****

**MODELING NBA CAREER LENGTH**

**AUTHORS:**

**ALEK LICHUCKI, JUSTIN LY, SONALI MAYER**

**Table of Contents**

Abstract

Data Source and Background Information

Research Question

Cleaning the Data

Kaplan-Meier Estimation Curves

Log Rank Test

Model Building

Model Checking

Residual Tests

C-Log-Log Plot

Interaction Term

Hazard Ratios and Confidence Intervals

Baseline Hazard Rates

Extension 1-

Extension 2 -

Conclusion/Discussion

**Abstract**

*In the NBA, the average career length is just 5.2 years. In this project, we build a Cox Proportional Hazards model to see how the decade they started playing affects their career length, using a dataset of about 4500 players.*

**Data Source and Background Information**

Our dataset is titled NBA players stats since 1950. It is adapted from Kaggle and was originally scraped from Basketball-Reference. The dataset contains aggregate individual statistics for 67 NBA seasons. It has information on 4500 players. The unit of observation is each player and the decade in which his career started, and the event of interest is their career coming to an end.

We are using two fixed covariates in our project.

Height - measured in inches

Weight - measured in pounds

The dataset also includes the following variables:

Year Start

Year End

Position - coded F (forward), C (center), or G (guard)

Birth Date

College

**Research Question**

We are interested in what factors influence NBA player career length. We thought the decade in which each player started his NBA career would be interesting to evaluate because this could help us understand what training techniques, trends, and time-centered factors shorten or lengthen career length of an NBA player.

**Data Exploration**

Before building the Cox Proportional Hazards model, we must first understand the dataset. By using the summary function, we found that the median number of years played in the NBA was 3.5 years. The minimum and maximum were 1 year and 23 years, respectively.

years\_played

Min. : 1.000

1st Qu.: 1.000

Median : 3.500

Mean : 5.201

3rd Qu.: 8.000

Max. :23.000

Upon using the quantile function, we found that 50% of NBA players stopped playing before 3.5 years.

0% 25% 50% 75% 100%

1.0 1.0 3.5 8.0 23.0

The mean career length from this dataset specifically, which includes data on players from 1950 to players who ended their careers in 2018, was 5.2 years.

**Cleaning the Data**

As mentioned in Background Information, we manually created two columns; the decade that each player began his career, and years played in the NBA. We also wrote code to change height to be measured in inches instead of feet and inches.

While cleaning the data, we created a column for Career Length using the Year Start and Year End columns. We also created a column for the decade that each player started his career using the Year Start column.

We created the covariates height group and weight group. Height group is coded 0 for men who are less than 72 inches tall, 1 for men between 72 and 84 inches, and 2 for men who are taller than 84 inches. Weight group is coded 0 for men who weigh less than 175 pounds, 1 for men between 175 and 225 pounds, and 2 for men who weigh more than 225 pounds.

In our data, censoring is done based on whether the player is a current player or not during the year 2018 because that is when the data was last updated.

We removed any row with missing values.

Below is the code we used to clean the data.

player\_data = read.csv("player\_data.csv")

work\_data = player\_data

work\_data = na.omit(work\_data) #getting rid of columns with missing values

for (i in 1:4544) {

work\_data$years\_played[i] = work\_data$year\_end[i] - work\_data$year\_start[i] + 1#this is to find the number of years played

if (work\_data$year\_end[i] == 2018){ #censoring is done based off of if they are a current player or not year 2018 because this is when the data was last updated

work\_data$cens[i] = 0

}

else{

work\_data$cens[i] = 1

}

}

for (i in 1:4544) { #making a column to show the decade when the player started

if (work\_data$year\_start[i] >= 1940 & work\_data$year\_start[i] < 1950){

work\_data$dec\_start[i] = 1940

}

else if ((work\_data$year\_start[i] >= 1950 & work\_data$year\_start[i] < 1960)){

work\_data$dec\_start[i] = 1950

}

else if ((work\_data$year\_start[i] >= 1960 & work\_data$year\_start[i] < 1970)){

work\_data$dec\_start[i] = 1960

}

else if ((work\_data$year\_start[i] >= 1970 & work\_data$year\_start[i] < 1980)){

work\_data$dec\_start[i] = 1970

}

else if ((work\_data$year\_start[i] >= 1980 & work\_data$year\_start[i] < 1990)){

work\_data$dec\_start[i] = 1980

}

else if ((work\_data$year\_start[i] >= 1990 & work\_data$year\_start[i] < 2000)){

work\_data$dec\_start[i] = 1990

}

else if ((work\_data$year\_start[i] >= 2000 & work\_data$year\_start[i] < 2010)){

work\_data$dec\_start[i] = 2000

}

else if ((work\_data$year\_start[i] >= 2010 & work\_data$year\_start[i] < 2020)){

work\_data$dec\_start[i] = 2010

}

}

for (i in 1:4544) { #this is to convert the height to inches

work\_data$height\_str[i] = toString(work\_data$height[i])

x = unlist(strsplit(work\_data$height\_str[i], split = "-"))

work\_data$height\_inches[i] = as.integer(x[1])\*12 + as.integer(x[2])

}

for(i in 1:4544){ #making weight groups to make analysis easier

if (work\_data$weight[i] < 175){

work\_data$weight\_group[i] = 0

}

else if (work\_data$weight[i] >= 175 & work\_data$weight[i] < 225){

work\_data$weight\_group[i] = 1

}

else if (work\_data$weight[i] >= 225){

work\_data$weight\_group[i] = 2

}

}

for (i in 1:4544){ #making height groups to make analysis more manageable

if (work\_data$height\_inches[i] < 72){

work\_data$height\_group[i] = 0

}

else if (work\_data$height\_inches[i] >= 72 & work\_data$height\_inches[i] < 84){

work\_data$height\_group[i] = 1

}

else if (work\_data$height\_inches[i] >= 84){

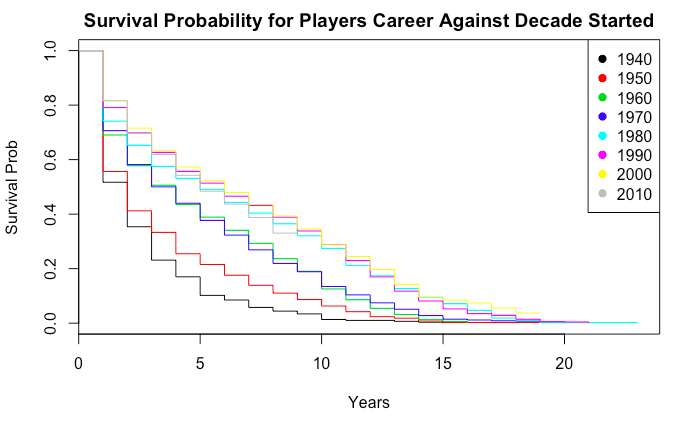
work\_data$height\_group[i] = 2

}

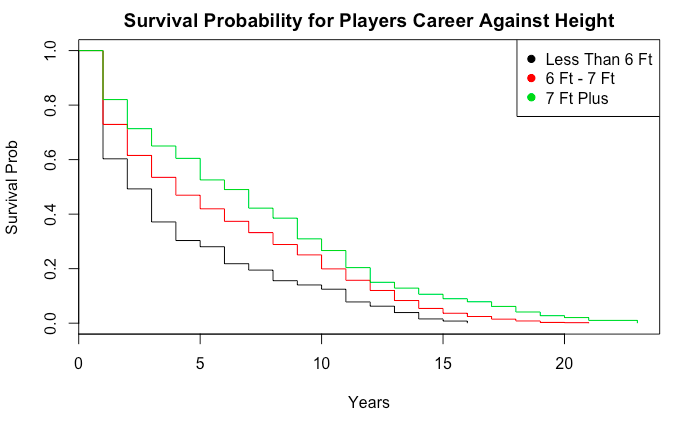
}

**Kaplan-Meier Estimation Curves**

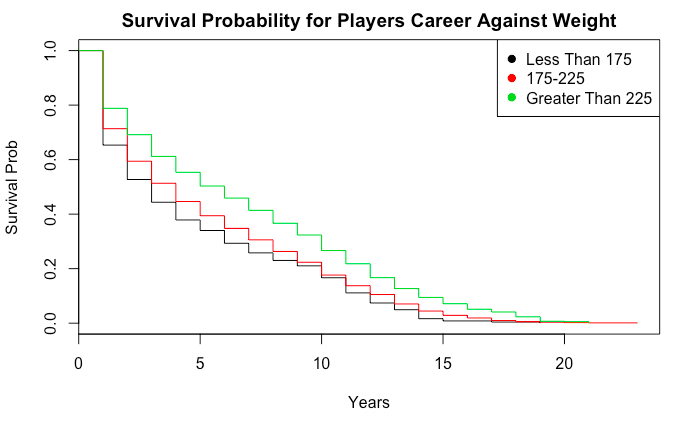
Next, we plotted Kaplan-Meier Estimation Curves to see the effects of each covariate on career length. Looking at the plots below, we can conclude that all three covariates affect career length.



Analyzing the graph directly above, we can conclude that career length has increased throughout time. Players who started in the 1940s and 1950s have significantly shorter career lengths than those who started in later decades. Career lengths look very similar throughout the 1960s and 1970s, but we observe another increase directly after. Players who started playing in the 1980s, 90s, 2000s, and 2010s have the best survival chance by far.



Looking at the estimate above, we can see that taller players tend to have a much better chance of having a long career in the NBA. Height has a direct correlation with career length.



While analyzing the survival curve above, we observe that weight is also directly correlated with career length. Following our observations about height, this makes complete sense because taller players tend to weigh more. Heavier players have longer career lengths than their lighter counterparts.

**Log Rank Test**

After creating Kapan-Meier survival curves, we conducted a Log Rank test on each variable.

All three p values were less than 0.05, so it is safe to say that all three of our covariates affect NBA career length significantly.

Call:

survdiff(formula = kap\_surv ~ work\_data$dec\_start)

N Observed Expected (O-E)^2/E (O-E)^2/V

work\_data$dec\_start=1940 294 294 145 152.1 203.0

work\_data$dec\_start=1950 381 381 234 92.0 123.5

work\_data$dec\_start=1960 478 478 396 17.0 23.4

work\_data$dec\_start=1970 684 684 591 14.7 21.5

work\_data$dec\_start=1980 634 634 731 12.9 20.1

work\_data$dec\_start=1990 689 687 818 21.0 33.5

work\_data$dec\_start=2000 689 596 766 37.8 58.2

work\_data$dec\_start=2010 695 319 391 13.3 19.2

Chisq= 467 on 7 degrees of freedom, p= <2e-16

Call:

survdiff(formula = kap\_surv ~ work\_data$height\_group)

N Observed Expected (O-E)^2/E (O-E)^2/V

work\_data$height\_group=0 136 133 95.4 14.78 18.97

work\_data$height\_group=1 4141 3716 3690.2 0.18 2.43

work\_data$height\_group=2 267 224 287.4 13.97 19.14

Chisq= 36.6 on 2 degrees of freedom, p= 1e-08

Call:

survdiff(formula = kap\_surv ~ work\_data$weight\_group)

N Observed Expected (O-E)^2/E (O-E)^2/V

work\_data$weight\_group=0 274 263 217 9.79 12.9

work\_data$weight\_group=1 2990 2729 2552 12.27 41.2

work\_data$weight\_group=2 1280 1081 1304 38.14 70.6

Chisq= 75.7 on 2 degrees of freedom, p= <2e-16

**Model Building**

After conducting our Log Rank tests, we begin to build our CoxPH model. We are building this model to pick the right covariates for analysis, based on their respective p values.

We used the step function in R to apply the backward elimination method.

We used the following code:

model\_full = coxph(kap\_surv ~ work\_data$dec\_start + work\_data$height\_group + work\_data$weight\_group)

step(model\_full, direction = "backward")

anova(model\_full)

Output:

Start: AIC=60263.94

kap\_surv ~ work\_data$dec\_start + work\_data$height\_group + work\_data$weight\_group

Df AIC

<none> 60264

- work\_data$height\_group 1 60268

- work\_data$weight\_group 1 60270

- work\_data$dec\_start 1 60558

Call:

coxph(formula = kap\_surv ~ work\_data$dec\_start + work\_data$height\_group +

work\_data$weight\_group)

coef exp(coef) se(coef) z p

work\_data$dec\_start -0.0148711 0.9852390 0.0008596 -17.300 < 2e-16

work\_data$height\_group -0.1434091 0.8663996 0.0594792 -2.411 0.01591

work\_data$weight\_group -0.0946924 0.9096527 0.0325791 -2.907 0.00365

Likelihood ratio test=378.6 on 3 df, p=< 2.2e-16

n= 4544, number of events= 4073

Analysis of Deviance Table

Cox model: response is kap\_surv

Terms added sequentially (first to last)

loglik Chisq Df Pr(>|Chi|)

NULL -30318

work\_data$dec\_start -30141 355.1399 1 < 2.2e-16 \*\*\*

work\_data$height\_group -30133 15.0080 1 0.0001071 \*\*\*

work\_data$weight\_group -30129 8.4741 1 0.0036023 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The step function did not remove any variables, which indicates that the full model is the best model to use. We also performed the log-likelihood test, and it indicated that all three variables are significant.

**Model Checking**

Because we are building a CoxPH model, we have to make sure all the covariates meet the Cox PH assumption. We used residual tests and C-log-log plots to check that our variables meet the Cox PH assumption.

**Residual Tests**

The function cox.zph tests for independence between residuals and time.

cox.zph(model\_full)

rho chisq p

work\_data$dec\_start 0.01491 0.9840 0.321

work\_data$height\_group -0.00643 0.1745 0.676

work\_data$weight\_group -0.00503 0.0996 0.752

GLOBAL NA 1.2184 0.749

All the p values are above .05 so they don't violate the Cox PH assumptions.

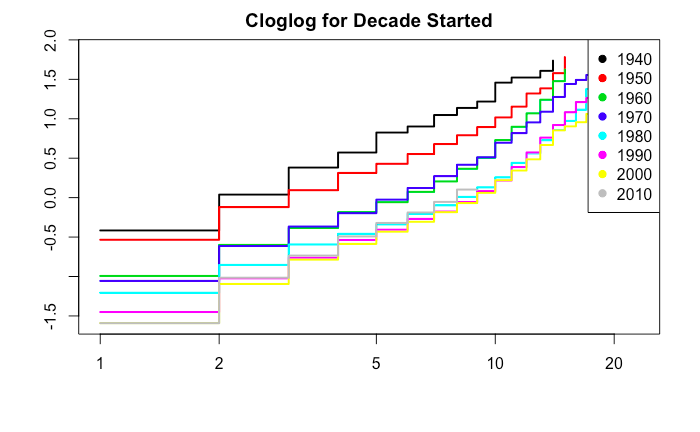
**C-Log-Log Plots**

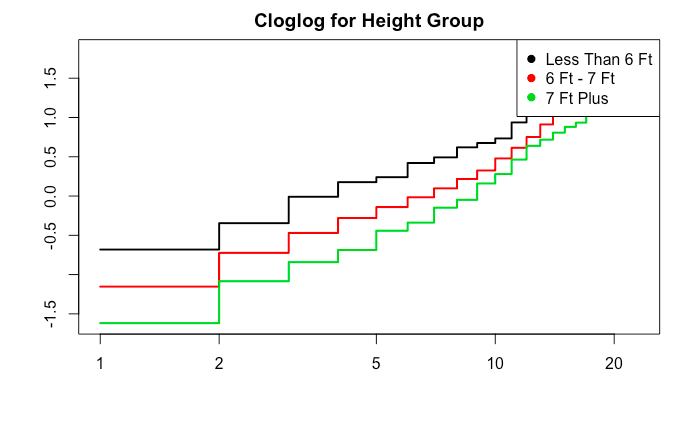
Next we check the variables via C-Log-Log plot. Below are our graphs.

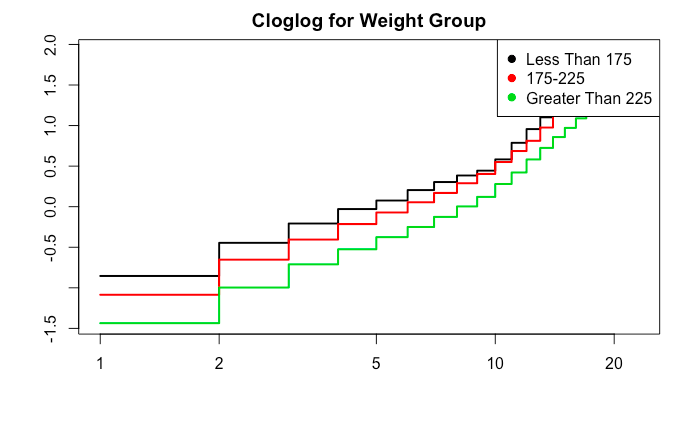
As a reminder, both height and weight are divided into three groups.

Height is coded 0 for men who are shorter than 72 inches, 1 for men between 72 and 84 inches, and 2 for men who are taller than 84 inches.

Weight is coded 0 for men who weigh less than 175 pounds, 1 for men between 175 and 225 pounds, and 2 for men who weigh more than 225 pounds.

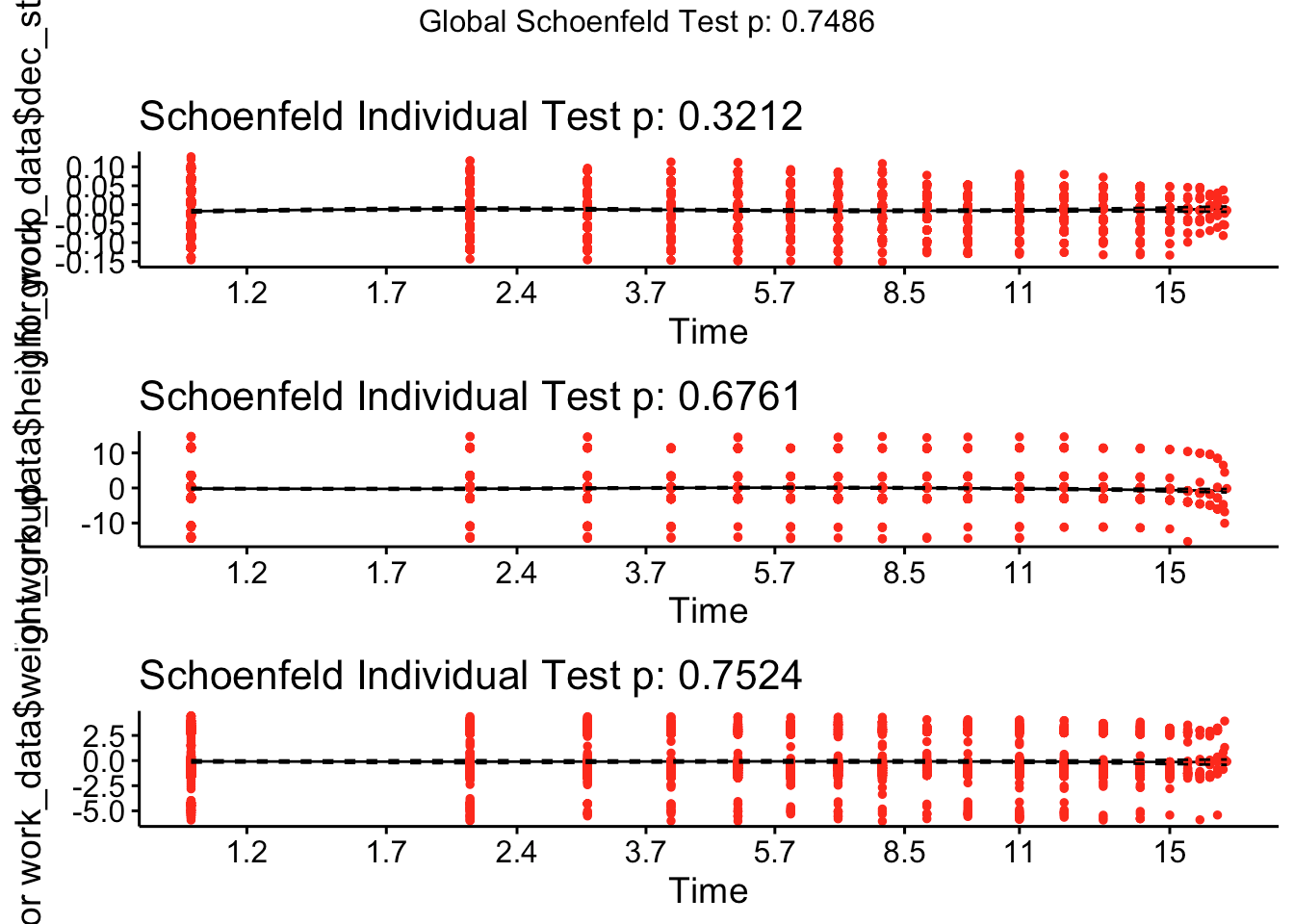
****

****

****

**Graphical Diagnostics**

ggcoxzph(cox.zph(model\_full))



We used the ggcoxzph functions to perform diagnostic checks on the full model.

These plots confirm what we concluded above from the cox.zph output, that these models do not violate any Cox PH assumptions.

**Interaction Term**

Now it is time for interaction terms in our model. The three potential interaction terms are height\_group\*weight\_group, height\_group\*dec\_start, and weight\_group\*dec\_start.

We ran the likelihood ratio test on each of the interaction terms and we found that there was only one significant interaction term, between dec\_start and weight\_group. Due to this we decided to include the interaction in our final model.

**The Final Model**

Code:

end\_model = coxph(kap\_surv ~ work\_data$dec\_start + work\_data$weight\_group\*work\_data$dec\_start + work\_data$height\_group + work\_data$weight\_group)

cox.zph(end\_model)

summary(end\_model)

Output:

rho chisq p

work\_data$dec\_start 0.01709 1.256 0.262

work\_data$weight\_group 0.01317 0.736 0.391

work\_data$height\_group -0.00778 0.249 0.618

work\_data$dec\_start:work\_data$weight\_group -0.01330 0.751 0.386

GLOBAL NA 2.334 0.675

Call:

coxph(formula = kap\_surv ~ work\_data$dec\_start + work\_data$weight\_group \*

work\_data$dec\_start + work\_data$height\_group + work\_data$weight\_group)

n= 4544, number of events= 4073

coef exp(coef) se(coef) z Pr(>|z|)

work\_data$dec\_start -2.239e-02 9.779e-01 1.970e-03 -11.362 < 2e-16

work\_data$weight\_group -1.300e+01 2.249e-06 3.052e+00 -4.261 2.04e-05

work\_data$height\_group -1.352e-01 8.736e-01 5.876e-02 -2.300 0.0214

work\_data$dec\_start:work\_data$weight\_group 6.519e-03 1.007e+00 1.541e-03 4.230 2.34e-05

work\_data$dec\_start \*\*\*

work\_data$weight\_group \*\*\*

work\_data$height\_group \*

work\_data$dec\_start:work\_data$weight\_group \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

work\_data$dec\_start 9.779e-01 1.023e+00 9.741e-01 0.9816461

work\_data$weight\_group 2.249e-06 4.446e+05 5.675e-09 0.0008914

work\_data$height\_group 8.736e-01 1.145e+00 7.785e-01 0.9801897

work\_data$dec\_start:work\_data$weight\_group 1.007e+00 9.935e-01 1.004e+00 1.0095855

Concordance= 0.598 (se = 0.006 )

Rsquare= 0.083 (max possible= 1 )

Likelihood ratio test= 396.1 on 4 df, p=<2e-16

Wald test = 420.3 on 4 df, p=<2e-16

Score (logrank) test = 430 on 4 df, p=<2e-16

**Hazard Ratios and Confidence Intervals**

We are using the ggforest() function to create a visual that demonstrates the Hazard Ratio and Confidence Interval for each covariate.

exp(coef) exp(-coef) lower .95 upper .95

work\_data$dec\_start 9.779e-01 1.023e+00 9.741e-01 0.9816461

work\_data$weight\_group 2.249e-06 4.446e+05 5.675e-09 0.0008914

work\_data$height\_group 8.736e-01 1.145e+00 7.785e-01 0.9801897

work\_data$dec\_start:work\_data$weight\_group 1.007e+00 9.935e-01 1.004e+00 1.0095855

Concordance= 0.598 (se = 0.006 )

Rsquare= 0.083 (max possible= 1 )

Likelihood ratio test= 396.1 on 4 df, p=<2e-16

For dec\_start the hazard ratio is .9779, this indicates that as decade start increases the players are at a lower hazard than the previous decade, meaning that the probability of having a longer career length increases. For weight\_group the hazard ratio is .000002249, this indicates that as the weight group increases there is a lower hazard. For height\_group the hazard ratio is .8736, meaning that as the height group increases the players have a high chance of having a longer career. Finally for the interaction between dec\_start and weight\_group the hazard ratio is 1.007, so for this interaction the players chance at having a longer career decreases.

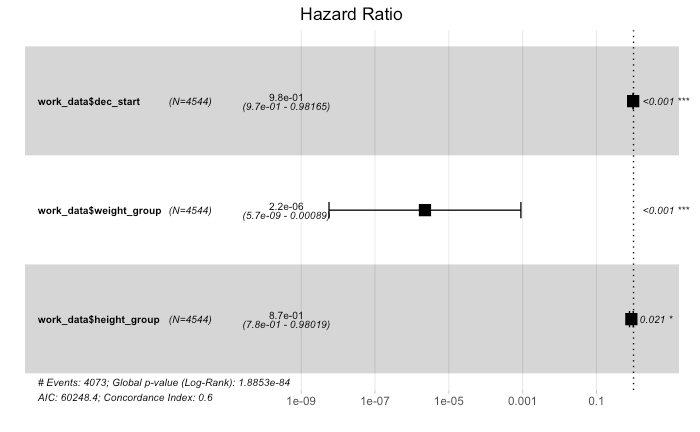
2.5 % 97.5 %

work\_data$dec\_start 9.740937e-01 0.9816460533

work\_data$weight\_group 5.674664e-09 0.0008914122

work\_data$height\_group 7.785211e-01 0.9801896834

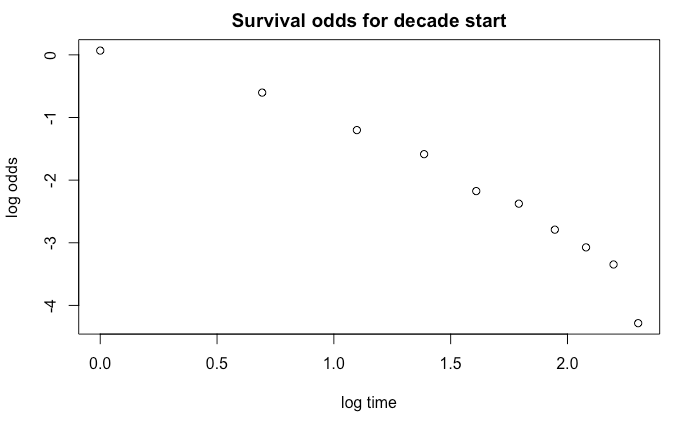
work\_data$dec\_start:work\_data$weight\_group 1.003505e+00 1.0095855163

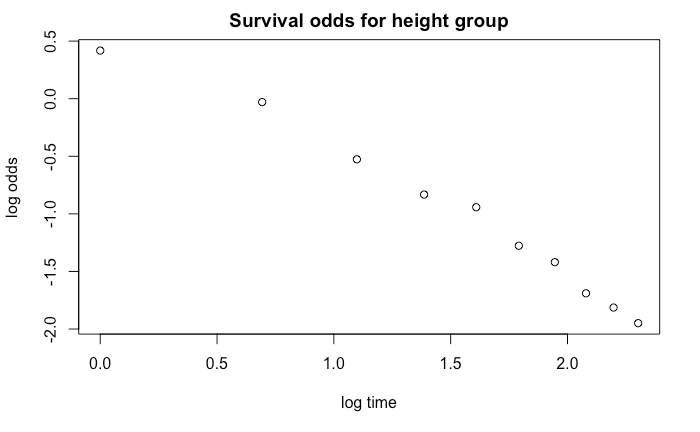


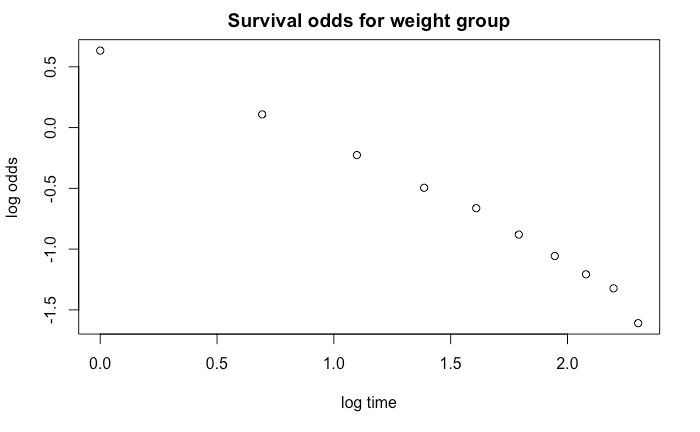
Above are the confidence intervals for dec\_start, height\_group, weight\_group, and the interaction term for dec\_start and weight\_group.

**Extension - AFT**

We built an AFT model. We draw an AFT model with log-logistic distribution so we can check our assumptions, by plotting the survival plot for decade started, height, and weight. Due to the log-logistic plots being straight lines we can use the log-logistic distribution for the AFT model.







Code:

loglig = survreg(formula = kap\_surv ~ work\_data$dec\_start + work\_data$height\_group + work\_data$weight\_group, dist = "loglogistic")

summary(loglig)

Output:

Call:

survreg(formula = kap\_surv ~ work\_data$dec\_start + work\_data$height\_group +

work\_data$weight\_group, dist = "loglogistic")

V`alue Std. Error z p

(Intercept) -2.42e+01 1.62e+00 -14.93 <2e-16

work\_data$dec\_start 1.28e-02 8.27e-04 15.46 <2e-16

work\_data$height\_group 1.10e-01 6.07e-02 1.80 0.0714

work\_data$weight\_group 9.27e-02 3.49e-02 2.66 0.0079

Log(scale) -5.00e-01 1.25e-02 -40.03 <2e-16

Scale= 0.607

Log logistic distribution

Loglik(model)= -11117.9 Loglik(intercept only)= -11279.6

Chisq= 323.46 on 3 degrees of freedom, p= 8.3e-70

Number of Newton-Raphson Iterations: 3

n= 4544

The acceleration factor we got for the difference in dec\_start was e^.0128 = 1.013. Based on this factor, the decade the player started playing in increases the chance that they have a longer career by a factor of 1.013.

The acceleration factor we got for the difference in height\_group was e^0.110 = 1.116. Based on this factor, the height group of the player lengthens the player's career by a factor of 1.116.

The acceleration factor we calculated for weight\_group is e^.0927 = 1.097. Based on this factor, the weight group of the player lengthens the player's career by a factor of 1.097.

**Conclusion**

In this project, we used a dataset of 4500 individual observations of NBA players. The dataset included information about each player's career start year, career end year, height, weight, position, and college. We knew we had to make some adjustments to answer our research question.

From there we had to censor the data. That process was done based off of if the player is an active player as of 2018. This date was decided on due to this being the last time the data was updated. We also chose to remove rows with missing values in them.

We then plotted the Kaplan-Meier curves for the overall data, then plotted the Kaplan-Meier curves for the different variables: dec\_start, weight\_group, and height\_group. We found that all three variables had a significant effect on career length through doing likelihood tests.

We then performed residual tests and created C-log-log tests to check our assumptions. Both the residual tests and the plots confirmed our assumptions and proved that all three of our variables of interest were significant, and there were no violations of the Cox Proportional Hazards assumptions.

We then looked to see if there were interaction terms for the model. We ran likelihood tests to see if there was a significant interaction for any of the terms and we found that there was one significant interaction between dec\_start and weight\_group. We then added this to our model, checked to see if the model was still valid. Once that was done we had our final Cox Proportional Hazards Model.

Our extension was an AFT model. The results confirmed that players who started their careers in later decades had longer careers overall. The results also confirmed that players who are taller have longer career lengths in the NBA, and players who weigh more have longer career lengths as well. We checked the AFT model using a stepwise log-likelihood test.

**Discussion**

The information we found and the conclusions we drew are extremely helpful in the real world, and would be extremely useful to the NBA. If humans can figure out factors that lengthen career lengths across all sports, we can change professional sports forever.

This leads to more questions though, because board members of NBA teams might not want to prolong career length. If highly paid players can play for longer, it costs the NBA more money.

Our analysis would greatly benefit from expanding using more possible covariates, by having a more complete dataset.