HIV and AIDs Survival Analysis

Clare Flynn

Introduction

In this analysis, I use data from a study testing the success of different HIV medication by Hammer et al. (1997). The response variables in the study are *time*, which indicates in days the time to death, AIDs diagnosis, or the end of the study, and *censor*, which indicates whether each patient made it to the end of the study without a diagnosis or death, or if the event occured before the end of the study. The explanatory variable of interest is tx, or treatment group, so whether each patient recieved the new treatment with protease inhibitor indinavir (IDV), or the control treatment without the IDV. IF this treatment is found to help delay the time to AIDs diagnosis or death, it could improve the quality of life of many people. There are also many other potential confounding variables recorded, including baseline CD4 count, CD4 count at HIV screening later in the study, sex, race, IV drug use history, hemophilia, Karnofsky score, months of prior zdv use, and age at enrollment. In this study, I will analyze how all of these different variables effect the time to death or AIDs diagnosis.

Methods and Results

I first break 1/4 of the data into a test set and 3/4 into a training sets. My variable analysis and model building will be based completely on the training set.

```
aids <- read.csv(file="AIDSdata.csv")
set.seed(30)
aids.subset <- sample(c(TRUE, FALSE), nrow(aids), replace=TRUE, prob=c(1/4,3/4))
aids.tst <- aids[aids.subset,]
aids.trn <-aids[!aids.subset,]
dim(aids.tst)
```

```
## [1] 210  16
```

```
dim(aids.trn)
```

```
## [1] 641  16
```

Variable Analysis

There are 587 uncensored data points in the training set, meaning that 587 people survived and were not diagnosed with AIDs by the end of the study period. There are 54 censored data points, meaning 54 people either died or were diagnosed with AIDs before the end of the study period. This indicates that only 8.4% of participants died or were diagnosed with AIDs during the one year of the study. Of those censored data, 19 of them died during the study, and 35 were diagnosed with AIDs. Since only 3% of the participants died during the study, I will be doing a survival analysis of time to death or AIDs diagnosis in order to better identify trends.

Table 2. χ^2 table for treatment vs strat2

	Censor	No	Yes	
Strat2	Low	214	40	
	High	373	14	X-sq = 27.699

Table 2. χ^2 table for treatment vs censor

Censor	No	Yes
Control	275	35

Censor	No	Yes	
Treatment	312	19	χ^2=5.6926

A chi-squared test of independence found that there is a relationship between censor and strat2 (p-adj = 8.508e-07, Table 1), so those with high strat2 counts had different rates of diagnosis or death than those with low counts. Another chi-squared test found a relationship between censor by treatment (p = 0.017, Table 2). However, when using Holm's adjusted p-value for multiple comparisons, the relationship between treatment and censor is no longer significant (p-adj = 0.085). There was also a relationship between hemophilia and censor (p-adj = 5.331e-13), though since only 3 participants had hemophilia, the assumptions are violated, so I do not feel comfortable drawing conclusions about hemophilia.

I next examined the relationship between the variables *cd4* and *strat2*, since they are both based on CD4 counts (Figure 1). The odds ratio is 1.078, so for each additional cd4 count, the odds of having a high strat2 score are 1.078 times what they were for one count lower. Based on the Wald's test, we are 95% confident that the true odds ratio is between (1.065, 1.093), so the initial cd4 count is a significant predictor of strat2. Because of this, I do not expect both variables to be needed in the model.

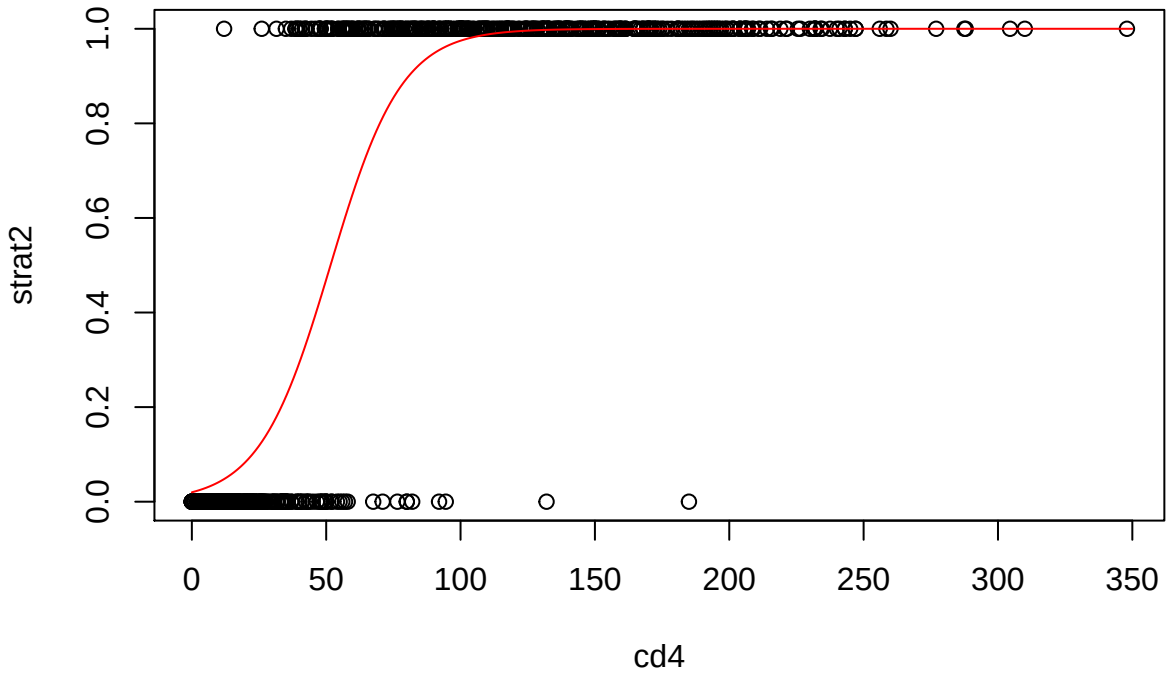

Figure 1. The later CD4 count categories (1 = high, 0 = low) by baseline CD4 count

Kaplan Meier

The following is a Kaplan-Meier Curve for all participants in the training set of the study (Figure 2). The curve is not very steep, since very few participants died or were diagnosed with AIDs during the study period. I then separated the Kaplan-Meier curves by treatment, to assess if treatment effects time to event, and strat2, to see the stage of the disease on time to event (Figure 3).

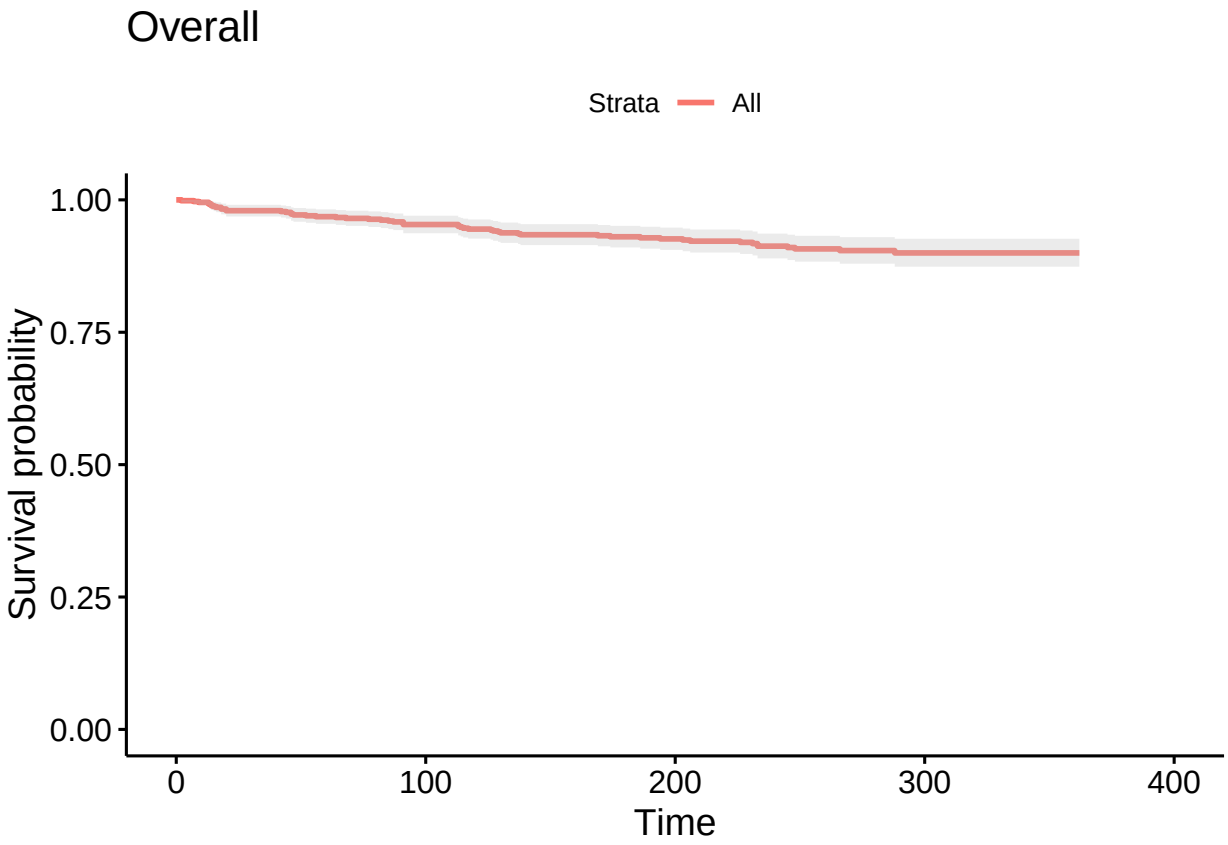

Figure 2. The Kaplan-Meier curve on all of the data where the event is diagnosis or death

```
aids.trn.surv <- survfit(Surv(aids.trn$time,  aids.trn$censor)~tx + strat2, data=aids.trn)
survminer::ggsurvplot(aids.trn.surv, conf.int=TRUE, censor=F) + ggtitle("Overall")
```

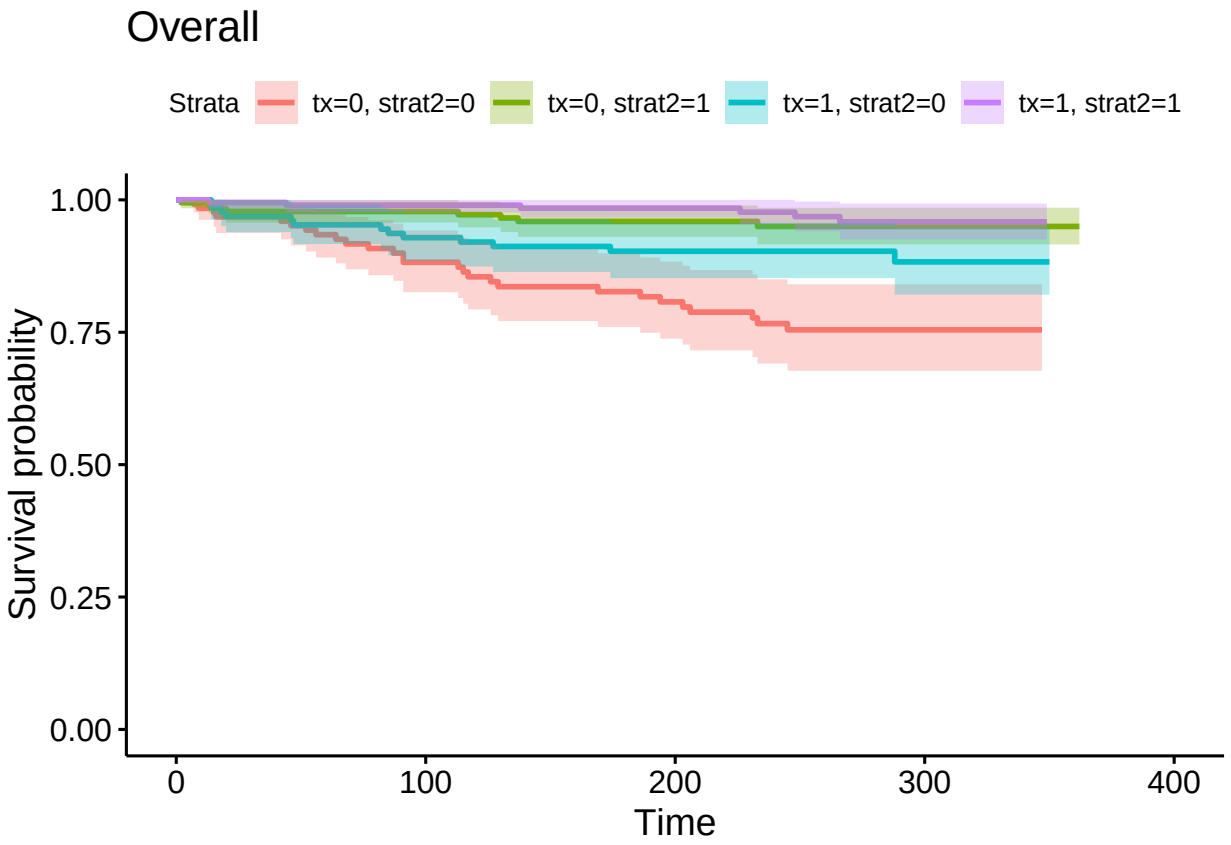

Figure 2. The Kaplan-Meier curve separated by *strat2* and treatment, where the event is diagnosis or death

There is a significant difference in survival time between the four groups based on the Log-Rank test and Wilcoxon test (both p = 4.26e-12, Figure 3). There is an especially noticeable difference in survival times for those with high strat2 counts- those on the treatment appeared to have much higher survival probabilities than those on the control. This indicates that there could be an interaction term needed between strat2 and treatment.

Cox PH

In order to pick a Cox Proportional Hazard model, I used a log-rank test to select the most effective, yet parsimonious model. I used both backwards and forward selection, starting with the following full model, and came up with the two models.

```
coxph(Surv(time,censor) ~ (tx + strat2 + sex + ivdrug  + cd4 +  age + karnof + raceth
                           + hemophil + priorzdv)^2, data=aids.trn, x=TRUE, y=TRUE)
```

I then took out the least significant term one by one, and calculated the drop in deviance for each new model. I continued this process until the drop was significant, then selected the previous model. For the forward selection model, I started with the null model. I then added the most significant term from the full model, and calculated the "drop" in deviance. When it was significant, I then went on and took that term out of the full model to find the next most significant term, and added that to the model. I continued this process until the "drop" in deviance was not significant, then used the previous model.

The best model from the drop in deviance test using backwards selection includes treatment, baseline cd4, and karnofsky score. Though I included interaction terms in the full model, the log-likelihood test determined none of them were necessary in the final model. I then used backwards selection on the full model without interactions to see if the results would be the same, and it did return the same model.

From forward selection process, the best model uses baseline cd4 and karnofsky score. Again, though interaction terms were included, none of them were selected for the final model.

Cross-Validation

I first tested the proportional hazard assumptions for each model, and found that they were not violated. I then used the package *pec* to cross-validate each of my models with the testing dataset. The model of *cd4*, *karnof*, and *tx*, found using backwards selection, had the highest C-index, meaning it had the most concordant pairs, so it predicted the test data the best. I submit the following model to be tested using the remainder of the dataset:

```
flynn.ph <- coxph(Surv(time,censor) ~ cd4 + karnof + tx, data=aids, x=TRUE, y=TRUE)
```

Something New

I used the Weibull model in order to do a more in depth survival analysis. The Weibull model is simial to Kaplan-Meier, but adds in the variables λ, p, and μ, where λ is a scale parameter, and p is a shape parameter and μ is the location parameter (Zhang 2016, Rodriguez 2010).

The equation for the instantaneous hazard (Stevenson 2009) is

$$h(t) = \lambda * p * t^{p-1}$$

The equation for the cumulative hazard (Stevenson 2009) is

$$H(t) = \lambda * t^{p}$$

The equation for the cumulative survival rate (Stevenson 2009) is

$$S(t) = e^{-(\lambda * t)^{p}}$$

And the equation for the pdf (Reliability Engineering Resources 2019) is

$$f(t) = p/\lambda((t-\mu)/\lambda)^{p-1} * e^{-((t-\mu)/\lambda)^{p}}$$

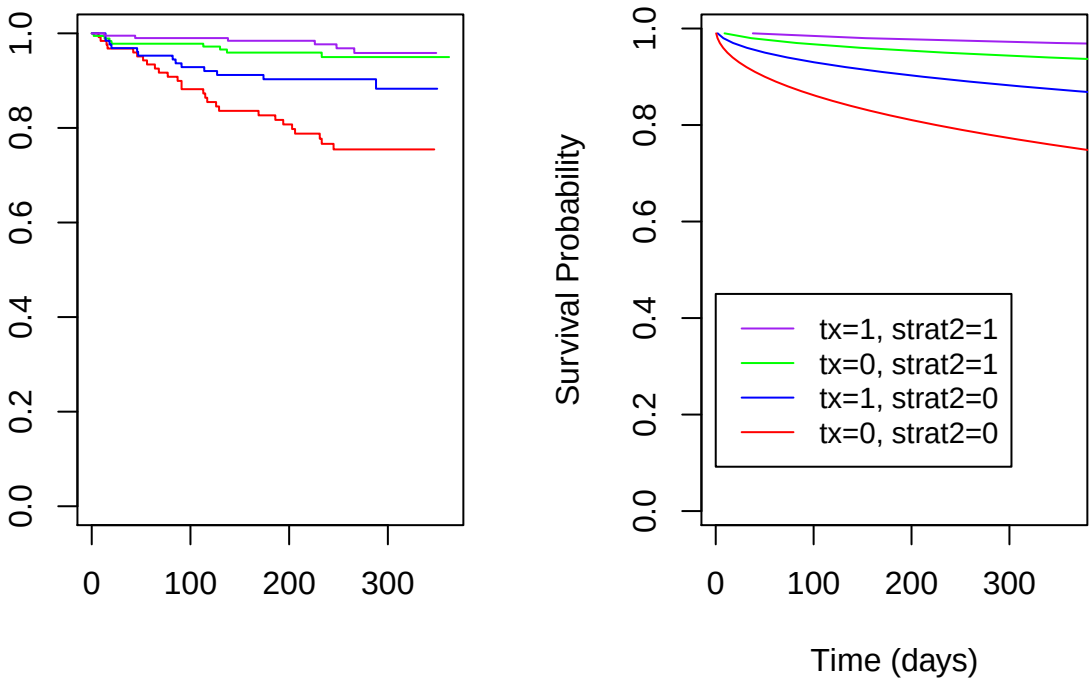

Figure 4. A comparison of Kaplan Meier survival curves (left) and Weibull curves over the one year of the study

```
##
## Call:
## survreg(formula = Surv(aids.trn$time, aids.trn$censor) ~ tx +
##     strat2, data = aids.trn, dist = "weibull", scale = 2)
##             Value Std. Error     z       p
## (Intercept) 8.416      0.374 22.51 < 2e-16
## tx          1.441      0.570  2.53   0.011
## strat2      2.979      0.621  4.80 1.6e-06
##
## Scale fixed at 2
##
## Weibull distribution
## Loglik(model)= -466.9   Loglik(intercept only)= -483.9
##  Chisq= 34 on 2 degrees of freedom, p= 4.1e-08
## Number of Newton-Raphson Iterations: 7
## n= 641
```

The summary returns that both the treatment and the cd4 at stratification were significant predictors of time of diagnosis or death.

The Weibull model has a location parameter, which starts the curve at the time of the first event. The purple curve, treatment and high strat2, doesn't start until about 75 days into the study, because no one no one in that group dies or is diagnosed with AIDs in the first 75 days.

These curves look very similar to the KM curves above, but are much smoother. This is a prediction of

Survival rates based on the actual data, whereas the KM curves are made from the actual data.

Discussion

The best model for predicting HIV/AIDs survivorship used baseline CD4 count, Karnofsky score, and treatment. Both CD4 count and Karnofsky score are measures of the health of the patient, and how far the disease has progressed by the start of the trial (VITAS Healthcare 2014). Essentially, these variables control for the different stages of HIV that participants entered the trial in.

It is very exciting that treatment was included in the model as a treatment of time to diagnosis or death. This indicates that there is a difference in time between those on the treatment vs those in the control group. In this study, those on the treatment had more time to live their lives without and AIDs diagnosis or death (Table 2). More analysis should be done to determine which stage of the disease this treatment works best for, but from this study, we can determine that the treatment shows a lot of promise.

References

Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., ... Cook, J. C. (1997). 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. New England Journal of Medicine, 337(11), 725-733. https://www.nejm.org/doi/full/10.1056/NEJM199709113371101

Reliability Engineering Resources. (2019). Life Data Analysis (Weibull Analysis). Retrieved from https://www.weibull.com/basics/lifedata.htm

Rodriguez, G. (2010). Parametric Survival Models. Retrieved from https://data.princeton.edu/pop509/ParametricSurvival.pdf

Stevenson, M. (2009, June 4). An Introduction to Survival Analysis. Retrieved from http://www.massey.ac.nz/massey/fms/Colleges/College%20of%20Sciences/Epicenter/docs/ASVCS/Stevenson_survival_analysis_195_721.pdf

VITAS Healthcare. (2014). Hospice Eligibility Guidelines for End-Stage HIV & AIDS. Retrieved from https://www.vitas.com/for-healthcare-professionals/evaluating-patients-for-hospice-and-palliative-care/clinical-hospice-guidelines-by-diagnosis/hiv-and-aids/

Zhang, Z. (2016). Parametric regression model for survival data: Weibull regression model as an example. Annals of Translational Medicine, 4(24), 484-484. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5233524/.