University of Warsaw
Faculty of Economic Sciences

Hubert Magdziak
Index: 429848

Identifying Key Factors Influencing Heart Disease Risk

Under the Guidance of

Prof. Rafał Woźniak

Warsaw, May 2024

Table of contents

1. Abstract

Heart disease remains a leading cause of mortality worldwide, necessitating a deeper understanding of its risk factors. This study aims to identify and analyze key variables influencing heart disease risk. Using a comprehensive dataset that includes demographic and clinical factors, we employed statistical methods and econometrics to assess the impact of each variable on being in a group of people with potential heart disease. Apart from the features in the dataset, we focused on feature engineering to find another relevant variables in order to increase explanatory power of the model.

The key part of the empirical research is the Probit model that was estimated to verify two hypotheses – main hypothesis about the influence of age of individual on probability of heart disease and the alternative hypothesis referring to the influence of specific chest pain type on the probability of observing heart disease. According to specific measures, the model enables to predict about 85,2% of the observations correctly. To quantitatively interpret the coefficients we calculated two types of marginal effects – marginal effects for mean observation and average marginal effects.

2. Introduction

With the global population growing rapidly, institutions face a growing need to develop automated tools for assessing various risks, such as credit default, customer churn, or heart disease. Understanding the key factors that significantly increase the risk of heart disease is crucial for developing effective prevention strategies. A thorough analysis of these factors could pave the way for creating a tool that utilizes patient data to identify individuals at higher risk of heart disease and automatically send them email prompts for medical evaluations. This proactive approach could enhance early intervention and lead to improved patient outcomes.

Such tools could have a considerable impact on national demographics and reduce the financial burden associated with combating heart disease. By encouraging at-risk individuals to seek medical attention earlier, we might be able to lower the rates of severe heart-related incidents and, consequently, the overall healthcare costs.

The Centers for Disease Control and Prevention (CDC) provides compelling statistics that underline the urgency of addressing heart disease. According to the CDC, one person dies every 33 seconds in the United States due to cardiovascular disease. Every year about 805,000 people in the United States have a heart attack, whereas of these about 605,000 are a first heart attack and the rest 200,000 happen to people who have already had a heart attack. Furthermore, heart disease incurs significant economic costs, with the United States spending approximately $239.9 billion each year from 2018 to 2019 to cover healthcare services, medications, and lost productivity.[1]

Given these significant statistics, implementing tools that leverage patient data to predict heart disease risk and prompt timely medical evaluations could play a vital role in reducing the prevalence and cost of heart disease in the long term. This approach could also foster a culture of preventive healthcare, shifting the focus from treating heart disease to preventing it, ultimately saving more lives and reducing the economic impact of cardiovascular disease.

[1] https://www.cdc.gov/heartdisease/facts.htm

3. Literature review

The article titled "Heart Disease Prediction using Exploratory Data Analysis" highlights the critical role heart disease plays in causing disability and premature death, emphasizing that heart attacks and other vascular diseases are leading contributors to these outcomes. The authors employed a clustering technique to identify significant factors that could predict heart disease, discovering that age, maximum heart rate, and the type of chest pain are key indicators. For instance, people with atypical angina type of chest pain who belong to the group diagnosed with heart disease tend to have notably higher maximum heart rates. This suggests a potential correlation between high heart rates and this specific type of chest pain in patients with heart disease. Conversely, individuals who experience non-angina chest pain type, typically not associated with classic heart-related issues, are also at increased risk of heart disease if they are older and have elevated maximum heart rates. This implies that even non-traditional symptoms like non-angina chest pain, when coupled with other factors, can indicate a higher risk of heart disease. The dataset used for this analysis is the subset of Cleveland heart disease dataset. Final set of observations chosen for the analysis consists of 209 observations. It includes seven independent features: age, type of chest pain, blood pressure, blood glucose levels, maximum heart rate, and results from electrocardiograms (ECGs) taken at rest, among others.

In the realm of heart disease research, most articles and papers focus on identifying key risk factors and providing explainable models to aid in diagnosis. However, some studies take a different approach, aiming to predict heart disease with high accuracy using advanced techniques. An example of this is the work by Ordas et al., titled "Heart Disease Risk Prediction Using Deep Learning Techniques with Feature Augmentation." This study employs deep learning to predict whether an individual is likely to have heart disease based on a set of clinical features. Early detection in heart disease is crucial, as it significantly influences a patient's survival rate. Ordás et. al.'s research addresses this critical need by utilizing deep learning models. The key variables used in their analysis include age, sex, cholesterol level, sugar level, and heart rate. By applying deep learning techniques, they were able to create a predictive model that outperforms other state-of-the-art methods by 4.4%, achieving a precision rate of 90%. This represents a substantial improvement, particularly in the context of heart disease, which has a high impact on public health. The dataset used for this deep learning model contains 918 samples, each with 11 clinical characteristics. Despite the relatively small number of features, the model's high accuracy

underscores the effectiveness of deep learning in medical applications. This level of precision is especially promising given the high stakes involved in diagnosing and treating heart disease.

Another source that enumerate important factors for heart disease is the article "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction" of Al-Alshaikh H. et. al. Although the article focus on predictive tools, in the introductory part we can read that the risk factors associated with heart disease include aspects such as advanced age, genetic predisposition, tobacco use, physical behaviors, substance misuse, elevated cholesterol levels, hypertension, sedentary lifestyle, obesity, diabetes, psychological stress, and inadequate hygiene practices.

In the "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare" article, the Cleveland heart disease dataset was used for the modelling. They proposed multiple algorithms to find relevant futures for the modelling. One of the algorithms - FCMIM FS – has selected futures like sex, chest pain type, resting blood pressure, serum cholesterol, resting electrocardiographic, maximum heart rate, exercise induced angina, old peak=ST depression induced by exercise to rest, the slope of the peak exercise ST segment, and thallium scan to be crucial for the analysis.

According to the literature and expert knowledge in the area of risk of heart disease we have decided to verify one main hypothesis and two alternative hypotheses:

H_1 : The probability of heart disease is significantly influenced by the age of the individual being analyzed.

H_2 : The probability of heart disease is significantly higher for individual with asymptomatic chest pain type.

4. Model

In this section, we aim to create a statistical model that will help us identify the factors that significantly contribute to heart disease. The outcome we're interested in is binary (1 – heart disease, 0 – normal). This type of outcome requires special models designed to predict probabilities, rather than continuous numerical values. To build a model that can predict the likelihood of heart disease, we have a couple of common options: the Logit model and the Probit model. Both of these are used in econometrics for binary dependent variables and are popular because they provide a clear framework for understanding relationships between predictors and a binary outcome. To determine which model to use, we will consider information criteria's such as Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Then we will proceed with chosen model and interpret marginal effects for special set of observations (marginal effects for mean, average marginal effect and marginal effect for specific observation). The ultimate goal is to use this information to draw meaningful conclusions and provide insights into heart disease risk factors. Due to the number of observations used in analysis – about 1000 - we will use 5% as the significance level alpha.

4.1. Data

For effective intrinsic analysis, it's crucial to use the data that is appropriate, representative, and unbiased. Although there are various beliefs in data science, one key principle stands out: the quality of your analysis is directly tied to the quality of your dataset. To ensure our analysis met these standards, we selected a real-world dataset with a large number of observations, minimizing the risk of violating assumptions about representativity and bias.

In the webpage where the dataset were published[2], we can read that this heart disease data is curated by combining 5 popular heart disease datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland
- Hungarian
- Switzerland
- Long Beach VA
- Statlog (Heart) Data Set

The dataset consists of 11 explanatory variables and binary dependent variable which are described in the Table 1.

[2] https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

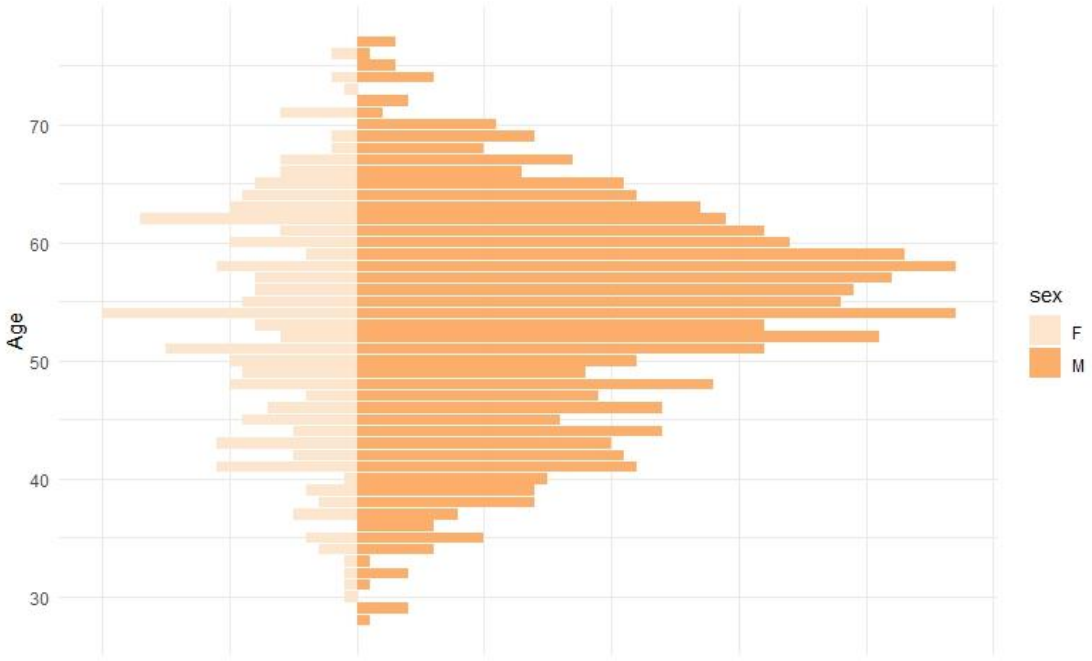
Table 1. Description of the variables used in the dataset.

No.	Variable	Unit	Data type
1	Age	years	Numeric
2	Sex	1 - male 0 - female	Binary
3	Chest pain type	1 – typical angina 2 – atypical angina 3 – non-anginal pain 4 – asymptomatic	Nominal
4	Resting blood pressure	mm Hg	Numeric
5	Serum cholesterol	mg/dl	Numeric
6	Fasting blood sugar	1 – fasting blood sugar > 120mg/dl 0 – fasting blood sugar ≤ 120mg/dl	Binary
7	Resting electrocardiogram results	0 – normal 1 – having ST-T wave abnormality 2 – showing probable or definite left ventricular hypertrophy by Estes' criteria	Nominal
8	Maximum heart rate achived	71 - 202	Numeric
9	Exercise induced angina	1 – yes 0 – no	Binary
10	Oldpeak = ST	depression	Numeric
11	The slope of the peak exercise ST segment	1 – upsloping 2 – flat 3 – downsloping	Nominal
Target variable	Target	1 – heart disease 0 – normal	Binary

Before proceeding to the model, it is good practice to conduct exploratory data analysis. This step is essential for uncovering inherent information and characteristics of the data we are working with.

To get full overview about the *age* variable we created an age pyramid to gain a comprehensive understanding (Figure 1.). The plot reveals a noticeable disparity in the number of observations across different ages. At first glance, it is evident that there are significantly more men than women. Focusing on the age distribution for females, it appears almost uniform for those aged 40 to 67. In contrast, the age distribution for males follows a right-skewed normal distribution, with the majority of observations concentrated in the 50 to 60 age group.

Figure 1. Age pyramid



Now let's delve into more details and examine summary statistics, distinguishing between sex and the presence or absence of heart disease. Since the data distributions via sex are not normal, we'll use the median to summarize certain numeric variables, as the median is more robust to skewed distributions than the mean. The variables summarized are *age*, *max.heart.rate* - maximum heart rate, and *cholesterol* (Table 2).

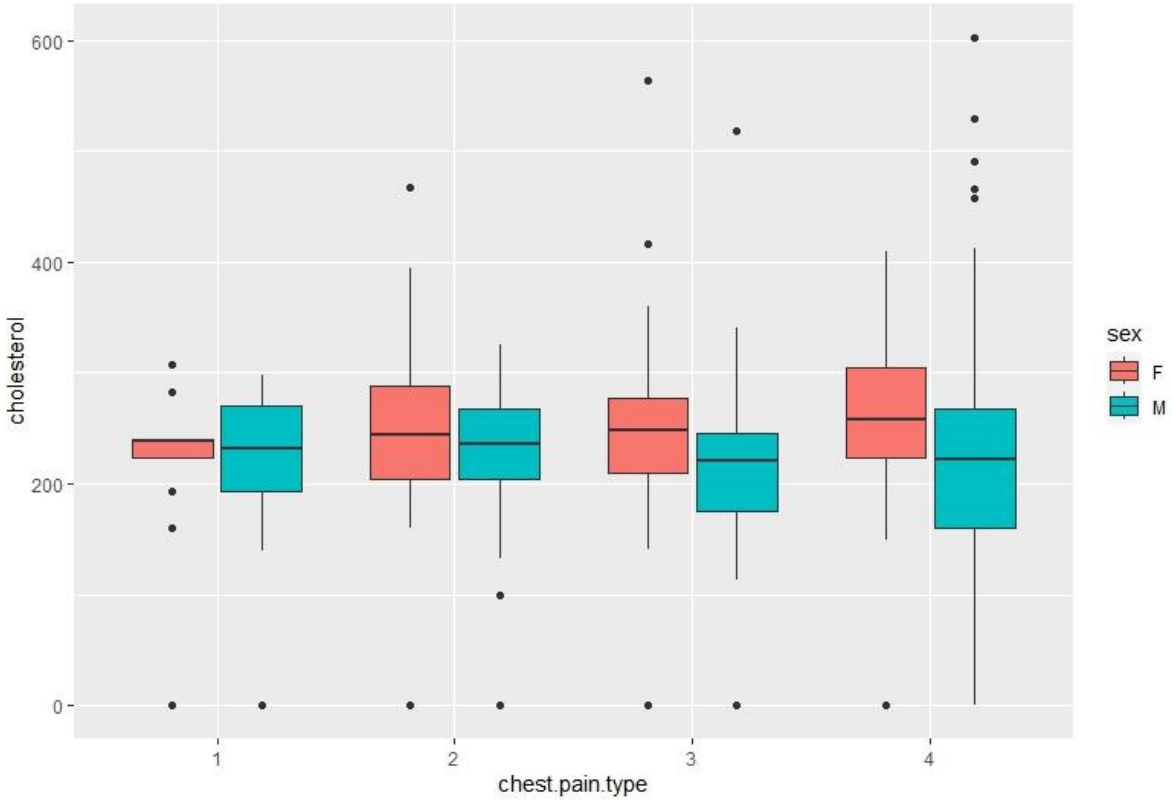
The median age is the highest for females with heart disease - 59 years old, while the lowest median age is observed for males without heart disease - 51 years old. For maximum heart rate, the highest median is found in females without heart disease at 157 bpm, whereas the lowest median is for males with heart disease at 126 bpm. Median cholesterol levels are highest in females with heart disease at 268, and the lowest in males with heart disease at 222.

Table 2. Summary statistics

Sex	***Heart Disease***	***Median of age***	***Median of maximum heart rate***	***Median of cholesterol***	***Number of observations***
Female	No	52	157	242	211
Female	Yes	59	144	268	70
Male	No	51	152	226	350
Male	Yes	57	126	222	559

Another variable that we analyzed is chest.pain.type according to cholesterol level. To visualize the distribution we used the boxplot (Figure 2.). The plot shows that the median cholesterol level is higher in females than in males across all groups. However, the difference is significant only for individuals with asymptomatic chest pain and non-anginal chest pain. Additionally, for women with typical angina chest pain, the interquartile range is relatively small compared to other groups.

Figure 2. Boxplot



4.2. Probit vs Logit

To begin our analysis, we need to select a suitable model for our specific type of study. Since our target variable is binary, with possible outcomes being 0 (no heart disease) or 1 (heart disease), the best modeling choices are the Probit and Logit models. To decide between them, we use two common metrics: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria evaluate model errors while penalizing for the number of variables used, with lower scores indicating better models.

Before calculating the information criteria, we assessed the model specifications using the Linktest and the Hosmer-Lemeshow test. The null hypothesis for both tests is that the specification of the model is correct, whereas alternative hypothesis states: specification of the model is incorrect. For both models Linktest result has statistically significant yhat, whereas yhat2 is statistically insignificant, therefore we fail to reject H_0. In Hosmer-Lemeshow test respectively, the p-values for both Logit and Probit are greater than significance level alpha (5%) – we fail to reject H_0. The conclusions based on summary statistics for these tests are shown in Table 3. Based on these results, we can state that both the Probit and Logit models are correctly specified (as indicated in Table 4).

For the Probit model, the AIC is 867.82 and the BIC is 959.28. The Logit model, on the other hand, has an AIC of 869.72 and a BIC of 961.17 (as shown in Table 4). Given these results, the Probit model appears to be the better choice, as it has lower values for both AIC and BIC. Therefore, we'll proceed with further analysis using the Probit model.

Table 3. Tests for specification – results.

Model	***Linktest (H_0)***	***Hosmer – Lemeshow (H_0)***
Probit	Fail to reject	Fail to reject
Logit	Fail to reject	Fail to reject

Table 4. Comparison of Probit and Logit models.

Model	***AIC***	***BIC***	***Specification***
Probit	867.82	959.28	Correct
Logit	869.72	961.17	Correct

4.3. Features selection

In the field of data science and econometrics, selecting the most suitable features for modeling is crucial. It is a key for creating effective models that can reveal meaningful patterns and relationships between the independent variables and the dependent variable. Among various methodologies for feature selection, such as backward selection, forward selection, and information criteria-based model comparisons (like AIC/BIC), the general-to-specific approach stands out for its methodical approach to refining the model by focusing on statistically significant variables.

The implementation of general-to-specific approach was done by estimating a model that includes all candidate variables. This initial model represents the full set of potential explanatory variables that we hypothesize might influence the dependent variable. Once the model is estimated, we evaluate the statistical significance of each variable. Typically, this is done using p-values, where a higher p-value indicates less statistical significance. The variable with the highest p-value is considered the most statistically insignificant. Having identified the variable with the highest p-value, remove it from the model. This process is based on the assumption that variables with high p-values are not contributing meaningfully to the model's explanatory power. To ensure that the removal of these variables doesn't impact the overall validity of the model, a Linear Hypothesis test is used to check whether the group of variables intended for removal is collectively statistically insignificant. If the test confirms joint insignificance, these variables can be safely dropped from the model. We repeat that process until the removal of any additional variables would cause the set of removed variables to become jointly statistically significant. Once the process reaches a point where removing additional variables would negatively impact the model's statistical significance, the remaining set of variables represents the final model. This model consists of the key features that are statistically significant and collectively explain the dependent variable.

Before applying the method on our dataset we decoded the nominal variables onto multiple binary variables. This results in obtaining *n-1* variables from the variable that contains *n* labels. Additionally we added second power of variable *age* (variable *age2*) and interaction of variables *max.heart.rate* and *age* (variable “max.heart.rate:age”) anticipating non-linear relationship between target variable and variable age. The Finally, that results in 17 different variables which will be taken into consideration while estimating initial model.

Using the general-to-specific approach on our dataset, the final outcome is a model with 10 statistically significant features: *sex*, *cholesterol*, *fasting.blood.sugar*, *max.heart.rate*,

exercise.angina, *oldpeak*, *chest.pain.type.asymptomatic*, *ST.slope.flat*, *ST.slope.downsloping* and *max.heart.rate:age*. The set of variables that turned out to be jointly statistically significant due to Linear Hypothesis testing are: *age*, *age2*, *resting.ecg.1*, *resting.ecg.2*, *resting.bp.s*, *chest.pain.type.atypical.angina*, *chest.pain.type.non.anginal* (Table 5).

Below we mention the Linear Hypothesis test results for the greatest number of variables that turned out to be jointly statistically insignificant.

Linear Hypothesis test results:

$$H_0: \beta_{\text{resting.ecg.1}} = 0 \wedge \beta_{\text{chest.pain.type.atypical.angina}} = 0 \wedge \beta_{\text{chest.pain.type.non.anginal}} = 0 \wedge \beta_{\text{age2}} = 0 \wedge \beta_{\text{age}} = 0 \wedge \beta_{\text{resting.ecg.2}} = 0 \wedge \beta_{\text{resting.bp.s}} = 0$$

$$H_1: \beta_{\text{resting.ecg.1}} \neq 0 \vee \beta_{\text{chest.pain.type.atypical.angina}} \neq 0 \vee \beta_{\text{chest.pain.type.non.anginal}} = 0 \vee \beta_{\text{age2}} \neq 0 \vee \beta_{\text{age}} \neq 0 \vee \beta_{\text{resting.ecg.2}} \neq 0 \vee \beta_{\text{resting.bp.s}} \neq 0$$

Results: p-value: 36% > 5% (significance level alpha), therefore we fail to reject the null hypothesis.

Table 5. Results of the general-to-specific approach. The final model is denoted as (6), whereas the initial one as (1).

	Dependent variable:					
	target					
	(1)	(2)	(3)	(4)	(5)	**(6)**
age	-0.117	-0.118	-0.115	-0.042		
	(0.078)	(0.078)	(0.077)	(0.033)		
sex	0.881***	0.880***	0.875***	0.870***	0.862***	**0.851*** **
	(0.127)	(0.127)	(0.126)	(0.126)	(0.125)	**(0.124)**
resting.bp.s	0.005*	0.005*	0.005*	0.005*	0.005	
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	
cholesterol	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***	**-0.002*** **
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	**(0.001)**
fasting.blood.sugar	0.529***	0.528***	0.525***	0.524***	0.524***	**0.537*** **
	(0.129)	(0.129)	(0.128)	(0.128)	(0.128)	**(0.127)**
max.heart.rate	-0.030**	-0.030**	-0.030**	-0.024*	-0.009***	**-0.009*** **
	(0.014)	(0.014)	(0.014)	(0.013)	(0.003)	**(0.003)**
exercise.angina	0.468***	0.467***	0.470***	0.466***	0.459***	**0.476*** **
	(0.115)	(0.115)	(0.114)	(0.114)	(0.114)	**(0.113)**
oldpeak	0.218***	0.218***	0.217***	0.219***	0.222***	**0.228*** **
	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)	**(0.054)**
chest.pain.type.atypical.angina	0.071	0.070				
	(0.232)	(0.232)				
chest.pain.type.non.anginal	0.080	0.080				
	(0.211)	(0.211)				
chest.pain.type.asymptomatic	1.079***	1.079***	1.013***	1.013***	1.028***	**1.013*** **
	(0.206)	(0.206)	(0.104)	(0.104)	(0.104)	**(0.103)**
resting.ecg.1	-0.028					
	(0.151)					
resting.ecg.2	0.165	0.170	0.166	0.167		
	(0.118)	(0.115)	(0.115)	(0.115)		
ST.slope.flat	1.183***	1.183***	1.180***	1.175***	1.171***	**1.159*** **
	(0.116)	(0.116)	(0.115)	(0.115)	(0.114)	**(0.113)**
ST.slope.downsloping	0.482**	0.482**	0.477**	0.471**	0.461**	**0.447** **
	(0.219)	(0.219)	(0.219)	(0.219)	(0.217)	**(0.217)**
I(age2)	0.001	0.001	0.001			
	(0.001)	(0.001)	(0.001)			
age:max.heart.rate	0.0005*	0.0005*	0.0005*	0.0004	0.0001**	**0.0001*** **
	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.00004)	**(0.00004)**
Constant	2.695	2.702	2.721	0.386	-1.958***	**-1.371*** **
	(2.896)	(2.893)	(2.890)	(1.887)	(0.528)	**(0.381)**
Observations	1,189	1,189	1,189	1,189	1,189	**1,189**
Log Likelihood	-415.910	-415.928	-416.004	-416.535	-418.415	**-419.823**
Akaike Inf. Crit.	867.820	865.856	862.008	861.071	860.830	**861.646**
Note:	*p<0.1; **p<0.05; ***p<0.01					

4.4. Model Validation

When the intrinsic features for modeling are chosen and the model is estimated, the next step for building successful model is model validation. This involves checking whether the assumptions our model relies on hold true. Doing so is essential because it helps us determine whether the coefficients in our model can be trusted and whether our conclusions can be inferred to the broader population. In our study, our final model is a Probit, specifically designed for binary dependent variables. One of the key assumptions underlying this model is that its specification is accurate. To verify this assumption, we utilized two important tests: the Linktest and the Hosmer-Lemeshow test. These tests are valuable tools for ensuring the reliability of our model's assumptions and, consequently, the validity of our findings. According to the Linktest results, the yhat is statistically significance, whereas yhat2 is not (Table 6.). Respectively, below we mention hypotheses and results for Hosmer-Lemeshow test.

Hosmer-Lemeshow test results:

H_0 : The specification of the model is correct

H_1 : The specification of the model is incorrect

Results: p-value = 16.3% > 5% (significance level alpha) – we fail to reject null hypothesis. Due to the test, the specification of the model is correct.

Concluding on the basis of both tests, the specification of the final model is correct.

Table 6. Results of the Linktest specification test for the final model.

	Estimate	***Std. Error***	***Z-value***	***P-value***	***Significance (Yes / No)***
yhat	0.567	0.0277	20.484	<2e-16	Yes
yhat2	-0.006	0.0126	-0.591	0.555	No

On the other hand we checked another important assumption which is lack of multicollinearity in the model. The tool that was used is Variance Inflation Factor (VIF), which provides a measure of multicollinearity among the independent variables. The interpretation of the results is as mentioned in Jong Hae Kim - "Multicollinearity and misleading statistical results". The rule of thumb works as if the variance inflation factor is higher or equal than 5, multicollinearity exists. Although the variance inflation factor helps to determine the presence of multicollinearity, it cannot detect the explanatory variables causing the multicollinearity.

According to the final model, all the variables have Variance Inflation Rate in the interval (1;1.77) (Table 7.). Therefore the assumption about lack of multicollinearity in the model is fulfilled.

Table 7. Variance Inflation Rate (VIF) for the variables in the final model.

Variable	***VIF***
Sex	1.078
Cholesterol	1.147
Fasting.blood.sugar	1.070
Max.heart.rate	1.763
Exercise.angina	1.212
Oldpeak	1.360
Chest.pain.type.asymptomatic	1.114
ST.slope.flat	1.350
St.slope.downsloping	1.386
Max.heart.rate:age	1.647

4.5. Model Performance

Having validated the model, we now turn to the crucial task of assessing its performance. To do that we will rely on 3 most commonly used statistics that are dedicated for model assessment – *McKelvey-Zavoina R2*, *Count R2*, *Adjusted Count R2*. The results are shown in Table 8.

Table 8. Model performance

Statistics	***Value***
McKelvey-Zavoina	0.672
Count R2	0.852
Adjusted Count R2	0.686

Due to the *McKelvey-Zavoina R2*, if a hidden variable would be observed, the model would explain about 67.2% of its total variability.

According to the *Count R2*, about 85.2% of the observations were predicted correctly by the model.

Interpreting the value of *Adjusted Count R2*, about 68.6% of the observations were predicted correctly only based on the variables characteristics that were used for modelling, no regards to the p* (threshold) level.

4.6. Marginal effects

To facilitate a quantitative interpretation of the parameters in the final model, we calculated the marginal effects. Table 9 presents the marginal effects evaluated at the mean observation, providing insights into how changes in each predictor influence the dependent variable at the mean values of the predictors. On the other hand, Table 10 displays the average marginal effects, which offer an overall summary of the predictor impacts averaged across all observations in the dataset. This dual approach allows for a more comprehensive understanding of the model's behavior and the relative importance of each predictor.

Table 9. Marginal effects for mean observation

Variable	*dF/dx*	*Std. Err.*	*Z*	*p-value*
Sex	0.325	0.044	7.457	0 ***
Cholesterol	- 0.0001	0.0002	-3.342	0 ***
Fasting.blood.sugar	0.216	0.004	4.817	0 ***
Max.heart.rate	- 0.004	0.001	-3.756	0 ***
Exercise.angina	0.190	0.043	4.474	0 ***
Oldpeak	0.109	0.019	5.678	0 ***
Chest.pain.type.asymptomatic	0.386	0.036	10.677	0 ***
St.slope.flat	0.404	0.036	11.307	0 ***
Max.heart.rage:age	0.00005	0.00002	2.774	0.006 **

According to the results in Table 9, we can derive several quantitative insights. For instance, a male with average characteristics has a probability of having heart disease that is approximately 32.5 percentage points higher than that of a female, holding all other factors constant (ceteris paribus). Additionally, an individual with average characteristics and a fasting blood sugar level greater than 120 mg/dl has a probability of having heart disease that is about 21.6 percentage points higher than someone with a fasting blood sugar level of 120 mg/dl or lower, again holding all other factors constant (ceteris paribus).

Alternatively, when we want to derive general insights, we can refer to Table 10. On average, a male has a probability of having heart disease that is approximately 18 percentage points higher than that of a female, holding all other factors constant (ceteris paribus). Additionally, on average, a person with a fasting blood sugar level greater than 120 mg/dl has a probability of having heart disease that is about 11.5 percentage points higher than someone with a fasting blood sugar level of 120 mg/dl or lower, again holding all other factors constant (ceteris paribus).

Table 10. Average marginal effects

Variable	*dF/dx*	*Std. Err.*	*Z*	*p-value*
Sex	0.180	0.044	7.457	0 ***
Cholesterol	- 0.0003	0.0002	-3.342	0 ***
Fasting.blood.sugar	0.115	0.004	4.817	0 ***
Max.heart.rate	- 0.002	0.001	-3.756	0 ***
Exercise.angina	0.106	0.043	4.474	0 ***
Oldpeak	0.054	0.019	5.678	0 ***
Chest.pain.type.asymptomatic	0.242	0.036	10.677	0 ***
St.slope.flat	0.258	0.036	11.307	0 ***
Max.heart.rage:age	0.00002	0.00002	2.774	0.005 **

5. Results

In the literature, age has been identified as a crucial factor in assessing the probability of heart disease. To capture the relationship between the dependent variable and age, we initially included the squared term of age (variable *age2*) in the model, as well as its interaction with the variable maximum heart rate (*max.heart.rate:age*). Although the final model shows that neither variable age nor age2 are statistically significant on their own (model (6) of Table 5), the interaction between age and maximum heart rate is statistically significant, with a positive coefficient. Therefore, based on our final model, we can assert that age significantly influences the probability of having heart disease and accept the hypothesis.

The analysis indicates that individuals with asymptomatic chest pain have a markedly increased likelihood of heart disease. In the final model (model (6) of Table 5), the variable representing asymptomatic chest pain type (*chest.pain.type.asymptomatic*) is statistically significant. The positive coefficient of 1.013 for this variable suggests that people experiencing asymptomatic chest pain have a higher probability of being classified in the group with heart disease. Therefore, we accept the hypothesis that asymptomatic chest pain is associated with a significantly higher probability of heart disease.

6. Findings

By combining data from five important datasets, we were able to validate two hypotheses stated at the beginning of this paper. Our conclusions indicate that the probability of heart disease is significantly influenced by the age of the individual being analyzed. Additionally, we demonstrated that individuals with asymptomatic chest pain have a significantly higher probability of heart disease.

A crucial part of the analysis was the calculation of marginal effects. Both the marginal effects for average characteristics and the average marginal effects provide substantial quantitative interpretations. For example, a male with average characteristics has a probability of having heart disease that is approximately 32.5 percentage points higher than that of a female, holding all other factors constant (ceteris paribus). Conversely, on average, a male has a probability of having heart disease that is approximately 18 percentage points higher than that of a female, holding all other factors constant (ceteris paribus).

Regarding the fasting blood sugar variable, an individual with average characteristics and a fasting blood sugar level greater than 120 mg/dl has a probability of having heart disease that is about 21.6 percentage points higher than someone with a fasting blood sugar level of 120 mg/dl or lower, holding all other factors constant. On the other hand, on average, a person with a fasting blood sugar level greater than 120 mg/dl has a probability of having heart disease that is about 11.5 percentage points higher than someone with a fasting blood sugar level of 120 mg/dl or lower, again holding all other factors constant.

In future analyses, it will be important to develop a transparent "glass-box" model that not only improves higher prediction accuracy but also provides intrinsic insights into the variables that are significant for assessing the likelihood of heart disease.

7. Bibliography

Al-Alshaikh, H.A., P, P., Poonia, R.C. et al. Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. Sci Rep 14, 7819 (2024). https://doi.org/10.1038/s41598-024-58489-7

García-Ordás, M.T., Bayón-Gutiérrez, M., Benavides, C. et al. Heart disease risk prediction using deep learning techniques with feature augmentation. Multimed Tools Appl 82, 31759–31773 (2023). https://doi.org/10.1007/s11042-023-14817-z

Hossain, Mohammad. (1998). AIC and BIC – The two competitive information criteria for model selection in economics and statistics. Journal of Social Science: Part II. 19. 133-140.

https://www.cdc.gov/heartdisease/facts.htm

https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

Khan, Aminulhaq. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare.

Kim, Jong. (2019). Multicollinearity and misleading statistical results. Korean Journal of Anesthesiology. 72. 10.4097/kja.19087.

Manu Siddhartha, November 5, 2020, "Heart Disease Dataset (Comprehensive)", IEEE Dataport, doi: https://dx.doi.org/10.21227/dz4t-cm36

R. Indrakumari, T. Poongodi, Soumya Ranjan Jena, Heart Disease Prediction using Exploratory Data Analysis, Procedia Computer Science, Volume 173, 2020, Pages 130-139, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.06.017.

8. Appendix

Link to the repository: https://github.com/Hubert-Magdziak/Advanced-Econometrics

The repository consists of:

- data used for the modelling - *heart_statlog_cleveland_hungary_final.csv*
- documentation for the data - *documentation.pdf*
- code to reproduce the findings - *code.R*
- code to calculate the Linktest - *linktest.R*[3]
- code to calculate the marginal effects - *marginaleffects.R*[4]
- .jpeg file of Figure 1. - *pyramid.jpeg*
- .jpeg file of Figure 2. - *boxplot.jpeg*

[3] written by dr Rafal Wozniak, Faculty of Economic Sciences, University of Warsaw
[4] written by dr Rafal Wozniak, Faculty of Economic Sciences, University of Warsaw

List of Tables and Figures

List of Tables

List of Figures