University of California, Santa Barbara
Department of Probability and Statistics

PSTAT 175 Final Project

Survival Analysis of NBA Career Length

Report by: Nathan Hwangbo, Austin Miles, Wen Tao Liao

Instructor: Adam Tashman

December 6th, 2019

Abstract

Basketball players have been prominent figures in social media. People often talk about their career in the field. Our project focuses on the questions: does round of drafting have an influence on the players' longevity of their career? How does position have an influence on the players' longevity of their career? Overall how does our covariates affect career length changed over time? We build a Cox proportional Hazards model to see if there is a significant difference between players who are round draft and players who are picked elsewhere.

Data source and Background information

https://www.basketball-reference.com/

Our data mainly focus on the NBA dataset which center around 1998 (Vince Carter's draft class), 2004 (Lebron/carmelo's draft class) and lastly 2009 - Steph Curry, James Harden draft class.

We have three fixed covariates: draft year of the player, position of the player on the team, and lastly draft round, an estimated of the player's skill on the game.

The variables are: draft year of the player (1998 ,2003 ,2009), position of the player on the team (Consolidated positions into Smalls (guards) and Bigs (everyone else), draft round, skill of the player (1 or 2) to proxy for skill

We consider an observation to be censored if the player is still active, or if they have played fewer than five games in the NBA. Our research question is concerned primarily of the career length of players who have made a living as a professional basketball player, which is not the case if they were cut from the NBA before 5 games. Then our dataset contains 165 players, 31 of whom are censored.

Research Question

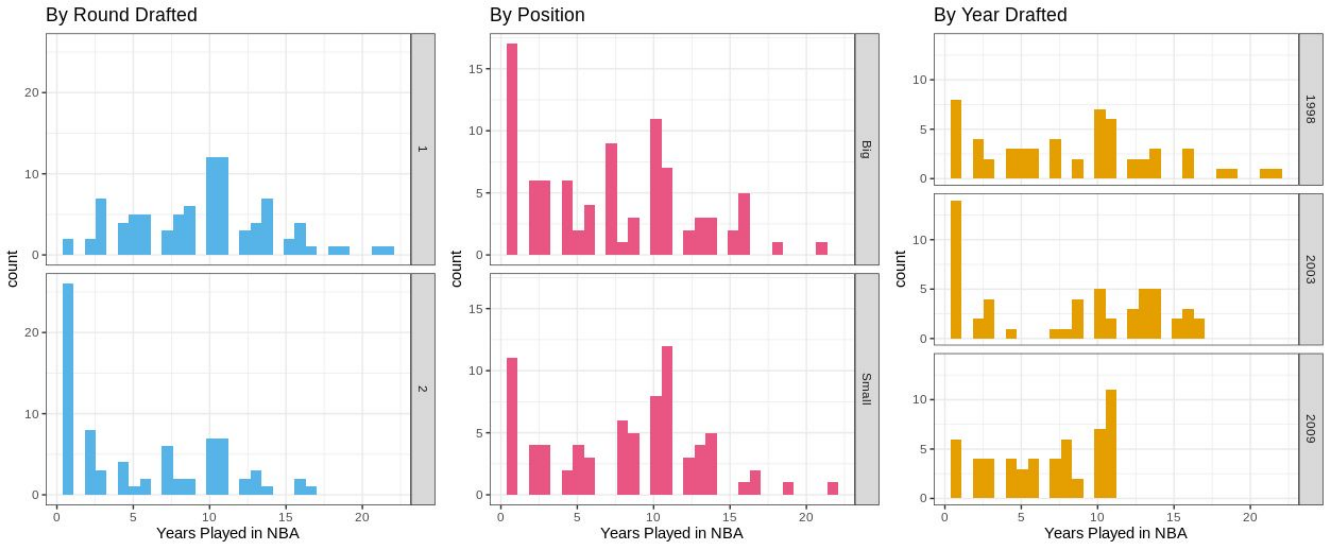
We are interested in how long does NBA careers last. What degrees does, does the round of the player, the position of the player and lastly draft year impact career length of the player impact on career length?

Moreover, we are interested in whether there is a relationship between NBA careers duration and these covariates, then we predict the NBA careers given their covariates combination.

Data Exploration

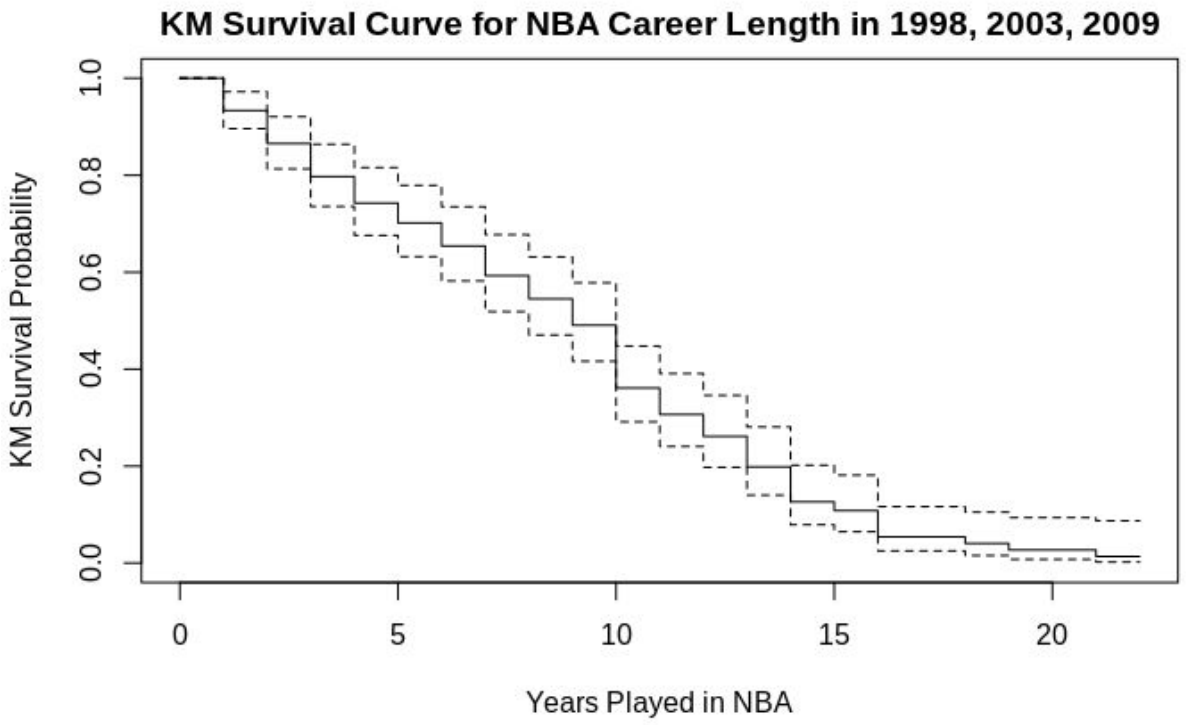
By using the data summary we find that out of that we observe 56 players drafted in 1998, 54 players drafted in 2003, and 55 players drafted in 2009. We also see that 88 of the 90 players drafted in the first round are observed while 77 of the 90 players drafted in the second round are observed. 89 “Big” players are observed and 76 “Small” players are observed. The median and mean are similar, 8 and 7.7 respectively.

```
  Year     Rd        Pos          Yrs
1998:56   1:88   Big   :89   Min.    : 1.000
2003:54   2:77   Small:76   1st Qu.: 3.000
2009:55                     Median : 8.000
                            Mean    : 7.727
                            3rd Qu.:11.000
                            Max.    :22.000
```

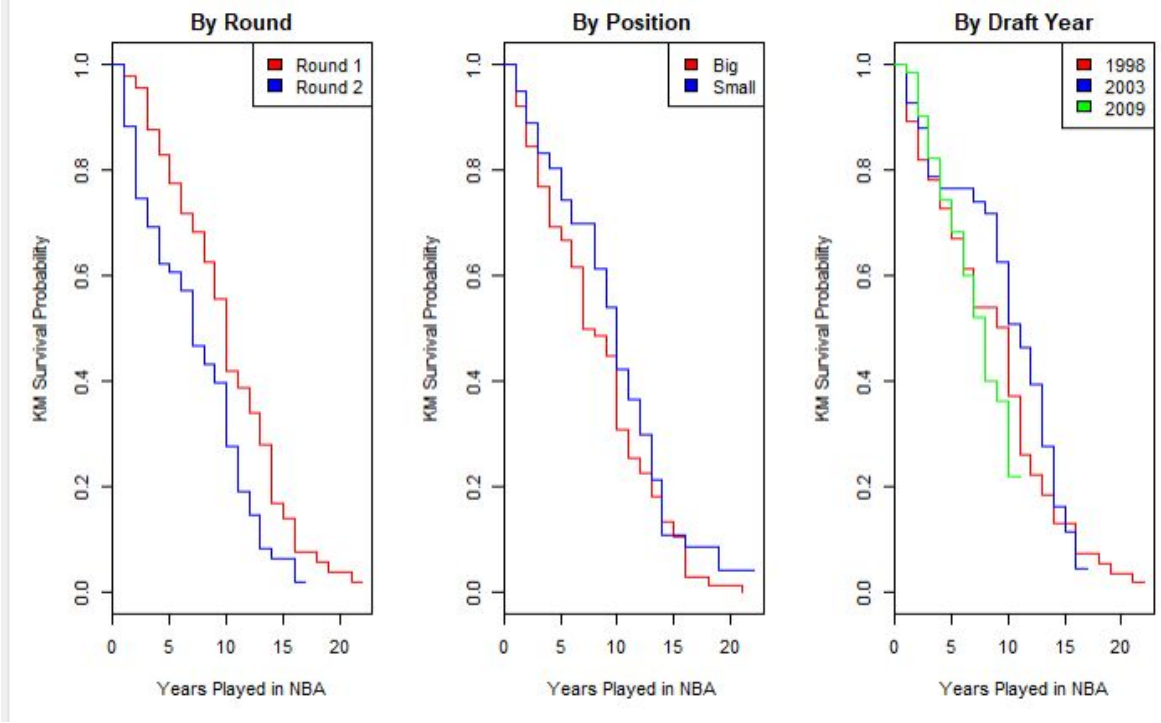


Kaplan-Meier estimation curves

First, we consider the Kaplan Meier curve with a 95% confidence interval for our entire population. We notice a fairly linear trend downwards, with the vast majority of careers lasting less than twenty years.



We plot the Kaplan-meier survival curves to visually analyze the effects of each covariate *Round, Position,* and *Year* on career length. By looking at the plots, we can conclude that there is a difference in the length of a player's career by the round they are drafted in. There is a slight difference in the career length between "Big" and "Small" players. Lastly we can see that players drafted in 2003 tend to have slightly longer careers than those drafted in 2009.



Log rank test

When we conduct a log rank test on each of the covariates, we see that *Round* is the only covariate which has a p-value less than 0.05 indicating that it has a significant effect on career length. *Position* and *Year* have p-value of 0.1 which is close to 0.05 however, we interpret this as covariates that do not have a significant effect on career length.

```
Call:
survdiff(formula = Surv(Yrs, event) ~ Rd, data = nba_draft)

      N Observed Expected (O-E)^2/E (O-E)^2/V
Rd=1 88       77       92      2.45      9.31
Rd=2 77       57       42      5.37      9.31

 Chisq= 9.3  on 1 degrees of freedom, p= 0.002
Call:
survdiff(formula = Surv(Yrs, event) ~ Pos, data = nba_draft)

           N Observed Expected (O-E)^2/E (O-E)^2/V
Pos=Big   89       76     67.9     0.968      2.29
Pos=Small 76       58     66.1     0.994      2.29

 Chisq= 2.3  on 1 degrees of freedom, p= 0.1
Call:
survdiff(formula = Surv(Yrs, event) ~ Year, data = nba_draft)

            N Observed Expected (O-E)^2/E (O-E)^2/V
Year=1998  56       53     52.7   0.00139   0.00282
Year=2003  54       42     50.0   1.28540   2.52821
Year=2009  55       39     31.3   1.92075   3.10759

 Chisq= 4  on 2 degrees of freedom, p= 0.1
```

Model Building

We begin by fitting a Cox Proportional Hazards model on all three covariates: Position, Draft Round, and Draft year. The output below shows that the model is significant, with an overall Likelihood Ratio Test p-value of 0.00361. Wald tests indicate that the only significant covariate in our model is the draft round, with the hazard ratio between players drafted in Round Two relative to Round One being 1.733, indicating that players drafted in the second round of the draft have a ~73% increase in hazard.

```
Call:
coxph(formula = Surv(Yrs, event) ~ Pos + Year + Rd, data = nba_draft)

            coef exp(coef) se(coef)      z      p
PosSmall -0.2805    0.7554   0.1783 -1.573 0.1157
Year2003 -0.1645    0.8483   0.2126 -0.774 0.4391
Year2009  0.3453    1.4124   0.2279  1.515 0.1297
Rd2       0.5501    1.7334   0.1788  3.076 0.0021

Likelihood ratio test=15.6  on 4 df, p=0.00361
n= 165, number of events= 134
```

Given our relatively small sample size (165), the normality assumption of the Wald Test might not be satisfied. We use the anova function in R to perform additional Likelihood Ratio Tests, comparing the models by adding covariates iteratively. We find adding Year and Position to our model do not add much improvement over the model which only includes Draft Round.

```
Analysis of Deviance Table
 Cox model: response is Surv(Yrs, event)
Terms added sequentially (first to last)

      loglik  Chisq Df Pr(>|Chi|)
NULL -552.30
Year -550.39 3.8110  2    0.14875
Pos  -549.08 2.6307  1    0.10482
Rd   -544.50 9.1555  1    0.00248 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

However, both Forwards and Backwards Stepwise indicate that it is still beneficial to control for Draft Year and Position, at least when using AIC as a model performance metric. Therefore, we choose to keep all three covariates in the model.

Forwards

```
Start:  AIC=1104.59
Surv(Yrs, event) ~ 1

       Df    AIC
+ Rd    1 1097.9
+ Pos   1 1104.3
<none>    1104.6
+ Year  2 1104.8

Step:  AIC=1097.88
Surv(Yrs, event) ~ Rd

       Df    AIC
+ Year  2 1097.5
+ Pos   1 1097.6
<none>    1097.9

Step:  AIC=1097.49
Surv(Yrs, event) ~ Rd + Year

       Df    AIC
+ Pos   1 1097.0
<none>    1097.5

Step:  AIC=1097
Surv(Yrs, event) ~ Rd + Year + Pos

Call:
coxph(formula = Surv(Yrs, event) ~ Rd + Year + Pos, data = nba_draft)

           coef exp(coef) se(coef)     z      p
Rd2       0.550     1.733    0.179  3.08 0.0021
Year2003 -0.165     0.848    0.213 -0.77 0.4391
Year2009  0.345     1.412    0.228  1.52 0.1297
PosSmall -0.280     0.755    0.178 -1.57 0.1157

Likelihood ratio test=15.6  on 4 df, p=0.004
n= 165, number of events= 134
```

Backwards

```
Start:  AIC=1097
Surv(Yrs, event) ~ Rd + Pos + Year

       Df    AIC
<none>    1097.0
- Pos   1 1097.5
- Year  2 1097.6
- Rd    1 1104.2
Call:
coxph(formula = Surv(Yrs, event) ~ Rd + Pos + Year, data = nba_draft)

           coef exp(coef) se(coef)     z      p
Rd2       0.550     1.733    0.179  3.08 0.0021
PosSmall -0.280     0.755    0.178 -1.57 0.1157
Year2003 -0.165     0.848    0.213 -0.77 0.4391
Year2009  0.345     1.412    0.228  1.52 0.1297

Likelihood ratio test=15.6  on 4 df, p=0.004
n= 165, number of events= 134
```

To see if interactions between covariates are significant, we build a model including the interaction terms. We find that again, round is the only significant covariate via Wald tests.

```
Call:
coxph(formula = Surv(Yrs, event) ~ Pos * Year * Rd, data = nba_draft)

                           coef exp(coef) se(coef)      z      p
PosSmall                -0.1005    0.9044   0.3838 -0.262 0.7934
Year2003                 0.1068    1.1128   0.3728  0.287 0.7744
Year2009                 0.7098    2.0336   0.4084  1.738 0.0822
Rd2                      0.8604    2.3642   0.3570  2.411 0.0159
PosSmall:Year2003        0.1362    1.1459   0.5397  0.252 0.8007
PosSmall:Year2009       -0.4259    0.6532   0.5820 -0.732 0.4643
PosSmall:Rd2             0.2820    1.3258   0.5773  0.488 0.6252
Year2003:Rd2            -0.3618    0.6964   0.5698 -0.635 0.5255
Year2009:Rd2            -0.2495    0.7792   0.5647 -0.442 0.6586
PosSmall:Year2003:Rd2   -1.1555    0.3149   0.8801 -1.313 0.1892
PosSmall:Year2009:Rd2   -0.2999    0.7409   0.8663 -0.346 0.7292

Likelihood ratio test=23  on 11 df, p=0.01765
n= 165, number of events= 134
```

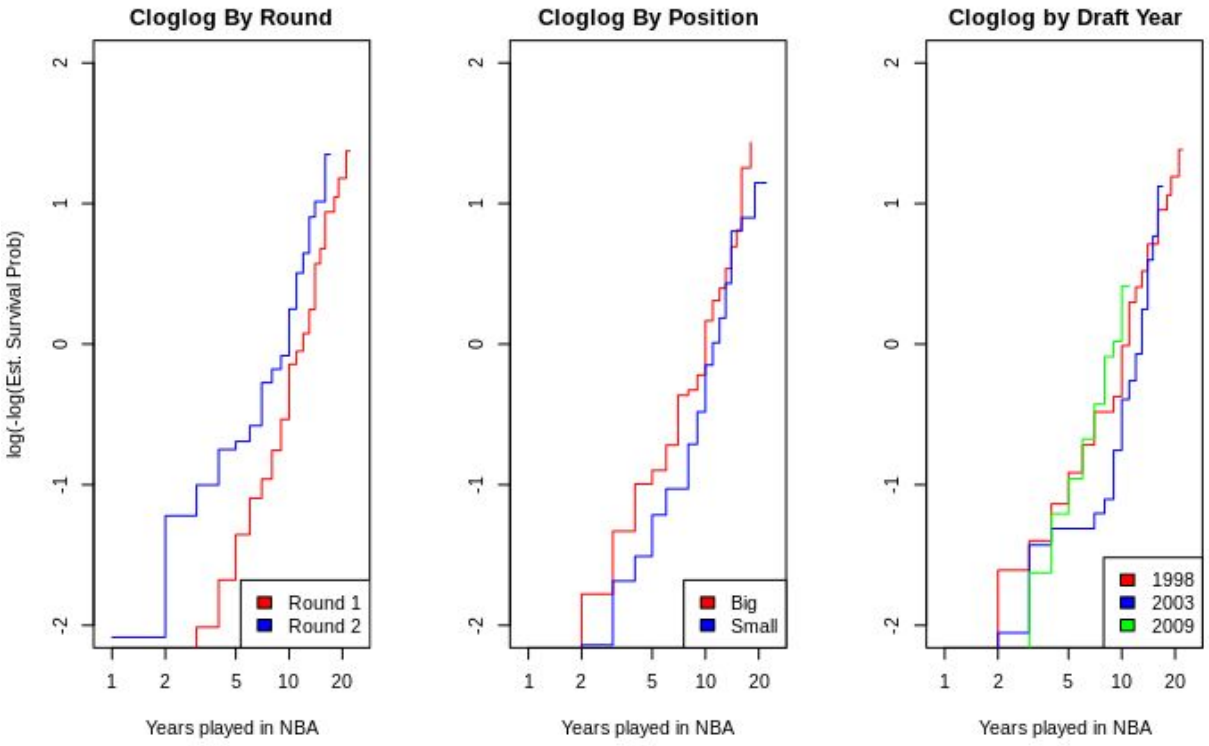
To see the interaction terms improve the model, we run a Likelihood ratio test between this interaction model with the base model above containing only draft year, position, and round. The result, shown below, indicate that the interaction model does not perform significantly better than the first order model, with a p value of 0.3878

```
Analysis of Deviance Table
 Cox model: response is  Surv(Yrs, event)
 Model 1: ~ Pos + Year + Rd
 Model 2: ~ Pos * Year * Rd
  loglik  Chisq Df P(>|Chi|)
1 -544.5
2 -540.8 7.4066  7    0.3878
```

For model interpretability purposes, and because none of the individual interaction terms seem particularly significant using the Wald test, we choose not to include any interaction terms. Note also that testing each of the interaction terms individually would exacerbate multiple hypothesis testing problems, and hence lead to misleading p-values. Multiple testing corrections are outside of the scope of this class, so we choose to retain the first order model only. Then our proposal model is a Cox Proportional Hazards model including the covariates Draft Year, Position, and Draft Round.

Model Checking

Before we decide on our final model, we validate the proportional hazards assumption for each of the covariates. The complementary log-log plots are shown below. For Draft Round, we see that the plots are further apart at the beginning than at the end, indicating that proportional hazards might not be satisfied. Overall, however, the lines look fairly parallel, so we would weakly conclude that the proportional hazards assumption is satisfied. Comparing Positions, we see that the lines cross at around 10 years. However, the violation is minimal, and the lines are fairly close throughout the time period studied, so we again weakly conclude that proportional hazards are satisfied. For Draft Year, however, we see lots of crossing, particularly between 2009 and the other years. Using this visual inspection, it seems as if the proportional hazards assumption mightn ot hold for this covariate.



To get a more empirical evaluation of the Proportional Hazards assumption, we run Schoenfeld Residual tests on the model, which effectively tests the null hypothesis that proportional hazards is satisfied. Looking at the p values for the individual covariates as well as the global model test, we find that all of the p values are large, with the smallest being 0.206. Then this test concludes that the proportional hazards assumption holds for all of our covariates.

	rho	chisq	p
PosSmall	0.0139	0.0268	0.870
Year2003	0.1080	1.5960	0.206
Year2009	0.1013	1.5009	0.221
Rd2	-0.0823	0.8894	0.346
GLOBAL	NA	3.2435	0.518

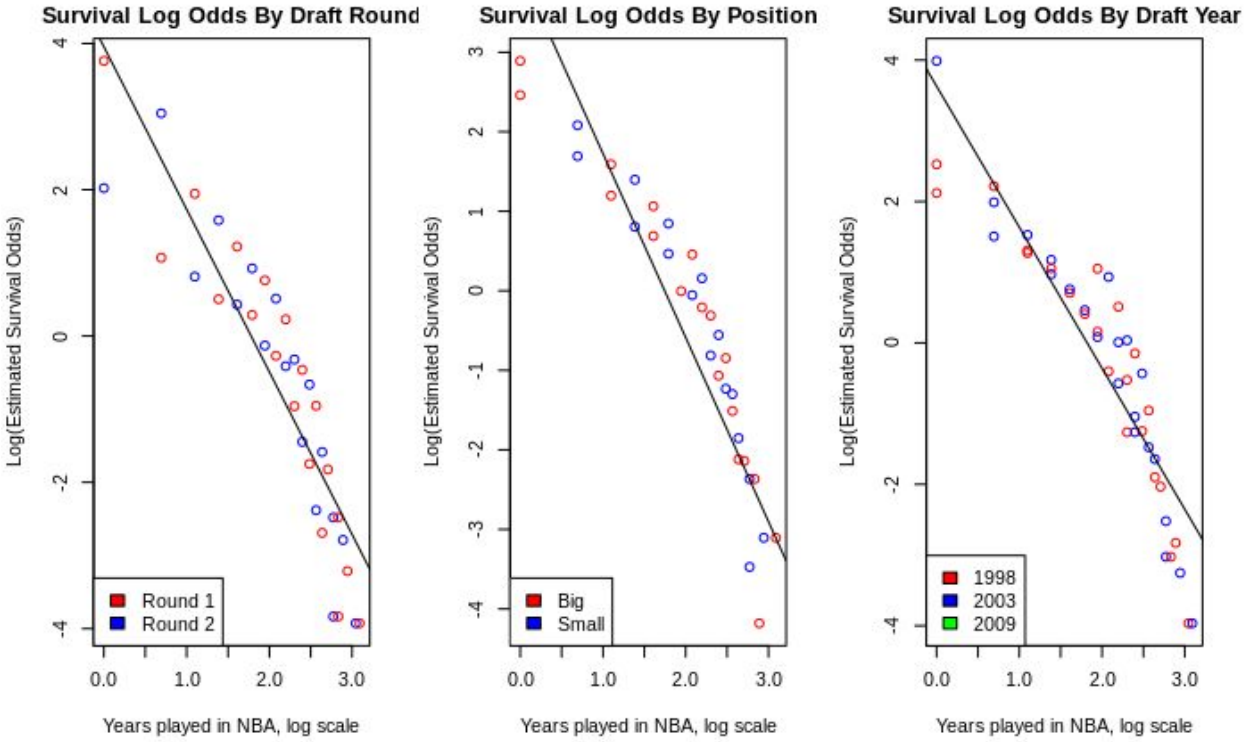
We note that the results from our visual inspection of the complementary log-log plot differs from the results of the test above in our evaluation of Draft Year. Because the Schoenfeld Residuals is a more empirical metric, we prefer these results, and claim that the proportional hazards assumption holds for all of our covariates.

Therefore, the Cox Proportional Hazards model is valid, and we conclude that our final model is the Cox Proportional Hazards model looking at career length using the covariates Draft Round, Draft Year, and Player Position.

Model Extension 1: Parametric Survival Models

From the response variable distribution charts shown in the Data Exploration section, we notice that career length tends to have heavy weight on early values (i.e. many players have short careers). Then out of the distributions we have studied in class, this suggests that our data would best fit the log-logistic distribution. To test this assumption, we use the fact that for log-logistic distributed data, the log survival odds will be approximately linear on the log-time scale. The plots below show this relationship across our three covariates, with the black line indicating the univariate linear regression fit for that variable. Comparing the points in the plot to the line, we find that the data might have some non-linear pattern, although the linear fit might be a sufficient approximation. One way to check this assumption is by looking at the distribution of the residuals of the linear fit. If they are approximately normally distributed, then the residuals are roughly symmetric and the linear fit might be a good enough approximation. Then to test for normality, we ran Shapiro-Wilkes tests on each of the three univariate linear fits, reporting p values of 0.103, 1.028e-7, and 0.1818 for Draft Round, Position, and Draft Year respectively.

Therefore, we conclude that the only covariate that strong violations normality is Position, and hence we exclude it from the log-logistic model.



Then our log-logistic model is fit using only the covariates Draft Round and Draft Year. The results of this fit are shown below. We find that this model is significant, with a p value of 0.005 for the model overall. Furthermore, we find again (via Wald tests) that Round is significant, controlling for Draft Year. The results below show that we have an Acceleration Factor of exp(-0.4456) = 0.64 for comparing players drafted in Round 1 vs Round 2. This means that, using this log-logistic model, that the probability that a player drafted in round 1 survives to time, say, 10 years, is approximately equal to the probability that a player drafted in round 2 survives to 10*.64 = 6.4 years, even after controlling for Draft Year.

```
Call:
survreg(formula = Surv(Yrs, event) ~ Year + Rd, data = nba_draft,
    dist = "loglogistic")
             Value Std. Error      z      p
(Intercept)  2.1699     0.1227  17.68 <2e-16
Year2003     0.1566     0.1633   0.96 0.3374
Year2009    -0.0280     0.1561  -0.18 0.8579
Rd2         -0.4456     0.1380  -3.23 0.0012
Log(scale)  -0.7847     0.0719 -10.91 <2e-16

Scale= 0.456

Log logistic distribution
Loglik(model)= -421.8   Loglik(intercept only)= -428.2
        Chisq= 12.85 on 3 degrees of freedom, p= 0.005
Number of Newton-Raphson Iterations: 4
n= 165
```

Model Extension 2: Leave One Out Linear Regression Models

The real novelty of Survival Analysis when compared to other modeling techniques is the ability to incorporate censored data into the model. We also noticed that the primary focus of the course has not been on predictive accuracy, but rather on model interpretation. These observations prompted us to question whether we could adapt a simple, interpretable model like linear regression to handle censored data.

Our idea for incorporating censored data into linear regression is as follows:

- Pick out all observations that were censored at a given time, say censored at time t.
- For the rest of the observations, compute our new response variable, “Career time remaining after t”, given by min{0, career length - t}.
- Fit a linear regression on the data using the same covariates as our final Cox Proportional Hazards model, excluding the censored observations
- Record model fit
- Predict on the censored observations, measuring RMSE in sample and out of sample.
- Repeat for all observations censored at a different time until all censored observations have been accounted for (we make 4 models because we only have censoring at times 1, 11, 17,and 22, as noted in the data exploration section)
- For each covariate, compute a weighted average of the coefficient by (1/RMSE) and (1/ coefficient p-value) (we want to put higher weight on models with low RMSE and low p-values, hence the weight being given by 1/RMSE and 1/p-value)
- Report these coefficients along with a Benjamini-Hochberg corrected p-values, controlling for the number of models we created. (note that these p-values don’t actually

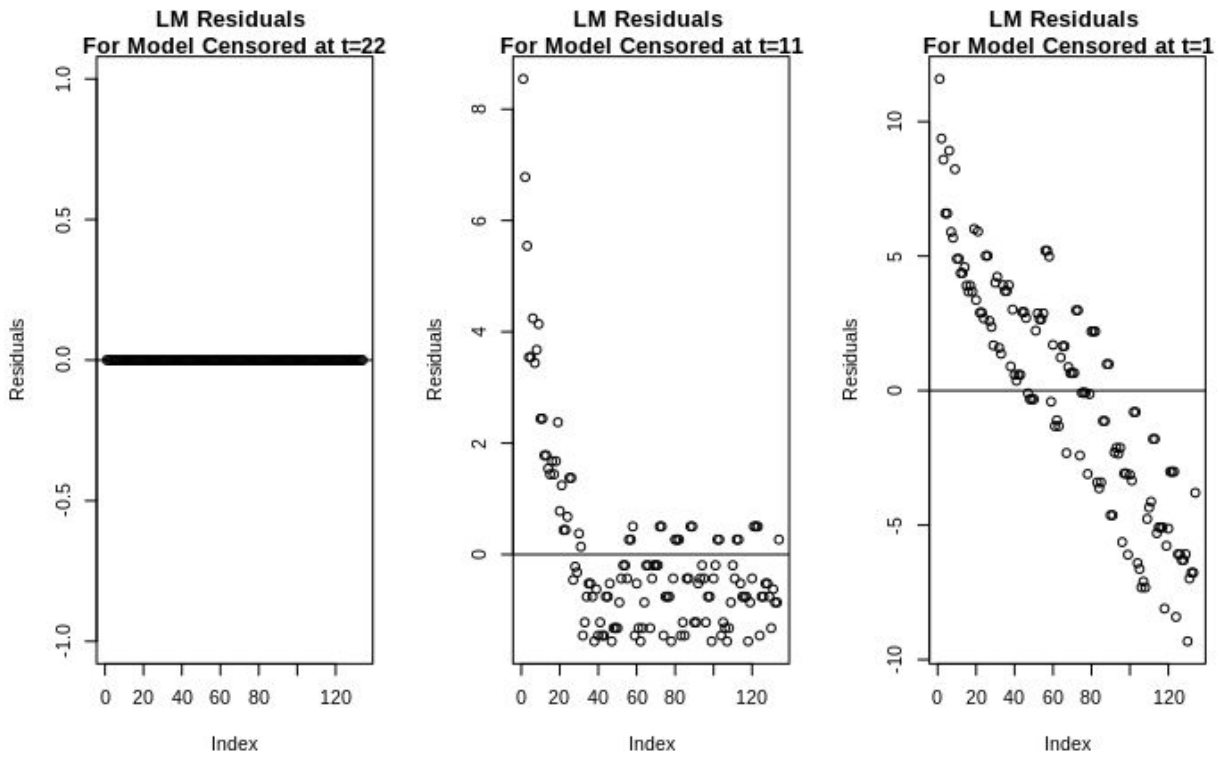
correspond to the coefficient value, but we think it is still useful as a rough measure of confidence.)

The results of this process are shown below. Using this method, it looks like Round is significant (confirming the results of our cox proportional hazards model), but we also notice that this model marks 2009 as potentially different than 1998. Unlike Cox Proportional Hazards, this model is not on a log scale, and the coefficients can be roughly interpreted as the expected difference in career length given the value of the covariate. For instance, the Round 2 coefficient of -1.57 indicates that we expect players drafted in round 2 to have careers that are 1.57 years shorter than their round 1 counterparts, controlling for Draft Year and Position. With the exception of Position (which has a very small coefficient value and a high p value), these coefficients are all in the same direction as our Cox Proportional Hazards model, which indicates that the two models are picking up on similar signals.

	term	avg_coef	avg_bh_p
	<chr>	<dbl>	<dbl>
1	Rd2	-1.57	0.0182
2	PosSmall	-0.0233	0.643
3	Year2003	0.211	0.547
4	Year2009	-1.66	0.0186

This was just a preliminary test to see what we could come up with using standard methods, and these results tend to match up with the models we have done in class.

To improve this model, we would recommend looking into using something like a generalized Poisson regression rather than linear regression, to deal with the discrete, non-negative nature of the data. We found that the fit of our model changes significantly based on the censorship time we use to create the response variable. For instance, Vince Carter (the only player from the 1998 draft class still active, and hence censored), we have that the response variable, min{0, 22 - time}, is always zero in our dataset because he has the longest career of anyone in our data. Then the linear model returns coefficients that are identically zero. While the resulting prediction is reasonable (that Carter will retire this year), the residuals plots below show that the model fit is very different than the models using different censoring times. We partially control for this variation by weighing the coefficients by p-value and RMSE (both indications of model fit), but more robust methods should be explored.



We would also look into finding a more clever way to combine the results of the model to get an interpretation of the coefficients. The weighted averages are a crude measure of doing the interpretation, especially because of the difficulty in finding an overall p value.

For being a first attempt at using methods from previous classes to censored data, we conclude that this extension is a success. This is partially due to the fact that this model confirms our results from the Cox Proportional Hazards model, but mainly because our work illustrates the difficulty of including censored data into the model without using the Survival Analysis methods taught in class.

Conclusion/Discussion

Overall, our project found our research questions on the variables that affect NBA career duration. We plotted the Kaplan-Meier curve to check if any of the covariates affects NBA career over time. With the curves indicating that only the round the players are drafted covariate had a major effect on length of a player's career. Another result is a slight difference in the career length between "Big" and "Small" players, but the p values were not significant.

To double check our result we then performed a log-rank test to confirm the significance of each covariate. With the result, we confirm that the significant covariate satisfies the model's assumption through C-log-log plot. Wald tests indicate that the only significant covariate in our model is the draft round, with the hazard ratio between players different round is1.733, Showing that players drafted in the second round of the draft have a ~73% increase in hazard. Our Forwards and Backwards Stepwise illustrate that it is still beneficial to control for Draft Year and Position, at least when using AIC as a model performance metric.

Our final model is the Cox Proportional Hazards model looking at career length using the covariates Draft Round, Draft Year, and Player Position. From the resulting cox-PH model, we again came up with the result that the round the players are draft covariate had a major impact on the NBA career duration while all of the other covariate p values were not significant with our population of the data set. Our conclusion comes from our 180 data sets maybe if our data set were much larger the other two covariate can be significant. When we look back at our research questions, we can conclude that NBA players' careers are independent of position they play and year of draft. NBA career has direct correlation with the player's ability on the game field which is a good sign for NBA players because they can focus on enhancing their ability of the game and don't have to worry about other variables. A player's skill is a good predictor of a player's career.