AUTHOR: MARTIN KATKOVČIN

PREDICTING SIGNS OF A HEART ATTACK

Each of us has certainly heard of a heart attack, whether as a very common cause of death in the elderly, but lately this sudden closure of the heart artery has been as common as possible even in people who are not expected to do so at all.

Thus, among the most common causes of myocardial infarction, increased blood pressure, obesity, smoking, excessive alcohol consumption, high blood clotting is inherently valid... We will deal with a large substance that belongs to one of the most common causes of the development of a heart attack is the deposition of cholesterol into the wall of the artery, through which this artery narrows, and subsequently at high blood flow an atherosclerotic canvas can burst, in this way there will be a partial or even complete closure of the artery and subsequent non-bleeding of the heart muscle.

In our research, which we carried out on data that came from trusted sources about patients from the hospital, who were measured real values, as well as values of quantities that are closely related to the development of a heart attack. Finally, we have chosen to investigate the relationship between the amount of cholesterol value and the age of a given patient, whether there is a real relationship between these quantities, or whether these quantities are also closely related to some other factors that can influence the development of a heart attack. The wording of the hypothesis we raised was: 'We assume that cholesterol levels increase with increasing age, which means that older people are more likely to have *a heart attack."* So, our zero hypothesis was that we assume that the value of cholesterol with increasing age remains the same, which we will try to refute using research methods and, through linear regressive models and statistics associated with them, confirm the original wording of the hypothesis.

One of the first things we wanted to start exploring was to prepare our data so that we could not experience collisions through deflected values, missing values, duplicate lines and problems of a similar type that the dataset may initially suffer from. After successful *data cleaning,* we did basic descriptive statistics as well as pair analysis, both between cholesterol and age, where we also had an indicator that simulated whether the patient had already had a heart attack.

To confirm that we had a good hypothesis and could have moved on with the research, we needed to calculate the Pearson correlation *coefficient,* which determines the degree of intensity of linear correlations between the selected pair of elements studied. This coefficient can be taken values from the range *<-1, 1>*, where *if* we have values that are non-zero negative, it means that if one of the variables rises and the other, on the contrary, the second quantity decreases, on the other hand, if we are in a non-zero non-negative plane, then it is a direct dependency, which means that the increase of one quantity also caused an increase in the other. and if we get a value of 0, the relationship between

the selected attributes does not exist. In our case, between the attributes cholesterol and age we acquired a value of $\rho \doteq 0.21$, which indicates to us that the relationship will be directly proportional here, but we can test that we take all the values of the coefficient β0 and β1, where we monitor whether the values are significant from each other, which means that the p-value for the F-test of the overall significance is less than the meaning level (p-value < 0.05), so we can reject the zero hypothesis and evaluate that our model provides a better fit which is for intercept-only model. This confirmed to us that we can reject the zero hypothesis, but only if we take the data that we have read from the original table, but when we generated 150 models at random, but it was enough for us to find one that did not meet this condition, so we could not reject the zero hypothesis, because p-value is a very strong value that indicates a very high probability of a zero hypothesis occurring. P-value, we also calculated using Shapiro's test, which is also a statistical test that can reject zero hypotheses, but we also confirmed the scenario that only in the original dataset can we refuse, but in other subsets (generated models) that we have created sometimes we can not, we also displayed this on *quantile-Quantile plots (QQ plots)*, which helped us only visualize, whether the attributes are from normal distribution or not. We must bear in mind that we can accept the zero hypothesis, we can only reject it or not prove it. β0 and β1 values representing the intercept (level constant) and slope values that come from the relationship for the linear regression curve (straight line), $Y \approx \beta_0 + \beta_1 X$, where we have also defined predictor (X) and response (Y) values, where the predictor is the patient's age and response is the patient's cholesterol value, where we can also generally describe that β1 is the regression coefficient on which the direction of inclination and steepness of the slope of the regression curve depend.

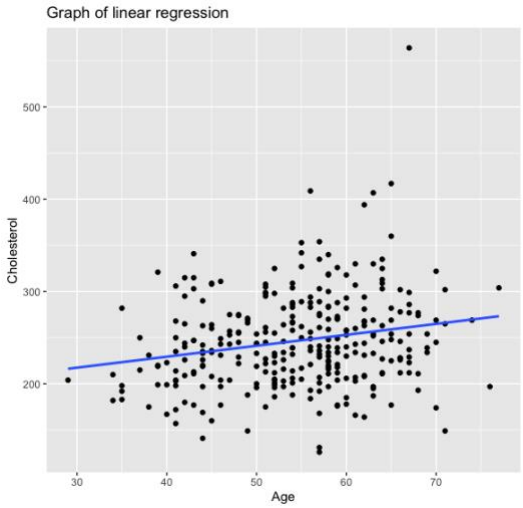

Figure 1. – Graph of linear regression model

In this way, we managed to display a graph of the linear regression model from the original data, where we can see, as we expected, also confirmed according to the slope coefficient of the curve, that the curve will only rise slightly, and even, we can read from the graph that from increasing age the cholesterol value also increases directly proportionally, but we can also see, that between the ages of 55 and 60 we can also see so-called residual clumps (residual – a measure of vertical distance from

the regression line – error between the predicted value and the observed actual value), and we can also say with certainty that from the age of 50 to 60 years most patients were in our original dataset. In the final phase of the research, we wanted to use the Cross-validation method, which estimates the error rate of the method for the created subsets of training observations during the fitting process. This method consists of 6 steps, where some of them we have already described above. The first step is to create subsets to apply the model. Subsets where we randomly select values for our model from the entire dataset of a certain size. In the second phase of the cross-validation method, we adapt the linear regression model to each subset that we created in the first step. Next, we retain coefficients β0 and β1 and residues from each model. The standard error of the coefficients β0 and β1 is calculated as follows. The standard deviation β0 gave us 18.79, where we had a residual average of 200, which is a negligible error, and also β1 we calculated a value of 0.36, where the values were around 1.5, is also a decent value of all the models tested. Through this step, we have calculated what the error rate will be at the slope of the curve and the level constant. In step five, we calculate the residual sum of squares (RSS) and residual standard error (RSE), which we will interpret and be shown on the graphs. RSS is a technique used in statistics to measure the amount of deviations in a data set, but we will be more interested in RSE, where we will see an average model error. It is used to express the degree of difference between the values that have been predicted by the model. In the graphs we can see that the average error is something on average around 52.5, with our residues gaining values on average around 200, where we completed about 25% (0.25) of the standard error, which ultimately is a low error value.

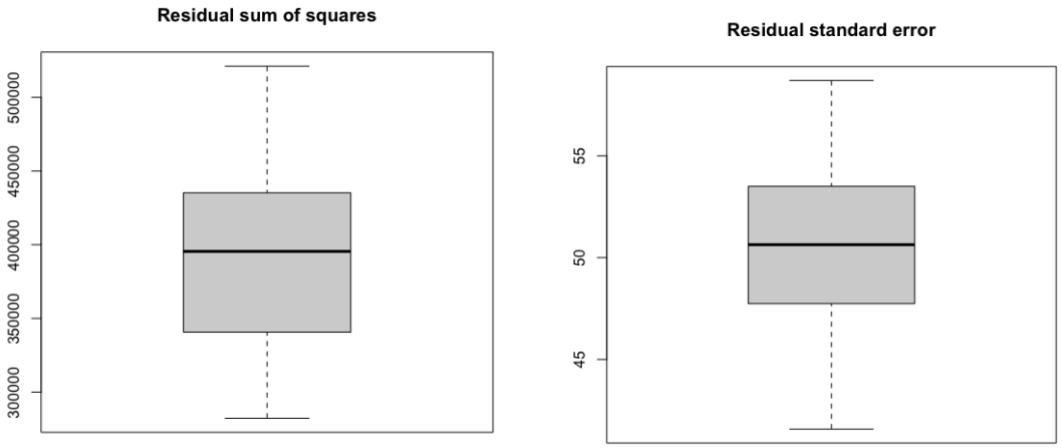

Figure 2. – RSS and RSME Graph

The last step *in the cross-validation* method for *displaying the relationship between response and predictor* via *Student't-Test*, which refuted the 0 hypothesis when we put our all the coefficient models in there, *so we came out with a p-value < of 0.05* several times less than the value of the extreme interval, so we can ultimately declare that we have managed to refute the hypothesis 0 and confirm the hypothesis 1, which read: 'We assume that cholesterol is increasing with increasing age, which means that older people are more likely *to have a heart attack.'*