Project 3, Analysis Plan

Joseph Froelicher

November 6, 2021

Introduction

Beginning in 1948, subjects were enrolled in a Heart disease study in the city of Framingham, Massachusetts. Our data come from a subset of that data consisting of the years 1956 - 1968. Participants were followed by hospital visits, participant contact and through death certificates for the occurrence of Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease.

Using the 10-year probability of stroked based on different risk profiles, the goal is to identify any risk factors for stroke based on our data. These are our two aims for this analysis:

1. Determine which risk factors are most associated with increased hazard of having a stroke
2. Analyze how much temporal change in the associated risk factors, and whether a more sophisticated temporal analysis is needed.

Data

Our data consist of $4,434$ participants who's data were collected in 6 year periods from approximately 1956 - 1968. Each participant has anywhere from one to three observations, depending on how many examinations they attended ($11,627$ observations). During the cleaning process, 32 participants were dropped due to having a previous stroke in the first period, so the usable data has $4,402$ participants. The list of outcomes collected include: Angina Pectoris, Myocardial Infarction, Atherothrombotic Infarction, Cerebral Hemorrhage (Stroke) or death. The data include information about binary sex (M/F), age, blood pressure, smoking status, medication use, cholesterol, BMI, glucose level, and diabetes status. There is missing data throughout the data set, primarily if four variables, High Density lipoprotein Cholesterol, Low Density Lipoprotein Cholesterol, BMI, and glucose. We are interested in the time to event of Cereberal Hemorrhage (stroke), in the first 10 years. In our data there were 2.52% who had a stroke in the 10 year period (n = 111).

Analysis Plan

To address aim 1, we are analyzing which risk factors are most highly associated with the hazard of stroke. There are several risk factors available, they include: age, systolic blood pressure, diastolic blood pressure, use of anti-hypertensive medication, current cigarette smoking, number of cigarettes per day, cholesterol, body mass index, glucose, diabetes status, and heart rate. As per the request of the principal investigator, these analysis will be stratified by sex. To assess the time to event of the hazard of stroke, Kaplan-Meier curves will be used. As mentioned above, events are strokes that occurred in the first 10 years of the study, and the time of each even is the minimum of: the event, death, and time to censoring (10 years). We will also build a Cox proportional hazards model to asses all of the associated risk factors in a model together. Then variable selection will be performed as outline in Tibshirani (1997) and Zhang & Lu (2007) using LASSO or Adaptive-LASSO for Cox's proportional hazards models.

To address aim 2, we will investigate the assumption of Cox's proportional hazards models that the hazards themselves are actually proportional. To assess the proportional hazards, we will make "$log(-log(S(t)))$" graphs, and check the Schoenfeld residuals. And if there appears to be some violation of the proportionality assumption, then we will interact the variables that are not proportional with time within a second proportional hazards model to reassess the hazard of stroke with selected risk factors. If necessary, we may also build generlaized linear models with each of the associated risk factors as outcomes, to assess the significance of time against each associated outcome.

Tables and Figures

	Level	Censored	Stroke	Missing (%)	Overall
N		4291	111		4402
Sex (%)	Men	1881 (43.8)	49 (44.1)	0	1930
	Women	2410 (56.2)	62 (55.9)	0	2472
Cholesterol		236.76 ± 44.59	246.83 ± 46.15	1.181	237.01 ± 44.65
Age		49.70 ± 8.63	56.53 ± 7.37	0	49.88 ± 8.66
sBP		132.10 ± 21.65	157.98 ± 31.62	0	132.76 ± 22.32
dBP		82.73 ± 11.73	94.54 ± 16.32	0	83.03 ± 12.01
Smoke (%)	No	2175 (50.7)	58 (52.3)	0	2233
	Yes	2116 (49.3)	53 (47.7)	0	2169
Cigs/day		8.99 ± 11.94	9.02 ± 12.58	0.704	8.99 ± 11.95
BMI		25.80 ± 4.05	27.00 ± 4.91	0.386	25.83 ± 4.08
Diabetees (%)	No	4183 (97.5)	100 (90.1)	0	4283
	Yes	108 (2.5)	11 (9.9)	0	119
Medication	No	4113 (97.1)	93 (86.9)	1.363	4266
	Yes	122 (2.9)	14 (13.1)	1.363	196
Heart Rate		75.89 ± 12.12	76.72 ± 12.27	0.023	75.91 ± 12.13
Death (%)	No	2864 (66.7)	13 (11.7)	0	2877
	Yes	1427 (33.3)	98 (88.3)	0	1525
Time to Event (Median)		3650 ± 555.09	2144 ± 555.09	0	3650 ± 614.08

Table 1. Descriptive Statistics for the Framingham heart data. Continuous variables have means reported, categorical variables have counts and percentages reported.

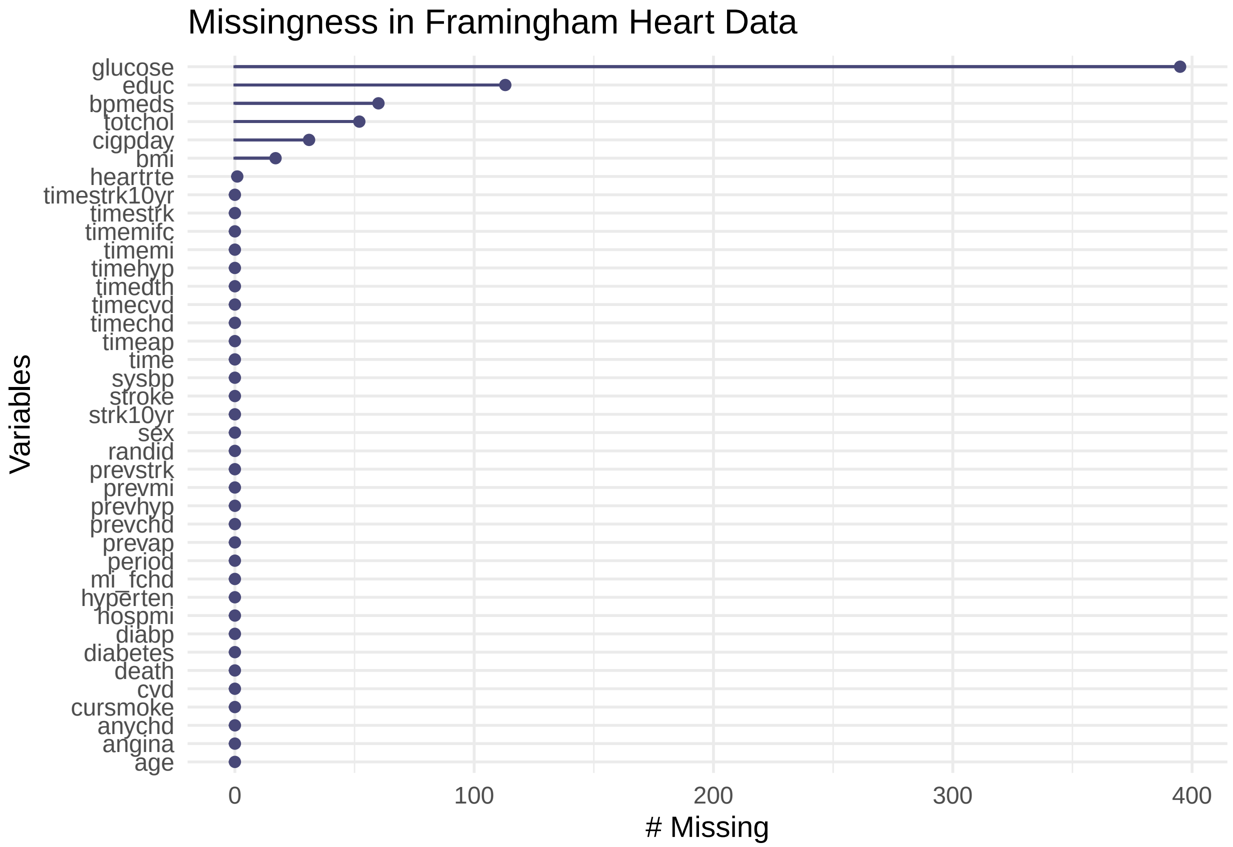

Figure 1. Missingness plot for Framingham Heart data.

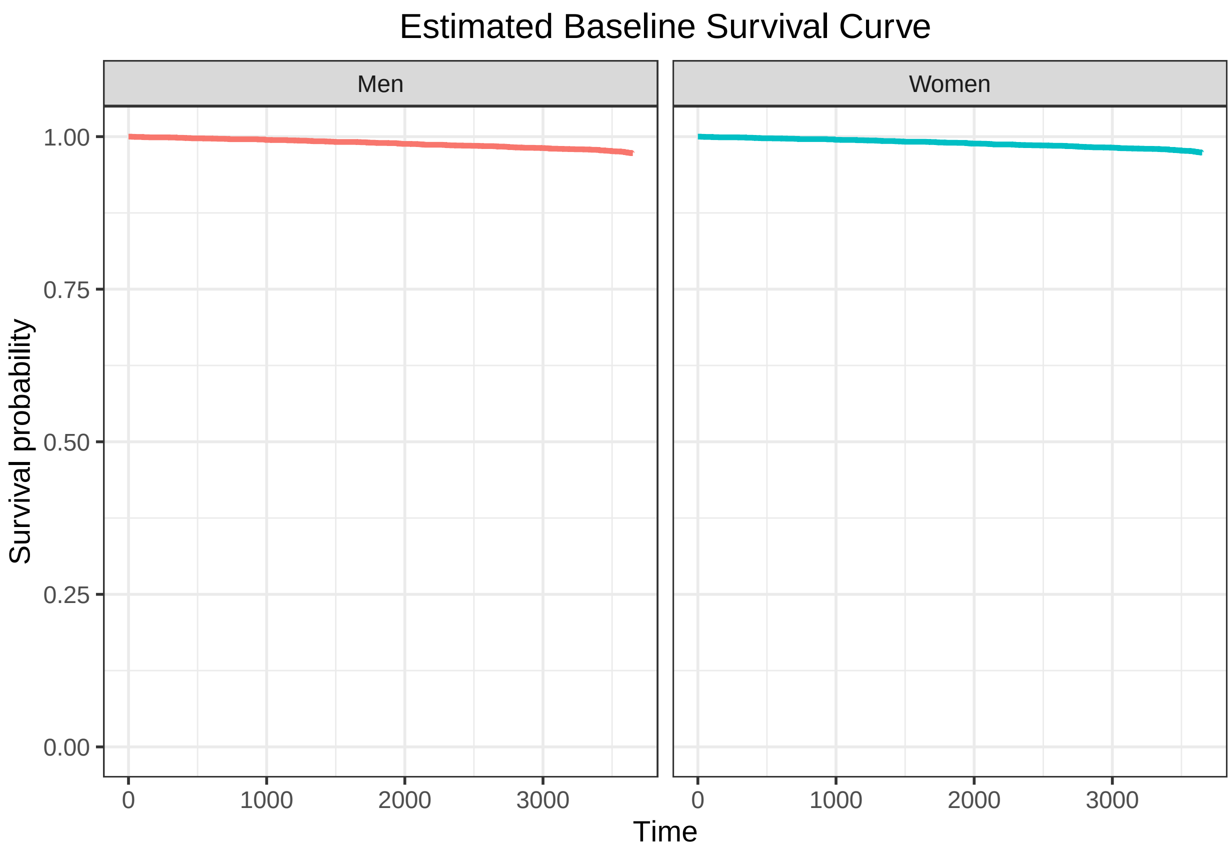

Figure 2. Estimate Baseline sruvival curves for men and women, with no covariates added.