International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Assessing the Risk of Road Traffic Fatalities Across Sub-Populations of a Given Geographical Zone, Using a Modified Smeed's Model

Christian A. Hesse[1], Francis T. Oduro[2], John B. Ofosu[1] & Emmanuel D. Kpeglo[1]

[1] Department of Mathematical Sciences, Faculty of Informatics and Mathematical Sciences, Methodist University College Ghana. P. O. Box DC 940, Dansoman – Accra, Ghana

[2] Department of Mathematics, College of Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Correspondence: Thomas C. A. Hesse, Department of Mathematical Sciences, Methodist University College, P. O. Box DC 940, Dansoman – Accra, Ghana. E-mail: akrongh@yahoo.com

Received: September 29, 2016 Accepted: October 24, 2016 Online Published: October 31, 2016

doi:10.5539/ijsp.v5n6p121 URL: http://dx.doi.org/10.5539/ijsp.v5n6p121

Abstract

Smeed (1949) provided a regression model for estimating road traffic fatalities (RTFs). In this paper, a modified form of Smeed's (1949) model is proposed for which it is shown that the multiplicative error term is less than that of Smeed's original model for most situations. Based on this Modified Smeed's model, Bayesian and multilevel methods are developed to assess the risk of road traffic fatalities across sub populations of a given geographical zone. These methods consider the parameters of the Smeed's model to be random variables and therefore make it possible to compute variances across space provided there is significant intercept variation of the regression equation across such regions. Using data from Ghana, the robustness of the Bayesian estimates was indicated at low sample sizes with respect to the Normal, Laplace and Cauchy prior distributions. Thus the Bayesian and Multilevel methods performed at least as well as the traditional method of estimating parameters and beyond this were able to assess risk differences through variability of these parameters in space.

Keywords: risk, Bayesian, multilevel, road traffic fatalities

1. Introduction

Smeed (1949) proposed a model for estimating road traffic fatalities (RTFs) in his paper. He showed that the formula

$$\frac{D}{N} = 0.0003\left(\frac{N}{P}\right)^{-\frac{2}{3}} \quad \text{.....(1)}$$

(were D = Number of RTFs, P = population size and N = number of vehicles in use) gave a fairly good fit to the data from 20 countries, including European countries, USA, Canada, Australia and New Zealand.

Ponnaluri (2012) used data from all states in India to develop seven different models for predicting RTFs and also examined if the individual models were more relevant for application. The seven models, including that of Smeed's, were tested for fitness with the actual data. Smeed's model was found to give the best fit. He showed that the original Smeed formulation cannot simply be discounted due to reasons cited by many researchers. This is because Smeed's model is ***parsimonious in parameter usage*.** According to Ponnaluri (2012), Smeed's model appears to be observation-driven, evidence-based, and logically valid in measuring the ***per vehicle fatality rate*.**

The predominant factors affecting RTFs are not the same as those of road traffic accidents (RTAs). Exposures to risk of RTFs (such as human error, environmental/weather, nature of the road and condition of vehicle) are predominant factors influencing road traffic accidents within a geographical region. However, the rate of RTFs is determined by vulnerability to risk (such as insufficient ambulance and emergency medical services, improper pre-hospital care for RTA trauma patients, inadequate safety mechanism in vehicles).

Exposure to risk of RTFs and vulnerability to risk of RTFs are not correlated. Thus, high exposure does not necessarily imply high vulnerability. For instance, Greater Accra Region in Ghana, with the highest exposure to the risk of RTF (due to high population and vehicular densities), has the lowest RTF rate among all the other 9 regions in Ghana. Whilst the three Northern regions of Ghana, with the lowest population density have the highest rate of RTFs (Hesse and Ofosu, 2015). Nigeria and Ghana have almost the same vehicular density. However, inhabitants of Nigeria are more vulnerable

to die as result of road traffic accidents. Developing countries, with only about 10% of the world motorization, account for about 85% of annual RTFs in the world (WHO, 2004, 2009). Thus, developed countries, though have ***greater exposure to risk of RTFs*** due to high vehicular density, however less vulnerable to RTFs compared to developing countries.

Two predominant factors that determine risk of RTFs in a geographical region are

(1) Safety mechanism in vehicles (such as anti-lock braking systems (ABS), air bags and seatbelts),

(2) Emergency medical services (such as Ambulance service).

One reason why developing countries are more vulnerable to risk of RTF is due to the fact that a large proportion of road traffic accident trauma patients in these regions do not have access to formal emergency medical services (Tiska, et al., 2002). Secondly, the ages of vehicles and availability of modern safety mechanisms in vehicles plying the roads in these regions have significant effect on the consequences of road traffic accidents. It is obvious that if greater attention is paid on improving road safety mechanisms (such as anti-lock braking systems (ABS), air bags, better design of cars and increased wearing of seatbelts in cars) there could be substantial benefits in reducing injuries and fatalities with respect to road traffic accidents in developing countries (Hesse, et al., 2014).

Smeed's model is of the form

$$\frac{D}{N} = \alpha\left(\frac{N}{P}\right)^{\beta} e, \quad \text{.....(2)}$$

where D = Number of RTFs, P = population size, N = number of vehicles in use, e = multiplicative error term, and α & β are parameters to be estimated. Equation (2) can be expressed as

$$Y = \alpha X^{\beta} e, \quad \text{.....(3)}$$

where, the predictor variable is $X = N/P = vehicular\ density$ and the dependent variable is $Y = D/N =$ *per vehicle fatality rate*.

The factors affecting RTAs correspond to exposure X while the factors affecting RTFs correspond to vulnerability given the same exposure. In Smeed's model exposure is measured by the variable X whereas vulnerability for a given X is captured by the parameters α and β.

Let X_1 (with $Y = Y_1$) and X_2 (with $Y = Y_2$) be two predictor variables of two geographical regions such that $X_1 = X_2$. If $Y_1 \neq Y_2$, then the different values of Y is not based on X but is due to the fact that α and β vary across the two geographical regions. It therefore follows that, the parameters of Smeed's model vary from one geographical region to another. Thus, one could use these parameters to assess variability of the risk of RTFs across geographical regions.

Smeed (1949) and other related studies by Ponnaluri (2012), Ghee *et al*., (1997), Bener and Ofosu (1991), Jacobs and Bardsley (1977), Fouracre and Jacobs (1977) used least squares regression (LSR) method to estimate the parameters. However, the LSR approach:

- does not allow the variability of the parameters,
- is *very sensitive* to violation of the normality assumption.

Thus, we need an estimation method that:

(1) is robust with respect to the assumptions of the model,

(2) could be used to estimate the variance of the parameters across geographical regions,

(3) enables us compare the risk of RTFs across the geographical regions.

As a general objective, therefore, this study aims at developing statistical methodology, based on Smeed's model, for assessing the risk of RTFs across sub-populations of a given geographical zone. The first specific objective is to develop a modified Smeed's model. Secondly, based on the modified Smeed model, the study seeks to develop and use

- the Bayesian analysis approach to derive an estimator, based on a prior distribution that is robust with respect to the normality assumption,
- the multilevel analysis approach to compare the risk of RTFs across geographical regions.

Finally, the study seeks to use data from Ghana to validate the developed method and to assess the robustness of the model.

2. Method

2.1 A Modified Smeed's Model

Smeed's model in Equation (2) measures per vehicle fatality rate. Multiplying both sides of (2) by N/P, we obtain

$$\frac{D}{P} = \alpha\left(\frac{N}{P}\right)^{\beta}\left(\frac{N}{P}e\right). \quad \text{.....(4)}$$

The modified Smeed's model of this study, which estimates the ***per capita fatality rate*** (also called [1]***public health risk indicator***), is of the form

$$\frac{D}{P} = \alpha\left(\frac{N}{P}\right)^{\beta} u, \quad \text{.....(5)}$$

where $u = \left(\frac{N}{P}e\right) < e$ provided $N < P$.

Table A1, in the Appendix, is an extract from the list of countries with ranks based on the number of road motor vehicles per 1,000 inhabitants. For every country in the world, except San Marino, the number of registered vehicles in use, *N*, is less than the population size, *P*. Since $N < P$ for most situations, it follows that the multiplicative error term *u* in the modified Smeed's model of this study is less than that of Smeed's original model, making the modified Smeed's model preferred.

The modified Smeed's model is ***intrinsically linear***. Thus, Equation (5) can be transformed to a linear model by a logarithmic transformation of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n. \quad \text{.....(6)}$$

For example, Equation (5) can be written in the form

$$\left.\begin{array}{l} \ln D = \ln\alpha + \beta \ln N + (1-\beta)\ln P + \ln u. \\ \text{or} \\ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \end{array}\right\} \quad \text{.....(7)}$$

where $y_i = \ln D$, $x_{i1} = \ln N$, $x_{i2} = \ln P$, $\beta_0 = \ln\alpha$, $\beta_1 = \beta$, $\beta_2 = \ln(1-\beta)$ and $\varepsilon_i = \ln u_i$. Another possible linear transformation of Equation (5) is of the form

$$\left.\begin{array}{l} \ln(D/P) = \ln\alpha + \beta \ln(N/P) + \ln u, \\ \text{or} \\ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \end{array}\right\} \quad \text{.....(8)}$$

where $\beta_0 = \ln\alpha$, $\beta_1 = \beta$, $x_i = \ln(N/P)$, $y_i = \ln(D/P)$ and $\varepsilon_i = \ln u_i$, $i = 1, 2, \ldots, n$.

The linear transformation in Equation (8) is preferred to that of Equation (7) because of the following reason. Since D/P is a risk indicator (known as ***Public Health Risk indicator***) used in epidemiological studies, it follows that any one-to-one relation of this indicator, such as $Y = \ln(D/P)$, can also be used as risk indicator of RTF. This is in sync with the general objective of this study (see Hesse & Ofosu, 2014).

2.2 Bayesian Approach to Estimation of Regression Parameters

In this Section, we develop, using the modified Smeed model, a Bayesian approach to derive an estimator, based on a given prior distribution, that is robust with respect to the normality assumption of the model.

The multiple linear regression model in (6), with *k* predictor variables, can be expressed as

$$y_i = \boldsymbol{\beta}'\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \ldots, n \quad \text{.....(9)}$$

where $\boldsymbol{x}_i' = (1, x_{1i}, x_{2i}, \ldots, x_{ki})$. It is assumed that the unknown parameter vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots \beta_k)$ is a value of some multivariate random variable with a multivariate prior distribution. The range of possible values that the regression coefficients $\beta_0, \beta_1, \ldots \beta_k$ can take is $-\infty$ to $+\infty$. Thus, the largest possible domain of the prior distribution is the set of all real numbers. This limits us to distribution which can take both negative and positive values. Therefore, the most suitable prior distributions are the bivariate Normal, Laplace and Cauchy distributions.

Two Bayesian methods were used in estimating the parameters in Equation (9). These are the 'conjugate prior' method

[1] National Road Safety Commission of Ghana (2011). Building and Road Research Institute (BRRI), *Road Traffic Crashes in Ghana*, Statistics

and the maximum a posteriori method which are discussed in the following sequel.

Conjugate Prior

In this section, we assume that the random variable Y, with components y_i, in Equation (9), has the normal distribution with mean $\boldsymbol{\beta}'\boldsymbol{x}$ and variance σ^2. Thus, the likelihood function will also follow a normal distribution. Since the normal distribution is conjugate to itself (or *self-conjugate*) with respect to a normal likelihood function, choosing a bivariate normal prior over $\boldsymbol{\beta}$ will ensure that the posterior distribution is also normal. The conditional p.d.f. of Y is then given by

$$f_Y\left(y_i|\boldsymbol{\beta}\right)=\frac{1}{\boldsymbol{\sigma}\sqrt{2\boldsymbol{\pi}}}\exp\left\{-\frac{1}{2\boldsymbol{\sigma}^2}\left(y_i-\boldsymbol{\beta}'\boldsymbol{x}\right)^2\right\},\quad |y_i|\geq 0. \qquad \text{.....(10)}$$

The likelihood function is given by (see Mettle et al., 2016)

$$f_{\boldsymbol{Y}}\left(\boldsymbol{y}|\boldsymbol{\beta}\right)=\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\boldsymbol{\beta}'\boldsymbol{x}_i\right)^2\right\},\ \boldsymbol{y}=(y_1,\,y_2,\,...,\,y_n). \qquad \text{.....(11)}$$

It is assumed that $\boldsymbol{\beta}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}=\left(\mu_0,\,\mu_1,...,\,\mu_k\right)$ and covariance matrix $\boldsymbol{\Sigma}$. Thus, the p.d.f. of $\boldsymbol{\beta}$ is

$$p\left(\boldsymbol{\beta}\right)=\frac{1}{2\pi}\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)\right\} \qquad \text{.....(12)}$$

where $\boldsymbol{\Sigma}^{-1}=\begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0k} \\ a_{10} & a_{11} & a_{12} & \cdots & a_{1k} \\ a_{20} & a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{k0} & a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}$. The posterior distribution can therefore be expressed as

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)=k\ f\left(\boldsymbol{y}|\boldsymbol{\beta}\right)p\left(\boldsymbol{\beta}\right)=k\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\boldsymbol{\beta}'\boldsymbol{x}_i\right)^2-\frac{1}{2}\left[\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)\right]\right\}. \qquad \text{.....(13)}$$

The function under the exponent in Equation (13) can be written as

$$\begin{aligned} Q(\boldsymbol{\beta}) \;=\; & \left(\frac{n}{\sigma^2}+a_{00}\right)\beta_0^2+\sum_{j=1}^{k}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}^2+a_{jj}\right)\beta_j^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}y_i+\sum_{l=1}^{k}a_{1l}\mu_l\right)\beta_0 \\ & -2\sum_{j=1}^{k}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}y_i+\sum_{l=1}^{k}a_{jl}\mu_l\right)\beta_j+2\sum_{j=0}^{k-1}\sum_{s=j+1}^{k}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}x_{si}+a_{js}\right)\beta_j\beta_s+v \qquad \text{.....(14)} \end{aligned}$$

where v is the constant term, independent of β_j. Therefore the posterior p.d.f. of $\boldsymbol{\beta}$ can be written as

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)\;=\;ke^{-\frac{1}{2}Q(\boldsymbol{\beta})}. \qquad \text{.....(15)}$$

Hence Equation (13) follows the multivariate normal distribution with mean vector given by

$$\boldsymbol{\mu}_{\boldsymbol{\beta}}=-\frac{1}{2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{C}, \qquad \text{.....(16)}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is a $(k+1)\times(k+1)$ matrix with inverse $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}=\left(m_{ij}\right)$ whose elements are given as

$$\left.\begin{aligned} m_{00} &= \frac{n}{\sigma^2}+a_{00}, \\ m_{j0} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}+a_{j0}, \quad j=1,2,...k, \\ m_{0j} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}+a_{0j}, \quad j=1,2,...k, \\ m_{ij} &= \frac{1}{\sigma^2}\sum_{l=1}^{n}x_{il}x_{jl}+a_{ij}, \quad i\neq j, \\ m_{ii} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ij}^2+a_{ii}, \quad j=1,2,...k. \end{aligned}\right\} \qquad \text{.....(17)}$$

and $\boldsymbol{C}$ is a column vector of order $(k + 1)$ with elements given as

$$\left.\begin{aligned} C_0 &= -2\left(\tfrac{1}{\sigma^2}\sum_{i=1}^{n} y_i + \sum_{j=1}^{k} a_{0j}\mu_j\right) \\ C_i &= -2\left(\tfrac{1}{\sigma^2}\sum_{i=1}^{n} y_i x_{li} + \sum_{j=1}^{k} a_{ij}\mu_j\right), \quad l = 1, 2, ..., k. \end{aligned}\right\} \quad \text{.....(18)}$$

Let $\hat{\boldsymbol{\beta}}_l = \left(\hat{\beta}_{0l},\ \hat{\beta}_{1l},\ ...,\ \hat{\beta}_{kl}\right);\ l = 1,\ 2,\ 3,\ ...,\ n$ be the l^{th} jackknife estimate of the regression. Then the estimate of the mean vector $\boldsymbol{\mu}$ of the random vector $\boldsymbol{\beta} = \left(\beta_0,\ \beta_1,\ ...,\ \beta_k\right)$ is given as $\hat{\boldsymbol{\mu}} = \left(\hat{\mu}_0,\ \hat{\mu}_1, ..., \hat{\mu}_k\right)'$, where

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\beta_{ji}, \quad j = 0,\ 1,\ ...,\ k. \quad \text{.....(19)}$$

and an estimate of the covariance matrix of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\mu}}_j\right)\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\mu}}_j\right)' = \left(\hat{a}_{ij}\right). \text{.....(20)}$$

The estimate of the standard error of the i^{th} coefficient, based on the Bayesian estimate is the square root of the i^{th} diagonal elements of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$.

Maximum a Posteriori Method

The goal here is to find the parameter estimates that maximizes the posterior probability of the parameters given the data. This corresponds to

$$\boldsymbol{\beta}_{MAP} = \arg\max_{\boldsymbol{\beta}}\ p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \quad \text{.....(21)}$$

We resort to sampling techniques, such as Markov chain Monte Carlo (MCMC), to get samples from the posterior distribution. The following algorithm is the description for the multivariate Metropolis Hastings procedure (Steyvers, 2011):

1. Set $t = 1$
2. Generate an initial value for $\beta_j \sim U(u_{1j},\ \mu_{2j}),\quad j = 0,\ 1,\ ...,\ k.$
3. Repeat

   $t = t + 1$

   Do a MH step on β_j,

   Generate a proposal $\beta_j^* \sim N(\beta_j,\ \sigma_j^2)$;

   Evaluate the acceptance probability $a = \min\left[1,\ \frac{p^*(\boldsymbol{\beta}|\boldsymbol{y})}{p(\boldsymbol{\beta}|\boldsymbol{y})}\right]$;

   Generate a u from a Uniform(0, 1) distribution

   If $u \le a,$ accept the proposal and set $\beta_j = \beta_j^*,\quad j = 0,\ 1,\ ...,\ k.$
4. Until $t = T$.

2.3 Multilevel Random Coefficient (MRC) Model

In this Section, we develop a Multilevel Analysis approach to estimate the regional distribution of parameters based on the modified Smeed's model and use them to compare the risk of RTFs across geographical regions.

Assuming the population is stratified into J geographical regions with n_j observations in each class, Equation (6) becomes

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + ... + \beta_{kj}x_{kij} + \varepsilon_{ij}, \\ &= \beta_{0j} + \sum_{l=1}^{k}\beta_{lj}x_{lij} + \varepsilon_{ij}, \qquad \begin{matrix} i = 1,\ 2,\ ...,\ n_j \\ j = 1,\ 2,\ ...,\ J \end{matrix} \end{aligned} \quad \text{.....(22)}$$

Across all geographical regions, $\beta_j = (\beta_{0j}, \beta_{1j}, ..., \beta_{kj})$ are assumed to have multivariate normal distribution (Hox, 2010). Thus, each β_{lj} ($l = 0, 1, 2, ..., k$) can be modeled as

$$\beta_{0j} = \gamma_{00} + \gamma_{01} z_j + u_{0j} \quad \text{.....(23)}$$

$$\beta_{lj} = \gamma_{l0} + \gamma_{l1} z_j + u_{lj}, \qquad l = 1, ..., k \text{ and } j = 1, 2, ..., J \quad \text{.....(24)}$$

From Equations (22), (23) and (24), we have

$$y_{ij} = \gamma_{00} + \gamma_{01} z_j + u_{0j} + \sum_{l=1}^{k} \left(\gamma_{l0} + \gamma_{l1} z_j + u_{lj}\right) x_{lij} + \varepsilon_{ij}, \qquad \begin{matrix} i = 1, 2, ..., n_j \\ j = 1, 2, ..., J \end{matrix} \quad \text{.....(25)}$$

$u_{lj} \sim N(0, \tau_l)$, $l = 0,1,...,k$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$. Y has the normal distribution with mean

$$\mu = \gamma_{00} + \gamma_{01} z_j + \sum_{l=1}^{k} \left(\gamma_{l0} + \gamma_{l1} z_j\right) x_{lij} \quad \text{.....(26)}$$

and variance

$$v = \tau_0 + \sum_{l=1}^{k} x_{lij}^2 \tau_l + 2 \sum_{l \neq r} x_{lij} x_{rij} \tau_{lr} + \sigma^2. \quad \text{.....(27)}$$

The parameters to be estimated are γ_{l0}, γ_{l1}, τ_l, $\tau_{lr} (l \neq r)$ and σ^2, $l = 0, 1, \ldots, k$, where $\tau_l = \text{var}(u_{lj})$, $\tau_{lr} = \text{cov}(u_{lj}, u_{rj})$, and $\text{var}(\varepsilon_{ij}) = \sigma^2$.

If τ_0 differs significantly from 0, then the parameters of the modified Smeed's model can be used to compare the risk of RTFs across the J geographical regions.

Equating the partial derivatives of the likelihood function to zero, we obtain the maximum likelihood estimators of the parameters γ_{l0}, γ_{l1}, τ_l, $\tau_{lr} (l \neq r)$ and σ^2 as $\hat{\gamma}_{l0}$, $\hat{\gamma}_{l1}$, $\hat{\tau}_l$, $\hat{\tau}_{lr} (l \neq r)$ and $\hat{\sigma}^2$ respectively.

3. Validation of Method Using Data from Ghana

In this section the study seeks to use data from Ghana to validate the

(1) Bayesian method and to assess the robustness of the model

(2) multilevel method and to compare the risk of RTFs across the 10 geographical regions.

3.1 Validation of Bayesian Method

(i) Conjugate Prior Method

Table A2, in the Appendix, gives the estimated population size and the number of motor vehicles and road traffic fatalities in Ghana (1991 – 2012). It can be seen that, the distribution of $\ln(D/P)$, with a Shapiro-Wilks normality p-value of 0.201, is closer to the normal distribution compared to that of $\ln(D)$ with a corresponding p-value of 0.086. This confirms that the logarithmic transformation in Equation (8) is preferred.

The 19 jackknife sample estimates of β_0 and β_1, based on the national data, derived from the values of y_i and x_i in Table A2 are given in Table A3. Based on Equations (19) and (20), jackknife estimate of the mean vector and covariance of the random vector $\boldsymbol{\beta}$ is computed as follows

$$\hat{\boldsymbol{\mu}} = (-8.3105,\ 0.3192) \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.001860 & 0.000504 \\ 0.000504 & 0.000139 \end{pmatrix}.$$

Based on Equations (17) and (18),

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \begin{pmatrix} 0.0017421 & 0.0004712 \\ 0.0004712 & 0.0001297 \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{C}} = \begin{pmatrix} 646278.208 \\ -2353969.324 \end{pmatrix}.$$

Thus, the posterior Bayes estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = -\tfrac{1}{2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{C} = \begin{pmatrix} -8.31048 \\ 0.319162 \end{pmatrix}. \quad \text{.....(28)}$$

Table 1 shows the coefficients estimates and the corresponding standard errors for the least square and the conjugate prior methods.

Table 1. Comparison of Coefficients of Least Square and Conjugate Prior Methods

	Methods			
	Least Squares		**Conjugate Prior**	
	Coefficient	Standard Error	Coefficient	Standard Error
β_0 = intercept	–8.31179	0.17386	–8.31048	0.04174
β_1 = coefficient of *x*	0.31879	0.04555	0.31916	0.01139
Coefficient of determination	0.7423		0.7423	

It can be seen from Table 2, that the estimated coefficients β_0 and β_1, are almost the same for the least squares and the conjugate prior methods. Both methods also reported the same coefficient of determination R^2. The conjugate prior estimates recorded comparatively very small standard errors; making the conjugate prior method preferred.

(ii) Maximum a posteriori method

Our objective here is to determine the parameter estimates that maximize the posterior distribution given the data with respect to the bivariate Normal, Laplace and Cauchy prior distributions.

Bivariate Normal prior distribution

The prior distribution in Equation (12) can be written in terms of ρ as

$$p(\beta_0,\ \beta_1|\,\boldsymbol{y}) = k\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{19}\left(y_i-\beta_0-\beta_1 x_i\right)^2-\frac{1}{2}q\right\}, \quad \text{.....(29)}$$

where $q=\frac{1}{1-\rho^2}\left\{\left(\frac{\beta_0-\mu_0}{\sigma_0}\right)^2-2\rho\left(\frac{\beta_0-\mu_0}{\sigma_0}\right)\left(\frac{\beta_1-\mu_1}{\sigma_1}\right)+\left(\frac{\beta_1-\mu_1}{\sigma_1}\right)^2\right\},\ -\infty<\beta_0<\infty,\ -\infty<\beta_1<\infty,$

$\sigma_0^2=\text{var}(\beta_0),\ \sigma_1^2=\text{var}(\beta_1).$

The Metropolis Hastings algorithm, above, is used to estimate the values of β_0 and β_1. The MATLAB code for the implementation of component-wise Metropolis sampler for the posterior distribution is as given in Listings 1 and 2 in the appendix.

Table 2 shows estimated values of β_0 and β_1 based on least squares, conjugate prior and maximum a posteriori methods. The results show that the estimated coefficients of β_0 and β_1 are almost the same for the least squares, conjugate prior and maximum a posteriori methods of estimates.

Table 2. Comparison of Coefficients of Least Squares, Conjugate Prior and Maximum a Posteriori Methods

	Methods		
	Least Square	**Conjugate prior**	**Maximum a posteriori**
β_0	–8.31179	–8.31048	–8.29094
(**Standard error**)	(**0.17386**)	(**0.04174**)	(**0.03978**)
β_1	0.31879	0.31916	0.32460
(**Standard error**)	(**0.04555**)	(**0.01139**)	(**0.01098**)

Laplace Prior Distribution

It is assumed that $\boldsymbol{\beta}=(\beta_0,\beta_1)$ has a bivariate Laplace distribution with mean vector $\boldsymbol{\mu}=(\mu_0,\ \mu_1)$. The joint p.d.f. is given by

$$f(\beta_0,\beta_1) = \frac{1}{4b_0b_1}e^{-\left[\frac{1}{b_0}|\beta_0-\mu_0|+\frac{1}{b_1}|\beta_1-\mu_1|\right]}, \quad \text{.....(30)}$$

$-\infty<\alpha<\infty,\ -\infty<\beta<\infty,\ b_0>0,\ b_1>0.$ Thus, the posterior distribution can be expressed as

$$p(\beta_0,\ \beta_1|\,\boldsymbol{y}) = k\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2-\frac{1}{\hat{b}_0}|\alpha-\hat{\mu}_0|-\frac{1}{\hat{b}_1}|\beta-\hat{\mu}_1|\right\}. \quad \text{.....(31)}$$

Using the above algorithm, the maximum a posteriori estimates of β_0 and β_1 to be –8.320085 and 0.317051, respectively,

with standard errors of 0.039047 and 0.010450.

Cauchy Prior Distribution

The bivariate random variable $\boldsymbol{\beta} = (\beta_0, \beta_1)$ has the Cauchy distribution if the p.d.f. can be expressed in the form given in the following form

$$f(\beta_0, \beta_1) = \frac{1}{2\pi}\left[\frac{1}{(\beta_0 - a)^2 + (\beta_1 - b)^2 + 1}\right]. \quad \text{.....(32)}$$

Thus, the posterior distribution can be expressed as

$$p(\beta_0, \beta_1 | \boldsymbol{y}) = k \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{19}(y_i - \beta_0 - \beta_1 x_i)^2\right\}\left\{\frac{1}{(\beta_0 - \hat{a})^2 + (\beta_1 - \hat{b})^2 + 1}\right\}. \quad \text{.....(33)}$$

The component-wise Metropolis-Hastings sampler for the posterior distribution based on the MATLAB codes, gave maximum a posteriori estimates of β_0 and β_1 to be –8.312857 and 0.317400, respectively.

The resulting posterior Bayesian estimates for the Normal, Laplace and Cauchy prior distributions are summarized in the Table 3. Given a sample size 19, the posterior Bayes estimate is reasonably consistent for the Normal, Laplace and Cauchy prior distributions.

Table 3. Posterior Bayesian estimates for different priors with a sample size of 19

	Prior distribution				
	Normal		**Laplace**		**Cauchy**
	Estimate	Standard Error	Estimate	Standard Error	
$\boldsymbol{\beta_0}$	–8.31048	0.04174	–8.32009	0.039047	–8.31286
$\boldsymbol{\beta_1}$	0.31916	0.01139	0.31705	0.010450	0.31740

Table 4 shows the posterior Bayesian estimates of β_0 and β_1 at four different sample sizes (5, 10, 15 and 19) using the Normal, Laplace and Cauchy prior distributions. It can be seen that, at sample sizes of 5 and 10, the posterior Bayesian estimates of β_0 and β_1 are not consistent across the three prior distributions used. Thus, the estimated values of β_0 and β_1 are said to be sensitive with respect to the prior distribution. At a sample size of 15 or more, the model becomes insensitive to the prior distribution. The relative influence of the prior distribution decreases while that of the data increases with a sample size of 15 or more. It can also be seen that the posterior Bayesian estimate is reasonably consistent for the Laplace prior distribution across all four sample sizes used. Even at a sample size of 5 where the normality assumption was violated, the estimates based on the Laplace prior distribution was robust. Thus, the Laplace prior distribution is preferred when the sample size is small.

Table 4. Bayesian estimates with respect to sample size and prior distribution

	Prior distribution					
Sample size	**Normal**		**Laplace**		**Cauchy**	
n	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$
5	–5.99608	0. 99041	–8.30608	0. 31923	–5.13317	1.01961
(**Standard error**)	(**0.67355**)	(**0.02767**)	(**0.61978**)	(**0.02195**)		
10	–8.29381	0.32272	–8.29637	0.32863	–7.72230	0.46478
(**Standard error**)	(**0.44057**)	(**0.01629**)	(**0.43884**)	(**0.01596**)		
15	–8.31195	0.31647	–8.29288	0.32266	–8.31034	0.31694
(**Standard error**)	(**0.36057**)	(**0.01328**)	(**0.35747**)	(**0.01298**)		
19	–8.31048	0.31916	–8.32009	0.31705	–8.31286	0.31740
(**Standard error**)	(**0.31916**)	(**0.01139**)	(**0.31705**)	(**0.01045**)		

3.2 Validation of Multilevel Method

Table A4, in the Appendix, shows the value of $x_{ij} = \ln(N_{ij}/P_{ij})$ and the corresponding values of $y_{ij} = \ln(D_{ij}/P_{ij})$ for the ten regions of Ghana. Instead of estimating a separate regression equation for each of the 10 regions in Ghana,

we wish to determine a single model for estimating regional distribution of RTFs. The collection of the regression parameters {$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_{10}$} is assumed to be a random sample of size 10 taken from a population whose distribution depends on the parameters γ_1, γ_2, δ_0, τ_0, τ_1, τ_{01} and σ^2, where $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j})$, $j = 1, 2, \ldots, 10$.

Equations (21), (22) and (23) can be written as

$$\left.\begin{array}{l} y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}, \\ \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j}, \\ \beta_{1j} = \gamma_{10} + u_{1j}. \end{array}\right\} \quad \begin{array}{l} i = 1, 2, \ldots, 19 \\ j = 1, 2, \ldots, 10 \end{array} \quad \text{.......................................(34)}$$

Combining the three equations, we obtain

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j + u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}, \quad j = 1, 2, \ldots, 10. \quad \text{..(35)}$$

$u_{0j} \sim N(0, \tau_0)$, $u_{1j} \sim N(0, \tau_1)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$. Y has the normal distribution with mean $\gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j$ and variance $v = \tau_0 + 2\tau_{01}x_{ij} + \tau_1 x_{ij}^2 + \sigma^2$. Thus, the pdf of Y given $X = x_{ij}$ is

$$f_Y\left(y_{ij} \middle| X = x_{ij}\right) = \frac{1}{\sqrt{2v\pi}} \exp\left[-\frac{1}{2v}\left(y_{ij} - \gamma_{00} - \gamma_{10}x_{ij} - \gamma_{01}\bar{x}_j\right)^2\right] \text{....................................(36)}$$

Three models are considered in the next section.

(i) The Unconditional Means Model, M_0

An unconditional means model does not contain any predictors, but includes a random intercept variance term for groups. In this section, we examine if there will be significant intercept variation (τ_0). If τ_0 does not differ significantly from 0, there may be little reason to use random coefficient modeling since simpler Ordinary Least Squares (OLS) modeling will suffice. Equation (34) therefore becomes

$$\left.\begin{array}{l} Y_{ij} = \beta_{0j} + \varepsilon_{ij}, \\ \beta_{0j} = \gamma_{00} + u_{0j} \end{array}\right\} \quad \text{..(37)}$$

Therefore

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij}. \quad \text{...... ....(38)}$$

Application of the nlme package in *R*, using data in Table A3, shows that there is significant intercept variation in terms of *y* scores across the 10 regions.

(ii) Random Intercept Model, M_1

In this model, it is assumed that the intercept β_{0j} vary across the 10 geographical regions whilst the slope β_{1j} remain constant. Equation (34), therefore, becomes

$$\left.\begin{array}{l} y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}, \\ \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j} \\ \beta_{1j} = \gamma_{10}, \end{array}\right\} \quad \text{........................ ....(39)}$$

Combine the three rows into a single equation,

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j + u_{0j} + \varepsilon_{ij}, \quad j = 1, 2, \ldots, 10. \quad \ldots \quad \text{...........................(40)}$$

The maximum likelihood estimates of the parameters, using data from Table A3 and nlme package in *R*, are given in Table 5.

(iii) Random slope model M_2

In section, we continue our analysis by trying to explain the third source of variation, namely, variation in the slope, τ_1. The model that we test is:

$$\left.\begin{array}{l} y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij}, \\ \alpha_j = \gamma_0 + \gamma_1\bar{x}_j + e_{\alpha j} \\ \beta_j = \delta_0 + e_{\beta j} \end{array}\right\} \quad \text{...(41)}$$

When we combine the three rows into a single equation in the form

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j + u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}, \quad j = 1, 2, \ldots, 10. \quad \text{..............................(42)}$$

Table 5 presents the parameter estimate and standard errors for the models M_0, M_1 and M_2. All the standard errors of the estimated parameters in model M_2 are smaller than the corresponding values of model M_1. Moreover, the deviance, which

measures the model misfit, is much lower in M_2 as compare to that of M_1 (Hesse, et al., 2014b) Thus, estimate parameters based on model M_2 is preferred.

Table 5. Comparison of models M_0, M_1 and M_2

Model	M_0: intercept only		M_1: with predictor		M_2: with predictor	
Fixed effect	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
γ_{00} = Intercept	-9.6888	0.1401	-10.0756	0.7426	-9.2341	0.2065
γ_{10} = coffiecient of x_{ij}			0.4591	0.0374	0.4459	0.0707
γ_{01} = coefficient of $\bar{x}_j$			-0.5448	0.1658	-0.3384	0.0516
Random part	Parameter	Standard Error	Parameter	Standard Error	Parameter	Standard Error
$\tau_0 = \text{var}(u_{0j})$	0.1891	0.2085	0.2094	0.1447	0.1545	0.1243
$\tau_1 = \text{var}(u_{1j})$					0.0382	0.0618
$\tau_{01} = \text{cov}(u_{0j}, u_{1j})$					0.0766	
$\sigma^2 = \text{var}(\varepsilon_{ij})$	0.1389	0.0855	0.0759	0.0632	0.0630	0.0576
Deviance	198.201		94.554		64.749	

The estimate of regional-level residuals $\hat{u}_{0j}$ and $\hat{u}_{1j}$ and the corresponding values of α and β for each region are given in Table 6.

Table 6. Estimate of regional-level residuals and the values of α and β

Regions	$\hat{u}_{0j}$	$\hat{u}_{1j}$	$\hat{\beta}_0$	$\hat{\beta}_j$	$\hat{\alpha}_j = e^{\hat{\beta}_0}$
Greater Accra	-0.273	-0.138	-8.709877	0.3083572	0.0001649
Ashanti	-0.168	-0.084	-8.073562	0.3614688	0.0003117
Western	-0.085	-0.041	-7.677551	0.4053849	0.0004631
Eastern	-0.470	-0.235	-7.930339	0.2109577	0.0003597
Central	-0.342	-0.170	-7.743066	0.2758323	0.0004337
Volta	-0.037	-0.020	-7.397244	0.4259363	0.0006129
Northern	0.427	0.214	-7.251897	0.6594775	0.0007088
Upper East	0.395	0.198	-7.400873	0.6439825	0.0006107
Upper West	0.703	0.353	-7.206664	0.7993004	0.0007416
Brong Ahafo	-0.152	-0.077	-7.694218	0.3686119	0.0004555

According to National Road Safety Commission (NRSC)[2] of Ghana 2011 report, two key national road traffic fatality indices required for characterization and comparison of the extent and risk of traffic fatality across the ten geographical regions of Ghana are *RTF*s per 100 accidents and *RTF per 100* casualties.

The last two columns of Table 7 give the means of RTFs per 100 accidents and RTFs per 100 casualties for each region from 1991 – 2009. This implies that the risk of dying as a result of road traffic fatality in Greater Accra is relatively low, recording an average rate of 5.7 road traffic fatalities per 100 accidents. Thus, out of every 100 road traffic accidents in the

[2] National Road Safety Commission of Ghana (2011). Building and Road Research Institute (BRRI), *Road Traffic Crashes in Ghana*, Statistics

Greater Accra, about 6 of the victims are likely to die (Hesse and Ofosu, 2015).

Table 7. Parameter estimates and Fatality indices

Regions	$\hat{\alpha}\times10^5$	$\hat{\beta}\times10^2$	*RTF per 100* Accident	*RTF per 100* Casualties
Greater Accra	16.5	30.836	5.7	7.7
Ashanti	31.2	36.147	17.8	12.2
Western	46.3	40.538	16.9	10.7
Eastern	36.0	21.096	19.9	9.7
Central	43.4	27.583	21.8	11.4
Volta	61.3	42.594	23.6	11.2
Northern	70.9	65.948	40.9	18.1
Upper East	61.1	64.398	27.3	17.0
Upper West	74.2	79.930	28.3	14.6
Brong-Ahafo	45.6	36.861	28.6	14.5

We wish to determine if strong positive correlation exist between the parameter estimates of the modified Smeed's model and the fatality indices based on NRSC definition of risk. The p-values in Table 8 show that there is strong positive correlation between the parameter estimates of the modified Smeed's model and the fatality indices. Thus, the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ of the modified Smeed's model can be used as risk indicators of RTFs in Ghana.

Table 8. Correlations coefficients

	$\hat{\alpha}$	$\hat{\beta}$	RTF per 100 Accident	RTF per 100 Casualties
$\hat{\alpha}$	1			
$\hat{\beta}$	0.8312 **(0.003)**	1		
RTF per 100 Accident	0.8424 **(0.002)**	0.6341 **(0.049)**	1	
RTF per 100 Casualties	0.7708 **(0.009)**	0.7610 **(0.010)**	0.9011 **(0.000)**	1

4. Conclusion

A modified Smeed's model,

$$\frac{D}{P} = \alpha\left(N/P\right)^{\beta} u,$$

has been developed. The multiplicative error term *u* in the modified Smeed's model of this study was found to be less than that of Smeed's, making the modified Smeed's model preferred. Using data from Ghana, it was confirmed that the modified Smeed's model for this studies, is relatively more accurate in estimating RTFs in Ghana than the Smeed equation.

Based on the modified Smeed's model of this study, the developed Bayesian method with respect to the Laplace prior distribution was found to be robust to violation of the normality assumption of the model. Using data from Ghana, the sensitivity of the Bayesian estimates at different sample sizes with respect to the Normal, Laplace and Cauchy prior distributions was assessed. At a sample size of 15 or more, the model becomes insensitive to the prior distribution. The posterior Bayesian estimate is consistent for the Laplace prior distribution across all four sample sizes. At a sample size of 5, the estimates based on Laplace prior distribution were robust with respect to violation of the normality assumption of the model.

The parameter estimates of modified Smeed's model can be used as risk indicator of RTFs across geographical regions provided there is significant intercept variation τ_0 of the regression equation across geographical regions. Using data from Ghana, it was shown that the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ across the 10 geographical regions can be used as *risk indicators of RTFs in Ghana*. Thus, the three Northern regions and the Brong-Ahafo region have the highest risk of RTFs.

References

Bener, A., & Ofosu, J. B. (1991). Road traffic fatalities in Saudi Arabia. *Journal of the International Association of Traffic and Safety Sciences, 15*, 35-8.

Fouracre, P., & Jacobs, G. D. (1977). *Further research on road accident rate in developing countries*. TRRL report LR 270. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Ghee, C. Silcock, D. Astrop, A., & Jacobs, G. (1997). *So cio-economic aspects of road accidents in developing countries.* TRL Report 247. Crowthorne: Transport Research Laboratory.

Hesse, C. A., & Ofosu, J. B. (2014). Epidemiology of road traffic accidents in Ghana. *European Scientific Journal, 10*(9), 370-381.

Hesse, C. A., & Ofosu, J. B. (2015). The Effect of Road Traffic Fatality Rate on Road Users in Ghana. *Research Journal of Mathematics and Statistics, 7*(4), 53-59. http://dx.doi.org/10.19026/rjms.7.2207

Hesse, C. A., Ofosu, J. B., & Oduro, F. T. (2014). A Bayesian Model for Predicting Road Traffic Fatalities in Ghana. *Mathematical Theory and Modeling, 4*(8), 1-9.

Hesse, C. A., Ofosu, J. B., & Lamptey, B. L. (2014). A Regression Model for Predicting Road Traffic Fatalities in Ghana. *Open Science Repository Mathematics*, Online(open-access), e23050497. http://dx.doi.org/10.7392/openaccess.23050497.

Hox, J. J. (2010). Multilevel analysis: Techniques and applications 2nd edition. Routledge Taylor and Francis Group.

Jacobs, G., & Bardsley, M. (1977). *Research on road accidents in developing countries*. Traffic engineering & control.

Mettle, F. O., Asiedu, L. Quaye, E. N. B., & Asare-Kumi, A. A. (2016). Comparison of Least Squares Method and Bayesian with Multivariate Normal Prior in Estimating Multiple Regression Parameters. *British Journal of Mathematics and Computer Science, 15*(1), 1-9. http://dx.doi.org/10.9734/BJMCS/2016/23145

Ponnaluri, R. V. (2012). Modeling road traffic fatalities in India: Smeed's law, time invariance and regional specificity. *International Association of Traffic and Safety Sciences,* 36, 75-82. http://dx.doi.org/10.1016/j.iatssr.2012.05.001

Smeed, R. (1949). Some statistical aspects of road safety research. *J. Roy Stats. Soc. Series-A, 12*(1), 1-23. http://dx.doi.org/10.2307/2984177

Steyvers, M. (2011). *Computational statistics with MATLAB.* University of California, Irvine, psiexp.ss.uci.edu/research/teachingP205C/205C.pd.

Tiska, M. A., Adu-Ampofo, M., Boakye, G., Tuuli, L., & Mock, C. N. (2002). A model of prehospital trauma training for lay persons devised in Africa. *Emergency Medical Journal*.

Copyrights