

# Case Study 2, Pt. 2

*Nathaniel Brown, In Hee Ho, Sarah Zimmermann*

*October 18, 2017*

## Introduction

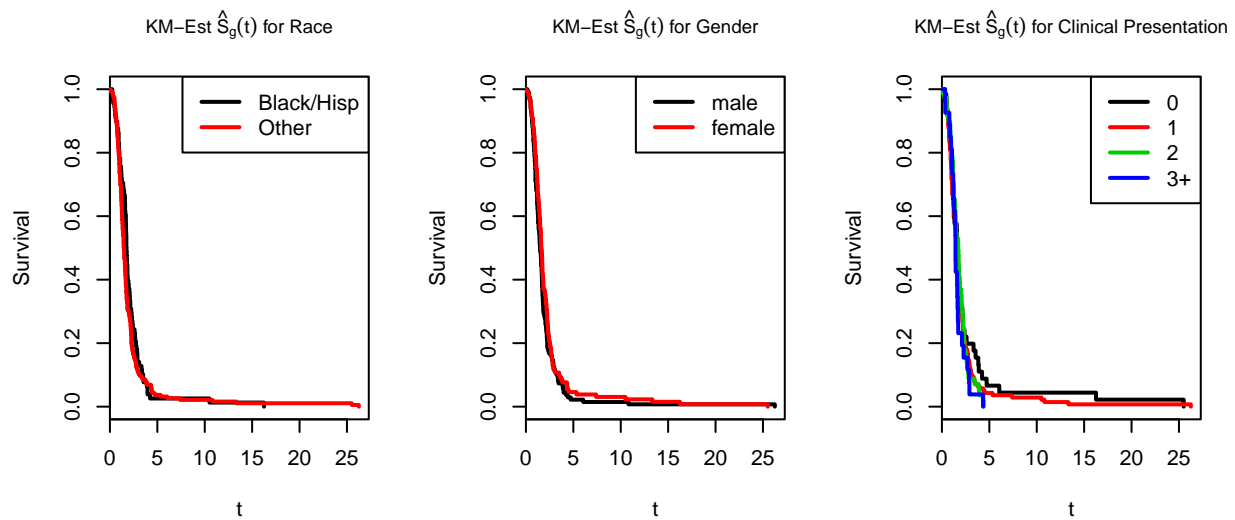
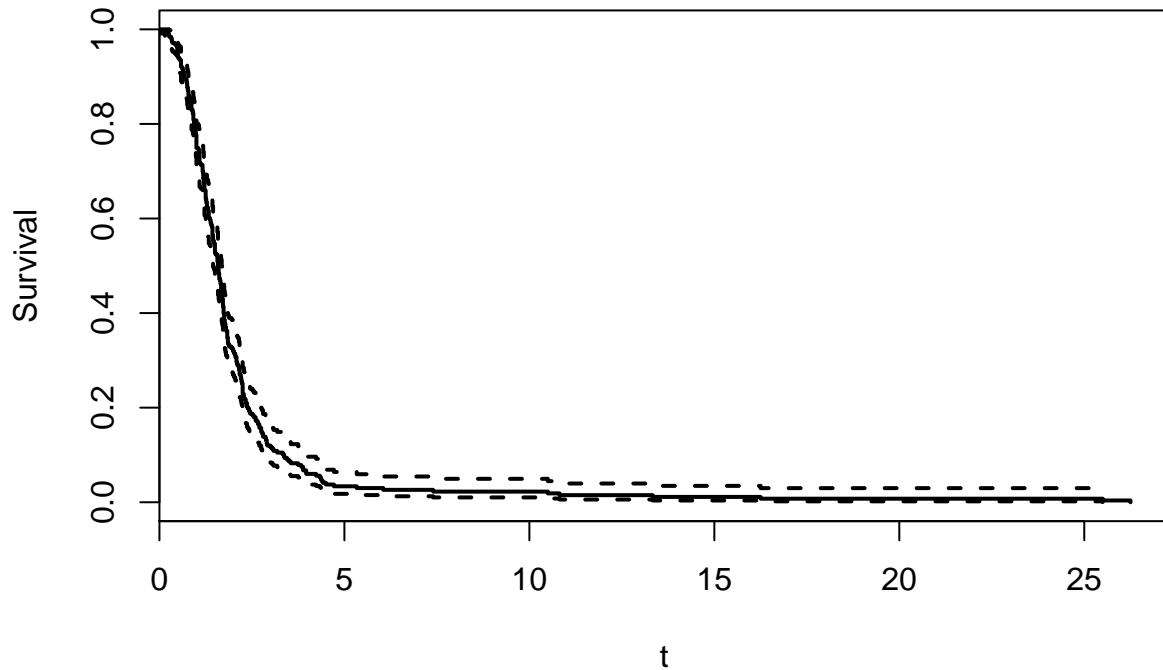
The data are from a study of time to critical neurological assessment for patients with stroke-like symptoms who are admitted to the emergency room. We are interested in the factors predictive of the time to assessment following admission to the ED for  $n=335$  patients with mild to moderate motor impairment. 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.

This report includes initial results for 2 different methods that were considered to analyze the data: Cox Proportional Hazard and Logistic Regression. There is analysis on model diagnostics, results, and problems which arise in the models. To conclude we assess the effectiveness of the work and steps to improve analysis for next week.

## Kaplan-Meier 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.

## Kaplan–Meier Estimate $\hat{S}(t)$ with CI



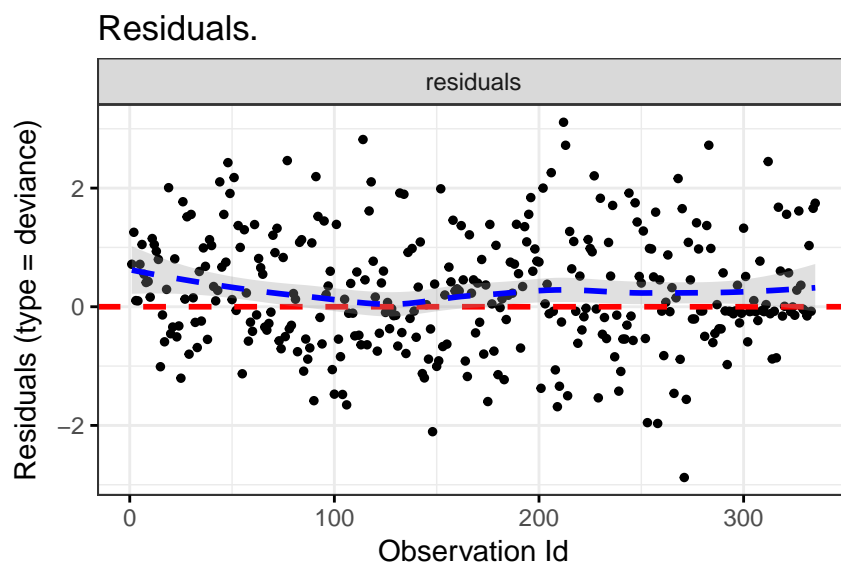
We now look at the separate estimated survival curves for different groups. From our Kaplan–Meier estimates for three different variables (race, gender, and the number of major stroke symptoms a patient shows), we observe that the survival curves are generally similar between 0–5 minute time range. However, there appears 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.

## Cox Proportional Hazard Model

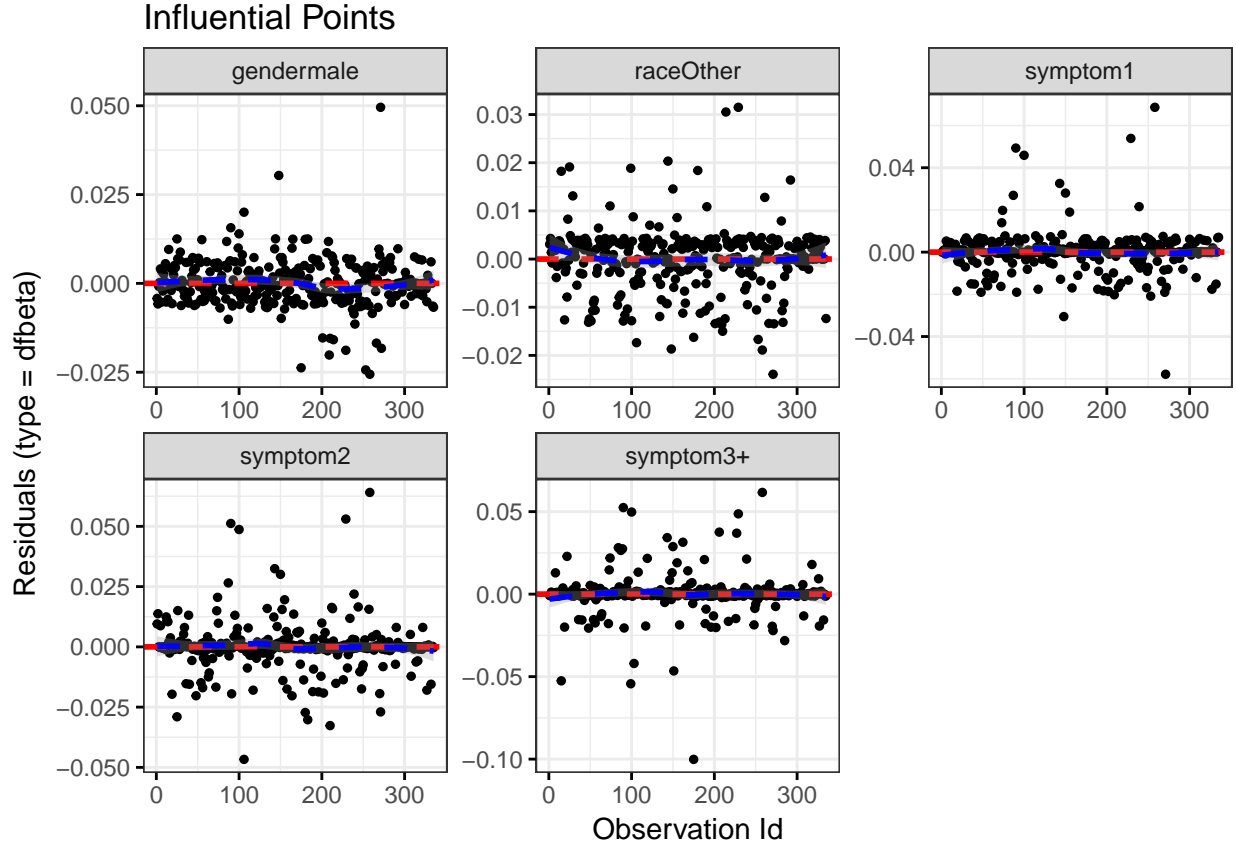
Cox regression looks into the effects of variables upon the event in the data set, which in this case is a CT scan. This model assumes that the effects of the predictor variables upon survival are constant over time and are additive in one scale. We fit the Cox proportional hazard models for waiting time until the event (CT scan) based on three variables: gender (female vs. male), race (Black or Hispanic vs. non-Black and non-Hispanic), and number of symptoms (0, 1, 2, or 3+).

### Diagnostics

Before moving on to analysis, we first check that our data satisfies the proportional hazard model assumptions. To do this, we investigate the residuals and look for any influential points.



The residual plot shows no apparent pattern. The points are generally evenly spread above and below 0.



Next, we investigate the influential points. For gender, although there are a few points close to or above  $\pm 0.025$ , all of the points are below 0.05 so we do not consider any of the points as influential. For race/ethnicity, there are points as high as 0.04, but since there are no points above/below 0.05, we do not consider any of the points to be influential. However, for clinical presentation (0, 1, 2 or 3+ symptoms), we observe that there are points as high as or higher than 0.05. We thus acknowledge that there are some points that can potentially affect the fit of the model.

	rho	chisq	p
gendermale	0.0428024	0.4949243	0.4817389
raceOther	-0.1043511	2.9934221	0.0836033
symptom1	0.0160123	0.0690484	0.7927274
symptom2	0.0931383	2.3895166	0.1221514
symptom3+	0.0642341	1.1439961	0.2848095
GLOBAL	NA	7.5916755	0.1802215

Finally, we check the assumption of proportional hazard using the `cox.zph()` function. Because none of the p-values for the covariates are significant (i.e. none of the p-values greater than 0.05), we can assume the hazards are proportional.

Since all of the assumptions of the Cox Proportional Hazard model are met, we now continue on to the actual model.

## Results

We are 95% confident that the wait time until scan for a male patient is 0.6790 to 1.092 minutes longer

than that of a female patient, if all else are held confident. Also, with 95% confidence level, we say that a non-black and non-hispanic patient waits for the examination 0.9052 to 1.527 minutes longer than a black or hispanic patient, with all other factors held constant. We are 95% confident that a patient who shows two of four major stroke symptoms will wait 0.8587 to 1.644 minutes longer than a patient with no notable symptoms, and that a patient with three or more symptoms will wait 0.8949 to 2.330 minutes longer for the neurological/CT scan than a patient with no symptoms.

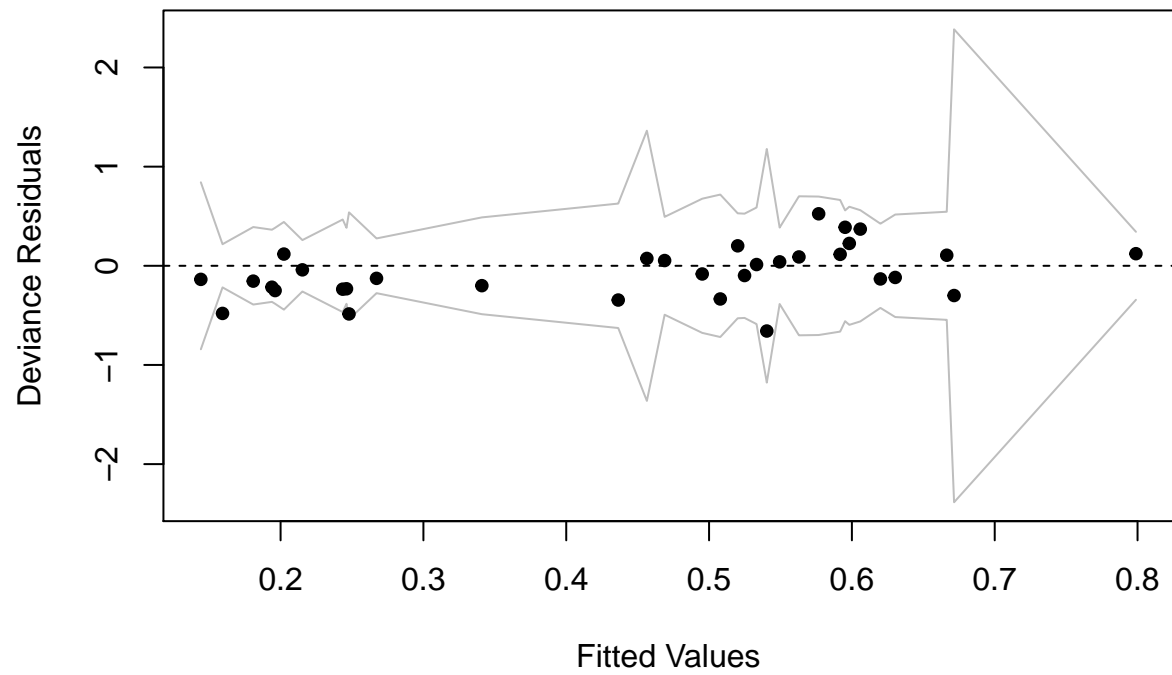
We note three things. First, that the 95% confidence interval contains 0 suggests there is no significant gender, race, or clinical presentation bias. Also, none of the variables are significant predictors in the model, which supports the idea that the wait response time is not influenced by gender, race, or clinical presentation. Finally, the R-squared value of 0.016 implies that only 1.6% of the variance is explained in the model. The summary statistics are presented below.

```
## Call:
## coxph(formula = Surv(nctdel, fail) ~ gender + race + symptom,
##       data = dat)
##
##      n= 335, number of events= 277
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## gendermale -0.1496   0.8610  0.1212 -1.235   0.217
## race0ther   0.1620   1.1758  0.1335  1.214   0.225
## symptom1    0.1724   1.1881  0.1657  1.040   0.298
## symptom2    0.1750   1.1913  0.1947  0.899   0.369
## symptom3+   0.3674   1.4439  0.2441  1.505   0.132
##
##              exp(coef) exp(-coef) lower .95 upper .95
## gendermale    0.861    1.1614    0.6790    1.092
## race0ther     1.176    0.8505    0.9052    1.527
## symptom1      1.188    0.8417    0.8587    1.644
## symptom2      1.191    0.8394    0.8134    1.745
## symptom3+     1.444    0.6925    0.8949    2.330
##
## Concordance= 0.538 (se = 0.021 )
## Rsquare= 0.016 (max possible= 1 )
## Likelihood ratio test= 5.45 on 5 df,  p=0.3635
## Wald test              = 5.43 on 5 df,  p=0.3659
## Score (logrank) test = 5.45 on 5 df,  p=0.3635
```

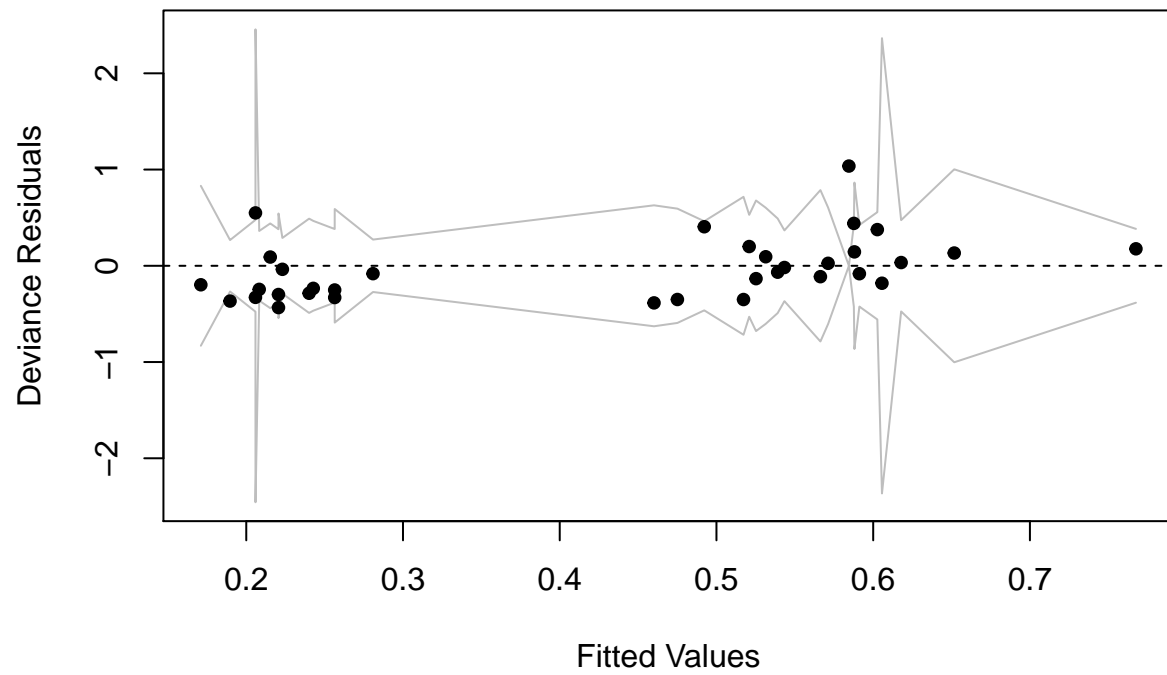
## Logistic Regression

To perform logistic regression on the data, we categorize the time-to-event variable into groups of 1 minute, with all events occurring 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 regression: 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.

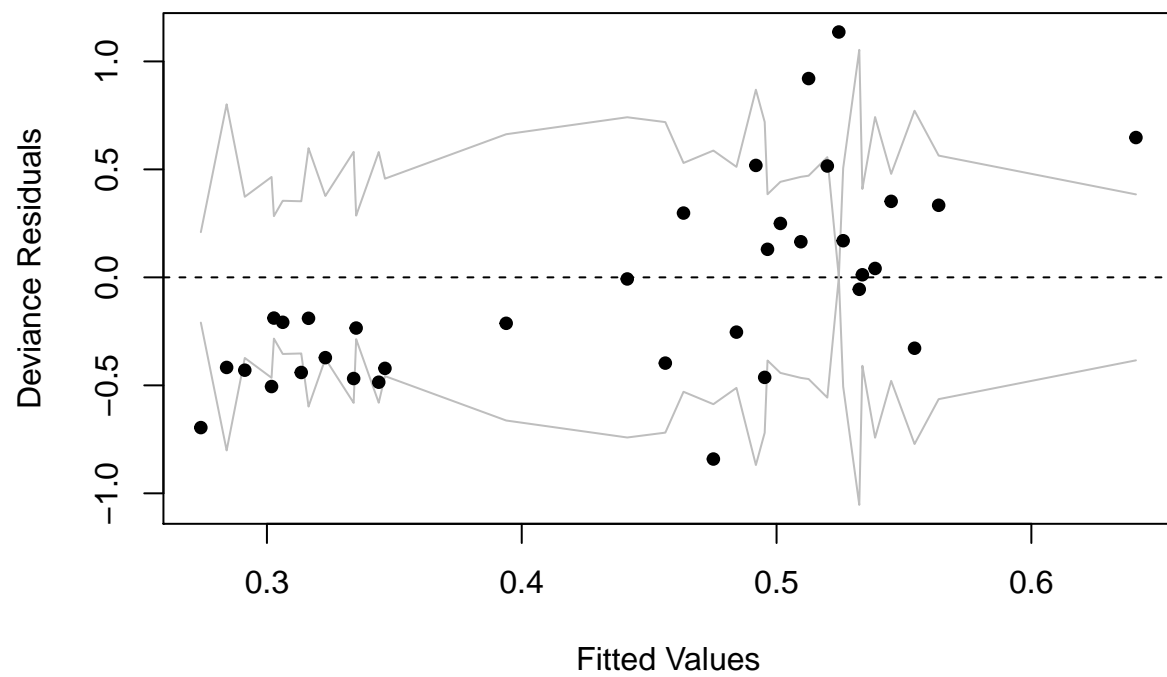
### OLS Logistic Regression Binned Residuals



**LASSO Logistic Regression Binned Residuals**



**Ridge Logistic Regression Binned Residuals**



	Deviance p-value
OLS	2e-04
LASSO Penalty	3e-04
Ridge Penalty	0e+00

	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



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.

## Discussion

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.

## Contributions

Nathaniel built the logistic regression model and the supporting visualizations and summary statistics. Sarah built and analyzed the Cox Proportional Hazard model and Kaplan Meier analysis. Inhee edited and added to the report.

## References

[http://influentialpoints.com/Training/coxs\\_proportional\\_hazards\\_modression\\_model-principles-properties-assumptions.htm#modmch](http://influentialpoints.com/Training/coxs_proportional_hazards_modression_model-principles-properties-assumptions.htm#modmch)  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/>  
<http://dwoell.de/rexrepos/posts/survivalKM.html>  
<http://www.sthda.com/english/wiki/cox-model-assumptions>