DA2 Term Project

Istvan Janco #2003877

01/03/2021

Abstract

The research is aimed to answer the question of whether and how performance metrics are affecting the salary of NBA players? The result is based on a data set of NBA salaries 1985 - 2018 and averages of player statistics (efficiency, number of games played, etc.). The main assumption is that the salaries and the performance stats are positively correlated. This analysis can be useful for general managers who are looking to extend contracts with existing players or sing new players. It suppose to provide an insight to what salary is to be expected for a player with certain characteristics and explore which players are over-valued and which players are under-valued.

1 Data

The data is restricted to a time frame of 2003 - 2018. The data was taken from the "data.world" website (https://data.world/datadavis/nba-salaries) and originally scraped from "basketballreference" webpage. The quality data is good, some missing values are present, however, there's not much possibility of systematic measurement error. Apart from removing the missing values, all the players who were on a payroll for less than 5 seasons were dropped from the dataset. The aim is to deal with random extreme observations, and establish a minimum base for comparison. The aim of the model is to compare players whose salaries are different due to performance metrics not due to the lack of time spent in the NBA.
In addition only, players who played a 100 (approx. 1 season) or more games while being under contract are qualified for the analysis. The reason is to exclude the possibility of extreme values due to lack of games played (e.g. 100% made free-throw shots over 20 games). The objective is to model average salaries on efficiency, by comparing players that are similar in some statistics but differ in terms of compensation. In order to conduct the analysis more effectively controls such as the number of games played and the draft in which a player was drafted were included. The table and figure below outlines the descriptive statistics and distributions of the aforementioned variables.

1.1 Variables Summary

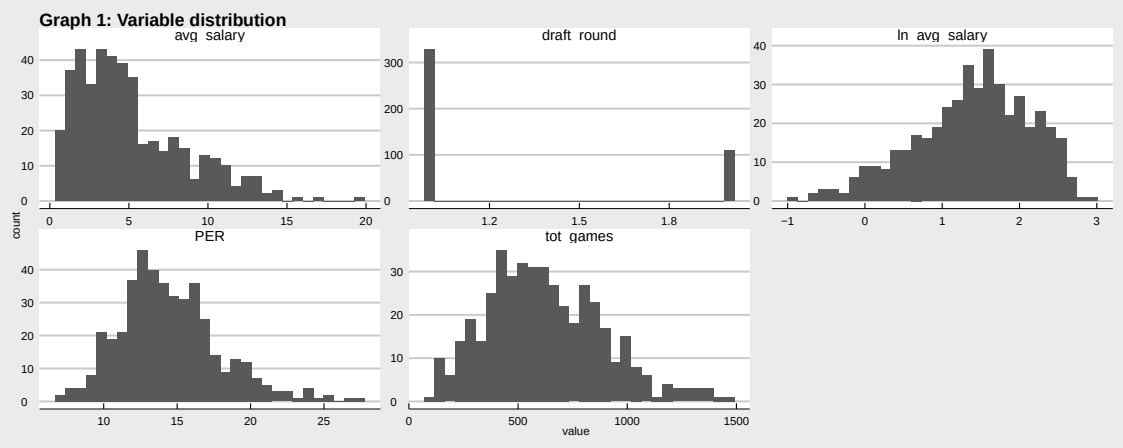
Table 1: Graph 1: Variable Summary

Average salary	PER	Tot games	Draft round
Min. : 0.4013	Min. : 7.00	Min. : 109.0	Min. :1.000
1st Qu.: 2.5120	1st Qu.:12.20	1st Qu.: 429.2	1st Qu.:1.000
Median : 4.3196	Median :14.15	Median : 603.0	Median :1.000
Mean : 5.1095	Mean :14.52	Mean : 624.9	Mean :1.251
3rd Qu.: 7.0576	3rd Qu.:16.38	3rd Qu.: 808.5	3rd Qu.:1.750

The outcome variable is the average salary of players per career. It was scaled down by $1M. The explanatory variable is the Player Efficiency Ratio (PER) which is per minute rating that was developed to measure overall player efficiency (see Appendix 1 for more detailed description and collinearity). The control variables

include the total games played for by a player for the whole career. In addition the draft round in which a player was selected is also incorporated.

1.2 Distribution of variables



The distribution of salary variable seems to be log normal, however, once log transformed the distribution starts skewing to the right. Other variables are distributed normally. The number of observations is 438.

1.3 Transformation of Variables

Once the variables were plotted against on another using scatter plots (see Appendix 2), the below points became evident.
1. The level-level model works best for the association between the average salary and PER. A quadratic term might be useful to capture the non-linearities, however it will complicates the interpretation. The benefit of it should be assessed relative to how complex the interpretation gets.
2. The simple model also works best for total games played, but there's an infection point, at about 500 games played, thus a spline with a knot at 500 can be used to tackle that issue 3. The draft round variable indicates a negative association between the draft round and average salary. The variable has only two categories thus it can be used as a dummy. The main reasons to use transformations are the following. From statistical perspective, the quadratic term for PER and linear spline for the total games played should help to offset the observed non-linear patterns. From substantive point of view, the interpretation should still be suitable for players as observations, despite introducing the quadratic term and adding a linear spline.

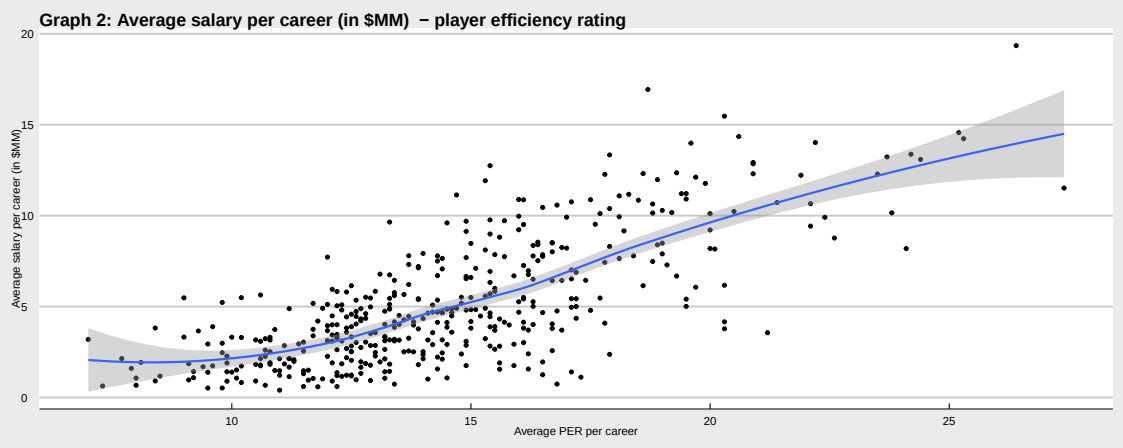
2 Model

The main aim is to regress the average NBA player salaries on player efficiency (see Appendix 2). By plotting the two variables against each other on a scatter plot, we can get a general impression of the association and the functional form.

Table 2: Modelling average NBA salaries on perfomance metrics

	Simple Model	Quadratic Model	Extended Model1	Extended Model2	Extended Model3
Intercept	−5.68900***	−2.21276	−6.41198***	−7.46500***	−7.07905***
	(0.47896)	(1.65077)	(0.43399)	(0.58158)	(0.61861)
PER	0.74385***	0.27715	0.62647***	0.62119***	0.60402***
	(0.03544)	(0.22925)	(0.03449)	(0.03432)	(0.03618)
PER_sq		0.01484			
		(0.00762)			
Total games			0.00388***		
			(0.00044)		
Total games <= 500				0.00675***	0.00681***
				(0.00106)	(0.00106)
Total games > 500				0.00298***	0.00284***
				(0.00062)	(0.00062)
Draft round					−0.54006*
					(0.23734)
R^2	0.55071	0.55652	0.62923	0.63433	0.63859
Num. obs.	438	438	438	438	438

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$



Graph 1 shows a upward trending, linear-like tendency. After and analyzing several models (see below), it was determined that the connection can be captured by using the following model. $AverageSalary = /beta_0 + \beta_1 * PER + \beta_2 * TotalGames * 1(TotalGames =< 500) + \beta_3 * TotalGames * 1(TotalGames > 500) + \beta_4 * DraftRound$

The model was chosen due to its ability to establish interpretation that works properly for players, despite introducing a linear spline and a dummy variable. In addition, the magnitudes of coefficients seem to be meaningful. From statistical perspective, the adjusted R^2 is the second highest among the models plus most of the variables are highly significant. The sample is also rather small (less than 500), thus a highly "tailored" model can lead to overfitting (see Appendix 3 for model fit).

2.1 Hypothesis Testing on Betas

The test hypothesis is the following $H_0 : \beta = 0$, $H_A : \beta \neq 0$ or not in our model. The estimated t-statistics for PER is 16.7, with p-value: $1.1432635 \times 10^{-48}$. The t-statistics for total games played under 500 is 6.41, with p-value: $3.8841436 \times 10^{-10}$, while for above 500 the t-statistics is 4.62 and the p-value is 5.0481785×10^{-6} Lastly, the draft round has a t-statistics of -2.28 and a p-value of 0.0233641 Choosing a significance level of p = 0.05. Based on the above, the H_0, can be rejected. This means that the average salary per career of

an NBA player is not uncorrelated with PER, total games played and draft round in which the player was selected. Based on the p-value being less than the significance level, the conclusion can be made that the sample data provides enough evidence to reject the null hypothesis. Changes in the independent variable are associated with changes in the dependent variable and control varailbes.

2.2 Residual Analysis

Table 3: Largest Negative Errors

Name	Average salary	Predicted salary	Residual
Alonzo Mourning	3.564590	10.090081	-6.525490
Marreese Speights	2.574359	6.993779	-4.419421
Brandan Wright	3.776013	8.095254	-4.319241

Table 4: Largest Positive Errors

Name	Average salary	Predicted salary	Residual
Gilbert Arenas	13.98695	7.770260	6.216687
Stephon Marbury	16.93594	8.602791	8.333147
Chandler Parsons	11.13643	4.220341	6.916088

After analyzing the above errors (Table 4 & 5), it evident that an increase of salaries over time, has an effect on the prediction of the model. It can be observed in the prediction with the larges negative error (Alonzo Mourning). It suggests that the issue might be attributed to the significant increase of the salaries from 2003 to 2018. However, in case of prediction with largest positive error observation (Carmelo Anthony) the issue might be lie in omitted variables. According to the model this player is overpaid, so perhaps in order to improve the prediction capability of the model, some additional variables (e.g. marketability) could be added. For more details on prediction uncertainty see Appendix 4.

2.3 Robustness analysis

In order to check whether the data is missing possible important patterns or if the conducted analysis is only true for this specific sample, the model was executed on an alternative sample of NBA salaries from 1988 to 2002. To provide an equal base for the comparison a subset of observations from the original data was used. It was sampled randomly from the original data on NBA salaries from 2003-2018.

The above summary suggests that the model is sample sensitive. One of the biggest differences is the R2. It looks like it's approximately 0.2 units lower for the new dataset. The total games played (500 >) spline explanatory variable has lost its significance. It indicates that the spline with a knot at 500 might not be a good predictor outside of the original dataset. In addition the RMSE is also smaller for the new data. This might indicate that in the period from 1988-2002 there was a smaller gap between players' salaries. As it was argued earlier some variables might be omitted (e.g. marketability), constant increase in salaries in the NBA also should be taken into account.

3 Causality and External Validity

Is there a possibility of causal relationship between player efficiency and salary? By creating a regression and adding the controls, a small progress has been made towards proving a possible causal relationship. However, once running the model on test sample of salaries and metrics from 1985 to 2002, it was clear that the model is sample sensitive. Perhaps the used functional form is not universal enough to uncover causal relationship between athletes' salaries and PER. In addition, PER is a composite metric, using weights to assign importance to statistics, like points scored, thus the stats are not treated equally, which might affect

Table 5: Model robustness check based on 2 samples

	Random Sample (original data)	Sample 1988-2002
Intercept	-7.01515^{***}	-1.44103^{**}
	(0.68628)	(0.52919)
PER	0.59727^{***}	0.27757^{***}
	(0.04051)	(0.03776)
Total games <= 500	0.00694^{***}	0.00044
	(0.00113)	(0.00065)
Total games > 500	0.00282^{***}	0.00096^{*}
	(0.00067)	(0.00042)
Draft round	-0.55899^{*}	-0.56908^{***}
	(0.25260)	(0.12321)
R^2	0.62874	0.39280
Num. obs.	389	389

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

the ability of the model to uncover causality. The analysis is also affected by possible omitted variables. More control variables should be added in order to bring the analysis closer to causality. These variables can include performance metrics such as win shares or more qualitative thing, like marketability of a player.

4 Summary

The relationship between NBA players' salaries and Player Efficiency was investigated. Two controls were added in order to account for an athlete's durability (Total games) and potential (Draft round). The functional form was altered using a linear spline (Total games) and a dummy variable (Draft round). Once running the model, it became evident that the continuously increasing salaries in the NBA can distort the predictions made by the model. For example, an athlete who played in the league during the beginning of the 2000s, would still get lower compensation than an athlete who competed during 2010s, even thought they can be similar on every other performance variable. There might be a possibility to offset this phenomena by applying the Purchasing Power Parity theory to the salaries and scaling them up to a reference year to normalize the value of the received compensation. These results might be useful for team managers who are trying to predict the possible market value of a prospective player.

Appendix

1. Variable description

- Outcome variable: average salary of NBA players for the career.
- Explanatory variable: Player Efficiency Ratio. Is a composite rating of a player's per-minute productivity. PER takes into account accomplishments, such as field goals, free throws, 3-pointers, assists, rebounds, blocks and steals, and negative results, such as missed shots, turnovers and personal fouls. The league average is 15.00 every season. It is important to mention that the PER might be more offense focused, thus some players who are better defenders, might be underrated.
- Control variable 1: Total games played, aiming to measure a player's durability.
- Control variable 2: Draft round in which a player was selected. The goal of the variable is measure potential of a player to earn more.

1.1. Collinearity

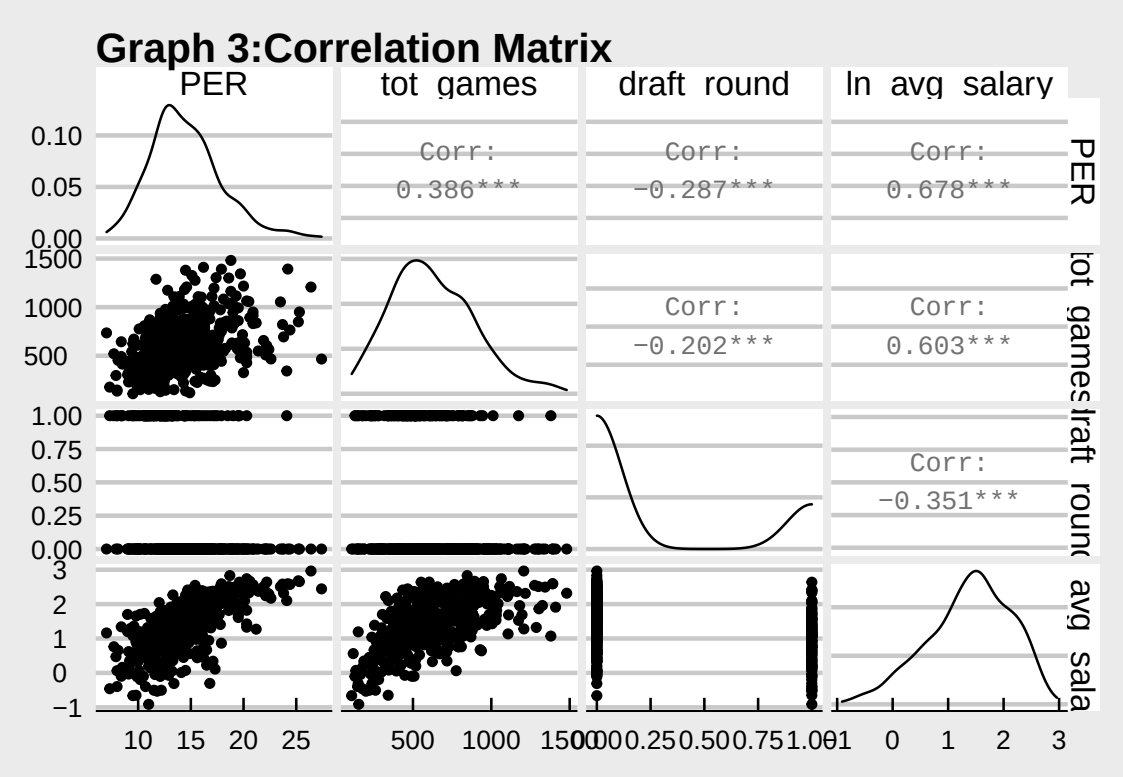
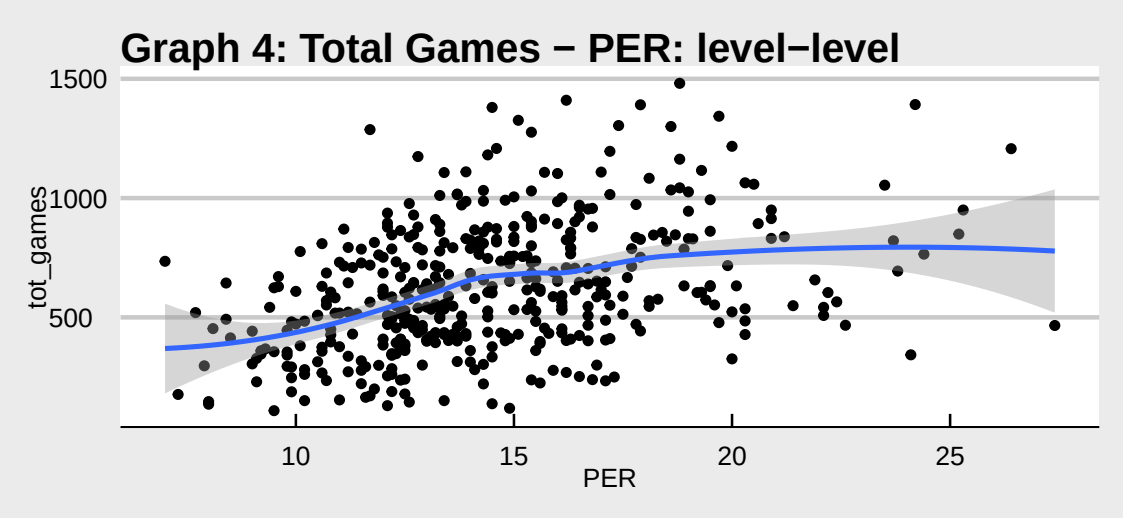


Table 6: Pairwise partial correlation coefficients

	PER	Tot games	Draft round	PERˆ2
PER	1.0000000	-0.0395034	-0.0705748	0.5782913
Tot games	-0.0395034	1.0000000	0.0104785	0.4961110
Draft round	-0.0705748	0.0104785	1.0000000	-0.1981077
PERˆ2	0.5782913	0.4961110	-0.1981077	1.0000000

Overall, it looks like there in no multicollinearity in the data. However we can check the following. The matrix suggests that there might be a correlation between the PER and the number of games played: level-level model without scaling



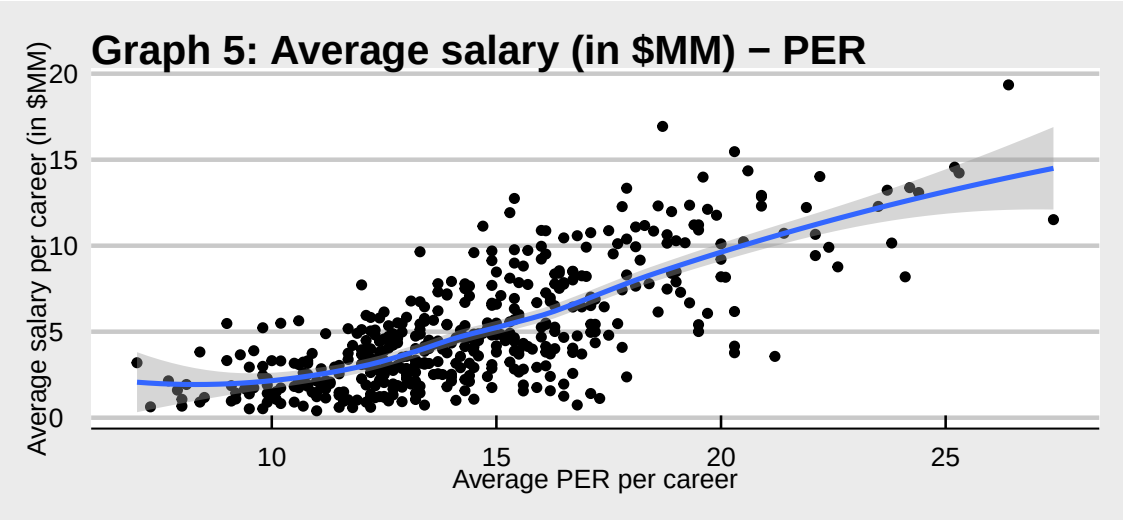
There might be a positive association between the number of games played and PER, however after PER = 14, it tends to be more and more flat. Collinearity is not expected between these variables.

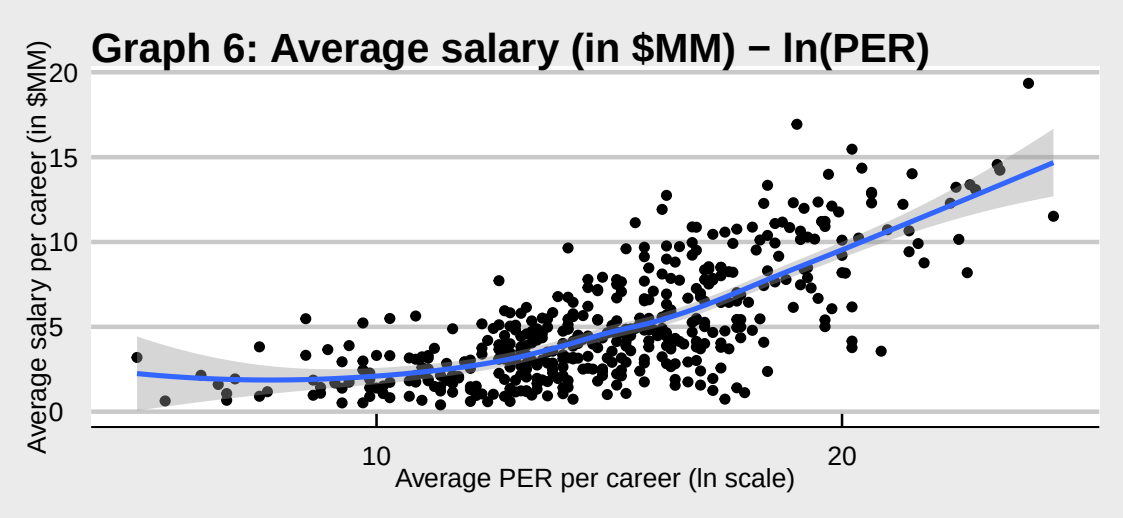
We can safely assume that there is no collinearity between the draft round and the number of games played. Draft round measures the potential of a player coming to the league, while the number of games played should measure the longevity of a player. It can be affected by many other factors such as injuries

No collinearity should be assumed for draft round and PER. Draft round is reflecting the accomplishments of a player before coming to the NBA, while PER is accumulated throughout a player's career.

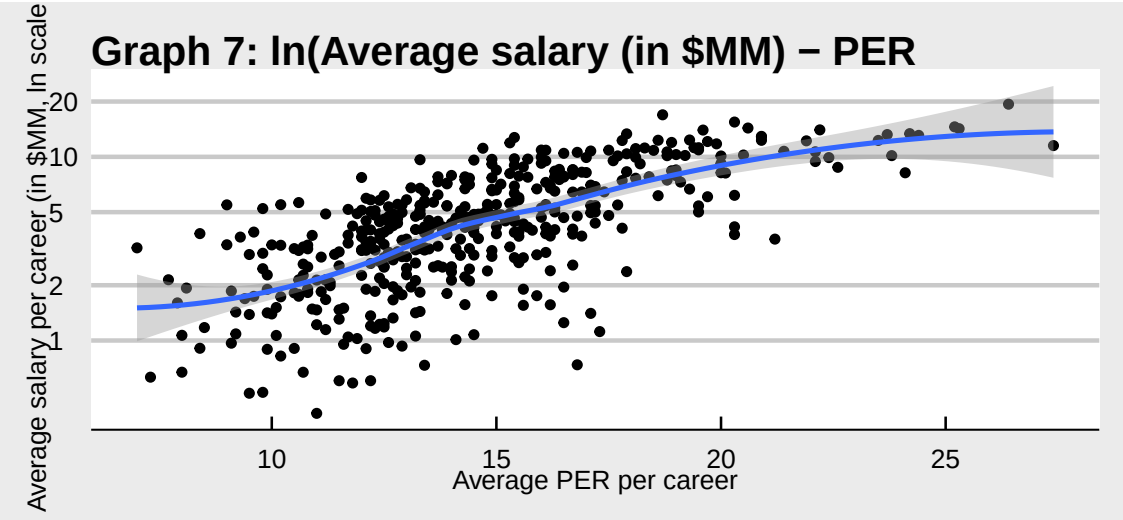
2. Distributions based on variable transformation

2.1. Average salary - PER: level-level model without scaling

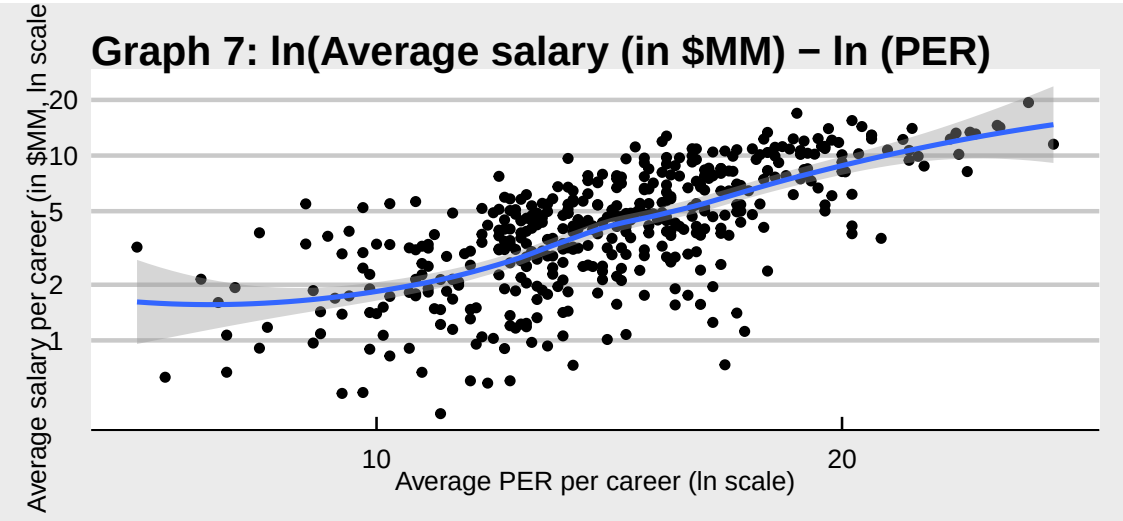




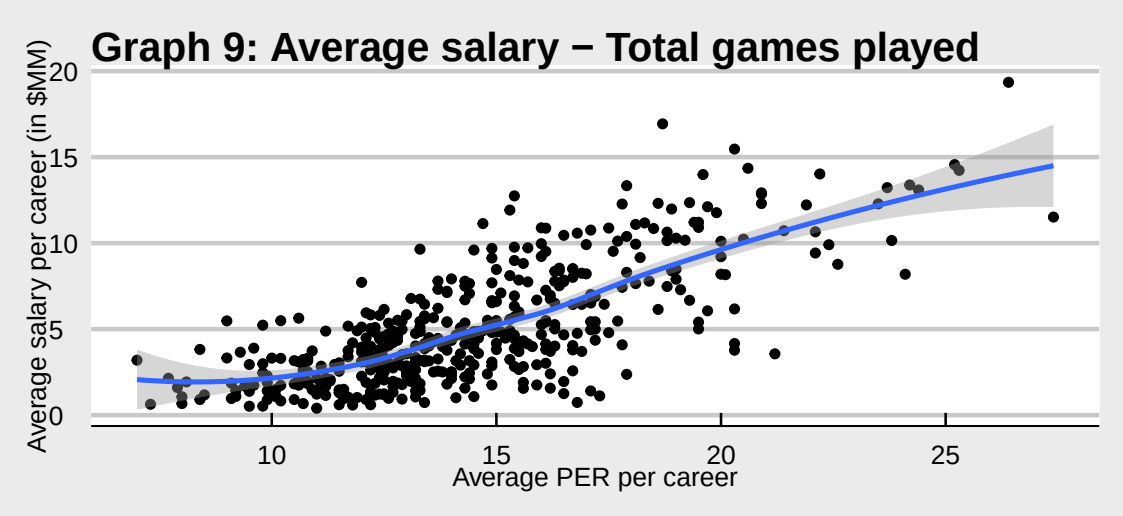
2.3. **3) ln(Average salary per career (in $MM)) - Average PER per career: log - level transformation**



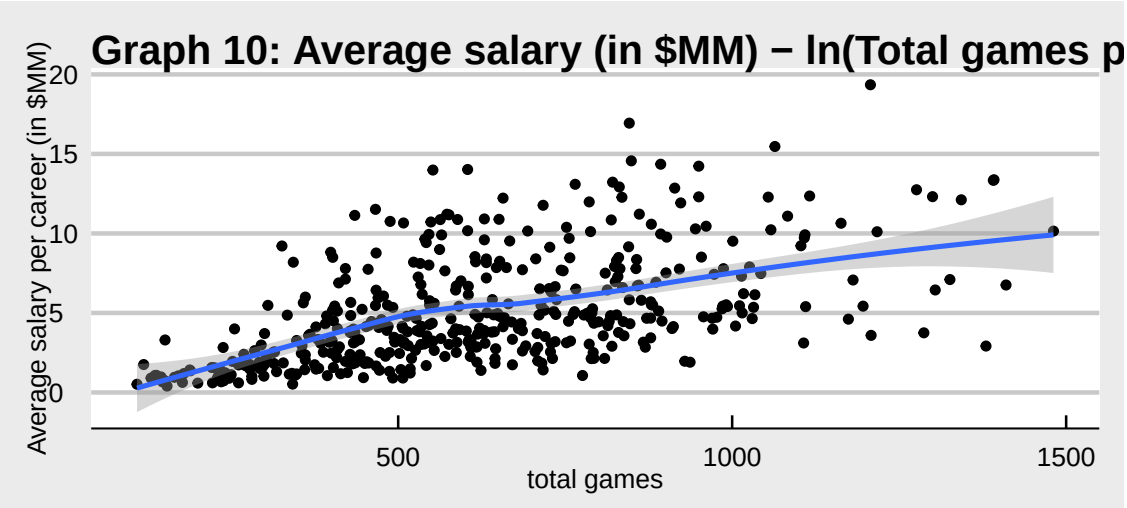
2.4. **# 4) ln(Average salary per career (in $MM)) - ln(average points scored per career): log log transformation**



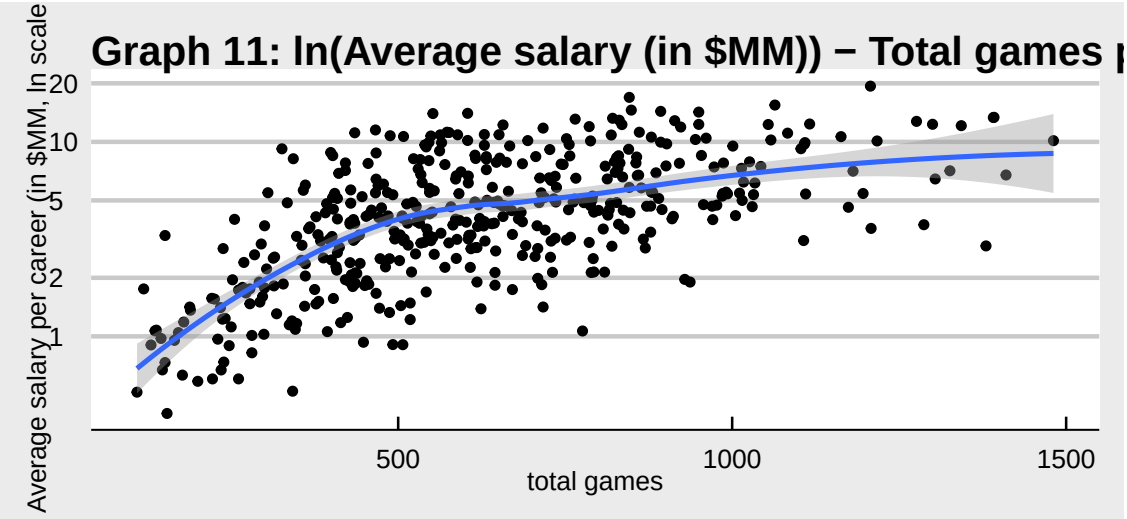
2.5. **Average salary - total games played : level-level model without scaling**



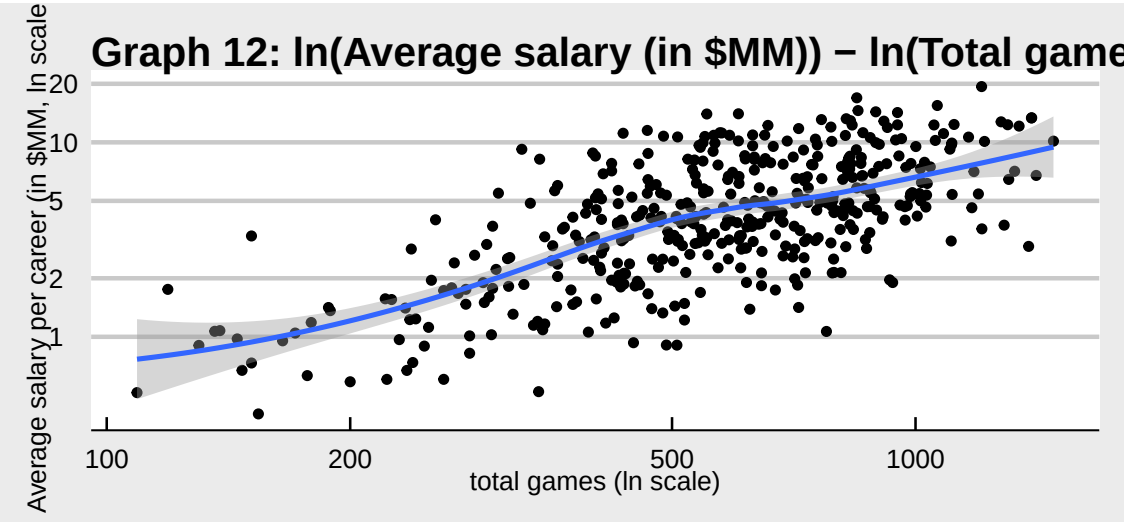
2.6. **Average salary per career (in $MM) - ln(total games played): level-log transformation**



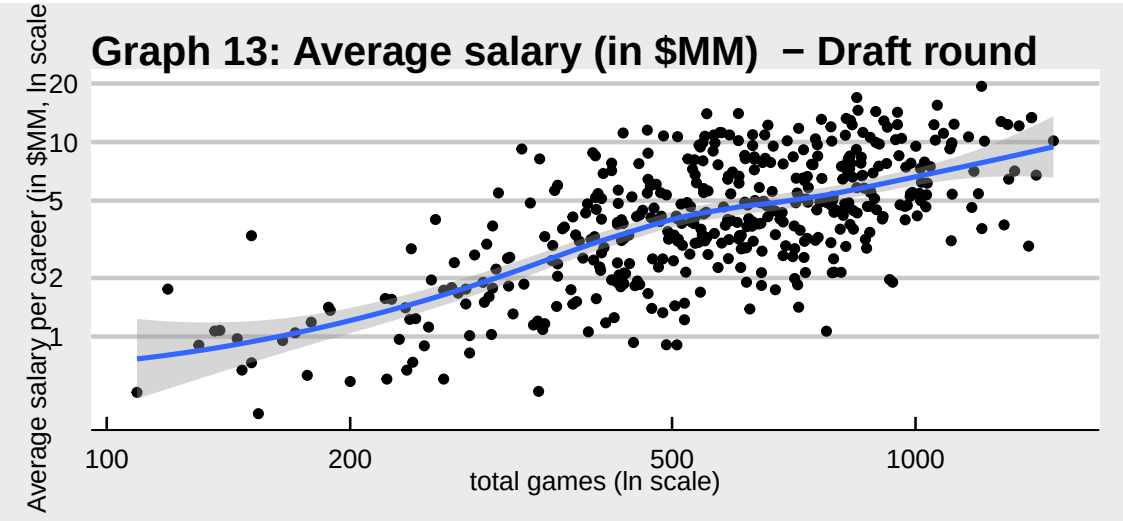
2.7. **ln(Average salary per career (in $MM)) - total games played : log - level transformation**



2.8. **ln(Average salary per career (in $MM)) - ln(total games played): log log transformation**



2.9 **Average salary per career (in $MM) - draft round: level-level model without scaling**



Conclusions on varaible transformations:

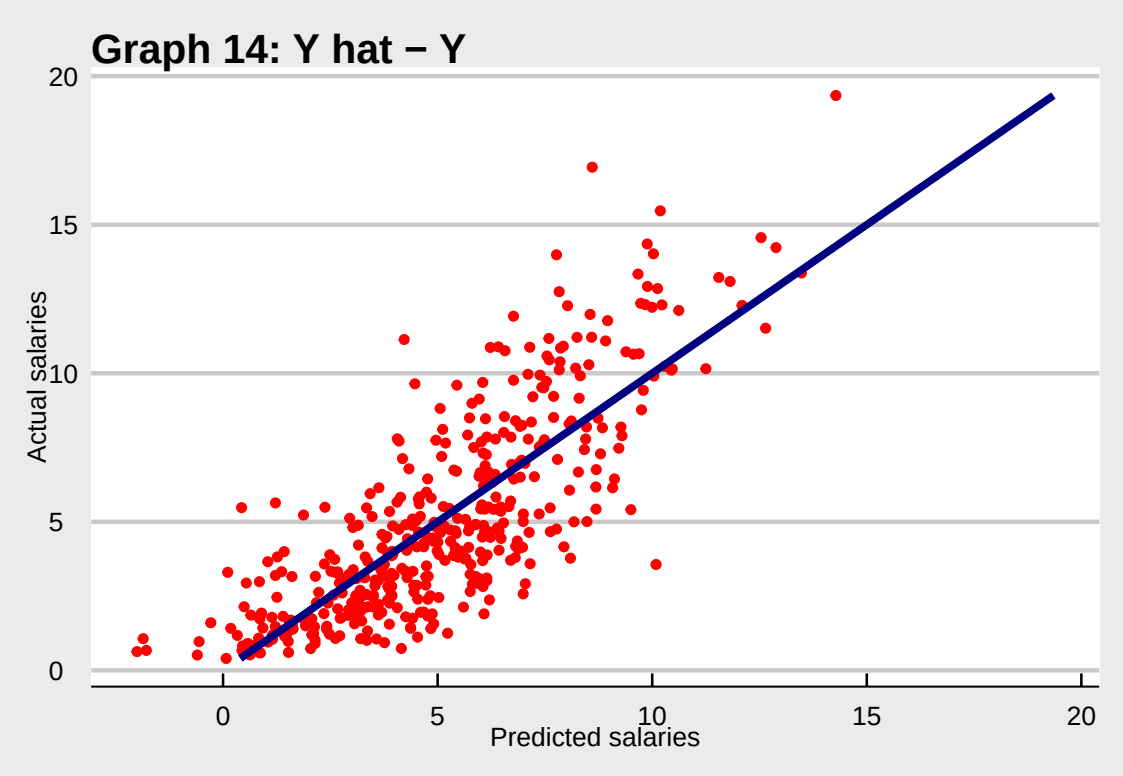
1) The level-level model seems to work best for the association between the average salary and PER. A quadratic term might be useful to capture the non-linearities, however it will complicate the interpretation. The benefit of it should be assessed relative to how complex the interpretation gets.

2) The simple model also works best for total games played, but there's an infection point, at about 500. A spline should be introduced to account for that.
3) The draft round variable indicates a negative association between the draft round and average salary. The variable can be used as a dummy.

Reasons to use transformations

1) Statistical: the quadratic term and spline should help to offset the observed non-linear patterns. In case of PER (quadratic term addition) and total games played (linear spline)
2) Substantive: while introducing the quadratic term and adding a linear spline, the interpretation should still be suitable for players as observations.

3. Model fit



The above plot suggests that the model is fairly good in predicting salaries, however there are still some errors, which might be attributed to possible omitted variables. In addition, we can see that the model considers a number of players to be overvalued.

3.1. BIC and AIC measures of model fit

Table 7: Bayesian Information Criterion

	df	BIC
Model 3	4	1908.095
Model 4	5	1908.107
Model 5	6	1909.054

The BIC suggests that by adding variables the model models is less likely to be true.

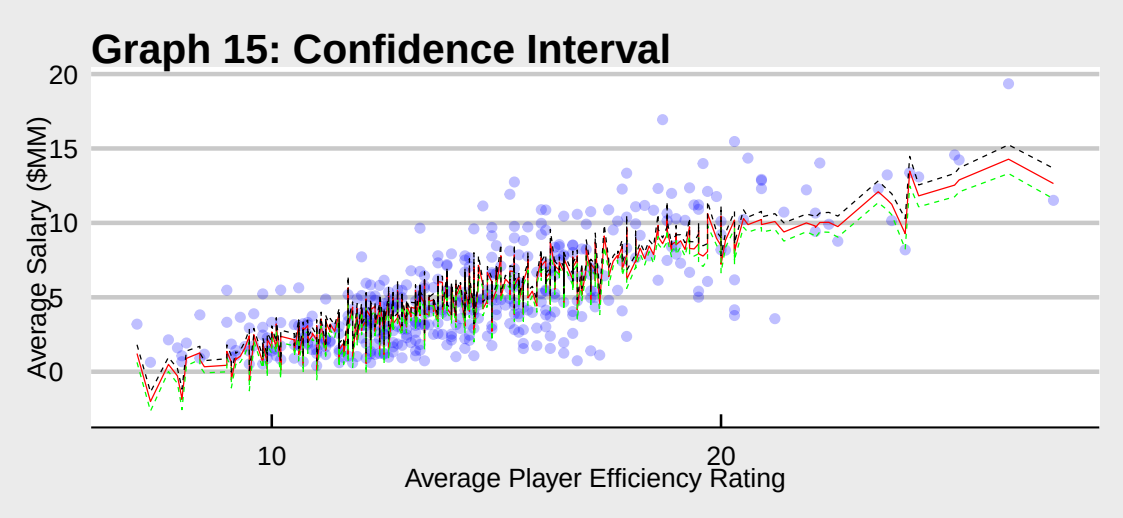
Table 8: Akaike Information Criterion

	df	AIC
Model 3	4	1891.766
Model 4	5	1887.696
Model 5	6	1884.561

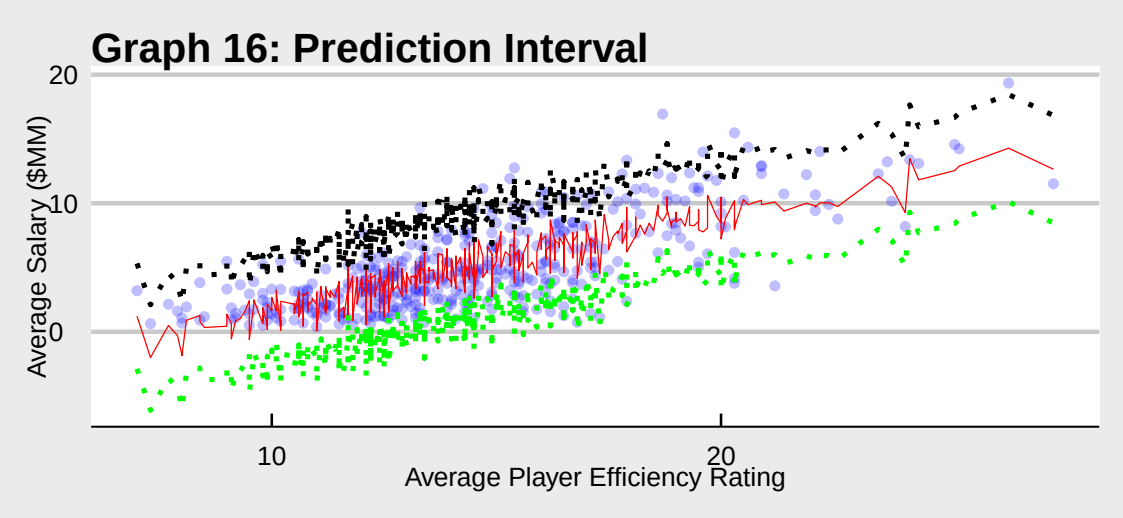
On the other hand the AIC decreased, once control variables were added, thus the added variables might actually be good controls.

4. Prediction uncertainty

4.1. Confidence Interval



4.2. Prediction Interval



The above suggests that the model might need calibration.