Modeling Socio-Economic Determinants of Traffic Fatalities

Amin Azimian  
University of Dayton

Deogratias Eustace  
University of Dayton, deustace1@udayton.edu

Follow this and additional works at: https://ecommons.udayton.edu/cee_fac_pub

Part of the Civil Engineering Commons, and the Transportation Engineering Commons

eCommons Citation  
https://ecommons.udayton.edu/cee_fac_pub/68

This Article is brought to you for free and open access by the Department of Civil and Environmental Engineering and Engineering Mechanics at eCommons. It has been accepted for inclusion in Civil and Environmental Engineering and Engineering Mechanics Faculty Publications by an authorized administrator of eCommons. For more information, please contact frice1@udayton.edu, mschlangen1@udayton.edu.
Abstract— the objective of this paper was to model socio-economic determinants of traffic fatalities across all U.S States. This goal was accomplished by employing the Geographically Weighted Regression (GWR) and global ordinary least squares model (OLS). The results demonstrated that the GWR model outperformed OLS model in terms of accuracy. Furthermore, it was found that population with travel time to work less than 20 minutes, population with no high school diploma, median income, population with age over 65 in labor force and high school graduates between 18-24 significantly contributed to traffic fatality rate

Index Terms— Traffic, Safety, Fatality Rate, Regression, Socio-economics.

I. INTRODUCTION

A. General

Over several decades, traffic growth has caused an increased number of traffic crashes, which are associated with economic losses and human sufferings. According to the National Highway Traffic Safety Administration [1] in 2016 there were a total of 34,247 fatal traffic crashes in the United States that resulted into 37,133 fatalities. Risk factors relating to the occurrence of fatal and injury severity of motor vehicle crashes have been extensively studied. Most studies [2-7] that have attempted to model the occurrences of traffic crashes and fatalities have been mainly confined to factors related to driver characteristics, roadway geometry characteristics, traffic characteristics, crash characteristics, and environmental characteristics. The driver characteristics usually modeled include driver age, gender, alcohol use, and drug impairment involvement. Roadway geometry factors mostly include horizontal and vertical alignments, roadway and shoulder widths, presence of work zone construction, and number of lanes. Traffic characteristics mainly include average daily traffic volume (ADT) and percent of trucks. For crash characteristics, factors usually considered include type of crash, manner of collision, and location where the crash occurred. The environmental characteristics include light condition, weather condition, day of the week and time when the crash occurred. However, some few studies [8-10] have attempted to model other factors such as socioeconomic factors that may play role in occurrences of traffic crashes and fatalities. Kirk et al. [8] explored the impacts of socio-economic factors and safety regulations have on statewide traffic crash rates in the state of Kentucky. Their study indicates that at the national level, socioeconomic factors such as poverty, income and education have a significant impact on traffic crash rates but when analyzed at the state level, they found that high school education attainment was the most significant indicator for elevated crash crashes.

Recently, many authors [11-13] proposed full Bayes (FB) hierarchical model to study traffic crashes over space and time. Although FB approaches accounts for the sources of uncertainties, but in some cases, the variables may not be converged after many iterations. In contrast, linear regression models have much lower running time and less space complexity.

B. Research Objectives

The objectives of this paper are two-fold: (i) Identification socioeconomic factors contributing to traffic fatality rates using both ordinary least squares linear regression model (OLS) and geographically weighted regression model (GWR), and (ii) consequently comparison of the results provided by the two models.

II. METHODOLOGY

A. Model Specifications

Regression analysis is a statistical process that figures out the relationship between a dependent variable (Y) and a set of one or more independent variables (X). The prediction of the dependent variable in a OLS assumes that the estimates apply universally disregarding the possibilities of the influence of some of the independent variables varying spatially. The OLS model can be represented as shown in Equation 1. For further study please

\[ y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i \]  

Where:

- \( y_i \) = dependent variable at location \( i \)
- \( x_{ij} \) = independent variables (j = 1, 2, …, m)
- \( \beta_i \) = model estimated coefficients
- \( \epsilon \) = error term

Therefore, the parameters for a linear regression model can be obtained by Equation 2:
Modeling Socio-economic Determinants of Traffic Fatalities

\[ \hat{B} = \left( X^T X \right)^{-1} X^T Y \]  
\[ (2) \]

Where:
- \( \hat{B} \) = vector of the parameter estimates
- \( X \) = matrix of independent variables with the values of 1 in the first column (corresponding to the intercept)
- \( Y \) = a column vector with the values of dependent variable
- \( X^TX \) = the variance-covariance matrix
- \( m \) = number of parameters in the model

\[ \hat{B}_i = \left( X^T W_i X \right)^{-1} X^T W_i Y \]  
\[ (4) \]

Where:
- \( \hat{B}_i \) = vector of the parameter estimates that describes a relationship in location \( i \) and is specific to that location
- \( W_i \) = a square matrix of weights relative to the position of location \( i \) in the study area
- \( X^T W_i X \) = a geographically weighted variance-covariance matrix

The square matrix, \( W_i \), is a matrix in which the diagonal entries are geographical weights and the off-diagonal entries are all zero.

\[ W_i = \begin{bmatrix} w_{11} & 0 & \ldots & 0 \\ 0 & w_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & w_{nn} \end{bmatrix} \]

The elements (weights) themselves are computed from a weighting scheme, which is also known as a kernel. A number of kernels are possible and one of the most typical ones has a Gaussian shape and is computed as shown in Equation 5:

\[ w_y = \exp \left( -\frac{d_y^2}{h^2} \right) \]  
\[ (5) \]

Where:
- \( d_y \) = geographical weight of the observation at location \( i \) in the dataset relative to the observation at location \( j \)
- \( d_y \) = distance between mean centers of locations \( i \) and \( j \)
- \( h \) = a quantity known as the bandwidth

In cases where the bandwidth is unknown or there is no prior justification for providing a particular bandwidth, Fotheringham et al. [14] recommends for the analyst to let the software choose an appropriate bandwidth. In this paper the type of kernel used to provide spatial weighting is a fixed kernel since the observations are randomly distributed in the study area and the bandwidth parameter was found by using cross-validation (CV) method that computes the bandwidth which minimizes a cross-validation score. This method automatically finds the bandwidth which gives the best prediction. According to Fotheringham et al. [14], a cross-validation score is essentially the sum of estimated predicted squared errors determined as shown in Equation 6. For a complete discussion of different types of kernels and cross validation methods, please refer to Fotheringham et al. (11).

\[ CV = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  
\[ (6) \]

Where:
- \( n \) = number of data points
- \( \hat{y}_i \) = prediction for the \( i \)th data point.

B. Data

In order to perform a regression analysis, traffic fatalities and socio-economic (population estimates) data for the year 2017 were obtained from the National Highway Traffic Safety Administration (NHTSA) and the U.S. Census Bureau websites, respectively. The dependent variable is “fatalities per 100 million vehicle miles of travel” in each State. Among many potential independent variables found in US Census Bureau, the following independent variables were selected using stepwise regression method (\( \alpha=0.05 \)). Each variable is grouped into smaller categories based on their frequency distribution. Variable statistics are shown in Table 1.

- Work: percent of population that arrives at work within 0-20 minutes.
- Nodiploma: percentage of population who do not hold high school diploma
- Income: median annual household income
- Labor: percent of population with age greater than 65 who are in labor force
- Highschl: percentage of high school graduates with age between 18 and 24 years old

Table 1: Descriptive statistics for the variables
III. RESULTS AND DISCUSSIONS

A. Analysis of The Assumptions

Normality Assumption: According to Bowerman and O’Connel [15], normality assumption holds if:
P(-1 ≤ ε_i ≤ 1) = 0.68 and P(-2 ≤ ε_i ≤ 2) = 0.95 were ε_i is a point estimate of the standardized residual. In this study about 70 percent of the standardized residuals are between -1 and 1, and about 94 percent of the standardized residuals are between -2 and 2. Therefore, normality assumption approximately holds.

Independence Assumption: Using Moran’s I function shown in Equation 7 it is possible to determine if any value of the dependent variable, fatality, is statistically independent from any other value of fatality. In general, a Moran's Index value near +1.0 indicates clustering while an index value near -1.0 indicates dispersion and a zero value indicates a random spatial pattern.

\[
I = \frac{\sum_{i} \sum_{j} w_{ij} (r_i - \bar{r})(r_j - \bar{r})}{\sum_{i} (r_i - \bar{r})^2}
\]

(7)

Where:
- \( I \) = Moran’s index value
- \( N \) = number of features (in this case 51)
- \( r_i \) and \( r_j \) = residuals related to features i and j
- \( \bar{r} \) = mean of residuals, i.e., 0.001
- \( w_{ij} \) = an element of a matrix of spatial weights

In this study Moran’s index value of \( I = 0.03 \) was obtained for the OLS model, which indicates a random spatial pattern. Therefore, this assumption holds. Likewise, for the GWR model, the Moran’s I index for the residuals was 0.04, which demonstrates that there is little evidence of any autocorrelation between each other.

Constant Variance Assumption: Constant variance means that for any value of the independent variable \( X_i \), the corresponding population of potential values of dependent variable has a variance that does not depend on the values of \( X_i \).

The constant variance assumption holds if the residual plots indicate the horizontal band appearance [15]. Figure 1 represents the residual plots with a horizontal band appearance, which demonstrates that constant variance assumption holds.

B. Model Parameter Estimates

Based on the correlation results shown in Table 2, none of the independent variables are highly correlated to each other, which is a desirable feature. Therefore, in order to account for the variation in traffic fatalities all independent variables were included in global linear regression model represented by Equation 1 was considered. The model in Equation 1 was calibrated using the OLS to produce the parameter estimates.

The result shows that all variables are significant at 0.05 level of significance. "Work" and “Income” are negatively associated with the fatality rate. This implies that increases in a U.S. State population whose travel time to work is less than 20 minutes will likely decrease the fatality rate in that U.S. State. Additionally, U.S. States with higher median income tend to have smaller crash rate. On the other hand, “Nodiploma”, “Highschl” and “Labor” are positively associated with the fatality rate. That is, increases in uneducated and young population will likely increase the fatality rate which could be due to lack of experience in young drivers. Furthermore, increases in labor forces who are 65 or older can increase the fatality rate. This might be because of poor reaction time in elderly drivers.

The following table shows the coefficient estimates of the regression model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Work</th>
<th>Income</th>
<th>Nodiploma</th>
<th>Labor</th>
<th>Highschl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>1.00</td>
<td>0.40</td>
<td>0.20</td>
<td>-0.19</td>
<td>-0.39</td>
</tr>
<tr>
<td>Income</td>
<td>1.00</td>
<td>-0.375</td>
<td>0.096</td>
<td>-0.112</td>
<td></td>
</tr>
<tr>
<td>Nodiploma</td>
<td>1.00</td>
<td>-0.354</td>
<td>-0.114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labor</td>
<td>1.00</td>
<td>-0.103</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highschl</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Plots of residual values versus independent variables and predicted values

Table 2: Variable Correlation Results
Table 3: Parameter Estimates for the OLS Model

| Variable | Estimate | Std. Error | t-value | Pr>|t| |
|----------|----------|------------|---------|-------|
| Fatality | 1.08     | 0.47       | 2.29    | 0.025 |
| Work     | -4.34    | 1.39       | -3.12   | 0.003 |
| income   | -0.18    | 0.038      | -4.69   | <0.0001 |
| Nodiploma| 13.05    | 3.65       | 3.58    | 0.0008 |
| Labor    | 68.31    | 31.41      | 2.17    | 0.035 |
| Highschl | 34.53    | 9.00       | 3.83    | 0.0004 |

Table 4: Parameter Estimates for the GWR Model

<table>
<thead>
<tr>
<th>State</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>$\hat{\beta}_3$</th>
<th>$\hat{\beta}_4$</th>
<th>$\hat{\beta}_5$</th>
<th>$\hat{\beta}_6$</th>
<th>$\hat{\beta}_7$</th>
<th>$\hat{\beta}_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>1.08</td>
<td>0.47</td>
<td>2.29</td>
<td>0.025</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK</td>
<td>-4.34</td>
<td>1.39</td>
<td>-3.12</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>13.05</td>
<td>3.65</td>
<td>3.58</td>
<td>0.0008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>68.31</td>
<td>31.41</td>
<td>2.17</td>
<td>0.035</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HI</td>
<td>34.53</td>
<td>9.00</td>
<td>3.83</td>
<td>0.0004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The dependent variable and the explanatory variables used in the GWR model are the same as those specified for the OLS model. Table 4 shows the parameter estimates for the GWR model. It can be seen that the all local $R^2$ are slightly higher than the $R^2$ in OLS model. This means that the GWR model fits data better than OLS model. Figure 2 shows the variation in the parameter estimates for each independent variable across the states. By examining Figure 2 we can see that local coefficients of explanatory variables reveal the influence of these variables in the GWR model, which varies over the United States with a strong west-east or east-west direction. In other word, the effects of “Nodiploma”, “Highschl” and “Income” on fatality rate in eastern States is higher than that in central and western States. In contrast, “Work” and “Labor” tend to significantly influence the fatality rate in western States compared to central and eastern States.

IV. CONCLUSION

In this paper the OLS and GWR models were employed to investigate the relationship between traffic fatality rates and some selected socio-economic factors across all U.S. states. The results indicate that global coefficient and local coefficients for each variable agree in terms of directionality, i.e., they are both either negative or positive for the same parameter estimated. However, the estimated $R^2$ in GWR is slightly higher than that in OLS model. Also, all independent variables; “Work”, “Nodiploma”, “Highschl”, “Labor” and “Income” are significant at $\alpha=0.05$. The effect of these variables on each state can be evaluated by the decision makers to determine whether any corrective actions are needed.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES


