Multiple Linear Regression

Saishab

November 19, 2021

Linear Regression

Linear regression is a simple statistical regression method used for predictive analysis.It shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable and the dependent variables.If we have only one independent variable and one dependent variable then it is said to be linear regression and if we have more than one independent variable then it is said to be multiple linear regression. You can get more Knowledge on Linear Regression from this article : https://www.digitalvidya.com/blog/linear-regression/.

The dataset contains observations on the percentage of people biking to work each day, the percentage of people smoking, and the percentage of people with heart disease.

Loading Data

```
data = read.csv("Data/heart.data.csv")
```

View Data

```
head(data)
```

```
##   X    biking   smoking heart.disease
## 1 1 30.801246 10.896608     11.769423
## 2 2 65.129215  2.219563      2.854081
## 3 3  1.959665 17.588331     17.177803
## 4 4 44.800196  2.802559      6.816647
## 5 5 69.428454 15.974505      4.062224
## 6 6 54.403626 29.333176      9.550046
```

Here we see we have unwanted Column X Lets remove it

```
df = subset(data,select = -c(X))
head(df)
```

```
##      biking   smoking heart.disease
## 1 30.801246 10.896608     11.769423
## 2 65.129215  2.219563      2.854081
## 3  1.959665 17.588331     17.177803
## 4 44.800196  2.802559      6.816647
## 5 69.428454 15.974505      4.062224
## 6 54.403626 29.333176      9.550046
```

Dimension of Dataframe : Which helps us to define Statistics of Data.

```
dim(df)

## [1] 498   3
```

So we have 498 rows(instances) and 3 Columns(Attributes).Lets See the Descriptive Statistics of our Data

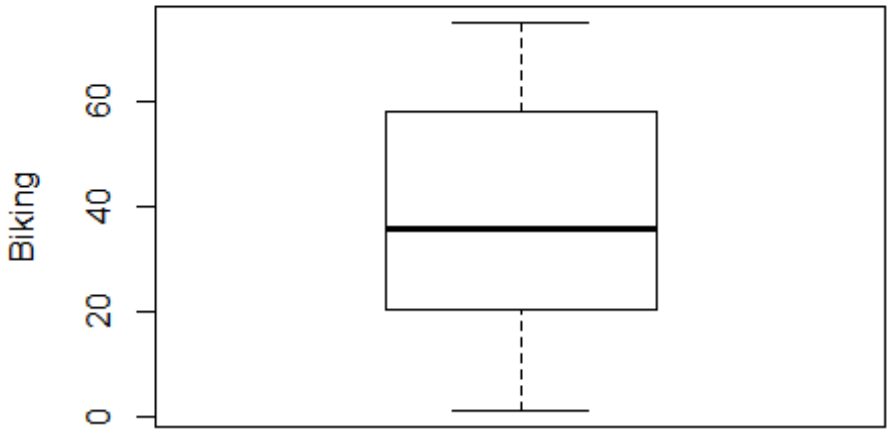
```
summary(df)

##      biking          smoking        heart.disease
##  Min.   : 1.119   Min.   : 0.5259   Min.   : 0.5519
##  1st Qu.:20.205   1st Qu.: 8.2798   1st Qu.: 6.5137
##  Median :35.824   Median :15.8146   Median :10.3853
##  Mean   :37.788   Mean   :15.4350   Mean   :10.1745
##  3rd Qu.:57.853   3rd Qu.:22.5689   3rd Qu.:13.7240
##  Max.   :74.907   Max.   :29.9467   Max.   :20.4535
```

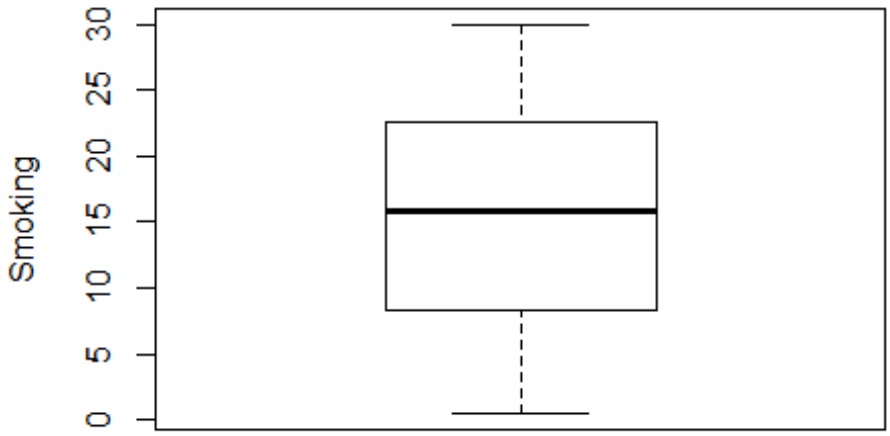
Here we can observe different Statistical Terms, 1. Min : Minimum Value in the attribute. This is also refered as 0th percentile as no value is less than this. 2. 1st Quartile : This is known as the lower or 25th empirical quartile, 25% of the data from our data set lies below this point.For Biking attribute we have 1st quartile as 20.205 which means 25% of or data falls below 20.205. 3. Median : Median is the mid value associated in the attributes , If the no. of observations(vectors) are odd then Median provides middle value, and if it is Even then provides average of two medians. 4. Mean : It provides average value of input vector. 5. 3rd Quartile : This is upper or 75th empirical quartile , 75% of the data from our data set have values less than 57.853 6. Max : Maximum Value in the attribute also refered as 100th Percentile as no value is more than it.

##Outliers Detection In Descriptive Stat of our data , we saw difference between maximum and minimum value is more this can happen because of outliers.Outliers are the data points which differes from the normal value in the dataset. As we can see we have 35.824 as median in Biking attribute and maximum value is 74.907 so this value might be the outlier in Biking , lets confrim this with Box Plot for each attributes.

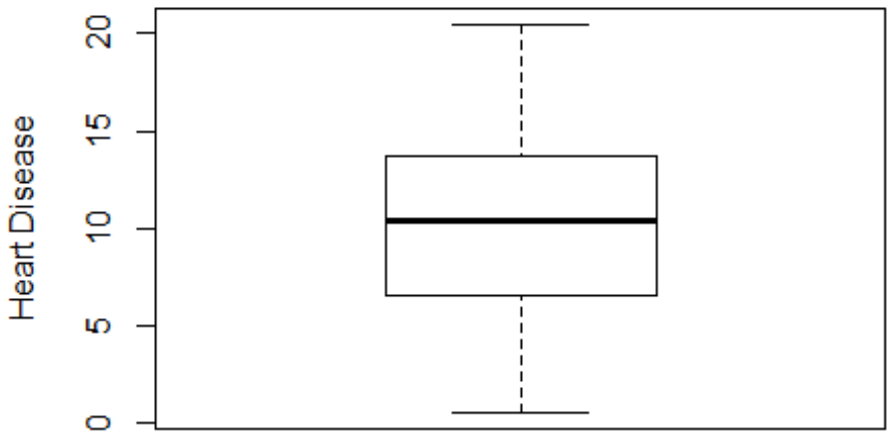
```
boxplot(df$biking,ylab = 'Biking')
```




```
boxplot(df$smoking, ylab = "Smoking")
```




```
boxplot(df$heart.disease,ylab = 'Heart Disease')
```



Here, we can see we don't have any outliers in the provided attributes.So we can move for Modeling of data. Since we are using Linear Regression so we should see the linearlity between the attributes and define them as dependent and independent varaiables

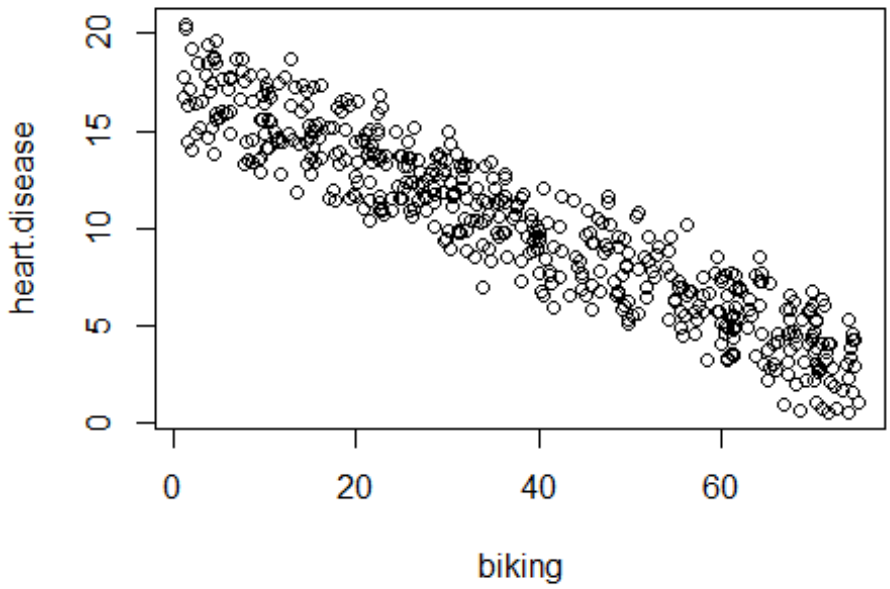
Smoking : Independent Varaiable Biking : Independent Varaiable Heart Disease : Dependent Varaiable

Problem Statement : Our dependent varaiable is dependent in independent varaiable, in simple word we are observing how well person's Heart Disease is dependent/ affected by their behavior of Smoking and Biking.

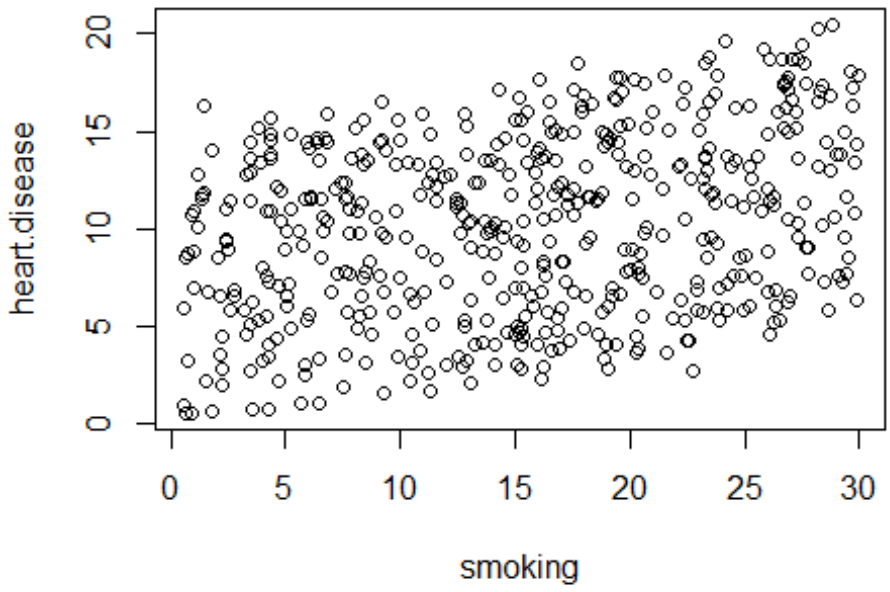
We can observe these dependencies by Plotting Scatter plot and using correlation fucntion.

1.Scatter Plot : A Scatter Plot plots the points that show the relationship between two sets of data. 2.Correlation : It shows degree in which relationship exist between variables. It ranges from (-1,1), -1 for strong negative relationship and +1 for strong positive relationship.

```
plot(heart.disease ~ biking , data = df )
```




```
plot(heart.disease ~ smoking , data = df)
```



Here from scatter plot, we can see the negative relationship between Heart Disease and Biking , and some sort of positive relationsip between Heart Disease and Smoking , lets check this with help of cor() function

```
cor(df$heart.disease, df$smoking)

## [1] 0.309131

cor(df$heart.disease,df$biking)

## [1] -0.9354555
```

The relation between Heart Disease and Smoking is 0.309131 which can be considered as +ve correlation. The relation between Heart Disease and Biking is -0.9354555 which is strong negative correlation.

Multicollinearity

This will be our problem if we have strong relationship between the independent variables, our independent varaibles should not have any form of dependencies within them

```
cor(df$biking,df$smoking)

## [1] 0.01513618
```

We see the correlation score is 0.0151 which is neglegible, so there's no any dependencies between the independent varaibles.

##Linear Model Our main task is to see the relationship between dependent varaiables and independent varaiables, so we fit a linear model with heart disease which is our dependent varaibale and biking and smoking is our independent variables. For this we use lm model. lm us used to fit linear models which is used for regression. The equation of Linear Regression is : y = mx + c , where m is our Coefficients, X our predictor and c is the contant i.e. y intercept.

```
linear_model = lm(heart.disease ~ biking + smoking ,data = df )
summary(linear_model)

##
## Call:
## lm(formula = heart.disease ~ biking + smoking, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1789 -0.4463  0.0362  0.4422  1.9331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.984658   0.080137  186.99   <2e-16 ***
## biking      -0.200133   0.001366 -146.53   <2e-16 ***
## smoking      0.178334   0.003539   50.39   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.654 on 495 degrees of freedom
## Multiple R-squared:  0.9796, Adjusted R-squared:  0.9795
## F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

Parameters : 1.Residuals : Residual is the difference between the observed value and the mean value that the model predicts for that particular observation. Here observation are the Min,1Q, Median,3Q,Max.

2.Coefficients and Intercepts :Coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In linear regression, coefficients are the values that multiply the predictor values.Intercepts are the expected mean value of Dependent variable(Y) when all dependent variables (X) is equal to 0.

```
cat('Heart Disease = ',linear_model$coefficients[1],' +
',linear_model$coefficients[2], '* Biking',' + ',
linear_model$coefficients[3],' *
Smoking')
```

```
## Heart Disease =  14.98466  +  -0.2001331 * Biking  +  0.1783339  *
## Smoking
```

Here : - Intercept : 14.98466 - Coefficients : -0.2001331 for Biking and 0.1783339 for Smoking

3.Signif.codes : They represent which attribute with p value is statisticaly significant.You can get more details from this article https://www.statology.org/significance-codes-in-r/

4. Residual standard Error : Residual Standard Error is measure of the quality of a linear regression fit and 495 degrees of freedom is given by the difference between the number of observations in sample and the number of variables in model.

5.Multiple R - Squared and Adjusted R - Squared : While using multiple indpendent varaiables in model the R square keeps on increasing. R square tells us about that how much variance is been explained by the model. Adjusted R - Squared calculate R square from only those variables whose addition in the model which are significant.

```

Conclusion of Project : 1.Estimated Effect of Biking on Heart Disease is -0.2001331 , means for every 1% increase in biking to work, there is decrease in heart disease by 0.20% . 2. Estimated effect of Smoking on Heart Disease is 0.1783339 , means there is 0.178% increase in Heart Disease for person who smokes.
```