Regression Analysis on Salary of NBA Players

Tianyi Zhou[1], Fengjie Chen[2], Yikun Zhang[2]
[1]Information School, [2]Departments of Statistics,
University of Washington

2019-03-06

Background

Background

- As the premier men's professional basketball league in the world, National Basketball Association (NBA) embraces high reputations in modern competitive sports and captures the eyes of millions of basketball fans worldwide.
- Big fans judge the performance of a superstar simply based on the point that he scored each game or in the whole season, while basketball mavens evaluate the capability and potential of a player by more statistics.
- Goal: Predict the salary of a NBA player in 2017 based on his season-long performance statistics.
- Evaluation metrics: We choose `Mean Sqaured Error(MSE)` as a metric to evaluate performace of a model in 5-fold cross validation.

Data Description

- Datasets available on Kaggle, skip procedure of collection.
- Combine team statistics with individual statistics, remove adundant predictors and impute missing values by our understandings of those statistics

Data Description

Table 1: Head of selected columns

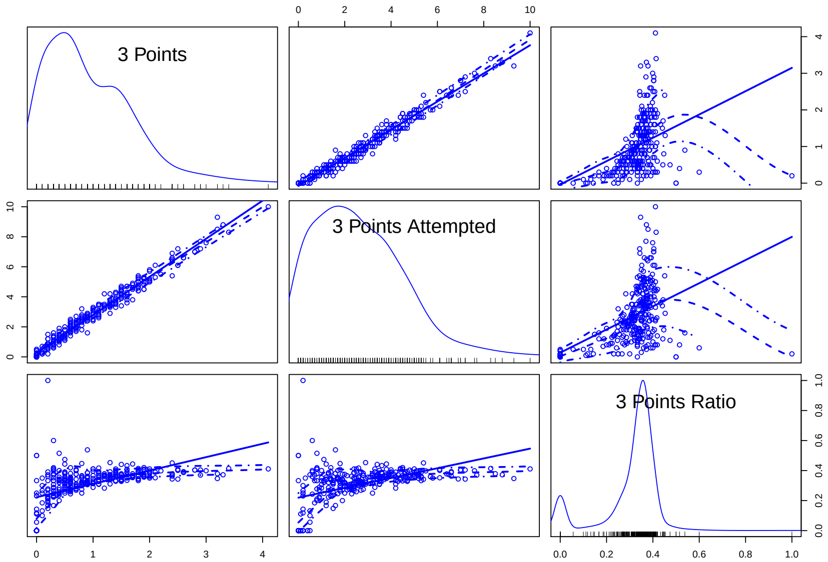
Player	Age	Turnover	Foul	Points	Salary
Russell Westbrook	28	1.9	2.3	31.6	26.5
James Harden	27	1.9	2.7	29.1	26.5
Isaiah Thomas	27	1.3	2.2	28.9	6.6
Anthony Davis	23	1.2	2.2	28.0	22.1
DeMarcus Cousins	26	1.5	3.9	27.0	17.0
Damian Lillard	26	1.3	2.0	27.0	24.3
LeBron James	32	1.6	1.8	26.4	31.0
Kawhi Leonard	25	1.1	1.6	25.5	17.6

Data Preprocessing

Data Preprocessing

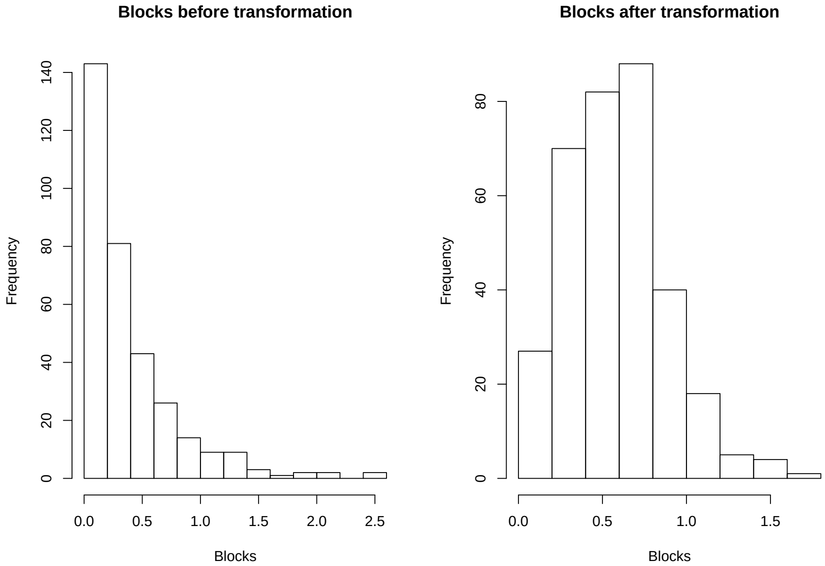
- Too many predictors (team,individual,social media etc.), relatively few samples, implicit relationships between predictors.
- Multicollinearity (Remove some predictors which can be determined by others, regression only on predictors with high correlation with responses,. . .)

Data Preprocessing



Data Preprocessing

After preprocessing, data is of shape 335×48



Exploratory Data Analysis

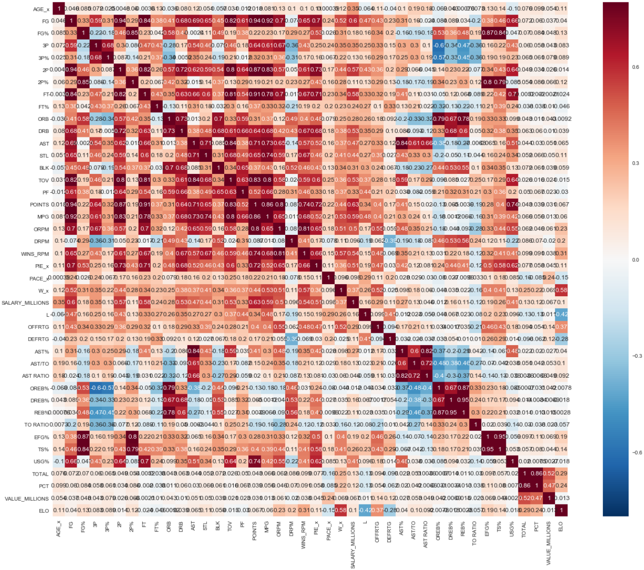
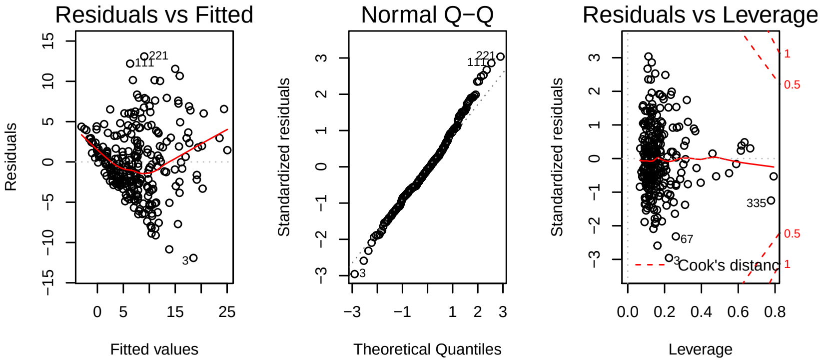


Figure 1: Correlation heatmap

Linear Regression

Linear Regression

- SALARY$\sim$ 42 continuous + 1 categorical with 6 levels (position)
- Leave one fold out, and train on other four folds

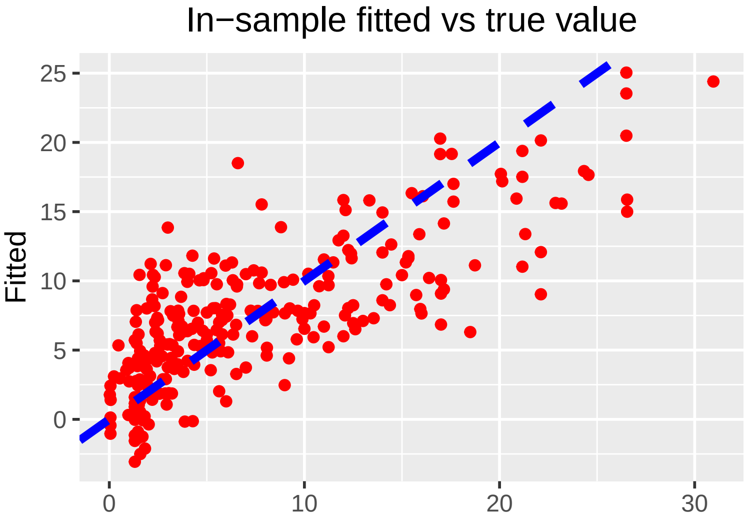


Linear Regression

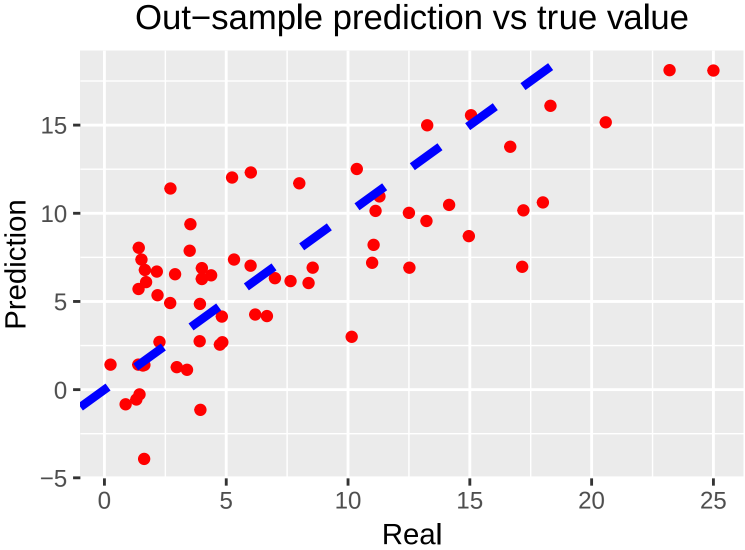
- The normal distribution seems plausible, and few outliers are presented.
- However, strong evidence is presented against constant-variance assumption, inference may not be solid.

Linear Regression

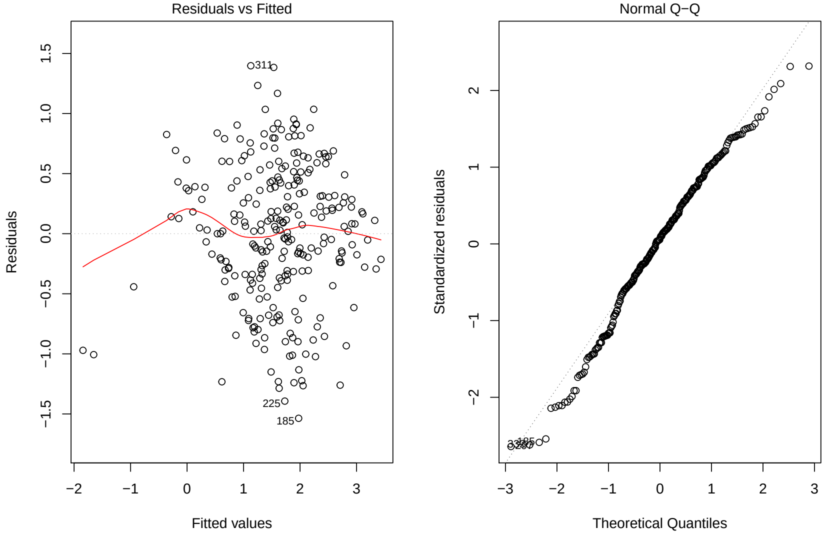
- conservative on the right-hand side, negative fitted value on the left-hand side. Corrected prediction or other models may be needed.



Linear Regression



Log Transformation of Response



Log Transformation of Response

- Age, 2 point%, Foul, Win, Offensive rating, Team value are significant at 5% level.
- Compared with null model, error sum of square shrinkages by 47%.

Linear Regression Results

- Linear model including all predictors: $\texttt{MSE} = 23.4$
- Correct the predicted value using $max(0, x)$, $\texttt{MSE} = 23.2$
- After log-scaling reponse, $\texttt{MSE} = 23.8$

Regularized Linear Models

Regularized Linear Models

Lasso Regression:

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$$

Ridge Regression:

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_2^2$$

- Lasso Regression: $\texttt{MSE} = 27.26$
- Ridge Regression: $\texttt{MSE} = 32.24$

Variants of Lasso (I)

Lasso (L^1 Penalty):

$$\begin{aligned} &\min_{\beta} ||y - X\beta||_2^2 + \lambda \sum_i |\beta_i| \\ &\equiv \min_{\beta} ||y - X\beta||_2^2 + f_\lambda(\beta) \end{aligned}$$

Variants of Lasso (I)

Lasso (L^1 Penalty):

$$
\begin{aligned}
&\min_{\beta} ||y - X\beta||_2^2 + \lambda \sum_i |\beta_i| \\
&\equiv \min_{\beta} ||y - X\beta||_2^2 + f_\lambda(\beta)
\end{aligned}
$$

SCAD Penalty:

$$\min_{\beta} ||y - X\beta||_2^2 + n \sum_i p_\lambda(\beta_i; a),$$

where for some $a > 2$ and $\lambda > 0$

$$p_\lambda(\beta_i; a) = \begin{cases} \lambda|\beta_i|, & \text{if } |\beta_i| \le \lambda, \\ -(\beta_i^2 - 2a\lambda|\beta_i| + \lambda^2) & \text{if } \lambda < |\beta_i| \le a\lambda, \\ (a+1)\lambda^2/2 & \text{if} |\beta_i| > a\lambda. \end{cases}$$

Variants of Lasso (II)

Lasso (Penalized or Lagrangian Form):

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$$

Variants of Lasso (II)

Lasso (Penalized or Lagrangian Form):

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$$

$\Longrightarrow$

$$\min_{\beta} ||y - X\beta||_2^2 \quad \text{subject to } ||\beta||_1 \leq s$$

Variants of Lasso (II)

Lasso (Penalized or Lagrangian Form):

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$$

$\Longrightarrow$

$$\min_{\beta} ||y - X\beta||_2^2 \quad \text{subject to } ||\beta||_1 \leq s$$

Dantzig Selector:

$$\min_{\beta} ||\beta||_1 \quad \text{subject to } ||X^T(y - X\beta)||_\infty \leq s,$$

where s is a tuning parameter.

Comparison of Models

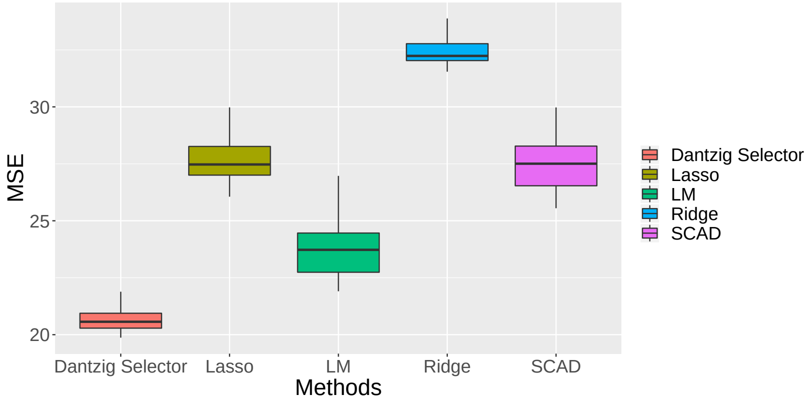


Figure 2: MSE of Different Models

Conclusions and Future Works

Conclusions and Future Works

Conclusions:

- Linear models seem to be reasonable and have relatively good predictability on the current dataset. (Compared with the null model, MSE decreases by 54%)
- Dantzig selector can further reduce MSE

Conclusions and Future Works

Conclusions:

- Linear models seem to be reasonable and have relatively good predictability on the current dataset. (Compared with the null model, MSE decreases by 54%)
- Dantzig selector can further reduce MSE

Future Works:

- Garner more predictors (e.g. social media stat), and collect more samples (from different years)
- Implement Grouped Lasso when more categorical variables are present

Thank you!