STA 440 Case 2

Annie Lott, Wuming Zhang, William Yang

October 18, 2017

Goals and Data Description

Getting treatment quickly after a stroke is crucial to a positive long-term prognosis. For emergency room patients who may have had a stroke, we are exploring the variables that influence the time it takes for these patients to receive a neurological assessment. The time from getting to the emergency room to receiving a neurological assessment, such as a CT scan, factors into the total time it takes for the patient to get treated for a stroke. This has a direct impact on the stroke patient's subsequent neurological health and survival.

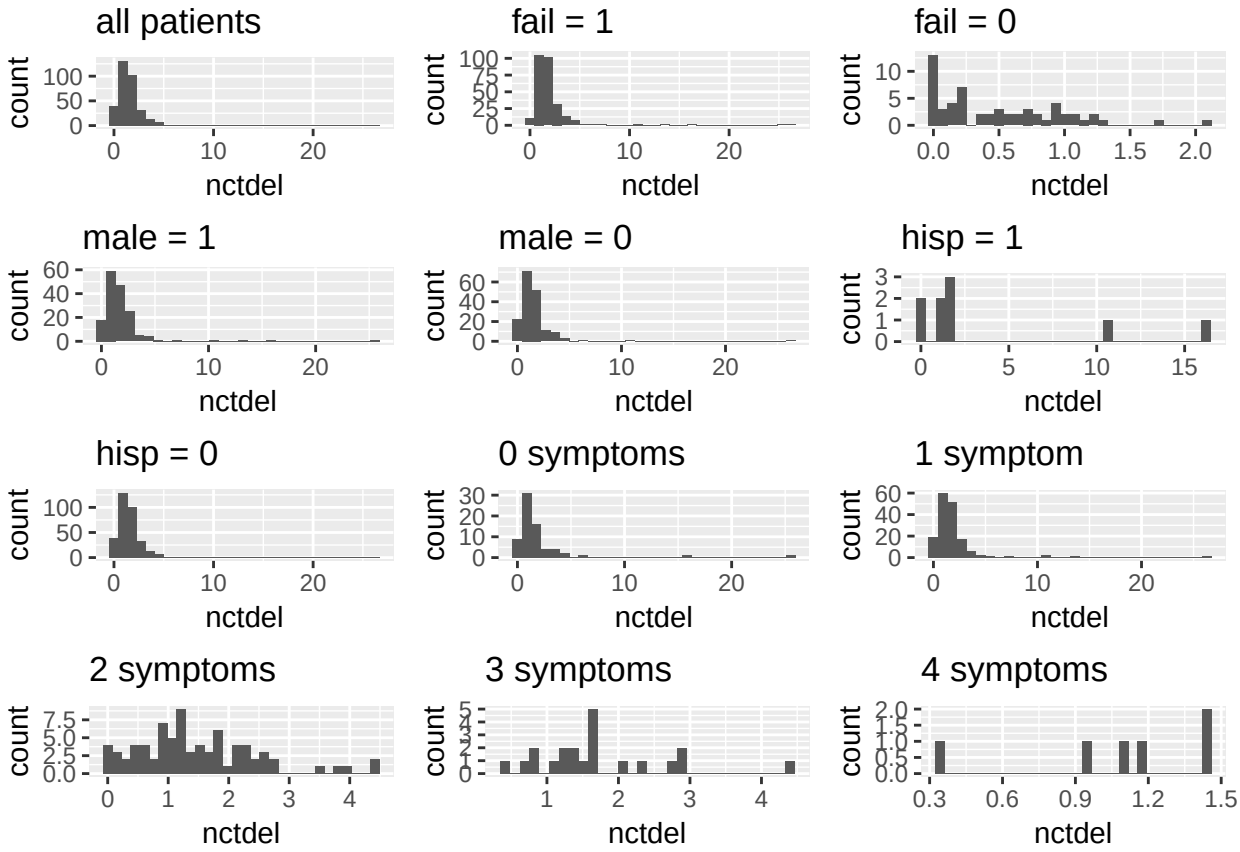
Based on data from 335 emergency room patients with mild to moderate motor impairment, possibly indicative of a stroke, we are analyzing whether sex, race, ethnicity, and the number of symptoms displayed affects the time until the patients receive a neurological assessment, or whether or not they receive a neurological assessment at all. Sex is given as a binary variable of whether the patient is male (1) or female (0), race is a binary variable of whether the patient is black (1) or not (0), ethnicity is a binary variable of whether the patient is hispanic (1) or not (0), and the number of symtoms ranges from zero to four, with binary variables for if the patient exhibited one, two, three, or four symptoms (1 for each number of symptoms if they did exhibit this number, 0 for each number if they didn't). The four possible symptoms include a headache, loss of motor skills or weakness, trouble talking or understanding, and vision problems. Thus, given these variables, our goal is to build a model that predicts the amount of time until neurological assessment for potential stroke patients based on clinical presentation (how many symptoms the patient seems to have), gender and race and ethnicity, and to perform inference on the impact of these variables based on the model.

Data Cleaning

For ease of analysis, we cleaned the data by introducing a numerical variable for the number of symptoms for each patient- zero, one, two, three, or four- rather than using binary variables for whether a particular number of symptoms were displayed or not. There were no missing data to recode for our analysis.

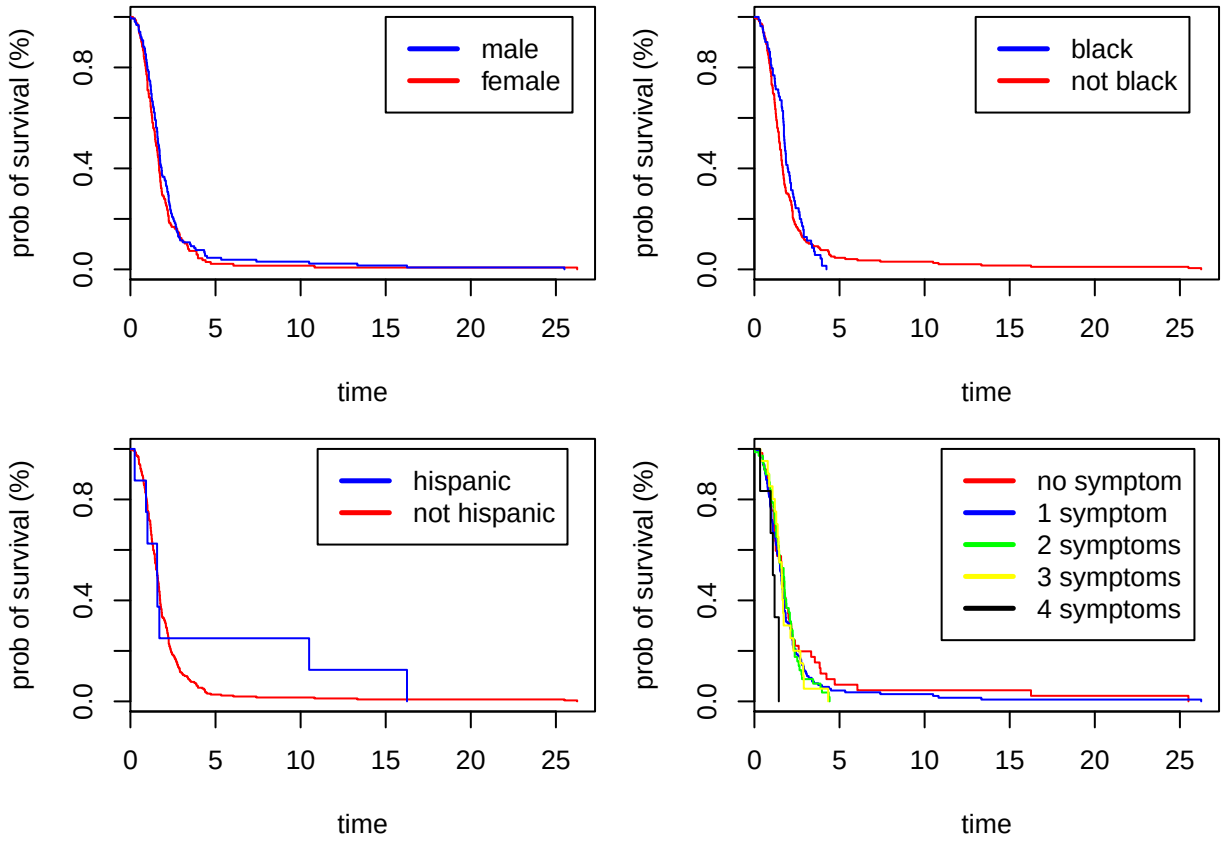
Exploratory Data Analysis

Histograms of each Indicator



To get a better sense of the distributions in the data, we first plot histograms of the nctdel times for each level of all the variables. We notice some general trends here - nctdel time seems to go down with more symptoms, the overall distribution of times for all patients is right-skewed with a mean around 1.5. However, we note that there are only 9 data points for hispanic patients, and only 6 for patients with all 4 symptoms. Inference around these variables will therefore have high variance.

Kaplan-Meier Analysis



Note that in our case, survival means that the patient still keeps waiting. The survival curves for males versus females look almost the same, suggesting that there is no significant difference in their chance of waiting at varied time. The survival curves for black patients versus non-black patients look a little different, with the probability of still waiting dropping off slightly more quickly for non-black patients as waiting time increases, but the difference is not substantial. There is a big difference between the survival curves for hispanic patients versus non-hispanic patients, but this difference may arise because there is so little data on hispanic patients. There seems to be large differences in the curves for the number of symptoms displayed, with most people having four symptoms being seen before about 2 time units, while people with no symptoms had a much less dramatic drop-off in the probability of still waiting, meaning that they were generally seen later.

Modeling Approaches

Cox Proportional Hazards Model

One modeling approach we can take is fitting a Cox proportional hazards model for survival time to the dataset. We can use this to measure the effect that each level of the predictor variables has on the time to neurological assessment.

```
## Call:
## coxph(formula = d3 ~ black + male + hisp + symptom, data = kellydat)
##
##   n= 335, number of events= 277
```

```
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## black   -0.14891   0.86164  0.13949 -1.068    0.286
## male    -0.14270   0.86701  0.12137 -1.176    0.240
## hisp    -0.29726   0.74285  0.37114 -0.801    0.423
## symptom  0.10860   1.11472  0.06959  1.561    0.119
##
##         exp(coef) exp(-coef) lower .95 upper .95
## black      0.8616     1.1606    0.6555     1.133
## male       0.8670     1.1534    0.6835     1.100
## hisp       0.7428     1.3462    0.3589     1.538
## symptom    1.1147     0.8971    0.9726     1.278
##
## Concordance= 0.536  (se = 0.021 )
## Rsquare= 0.017   (max possible= 1 )
## Likelihood ratio test= 5.81  on 4 df,   p=0.2137
## Wald test            = 5.75  on 4 df,   p=0.2187
## Score (logrank) test = 5.77  on 4 df,   p=0.2167
```

An initial implementation of the model on all the predictors does not identify any of them as having a significant effect. We will use other modeling approaches, generalized linear modeling with and without kernel regression, to verify if indeed the variables of clinical presentation, race, ethnicity, and gender have no impact on the time until clinical assessment for patients in the ER who are displaying symptoms potentially indicative of a stroke.

Generalized Linear Modeling without Kernel Regression

Generalized linear modeling is a form of inference that abstracts linear regression so that the error distribution of the response variable does not have to be normal. In a generalized linear model, the linear predictor for the explanatory variables is related to the expected value of the response variable through the link function, which in this case is the log odds. We use the log odds as the link function because we are implementing logistic regression through the generalized linear model paradigm. The only relevant assumption made for our generalized linear model is the the wait times for patients are independent of each other, which we are comfortable making. We do not need to assume that the residuals are normally distributed or that they have equal variance.

To implement generalized linear modeling, we needed to transform the data into a discrete format suitable for applying this technique, while taking into account censoring. We divided the times into uneven bins, with the first six bins of width 0.5, the seventh and eighth bins of width 1, and the rest of the bins of width 5. We did this because most of the wait times are low, but the wait time distribution is also right skewed. We used a data transformation method suitable for generalized linear modeling that centered around classifying each wait time into one of these bins.

```
##
## Call:
## glm(formula = assessment ~ p1 + p2 + p3 + p4 + p5 + p6 + p7 +
##     p8 + p9 + p10 + p11 + p12 + p13 + male + black + hisp + sn1 +
##     sn2 + sn3 + all4, data = data_pp)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.55582  -0.28347  -0.11408  -0.00026   1.03387
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.99564    0.28000   3.556 0.000392 ***
## p1          -0.96557    0.28050  -3.442 0.000597 ***
## p2          -0.82812    0.28063  -2.951 0.003230 **
## p3          -0.71756    0.28096  -2.554 0.010774 *
## p4          -0.59150    0.28158  -2.101 0.035883 *
## p5          -0.58673    0.28289  -2.074 0.038292 *
## p6          -0.63936    0.28525  -2.241 0.025186 *
## p7          -0.49804    0.28822  -1.728 0.084248 .
## p8          -0.55580    0.29656  -1.874 0.061155 .
## p9          -0.65271    0.30929  -2.110 0.035039 *
## p10         -0.47834    0.32342  -1.479 0.139402
## p11         -0.63217    0.36140  -1.749 0.080513 .
## p12         -1.00000    0.39514  -2.531 0.011510 *
## p13               NA         NA      NA       NA
## male        -0.02981    0.02299  -1.297 0.195025
## black       -0.02363    0.02599  -0.909 0.363392
## hisp        -0.06933    0.06539  -1.060 0.289247
## sn1          0.03853    0.03011   1.280 0.200920
## sn2          0.03520    0.03543   0.993 0.320691
## sn3          0.04435    0.05087   0.872 0.383478
## all4         0.27264    0.10547   2.585 0.009856 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1561325)
##
##     Null deviance: 213.17  on 1201  degrees of freedom
## Residual deviance: 184.55  on 1182  degrees of freedom
## AIC: 1200.8
##
## Number of Fisher Scoring iterations: 2
```

Based on the summary of the generalized linear model without kernel regression, only one explanatory variable coefficient is significant, the binary variable of whether the patient has all four stroke symptoms or not. The rest of the explanatory variables do not have statistically significant effects on the wait time.

Generalized Linear Modeling with Kernel Regression

To set up our generalized linear model with kernel regression, we first needed to transform the data in a similar fashion to the data conversion for our generalized linear modeling without kernel regression, but this time using bins of equal width for the wait times. We used seven kernels for the regression, and weighted the kernels for the lower bins (corresponding to the shorter wait times) more than the higher bins (corresponding to the longer wait times), because more patients had shorter wait times until clinical assessment. Again, we assume that the response variables for the patients are independent, and we don't need to check for normality or homoscedacity of the residuals for this model.

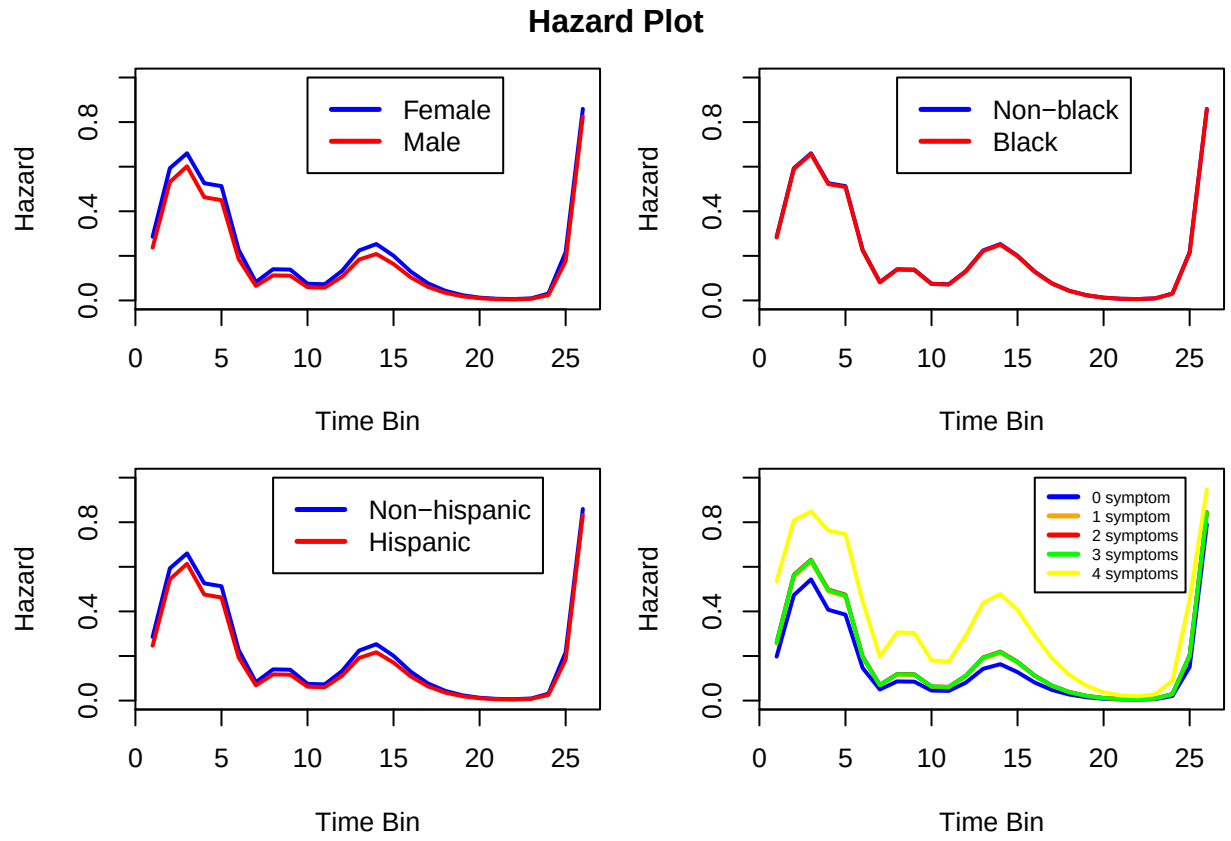
```
##
## Call:
## glm(formula = y ~ 0 + ., family = "binomial", data = d2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4833  -0.8204  -0.6693   1.0284   2.5741
##
```

```
## Coefficients: (4 not defined because of singularities)
##         Estimate Std. Error z value Pr(>|z|)
## X1    -1.757e+83  2.650e+83  -0.663   0.5074
## X2            NA         NA      NA       NA
## X3            NA         NA      NA       NA
## X4            NA         NA      NA       NA
## X5            NA         NA      NA       NA
## X6     3.510e+13  5.296e+13   0.663   0.5074
## X7    -1.299e+10  1.960e+10  -0.663   0.5074
## X8     7.507e+03  1.088e+04   0.690   0.4901
## X9    -5.159e+01  2.373e+01  -2.174   0.0297 *
## X10   -4.549e+01  2.479e+01  -1.835   0.0665 .
## X11   -7.095e+01  3.347e+01  -2.120   0.0340 *
## X12   -1.080e+02  4.948e+01  -2.183   0.0290 *
## X13   -3.883e+02  1.801e+02  -2.156   0.0310 *
## X14   -4.564e+02  2.188e+02  -2.086   0.0370 *
## male  -2.519e-01  1.738e-01  -1.449   0.1473
## black -1.527e-02  1.986e-01  -0.077   0.9387
## hisp  -2.017e-01  4.565e-01  -0.442   0.6586
## sn1    3.716e+01  1.801e+01   2.063   0.0391 *
## sn2    3.719e+01  1.801e+01   2.065   0.0389 *
## sn3    3.716e+01  1.802e+01   2.063   0.0391 *
## all4   3.833e+01  1.803e+01   2.126   0.0335 *
## sn0    3.683e+01  1.801e+01   2.045   0.0409 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 944.07  on 681  degrees of freedom
## Residual deviance: 782.63  on 663  degrees of freedom
## AIC: 818.63
##
## Number of Fisher Scoring iterations: 25
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 0 + (X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + male + black + hisp + sn1 + sn2 + sn3 +
##     all4 + sn0)
## Model 2: y ~ 0 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + male + black + hisp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       663     782.63
## 2       668     794.92 -5  -12.293  0.03099 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 0 + (X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + male + black + hisp + sn1 + sn2 + sn3 +
##     all4 + sn0)
## Model 2: y ~ 0 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + sn0 + sn1 + sn2 + sn3 + all4
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       663     782.63
## 2       666     784.85 -3  -2.2176   0.5285
```



Based on the summary of the generalized linear model with kernel regression, none of the coefficients for the explanatory variables of race, ethnicity, gender, or number of symptoms are significant, contrasting with the results for the generalized linear model without kernel regression.

However, we do see significance in the number of symptoms a patient has when all 4 symptoms are present. This is evident in hazard plots above comparing the levels of different explanatory variables. In each of the plots above, we plot a hazard curve for the different values of each predictor variable. We see that for Male, Black, and Hispanic, the different levels of the predictors result in very similar hazard curves that overlap heavily. This suggests that the different values of these predictors don't significantly affect the chance that a patient gets evaluated at any point in time. However, the hazard curve for a patient with 4 symptoms present is significantly different from the hazard curves for patients with 0-3 symptoms. This suggests that a patient displaying 4 symptoms has a higher chance at each point in time to be evaluated.

For the plots above, we fixed the predictors not being plotted as arbitrary values. However, fixing the other variables to different values also produces similar results.

Conclusions

After an analysis of our Cox proportional hazard model, our generalized linear model created without kernel regression, and our generalized linear model generated with linear regression, we see that none of the predictors had any statistically significant correlation with the wait time until clinical assessment for a stroke, with a single exception. The predictors for wait time included whether or not the patient was black, whether or not the patient was Hispanic, whether the patient was male or female, and whether the patient displayed one,

two, three, or four symptoms of a stroke. The only predictor with a statistically significant coefficient at an alpha of 0.05 was the variable for whether or not the patient displayed all four symptoms of a stroke, and this was only significant for the generalized linear model without kernel regression. The other models did not show statistical significance for the correlation of this explanatory variable with the wait time. We therefore conclude that none of the predictors in this study have an impact on the wait time until clinical assessment for a stroke, except for clinical presentation, which may have some effect when the patient displays all four symptoms of a stroke.

Although our models indicated that most of this study's explanatory variables were not significantly associated with wait time, this outcome may be a result of having too little data. For example, there were only nine Hispanic patients in this data set. If there had been more Hispanic patients, perhaps our models would have shown a statistically significant difference between the wait times of Hispanic and non-Hispanic patients. The problem of having too few observations affects the reliability of our results, but the only solution is to gather more data.

Even though our models show that most explanatory variables have no statistically significant impact on wait times, this result has positive implications even though it may initially seem uninteresting. Based on our analysis, the predictors of race, ethnicity, and gender have no impact on wait times, indicating that the selection process for clinical assessment in the ER is non-discriminatory. The only factor that may be relevant in influencing wait time is whether or not the patient has all four symptoms of a stroke. We would expect and hope that a patient with more stroke symptoms would have to wait less time for a clinical assessment, because it is even more urgent that this patient see a doctor. We have found evidence that suggests that a patient with stroke symptoms need not worry about discrimination in the emergency room based on race, ethinicity, or gender, and the only variable that may impact wait time is the number of symptoms exhibited.

Goals for Further Analysis

In our next report, we hope to visualize the hazard curves for our models to better understand the process of waiting for a clinical assessment for a stroke. We also need to analyze different variations of each model to find the best fit. It may be that if we were to choose different kernel weights for our generalized linear model with kernel regression, the predictor of having all four symptoms of a stroke may show a statistically significant association with wait times in the ER.

References

"Introduction to Generalized Linear Models." *STAT 504*, Eberly College of Science at Penn State, onlinecourses.science.psu.edu/stat504/node/216.

Contributions

William and Wuming both helped transform the data for implementing the generalized linear model and the generalized linear model with kernel regression. Wuming and William wrote the code to use generalized linear models with and without kernel regression as well. Annie wrote the text for the conclusion and model approaches as well as analyzed the model.