Assignment 2 - Data Analysis 2 and Coding with R

David Utassy

01/01/2021

Abstract

In this assignment, I have fitted a logit model on the dataset which measures 1000 patients' attributes and whether they have heart disease or not. That model uncovers some interesting patterns between the variables and might predict the probability of a new patient with the given variables. This model can not be generalized to the whole population because the dataset is clearly not representative, but it might be used as a compass between the variables.

Introduction

In this project I address the question, how to predict from some measurable variables whether someone has heart disease. This is an important question as according to the World Health Organization, cardiovascular diseases are the leading cause of death globally. Of course, this topic requires broad knowledge on the topic. To get familiar with the problem I recommend this article, which describes the theory in an understandable way.

Data

This dataset is hosted on Kaggle Heart Disease Dataset, it is from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long each. The original data source is the UCI Machine Learning Repository, which contains 76 attributes, but all published experiments refer to using a subset of 14 of them, so as mine.

In the dataset, every row is a person with the given parameters in each column. The number of the observations is 1025 originally, which will decrease to 1000 after omitting some with missing values. This dataset was pretty clean already, only a few modifications were needed to make the analysis more convenient.

The most challenging part of this project is to understand somehow the variables I have. In order to get the reader somewhat familiar with them, I introduce them in a few sentences in an easy to understand form. For further details, I recommend the previously mentioned online sources.

1. Target: (binary) whether the patient has heart disease or not (originally 0: yes, 1: no) but I made a modification, therefore 0: no, 1: yes
2. Age: (quantitative) the age of the patient
3. Gender: (binary) The gender of the patient (0:female, 1: male)
4. Chest pain type (qualitative)
    - Value 0: asymptomatic (no symptoms)
    - Value 1: atypical angina

- Value 2: pain without relation to angina
- Value 3: typical angina

5. Resting blood pressure in mm Hg (qualitative)
6. Serum cholesterol in mg/dl: The level of cholesterol in the blood
7. Fasting blood sugar > 120 mg/dl: (binary) Whether the level of blood sugar is under or above the threshold (0: under, 1:above)
8. Resting electrocardiographic results (qualitative)
    - Value 0: probable left ventricular hypertrophy
    - Value 1: normal
    - Value 2: abnormalities in the T wave or ST segment
9. Maximum heart rate achieved (qualitative)
10. Exercise-induced angina: (binary) whether pain occurred during exercise (0: no, 1: yes)
11. Oldpeak: (quantitative) Decrease of the ST segment during exercise according to the same one on rest. (0 is optimal)
12. The slope of the peak exercise ST segment: (qualitative) the slope of a specified line segment on the ECG
    - Value 0: descending
    - Value 1: flat
    - Value 2: ascending
13. Number of major vessels (0-3) colored by fluoroscopy: (qualitative, ordered) Number of narrow vessels, ideally should be 0.
14. Results of the blood flow observed via the radioactive dye
    - Value 1: fixed defect (no blood flow in some part of the heart)
    - Value 2: normal blood flow
    - Value 3: reversible defect (a blood flow is observed but it is not normal)

variable	mean	median	std	min	max	skew
Heart disease (0: no, 1: yes)	0.49	0.0	0.50	0	1.0	0.03
Age	54.61	56.0	9.04	29	77.0	-0.26
Gender (0:female, 1:male)	0.69	1.0	0.46	0	1.0	-0.83
Resting blood pressure (mm Hg)	131.59	130.0	17.71	94	200.0	0.74
Cholesterol level (mg/dl)	247.00	240.5	51.70	126	564.0	1.06
Fasting blood sugar > 120 mg/dl (0: under, 1:above)	0.15	0.0	0.35	0	1.0	2.01
Maxium heart rate during the stress test	148.98	152.0	23.08	71	202.0	-0.50
Exercise-induced angina (0: no, 1: yes)	0.34	0.0	0.47	0	1.0	0.68
Oldpeek	1.09	0.8	1.18	0	6.2	1.18
Number of main blood vessels	0.70	0.0	0.94	0	3.0	1.12

In order to get to know my variables, I have made some basics histograms. From those, it was visible that some variables might need a log transformation. In the appendix, we can see the result of some possible transformations, but in the end, I decided to stay with the original scales as the modifications probably would not pay off.

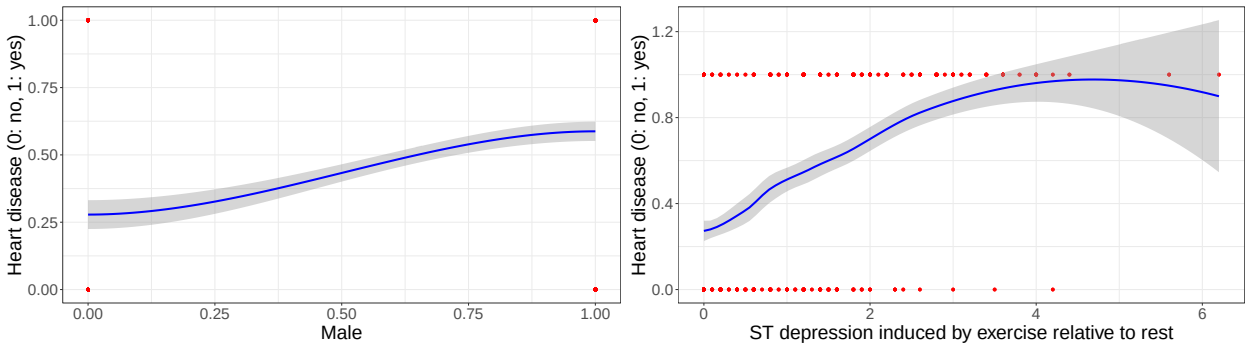
The table above contains the summary statistics of the variables I used in the analysis. To highlight some patterns in our data, the mean of the presence of heart disease is almost 50%, which could be good for prediction, however, 70% of the observations are male, which is not representing the population well. The age variable also shows that the sample is not really representing a whole population, but probably a subset of people who has to do something with cardiovascular problems. The other variables are more like medical terms, therefore I would not go into detailed preconceptions.

Data quality issues

However this dataset is clean, quality is a different thing. This dataset is from 1988 from different parts of the world. We do not know much about the collection of this data, but it was a medical project, therefore I would say, that most patients had some connection with cardiovascular problems. This argument would explain the ratio of men and the age distribution of the dataset as well. The age and the sample of this dataset mean that the external validity is not perfect of this dataset if we want to generalize to the whole population.

Model

In order to create a proper model, it worth to visualize the association between variables. In the appendix, I plotted all the meaningful associations.

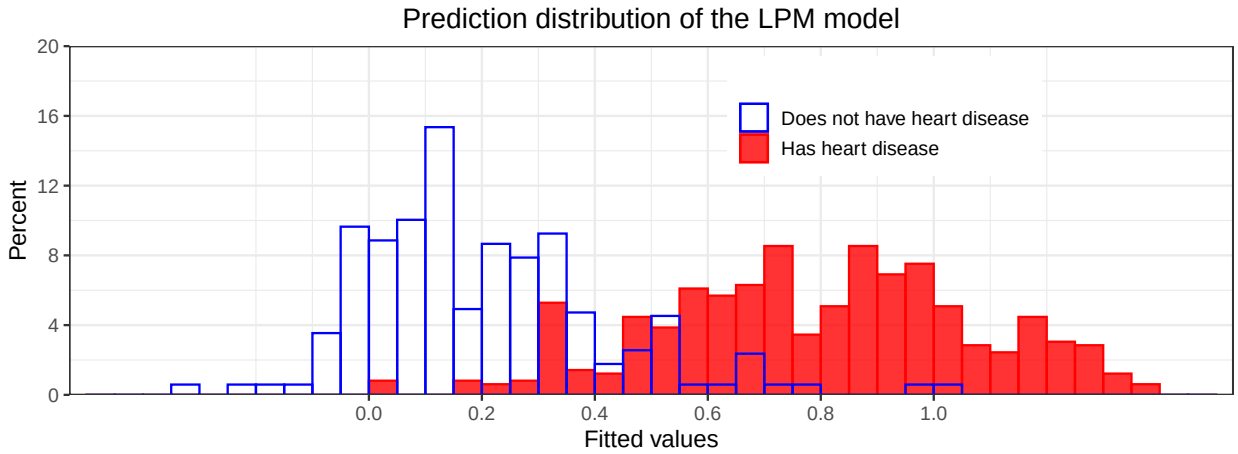


On the plots above we can see two plots from the ones mentioned before. It can be easily seen that men have heart disease more likely than women and that ST depression induced by exercise has a positive correlation with heart disease as well. These are just two variables out of 13 but in the appendix, the association of other variables with the presence of heart disease can be observed. It turns out, that almost all of the provided variables have some correlation with the presence of heart disease. Therefore I will try to use all of them in an LPM, logit, and probit model.

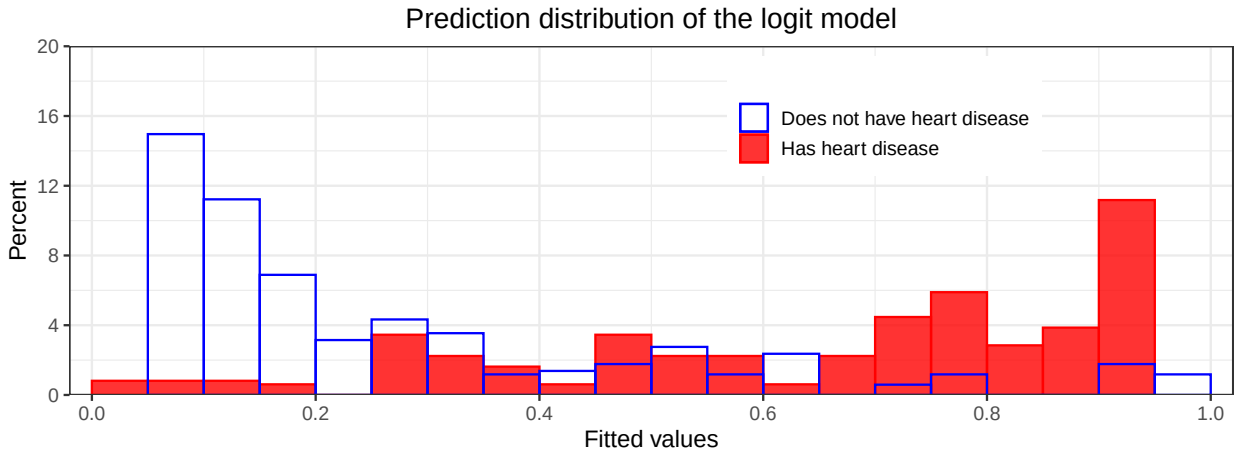
In the experimenting phase, I have created different models to capture the data in my sample. In the appendix, one can see the summary table of three different LPM models. At the bottom of that table, we can compare the three models with the R^2 attribute. According to that, it is clear, that we should go on with the model that has all the variables.

In order to handle nonlinearities, I used piecewise linear splines according to the already mentioned plots. Furthermore to handle quantitative variables I used them as a factor (which creates n-1 dummy variables in the model).

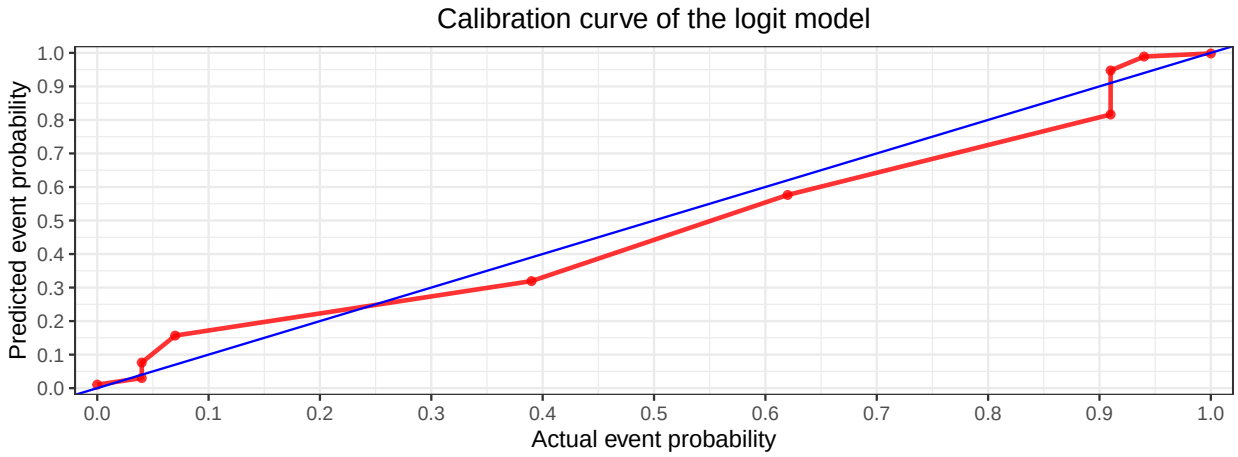
In order to get closer to possible prediction, we want our model to give the probability between 0 and 1 on the presence of heart disease in a patient. In our case with the LPM model, we are in trouble as it gives predictions under zero and above one as well.



The solution is to use logit or probit model instead of LPM. This will predict only probabilities between 0 and 1. To be able to interpret the logit and probit models we should calculate the marginal differences as well. In the appendix, we can see the comparison of the LPM the Logit, and he Probit model. At the bottom of the table, we can also compare them in their "goodness". According to the AIC and BIC I have chosen the logit model.



As it is visible on the prediction distribution of the logit model, it is kind of OK. All the values are in the range of [0, 1] and there is a pattern that I was looking for. The logit and the probit model are nonlinear by definition, therefore they produce a nonlinear prediction curve as it is plotted in the appendix. This is the price of forcing predictions into the range [0,1].



From the calibration curve of the logit model above we can say, that the model is well-calibrated and the calculated bias of the model is almost zero, thus the model is unbiased.

To analyse the lowest and highest percent probabilities I used the LPM model as from the prediction distribution of the logit model we can see, that there are some errors in the lowest and highest percent predictions, which would affect my results.

Top 1 percent (ones with the likelihood of heart disease)

statistics	Age	Male	Chol. lev.	Rest. blood pres.	Max. h. rate	Ex. Ang.	Num ves.
mean	59.10	1	259.30	131.90	123.80	1	2.40
median	60.00	1	267.00	128.00	131.00	1	2.00
sd	3.28	0	31.35	10.81	18.25	0	0.52

Bottom 1 percent (ones with a probably healthy cardio vascular system)

statistics	Age	Male	Chol. lev.	Rest. blood pres.	Max. h. rate	Ex. Ang.	Num ves.
mean	46.90	0.10	232.60	113.00	174.70	0	0
median	46.00	0.00	204.00	105.00	172.00	0	0
sd	6.38	0.32	49.32	18.11	5.06	0	0

In the tables above we can observe the summary statistics (the most interpretable variables) of the lowest and highest percent probabilities.

We can interpret, that the most endangered ones are men with age around 59, with cholesterol level around 260, with resting blood pressure around 130, with maximum heart rate 125 during exercise, with angina occurred during exercise and with 2-3 narrow vessels in their heart.

The most healthy ones are women with age around 47, with lower cholesterol level and blood pressure, but interestingly higher maximum heart rate during exercise. To explain this strange behavior I have made some research and some publications suggested, that maximum heart rate gets lower with age which means that the worst group has a lower maximum heart rate on average which explains my finding.

Robustness check

In order to check the robustness of the model, I separated my data set into training and test sample (4/5 and 1/5 ratio). After training the model on the training sample we can compare the coefficients of the model

(comparison table in the appendix). We can conclude that there are some slight differences, but the main pattern is the same. Furthermore, the prediction distribution of this model (in the appendix) also shows a similar pattern to the original one. Therefore I would say, that my logit model is robust within this dataset, however, we should highlight that this is not equal with external validity at all.

Generalization

In the data quality part, it has been already stated, that the external validity of this dataset is not very good if we want to generalize our findings to the population as it is not representative in many variables. I already mentioned age and gender, moreover the average and median cholesterol level in the data sample is also way above the optimal value which further strengthens the argument, that probably most of the patients in this dataset have health problems. On the other hand, this model can be generalized for people who have some symptoms and reach out to medical organizations. In these cases, this model can give advice on the seriousness of the problem.

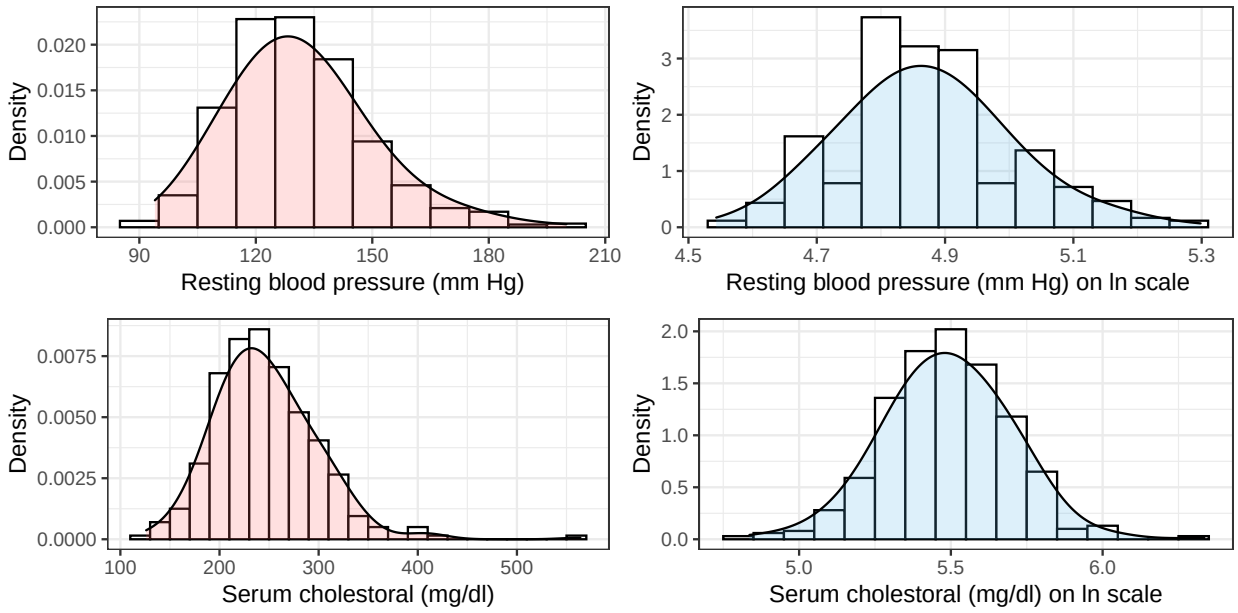
Summary

As a conclusion, we can interpret the most interesting variables in the final model (we can state the following statements on a 95% confidence interval or higher).

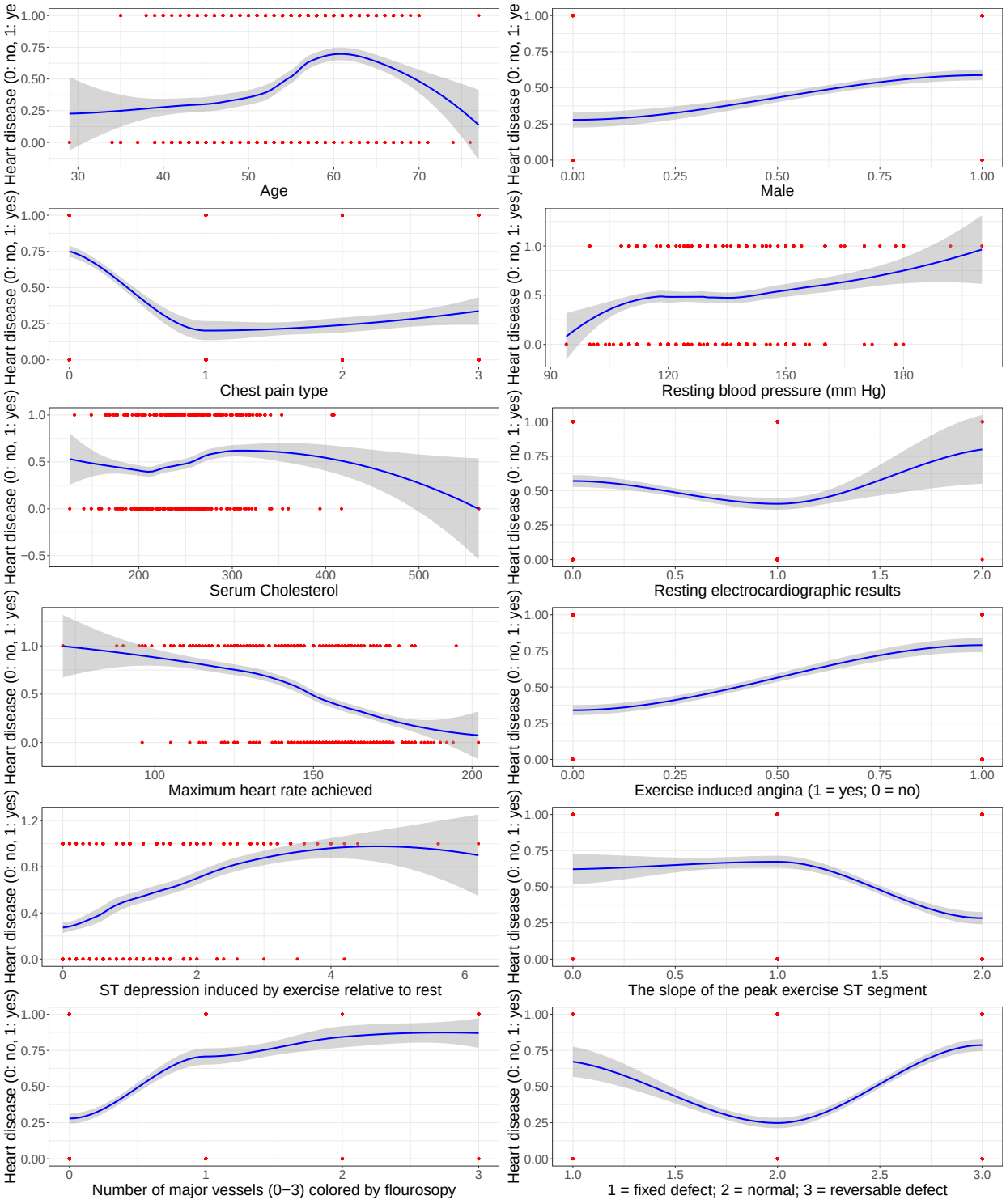
- If the patient is male, he has 17% more probability on average to have heart disease than women.
- Between the age of 52 and 61 if the patient is older by one year, on average s/he has 2.3% more probability of having a heart disease.
- Below 115 (Hg mm), if the patient has one unit higher resting blood pressure, on average s/he has 1% more probability of having a heart disease.
- If the patient has exercise-induced angina s/he has 9% more probability of having heart disease.
- Under 4 pieces, if the patient has one more narrow vessel, s/he has 14% more probability of having a heart disease.

Some of these variables are just symptoms, some of them are measurable medical variables, and some of them can be both. For this reason, in my opinion, this model is mostly good to uncover the significance of each variable in connection with the presence of heart disease, however, it can make some suggestions to diagnose symptoms in order to identify life-threatening problems faster. On the other hand, I would like to highlight that with models like this, we should always be really careful not to make any false-negative decisions!

Appendix



Possible variable transformations


Heart disease (0: no, 1: yes)
Age
Male
Chest pain type
Resting blood pressure (mm Hg)
Serum Cholesterol
Resting electrocardiographic results
Maximum heart rate achieved
Exercise induced angina (1 = yes; 0 = no)
ST depression induced by exercise relative to rest
The slope of the peak exercise ST segment
Number of major vessels (0−3) colored by flourosopy
1 = fixed defect; 2 = normal; 3 = reversable defect

	LPM model 1	LPM model 2	LPM model 3
Intercept	0.28^{***}	-0.18	-0.29
	(0.03)	(0.19)	(0.56)
Gender	0.31^{***}	0.33^{***}	0.17^{***}
	(0.03)	(0.03)	(0.03)
Age below 52		0.01	-0.01^{**}
		(0.00)	(0.00)
Age between 52 and 61		0.05^{***}	0.02^{***}
		(0.01)	(0.00)
Age above 61		-0.03^{***}	-0.03^{***}
		(0.01)	(0.01)
Resting blood pressure under 115			0.01^{**}
			(0.00)
Resting blood pressure between 115 and 130			-0.01^{*}
			(0.00)
Resting blood pressure above 130			0.00^{*}
			(0.00)
Serum cholesterol under 210			-0.00
			(0.00)
Serum cholesterol between 210 and 270			0.00^{*}
			(0.00)
Serum cholesterol between 270 and 430			-0.00
			(0.00)
Serum cholesterol above 430			-0.00
			(0.00)
Fasting blood sugar above 120 mg/dl			-0.04
			(0.03)
Maximum heart rate achieved below 140			-0.00^{*}
			(0.00)
Maximum heart rate achieved above 140			-0.00
			(0.00)
Exercise-induced angina occured			0.12^{***}
			(0.03)
Oldpeak below 4			0.03^{**}
			(0.01)
Oldpeak above 4			-0.03
			(0.07)
Number of major vessels			0.14^{***}
			(0.01)
R^2	0.08	0.19	0.58
Adj. R^2	0.08	0.19	0.57
Num. obs.	1000	1000	1000

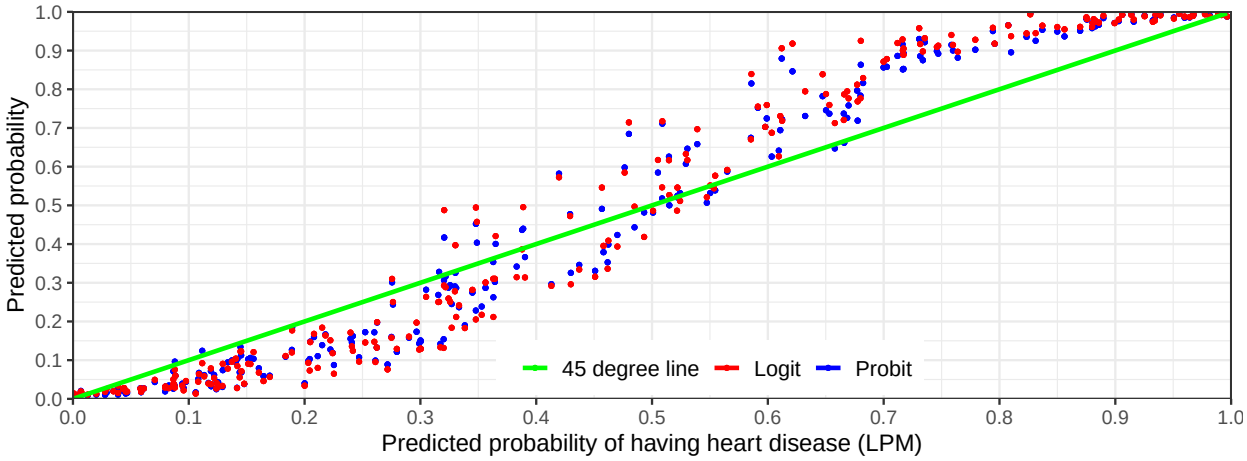
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Comparing LPM models with different number of variables (quantitative variables are hidden from the third model to fit onto one page but I am controlling the model on them as well)

	LPM	Logit	Logit margins	Probit	Probit margins
Intercept	−0.29	−9.93		−6.08	
	(0.56)	(6.24)		(3.32)	
Gender	0.17***	1.87***	0.17***	1.02***	0.18***
	(0.03)	(0.34)	(0.03)	(0.17)	(0.03)
Age below 52	−0.01**	−0.10**	−0.01**	−0.04*	−0.01***
	(0.00)	(0.04)	(0.00)	(0.02)	(0.00)
Age between 52 and 61	0.02***	0.23***	0.02***	0.11***	0.02***
	(0.00)	(0.05)	(0.01)	(0.03)	(0.00)
Age above 61	−0.03***	−0.32***	−0.03***	−0.17***	−0.03***
	(0.01)	(0.06)	(0.01)	(0.03)	(0.00)
Resting blood pressure under 115	0.01**	0.13**	0.01***	0.07**	0.01***
	(0.00)	(0.05)	(0.00)	(0.02)	(0.00)
Resting blood pressure between 115 and 130	−0.01*	−0.05	−0.00*	−0.04*	−0.01**
	(0.00)	(0.03)	(0.00)	(0.02)	(0.00)
Resting blood pressure above 130	0.00*	0.03*	0.00*	0.02**	0.00**
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Serum cholesterol under 210	−0.00	−0.00	−0.00	−0.00	−0.00
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Serum cholesterol between 210 and 270	0.00*	0.01*	0.00*	0.01*	0.00*
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Serum cholesterol between 270 and 430	−0.00	0.01	0.00	0.00	0.00
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Serum cholesterol above 430	−0.00	−0.10	−0.49***	−0.04	−0.48***
	(0.00)	(3.80)	(0.01)	(1.01)	(0.02)
Fasting blood sugar above 120 mg/dl	−0.04	−0.51	−0.05	−0.26	−0.04
	(0.03)	(0.34)	(0.03)	(0.18)	(0.03)
Maximum heart rate achieved below 140	−0.00*	−0.02	−0.00	−0.01	−0.00
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Maximum heart rate achieved above 140	−0.00	−0.02*	−0.00*	−0.01*	−0.00*
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Exercise-induced angina occured	0.12***	0.96***	0.09***	0.56***	0.10***
	(0.03)	(0.26)	(0.03)	(0.14)	(0.03)
Oldpeak below 4	0.03**	0.36**	0.03*	0.21**	0.03**
	(0.01)	(0.14)	(0.01)	(0.07)	(0.01)
Oldpeak above 4	−0.03	−0.54	−0.05	−0.18	−0.03
	(0.07)	(1.16)	(0.04)	(0.58)	(0.04)
Number of major vessels	0.14***	1.53***	0.14***	0.76***	0.13***
	(0.01)	(0.17)	(0.03)	(0.09)	(0.02)
R^2	0.58				
Adj. R^2	0.57				
Num. obs.	1000	1000	1000	1000	1000
AIC		638.48	638.48	647.70	647.70
BIC		775.90	775.90	785.12	785.12
Log Likelihood		−291.24	−291.24	−295.85	−295.85
Deviance		582.48	582.48	591.70	591.70

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Comparing LPM model with Logit and Probit models (quantitative variables are hidden from the third model to fit onto one page but I am controlling the model on them as well)

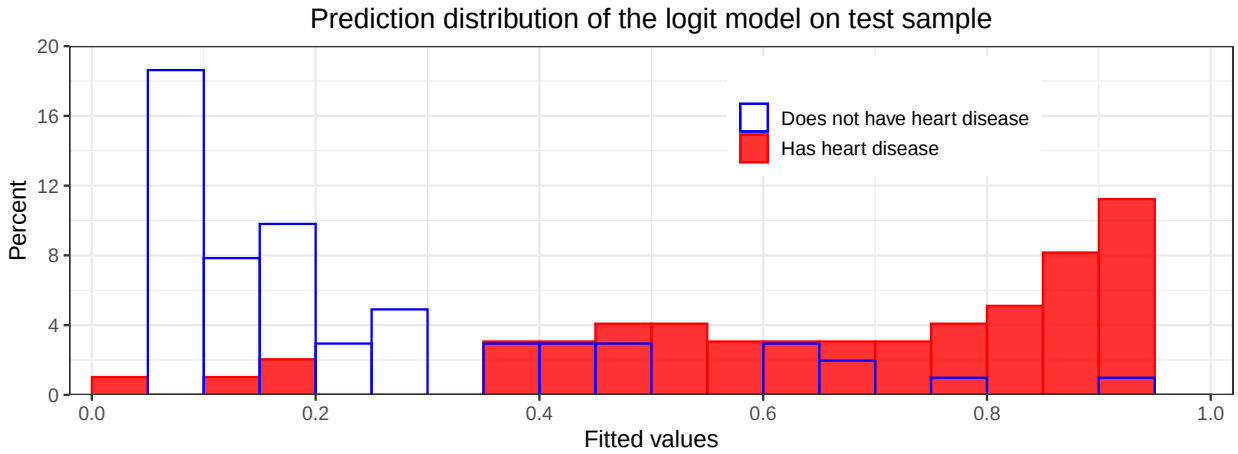


Comparing LPM model with Logit and Probit model predictions

	Logit	Logit robust	Logit margins	Logit robust margins
Intercept	−9.93	−12.10		
	(6.24)	(7.50)		
Gender	1.87***	2.00***	0.17***	0.19***
	(0.34)	(0.37)	(0.03)	(0.03)
Age below 52	−0.10**	−0.08*	−0.01**	−0.01*
	(0.04)	(0.04)	(0.00)	(0.00)
Age between 52 and 61	0.23***	0.22***	0.02***	0.02***
	(0.05)	(0.05)	(0.01)	(0.01)
Age above 61	−0.32***	−0.27***	−0.03***	−0.03***
	(0.06)	(0.06)	(0.01)	(0.01)
Resting blood pressure under 115	0.13**	0.14**	0.01***	0.01**
	(0.05)	(0.06)	(0.00)	(0.00)
Resting blood pressure between 115 and 130	−0.05	−0.06	−0.00*	−0.01*
	(0.03)	(0.03)	(0.00)	(0.00)
Resting blood pressure above 130	0.03*	0.03*	0.00*	0.00*
	(0.01)	(0.01)	(0.00)	(0.00)
Serum cholesterol under 210	−0.00	−0.01	−0.00	−0.00
	(0.01)	(0.01)	(0.00)	(0.00)
Serum cholesterol between 210 and 270	0.01*	0.01	0.00*	0.00
	(0.01)	(0.01)	(0.00)	(0.00)
Serum cholesterol between 270 and 430	0.01	0.01	0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.00)
Serum cholesterol above 430	−0.10	−0.11	−0.49***	−0.49***
	(3.80)	(4.66)	(0.01)	(0.01)
Fasting blood sugar above 120 mg/dl	−0.51	−0.47	−0.05	−0.04
	(0.34)	(0.37)	(0.03)	(0.03)
Maximum heart rate achieved below 140	−0.02	−0.01	−0.00	−0.00
	(0.01)	(0.01)	(0.00)	(0.00)
Maximum heart rate achieved above 140	−0.02*	−0.02	−0.00*	−0.00*
	(0.01)	(0.01)	(0.00)	(0.00)
Exercise-induced angina occured	0.96***	0.91**	0.09***	0.09**
	(0.26)	(0.29)	(0.03)	(0.03)
Oldpeak below 4	0.36**	0.38*	0.03*	0.04*
	(0.14)	(0.15)	(0.01)	(0.01)
Oldpeak above 4	−0.54	−0.29	−0.05	−0.03
	(1.16)	(1.38)	(0.04)	(0.05)
Number of major vessels	1.53***	1.39***	0.14***	0.13***
	(0.17)	(0.18)	(0.03)	(0.03)
AIC	638.48	538.99	638.48	538.99
BIC	775.90	670.16	775.90	670.16
Log Likelihood	−291.24	−241.50	−291.24	−241.50
Deviance	582.48	482.99	582.48	482.99
Num. obs.	1000	800	1000	800

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Comparing model with the model created for robustness check (separated training and test sample (4/5 and 1/5))



The prediction distribution of the model created for robustness check (separated training and test sample (4/5 and 1/5))