Kayson Hansen, Nolan Mejia, John Belardi, Will Newton
MS&E 125: Introduction to Applied Statistics
Final Report

Moneyball: Basketball Edition

Introduction and Motivation

The distribution of salaries for players in the National Basketball Association (NBA) is known to be highly unequal. While select superstars demand salaries of over $40 million per year, the most common salary for the remainder of players falls in the $1-$3 million range (ESPN, *The Hoops Geek*). The NBA's Gini coefficient lends credence to this inequality. The Gini Coefficient, a measure of income inequality ranging from a perfectly equal distribution of 0 to a completely unequal distribution of 1, was calculated at approximately 0.5 for the 2022 NBA season, indicating a highly unequal distribution of wealth ("Gini Index", *The Hoops Geek*). For reference, the United States, often criticized for having high-income inequality, had a Gini coefficient of around 0.47 in 2022 (*Siripurapu*, *The Hoops Geek*). While it is hard to feel bad for anyone making 7-figures or more, within this highly unequal distribution there are certainly NBA players whose production merits a higher salary. By the same token, there are undoubtedly players who make more money than they deserve.

Our group sought to determine if we could use on-court statistical performance to identify the most underpaid and overpaid players in the NBA for the 2021-22 season. We hypothesized that players typically considered "superstars"-- those paid in the upper echelon of the league– will be identified as overpaid based on their statistical performance. We believed the enormous difference between the salary of superstars and the salary of average NBA players would make it nearly impossible for the statistical performance of superstars to exceed that of regular NBA players by a similar difference. Similarly, we expected certain players who recently entered the NBA on entry-level contracts to be considered underpaid due to the pay-ceiling limitations of their contracts.

We analyzed the impact of traditional, advanced, and clutch time statistics on player salary using correlation and regression techniques to generate "fair" salary predictions for players. We then used negative and positive deviations from this "fair" salary to determine underpaid and overpaid players. As we hypothesized, numerous players identified as overpaid were superstars that ranked in the top echelon of salaries such as Russell Westbrook, Damian Lillard, and Paul George, while numerous players identified as underpaid were high-level statistical performers on entry-level contracts at the time such as Trae Young, Ja Morant, and Lamelo Ball. We also found using k-fold cross validation that our models using advanced statistics tended to fit the salary data more closely (lower RMSE) than the models using traditional statistics.

Our analysis is especially relevant for NBA franchises navigating contract negotiations and the free-agent market for players. Paying players appropriately for their statistical output will help teams optimize their cap-space and build a winning roster that will drive revenue.

Relevant work

A 2018 study titled "NBA Players' Pay and Performance: What Counts?" used a multiple regression model and backward stepwise regression to identify the most significant statistics on NBA player salary for the 2017-18 season. The study found points, years in the league, rebounds, assists, and fouls to be the most significant factors on player salary. The authors were surprised 3-point field goals made was not significant considering the League's increasing reliance on the 3-pointer as a primary means of scoring, and the commonly-cited advanced statistic PER, player efficiency rating, was also surprisingly found insignificant (Sigler).

Data and Methods

We obtained raw data from Basketball Reference, the NBA, Spotrac, and ESPN for the statistics and salaries of players in the 2021-22 season. We generated *six* key CSV files for our proceeding analysis: total player stats, players stats per 100 possessions, player stats per 36 minutes, player advanced stats, total player clutch stats, and player salaries. These CSV files were imported into Pandas dataframes and the, munged, as follows:

- Duplicates (players with multiple rows because of a midseason team switch) combined into one entry
- Dollar signs and commas removed from player salaries CSV
- Inner join between player salaries CSV and each of the other five player statistics CSV's.

To avoid outliers, players that appeared in minimal games were filtered out – but this threshold changed across analyses.

After data munging, we performed extensive exploratory data analysis and then different statistical methods of analysis. First, we analyzed a combined CSV for traditional stats (total statistics, stats per 100 possessions, and stats per 36 minutes) and a separate CSV for advanced stats. We analyzed the correlation with salary for each stat type and performed linear regressions on each stat type. After we fit linear regression models for our traditional and advanced stat categories, we fit multiple regression models. First, we fit models using the top 6 most highly correlated stats with salary per stat type. Then, we isolated the stats with statistically significant p-values (<0.05) and fit another multiple regression model only including these significant stats. We then performed 5-fold cross validation to analyze the average predicted RMSE of each model and determine which fit the data most closely.

IV. Results and Discussion

IV.I *Correlation Analysis*

To quantify the relationship between player statistical performance and salary, correlation was calculated. Statistics with a high correlation to salary, furthermore, were used as predictor variables (independent variables) in our upcoming multiple linear regression models. Figure 1 shows the results of the correlation analysis for certain traditional statistics.

Statistic	Correlation with Salary	Statistic	Correlation with Salary	Statistic	Correlation with Salary
PTS	0.565	AST	0.414	FG%	0.055
FGA	0.506	TOV	0.359	3P%	0.029
FGM	0.506	FT%	0.214	STL	0.024
FTM	0.502	3PM	0.199	REB	0.023
MIN	0.500	3PA	0.171	BLK	-0.042
FTA	0.428	DREB	0.132	OREB	-0.150

Figure 1. Selected traditional statistics and their correlation with salary.

These results were somewhat surprising. Traditionally valued indicators of ability, such as field goal percentage (FG%) and three point percentage (3P%), showed little to no positive correlation with salary. It was the stats associated with high usage, such as total points scored (PTS), field goals attempted (FGA), minutes (MIN), and even turnovers (TOV) that were the highest positively correlated with salary. Figure 2 shows the results of a similarly designed correlation analysis for certain advanced statistics.

Statistic	Correlation with Salary	Statistic	Correlation with Salary	Statistic	Correlation with Salary
VORP[1]	0.647	OWS[2]	0.496	PER[3]	0.471
WS[4]	0.540	DWS[5]	0.491	WS/48[6]	0.284

Figure 2. Selected advanced statistics and their correlation with salary.

The higher correlations with salary – compared to traditional statistics – were expected. Advanced statistics are not a single performance measure; they incorporate a wide range of factors, providing a more holistic evaluation of a player's impact. It is no surprise, therefore, that

[1] Value over Replacement Player (VORP): estimated points per 100 possessions that a player adds relative to a replacement-level player

[2] Offensive Win Shares (OWS): estimated number of wins that a player contributes due to their offense

[3] Player Efficiency Rating (PER): estimated per-minute productivity of a player

[4] Win Shares (WS): estimated number of wins that a player contributes due to their overall performance

[5] Defensive Win Shares (DWS): estimated number of wins that a players contributes due to their defense

[6] Win Shares per 48 Minutes (WS/48): estimated number of wins per 48 minutes that a player contributes due to their overall performance

they are more related to salary than traditional statistics. One noteworthy result was the similarity in correlations to salary between OWS and DWS. This, at least initially, indicates that offensive and defensive performance matter about the same when predicting or setting a player's salary. Figure 3 shows the results of a final correlation analysis for the same traditional statistics used in Figure 1 now limited only to clutch time[7] performance.

Statistic	Correlation with Salary	Statistic	Correlation with Salary	Statistic	Correlation with Salary
FGA	0.604	FTA	0.449	3PM	0.240
PTS	0.555	MIN	0.440	BLK	0.186
FGM	0.545	3PA	0.376	FG%	0.125
AST	0.493	FT%	0.345	OREB	0.123
FTM	0.467	REB	0.309	STL	0.115
TOV	0.454	DREB	0.307	3P%	0.092

Figure 3. Selected clutch time traditional statistics and their correlation with salary.

15 of the 18 correlations with salary were higher for clutch time statistics than overall statistics. This indicates an overall stronger relation between clutch time statistical performance and salary than overall performance and salary. This makes sense. Players are valued by and paid to perform in the clutch. High contracts are earned with play when it is all on the line and matters the most.

IV.II *Multiple Linear Regression Analysis*

To generate a model(s) that predicts salary based on certain (traditional, advanced, or clutch time traditional) statistical performance, multiple linear regression was used. The same process was used to generate each model:

- Fit an initial model with the six statistics with the highest correlation to salary (as determined in the preceding correlation analysis) serving as the predictor variables and salary serving as the response variable.
- Investigate the p-values associated with the coefficients of the predictor variables. Remove any predictor variables from the model with p-values greater than 0.05 (statistical significance threshold).
- Generate a final, refined model.

Using this final model, predicted salaries were calculated for all players in our dataset, according to their relevant statistics. Residuals (difference between predicted salary and actual salary) were then calculated. High residuals (actual salary much greater than predicted salary) suggest overpaid players while low residuals (predicted salary much greater than actual salary) suggest underpaid players. These outlier residuals and players are displayed in residual plots.

[7] Clutch time, per the NBA, is defined as the game time where the scoring margin is within 5 points with 5 or fewer minutes left.

Figure 4 shows the generated model for salary derived from the six highest performing traditional statistics in the correlation analysis; points (PTS), field goals made (FGM), and free throws attempted (FTA) were dropped because of high p-values. Figure 5 identifies overpaid and underpaid players, according to this model.

predicted salary = -8,512,000[8] *+ (4,011*[9] ** MIN) + (440,100 * FGA) + (1,292,000 * FTM)*

Figure 4. The linear regression equation for predicted salary based on overall minutes played, field goals attempted, and free throws made.

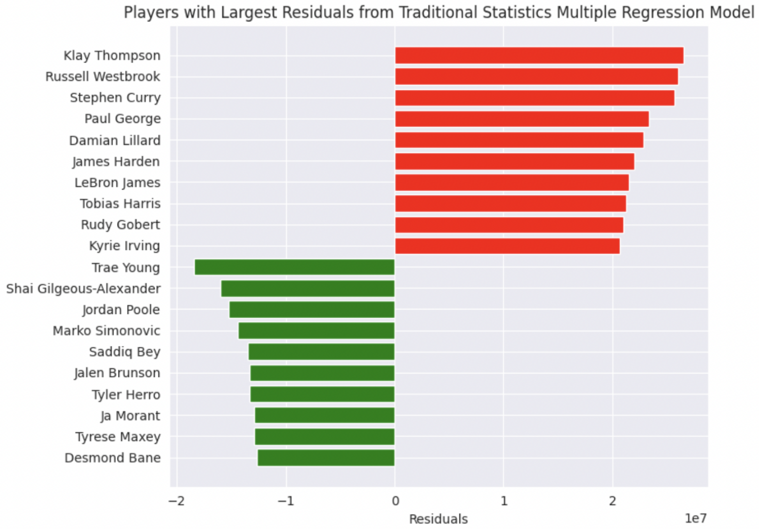

Figure 5. Ten most positive (red) and ten most negative (green) residuals for the model identified in Figure 4.

An expected trend emerged: younger, high performing players were labeled underpaid while older, declining players were labeled overpaid. This is because of the current NBA contract structures, as governed by the league's collective bargaining agreement. When a player is drafted, their draft position corresponds to a pre-set rookie contract of specific length and salary. These are usually 2 (guaranteed) to 4 year (team options) contracts for $2-12 million/year. While still lucrative by everyday standards, this is a complete bargain for young superstar players. Superstars – once free of their rookie deals – are able to sign massive (sometimes 'supermax') contracts that regularly exceed $30 million/year for 5 years (guaranteed). NBA sensation Ja Morant recently signed a 5 year/$193 million contract with the Memphis Grizzlies after his 4 year/$40 million rookie contract expired with the team. In the 2021-22 season (the season of our analysis), he was in the final year of his team-friendly rookie deal, explaining his underpaid status in Figure 5.

Figure 6 shows the generated model for salary derived from the six advanced statistics tracked; offensive win shares (OWS), defensive win shares (DWS), and win shares (WS) were

[8] Example Interpretation of Intercept Coefficient: The predicted salary of a player with 0 minutes played, 0 field goals attempted, and 0 free throws made per 100 possessions is $-8,512,000.

[9] Example Interpretation of Minutes Coefficient: The predicted increase in player salary for each additional minute played per 100 possessions is $4,011.

dropped because of high p-values. Figure 7 identifies overpaid and underpaid players, according to this model.

*predicted salary = -292,600 + (764,200 * PER) + (5,361,000 * VORP) -*
*(66,520,000 * WS/48)*

Figure 6. The linear regression equation for predicted salary based on clutch time field goals attempted, assists, turnovers, and free throws made.

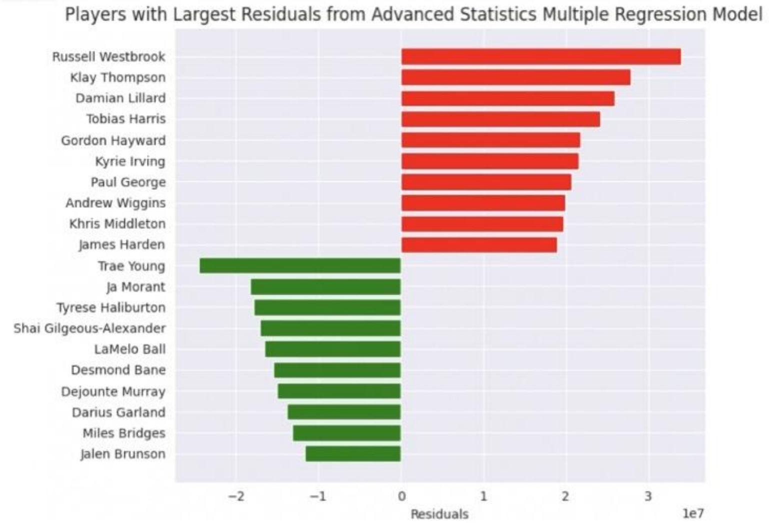

Figure 7. Ten most positive (red) and ten most negative (green) residuals for the model identified in Figure 6.

The same trend held when running the multiple linear regression analysis with advanced statistics. 9 of the 10 most negative residuals corresponded to players signed to rookie contracts; their average age was 23. Conversely, none of the 10 most positive residuals corresponded to players signed to rookie contracts; their average age was 32. Many of the identified overpaid players were, at the time of their contract signing, highly productive players but have since declined. After an MVP season in 2016-17, Russell Westbrook signed the largest contract in NBA history (at the time), earning 6 years/$233 million in guaranteed money. In the contract's lifespan, he would only go on to make 2 All-NBA teams and was out of a starting role by the final year of the deal. Westbrook and the other overpaid players' cases should serve as warnings to teams considering massive contracts for aging superstars.

Figure 8 shows the generated model for salary derived from the six highest performing clutch time traditional statistics in the correlation analysis; points (PTS) and field goals made (FGM) were dropped because of high p-values. Figure 5 identifies overpaid and underpaid players, according to this model.

*predicted salary = 545,100 + (3,182,000 * FTM) + (5,851,000 * FGA) +*
*(7,122,000 * AST) + (8,609,000 * TOV)*

Figure 8. The linear regression equation for predicted salary based on clutch time field goals attempted, assists, turnovers, and free throws made.

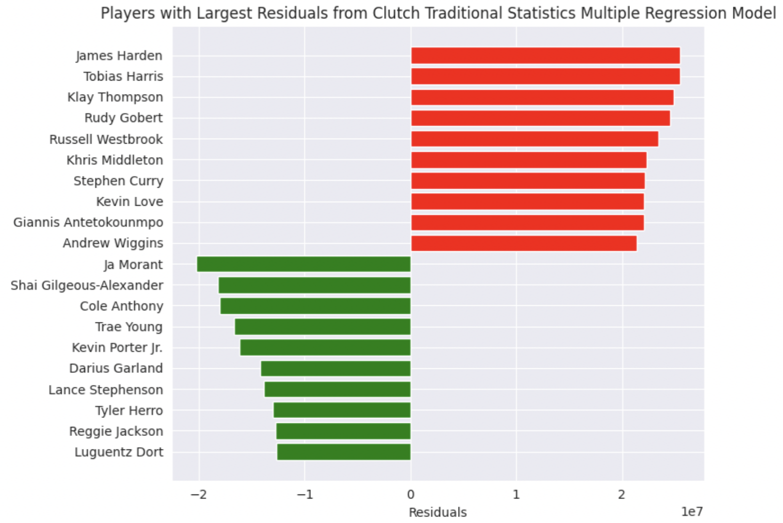

Figure 9. Ten most positive (red) and ten most negative (green) residuals for the model identified in Figure 8.

The multiple linear regression with clutch statistics identifies many of the same names and prototypes as overpaid but suggests one new, differently molded name as underpaid: Reggie Jackson. Jackson, in the 2021-22 season, was a 31-year old playing on a non-rookie 2 year/$22 million deal for the Los Angeles Clippers. An inconspicuous 10 year veteran, Jackson averaged a head-turning 3 points in clutch time (per 5 minutes). While his modest $11 million annual salary ranked 110th in the league, his clutch time points ranked 21st in the league – a nearly 90 spot overperformance! Famous NBA analyst Skip Bayless, similar to us, noticed Jackson's clutch gene, tweeting "I'd trust Reggie Jackson in the clutch more than Lebron James."

All three multiple linear regression models, either incorporating traditional, advanced, or clutch statistics, revealed largely the same trend: high performing, young players on rookie contracts are some of the NBA's most underpaid players and declining, old players on non-rookie contracts are some of the NBA's most overpaid players.

IV.III *5-Fold Cross Validation Analysis*

To assess our multiple regression models' performance on our data and our key decision to remove any non-statistically significant features from these models, 5-fold cross validation (to estimate each model's RMSE) was used. Recall that in our process for multiple linear regression analysis, we first fit a model with all six statistics with the highest correlation to salary and then removed any non-statistically significant statistics (p-value greater than 0.05) from the model to form a final, refined model. In our forthcoming analysis, we calculated the RMSE for both of these models – the non-refined and refined – to see if this was a wise decision. Standard 5-fold cross validation procedures were followed. 5 trials per statistical category were run, and each individual trial's average RMSE was further averaged among the 5 total runs for a more accurate average prediction. Figure 10 displays the results from the 5-fold cross validation analysis.

Statistical Category (used for model)	Non-Statistically Significant Predictors Removed (from model)?	Averaged Predicted Out-of-Sample RMSE (10^6)
Traditional Statistics	No	7.36
	Yes	7.35
Advanced Statistics	No	6.84
	Yes	6.80
Clutch Statistics	No	7.60
	Yes	7.57

Figure 10. Average RMSE across five cross-validation trials for the non-refined and refined multiple linear regression models.

These results indicate that the refined regression models (removed non-statistically significant features) fit the salary data slightly more closely than the non-refined ones (contained non-statistically significant features). We hypothesize that the models with only statistically significant features were able to capture the underlying trends in the data approximately as well as the models with all top-6 correlation features while having lower variance by avoiding some overfitting that the more complex model may have fallen victim to.

The table shows multiple regression using clutch stats was the worst fit to the data. We were surprised to see clutch-time performance as the worst fit to the data. Clutch performance is crucial to the final outcome of a game, and thus we expected compensation to reflect the increased importance of performance during outcome-deciding minutes of close games. An explanation for clutch performance's diminished importance on predicted player salary is that it is a relatively small component of the entire game. Clutch time only occurs during the last 5 minutes of a game within 5 points, meaning that if every game played fits these constraints, clutch time still only comprises roughly 10% of total action. Considering many NBA games do not fit these constraints, the vast majority of player performance occurs outside of clutch time and therefore will be considered with greater weight than only clutch performance.

The table also shows multiple regression using advanced stats was the best fit to the data. This outcome came as no surprise to us. Advanced stats are able to more accurately depict the entire picture of a player's worth than traditional stats by taking into account player efficiency, adjusting for differences in team playstyle, and standardizing player performance to allow for value-added metrics.

Conclusion

Our results indicate that PER, VORP, and WS/48 are the most significant statistics on predicting NBA player salary. This finding affirmed our hypothesis that players considered superstars will be identified as the most overpaid while players on entry-level contracts will be identified as the most underpaid: Stars such as Russell Westbrook, Klay Thompson, and Damian

Lillard were identified as the most overpaid by these stats, while players on rookie contracts such as Trae Young, Ja Morant, and Tyrese Haliburton were identified as the most underpaid.

We note certain limitations to our analysis. Previous research has indicated that the underlying trends in the data may not be suited for linear regression models and has also pointed to non-statistical factors such as player popularity and reputation as being potentially significant factors to salary (Papadaki). Additionally, our analysis only focused on one year of player performance despite contracts being multi-year player investments. For example, the identification of Klay Thompson as an overpaid player may be premature considering we only analyzed stats from his first season back after tearing his ACL. He is likely to improve his performance over the remainder of his contract as he continues to rehabilitate his injury. Lastly, the restrictions of the NBA's collective bargaining agreement on player contracts provided an obstacle by artificially deflating entry-level player compensation. We would like to see the NBA transition to more incentive-based contracts for entry-level players so that statistical overperformers can be appropriately paid.

Despite these limitations, we believe our analysis uncovers some of the underlying statistics that impact player salary and provides a solid foundation for future considerations on the topic.

Large Language Model Attribution: ChatGPT was used to help develop code that was utilized to produce our figures. Analysis was conducted by ourselves.