A Data-Driven Approach to Understanding the Factors that Drive Basketball Player Performance

Nathan Bick
Department of Mathematics and Statistics
Georgetown University
nb845@georgetown.edu

Juna Luzi
Department of Mathematics and Statistics
Georgetown University
jl2845@georgetown.edu

Chris McMinimy
Department of Mathematics and Statistics
Georgetown University
cm1893@georgetown.edu

Abstract

This paper explores three different regression techniques to model the relationship between NBA player statistics and their average points, assists, and total rebounds per game, the most highly cited trio of stats. The measurement of basketball player performance is inherently multivariate and necessitates the use of multiple in-game statistics for a comprehensive evaluation. We adopt several multivariate regression models, with the predicting vector as our response variable. Our approach commences with the selection of predictors informed by domain knowledge and subsequently incorporates methods of feature selection, namely principal component analysis (PCA) and Lasso regularization. Our results indicate that the domain-knowledge model performs adequately; however, it does not match up to the performance of the other two models. Nonetheless, the interpretability of the domain-knowledge model makes it a valuable tool. For instance, our findings reveal that player salary is a reasonable predictive variable for overall performance, as measured by the vector of points, assists, and rebounds.

Code: https://github.com/najuzilu/math-652-project

MATH-652: Multivariate Data Analysis, Spring 2023, Georgetown University, DC. (LateX template borrowed from NIPS 2017.)

1 Introduction

Basketball is one of the most popular sports worldwide, and it generates a significant amount of rich data that can be used for various analyses. In this paper, we predict three key player performance metrics using basketball player and game data, with particular attention paid to the relationship between in-game performance and player compensation. Specifically, we examine data from the National Basketball Association (NBA) over the past 38 seasons, focusing on three performance metrics of individual players.

Our objective is to construct a multivariate predictive model utilizing the provided dataset for the performance metrics of interest, namely, points scored, total rebounds, and assists. The aim is to develop a rigorous and comprehensive approach that captures the underlying complexity of these metrics while ensuring high levels of predictive accuracy. Initially, we develop a multivariate multiple regression model that incorporates features of the dataset we think are most important. Subsequently, two dimensionality reduction techniques, specifically principal component analysis (PCA) and Lasso regression, are examined to identify the optimal approach for capturing the variability of the data.

Ultimately, the obtained results are integrated into our predictive model with the intention of enhancing its overall performance. By using multivariate techniques, we identify patterns and relationships in the data that would be difficult to discern with traditional univariate analyses. This paper provides valuable insights into the factors that contribute to team success in the NBA and has practical implications for coaches, players, and management in the NBA. By identifying key factors that contribute to team success, this paper can aid in decision-making related to player acquisition and development, as well as team strategy. Additionally, the methods employed in this study can be applied to other sports and industries, where data analysis can provide valuable insights for decision making.

2 Description of Problem

In this paper, we attempt to accurately predict a vector of important basketball player performance statistics using several regression models to investigate the relationship of variables such as player compensation with in-game performance. Our primary method for this investigation is multivariate multiple regression. This technique aims to explain the variability present in one set of variables, called the response, as a linear combination of another set of variables, called predictors. The model provides us with coefficients that help explain the relationship between these predictors and the response. In addition, accurate predictions of player performance might be of interest to teams scouting for prospective talent, coaches interested in using quantitative methods to improve player performance, and sports gamblers.

3 Dataset and Features

The data collection process involved extracting player metadata, salary information, and gamelog data from the Basketball-Reference.com website using Python for web scraping. Despite the availability of the data on the website, the process of extracting it was not without its challenges. In particular, the website implemented bot detection measures that prevented rapid and systematic data extraction. To circumvent this obstacle, the scraping pace had to be reduced, which in turn slowed down the overall data collection process.

To address these challenges, we leveraged cloud computing resources like AWS Lambda to invoke multiple concurrent functions, each with its own IP address, to execute web scraping at a reduced scraping pace. Figure 1 illustrates the architecture of our data pipeline. This allowed us to increase the overall speed of data collection by enabling up to 1,000 concurrent web scraping functions to run simultaneously. Through this implementation, we were able to execute a total of 32,062 queries in just 32 minutes, significantly reducing the time required to collect the data. Despite the challenges posed by the website's bot detection measures, our data collection process was able to successfully extract a comprehensive and accurate dataset.

The dataset is composed of three distinct sets of data. The first includes data on individual players, such as their position, height, weight, and the college they attended. The second set is salary data, which reports the yearly payment for each player. Finally, the third set comprises game logs that detail player performance data throughout the regular season and playoffs. The game logs track various statistics, including minutes played, field goals made, three-point shots made, free throws made, attempted shots, and percentage of shots made. The game logs also include data on rebounds, assists, steals, blocks, turnovers, individual fouls committed, points scored, and other game-specific details.

These three sets of data are collected and reported on different intervals. Player data consist of single values, assumed to be constant over the player's career. Salary data is reported annually, while the game logs contain multiple entries per year. We create a usable dataset by calculating annual averages for each player's statistics in the game logs and pairing them with the annual salary data. We assume the height, weight, and position remain constant throughout their career. Our final annual dataset comprises 11,354 observations of 30 features, which will be referred to as our data for the remainder of this paper.

3.1 Relationships between Variables of Interest

We examined the correlations between the players' physical attributes (e.g., height, weight), performance metrics (e.g., points per game, rebounds, field goal percentage), and salary. Figure 2

represents the correlation matrix between various measures with the off-diagonal elements showing the correlation between pairs of variables.

The results indicate a strong positive correlation between height and weight, with taller players generally being heavier. This seems to suggest that larger body size is advantageous in basketball.

Another strong correlation observed is between field goals, field goal attempts, and minutes played. Specifically, players who played more minutes tend to take more shots and make more shots, possibly indicating greater skill and efficiency at scoring. The correlation coefficient of 1 indicates a perfect positive linear relationship between field goals and field goal attempts, which is reasonable given that players who take more shots are more likely to make more shots.

Additionally, free throws and free throw attempts are strongly correlated with field goals and field goal attempts. This may be explained by the fact that free throws are often rewarded when a player is fouled while attempting a shot, and field goals and field goal attempts are two common types of shots in basketball.

Lastly, we find a strong correlation between turnover and minutes played, field goals, and field goal attempts. Players who played more minutes and took more shots were more likely to commit turnovers, potentially due to greater involvement in the offense and more frequent ball handling.

3.2 Normality and Transformations

Prior to conducting any regression analysis, it is essential to examine the distribution of the data to ensure that it meets the normality assumption. Although the least squares method does not require normality to fit a model, any inference drawn from the regression analysis is predicated on the assumption that the data are normally distributed. In the current study, since the dataset is large, the Shapiro-Wilk test is not reliable for assessing normality $(Royston, 1992)$. Therefore, we rely on visual analysis of histograms and Q-Q plots of each variable, which can be found in Figure 3.

Based on our visual inspection, it is evident that only height, minutes played, and weight are near-normally distributed, while the remaining variables display various forms of non-normality. For instance, salary follows a power law distribution, while variables such as three-point attempts, have heavy tails and variables like free throws or personal fouls are skewed. We apply various power transformations to bring the variables closer to normality, depending on the type and severity of skewness. The resulting histograms and Q-Q plots after transformation are presented in Figure 4. After transformation, the variables points scored, total rebounds, and salary now appear near-normal.

4 Methods

Measuring in-game basketball performance is multivariate by definition. There are many different in-game tasks, particularly scoring points, getting rebounds, and getting assists, which are necessary to more fully describe performance. This vector is called a player's "stat line," and although there are additional variables which could be included in the vector (steals, blocks, etc), we felt that there is a tradeoff between the complexity of the model and its usefulness for explaining our relationships of interest - in other words, the set of coefficients would grow large and unwieldy for easy interpretation.

We employ three methods to explore the relationships between our key player performance metrics and the remaining variables in our dataset. The first method we employ is multivariate multiple regression using domain knowledge to inform the set of variables included in the model. Multivariate multiple regression is a comparatively simple model with outputs that readily lend themselves to interpretation. We pay particular attention to the coefficients and their associated p-values which measure whether the coefficient value is significantly different from zero.

After that, we use two variable selection techniques that allow us to confirm or question our previously selected set of variables. The first of these methods is to employ principal component analysis to both reduce the dimensionality of our data as well as look for broader structure within the data that can explain its variability. Finally, we employ a lasso regression as an alternative method to reduce our data down to a smaller set of features.

When assessing the relative quality of different models, we consider the root mean squared error (RMSE), which measures the average squared deviation of fitted values from observations. We also use the Akaike Information Criterion (AIC), which assesses model performance vs model complexity. For both RMSE and AIC, lower values indicated a better model. We also examine the amount of model variation explained using adjusted R^2. Finally, we examine the residuals to check whether or not the model meets the normality assumptions of linear regression.

4.1 Multivariate Multiple Regression

Our first regression model is determined using domain knowledge to select variables that we would intuitively expect to be predictive of in-game performance. The model provides a way for us to test the significance of these variables.

The model specification is:

$$[Points, Rebounds, Assists]^T = \hat{\beta}X + \epsilon$$

where these predictors are transformed using polynomials as follows (determined iteratively via inspection):

- $pts = pts^{(1/2)}$
- $ast2 = ast^{(1/2)}$
- $trb = trb^{(1/2)}$
- $salary = salary^{(1/10)}$
- $wt = wt^2$
- $gs = gs^{(1/3)}$
- $g = g^{(1/2)}$

We use AIC, R-squared, and MSE to measure and compare the performance of the models.

Intuitively, we use salary as a stand-in to measure overall performance. This assumes that basketball compensation is meritocratic, but we will test this via the regression. The model also includes basic physical or demographic and game participation variables as well as position. We do not include other in-game performance variables as predictors, reserving those as potential future response variables.

4.2 Principal Component Analysis

We calculated average player statistics using 20 features including height, weight, salary, and other player performance metrics.[1] Our sample includes data from players across different positions and skill levels to ensure a diverse and representative sample. To ensure that the resulting principal components capture other relevant information in the data that may be useful for the regression model, we excluded the variables that are directly related to the outcome variables from the PCA. Due to the limited availability of non-missing seasonal data from 1980-1983, the PCA is initiated from 1984-2021, encompassing a total of 38 seasons.

Our analysis includes exploration of player metrics across all seasons. To perform the PCA, we standardized the features to ensure that all variables were given equal importance and to prevent variables with large variances from dominating the analysis.[2]

[1]The following is a complete list of input features used in the PCA: height, weight, salary, minutes played, field goals, field goals attempts, field goals percentage, three-point field goals, three-point field goals attempts, three-point field goals percentage, free throws, free throw attempts, free throws percentage, offensive rebounds, defensive rebounds, total rebounds, steals, blocks, turnovers, and personal fouls.

[2]Figure 5 provides a summary of the feature variables and demonstrates the significance of standardization in the analysis.

4.3 Lasso

Lasso regression is a feature-reduction technique that works by introducing a penalty term λ into the regression equation. The lasso model is fit to minimize the following equation

$$\frac{1}{N}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1$$

To fit the lasso regression we choose a value for the penalty parameter that offers a good tradeoff between the number of features retained in the model, and the percent of total variance explained. Figure 6 plots the number of features, and their magnitudes, for different values of the penalty parameter. We chose the value of the penalty parameter based on our judgment about what constitutes an appropriate tradeoff between model simplicity and model explanatory power. In subsequent work we might employ a more systematic approach, such as using cross-validation to select the optimal value for the penalty parameter.

5 Results

5.1 Multivariate Multiple Regression

Judging from the residual plots in the appendix, they are normally distributed and not correlated with the fitted values. We observed the following noteworthy relationships for the key predictor variables. All the following are statistically significant at the 0.01 level, except some of the position variables (see coefficient table in Appendix for more details).

* Salary: Positive relationship for all response variables. This suggests that the higher paid players do in fact perform better in games.

* Height: Negative relationship with points, assists but positive for rebounds. This makes sense given the different in-game roles for guards vs forwards and centers.

* Weight: Significant but very small impact to the response such that we would interpret this to be of little impact.

* Position: as an enum variable each value is represented as a binary. We see that the effects of position make a lot of sense given the typical in-game roles they play. Guard have the most strong relationship with assists, centers and forwards the most with rebounds, while all positions are positive with respect to the points.

We fit a regression using $mean\_salary$, height, weight, position, $games\_started$, $total\_games$, and $minutes\_played$ as the explanatory variables. These variables were chosen by us as promising predictors of player performance. This model accounts for around 79%, 64%, and 74% of the

variation in $mean\_points$, $mean\_assists$, and $mean\_total\_rebounds$ respectively, as measured by adjusted R-squared. This model had RMSE values of 2.76, 1.20, and 1.29 for the three target variables. It has an AIC of approximately 350,721. This model is the best predictor of $mean\_assists$ as measured by adjusted R^2. It is outperformed by the PCA and Lasso regressions in terms of the other two dependent variables. It has the highest AIC of the three models. The model summary can be found in Figure 7, and scatter plots of the residuals can be found in Figure 8. The residuals show significant heteroskedasticity.

5.2 Principle Component Analysis Regression

5.2.1 Average Player Performance Across Seasons

The initial PCA analysis was conducted using basketball player performance data across all seasons. The scree plot presented in Figure 9 was used to determine the number of factors to include in the analysis, and it was determined that three factors were appropriate based on the proportion of total sample variance. The first factor accounted for 42% of the total sample variance, the second factor accounted for 23%, and the third factor accounted for 8%, with a cumulative proportion of 73% for all factors.

Figure 10 displays the results of the PCA analysis on the basketball player performance metrics. The loading values for the first factor were mostly negative, except for the mean three-point field goal attempts, which had a loading close to 0. This factor appears to reflect overall player performance. The second factor appears to distinguish between offensive and defensive player roles. The red loadings in factor 2 relate to a player's offensive performance and playing time, while the blue loadings relate to a player's physical attributes and defensive performance. The third factor seems to capture the differences between player roles and positions. The red loadings in factor 3 are associated with positions such as center and power forward, which require players to be tall, strong, and have good defensive skills. The blue loadings in factor 3 are often associated with positions such as point guard and shooting guard, which require players to be quick, agile, and skilled in ball-handling and creating scoring opportunities.

5.2.2 PCA Regression

We fit a regression using three principal components as explanatory variables. This model accounts for around 94%, 58%, and 90% of the variation in $mean\_points$, $mean\_assists$, and $mean\_total\_rebounds$ respectively, as measured by adjusted R^2. This model had RMSE values of 1.47, 1.29, and 0.80 for the three target variables. This model has an AIC of approximately 325,046, which is lower than the AIC of the domain-expertise regression model, but higher than the AIC of the lasso regression model. This model performs well at predicting all three dependent variables, but is

not the best at any of the three. This is in contrast to the Domain knowledge regression model and the lasso regression model which both appear to focus their predictions on a subset of the dependent variables. The model summary can be found in Figure 12, and scatter plots of the residuals can be found in Figure 13. Heteroskedasticity is still present in the residuals, but it is attenuated compared to the residuals in the domain knowledge regression.

5.3 Lasso Regression

We fit a Lasso regression with 4 features, which explains about 99%, 38%, and 96% of the variation in $mean\_points$, $mean\_assists$, and $mean\_total\_rebounds$ respectively, as measured by adjusted R-squared. The features in this regression are $mean\_fieldgoals$, $mean\_fieldgoals\_attempts$, $mean\_free\_throws$, and $mean\_defensive\_rebounds$. This model has RMSE values of 0.55, 1.57, and 0.51 for $mean\_points$, $mean\_assists$, $mean\_total\_rebounds$ respectively. It has an AIC of approximately 295,726, the lowest of the three models. It is the best model of $mean\_points$ and $mean\_total\_rebounds$, while the PCA model does better with $mean\_assists$. Overall it appears to be the best model as it has the lowest AIC, and is the best predictor of two of our three variables of interest. The model summary can be found in Figure 14, and scatter plots of the residuals can be found in Figure 15. These residuals show the least structure of the three regressions, but heteroskedasticity is still noticeable.

6 Discussion

In this project we sought to accurately characterize the relationships between three NBA player performance statistics: points, assists, and rebounds, and a number of other available statistics. We chose these three statistics as our dependent variables because they are generally considered the most important summaries of player performance. We began by fitting a multivariate multiple linear regression relying on our knowledge of basketball to choose the regressors. We then compared this regression to two regressions fit using dimensionality reducing techniques: Principle Component Analysis, and Lasso regression. We compared the results of our three regressions using several metrics: AIC, RMSE, and adjusted R^2. We found that the lasso regression most accurately modeled the variation points, and rebounds. The domain knowledge regression most accurately modeled the variation in assists. The lasso regression was the most efficient model, as measured by AIC.

6.1 Limitations

By aggregating observations across over 38 seasons of play, we implicitly assumed that the way basketball is played, and thus the predictive relationships in the data are constant over time. This is not likely to be true. For example, three-point shooting has become increasingly popular in the

modern era as evidenced by the success of Stephen Curry and the Golden State Warriors. Our model may over or understate the importance of various predictors for any given era.

To enhance the accuracy of the analysis, it may be preferable to perform PCA on each season individually. This approach would allow for the identification of seasonal variability, resulting in a more redefined depiction of the data. Additionally, this approach would facilitate the comparison of PCA results across seasons, enabling the identification of any patterns or trends in player performance.

In the Lasso regression results, we noticed that many of the selected predictors are related to the response variables by definition and so are highly correlated. For example, the lasso chooses three variables highly correlated with the dependent variable $mean\_points$. These are $mean\_field\_goals$, $mean\_field\_goals\_attempts$, and $mean\_free\_throws$. Our results are therefore less interesting than they may have been had we excluded these regressors. In a future analysis, it would be beneficial to use the results of our exploratory data analysis to loosely filter certain variables before continuing with the more unsupervised method like Lasso, thus reducing the risk of building a model on "spurious correlation".

6.2 Future Work

Relaxing the assumption of constant relationships over time may prove to be the most fruitful area of potential future work, as this could more realistically capture the relationships in the data.

Additionally, we might consider additional variables in our response vector. The are other significant variables such as blocks, steals, turnovers, and personal fouls, to name a few. We could continue the above analyses with this extended response vector. This would more fully capture the in-game performance but would lead to a more complex model.

More broadly, we might explore alternative modeling techniques that do not rely so much on linearity assumptions, including tree methods.

References

"Basketball Statistics and History of Every Team and NBA and WNBA Players." www.basketball-reference.com/

Johnson, R.A. and Wichern, D.W. (2002) Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey.

Royston, P. Approximating the Shapiro-Wilk W-test for non-normality. Stat Comput 2, 117–119 (1992). https://doi.org/10.1007/BF01891203

Appendix

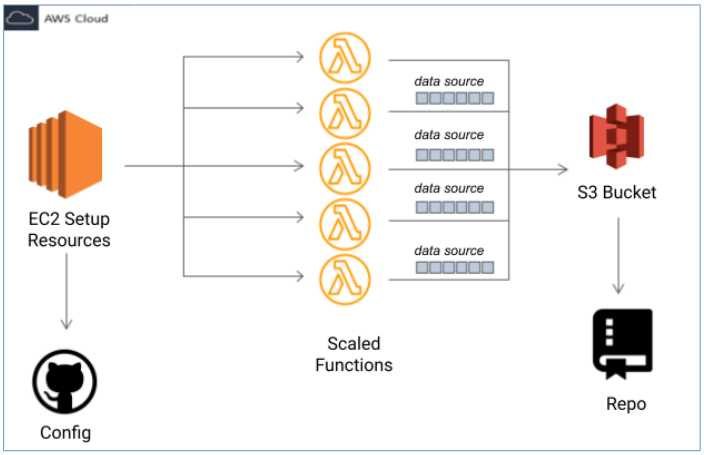


Figure 1: Data Pipeline Architecture

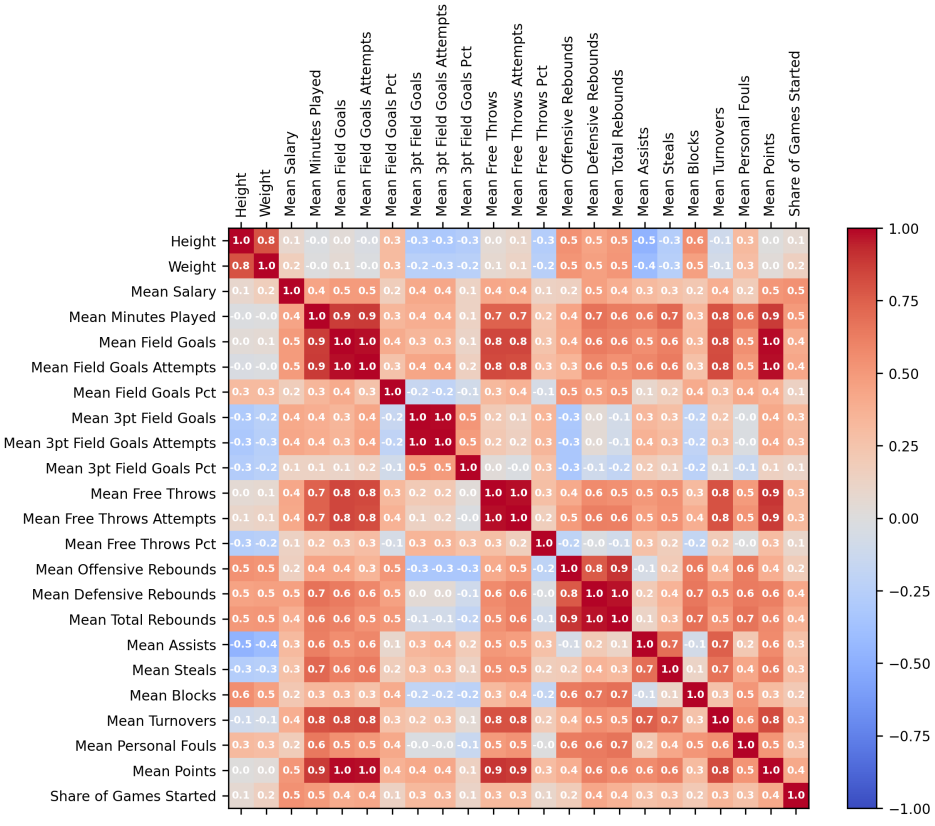
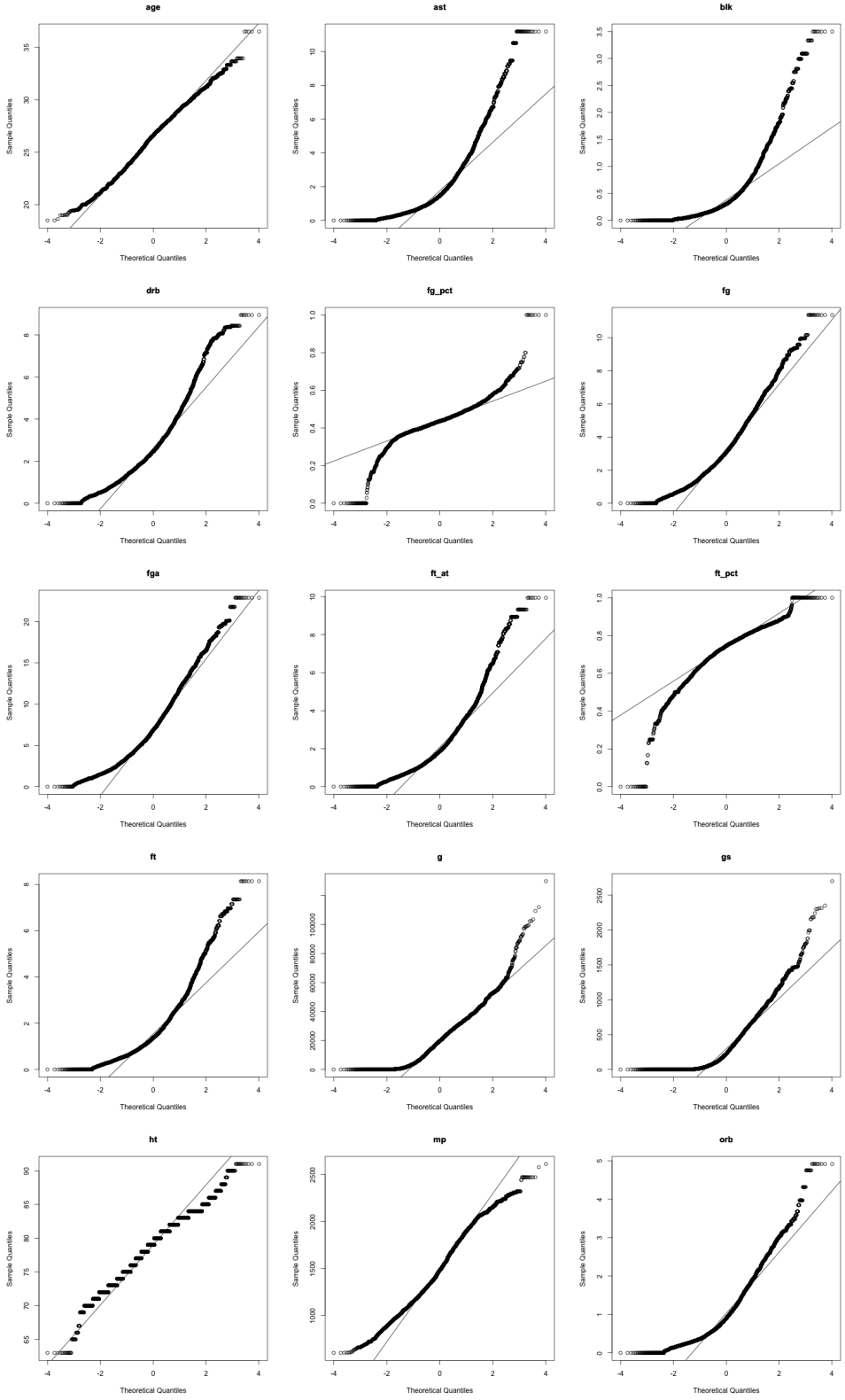


Figure 2: Correlation Matrix of Features


age
ast
blk
drb
fg_pct
fg
fga
ft_at
ft_pct
ft
g
gs
ht
mp
orb
Sample Quantiles
Theoretical Quantiles

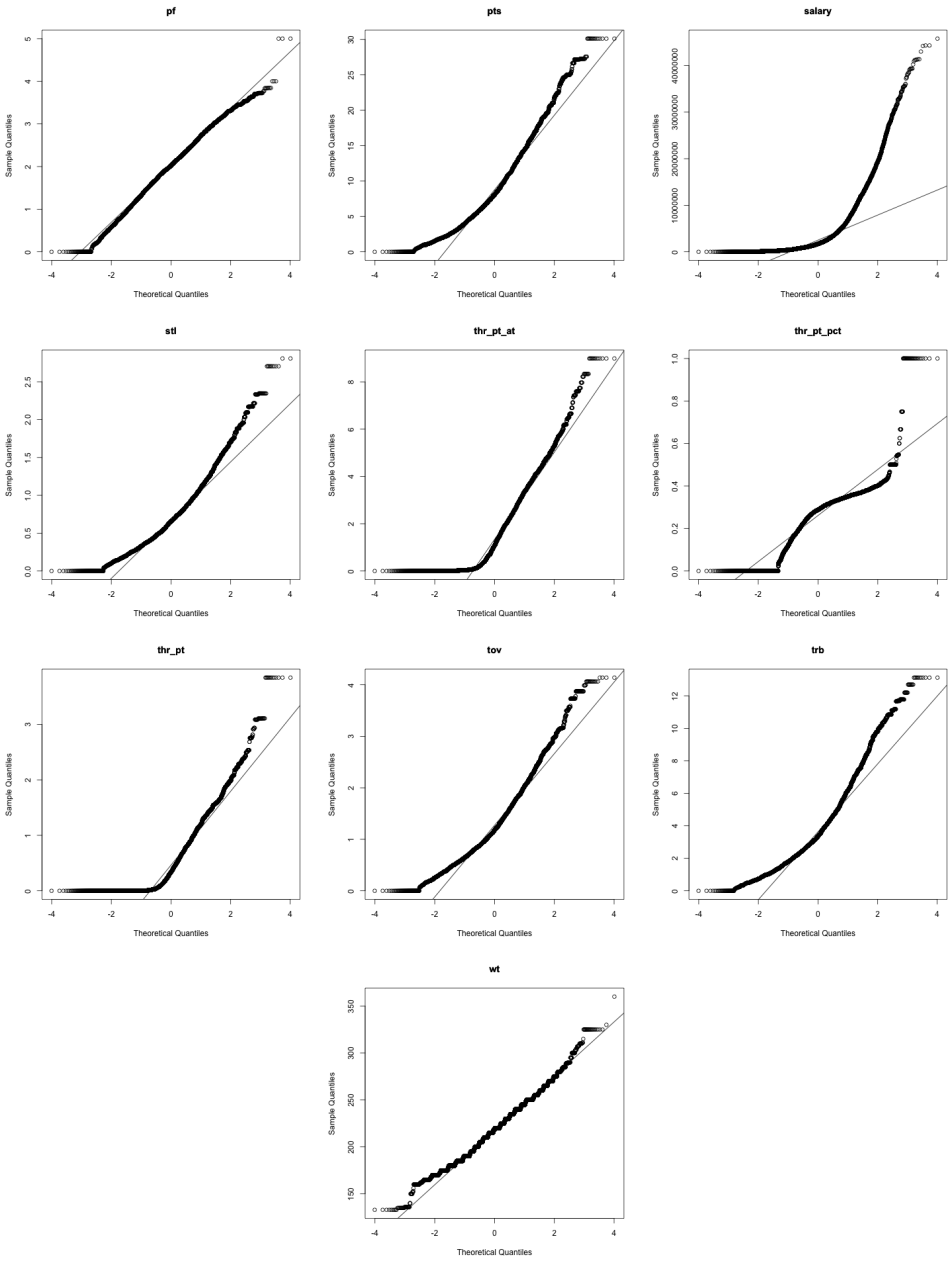


Figure 3: Q-Q Plots

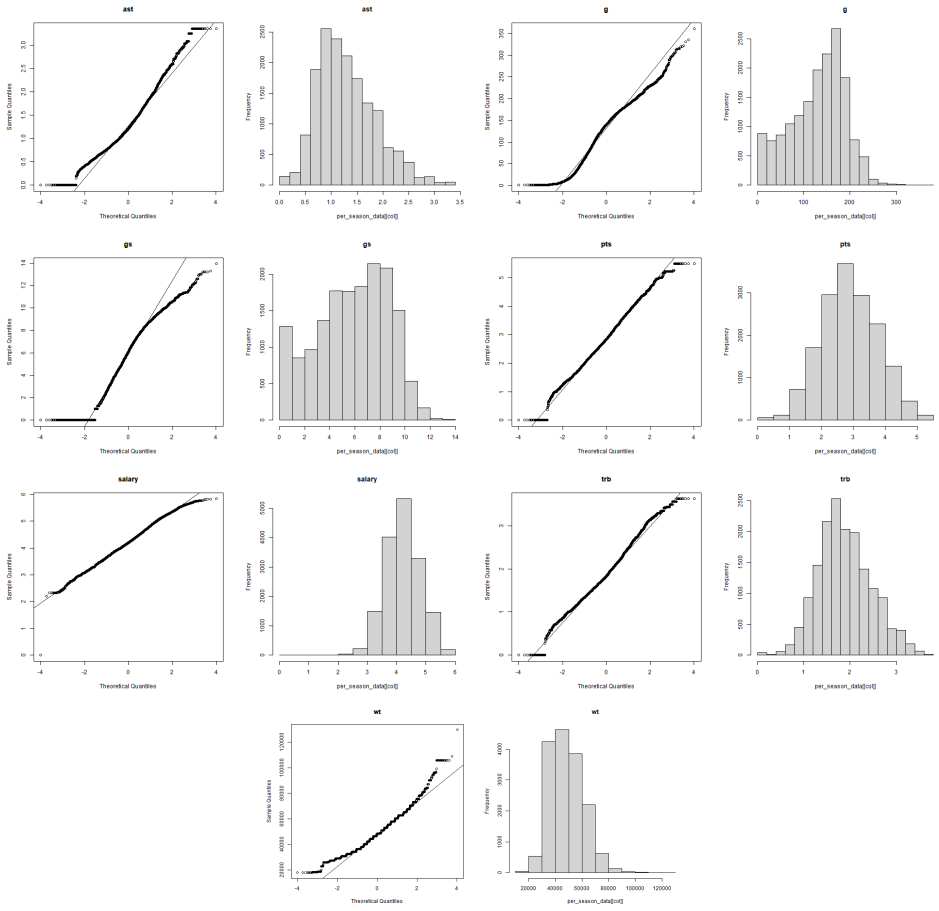


Figure 4: Q-Q Plots of Transformed Variables

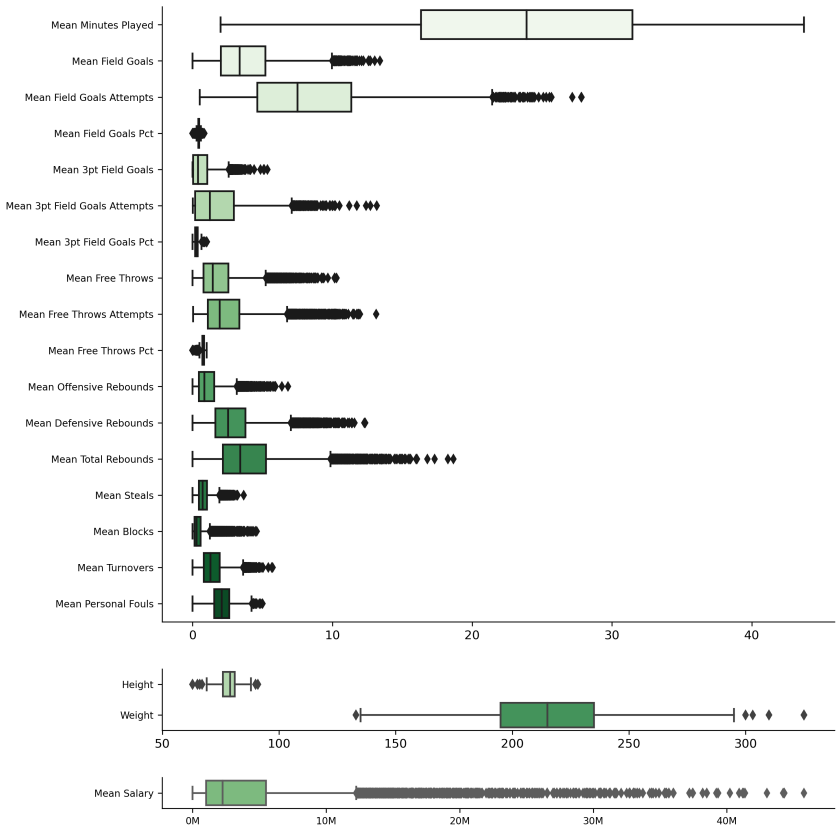


Figure 5: Performance Metrics of Players Before Standardization

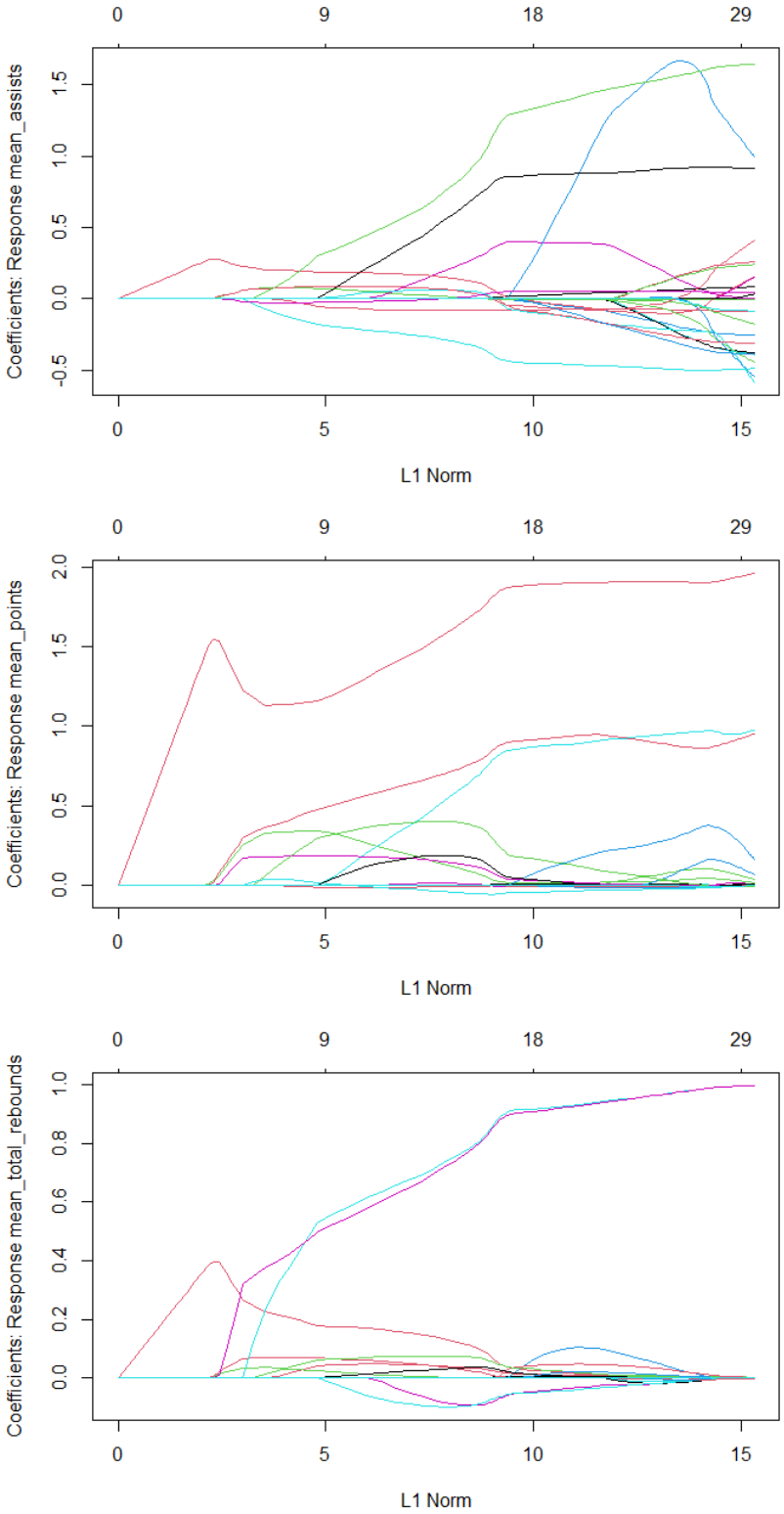


Figure 6: Number of Features and Magnitudes for Varying Penalty Parameter

Domain Knowledge Regression

	Dependent variable:		
	mean_points	mean_assists	mean_total_rebounds
	(1)	(2)	(3)
mean_salary	0.00000***	0.00000***	0.00000***
	(0.000)	(0.000)	(0.000)
height	0.072***	-0.164***	0.042***
	(0.016)	(0.007)	(0.007)
weight	0.012***	-0.008***	0.015***
	(0.002)	(0.001)	(0.001)
positionC-F	0.357**	-0.127*	0.208***
	(0.155)	(0.068)	(0.072)
positionF	0.703***	-0.573***	-1.229***
	(0.124)	(0.054)	(0.058)
positionF-C	0.373***	-0.434***	-0.177***
	(0.129)	(0.057)	(0.060)
positionF-G	1.537***	-0.288***	-2.072***
	(0.165)	(0.072)	(0.077)
positionG	1.565***	0.180**	-2.612***
	(0.180)	(0.079)	(0.084)
positionG-F	1.401***	-0.470***	-2.472***
	(0.160)	(0.070)	(0.075)
games_started	-0.021***	-0.003***	-0.002***
	(0.001)	(0.001)	(0.001)
total_games	-0.003***	-0.0002	-0.0004
	(0.001)	(0.001)	(0.001)
mean_min_played	0.550***	0.125***	0.173***
	(0.003)	(0.001)	(0.002)
Constant	-12.328***	14.069***	-4.989***
	(1.248)	(0.545)	(0.582)
Observations	11,354	11,354	11,354
R^2	0.794	0.635	0.745
Adjusted R^2	0.794	0.635	0.744
Residual Std. Error (df = 11341)	2.755	1.203	1.285
F Statistic (df = 12; 11341)	3,650.109***	1,644.234***	2,756.573***
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 7: MLM Regression: Results

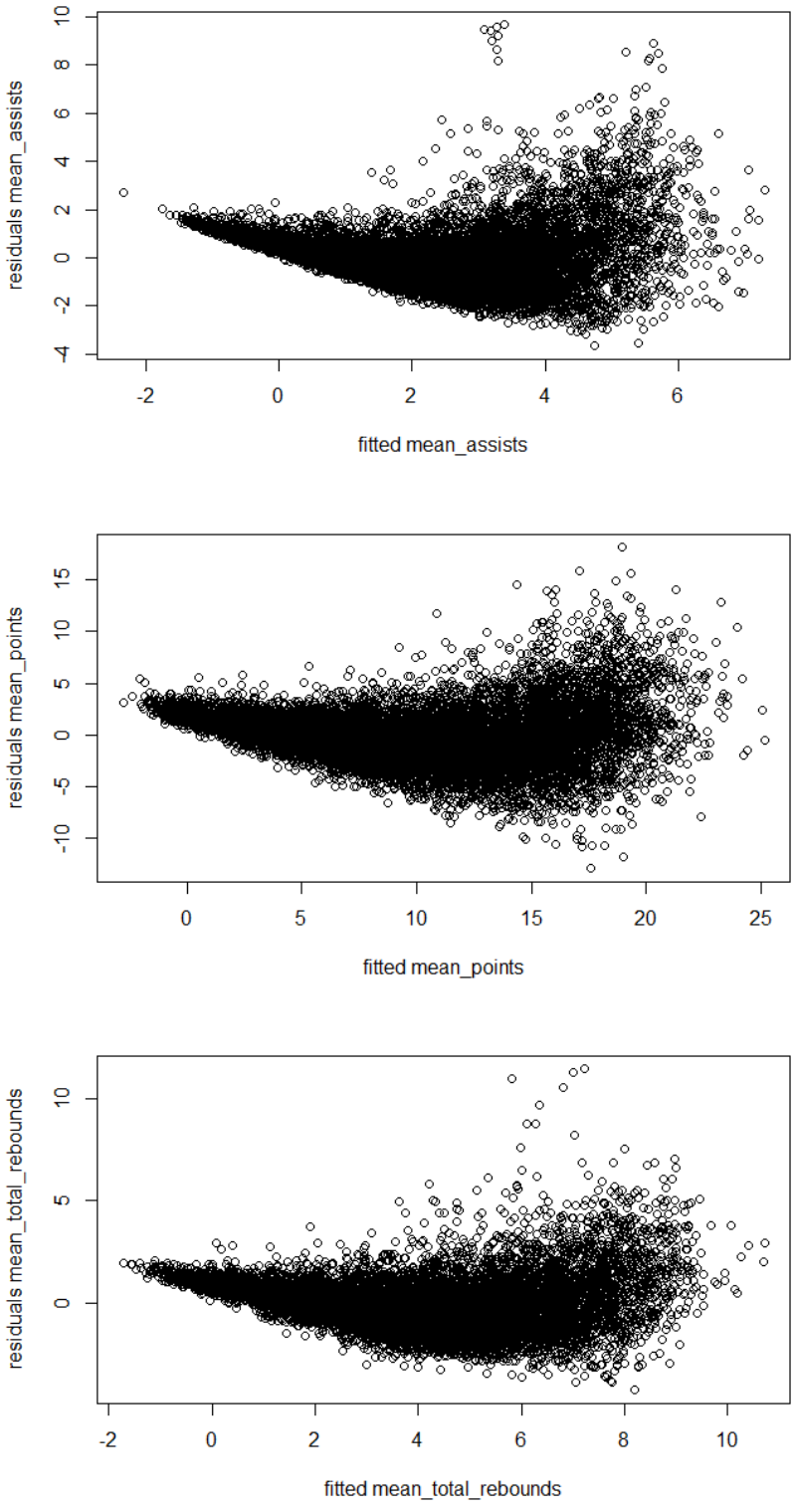


Figure 8: MLM Regression: Residual Plots

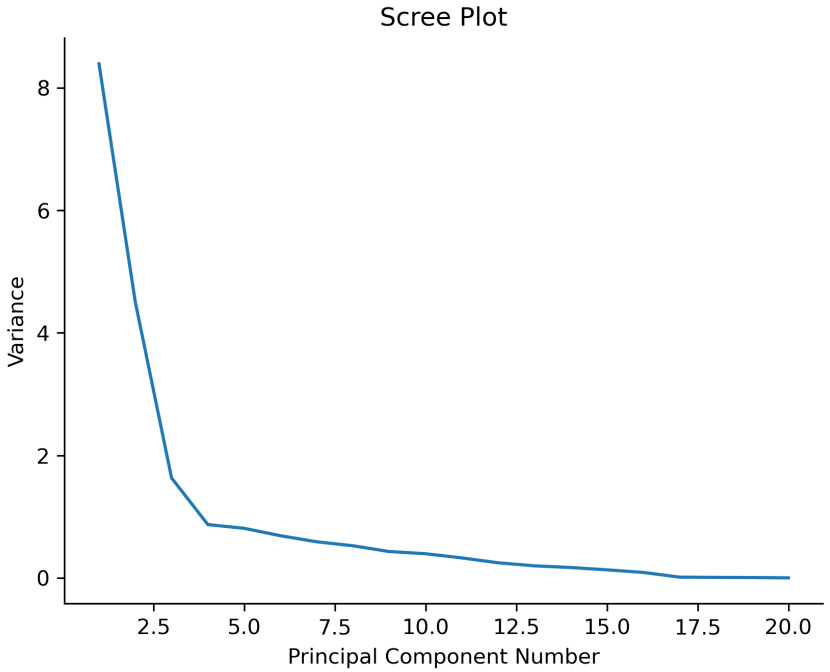


Figure 9: PCA: Scree Plot

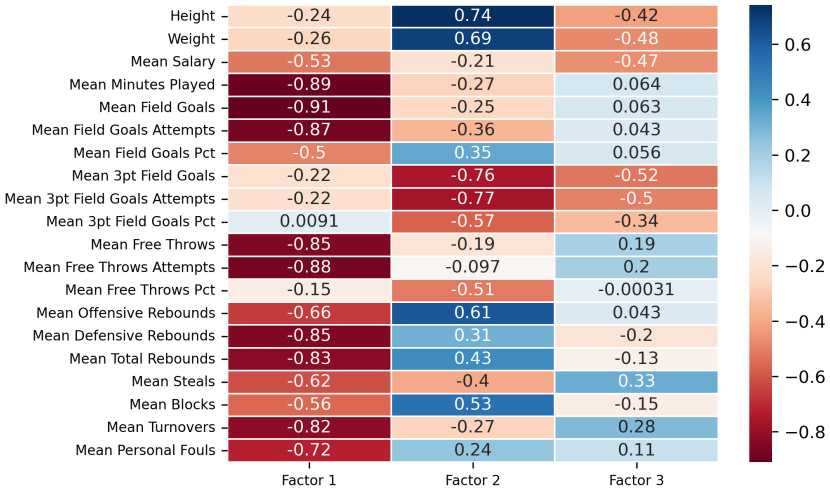
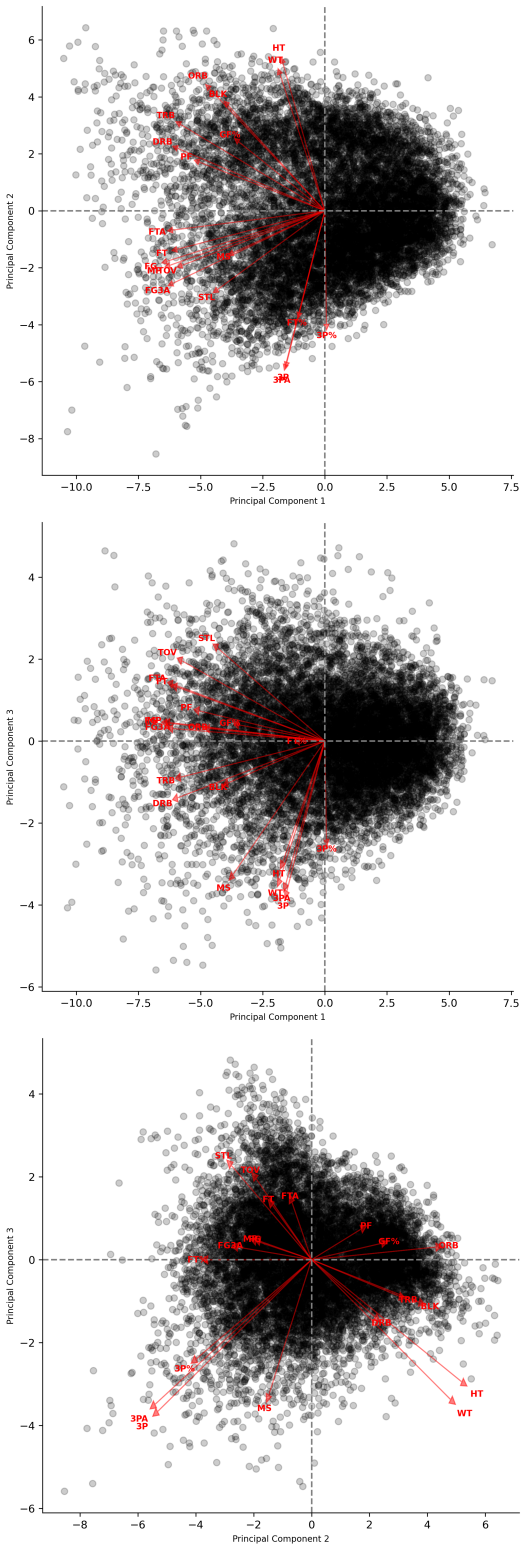


Figure 10: PCA: Results

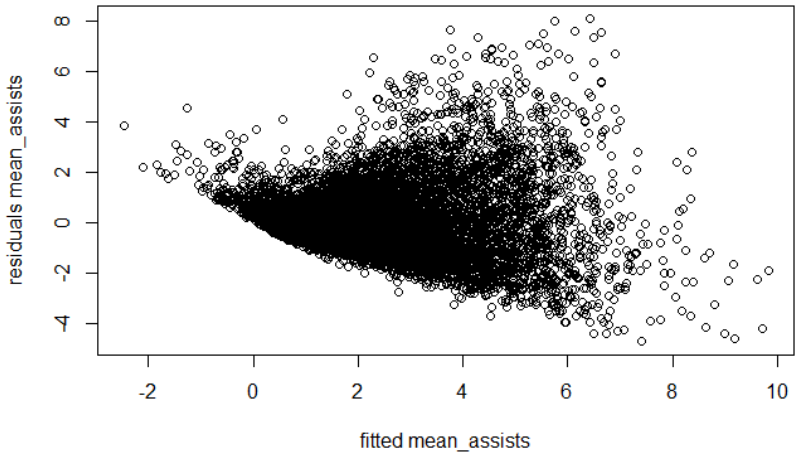

Principal Component 2
Principal Component 1
Principal Component 3
Principal Component 1
Principal Component 3
Principal Component 2

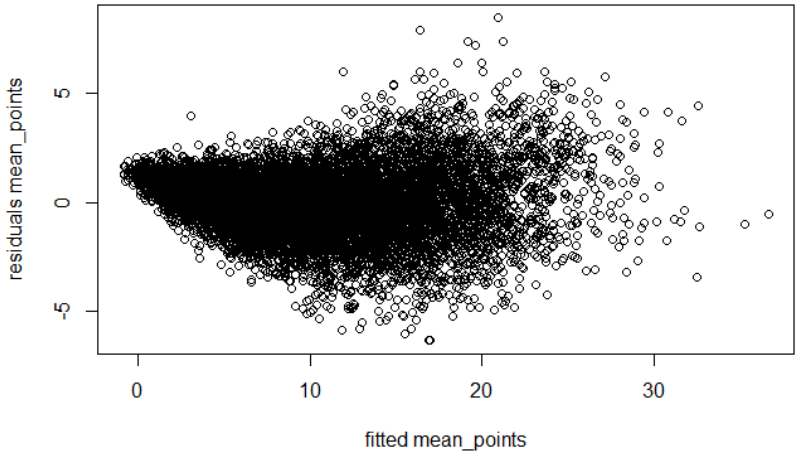
PCA: Biplot

PCA Regression

	Dependent variable:		
	mean_points	mean_assists	mean_total_rebounds
	(1)	(2)	(3)
scores_pc1	2.140***	0.419***	0.700***
	(0.005)	(0.005)	(0.003)
scores_pc2	0.730***	0.464***	-0.775***
	(0.008)	(0.007)	(0.005)
scores_pc3	-0.499***	-0.672***	0.826***
	(0.014)	(0.012)	(0.008)
Constant	10.106***	2.317***	4.024***
	(0.014)	(0.012)	(0.008)
Observations	11,354	11,354	11,354
R^2	0.942	0.578	0.901
Adjusted R^2	0.942	0.578	0.901
Residual Std. Error (df = 11350)	1.466	1.293	0.800
F Statistic (df = 3; 11350)	61,173.110***	5,180.115***	34,421.780***
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 12: PCA Regression: Results





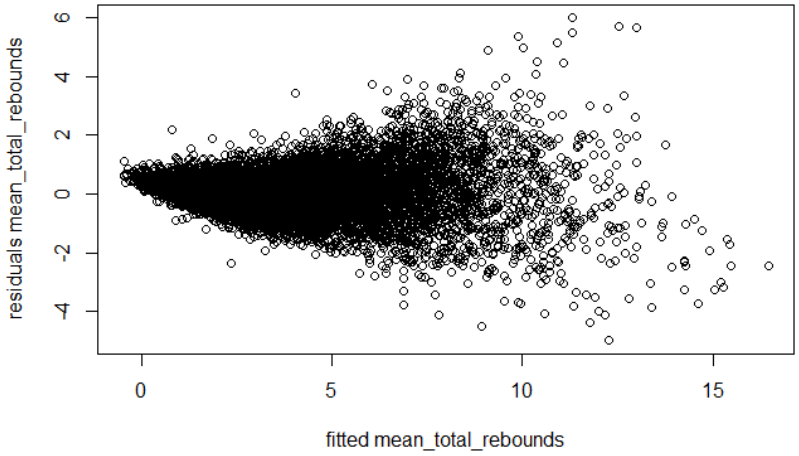


Figure 13: PCA Regression: Residual Plots

Lasso Regression

	Dependent variable:		
	mean_points	mean_assists	mean_total_rebounds
	(1)	(2)	(3)
mean_field_goals	1.474***	-0.334***	0.410***
	(0.013)	(0.038)	(0.012)
mean_field_goals_attempts	0.363***	0.400***	-0.224***
	(0.006)	(0.017)	(0.006)
mean_free_throws	0.852***	0.214***	0.049***
	(0.006)	(0.018)	(0.006)
mean_defensive_rebounds	-0.040***	-0.219***	1.359***
	(0.004)	(0.011)	(0.004)
Constant	-0.003	0.490***	0.270***
	(0.012)	(0.034)	(0.011)
Observations	11,354	11,354	11,354
R^2	0.992	0.381	0.959
Adjusted R^2	0.992	0.381	0.959
Residual Std. Error (df = 11349)	0.548	1.566	0.514
F Statistic (df = 4; 11349)	345,379.500***	1,745.888***	66,595.430***
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 14: Lasso Regression: Results

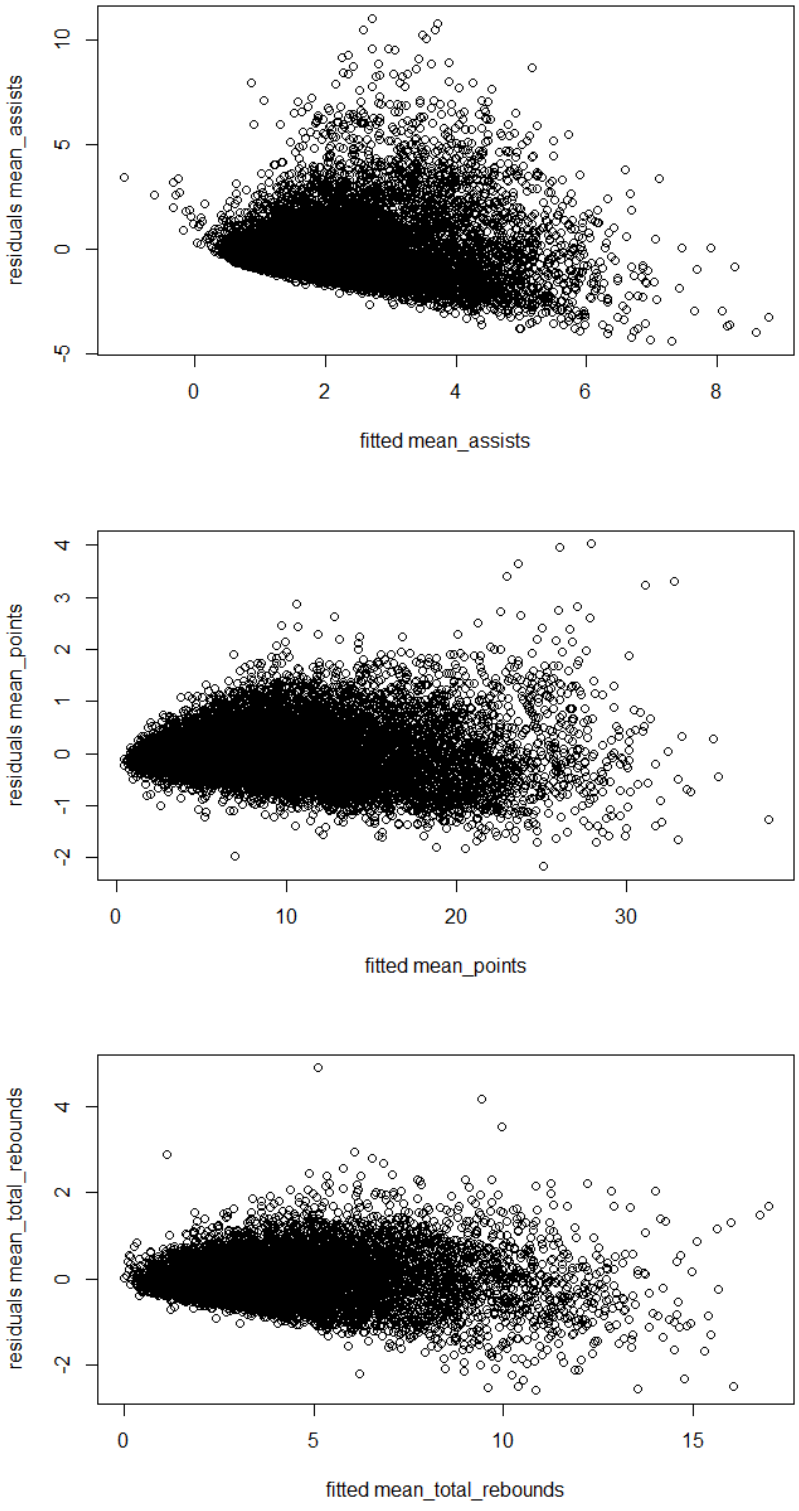


Figure 15: Lasso Regression: Residual Plots