Survival Analysis of AIDS

Viral Desai, Fady Naeimm, Caleb Rovell

Introduction

With an epidemic as deadly and as widespread as AIDS, it is important for us to understand the factors that significantly effect the survival rates of those effected. The data set of interest focuses on AIDS patients located in several states in Australia. The 2,843 patients were studied before July 1st, 1991 and we used the resulting data to analyze how different factors effect each person's survival probability. The data comes from: Dr P. J. Solomon and the Australian National Centre in HIV Epidemiology and Clinical Research. Our potential covariates of interest are: the state of origin (state), sex of the patient (sex), time elapsed from the time of diagnosis till death (time), and the way in which the AIDS was transmitted. The four states of origin include New South Wales (NSW), Queensland (QLD), Victoria (VIC), and other smaller areas in Australia (other). Types of transmissions include: male homosexual or bisexual contact (hs), a hs intravenous drug user (hsid), female or heterosexual male intravenous drug user (id), heterosexual contact (het), haemophilia or coagulation disorder (haem), receipt of blood components or tissue (blood), mother with or at risk of HIV infection (mother), and other or unknown transmissions (other). The sex variable is divided up amongst males and females, with 89 female observations and 2754 male observations. The time covariate was one that we created to get the time from diagnosis till their death.

Questions of Interest

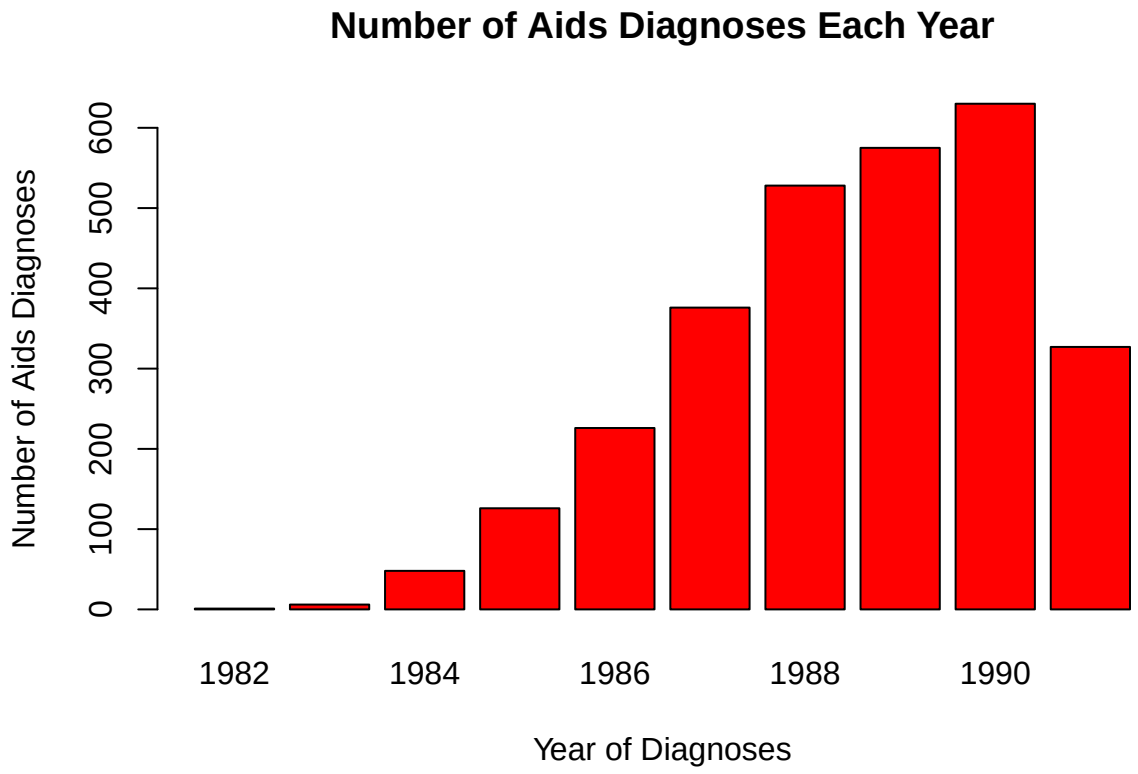
1. Is there siginificant evidence to say that men or women have a higher chance of surviving with AIDS? Can we find a coefficient and confidence interval for the parameters?
2. Does the date of diagnosis affect survival probability?
3. What is the best fitting model for this dataset?
4. Is there a significant difference between hazard ratios based on sex, from start to diagnosis and diagnosis to death?

Survival Analysis Method

1. In order to find if there is a difference in survival probabilities between sexes we will first plot a KM estimate to see how closely they appear. From that we will check the cox PH assumption using a log log plot so we can get a coxph model whose parameters we can estimate.
2. We first split the data into two parts, first being patients diagnosed in the first half of the study, second being the patients diagnosed in the second half of the study. We then used a Kaplan-Meier on these two covariates to determine if there was an increase in survival probability of patients as time went on.
3. We compared the full model containing every possible covariate to each combination of models with a different covariate removed each time. After, we used the likelihood ratio test to find the difference between the full model and each of the reduced models. Then we found the p-value for each LRT function. Based off of these p-values we determined which covariates were signifcant in order to find the best fitting model.
4. To find if there is a difference between hazard rates from start to diagnosis and diagnosis to death we will divide our aids dataset into two parts. We will call the first dataset episode 1 from start to daignosis and episode 2 from diagnosis to death. Then we will do a log likelihood test to see if its significant.

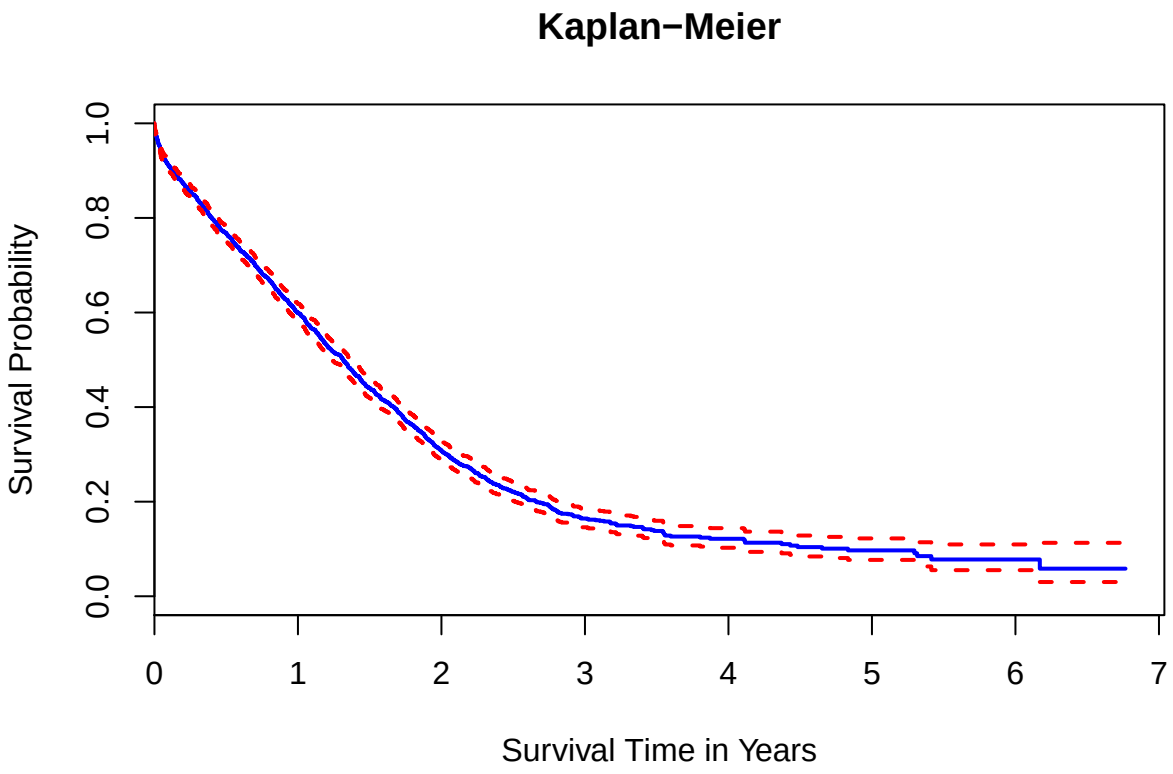
Standard Visualizations

```
E2 = length(diag[which(diag < 8401)])
E3 = length(diag[which(diag >= 8401 & diag < 8766)])
E4 = length(diag[which(diag >= 8766 & diag < 9132)])
E5 = length(diag[which(diag >= 9132 & diag < 9497)])
E6 = length(diag[which(diag >= 9497 & diag < 9862)])
E7 = length(diag[which(diag >= 9862 & diag < 10227)])
E8 = length(diag[which(diag >= 10227 & diag < 10593)])
E9 = length(diag[which(diag >= 10593 & diag < 10958)])
N0 = length(diag[which(diag >= 10958 & diag < 11323)])
N1 = length(diag[which(diag >= 11323 & diag < 11688)])
barplot(c(E2, E3, E4, E5, E6, E7, E8, E9, N0, N1), names.arg = c('1982', '1983', '1984', '1985', '1986'
```



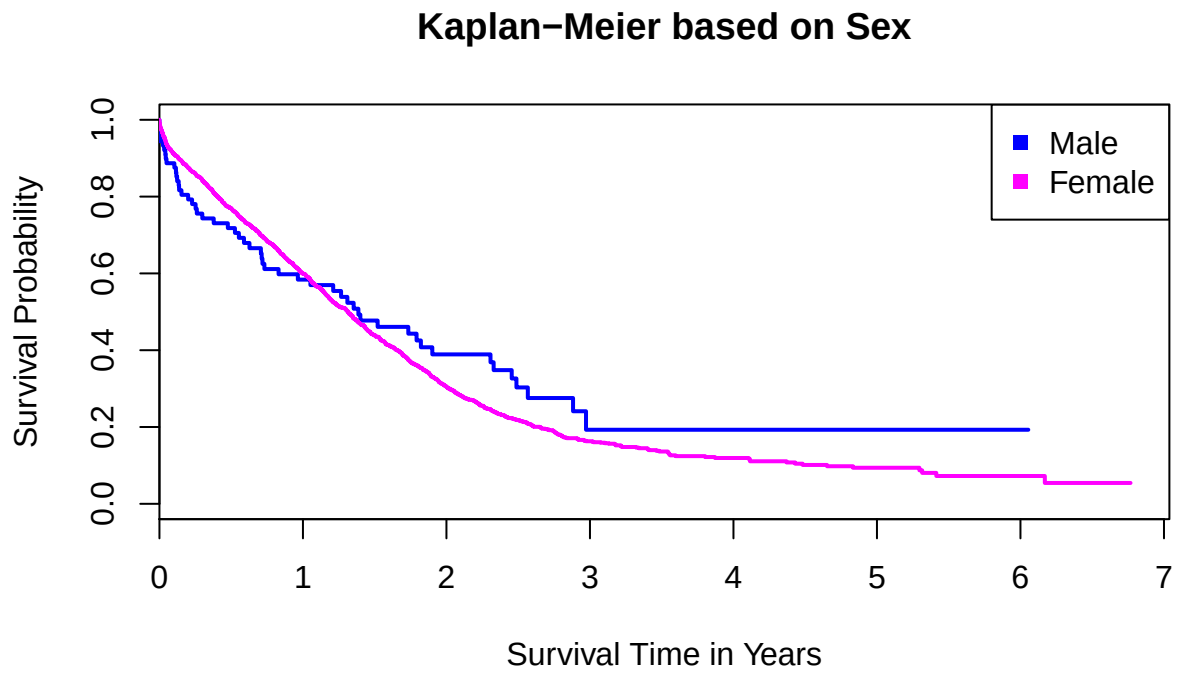
This basic histogram tells us how our dataset is divided. Since our dataset is quite large but length of study is small, we just wanted to see if their was any trend to the number of people diagnosed each year.

```
plot(survfit(Surv(Aids2$time, Aids2$newstatus)~1), col = c(4,2,2), ylab = 'Survival Probability', xlab =
```



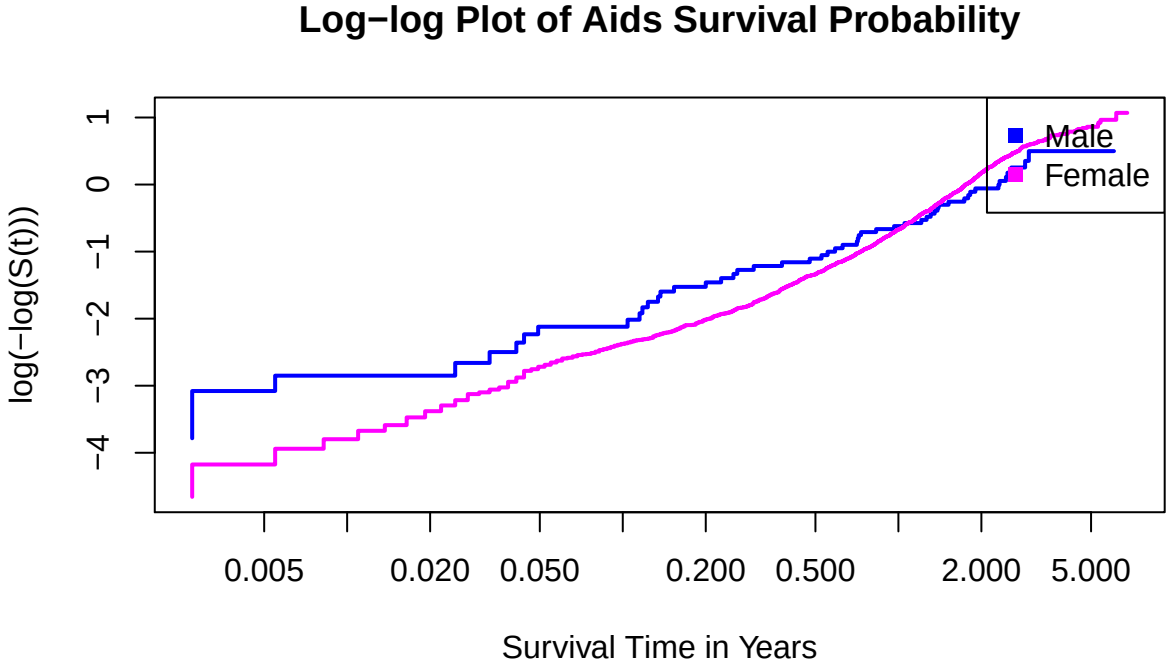
Question 1: Is there siginificant evidence to say that men or women have a higher chance of surviving with AIDS? Can we find a coefficient and confidence interval for the parameters?

```
#plotting KM estimate based on sex
plot(survfit(Surv(Aids2$time, Aids2$newstatus)~sex), ylab = 'Survival Probability', xlab = 'Survival Ti
legend("topright", legend = c("Male", "Female"), col = c(4,6), pch = 15)
```



As we can see from the graph, the lines intersect at a certain point which implies that there is no difference between the sexes. To prove this assumption, we must first see if the PH assumption is met through a log log plot of our estimator. If we see a parallel linear trend then we can assume that the PH assumption is met. From there we will then proceed with a Cox Proportional Hazard model to see if our assumption is correct.

```
plot(survfit(Surv(Aids2$time, Aids2$newstatus)~sex), fun = "cloglog", ylab = 'log(-log(S(t)))',
     xlab = 'Survival Time in Years', col = c(4, 6), lwd = 2, main = "Log-log Plot of Aids Survival Prob
legend("topright", legend = c("Male", "Female"), col = c(4,6), pch = 15)
```



From our log log plot we want to see the lines distance remain constant. In our case our estimators for Male and Female intsersect meaning that they do not fit the PH assumption. One reason for this could be due to the number of observations we have. We only have 89 observations for females and have 2754 for males. This

could be significant but we should use a Likelihood Ratio test to determine its significance.

```
#calculating a p-value to see if there is a difference between sex
summary(coxph(Surv(Aids2$time, Aids2$newstatus)~sex))
```

```
## Call:
## coxph(formula = Surv(Aids2$time, Aids2$newstatus) ~ sex)
##
##   n= 2843, number of events= 1761
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## sexM 0.1266    1.1349   0.1396 0.907    0.365
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sexM     1.135     0.8811    0.8632     1.492
##
## Concordance= 0.499  (se = 0.002 )
## Rsquare= 0   (max possible= 1 )
## Likelihood ratio test= 0.86  on 1 df,   p=0.3551
## Wald test            = 0.82  on 1 df,   p=0.3647
## Score (logrank) test = 0.82  on 1 df,   p=0.3643
```

```
survdiff(Surv(Aids2$time, Aids2$newstatus)~sex)
```

```
## Call:
## survdiff(formula = Surv(Aids2$time, Aids2$newstatus) ~ sex)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F   89       53     59.9     0.795     0.826
## sex=M 2754     1708   1701.1     0.028     0.826
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.363
```

```
#95% CI for beta
c(0.1266 - 1.96*(0.1396), 0.1266 + 1.96*(0.1396))
```

```
## [1] -0.147016  0.400216
```

```
##So our 95% confidence for our estimation is (-0.147016, 0.400216).
#95% CI for hazard rate
exp(c(0.1266 - 1.96*(0.1396), 0.1266 + 1.96*(0.1396)))
```

```
## [1] 0.8632802 1.4921470
```

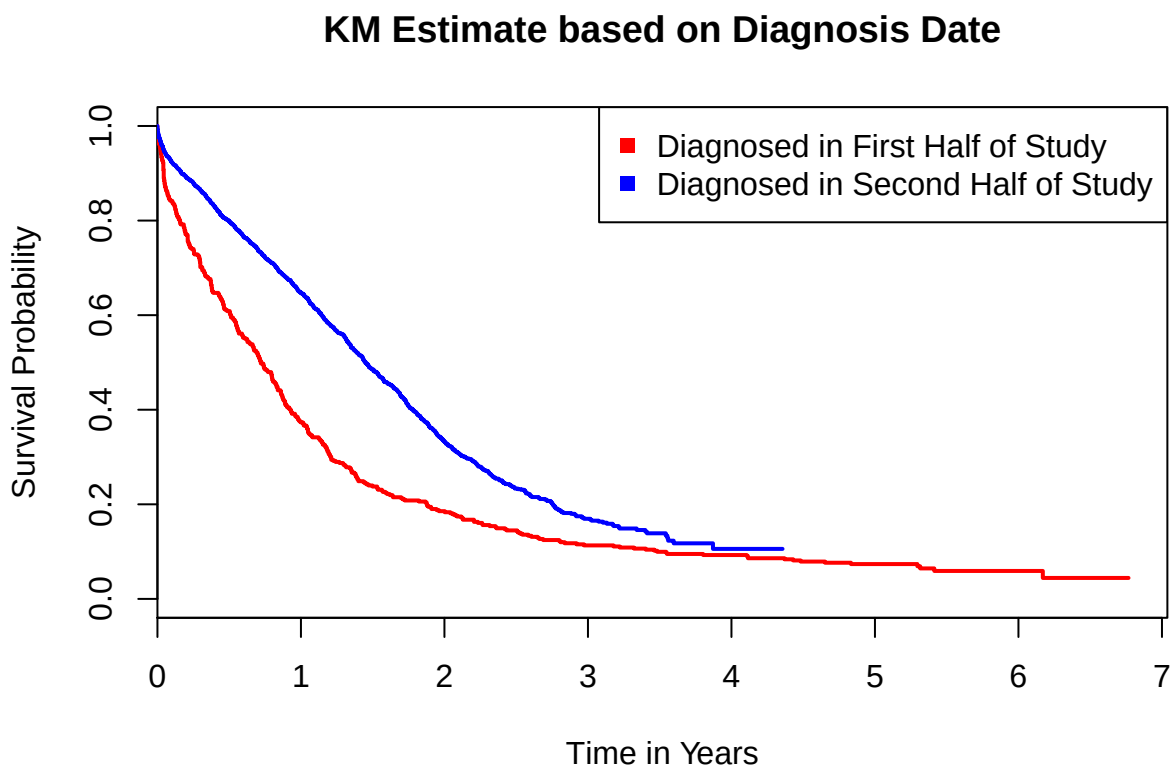
Since both of these tests give a p-value greater than $\alpha = 0.05$, we are able to conclude that there is no significant difference in survival probabilities between sexes. Our summary also gives us the hazard rate of males as 1.1349. This means that males' hazrad rate increases by 13.49% compared to women. Basically this means that at any point in time, women have 13% higher chance of survival. Our 95% confidence interval for our β estimate is (-0.147016, 0.400216). Meaning our hazard probability estimate = (0.8632802 1.4921470). Since this interval includes 1 our conclusion from the p-value was correct.

Question 2: Does the date of diagnosis affect survival probability?

```
#checking the middle point of diagnosis day
(max(Aids2$diag) - min(Aids2$diag)) / 2
```

```
## [1] 1600.5
```

```
#new column label A for before median date and B for after
Aids2['EarlyLate'] <- c(0)
#A = those diagnosed before 9902 and B = those diagnosed after
Aids2$EarlyLate <- ifelse(Aids2$diag >= 8302 & Aids2$diag < 9902, 'A', 'B')
#plotting KM estimate based on date of diagnosis
plot(survfit(Surv(time, newstatus) ~ EarlyLate, data = Aids2), col = c(2,4), lwd = 2, xlab = "Time in Y
legend("topright", legend = c("Diagnosed in First Half of Study", "Diagnosed in Second Half of Study"),
```




```
summary(coxph(Surv(Aids2$time, Aids2$newstatus)~ Aids2$EarlyLate), data = Aids2) #95% CI for estimate
```

```
## Call:
## coxph(formula = Surv(Aids2$time, Aids2$newstatus) ~ Aids2$EarlyLate)
##
##   n= 2843, number of events= 1761
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## Aids2$EarlyLateB -0.49277   0.61093  0.05759 -8.557   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## Aids2$EarlyLateB    0.6109      1.637    0.5457    0.6839
##
```

```
## Concordance= 0.555  (se = 0.005 )
## Rsquare= 0.023   (max possible= 1 )
## Likelihood ratio test= 66.85  on 1 df,   p=3.331e-16
## Wald test            = 73.22  on 1 df,   p=0
## Score (logrank) test = 74.61  on 1 df,   p=0
```

```
c(-0.49277 - 1.96*(0.05759), -0.49277 + 1.96*(0.05759))
```

```
## [1] -0.6056464 -0.3798936
```

```
#95% CI for hazard prob
exp(c(-0.49277 - 1.96*(0.05759), -0.49277 + 1.96*(0.05759)))
```

```
## [1] 0.5457216 0.6839342
```

In our analysis of patients diagnosed in the second half of the study, we can see that their survival probability increases. Unfortunately, we cannot say with certainty why this is, but we could suggest that this may have been due to better technology as time went on. In our earlier graph we can see that the number of people with AIDS has consistently increased. However, due to some factor, patients' survival rates have increased. From our summary we find that our β estimate = -0.49277. Meaning our 95% CI for β = (-0.6056464, -0.3798936). By raising e to our β estimate we get a 95% CI for our hazard rate = (0.5457216, 0.683934). Since this interval does not include 1 we can safely conclude that there is a difference between the early and late group.

Question 3: What is the best fitting model for this dataset?

```
reduced = coxph(Surv(time, newstatus)~1, data = Aids2)
full = coxph(Surv(time, newstatus)~state+sex+T.categ+age, data = Aids2)
step(reduced, scope = list(lower = reduced, upper = full))
```

```
## Warning in is.na(fit$coefficients): is.na() applied to non-(list or vector)
## of type 'NULL'
```

```
## Start:  AIC=24951.14
## Surv(time, newstatus) ~ 1
```

```
## Warning in is.na(fit$coefficients): is.na() applied to non-(list or vector)
## of type 'NULL'
```

```
## Warning in is.na(fit$coefficients): is.na() applied to non-(list or vector)
## of type 'NULL'
```

```
##           Df   AIC
## + age      1 24916
## + T.categ  7 24928
## <none>       24951
## + state    3 24951
## + sex      1 24952
##
## Step:  AIC=24916.13
## Surv(time, newstatus) ~ age
```

```
## Warning in is.na(fit$coefficients): is.na() applied to non-(list or vector)
## of type 'NULL'
```

```
##           Df   AIC
## + T.categ  7 24901
## <none>       24916
```

```
## + state    3 24917
## + sex      1 24918
## - age      1 24951
##
## Step:  AIC=24901.39
## Surv(time, newstatus) ~ age + T.categ
##
##           Df   AIC
## <none>       24901
## + state    3 24903
## + sex      1 24903
## - T.categ  7 24916
## - age      1 24928
```

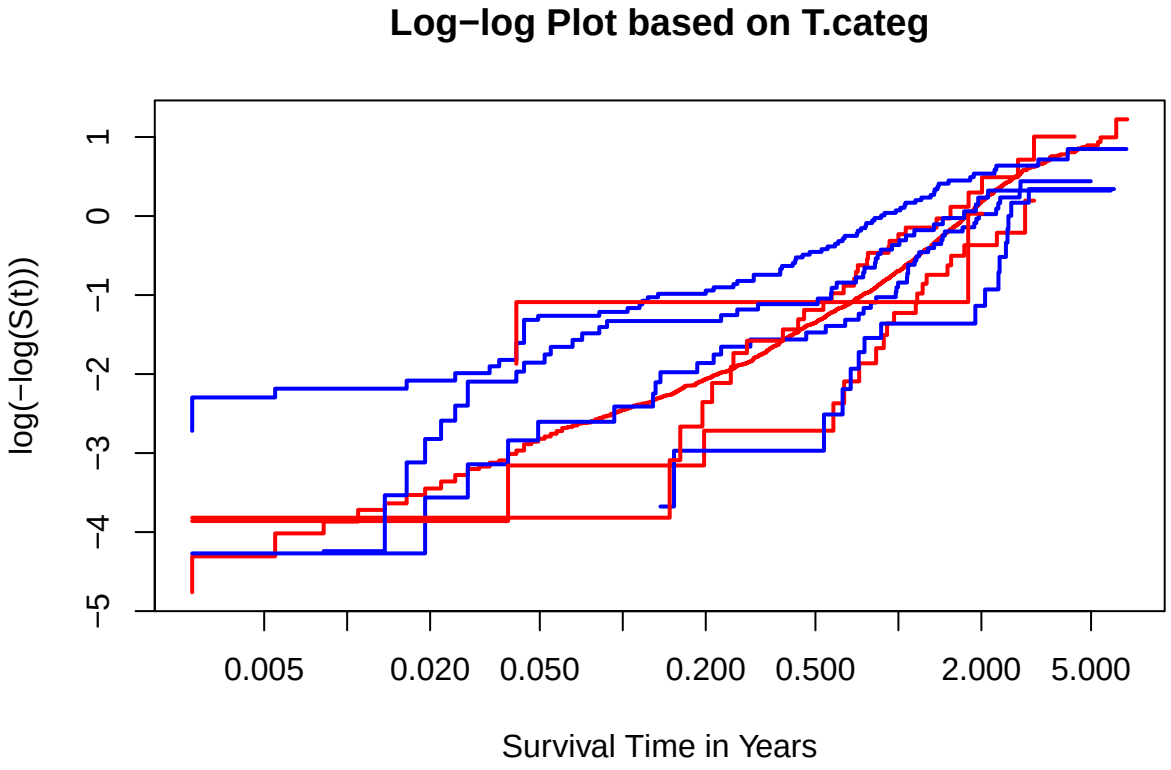
```
## Call:
## coxph(formula = Surv(time, newstatus) ~ age + T.categ, data = Aids2)
##
##                   coef exp(coef) se(coef)     z       p
## age            0.01331   1.01340  0.00249  5.34 9.2e-08
## T.categhsid   -0.07938   0.92369  0.15201 -0.52  0.6015
## T.categid     -0.48203   0.61753  0.23174 -2.08  0.0375
## T.categhet    -0.70760   0.49283  0.24396 -2.90  0.0037
## T.categhaem    0.34035   1.40544  0.18840  1.81  0.0708
## T.categblood   0.37530   1.45542  0.12092  3.10  0.0019
## T.categmother  0.28417   1.32866  0.58397  0.49  0.6265
## T.categother   0.08045   1.08378  0.16074  0.50  0.6167
##
## Likelihood ratio test=65.8  on 8 df, p=3.42e-11
## n= 2843, number of events= 1761
```

The purpose of the AIC is to estimate the quaality of all the models compared to eachother. Thus, the lowest AIC value provides us with the best model. Using this, we are able to select our final model to include age and transmission categories as our covariates.

```
final_model <- coxph(Surv(time, newstatus) ~ T.categ, data = Aids2)
#checking if our PH assumptions are met
cox.zph(final_model)
```

```
##                    rho   chisq        p
## T.categhsid   -0.01464  0.3774 5.39e-01
## T.categid      0.01058  0.1970 6.57e-01
## T.categhet     0.03648  2.3364 1.26e-01
## T.categhaem    0.00264  0.0122 9.12e-01
## T.categblood  -0.13315 31.8290 1.68e-08
## T.categmother -0.01711  0.5151 4.73e-01
## T.categother  -0.06345  7.0779 7.80e-03
## GLOBAL               NA 41.8191 5.63e-07
```

```
##Since our PH assumptions are not met we should check our covariates' log log graph.
#log log plot based on transmission category
plot(survfit(Surv(time, newstatus) ~ T.categ, data = Aids2), col = c(2,4), lwd = 2, fun = "cloglog",
     ylab = 'log(-log(S(t)))', xlab = 'Survival Time in Years', main = "Log-log Plot based on T.categ")
```



As we can see from these graphs our PH assumptions are clearly violated since their is non-parallel lines between the any of the different coviarate factors so a cox ph model isn't exactly the best model. Since our assumptions are not met we should stratify age so we estimate a baseline hazard function for the entire age category.

```
## The following objects are masked from Aids2 (pos = 3):
##
##     age, death, diag, sex, state, status, T.categ
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, newstatus)
## Terms added sequentially (first to last)
##
##          loglik Chisq Df Pr(>|Chi|)
## NULL    -6067.3
## T.categ -6050.6 33.42  7  2.211e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By using a stratified model we can clearly see that certain transmission categories have a statistically significant effect on Aids survival probability.

Question 4: Is there a significant difference between episode and sex from start to diagnosis and diagnosis to death?

```
#copying Aids2 dataset
#first dataframe from 0 to diagnosis
Aids_part1 <- Aids2[,]
Aids_part1['start'] <- c(0)
```

```r
Aids_part1['stop'] <- Aids_part1$diag
Aids_part1['episode'] <- c(1)
#second dataframe from diagnosis to death
Aids_part2 <- Aids2[,]
Aids_part2$time <- Aids_part2$time * 365
Aids_part2['start'] <- Aids_part2$diag
Aids_part2['stop'] <- Aids_part2$death
Aids_part2['episode'] <- c(2)
#r binding both datasets into ONE dataframe
new_aids <- rbind(Aids_part1, Aids_part2)
new_aids$episode <- factor(new_aids$episode)
#showing first 10 rows of our data
head(new_aids)
```

```
##   state sex  diag death status T.categ age newstatus      time EarlyLate
## 1   NSW   M 10905 11081      D      hs  35         1 0.4821918         B
## 2   NSW   M 11029 11096      D      hs  53         1 0.1835616         B
## 3   NSW   M  9551  9983      D      hs  42         1 1.1835616         A
## 4   NSW   M  9577  9654      D    haem  44         1 0.2109589         A
## 5   NSW   M 10015 10290      D      hs  39         1 0.7534247         B
## 6   NSW   M  9971 10344      D      hs  36         1 1.0219178         B
##   start  stop episode
## 1     0 10905       1
## 2     0 11029       1
## 3     0  9551       1
## 4     0  9577       1
## 5     0 10015       1
## 6     0  9971       1
```

```r
#showing last 10 rows of our data
tail(new_aids)
```

```
##      state sex  diag death status T.categ age newstatus time EarlyLate
## 5681 Other   M 11359 11504      A      hs  27         0  145         B
## 5682 Other   M 11475 11504      A     het  46         0   29         B
## 5683 Other   F 11420 11504      A     het  34         0   84         B
## 5684 Other   M 11496 11504      A    haem  49         0    8         B
## 5685 Other   M 11460 11504      A      hs  55         0   44         B
## 5686 Other   M 11448 11504      A      hs  37         0   56         B
##      start  stop episode
## 5681 11359 11504       2
## 5682 11475 11504       2
## 5683 11420 11504       2
## 5684 11496 11504       2
## 5685 11460 11504       2
## 5686 11448 11504       2
```

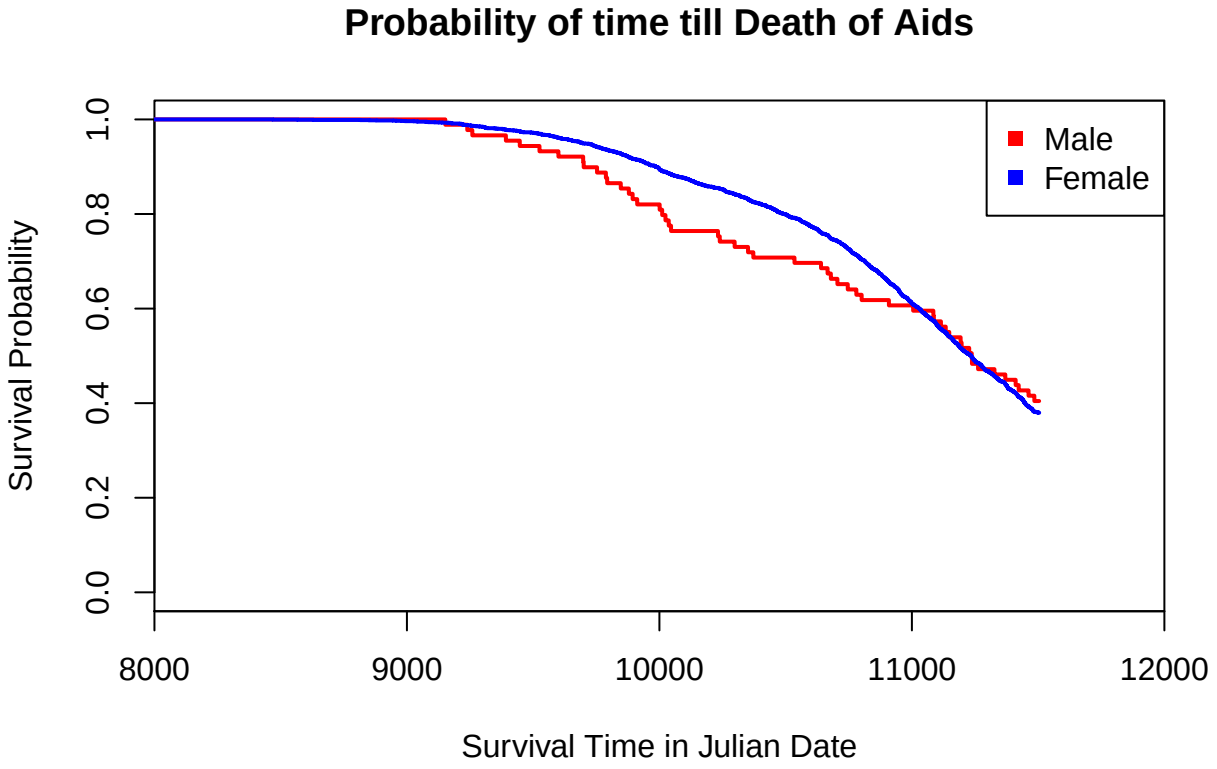
```r
#our model from start stop to finish
anova(coxph(Surv(start,stop,newstatus) ~ strata(episode) * sex, data = new_aids))
```

```
## Warning in Surv(start, stop, newstatus): Stop time must be > start time, NA
## created
```

```
## Warning in Surv(start, stop, newstatus): Stop time must be > start time, NA
## created
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(start, stop, newstatus)
## Terms added sequentially (first to last)
##
##                       loglik  Chisq Df Pr(>|Chi|)
## NULL                  -23910
## sex                   -23910 0.1794  1     0.6719
## strata(episode):sex -23909 1.5516  1     0.2129
```

Since our p-value for interaction between episode and sex is 0.2129 we can say there is not a significant of hazaard raters between sexes from start to diagnosis and from diagnosis to death. This means that there wasn't a significant difference between sexes from birth to diagnosis date and from diagnosis date to death.



This is a plot of the marginal model calculating the time from study entry to second event or death in this case. (Time started at 0 but we cut it off so we could actually see the ending of the graph clearly.)

Conclusion

By first creating basic visualizations, we were able to see from our data that more people were diagnosed later on in the study than in the beginning and a simple kaplan meier estimate from our data. Then, by using a Cox Proportional Hazards model we were able to prove that there is no statistically significant difference in survival probabilities between the 2 sexes. In addition, we were able to split our data into two categories one for observations between 1982-1987 and the other from 1987-92 which showed us that patients diagnosed with AIDS in the second group had a high survival probabilty. Our greatest challenge with working on this dataset was the use of Julian dates which severely hindered us from trying to split our dataset by using time variate covariates instead of simply splitting the dataset into groups. Something we wish we had been able to add to our report was to plot a weibull distrubtion of those diagnosed after 1987, so that we are able to extend our survival probabibilty.

References

Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.