A Team-Focused Analysis of the 2018-19 NBA Season

Yixuan (Sherry) Wu, Adam Ginsburg, and Brandi Ginn

December 8, 2019

Contents

1 Abstract

Through a variety of media and technologies, aspects of popular American culture have had a major impact on different nations across the world. With regard to entertainment, controversy, and displays of athletic pageantry, there may not be a sport as universally captivating as basketball.

In our analysis, we tackled two major aspects of the NBA: average per game attendance, and making the playoffs. For our analysis of attendance, we investigated basketball-related and socio-economic factors of the cities in which teams play - and found that the number of Facebook fans a team has, the number of All-Stars on a team, the overall transit score (a composite index score that takes into account the accessibility of a city via transit, the quality of transit in a city, the percentage of jobs in a metropolitan area accessible by transit), and the number of transit routes in a city are all significantly associated with average attendance per game.

For our analysis of making the playoffs, we delved into the data to see what factors affect it. As it turns out, in our final model, neither of the two predictors, combined player impact estimate (also termed PIE, and is a composite statistic that takes into account each player's statistical impact on the game) and contested 2-point shots, were significantly associated with making the playoffs.

2 Introduction

From player personalities to off-court drama to the incredible skill shown nearly every night by some of the world's most impressive athletes, the popularity of American basketball–and the National Basketball Association in particular–has skyrocketed over the past 30 years. In that same time frame, the wave of mathematical analysis crashed over sports, as a result of the "Moneyball revolution".

For our research, we want to marry the growing popularity of American basketball with the statistical revolution that has taken over sports. We are interested in investigating the relationships between a team's statistics and their achieved success in a single season. For the purpose of this project, we are measuring success in two different ways: success as it relates to making the playoffs and success as it relates to average attendance in the 2018-2019 season.

As basketball viewership has increased in the past few years (Sprung, 2019), we want to understand more about the NBA fan base as well as the unique conditions under which NBA fans attend the games. Thus, it is important that we capture in our research the socio-economic and geographic data that impacts the lives of the growing population of basketball spectators. We are including socio-economic and geographic data in our analysis to contextualize research and provide a holistic understanding of factors that impact a team's success. Our questions are as follows: What factors determine whether a team can make it to the playoffs? And, what factors, including basketball-related statistics and socio-economic statistics of the major metropolitan cities that house NBA teams, most affect average attendance over the course of a season? These questions will be answered using a multivariate logistic regression and a multivariate linear regression. To report our findings, our methods and analysis will be divided into two sections: Take Me Out to the (Basket)Ball Game and Making it to the Playoffs.

One variable of interest to the present research is the number of wins a team can have in a given season. Other statisticians have also conducted research looking at wins. For example, Yang (2015) looks at how player's impact estimate (PIE) affects the wins ratio over time. Kotecki (2014) analyzes the effect of home court advantages on wins of a team. However, as hundreds of

variables capture different aspects of a team's performance, some combinations of variables have not yet been tested. Therefore, more research is needed to analyze what can contribute to a team's success in making the playoffs – and that is one of the questions we will be exploring in this paper.

3 Methods

3.1 Introduction of Data

For the model that associates average attendance per game with basketball-related and sociological factors, we aggregated data from NBA.com (for basketball-related statistics), from Basketball Reference (for attendance statistics), from Statista (for NBA teams' expenses on salaries), from Ticket IQ (for average midseason ticket price), from Trackalytics (for Facebook fans per team), from All Transit (for the composite index score and the number of major routes per city), and the U.S. census bureau (For unemployment, poverty, and median income in American cities) and the Ontario census bureau (for unemployment, poverty, and median income in Toronto).

3.2 "Take Me Out to the (Basket)Ball Game"

3.2.1 Dependent Variable

As discussed above, one of the major goals of this project was to delve into major factors—both basketball-related and sociological—affecting attendance per game in cities across the NBA. When first looking at this problem, a Poisson model, with attendance per game offset by arena capacity, seemed most natural to assess the attendance per game rate. However, after cursory analysis of potential Poisson models, it became clear that such a model would be unrealistic, as, in the Poisson models we fit, more than 66% of the 30 cities were influential points. Thus, analyzing such a Poisson model was untenable.

Despite initial disappointment at having to eschew the Poisson model, a linear model more

than suffices at analyzing the factors affecting average attendance of NBA games. It became clear that the data met the model assumptions of the linear model (as seen below) and, in fact, made interpreting the model remarkably straightforward and simple to digest.

Average attendance per game, computed by dividing total attendance at a team's home arena by its 41 home games, measures the attendance reported by each team after every game, which is comprised of the amount of tickets sold each game by each team. While the size of each NBA stadium differs slightly, the discrepancies are not very large; therefore, we were comfortable eschewing an offset and conducting a straightforward multiple linear regression model.

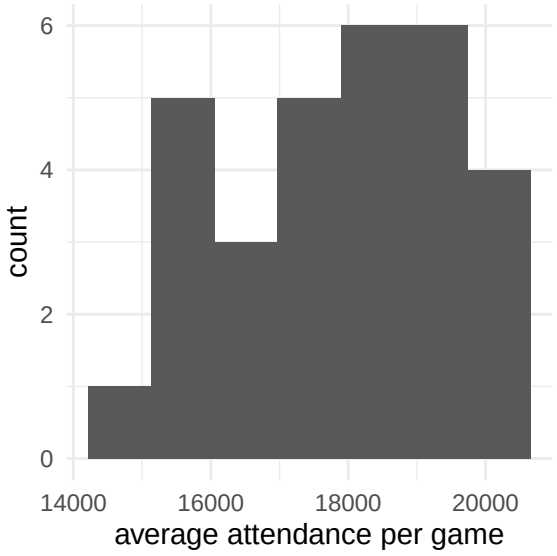


Figure 1: Histogram of average attendance per game

Average attendance per game is approximately normal, though skewed slightly left, as shown by the histogram in *Figure 1* and the box plot in *Figure 2*. It has a minimum average attendance per game of 14,941, a maximum attendance per game of 20,447, a median of 18,130 people, a mean of 17,852 people, and a standard deviation of 1641.732 people per game. We will delve more into the characteristics of the response variable in the model assumptions.

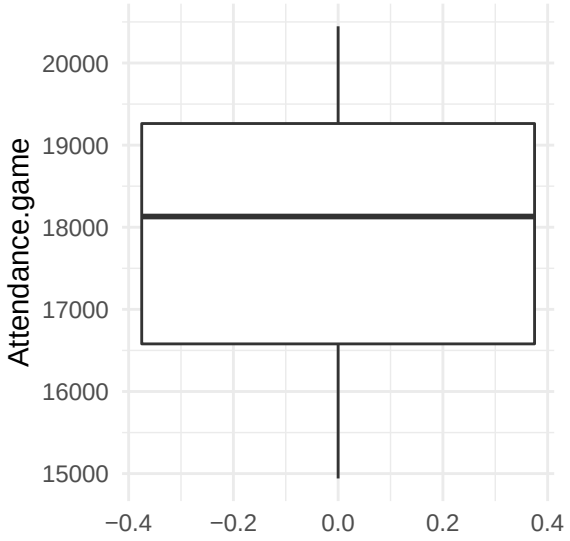


Figure 2: Boxplot of average attendance per game

3.2.2 Independent Variables

In analyzing potential models, we identified 14 initial covariates of interest that we intended to use to build our model. Below please find the 14 covariates, reasons for including them, and some important summary statistics about each potential covariate. For histograms of each potential covariate, please see figure 3,

Midseason secondary market ticket price, chosen because the ticket price ostensibly affects fan demand to attend games. It is approximately normal, though slightly right-skewed, with a median of 135, a mean of 168.10, and a standard deviation of 110.24.

Total money spent on salary per team, chosen to see if there is any relation between amount of money invested in a team and the attendance of a game. It is approximately normal with a median of $121,508,324$, a mean of \$120,571,121, and a standard deviation of \$15,312,423.

Number of All Stars per team, chosen to see if star power improves average attendance per game. It is approximately normal, with a mean of 0.9, a median of 1, and a standard deviation of 0.76.

Win percentage per team, chosen to see if average attendance is better for winning teams. It is approximately normal, with a median of 0.506%, a mean of 0.5%, and a standard deviation of

14.7%.

Pace (the number of possessions a team uses per game), chosen to see if a team that plays faster entices fans to come to games with its quicker play. It is approximately normal, with a median of 100.47, a mean of 100.66, and a standard deviation of 2.15.

Fast break points per game (a measure of the amount of points a team scores in the open court on a broken play), also chosen to see if faster, more loose play, has an effect on average attendance. It is approximately normal, though skewed slightly left, with a median of 15.98 points, a mean of 15.36 points, and a standard deviation of 2.62 points.

The All-transit index score (a composite variable, with a range of 0 to 10, composed by All-Transit that takes into account connectivity, access to jobs, frequency of service, and other sociological factors), chosen as a sociological factor to see if ease of access around a city—which ostensibly affects mobility, both in access to jobs and in direct access to the arena, affects attendance per game. It is approximately normal, with a median of 5.3, a median of 5.097, and a standard deviation of 1.55. It should be noted that the city of Toronto was not included in the All-Transit dataset (because it is not an American city), and was therefore removed.

The number of major transit routes per metropolitan area, chosen to see if direct ease of access around the metropolitan area has any effect on attendance per game. It is skewed slightly right, with a median of 5 routes, a mean of 5.621 routes, and a standard deviation 3.33 major transit routes.

Population of a metropolitan area, chosen to see if population of an area has any association with attendance per game.

Facebook fans of an NBA team, gathered by assessing the amount of people who "like" a team's profile on Facebook in June 2019 and chosen to see if online popularity of a particular team has any association with actual game attendance. It is skewed right, with a median of 2.725 million fans, a mean of 5.577 million fans, and a standard deviation of 5.35 millions fans.

Unemployment of a metropolitan area, chosen as a sociological factor to assess if the overarching economic health of an area had any impact on average attendance. It is approximately normal, with a median of 4.9%, a mean of 4.84%, and a standard deviation of 0.95%.

Median income of families in a given metropolitan area, chosen as a sociological factor to assess if median family income in a particular metropolitan area has any association with average attendance per game. It is approximately normal, if skewed slightly right, with a median of $69,454$, a mean of 69,740, and a standard deviation of $13,373.79$.

The poverty rate (percentage of residents living below the poverty line) in a given metropolitan area, chosen, like unemployment, to assess if economic health of an area has any effect on average attendance. It is approximately normal, with a median of 12.3%, a mean of 12.28% and a standard deviation of 2.73%.

Except for the fact that Facebook fans, tickets price, and scores each is weakly positively associated with attendance, according to the scatter plot (*Figure 4*), the other variables do not, at first glance, appear to be associated with attendance.

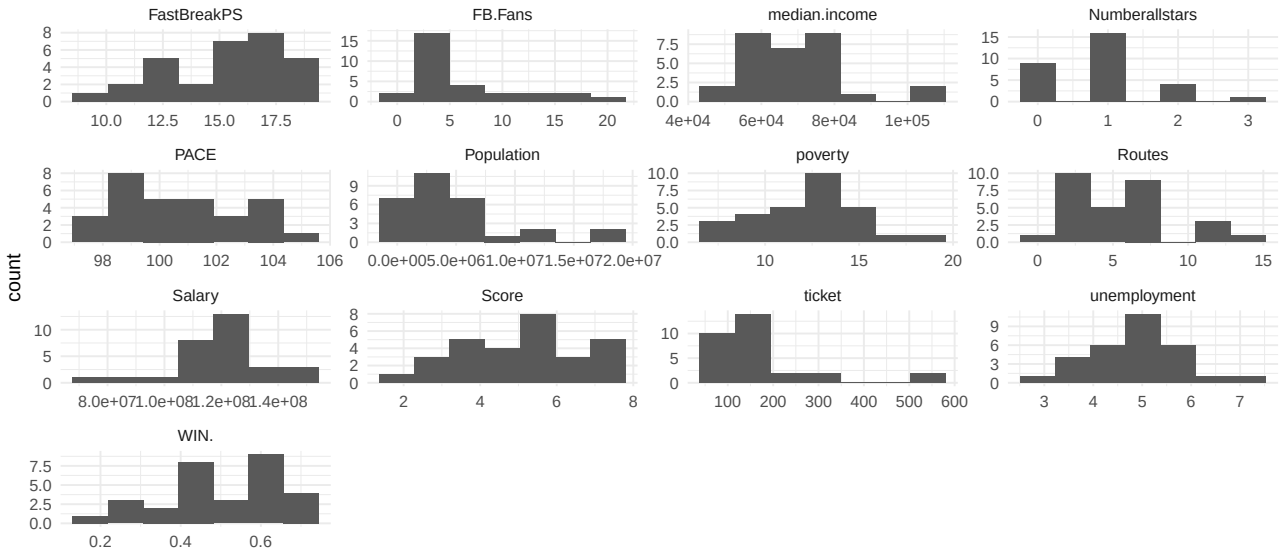


Figure 3: Histogram

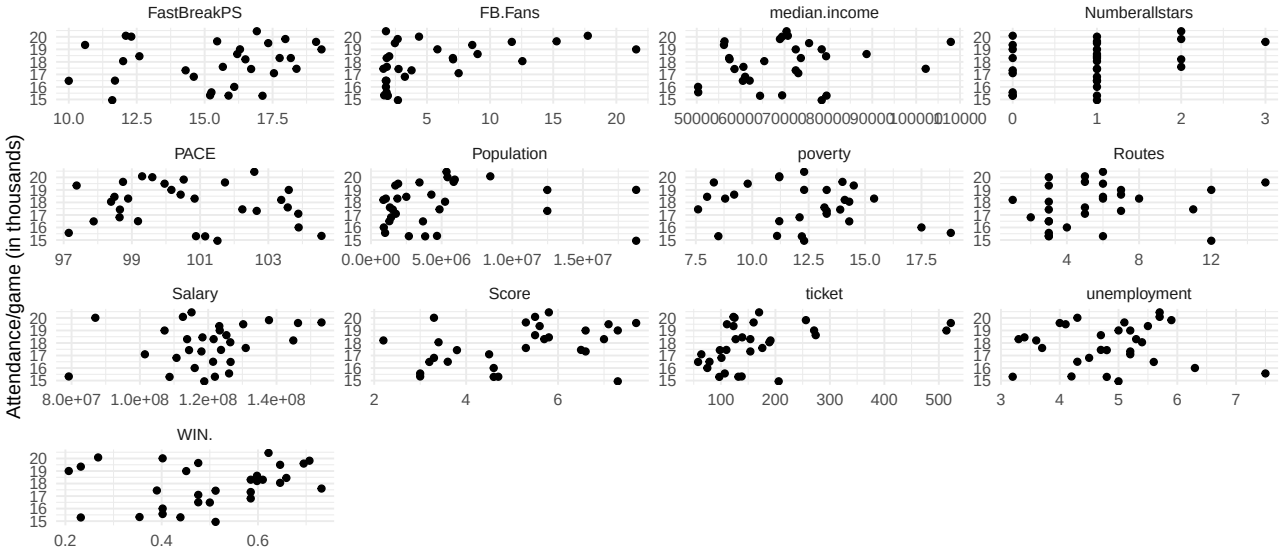


Figure 4: Plot of each IV vs. Attendance per Game

3.2.3 Procedure

The goal of the first analysis–determining factors that affect average attendance per game–is to check if any of 14 initial covariates are strongly correlated with average attendance per game. To do this, we will first analyze simple regressions to pick out variables that potentially be influential, then use model selection to select the best model by analyzing BIC, adjusted R-squared, and residual standard deviation of those models–while also making sure not to over fit the model. Once a model is selected, we will delve into the model assumptions–normality, equal variance, collinearity, and independence–to ensure that a linear model is truly appropriate for the data. Once the model assumptions are met, we will analyze outliers and influential points to determine if these should be removed. After we are satisfied with the final model, we will discuss the results and conclusions of the model.

3.3 Making the Playoffs

3.3.1 Dependent Variables

As introduced above, another goal of this project is to determine variables correlated to whether NBA teams have made the playoffs or not, which is the dependent variable. It is binary, meaning that the team either makes the playoffs or not. In the NBA, there are the Eastern and Western Conferences, and the eight teams in each conference which accrue the most amount of wins make the playoffs. For the purpose of this analysis, the eight teams in each conference that made the playoffs during the 2018-19 season are assigned the value of 1, and the rest of 14 teams in the league are assigned the value of 0. Therefore, given the nature of the dependent variable, a binary logistic regression will be used to model our dependent variable. The general model is given as:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where p is the probability of making the playoffs.

3.3.2 Independent Variables (IV)

Although there are many factors or statistics that can be considered, we initially chose 17 variables to look at, because we speculate that all those 17 variables may be important. After fitting a simple logistic regression against the DV using each variable, 5 variables are left with p value of lower than 0.05. The five variables are drive field goal percentage (FG%), catch and shoot (C&S) FG%, pull up FG%, elbow touch FG%, and PIE. Histograms of each IV are shown in *Figure 5*. Two other variables, contested 3 point shots and contested 2 point shots, have coefficients with p-values less than 0.1 but greater than 0.05. They are also considered in our selection, because it is possible that they become significant after fitted together with other variables. It is true that a variable, when holding others constant, can become significant, even when the variable, by itself, does not significantly correlate with the dependent variable. However, given the time constraint of this project, it is almost impossible to look at all 17 variables closely. Therefore, we narrowed down the scale by looking at simple logistic regressions

for each variable. This is not the best way to narrow down variables, but is the optimal way given our time.

The Drive FG% refers to the percentage of shots made on drives to the basket. It is right-skewed with a mean of 47.3% and a standard deviation of 2.64%. C&S FG% refers to the percentage of makes on catch and shoot shots. It is right-skewed with a mean of 37.3% and standard deviation of 2.03%. The pull up FG% refers to the percentage of pull up shots made. It is right-skewed with a mean of 37.2% and a standard deviation of 1.97%. The elbow touch FG% refers to the percentage of shots made when taken from the elbow–which is the free throw line extended. It is left-skewed with a mean of 56.0% and a standard deviation of 3.29%. PIE refers to player impact estimate, which measures a player's contribution of total statistics in a game. It is calculated by an aggregate of the stats of a player, divided by the accumulated stats of the enter game. It is roughly symmetrical with a mean of 50.0 and a standard deviation of 2.99. The contested 3 pt shot refers to "the number of times a defensive player or team closes out and raises a hand to contest a 3 point shot prior to its release." It unimodal/uniform with a mean of 24.4 with a standard deviation of 2.01. The contested 2 pt shot (the same statistic, but for 2 point attempts) is a bit right skewed with a mean of 38.9 and a sd of 2.88.

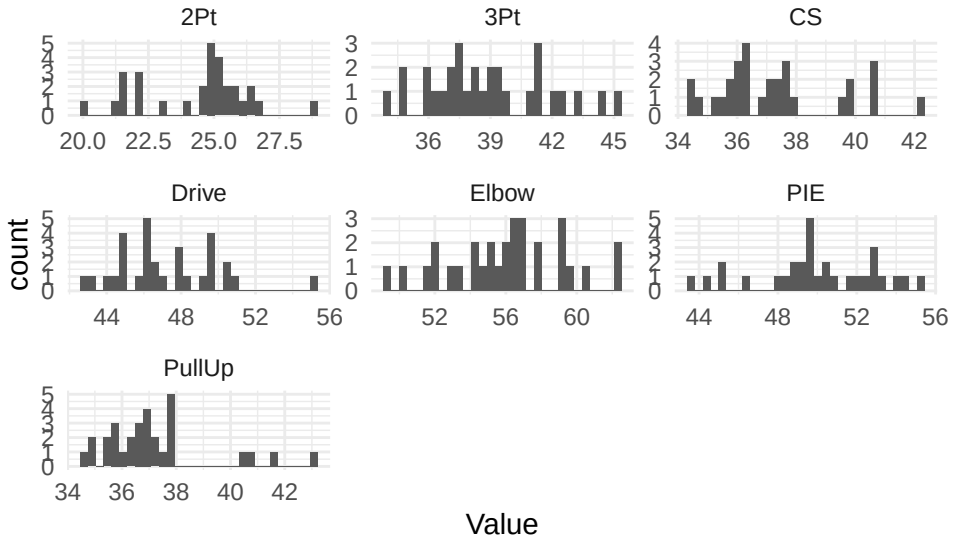


Figure 5: Histogram of Each IV

3.3.3 Procedure

The goal of the analysis is to check if any of the 7 variables selected above are strongly correlated to making the playoffs or not. To achieve this goal, one needs to first gain a better understanding of each IVs and in relation to the DV. The plots of each IV with the dependent variable (DV) are shown in *Figure 6*. From each graph, we can roughly say that, as the value of an independent variable increases, it is more likely that a team makes into playoffs (even though there is variation within each scatter plot), except for 2 pt shots contested, which has almost an opposite relationship with the DV.

Since simple logistic regressions are already conducted during the preliminary IV selection, the next step is to potentially incorporate more variables in the model. In the following pages, bivariate and trivariate models will be fitted. The fit will be discussed, and one final model will be selected based on criteria such as AIC, BIC, and sigma squared. The results from the regression estimates will be compared against actual results to assess model's reliability. Finally, conclusion, limitations, and other concerns will be discussed.

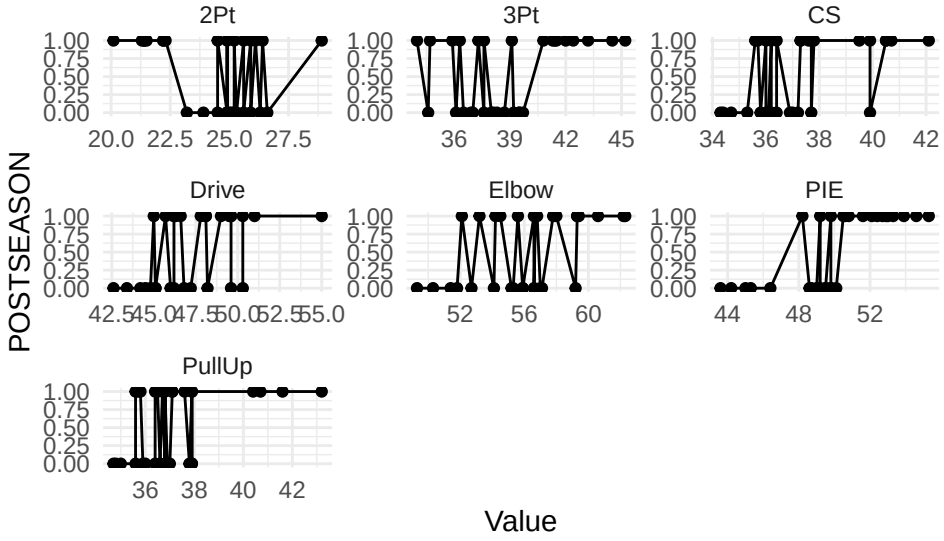


Figure 6: Plot of Each IV against the DV

4 Results:

4.1 "Take Me Out to the (Basket)Ball Game"

4.1.1 Model Selection

First, we built a preliminary linear model with variables that appeared to correlate relatively strongly with average attendance per game in simple regressions. This led to a model with 6 parameters (including the intercept): average midseason ticket price, Number of all-stars, Transit Score, Facebook Fans, median income of the city.

Table 1: Best Model Selected vs. Other considered models

	Dependent variable:		
	Attendance.game		
	(1)	(2)	(3)
ticket		−6.299	
		(4.009)	
PACE			−161.623
			(108.449)
Score	784.841**	425.076*	801.617***
	(289.251)	(213.965)	(282.387)
Routes	−290.669**		−278.545*
	(137.958)		(134.823)
FB.Fans	136.307***	207.483***	131.565***
	(44.310)	(63.248)	(43.341)
median.income		0.001	
		(0.025)	
Numberallstars	770.033**	877.050**	845.886**
	(341.541)	(405.717)	(337.036)
Constant	13,979.670***	14,657.310***	30,057.520**
	(1,014.247)	(1,437.103)	(10,833.540)
Observations	29	29	29
R^2	0.498	0.468	0.543
Adjusted R^2	0.415	0.352	0.443
Residual Std. Error	1,244.677 (df = 24)	1,309.799 (df = 23)	1,214.174 (df = 23)
F Statistic	5.963*** (df = 4; 24)	4.043*** (df = 5; 23)	5.457*** (df = 5; 23)
Note:			*p<0.1; **p<0.05; ***p<0.01

Then, we ran a stepwise package to build and identify other potential predictive models. The

adjusted R^2 led to the selection of a model, with 6 parameters, that included pace, transit score, number of transit routes, number of Facebook fans, and number of All Stars.

To avoid over fitting because we only have 30 observations, we use the BIC measure, which penalizes more harshly on the addition of variables. Thus, a model that includes transit score, number of transit routes, number of Facebook fans, and number of All Stars. The summary of all 3 models are shown in *Table 1.* There are only 29 observations considered in the models because Toronto's data is not available for some of the variables considered (as detailed in the methods section).

Although the second model had the lowest residual standard error, and larger adjusted r-squared, in taking over-fitting into account, we ultimately selected the third model, which has the lowest BIC score (510.36). Furthermore, the difference in adjusted R-squared between the "Second Model" and "Third Model" (0.443 compared to 0.415) and in the residual standard error (1214 on 23 df to 1245 on 24 df) is not so great. Ultimately, our concern in overfitting the model overrode these relatively small differences, and we settled on the "Third Model" as our final model.

The general model of a multiple linear regression is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \epsilon$$

where $\epsilon \sim (0, \sigma^2)$

Our estimated model equation is:

$\hat{Y} = 13979.67 + 787.84X_1 - 290.67X_2 + 136.31X_3 + 770.03X_4$ where X_1 is All Transit's index score , X_2 is Transit routes, X_3 is Facebook fans, X_4 is number of all stars, and $\hat{Y}$ is the expected attendance per game.

4.1.2 Model Assumptions Check, Multicollinearity & Influential Points

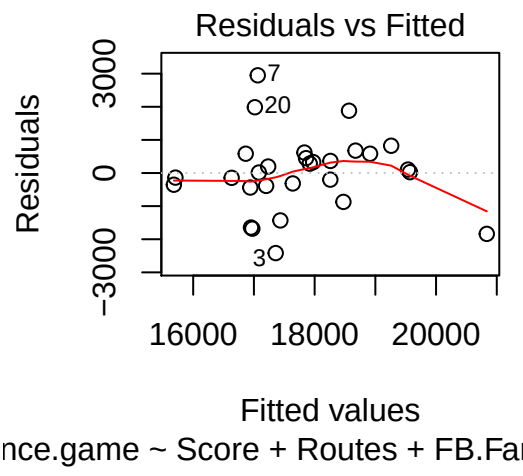
4.1.2.1 Model Assumptions

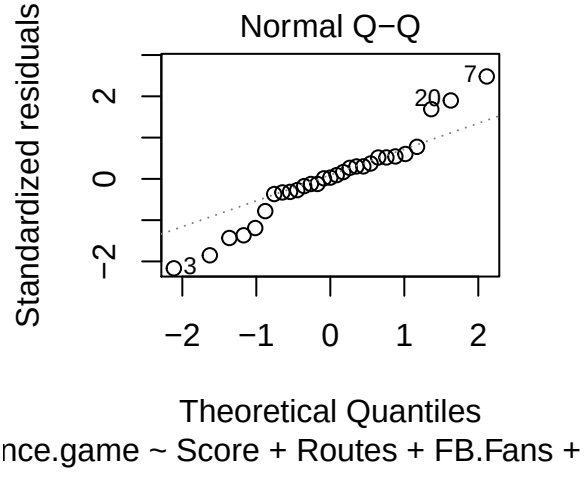
Independence: When looking at the 30 observations of the response variable, average

attendance per game, we do acknowledge that attendance at a particular game is sometimes dependent on the team that an opponent is playing. For example, if the Phoenix Suns, a team that, in the 2018-19 season, did not have a winning record and does not usually sell out its arena, played the Golden State Warriors, fans presumably flocked to the arena to see Warriors superstars Stephen Curry and Kevin Durant in action. However, while there does appear to be some slight dependence, every team in the NBA plays every other team at home at least once, so we would expect such dependence to average out, or cancel out, over the course of the season.

Equal Variance: When we look at the residuals vs. fitted plot (*Figure 4.5*), even though there is some variation, there does not appear to be any discernible trend-line. Thus, the equal variance assumption is met.

Normality: Then looking at Figure 4.5, we assess the normality assumption of the model. While the standardized residuals largely appears to adhere to the QQline diagonal, there does appear to be some deviance from the line. Thus, to be sure, we will conduct a Shapiro-Wilk test for normality. The Shapiro-Wilk test has the null hypothesis that the distribution of the sample came from a normally distributed population, and the alternative hypothesis that the distribution of the sample did not come from a normally distributed population. With a test-statistic of 0.949 (p-value = 0.166 $>$ 0.05), we fail to reject the null hypothesis, and conclude that there is no evidence that the sample did not come from a normally distributed population. Therefore, the normality assumption is met.





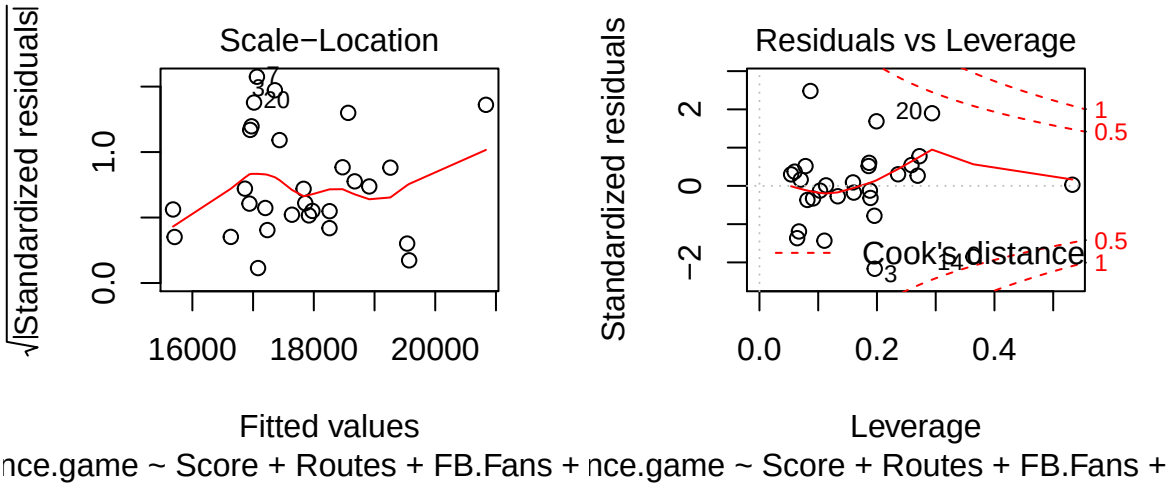


Figure 4.5. Residual Plot

4.1.2.2 Multicollinearity

Because All Transit's Index and number transit routes in an area might appear to be related, we needed to test for multicollinearity in the model to ensure that the covariates were not unduly related to one another. Here are our results for the Variance Inflation Factor (VIF), which tests for multicollinearity (a VIF of 10 indicates serious multicollinearity; a VIF of 5 is, conservatively, cause for concern):

Transit Score: 3.654534; Routes: 3.818569; Facebook Fans: 1.041602; Number of All Stars: 1.163193;

Because the respective VIFs for the variables do not exceed 10 or even 5, we can safely say that multicollinearity is not an issue with this model.

4.1.2.3 Influential Points

For outliers, there appear to be two observations, observation 3 (the Brooklyn Nets) and observation 7 (the Dallas Mavericks), whose standardized residuals lie more than two standard deviations (albeit barely more than two) away from 0. The Nets are a low outlier, presumably because they have the lowest attendance in the league, while the Mavericks, with the second highest attendance in the league, are a high outlier. However, both are not such egregious

outliers to warrant altering the model.

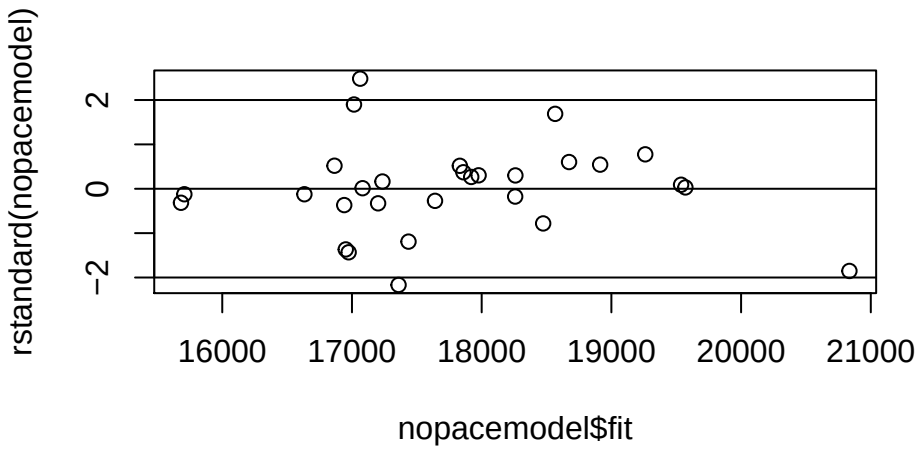


Figure 7: Outlier Plot

There are four observations—observation 7 (Dallas Mavericks), observation 10 (Golden State Warriors), observation 14 (Los Angeles Lakers), and observation 21 (the Oklahoma City Thunder)—that are flagged as influential.

The Mavericks were flagged as influential because of a DFFIT of 0.867. However, because that does not exceed the threshold of 1 (which is typically used for small datasets), we do not necessarily consider it influential.

The Warriors were flagged as influential because of an unusually high number of All Stars (they have 3), while the Lakers were flagged as influential because of an unusually large number of Facebook fans (21.57 million). The Thunder were flagged as influential because an unusually low transit score—the lowest in the study.

Because all of the points that were influential were only flagged by one measure of influence and not multiple measures, given our small number of obseravtions, we are not particularly concerned about their combined effect on the model. We still compared the models with and without the influential points.

Regular model: $\hat{Y} = 13979.67 + 787.84X_1 - 290.67X_2 + 136.31X_3 + 770.03X_4$ where X_1 is

All Transit's index score, X_2 is Transit routes, X_3 is Facebook fans, X_4 is number of all stars, and $\hat{Y}$ is the expected attendance per game.

Model with influential observations removed: $\hat{Y} = 12617.8 + 1017.1X_1 - 322.8X_2 + 194.4X_3 + 787.7X_4$ where X_1 is All Transit's index score, X_2 is Transit routes, X_3 is Facebook fans, X_4 is number of all stars, and $\hat{Y}$ is the expected attendance per game.

The coefficients do not change by a substantially large amount. Therefore, we keep our original model.

4.2 Making the Playoffs

4.2.1 Simple Logistic Regression

A brief result of all 5 simple logistic regression is shown in *Table 2*. They all have significant coefficients except contested 3 pt and 2 pt shots.

4.2.2 Model Selections

Since package `leaps` do not work for logistic regression, another package that is almost equivalent is `bestglm`. Using both AIC and BIC criteria, the best model provided has predictors of PIE and contested 2 pt shot, as the predictors. The fitted model is shown in the table *3* in comparison to the pure PIE and contested 2 pt shot models. The coefficients are not significant, potentially due to low number of observations. Therefore, if we want to keep a model that we can interpret the result, we would choose the model with player impact estimate as the only variable.

The fitted equation has a form of:

$$\ln(\frac{\hat{p}}{1-\hat{p}}) = -1280.81 + 39.42X_1 - 27.82X_2$$

where X_1 is the player impact estimates, X_2 is the contested 2 pt shots, and $\hat{p}$ is the estimated probability of making into playoffs.

Table 2: Simple Logistic Regression

	Dependent variable:						
	POSTSEASON						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Drive	0.473**						
	(0.208)						
CS		0.640**					
		(0.282)					
PullUp			0.627*				
			(0.333)				
Elbow				0.355**			
				(0.156)			
PIE					1.416**		
					(0.619)		
'3Pt'						0.281*	
						(0.155)	
'2Pt'							−0.391*
							(0.222)
Constant	−22.119**	−23.591**	−23.049*	−19.695**	−70.605**	−10.768*	9.702*
	(9.759)	(10.412)	(12.244)	(8.725)	(30.855)	(5.991)	(5.476)
Observations	30	30	30	30	30	30	30
Log Likelihood	−17.099	−16.840	−17.777	−17.185	−9.578	−18.765	−18.877
Akaike Inf. Crit.	38.198	37.679	39.554	38.370	23.156	41.530	41.754

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: Best Model Selected vs. Simple Model

	Dependent variable:		
	POSTSEASON		
	(1)	(2)	(3)
PIE	39.419	1.416**	
	(31,181.130)	(0.619)	
'2Pt'	−27.816		−0.391*
	(32,875.630)		(0.222)
Constant	−1,280.808	−70.605**	9.702*
	(1,291,878.000)	(30.855)	(5.476)
Observations	30	30	30
Log Likelihood	−0.000	−9.578	−18.877
Akaike Inf. Crit.	6.000	23.156	41.754
Note:			*p<0.1; **p<0.05; ***p<0.01

If we want to look at the model with interpretable coefficients, then the fited equation has a form of:

$$\ln(\frac{\hat{p}}{1-\hat{p}}) = -70.605 + 1.416X_1$$

where X_1 is the player impact estimates, and $\hat{p}$ is the estimated probability of making into playoffs.

4.2.3 Model Assumptions, Multicollinearity, & Influential Points

Structure of the dependent variable: the dependent variable only takes either 0 or 1, namely making the playoffs or not.

Independence: the dependent variable is not independent because one team making into playoffs means that out of all other teams in the same league, there is one fewer team making into playoffs. However, this problem is common when trying to capture wins or making the playoffs for sports when using parametric methods. It is discussed more later.

Multicollinearity: With a vif of 1.45, there is no multicollinearity between the two variables.

Influential Points: from the cook's distance plot (*Figure 8*), only observation 21 has an observable cook's distance, but even it has a really low Cook's distance of a little over 2.0e-05.

Therefore, we can conclude that there is not influential points.

Independent Variale vs. Log Odds: the plot of each independent variable against the log odds are shown below (*Figure 9*), and linearity can be assumed for both variables.

Sample size: logistic regression generally requires a large dataset. However, there are only 30 observations. There will be more discussion in the following section.

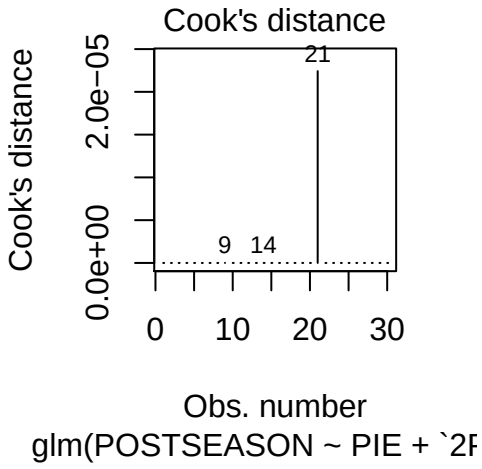


Figure 8: Cook's Distance

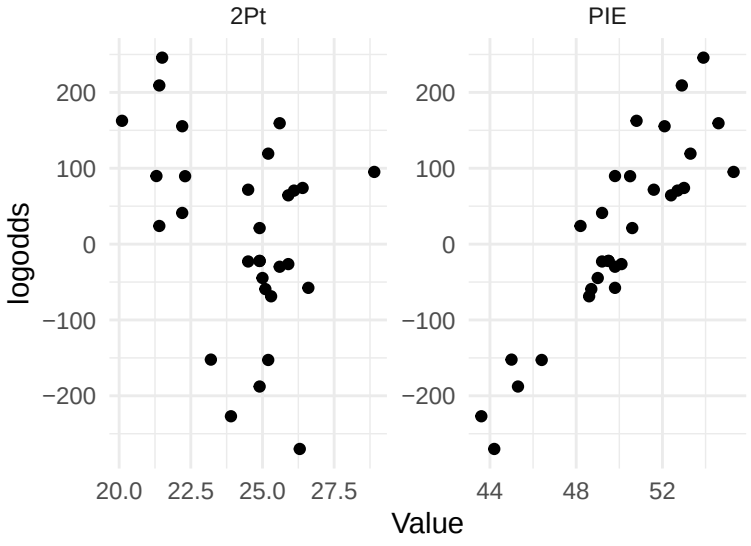


Figure 9: Independent Variables versus Log Odds Plot

5 Discussion

5.1 "Take Me Out to the (Basket)Ball Game"

The final model includes all transit index score, transit routes, Facebook fans, number of all stars. All of the covariates in this model are significant at the 0.05 confidence level.

According to the model, for teams in metropolitan areas with the same number of transit routes, Facebook Fans, and number of all stars, for every 1 point increase in All Transit's Composite Transit Index score, we would expect average attendance per game to increase by 784.84 people, on average. This is understandable. In essence, this interpretation states that, in cities for which it is easier to connect to public places—like arenas or places of employment—attendance at basketball games is expected to be higher than cities for which attendance is lower. Ease of access to the arena, then, facilitates the ability to fill the arena up and get more people in their seats.

On the other hand, according to the model, for teams in metropolitan areas with the same transit score, Facebook Fans, and number of all stars, for every 1 transit route increase, we would expect average attendance per game to decrease by 290.67 people, on average. This statistic indicates the number of major transit routes available in a metropolitan area does not necessarily ensure the quality of such transit. For example, a city like Minneapolis, though it has a relatively large number of transit routes (6 routes), has a decidedly below average All Transit Composite Transit Index Score (4.7). Thus, the fans may find it harder to get to the arena because of the inefficiency of the transit routes. Also, more transit routes may indicate a more sprawling metropolitan area—where fans decide they would rather stay at home and watch the games rather than risk a long trip to them.

Additionally, according to the model, controlling for other variables, for every 1 million person increase in Facebook fans, we would expect average attendance per game to increase by 136 people, on average. Teams with a stronger and more rapid online fanbase would be expected to have more fans show up in person for the games. However, in reality, an increase of 136 people for every one million fans, while intriguing, would not appear to have an overly large impact on

attendance per game.

Lastly, according to the model, for teams in metropolitan areas with the same transit score, number of transit routes, and number of Facebook fans, for every one player increase in All Stars, we would expect average attendance per game to increase by 770.03 people, on average. The results of this interpretation also make sense for two reasons. First, having more All Stars on a team usually means the team is better, and common sense dictates that fans will want to come watch a good team rather than a bad team, all other factors being equal. Second, star power is a powerful and alluring entity for fans. After all, the NBA is a league of stars, with Lebron James, Stephen Curry, James Harden, and others having become ubiquitous in popular culture. Thus, fans will want to come and watch stars perform—and, if a team has these stars, they will be more compelled to go to the games.

In this study, we were limited by our relatively small sample size, as there are only 30 NBA teams to analyze. This constrained some of our analysis to over fit our models. Furthermore, we were constrained because the season just concluded in June 2019, so not all statistics about the season and, sociologically, about the metropolitan area are available. For example, Toronto's official poverty rate in 2018 has not yet been released; we used an estimate provided by a respected nonprofit. Additionally, the NBA has not yet released its report on average ticket prices from the previous season, which is why we had to use average price on the secondary market. Furthermore, we were constrained in variable choice because of multicollinearity, as certain variables—like points scored, are directly associated with assists or field goal percentage. Thus, our analysis was certainly constrained by the availability of suitable data to analyze.Lastly, as acknowledged earlier in the model assumptions, we recognize that the recorded observations cannot be completely independent of each other, as attendance per game does depend, to an extent, on the opponent a team is playing. However, we believe that, over the course of a season, such dependence—because it affects all teams—cancels out.

In terms of potential areas for future study, it would be interesting to test other factors—like the amount of public money poured into the stadium, the political leaning of a city/state where a team plays, or in-arena entertainment features—affect average attendance per game.

5.2 Making the Playoffs

According to the significance of the coefficient, the final model contains only the player impact estimate (PIE). We can conclude that as the PIE increases for 1 point, On Average, the chance of making the playoffs is multiplied by $\exp(1.416) = 4.12$. This makes sense because the better the players on a team are, then it is more likely that the team does well.

According to he AIC and BIC criteria, the final model contains the variable player impact estimate (PIE) and contested 2 pt shots. Even though this model has the lowest AIC and BIC, the p-value of all the coefficients are close to 1. The insignificance of the p-value is potentially due to the fact that there are only 30 observations in the model. Since both coefficients are not significant (though they each were significant in simple logistic regression with response variable), we cannot interpret the corresponding coefficients.

Furthermore, another challenge with the model fitted is that the observations, in terms of making the playoffs, are potentially dependent on each other. One team in the western conference making into the playoffs means that another team in the same conference will lose the opportunity to make it into the playoffs. However, although the independence assumption is challenged, the fact that this occurs throughout the league, and for every team, means that the challenge was not large enough to halt our analysis. After all, any analysis of wins or making the playoffs in the NBA needs to grapple with this reality (previous analyses, as cited in our introduction, did so by acknowledging the shortcoming), and we are not an exception.

5.2.0.1 Areas for Future Study

Therefore, a potential area for future study would be to analyze the players instead of team, though a player's status is strongly dependent on other players of the same team, thus creating more problems. Other non-parametric methods like K-nearest neighbor and decision trees could also be considered in the future to better model the data. For future reference, one can also consider redoing the analysis for data of different leagues and comapre if the models are the same.

6 Bibliography

"2018-19 NBA Season Summary." Basketball Reference, https://www.basketball-reference.com/leagues/NBA_2019.html.

Archives Canada. "Censuses." Library and Archives Canada, 2 Dec. 2019, https://www.bac-lac.gc.ca/eng/census/Pages/census.aspx.

Kotecki, Jason, "Estimating the Effect of Home Court Advantage on Wins in the NBA" (2014). Honors Projects. Paper 124. http://digitalcommons.iwu.edu/econ_honproj/124

Statista. "National Basketball Association Teams Ranked by Player Expenses (Salaries)* (in Million U.S. Dollars)." Statista, Statista Inc., 6 Feb 2019, https://www.statista.com/statistics/193709/player-expenses-of-nba-teams/

Sprung, S. (2019, March 7). Inside The NBA's Push To Make Basketball The World's Most Popular Sport. Retrieved from https://www.forbes.com/sites/shlomosprung/2019/03/04/nba-china-ceo-derek-chang-takes-us-inside-nbas-push-to-make-basketball-worlds-most-popular-sport/#6ed0cde151b0

"Team Stats." NBA Stats, 20 July 2019, https://stats.nba.com/teams/traditional/?sort=W_PCT&dir=-1.

"The Most Liked NBA Teams on Facebook." Trackalytics, 15 Sept. 2019, https://www.trackalytics.com/the-most-liked-nba-teams-on-facebook/page/1/.

"Transit Rankings." AllTransit, https://alltransit.cnt.org/rankings/.

US Census Bureau. "Census Urban and Rural Classification and Urban Area Criteria." The United States Census Bureau, 2 Dec. 2019, https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html.

Yang, Y. (S. (2015, May). stat.berkeley.edu. Retrieved from https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang _Thesis.pdf.