Regression Analysis and Statistical Approach on Socio-Economic Data

Syeda Farha Shazmeen[1], Mirza Mustafa Ali Baig[2], M. Reena Pawar[3]

Abstract

Statistical analysis and data mining addresses the broad area of data analysis, including data mining algorithms, statistical approaches and practical application. Statistical methods and algorithms can be combined to form the basic idea such as factor analysis, function based data analysis etc. and thorough experimental evaluations show that the results are highly effective and efficient.

Keywords

Linear Regression, Statistics, SPSS, Data Normalization.

1. Introduction

Data mining can be viewed as an extension of statistical analysis techniques used for exploratory analysis and incorporating new techniques [1]. Regression analysis is an important statistical method for the analysis of socio economic data. It also helps us in recognition and categorization of relationships between various factors. ARIMA (auto–regressive integrated moving average), long-memory time-series modeling, and auto-regression are popular methods for such analysis [2].

Communities and Crime Dataset

The data combines socio-economic data[3] from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. Many variables are included so that algorithms that select or learn weights for attributes could be tested. However, clearly unrelated attributes were not included; attributes were picked if there was any plausible connection to crime (N=122), plus the attribute to be predicted (Per Capita Violent Crimes). The variables included in the dataset involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units. Many of these omitted communities were from the Midwestern USA. Data is described below based on original values. All numeric data was normalized into the decimal range 0.00-1.00 using an Unsupervised, equal-interval binning method. Attributes retain their distribution and skew (hence for example the population attribute has a mean value of 0.06 because most communities are small). A limitation was that the LEMAS survey was of the police departments with at least 100 officers, plus a random sample of smaller departments. For our purposes, communities not found in both census and crime datasets were omitted. Many communities are missing LEMAS data.

Regression

Regression analysis is used when researchers want to predict a continuous dependent variable (DV) from a number of independent variables(IV). The purpose of regression analysis [4, 5] is to come up with an equation of a line that fits through that cluster of points with the minimal amount of deviations from the line. The deviation of the points from the line is called "error."

Pre-processing

The pre-processing includes data Cleaning, data Transformation and data Selection. In data cleaning the missing values are eliminated by using the minimum, maximum or average of the attributes. After eliminating the missing values, to improve the Regression factor, the data transformation is used by applying the Z-score Normalization.

$$x_i' = \frac{sd_{user}}{sd} \times (x_i - \overline{X}) + \overline{X}_{user}$$

$$x_i' = \frac{sd_{user}}{sd} \times (x_i - \overline{X}) + \overline{X}_{user}$$

$$N\left(\overline{X}_{user}, \frac{sd_{user}^2}{w_i}\right)$$

After a z-score transformation, the following statistics are updated

1. Number of missing values
$$N_{x'}^{missing} = N_x^{missing}$$
2. Number of valid values
$$N_{x'} = N_x$$
Min value:
$$min(x_i') = \frac{sd_{user}}{sd} \times (minx_i - \overline{X}) + \overline{X_{user}}$$
Max value:
$$max(x_i') = \frac{sd_{user}}{sd} \times (maxx_i - \overline{X}) + \overline{X_{user}}$$

3. Mean: $\overline{x'} = \overline{x_{user}}$
   Standard deviation: $sd(x') = sd_{user}$
4. Skewness is the measure of the asymmetry of a distribution; the normal distribution is symmetric and has skewness.
   Skewness: $skew(x') = skew(x)$

2. Regression

Regression analysis can imply a broader range of techniques that ordinarily appreciated. Statisticians commonly define regression so that the goal is to understand "as far as possible with the available data how the conditional distribution of some response y varies across sub populations determined by the possible values of the predictor or predictors".

If the DV have two form, then logistic regression [6,7] can be used. Regression is used for building advanced data mining model, the applications range from assessing experimental data, through statistical and econometric. It is mainly used for estimating a relationship between many attributes. While being effective and relatively simple method, regression can be applied only for data that are internally dependable.

The IV's used in regression can be either continuous or dichotomous. The variables which are having two levels those variables can be used in regression analysis if not they must be converted in two levels. Usually, regression analysis is used with naturally-occurring variables, even though we can use regression with experimentally manipulated variables. One important thing to consider is regression analysis is that underlying relationships among the variables cannot be determined. While the terminology is such that we say that X "predicts" Y, we cannot say that X "causes" Y.

Regression analysis[8] also has an assumption of linearity. Linearity means that there is a straight line relationship between the IVs and the DV. This assumption is important because regression analysis only tests for a linear relationship between the IVs and the DV. Any nonlinear relationship between the IV and DV is ignored. You can test for linearity between an IV and the DV by looking at a bivariate scatter plot (i.e., a graph with the IV on one axis analytics, and collaboration and deployment (batch and automated scoring services) 1968 after being developed by Norman H. Nie, Dale H. Bent, and C. Hadlai Hull. SPSS [14] is among the most and the DV on the other). If the two variables are linearly related, the scatter plot will be oval.

Linear Regression

Linear regression is a statistical [9] procedure for predicting the value of a DV from an IV when the relationship between the variables can be described with a linear model [10,11]. Simple linear regression is when researchers want to predict values of one variable, given values of another variable. The relationship is typically expressed in terms of a mathematical equation such as Y = b + mX.

3. Implementation

The proposed technique is implemented using SPSS and STATISTICA, it is a statistics and analytics software package developed by Stat Soft. Statistica provides data analysis, data management, statistics, data mining, and data visualization procedures. Statistica web product categories include Enterprise (for use across a site or organization), Web-Based (for use with a server and browser), Concurrent Network Desktop, and Single-User Desktop.

Statistica originally derives from a set of software packages and add-ons that were initially developed during the mid 1980's by Stat Soft. Following the 1986 release of CSS (Complete Statistical System) and the 1988 release of MacSS (Macintosh Statistical System), the first DOS version of Statistica (trademarked in capitals as Statistica) was released in 1991.

Statistica 5.0 was released in 1995. It operates on both the new 32-bit Windows 95/NT and the older version of Windows (3.1). It featured many new statistics [12] and graphics procedures, a word-processor-style output editor (combining tables and graphs), and a built-in development environment that enabled the user to easily design new procedures (e.g., via the included Statistica Basic language) and integrate them with the Statistica system. To analyze the regression analysis, SPSS Statistics [13] is a software package used for statistical analysis. It is now officially named "IBM SPSS Statistics", Data mining (IBM SPSS Modeller), Text SPSS Statistics (originally, Statistical Package for the Social Sciences, later modified to read Statistical Product and Service Solutions). It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. The original SPSS manual has been described as one of "sociology's most influential books". In addition to statistical analysis, data management and data documentation (a metadata dictionary is stored in the data file) are features of the base software. SPSS was released in its second version in 1972 and its company name is INDUS Nomi.

Rapid Miner [15] is an open source-learning environment for data mining and machine learning. This environment can be used to extract meaning from a dataset. There are hundreds of machine learning operators to choose from, helpful pre and post processing operators, descriptive graphic visualizations, and many other features. It is available as a stand-alone application for data analysis and as a data-mining engine for the integration into own products.

4. Experimental Analysis

'Model summary' table, which provides information about the regression line's ability to account for the total variation in the dependent variable demonstrates that the observed y-values are highly dispersed around the regression line. ANOVA table presents the F-Statistics; the most commonly used significance threshold is .05, which means that the variable or model would be significant at the 95% level.

Model Summary
Table 1(a): This table indicates the model summary of data with missing values.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.837	.700	.683	.13119

ANOVA
Table 1(b): This table indicates the F-statistics of data with missing values.

Model	Sum of Squares	df	Mean Square	f	Sig.
1Regression	75.706	106	.714	41.4	.000^{b}
Residual	32.478	1887	.017		
Total	108.184	1993			

Model Summary
Table 2(a): This table indicates the model summary of without missing values

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997	.993	.797	.12325

ANOVA
Table 2(b): This table indicates the F-statistics of data without missing values.

Model	Sum of Squares	df	Mean Square	f	Sig.
1Regression	8.995	117	.077	5.06	.061^{b}
Residual	.061	4	.015		
Total	9.056	121			

Model Summary
Table 3(a): This table indicates the model summary of data after applying factor analysis.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899	.809	.782	.12788

ANOVA
Table 3(b): This table indicates the F-statistics of data after applying factor analysis.

Model	Sum of Squares	df	Mean Square	f	Sig.
1 Regression	7.323	15	.488	29.85	.000^{b}
Residual	1.733	106	.016		
Total	9.056	121			

From the above table i.e. model summary and anova, we can observe that R Square and F-Statistics have been improved by applying factor analysis techniques.

Regression Graphs

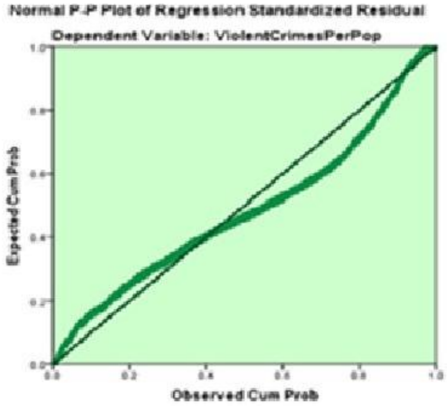
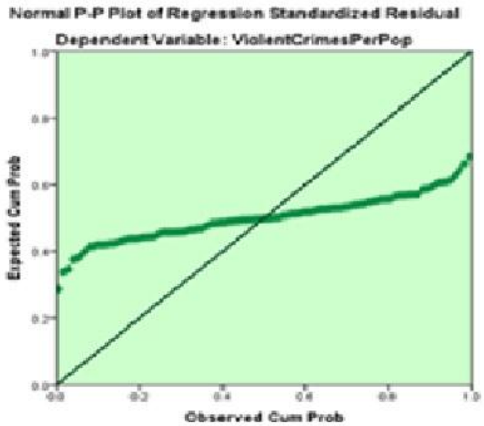
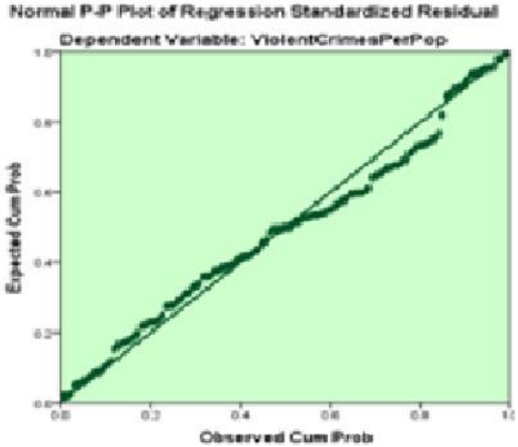


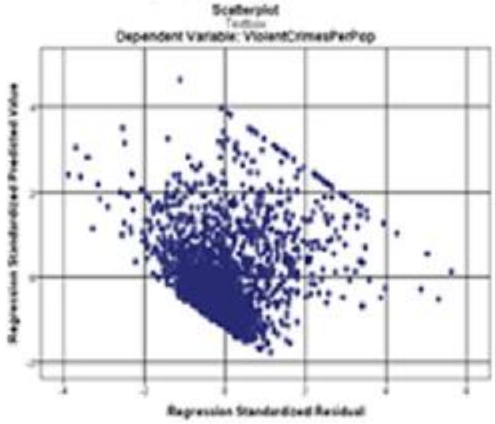
Fig (1.a):-Correspond to Data with missing values



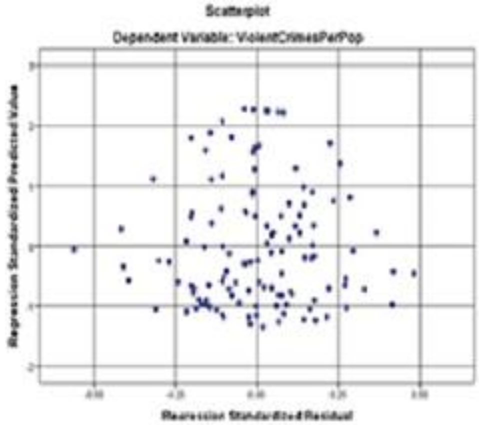
Fig(1.b):-Correspond to data without missing values.



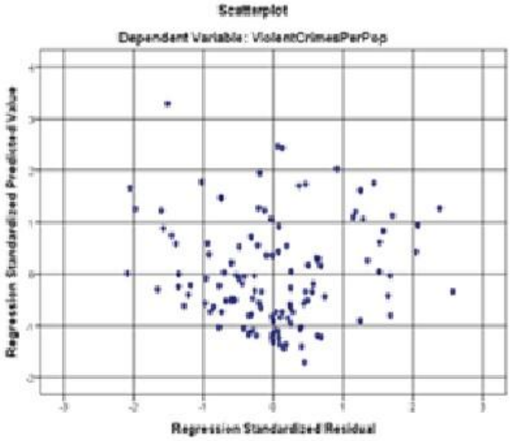
Fig(1.c):-Correspond to data after applying factor reduction



Fig(2.a):-Embody the data with missing values.



Fig(2.b):-Embody the data without missing values.



Fig(2.c):-Embody the data after applying factor analysis.

Scatter plots: These are useful for plotting multivariate data. They can help you determine potential relationships among scale variables.

5. Conclusion and Future scope

This paper presents a better approach for Regression. We have applied statistical approach to linear regression which leads to more accurate regression results. For future work we planned to extend our research in following directions, to find more efficient results of regression and to work on alternative statistical tools which can lead to accurate results.

References

[1] "On the Impact of Knowledge Discovery and Data Mining". Kirsten Wahlstrom, John F. Roddick, School of Computer and Information Science University of South Australia. Australian Computer Society, at the 2nd Australian Institute of Computer Ethics Conference (AICE2000), Canberra.

[2] "A Regression-Based Temporal Pattern Mining Scheme for Data Streams" Wei-Guang Teng, Ming-Syan Chen, Electrical Engineering Department National Taiwan University, Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003.

[3] Distributed, Parallel, and Cluster Computing (cs.DC), From Social Data Mining to Forecasting Socio-Economic Crisis", Dirk Helbing, Stefano Balietti. The European Physical Journal - Special Topics Volume 195, Number 1, 3-68, DOI: 10.1140/epjst/e2011-01401-8, Computers and Society (cs.CY); Databases (cs.DB).

[4] "A Regression-Based Temporal Pattern Mining Scheme for Data Streams" Wei-Guang Teng, Ming-Syan Chen, Electrical Engineering Department National Taiwan University, Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003.

[5] "An Introduction to Logistic Regression Analysis and Reporting", chao-ying joanne peng Kuk lida lee Gary m. Ingersoll, Indiana University-Bloomington September/October 2002 [Vol. 96(No. 1)].

[6] Hosmer, D. and Stanley, L. (1989). Applied Logistic Regression, John Wiley and Sons, Inc.

[7] "Applied Regression Analysis: A Research Tool, Second Edition", John O. Rawlings Sastry G. Pantula David A. Dickey Springer, Second Edition, Department of Statistics, and North Carolina State University.

[8] "Measuring Bank Efficiency: A Meta-Regression Analysis ", Zuzana Iršová, Tomáš Havránek, PRAGUE ECONOMIC PAPERS, 4, 2010, Charles University in Prague, Institute of Economic Studies.

[9] "Statistical literacy guide a basic outline of regression analysis", David Knott & Paul Bolton, David Knott & Paul Bolton, March 2009.

[10] Linear Regression Models", Dr. Maher Khelifa, SPSS for Windows Intermediate & Advanced Applied Statistics, Zayed University Office of ResearchSPSS for Windows®Workshop Series.

[11] "Logistic Regression Models in Sociological Research Vernon Gayle", Paul S. Lambert University of Stirling 20th April 2009 [Edition 1.1], DAMES Node, Technical Paper 2009.

[12] "Quantitative Software Management ", Paul below, Inc. (QSM), 2011.

[13] Statistica'12, Stat Soft http://www.statsoftindia.com/.

[14] "Data mining and statistics", Gain a competitive advantage, data mining and statistics: gain a competitive advantages- SPSS.

[15] Rapid Miner-http://rapid-i.com/.

Syeda Farha Shazmeen received her Master's in Software Engineering from JNTU HYD in the year 2010. Now working as an Assistant Professor in the department of Computer Science, Balaji Institute of Technology and Science, Warangal, India. Published 3 papers in IEEE International Conferences and International Journals. Her research interest includes Data mining, Data Bases, Semantic web, and Machine Learning.

Mirza Mustafa Ali Baig received bachelor degree from Balaji Institute of Technology and sciences in the year 2013. Published a paper in IOSR Journal. His interest includes Data Mining, Statistics, Data Bases, Text analytics, Machine learning and Computer Networks.

M. Reena Pawar received bachelor degree from Balaji Institute of Technology and sciences in the year 2013. Published a paper in IOSR Journal. Her interest includes Data mining, Text analytics, Machine learning and Databases.