Proceedings of Science and Mathematics
Faculty of Science,
Universiti Teknologi Malaysia
Vol. 2, 2021, page 177-185

Statistical Analysis of Road Accident in Malaysia

[1]Muhamad Latif Zulhizad and [2]Noraslinda Mohamed Ismail

[1,2]Department of Mathematical Sciences
Faculty of Science, Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia.

e-mail: [1]latifzulhizad46@gmail.com , [2]noraslinda@utm.my

Abstract This research focus on the analysis of cases and casualty of road accident in Malaysia. Multivariate Analysis of Variance (MANOVA) is used to see whether there exists a significant difference for the rate of cases and casualty of road accident by years and states. Simple Linear Regression also used to predict the number of cases and casualty with type of road user and cause of road accident as the independent variables. The experimental result will help the Malaysian Road Safety Department (MRSD) and Royal Malaysian Police to find a solution to reduce the number of cases and casualty of road accidents in Malaysia

Keywords road accident; multivariate analysis of variance; simple linear regression; cases; casualty;

1 Introduction

One of the leading causes of death and injury in Malaysia is traffic accidents. Every week many people get involve with road accident whether injured or fatally. In these kinds of cases, the position of healthcare organizations is becoming critical, as the first few minutes after the "golden hour" accident are very precious and vital. Many lives may be saved and injuries averted if crash victims were given quick medical attention. Accidents resulting from the extension of air, ground water, river transportation, and road transport are the most common in terms of occurrence and scale, as well as human and economic losses. It is described that the transport mechanism is a transfer of persons and goods from one place to another. Transportation can take many forms, including air, train, sea, water, cable, and space travel. A road is a visible route or path that connects two or more points. To make travel easier, roads are generally sanded, paved, or otherwise prepared. An automobile with great adaptability and modest power is the most popular road vehicle. It makes for more effective travel at a lower cost.

Road accident can be divides into three type which is fatal, serious, and minor. Fatal is the result of the accident that lead to death. Serious is where the accident is not deadly but the person got inflict harm and injury and minor is accident which person get minor injury or no damage. Research has shown that there are three factors that can cause road accident which is human, mechanical and environment. Hansen and Benjamin [1]said that human is the top cause of road accident which mostly the case where the driver intoxication themselves with drugs and alcohol.

This research mainly focused on Malaysia road traffic accidents. It includes the rate of accident cases as well as the rate of casualties (injured and dead) in each Malaysian state. This study also looks into the causes of road accidents and the types of road users that are involved in them. The objective of this research is to see whether exists significance difference for cases and casualty of road accidents within year from 2014 to 2020 and all states of Malaysia. This research also wants to predict the number of cases and casualty of road accident in Malaysia with type of road user and total cause of road accident as the independent variables.

2 Literature Review

2.1 Road Traffic in Malaysia

Road transport is a necessity which appears to be of great advantage for country and an individual in particular to enhance access to work, economic, educational and health center. But there is some negative effect about road transport which is road accident where it can happen to person, families and communities that can lead to injury and deaths. As one of the developing countries, Malaysia has acknowledged road safety as a critical problem that should be approached. This is because according [2], the World Health Organization and World Bank has report that most of the nation in the world are severely affected by the fatalities of the road traffics accidents where about 90% of the accident leads to death.

In the past decade, Musa MF [2] said that the total of traffic accidents, which in 2007 and 2016, increased by approximately 41 percent, amounted to 369.319 and 521.466 accidents respectively, according to the Malaysian Transportation Ministry. Therefore from 6,282 to 7,152 the overall number of deaths is raised. Over the same period, there has been a rise of 666,027 to 960,569 in overall vehicle involving road crashes (around 44 percent). During the year 1997–98, according to Ministry of Health Malaysia, accident was the third most common reason for hospitalization and the fourth most common cause of death, behind heart disease, septicemia, and cerebrovascular accident. The Government takes this injury data seriously and has taken measures to implement action by 5E methods to tackle the problem of road safety in a holistic manner.

According to Singh [3], many causes have triggered road accidents that are primarily linked to the users, the environment and the vehicles. The driver's gender, age, experience, physical and behavior attributes will make a major contribution towards crash chance and severity. In a case study, Wedagama [4] said that it shown that crashes with pedestrians and other vehicles can lead to fatal vehicle accidents. Traffic and transport characteristics including traffic volume and structure and speed estimation can be used in many models for road accidents as the typical variables. It being observer that most road accidents that happen in worldwide occur because the role of intoxication with drugs and alcohol by the driver [1]. One by one, reports of motorists driving under the influence of alcohol or drugs sparked fury among those who demanded that the government adopt a comprehensive approach to risky drivers [3].

2.2 Analysis using Multivariate Statistics

According Muller [5], Multivariate statistics is quite flexible and can be used for following models such as ANOVA, ANCOVA, MANOVA, MANCOVA and repeated measures models with and without time-varying and time-constant covariates. For the road accident analysis, we focus using MANOVA to analyze the value of road accident in Malaysia. Multivariate analysis of variance (MANOVA) can have one or several grouping variables, however several consequence variables will be included (say P in number). It is the influence of the grouping variable(s), which interests the researcher using MANOVA techniques, in gathering the results variables.

There is One-Way MANOVA and Two-Way MANOVA. From a website [6], One-Way MANOVA is used to determine if variations occur in more than one continuous dependent variable between independent groups. In this aspect, it varies from an ANOVA one direction, which calculates only one element depending on it. Two-Way MANOVA is a case where there are two or more dependent variables which is an expansion of the two-way ANOVA [6]. The main purpose of the two-way MANOVA is to understand if occur interaction between two independent variables on the two or more dependent variables.

2.3 Predicting using Linear Regression Models

Many researchers use regression analysis for prediction and forecasting. Regression analysis is a technique for predictive modelling that estimates the connection between two or more variables. Recall that a correlation analysis does not assume that two variables are causally related. In correlation analysis, [7] Kumari state that the correlation coefficient is a dimensionless number whose value ranges from −1 to +1. A value toward −1 indicates inverse or negative relationship, whereas towards +1 indicate a positive relation.

The analysis of regression focuses on the relationship between the dependent variable and the independent variable (predictors). J. Ramsey[8] said that the average value of Y for a certain X value is also predicted by mathematical researchers using a straight line in a linear relation (called the regression line). You would be able to link the value for X and estimate the average value for Y if you know the direction and y-intercept of the regression line. In other words, average Y from X is expected. Correlation and linear regression are used to examine the relationship between two variables. Correlation and regression both illustrate the strength of a linear connection between two variables, but regression does it in the form of an equation.

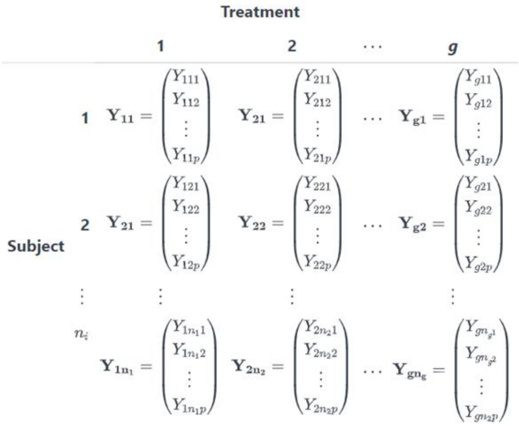
3 Methodology

3.1 One-Way Multivariate Analysis of Variance (One-Way MANOVA)

One-Way MANOVA is known as a statistical test which tests the relationship between two different variables which can be seen as a set of numeric variables and a single categorical variable.

3.1.1 Formulation of One-Way MANOVA

Suppose that we have data on p variables which we can arrange in a table such as the one below:



		Treatment			
		1	2	⋯	g
	1	$\mathbf{Y}_{11} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{11p} \end{pmatrix}$	$\mathbf{Y}_{21} = \begin{pmatrix} Y_{211} \\ Y_{212} \\ \vdots \\ Y_{21p} \end{pmatrix}$	⋯	$\mathbf{Y}_{g1} = \begin{pmatrix} Y_{g11} \\ Y_{g12} \\ \vdots \\ Y_{g1p} \end{pmatrix}$
Subject	2	$\mathbf{Y}_{21} = \begin{pmatrix} Y_{121} \\ Y_{122} \\ \vdots \\ Y_{12p} \end{pmatrix}$	$\mathbf{Y}_{22} = \begin{pmatrix} Y_{221} \\ Y_{222} \\ \vdots \\ Y_{22p} \end{pmatrix}$	⋯	$\mathbf{Y}_{g2} = \begin{pmatrix} Y_{g21} \\ Y_{g22} \\ \vdots \\ Y_{g2p} \end{pmatrix}$
	⋮	⋮	⋮		⋮
	n_i	$\mathbf{Y}_{1n_1} = \begin{pmatrix} Y_{1n_11} \\ Y_{1n_12} \\ \vdots \\ Y_{1n_1p} \end{pmatrix}$	$\mathbf{Y}_{2n_2} = \begin{pmatrix} Y_{2n_21} \\ Y_{2n_22} \\ \vdots \\ Y_{2n_2p} \end{pmatrix}$	⋯	$\mathbf{Y}_{gn_g} = \begin{pmatrix} Y_{gn_{g}1} \\ Y_{gn_{g}2} \\ \vdots \\ Y_{gn_2p} \end{pmatrix}$

Figure 1 show the randomized block design data

In this multivariate case the scalar quantities, of the corresponding table in ANOVA, are replaced by vectors having p observations.

Notation:

Y_{ijk}= Observation for variable k from subject j in group i. These are collected into vectors:

$$Y_{ij} = \begin{pmatrix} Y_{ij1} \\ Y_{ij2} \\ \vdots \\ Y_{ijp} \end{pmatrix} = \text{Vector of variables for subject } j \text{ in group } i \quad (1)$$

n_i = the numbers of subjects in group i

$$N = n_1 + n_2 + \cdots + n_g = \text{Total sample size} \quad (2)$$

We will testing's the null hypothesis so that the group mean vectors are all equal to one another. Mathematically this is expressed as:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_g$$

The alternative hypothesis:

$H_a: \mu_{ik} \neq \mu_{jk}$ for at least one $i \neq j$ and at least one variable k

This says that the null hypothesis is false if at least one pair of treatments is different on at least one variable.

3.1.2 Assumptions of One-Way MANOVA

There are some assumptions we need to go through when using MANOVA.

1. The observation must be sampled randomly and independently from the population
2. There is an interval measurement for each dependent variable
3. For each group of the independent variables, the dependent variables need to be multivariate normally distributed.
4. The population covariance matrices are equal for each group (this is the extension of homogeneity of variances for univariate ANOVA)

3.2 Linear Regression Model

Regression analysis is often used to model the relation between a single variable Y and one or more predictors. When we have one predictor, we call it simple linear regression. An estimation of the true relationship based on real measurements is a statistical regression model based on multiple assumptions on the distributive characteristics of variables X and Y

3.2.1 Linear Regression Formula

The linear regression formula is:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (3)$$

where y = predicted value for the dependent variable

x = independent variable (the variable that we expect influencing y)

β_0= the intercept (predicted of y when x is 0)

β_1= regression coefficient (how much y change as x increase)

ε= error of the estimate

Linear regression determines the line of best fit by searching the regression coefficient, which minimizes the cumulative model error.

3.2.2 Assumptions for Linear Regression Models

There are four assumptions:

1. Linearity: The relationship between the independent (x) and dependent (y) variable is linear
2. Homoscedasticity: The size for the prediction of error doesn't change significantly across the values of the independent variable.
3. Independence: The findings in the dataset have been obtained using statistically appropriate sampling techniques, and the correlations between observations are not concealed.
4. Normality: The data follow normal distribution

4 Results and Discussions

4.1 Analysis Using One-Way Multivariate Analysis of Variance (MANOVA)

For this research, the data we used is from 2014 until 2020 and all states of Malaysia that combine into three categorize which is North, Central and South states. The assumption for the MANOVA is check before the analysis. The data for the rate of the cases and casualty is sampled randomly and independently from the population. The data also has interval for the rate of the cases and casualty. Last but not least, checking the normality of the data. Shapiro-Wilk test is use to test the data.

Table1: Tests of Normality

	Shapiro-Wilk		
	Statistic	df	Sig.
Cases	.901	21	.054
Casualty	.942	21	.239

From result above, since the value of significance is greater than 0.05 for both cases and casualty. we can state that the data for the rate of cases and casualty of road accident is normally distributed. Since the data follow all the assumption for the MANOVA, we can continue with the analysis for the road accident. For the analysis, we focus to use the Wilk's Lambda test to see if the data for the independent and dependent are significant. We using the value alpha = 0.05.

Table 2: The value of the significance

Independent Variables	Significance of Wilks' Lambda
Years	0.650
States	0.000

The value of the significance of the Wilks' Lambda for years is more than alpha value while for states the value is less than alpha value. Since the value of the significance of the Wilk's Lambda for years is greater than 0.05, we can say that year is not important to determine the differences for the dependent variables but for states, it important since the value of the Wilk's Lambda is less than p-value. From here we can see that there are no statistically significant differences based on years but for states, there exist a significant difference for the rate of the road accident. The follow-up test is continuing for states since there exists significant differences in the analysis. The follow-up test are tests of between-subject effect and Tukey HSD Post Hoc tests.To see the significant difference for cases and casualty, we use test off between-subject effects.

Table 3: The tests of between-subject effects

Tests of Between-Subjects Effects						
Source	D. Variable	Sum of Squares	df	Mean Square	F	Sig.
State	Cases	.113	2	.056	46.370	.000
	Casualty	.000	2	.000	26.023	.000

Based on table 3, we can see that the rate of road accident in Malaysia has a statistically significant effect on both cases and casualty since the value of the significance = 0.000 which is less than p-value (0.05),

$$F(2{,}18) = 46.730\ , p < 0.05 \text{ [For cases]}$$
$$F(2{,}18) = 26.023\ , p < 0.05 \text{ [For casualty]}$$

Since there was a significant difference for the independent variable and both dependent variables, we can accept the significant difference at $p < 0.05$. Therefore, at least one group differ significantly with other group for the number of cases and casualty.

Tukey HSD were used to see the significant difference for each categorized.

Table 4: The result of the Tukey HSD

Multiple Comparisons					
Tukey HSD					
Dependent Variable	(I) State	(J) State	Mean Difference (I-J)	Std. Error	Sig.
Case	North	Central	-.171714*	.0186479	.000
		South	-.040329	.0186479	.105
	Central	North	.171714*	.0186479	.000
		South	.131386*	.0186479	.000
	South	North	.040329	.0186479	.105
		Central	-.131386*	.0186479	.000
Casualty	North	Central	.008171*	.0011335	.000
		South	.003814*	.0011335	.009
	Central	North	-.008171*	.0011335	.000
		South	-.004357*	.0011335	.003
	South	North	-.003814*	.0011335	.009
		Central	.004357*	.0011335	.003

Based on table 4, we can see that the mean scores for cases were statistically different between North state and Central state, and Central state and South state with $p < 0.05$ and the confidence interval also show that the value significant difference but not between North state and South state. For north and south state, the value of the significance is 0.105 and it greater to the p-value which 0.05. From there, we can see that there are not many significant differences for the North and South states which we can see from the confidence interval also because the lower and upper bound have different sign value.

For casualty, the mean scores between all comparison were statistically significant since all comparison give value $p < 0.05$ and the interval for the comparison show the same sign for every comparison. These differences can be easily visualized by plot generated below,

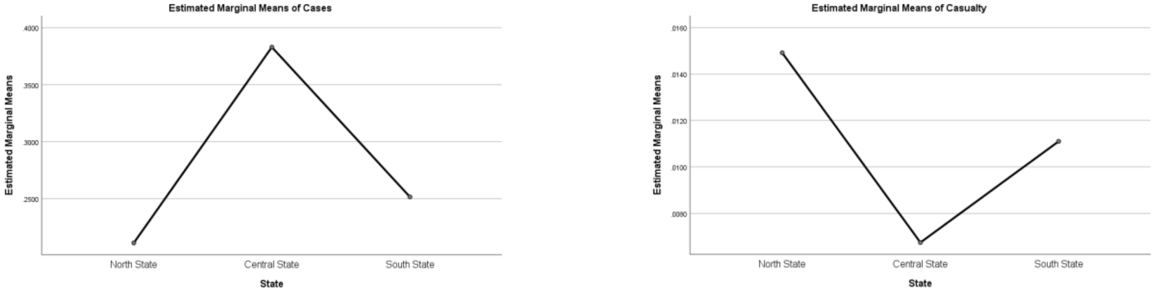


Figure 1: The mean for the significant difference for road accident in Malaysia based on states for cases and casualty

Based on figure 1, we can see more about the differences value of the mean from each category of states for cases and casualty of road accidents in Malaysia. All group show many significant differences for the mean for both cases and casualty and just for one group show a less differences which for north and south states for cases.

4.2 Predicting using Simple Linear Regression

The independent variable is different for cases and casualty of road accident. For cases the independent variable is type of road user and for casualty is total cause of road accident. To see whether is there any linear relationship between the variables, we check the correlation for the model.

Table 4 show the result of Pearson Correlation for both variables

	Pearson Correlation	Sig
Total cases and type of road user	-0.659	0.032
Total casualty and total cause of road accident	0.761	0.011

Based on table 4, total cases and type of road user show strong negative linear relationship and for total casualty and total cause of road accident show strong positive linear relationship. Since the value of the significance is less that the p-value (0.05), there is enough evidence to show that the value of the correlation is significant which indicate the dependent and independent variable has a linear relationship.

From here, the prediction equation can be made to predict the value of the cases and casualty of road accident in Malaysia.

Table 5: The coefficients value for total case and type of road user

Coefficients					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	607735.972	72800.495		8.348	.000
Type of road user	-6.027	3.802	-.659	-1.585	.032

Table 6: The coefficient value for the total casualty and total cause of road accident

Coefficients					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-37604.140	16980.309		-2.215	.058
Cause of road accident	9.165	2.765	.761	3.314	.011

Based on table 5 and 6, we can create the predicting equation from the unstandardized coefficients from the table and the equation in (3).

For cases and type of road user (Model 1):

Prediction equation $=> y = -6.027x + 607735.97$ (4)

Where y = total of road accident cases

x = total type of road user

For casualty and cause of road accident (Model 2):

Prediction equation $=> y = 9.165x - 37604.14$ (5)

Where y = total of road accident casualty

x = total causes of road accident

5 Conclusion

From the discussion of this research study, it can be stated that there is no significant difference in rate of cases and casualties of road accidents from 2014 to 2020, but there is a significant difference in states for rate of cases and casualties of road accidents, with the exception of paired category states, which are North states and South states for cases. For the predicted models, for both cases and casualty of road accident. We use 10 years to see the relationship between the variables and from the analysis, we can see that both have strong relationship between their independent variables (type of road user and cause of road accident). The models can be used to predicted the number of cases and casualty for road accident in years.

6 References

[1] Hansen and Benjamin, "Punishment and Deterrence: Evidence from Drunk Driving," *American Economic Review,* 2015.

[2] Musa MF, "The impact of roadway conditions towards accident severity on federal roads in Malaysia," *Plos One,* 2020.

[3] S. Singh , "Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey," *National Highway Traffic Safety Administration,* 2015.

[4] D. Wedagama, "Estimating the Influence of Accident Related Factors on Motorcycle Fatal Accidents using Logistic Regression (Case Study: Denpasar-Bali)," *Civil Engineering Dimension,* pp. 106-112, 2010.

[5] K. Muller, L. Lavange, S. Landesman-Ramey and C. Ramey, "Power Calculation for general linear multivariate models including repeated measures applications," *Journal of the American Statistical Association,* 1992.

[6] "Laerd Statistics," 2018. [Online]. Available: https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php.

[7] K. Kumari and S. Yadav, "Linear Regression Analysis study," *Journal of the Practice of Cardiovascular Sciences,* 2018.

[8] D. J.Rumsey, "dummies," 25 January 2016. [Online]. Available: https://www.dummies.com/education/math/statistics/using-linear-regression-to-predict-an-outcome/.