Lin Zhou
MA 678 Midterm Project
11/10/2020

Severity analysis for US traffic accidents

Abstract:
In this project, I am going to explore the relationships between severity of car accidents (the severity means the traffic severity rather than drivers and passengers' injury) and other possible related variables based on the accidents happened in the United States.

Introduction:
Background: Nowadays, car is the most popular vehicle people using in their daily lives. Even though the generation of car has been updated over time with development of technology, there still exist a large amount of car accidents happened every year.
Dataset: The dataset I found on the internet has been collected in real-time, using multiple Traffic APIs. It is a countrywide car accident dataset, which covers 49 states of the USA through February 2016 to June 2020. Generally, there are a bout 3.5 million accident records in this dataset. In the first place, I would like to clean the data to find suitable part of the dataset which can be used to fit the model. Then I will do some EDA which can give us some general description of the dataset. Moreover, I will try to fit and check the model, make some predictions.

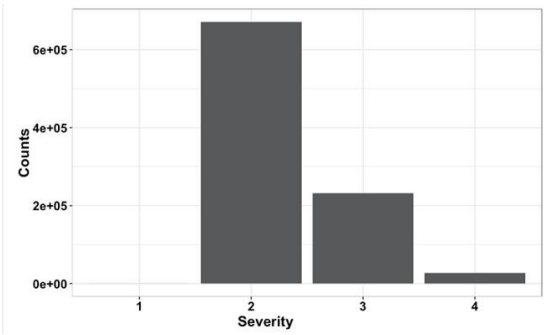
Method:
Data cleaning: The whole dataset contains total 3513617 car accidents. After going through the data, I found that the date jumps in somewhere 2018. Since variables like seasons, weathers could be a factor which might affect the accidents' severities, it is necessary to choose a consecutive year to make sure enough observations for analysis. As a result, in this case, I would like to select observations of the whole year 2019. The start time and end time of traffic severity (delay) are also included in the dataset. Since my object is to find the variables related to the severity, these kinds of information are already converted to the scale of severity. As a result, I would not use these data as predictive variables. Then I am going to check the missing data or NA values. Firstly, I select other possible related variables, then using df_status function to check the NA proportions for these variables in the dataset.

variable <chr>	q_zeros <int>	p_zeros <dbl>	q_na <int>	p_na <dbl>	q_inf <int>	p_inf <dbl>	type <chr>	unique <int>
Humidity...	0	0.00	6577	0.70	0	0	numeric	100
Visibility.mi.	597	0.06	6793	0.72	0	0	numeric	63
Wind_Speed.mph.	133106	14.13	48401	5.14	0	0	numeric	99
Precipitation.in.	666397	70.74	207455	22.02	0	0	numeric	172
Traffic_Signal	0	0.00	0	0.00	0	0	character	2

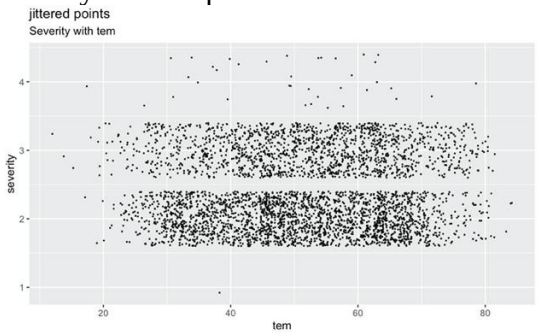
Part of the output shows above. The p_na denotes the percentage of NA values of the variable within the dataset. Finally, I filter the variables which have less than 10% NA values and delete those NA values. Then I use the unique function to check whether all the observations under each variable have correct record. I find that there are some "spaces" under the variable Civil_twilight which is supposed to contain only two records (Day and Night). After I delete all the "space" rows, at this point, I have got a clean dataset of total 830444 observations.

Variable description: The response variable I am going to explore is Severity. It shows the severity of the accident with numbers from 1 to 4, where 1 indicates the least impact on traffic, 4 indicates the most. Some possible related predictive variables are: Civil_Twilight: It shows the period of day (day or night) based on civil twilight. Temperature: It shows the temperature (in Fahrenheit). Humidity: It shows the humidity (in percentage). Traffic_Signal: A POI annotation which indicates presence of traffic signal in a nearby location. Visibility: It shows visibility (in miles). Weather_Condition: It records the weather condition of the accident day.
Simple EDA: After I have the clean dataset, I would like to use ggplot to make the barplot for the severity and other plots for general descriptions.

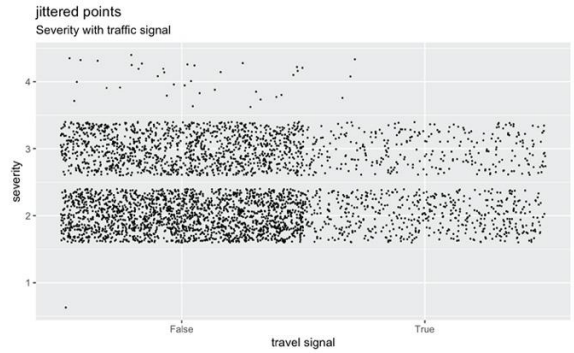
For severity:



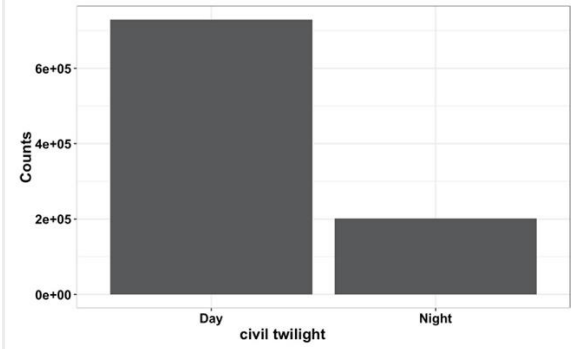
For severity and temperature:



For severity and traffic signal



For severity and civil twilight



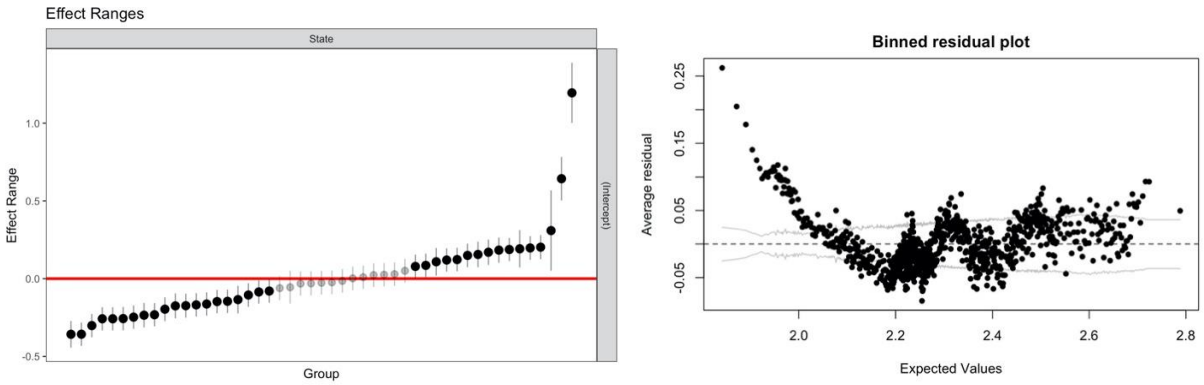
From the graph, severity level 1 and 4 rarely happened in the accidents, level 2 is the most common situation. Since there are about 1 million observations, the graph would not be informative if I put all points on it. As a result, I choose first 4000 observations to make some plots. From the third graph, we can see most of the traffic severity caused by car accidents happened with no traffic signals. From the last graph, the situations usually happened at daytime based on civil twilight. That somehow makes sense since there should be more cars on the roads at daytime.
Model selection: The locations of accidents are recorded in the dataset detailed in street, city, county, state. It seems that HLM (Hierarchical Linear Model) might be a good potential choice for the dataset. As a result, firstly, I would like to try to fit dataset using HLM.
Model fit: To fit HLM model, I would like to use lmer function of lme4 package in R. First and foremost, I will fit the empty model.
The empty model didn't contain any fixed effects beyond intercept. The results from the empty model are used to calculate the variance in the outcome. In this case, it can be used to calculate

the intraclass correlation coefficient (ICC) which is the variance of the outcome that occurs at level2 (states) versus level1(single observations).

```
    Adjusted ICC: 0.219
Conditional ICC: 0.219
```

The ICC result shows above, it tells us that 21.9% of total variability exists at the state level. Then I would like to try random intercepts model. It is shown as two parts of the regression model, one at the observation level and one at state level.

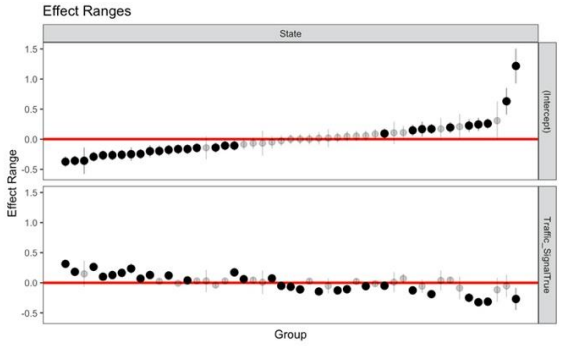
For standard regression model:
Severity = b_intercept + b_visibility * visibility + (b_other_variable * other_variable) + ϵ
where $\epsilon \sim N(0,\sigma)$
For random intercepts mode:
Severity = (b_intercept + effect_state) + b_visibility * visibility + (b_other_variable * other_variable) + ϵ
where $\epsilon \sim N(0,\sigma)$, effect_state $\sim N(0,\sigma)$

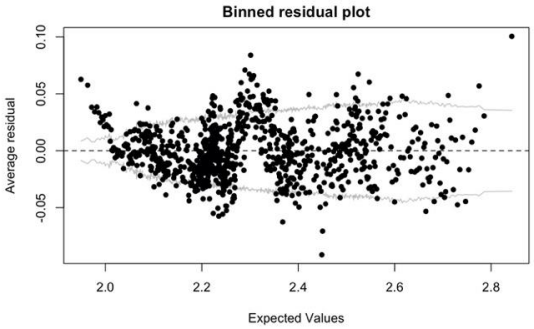


The left graph above is the estimated random effects for each state and their interval estimates. The right graph is the binned residual plot for the model. In this model, the state is the only random variable and the rest of variables are set as fixed effects. From the graph, there are lots of observations fall outside the confidence bands, mostly clustered at expected value of 2 and 2.2. The residual plot does not seem good.
Then I would like to add some interactions between predictive variables to see if the model can be better.
Severity = b*humidity + b*temperature + b*visibility + b*traffic_signal + b*weather_condition + b*humidity*visibility + b_intercept+effect_state
Also, I plot the residual plot. Unfortunately, the graph is still not good at this time.

Then I would like to fit the random slope models.

This time I would like to add random effect of traffic signals. And all other variables are recognized as fixed effects.



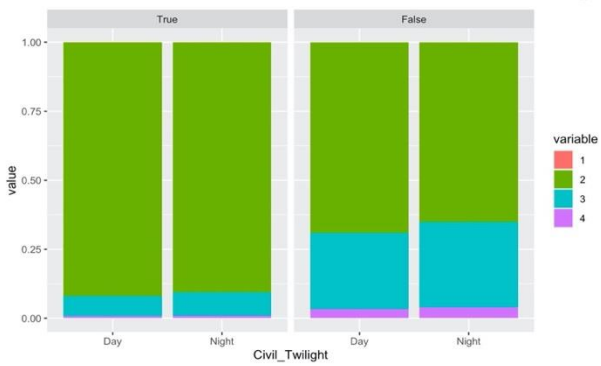


The left graph shows the estimated random effects based on two factors (slope for traffic signals and intercept for states). The right graph shows the binned residual plot of the model. From the plot, we can see most of the observations fall inside the confidence bands. It is much better than the previous models.
Then I would like to try another random slope model. This time, I choose civil twilight as the random effect.

	df <dbl>	AIC <dbl>
fita1	15	1169141
fita2	16	1166661

From the AIC graph, we can see the second model has lower AIC. Moreover, I also check the anova outputs of these models. I choose the second model since it has lower AIC and its p value is significant.
Interpretation: Fixed Effects: The expected value of severity for the accidents happened with no traffic signal, daytime based on civil twilight, zero temperature humidity visibility under clear weather condition is 2.44.
Random Effects: sd(intercept): There is a substantial extent of between states differences in the expected value of severity (0.05).
Sd (civil_twilight): There is a substantial extent of between states differences in the average within states association between civil twilight and states.

Alternative method: The difference between states might be negligible for the dataset. Thus, I would like to regard the whole observations as a single group. The response variable is severity with levels from 1 to 4. It is reasonable to use ordinal logistic regression to fit the model. As a result, I would like to try the ordinal logistic regression model this time.
To fit an ordinal logistic regression model, I use the polr function from MASS package. My plan is to fit the model with different combinations of predictive variables and interactions. And then I would like to check which models are better and make common sense. I would like to use AIC as the standard to evaluate models. Firstly, I would like to start from the simplified model.
Fit1: fit the model with 2 factor variables, Traffic signal + Civil twilight
I visualize the results by plotting predicted probability. This is simple since we only have 4 situations and total 16 probabilities.



Fit2 - 6: fit the model with response variable severity and predictive variables humidity, temperature, visibility, weather condition, traffic signal, civil twilight.

From the output, we can get the regression coefficients with their values, standard errors and t value, residual deviance and AIC.

	df <dbl>	AIC <dbl>
fit1	12	1127781
fit2	5	1093134
fit3	8	1092715
fit4	14	1088215
fit5	15	1086885
fit6	15	1088149

	Value	Std. Error	t value	p value
as.factor(Weather_Condition)Cloudy	-0.2637832437	7.387787e-04	-357.053118	0.000000e+00
as.factor(Weather_Condition)Fair	-0.4268027538	4.260746e-03	-100.170895	0.000000e+00
as.factor(Weather_Condition)Light Rain	0.0276793010	1.554556e-04	178.052737	0.000000e+00
as.factor(Weather_Condition)Mostly Cloudy	-0.0975569329	1.846781e-03	-52.825407	0.000000e+00
as.factor(Weather_Condition)Overcast	0.0947422669	1.755352e-04	539.733783	0.000000e+00
as.factor(Weather_Condition)Partly Cloudy	-0.2060611409	1.082620e-03	-190.335613	0.000000e+00
Visibility.mi.	0.0206924406	1.056541e-03	19.585085	2.072611e-85
as.factor(Traffic_Signal)True	-1.6489776095	9.054701e-05	-18211.286861	0.000000e+00
as.factor(Civil_Twilight)Night	0.2240829466	8.213691e-04	272.816377	0.000000e+00
Humidity...	-0.0150242379	NaN	NaN	NaN
Temperature.F.	-0.0105750460	3.637777e-05	-290.700764	0.000000e+00
Humidity...:Temperature.F.	0.0002400564	NaN	NaN	NaN
1|2	-9.6668133870	1.342303e-04	-72016.615508	0.000000e+00
2|3	0.1054374693	1.470129e-02	7.171987	7.391711e-13
3|4	2.6984482434	1.733096e-02	155.700996	0.000000e+00

The AIC of each model shows above, fit5 has the smallest AIC which means it is the best among these models. As a result, I would like to choose fit5. The significance of coefficients and intercepts shows on the right.

Interpretation of coefficients: mathematically, the intercept 2|3 (0.11) can be interpreted as the log of odds of the severity at level 1,2 versus the severity at level 3,4.

Discussion: In general, the coefficients of the variables indicted that setting travel signals might decrease the level of severity of traffic delay caused by accidents. Besides, daytime based on civil twilight tends to have longer delay than nighttime. Weather conditions and other factors related to weather seem do not have too much impact on traffic severity.

Limitations: Since the dataset I use is the simplified one which does not contain all the variables of the original dataset, I might miss some decisive variables which might critically affect the severity. For example, the real-time traffic flow is an obvious factor which has the impact on traffic severity. Besides, variables like road width, whether it is the time for commuting can also be crucial.

Future direction: I would like to search more materials on the Internet and find more potential variables. Besides, more types and strategies like cross-level interactions, can be tried to make the model more fitted.

Reference: Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Jeremy Albright https://tutorials.methodsconsultants.com/posts/estimating-hlm-models-using-r-part-1/

Akanksha Rawat https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5