Case 2 Part 3

Sonia Xu, Grant Goettel, Ian Hua

October 27, 2017

Introduction

The following data is from a study of time to critical neurological assessment for patients with stroke-like symptoms who were admitted to the emergency room. The purpose of the analysis is to perform inferences on the impact of clinical presentation (reported number of makor stroke symptoms), gender, and race on time to neurological asessment. This paper fits a model that identifies the differences in wait time to neurological assessment based on these features of interest. A final model of kernel regression with a response of time at failure, continuous feature of nctdel estimated via kernel regression, and categorical features of black, hispanic, male, number of symptoms.

Dataset

The dataset contained information about the amount of time elapsed from arrival at the ER to the assessment, and whether or not the patient received a CT scan. It also contained data about whether each patient was male or female, whether they were black or hispanic or not, and the amount of major symptoms of a stroke they had upon arrival (out of 4 main symptoms). Rather than have four separate indicator variables showing whether they had 1, 2, 3, or 4 symptoms, we created a single numerical variable containing the amount of symptoms.

Methodology

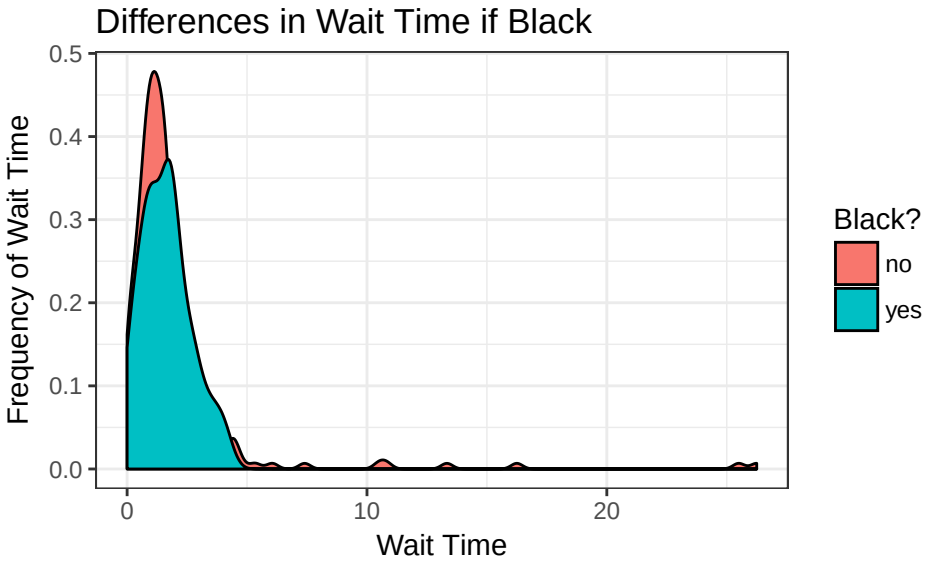
To assess the impact of clinical presentation, gender and race on time to neurological assessment, multiple models and analyses were explored to find the best model that fit the data. After exploring Cox Proportional Hazards, Kaplan-Meier Estimate, Random Forest, and Kernel Regression, the Kernel Regression Model seemed to perform the best despite its flaws. We chose 10 bins based on quantiles to account for the uneven frequencies of the data.

EDA

Racial Bias

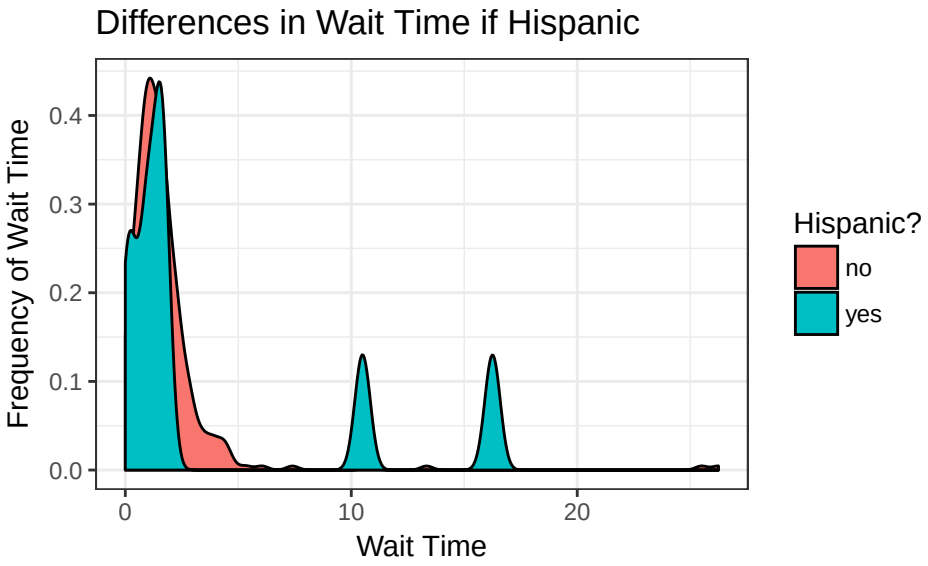
Black vs. Non-Black

There could potentially be bias here, but more analysis should be done. Non-blacks appear to have shorter wait times, but some non-blacks have extremely long wait times.



Hispanic vs. Non-Hispanic

There could potentially be bias here, but the lack of sample size for Hispanics appears to be an issue.



Gender Bias

There could potentially be bias here, but more analysis should be done. Females appear to have shorter wait times.

Final Model

We decided to create a Kernel Regression model with time at failure as the response to explore the robustness of this new model.

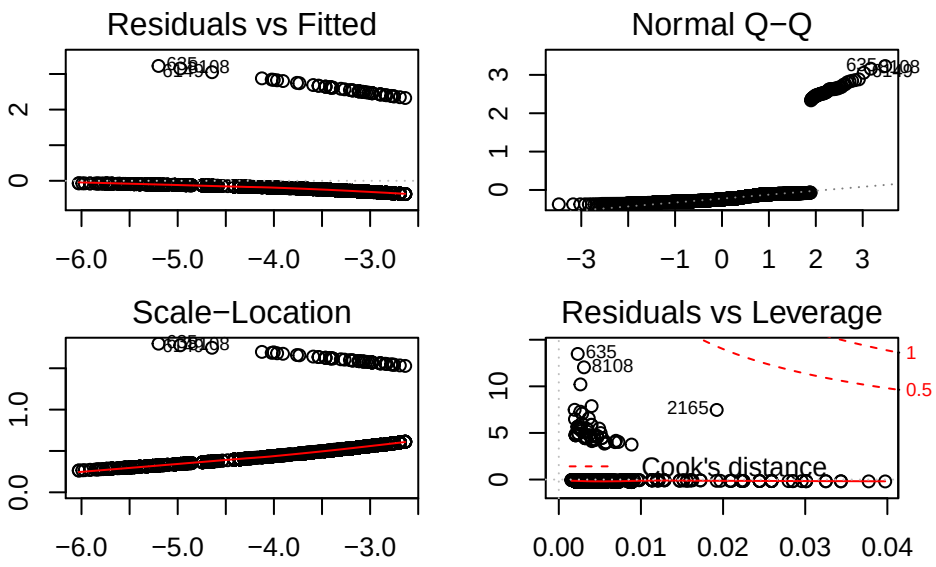
Kernel Regression with 10 Bins

10 bins were calculated to fit the kernels. The bins are unevenly spaced because the data has a higher concentration of points for the feature nctdel between 0 and 2, even though its range is (0,26.25). We chose 10 bins because we wanted to be careful of the sample size with only 335 observations. If we split that over too many bins, the sample sizes in the individual kernels would be uncomfortably small for regression. The model has time at failure as the response, a kernel estimation of nctdel, and the features male, black, hispanic, and count_sn. The bin levels are (-Inf,0], (0,0.357], (0.357,0.737], (0.737,0.983], (0.983,1.21], (1.21,1.49], (1.49,1.73], (1.73,2.16], (2.16,2.8], (2.8, Inf]. A summary of the model noted some significance for the feature nctdel (for the full summary, Appendix A).

Bins

Model Checks

To understand how well the model fits, we performed goodness of fit tests and a model check assumptions.



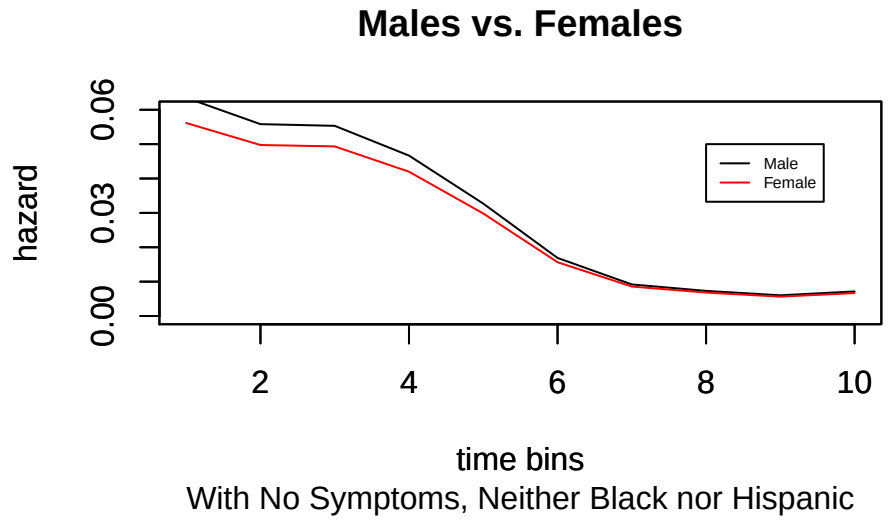
Looking at the Residuals vs. Fitted graph, the points do not exhibit a pattern and are relatively evenly distributed. With the Residuals vs. Leverage plot, the points are not homoscedastically distributed. In addition, data no longer tends to be normal the longer the patient stays waiting, as shown in the Normal Q-Q plot. Generally, the model seems to decently fit our data and is exceptionally better than previous models.

The model fits the true dataset 97.0854271% of the time when predicting for failure over the entire dataset, implying a decent model which holds much potential given the assumptions hold.

Discussion

Gender

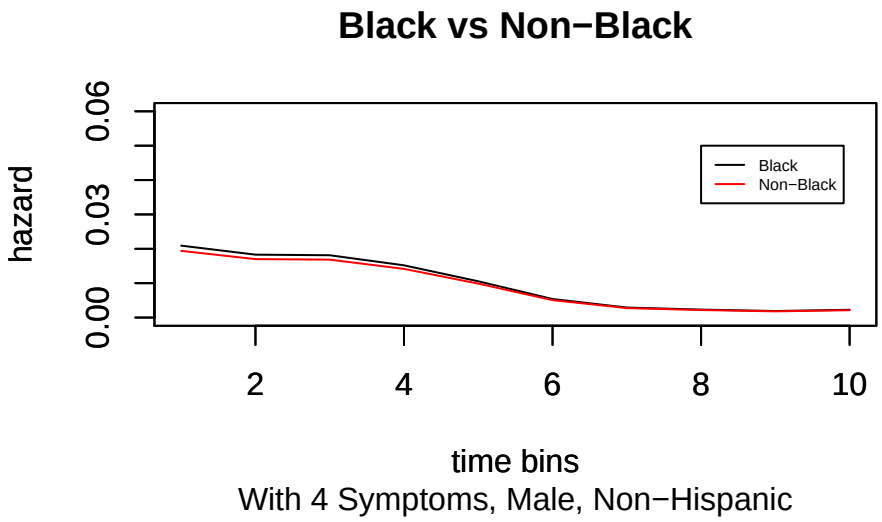
Looking at the survival function of Males (with no symptoms, neither black nor male) against Females (with the same characteristics), it reaffirms the EDA that females tend to have a lower time until assessment than males.



Race

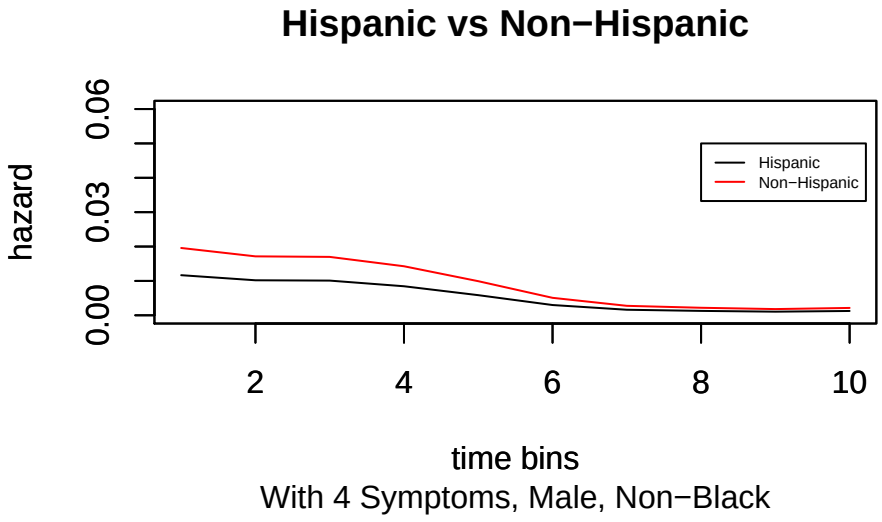
Black

Again, this survival function reaffirms that Non-Black patients have shorter wait times than Black patients (with features 4 symptoms, male, and non-hispanic held constant).



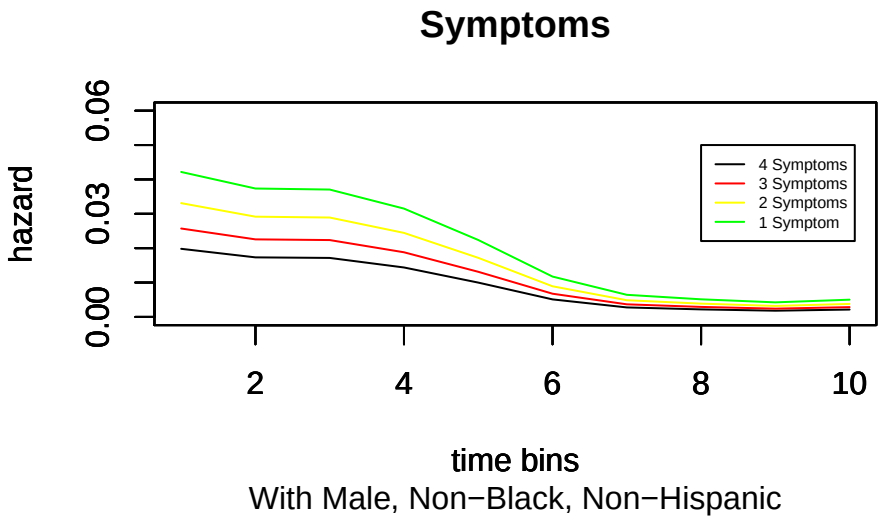
Hispanic

Hispanic patients tend to have a shorter wait time than Non-Hispanic patients.



Symptoms

Patients who exhibit 4 symptoms have shorter wait times than those who have 3,2, and 1.



Recommendations

Based on the survival function produced from the fitted model, there appears to be a positive bias towards females, non-Black patients, and Hispanic patients. In a perfect world, assesssment bias should only depend on the number of symptoms–intuitively, the more symptoms, the greater urgency, so a shorter wait time. Further qualitative studies should be conducted to understand why neurological assessments favor females, non-Black, and Hispanic patients.

A kernel regression hazard model is dependent on all the observations; a slight change in some observations with high leverage could potentially change the model quite a bit. Reproducing this study on various differing datasets could prove beneficial in discovering a better model. It would be also be interesting to test different percentiles for the binning of the model.

Appendix A

```
##
## Call:
## glm(formula = y ~ 0 + ., family = "binomial", data = d2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3721  -0.2933  -0.2483  -0.1553   3.2264
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## X1        -9.00532    1.13955  -7.903 2.73e-15 ***
## X2        -8.19166    1.24270  -6.592 4.34e-11 ***
## X3       -14.23878    2.49426  -5.709 1.14e-08 ***
## X4       -16.62200    5.76268  -2.884  0.00392 **
## male      -0.06098    0.26801  -0.228  0.82001
## black      0.05580    0.30129   0.185  0.85307
## hisp      -0.53802    1.02553  -0.525  0.59984
```

```
## count_sn  -0.28731    0.16191  -1.775  0.07597 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2758.73  on 1990  degrees of freedom
## Residual deviance:  502.49  on 1982  degrees of freedom
## AIC: 518.49
##
## Number of Fisher Scoring iterations: 8
```

Contributions

Ian Hua: Discussion, Methodology, Model Checks

Grant Goettel: EDA, Dataset, Binning

Sonia Xu: Hazards, Final Model, Introduction