Modeling Injuries in Car Accidents

MTH 5387 - Applied Regression Analysis

Brady Lamson

Fall 2024

CU Denver

Introduction

Background and Motivation

Cars dominate our lives here in America. They're extremely important to how we live our lives. They also result in far more deaths and injuries than could ever possibly be acceptable for our primary form of transporation in the modern age. According to the CDC, "deaths from crashes in 2022 resulted in over $470 billion in total costs - including medical costs and cost estimates for lives lost" (Source). On the upside, things have gotten better over time. According to data visualizations created by the Insurance Institute for Highway Safety (IIHS), the number of motor vehicle crash deaths per 100,000 people since 1975 has shown a fairly steady decline (Source). Things are getting better, but that doesn't mean we shouldn't keep improving. Car deaths also aren't the only measure of driving risk, injuries are also an important thing to look at.

When it comes to this topic in Colorado, this state isn't doing too bad. We tend to rank around the middle for deaths and injuries in the country. We could always be doing better though. From 2021-2023 Colorado saw around 100,000 car accidents per year. Of those, a quarter resulted in some degree of injury and around 700 people were killed each year. (Colorado Department of Transportation)

With this project I'm taking a step back from the individual accident level and looking at counties. Are there any variables we can use that could help us understand what results in more injuries in the state? Do socioeconomic factors play into this at all? Do poorer counties see more injuries? How about counties where commute times are longer? Any additional understanding we can glean about what's putting us at risk behind the wheel could potentially save lives and prevent life changing injuries.

Data Tables and Sources

For the work done here I am pulling data from 4 separate sources. The main tables I'm using come from the Colorado Department of Transportation and contain individual records of every single motor accident in the state from 2021 to 2023. To account for heavily populated coun-

ties having inflated injury values I pull in population data from the Colorado Department of Local Affairs. For additional demographic data I pull from two sources. One source contains information about the median household income for each county and the other has information about the average commute time per county. The first comes from the National Institute on Minority Health and Health Disparities and the latter from Opendatasoft. Both are simply cleaned tables that were originally queried from the Census Bureau api. I used those sources as I was struggling to get the correct information from the api myself. All relevant data sources and files are included in the github repository in the appendix. As for the type of project I believe this would count as an observational study as I'm looking at car accidents that have already occured.

It is worth noting that the data I am using for average commute time is out of date. The data I'm using is from 2017 and I'm applying it to data from 2021-2023. Any attempt I made to get up to date information on this variable through the census api resulted in only one county having any information. Therefore it isn't ideal but I opted to include this information anyway despite it being slightly out of date. If this was a more formal study this wouldn't be acceptable, but I feel this should be okay for a class project.

Results

Data Exploration

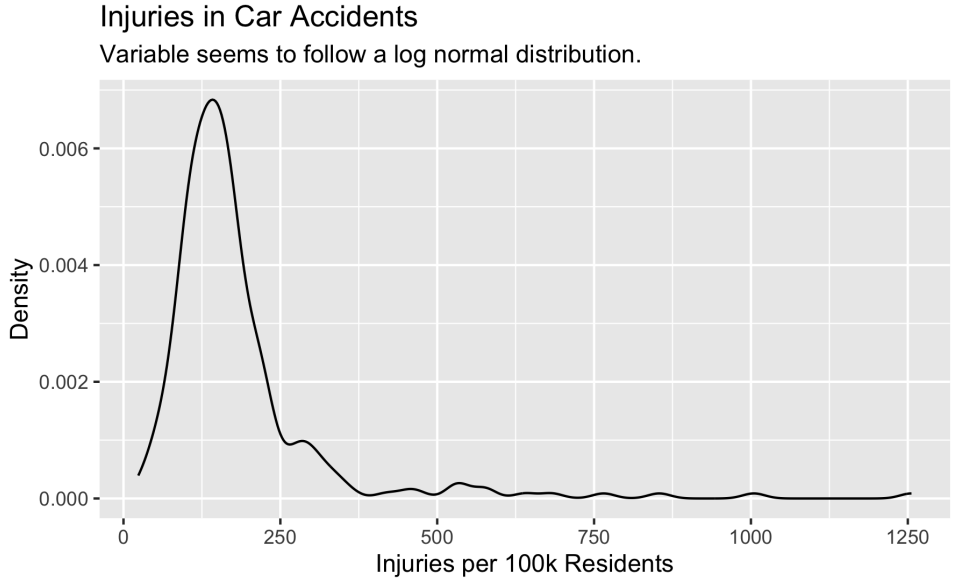
The data taken from the Colorado Department of Transportation has one row per accident, containing 295445 total observations. Heavy aggregation and preprocessing was done to prepare this data for modeling. In particular, these rows ended up being grouped by county and by season bringing the number of observations down to 256. That is 64 counties and 1 row for each season.

Response Variable: Injuries

To start we will look at our response variable which represents the average number of injuries per 100k residents per year.

Variable	Data Type	Description
County	string	Name of Colorado county
Season	factor	Spring, Summer, Fall, Winter
Deaths	continuous numeric	# of deaths per 100k residents
Injuries	continuous numeric	# of injuries per 100k residents
Bad weather accidents	continuous numeric	# of accidents in poor weather per 100k residents
Median household income	continuous numeric	Median household income for county residents
Mean commuting time	continuous numeric	Mean commuting time for county residents (minutes)

Table 1: Variables Used in the Analysis



It seems that, on average, a county in Colorado sees approximately 150 injuries in car accidents on any given season per year per 100 thousand residents. We can also see that the distribution of injuries appears log normal. We see most values hovering between 100 and 200 with a very long right tail. As we need our response variable to be normally distributed this immediately puts a log transformation into consideration. It is worth noting that a log transformation of the response variable does complicate model interpretation slightly and so should not be done without reason.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Untransformed	0	116.09	150.63	183.73	197.19	1255.83
Transformed	3.159	4.754	5.015	5.048	5.284	7.136

Table 2: Numerical summary of Injuries, with and without log transformation

Instead of including a plot showing the transformed variable, I feel this numerical summary will suffice. The more extreme values have been pulled inline with the rest of our data which is shown by the mean not being as far from the median. The resulting density plot also does appear far more normal though is not included here. We will look at more evidence in the structural section of this report later.

Predictor Variable Candidates

Predictor	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Deaths	0	2.10	4.19	10.71	9.62	179.40
Bad Weather Accidents	0	28.33	51.37	93.43	97.23	904.07
Income	34578	56303	65976	70543	85228	139010
Commute Time	11.80	17.23	20.05	21.17	23.27	42.40

Table 3: Numerical summary of predictor candidates.

As for season, that one is set up such that each season shows up 64 times in the dataset. So I opted to not include a summary of that.

For this model we have 5 candidates predictors. When examining bivariate relationships between each predictor and the response we can already begin to guess which of these candidates will be included in the final model. Both deaths and bad weather accidents seem to have a positive linear relationship with injuries. However, neither income or commute time seem to have much of a linear relationship at all. All of these bivariate plots also included season as a dimension via color. What I saw did not seem to indicate any noticable groups based on season. I used this information to decide that I was probably going to stick to a parallel lines model to avoid additional complexity.

As for season and injury, at first glance there didn't seem to be much of a relationship there as well. However, there are some differences. From most injuries to least we see summer at the top, followed by fall, winter and lastly is spring with the least number injuries. Whether these differences would be substantial enough to warrant inclusion didn't seem obvious to me.

Variable Selection

I performed the variable selection process twice. First with the untransformed response and second with the transformed response. My preference in variable selection was for as simple a model as possible as I'm primarily interested in interpretation. Therefore any complexity I can avoid is worth investigating.

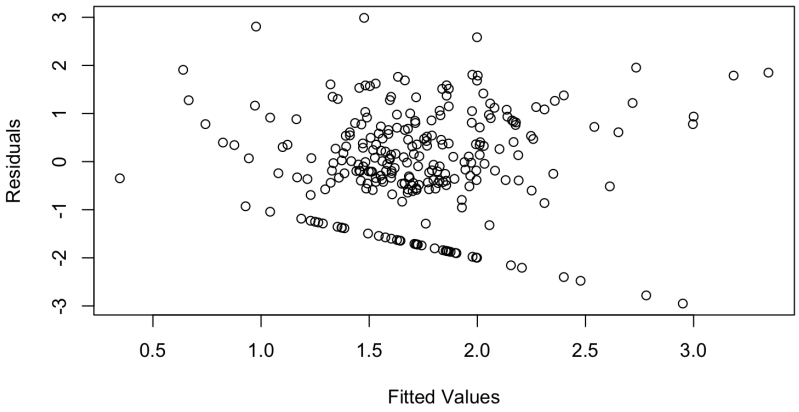
For the first run I looked at two different methods. First I performed a stepwise selection and second was a forward selection using the BIC statistic. I used these two methods as I felt they would result in a different number of predictors and was curious to see how different these models would look. I was mostly interested in the forward selection with BIC however as I only wanted to add predictors if it was necessary. Both models ended up being the exact same. They included bad weather accidents, deaths, season and mean commuting time. So, all the predictors but median income. The model had an adjusted r squared of approximately 0.65.

For the second process using the log transformed response I only used the forward selection with the BIC statistic. My plan was to later compare it against the full model to compare their performance. The model we got from this used season, deaths and bad weather accidents as the predictors. It's adjusted r squared was far worse than the untransformed model however, at 0.48. I don't see this as an enormous downside however as this transformation resolves some structural issues we'll discuss soon.

Structural Checks and Influential Observations

Brief Tangent: Deaths as the response

This is where things get interesting. I want to take a step away from the injuries model for a moment and return to the problem proposed at the start of this project. Originally, the deaths variable was my response. I went through a very similar methodology during the variable selection process and this is the stage where the death model ceased being a viable option. The model I got had the log of injuries and mean commuting time as predictors and the log of deaths + 1 as a response. To be frank, the log transformations in this model were desperate attempts to fix some severe issues that I was unable to figure out at first.



What we see in this residual plot is a line of values going down very consistently. Upon investigation, every single one of these points was an observation with zero deaths. About 16% of the dataset was 0 and that resulted in a model with this trend both before and after a log transformation. 0 is a very reasonable value for the response as well so there is no justifying removal of these rows from the data. My model using deaths as a response violated some key assumptions about the errors and the results here were disasterous. This is what motivated the pivot to using injuries as the response variable. I wanted to be sure to include this aside in the report as it was the most interesting part of the process and helped inform my modeling strategy with injuries later.

What I will say is the residual plot looks fairly well behaved aside from that structural issue. I don't think linear regression is entirely useless here. One possible fix that is outside the scope of this class is to use two models. The first is a logistic regression model that assesses if the deaths response is nonzero and only then do we use a linear regression. That allows us to use the full dataset and properly handle the zeros.

Back to Injuries

Returning to injuries, our structural checks go a lot more smoothly. I performed structural checks on both the transformed and untransformed response models. The residual plot of the untransformed model looked okay but there seemed to be some patterns in how the residuals behaved as the fitted values changed. Looking at the qq plot was also concerning as I noticed

some very extreme standardized residuals at the tails. There was also noticable amounts of curving throughout the center.

As for the transformed model, things appeared to behave a lot better. Though I can't confidently claim that the errors have constant variance, the transformed model makes me far less concerned. There does seems to be a narrowing in variance as the fitted values grow more extreme however. The qq plot looks a lot better as well. There aren't any particuarly extreme values anymore and, though the plot doesn't follow the line exactly, it does follow it a lot better. This was enough evidence for me to commit to the transformed model at this point. From this point forward I completely abandon the untransformed model.

Influential Observations

County	Season	Population	Injuries per 100k/year	Injuries per year
Mineral	Fall	929	1255.8	11.7
Mineral	Spring	929	1004.7	9.3
Hinsdale	Spring	776	43.0	0.67

Table 4: Some examples of small counties.

Checking for outliers gives us a few points worth considering. Index 225, 53 and 56. Using the outlier test function from the api2lm package pointed me towards index 56. This index is for Crowley county in the winter. Crowley county is a county with around 5600 people. Of note in this observation is the number of injuries at 23.55, which is both the minimum value for this variable and far below the first quantile of 116. I recommend referring back to table 2 if a refresher on the summary for injuries is needed.

Looking at leverage points brings up index 161 and 162. None of the leverage values exceed 0.30 so they don't seem to be concerning but they are still worth investigating. Looking at these rows gives us two seasons from Mineral county. Mineral county has an extremely tiny population of around 929 people. Of note in this row is the number of injuries. One row is the max for injuries at 1255, far exceeding the 3rd quantile of 197. The other row isn't far behind with a value of 1004. The thing to point out is the injuries variable is scaled by the population. Injuries really represents injuries per 100k people. The unscaled number of injuries is very small, one row has 9 injuries and the other has 12. These values just get scaled extremely high

by the small population.

This happens again when we look at influential points. Mineral county comes up again but now we also get Hinsdale county which is an even smaller county with 776 people. Hinsdale in the spring has an injuries value of 43, but when we unscale that value we see this county gets less than 1 injury per year in the spring. There are a couple big takeaways from this information. First is that none of these points are so drastic that we need to remove them or handle them in any way. However these do point out a flaw in my modeling strategy. My handling for population didn't account for such small counties. It's also possible these smaller counties would be better served with a totally different type of model. Since we see such small numbers of injuries something more appropriate for discrete counts would likely perform better here.

Model Inference and Interpretation

For this portion I wanted to compare my fairly simple model to the complete model and see how much performance I lose if any. To be more specific, I chose to compare the model with every predictor against a model only including deaths, season and bad weather accidents. I ran these both through a 100 fold cross validation. I chose 100 folds arbitrarily because I've never used that number before. Comparing the RMSE of the two models, they're both nearly identical with rounded values of 0.34. The MAE is similar, with the reduced model having a slightly higher value of 0.30 against 0.29 for the complete model.

These results lead me to believe that the rest of the predictiors add very little to the effectiveness of the model. I was willing to accept a decent drop in performance for the sake of interpretation but the results indicate that no such sacrifice is necessary.

Regression Coefficients

Bounds	Spring	Summer	Winter	Deaths	Bad Weather Accidents
Upper	-0.04	0.32	-0.21	0.010	0.003
Lower	-0.31	0.04	-0.51	0.001	0.002

Table 5: 95% Confidence intervals for regression coefficients.

	Intercept	Spring	Summer	Winter	Deaths	Weather
Untranslated	4.83	-0.18	0.18	-0.36	0.04	0.003
Translated	125.87	0.84	1.20	0.70	1.004	1.003

Table 6: Regression coefficients before and after undoing log transformation.

To work with these regression coefficients we need to recall that we used a log transformation on our response variable. This changes the interpretation of the regression coefficients quite a bit. You can't directly interpret -0.18 for spring as an example. What we need to do is take each coefficient, β_i, and undo the log transformation by putting it in the exponent of e like e^{β_i}. From here the values give us a proportional increase or decrease of the response depending on if they're greater than or less than 1. I have the coefficients translated but I will be leaving the confidence intervals as is. As an example, if summer had a coefficient of 1.5 we would expect to see an increase in injuries of 50% compared to the fall. If that coefficient was 0.5 we would expect to see a (1-.5) or 50% decrease in injuries compared to the fall.

Let us start with the intercept. We expect a typical Colorado county on any given year in the fall with 0 deaths and 0 bad weather accidents to see around 125.87 injuries per 100k people. Next up is interpreting season. I find the interpretation of this one to be a little awkward. I'm unsure if we want to compare two counties with the same X values or compare the same county in two seasons. Let's assume the former despite the wording being a little awkward. Let there be two Colorado counties with the same number of deaths and bad weather accidents but one is in the fall and one is in the spring. We expect the county in the spring to see $1-0.84=16\%$ less injuries per 100k residents than the county in the fall. If that spring county was instead in the summer, we would then expect to see a 20% increase in the number of injuries per 100k residents. Lastly, for winter, we would expect a $1-0.7=30\%$ decrease in the number of injuries per 100k residents compared to the fall. Deaths and bad weather accidents both have a nearly identical coefficient.

Let there be two Colorado counties in the same season and with the same number of bad weather accidents, however one has one additional death per 100k residents. We expect the county with the additional death to have 0.4% more injuries per 100k residents. We can interpret bad weather accidents in the same vein just by flipping the wording, though we see 0.3% more injuries per 100k residents instead.

Looking at the confidence intervals of our coefficients, it's a pleasant surprise to see that none of these intervals contain 0. This means that, at least at the 95% confidence level, that 0 isn't a reasonable value for them. This is reflected in the effect plots for all of the regression coefficients as well. Deaths and bad weather accidents both show a positive relationship with injuries given the other regressors are held at their typical values. Each season also shows a meaningful difference from the other, with none of the confidence intervals seeming to overlap between the 4. All of this, combined with our other results compared against the complete model, indicate that we have a good selection of regressors.

Collinearity

As for collinearity, none of the predictors had VIF values that indicated an issue. The highest generalized VIF score was 1.48 for bad weather accidents. Therefore, we don't seem to have much infighting between our regressors and we can be fairly confident in the interpretation of the values that have been put forth here given the model being discussed and the data that it was fit to.

Conclusion

The conclusions I can draw from this project are somewhat interesting. We see that summer is the season with the largest coefficient which is surprising at first glance. We expect it to have the most injuries per 100k residents. I assumed it would be winter which is actually the second smallest. I believe this is due to the frequency of travel in the summer and the lackthereof in the winter. As my analysis does not in any way examine the number of accidents I can draw no deeper conclusions. I believe it would be intersting to examine which season sees the highest proportion of accidents that result in injury.

What I find interesting is a variable that I did not include in the model. I was surprised that there was seemingly no linear relationship between commute times and injuries. It's worth noting that this doesn't necessarily mean there isn't a relationship between commute time and accidents in general, but there not being one for injuries was shocking. I was prepared to

advocate for systems like remote work which reduce traffic and injury risk but can draw no such conclusions here. However, my commute time data is also out of date so it's possible this has changed. As such, it's unwise to read too deeply into this variable one way or the other.

What I will say is the main dataset I used for this project is fascinating. I had to do a lot of aggregation to make it work for this project, but I believe there are so many insights to dig into here if we go further. This dataset contains so many variables related to the specific location of the accident. Using that information we could see which streets or intersections have the highest proportions of accidents that result in injury or death. We could see how that varies month to month or over seasons and use that information to inform variable traffic rules for those specific areas. We could find problematic areas that could benefit from modernized infrastructure. Accidents will always happen, but reducing the number of them that result in injury or death is always valuable and can always be improved on.

Though it's difficult for me to glean too much from my model specifically, I see so much potential in this dataset now. This is why it's important that this type of information is recorded so diligantly. It's not hard to imagine a world where our government doesn't maintain this data at all and we would be so much worse off without it. Infact I was unable to use two fascinating variables related to suspected alcohol and marijuana use because only one or two counties even collects that information. We should continue to fund and advocate for the collection of this data by public institutions that benefit all of us. I worry so much data collection is in the hands of private businesses now and that keeps data from truly working for the greater public. I find this extremely concerning with the upcoming administrations plans to gut and destroy many government agencies that have been collecting valuable data for decades. There's so much good this kind of data does behind the scenes that we take for granted. We need to protect it.