Using Multiple Regression for Pedestrian Crash Severity Modeling

Kirsten Johnson
kirstenjohns@umass.edu

UMassAmherst

Data Analytics and Computational Social Science Program

Introduction

At MassDOT, traffic crashes are never referred to as accidents, as the word "accident" implies that a collision occurs by random happenstance. However, the vast majority of crashes are not random. Pedestrian safety is becoming a much bigger topic of conversation in the transportation community as cities are creating more enhanced multimodal infrastructure. Pedestrians are considered vulnerable road users, meaning they are prone to a high risk of injury in any vehicular collision. Since I work as a Traffic Data Specialist for MassDOT, I decided to focus my project on pedestrian crashes in Massachusetts.

Research Questions

- Is there an ideal categorization method for modeling crash severity as an outcome variable?
- What crash factors are most significant in predicting pedestrian crash severity?

I hypothesize that 1) categorizing crash severity as a dummy variable will yield a better fitting model than categorizing it on an ordinal numeric scale with multiple levels and 2) factors related to driver behavior are more significant than roadway characteristics or other external factors in predicting pedestrian crash severity.

Data

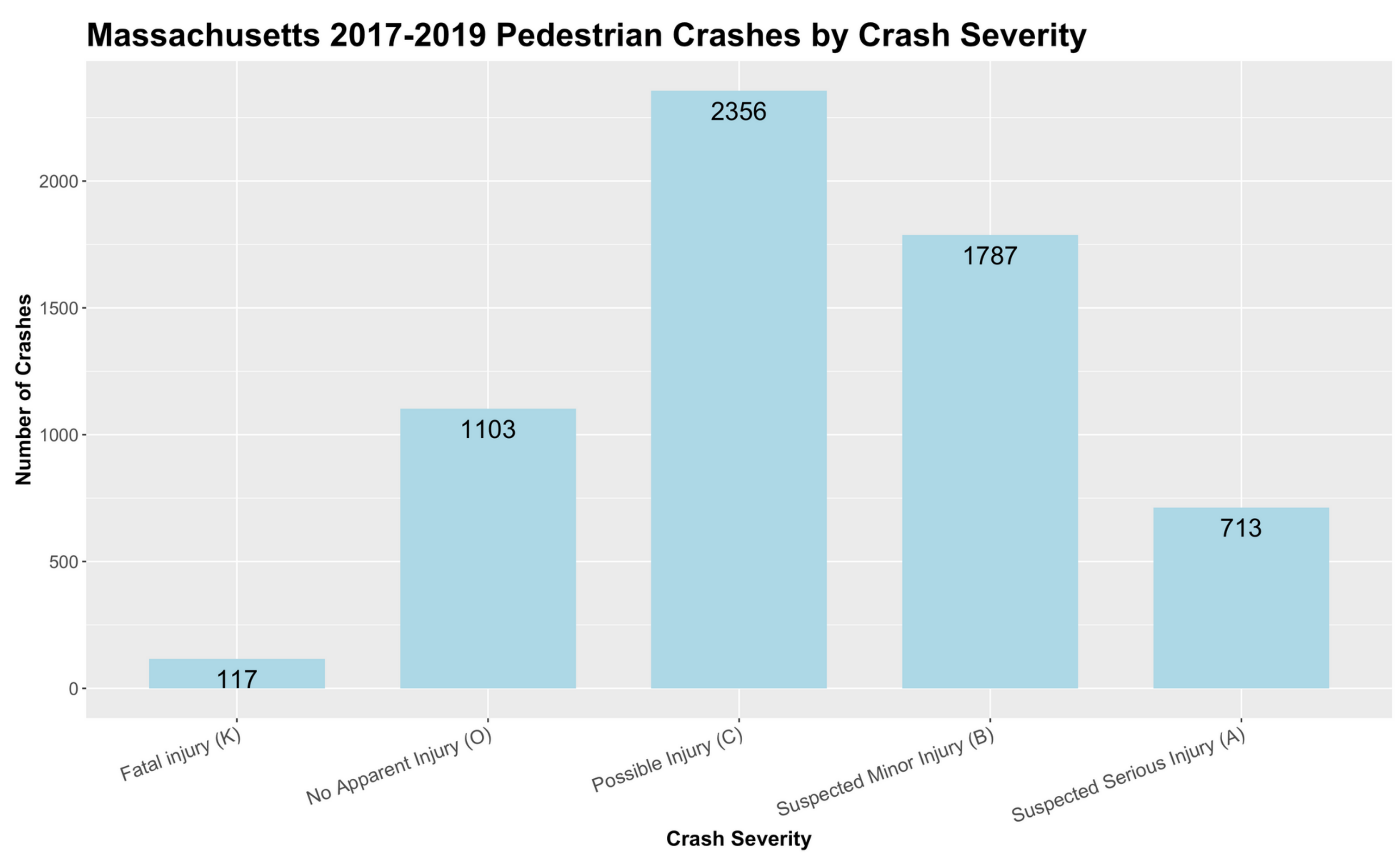

Figure 1. Pedestrian Crash Distribution by Severity using KACBO scale

There were a total of 6,079 Massachusetts crashes involving a pedestrian that took place between 2017 and 2019. Crash data are derived from the reports that Massachusetts local, state, MBTA Transit and some campus police agencies send to the MassDOT Registry of Motor Vehicles (RMV) division. The data was obtained from the MassDOT IMPACT crash data portal using the Crash Query and Visualization tool.

Due to inconsistencies among police reporting, extensive filtering and recoding was necessary for the variables of interest. As a result, some of the statistical findings may have been influenced.

- **Outcome variables:** Two crash severity categorization variables were created to compare as outcome variables. In MA, crashes are categorized using the KABCO Injury Classification Scale (Fig.1). However, for prioritizing safety projects, KA (Fatal & Serious Injury) crashes are given more weight. To determine which scale is optimal, an ordinal numeric variable was created for the KABCO scale, and a dummy was created for KA or not KA crashes.
- **Explanatory variables:** 5 outcome variables were determine by reviewing the crash data. While the data has numerous potential variables, variable selection was limited due to poor data quality. There were numerous categories contained in each of the variables chosen. To avoid model issues, a dummy variable was created for predictor: *Pavement friction impaired, road contributing circumstance, ambient light, traffic control device, and improper driving behavior.*
- **Training Data:** The data was split into 70% training and 30% test to evaluate the models.

Methods

Crash Severity Model Diagnostics

- Each version of the crash severity outcome variable was fit to an MLR model using all explanatory variables.
- The models were evaluated via diagnostics plots using the assumption of linearity and normality.
- Whichever model had better fit using the given variables would be further evaluated for predictor optimization.

Pedestrian Crash Severity Predictors

- Using the KABCO scale as the outcome variable, backwards elimination was used to see if the model could be improved by systematically selecting the most statistically significant variables.
- The backwards elimination model was compared with the model using all explanatory variables.

Analysis

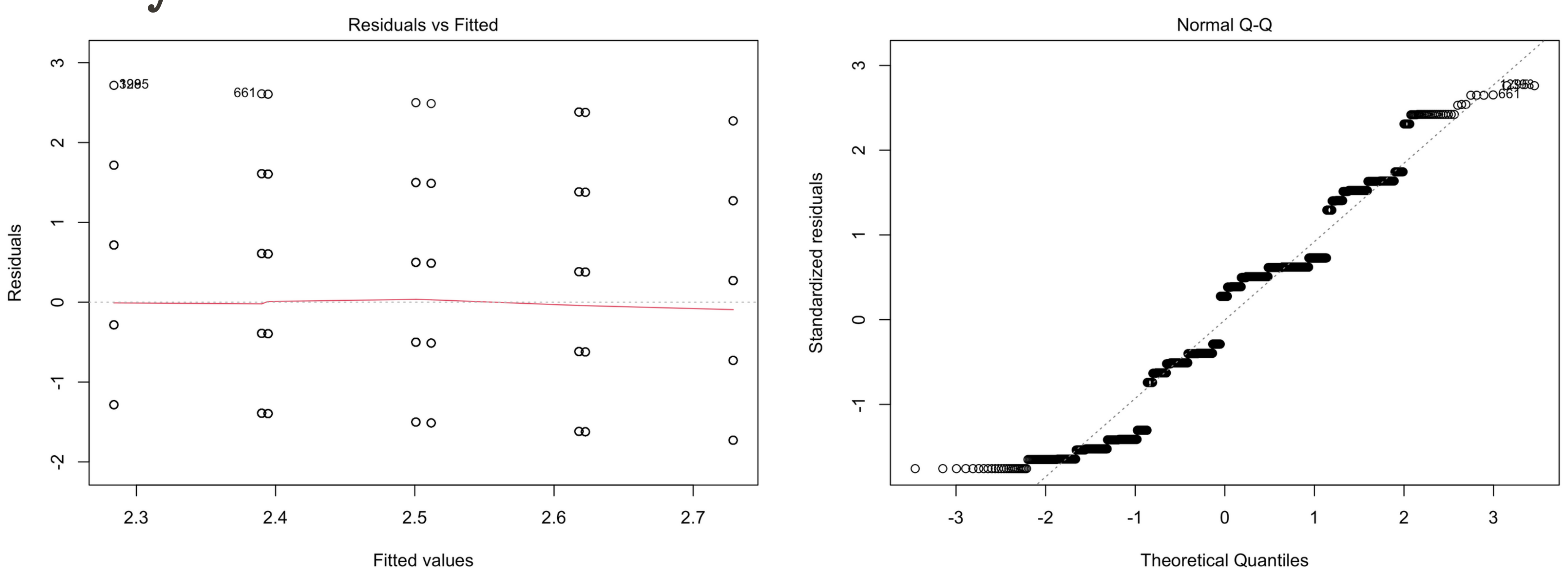

Figure 2. Backwards elimination diagnostics plots. Residuals vs Fitted shows the linearity assumption is upheld, Normal Q-Q plot shows some normality.

Hypothesis 1- Outcome Variable Evaluation

- When comparing the diagnostics plots, the KABCO numeric scale was found to be a better categorization method than the KA dummy variable. Figure 2 shows diagnostics for the KACBO Crash Severity level model.
- While the regression assumptions do look to be maintained, the model is not perfect as there are clear outliers.
- The Residuals vs Fitted plot points form a very horizontal formation, likely due to the limited values for the variables. Still, this looks to confirm the linearity assumption.
- The Normal Q-Q plot shows the points to be pretty linear, confirming the normality assumption. Some clear clustering is apparent, again speaking to the limited categories for the variables.

Hypothesis 2 - Predictor Evaluation

- Using training data, the backwards elimination model determined that ambient light, traffic control device presence, and improper driving were the most statistically significant crash severity predictors as they had the highest p-values (Table 1).
- PRESS, AIC, and BIC favor the backwards elimination model.
- The R^2 and Adjusted R^2 for both models is very small, meaning the predictors don't really explain how much variance there is for injury severity level.
- The F Statistic for both models is very significant, meaning our model explains the outcome better than if no predictors were present
- As a final model evaluation test, the backwards elimination model was fit using the remaining 30% test data and compared to the expected values(Fig. 3).

Table 1 shows two Ordinary Least Squares (OLS) models for Injury Severity Level, one using backwards elimination and one using all predictor variables.. Coefficients and 95% confidence intervals are on the same row. Summary statistics are reported for model comparison.

Table 1. Regression Results

	Dependent variable:	
	Injury Severity Level	
	All Predictors	Backwards Elimination
	(1)	(2)
Pavement Friction Impaired	-0.042 (-0.045, -0.040)	
Road Contributing Circumstance	-0.081 (-0.085, -0.077)	
Ambient Light	-0.171^{***} (-0.173, -0.169)	-0.164^{***} (-0.166, -0.162)
Traffic Control Device	-0.096^{***} (-0.098, -0.094)	-0.097^{***} (-0.099, -0.096)
Improper Driving	0.153^{***} (0.151, 0.155)	0.153^{***} (0.151, 0.155)
Constant	2.598^{***} (2.596, 2.600)	2.582^{***} (2.580, 2.584)
AIC	11824.04	11822.68
BIC	11868.53	11854.45
PRESS	4012.704	4011.383
Observations	4,253	4,253
R^2	0.014	0.014
Adjusted R^2	0.013	0.013
Residual Std. Error	0.971 (df = 4247)	0.971 (df = 4249)
F Statistic	12.248^{***} (df = 5; 4247)	19.532^{***} (df = 3; 4249)
Note:		*p<0.1; **p<0.05; ***p<0.01

Conclusion

Using a numeric ordinal scale rather than a dummy variable for crash severity improved the model significantly. However, this may not stay consistent if additional variables are added or if variables were coded differently. Driver behavior was just as statistically significant as other crash characteristics, so I failed to reject both of my null hypotheses.

Given there were multiple significant p-values for the predictors and F Statistics for the models, it is clear that the variables chosen do contribute to explaining some aspect of pedestrian crash severity. Still, the model could not account for the "noise" in the data and failed to explain much in the variation of the crash severity level. This means more variables are needed if the model were to be used for predicting crash severity. At a glance, the results may not look groundbreaking. However, the findings confirm just how needed a model-based approach to traffic safety is needed. There are countless factors that need to be accounted for in order improve crash severity modeling and ultimately work towards reducing roadway fataltities

References

1. Yannis, George, et al. "Vulnerable Road Users: Cross-Cultural Perspectives on Performance and Attitudes" ScienceDirect | Science, Health and Medical Journals, Full Text Articles and Books., International Association of Traffic and Safety Sciences, Oct. 2020, https://www.sciencedirect.com/science/article/pii/S0386111220300716#
2.https://safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf
3. Crash Query and Visualization tool - https://apps.impact.dot.state.ma.us/cdv/