Untitled

Nathaniel Brown, In Hee Ho, Sarah Zimmermann

October 19, 2017

Introduction

The data are from a study of time to critical neurological assessment for patients with stroke-like symptoms who are admitted to the emergency room. We are interested in the factors predictive of the time to assessment following admission to the ED for n=335 patients with mild to moderate motor impairment. The goal of the analysis is to perform inferences on the impact of clinical presentation, gender, and race (Black, Hispanic, and others) on time to neurological assessment, where clinical presentation is measured as the number of the four major stroke symptoms: headache, loss of motor skills or weakness, trouble talking or understanding, and vision problems. However, as discussed in our previous report, we group Blacks and Hispanics together, and number of symptoms of 3 and 4 together, due to their small sample size.

Methods

The team has cleaned, understood, and modeled these time to critical neurological assessment for patients with stroke-like symptoms data in order to solve the scientific problem of exploring if gender, race/ethnicity, and clinical presentation have an affect on wait list to assessment. To do so the team has approached the problem as such:

1. Data Exploration

- Read in the data
- Explore summary statistics of data
- Visualize the data

2. Create inital models including OLS, Ridge, and LASSO

- Diagnostics
- Results
- Survival Curves
- Interpretation
- Assess sucess

3. Create final models with kernel regression

- Diagnostics
- Results
- Survival Curves
- Interpretation

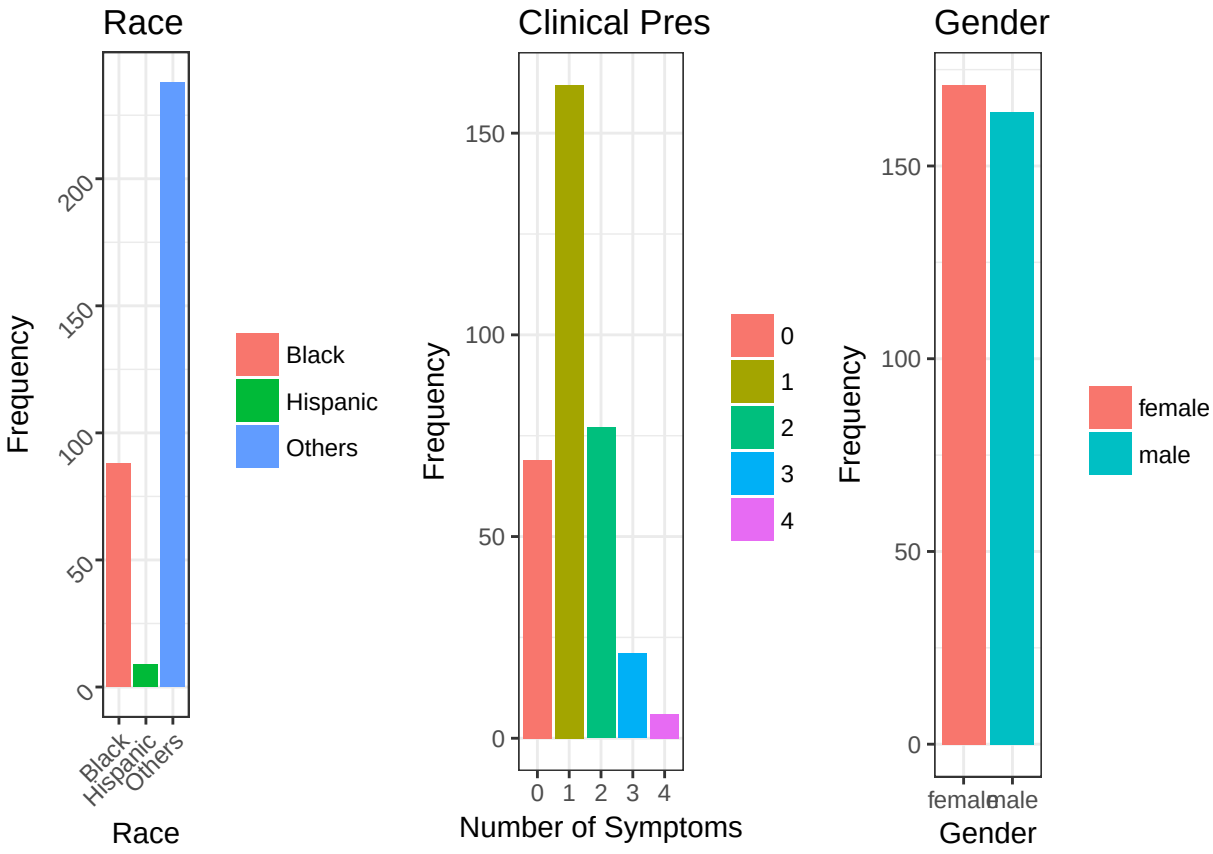
4. Final recommendations and insight

Data Exploration

Variables

The original data set contains 335 observations across 9 variables. They are defined as:

Variable Name	Short Description	Type
nctdel	min of neurologist time to assessment & CT scan from arrival at ER	continous
fail	1 if got neurologist/CT scan & 0 otherwise	categorical
male	1 if male, 0 if female	categorical
black	1 if black, 0 if not black	categorical
hisp	1 if hispanic, 0 if not hispanic	categorical
sn1	0/1 indicator 1 main symptom	categorical
sn2	0/1 indicator 2 main symptoms	categorical
sn3	0/1 indicator 3 main symptoms	categorical
all4	0/1 indicator all main sumptoms	categorical



To visualize the data we have provided the above graphs which show the frequency of the characteristics patients can have. We see most (above 250) of the patients are non-black and non hispanic and less than 100

are either black or hispanic. The most common number of symptoms is 1 then 0, 2, 3, and 4. Finally the gender split between male and females is generally even; however, there are more females in the data then male.

Before moving on it should be noted there are no missing values or apparently out of range values in our data, and therefore we did not clean the data in any way. We did group some of the characteristics of patients because of sample size, however. In the following analysis we grouped black and hispanics in one category for race/ethnicity and non black and non hispanics in the other. This decision was based on the fact there are only 9 hispanics in the data set which is too small of a population to make meaningful conclusions. Next, symptom is a 4 level variable: "0" for those who had no symptoms, "1" for those who had 1 symptom, "2" for those who had 2 symptoms, and "3+" for those who had 3 or more symptoms. We grouped those patients which 3 or 4 symptoms together because there were only 6 individuals out of 335 observations in the dataset who had 4 symptoms, which is very small a population size to draw any conclusions on.

We now continue on in our report to further understand the data specifically the exploratory data analysis we found useful when later we created data models.

Exploratory Data Analysis

PUT USEFUL EDA HERE

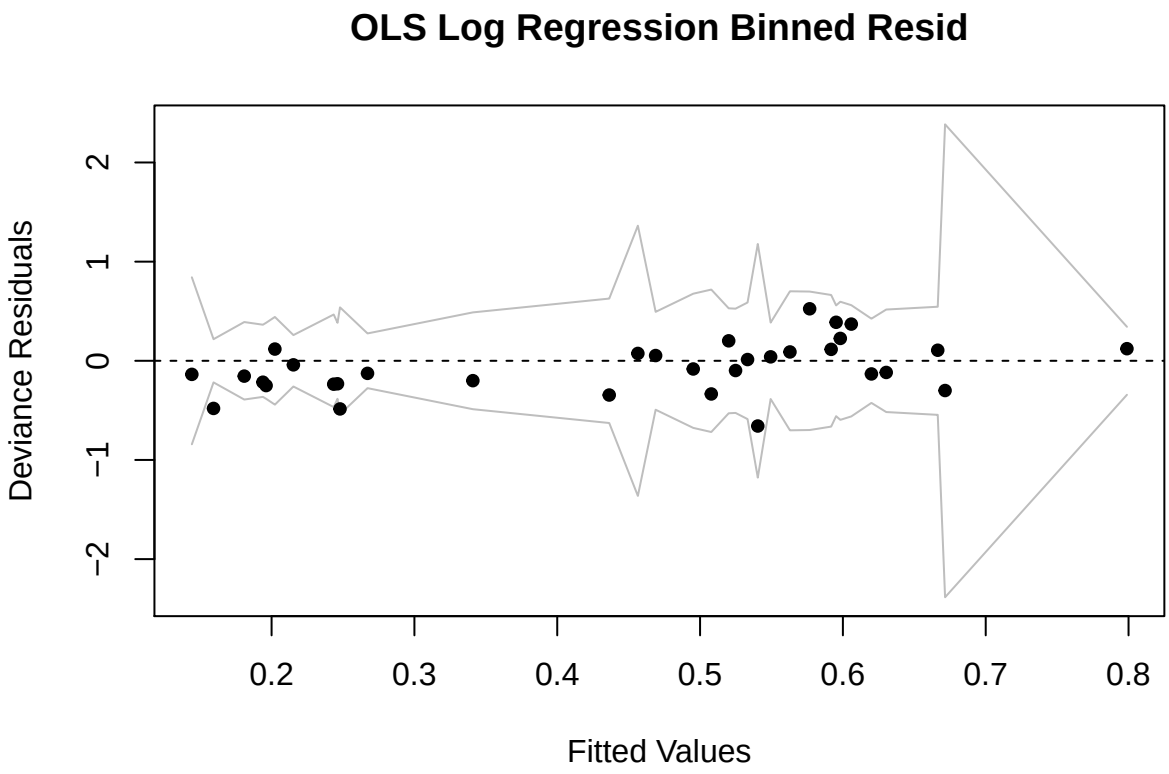
Initial Model Exploration

Now that we have visualized each variable along with the relationship between the variables we will continue on to modeling our data.

To perform logistic regression on the data, we categorize the time-to-event variable into groups of 1 minute, with all events occuring after 5 minutes grouped together, since the sample size is low after 5 minutes, with only 9 observations. Our predictors consist of these time categories, as well as the aforementioned categories of race, gender, and clinical presentation. The binned residual plots, deviance test results, and coefficients are reported below. We attempt three approaches to logistic regresion: Ordinary Least Squares (OLS), LASSO, and Ridge. The glmnet package does not provide standard errors for its coefficients, so we cannot report confidence intervals for the estimates.We choose to move forward with the OLS model because of the fit of the model to the data. See exploration below.

Diagnostics

We check the diagnostics to see if the model properly fits the data. We see the residuals are spread evenly above and below 0 and have no apparent pattern. Overall, the logistic models require normality and independence and from the residuals we can generally say the model meets these assumptions.



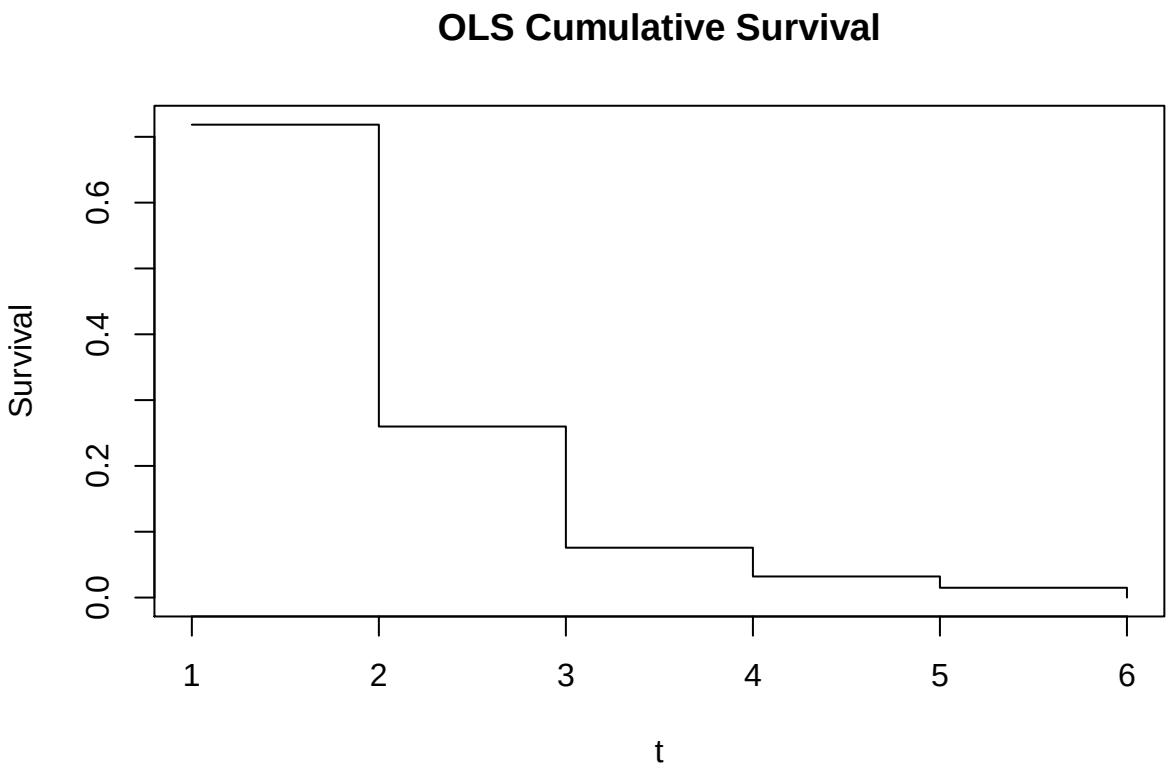
Results

Below are the results of the OLS.

The OLS does not find much difference between these groups; however, in the OLS model, the only coefficients that do not contain zero in their 95% confidence interval are X1 and X3. We cannot confidently claim that any of these factors (gender, race, clinical presentation) are predictive of the time to assessment.

	Lower	Upper
symptom0	-1.2348	0.1283
symptom1	-0.8128	0.4192
symptom2	-0.9683	0.3673
raceother	-0.2452	0.4814
male	-0.6261	0.0439
X1	-1.6083	-0.2653
X2	-0.1101	1.2464
X3	0.1159	1.6606
X4	-0.6474	1.2667
X5	-1.0553	1.3369
X6	-926.4905	958.4814

Survival Curves



CREATE INDIVIDUAL SURVIVAL CURVES HERE

Interpretation

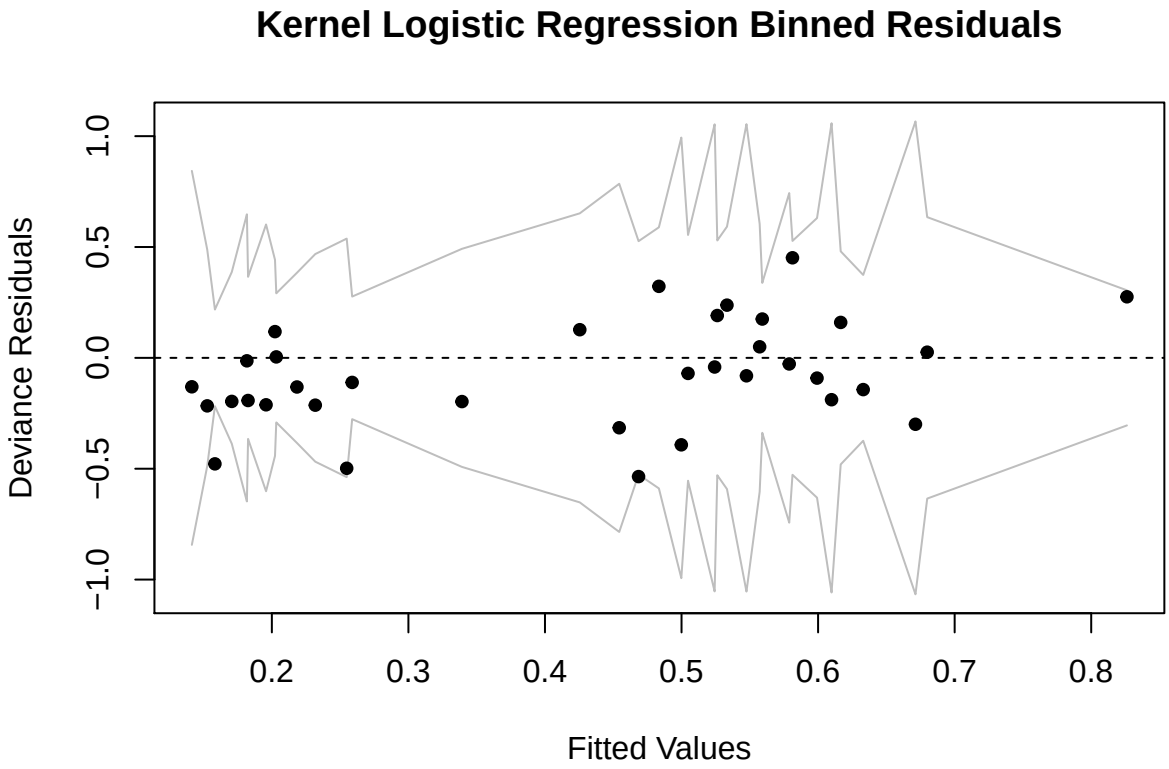
The models we build for this analysis do not fit the data well. This can be due to a relatively small sample size of 335 patients, or the possibility that there is no measurable difference between races, genders, and clinical presentation in time to treatment. In future analysis, we will attempt to fit more flexible models, such as generalized additive models with kernel smoothing.

Final Model Exploration

In order to better model the data, the group moved forward with a more flexible model: kernel logistic regression. A kernel model can possibly better fit the data because kernel regression create smaller divisions or bins of the data and then fit the data within each bin. This new approach might better model the data as we hope to see some of the coefficients not close to 0 therefore suggesting that clinical presentation, gender and race have an influence on wait time.

Diagnostics

As in any model we first check the diagnostics of the model to see if it properly fits the data. In kernel logistic regression we check if the linearity and normality assumptions are met. Since we see no general pattern in the residuals and an even spread above and below 0 we believe we can move forward with the model.



Results

The coefficients of this model suggest wait time changes with number of symptoms, race, and gender; however, since the coefficients are small and close to 0 they do not suggest that much influence on wait time. Nevertheless, the more the symtoms the shorter the wait time, if the patient is non black or non hispanic the wait time will be shorter, and finally if the patient is male the wait time is expected to be shorter as well.

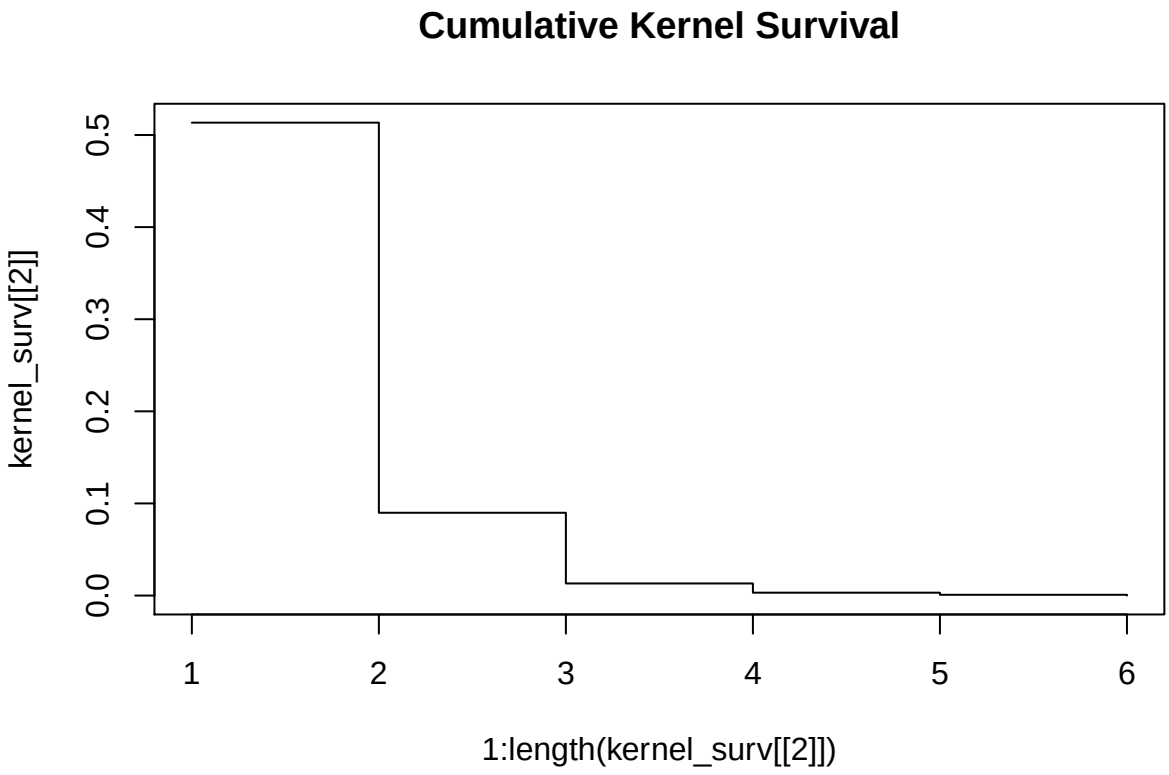
	Lower	Upper
symptom0	-1.288600e+00	9.280000e-02
symptom1	-9.404000e-01	3.568000e-01
symptom2	-1.304200e+00	2.908000e-01
raceother	-2.377000e-01	5.033000e-01
male	-6.397000e-01	4.070000e-02
k1	-8.400077e+12	2.702965e+13
k2	-2.002199e+14	6.222287e+13
k3	-6.222326e+13	2.002211e+14
k4	-2.709698e+13	8.421002e+12
k5	-1.542361e+11	4.962986e+11
k6	-1.256005e+09	4.080909e+08

	Lower	Upper
k7	-2.761008e+07	2.789802e+07

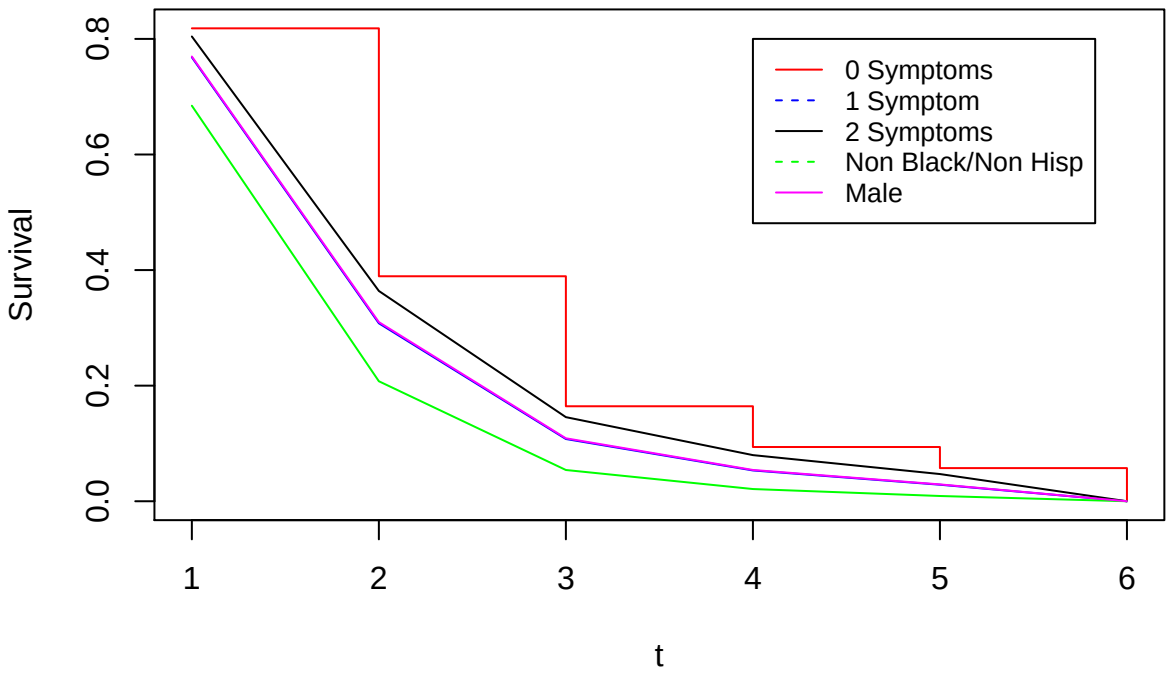
We will now visualize these results by creating a survival curves for the total population and survival curves for patients with 0/1/2/3+ symptoms, male/female, and black+hispanic/non-black+ non-hispanic.

Survival Curves

Below is the survival curve for the entire population despite race, gender, and clinical presentation. We will compare the population survival curve to the survival curves based on gender, race, and clinical presentation. If there is a difference between the total population survival curve and the other survival curves this is a way to tell if gender, race, or clinical presentation have an influence on wait time.



Here we see the survival curves based on 0 symptoms, 1 symptom, 2 symptoms, gender, and race.



Interpretation

Discussion

why nothing is significant:

```
## # A tibble: 4 x 6
##   symptom     mean     n        sd    lower    upper
##     <chr>    <dbl> <int>     <dbl>    <dbl>    <dbl>
## 1       0 1.560370    45 0.8675425 1.306892 1.813849
## 2       1 1.547995   133 0.7804779 1.415350 1.680640
## 3       2 1.618750    56 0.7784150 1.414871 1.822629
## 4      3+ 1.493333    25 0.6746227 1.228881 1.757785
```

```
## # A tibble: 2 x 3
##   gender     mean   median
##    <chr>    <dbl>    <dbl>
## 1 female 1.516541 1.433333
## 2   male 1.606217 1.566667
```

```
## # A tibble: 2 x 3
##                race     mean   median
##               <chr>    <dbl>    <dbl>
## 1 Black or Hispanic 1.727556 1.716667
## 2             Other 1.491938 1.383333
```

References

https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

http://influentialpoints.com/Training/coxs_proportional_hazards_regression_model-principles-properties-assumptions.htm#modmch

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/

http://dwoll.de/rexrepos/posts/survivalKM.html

Credits

Old survival curves.. not sure if we need these # {r} # plot(survfit(Surv(timecat, fail) ~ raceother + male, data = datcat_X), # main=expression(paste("Kaplan-Meier Estimate ", hat(S)(t), " with CI")),xlab="t", ylab="Survival", lwd=2) # # # {r} # plot(survfit(Surv(timecat, fail) ~ raceother + male, data=datcat_X) , xlab="Survival Time", # ylab="% Surviving", yscale=100, col=c("red","blue", "black", "green"), # main="Survival Distributions") # legend("topright", title="Legend", c("Black/Hisp", "Non Black/Hisp" ,"Male", "Female"), # fill=c("red", "blue", "black", "green")) # # survdiff(Surv(timecat, fail) ~ raceother + male, data=datcat_X) #

Mice stuff

```
{r} # data <- read.table("kellydat.txt", header=T) # data$race
= 0 # data$race[data$black==1|data$hisp==1] = 1 # data$sn0 = 0
# data$sn0[data$sn1==0 & data$sn2==0 & data$sn3==0 & data$all4==0]
= 1 #  # data.imp = data # data.imp$nctdel[data.imp$fail == 0]
= NA #  # #md.pattern(data.imp[!is.na(data.imp$nctdel),]) #  #
tempData <- mice(data.imp,m=5,maxit=50,meth='pmm',seed=500,
print=FALSE) # # methods:  # # 2l.norm / 2l.pan / 2lonly.mean
/ 2lonly.norm / 2lonly.pmm /  # # cart / fastpmm / lda / logreg
/ logreg.boot / mean / midastouch / norm / norm.boot / norm.nob
/  # # norm.predict / passive / pmm / polr / polyreg / quadratic
/ rf / ri / sample #  # #tempData$imp$nctdel #  # data.imp <-
complete(tempData) #  # hist(log(data.imp$nctdel + 0.1)) #  #
fit <-lm(log(nctdel+0.1) ~ sn0 + sn1 + sn2 + race + male, data
= data) #
```

##Diagnostic

```
{r} # library(VIM) # library(mice) #  #  # ##looking for pattern
of missing data # md.pattern(data.imp) #  # #visualizations:
# aggr_plot <- aggr(data, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram
of missing data","Pattern")) #  # marginplot(data.imp[c(1,2)])
# marginplot(data[c(1,3)]) # marginplot(data[c(1,4)]) # marginplot(data[
# marginplot(data[c(1,6)]) # marginplot(data[c(1,7)]) # marginplot(data[
# marginplot(data[c(1,9)]) # marginplot(data[c(1,10)]) #  #  #
# #visualize distribution of original and imputed data= check
```