CoxPH

Amy Watt

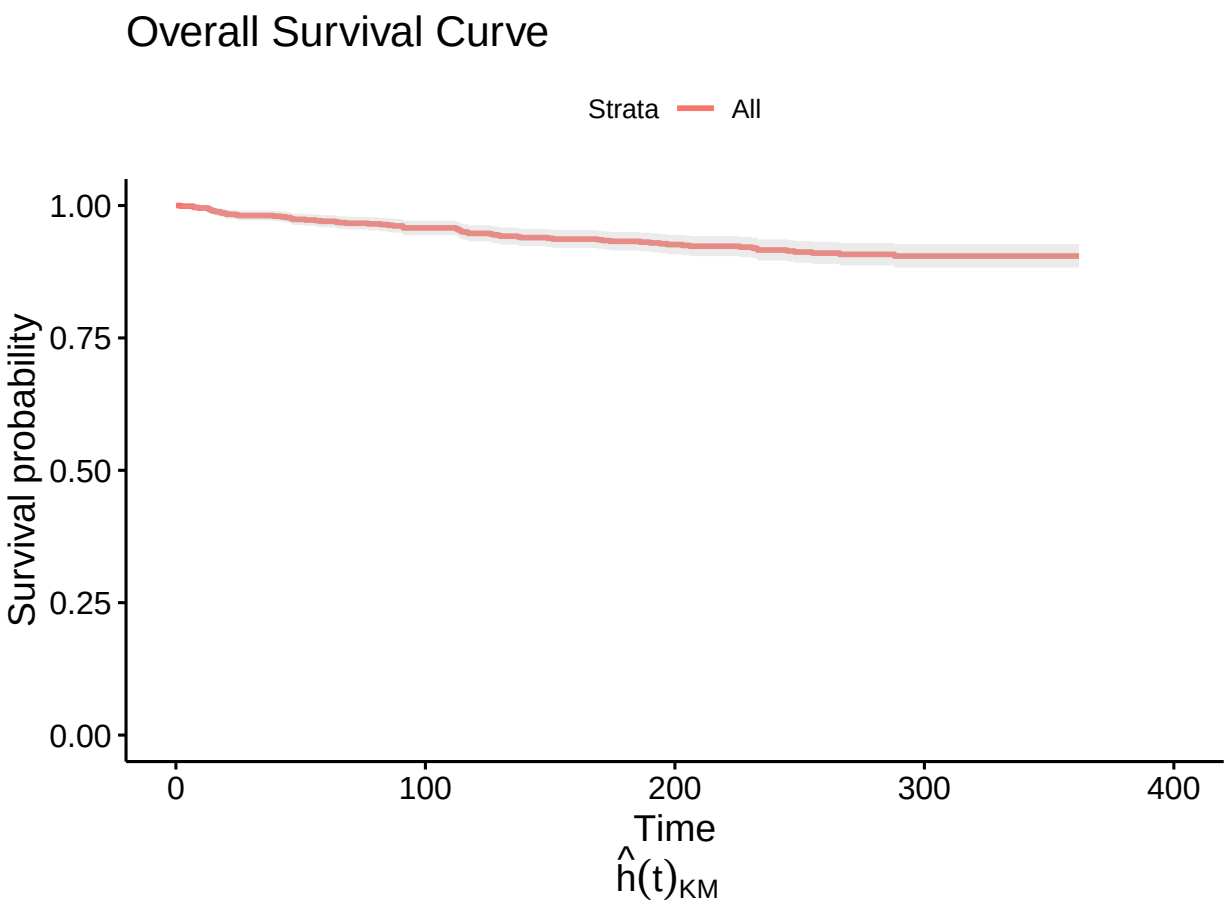
4/12/2019

The Cox Proportional Hazards model describes the relationship between survival of an individual based on one or more explanatory variables (covariates). Thus, it can help estimate the effectiveness of treatment on survival, and can provide an estimate of the hazard function (the liklihood of the event occuring at any given point in time) based on the covariates. I will select explanatory variables from the AIDS study to build a Cox model to predict survival information related to the significant explanatory variables, as well as to construct an estimates of the hazard function to predict the liklihood of an individual experiencing AIDS diagnosis or death. The Cox model relies on the assumption that there are proportional hazards (the ratio of hazards for any two individuals is constant over time). This is because the Cox model is built on the following: $h_i(t) = h_0(t)e^{\beta x_i}$. Then, the hazard ratio for individuals i and j is $\frac{h_0(t)e^{\beta x_i}}{h_0(t)e^{\beta x_j}} = e^{\beta(x_i - x_j)}$, so there are proportional hazards for any two individuals, independent of time. Additionally, e^{β_k} can be interpreted as the hazard ratio associated with a one unit increase in covariate k. Because the Cox model is build upon a proportional hazards assumption, it is important to investigate and test whether there are proportional hazards in a proposed model.

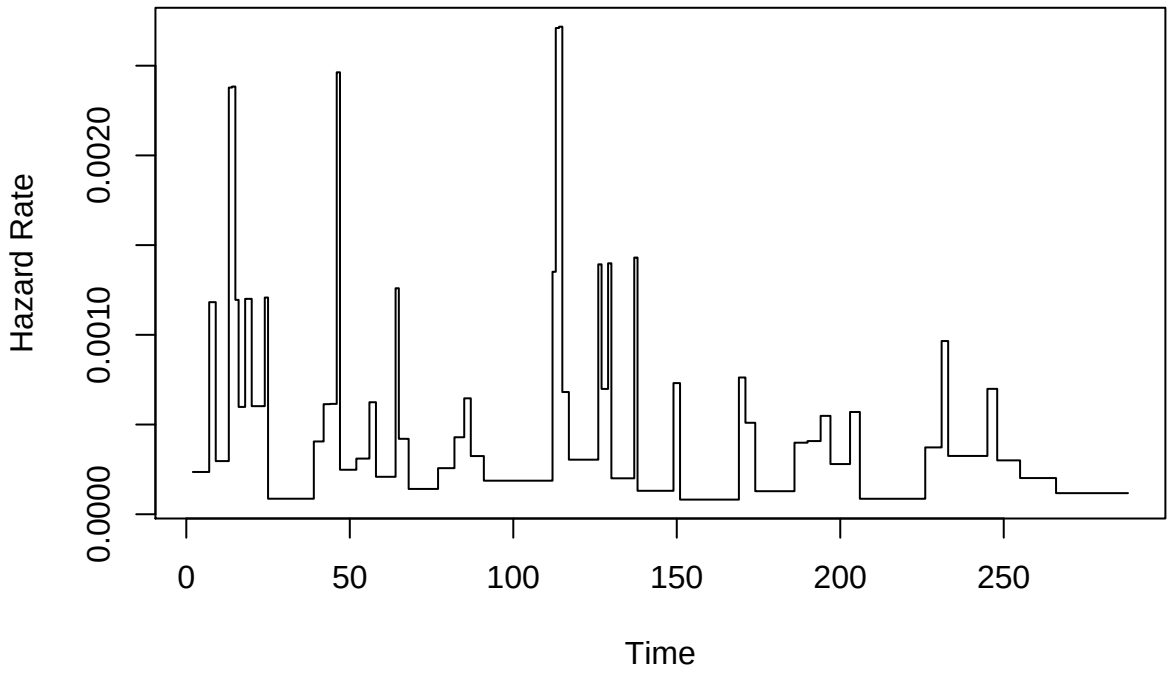
Under the proportional hazards assumption, the β coefficients are determined with maximum liklihood estimation. $L(\beta) = \prod_{i=1}^{n} (\frac{e^{\beta x_i}}{\sum_{k:t_k > t_i} e^{\beta x_k}})^{\delta_i}$ where δ_i is an indicator for events. $b = \hat{\beta}$ is determined by taking the log-liklihood and setting partial derivatives with respect to β equal to 0. When proportional hazards are violated, the hazard ratio is dependent on time. Thus, $h_i(t) = h_0(t)e^{\beta_1 + \beta_2 x_i(t)}$. We want to test whether $\beta_2 = 0$, which means that there are proportional hazards. This is done by the r function cox.zph.

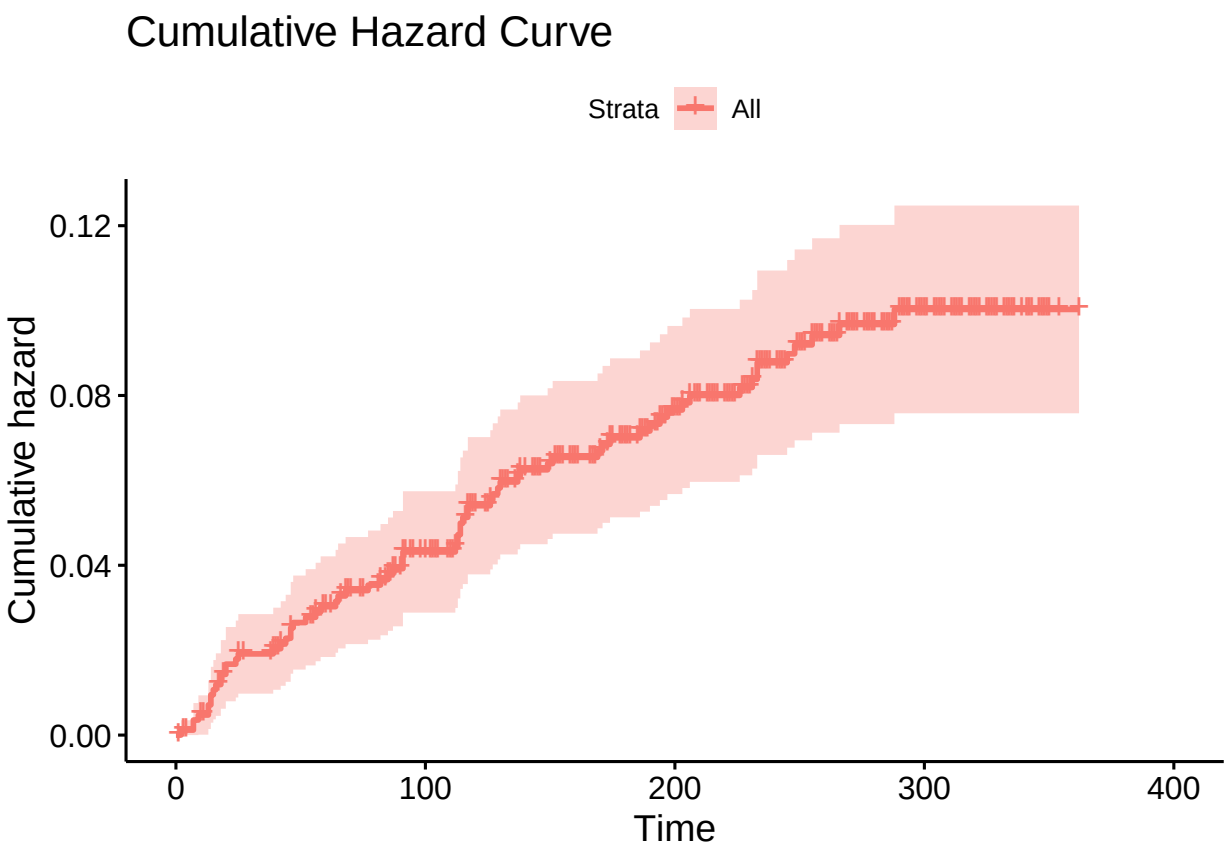
Additional work on what the cox.zph function does to test the hypothesis that $\beta_2 = 0$ will be explored by looking at the r documentation.

The following plots display the survival function, hazard function (to be estimated by the cox model), and cumulative hazard function for the AIDS study.


Overall Survival Curve
Strata
All
1.00
0.75
0.50
0.25
0.00
Survival probability
0
100
200
300
400
Time

$\hat{h}(t)_{KM}$

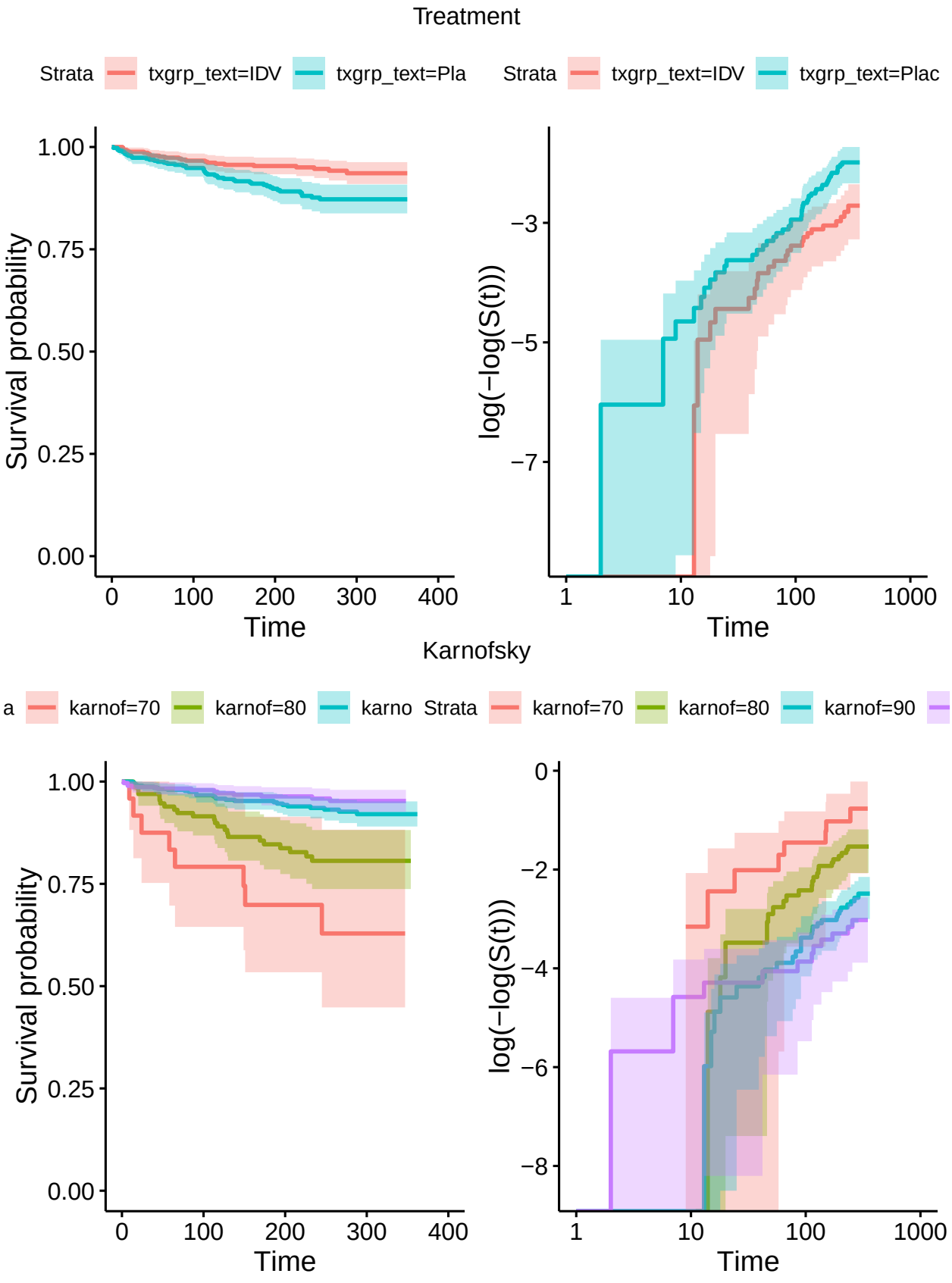

Hazard Rate
0.0000
0.0010
0.0020
0
50
100
150
200
250
Time

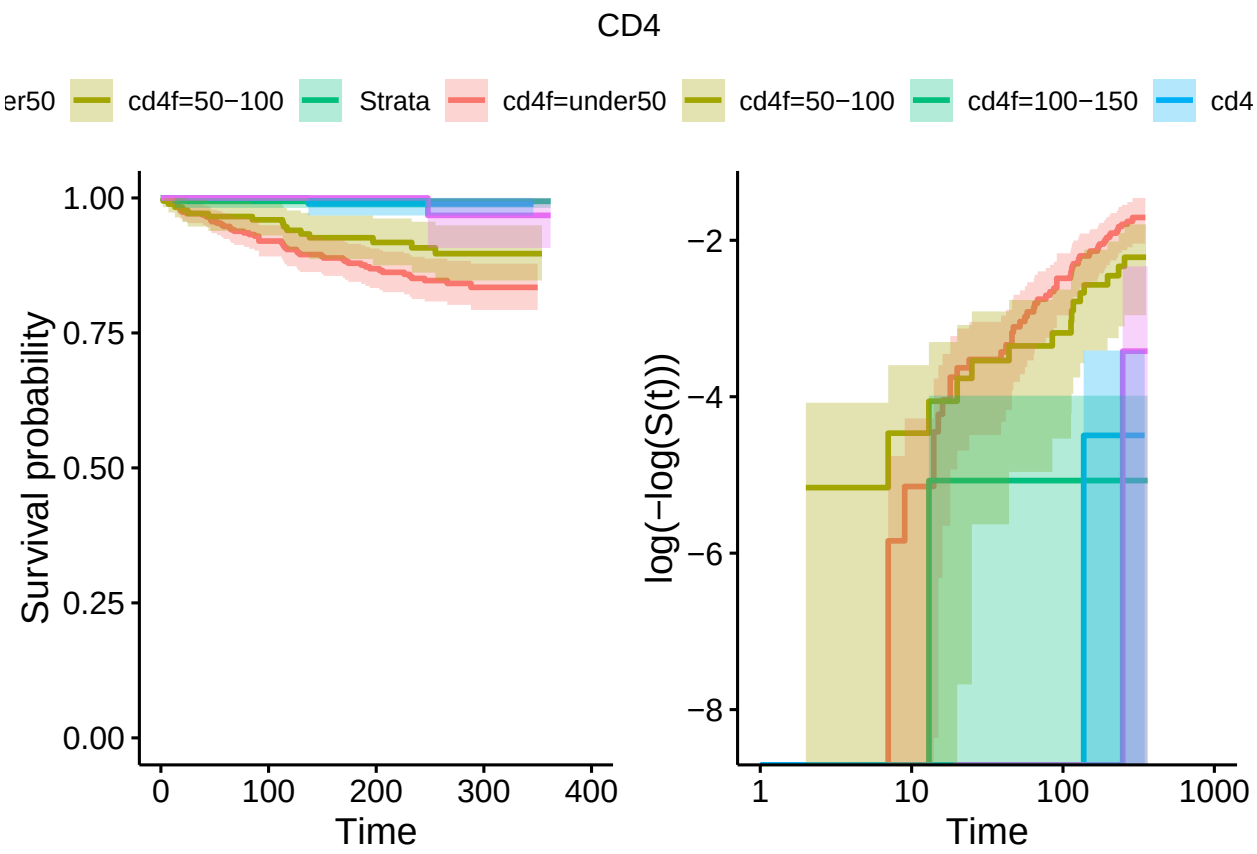


The table below displays the significance of each covariate individually when incorporated into a model.

```
##              beta HR (95% CI for HR) wald.test p.value
## txgrp      -0.76   0.47 (0.28-0.77)       8.9  0.0028
## sex          0.2     1.2 (0.65-2.3)      0.39    0.53
## raceth    0.0036       1 (0.77-1.3)         0    0.98
## ivdrug     -0.13    0.88 (0.62-1.2)      0.52    0.47
## hemophil    0.27     1.3 (0.41-4.2)      0.21    0.65
## karnof    -0.081    0.92 (0.9-0.95)        34 6.9e-09
## cd4       -0.017   0.98 (0.98-0.99)        32 1.9e-08
## priorzdv -0.0032         1 (0.99-1)      0.48    0.49
## age         0.017         1 (0.99-1)       1.8    0.18
```

The individual covariates txgrp, karnof, and cd4 have significant correlation coefficients. The order of covariates in order of most to least significant is karnof, cd4, txgrp, age, ivdrug, prior zdv, sex, hemophil and raceth. Now, to explore the proportional hazards assumption, I will plot the complimentary log log curves for the most significant variables. Under perfect proportional hazards, at any point in time, the difference between any two curves is constant. When proportional hazards is violated, the curves will exhibit a significant cross.


Treatment
Strata txgrp_text=IDV txgrp_text=Pla
Strata txgrp_text=IDV txgrp_text=Plac
Survival probability
1.00
0.75
0.50
0.25
0.00
0 100 200 300 400
Time
log(−log(S(t)))
−3
−5
−7
1 10 100 1000
Time
Karnofsky
a karnof=70 karnof=80 karno
Strata karnof=70 karnof=80 karnof=90
Survival probability
1.00
0.75
0.50
0.25
0.00
0 100 200 300 400
Time
log(−log(S(t)))
0
−2
−4
−6
−8
1 10 100 1000
Time



When treatment is the variable, the complimentary log log curves clearly do not cross, indication proportional hazards. When karnofsky score is the variable, scores of 90 and 100 overlap a bit, but are essentially the same curve (and the confidence intervals are very large and overlapping), so proportional hazards holds. When CD4 is the variable, the curves for the higher categories have some overlap, but once again have extremely large and overlapping confidence intervals, so we can assume proportional hazards. A multivariate model will be fit with the three significant covariates in the order of most to least significant, as all appear to have proportional hazards.

I will build a cox ph model using onlu txgrp as a covariate

```
## # A tibble: 1 x 7
##   term  estimate std.error statistic p.value conf.low conf.high
##   <chr>    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 txgrp   -0.762     0.255     -2.98 0.00284    -1.26    -0.262
```

```
## [1] 0.002073565
```

The estimate for the treatment coefficient is -0.76225, which means that there is a reduction in hazard as treatment group changes from placebo to IDV. The LRT for adding in txgrp as a variable results in a p-value of 0.002, so it is included in the model.

I will build a cox ph model using txgrp + karnof as the covariates.

```
## # A tibble: 2 x 7
##   term   estimate std.error statistic       p.value conf.low conf.high
##   <chr>     <dbl>     <dbl>     <dbl>         <dbl>    <dbl>     <dbl>
## 1 txgrp   -0.797     0.255      -3.12 0.00181         -1.30    -0.296
## 2 karnof  -0.0805    0.0137     -5.89 0.00000000396   -0.107   -0.0537
```

```
## [1] 1.320752e-08
```

The estimate for the karnofsky coefficient is -0.080462, so hazard decreases as karnosky score increases. The LRT for adding in karnof as a variable results in a p-value of 1.320752e-08, so it is included in the model.

I will build a cox ph model using txgrp + karnof + cd4 as the covariates.

```
## # A tibble: 3 x 7
##   term   estimate std.error statistic    p.value conf.low conf.high
##   <chr>     <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
## 1 txgrp   -0.681    0.256       -2.66 0.00787     -1.18    -0.179
## 2 karnof  -0.0574   0.0138      -4.16 0.0000318   -0.0845  -0.0304
## 3 cd4     -0.0146   0.00307     -4.76 0.00000197  -0.0206  -0.00860
```

```
## [1] 7.403133e-09
```

The estimate for the cd4 coefficient is -0.014622, so hazard decreases as cd4 increases. The LRT for adding in cd4 as a variable results in a p-value of 7.403133e-09, so it is included in the model.

I will build a cox ph model using txgrp + karnof + cd4 + age as the covariates.

```
## # A tibble: 4 x 7
##   term   estimate std.error statistic    p.value conf.low conf.high
##   <chr>     <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
## 1 txgrp   -0.697    0.256       -2.72 0.00654    -1.20     -0.195
## 2 karnof  -0.0559   0.0139      -4.02 0.0000575  -0.0831   -0.0287
## 3 cd4     -0.0151   0.00313     -4.84 0.00000132 -0.0213   -0.00900
## 4 age      0.0213   0.0138       1.54 0.123      -0.00579   0.0483
```

```
## [1] 0.1295928
```

The confidence interval for the coefficient for age includes 0, so it is unclear what effect age has on hazard.The LRT for adding in age as a variable results in a p-value of .13, so it is not included in the model. The covariates txgrp, karnof and cd4 should be included in the model built using forward selection. Next, I will create a model where cd4 is split into factors to see if it shoud be a categorical or continuous variable.

```
## # A tibble: 6 x 7
##   term        estimate std.error statistic    p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
## 1 txgrp        -0.694     0.256      -2.72 0.00662     -1.20     -0.193
## 2 karnof       -0.0625    0.0138     -4.52 0.00000608  -0.0896   -0.0354
## 3 cd4f50-100   -0.310     0.298      -1.04 0.299       -0.895     0.275
## 4 cd4f100-150  -2.94      1.01       -2.90 0.00368     -4.92     -0.956
## 5 cd4f150-200  -2.39      1.01       -2.36 0.0181      -4.38     -0.409
## 6 cd4fover200  -1.90      1.01       -1.87 0.0609      -3.89      0.0866
```

```
## [1] 2.559148e-06
```

The p-value of 2.559148e-06 shows that the cd4 categories should be included in the model.The log(HR) will be linear in cd4 if it should be linear. Thus, $\frac{e^{b_2}}{e^{b_1}} = \frac{e^{b_3}}{e^{b_2}} = \frac{e^{b_4}}{e^{b_3}}$, so $b_2 - b_1 = b_3 - b_2 = b_4 - b_3$.

```
## [1] 0.494331
```

```
## [1] 0.5440384
```

```
## [1] -2.628251
```

Because there are constant 50 cd4 gaps between the groups. The relationship holds for the upper 3 cd4 factors, but the value from moving from the lowest to the next lowest cd4 group is -2.628251 as compared to around 0.5, so cd4 should be kept as factors. Next, I will create models with interaction: first with karnof and cd4f interacting, and then with txgrp and cd4 interacting.

```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 4,5,8,9 ; beta may be infinite.
```

```
## # A tibble: 10 x 7
##    term          estimate std.error statistic   p.value  conf.low conf.high
##    <chr>            <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
##  1 txgrp          -0.661    2.56e-1  -2.58      9.91e-3   -1.16     -0.159
##  2 karnof         -0.0762   1.54e-2  -4.94      7.69e-7   -0.106    -0.0460
##  3 cd4f50-100     -5.59     3.46e+0  -1.62      1.06e-1  -12.4       1.19
##  4 cd4f100-150  -170.       2.86e+4  -0.00596   9.95e-1 -Inf       Inf
##  5 cd4f150-200  -176.       3.99e+4  -0.00442   9.96e-1 -Inf       Inf
##  6 cd4fover200    -0.611    1.37e+1  -0.0448    9.64e-1  -27.4      26.2
##  7 karnof:cd4f~    0.0595   3.82e-2   1.56      1.20e-1   -0.0155    0.134
##  8 karnof:cd4f~    1.69     2.86e+2   0.00592   9.95e-1 -558.      562.
##  9 karnof:cd4f~    1.76     3.99e+2   0.00440   9.96e-1 -Inf       Inf
## 10 karnof:cd4f~   -0.0133   1.51e-1  -0.0880    9.30e-1   -0.310     0.284
```

```
## [1] 0.2164783
```

The p-value of 0.2164783 indicates that interaction between karnof and cd4f is not necessary in the model.

```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 3,4,5,8,9,10 ; beta may be infinite.
```

```
## # A tibble: 10 x 7
##    term          estimate  std.error statistic  p.value  conf.low conf.high
##    <chr>            <dbl>      <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
##  1 txgrp          -0.664      0.298   -2.23     2.57e-2   -1.25     -0.0805
##  2 cd4f50-100      0.199      0.894    0.222    8.24e-1   -1.55      1.95
##  3 cd4f100-150   -36.2     8547.      -0.00423  9.97e-1 -Inf       Inf
##  4 cd4f150-200    14.8     4988.       0.00296  9.98e-1 -Inf       Inf
##  5 cd4fover200   -37.2    14236.      -0.00262  9.98e-1 -Inf       Inf
##  6 karnof         -0.0622     0.0138  -4.49     7.08e-6   -0.0893   -0.0351
##  7 txgrp:cd4f50~  -0.391      0.656   -0.595    5.52e-1   -1.68      0.895
##  8 txgrp:cd4f10~  17.0     4273.       0.00399  9.97e-1 -Inf       Inf
##  9 txgrp:cd4f15~ -16.6     4988.      -0.00334  9.97e-1 -Inf       Inf
## 10 txgrp:cd4fov~  18.1     7118.       0.00254  9.98e-1 -Inf       Inf
```

```
## [1] 0.5398977
```

The p-value of 0.5398977 indicates that interaction between txgrp and cd4 is not needed in the model.

The best candidate models are those with txgrp, karnof, and cd4 as either a categorical or continuous variable.

Now, I will test whether the proportional hazards assumption holds with the two best models.

```
##            rho chisq     p
## txgrp  -0.0893 0.550 0.459
## karnof -0.0594 0.237 0.626
## cd4     0.1585 1.555 0.212
## GLOBAL      NA 2.060 0.560
```

```
##                 rho  chisq      p
## txgrp       -0.0848 0.4947 0.4818
## karnof      -0.0326 0.0695 0.7920
## cd4f50-100  -0.0293 0.0558 0.8133
## cd4f100-150 -0.1758 2.0837 0.1489
## cd4f150-200  0.0766 0.3960 0.5292
## cd4fover200  0.2045 2.8965 0.0888
## GLOBAL           NA 6.1205 0.4098
```

From the cox.zph test, the p-value for all covariates is above 0.05, so the proportional hazards assumption is met and we cannot reject the null hypothesis that the hazards ratio is dependent on time.