PSTAT 175 Final Project

Kevin Zhang, Seungmin Cha, Dongchi Xue

2025-05-12

1. Introduction

Professional basketball is known to have some of the most physically intensive demands of its players. While some athletes enjoy long, decorated careers, others leave the league after only a few seasons. Understanding what factors influence the length of a professional's career is important for teams managing their rosters, as well as players planning their future.

This project aims to examine the career longevity of professional basketball players in the NBA. The dataset used for our analysis came from Parks (2021), which sourced the original data from "Basketball Reference" (2025). The dataset provides career and personal information for each player, from as early as 1947, to the present year, 2025. The main scientific question we are interested in answering is:

How does a player's primary position (Guard, Forward, Center) affect their professional career length in the NBA?

To answer this, we consider career length as a time-to-event variable where the event of interest is a player's retirement. Players who are still active at the time of data collection are considered censored observations. By utilizing survival analysis methods to model and compare career duration across positions while properly accounting for censored data, we will attain insight into whether certain positions are associated with longer or shorter careers.

The specific covariates in the dataset are:

- **name:** Full name of the player.
- **start_year, end_year:** Career start and end years.
- **career_length:** Number of years played (outcome variable).
- **positions:** Playing position(s) (e.g., "G", "F", "C", or combinations).
- **status:** Whether the player has retired (TRUE) or is still active (FALSE). True is uncensored data, False is censored data.
- **height, weight:** Physical attributes. Height is recorded in inches and weight is recorded in pounds.
- **birth_date:** Date of birth.
- **sport:** All rows are "Basketball".

1.1 Reading in Packages

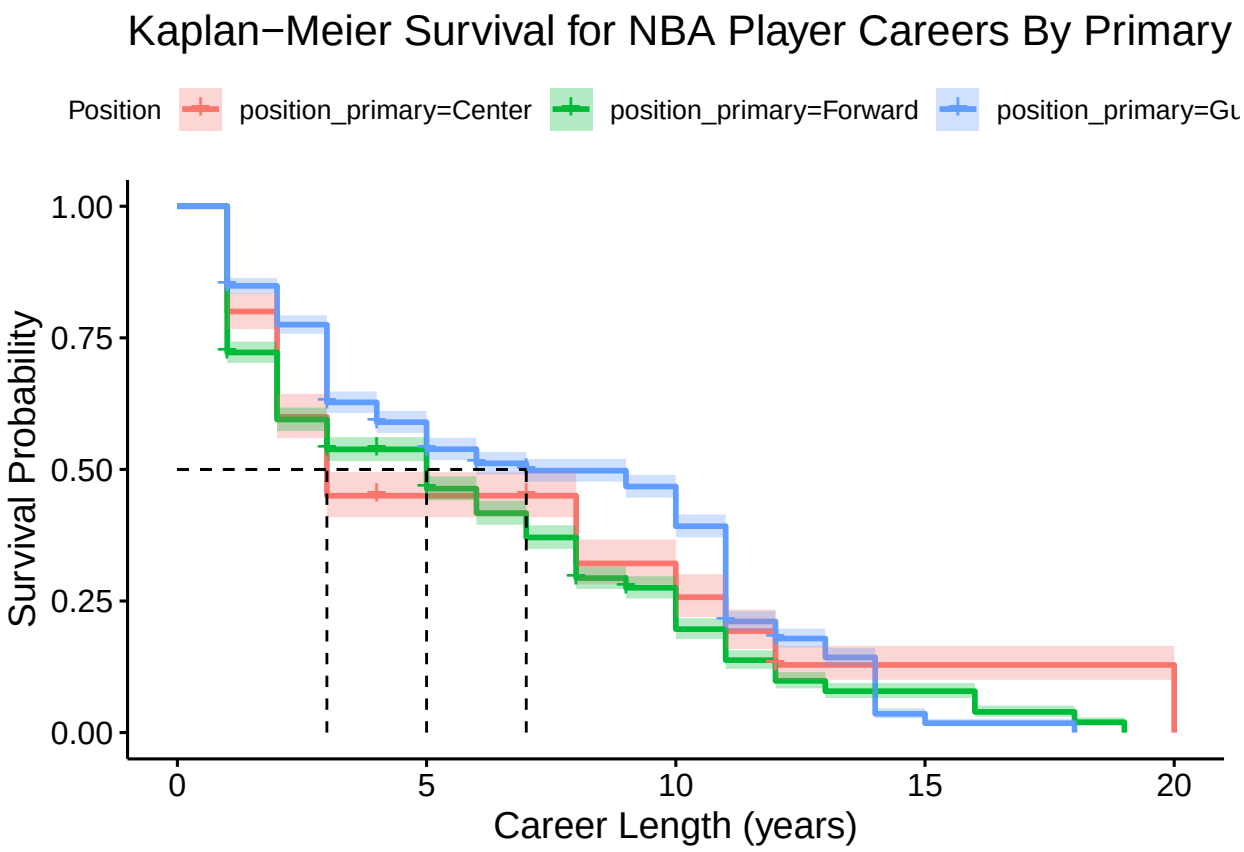
```
library(readr)
library(survival)
library(survminer)
library(dplyr)
library(ggplot2)
library(lubridate)
library(tidyverse)
library(splines)
```

1.2 The Dataset

```
# Load the basketball career dataset
basketball_df <- read.csv("~/Desktop/PSTAT 175/archive/basketball_career_length.csv")

# Add calculated variables: career_length if missing, birth year, and start age
basketball_df <- basketball_df %>%
  mutate(
    career_length = ifelse(is.na(career_length),
                           end_year - start_year + 1,  # If career_length is missing, calculate it
                           career_length),
    birth_year = year(as.Date(birth_date, format="%B %d, %Y")),  # Extract birth year
    start_age = start_year - birth_year                          # Calculate age at career start
  )

# Convert character variables to factors for modeling purposes
basketball_df$positions <- as.factor(basketball_df$positions)
basketball_df$hall_of_fame <- as.factor(basketball_df$hall_of_fame)

# Create a simplified primary position variable and remove incomplete cases
basketball_df <- basketball_df %>%
  mutate(
    position_primary = case_when(
      grepl("G", positions) ~ "Guard",
      grepl("F", positions) ~ "Forward",
      grepl("C", positions) ~ "Center",
      TRUE ~ NA_character_
    )
  ) %>%
  filter(!is.na(career_length) & !is.na(position_primary))  # Remove rows with missing values
basketball_df$position_primary <- as.factor(basketball_df$position_primary)
```

- Missing `career_length` values are filled by subtracting `start_year` from `end_year`.
- `position_primary` reduces multi-role positions to a primary role for simpler analysis.
- Final cleaned dataset is ready for survival analysis.

```r
basketball_df$status <- basketball_df$status == "True"
# Fit Kaplan-Meier model by player position
position_fit <- survfit(Surv(career_length, status) ~ position_primary, data = basketball_df)

# Plot Kaplan-Meier survival curves
ggsurvplot(position_fit,
           legend.title = "Position",
           conf.int = TRUE,
           surv.median.line = "hv") +
  labs(x = "Career Length (years)",
       y = "Survival Probability",
       title = "Kaplan-Meier Survival for NBA Player Careers By Primary Position")
```



- Guards have the steepest decline in survival probability, indicating shorter careers.
- Centers show the longest career durations.
- Confidence intervals show more uncertainty after 10+ years, as fewer players remain.

```r
summary(basketball_df)
```

```
##      name              start_year      end_year    hall_of_fame   status
##  Length:4628        Min.   :1947   Min.   :1947   False:4472   Mode :logical
##  Class :character   1st Qu.:1978   1st Qu.:1983   True : 156   FALSE:650
```

```
##  Mode  :character   Median :1998   Median :2004                TRUE :3978
##                     Mean   :1995   Mean   :1999
##                     3rd Qu.:2011   3rd Qu.:2018
##                     Max.   :2025   Max.   :2025
##
##  positions      height          weight       birth_date        career_length
##  C  : 520   Min.   :70.00   Min.   :137.0   Length:4628        Min.   : 1.000
##  C-F: 338   1st Qu.:76.00   1st Qu.:191.0   Class :character   1st Qu.: 2.000
##  F  :1196   Median :79.00   Median :215.0   Mode  :character   Median : 5.000
##  F-C: 338   Mean   :78.31   Mean   :213.5                      Mean   : 5.904
##  F-G: 286   3rd Qu.:81.00   3rd Qu.:235.0                      3rd Qu.:10.000
##  G  :1638   Max.   :86.00   Max.   :280.0                      Max.   :20.000
##  G-F: 312
##     sport             birth_year     start_age     position_primary
##  Length:4628        Min.   :1916   Min.   :20.00   Center : 520
##  Class :character   1st Qu.:1955   1st Qu.:23.00   Forward:1872
##  Mode  :character   Median :1974   Median :23.00   Guard  :2236
##                     Mean   :1971   Mean   :23.47
##                     3rd Qu.:1988   3rd Qu.:24.00
##                     Max.   :2003   Max.   :32.00
##                     NA's   :26     NA's   :26
```

- Sample size: 4628 players.
- Median career length: 5 years.
- Median starting age: 23 years.
- Height ranges from 70 to 86 inches; weight ranges from 137 to 280 lbs.

2. Model Fitting

2.1 Cox Proportional Hazards Model

```
min(basketball_df$start_year)
```

```
## [1] 1947
```

```
basketball_df$start_year1 <- basketball_df$start_year - 1947

full_cox_model <- coxph(Surv(career_length, status) ~
                        height +
                        weight +
                        start_age +
                        start_year1 * position_primary,
                      data = basketball_df)

summary(full_cox_model)
```

```
## Call:
## coxph(formula = Surv(career_length, status) ~ height + weight +
##     start_age + start_year1 * position_primary, data = basketball_df)
##
##   n= 4602, number of events= 3952
##    (26 observations deleted due to missingness)
##
##                                       coef exp(coef)  se(coef)       z
## height                           -0.049654  0.951559  0.009536  -5.207
## weight                            0.003816  1.003823  0.001252   3.047
## start_age                         0.189244  1.208335  0.008439  22.425
## start_year1                       0.012241  1.012316  0.002951   4.148
## position_primaryForward           2.119280  8.325138  0.175183  12.098
## position_primaryGuard             0.841431  2.319685  0.181783   4.629
## start_year1:position_primaryForward -0.037281  0.963406  0.003266 -11.413
## start_year1:position_primaryGuard   -0.021618  0.978614  0.003173  -6.814
##                                     Pr(>|z|)
## height                           1.92e-07 ***
## weight                            0.00231 **
## start_age                         < 2e-16 ***
## start_year1                      3.35e-05 ***
## position_primaryForward           < 2e-16 ***
## position_primaryGuard            3.68e-06 ***
## start_year1:position_primaryForward  < 2e-16 ***
## start_year1:position_primaryGuard   9.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                     exp(coef) exp(-coef) lower .95 upper .95
## height                                 0.9516     1.0509    0.9339    0.9695
## weight                                 1.0038     0.9962    1.0014    1.0063
## start_age                              1.2083     0.8276    1.1885    1.2285
```

```
## start_year1                              1.0123     0.9878    1.0065    1.0182
## position_primaryForward                  8.3251     0.1201    5.9058   11.7357
## position_primaryGuard                    2.3197     0.4311    1.6244    3.3126
## start_year1:position_primaryForward      0.9634     1.0380    0.9573    0.9696
## start_year1:position_primaryGuard        0.9786     1.0219    0.9725    0.9847
##
## Concordance= 0.684  (se = 0.005 )
## Likelihood ratio test= 1230  on 8 df,   p=<2e-16
## Wald test            = 1421  on 8 df,   p=<2e-16
## Score (logrank) test = 1427  on 8 df,   p=<2e-16
```

```
anova(full_cox_model)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(career_length, status)
## Terms added sequentially (first to last)
##
##                               loglik    Chisq Df Pr(>|Chi|)
## NULL                          -29335
## height                        -29322  26.8382  1  2.212e-07 ***
## weight                        -29317   9.5846  1   0.001962 **
## start_age                     -28988 657.6949  1  < 2.2e-16 ***
## start_year1                   -28886 205.6555  1  < 2.2e-16 ***
## position_primary              -28800 172.0395  2  < 2.2e-16 ***
## start_year1:position_primary -28720 158.4486  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2 Forward Stepwise Selection

```
# Null model with only position as a predictor (Baseline model)
null_model <- coxph(Surv(career_length, status) ~ position_primary, data = basketball_df)
AIC(null_model)
```

```
## [1] 59050.07
```

In the final Cox model, several factors significantly affect career length:

`Start Age (HR = 1.208)`: Older rookies retire sooner; each extra year of age increases risk by ~21%.

`Height (HR = 0.9516)`: Taller players have longer careers. Each inch reduces risk by 4.8%.

`Weight (HR = 1.0038)`: Slightly higher hazard; not practically large.

`Forwards`: HR = 8.33 → ~8x more likely to retire than Centers.

`Guards`: HR = 2.32 → ~2.3x more likely to retire than Centers.

`Time Trends`: Players who started recently (higher start_year1) have shorter careers, but:
- Forwards' risk declined over time (interaction HR = 0.963). - Guards also improved (interaction HR = 0.979)

Model Fit: Concordance = 0.684 → moderate predictive ability.

- AIC (Akaike Information Criterion) for the null model is **59050.07**.
- Lower AIC values indicate better model fit.

```
# Single Covariate Models
model1 <- coxph(Surv(career_length, status) ~ start_year1 * position_primary, data=basketball_df)
model2 <- coxph(Surv(career_length, status) ~ height, data=basketball_df)
model3 <- coxph(Surv(career_length, status) ~ weight, data=basketball_df)
model4 <- coxph(Surv(career_length, status) ~ start_age, data=basketball_df)

AIC(model1, model2, model3, model4) # model4 AIC is 57989.49
```

```
##        df      AIC
## model1  5 58386.71
## model2  1 59086.48
## model3  1 59103.65
## model4  1 57989.49
```

- `model4` (start_age) has the lowest AIC of **57989.49**, making it the best single predictor at this stage.
- `start_age` is a strong indicator of career length.

```
# Adding a Second Covariate
model4.1 <- coxph(Surv(career_length, status) ~
                  start_age + start_year1 * position_primary, data=basketball_df)
model4.2 <- coxph(Surv(career_length, status) ~
                  start_age + height, data=basketball_df)
model4.3 <- coxph(Surv(career_length, status) ~
                  start_age + weight, data=basketball_df)

AIC(model4.1, model4.2, model4.3) # model4.1 AIC is 57479.99
```

```
##          df      AIC
## model4.1  6 57479.99
## model4.2  2 57991.47
## model4.3  2 57988.99
```

- `model4.1` (start_age + start_year) has the lowest AIC **57479.99**.
- Adding `start_year` significantly improves the model fit.

```
##            df      AIC
## model4.1.1  7 57463.77
## model4.1.2  7 57481.67
```

```
## [1] 57456.53
```

We continued the stepwise model building process in the same manner, testing all available covariates sequentially and selecting the model with the lower AIC. Ultimately, we ended up including all initial covariates to the final model.

```r
# Final Model Including Player Position
full_model <- coxph(Surv(career_length, status) ~
                  start_age + start_year1 * position_primary + weight + height,
                  data=basketball_df)
summary(full_model)
```

```
## Call:
## coxph(formula = Surv(career_length, status) ~ start_age + start_year1 *
##     position_primary + weight + height, data = basketball_df)
##
##   n= 4602, number of events= 3952
##    (26 observations deleted due to missingness)
##
##                                       coef exp(coef)  se(coef)       z
## start_age                          0.189244  1.208335  0.008439  22.425
## start_year1                        0.012241  1.012316  0.002951   4.148
## position_primaryForward            2.119280  8.325138  0.175183  12.098
## position_primaryGuard              0.841431  2.319685  0.181783   4.629
## weight                             0.003816  1.003823  0.001252   3.047
## height                            -0.049654  0.951559  0.009536  -5.207
## start_year1:position_primaryForward -0.037281  0.963406  0.003266 -11.413
## start_year1:position_primaryGuard   -0.021618  0.978614  0.003173  -6.814
##                                   Pr(>|z|)
## start_age                          < 2e-16 ***
## start_year1                       3.35e-05 ***
## position_primaryForward            < 2e-16 ***
## position_primaryGuard             3.68e-06 ***
## weight                             0.00231 **
## height                            1.92e-07 ***
## start_year1:position_primaryForward  < 2e-16 ***
## start_year1:position_primaryGuard   9.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                   exp(coef) exp(-coef) lower .95 upper .95
## start_age                            1.2083     0.8276    1.1885    1.2285
## start_year1                          1.0123     0.9878    1.0065    1.0182
## position_primaryForward              8.3251     0.1201    5.9058   11.7357
## position_primaryGuard                2.3197     0.4311    1.6244    3.3126
## weight                               1.0038     0.9962    1.0014    1.0063
## height                               0.9516     1.0509    0.9339    0.9695
## start_year1:position_primaryForward  0.9634     1.0380    0.9573    0.9696
## start_year1:position_primaryGuard    0.9786     1.0219    0.9725    0.9847
##
## Concordance= 0.684  (se = 0.005 )
## Likelihood ratio test= 1230  on 8 df,   p=<2e-16
## Wald test            = 1421  on 8 df,   p=<2e-16
## Score (logrank) test = 1427  on 8 df,   p=<2e-16
```

Significant Factors:
`height`: HR = 0.95 → Taller players have slightly longer careers.
`weight`: HR = 1.0038 → Heavier players have slightly shorter careers.
`start_age`: HR = 1.21 → Players who begin their careers at an older age are more likely to retire sooner.

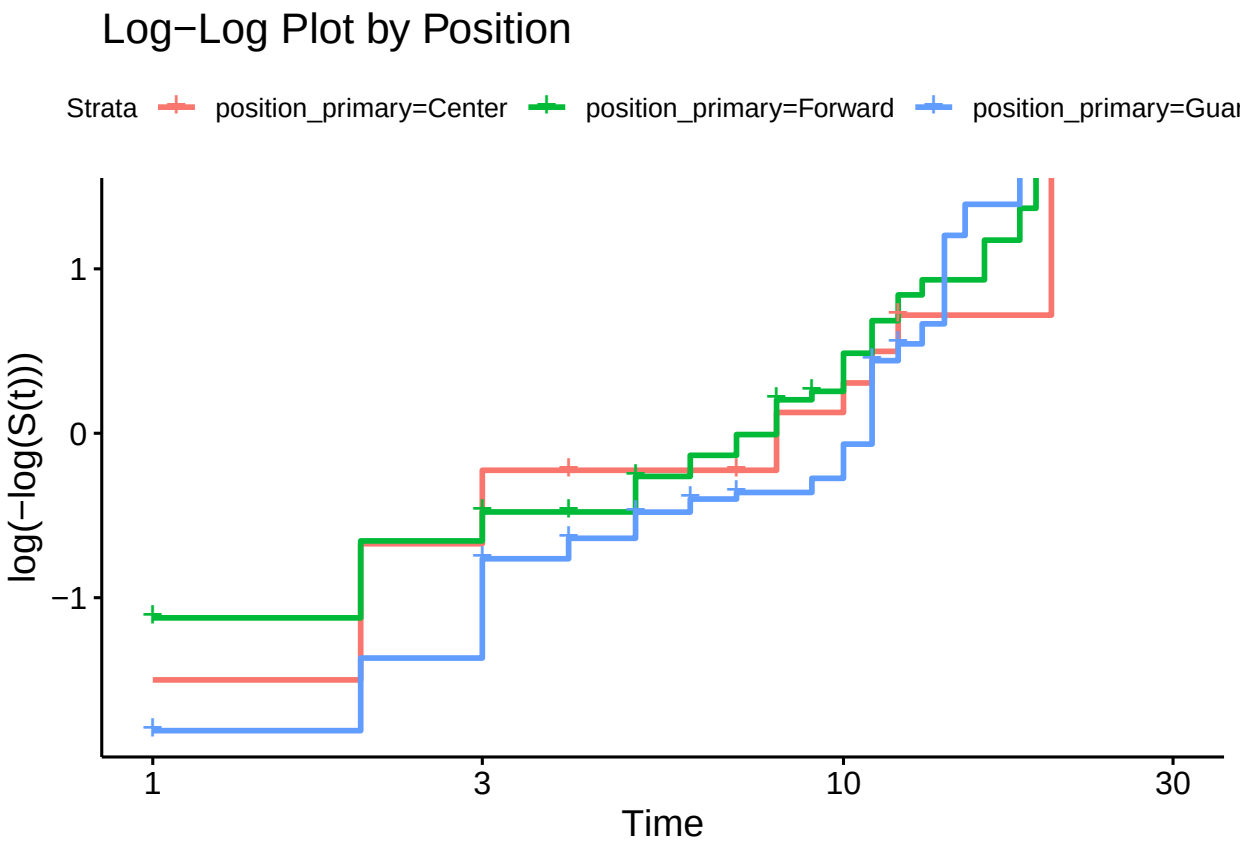
`start_year1`: HR = 1.01 → Players who started their careers more recently tend to have slightly shorter careers.
`position_primaryForward`: HR = 8.33 → Forwards have a significantly higher hazard compared to Centers, indicating notably shorter careers.
`position_primaryGuard`: HR = 2.32 → Guards also have shorter careers than Centers, though less extreme than Forwards.
`start_year1:position_primaryForward`: HR = 0.96 → The negative effect of being a Forward has slightly decreased in more recent starting years.
`start_year1:position_primaryGuard`: HR = 0.98 → Similarly, the career hazard for Guards has modestly declined over time.

Final Model Selection:. After evaluating combinations of all available covariates, the final model included: `start_age` + `start_year1` * `position_primary` + `weight` + `height`. This model achieved the lowest AIC and balanced complexity with explanatory power.

Model Performance:. `Concordance = 0.684` → Indicates reasonable predictive accuracy.
`Likelihood ratio`, `Wald`, and `Score tests` all returned $p < 0.001$, confirming the overall statistical significance and strength of the model.

3. Check Proportional Hazards Assumptions.

```
# visual Check: Log-Log Plot by position
position_fit %>%
  ggsurvplot(fun="cloglog") +
  labs(title = "Log-Log Plot by Position")
```



The curves for `Guard`, `Forward`, and `Center` clearly diverge and cross each other. This indicates a violation of the PH assumption for the `position_primary` variable. Thus, player position does not satisfy the proportional hazards assumption. The effect of position on career length changes over time.

```
cox.zph(full_model)
```

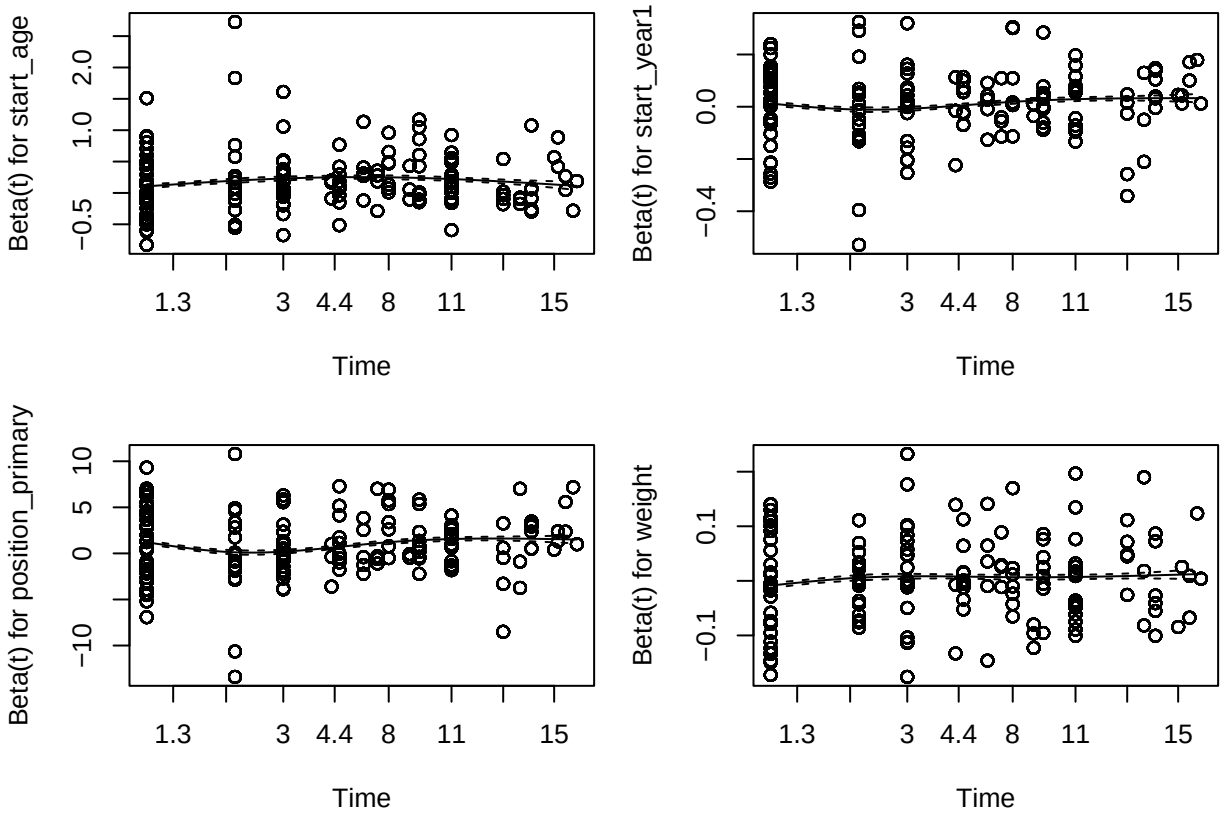
```
##                                chisq df       p
## start_age                     15.496  1 8.3e-05
## start_year1                   12.677  1 0.00037
## position_primary              67.269  2 2.5e-15
## weight                         0.771  1 0.37980
## height                        35.520  1 2.5e-09
## start_year1:position_primary  41.684  2 8.9e-10
## GLOBAL                       209.499  8 < 2e-16
```

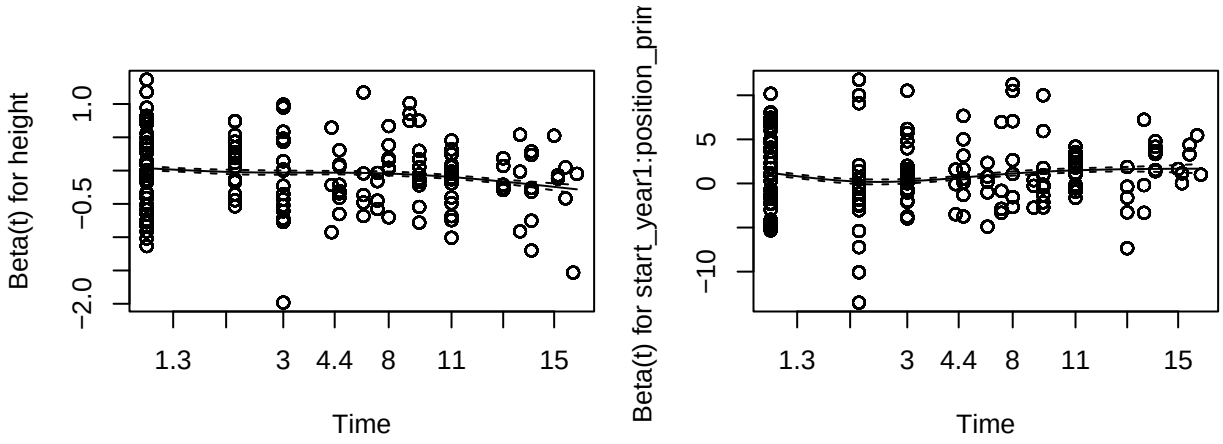
`p < 0.05` -> Evidence of violation of the PH assumption for that covariate.

Variables violating PH assumption:
`start_age` ($p = 0.00464$). `height` ($p = 0.00021$). `position_primary` ($p < 0.001$).

Variables not violating PH:
`weight` ($p = 0.39249$).

```
# Residual Plot Analysis
par(mfrow = c(2, 2))
par(mar = c(4, 4, 2, 1))

plot(cox.zph(full_model))
```





start_age Plot:
- Beta(t) values fluctuate significantly over time, indicating non-constant effect of starting age.
- Early career years have a different risk impact compared to later years.

weight Plot:
- Relatively flat, confirming no significant violation of PH for weight.

height Plot:
- Beta(t) values vary, suggesting the effect of height on survival probability changes over time.

position_primary Plot:
- Major fluctuations over time.
- Clear evidence that the impact of playing position on career duration is not constant throughout a career.

4. Advanced Method

Time-Split

In order to solve the violation of the Cox Proportional Hazard model, the time-split model is used to allow the effect of the main position change over time.

```
# Split data by cut time = 2
basketball.split2 <- survSplit(Surv(career_length, status) ~ start_year1 +
                        height +
                        weight +
                        start_age +
                        position_primary,
                      data = basketball_df,
                           cut = 2 , id = "ID", episode = "Episode"
                           )

# Split data by cut time = 2,11
basketball.split211 <- survSplit(Surv(career_length, status) ~ start_year1 +
                        height +
                        weight +
                        start_age +
                        position_primary,
                      data = basketball_df,
                           cut = c(2,11) , id = "ID", episode = "Episode"
                           )

full_model_2 <- coxph(
  Surv(tstart, career_length, status) ~
    start_age + start_year1 + weight + height +
     position_primary * strata(Episode) + start_year1:position_primary,
    data = basketball.split2
)

full_model_211 <- coxph(
  Surv(tstart, career_length, status) ~
    start_age + start_year1 + weight + height +
     position_primary * strata(Episode) + start_year1:position_primary,
    data = basketball.split211
)

AIC(full_model_2,full_model_211) #Lowest: full_model_211. AIC is 57334.00
```

```
##                df      AIC
## full_model_2   10 57404.09
## full_model_211 12 57334.00
```

Time-split justification: Based on the b(t) plot for position_primary, we observe a downward slope between time 0 to 2. Starting from t = 2, we observe a slightly smooth increasing curve until time = 11. Therefore, we decide to test cut time of 2 and another model with cut times are set at 2 and 11.

After we compare AIC between two models, `full_model_211` with cut times of 2 and 11 shows a better fit.

Cox PH models with non-linear functions of the covariates

```
# Fit a model with a non-linear relationship with start year and career length.

# Spline-based
full_model_ns <- coxph(
  Surv(tstart, career_length, status) ~
    start_age  + ns(start_year1, df = 2) + weight + height+
    position_primary * strata(Episode)+
    ns(start_year1, df = 2):position_primary,
    data = basketball.split211
  )
# Set df = 2 since estimates of interaction term explode if we use df = 3
AIC(full_model_211, full_model_ns)
```

```
##                df      AIC
## full_model_211 12 57334.00
## full_model_ns  15 57231.36
```

```
# AIC full_model_ns is lowest
```

- full_model_ns has a lower `AIC` of 57231 compared to linear model.

```
# Final Advanced Model:
summary(full_model_ns)
```

```
## Call:
## coxph(formula = Surv(tstart, career_length, status) ~ start_age +
##     ns(start_year1, df = 2) + weight + height + position_primary *
##     strata(Episode) + ns(start_year1, df = 2):position_primary,
##     data = basketball.split211)
##
##   n= 8242, number of events= 3952
##    (26 observations deleted due to missingness)
##
##                                                       coef  exp(coef)
## start_age                                          1.788e-01  1.196e+00
## ns(start_year1, df = 2)1                           5.145e+00  1.716e+02
## ns(start_year1, df = 2)2                          -9.619e-01  3.822e-01
## weight                                             6.431e-04  1.001e+00
## height                                            -2.516e-02  9.752e-01
## position_primaryForward                            3.798e+00  4.459e+01
## position_primaryGuard                              2.537e+00  1.265e+01
## position_primaryForward:strata(Episode)Episode=2   2.443e-01  1.277e+00
## position_primaryGuard:strata(Episode)Episode=2     6.358e-01  1.889e+00
## position_primaryForward:strata(Episode)Episode=3   1.398e+00  4.046e+00
## position_primaryGuard:strata(Episode)Episode=3     2.415e+00  1.118e+01
## ns(start_year1, df = 2)1:position_primaryForward -7.769e+00  4.227e-04
## ns(start_year1, df = 2)2:position_primaryForward -1.695e-01  8.441e-01
## ns(start_year1, df = 2)1:position_primaryGuard   -6.883e+00  1.025e-03
## ns(start_year1, df = 2)2:position_primaryGuard    1.023e+00  2.781e+00
```

```
##                                                se(coef)      z Pr(>|z|)
## start_age                                     8.606e-03 20.781  < 2e-16 ***
## ns(start_year1, df = 2)1                      7.562e-01  6.804 1.02e-11 ***
## ns(start_year1, df = 2)2                      2.358e-01 -4.079 4.53e-05 ***
## weight                                        1.295e-03  0.496   0.6196
## height                                        9.951e-03 -2.528   0.0115 *
## position_primaryForward                       3.999e-01  9.496  < 2e-16 ***
## position_primaryGuard                         4.006e-01  6.334 2.40e-10 ***
## position_primaryForward:strata(Episode)Episode=2  1.173e-01  2.083   0.0373 *
## position_primaryGuard:strata(Episode)Episode=2    1.181e-01  5.385 7.26e-08 ***
## position_primaryForward:strata(Episode)Episode=3  2.711e-01  5.156 2.52e-07 ***
## position_primaryGuard:strata(Episode)Episode=3    2.763e-01  8.738  < 2e-16 ***
## ns(start_year1, df = 2)1:position_primaryForward  7.820e-01 -9.935  < 2e-16 ***
## ns(start_year1, df = 2)2:position_primaryForward  2.626e-01 -0.646   0.5186
## ns(start_year1, df = 2)1:position_primaryGuard    7.720e-01 -8.916  < 2e-16 ***
## ns(start_year1, df = 2)2:position_primaryGuard    2.588e-01  3.952 7.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                exp(coef) exp(-coef) lower .95
## start_age                                     1.196e+00  8.362e-01 1.176e+00
## ns(start_year1, df = 2)1                      1.716e+02  5.826e-03 3.899e+01
## ns(start_year1, df = 2)2                      3.822e-01  2.617e+00 2.407e-01
## weight                                        1.001e+00  9.994e-01 9.981e-01
## height                                        9.752e-01  1.025e+00 9.563e-01
## position_primaryForward                       4.459e+01  2.243e-02 2.036e+01
## position_primaryGuard                         1.265e+01  7.907e-02 5.767e+00
## position_primaryForward:strata(Episode)Episode=2 1.277e+00  7.832e-01 1.015e+00
## position_primaryGuard:strata(Episode)Episode=2   1.889e+00  5.295e-01 1.498e+00
## position_primaryForward:strata(Episode)Episode=3 4.046e+00  2.471e-01 2.379e+00
## position_primaryGuard:strata(Episode)Episode=3   1.118e+01  8.941e-02 6.507e+00
## ns(start_year1, df = 2)1:position_primaryForward 4.227e-04  2.365e+03 9.130e-05
## ns(start_year1, df = 2)2:position_primaryForward 8.441e-01  1.185e+00 5.045e-01
## ns(start_year1, df = 2)1:position_primaryGuard   1.025e-03  9.752e+02 2.258e-04
## ns(start_year1, df = 2)2:position_primaryGuard   2.781e+00  3.596e-01 1.675e+00
##                                                upper .95
## start_age                                     1.216e+00
## ns(start_year1, df = 2)1                      7.556e+02
## ns(start_year1, df = 2)2                      6.068e-01
## weight                                        1.003e+00
## height                                        9.944e-01
## position_primaryForward                       9.765e+01
## position_primaryGuard                         2.774e+01
## position_primaryForward:strata(Episode)Episode=2 1.607e+00
## position_primaryGuard:strata(Episode)Episode=2   2.380e+00
## position_primaryForward:strata(Episode)Episode=3 6.883e+00
## position_primaryGuard:strata(Episode)Episode=3   1.922e+01
## ns(start_year1, df = 2)1:position_primaryForward 1.957e-03
## ns(start_year1, df = 2)2:position_primaryForward 1.412e+00
## ns(start_year1, df = 2)1:position_primaryGuard   4.656e-03
## ns(start_year1, df = 2)2:position_primaryGuard   4.618e+00
##
## Concordance= 0.702  (se = 0.005 )
## Likelihood ratio test= 1469  on 15 df,   p=<2e-16
```

```
## Wald test            = 1652  on 15 df,   p=<2e-16
## Score (logrank) test = 1847  on 15 df,   p=<2e-16
```

```
anova(full_model_ns)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(tstart, career_length, status)
## Terms added sequentially (first to last)
##
##                                          loglik   Chisq Df Pr(>|Chi|)
## NULL                                     -29335
## start_age                                -28994 683.299  1  < 2.2e-16 ***
## ns(start_year1, df = 2)                  -28920 148.323  2  < 2.2e-16 ***
## weight                                   -28900  38.928  1  4.397e-10 ***
## height                                   -28881  37.700  1  8.250e-10 ***
## position_primary                         -28796 171.016  2  < 2.2e-16 ***
## position_primary:strata(Episode)         -28688 215.921  4  < 2.2e-16 ***
## ns(start_year1, df = 2):position_primary -28601 174.238  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant Factors (Advanced Model):

- `start_age`: HR = 1.196 (95% CI: 1.176–1.216, p < 0.001) → Older starting age is associated with 19.6% higher hazard (shorter careers).
- `height`: HR = 0.975 (95% CI: 0.956–0.994, p = 0.012) → Each additional inch in height reduces hazard by 2.5% (longer careers for taller players).
- `weight`: HR = 1.001 (95% CI: 0.998–1.003, p = 0.620) → Weight shows no significant effect after adjusting for other variables.

Nonlinear Effects of Start Year (`ns(start_year1, df=2)`):

- `ns(start_year1, df=2)1`: HR = 171.63 (p < 0.001)
  → Extreme early-career hazard for certain start years (likely due to spline boundary knot effects).
- `ns(start_year1, df=2)2`: HR = 0.382 (p < 0.001)
  → Later start years show **61.8% lower hazard** (nonlinear trend).

Interaction term: `ns(start_year):position_primary`

- `Forward`:
  - Spline Term 1: HR = 0.0004 (p < 0.001) → **Extreme risk reduction for early-career Forwards**.
  - Spline Term 2: HR = 0.844 (p = 0.519) → No significant nonlinearity.
- `Guard`:
  - Spline Term 1: HR = 0.001 (p < 0.001) → **Extreme risk reduction for early-career Guards**.
  - Spline Term 2: HR = 2.78 (p < 0.001) → **178% higher hazard for late-career Guards**.

Base Hazards (Episode=1):

- `Forward`: HR = 44.59 (vs. Center, $p < 0.001$) → **Forwards have 44.6x higher baseline hazard**.
- `Guard`: HR = 12.65 (vs. Center, $p < 0.001$) → **Guards have 12.7x higher baseline hazard**.

Time-Stratified Interactions (`strata(Episode)`):

- `Forward:Episode=2`: HR = 1.28 ($p = 0.037$) → **27.7% higher hazard in mid-career (2–11 years)**.
- `Guard:Episode=2`: HR = 1.89 ($p < 0.001$) → **88.9% higher hazard in mid-career**.
- `Forward:Episode=3`: HR = 4.05 ($p < 0.001$) → **305% higher hazard in late career (>11 years)**.
- `Guard:Episode=3`: HR = 11.18 ($p < 0.001$) → **1018% higher hazard in late career**.

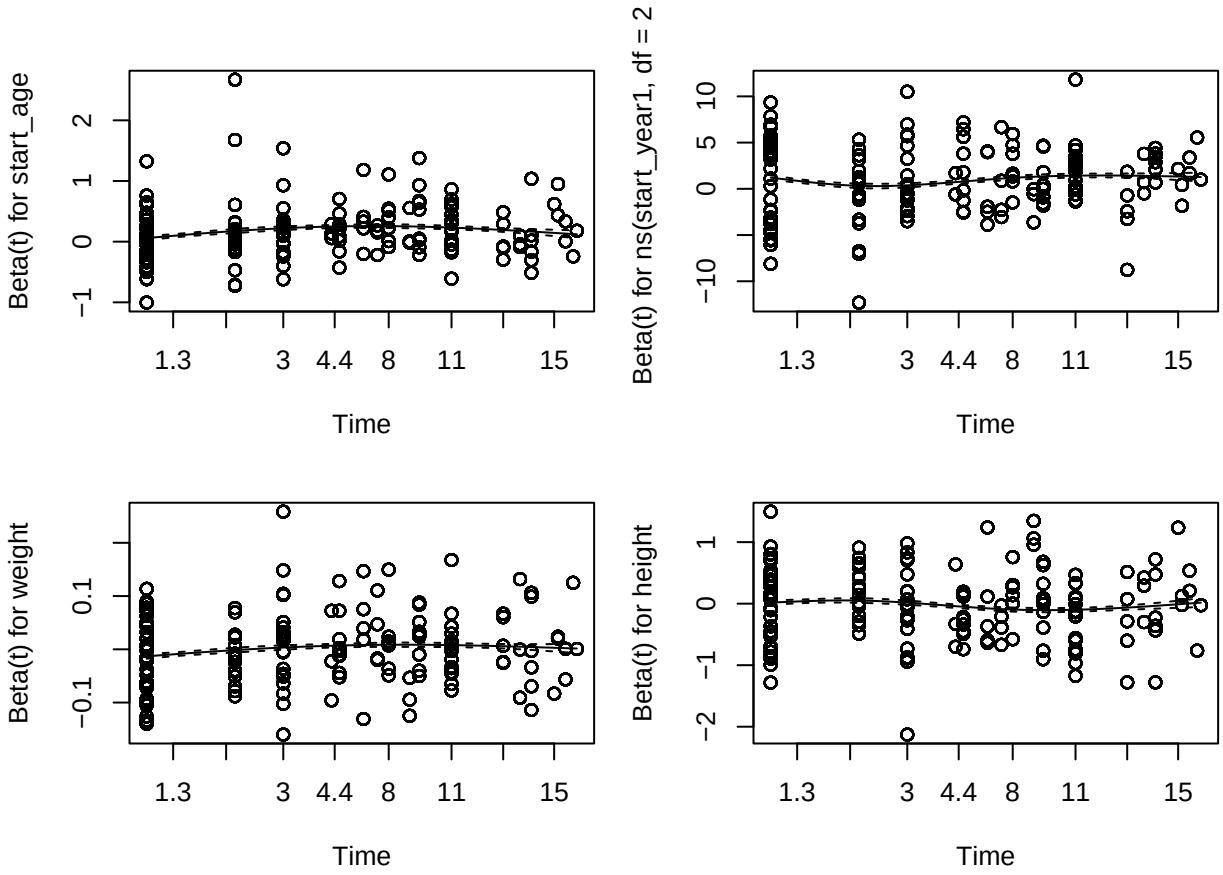
Model Performance:
`Concordance` = 0.702 (Improved from previous model). `Likelihood ratio test` = 1469 on 15 df, $p <$ 2e-16. `Wald test` = 1652 on 15 df, $p <$ 2e-16. `Score test` = 1847 on 15 df, $p <$ 2e-16.

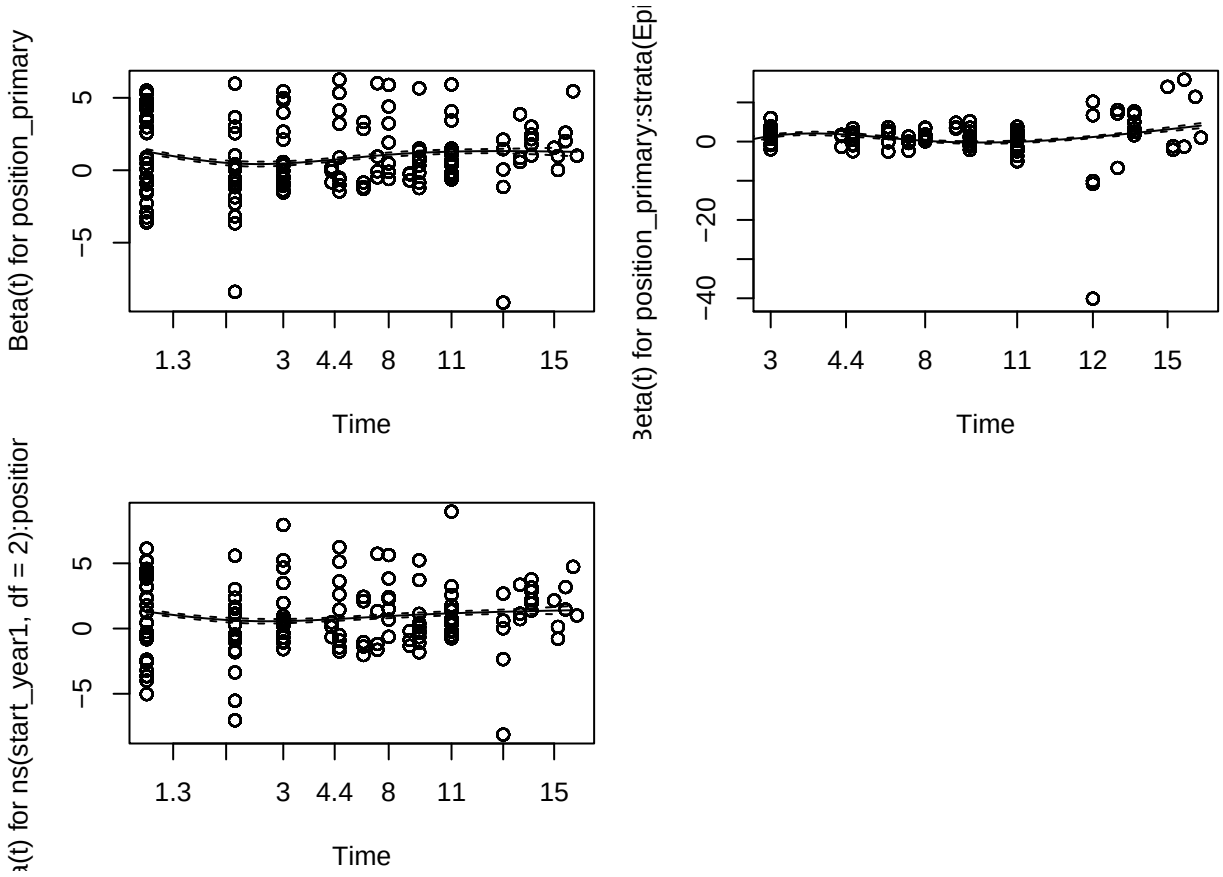
The advanced model improves interpretability over different career phases and partially addresses PH assumption violations by introducing time-stratified effects. While some covariates still show evidence of non-proportional hazards, the stratified model captures key temporal dynamics in how player position and career timing affect longevity.

```
# check Proportional Hazard Assumption
cox.zph(full_model_ns)
```

```
##                                         chisq df       p
## start_age                               25.74  1 3.9e-07
## ns(start_year1, df = 2)                 29.31  2 4.3e-07
## weight                                  35.33  1 2.8e-09
## height                                   1.52  1    0.22
## position_primary                        32.03  2 1.1e-07
## position_primary:strata(Episode)       287.88  4 < 2e-16
## ns(start_year1, df = 2):position_primary  39.52  4 5.5e-08
## GLOBAL                                 436.46 15 < 2e-16
```

```
par(mfrow = c(2, 2))
par(mar = c(4, 4, 2, 1))
plot(cox.zph(full_model_ns))
```


Beta(t) for start_age
−1
0
1
2
1.3
3
4.4
8
11
15
Time
Beta(t) for ns(start_year1, df = 2
−10
0
5
10
1.3
3
4.4
8
11
15
Time
Beta(t) for weight
−0.1
0.1
1.3
3
4.4
8
11
15
Time
Beta(t) for height
−2
−1
0
1
1.3
3
4.4
8
11
15
Time



Interpretation of Plots:.

`start_age`: The plot shows noticeable fluctuations in beta(t), especially in early time periods. This suggests the effect of age at career start is not constant over time.

`start_year1`: The slope deviates from zero, indicating that the effect of debut year changes during a player's career.

`position_primary`: The beta(t) line varies and crosses over time, confirming that the effect of player position violates the PH assumption.

`weight`: Although significant in the test, the plot shows some mild deviations from flatness; this may be due to the large sample size amplifying small effects.

`height`: The beta(t) line is relatively flat, suggesting that height has a constant effect over time.

`start_year1:position_primary`: Shows mild deviations, but the test result (p = 0.08327) suggests no strong violation.

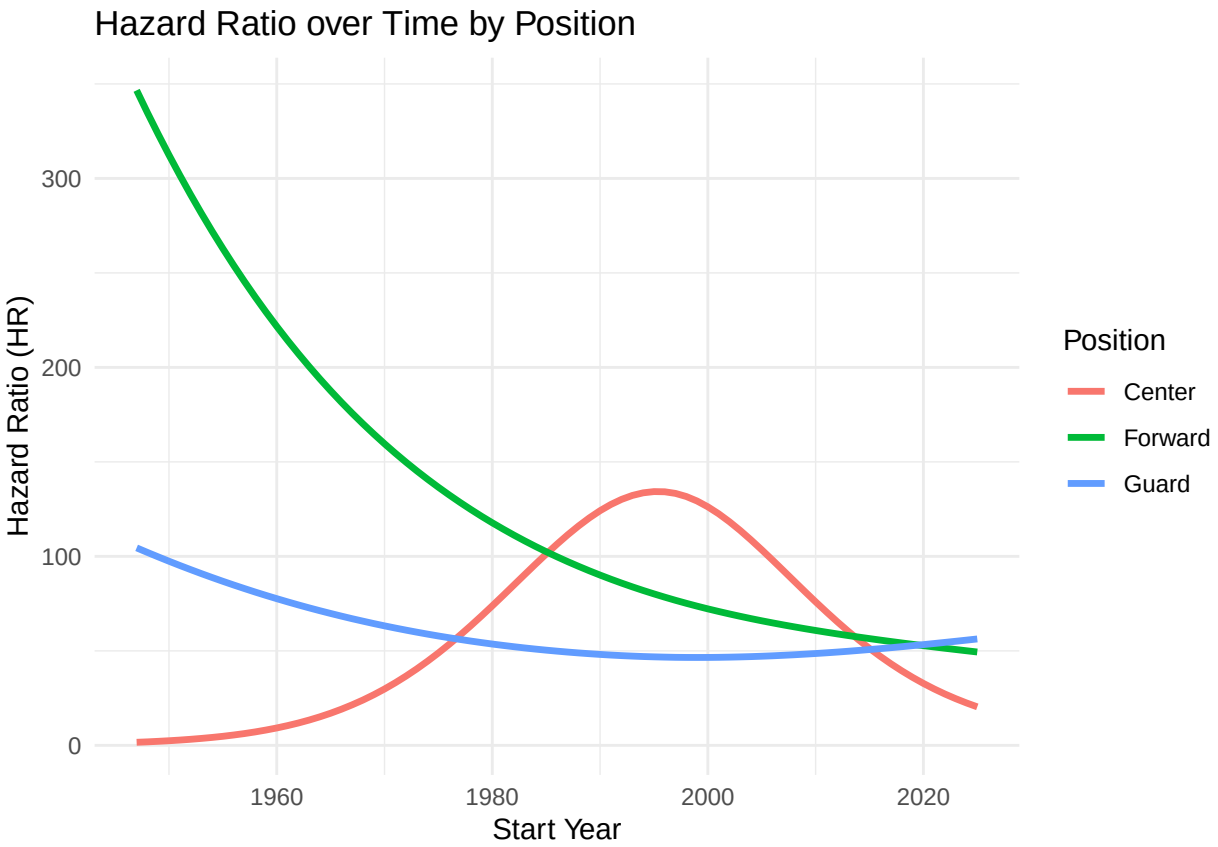
`position_primary:Episode`: The effect of position varies across episodes, which is expected given the interaction. The violation is due to meaningful changes in hazard over different career stages.

While the advanced model improves flexibility and incorporates time-varying effects via stratification and interactions, some covariates (especially ***`start_age`****,* ***`start_year1`****, and* ***`position_primary`****) continue to exhibit non-proportional behavior. Nevertheless, the stratified framework allows us to capture key dynamic effects and improves interpretability across career phases.*

Plot HR vs Start Year by Position

```
# Fit the model (no strata version)
model_no_strata <- coxph(
  Surv(tstart, career_length, status) ~
```

```
    start_age + ns(start_year1, df = 2) + weight + height +
    position_primary + ns(start_year1, df = 2):position_primary,
  data = basketball.split211
)

# Create prediction grid
pred_grid <- expand.grid(
  start_age = mean(basketball.split211$start_age, na.rm = TRUE),
  start_year1 = seq(min(basketball.split211$start_year1), max(basketball.split211$start_year1), by = 1),
  position_primary = c("Center", "Forward", "Guard"),
  weight = mean(basketball.split211$weight, na.rm = TRUE),
  height = mean(basketball.split211$height, na.rm = TRUE)
)

# Match factor levels
pred_grid$position_primary <- factor(pred_grid$position_primary,
                                     levels = levels(basketball.split211$position_primary))

# Predict linear predictor and compute HR
pred_grid$lp <- predict(model_no_strata, newdata = pred_grid, type = "lp")
pred_grid$HR <- exp(pred_grid$lp)
pred_grid$start_year <- pred_grid$start_year1 + 1947

# Plot
ggplot(pred_grid, aes(x = start_year, y = HR, color = position_primary)) +
  geom_line(size = 1.2) +
  labs(title = "Hazard Ratio over Time by Position",
       x = "Start Year", y = "Hazard Ratio (HR)", color = "Position") +
  theme_minimal()
```



This plot illustrates how the risk of career end changes over time for each position:
`Forwards` show the steepest drop in hazard, indicating greatly improved longevity since the 1970s.

`Guards` maintain a moderate but steady decline in hazard.

Centers show a temporary peak in risk around 2000, followed by recovery. Overall, the model captures meaningful non-linear trends in how position and start year interact to affect career duration.

5. Conclusion

```
summary(full_model)
```

```
## Call:
## coxph(formula = Surv(career_length, status) ~ start_age + start_year1 *
##     position_primary + weight + height, data = basketball_df)
##
##   n= 4602, number of events= 3952
##    (26 observations deleted due to missingness)
##
##                                        coef exp(coef)  se(coef)       z
## start_age                           0.189244  1.208335  0.008439  22.425
## start_year1                         0.012241  1.012316  0.002951   4.148
## position_primaryForward             2.119280  8.325138  0.175183  12.098
## position_primaryGuard               0.841431  2.319685  0.181783   4.629
## weight                              0.003816  1.003823  0.001252   3.047
## height                             -0.049654  0.951559  0.009536  -5.207
## start_year1:position_primaryForward -0.037281  0.963406  0.003266 -11.413
## start_year1:position_primaryGuard   -0.021618  0.978614  0.003173  -6.814
##                                    Pr(>|z|)
## start_age                           < 2e-16 ***
## start_year1                        3.35e-05 ***
## position_primaryForward             < 2e-16 ***
## position_primaryGuard              3.68e-06 ***
## weight                              0.00231 **
## height                             1.92e-07 ***
## start_year1:position_primaryForward  < 2e-16 ***
## start_year1:position_primaryGuard   9.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                    exp(coef) exp(-coef) lower .95 upper .95
## start_age                             1.2083     0.8276    1.1885    1.2285
## start_year1                           1.0123     0.9878    1.0065    1.0182
## position_primaryForward               8.3251     0.1201    5.9058   11.7357
## position_primaryGuard                 2.3197     0.4311    1.6244    3.3126
## weight                                1.0038     0.9962    1.0014    1.0063
## height                                0.9516     1.0509    0.9339    0.9695
## start_year1:position_primaryForward   0.9634     1.0380    0.9573    0.9696
## start_year1:position_primaryGuard     0.9786     1.0219    0.9725    0.9847
##
## Concordance= 0.684  (se = 0.005 )
## Likelihood ratio test= 1230  on 8 df,   p=<2e-16
## Wald test            = 1421  on 8 df,   p=<2e-16
## Score (logrank) test = 1427  on 8 df,   p=<2e-16
```

Key Findings from the Final Cox Model:
The final model included start_age, start_year1, position_primary, height, and weight, along with interaction terms between start_year1 and position_primary.

`Start Age` (HR = 1.2083, 95% CI: [1.1885, 1.2285]).
- Each additional year in starting age increases the hazard of retirement by approximately 21%.

- This confirms that players who debut later tend to have shorter careers.
`Height` (HR = 0.9516, 95% CI: [0.9339, 0.9695]).
- Taller players have a lower risk of retirement, implying longer career spans.
`Weight` (HR = 1.0038, 95% CI: [1.0014, 1.0063])
- Heavier players show a slightly increased hazard of career end.

Position Effects (compared to Centers):
`Forwards`: HR = 8.3251, 95% CI: [5.9058, 11.7357].
- Substantially higher risk of retirement, indicating shorter careers.
`Guards`: HR = 2.3197, 95% CI: [1.6244, 3.3126]
- Also significantly shorter careers than Centers.

Time Trend Effects:.
`start_year1` (HR = 1.0123, 95% CI: [1.0065, 1.0182])
- Players starting in more recent years have slightly shorter careers.

Interaction Terms:.
`start_year1 * Forward`: HR = 0.9634, 95% CI: [0.9573, 0.9696]
`start_year1 * Guard`: HR = 0.9786, 95% CI: [0.9725, 0.9847].
- These interactions suggest that the career disadvantage for Guards and Forwards has decreased modestly in recent decades.

Model Performance:.
Concordance = 0.684 (SE = 0.005).
- Indicates moderate predictive accuracy of the model.
Likelihood Ratio Test = 1230 on 8 df, p < 2e-16.
Wald Test = 1421 on 8 df, p < 2e-16.
Score Test = 1427 on 8 df, p < 2e-16
- These metrics confirm strong overall model fit.

The Cox model shows that `age`, `position`, and `start_year` strongly impact NBA career length. Forwards and Guards face much higher risks of early exit, though recent players show improving trends. Stratified models and splines better capture non-proportional effects. These findings support tailored career planning by position and debut timing.

6. References

Parks, Kevin (2021). *Athlete Career Length Dataset.* Kaggle. http://www.kaggle.com/datasets/kevinparks/athlete-career-length

"Basketball Reference" (2025). *Basketball Statistics & History of every Team & NBA and WNBA players.* https://www.basketball-reference.com/