Modeling the number of car theft using Poisson regression

Malina Zulkifli, Agnes Beh Yen Ling, Maznah Mat Kasim, and Noriszura Ismail

Citation: AIP Conference Proceedings **1782**, 050018 (2016);
View online: https://doi.org/10.1063/1.4966108
View Table of Contents: http://aip.scitation.org/toc/apc/1782/1
Published by the American Institute of Physics

Modeling the Number of Car Theft Using Poisson Regression

Malina Zulkifli[1, a)], Agnes Beh Yen Ling[2, b)], Maznah Mat Kasim[3, b)] and Noriszura Ismail[4,d)]

[1, 2, 3]School of Quantitative Sciences, College of Arts and Science, Universiti Utara Malaysia, 06010 Sintok, Kedah, MALAYSIA
[4]School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, MALAYSIA

[a)]Corresponding author: malina@uum.edu.my
[b)]agnesbeh92@gmail.com
[c)]maznah@uum.edu.my
[d)]ni@ukm.edu.my

Abstract. Regression analysis is the most popular statistical methods used to express the relationship between the variables of response with the covariates. The aim of this paper is to evaluate the factors that influence the number of car theft using Poisson regression model. This paper will focus on the number of car thefts that occurred in districts in Peninsular Malaysia. There are two groups of factor that have been considered, namely district descriptive factors and socio and demographic factors. The result of the study showed that Bumiputera composition, Chinese composition, Other ethnic composition, foreign migration, number of residence with the age between 25 to 64, number of employed person and number of unemployed person are the most influence factors that affect the car theft cases. These information are very useful for the law enforcement department, insurance company and car owners in order to reduce and limiting the car theft cases in Peninsular Malaysia.

INTRODUCTION

Crime can be known as an act that break the law and result in punishment. Crime is the act that seen to be against the society but not only just to a specific person so that it also known as a type of harmful behavior to the society. In Malaysia, Royal Malaysian Police (RMP) uses a phrase of "Index Crime" to quantify crime. The RMP divided the crime into two major categories which is violent crime and property crime. Each type of crime comes with different sociological, demographics profile and phenomena.

Violent crime is defined as offences which involve threat of force while property crime defined as taking others property with no force against the victims. Violent crime are such as rape, kidnapping, murder, and voluntarily that causing hurt, while property crime are such as burglary, vehicle theft, housebreaking and theft and other kind of theft. Therefore, violent crime is more dangerous than property crime because the percentage of violent crime will causing death is higher.

The 4th International Conference on Quantitative Sciences and Its Applications (ICOQSIA 2016)
AIP Conf. Proc. 1782, 050018-1–050018-8; doi: 10.1063/1.4966108
Published by AIP Publishing. 978-0-7354-1444-0/$30.00

In Malaysia, violent crime attract the most attention of both media and public. Whereas, property crimes take for about ninety percent from all crimes that reported during the thirty years of 1980-2009 [1]. According to Sidhu [2], property crime had shown a slightly unpredictable growth pattern and it can be seen to be the main donor of the total index crime in our country.

Comparing to other country, property crimes is the most common committed crime that happen in the United States. There are two leading data collection agencies that collect and publish the data of property crime in United States which are the Uniform Crime Reports (UCR) of the Federal Bureau of Investigation (FBI) [3] and the National Crime Victimization Survey (NCVS) of the Bureau of Justice Statistics (BJS). The Table 1 and Figure 1 show the property crime rate in Unites State from 2000 to 2014. From the property crime rate graph that collected by FBI we can see that the trend of property crime rate decreased in the fourteen years from year 2000 to 2014. The decreases of the trend is very small which not significant enough to reduce the anxiety among the residents.

Table 1. Property crime rate in United State from 2000 to 2014

Year	Population	Property Crime	Property Crime Rate/100000
2000	281,421,906	10,182,586	3,618.3
2001	285,317,559	10,437,480	3,658.1
2002	287,973,924	10,455,277	3,630.6
2003	290,690,788	10,442,862	3,591.20
2004	293,656,842	10,319,386	3,514.1
2005	296,507,061	10,174,754	3,431.5
2006	299,398,484	9,983,568	3,334.5
2007	301,621,157	9,843,481	3,263.5
2008	304,374,846	9,767,915	3,211.5
2009	307,006,550	9,337,060	3,036.1
2010	309,330,219	9,112,625	2,945.90
2011	311,587,816	9,052,743	2,905.40
2012	313,873,685	9,001,992	2,868.00
2013	316,497,531	8,650,761	2,733.30
2014	318,857,056	8,277,829	2,596.10

Source : FBI UCS Annual Crime Reports

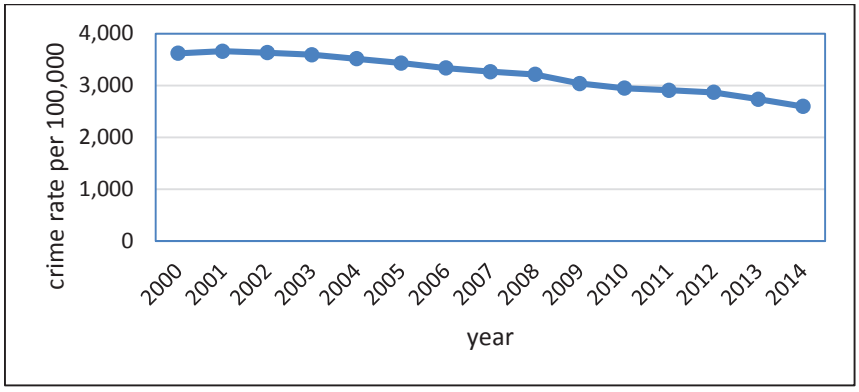

Figure 1. The property crime offense in twenty years trend of United State from 1991 to 2010

In Malaysia, property crime is also known as a serious crime that is common to happen. According to Sidhu [2], violent crimes is only 10% of the crimes reported in year and property crimes is known as the majority crimes as it is accounted for 90%. The vehicle theft is composing one half of the property crimes.

There are many studies done on property crime. From the studies, we know that the crime rates tend to increase due to the gap between the rich and the poor [4]. Besides that, we also know that the pattern of crime are affect by the demographic area and the environment [5].The most developed area is believed to have the higher crime rate compared to the undeveloped area. For example, the urban areas such as large cities or town have the higher crime rate compared to rural areas such as countryside. This situation happened because of several factors such as environmental characteristic, economics, social, political and demographics [6][7][8].

Specifically, there are not much scientific research has been done in identifying the factors that influence the car theft crime in Peninsular Malaysia. Most studies conducted previously focused on the use of qualitative methods to identify the factor that affecting the property crime. Based on the literature, most studies involving count data using the Poisson distribution. Therefore, this study fit the Poisson regression model to claim count data and covariates involved to investigate the effect of covariates on the car theft crime.

The analysis in quantitative method let researcher determine the important facts from the research data such as the preference trends, the difference between the groups and the demographics. The quantitative approach has two significant advantages. First, it can be administered and evaluated quickly. Second, numerical data obtained through this approach facilitates comparisons between organizations or groups, as well as allowing determination of the extent of agreement or disagreement between respondents [9]. The data of quantitative method can be collected and analyzed rapidly. By using the real data, the result of the studies will be more accurate, specific and significant by using the quantitative method.

For a vehicle theft data also known as a count data provides the number of exposures and the number of car theft with related covariates, a Poisson regression model can be applied to predict the estimates of the regression parameters which can be used as possible indicators of a crime index. Based on the actuarial and insurance literatures, Poisson regression model has been widely used for modeling claim count data, and such examples can be found in Aitkin et. al [10] and Renshaw [11], who fitted the Poisson to two different sets of U.K. motor claim count data. For insurance researchers and practitioners, the Poisson regression model has been considered as practical and convenient as the model allows statistical inferences and hypothesis tests to be determined by statistical theories. Besides that, the Poisson regression has been used in many fields such as in medical, biological sciences, general health and social sciences. The application of Poisson regression in the study of crime still not been fully explored in Malaysia.

The objective of this paper is to model the number of car theft that occurred by districts in Peninsular Malaysia using the Poisson regression model and to evaluate the significant covariates of the car theft crime. In this study, the Poisson regression model is used to fit the car theft count data in Peninsular Malaysia. Several covariates obtain from literature and expert opinions are considered to determine the most influence factor that contribute to the car theft crime.

METHODOLOGY

Data

This study used a secondary data of the number of private car theft obtained from the Insurance Services Malaysia (ISM)[12] for four years period and the information of socio and demographic provided by and the Statistics Department of Malaysia [13]. The covariates considered in the study are obtained from the literature review and also from the expert opinions. The variables that considered in this paper can be categorized into two groups, namely district descriptive factors and socio and demographic factors are shown in Table 2.

Table 2. District descriptive factors

Factor	Classes
Density	No. of population divide by width area per district
	Low
	Moderate
	High
Ethnic	No. of ethnic per district.
Bumiputera	Low
Chinese	Moderate
Indian	High
Others	
Migrant	No. of migrant per district
	Low
	Moderate
	High
Employment	No. of person work per district
	Low
	Moderate
	High
Unemployment	No. of person do not work per district
	Low
	Moderate
	High
Category of age	No. of person age 25-64 per district
	Low
	Moderate
	High

Poisson Regression Model

Basically, let the random variable Y_i is the number occurrences that has a Poisson distribution, then the density function is,

$$P\left(Y_i = y_i\right) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{\mathrm{y}_i!}, \qquad y_i = 0,1,2,.... \tag{1}$$

with parameter λ_i is the mean of the Y_i and the variance of the random variable Y_i is $V\left(Y_i\right) = E\left(Y_i\right) = \lambda_i$.

In the Poisson regression model, the response variable, Y_i represents the number of car theft of *i*-th district. Therefore, the effect of the explanatory variable x_i on the response variable Y_i can be written in the form of multiplication regression model,

$$E\left(Y_i \mid \mathbf{x}_i\right) = \lambda_i = e_i \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right) \tag{2}$$

with e_i represents the number of exposure which is the policyholders, $\mathbf{x}_i$ is a vector of variables that sized p x 1 , and $\boldsymbol{\beta}$ is a vector of regression parameter that sized p x 1. In this study, λ_i is a conditional mean of the response variable, Y_i .

The main concerns here is the conditional mean of (2) is not a linear function of $\boldsymbol{\beta}$. Therefore, a non-linear transformation is necessary to generalize the conditional mean by taking the log of both parts and found that

$$\log_e\left(\lambda_i\right) = \log\left(e_i\right) + \mathbf{x}_i^T\boldsymbol{\beta} \tag{3}$$

where $\log_e\left(\lambda_i\right)$ is known as canonical function or the link function. It's named as that because $\log_e\left(\lambda_i\right)$ is the original parameter for the Poisson distribution when it expressed in exponential form [14]. $\log\left(e_i\right)$ is known as offset. Offset means that we assumed the number of car theft is proportionate to the number of exposures or policyholders. The log likelihood function of the Poisson regression model is given by

$$\log L\left(\boldsymbol{\beta}\right) = \sum_i y_i \log\left(\lambda_i\right) - \lambda_i - \log\left(y_i!\right) \tag{4}$$

The estimates of the parameters can be obtained by maximizing the $L\left(\boldsymbol{\beta}\right)$ on the $\boldsymbol{\beta}$ using a maximum likelihood method

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{y_i - \lambda_i}{\lambda_i}\frac{\partial \lambda_i}{\partial \beta_j} = 0 \qquad \text{for } j = 1,\ldots,p \tag{5}$$

where p is the number of regression parameters. The probability of likelihood is equal to the Weighted Least Squares Iteration procedure [15].

In the Poisson regression model, dispersion, α often happen. This dispersion can be over-dispersion, underdispersion or no-dispersion. Dispersion is obtained by dividing statistics deviance to its degrees of freedom. If no-dispersion, then α=1. If α>1 or α<1, then there will be existence of over-dispersion or under-dispersion in a Poisson regression model.

Generally, the Poisson regression model for the main effect in this study can be written as

$$\begin{aligned} Log\left(\text{number of car theft}\right) = \beta_0 &+ \log\left(\text{exposure}\right) \\ &+ \beta_1\left(\text{density\_moderate}\right) + \beta_2\left(\text{density\_high}\right) \\ &+ \beta_3\left(\text{Bumiputera\_moderate}\right) + \beta_4\left(\text{Bumiputera\_high}\right) \\ &+ \beta_5\left(\text{Chinese\_moderate}\right) + \beta_6\left(\text{chinese\_high}\right) \\ &+ \beta_7\left(\text{Indian\_moderate}\right) + \beta_8\left(\text{Indian\_high}\right) \\ &+ \beta_9\left(\text{Others\_moderate}\right) + \beta_{10}\left(\text{Others\_high}\right) \\ &+ \beta_{11}\left(\text{Migrant\_moderate}\right) + \beta_{12}\left(\text{Migrant\_high}\right) \\ &+ \beta_{13}\left(\text{Employment\_moderate}\right) + \beta_{14}\left(\text{Employment\_high}\right) \\ &+ \beta_{15}\left(\text{Unemploy\_moderate}\right) + \beta_{16}\left(\text{Unemploy\_high}\right) \\ &+ \beta_{17}\left(\text{age\_moderate}\right) + \beta_{18}\left(\text{age\_high}\right) \end{aligned}$$

RESULT AND DISCUSSION

The analysis of independent variables (explanatory variables) for each district is used to determine the factors that affected the number of car theft in Peninsular Malaysia. In this study, the Poisson regression is used to model the relationship between the independent variables with the number of car theft in the district.

Table 3 provides more detailed information about the test for each covariate considered. The parameters, log likelihood, AIC and BIC for the fitted regression model with covariates that are significant at 5% level of significance are shown in Table 3. Based on the *p*-value, the results indicate that population density and the number of houses in a district even moderate or high has no significant effect on the car thefts in the district. Based on ethnic, if there are moderate numbers of Indian in a district there are also no any significant effect on car theft. But if there are high number of Indian in the district, it tends to have more cases of car theft in the area. However, the moderate and high Bumiputera composition, Chinese composition, Others ethnic composition, foreign migration, number of residence with the age between 25 years to 64 years, number of employed person and number of unemployed person play an important role in affecting the number of car theft in any district.

Table 3. Estimate parameters for Poisson regression model

Parameter		Est.	t-ratio	*p*-value	
Intercept		-2.49	-15.38	0.00	
Bumiputera:	Moderate	-0.50	-2.44	0.01	*
	High	0.62	2.72	0.00	*
Chinese:	Moderate	-0.76	-3.83	0.00	*
	High	-1.46	-5.30	0.00	*
India:	Moderate	0.01	0.04	0.49	
	High	0.96	5.00	0.00	*
Others:	Moderate	-0.67	-3.96	0.00	*
	High	-0.86	-4.72	0.00	*
Migrant:	Moderate	0.66	7.89	0.00	*
	High	0.53	13.17	0.00	*
No. of houses:	Moderate	0.19	0.78	0.22	
	High	0.12	0.36	0.36	
Density:	Moderate	0.14	0.97	0.17	
	High	0.08	0.44	0.33	
Adult (age 25-64):	Moderate	-1.22	-4.30	0.00	*
	High	-0.85	-1.99	0.02	*
Employed:	Moderate	0.77	5.70	0.00	*
	High	1.59	5.99	0.00	*
Unemployed:	Moderate	-1.34	-8.89	0.00	*
	High	-2.17	-8.02	0.00	*
Log likelihood			-443.9737		
AIC			929.9474		
BIC			980.2309		

* significant at 5% of significance level.

Table 4 shows the value of the parameter estimate, log likelihood, AIC and BIC for the Poisson regression model with significant covariates obtained from Table 3 at 5% significance level.

Table 4. Parameter estimate, log likelihood, AIC and BIC for the Poisson regression model with significant covariates

Parameter		Est.	t-ratio	*p*-value	
Intercept		-2.49	-15.38	0.00	
Bumiputera:	Moderate	-0.50	-2.44	0.01	*
	High	0.62	2.72	0.00	*
Chinese:	Moderate	-0.76	-3.83	0.00	*
	High	-1.46	-5.30	0.00	*
India:	High	0.96	5.00	0.00	*
Others:	Moderate	-0.67	-3.96	0.00	*
	High	-0.86	-4.72	0.00	*
Migrant:	Moderate	0.66	7.89	0.00	*
	High	0.53	13.17	0.00	*
Adult (age 25-64):	Moderate	-1.22	-4.30	0.00	*
	High	-0.85	-1.99	0.02	*
Employed:	Moderate	0.77	5.70	0.00	*
	High	1.59	5.99	0.00	*
Unemployed:	Moderate	-1.34	-8.89	0.00	*
	High	-2.17	-8.02	0.00	*
Log likelihood			-443.9737		
AIC			929.9474		
BIC			980.2309		

* significant at 5% of significance level.

Based on Table 4, the estimated value shown in this table reflect the magnitude of the effect of covariates on the car theft cases. The higher value shows the covariate had a great influence on the car theft, and vice versa. For example, the estimate value of the covariates of high composition of employed people in the district is 1.59 can be said that the area with high composition of employed people is a major factor contributing to the car theft in the area of Peninsular Malaysia.

In addition, the details of the relationship for each covariate on the response variables can also be obtained based on the indicators provided in the estimated value of the regression coefficient. For example, the covariates migrant, all of the estimated value of the regression coefficient is positive. This indicates that there is a direct relationship between covariate migrant and the car theft which means the number of car theft increase if the number of migrant increase. For the covariate unemployed, both estimated value gives a negative sign which it indicates that the number of car theft reduced when a district has more unemployed people.

CONCLUSION

This paper has applied a Poisson regression model to the car theft count data. The number of car theft in Peninsular Malaysia is affected by several factors. The most influence factors that contribute to the car theft crime in a district are high composition of Indian ethnic, moderate composition of foreign migrant, high number of residence with the age between 25 years to 64 years and a high number of employed person.

The study of the relationship through fitting Poisson regression model between the area and the risk or even more easily referred to as a socio and demographic factors can give information to the local administrators and the owner of the vehicle about the factors that should be considered in addressing the problem of car theft. By increasing the level of awareness and concern about the factors that affect the theft of the vehicle can help reduce the risk of the vehicle theft. For academics, by fitting Poisson regression model led to understand more complex models such as over-dispersion model of Negative Binomial regression model and the Generalized Poisson regression model.

This paper focus on the district descriptive factor and a few factors in socio and demographic characteristics. Therefore, the other characteristics are recommend to be included in the analysis of the car theft for the next

research. For example, may be the researcher can include the factors that involve the car characteristics and also the socio economics factors. So that, the analysis will be more accurate when there are more covariates are considered in the analysis. In addition, this study can be extended by focusing on the car theft crime in the district in east Malaysia; Sabah and Sarawak. Maybe there is a difference in terms of the factors that contribute to the car theft crime based on terrain and different demographics in Sabah and Sarawak.

ACKNOWLEDGMENTS

This research is financed by the Fundamental Research Grant Scheme (Code: FRGS/2/2013/SS08/UUM/03/1).

REFERENCES

1. Z. Malina, I. Noriszura, & R. Ahmad Mahir, Applied Mathematics & Information Sciences, 7, No. 2L, 389-395 (2013).
2. A. S. Sidhu, Journal of the Kuala Lumpur Royal Malaysia Police College(No. 4), 25, 1-28 (2005)
3. FBI UCS Annual Crime Reports, *Crime in the U.S 2014* (UCR Publication, FBI, 2014)
4. A. A. J. Baker, Jornal of Population Research, 95 (2012).
5. N. James, *How Crime in the United State is Measure* (CRS Report for Congress, United State, 2008).
6. S. Perreault, J. Savoie and F. Bédard, Crime and Justice Research, no. 85-561-M — No. 011 (2008)
7. J. Savoie, F. Bédard and K. Collins, Crime and Justice Research Paper Series, Statistics Canada Catalogue no. 85-561-MIE. No. 7. Ottawa. (2006)
8. J. A. Gyamfi, Journal of Criminal Justice. 30, no. 3 (2002)
9. C. A. Yauch and H. J. Steudel, Organizational Research Methods , Vol. 6 No. 4, 465-481 (2003)
10. M. Aitkin, D. Anderson, B. Francis, & J. Hinde, *Statistical modelling in GLIM* (New York: Oxford University Press, 1990)
11. A. E. Renshaw, ASTIN Bulletin, 24(2): 265–285 (1994)
12. *Industry Insight: Private Car Theft* (Insurans Services Malaysia, 2013).
13. *Banci Penduduk dan Perumahan Malaysia: Taburan Penduduk dan Ciri-ciri Asas Demografi* (Jabatan Perangkaan Malaysia, 2010).
14. G. H. Dunteman, & M. R. Ho, *An introduction to generalized linear models* (Thousand Oaks, Calif.: Sage Publications, 2006).
15. J. A. Nelder and R. W. M. Wedderburn, Journal of the Royal Statistical Society, Series A (General), Vol. 135, No. 3 (1972)