# 

# 

# 

# **Analysis of Framingham Heart Study Data using Nonparametric Methods**

Group Members:

Zhuo Chen (zc2442);

Hang Qin (hq2158);

Caspar Chen (jc5067);

Jiyuan Liu (jl5217)

**I. Introduction**

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free-living subjects in the community of Framingham, Massachusetts. 5,209 patients were enrolled in 1948, and now on the third generation is the majority. People are given questionnaires and exams every two years in order to gather data, including Physical characteristics, Behavioral characteristics, and Test results.

Our dataset has 5,209 observations and 16 variables. Among all records, 5039 of these are complete cases, and the rest have some missing values. Moreover, we define a new response variable Y: Response, which is the indicator function I(AgeCHDdiag-AgeAtStart <= 10). It means the 10-year risk of coronary heart disease(CHD) of an individual patient. We mainly study the relationship between Response and 14 risk factors, including five qualitative variables and nine quantitative variables.

First, we explored whether there existed a significant correlation between every single variable and the response variable in nonparametric methods. Then, we use the risk factors that are associated with Response from nonparametric tests to build a regression model to predict whether a participant would develop CHD in 10 years. During the study, we mainly focus on four sections. First, we explore whether non-smokers have a higher survival rate than smokers and whether the female has a higher survival rate than the male by using Survival Analysis. In the second part, we perform three correlation tests and two sample tests between all quantitative variables (AgeatStart, Weight, Cholesterol) and 10-yr CHD. We also study association between qualitative variables and 10-yr CHD using Fisher’s exact test in the third section. In the last part, we build a logistic regression model to predict whether participants would develop CHD in 10 years.

We have used multiple methods from our course including K-M Curve; Log-Rank test; Cox model and partial likelihood test; Fisher’s Exact Test; Correlation Test; Two-sample t-Test; Kruskal-Wallis Test;Wilcoxon Rank Sum Test & Hodges-Lehmann Estimator; F Test.

**II. Results**

## 1. Survival Analysis

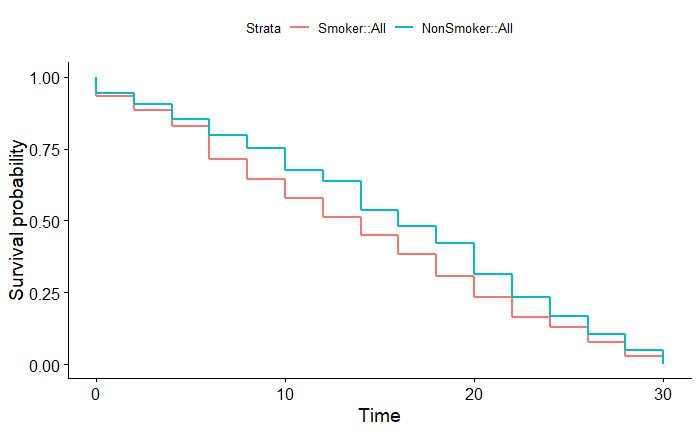
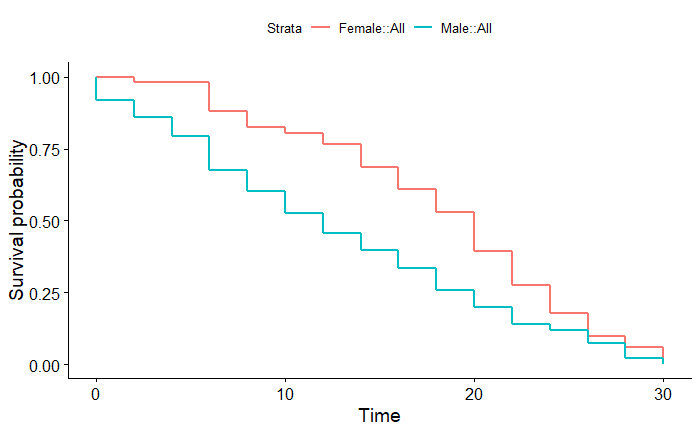
For the survival analysis of the Framingham Heart Study, we mainly want to explore the answers to the following two questions:

1. Will non-smokers generally have higher “survival rate” than smokers?
2. Will female generally have higher “survival rate” than the male?

We will try to answer these two questions using K-M curve, log-rank test, Cox model and partial likelihood ratio test.

But first of all, How do we define the “survival rate” here in our context? In our case, we want to explore if smoking or gender will affect patients getting CHD after the patients entering the research. So our survival rate is now measured in terms of the probability of getting CHD.

Thus, in the KM curve, Y-axis is the probability of the patient getting CHD. And the X axis will be the time which is calculated by AgeGetCHD-AgeAtStart. Age at the start is just the patient’s age when the patient enters the research. AgeGetCHD is the patient’s age when the patient gets CHD.

**KM Curve Smoking**  **KM Curve Gender**

As can be seen from the two graphs above, generally male have a higher probability of getting CHD than female. Also, smokers generally have a higher probability of getting CHD than non-smokers.

To further determine what we observe from the two KM Curves, we also did the Log-Rank test and the Cox model and partial likelihood ratio test. The outputs can be found in the Figure 0.1, 0.2[[1]](#footnote-0). Our conclusion drew from the results are the p-values of these tests are all smaller than 0.05. Therefore, we conclude that female generally have a lower probability of getting CHD than male. And non-smokers generally have a lower probability of getting CHD than smokers. Smoking or gender will significantly affect patients getting CHD after the patients entering the research.

## 2. Correlation Test & Two-sample test between Quantitative variables and Response Y

## Correlation Test[[2]](#footnote-1)

There are 5,047 observations excluding missing data. The three correlation test is conducted between all quantitative variables (AgeatStart, Weight, Cholesterol, etc.) and Response Y. Y is defined as Y = 1 when people get Coronary Heart Disease in ten years; Y = 0, when people do not get Coronary Heart Disease in ten years.

From the three correlation test, we found that the Age at start variable when people first entered the study is the most significant among all, which is moderately positive correlated between whether people get CHD in ten years or not, and then following by Systolic BP, Diastolic BP, and Cholesterol. These three quantitative variables can be summarized as body indexes of a human. Weight also plays an important role which can be used for further variable selections.

Thus, we decide to use these body indexes for further two-sample statistical diagnose as below. We now divide the data into two groups, the group without CHD have 4635 observations while the group with CHD have 412 observations.

Two-sample t test[[3]](#footnote-2)

For all three variables, since p-value < 2.2e-16, at 5% we reject the null hypothesis and in favor of the alternative to concluding significance. For a one-tailed test, the conclusion is that people do not get CHD in ten years at the end generally have less those body indexes than people get CHD in ten years when they enter the experiment at the very beginning.

Kruskal-Wallis Test[[4]](#footnote-3)

For Cholesterol between two groups, KW stat = 79.682. At 5% significance level, we reject the null hypothesis and in favor of the alternative and conclude that the Cholesterol in people without CHD and with CHD are non-identical populations.

Wilcoxon Rank sum test & Hodges-Lehmann estimator[[5]](#footnote-4)

For the two sample Wilcoxon test, what we are conducting is the Mann-Whitney test. Since p-value < 2.2e-16 for all three variables, significance found at 5% between people’s Cholesterol, Diastolic blood pressure, Systolic blood pressure between two groups. For a one-tailed test, we found that people do not get CHD in ten years at the end generally have less those body indexes than people get CHD in ten years when they enter the experiment at the very beginning.

### F-test[[6]](#footnote-5)

In order to test the variance of the two groups, we conducted F-test. Since p-value for all three variables is less than 5%, there is enough evidence to conclude the significance between variances of the two groups.

## 3. Fisher’s exact test for contingency tables

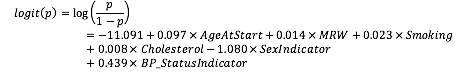
We perform Fisher’s exact test for the contingency tables (Figure 6.1-5[[7]](#footnote-6)) to study the association between 10-yr CHD and some qualitative variables.

The results showed that 10-yr CHD is associated with sex. Meanwhile, a male is more likely to develop CHD than a female. Results also indicated that 10-yr CHD is associated with cholesterol status, blood pressure status, weight status, smoking status, which is similar to the results of the association between 10-yr CHD and a quantitative variable.

4. Regression model

After previous analysis of the association between 10-yr CHD and quantitative and qualitative variables, we want to make a logistic regression model to predict whether participants would develop CHD in 10 years after they join the study. Six risk factors, including AgeAtStart, MRW, Smoking, Cholesterol, Sex (1- female, 0 - male) and BP\_Status (0 - optimal, 1 - normal, 2 - high) are chosen so that there is no high correlation between the predictor variables. Result of the regression model is summarized in the Figure 7[[8]](#footnote-7). It shows the same result as previous nonparametric tests: all these risk factors are significant.

Then, we can use the model:



to predict the possibility of getting CHD in 10 years.

## **III. Conclusion**

By using non-parametric methods, we conclude that the male has a higher probability of getting CHD than female. Also, Smokers generally have a higher probability of getting CHD than non-smokers. Based on three correlation tests and two-sample tests, we found that the start Age was the most significant variable among all. People who do not get CHD in ten years in the end generally have less those body indexes (Systolic BP, Diastolic BP, and Cholesterol) than people get CHD in ten years when they enter the experiment at the very beginning. According to the Fisher’s exact test, the results showed that 10-yr CHD is associated with cholesterol status, blood pressure status, weight status, smoking status, sex status. Lastly, we build a regression model using six risk factors. It shows similar results as previous nonparametric test that all six variables are statistically significant.

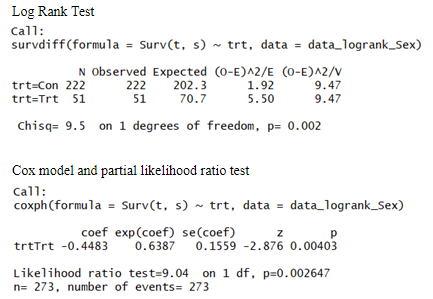
Our analysis of Framingham Heart Study is based on a subset of data that only include 16 variables when participants first participated the study back in 1948. With more complete data, we may study association between 10-yr CHD and more various risk factors, like use of antihypertensive medication, stroke history and diabetic or not which are possible risk factors for CHD. We may also study relationship of risk factors and progress of CHD if we get data of each participant during their follow-up exam.

## Appendix

### Figure 0.1. Survival Analysis Output 1

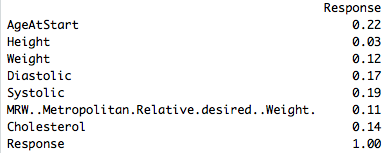
### 

### Figure 0.2. Survival Analysis Output 2

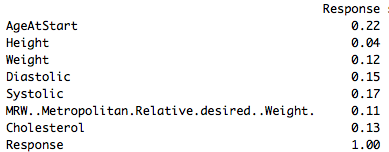


### Figure 1. Correlation Test

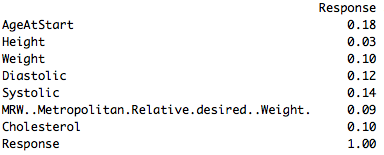
Pearson’s r



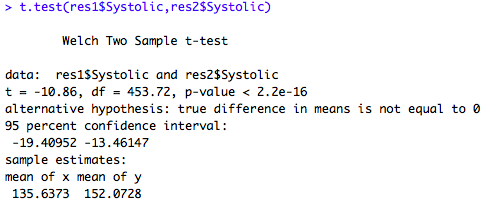
Spearman’s ρ

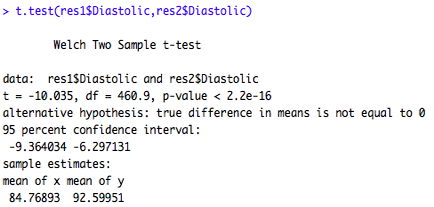


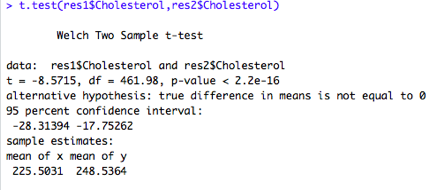
Kendall’s tau



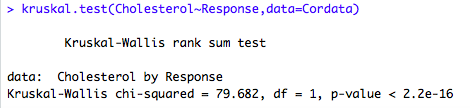
### Figure 2. Two-sample t test

****

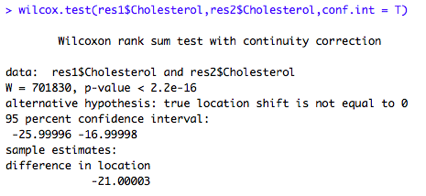
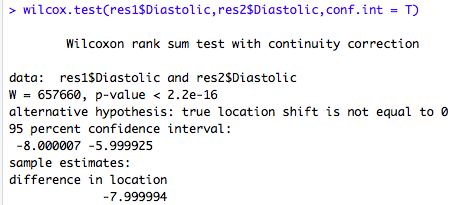
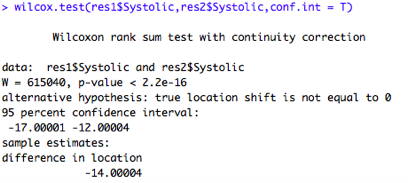
****



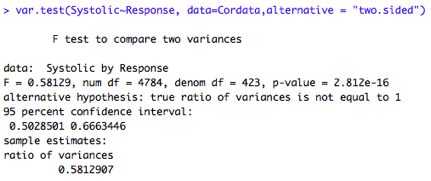
### Figure 3. Kruskal-Wallis Test

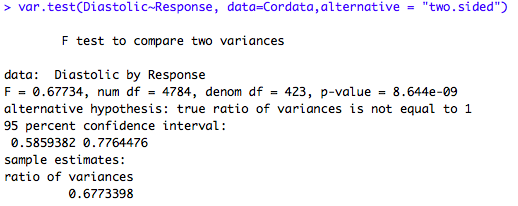


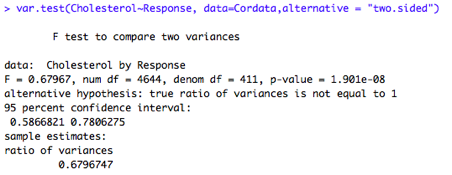
### Figure 4. Wilcoxon Rank sum test & Hodges-Lehmann estimator



### Figure 5. F-test







### Figure 6.1. Contingency table of Sex and 10-yr CHD

|  |  |  |  |
| --- | --- | --- | --- |
|  | Have CHD in 10 Years | Not have CHD in 10 Years |  |
| Female | 137 | 2627 | 2764 |
| Male | 275 | 2000 | 2275 |
|  | 412 | 4627 | 5039 |

**Figure 6.2. Contingency table of Cholesterol status and 10-yr CHD**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Have CHD in 10 Years | Not have CHD in 10 Years |  |
| Desirable | 66 | 1331 | 1859 |
| High | 220 | 1563 | 1397 |
| Borderline | 126 | 1733 | 1783 |
|  | 412 | 4627 | 5039 |

**Figure 6.3. Contingency table of Blood pressure status and 10-yr CHD**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Have CHD in 10 Years | Not have CHD in 10 Years |  |
| Optimal | 23 | 743 | 2198 |
| Normal | 109 | 1966 | 2075 |
| High | 280 | 1918 | 766 |
|  | 412 | 4627 | 5039 |

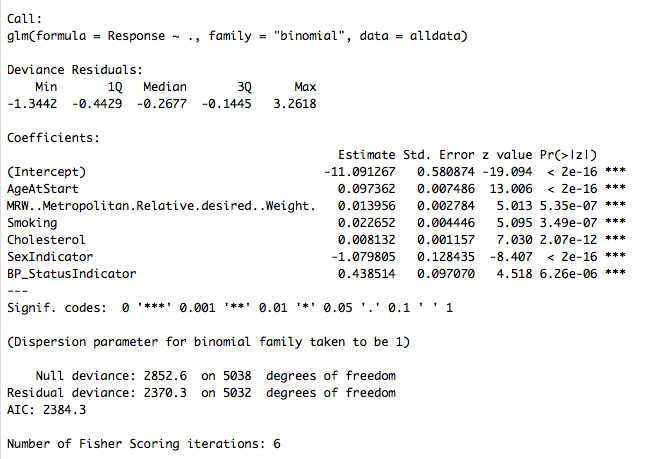
**Figure 6.4. Contingency table of Weight status and 10-yr CHD**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Have CHD in 10 Years | Not have CHD in 10 Years |  |
| Normal | 71 | 1356 | 1427 |
| Overweight | 335 | 3101 | 3436 |
| Underweight | 6 | 170 | 176 |
|  | 412 | 4627 | 5039 |

**Figure 6.5. Contingency table of Smoking status and 10-yr CHD**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Have CHD in 10 Years | Not have CHD in 10 Years |  |
| Non-smoker | 180 | 2253 | 2433 |
| Light (1-5) | 35 | 526 | 561 |
| Moderate (6-15) | 37 | 525 | 562 |
| Heavy (16-25) | 97 | 930 | 1027 |
| Very Heavy (>25) | 63 | 393 | 456 |
|  | 412 | 4627 | 5039 |

### Figure 7. Logistic Regression Output



1. [Figure 0.1. Survival Analysis Output 1](#_dh3hogldu0st) & [Figure 0.2. Survival Analysis Output 2](#_lfaupc2nov) [↑](#footnote-ref-0)
2. [Figure 1. Correlation Test](#_wiueb6yng370) [↑](#footnote-ref-1)
3. [Figure 2. Two-sample t test](#_x5g2o7sxu4a) [↑](#footnote-ref-2)
4. [Figure 3. Kruskal-Wallis Test](#_gwusq96qt5n6) [↑](#footnote-ref-3)
5. [Figure 4. Wilcoxon Rank sum test](#_seshf7ddr24h)  [↑](#footnote-ref-4)
6. [Figure 5. F-test](#_qufynvaohoxa) [↑](#footnote-ref-5)
7. [Figure 6.1. Contingency table of Sex and 10-yr CHD](#_gh7w6kmua2h) [↑](#footnote-ref-6)
8. [Figure 7. Logistic Regression Output](#_q5qefk3zgszl) [↑](#footnote-ref-7)