Statistics about heart attacks

Jason Liu
Science
UAntwerpen
Middelheim, Antwerp
Jason.Liu@student.uantwerpen.be

***Abstract*— The aim of the study was to describe factors associated with trends over time survival after hospital admission for acute myocardial infarction.**

***Index terms*—Paper, Report, Research**

I. Dataset

The dataset got formatted as follows:

- Row 1, 3, 9, 11 and 17 got deleted based on student number "20213082".

II. Questions

A. Study and discuss the distribution of the variable separately. Discuss appropriate graphics for this purpose performances. Also formally check whether the data is normally distributed. If this is not the case, in what way do the data deviate from normally distributed facts. Can you transform the data into normally distributed data? Discuss.

Normality can be tested in different ways, we will test this statistically using some test and graphically using some graphs.

1) *Los:*

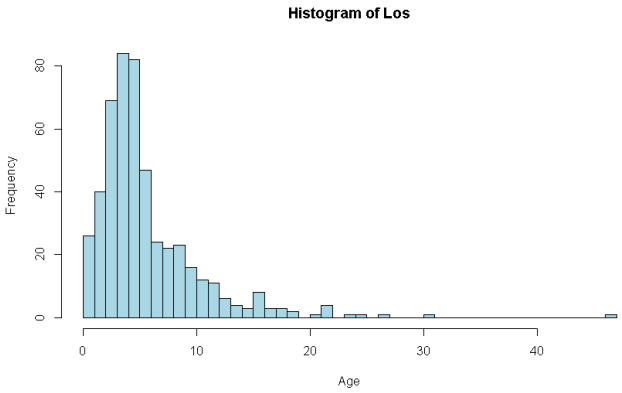
Statistical test: Let's say our null hypothesis is that it is normally distributed. That means our alternative hypothesis is not. Also we will make use of $\alpha = 0.05$.

We make use of a Shapiro-Wilk test to test if it is normally distributed.

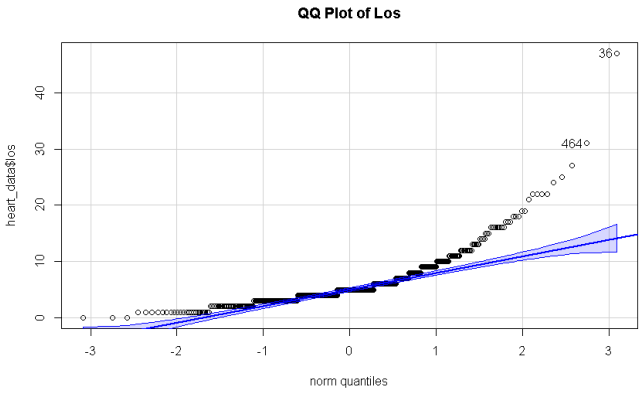
```
Shapiro-Wilk normality test

data:  heart_data$los
W = 0.76626, p-value < 2.2e-16
```

We can conclude from the output of the p-value, that it is not normally distributed, because p-value $< \alpha$.

Graphical test:

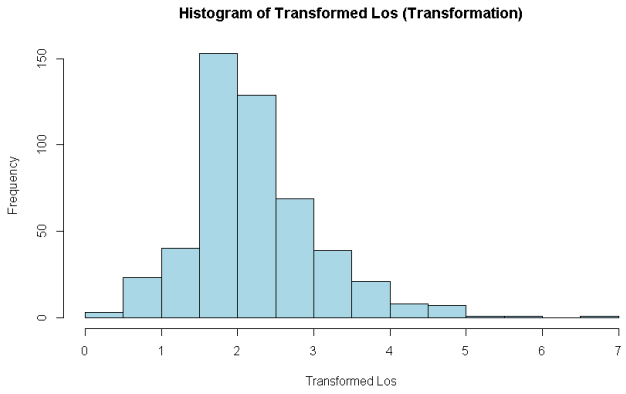


From the bar chart we can tell that the data is right-skewed. This indicates that it is not normally distributed.

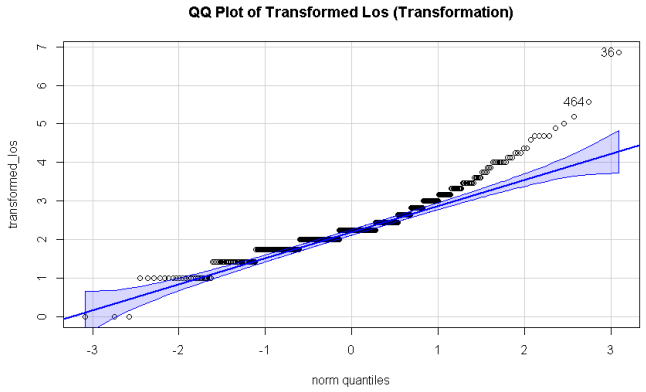


The QQ plot gives the same indication, because the data forms a cup shape, we can tell from that it is right skewed.

Transformation: We can apply square root transformation and see what this results.



It does resemble closer to a normal distribution, but it is not quite there yet.



Looking at the QQ plot, we can see the cup form again, indicating it is still right-skewed.

```
Shapiro-Wilk normality test

data:  transformed_los
W = 0.93406, p-value = 5.64e-14
```

The test results are showing similar results, saying that it is still not normally distributed.

Applying more transformation than this, is not possible, resulting in infinite y values (log10, 1/x).

B. Check whether there is a connection between the type of myocardial infarction and the discharge status from the hospital after admission. Perform an appropriate test.

We will make use of a chi-square test to check the correlation between these two binary variables.

Null Hypothesis (H0): The two binary variables are independent. Alternative Hypothesis (H1): The two binary variables are not independent.

```
Pearsons Chi-squared test with Yates continuity
correction

data:  table(heart_data$mitype, heart_data$dstat)
X-squared = 0.02069, df = 1, p-value = 0.8856
```

The p-value $> \alpha$, so in other words we do not reject the null hypothesis. So the two binary variables are independent and have no correlation.

C. Can you predict the BMI from the patient's age? Please answer this question thoroughly and as completely as possible.

So before we test any corelation, we test if both variables are normally distributed.

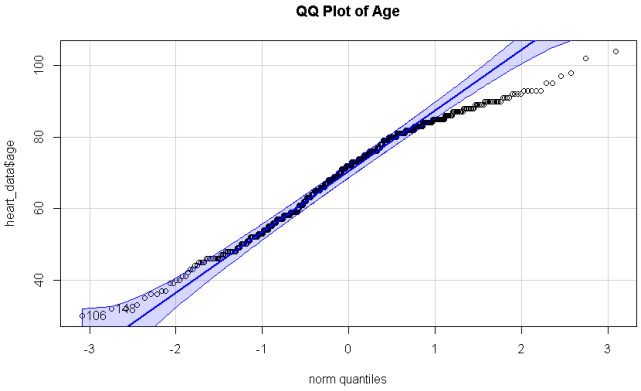
1) *Age:* Let's say our null hypothesis is that it is normally distributed. That means our alternative hypothesis is not. Also we will make use of $\alpha = 0.05$.

We make use of a Shapiro-Wilk test to test if it is normally distributed.

```
Shapiro-Wilk normality test

data:  heart_data$age
W = 0.97381, p-value = 9.633e-08
```

P-value $< \alpha$, this concludes that it is not normally distributed and we can disregard the null hypothesis.

We can check the QQ-plot, to check where it deviates from a normal distribution.



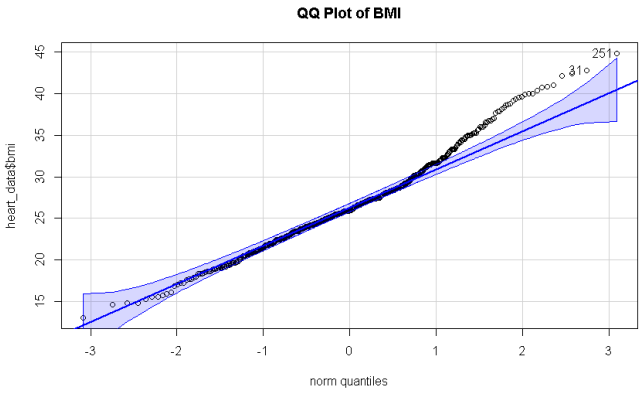
The QQ-plot tells us the following:

- The data forms more of a cap shape, that result indicates it is more left-skewed.

2) *BMI:* The same null hypothesis and alternative hypothesis can be applied for BMI.

```
Shapiro-Wilk normality test

data:  heart_data$bmi
W = 0.97965, p-value = 2.111e-06
```

Again here, we can see it is not normally distributed. P-value is again less than α, indicating that we can disregard the null hypothesis.



Again we can form a QQ-plot to check where it deviates. We can see it has a slight tail and indeed it is not normally distributed.

3) *Corelation test:* We apply two tests, both can be used to check if two non bivariate variables are corelated.

```
Spearmans rank correlation rho

data:  heart_data$age and heart_data$bmi
S = 28469120, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.4083528
```

```
Kendalls rank correlation tau

data:  heart_data$age and heart_data$bmi
z = -9.1831, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.2789187
```

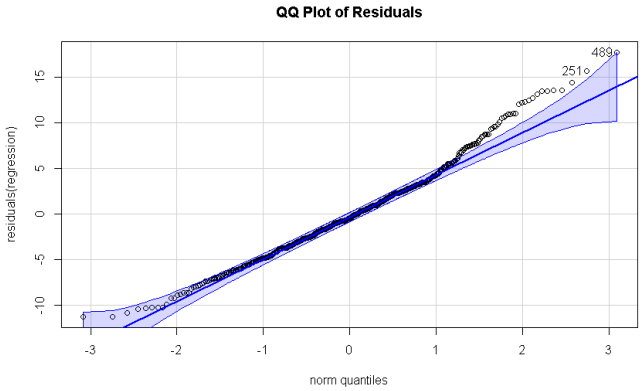
The null hypothesis is that both are independent, the alternative hypothesis indicates the opposite (they are dependent).

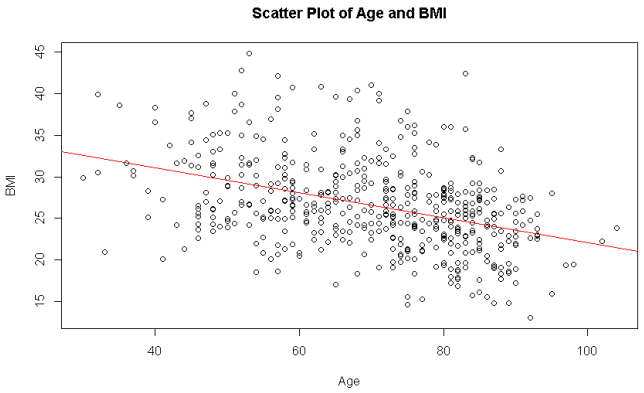
From both test we can see that the alternative hypothesis is true, because p-value $< \alpha$.

They have a slight corelation, because the corelation of both test are rather low.

4) *Linear regression:* Now we can perform some linear regression to check more information regarding the relationship of BMI and age.

```
Call:
lm(formula = heart_data$bmi ~ heart_data$age)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2485  -3.4354  -0.3852   2.8019  17.7449

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.09715    1.09944   33.742   <2e-16
***
heart_data$age -0.15011    0.01542  -9.737   <2e-16
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 4.967 on 493 degrees of
freedom
Multiple R-squared:   0.1613,   Adjusted R-squared:
0.1596
F-statistic:  94.8 on 1 and 493 DF,  p-value: < 2.2e-16
```





Looking at the residuals we can conclude following:

- Min (-11.2485): This is the minimum residual, meaning the largest negative difference between an observed value and the predicted value. It suggests that for at least one data point, the predicted BMI was 11.2485 units higher than the observed BMI.
- 1Q (-3.4354): This is the first quartile of the residuals, meaning that 25% of the residuals are less than or equal to −3.4354. It indicates that a quarter of the predictions are off by more than 3.4354 units on the lower side.
- Median (-0.3852): This is the median residual, which means that half of the residuals are less than −0.3852

and half are greater. A median close to zero indicates that the model does not have a systematic bias of overestimating or underestimating the BMI.

- 3Q (2.8019): This is the third quartile of the residuals, meaning that 75% of the residuals are less than or equal to 2.8019. It indicates that three-quarters of the predictions are within 2.8019 units of the observed values on the higher side.
- Max (17.7449): This is the maximum residual, meaning the largest positive difference between an observed value and the predicted value. It suggests that for at least one data point, the predicted BMI was 17.7449 units lower than the observed BMI.

Overall let's we can see that the Min and Max difference is rather wide. This observation tells us that some prediction are far off from the observed values.

Next, the median residual is close to zero (-0.3852), which suggests that the model does not have a significant systematic bias. However, the first quartile is much closer to zero than the third quartile, indicating that the model's errors are not perfectly symmetric around zero.

Let's look at coefficients:

- (Intercept): The intercept is 37.09715, which is the estimated BMI when age is 0.
- heart_data$age: The coefficient for age is −0.15011. This indicates that for each additional year of age, the BMI is expected to decrease by approximately 0.15011 units.

Next, we can tell from the significance codes, that this relationship is 99% real. This improves the trust we can have for this model.

Residual standard error tells us the following:

- This value indicates the average amount by which the observed BMI values deviate from the predicted values.
- Freedom of 493 degrees.

The R-squared values tells us the following:

- Multiple R-squared (0.1613): This indicates that approximately 16.13% of the variability in BMI can be explained by age.
- Adjusted R-squared (0.1596): This value adjusts the R-squared value for the number of predictors in the model. It is slightly lower than the R-squared, suggesting that the model has only one predictor and the adjustment is minimal.

Last but not least, F-statistic:

- The F-statistic tests whether at least one of the predictors is significantly related to the response variable. The high F-statistic (94.8) and the very low p-value (< 2.2e-16) indicate that the model is statistically significant.

5) *Conclusion:* We can conclude that age does have an influence on BMI but that the influence is rather minimal. The influence has a negative or inverse impact, the older we get, the less the BMI becomes.

Looking at the scatter plot, we can conclude similar things. One factor we have to account for is that other variables need to be accounted for.

References