DE GRUYTER OPEN

Slovak Journal of Civil Engineering Vol. 25, 2017, No. 4, 12 – 18

DOI: 10.1515/sjce-2017-0018

DEVELOPING A MODEL FOR FORECASTING ROAD TRAFFIC ACCIDENT (RTA) FATALITIES IN YEMEN

Dr. Fareed M. A. KARIM[1*], Dr. Ali ABDO SALEH[1], Dr. Aref TAIJOOBUX[2], Dr. Marko ŠEVROVIĆ[3]

Abstract

The aim of this paper is to develop a model for forecasting RTA fatalities in Yemen. The yearly fatalities was modeled as the dependent variable, while the number of independent variables included the population, number of vehicles, GNP, GDP and Real GDP per capita. It was determined that all these variables are highly correlated with the correlation coefficient ($r \approx 0.9$); in order to avoid multicollinearity in the model, a single variable with the highest r value was selected (real GDP per capita). A simple regression model was developed; the model was very good (R2=0.916); however, the residuals were serially correlated. The Prais-Winsten procedure was used to overcome this violation of the regression assumption. The data for a 20-year period from 1991-2010 were analyzed to build the model; the model was validated by using data for the years 2011-2013; the historical fit for the period 1991 - 2011 was very good. Also, the validation for 2011-2013 proved accurate.

Address

[1] Faculty of Engineering, University of Aden, Yemen
[2] General Surgeon, Al-Gumohria Hospital, Aden, Yemen
[3] Head of Transport Planning Department, Faculty of Traffic and Transport Sciences, Zagreb University, Croatia

* ***Corresponding author:*** far_krm@yahoo.com

Key words

- RTA Fatalities,
- Forecasting RTA Fatalities,
- RTA Fatalities in Yemen,
- Regression Analysis,
- Prais-Winsten procedure.

1 INTRODUCTION

Since the unification of the northern and southern sectors of the country in 1990, Yemen has undergone rapid growth in road construction and motorization. During the 20-year period from 1991-2010, the population increased from 12.57 to 23.6 million, and registered vehicles increased from 0.334 to 1.36 million (yearly statistical books: 1991-2013); this means that vehicle ownership increased from 26.59 vehicles per 1000 persons to 57.6 vehicles per 1000 persons.

During the same period the number of fatalities increased from 1274 to 2959 (yearly statistical books: 1991-2013) as shown in Figure (1). This figure clearly demonstrates the seriousness of road traffic problems in Yemen, which continue to worse primarily due to the ever-increasing growth in motor vehicles.

Figure (2) shows the changes in vehicle ownership and deaths per 10,000 vehicles during the last 20 years.

The number of deaths per 10,000 vehicles has shown a downward trend from 38.12 to 21.77; this reduction in fatalities per vehicle does not necessarily indicate improvement in safety conditions on the road. As the per-capita availability of motor vehicles increases, fatalities per vehicle always decrease (Smeed, 1949), (Kopits and Cropper, 2005), (Mohan et al, 2009); therefore, a decrease in this ratio is not necessarily an indicator of road safety conditions.

1.1 Earlier Studies

One of the pioneering works in this regard was done by (Smeed, 1949); (Jacobs and Hutchinson, 1973) modified Smeed's model for the developing countries. However, the Smeed model did not provide a good predictive model for the data related to Yemen.

(Ameen and Nagi, 2001) undertook to develop Road Traffic Accident (RTA) fatality models in Yemen; however, these models included many independent variables which are difficult to forecast accurately within the context of life in Yemen such as Qat (a locally grown stimulant), the consumption of Qat, which is cultivated randomly by

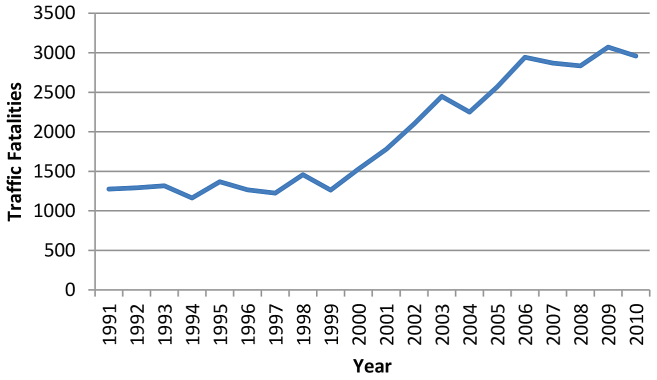

Fig. 1 *Traffic Fatalities from 1991 to 2010.*

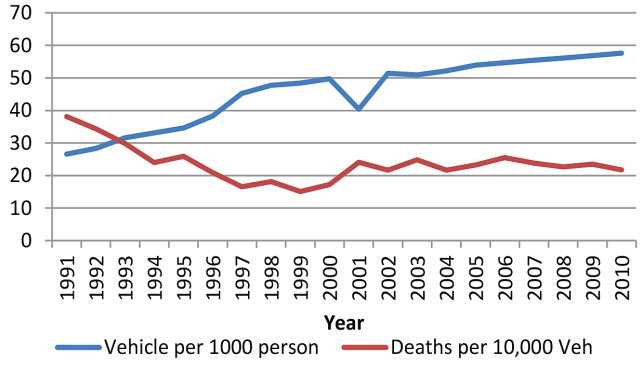

Fig. 2 *Vehicle ownership and fatality rate during 20 years.*

local people without any control from the government as measured by the area (acres grown per year), the number of hospital beds in Yemen, the annual maintenance cost in US$ per kilometer of rural roads, etc. Therefore, this study aims to develop a simple model with variables that can be easily acquired and forecasted in the context of life in Yemen.

2 MODEL

2.1 Data for calibrating the model

Limited data available in Yemen regarding RTA accidents, such as yearly fatalities, injuries, accidents, number of vehicles and the population as well as the Gross National Product (GNP) per capita is published by the Central Statistical Organization in Yemen (Yearly statistical books: 1991 – 2013). Other data such as GDP and real GDP (measured Purchasing Power Parity (PPP) $) are obtained from the World Bank website (World Bank, 2016).

Data from 1991 to 2010 was used for calibration of the model, while data from 2011 to 2013 was used for the validation of the model.

The yearly fatalities were modeled as the dependent variable, while the number of independent variables included the population, number of vehicles, GNP, GDP and Real GDP per capita (measured in $ Purchasing Power Parity (PPP)).

It was determined that all these variables are highly correlated with the correlation coefficient ($r \approx 0.9$) and tolerance values approaching zero as shown in Table (1); in order to avoid multicollinearity in the model, a single explanatory variable with the highest r value was selected (i.e., Real GDP per capita).

The real gross domestic product per capita (Real GDP) is used as a proxy for income (i.e., the average value of production per person). Previous researchers have noted that disposable income can have a positive or negative effect on safety (Fuchs, 1974). Real GDP may affect both exposure and the risk of a fatal crash. Gross Domestic Product per capita was included because past research (Kopits and Cropper, 2005 and 2008) has shown that it is strongly related to fatality counts.

2.2 Calibration of the Model

The Ordinary Least-Squares (OLS) regression model, which describes yearly fatalities as a function of Real GDP, has been developed as shown in Figure (3). The Fatalities and Real GDP ($) data for each year are available in the Appendix.

Tab. 1 *Correlation Matrix and Multicollinearity Statistics.*

Correlation matrix:						
Variables	Fatalities	Population	Vehicles	GNP ($)	GDP ($)	Real GDP($)
Fatalities	**1.000**	0.929	0.909	0.885	0.910	0.957
Population	0.929	**1.000**	0.989	0.816	0.875	0.993
Vehicles	0.909	0.989	**1.000**	0.789	0.855	0.980
GNP ($)	0.885	0.816	0.789	**1.000**	0.971	0.862
GDP ($)	0.910	0.875	0.855	0.971	**1.000**	0.914
Real GDP($)	0.957	0.993	0.980	0.862	0.914	**1.000**
Multicolinearity statistics:						
Statistic	Fatalities	Population	Vehicles	GNP ($)	GDP ($)	Real GDP($)
Tolerance	0.039	0.003	0.020	0.044	0.024	0.002
VIF	25.538	311.958	50.393	22.587	41.826	486.287

Tab. 2 *Output of a simple regression model.*

Goodness of fit statistics:					
R^2	Adj. R^2	MSE	RMSE	MAPE	DW
0.916	0.912	45515.194	213.343	10.514	0.827
Analysis of variance:					
Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	1	8972903.309	8972903.309	197.141	< 0.0001
Error	18	819273.491	45515.194		
Corrected Total	19	9792176.800			
Model parameters:					
Source	Value	Standard error	t	Pr > |t|	
Intercept	-1451.083	246.772	-5.880	< 0.0001	
Real GDP($)	1.054	0.075	14.041	< 0.0001	
Equation of the model: Fatalities = -1451.083+1.054*Real GDP($)					

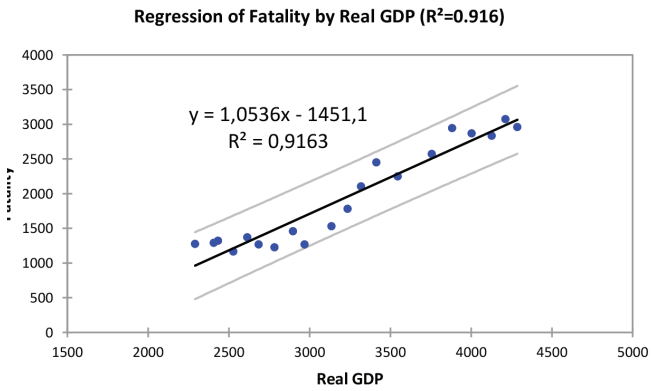

Fig. 3 *Simple Regression Model for Forecasting Yearly Fatalities.*

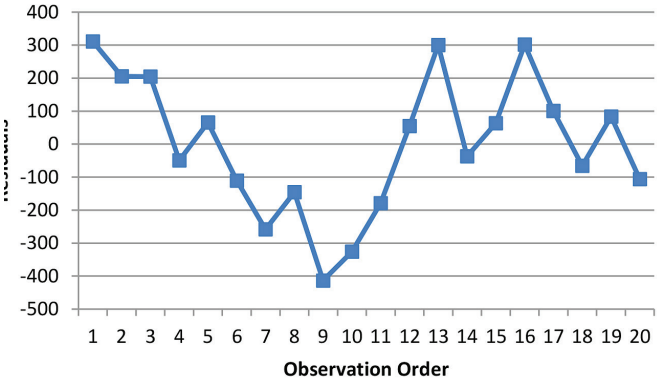

Fig. 4 *Residuals versus Data Order for the OLS Models.*

Table (2) shows the output of the simple regression model using Excel Software 2007 (Microsoft Office, 2007).

This output provides us with a great deal of information about this model. First, the adjusted R^2 is high at 0.912, meaning that nearly 91% of the variations in yearly fatalities are explained by real GDP. In other words, this model is very useful in predicting yearly RTA fatalities. The Mean Absolute Percentage Error (MAPE) is 10.51%, which is good. The ANOVA table shows the F-statistic is relatively large and significant. The standard error estimate of the coefficient is less than half the size of the coefficient, and the t-values are highly significant.

However, the Durbin-Watson statistic turns out to be 0.827; the DW statistic tests the hypothesis that the residuals (ε_t) from an ordinary least squares (OLS) estimation are not autocorrelated (Montegomery and Peck et al, 2001). Since its value is less than 2, a test for positive autocorrelation should be conducted. From the Durbin-Watson statistic table (Montegomery and Peck et al, 2001), when k=1(-number of independent variables) and observations (N) =20, for a 5% error level, d_l=1.201 and d_u=1.411. Since the Durbin-Watson statistic of 0.827 from the above results, is lower than 1.411, we reject the null hypothesis of no autocorrelation (with a 5% error level) and accept the alternative hypothesis of serial autocorrelation. Figure (4) shows the residual order of the model; it is clear from the figure that the residuals are autocorrelated. The Appendix shows the values of the residuals from the OLS Model.

If autocorrelation is detected, the estimated variances of the Ordinary Least Square (OLS) estimators are biased; they tend to underestimate the true variances and standard errors, and thus inflate the t values, thus potentially leading to the erroneous conclusion that the coefficients and other estimators are statistically different from 0; as a result, the usual F and t tests are not reliable. The formula used to compute the error variance (σ^2) is a biased estimator; it usually underestimates the actual variance in the error. Thus, the estimated R^2 will not be a reliable estimate of the true R^2 (Gujrati, 2004). Therefore corrective action is required.

Various transformations are carried out for the variables, including Box-Cox transformations (Box and Cox, 1964) to improve the model and the Durbin-Watson statistic; however, the results did not improve.

2.3 Modelling with Autocorrelated Residuals

Since we have autocorrelation in the residuals, a correction procedure is necessary if the model is to be used for forecasting. There are several methods available to take care of serial correlation in a linear model. One of them is the Prais–Winsten estimation (Website,

2016); it is a modification of the Cochrane–Orcutt estimation (Website, 2016) in the sense that it does not lose the first observation and leads to more efficiency as a result. In this study the Prais-Winston procedure is used to remove the serial correlation.

2.4 Prais–Winsten Procedure (website 2016)

Consider the model $y_t = \alpha + X_t \beta + \varepsilon_t$

where y_t is the time series of interest at time *t*; β is a vector of the coefficients; X_t is a matrix of the explanatory variables; and ε_t is the error term. The error term can be serially correlated over time: $\varepsilon_t = \rho\, \varepsilon_{t-1} + e_t$, $|\rho| < 1$; and e_t is white noise. In addition to the transformation of the Cochrane–Orcutt procedure, which is

$$y_t - \rho y_{t-1} = \alpha\,(1-\rho) + \beta(x_t - \rho X_{t-1}) + e_t.$$

for t = 2,3,...,T, the Prais-Winsten procedure makes a reasonable transformation for t=1 in the following form:

Tab. 3 *Iterations of Calculating the Rho (ρ) value.*

Iteration History

	Rho (AR1)		Durbin-Watson	Mean Squared Errors
	Value	Std. Error		
0	.520	.207	1.842	33075.029
1	.555	.202	1.916	32857.086
2	.560	.201	1.926	32834.402
3	.561	.201	1.927	32831.155
4	.561	.201	1.928	32830.665
5	.561	.201	1.928	32830.591
6	.561	.201	1.928	32830.579
7[a]	.561	.201	1.928	32830.578

The Prais-Winsten estimation method is used.
a. The estimation terminated at this iteration, because all the parameter estimates changed by less than .001.

$$\sqrt{1-\rho^2}\,y_1 = \alpha\sqrt{1-\rho^2} + (\sqrt{1-\rho^2}\,X_1\beta + \sqrt{1-\rho^2}\,\varepsilon_1$$

Then the usual least squares estimation is performed.

2.5 Estimation procedure

To perform the estimation in an efficient way, it is necessary to look at the auto-covariance function of the error term considered in the model above:

$$Cov\,(\varepsilon_t, \varepsilon_{t+h}) = \frac{\rho^h}{1-\rho^2} \quad , for\ h = 0 \mp 1 \mp 2, \ldots$$

Now it is easy to see that the variance–covariance matrix, Ω , of the model is

$$\Omega = \begin{bmatrix} \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} & \frac{\rho^2}{1-\rho^2} & \cdots & \frac{\rho^{T-1}}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} & \cdots & \frac{\rho^{T-2}}{1-\rho^2} \\ \frac{\rho^2}{1-\rho^2} & \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & \cdots & \frac{\rho^{T-2}}{1-\rho^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\rho^{T-1}}{1-\rho^2} & \frac{\rho^{T-2}}{1-\rho^2} & \frac{\rho^{T-3}}{1-\rho^2} & \cdots & \frac{1}{1-\rho^2} \end{bmatrix}$$

Now having ρ (or an estimate of it), we can see that

$$\Theta = (Z'\Omega^{-1}Z)^{-1}\,(Z'\,\Omega^{-1}\,Y),$$

where **Z** is a matrix of observations on the independent variable (X_t, *t* = 1, 2, ..., *T*), including a vector of ones; **Y** is a vector stacking the observations on the dependent variable (X_t, *t* = 1, 2, ..., *T*); and Θ includes the model parameters.

To forecast the future yearly fatalities, SPSS 20 software (SPSS 20.2016) is used with a Special SPSS Syntax command to generate the Prais-Winsten model. Table (3) shows the interactive process to calculate the Rho (ρ) value. Table (4) shows the output of the Prais-Winsten model.

Our forecasting model is:

$$y_t = \rho\, y_{t-1} + b_0\,(1-\rho) + b_1(\,X_t - \rho\, X_{t-1})$$

Tab. 4 *Output of the Prais-Winsten Model.*

Goodness of fit statistics:					
R^2	Adjusted R^2	Std. Error of Estimate		MAPE	DW
0.797	0.773	181.192		8.29%	1.928
Analysis of variance:					
Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	1	2189355.803	2189355.803	66.686	< 0.0001
Error	17	558119.821	32830.578		
Corrected Total	18	2747475.623			
Model parameters:					
Source	Value	Standard error	t	Pr > |t|	rho
Intercept	-1269.427	405.777	8.166	.000	0.561
Real GDP($)	1.001	.123	-3.128	.006	
Equation of the model:					
Fatalities* = -1269.427+1.001*Real GDP($)*					

where, y_t = yearly fatalities in year t
y_{t-1} = yearly fatalities in year t-1
ρ = 0.561 (autocorrelation coefficient)
b_0 = -1269.427 (Constant)
b_1 = coefficient of Real GDP (1.001)
X_t = Real GDP in year t
X_{t-1} = Real GDP in year t-1

Therefore, the final model is:

$$y_t = -557.28 + 0.561y_{t-1} + 1.001(X_t - 0.561 X_{t-1})$$

From the table above, we can see that the final model, which is corrected for autocorrelation, gives an adjusted R^2 value of 0.773 for the transformed variables. The DW statistic has increased from 0.822 to 1.928, which is above 1.411 (tabulated value) and indicates the autocorrelation in the residuals has been corrected. The mean absolute percentage error (MAPE) decreased from 10.5% to 8.29%, which is considered to be very good. The Appendix shows the residuals from the Prais-Winston procedure.

As per regression assumptions, the residuals should also be normally distributed and homoscedastic (homogeneity of error variance) (Gugrati, 2004). The Breusch-Bagan test is used to test the homoscedasticity of the residuals, while the Shapiro-Wilk test is used to test the normality of the residuals.

Breusch-Pagan Test for Homoscedasticity (Gugrati, 2004):

Test Interpretation:
H_0: The Residuals are homoscedastic
H_1: The Residuals are heteroscedastic

LM (Observed Value)	0.045
LM (Critical Value)	3.841
DF	1
p-value (Two-tailed)	0.831
alpha	0.05

LM =Lagrange multiplier

As the computed p-value is greater than the significance level of alpha = 0.05, one cannot reject the null hypothesis H_0.

Figure (5) shows the residuals from the Prais-Winsten Model; it is clear that the residuals are homoscedastic.

Shapiro-Wilk Test for Normality (Shapiro and Wick, 1965):

Test Interpretation:
H_0: The variable from which the sample was extracted follows a normal distribution.
H_1: The variable from which the sample was extracted does not follow a normal distribution.

Statistic	0.970
p-value (Two-tailed)	0.831
alpha	0.050

As the computed p-value is greater than the significance level of alpha=0.05, one cannot reject the null hypothesis H_0.

Figure (6) shows the normal P-P plots of the residuals from the Prais-Winsten Model; it is clear that the residuals are normal.

Figure (7) illustrates how well the model generates the historical data series by using the real GDP to predict the yearly RTA fatalities.

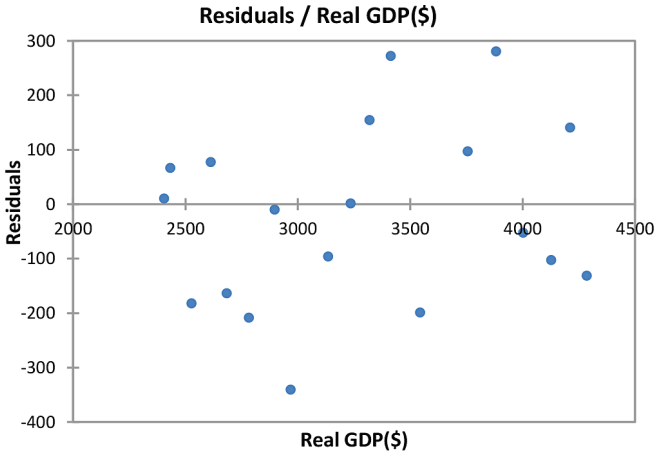

Fig. 5 *Residuals from the Prais-Winsten Model.*

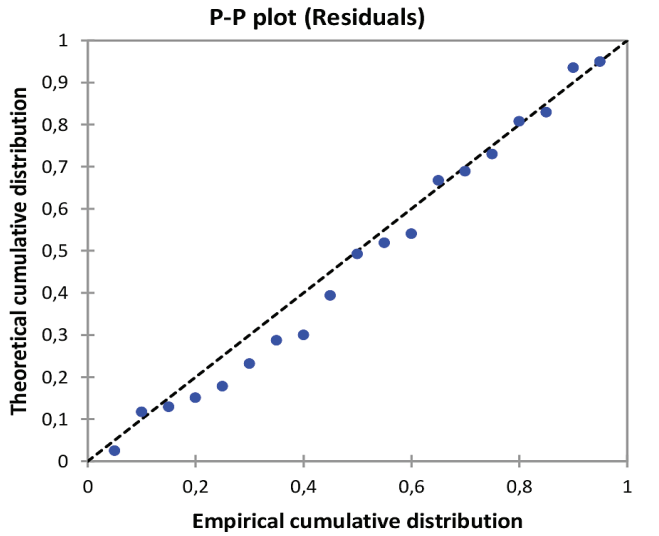

Fig. 6 *Normal P-P plots for Residuals.*

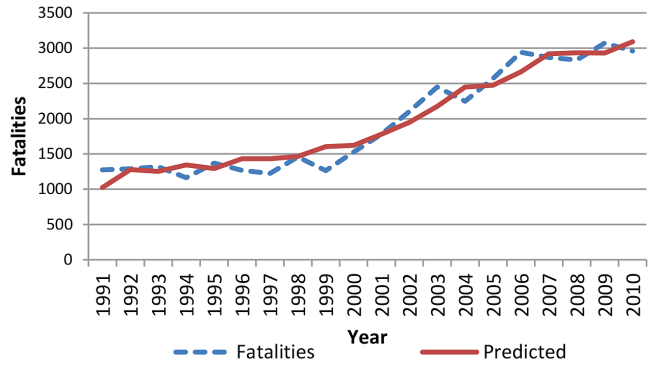

Fig. 7 *Actual Versus Forecasted Fatalities.*

This shows the predictive accuracy of the model in comparing the predicted values to the actual series. The lines show that the model does quite well.

3 VALIDATION OF THE MODEL

Once the models are ready, they are validated before using them as policy tools. This is done by predicting the yearly fatalities in known conditions. Since we know the yearly fatalities and Real GDP

per Capita for the years 2011 to 2013, and if the calibrated model can predict the yearly fatalities for the 3 years mentioned above with reasonable accuracy, then the models are assumed to correctly predict any future fatalities.

Table (5) below shows the yearly actual and predicted fatalities developed by the model for the years 2011-2013. As can be seen from the table, the validation has proved to be accurate (the yearly fatalities data beyond 2013 is unreliable due to the civil war in Yemen because of underreporting and hence discarded from the analysis).

Tab. 5 *Results from Validation of the Model.*

Year	Real GDP	Fatalities	Forecasted Fatalities	Difference
2011	3616.24	2152	2315	-208.44
2012	3673.87	2382	2388	-39.18
2013	3784.64	2494	2508	-43.93
Mean Absolute Percentage Error (MAPE)				2.81%

4 CONCLUSION

Unlike developed countries, the fatality rate in Yemen has an increasing trend. This research developed a statistical model that can be used in the prediction of the expected number of fatalities in Yemen with data that can be acquired and forecasted easily. This model developed a relationship between the yearly fatalities and the Real GDP per capita. The time series data of the fatalities for a 20-year period (1991-2010) is used to calibrate the regression model; the fit is very good (MAPE=8.29%). The model validated the use of 3 years of data (2011-2013) and was found to be accurate (MAPE= 2.81%).

This statistical modeling will serve as a guide to policy makers and the government in reviewing and formulating solid preventative measures, comprehensive legislation, and enforcement of road traffic safety laws. One of the major shortcomings of road traffic accidents in Yemen is the lack of recorded traffic data. There is an urgent need to improve the accuracy of police data-collecting procedures so that necessary information is available for scientific analysis.

APPENDIX

Year	Fatalities (Y)	Real GDP $ (X)	OLS Regression Residuals	Prais-Winston Residuals
1991	1274	2291.12	311.18	250.01
1992	1290	2406.69	205.42	10.14
1993	1317	2433.38	204.29	66.32
1994	1163	2528.06	-49.46	-182.62
1995	1369	2614.02	65.97	76.83
1996	1267	2685.10	-110.92	-163.59
1997	1223	2783.39	-258.47	-208.88
1998	1456	2897.51	-145.71	-10.28
1999	1264	2969.21	-413.25	-340.65
2000	1527	3135.57	-325.53	-96.28
2001	1779	3236.07	-179.41	1.03
2002	2101	3319.70	54.47	154.43
2003	2447	3414.53	300.56	271.90
2004	2248	3545.74	-36.68	-199.24
2005	2570	3756.53	63.23	97.00
2006	2942	3883.05	301.93	280.12
2007	2868	4004.22	100.27	-52.75
2008	2833	4128.41	-65.58	-102.57
2009	3071	4212.59	83.73	140.50
2010	2959	4286.40	-106.04	-131.59
Mean Absolute Percentage Error (MAPE)			10.51%	8.29%

REFERENCES

Box, G. E. P. & Cox, D. R. (1964) "*An Analysis of Transformations".* Journal of the Royal Statistical Society, Series B, 1964, 26, 211-252.

Gujrati D., (2004) *Basic Econometrics*, 4th Edition, The McGraw-Hill Companies.

Fuchs, V., (1974) *Some economic aspects of mortality in developed countries.* In: Perlman, M. (Ed). The Economics of Health and Medical Care. Macmillan, London, 174-193.

Jacobs G. D. and Hutchinson P. (1973) *A study of accident rates in developing countries.* TRRL Report LR546. Transport and Road Research Laboratory.

Jamal R. M. Ameen, and Jamil A. Naji, (2001) *Causal models for road accident fatalities in Yemen.* Accident Analysis and Prevention 33.

Kopits, E. and Cropper, M. (2005) *Traffic fatalities and economic growth.* Accident Analysis and Prevention, 37, 169-178.

Kopits, E., Cropper. M., (2008) *Why have traffic fatalities declined in industrialized countries for pedestrians and vehicle occupants.* Journal of Transport and Economics and Policy 42, 129-154.

Microsoft Office Excel (2007), www.microsoft.com. Cited on 25/2/2016.

Mohan, D. et el. (2009) *"Road Safety in India: Challenges and Opportunities",* University of Michigan Transportation Research Institute.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001) *Introduction to Linear Regression Analysis.* 3rd Ed, NY, NY: John Wiley & Sons.

Shapiro, S. S.; Wilk, M. B. (1965) *"Analysis of Variance test for normality (Complete samples)".* Biometrika 52 (3-4):591-611.

Smeed R. J., (1949) *Some statistical aspects of road safety research*. J. Roy. Stat. Soc. Ser. A 112, 1-23.

World Bank Site: http://data.worldbank.org/country/yemen-republic. Cited on 14/2/2016.

Yearly traffic accident statistics (1991-2013) published by the Central Statistical Organization, Sana'a, Republic of Yemen: Ministry of Planning and International Cooperation, www.cso-yemen.org. Cited in 2015.

http://en.m.wikipedia.org/wiki/Prais%E2%80%93Winsten_esttimation. Cited on 15/1//2016.

http://en.m.wikipedia.org/wiki/Cochrane%E2%80%93Orcutt_estimation. Cited on 3/1/2016.

SPSS 20 User Guide, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/amos/20.0/en/Manuals/IBM_SPSS_Amos_User_Guide.pdf . Cited on 14/2/2016. Cited on 5/2/2016.