East-West J. of Mathematics: Vol. 25, No 2 (2024) pp. 145-171
https://doi.org/10.36853/ewjm0406

A STATISTICAL ANALYSIS OF BANGKOK TRAFFIC ACCIDENT DATA

Kanokwan Channgam[*], **Suntaree Unhapipat**[*]
and
Nabendu Pal[**]

[*] *Dept. of Mathematics, Faculty of Science,*
Mahidol University, Bangkok, Thailand

[**] *Faculty of Mathematics and Statistics,*
Ton Duc Thang University, Ho Chi Minh City, Vietnam and
Department of Mathematics, University of Louisiana at Lafayette,
Lafayette, Louisiana, USA

Abstract

A nonlinear regression model has been used to explain the adjusted monthly number of road accidents per 100,000 people (referred as "Adjusted Road Accident Rate") occurred within Bangkok, Thailand, from 2010 to 2016 in term of seven other variables. The extensive analysis is quite interesting, as it indicated that the main and overwhelming factor that contributed to accident rate is the total number of vehicles in the city. In addition, three other variables, namely - mean temperature, the total amount of rainfall and the number of holidays, also contributed to the accident rate, albeit mildly. The implications of this study is quite profound in the sense that if the city planners and policy makers want to reduce the accident rate then total number of vehicles need to be reduced, perhaps through investing more resources into public transportation and/or increasing vahicular fees which may discourage more vehicles on the streets. Other major cities in Southeast Asia, such as Kolkata, Dhaka, Ho Chi Minh City, etc. having similar characteristics as Bangkok can draw lessons from this study.

Key words: Road safety, Road traffic accident, Nonlinear regression model.
Mathematics Subject Classification (MSC) 2020: 62G09, 62F40, 62E17.

1 Introduction

1.1 Preliminaries

The attacks on the twin towers of the World Trade Center, New York, on September 11, 2001, were the deadliest terrorist acts on a single day in modern history, and over 3,000 people perished on that day (see Tanaboriboon and Satiennam (2004)). Unfortunately, almost the same number of people get killed by road accidents worldwide each day, but it doesn't draw much public attention. This figure does not include the number of injured and disabled, which is over 50,000 per day. Since 2007, more than 1.2 million people have been killed due to road accidents per year worldwide. In addition, another 20 to 50 million people in 2010, and more than 50 million in 2013 have sustained nonfatal injuries or became disabled as a result of road accidents. Obviously, from a public health point of view, road accidents constitute a major concern for human life and safety.

Another unfortunate aspect of the road hazard is that it affects the low and middle income countries with rates nearly double than what it is for high-income countries. (The World Bank Atlas method was used to categorized the countries in terms of the annual gross national income (GNI) into bands, such as: low-income = US$ 1005 or less, middle-income = US$ 1,006 to US$ 12,275 and high-income = US$ 12,276 or more). Economically disadvantageous families are hardest hit by both direct costs (such as hospitalization, rehabilitation, etc.) and indirect costs (such as lost wages, economic, opportunities, etc.) that result from the road injuries. This is due to the fact that financially weaker segment of the society uses more risky modes of road transportations. The road accident injuries are estimated to cost the low and middle income countries somewhere between 1 to 2% of their gross national product (GNP), totaling over an estimated US$ 100 billion a year (source: World Health Organization (2013)).

1.2 Thailand Road Accident Situation

The yearly road accidents death rate in Thailand, which is about 36.2 per 100,000, is currently the second most lethal in the world after Libya, and the highest in Asia, according to the World Health Organization. It is more than 10 times of Singapore (3.6 per 100,000) which has a much higher standard of living. Thailand's road accident death rate is approximately 8 times of Japan's (4.7 per 100,000) which is an economic leader with the third highest GDP in the world since 2010 (Riley and Sherman (2017)). In addition, Thailand's road accident death rate is more than twice that of the global average rate (17.4 per 100,000). Thailand loses approximately 3.4% of her GNP as a result of road accidents (see Luathep and Tanaboriboon (2005) and World Health

Organization (2015)). The total annual economic losses due to road accidents in Thailand have been presented in the following Table1 (The World Bank (2017)).

Table 1: Yearly Thailand Road Accident Data and Corresponding Economic Losses

Year	Annual number of road accident	Annual adjusted road accident rate	Thailand GNP (billion US$)	Total estimated anual economic lose due to road accidents $\approx 3.4\%$ (billion US$)
2010	14879	257.0672	850.8773	28.9298
2011	22334	387.4761	891.9839	30.3275
2012	24085	417.5388	952.5483	32.3866
2013	24483	425.0555	975.5514	33.1688
2014	25972	450.4211	1014.8	34.5032
2015	28476	493.4465	1056.1	35.9074
2016	33410	577.4857	1106.8	37.6312

Bangkok is not only the capital city of Thailand, but also the most populous city in Thailand. Among the world's metros it also experiences one of the worst traffic jams, and witnesses one of the highest road accident rates (see Fernquest (2017)).

The following are just some of typical daily newspaper reporting about road accidents in and around Bangkok:

-On 14th April 2016: Four people were killed and six others injured when a speeding car crashed into a tree in Chatuchak district, Bangkok (reported by The Nation website).

-On early Monday 27th March 2017: Twoadults and a girl were killed and six others injured after their sedan overturned on a road in Bangkok's Nong Chok district (reported by The Nation website).

- On 29th April 2017: Pickup truck which carried laborers turned upside down because shaft was broken and many laborer were injured (reported by the Accident Alert Network website).

-On 22nd May 2017: Two people were killed and three others injured when a van collided with car cleaner at tollway (reported by the Accident Alert Network website).

1.4 Background of the current (Bangkok Road Accident) Study

The primary objective of this study had been to explain the road accident rate in Bangkok in terms of other variables to see if some variable, other than the natural ones, can be found as a major contributing factor. However, the road accident rate, as we will see in Section 2 and 3, varies by months for various reasons. Therefore, we decided to look at the monthly road accident rate. But since the months of a year have different number of days which may influence the monthly accident rates, so we decided to use the monthly adjusted road accident rate (henceforth referred as 'Adj-RAR') as a uniform measure to study the accident rate, which essentially looks at the number of accidents per 100,000 individuals over a period of 30 days. (Note that the difference between a non-Leap year February and March is 3 days, which is about 10%, and thus influence the accident rate accordingly if not adjusted properly.) Similarly, other suitable observed variables have been adjusted as shown in Secton 3.

The monthly adjusted road accident rate (Adj RAR) per 100,000 is defined as:

$$\text{RAR} = \frac{\text{Total number of accidents per month}}{\text{populaion size}} \times 100.000 \tag{1}$$

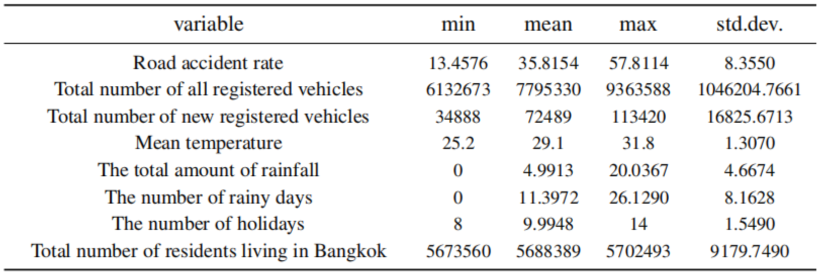
$$\text{Adj RAR} = \text{RAR} \times \frac{30}{\text{number of days in the month}} \tag{2}$$

The Thai Road Accident Data Center for Road Safety Culture, called Thai RSC (Thairsc (2015)) is one of the few organizations that collect data on road accidents in Thailand. It provides a whole array of useful information, such as - the number of accidents, the number of injuries, the number of disabilities, and the number of fatalities for each province per month, etc. For this study, Thai RSC provided the information about all road accidents occurred within the municipality of Bangkok City from 2010 to 2016. We use the monthly adjusted road accident rate (Adj-RAR) using the above equation (2) apart from the total number of accidents because we consider that the total city population size has an effect on the number of accidents. If population size increases (decreases) in a short span of time, then the number of accidents may increase (decrease) higher (lower) than the proportionate rate since the road capacities or logistics take longer to adjust with the varying population size. Hence, the accident rate (through Adj-RAR) may provide more useful insight than the actual total number of accidents into the dynamics of the road accident situation. Information on the total number of all registered vehicles and the total number of new registered vehicles have been obtained from the Planning Division, Transportation Statistics sub-division (apps.dlt.go.th/statistics_web/statistics.html).

Further, information related to weather, such as - mean temperature, the total amount of rainfall, the number of rainy days, etc. have been obtained from the meteorologicale department. Total number of residents living in Bangkok has been procured from the Official Statistics Registration System (see http://stat.dopa.go.th/stat/statnew/upstat age.php). This study has used the number of all register vehicles per month by interpolating the reported annual data.

The following Table 2 is a summary of the complete data used in this research work. The full dataset is presented in the appendix.

Table 2: Summary Characteristics of the Relevant Variables (per month) over 84 Months (January 2010 - December 2016)



variable	min	mean	max	std.dev.
Road accident rate	13.4576	35.8154	57.8114	8.3550
Total number of all registered vehicles	6132673	7795330	9363588	1046204.7661
Total number of new registered vehicles	34888	72489	113420	16825.6713
Mean temperature	25.2	29.1	31.8	1.3070
The total amount of rainfall	0	4.9913	20.0367	4.6674
The number of rainy days	0	11.3972	26.1290	8.1628
The number of holidays	8	9.9948	14	1.5490
Total number of residents living in Bangkok	5673560	5688389	5702493	9179.7490

1.5 A Summary of the current Study

In this study our primary objective has been to see if the monthly adjusted RAR can be explained by extraneous factors, ranging from weather related variables to man-made ones. First we wanted to see if the monthly Adj-RAR has remained same for all the months over the years during the study period. Next, we have investigated if a suitable multiple regression model could be used to explain Adj-RAR in terms of the other observed variables. Interestingly, a nonlinear regression model has been found to provide quite a good fit as it explains about 90% of the total variability of Adj-RAR. Among all the explanatory variables, the total number of vehicles (a man-made variable) is found to have an overwhelming influence in explaining the accident rate. Three other variables two of which are weather related also have some minor influence on the accident rate. The implications of this research is quite profound in the sense that if the city planners and policy makers want to reduce the accident rate then total number of vehicles need to be reduced too, perhaps through investing more resources into public transportation and/or increasing vehicular fees which may discourage more vehicles on the streets.

Section 2 provides data description and carries out a detailed analysis to see if the mean monthly Adj RAR is same across the months. In Section 3 we provide a step by step approach to build a suitable regression model to explain the variable of interest. Section 4 draws a conclusion to this study by observing overall trends and suggested recommendations. Excess tables and derivations have been relegated to the Appendix.

2 Statistical Data Analysis

2.1 Description of the Current Dataset

In this section first we describe the variables clearly which we have been able to observe from January 2011 to December 2016. Our primary goal is to study the monthly adjusted road accident rate (adj RAR), and how it can be explained by other useful variables. The variables of interest are defined as follows. Y = monthly adjusted road accident rate (adj RAR)

X_1 = adjusted total number of all vehicles in a month

X_2 = adjusted total number of new vehicles in a month

X_3 = monthly mean temperature (celcius)

X_4 = average total amount of rainfall (mm) in a month,

X_5 = adjusted number of rainy days in a month;

X_6 = adjusted number of holidays in a month;

X_7 = adjusted total number of residents living in Bangkok in a month.

For monthly total number of all vehicles and monthly total number of residents living in Bangkok, we interpolated them from the annual figures.

$$\mathrm{Y} = \text{adj RAR} = \text{RAR} \times \frac{30}{\text{number of days in the month}}$$

$$X_1 = \text{total number of all vehicles in a month} \times \frac{30}{\text{number of days in the month}}$$

$$X_2 = \text{total number of new vehicles in a month} \times \frac{30}{\text{number of days in the month}}$$

$$X_3 = \text{monthly mean temperature}$$

$$X_4 = \frac{\text{total amount of rainfall (mm) in a month}}{\text{number of days in the month}}$$

$$\mathrm{X}_5 = \text{number of rainy days in a month} \times \frac{30}{\text{number of days in the month}}$$

$$\mathrm{X}_6 = \text{number of holidays in a month} \times \frac{30}{\text{number of days in the month}}$$

$$\mathrm{X}_7 = \text{total number of residents living in Bangkok in a month} \\ \times \frac{30}{\text{number of days in the month}}$$

We begin our study by plotting monthly adjusted RAR against time (= month) in Figure 1. Since the Thai RSC started collecting data only in early 2010, and it didn't become fully functional in terms of data collection resources till the end of that year, the data for the year 2010 is not reliable. Therefore, the data for 2010, as shown in Figure 1 and Figure 2, are not very reliable.

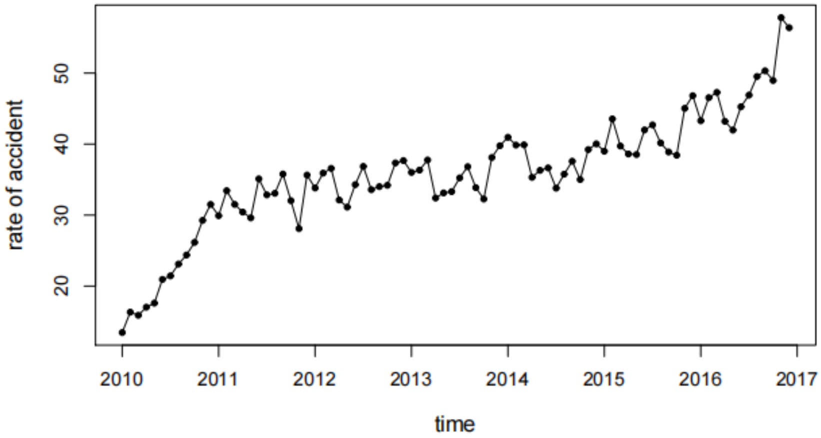


Figure 1: Adj Road Accident Rate in 2010-2016

Figure 1 plots Y against time (month of a year starting with January 2010 (2010 data have been included in this plot only for the sake of completeness)). Beginning with 2011 January, the plot shows, more or less, an upward trend over time.

Next, Figure 2 shows the monthly adjusted RAR (i.e., Y) for each year plotted against each month. Though the yearly lines show an upward movement as the year progressed, no discernible pattern seems to exist over months.

In the folowing section we study whether the mean monthly RAR over years has remained constant across the twelve months. This has been done using both the parametric as well as nonparametric ANOVA. In subsection 3.4 we build a regression model to explain Y interm of X_i is (i=1,2,...,7).

It needs to be pointed out that in November 2011 there was a historic flood in Thailand which affected the City of Bangkok adversely. There was a great loss of life and materials as many sections of the city got inundated by several feet of water throwing life completely out of balance. That's why the data point for November 2011 has been excluded from most of the analysis here. However, in Subsection 2.2, the statistical analysis has been done by including the November 2011 data point (in 2.2.1) as well as by excluding that data point (in 2.2.2) just to show how much this single outlier can impact the analysis.

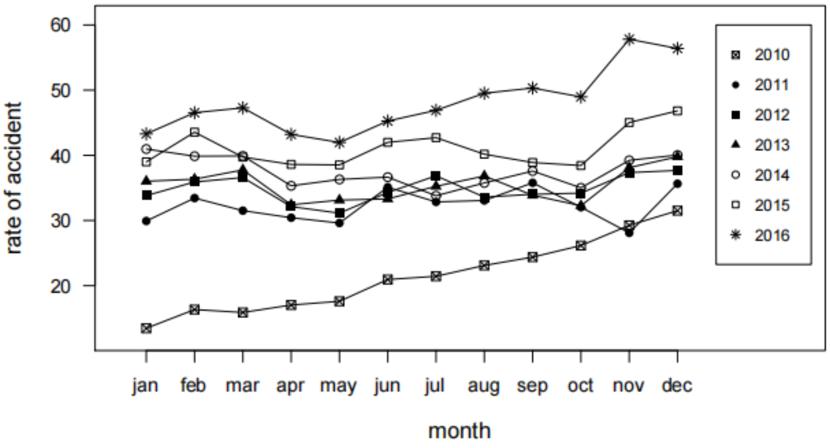


Figure 2: Annual Adj Road Accident Rate in 2010-2016

2.2 Checking Monthly adj. RAR Against Month

2.2.1 Using 72 observations (January 2011 - December 2016)

The standard parametric ANOVA have been carried out to see if the mean monthly adj RAR is same across the months. The resultant sum of squares decomposition is given in Table 3.

Table 3: Parametric ANOVA with 72 observations (including November 2011)

Source	**d.f.**	**Sum of Squares**	**Mean Square**	F**-statistic**	**Pr**$(> F)$
Month	11	302.267	27.479	0.736	0.700
Error	60	2240.255	37.338		
Total	71	2542.522			

The above ANOVA shows that there is no significant difference in mean adj RAR among the twelve months. However, this is based on the two key model assumptions which are now verified below. Figure 3 shows the plot of the residuals against the corresponding ordered observations, and it clearly shows an increasing trend instead of a total random scatter-plot.

The below Figure 4 shows the quantile-quantile (”Q-Q”) plot of the 72 residuals under the normality and homoscedaticity assumptions.

Next, provide the results of specific tests to test the two key model assump-

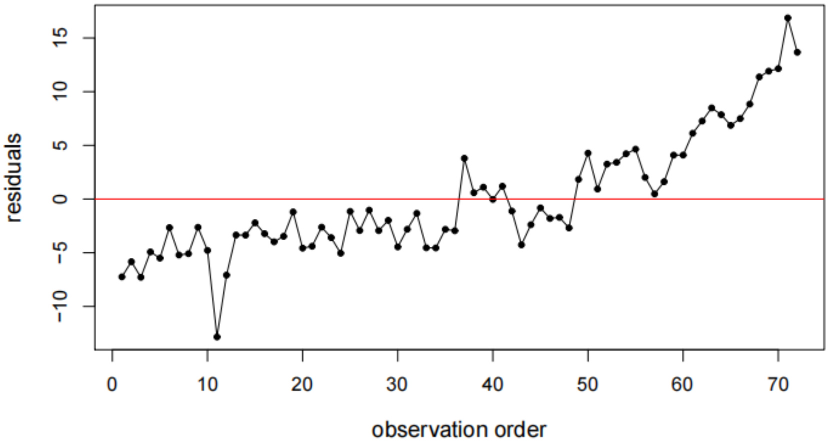


Figure 3: Residuals versus The Order Of The Data

tions.

1. Tests for the Normality Assumption: (1.1) Anderson-Darling normality test: $A = 2.0309$, p-value $= 3.206 \times 10^{-5}$.
(1.2) Shapiro-Wilk normality test: $W = 0.92927$, p-value $= 0.000569$. Thus, there appears to have a strong evidence against the normality assumption.

2. Tests for the Homoscedasticity Assumption:

(2.1) Levene's Test for Homogeneity of Variance (center = mean): F-statistic $= 17.553$, p-value $= 8.022 \times 10^{-5}$.

(2.2) Levene's Test for Homogeneity of Variance (center = median):F-statistic =10.61, p-value =0.001737.

Again, the above tests indicate that possibly the variances are not same over the months.

In the light of the above test results on the model assumptions, it is imperative that we carry out a nonparametric version of the ANOVA, i.e., Kruskal-Wallis test.

Kruskal-Wallis Rank Sum Test: The test statistic value =8.7557,df=11 and p-value = 0.6444 . Thus one fails to reject the null hypothesis which implies that the mean (over the years) monthly adj RAR is same across the months.

2.2.2 Using 71 observations (January 2011 - December 2016 exclude November 2011)

The standard parametric ANOVA have been carried out to see if the mean monthly RAR is same across the months. The resultant sum of squares decomposition is given in Table 4.

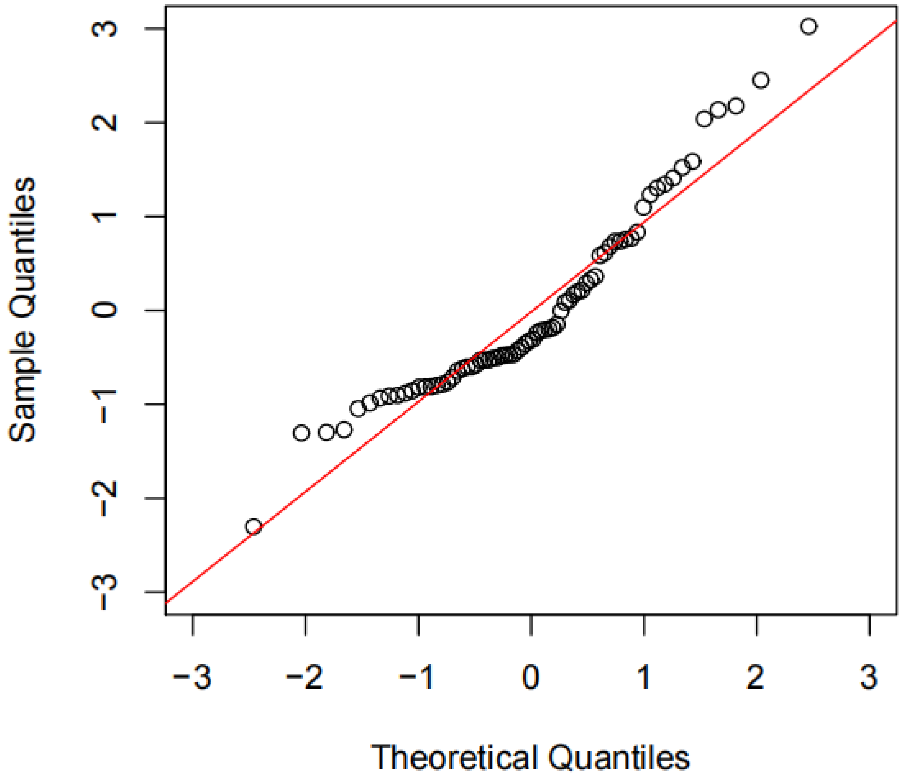


Figure 4: Normal Q-Q Plot

Table 4: parametric ANOVA with 71 observations (excluding November 2011)

Source	d.f.	Sum of Squares	Mean Square	F-statistic	Pr($> F$)
Month	11	396.3861	36.0351	1.0410	0.424
Error	59	2042.3050	34.6153		
Total	70	2438.6911			

The above ANOVA shows that there is no significant difference in mean RAR among the twelve months. However, this is based on the two key model assumptions as discussed in 2.2.1.

The Figure 5 shows the plot of the residuals against the corresponding ordered observations, which again shows an increasing trend instead of a random scatter-plot.

Figure 6 shows the quantile-quantile (”Q-Q”) plot of the 71 residuals under the normality and homoscedasticity assumptions.

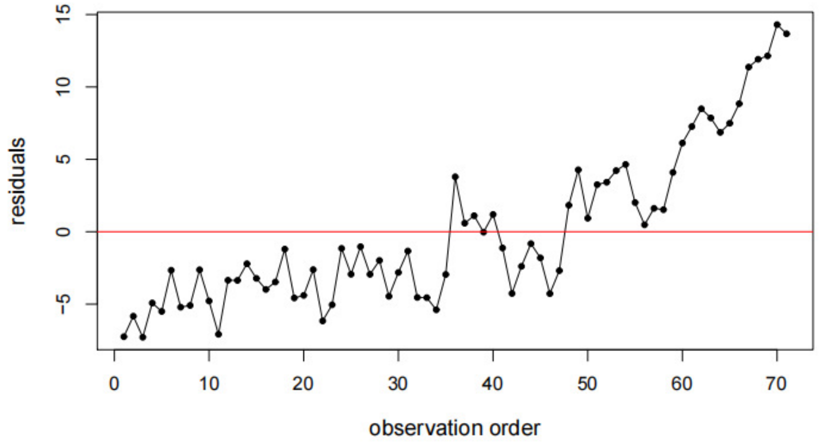


Figure 5: Residuals versus The Order Of The Data

In the following we provide the results of normality and homoscedasticity assumptions. 1. Tests for the Normality Assumption:

(1.1) Anderson-Darling normality test: A=2.2567, p-value $=8.852\times10^{-6}$.

(1.2) Shapiro-Wilk normality test:$W = 0.90497$,p-value $=5.457\times10^{-5}$. Thus, there appears to have a strong evidence against the normality assumption.

2. Tests for the Homoscedasticity Assumption:

(2.1) Levene’s Test for Homogeneity of Variance (center = mean): F-statistic =29.26, p-value = 8.587×10^{-7}.

(2.2) Levene’s Test for Homogeneity of Variance (center = median): F-statistic =17.034, p-value $=1.011\times10^{-4}$.

Again, the above tests indicate that possibly the variances are not same over the months.

In the light of the above test results on the model assumptions, it is imperative that we carry out a nonparametric version of the ANOVA, i.e., Kruskal-Wallis test.

Kruskal-Wallis Rank Sum Test: The test statistic value =11.171,df=11, and ..-value =0.4291. Thus one fails to reject the null hypothesis which implies that

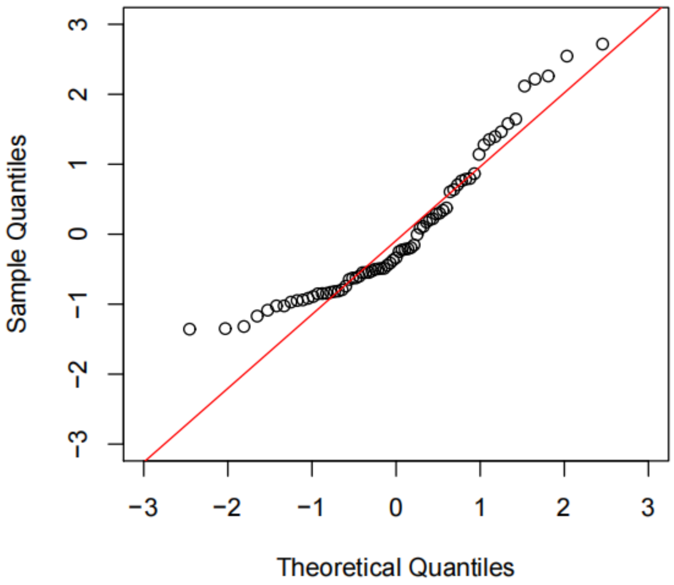


Figure 6: Normal Q-Q Plot

the mean (over the years) monthly adj RAR is same across the months.

Remark 2.1 The take away message of this subsection is that the mean (over years) monthly adj RAR doesn't seem to differ significantly from month to month as pointed out by both the parametric as well as the nonparametric ANOVA methods.

3 Model Estimation Results

3.1 The Correlation Matrix and Scatter Plot

In section 2.2 we simply studied the variable Y singularly, and observed its behavior over the months. In this section we study all the eight variables (Y and X_1 through X_7) simultaneously to extract information about interdependency among the variables.

We start our effort to see the hidden connections among the eight variables by computing the pairwise Pearson's correlation coefficients as presented in the sample correlation matrix below. (It is enough to focus on the upper triangular part since the matrix is symmetric.) The correlation matrix has been obtained from the 71 monthly observations after dropping the November 2011 observation. (A similar computation based on all the 72 observations, without dropping the November 2011 observations gave nearly identical matrix except that correlation between Y and X_2, Y and X_5, X_1 and X_2, and X_2 and X_6 become non-significant. However, the November 2011 observation was dropped since the city traffic was adversely affected by the historic flood.) Only nine correlations among the total 28 correlations have been found to be significantly different from zero (using the standard t-test), and these nine correlation coefficients have been bold-faced. For example, correlation coefficient between Y and X_1 is 0.7922 , the correlation coefficient between Y and X_2 is -0.2737, etc.

The first row of $\hat{\rho}$ contains the simple linear correlations of the dependent variable with each of the independent variables. The two variables, - adjusted total number of all registered vehicles (i.e., X_1) and adjusted total number of residents living in Bangkok (i.e., X_7) have reasonably high linear correlations with road accident rate. They would account for 62.76% ($r^2 = 0.7922^2$) and 14.06%, respectively.

The Figure 7 is the scatter plot of all pairs of eight variables (i.e., Y and X_1 through X_7). It reveals the linear relationship or association between two variables at a time.

3.2 Regression Model

We start with a multiple linear regression model by using all the independent variables ($X_1, X_2, ..., X_7$). The outcome of the full model is summarized in Table 5a followed by the ANOVA in Table 5b.

The ANOVA in Table 5b shows that the seven independent variables do have a significant combined contribution in explaining Y. But, the Table 5a indicates that most of these variables may not have individual significant contribution as seen from their p-values. Therefore, a step-wise regression model building approach is taken starting with the most contributing variable, that is variable X_1.

The equation of simple linear regression model is

$$\hat{Y} = -3.296 + 5.165 \times 10^{-6} X_1. \tag{3}$$

The above Figure 8 shows the plot of real value and predicted value (result from Table 6a) against time. In this Figure 8, the red line represent the value from equation (3). It is not a straight because the X axis in the plot is time, it is not the value of X_1.

The following Figure 9 of the residuals from the above linear regression (3) shows that the model is missing some important components, since there is a clear pattern of residuals in general decreasing and then increasing. The Q-Q plot of the residuals as well as the usual model assumption checks are provided in the Appendix (Figure 14).

As it can be seen in Figure 8, the simple linear regression of Y on X_1 is not a good one. It appears that a third degree polynomial of X_1 may provide a

Figure 7: The Scatter Plot between Each Pair Of All Variables

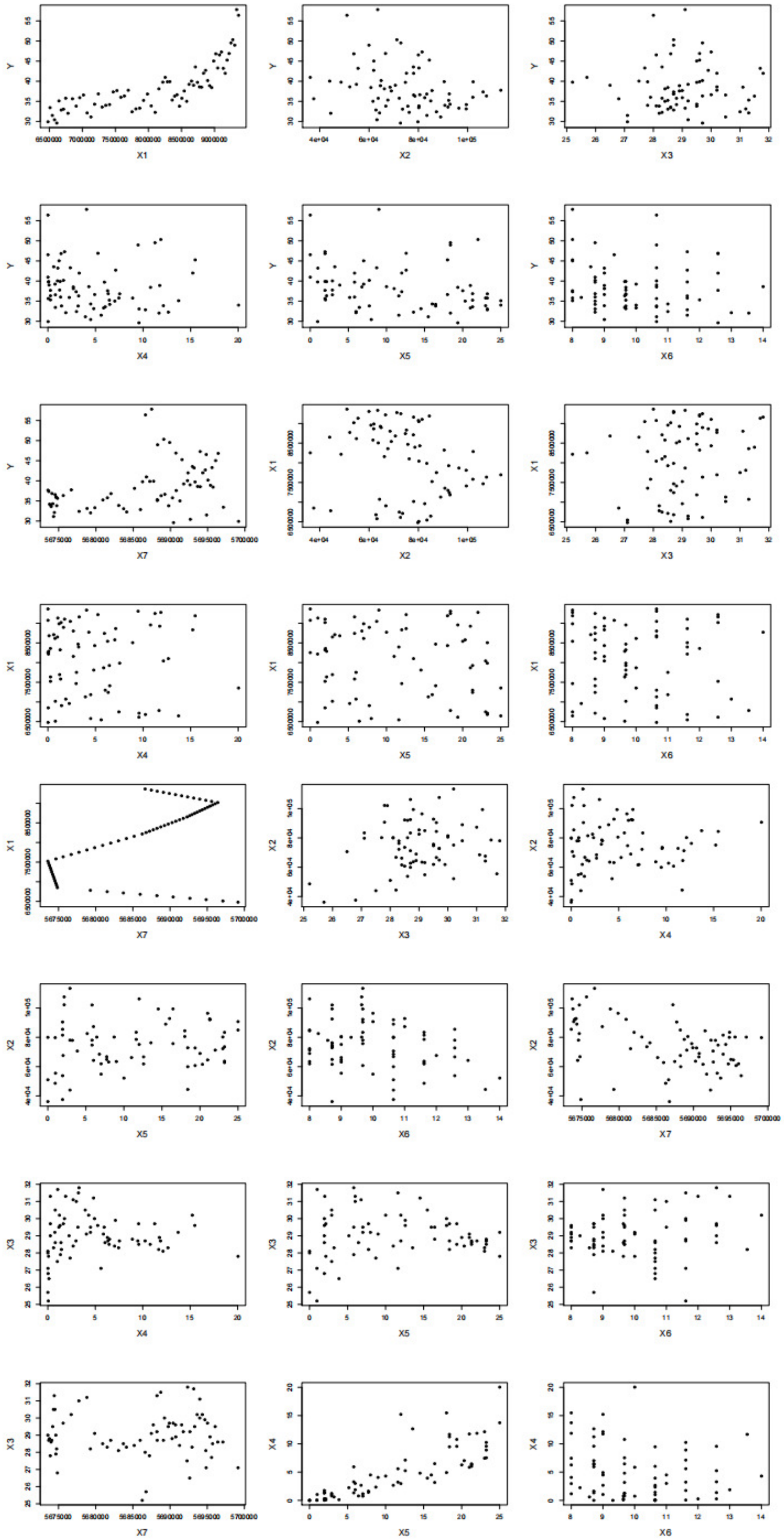


Fig7(cont.)

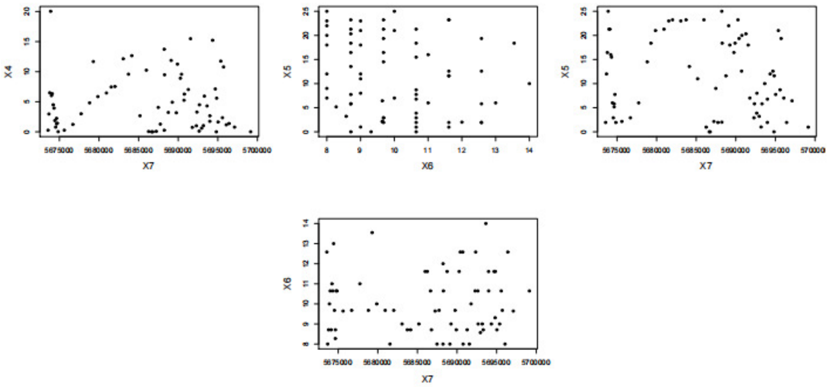


Table 5a: Results of the Multiple Linear Regression Model with Seven Independent Variables.

Variable	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	t value	Pr(> $\|t\|$)	Partial SS
(Intercept)	5.805×10^{2}	3.819×10^{2}	1.520	0.1335	-
X_1	5.263×10^{-6}	5.742×10^{-7}	9.166	3.34×10^{-13}	1530.6
X_2	-3.242×10^{-5}	3.059×10^{-5}	-1.060	0.2933	10.0
X_3	-4.364×10^{-1}	3.484×10^{-1}	-1.253	0.2149	36.2
X_4	1.512×10^{-1}	1.591×10^{-1}	0.951	0.3455	10.9
X_5	-1.845×10^{-1}	9.387×10^{-2}	-1.966	0.0537	35.5
X_6	-4.684×10^{-1}	3.027×10^{-1}	-1.547	0.1268	27.2
X_7	-9.909×10^{-5}	6.734×10^{-5}	-1.471	0.1462	26.2

Table 5b: Analysis of Variance

Source	d.f.	Sum of Squares	Mean Square	F-statistic	Pr(> F)
Regression	7	1676.64	239.52	19.80	9.511×10^{-14}
Error	63	762.05	12.10		
Total	70	2438.69			

R-squared = 0.6875 , Adjusted R-squared = 0.6528

Table 6a: Results of The Simple Linear Regression Model of X_1.

Variable	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	t value	Pr(> $\|t\|$)	Partial SS
(Intercept)	-3.296	3.886	-0.848	0.399	-
X_1	5.165×10^{-6}	4.790×10^{-7}	10.784	$< 2.2 \times 10^{-16}$	1530.6

Table 6b: Analysis of Variance

Source	d.f.	Sum of Squares	Mean Square	F-statistic	Pr(> F)
Regression	1	1530.6	1530.6	116.30	$< 2.2 \times 10^{-16}$
Error	69	908.1	13.20		
Total	70	2438.7			

R-squared = 0.6276 , Adjusted R-squared = 0.6222

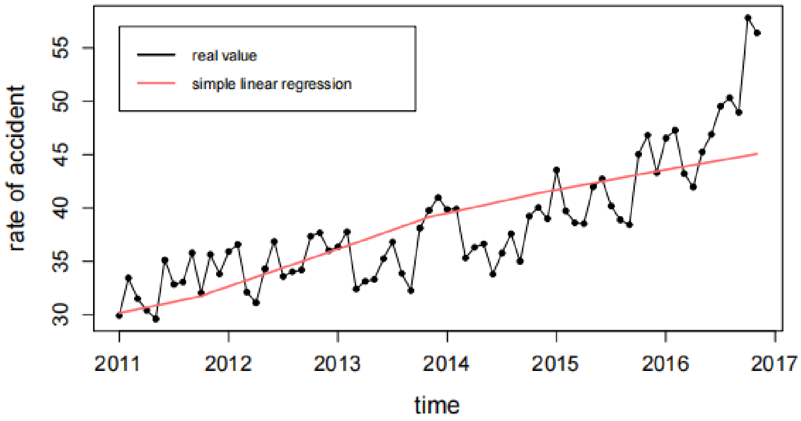


Figure 8: Predicted Value of adj RAR by Simple Linear Regression Of X_1

better fit as shown in the following Tables 7a and 7b.

The coefficients of both the quadratic as well as cubic terms are indeed significant. Also, the R-square improves significantly from 62.76% to 83.60%, a jump of more than 20%. The cubic polynomial regression model using X_1 only is given in the following equation (4). The equation of cubic polynomial of X_1 model is

$$\hat{Y} = -1492 + 0.0006013X_1 - 7.883 \times 10^{-11}X_1^2 + 3.442 \times 10^{-18}X_1^2. \tag{4}$$

Figure 10 shows the plot between the predicted value from the cubic polynomial model against time. In addition, this plot also shows the predicted value from the simple linear regression model against time (for a visual comparison between two models).

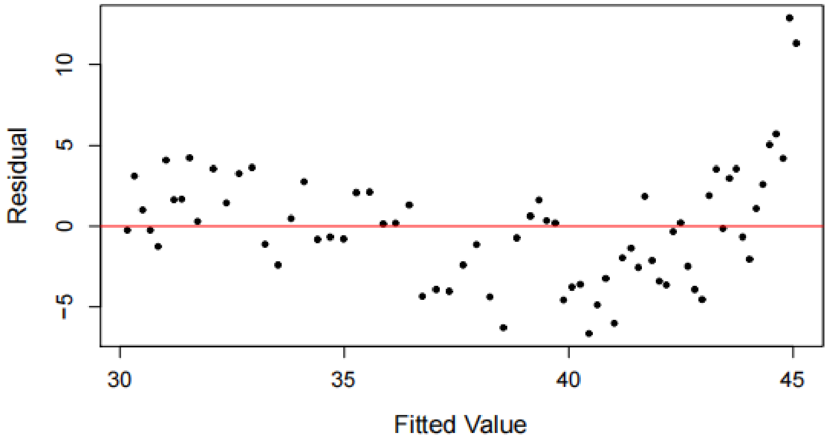


Figure 9: Residuals Versus the Fitted Values

Table 7a: Results of the Cubic Polynomial of X_1.

Variable	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	t value	Pr(> $\|t\|$)	Partial SS
(Intercept)	-1.492×10^3	2.940×10^2	-5.075	3.29×10^{-6}	-
X_1	6.013×10^{-4}	1.124×10^{-4}	5.348	1.16×10^{-6}	1530.6
X_1^2	-7.883×10^{-11}	1.424×10^{-11}	-5.536	5.57×10^{-7}	309.9
X_1^3	3.442×10^{-18}	5.974×10^{-19}	5.763	2.28×10^{-7}	198.2

Table 7b: Analysis of variance

Source	d.f.	Sum of Squares	Mean Square	F-statistic	Pr(> F)
Month	3	2038.71	679.57	113.83	$< 2.2 \times 10^{-16}$
Error	67	399.98	5.97		
Total	70	2438.69			

R-squared = 0.8360, Adjusted R-squared = 0.8286

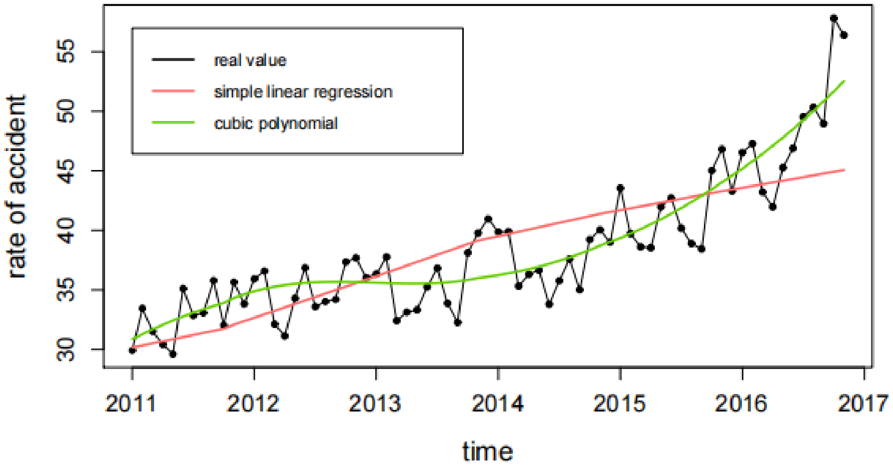


Figure 10: Predicted Value of adj RAR by Cubic Polynomial of X_1

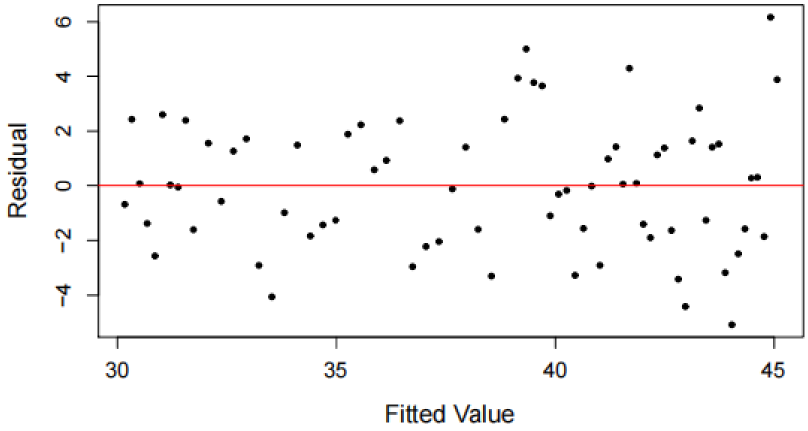
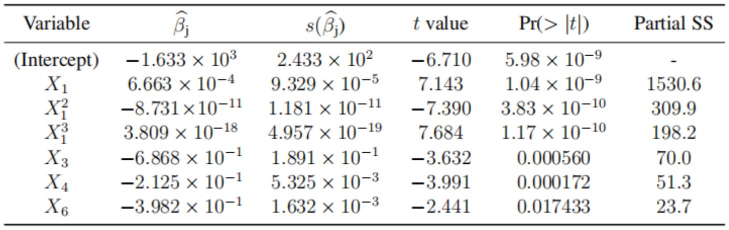


Figure 11: Residuals Versus the Fitted Values

The above Figure 11 shows equally spread residuals around a horizontal line without any distinct pattern. For the cubic polynomial model the residual plot looks better. The Q-Q plot of the residuals as well as the usual model assumption checks are provided in the Appendix (Figure 15).

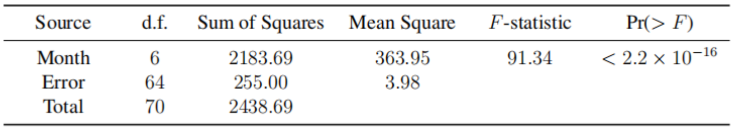
Next, we apply the cubic polynomial regression model above by adding all other independent variables to see if they can provide any extra information.

Table 8a: Results of Multiple Regression Model with Cubic Polynomial in X_1 and X_3, X_4, and X_6 in Linear Terms.



Variable	$\widehat{\beta}_j$	$s(\widehat{\beta}_j)$	t value	Pr(> $\|t\|$)	Partial SS
(Intercept)	-1.633×10^3	2.433×10^2	-6.710	5.98×10^{-9}	-
X_1	6.663×10^{-4}	9.329×10^{-5}	7.143	1.04×10^{-9}	1530.6
X_1^2	-8.731×10^{-11}	1.181×10^{-11}	-7.390	3.83×10^{-10}	309.9
X_1^3	3.809×10^{-18}	4.957×10^{-19}	7.684	1.17×10^{-10}	198.2
X_3	-6.868×10^{-1}	1.891×10^{-1}	-3.632	0.000560	70.0
X_4	-2.125×10^{-1}	5.325×10^{-3}	-3.991	0.000172	51.3
X_6	-3.982×10^{-1}	1.632×10^{-3}	-2.441	0.017433	23.7

Table 8b: Analysis of Variance



Source	d.f.	Sum of Squares	Mean Square	F-statistic	Pr(> F)
Month	6	2183.69	363.95	91.34	$< 2.2 \times 10^{-16}$
Error	64	255.00	3.98		
Total	70	2438.69			

R-squared = 0.8954, Adjusted R-squared = 0.8856

The equation of the Cubic Polynomial of X_1, X_3, X_4 and X_6 in linear term is

$$\hat{Y} = -1633 + 0.0006663X_1 - 8.731 \times 10^{-11}X_1^2 + 3.809 \times 10^{-18}X_1^3 - 0.6868X_3 - 0.2125X_4 - 0.3982X_6. \quad (5)$$

Figure 12 shows the plot of the predicted value from equation 5. In addition, this plot also show the predicted value from simple linear regression model and cubic polynomial of X_1 model against time for a visual comparision among the three models.

The following Figure 13 shows almost equally spread out residuals around a horizontal line without any distinct pattern. For the multiple regression model with cubic polynomial in X_1 and X_3, X_4, and X_6 in linear terms the residual plot looks much better. The Q-Q plot of the residuals as well as the usual model assumption checks are provided in the Appendix (Figure 16).

We have also used AR(1) term in the above model to see if that can provide any further improvement. But it didn't help.

4 Conclusion

In this section we summarize our findings of our traffic accident data analysis based on the final regression model (5).

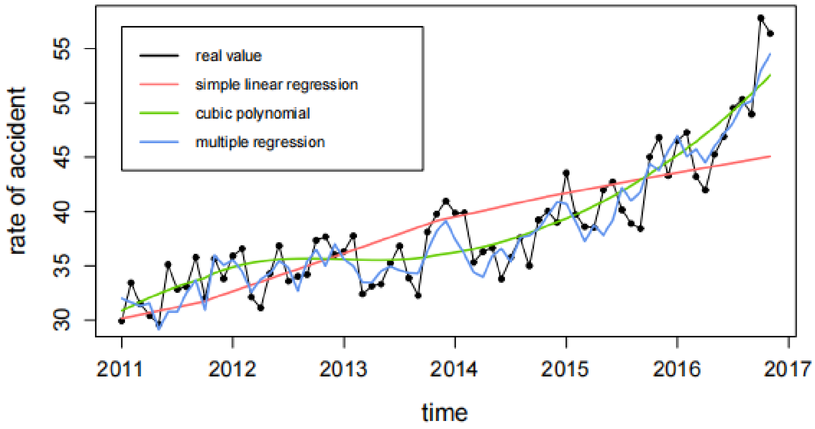


Figure 12: Y versus Model

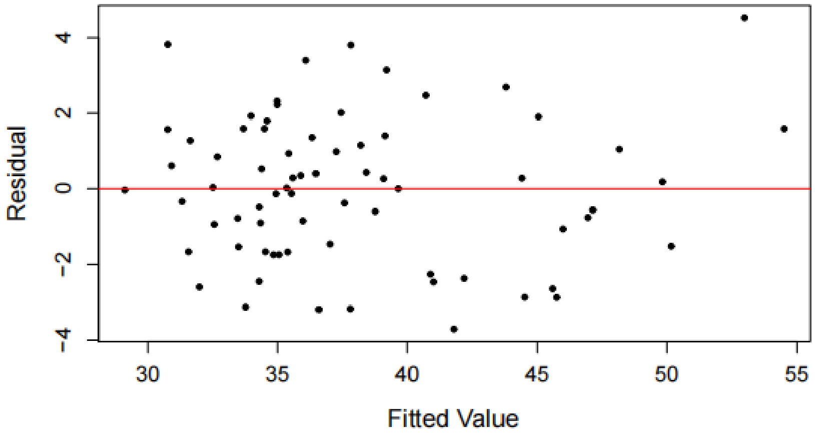


Figure 13: Residuals Versus the Fitted Values

(a) The independent variables which are found to have significant contribution in explaining Y (i.e., monthly adjusted RAR) are X_1 (i.e., adjusted total number of all vehicles per month), X_3 (i.e., monthly mean temperature), X_4 (i.e., average total amount of rainfall (mm) in a month), and X_6 (i.e., adjusted number of holidays per month, including weekends).

(b) The above four independent variables, including the cubic polynomial in X_1, have a combined R-square values of 0.8954 , in explaining Y. In other words, nearly 90% of the variability in adjusted monthly road accident rate can be attributed to the above four independent variables. The relative contribution of the four independent variables in explaining Y, based on the step-wise regression model build-up, can be explained as follows: (i) X_1 contributes (through its cubic polynomial) about 83.60%; (ii) X_3 contributes about 2.87% (or, approximately 3%); (iii) X_4 contributes about 2.10% (or, approximately about 2%); and (iv) X_6 contributes about 0.97% (or, approximately 1%).

(c) A further look at the final multiple regression model (5) reveals that when X_1 is "small", then the impact of the marginal effect of X_1 can be negative, i.e., the adjusted RAR may go down. In other words, more vehicles may help the commuters which may decrease the number of pedestrians getting hit by the cars. But when X_1 itself is high, then it has a positive impact on Y, i.e., with a higher number of vehicles already in the city, further increase in the number of vehicles can increase the number of accidents. This is what we are seeing in the plot. From the year 2015 there has been a sudden spurt in the total number of vehicles which is causing more traffic accidents. All the other three independent variables (X_3, X_4 and X_6) have negative impact on Y. That means, as the mean temperature soared, it kept drivers and pedestrians indoor, and caused less traffic accidents. Similarly, more rainfall had a dampening effect on the overall traffic accidents as it kept vehicles off road. Similarly, total number of holidays decreased the traffic accident rates. But note that the marginal effects of these three variables are really small, about 3%,2% and 1% respectively.

(d) Out of the four independent variables the last three (i.e., X_3, X_4 and X_6) are "natural" (i.e., human interference is either nonexistent or minimal). On the other hand, the biggest contributing variable is X_1 which is totally "man-made" (i.e., can be controlled if desired).

(e) Therefore, the overall traffic accidents can be reduced perhaps by reducing the total number of vehicles in the City of Bangkok. Possibly this statistical research finding calls for more funding and upgrading the public transportation system which can discourage more vehicles to be on the roads. Other metros in Southeast Asian region, such as Kolkata, Dhaka, Ho Chi Minh City, etc. can take lessons from this study for better traffic management in order to reduce road accidents.

Acknowledgements

This research project, titled, 'A Statistical Analysis of Bangkok Traffic Accident Data' was undertaken and completed by Ms. Channgam, while she visited the University of Louisiana at Lafayette, USA, with a generous research grant from the Science Achievement Scholarship of Thailand (SAST) Project of the Royal Thai Government. Ms. Channgam would like to express her deep gratitude and appreciation to Professor Nabendu Pal (Department of Mathematics, University of Louisiana at Lafayette) for making the local arrangements for her to undertake this research under his supervision. The final version of this research was prepared in 2019, but the subsequent pandemic disruption delayed the dissemination of this research project. This paper was presented as an invited talk by the third author (Nabendu Pal) at the Workshop on Mathematical Research, Education and Applications (WMREA - NTTU & DthU 2024), jointly organized by Nguyen Tat Thanh University & Dong Thap University,15th - 18th July 2024. The author is indebted to the organizers for their kind hospitality. Also, the presentation materials were prepared while the third author was visiting the Vietnam Institute for Advanced Study in Mathematics (VIASM) in Summer of 2024.The author is very much appreciative of the support received from the VIASM administration.

Appendix

Figure 14 shows the normal Q-Q plot of the residuals from Table 6a based on the 71 observations. In the following we provide the results of normality and homoscedasticity tests.

1. Normality Assumption (1.1) Anderson-Darling normality test: $A = 0.52514, p-$value =0.175.

(1.2) Shapiro-Wilk normality test: $W = 0.94791, p-$value =0.00522. Thus it is question, we cannot conclude that.

2. Homoscadasticity

(2.1) Levene's Test for Homogeneity of Variance (center = mean): F-statistic = 0.5748,p-value =0.4509.

(2.2) Levene's Test for Homogeneity of Variance (center = median):F-statistic = 0.3365,p-value =0.5638.

The above test indicates that the variances are constant.

Figure 15 shows the normal Q-Q plot of the residuals from Table 7a based on the 71 observations. In the following we provide the results of normality and homoscedasticity tests.

1. Normality Assumption

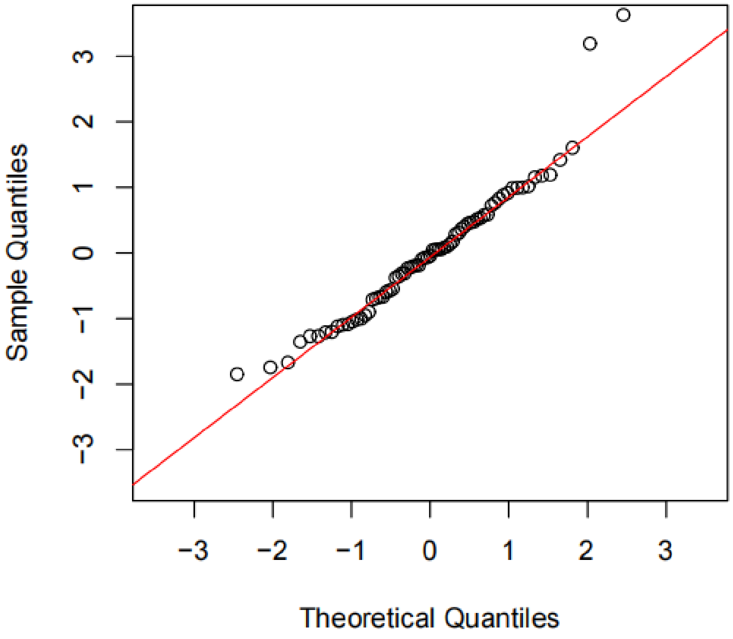


Figure 14: Normal Q-Q Plot

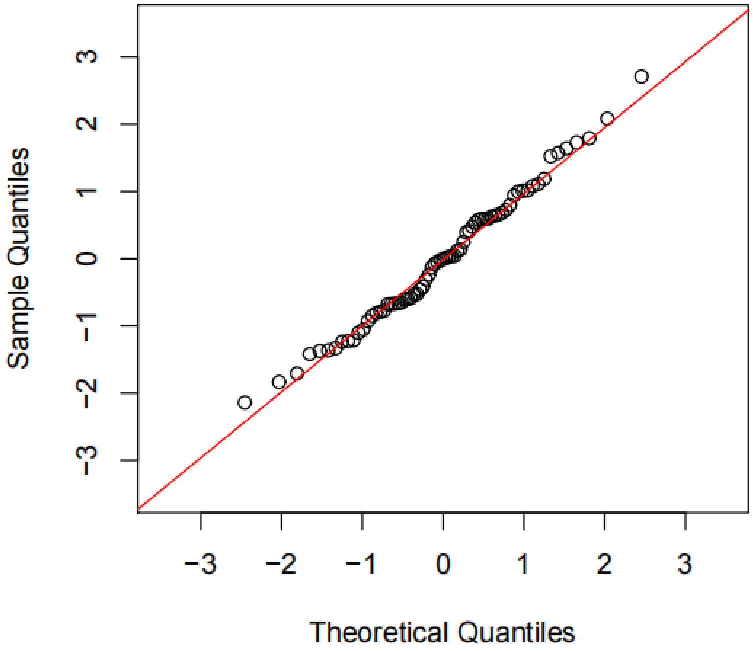


Figure 15: Normal Q-Q Plot

(1.1) Anderson-Darling normality test: A=0.31408,p-value =0.538.

(1.2) Shapiro-Wilk normality test: W =0.98812,p-value =0.7442. Thus it is no evidence against the assumption that the errors follow a normal distribution.

2. Homoscadasticity

(2.1) Levene's Test for Homogeneity of Variance (center = mean):F-statistic = 0.7063,p-value =0.4036.

(2.2) Levene's Test for Homogeneity of Variance (center = median):F-statistic = 0.381,p-value =0.5391.

The above test indicates that the variances are constant.

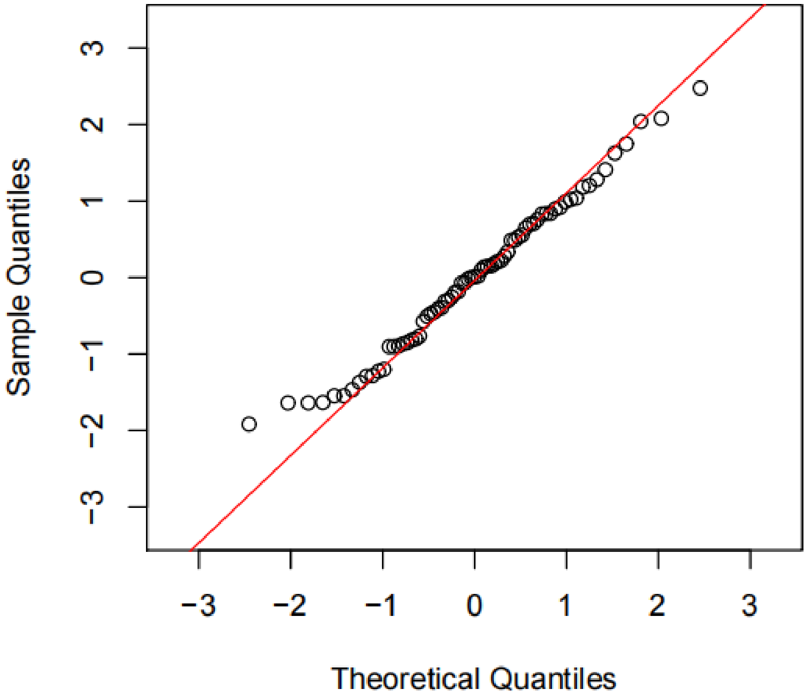


Figure 16: Normal Q-Q Plot

Figure 16 shows the normal Q-Q plot of the residuals from Table 8a based on the 71 observations. In the following we provide the results of normality and homoscedasticity tests.

1. Normality Assumption (1.1) Anderson-Darling normality test: A=0.21879,p-value =0.8321.

(1.2) Shapiro-Wilk normality test: $W = 0.98536, P-$value =0.5801.

Thus it is no evidence against the assumption that the errors follow a normal distribution.

2. Homoscadasticity

(2.1) Levene's Test for Homogeneity of Variance (center = mean):f-statistic = 0.0212,$p-$value =0.8847.

(2.2) Levene's Test for Homogeneity of Variance (center = median):$F-$statistic = 0.0109,p-value =0.917.

The above test indicates that the variances are constant.

References

[1] Accident Alert Network. (29 April 2017). Pickup truck turned upside down because shaft I brokens. [online] Available at: http://www.thairsc.com/th/BigAccDetail.aspx?qid=46067[Accessed 27 July 2017].

[2] Accident Alert Network. (22 May 2017). van collided with car cleaner. [online] Available at: http://www.thairsc.com/th/BigAccDetail.aspx?qid=46113[Accessed 27 July 2017].

[3] Augmented Dickey-Fuller Unit Root Tests. [online] Available at: https://faculty.smu.edu/tfomby/eco6375/bj%20notes/adf%20notes.pdf

[4] Brijs T. and Karlis D. (2008). Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accident Analysis and Prevention, 40, 1180-1190.

[5] Dettlin Marcel. (2013). Applied Time Series Analysis. Zurich.

[6] Garrido R., Bastos A., Almeida A. and Elvas J. (2014). Prediction of road accident severity using the ordered probit model. Transportation Research Procedia, 3, 214 223.

[7] Gregory W. Corder and Dale I. Foreman. (2014). Nonparametric statistics : a stepbystep approach. second edition. Hoboken, New Jersey.

[8] History. com staff. 9/11: Timeline of Events. [online] Available at: http://www. history.com/topics/911timeline[Accessed 10 August 2017].

[9] John O. Rawlings, Sastry G. Pentula, David A. Dickey.(1998). Applied Regression Analysis: A Research Tool. 2nd ed. New York: Springer.

[10] Jon Fernquest. (20 Feb 2017). Bangkok traffic jams among world's worst. [online] Available at: http://www.bangkokpost.com/learning/advanced/1201724/bangkok-traffic-jams-among-worlds-worst[Accessed 27 July 2017].

[11] Kpss test. [online] Available at: https://www.researchgate.net/file.PostFileLoader.html?id=58c82201b0366d1894358fc1&assetKey=AS%3A471857125171200%401489 510913553.

[12] Lee W., Lee H. , Hwang S. , Kim H. , Lim Y., Hong Y. , Ha E. and Park H. (2014). A time series study on the effects of cold temperature on road traffic injuries in Seoul, Korea. Environmental Research, 132, 290-296.

[13] Mohamed Ahmed Zaid.(2015). Correlation and Regression Analysis TEXTBOOK. Ankara: Diplomatik Site.

[14] Official Statistics Registration System. Official Statistics Registration System. [online] Available at: http://stat.dopa.go.th/stat/statnew/ upstat_age.php [Accessed 27 May 2017].

[15] Ratanavaraha V. and Suangka S. (2014). Impacts of accident severity factors and loss values of crashes on expressways in Thailand. IATSS Research, 37, 130-136.

[16] Rate of accidents. [online] Available at: http://www.thairsc.com/[Accessed 10 August 2017].

[17] Riley C. and Sherman I. (18 January 2017) World's largest economies. [online] Available at: http://money.cnn.com/news/economy/world_economies_gdp/index.html [Accessed 15 August 2017].

[18] Romer M. (2017). STAT 510: Applied Time Series Analysis. [online] Available at: https://onlinecourses.science.psu.edu/statprogram/stat510[Accessed 5 July 2017].

[19] Simon L. and Heckard R. (2017). STAT 501: Regression Methods. [online] Available at: https://onlinecourses.science.psu.edu/statprogram/stat501[Accessed 5 July 2017].

[20] Tanaboriboon Y. (2005). Determination of economic losses due to road crashes in Thailand. Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 3413 3425,2005.

[21] Tanaboriboon Y. and Satiennam T. (2004). Traffic accident in Thailand. IATSS RESEARCH Vol. 29 No.1, 2005. [22] ThaiRSC. (2015). Statistical Report. [online] Available at: http://rvpreport.rvpeservice.com/viewrsc.aspx?report=0486&session=16 [Accessed 17 May 2017].

[22] The Nation. (14 April 2016). Four killed, six injured in Bangkok road accident. [online] Available at: http://www.nationmultimedia.com/detail/breakingnews/30283938 [Accessed 27 July 2017].

[23] The Nation. (27 March 2017). Three killed, four others injured in Bangkok car accident. [online] Available at: http://www.nationmultimedia.com/news/breakingnews/30310388[Accessed 27 July 2017].

[24] The World Bank. (2017). Thailand GDP. [online] Available at: http://data.worldbank .org/country/thailand[Accessed 17 August 2017].

[25] Transportation Statistics Group. Transport statistics. [online] Available at: http://apps. dlt.go.th/statistics_web/statistics.html[Accessed 27 May 2017].

[26] World Health Organization. (2013). Global status report on road safety 2013. Geneva: Management of Noncommunicable Diseases, Disability,Violence and Injury Prevention (NVI).

[27] World Health Organization. (2015). Global status report on road safety 2015. Geneva: Management of Noncommunicable Diseases,Disability,Violence and Injury Prevention (NVI).

[28] (2010). Timeline: Thailand's political crisis. [online] Available at: http://www.cnn.com/2010/WORLD/asiapcf/05/17/thailand.timeline/index.html#[Accessed 17 August 2017].

[29] (2017). 2010 Thai political protests. [online] Available at: https://en.wikipedia.org/wiki/2010_Thai_political_protests[Accessed 17 August 2017].