Injury Severity Caused by Traffic Accidents

Alyssa Cable, Kamie Champlin, Darryl Gatrell, and Clinton Williams

April 17, 2020

Introduction

Traffic accidents impose high human and monetary costs on society, but the irony is that most accidents are caused by human error and not by chance. The National Highway Traffic Safety Administration (NHTSA) has actively advocated for drivers to avoid driving under the influence of alcohol or other judgement impairing substances in order to keep the accident numbers and injuries low, but we know that there are many other factors that also contribute to the occurrence of auto accidents. Insurance companies repeatedly say that young drivers are at a higher risk, thus their insurance rates are higher. Driving at night diminishes visibility and makes it harder to see potential issues. Parents buy their children older models of cars because young adults are more likely to total their vehicles. While older models are generally cheaper, they are also commonly believed to be more dangerous. Because there are so many unknown factors that play a part in these situations, we want to find answers on how one can best limit the severity of traffic accident injuries.

Our final project explores traffic accidents that resulted in the condition of a driver having no injury at all, suffered a fatality, or anywhere in between. Using variable selection and multiple linear regression techniques, we will analyze the relationship between many of these variables and the severity of the accident in regards to human injury. The motive behind this project is to find out what relationship certain factors have to how severe an injury is after an automobile crash, in order for the human population to be safer and smarter on the roads.

Data

We will be analyzing public data published by NHTSA that was collected using the Fatality Analysis Reporting System (FARS). FARS is a nationwide census providing NHTSA, Congress and the American public yearly data regarding injuries suffered in motor vehicle traffic crashes (NHTSA, 2020). We will be analyzing data from 2010-2018, making the results of our analysis applicable to our current day. This data set includes 710,265 observations of traffic accidents that occurred in the 50 United States of America and some outlying territories. We will obtain a model from data from 2010-2014 that was collected for the state of Utah, and preserve data from years 2015-2018 as a test set. This splitting of the data reserves 2,586 observations in the training data set and 2,505 observations in the test data set.

This data set includes both qualitative and quantitative variables, and the model will be fit to the response variable Injury Severity. The values of this variable range from no injury to fatal injury as a result of the accident (see Table 1). We choose to remove data where the injury severity variable is greater than 5 (5=injured but severity unknown, 6=died prior to crash, 7=blank, 8=not reported, 9=unknown/not reported). This choice is made because these points accounted for a very small percentage of the data and their values made no difference in predicting how severe an individual's injury is. The response appears to have short-tailed distribution and exhibits bi-modality, see Figure 1. This is unlikely to have any negative effects on our regression techniques. This will be discussed further in our analysis of the original model.

Table 1: Levels of Response Variable

0	No Injury/No Apparent Injury
1	Possible Injury
2	Non-Incapacitating Evident Injury
3	Incapacitating Injury/Suspected Serious Injury
4	Fatal Injury

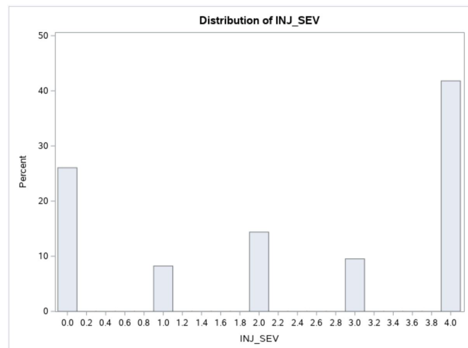

Figure 1: Histogram of Injury Severity

Table 2: Variable Names

DEAD	Did the passenger die or not.
MINS_MIDNIGHT	Number of minutes from midnight that the crash occurred.
VE_FORMS	Number of vehicles involved in accident.
MAN_COL	Manner of collision
SCH_BUS	Was the crash school bus related.
BODY_TYP	Body/size of the car.
MOD_YEAR	Model year of automobile.
TOW_VEH	Whether or not vehicle was towing another unit or vehicle.
FIRE_EXP	Whether or not the accident involved a fire explosion.
AGE	Age of driver
SEX	Sex of driver (1=Male, 2=Female)
INJ_SEV	Severity of injuries received
DRINKING	Whether or not driver of vehicle was drinking.
DRUGS	Whether or not driver of vehicle was using drugs.
HOSPITAL	Mode of transportation to hospital or medical facility.
RACE	Race of driver.
IMPACT1	Initial point of impact.

We begin with roughly 60 predictor variables, ranging from age of driver to what time of day the crash happened. After examining these variables, we are able to narrow them down to 17 by using human judgement. Table 2 lists each of those 17 variables and an explanation of each. With the data trimmed down to a workable size, we can perform further analysis in order to fit a regression model for injury severity. For the remainder of this paper, whenever refer to the "orignal" or "raw" data, we are speaking about the data for the state of Utah with 17 different variables.

Original Model Analysis

Upon fitting a regression model to the raw data, there is little evidence of heteroskedasticity in the residuals. However, we find the data to follow a bi-modal distribution (see Figure 2). In order to address this problem, we run multiple common transformations on the response variable, including a log transformation, square root transformation, and inverse square root transformation. Each transformation only makes the bi-modal distribution more prominent and fails to change the QQ plot and push our data towards normality. This suggests that the data could be non-linear. The attempts made to fit this data using methods other than linear regression can be seen in this paper under the section **Alternative Regression Methods**. Ultimately, we choose to continue with an Ordinary Least Squares approach, as the response variable is quantitative, and determine if a model exists that will help us to predict injury severity as a linear function of other variables.

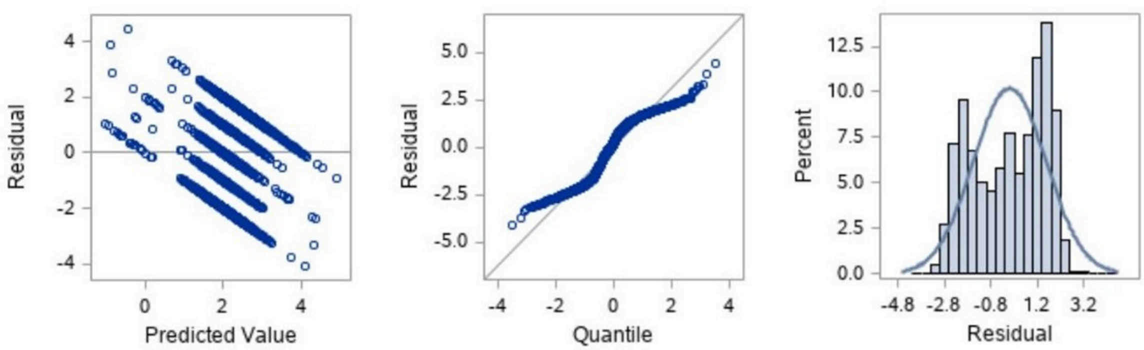

Figure 2: Fit Diagnostics of Raw Data

Variable Selection

Upon exploring the variables listed in Table 1, we decide to remove the variable DEAD. This is a confounding variable, as the response will always be "severe" when the individual dies. After removing the variable, there are no changes in our normality checks, though we notice that our predictive abilities went down, as expected. In order to improve the predictive capability and interpretability of our model, we want to retain only predictors that are significant in explaining our response variable. We use variable selection techniques in order to further slim the group of variables we had manually selected.

We briefly attempted to perform a LASSO variable selection test, but experienced difficulties in regards to the qualitative variables. For this reason, we use stepwise selection in order to choose variables for a linear model, using a significance level of 0.1 as our threshold for both entry and removal. The significant variables we received from this analysis are as follows: VE_FORMS, AGE, MOD_YEAR, FIRE_EXP, TOW_VEH, BODY_TYP, MINS_MIDNIGHT, SEX, IMPACT1.

We explored some interactions terms that were of particular interest. The results for a test of significance are given in Table 3. Because no significance was found, we decided not to include any interaction terms.

Table 3: Significance of Interaction Terms

Interaction Term	P-value
Age and Year	0.1351
Age and Body Type	0.9152
Year and Body Type	0.1683

Regression Model

After completing the stepwise variable selection process, we obtained an equation containing the intercept and nine predictor variables, defined in Table 4. $\hat{Y}$ is the variable associated with the response of injury severity from an automobile collision.

$$\hat{Y} = 52.94 - 0.20X_1 + .01X_2 - .49X_3 - 0.03X_4 + 1.38X_5 + 0.01X_6 - 0.0003X_7 + .14X_8 - 0.01X_9$$

Table 4: Model Variable Names

Variable	Name
X_1	VE_FORMS
X_2	AGE
X_3	TOW_VEH
X_4	MOD_YEAR
X_5	FIRE_EXP
X_6	BODY_TYP
X_7	MINS_MIDNIGHT
X_8	SEX
X_9	IMPACT1

Using an ordinary least squares regression model allows for simple interpretation of our coefficients. For example, if we were to take the variable of Age, which corresponds to X_2, we see that we have obtained a coefficient of positive .01. This means for every one year increase in age, we can expect an increase in injury severity of 0.01, holding all else constant. This interpretation can be applied to all of the other variables listed above. Two of the other large predictors are X_3, which corresponds to what was being towed behind the vehicle, and X_5, which is whether or not a fire had to be extinguished at the scene of the crash. TOW_VEH has a coefficient of -0.49, meaning holding all else constant, for each increase in number of trailers, we expect a decrease of 0.49 in injury severity. FIRE_EXP has a coefficient of 1.38, meaning holding all else constant, if a fire had to be extinguished at the scene of the accident, we expect there to be an increase of 1.38 in injury severity.

Outliers and Influential Points

Initial examination of outliers and influential points was not possible due to the large amount of data involved, so we revisit this subject following variable selection and the trimming of our data. When examining each of our predictor variables, we found some outliers in the data, but we assume that most of them simply come from extreme observations and do not present issues in the accuracy of our model. However, in the variable "Minutes after Midnight", we find an outlier at about 6,000 minutes. Because there only 1,400 minutes in a day, we know that this is a clerical error in the data and have sufficient evidence to support removing this value in order to more accurately model the data.

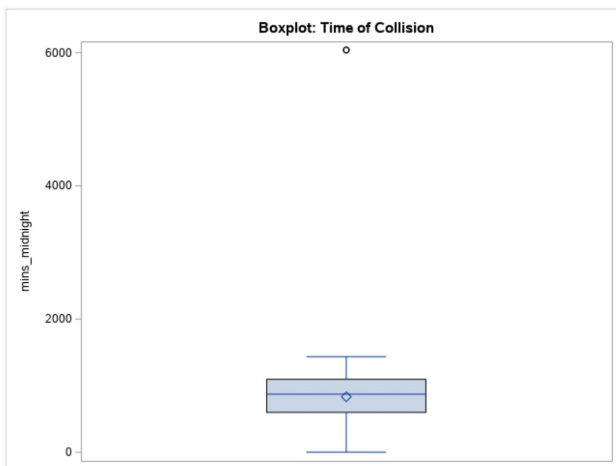

Figure 3: Box Plot for Minutes After Midnight

Overall, after selecting which variables will be included in our final regression model, there are no apparent outliers in the data. We can see in the scatter plot of the residuals and normal probability plot below (Figures 4 and 5) that all points seem to follow the distribution as expected.

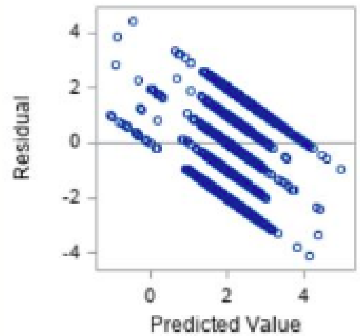

Figure 4: Scatter Plot of Residuals After Variable Selection

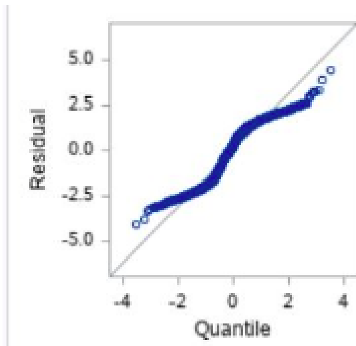

Figure 5: Normal Probability Plot of Data After Variable Selection

However, we do find multiple influential points when examining our leverage and Cook's distance plots, as seen in Figures 6 and 7. Because our data include over 2,500 observations, it is difficult for us to identify why these points are influential. But they are relatively few in comparison with the overall amount of data we have, and no one point seems to be significantly more influential than the others, according to the Cook's distance plot in Figure 5. Due to the lack of dramatic points of interest, we conclude that these few points will only slightly influence the values we obtain for our parameter estimates (β_k) and/or the predicted values we obtain for injury severity ($\hat{Y}$), especially when considering the large quantity of observations in the data set.

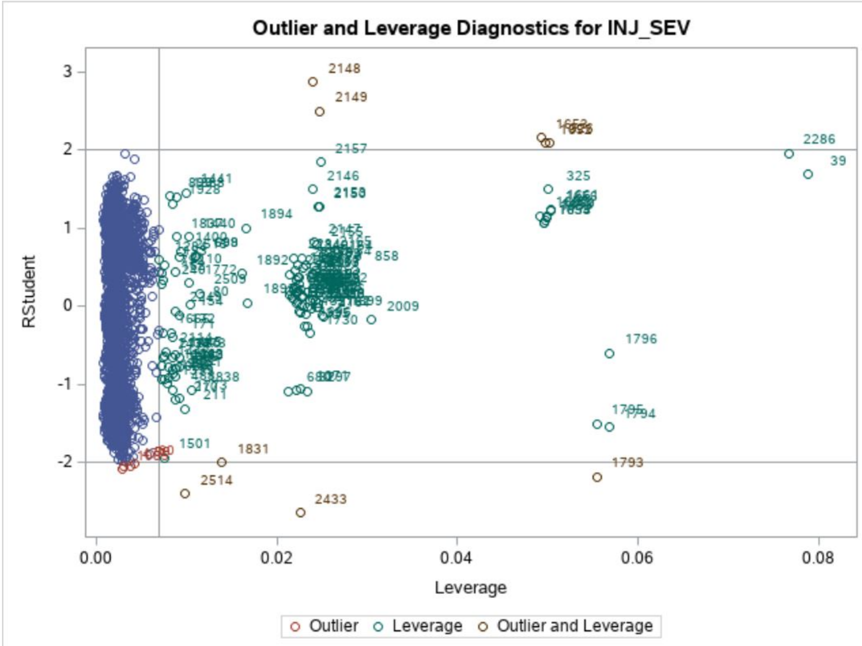

Figure 6: Leverage Plot

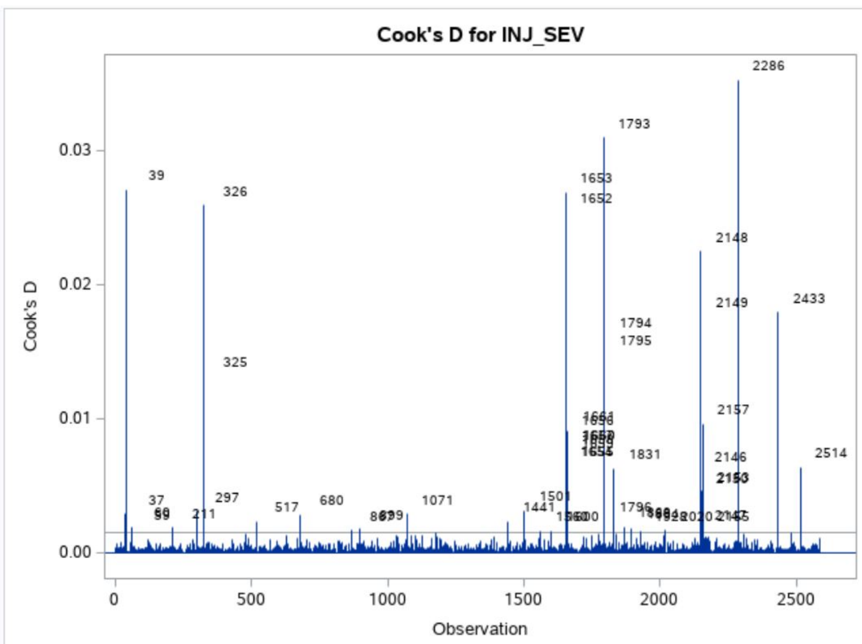

Figure 7: Cook's Distance Plot

Multicollinearity

Before using the model to make assumptions, we must check for multicollinearity in the data. If present, it would be necessary to account for it in final inference, giving us less validity overall. Upon running a test to find the Variance Inflation Factor for each variable, we find each of those values to be close to 1 (see Figure 5), suggesting no evidence of multicollinearity. This is confirmed further upon examining the condition indices and finding that variables 1-8 have a value of below 10. We find one variable with a condition index greater than 10, but it doesn't have multiple values associated with proportion of variance over 50%, suggesting that multicollinearity will not be an issue. Thus, we move forward with our model without accounting for any multicollinearity.

Table 5: Checks for Multicollinearity

Variable	VIF	Eigenvalue	CI
VE_FORMS	1.02601	1	1.00000
AGE	1.02399	2	2.48789
TOW_VEH	1.09209	3	2.52269
MOD_YEAR	1.02156	4	3.54804
FIRE_EXP	1.01932	5	3.68933
BODY_TYP	1.17580	6	4.40847
MINS_MIDNIGHT	1.02140	7	5.42188
SEX	1.07137	8	6.74683
IMPACT1	1.01919	9	10.77287
		10	969.38981

Final Model Assumptions

In order to ensure that the final model satisfies assumptions, we must be sure that the model itself is significant. We look at the Analysis of Variance table and find an F-statistic value of 39.18, with a corresponding P-value of < 0.0001. This confirms the presence of a statistically significant relationship between our final fitted model and predicted variable of injury severity.

We use the plots attached to check the normality and constant variance assumptions, as seen in Figure 8. Looking first at the plot of the Residuals vs Predicted Values, we identify the presence of both categorical and discrete data as we see four lines running from the top left corner to bottom right. This is expected due to the nature of the data and tells us that we have roughly constant variance, as all lines are about the same length. This conclusion is reinforced by the large amount of data we have, suggesting a constant variance overall.

The curve found in the QQ-plot presents potential normality concerns. We note from the histogram that the issue with bi-modal distribution present in the crude original model is largely unchanged, supporting the suspected violation of normality. However, no amount of transformations we can perform will overcome this violation, as seen in the crude model assumptions mentioned previously. Although the normality assumption is not satisfied, we are still confident in our ability to draw inference from the model.

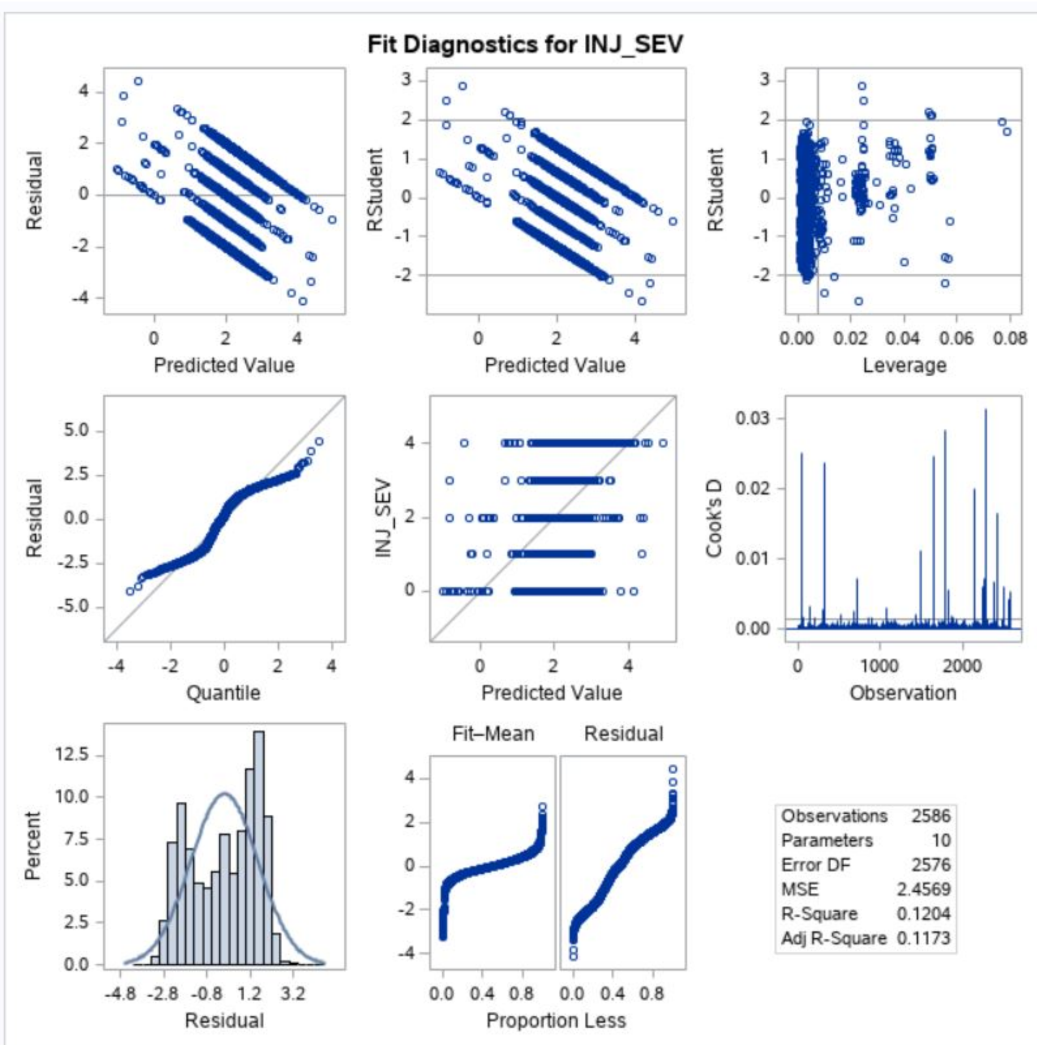

Figure 8: Post Selection Fit Diagnostics

The model has an Adjusted R-Squared value of 0.1173, meaning that the model accounts for 11.73% of the variance in the overall data. With only 11.73% of the variance being explained by the model, we will not have perfect accuracy in drawing conclusions and making predictions. However, it is important to note that for non-experimental data such as we are working with, it is expected to see an R-Squared value of less than 30%, so we should not be alarmed by the R-Squared value obtained.

Model Accuracy

In order for us to conclude that our model is truly effective in prediction, we compare the validation error of our model to that of a model with no parameter estimates and to a model with all 17 original parameter estimates included. We use the test set of data that we earlier reserved in order to perform this analysis with data we haven't seen before. This is important because we should be able to predict trends for all data, not only the data that we created our model with.

On comparing these values, we see that the validation error (MSPR) for the model with only the intercept term is 2.797, while the model with all 17 predictors has an MSPR of 2.501 and our reduced model has an MSPR of 2.477. This tells us that the model containing the 17 original predictors has better predictive capabilities than simply using the mean value for injury severity, and that the model with the variables we have selected is better than simply using all 17 predictors we began with.

Alternative Regression Methods

The low R-Squared values observed in our linear regression models is evidence that the linear model was not adequate in describing the data. The data also follow a bi-modal distribution, suggesting non-normality, but improvement was not seen following various transformations of the data. Many of the variables exhibited extremely skewed distributions and not all of them benefited from transformations. Additionally, our response variable, though quantitative, is discrete. Furthermore, it is not possible to define the distance between "no injury" and "minor injury" in a way that is comparable to the distance between "minor injury" and "severe injury". We attempted to achieve a richer description of the data through alternative regression methods. Given the structure of our response variable, we tested multi-class logistic regression but discovered our data to be unsuitable for this method without taking extreme measures to correct some issues specific to the data. Due to these complications, we chose not to use logistic regression to model the data.

Our next alternative was a tree based method. Regression trees assign a prediction to all values that fall into a certain category, which seems to parallel the structure of the response variable. We first constructed a regression tree using all the variables mentioned in Table 2 excluding the DEAD variable. The fully grown tree had 604 nodes but was pruned back to just 15. The most important variables in the tree are the time of admittance to hospital, the body type / size of vehicle, the place of initial impact, if drugs were involved, the age of the driver, the time of day, and the model year of the vehicle.

In general, when patients are admitted early in the morning or late at night they are more likely to have sustained more serious injuries, possibility due to low visibility at these times. Smaller cars, with exceptions for trucks and tractors, are associated with higher levels of injury severity, especially for collisions that occur perpendicular to the vehicle (t-bone). Injuries are more severe when the driver is elderly or very young. Front to front (head on) collisions and end swipes are much more likely to cause severe injury than rear ends, glancing or angled collision, and collisions involving few vehicles are more likely to result in injury than collisions involving many vehicles. An interesting finding is that injury severity is much lower for drivers under the influence of drugs.

The Average Squares for Error (ASE) for this tree was 1.8290. This is significantly lower than the MSE found using linear methods. This is evidence that a non-linear method is more suited for this data set.

We validated this regression tree, constructed from the training data set, with the test data set and found an ASE of 2.0163. This was lower than the MSPR found using linear regression models. This is further evidence that a non-linear model is more appropriate for this data. However, for the purposes of our analysis and in order to obtain a model that could be interpreted simply, we chose to continue to use multi-variable linear regression.

Although we chose to use a regression tree for its transparency, we also performed a Random Forest routine on the data using the same variables as the regression tree. We found that the decrease in MSE associated with each variable was low. This shows that each variable individually has little predictive power which may be why simple linear models are unable to adequately model the data.

Conclusion

Using FARS survey data, a model was constructed to predict the severity of car accident injuries. Multiple methods were used to predict injury severity, and the differing results among the models enabled us to gain insight about the data.

We first fit a multi-variable linear regression model on the data. After variable selection our final model included the following variables: number of cars involved in the crash, age of driver, how many and what type objects the vehicle was towing, the model year, the body type, if the vehicle crash resulted in a fire at the scene of the accident, the time of day, the gender of the driver, and the place of initial impact. Government legislation and law enforcement in Utah can focus on these variables to reduce the severity on injuries sustained in a car crash.

Although it may seem counter-intuitive that more cars involved in a crash leads to less severe injuries, it could be due to these accidents generally occurring during a period with high congestion– leading to lower overall speeds. The model also suggests that young drivers are more likely to be seriously injured, suggesting that young drivers need more experience and training. Accidents involving vehicles that are towing cargo tend to suggest slower moving vehicles, leading to a lower chance of severe injury for the driver. The same can be said for vehicles that easily catch on fire. The model also suggests that females are more likely to sustain serious injuries than men, meaning vehicle companies may need to invest more research into this difference and design additional safety features for female drivers.

Our models are evidence that the FARS national survey is inadequate in predicting injury severity. The selection of variables in this survey cannot provide a satisfactory relation to injury severity on their own. We believe that additional information is necessary for three reasons: the current selection of variables has low predictive accuracy, some important variables were absent from the survey, and these absent variables may have important interactions with the current selection of variables.

Our linear regression models had R-Squared values below 20%. This is high enough to make some inference about the variables in the model, but too low to make general rules of thumb regarding these variables. When we fit a regression tree to the data, we noticed that the reduction in MSE was very small at every split. This also suggests that none of the variables have significant predictive power.

Looking at the variables selected in this survey, the speed of the collision was notably absent. We believe that variables such as speed, especially the difference between the driver's speed and the posted speed limit, would provide increased predictive power to the model. The degree to which a driver exceeded the speed limit would not only provide extra information, but it may have interesting interaction effects with other variables. For example, the age of the driver was an important part of our model, but there may be a significant interaction between speed and age which could be characterized in the model.

For the reasons stated above, future research should include an analysis of speed and its possible interaction with the current selection of variables found in the FARS crash survey. The current analysis used a selection of these variables to model vehicle crashes in the state of Utah. It is possible that this selection is not common to all states in the US. Further research may find that different states must focus on different selections of variables.

References

NHTSA (2020). Fatality analysis reporting system (fars).

Appendix

```
/* 2010 - 2014 Data */
FILENAME REFFILE ’/home/alyssacable250/EPG194/PERSON2010.CSV’;
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2010; GETNAMES=YES; RUN;

FILENAME REFFILE ’/home/alyssacable250/EPG194/PERSON2011.CSV’;
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2011; GETNAMES=YES; RUN;

FILENAME REFFILE ’/home/alyssacable250/EPG194/PERSON2012.CSV’;
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2012; GETNAMES=YES; RUN;

FILENAME REFFILE ’/home/alyssacable250/EPG194/PERSON2013.CSV’;
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2013; GETNAMES=YES; RUN;

FILENAME REFFILE ’/home/alyssacable250/EPG194/PERSON2014.CSV’;
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2014; GETNAMES=YES; RUN;


data work.fullTrain; set work.person2010(drop=MCYCL_DS) work.person2011
work.person2012 work.person2013 work.person2014;
mins_midnight = hour*60 + minute;
if DEATH_YR in (’8888’, ’9999’) then DEAD=0; else DEAD=1;
drop ALC_DET ALC_RES ALC_STATUS ATST_TYP AIRBAG CARBUR CYLINDER CERT_NO COUNTY
DISPLACE DEATH_DA DEATH_MO DEATH_YR DEATH_HR DEATH_MN DEATH_TM DAY DRUG_DET
DSTATUS DRUGTST1 DRUGTST2 DRUGTST3 DRUGRES1 DRUGRES2 DRUGRES3 DOA EJ_PATH
EMER_USE EJECTION EXTRICAT FUELCODE HOUR HARM_EV HISPANIC IMPACT2 LAG_HRS
LAG_MINS LOCATION MCYCL_DS  MCYCL_CY MCYCL_WT MONTH MINUTE MAK_MOD MAKE N_MOT_NO
PER_NO PER_TYP P_SF1 P_SF2 P_SF3 ROLLOVER REST_USE REST_MIS ST_CASE STR_VEH
SER_TR SPEC_USE SEAT_POS TIRE_SZE TON_RAT TRK_WT TRKWTVAR VEH_NO VIN_REST VIN_BT
VIN_LNGT VINMODYR VINTYPE VINMAKE VINA_MOD VIN_WGT WHLDRWHL WGTCD_TR WHLBS_LG
WHLBS_SH WORK_INJ;
run;


data work.selectTrain;
set work.fulltrain;
keep DEAD MINS_MIDNIGHT VE_FORMS MAN_COLL SCH_BUS BODY_TYP MOD_YEAR TOW_VEH
FIRE_EXP AGE SEX INJ_SEV DRINKING DRUGS HOSPITAL RACE IMPACT1 STATE;
if MOD_YEAR in (’9998’, ’9999’) then MOD_YEAR = .;
if AGE in (’998’, ’999’) then AGE = .;
where INJ_SEV in (0,1,2,3,4) and STATE=49;
if cmiss(of _ALL_) then delete;
run;

data work.selectTrain;
set work.selectTrain;
```

```
if mins_midnight < 1440;
run;


/* 2015-2018 Data */
FILENAME REFFILE '/home/alyssacable250/EPG194/person2015.csv';
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2015; GETNAMES=YES; RUN;

FILENAME REFFILE '/home/alyssacable250/EPG194/PERSON2016.csv';
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2016; GETNAMES=YES; RUN;

FILENAME REFFILE '/home/alyssacable250/EPG194/PERSON2017.csv';
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2017; GETNAMES=YES; RUN;

FILENAME REFFILE '/home/alyssacable250/EPG194/PERSON2018.csv';
PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=WORK.person2018; GETNAMES=YES; RUN;


data work.fullTest;
set work.person2015 work.person2016 work.person2017 work.person2018;
mins_midnight=hour*60 + minute;
if DEATH_YR in ('8888', '9999') then DEAD=0; else DEAD=1;
drop ALC_DET ALC_RES ALC_STATUS ATST_TYP AIRBAG CARBUR CYLINDER CERT_NO COUNTY
DISPLACE DEATH_DA DEATH_MO DEATH_YR DEATH_HR DEATH_MN DEATH_TM DAY DRUG_DET
DSTATUS DRUGTST1 DRUGTST2 DRUGTST3 DRUGRES1 DRUGRES2 DRUGRES3 DOA EJ_PATH
EMER_USE EJECTION EXTRICAT FUELCODE HOUR HARM_EV HISPANIC IMPACT2 LAG_HRS
LAG_MINS LOCATION MCYCL_DS  MCYCL_CY MCYCL_WT MONTH MINUTE MAK_MOD MAKE N_MOT_NO
PER_NO PER_TYP P_SF1 P_SF2 P_SF3 ROLLOVER REST_USE REST_MIS ST_CASE STR_VEH
SER_TR SPEC_USE SEAT_POS TIRE_SZE TON_RAT TRK_WT TRKWTVAR VEH_NO VIN_REST VIN_BT
VIN_LNGT VINMODYR VINTYPE VINMAKE VINA_MOD VIN_WGT WHLDRWHL WGTCD_TR WHLBS_LG
WHLBS_SH WORK_INJ;
run;

data work.selectTest;
set work.fullTest;
keep DEAD MINS_MIDNIGHT VE_FORMS MAN_COLL SCH_BUS BODY_TYP MOD_YEAR TOW_VEH
FIRE_EXP AGE SEX INJ_SEV DRINKING DRUGS HOSPITAL RACE IMPACT1 STATE;
if MOD_YEAR in ('9998', '9999') then MOD_YEAR=.;
if AGE in ('998', '999') then AGE=.;
where INJ_SEV in (0, 1, 2, 3, 4) and STATE=49;
if cmiss(of _ALL_) then delete;
run;


/* Attempt at Ordinal Logistic Regression */
proc sort data=work.selectTrain; by descending INJ_SEV; run;
proc logistic data=work.selectTrain;
class ROAD_FNC MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX IMPACT1;
```

```
model INJ_SEV(order=data) = AGE VE_FORMS MAN_COLL IMPACT1 SCH_BUS BODY_TYP
MOD_YEAR TOW_VEH FIRE_EXP SEX MINS_MIDNIGHT / link=glogit;
run;

/* running this shows which variables have the same amount of observations */
proc means data=work.selectTrain; run;

/* Histogram of Response */
proc univariate data=work.selectTrain;
histogram INJ_SEV;
title1 "Histogram: Severity of Injury";
run;


/* Box Plots */
proc sgplot data=work.selecttrain;
vbox mins_midnight; title1 "Boxplot: Time of Collision";
run;

proc sgplot data=work.selecttrain;
vbox VE_FORMS; title1 "Boxplot: Number of Vehicles Involved";
run;

proc sgplot data=work.selecttrain;
vbox MOD_YEAR; title1 "Boxplot: Model Year of Vehicle";
run;

proc sgplot data=work.selecttrain;
vbox AGE; title1 "Boxplot: Age of Individual Involved";
run;

proc sgplot data=work.selecttrain;
vbox INJ_SEV; title1 "Boxplot: Severity of Injury";
run;

proc reg data=work.selectTrain
plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP
MINS_MIDNIGHT SEX / vif collin;
run;

/* Transformation attempts */
data work.selectTrain; set work.selectTrain;
LOG_INJ_SEV = log(INJ_SEV);
run;

proc reg data=work.selectTrain;
model LOG_INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP
```

```
MINS_MIDNIGHT SEX / vif collin;
run;

proc reg data=work.selectTrain;
model INJ_SEV = AGE VE_FORMS MAN_COLL IMPACT1 SCH_BUS BODY_TYP
MOD_YEAR TOW_VEH FIRE_EXP SEX MINS_MIDNIGHT;
store RegModel3;
run;


/* Variable Selecion */
proc reg data=work.selectTrain;
model INJ_SEV = AGE VE_FORMS MAN_COLL IMPACT1 SCH_BUS BODY_TYP MOD_YEAR TOW_VEH
FIRE_EXP SEX MINS_MIDNIGHT / selection=stepwise slentry=.10 slstay=.10;
title1 'Stepwise Selection';
run;

/* regression model with 9 variable after stepwise selection */
proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP
MINS_MIDNIGHT SEX IMPACT1 / vif collin;
output out=trainout residual=resid predicted=pred;
store regModel;
run;

/* regression model with 6 variables after stepwise selection */
proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP / vif collin;
output out=trainout residual=resid predicted=pred;
store regModel;
run;


/* checking for interactions */
data work.selectTrain; set work.selectTrain;
AGE_YEAR = AGE*MOD_YEAR;
AGE_BOD = AGE*BODY_TYP;
YEAR_BOD = MOD_YEAR*BODY_TYP;
run;

proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP MINS_MIDNIGHT SEX
IMPACT1 AGE_YEAR AGE_BOD YEAR_BOD;
title "All three interactions";
run;

proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP MINS_MIDNIGHT SEX
```

```
IMPACT1 AGE_YEAR AGE_BOD;
title "No Year_Bod interactions";
run;

proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP MINS_MIDNIGHT SEX
IMPACT1 AGE_YEAR YEAR_BOD;
title "No Age_Bod interactions";
run;

proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP MINS_MIDNIGHT SEX
IMPACT1 AGE_BOD YEAR_BOD;
title "No Age_Year interactions";
run;

proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP MINS_MIDNIGHT SEX
IMPACT1 AGE_YEAR;
title "Only Age_Year interactions";
run;

proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP MINS_MIDNIGHT SEX
IMPACT1 AGE_BOD;
title "Only Age_Bod interactions";
run;

proc reg data=work.selectTrain;
model INJ_SEV = VE_FORMS AGE TOW_VEH MOD_YEAR FIRE_EXP BODY_TYP MINS_MIDNIGHT SEX
IMPACT1 YEAR_BOD;
title "Only Year_Bod interactions";
run;



/* calculating MSPR's */
proc plm restore=regModel;
score data=work.selectTest out=new_test predicted;
run;

data new_test; set new_test;
MSE = (INJ_SEV - predicted)**2;
run;

proc means data=new_test;
var MSE;
title "MSE of Test Data";
```

```
run;


proc reg data=work.selectTest;
model INJ_SEV = ;
output out=trainout residual=resid predicted=pred;
store regModel1;
run;

proc plm restore=regModel1;
score data=work.selectTest out=new_test2 predicted;
run;

data new_test2; set new_test2;
MSE = (INJ_SEV - predicted)**2;
run;

proc means data=new_test2;
var MSE;
title "MSE of Test Data with Intercept";
run;


proc plm restore=RegModel3;
score data=work.selectTest out=new_test3 predicted;
run;

data new_test3; set new_test3;
MSE = (INJ_SEV - predicted)**2;
run;

proc means data=new_test3;
var MSE;
title "MSE of Original Model";
run;


/* Random Forests */
proc hpforest data=work.selectTrain seed=12345 scoreporle=oob;
input AGE VE_FORMS MAN_COLL IMPACT1 SCH_BUS BODY_TYP MOD_YEAR TOW_VEH
FIRE_EXP SEX MINS_MIDNIGHT;
target INJ_SEV;
ods output FitStatistis=fitstats VariableImportance=varimp;
run;

data varimp; set varimp;
VarOrder=_n_;
run;
```

```
proc sgplot data=varimp;
scatter x=MSEOOB y=VarOrder / markerchar=Variable
markercharattrs=(size=12);
yaxis reverse;
refline 0 / axis = x LINEATTRS=(pattern=2);
run;


/* forest with 9 vars */
proc hpforest data=work.selectTrain seed=12345 scoreporle=oob;
input VE_FORMS MAN_COLL IMPACT1 SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX;
target INJ_SEV;
ods output FitStatistis=fitstats VariableImportance=varimp;
title1 "9 Variables";
run;

/* forest with 7 vars */
proc hpforest data=work.selectTrain seed=12345 scoreporle=oob;
input VE_FORMS IMPACT1 SCH_BUS BODY_TYP TOW_VEH FIRE_EXP;
target INJ_SEV;
ods output FitStatistis=fitstats VariableImportance=varimp;
title1 "7 Variables";
run;

/* forest with 5 vars */
proc hpforest data=work.selectTrain seed=12345 scoreporle=oob;
input VE_FORMS IMPACT1 BODY_TYP TOW_VEH FIRE_EXP;
target INJ_SEV;
ods output FitStatistis=fitstats VariableImportance=varimp;
title1 "5 Variables";
run;


/* Fit a regression tree */
proc hpsplit data=work.selectTrain seed=123 maxdepth=15 maxbranch=2
plots = zoomedtree(nodes = ('0' '1' '2' 'F' 'G' 'N') depth = 3);
model INJ_SEV = MINS_MIDNIGHT VE_FORMS MAN_COLL SCH_BUS BODY_TYP MOD_YEAR TOW_VEH
FIRE_EXP AGE SEX DRINKING DRUGS HOSPITAL IMPACT1 STATE;
prune costcomplexity (leaves=15);
code file='/home/alyssacable250/EPG194/tree.sas';
/* This saves the tree to a file (need to change the path) */
run;

/* Call the test data and include the tree, this will make predictions on the tree */
data scored; set work.selectTest;
%include '/home/alyssacable250/EPG194/tree.sas';
run;
```

```
/* Now calculate the MSPR as we did in OLS */
data testTree;
set scored;
ASE = (INJ_SEV - P_INJ_SEV)**2;
run;

proc means data = testTree;
var ASE;
run;

proc hpsplit data=work.selectTrain seed=123 maxdepth=15 maxbranch=2;
class MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX IMPACT1;
model INJ_SEV = MINS_MIDNIGHT VE_FORMS MAN_COLL SCH_BUS BODY_TYP MOD_YEAR TOW_VEH
FIRE_EXP AGE SEX DRINKING DRUGS HOSPITAL RACE IMPACT1 STATE;;
code file='/home/alyssacable250/EPG194/tree1.sas';
/* This saves the tree to a file (need to change the path) */
run;

data scoredTrain;
set work.selectTest;
%include '/home/alyssacable250/EPG194/tree1.sas';
run;

/* Now calculate the MSPR as we did in OLS */

data testTree;
set scoredTrain;
ASE = (INJ_SEV - P_INJ_SEV)**2;
run;

proc means data = testTree;
var ASE;
run;

/* Regression Tree */
proc hpsplit data=work.selectTrain seed=12345 maxdepth=15 maxbranch=2;
class MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX IMPACT1;
model INJ_SEV = AGE VE_FORMS ROAD_FNC MAN_COLL IMPACT1
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
MINS_MIDNIGHT;
output out=out2;
run;

/* AGE */
proc glmmod data=work.selectTrain outdesign=GLMDesign outparm=GLMParm NOPRINT;
class DEAD MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX DRINKING DRUGS HOSPITAL
RACE IMPACT1;
model AGE = DEAD VE_FORMS ROAD_FNC MAN_COLL
```

```
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
INJ_SEV MINS_MIDNIGHT IMPACT1;
run;

proc reg data=work.selectTrain plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model AGE = DEAD VE_FORMS MAN_COLL
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
INJ_SEV MINS_MIDNIGHT IMPACT1/ vif collin;
output out=trainout residual=resid predicted=pred;
store regModel;
run;

/* Mins_midnight */
proc glmmod data=work.selectTrain outdesign=GLMDesign outparm=GLMParm NOPRINT;
class DEAD MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX DRINKING DRUGS HOSPITAL
RACE IMPACT1;
model Mins_midnight = Age DEAD VE_FORMS ROAD_FNC MAN_COLL
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX IMPACT1
INJ_SEV;
run;

proc reg data=work.selectTrain plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model Mins_midnight =  AGE DEAD VE_FORMS MAN_COLL IMPACT1
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
INJ_SEV / vif collin;
output out=trainout residual=resid predicted=pred;
store regModel;
run;

/* VE_FORMS */
proc glmmod data=work.selectTrain outdesign=GLMDesign outparm=GLMParm NOPRINT;
class DEAD MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX DRINKING DRUGS HOSPITAL
RACE IMPACT1;
model VE_FORMS = AGE DEAD mins_midnight ROAD_FNC MAN_COLL IMPACT1
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
INJ_SEV;
run;

proc reg data=work.selectTrain plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model VE_FORMS = AGE DEAD mins_midnight MAN_COLL IMPACT1
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
INJ_SEV / vif collin;
output out=trainout residual=resid predicted=pred;
store regModel;
run;

/* MOD_YEAR */
proc glmmod data=work.selectTrain outdesign=GLMDesign outparm=GLMParm NOPRINT;
```

```
class DEAD MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX DRINKING DRUGS HOSPITAL
RACE IMPACT1;
model MOD_YEAR = AGE DEAD VE_FORMS ROAD_FNC MAN_COLL IMPACT1
SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX
INJ_SEV Mins_midnight;
run;

proc reg data=work.selectTrain plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model MOD_YEAR = AGE DEAD VE_FORMS ROAD_FNC MAN_COLL IMPACT1
SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX
INJ_SEV mins_midnight/ vif collin;
output out=trainout residual=resid predicted=pred;
store regModel;
run;

/* INJ_SEV */
proc glmmod data=work.selectTrain outdesign=GLMDesign outparm=GLMParm NOPRINT;
class DEAD ROAD_FNC MAN_COLL SCH_BUS BODY_TYP TOW_VEH FIRE_EXP SEX DRINKING DRUGS
HOSPITAL RACE IMPACT1;
model INJ_SEV = AGE DEAD VE_FORMS ROAD_FNC MAN_COLL IMPACT1
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
MINS_MIDNIGHT;
run;

proc reg data=work.selectTrain plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model INJ_SEV = AGE VE_FORMS MAN_COLL IMPACT1
SCH_BUS BODY_TYP MOD_YEAR TOW_VEH FIRE_EXP SEX
MINS_MIDNIGHT / vif collin;
output out=trainout residual=resid predicted=pred;
store regModel;
run;
```