* **CRIME RATE PREDICTION USING MULTIPLE LINEAR REGRESSION**
* Sean Andrei H. Marasigan
* BSCS – III
* Manuel S. Enverga University Foundation
* sh.marasigan @mseuf.edu.ph

# INTRODUCTION

Crime is an unlawful and punishable act committed by the guilty which endangers public safety at different levels [1]. Crime is an act performed or omitted in breach of the legislation banning or prescribing it and for which a prosecution is enacted on judgment [2]. In the Philippines, the Revised Penal Code (Republic Act. No. 3815) is a major statute that characterizes and punishes criminal offences.

Crime can be affected by a number of factors, predictors, or variables. With this, prediction of crime rate can be within one’s reach. In addition of numerous technological advances in data analysis, predicting the very elusive concept of crime can be possible. In this paper, factors such as age and police funding are taken into account in predicting crime rate and violent crime rate.

Using multiple linear regression, an estimation of crime rate and violent crime rate will be predicted or determined. Moreover, the relationship of crime rate and various predictors will be identified through a regression model.

# OBJECTIVES OF THE STUDY

The objective of the study is to predict the crime rate and violent crime rate and how does it change using predictors such as annual police funding, percentage of people in various age groups, and their highest educational attainment. These said factors or predictors will help in determining the crime rate and violent crime rate. This is done by first analyzing the data and cleaning up the data for plotting, then interpreting the results and conclusion.

# DATA

A picture containing calendar

Description automatically generated

The dataset consists of seven (7) columns and 50 rows. As you can see, the column names for the given dataset are not yet defined or renamed. The descriptions for each column are listed below:

**X1 –** total overall reported crime rate per 1 million residents

**X2 –** reported violent crime rate per 100,000 residents

**X3 –** annual police funding in $/resident

**X4 –** percentage of people 25 years+ with 4 yrs. of high school

**X5 –** percentage of people 16- to 19-year-olds not in high school and not high school graduates.

**X6 –** percentage of people18 to 24-year-olds in college

**X7 –** percentage of people 25 years+ with at least 4 years of college.

The dataset can be accessed through here: <https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr06.html>

# ALGORITHM – *PSEUDOCODE*, MODEL



These are the libraries needed for this paper.



Importing the dataset and viewing the dataset.

Scatter chart

Description automatically generated

Checking for the summary of the dataset is practical in determining the central tendencies of a numerical dataset.

A picture containing scatter chart

Description automatically generated

Next, it has been observed that the columns in the dataset are not identified. Renaming each of the column is practical for ease and to minimize human error.

Text

Description automatically generated

After executing the above code, checking the dataset again is a must.

Scatter chart

Description automatically generated

After preparing the dataset, a regression model is now created or built. The estimation of crime rate and violent crime rate will now be predicted. The factors that will be used are police funding, various age groups, and their highest educational attainment.



A picture containing text, receipt, screenshot

Description automatically generated

As you might notice already, looking at the model there is only one (1) predictor that greatly affects the prediction, it is the police funding. This is through paying attention to the the p-value of the coefficients. In simple terms, a p-value indicates whether or not you can reject or accept a hypothesis. The hypothesis, in this case, is that the predictor is not meaningful for your model. A standard way to test if the predictors are not meaningful is looking if the p-values smaller than 0.05. This is apparent in the p-value of the police\_fund column which it has a p-value of 0.000884

The formula can now be written like this:

# RESULTS AND DISCUSSION

The summary of the model shows the metrics that are used to check how well the model fits the dataset such as residual standard error (RSE), r-squared (R2), and F-statistic. As you can see the F-statistic model and r-squared value is low, but the p-value indicates that at least one (1) of the predictors in the model is significant in predicting the overall crime rate. Based on the r-squared value, the model explains only 25.79% of data variability.

Text

Description automatically generated with medium confidence

Although the model is not a good fit, the prediction error rate of the model is low. This indicates that the predicted results is somehow accurate.



Now, we plot our model using scatter plot and drawing a regression line.



Chart, scatter chart

Description automatically generated

Based on the regression plot above, it can be observed that slope of the regression is low, and the points are scattered erratically but there is a slight hint of cluster in the bottom left of the graph near the origin.

# CONCLUSION

Based on the dataset and model created above, crime rate and violent crime rate is heavily affected by the amount of funding the police or authorities have. Factors such as age and education attainment contributed less or not as much significant. In addition, the number of entries in the dataset is still insufficient to make a proper prediction because the dataset consists of only 50 rows which can another cause for inaccuracy in the model.

# REFERENCES

[1] S. N. Bezugly, G. Y. Lesnikov, A. S. Dugenets, Y. I. Ivanenko, and Y. Y. Tischenko, “Comparative and Legal Study of the Concept of Crime under Legislation of CIS Countries,” 2019.

[2] R. Sandy, H. Baculinao, and R. F. Ceballos, “An Analysis on the Location and Type of Index Crimes in the Philippines,” Oct. 2019.