Predicting voting behavior for the NBA Most Valuable Player Award

Christopher Vela and Daniel Schwartz

Advanced Data Analysis, Spring 2014

Final Project

Abstract

In this project we consider which metrics best predict total MVP votes for basketball players in the NBA. We used multiple regression techniques on a data set that included 20 years of MVP candidates. Ridge and step-wise regression techniques were used to find a model minimizing in-sample predictive error. Step-wise regression yielded a lower error of 47 percent in predicting the winner, while ridge regression yielded an error rate of 52 percent. We identified player salary as well as several performance metrics to be most associated with MVP success.

I. Introduction

At the end of each season, the National Basketball Association designates the most valuable player (MVP) based on a voting system in which select basketball journalists assign ranks to who they believe are the five best players for the season. Points are then assigned inversely with rank for each ballot cast. The rank-point equivalencies are as follows:

rank for a given ballot	**points**
1st	10
2nd	7
3rd	5
4th	3
5th	1

The award is given to the player receiving the most points.

In recent years, the MVP has been the subject of much criticism from fans and sports commentators. Wrote *Fansided* writer Phil Daniels: "The NBA fails to concretely define the criteria for the Maurice Podoloff Trophy winner. The only criterion is that the MVP is awarded annually to the 'most valuable player.' Value is not defined. Scope is not limited" (Daniels 1). Many have also claimed that the award is given based on popularity and media attention rather than the performance of a player.

To investigate voting behavior and identify important predictors of success in the MVP race, we examine historical data from 1986-2012, looking at metrics in several areas: player age, position, individual performance, contribution in the context of a team and salary. We assume here that player salary functions as a proxy variable for understanding popularity and overall market demand for a particular player. We narrow our focus to the subset of players receiving positive point counts (typically a handful of players each season). Our question of interest can then be distilled to: "Among the top-performing players each season, which characteristics are best associated with MVP success?" Data pertaining to basketball metrics, player age, position and MVP voting was downloaded from `www.basketball-reference.com`. Player salary data was taken from `www.eskimo.com/~pbender`.

Two seasons of data (1998-1999 and 2011-2012) were excluded from our analysis as player negotiations prevented completion of these seasons' games. The 1987-1988 and 1990-1991 seasons were also excluded as player salary data could not be found. Total point counts per player were normalized by the total number of points available since the total number of ballots cast differs from year to year, giving a response variable that can be interpreted as the proportion of the vote captured by a particular player. Player salaries were normalized by the cumulative league salary for the particular year to account for inflation.

After aggregating all of the positive point-receiving players over 23 seasons we are left with a sample size of n = 374. This method takes into account two important assumptions. The first is that the tastes and preferences of basketball media sources participating in the vote are stable over time. With the advent of sabermetrics and newer ways to analyze sports data (see the Michael Lewis book, "Moneyball"), it's certainly possible that new trends could be emerging that are influencing how players are perceived. We ignore this possibility in weighing the samples equally over time for model-building. The second assumption is that player performance has no time-dependence. We assume the performance of a player like Lebron James, for instance, who has been successful in the MVP race over consecutive seasons, is not influenced by previous seasons' success.

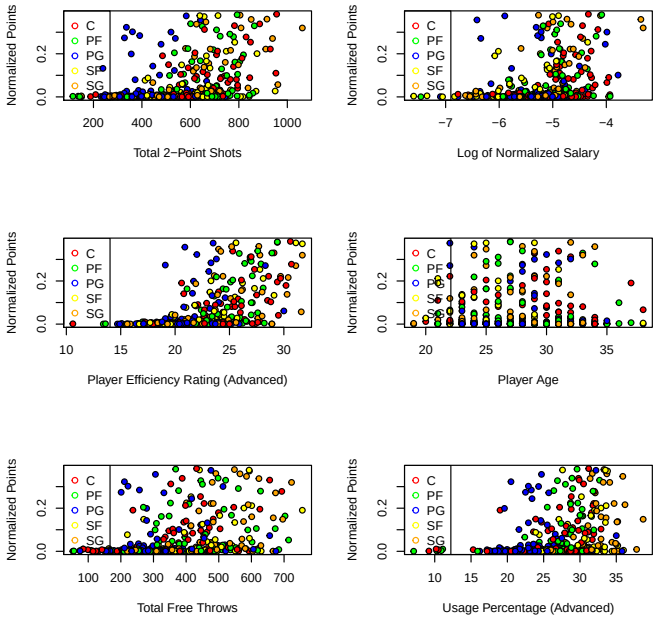
Our paper is organized into four sections. The first comprises an exploratory analysis of this rich data set and clues from the data structure that led us to deploy regression-based tools. The second part focuses on our results from ridge regression as well as model-checking, statistical analysis and model validation using in-sample prediction error. The third part focuses on results from implementing stepwise regression, along with a discussion of model-checking, analysis and predictive error. We conclude with a discussion of our findings using these two methods and interpretation.

II. Data Structure and Exploratory Analysis

For both the full (23 seasons of player data) and recent (10 most recent seasons) data sets, there are p = 46 covariates falling into the following five categories:

AGE (1)	continuous variable for player age
TOTALS (24)	continuous variables, individual player performance in season
POSITION (1)	categorical variable, five possible positions
ADVANCED (19)	continuous variables, tracking individual player contribution to the team in season
SALARY (1)	continuous variable, normalized player salary in season

Our exploratory analysis of the full data set begins with simple 2D scatter plots of normalized point counts plotted against various predictors, color-coded for player position:



Several linear relationships were observed through visual inspection of plots of the response variable against the predictors, prompting us to start with multiple linear regression models. Assuming a high degree of

correlation among the covariates (the best players should have higher metrics, and vice versa), we then examined a pairwise scatterplot matrix of some of the predictors, which confirmed our initial suspicions:

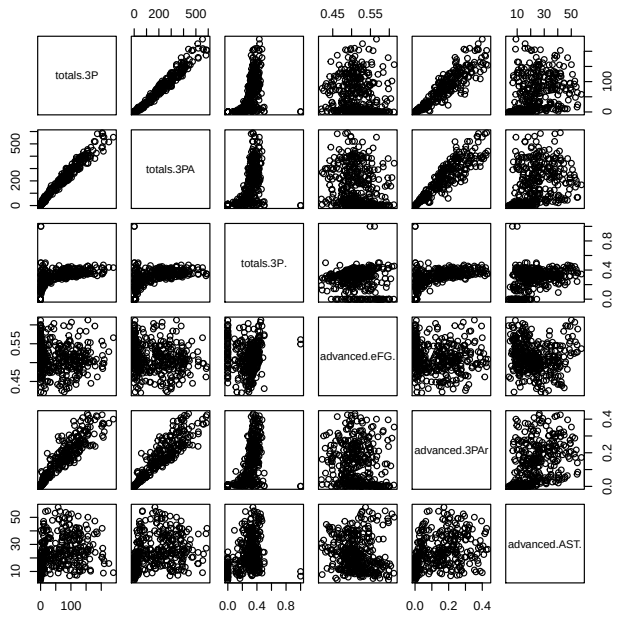


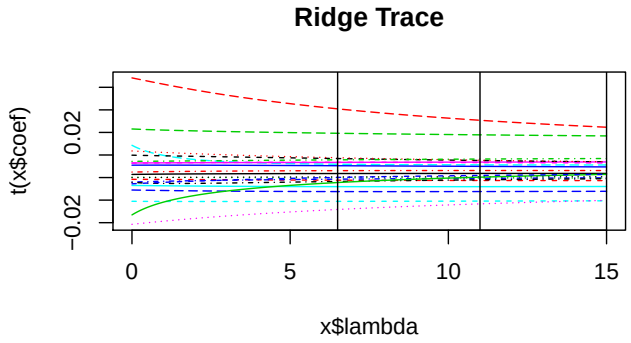
Figure 1: totals.3P is the number of successful 3-Point shots. totals.3PA is the number of attempted 3-Point shots. totals.3P. is the percentage of successful 3-Point attempts. advanced.eFG. is the proportion of effective field goals. advanced.3PAr is the number of field goals attempted from the 3-Point range. advanced.AST is a rating of player assists in point production.

Multicollinearity was examined further by computing variance inflation factors (VIF) associated with estimated standardized regression coefficients. We examined VIF values for predictors in a preliminary model based on our intuition of which predictors would be critical to predicting success. Inspecting this preliminary model, we found several VIF values greatly exceeding 10, indicating serious multicollinearity among the predictors. The mean VIF value was 18.806, implying that the expected sum of squared errors here is close to 20 times greater than it would be if these particular covariates were uncorrelated (Kutner, Nachtsheim and Neter 409).

X	VIF associated with estimated coefficient
totals.Age	1.464878
totals.MP	7.121350
totals.GS	3.337711
totals.3P.	1.789849
totals.2PA	119.688765
Salary	1.504114
advanced.TRB	7.441620
advanced.TOV.	3.405404
advanced.DWS	4.082178
totals.Pos	26.899707
advanced.WS48	4.257684
totals.BLK	4.539006
totals.PF	2.545500
totals.FT	6.580449
advanced.eFG.	8.821838
advanced.DWS	4.082178
totals.PTS	24.857939
totals.2P	106.079415

III. Implementation of Ridge Regression and Results

We implemented ridge regression on a subset of 20 standardized predictors, many of which we assumed to be important in MVP success based on exploratory data analysis as well as our own basketball intuition. Ridge trace was employed to select three candidate models based upon stabilization of the magnitudes of the estimates, as well as minimization of variance inflation factors for the coefficients.

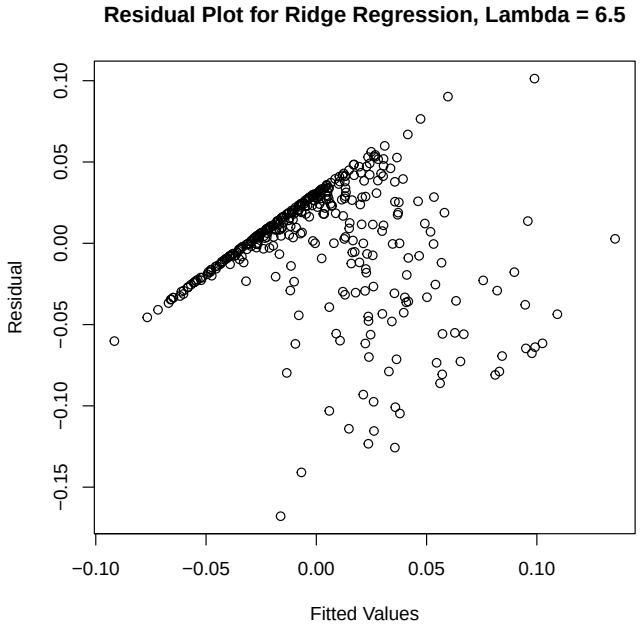


Predictive performance of the three candidate models was assessed in two ways. We first computed the estimated mean-squared error on our sample. Mean-squared error was minimized for the first ridge regression model ($\lambda = 6.5$).

λ	prediction error (in-sample mean-squared error)
6.5	0.001346844
11	0.001356818
15	0.00136498

Applying these candidate ridge models to the sample, we then determined the misclassification rate for each by looking at the proportion of seasons where the model incorrectly predicted the MVP for that season. The computed misclassification error rate was found to be equal and approximately 50 percent for all three ridge models under consideration (error = 0.5217391). Our final model was thus selected based on minimum mean-squared error ($\lambda = 6.5$).

For model-checking and diagnostics, we examined a plot of the residuals for the first ridge model to check the assumptions of constant variance and independence of the error terms.



The residuals appear to be scattered randomly about the zero line, indicating that the model assumptions for ridge regression (constant variance of the error terms, independence) have been satisfied. There is a strong linear trend for several residuals, suggesting a possible lack of independence of the error terms. Checks for normality of the error terms were not performed as there is no underlying distributional assumption for ridge regression.

We conclude this section by deriving 95 percent confidence intervals for the estimated standardized ridge coefficients by bootstrapping. We use the reflection method (Kutner, Nachtsheim and Neter 460) using random X sampling to sample (with replacement) 500 bootstrap samples. Potential issues with model assumptions (evidenced by the residual plot) led us to use random X sampling for the bootstrapping procedure.

Of the predictors used in this procedure, player salary yields the most positive coefficient estimate, indicating that this variable is highly predictive of MVP success compared to the performance metrics studied.

X	Est Std Ridge Coef	95% LB	95% UB
totals.Age	5.725e-04	-3.179e-04	1.681-03
totals.G	3.608e-04	9.631e-05	1.604e-03
totals.GS	-3.436e-04	-1.143e-03	-3.136-04
totals.3P.	-1.840e-03	-2.984e-02	3.112e-02
totals.2PA	1.988e-05	-8.327e-06	4.901e-05
Salary	1.270e+00	-2.910e-01	2.060e+00
advanced.TRB.	1.595e-03	4.958e-04	2.951e-03
advanced.TOV.	3.282e-03	2.102e-03	5.4512e-03
advanced.DWS	2.727e-03	-3.583e-03	5.342e-03
totals.PosPF	-1.566e-03	-1.835e-02	5.406e-03
totals.PosPG	2.414e-02	7.449e-03	4.229e-02
totals.PosSF	7.153e-03	-1.107e-02	1.845e-02
totals.PosSG	1.205e-02	-9.094e-03	2.372e-02
advanced.WS48	5.832e-01	5.949e-01	8.943e-01
totals.BLK	7.548e-05	-9.203e-06	1.775e-04
totals.PF	-2.209e-04	-3.392e-04	-1.344e-04
totals.FT	-1.490e-04	-2.459e-04	-1.390e-04
advanced.eFG.	-2.458e-01	-5.944e-01	-3.044e-01
totals.PTS	7.263e-05	4.781e-05	9.901e-05
totals.2P	1.445e-05	-9.694e-06	8.883e-05
totals.MP	3.520e-06	-9.266e-06	2.380e-05

III. Forward and Backward Regression

Forward and backward regression methods were implemented using the full set of predictors. Two candidate models were selected (one for forward, one for backward) based on minimization of the AIC criterion (the minimum AIC for forward stepwise regression for our final selected model was -2021.76). For each candidate model, we examined the adjusted R^2 and BIC as well. Based on these two criteria, we then found for each of the forward and backward step regression, that the BIC and adjusted R^2 showed the same variables for each respective model. We then looked at the MSE for the orignal forward and backward model, as well as the 2 new models based on the new criteria. The model selected from forward regression mininmized the estimated MSE at 0.00414.

Looking at the residuals versus fitted values for each model, we saw that they were all very similiar in trend and closely scattered around zero, demonstrating that the assumptions of independence and constant variance for linear models were satisfied. On inspection of the normal quantile-quantile plot for this model, we see that the trend is mostly positive and linear, suggesting that the normality assumption of the error term for linear models has been satisfied. Computations of Cook's Distance for each sample are all less than 1, indicating that no individual sample is unduly influencing the regression model fitted.

We then computed the misclassification rate. Our model selected from forward stepwise regression gave a misclassification rate of 0.4782609 - a slight improvement on the model chosen from ridge regression.

X	Est Coef	Std Error	2-sided p-value	95% LB	95% UB
(Intercept)	5.647e-01	1.650e-01	0.001	2.402e-01	8.891e-01
advanced.WS	3.410e-02	3.504e-03	$<$ 2e-16	2.721e-02	4.099e-02
Salary	3.140e+00	9.087e-01	0.001	1.353e+00	4.927e+00
advanced.USG.	-2.975e-04	2.908e-03	0.919	-6.016e-03	5.421e-03
totals.FT	-3.054e-04	9.193e-05	0.001	-4.862e-04	-1.246e-04
advanced.TOV.	6.238e-03	3.025e-03	0.040	2.889e-04	1.219e-02
advanced.TS.	-1.028e+00	1.765e-01	1.290e-08	-1.375e+00	-6.806e-01
totals.PF	-3.448e-04	1.009e-04	0.001	-5.433e-04	-1.464e-04
advanced.AST.	-1.252e-03	5.303e-04	0.019	-2.295e-03	-2.093e-04
totals.FG	-9.456e-05	1.643e-04	0.565	-4.177e-04	2.286e-04
totals.MP	-1.489e-04	2.891e-05	4.310e-07	-2.058e-04	-9.203e-05
advanced.PER	-4.903e-03	3.262e-03	0.134	-1.132e-02	1.513e-03
totals.TOV	5.163e-04	2.087e-04	0.014	1.060e-04	9.266e-04
advanced.STL.	-1.492e-02	6.075e-03	0.015	-2.686e-02	-2.970e-03
totals.PTS	1.946e-04	8.963e-05	0.031	1.832e-05	3.709e-04

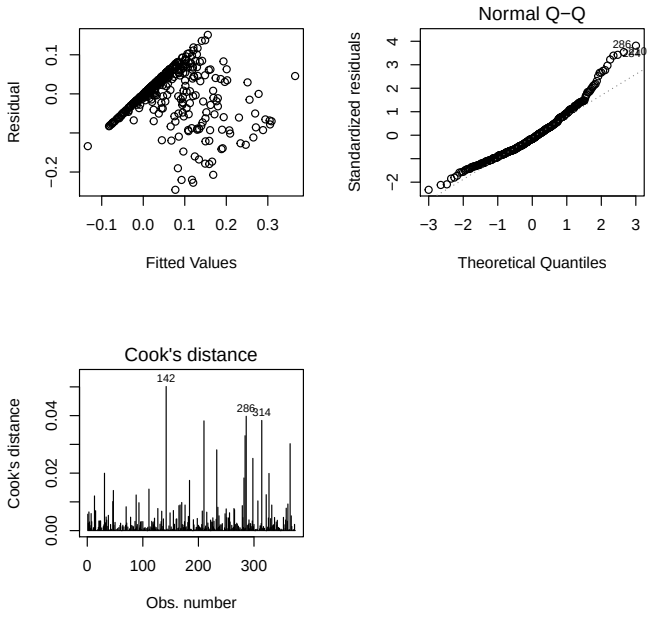


Figure 2: For the model selected from forward stepwise regression: Plot of the residuals (top-left), Normal Q-Q Plot (top-right), Cook's Distance (bottom)

IV. Conclusion

In our two models, we identified several important predictors that were most associated with MVP success. In both the ridge regression and forward stepwise regression models, the estimated coefficient for player salary was found to be statistically significant (for forward stepwise, the p-value was 0.001). In our ridge model, we noticed other key predictors associated with success. The estimated coefficients for the player positions of Point guards and Shooting guards (totals.PosPG and totals.PosSG, respectively) were both relatively large in magnitude compared to the other estimates. This could be tied to guards historically winning the MVP award and the tendency for guards to be the leaders of team and creating plays for their teammates. Win share per 48 minutes (advanced.WS48) was another key variable and this could be tied to how much a player contributes to a team winning. In the stepwise model, we notice that win share (advanced.WS) was also a key variable, as well as true shooting percentage (advanced.TS, positively associated) and estimated steals (advanced.STL, negatively associated) for a player. Large positve values for True shooting percentage and

win share might correspond to a player who takes smarter shots on the court.

The forward step-wise model improved our prediction error by 5 percent. Interestingly, stepwise regression resulted in a more parsimonious model compared to the ridge model. Checking both models, we saw that the residuals were similar in trend and were scattered randomly around the zero line. Yet for both these residuals plots there seemed to be a strong linear trend for negative fitted values. The Normal Q-Q Plot for the forward stepwise model demonstrated a strong linear trend with a slight left skew. This suggests that there might be a slight departure in normality of the error terms. The Cook's distance (less than 1 for all observations) for the forward model also showed that there were no highly influential points, indicating that none of the observations were outliers.

Both models validated our hypothesis that salary could be tied to MVP voting success and could indicate that player salary correlates to players who play in larger markets and get more exposure in the media. Our models show that popularity is highly indicative of MVP success and this statistical analysis could be used to both accurately predict MVP success and validate criticism from some in the media that the MVP award is a popularity contest not based on performance. Finally, applying our model to this year's player statistics, we predict that Kevin Durant will win the MVP with 29 percent of the vote, with Lebron James in close second place with 20 percent of the vote.

Code and Author Contributions

Our data set as well as R code for data cleaning can be found here: `www.github.com/velaraptor/nba_mvp_metric` under the folder "mvp data". This Latex document was created using R Sweave. Code for this document, which includes all computations for analysis, can be found through the same github link under "report.Rnw". C.V. handled data cleaning and coding/analysis for stepwise regression. D.S. conducted the exploratory analysis and handled coding/analysis for ridge regression. Paper and presentation were prepared jointly.

References

1. `www.basketball-reference.com`
2. `www.eskimo.com/~pbender`
3. Härdle, W., Simar, L. (2012) Applied Multivariate Statistical Analysis. 3rd ed., Springer Verlag, Heidelberg. ISBN 978-3-642-17228-1, e-ISBN 978-3-642-17229-8 (539 p), DOI:10.1007/978-3-642-17229-8
4. Kutner, M., Nachtsteim, C., Neter, J. (2004) Applied Linear Regression Models. 4th ed., McGraw-Hill Irwin.
5. `http://fansided.com/2014/04/22/kevin-durant-lebron-james-meaning-mvp/#!Hf1Gk`