11-2018

Modeling Socio-Economic Determinants of Traffic Fatalities

Amin Azimian
University of Dayton

Deogratias Eustace
University of Dayton, deustace1@udayton.edu

Modeling Socio-economic Determinants of Traffic Fatalities

Amin Azimian, Deogratias Eustace

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 wok 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_i). 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

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ji} + \varepsilon_i \qquad (1)$$

Where:

y_i = dependent variable at location i
x_{ji} = independent variables (j = 1, 2, …, m)
β_i = model estimated coefficients
ε = error term

Therefore, the parameters for a linear regression model can be obtained by Equation 2:

Manuscript received November 03, 2018

Amin Azimian, Department of Civil and Environmental Engineering & Engineering Mechanics, University of Dayton, Dayton, OH, USA,
Deogratias Eustace, Department of Civil and Environmental Engineering & Engineering Mechanics, University of Dayton, Dayton, OH, USA

$$\bar{B} = \left(X^T X\right)^{-1} X^T Y \quad (2)$$

Where:

$\bar{B}$ = 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
$X^T X$= the variance-covariance matrix
m = number of parameters in the model

GWR is used to calibrate multiple regression models that allow different relationships that exist at different locations. The underlying concept of GWR is that observations which are nearer to a particular location should have a greater weight in the estimation than observations that are further away. The GWR model shown in Equation 3 is used to estimate parameters in the study area that relate the dependent variable with a set of independent variables, which have been measured locally, e.g., for each state in the United States.

$$y_i = \beta_{0i} + \sum_{j=1}^{m} \beta_{ji} x_{ji} \quad (3)$$

The parameter estimates for GWR are solved using a weighting system as shown in Equation 4, where the weights are inured on the location i.

$$\bar{B}_i = \left(X^T W_i X\right)^{-1} X^T W_i Y \quad (4)$$

Where:

$\bar{B}_i$ = 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.

$$W_i = \begin{bmatrix} w_{i1} & 0 & 0 & \ldots & 0 \\ 0 & w_{i2} & 0 & \ldots & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & \ldots & w_{in} \end{bmatrix}$$

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:

$$w_{ij} = Exp\left(-\frac{d_{ij}}{h}\right)^2 \quad (5)$$

Where;

W_{ij} = geographical weight of the observation at location i in the dataset relative to the observation at location j
d_{ij} = 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*).

$$CV = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \quad (6)$$

Where:

n = number of data points
$\hat{y}_i$ = prediction for the ith 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 “fatality *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 (α=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

Variable	*Mean*	*Std dev.*	*Max*	*Min*
Fatality	1.20	0.33	2.01	0.61
work	15.44	2.37	21.7	8.9
income	5.03	0.85	6.93	3.66
nodiploma	2.78	0.85	4.4	1.5
labor	1.74	0.09	0.6	0.2
highschl	3.10	0.30	4.3	2.6

No. Observations: 51

III. Results and Discussions

A. Analysis of The Assumptions

Normality Assumption: According to Bowerman and O'Connel [15], normality assumption holds if:
$P(-1\leq \varepsilon_i \leq 1) = 0.68$ and $P(-2 \leq \varepsilon_i \leq 2) = 0.95$
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.

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(r_i - \bar{r})(r_j - \bar{r})}{\sum_i (r_i - \bar{r})^2} \qquad (7)$$

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.

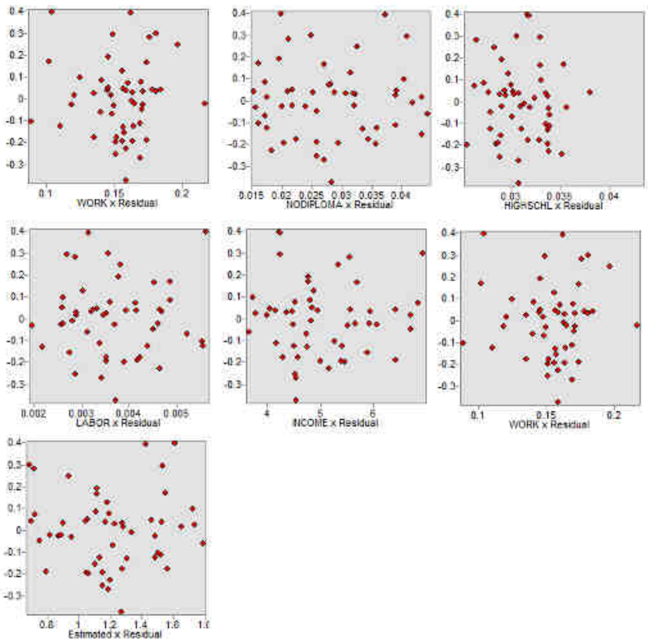

Figure 1: Plots of residual values versus independent variables and predicted values

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.

Table 2: Variable Correlation Results

Variable	*Work*	*Income*	*Nodiploma*	*Labor*	*Highschl*
Work	1.00	0.40	0.20	-0.19	-0.39
income		1.00	-0.375	0.096	-0.112
Nodiploma			1.00	-0.354	-0.114
Labor				1.00	-0.103
Highschl					1.00

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

State	β_0	β_1	β_2	β_3	β_4	β_5	Y_1	R^2
AL	2.171	-4.583	13.768	-0.173	62.866	33.293	1.467	0.751
AK	2.216	-4.169	10.983	-0.190	80.803	33.540	1.268	0.750
AZ	2.241	-4.143	11.880	-0.192	67.907	30.732	1.191	0.743
AR	2.194	-4.524	13.325	-0.178	63.939	32.411	1.736	0.749
CA	2.241	-4.002	11.402	-0.195	70.993	31.003	1.080	0.742
CO	2.229	-4.346	12.387	-0.186	67.126	31.247	0.737	0.745
CT	2.118	-4.797	14.618	-0.162	63.719	35.715	0.758	0.755
DE	2.126	-4.753	14.496	-0.164	63.349	35.317	1.105	0.754
DC	2.133	-4.739	14.393	-0.165	63.426	35.040	0.803	0.754
FL	2.149	-4.603	14.165	-0.169	61.654	34.132	1.265	0.753
GA	2.156	-4.623	14.032	-0.170	62.558	33.925	1.316	0.753
HI	2.230	-2.843	8.705	-0.219	84.121	32.211	1.095	0.757
ID	2.235	-4.249	11.880	-0.190	70.348	31.348	1.484	0.744
IL	2.182	-4.615	13.531	-0.174	64.685	33.001	1.058	0.750
IN	2.171	-4.649	13.735	-0.172	64.357	33.439	1.266	0.751
IA	2.197	-4.577	13.226	-0.177	65.642	32.478	1.118	0.749
KS	2.213	-4.467	12.889	-0.181	65.603	31.797	1.127	0.747
KY	2.167	-4.641	13.831	-0.172	63.758	33.590	1.657	0.751
LA	2.193	-4.495	13.371	-0.178	63.040	32.364	1.540	0.749
ME	2.110	-4.851	14.733	-0.161	64.292	36.108	1.255	0.755
MD	2.132	-4.742	14.406	-0.165	63.440	35.079	0.694	0.754
MA	2.115	-4.809	14.661	-0.162	63.795	35.847	0.796	0.755
MI	2.170	-4.697	13.741	-0.172	65.239	33.579	1.148	0.751
MN	2.199	-4.609	13.162	-0.177	66.643	32.513	0.697	0.748
MS	2.183	-4.544	13.549	-0.176	63.135	32.797	1.794	0.750
MO	2.194	-4.556	13.311	-0.177	64.701	32.506	1.190	0.749
MT	2.229	-4.387	12.228	-0.186	69.508	31.567	1.631	0.745
NE	2.215	-4.479	12.801	-0.182	66.485	31.801	1.224	0.747
NV	2.239	-4.114	11.642	-0.193	70.274	31.058	0.957	0.743
NH	2.116	-4.821	14.637	-0.162	64.076	35.810	0.865	0.754
NJ	2.124	-4.770	14.525	-0.164	63.544	35.426	0.736	0.754
NM	2.234	-4.272	12.292	-0.188	66.361	30.940	1.566	0.745
NY	2.131	-4.782	14.404	-0.165	64.169	35.167	1.073	0.754
NC	2.140	-4.690	14.290	-0.167	62.889	34.676	1.222	0.754
ND	2.214	-4.538	12.783	-0.181	67.797	32.026	1.552	0.747
OH	2.158	-4.692	13.970	-0.170	64.134	34.000	1.175	0.752
OK	2.211	-4.451	12.946	-0.181	64.828	31.766	1.526	0.747
OR	2.236	-4.117	11.514	-0.193	72.078	31.428	1.050	0.743
PA	2.138	-4.746	14.299	-0.166	63.887	34.842	1.160	0.753
RI	2.113	-4.806	14.694	-0.162	63.660	35.923	1.133	0.755
SC	2.145	-4.662	14.207	-0.168	62.646	34.410	1.423	0.753
SD	2.214	-4.506	12.785	-0.181	67.153	31.903	1.494	0.747
TN	2.170	-4.615	13.773	-0.173	63.491	33.404	1.515	0.751
TX	2.219	-4.373	12.791	-0.184	64.238	31.361	1.340	0.747
UT	2.237	-4.225	11.972	-0.190	68.752	31.049	0.961	0.744
VT	2.121	-4.815	14.558	-0.163	64.224	35.603	1.176	0.754
VA	2.139	-4.711	14.294	-0.167	63.257	34.741	0.906	0.753
WA	2.234	-4.191	11.612	-0.191	72.381	31.619	0.896	0.744
WV	2.148	-4.701	14.147	-0.168	63.617	34.393	1.733	0.753
WI	2.186	-4.646	13.443	-0.175	65.767	32.966	1.139	0.749
WY	2.229	-4.364	12.301	-0.186	68.333	31.389	1.509	0.745

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

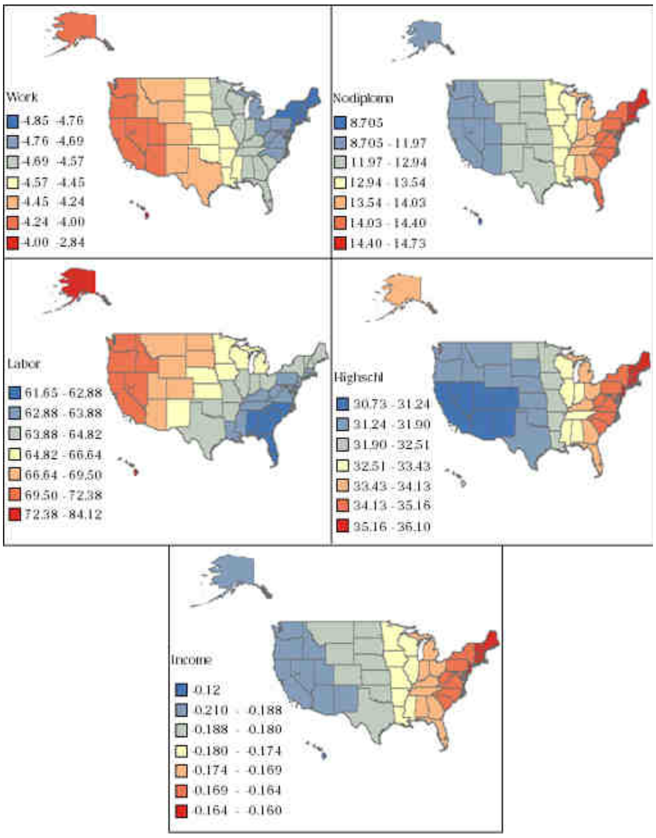

Figure 2: Local variation of parameter estimates by states

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 α=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

[1] NHTSA. *Overview: Traffic Safety Facts 2017 Data*. Pub. DOT HS 811 753, National Highway Traffic Safety Administration, United States Department of Transportation, Washington, D.C., 2013.

[2] Ivan, J. and P. O’Mara. Prediction of Traffic Accident Rates Using Poisson Regression. In *Proceedings of the 76th Annual Meeting of the Transportation Research Board*, National Research Council, Washington, DC, January 1997.

[3] Mensah, A., E. Hauer. Two Problems of Averaging Arising in the Estimation of the Relationship between Accidents and Traffic Flow. In *Journal of Transportation Research Board,* 1635, 1998, pp. 37–43.

[4] Abdel-Aty, M., A. Radwan. Modeling Traffic Accident Occurrence and Involvement. In *Accident Analysis and Prevention*, 2000, 32, pp. 633–642.

[5] Persaud, B., D. Lord and J. Palmisano. Calibration and Transferability of Accident Prediction Models for Urban Intersections. *In Journal of Transportation Research Board,* 1784, 2002, pp. 57–64.

[6] Eustace, D., V.K. Indupuru and P. Hovey. Identification of Risk Factors Associated With Motorcycle Related Fatalities in Ohio. In Journal of Transportation Engineering, 137, 2011, pp. 437-480.

[7] Ackaah, W., M. Salifu. Crash Prediction Model for Two-lane Rural Highways in the Ashanti Region of Ghana. In *IATSS Research.* 35, 2011, pp. 34–40.

[8] Kirk, A.J, J.G. Pigman and K.R. Agent. *Socio-economic Analysis of Fatal Crash Trends*, Final Report, KTC-05-39/TA19-05-1F, Kentucky Transportation Center, University of Kentucky, Lexington, KY, 2005.

[9] Aguero-Valverde, J. and P.P. Jovanis. Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. In *Accident Analysis and Prevention*, 38, 2006, pp.618-625.

[10] Babcock, M.W., T.J. Zlatoper and A.M. Welki. Determinants of Motor Vehicle Fatalities: A Kansas Case Study. In *Journal of the Transportation Research Forum*, 47, 2008, pp. 89–106.

[11] Aguero-Valverde, J., 2013. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. *Accident Analysis & Prevention*, *59*, pp. 365-373.

[12] Boulieri, A., Liverani, S., de Hoogh, K. and Blangiardo, M., 2017. A space–time multivariate Bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(1), pp.119-139.

[13] Liu, C. and Sharma, A., 2018. Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. *Analytic methods in accident research*, *17*, pp.14-31.

[14] Fotheringham, A.S., C. Brunsdon and M. Charlton. *Geographically Weighted Regression*, John Wiley & Sons, LTD, UK, 2002.

[15] Bowerman, B.L. and R.T. O'Connel. *Linear Statistical Models: An Applied Approach*, Duxbury Press, Belmont, CA, 1990.