Predicting Crime using Multiple Regression Analysis

Amartya Banerjee
4th Year
Computer Science and Engineering
IEM , Salt lake
amartya_banerjee@yahoo.com

Arunangshu Pal
4th Year
Computer Science and Engineering
IEM , Salt lake
arunangshumail@gmail.com

Subhabrata Sengupta
Assistant Professor
Computer Science and Engineering
IEM , Salt lake
ssg365@gmail.com

Abstract- Using multiple regression analysis, find the most suitable equation for different crimes analyzing the previous crime records and the factors.

I. INTRODUCTION

Data Mining is an analytic process designed to explore large amount of data in search of consistent patterns and relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Here variables refer to the factors on which the very thing depends whose analysis is being done. The ultimate goal of determining is prediction and predictive data mining is the most common type of data mining which issued here to analyse the crime records. The process of data mining consists of three stages: (1) The initial exploration, (2) Model building or pattern identification and (3) Deployment.

A. Stage 1: Exploration:

This stage usually starts with data preparation which may involve cleaning data, selecting data which are needed, giving the data a structured form. Basically we need to find all the factors on which it depends. Among all the data, collecting only the data which are related to these factors are called collecting subset of data. Generally these data are available in unstructured form. So we need to transform it to a structured form. Until and unless we give these data a structured form we cannot use it for analysing.

B. Stage 2: Model building and pattern identification:

This stage involves considering various models and choosing the best one based on their predictive performance .This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. In this very stage we analyse the data and find a predictive model which gives the values which are very close to the original value. Here we mainly concentrate on finding the best
suited model so that the summation of the total difference of the original value and the value we get from the model is least. So we need to make sure that the standard deviation is minimal.

Stage 3: Deployment:

That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome. We represent the predictive model using an equation where the (x_1, x_2…..x_i) are the factors and (y) is the value on which the analysis is done. In the model building stage we generate the best suited equation and use it for prediction. Thus giving the values of the factors we can easily get the future values of y. We may also represent it using different user interface like line graph, pie chart etc.

II. WORKING PRINCIPLE

For finding out the predictive equation we can use several methods. The one which we had followed is regression analysis. There are two parts of this regression analysis. One is bivariate regression where the dependent variable depends on a single independent variable. The one which we have used is multiple regression analysis where the dependent variable (crime value) depends on several independent variables (factors on which the crime depends).

A multiple regression model is a linear model with many predictors. In general, we write the model as

$Y' = b_0 + b_1X_1 + b_2X_2 + b_3X3 + …….. + b_kX_k$;

Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors.

More precisely, multiple regression analysis helps us to predict the value of Y for given values of X_1, X_2, X_3, ……. X_k. Multiple regression analysis is used when one is interested in predicting a continuous dependent variable from a number of independent variables. Here b0 is the intercept and b_1, b_2, b_3, …, b_k are analogous to the slope in linear regression equation and are also called regression coefficients. They can be interpreted the same way as slope. The value of these regression coefficients need to be found out using algorithm.

The basic methodology is to find out the Y' so that the summation of the squares of the difference of original value and predicted value is least. As Y is dependent on b_0,b_1,b_2…b_X thus we find the values of this regression coefficients

maintaining the above criterion. We select b1, b2, ..., b_X that minimize the sum of the squared residuals:

Sum of Squared Residuals = $e_1^2 + e_2^2 + e_3^2 + e_4^2 + \ldots\ldots\ldots\ldots\ldots\ldots e_K^2$

Sum of Squared Residuals = $\sum_{i=1}^{K} e_i^2$

Here e_i represents the difference between the predicted value and the original values for the i^{th} data record.

Sum of Squared Residuals = $\sum_{i=1}^{K} (Y_i - Y'_i)^2$

Thus to get the desired equation for predicting the crime our main objective is to find out the regression coefficients such that sum of squared residuals is least. So by getting these constants we can form our equation using which we can easily predict the crime values.

III. Proposed Algorithm

A. Algorithm to find the regression equation:

In this project there are two portions where two algorithms are used. The 1st one is to find the regression coefficients so that the sum of the squared residuals is least and find the proper equation of regression line.

Another is to find the very factor on which the crime depends most among all the factors so that more focus can be given on that factor to reduce the crime rate.

The multiple linear regression model which is used can be expressed as:

$Y' = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k + e;$

Suppose we are doing this crime record analysis using n number of records. In that case there would be n number of Y values and for each Y there would be k number of X values.

Thus we can write the above mentioned equation for n observations:

$$Y_1' = b_0 + b_1X_{11} + b_2X_{12} + b_3X_{13} + \ldots + b_kX_{1k} + e_1$$
$$Y_2' = b_0 + b_1X_{21} + b_2X_{22} + b_3X_{23} + \ldots + b_kX_{2k} + e_2$$
$$Y_3' = b_0 + b_1X_{31} + b_2X_{32} + b_3X_{33} + \ldots + b_kX_{3k} + e_3$$
$$\vdots$$
$$\vdots$$
$$\vdots$$
$$\vdots$$
$$Y_n' = b_0 + b_1X_{n1} + b_2X_{n2} + b_3X_{n3} + \ldots + b_kX_{nk} + e_n$$

We can represent these equations in matrix format.

The Y matrix represents :

$$\begin{bmatrix} Y1 \\ Y2 \\ \vdots \\ \vdots \\ Yn \end{bmatrix}$$

The X matrix represents :

$$\begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & X_{23} & \ldots & X_{2k} \\ 1 & X_{31} & X_{32} & X_{33} & \ldots & X_{3k} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & \ldots & X_{nk} \end{pmatrix}$$

The b matrix represents :

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_n \end{pmatrix}$$

The e matrix represents :-

$$\begin{pmatrix} e_0 \\ e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{pmatrix}$$

Or Y= X b + e;

Here we used the least-squares approach to estimate of the b's in the

Fixed-x model. For the parameters $b_0, b_1, \ldots, b_k$, we seek estimators that minimize the sum of squares of deviations of the n observed y's from their predicted values Y';

We need $b_0, b_1, b_2 \ldots b_k$ that minimize

$\sum_{i=1}^{K} e_i^2$

This means,

$$\sum_{i=1}^{n} (Y_i - Y`_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_1 - b_2 X_2 - b_3 X_3 - \ldots\ldots\ldots\ldots - b_k X_k)^2$$

Note that the predicted value $Y' = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k + e;$

Estimates $E(y_i)$, not y_i. A better notation would be E(yi), but yi'is commonly used. To obtain the least-squares estimators, it is not necessary that the prediction equation

$Y' = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k + e;$be based on $E(y_i)$.

It is only necessary to postulate an empirical model that is linear in the b's, and the least-squares method will find the ―best‖ fit to this model.

To find the values of $b_0, b_1, \ldots, b_k$ that minimize the equation we could differentiate

$\sum_{i=1}^{K} e_i^2$

With respect to each b_j and set the results equal to zero to yield k +1 equations that can

be solved simultaneously for the b_j's. However, the procedure can be carried out in more

compact form with matrix notation. The result is given in the following theorem 12 Predicting crime using Multiple Regression Analysis.

If Y= X b + e; where X is n * (k +1) of rank k + 1 , n, then the value of

$b = (b_0, b_1, \ldots, b_k)'$ that minimizes the given equation is

$$b = (X'X)^{-1} X'Y$$

To examine the structure of X'X and X'Y, the (k +1)*(k -1) matrix X'X can be obtained as products of columns of X; similarly, XY contains products of columns of X and Y:

$$X'X = \begin{pmatrix} n & \sum_i X_{i1} & \sum_i X_{i2} & \dots & \sum_i X_{ik} \\ \sum_i X_{i1} & \sum_i X^2_{i2} & \sum_i X_{i1}X_{i2} & \dots & \sum_i X_{i2} X_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_i X_{ik} & \sum_i X_{i1}X_{ik} & \sum_i X_{i2}X_{ik} & \dots & \sum_i X^2_{ik} \end{pmatrix}$$

$$X'y = \begin{pmatrix} \sum_i y_i \\ \sum_i X_{i1}\, y_i \\ \vdots \\ \vdots \\ \sum_i X_{ik} y_i \end{pmatrix}$$

B. Algorithm to find the most affecting factor:

From crime record analysis we can find that all kinds of crime are dependent on some factors. But the degree of dependency may vary. A particular crime may depend on a particular factor mostly among all the other factors. Our objective is to find that very factor on which the crime depends most. And to find that we use correlation factor. In statistics, the **Pearson product-moment correlation coefficient** (sometimes referred to as the **PPMCC** or **PCC** or **Pearson's *r***) is a measure of the linear correlation (dependence) between two variables *X* and *Y*, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables.

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. The formula for ρ is

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\operatorname{cov}(X,Y) = E[(X-\mu_x)(Y-\mu_x)]$$

Then the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{E[(X-\mu_x)((X-\mu_y)]}{\sigma_X \sigma_Y}$$

where:

cov is covariance and σ_x is standard daviation of x ,σ_y standard deviation of y , μ_x is the mean of x and μ_y is the mean of y and E is the expectation.

The formula for ρ can be expressed in terms of uncentered moments. Since

$\mu_x = E(X)$

$\mu_y = E(Y)$

$$\sigma_x^2 = E[(X-E(X))^2] = E(X^2) - E(X)^2$$

$$\sigma_y^2 = E[(Y-E(Y))^2] = E(Y^2) - E(Y)^2$$

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}}$$

Thus using this formula for every factor with the crime we can get the correlation factor of every factor. And from that we can easily find out the one on which the crime depends most. And it would be the factor for which the correlation will be greatest.

IV. Experimental Result

First of all we collected all the data regarding the registered rape cases in West Bengal. Then we took some factors and collected data on those too. We found that such crime depends **on GDP per capita, population** and **level of education**. For level of education we took GER ratio which means **Gross Enrolment Ratio**. It signifies Enrolment in under graduate. These are the three factors which we chose.

Year	GDP per capita	Population	GER	Rape cases
2004	41217	83544280	71.17	1661
2005	46952	84659060	72.01	2085
2006	52665	85773839	72.86	2045
2007	67797	86888618	73.7	2409
2008	66104	88003398	74.54	1790
2009	72780	89118177	75.39	1748
2010	89894	90232956	76.23	2395
2011	97672	91347736	77.08	1870

Using these data we used our analysis algorithm and found out the predictive equation which suits most. Firstly we formed the matrixes and then applied our algorithm.

The Matrixes are:

$$Y = \begin{pmatrix} 1661 \\ 2085 \\ 2045 \\ 2409 \\ 1790 \\ 1748 \\ 2395 \\ 1870 \end{pmatrix}$$

$$X=\begin{pmatrix} 41217 & 83544280 & 71.17 \\ 46952 & 84659060 & 72.01 \\ 52665 & 85773839 & 72.86 \\ 67797 & 86888618 & 73.7 \\ 66104 & 88003398 & 74.54 \\ 72780 & 89118177 & 75.39 \\ 89894 & 90232956 & 76.23 \\ 97672 & 91347736 & 77.08 \end{pmatrix}$$

By using our derived formula to calculate the regression coefficients $b = (X'X)^{-1} X'Y$ we find it out.

$$b=\begin{pmatrix} 286.548300258699 \\ 0.0422413816613923 \\ -0.00254022395190306 \\ 2981.68261096212 \end{pmatrix}$$

The final b matrix gives the constants in a matrix form.
$b_0 = 286.548300258699$
$b_1 = 0.0422413816613923$
$b_2 = - 0.00254022395190306$
$b_3 = 2981.68261096212$
The predictive equation which we found from analysing the crime records is:

$Y = (286.548300258699) + (0.0422413816613923)* x1 + (-0.00254022395190306)* x2 + (2981.68261096212)*x3$

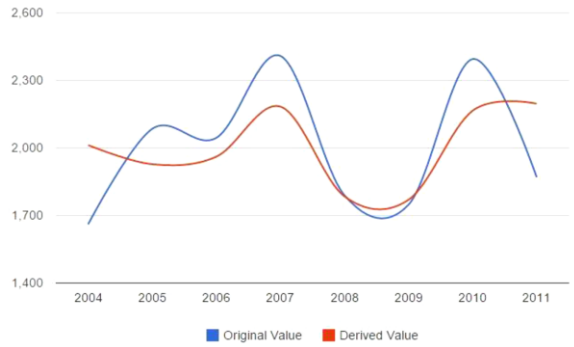
After finding the regression line equation we also calculated the correlation factor for the factors with the crime values. For this we used the above mentioned algorithm.
Then we calculated the correlation between the **registered rape record** and **population**
X Values $\Sigma = 699568064$ Mean = 87446008
$\Sigma(X - Mx)2 = SSx = 52194786634698$
Y Values $\Sigma = 16003$ Mean = 2000.375
$\Sigma(Y - My)2 = SSy = 571979.875$
The value of r is 0.1533

I. Performance Analysis

After finding the regression predictive equation we did performance analysis. For that we calculated value of the rape cases using the predictive equation for the years which we already know. After that we calculated the variation of the prediction value with the original value. And the result seemed to be very promising.

YEAR	Original Value	Estimated Value	Deviation (%)
2004	1661	2012.78	-21.18
2005	2085	1927.85	7.53
2006	2045	1961.27	4.09
2007	2409	2183.84	9.34
2008	1790	1785.15	0.7
2009	1748	1769.80	-1.25
2010	2395	2165.54	9.58
2011	1870	2196.73	-17.47

Here we have shown graphically how much the estimated and the original value have deviated in a graphical manner.
X axis – YEAR
Y axis – Registered Rape Cases



We can clearly observe how close these points are in the graph. From this we can clearly understand that the predictive model has been able to predict the values of the crimes at its level. Our main objective was to find the equation such a way so that the difference between the predicted value and the original value is less. This would eventually make the square of those differences lest.
This is one of the test cases. We can find different predictive equations for different crimes. As each crime has different set of affecting factors. The correlation factor gets changed with the data. So does the equation. In this manner we can find out different model for different crimes. It can vary from state to state for same crime also.

V. Future Scope

There are several futures scopes to this thesis. One of the sides of perfection is the number of independent variables rather the factors on which the crime depends. We can certainly survey and find out many more factors on which the crime depends. And increase the number of dependent variable. In this way the productiveness of a predictive model would increase.
We can divide each factor in to many subparts. There could be a possibility of variables like:

A. Demographic Variables

1. Total population in thousands 2. Population density (population per square mile) B. Population Composition 3. Percentage population non-white 4. Percentage population under age 185. Percentage families headed by female with own children under age 18 6. Percentage primary male individuals of total population**.**

B. Population Shift

7. percent change in total population,. 8. percent change in white population. 9. percent change in non-white population.

C. Housing Variables

10. Percent housing owner-occupied, 1970. 11. percent housing crowded (more than one person per room), 1970. 12. percent housing substandard (lacking some or all plumbing), 1970.

D. Income and Living Standards

13. Per capita personal income (SMSA). 14. Median value of owner-occupied housing. 15. median monthly rent, 1970. 16. per capita retail sales (proxy of consumerism),

E. Physical Structure of the Community

17. Percent assessed value of single-family houses,. 18. Percent assessed value of

Besides these we can also do this prediction for many more crimes. And find pattern on their beings.

REFERENCES

Theory :

[1] Yong Hyo Cho, The university of Akron, Akron. "A Multiple Regression Model for the Measurement of the Public Policy Impact on Big City Crime."

[2] Theresa Hoang Diem Ngo , La Puente, CA. The Steps to Follow in a Multiple Regression Analysis.

[3] New York University, Stern School of Business. Multiple Regression Basics.

[4] Alvin C. Rencher and G. Bruce Schaalje, Department of Statistics, Brigham Young University, Provo, Utah. LINEAR MODELS IN STATISTICS

[5] David M Diez , Christopher D Barr , Mine Cetinkaya-Rundel , Introductory Statistics With Randomization and Simulation

[6] Shyam Varan Nath , Oracle Corporation. Crime Pattern Detection Using Data Mining.

[7]. Dr: Zakaria Suliman Zubi , Sirte University, Faculty Of Science, Computer Science Department. Ayman Altaher Mahmmud , The Libyan Academy, Information Technology Department, Tripoli, Libya . Using Data Mining Techniques to Analyze Crime patterns in the Libyan National Crime Data

[8]. Malathi. A , Assistant professor, Department of Comuter Science, Govt Arts College, Coimbatore, India . Dr. S. Santhosh Baboo , Readers, Department of Computer Science, D. G. Vaishnav College, Chennai, India . Anbarasi. A , Associative Professor, Department of Civil Engineering, Kumaraguru college of Technology, Coimbatore. An intelligent Analysis of a City Crime Data Using Data Mining

M. Todd Henderson, Justin Wolfers, Eric Zitzewitz, University of Chicago. Predicting Crime

[9] Crime prediction and prevention , IBM Software Group

[10] Hua Liu, CSG Systems, Inc., One Main Street, Cambridge, MA 02142, USA . Donald E. Brown, Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22903, USA. C riminal incident prediction using a point-pattern-based density model

[11] Nikhil Dubey, M.Tech. Research scholar Dept. of Computer Science and Engg. Technocrats Institute of Technology, Bhopal. Setu Kumar Chaturvedi, Professor & Head Dept. of Computer Science and Engineering Technocrats Institute of Technology, Bhopal. A Survey Paper on Crime Prediction Technique Using Data Mining

[12] Stephen Schneider, Ph.D.Adjunct Professor,School of Justice Studies,Ryerson University,Research and Statistics Division. PREDICTING CRIME: A REVIEW OF THE RESEARCH

[13] Elizabeth R. Groff National Institute of Justice ,Nancy G. La Vigne The Urban Institute . FORECASTING THE FUTURE OF PREDICTIVE CRIME MAPPING

[14] S.Yamuna, M.Phil (CS) Research Scholar School of IT Science, Dr.G.R.Damodaran College of science Coimbatore. N.Sudha Bhuvaneswari, Associate Professor MCA, Mphil (cs), (PhD) School of IT Science Dr.G.R.Damodaran College of science Coimbatore.
Datamining Techniques to Analyze and Predict Crimes

[16] Malathi. A, Assistant Professor Post Graduate and Research department of Computer Science, Government Arts College,Coimbatore, India , Dr. S. Santhosh Baboo , Reader, Post Graduate and Research Department of Computer Science,D.G Vaishnav College,Chennai, India . An Enhanced Algorithm to Predict a Future Crime using Data Mining

Data :

[1] National Crime Records Bureau. (http://ncrb.gov.in/)
[2] data.gov.in (https://data.gov.in/)
[3] Wikipedia (https://en.wikipedia.org/)