Case 2 Part 2

Sonia Xu, Grant Goettel, Ian Hua

October 18, 2017

Overview: Model Explorations

Multiple models and analyses were explored to find the best model that fit the neurological assessment data. Methods explored included:

1. Cox Proportional Hazards Model
2. Kaplan-Meier Estimate
3. Random Forest
4. Kernel Regression

Cox Proportional Hazards Model

A Cox Proportional Hazards model was fit with nctdel as the response, and the features male, black, hispanic, and coun_sn as a factor (Appendix A for summary).

Looking at the Cox Model, only one of the coefficients are significant in detecting the nctdel waiting time–if the number of symptoms = 4. Either the data is not informative or the model is not a good fit of the data.

To assess the goodness of fit of the model, we check to see if the model fits the Cox Proportional Hazards Model Assumptions of no influential outliers, linearity, and homoscedascity of Schoenfeld residuals. The dataset does not satisfy the linearity assumption, so this implies that the Cox Proportional Hazard Model may not be the best model for the dataset (Appendix B).

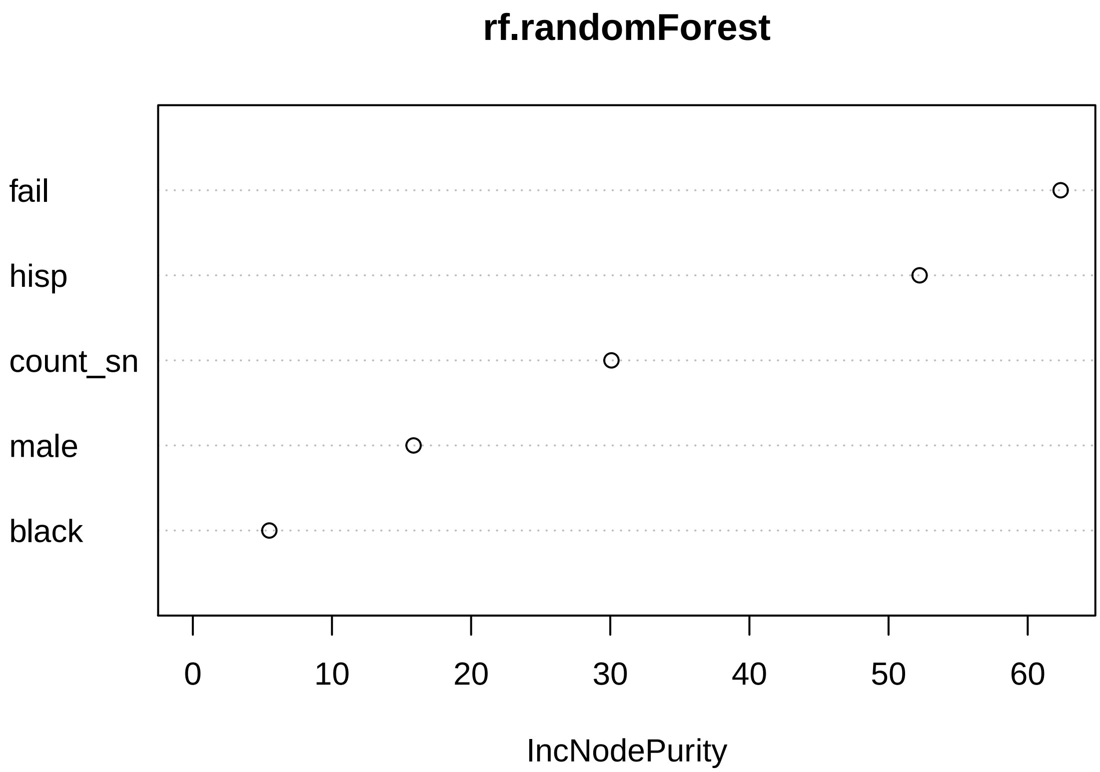
Kaplan-Meier Estimate

The Kaplan-Meier estimate can take into account the fail variable properly in the dataset. It is clear that sample size is an issue for some of categories but for the most part we can observe general relationships between categories in our dataset. It seems that females are treated slightly faster than males although not significantly. Non-blacks appear to be treated faster than blacks with a 95% CI for non-blacks between 1.33 and 1.62 while a 95% CI for blacks is between 1.68 and 2.08. This appears to be significant bias, however there appears to be no bias with regard to hispanics. Finally, by looking at the graph for the Kaplan-Meier estimate on the dataset based on count of symptoms, we observe that symptom count does appear to be a major factor in wait time, especially when all four symptoms are present (Appendix D).

Random Forest for Two Different Responses

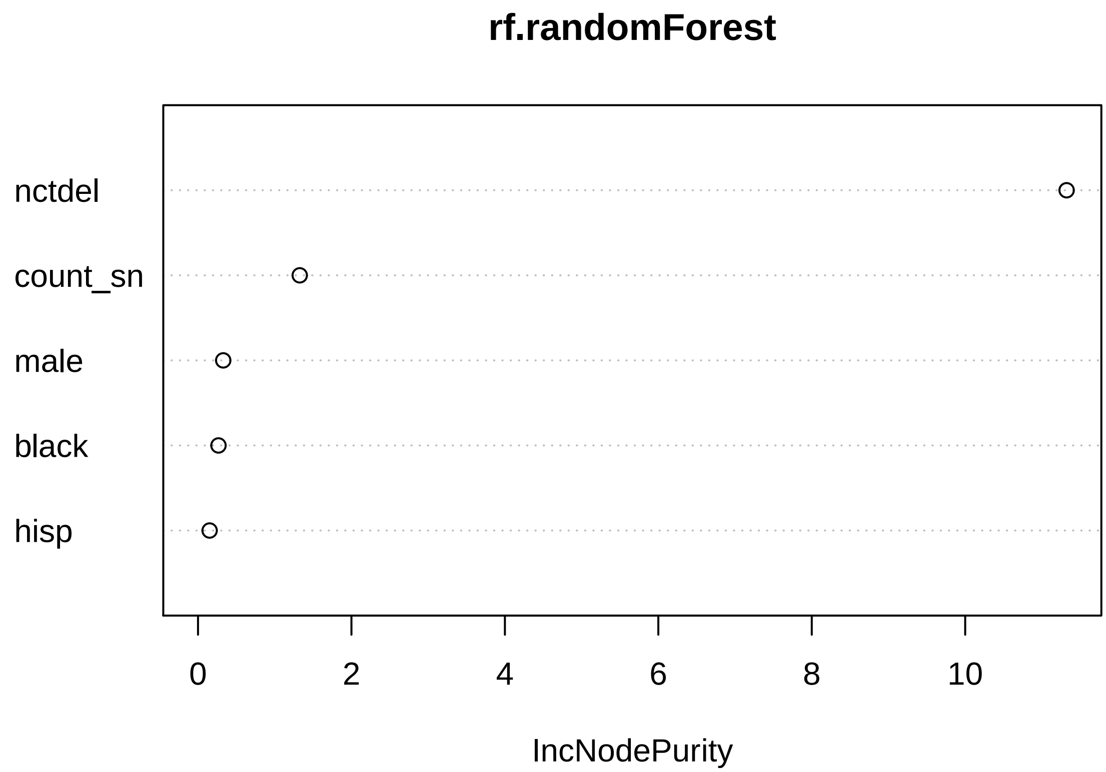
We built a random forest tree to identify the most significant features. In doing so, we realized that changing the response could change the significance ranking of features.

nctdel: Wait Time



Based on the variable importance plot, the most significant variables for determining nctdel wait time are in the order: fail, hispanic, count of symptoms, sex, and black.

Failure

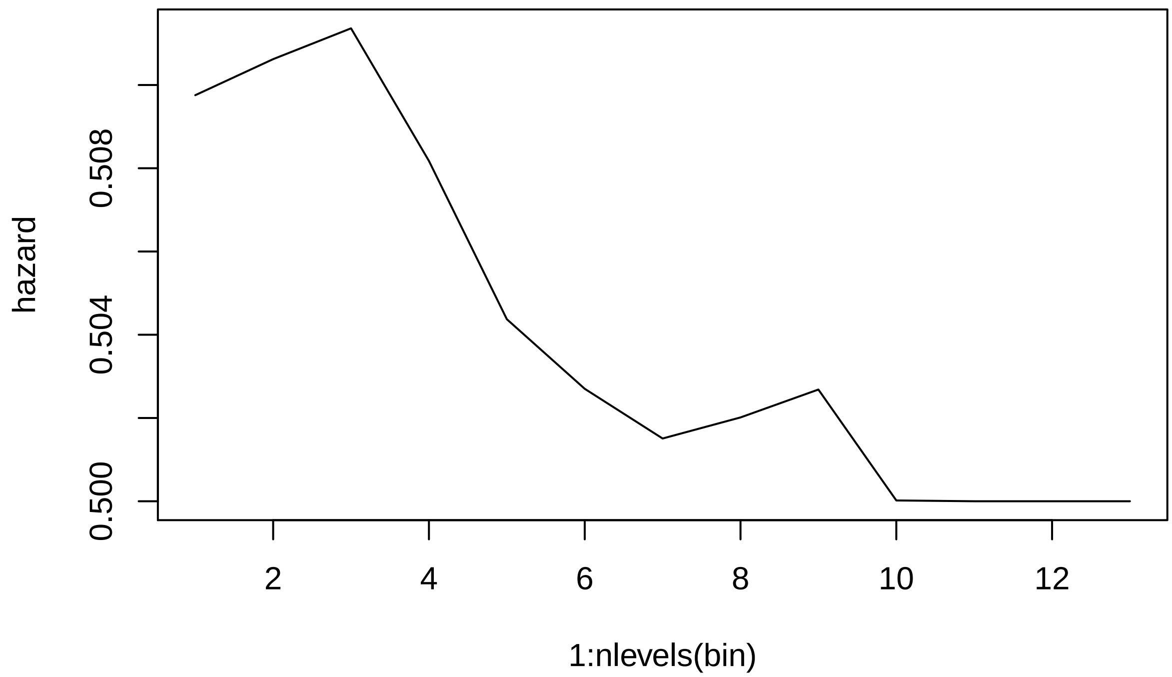


However, when changing the response to failure (0/1), the most significant predictors are nctdel, count_sn, and male. Being hispanic is less important.

Originally, for the CPH model and KM model, we used nctdel as the response, and noticed goodness-of-fit issues. We decided to create a Kernel Regression model with failure as the response to explore the robustness of this new model.

Kernel Regression with 13 Bins

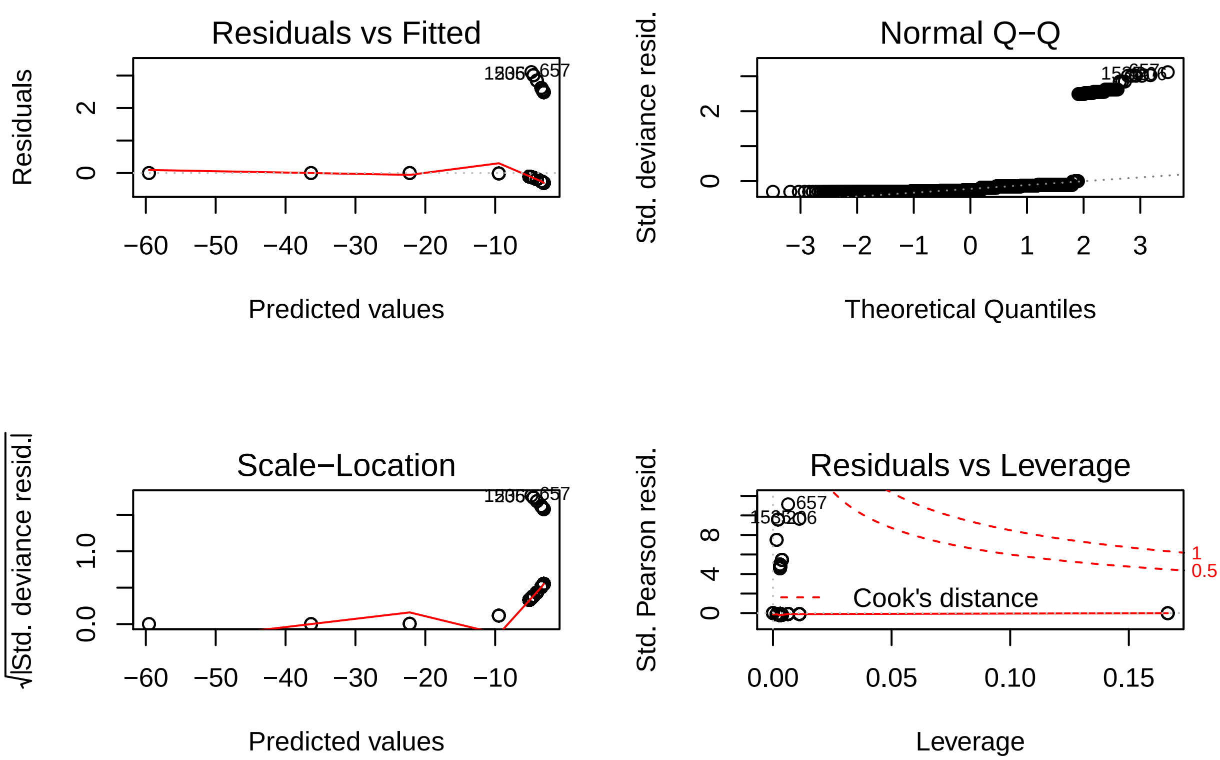
13 bins were calculated to fit the kernels. The bins are unevenly spaced because the data has a higher concentration of points for the feature nctdel between 0 and 2, even though its range is (0,26.25). The model has fail as the response, a kernel estimation of nctdel with 4 knots, and the features male, black, hispanic, and coun_sn. The bin levels are (-Inf,0], (0,0.3], (0.3,0.7], (0.7,1], (1,1.1], (1.1,1.3], (1.3,1.6], (1.6,1.9], (1.9,10], (10,13], (13,15], (17, Inf]. A summary of the model noted some significance for the feature nctdel (for the full summary, Appendix C).



Looking at the hazard plot, after bin 6 (1.1,1.3], the survival log odds are significantly lower.

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. However, the points are not evenly distrbuted, so they are heteroscedastic. Similarly, with Residuals vs. Leverage, the points are not homoscedastically distributed. While this model better fits the data than previous models, this model can also be improved.

Overall, the model fits the true dataset 71.641791% of the time when predicting for failure over the entire dataset, which reaffirms the fact that the model can be improved.

Conclusion

Overall, most of the models explored were average at best. The best model for the dataset currently is the Kernel Regression Model. For next week, more model exploration and testing will be conducted to improve the model fit.

Appendix

A

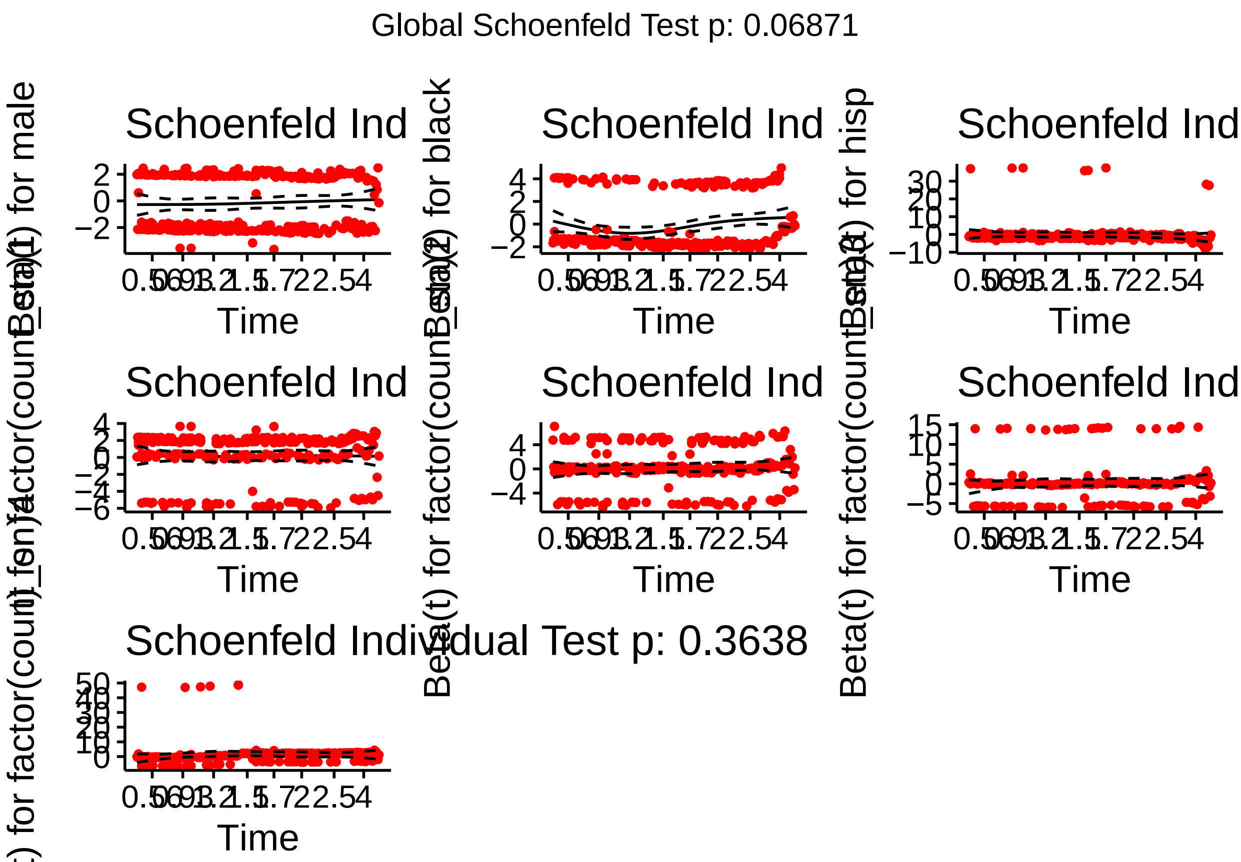
CPH Model Summary

```
## Call:
```

```
## coxph(formula = Surv(nctdel, fail) ~ male + black + hisp + factor(count_sn),
##     data = kelly)
##
##   n= 335, number of events= 277
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## male              -0.1418    0.8678   0.1219 -1.163  0.24467
## black             -0.1442    0.8657   0.1397 -1.032  0.30184
## hisp              -0.3072    0.7355   0.3713 -0.827  0.40799
## factor(count_sn)1  0.1660    1.1806   0.1663  0.998  0.31816
## factor(count_sn)2  0.1643    1.1786   0.1963  0.837  0.40269
## factor(count_sn)3  0.2007    1.2223   0.2673  0.751  0.45278
## factor(count_sn)4  1.2187    3.3829   0.4395  2.773  0.00556 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## male                 0.8678     1.1524    0.6833     1.102
## black                0.8657     1.1551    0.6584     1.138
## hisp                 0.7355     1.3596    0.3553     1.523
## factor(count_sn)1    1.1806     0.8471    0.8522     1.635
## factor(count_sn)2    1.1786     0.8485    0.8021     1.732
## factor(count_sn)3    1.2223     0.8182    0.7238     2.064
## factor(count_sn)4    3.3829     0.2956    1.4294     8.006
##
## Concordance= 0.541  (se = 0.021 )
## Rsquare= 0.028   (max possible= 1 )
## Likelihood ratio test= 9.5  on 7 df,   p=0.2186
## Wald test            = 11.16  on 7 df,   p=0.1318
## Score (logrank) test = 11.94  on 7 df,   p=0.1025
```

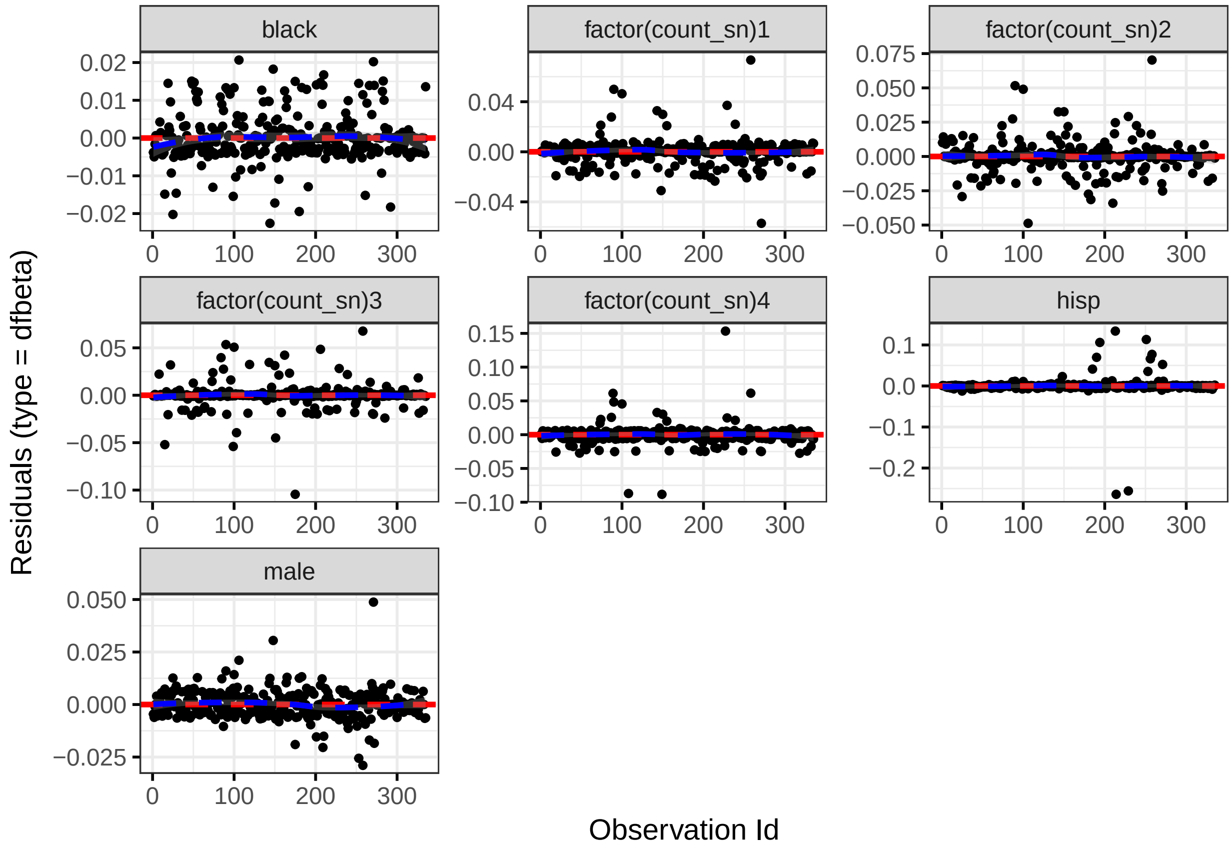
B

Schoenfeld Residuals



From the graphical inspection, there exists a pattern (slight curve in tails) with time for the feature black. The assumption of proportional hazards appears to be supported for the covariates male, each factor of the symptoms, and hispanic.

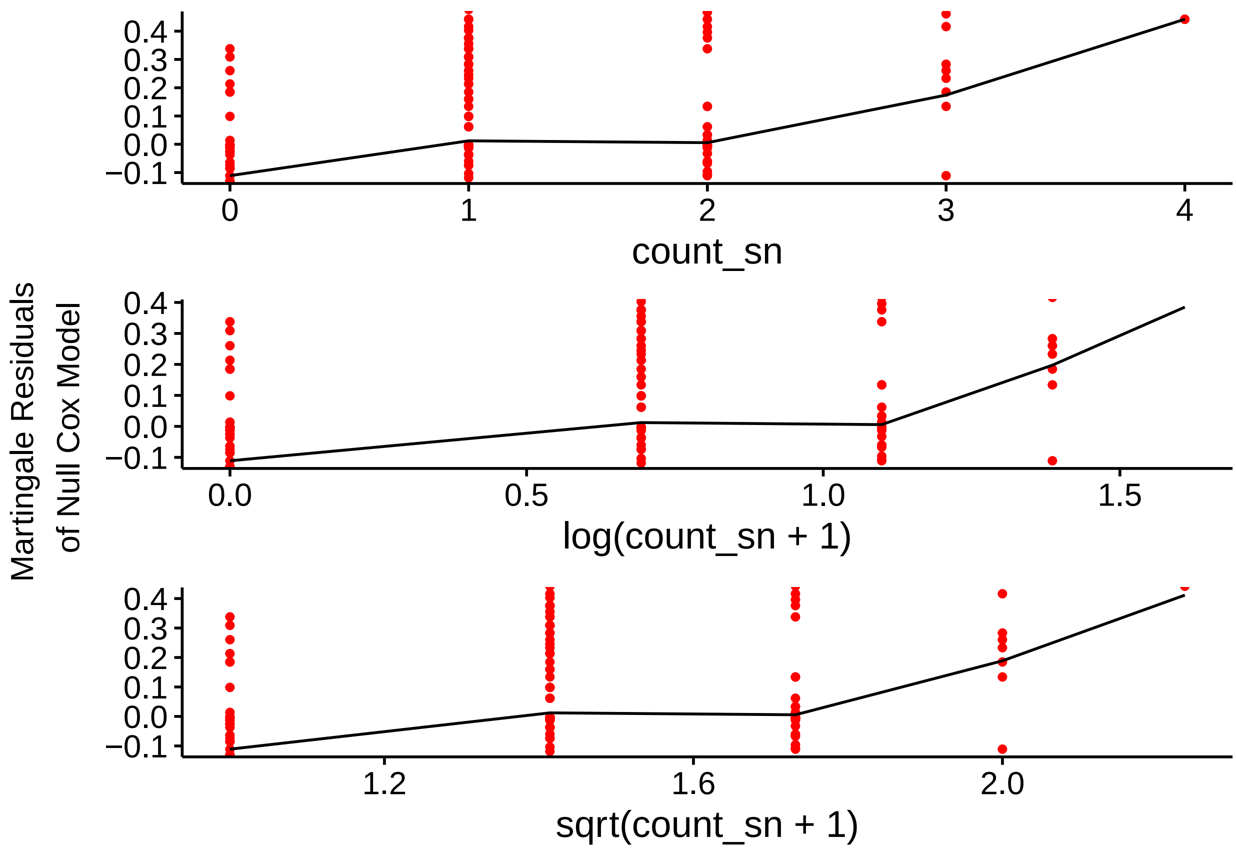
Test for Outliers



Most of the plots show no trends, so no points are significantly influential.

Linearity

```
## Warning: arguments formula is deprecated; will be removed in the next
## version; please use fit instead.
```



The feature count_sn does not follow a linear trend, so it breaks the linearity assumption for a CPH Model.

C

Kernel Regression Summary

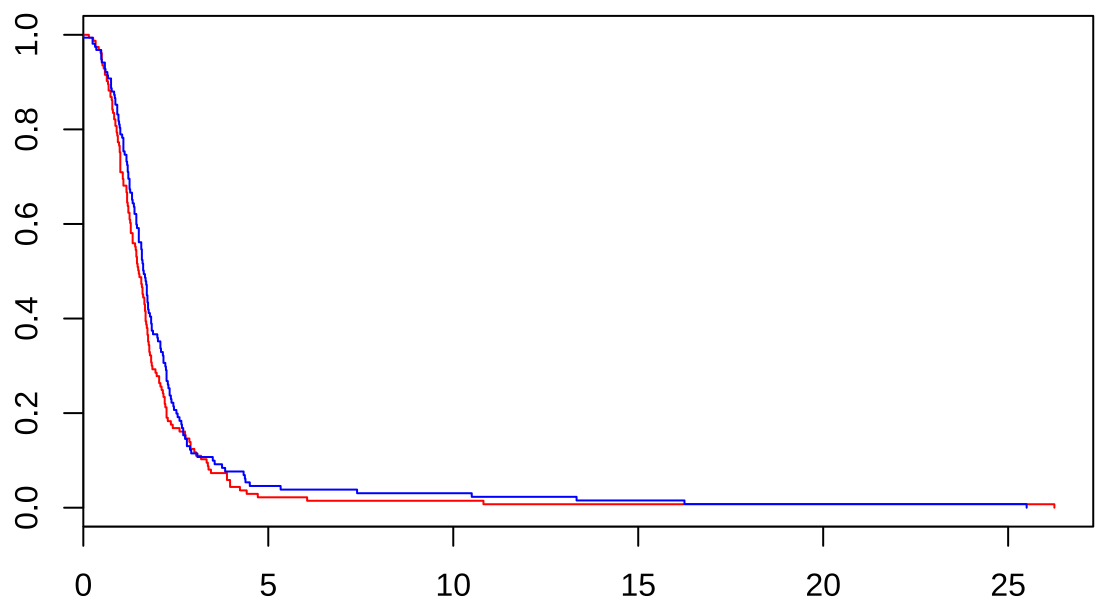
```
##
## Call:
## glm(formula = y ~ 0 + ., family = "binomial", data = data.frame(d2_full))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3050  -0.2947  -0.2579  -0.1474   3.1054
##
## Coefficients:
##          Estimate Std. Error z value Pr(>|z|)
## X1         3.9651     2.2936   1.729  0.08385 .
## X2        -3.5044     1.3006  -2.695  0.00705 **
## X3         5.1603     6.0816   0.849  0.39615
## X4       -24.0796    92.2309  -0.261  0.79403
## male      -3.3568     4.1214  -0.814  0.41538
## black      1.7453     5.8936   0.296  0.76713
## hisp      -0.8621     4.1038  -0.210  0.83361
## count_sn  -1.8818     2.2354  -0.842  0.39989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2828.04  on 2040  degrees of freedom
## Residual deviance:  507.79  on 2032  degrees of freedom
## AIC: 523.79
##
## Number of Fisher Scoring iterations: 13
```

D

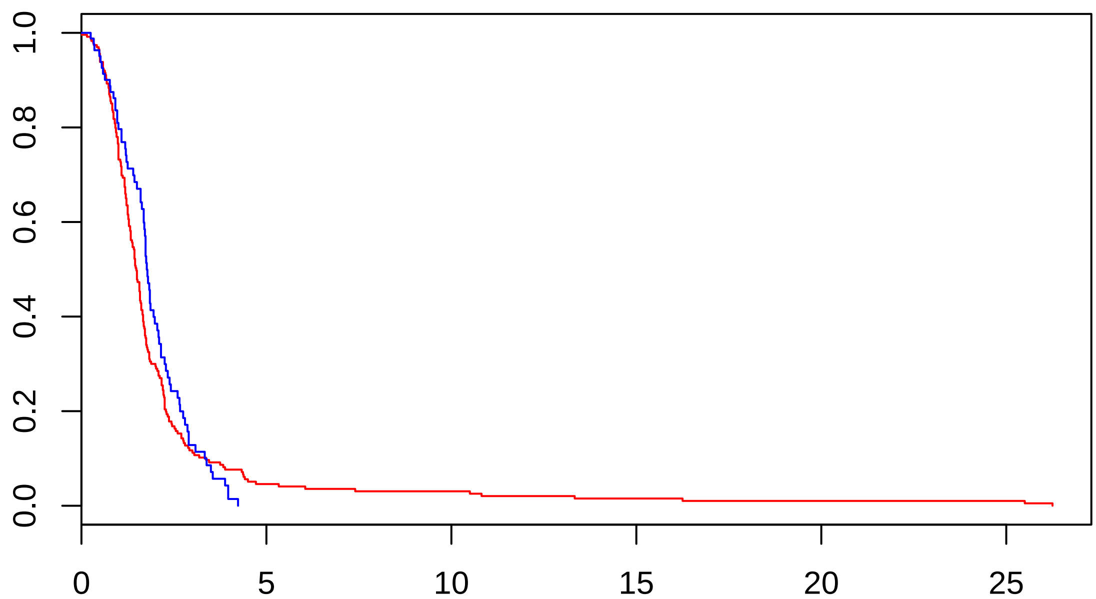
Kaplan Meier Estimate for SubCategories

The following is the Kaplan-Meier estimate for all subcategories present in the data set. Larger sample sizes

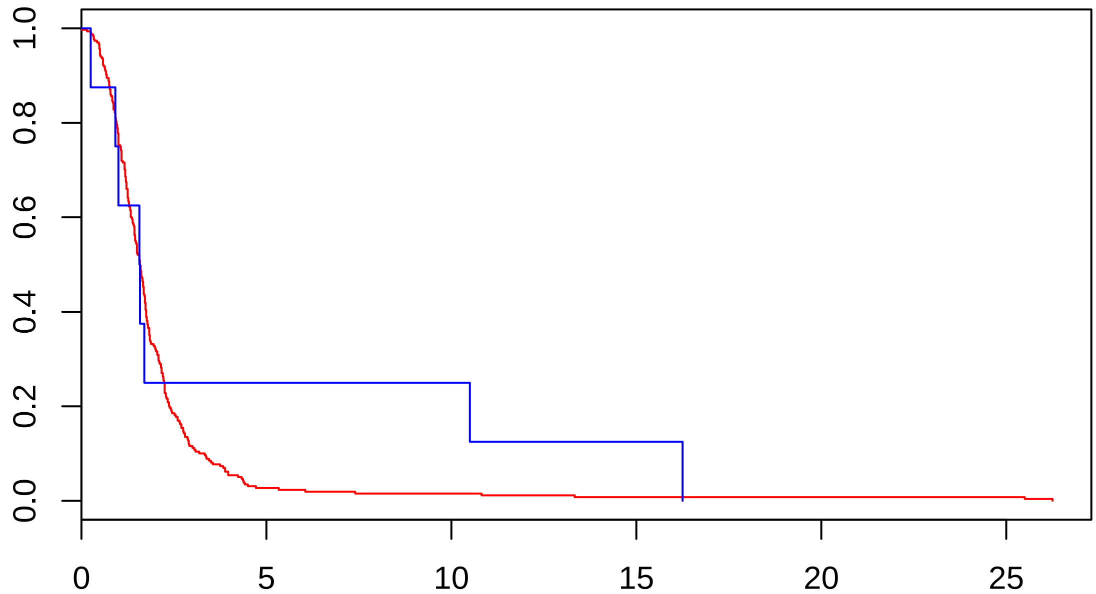


would be useful.

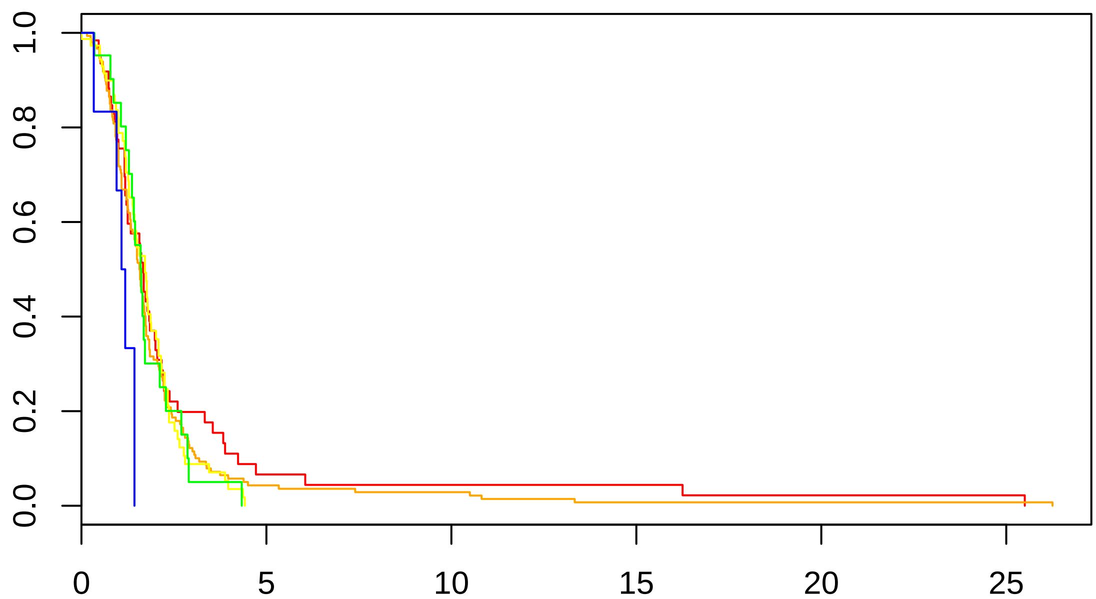
```
## Call: survfit(formula = d2 ~ kelly$male)
##
##                n events median 0.95LCL 0.95UCL
## kelly$male=0 171    141   1.50    1.33    1.68
## kelly$male=1 164    136   1.63    1.50    1.80
```




```
## Call: survfit(formula = d2 ~ kelly$black)
##
##                  n events median 0.95LCL 0.95UCL
## kelly$black=0 247    205   1.48    1.33    1.62
## kelly$black=1  88     72   1.77    1.68    2.08
```




```
## Call: survfit(formula = d2 ~ kelly$hisp)
##
##                  n events median 0.95LCL 0.95UCL
## kelly$hisp=0 326    269   1.58    1.45     1.7
## kelly$hisp=1   9      8   1.57    1.00      NA
```




```
## Call: survfit(formula = d2 ~ kelly$count_sn)
##
##                    n events median 0.95LCL 0.95UCL
## kelly$count_sn=0  69     50   1.67    1.25    2.00
## kelly$count_sn=1 162    142   1.57    1.43    1.70
## kelly$count_sn=2  77     59   1.72    1.43    2.02
## kelly$count_sn=3  21     20   1.62    1.37    2.28
## kelly$count_sn=4   6      6   1.13    0.95      NA
```

```
## Call: survfit(formula = d2 ~ kelly$male + kelly$black + kelly$hisp +
##     kelly$count_sn)
##
##                                                                   n events
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=0 28     24
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=1 57     47
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=2 28     21
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=3  8      8
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=4  3      3
## kelly$male=0, kelly$black=0, kelly$hisp=1, kelly$count_sn=0  1      1
## kelly$male=0, kelly$black=0, kelly$hisp=1, kelly$count_sn=1  3      3
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=0 11      8
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=1 16     14
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=2 11      7
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=3  4      4
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=4  1      1
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=0 23     13
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=1 58     53
```

```
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=2 24     20
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=3  7      6
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=4  2      2
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=0  1      1
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=1  2      2
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=2  2      1
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=0  5      3
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=1 26     23
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=2 12     10
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=3  2      2
##                                                              median 0.95LCL
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=0  1.250   1.000
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=1  1.567   1.267
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=2  1.283   1.183
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=3  1.433   1.283
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=4  1.083   0.333
## kelly$male=0, kelly$black=0, kelly$hisp=1, kelly$count_sn=0  1.567      NA
## kelly$male=0, kelly$black=0, kelly$hisp=1, kelly$count_sn=1  1.000   0.917
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=0  1.983   0.967
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=1  1.733   1.600
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=2  1.783   1.400
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=3  2.275   0.783
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=4  1.433      NA
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=0  1.683   1.583
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=1  1.383   1.217
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=2  1.883   1.267
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=3  1.667   1.600
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=4  1.192   0.950
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=0 16.250      NA
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=1  6.042   1.583
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=2  0.250      NA
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=0  2.383   0.967
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=1  1.750   1.250
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=2  2.150   1.717
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=3  0.775   0.350
##                                                              0.95UCL
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=0     2.17
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=1     1.75
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=2     1.80
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=3       NA
## kelly$male=0, kelly$black=0, kelly$hisp=0, kelly$count_sn=4       NA
## kelly$male=0, kelly$black=0, kelly$hisp=1, kelly$count_sn=0       NA
## kelly$male=0, kelly$black=0, kelly$hisp=1, kelly$count_sn=1       NA
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=0       NA
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=1     3.38
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=2       NA
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=3       NA
## kelly$male=0, kelly$black=1, kelly$hisp=0, kelly$count_sn=4       NA
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=0       NA
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=1     1.58
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=2     2.33
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=3       NA
## kelly$male=1, kelly$black=0, kelly$hisp=0, kelly$count_sn=4       NA
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=0       NA
```

```
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=1      NA
## kelly$male=1, kelly$black=0, kelly$hisp=1, kelly$count_sn=2      NA
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=0      NA
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=1    2.28
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=2      NA
## kelly$male=1, kelly$black=1, kelly$hisp=0, kelly$count_sn=3      NA
```

Contributions

Grant: Ian: Kaplan Meier Sonia: Everything else