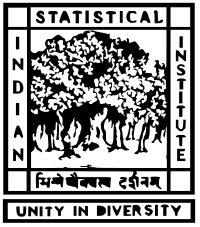
**INDIAN STATISTICAL INSTITUTE**



***ASSIGNMENT 2:***

***Analyzing bivariate data***

**NAME:** SUBHRAJYOTY ROY

**ROLL:**  BS-1613

**DEPT:** B.STAT YEAR- I

**SUBJECT:** STATISTICAL METHODS I

Introduction

I was trying out some datasets and their plots to see whether I can find something interesting, and luckily I came across to a dataset which showed that Smokers tend to have lower heights. This observation really felt me uneasy and thought that it may be connected through genetics. I looked all the data in the sets about some different things and dug more into the observations to analyze the connection between these two.

Collection of Data

The dataset I found from the website http://nargurd.com/lss/. The dataset is formed with Lung capacity (in cc), Height (in inches), Age (last birthday convention), Exercise time(in minutes per day), Degree of Smoking, Gender in about 150 people.

Summary of Data

I summarize the following frequency table from the dataset, and put them in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Frequencies** | **Addicted Smoker** | **Non- addicted smoker** | **Non-smoker** |
| **Male** | 30 | 33 | 20 |
| **Female** | 31 | 28 | 19 |

Summarizing the basic univariate measures of the variables, I have put them in a table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Lung Capacity (cc) | Height | Exercise | Age |
| Mean | **5325.60** | **68.23** | **21.35** | **46.42** |
| Std. Deviation | **410.48** | **3.45** | **8.91** | **13.98** |
| Min | **4233.71** | **58.93** | **0.00** | **19.00** |
| Max | **6261.00** | **76.61** | **40.29** | **82.14** |

Summarizing the bivariate measures of the variables, I have put the Pearson's product moment correlation coefficients of every possible pairs in another table similar to the above one.

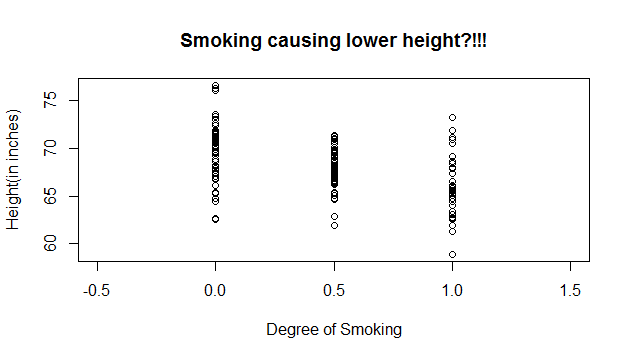
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlation Coefficient (r)** | Height | Exercise time | Age | Lung capacity |
| Height | 1 | 0.46356686 | 0.07136433 | 0.74270302 |
| Exercise time | 0.46356686 | 1 | 0.1600506 | 0.6204149 |
| Age | 0.07136433 | 0.1600506 | 1 | 0.10740713 |
| Lung capacity | 0.74270302 | 0.6204149 | 0.10740713 | 1 |

Judging the above table, and the plots (given later), I became interested in finding some more correlation between the conditional groups of people. Here are the details:

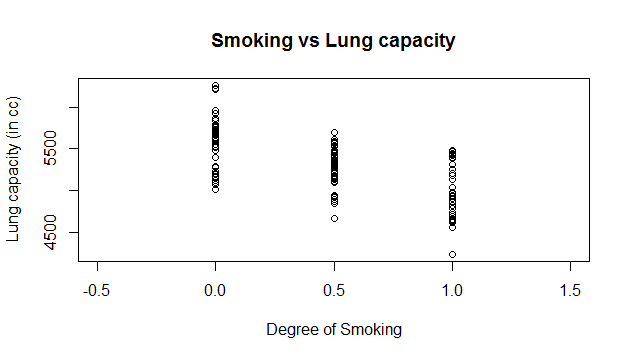
* Correlation coefficient between Exercise time and Lung capacity for males is 0.8145902, a strong linear relationship.
* Correlation coefficient between Exercise time and Lung capacity for females is 0.5497824, a weak linear relationship.
* Correlation coefficient between Exercise time and Lung capacity for smokers is 0.161776, almost no linear relationship. It is guessed that for smokers, Lung capacity doesn't affect much to their exercise time, probably because smokers have low Lung capacity and hence they cannot undergo through a long amount of time in exercising.
* Correlation coefficient between Exercise time and Height for smokers is 0.2588487, an extremely less linear relationship. It is guessed that for smokers, as they cannot afford a lot of time in exercising, Height is not at all affected.

Analysis of the Data

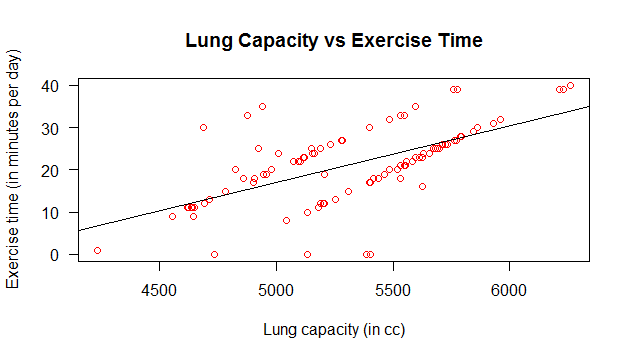
Firstly, I would like to point out the interesting relationship between Smoking and the height of the people through the figure (**Note:** Index 0 means non-smoker, 0.5 means non-addicted smoker and 1 means addicted smoker) below:



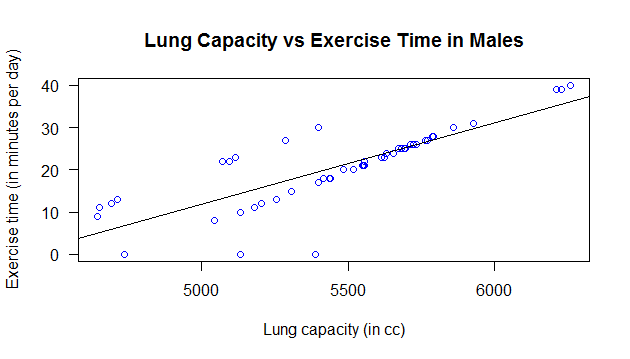
It is clearly seen that as the degree of smoking increases, the tendency of height goes down. But, it is highly impracticable to think that Smoking directly causes Lower height. Hence I dig further into the relationships of different variables of the dataset so that this seems plausible. It is known that smoking may cause Lower lung capacity and here is the scatter plot I found:

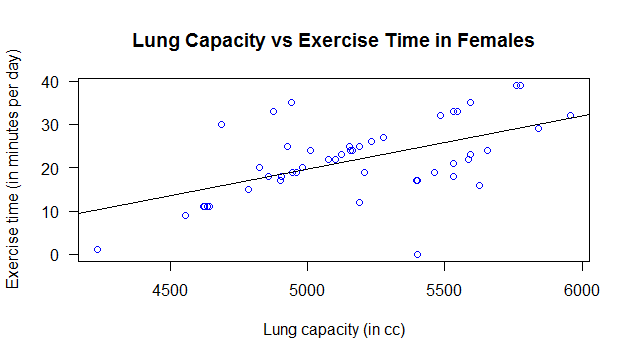


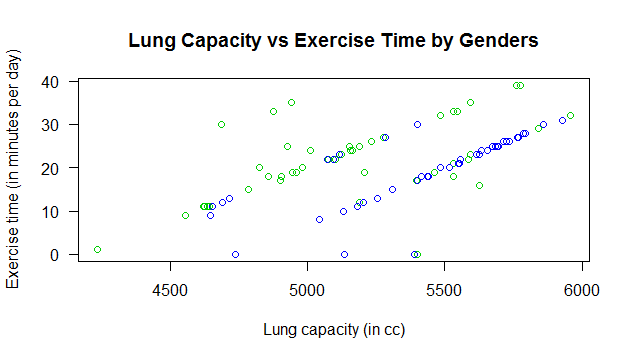
Here, the negative relationship between degree of smoking and Lung capacity (measured in c.c.) seems reasonable. Next, we should focus on how Lung capacity affects height of the people. These two things also seems uncorrelated but we will see how the correlation comes obvious as we bring in more variables. Below I have presented the scatter plot of Lung capacity and Exercise time.



In the above picture, the positive relationship between Lung capacity and Exercise time is reasonable due to the fact that who has more Lung capacity, he has more stamina and can keep up Exercises for a longer period of time. But, we also see that the regression line predicting exercise time by lung capacity is not well fitted as there are two groups of people differ by gender forming the linear relationship. Next few scatter plots shows the fact.



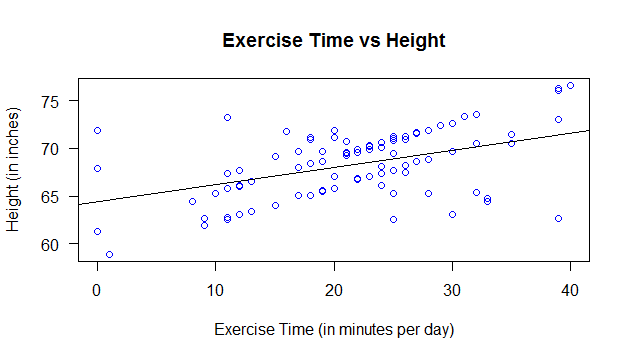




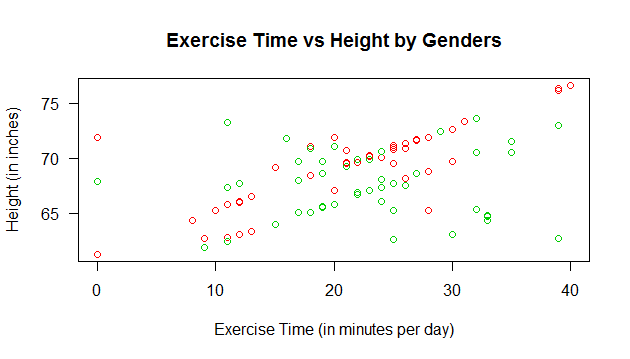
**N.B.-** *In last figure, green points are females while blue points are for males.*

From the very last figure, We can get a surprising result that with the same Lung capacity, a woman can do more amount of exercise than a man. This is probably because men's exercises are more intense and due to males' more BMR rate.

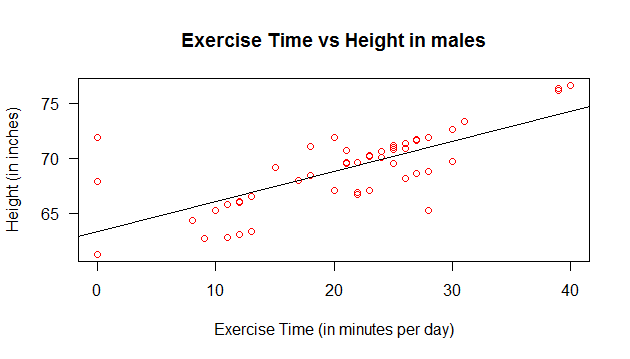
Now we see that lung capacity can affect exercise time of people. Now, we shall move towards showing the relationship between Exercise time and Height. Here, I have presented the following figure:

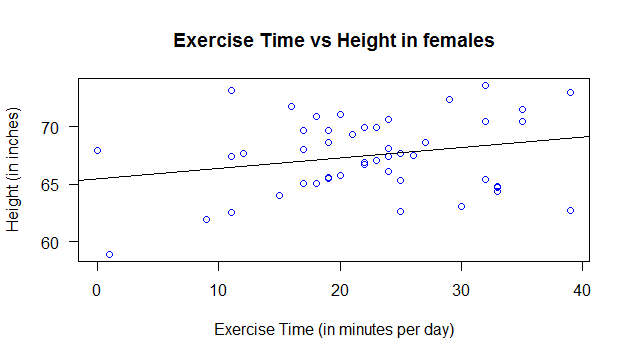


As we can guess, there is a strong positive correlation between Exercise time and the heights. It is because due to the exercises, peoples muscles gets stretched over time and their height increases. Also, the regression lines here, too, fit not so well. So, again the groups are divided by genders and the scatter plots found are presented below.



**N.B.-** *In last figure, green points are females while red points are for males.*





From above plots we see that Exercise time has a strong affect on Heights for males, while for females the relationship is not so strong. (suggested by the correlation coefficient between exercise time and height is about 0.5). For females, overall there is a positive trend in heights as Exercise time increases.

now, we see that the relationship between Lung capacity and Height is strong (suggested by the fact that correlation coefficient is 0.74270302) as the grouping of gender to bring in the third variable Exercise time collides for both the cases. Hence, at last we get the causing chain: Smoking causes lower Lung capacity, low lung capacity causes lower time given for exercise, and lower exercise time means lower height. But, clearly smoking does not causes lower height, it is just that both of them follow the same trend under the distributions of some intermediate variables.

Conclusions

From the above interpretation and analysis of the data we see that there are some unrelated variables that are correlated as seen in the scatter plots or by correlation coefficients, but one variable does not directly causes the other. Actually, if we bring in some intermediate variables to measure the relationship between the variables one at a time, i.e. we proceed one step at a time, then the causation seems pretty obvious through a chain of causation. As we can see for the aforementioned dataset, we have the following causation chain:

No Apparent causation and relationship

THANK YOU

------------------