AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Stephen Kiilu | Assignment Number: RWR1

Course: Regression with R | Date: December 5, 2020

Question 1

From my analysis :
Our data has three variable namely sex, age and Death rates per 10,000.
The sample size is 62. The average age is 35 years with standard deviation from mean of 9 years. The median age is equivalent to mean .
The average death rate per 10,000 is 12.65 and the middle death rate(median) is 10. The death rate per 10,000 has a spread of 8.95 from the mean . The correlation between age and death rate is 0.82. This implies that there is a strong positive linear relationship between the two. This means that as age is increasing the death rate is increasing.
The following graphics describes more about our analysis visually.

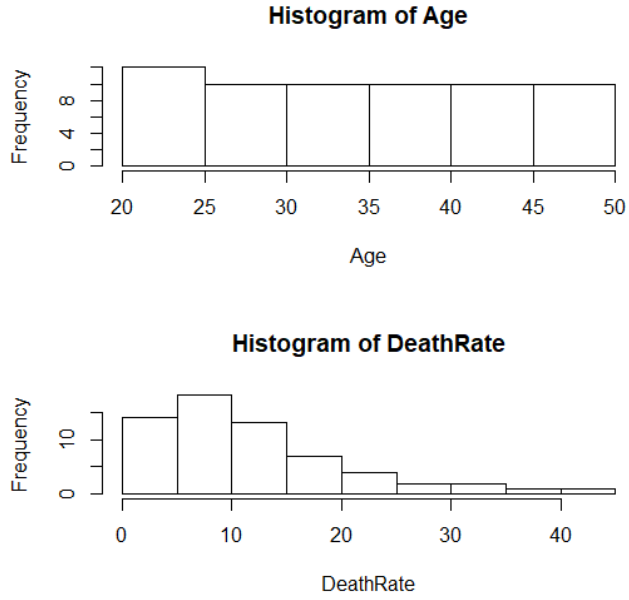

Figure 1: side by side histogram of age and death rate

From the the histogram of age, we have mode i.e highest counts of age are occurring between age 20-25 and evenly distributed among intervals of 5. From the histogram of death rate, the

maximum frequency of death rates is at interval 5-10 deaths rates per 10,000 and it decreasing in the other intervals of 5 from right to left.

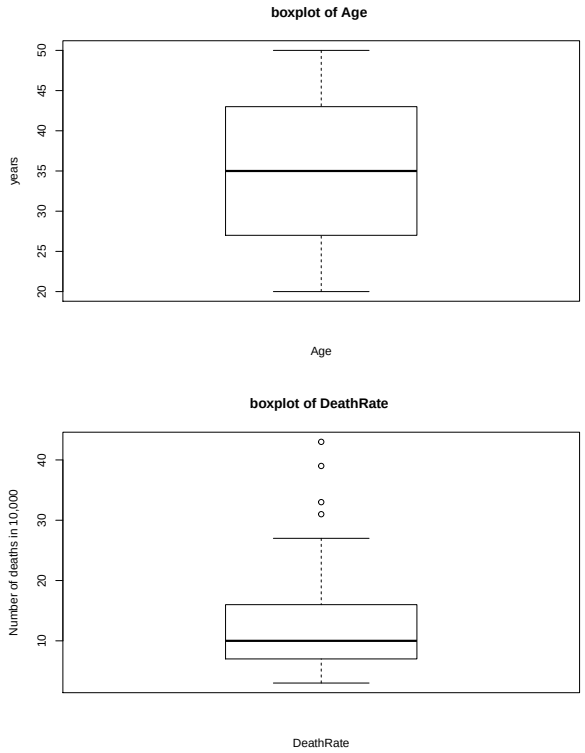

Figure 2: side by side box plot of age and death rate

From the box plot of age the median age is around 35 years and the spread from the median is fairly equal with no outliers in the data. From the boxplot of death rate, the median is around 10 deaths per 10,000 with the data more spread above the median. This indicates that we have more observation lying above our middle value.We have several outliers in our data, this may cause some question on how the data was collected.

Question 2

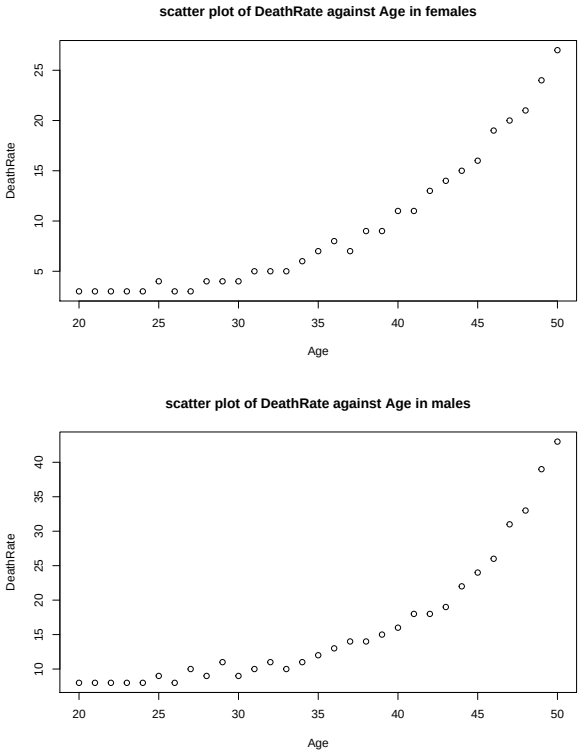

Figure 3: side by side scatter diagram of death rates against age in females and males

From the scatter diagrams above, it clear that there is a positive relationship between age and death rates in both males and females. This shows that in both females and males, there is an increase in number of death rates with an increase in number of years.
The coefficient of correlation of age and death rates in females and male is 0.92 and 0.88 respectively. This shows that there is a strong linear relationship between age and death rates in 10,000 in both females and males. We can interpret this to mean that an increase in number of years corresponds to increase in number of deaths in 10,000.

Question 3

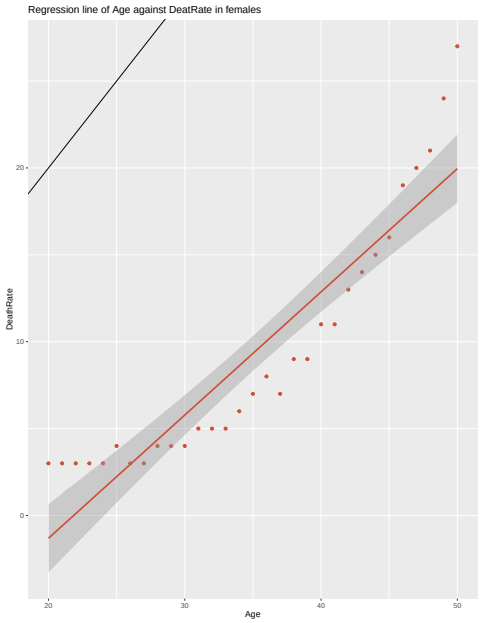

Figure 4: Linear regression line

The diagram above shows the regression equation of age against death rates in females. The regression model is given by:

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0.$$

where $\hat{y}$ is the death rates per 10,000, $\hat{\beta}_1$ is the slope of the regression line, x is the age and $\hat{\beta}_0$ is the intercept of the regression line. Our regression model is therefore:

$$\hat{y} = 0.7089x - 15.4879.$$

Question 4

We replace age 51 in our regression model

$$\hat{y} = 0.7089x - 15.4879.$$

and obtain the predicted death rate in females at aged 51 as 20.66.

Question 5

We are going to use R^2, the coefficient of determination and mean standard error (MSE) to asses the quality of our model. From our data R^2 is 0.852 which means that the around 85% of total variation in the data is explained by our model. The R^2 is large and is an indication that our model is good. But when we look at our MSE, which from our data is around 7, and for a good model we expect mean squares error to be minimum and close to zero. This cast doubts in the quality of our model. Basing on our MSE we can conclude that our model is not the best and can be improved. Some of ways to improve our model include; adding more data, normalize the data, check for collinearity problem and checking for outliers.