AIDS Patient Survival Analysis

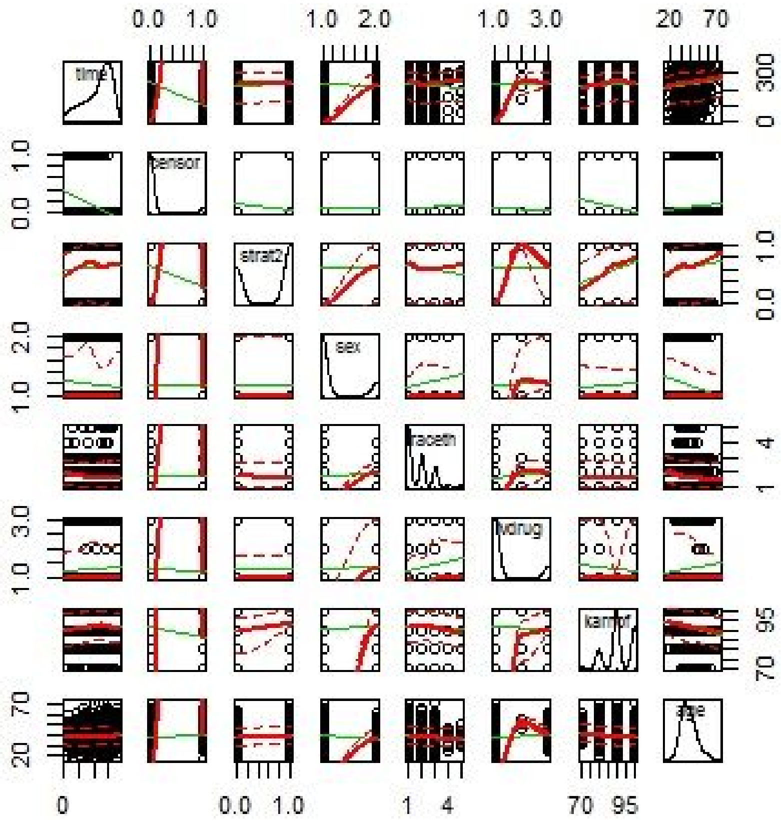
Lam Vu

July 10, 2017

Introduction:

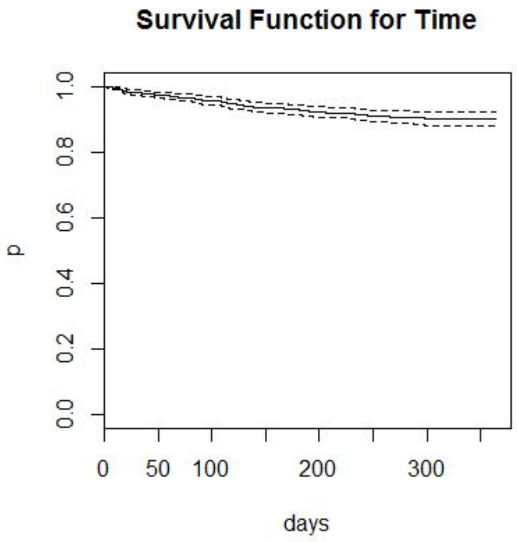
The goal in this analysis of the AIDS Clinical Trials Group Study 320 Data found here: https://www.umass.edu/statdata/statdata/data/actg320.txt, is to estimate the patient's chance of survival as a function of time. The data consists of the 16 variables of which in this study we will only concentrate on id, time(time to AIDS diagnosis or death), censor(1=AIDS,0=otherwise), strat2 (CD4 count: 0 = CD4 <= 50, 1 = CD4 > 50), sex(1=Male,2=Female), raceth(1=White Non-Hispanic, 2=Black Non-Hispanic, 3= Hispanic, 4=Asian,Pacific Islander, 5=American Indian,Alaskan Native, 6=Other/unknown), ivdrug(IV drug use history: 1=Never, 2=Currently, 3=Previously), karnof(Karnofsky Performance Scale: 100 = Normal;no complaint no evidence of disease, 90 = Normal activity possible; minor signs/symptoms of disease, 80 = Normal activity with efforts, some signs/symptoms of disease, 70 = Cares for self; normal activity/active work not possible), and age.

Analysis:



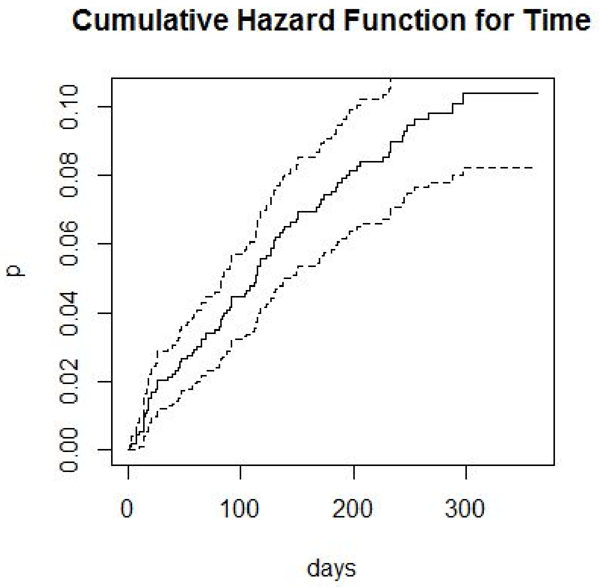
Using a scatterplot matrix to help explore the relationships between variables. We can also see the distribution of each variable. There are more male than female in the data. Between the race in the data the Asian,Pacific Islander, American Indian,Alaskan Native, and Other/unknown seems under represented. Karnof is unevenly distributed with patients scoring 70 and 80 most often. We also notice the Age is approximately normally distributed. In our censored variable 96 patients contracted AIDS and 1055 patients did not. There does not seem to be any variables that have any correlations.

Using survival library function Surv() to create combined object linking censored flag of AIDS to time. I created Kaplan-Meier curve with 95% confidence interval to figure out proportion of individuals who don't have AIDS.

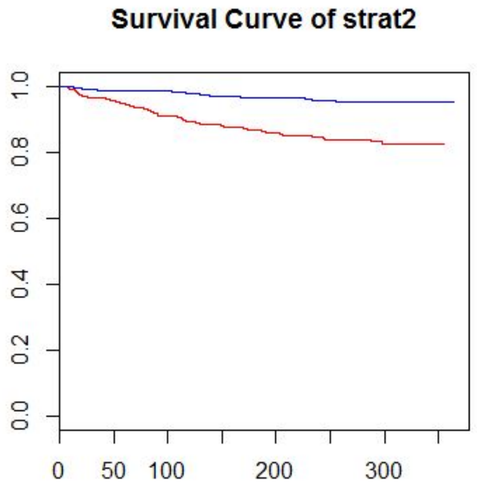


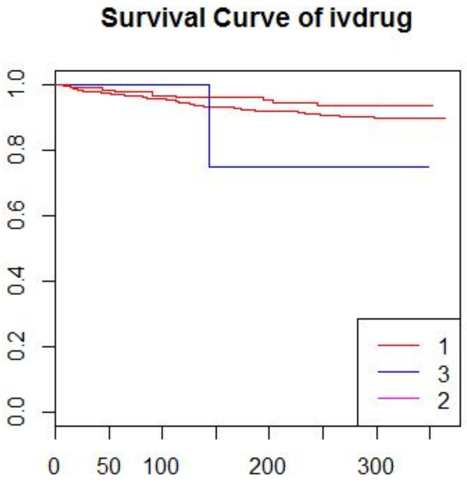
```
Call: survfit(formula = censurv ~ 1, data = data, type = "kaplan-meier")
       n     events     *rmean *se(rmean)
  1151.00      96.00     340.03       2.38
```

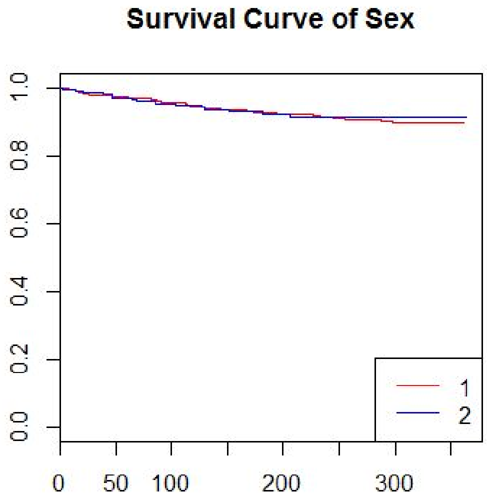
Below is a cumulative hazard plot which illustrates the risk of contracting AIDS with a 95% confidence interval. Cumulative event of having aids is calculated (f(y)=1-y)

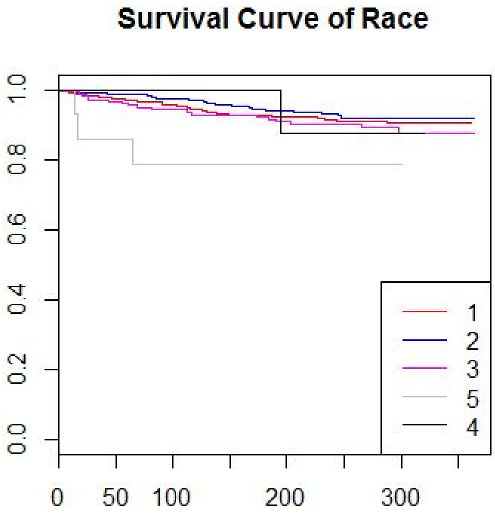


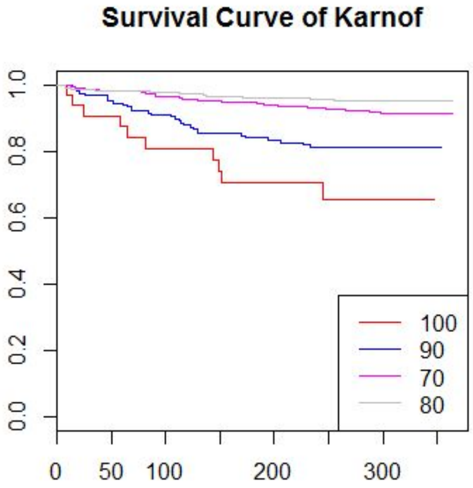
Now I find categorical variables that have impact on the response. Below are survival curves of different categories, in each plot the levels of the variable with the higher lines indicates a greater survival rate.

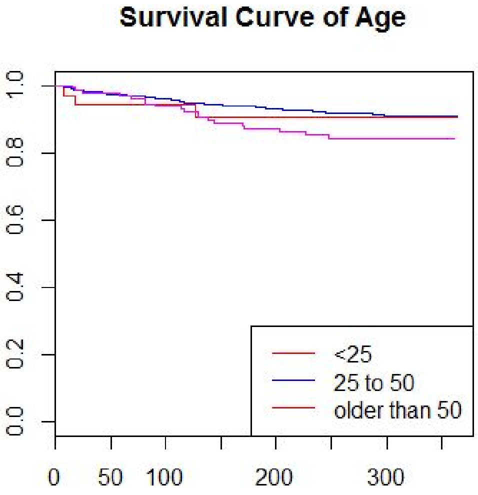












From the plots above we can see that in the survival curve of stat2, the group with CD4<=50 have a higher survival rate. In the survival curve of ivdrug, group 2(currently on IV drug) has a higher rate of survival overall and the patients in group 3(previously on IV drug) has the lowest rate of survival. From the survival curve of sex, there does no seem to be any significant differences between the sex. From the survival curve of race, most race are similar except for group 5 (American Indian/Alaskan Natives) who have a significantly lower rate of survival. From the survival curve of Karnofsky Performance Scale we can see the survival rate of groups with lower Karnof score has a higher survival rate. From the survival curve of age, there does not seem to be a significant difference but older patients seem to have a lower survival rate.

```
Exp        Weibull      Gamma    GenGamma Lognormal Gompertz
-856.5998 -852.9367 -853.0663 -851.2834 -851.4014 -851.2824
```

I then find the distribution of the survival function that bests fit the data. Above are the AIC of the difference distributions, GenGamma and Lognormal produced the lowest AIC. Since the Lognormal distribution is simpler in characteristic and the difference between the AIC is not too large, I conclude the distribution that best fit the data is a lognormal distribution.

I then fit a multiple logistic models using AIC as my criteria I use the different step functions and their results all agree upon a model that includes time,strat2,ivdrug,karnof, and age: censor ~ time + strat2 + ivdrug + karnof + age

---

Interpretation:

From the Kaplan-Meier curve with the 95% confidence interval we fitted using the data we noticed that only about 10% of the patients will contract AIDS. Which meant the chance of contracting AIDS is generally low. From the Cumulative Hazard function with the 95% confidence interval we can see that the risk of contracting AIDS is approximately 10% and from the confidence interval we see the interval increases as time increases. From the separate survival curves we fitted different categories of interest to compare each levels in the category. For each category we can see the group with CD4<=50 have a higher survival rate, group 2(currently on IV drug) has a higher rate of survival overall and the patients in group 3(previously on IV drug) has the lowest rate of survival. From the survival curve of sex, there does no seem to be any significant differences between the sex. From the survival curve of race, most race are similar except for group 5 (American Indian/Alaskan Natives) who have a significantly lower rate of survival. From the survival curve of Karnofsky Performance Scale we can see the survival rate of groups with lower Karnof score has a higher survival rate. From the survival curve of age, there does not seem to be a significant difference but older patients seem to have a lower survival rate. I then fit different distributions to the dataset using the AIC as my criteria and concluded the lognormal distribution to be the best fit for the dataset. The multiple logistic regression models used resulted in a model, censor ~ time + strat2 + ivdrug + karnof + age.

Summary:

In conclusion from the analysis of our data we can estimate the probability of patients contracting AIDS to be 10%. When comparing different categorical variables in the data we find for each level in the category the patients with the higher survival rate are from group with

CD4<=50 have a higher survival rate, group 2(currently on IV drug) has a higher rate of survival overall and the patients in group 3(previously on IV drug) has the lowest rate of survival. From the survival curve of sex, there does no seem to be any significant differences between the sex. From the survival curve of race, most race are similar except for group 5 (American Indian/Alaskan Natives) who have a significantly lower rate of survival. From the survival curve of Karnofsky Performance Scale we can see the survival rate of groups with lower Karnof score has a higher survival rate. From the survival curve of age, there does not seem to be a significant difference but older patients seem to have a lower survival rate. The lognormal distribution is the best fit for the dataset using the AIC. The multiple logistic models used resulted in a model censor ~ censor ~ time + strat2 + ivdrug + karnof + age.