

Using Generalized Linear Regression models to Determine the effect of ride-sourcing demand on road crashes

Tong Zhu
Ria Kontou, PhD
University of North Carolina at Chapel Hill
Department of Statistics and Operation Research

October 29, 2019

Abstract

The ride-sourcing services, such as Uber and Lyft type services have become a popular discussion topic related to road crashes. There is evidence showing ride-sourcing is associated with the reduction of alcohol-involved crashes. At the same time, Uber/Lyft drivers crashing contributes to congested city centers and associates with an increase in crashes. After identifying some key research gaps related to the existing crash modeling such as lack of insightful modeling of crash rate, demographic, and vehicle ownership variables, the objective of this report is to narrow these gaps by systematically developing general linear regression models for road crashes modeling. The aim of this study is to determine the effect of ride-sourcing in Austin, TX. To reach the conclusion, this study implemented Poisson regression, Negative Binomial regression, and Zero-inflated Negative Binomial regression. The Zero-inflated Negative Binomial regression model fits the data the best, and all of these three models shows the period(ride-sourcing) factor is statistics significant, which means having ride-sourcing services increases the road crashes in Austin TX.

Keywords: Poisson regression model, Negative-Binomial regression model, Zero-inflated Negative Binomial regression model

Contents

1	Introduction	3
2	Data	3
2.1	Data structure	3
2.2	Data preprocessing	3
3	Data analysis and Modeling	6
3.1	Poisson regression model	6
3.2	Negative Binomial regression model	8
3.3	Zero-inflated Negative Binominal regression model	9
4	Summary	11

List of Tables

1	Variables description	3
2	Potential independent variables	6
3	Summary of Poisson regression	7
4	Result of over-dispersion test	7
5	Summary of Negative Binomial regression model	8
6	Vuong test output 1	9
7	Vuong test output 2	10
8	Summary of Zero-inflated Negative Binomial regression model	11

List of Figures

1	Distribution of crashes	4
2	Distribution of independent variables	4
3	The transformed distributions of Gasprice and MedHHInc	5
4	The transformed distributions of PopDensity and TotalTrips	5
5	Correlation between the independent potential variables	6
6	Fit of Zero-Inflated Negative Binomial model	10

1 Introduction

Uber and similar rideshare services are rapidly dispersing in cities across the United States and beyond. Research has shown own-demand services like Uber and Lyft provide safe rides home have reduced the number of traffic deaths associated with alcohol drinking. In cities, that is translating into more cars on the road. Research suggested it is linked with more congestion and traffic deaths, (see Barrios et al. 2018). Several studies determined the relationship between ride-sourcing entry and road fatalities. Some of them use the difference-in-differences method by comparing two groups over two time periods. The other studies use time-series forecasting (ARIMA) to analysis road traffic deaths. However, these studies did not consider demographic variables, such as income, population and employment density.

In this paper, the study developed Generalized Linear Models (GLM) approach that can be applied to crash data to determine the effect of ride-sourcing on the Austin TX case study. The methodology for this study consists of three steps. First, the description and processing of the data (section 2). Next, building statistical count models to the access road data; Poisson regression (section 3.1), Negative Binomial models account for overdispersion (section 3.2), and Zero-inflated Negative Binomial Poisson model(section 3.3) address the overabundance of sites with zero observations. Lastly, the study conclusions are presented.

2 Data

2.1 Data structure

The study approach for developing Generalized Linear Models (GLM) are demonstrated using a database of road crashes, census-tract data obtained from 218 locations of Austin from January 2012s to July 2017s. There are 8720 observations of 14 variables. Each row of the data represents a specific location in a specific time unit (week). The following table describes the definition of these variables:

Table 1: Variables description

Variables	Description
Geoid	Different local address code
Time	Month-year indicator
GasPrice	Gas price per gallon
Crashes	road crashes
Fatalities	All fatalities in crashes
Injuries	All the injuries in crashes
DUI	Arrests due to drunk driving
TotalTrips	Index capturing exposure to total travel in a region at certain time unit
MedHHInc	Median income
VehOwn0	Percentage of 0 vehicle ownership
PopDensity	Population density (people per square feet)
Airport	Airport location indicator
Period	1: No ridesourcing, 3: Ridesourcing

The time indicator is not correct in this database, therefore it will not be considered in this study. Because Fatalities, Injuries, and DUI only occur when road crashes happened, they will not be included as the independent variables. In this study, we would like to choose Crashes as dependent variable and choose GasPrice, TotalTrips, MedHHInc, VehOwn0, PopDensity, Airport, and Period as potential independent variables.

2.2 Data preprocessing

The figure 1 below shows the distribution of crashes. The crash frequency distribution demonstrates that the number of crashes ranges from a minimum of 0 to a maximum of 49 crashes over 6 years.

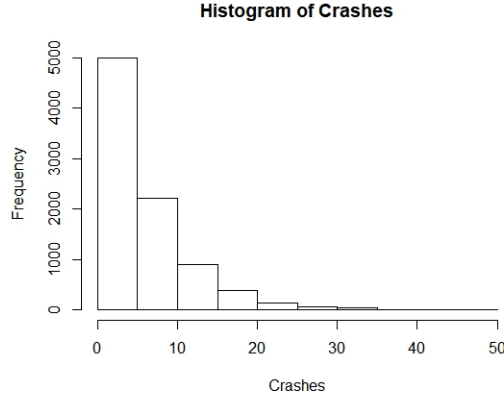


Figure 1: Distribution of crashes

The road crash is count data and it looks approximately close to Poisson distribution. From the plot above, we can see about half of the data has zero road crashes. The large number of zero crashes also led us to apply Zero-inflated Negative Binomial models to the modeling of the road crashes number considering two underlying processes. We further discuss this in section 3.3. The following is a brief visualization of each independent variables in the dataset.

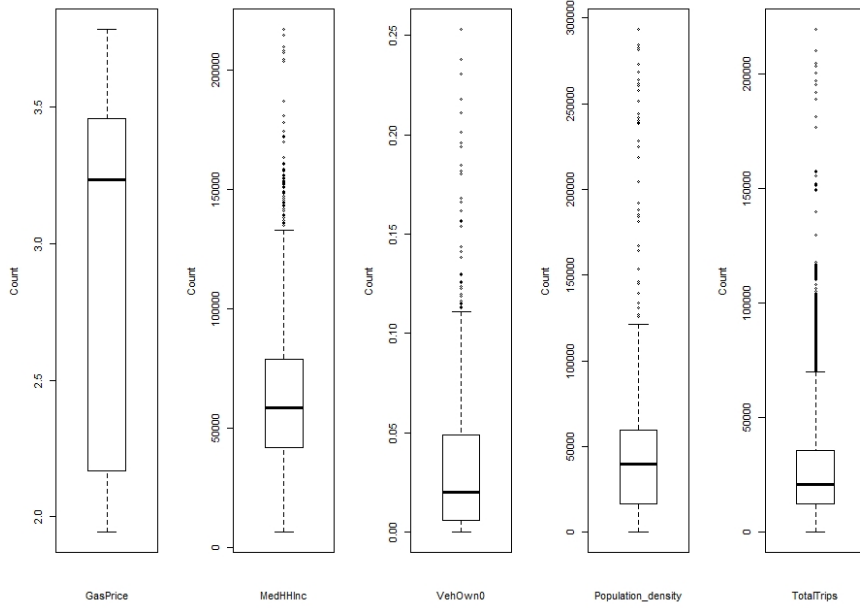


Figure 2: Distribution of independent variables

The plots in figure 2 are boxplots. A boxplot displays the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). The thick line on the middle of the box is the median value of the data, and the box represents interquartile range of the data. The outliers, which larger than $Q3 + 3 \cdot IQR$ or smaller than $Q1 - 3 \cdot IQR$, are described as dots in the plot.

Both Gas price variable and PopDensity variable skew to left, but only PopDensity has lots of outliers. The other three variables are right-skewed with many large outliers. To respond to the skewness towards large values in these variables, the logarithmic transformation, and square root transformation were taken into consideration. Here we chosed logarithm with base 10 which is useful in calculations. The logarithm of numbers greater than 1 that differ by a factor of a power of 10 all have the same fractional part.

The distributions of transformed data are shown in following figures.

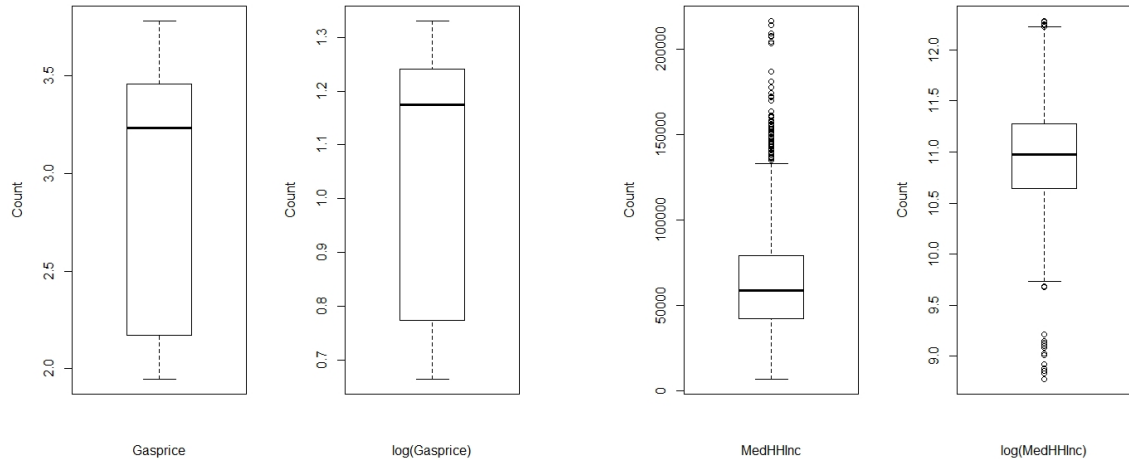


Figure 3: The transformed distributions of Gasprice and MedHHInc

After logarithmic transformation, even more data falls to the right side, but the MedHHInc variable is closer to normal distribution. Many large data in MedHHInc be scaled down. There are less outliers in MedHHInc variable after logarithmic transformation.

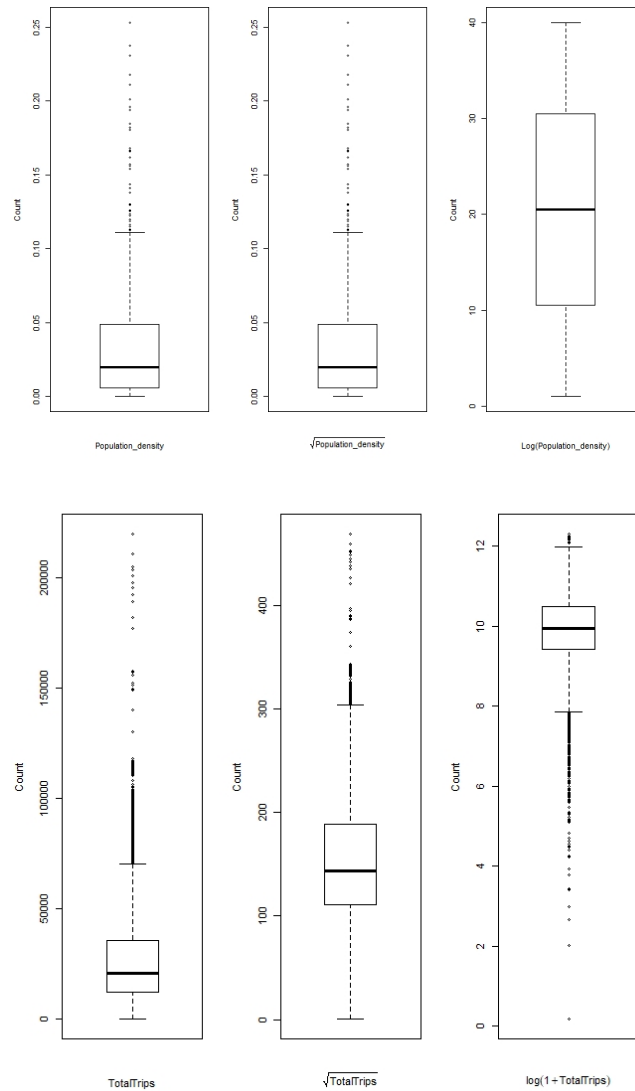


Figure 4: The transformed distributions of PopDensity and TotalTrips

Logarithmic transformation transformed the populationdensity variable into more normally distributed. There are lots of zero in TotalTrips and $\log(0)$ is undefined, so we added 1 to TotalTrips before logarithmic transformation. After transformation, the skewness of TotalTrips be improved a little bit. The transformation of independent variables are not necessary in generalized linear regression, but if we can scale the data into more normally distributed, it will improve the model processing. Here, we decided not to do any transformation on Gasprice variable. The potential independent variables are listed in the following table:

Table 2: Potential independent variables

Variables	Description
GasPrice	Gas price per gallon
Log(TotalTrips+1)	Log(Index capturing exposure to total travel in a region at certain time unit+1)
Log(MedHHInc)	Log(Median income)
VehOwn0	Percentage of 0 vehicle ownership
LogPopDensity	Log(population density (people per square feet))
Airport	Airport location indicator
Period	1: No ridesourcing, 3: Ridesourcing

Whenever two supposedly independent variables are highly correlated, it will be difficult to assess their relative importance in determining some independent variables. Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when we fit the model and interpret the results. The correlations coefficients between the possible pairs of these potential independent variables are shown below.

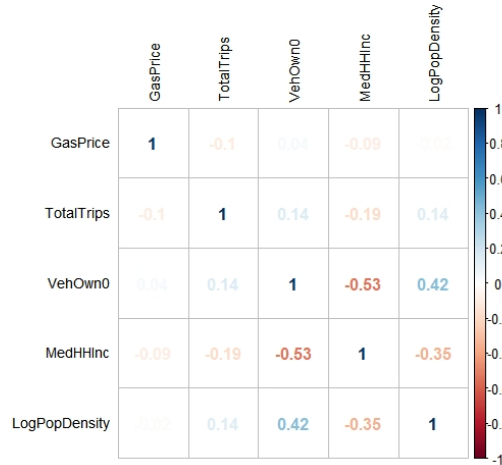


Figure 5: Correlation between the independent potential variables

Positive correlations are displayed in blue and negative correlations in red color. The color intensity and the number represent the correlation coefficients. There is no strong correlation between any of these independent variables(i.e., correlation of 0.6 or more). We do not need to worry about the multicollinearity in regression.

3 Data analysis and Modeling

3.1 Poisson regression model

The Poisson distribution is more frequently applied to models with count data. Poisson regression assumes the response variable Y has a Poisson distribution and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. In this case, we assumed the road crash followed Poisson distribution. The probability mass function of a Poisson crash random variable is given by

$$P(y_i) = \frac{e^{-\mu_i} \mu_i^y}{y!}, \quad y = 0, 1, 2, \dots$$

where,

$P(y_i)$: the probability of y crashes occurring on location i during a period of time

μ_i : the expected crash frequency on location i

A crash frequency analysis was conducted on the data-set using the Poisson regression model, and the results of the Poisson model fit were obtained by using the R software as shown in Table 3.

Table 3: Summary of Poisson regression

	Estimated coefficient	P-value
GasPrice	0.103	1.11e-07
Log(TotalTrips+1)	0.507	6.71-e05
Log(MedHHInc)	-0.137	2e-16
VehOwn0	1.53	2e-16
LogPopDensity	-0.120	2e-16
Airport	-0.189	0.0035
Period	0.206	4.55e-09

The positive sign for a estimated coefficient indicates that road crashes will increase as the value for the variables related to that parameter increase. Similarly, the negative sign indicates this factor will decrease the road crashes, while keeping other factors the same. All of these 7 variables' p-value are less than the statistic threshold (0.05). Because we've tested 7 different potential predictors, the likelihood of finding an erroneous significant effect (purely by random chance) has now ballooned to approximately 0.3. After using Bonferroni's method, we got new threshold of significance (p-value $< .007 = 0.05/7$), maintaining our 95 confidence in our set of analyses as a whole. The VehOwn0 has the largest estimated coefficient and smallest p-value which indicates it is the most significant factor with biggest positive impact on the road crashes. GasPrice variable has relative bigger p-value compare to other variables, and it is not significant after we got the new threshold of significance (p-value $< .007$). The period predictor is the factor we want to analysis, and it represent whether there is ride-sourcing. The period factor is statistic significant and the positive estimated coefficient means that when there exist ride-sourcing in this period, the expected log count of the number of road crashes increases by 1.937.

The Poisson model assumes that the mean equals the variance, and hence can't handle the over-dispersion nature of the crash data when the variance exceeds the mean. In statistics, overdispersion is the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model. We tested whether the observed variance is higher than the variance of this Poisson model. This test was applied using "dispersiontest" function in R-statistic software (version 3.4.) package AER. The dispersion test assesses the hypothesis that the mean equals the variance holds against the alternative that the variance is of the form:

$$Var(y) = (1 + \alpha)\mu = dispersion * \mu$$

Over-dispersion corresponds to $\alpha > 0$ and under-dispersion to $\alpha < 0$. The coefficient α can be estimated by an auxiliary ordinary least squares (OLS) regression and tested with the corresponding Z statistic which is asymptotically standard normal under the null hypothesis.

Table 4: Result of over-dispersion test

α	P-value
3.632	2.2e-16

Here we clearly see that there is evidence of over-dispersion ($\alpha > 1$) which speaks quite strongly against the assumption of equidispersion (i.e. $\alpha=0$). It indicates that the Poisson regression model might not be well suited for these data, apparently because of the over-dispersion in the data count, that cannot be handled effectively by the Poisson regression. One way to deal with the over-dispersed count data is to apply Negative Binomial regression (Washington et al., 2011).

3.2 Negative Binomial regression model

Negative Binomial model is a generalization of Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model. The Negative Binomial model is derived by rewriting the Poisson parameter for each observation i as $\mu_i = e^{\beta X_i + \epsilon_i}$ where e^{ϵ_i} is a gamma-distribution error term with mean 1 and variance α . The addition of this term allows the variance to differ from the mean as:

$$Var[y_i] = E[y_i][1 + \alpha E[y_i]]$$

The Poisson regression model is a limiting model of the Negative Binomial regression model as α approaches zero, which means the variance and mean are the same (Lord and Mannering, 2010). The Negative Binomial model can be expressed as:

$$P(y_i) = \left(\frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \right) \left(\frac{1}{1 + \alpha \mu} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha \mu}{(1 + \alpha \mu)} \right)^{y_i}$$

where,

$$\text{mean: } \mu = Var(e^{x\beta})$$

$$\text{Variance: } Var(y_i) = \mu + \alpha \mu^2$$

The following table shows the summary of Negative Binomial regression model for road crashes:

Table 5: Summary of Negative Binomial regression model

	Estimated coefficient	P-value
GasPrice	0.160	0.001
Log(TotalTrips+1)	0.009	2e-16
Log(MedHHInc)	-0.192	2e-16
VehOwn0	2.482	2e-16
LogPopDensity	-0.139	2e-16
Airport	0.145	0.1249
Period	0.276	3.3e-05

There are 6 most important factors for both models: GasPrice, Log(TotalTrips+1), Log(MedHHInc), VehOwn0, LogPopDensity and Period (ride-sourcing). Log(TotalTrips+1) and VehOwn0 has positive effects on road crashes, which is in accord with common sense. Less vehicle ownership means less cars on the road which will decrease the road crashes. More trips and more cars associate with more road crashes. Conversely, MedHHInc have negative effects on road crashes, which means accident likelihood is greater on low MedHHInc area. One possible reason for this could be high-income people drive more safely. In particular, the variable period (ride-sourcing) has a coefficient of 0.276, which is statistically significant. This means that when there exist ride-sourcing in this period, the expected log count of the number of road crashes increases by 0.276.

Here, a Vuong test is considered for the model selection between Poisson regression model and Negative Binomial regression model. Proposed by Quang Vuong (1989), this test considers a better model with the individual log-likelihoods significantly higher than the ones of its rival by using the Kullback–Leibler information criterion. With the likelihood of each individual record from both models, Vuong test is calculated with the formulation given below:

$$Vuong \text{ statistics} = \frac{LR(model1, model2 - C)}{\sqrt{N * V}} \quad N(0, 1)$$

where,

LR(...): the summation of the individual log-likelihood ratio between 2 models.

C: a correction term for the difference of DF (Degrees of Freedom) between 2 models.

N: is the number of records.

V: is the variance of the individual log-likelihood ratio between 2 models.

The null hypothesis is that both classes of distributions are equally far from the true distribution. If this is true, the log-likelihood ratio should (asymptotically) have a Normal distribution with mean zero. If the null hypothesis is false, and one class of distributions is closer to the "truth", the test statistic goes to +/-infinity with probability 1, indicating the better-fitting class of distributions (Vuong, Q.H. 1989). The output of Vuong test between Poisson regression model (PM) and Negative-Binomial regression model (NBM) is given below:

Table 6: Vuong test output 1

	H_A	P-value
Raw	NBM > PM	2.22e-16
AIC-corrected	NBM > PM	2.22e-16
BIC-corrected	NBM > PM	2.22e-16

A large, positive test statistic provides evidence of the superiority of Negative Binomial model over Poisson model (i.e. P-value < 0.05). The corrected Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are information-based criteria that assess model fit. They are both methods of assessing model fit penalized for the number of estimated parameters. Each of the three tests performs similarly if the Negative-Binomial model fits the data better. Either using the raw results or using the AIC- or BIC-corrected results, we can conclude These three tests provide the same result which give strong evidence that the Negative Binomial model is much better than the Poisson model.

The Negative Binomial regression model is not the only alternative option that allows for over-dispersion. The Zero-inflated Negative Binomial model also allows for over-dispersion. The over-dispersion could also be caused by the excess zeros by an additional data generating process. In this situation, Zero-inflated model should be considered since road crashes has a large number of zero counts.

3.3 Zero-inflated Negative Binominal regression model

Zero-inflated Negative Binomial regression is for modeling count variables with excessive zeros and it is usually for overdispersed count outcome variables. The model assumes that excess zeros are due to two different processes. A logit model to model which of the two processes the zero outcome is associated with and a count model. In this case, a Negative Binomial model is used to model the count process. The expression of the likelihood function depends on whether the observed value is a zero or greater than zero. From the logistic model of $y_i > 1$ versus $y_i = 0$:

$$p = \frac{1}{1 + e^{-x'_i \beta}}$$

and

$$1-p = \frac{1}{1 + e^{x'_i \beta}}$$

then

$$\mathcal{L} = \begin{cases} \sum_{i=1}^n \left[\ln(p_i) + (1-p_i) \left(\frac{1}{1+\alpha\mu_i} \right)^{\frac{1}{\alpha}} \right] & \text{if } y_i = 0 \\ \sum_{i=1}^n \left[\ln(p_i) + \ln \Gamma \left(\frac{1}{\alpha} + y_i \right) - \ln \Gamma(y_i + 1) - \ln \Gamma \left(\frac{1}{\alpha} \right) + \left(\frac{1}{\alpha} \right) \ln \left(\frac{1}{1+\alpha\mu_i} \right) + y_i \ln \left(1 - \frac{1}{1+\alpha\mu_i} \right) \right] & \text{if } y_i > 0 \end{cases}$$

We use the pscl package in R to run a Zero-inflated Negative Binomial regression.

The Vuong test between Zero-inflated Negative Binomial model (ZNBM) and negative binomial model (NBM) suggests that the Zero-inflated Negative Binomial model is a significant improvement over a standard Negative Binomial model.

Table 7: Vuong test output 2

	H_A	P-value
Raw	ZNBM > NBM	0.001
AIC-corrected	ZNBM > NBM	0.013
BIC-corrected	ZNBM > NBM	0.33

The large positive test statistic provides evidence of the superiority of Zero-inflated Negative Binomial model over than ordinary Negative Binomial model under the null that the models are indistinguishable (i.e. P-value < 0.05). Under BIC-corrected test, the Zero-inflated Negative Binomial model is not significant better than the ordinary Negative Binomial regression model. It can be explained by if we include less significant variables in the model, the ordinary Binomial model will fit the data as well as the Zero-inflated Negative Binomial model.

To judge the goodness of fit of a model with estimated parameters to observations, a natural idea is to assess whether observed frequencies match expected frequencies from the model. To visualize the goodness of fit of the Zero-inflated Negative Binomial model, the "rootogram" function from R package countreg was applied. The rootogram compares observed and expected values graphically by plotting histogram-like rectangles or bars for the observed frequencies and a curve for the fitted frequencies, all on a square-root scale.

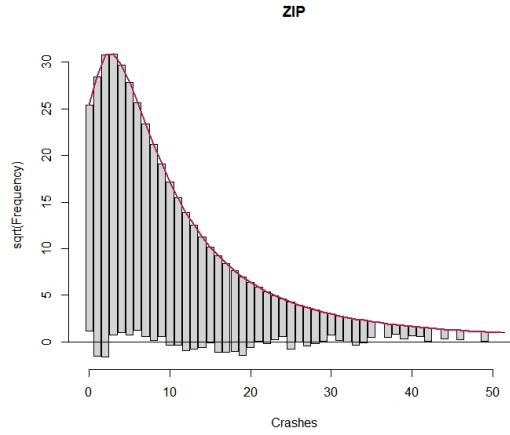


Figure 6: Fit of Zero-Inflated Negative Binomial model

Expected counts, given Zero-inflated Negative Binominal model, are shown by the thick red line, observed counts are shown as bars, which in a hanging rootogram are show hanging from the red line of expected counts. On the x-axis we have the count bin, 0 count, 10 count, 20 count, etc, on the y-axis we have the square root of the observed or expected count. If a bar does not reach the zero line then the model over predicts a particular count bin, and if the bar exceeds the zero line it under predicts. Here, the curve representing expected frequencies closely tracks the histogram representing observed frequencies, there are also no clear patterns in the hanging and suspended versions. All this indicates that the model fits well.

The Table 8 in the next page describes the count part of Zero-inflated Negative Binomial regression coefficients for each of the variables along with standard errors, z-scores, and p-values for the coefficients.

Comparing to ordinary Negative Binomial regression model, We got the same important factors from Zero-inflated Binomial model. All of these 6 factors in the count part of the Zero-inflated Negative Binomial regression model predicting number of road crashes (count) are significant predictors (i.e. P-value .0083 = 0.05/6)). The positive coefficient represents there exist positive correlation between the factor with the road crashes. It is not surprising that more total trips (TotalTrips) and the higher Percentage of zero vehicle ownership (VehOwn0) are associated with more road crashes in all three models. It might be explained by that the road crashes are less likely occur in residential areas. The residential area have low percentage of zero vehicle ownership and high traffic flows. In particular, the ride-sourcing (Period) also positive related to the road crashes in all three models. Among those exist road crashes, having ride-sourcing (Period) increases the expected rate of road crashed by $0.32 = e^{0.28} - 1$, holding other variables constant.

Table 8: Summary of Zero-inflated Negative Binomial regression model

Count model coefficients		
	Estimated coefficient	P-value
Intercept	0.375	0.254
GasPrice	0.169	0.0005
Log(TotalTrips+1)	0.406	2e-16
Log(MedHHInc)	-0.181	2e-16
VehOwn0	2.150	2e-16
LogPopDensity	-0.128	2e-16
Period	0.280	2.91e-05
Log(theta)	0.948	2e-16

4 Summary

The objectives of this research is to determine the effect of ride-sourcing in Austin, TX. This study approaches this by using Poisson regression model, Negative Binomial regression model and Zero-inflated Negative Binomial regression model. These models relate road crashes to the potential contributing factors. The results show the Zero-inflated Negative-Binomial fits the data very well, and all of these three models captured explanatory variables that have significant effects on road crashes. The ride-sourcing variable (Period) significantly impacts road crashes in Austin, TX.

In this study we did not include seasonal factors in the model. There are more vehicles on roads in summer months than in winter months. In future work, it would be worth extending the analysis to include seasonality in road crashes.

References

- Vuong, Q.H. 1989. "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica*.
- Greene, William H. (2003). *Econometric Analysis* (Fifth ed.). Prentice-Hall.
- Lambert, Diane (1992). "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing". *Technometrics*.
- Hilbe, J. M. (2007). *Negative Binomial Regression*. Cambridge University Press.
- Aho, K.; Derryberry, D.; Peterson, T. (2014), "Model selection for ecologists: the worldviews of AIC and BIC", *Ecology*.
- Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications. Everitt, B. S. and Hothorn, T. *A Handbook of Statistical Analyses Using R*.
- Functions, data sets, examples, demos, and vignettes for the book Christian Kleiber and Achim Zeileis (2008), *Applied Econometrics with R*, Springer-Verlag, New York.
- Kleiber, C., and Zeileis, A. (2016). Visualizing count data regressions using rootograms. Available at: <http://arxiv.org/abs/1605.01311>.