Determinants of Heart Disease, A Precursor to Machine Learning

Eduardo Abdala

12/2/2019

Introduction

In the medical field, a strong knowledge of statistics is necessary to ensure that the assumptions and inferences we make in patient data are mathematically sound. Making incorrect inferences on patient sample data can have profound negative effects. Today, with the widespread use of machine learning, we can make better predictions on data, but we have to first explore the data and make hypotheses for testing. The following analysis is an attempt at describing feature importance through statistical inference. Ultimately, researchers of heart disease would perform tests to learn more about their data and then run the data through various machine learning algorithms to predict the presence of heart disease in a way that is too computationally intensive for regular methods.

Project Goals:

Provide a survey of various statistical inference methods with the following:

1. Check for normality in data
2. Check for feature independence in the data (Chi-Square)
3. Use permutation tests for hypothesis testing
4. Interpret the results and give a conclusion

Description of the data

This dataset is taken from kaggle.com and sourced from the UC Irvine Machine Learning Repository. We have a dataset of n = 303 with 13 features and one target column used to train a classifier machine learning algorithm.

The features are as follows:

1. age
2. sex (male = 1, female = 0)
3. cp = chest pain type (values 1-4)
4. trestbps = resting blood pressure (mm Hg)
5. chol = serum cholesterol (mg/dl)
6. fbs = fasting blood sugar (> 120 mg/dl)
7. restecg = resting ECG results (values 0-2)
8. thalach = max heart rate acheived
9. exang = exercised induced angina
10. oldpeak = ST depression induced by exercise
11. slope = slope of the peak exercise ST segment
12. ca = number of major vessels colored by flouroscopy (values 0-3)
13. thal: 3 = normal, 6 = fixed defect, 7 = reversable defect
14. target = presence of heart disease (absent = FALSE, present = TRUE)

Analysis was done in RStudio using R v.3.6.1

Exploratory Data Analysis (EDA)

We first observe the distributions of our sample data (Figure 1). We are working with mostly categorical data with a few numeric variables. As for the numeric features, we see that our data is heavily skewed, with the exception of age. To check for normality, we draw a qqplot. The output (Figure 2) clearly tells us that the data is not normal because our data points do not all fall along the diagonal line. This is expected because researchers studying heart disease will favor older patients.

Figure 1

```
##       age        sex     cp         trestbps          chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
##  restecg    thalach      exang      oldpeak         slope
##  0:147   Min.   : 71.0   0:204   Min.   :0.00   Min.   :0.000
##  1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00   1st Qu.:1.000
##  2:  4   Median :153.0           Median :0.80   Median :1.000
##          Mean   :149.6           Mean   :1.04   Mean   :1.399
##          3rd Qu.:166.0           3rd Qu.:1.60   3rd Qu.:2.000
##          Max.   :202.0           Max.   :6.20   Max.   :2.000
##        ca         thal      target
##  Min.   :0.0000   0:  2   Mode :logical
##  1st Qu.:0.0000   1: 18   FALSE:138
##  Median :0.0000   2:166   TRUE :165
##  Mean   :0.7294   3:117
##  3rd Qu.:1.0000
##  Max.   :4.0000
```

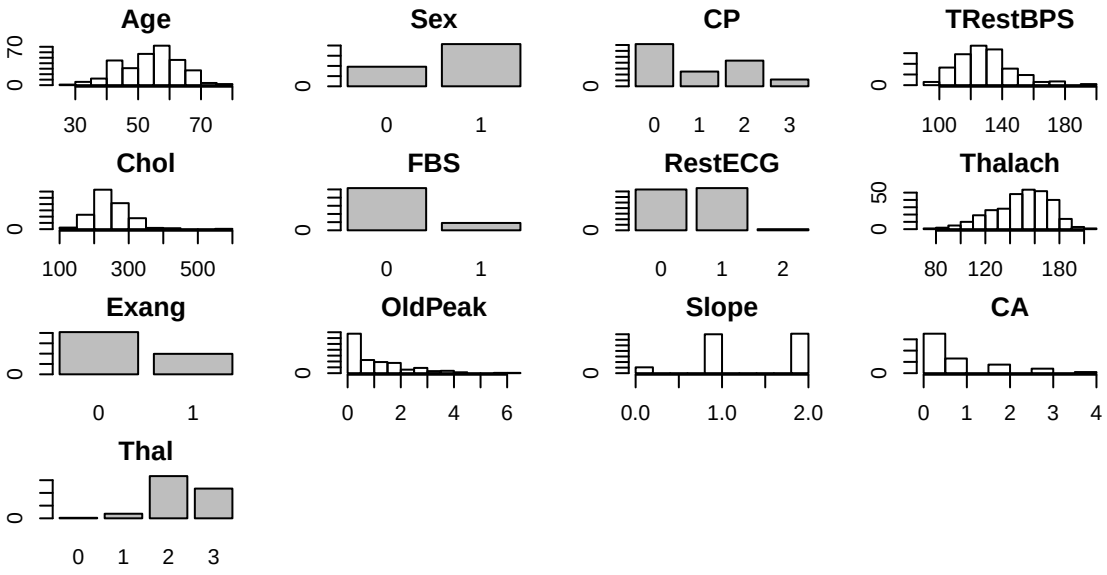
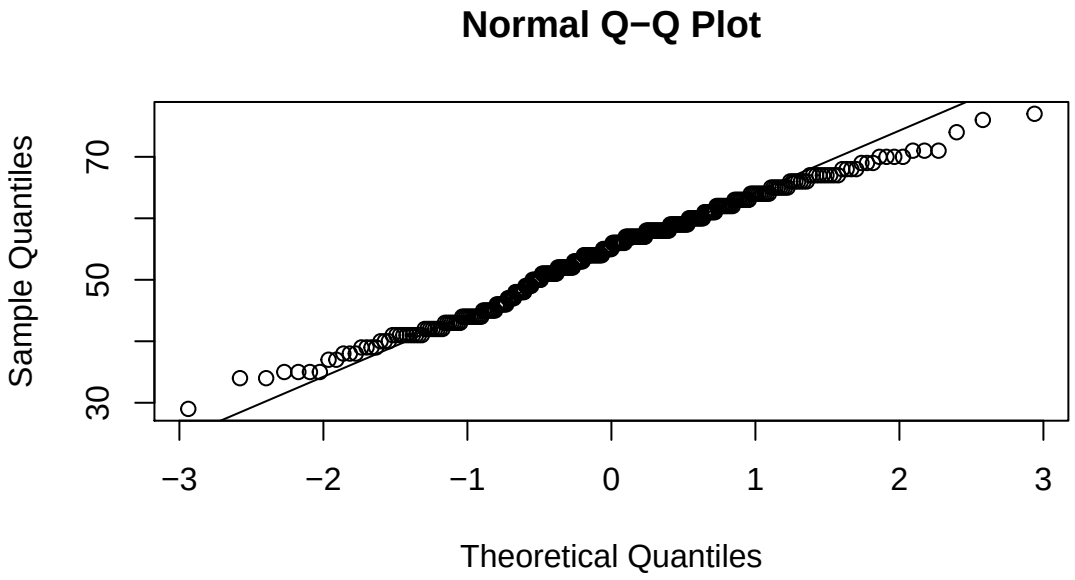


Figure 2



Feature Analysis - Numeric Data

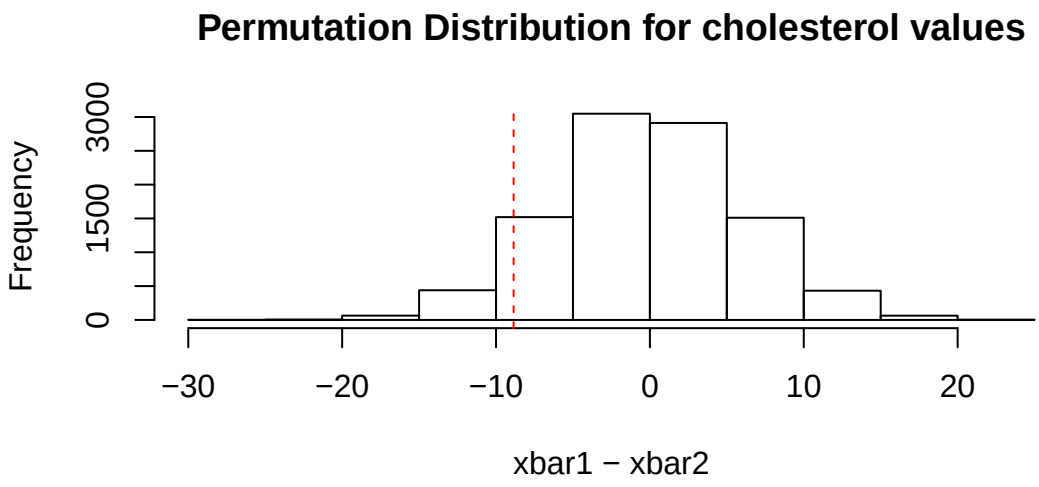
This data set has 13 factors that could influence the presence of heart disease in a patient, some more than others. To see if we can remove some variables from consideration and lower the amount of analysis we have to do, we create a correlation matrix. Features with high correlation are more lienarly dependent and have the same effect on out target column. If we find such variables, we pick one out of the two.

```
##             age trestbps  chol thalach oldpeak slope    ca
## age        1.00     0.28  0.21   -0.40    0.21 -0.17  0.28
## trestbps   0.28     1.00  0.12   -0.05    0.19 -0.12  0.10
## chol       0.21     0.12  1.00   -0.01    0.05  0.00  0.07
## thalach   -0.40    -0.05 -0.01    1.00   -0.34  0.39 -0.21
## oldpeak    0.21     0.19  0.05   -0.34    1.00 -0.58  0.22
## slope     -0.17    -0.12  0.00    0.39   -0.58  1.00 -0.08
## ca         0.28     0.10  0.07   -0.21    0.22 -0.08  1.00
```

After running this code, we find that no numeric variables have a correlation greater than .9. We move on to other methods.

For a long time, we have been told that high cholesterol is a good indicator for heart disease. We can run a permutation test on our serum cholesterol (LDL + HDL + triglycerides) data to see if there is a difference between the mean serum cholesterol of those with heart disease and those without. Our hypotheses are: $H_o : \mu_1 = \mu_2$ and $H_A : \mu_1 \neq \mu_2$. We assume the null hyothesis is true and pool all of our patient data together, resample them without replacement, and calculate our test statistic. In doing this N times, we can compare the test statistic distribution and then compare the observed test statistic against the distribution.

```
##    FALSE     TRUE
## 251.0870 242.2303
```




```
## [1] 0.1402
```

Above are the mean serum cholesterol values for the target groups. The result for this two sided permutation test is a p-value greater than 0.05. Thus we find that the two groups did not have a significantly different serum cholesterol level. This is an interesting contrast to the information that doctors give us. The explanation could be that this calculation includes "good" cholesterol (HDL) and "bad" cholesterol (LDL). Maybe splitting up the values changes the result. Another explanation could be that high cholesterol does influence heart disease, but it needs other factors as well.

Chi-square: Analysis of Categorical Data

Suppose that we want to see if there is a link between sex and the presence of heart disease. The Chi-square test for independence is the necessary method to see if there is a dependency between two factors. Note: Our sample size and expected counts for all cells are large enough, making this test appropriate.

```
##
##     FALSE TRUE
##   0    24   72
##   1   114   93
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sexcontingency
## X-squared = 22.717, df = 1, p-value = 1.877e-06
```

A low p-value is observed and the conclusion is that there is a dependency between sex and the presence of heart disease. In this sample, of the 96 females, 72 had heart disease (75%). On the other hand, 45% of our sample's males had heart disease. Clearly there is something else happening in the data. It would not make sense to stop here and say that being female is one of the causes of heart disease. While the argument can be made that women are more likely to have heart disease, we can look at the other factors to see if we can find other significant results.

If we do this for resting ECG results, we get a similar result.

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 9999
##  replicates)
```

```
##
## data:  ecgcont
## X-squared = 10.023, df = NA, p-value = 0.0036
```

As we run more tests, we start to see that some variables are giving us significant results while others are not. Further, we introduced the problem of certain features such as sex and serum cholesterol giving us answers that do not quite make sense logically. Humans cannot empirically weigh all of the significant factors together without other mathematical models. The solution is to pool all of these features together and create a prediction.

Future Directions

Only some of the variables were considered in this analysis. In the real world, more variables will be considered and some multivariate statistical methods would be used to get a better picture of the causes of heart disease. Here, we set the foundation for introductory study into this complicated dataset.

We learned that women are significantly overrepresented in the sample with heart disease. We also learned that cholesterol is either not a cause of heart disease or that it needs to be combined with other factors. Lastly, we learned that resting ECG results also have a part to play in predicting heart disease.

With more time, a decision would be made as to what features will be included in a machine learning model. These decisions (a process which is called feature engineering) will influence the accuracy of the model. This is why it is imperative that a data scientist have a strong knowledge of mathematical models, not just knowledge of classifier algorithms. In the medical field, accuracy can mean the difference between life and death.

References

Chihara, Laura, and Tim Hesterberg. Mathematical Statistics with Resampling and R. Wiley, 2019.