PSTAT 196 Week 1 Write-Up

Terry Jiang, John Randazzo, Jason Freeberg

4/17/2017

## Exploring Our Data

* We are using the following covariates in our analysis: sex, Socio.Economic, alcohol, cigar, BMInew.
* Our time variable is DURATION
* For each dataset, we define a binary event variable, event = as.numeric(Partial\_code\_ff)
* Our transitioned dataset is 113167x17
* Our censored dataset is 13498x17
* Our full dataset is therefore 126665x17
* We also need to take care of the occasional duplicate measure for Socio.Economic reclassification.

## Further Summary Statistics

Number of men and women in the study, and then their relative proportions:

## Male Female  
## Trasnsitioned 46296 66871  
## Censored 5082 8416  
## All 51378 75287

## Male Female  
## Trasnsitioned 0.4090945 0.5909055  
## Censored 0.3765002 0.6234998  
## All 0.4056211 0.5943789

Socio-economic classification counts for each dataset:

## 1 2 3 4 5 mean  
## Trasnsitioned 31082 26174 23868 19285 12758 2.615285  
## Censored 3416 2968 2794 2442 1878 2.733146  
## All 34498 29142 26662 21727 14636 2.627845

Summary statistics for BMI - These are a bit wonky. We will revisit this table a bit later in the report.

## Min. 1st Qu. Median Mean 3rd Qu. Max.  
## Trasnsitioned 0.2 24.1 26.8 38.66 30.4 66000  
## Censored 0.5 22.3 25.0 39.09 28.4 21920  
## All 0.2 23.9 26.6 38.71 30.2 66000

Percentages of smokers and drinkers in each dataset:

## Smoker Used Alcohol  
## Trasnsitioned 0.1597816 0.3570034  
## Censored 0.2122537 0.3087865  
## All 0.1653732 0.3518652

Summary statistics for DURATION:

## Min. 1st Qu. Median Mean 3rd Qu. Max.  
## Trasnsitioned 0 362 1122 1609 2409 41180  
## Censored 0 330 1305 1881 2894 39830  
## All 0 360 1137 1638 2456 41180

## A Few Words About BMI and Other Data Management Issues

We found some funky BMI values for certain subjects in the study.

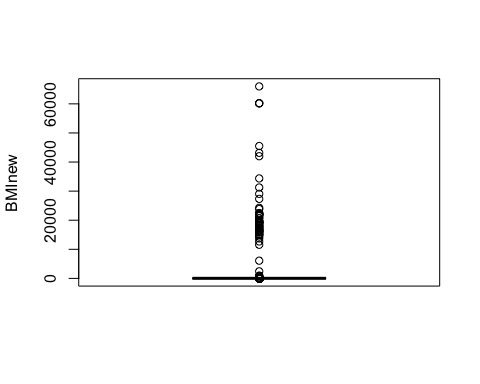
max(full$BMInew)

## [1] 66000

min(full$BMInew)

## [1] 0.2

boxplot(full$BMInew, ylab = "BMInew")



The population for our study consists of people with an existing history of moderate hypertension; a health problem generally associated with an unhealthy body weight. However, we believe that a BMI of 0.2 or 66000.0 is likely the product of clerical data entry error, and inspecting the boxplot confirms that there are many other outlandishly incorrect BMI measures in our dataset. We need to remove these broken values:

full2 <- full  
full2$BMInew <- ifelse(full2$BMInew > 100 | full2$BMInew < 12, NA, full2$BMInew)  
full2 <- na.omit(full2)

After considering the origins of our data and the range of values that a person with moderate hypertension's BMI can take, we decide to omit any BMI measures under 12.00 or above 100.00.

With junk values removed, our new dataset, full2, is now 126339x17.

The new table of summary statistics for each dataset's BMI measures is as such:

## Min. 1st Qu. Median Mean 3rd Qu. Max.  
## Trasnsitioned 12.1 24.1 26.8 27.69 30.4 98.5  
## Censored 12.4 22.3 25.0 25.86 28.4 93.1  
## All 12.1 23.9 26.7 27.50 30.2 98.5

## Exploratory Kaplan-Meier Estimates

Now that we have a good feel for our dataset, we are ready and eager to begin our survival analysis. We compute Kaplan Meier estimates contingent on each covariate's different levels, as well as a general KM estimate for our entire population.

