2018

EFFECT OF SOCIOECONOMIC AND DEMOGRAPHIC FACTORS ON KENTUCKY CRASHES

Aaron Berry Cambron
University of Kentucky, aaron.cambron@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0002-0022-5846

Digital Object Identifier: https://doi.org/10.13023/etd.2018.482

EFFECT OF SOCIOECONOMIC AND DEMOGRAPHIC FACTORS ON KENTUCKY CRASHES

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in Civil Engineering in the
College of Engineering
at the University of Kentucky

By

Aaron Cambron

Lexington, Kentucky

Director: Dr. Nikiforos Stamatiadis, Professor of Civil Engineering

Lexington, Kentucky

2018

https://orcid.org/0000-0002-0022-5846

ABSTRACT OF THESIS

EFFECT OF SOCIOECONOMIC AND DEMOGRAPHIC FACTORS ON KENTUCKY CRASHES

The goal of this research was to examine the potential predictive ability of socioeconomic and demographic data for drivers on Kentucky crash occurrence. Identifying unique background characteristics of at-fault drivers that contribute to crash rates and crash severity may lead to improved and more specific interventions to reduce the negative impacts of motor vehicle crashes. The driver-residence zip code was used as a spatial unit to connect five years of Kentucky crash data with socioeconomic factors from the U.S. Census, such as income, employment, education, age, and others, along with terrain and vehicle age. At-fault driver crash counts, normalized over the driving population, were used as the dependent variable in a multivariate linear regression to model socioeconomic variables and their relationship with motor vehicle crashes. The final model consisted of nine socioeconomic and demographic variables and resulted in a R-square of 0.279, which indicates linear correlation but a lack of strong predicting power. The model resulted in both positive and negative correlations of socioeconomic variables with crash rates. Positive associations were found with the terrain index (a composite measure of road curviness), travel time, high school graduation and vehicle age. Negative associations were found with younger drivers, unemployment, college education, and terrain difference, which considers the terrain index at the driver residence and crash location. Further research seems to be warranted to fully understand the role that socioeconomic and demographic characteristics play in driving behavior and crash risk.

KEYWORDS: Multilinear Regression, Crash Rate, Socioeconomics, Demographics, At-Fault, Zip Code

Aaron Cambron

11/05/2018

Date

EFFECT OF SOCIOECONOMIC AND DEMOGRAPHIC FACTORS ON KENTUCKY CRASHES

By

Aaron Cambron

Nikiforos Stamatiadis

Director of Thesis

Timothy Taylor

Director of Graduate Studies

11/05/2018

Date

DEDICATION

To *procrastination*, my constant companion of 23 years, without whom this report may have been done a bit sooner.

ACKNOWLEDGMENTS

While ultimately my own work, this thesis is the product of many additional ideas, insights, direction, and assistance from those around me. First, I would like to thank my Thesis Chair, Dr. Nikiforos Stamatiadis, without whom I may not have turned out this well. Besides being a mouthful to say, Dr. Stamatiadis has meant a great deal to me over my five and a half year tenure at the University of Kentucky. He has provided constant guidance and advice, has kept me on track and on task, and has bestowed onto me a small piece of his vast world knowledge.

Next, I would like to thank the rest of my thesis committee, Dr. Reg Souleyrette and Dr. Greg Erhardt, both of whom have provided much needed motivation and guidance. I would also like to thank my co-workers Dr. Eric Green, Erin Lammers, and William Staats for providing technical advice and sympathetic ears.

I would like to thank everyone at the University of Kentucky who has befriended, guided, supported, or annoyed me over the past five years. I am the product of our relationships.

Finally, I would like to thank my family: my mother and father for all that I have and all that I am, my brother and sister for their love and support, and my girlfriend Katie who really learned to love me as the busy and procrastinating grad student that I've become.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Vehicle crashes are a major cause of injuries and fatalities in the U.S. and around the world even though roadway safety has been improving globally for the several decades. The World Health Organization (WHO) estimates that more than 1.25 million people are killed each year as the result of road traffic crashes (WHO 2018). In the United States alone, there were 37,461 traffic related fatalities in 2016, a 5.6 percent increase since 2015 (NHTSA 2016). Crash data from Kentucky indicate a higher percent increase from 2015 than the national average – an increase of 10 percent from 761 to 834 fatalities (KTC 2016). In addition, Kentucky has a higher overall crash rate per population than the national average. In 2016, the National Highway Traffic Safety Administration (NHTSA) estimated 22.5 crashes per 1,000 residence population for the country, while Kentucky had a rate of 37.3. NHTSA also estimated that all traffic crashes in 2010 incurred a comprehensive cost of $836 Billion on the U.S. economy (NHTSA 2016).

Prior research has shown that the highest contributing factors to crashes are related to human aspects (AASHTO 2010). Even though a relatively small percentage of motor vehicle crashes are related to highway conditions, safely designed highways could lessen the severity of injuries when crashes occur. The vehicle is not a contributing factor in many crashes; however, more safely designed vehicles have had beneficial effects in reducing injuries resulting from crashes. Various programs targeting safety improvements, such as increased seat belt use and alcohol enforcement, have been recently implemented throughout the United States, but may not have the same effectiveness in Kentucky. However, it appears that fewer benefits have been gained as a result of those programs in the Southeast overall (Stamatiadis and Puccini 1999). Moreover, it is reasonable to believe that design standards are in general similar among the various states, and thus these differences may be attributed to other factors.

A plausible explanation of the increased crash rates in Kentucky may be the differences in a variety of socioeconomic characteristics of the state compared with other states. Based on statistics from the Bureau of the Census, Kentucky has lower percentages of high school completion and university attainment than the national average (U.S. Census Bureau 2018). With respect to income characteristics, most of the counties have a median family income 19 percent lower than the national median income, they are at the bottom of the national rankings with respect to both income and disposable income per capita, and they have one of the largest percentages among the states of persons below the poverty level. These types of socioeconomic characteristics could influence highway safety by affecting the age of vehicles owned (older, less safe vehicles), the condition of these vehicles (not properly maintained), the attitudes of the drivers toward safety and risk-taking behaviors, and the level of driving education available to people (Stamatiadis and Puccini 1999).

At the same time, the fact that most areas in the region are considered rural areas may also contribute to these increased rates. About 42 percent of the Kentucky population is classified as rural, compared to the national average of 27 percent. (U.S. Census Bureau 2018). Historical crash data indicate that the fatality rates are twice as high in rural as in urban areas (KTC 2016). Thus, the higher fatality rates in rural areas combined with the larger rural area population in Kentucky may also help explain these higher crash rates.

This research seeks to examine the impacts of socioeconomic and demographic factors on at-fault drivers in the state of Kentucky. Crash data will be obtained from the Kentucky State Police and demographic information of drivers will be acquired from the U.S. Census Bureau's American Community Survey linked to the residence zip code of the driver. A multivariate linear regression will be used to identify significant variables as they relate to at-fault driver crash rates in each zip code.

CHAPTER 2. LITERATURE REVIEW

A significant research effort has been undertaken globally to investigate the role and possible contribution of socioeconomic and demographic factors on crash occurrence. Some of the methods investigate demographics surrounding the crash location, while others use surrogate descriptors associated with residence location of the drivers involved in a crash. The following sections discuss past research efforts to identify significant socioeconomic and demographic factors that could explain crash involvement as well as methods used to investigate them.

2.1 Socioeconomic and Demographic Variables

Various socioeconomic and demographic variables have been examined in the past to identify their potential contribution on crash occurrence. Prior research shows some common threads among explanatory variables, which agree with a priori expectations: income, poverty, employment, education, rurality, and driver age all seem to have an impact. Hasselberg et al. (2005) determined that drivers with a relatively low educational attainment level show an excess risk of both road-traffic crashes and of crashes leading to fatality or serious injury. Their study also estimated that 33 percent of minor injuries and 53 percent of severe injuries would be avoided if all subjects had the same injury rate as subjects with a higher education. Similarly, Zephaniah et al. (2018) concluded that "a more educated population contributes to a lower DUI crash rate for a given postal code." Their study found that the percentage of college educated women and the overall percentage of residents with at least a high school education in a postal code reduced the occurrences of DUI crashes. These two studies used the characteristics of the driver's residence location and showed that a higher education has a positive impact, i.e., reduction, on vehicle crashes. Conversely, Lourens et al. (1999) conducted a study based on data from a traffic survey conducted each year in the Netherlands. Lourens found that while adjusting for annual mileage of a driver, education did not have a significant impact on accident involvement.

Both income and poverty were cited as relevant predictors for crash related analysis from several sources. It should be noted though that income and poverty could be closely related, as poverty status is generally based on income below a certain level. Lee et al. (2014) investigated the relationship between at-fault driver residence characteristics and all types of crashes for three years of data in Florida. They found that Median Family Income had a negative relationship with the number of at-fault drivers, indicating that drivers from lower income communities are more likely to be responsible for a crash occurrence. Maciag (2014) indicated that within metro areas, low-income tracts recorded pedestrian fatality rates approximately twice that of more affluent neighborhoods; high poverty rate tracts revealed a similar trend. Aguero-Valverde et al. (2005) also concluded that percent of population under the poverty level had a highly significant and positive correlation with crash risk when using a negative binomial prediction model. In contrast to this, Noland and Laham (2018) concluded that income did not have an impact on the probability of dying from a motor vehicle crash. This was determined with data from the National Longitudinal Mortality Study.

Rural areas are generally cited as having higher fatality crash rates than urban areas and a large portion of previous research dealt with the levels of rural and urban components of a region. Muelleman et al. (1986) investigated fatal motor vehicle crash characteristics to determine the crash characteristics related to population density. They concluded that the fatality rate per 100 million Vehicle Miles Traveled (VMT) was 44 percent higher ($p < 0.001$) in rural than urban areas. They also noted that rural areas are not homogeneous and comparisons based only on urban/rural groupings can obscure important differences between urban and rural areas as well as variations within rural areas. Similarly, Zwerling et al. (2005) found that fatal crash incidence density was more than two times higher in rural than in urban areas.

Employment has been cited in several forms either as unemployment rates, portion of people working from home or portion of unskilled workers. Adanu et al. (2017) found that unemployed drivers were shown to have a probability of 0.23 of being at-fault in a crash and the probability of being at-fault in a serious injury crash was 0.57. They suggested that the odds of an unemployed driver being at-fault for a serious crash were 1.32 times higher than a driver who was employed, self-employed or retired. Factor et al. (2008) used a sample of the Israeli population with detailed socioeconomic data and nine years of crash data for their analysis. They found that non-skilled workers are over-involved in fatal crashes relative to their size in the total population of all workers. Conversely, Lee et al. (2014) found that the higher proportion of the population working from home resulted in a lower number of at-fault drivers, though it was proposed that this is the result of travel exposure. Finally, Noland and Quddus (2004) examined traffic fatalities in England aggregated to the census ward. They determined that areas of higher employment density (employed people per land area) actually resulted in more traffic casualties. They proposed that this is most likely due to an increased level of street activity in higher employment dense areas, resulting in more pedestrian related crashes.

Chen et al. (2010), Factor et al. (2008), and Hanna et al. (2012) all indicated that undesirable crash results, such as more crashes or higher fatality rates, were present for young or new drivers, but there was some variation about the impact of elderly drivers. Lee et al. (2014) determined that a larger proportion of elderly population decreases the likelihood of drivers being at-fault, while Aguero-Valverde et al. (2005) concluded that age groups below 25 and over 65 have a positive association with crash risk. Kim et al. (1998) found that young drivers are more likely to be classified as at-fault, and more specifically that young (less than age 25) males driving pickup trucks were the most likely subgroup to be at-fault in a crash.

Additional variables identified as significant include driving under the influence (DUI)/impaired driving (Adanu et al. 2018; La Torre et al. 2007; Muelleman et al. 1986), commute time to work (Adanu et al. 2017; Lee et al. 2014), and marital status (Factor et al. 2008).

2.2 Analysis Methods

The negative binomial distribution is a discreet probability distribution which is often used when dealing with crash counts and negative binomial regressions are used to model crash counts for a roadway segment. Noland and Quddus (2004) used negative binomial count data models to analyze the associations between demographic factors (such as land use types, road characteristics and area-wide demographics including the level of social deprivation) with traffic fatalities and serious or slight injuries. The social deprivation is an index developed in the United Kingdom consisting of six socioeconomic factors: income, employment, health deprivation and disability, education skills and training, housing, and geographical access to services. They used the census block in England as a spatial unit of the crash location in order to connect these demographics with crash fatalities. More recently, the Highway Safety Manual recommends developing Safety Performance Functions (SPFs) using negative binomial regressions which are primarily based on Average Annual Daily Traffic (AADT) for homogeneous roadway segments. However, Ivan et al. (2016) demonstrated an alternative in predicting crashes on local roads where the traffic volumes are not available. The study estimated SPFs for local road intersections and segments at the Traffic Analysis Zone (TAZ) level using socio-demographic and network topological data. There are approximately 1,800 TAZs in Connecticut which were then clustered into six analysis groups based on land use and population density. SPFs were developed using Poisson regression models which can predict intersection and segment crashes within each TAZ using the number of intersections and the total local roadway length, respectively.

Various other forms of regression modeling have been used in crash analysis. La Torre et al. (2007) and Rivas-Ruiz et al. (2007) used multiple linear regression in their analysis. La Torre et al. (2007) investigated the association between regional differences in traffic crash mortality case fatality and crash rates with socio-demographic factors and variables describing road behavior, vehicles, infrastructure and medical care in Italy, while Rivas-Ruiz et al. (2007) utilized a backwards stepwise elimination approach to study the variability of Road Traffic Injury (RTI) mortality on Spanish roads adjusted for Vehicle Kilometers Traveled in each Spanish province. Both studies found some significance in area wide socioeconomic factors, such as employment rates, alcohol use, and education levels.

Some have found other regression models to be more useful such as logistic and lognormal regressions. The logistic regression is typically used to describe a discrete variable. Noland and Laham (2018) used a multinomial logit model in conjunction with data from the National Longitudinal Mortality Study to determine the effects of socioeconomics on the probability for a motor vehicle fatality in comparison to other causes of death. Factor et al. (2008) created a binary response variable to describe crash fatality level. The model used demographic factors to predict the probability of being involved in a fatal crash versus a non-fatal crash. The research linked nine years of injury and fatal road-crash records with census data and used several socioeconomic factors all grouped into discrete categories, such as gender, education groups, and age groups. Similarly, Hanna et al. (2012) considered fatal crashes involving unlicensed young drivers (under age 19) in the United States using conditional and unconditional logistic modeling. This analysis was based on the urbanicity (which categorizes all US counties as urban, suburban or rural based on population and proximity to metropolitan areas) and the Townsend Index of Relative Material Deprivation (which serves as a proxy measure for socioeconomic status based on access to local goods, services, resources and amenities.)

To allow for the simultaneous study of driver characteristics and region information, Adanu et al. (2017) used multilevel logistic modeling, which recognizes "the hierarchical structure in data and also provide[s] information to compute the amount of variability in the data attributable to each level of the hierarchy." They also created a binary response variable which identifies the crashes as fatal or non-fatal. Kim et al. (1998) used a binary response variable to study the fault of a driver. They employed log-linear modeling which allows for the analysis of joint relationships among categorical variables, in this case age, gender, and vehicle type. As opposed to a logistic regression which relies on an additive model, the log-linear allows for a multiplicative model which improves the comparisons of variable combinations.

Other methods such as spatial analysis have also been used in crash analysis utilizing socioeconomic factors. Brown et al. (2016) considered the residential locations of at-risk drivers (drivers reported as contributing to fatal crashes) and the demographic characteristics associated with those residential locations at the Census Block Group level. Higher risk block group (more than 8 at-risk drivers per 1,000 driving population) socioeconomic variables were compared to those of lower risk groups to determine trends. This study used a cluster analysis creating hot spots of high or low risk areas that can be targeted for specific safety programs. Of note here is the fact that this study examined demographic characteristics tied to the driver's home location instead of the commonly used method of socioeconomic characteristics tied to the crash location. Kocatepe et al. (2017) used hotspots to investigate the exposure of different age groups to severe injury crashes in the Tampa Bay region. The severity-weighted crash hotspots were identified using Getis-Ord Gi method weighted by the number of severely injured occupants involved in each crash. The study examined the proximity of residents in different age groups (17 and younger, 18 to 21, 22 to 64 and 65 and older) to severity-weighted crash hotspots. The results revealed that age, ethnicity, education, poverty level, and vehicle ownership have an effect on crash injury exposure.

Finally, a less defined but seemingly widely used method for this type of research simply involves separating crash or socioeconomic data into groups and comparing them with descriptive statistics. Abdalla et al. (1997) studied the effect of driver social circumstance on crash occurrence and casualty by linking crash records and census data in the Lothian Region, Scotland. The research showed a correlation between fatal crashes and a driver's distance from home. Socioeconomic variables were bundled into a Deprivation Index and postal codes were separated into the most affluent and most deprived in order to compare traffic casualties normalized by population. Similarly, Blatt et al. (1997) considered fatal crashes occurring in rural areas, with a focus on the residential location of the driver. Five years of crash data from the Fatality Analysis Reporting System (FARS) was linked with driver home zip code and other factors, including driver age, gender and blood alcohol concentration. Five levels of population density were identified for classifying each driver's residence location, including rural, small town, second city, suburban, and urban; other driver characteristics were divided into social clusters (age groups, for example). Using geodemographic analysis, the percentage of drivers in fatal crashes in each social cluster was compared to the base population of that social cluster. In additional research involving traffic fatalities, Maciag (2014) investigated the differences in demographics of census tracts in relation to pedestrian fatalities in that tract. Census tracts were broken into categories by income and poverty to allow for a direct comparison of pedestrian fatalities.

2.3 Summary

In conclusion, the most prominent factors that seem to be relevant to crash occurrence investigation are income, education level, poverty percentage, employment, driver's age, and the rurality of an area. Education and income are typically negatively correlated with crash response, while poverty is positively correlated, and employment varies across studies. Young drivers, and areas with a high proportion of young drivers, tend to have a higher proportion of crashes and fatalities, and in general crashes in more rural areas seem to be more fatal.

To investigate the role of these factors on crash occurrence, many different methods have been used, and while all of the considered methods are valid, there is still a wide range of analysis practices for relating socioeconomic characteristics with crash data. Many forms of regression techniques have been applied, as well as spatial statistics, clustering, and comparative grouping. Most of the relevant research has examined fatal crashes in some form and the socioeconomic variables in question often pertain to the location of the crash rather than the residential location of the driver.

CHAPTER 3. DATA

To examine the characteristics of drivers involved in crashes, two types of data are required: the number of at-fault drivers aggregated to a spatial unit, and the socioeconomic and demographic characteristics associated with that unit. The smallest common unit that can be easily obtained for these two datasets in Kentucky is the zip-code. The Kentucky State Police (KSP) record the drivers' 5-digit home zip code for each crash, and the US Census Bureau aggregates the American Community Survey to the zip code level as well. For each Kentucky zip code, a database record can be developed that contains the pertinent demographic and socioeconomic data augmented with the number of at-fault drivers involved in a crash from that zip code. Once crash data was obtained from KSP, driver records were pared down to those occurring in valid Kentucky zip codes, of which there are 960. These records along with the American Community Survey were all linked into a single database for analysis.

A total of 463,116 crashes occurred in Kentucky between 2012-2016 in which a driver could be identified as at-fault and a valid Kentucky zip code was recorded for that driver's residence. Comparatively, the full database of driver crash records with Kentucky zip codes in that time span contained 751,634 records in total. The at-fault driver is identified through use of the Human Factor Code in the police report (Stamatiadis and Deacon 1997; Chandraratna and Stamatiadis 2007). If the human factor code is recorded for only one driver in a crash, that driver is most likely the main contributor to the crash as determined by the police officer. This method includes all single vehicle crashes and all the at-fault drivers for multivehicle crashes. For multivehicle crashes in which a human factor code was recorded for neither or both drivers, these driver records were eliminated from the analysis. Lee et al. (2014) used the police citation record to determine at-fault drivers, but this significantly reduces the sample size as citations are often not issued for crashes. Records containing home zip codes outside of Kentucky were not used for this analysis.

Several socioeconomic and demographic characteristics were identified as potential independent variables for this analysis based on the literature review. These include factors related to income, employment, and education for each zip code. In total, 16 variables that describe relevant socioeconomic and demographic factors were considered for the analysis and are summarized in Table 3.1.

Table 3.1 Socioeconomic and Demographic Variable Stats Summary

Socioeconomic/Demographic	Mean	Median	Stdev	Max	Min
Median Income ($)	41599.79	38990.50	17524.55	250000.00	10881.00
Mean Income ($)	52517.65	48984.00	25980.38	541084.00	10423.00
Unemployment Rate (%)	9.85	7.60	10.19	100.00	0.00
Percent Families Below Poverty Line (%)	23.49	21.10	15.18	100.00	0.00
Percent Below Poverty Line (≤18yrs) (%)	31.03	27.80	22.52	100.00	0.00
Percent High School Graduate (18-24yrs) (%)	81.90	85.75	19.44	100.00	0.00
Percent High School Graduate (≥25yrs) (%)	78.51	79.95	12.20	100.00	0.00
Percent with College Degree (≥25yrs) (%)	22.10	19.50	14.54	100.00	0.00
Percent Divorced or Separated (%)	15.72	14.90	10.78	100.00	0.00
Average Terrain Index	94.08	94.39	2.61	99.93	83.13
Average Terrain Difference	7.79	7.64	2.57	53.56	0.00
Percent Elderly (>65) (%)	17.04	15.60	11.86	100.00	0.00
Percent Young (15-24) (%)	12.56	11.90	8.42	99.10	0.00
Percent Rural (%)	79.47	100.00	34.70	100.00	0.00
Mean Travel Time to Work (minutes)	27.34	26.00	8.01	76.40	8.50
Average Vehicle Age (yrs)	10.09	10.15	1.02	15.33	0.00

Source: U.S. Census Bureau 2018, Kentucky State Police 2018

Of these variables, only Average Terrain Index, Average Vehicle Age, and Average Terrain Difference were obtained through sources other than the 2016 American Community Survey (U.S. Census Bureau 2018). Terrain Index is a proxy for roadway curviness based on the total length compared to the point-to-point length of all roadway segments in Kentucky (Staats et al. 2015). These values were averaged over zip code boundaries. The Average Terrain Difference can be calculated as the difference between the terrain index for the driver's residency (TI_{home}) and the crash location (TI_{crash}) multiplied by a factor of ten. This could provide an indication of a driver's lack of familiarity with the roadway environment based on their experience driving in their home environment. Average Vehicle Age was calculated from the crash records. In each crash, the model year of the vehicle is recorded for all units involved and the vehicle age was determined subtracting this from the crash year. The age of all vehicles was averaged across each zip code to allow for an estimated average vehicle age for the zip code.

Drug related data was also considered as a potential predicting variable for crash rates; however, drug data is difficult to obtain for such a small spatial unit. Drug overdose death rate was the primary data considered for this research, but due to privacy concerns, over 500 out of 960 zip codes were suppressed as the result of overdose related deaths being less than five for that zip code (i.e. privacy was not a concern when a zip code had more than five overdose deaths). Due to such a reduced sample size, drug related factors were not included in this phase of the research.

Driving exposure is an integral part of modeling the potential for vehicle crashes. Current methodology in the Highway Safety Manual recommends that crashes should be primarily modeled using AADT and analyzed at the segment level. A roadway segment is a continuous homogeneous portion of the roadway. Vehicle crashes tend to follow a negative binomial trend with AADT. In other words, more miles driven is positively correlated with the number of crashes that occur on a segment up to a point. However, to account for the socioeconomic and demographic variables, zip codes and not road segments, must be used as a spatial unit to connect crashes to those demographic predictors. Unfortunately, the use of zip codes as a spatial unit renders the use of AADT as a predictor nearly impossible. As alternative to consider exposure and thus normalize the number of crashes occurring in a zip code, the population above the age of 15-years-old was used as a near estimate of the driving population. This information is taken directly from the American Community Survey. Using this normalization, a crash rate of at-fault drivers can be developed for each of the zip codes in Kentucky. For each zip code (Z), the crash count is the number of crashes between the years 2012 and 2016 which involved an at-fault driver whose home location is zip code Z. This count is divided by the driving population in thousands of each zip code Z to produce a crash rate. For example, if the driving population of a zip code is 6,000, and there have been 900 at-fault crashes with drivers from that zip code in five years, the rate would be 150 crashes per 1,000 driving population per 5 year period.

Once the initial database was developed, it was examined for outliers. For example, several zip codes were reporting high crash rates, upwards of 2,000 crashes per 1,000 driving population. This would indicate that on average, every person in that zip code had two crashes in the five-year period considered. Most of these outliers were the result of a very low zip code population, so to account for this, zip codes with a total population of less than 20 people were excluded from the database. This eliminated 17 of the zip codes. It should be noted that such outliers may indicate specific problem driver groups and their repeated crashes could be associated to their socioeconomic characteristics. In order to examine this, a separate analysis was conducted that examined these initially excluded zip codes.

CHAPTER 4. METHODOLOGY

4.1 Multilinear Regression Model

To better understand how human factors can contribute to crash occurrence, this research utilized a multiple linear regression with socioeconomic and demographic characteristics as the explanatory variables. Crash counts were aggregated to every zip code in the state and normalized by driving population of that zip code. Additionally, socioeconomic and demographic variables were obtained from the US Census Bureau and connected to a singular database via the zip code. With the zip code as a consistent spatial unit of analysis, multiple linear regression was recognized to be a straight forward and useful approach for this analysis.

Multiple linear regression is a standard statistical technique that has been used by La Torre et al. (2007) and Ruiz et al. (2007) to relate socioeconomic and demographic characteristics to crash response. To determine if the socioeconomic and demographic variables were viable for linear regression, each potential variable was plotted with the response variable, i.e., crash rate. A regression analysis was conducted for each of the potential predictor variables alone to identify any trends. Linear and quadratic curves were fit to each of the regressions, and the p-value (significance) was compared for a linear versus a quadratic fit.

After fitting both a linear and a quadratic line to each of the scatter plots, the R-square values for both lines can be compared. The R-square indicates the proportion of the variance in the dependent variable that is predictable from the independent variable. For most of the independent variables, the R-square values were low (ranging from 0.000 to 0.118) for both the linear and quadratic curve fits. This means that on their own, none of the potential predictor variables (shown in Table 1) can significantly explain the change in crash rates among zip codes. Examining the plots, it is evident that each variable shows some degree of relation to crash rates using either a linear or a quadratic fit. In addition to a linear and a quadratic term, some variables were also considered for a logarithmic fit. Primarily this fit was considered in order to convert Median Income on a smaller, more workable scale; however, none of the variables showed any strong correlation with a logarithmic fit, so it was not used.

One of the highest R-square values (11.8 percent with p-value 0.000 for quadratic and 4.4 percent with p-value 0.000 for linear) shown among the predictor variables is for the “Percent Divorced and Separated.” With many variables following the trend of both a significant linear and a significant quadratic fit, each variable was transformed by squaring it, effectively allowing the use of a quadratic variable in a linear model. Figure 4.1 shows the regression fits for "Percent Divorced and Separated" as the independent variable.

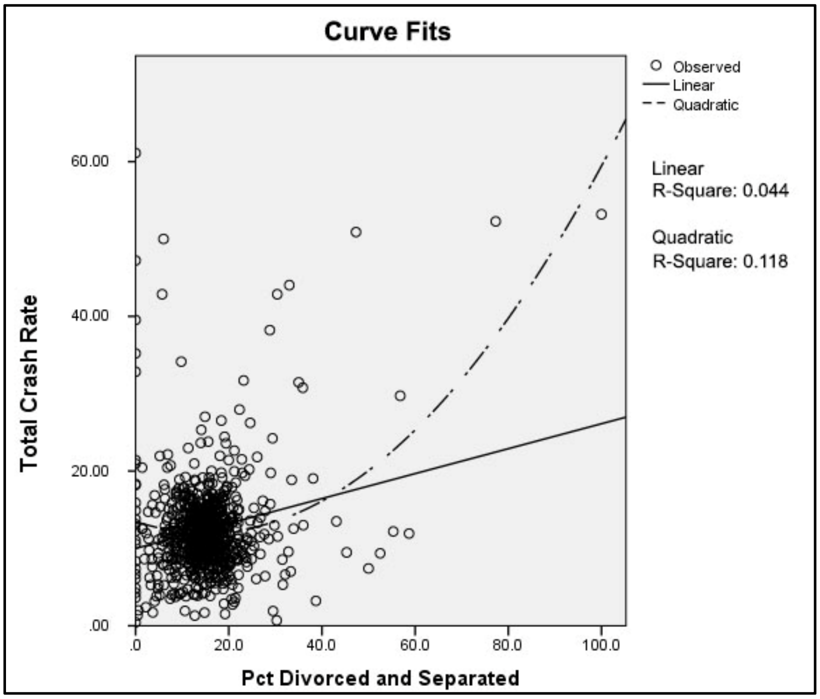

Figure 4.1 'Percent Divorced and Separated' Curve Fits

For each variable, the plots were examined for the goodness of curve fits and for the presence of outliers. There were outliers, but the data from the outlying zip codes still appears to be accurate. Notably, a zip code near Louisville had extremely high numbers for mean and median income (>$250,000 and $541,084 respectively) as well as high unemployment and high education levels. This zip code is however a very small wealthy neighborhood and the data seems appropriate.

The statistical package SPSS by IBM was used for the regression analysis in this study (IBM 2016). SPSS allows for the development of a multivariate linear model while introducing one independent variable at a time. A stepwise regression can be used to optimize many variables, but the stepwise will primarily consider the R-square value without considering the significance of the variables or their multicollinearity with each other. To build this regression model, each variable was manually considered for inclusion or exclusion based on its contribution to the R-square, its significance, and its multicollinearity. The first model was developed using the variable with the highest R-square value from the earlier scatter plots, Percent Divorced and Separated, as a starting point. Both the linear and quadratic terms were included as independent variables for the dependent variable, crash rate. As both the linear and the quadratic terms were significant, both were initially left in the model, though using both terms will not be allowed in the final model due to the presence of multicollinearity. After Percent Divorced and Separated, the potential predicting variables were included in descending order of R-square. Both the linear and quadratic terms were considered for inclusion. If the R-square value decreased or if one of the variables was notably not significant (having a t-statistic with a p-value of greater than .05), then it was not included. The R-square value quantifies the variance in the dependent variable that is predictable from the independent variable, and it was chosen as the indicator of model vitality for this research due to its widely understood meaning. R-square is simple and can be easily interpreted by most levels of familiarity with statistics. Other metrics such as AIC or BIC could also be used. Each of these simply accounts for a different penalty based on model parameters.

After all of the variables had been considered for inclusion, the collinearity of each was considered to determine if it warranted addressing. The Variance Inflation Factor (VIF) was used to assess multicollinearity of independent variables in the model. The VIF is the reciprocal of the tolerance, computed as:

$$VIF = \frac{1}{(1 - R^2)}$$

with R^2: the Coefficient of Determination.

The VIF is always greater than or equal to 1. There is no formal VIF value for determining presence of multicollinearity but often a VIF that exceeds 10 is regarded as indicating that multicollinearity should be addressed. A value of 10 would indicate that the standard errors are larger by a factor of 10 than they would normally be without inter-correlations between the predictor of interest and the remaining predictor variables in the multiple regression analysis. For this model, if a VIF exceeded 10 for an independent variable, then the variable was considered for removal. Sometimes removing only the linear or only the quadratic term for that variable addressed the problem. In some cases, there was multicollinearity between the variables which predict similar demographics such as Mean and Median Household Income. When Mean Income was considered for inclusion, it increased the R-square of the model, but the VIF for the Mean and the Median Income variables increased beyond 10, and as a result, the Median Income was left out of the initial model.

The Census data is not always complete for each zip code. Some records such as Income might be available for a particular zip code while other variables were simply not obtained during the census. Additionally, the default setting when running the regression is to exclude records that were missing one of the variable fields; a process called "listwise" in SPSS. In this case, the data is more complete, but the overall sample size may be reduced. Another option is to remove records based on the presence or absence of each particular variable. This approach is called in SPSS "pairwise." This allows for the use of records that contain missing data. Pairwise deletion does not include a particular variable when it has a missing value, but it can still use the record when analyzing other variables with non-missing values. This however can result in different statistics being based on different subsets of variables, which can be problematic. For this model, when the regression was run with pairwise deletion, the overall R-square value increased but the residual plots showed notable linear trends, and several of the otherwise significant variables became insignificant. After a final model was determined using listwise deletion, the pairwise deletion method was tested again, and the resulting R-square notably decreased.

CHAPTER 5. MODEL RESULTS

5.1 Primary Regression Model

After considering each variable and each term, linear and quadratic, for inclusion in the model, the first model was constructed optimizing R-square, VIF, and variable significance. The first model has an R-square value of 0.318 and a standard estimate error of 345.28. The variables included in the model are shown in Table 5.1, along with their respective coefficients, significance values, and VIF.

Table 5.1 Initial Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-2619.032	362.538		-7.224	.000		
Unemployment Rate	-7.021	2.223	-.112	-3.158	.002	.855	1.170
Percent High School Graduate (18-24)	2.157	.771	.095	2.799	.005	.926	1.080
Percent High School Graduate (≥25)	8.010	2.004	.191	3.996	.000	.472	2.117
Percent with College Degree (≥25)	-7.116	1.599	-.215	-4.452	.000	.462	2.163
Percent Young	-9.643	2.647	-.123	-3.643	.000	.947	1.056
Mean travel time to work (minutes)	5.327	1.949	.102	2.734	.006	.768	1.302
Mean Income Squared	1.389E-8	.000	.393	10.575	.000	.779	1.284
Average Terrain Index Squared	.323	.037	.351	8.783	.000	.674	1.484
Average Vehicle Age Squared	2.840	.937	.106	3.030	.003	.882	1.134
Average Terrain Difference Squared	-.337	.129	-.092	-2.625	.009	.876	1.142

Each variable in the model and the model as a whole is significant (all p-values are less than the 0.05 level of significance). Additionally, the VIF for all included variables fall below 3 indicating an acceptable level of variable multicollinearity for this model. Based on the Beta Coefficients, which have been standardized to a common scale among variables, the variable with the largest effect on crash rates is the squared term of "Mean Income" with a coefficient of 0.393, followed by the squared term for "Average Terrain

Index" with a coefficient of 0.351. This indicates that for every unit increase in each of these variables, crashes would increase.

To test the homoscedasticity of the model, a scatter plot was developed of the regression residuals as shown in Figure 5.1 below. The scatterplot allows for a visual examination of the assumption of homoscedasticity between the predicted values and the error for those values. If the variance of the standardized residuals is nearly the same for all predicted scores, then the assumption of homoscedasticity is met, and the chances for making Type I (false positive) and a Type II (false negative) errors are greatly reduced.

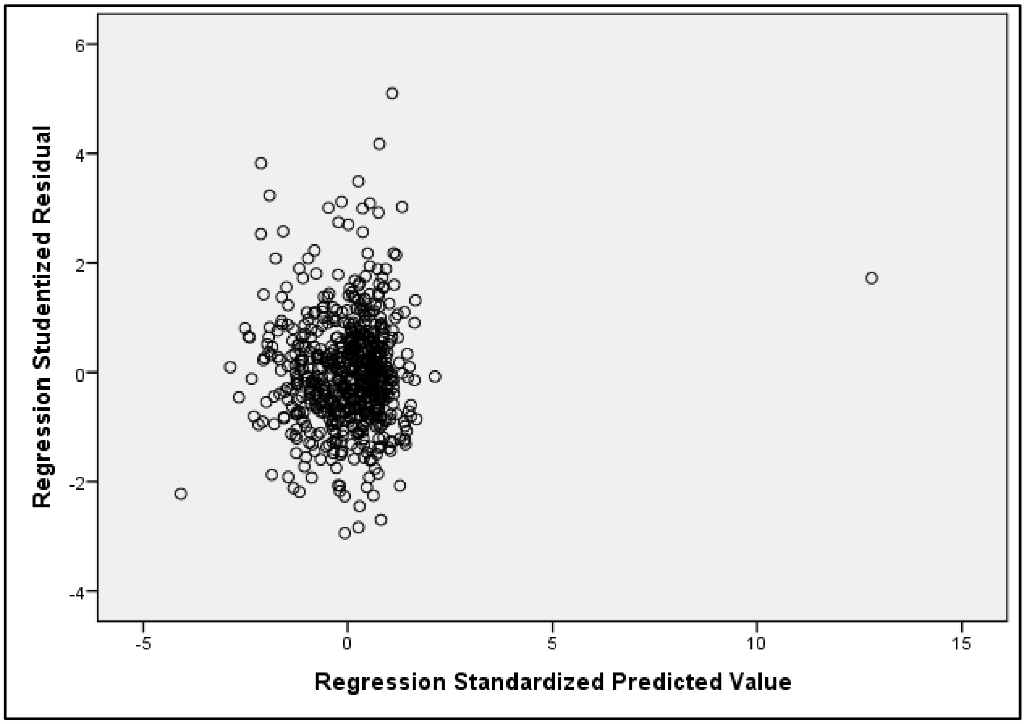

Figure 5.1 Initial Residual Plot

Figure 5.1 shows a generally randomly distributed residual plot, except for a large outlier. This outlier was identified to be the result of zip code 40025, a previously discussed zip code north of Louisville. This zip code not only has an exceptionally high median income, but a relatively small population (45 persons), high crash rate (428 crashes per 1,000 driving population), high education rate (97 percent college graduates), and high unemployment rate (17.6 percent). For these reasons, it was analyzed separately to determine whether it should be included in the data set. To do this, the human factor codes were examined for each crash that happened from 2012-2016. If a pattern became apparent (e.g. the majority of crashes attributed to drunk driving), then the zip code might merit inclusion in the model. Table 5.2 shows the primary human factor code listed for each crash in zip code 40025.

Table 5.2 Human Factors in Outlying Zip Code

Code	Human Factor	Number of Occurrences
4	Distraction	2
5	Drug Involvement	1
8	Failed to Yield Right of Way	2
11	Following Too Close	1
12	Improper Backing	1
14	Inattention	2
17	Misjudge Clearance	1
18	Not Under Proper Control	1
23	Turning Improperly	1
97	Other	2
99	None Detected	1

Table 5.2 shows no immediately discernable pattern which might hint to the human factor contribution in crash rates and therefore zip code 40025 will be excluded from the data set.

Running the model without this zip code, the R-square drops to 0.279, which is notably lower for the exclusion of a single data point. This means that zip code 40025 was having an undue effect on the fit of the model, likely due in part to its large income figures. In addition, with the exclusion of this point, the Mean Income squared term became insignificant in the model. Otherwise, the variable coefficient signs and magnitudes remained nearly the same. Shown in Table 5.3 is the new list of model variables, coefficients, p-values, and VIFs.

Table 5.3 Primary Model Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-2653.992	360.417		-7.364	.000		
Unemployment Rate	-7.827	2.207	-.129	-3.546	.000	.857	1.167
Percent High School Graduate (18-24)	2.109	.766	.095	2.753	.006	.955	1.047
Percent High School Graduate (≥25)	8.218	1.957	.203	4.199	.000	.485	2.060
Percent with College Degree (≥25)	-5.102	1.466	-.162	-3.480	.001	.523	1.911
Percent Young	-11.846	1.814	-.230	-6.531	.000	.916	1.091
Mean travel time to work (minutes)	5.967	1.937	.119	3.080	.002	.757	1.321
Average Terrain Index Squared	.327	.037	.367	8.958	.000	.676	1.478
Average Vehicle Age Squared	2.831	.917	.110	3.086	.002	.888	1.126
Average Terrain Difference Squared	-.323	.127	-.091	-2.549	.011	.896	1.116

As with the first model, the VIFs and significance values are all within acceptable levels. Figure 5.2 shows a new residual plot for the model, excluding the outlier.

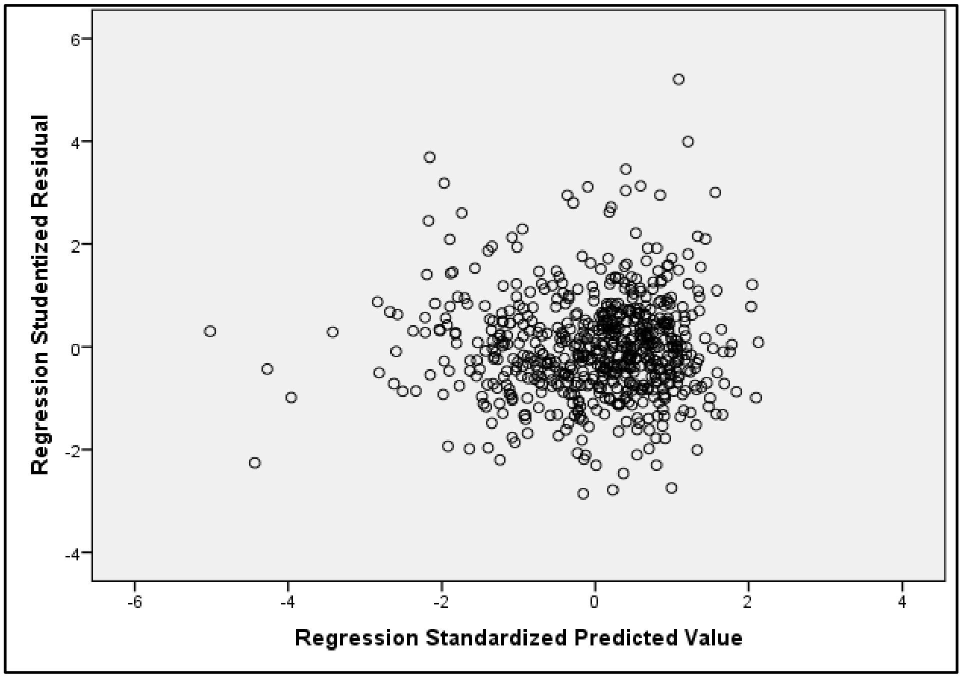

Figure 5.2 Final Residual Plot

The variance of the residuals remains relatively constant as the predicted values increase. For this reason, it is assumed that the model is homoscedastic, and the chances for making Type I (false positive) and a Type II (false negative) errors are greatly reduced. Though the R-square of the model is reduced with the exclusion of this outlier, the residual plot indicates that the overall worth of the model is improved.

In addition to homoscedasticity, the model must also be analyzed for the normality of its errors. A model with non-normal errors could result in incorrect inferences about variable relationships. A P-P plot shown as Figure 5.3 below was used to analyze the error normality of this model.

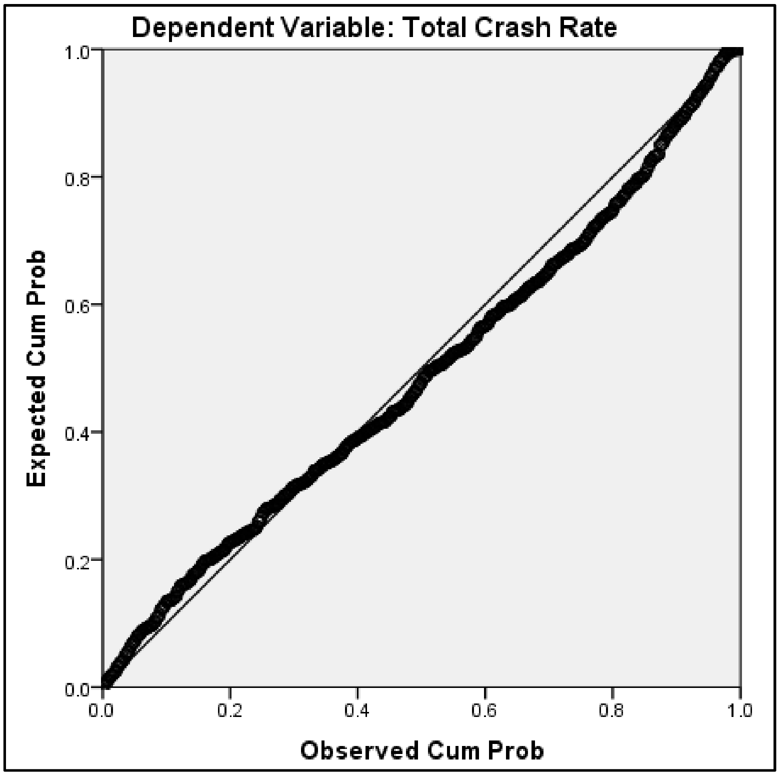

Figure 5.3 Error Normality Plot

The errors shown in the plot generally follow the line corresponding to a normal distribution. For this reason, it is assumed that the errors are normally distributed, and the risk of making invalid inferences is greatly reduced.

5.2 Interactions

An important aspect of the analysis is the examination of the interaction among the variables included in the model. A variable interaction is simply the product of two predictor variables. Certain variable interactions may cause a dramatic increase in the vitality of the model, in this case the R-square value. A program developed by the University of Kentucky was used to examine the impact of variable interactions (Lambert 2018). This program is designed to run a regression on the given variables and cycle through all of the potential interactions while optimizing a specified criterion such as R-square. The program is called "A Feasible Solution Algorithm (FSA) for Finding Interactions."

The program identified several variable interactions that when added to the current model increased the overall R-square. The interaction with the greatest effect on the R-square is a product of "Median Income" and "Average Terrain Index." These two terms and their product term must be added to the model when examining the impact of an interaction. With both terms and their interaction added to the model, the R-square value increased from .279 to .313; however, the VIF values for all three added terms and some terms already in the model became extremely high and the interaction inevitably was not included in the model.

Two additional interactions were also determined to increase the overall R-square of the model. These include the product of "Percent with a College Degree" with "Average Vehicle Age" and the product of "Percent Rural" with "Percent Divorced and Separated." Both interaction terms increased the R-square from 0.279 to 0.299 and 0.298 respectively, but both also had similar issues with VIF, indicating an unacceptable level of multicollinearity, and were not included in the final model. Though none of these terms were included in the final model, the increase in R-square from the interactions is still relevant.

5.3 Different Crash Rates

Though the socioeconomic and demographic characteristics associated with each crash are generally only available at the zip-code level, there are some variables which can be obtained at the driver level. For example, driver age can be included as a variable at the zip code level, indicating the percentage of certain age groups among the total population, but driver age is also a variable available for each driver record and thus it can be analyzed with less aggregation. One way of conducting this analysis is to average the age of each at-fault driver over the zip code. An alternative way is to group drivers into age groups and develop specific models for each subgroup.

The database was divided based on the driver's gender (male or female) and age (groups for 16-24, 25-44, 45-64, and over 65 years old). The regression was recalculated with each of these subgroups to determine if any of these variables examined could improve the predictive power of the model. Table 5.4 summarizes the R-square values for each of these models.

Table 5.4 R-square for different subgroups

Model	R-square
All Records	0.279
Male	0.285
Female	0.228
Age (16-24)	0.222
Age (25-44)	0.236
Age (45-64)	0.222
Age (65+)	0.196

From the initial R-square of 0.279 using all crash rates, the only model improvements came from male driver crash rates, with an R-square of 0.285. Using only male drivers was a very small improvement and it rendered the Average Terrain Difference insignificant. Without this variable, the model falls to an R-square of 0.282, meaning that these socioeconomic variables will predict male at-fault drivers slightly better than all drivers.

CHAPTER 6. DISCUSSION

The goal of this research was to examine the potential predictive ability of socioeconomic and demographic data for drivers on crash occurrence and determine whether any specific such variables could explain the differences in crash rates (as defined here) among zip codes. Some of the correlations are intuitive, while others may need more research to fully understand their ramifications.

The final model developed here has an R-square value of 0.279, which means that trying to predict crash rates with a model based solely on these socioeconomic and demographic variables is not fully practical at this stage. It should be noted, though, that each variable in the model is statistically significant, meaning that at a 95 percent confidence level, the impact of the variables is not random. While the model may not accurately predict crash rates at the zip code level, it still allows for meaningful interpretation of the relationship between the model variables and crashes.

For each variable in the model, the sign of its coefficient describes the relationship of the variable with crash rates. If the coefficient is positive, an increase in that variable will coincide with an increase in crash rates and if the coefficient is negative, the crash rates will decrease. For this model, each variable describes a different socioeconomic or demographic factor, so the sign of the variable must be interpreted accordingly. For example, employment can be reported as an employment or an unemployment rate, and the direction of the relationship with the dependent variable would be different for each. All the model variables and their respective coefficients are listed in Table 6.1.

Table 6.1 Model Variable Coefficients

Independent Variables	Unstandardized Coefficients
Average Terrain Index Squared (*ATI*)	0.327
Average Terrain Difference Squared (*ATD*)	-0.323
Percent Young (*PY*)	-11.846
Mean travel time to work (minutes) (*MTT*)	5.967
Unemployment Rate (*UR*)	-7.827
Percent with College Degree (≥25) (*CD25*)	-5.102
Percent High School Graduate (18-24) (*HS18*)	2.109
Percent High School Graduate (≥25) (*HS25*)	8.218
Average Vehicle Age Squared (*AVA*)	2.831

The coefficients shown here have both positive and negative signs indicating their relationships with the predicted variable, Crash Rate. Along with the intercept, the crash rate equation can be given as:

$$crash\ rates = -2653.992 + [ATI] * (0.327) + [ATD] * (-0.323) + [PY] * (-11.846) + [MTT] * (5.967) + [UR] * (-7.827) + [CD25] * (-5.102) + [HS25] * (8.218) + [HS18] * (2.109) + [AVA] * (2.831)$$

The coefficients reveal that Average Terrain Index Squared has a positive correlation with crash rates, meaning an increase in the Average Terrain Index will result in more crashes. The Average Terrain Index typically ranged from 75 to 100, with 100 being a zip code with entirely straight segments of road. Even though this correlation could be deemed counterintuitive, one needs to consider the location of such straighter roadway segments and their implications. For example, the straighter roads are found more often in urban settings and on larger arterial roads such as interstates. Both of these have a higher amount of travel than a rural curvy road and frequently drivers may drive at higher speeds and thus a positive correlation with terrain may simply point to other factors. The Average Terrain Index was also used to create a variable of the difference between a driver's home terrain and the terrain of the crash site. In this model, the Average Terrain Difference Squared term showed a negative association with crash rates in a zip code, indicating that on average, the crash rate increases when drivers are having crashes in similar terrain to that of their residence. However, this could be indicative of many other factors, such as drivers crashing at higher rates closer to home, or drivers becoming more attentive on unfamiliar roads. This variable may merit further investigation on a crash-by-crash basis.

Percent Young has a negative correlation with crash rates in this model. The Percent Young is the proportion of people ages 15-24 as compared to the total population. The negative coefficient sign indicates that a decrease in the percent of young drivers in a zip code would result in a higher crash rate for a zip code. This is somewhat unexpected, as a higher population of young drivers is typically associated with more crashes. In addition, past literature primarily indicates a positive association with young population and crashes or fatalities. Though most of the literature focuses on changes in fatality rates, Aguero-

Valverde et al. (2005) concluded that age groups below 25 and over 65 have a positive association with crash risk.

The Mean Travel Time to Work variable shows an expected relationship with crashes: the at-fault driver crash rates are higher in relation to a longer commute time to work. Both Adanu et al. (2017) and Lee et al. (2014) corroborate this result. Adanu et al. (2017) determined that an increase in travel time for a given postal code increased the probability of serious injury crashes, while Lee et al. (2017) showed that a shorter commute time corresponded to a smaller number of at-fault drivers. In both cases, it was concluded that commute time is a measure of exposure and that longer commute times will increase the number of crashes in an area.

Driving exposure may also explain the relationship of Unemployment Rate in a given zip code with crashes. This model shows a negative correlation with Unemployment Rate, indicating a higher rate of employment will correspond with a higher crash rate. Past research pertaining to this socioeconomic factor has shown inconclusive findings; however, like Mean Travel Time to Work, Unemployment Rate may also be predicting driving exposure. A high Unemployment Rate could indicate less day-to-day driving activity within a zip code, which would result in a lower crash rate.

Similar to findings from Hasselberg et al. (2005) and Zephaniah et al. (2008), this model shows a negative relationship between education level and crash rates. As the Percent with a College Degree increases for a zip code, crash rates are decreased, which points to a positive impact of higher education on driving ability. Two additional education terms in the model however show a different relationship. The Percent of High School Graduates in a zip code has a positive association with crash rate, which indicates that an increase in the number of high school graduates would actually result in a higher crash rate for a zip code. This relationship holds for drivers in the 18-24 age group and those over 25 years old. Despite the divide between high school and college attainment, there are some reasons why this relationship might happen. For example, a higher level of high school graduates in a zip code may indicate a higher level of employment, which has also been shown to increase crash rates.

The final variable in the model, Average Age of Vehicle Squared, has a positive association with crash rates, indicating that drivers of older vehicles are more frequently involved in crashes. As a vehicle ages, certain functional systems such as brakes, tires, and wipers become more likely to underperform or fail, and in many cases, this could be a contributing factor in crashes. The age of vehicles, and consequently the safety of drivers in an area might also be related to the average income in an area, as lower income drivers are potentially less equipped to afford vehicle maintenance and repairs.

Several of the variable relationships in this model contradict a priori expectations from previous literature. For example, several articles noted that the proportion of young people had a positive correlation with crashes and crash severity, but this model indicates otherwise. In addition, past research indicated a negative association between education and crashes, but the results of this work are mixed, pointing to a negative relationship with college attainment but a positive relationship with high school attainment.

Based on prior research and the results of this regression, the Mean Travel to Work seems to be a proxy for driving exposure and it is theorized that Unemployment Rate may be pointing to the same concept. In addition, the Average Terrain Index variable shows higher crash rates on straighter roads indicating it may also have some interaction with driving exposure or even driving speed.

CHAPTER 7. CONCLUSIONS AND RECOMMENDATIONS

The analysis undertaken in this report shows both positive and negative correlations between socioeconomic variables and crash rates, some of which met expectations while others contradicted the findings of prior research as discussed previously. Further research seems to be warranted to fully understand the role that socioeconomic and demographic characteristics play in driving behavior and crash risk.

One way to further this research is to address driving exposure. This can be done by normalizing the crash counts by VMT instead of population. The current Highway Safety Manual methodology develops SPFs which analyze crashes on a roadway segment level and take AADT into account for that segment. Though the crashes have been normalized by population of the zip code, the total miles traveled in that zip code may play a large role in crash estimation. It should be noted though that such estimates are impossible to obtain at the level required for such an analysis. One way to include VMT in conjunction with socioeconomic data is to geocode at-fault driver crashes to the road segment level and analyze the average socioeconomic and demographic data on that segment. Using residence zip codes of the at-fault drivers crashing on each segment, an average socioeconomic variable can be calculated for that segment. An SPF can then be based on the AADT of the roadway segment in conjunction with socioeconomic and demographic data. However, such aggregation may mask the detailed information required for estimating the effect of socioeconomic factors on crash occurrence.

Moreover, while this report studied the number of at-fault driver crashes, the fault of the driver could be further analyzed with other techniques such as a log-linear model (Kim et. al 1998) or a logistic regression. The logistic regression approach would involve building a model to predict the fault of a driver based on socioeconomics and would include the entire database of crashes (both at-fault and not at-fault), effectively doubling the sample size. In addition, it would allow for the inclusion of certain variables which can be obtained from the crash record (e.g. crash type, roadway conditions, gender, and age) without the need for aggregation, as the driver would be the analysis unit.

With more research, it is possible to further identify unique background characteristics of at-fault drivers that contribute to crash rates and crash severity. A more in depth understanding of these relationships will lead to improved and targeted intervention measures to reduce the negative impacts of motor vehicle crashes.

REFERENCES

Abdalla, I. M., Raeside, R., Barker, D., & McGuigan, D. R. (1997). An investigation into the relationships between area social characteristics and road accident casualties. *Accident Analysis & Prevention*, *29*(5), 583-593.

Adanu, E., Penmetsa, P., Jones, S., & Smith, R. (2018). Gendered Analysis of Fatal Crashes among Young Drivers in Alabama, USA. *Safety*, *4*(3), 29.

Adanu, E. K., Smith, R., Powell, L., & Jones, S. (2017). Multilevel analysis of the role of human factors in regional disparities in crash outcomes. *Accident Analysis & Prevention*, *109*, 10-17.

Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention*, *38*(3), 618-625.

American Association of State Highway and Transportation Officials. Transportation Research Board. Task Force on Development of the Highway Safety Manual, & Transportation Officials. Joint Task Force on the Highway Safety Manual. (2010). *Highway safety manual* (Vol. 1). AASHTO.

Blatt, J., & Furman, S. M. (1998). Residence location of drivers involved in fatal crashes. *Accident Analysis & Prevention*, *30*(6), 705-711.

Brown, K., Sarasua, W. A., & Ogle, J. H. (2016). *Safety Planning: Analysis of the Socio-Economic and Demographic Characteristics of At-Risk Driver Residential Areas in South Carolina* (No. 16-5034).

Chandraratna, S., & Stamatiadis, N. (2009). Quasi-induced exposure method: evaluation of not-at-fault assumption. *Accident Analysis & Prevention*, *41*(2), 308-313.

Chen, H. Y., Ivers, R. Q., Martiniuk, A. L. C., Boufous, S., Senserrick, T., Woodward, M., Stevenson, M., & Norton, R. (2010). Socioeconomic status and risk of car crash injury, independent of place of residence and driving exposure: results from the DRIVE Study. *Journal of Epidemiology & Community Health*, *64*(11), 998-1003.

Factor, R., Mahalel, D., & Yair, G. (2008). Inter-group differences in road-traffic crash involvement. *Accident Analysis & Prevention*, *40*(6), 2000-2007.

Hanna, C. L., Laflamme, L., & Bingham, C. R. (2012). Fatal crash involvement of unlicensed young drivers: county level differences according to material deprivation and urbanicity in the United States. *Accident Analysis & Prevention*, *45*, 291-295.

Hasselberg, M., Vaez, M., & Laflamme, L. (2005). Socioeconomic aspects of the circumstances and consequences of car crashes among young adults. *Social science & medicine*, *60*(2), 287-295.

IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

Ivan, J., Burnicki, A., Wang, K., & Mamun, S. (2016). *Improvements to road safety improvement selection procedures for Connecticut* (No. JHR 16-328).

Kentucky State Police. (2018). KY Collision Analysis. Retrieved January, 2018, from crashinformationky.org/.

Kim, K., Li, L., Richardson, J., & Nitz, L. (1998). Drivers at fault: influences of age, sex, and vehicle type. *Journal of Safety Research*, *29*(3), 171-179.

Kocatepe, A., Ulak, M. B., Ozguven, E. E., Horner, M. W., & Arghandeh, R. (2017). Socioeconomic characteristics and crash injury exposure: A case study in Florida using two-step floating catchment area method. *Applied geography*, *87*, 207-221.

Lambert, J. (2018). A Feasible Solution Algorithm (FSA) for Finding Interactions. Retrieved June, 2018, from shiny.as.uky.edu/mcfsa/.

La Torre, G., Van Beeck, E., Quaranta, G., Mannocci, A., & Ricciardi, W. (2007). Determinants of within-country variation in traffic accident mortality in Italy: a geographical analysis. *International journal of health geographics*, *6*(1), 49.

Lee, J., Abdel-Aty, M., & Choi, K. (2014). Analysis of residence characteristics of at-fault drivers in traffic crashes. *Safety science*, *68*, 6-13.

Lourens, P. F., Vissers, J. A., & Jessurun, M. (1999). Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. *Accident Analysis & Prevention*, *31*(5), 593-597.

Maciag, M. (2014). Pedestrian Deaths in Poorer Neighborhoods Report. *Governing Magazine: State and Local Government News for America's Leaders*. Retrieved from www.governing.com/gov-data/pedestrian-deaths-poor-neighborhoods-report.html.

Muelleman, R. L., & Mueller, K. (1996). Fatal motor vehicle crashes: variations of crash characteristics within rural regions of different population densities. *Journal of Trauma and Acute Care Surgery*, *41*(2), 315-320.

National Highway Traffic Safety Administration. (July 2, 2018). Fatality Analysis Reporting System (FARS). Retrieved August 25, 2018, from www.nhtsa.gov/FARS.

Noland, R. B., & Laham, M. L. (2018). Are low income and minority households more likely to die from traffic-related crashes?. *Accident Analysis & Prevention*, *120*, 233-238.

Noland, R. B., & Quddus, M. A. (2004). A spatially disaggregate analysis of road casualties in England. *Accident Analysis & Prevention*, *36*(6), 973-984.

Rivas-Ruiz, F., Perea-Milla, E., & Jimenez-Puente, A. (2007). Geographic variability of fatal road traffic injuries in Spain during the period 2002–2004: an ecological study. *BMC Public Health*, *7*(1), 266.

Staats, W. N. (2016). Estimation of Annual Average Daily Traffic on Local Roads in Kentucky.

Stamatiadis, N., & Deacon, J. A. (1997). Quasi-induced exposure: methodology and insight. *Accident Analysis & Prevention*, *29*(1), 37-52.

Stamatiadis, N., & Puccini, G. (1999). Fatal crash rates in the Southeastern United States: why are they higher?. *Transportation Research Record: Journal of the Transportation Research Board*, (1665), 118-124.

University of Kentucky and Kentucky State Police. (2017). Kentucky Traffic Collision Facts 2016. Retrieved September 2, 2018, from http://ksponline.org/pdf/KY_Traffic_Collision_Facts_2016.pdf

United States Census Bureau. (2016). 2012-2016 American Community Survey 5-year estimates. Retrieved December, 2017, from http://factfinder2.census.gov.

World Health Organization. (19 Feb. 2018). Road Traffic Injuries. Retrieved September, 2018, from www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Zephaniah, S., Jones, S., Smith, R., & Weber, J. (2018). Spatial Dependence among Socioeconomic Attributes in the Analysis of Crashes Attributable to Human Factors.

Zwerling, C., Peek-Asa, C., Whitten, P. S., Choi, S. W., Sprince, N. L., & Jones, M. P. (2005). Fatal motor vehicle crashes in rural and urban areas: decomposing rates into contributing factors. *Injury Prevention*, *11*(1), 24-28.

VITA

AARON BERRY CAMBRON

EDUCATION

BSCE	University of Kentucky; Lexington, Kentucky. May 2017
2013 -2017	Bachelor of Science in Civil Engineering. GPA: 3.53

HONORS AND AWARDS

2018 Distinguished Service Award, Engineers Without Borders - University of Kentucky

2016 William N. and Ocie M. Downey Scholarship ($1000)

2015 William Harp Memorial Scholarship ($500)

2013-2017 UK Provost Scholarship ($1500/year)

2013-2017 Kentucky Educational Excellence Scholarship ($2500/year)