Exam 3 Take Home

Tyler Trupke

April 21, 2019

1. Data set being used

NBASAL (information regarding NBA palyers and their salaries).

2. Research question

I want to examine the relationship between NBA salaries and points per game. Specifically, I want to see determine the only in-game statistic which affects wage is points per game. The three major basketball statistics are points, rebounds, and assists. I believe that getting more rebounds or having more assists will have no effect on wage when points per game is controlled for.

3. Variables of interest

Dependent variable: Wage. In this data set wage is measured in annual salary, in thousands of dollars. So, if a data point has wage = 650, that means they made $650,000 for that year. The data set also includes a log(wage) variable, which could be used also, to examine the percent changes in wage rather than the nominal values.

Explanantory (independent) variable: Points per game, denoted "points" in the data set. This is simply an average of the total points they scored on a season divided by the number of games they played. I expect the relationship between points and wage to be positive, because if you score more points per game you are likely a better player, meaning you should earn a higher wage.

Control variables: One variable I believe should be controlled for is experience. The reason for this is that as a rookie on your first contract, you are only allowed to make a certain amount of money, regardless of your performance in games. Simliarly, there is a veteran's minimum contract that is given out to a significant amount of veteran NBA players, regardless of their performance as well. By controlling for this variable it ensures that every player has the same opportunity to make money based on their performance.

I also will be controlling for minutes per game. I believe this will have a strong correlation with wage and points per game, so it should not be excluded from the model. The reason for this correlation is that players who play more minutes per game are the starters, who are the team's best players, meaning they score a lot of points and earn the highest wages. By holding this constant in the model the partial effect of points per game on wage will be better explain by the estimates, because it will assume everyone is playing equal minutes.

In my unrestricted model I will be including two other variables, assists per game and rebounds per game. The null hypothesis I will be working with is that points per game is the only signifcant statistic regarding wages, and that assists per game and rebounds per game will not matter. I will be doing a regression of this unrestricted model and then a restricted model which does not include these two variables, then performing an F-test to determine the statistical significance.

4. Models

Unrestricted model:

$$lwage = \beta_0 + \beta_1 points + \beta_2 assists + \beta_3 rebounds + \beta_4 exper + \beta_5 avgmin + u \tag{1}$$

Restricted model:

$$wage = \gamma_0 + \gamma_1 points + \gamma_2 exper + \gamma_3 avgmin + u \tag{2}$$

A few notes here: I will be using log(wage) for my dependent variable. A log-level model will be easiest to understand in this case, since wage is in thousands of dollars it could be confusing otherwise. Making it a percent change will be much easier to understand and make it easier to draw reasonable conclusions.

$$H_0 : \beta_2 = 0, \beta_3 = 0 \tag{3}$$

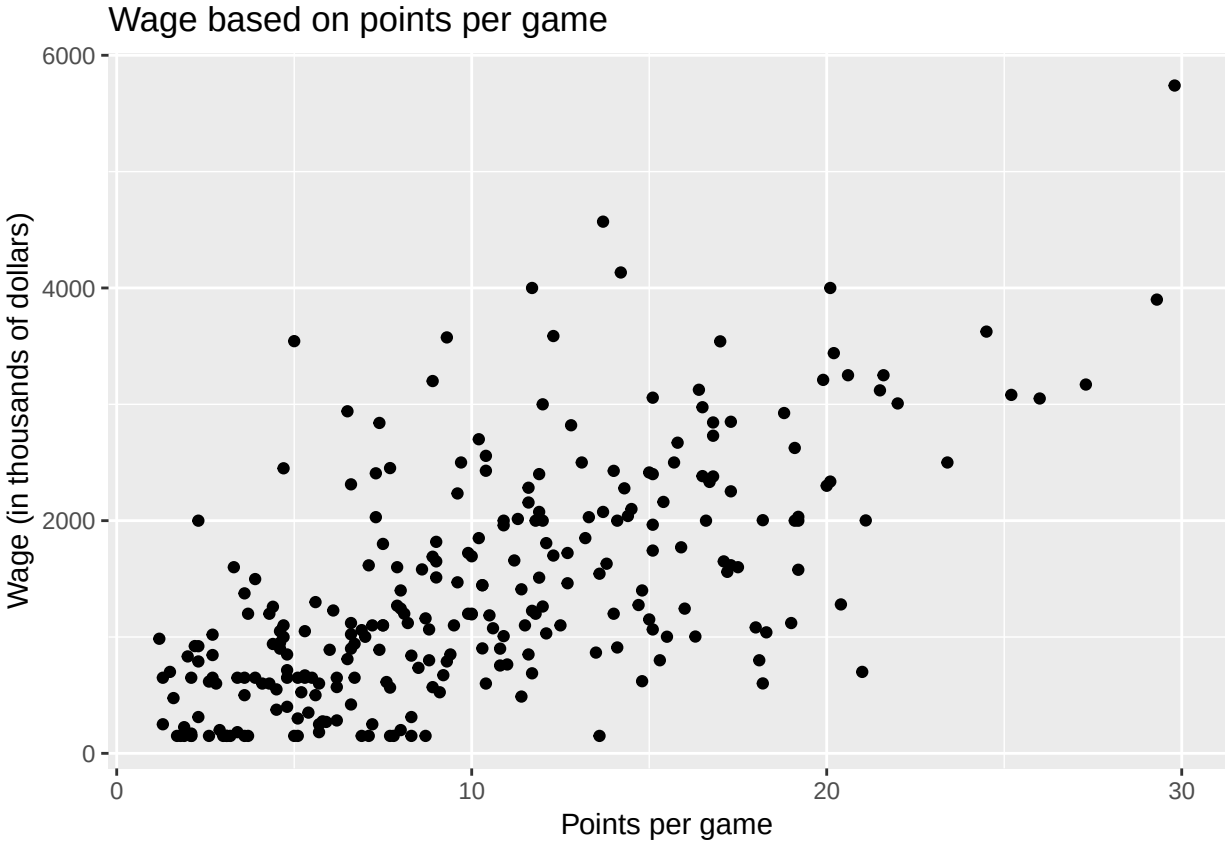
5. Summary Statistics

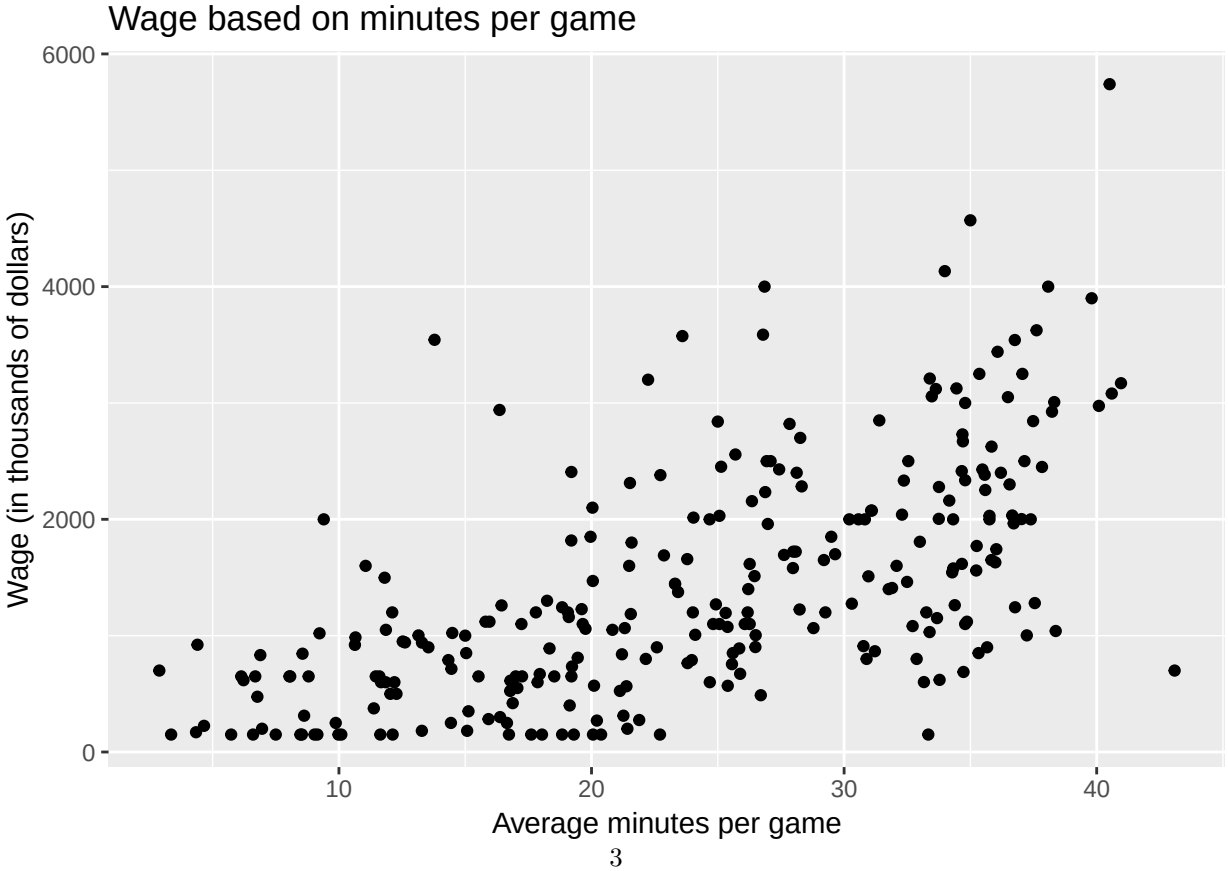
Table 1: Summary statistics of NBA player performance and salaries

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
wage	269	1,423.828	999.774	150.000	650.000	2,014.500	5,740.000
exper	269	5.119	3.400	1	2	7	18
points	269	10.210	5.901	1.200	5.400	14.200	29.800
rebounds	269	4.401	2.893	0.500	2.300	5.500	17.300
assists	269	2.409	2.093	0.000	0.900	3.400	12.600
avgmin	269	23.979	9.731	2.889	16.731	33.256	43.085
lwage	269	6.952	0.881	5.011	6.477	7.608	8.655
expersq	269	37.721	46.537	1	4	49	324

One thing to note here is that this data is from the 1993-94 season. After doing some research, I found the year that this data set matched up to (by using the max points, rebounds, and assists per game). Player salaries have increased greatly since then, so the maximum wage of 5,740 ($5.74 million) is much lower in this data set than it would be today. This points to another reason why we should be using log(wage) in our model, because a percentage increase will be much more helpful in our present day, compared to a nominal increase based on 1993's level of salaries and supply of money in the league.

6. Data plots


Wage based on points per game
Wage (in thousands of dollars)
6000
4000
2000
0
0
10
20
30
Points per game


Wage based on minutes per game
Wage (in thousands of dollars)
6000
4000
2000
0
10
20
30
40
Average minutes per game

Analysis: The first graph looks pretty much exactly as one would expect. There is a clear positive correlation between points per game and wage, as I mentioned earlier would be true. In fact, the NBA's highest paid player also scored the most points per game (David Robinson in this case). But what the graph cannot tell us is how strong the partial effect of points per game is on wages.

The second graph shows that there is also a positive correlation between minutes per game and wage, which I mentioned earlier as well. However, it is intersting to see that the player with the most minutes per game (Latrell Sprewell) was nowhere near being the highest paid player, and looks to be even below the average salary. The graph has a few distinct regions, one being the region of about 33 to 38 minutes played per game. This is where the team's starters would fall, and it shows that they generally have higher wages than the bench players, but there is still plenty of variation in the plot. So while there is a positive correlation, minutes per game cannot explain all of the variation in wage.

7. Regressions

Table 2: Unrestricted and restricted regression of points per game on wage

	Dependent variable:	
	lwage	
	(1)	(2)
points	0.038^{***}	0.037^{***}
	(0.014)	(0.014)
assists	-0.006	
	(0.029)	
rebounds	0.035	
	(0.021)	
exper	0.072^{***}	0.072^{***}
	(0.012)	(0.012)
avgmin	0.026^{**}	0.032^{***}
	(0.012)	(0.009)
Constant	5.429^{***}	5.425^{***}
	(0.119)	(0.115)
Observations	269	269
R^2	0.506	0.497
Adjusted R^2	0.497	0.492
Residual Std. Error	0.625 (df = 263)	0.628 (df = 265)
F Statistic	53.925^{***} (df = 5; 263)	87.353^{***} (df = 3; 265)
Note:		*p<0.1; **p<0.05; ***p<0.01

Understanding the regression: Since this is a log-level model, the estimates represent a percent change in wage. For example, $\hat{\beta}_4 = .072$ means an extra year of experience increases wage by 100*.072, or 7.2%, holding all other variables equal. It is also worth noting that the constants $\hat{\beta}_0$ or $\hat{\gamma}_0$ are not worth examining, because that would mean a player has zero experience, scores zero points and plays zero minutes. The regression table shows that besides $\hat{\beta}_2$ or $\hat{\beta}_3$, all OLS estimates have statistical significance at the 1% level. This means there is only a 1% chance we would incorrectly determine this result.

8. Analysis and conclusions

a. Testing the null hypothesis, $H_0 : \beta_2 = 0, \beta_3 = 0$

The question I wanted to examine was whether or not points per game was the only relevant on court statistic affecting wages. Based on the fact that $\hat{\beta}_2$ and $\hat{\beta}_3$ were not statistically significant, the null hypothesis seems to be true. However, we should do an F-test to confirm H_0 is valid.

$$F = \frac{(0.506 - 0.497)/2}{(1 - 0.506)/263} = 2.396 \tag{4}$$

At the 5% significance level, the critical value is 19.49. Therefore we cannot reject the null hypothesis, as we presumed. This means that two players scoring the same amount of points per game cannot affect their wage by getting more rebounds or assists. The only way to increase their wage is to score more points.

b. Preferred model

After performing the F-test it is clear that the preferred model is the restricted model, which removes the irrelevant variables *assists* and *rebounds*. Even when they were included, the point estimates of the relevant variables did not change that much. This shows us an example of the question we posed many times in class, regarding how including a random variable affects bias. This model confirms it will not bias the estimates, but will only make the unrestricted model inefficient.

c. Other analysis (referring to the restricted model)

The model shows that increasing points per game by just one point increases your wage by 3.7%, after controlling for experience and minutes per game. So increasing your points per game by 5 points, which is not unreasonable from one year to the next, would lead to an 18.5% increase in wage, a substantial amount. Since we used a percent change model, it is not unreasonable to think that this model using data from 1993 would still be fairly accurate in today's NBA. This explains some of the reason why the superstars of the league make way more money than the average NBA starter. Their wages are based strongly on their performance, meaning the best of the best are deserving of their huge salaries.