Lovicu, Gian-Piero
Aguilar, Ivan

Measuring NBA Team Success: Salary Concentration and Player Performance

1. Introduction

In the US film 'Moneyball' a coach takes a struggling Major League Baseball team on a record winning streak by choosing a team of players based on statistical analysis and value for money. The lesson from the film is that a team does not necessarily need to be filled with (expensive) superstars to find success. Taking inspiration from the film and the broader concept of making decisions in sport based on data analysis (a concept called sabermetrics popularised by baseball statistician Bill James), we analyse a set of National Basketball Association (NBA) data over a period of 12 years (2005-2017) to consider the drivers of team success. Specifically, we are interested in whether a player's salary is a reasonable indicator of their contribution to a team's success and whether the highest earning players in a team are the most important contributors. We consider the following questions:

- Which players and performance metrics are most important for explaining a team's success? Does the performance of higher paid players tend to matter more for a team's success?
- Do teams that spend a larger share of their salary pool on their top players win more games? Or are teams with a more balanced distribution of salaries more successful?

The performance contributions of the two or three most highly paid players in the team appear to matter most for a team's success. Offensive performance metrics such as field goal percentage, points scored and assists tend to contribute most to team wins and their contribution tends to increase in a player's share of the salary pool. In addition, there is some evidence that higher turnovers, including outside the most highly paid players in a team, is correlated with teams that win fewer games. In addition, teams that concentrate their salary expenditure on fewer players tend to win fewer games on average though the magnitude of this effect is not particularly large. Together these findings support the notion of the 'big three' in basketball, where a team's success is reflected mostly in the performance of the two or three best players in the team.

To conduct this analysis, we utilise several methods from the course including: K-means clustering, penalised likelihood, Bayesian model selection and robust regression techniques (quantile regression and mixed effects models). The input to these algorithms is a combination of player performance and salary data and we use these covariates to explain and predict the number of games a team wins in a season and a player's contribution to those wins (as measured by player win shares, see more in the data section).

2. Related Work

The related studies we considered mainly focused on game-level data inputs to predict game outcomes, either for regular season, playoffs or simulations. Studies tend to report an accuracy of between 0.6 and 0.7, reflecting the inherent uncertainty in sports outcomes. In our case even though we will develop the capacity to predict season wins and player win shares, that will not be our main objective because we will focus on identifying the most successful configuration for salary share. In terms of methods and techniques, we were able to find ample materials on machine learning approaches using Markov chains, decision trees and simple regression models, but did not find approaches that used regularisation modelling techniques.

Maximum entropy method is used as a non-regression approach combined with the use of k-means clustering which we will also be utilising as an initial heuristic in our study (Nasser). Nasser also compared more traditional prediction methods, such as neural networks and decision trees predictions .

In some cases, papers predetermined the features used in models to approximate game outcome prediction. Some studies also added player salary as a covariate, as in our case (Gonzalez Dos Santos et al.). Gonzalez Dos Santos et al also explored the concept of the big-two or big-three in reference to the most important players of each team, which we also consider: *Due to the top players' significantly impacting the outcome of games, many NBA teams prioritise trying to recruit two or three top* *players to their roster. These players are often referred to as the "Big Two" or the "Big Three", and are generally considered the most important players for team success.*

Studies have also considered the autocorrelation in game outcomes by building hidden Markov chains and analysing hot/cold streaks together with general sabermetrics (Madhavan). Others have not explored game outcomes directly, but a total winning percentage, which uses classic metrics and regression models and shares inspiration with our work from the 'Moneyball' movie and sabermetrics (He et al.)

3. Dataset

We collect NBA data from three sources: performance metrics, win shares and salaries for each player across all NBA teams across 12 seasons (2005-2017).[1] These data sets were publicly available to download from third party websites (see references for links), though the data all originally come from the nba.com/stats and basketball-reference.com websites. The data are joined by player, team and season to obtain a performance and salary profile for each player in each season, also capturing players that played in multiple teams. Table 1 summarises the features of our data.

The variable of interest is the number of games won by a team, which we consider two measures for. The first is the total games won by a team in a season. The second is a metric called 'win shares', which allocates shares of each game a team wins to the players involved (this is a complex calculation based on a similar measure developed by Bill James for baseball). The sum of win shares for each player in the team roughly adds to a team's total wins for the season. With these two variables of interest, we can approach our questions from the perspective of the team (overall team wins) individual players (player win shares within a team). The season wins data looks approximately normally distributed (Chart 1), though for win shares there is a long right tail of high win shares (Chart 2).

Chart 1. Histogram of season wins

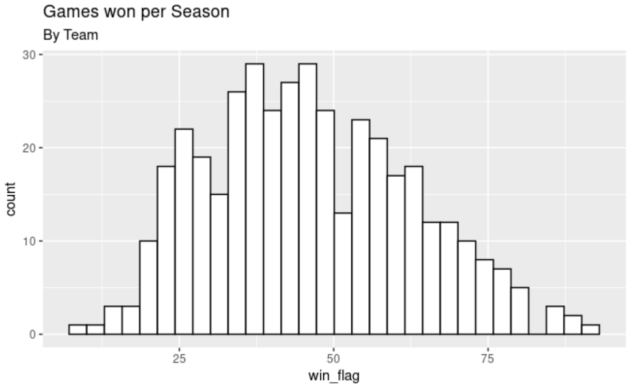
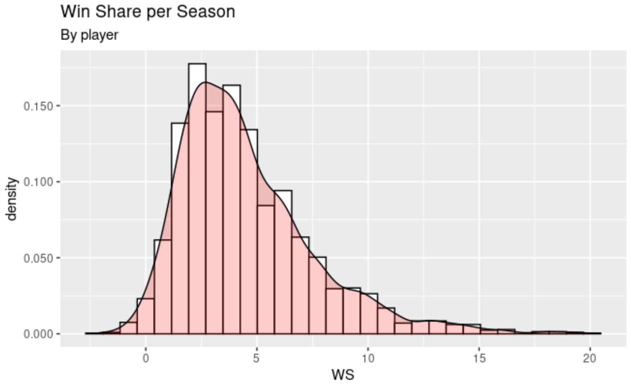


Chart 2. Histogram of win shares



Our performance covariates consist of a relatively set of basic performance metrics covering offence, defence and shooting accuracy. Although the NBA also provides a range of composite summary metrics aggregating many different aspects of player performance (e.g. plus-minus boxscore, player efficiency rating, etc.), because of their complexity these are challenging to interpret from within a model. Though only considering only simple metrics may come at the expense of some predictive accuracy (since they contain less information than composite metrics), they help us to more clearly consider our hypothesis.

The salary data consists of several measures of player and team salary. By player, the data has the team salary pool and a player's share of the overall team wage bill. A player's share of the wage bill is the proxy we use to signal a player's standing in the team. At the team level, we consider the overall wage bill and a measure of concentration used widely in economics (the Herfindahl-Hirschman Index or HHI). The salary HHI helps to identify teams that spend a large portion of their budget

[1] Teams that changed cities/franchises during our sample, we map them to their latest city/franchise. For example, in 2012 the New Jersey Nets became the Brooklyn Nets, but we have corrected this so that the team is labelled as the Brooklyn Nets across the whole sample. This is relevant for the section on hierarchical modelling.

on relatively few players. Chart 3 shows there is a negative correlation between the salary HHI and the games won in each season. However, we can also see a number of outliers in the data - teams with a highly concentrated wage bill that don’t win many games. Indeed, Chart 3 shows that the HHI data have a long right tail. So although this provides some initial indication that high salary concentration is correlated with poorer performance, we must also be aware of these outliers in the data.

Table 1. Performance and salary metrics

Name	Type	Description
season_wins	-	Number of games won by a team in a season
WS	-	Win shares; an estimate of the share of team wins contributed by individual players. See Calculating Win Shares for more information.
PTS	Off	Points (per 10 mins played)
AST	Off	Assists (per 10 mins played)
OREB	Off	Offensive rebounds (per 10 mins played)
BLK	Def	Blocks (per 10 mins played)
DREB	Def	Defensive Rebounds (per 10 mins played)
STL	Def	Steals (per 10 mins played)
TO	Off/Def	Turnovers (per 10 mins played)
FG_PCT	Shoot	Field Goal Percentage; the formula is FG / FGA (goals/attempts).
FT_PCT	Shoot	Free Throw Percentage; the formula is FT / FTA (free throws/attempts.
FG3_PCT	Shoot	3-Point Field Goal Percentage; the formula is 3P / 3PA (goals/attempts).
salary	Salary	Player’s salary for each season (for players who switched teams during the season, we carry back their salary at their new team for the following season)
salary_share	Salary	Player’s share of overall wage bill (top 8 players)
salary_total	Salary	Teams wage bill in a given season (all players)
salary_hhi	Salary	Salary Concentration :Herfindahl-Hirschman Index: $hhi_i = \sum_{p=i}^{8} (salary\_share_p)^2$

Chart 3. Salary concentration and seasons wins

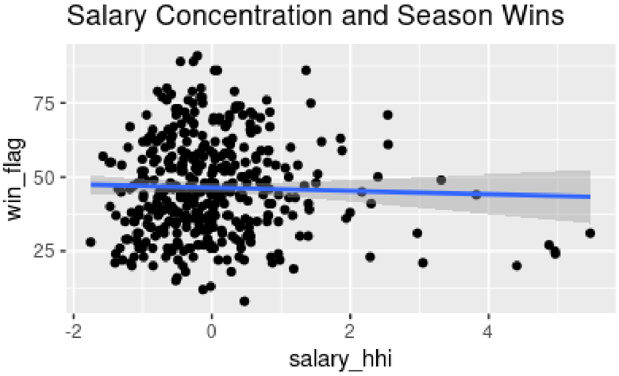
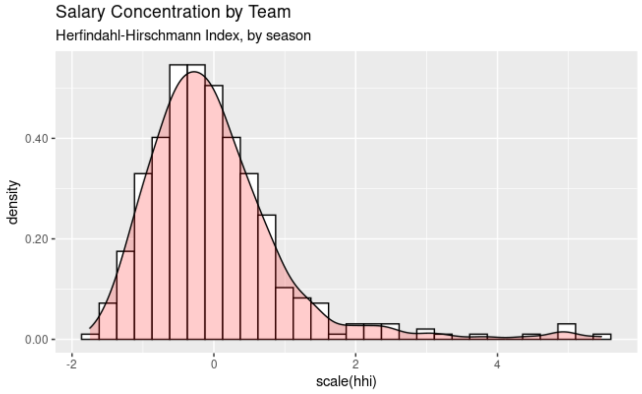
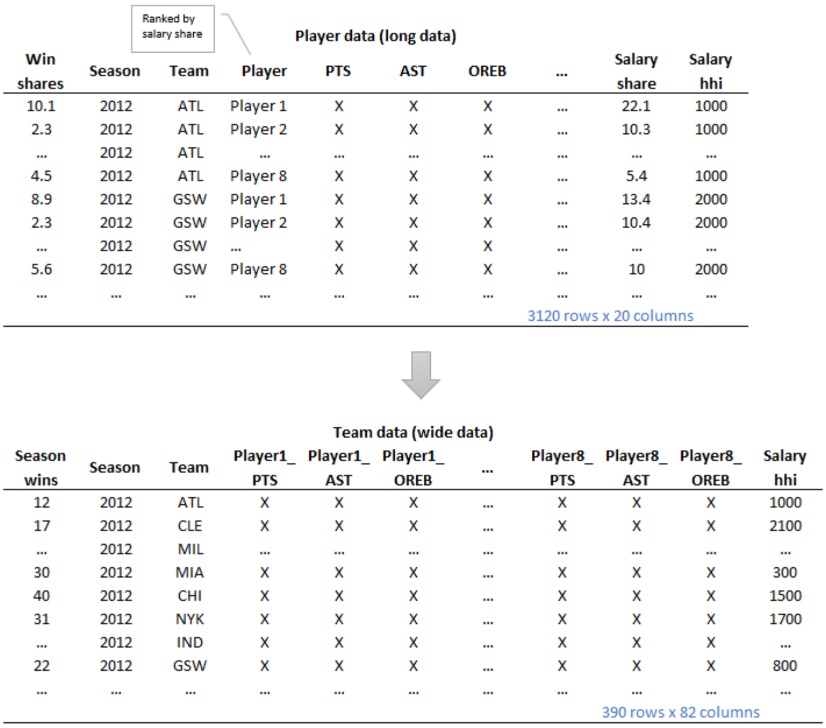


Chart 4. Salary concentration by team



These data form two data frames: a player-level data frame (a row for each player in each team and season) and a team-level data frame (a row for each team in each season). For the player-level dataframe, our variable of interest is player win shares. We filter the data for 8 players who played the most minutes for each team in each season (this covers 75-80 per cent of total minutes played, on average). We then rank these players 1 to 8 by the share of the salary pool they receive (player 1 receives the highest salary share, player 8 the lowest). To obtain the team-level dataframe, we pivot the performance metrics by the player’s salary rank and our variable of interest becomes season wins for each team. Figure 1 below shows the relationship between the player and team-level dataframes. In each data frame all covariates are scaled to have zero mean and unit variance to enable us to compare the coefficients in our models.

Figure 1: mock-up of our player- and team-level data frames



Ranked by salary share

Player data (long data)

Win shares	Season	Team	Player	PTS	AST	OREB	...	Salary share	Salary hhi
10.1	2012	ATL	Player 1	X	X	X	...	22.1	1000
2.3	2012	ATL	Player 2	X	X	X	...	10.3	1000
...	2012	ATL	...	...	...	...	...	...	...
4.5	2012	ATL	Player 8	X	X	X	...	5.4	1000
8.9	2012	GSW	Player 1	X	X	X	...	13.4	2000
2.3	2012	GSW	Player 2	X	X	X	...	10.4	2000
...	2012	GSW	...	X	X	X	...	...	...
5.6	2012	GSW	Player 8	X	X	X	...	10	2000
...	...	...	...	...	...	...	...	...	...

3120 rows x 20 columns

Team data (wide data)

Season wins	Season	Team	Player1_ PTS	Player1_ AST	Player1_ OREB	...	Player8_ PTS	Player8_ AST	Player8_ OREB	Salary hhi
12	2012	ATL	X	X	X	...	X	X	X	1000
17	2012	CLE	X	X	X	...	X	X	X	2100
...	2012	MIL	...	...	...	...	...	...	...	...
30	2012	MIA	X	X	X	...	X	X	X	300
40	2012	CHI	X	X	X	...	X	X	X	1500
31	2012	NYK	X	X	X	...	X	X	X	1700
...	2012	IND	X	X	X	...	X	X	X	...
22	2012	GSW	X	X	X	...	X	X	X	800
...	...	...	...	...	...	...	...	...	...	...

390 rows x 82 columns

4. Data analysis

This section consists of some exploratory clustering on the player-level data frame, penalised likelihood and bayesian model selection on the team-level dataframe and finally robust regression (quantile regression and hierarchical modelling) on the player-level dataframe. We also conduct OLS regressions on the player- and team-level dataframe as a baseline to compare with our other methods. To compare models we use root mean squared error (in- and out-of-sample), normalised by the mean of our predictor. Out-of-sample predictions are made using 10-fold cross-validation. Where appropriate, R-squared is also reported for our models. For each model, we will consider the implications for both of our research hypotheses.

Exploratory clustering (team-level dataframe)

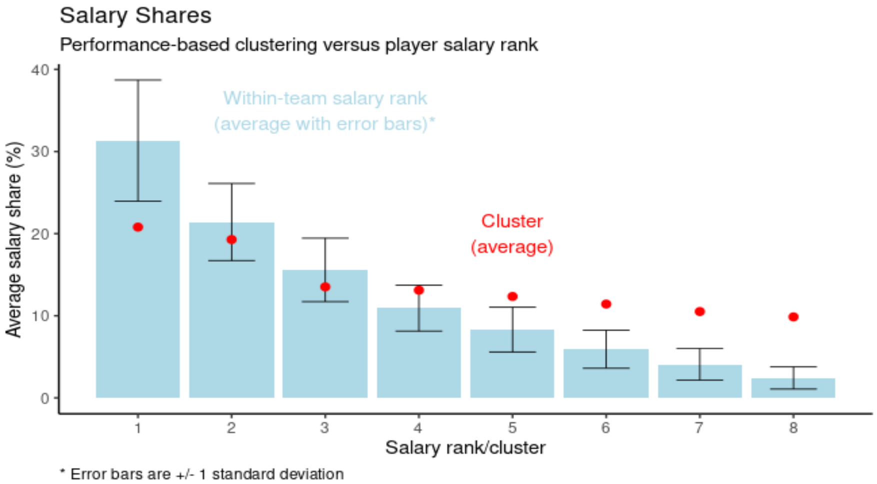
A key assumption in our analysis is that a player's share of the team salary pool is a reasonable measure of their standing in the team. Indeed, we have constructed our team-level dataframe by ranking the players based on salary share and organising them across teams and seasons (i.e. compare all players with the highest salary share in each team and season). To consider whether this ranking is a reasonable approach, we assign players to 8 clusters using K-means, based only on their performance metrics (basic performance data from Table 1 and win shares) and look at the average salary shares within each cluster.

Comparing the average salary share in each cluster with the average salary share in each player rank (we treat the latter as our 'true' label) can then shed light on whether ranking players by their salary share is an appropriate approximation of their standing in a team. This is of course an imperfect comparison, since we cannot impose that each cluster has the same number of players, or that each player within a team/season be allocated to a different cluster. However, if we see that some clusters exhibit a higher salary share than others (without explicitly passing the clustering algorithm any information on salary) this provides some confidence that pivoting performance data by a player's salary rank is a reasonable way to construct our team-level dataframe.

Chart 5 shows a summary of this exercise. The blue bars show the average salary share for each rank within a team/season (with error bars for +/-1 standard deviation) and the red dots represent the average salary share in each cluster. The top 4 clusters clearly correspond with a higher average salary share. The magnitude within these clusters is also comparable to the average salary share in ranks 1 to 4. Spot checks also indicate that many of the highest and second highest paid players (we looked at MVPs across a few seasons) end up in the top 2 clusters. This gives us confidence that among the higher paid players at least, ranking by salary share is a meaningful exercise.

In the lower ranks, the clustering doesn't exhibit much difference by salary share and this is not altogether surprising because the separation of these players by salary is marginal compared with the higher ranks. This could indicate we might not expect much variation in the contribution of 'middle-of-the-pack' players to a team's performance and we will explore this idea further below.

Chart 5. Performance based clustering vs player salary rank



Team-level data - regression models

In this section, we will use regularisation techniques on our team-level data frame to consider which players (ranked by salary share) and performance metrics are most highly correlated with a team's success. First, we consider a penalised likelihood model (using an L1 penalty) and then to supplement this we also conduct Bayesian model selection (BMS). Since the team-level data has around 80 covariates and only 400 observations, regularisation can help us to identify which variables appear most important for explaining variation in a team's season wins (though we do not claim causality). We also make predictions of season wins with our models and report normalised RMSEs. However, since our data is contemporaneous (season wins are compared with performance and salary metrics from that same season) we are more interested in interpreting the coefficients from the models.

Summarising the results in the models that follow, we find some evidence that:

- The performance of the highest ranked players (by salary share) in general have the most explanatory power for season wins. The performance metrics the appear most important include the field goal percentage (FG_PCT), turnovers (TO), points (PTS), assists (AST) and blocks (BLK)
- When players just outside the top 3 turnover (TO) the ball more, this is correlated with fewer season wins for their team.
- Salary concentration (salary_hhi) is negatively correlated with season wins, though we require further evidence to support this given the outliers observed in these data.

Equation 1 shows the basic specification, which is the same across all the team-level models:

$$w_i = \alpha + \sum_{m=1}^{10} \sum_{p=1}^{8} \beta_{mp} x_{mpi} + \gamma hhi_i + \varepsilon_i \qquad (1)$$

where:

$w_i = season\ wins$

$X_{mp} = vector\ of\ observations\ for\ performance\ metric\ 'm'\ and\ player\ 'p'$

$m = (1, \dots, 10)\ (performance\ metrics\ from\ table\ 1)$

$p = (1, ..., 8)\ (players\ 1\ to\ 8,\ ranked\ by\ salary\ share)$

$i = (1, 2, ..., 380)\ (observations\ for\ each\ team\ in\ each\ season)$

$\varepsilon_i \sim N(0, \rho I)\ (gaussian\ iid\ errors)$

$\alpha = intercept$

Ordinary least squares

First, we run a baseline OLS model to compare with our other modelling approaches. Table 2 shows that the ordinary least squares model has an R-squared of around 0.6, though we see evidence of overfitting, as both the out-of-sample and adjusted R-squared measures are considerably lower. Considering normalised RMSEs on both the fitted values and cross-validated predictions tells a similar story. The full results of the model can be found in the model appendix.

Table 2: Model performance team-level data - Ordinary Least Squares

	In-sample	Out-of-sample	Adjusted
R-squared	0.58	0.33	0.47
nRMSE	0.23	0.30	

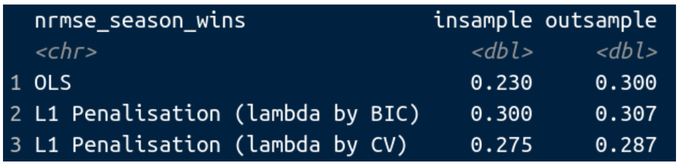
Penalised likelihood (L1 penalty)

Next, we undertake penalised likelihood on the team-level dataframe. The model specification is the same as for OLS, except in this instance we are optimising over a loss function that includes an L1 penalty. Since we are interested in interpreting the coefficients of the model, we take a conservative approach and set the regularisation parameter (lambda) according to the smallest Bayesian Information Criteria (BIC). Chart 6 shows the variables selected by this model: 12 out of the 80 or so covariates were selected (this includes an intercept not shown on the chart). The chart also includes confidence intervals (constructed using the method from Lee *et al* (2016)) and labels the variables that are statistically significant (see model appendix for full model).

The FG_PCT (field goal shooting percentage) of players 1, 2 and 3 all have a positive and statistically significant correlation with a team's season wins. The magnitude of the coefficient is also in the order we would expect: the higher a player ranks, the larger the coefficient. Interestingly, higher turnovers among some of the lower ranked players (player 4 and 5) has a large negative correlation with season wins. This could suggest that worse performance among 'middle-of-the-pack' players may contribute negatively to season wins.

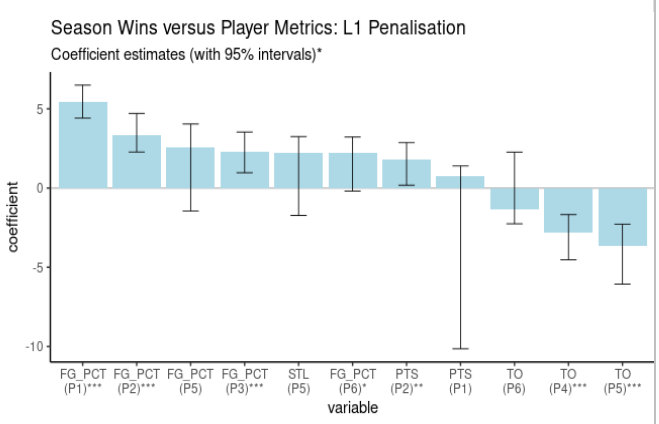
Table 3 shows that the L1 penalised likelihood model has a higher normalised RMSE than OLS, though the gap between the two is very small out-of-sample. This suggests that selecting fewer covariates has corrected some of the overfitting present in the OLS model, though its overall fit of the model is still pretty poor. If we set the regularisation parameter by cross-validation (instead of via BIC) the RMSE improves, which is typical when making predictions.

Table 3: Model performance team-level data - L1 Penalisation



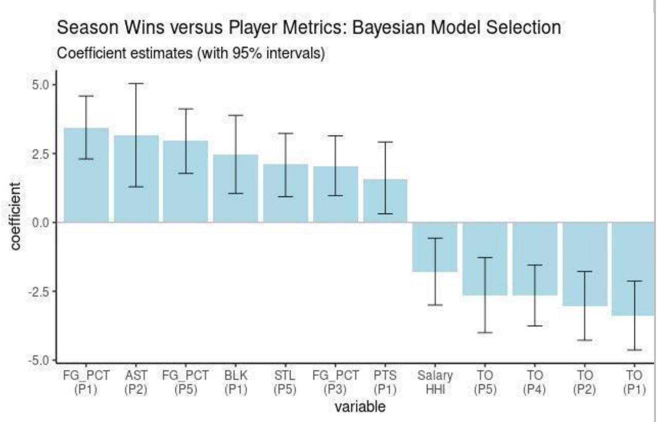
```
  nrmse_season_wins                    insample outsample
  <chr>                                   <dbl>     <dbl>
1 OLS                                     0.230     0.300
2 L1 Penalisation (lambda by BIC)         0.300     0.307
3 L1 Penalisation (lambda by CV)          0.275     0.287
```

Chart 6. Season wins vs player metrics for L1 penalisation



* Intervals calculated using post-selection inference method from in Lee et al (2016).
***, ** and * indicate statistical significance at 1, 5 and 10% levels.

Chart 7. Season wins vs player metrics for BMS



Only showing coefficients with marginal posterior inclusion probability above 90%

Bayesian Model Selection

Next, we run the same model using Bayesian model selection (BMS). Our assumptions to obtain the posterior include:

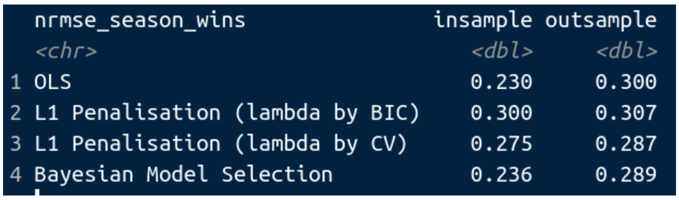
- A standard gaussian likelihood: $N(Y\ ;\ X\beta\ ,\ \rho I)$
- Conjugate gaussian (Zellner shrinkage prior): $p(\beta\mid\rho)\ =\ N(\beta;0,\ \rho g\left(X^{T}X\right)^{-1})\ where\ g\ =\ 0.01$
- An inverse gamma prior on the variance: $\rho\ =\ IG(0.01,\ 0.01)$
- A model selection prior: $beta\_binomial(1,\ 1)$

To set the value for the shrinkage parameter in our Zellner prior, we conduct prior elicitation. A low value of around 0.01 for this parameter implies a theoretical R-squared of around 0.5, which appears reasonable given the R-squared from the OLS regression. This leads to less regularisation than in the penalised likelihood setting and most of the posterior model probability (around 90%) is concentrated in a single model. The model convergence diagnostics demonstrate that the gibbs sampling algorithm converged successfully to a stationary distribution. Further details (including the charts) can be found in the model appendix.

Chart 7 shows the coefficients calculated using Bayesian model averaging with 95 per cent credible intervals. The plot only includes the variables with a marginal posterior probability of above 90 per cent (excluding the intercept). The results from this model shows fairly consistent results compared with L1 penalisation - there are large positive coefficients for FG_PCT across players 1 and 3 and large negative coefficients for turnovers (TO) for players 4 and 5. However, the BMA also suggests that other performance indicators across the top players are important for season wins, including points scored (PTS) and blocks (BLK) (player 1) and assists (P2). Interestingly turnovers (TO) also feature in the model for players 1 and 2. From a performance perspective, this emphasises that the performance metrics of the most highly paid players matter most for explaining season wins, including when they perform poorly (i.e. turnover the ball more). From the perspective of salary, we see that salary concentration is correlated with worse team performance, though before taking too much confidence in this result it is important to check that this is not being driven by the outliers we observe. The full model results are available in the model appendix.

The normalised RMSEs for the Bayesian model averaging do the best in predicting season wins across the team-level models, though not by a large margin.

Table 4: Model performance team-level data



```
  nrmse_season_wins                  insample outsample
  <chr>                                 <dbl>     <dbl>
1 OLS                                   0.230     0.300
2 L1 Penalisation (lambda by BIC)       0.300     0.307
3 L1 Penalisation (lambda by CV)        0.275     0.287
4 Bayesian Model Selection              0.236     0.289
```

Player-level data - regression models

This section uses robust regression techniques on our player-level data frame to support the results gleaned from the team-level data. First we run a regular OLS model as a baseline and then robustify it by running median regression and some mixed effects models to see if our results change. In each of these models, we interact each performance metric with a player's salary share to consider whether the contribution of performance metrics varies according to a player's salary share. We also include salary concentration to explore whether a more balanced distribution of salary is correlated with higher win shares across players in a team. Again, we are less interested in predictions than in interpreting the coefficients of our model, though we report normalised RMSEs for each model.

Equation 2 shows the basic specification for the model used in this section. Where appropriate we will discuss modifications to the specification as needed.

$$ws_i = \alpha + \sum_{m=1}^{10} \beta_m x_{mi} + \beta_{m2} x_{mi} z_{2i} + \sum_{s=1}^{3} \gamma_s z_{si} + \varepsilon_i \qquad (2)$$

where:

$ws_i = win\ shares$

$X_m = vector\ of\ observations\ for\ performance\ metric\ 'm'$

$m = (OREB,\ DREB,\ PTS,\ BLK,\ STL,\ TO,\ AST,\ FG\_PCT, FT\_PCT,\ FG3\_PCT)$

$z_s = vector\ of\ observations\ for\ salary\ metric\ 's'$

$s = (salary\_hhi,\ salary\_share,\ salary\_total)$

$X_m z_s = vector\ of\ observations\ for\ performance: salary\_share\ interactions$

$i = (1, 2, ..., 3120)\ (observations\ for\ each\ player, team,\ season\ combo)$

$\varepsilon_i \sim N\ (0,\ \rho I)\ (gaussian\ iid\ errors)$

$\alpha = intercept$

Median regression

In the data section, we saw that both player win shares and salary concentration have a number of positive outliers (i.e. players with abnormally high win shares and teams with high salary concentration). Since OLS assumes our data is approximately normally distributed, this feature of the data could be driving some of our results. To address this, we will re-run our model using median regression (with bootstrapped standard errors) to see if it changes our results. The model specification does not change, we simply optimise over a different loss function that penalises our outliers less than OLS.

Full details of the OLS and Quantile regression models (coefficients, etc.) can be found in our model appendix.

Hierarchical (mixed effects) models

Our data have a clear hierarchy: within seasons and within teams. OLS assumes that all of our observations are independent, but there could be trends within a season or within a team (or both). For example, the total salary bill increases over the seasons in our sample across all teams. In addition, the success of teams with a large salary budget could be different than teams with a small salary budget. As a result, we re-run our model using two types of mixed effects: one that varies the intercepts in a nested season:team structure (to capture the hierarchy in our data) and another that varies the slopes of the salary concentration coefficient by team and intercepts by season (to see if the relationship between salary concentration and win shares differs across teams in the league). The model specifications change as follows:

Equation 3: Nested intercepts

$$w_{si} = \alpha_{uv} + \sum_{m=1}^{10} \beta_m x_{mi} + \beta_{m2} x_{mi} z_{2i} + \sum_{s=1}^{3} \gamma_s z_{si} + \varepsilon_i \qquad (3)$$

$\alpha_{uv} = intercept\ varying\ by\ season{:}\ team$

$u = (2005, 2006, ..., 2016)\ (seasons)$

$v = (ATL, ..., UTA)\ (teams)$

Equation 4: varying salary concentration slopes and varying season intercept

$$w_{si} = \alpha_v + \sum_{m=1}^{10} \beta_m x_{mi} + \beta_{m2} x_{mi} z_{2i} + \sum_{s=2}^{3} \gamma_s z_{si} + \gamma_{1v} z_{1i} + \varepsilon_i \qquad (4)$$

$\alpha_u = intercept\ varying\ by\ season$

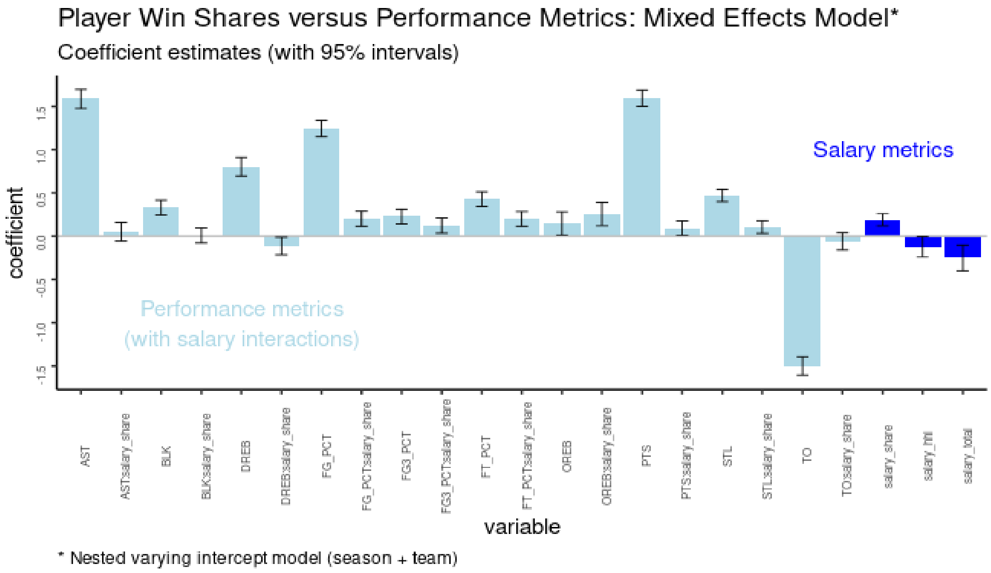
$u = (2005, 2006, ..., 2016)\ (seasons)$

$v = (ATL, ..., UTA)\ (teams)$

Results

The coefficient estimates are similar across OLS and each of our robust regression specifications. Chart 8 shows the results for the nested intercepts model (the best performer by RMSE), which places the performance coefficients side-by-side with their salary interactions to give an overall sense of the importance of that performance metric. Table 3 shows each coefficient and their statistical significance for each of the models. Adding the mixed effects improves the fit of the models (in and out-of-sample), though more so for the varying intercepts model.

Chart 8. Player wins share vs performance metrics for Mixed Effects model



Points scored (PTS), field goal accuracy (FG_PCT), assists (AST) and turnovers (TO) appear to have the most explanatory power for win shares, though all of the performance metrics are statistically significant. The importance of these particular metrics is consistent with the results from the team-level models. The coefficient signs and magnitudes are similar across all models. In the case of the median regression model, this provides confidence that our results are not driven by the positive outliers in the win shares data.

A player's salary share has a positive, statistically significant coefficient and consistent with the team-level model the coefficient on salary concentration (salary_hhi) is negative and statistically significant. The sign and magnitude of the salary_hhi coefficient is robust to median regression, giving us confidence that the result is not driven by the right tail in the data. Interestingly, the salary pool of the team has a negative coefficient, indicating that spending more on salaries does not necessarily lead to more wins.

In addition, the interactions all suggest that the performance effects are increasing in magnitude with salary share (except for defensive rebounds) and many are statistically significant, in particular for PTS and FG_PCT. This is an interesting result and suggests that the performance of players earning a higher share of the salary pool matters more for explaining the variation in their win shares and confirms the big two/three concept which we will discuss further below.

Table 6: Coefficients for models on player-level data

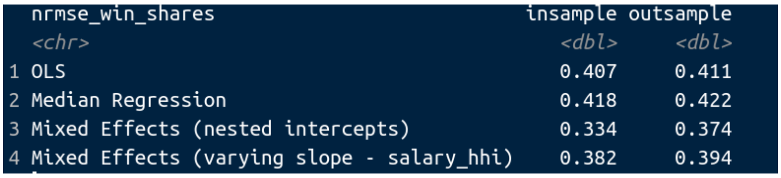
	OLS	Median	Nested	Varying
(Intercept)	4.43 ***	4.21	4.48 ***	4.46 ***
OREB	0.16 *	0.28	0.15 *	0.15 *
DREB	0.78 ***	0.69	0.8 ***	0.81 ***
AST	1.57 ***	1.55	1.59 ***	1.59 ***
STL	0.45 ***	0.44	0.47 ***	0.48 ***
BLK	0.32 ***	0.31	0.33 ***	0.32 ***
TO	-1.5 ***	-1.56	-1.5 ***	-1.54 ***
PTS	1.55 ***	1.35	1.59 ***	1.59 ***
FG_PCT	1.28 ***	1.19	1.24 ***	1.28 ***
FG3_PCT	0.25 ***	0.2	0.23 ***	0.27 ***
FT_PCT	0.38 ***	0.42	0.43 ***	0.39 ***
salary_share	0.17 ***	0.27	0.19 ***	0.18 ***
salary_hhi	-0.13 ***	-0.14	-0.12 *	-0.16 .
salary_total	-0.31 ***	-0.33	-0.24 ***	-0.26 ***
OREB:salary_share	0.26 ***	0.31	0.26 ***	0.25 ***
DREB:salary_share	-0.09	-0.04	-0.11 *	-0.1 .
AST:salary_share	0.1 .	0.16	0.05	0.08
STL:salary_share	0.12 **	0.07	0.11 **	0.13 **
BLK:salary_share	0.03	-0.03	0.01	0.02
TO:salary_share	-0.11 .	-0.16	-0.06	-0.07
PTS:salary_share	0.18 ***	0.23	0.09 *	0.12 *
FG_PCT:salary_share	0.22 ***	0.21	0.2 ***	0.23 ***
FG3_PCT:salary_share	0.1 *	0.1	0.12 **	0.09 *
FT_PCT:salary_share	0.19 ***	0.16	0.2 ***	0.22 ***
R-squared	0.64		0.73	0.68
R-squared marginal			0.63	0.64
R-squared (CV out-of	0.63		0.7	0.66

* p-values not available

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In terms of predictive performance, the nested intercepts model performs best both in- and out-of-sample by normalised RMSE. We also report cross-validated out-of-sample R-squared for the hierarchical models and our value of 0.65 to 0.7 is comparable with other studies predicting game outcomes and win shares.

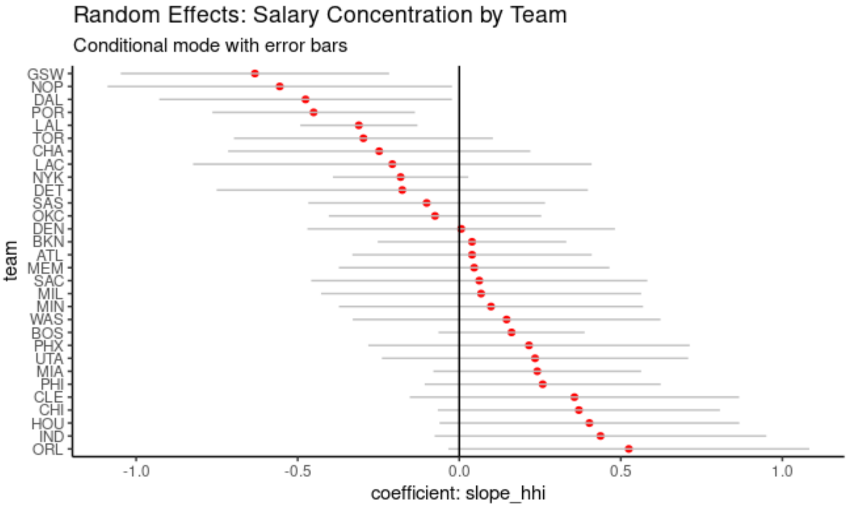
Table 6: Model performance for player-level data



```
  nrmse_win_shares                              insample outsample
  <chr>                                            <dbl>     <dbl>
1 OLS                                              0.407     0.411
2 Median Regression                                0.418     0.422
3 Mixed Effects (nested intercepts)                0.334     0.374
4 Mixed Effects (varying slope - salary_hhi)       0.382     0.394
```

Finally, for the mixed effects model with varying slopes it is interesting to observe how the coefficient on salary concentration varies across each of the teams over the 12 seasons analysed. The chart gives us an idea of how salary concentration is correlated with win shares within each team. For example, the Golden State Warriors (GSW) tended to perform worse in seasons where their salary bill is more concentrated. The GSW were a very successful team over the latter years in our sample (top 4 in season wins for 2014-2017) and this corresponded with a more balanced distribution of their salary pool. In particular, their lineup during these years featured several high-profile players who each earned a significant share of the salary pool - for example, Kevin Durant, Klay Thompson and Draymond Green. However, the total salary pool of the GSW was in the middle of the pack for the years they were most successful and only around two thirds of it was distributed among these top three players (recall in our model the total salary pool is negatively associated with player win shares). This is case study is consistent with the hypothesis that the 'big two or three' players matter most for team success.

Chart 9. Salary concentration coefficient by team



5. Discussion

Across the team- and player-level data we see consistent results across our models. They imply that player's earning a higher share of a team's salary pool tend to contribute more wins to a team (both via season wins for the team and player win shares). In terms of performance metrics, points scored, field goal percentage, assists and turnovers from the top players are the most relevant in our models. We also find some evidence that more turnovers among some of the lower earning players in the team is negatively correlated with team wins. In almost all our models the coefficient on salary concentration is negative and statistically significant, though its magnitude is not particularly large. Nevertheless, this provides some evidence that teams spending too much of their salary pool on one or two players tend to win fewer games. Taken together with the performance metrics, it appears the teams that are most successful may spread their salary budget across their most highly rated players, but teams that rely too much on one superstar player may not fare so well. This supports the idea of the 'big two' or 'big three' in basketball.

6. Appendix

Additional work. Full model tables, model diagnostics and additional charts are in the model-appendix.html file (submitted with this document).

References

Davis, Chris. "NBA Salaries - dataset by datadavis | data.world." Data.World, https://data.world/datadavis/nba-salaries. Accessed 20 December 2021.

Goldstein, Omri. “NBA Players stats since 1950.” Kaggle, 27 April 2018, https://www.kaggle.com/drgilermo/nba-players-stats. Accessed 20 December 2021.

Gonzalez Dos Santos, Teno, et al. “Predicting Season Outcomes for the NBA.” Predicting Season Outcomes for the NBA, https://www.ida.liu.se/research/sportsanalytics/projects/conferences/MLSA21-basketball/MLSA21-paper.pdf. Accessed 30 11 2021.

He, Wei-De, et al. “Predicting the NBA Winning Percentage Based on the Linear Regression Model.” Predicting the NBA Winning Percentage Base on the Linear Regression Model, https://www.jstage.jst.go.jp/article/pjsai/JSAI2020/0/JSAI2020_1K5ES201/_pdf/-char/ja.

Lauga, Nathan. “NBA games data.” Kaggle, 18 November 2021, https://www.kaggle.com/nathanlauga/nba-games?select=games_details.csv. Accessed 20 December 2021.

Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection

inference, with application to the lasso. The Annals of Statistics, 44(3):907–927, 2016.

Madhavan, Vashisht. “Predicting NBA Game Outcomes with Hidden Markov Models.” Predicting NBA Game Outcomes with Hidden Markov Models, https://vashishtmadhavan.github.io/pdf/hmm_nba.pdf. Accessed 30 11 2021.

Nasser, Kimbugwe. “Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle.” Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle, 2016, https://www.researchgate.net/publication/312236952_Predicting_the_Outcome_of_NBA_Playoffs_Based_on_the_Maximum_Entropy_Principle. Accessed 30 11 2021.