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

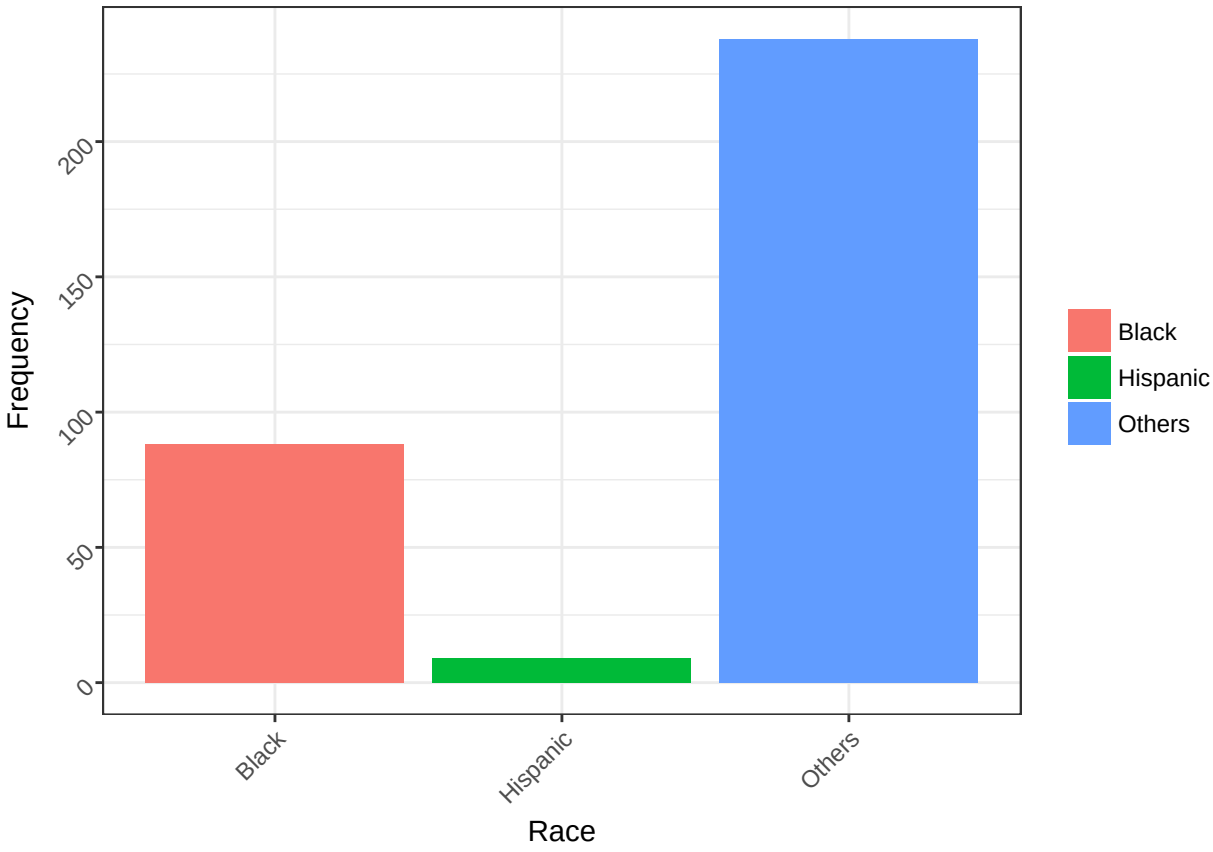
Data Exploration

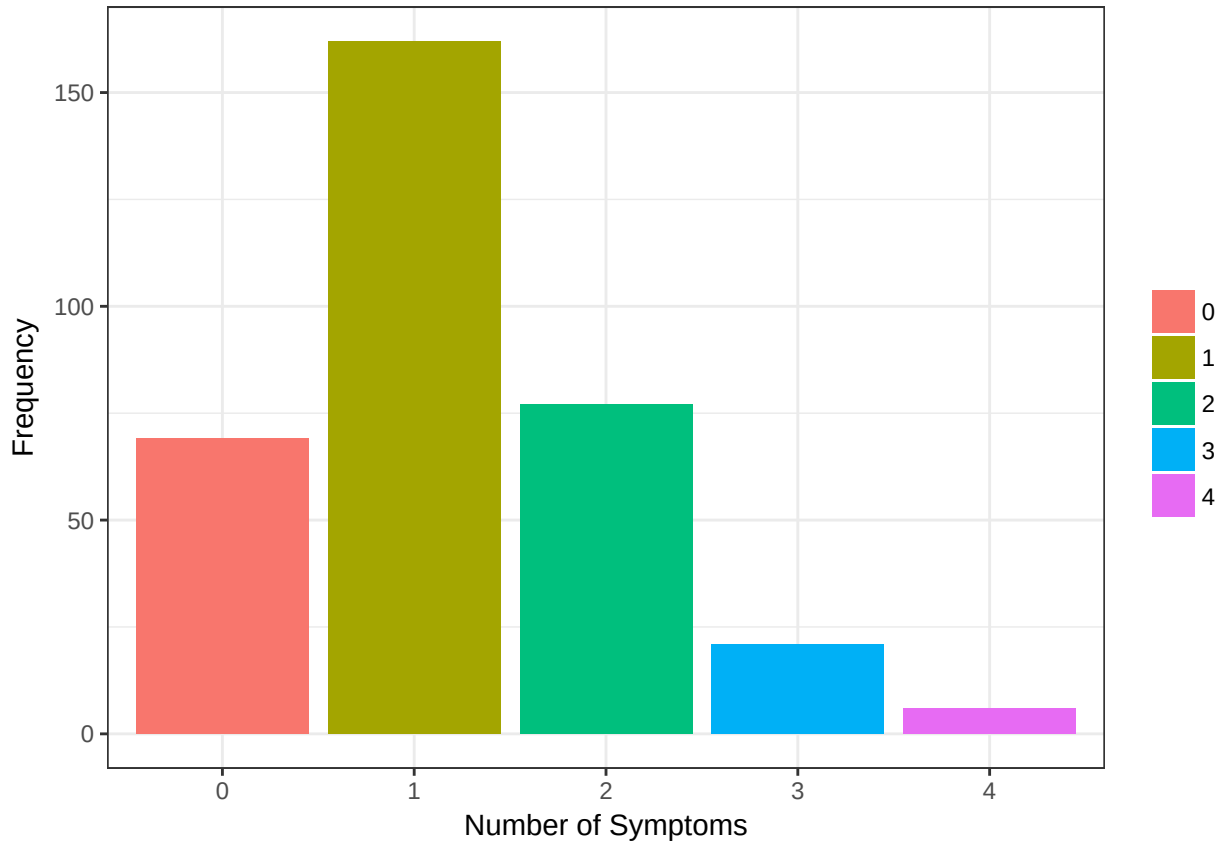
Variables

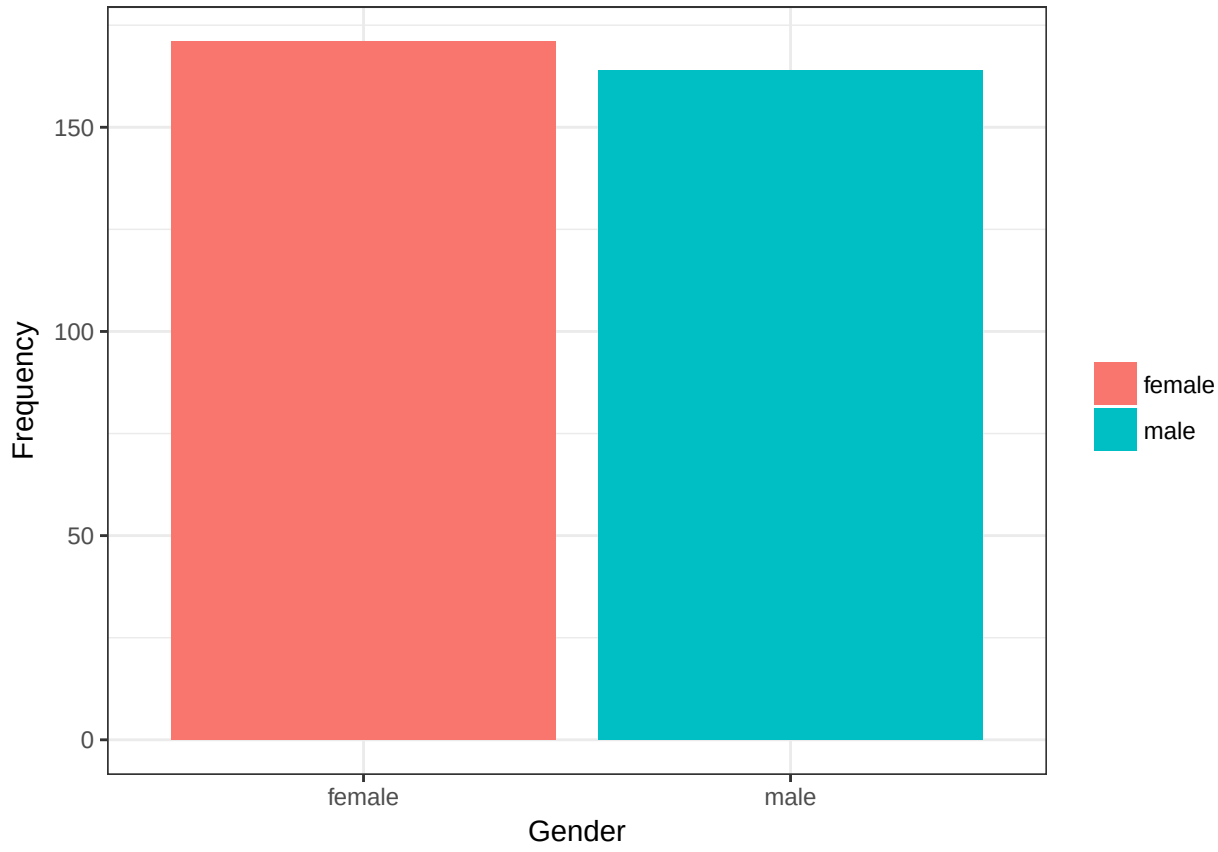
The original data set contains 335 observations across 9 variables. They are defined as:

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

There are no missing values or apparently out of range values in our data, and therefore we had no need to clean the data. However, we created variables to better visualize the data. Race is a 2-level variable (Black/Hispanic and Other) which represents all those in the dataset who are Black or Hispanic in one category and non-Black and non-Hispanic in the other. We decided to group Black and Hispanic together because Blacks and Hispanics had relatively small sample sizes, constituting 97 out of 335 patients observed. Particularly, there were only 9 Hispanics, which is a population too small to represent the group and therefore difficult to draw conclusions upon. 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.


Frequency
200
150
100
50
0
Black
Hispanic
Others
Race
Black
Hispanic
Others

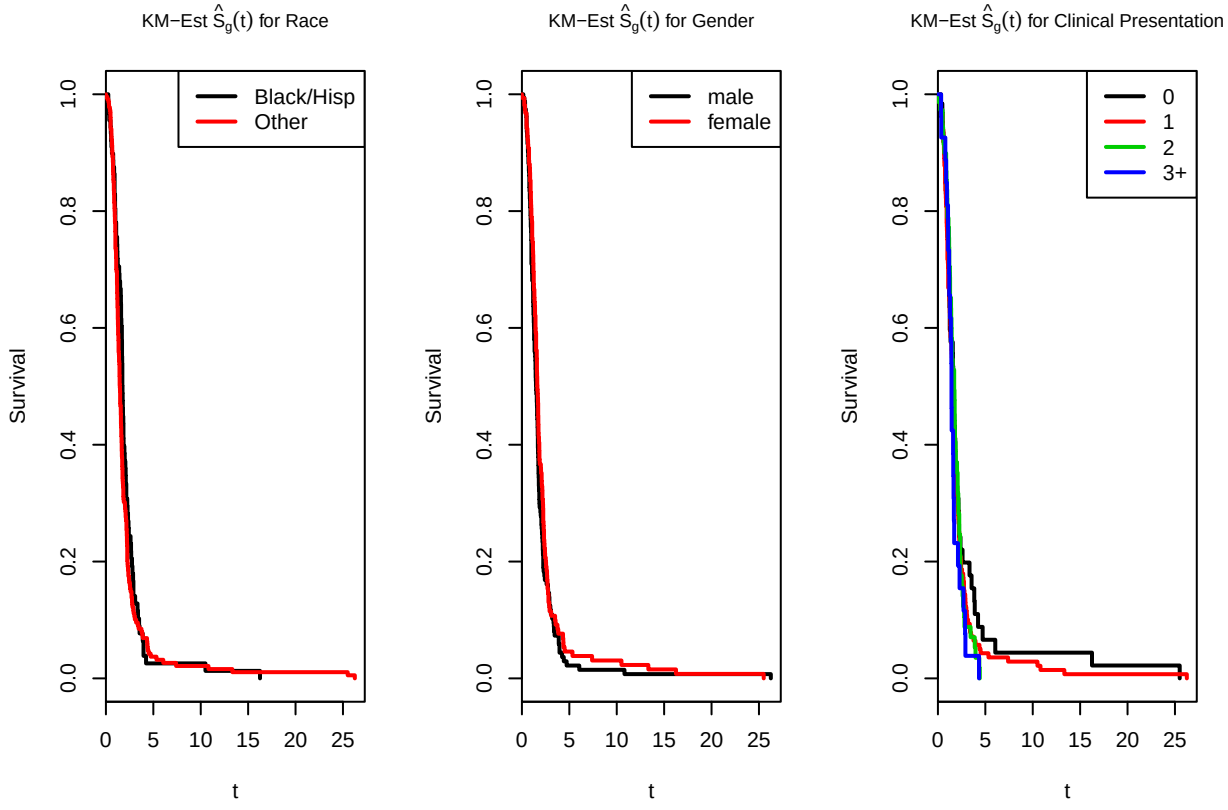

150
100
50
0
Frequency
0
1
2
3
4
Number of Symptoms
0
1
2
3
4



Exploratory Data Analysis

We first look at the overall survival curve. The curve below shows the time the entire population - regardless of race, gender, or clinical condition - will wait for the examination, as well as its 95% confidence interval.

We now look at the separate estimated survival curves for different groups. From our Kaplan-Meier estimates for three different variables (race, gender, and the number of major stroke symptoms a patient shows), we observe that the survival curves are generally similar between 0-5 minute time range. However, thereappears a difference between the estimated survival curves for different number of symptoms a patient shows; specifically, people showing more symptoms tend to wait a shorter amount of time for their treatment. Overall, the survival curves are proportional, which is a key assumption to fit a Cox Proportional Hazard model.

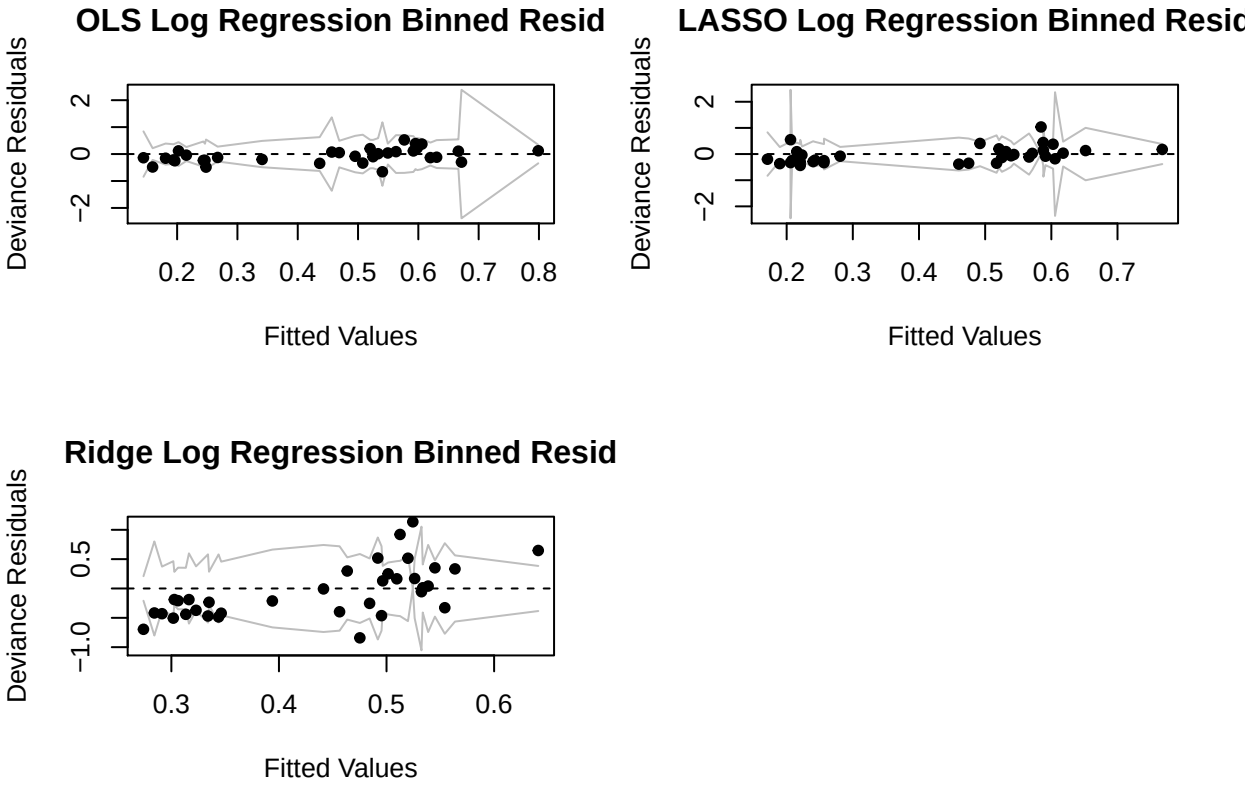


Initial Model Exploration

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.

Diagnostics

We check the diagnostics to see if the model properly fits the data.Here we see a no pattern to any of he residuals and the points generally spread evenly above and below 0. We continue with our analysis because the assumptions of indpendence and normality are generally met.



Results

Below are the results of the OLS, LASSO Penalty and Ridge Penalty respectively. The output shows the coefficients for

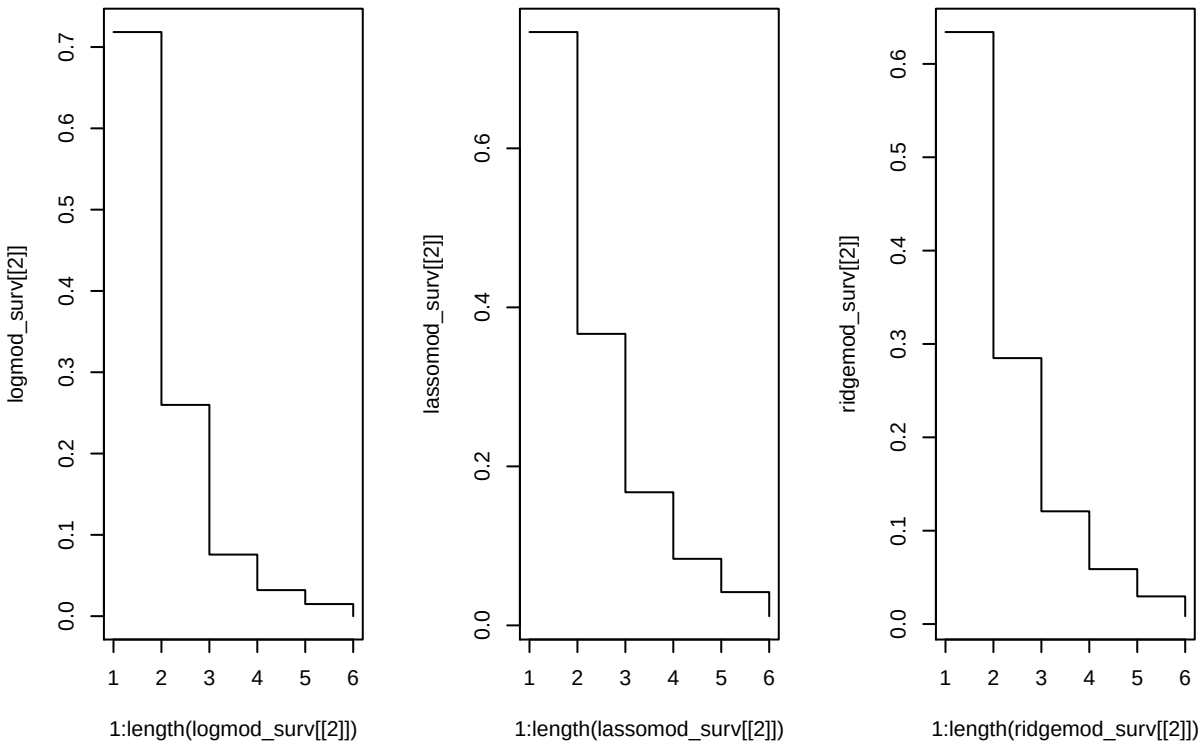
	Deviance p-value
OLS	0
LASSO Penalty	0
Ridge Penalty	0

	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

	LASSO Estimate
(Intercept)	0.0000
symptom0	0.0000
symptom1	0.0000
symptom2	0.0000
raceother	0.0000
male	0.0000
X1	-1.0788
X2	0.0347
X3	0.1736
X4	0.0000
X5	0.0000
X6	0.9557

	Ridge Estimate
(Intercept)	0.0000
symptom0	-0.1646
symptom1	-0.0401
symptom2	-0.0893
raceother	-0.0588
male	-0.1393
X1	-0.5499
X2	0.2039
X3	0.3068
X4	0.0513
X5	-0.0139
X6	0.9175

Survival Curves



Interpretation

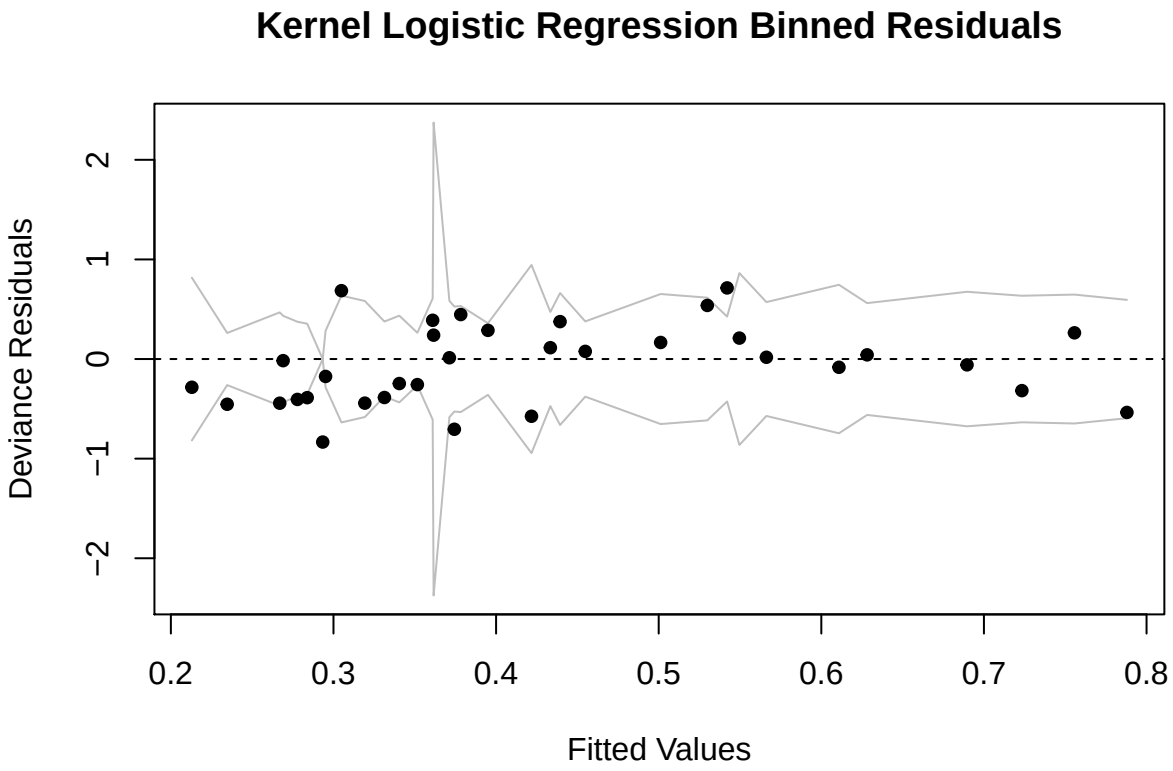
All three of our logistic regression approaches-OLS, LASSO, and Ridge-do not find much difference between these groups. In the OLS model, the only coefficients that do not contain zero in their 95% confidence interval are X1 and X3. In the LASSO and Ridge models, the majority of coefficients are close to zero or exactly zero, and all of the models fit the data poorly according to the deviance test. The residual plots also suggest a poor fit for all the models other than the OLS model. Therefore, we cannot confidently claim that any of these factors are predictive of the time to assessment.

Next Steps

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

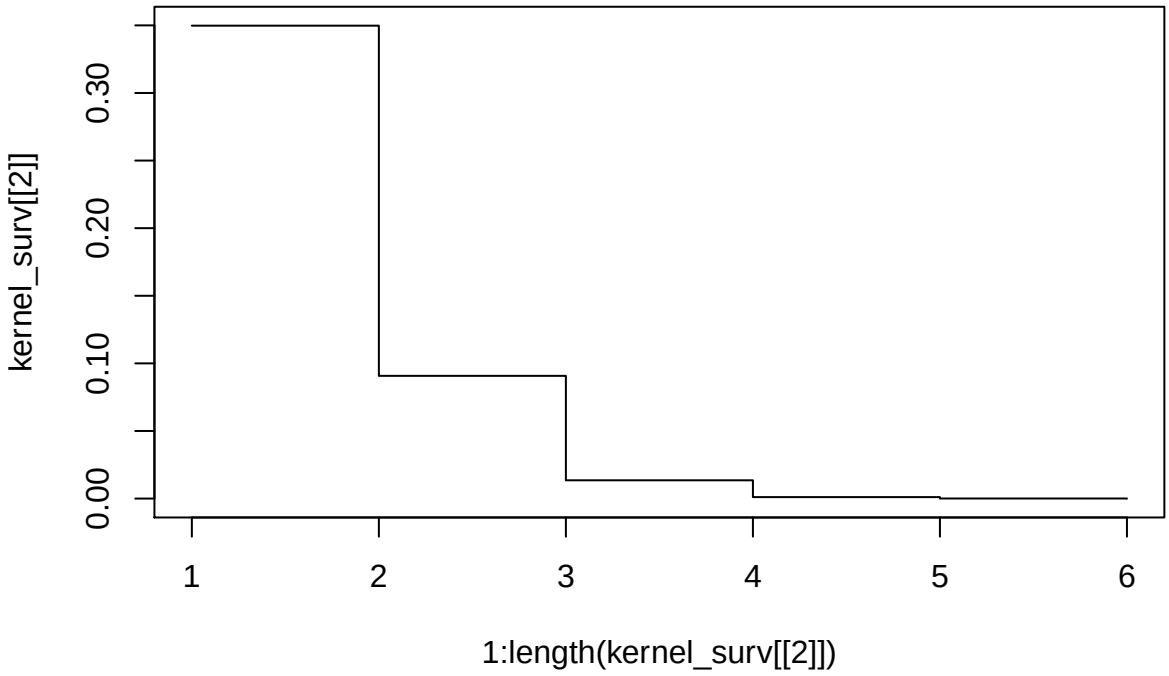
describe kernel regression

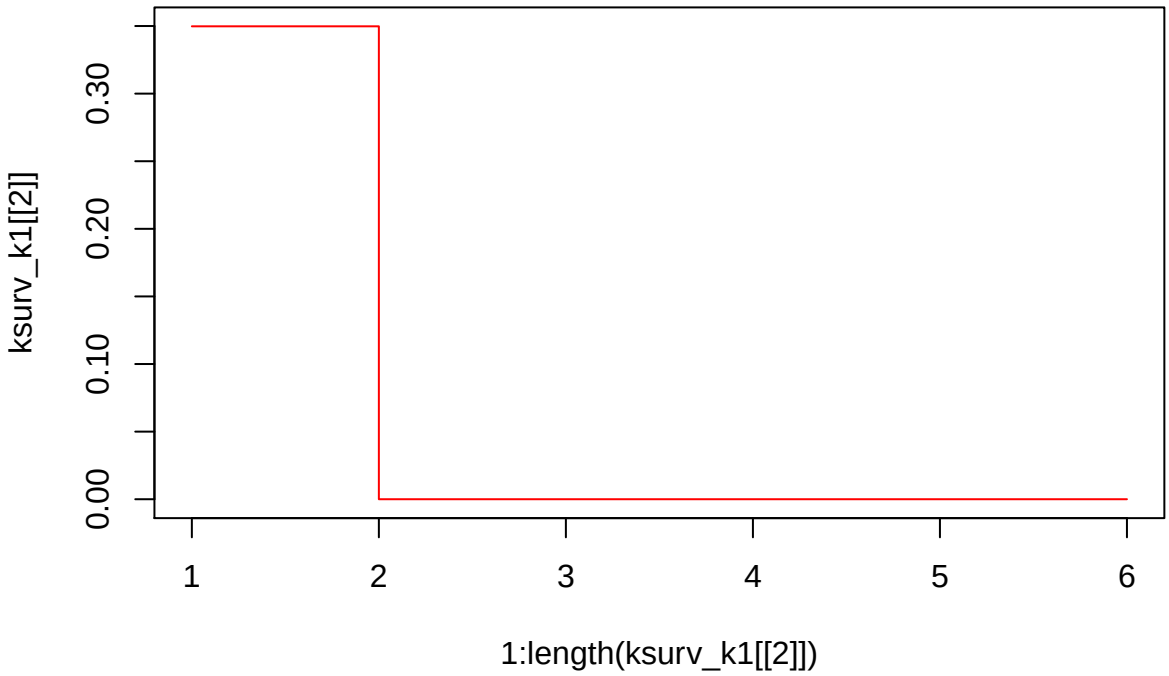


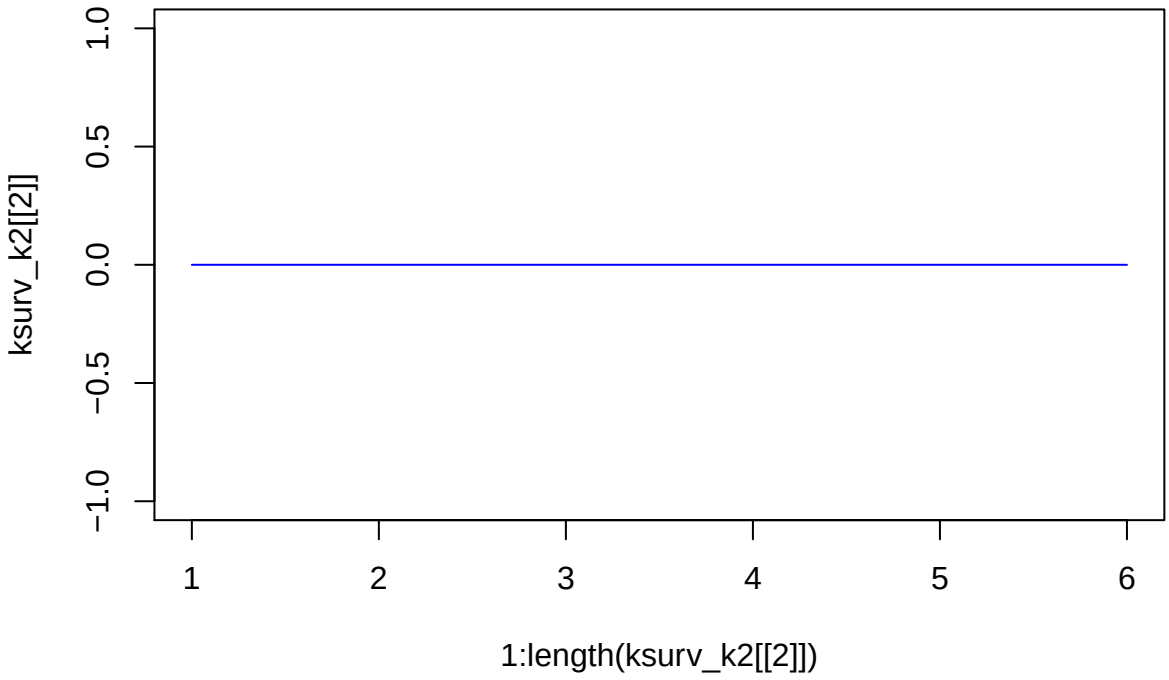
Results

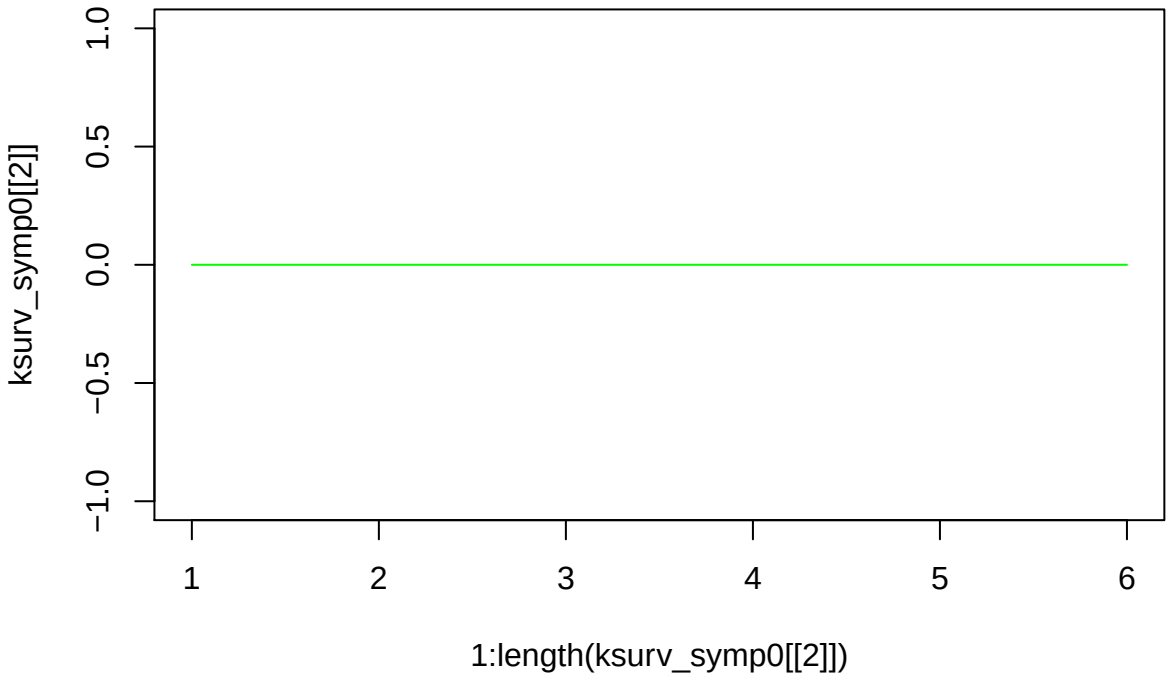
	Lower	Upper
symptom0	-1.3827	-0.0953
symptom1	-0.9360	0.2222
symptom2	-1.0734	0.1903
raceother	-0.2915	0.4013
male	-0.5736	0.0674
k1	-5.6259	-0.1256
k2	5.8663	13.2598

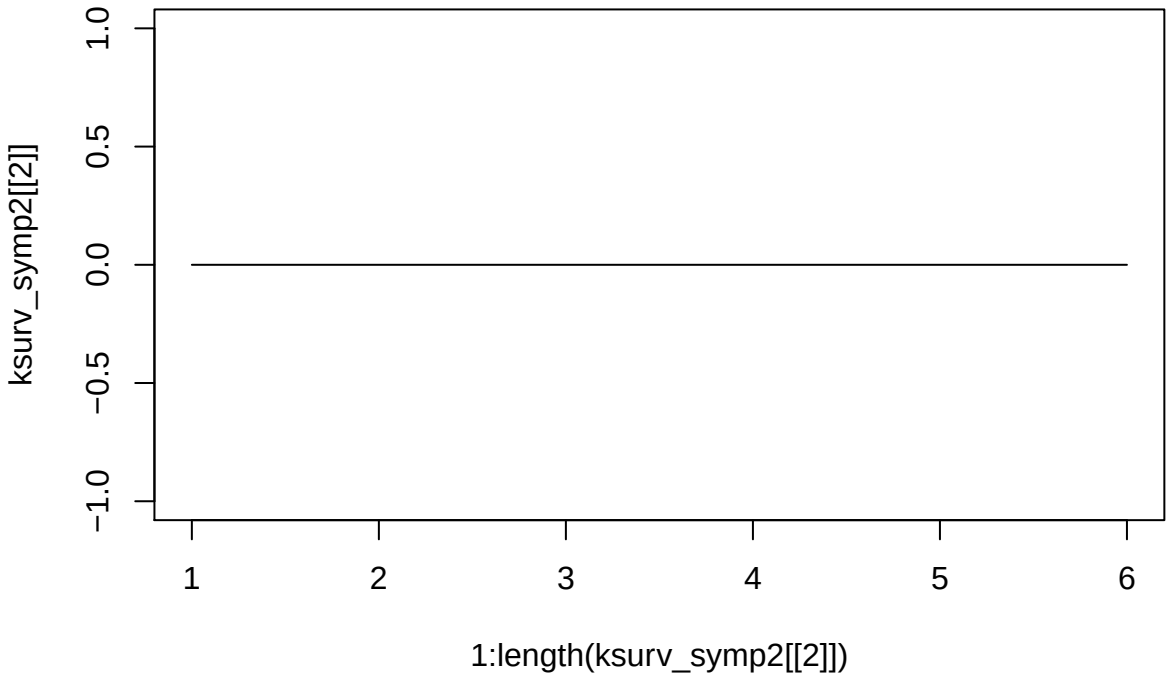
Survival Curves

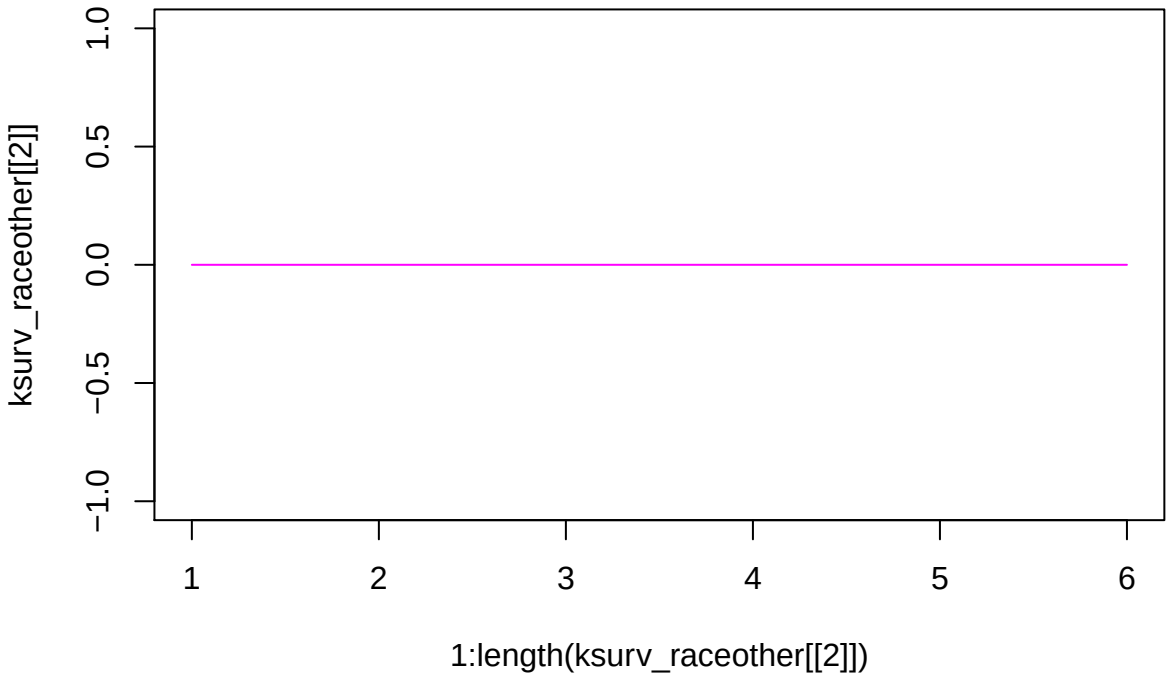


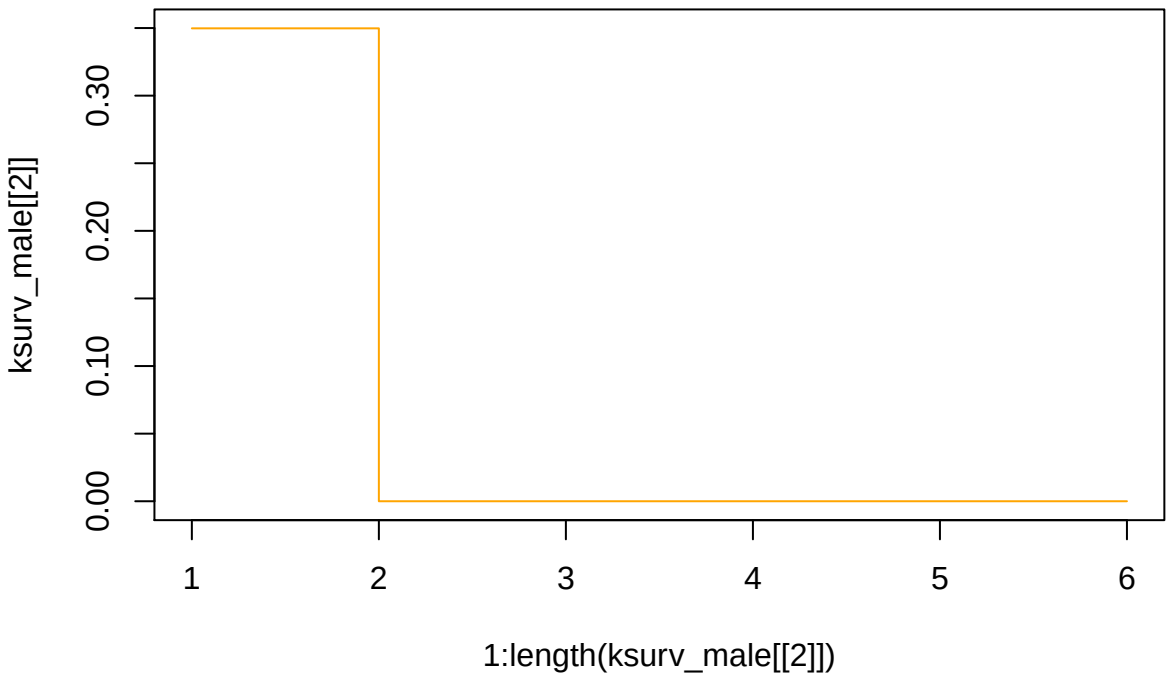



ksurv_k2[[2]]
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
1:length(ksurv_k2[[2]])


ksurv_symp0[[2]]
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
1:length(ksurv_symp0[[2]])


ksurv_symp2[[2]]
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
1:length(ksurv_symp2[[2]])


ksurv_raceother[[2]]
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
1:length(ksurv_raceother[[2]])



###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
```