Regression Analysis on Salaries of NBA Players

Yikun Zhang[1], Fengjie Chen[1], and Tianyi Zhou[2]

[1]Department of Statistics, University of Washington [2]Information School, University of Washington

Problem

National Basketball Association (NBA) embraces high reputations in modern competitive sports and captures the eyes of millions of basketball fans worldwide. Fans judge the performance of a superstar simply based on the point that he scored in a season, while basketball mavens evaluate a player by more professional statistics.

Our goals are two folds:

- Predict annual salaries of players based on team and individual statistics.
- Seek for most informative predictors.

Evaluation metrics: *Mean Sqaured Error (MSE)* through 5-fold cross validation.

Simple Linear Regression

Model 1: Salary $\sim$ All predictors

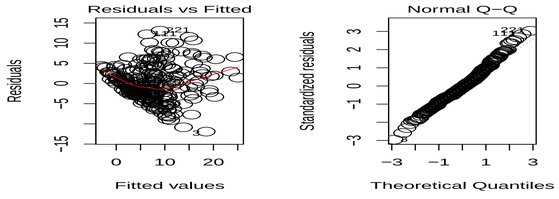

Fig 3: Residual Plots of Model 1

- The normality assumption seems plausible, and few outliers are presented.
- Variances of residuals tend to increase, so the constant-variance assumption is doubtful.

Model 2: log**(Salary)** $\sim$ **All predictors**

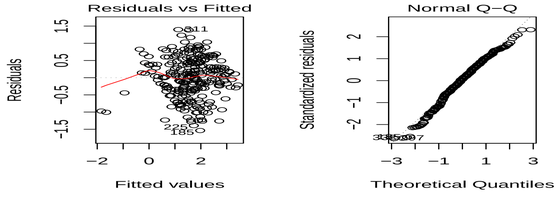

Fig 4: Residual Plots of Model 2

- Within a wide range of fitted values, the bandwidth of residuals remains unchanged, and the normality assumption seems plausible.
- Age, 2 points%, Foul, Win, Offensive rating, Team value are significant at 5% level.
- Compared with null model, the residual sum of square is reduced by **47%**.

Data Preprocessing

Our data is available at *Kaggle*. We extracted team and individual statistics from many other statistics (*e.g.* social stats) for future use.

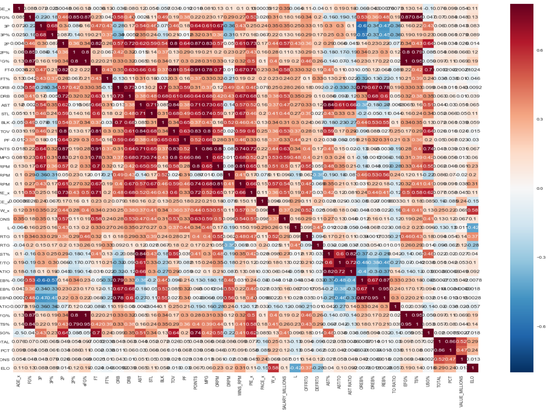

Fig 1: Heatmap of Variables

Our data preprocessing includes:

- Remove duplicated columns with different alias but same value.
- Impute missing values based on relationships between variables.
- Apply log or square root transformation to skewed features to achieve normality.
- Delete highly correlated variables.

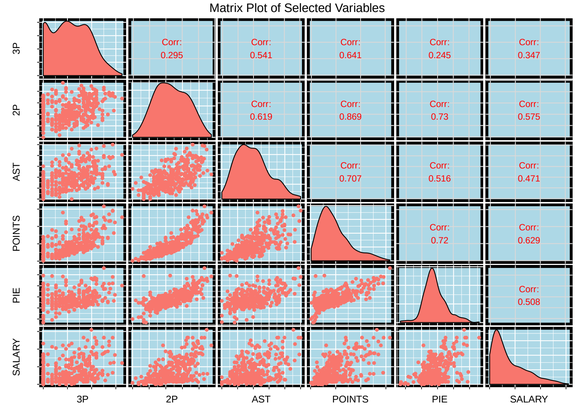

Fig 2: Matrixplot of some features

After preprocessing, 335 observations and 43 variables are retained for further use.

Regularized Linear Models

Ridge Regression:

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

Lasso Regression:

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

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

Model	LM	LM(corrected)	LogLM	Lasso	Ridge
MSE	23.4	23.2	23.8	27.2	32.2

So far, the best model is linear regression with outputs corrected by $\max(0, y)$. To obtain better predictability, two variants of *Lasso* are implemented.

(1) SCAD Penalty:

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

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

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

(2) Dantzig Selector:

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

where s is a tuning parameter.

Results

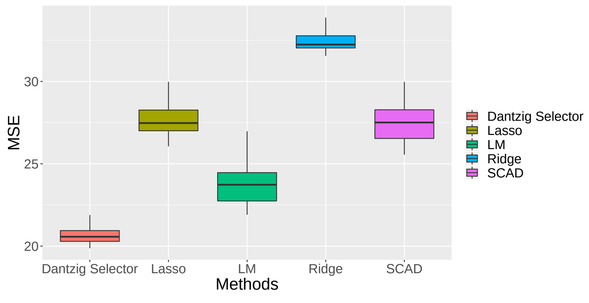

Fig 5: Cross-Validation MSEs (Repeated 50 Times)

Selected Features by Regularized Models		
Lasso	SCAD	Dantzig Selector
Age	Age	Age
Defensive RB	Field Goals	Personal Fouls
Points		Points
		Offensive RPM
		Wins RPM
		Player Impact Factor
		Pace Factor
		Usage Percentage

* RPM: Offensive Real Plus-Minus. Red features: Positive coefficients; Cyan features: Negative Coefficients.

Conclusions and Future Work

Conclusions:

- Linear models seem to be reasonable and embrace a relatively good predictability on the current dataset. (Compared with the null model, Dantzig Selector decreases MSE by 54%)
- The features, *Age* and *Points* are selected more than twice among three regularized models, which are positively correlated the *salary*.

Future Work:

- Incorporate more features from different sources (e.g. social media stats) in order to generalize and improve the predictability of our model.
- Apply nonparametric algorithms (e.g. tree-based methods) to capture interaction effects.

References

[1] Emmanuel Candes and Terence Tao (2007). *The Dantzig selector: Statistical estimation when p is much larger than n.* The Annals of Statistics 35, No. 6, 2313-2351.

[2] Jianqing Fan and Runze Li (2001). *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties* Journal of the American Statistical Association, 96:456, 1348-1360