

Survival Analysis of Worsening Hypertension

PSTAT 196 Report

John Randazzo, Jason Freeberg, Ziyi Jiang

4/16/2017

Introduction

The following is a full report on our data analysis project for the course PSTAT 196: Projects in Actuarial Science. Our analysis was completed during the Spring Quarter of 2017 under the supervision of Professor Ian Duncan and graduate student Shannon Nicponski at The University of California, Santa Barbara.

About the Data

Our dataset comes from a longevity study performed in the United Kingdom from March 2006 to May 2014. Our team's subset of the original data is concerned only with high-risk patients developing severe hyper-tension. All participants in our dataset were already diagnosed with moderate hypertension. Our task was to model the time until diagnosis of sever hypertension—which lends itself to a classic Survival Analysis problem. Our dataset contains 126665 observations and 13 variables. A complete breakdown of the the variables is provided below.

Table 1. A breakdown of all covariates in the original dataset

Name	Explanation	Range
UNQID	A unique ID for each individual	
Sex	The patient's sex	1:
Socio-Economic	A categorical range indicating economic status	1: Affluent -> 5: Deprived
Socio-Econdate	The date of the patient's economic evaluation	
EventDate	The patient's event date	
BMI	The Body Mass Index of the individual	
Alcohol	A binary indicator for drinking habits	1: Drinker, 0: Non-Drinker
Cigar	A binary indicator for smoking habits	1: Smoker, 0: Non-Smoker
YOB	The year of birth of the individual	
Duration	Time in days spent in the study	[0, ???]
Partial_code_pre	Enter code	Always 2000
Partial_code_ff	Transition Code	2000 := Censored, 3000 := Transitioned
.	.	
Age_Pre	Age upon entry to the study	
Age_Post	Age at the EventDate	

As you can see, some variables (such as YOB, duration, Age_Pre, Age_Post) are redundant. When our team began the modeling phase, we did not use every available covariate listed in the above table.

Objectives

Our team set the following objectives to complete by the end of the quarter.

1. Thoroughly explore the dataset for discrepancies or interesting trends
 - Make adjustments as necessary
2. Fit a Cox-PH model to the data
 - Test model assumptions through Schoenfeld Residuals and Log-Log plots
 - Narrow model to the most significant predictors
3. Attempt k-fold cross validation of the decided Cox model
4. If time allows, we chose to pursue the following as well:
 - Non-parametric modeling techniques

Exploratory Analysis

Once our group recieved the dataset, we thuroughly explored the variables using two-way tables and visualizations.

Tabluar Analysis

Table 2. Censorship by Sex

	Male	Female		Male	Female
Transitioned	46296	66871	Transitioned	0.409	0.591
Censored	5082	8416	Censored	0.377	0.623
All	51097	74949	All	0.405	0.595

Table 3. Censorship by Socioeconomic Status

	1	2	3	4	5
Transitioned	0.275	0.231	0.211	0.170	0.113
Censored	0.253	0.220	0.207	0.181	0.139
All	0.272	0.230	0.211	0.172	0.116

Table 4. Censorship by BMI Quartile

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Trasnsitioned	0.2	24.1	26.8	38.66	30.4	66000.0
Censored	0.5	22.3	25.0	39.09	28.4	21920.0
All	12.1	23.9	26.7	27.49	30.2	74.5

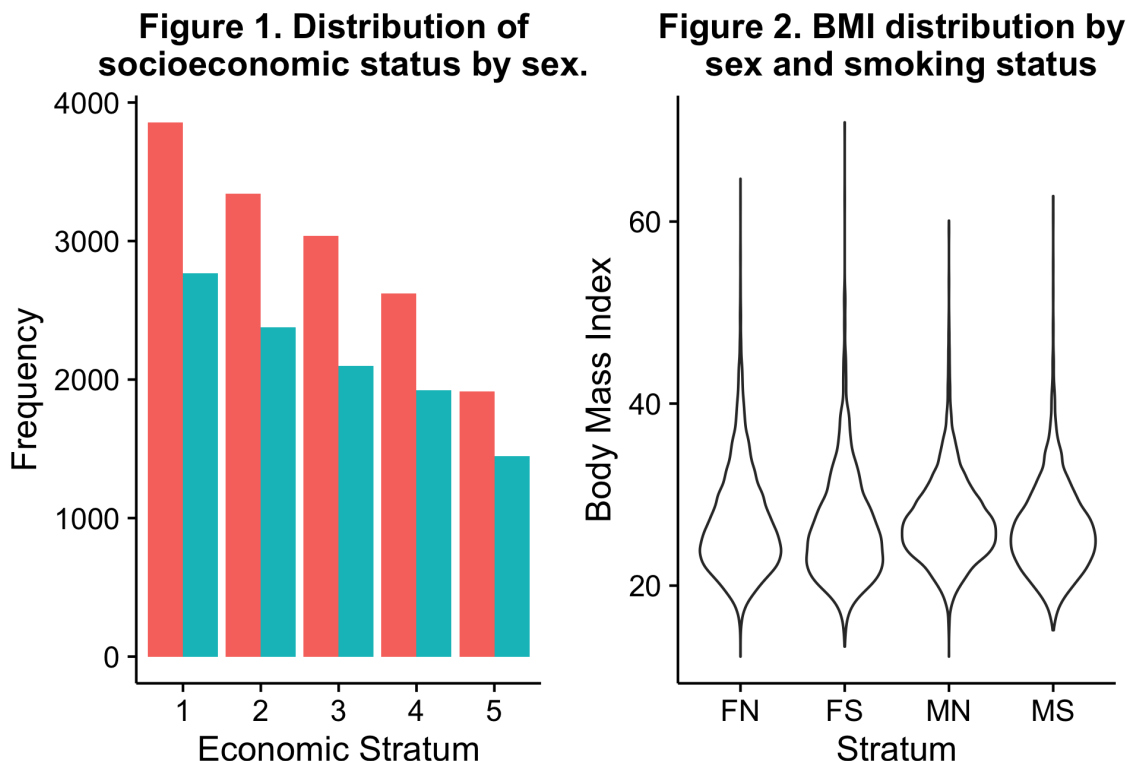
Discrepancies

Table 4 shows some *very* concerning BMI values. The maximum ranges as high as 66000, and some as low as 10. A reasonable range for BMI values is between 18 and 60. We concluded that this was likely the result of an error during data entry or data transfer. Luckily, there were only 326 records that were outside of [12, 100], so we opted to remove those observations which were outside of that range. We decided on 100 as the upper limit because it *is* possible for a human to have a BMI close to 100. Furthermore, our data is concerned with patients that already developed moderate hypertension. So it is absolutely possible that an individual has a BMI between 60 and 100.

Next, we found that there *are* some redundant records for a single individual (multiple occurrences of a single UNQID). This is because some individuals had their economic status re-evaluated during the study. Luckily, this accounted for only a small set of the data: 278 records. Our team decided to keep only the earliest records and drop any duplicate record with the re-evaluated socioeconomic status. Lastly, we found roughly 200 records with negative age values. Professor Ian Duncan, who had worked with the dataset in the past, advised us to remove all observations with either negative age values or survival times equal to zero. We followed his advisement.

Further Analysis

After removing observations with outlandish BMI values, duplicate observations, negative ages, and zero survival time, we continued our exploratory analysis. By this point in the project, our dataset contains 25387 rows.



In Figure 1, red corresponds to women and blue to men. We can see that there are more females in the dataset—roughly 3 women for every 2 men. The distributions of economic status are similar between the sexes. Figure 2 is a violin plot, so a wider horizontal portion corresponds to a more dense concentration of data at

that value. It seems that women's BMI values are distributed more widely, while the men's BMI values are more densely distributed about their respective means.

Modeling

Now that we have explored the dataset, we begin our modeling phase. First, our team constructed simple Kaplan-Meier estimate plots for each stratum of variables concerning alcohol consumption, Body Mass Index, sex, and cigarette consumption. We then moved on to construct two Cox Proportional-Hazards models: one with every covariate available, and a second with only those covariates which satisfied the proportional-hazards assumption. We constructed the reduced model to properly interpret the interactions. We constructed the full model in the hopes that it may provide some predictive power for future cases of severe hypertension. To test the predictive capabilities of the full model, we used cross validation on a single fold. All these steps are discussed in the following sub-sections.

Kaplan Meier Plots

Here we have Kaplan-Meier plots for each of the covariates listed in the section introduction.

Figure 3. KM Estimates by Alcohol Consumption

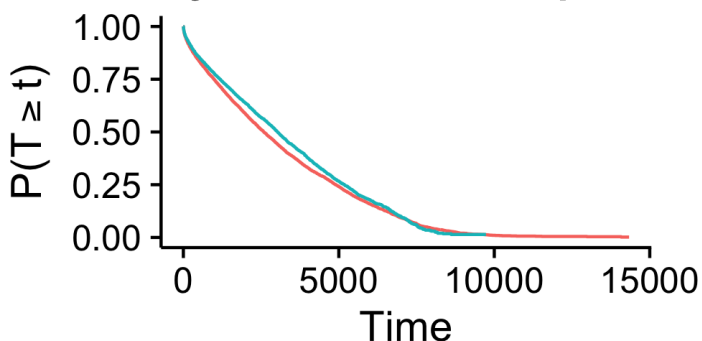


Figure 4. KM Estimates by BMI Level

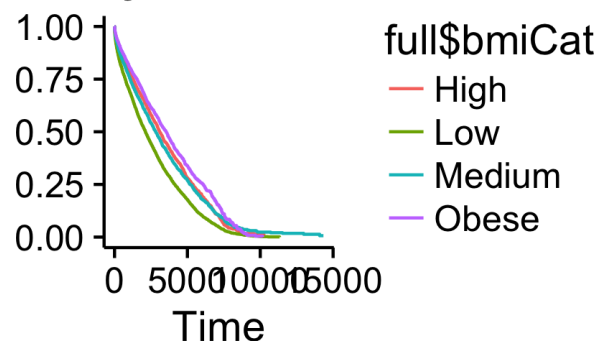
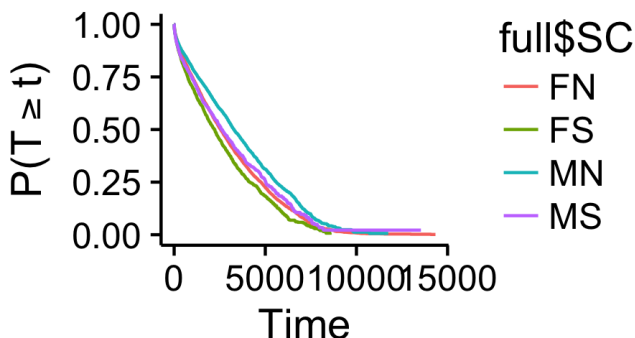


Figure 5. KM Estimates by Sex and Cigarette Usage



Fix these plots. Either one per row, or move legends into the open cell on bottom right (if that's possible).

The factor levels in Figure 4 are broken down by quantile; please see Table 5 shown below. In Figure 5, the first letter corresponds to the sex, and the second letter corresponds to their smoking status (smoker or

nonsmoker).

Needs a title

Quantile Name	BMI Range
Low	[12.1, 23.9)
Medium	[23.9, 26.7)
High	[26.7, 30.2)
Obese	[30.2, 74.5]

Cox Modeling