Taylor Swift (most likely) Cannot Dunk a Basketball

A while back, I embarked upon a project to try to definitively conclude who was the most overpaid NBA basketball player last season (2022-23). I undertook this project for a variety of reasons. I wanted to improve my statistical ability; I wanted to be able to bait my friends into an argument where I was intimately familiar with the data (and would win); and I wanted to build something completely on my own.

The size and the scale of the project quickly grew out of hand. The initial results I had obtained from running single and multivariable linear regressions had not been all that great. Obviously, when obtaining the data, a large amount of data cleaning and refinement had to be completed. I had to set various requirements such as games played in a season.

Then, there are an ample amount of statistics for NBA basketball. Box scores are recorded every night and provide the basis for forming all of the basic statistics recorded in basketball, and give the building blocks for combining statistics to produce ever more complex stat measures.

This is really where the project broke down. Of course, the many multitudes of combinations of stats choices and model choices one could use in any given analysis creates huge complexity that makes it difficult for one person to undertake such a project alone. However, for me, my awareness opened up to the difficulty of the project I had taken on. Yes, there are many variables to consider. But, what's more important than the sheer scale of this proposition is to realize the confounding variables in both the input and the output of the data itself (with the input being the statistical measures and the output being salary). So let's embark on a thought process of what we might do in this analysis.

Let's start with the inputs. I think a simple and natural place to start for this kind of analysis would be with points. Points are how you win games in basketball, so they should be a useful predictor in salary. But wait, you say, what about players who maybe missed some games? They could still be incredibly valuable, but be punished because of a stretch of missed games. Now I could point out to you that if you were to take two identical basketball players and one was more injury prone than the other, the less injury prone one would have to be more valuable. That player would be more likely to be there when you need them. But, I don't do this and I grant the point that we shouldn't punish players for some missed games. Then you suggest that we use points *per game* as a better indicator of value in a player.

Now we try to use points per game as our predictor of value. But, as we run this analysis another thought pops into our heads. What about a player who is so good that his team is usually up by the end of the third quarter and he sits for the rest of the game? We don't capture a player's performance if this happens. We also now have the opposite of the previous problem in that we are ignoring a player's missed games. We're also punishing players that don't get that many minutes. A player who comes off the bench will generally score less points than their

counterpart who starts. Sometimes, this happens to good players who are on teams with a superstar who starts at their position.

So, we move on to yet another measure of points. We now use points per 100 possessions. We take all of the points a player has scored in the season, divide it by the number of possessions, and multiply it by 100. Now, we have a measure that eliminates the bias towards points that could be affected by the playing time a player has in a game. But, we ask, can this stat be biased in any way? We think about players who play in garbage time. When a game is generally thought to be decided in some way (i.e. a certain point differential with some small amount of remaining time to play), coaches often empty the “ends” of their benches to play out the remaining minutes. You often end up with lower level players playing against lower level players in these last remaining minutes. This stat would then push up these players points per 100 possession statistics that could potentially make them look as good or better than their counterparts.

I think it’s becoming clear that any stat that we use, any combination of stats that we use, will have some sort of drawback to it. This is true for offensive stats, defensive stats, and advanced stats as well. We don’t want to feed a model too much data so that it becomes unusable and overfit to the data. We are trying to identify outliers and we need our model to try to get as close to a “true” estimate of value as possible.

So we make concessions and identify different groups of variables that we would like to run our model on. If we start with linear, we need to think about how our data is distributed. Is the pay scale in the NBA linear? It most certainly is not. We can start by simply identifying that there are league minimums and maximums that a player can be paid. Rookie contracts are also something to consider. A rookie is signed to a deal that is specified based upon draft position. So if a player immediately comes into the league and plays at an elite level, he will be compensated at a rookie level. This complicates our efforts to try to determine a true value based on stats alone, we probably need to compensate for rookie status as well. The most valuable player in the league (whoever you would like to say this is) will be compensated a level akin to several other players even if he has the best season in recorded history.

In addition, the NBA player market is truly a market as well. What I mean by this, is that teams are bidding for players in relatively illiquid situations. This can cause players to be compensated at a level above what they are really worth simply because several teams have a need to fill a position and this player is the best available. Players’ original drafting teams can be the most at risk for this because they are usually able to offer the most money to a player (this is a structural feature of NBA contracts). We have yet another confounding variable to think about.

Without taking too much time discussing every possible issue, I want to talk about the final boss of this project. As I was working through my analysis, basketball was always on my mind. One day, Michael Jordan popped into my head, and I had this thought: “What would be the value of Michael Jordan in today’s NBA?” To be clear, I mean Michael Jordan as he is right now.

I think he would be worth quite a lot of money. Imagine if he signed a 10-day contract with a team. Do you think the three or so games he would play in would sell out? I think they almost certainly would. People want to see Michael Jordan because he is Michael Jordan. Teams would pay a pretty penny to have him do this. Michael Jordan was a great basketball player, however, although he could probably still kick my but in a game of pickup, he is nowhere near a low level NBA player today.

As I was going through this thought process, something kind of incredible happened. Taylor Swift started dating Travis Kelce and publicly attending Chiefs games. Out of nowhere, many people who had no interest in football started tuning in to the games, hoping to get a glimpse of Taylor.

Taylor Swift (as far as I'm aware) cannot play football in the slightest. No slight intended, I just don't think she's producing anything valuable on the field. But, her value to the Chiefs is undeniable. She produces jersey sales, ticket sales,and watch time.

We see a similar sort of phenomena with NBA players. Lebron James will be entering his 21st season soon, and is in the twilight of his career. Lebron still produces at a high level, but next year is likely to see a decline. He just is no longer capable of doing the things he once did. But, he still demands a premium. Why? His jersey sales consistently rank among the highest in the league; he brings people to the game (in-person or on TV). There is an aura to Lebron James that attracts the masses.

With this in mind, I had to question if stats alone would really be able to tell me what I was searching for. The more I thought about it, the more I was certain the answer was no. We established that all of our statistics have confounding variables. As such, we would likely be able to establish any number of players as the most overpaid in the league. Obviously, there is a strong correlation between a players' performance and their "marketable" value. Perhaps a more thorough approach would consider a player's entire career to produce an estimate of their current fair salary. Even then, I think there would be a fair number of intangibles that would be difficult to capture. While I don't have the time or resources to currently undertake such a project, I'm not turned off by the difficulty or complexity of such a task. If anything, I'm more interested now than when I started the project.

I'll leave it with this. The epitome of the NBA is the dunk. It's raw and pure athleticism in an incredible display. I'm pretty sure Taylor Swift can't dunk a basketball. But, imagine if she announced she was competing in this year's dunk contest. Do you think more or less people would watch it than are going to currently?

Maybe a team should try to sign her to a 10-day.