**Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

The proportionality of Height with Lung Capacity

Submitted by

Abhishek Narayan 1MS15CS004

Abishek Padaki 1MS15CS005

Ankitha V Nandan 1MS15CS019

Divya Tantri 1MS15CS042

Ahish SM 1MS14CS010

*Bachelor of Engineering in Computer Science & Engineering*

Under the guidance of:

Sowmya B.J.

Designation Assistant Professor

Dept name Computer Science Engineering

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

<www.msrit.edu>, **20177**

INTRODUCTION

CLASSIFICATION

Classification is one of the data mining functionalities and the functionalities also includes

Characterization, Discrimination, Association Analysis, Prediction, Clustering, Outlier Analysis,

Evolution and Deviation Analysis

Classification analysis is the organization of data in given classes. Also known as supervised

classification, the classification uses given class labels to order the objects in the data collection.

Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

WHY IS CLASSIFICATION IMPORTANT?

After starting a credit policy, the Our Video Store managers could analyze the customer

behaviours, their credit, and label accordingly the customers who received credits with three

possible labels &quot;safe&quot;, &quot;risky&quot; and &quot;very risky&quot;.

The classification analysis would generate a model that could be used to either accept or reject

credit requests in the future.

PREDICTION

Prediction is another important data mining functionality. There are two major types of

predictions: one can either try to predict some unavailable data values or pending trends, or

predict a class label for some data. The latter is tied to classification. Once a classification model

is built based on a training set, the class label of an object can be foreseen based on the attribute

values of the object and the attribute values of the classes.

WHY IS PREDICTION USED?

Prediction is however more often referred to the forecast of missing numerical values, or

increase/ decrease trends in time related data. The major idea is to use a large number of past

values to consider probable future values.

CLASSIFICATION METHODS

 Classification according to the type of data source mined:

According to the type of data handled such as spatial data, multimedia data, time-

series data, text data, World Wide Web, etc.

 Classification according to the data model drawn on:

Based on the data model involved such as relational database, object-oriented

database, data warehouse, transactional, etc.

 Classification according to the king of knowledge discovered:

Based on the kind of knowledge discovered or data mining functionalities, such as

characterization, discrimination, association, classification, clustering, etc. Some

systems tend to be comprehensive systems offering several data mining

functionalities together.

 Classification according to mining techniques used:

Categorizes data mining systems according to the data analysis approach used such

as machine learning, neural networks, genetic algorithms, statistics, visualization,

database-oriented or data warehouse-oriented, etc.

degree of user interaction involved in the data mining process such as query-driven

systems, interactive exploratory systems, or autonomous systems.

PREDICTION

The prediction of continuous values can be modelled by statistical techniques of regression. Many problems can be solved by linear regression.

Multiple regression is an extension of linear regression involving more than one predictor variable.

In statistics nonlinear regression is a form of regression analysis in which observational data are

modelled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation.

DISADVANTAGES OF NON-LINEAR REGRESSION

Nonlinear regression can be a powerful alternative to linear regression but there are disadvantages. nonlinear regression allows a nearly infinite number of possible functions, it can be more difficult to setup. It is easier to use linear regression model.

DATA SET DESCRIPTION

The Source of our Data Set

The Lung Capacity data set, was sourced from the internet.

ATTRIBUTE DESCRIPTION

• AGE: The Attribute age, helps us determine the expected

lung capacity of a male/female of a particular age, which

is then used to compare with, the male/female’s lung

capacity we are recording in our data set.

• LUNG CAP: Lung Capacity, Suggests to us, the volume of oxygen that

can be stored, by a person’s lung's with a single breath.

• HEIGHT: Provides us, with the height of the person,

who’s data we’re recording. Data such as his Lung

capacity, and whether he’s a Smoker.

• GENDER: Gender is one of the key factors in helping us

correlate a male/female’s average lung capacity, with the

lung capacity of the male/female being recorded in our

dataset.

• SMOKE: This attribute Suggests to us, if the male/female

is a Tobacco smoker or a non smoker. This is one of the

main factors that is responsible in determining the

adverse effects of smoking on a male/female’s lung

capacity.

SIZEOF DATASET

No of bytes : 60 KB

No of tuples : 655

INFERENCES DRAWN

1. The first inference we are drawing from our analysis on the lung capacity data set is, lung capacity is directly proportional to the height of the male/female.

2. The second inference we are drawing is, the amount of

smoking is inversely proportional to the male/female’s lung capacity.

3. The third inference we are making is, the age of a male/female is inversely proportional to their lung capacity.

4.From height dataset,we can approxmiate average height in that country,find out the lifestyle,etc.

5.From lung capacity data,we can use it in Medical Fields for lung trasnplants,operations,etc.

6.From the smoking/non smoking values,we can predict the brand's profit and predict its sales.

NON-LINEAR REGRESSION

* Polynomial Regression can be modeled by adding Polynomial terms to the basic Linear model
* By applying transformations to the variables,we can convert the non-linear model into a linear one that can be solved by the method of least squares.

TRANSFORMATION OF A POLYNOMIAL REGRESSION MODEL TO A LINEAR REGRESSION MODEL

Consider a cubic polynomial relationship given by:

y=w0+w1x+w2x2+w3x3

To convert this equation to form,we define new variables:

x1=x

x2=x2

x3=x3

First equation can then be converted to linear form by applying the above assignments,resulting in the equation

y=w0+w1x+w2x2+w3x3

which is easily solved by the method of least squares.

CODE

summary(LungCapData2)

library(ggplot2)

ggplot()+

geom\_point(aes(x=LungCapData2$Height, y=LungCapData2$LungCap), color='red')+

ggtitle('Lung Cap Vs. Height')+

xlab('Height')+

ylab('LungCap')

linear\_regressor = lm(LungCapData2$LungCap ~ LungCapData2$Height)

linear\_regressor

summary(linear\_regressor)

#plot(linear\_regressor)

ggplot()+

geom\_point(aes(x=LungCapData2$Height, y=LungCapData2$LungCap), color='red')+

geom\_line(aes(x=LungCapData2$Height, y=predict(linear\_regressor)), color='darkgreen',lwd=1)+

ggtitle('Lung Cap Vs. Height')+

xlab('Height')+

ylab('LungCap')

quad\_regressor = lm(LungCapData2$LungCap ~ poly(LungCapData2$Height, degree = 2, raw = T))

quad\_regressor

summary(quad\_regressor)

#plot(quad\_regressor)

ggplot()+

geom\_point(aes(x=LungCapData2$Height, y=LungCapData2$LungCap), color='red')+

geom\_line(aes(x=LungCapData2$Height, y=predict(linear\_regressor), color = "Linear"), lwd=1)+

geom\_line(aes(x=LungCapData2$Height, y=predict(quad\_regressor), color = "Quad"), lwd=1)+

scale\_color\_manual("", breaks = c("Linear", "Quad"),values = c("Linear"="darkgreen", "Quad"="purple"))+

ggtitle('Lung Cap Vs. Height')+

xlab('Height')+

ylab('LungCap')

cube\_regressor = lm(LungCapData2$LungCap ~ poly(LungCapData2$Height, degree = 3, raw = T))

cube\_regressor

summary(cube\_regressor)

ggplot()+

geom\_point(aes(x=LungCapData2$Height, y=LungCapData2$LungCap), color='red')+

geom\_line(aes(x=LungCapData2$Height, y=predict(linear\_regressor), color = "Linear"),lwd=1)+

geom\_line(aes(x=LungCapData2$Height, y=predict(quad\_regressor), color = "Quad"), lwd=1)+

geom\_line(aes(x=LungCapData2$Height, y=predict(cube\_regressor), color = "Cube"), lwd=1)+

scale\_color\_manual("", breaks = c("Linear", "Quad", "Cube"),values = c("Linear"="darkgreen", "Quad"="purple", "Cube"="cyan"))+

ggtitle('Lung Cap Vs. Height')+

xlab('Height')+

ylab('LungCap')

anova(quad\_regressor,cube\_regressor)

#deg4\_regressor = lm(LungCapData2$LungCap ~ poly(LungCapData2$Height, degree = 4, raw = T))

#deg4\_regressor

#summary(deg4\_regressor)

#ggplot()+

# geom\_point(aes(x=LungCapData2$Height, y=LungCapData2$LungCap), color='red')+

#geom\_line(aes(x=LungCapData2$Height, y=predict(linear\_regressor), color = "Linear"),lwd=1)+

#geom\_line(aes(x=LungCapData2$Height, y=predict(quad\_regressor), color = "Quad"), lwd=1)+

#geom\_line(aes(x=LungCapData2$Height, y=predict(cube\_regressor), color = "Cube"), lwd=1)+

#geom\_line(aes(x=LungCapData2$Height, y=predict(deg4\_regressor), color = "Degree4"), lwd=1)+

#scale\_color\_manual("", breaks = c("Linear", "Quad", "Cube", "Degree4"),values = c("Linear"="darkgreen", "Quad"="purple", "Cube"="cyan", "Degree4"="Magenta"))+

#ggtitle('Lung Cap Vs. Height')+

#xlab('Height')+

#ylab('LungCap')

library(caTools)

#set.seed(47)

split <- sample.split(LungCapData2$LungCap, SplitRatio = 0.8)

split

training\_set <- subset(LungCapData2,split==T)

training\_set

testing\_set <- subset(LungCapData2,split==F)

testing\_set

poly\_regressor = lm(formula = LungCap ~ poly(Height, degree = 2, raw = T), data = training\_set)

poly\_regressor

summary(poly\_regressor)

ggplot()+

geom\_point(aes(x=training\_set$Height,y=training\_set$LungCap),color='red')+

geom\_line(aes(x=training\_set$Height,y=predict(poly\_regressor,newdata = training\_set)),color='blue', lwd=1)+

ggtitle('LungCap vs Height(trainingset)')+

xlab('Height')+

ylab('LungCap')

ggplot()+

geom\_point(aes(x=testing\_set$Height,y=testing\_set$LungCap),color='green')+

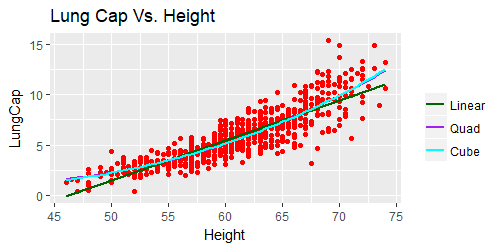
geom\_line(aes(x=training\_set$Height,y=predict(poly\_regressor,newdata = training\_set)),color='blue',lwd=1)+

ggtitle('LungCap vs Height(testingset)')+

xlab('Height')+

ylab('LungCap')

SNAPSHOT OF CODE WITH GRAPH



Linear, Quadratic and Cubic graph:

X axis -> Height

Y axis -> Lung Capacity

X acts as the response variable

Y acts as the explanatory variable

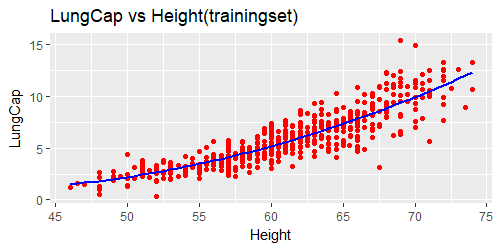
Cubic equation gives a best fitted line than quadratic equation gives a best fitted line than linear equation. Hence we can conclude that as the degree of the equation increases the accuracy increases.

Linear Equation: ax + b

Quadratic Equation: ax^2 + bx + c

Cubic Equation: ax^3 + bx^2 + cx + d

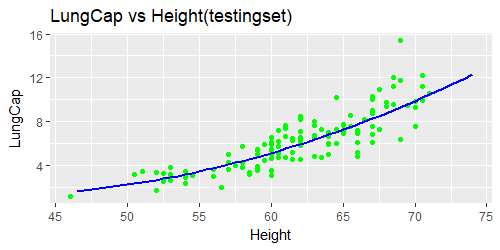
a,b, c,d are the co-efficients



X axis -> Height

Y axis -> Lung Capacity

The above graph is plotted for a training set that has been taken from the Lung capacity data we have. This is the set of tuples on which the data model has been constructed on.



X axis -> Height

Y axis -> Lung Capacity

The graph above is plotted based on the testing set, in order to check the working of the model that we have constructed.

SOCIAL IMPACT

Statistical justification was defined as the presence of effect modification by race/ethnicity among never-smoking participants without respiratory disease or symptoms and was tested with interaction terms for race/ethnicity (× age and height) in regression models. There was no evidence of effect modification by race/ethnicity for forced expiratory volume in 1 second, forced vital capacity, or the forced expiratory volume in 1 second/forced vital capacity ratio among white, African-American, and Mexican-American men or women on an additive scale or a log scale. Interaction terms for race/ethnicity explained less than 1% of variability in lung function. The mean lung function for a given age, gender, and height was the same for whites and Mexican Americans but was lower for African Americans. Determination of the most appropriate forms of the equations for height, for sitting height, and for Quetelet Index was accomplished for each of the four groups by examination of plots of moving averages of the dependent variable versus the independent variable and by polynomial regression including the linear and the quadratic tests. When necessary to avoid problems of co-linearity, the independent variables were centered prior to regression analysis by subtracting the sample mean of that variable from each persons value .

As with height, most lung volumes had significant linear relationships with sitting height without

statistically significant quadratic relationships. In general the males and the Whites had larger sitting heights coefficients than did the females and the Blacks

. The degree of difference among the coefficients was similar to that for height when comparing

races, but was larger than that for height when comparing genders.

Since inhaling cigarette smoke has been shown to produce acute changes in the lung including

alterations in resistance to airflow, cough, and irritation of the airway, the early stage of smoking

might affect the respiratory function of youths. However, there have been few studies which have

investigated the effect of smoking on pulmonary function in adolescents.

Therefore, to clarify the effect of smoking on the respiratory function of smoking and non-smoking youths, we measured and compared their chest expansion, the lung function test using a spirometer, and respiratory muscle strength to learn more about the dangers of cigarette smoking.