

8-5-2012

An Analysis of the Career Length of Professional Basketball Players

Kwame D. Fynn

Macalester College, kwamefynn@gmail.com

Morgan Sonnenschein

sonnenschein.morgan@gmail.com

Follow this and additional works at: <http://digitalcommons.macalester.edu/macreview>

 Part of the [Applied Mathematics Commons](#), and the [Survival Analysis Commons](#)

Recommended Citation

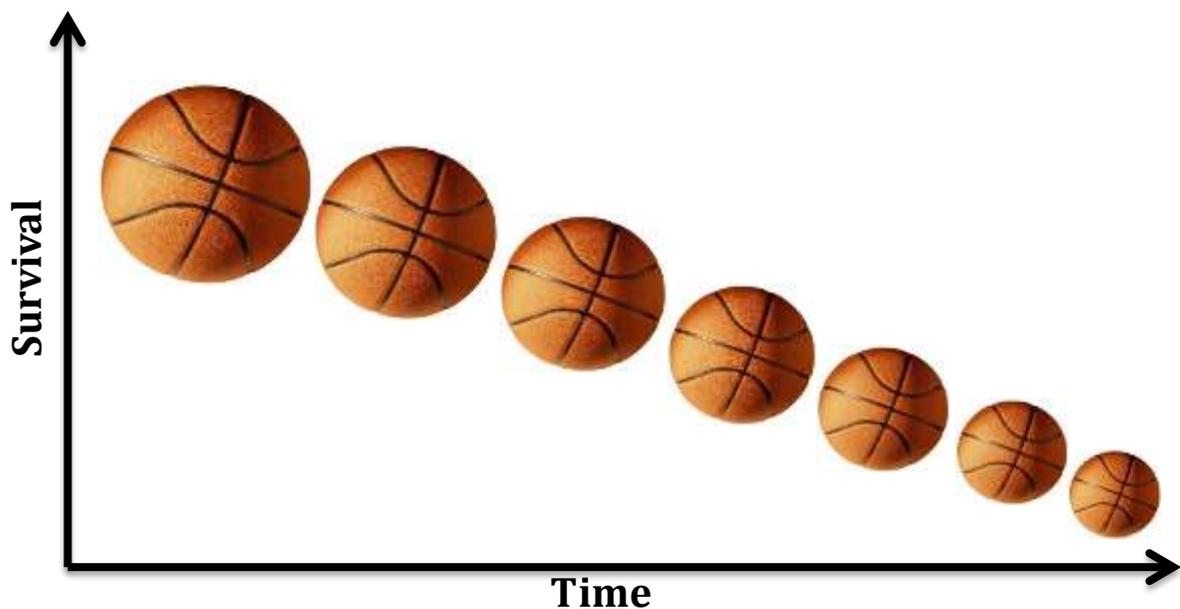
Fynn, Kwame D. and Sonnenschein, Morgan (2012) "An Analysis of the Career Length of Professional Basketball Players," *The Macalester Review*: Vol. 2: Iss. 2, Article 3.

Available at: <http://digitalcommons.macalester.edu/macreview/vol2/iss2/3>

This Article is brought to you for free and open access by the DigitalCommons@Macalester College at DigitalCommons@Macalester College. It has been accepted for inclusion in The Macalester Review by an authorized administrator of DigitalCommons@Macalester College. For more information, please contact scholarpub@macalester.edu.

An Analysis of the Career Length of Professional Basketball Players

Kwame Fynn & Morgan Sonnenschein
December 16, 2011



An interesting problem in professional basketball is predicting how long a player remains in the NBA League. Previous research on this problem has focused on factors such as race, performance in games, and size. We propose to analyze career duration in the NBA based on awards won, position played and biological variables such as height. Using Accelerated Failure Time models, Cox Proportional Hazards models and Kaplan-Meier analyses, we determine that both height and number of awards won lengthen career duration; however, only certain player positions significantly affect career length of a player.

Section 1: Introduction

Professional basketball in the United States is one of the largest sports industries nationwide. The National Basketball Association (NBA) consists of 30 teams and is arguably the most prominent professional basketball league in the world. Research of survival analysis in the NBA has focused on specific topics such as how long a player survives within a particular franchise (Staw and Hoang, 1995), survival within each game (i.e. minutes played per game, *ibid*), and career length (Groothuis and Hill, 2004). An immense number of factors can affect the overall length of a player's career; however few survival analyses of career length in the NBA exist. In this paper, we propose to analyze the length of a professional basketball player's career and expand upon existing research on that topic.

Groothuis and Hill seek to determine whether or not players of different races (white or black) have larger hazards (i.e. are at a greater risk of shortening their careers). The assessment used performance-based data stratified by race as possible explanatory variables (e.g. points per minute, free throw percentage). Much of the paper has to do with the mathematics of their equation for the hazard used to analyze their data, which is specifically tailored to help answer their question. Much of this math is beyond our grasp and will not be discussed here. By the end of the paper, the authors conclude that race does not play a significant role in NBA career length. In addition to their main conclusion, an intermediate observation relevant to our study is also made: larger and better performing players tend to stay in the NBA longer.

Our proposed analysis differs from that of Groothuis and Hill in three major ways. First, the data used in this paper includes all professional basketball players across all years available (1947-2011). While this data set does not necessarily focus on current or more recent players, it provides an opportunity to examine career length of retired players and eliminates the possible hassle of left truncation. Secondly, we use the number of individual awards as our measure of performance rather than individual player statistics such as field goal percentage, points per game, and number of games played. We also include the positions of players as an explanatory variable.

The goal our study is to determine performance and biological variables that may affect the career length of a player. Analysis of these variables may reveal a significant difference in the careers of players who have these traits and those that do not.

Our paper proceeds as follows: in Section 2, we discuss the data obtained and the methods used to examine it. Section 3 presents the results we obtain from our data analysis. In Section 4, we discuss our results relative to previous literature and its practicality. We conclude in Section 5 by providing a summary, the limitations and possible extensions of the paper.

Section 2: Data and Methods

Data

The data we obtained is of all basketball players (ABA and NBA Leagues) from 1947 to 2011 from www.basketball-reference.com. Our data consists mostly of biological data of all the players. As previously mentioned, to obtain performance data, we included the number of awards a player has won individually. Due to the immense size and variability of our data, both parametric and non-parametric regression models were used to determine which model best fits our data set. This process enabled us to generate the necessary models to analyze our topic of interest. Additionally, by including players who have not yet retired, our data is right-censored. Table 1 below shows a summary of all our variables with abbreviations in parenthesis.

Variable Name	Explanation	Units
Name	The name of each player	-
From	Commencement of NBA Career	-
To	End of NBA Career or 2011 if still playing	-
Duration	How long a player has played for	Years
Status*	If the player's career has ended or not	1=Retired, 0=Still Playing
Position*	Assigned Team Position	G=Guard, F=Forward, C=Center
DayOfBirth (DOB)	Day the player was born	-
MonthOfBirth (MOB)	Month the player was born	-
YearOfBirth (YOB)	Year the player was born	-
AgeAlive	Age of player if alive	Years
Feet	Height of player (Feet Only)	Feet
Inches	Height of player (Inches above number of feet)	Inches
HeightInches	Total Height of player	Inches
Weight	Weight of player	Pounds
WeightC	Categorical weight of player	1=0-150, 2=151-200, 3=201-250, 4=251-300, 5=301-350
Awards	Total number of awards won	-
AwardsYN*	If a player has won at least one award	0=No award, 1=at least one award
College	Name of player's college/university	-
HallOfFame (HOF)	If the player is enrolled in the NBA Hall of Fame	1=Yes, 0=No

*Indicates assumptions have been made or needs further explanation

Table 1: A summary of all variables included in our data set, their respective explanation and units

Assumptions/Further Explanation:

Status: Based on the ‘To’ variable, we assume that all players who have played in 2011 (even though some may have retired) are currently in the league¹. This assumption implies players like Shaquille O’Neal who retired in 2011 is a current player.

Position: Some players are assigned more than one position in a team; however, the primary position of a player could not be obtained. For example, F-C and G-F.

AwardsYN: These awards include (a) Most Valuable Player, (b) Defensive Player of the Year, (c) Sixth Man of the Year, (d) Most Improved Player, (e) Finals Most Valuable Player, (f) All-Star Game Most Valuable Player, (g) Comeback Player of the Year, and (h) Rookie of the Year. (All individual awards were used to account for both the improvement and talent of a player).

Methods

We begin our analysis with a comparison of parametric (normal, weibull, exponential and log-normal) distributions to the empirical cumulative distribution function to determine the fitness of a parametric model to our data set. From previous research, biological variables that have been strong

¹ We make this assumption due to unavailable data for all players who retired after the 2010-2011 season.

predictors of career duration include the size of a player. We expand our analysis by fitting non-parametric models (Kaplan-Meier curves) for both our performance and biological variables. We further use Mantel-Cox log-rank tests and observe the stochastic ordering of curves to compare survival distributions for variables that may have a relatively strong difference in their Kaplan-Meier Curves. For all Mantel-Cox tests, we assume a significant level of 5%, Null Hypothesis (H_0): No difference in survival curves and Alternative Hypothesis (H_A): Difference in the survival curves.

With variables we found to be significantly related to the length of a professional basketball player's career, we apply multivariate parametric regressions using accelerated failure time (AFT) models. We generate an Akaike Information Criterion to determine the quality of fit for each of the models created. Cox proportional hazards are also used as another way to examine career duration. This would test whether or not players with different traits are systematically lasting longer in professional basketball relative to those who do not have these traits. We test the adequacy of the proportional hazards assumption using Schoenfeld residuals and a formal test for proportional hazards.

Finally, we look to see whether or not our models manifest collinear variables. A covariate related to another in the same model might make the model significant but not the coefficients within the model.

Section 3: Results

From our analysis, we realize that the Weibull, lognormal and exponential distribution fit our perceived failure time (Duration) relatively better than the normal distribution as made evident in Figure 1 below. Moving on, we used non-parametric models in addition to the parametric models that provide a good fit to our data set.

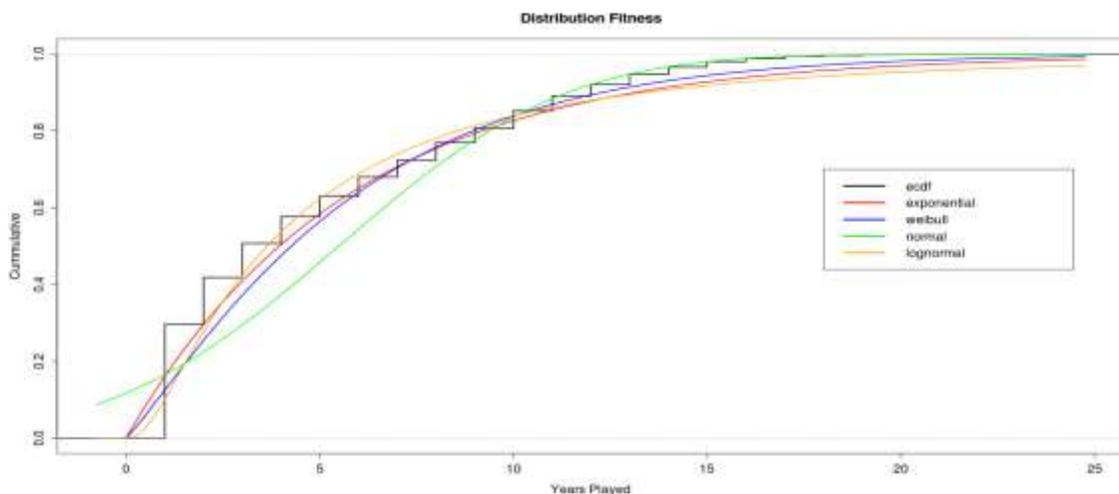


Figure 1: A graph showing the empirical cumulative distribution function of the career length of professional basketball players and estimates using the Weibull, Exponential, Normal and Log-Normal distributions'.

We observe major overlaps of the confidence intervals of the Kaplan-Meier curve (interval overlaps signify similarities) for position and categorical weight of a player. However, we do not completely rule out these variables as an insignificant measure of differences in career length of a player. The proximity of the Kaplan-Meier curves may be due to minor variations in the different categories in a given variable. From the stochastically ordered Kaplan-Meiers in Figure 2, we realize that at every given time point the career length of a player regarding their height and performance is significantly different. Seven footers tend to have the longest career with a median duration of 6 years whereas 6 footers and 5 footers have medians of 4 years and 3 years respectively.

Although Hall of Fame status is a strong measure of survival, we fail to use this variable to predict career longevity. Players retire before getting inducted into the Hall of Fame, and so this does not answer our question of interest. It rather offers a reverse causality to our analysis. (*Using Hall of Fame as a variable analyzes how long a career must be to get inducted into the Hall of Fame*). Players with at least one award have a significantly longer career than players with no awards. On average, players with at least one award have median career duration of 14 years whereas players with no awards have median career duration of 4 years. By comparing survival curves, we observe a close similarity between players who have won an award and players who have been inducted in the Hall of Fame.

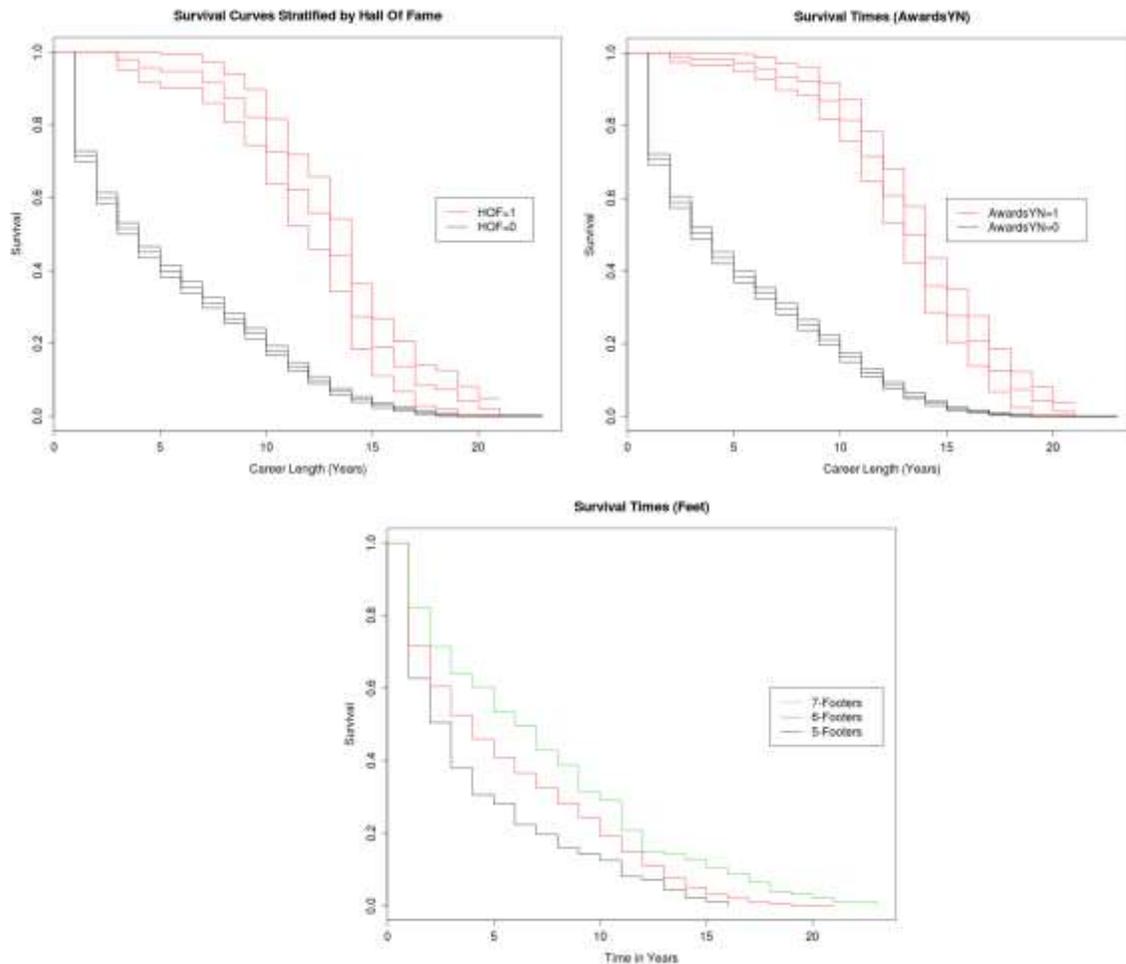


Figure 2: Kaplan-Meier curves stratified by Feet, AwardsYN and HOF.

Variable	Category	Median	Confidence Interval	Test Statistic	P-value
Feet	7	6	[5,8]	26.05	0.000002
	6	4	[4,4]		
	5	3	[2,3]		
AwardsYN	1	14	[13,14]	205.4	0
	0	4	[3,4]		

Table 2: Summary of Feet and Awards from a Kaplan-Meier and Log-Rank Test analysis.

An important concept we will be using to examine our models is the hazard ratio as an indicator of a player’s relative longevity. That is, a player who survives longer than another has a hazard ratio of less than one. Hazard ratios come into our analysis with the use of Cox proportional hazards.

We compare our accelerated failure time models using an Akaike Information Criterion (AIC). A lognormal model using HeightInches, Awards and Position has the lowest AIC of 18449.4. From Table 3, we rank our AIC from the different models in ascending order. With lognormal models occupying the top three ranks, we believe lognormal models offer the best fit for our data. From our model, all our coefficients are positive and significant. However, by checking for collinearity between variables, we observe a direct relationship between a player’s height and his position (Table 5). This allows us to eliminate the HeightInches variable since it can be accounted for using Position.

Model 1: *survreg(Surv(Duration, Status)~HeightInches + AwardsYN + Position, dist = "A")*

Model 2: *survreg(Surv(Duration, Status)~Feet + AwardsYN + Position, dist = "A")*

Model 3: *survreg(Surv(Duration, Status)~HeightInches + Awards + Position, dist = "A")*

Model 4: *survreg(Surv(Duration, Status)~Feet + Awards + Position, dist = "A")*

"A" represents a Log – Normal, Exponential or Weibull distribution

Model	Akaike Information Criterion
Log-Normal 3	18449.4
Log-Normal 2	18459.0
Log-Normal 4	18616.8
Weibull 1	18694.4
Weibull 2	18808.6

Table 3: The lowest 5 AIC for the different models we run.

Variable	Coefficient	P-value
Awards	0.312	4.71E-36
PositionF	0.205	5.85E-04
PositionF-C	0.721	4.70E-29
PositionF-G	0.886	4.46E-32
PositionG	0.738	3.27E-22
HeightInches	0.088	1.06E-45

Table 4: Summary of results from a Log-Normal distribution using Model 3

Position	HeightInches				
	0%	25%	50%	75%	100%
C	76	82	83	85	91
F-C	74	79	81	82	88
F	73	78	79	81	85
F-G	70	75	77	78	81
G	63	73	75	76	81

Table 5: Quantiles of Position by HeightInches

After re-running our model exclusive of the HeightInches variable, winning an award and playing two positions has a significantly positive effect on the length of a player’s career. On the contrary, in comparison to Centers, Forwards and Guards have a shorter survival time. A summary of our results is provided in Table 6 below.

Variable	Coefficient	P-value
Awards	0.328	2.58E-38
PositionF	-0.162	3.13E-03
PositionF-C	0.476	6.34E-14
PositionF-G	0.295	4.05E-06
PositionG	-0.042	4.39E-01

Table 6: Summary of Results from a lognormal model using Awards and Position

Additionally, from Table 7 (a) with a hazard ratio greater than one, relative to Centers, Forwards have a higher likelihood of exiting the league. On the contrary, players with two positions (Forward-Centers and Forward-Guards) and more awards have a lower likelihood of exiting the league as implied by hazard ratios of less than one.

We implement a formal test to assess if the hazard ratios’ between the position of a player and the number of award he has won is proportional over time. From Table 7 (b), although players with more awards and players who play two positions have a lower likelihood of exiting the league, their hazard ratio increases over time as shown by their positive statistically significant value of ρ (ρ measures the correlation of Schoenfeld residuals over time). However, relative to Centers, Forwards and Guards have a proportional hazard ratio as depicted by their insignificant p-values.

This implies that overall, although players with two positions and players with more awards are likely to survive longer in the league, with time, their chances of survival will decrease at an increasing rate.

Variable	Coefficient	$e^{Coefficient}$ (HR)	P-value
Awards	-0.4753	0.622	0.0E+00
PositionF	0.1666	1.181	5.6E-03
PositionF-C	-0.3691	0.691	7.5E-08
PositionF-G	-0.1821	0.833	8.3E-03
PositionG	0.0596	1.061	3.1E-01

(a)

Variable	ρ	P-value
Awards	0.1151	0.0E+00
PositionF	-0.0292	8.27E-02
PositionF-C	0.0701	3.35E-05
PositionF-G	0.0653	1.08E-04
PositionG	-0.0068	6.85E-01

(b)

Table 7: Cox-PH test and a formal test of PH assumption using Awards and Position

We check for the persistency of our hazard ratio using Scheonfeld residual plots (Figure 3). Our plots are consistent with our previous analysis. For variables which violate the assumption of a proportional hazard, we observe that the hazard ratio is previously overestimated then underestimated for Forward-Centers and Forward-Guards. It is unclear if there is any overestimation of the hazard ratio for Awards since it is evenly scattered around zero; however, as time increases, the hazard ratio becomes underestimated.

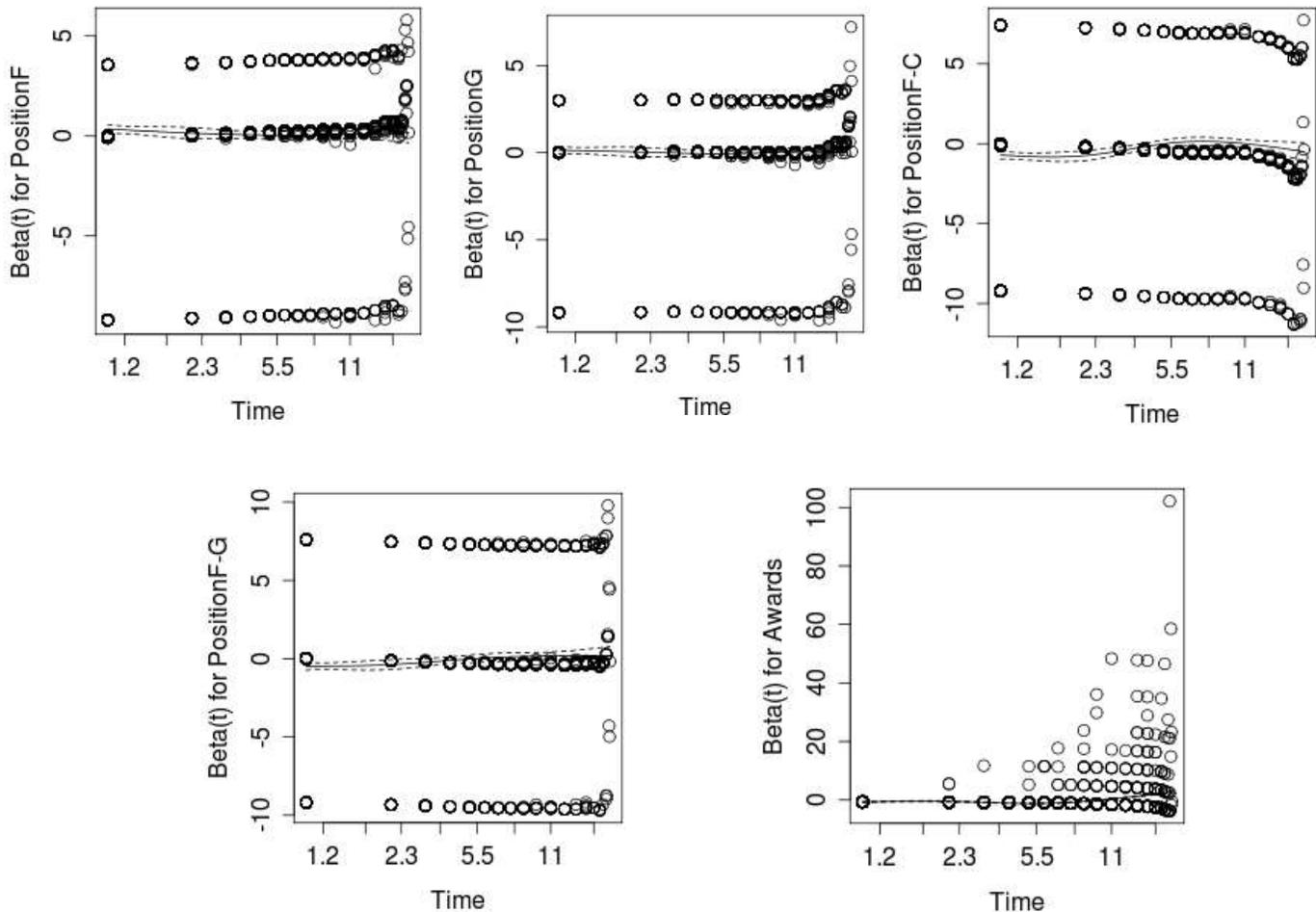


Figure 3: Schoenfeld residuals of the Cox-PH model in Table 7

Section 4: Discussion

We prefer to eliminate HeightInches rather than Position for practical purposes. As shown in Table 5, Position changes relative to HeightInches thus people of different heights are likely to play different positions. From Example 1 below, according to Table 4, a Guard or Forward that is the same height as a Center will survive longer; however, such an occurrence will rarely happen as evidenced by Table 5.

Example 1: Height: Player 1 = Player 2 = 7'0. Position: Player 1 = C, Player 2 = G or F.

From our results, we realize a player's height and number of awards won has a positive effect on his career duration. Taller players tend to survive longer as made evident from the career duration of Centers in the league. Additionally, players with two positions have a significantly longer survival time relative to all other players in the league. We believe this is attributed to the fact that players with two positions are more versatile and have a diverse skill set. Interestingly, we find a relationship between the number of awards won and the position of a player. Players with two positions have the highest award per player, with Forward-Centers having the largest value of 0.08. From our Cox PH models, there is no proportional hazard in any of the significant variables we include in our model. This implies that, relative to time, the hazard ratio is non-constant. We find it intriguing that players with two positions have a longer survival but that their chances of survival will decrease at an increasing rate. We believe this may be because Centers' have a different set of skills and in the long

run, improving their chances of survival in the league. Our results are consistent with previous research from Groothuis and Hill, 2004, who determine the size of a player to be a strong predictor of his career length. We have improved on this research by basing our results on the position of a player.

In our paper, we have been able to determine both biological and performance based variables to determine the longevity of a professional basketball player's career. We fail to use induction into the Hall of Fame as a proxy because a player's career has to end in order to get inducted into the Hall of Fame. However, a relationship exists between the number of awards won and Hall of Fame enrollment. There is a positive R^2 value of 0.445 between Hall of Fame status and number of awards won, but it is not statistically significant. Hall of Fame players on average have won 1.93 awards. From Figure 3 below, the side-by-side boxplots depict the relationship between Hall of Fame and Awards. Interestingly, with the exception of Mel Daniels, current and recently retired players, all players who have won over five awards are members of the Hall of Fame and played for at least ten years. As a result of the collinearity, we prefer not to include both variables in a model simultaneously.

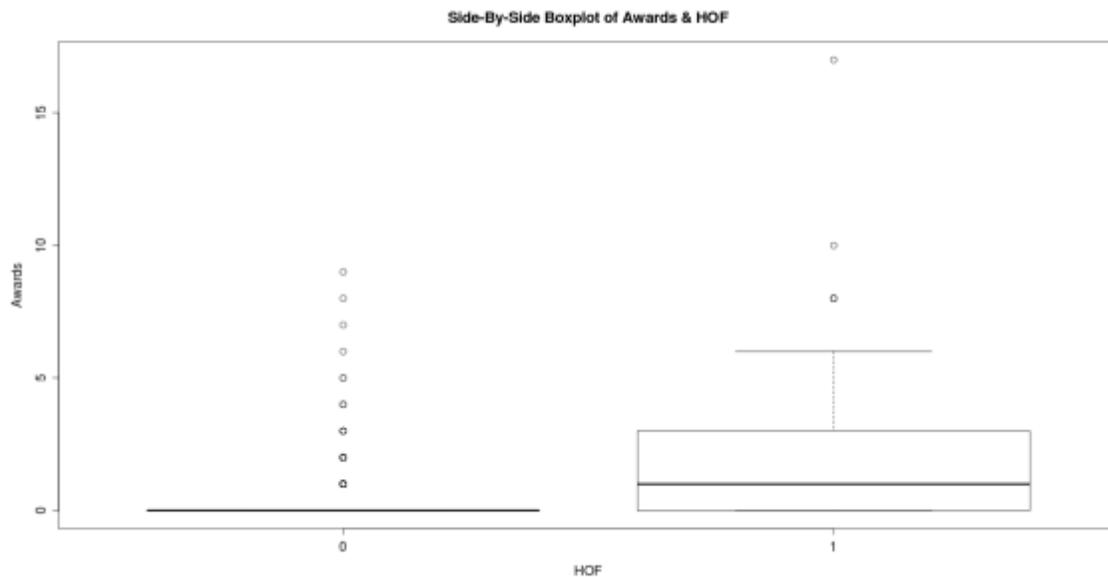


Figure 3: A Scatter plot of Awards by Hall of Fame Status

Conceptually, scarcity might explain why taller players survive longer in the league. In the NBA, it is advantageous to have a tall player play the role of a center or forward. Their height gives them the advantage on both offensive and defensive plays to account for blocking shots, rebounding and easily making shots close to the rim in the presence of an opponent. Since relatively fewer players in the league play Center and Forward-Center, regardless of their performance, such players are likely to be in high demand for a longer period of time in comparison to shorter players. Additionally, we believe winning an award increases the desirability of a player, so AwardYN and Awards are good measures of players' performances.

Section 5: Conclusion

In this paper, we analyze which variables affect the career duration of professional basketball players. The variables that are most explanatory of a player's career are the number of individual awards they have won and the position they play. To select our variables and models, we used

Kaplan-Meier curves, Mantel-Cox tests, and accelerated failure time models. We further analyzed if the relationship between the careers of players was constant over time using the Cox proportional hazard method. We came to our conclusion by taking all the above-mentioned techniques into account. We do not try to predict how long a player will survive in the NBA; rather, we are trying to predict what variables affect the career length of professional basketball players.

Our data set not being time-varied is a limitation since it does not account for the exact time a player wins an award. Although we do not have individual performance variables such as points per game to determine performance factors that affect career length with respect to time, we believe the number of awards won is a better measure. Individual performance variables vary with time and skill, and thus may predict different survival durations for different time frames. Additionally, due to the varying skill set of players, we might obtain biased results by using per game performance variables. This begs the question of what aspect of basketball-(points, assists blocks, steals, etc.) is most important to career duration.

One major difference we believe we could have used to improve our results would be to use data of players prior to being enrolled in a professional basketball league. We believe it will be interesting to relate the performance of a player in high school and/or college to his performance in the NBA to determine if there is a relationship between high school and/or college performance and career length in the league.

To improve on our analysis, more performance variables that vary with time can be used to determine the career length of a player. Additionally, it would be interesting to determine if the 'hot-hand' fallacy will be a significant explanatory variable. Related areas of interest would be extending our analysis to other sports such as baseball, football and soccer to find the different variables that may affect the duration of those players' careers.

In the presence of variables we determine to affect the Hall of Fame status of a player, we find an interesting reverse impact of a height on Hall of Fame status. From our results, HeightInches is positively related to career length and members of the Hall of Fame usually have a long career. By plotting a linear model of HOF using Awards, Duration, and HeightInches, the number of awards and career duration are positively related to Hall of Fame status at a significant level with coefficients of 0.095 and 0.006 respectively. However, HeightInches is negatively related to Hall of Fame status at a significant level with a coefficient of -0.002. By eliminating the Duration variable, HeightInches remains negatively related however statistically insignificant. (Coefficient=-0.0007 and a p-value of 0.216). This is consistent with a probability analysis which showed a significantly negative relationship between HeightInches and HOF in the presence of performance variables (a coefficient of -0.177) (Hall of Fame Probability).

References

1

Staw, Barry M., and Ha Hoang. "Sunk Costs in the NBA: Why Draft Order Affects Playing Time and Survival in Professional Basketball." *Administrative Science Quarterly* . 40.3 (1995): 474-94. Web. 21 Nov. 2011.

2

Groothuis, Peter A., and J. Richard Hill. "EXIT DISCRIMINATION IN THE NBA: A DURATION ANALYSIS OF CAREER LENGTH." *Economic Inquiry* 42.2 (2004): 341-349. *EBSCO MegaFILE*. EBSCO. Web. 14 Oct. 2011.

3

"Basketball-Reference.com." N.p., 2011. Web. 21 Nov 2011.

4

"Hall of Fame Probability." *basketball-reference*. N.p., n.d. Web. 6 Dec 2011. <http://www.basketball-reference.com/about/hof_prob.html>