**Acknowledgement**

      We take to opportunity to express our gratitude to each one of them. Here by we thank our HOD and department of CSE for giving this opportunity to do a mini project in this semester.

We express our deepest gratitude to our, Mrs. Chaithra H V, Department of Computer Science and Engineering for her valuable suggestions, that enabled us to complete this project successfully.

       We also want to thank the department and faculty for being highly approachable. From this project we could understand the application of algorithms in our day to day life.

**ABSTRACT**

Predictive analytics encompasses a variety of statistical techniques from [predictive modelling](https://en.wikipedia.org/wiki/Predictive_modelling), [machine learning](https://en.wikipedia.org/wiki/Machine_learning), and [data mining](https://en.wikipedia.org/wiki/Data_mining) that analyze current and historical facts to make predictions about future or otherwise unknown events.

In business, predictive models exploit [patterns](https://en.wikipedia.org/wiki/Pattern_detection) found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding [decision making](https://en.wikipedia.org/wiki/Decision_making) for candidate transactions.

The defining functional effect of these technical approaches is that predictive analytics provides a predictive score (probability) for each individual (customer, employee, healthcare patient, product SKU, vehicle, component, machine, or other organizational unit) in order to determine, inform, or influence organizational processes that pertain across large numbers of individuals, such as in marketing, credit risk assessment, fraud detection, manufacturing, healthcare, and government operations including law enforcement.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SLNO** | **CONTENTS** | **PAGE NUMBER** |
| 1. | INTRODUCTION | 4 |
| 2. | WORKING | 5 |
| 3. | CASE STUDY | 6 |
| 4. | PSEUDOCODE | 7-8 |
| 5. | CODE | 9 |
| 6. | RESULT | 10-11 |
| 7. | CONCLUSION | 12 |

**1. INTRODUCTION**

[Predictive models](https://en.wikipedia.org/wiki/Predictive_modeling) are models of the relation between the specific performance of a unit in a sample and one or more known attributes and features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models in many areas, such as marketing, where they seek out subtle data patterns to answer questions about customer performance, or fraud detection models. Predictive models often perform calculations during live transactions, for example, to evaluate the risk or opportunity of a given customer or transaction, in order to guide a decision. With advancements in computing speed, individual agent modelling systems have become capable of simulating human behaviour or reactions to given stimuli or scenarios.

The available sample units with known attributes and known performances are referred to as the "training sample". The units in other samples, with known attributes but unknown performances, are referred to as "out of [training] sample" units. The out of sample units bear no chronological relation to the training sample units. For example, the training sample may consists of literary attributes of writings by Victorian authors, with known attribution, and the out-of sample unit may be newly found writing with unknown authorship; a predictive model may aid in attributing a work to a known author. Another example is given by analysis of blood splatter in simulated crime scenes in which the out of sample unit is the actual blood splatter pattern from a crime scene. The out of sample unit may be from the same time as the training units, from a previous time, or from a future time.

**2. Analytical Technique Used**

Linear regression model:

The [linear regression model](https://en.wikipedia.org/wiki/Linear_regression_model) analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. This relationship is expressed as an equation that predicts the response variable as a linear function of the parameters. These parameters are adjusted so that a measure of fit is optimized. Much of the effort in model fitting is focused on minimizing the size of the residual, as well as ensuring that it is randomly distributed with respect to the model predictions.

The goal of regression is to select the parameters of the model so as to minimize the sum of the squared residuals. This is referred to as [ordinary least squares](https://en.wikipedia.org/wiki/Ordinary_least_squares) (OLS) estimation and results in best linear unbiased estimates (BLUE) of the parameters if and only if the [Gauss-Markov](https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem) assumptions are satisfied.

Once the model has been estimated we would be interested to know if the predictor variables belong in the model—i.e. is the estimate of each variable's contribution reliable? To do this we can check the statistical significance of the model's coefficients which can be measured using the t-statistic. This amounts to testing whether the coefficient is significantly different from zero. How well the model predicts the dependent variable based on the value of the independent variables can be assessed by using the R² statistic. It measures predictive power of the model i.e. the proportion of the total variation in the dependent variable that is "explained" (accounted for) by variation in the independent variables.

**3. CASE STUDY**

* In this case we have taken an example of a classification of people based on their injury and deaths due to vehicular accidents.
* Classification of people as the death count.
* Implementation is done is R programming language.

**DATA SET**

|  |  |
| --- | --- |
| **YEAR** | **DEATHS** |
| **2005** | **560250** |
| **2006** | **602230** |
| **2007** | **627784** |
| **2008** | **643053** |
| **2009** | **641118** |
| **2010** | **662025** |
| **2011** | **653897** |
| **2012** | **647925** |
| **2013** | **681902** |
| **2014** | **693484** |
| **2015** | **705065** |
|  | |

**4. PSEUDOCODE**

**# Multiple Linear Regression Example**   
fit <- lm(y ~ x1 + x2 + x3, data=mydata)  
summary(fit) # show results

**# compare models**  
fit1 <- lm(y ~ x1 + x2 + x3 + x4, data=mydata)  
fit2 <- lm(y ~ x1 + x2)  
anova(fit1, fit2)

Let us start with loading the data set in the R.

**#Read data set**

data <- read.excel("data.xlsx")

Here dependent variable is Profit and interdependent variable is Population. So let us set dependent variable Y and independent variable x.

**#Dependent variable**

y <- data$profit

#Independent variable

x <- data$population

The objective of linear regression is to minimize cost function

[costFunction](https://i1.wp.com/pingax.com/wp-content/uploads/2013/11/costFunction.png)

Where hypothesis hΘ(x) is given by the linear model

[LinearModel](https://i2.wp.com/pingax.com/wp-content/uploads/2013/11/LinearModel.png)

To take into account the intercept term Θ0, we add an additional first column to x and set it to all ones. This allows us to treat Θ0 as simply another feature.

Let us first add ones to x, also initialize Θ0 and Θ1 to zero and calculate cost using above equation.

**#Add ones to x**

x <- cbind(1,x)

**# initalize theta vector**

theta<- c(0,0)

**# Number of the observations**

m <- nrow(x)

**#Calculate cost**

cost <- sum(((x%\*%theta)- y)^2)/(2\*m)

For initial value of the theta parameter cost is 32.07, our objective is to minimize cost by updating the values of the thetas. One way to do this is to use batch gradient descent algorithm. We update the values of the thetas by iterating following equation.

**5. CODE**

library(readxl)

input <- read\_excel("C:/Users/Prajwal/Desktop/DW/input.xlsx")

View(input)

# Create the data for the chart.

v <- c(55000,64000,59000,69000,54541)

# Give the chart file a name.

png(file = "C:/Users/Prajwal/Desktop/DW/line1.jpg")

# Plot the bar chart.

plot(v,type = "o", col = "red", xlab = "Year", ylab = "Deaths",xlim = c(2004,2016), ylim = c(55000,75000),

main = "Vehicular Accident in India")

# Create the predictor and response variable.

x <- c(2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015)

y <- c(560250,602230,627784,643053,641118,662025,653897,647925,681902,693484,705065)

relation <- lm(y~x)

# Give the chart file a name.

png(file = "linearregression.png")

#predict

a <- data.frame(x = 2013)

result <- predict(relation,a)

print(result)

# Plot the chart.

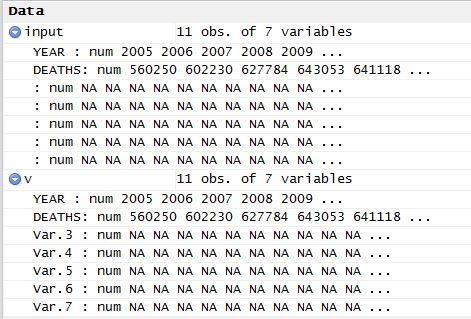
plot(y,x,col = "blue",main = "DEATH AND ACCIDENT REGRESION",

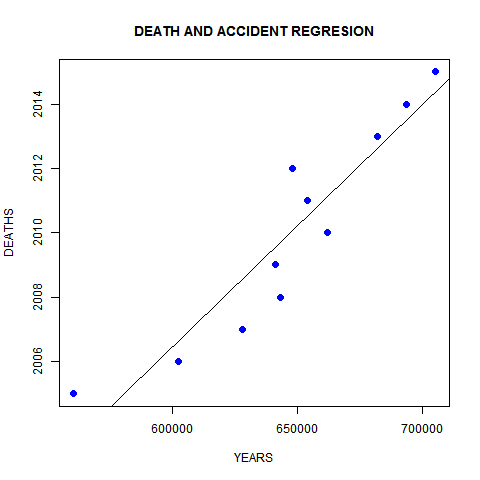
abline(lm(x~y)),cex = 1.3,pch = 16,xlab = "YEARS",ylab = "DEATHS")

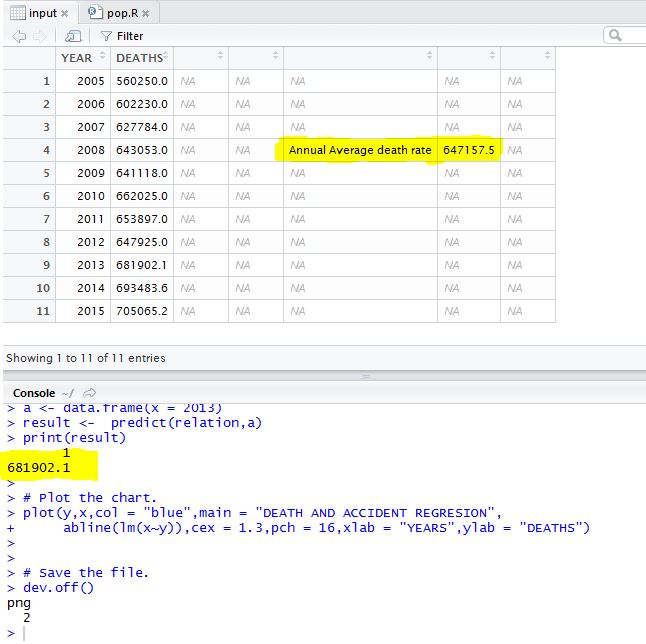
# Save the file.

dev.off()

**6. RESULTS**

****



****

**7. CONCLUSION**

Success is better decision making. Previously with low volumes of data, intuitive decision making would work. As the data size has grown to incredible proportions, human ability to make completely intuitive decisions has been reduced. As a result, data-driven decision making has become more prevalent to ensure a reasonable path for success. This situation makes sense as it is easy to see that data are not diminishing but rather increasing.

These data-driven decisions are based often on quantitative models created using a typical closed-loop process: a cycle. The cycle described in this book includes:

* Problem definition and identification
* Design and build of an analytical framework, if there isn’t one available
* Data management, reporting, and visualization
* Analysis to produce models
* Execution and testing
* Feedback

**8. REFERENCES**

* Wikipedia.org
* Tutorialspoint.org
* http://www.kdnuggets.com