EXPLORING THE ROLE OF CORRELATIONS FOR ANALYZING THE MALAYSIAN ROAD ACCIDENT PROBLEM

Prof. Madya Dr. Omar Mohd. Rijal*
Chang Yun Fah**
Norliza Mohd. Noor***

*Institute of Mathematical Science, Faculty of Science, University Malaya,
Lembah Pantai, 59000 Kuala Lumpur.
**Faculty Of Engineering, Multimedia University, Cyberjaya, Putrajaya, Kuala Lumpur.
***Electrical Eng. Course, Diploma Program, Univ. Teknologi Malaysia,
Jalan Semarak, 54100 Kuala Lumpur

ABSTRACT

The belief or misconception that correlation may mean causation has seen widespread use of various types of correlations in the analysis of numerical data. The role of correlation, in particular partial correlations, in specific areas of data analysis; approach of problem, interrelationships of variables and modeling will be considered in a particular case study, namely the Malaysian road accident problem.

Introduction

Within the period of 1995 to 1998, considerable attention has been focused on the problem of increasing trends of road accidents and mortality in Malaysia. As an effort to handle the problem, the government of Malaysia has formed the Road Safety Council (RSC) to study and organize activities related to road safety awareness together with the police and Road Transport Department. The work done in this paper is a continuing effort of an earlier consultation problem with the RSC.

There are four reasons to motivate this study. Firstly, the direct initial interpretation of the statistical data does not necessarily reflect the true status of the road accident phenomena. A comparison of two interpretations will be provided in the next section, which we will show to have very different implications on decision-making. Secondly, there exist different approaches in handling the problem, a result of different initial interpretations. Finally the social costs in terms of loss of lives, and financial costs [1], [2] in terms of damages to vehicles and property should help put the problem of road accidents in correct perspective.

The data set

The data used in this study is given in Table 1. Our road accident data was obtained from [1] the traffic division of the Malaysian Royal Police (PDRM). The data is from two sources:

(i) The computer system of PDRM (Traffic/RSC/UPM/TRL) called Microcomputer Accident Analysis Package (MAAP) located in the headquarters of PDRM Traffic Division at Bukit Aman, Kuala Lumpur, and

(ii) The PDRM annual report "Laporan Perangkaan Kemalangan Jalan Raya Malaysia".

The explanatory variables initially selected from the PDRM data set coincides with variables from other studies noted in an earlier literature survey done in [3]. In particular the variables Age (x_1), Race (x_2), Number of accident due to alcohol (x_3), Experience (x_4) and Sex (x_5) were selected.

Types of correlations

Many types of correlations have been defined for a variety of applications. We concentrate on the following four types for our study. Firstly, we define the popular Pearson's sample correlation as;

$$r_{xy} = \frac{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right]}{\sqrt{\left\{\hat{\sigma}_x^{\,2} \cdot \hat{\sigma}_y^{\,2}\right\}}} \qquad \text{(E1)}$$

for a given data set $\left[(x_i, y_i); i = 1,\ldots,n\right]$ such that $\hat{\sigma}_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, and $\hat{\sigma}_y^{\,2}$ similarly defined.

The sample statistic r_{xy} is the estimator of the actual population correlation ρ_{xy}. A useful result is that if we may assume bivariate normality of x and y, then if $\rho_{xy} = 0$ we have

$$T(x,y) = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^{\,2}}} \sim t(n-2) \qquad \text{(E2)}$$

That is $T(x,y)$ has a t-distribution with $n-2$ degrees of freedom. Here n denotes the sample size.

The second type of correlation is called the partial correlation. Let $\underline{x}^T = (x_1, x_2, \ldots, x_p)$ and $\underline{y}^T = (y_1, y_2, \ldots, y_q)$ denote $p+q$ random variables.
Suppose

$$\underline{v}^T = (y_1, y_2, \ldots, y_q, x_1, x_2, \ldots, x_p) \qquad \text{(E3)}$$

has a $p+q$ multivariate normal distribution. Then [4] show that;

$$\text{cov}(\underline{y} \mid \underline{x}) = \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{yx}^T \qquad \text{(E4)}$$

where $\Sigma_x = \text{cov}(\underline{x})$, $\Sigma_{xy} = \text{cov}(\underline{x}, \underline{y})$ and $\Sigma_y = \text{cov}(\underline{y})$. (E5)

For the case $q = 1$ in (E3); and re-label y_1 as y, clearly (E4) gives us $Var(y \mid \underline{x})$. Now we may define the multiple correlations between y and $\underline{x}^T = (x_1, \ldots, x_p)$ as

$$\rho_{y\cdot\underline{x}}^2 = \frac{V(y) - Var(y \mid \underline{x})}{Var(y)} \qquad \text{(E6)}$$

using (E4), the partial correlation between y and x_i given $\underline{x}^* = (x_j \mid j = 1,\ldots,5, j \neq i)$ is

$$\rho_{yx_i|\underline{x}^*} = \frac{\text{cov}(y, x_i \mid \underline{x}^*)}{\{Var(y \mid \underline{x}^*).Var[x_i \mid \underline{x}^*]\}^{1/2}}. \qquad \text{(E7)}$$

The t-test for $H_0 : \rho_{yx_i|\underline{x}^*} = 0$ against $H_1 : \rho_{yx_i|\underline{x}^*} \neq 0$ is carried out using the statistic

$$T = \frac{r_{yx_i|\underline{x}^*} \cdot \sqrt{n-q-2}}{\sqrt{1-r^2_{yx_i|\underline{x}^*}}} \qquad \text{(E8)}$$

which has a t distribution under H_0 with n-q-2 degrees of freedom [5]. Here q is the dimension of $\underline{x}^*$. The critical region of this test is given by

$$|T| \geq t_{n-q-2,\beta} \quad \text{where} \quad \beta = 1 - \frac{\alpha}{2}$$

Finally, canonical correlation investigates the relationship between two linear combinations of variables. Consider $\underline{v}$ as defined in (E3) that is $\underline{v}^T = (\underline{y}^T, \underline{x}^T)$.

Define $\eta = a_1 y_1 + \ldots + a_q y_q = \underline{a}^T \underline{y}$ and $\phi = b_1 x_1 + \ldots + b_p x_p = \underline{b}^T \underline{x}$.

The task here is to find $\underline{a}$ and $\underline{b}$ such that the correlation between η and ϕ is maximum. If $\rho_{\eta\phi}$ is large then either of $y_1, \ldots, y_q$ or $x_1, \ldots, x_p$ may be required as explanatory variables. Details of estimation procedure and relevant tests may be found in [3] and [4].

Simple correlations and one approach to the problem

Looking at Table 1 and Figure 1 clearly show that the number of accidents in 1997 is about seven times the accident figures in 1970. If the opinion that the increasing number of accident (say y) and death is important and cause for concern, a sensible question to ask is what variables is associated or correlated to y. This in fact is a current approach (see [6] and [7]) at understanding the traffic accident problem. Hence if y is found to have strong correlations, with another variable, say number of vehicles (x), hence r_{yx} large would strongly suggest x causes y.

In other words, if the traffic accident problem is stated as "increasing y-values is of concern" than the solution sought after is "find variables $x_1, x_2, \ldots, x_k$ correlated to y" and make subsequent decisions, e.g. increase or decrease x_j based on size of correlations, i.e. r_{yx}. An obvious problem is the choice of k, namely the number of explanatory variables.

Partial and multiple correlations for a second approach

A re-look at Table 1 will show that approximately 2 percent (number of accident/number of registered vehicle) of vehicles are involved in road accidents every year. Also, about 0.1% (number of death/number of registered vehicle) of the victims died since 1970. These consistent rates suggest that only a small number of road users (motorist) contributed to the road accident figures. If these consistent rates are regarded as meaningful, then the focus should be on the drivers. This helps in reducing greatly the number of the explanatory variables that need be considered.

The initial selection (see 'The data set' section) of variables led to the choice of Age (x_1), Race (x_2), Number of accident due to alcohol (x_3), Experience (x_4) and Sex (x_5). The following procedure was used to investigate if all x_1, x_2, x_3, x_4 and x_5 were needed:

(i) Choose $x_t, t = 1,\ldots,5$ for ρ_{Yx_t} maximum.

(ii) Choose $x_u, u \neq t, u = 1,\ldots,5$ for $\rho_{Yx_u|x_t}$ maximum.

(iii) Choose $x_v, v \neq t, v \neq u, v = 1,\ldots,5$ for $\rho_{Yx_v|x_t x_u}$ maximum.

(iv) Repeat process until the next partial correlation is 'small' enough.

(v) Finally calculate the multiple correlation $\rho_{Y\underline{w}}$ where $\underline{w} = (x_t, x_u, x_v, \ldots)^T$.

(E9)

In our study, the criterion for normality of data, and hence used of (E4) was strictly adhered to, see [3]. The final result showed in [3] only variable x_3 is significant (using the test involving (E8)). We note the correlation between Y and x_3 is 0.508.

Comparison Between Approaches; Canonical Correlations

The relatively low values of ρ_{Yx_3} suggest an investigation on x_3 which may be related to other explanatory variables (derived from another approach). In particular [6] and [7] studied variables such as Locations ($A1$), Type of vehicles ($A2$), Light/Visibility conditions ($A3$) and Number of Registered Vehicles ($A4$).

The study in [3] showed that the two linear combination;

$$\eta_j = x_3$$

and $\phi_j = 0.0016A1 + 0.9995A2 + 0.0199A3 + 0.0127A4$

are **independent**. This result strengthened the belief that x_3 is a significant variable that may be used to explain the variability of y.

Some Other Methods In Retrospect

The correlation study clearly showed that the 'choice' of explanatory variables is crucial to the results of the study. In similar fashion, we look briefly at two other popular methods of variable selection, namely graphical method and regression.

Graphs are often used to display data. Graphical displays visually exhibit the variation of, say, mortality rates (y^*) with respect to Age (x_1^*). Further a (y^*, x_1^*) plot for males, and another for females show variation of y^* with respect to Age **and** Sex. Suppose x_3^* denote Race, Colton and Buxbaum [8] provides 4 plots showing the variations of y^* with respect to x_1^*, x_2^*, x_3^* (x_2^* denotes Sex). These four plots showing variations of y^* may be regarded as an accident model. In general when the number of explanatory variables is large, the number of possible plots is also large making visual interpretations difficult if not misleading. One possible solution is to use the procedure given in (E9) before deciding on which plots to study.

Regression models are a natural extension from the use of correlations. Radin Umar [7] uses log-linear models. Again an appropriate application of (E9) may help obtain simple models.

Conclusion and Further Remarks

Important decisions are made by interpreting correlations. The particular case study has shown the 'direct' use of simple correlations is unwise. Since the number of explanatory variables is usually large, important variable relationships cannot be ignored, and partial correlations with multiple correlations offer one method to account for such interrelationship.

Further correlations may not be correctly used for at least two reasons. Firstly, the assumption of normality is not checked. Secondly, appropriate hypothesis testing are often not done or reported.

In general the idea, methods and application of correlations need to be handle with care and 'deeper' understanding of the data. Causation should be defined (see for example Cox [9]) and not assumed. If definitions of causation are difficult or not obvious we recommend that a problem involving the analysis of numerical data be studied from more than one perspective. Even if causation is implied, mathematical models such as regression models should apply a correlation study such as that in (E9) before performing the usual process of estimation and predictions.

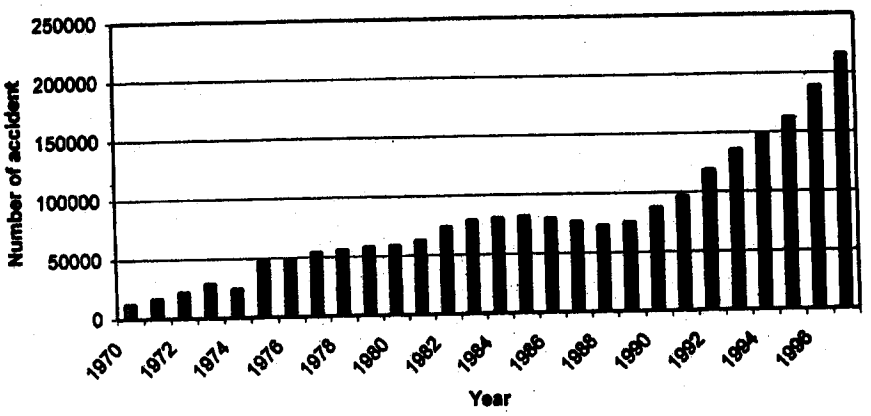

Figure 1: Number of road accident in Malaysia from 1970 to 1997.

Table 1: General Road Accident Statistics in Malaysia (Source: PDRM, 1998).

Year	Population	Vehicles Registered	No. Of Accidents	Accident Rate	No. Of Death	Death Rate	Total Casualties	Casualties Rate
1970	9000399	669294	12704	0.01898	579	0.00087	7621	0.01139
1971	9133506	730035	16847	0.02308	1548	0.00212	8481	0.01162
1972	9873623	802831	22151	0.02759	1712	0.00213	10716	0.01335
1973	10130672	939951	29286	0.03116	1922	0.00204	16602	0.01766
1974	10434592	1090279	24581	0.02255	2303	0.00211	13332	0.01223
1975	10438137	1267119	48233	0.03807	2317	0.00183	19440	0.01534
1976	10472544	1429845	48291	0.03377	2405	0.00168	19327	0.01352
1977	10716642	1621271	54222	0.03344	2512	0.00155	20305	0.01252
1978	10944500	1829958	56021	0.03061	2561	0.00140	21659	0.01184
1979	11188630	1989391	57931	0.02912	2607	0.00131	22611	0.01137
1980	11442086	2357386	59084	0.02506	2568	0.00109	22404	0.00950
1981	14128354	2901182	63192	0.02178	2769	0.00095	22303	0.00769
1982	14506589	3246790	74096	0.02282	3266	0.00101	22820	0.00703
1983	14886729	3594943	79150	0.02202	3550	0.00099	26557	0.00739
1984	15437683	3941036	80526	0.02043	3637	0.00092	25552	0.00648
1985	15866592	4243142	82059	0.01934	3603	0.00085	23924	0.00564
1986	16278001	3523674	79804	0.02265	3525	0.00100	23257	0.00660
1987	16527973	3674482	76882	0.02092	3320	0.00090	21799	0.00593
1988	16921300	3865711	73250	0.01895	3335	0.00086	22538	0.00583
1989	17376800	4155197	75626	0.01820	3773	0.00091	30037	0.00723
1990	17812000	4547417	87999	0.01935	4048	0.00089	29814	0.00656
1991	18178100	4942040	96513	0.01953	4331	0.00088	30107	0.00609
1992	18606000	5259836	118554	0.02254	4557	0.00087	36262	0.00689
1993	19050000	5656037	135995	0.02404	4666	0.00082	41686	0.00737
1994	19494000	6166432	148801	0.02413	5159	0.00084	48503	0.00787
1995	20096700	6802375	162491	0.02389	5712	0.00084	52152	0.00767
1996	21169000	7686684	189107	0.02460	6304	0.00082	53475	0.00696
1997	21665600	8550469	215632	0.02522	6302	0.00074	56574	0.00662
1998	22179500	9141357	211037	0.02309	5740	0.00063	55704	0.00609

REFERENCES:

[1] Polis Diraja Malaysia, Laporan Tahunan Cawangan Trafik Bukit Aman, Kuala Lumpur, PDRM 1995-1998.

[2] Dawson, R.F.F., "Cost of Road Accidents in Great Britain", Crowthorne: Road Research Laboratory, Ministry of Transport, RRL Report LR79, 1967.

[3] Chang, Y. F., "A Statistical analysis of the Malaysian Road Accident Problem", An M. Sc. Thesis, University of Malaya, 2001.

[4] Anderson, T. W., " An Introduction To Multivariate Analysis", J. Wiley & Sons, New York, 1958.

[5] Kleinbaum, D. G., Kupper, L.L. and Muller, K. E., "Applied Regression Analysis And Other Multivariable Methods", Boston: PWS-Kent Publishing Co., 1998.

[6] Radin Umar, R. S., MacKay, G. M. and Hills, B. L., "Preliminary Analysis of Exclusive Motorcycle Lanes Along The Federal Highway F02, Shah Alam, Malaysia, Journal of IATSS Research, Vol. 19, no.2, Japan, pp. 93-98, 1995.

[7] Radin Umar, R. S., "Model Kematian Dan Kecederaan di Malaysia Unjuran Tahun 2000", Kertas Dasar Keselamatan Jalan Raya, Kementerian Pengangkutan, Universiti Putra Malaysia, Unit Kajian Kemalangan Jalan Raya, 1996.

[8] Colton, T. Buxbaum, R.C., "Motor Vehcle Inspection And Motor Vehicle Accident Mortality", American Journal Of Public Health, pp. 1090-1099, 1968.

[9] Holland, P.W., "Statistics and Causal Inference (with discussion)", Journal American Statistical Association, 81, pp. 945-970, 1986.

[10] Cox, D. R., "Causality: Some Statistical Aspects", Journal of Royal Statistical Society, A, 155, part 2, pp. 291-301, 1992.