Star Power

RJ Cass

Introduction

'Star Power' is the idea that a certain person, due to having something along the lines of 'celebrity status', can influence how many people watch a given event. In this case, we care about trying to capture the effect of Star Power on NBA All-star game viewership by looking at a variety of explanatory variables and seeing how they affect the number of people watching the game.

Data

For this analysis we are using the NBA TV Audience dataset. This dataset consist of data aggregated per NBA market (more or less the geographical region), and year. The data are aggregated by NBA Market Region (basically, areas where the population receives similar mdeia programming) by year. The list of Market Regions are provided in Table 1. For reach region, we have yearly metrics about the region, the players in that region, and the number of people who watched the All-star game that year from that region. A summary of the regional-aggregated data is provided in Table 2 (the summary stats for local/secondary are generated ignoring 0s which indicate the region did not have a team in that category). A summary of the year-aggregated data is provided in Table 3.

Table 1: *List of all NBA Market Regions*

Region
Albuquerque.Santa.Fe
West.Palm.Beach.Ft..Pierce
Norfolk.Portsmth.Newpt.Nws
Charlotte
Detroit
Greenvll.Spart.Ashevll.And
Sacramnto.Stkton.Modesto
Memphis
Dayton
Houston
Philadelphia
Raleigh.Durham..Fayetvlle.

(continued)

Region
Kansas.City
Washington..DC..Hagrstwn.
Denver
Pittsburgh
New.York
Portland..OR
Louisville
Phoenix..Prescott.
Baltimore
San.Francisco.Oak.San.Jose
Chicago
Salt.Lake.City
Atlanta
Los.Angeles
Ft..Myers.Naples
Milwaukee
Knoxville
San.Antonio
Birmingham..Ann.and.Tusc.
Jacksonville
Seattle.Tacoma
Greensboro.H.Point.W.Salem
Austin
Cincinnati
Orlando.Daytona.Bch.Melbrn
Providence.New.Bedford
Dallas.Ft..Worth
Nashville
Buffalo
Oklahoma.City
San.Diego
Las.Vegas
Indianapolis
Cleveland.Akron..Canton.
Richmond.Petersburg
New.Orleans
Hartford...New.Haven
Boston..Manchester.
Minneapolis.St..Paul
St..Louis
Tampa.St..Pete..Sarasota.
Miami.Ft..Lauderdale

(continued)
Region
Columbus..OH
Tulsa

Table 2: *Summary Statistics of the Market Region-based metrics*

Variable	Mean	SD	Min	Max
Year	2010.52	4.61	2003	2018
all.tvs	1411414.55	1198681.09	413730	7515330
aud	76983.92	90290.16	4426	659654
local.atbreak	0.51	0.15	0.125	0.923
local	1.52	0.74	1	4
local.start	1.21	0.52	1	4
local.host	1	0	1	1
secondary.atbreak	0.54	0.16	0.125	0.849
secondary	1.83	0.99	1	6
secondary.start	1.23	0.58	1	4
secondary.host	1	0	1	1

Table 3: *Summary Statistics of the Year-based metrics*

Variable	Mean	SD	Min	Max
MaxPER	29.93	1.34	27.3	31.7
PER10best	22.83	0.95	21.65	24.85

As shown in Table 2, we have data ranging from the year 2003 to 2018. We have data on the number of TVs within a region, as well as the total estimated viewership of the NBA All-star game. This viewership is the 'Nielsen TV Audience' number, meaning there is a portion of the population that has devices on their TV to measure what they watch, and that sample is used to estimate the total viewing audience. There is a fairly large spread of TV audience across the regions, due to some regions having vastly larger populations than others (ie. New York vs. a more rural region).

At the region level, the metrics are split by whether the region has a local NBA team, or whether they are considered a second market (ie. a neighboring reigon has a team). For example, Cleveland, OH has the NBA team Cleveland Cavaliers, so that region has a local team. Cincinnati has no NBA team, but their secondary team is also Cavaliers. For each region we have a record of whether or not they have a local team (local) or a secondary team (secondary). We have indications of what the win percentage of those teams were over the season (local.atbreak, secondary.atbreak), whether or not those team hosted the All-star game (local.host, secondary.host), and whether they had a player in the starting lineup for the All-star game (local.start, secondary.start).

Finally, the data contain metrics on the All-star players for that year. These metrics are aggregated on the 'year' level, not the market level as they measure the values for the All-star players that year. Specifically, it contains metrics involving the PER (Player Efficieny Rating). PER is a measure of a players average performance over a season. PER is calculated via: (Positive Actions - Negative Actions)/minutes_played. A value of 15 is the league average, and all individual values are normalizdd to that value (so a value over 15 indicates better than the league average, a number less than 15 is under the average). The dataset contains the MaxPER for that year (so the highest PER achieved by an All-star player that year) and the PER10best (the PER of the 10th best player, or the 10th best PER of the All-stars that year).

Model

Using the available data, we constructed a linear model represented by the following equation:

$$\begin{aligned} ln(Y_{it}) =& \beta_{0i} + \beta_1 ln(Y_{i,t-1}) + \beta_2 ln(Y_{i,t-2}) + \beta_3 maxPER_t + \beta_4 10thBestPER_t \\ & + \beta_5 Z_i AllStars_{it} + \beta_6 Z_i Starters_{it} + \beta_7 Z_i Team_{it} + \beta_8 Z_i Host_{it} + \beta_9 (1-Z_i) AllStars_{it} \\ & + \beta_{10}(1-Z_i) Starters_{it} + \beta_{11}(1-Z_i) Team_{it} + \beta_{12}(1-Z_i) Host_{it} + \eta_i + \nu_{it} \end{aligned}$$

In this model we are assuming that the error terms (η_i and ν_{it}) are normally distributed, that each variable has a linear relationship to the response variable (again, the natural log of viwership in this case). We also assume that each variable is independent of the others, and that there is equal variance across all factors.

To begin understanding the model, we first need to identify some of the notation. The subscripts refer to the market (subscript i) and the year (subscript t). So, $Variable_{it}$ refers to the given variable for $Market_i$ in $Year_t$. In this model, Y_{it} is our response variable (viewing audience size). Note that the model relates the natural log of Y_{it} with the meaning that given a model prediction for a specific team/year, to get the actual viewing audience size the predicted value would need to be back-transformed into regular 'viewers' units.

Each term in this model has a coefficient (the β_x). This coefficient is the 'multiplier' for that spcific variable, meaning that, holding all other variables constant, if that variable were to increase by 1 unit, the response ($ln(Y_{it})$) would increase by β_x. For example, for the term $\beta_3 maxPER_t$, if we were to keep all the other variables constant (ie. same region, same host, same All-stars, etc.) and raised the MaxPER for that year by 1, the $ln(Y_{it})$ would increase by β_3. It's important to note that in this model we are assuming that none of the different variables are correlated (that none of them affect any of the others). For variables such as the PER metrics, that assumption is a little more hazy.

The model includes two terms for the viewership of the previous years $\beta_1 ln(Y_{i,t-1}) + \beta_2 ln(Y_{i,t-2})$. This means that the viewership in a market in a given year is dependent on the natural log of the viewership of the previous 2 years (ie. viewership for a market in 2013 is dependent on the viewershipo numbers from 2011 and 2012).

The other terms in the model are as described previously, with $maxPER_t$ being the highest PER of the All-stars in a given year and $10thBestPer_t$ being the 10th best PER of the All-stars in a given year. These variables change only per year, and not market.

The $AllStars_{it}$, $Starters_{it}$, $Team_{it}$ and $Host_{it}$ terms are the market level metrics, indicating whether for a given year their team had any All-stars in the game ($AllStars_{it}$), whether any of them were in the starting lineup ($Starters_{it}$), what their win percentage is ($Team_{it}$), and whether that market hosted it ($Host_{it}$). The Z_i and $(1 - Z_i)$ terms are just indicators of whether it refers to the local team (Z_i) or the secondary team $(1-Z_i)$. As noted previously, there are some markets that have multiple primary/secondary teams. In these cases, the model is built using only the top-rated team in each category.

The final terms in the model are the 'uncertainty' terms. We have the term η_i which indicates the randomness/variability within a given market. The ν_{it} term refers to the variability within a given market in a given year.

Other Considerations

As mentioned previously, there are some interesting tidbits to consider in this data. For example, the dataset starts in 2003, which also happened to be the final year of Michael Jordan played in the NBA. This is something that likely had an impact on viewership but is not captured in the data. There have been other significant events in history which are likely to have an impact: though not including in this dataset, things such as the COVID-19 pandemic likely had a large impact on viewership and the model currently does not include any way to indicate such events (besides assuming it to be random uncertainty).