Math 150 - Methods in Biostatistics - Project Final Submission

Victor Machado

April 19, 2019

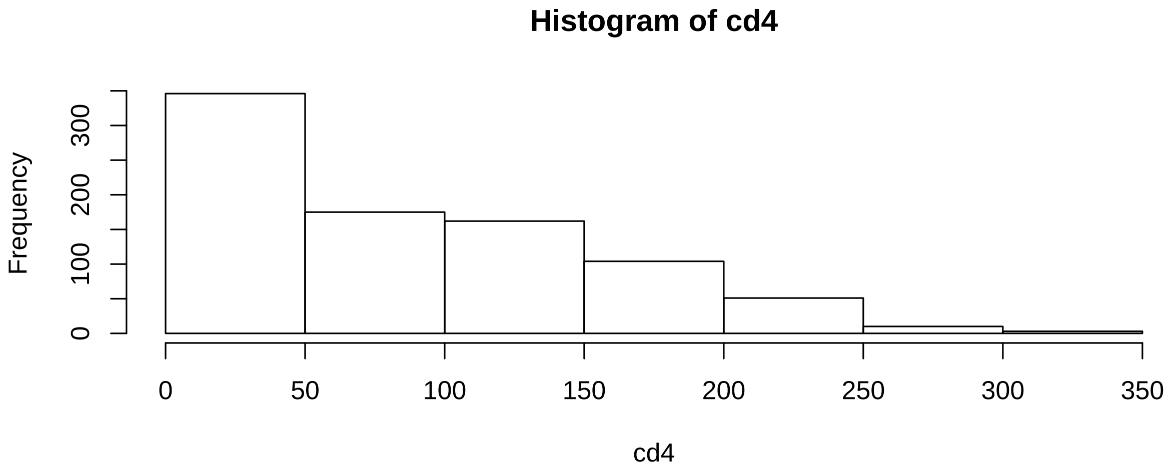
```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=3.5, fig.width=8,
                      fig.align = "center")
library(tidyverse)
library(broom)
library(tidylog)
library(survival)
library(survminer)
library(readr)
library(boot)
```

Exploratory Data Analysis

The CD4 count is a "snapshot of how well your immune system is functioning". CD4 cells are white blood cells that fight infection. The more a person has the btter, these are the cells that the HIV virus kills. When CD4 count drops below 200, a person is diagnosed with AIDS. A normal range for CD4 cells is about 500-1,500 cells/mm3. The following histogram shows that more than half of the patients in the study had this count at lower than 150 cells/mm3.

(https://www.hiv.va.gov/patient/diagnosis/labs-CD4-count.asp)

```
AIDSdata <- read_csv("AIDSdata.csv")
cd4 <- AIDSdata$cd4
hist(cd4)
```



The following barplot shows that most of the patients in the study were males 84%. According to the CDC, from 1996-2000, out of all the people with AIDS 77.4% were male.

(https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5021a2.htm#tab1)

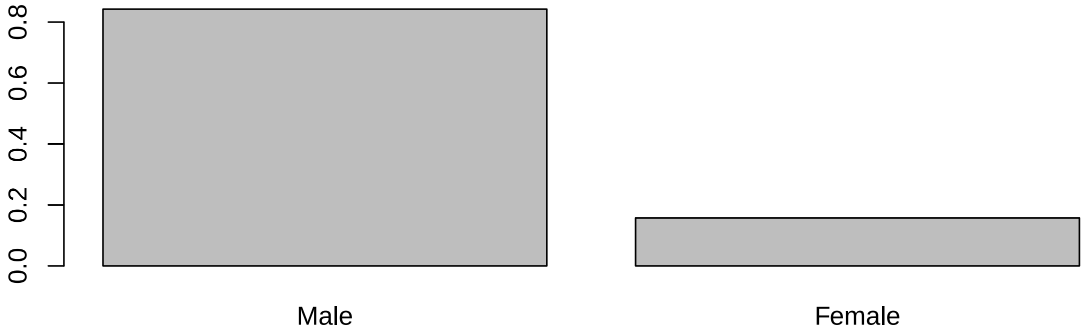
```r
sex <- factor(AIDSdata$sex, labels=c("Male", "Female"))
summary(sex)
```

```
##   Male Female
##    717    134
```

```r
prop.table(table(sex))
```

```
## sex
##      Male    Female
## 0.8425382 0.1574618
```

```r
barplot(prop.table(table(sex)))
```



In the study, 53% of the patients who participated were White, 28% of the patients were Black, 18% of the patients were Hispanic, 1% of the patients were Asian or Pacific Islander and 1% of the patients were American Indian or Alaskan Native.

However, according to the CDC, from 1996-2000, the highest rate of people with AIDS by race was Black with 44.9%, then White with 34%, Hispanic with 19.7%, Asian/Pacific Islander with 0.8% and American Indian/Alaska Native with 0.4%.

(https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5021a2.htm#tab1)

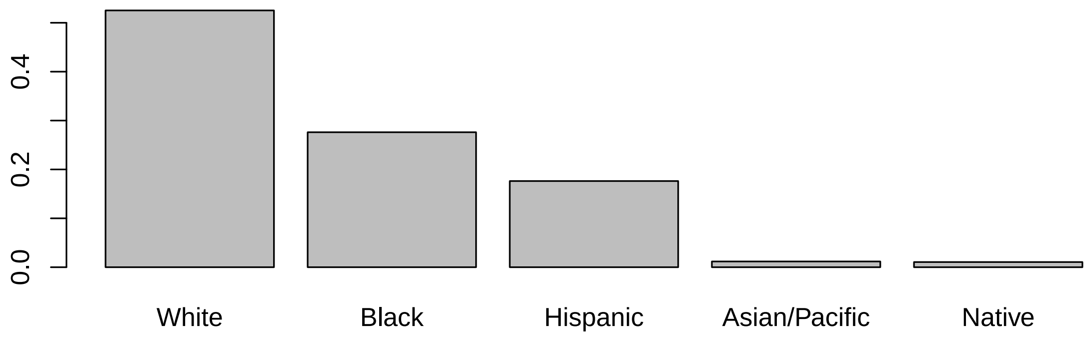
```r
ethnicity <- factor(AIDSdata$raceth, labels=c("White", "Black", "Hispanic", "Asian/Pacific", "Native"))
summary(ethnicity)
```

```
##         White         Black      Hispanic Asian/Pacific        Native
##           447           235           150            10             9
```

```r
prop.table(table(ethnicity))
```

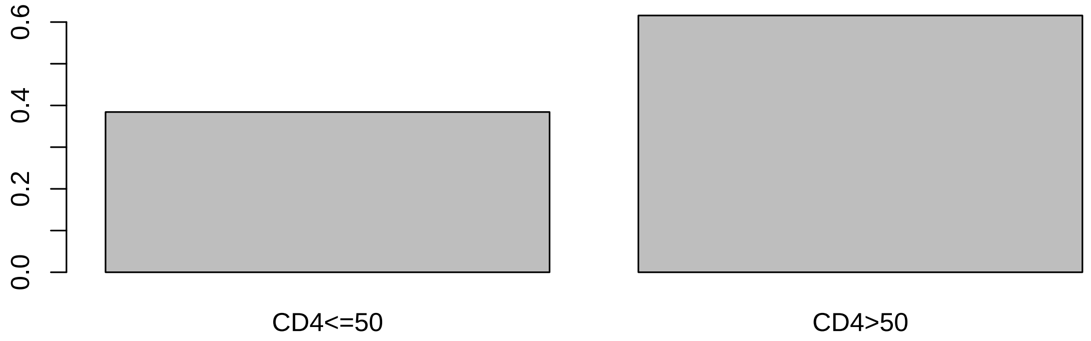
```
## ethnicity
##         White         Black      Hispanic Asian/Pacific        Native
##    0.52526439    0.27614571    0.17626322    0.01175088    0.01057579
```

```r
barplot(prop.table(table(ethnicity)))
```



A total of 1156 patients not previously treated with lamivudine (3TC) or protease inhibitors were stratified according to CD4 cell count (50 or fewer vs 51 to 200 cells per cubic millimeter) and then were randomly assigned to one of two daily regiments:
600mg of zidovudine (ZDV) and 300 mg of lamivudine (3TC) or
The same regiment along with 2400 mg of indivanir.
Stavudine could be substituted for zidovudine.
The primary end point was the time to the development of acquired immunodeficiency syndrome (AIDS) or death.

```
stratified_data <- factor(AIDSdata$strat2, label=c("CD4<=50","CD4>50"))
barplot(prop.table(table(stratified_data)))
```

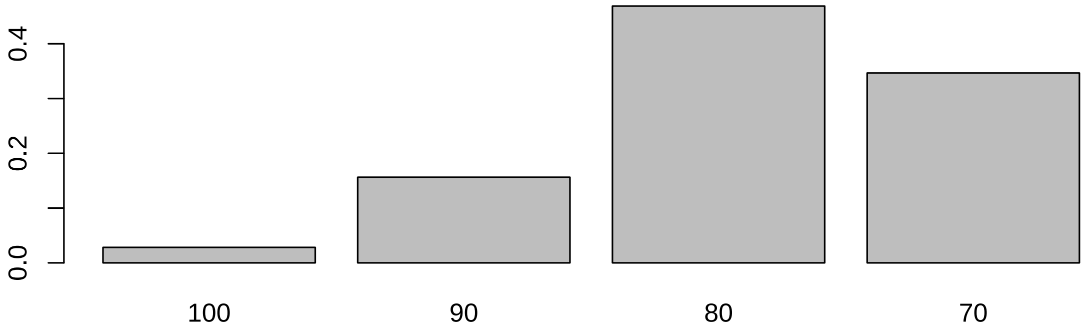


The patients, had to be more than 16 years old and had to have laboratory documentation of HIV-1 infection, a CD4 cell count of 200 per cubic millimeter or less within the 60 days before entry into the study, at least 3 months of zidovudine treatment, no more than 1 week of prior lamivudine treatment, no prior treatment with protease inhibitors and a Karnofsky score of at least 70.

Below we can see the proportion of patients according to their karnofsky score, where
100 means normal no complaint/no evidence of the disease 90 means normal activity possible; minor signs/symptoms of the disease 80 means normal activity with effort;some signs/symptoms of the disease 70

means care for self; normal activity/active work not possible

```
karnof <- factor(AIDSdata$karnof, label=c("100", "90", "80", "70"))
barplot(prop.table(table(karnof)))
```



IV drug use history was also kept track of, 84% of the patients in the study had never used IV drugs and about 16% had previously used IV drugs or never used it. Meaning many of these patients were using IV drugs for the first time.

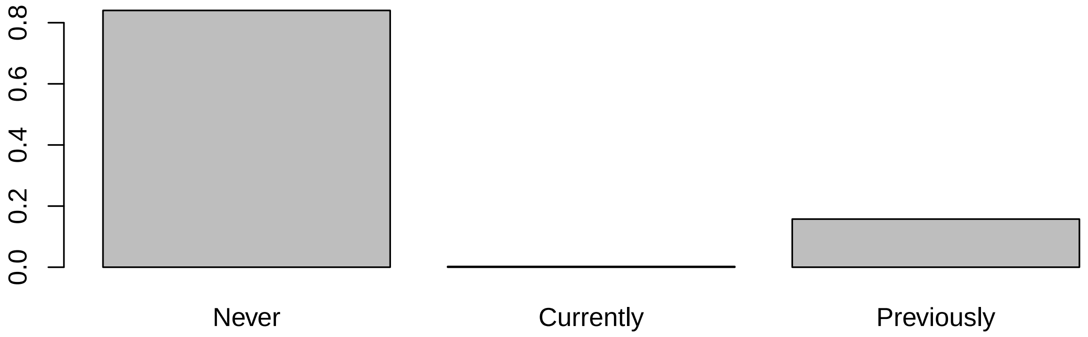
```
ivhistory <- factor(AIDSdata$ivdrug, label=c("Never", "Currently", "Previously"))
summary(ivhistory)
```

```
##      Never  Currently Previously
##        715          2        134
```

```
prop.table(table(ivhistory))
```

```
## ivhistory
##       Never   Currently  Previously
## 0.840188014 0.002350176 0.157461810
```

```
barplot(prop.table(table(ivhistory)))
```



Cox PH Model

First, we must check the proportional hazards assumption for treatment KM curves.

```
AIDSdata <- read_csv("AIDSdata.csv")
survdiff(Surv(time, censor)~tx, data=AIDSdata)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ tx, data = AIDSdata)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## tx=0 422       46     33.3      4.83      9.35
## tx=1 429       23     35.7      4.51      9.35
##
##  Chisq= 9.3  on 1 degrees of freedom, p= 0.002
```

This log-rank test, indicates that the survival times are different in both groups.

Next we can look at the Schoenfeld residuals to test the proportional hazards assumtpion for each covariate included in a Cox regression model fit. The funtion cox.zph correlates the corresonding set of scaled Schoenfeld residuals with test, to test for independence between residuals and time. ()

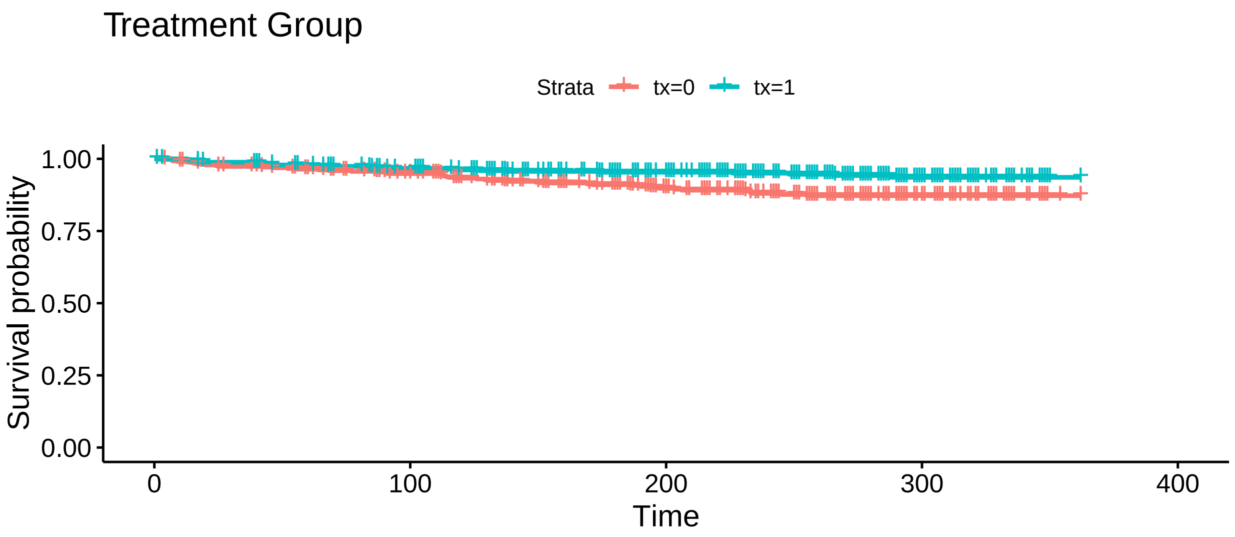
```
aids.cox1 <- coxph(Surv(time, censor)~tx + karnof + cd4, data=AIDSdata)
test.ph <- cox.zph(aids.cox1)
test.ph
```

```
##            rho chisq     p
## tx     -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
```

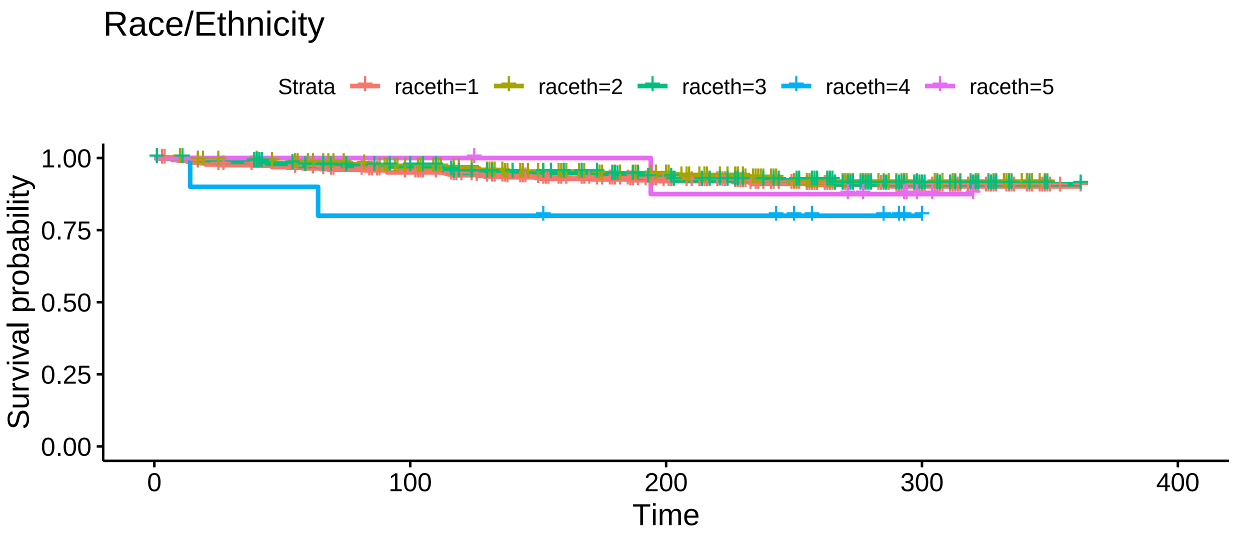
From this output, we can conclude that the test is not statistically significant. Therefore, we assume proportional hazards.

Before choosing a model it is interesting to look at the Kaplan-Meier curves based on treatment groups, race and ethnicity and sex just to see if there are any huge differences between these KM curves.

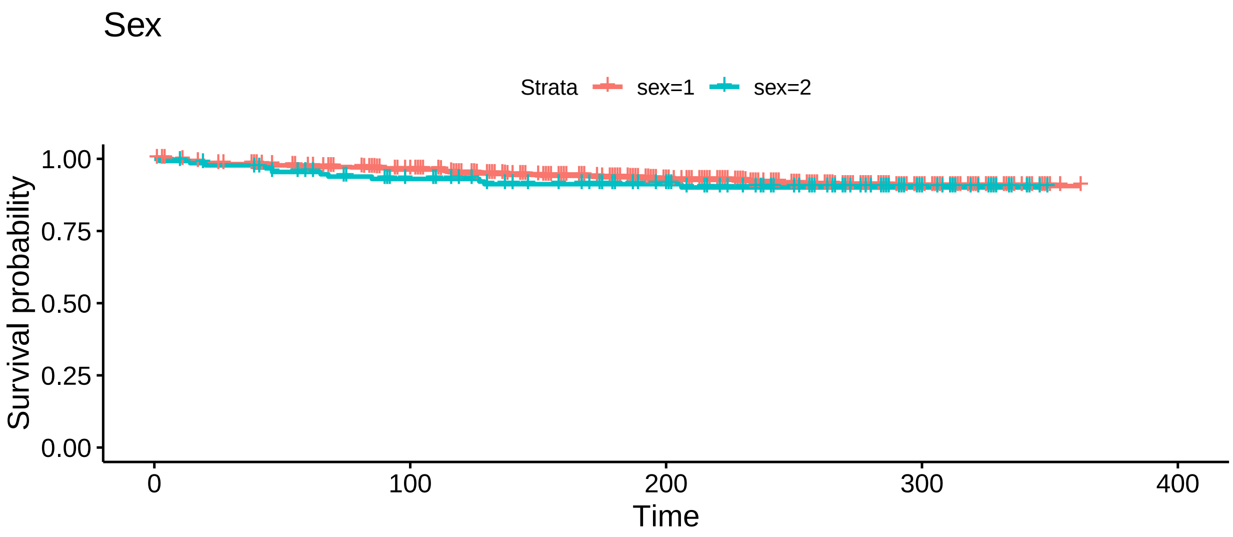
```
ggsurvplot(survfit(Surv(time, censor)~tx, data=AIDSdata)) + ggtitle("Treatment Group")
```



```r
ggsurvplot(survfit(Surv(time, censor)~raceth, data=AIDSdata)) + ggtitle("Race/Ethnicity")
```



```r
ggsurvplot(survfit(Surv(time, censor)~sex, data=AIDSdata)) + ggtitle("Sex")
```



We can see that for sex and race/ethnicity the functions cross which make them violate the proportional hazards assumption so we will not include them in the model. Whereas, for the treatment group option we do not see them crossing which validates our earlier test to assume proportional hazards for this variable. We also see that patients who had treatments that include IDV have a higher survival probability than patients who did not have treatments that include IDV.

After much testing and model selecting strategies (not including the new information). I reached the following model:

```
aids.cox = coxph(Surv(time, censor)~tx + karnof + cd4, data=AIDSdata)
summary(aids.cox)
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ tx + karnof + cd4, data = AIDSdata)
##
##   n= 851, number of events= 69
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## tx     -0.680710  0.506258  0.256155 -2.657  0.00787 **
## karnof -0.057422  0.944195  0.013804 -4.160 3.18e-05 ***
## cd4    -0.014622  0.985485  0.003074 -4.757 1.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## tx        0.5063      1.975    0.3064    0.8364
## karnof    0.9442      1.059    0.9190    0.9701
## cd4       0.9855      1.015    0.9796    0.9914
##
## Concordance= 0.791  (se = 0.026 )
## Likelihood ratio test= 75.21  on 3 df,   p=3e-16
## Wald test            = 61.56  on 3 df,   p=3e-13
## Score (logrank) test = 71.72  on 3 df,   p=2e-15
```

After running our baseline model building algorithm from class, I ended up concluding that the variables strat2, sex, race/ethnicity, ivdrug, hemophil and priorzdv were to be taken out of the model.

I expected race and ethnicity to affect the survival of the participants, however the model selection determined these variables to be insignificant. The results would suggest that race and ethnicity are not related to the risk of death or progression to AIDS.

The Wald statistic value is marked "z" above. It corresponds to the ratio of each regression coefficient to its standard error. It evaluates whether the β coefficient of a variable is statistically different from 0. From the values above, we can conclude that the variables tx, karnof and cd4 have highly significant coefficients.

Next we can interpet the "coef" column, first we must understand that a positive sign means that the hazard is higher, and, thus, for subjects with higher values of that variable there is a higher risk. Since the txgrp variable is -0.6807 we can interpet this as saying that subjects with treatment including IDV have a lower risk of death than subjects with treatment not including IDV, the other two values are close to zero so we will not interpet their regression coefficients as meaning anything significant.

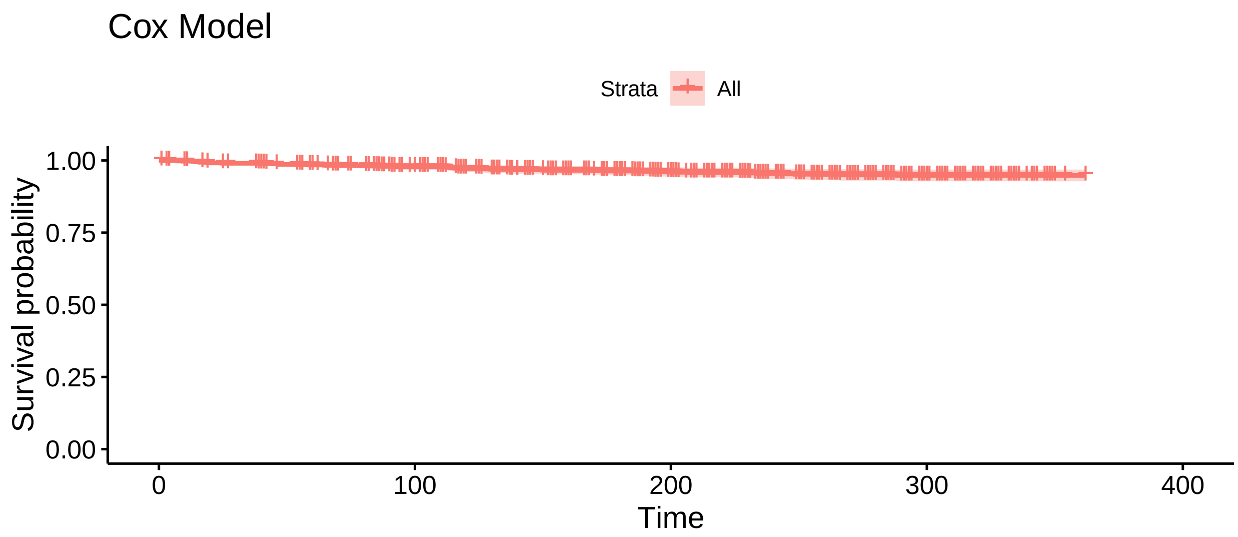
Now we can interpet the hazard ratios (exp(coef)), gives the effect size of the different variables. In this case, being a subject with treatment including IDV reduces hazard by a factor of 0.5063, having a high karnof score reduces the hazard by a factor of 0.9442 and having a high cd4 count reduces the hazard by a factor of 0.9855. Since the last two values are close to 1 it means that karnof score and cd4 count make smaller contribution to the reduction of the hazard than the treatment group.

Finally, the output gives p-values for three different tests for overall significance of the model. The three methods will give similar results for a large enough N. Since all three tests have significant p-values, we can conclude that there is a difference between survival times of the control group and the treatment group. It also indicates that the model is significant, meaning that the null hypothesis that β_i are all equal to 0 is rejected.

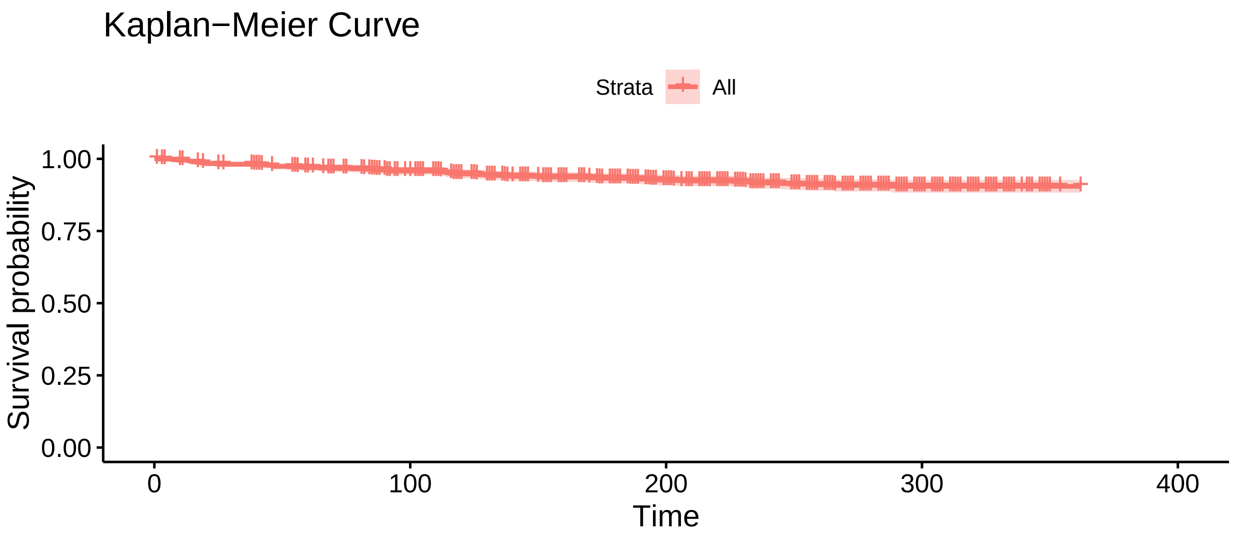
Now that we have fit a Cox model to the data, we can visualize the predicted survival proportion at any given point in time for a particular group.

And we can compare this to the KM survival curve for the entire sample. It seems as if our model is over estimating the survival probability at a given time compared to the KM curve.

```
ggsurvplot(survfit(aids.cox, data = AIDSdata)) + ggtitle("Cox Model")
```



```
ggsurvplot(survfit(Surv(time, censor)~1, data=AIDSdata)) + ggtitle("Kaplan-Meier Curve")
```



AIC & BIC for model selection on Cox PH model

The Akaike information criterion (AIC) is an estimator of the quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. AIC provides a method for model selection different from what we have discussed in class. It can be more formally defined, suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the AIC value is

$$AIC = 2k - 2\ln(\hat{L})$$

Given a set of candidate models for the data, the preferred model is the one with the lowest AIC value.

The Bayesian information criterion (BIC) is a criterion for model selection over a finite set of models; the model with the lowest BIC is the preferred model. This criterion is closely related to the AIC. When fitting models, it is possible to make the model fit better by adding parameter, but doing this may result in overfitting. BIC attempts to resolve this problem by having a penalty term for the number of parameters in the model. The BIC is formally defined as

$$BIC = \ln(n)k - 2\ln(\hat{L})$$

where $\hat{L}$ is once again the maximum value of the likelihood function for the model, k is the number of estimated parameters in the model and n is the amount of data points in the data.

The main difference between BIC and AIC comes with a different penalty for the number of parameters. With AIC the penalty is $2k$ wheras, with BIC the penalty is $\ln(n)k$. Researchers have argued that AIC and BIC are appropriate depending on the task. AIC tries to select the model that most adequately describes an unknown, high dimensional reality. Whereas, BIC tries to find the true model among the set of candidates.

```
model0 <- coxph(Surv(time, censor)~tx + ivdrug +  karnof + cd4 + age, data=AIDSdata)
model1 <- coxph(Surv(time, censor)~tx + karnof + cd4 + priorzdv, data=AIDSdata)
model2 <- coxph(Surv(time, censor)~tx + karnof + cd4, data=AIDSdata)
model3 <- coxph(Surv(time, censor)~tx + karnof, data=AIDSdata)
model4 <- coxph(Surv(time, censor)~tx, data=AIDSdata)
model5 <- coxph(Surv(time, censor)~tx + cd4 * karnof, data=AIDSdata)
extractAIC(model0)
```

```
## [1]   5.0000 835.3762
```

```
extractAIC(model1)
```

```
## [1]   4.0000 838.0202
```

```
extractAIC(model2)
```

```
## [1]   3.0000 836.0552
```

```
extractAIC(model3)
```

```
## [1]   2.0000 867.4811
```

```
extractAIC(model4)
```

```
## [1]   1.0000 897.7817
```

```
extractAIC(model5)
```

```
## [1]   4.0000 837.1484
```

```
BIC(model0)
```

```
## [1] 846.5468
```

```
BIC(model1)
```

```
## [1] 846.9566
```

```
BIC(model2)
```

```
## [1] 842.7576
```

```
BIC(model3)
```

```
## [1] 871.9494
```

```
BIC(model4)
```

```
## [1] 900.0158
```

```
BIC(model5)
```

```
## [1] 846.0848
```

One model I was particularly interested in was whether or not karnof and cd4 had interaction in the sample. However, after looking at BIC and AIC scores for these models, it is clear that the model I reached after doing the model building algorithm we established in class was the best in terms of these two metrics. It is interesting to note that the model labeled "model0" has a better AIC score than every other model I tried, but its BIC score is larger than "model2" my chosen model.

I thought it would also be interesting to look at these scores for different terms interacting.

```
model0 <- coxph(Surv(time, censor)~tx * ivdrug *  karnof * cd4 * age, data=AIDSdata)
model1 <- coxph(Surv(time, censor)~tx * ivdrug *  karnof * cd4, data=AIDSdata)
model2 <- coxph(Surv(time, censor)~tx * karnof * cd4, data=AIDSdata)
model3 <- coxph(Surv(time, censor)~tx * karnof, data=AIDSdata)
model4 <- coxph(Surv(time, censor)~tx, data=AIDSdata)
extractAIC(model0)
```

```
## [1]  31.0000 865.9038
```

```
extractAIC(model1)
```

```
## [1]  15.0000 850.4652
```

```
extractAIC(model2)
```

```
## [1]   7.0000 838.2498
```

```r
extractAIC(model3)
```

```
## [1]   3.0000 869.4803
```

```r
extractAIC(model4)
```

```
## [1]   1.0000 897.7817
```

```r
BIC(model0)
```

```
## [1] 935.1611
```

```r
BIC(model1)
```

```
## [1] 883.9768
```

```r
BIC(model2)
```

```
## [1] 853.8885
```

```r
BIC(model3)
```

```
## [1] 876.1826
```

```r
BIC(model4)
```

```
## [1] 900.0158
```

Something I was thinking about that may be interesting would be to look at the BIC and AIC value for all models with only 1 variable, 2 variables, all the way up to 5 variables. The models with 2 variables being all 5 choose 2 combinations of the explanatory variables tx, ivdrug, karnof, cd4 and age since those had the lowest p-values in the cox model. This proved to be difficult for me to do in R, but it would be interesting to look at information relating to this.