

MAHARASHTRA EDUCATION SOCIETY’S

ABASAHEB GARWARE COLLEGE

‘**Efficacy of Lungs: A Statistical Study’**



SUBMITTED TO

DEPARTMENT OF STATISTICS IN THE FULLFILMENT OF T.Y.B.Sc.

2019-2020

**CERTIFICATE**

This is to certify that the project report entitled ‘**Efficacy of Lungs : A Statistical Study’** is being submitted by **Madhura** **Pote** (9407), **Rujuta Kulkarni** (9409) , **Janaki Inamdar**  (9410) , **Aparna Patil** (9411) and **Jyoti Wagh**  (9412) as a partial fulfilment for the award of the degree of the Bachelor of Science (B.Sc.). This is a record of bonafide work carried out by them under supervision and guidance.

Prof. Rutuja Joshi Prof. Sandesh Kurade

**Project Guide**  **Head Dept. of Statistics**

Place: Pune

Date:

**ACKNOWLEDGEMENT**

We have satisfaction upon completion of our project work ‘**Efficacy of Lungs : A Statistical Study**’ at the Department of Statistics of MES Abasaheb Garware College, Pune during the academic year 2019-2020.

We express our gratitude towards **Prof. Rutuja Joshi** and **Prof. Apoorva Pali** for their valuable guidance during completion of this project work.

We are thankful to **Prof. Sandesh Kurade** (Head, Department of Statistics) for providing all the necessary help, guidance and environment and co-operation required for the same. Also, the teaching staff of Department of Statistics for their valuable support and co-operation during completion of this project.

We would also like to thank our classmates and family members and also nearby people for their help and encouragement in our project.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **SR. NO** | **TITLE** | **PAGE NO** |
| 1 | Motivation | 5 |
| 2 | Abstract | 6 |
| 3 | Statistical definitions | 7 |
| 4 | Technical statement | 13 |
| 5 | Statistical statement | 13 |
| 6 | Methodology and data collection | 14 |
| 7 | Data | 17 |
| 8 | Exploratory Data Analysis | 18 |
| 9 | Statistical analysis | 29 |
| 10 | Conclusions | 53 |
| 11 | Scope and Limitations | 54 |
| 12 | References | 55 |
| 13 | Appendix | 56 |

**MOTIVATION**

In day to day life we go to small hill like “Parvati” or “taljai” or any fort like “sinhagad”. There are many people of different age, height, and weight. So, when we observe them, we notice that most of the people who have age greater than 40 or 50 feel very exhausted even for small ascent. And for kids, teenagers or people who have age less than 40 don’t feel that much exhausted for same or more ascent. So, by observing this we were eagerly want to know what is the reason behind it. Initially we just knew that it is something related to breathing rate. After doing more search on it we came to know that it was related to the lung capacity.

Then many questions raised in mind like how to calculate lung capacity, what is its range, whether our lung capacity is within range or not etc. After reading various published papers and available information on the internet we found the method to measure the lung capacity as “lung capacity using water spirometer”.

Everyone including teachers found that this topic and its method is different and attractive. So then, we decided this topic for our project and started the work on it.

Actually, in medical field the lung capacity can be calculated using the mechanical spirometer but that spirometer is very costly and it gives the required values of inspiratory reserve volume (IRV), expiratory reserve volume (ERV) and vital capacity separately. In order to obtain the total lung capacity, we have to add IRV, ERV and vital capacity so it is little bit time consuming. Hence, in order to reduce time and cost we have used the water spirometer as it provides the actual lung capacity directly and it is cheapest among all and anyone can make this water spirometer at home.

**ABSTRACT**

The total lung capacity of an individual depends upon the various factors like age, height, weight, profession, ethnicity, etc. overall, we have total data set of 310 people consisting the variables mentioned as above. The average value of lung capacities of males and females are obtained by one sample t test and are **2.85 Litre for females and 3.36 litre for males.**

The distribution of lung capacity is **non normal distribution** proved by Shapiro test perform in the statistical analysis. The relationship between lung capacity and age, height, weight is nonlinear. We have built the best possible regression model for obtaining the lung capacity of an individual. For obtaining the best regression model we use the weighted least square method, Box-Cox method and all required transformations.

The average lung capacities differ significantly for all age groups, height groups, profession groups, and pulse rate groups but remain similar for all BMI groups and this was done by Kruskal Walli H test. We have given the graphical base for Mann Whitney Test using box plot.

There is **strong association** present between **lung capacity and the profession** but this data set does not show any association between smoker’s lung capacity and non-smoker’s lung capacity.

**KEYWORDS:** Regression analysis, Box-Cox method, Weighted least square method, Kruskal Walli H test, Mann Whitney test, Shapiro Wilk test, Kolmogorov Smirnov test,

**STATISTICAL DEFINITIONS**

1. **REGRESSION ANALYSIS**

In statistical analysis the regression analysis is the most powerful technique for estimating the relationship between independent variable often called regressors and dependent variables often called response variables. Using regression analysis one can fit regression model for the purpose of forecasting and prediction. The parameters of regression model can be estimated by ordinary least square method.

ASSUMPTIONS OF LINEAR REGRESSION MODEL: -

* The relationship between response variables and regressor variables should be linear.
* The error term has zero mean.
* The error term has constant variance.
* The errors are uncorrelated.
* The errors are normally distributed.

1. **BOXCOX METHOD**

Box-cox method is used to transform response variable to correct non normality or non-constant variance which is one of the assumptions of linear regression model. The useful class of transformation is called as power transformation where is the parameter to be determined.

The best procedure to estimate is given below –

= , if

= ln y , if

1. **WEIGHTED LEAST SQUARE METHOD**

Weighted least square method is suitable when variance of uncorrelated errors is not constant. It is either increasing or decreasing function of response variable. In this case the linear regression model can be fitted by the method of weighted least square.

In this method of deviation, the deviation produced between the observed and the expected value of response is multiplied by corresponding weights. For using weighted least square method the weights must be known. The residual diagnostic analysis indicates that var () may be function of one of the regressors.

Say, **var () =** in this case weights **=**

If response variable is actually an average of observations at and if all original observations have constant variance then **var () = .** Thus **=**

Sometimes the primary source of error is measurement error and different observations are measured by different instruments of unequal but known accuracy then weights could be chosen inversely proportional to variances of measurement error.

1. **KRUSKAL WALLI H TEST**

In situations where the normality assumption of errors is not justified or failed the Kruskal Wallis H test is an alternative procedure to the F test used in analysis of variance.

The Kruskal Wallis H test is use for testing the equality of treatment means. The test procedure is given below

= The treatments mean does not differ significantly.

vs

= The treatments mean differ significantly.

Test statistic:

H=

Where is number of observations for treatment.

N = total number of observations

If there **are ties** in the observations

**=**

If there are **no ties** in the observations

**=**

**H=**

**Decision criteria:**

If **H ≥** then **reject** accept otherwise. Where α is level of significance.

1. **MANN – WHITNEY TEST**

The Mann-Whitney U test is a non-parametric test that can be used in place of an unpaired t-test. It is used to test the null hypothesis that two samples come from the same population (i.e. have the same median) or, alternatively, whether observations in one sample tend to be larger or smaller than observations in the other or not equal to the observations in the other population.

Test statistic is given by:

Suppose X and Y are two populations

Hypothesis to be tested,

: = against

: <

: >

≠

Let us consider ==.

Let U denote number of X observations below . under the null hypothesis the probability distribution of U is given by:

**U=u)=**  u=0,1, 2…..., min()

Where t= , N= (+)

1. **CRAMER’S V TEST:**

**Cramer’s V** is a measure of [association](https://en.wikipedia.org/wiki/Association_(statistics)) between two [nominal variables](https://en.wikipedia.org/wiki/Nominal_data#Nominal_scale). It is based on [Pearson's chi-squared statistic](https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test#Calculating_the_test-statistic).

**Test statistic:**

**V=**

Where q is min (row and column)

Testing criteria:

|  |  |
| --- | --- |
| **Value of V** | **Interpretation** |
| V>0.25 | Very strong |
| V>0.15 | Strong |
| V>0.10 | Moderate |
| V>0.05 | Weak |
| V>0 | No or very weak |

1. **SHAPIRO-WILK TEST**

The test procedure is given below:

= The observations (, , …... ) came from normal distribution.

**Vs**

=The observations (,,…... ) does not came from normal distribution.

Test statistic:

**W=**

Where order statistic, i.e. the smallest number in the sample.

**= sample mean**

The coefficients are given by,

**(,…….=,**

Where C is vector norm

**C= =**

And the **vector m=**

is made of the expected values of the order statistics of independent and identically distributed random variates sampled from the standard normal distribution; finally, {\displaystyle V}V is the covariance matrix of those normal order statistics.

**Rejection criteria:**

**If p value is less than the chosen value of level of significance α then we reject null hypothesis.**

1. **MULTIPLE LINEAR REGRESSION ANALYSIS**

A regression model which involves more than one regressor variable is called multiple linear regression model. In general, the response Y may be related to K regressors then, model becomes,

Y=+++…...++

Where, is the regressor’s coefficient.

Assumptions:

1. Errors are uncorrelated with mean zero and constant variance i.e.
2. Errors are normally distributed.
3. Rank of matrix of regressors is K

**TECHNICAL STATEMENT**

This project aims at examining the relationship between the lung capacity of an individual with various factors like age, height, weight, pulse rate, profession etc.

* **Lung Capacity (Breathing capacity):** It refers to the volume of air in the lungs at different phases of the respiratory cycle and measured in litres. The average total lung capacity of an adult human male is about 6 litres of air. The average human respiratory rate is 30-60 breaths per minute at birth, decreasing to 12-20 breaths per minute in adults. Human lung size is determined by genetics, gender, and height. At maximal capacity, an average lung can hold almost six litres of air, but lungs do not usually operate at maximal capacity. Air in the lungs is measured in terms of lung capacities. Capacity is any two or more volumes (for example, how much can be inhaled from the end of a maximal exhalation).
* Pulse rate:

The pressure wave that is generated by the contraction of the heart and travels through the arterial system is called a pulse. The rate at which pulses occur is the Pulse Rate (PR). It is typically measured as the average number of pulses that occur per minute, and is expressed in units of /min – pulses per minute.

For normal adult PR is considered to be in the range of 60 to 100 /min.

**STATISTICAL STATEMENTS**

With reference to the data set attached with the report, the objective is to observe the effect of various factors like age, height, weight, pulse rate, profession on lung capacity using various statistical tests like Kruskal Walli, Mann Whitney test. Also, to obtain the best regression model for forecasting purpose using multiple linear regression analysis.

It is of interest to check the association between the lung capacity with any of the regressors using chi square and Cramer’s v test.

**METHODOLOGY AND DATA COLLECTION**

Since this project involves the execution of an experiment, it came with its own set of challenges and practical difficulties. We conducted this experiment at various places such as in our college campus, societies etc.

**Instruments required for collecting observations of lung capacity of a person:**

* 5L water bottle, pipe, water tub and water
* measuring cylinder
* stopwatch or clock
* tape

**Procedure of collecting observations of lung capacity of a person:**

* Measure the pulse rate of an individual before performing an experiment.
* We made 20 markings on 5L water bottle, with each line shows marking of 250 ml.
* 5L water bottle should be calibrated for measurement purpose: for calibrated, fill it with 250ml water, until it has no air bubble in it
* Fill water tub with water.
* Then the calibrated 5L water bottle placed into water tub in vertical inverted position in such a way that, no water in the calibrated water bottle should be mixed with the water tub.
* After that, one side of pipe was inserted into the inverted calibrated 5L water bottle and the other side of the pipe was given to the individual whose lung capacity will be measure.
* Now the individual should exhale totally in only one breath through that particular pipe so that water gets displaced from the water bottle in to the water tub.
* Then we have removed the water bottle from the water tub but have taken care that no water was displaced from bottle at the time of removal and we have measured the lines up-to that line where the remaining water surface meets after completing the experiment.
* So, the total counted lines multiplied by 250 gives the lung capacity of an individual.

**A GLIMPSE….**

Requirements for making Spirometer

Gender: FEMALE

AGE: 53 years

HEIGHT: 1.64592 metre

LUNG CAPACITY: 2.2L

GENDER: MALE

AGE:33

HEIGHT:1.70688 metre

LUNG CAPACITY: 3.5L



GENDER: FEMALE

AGE:20

HEIGHT:1.6764

LUNG CAPACITY: 4.2L



GENDER: MALE

AGE: 58

HEIGHT:1.81 metre

LUNG CAPACITY: 3.2L



GENDER: FEMALE

AGE:

HEIGHT:1.64592 metre

LUNG CAPACITY: 3.3L



GENDER: MALE

AGE: 28

HEIGHT:1.76784 metre

LUNG CAPACITY: 4.5L

**DATA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr. no** | **Age** | **Height**  **(in m)** | **Weight**  **(in kg)** | **Gender** | **Lung capacity (in litre)** |
| 1 | 15 | 1.65 | 63 | f | 3.30 |
| 2 | 15 | 1.55 | 40 | m | 3.88 |
| 3 | 15 | 1.60 | 42 | m | 2.38 |
| 4 | 16 | 1.64 | 48 | m | 2.90 |
| 5 | 16 | 1.54 | 46 | f | 2.40 |
| 6 | 16 | 1.46 | 48 | m | 2.90 |
| 7 | 16 | 1.71 | 43 | m | 4.10 |
| 8 | 16 | 1.58 | 57 | f | 2.60 |
| 9 | 16 | 1.71 | 55 | m | 3.40 |
| 10 | 17 | 1.87 | 55 | m | 3.95 |
| 11 | 17 | 1.65 | 60 | f | 3.30 |
| 12 | 17 | 1.65 | 46 | m | 3.00 |
| 13 | 17 | 1.69 | 53 | m | 3.90 |
| 14 | 17 | 1.63 | 49 | m | 3.40 |
| 15 | 17 | 1.65 | 51.5 | m | 3.50 |
| 16 | 17 | 1.65 | 54 | f | 3.90 |
| 17 | 17 | 1.83 | 75 | m | 4.20 |
| **.** | **.** | **.** | **.** | **.** | **.** |
| **.** | **.** | **.** | **.** | **.** | **.** |
| **.** | **.** | **.** | **.** | **.** | **.** |
| 303 | 67 | 1.52 | 59 | m | 3.05 |
| 304 | 67 | 1.77 | 65 | m | 3.00 |
| 305 | 69 | 1.48 | 49 | f | 2.63 |
| 306 | 71 | 1.70 | 92.5 | m | 2.50 |
| 307 | 72 | 1.76 | 59 | m | 2.25 |
| 308 | 72 | 1.64 | 69 | m | 3.60 |
| 309 | 72 | 1.56 | 50 | f | 3.10 |
| 310 | 73 | 1.62 | 55 | m | 3.13 |

We have total data set of 310 group of people. For that group we have taken the data of age (in yrs.), height (in m), weight (in kg). Pulse rate were measured at the time of performing an experiment. The lung capacity estimated using water spirometer which is in litres.

**EXPLORATORY DATA ANALYSIS**

1. **Descriptive statistics:**
2. **For age:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **age** | | |
|  | **male** | **female** | **total** |
| **Mean** | 36.6203 | 33.2303 | 34.9581 |
| **Standard Error** | 1.2408 | 1.1242 | 0.8431 |
| **Median** | 34.5000 | 32.0000 | 33.0000 |
| **Mode** | 20.0000 | 20.0000 | 20.0000 |
| **Standard Deviation** | 15.5971 | 13.8600 | 14.8440 |
| **Sample Variance** | 243.2689 | 192.0989 | 220.3574 |
| **Kurtosis** | -0.7952 | -0.6576 | 0.6775 |
| **Skewness** | 0.5057 | 0.5944 | 0.5709 |

**For the data corresponding to age of females the distribution is platykurtic, for age of males it is platykurtic, and for irrespective of gender it is leptokurtic. Average age for male is 37 years and for females is 33 years.**

1. **For lung capacity:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lung capacity** | | | |
|  | **females** | **males** | **total** |
| **Mean** | 2.8530 | 3.3698 | 3.1164 |
| **Standard Error** | 0.0358 | 0.0474 | 0.0332 |
| **Median** | 2.9000 | 3.4000 | 3.0750 |
| **Mode** | 3.0000 | 3.0000 | 3.0000 |
| **Standard Deviation** | 0.4412 | 0.5964 | 0.5855 |
| **Sample Variance** | 0.1946 | 0.3557 | 0.3428 |
| **Kurtosis** | 0.3909 | -0.8253 | -0.5319 |
| **Skewness** | 0.1945 | -0.1571 | 0.2653 |

**For the data corresponding to lung capacity of females the distribution is leptokurtic, for lung capacity of males it is platykurtic, and for irrespective of gender it is platykurtic. Average lung capacity for female is 2.85L, for male 3.36 L.**

1. **For BMI**

|  |  |  |  |
| --- | --- | --- | --- |
| **BMI** | | | |
|  | **female** | **male** | **Total** |
| **Mean** | 22.9391 | 23.5064 | 23.2282 |
| **Standard Error** | 0.3782 | 0.3381 | 0.2532 |
| **Median** | 22.5285 | 23.2673 | 22.8178 |
| **Mode** | 20.2395 | 20.8209 | 20.2395 |
| **Standard Deviation** | 4.6627 | 4.2499 | 4.4588 |
| **Sample Variance** | 21.7407 | 18.0613 | 19.8815 |
| **Kurtosis** | 1.6724 | -0.1015 | 0.8367 |
| **Skewness** | 0.9504 | 0.4824 | 0.7206 |

**For the data corresponding to BMI of females the distribution is leptokurtic, for BMI of males it is platykurtic, and for irrespective of gender it is leptokurtic. Average BMI for female is 22.93, for male 23.50.**

1. **Multiple bar diagram for age:**

**From above plot it is clear that there is a greater number of females from the age group (15 to 24) were present.**

1. **Multiple bar diagram for BMI:**

**From above part it is observe that there is a greater number of females with their BMI lies between (20 to 25)**

1. **Multiple bar diagram for lung capacity and height:**

**Average lung capacity is greater in the height group (1.79 to 1.89) metre for males and also for females.**

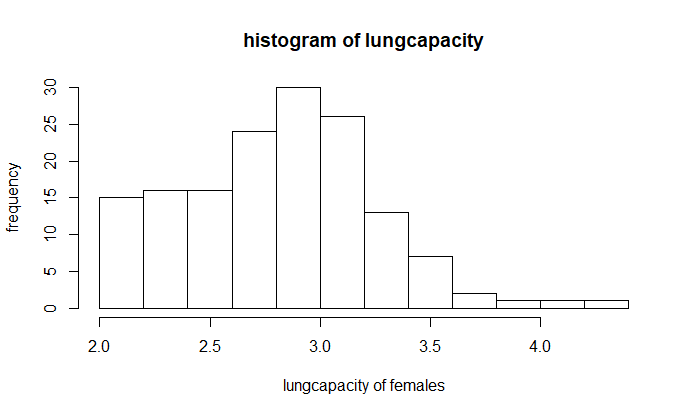
1. **Multiple bar diagram for average lung capacity and age:**

**Average lung capacity is greater in the age group (15-24) for males and for females it is in (15-24) and (25-34) age groups.**

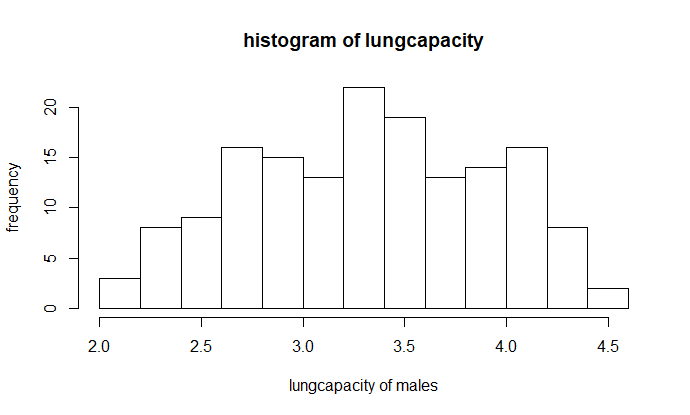
1. **Multiple bar diagram for average lung capacity and weight:**

**Average lung capacity is greater for the (101-111) weight group in males and for females it is in (68-78) age group.**

1. **Histogram of lung capacity of females:**



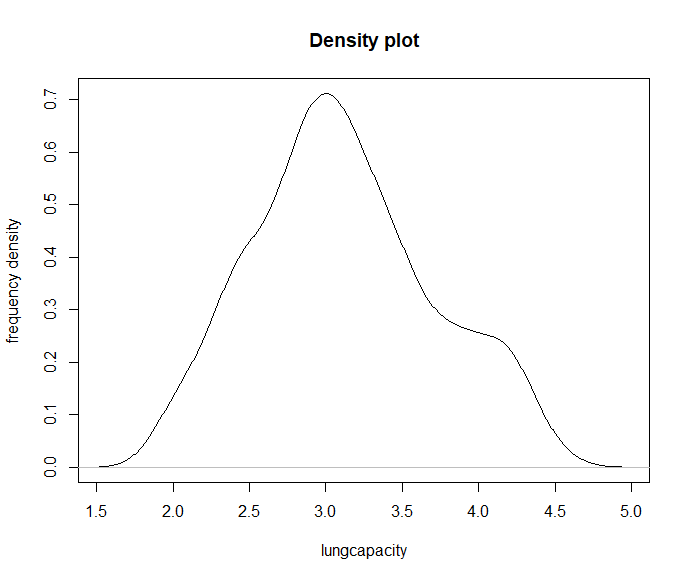
1. **Histogram of lung capacity of males:**



1. **Pie chart denoting the frequency of professions:**

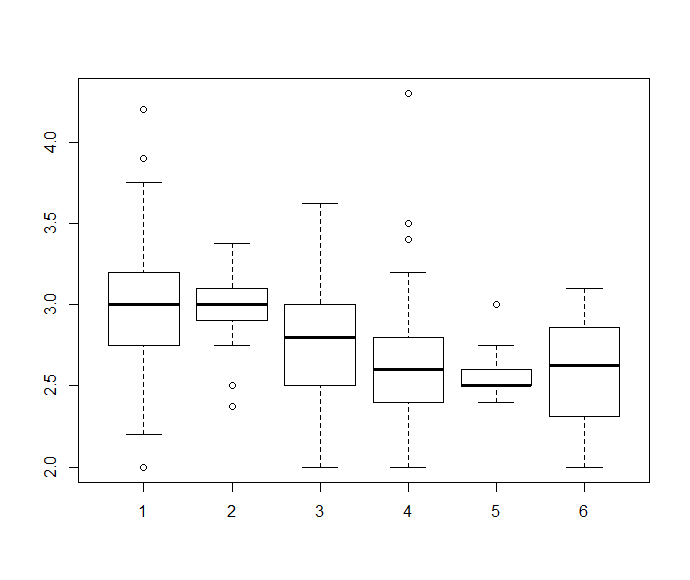
**From above chart it is clear that there is a greater number of students were present in the data. And then the people who does the service were present in greater number.**

1. **Density plot for lung capacity:**



From this graph it is clear that distribution of lung capacity is not normal.

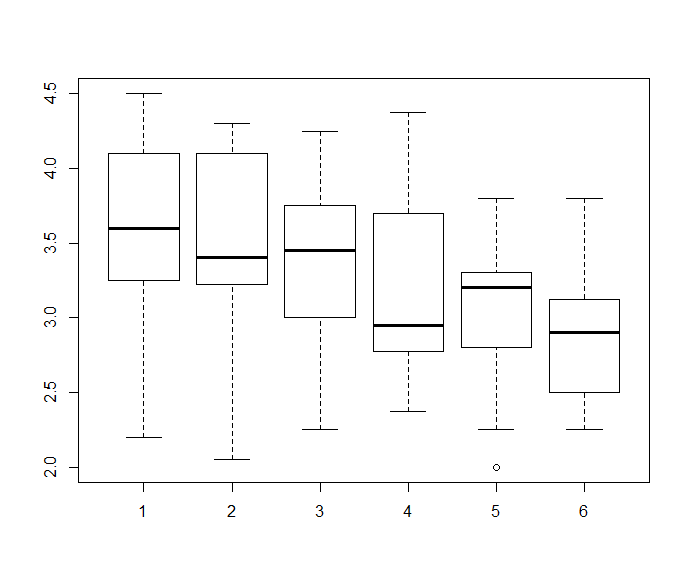
1. **BOXPLOT FOR LUNG CAPACITY FOR FEMALES WITH AGE WISE GROUP**



|  |  |  |
| --- | --- | --- |
| **CLASS NUMBER** | **AGE CRITERIA** | **SKEWNESS** |
| Class1 | 15 to 24 | Negatively skewed |
| Class2 | 25 to 34 | Symmetric |
| Class3 | 35 to 44 | Negatively skewed |
| Class4 | 45 to 54 | Symmetric |
| Class5 | 55 to 64 | Positively skewed |
| Class6 | 65 to 74 | Negatively skewed |

Remark: From boxplot it is clear that in class 1 have total 3 outliers, class2 have 2, class4 have 3 and class5 have 1 outlier. Class1 have 2 womans with higher lungcapacity and one woman have lower lungcapacity than others in that particular group. Simmillarly in class2 two womans have lower lungcapacities than others in that respective class. In class 4 and 5 there are 3 and 1 womans have higher lungcapacities respectively in that particular group.

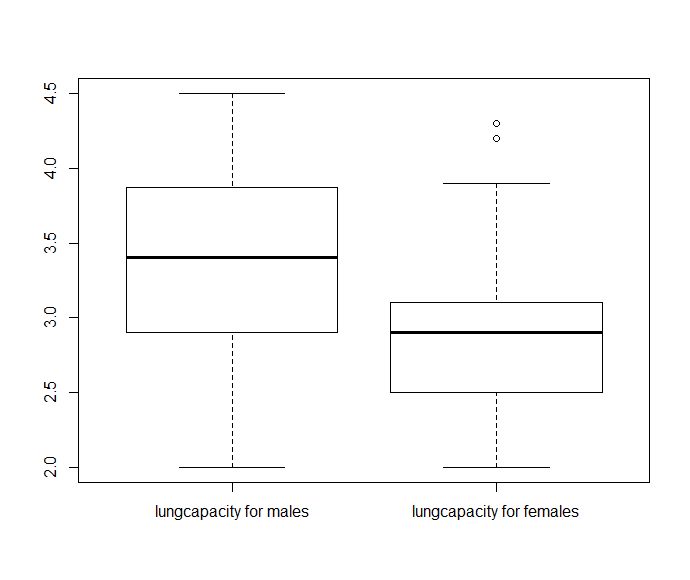
1. **BOXPLOT OF LUNG CAPACITY FOR MALES WITH AGE WISE GROUPS**



|  |  |  |
| --- | --- | --- |
| **CLASS NAME** | **AGE GROUP** | **SKEWNESS** |
| Class1 | 15 to 24 | Positively skewed |
| Class2 | 25 to 34 | Positively skewed |
| Class3 | 35 to 44 | Negatively skewed |
| Class4 | 45 to 54 | Positively skewed |
| Class5 | 55 to 64 | Negatively skewed |
| Class6 | 65 to 74 | Negatively skewed |

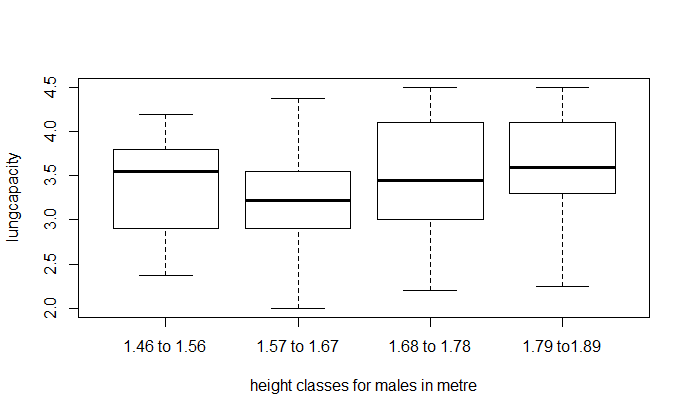
Remark: From boxplot it is clear that only for fifth class there is only one outlier that is there is a men who have lower lung capacity than remaining ones present in that particular group.

1. **BOXPLOT FOR LUNG CAPACITY GENDER WISE**



Here from above boxplot it is clear that for males it is symmetric and for females it is negatively skewed. Also for females there are two outliers representing that there are two females who have highest lung capacity approximately range between 4 to 4.5 L.

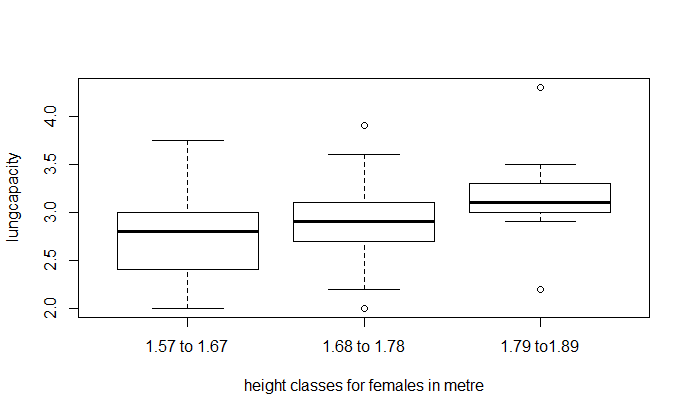
1. **BOXPLOT FOR LUNGCAPACITY OF MALES SORTED WITH HEIGHT WISE GROUPS:**



|  |  |  |
| --- | --- | --- |
| **CLASS NAME** | **HEIGHT GROUP** | **SKEWNESS** |
| Cl2 | 1.46 to 1.56 | Negatively skewed |
| Cl3 | 1.57 to 1.67 | Symmetric |
| Cl4 | 1.68 to 1.78 | Positively skewed |
| Cl5 | 1.79 to 1.89 | Positively skewed |

**-**

1. **BOXPLOT FOR LUNGCAPACITY OF FEMALES SORTED WITH HEIGHT WISE GROUPS**



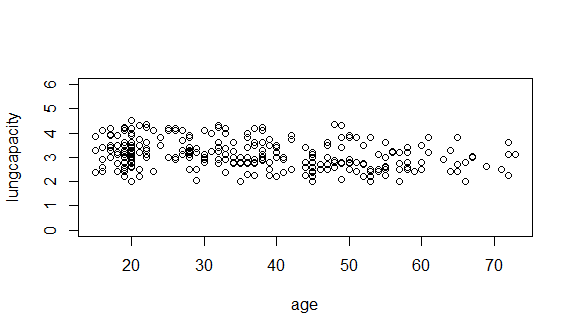
|  |  |  |
| --- | --- | --- |
| **CLASS NAME** | **HEIGHT GROUP** | **SKEWNESS** |
| Cl1 | 1.57 to 1.67 | Symmetric |
| Cl2 | 1.68 to 1.78 | Positively skewed |
| Cl3 | 1.79 to 1.89 | Symmetric |

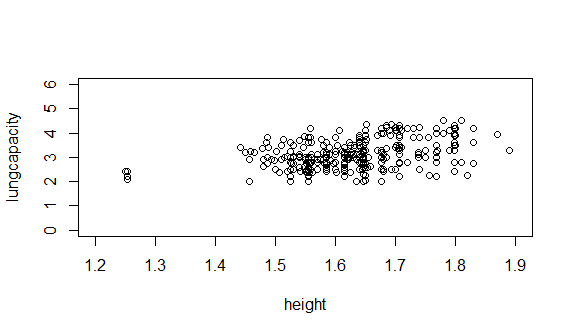
Third and fourth class shows the 2 outliers in each one is specifying lower lung capacity and the other is specifying higher lung capacity.

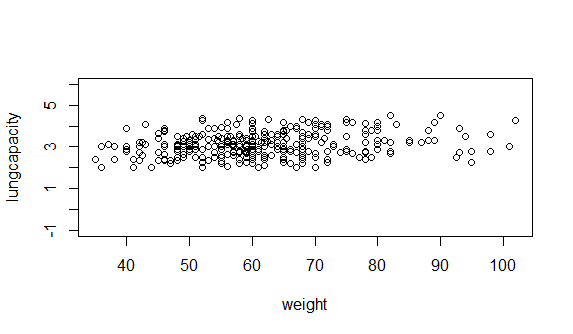
**STATISTICAL ANALYSIS**

Before beginning the analysis let, response variable is lung capacity and regressors are age, height, weight. (The required R code is written in appendix.)

1. **TO CHECK LINEARITY ASSUMPTION BETWEEN LUNG CAPACITY AND AGE, HEIGHT, WEIGHT:**







**From the above plots, age, height, weight are moderately linearly related with lung capacity.**

1. **CORRELATION TEST:**

|  |  |  |
| --- | --- | --- |
|  | Height | Weight |
| AGE | -0.02662291 | 0.2946949 |

|  |  |  |
| --- | --- | --- |
|  | AGE | Weight |
| Height | -0.02662291 | 0.4878669 |

|  |  |  |
| --- | --- | --- |
|  | Age | Height |
| Weight | 0.29469493 | 0.40054858 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Lung capacity** | **Age** | **Height** | **Weight** |
| **r** | -0.2498987 | 0.4005486 | 0.2352809 |
| **t statistic** | 4.5294 | 7.6719 | 4.2484 |
| **Df** | 308 | 308 | 308 |
| **p value** | 4.234e-06 | 1.124e-13 | 1.428e-05 |
| **Decision** | **negative**  **correlation** | **positive**  **correlation** | **positive**  **correlation** |

Since all p values are less than 0.05 los, on the basis of **correlation t test we conclude that age, height, weight are linearly correlated with lung capacity.**

|  |
| --- |
| From above tables it is clear that age is negatively correlated with height and lung capacity with 2.66% and 24.98% respectively. Age is positively correlated with weight with 29.46%.  Height is positively correlated with weight and lung capacity with 48.79% and 40.05% respectively.  Height is negatively correlated with age with 2.6%.  Weight is positively correlated with age, height, lung capacity with 29.45%, 40.05%, 23.52% respectively.  (age, height, weight are linearly correlated with lung capacity.) |
|  |
| |  | | --- | |  | |

1. **EIGEN ANALYSIS FOR CHECKING MULTICOLLINEARITY**:

Multicollinearity can also be treated from eigen values of correlation matrix of x for 3 regressors. Hence for model there will be 3 eigen values say , ,.

Let k be the condition matrix which is retio of max eigen value to the minimum eigen value.

**If k <100** then there will be **no** serious multicollinearity present in the data,

If **100 k1000** indicates **moderate** multicollinearity in the data.

**If k>1000** indicated **severe problem** of multicollinearity.

|  |  |
| --- | --- |
| **Lambda** | **Eignen values** |
|  | 1.5584441 |
|  | 1.0235537 |
|  | 0.4180022 |

**Condition number and condition index:**

|  |  |
| --- | --- |
| k | 3.728316 |
|  | 1 |
|  | 0.6567 |
|  | 0.2682 |

From this it is clear that **condition number is less than 100** which indicates **no serious**

**Problem of multicollinearity** in the data set. All condition indices are also less than 1000 hence,

**Regressors are not dependant.**

**values are listed below**:

|  |  |  |
| --- | --- | --- |
| Response variable | Multiple R square value |  |
| Age | 0.1249 | 1.498801 |
| Height | 0.2698 | 1.369488 |
| weight | 0.3328 | 1.498801 |

**From above table we observe that no value of is greater or equal to 5 hence multicollinearity does not exist for this data set.**

1. **NORMALITY TEST:**
2. SHAPIRO WILK TEST

Shapiro-Wilk normality test

data: lungcapacity

**W = 0.97741, p-value = 0.01326**

Here p value is less than 5% level of significance hence the response variable

that is **lung capacity is not normally distributed**

1. **TEST FOR CHECKING OBSERVATIONS ARE RANDOM OR NOT**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Factor | Statistic | Runs |  |  | P value |
| Age | -17.321 | 2 | 151 | 152 | 2.2e-16 |
| Height | -0.45511 | 152 | 155 | 155 | 0.649 |
| Weight | -3.5743 | 116 | 142 | 149 | 0.0003512 |
| lungcapacity | -2.8562 | 55 | 69 | 73 | 0.004288 |

Here observations of **height** are **random** as p value is **greater than 0.05** los but all others are

don’t have random observations.

1. **MULTIPLE REGRESSION MODEL**: (**without intercept term**)

Since intercept term is not significant for the models hence, we removed it.

**Assumptions:**

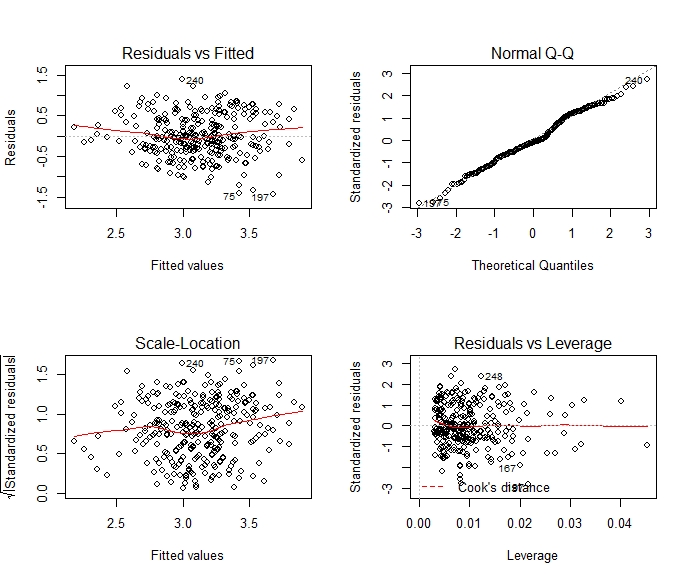
1. Errors are uncorrelated with mean zero and constant variance i.e.
2. Errors are normally distributed.
3. Rank of matrix of regressors is K

For this sample data, **intercept is not significant** and because of that value of  **is very low.** The variability of lung capacity explained by corresponding regressors is just **23.7%** hence we have decided to **eliminate intercept** from the model and then **refitted the model** and summary is given below:

|  |  |  |
| --- | --- | --- |
| Regressors | value | Pr(>|t|) |
| Age | -0.011288 | < 6.11e-08 |
| Height | 1.883206 | < 2e-16 |
| Weight | 0.007276 | <0.00571 |

**Multiple R-squared: 0.9741**, **Adjusted R-squared: 0.9738**

F-statistic: 3842 on 3 and 307 DF, p-value: < 2.2e-16



From above residuals plots, by observing first plot (**residuals vs fitted**) it is clear that there is a **“double bow”** pattern showed by the errors and hence the assumption of **homoscedasticity**  **fails** that is errors does **not have constant variance**.

In Normal Q-Q plot all observations are lie around the dotted line and therefore **normality**  condition **holds** approximately.

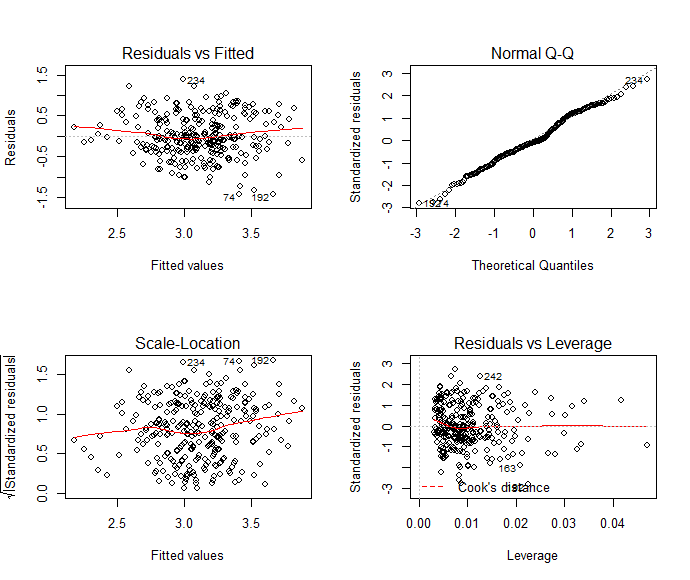
For removing the “double bow” pattern we have tried to remove the preferable outliers from the data set and then again fitted the corresponding model and we have observed the pattern, multiple value, adjusted value.

**Table1:**

**For getting the correct pattern we have removed the outliers from the data set and again fitted the respective model.**

|  |  |  |
| --- | --- | --- |
| **Model number** | **Adjusted** | **Multiple** |
| Model1 | 0.974 | 0.9742 |
| Model2 | 0.9742 | 0.9744 |
| Model3 | 0.974 | 0.9743 |

Even though adjusted slightly increased in model1, model2, model3 than the previous model the residual pattern does not changed from the double bow pattern.

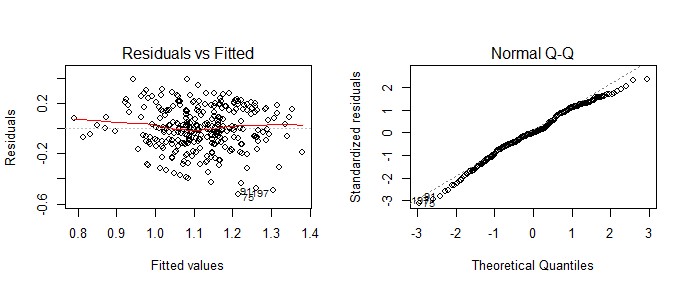


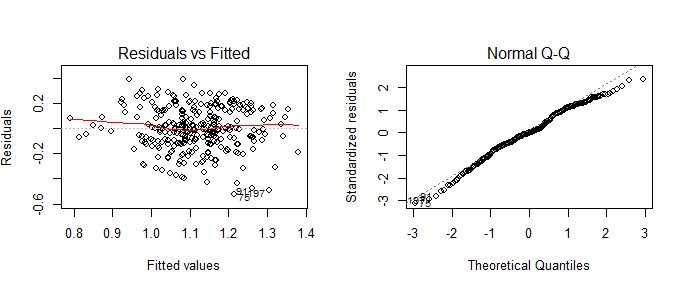
**Transformations** used on response and regressors in order to **remove “double bow”** pattern are as follows:

|  |  |  |
| --- | --- | --- |
| **Transformation** | **Multiple** | **Adjusted** |
| ln (lungcapacity) | 0.9784 | 0.9782 |
| ln (lungcapacity)  ln(age)  ln(height)  ln(weight) | 0.9784 | 0.9782 |
| log (age)  log(height)  log(weight) | 0.9738 | 0.9736 |
| 1/lungcapacity  1/age  1/height  1/weight | 0.959 | 0.9586 |

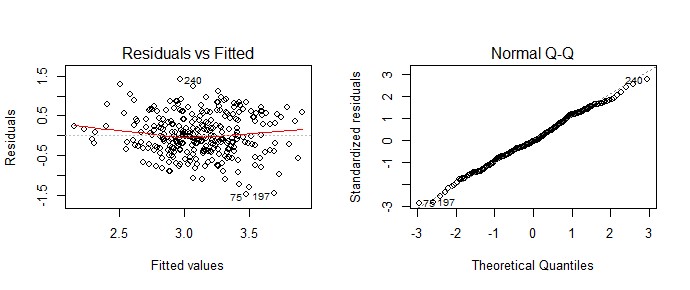
Even though **variability in lung capacity explained by age, height, weight is high** almost it ranges from **95% to 98%,** the residuals plots for corresponding models have

**“Double bow pattern”** and also normality plot seems **very skewed** in the last transformation. Plots are shown on the next page.

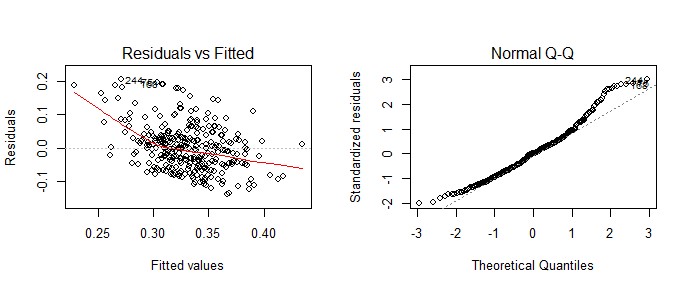
1. **Transformation used: ln (lung capacity)**
2. **Transformation used: ln (lung capacity, age, height, weight)**

****

1. **Transformation used: log (age, height, weight)**

**-**

1. **Transformation used: (1/lung capacity, 1/age, 1/height, 1/weight)**

****

All above transformations are not useful for correcting the double bow pattern from the residuals plot**.**

1. **The next transformation we have used is Box-Cox method.**

|  |  |  |
| --- | --- | --- |
| **Regressors** | **Coefficients** | **Pr(>|t|)** |
| **Age** | -0.006319 | <2.89e-08 |
| **Height** | 0.890802 | <2e-16 |
| **Weight** | 0.004129 | <0.00407 |

|  |
| --- |
| Residual standard error: 0.2801 on 307 degrees of freedom  **Multiple R-squared: 0.9662,** **Adjusted R-squared: 0.9658**  F-statistic: 2921 on 3 and 307 DF **p-value: < 2.2e-16**    This plot is to find value of so that can be calculated and can be use as new response.  Here value of  **is 0.4646465** |
|  |
| |  | | --- | |  | |

Both plots can be plotted using new response values of lung capacity obtained from box cox method. But here also “double bow” pattern is present.

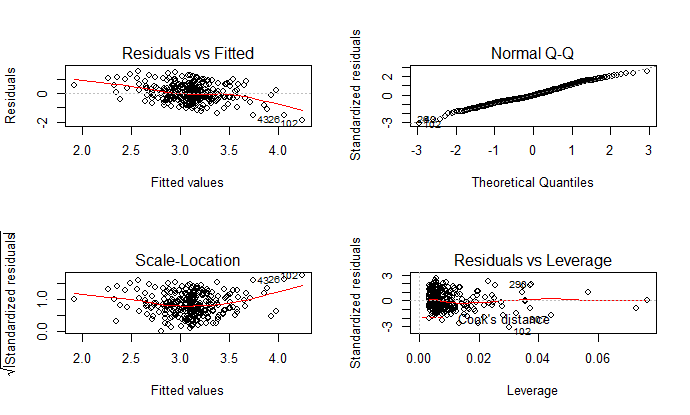
1. **POLYNOMIAL REGRESSION MODEL**

Then we have tried to fit the **polynomial regression** model so for that instead of considering all regressors age, height, weight, we have considered only **age** and **BMI** as it is depending upon height and weight.

|  |  |  |
| --- | --- | --- |
| Regressors | Estimates | Pr(>|t|) |
| Age | 0.0742361 | <2e-16 |
| BMI | 0.1516669 | <2e-16 |
| Age\*BMI | -0.0036127 | <2e-16 |

**Multiple R-squared: 0.9643, Adjusted R-squared: 0.964**

F-statistic: 2765 on 3 and 307 DF **p-value: < 2.2e-16**

****

Here also residual plot shows that the **homoscedasticity assumption fails.**

1. **WEIGHTED LEAST SQUARE METHOD:**

Response variable: lung capacity

Regressors: Age and BMI

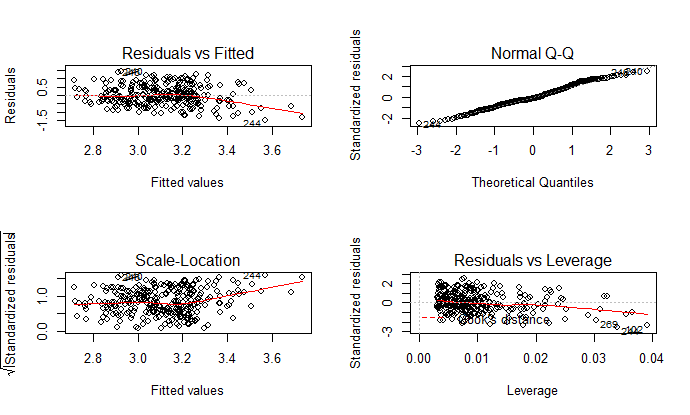
To correct the inequality of variance problem, we must know the weights. We note from examining the data in that there are several sets of regressors values that are “near neighbours,” that is, that have approximate repeat points. We will assume that these near neighbours are close enough to be considered repeat points and use the variance of the responses at those repeat points to investigate how Var(lungcapacity) changes with regressors. We have estimated the variance of lung capacities and reciprocal of that gives corresponding weights.

|  |  |  |
| --- | --- | --- |
| Regressors | Estimates | Pr(>|t|) |
| Age | -0.013167 | <9.49e-08 |
| BMI | 0.083606 | < 2e-16 |
| Weights | 4.719679 | < 2e-16 |

Residual standard error: 0.5844 on 307 degrees of freedom

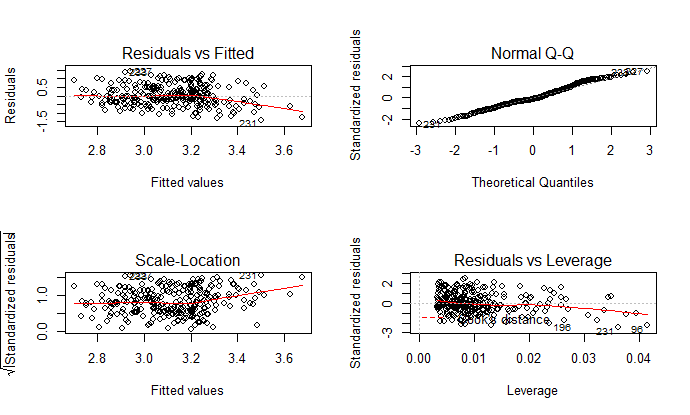
**Multiple R-squared: 0.9664, Adjusted R-squared: 0.966**

F-statistic: 2940 on 3 and 307 DF, p-value: < 2.2e-16



**Here double bow pattern has removed but there are some influence points which are pulling the regression line in its direction so we have decided to remove those from the data set. After removing those points, we got the following estimates and corresponding plots.**

|  |  |
| --- | --- |
| **Regressors** | **Estimates** |
| Age | -0.01336 |
| BMI | 0.08112 |
| Weights | 4.91129 |



So here double bow pattern is totally gone from the residual plots , normality is better than all other above plots so weighted least square methods gave the best regression model for forecasting purpose.

The best regression model is:

**Lung capacity= -0.01336\*age + 0.08112\*BMI +4.91129\*weights**

**KRUSKAL WALLI H TEST: (all corresponding hypothesis)**

1. **FOR AGE**

: Average lung capacity is same for all age groups.

: Average lung capacity is differ significantly.

**Kruskal-Wallis rank sum test**

Kruskal-Wallis chi-squared = 26.715, df = 5**, p-value = 6.481e-05**

As **p value is less than 0.05 level of significance** so we are unable to accept null hypothesis.

**Hence average lungcapacity differ significantly for all age groups**

1. **FOR PULSE RATE**

: Average lung capacity is same for all pulserate groups.

: Average lung capacity is differing significantly for all pulserate groups

**Kruskal-Wallis rank sum test**

Kruskal-Wallis chi-squared = 0.93028, df = 4, **p-value = 0.9202**

Here **p value is greater than the 0.05 level of significance** therefore we accept the null hypothesis.

Hence, **the average lungcapacity is same for all pulse rate groups.**

1. **For BMI**

: Average lung capacity is same for all groups according to BMI.

: Average lung capacity is differing significantly for all groups according to BMI

**Kruskal-Wallis rank sum test**

Kruskal-Wallis chi-squared = 5.2223, df = 4, p-value = 0.2652

Here **p value is greater than 0.05** level of significance so we accept null hypothesis.

Hence **average lungcapacity is same for all groups according to BMI.**

1. **FOR PROFESSION:**

: Average lung capacity is same for all professions.

: Average lung capacity is differing significantly for all professions.

**Kruskal-Wallis rank sum test**

Kruskal-Wallis chi-squared = 19.604, df = 3, p-value = 0.0002051

Here **p value is less than 0.05** level of significance and therefore we are unable to accept null hypothesis.

Hence**, average lung capacity is differing significantly for professions.**

1. **For height**

: Average lung capacity is same for all height groups.

: Average lung capacity is differing significantly for all height groups.

**Kruskal-Wallis rank sum test**

Kruskal-Wallis chi-squared = 52.767, df = 5, **p-value = 3.756e-10**

Here **p value is less than 0.05** level of significance so we are unable to accept the null hypothesis. Hence, **average lung capacity is differing significantly for all height groups.**

1. **For weight**

: Average lung capacity is same for all weight groups.

: Average lung capacity is differing significantly for all weight groups.

**Kruskal-Wallis rank sum test**

Kruskal-Wallis chi-squared = 15.955, df = 6, **p-value = 0.014**

Here, **p value is less than 0.05** level of significance so we are unable to accept the null hypothesis.

Hence, **average lungcapacity is significantly different for all groups according to weights.**

**Conclusion table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Regressors** | **Statistic** | **P value** | **Decision for lung capacity** |
| Age | 26.715 | 6.481e-05 | **Differ** significantly |
| Pulse rate | 0.93028 | 0.9202 | Does **not differ** significantly. |
| BMI | 5.2223 | 0.2652 | Does **not differ** significantly. |
| Profession | 19.604 | 0.0002051 | **Differ** significantly |
| Height | 52.767 | 3.756e-10 | **Differ** significantly |
| Weight | 15.955 | 0.014 | **Differ** significantly |

**Notation table for Mann –Whitney test:**

|  |  |
| --- | --- |
| **Name** | **Class limit** |
| **Age1** | **15-24** |
| **Age2** | **25-34** |
| **Age3** | **35-44** |
| **Age4** | **45-54** |
| **Age5** | **55-64** |
| **Age 6** | **65-74** |

|  |  |
| --- | --- |
| **Name** | **Class limit** |
| **BMI1** | **14-19** |
| **BMI2** | **20-25** |
| **BMI3** | **26-31** |
| **BMI4** | **32-37** |
| **BMI5** | **38-43** |

**MANN – WHITNEY TEST:**

: median difference for pairs of observations is zero.

: median difference for pairs of observations is not zero.

**FOR AGE:**

**Median for lung capacity of sample with respect to age is 3.075L**

|  |  |  |  |
| --- | --- | --- | --- |
| **Class combination** | **Test statistic** | **P value** | **Decision for** |
| Age1, Age2 | 2872.5 | 0.441 | Accept |
| Age1, Age3 | 3378 | 0.1763 | Accept |
| Age1, Age4 | 3533 | 0.000505 | Reject |
| Age1, Age5 | 1808 | 0.01055 | Reject |
| Age1, Age6 | 926 | 0.03645 | Reject |
| Age2, Age3 | 2043.5 | 0.0456 | Reject |
| Age2, Age4 | 2167 | 2.551e-05 | Reject |
| Age2, Age5 | 1102 | 0.001398 | Reject |
| Age2, Age6 | 571 | 0.006106 | Reject |
| Age3, Age4 | 1751 | 0.04172 | Reject |
| Age3, Age5 | 909 | 0.09949 | Accept |
| Age3, Age6 | 455.5 | 0.2011 | Accept |
| Age4, Age5 | 627 | 0.8049 | Accept |
| Age4, Age6 | 326 | 0.9932 | Accept |
| Age5, Age6 | 179 | 0.7767 | Accept |

**Decision criteria** for null hypothesis is:

Reject the null hypothesis if the p value is less than or equal to 0.05 level of significance.

Combinations (Age1, Age4), (Age1, Age5), (Age1, Age6), (Age2, Age3), (Age2, Age4) (Age2, Age5), (Age2, Age6), (Age3, Age4) **based on sample corresponding population median may be considered as unequal.**

**FOR BMI:**

**Median for lung capacity of sample with respect to BMI is: 3.1 L**

|  |  |  |  |
| --- | --- | --- | --- |
| **Class combination** | **Test statistic** | **P value** | **Decision for** |
| BMI1, BMI2 | 3543 | 0.9788 | Accept |
| BMI1, BMI3 | 1444.5 | 0.6432 | Accept |
| BMI1, BMI4 | 242.5 | 0.6431 | Accept |
| BMI2, BMI3 | 4326.5 | 0.4713 | Accept |
| BMI2, BMI4 | 730.5 | 0.5501 | Accept |
| BMI3, BMI4 | 267 | 0.7826 | Accept |

For BMI**, all combinations of classes based on samples, corresponding population median may be considered as equal .**

**FOR PROFESSION:**

**Sample medians**

|  |  |
| --- | --- |
| **Profession** | **Median** |
| **Job** | **3.1 L** |
| **Housewife** | **2.8 L** |
| **Retired** | **2.75 L** |
| **student** | **3.15 L** |
| **Business** | **3.3 L** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Class combination** | **Test statistic** | **P value** | **Decision for** |
| Job, business | 941.5 | 0.09564 | Accept |
| Job, Housewife | 3697.5 | 0.001077 | Reject |
| Job, Retired | 1265.5 | 0.05645 | Accept |
| Job, Student | 5897.5 | 0.1933 | Accept |
| Business, Housewife | 546.5 | 0.001746 | Reject |
| Business, Retired | 188.5 | 0.0184 | Reject |
| Business, Student | 960 | 0.4116 | Accept |
| Housewife, Retired | 267.5 | 0.8125 | Accept |
| Housewife, Student | 1122.5 | 0.0001773 | Reject |
| Retired, Student | 422.5 | 0.02818 | Reject |

**Decision criteria for null hypothesis is:**

Reject the null hypothesis if the **p value is less than or equal to 0.05 level of significance.**

Class combinations (job, housewife), (business, housewife), (business, retired), (housewife, student), (retired, student) have **significantly different population median**.

**TO TEST WHETHER AVERAGE SAMPLE LUNGCAPACITY IS EQUAL TO THE POPULATION LUNG CAPACITY:**

1. **: lungcapacity for sample value for females is equal to the lungcapacity for population value for females that is 4.2 L.**

**Vs**

**: lungcapacity for sample value for females is not equal to the lung capacity for population value for females that is 4.2 L.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test statistic (|t|)** | **P value** | **95% CI** | **mean of sample** |
| 37.644 | <2.2e-16 | (2.78, 2.92) | 2.85 |

Here p value is **less than 0.05** level of significance hence, we are unable to accept the null hypothesis.

Hence, **lung capacity for sample value for females is not equal to the lung capacity for population values for males.**

1. **: lungcapacity for sample value for males is equal to the lung capacity for population value for males that is 6L.**

**Vs**

**: lung capacity for sample value for males is not equal to the lung capacity for population value for males that is 6L.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test statistic (|t|)** | **P value** | **95% CI** | **mean of sample** |
| 55.434 | <2.2e-16 | (3.27, 3.46) | 3.36 |

Here p value is **less than 0.05** level of significance hence, we are unable to accept the null hypothesis.

Hence, **lung capacity for sample value for males is not equal to the lung capacity for population values for males.**

**INDEPENDENCE OF ATTRIBUTES**

1. **Test for checking association of smoking and lung capacity**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Smoking** | **Non smoking** | **Occasionally** | **Total** |
| **A=2 to 2.5** | 4 | 53 | 2 | 59 |
| **B=2.6 to 3.1** | 2 | 108 | 7 | 117 |
| **C=3.2 to 3.7** | 2 | 77 | 3 | 82 |
| **D=3.8 to 4.3** | 1 | 49 | 0 | 50 |
| **E=4.4 to 4.9** | 0 | 2 | 0 | 2 |
| **Total** | 9 | 289 | 12 | 310 |

: No association between smoking and lung capacity.

: There is an association between smoking and lung capacity.

**X-squared = 7.5131**, df = 8, **p-value = 0.4824**

**Here p value is greater than 0.05 level of significance, therefore we accept null hypothesis.**

**There is no association between smoking and lung capacity.**

1. **Test for checking association between profession and lung capacity.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Job** | **Business** | **Housewife** | **Retired** | **Student** | **Total** |
| **A** | 24 | 1 | 13 | 5 | 14 | 57 |
| **B** | 50 | 7 | 19 | 6 | 32 | 114 |
| **C** | 42 | 3 | 6 | 3 | 26 | 80 |
| **D** | 22 | 7 | 1 | 0 | 19 | 49 |
| **E** | 0 | 0 | 0 | 0 | 2 | 2 |
| **Total** | 138 | 18 | 39 | 14 | 93 | 302 |

NOTE: A= 2 to 2.5, B=2.6 to 3.1, C=3.2 to 3.7, D=3.8 to 4.3, E=4.4 to 4.9

**: No association between profession and lung capacity.**

**: There is an association between profession and lung capacity.**

**X-squared = 32.164**, df = 16, **p-value = 0.00952**

Here **p value is less than the 0.05** level of significance, therefore we are unable to accept the null hypothesis**.**

Hence, **there is an association between profession and lung capacity.**

**TO CHECK EXACT AMOUNT OF ASSOCIATION BETWEEN PROFESSION AND LUNG CAPACITY:**

**Cramer’s V test:**

**Test statistic:**

**V=**

**= 32.164**

**n= 302**

**q=5**

**Test statistic: 0.1631742**

The value of y indicated the test statistic value for Cramer’s v test

Conclusion:

Since, **test statistic value is greater than 0.15** and hence there is a **strong association** between lung capacity and profession.

**CONCLUSIONS**

1. Age, height, weight are linearly related with lung capacity.
2. Age is **negatively correlated** with lung capacity whereas height and weight are **positively correlated** with lung capacity.
3. The lung capacity is **not normally** distributed.
4. The best possible regression model for forecasting purpose is given as follows:

**Lung capacity= -0.01336\*age + 0.08112\*BMI +4.91129\*weights**

Where weights can be obtained using:

sy=-0.006217\*age+0.138595\*BMI

weights=1/sy

1. Lungcapacity for sample data of females is 2.85L and for sample data of males is 3.36L
2. Average lung capacity is **differing significantly** for all age, height, weight, profession groups but does **not differing significantly** for pulse rate and BMI.
3. For **age classes** [(15 to 24) and (45 to 54)], [(15 to 24) and (55 to 64)], [(25 to 34), (35 to 44)], [(25 to 34),(45 to 54)], [(25 to 34),(55 to 64)], [(25 to 34),(65 to 74)], [(35 to 44), (45 to 54)] based on sample corresponding population median for lung capacity may be considered as unequal.
4. For profession wise groups consist (job and housewife), (business, housewife),

(business, retired), (housewife, student), (retired, student) **have significantly different population median for lung capacity.**

1. For all BMI based on samples, population median for lung capacity may be considered as equal**.**
2. There is **no association** between **smoking and lung capacity.**
3. There is a **strong association** between **profession and lung capacity**.

**SCOPE AND LIMITATIONS**

* **SCOPE**

The regression model obtained in this project can be used in medical field for obtaining the lung capacity of an individual by just substituting the values of his or her age, BMI and weights can be calculated as per the given formula without performing any experiment.

Water spirometer which is used in this project can be made easily by anyone and is cheapest among all measuring instruments which are used in the medical field. This type of procedure for estimating lung capacity is financially acceptable.

* **LIMITATIONS**
* Observations of lung capacity do not follow Normal Distribution.
* The experimental procedure is time consuming as we need approximately 7-8 minutes per observation.
* Since some group of people did not understand the exact procedure of the experiment, they inhaled water from the pipe and further they did not co-operate with the procedure of the experiment.
* Some individuals were not at rest before performing the experiment which might be affects their lung capacity value.
* For information such as Age, Height, Weight, Profession we have to rely on the information provided by individuals.

**REFERENCES**

1. **GUPTA S. C AND KAPOOR V. K FUNDAMENTAL OF APPLIED STATISTICS S. CHAND SONS, New Delhi.**
2. **Introduction to Linear Regression Analysis (fifth edition) by DOUGLAS C. MONTGOMERY, ELIZABETH A. PECK, G. GEOFFREY VINING.**
3. <https://www.researchgate.net/figure/nterpretation-of-Phi-and-Cramers-V_tbl2_326885374> (for Cramer’s V test)
4. <https://en.wikipedia.org/wiki/Lung_volumes> (ideal lung capacity values for population)

**APPENDIX**

All required R commands

1. For finding correlation:

> age=matrix(age,ncol=1,nrow=310)

> height=matrix(height,ncol=1,nrow=310)

> weight=matrix(weight,ncol=1,nrow=310)

> lungcapcity=matrix(lungcapcity,ncol=1,nrow=310)

> vector=cbind(age,height,weight,lungcapcity)

> r=cor(vector)

> r

1. Eigen analysis for checking multicollinearity:

> vector=cbind(age,height,weight,lungcapcity)

> r=cor(vector)

> eigen(r)

> condition\_no=1.5584441/0.4180022

> condition\_no

[1] 3.728316

> k2= 1.0235537/1.5584441

> k2

[1] 0.6567792

> k3=0.4180022/1.5584441

> k3

[1] 0.2682176

1. Shapiro test

> lungcapcity=scan("clipboard")

Read 310 items

> shapiro.test(lungcapacity)

1. **TEST FOR RANDOMNESS**

>library(randtests)

> runs.test(age)

> runs.test(weight)

> runs.test(height)

> runs.test(lungcapacity)

1. **MULTIPLE REGRESSION MODEL:**

> age=scan("clipboard")

Read 310 items

> height=scan("clipboard")

Read 310 items

> weight=scan("clipboard")

Read 310 items

> lungcapcity=scan("clipboard")

Read 310 items

> age=matrix(age,ncol=1,nrow=310)

> height=matrix(height,ncol=1,nrow=310)

> weight=matrix(weight,ncol=1,nrow=310)

> lungcapcity=matrix(lungcapcity,ncol=1,nrow=310)

> vector=cbind(age,height,weight,lungcapcity)

> model=lm(vector[,4]~vector[,1]+vector[,2]+vector[,3]-1

> summary(model)

> plot(model)

1. **For removing outliers and for fitting model (table 1)**

> vector1=vector[c(-106,-158,-168),]

> model1=lm(vector1[,4]~vector1[,1]+vector1[,2]+vector1[,3]-1)

> vector2=vector1[c(-109,-295,-256),]

> model2=lm(vector2[,4]~vector2[,1]+vector2[,2]+vector2[,3]-1)

> vector3=vector2[c(-13,-298,-233),]

> model3=lm(vector3[,4]~vector3[,1]+vector3[,2]+vector3[,3]-1)

1. **For box cox method:**

> age=scan("clipboard")

Read 310 items

> height=scan("clipboard")

Read 310 items

> weight=scan("clipboard")

Read 310 items

> lungcapcity=scan("clipboard")

Read 310 items

> model9=lm(lungcapcity~age+height+weight-1)

> library(MASS)

> tra=boxcox(model9)

> best.lam=tra$x[which(tra$y==max(tra$y))]

> best.lam

[1] 0.4646465

> new\_response=scan("clipboard")

Read 310 items

> model0=lm(new\_response~age+height+weight-1)

> model0

summary(model0)

1. **for polynomial regression model:**

> age=scan("clipboard")

Read 310 items

> bmi=scan("clipboard")

Read 310 items

> lungcapacity=scan("clipboard")

Read 310 items

> model11=lm(lungcapacity~age+bmi+age\*age+bmi\*bmi+age\*bmi-1)

> summary(model11)

> par(mfrow=c(2,2))

> plot(model11)

1. **FOR WEIGHTED LEAST SQUARE METHOD:**

> age=scan("clipboard")

Read 310 items

> bmi=scan("clipboard")

Read 310 items

> lungcapacity=scan("clipboard")

Read 310 items

> vectors=cbind(age,bmi,lungcapacity)

> model=lm(vectors[,3]~vectors[,1]+vectors[,2]-1)

> model

1. **For removing the influence points from the data set:**

> weights=matrix(weights,nrow=310,ncol=1)

> m=cbind(age,bmi,lungcapacity,weights)

> m1=m[c(-174),]

> model1=lm(m1[,3]~m1[,1]+m1[,2]+m1[,4]-1)

> m2=m1[c(-238),]

> model1=lm(m2[,3]~m2[,1]+m2[,2]+m2[,4]-1)

> m3=m2[c(-170),]

> model1=lm(m3[,3]~m3[,1]+m3[,2]+m3[,4]-1)

> m4=m3[c(-249),]

> model1=lm(m4[,3]~m4[,1]+m4[,2]+m4[,4]-1)

> m5=m4[c(-73),]

> model1=lm(m5[,3]~m5[,1]+m5[,2]+m5[,4]-1)

> m6=m5[c(-270,-50,-210),]

> model1=lm(m6[,3]~m6[,1]+m6[,2]+m6[,4]-1)

> m7=m6[c(-246),]

> model1=lm(m7[,3]~m7[,1]+m7[,2]+m7[,4]-1)

> m8=m7[c(-273,-5,-37),]

> model1=lm(m8[,3]~m8[,1]+m8[,2]+m8[,4]-1)

> m9=m8[c(-86,-145,-288),]

> model1=lm(m9[,3]~m9[,1]+m9[,2]+m9[,4]-1)

> m10=m9[c(-166,-259,-9),]

> model1=lm(m10[,3]~m10[,1]+m10[,2]+m10[,4]-1)

> m11=m10[c(-148),]

> model1=lm(m11[,3]~m11[,1]+m11[,2]+m11[,4]-1)

> plot(model1)

>model1

1. **For Kruskal test**:

For age:

> age=scan("clipboard")

Read 310 items

> lungcapacity=scan("clipboard")

Read 310 items

> class1m=lungcapacity[age>=15 & age <=24]

> class2m=lungcapacity[age>=25 & age<=34 ]

> class3m=lungcapacity[age>=35 & age<=44]

> class4m=lungcapacity[age>=45 & age<=54]

> class5m=lungcapacity[age>=55 & age<=64]

> class6m=lungcapacity[age>=65 & age<=74]

> list1=list(class1m,class2m,class3m,class4m,class5m,class6m)

> kruskal.test(list1)

1. **For pulse rate:**

> pulse\_rate=scan("clipboard")

Read 310 items

> lungcapacity=scan("clipboard")

Read 310 items

> cl1=lungcapacity[pulse\_rate>=60 & pulse\_rate<=79]

> cl2=lungcapacity[pulse\_rate>=80 & pulse\_rate<=99]

> cl3=lungcapacity[pulse\_rate>=100 & pulse\_rate<=199]

> cl3=lungcapacity[pulse\_rate>=100 & pulse\_rate<=119]

> cl4=lungcapacity[pulse\_rate>=120 & pulse\_rate<=139]

> cl5=lungcapacity[pulse\_rate>=140 & pulse\_rate<=159]

> list2=list(cl1,cl2,cl3,cl4,cl5)

> kruskal.test(list2)

1. **For BMI:**

> bmi=scan("clipboard")

Read 310 items

> lungcapacity=scan("clipboard")

Read 310 items

> class1bmi=lungcapacity[bmi>=14 & bmi<=19]

> class2bmi=lungcapacity[bmi>=20 & bmi<=25]

> class3bmi=lungcapacity[bmi>=26 & bmi<=31]

> class4bmi=lungcapacity[bmi>=32 & bmi<=37]

> class5bmi=lungcapacity[bmi>=38 & bmi<=43]

> list3=list(class1bmi,class2bmi,class3bmi,class4bmi,class5bmi)

> kruskal.test(list3)

1. **For profession**:

> job=scan("clipboard")

Read 138 items

> business=scan("clipboard")

Read 18 items

> housewife=scan("clipboard")

Read 40 items

> retired=scan("clipboard")

Read 14 items

> student=scan("clipboard")

Read 95 items

> list4=list(business,housewife,retired,student)

> kruskal.test(list4)

1. **For height**

> height=scan("clipboard")

Read 310 items

> lungcapacity=scan("clipboard")

Read 310 items

> cl1=lungcapacity[height>=1.24 & height<=1.34]

> cl=lungcapacity[height>=1.35 & height<=1.45]

> cl3=lungcapacity[height>=1.46 & height<=1.56]

> cl4=lungcapacity[height>=1.57 & height<=1.67]

> cl5=lungcapacity[height>=1.68 & height<=1.78]

> cl6=lungcapacity[height>=1.79 & height<=1.89]

> list5=list(cl1,cl,cl3,cl4,cl5,cl6)

> kruskal.test(list5)

1. **For weight**

> weight=scan("clipboard")

Read 310 items

> lungcapacity=scan("clipboard")

Read 310 items

> clw1=lungcapacity[weight>=35 & weight<=45]

> clw2=lungcapacity[weight>=46 & weight<=56]

> clw3=lungcapacity[weight>=57 & weight<=67]

> clw4=lungcapacity[weight>=68 & weight<=78]

> clw5=lungcapacity[weight>=79 & weight<=89]

> clw6=lungcapacity[weight>=90 & weight<=100]

> clw7=lungcapacity[weight>=101 & weight<=111]

> list6=list(clw1,clw2,clw3,clw4,clw5,clw6,clw7)

> kruskal.test(list6)

1. **For Mann Whitney**

>Wilcoxon.test(x,y)

1. **FOR INDEPENDENCE**

> smoking=c(4,2,2,1,0)

> non\_smoking=c(53,108,77,49,2)

> occasionally=c(2,7,3,0,0)

> x=cbind(smoking,non\_smoking,occassionally)

> chisq.test(x)

1. **FOR INDEPENDENCE OF PROFESSON**

> j=c(24,50,42,22,0)

> b=c(1,7,3,7,0)

> hw=c(13,19,6,1,0)

> re=c(5,6,3,0,0)

> s=c(14,32,26,19,2)

> x=cbind(j,b,hw,re,s)

> chisq.test(x)

1. **Cramer’s v test**

> x=cbind(j,b,hw,re,s)

> sum(x)

[1] 302

> sum(j)+sum(b)+sum(hw)+sum(re)+sum(s)

[1] 302

> q=min(nrow(x),ncol(x))

> q

[1] 5

> y=sqrt(32.164/(302\*(4)))

> y

1. **For correlation t test:**

>cor.test(height,lungcapacity,alternative ="greater")

> cor.test(weight,lungcapacity,alternative ="greater")

> cor.test(age,lungcapacity,alternative ="less")