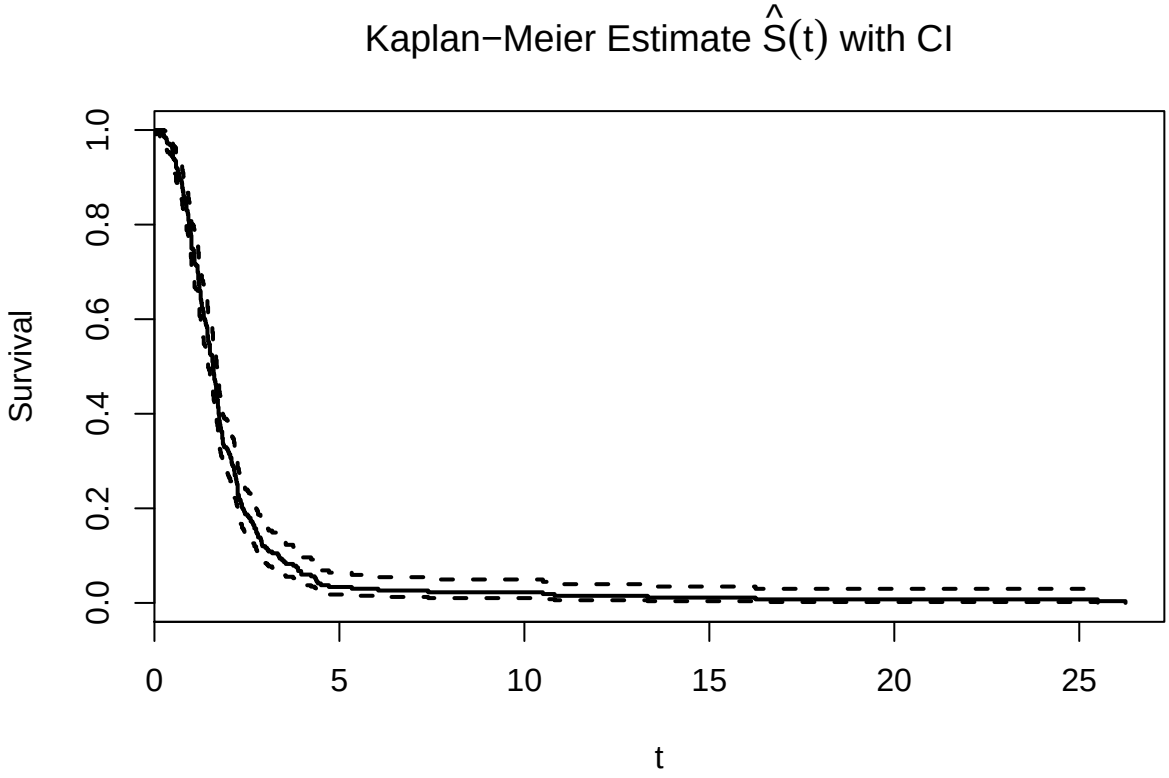
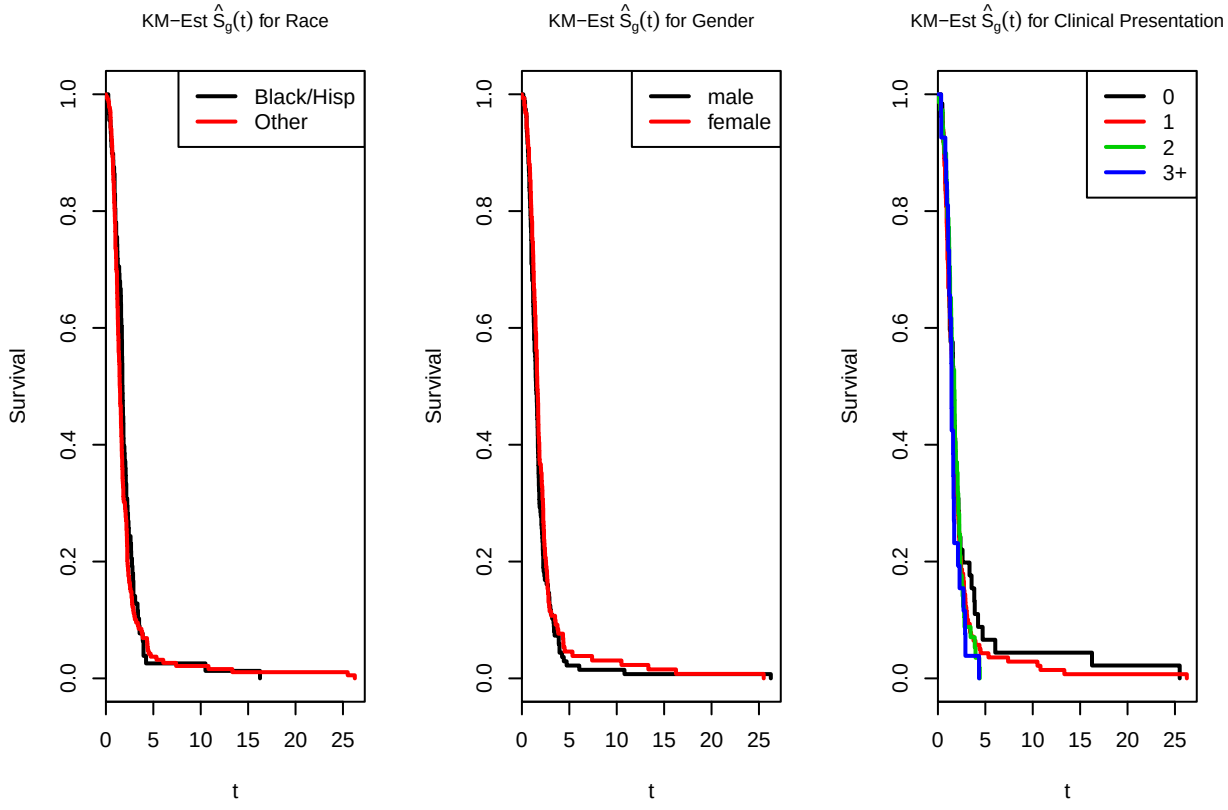
Case Study 2, Pt. 2

Nathaniel Brown, In Hee Ho, Sarah Zimmermann

October 18, 2017

Kaplan-Meier Analysis for Race, Gender, and Clinical Presentation



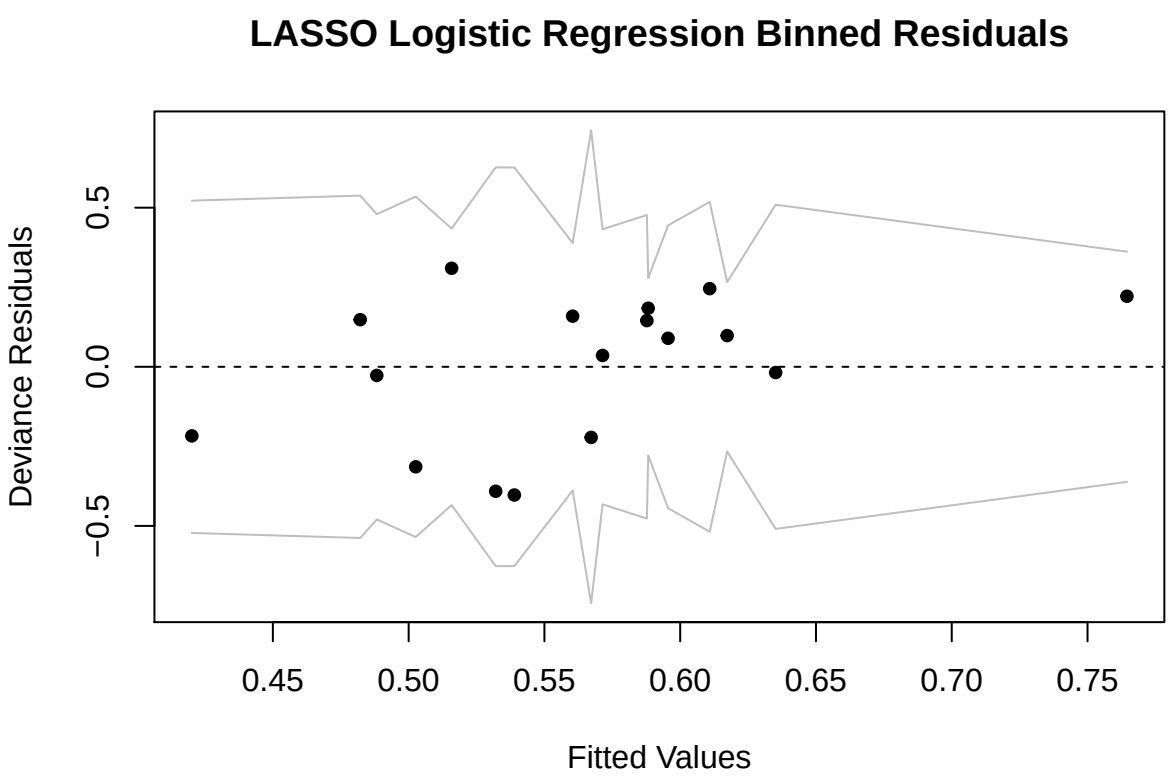


copy and paste some survival interpretations from the first submission. also mention that these lines need to be proportional for Cox. also re-order the cox stuff up here and logistic stuff last.

Logistic Regression

```
## NULL
```


OLS Logistic Regression Binned Residuals
Deviance Residuals
0.5
0.0
−0.5
0.4
0.5
0.6
0.7
0.8
Fitted Values


LASSO Logistic Regression Binned Residuals
Deviance Residuals
0.5
0.0
−0.5
0.45
0.50
0.55
0.60
0.65
0.70
0.75
Fitted Values

To perform logistic regression on the data, we categorized the time to event variable into groups of **BLANK** minutes, with all events occuring after **BLANK** minutes representing one group, since the sample size was so lo after that time (only 9 observations). The LASSO Regularization has the best fit. (the best fit depends on bin size and idk what bin size to use).

The ridge looks like it has cleanest residuals (for now) ###Blah blah, more insights and p-values from logistic stuff. positive and negative coefficients.

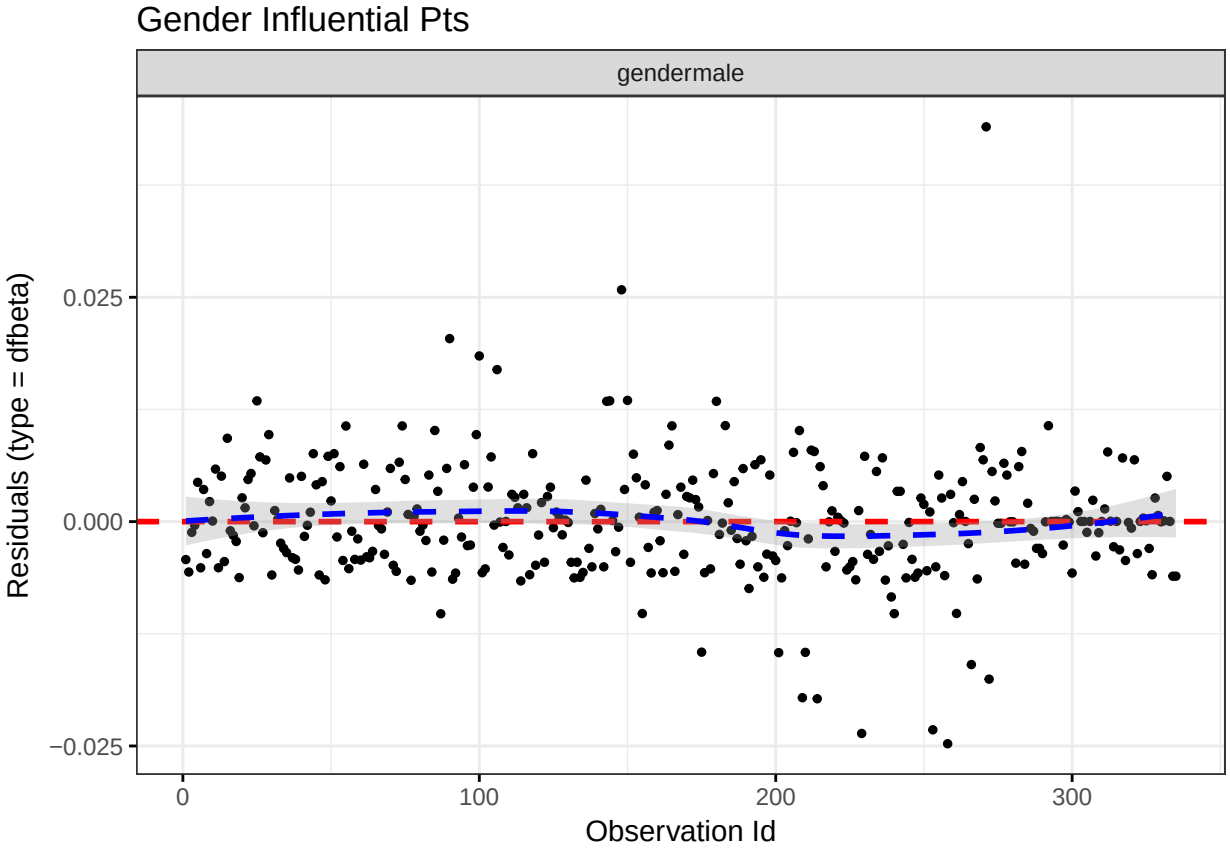
Cox Proportional Hazard

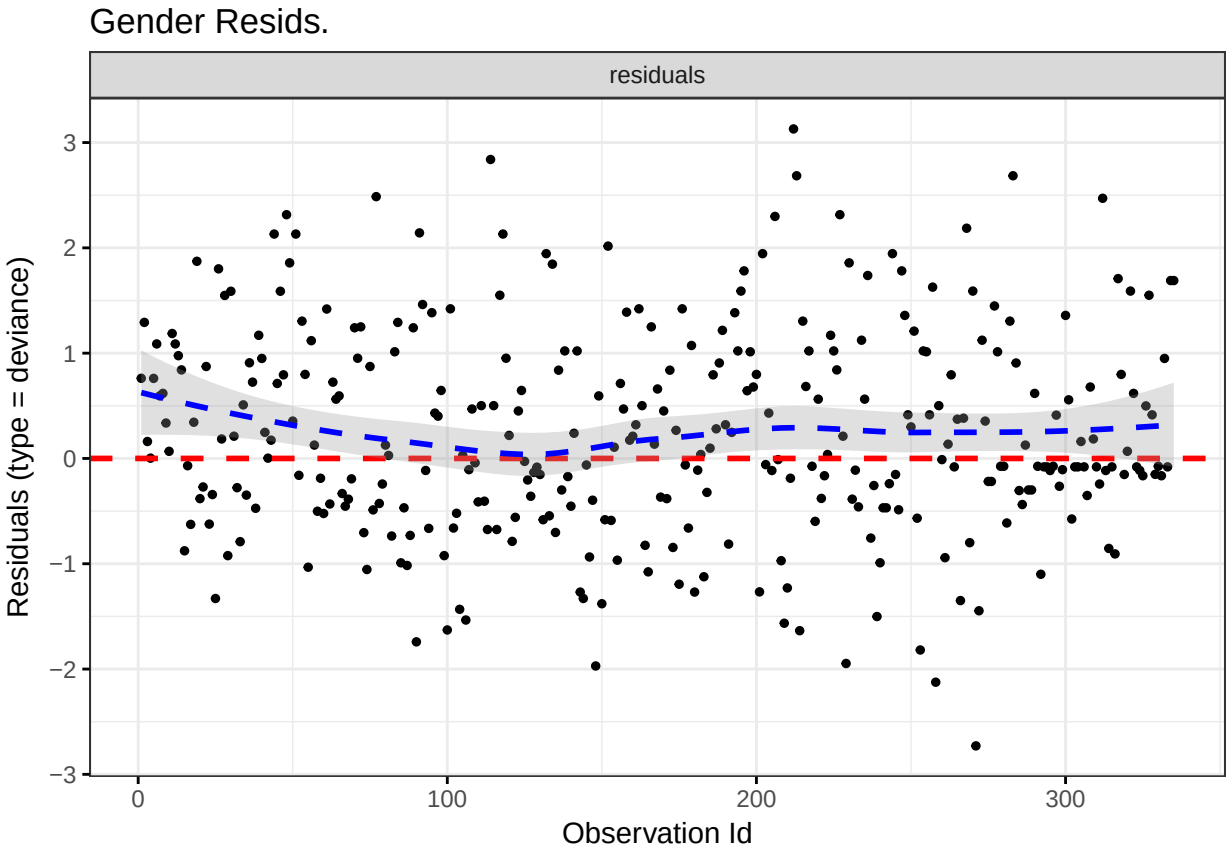
Below are models using Cox Proportional Hazard Model. This type of regression looks into the effects of variables upon the event in the data set which in this case is a CT scan. This model does assume the effects of the predictor variables upon survival are constant over time and are additive in one scale. We began by creating the models. The three listed below are the Cox Ph for waiting time until event (CT scan) based on gender (female vs male), race (black + hispanic vs non black + non hispanic), and number of symptoms (0,1,2, or 3+). Before analyzing any of these model we checked to see if the models are appropriate for the data by looking at the assumptions. Discussion is below.

Gender

Looking at the diagnostics for the model gender influence on wait time to CT, we first investigated the influential points and then residuals of the graph. Although there are a few close to or above +/-.025 all of the points are below .05 so we do not consider any of the points influential. Additionally, looking at the residuals we see no pattern in the graph and a generally even spread number of points above and below 0. Since we see the residuals are independent of time. Finally we checked the assumption of proportional hazard. From the output, the test is not statistically

significant (p=.447) for each of the covariates; therefore, we can assume the proportional hazards.

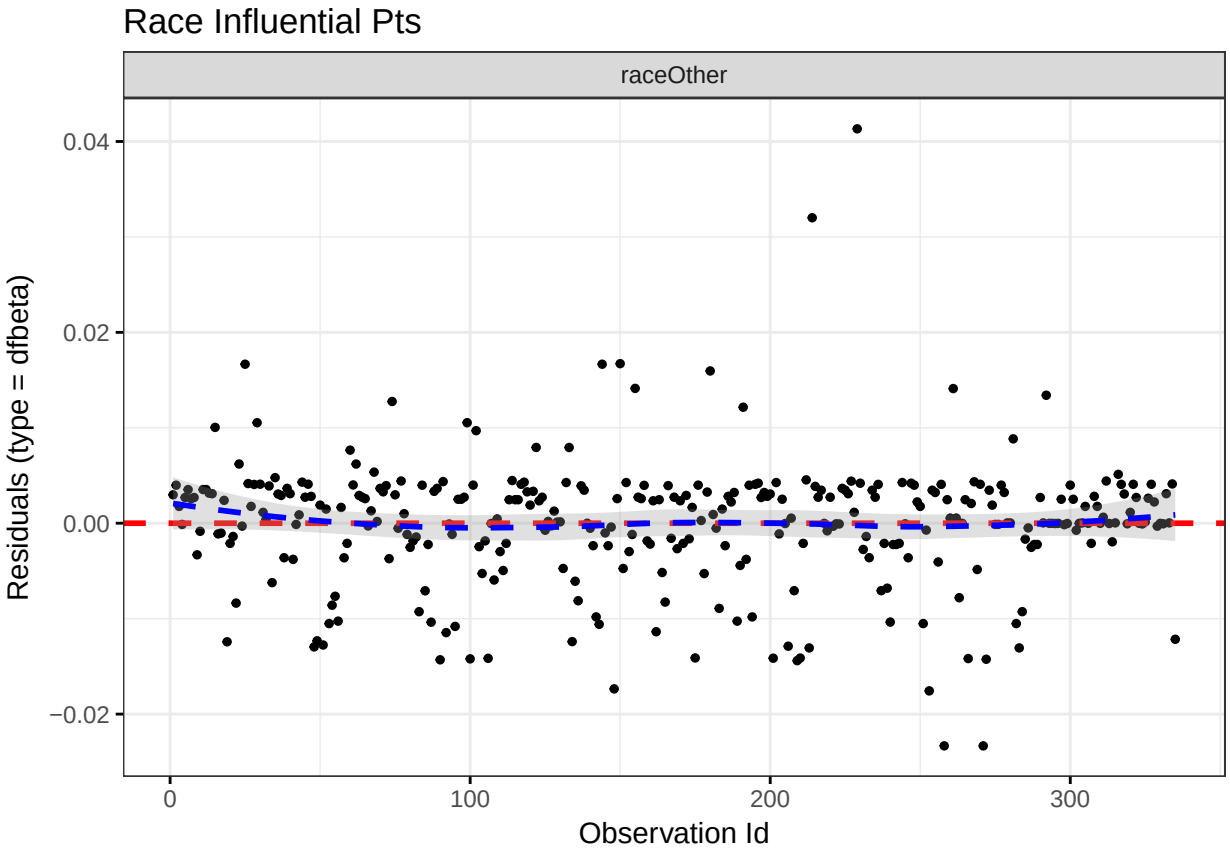


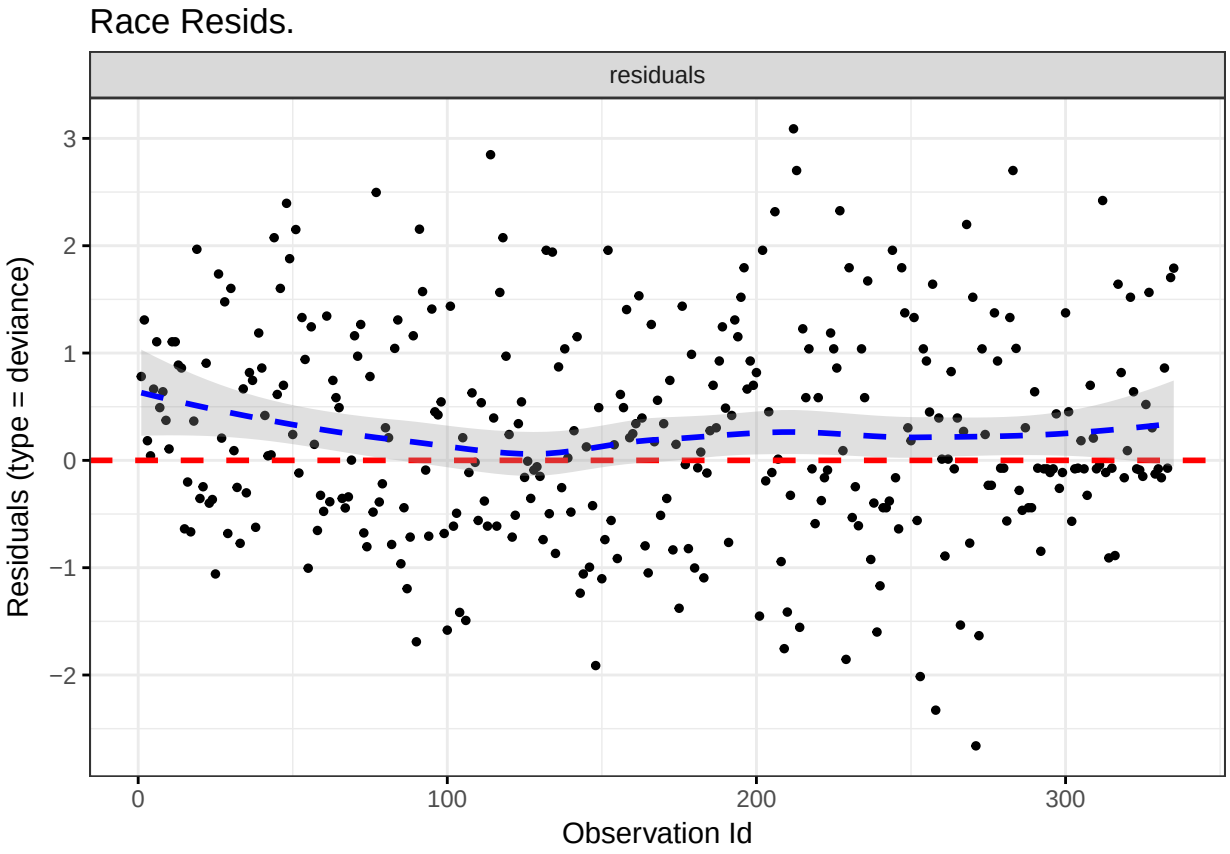



```
##              rho chisq     p
## gendermale 0.046 0.579 0.447
```

Race/Ethnicity

Looking at the diagnostics for the model race/ethnicity influence on wait time to CT the results were very similar to the analysis for gender above. Looking into the influential points, there are points as high as .04; however, since there are no points are above/below .05 so we do not consider any of the points influential. Additionally, looking at the residuals we see no pattern in the graph and a generally even spread number of points above and below 0. Since we see the residuals are independent of time. Finally we checked the assumption of proportional hazard. From the output, the test is not statistically significant (p=.0802) for each of the covariates; therefore, we can assume the proportional hazards.

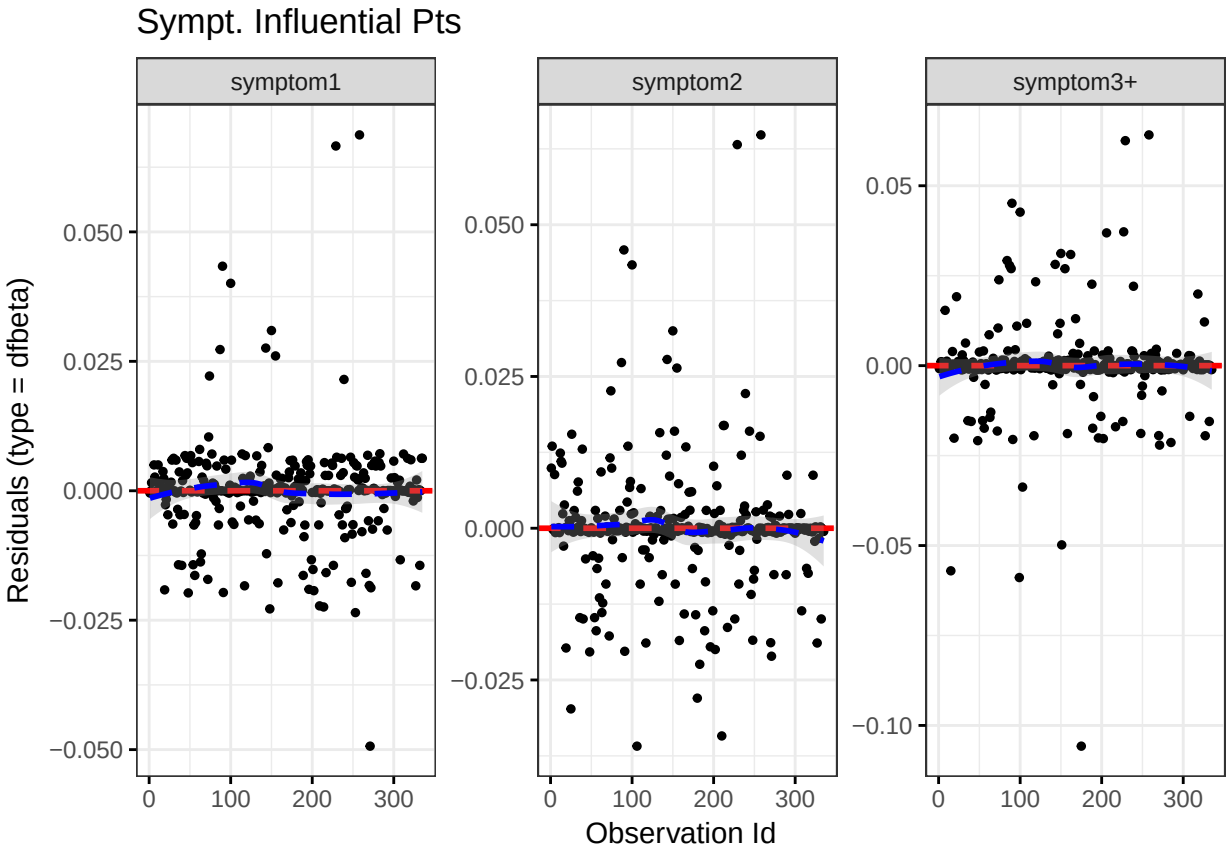

Race Influential Pts
raceOther
0.04
0.02
0.00
−0.02
Residuals (type = dfbeta)
0
100
200
300
Observation Id

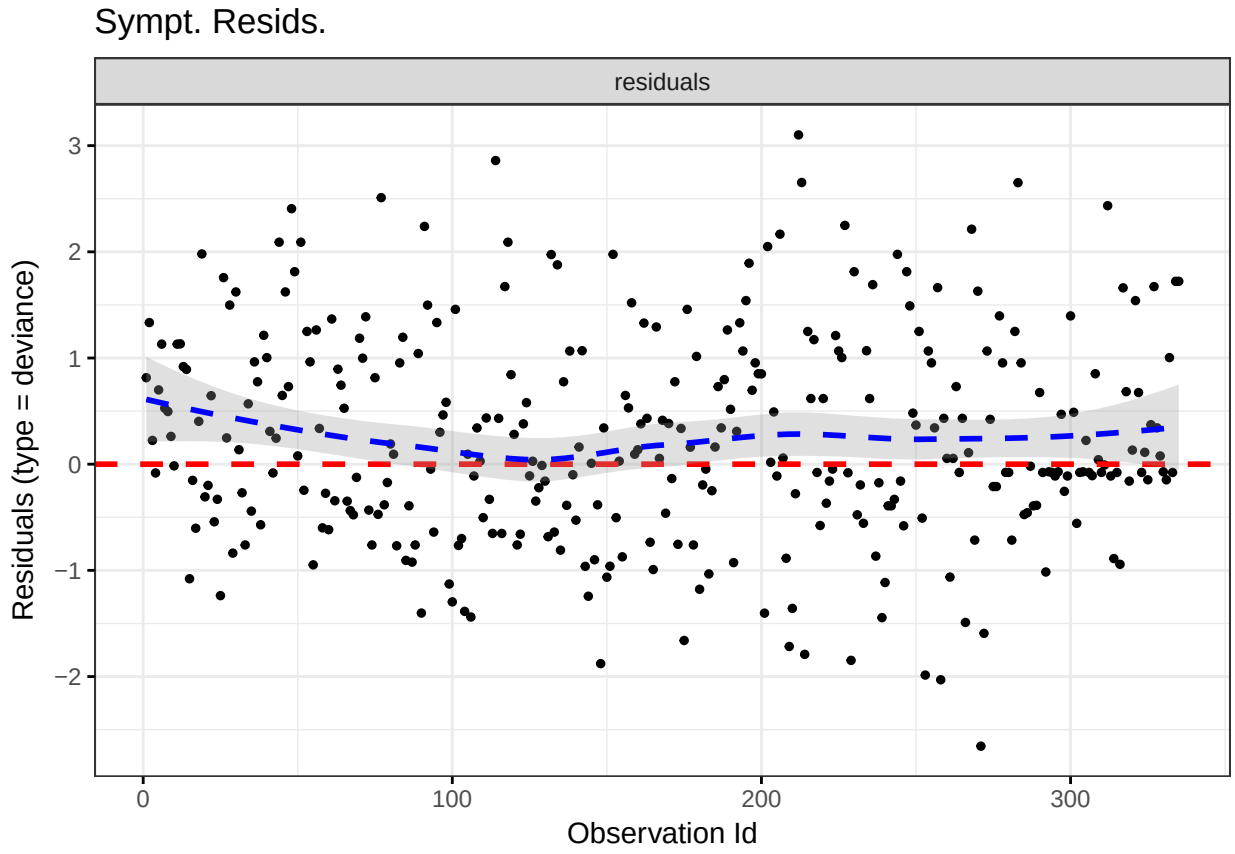



```
##              rho chisq      p
## raceOther -0.106  3.06 0.0802
```

Clinical Presentation

Looking at the diagnostics for the model clinical presentation influence on wait time to CT we do see different results than the results for gender and race/ethnicity above. Looking into the influential points, this time there are points as high as .05. We need to remove these points from the data since we consider these points affect the fit of the model. After removing the points looking at the residuals we see no pattern in the graph and a generally even spread number of points above and below 0. Since we see the residuals are independent of time. Finally we checked the assumption of proportional hazard. From the output, the test is not statistically significant (p=.0802) for each of the covariates; therefore, we can assume the proportional hazards.


Sympt. Influential Pts
symptom1
symptom2
symptom3+
Residuals (type = dfbeta)
0.050
0.025
0.000
−0.025
−0.050
0.05
0.00
−0.05
−0.10
0
100
200
300
Observation Id




```
##              rho chisq      p
## symptom1  0.0245 0.164 0.6855
## symptom2  0.1031 2.965 0.0851
## symptom3+ 0.0638 1.128 0.2882
## GLOBAL        NA 4.205 0.2402
```

Results

Looking at the summary we see if the patient is a male the wait time until CT scan will decrease by .1499 time units; however,

```
## Call:
## coxph(formula = Surv(nctdel, fail) ~ gender, data = dat)
##
##   n= 335, number of events= 277
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## gendermale -0.1499    0.8608   0.1207 -1.241    0.214
##
##            exp(coef) exp(-coef) lower .95 upper .95
## gendermale    0.8608      1.162    0.6794     1.091
##
## Concordance= 0.526  (se = 0.018 )
## Rsquare= 0.005   (max possible= 1 )
## Likelihood ratio test= 1.54  on 1 df,   p=0.2145
## Wald test            = 1.54  on 1 df,   p=0.2145
## Score (logrank) test = 1.54  on 1 df,   p=0.2141
```

```
## Call:
## coxph(formula = Surv(nctdel, fail) ~ race, data = dat)
##
##   n= 335, number of events= 277
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## raceOther 0.1673    1.1821   0.1333 1.255     0.21
##
##           exp(coef) exp(-coef) lower .95 upper .95
## raceOther     1.182      0.846    0.9103     1.535
##
## Concordance= 0.531  (se = 0.016 )
## Rsquare= 0.005   (max possible= 1 )
## Likelihood ratio test= 1.61  on 1 df,   p=0.2045
## Wald test            = 1.57  on 1 df,   p=0.2096
## Score (logrank) test = 1.58  on 1 df,   p=0.209
```

```
## Call:
## coxph(formula = Surv(nctdel, fail) ~ symptom, data = dat)
##
##   n= 335, number of events= 277
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## symptom1  0.1600    1.1735   0.1655 0.967    0.334
## symptom2  0.1561    1.1690   0.1947 0.802    0.422
## symptom3+ 0.3755    1.4557   0.2442 1.538    0.124
##
##           exp(coef) exp(-coef) lower .95 upper .95
## symptom1      1.173     0.8522    0.8484     1.623
## symptom2      1.169     0.8554    0.7982     1.712
## symptom3+     1.456     0.6870    0.9020     2.349
##
## Concordance= 0.521  (se = 0.019 )
## Rsquare= 0.007   (max possible= 1 )
## Likelihood ratio test= 2.37  on 3 df,   p=0.4986
## Wald test            = 2.41  on 3 df,   p=0.4915
## Score (logrank) test = 2.42  on 3 df,   p=0.4892
```

Contributions

Nathaniel built the logistic regression model.

References

http://influentialpoints.com/Training/coxs_proportional_hazards_modression_model-principles-properties-assumptions.htm#modmch

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/

http://dwoll.de/rexrepos/posts/survivalKM.html

http://www.sthda.com/english/wiki/cox-model-assumptions