Assumptions

Steven Herrera and Ethan Shen

11/09/2018

Logistic Regression Assumptions

After looking at exploratory data analysis, found on our Project RMD file, we have concluded with the following model that we will use as our final model:

```
newten <- ten
final <- glm(status ~ minutes + ht + rankpoints + ace + df, family = binomial,
             data = newten)
kable(tidy(final), format = "markdown", digits = 6)
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.197141	1.950891	2.663983	0.007722
minutes	-0.008897	0.002395	-3.714847	0.000203
ht	-0.023129	0.010416	-2.220635	0.026376
rankpoints	0.000184	0.000035	5.237843	0.000000
ace	0.094100	0.017519	5.371298	0.000000
df	-0.169342	0.035503	-4.769779	0.000002

Model Assessment

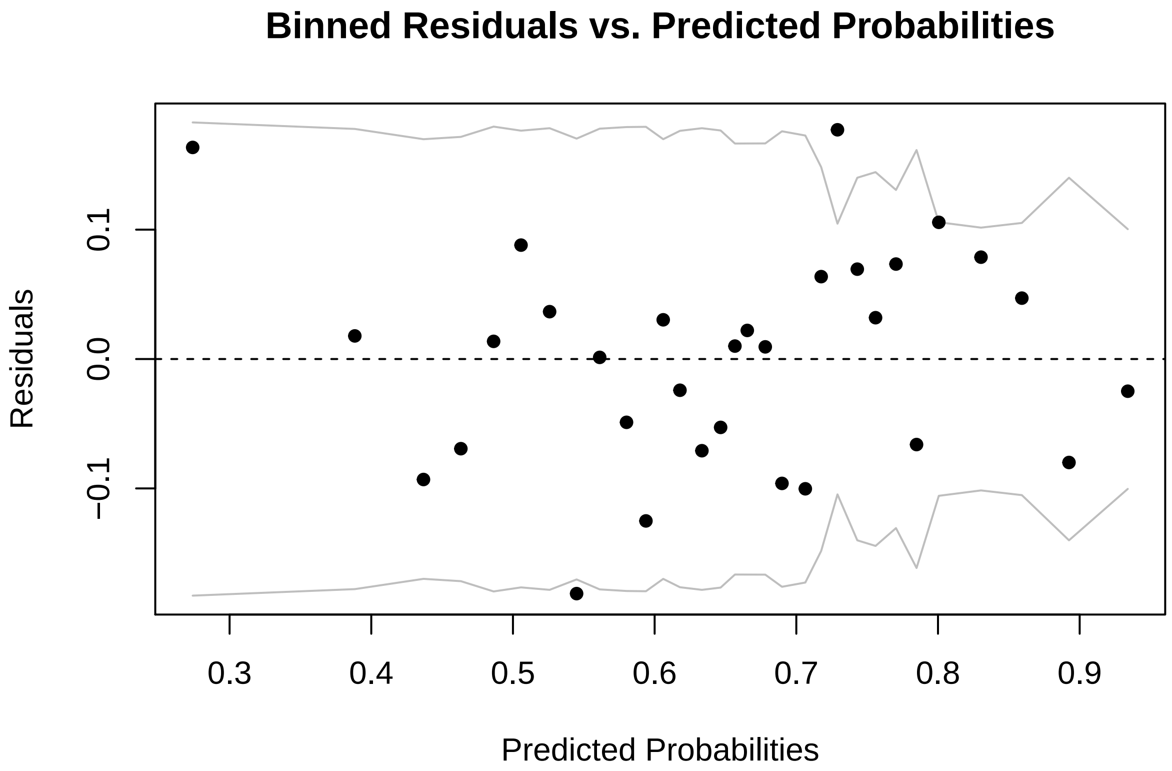
Binned Plots with Residuals vs Predicted

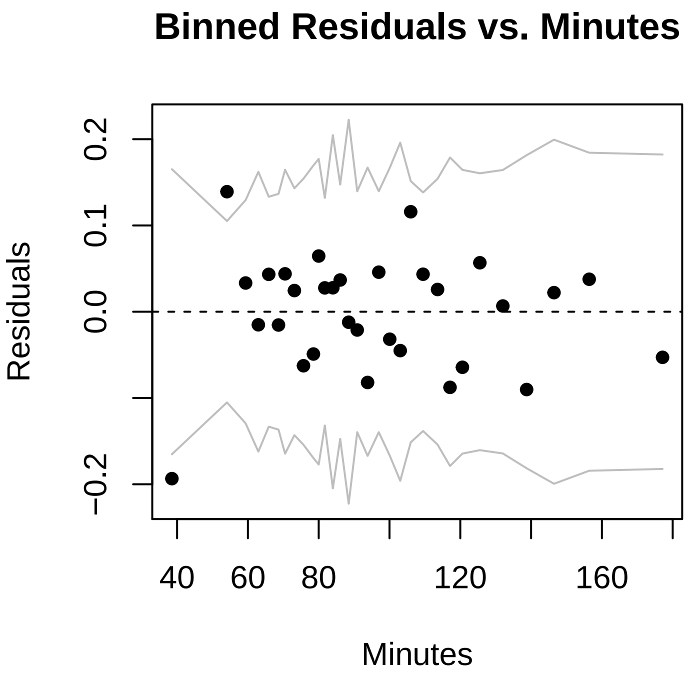
We will further our investigation of whether our new model follows the key model assessment characteristics:

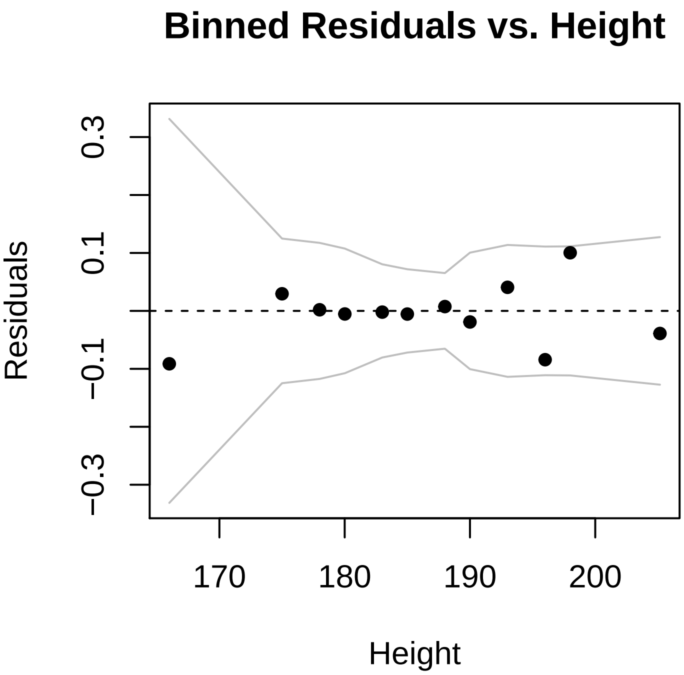
- Good binned residual vs. predicted plot
- Good binned residual vs. numerical explanatory plots
- Large area under the ROC curve

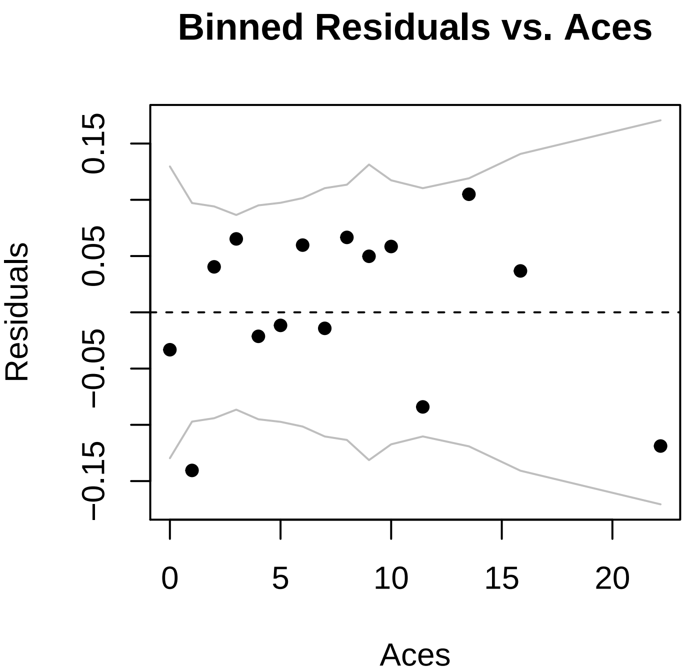
```
newten <- newten %>% mutate(Residuals = residuals.glm(final,type="response"),
                          Predicted = predict.glm(final,type="response"))

binnedplot(newten$Predicted, newten$Residuals,xlab="Predicted Probabilities",
           ylab="Residuals",main="Binned Residuals vs. Predicted Probabilities")
```

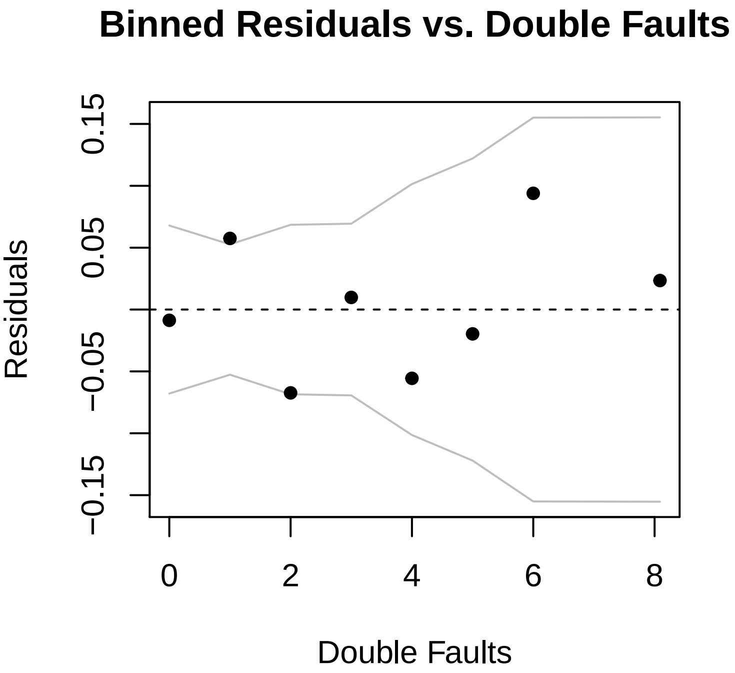

Binned Residuals vs. Predicted Probabilities
Residuals
0.1
0.0
−0.1
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Predicted Probabilities


Binned Residuals vs. Minutes
Residuals
0.2
0.1
0.0
−0.2
40
60
80
120
160
Minutes

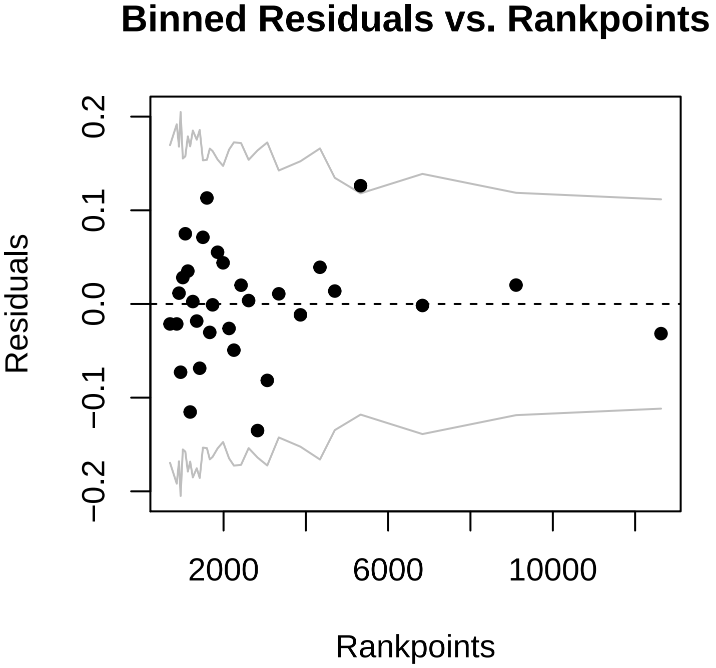

Binned Residuals vs. Height
Residuals
0.3
0.1
−0.1
−0.3
170
180
190
200
Height



Binned Residuals vs. Aces

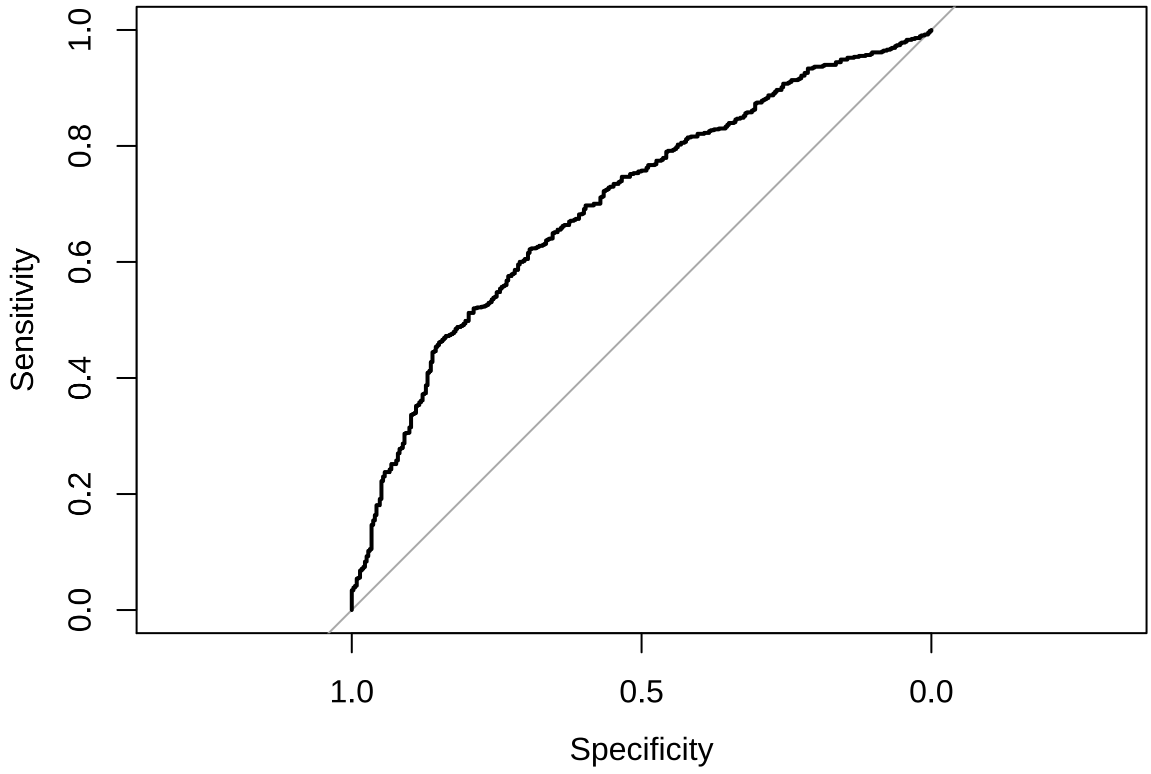


Binned Residuals vs. Double Faults



Binned Residuals vs. Rankpoints

```
ROC.newten <- roc(newten$status,newten$Predicted,plot=T)
```



```
ROC.newten$auc
```

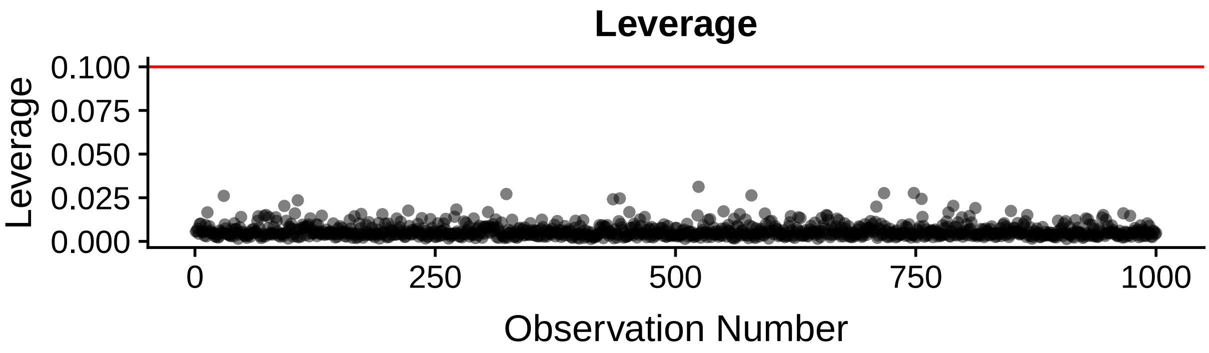
```
## Area under the curve: 0.6996
```

Influential Points

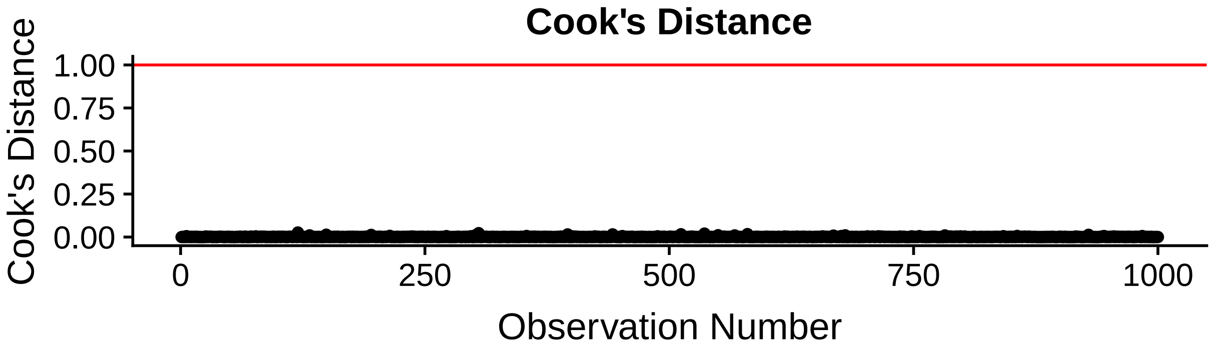
```
newten <- newten %>%
  mutate(leverage = hatvalues(final),
         cooks = cooks.distance(final),
         stand.resid = rstandard(final),
         obs.num = row_number())
```

Leverage and Cook's Distance

```
ggplot(data=newten, aes(x=obs.num,y=leverage)) +
  geom_point(alpha=0.5) +
  geom_hline(yintercept=0.1,color="red")+
  labs(x="Observation Number",y="Leverage",title="Leverage")
```



```r
ggplot(data=newten, aes(x=obs.num,y=cooks)) +
  geom_point() +
  geom_hline(yintercept=1,color="red")+
  labs(x="Observation Number",y="Cook's Distance",title="Cook's Distance")
```



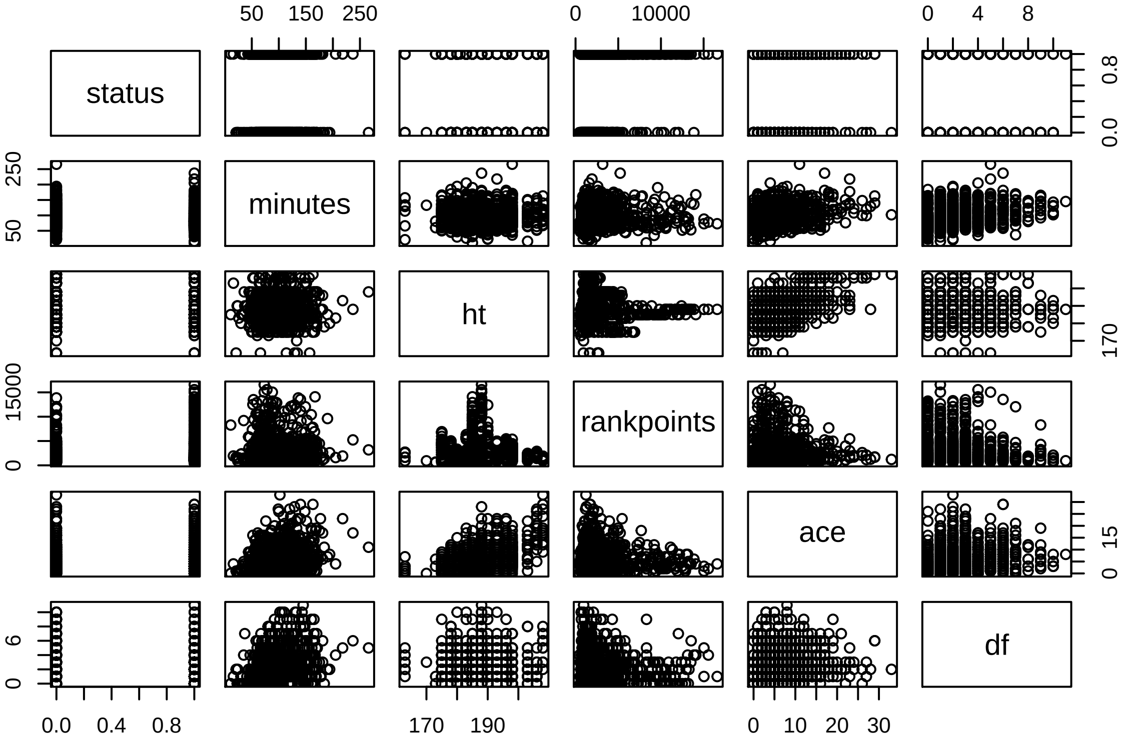
VIF and Multicollinearity

```r
tidy(vif(final))
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")
```

```
## # A tibble: 5 x 2
##   names          x
##   <chr>      <dbl>
## 1 minutes     1.19
## 2 ht          1.28
## 3 rankpoints  1.02
## 4 ace         1.41
## 5 df          1.08
```

```r
pairs(status ~ minutes + ht + rankpoints + ace + df, data = newten)
```



Conclusion

With VIF values less than 10 and a scatterplot matrix that does not show obvious linear relationships between the explanatory variables, observations that are under the leverage and Cook's distance threshold, and binned residual plots that satisfy the assumptions, we have cleared model assessment and assumptions for the following final model:

```
kable(tidy(final), format = "markdown", digits = 6)
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.197141	1.950891	2.663983	0.007722
minutes	-0.008897	0.002395	-3.714847	0.000203
ht	-0.023129	0.010416	-2.220635	0.026376
rankpoints	0.000184	0.000035	5.237843	0.000000
ace	0.094100	0.017519	5.371298	0.000000
df	-0.169342	0.035503	-4.769779	0.000002