Heart Disease Estimation

Consulting Report for the GP

Ningkai Zheng

Words Count: 1836

Table of Contents

1 Introduction

As more and more people go to the GP for heart-related problems, an efficient preliminary heart disease diagnosis appears to be necessary. This report is focused on building a simplified estimation model, which aims to diagnose whether a future patient has a high probability of suffering coronary heart disease or not, with relatively high accuracy and high efficiency in practice. The method we used in Excel is *Logistic Regression* and *Feature Selection*. Logistic regression helps us find the impact of each factor on the diagnosis result, while feature selection contributes to filtering out unrepresentative factors.

2 Data Description

The sample data was recently collected by cardiology doctors at the GP in one month, containing 12 dimensions that can be used to estimate the probability of having heart disease. The 12 dimensions include age, gender, chest pain type, resting blood pressure and other factors of the professional aspect (Figure 1, Appendix 1). Note that some factors' values are binary data. There are 502 patients recorded in the data, about half of whom are diagnosed with heart disease. We have applied the appropriate technical method to ensure that there is no duplicate patient ID.

3 Theoretical Basis and Assumptions

3.1 Logistic Regression

Logistic regression is suitable for determining yes or no questions, and it helps us find the model that most fit the data trend. In the algorithm, it aims to maximise the total *log-likelihood* (LL). And LL is calculated based on the coefficients of the factors, so we can get a certain list of coefficients that fit the data most while maximising LL. The formulas of the model given by logistic regression are as follows:

$$Probalitity\ of\ something\ happening = \frac{1}{1+e^{-g}}$$

$$g = b_0 + b_1 \times Factor_1 + b_2 \times Fctor_2 + \cdots + b_i \times Factor_i$$

This is also known as one of the logistic regression's assumptions. In these formulas, *e* is the natural constant with a value of about 2.72, *g* is an intermediate variable, and b_i is the coefficient of factor i. Mathematically, the *Probability of something happening* will increase as *g* increases.

Logistic regression's another assumption is that the data follows a *Bernoulli distribution* (i.e. yes or no). Estimating whether a patient has heart disease or not conforms to this assumption.

Other less essential assumptions include *Independence of observations*, *No inappropriately high collinearity between predictors*, *Fully represented (no sparse) data matrix*, *Perfect measurement*, *Accurate model specification*, *Data free from inappropriately influential cases* (Osborne, 2015, p.86), which we are not going to illustrate here.

3.2 Feature Selection

Feature selection helps us find the most relevant factors so that we can simplify the model. There are three types of feature selection algorithm, and we used the technique called *wrapper* (Chen et al., 2020, p.2). In Excel, the algorithm aims to maximise the total correct number of estimations by changing the set of selected factors. The maximisation process will evaluate all the feature subsets and eventually generate the optimal solution.

4 Data Analyses

We select 50 patients randomly from the original data to train and build the model, while the rest are used to test the model's accuracy.

4.1 Logistic Regression

4.1.1 Modelling

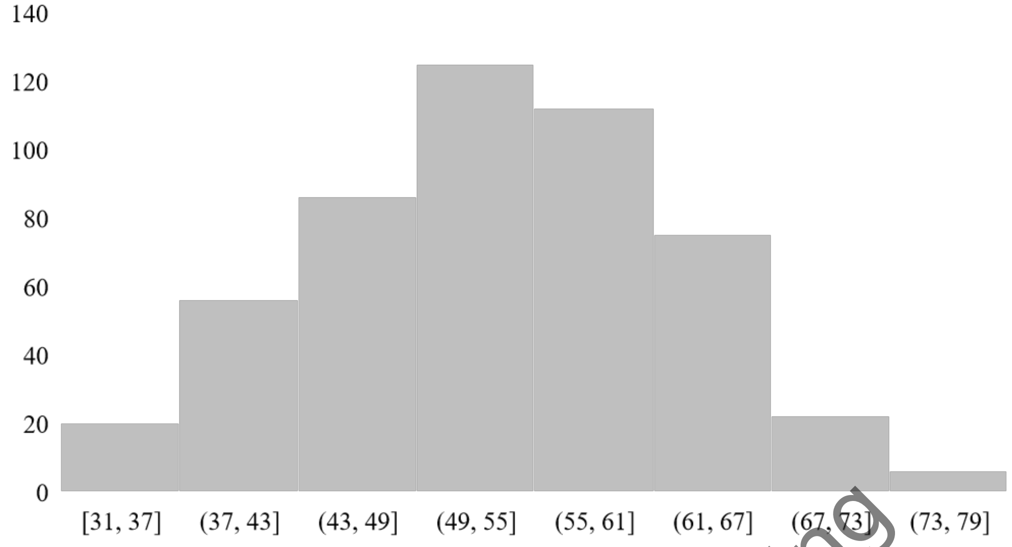
Firstly, we use logistic regression with all 12 factors being considered. We obtain a set of coefficients and use them to construct the model below (Model 1).

$$Probalitity\ of\ having\ heart\ disease = \frac{1}{1+e^{-g}}$$

$$\begin{aligned} g \;=\; 3.7193\; &+ (-0.0064 \times Age) \\ &+ (-6.4489 \times Gender) \\ &+ (-3.7263 \times Chest\ pain\ type) \\ &+ (5.0000 \times Resting\ blood\ pressure) \\ &+ (0.6900 \times Cholesterol) \\ &+ (-1.8963 \times Resting\ blood\ sugar) \\ &+ (3.5168 \times Resting\ ECG\ result_{LVH}) \\ &+ (1.8876 \times Resting\ ECG\ result_{ST}) \\ &+ (-6.3630 \times Maximum\ heart\ rate) \\ &+ (-0.3520 \times Exercise\ induced\ angina) \\ &+ (0.6111 \times Previous\ peak) \\ &+ (-3.2757 \times Slope) \end{aligned}$$

Now we can use this model to do the estimation. The model shows that factors with negative coefficients will lead to a low probability of having heart diseases, such as age and gender. More precisely, this means the older and females are less likely to have heart diseases. Obviously, "the older, the less likely to get heart disease" is counter-intuitive (Janna et al., 2020, para.11). We speculate that it is because the collected data lacks young and old patients (Figure 1), and this centralised tendency is also evident in the 50 patients we selected (Figure 2, Appendix 1). Even more of the values at both ends get thrown out in the train data. Hence, middle-age causes this abnormal trend. Fortunately, the absolute value of the coefficient of age is quite small compared to other factors, which means the data distortion is not significant.

Figure 1 Histogram of Age Distribution in the Collected Data



How to explain the missing data of both ends? The data is collected from patients visiting the GP for heart-related issues. However, the younger rarely suffer from heart disease and the older with heart disease are likely to die in their middle age. Hence, people from these two groups hardly come to the GP for diagnosing heart disease, so this is why we lack the data from these two generations.

Factors like resting blood pressure and cholesterol, which have positive coefficients, correlate positively with the probability of having heart disease. Except for the age and gender, the effectiveness of other factors' coefficients should be further discussed with cardiology doctors from the GP.

4.1.2 Evaluation

It is essential to test the model using the 452 patients to evaluate its accuracy. We estimate that the patient has heart disease if his/her probability of having heart disease is greater than 0.5, then compare the estimated outcomes to the actual diagnosis results. Finally, we get a correct rate of 78.10%, which is relatively high.

However, we can not claim that this accurate model is the optimal solution for estimating in practice. The reason is that the GP need to spend much time collecting data of the 12 dimensions, which is apparently not an efficient approach. To address this issue, we need to see if it is possible to reduce the factors in the model while maintaining a high correct rate.

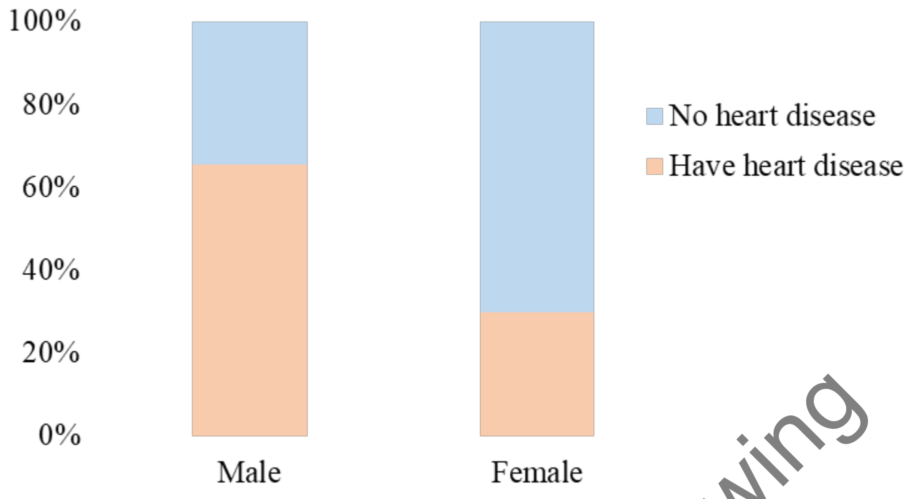
4.2 Feature Selection to Aid Logistic Regression

4.2.1 Simplification

In the first trial, we set the least number of selected factors to be 1, and Excel shows that only *gender* should be considered in this way, with the maximum total correct number of estimations being 35. We conclude from the original data that males indeed have a much more significant proportion of having heart disease than females (Figure

2), so it is rational to reach this result, although it appears to be arbitrary to estimate whether a patient has heart disease or not by only considering gender.

Figure 2 Rates of Heart Disease by Gender



Furthermore, we apply the logistic regression again and construct the new model below (Model 2).

$$Probability\ of\ having\ heart\ disease = \frac{1}{1+e^{-g}}$$
$$g = 0.7802 + (-2.1665 \times Gender)$$

The negative coefficient of gender also indicates that females are less likely to have heart disease. We use the 452 patients to test the model and get a correct rate of 65.93%, about 12% less than the previous model. However, we only require gender information to complete the estimation, which is much quicker and easier to obtain.

4.2.2 Accuracy Improvement

Can we construct a model with an accuracy similar to Model 1 by adding a few factors back to Model 2? We set the least number of selected factors to equal or greater than two and rerun the feature selection programme. This time Excel indicates that we need *gender*, *chest pain type* and *slope* in the model. Prediction made through the selected factors still reaches the total correct number of 35, suggesting that this feature selection is a multi-solution problem. The model with these three factors (Model 3) built by logistic regression is:

$$Probability\ of\ having\ heart\ disease = \frac{1}{1+e^{-g}}$$
$$\begin{aligned} g = 3.4430 &+ (-3.5523 \times Gender) \\ &+ (-3.2924 \times Chest\ pain\ type) \\ &+ (-3.3166 \times Slope) \end{aligned}$$

When predicting 452 patients' heart disease situations, this model's correct rate turns out to be 76.99%, which is quite close to the first model's accuracy. Compared to the second model, we achieve 10% progress in the estimation's correct rate with only two factors being further considered. This model is entirely accurate and efficient, so we do not need to put in the effort to add more factors to improve it.

As it is shown that this feature selection has multiple optimal solutions, we still need to find if there is an optimal solution when selecting two factors. Nevertheless, the feature selection programme shows the maximised total correct number is 25 this time, which means the set of factors selected is inferior to those mentioned above—the hope of making the model more efficient dashes.

5 Limitations

5.1 Data Imperfection

The data we use to build the model is collected in one month in the local GP, so we cannot measure the seasonal and environmental influences. Therefore, the estimation model may not be suitable for the whole year and GP in other areas. This problem can be solved when we make the best use of big data in the national healthcare system (Karalee, 2014).

As aforementioned, the data is collected from patients who visited the GP, so it lacks records from the younger and older. Thus, Model 1 may be misleading to the public. We should clarify that this model does not capture the effect of age at all. To build a more comprehensive model, we need massive data from different age groups.

5.2 Technical Defects

Since we do not apply advanced machine learning techniques in the process of building the model, the model's accuracy is limited by the 50 patients we randomly selected from the original data. By applying more advanced machine learning techniques, the model's accuracy will be increased based on the 502 patients and further improved as the data increases in the future.

6 Suggestions

6.1 General Suggestions

Based on Model 2, we recognise that males are much more likely to get heart disease. Therefore, the local GP is supposed to encourage males to have regular physical check-ups for heart-related issues. This is not to say that females do not need to have physical check-ups. Females should also keep this good habit, but generally with a lower frequency.

6.2 Estimation Model

For the patients who come to the GP, Model 3 should be considered. We build the estimation table in the Excel spreadsheet (Figure 3, Appendix 1), in which you will get the predicted heart disease diagnosis once you enter the *gender*, *chest pain type* and *slope* information of a patient using the dropdown list. Patients with a probability of having heart disease greater than 0.5 are classified in the *High Probability* group and highlighted. Plus, no duplicate values can be entered in the *Patient ID* column.

7 Conclusion

In conclusion, for the sake of public health in the local area, the GP should encourage residents to have regular physical check-ups related to heart disease, especially male residents, which will inevitably lead to a further increase in the number of patients visiting the GP. Thus, we advise the local GP to use the estimation model we provide in preliminary heart disease diagnosis, through which the GP will refer patients with a high probability of having heart disease to the general hospital in an efficient way. However, we still need to collect more comprehensive data and apply advanced machine learning techniques to pursue a better estimation model.

Appendix 1

Figure 1:

Factor	Variable Type	Notes
Age	Numeric	
Gender	Binary: 0 for male, 1 for female	
Chest pain type	Binary: 0 for regular, 1 for irregular	
Resting blood pressure	Numeric	mmHg/100
Cholesterol	Numeric	mg/dl/100
Resting blood sugar	Numeric	
Resting ECG result-LVH	Binary: 1 for lower heart rate; 0 otherwise	
Resting ECG result-ST	Binary: 1 for strained; 0 otherwise	
Maximum heart rate	Numeric	Rate/100
Exercise-induced angina	Binary: 0 for no, 1 for yes	
Previous peak	Numeric	Peak exercise level in standard deviations
Slope	Binary: 0 for flat, 1 for up	Slope of the heart rate with exercise

Figure 2:

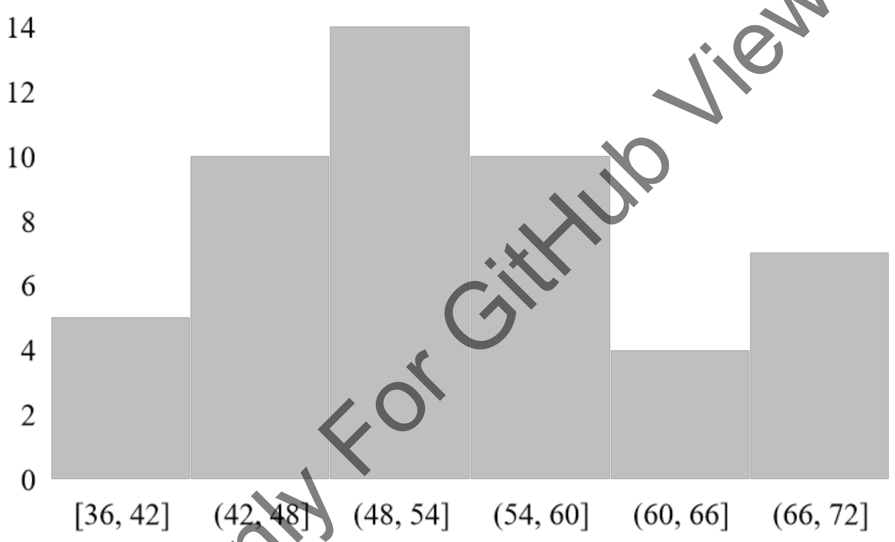


Figure 3:

Patient ID	Gender	ChestPainType	ST_Slope	HeartDisease Predicted
1	Male	Regular	Flat	High Probability
2	Female	Regular	Flat	Low Probability
3	Female	Regular	Up	Low Probability
4	Male	Regular	Up	High Probability
5	Female	Irregular	Flat	Low Probability
6	Male	Irregular	Up	Low Probability
7				

Appendix 2

Bibliography

Osborne, J.W., 2015. Best practices in logistic regression, Los Angeles: SAGE.

Chen, C.W. et al., 2020. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert systems, 37(5), pp.e12553-n/a.

Jaana, R., Jonathan, W., Sven, S., Katherine, L., Shubham, S., Martin, D., Penelope, D., Kristin, A.R., Matthias, E., Matt, W. and Aditi, R., 2020. *Prioritizing health: A prescription for prosperity* [Online]. McKinsey Global Institute. Available from: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/prioritizing-health-a-prescription-for-prosperity [Accessed 7 November 2021]

Karalee, C., Stefan, L., John, L., Neil, S. and Anna, V., 2014. *Making Big Data Work: Health Care Payers and Providers* [Online]. Boston Consulting Group. Available from: https://www.bcg.com/publications/2014/making-big-data-work-health-care-payers [Accessed 7 November 2021]