AIDS Data Analysis: Rough Draft

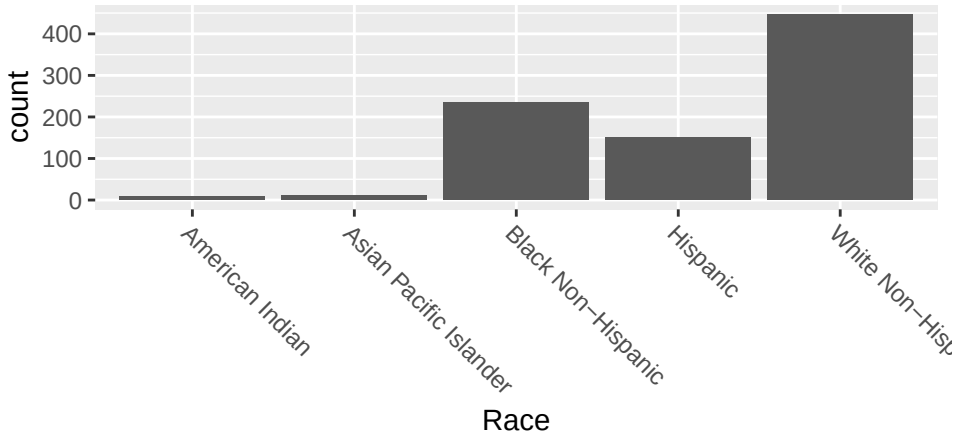
Jack Hanley

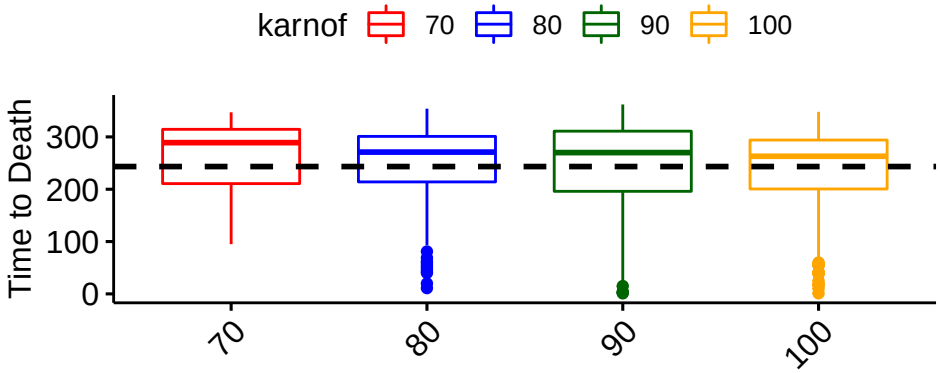
April 26, 2019

Introduction

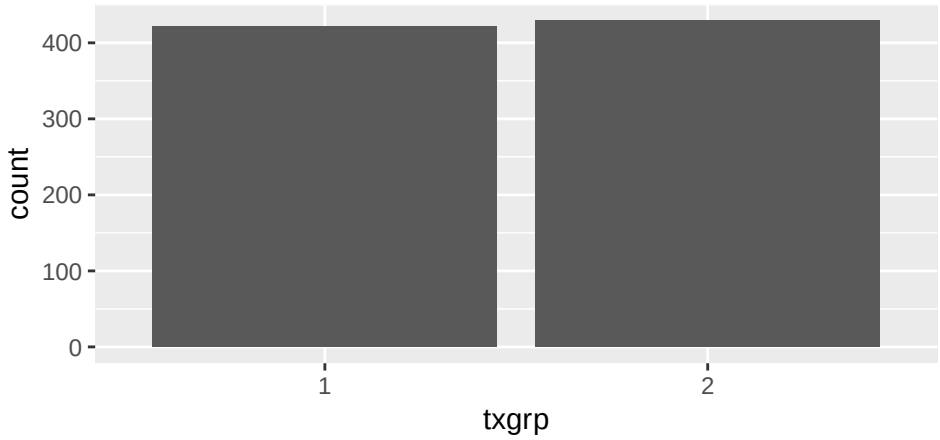
The data for this analysis comes from the study A Controlled Trial of Two Nucleoside Analogues plus Indinavir in Persons with Human Immunodeficiency Virus Infection and CD4 Cell Counts of 200 per Cubic Millimeter or Less by Hammer et al. The principal outcome measure was the time from entering the trial to AIDS defining event (diagnosis) or death. Using this data collected, I will attempt to fit a Cox PH model to help predict the survival of individuals with AIDS given their use antiretroviral drugs such as indinavir (IDV), open label zidovudine (ZDV), stavudine (d4T) and lamivudine (3TC). [1]

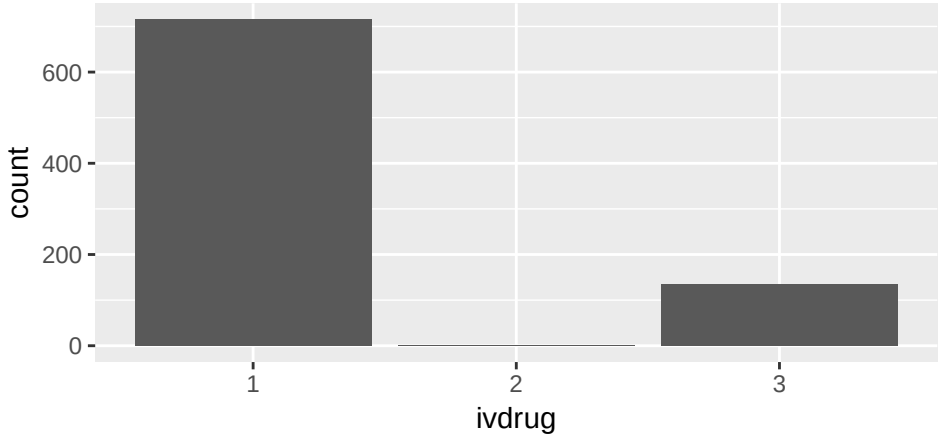
Exploratory Data Analysis

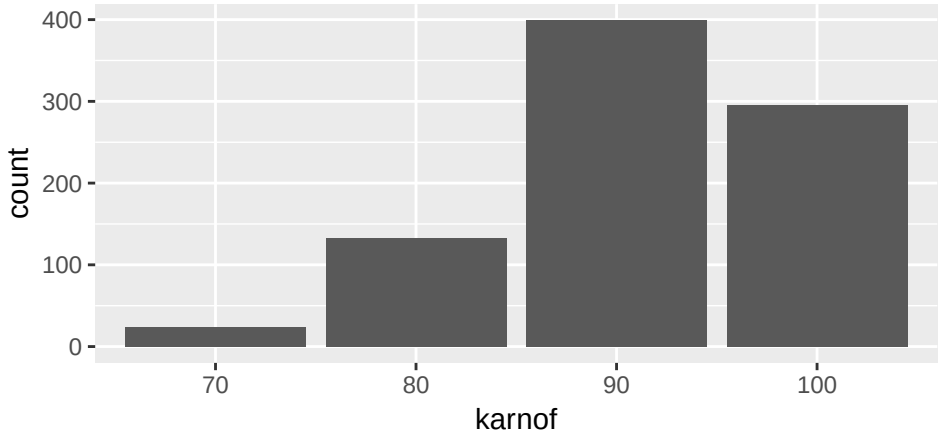


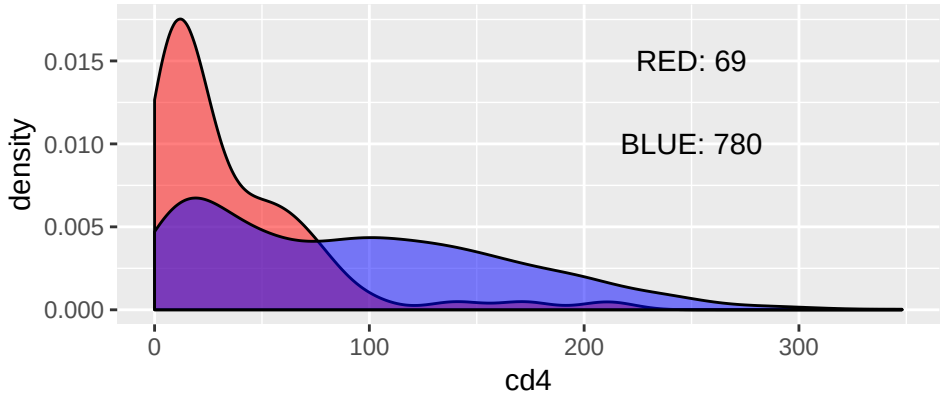


[1] For more information, visit: https://clinicaltrials.gov/ct2/show/NCT00000841 and http://www.nejm.org/doi/full/10.1056/NEJM199709113371101

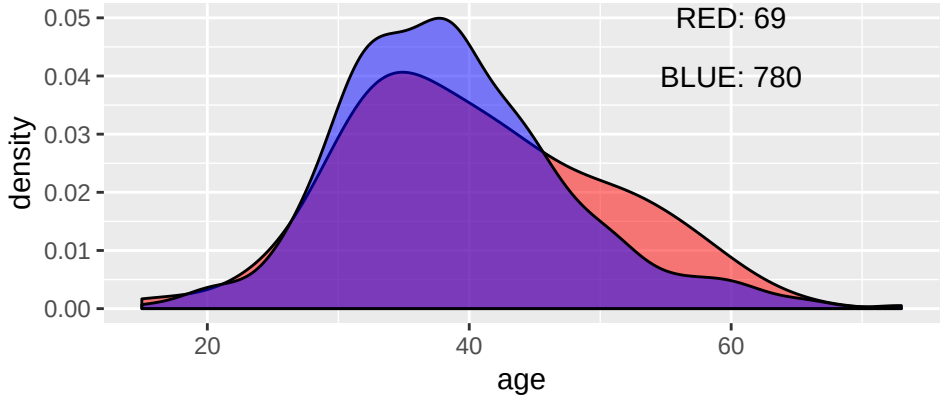

400
300
200
100
0
count
1
2
txgrp


600
400
200
0
count
1
2
3
ivdrug

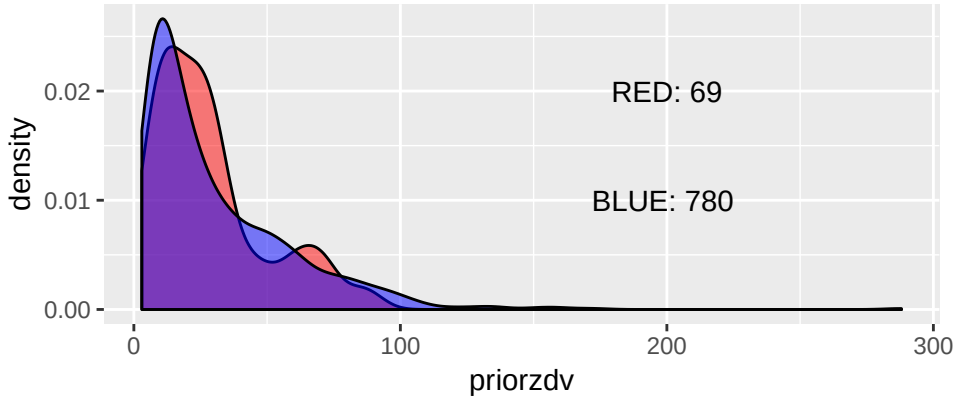

400
300
200
100
0
count
70
80
90
100
karnof



RED: AIDS defining diagnosis/death | BLUE: Otherwise



RED: AIDS defining diagnosis/death | BLUE: Otherwise



RED: AIDS defining diagnosis/death | BLUE: Otherwise

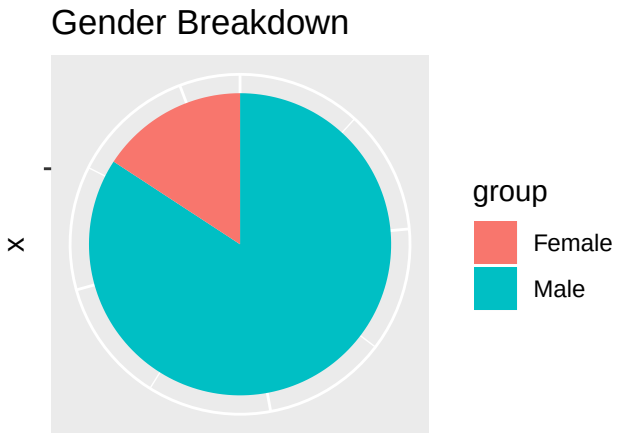
ivdrug	censor	n
1	0	655
1	1	60
2	0	2
3	0	125
3	1	9

Overview of Variables

Before we begin any form of modeling, it is important to first get a better understanding of the data at our disposal through the use of exploratory data analysis. Through exploratory data analysis, we can get a sense of how our variables are distributed and if there are any clear outliers.

Looking at the `raceth` variable, it appears that the data is dominated by white, hispanic, and black individuals. Despite this, each racial group sees at least one death in `censor`. Although we cannot use this study to generalize for American Indian and Asian Pacific-Islander populations, no other aspects of their observations are out of the ordinary and should be left in the dataset.

The same cannot be said of the `ivdrug` variable, which indicates the degree to which the participant used IV drugs (never, in the past, or currently using). Only two individuals out of the entire dataset. The lack of any deaths/diagnoses in this group will screw up any proportional hazards assumptions that would be necessary to construct a Cox Proportional Hazards model. With the implications for a possible COx PH model and the lack of event instances (neither `censor` nor `censor_d`) in mind, I will be removing these observations. Outside of `ivdrug = 2`, all factors have at least one instance of each event (`censor` or `censor_d` equal to 1 and 0).



Another variable of interest is `sex`. Approximately 84% of individuals in the study are male. From the CDC, 76% of all United States adults with HIV were male, so while this number is high, it is not unrealistic [2]. It does, however, limit our ability to make inferences about female HIV individuals. Any results of our models will likely be generalizable to male adult HIV victims, but less so for female adult HIV victims.

Health of the individual seems like it should play a role as well. This is represented by the indivdual's Karnofsky Performance Score in the variable `karnof`. Values range from 70: Cares for self; normal activity/ active work not possible,

to 100: Normal;no complaint no evidence of disease. However, regardless of initial Karnofsky score, the time to death distributions seem fairly similar. Additionally, all karnof-score mean time-to-events are above the total dataset time-to-event.

Other Notes of Interest:

- Even though `txgrp` is listed as having 4 possible indicators in the study documentation, only two indicators, the ones for ZDV + 3TC and ZDV + 3TC + IDV, are used.
- The max of`priorzdv` is 288, with the next closest value at 172. This indicates that the individual was using open label zidovudine (ZDV) for exactly 24 years prior. This could potentially be an outlier,

[2]CDC stat from https://www.cdc.gov/hiv/group/gender/men/index.html

but there is no precedent for a reasonable length of time to be taking ZDV outside of the study. The individual was left in the dataset.

- All continuous variables in the dataset, `age`, `priorzdv`, and `cd4` are skewed left, with both `priorzdv`, and `cd4` heavily skewed. This suggests that they may need to be transformed for modeling purposes.

#Preliminary Modeling

Now that we have given ourselves a sense of what our data looks like and what our results might imply, we can begin some preliminary modeling.

Since there are more events in `censor` than `censor_d`, and therefore likely contain more information, we will be using it along with `time` in our models. We start off with a full model `cox1 <- coxph(Surv(time, censor) ~ tx + txgrp + strat2 + sex + raceth + ivdrug + hemophil + karnof + cd4 + priorzdv + age, data = aids)`, then use `stepAIC` from the MASS package to get a general sense of what kind of variables stay in the model. The output, is `aic.cox`: `coxph(formula = Surv(time, censor) ~ tx + ivdrug + karnof + cd4 + age, data = aids)`. From this point, we will use the likelihood ratio test to see if we can eliminate any other superfluous variables. Comparing `aic.cox` and `cox2: coxph(formula = Surv(time, censor) ~ tx + karnof + cd4, data = aids)` gives us an LRT p-value of 0.09. Although it is close, we cannot reject the null, and are therefore ok with dropping the `sex` and `ivdrug` variables.

```
## Call:
## coxph(formula = Surv(time, censor) ~ tx + karnof + cd4, data = aids)
##
##   n= 849, number of events= 69
##
##               coef exp(coef) se(coef)     z Pr(>|z|)
## tx1       -0.67646   0.50841  0.25761 -2.63  0.00864 **
## karnof80  -0.48840   0.61361  0.41228 -1.18  0.23617
## karnof90  -1.18437   0.30594  0.41262 -2.87  0.00410 **
## karnof100 -1.63580   0.19480  0.46490 -3.52  0.00043 ***
## cd4       -0.01448   0.98563  0.00307 -4.71  2.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## tx1           0.508       1.97    0.3069     0.842
## karnof80      0.614       1.63    0.2735     1.377
## karnof90      0.306       3.27    0.1363     0.687
## karnof100     0.195       5.13    0.0783     0.485
## cd4           0.986       1.01    0.9797     0.992
##
## Concordance= 0.791  (se = 0.026 )
## Rsquare= 0.085   (max possible= 0.656 )
## Likelihood ratio test= 75.4  on 5 df,   p=8e-15
## Wald test            = 62.2  on 5 df,   p=4e-12
## Score (logrank) test = 79.7  on 5 df,   p=1e-15
```

At this point, we none of the LRTs for dropping any of the remaining variables were unsignificant, so we do not feel comfortable dropping them. Our final model is therefore: `cox2 : coxph(formula = Surv(time, censor) ~ tx + karnof + cd4, data = aids)`. It should also be noted that only one of our variables indiating which drug was used by the individuals was included in the final model (`tx`).

```
##               rho chisq     p
```

```
## tx1       -0.1031 0.735 0.391
## karnof80   0.0450 0.142 0.706
## karnof90   0.0541 0.208 0.649
## karnof100 -0.0545 0.197 0.657
## cd4        0.1519 1.452 0.228
## GLOBAL         NA 3.267 0.659
```

Unfortunately, from the output of `cox.zph`, we can see that the Proportional Hazards assumption is not one we can really make, considering the high p-values for each of the variables in our model. While it's certainly possible that the model could perform well with predicting survival, we cannot interpret any of our coefficients in the context of a Porportional Hazards model. However, both the Log-Rank and Wolcoxon tests proved significant, lending to the fact that our model may indeed have some merit.

#Something New: Gradient Boosting via Concordance Index

Background

What is Boosting?

The basic idea of boosting as a regression tool is crowdsourcing. With boosting, you essentially use a bunch of poor learners, such as simple linear regression or a simple regression tree to fit your data. Since these are simple learners, any signficant results you get are almost certainly indicative of signal rather than noise. Therefore, when lots of these simple learners are saying the same thing, we have a pretty good idea that what they're saying is signal.

On a more applied level, we use the findings of each of these 'dumb' learners by incorporating the residuals into the model. This process of first making a model, modeling the residuals, and then creating a new model can be generalized in the following format [3]:

$$F_1(x) = y$$

$$h_1(x) = y - F_1(x)$$

$$F_2(x) = F_1(x) + h_1(x)$$

This process can then be generalized for as many estimations as you want.

What is Gradient Boosting in the context of Survival Analysis?

Although the $h_m(x)$ function in the boosting algorithm above does not specify what type of learner is required (a powerful feature), survival analysis is different. For starters, we do not have residuals in Survival Analysis, and our data can be heavily right-censored. Gradient Boosting of Survival Analysis is therefore not a trivial application. Additionally, our prediction given new data is a hazard function. Rather than residuals though, Gradient Boosting of Survival Analysis accepts theses differences and instead attempts to optimize an approximation of concordance index.

Although the concordance index is a discrete value, it cannot be optimized directly. Instead, a smoothed concordance index is approximated by adopting a logistic sigmoid function. While this makes the function differentiable, the function is neither convex nor concave, and can lead to varying local optima.[4]

[3] Boosting information http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/
[4] Gradient Boosting Survival Analysis: https://www.hindawi.com/journals/cmmm/2013/873595/

Implementation

```
##       Iter       Train Loss   Remaining Time
##          1         369.0005            2.74s
##          2         369.0003            2.29s
##          3         369.0002            2.10s
##          4         369.0001            2.01s
##          5         369.0000            1.80s
##          6         368.9999            1.80s
##          7         368.9998            1.70s
##          8         368.9997            1.60s
##          9         368.9996            1.55s
##         10         368.9995            1.56s
##         20         368.9984            1.45s
##         30         368.9973            1.36s
##         40         368.9962            1.30s
##         50         368.9952            1.20s
##         60         368.9941            1.11s
##         70         368.9931            1.02s
##         80         368.9920            0.93s
##         90         368.9910            0.86s
##        100         368.9900            0.78s
##        200         368.9802            0.00s

##
## Concordance w/ Predictions:

## 0.5005074440647667

##
## FEATURE IMPORTANCE:

## tx=1  ==  0.07228451772488653
## txgrp=2  ==  0.07228480449387613
## strat2=1  ==  0.0
## sex=1  ==  0.05400915712864702
## raceth=Asian Pacific Islander  ==  0.0
## raceth=Black Non-Hispanic  ==  0.0
## raceth=Hispanic  ==  0.0
## raceth=White Non-Hispanic  ==  0.0
## ivdrug=2  ==  0.0
## ivdrug=3  ==  0.0
## hemophil=1  ==  0.012370865413057942
## karnof=80  ==  0.0
## karnof=90  ==  0.0
## karnof=100  ==  0.0
## cd4  ==  0.6135366749620016
## priorzdv  ==  0.17551398027753073
## age  ==  0.0
```

As we can see with the output of the model, the loss at each stage/concordance leave much to be desired. Intererstingly, though, the model is indicating that the most signal is coming from `cd4`, which was the most signficant variable in our Cox PH model. It's likely that this method could create a viable model, but better feature selection/manipulation may be needed.