Doing Without Speed Limits: Report on Montana Highway Patrol

Under the Guidance of
Dr. Niamh Cahill
Assistant Professor
Department of Mathematics and Statistics
National University of Ireland, Maynooth

Submitted By
Srishti Kakkar
MSc. Data Science and Analytics
19250263
Batch 2019-20

Submitted in Partial Fulfilment of the Requirement for the Degree of Master of Science

CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF TABLES

CHAPTER 2

CHAPTER 4

CHAPTER 5

STATEMENT OF ORIGINALITY

I have read and understood the Department policy on plagiarism and I certify that work demonstrated in this thesis titled is my own and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education.

I confirm that:

1. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.
2. A source of all figures or tables has been provided in the document which is not my work.

Srishti.

Srishti Kakkar
19250263
MSc. Data Science & Analytics
National University of Ireland, Maynooth

ACKNOWLEDGEMENT

All the way through, working on this thesis I received countless support and assistance. I would like to recognise and thank my supervisor, **Dr. Niamh Cahill** for all her efforts and continuous guidance. She helped me throughout the project and carry out the analysis for the same. She guided me by providing valuable feedback and encouraged me to deliver the work done in the thesis. She has been very responsive and resolved every query that I had very quickly and suggesting new ways to target the problems and helped to become better in my skills than I was on day one of the thesis.

I would like to give special thanks to **Dr. Catherine Hurley** for the help she has provided me through the modules of R for Data Analytics and Machine Learning.

At last, I would like to express my gratitude towards National University of Ireland, Maynooth for providing me with all the necessary facilities which facilitated this research work.

Srishti

Srishti Kakkar

ABSTRACT

Stanford Open Policing Project is an initiative in collaboration with various Police Departments and an analytical approach used in order to understand the relationship between the population of various states of U.S.A and police officers. The Stanford Open Policing has produced extremely important resources that can help in detecting various issues existing in the country like racial profiling, increasing fatalities, and gender biasing and illegal activities. We present an analysis and modelling techniques, an attempt to deal with the over-speeding problem that Montana state is facing since the government passed the rule that abolished the day-time speeding limit. The dataset contains various attributes and through the analysis performed we find the best set of variables that associate with the state's speeding violation problem. We look for modelling strategies to fit data variations and are able to achieve the fit capturing most of the variation exhibited by our dependent variable (number of speeding stops for each outcome group: warning, citation and arrest) . The dataset used for analysis is over-dispersed and the results of Poisson Model will lead to make incorrect inferences. Alternative method of dealing with over-dispersed count data like negative binomial is used. After comparing each fitted model against their AICs, negative binomial model is found to best fit the data. Also, the full model with negative binomial fit displayed better results than the pruned negative binomial models.

Keywords: Montana, speeding limit, Negative Binomial, Poisson Regression, over-dispersed, AIC.

.

CHAPTER 1

INTRODUCTION

The Stanford Open Policing Project (https://openpolicing.stanford.edu/) is collaboration between the Stanford Computational Journalism Lab and the Stanford Computational Policy Lab, and it has been the lead entity in dealing informally with the Montana Highway Patrol (MHP) for the past two years. This contact is mainly the result of general discussions related to the traffic stop data collection and statistical analysis and how this data is applied to the various issues like racial profiling (biased police), gender biasing and road rule violations. The project was formalized in late 2015 as a research contract between MHP and Stanford Computational Policy Lab. The development of this resource has been marked by challenges. Many counties don't collect demographic details on the drivers that the police are pulling over. The ones that collect the information may not always release the details. Even when these counties provide information, the way they collect and process data varies widely, creating difficulties for standardization of information.

Immediately following the execution of the contract and the approval through MHP, the Open Policing Project Team began working closely with MHP project advisor and data managers on issues related to collection and analysis of traffic stop data (Gaffney, et al., 2013). Initial issues arose with the format of the data files, the coding and the labelling of data and the way data is transmitted. These problems involved tremendous work to translate raw data files into a standard format, and to merge a variety of different files into a format that facilitates research. Thus, post transformations, the data were made available for analysis; extract significant disparities in policing to attempt to understand the relationship between police tactics and public behaviour. Montana was the first and only state to abolish the numerical daytime speed limit from its interstate highways.

The Government adopted the "Simple Law" of daytime driving "fair and prudent" (King & Sunstein, 1999). The Policy sparked a vigorous debate about its implications, the relationship between practice and legislation, and the preference of rules and principles in key conduct laws. With an area of 147,047 square miles, Montana remains as the fourth largest province. Moreover, the population of Montana is approximately 870,000, making it the third most sparsely inhabited nation with an average of just six residents per square mile (Mueller, 1996). Though, the consequences of this change were not very significant in terms of driving habits. The percentage of Motorists driving over 80 mph rose from 2 per cent to 5 per cent, the overall average speed rose negligibly from 72 to 74 mph within the first three years of the rule. But traveling Montana's vast expanse across its interstate highways can be risky. The state has a highway of 88,000 miles. There is no separate state police force and only 212 Highway Patrol officers are monitoring the roads (Cornwell, 1996). Montana had a fatality rate of 2.3 per 100 million vehicles miles which is slightly over the national average of 1.8 deaths per 100 million miles of cars. In fact, the deadliest stretch of road in the U.S.A and two of the three most dangerous highway regions of the country are found in Montana, with fatality figures being 6.206 and 4.031 per 100 million miles, respectively.

In short, Montana's wide-open spaces afford enough reach, and threat to the public of the motoring public (Kowal, 1995). Another interesting fact about Montana remains the speed magnet attracting more and more tourists. Tourists are a key exception to the relentless driving actions of motorists. The removal of the state speed limit has turned Montana into a national level Magnet. Montana attracts an estimated 7 million tourists a year. Most of them came to be known, or more aptly mocked, as "pace visitors". Road safety statistics of the Montana Highway Patrol tend to support perceptions of tourists as the most dangerous drivers. From 1996 onwards the number of out-of-state vehicles involved in fatal accidents increased by 38%, the number of road fatalities involving out-of-state accidents increased by 38% and vehicles increased by 69% in the same one-year period (Brooke, 1995). The statistics point to the same conclusion that the implementation of the new rule- “Simple Law” brought some serious tensions into the state and invited unsafe drivers from neighbouring states bringing a significant rise in fatalities.

This report is thus based on the analysis performed using the data provided through the Open Policing Project which contain traffic stop data from the period January 2009 to through December 2016. The dataset contains more than 0.7 million stop records and the demographics related with it. Through the study of relevant facts and statistics, the report proposes to detect the factors influencing the speeding violation for the state based on the attributes like subject race, gender, type of the vehicle and time. It will also break down the analysis to look at the 3 groups for the Outcome of each stop made separately, to explore how drivers' demographic factors influence the decision-making of the police on speeding stops for the arrest-made, ticket issued

or left with a warning. As the dataset contains the geographic coordinates for the stops conducted, it is expected to provide more useful insights to the results. The following report is structured as follows. Section 2 explains the data and the exploratory analysis performed to gain better inferences. Section 3 describes the various methods used. Section 4 displays the results. Section 5 contains findings and discussions.

CHAPTER 2

DATA

Montana is the fourth largest city area wise with a population of 1,068,778, the 8th least populous and third least densely populated of the 50 states in USA. The MHP traffic stop data was obtained from Stanford Open Policing Project website with the date range from 01/01/2009 to 31/12/2016. Most of the data was followed by geographic coordinates and time information. There were 825,111 traffic stop reports, including 545998 speeding stops. There were many types of violations for which the Montana state population were being stopped; therefore, the violations falling in the same category were further collapsed into 1 category. For instance, "Speed over legal", "Truck speed", "Reckless driving", "Careless driving", "Speed over basic rule" were collapsed under the category of "Speeding". The result or the outcome of a traffic stop can be a warning, an arrest or citation that is a ticket received. Finally, to classify factors that affect speeding output through regression analysis, 545998 speeding traffic stop records with established driver gender (male or female), Vehicle Type (Sedan, Pickup, Sport Utility etc.), Time and geographic coordinates will be used in this research. A summary of the total speeding stops can be seen in Table 1.

Table 1: Summary of attributes used in analysis & modelling

Variable	Definition	Values

Outcome	The outcome of a speeding traffic stop	619 for arrest;
		1655 for warning;
		2325 for citation.

Subject Sex	Gender of the driver	1800 for female driver; 2799 for male driver.

Hour	**Hour of day when speeding stop occurred**	**0,7038; 1,4966; 2,2903; 3,2324,86; 5,726; 6,3615; 7,22817; 8,35166; 9,33875; 10,34839; 11,28615; 12,24137; 13,32406; 14,46606; 15,54636; 16,46812; 17,37771; 18,36588; 19, 26296; 20,20543; 21,17273; 22,15549; 23,12499**

Vehicle Type	The type of the vehicle stopped	158991 for Sedan;
		126932 for Pickup;
		110004 for Sport Utility;
		23431 for Van;
		16453 for Coupe.

District	The MHP is divided into 8 districts	1,54289; 2,56335; 3,60440; 4,51776; 5,87328; 6,69142; 7,61126; 8,105562

2.1 EXPLANATORY AND OUTCOME VARAIBLES

2.1.1 Traffic Stop Analysis by Statewide Racial Composition: For the data, analysis was implemented using disparity index analysis to examine whether any groups may be potentially overrepresented in stops by Montana Highway Patrol officers. The disparity index analysis compares stops of a particular group to their proportion of the driving population to determine if potential overrepresentation is occurring. It is important to note that this analysis has limitations, as a finding of overrepresentation cannot indicate whether potential differences are due to bias (as there are several potential explanations for overrepresentation that cannot be examined in disparity index analysis) (Gaffney, et al., 2013). More importantly, disparity index results are subject to error when less than 15% of the population is people of colour, which can lead to a false overrepresentation. As people of colour in Montana comprise less than 15% of the total population, these results should be treated with extreme caution.

Formula for Disparity Index used: - $\frac{Stop\,\%\,for\,each\,group}{Population\,\%}$

Despite limitations, disparity index analysis is performed to analyse potential overrepresentation in order to help with MHP efforts to examine this important issue. To interpret disparity index, value of disparity index for each group greater than 1 indicates over-representation, a value less than 1 indicates under-representation and a value equal to 1 indicates no disparity for the group of population. The disparity index analysis revealed potential overrepresentation for some groups in state-wide stops.

Table 2: Montana State-wide Percentage of Stops by Race/Ethnicity and Demographics

Subject Race	Number of stops	Percentage stops	Population %
asian/pacific islander	6700	0.8121094	6.6
black	8805	1.0672572	0.6
hispanic	16055	1.9460323	3.5
unknown	2633	0.3191469	0.1
white	752035	91.1544317	89.0

Table 3: Disparity Index for Montana for each Race group

Subject Race	Disparity Index
White	1.02
Black	1.76
Hispanic	0.55
Asian/Pacific islander	0.12
Other	5.8
Unknown	3.1

The disparity indices suggest that Hispanic and Asian drivers were underrepresented in stops, Black and White drivers may be overrepresented in stops (see Table 2 and Table 3). For example, a total of 7, 52,035 (2009-2016) White drivers were stopped accounting for 91.0% of all traffic stops, whereas Whites comprise 89% of the driving population in Montana. The disparity index for Whites is therefore 1.02 (91/89). A disparity index of 1.02 means that Whites were stopped at nearly the rate to be expected based on their proportion of the driving-age population from the 2010 Census. Conversely, Hispanic drivers account for 2% of all traffic stops, but 3.5% of the population. The disparity index for Hispanics is therefore 0.5 signifying that Hispanics were stopped at a rate less than expected based on their proportion of the driving-age population. Similarly, Asians with a disparity index of 0.12 were stopped at a rate less than expected. Black drivers, however, are potentially stopped at a rate greater than expected with a disparity index of 1.76. The 5% threshold, used in past analyses to determine whether there was evidence of systemic bias, is not exceeded suggesting minimal differences in stops by race/ethnicity. But due to the limitations of the method performed, in the final report there will be an attempt to investigate different methods to conclude the fact that there do not exists biasing in Montana.

2.1.2 Driver Gender: Young drivers are carefree and can be the ones who indulge in more rule-breaking activities. Similarly, male drivers are more likely to participate in "risky driving actions" than female drivers, and Lange, et al. (2005) has also recorded this in 65 mph speed zones. Since MHP provides access to the information available on each driver's license, as well as they performed direct observation and conversation, it is likely that taking gender into account will give more accurate results. Looking at the Figure 1 of licensed drivers by gender, the number of licensed females is similar to number of the licensed male drivers. But when we compare them with traffic stop data, the number of male driver stops are notably higher than those of female driver stops (see table 4 below).

Figure 1: Number of licensed drivers in Montana in 2016, by gender

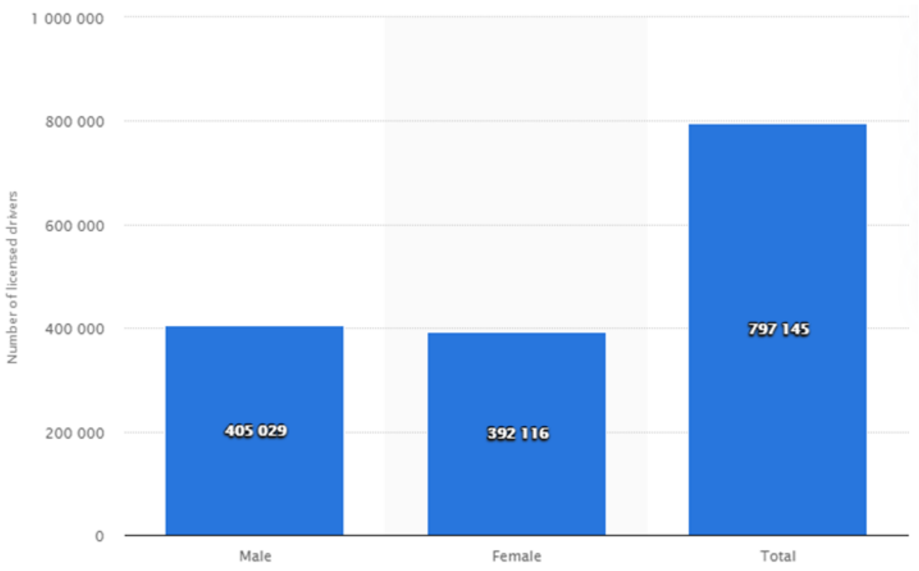
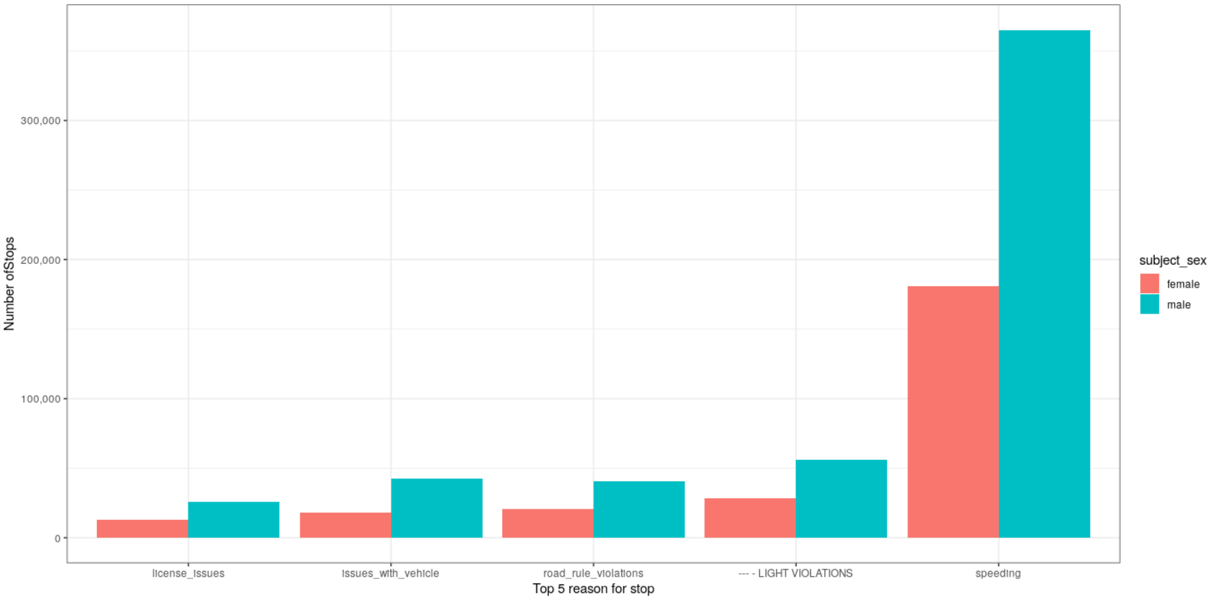


Table 4: Number of Traffic Stops by Gender

Driver Gender	Traffic Stops
Male	556934
Female	268065

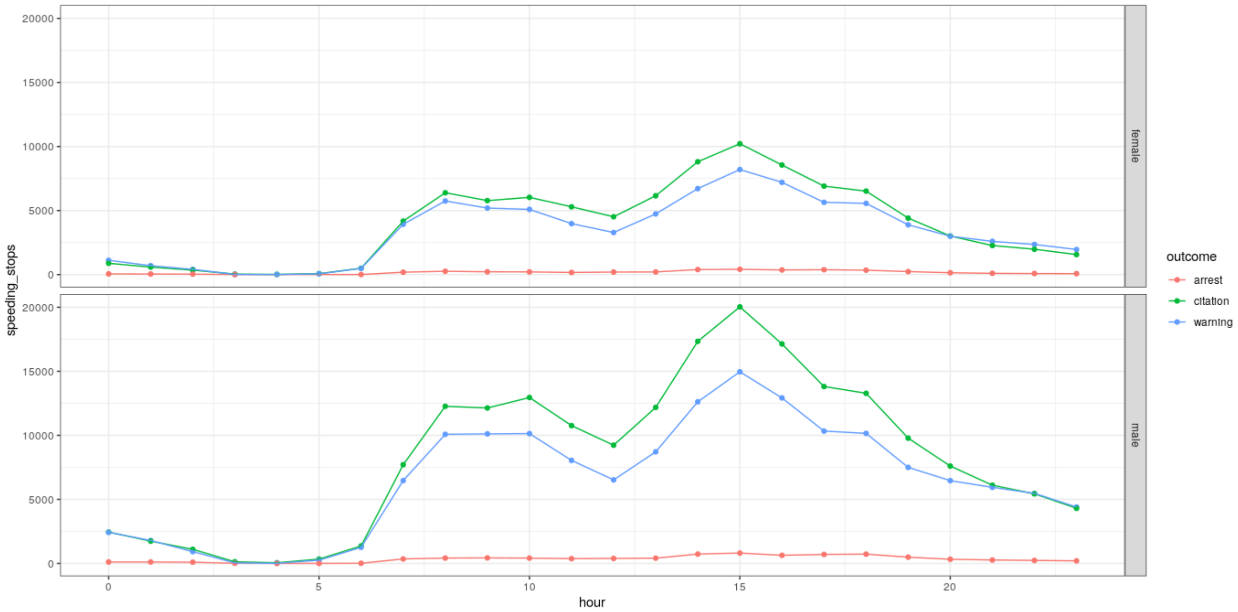
2.1.3 Reason for Stop: Traffic stops can be classified into two types of violation: non-moving offenses, such as parking in restricted areas and broken headlights, and moving offenses, such as driving under influence and speeding. These violations may lead to very different outcomes and penalties for their risks. For example, speeding may be more likely to get a ticket than a broken headlight. Hence is also important to evaluate traffic stop data separately by form of violation in law enforcement analysis in order to eliminate potential confounding effects (Sharma & Liu, 2019). Usually, most of the current studies evaluate all traffic stop data together and very few aims at analysing the two forms separately. Speeding is one of the most dangerous and frequent violations that lead to traffic accidents. In 2015, 27% of road deaths were related to speeding accidents in the USA and the number of fatalities due to over-speeding increased from 2014 to 2016 (Administration, 2008). Reason for Stop variable records various violations committed by the Montana population. On exploring the driver gender against reason for stop (figure 2 below), it is noticed that men were more likely to be stopped for speeding by police officers. Further analysis of the speeding stops of the Montana Highway Patrol from 2009 to 2016 indicated that for minority drivers the speed stops were less likely to occur compared to white drivers.

Figure 2: Reason for Stop by Gender



2.1.4 Outcome and Hour of the day: Several useful observations of importance can be found through the analysis on these two measures shown in figure 3. First, instead of most speeding stops performed by police occurring at night, they tend to be at an unusual hour of the day that is, 3 p.m.; one way to look at the possible reasons can be further analysing the location along with the hour. Also, as Montana faces over-speeding issue since the implementation of “simple rule” and abolishing of day-time speeding limit, the number of arrests made is negligible as compared to warning or citation. If Montana is among the top state in terms increasing fatalities and fatal accidents caused due to over-speeding, then why aren't the MHP authorities taking any legal actions against the population committing the violation?

Figure 3: Number of speeding stops by Hour, Driver Gender and Outcome



CHAPTER 3

METHODOLIGIES

The summaries given so far describe number of speeding stops for the whole city of Montana. But suppose that the police make more stops in high-crime areas and the outcome of speeding stop (warning, citation and arrest) varies but treat both the genders equally within any locality. Then the citywide ratios could show significant differences between genders even if stops were determined entirely by location rather than by gender. To separate these two kinds of predictors, we performed multilevel analyses using the city's 8 districts. Allowing district-level effects is consistent with theories of policing such as "broken windows" that emphasize local, neighbourhood-level strategies (Sousa & Kelling, 2006). We divided our analysis further into three groups based on the outcome of speeding stops as it would be useful to see the effect of each outcome group on our response (speeding stops) based on the driver gender, district and hour.

3.1 GENERALIZED LINEAR MODEL (GLM)

The central idea of multiple regression models is that the residuals have a normal distribution (Cohen, et al., 2003). Typically, the residuals obey the distribution of the output variable, which is also not usually distributed or even symmetrical for the count variables (Figure 4). In addition, residual variance also decreases as the variable predictor decreases. Therefore, the use of traditional regression methods with count outcome variables can lead to the bias of the parameter, the standard error and confidence interval. In the end, this may lead an educational researcher to make incorrect inferences and bad decisions. Alternatively, regression models which embody the count nature of the outcome variable (and the subsequent nature of the model residuals) produce more accurate results. By making use of GLM, models can explicitly manage the distribution of count variables. The GLM (Lee & Nelder, 1998) is a technique designed to manage regression models with a number of output variable forms. Generalized Linear Model's require two components: a proper distribution specification for residuals and a function to relate the outcome and a linear combination of the predictor variables.

3.1.1 LINEAR REGRESSION (GAUSSIAN FAMILY)

Linear regression belongs to the Gaussian family. The relation function i.e. the link function, g is the identity, and the density f is the normal distribution. It is the simplest example of a GLM, but it has several applications and many advantages over other groups. Basically, it is faster and requires more stable computation. For Gaussian models the dependency between the response y and the covariate vector x as a linear function: $\acute{y}=x^T \beta+\beta0$. The model is built through solving the least square problem, which is equivalent to maximizing the probability for the Gaussian family.

$$\max_{\beta,\beta 0} - \frac{1}{2N} \sum\nolimits_{i=1}^{N} (xi^T \beta + \beta 0 - yi)^2 - \lambda \left(\alpha \left|| \beta |\right|_1 + \frac{1}{2(1-\alpha)} \right) \left|| \beta |\right|^2 2$$

Equation (1)

We modeled the number of speeding stops, y, using linear regression as follows:

$E(y) = \mu = g^{-1}(\beta_{0\,+}\ \beta_1$ *subject_sexmale+ β_2 *hour+ β_3 *districtII $_+$ β_4 *districtIII $_+$ β_5 *districtIV + β_6 *districtV+ β_7 *districtVI + β_8 *districtVII + β_9 *districtVIII)

$= g^{-1}(X\beta)$

Where, E(y) is the expected value of y that is the number of speeding stops for each outcome group; Xβ is the linear predictor, a linear combination of unknown parameters β; g is the link function and intercept i.e. β_0 is the #speeding stops for female driver in district I (baseline).

3.1.2 POISSON REGRESSION

The Poisson distribution is the most general type of distribution for count variables, as seen by an example in Figure 4 which is produced by generating random number series. The Poisson distribution is used because it is a matrix of probability meant for non-negative integers. It is defined by a single parameter, λ, which estimates the distribution for mean and variance, thus fully regulating the shape of the distribution. When λ is close to 0, the distribution becomes very positively skewed, but as the λ increases, the distribution becomes less distorted and becomes closer to the normal distribution (see Figure 5).

Figure 4: Plot of count variable x distribution (Poisson)

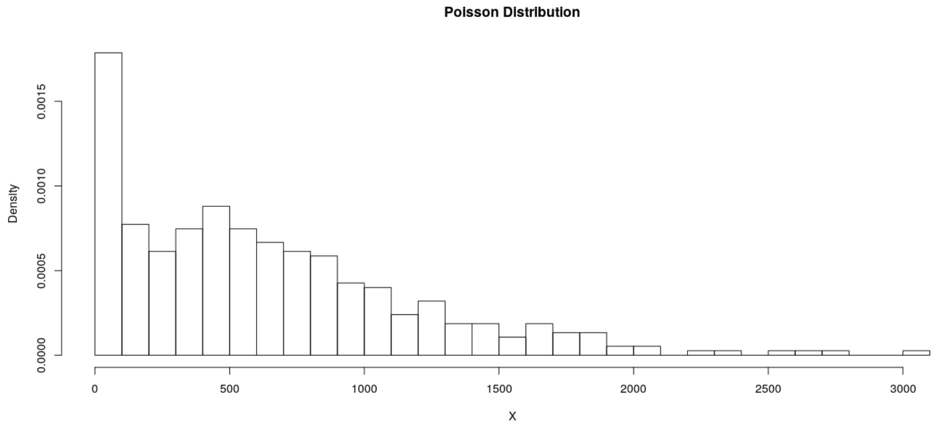
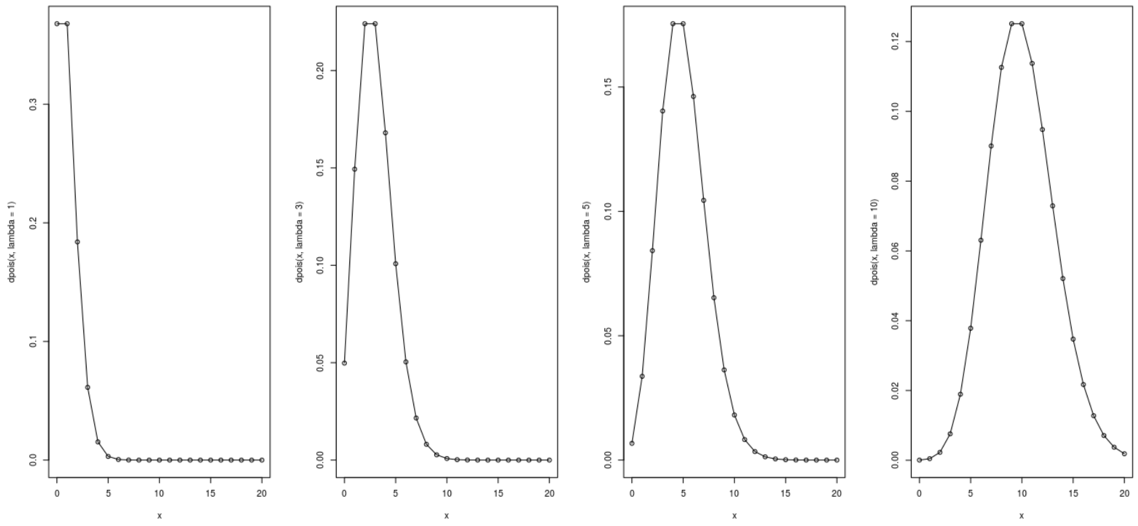


Figure 5: Plots of Poisson variable distributions with different values of λ.



The main variations between the Poisson regression and its counterpart to standard regression are many. The main distinction is the assumption of distribution; the Poisson regression model believes the residuals are obeying a distribution of Poisson rather than a normal distribution. Next, the predictor variables are related to the product of natural log transformation (Cameron & Trivedi, 2013), similar to what is done in Logistic regression (David W & Stanley, 2000). The log transformation ensures that the expected values of the regression model are never negative.

For a simple Poisson regression, the model is,

$$\underbrace{\ln(\lambda_i) = \mu_i}_{\substack{\text{Predicted Count} \\ \text{(Transformed by Link)}}} = \underbrace{a + bX_i}_{\substack{\text{Structural} \\ \text{(Original)}}}$$

Equation (2)

In the equation 2, X is a predictor variable, i is a set of observations with the same values for X, a and b are the intercept and slope, respectively, and μ_i is the predicted value of the outcome variable for all respondents whose X value is X_i. Since the mean of the Poisson distribution is λ and the link function for the Poisson regression is the natural log, the equation 2 shows that the mean of the regression equation, μi is equal to $\ln(\lambda i)$. Returning the output variable to its original count scale requires the transformation of the structural part of defined equation 2 by the inverse of the link function. The exponent function is the inverse of the natural log equation. Thus, the result that we get is,

$$\underbrace{\lambda_i = \exp(\mu_i)}_{\text{Predicted Count (Original)}} = \underbrace{\exp(a + bX_i)}_{\text{Structural (Transformed by Inverse Link)}}$$

Equation (3)

For each outcome group, we modelled the number of speeding stops, y, using a Poisson regression model and number of speeding stops for female driver in district I as a baseline,

Y = # speeding stops, X1 = subject_sex, X2= hour, X3=district

Assume $Y \sim \text{Poisson}(\mu)$, $E[Y] = \mu$, $Var[Y] = \mu$.

The model is $\log(\mu i) = \beta_0 + \beta_1$ *subject_sexmale+ β_2 *hour+ β_3 *districtII + β_4 *districtIII + β_5 *districtIV + β_6 *districtV+ β_7 *districtVI + β_8 *districtVII + β_9 *districtVIII

The model fit yields an estimated mean number of speeding stops $\mu\hat{}$ as:

$\mu\hat{} = \exp(\beta_0 + \beta_1$ *subject_sexmale+ β_2 *hour+ β_3 *districtII + β_4 *districtIII + β_5 *districtIV + β_6 *districtV+ β_7 *districtVI + β_8 *districtVII + β_9 *districtVIII)

3.1.3 NEGATIVE BINOMIAL REGRESSION

The Poisson distribution presumes the variable's mean and variance is the same. Count variables may not obey this rule particularly if there are more zeros or higher than expected values. It is termed overdispersion, resulting in the variance (v) of a measure being significantly greater than the mean (λ). In the GLM regression, overdispersion can be implemented by calculating the sum of extra variance. One way to do so is by using a negative binomial (NB) distribution of residues. The variance of the NB distribution models as

$$v = \lambda + \lambda^2/\theta$$

Equation (4)

where θ is a dispersion parameter (HOEF1 & BOVENG, 2007).

It turns out to be mathematically easy to conclude that θ has a gamma distribution with α and β parameters. This distribution has mean α/β and variance α/β^2, so we take $\alpha = \beta = 1/\sigma^2$, which makes the mean of the unobserved effect one and its variance σ^2. With this knowledge, we can measure the unconditional distribution of the result i.e. our response, which turns out to be the negative binomial distribution. The density is written in terms of the parameters α, β and μ as shown below in the equation 4, although it should be noted that in our case $\alpha = \beta = 1/\sigma^2$, there is only one more parameter compared to the Poisson model.

$$\Pr\{Y = y\} = (\Gamma(\alpha+y)/ y!\Gamma(\alpha)) (\beta^{\alpha}\mu^{y}/ (\mu + \beta)^{\alpha+y})$$

Equation (5)

This distribution is known as the number of failures before the kth success in a series of separate Bernoulli trials with an equal likelihood of success π. The density corresponding to this definition can be obtained from the expression referred to above by setting α = k and π = β/ (μ+β). The negative binomial distribution of $\alpha = \beta = 1/\sigma^2$ has a mean, E(Y) = μ and variance, $\text{var}(Y) = \mu(1 + \sigma^2 \mu)$. If $\sigma^2 = 0$ there is no unobserved heterogeneity and the Poisson variance is obtained. If $\sigma^2 > 0$, then the variance is greater than the mean. The negative binomial distribution is over-dispersed according to Poisson (Rodrıguez, 2013).

The form of the model equation for negative binomial regression is the same as that for Poisson regression. The log of the outcome is predicted with a linear combination of the predictors:

log(#speedingstops) = β_0 + β_1 *subject_sexmale+ β_2 *hour+ β_3 *districtII + β_4 *districtIII + β_5 *districtIV + β_6 *districtV+ β_7 *districtVI + β_8 *districtVII + β_9 *districtVIII

This implies:

#speedingstops = exp(β_0 + β_1 *subject_sexmale+ β_2 *hour+ β_3 *districtII + β_4 *districtIII + β_5 *districtIV + β_6 *districtV+ β_7 *districtVI + β_8 *districtVII + β_9 *districtVIII)

The dispersion parameter θ in negative binomial regression does not affect the expected counts, but it does effect the estimated variance of the expected counts.

3.1.4 MODEL COMPARISON

A significant feature of all models of regression is to evaluate how well the model matches the results, either by comparing the real values with the expected values of the model or by comparing a model with competing models. In our analysis below we demonstrate the latter approach. Usually R^2 is used in typical regressions as a measure of how close the actual values are to the predicted values. Although there are pseudo-R^2 values for count regression models, they have the challenges same as pseudo-R^2 for logistic regression, such as not really calculating variance and generating multiple formulae with varying values. Therefore, since there is a finite number of a possible outcome value for count regression models, we analyse model-data fit by evaluating the raw difference between the expected counts and actual counts at each outcome value.

If we have to compare different models, we make use of Information criterion based fit indices which are proven to be helpful (Burnham & Anderson, 2002). These fit measures' basic concept is to choose the simplest models that can represent the data well (Sherman & Funder, 2009). The Akaike Information Criterion (AIC) is a commonly used measure from the information theoretical tradition. AIC combines the goodness of fit of the model to the data with a punishment (penalty factor) for the difficulty of the model. The general approach of using the AIC is to pick the one which has the lowest AIC value. Specific AIC values cannot be represented explicitly because they contain unknown constants and are heavily influenced by sample size. Often such artificial changes in AIC values that make it seem that many models theoretically tend to have very similar AIC values, but some models match the data somewhat better than others. The AIC values for a group of models can be converted to add up to the number one, and they transform into probabilities. Such values are called Akaike weights, which are usually translated as the chance of modelling a given model being the better one for the data out of all the models compared.

Schwartz's Bayesian Knowledge Criterion (BIC) is another model-fit metric that penalizes models for complexity. While it is not related scientifically to knowledge theory, it can also be useful in model picking criteria. The general approach of using the BIC is to pick the one which has the lowest BIC value. The BIC appears to be excessively cautious for limited sample sizes (i.e., favours models with very few variables), but when the sample size is high it appears to choose the right model if a range of different models contains the true model.

CHAPTER 4

RESULTS

4.1 LINEAR REGRESSION (GAUSSIAN FAMILY)

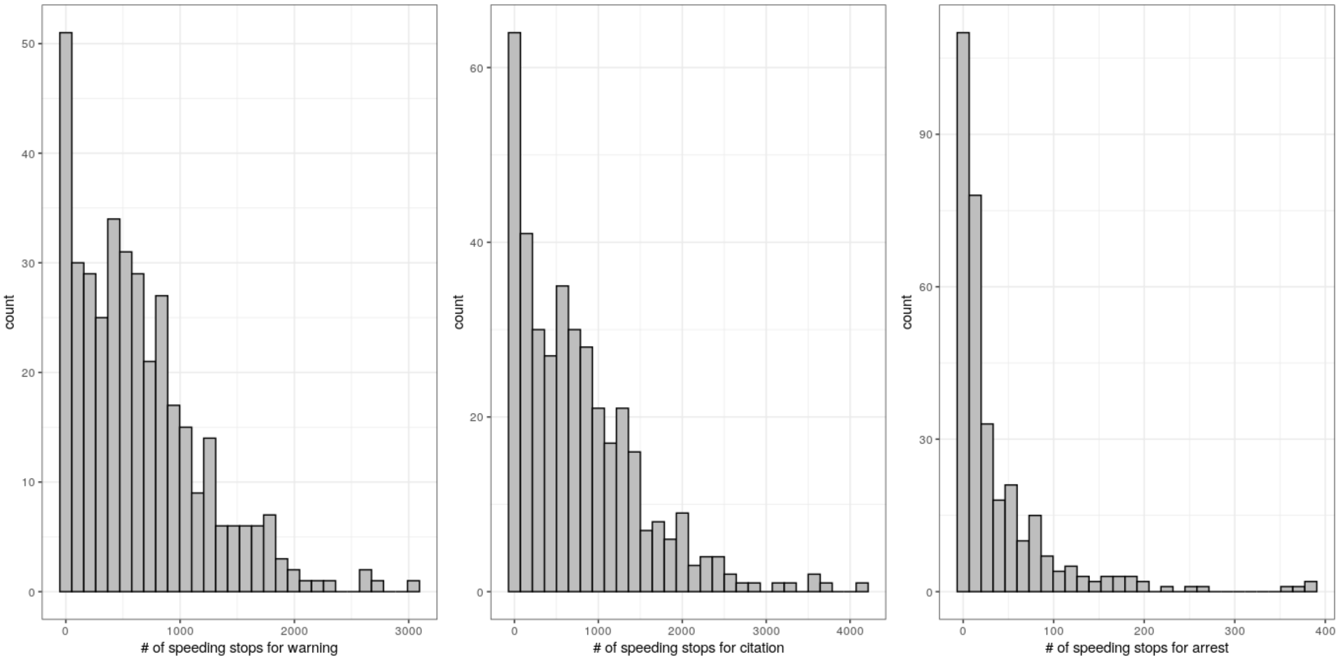
For the given data, this report is interested in predicting the number of speeding stops in each group of the outcome variable (warning, citation and arrest) by driver gender, hour and district. As a baseline, we fit a typical regression model to the data, i.e., a model that assumes the residuals follow a normal distribution. Often, these regression parameters are estimated through ordinary least squares (OLS). With normally-distributed residuals, OLS and maximum likelihood (ML) parameter estimates are the same (Kutner, et al., 2005). For consistency with the other models we fit, we used ML estimation for this model.

The results are shown in Table 5 to each of the three outcome groups. The intercept in the model is interpreted as the predicted number of speeding stops for each of the outcome group for a female driver and who is a resident of district I. The regression coefficients are interpreted as any other unstandardized coefficients from a typical regression.

Table 5: Linear Regression Model Summary

	Outcome		
	Warning	**Citation**	**Arrest**
Variables	**Estimate**		
Intercept	-39.501	-85.868	-14.8280
Subject_sexmale	394.574	537.681	23.7262
Hour	30.306	37.269	1.4822
DistrictII	-41.683	108.451	-16.0882
DistrictIII	-75.925	99.500	115.0537
DistrictIV	-100.897	54.792	-10.9016
DistrictV	299.512	391.264	16.8547
DistrictVI	332.774	-29.271	33.8733
DistrictVII	166.198	-9.500	-11.8651
DistrictVIII	405.921	642.667	31.8549

Figure 6: Histogram for the number of speeding stops in each group of outcome



The typical regression model which is the linear regression assumes that the residuals follow a normal distribution. A plot of the residuals for typical regression model is shown in Figure 7 and clearly shows they do not follow a normal distribution for outcome group, warning and citation. Another plot that is useful to examine is to compare the residuals to the predicted values. There should be no relationship between these two values, so the LOWESS line should be horizontal and close to zero (for more about LOWESS lines, see (Trexler & Travis, 1993)).

Figure 8 shows plots of the residuals vs. the predicted values. The typical regression shows that the horizontal line is not centered around 0, so linearity assumption might be violated. Also, as we move towards higher predicted values, the variance also becomes non-constant with presence of outliers. Looking at the Q-Q plot, we can see the points move away around the tails so normality seems to be violated too.

Figure 7: Residuals for Baseline regression model (Linear Model)

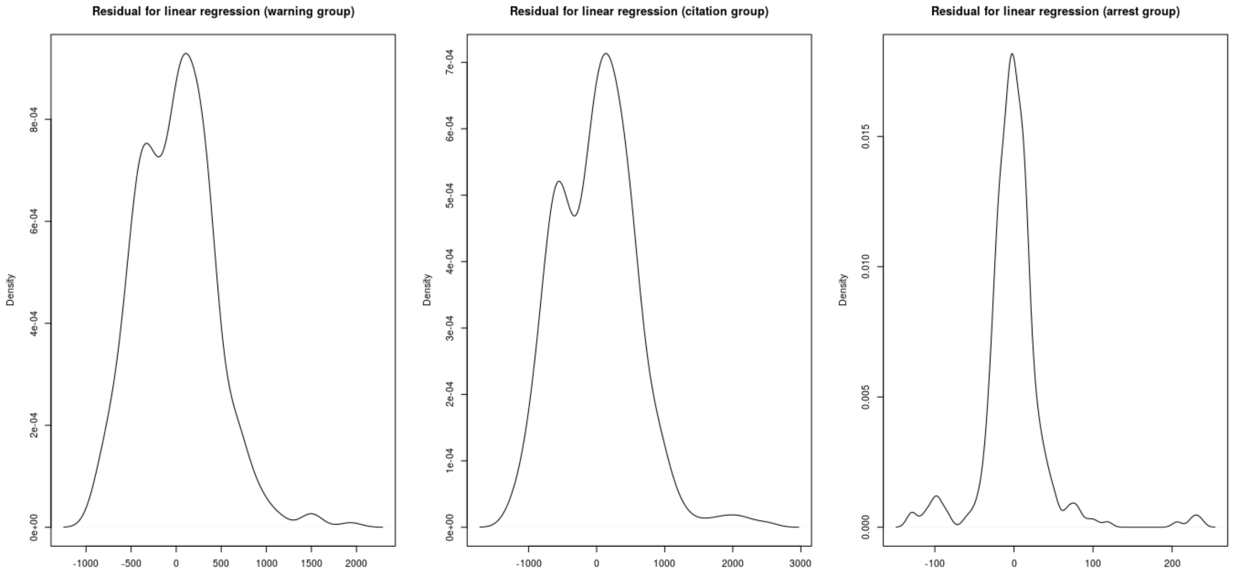
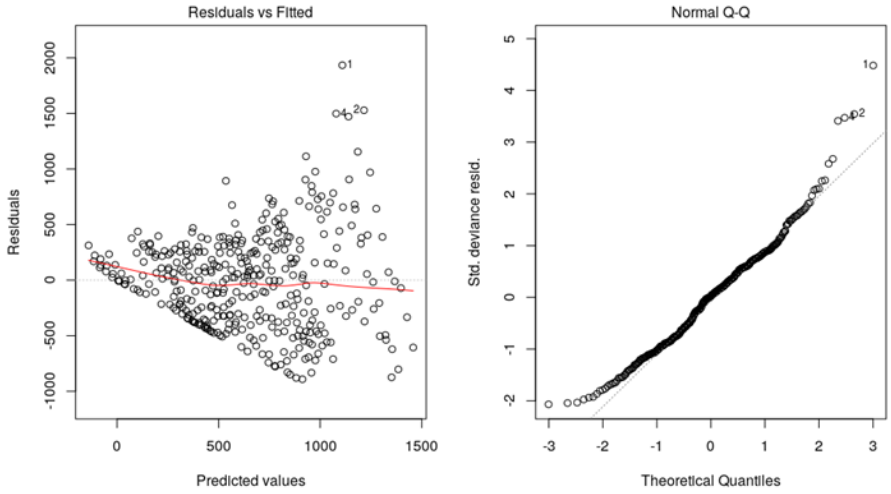


Figure 8: Predicted values vs. residual plot. LOWESS line is in red color



4.2 POISSON REGRESSION

We fit the Poisson regression model using the same predictors with number of speeding stops as dependent variable for each of the outcome group as we used with the baseline regression model. The results from the Poisson regression are shown in the Table 6, but the log link makes these values hard to interpret. This can be remedied by exponentiating the value. As with the baseline regression model, the intercept represents the predicted number of speeding stops for female drivers who all are residents of district I. topped.] The parameters of most interest are used in the regression analysis, the rate of speeding stops for each district, driver gender and hour. We display these graphically in Figure 9. Figure 9 shows the rate of speeding stops—those associated with outcome (warning, citation and arrest)—male drivers were much more likely to be stopped for speeding than females, in all categories of outcome. We see, ~90% (exp (0.65)) increase in the number of speeding stops for males compared to females under arrest group. For arrest outcome, district 3 residents were stopped 8 times higher than district 1. It seems that there is ~80% increase in the number of warning stops in district 8 compared to 1. Whereas, there is a ~20 % decrease in district 4 compared to 1.

The percentage change in the expected counts method of interpretation requires two values: the regression coefficient and the desired amount of change in the variable. For the number of warning stops in district II for male driver residents, the regression coefficient is −0.087014. For the amount of change we use one SD, which is 9.98. Plugging those values into Equation (6) produces

$$100 \times [\exp(-0.087014 \times 9.98) - 1] = -55$$

Equation (6)

Meaning there is a 55% decrease in the expected speeding stops value for a one SD increase.

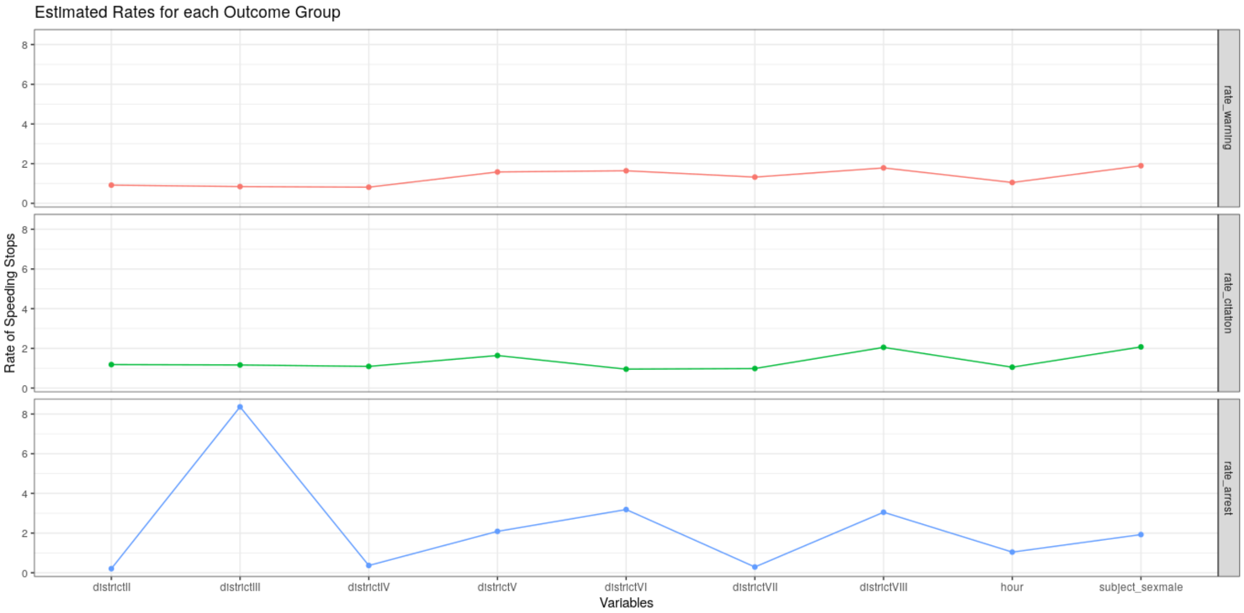
Table 6: Poisson Regression Model Summary

	Outcome		
	Warning	**Citation**	**Arrest**
Parameter	**Estimates**		
Intercept	5.2478673	5.3584377	1.815143
Subject_sexmale	0.06394409	0.7292324	0.654589
Hour	0.0486969	0.0495460	0.040911
DistrictII	-0.0870149	0.1671342	-1.584473
DistrictIII	-0.1693891	0.1507432	2.123759
DistrictIV	-0.2070352	0.0858042	-1.001740
DistrictV	0.4578051	0.4936519	0.735703
DistrictVI	0.4951215	-0.049057	1.158789
DistrictVII	0.2787305	-0.0156559	-1.235004
DistrictVIII	0.5801246	0.7182593	1.115295

Parameter	**Exp(Estimates)**		
Intercept	190.1602782	212.3928574	6.1419541
Subject_sexmale	1.8954208	2.0734885	1.9243516
Hour	1.0499021	1.0507940	1.04175960
DistrictII	0.9166634	1.1819129	0.2050557
DistrictIII	0.8441804	1.1626980	8.3625126
DistrictIV	0.81299910	1.0895929	0.3672397
DistrictV	1.5806009	1.6382883	2.0869488
DistrictVI	1.6406975	0.9521376	3.1860734
DistrictVII	1.3214512	0.9844660	0.2908337
DistrictVIII	1.7862611	2.0508602	3.0504667

Parameter	**Standard Error**		
Intercept	0.0082387	0.0074779	0.045377
Subject_sexmale	0.0043102	0.0039438	0.019170
Hour	0.0003079	0.0002765	0.001384
DistrictII	0.0093302	0.0079464	0.103334
DistrictIII	0.0094680	0.0079601	0.040946
DistrictIV	0.0096148	0.0080827	0.076471
DistrictV	0.0081909	0.0074315	0.047238
DistrictVI	0.0081759	0.0083572	0.044375
DistrictVII	0.0085120	0.0082866	0.083097
DistrictVIII	0.0080206	0.0071187	0.044670

Figure 9: Estimated rates at which people of different districts were stopped for different categories of outcome, as estimated from Poisson regression using districtI and female driver as a baseline. Rates are plotted on an exponent scale.



As the Poisson regression model assumes the mean-variance of the variable to be equal, but it is not always true. In this case, for warning outcome group, our residual deviance is 110477 for 365 degrees of freedom (df). The rule of thumb is that the ratio of deviance to df should be 1, but it is 302.6, indicating severe overdispersion. Similarly, the ratio is much greater than 1 for the rest of the outcome groups. This can be done more formally by performing the dispersion test through different packages available in R software (DHARMa or AER) (Dormann, 2016).

Overdispersion test

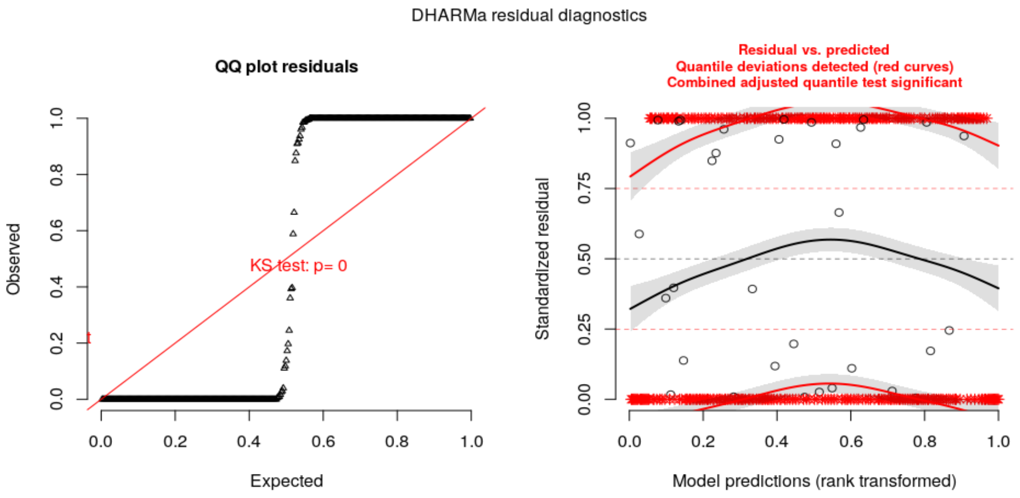
Data: Poisson regression, z = 16.115, p-value < 2.2e-16

Alternative hypothesis: true alpha is greater than 0

Sample estimates: **dispersion 257.5345**

The value for dispersion parameter is 257.5345 which is lower than 302.6 (it was a rule of thumb!), but the result is the same: substantial overdispersion. The advantage of using the method for test is that we can additionally visualize overdispersion.

Figure 10: Residual Diagnostics



We see various signals of overdispersion (see figure 10 above)

- QQ: s-shaped QQ plot, distribution test (KS) significant
- QQ: Dispersion test is significant
- QQ: Outlier test significant
- Res ~ predicted: Quantile fits are spread out too far

4.3 NEGATIVE BINOMIAL REGRESSION

The results from the negative binomial (NB) regression are shown in the Table 7. The NB model is very similar to the Poisson model, thus the NB model's coefficients are interpreted in the same way as the Poisson regression. The main difference between the NB and Poisson models is that the NB model allows for more variability (dispersion) in the outcome by not assuming the mean and variance of the residuals are the same.

Table 7: NB Regression Model Summary

	Outcome		
	Warning	**Citation**	**Arrest**
Parameter	**Estimates**		
Intercept	4.842275	4.80746	1.717132
Subject_sexmale	0.633694	0.70943	0.752518
Hour	0.083419	0.09666	0.044059
DistrictII	-0.01421	0.09828	-1.56125
DistrictIII	-0.24255	0.10214	2.129202
DistrictIV	-0.24255	0.12544	-0.97419
DistrictV	0.324012	0.38086	0.664763
DistrictVI	0.522108	-0.04774	1.221581
DistrictVII	0.258222	-0.02485	-1.27354
DistrictVIII	0.517715	0.65060	1.052063

Parameter	**Exp(Estimates)**		
Intercept	126.7574311	122.4202465	5.5685362
Subject_sexmale	1.8845596	2.0328387	2.1223371
Hour	1.0869973	1.1014833	1.0450436
DistrictII	0.8674997	1.1032697	0.2098735
DistrictIII	0.7846186	1.1075324	8.4081583
DistrictIV	0.8502095	1.336488	0.3774969
DistrictV	1.3826640	1.4635384	1.9440299
DistrictVI	1.6855775	0.9533810	3.3925470
DistrictVII	1.2946266	0.09754585	0.2798391
DistrictVIII	1.6781890	1.9166978	2.8635512

Parameter	**Standard Error**		
Intercept	0.176771	0.18028	0.143111
Subject_sexmale	0.102290	0.10522	0.082945
Hour	0.007415	0.00761	0.006078
DistrictII	0.205398	0.21071	0.192900
DistrictIII	0.203245	0.20959	0.152476
DistrictIV	0.205398	0.20959	0.170371
DistrictV	0.203169	0.21185	0.155708
DistrictVI	0.205319	0.20961	0.153537
DistrictVII	0.204232	0.20961	0.173870
DistrictVIII	0.204208	0.20955	0.153933

Consequently, the NB model estimates one extra parameter than the Poisson model: a overdispersion parameter (see Equation 4). The value for overdispersion parameter for the speeding stop for each outcome group data is between 1 and 2. Since the NB and Poisson models are so similar, it is not surprising that the regression coefficients for the two models are very close. The standard errors, however, are larger for the NB model reflecting its larger residual variance.

The plots of the predictor variables against the standardized residuals are shown in Figure 12. Based on visual inspection, we determined that the residual distributions were approximately the same across levels of the predictor variables. We noticed that there are discrepancies for the observations in the hour distribution, which produces a dissimilar pattern of residuals for other predictor variable. On the whole, the residual patterns across all predictor variables from the NB model are acceptable. When performed overdispersion test on each of the outcome group, the overdispersion looks much better and QQ plot looks better than before and residuals also do not exhibit a pattern like before (see figure 11).

Figure 11: Residual Diagnostics.

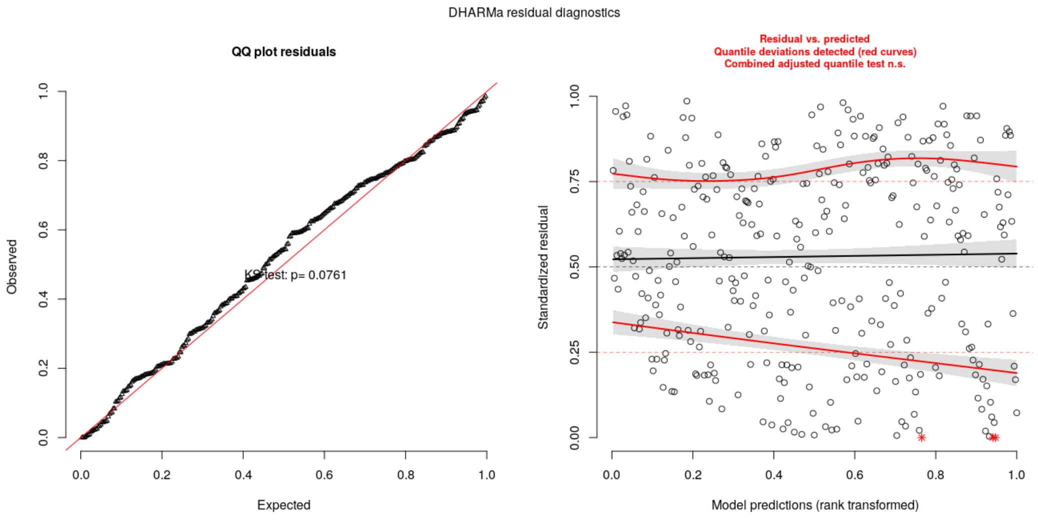
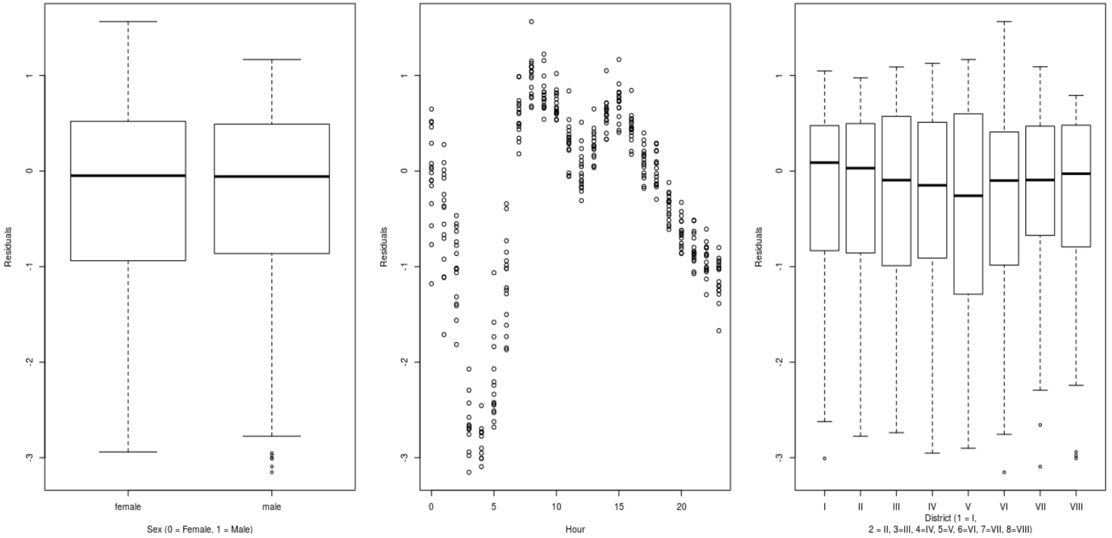


Figure 12: Plots of negative binomial model predictors by residuals.



4.4 MODEL COMPARISON

To examine model fit, we first compared the AIC and BIC values for the Negative Binomial with Poisson model to those from the baseline regression model (see Table 8). Second, we compared and overlaid the distributions for the model fits to the actual data. The Negative Binomial model appears to do a much better job capturing the speeding stops data than the typical and Poisson regression model. The AIC values for NB model are smaller than those for the Poisson model, indicating that the NB model fits the data somewhat better than the Poisson model. The amount of difference is very large as there appears to be enough overdispersion in the speeding stops variable that the Poisson model was not able to capture the variance as well as the NB model. This result is echoed in Figure 13 which shows that the NB model provides more accurate fit to the data distribution than the Poisson regression.

Figure 13: Histogram of speeding stops with overlaid predicted probabilities from each regression model (Blue=Linear, Red= Poisson, Green=NB)

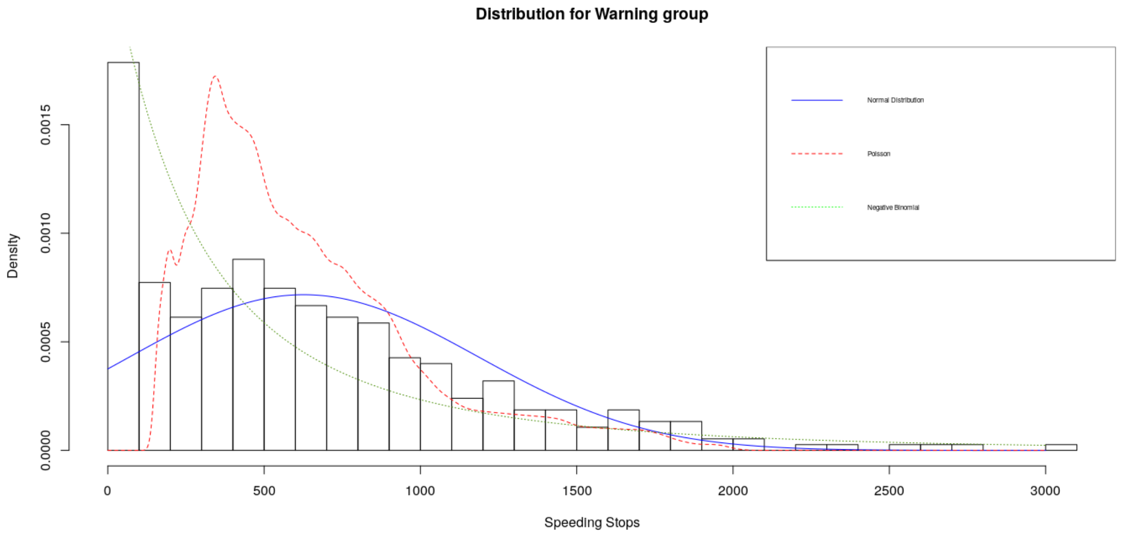


Table 8: Comparison of Regression Model for each outcome group (Warning, Citation, and Arrest)

Models	AIC Value	AIC Delta	AIC Likelihood	AIC Weight
Normal	5635.725	158.6018	3.631196e-35	3.631196e-35
Poisson	113322.89	107845.7742	0.000000e+00	0.000000e+00
NB	5477.123	0.0000	1.000000e+00	1.000000e+0

Models	AIC Value	AIC Delta	AIC Likelihood	AIC Weight
Normal	5952.288	259.2803	4.988942e-57	.988942e-57
Poisson	163548.07	157855.069	0.000000e+00	0.000000e+00
NB	5693.008	0.0000	1.000000e+00	1.000000e+00

Models	AIC Value	AIC Delta	AIC Likelihood	AIC Weight
Normal	3349.982	830.6726	4.185368e-181	4.185368e-181
Poisson	7065.590	4546.2799	0.000000e+00	0.000000e+00
NB	2519.310	0.0000	1.000000e+00	1.000000e+00

CHAPTER 5

DISCUSSION

Law enforcement is critical to the prevention of traffic accidents, but Montana state government passed a rule that went against all the odds of existing traffic laws. This study explores the situation of the state after the government abolished the day-time speeding limit and a penalty charged as small as $5 by identifying factors that influence the speeding traffic stops for each outcome issued by the MHP from 2009 to 2016. This study is different from the existing researches as it takes into account only the number of speeding stops leading to warnings, citations (ticket) or arrest occurring in Montana in order to exclude possible confounding effects. The negative binomial regression model is used to better compensate for the over-dispersion and highly skewed dependent variable. It is found that, with the increase in the speeding stops, traffic stops are more likely to lead to tickets and warnings and negligible number of arrests made. Male drivers are more likely to receive a ticket than a warning as compared to females as a result of speeding stop, which may imply the possible presence of gender biasing. In addition, speeding traffic stops at 3 p.m. are more likely to lead to tickets. Finally, the hot areas of state are identified, where drivers are more likely to get tickets issued or get arrested at speeding traffic stops. These areas are located mainly the counties in district III, VI, VIII and along some major highways.

All the models fit thus far assume that all the predictor variables are needed for each outcome group. An examination of the values in Table 9 indicates the same. Specifically, it appears that only the driver gender and hour variables might be needed. Consequently, we pruned the model removing each variable singly and in sets. The BIC indicted that the model with only the driver gender and hour variables fit the best, while the AIC indicated that all three variables: driver gender, hour and district should be kept. We opted to keep all the three variables (full model) in the model as indicated by AIC.

Table 9: Comparison of Negative Regression Model for warning group against each predictor

Models	BIC Value	BIC Rank	AIC Value	AIC Delta	AIC Likelihood	AIC Weight
Full	5520.319	2	5477.123	0.00000	1.000000e+00	9.997962e-01
Sex/Hour	5509.827	1	5494.119	16.99626	2.038493e-04	2.038078e-04
Hour/District	5550.404	3	5511.134	34.01136	4.116496e-08	4.115657e-08
Sex/District	5580.890	4	5541.620	64.49747	9.875358e-15	9.873345e-15

This research is limited as it only tackles speed, just one of many legitimate reasons for a stop to traffic. We conjecture that pace (over-speeding) is by far the most common reason to stop the vehicle but because the traffic stop data lists only that the offenses that comes under the category of a moving breach, without specifying we cannot examine this possibility empirically for the offence. However, due to the absence of geographic distribution of the state, the study could not perform spatial analysis. It is also important to consider that there might be presence of significant spatial correlations across traffic stops and taking these correlations into account may give more accurate results as it will give better insights on the police behavior, where some of the officers may be more strict or biased towards some groups than others and will be able to locate hot spots of over-speeding more clearly. Thus, spatial analysis and performing spatial regression can be very helpful for the government of Montana for taking the necessary measures to change the current policies for future improvement

.

CHAPTER 6

CONCLUSION

The fundamental law of Montana appears to be a straightforward proposition: on the highways and interstates of the state, motorists will travel in a "Reasonable and Cautious" manner. Yet the 5effect of the Law on America's fourth largest city, its highways, its courtrooms and its inhabitants proved to be something less than easy. The most discernible consequence of the Law is undoubtedly that Montana has become a "pace trap," drawing visitors to see the interstates of Montana just like race tracks. It is also clear that the simple rule generated a variety of complications. Within the Basic Law, road safety, fuel economy, air quality and Montana's image all suffered.

In this study, we present through our careful analysis on various reason for stops that Montana has an over-speeding issue. First, our results suggest that the estimated effects of the law that the government of Montana passed in 1995 are quite sensitive on the population as it increases the risk of fatalities. We also tried to comment on the behaviour of police through carrying out an analysis on the population of Montana in comparison to the total number of traffic stops for each ethnicity. In particular, these results suggest that the state is not suffering through any racial discrimination. However, the method used for analysis has some limitations and thus, we cannot necessarily imply that MHP is not acting in unfair or racist manner.

Second, this study also provided evidence that the overall effect of over- speeding is higher in some districts as compared to others but still there is no sign of action through MHP department to safeguard the lives of people residing in such hotspot areas. Third, we presented evidence that these increases in speeding stops differ between males and females, with males having higher speeding stops and getting more tickets and warnings as compared to female counterpart. These findings may signal towards the basis for gender biasing existing in Montana, but to confirm the same more statistical analysis has to be performed. We also found that the speeding stops reach their peak post office/school hours suggesting some unusual movement of the state population. When we tried to explore further this finding with location, it resulted that these stops occur mostly on highways or roadways that connects to highways. An understanding of these findings should be an important part of an informed public debate on the desirability of this no speeding limit on day-time policy change. However, the pattern of these results also suggests the need for a better understanding of the structural relationship between speed limits, driving behaviour, driving ability and traffic safety. In particular, a clear understanding of why higher speed stops are more often particularly for men and fatality risks for women and the elderly in accident prone areas may be quite useful since it could suggest ways in which the unfortunate state of affair associated with higher speeding stops can be reduced.

APPENDIX: R CODE

```
#loading required libraries
library(dplyr)
library(tidyverse)
library(lubridate)
library(forcats)
library(gridExtra)
library(MASS)
library(pscl)
library(AER)
library(devtools)
library(DHARMa)

 # import data
montana<- read_csv("mt_statewide_2020_04_01.csv")

#Having a look at dimensions of data and names. Have a peek at the datatypes of columns
dim(montana)
names(montana)
str(df)

#Learning about missing values in the dataframe:
na_count <-sapply(montana, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
print(na_count)

#adding new variable district to the dataset, used as a predictor in Modeling Methods
montana<-montana%>%
mutate(district = ifelse(county_name %in% c(" Mineral County", "Missoula County", "Ravalli
County", "Sanders County"),"I", ifelse(county_name %in% c("Cascade County", "Fergus County",
"Golden Valley County", "Judith Basin County", "Musselshell County", "Petroleum County", "Teton
County", "Wheatland County"), "II",  ifelse(county_name %in% c("Beaverhead County", "Deer
Lodge County", "Granite County",  "Jefferson County",  "Lewis & Clark County", "Madison County",
"Powell County", "Silver Bow counties County"), "III", ifelse(county_name %in% c(" Big Horn
County",  "Carbon County", "Stillwater County", "Sweet Grass County", "Yellowstone County"),"IV",
ifelse(county_name %in% c("Carter County ", "Custer County",  "Daniels County", "Dawson County",
"Fallon County", "Garfield County", "McCone County", "Powder River County", "Prairie County",
"Richland County", "Roosevelt County", "Rosebud County", "Sheridan County", "Treasure County",
"Valley County", "Wibaux County"),"V",ifelse(county_name %in% c("Flathead County","Lake
County","Lincoln County"),"VI",ifelse(county_name %in% c("Broadwater County", "Gallatin
County", "Madison County", "Meagher County", "Park County"),"VII","VIII"))))))))

#Collapsing various categories of Reason for Stop Attribute into similar groups for better analysis
montana$reason_for_stop<-montana$reason_for_stop %>% fct_collapse(speeding = c("--- - SPEED
OVER LEGAL","--- - TRUCK SPEED", "--- - RECKLESS DRIVING","--- - CARELESS
DRIVING","CARELESS DRIVING","--- - SPEED BASIC RULE","SPEED","--- - HIT AND RUN","RECKLESS
DRIVING"))
```

```
montana$reason_for_stop<-montana$reason_for_stop %>% fct_collapse(toll_related =
c("EXPIRED TAG ( - MONTHS OR LESS )","OTHER TAG/REGISTRATION VIOLATIONS", "TOLL
EVASION","EXPIRED TAG ( MORE THAN - MONTHS )" ))

montana$reason_for_stop<-montana$reason_for_stop %>% fct_collapse(issues_with_vehicle =
c("--- - OTHER DEFECTIVE EQUIPMENT","FAULTY EQUIPMENT", "NO REGISTRATION","LOADS ON
VEHICLE","--- - DEFECTIVE BRAKES","--- - DEFECTIVE EXHAUST","--- - DISPLAYING ONLY ONE
LICENSE PLATE" ))

montana$reason_for_stop<-montana$reason_for_stop %>% fct_collapse(license_issues = c("--- -
VEHICLE LICENSE","OTHER DRIVERS LICENSE VIOLATIONS", "--- - FOREIGN LICENSE","--- -
DRIVERS LICENSE", "DRIVING WHILE LICENSE SUSPENDED/REVOKED","--- - OVER LICENSE
CAPACITY","--- - SUSPENDED OR REVOKED D/L" ))

montana$reason_for_stop<-montana$reason_for_stop %>% fct_collapse(criminal_offence =
c("OTHER CRIMINAL TRAFFIC VIOLATIONS","PROPERTY CRIMES", "OTHER CRIMINAL
VIOLATIONS","CRIMES AGAINST PERSONS", "--- - CHILD RESTRAINT","--- - FURNISHING LIQUOR
TO MINORS","--- - DRUGS","--- - ILLEGAL POSSESSION OF LIQUOR"))

montana$reason_for_stop<-montana$reason_for_stop %>% fct_collapse(road_rule_violations =
c("--- - FAIL TO / IMPROPER SIGNAL","--- - IMPROPER PASSING", "--- - CROSSING
DIVIDER/BARRIER","--- - IMPROPER LANE TRAVEL", "--- - PEDESTRIAN","FOLLOWING TOO
CLOSELY","DRIVING WITHOUT HEADLIGHTS","DRIVING ON WRONG SIDE OF
ROAD","OBSTRUCTING THE ROADWAY","IMPROPER TURN ","IMPROPER PASSING","--- -
IMPROPER TURN","--- - DRIVING OVER CENTERLINE","--- - FAIL TO STOP - SIGN OR LIGHT","--- -
RIGHT OF WAY","IMPROPER START"))

#Performing Exploratory Analysis
#Montana Statewide Percentage of Stops by Race/Ethnicity ------------------------------------------------
df<-montana %>%
  dplyr::select(subject_race)%>%
  drop_na()%>%
  group_by(subject_race) %>%
  summarise(number_of_stops=n())%>%
  ungroup()%>%
  mutate(percentage_stops=(number_of_stops/sum(number_of_stops))*100)

#Reason for Stop by Gender ------------------------------------------------------------------------------------------
df1<-montana %>%
  dplyr::select(reason_for_stop,subject_sex)%>%
  drop_na()%>%
  group_by(reason_for_stop,subject_sex)%>%
  summarise(stops=n())%>%
  arrange(desc(stops))%>%
  head(10)%>%
  ungroup()

ggplot(df1,aes(x=reorder(reason_for_stop,stops),y=stops,fill=subject_sex))+
  geom_bar(stat="identity",position = "dodge")+
  scale_y_continuous(labels=scales::comma)+
  theme_bw()+
  xlab("Top 5 reason for stop") +
  ylab("Number ofStops")

#Number of Speeding Stops by Hour and Outcome -------------------------------------------------------------
df2<-montana %>%
  dplyr::select(reason_for_stop,time,outcome)%>%
  drop_na()%>%
  group_by(reason_for_stop,hour=hour(time),outcome)%>%
  summarise(speeding_stops=n())%>%
  arrange(desc(speeding_stops))%>%
  filter(reason_for_stop=="speeding")%>%
  ungroup()

ggplot(df2,aes(x=hour,y=speeding_stops,col=outcome))+
  geom_point()+
  geom_line()+
  scale_y_continuous(labels=scales::comma)+
  theme_bw()+
  xlab("Time at which speeding stop occurred") +
  ylab("Number of Speeding Stops")

#Number of speeding stops by Driver Gender and Outcome --------------------------------------------------
df3<-montana %>%
  dplyr::select(reason_for_stop,subject_sex,outcome,time)%>%
  drop_na()%>%
  group_by(reason_for_stop,subject_sex,outcome,hour=hour(time))%>%
  summarise(speeding_stops=n())%>%
  arrange(desc(speeding_stops))%>%
  filter(reason_for_stop=="speeding")%>%
  ungroup()

ggplot(df3, aes(hour, speeding_stops,col=outcome)) +
 geom_point()+
  geom_line()+
  facet_grid(subject_sex~.)+
  theme_bw()

#generating poisson distribution with different values of  λ, the parameter describing the rate, that
is the mean of the distribution ---------------------------------------------------------------------------------------
x <- seq(0,20,5)
plot(x, dpois(x, lambda=1),type = "o")
plot(x, dpois(x, lambda=3),type="o")
plot(x, dpois(x, lambda=5),type="o")
plot(x, dpois(x, lambda=10),type="o")

#Histogram for the number of speeding stops in each group of outcome (knowing the distribution
of our response under each outcome group) ----------------------------------------------------------------------
g0 <- ggplot(data=speeding_stops_a,aes(x=speedingstops))
```

```
g1<-g0 +geom_histogram(fill="grey",colour="black") + theme_bw() + xlab("# of speeding stops
for arrest")
g2 <- ggplot(data=speeding_stops_c,aes(x=speedingstops))
g3<-g2 +geom_histogram(fill="grey",colour="black") + theme_bw() + xlab("# of speeding stops
for citation")
g4 <- ggplot(data=speeding_stops_w,aes(x=speedingstops))
g5<-g4 +geom_histogram(fill="grey",colour="black") + theme_bw()+ xlab("# of speeding stops
for warning")
grid.arrange(g5, g3, g1, nrow = 1)

#preparing data for modeling for each outcome group (Warning, Citation and Arrest) with
variables of interest
speeding_stops_w<-montana%>%
 dplyr::select(reason_for_stop,subject_sex,time,outcome,district)%>%
 group_by(reason_for_stop,subject_sex,outcome,hour=hour(time),district)%>%
 summarise(speedingstops=n())%>%
 filter(reason_for_stop=="speeding")%>%
 filter(outcome=="warning")%>%
 arrange(desc(speedingstops))%>%
 drop_na()%>%
 ungroup()

speeding_stops_c<-montana%>%
 dplyr::select(reason_for_stop,subject_sex,time,outcome,district)%>%
 group_by(reason_for_stop,subject_sex,outcome,hour=hour(time),district)%>%
 summarise(speedingstops=n())%>%
 filter(reason_for_stop=="speeding")%>%
 filter(outcome=="citation")%>%
 arrange(desc(speedingstops))%>%
 drop_na()%>%
 ungroup()

speeding_stops_a<-montana%>%
 dplyr::select(reason_for_stop,subject_sex,time,outcome,district)%>%
 group_by(reason_for_stop,subject_sex,outcome,hour=hour(time),district)%>%
 summarise(speedingstops=n())%>%
 filter(reason_for_stop=="speeding")%>%
 filter(outcome=="arrest")%>%
 arrange(desc(speedingstops))%>%
 drop_na()%>%
 ungroup()

#fitting regression models and summary of results
# normal theory regression using maximum likelihood
normal <- glm(speedingstops ~ subject_sex+hour+district,
        data = speeding_stops_w, family = gaussian(link="identity"))
normal_c <- glm(speedingstops ~ subject_sex+hour+district,
         data = speeding_stops_c, family = gaussian(link="identity"))
normal_a <- glm(speedingstops ~ subject_sex+hour+district,
         data = speeding_stops_a, family = gaussian(link="identity"))
```

```
summary(normal)
summary(normal_c)
summary(normal_a)

#poisson regression model and summary of results
summary(mp <- glm(speedingstops ~ subject_sex+hour+district,
family=poisson,data=speeding_stops_w))

summary(mp_c <- glm(speedingstops ~ subject_sex+hour+district,
family=poisson,data=speeding_stops_c))

summary(mp_a <- glm(speedingstops ~ subject_sex+hour+district,
family=poisson,data=speeding_stops_a))

#negative binomial regression and summary of results
summary(mnb <- glm.nb(speedingstops ~ subject_sex+hour+district, data=speeding_stops_w))
summary(mnb_c <- glm.nb(speedingstops ~ subject_sex+hour+district, data=speeding_stops_c))
summary(mnb_a <- glm.nb(speedingstops ~ subject_sex+hour+district, data=speeding_stops_a))

#Model Diagnostics --------------------------------------------------------------------------------------------------
# normal residuals density plot
par(mfrow=c(1,3))
plot(density(residuals(normal)),main = "Residual for linear regression (warning group)")
plot(density(residuals(normal_c)),main = "Residual for linear regression (citation group)")
plot(density(residuals(normal_a)), main = "Residual for linear regression (arrest group)")

#Predicted values vs. residual plot. ---------------------------------------------------------------------------------
plot(predict(normal, type="response"), residuals(normal), main="Typical
Regression", ylab="Residuals", xlab="Predicted")
abline(h=0,lty=1,col="gray")
lines(lowess(predict(normal,type="response"),residuals(normal)), lwd=2, lty=2)

plot(predict(normal_c, type="response"), residuals(normal_c), main="Typical
Regression", ylab="Residuals", xlab="Predicted")
abline(h=0,lty=1,col="gray")
lines(lowess(predict(normal_c,type="response"),residuals(normal_c)), lwd=2, lty=2)

plot(predict(normal_a, type="response"), residuals(normal_a), main="Typical
Regression", ylab="Residuals", xlab="Predicted")
abline(h=0,lty=1,col="gray")
lines(lowess(predict(normal_a,type="response"),residuals(normal_a)), lwd=2, lty=2)

#over-dispersion test on poisson regression model ----------------------------------------------------------------
dispersiontest(mp,trafo = 1)
sim_mp <- simulateResiduals(fittedModel = mp,n = 250,refit = T)
testDispersion(sim_mp)
plot(sim_mp)

# diagnostic plots for negative binomial model ------------------------------------------------------------------
par(mfrow=c(1,3))
```

```
plot(as.factor(speeding_stops_w$subject_sex),resid(mqp),xlab="Sex (0 = Female, 1 = Male)",
    ylab="Residuals")
plot(speeding_stops_w$hour,resid(mqp),xlab="Hour", ylab="Residuals")
plot(as.factor(speeding_stops_w$district),resid(mqp),xlab="District (1 = I,
2 = II, 3=III, 4=IV, 5=V, 6=VI, 7=VII, 8=VIII)", ylab="Residuals")

#Estimated rates at which people of different districts were stopped for different categories of
outcome, as estimated from Poisson regression using districtI and female driver as a baseline.
#Rates are plotted on an exponent scale. -----------------------------------------------------------------------
a1<-data.frame(exp(mnb$coefficients))
b1<-data.frame(exp(mnb_c$coefficients))
c1<-data.frame(exp(mnb_a$coefficients))
a1$rate_warning<-a1$exp.mnb.coefficients.
a1$parameters<-row.names(a1)
a1<-a1[-1]
b1$rate_citation<-b1$exp.mnb_c.coefficients.
b1$parameters<-row.names(b1)
b1<-b1[-1]
c1$rate_arrest<-c1$exp.mnb_a.coefficients.
c1$parameters<-row.names(c1)
c1<-c1[-1]
d1<-cbind(a1,b1,c1)
d1<-d1[-1,]
d1<-d1[,c(-2,-4)]
data_long1 <- gather(d1, outcome, rate, rate_warning:rate_arrest, factor_key=TRUE)
data_long1$parameters<-as.factor(data_long1$parameters)
ggplot(data = data_long1, aes(x=parameters, y=rate,col=outcome,group=1)) +
  geom_line()+geom_point()+
  ggtitle("Estimated Rates for each Outcome Group") +
  xlab("Variables")+
  facet_grid(outcome ~ .)+
  ylab("Rate of Speeding Stops")+
  theme_bw()+
  theme(legend.position = "none")

#Model Comparison
#overlaying model fit on distribution of data to check best fitted model
# predicted values for linear regression (done for warning group)
normal.y.hat <- predict(normal, type = "response")
normal.y <- normal$y
normal.yUnique <- 0:max(normal.y)
normal.nUnique <- length(normal.yUnique)
phat.normal <- matrix(NA, length(normal.y.hat), normal.nUnique)
dimnames(phat.normal) <- list(NULL, normal.yUnique)
for (i in 1:normal.nUnique) {
  phat.normal[, i] <- dnorm(mean = normal.y.hat, sd = sd(residuals(normal)),x =
                        normal.yUnique[i])
}
# mean of the normal predicted probabilities for each value of the outcome
phat.normal.mn <- apply(phat.normal, 2, mean)
```

```
#predcitions for poisson and negative binomial model
phat.pois <- predprob(mp)
phat.pois.mn <- apply(phat.pois, 2, mean)
phat.nb <- predprob(mnb)
phat.nb.mn <- apply(phat.nb, 2, mean)

#overlaying predictions on data distribution to check best fit model ---------------------------------------
hist(speeding_stops_w$speedingstops, probability = TRUE,
    main = "Distribution for Warning group", xlab = "Speeding Stops",
    breaks = 30)
lines(x = seq(0,3000,length.out = 3044), y = phat.normal.mn, pch = 18,col = "blue", lty = 1)
lines(x = seq(0, 3000, length.out = 3044), y = phat.pois.mn,pch = 19, col = "red", lty=2)
lines(x = seq(0, 3000, length.out = 3044), y = phat.nb.mn, pch = 22, col = "chartreuse4", lty=3)
# Add a legend to the plot
legend("topright", legend=c("Normal Distribution", "Poisson", "Negative Binomial"),
      col=c("blue", "red", "green"), lty = 1:3, cex=0.5)

#Model Comparsion using AIC (warning group) -----------------------------------------------------------------
compare.models <- list( )
compare.models[[1]] <- normal
compare.models[[2]] <- mp
compare.models[[3]] <- mnb
compare.names <- c("Normal","Poisson","NB")
compare.results <- data.frame(models = compare.names)
compare.results$aic.val <- unlist(lapply(compare.models, AIC))
compare.results$aic.delta <- compare.results$aic.val-min(compare.results$aic.val)
compare.results$aic.likelihood <- exp(-0.5* compare.results$aic.delta)
compare.results$aic.weight <-compare.results$aic.likelihood/sum(compare.results$aic.likelihood)

#Selecting best features for negative binomial regression model(done for warning group) -----------
cand.models <- list( )
cand.models[[1]] <- glm.nb(speedingstops ~ subject_sex+hour+district, data = speeding_stops_w)
cand.models[[2]] <- glm.nb(speedingstops ~ subject_sex+hour, data = speeding_stops_w)
cand.models[[3]] <- glm.nb(speedingstops ~ hour+district, data
                    = speeding_stops_w)
cand.models[[4]] <- glm.nb(speedingstops ~ subject_sex+district,
                    data = speeding_stops_w)
model.names <- c("Full", "sex/hour","hour/district","sex/district")
names(cand.models) <- model.names
results <- data.frame(models = model.names)
results$bic.val <- unlist(lapply(cand.models, BIC))
results$bic.rank <- rank(results$bic.val)
results$aic.val <- unlist(lapply(cand.models, AIC))
results$aic.delta <- results$aic.val-min(results$aic.val)
results$aic.likelihood <- exp(-0.5* results$aic.delta)
results$aic.weight <- results$aic.likelihood/sum(results$aic.likelihood)
results <- results[rev(order(results[, "aic.weight"])),]
results$cum.aic.weight <- cumsum(results[, "aic.weight"])
```

REFERENCES

Administration, N. H. T. S., 2008. Traffic safety facts: 2008 data. Washington, DC: National Highway Traffic Safety Administration..

Beaujean, A. A. & Grant, M. B., 2016. Tutorial on Using Regression Models with Count Outcomes using R. Volume 21.

Brooke, J., 1995. *Life Without Speed Limits: No Rush to the Fast Lane,* s.l.: New York Times.

Burnham, K. & Anderson, D., 2002. Model selection.

Cameron, A. & Trivedi, P., 2013. Regression analysis of count data. *Cambridge university press,* Volume 53.

Cohen, J., Cohen, P. & Stephen, G., 2003. Applied multiple regression/correlation analysis for the behavioral sciences. *West, and Leona S. Aiken,* Volume 3.

Cornwell, T., 1996. Bomber from the Backwoods?. *INDEPENDENT .*

David W, H. & Stanley, L., 2000. Apllied Logistic Regression. *Wiley Publishing.*

Dormann, C. F., 2016. Overdispersion, and how to deal with it in R and JAGS.

Gaffney, M., Hoard, S. A. & Eliason, S., 2013. *MHP Traffic Stop Data Analysis Project,* s.l.: Washington State Univeristy.

GELMAN, A. & Jeffrey, F., n.d. An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias.

HOEF1, J. M. V. & BOVENG, P. L., 2007. QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION. *Ecological Society of America,* 88(11), pp. 2766-2772.

King, R. & Sunstein, C. R., 1999. Doing Without Speed Reports. *Boston University Law Review,* 79(1), pp. 155-194.

Kowal, J., 1995. Sky's the Limit on Montana Highways. *TIMES UNION.*

Kutner, M., Nachtsheim, J. & J., N., 2005. A Review of:"Applied Linear Regression Models". *H.ayden, R.W,* xvii(4), p. 701.

Lange, J., Johnson, M. & Voas, R., 2005. Testing the racial profiling hypothesis for seemingly disparate traffic stops on the New Jersey Turnpike. *JusticeQ,* 22(2), pp. 193-223.

Lee, Y. & Nelder, J., 1998. Generalized linear models for the analysis of quality-improvement experiments. *Canadian Journal of Statistics,* 26(1), pp. 95-105.

Mueller, M., 1996. Montana's Wild Frontier Beckons Actors, Writers and Reactionaries. *BOSTON HERALD,* Volume 6.

Project, S. O. P., n.d. [Online]
Available at: https://openpolicing.stanford.edu/

Rodrıguez, G., 2013. Models for Count Data With Overdispersion. *Princeton Edu.*

Sharma, A. & Liu, C., 2019. Are you going to get a ticket or a warning for speeding? An autologistic regression analysis in Burlington, VT. *Transportation Research Interdisciplinary Perspectives,* Volume 1.

Sherman, R. & Funder, D., 2009. Evaluating correlations in studies of personality and behavior: Beyond the number of significant findings to be expected by chance. *Journal of Research in Personality,* 43(6), pp. 1053-1063.

Sousa, W. & Kelling, G., 2006. Of “broken windows,” criminology, and criminal justice. Police innovation: Contrasting perspectives. pp. 77-97.

Trexler, J. & Travis, J., 1993. Nontraditional regression analyses. *Ecology,* 74(6), pp. 1629-1637.