# ST525 HW 8

Nora Quick

## Question 1

The term proportional hazard means that the event of an event happening is the same for all individuals but depending on covariates such as taking a drug to prevent cancer will change the likelihood of the hazard. In other words someone's hazard at time t depends on many different factors that could increase or decrease the hazard.

We can check whether the proportional hazard assumption is valid by graphing the data. It is easiest to see if the assumptions are met when the hazard rate is linear and the survival probability of two variables is similar but usually one is slightly higher/lower than the other.

## Question 2

### Part (a)

Model 1: $h(t) = h_0(t)exp(\beta_1 P27 + \beta_2 CYCLINE + \beta_3 NODES + \beta_4 SIZE2 + \beta_5 SIZE3 + \beta_6 Age + \beta_7 Year)$

Model 2: $h(t) = h_0(t)exp(\beta_1 P27 + \beta_2 CYCLINE + \beta_3 NODES + \beta_4 SIZE2 + \beta_5 SIZE3 + \beta_6 Age + \beta_7 Year + \beta_8(NODES * P27) + \beta_9(NODES * CYCLINE))$

### Part (b)

For coefficient NODES*CYCLINE shows that the estimated risk of death is 0.7718 (exp(-0.259)=0.7718) times lower for patients with no cancer spread to their lymph nodes and a normal protein cycline.

### Part (c)

I would normally say the wald statistic and while I think it is still relevant based on the output we're given the p-value seems to be the best indication of significance. For NODESxP27 the p-statistic is 0.96 indicating no significance with a DF of 1. For NODESxCYCLINE it has a p-value of 0.58 which is also not statistically significant with a DF of 1.

### Part (d)

```
(5+5+(1.357*5)+(0.664*5)+(0.867*5)+(-0.125*5)+(0.198*5)+(-0.027*5)+(-0.259*5))/45
```

```
## [1] 0.5194444
```

### Part (e)

I would use the estimates, wald statistic, and p-value to determine which fit is better. Each model has only 1 degree of freedom for each covaraite.

Model 1 has 5/7 p-values with significant values and model 2 has 5/9 p-values with significance. Both have only 1 DF for each covariate and the wald test reflects the p-values. Therefore, I would conclude that model 1 is the better for for the data.

### Part (f)

I would do a ggplot2 survival plot to see the pattern. I think this would be because because we want to see the survival of patients with the different covariates and it could be a great visual way to see what causes faster deteriation.

I would also likely plot of normal plot from the data to see if there was a pattern.

Additionally, I know we've done plots for KM and after doing some reading for week 10's first discission it would be a good decision to plot that as well.

## Question 3

```
rossi <- read.csv('Rossi.csv')
head(rossi)
```

```
##   id week arrest fin age  race wexp        mar paro prio educ
## 1  1   20      1   0  27 black    0 notmarried    1    3    3
## 2  2   17      1   0  18 black    0 notmarried    1    8    4
## 3  3   25      1   0  19 other    1 notmarried    1   13    3
## 4  4   52      0   1  23 black    1    married    1    1    5
## 5  5   52      0   0  19 other    1 notmarried    1    3    3
## 6  6   52      0   0  24 black    1 notmarried    0    2    4
```

### Part (a)

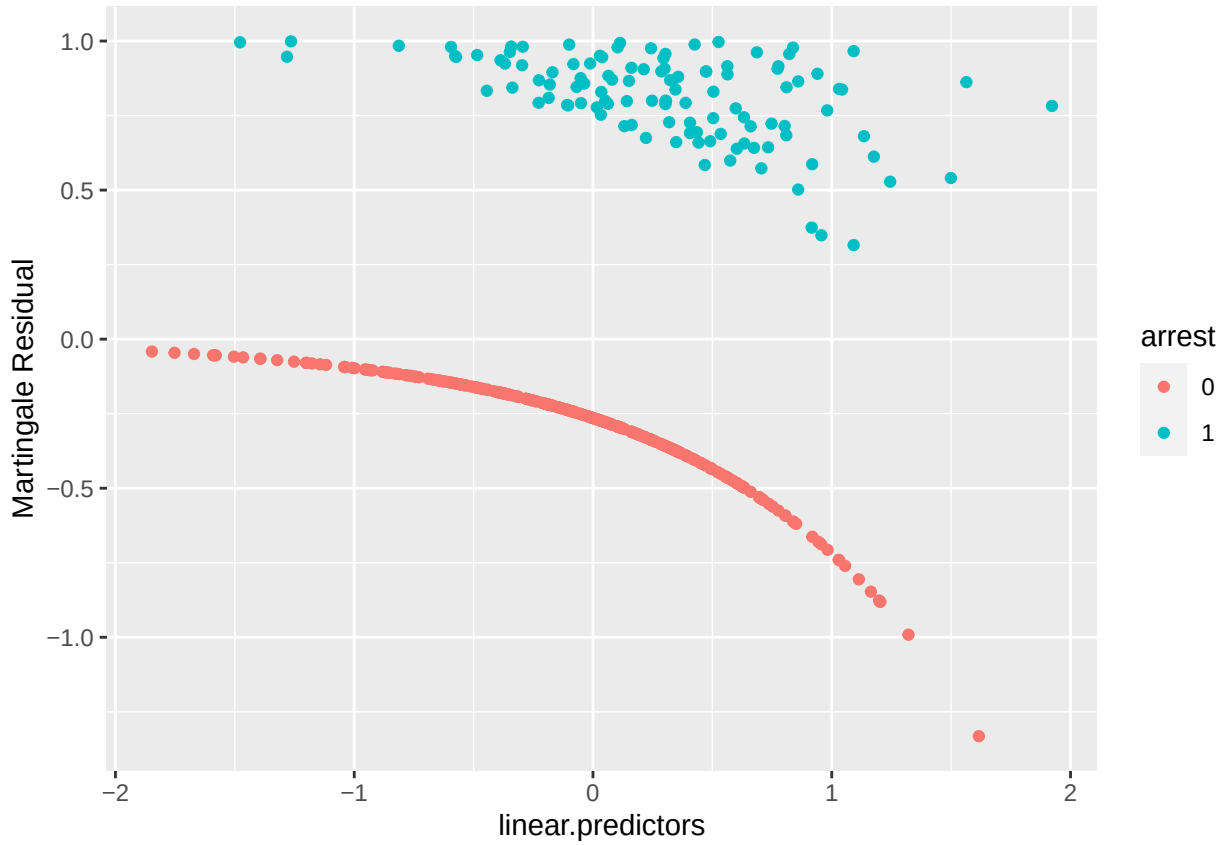
```
fit1 <- coxph(formula = Surv(week , arrest) ~ fin + age + race + wexp + mar + paro + prio + educ,
              data = rossi)
summary(fit1)
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
##     mar + paro + prio + educ, data = rossi)
##
##   n= 432, number of events= 114
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## fin          -0.35963   0.69794  0.19180 -1.875  0.06079 .
## age          -0.05768   0.94395  0.02187 -2.638  0.00835 **
## raceother    -0.34554   0.70784  0.30907 -1.118  0.26356
```

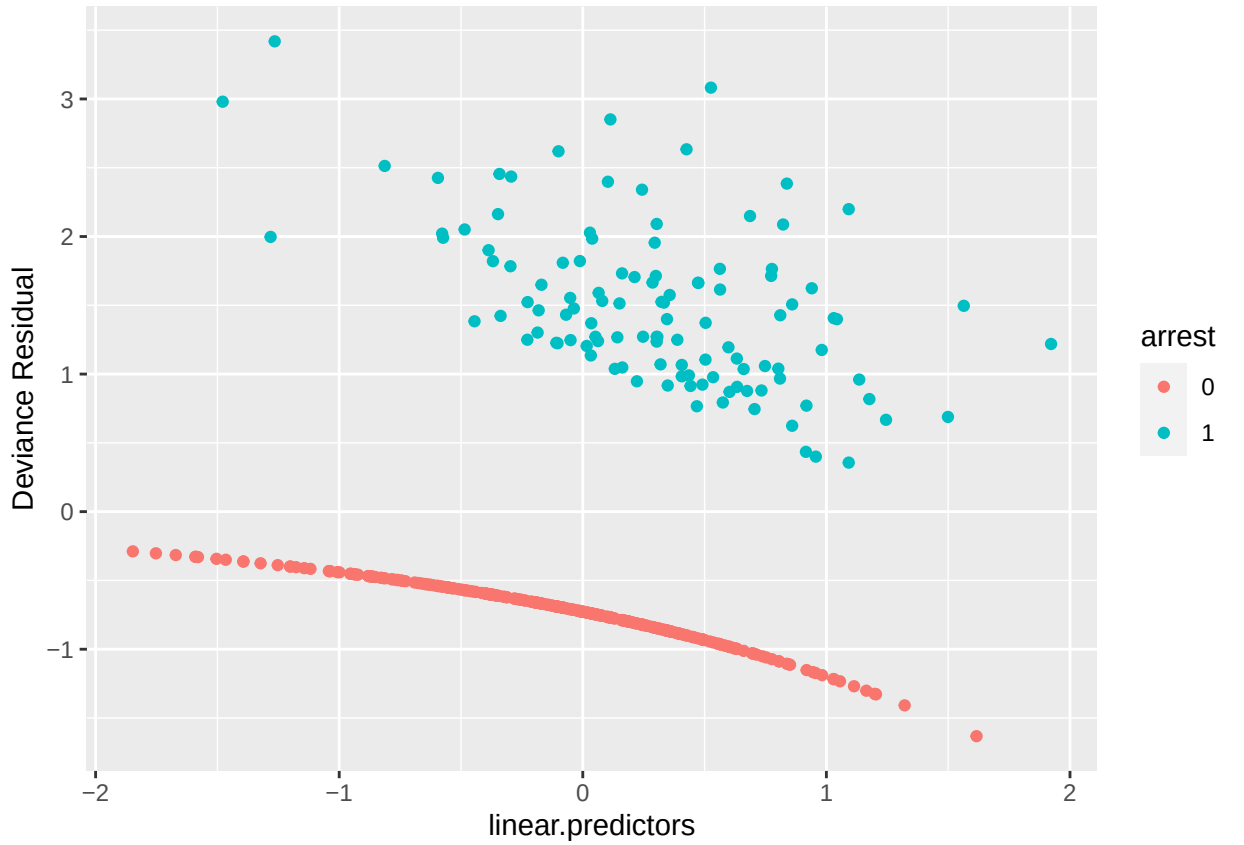
```
## wexp          -0.11439   0.89191  0.21311 -0.537  0.59145
## marnotmarried  0.42496   1.52953  0.38209  1.112  0.26605
## paro          -0.08991   0.91401  0.19568 -0.459  0.64589
## prio           0.08469   1.08838  0.02919  2.902  0.00371 **
## educ          -0.18578   0.83046  0.13153 -1.412  0.15782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## fin              0.6979     1.4328    0.4792    1.0164
## age              0.9440     1.0594    0.9044    0.9853
## raceother        0.7078     1.4128    0.3862    1.2972
## wexp             0.8919     1.1212    0.5874    1.3543
## marnotmarried    1.5295     0.6538    0.7233    3.2344
## paro             0.9140     1.0941    0.6229    1.3413
## prio             1.0884     0.9188    1.0279    1.1525
## educ             0.8305     1.2042    0.6417    1.0747
##
## Concordance= 0.656  (se = 0.026 )
## Likelihood ratio test= 35.35  on 8 df,   p=2e-05
## Wald test            = 33.74  on 8 df,   p=5e-05
## Score (logrank) test = 35.1  on 8 df,   p=3e-05
```

## Part (b)

```
rossi$res_mar <- residuals(fit1, type = 'martingale')
rossi$res_dev <- residuals(fit1, type = 'deviance')
rossi$linear.predictors <- fit1$linear.predictors
ggplot(data = rossi) + geom_point(aes(x = linear.predictors, y = res_mar,
                                      color = factor(arrest))) +
  ylab('Martingale Residual') + labs(color='arrest')
```




```
ggplot(data = rossi) + geom_point(aes(x = linear.predictors, y = res_dev,
                                      color = factor(arrest))) +
  ylab('Deviance Residual') + labs(color='arrest')
```




The Matringale should be more linear based on the lecutres and while the deviance curve is more linear I would say that this is not the best fit.

## Part (c)

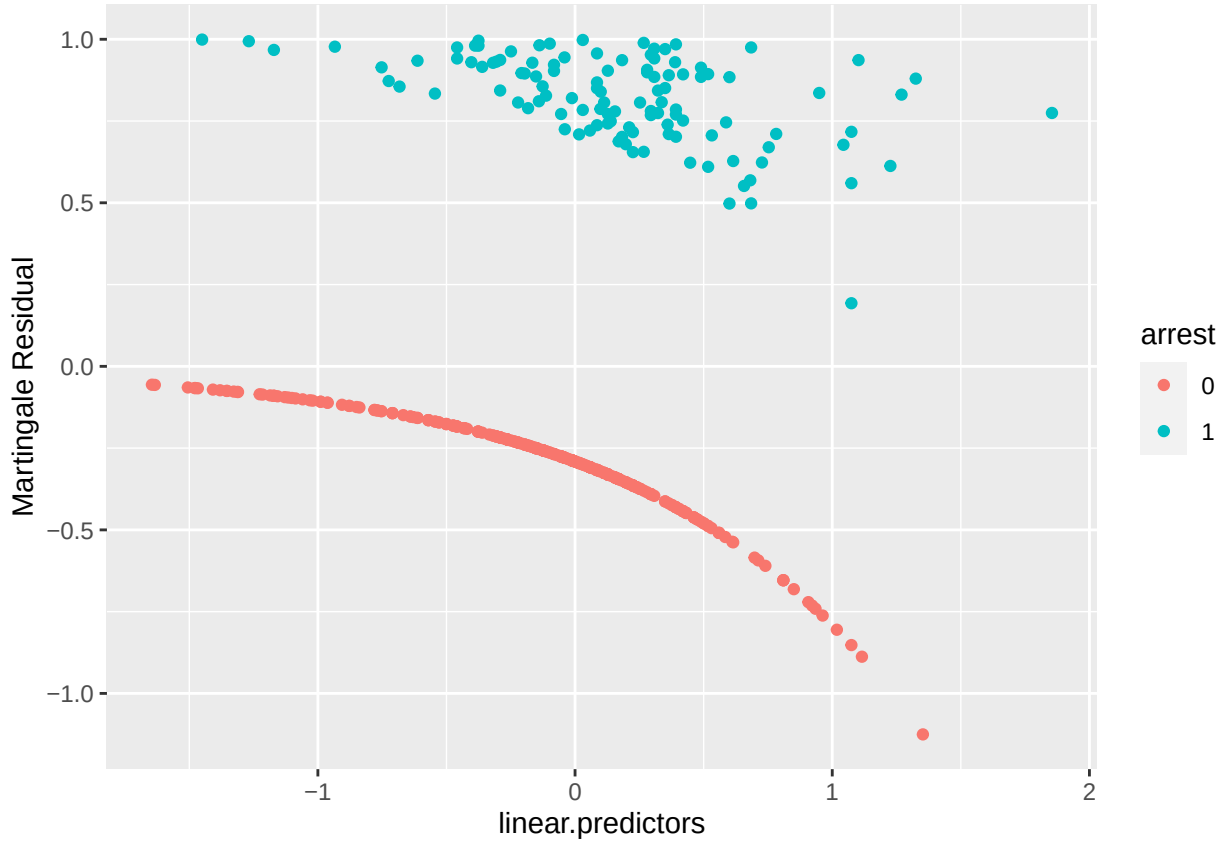
```
fit2 <- coxph(formula = Surv(week , arrest) ~ age + race + prio, data = rossi)
summary(fit2)
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ age + race + prio, data = rossi)
##
##   n= 432, number of events= 114
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## age       -0.06980   0.93258  0.02089 -3.341 0.000834 ***
## raceother -0.32340   0.72368  0.30643 -1.055 0.291244
## prio       0.09747   1.10238  0.02700  3.610 0.000306 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          0.9326     1.0723    0.8952    0.9716
## raceother    0.7237     1.3818    0.3969    1.3194
## prio         1.1024     0.9071    1.0456    1.1623
```

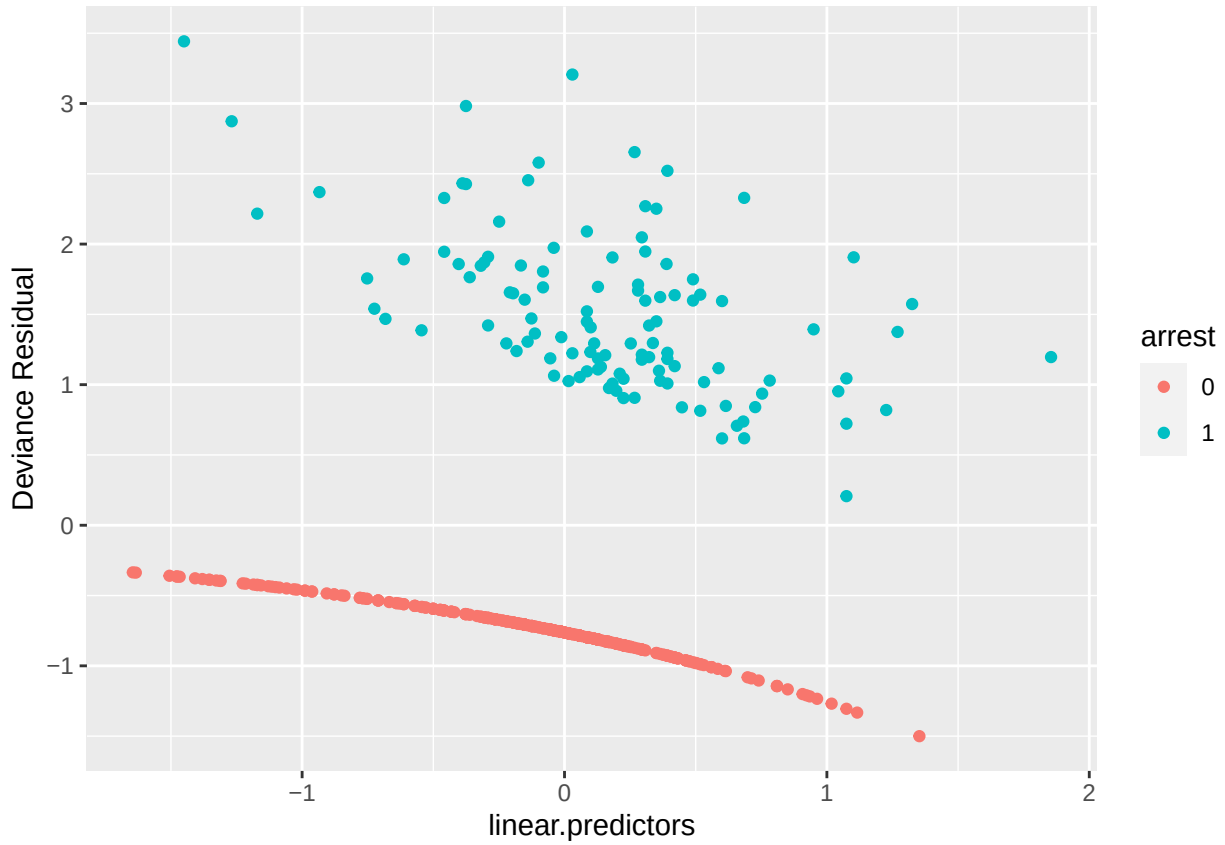
```
##
## Concordance= 0.631  (se = 0.028 )
## Likelihood ratio test= 26.89  on 3 df,   p=6e-06
## Wald test            = 25.66  on 3 df,   p=1e-05
## Score (logrank) test = 26.78  on 3 df,   p=7e-06
```

## Part (d)

```
rossi$res_mar <- residuals(fit2, type = 'martingale')
rossi$res_dev <- residuals(fit2, type = 'deviance')
rossi$linear.predictors <- fit2$linear.predictors
ggplot(data = rossi) + geom_point(aes(x = linear.predictors, y = res_mar,
                                      color = factor(arrest))) +
  ylab('Martingale Residual') + labs(color='arrest')
```




```
ggplot(data = rossi) + geom_point(aes(x = linear.predictors, y = res_dev,
                                      color = factor(arrest))) +
  ylab('Deviance Residual') + labs(color='arrest')
```




Based on these graphs as well it doesn't appear to be the best fit.