Carnegie Mellon University

NBA Project - Progress Report 2

Team: Andrew Liu, Willis Lu, Reed Peterson
Advisor: Brian MacDonald
Client: Kostas Pelechrinis

Introduction - Team

Andrew Liu

Willis Lu

Reed Peterson

Carnegie Mellon University

Introduction - Faculty Advisor

- Special Faculty at Carnegie Mellon's Department of Statistics
- Has experience working with ESPN on NBA analytics.

Carnegie Mellon University

Introduction - Client

- Client: Kostas Pelechrinis
- Associate Professor at University of Pittsburgh
- Presenter at MIT Sloan Sports Analytics Conference
- Very high technical knowledge

Project Overview

Is there a way to accurately measure an NBA player's performance? There are so many ways to statistically measure a player. One such method is through box score +/-. Our goal is to use additional data such as contract value, team rating, and player history, to calculate +/- posteriors for a player. This research has a wide variety of applications, such as within the gambling industry and for sports fans in general.

1. Can we predict how good a player is using +/-?
2. Can our +/- predictions account for team strength?
3. Additional questions can come up based on any interesting findings. For example, we might want to see if coaching impacts +/-.

Primary Datasets

NBA Data

- Games data from 538
- NBA Play by play data
  - Original data from https://eighthirtyfour.com/data

Priors

- Contract data
- Team ratings
- +/- per 100 possessions

Carnegie Mellon University

Contract Data

Scraped using Python and BeautifulSoup Package. Original source: spotrak.com.

Contains information such as team, player name, position, year of contract, age, contract value, and type of contract. Data only available

Additional data found on Kaggle, but it does not include some key information (i.e. type of contract).

Team	Year	Name	Age	Pos	Contract Value	Type
Atlanta Hawks	2018	Kent Bazemore	29	SG	$18,089,887	Cap Space
Atlanta Hawks	2018	Miles Plumlee	30	C	$12,500,000	Bird
Atlanta Hawks	2018	Dewayne Dedmon	29	C	$7,200,000	Cap Space
Atlanta Hawks	2018	Trae Young	20	PG	$5,356,440	Rookie
Atlanta Hawks	2018	Alex Len	25	C	$4,350,000	Room
Atlanta Hawks	2018	Taurean Prince	24	SF	$2,526,840	Rookie
Atlanta Hawks	2018	Justin Anderson	24	SG	$2,516,048	Rookie
Atlanta Hawks	2018	John Collins	21	PF	$2,299,080	Rookie

Carnegie Mellon University

Games Data

Our games data comes from 538's study on NBA Elo rankings.

https://github.com/fivethirtyeight/data/tree/master/nba-forecasts

This dataset contains game by game elo ratings all the way back to the 1946 NBA Season. The only variables we will be using are the game scores.

	date	season	neutral	playoff	team1	team2	elo1_pre	elo2_pre	elo_prob1	elo_prob2	elo1_post	elo2_post	carm.elo1_pre	carm.elo2_pre	carm.elo_prob1
1	2017-10-17	2018	0		CLE	BOS	1647.990	1532.470	0.7756739	0.2243261	1650.129	1530.331	1648	1549	0.7603643
2	2017-10-17	2018	0		GSW	HOU	1760.610	1574.467	0.8385078	0.1614922	1751.819	1583.258	1761	1675	0.7474946
3	2017-10-18	2018	0		WAS	PHI	1565.684	1379.576	0.8384814	0.1615186	1567.534	1377.726	1549	1478	0.7203165
4	2017-10-18	2018	0		ORL	MIA	1390.229	1552.810	0.4109011	0.5890989	1400.664	1542.375	1458	1483	0.5986336
5	2017-10-18	2018	0		IND	BRK	1502.885	1405.034	0.7574814	0.2425186	1506.961	1400.958	1406	1381	0.6756820
6	2017-10-18	2018	0		DET	CHO	1456.655	1473.216	0.6178213	0.3821787	1464.993	1464.879	1427	1542	0.4844012

	carm.elo_prob2	carm.elo1_post	carm.elo2_post	raptor1_pre	raptor2_pre	raptor_prob1	raptor_prob2	score1	score2
1	0.2396357	1650.309	1546.691	NA	NA	NA	NA	102	99
2	0.2525054	1753.884	1682.116	NA	NA	NA	NA	121	122
3	0.2796835	1552.479	1474.521	NA	NA	NA	NA	120	115
4	0.4013664	1464.398	1476.602	NA	NA	NA	NA	116	109
5	0.3243180	1411.729	1375.271	NA	NA	NA	NA	140	131
6	0.5155988	1439.104	1529.896	NA	NA	NA	NA	102	90

Team Ratings

We utilize a simple linear regression that uses game data and tries to predict point differential given the variables team and home court advantage. This creates a coefficient for each team that we use as player ratings.

-Dimensions: 30 observations and 2 variables.

-These team ratings will be used in our Bayesian Regression. We will likely add additional features to our team rating model.

	team	rating
17	MIL	16.3807222
14	LAL	14.6040696
13	LAC	14.5170573
2	BOS	14.2242103
28	TOR	13.8739727
7	DAL	12.0453683
16	MIA	11.5500537
11	HOU	11.0840935
29	UTA	10.6306580
8	DEN	10.2444927
23	PHI	9.6917090
21	OKC	9.5921958

Carnegie Mellon University

Shifts Data

A “shift” is a period of time in an NBA game where the same 10 players are on the court with no substitutions

We reformatted play-by-play data from eightthirtyfour[1] as shift data to track the +/- of each shift

Shifts are normalized by recording +/- per 100 possessions, where the number of possessions in each shift is calculated from this common formula: https://www.nbastuffer.com/analytics101/possession/

1. https://eightthirtyfour.com/data

Carnegie Mellon University

Methods

- Simple Linear Regression
    - Used to acquire initial team ratings
- Bayesian Regression
    - Using team ratings as our prior, project point differential for individual games
        - *Work in progress*
    - Toy example with arbitrarily chosen prior
        - *Goal is to familiarize ourselves with Bayesian regression and the PyMC3 package*

Linear Regression

We use simple linear regression to create team rating for priors. We regress point differential on two variables (team and location).

In comparison, 538 has created an elo system that we could potentially use as priors as well. This elo system is calculated much the same way as our team ratings. Except, they include margin of victory, a k-factor (tuned to determine how quickly each game will affect elo), home court, and year-to-year carry-over.

Applications

-to be used in our Bayesian regressions

-can tell us how good teams are in the regular season

-will allow us to adjust player ratings in accordance to their team ratings.

Bayesian Regression

Bayesian Regression will be implemented in Python with the pymc3 package.

Our data is structured such that the dependent variable is +/- per 100 possessions, and the predictors are each NBA player in the league for a given season

- So our design matrix X is NxM dimensional, where N is the number of shifts in a season and M is the number of players in the league.
- For each shift, there will be five 1's, five -1's and M-10 zeros, where the 1's correspond to the five players on the home team during a given shift and the -1's correspond to the five players on the away team during a given shift

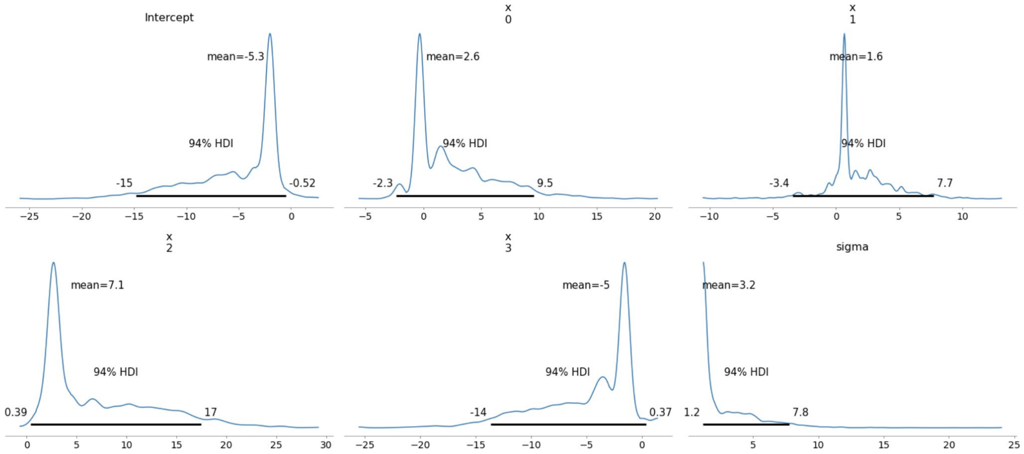
Prior distributions for each covariate can be specified (in this case, each covariate is a player), and the result is a distribution for the coefficient for each player, representing the distribution of their +/-

Carnegie Mellon University

Results

We have contract priors and team ratings for the 2018-2019 NBA Season.

We have a general framework for implementing Bayesian regression, and our skeleton model has run successfully.



Challenges/Next Steps

-Compare results from Bayesian regression to results from ridge regression and standard linear regression

-Does player changing teams change their ratings?

-Can we include coaching in the player rating?

-Does resting players increase the player's performance?

-Raptor ratings from 538 as a prior. Already in per 100 possessions.

-Predict team offensive/defensive ratings at end of season. Compare this to weighted average player rating based on minutes played to overall offensive/defensive ratings.

Thank You!

Q & A