Prediction of **Critically Polluted Industrial Areas by CEPI using** Regression analysis

Vikas Ojha Aditya Gupta Prajwal Katyal

Department of Computer Science and Engineering, MRIIRS

ABSTRACT:

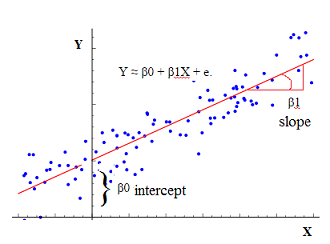
This paper will provide a perspective on industrial pollution and it’s impact on environment. In this research paper Comprehensive Environmental Pollution Index is used for analysis and prediction. The Central Pollution Control Board (CPCB) has developed a Comprehensive Environmental Pollution Index (CEPI). CPCB has done a nation wide environmental assessment of Industrial Clusters based on CEPI and 43 such industrial clusters having CEPI greater than 70, on a scale of 0 to 100, has been identified as critically polluted. **We have analysed the adverse effects of industries on Land, Water and Air** .**The calculated coefficient of CEPI score for Air ,Water and Land are as follows respectively:21.3824, 31.4203,-7.55024.**

We predicted new CEPI score according year2009,2011,2013are:89.9174,88.2628,86.84101 for critically polluted industrial area using Multiple regression analysis. R tools are used thoroughly for entire prediction and analysis of this research project.

KEYWORD: Multivariable regression, CEPI: Comprehensive Environmental Pollution Index, PIA: Polluted Industrial Areas ,Air pollution, Water pollution ,Land pollution.

INTRODUCTION:

Regression analysis is a family of statistical tools that can help sociologists better understand and predict the way that people act and interact. Regression analysis is used to build mathematical models to predict the value of one variable from knowledge of another. Although statistical methods of correlation offer researchers techniques to help them better understand the degree to which two variables are consistently related, such knowledge alone is typically insufficient to predict behavior. Simple linear regression allows the value of one dependent variable to be predicted from the knowledge of one independent variable. Multiple linear regression can be used to develop models to predict the value of a dependent variable from the knowledge of the value of more than one independent variable. Regression analysis allows researchers to build mathematical models that can be used to predict the value of one variable from knowledge of another. There are a number of specific regression techniques that can be used by sociologists to model real-world behavior.



**Figure 1: Linear Regression**

These include:

1)Simple linear regression analysis, which allows the modeling of two variables, one independent and one dependent

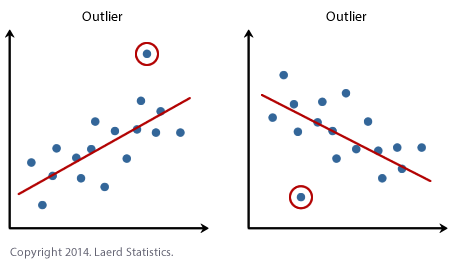
2)Multiple linear regression analysis, which allows the modeling of two or more independent variables to predict one dependent variable

3)Multiple curvilinear regression, where the relationship between variables is nonlinear (e.g., quadratic)

1.1 Simple linear regression

Statistics offers sociology researchers a number of correlation techniques to help them better understand the degree to which two variables are consistently related. For example, correlation can help one understand the relationship between educational level and income level. Correlation coefficients show the degree of relationship between two variables with a value between zero and one. A correlation of 1.0 shows that the variables are completely related and a change in the value of one variable will signify a corresponding change in the other, while a correlation of 0.0 shows that there is no relationship between the two variables and that knowing the value of one variable will tell us nothing about the value of the other.

correlation coefficient also shows how the two variables are related. A positive correlation means that as the value of one variable increases, so does the value of the other variable. A negative correlation, on the other hand, means that as the value of one variable increases, the value of the other variable decreases.



**Figure 2: Outliers**

1.2 Multivariable regression

Multivariate linear regression, which allows the simultaneous examination of several dependent variables. Multivariate polynomial regression, which can be used to account for nonlinear relationships.

LITERATURE REVIEW:

There are many tools available for the regression analysis for example excel has some extension tools which allows us to do regression analyisis in the excel itself. Then there are statisticalsoftware like SPSS or NCSS which gives this option.The facility provided in the excel requires latestexcel package and we need to download analysis toolkitto run regression analysis in the excel. Excel providesoption to select dependant and independent variableafter providing the data and it gives the result in form ofstatistical reports which includes anova test report andgeneral terms related with the regression analysis like R square value, Multiple F value etc. It also plots the givenpoint on the 2D surface and gives the approximateregression analysis line.Regression Analysis alludes to a gathering ofsystems for considering the connections among at least two variables in view of an example. NCSS makes itsimple to run either a basic direct regression analysis ora complex regression analysis, and for an assortment ofresponses. NCSS has present day graphical andstatistical devices for concentrating on residuals,multicollinearity, decency of-fit, model estimation,relapse diagnostics, subset determination, investigationof variance, and numerous different perspectives thatare particular to kind of regression being performed.Manufactured by NCSS, LLC, it is a statistical packagewhich was built in 1981 by Jerry Hintze. NCSS, LLC hassome expertise in giving measurable software forstatistical analysis to businesses, academics institutions,and researchers. It likewise delivers PASS Sample SizeSoftware which is utilized as a part of logical studyarranging and assessment.Statistical Package for Social Sciences (SPSS) is asoftware package utilized for statistical investigation.SPSS is a generally utilized program for factualinvestigation as a part of social science. It is likewiseutilized by health researchers, marketing organizations,government, market researchers and others. The firstSPSS manual (1970) has been portrayed as one of"humanism's most compelling books" for permittingcommon analysts to do their own measurable analysis.Adding statistical analysis, information administrationand information documentation are components of thebase programming software. The base software consistsof the following statistics: Bivariate Statistics, Predictionfor acknowledging groups, prediction for statisticaloutcomes, Descriptive Statistics. SPSS has a lot offeatures which can be accessed by pull down menus or itcan be customized with a restrictive 4GL command syntax.

Application technique:

The Central Pollution Control Board (CPCB) has developed a Comprehensive Environmental Pollution Index (CEPI). CPCB has done a nation wide environmental assessment of Industrial Clusters based on CEPI and 43 such industrial clusters having CEPI greater than 70, on a scale of 0 to 100, has been identified as critically polluted.

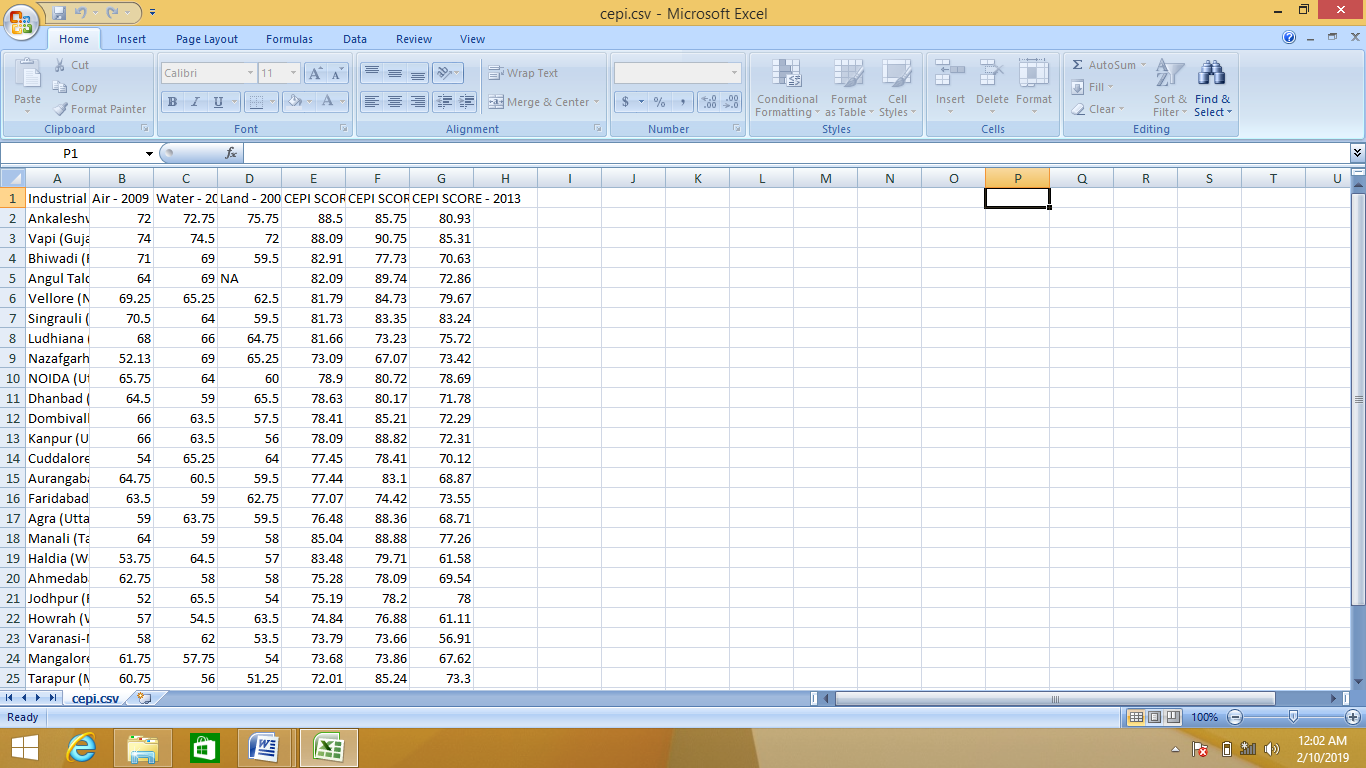
The data refers to the Comprehensive Environmental Pollution Index (CEPI) scores of the critically polluted industrial clusters/areas. The index captures the various dimensions of environment including air, water and land.

Comprehensive Environmental Pollution Index (CEPI), which is a rational number to characterize the environmental quality at a given location following the algorithm of source, pathway and receptor have been developed.

The overall CEPI is presented in the alpha-numeric form stating the score along with the status of **A**ir, **W**ater and **L**and environment in terms of subscript as **c**ritical / **s**evere/ **n**ormal. A sub-index score of more than 60 shows a critical level of pollution in the respective environmental component, whereas a score between 50–60 shows a severe level of pollution with reference to the respective environmental component.

**Critically Polluted Industrial clusters / areas: Dataset**

Industries are growing at common centers/estates/parks as the resources, manpower, transportation and marketing are feasible. Generally medium and small scale industries are developed at such areas and form industrial clusters. Due to lack of awareness and ignorance of waste management technologies, environmental pollution has been resulted to the surrounding environment. So, such industrial areas have to be assessed for improving the quality of the environment. With this objective CPCB developed Comprehensive Environmental Pollution Index (CEPI) to find out an Index value to characterize quality of the environment. In 2009, 88 prominent industrial clusters were identified in consultation with the MoEF&CC for CEPI analysis. Out of identified 88 prominent industrial clusters, 43 industrial clusters in 17 States having CEPI score of 70 and above are identified as Critically Polluted Areas (CPAs). Further, 32 industrial clusters with CEPI scores between 60 & below 70 are categorized as Severely Polluted Areas (SPAs).To assess the environmental quality in the Polluted Industrial Areas (PIAs), monitoring is carried out by CPCB through recognized environmental laboratory periodically and CEPI is assessed based on the recorded monitoring data. The evaluated CEPI reflects the environmental quality of the industrial areas and also serves as a yardstick to assess the progress achieved in the implementation of action plans. So far, three rounds of monitoring have been undertaken by CPCB (2009, 2011, 2013) based on which CEPI assessment was done.

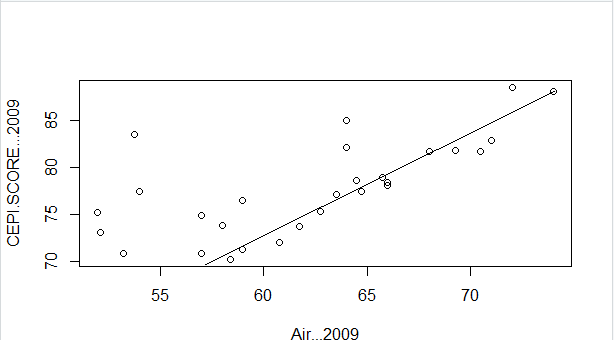


**Figure 3: Dataset**

Visualization:

1) Correlation between CEPI SCORE..2009 and air 2009

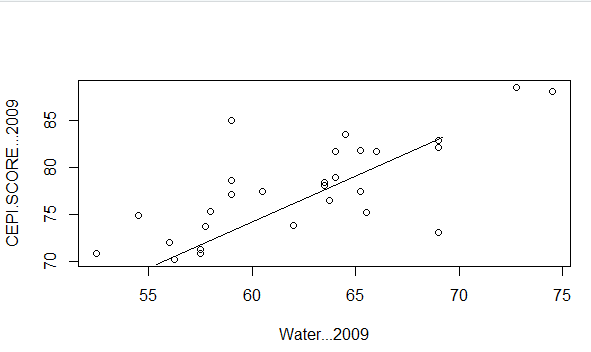




**Figure 4: Air vs CEPI (2009)**

2) Correlation between water2009 and CEPI score 2009

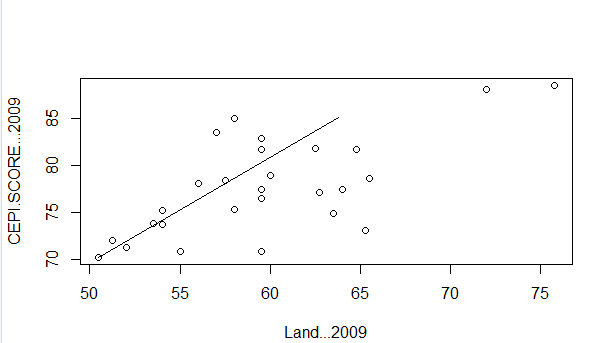




**Figure 5: Water vs CEPI (2009)**

3) Correlation between CEPI SCORE..2009 and land 2009

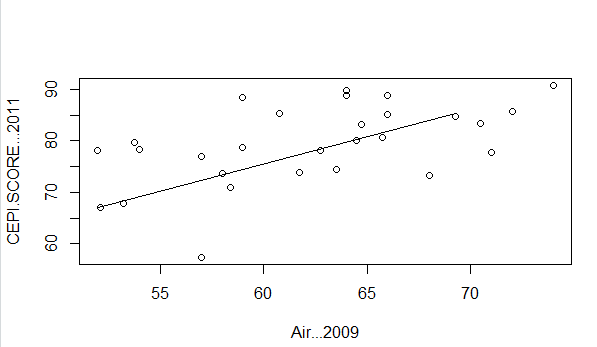




**Figure 6: Land vs CEPI(2009)**

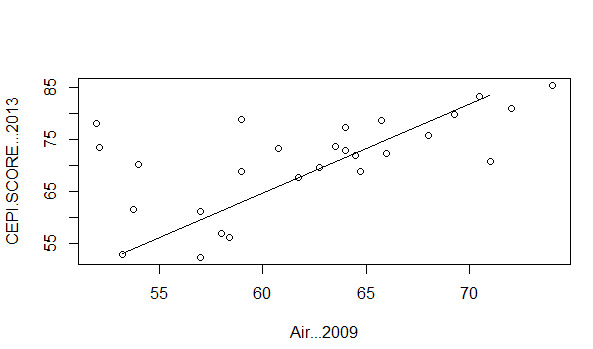
4) Correlation between CEPI SCORE..2011 and air 2009





**Figure 7: Air vs CEPI (2011)**

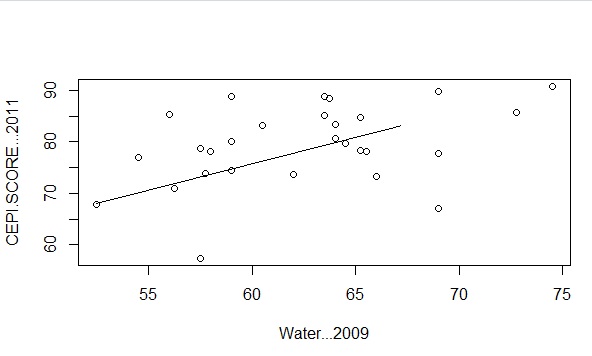
5) Correlation between CEPI SCORE..2013 and air 2009

**Figure 8: Air vs CEPI (2013)**

6) Correlation between CEPI SCORE..2011 and water 2009

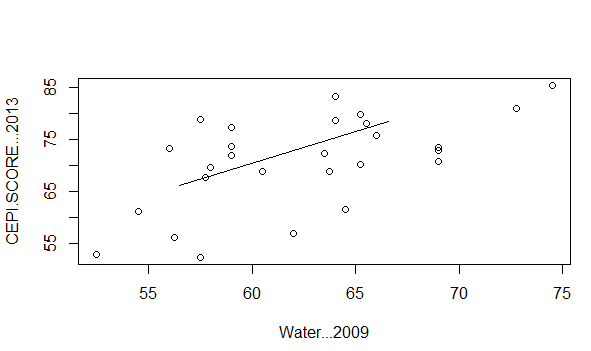




**Figure 9: Water vs CEPI (2011)**

7) Correlation between CEPI SCORE..2013 and air 2009

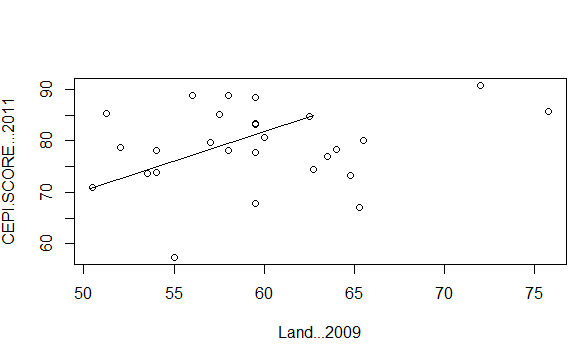




**Figure 10: Air vs CEPI (2013)**

8) Correlation between CEPI SCORE..2011 and land 2009

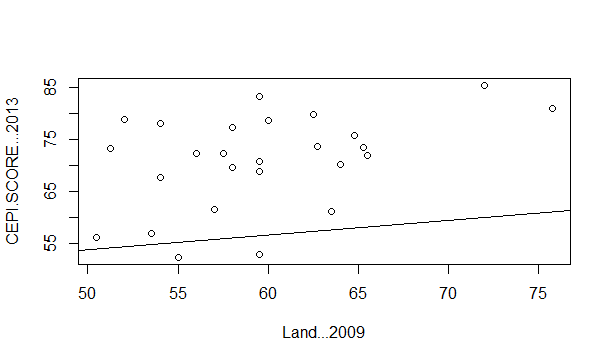




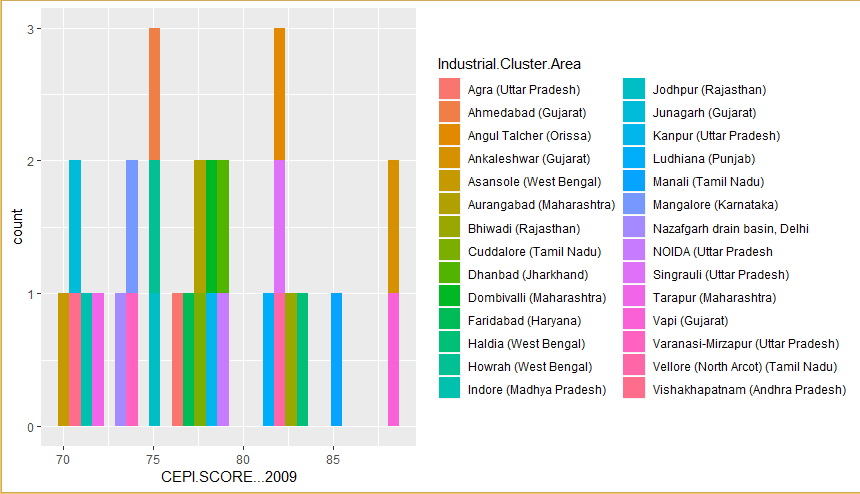
**Figure 11: Land vs CEPI (2011)**

9) Correlation between CEPI SCORE..2013 and air 2009



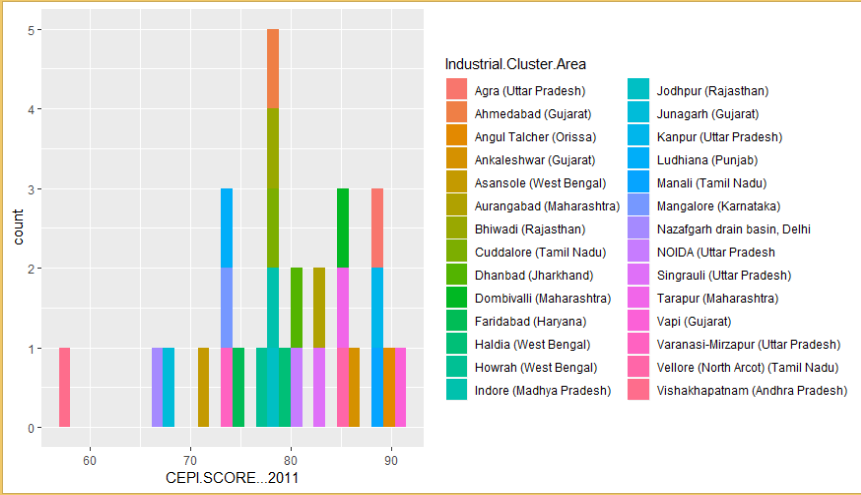


**Figure 12: Land vs CEPI (2013)**



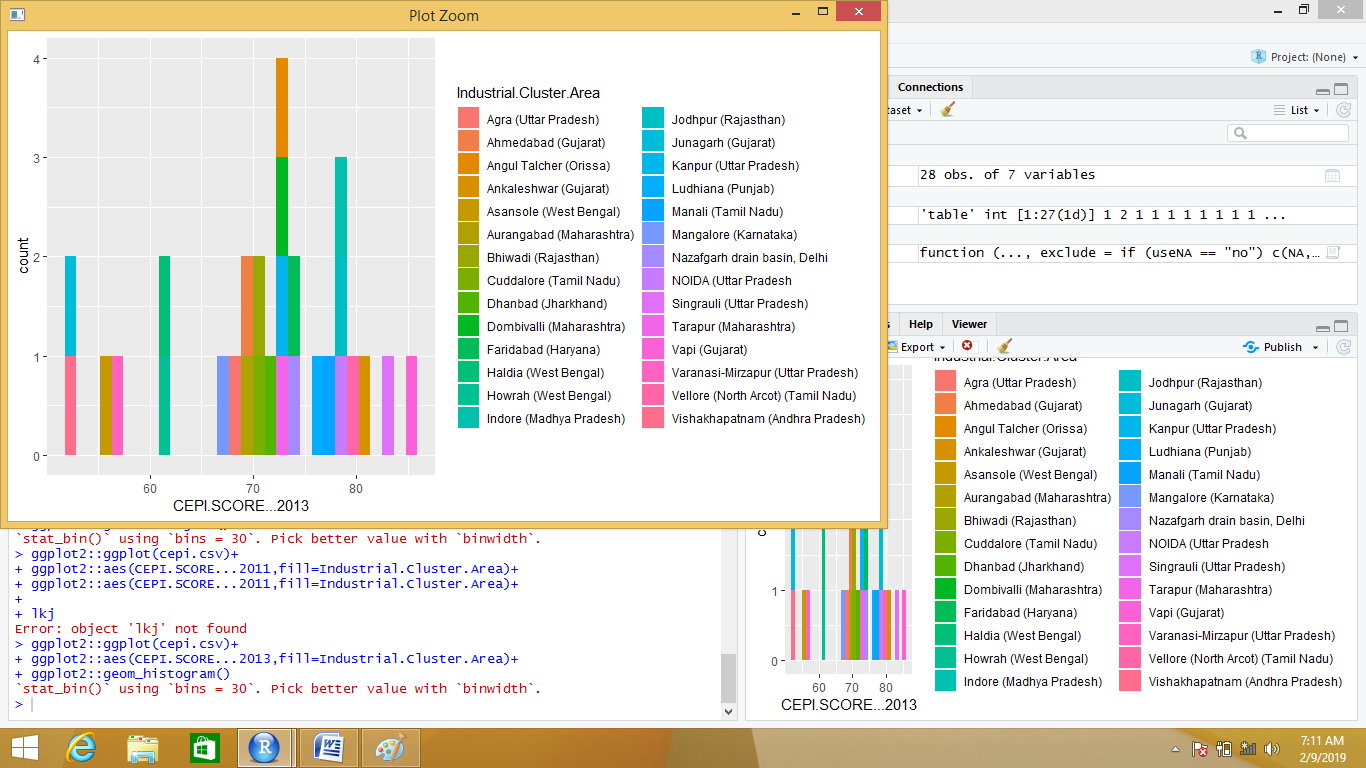
**Figure 13:ggplot of CEPI score 2009**





**Figure 14:ggplot of CEPI score 2011**





**Figure 15: ggplot of CEPI score 2013**



METHOD:

In this dataset we apply multiple regression .  In multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression is −

y = a + b1x1 + b2x2 +...bnxn

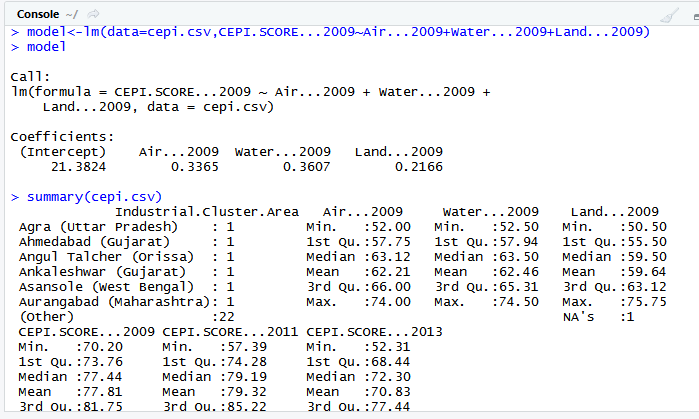
* **y** is the response variable.
* **a, b1, b2...bn** are the coefficients.
* **x1, x2, ...xn** are the predictor variables.

We create the regression model using the **lm()** function. The model determines the value of the coefficients using the input dataset(cepi.csv). Next we can predict the value of the response variable(CEPI SCORE of 2009,11&13) for a given set of predictor variables(air, water and land pollution) using these coefficients. Consider the data set "cepi.csv" available in the R environment. It gives a comparison between different car models in terms OfAir…2009,water…2009,land…2009,cepiscore2009,cepiscore2011,cepiscore2013.some more parameters.

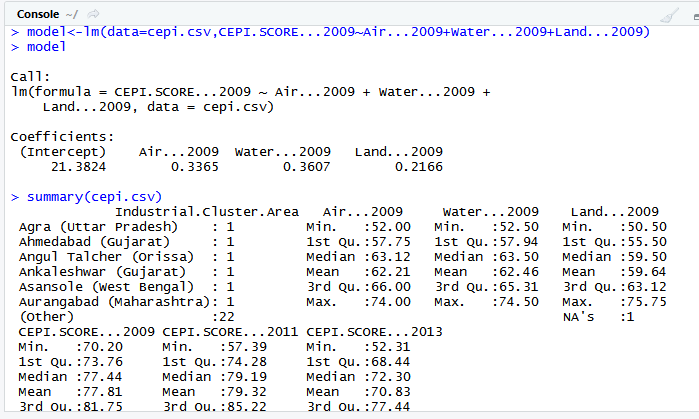
The goal of the model is to establish the relationship between CEPISCORE as a response variable with Air..2009,water..2009,land..2009 as predictor variables. We create a subset of these variables from the cepi.csv data set for this purpose.

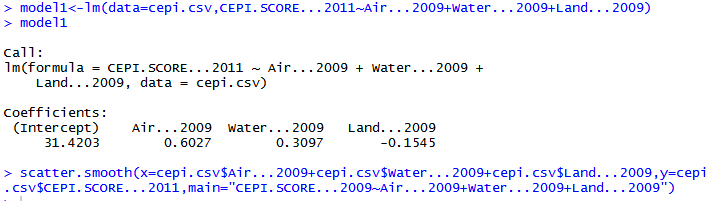
Results:

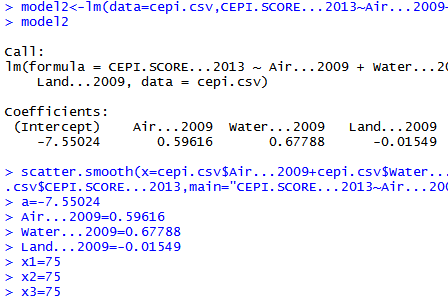
[1].summary of dataset: we summarize the mean, median,mode,1st quartile ,3rd quartile, max,min for each data.



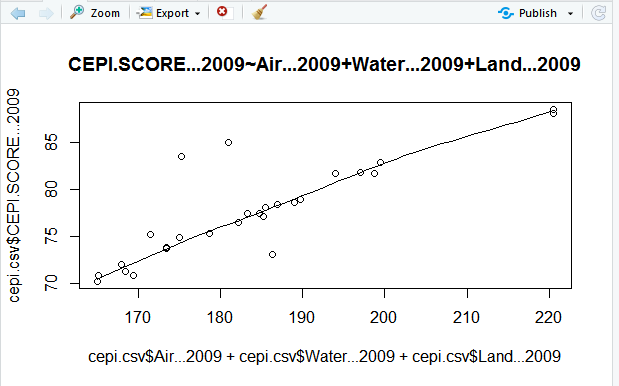
### [2].Create Relationship Model & get the Coefficients:



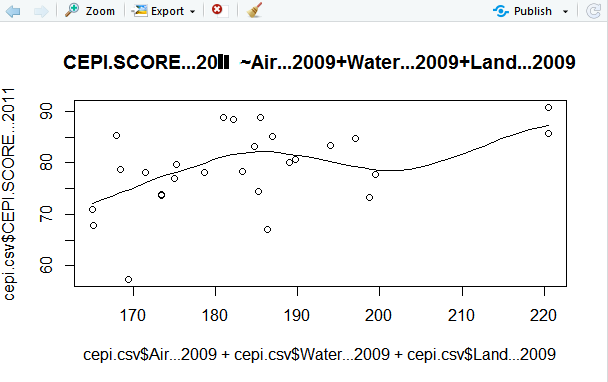




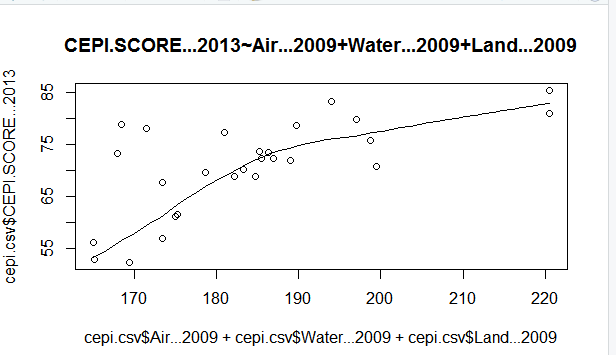
[3]. scatter plot of CEPI score 2009,2011 and 2013 with air, water and land pollution :



**Figure 16:scatter plot of CEPI** **score2009**



**Figure 17:scatter plot of CEPI** **score2011**

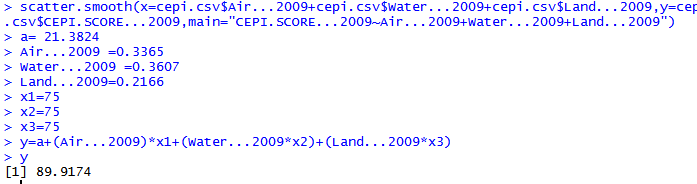


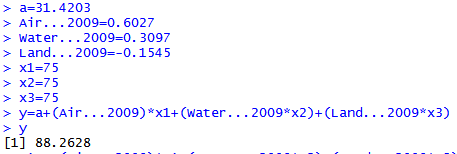
**Figure 18:scatter plot of CEPI** **score2013**

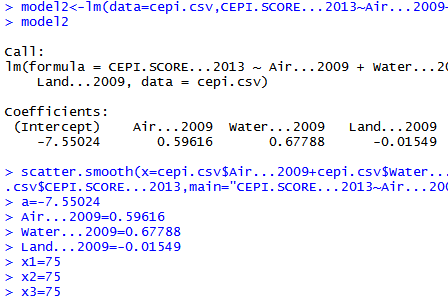
[4] Prediction: Based on the above intercept and coefficient values, we create the mathematical equation.

Y = a+XAir.2009.x1+XWater.2009.x2+XLand2009.x3

We can use the regression equation created above to predict the CEPI score when a new set of values for , Air2009,water2009,land2009 is provided. For a CEPI score with Air2009 = 75, Water2009 = 75and land2009 = 75 the predicted CEPI SCORE .







D:\technical seminar\ts18.PNG

CONCLUSION: In this paper we predict CEPI score according historical data using multiple regression techniques. When we plot the histogram we can see predictive value. According to CEPI score of 2009,2011 and 2013 the range is:89 to 90,88 to 89 and 86 to 87. So after getting predictive value we can say that CEPI score will decrease. Because the correlation coefficient Of CEPI score are decreased and thus show negative relationship. So we can say that in the upcoming year, CEPI score of critical polluted industrial area will decrease. And therefore according to the prediction these industries will be less harmful for environment.

REFERENCES:

[1] Arora, P.N. and Arora, S. (2009). Statistics for Management. New Delhi: S. Chand & Company Limited. ISBN: 81-219-2285-2.

[2] Allen, D. M. and Cady, F. B. (1982), ***Analyzing Experimental Data by Regression***, Belmont, CA: Lifetime Learning Publications.

[3] Bock, R. D. (1975), ***Multivariate Statistical Methods in Behavioral Research***, New York: McGraw-Hill.

[4] Morrison, D. F. (1976), ***Multivariate Statistical Methods***, Second Edition, New York: McGraw-Hill.

[5] Mosteller, F. and Tukey, J. W. (1977), ***Data Analysis and Regression***, Reading, MA: Addison-Wesley.

[6] [cpcb.nic.in/comprehensive-environmental-pollution-index-cepi/.](http://cpcb.nic.in/comprehensive-environmental-pollution-index-cepi/)

[[7]](http://cpcb.nic.in/comprehensive-environmental-pollution-index-cepi/)[cpcbenvis.nic.in/industrial\_pollution.html.](http://cpcbenvis.nic.in/industrial_pollution.html)