accuracy in predicting severe crashes and predicting not severe crashes
- Maximize overall accuracy (biases towards the more common category)

Accuracy Rates



Balancing Two Types of Accuracy



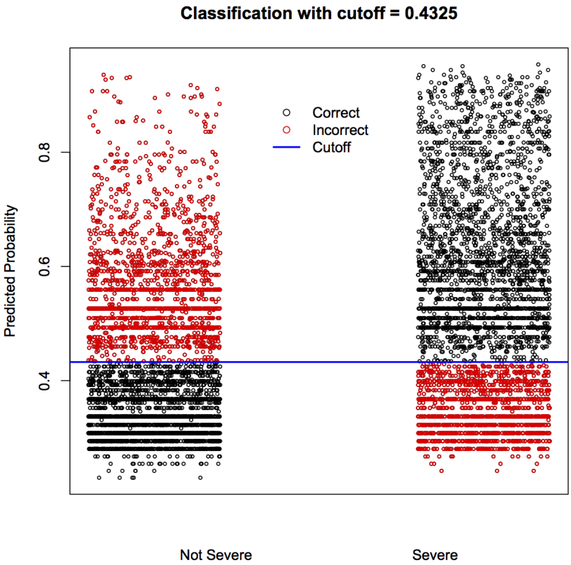
Classification

Table: Confusion Matrix, cutoff = 0.4325

	Classified Not Severe	Classified Severe
True Not Severe	2906	1643
True Severe	1455	2599

- Sensitivity $= 2599/(2599+1455) = 0.6388$
- Specificity $= 2906/(2906+1643) = 0.6411$

Classification



Research questions

Recall our first research question was what makes a severe accident more likely.

To answer this question, we'll look at the model coefficients.

Results

Effect	Estimate	2.5 %	97.5 %
(Intercept)	-1.276	-1.446	-1.107
1+ Air Bag Deployed	0.781	0.688	0.875
No Restraint Used	1.390	1.197	1.588
Unknown Restraint Used	0.456	0.221	0.695
Alcohol	0.544	0.397	0.692
Speed Limit	0.013	0.010	0.017
Night - Lit Roads	0.373	0.252	0.493
Left Curve	0.418	0.226	0.611

Baseline Levels: No airbag deployment, known restraint used, no alcohol involvement, speed limit of 0, daytime conditions, straight road.

Interpretation of effects

Interpretation:

- Categorical effects are relative to baseline level.
- When at least one airbag is deployed, severe crashes are $e^{.781} = 2.18$ times more likely.
- For every one mph increase in speed limit, severe crashes are $e^{.013} = 1.013$ times more likely.
- All confidence intervals do not include zero $\rightarrow$ significance.

What makes a severe accident more likely?

Contributors to severe accidents

Contributor	Multiplier	2.5 %	97.5 %
1+ Air Bag Deployed	2.184	1.990	2.398
No Restraint Used	4.014	3.310	4.896
Unknown Restraint Used	1.579	1.247	2.003
Alcohol	1.723	1.487	1.998
Speed Limit (per 5mph)	1.069	1.050	1.089
Night - Lit Roads	1.451	1.286	1.638
Left Curve	1.519	1.254	1.842

- Not wearing seat belts increases odds of severe car crash the most.
  - No difference between most other restraint types
- Air bag deployment also increases odds of a severe crash.
  - Confounded?
- Left curves are more associated with severe accidents.

What is the probability of a severe crash for different groups?

- We can explore the probability of a severe crash by group
- Compare estimated probabilities across all groups
- 48 possible combinations of categorical variables
- Hold speed limit constant at 45 mph

Probabilities by group

Table: Estimated probabilities of a severe crash for worst, best, alcohol only, no seatbelt only, and night driving only groups (speed limit held constant at 45 mph).

	Best	Worst	Alcohol	No seatbelt	Night driving
estimated probability	0.337	0.944	0.467	0.671	0.425
dark with lights	0	1	0	0	1
alcohol involved	0	1	1	0	0
no restraint used	0	1	0	1	0
unknown restraint used	0	0	0	0	0
1+ airbag deployed	0	1	0	0	0
road curved left	0	1	0	0	0

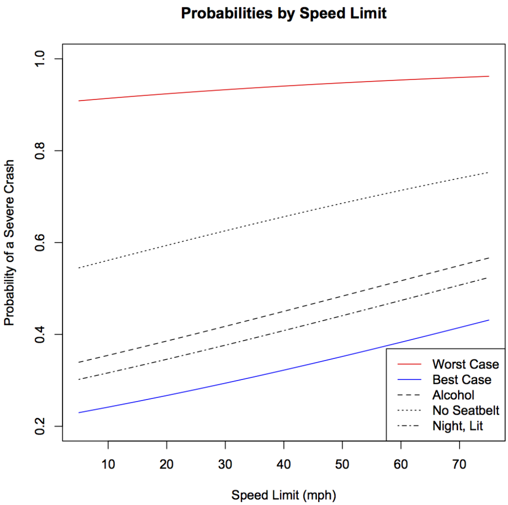
Worst Case Scenario

- Dark road with lights
- Alcohol is involved
- No restraint used
- At least one airbag deploys
- Road curves left
- Probability increases as speed limit increases

Best Case Scenario

- Road is not dark and lit
- Alcohol is not involved
- Either lap, shoulder, lap/shoulder, or motorcycle helmet used (or restraint is not applicable)
- No airbags deploy (or airbag is not applicable)
- Road is straight (or curves not to the left)
- Probability decreases as speed limit decreases

What about speed limit?



Remarks

Logistic regression model succeeded in estimating the relationship between independent variables and the probability of a severe crash for different groups.

Model fit is adequate and the assumptions of the model are satisfied.

Weaknesses

- Variable selection is subjective (bias vs. variance)
- "Unknown" variables may create problems (i.e. unknown restraint used)
- Model seems to be missing important variables

Future Questions

- New data set, build a model for prediction
- Try different cutoffs for classification
- Estimate the real-world costs of the two types of errors

References

1. komonews.com *1 dead, 1 under arrest in 3-vehicle crash in Lynnwood*, Jan 13th, 2019
2. NHTSA public traffic crash data 2016