Prediction of Crime

Gustavo L. Serrano Malavé
Master of Engineering in Computer Engineering
Dr. Nelliud Torres
Electrical and Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract *— Crime has always been a controversial topic amongst society. The prediction of criminal events could help law enforcement units to be more alert in the areas, reducing or preventing the events from ever happening. Even in cases when the crime is committed, law enforcement would be already deployed in the area reducing the chances of the criminal getting away. Historical data was retrieved from the databases of the government of Puerto Rico, where only 3 years of data were available. A fit model regression considering categorical and ordinal variables resulted in an equation that validated our data with a degree of error. Although, the equation estimates the location on the next crime, its accuracy could be improved with more data and analysis of the factors that influence crime. The program is not intended to determine the location and time with 100% accuracy, but to give a good estimate of these for a crime.*

Key Terms *— Crime Prediction, Crime Statistics, Government Open Data, Regression Models.*

INTRODUCTION

Crime will always exist but if we can predict them, we can reduce them by having the right people at the right place at the right time. Predicting crimes will change society drastically in the best of ways. Innocent people's suffering will be reduced and happiness should increase. With these predictions perhaps we can understand patterns in when and where crimes are chosen to be committed.

METHODOLOGY

In hopes of predicting crime accurately, 2 years of historical data were used to generate an equation. The third and current year was used as a validation resource, in order to verify the accuracy of the equation generated. The data is comprised of the crimes reported to the Police of Puerto Rico, retrieved from Puerto Rico Government Open Data Portal. The data collected consisted of the following:

- Date of the crime (day/month/year)
- Time of crime
- Type of Crime
- Location (X and Y coordinates)
- Precinct

The date was modified on the local data in order to see a trend and predict a pattern. The year was removed from the date to identify trends by day of the year. As an example consider the following:

Every year on February 14th, the Closed Door Hotel is booked solid. If we were to plot historical data of the previous 4 years to look at a trend on February 14th, we wouldn't be able to see it clearly as each date would contain the year, making each date a unique value per year.

Figure 1
Identifying a Trend without Excluding the Year

As we can clearly see from the example the exclusion of the year on the date can influence how the data is seen and analyzed.

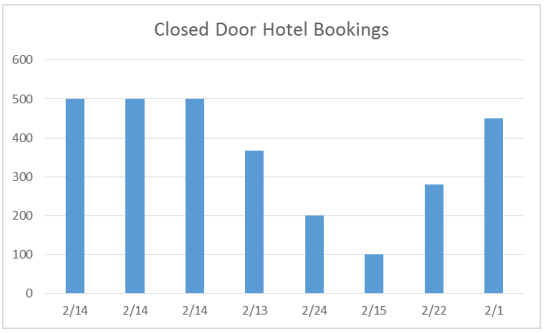

Figure 2
Identifying a Trend Excluding the Year

The time of day was also manipulated into a 24 hour format to remove any mix-up between am/pm times. This change in format also avoids implementing an additional variable for the am/pm category.

The precinct was modified as a categorical variable into a discrete numerical value because categorical values pose limitations. For example, if the data contains too many categories, several categories would need to be combined into one. In addition, there may be a need to use a data mining techniques that require numeric data rather than categorical data [1].

Additionally the X and Y coordinates were used as the response variable, also known as Y in regression models.

The software used to aid the manipulation and analysis of the data were the following:

- Arena – Input Analyzer
- MathCad
- Minitab
- Excel

Arena is a simulation tool but it also has the capability of recognizing distributions in a set of data with Input Analyzer. This tool served as an initial stage in finding out the distributions per variable. Having the distribution of the data, the expected values of the variables and their probabilities could be computed.

MathCad was used as calculation sheet because of its capabilities as a computer algebra system and its simultaneous use of engineering applications as integrals and probabilities. During the early stages of the data manipulation, integrals were used to estimate the probability of the crime, date and time occurring. The problem with calculating the probabilities was that it didn't consider the correlation between the variables and therefore would produce inaccurate results.

Although Excel and Minitab are both spreadsheet tools, these were used differently based on their features and friendliness. Minitab computes regressions and can chart advanced graphics when compared to excel. Both tools were used equally but generally Excel was friendlier when manipulating the data and plotting it.

Statistical methods are required to ensure that data are interpreted correctly and that apparent relationships are meaningful and not simply a chance occurrences [2].

A regression model has an output equation which is used to describe the statistical relationship between one or more predictors (X's) and the response variable (Y) to predict new observations. A multiple linear regression examines the linear relationships between the response and two or more predictors.

If there is a large amount of variables affecting your response, before fitting a regression model with all the predictors, it is recommended to use a stepwise technique to screen out predictors not associated with the responses. [3]

While linear regression is the most commonly known regression, due to the nature of the data this type of regression could not be used. A scatterplot of the data confirms the non-linearity of the data. The scatterplot took the shape of the island, considering the data we are handling, this is to be expected.

Linear regression has a basic set of assumptions:

1. Linear relationship
2. Continuous variables
3. No auto-correlation
4. Homoscedasticity
5. Normal Errors

Linear relationship refers to the behavior of the data when plotted as shown in the following figure:

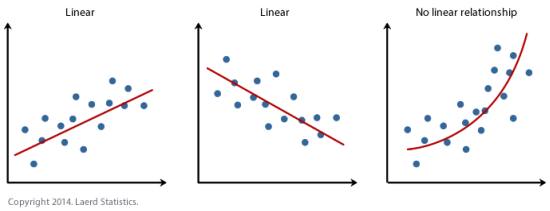

Figure 3
Examples of Linear and Non-Linear Relationships

This brings us to another assumption which was the homoscedasticity. Meaning that the variances along the line of best fit should remain similar as you move along the line which is essentially removing outliers from the data and maintaining the shape along the line [4].

Additionally, the variables must be continuous in the matter that they could assume an infinite number of possible values. This was an incorrect assumption for our data, therefore the initial data manipulation.

The variables couldn't be auto correlated, which refers to the correlation between members of a series of numbers arranged in time. This means that the correlation of a time series with its own past and future values should be avoided because there will be a tendency for a system to remain in the same state from one observation to the next.

Finally, check that the residuals (errors) of the regression line are approximately normally distributed. Two common methods to check this assumption include using either a histogram or a P-P Plot [5].

As you evaluate models, check the residual plots because they can help you avoid inadequate models and help you adjust your model for better results. For example, the bias in underspecified models can show up as patterns in the residuals [3].

Choosing a different type of regression analysis depends on the characteristics of the data.

In this case a model that took ordinal/categorical data into consideration instead of all continuous variables was required therefore the proportional odds model was an option. The proportional odds cumulative logit model is possibly the most popular model for ordinal data. This model uses cumulative probabilities up to a threshold, thereby making the whole range of ordinal categories binary at that threshold. The response being Y=1,2,..., J where the ordering is natural. The associated probabilities are $\{\pi 1, \pi 2,..., \pi J\}$, and a cumulative probability of a response less than equal to j is: $P(Y \leq j)=\pi 1+\ldots+\pi j$ [6].

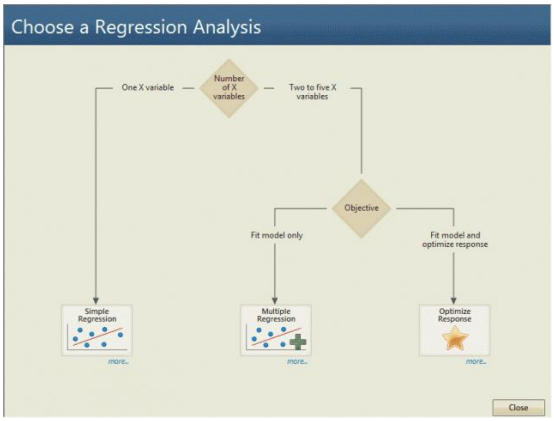

Figure 4
Choosing a Regression Analysis by Minitab

Although a polytomous logistic regression instead of having a binary logistic regression, where the response is a binary variable with 'success' and 'failure' being only two categories, it can be extended to handle polytomous responses. Polytomous referring to r > 2 categories. When r = 2, Y is dichotomous and we can model log of odds that an event occurs or does not occur. For binary logistic regression there is only 1 logit that can be formed: $\text{logit}(\pi)=\log(\pi 1-\pi)$ Polytomous regression allows categorical variables, instead of being a quantitative explanatory variable [7].

The results in a regression analysis differ from the normal statistics analysis like the p-value. For each variable, test the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response [8].

Regression coefficients represent the mean change in the response when holding other predictors in the model constant. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model.

The constant is necessary when your regression line naturally does not go through the origin (0). If the constant is left out then the regression will be forced to go through the origin which means that all the variables must equal zero when the response (Y) is 0 [9].

To determine how well the model fits the data, look into the R-squared (R^2). R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination for multiple regression. It is the percentage of the response variable variation that is explained by a linear model. It is a common error to assume a low R^2 is bad but in some fields, it is entirely expected. For example, any field that attempts to predict human behavior, as is our case; typically have R^2 values lower than 50%. Humans are simply harder to predict than physical processes. Furthermore, if the R^2 value is low but the predictors are statistically significant, conclusions can still be drawn on how the response changes based on the predictor. Regardless of the R^2, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant [10].

An additional test which should be looked at in a regression is the F-test. Unlike t-tests that can assess only one regression coefficient at a time, the F-test can assess multiple coefficients simultaneously. It compares a model with no predictors to the model that you specify [11].

The hypotheses for the F-test of the overall significance are as follows:

Null hypothesis: The fit of the intercept-only model and your model are equal.

Alternative hypothesis: The fit of the intercept-only model is significantly reduced compared to your model.

The language chosen to implement the program that used the regression model equations was Java with the IDE Eclipse. This language and IDE were chosen mainly due to the developer's expertise. To handle the inputs required from excel spreadsheets, a specialized library was required. For this task the library JExcelApi [12] was researched and determined to be the best option. The selection process was done by determining who combined simplicity with effectiveness the best. Inputs into the program were selected by availability and need. Not all available inputs were used because they made no impact on the whole. The main inputs for the program were the historical data obtained from the Puerto Rican government databases. The equations obtained from Minitab were also essential inputs.

RESULTS

The regression model outputted the following equations:

$$Points\ of\ Location = C + DY + CT + P \quad (1)$$

Where C is a constant, DY is the day of the year multiplied but its corresponding factor, CT is the type of crime multiplied but the corresponding factor, and P is the precinct in charge of the area multiplied by its proper factor.

There are separate equations for the latitude and longitude coordinates but they follow the same structure, with different constants.

These equations where then inserted into a program that ran all possible combinations of days, crimes, and precincts to output a location. These locations along with their combination of variables were then compared to the test data saved from the latest year to determine accuracy.

For easier readability, different levels of accuracy were stored in separate files, while the full set of predictions were kept apart as well.

Guidelines

A more in depth analysis could be made if more data were available, at the moment only the 3 years used were available. In statistical analysis a regularly used estimate is to use 30 data points, in our case this could mean 30 years. *Please consider* that crime in 30 years even 10 years could change dramatically and would not be accurately predicted by an equation that takes into consideration these inputs. [13] A better approach would be to estimate the amount of data required is by calculating the sample size as follows:

$$Sample\ Size = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1+(\frac{z^2 \times p(1-p)}{e^2 N})} \quad (2)$$

Where the variables represent the following:

Population Size = N | *Margin of error* = e | *z-score* = z [14] (3)

Additional studies of the criminal mind could result in an increase of the accuracy of this tool, as suggested by the FBI the population and other aspects of crime must be considered as variables that affect crime rates. [15]

DISCUSSION

The regression model equations gave us a long list of coefficients attached to each individual factor. These factors were crime type, day of the year, and precinct in charge. In total the equation had 389 factors each: 365 days, 10 crime types, 13 precincts, and 1 constant.

As these equations were run through the program to obtain the predictions, we obtained a long list of predictions, a total of 47450 predictions. Of course not all of these predictions will be accurate, but a very good number of them will have differences of 1% or less. The difference is calculated with the following equation:

$$Difference = \frac{|Theoretical\ Value - Practical\ Value|}{\frac{Theoretical\ Value + Practical\ Value}{2}} * 100 \quad (4)$$

The factors in the prediction equations have smaller impact on the equation outputs compared to the constant, but they do bring up noticeable changes in the locations obtained using the coordinates. Although they may seem small, they are hugely significant.

CONCLUSION

The prediction program outputted a long list of predictions that had differences of 1% or less from the expected values. These results were obtained from the equations extracted from the regression model. These equations could be improved by adding factors that were not available at this time.

ACKNOWLEDGEMENTS

I'd like to thank my mentor, Dr. Nelliud Torres, for all of his support and guidance.

I'd also like to thank Dr. Alfredo Cruz, director of the masters program.

REFERENCES

[1] Frontline Systems, Inc. (2015, December 12). *Transform Categorical Data* [Online]. Available: http://www.solver.com/transform-categorical-data.

[2] S. Sherwood (2015, December). *Statistics* [Online]. Available: http://web.science.unsw.edu.au/~stevensherwood/120b/Statistics.pdf.

[3] Minitab. (2015, December). *Minitab Support* [Online]. Available: http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/basics/types-of-regression-analyses/.

[4] Laerd Statistics. (2015, December). *Linear Regression Analysis using SPSS Statistics* [Online]. Available: https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php.

[5] D. R. Dawdy and N. C. Matalas, "Statistical and probability analysis of hydrologic data, part III: Analysis of variance, covariance and time series," in *Handbook of applied hydrology, a compendium of water-resources technology*, NY, McGraw-Hill Book Company, 1964, pp. 8.68-8.90.

[6] PSU. (2015, December). *The Proportional-Odds Cumulative Logit Model* [Online]. Available: https://onlinecourses.science.psu.edu/stat504/print/book/export/html/176.

[7] PSU. (2015, December). *Polytomous (Multinomial) Logistic Regression* [Online]. Available: https://onlinecourses.science.psu.edu/ stat504/node/172.

[8] J. Frost. (2013, July 1). *How to Interpret Regression Analysis Results: P-values and Coefficients* [Online]. Available: http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients.

[9] J. Frost. (2013, July 11). *Regression Analysis: How to Interpret the Constant (Y Intercept)* [Online]. Available: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-to-interpret-the-constant-y-intercept.

[10] J. Frost. (2013, May 30). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* [Online]. Available: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.

[11] J. Frost (2015). What Is the F-test of Overall Significance in Regression Analysis? [Online]. Available: http://blog.minitab.com/blog/adventures-in-statistics/what-is-the-f-test-of-overall-significance-in-regression-analysis.

[12] E. H. Jung. (2014, July 8) *JExcelApi* [Online]. Available: http://sourceforge.net/projects/jexcelapi/.

[13] L. E. Ohlin, "Effect of Social Change on Crime and Law," Notre Dame Law Review, 1968, pp. 834-846.

[14] Australian Bureau of Statistics. (2016, December). *Sample Size Calculator* [Online]. Available: http://www.nss.gov.au/nss/home.nsf/pages/Sample+size+calculator.

[15] U.S. Department of Justice — Federal Bureau of Investigation. (2010, September). *Variables Affecting Crime* [Online]. Available: https://www2.fbi.gov/ucr/cius2009/about/variables_affecting_crime.html.