QTM 220
Dr. Hirshberg
Jonathan Lin, Justin Ross, Sean Hong, Chris Li, Ben Schwartz

Final Write-up: Estimation of Salary Given Differences in Draft Picks in the NBA

Introduction

The age of technology and statistics have thrusted sports analytics into mainstream analytics. Possible analyses include how hand size affects player shooting performance or the use of a player's efficiency score as a predictor of performance. The main goal of our project is to find how a player's initial NBA draft pick number affects the player's salary, depending on their performance during the 2 year post-draft contract. We will attempt to measure this effect by observing how salary changes if we increase the draft pick number for each player by one. Thus, by holding all other variables constant, we will be able to measure the treatment effect of draft pick number on player's first post-draft contract salary. In determining this treatment effect, we will highlight the importance of draft-pick number when it comes to a player's monetary success later on. Thus, salary is able to be used as a proxy for performance.

Our dataset consists of real information on players that were drafted into the NBA between 1997 and 2013, including information such as draft pick number, team winning percentage three years after draft year (season of first offered contract), player salary of first offered contract, and average minutes per game played during season of first contract. As previously stated, we would like to see the effect on salary if a player was picked 1 pick later in the draft. Thus, our treatment is shifting a player one pick down. Our estimation in terms of potential outcomes is $E[E(Y_i(Pk_i)) - E(Y_i(Pk_i+ 1))]$ as we want to see the treatment effect on salary based on shift in draft pick. To obtain this with the data and variables that we had collected, we have defined our estimate in terms of what we observed as $E[E(Y_i \mid Pk_i = Pk, MPG, WP) - E(Y_i \mid Pk_i = Pk + 1, MPG, WP)]$ in which we take a player's draft pick number (Pk), minutes per game (MPG), and winning percentage (WP) as inputs.

Our estimation has a variety of implications. On the team's side, our estimation will allow teams to make informed decisions on the cost benefit of keeping or trading 2nd contract players. From the player's perspective, our estimation will allow players to gain insight into a sort of expected salary decrease/increase based on their pick (pre-draft motivation for a player). We hope our estimation is significant enough that it will provide real-world benefits such as those listed above.

Model

While our original dataset has missing values for some years and draft picks, we will act as if there is a meaningful population outside our sample dataset from which we draw our data. We will run a partially interactive model regression model on our sample data using minutes per game (*MPG*), draft pick (*Pk*), and team winning percentage (*WP)* as our model's features. We believe that the features MPG and Pk will have a negative interaction and the same goes for features WP and Pk. Our model, with the calculated coefficients, that we have used to fit this data is

Regression Coefficients (real data model)

$Salary$ = 80920.621(MPG_i) -19432.653(Pk_i) - 414297.286(WP_i) -40408.833(Pk:WP) -1735.515(Pk: MPG) + 705832.737(intercept)

From our model, we see that minutes per game and pick number has a positive effect on player salary while winning percentage has negative effects on salary. This is mostly intuitive as we can assume that a player with more minutes played per game and with a lower draft pick number will be more highly coveted and thus receive a higher salary. However, to our initial surprise, winning percentage has a negative effect on salary, which probably means that a successful team is less apt to pay a higher salary to an individual player if the entire team plays well. In other words, a team will perform better with multiple average salary players than a team with one drastically paid player and multiple below average players.

Fake Data Generation

Generating fake data is important in answering our question as it allows us to randomize data in a way that we believe makes sense and of which we also know the distribution. To generate our data we did the following steps:

1. We first grouped players into 6 categories based on team WP and player MPG.
    a. Teams with a winning percentage >= .5 were classified as *good.* On these teams:
        i. Players with 32-48 MPG were classified as *high*, in terms of minutes played.
        ii. Players with 16-32 MPG were classified as *medium*.
        iii. Players with 0-15 MPG were classified as *low*.
    b. Teams with a winning percentage <.5 were classified as *bad*. On these teams:
        i. Players with 32-48 MPG were classified as *high*, in terms of minutes played.
        ii. Players with 16-32 MPG were classified as *medium*.
        iii. Players with 0-15 MPG were classified as *low*.
    This stratification yielded the following categories: *goodhigh, goodmedium, goodlow, badhigh, badmedium,* and *bad low.*
2. For each player, we generated 60 possible salaries based on draft pick. We assume [*DP*] is a vector of length 60 containing all integers 1-60.
    a. For players in the *low, medium,* and *high* MPG categories, we established base salaries (*base*) as $1 million, $2 million, and $3 million, respectively.
    b. Thus for each player MPG category, we generate a vector of 60 possible salaries [*FakeSalariesMPG*] using the following function:

$$[\mathit{FakeSalariesMPG}] = \mathit{base} \times (2 - ([\mathit{DP}]/60)$$

This assumes that players drafted earlier and with a higher MPG will have a higher salary.

3. We created a range value of possible salaries based on the team's performance as winning percentage (*WP).* We will once again assume [*DP*] is a vector of length 60 containing all integers 1-60.
   a. For players in teams part of the *good* category, we will set our *multiplier* as 2. For players in teams part of the *bad* category, we will set our *multiplier* as 1.
   b. Thus, for each team WP category, we generate a vector of 60 possible salaries [*FakeSalariesWP*] using the following function:

$$[\mathit{FakeSalariesWP}] = \mathit{multiplier} \times (1.1 - ([\mathit{DP}]/60)$$

This function assumes that teams with higher winning percentages value their players more and are willing to pay a higher salary.

4. Lastly, we combined both functions from parts 2 and 3 to get a salary range for each draft pick 1-60 for each of our 6 categories.
   a. We used the following function *Min* in which we get a vector of 60 salaries that determine the minimum salary for each draft pick in a "fake data" category :

$$\mathit{Min} = [\mathit{FakeSalariesMPG}] - [\mathit{FakeSalariesWP}]$$

   b. We used the following function *Max* in which we get a vector of 60 salaries that determine the maximum salary for each draft pick in a "fake data" category :

$$\mathit{Max} = [\mathit{FakeSalariesMPG}] + [\mathit{FakeSalariesWP}]$$

5. Using this range, we constructed a uniform distribution from which we could take samples from and assigned 60 different salaries for each player based on picks 1-60.
6. The pick that each player received to generate the random salary was given by a vector or probabilities for each of the 60 picks for each player.
   a. This equation was used to sample the salary from the uniform distribution:

$$\text{Prob(Draft Pick)} = \frac{1 + MPG/[DP]}{sum(1 + MPG/[DP]}$$

Confidence Intervals

Our estimate for the fake data equated to $28488.98, which is relatively similar to the estimate of the real data. The difference in estimates reveals that, on average, the salary of players in the fake data decreases *less* than the salary of real players as draft picks decrease. For the 95% confidence interval of fake data, our target value is $28488.98 and the bounds for our confidence interval are [$28591.19, $28386.77].

From our real dataset, the target value was $36329.97 and our 95% confidence interval was [$36036.48, $36623.46]. The confidence intervals from our fake and real data are relatively comparable. This emphasizes the fact that being picked 1 pick later will decrease a player's first offered contract salary by ~$30,000. Our fake data did not take into account the politics and negotiations that come with determining both pick number and offered contract salary, hence why we do see a difference between our fake data estimate and real data estimate.

Implications

Our estimate has several implications for NBA players and the future of the league. That being said, it is important to remember the presence of confounders that we chose to hold constant for the sake of this study (injuries, franchise cap space, skill sets, position depth, etc.). Our estimate provides information about the relative worth of a draft pick in terms of future salary, this allows teams to plan years in advance with more accuracy. Teams need to make sure they have enough cap space in their budgets to make signings and roster movers, so a lot of planning is required. Having a quantifiable value for the worth of a draft pick position helps during the drafting process because teams can make more informed decisions when trading draft picks. Additionally, having a ballpark figure for the future salary of a player allows teams to make moves in free agency with more confidence.

Figures

This figure is a boxplot with a pick number on the x-axis and the difference in salary between our real and fake data on the y-axis. From this plot we can see the median salary difference for each draft pick 1-60, along with the maximum, minimum, first quartile, and third quartile values. It makes sense that the distribution of salary differences hovers around 0 because our real and fake data are attempting to tell the same story.

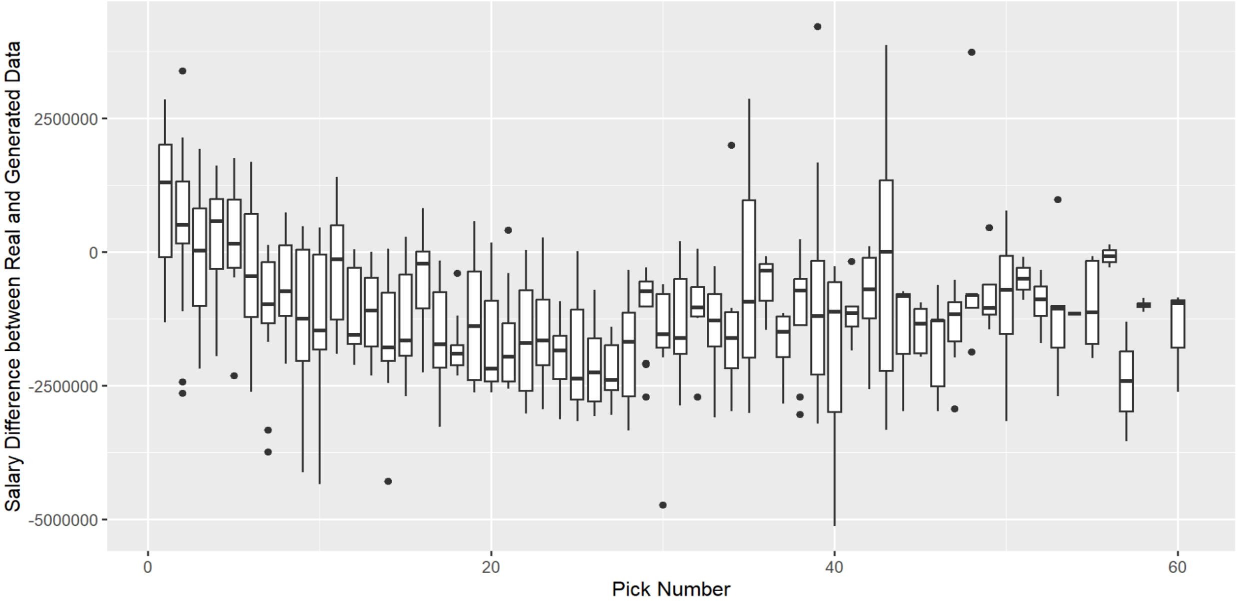

Figure 1: Real Data Salary - Fake Data Salary Distribution at Each Pick

The next graph depicts the salary distribution per draft pick for the real data. Obviously we see a steady decline in salary as draft pick increases and overall talent decreases. Around pick 30 we start to see some randomization of salaries which continues for the remaining pick. One possible explanation for this could be that players that get picked in later rounds have less "guaranteed" talent, and there's more variability in the outcome of the player's performance.

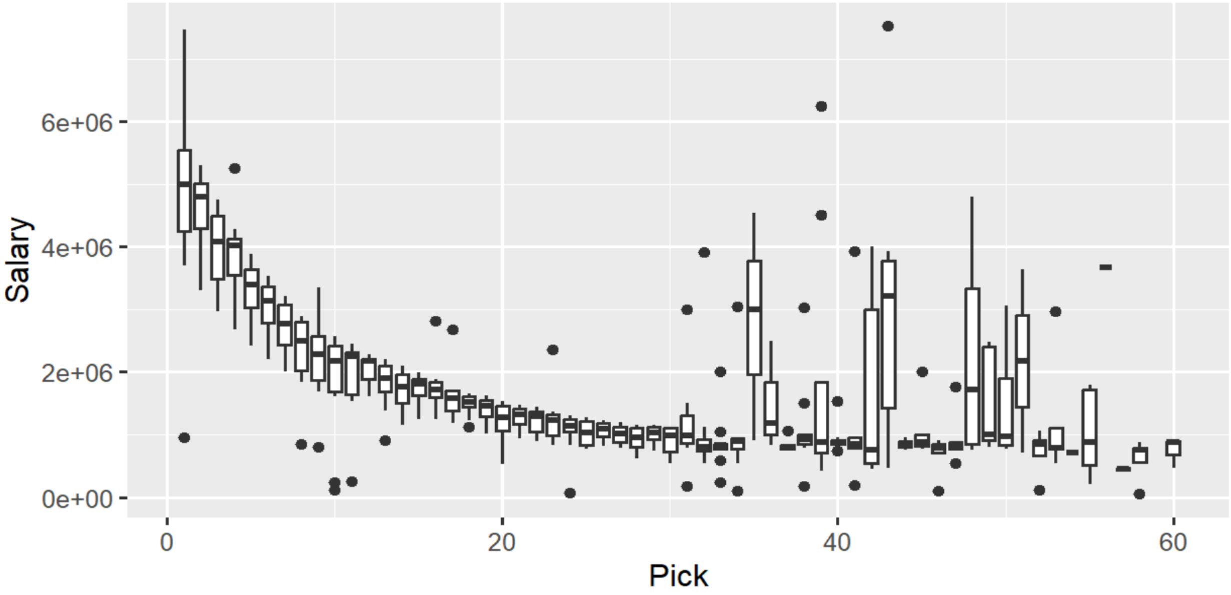

Figure 2: Distribution of Salary at Each Pick for Real Data

This figure displays the salary distribution for each pick using fake data. Similar to the real data we see a downward trend, this time with a fairly constant slope.

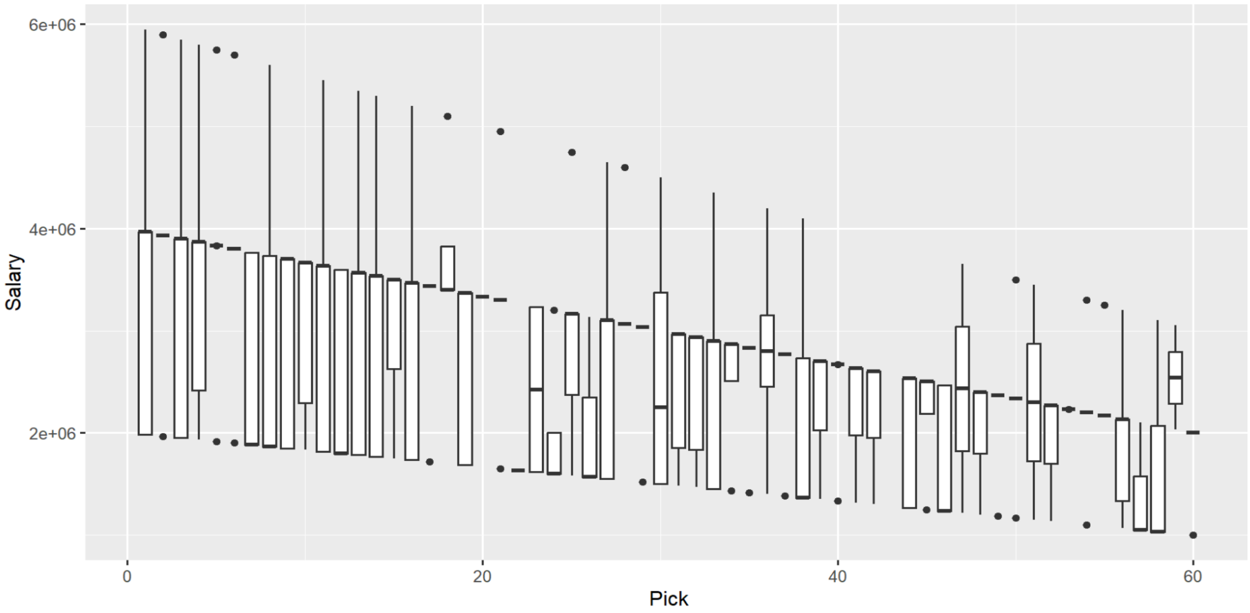

Figure 3: Distribution of Salary at Each Pick for Fake Data

This graph is showing the difference in salary of players at pick *n* from the salary of players at pick *n + 1* in the real data. Visually, we can see the mean of this distribution to be at approximately 35000, which reinforces the value we calculated as our estimate, 36329.97.

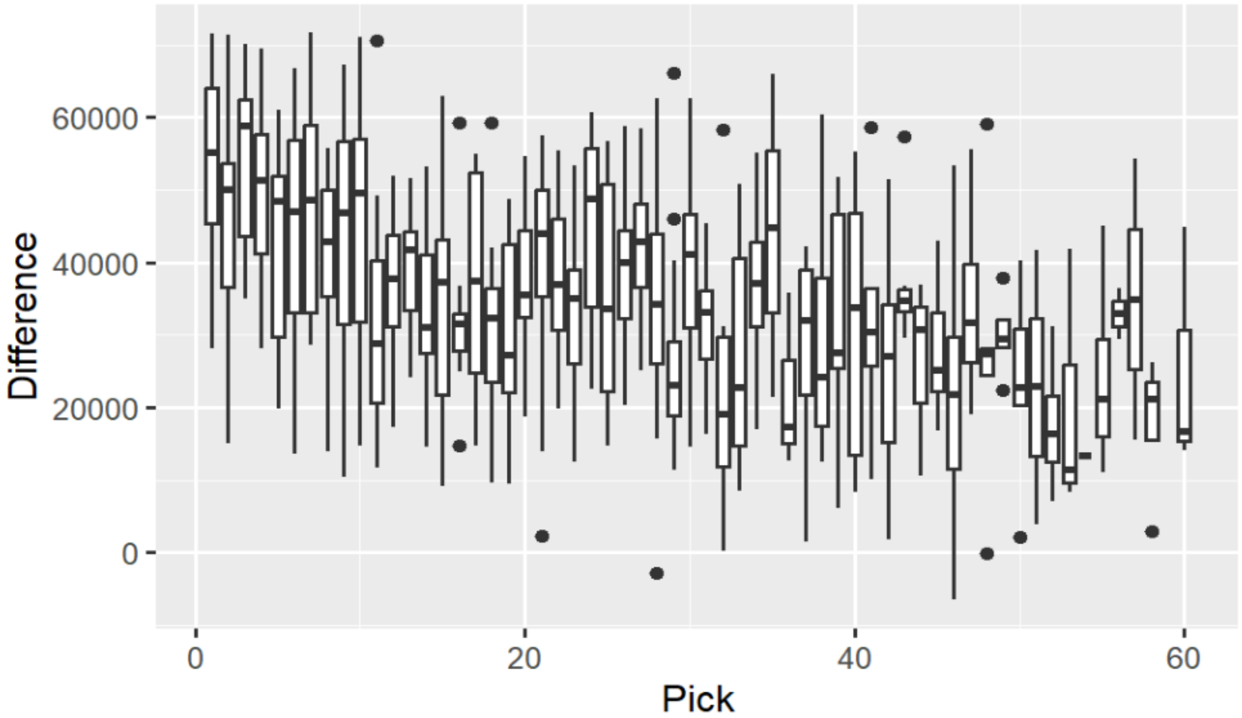

Figure 4: Difference in Salary at Pick - Salary at Pick + 1 for Real Data

The final graph, Figure 5, shows the difference in salary of players at pick *n* from the salary of players at pick *n + 1* from the fake data (randomized uniformly). The mean of this distribution looks to be about 30000, which correlates nicely with the fake data estimate, 28488.98.

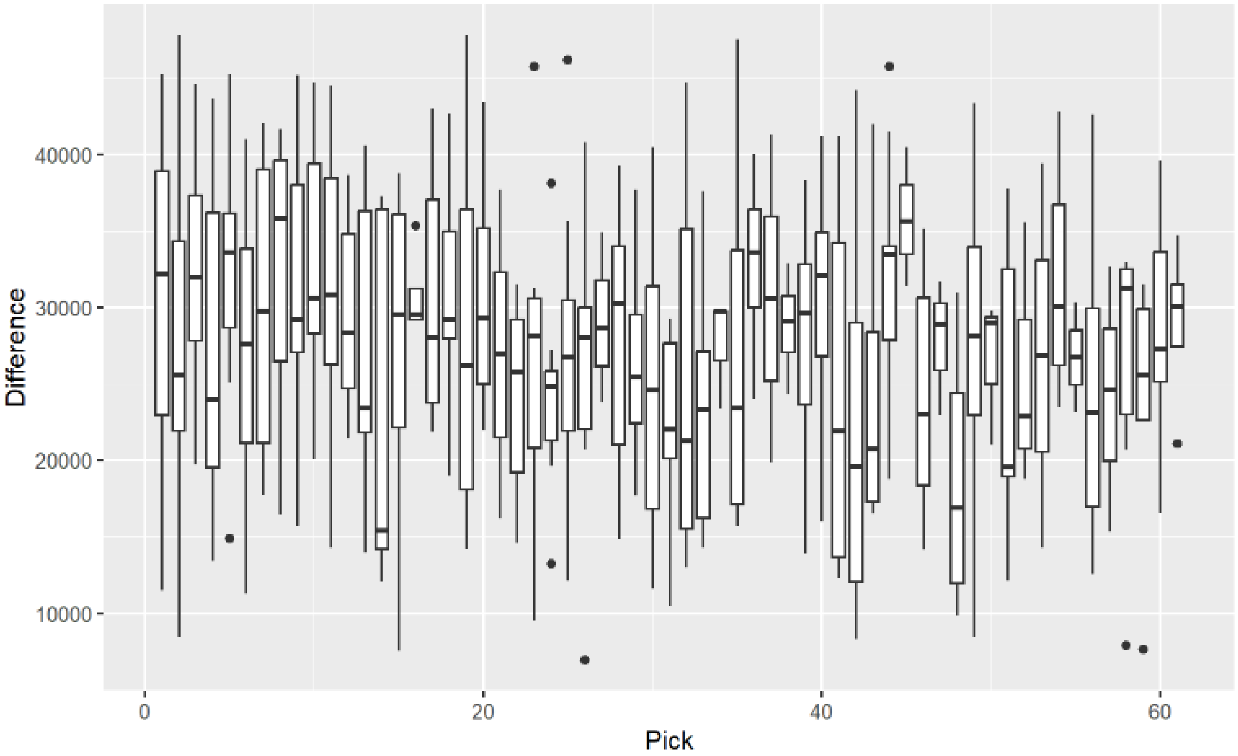

Figure 5: Difference in Salary at Pick - Salary at Pick + 1 for Fake Data