Florida Scrub-Jay Survival

Erin Feichtinger

Monday, February 21, 2016

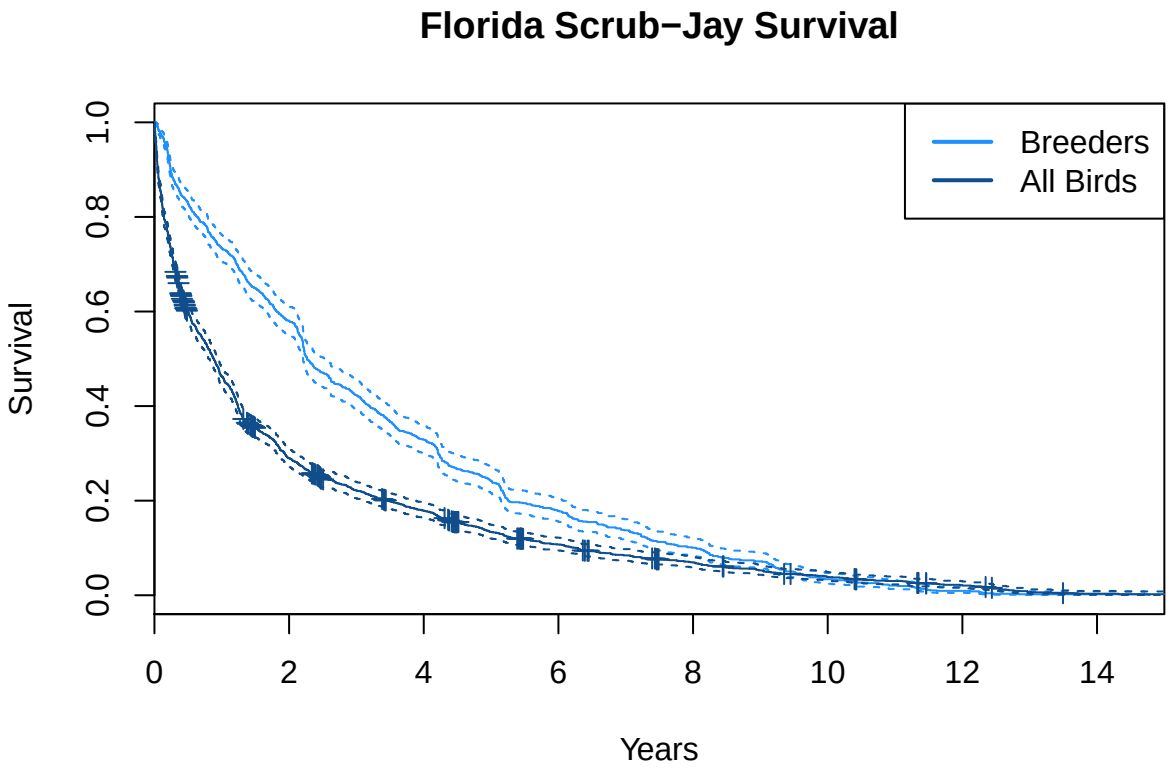
The last report I made on this research project was on January 25, 2016. I estimated survival over time using the Kaplan-Meir estimate. I constructed a survival curve for all known-age birds in the population from fledge date to date last seen. Then, I made a survival curve for breeders in the population.

I realize my error on the procedure for determining if first time breeders have a different survival probability than experienced breeders. I was not correctly thinking about what censorship is. Now, I think I understand. I am using years of experience as a covariate in the models where all breeders are included (both known and unknown age). The input data will have multiple rows per individual jay (if the bird bred more than once), the date first recorded as a breeder and date last seen or still alive and some other information about each bird.

The following output is from two subsets of the jay data from 1981 to 2015. The first data set includes all known-age birds with a fledge date and a date last observed. The second data set includes all breeders in the population from 1981 to 2015. This has both known and unknown age birds. Unknown age birds have a minimum age, however. These data are mainly for practice. Next step is to add breeder experience to the breeder data set (keep reading!).

```
## Loading required package: Matrix
```

```
## Loading required package: quadprog
```



The figure generated has two K-M curves on it. One for all known-age birds and the other is for breeders, which includes unknown-age birds. The curves do have different shapes, especially in the beginning of the interval. As I'm writing this and looking at the graph, I had a thought. Does it matter that for the breeder curve, the time period starts when birds become breeders, so year 0 is at least age 1. However, for all birds, year 0 is when they fledged. I think I know what I need to do to work this out. I would have expected to see a steep decline in survival at first because young birds are more vulnerable to predation than older ones generally.

Just as a place to start, let's fit models with an intercept only using both sets of data (all birds and breeders only) - these data primarily for practice

```
#All known age birds first

#Cox model
all.cox <- coxph(survobj~1, data= birds2)
#Exponential
all.exp <- survreg(Surv(birds2$days, birds2$censorship)~1, dist="exponential")
#Weibull
all.weib <- survreg(survobj~1, dist = "weibull")

summary(all.cox)
```

```
## Call:  coxph(formula = survobj ~ 1, data = birds2)
##
## Null model
##   log likelihood= -14567.93
##   n= 2309
```

```
summary(all.exp)
```

```
##
## Call:
## survreg(formula = Surv(birds2$days, birds2$censorship) ~ 1, dist = "exponential")
##             Value Std. Error   z p
## (Intercept)  6.59     0.0216 305 0
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -16267.5   Loglik(intercept only)= -16267.5
## Number of Newton-Raphson Iterations: 5
## n= 2309
```

```
summary(all.weib)
```

```
##
## Call:
## survreg(formula = survobj ~ 1, dist = "weibull")
##             Value Std. Error    z         p
## (Intercept) 0.444     0.0339 13.1  2.87e-39
## Log(scale)  0.417     0.0170 24.6 2.48e-133
```

```
##
## Scale= 1.52
##
## Weibull distribution
## Loglik(model)= -3250.7   Loglik(intercept only)= -3250.7
## Number of Newton-Raphson Iterations: 6
## n= 2309
```

```
#Breeders only
breed.cox <- coxph(my.survyr~1, data= brdrs.new)
breed.exp <- survreg(Surv(brdrs.new$days, brdrs.new$censorship)~1, dist="exponential")
breed.weib <- survreg(my.survyr~1, dist = "weibull")

summary(breed.cox)
```

```
## Call:  coxph(formula = my.survyr ~ 1, data = brdrs.new)
##
## Null model
##   log likelihood= -5746.62
##   n= 976
```

```
summary(breed.exp)
```

```
##
## Call:
## survreg(formula = Surv(brdrs.new$days, brdrs.new$censorship) ~
##     1, dist = "exponential")
##             Value Std. Error   z p
## (Intercept)  7.09      0.032 221 0
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -7893.2   Loglik(intercept only)= -7893.2
## Number of Newton-Raphson Iterations: 4
## n= 976
```

```
summary(breed.weib)
```

```
##
## Call:
## survreg(formula = my.survyr ~ 1, dist = "weibull")
##               Value Std. Error     z         p
## (Intercept)  1.1946     0.0330 36.22 3.42e-287
## Log(scale)  -0.0201     0.0258 -0.78  4.36e-01
##
## Scale= 0.98
##
## Weibull distribution
## Loglik(model)= -2133.9   Loglik(intercept only)= -2133.9
## Number of Newton-Raphson Iterations: 7
## n= 976
```

Now, what to make of the output? The models above are with no predictors so it should just estimate the intercept. I think what the output is saying is that the intercept is not zero, and these models are not very good. I would like to spend some time reviewing this with Gordon. A while back Gordon showed me something using the car package but I don't remember exactly what it was. I think he was able to get more information on the estimates in the model and somehow he got an object that showed all the years that are estimated in the model. I don't remember how he did that, though.

```
#Fit a Cox PH model with sex as a predictor
brsex.cx <- coxph(my.survyr~brdrs.new$Sex, data = brdrs.new)
summary(brsex.cx)
```

```
## Call:
## coxph(formula = my.survyr ~ brdrs.new$Sex, data = brdrs.new)
##
##   n= 976, number of events= 976
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## brdrs.new$SexM -0.03615   0.96449  0.06424 -0.563    0.574
##
##                exp(coef) exp(-coef) lower .95 upper .95
## brdrs.new$SexM    0.9645      1.037    0.8504     1.094
##
## Concordance= 0.497  (se = 0.009 )
## Rsquare= 0   (max possible= 1 )
## Likelihood ratio test= 0.32  on 1 df,   p=0.5736
## Wald test            = 0.32  on 1 df,   p=0.5736
## Score (logrank) test = 0.32  on 1 df,   p=0.5736
```

The model output says sex is not a significant predictor of hazard (I think). This is a little surprising but I wonder if I will see a different result when age is a covariate. I would think younger breeder females would have a higher risk than older ones because females tend to disperse farther into potentially unfamiliar territory.

```
#Fit a Cox PH model with first year as a breeder as predictor
bryr.cx<- coxph(my.survyr~brdrs.new$Yr, data=brdrs.new)
summary(bryr.cx)
```

```
## Call:
## coxph(formula = my.survyr ~ brdrs.new$Yr, data = brdrs.new)
##
##   n= 976, number of events= 976
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## brdrs.new$Yr -0.01627   0.98386  0.00359 -4.533 5.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## brdrs.new$Yr    0.9839      1.016     0.977    0.9908
##
## Concordance= 0.532  (se = 0.011 )
## Rsquare= 0.021   (max possible= 1 )
## Likelihood ratio test= 20.28  on 1 df,   p=6.695e-06
```

```
## Wald test            = 20.55  on 1 df,   p=5.812e-06
## Score (logrank) test = 20.64  on 1 df,   p=5.535e-06
```

Here, this suggests that the year in which a bird first bred is significant, meaning there is year to year variation (I think). Not really surprising given that we already know there can be a lot of variation from year to year in conditions.

```
#Fit a Cox PH model with age/min age at first breeding as predictor
brage.cx <- coxph(my.survyr~brdrs.new$AgeFirstBreed)
summary(brage.cx)
```

```
## Call:
## coxph(formula = my.survyr ~ brdrs.new$AgeFirstBreed)
##
##   n= 958, number of events= 958
##    (18 observations deleted due to missingness)
##
##                             coef exp(coef) se(coef)     z Pr(>|z|)
## brdrs.new$AgeFirstBreed 0.08377   1.08737  0.03328 2.517   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## brdrs.new$AgeFirstBreed     1.087     0.9196     1.019     1.161
##
## Concordance= 0.51  (se = 0.01 )
## Rsquare= 0.006   (max possible= 1 )
## Likelihood ratio test= 6  on 1 df,   p=0.01432
## Wald test            = 6.34  on 1 df,   p=0.01183
## Score (logrank) test = 6.32  on 1 df,   p=0.01193
```

I think the output is telling us that age at first breeding is significant but I would like to wait until I get the "new" data set to really say something about this. I have to look at the relationship between years breeding and age to see how correlated they are. Make a plot of years breeding vs. age to see what the relationship looks like.

Also this week, I consulted with Dr. Angela Tringali, the new post doc/lab manager in the Avian Ecology lab at Archbold on some questions I had about the data.

On time to event models and censored regression

1. How do I test the assumption that the hazards are proportional?
2. How to build the likelihood function with right censored data points?

As far as I know, the only censored data in the jay data are birds that are still alive and these would be right-censored. From Gordon's book chapter, I have come to realize that with censored regression we can estimate the probability that any value of y is censored. Do I need to worry about this for the jay data? He also writes that we need a likelihood function that uses information about the censored data. Eqn 5.4 in the book is the log-likelihood for N data points where k of which are censored.

I had to refresh my memory on concepts related to survival modeling, including the following:

*Survival function S(t) is the probability an individual with survive beyond time t, $\Pr\{T > t\}$, which is also the compliment of the cumulative distribution function $F(t) = \Pr\{T < t\}$ which is the probability that the event occurs by time t.

*Hazard function h(t) is the instantaneous rate of the even occurring/instantaneous risk of dying, or other words, the number of individuals experiencing event in interval beginning with t divided by the number of individuals surviving at time t times the interval width.

*Cumulative hazard - sum of risks faced in the duration 0 to t.

Question - Are the jay data an example of Type I censoring?

Example: If the unit i is observed for some time and it dies at time t, its contribution to the likelihood function is the density at that duration, which is the product of the survivor and hazard functions (Rodriguez 2007).
I definitely need some more guidance on this whole likelihood business and how it fits into my modeling. Let's talk about this Gordon!

Model types

*Accelerated Life Models

*Weibull Model - parametric approach

*Exponential - parametric approach

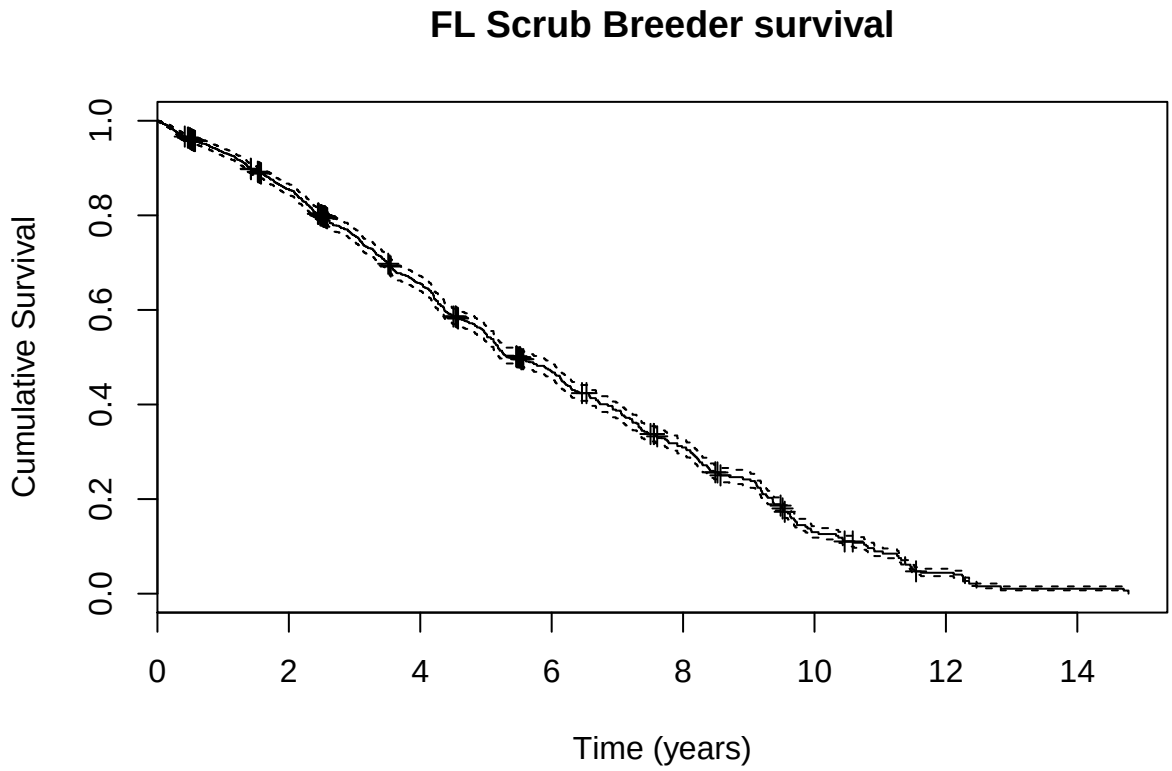
*Cox Proportional Hazards - semi parametric, baseline hazard is non-parametric

*Cox models with time-dependent and time-varying effects

In the R package survival, the survival object Surv has time and event as a status indicator. The info page says "for right censored data, this is the follow up time". Is this the whole time a jay was alive? Or do I use time and time2 and put in the start date and the end date?

```
library(survival)
library(car)
library(kinship2)

##Read in CSV file of male and female breeders with mulitple rows for each bird
bird.df <- read.csv("Erin_Breeders_All_Years.csv")
#str(bird.df)

#remove duplicates  - for years where there was more than one nest in a year
jay.df<- bird.df[!duplicated(bird.df),]
#str(jay.df)

colnames(jay.df)[1] <- "ID"
colnames(jay.df)[2] <- "Band"

#convert dates to date format
jay.df$MinDate <- as.Date(jay.df$MinDate, format = "%m/%d/%Y")
jay.df$LastObsDate <- as.Date(jay.df$LastObsDate, format = "%m/%d/%Y")

#subtract dates to get number of days
date.diff<- jay.df$LastObsDate-jay.df$MinDate

#and survival period in years, account for leap year
jay.df["Yrs"] <- date.diff/365.25

#very important piece of code for the model to work properly, remove any
```

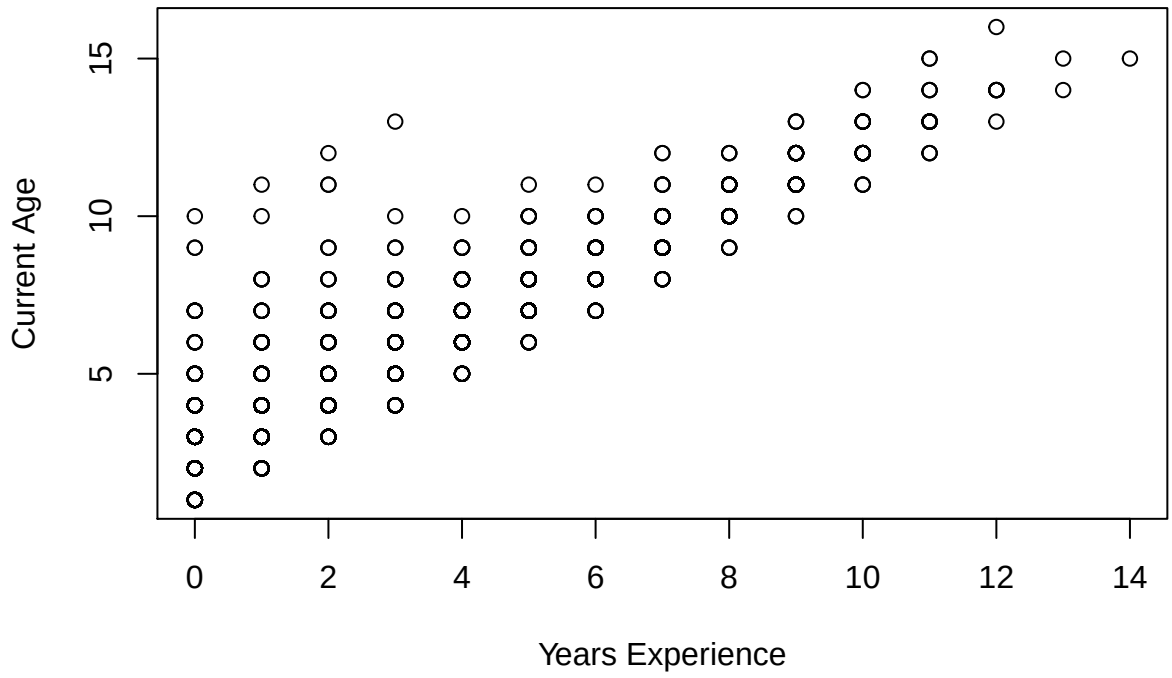
```
#weird entries like birds that have negative years of experience or a negative
#survival interval
jay.df <- subset(jay.df, jay.df$YrsExp >= 0 & jay.df$Yrs > 0)
#str(jay.df)

#add column for censorship status, in survival package - 0=alive, 1=dead
jay.df["censorship"] <- 1
#If last observed date = 10/14/2015, 0 for still alive today
jay.df$censorship[which(jay.df$LastObsDate=="2015-10-14")]<-0


#change back to numeric for survival object
jay.df$MinDate <- as.numeric(jay.df$MinDate)
jay.df$LastObsDate <- as.numeric(jay.df$LastObsDate)
jay.df$Yrs <- as.numeric(jay.df$Yrs)

#Create survival object - IS THIS CORRECT??
jay.ob <- Surv(jay.df$Yrs, jay.df$censorship, type =c('right'))
jay.lifetab <- survfit(jay.ob~1)
jay.fit <- plot(jay.lifetab, xlab = "Time (years)",
                ylab = "Cumulative Survival", main = "FL Scrub Breeder survival")
```



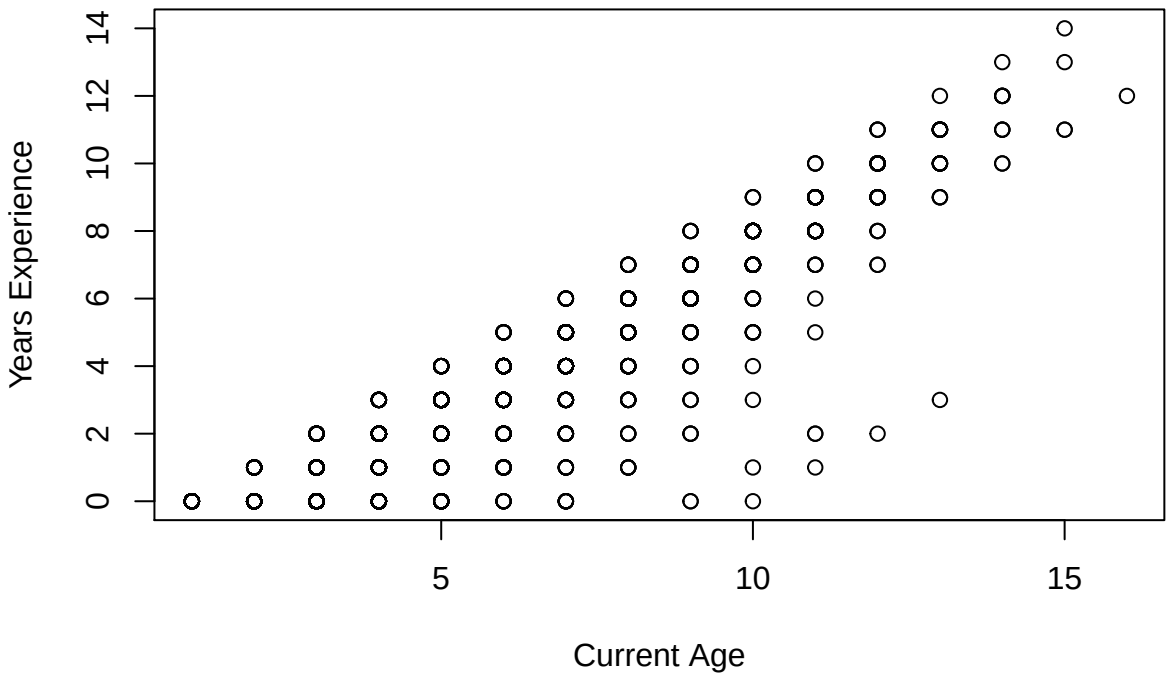
So, the cumulative survival of breeders looks linearish over time (if I did this correctly, that is!). Compare to the first figure. Just a note, the breeder data set that I used in the first figure only has one row per individual with a date starting breeding and a death or censorship date (meaning still alive as of 10/14/2015). This

data set however, has multiple rows for individuals for each year bred. That way I have a column for age and breeder experience at any given year for each bird.

```
plot(jay.df$YrsExp, jay.df$CurrentAge, xlab = "Years Experience",
     ylab = "Current Age")
```




```
plot(jay.df$CurrentAge, jay.df$YrsExp, xlab = "Current Age",
     ylab = "Years Experience")
```



I've been doing some mental gymnastics on which way this should go. My first thought is that I want to ask if your years of experience is a function of age? As birds get older, the number of years of experience increases (obviously, it has too). However, there are several possible ages at each level of experience. Obviously, the possible ages a bird could be decreases as breeder experience increases, for example, a bird with 12 years of experience has to be at least 12 years old, but it could be 13, 14, or 15. Looking at the first plot (years experience on the x axis), we see that there are birds from age 1 to 10 breeding for the first time. So, I think this means we can look at these two things separately?

```
#Fitting some basic Cox Models
#First Cox Model
cox1 <- coxph(jay.ob ~ YrsExp, data = jay.df)
cox2 <- coxph(jay.ob ~ Sex, data = jay.df)
cox3 <- coxph(jay.ob ~ CurrentAge, data = jay.df)
summary(cox1)
```

```
## Call:
## coxph(formula = jay.ob ~ YrsExp, data = jay.df)
##
##   n= 3623, number of events= 3168
##
##             coef exp(coef)  se(coef)      z Pr(>|z|)
## YrsExp -0.193522  0.824052  0.007145 -27.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## YrsExp    0.8241      1.214    0.8126    0.8357
```

```
##
## Concordance= 0.72  (se = 0.006 )
## Rsquare= 0.21   (max possible= 1 )
## Likelihood ratio test= 853.4  on 1 df,   p=0
## Wald test            = 733.5  on 1 df,   p=0
## Score (logrank) test = 774.7  on 1 df,   p=0
```

```
summary(cox2)
```

```
## Call:
## coxph(formula = jay.ob ~ Sex, data = jay.df)
##
##   n= 3623, number of events= 3168
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## SexM -0.22138   0.80141  0.03606 -6.14 8.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## SexM    0.8014      1.248    0.7467    0.8601
##
## Concordance= 0.512  (se = 0.005 )
## Rsquare= 0.01   (max possible= 1 )
## Likelihood ratio test= 37.73  on 1 df,   p=8.107e-10
## Wald test            = 37.7  on 1 df,   p=8.258e-10
## Score (logrank) test = 37.84  on 1 df,   p=7.66e-10
```

```
summary(cox3)
```

```
## Call:
## coxph(formula = jay.ob ~ CurrentAge, data = jay.df)
##
##   n= 3559, number of events= 3119
##    (64 observations deleted due to missingness)
##
##                 coef exp(coef)  se(coef)      z Pr(>|z|)
## CurrentAge -0.164576  0.848253  0.006738 -24.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## CurrentAge    0.8483      1.179    0.8371    0.8595
##
## Concordance= 0.694  (se = 0.006 )
## Rsquare= 0.17   (max possible= 1 )
## Likelihood ratio test= 665.2  on 1 df,   p=0
## Wald test            = 596.6  on 1 df,   p=0
## Score (logrank) test = 619.5  on 1 df,   p=0
```

```
extractAIC(cox1)
```

```
## [1]     1.00 44612.17
```

```
extractAIC(cox2)
```

```
## [1]     1.00 45427.84
```

```
extractAIC(cox3)
```

```
## [1]     1.0 44007.9
```

My suspicion is that I did not do something correctly because the degrees of freedom used is one. That doesn't seem right...

```
AFT.exp <- survreg(jay.ob ~ YrsExp, data = jay.df, dist = "exponential")
summary(AFT.exp)
```

```
##
## Call:
## survreg(formula = jay.ob ~ YrsExp, data = jay.df, dist = "exponential")
##             Value Std. Error    z        p
## (Intercept)  1.44    0.02543 56.6 0.00e+00
## YrsExp       0.13    0.00716 18.2 5.48e-74
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -8776.5   Loglik(intercept only)= -8962.9
##  Chisq= 372.74 on 1 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 4
## n= 3623
```

```
AFT.weibull <- survreg(jay.ob ~ YrsExp, data = jay.df, dist = "weibull")
summary(AFT.weibull)
```

```
##
## Call:
## survreg(formula = jay.ob ~ YrsExp, data = jay.df, dist = "weibull")
##              Value Std. Error     z         p
## (Intercept)  1.558    0.01512 103.1  0.00e+00
## YrsExp       0.105    0.00416  25.3 1.77e-141
## Log(scale)  -0.554    0.01413 -39.2  0.00e+00
##
## Scale= 0.574
##
## Weibull distribution
## Loglik(model)= -8180.2   Loglik(intercept only)= -8548.6
##  Chisq= 736.94 on 1 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 6
## n= 3623
```

Well, not really sure where to go from here. I don't know how to interpret the model output or IF I even did it correctly! The df seems suspect.