Survival Analysis of NFL Running Backs

John Randazzo, Kevin Wang, Brian Luu
UCSB

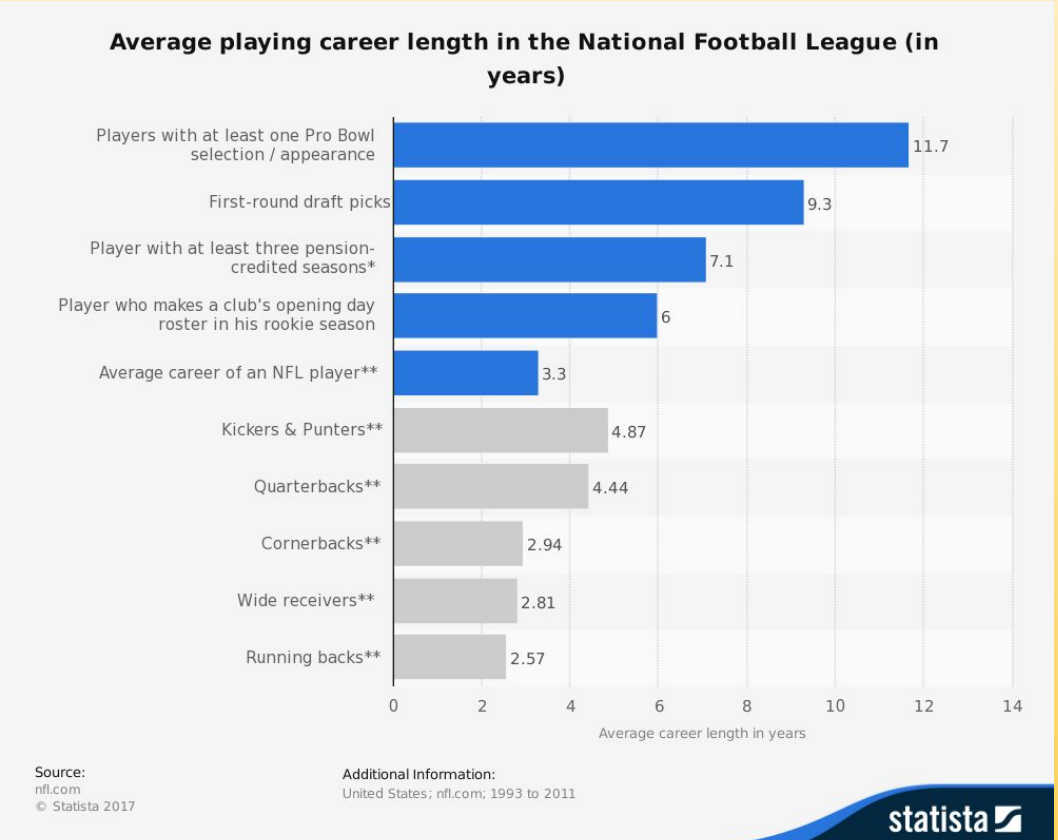
The nature of this project

- Goal: Identify significant predictors (such as career averages, accolades, and physical measurements) of career longevity in NFL running backs using survival analysis methods
- We use methods in Python to obtain and process our data, and perform our analysis and create data visualizations with R
- Why is this of interest?

Motivation - RBs don't last very long in the NFL

Motivation - RBs don't last very long in the NFL

Motivation - RBs don't last very long in the NFL



Data Wrangling and Processing

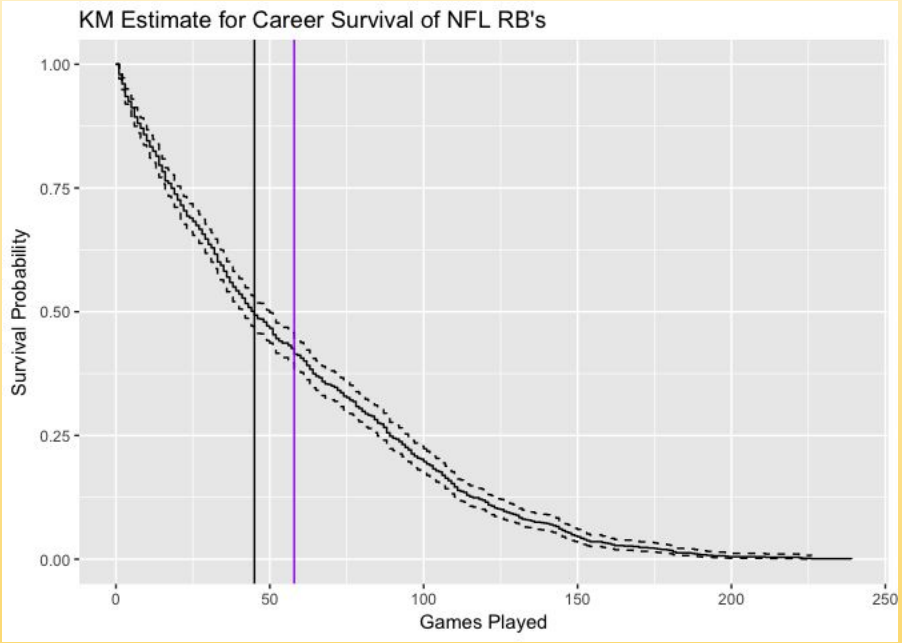
- Extracted career statistics of 1500 RBs drafted into NFL between 1976 - 2016 off of pro-football-reference.com
- We figured that physical measures play a role in a RB's career length, so we used Python's Beautiful Soup to parse each player's unique pro-football-reference.com web page and extract their height and weight
- Post-processing, we are left with 1037 RBs
- Full details available on Github

A Primer on Survival Analysis

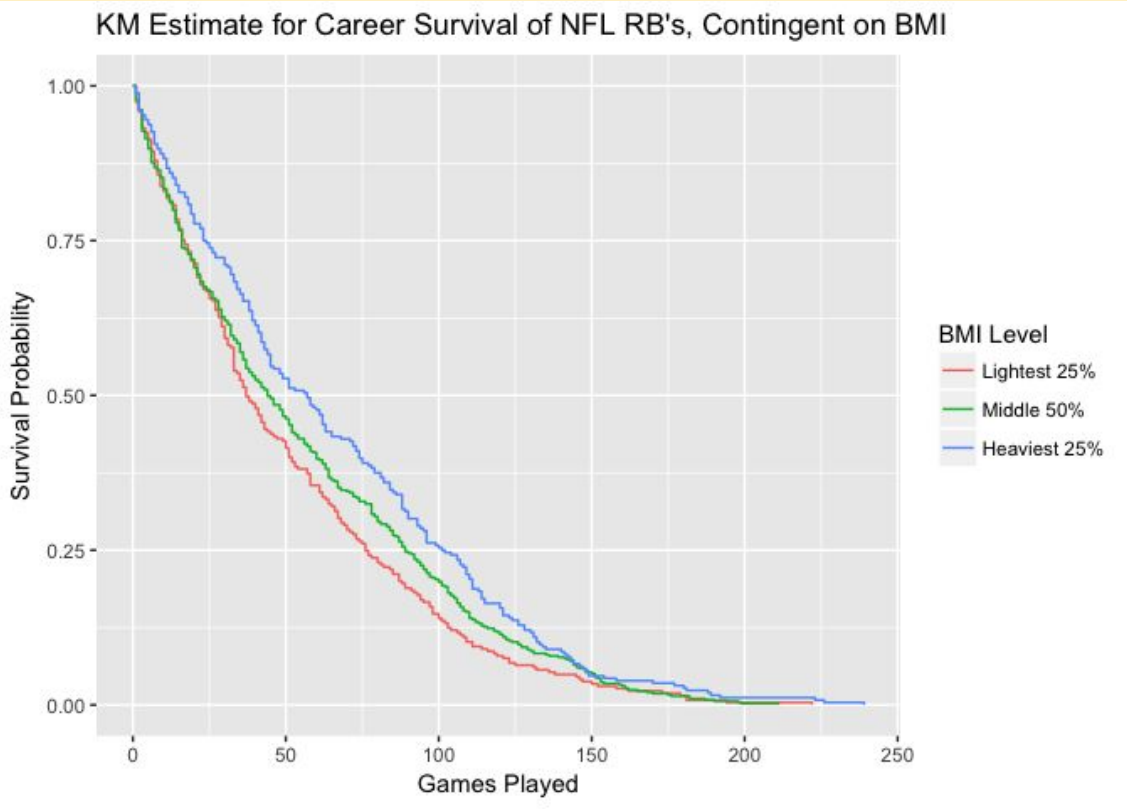
- Survival analysis concerns itself with time until a particular event of interest
- Here, the event is retirement from the NFL, but in most applications, we are waiting for something to die…. More appropriate name is “Death Analysis”
- Two important variables:
  - Time until event or censoring (in our study, the amount of games played during career)
  - Binary variable indicating whether subject experienced event
    - 0 : censored (left the study or did not experience event during study)
    - 1 : subject experienced event of interest

Kaplan-Meier Curves

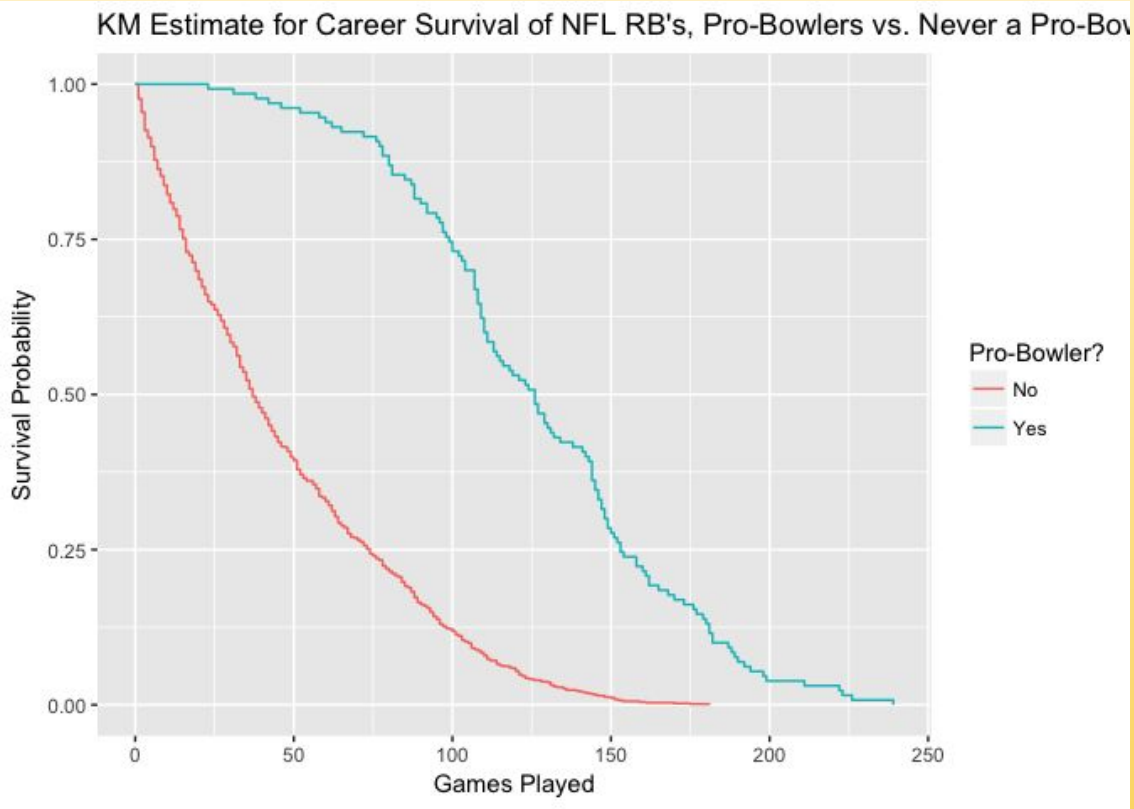
- Estimates survival function at a given time (number of games) using the ratio of players that retired at that time compared to the amount that had not retired by then
- Purple: Mean career length (~53 games)
- Black: Median career length (~47 games)



Kaplan-Meier Curves



Kaplan-Meier Curves



The Cox Proportional Hazards Model

- The KM estimator is limited because it only considers a single homogenous risk group at a time
- The Cox PH model is the “industry standard” of survival models
- The hazard function given by the Cox PH model takes the form:

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = \lambda_0(t)\exp(X_i \cdot \beta).$$

- The above expression gives the hazard function at time *t* for subject *i* with covariate vector (explanatory variables) X_*i*.
- This allows us to evaluate different covariates’ effects on career longevity.

Our Cox Proportional Hazards Model

Covariate	Coefficient Value	e^(Coefficient)	P-value
BMI	-0.07703	0.92586	8.71 * 10^(-8)
YPC (Yards / Carry)	-0.20421	0.81529	8.02 * 10^(-12)
Draft Age	0.17489	1.19111	5.52 * 10^(-5)

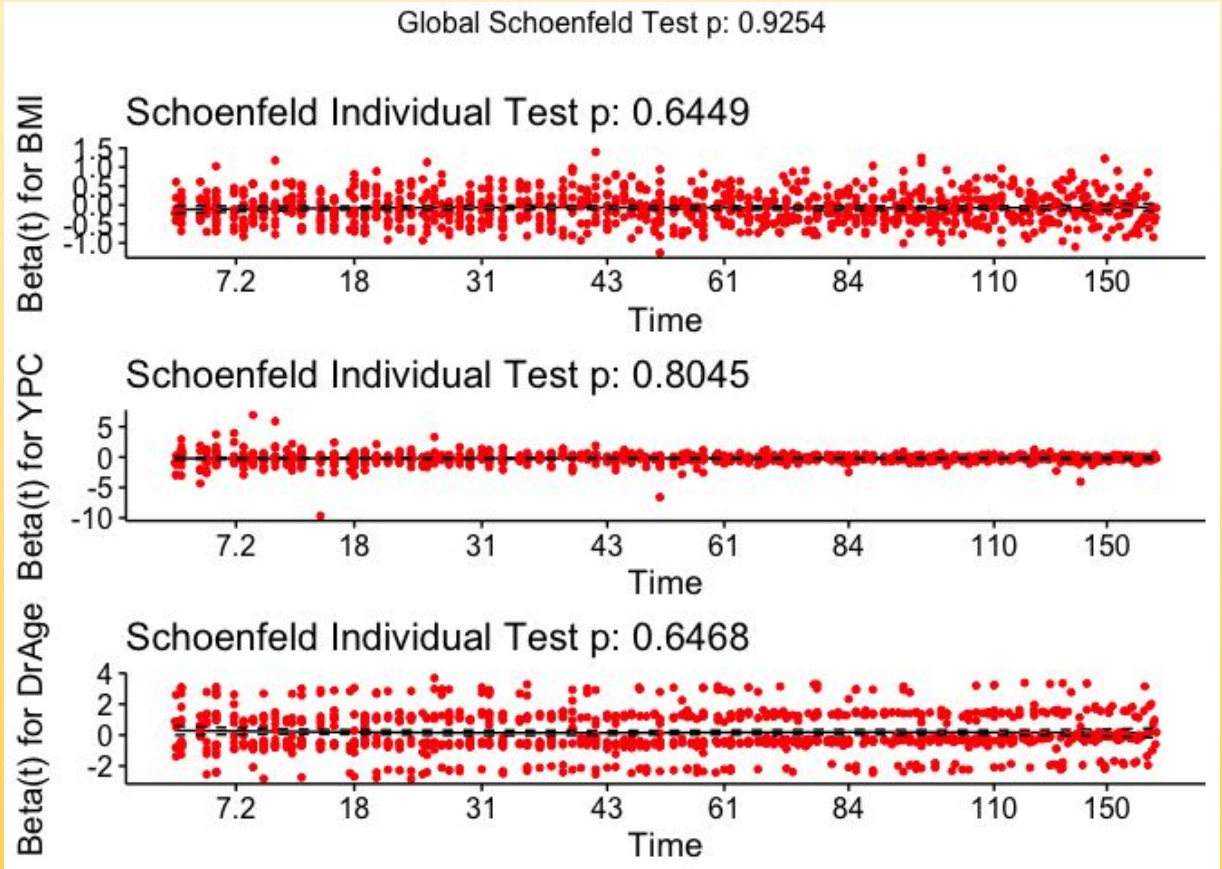
$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = \lambda_0(t)\exp(X_i \cdot \beta).$$

- Interpretation: A player with an additional unit of X is e^(Coefficient)*100 % as likely at any given time to retire as a player without the additional unit of X, all other measures held equal

The Proportional Hazards Assumption

- The key assumption in the Cox PH model is that of Proportional Hazards
- In particular, we assume that covariate effects are fixed throughout the study, ie that the effects of the predictors are TIME-INDEPENDENT
- This is a very tenuous assumption in most settings, as most variables of interest exhibit some sort of time-dependence.
- Never trust a Cox PH model in a paper that does not at least acknowledge the existence of this assumption!!!!!!!!!!

Testing Our Model for Proportional Hazards

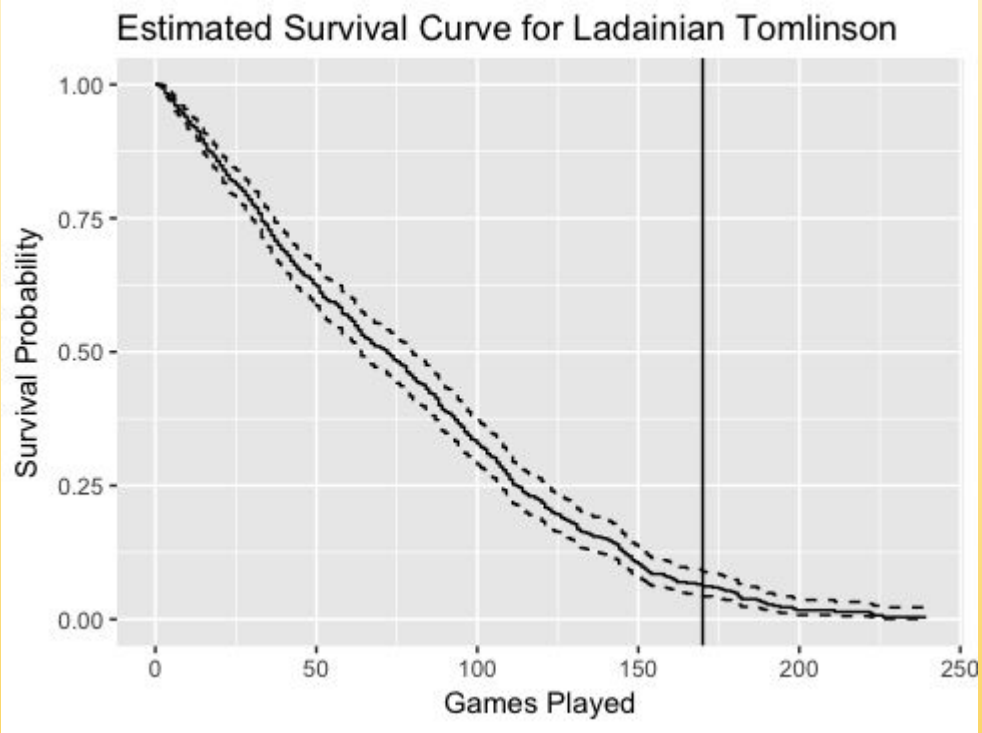


Wow!

- In my experience with survival analysis, I have never seen a model's predictors behave as nicely as this one with regards to the PH assumption
- Usually we have to either throw out the most significant predictors altogether because they are fundamentally incompatible with the assumption, or we have to find a clever workaround to model them… but not this time!
- Now that we have confirmed our model meets the assumption, we can visualize the effects of the covariates on estimated career length

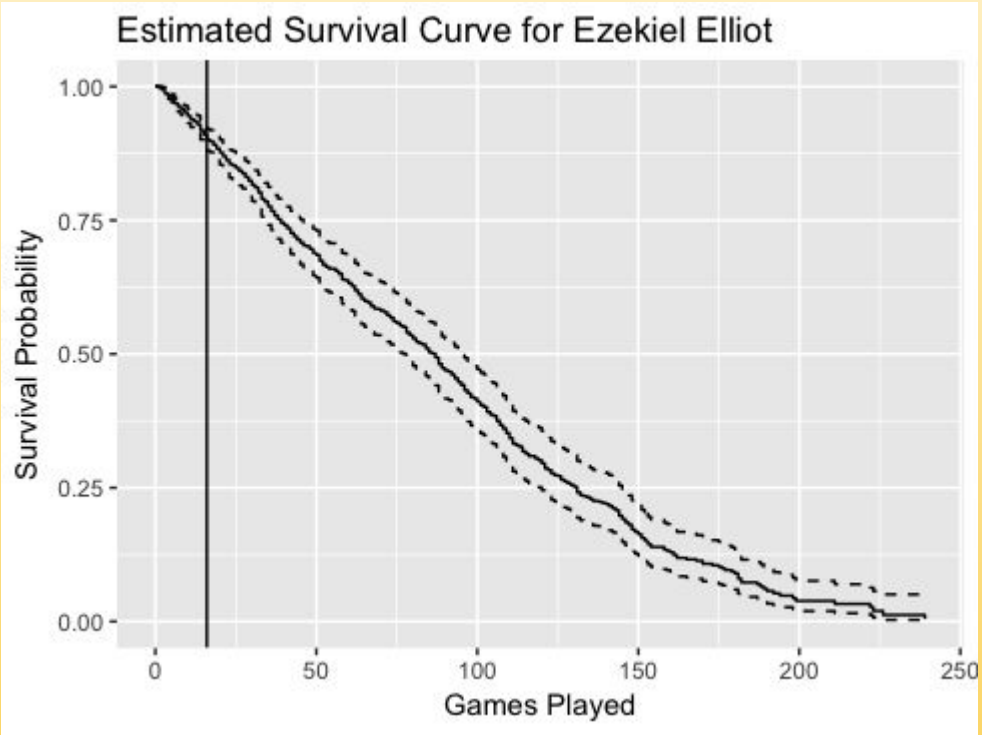
Fun With Our Model: A Tale of 3 RBs

- We can use our model to estimate real-life career survival probabilities given a player's measures for BMI, Draft Age, and YPC
- Here is a survival curve as determined by the Cox model for SD Chargers legend Ladainian Tomlinson
- Black line denotes actual number of games played



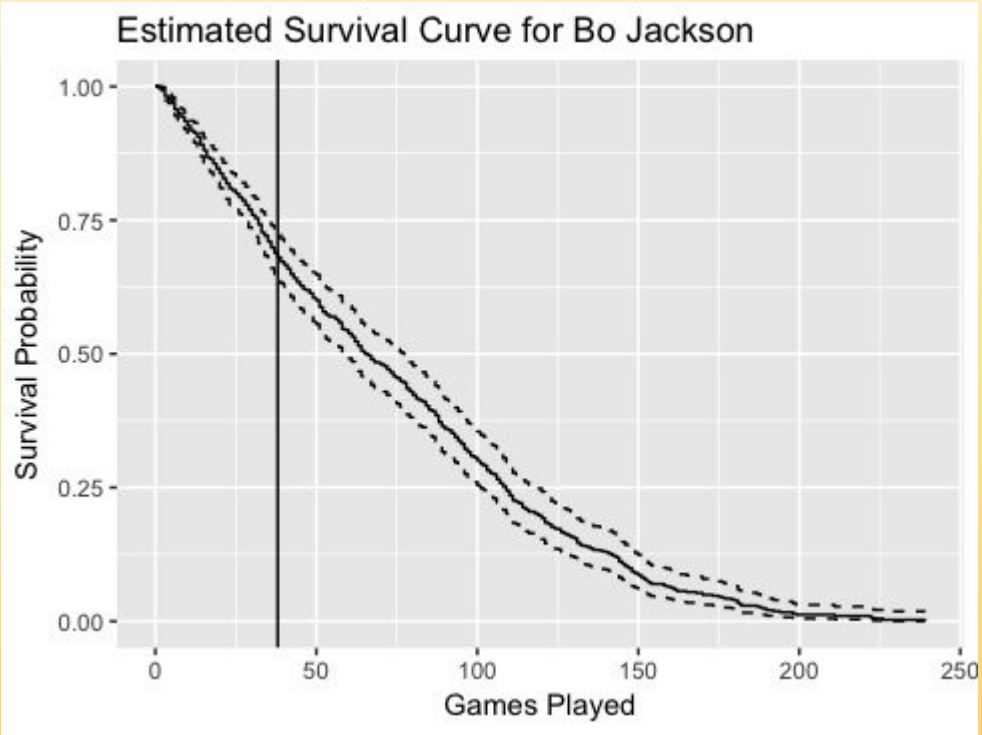
Fun With Our Model: A Tale of 3 RBs

- The most electrifying RB of the 2016-17 season, Ezekiel Elliott of the Dallas Cowboys
- Only played 1 season (16 games) when this project was completed
- Hopefully he can keep himself out of trouble and have a lengthy career



Fun With Our Model: A Tale of 3 RBs

- Bo Jackson, arguably the greatest pure athlete in American history
- Only lasted 3 seasons but also moonlighted as a star MLB player
- Career tragically ended by a catastrophic hip injury, which may have been catalyzed by his (alleged) steroid use



Conclusions

- BMI and YPC express negative effects on our hazard function as determined by our Cox model, so a denser player with higher YPC is expected to have a longer career on average
- Draft Age has a positive effect on career hazard, ie a player that is drafted at a more advanced age is less likely to have a long stint as an NFL player
- All three covariates align beautifully with the PH assumption for the Cox model
- Interesting covariates for future work: salary data, Wonderlic test score, concussions/game

Thank you for listening!

Github: johnrandazzo/surv_nflrb

Follow me on twitter: @johnrandazz0