STA 440 Case 2

Annie Lott, Wuming Zhang, William Yang

October 1, 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.

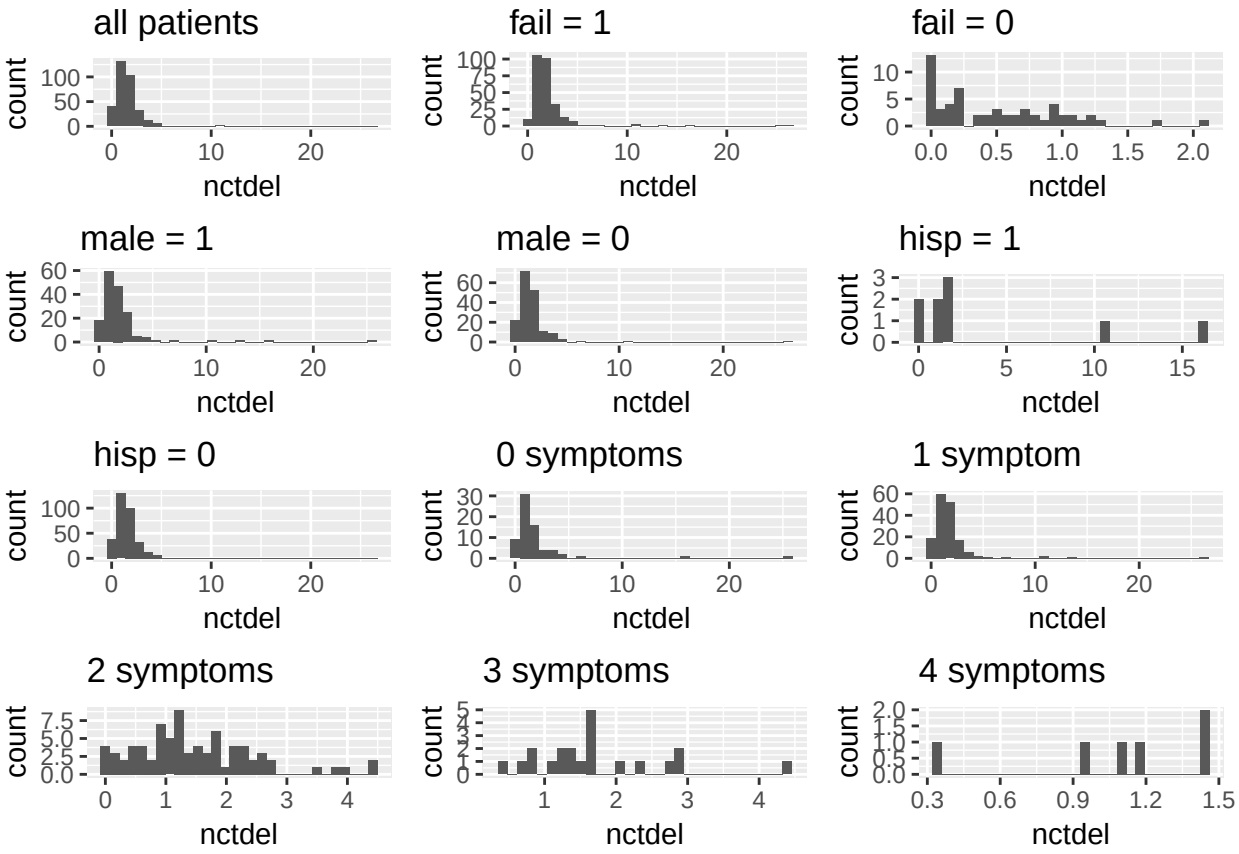
Based on data from 335 emergency room patients with mild to moderate motor impairment, possibly indicative of a stroke, we are analyzing whether sex, race, ethnicity, and the number of symptoms displayed affects the time until the patients receive a neurological assessment, or whether or not they receive 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.

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.

Data Cleaning

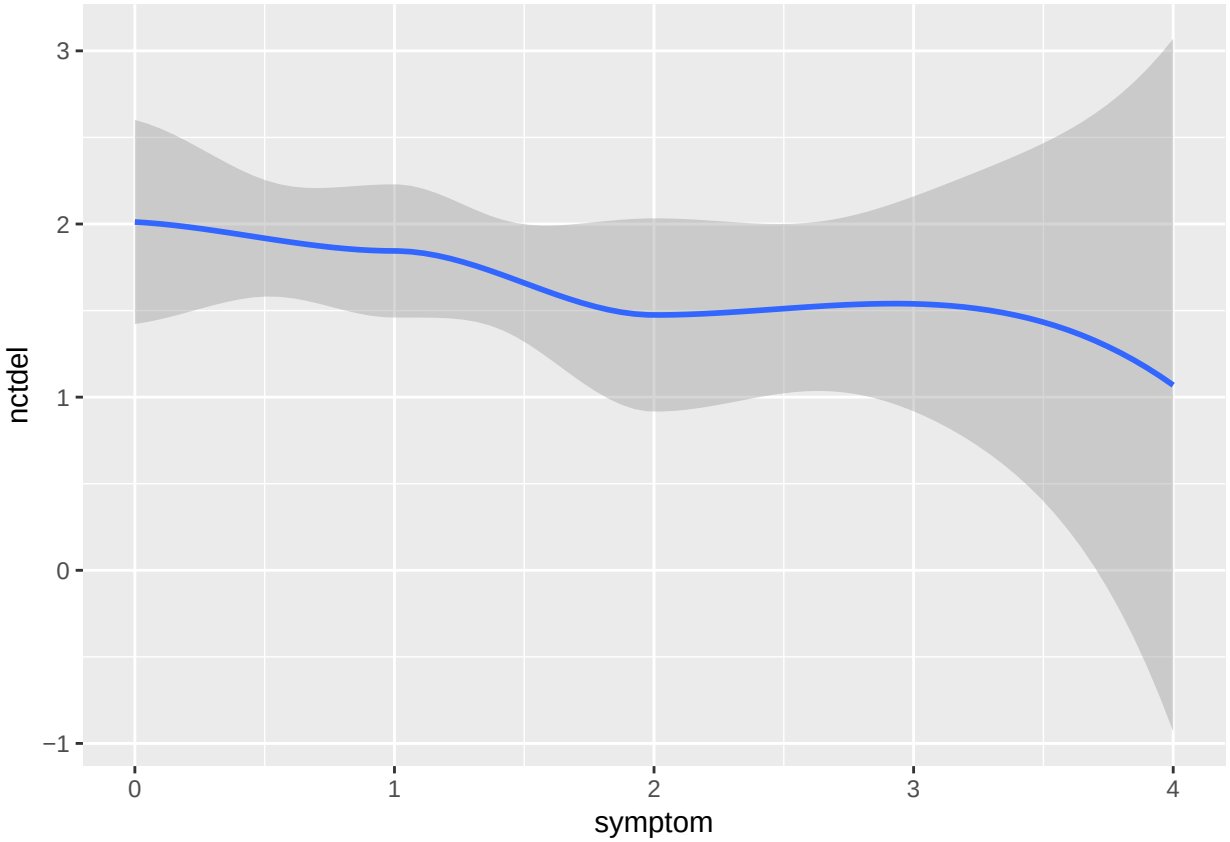
Exploratory Data Analysis

Histograms of each Indicator



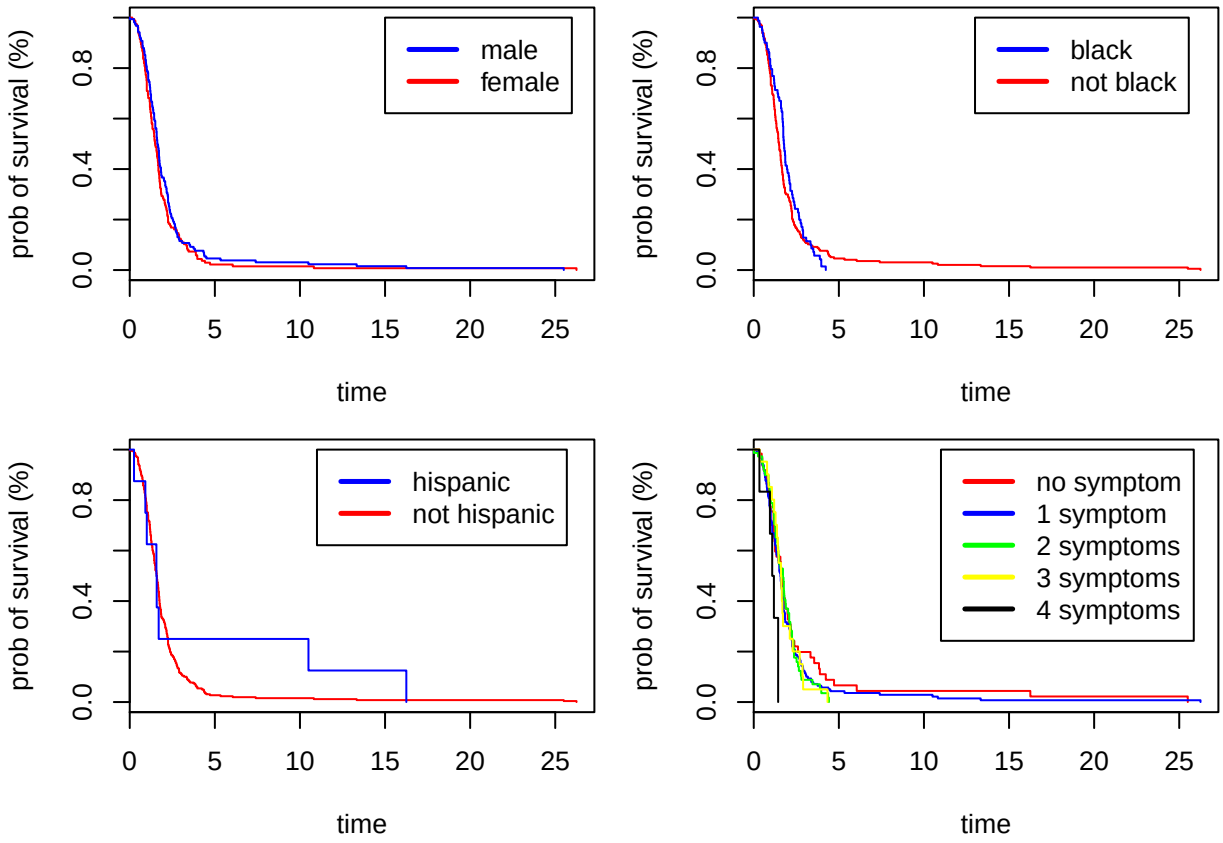
To get a better sense of the distributions in the data, we first plot histograms of the nctdel times for each level of all the variables. We notice some general trends here - nctdel time seems to go down with more symptoms, the overall distribution of times for all patients is right-skewed with a mean 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.

Correlation between Waiting Time and Number of symptoms



To further explore the possible correlation between the nctdel time and number of symtons, we plot a smooth fitted line of nctdel time vs symptom. We can observe that the nctdel time is likely to be shorter for patients with more symptoms, but we don't know with certainy that this relationship is statistically significant.

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.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  nctdel by zero.four
## W = 235.5, p-value = 0.5844
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  nctdel by male
## W = 12734, p-value = 0.146
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  nctdel by black
## W = 9657, p-value = 0.1207
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  nctdel by hisp
## W = 1442.5, p-value = 0.9333
## alternative hypothesis: true location shift is not equal to 0
```

To perform an initial inferential analysis evaluating the impact of clinical presentation, gender, ethnicity, and race on time to neurological assessment, our first thought was to use two-sample t-tests to compare the time to neurological assessment for the levels of this study's different factors (sex, race, ethnicity, and number of symptoms). However, two-sample t-tests require normal distributions of the wait time for the levels of the variables, and we know from the exploratory data analysis that many of these distributions are instead right-skewed. Additionally, these distributions are cut off at zero, because there is no such thing as negative wait times. We therefore used a nonparametric Mann-Whitney-Wilcoxon test to investigate the differences in wait time by variable (R Tutorial 2017). We assumed that the distributions are independent, which is reasonable, but we need not assume that they are normal. The results of the Mann-Whitney-Wilcoxon tests show that there is no significant difference in wait time for males versus females, black people versus non-black people, hispanic people versus non-hispanic people, and people showing no symptoms versus people showing four symptoms. We only compared people displaying no symptoms, the minimum, to people showing four symptoms, the maximum, rather than comparing all the combinations of zero, one, two, three, or four symptoms because it is bad practice to run too many tests on one variable. We thought that if there was a difference in wait times among people showing different symptoms, it would be most evident between those with no symptoms and those with the maximum number of symptoms. However, there seems to be no significant differences in wait time associated any of the variables, in our preliminary analysis. This preliminary analysis is limited in that it only considers differences in the mean wait time between categories, rather than the survival curves overall.

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 naive implementation of the model on all the predictors does not identify any of them as having a significant effect, so adjustments to the model and data will be necessary to produce meaningful results.

Other Modeling Approaches

While we have decided to use the Cox Proportional Hazard Model to represent our data, there are other possible modeling approaches, such a Bayesian survival analysis based on the Weibull distribution. In this approach, a prior is given for a Weibull regression model, and a posterior is derived using Markov Chain Monte Carlo. We will use the Cox Proportional Hazard Model because it is simpler to implement.

References

'Mann-Whitney-Wilcoxon Test.' R Tutorial, 2017, www.r-tutor.com/elementary-statistics/non-parametric-methods/mann-whitney-wilcoxon-test.

Contributions

William plotted and analyzed histograms of each level of each indicator as part of the EDA and helped write potential modeling approaches. Wuming plotted the correlation between assessment time and number of symptoms curve and the Kaplan-Meier curve and helped write up some discussion about the Kaplan-Meier curve. Annie wrote the Goals and Data Description section and contributed the section on the Mann-Whitney-Wilcoxon test, and helped with the Modeling Approaches section.