Comparison of Generalized Linear Model (GLM) and Generalized Estimation Equation (GEE) for Modelling Road Accidents from National Data Sets: A Case Study of Great Britain

A.Qadeer Memon

Centre for Transport Studies, University College London, United Kingdom

ABSTRACT

The objective of this paper is to investigate statistical models for road accident data of Great Britain. We compared the results of Generalized Linear Model (GLM) and Generalized Estimation Equation (GEE) techniques with Poisson and negative binomial regression. GEE with negative binomial regression was preferred because it can accommodate the over-dispersion and serial correlation in data. It is found that increase in traffic flow and road length will cause more accidents whereas Time variable will have negative influence on number of daily accidents. It was also concluded that Urban A roads had at least 8 times higher risk than Motorways on all days of week.

Key words: Generalized Linear Model, Generalized Estimation Equation, STATS 19 data.

1. Introduction

Various safety improvement programmes are designed by the planning and development agencies to reduce the number and severity of road traffic accidents. Numbers of accidents are estimated by using the accident prediction models. These models relate the expected number of accidents to some available explanatory variables. Based on the modelling results new road safety initiatives are proposed to improve road safety. If the results drawn from the modelling are not appropriate it may result in reduced road safety and the loss of resources.

There are several techniques available for estimating the number of accidents. In earlier research relationship between accidents and other variables was found by using the conventional multiple regression technique. Multiple regression method assumes that dependent variable is continuously and normally-distributed with a constant variance. Conventional multiple linear regression technique lacks the distributional property necessary to describe adequately random, discrete and non-negative events such as traffic accidents. Various studies like Miaou (1993), Miaou and Lum (1993) have shown that test statistics derived from these models are questionable. In recent studies by Hadi et al, (1995), Anis (1996) significant advances have been made to describe the discrete count traffic accident data and to produce more accurate and reliable models through the use of Generalized Linear Models (GLM) with Poisson and negative binomial distributions.

Maher and Summersgil (1996) showed that the variance of the count data is found to be higher than mean. The extra variation is known as over-dispersion. When using the Poisson regression in presence of over-dispersion, model

parameter estimates will still be close to the true values, but their variance tends to be under-estimated and significance levels of the estimated coefficients will be overstated. In order to address the issue of over-dispersion Abdel-Aty and Radwan (2000), Guevara et al (2004), McCarthy (2005) used negative binomial distribution which allows variance to exceed the mean.

Another important issue that arise in time-series of accident data is the presence of serial correlation. Time-series observations of multiple years of cross-sectional data on highway accident occurrence are often available from public domain. In the presence of serial correlation, the efficiency of the parameter estimates comes into question. Ulfarsson and Shankar (2003) used Negative Multinomial (NM) model to account the serial correlation present in the models. Lord and Parsad (2000) used Generalized Estimation Equation (GEE) to handle the temporal correlation in data. More recently Bossche (2006) used ARIMA models and regression models with ARIMA errors by considering the count data as time-series data. He used the calendar data with seasonal trend. Wang et al (2006) used Generalized Estimating Equations (GEE) to accommodate serial correlation in data for modelling accidents at different intersections.

In Great Britain STATS 19 data have been used in modelling the accidents at national, county and city levels in various studies some of which are as under;

Edwards (1998) used STATS 19 data for analysing the relationship between weather conditions and accident occurrence at county level by using monthly and yearly aggregated data. She used multiple regression to model accidents. The results showed a consistent pattern to weather related accidents

which followed the broad changes in the weather from north to south and west to east.

Noland and Quddus (2005) carried out spatial analysis with the aim of examining the congestion effect on the road safety by using the enumeration district data of London. Negative binomial models were used to examine whether factors affecting casualties differed during congested time periods as opposed to un-congested time periods. It was found that spatial differences between inner and outer London were minor and the differences between the congested and un-congested time period models were not conclusive.

2. Aim of Research

The aims of this research are as under;

- To identify the suitable technique to model the road accidents at national level by using the national accident data sets of Great Britain.
- To compare the results of models to estimate the daily road accidents by road class in Great Britain.
- To identify relationship between number of accidents and road class, road length, flow per day, day of week and month.
- To compare the risk estimated for different combinations of road and Day of week.

Generalised Linear Model (GLM) and Generalised Estimation Equation (GEE) is used to develop the accident prediction models at national level. The national accident data set of STATS 19 is used for this study. The data set is prepared by

extracting the information about daily accidents by road class from STATS 19 data from 1999 to 2002. The results of this research will help various planning and rescue agencies to develop road safety intervention programs and identify the significant variables in a better way. Because using wrong method for estimating the daily accidents may lead to draw wrong inferences and this will result in putting more emphasis on those explanatory variables which are actually less significant. The results will also enable the agencies to allocate the resources in a better way by anticipating how many accidents are likely to occur on any day of week by road class throughout the study area. The results of this study may also help to promote safer usage of road.

3. DATA USED

The road accident statistics in the Great Britain are compiled by police. All roads accidents involving human death or personal injury occurring on the highway are notified to police within 30 days of occurrence. For each road accident which has occurred, police authorities complete a STATS 19 form which provides details of the accident circumstances, information for each vehicle which was involved and information of each person who was injured in accident. This whole data set is maintained by the Department for Transport. The four years accident data from 1999 to 2002 was used for modelling the accidents. The road classification of STATS 19 data and traffic flow data which is obtained from Department for Transport is not same. The roads are classified as Motorway, A, B, C and Unclassified in the STATS 19 data. However the roads are classified as Motorways, Rural A, Urban A, Rural minor, and Urban minor road in available

traffic flow data. Thus in order to make joint use of this data, classification of roads is rearranged by using speed limit data as shown in Table 1.

MS Access queries were used to select accidents for each road class. This was achieved by using the road classification of STATS 19 data and speed limit of the different roads. These access files were exported to SPSS to develop a new data set which consists of the information about all the numbers of daily road accidents by road class which occurred during the four year period from 1st January 1999 to 31st December 2002 on all 5 classes of roads. The data set consisted of 7,305 observations. Each observation represents daily accidents by road class for whole Great Britain. Further explanation about the number of observation in the data set is given in Table 2. Road length and Daily traffic flow by road class in billion vehicle kilometres was obtained from the Department for Transport. Daily and monthly corrections for all motor vehicles on Motorways, Non built-up roads and Built-up roads were applied to the daily traffic flow to adjust the variation in the flow by different days of week and month of year.

4. METHODOLOGY

Generalized Linear Model (GLM) and Generalized Estimation Equation (GEE) with Poisson and negative binomial regression were used. The description of the both models is described below:

4.1 Generalized Linear Model (GLM)

The Generalized linear model (Nelder, 1983) extends the standard linear regression model while retaining some of its distinctive features. Generalized Linear Model for y_i has following three parts:

a. distributional assumption

b. systematic component

c. link function

4.1.1 Distributional assumption

Generalized Linear Model (GLM) assumes that the response variable has a probability distribution belonging to the exponential family which includes normal, Bernoulli, binomial and Poisson distributions. The distributional assumption specifies the random component of the model and the probabilistic mechanism by which the responses are assumed to be generated. The variance of the response variable is expressed in terms of the product of a single scale or dispersion parameter ϕ and a variance function, denoted by $\nu(\mu_i)$, where $\phi > 0$. The variance function $\nu(\mu_i)$, describes how the variance of the response is functionally related to the mean of the response.

$$Var(Y_i) = \phi \nu(\mu_i) \qquad (1)$$

4.1.2 Systematic component

Systematic component of Generalized Linear Model specifies the effects of the covariates on mean of Y_i which can be expressed as linear predictor, η_i

$$\eta_i = \beta 1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots\ldots\ldots\ldots + \beta_p X_{ip} = \sum_i \beta_i X_i \qquad (2)$$

where typically $X_{i1} = 1$ for all i and then β_I is the intercept. The linear predictor is simply a linear combination of the unknown regression coefficients, $(\beta_1, \beta_{2,} ....., \beta_p)$ and the covariates, X_i.

4.1.3 Link function

The link function applies a transformation to the mean and then links the covariates, via the linear predictor, to the transformed mean of the distribution of the responses,

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + ...+ \beta_p X_{ip} = \sum_{k=1}^{p} \beta_k X_{ik} = X_i' \beta \qquad (3)$$

The use of log link in the Poisson distribution allows the mean to be greater than zero which is the requirement for Poisson distribution.

4.2 Generalized Estimation Equation (GEE)

The GEE approach is based on the concept of estimating equations and provides a very general approach for analysing the correlated responses. Liang and Zegar (1986) extended the GLM to GEE by replacing the identity matrix with a correlation matrix. The covariance matrix is defined by

$$V(u_i) = \left[D(V(u_{it}))^{\frac{1}{2}} R(\alpha)_{(n_i \times n_i)} D(V(\mu_{it}))^{\frac{1}{2}} \right]_{n_i \times n_i} \qquad (4)$$

where $R(\alpha)$ denotes the within panel correlation matrix, whereas in GLM the within panel correlation is represented by identity matrix.

Some of the correlation structures are described as under;

4.2.1 Independent structure

Independent structure is defined as

$$R_{uv} = \begin{bmatrix} 1 & if\ u = v \\ 0 & otherwise \end{bmatrix} \qquad (5)$$

4.2.2 Exchangeable structure

Exchangeable structure assumes a common correlation among the observations within the panel. In this case α is scaler and working correlation matrix has following structure

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \cdots\cdots & \alpha \\ \alpha & 1 & \cdots\cdots & .\alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots\cdots & 1 \end{bmatrix} \qquad (6)$$

The GEE with an exchangeable correlation structure uses the estimated Pearson residuals from the fit of the model to estimate the common correlation parameter. The estimate of α using these residuals is

$$\alpha = \frac{1}{\phi} \sum_{i=1}^{n} \left\{ \frac{\sum_{u=1}^{n_i} \sum_{v=1}^{n_i} r_{iu}\ r_{iv} - \sum_{u=1}^{n_i} r_{iu}^{\ 2}}{n_i (n_i - 1)} \right\} \qquad (7)$$

4.2.3 Autoregressive correlation

Autoregressive structure assumes the time dependence for the association if the repeated observations with in the panels have a natural order. In this case α is vector and correlation is estimated by using the Pearson residuals from the fit of the model.

$$\alpha = \frac{1}{\phi}\left[\sum_{i=1}^{n}\left(\frac{\sum_{t=1}^{n_i-0} r_{i,t}\, r_{i,t+0}}{n_i}, \ldots\ldots\ldots, \frac{\sum_{t=1}^{n_i-k} r_{it}\; r_{i,t+k}}{n_i}\right)\right] \tag{8}$$

4.3 Durbin Watson Statistics

Durbin Watson Statistic is used to test for the presence of first order autocorrelation in the residuals of a regression equation. The test compares the residuals for time period t with the residuals from time period t-1 and develops a statistic that measures the significance of the correlation between successive comparisons. The formula for the statistic is:

$$d = \frac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n}\left(e\right)^2} \tag{9}$$

d represents Durbin Watson statistic,

e represents residual

t represents time

Durbin Watson Statistic is used to test the presence of both positive and negative correlation in the residuals. The regions and the values of the Durbin Watson Statistic for the acceptance or rejection of null hypothesis that no significant correlation is present in residuals are shown in Table 3.

5. Model Development

5.1 Variables used

Following variables were incorporated into the model to estimate the daily number of accidents by road class in Great Britain;

- log value of the daily traffic flow as variate (traffic flow is measured in terms of vehicle kilometres travelled)
- log value of the road length as variate (road length are measured in miles)
- Day of week (7 days of week)
- Road class (5 classes of road)
- Time (1 to 1406, 1st January 1999 is denoted by 1 and 31st December 2002 is denoted by 1406)

5.2 Models developed

Following three models were developed by using Generalized Linear Model and Generalized Estimation Equation with both Poisson and negative binomial distributions.

Model 1: Constant + Log (Daily Traffic flow) + log (Road length)

Model 2: Model 1 + Day of week + Road class + Time

Model 3: Model 2 + Day of week.Road class.Log (Daily Traffic flow)

5.3 Deviance residuals

The deviance residual is the increment to overall deviance by each observation. The deviance is used to compare the fitness of model. The value of 1 for the deviance per degree of freedom is considered to be ideal fit for the model. The standardized deviance residual for Poisson and negative binomial models is calculated as under:

Poisson:
$$d_i^{\,2} = 2\left\{ y_i \ln\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i) \right\} \quad (10)$$

Negative Binomial
$$d_i^{\,2} = 2\, y_i \ln\left(\frac{y_i}{\mu_i}\right) - \frac{2}{\alpha}(1+\alpha\, y_i)\ln\left(\frac{1+\alpha\, y_i}{1+\alpha\, \mu_i}\right) \quad (11)$$

Standardized deviance residual:
$$D_{ri} = sign(y_i - \mu_i)\sqrt{d_i^{\,2}} \quad (12)$$

where α is the scale parameter

y_i is the number of accidents occurred

μ_i is the number of accidents predicted

5.4 Akaike Information Criterion (C_A)

Akaike Information Criterion (C_A) is used to compare different models. The comparison can also be made between non nested models. Lower value of the C_A shows the better fit of the model.

$$C_A = \frac{-2L(M_k) + 2p}{n} \tag{13}$$

where $L(M_K)$ is the log likelihood for model k

p is the number of predictors

n is total number of observations

5.5 Bayesian Information Criterion (C_B)

Bayesian Information Criterion (C_B) is an alternate to Akaike Information Criterion. In this criterion the saturated model has zero criterion. Smaller value of the (C_B) shows the better fit of the data.

$$C_B = D(M_k) - (df)\ln(n) \tag{14}$$

where $D(M_k)$ is the deviance of model k

d_f is the residual degrees of freedom

6. GOODNESS OF FIT

6.1 Generalized Linear Model with Poisson regression

Initial effort to develop the Generalized Linear Model with Poisson regression was made by using the available variables in our data set. Three different models as shown in section 5.2 were developed. This first model was developed by using the logarithm values of daily Traffic flow and Road length. Model 1 is found to be relatively over-dispersed as the deviance per degree of freedom is found to be 58.68 which is significantly large. In model 2 more variables were added to

reduce the mean deviance per degree of freedom. The variables of Day of week and Road class were added which significantly reduced the mean deviance of the model to 4.51.

In model 3 interaction variables of Daily traffic flow, Road class and Day of week were used to identify their effect on the daily number of accidents by road class. This slightly improved the deviance per degree of freedom to 4.31. Model 3 was preferred due to the better values of the total deviance, log likelihood, Bayesian information criterion (C_B) and Akaike information criterion (C_A) in comparison to model 1 and 2. The comparative results of the three models developed with Poisson distribution are shown in Table 4. The results suggest that data is over-dispersed relative to Poisson process.

6.2 Generalized Linear Model with Negative Binomial regression

The Generalized Linear Model with Poisson distribution was found to be over-dispersed as deviance per degree was found to be higher than 1. In order to account the over-dispersion in the data, Generalized Linear Model with negative binomial regression was used. Model 1 which is developed with values of log of traffic flow and road length showed the mean deviance of 1.09 per degree of freedom. Although the values of Bayesian information criterion (CB) of model 2 were slightly better than all models but model 3 was selected because of better Akaike information criterion (CA), log likelihood values and the need to explore the effect of interaction variables on the number of accidents. The mean deviance per degree of freedom for model 3 was obtained as 1.06. The results of all the three models developed by using negative binomial regression are shown in Table 4.

6.3 Durbin Watson Statistics

As the data set contains cross-sectional time-series data so it is assumed that serial correlation exists in the data and due to the presence of serial correlation the t values of the GLM are affected. Durbin Watson Statistic was used to check whether the autocorrelation exists among the residuals of each panel. Each road class is considered to be a panel which consists of four years time-series data from 1st January 1999 to 31st December 2002. Dwstat command was used in STATA software to calculate the values of Durbin Watson Statistic after the glm command. The Durbin Watson statistic results of GLM for Poison and negative binomial were found to be same. Based on the obtained results of the test for all classes of road, the null hypothesis of no autocorrelation among the residuals was not accepted for all models for all classes of road. The results of the Durbin Watson Statistic for all models are shown in Table 5. After comparing these results with the values shown in Table 3 it is concluded that positive autocorrelation exists in the residuals.

6.4 Generalized Estimation Equation with Negative Binomial Regression

Due to the presence of autocorrelation in residuals which can affect the significance level of explanatory variables Generalized Estimation Equation (GEE) was used to estimate the coefficient and t values of model 3 by considering the existing data as combination of panel and time-series data. The correlation structure of autoregressive order 1 (AR1) for the residuals was considered within the panel. Comparison was carried out between the coefficients and t values estimated by Generalized Estimation Equation with AR

(1) error structure and by Generalized Linear Model (GLM) with negative binomial distribution as shown in Table 6.

The coefficients of Daily Traffic flow, Thursday, Rural A, Urban A, Rural minor, Urban minor, Time and few other interaction variables were found to be statistically significant in both GEE and GLM. The highlighted results in the Table 5 show that the t values for some of the variables changed significantly. It was found that out of total 48 variables in the model the variable of Wednesday and interaction variable of Tuesday.UrbanA.Log (Daily Traffic flow) which were found to be significant in GLM changed to be insignificant in GEE. However there were 6 more variables which were insignificant in GLM but turned to be significant in GEE. The change in the t values suggests that if the presence of the serial correlation in the data is neglected then it may lead to wrong inferences and will result in putting more emphasis on those variables which are actually less significant variables and may result in loss of resources. Further details of the coefficients, t values and their comparison are given Table 6.

7. ESTIMATION OF RISK PER VEHICLE KILOMETER OF TRAVEL

The numbers of accidents for each day of week by road class were estimated by using the selected model. The predicted accident values were used to estimate the risk of accidents per billion vehicle kilometre of travel. Because of space limitation the accidents and risk are estimated for average values rather than every day. It follows as under;

7.1 Predicting Accidents

The coefficients of Generalized Estimation Equation (GEE) with negative binomial for model 3 were used to estimate the average accidents for all days of week on all classes of road. In the first step the average values of the Daily traffic flow, Road length and Time were calculated from the existing data. The accidents modelled by GEE closely matched the results derived from the raw data set. The results suggest that Urban A roads will have more accidents than all other road classes. Friday will have more accidents while Sunday will have fewer accidents on all classes of road. On Motorways highest number of accidents will occur on Friday with 31 accidents across whole Great Britain whereas lowest number of accidents on Motorways will be observed on Saturday. On Urban A roads highest number of accidents will reach to 232 on Friday. The highest number of accidents on Sunday was estimated to be 145 which occurred on Urban A. The number of accidents estimated for all classes across whole Great Britain is shown in Table 7.

7.2 Risk of Accident per Billion Vehicle Kilometre of Travel

After predicting the accidents the risk was estimated by dividing with the respective traffic flow. Following results were obtained;

- Although Sunday was estimated with second lowest number of accidents on Motorway but the risk per billion vehicle kilometre of travel was found to be second highest for Sunday on Motorways. The highest risk on Motorways was on Friday which had 106 accidents per billion vehicle kilometres of travel.

- On Rural A roads Sunday had the lowest number of accidents but it had the highest risk of accidents per unit of travel. It was estimated that on Rural A roads Sunday had 295 accidents per billion vehicle kilometre of travel whereas the lowest risk of accidents per unit of travel on Rural A roads was found to be on Tuesday.

- Monday will have the highest risk of accidents on Urban A roads with 952 accidents per billion vehicle kilometres of travel despite Friday having highest number of accidents. Sunday was found to have lowest risk per unit of travel on Urban A roads.

- On Rural minor roads Sunday being the safest day in terms of number of accidents had highest risk with 374 accidents. The safest day in terms on risk on Rural minor roads is Tuesday with 312 accidents per billion vehicle kilometres of travel.

- Monday is estimated to be with higher risk on Urban minor roads with 904 accidents per billion vehicle kilometres of travel. Sunday was found to be safer in terms risk per travel and number of accidents on Urban minor roads.

7.3 Comparison of the risk per billion vehicle kilometre on Motorways with other road classes

Table 8 shows the comparison of the risk between Motorways and other road classes on different days of the week. It shows that:

- On Rural A roads the risk per unit of travel is at least two times higher than Motorways.

- On Urban A roads the risk per unit of travel is at least 8 times higher than Motorways. Tuesday, Wednesday, Thursday are having 10 times greater risk of accident per unit of travel.

- On Rural minor roads the risk is at least 3 times higher than Motorways on all days of week.

- Urban minor roads are at least having 7 times higher risk of having accidents than Motorways.

8. Comparison of the number of accidents occurred and Predicted, Standardized deviance residual and Cumulative distribution function of Standardized deviance residuals

Graphs of the model 3 for accident occurred and predicted, standardized deviance residuals and cumulative distribution graph of the standardized deviance residual are shown in Figure 1. It is observed that there is no significant difference in terms of accidents prediction between the Generalized Linear Model and Generalized Estimation Equations with Poisson and negative binomial distributions.

The standardized deviance residual graph of Generalized Estimation Equations with negative binomial distributions is shown in Figure 1 which clearly shows the pattern at the end of each category of road class where the standardized deviance residual reaches its highest negative value which indicates the higher variability in the number of accidents during that time period. This standardized deviance residual is comparatively higher for Urban A and Urban minor roads than Rural A and Rural minor categories.

The cumulative distribution of the standardized deviance residuals shows that about 80 percent of the observations standardized residuals lie between -2 to 2.

9. Conclusion

The purpose of this study was to compare the performance of Generalized Linear Model and Generalized Estimation Equation with Poisson and negative binomial regression. A further objective was to formulate a model from the national accident data set for estimating the number of accidents which can be used by planning and road safety organizations for improving the road safety. In this case it is found that serial correlation exists in the residuals due to the time-series effect of the observations which affects the significance levels of the variables. In order to draw inferences from such models for policy or improving road safety purposes suitable method should be applied which can account for the serial correlation. Otherwise the inferences drawn from such models will result in loss of resources. In this case no particular difference was observed between the coefficient obtained by GLM or GEE. However the change in the t value of various variables was observed.

From the modelling results it is also found that Urban minor roads are having the highest number of road accidents where as Motorways have comparatively few accidents. The increase in the Daily Traffic flow and Road length will cause more accidents. It was also found that Friday will have more accidents whereas Sunday will have fewer accidents.

It is also concluded that despite having lowest accidents on Sunday the risk per billion vehicle kilometre of travel is highest on Rural A and Rural minor roads. For Urban A and Urban minor roads the risk of accidents per unit of travel is

highest on Monday. The results also conclude that Motorway are having the lowest risk of accidents and Urban A roads are having at least 8 times higher risk per unit of travel than Motorway.

Acknowledgments

The author is highly thankful to his advisor Professor Benjamin Heydecker, University College London for providing valuable guidance and timely suggestions for this study.

References:

Abdel-Aty, M and Radwan, AE, (2000) 'Modelling Traffic Accident Occurrence and Involvement', *Accident Analysis and Prevention*, Volume 32, issue (5), pp 633-642.

Anis, G.. (1996) 'An Application of Generalized Linear Modelling to the Analysis of Traffic Accidents', *Traffic Engineering Control*, Volume 37, issue (12), pp 691-696.

Bossche, F.V., Wets, G.., Brijs, T. (2006), 'Predicting Road Crashes Using Calendar Data'. Paper presented at the 85th Meeting of Transportation Research Board, Washington, DC, January 2006.

Edwards, J.B. (1998) 'The Relationship between Road Accident Severity and Recorded Weather', *Journal of Safety Research*, Volume 29, issue (4), pp 249-262.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004) *Applied Longitudinal Analysis*, Wiley & Sons press, USA.

Guevara, FL., Washington SP., (2004) 'Forecasting Crashes at the Planning Level Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona', *Transportation Research Record*: Journal of the Transportation Research Board, No.1897, TRB, National Research Council, Washington, D.C., pp 491-499.

Hadi, MA., Aruldhas, J., Chow, LF., Wattleworth, JA., (1995) 'Estimating Safety Effects of Cross-section Design for Various Highway Types using Negative Binomial Regression' *Transportation Research Record*: Journal of the Transportation Research Board, No. 1500, TRB National Research Council, Washington, DC, pp.169-177.

Hardin, J and Hilbe, J, (2001) *Generalized Linear Models and Extensions*, STATA Press, USA.

Hardin, J.W., Hilbe J.M., (2003) *Generalized Estimation Equations*, Chapman & Hall, USA.

Lord,D., Persaud, BN., (2000) 'Accident Prediction Models with and without Trend: Application of the Generalized Estimation Equations Procedure', *Transportation Research Record:* Journal of the Transportation Research Board, No. 1717, TRB National Research Council, Washington, DC pp 102-108.

Liang. K.Y. and Zeger S.L. (1986) 'Longitudinal Data Analysis Using Generalized Linear Models', *Biometrica*, Volume 73, pp 13-26.

Mahar, SP., Summersgill, J., (1996) 'Comprehensive Methodology for the Fitting of Predictive Accident Models', *Accidents Analysis and Prevention*, Volume 28, issue (3), pp 281-296.

McCarthy, P.S. (2005) ‘Public Policy and Alcohol Related Crashes among Old Driver’, viewed March 2005,

www.econ.gatech.edu/papers/mccarthy_CADTS_paper_SOEwebst_091402.pdf/.

McCullagh, P., Nedler, J.A., (1983) *Generalized Linear Models*, Chapman and Hall, USA.

Miaou, SP., Lum, H, (1993) ‘Modelling Vehicle Accidents and Highway Geometric Design Relationships’, *Accidents Analysis and Prevention*, Volume 25, issue (6), pp 689- 709.

Miaou, SP., (1994) ‘Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson and Negative Binomial Regressions’, *Accidents Analysis and Prevention*, Volume 26, issue (4), pp 471-482.

Noland, R.B., and M.A. Quddus., (2005) ‘Congestion and safety: A Spatial Analysis of London’, *Transportation Research Part A*, Volume 39, pp 737-754.

STATA Press, (2005) *STATA Longitudinal/ Panel dat*a, Reference manual, Release 9, USA.

Ulfarsson, GF., Shankar, VN., (2003), ‘An Accident Count Model Based on Multi-year Cross- sectional Roadway Data with Serial Correlation’ ,*Transportation Research Record*, Volume 1840, pp 193-197.

Wang, X., Abdel_Atey, M., (2006) ‘Crash Estimation at Signalized Intersections along Corridors: Analyzing Spatial Effect and Identifying Significant Factors’. Paper presented at the 85th Meeting of Transportation Research Board, Washington, DC. January 2006.

Table 1: Criteria for rearranging the Road Classification

S.No	Roads reclassified		
	New classification	Criteria	
		STATS 19 data classification	Speed
1	Motorway	Motorway	-
2	Rural A	A (M) or A	>40
3	Urban A	A (M) or A	<= 40
4	Rural Minor	B or C or Unclassified	> 40
5	Urban Minor	B or C or Unclassified	<= 40

Table 2: Explanation of Number of Observation in Data Set

S.No	Year	Total days in year	Road class	Total Number of observations
1	1998	365	5	1825
2	1999	365	5	1825
3	2000	366	5	1830
4	2001	365	5	1825
5	2002	365	5	1825
Total				**7305**

Each observation represents number of accidents by road class across whole Great Britain

Table 3: Regions and the values of Durbin Watson Statistic for Acceptance and Rejection of the Null Hypothesis

Zero to d_1	d_1 to d_u	d_u to $(4-d_u)$	$(4-d_u)$ to $(4-d_1)$	$(4-d_1)$ to 4
Reject Null H_o POSTIVE Autocorrelation	Neither Accept or reject	Accept the Null Hypothesis	Neither accept or reject	Reject Null $H_{o:}$ NEGATIVE Autocorrelation

Significance Points of d_1 and d_u at 5%

	K=1		**K=2**		**K=3**		**K=4**		**K=5**	
n	d_1	d_u	d_1	d_u	d_1	d_u	d_1	d_u	d_1	d_u
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
100+	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

K represents the number of independent variables in the equation

n represents the number of observations in the data set

Table 4: Comparison of the results of Generalized Linear model with Poisson and Negative binomial Distributions

	Generalized linear model with Poisson distribution			Generalized linear model with Negative binomial distribution		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
No. of obs.	7305	7305	7305	7305	7305	7305
Residual	7302	7291	7257	7302	7291	7257
Scale	1	1	1	0.549	0.032	0.029
(1/df)	58.68	4.51	4.31	1.09	1.055	1.060
Total	428533.	32905.7	31319.4	7932.3	7693.6	7694.5
Log	-237333	-39519.4	-38726.2	-41625.4	-32115.7	-31894.6
C_B	363572	-31957.3	-33241.1	-57028.6	-57169.4	-56866
C_B	64.97	10.82	10.61	11.39	8.79	8.74

C_B represents Bayesian information criterion

C_A represents Akaike information criterion

Table 5: Results of estimated values of the Durbin Watson Statistic

	Motorway	**Rural A**	**Urban A**	**Rural Minor**	**Urban Minor**
Model 1	1.52	1.36	1.16	1.35	1.24
Model 2 & 3	1.48	1.29	1.05	1.39	0.98

Table 6: Comparison of the coefficients and t values of Model 3 with GEE and GLM by using Negative binomial distribution

Variables	GEE Results		GLM Results	
	Coefficient	t-value	Coefficient	t-value
Constant	-3.11283	3.33	-4.632	-1.78
Log (flow)	0.219	3.26	0.194	2.96
Log (Road length)	0.091	0.72	0.013	0.13
Monday	1.031	0.57	1.04	0.54
Tuesday	-0.8598	-0.42	2.062	0.93
Wednesday	3.133	1.56	6.234	2.83
Thursday	5.065	2.87	7.238	3.27
Saturday	-2.369	-1.48	0.535	0.03
Sunday	-7.032	-3.95	-3.510	-1.79
Rural A	10.448	3.26	9.832	4.16
Urban A	8.523	2.67	9.177	3.82
Rural Minor	8.828	2.80	8.253	3.66
Urban Minor	8.944	2.86	8.355	3.63
Time	-0.0000293	-3.32	-0.0000253	-3.94

Comparison of the coefficients and t values of interaction variable of Day of week. Road class and Log (flow) of Model 3

	Motorway		Rural A		Urban A		Rural Minor		Urban Minor	
	GEE	GLM	GEE	GLM	GEE	GLM	GEE	GLM	GEE	GLM
Monday	0.016	0.015	-0.46	-0.29	-0.32	-0.22	-0.42	-0.24	-0.33	-0.16
	0.09	1.12	-4.35	-2.76	-3.16	-2.06	-5.56	-2.87	-3.43	-1.57
Tuesday	0.109	0.09	-0.37	-0.34	-0.22	-0.27	-0.32	-0.29	-0.24	-0.21
	0.60	0.63	-2.94	-2.79	-1.76	-2.08	-3.42	-2.93	-1.95	-1.68
Wednesday	-0.09	-0.11	-0.57	-0.55	-0.43	-0.5	-0.53	-0.51	-0.44	-0.42
	-0.05	-0.81	-4.47	-4.53	-3.32	-3.79	-5.52	-5.17	-3.51	-3.39
Thursday	-0.19	-0.17	-0.66	-0.60	-0.53	-0.53	-0.63	-0.56	-0.54	-0.47
	-1.07	-1.15	-5.27	-4.92	-4.10	-4.17	-6.59	-5.63	-4.29	-3.37
Friday	0.079	0.214	-0.40	-0.23	-0.26	-0.15	-0.36	-0.17	-0.28	-0.099
	0.44	1.46	-3.31	-1.85	-2.12	-1.19	-3.83	-1.69	-2.28	-0.78
Saturday	0.188	0.199	-0.29	-0.23	-0.15	-0.17	-0.24	-0.19	-0.17	-0.111
	1.09	1.44	-2.97	-2.37	-1.47	-1.56	-4.28	-2.54	-1.74	-1.08
Sunday	0.432	0.385	-0.06	-0.06	0.08	0.008	0	0	0.059	0.059
	2.64	3.31	-0.69	-0.85	1.07	0.11	0	0	0.77	0.80

Italic and bold values *represent the t values,* ***Shaded colour*** *represents the statistically significant variables* ***Red colour values*** *shows the variables either change from significant to insignificant or insignificant to significant*

Table 7: Estimation of risk per billion vehicle kilometres of travel and number of accidents predicted by GEE model

	Road classification				
	Motorway	**A Roads**		**Minor Roads**	
		Rural A	**Urban A**	**Rural**	**Minor**
Monday	**96.06** (24.73)	**274.16** (94.55)	**951.20** (199.29)	**344.13** (53.21)	**903.47** (265.28)
Tuesday	**86.16** (22.29)	**240.65** (90.16)	**892.65** (203.87)	**312.18** (52.49)	**841.03** (268.98)
Wednesday	**86.90** (23.02)	**241.33** (91.96)	**882.90** (207.07)	**319.09** (54.55)	**817.73** (268.15)
Thursday	**86.32** (23.47)	**241.98** (93.78)	**880.69** (209.51)	**316.34** (55.01)	**838.12** (279.23)
Friday	**105.60** (30.26)	**266.47** (106.73)	**932.62** (231.49)	**344.97** (62.01)	**879.69** (305.69)
Saturday	**100.04** (20.32)	**258.84** (90.34)	**907.15** (191.16)	**334.48** (52.40)	**855.30** (251.67)
Sunday	**102.2** (20.76)	**294.54** (84.69)	**805.00** (145.20)	**373.5**1 (48.20)	**740.63** (187.30)

Bold values shows the risk per billion vehicle kilometre of travel
Value in ( ) shows the predicted value of number of accidents for that class

TABLE 8: Comparison of the risk per billion vehicle kilometre on Motorways with other road classes

Day of week	Comparison of risk of Motorways with other road classes on different days of week			
	Motorway & Rural A	**Motorway & Urban A**	**Motorway & Rural Minor**	**Motorway & Urban Minor**
Monday	2.85	9.90	3.58	9.40
Tuesday	2.79	10.3	3.62	9.76
Wednesday	2.78	10.16	3.67	9.41
Thursday	2.80	10.20	3.66	9.71
Friday	2.52	8.83	3.26	8.33
Saturday	2.58	9.06	3.34	8.54
Sunday	2.88	7.87	3.65	7.24

Figure 1: Comparison of the results of Accident occurred and predicted and Standardized deviance residuals and cumulative distribution function