Factors that Influence the Chance of Heart Attack

Bihan Lu

08/12/2020

Abstract

As the world continues to grow, the Internet and technologies develop rapidly. People tend to have long working hours and more pressure during their daily life. Meanwhile, a heart attack has always been the most significant substantial hidden danger to human health since the heart is considered as the most crucial feature of living-beings. In this paper, a logistic regression model is fitted on "target" from "Heart" dataset. (Naresh, 2020) to determine the factors that cause a high chance of heart attack.

Keywords: Heart Disease, Heart Attack, Health Care, Heart Condition, Health, Health Condition, Logistic Regression, Binary Classification

This study is also available at: https://github.com/tiffanyylu/Factors-that-Influence-the-Chance-of-Heart-Attack

Introduction

The speed of life has now become more compact with the advent of remote equipment and communication technology, allowing people to work longer, and the pressure on employees is also rising rapidly. At the same time, people generally do not take good care of their bodies in such an atmosphere, leading to the onset of different diseases. Cardiac disease is one of the most common circulatory system disorders, which often cannot be stopped from dealing a fatal blow to patients because of its sudden, unpredictable occurrence. It is an immense health threat.

Therefore, I want to identify the high-risk and low-risk populations in this project and find out which factors are more significant relatively. By using supervised learning through 14 variables that may be correlated with heart disease, such as the age of the patient, gender, type of chest pain, resting blood pressure, fasting blood sugar, and so on, statistical models are generated and analyzed.

In this paper, one dataset, called it "Heart" dataset (Naresh, 2020), will be used to investigate what factors are positively correlated to the chance of heart attack. The detail about data will be discussed in the Data section. In the Model section, the logit behind logistic regression and the model equation will be addressed there. Results of the data and model analysis, including summary tables and graphs, are provided in the Results section. Interpretation of the results, weaknesses, and further steps will all be discussed in the following Discussion section. The reference section is at the end of this paper.

Data

The original dataset can be found on the UCI Machine Learning Repository; it is collected by four Medicinae Doctors (M.D.).

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

These 303 observations with 76 attributes in the dataset are donated by real patients and collected by the above four Doctors. Some missing values exist in the original dataset with 76 variables, but all published studies apply to a subset of 14 of these. Therefore, for this paper, a shorter dataset (Naresh, 2020) is used, and it can be found in kaggle.com. This dataset with 303 cases and 14 variables does not contain any missing values. The *target* variable refers to the patient's heart disease presence, with 0 representing less chance of heart attack, and 1 represents more chance of heart attack. The variable *target*, evaluated and created by the four M.D, is the predicted attribute and the main object that we want to study. The other 13 variables from the cases of real patients are listed below with their corresponding meanings.

1. *age*: the age of the patient
2. *sex*: the biological gender of the patient (1 = male; 0 = female)
3. *cp*: the chest pain type (values 0,1,2,3);
   - Value 0: typical angina
   - Value 1: atypical angina
   - Value 2: non-anginal pain
   - Value 3: asymptomatic
4. *trestbps*: the resting blood pressure (in mm Hg on admission to the hospital)
5. *chol*: the serum cholestoral in mg/dl
6. *fbs*: whether the fasting blood sugar > 120 mg/dl (1 = TRUE; 0 = FALSE)
7. *restecg*: resting electrocardiographic results (values 0,1,2)
   - Value 0: normal
   - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. *thalach*: the maximum heart rate achieved
9. *exang*: exercise induced angina (1 = yes; 0 = no)
10. *oldpeak*: ST depression induced by exercise relative to rest
11. *slope*: the slope of the peak exercise ST segment (values 0,1,2)
    - Value 0: upsloping
    - Value 1: flat
    - Value 2: downsloping
12. *ca*: number of major vessels (0-3) colored by fluoroscopy
13. *thal*: thalassemia with 0 = normal; 1 = fixed defect; 2 = reversable defect

All of these 14 variables are used throughout this paper since each factor is distinct and meaningful. Table 1 below shows the data visualization of the first ten observations of the data (Naresh, 2020) used in this paper. The data summary table and some plots about this data are in the result section.

Table 1: Data Visualization

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Model

To find out the factors that affect the level of chance of heart attack, a multiple linear regression(MLR) model and a logistic regression model are generated in R Markdown using functions $lm()$ and $glm()$. The predictors are chosen based on the significance (at a 90% confidence interval) of two original models with all 13 variables as predictors in both methods. Only those with a p-value<0.1 are chosen. Here, p-value <0.1 means that the corresponding predictor is significant and will impact the response variable *target*. The summary table of the original model (Table 3) is in Appendix #1. Thus, based on Table 5, *sex*, *cp*, *trestbps*, *thalach*, *exang*, *oldpeak* and *ca* are selected as the predictors for MLR model. For logistic regression model, *sex*, *cp*, *exang*, *oldpeak* and *ca* are chosen based on Table 6 in Appendix #2. The model assumptions are checked after generating the models. Although there are some violations, we assume both models satisfy the model assumptions.

Model 1 (MLR)

$$y_{target} = \beta_0+\beta_1 X_{sex_1}+\beta_2 X_{cp_1}+\beta_3 X_{cp_2}+\beta_4 X_{cp_3}+\beta_5 X_{trestbps}+\beta_6 X_{thalach}+\beta_7 X_{exang_1}+\beta_8 X_{oldpeak}+\beta_9 X_{ca}+\epsilon_i$$

Since *sex*, *cp* and *exang* are categorical variables, dummy variables are used in the regression and the meaning of each dummy variable is listed below: - X_{sex_1} = sex is *male* - X_{cp_1} = the chest pain type is *atypical angina* - X_{cp_2} = the chest pain type is *nonanginal pain* - X_{cp_3} = the chest pain type is *asymptomatic* - X_{exang_1} = exercise induced angina

y_{target} (level of chance of heart attack) is the response and goal in this multiple linear regression. β_0 represents the intercept, it is simple a constant value. β_1 represents the average difference between sex is *male* ($X_{sex_1} = 1$) and sex is $female$ ($X_{sex_1} = 0$). Similary, β_2, β_3, and β_4 represent the corresponding values that the response changes when the chest pain type is *atypical angina*($X_{cp_1} = 1$), *nonanginal pain*($X_{cp_2} = 1$) and *asymptomatic*($X_{cp_3} = 1$). β_7 represents the average difference in the response between exercise induced angina and not exercise induced angina. β_5, β_6, β_8 and β_9 represents the average change in the response if the resting blood pressure, the maximum heart rate achieved, ST depression induced by exercise relative to rest, and number of major vessels colored by fluoroscopy increase by one unit respectively. After getting the model, we classify the $y_{target} \geq 0.5$ as more chance of heart attack and $y_{target} < 0.5$ as less chance of heart attack.

Model 2 (Logistic)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{sex_1} + \beta_2 X_{cp_1} + \beta_3 X_{cp_2} + \beta_4 X_{cp_3} + \beta_5 X_{exang_1} + \beta_6 X_{oldpeak} + \beta_7 X_{ca}$$

$log(\frac{p}{1-p})$ is the log odds, which is the odds of the level of chance of heart attack. Here p is the probability of the level of chance of heart attack. β_0 is the intercept term. β_1, β_2, β_3, β_4 and β_5 are the coefficients of dummy variables, representing the change in log odds if sex is male, the chest pain type is *atypical angina*, the chest pain type is *nonanginal pain*, the chest pain type is *asymptomatic*, and it is exercise-induced angina respectively. β_6 and β_7 represent the change in log odds for every one unit change increase in ST depression induced by exercise relative to the rest and number of major vessels colored by fluoroscopy.

Results

- Data

Table 2: Summary Table of Different Chest Pain Type

cp	prop	chance_of_heart_attack	avg_age	avg_trestbps	avg_chol	avg_thalach	avg_oldpeak
0	0.472	0.273	55.692	132.021	250.133	140.538	1.383
1	0.165	0.820	51.360	128.400	244.780	162.420	0.316
2	0.287	0.793	53.517	130.379	243.172	155.609	0.798
3	0.076	0.696	55.870	140.870	237.130	155.957	1.391

Table 2 shows the summary statistics of different chest pain type with the proportion of each chest pain type, average age, average resting blood pressure, average serum cholesterol, the average maximum heart rate achieved, and the average ST depression.

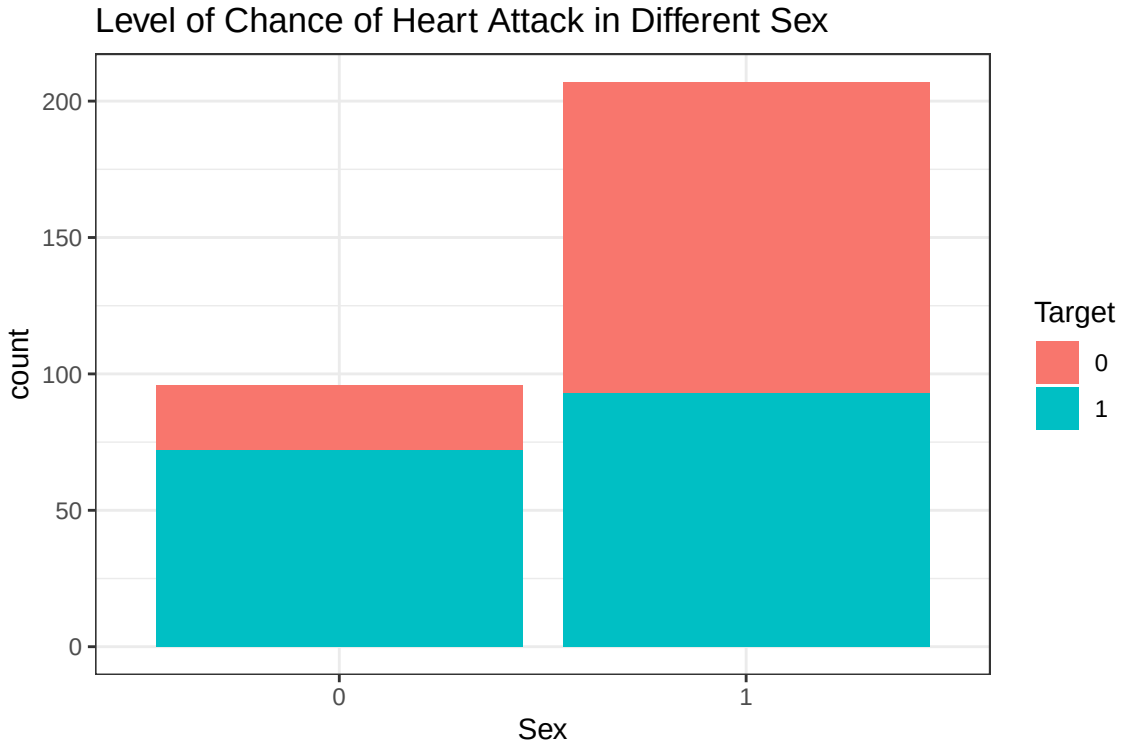

Figure 1: Level of Chance of Heart Attack in Different Sex

Figure 1 shows the different level of chance of heart attack in different sex. (*sex*: 0 = female, 1 = male; *target*: 0 (in red) = less chance of heart attack, 1 (in blue) = more chance of heart attack) The x-axis is sex and y-axis indicates the corresponding count of each group.

- Model

Table 3: Multiple Linear Regression Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	0.575	0.229	2.509	0.013
as.factor(sex)1	-0.205	0.046	-4.488	0.000
as.factor(cp)1	0.236	0.067	3.503	0.001
as.factor(cp)2	0.290	0.055	5.272	0.000
as.factor(cp)3	0.306	0.086	3.570	0.000
trestbps	-0.002	0.001	-1.848	0.066
thalach	0.003	0.001	3.018	0.003
as.factor(exang)1	-0.151	0.052	-2.885	0.004
oldpeak	-0.081	0.020	-3.987	0.000
ca	-0.104	0.021	-4.845	0.000

Table 3 is the summary table for MLR model with 7 predictors. The intercept, coefficients of the 7 predictors and corresponding p-value are listed. Based on this table, we can write out the following model equation.

$$\hat{y} = 0.58 - 0.21X_{sex_1} + 0.24X_{cp_1} + 0.29X_{cp_2} + 0.31X_{cp_3} + 0.002X_{trestbps} + 0.003X_{thalach} - 0.15X_{exang_1} - 0.08X_{oldpeak} - 0.104X_{ca}$$

Table 4: Logistic Regression Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	1.961	0.435	4.511	0.000
as.factor(sex)1	-1.412	0.389	-3.625	0.000
as.factor(cp)1	1.350	0.487	2.773	0.006
as.factor(cp)2	2.090	0.419	4.987	0.000
as.factor(cp)3	2.016	0.609	3.313	0.001
as.factor(exang)1	-1.222	0.372	-3.283	0.001
oldpeak	-0.806	0.181	-4.454	0.000
ca	-0.763	0.166	-4.595	0.000

Table 4 is the summary table for the logistic model with 5 predictors. The intercept, coefficients of the 5 predictors and corresponding p-value are listed. Based on this table, we can write out the following model equation.

$$log(\frac{\hat{p}}{1-\hat{p}}) = 1.96 - 1.41X_{sex_1} + 1.35X_{cp_1} + 2.09X_{cp_2} + 2.02X_{cp_3} - 1.22X_{exang_1} - 0.81X_{oldpeak} - 0.76X_{ca}$$

Discussion

Data

- From Table 2, we can see that chest pain type 0, typical angina, has a large proportion of the data and less chance of heart attack, and the average maximum heart rate achieved is significantly lower than the other 3 types of chest pain. However, for chest pain type 1, atypical angina, it has the highest heart attack chance among these four different types of chest pain with the lowest average *oldpeak* value 0.316 and a relatively higher average maximum heart rate (162.43).
- Figure 1 shows that there are approximately 96 females and 207 females, where the first bar represents females, and the second bar represents females. Figure 1 indicates that about 75% of females have a high risk of heart attack, and about 45% of males have a high risk of a heart attack. Thus, in our sample data, females have a higher chance of heart attack than males.

Model

- We can see that based on the result of the MLR model, factors of sex, chest pain type and the whether it is exercise induced angina or not affect the level of chance of heart attack the most. If the sex of patient is male, then the level of chance decreases by 0.21. If one's chest pain type is *atypical angina*, the level of chance increases by 0.24. If one's chest pain type is *nonanginal pain*, the level of chance increases by 0.29 and if the chest pain type is *asymptomatic*, the level of chance increases by 0.31 which increases by the largest value among the other chest pain types. When it is exercise induced angina, the level of chance of heart attack decreases by 0.15. Also, when the resting blood pressure, the maximum heart rate achieved, ST depression induced by exercise relative to rest, and number of major vessels colored by fluoroscopy increase by one unit, the average changes in the response are 0.002, 0.003, -0.08 and -0.104, respectively.
- The p-values in Table 4 indicates that all predictors of the logistic regression are significant. From the logistic model equation in the result section, we can find the probability of the level of chance of heart

attack of a patient by calculating $\hat{p}$. Based on the logistic regression model, if sex is male, the log odds decreases by 1.41. If the chest pain type is *atypical angina*, *nonanginal pain*, and *asymptomatic*, the log odds increases by 1.35, 2.09 and 2.02 respectively. When it is the case of exercise induced angina, the log odds decrease by 1.22. For every one unit increase in ST depression induced by exercise relative to rest and number of major vessels colored by fluoroscopy, the log odds decrease by 0.81 and 0.76, respectively. In this case, all predictors (*sex*, *cp*, *exang*, *oldpeak* and *ca*) have a significant impact on the level of chance of heart attack.

- The AIC for the MLR model is 252.7066, while the AIC score for the logistic regression model is 254.3197. Therefore, if the AIC criterion is used in model selection, the MLR model has a relatively better fit than the logistic regression model in this case. The difference in the AIC score is quite small which means that in practice, these two model have similar goofness of fit. However, the R^2 of the MLR model is relatively low which means the predictors in the MLR model explain a relatively small proportion of variation of the target. Since target is a binary variable and the all predictors in the logistic regression model are significant. The logistic regression model is chosen to be the final model to predict the level of heart attack of other patients. Sex, chest pain types, whether exercise induced angina, ST depression induced by exercise relative to rest and number of major vessels (0-3) colored by fluoroscopy are critical factors that accounts for the level of heart attack.

Weaknesses and Future Work

- The size of the dataset is relatively small, which means the data may not be fully representative of the population. When fitting models, some assumptions are violated which may affect the model results.
- Generating another statisticcal model using a different method on this analysis is a good idea. On the other hand, the casual inference can be discussed through propensity score matching in this study since the data is observational study instead of experimental study. Using this approachm, the causal inference between the essential factors and the level of heart attack can be determined.

References

- bhat, Naresh. Health Care: Data Set on Heart Attack Possibility. 25 June 2020, www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility.
- UCI Machine Learning Repository: Heart Disease Data Set, archive.ics.uci.edu/ml/datasets/heart Disease.1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.
- Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra
- David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom

Appendix

1.

Table 5: Original Model Summary (MLR)

term	estimate	std.error	statistic	p.value
(Intercept)	0.577	0.379	1.523	0.129
age	-0.001	0.003	-0.193	0.847
as.factor(sex)1	-0.157	0.049	-3.179	0.002
as.factor(cp)1	0.178	0.066	2.699	0.007
as.factor(cp)2	0.243	0.054	4.485	0.000
as.factor(cp)3	0.271	0.084	3.234	0.001
trestbps	-0.002	0.001	-1.258	0.209
chol	0.000	0.000	-0.855	0.393
as.factor(fbs)1	0.023	0.059	0.385	0.701
as.factor(restecg)1	0.059	0.042	1.399	0.163
as.factor(restecg)2	-0.045	0.181	-0.248	0.804
thalach	0.002	0.001	1.725	0.086
as.factor(exang)1	-0.112	0.051	-2.190	0.029
oldpeak	-0.054	0.023	-2.316	0.021
as.factor(slope)1	-0.103	0.088	-1.161	0.246
as.factor(slope)2	0.017	0.097	0.174	0.862
ca	-0.093	0.022	-4.330	0.000
as.factor(thal)1	0.156	0.262	0.596	0.552
as.factor(thal)2	0.243	0.250	0.975	0.330
as.factor(thal)3	0.018	0.251	0.073	0.942

2.

Table 6: Original Model Summary (MLR)

term	estimate	std.error	statistic	p.value
(Intercept)	1.101	3.365	0.327	0.744
age	-0.001	0.024	-0.024	0.981
as.factor(sex)1	-1.515	0.521	-2.906	0.004
as.factor(cp)1	0.983	0.564	1.743	0.081
as.factor(cp)2	1.945	0.477	4.076	0.000
as.factor(cp)3	2.016	0.651	3.098	0.002
trestbps	-0.017	0.011	-1.596	0.111
chol	-0.004	0.004	-1.114	0.265
as.factor(fbs)1	0.176	0.566	0.312	0.755
as.factor(restecg)1	0.570	0.375	1.523	0.128
as.factor(restecg)2	-0.277	2.267	-0.122	0.903
thalach	0.017	0.011	1.596	0.111
as.factor(exang)1	-0.763	0.426	-1.791	0.073
oldpeak	-0.489	0.226	-2.167	0.030
as.factor(slope)1	-0.720	0.863	-0.834	0.404
as.factor(slope)2	0.202	0.938	0.215	0.830
ca	-0.833	0.204	-4.078	0.000
as.factor(thal)1	1.815	2.379	0.763	0.446
as.factor(thal)2	1.853	2.290	0.809	0.418
as.factor(thal)3	0.473	2.301	0.206	0.837