SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

The Rise of Chef Curry

Studying Advanced Basketball Metrics with Quantile Regression in SAS®

USERS PROGRAM

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

The Rise of Chef Curry: Studying Advanced Basketball Metrics with Quantile Regression in SAS®

Taylor K. Larkin and Denise J. McManus
The University of Alabama

Introduction

In the 2015-2016 season of the National Basketball Association (NBA), the Golden State Warriors achieved a record-breaking 73 regular season wins. This accomplishment could not have been done without their reigning Most Valuable Player (MVP) champion Stephen Curry (featured in Animation 1) and his historic shooting performance (see Figures 1 and 2). Shattering his previous NBA record of 286 three point shots made during the 2014-2015 regular season, he accrued an astounding 402 in the next one. With increased emphasis on the advantages of the three point shot and guard-heavy offenses in the NBA today, organizations are naturally eager to investigate player statistics related to shooting at long ranges, especially for the best of shooters. Furthermore, the addition of more advanced data collecting entities such as SportVU invites an incredible opportunity for data analysis, moving beyond simply using aggregated box scores. This work uses quantile regression within SAS® 9.4 to explore the relationships between the three point shot and other relevant advanced statistics, including some SportVU player tracking data, for the top percentile of three point shooters from the 2015-2016 NBA regular season.

Why Quantile Regression?

One of the most popular techniques for exploring explanatory relationships in observational data is linear regression. With this method, the idea is to find a set of predictor variables that best explain the conditional mean of a response variable. In some cases, it may be advantageous to explore different parts of the response's distribution, particularly when the extreme values are important. In these situations, quantile regression (Koenker & Bassett, 1978) is a useful alternative. This technique allows for the estimation of effects at specific quantiles (or percentiles) of the response variable. Mathematically, it minimizes the following objective function for a given quantile τ:

$$\min_{\beta}\Big[\sum_{i\in\{i:y_i\geq \mathbf{x}_i'\beta\}}^{n}\tau|y_i-\mathbf{x}_i'\beta|+\sum_{i\in\{i:y_i<\mathbf{x}_i'\beta\}}^{n}(1-\tau)|y_i-\mathbf{x}_i'\beta|\Big]$$

for some data $\{y_i, \mathbf{x}_i, i=1,...,n\}$ (Chen, 2005). Advantages to using quantile regression include: invariance to monotonic transformations, no distributional or constant variance assumption on the errors, and robustness towards outliers. Moreover, it is possible to find predictive relationships at specific quantiles that otherwise would not be significant when modeling the conditional mean (Cade & Noon, 2003). Therefore, as presented in this study, one can use quantile regression to identify the most important advanced player statistics associated with the best three point shooters in the NBA, as oppose to trying to find those associated with the average three point shooter.

Animation 1: Select Stephen Curry highlights from this past season ([NBA], 2016).

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

Visualizations and Data Collection

Taylor K. Larkin and Denise J. McManus
The University of Alabama

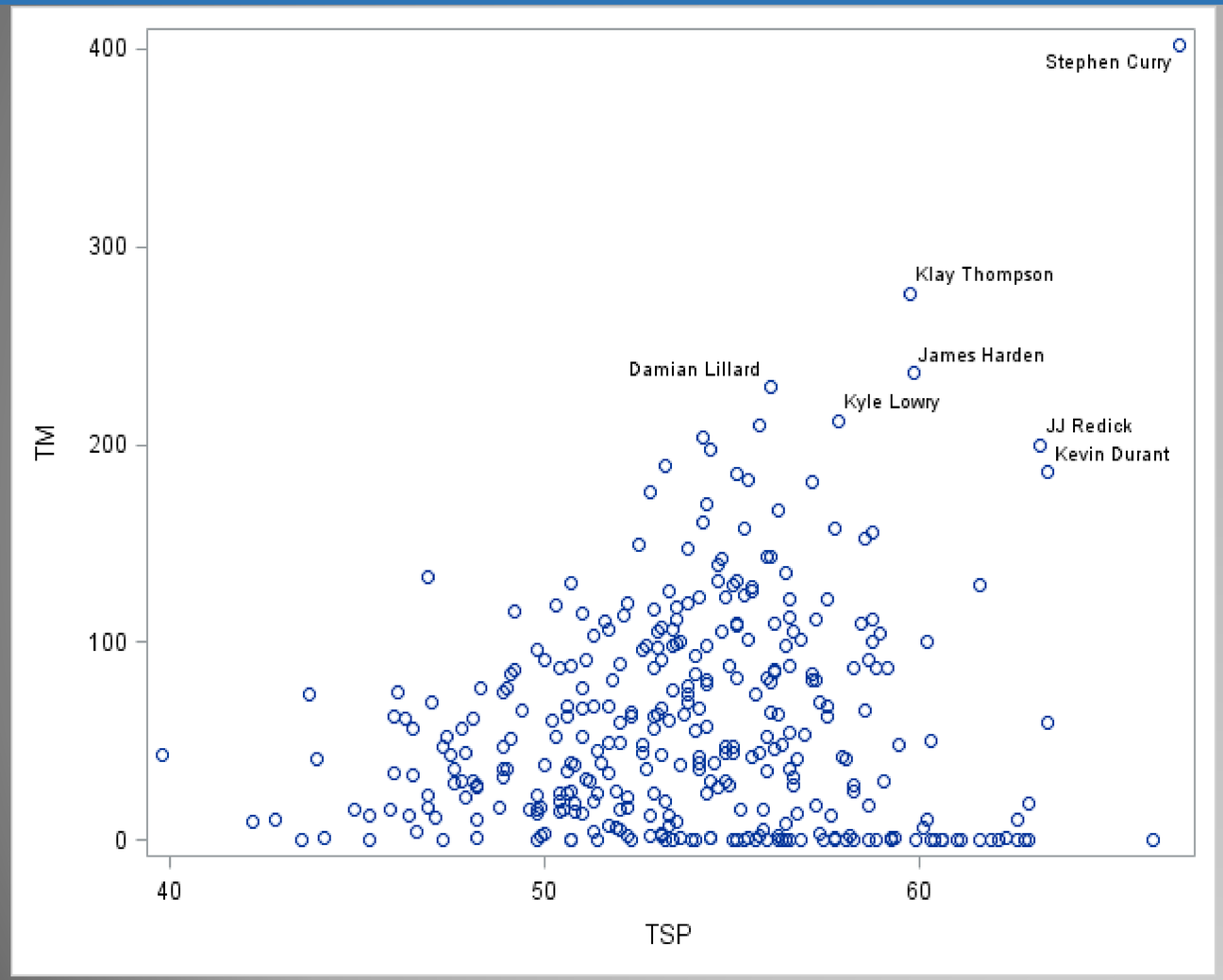

Figure 1: Number of three point shots made versus true shooting percentage for the 2015-2016 season.

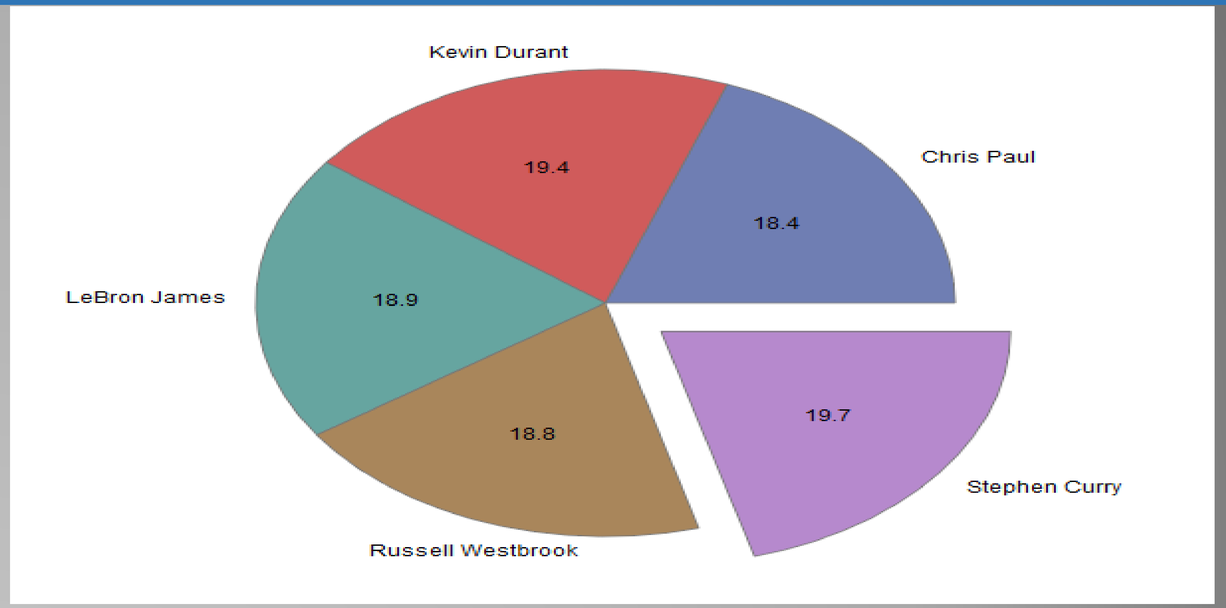

Figure 2: Pie chart of the top 5 players with the highest Player Impact Estimate (PIE). PIE measures the overall contribution a player has in a game for positive and negative events (e.g. points, rebounds, fouls, etc.)

Selecting Data for Analysis

Thanks to the STATS SportVU player tracking system, the most advanced data such as a player's spatial characteristics during a game can be captured. This system uses six cameras to track real-time positions and ball activity 25 times per second in each arena, resulting in a wealth of new information (SportVU Player Tracking, n.d.). Because STATS is an official NBA Partner, much of these advanced statistics can be found online via the NBA's statistics page (NBA, n.d.). All the data for this study can be found here. Because of the amount of data present, a diverse set of advanced player statistics is manually selected. Care is taken to ensure that no two statistics have a correlation coefficient greater than 0.75 in magnitude. Since the goal is to investigate the advanced statistics associated with good three point shooters, the number of three pointers made (*TM*) serves as the response variable. Furthermore, only players who played at least half of the regular season are considered. The resulting efforts yield a dataset of 26 predictor variables for modeling *TM* for 342 players during the 2015-2016 regular season.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

THE UNIVERSITY OF ALABAMA® | Culverhouse College of Commerce

Dataset and Experimental Set-up

Taylor K. Larkin and Denise J. McManus
The University of Alabama

PROC QUANTSELECT

In order to simplify the analysis further, a data-driven variable selection step is included. Using PROC QUANTSELECT, one can empirically decide which predictor variables to include for a subsequent quantile regression model. In this study, the adaptive lasso procedure is used (Zou, 2006) to determine the useful predictor variables at the 95th percentile. All model selection options are set to "validate" which chooses the model with the smallest average check loss on an independently held-out sample of data. The partition is set to be a 70/30 training/validation split, respectively. The stopping horizon in the model selection process is set to three in order to find an adequate local minimum. Since the number of observations is small, the variable selection procedure may yield a different set of predictor variables, depending on a particular data partition. To circumvent this issue, a macro do loop is written to perform PROC QUANTSELECT 1,000 times at different random seeds. The total number of times a predictor variable is chosen during the 1,000 iterations is then calculated using PROC FREQ. Those which are not chosen at least 10% of the time are omitted from further analysis. The final set of 11 predictor variables can be found in Table 1.

PROC QUANTREG

To fit a quantile regression model, PROC QUANTREG can be used. Not only can this procedure execute quantile regression for some specified quantile level, but it can also fit multiple models across varying percentiles. This is known as quantile process regression. By plotting the predictor variables across a range of quantile levels, one can explore the effects at different parts of the conditional distribution of *TM*. Furthermore, along with the estimated coefficients, the confidence intervals can be graphed. For this study, the confidence intervals are estimated via the resampling option which uses Markov chain marginal bootstrap (Kocherginsky, He, & Mu, 2005). Using this method to calculate confidence intervals only requires the assumption that the observations are independent. Since each observation represents a different player, this is reasonable to enforce. To increase stability, the number of repeats is set to 10,000. Starting at the 5th and ending at the 95th percentile, a quantile regression model is estimated in 5% increments. The predictor variables are standardized prior to running PROC QUANTREG to more fairly compare the coefficients. The estimated coefficients, the respective confidence intervals, and t-values are shown in Figures 3 and 4.

Predictor Variable	Description
Avg_Reb_Dist	The average distance of a rebound measured in feet per game.
Avg_Sec_Per_Touch	The average number of seconds per touch by a player per game.
Dist_Mi	Distance ran by a player measured in miles per game.
Drives	The number of times a player drives to the basket per game.
Elbow_Touch	The number of touches made by a player at the elbow per game.
Net_Rtg	Difference between a player's offensive and defensive ratings. Offensive/defensive ratings represents the number of points a team scores/allows while that player is on the court per 100 possessions.
Post_Touch	The number of touches made by a player at the post per game.
TO_Rt	The number of turnovers a player averages per 100 of their own possessions.
Touches	The number of times a player touches and possesses the ball during the game per game.
TSP	Measure of shooting efficiency when taking into account two point, three point, and free throw shots.
USGP	The percentage of a team's offensive possessions that a player uses while on the court per game.

Table 1: Data dictionary for the final set of predictor variables included for the quantile regression model.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

Quantile Process Regression Results

Taylor K. Larkin and Denise J. McManus
The University of Alabama

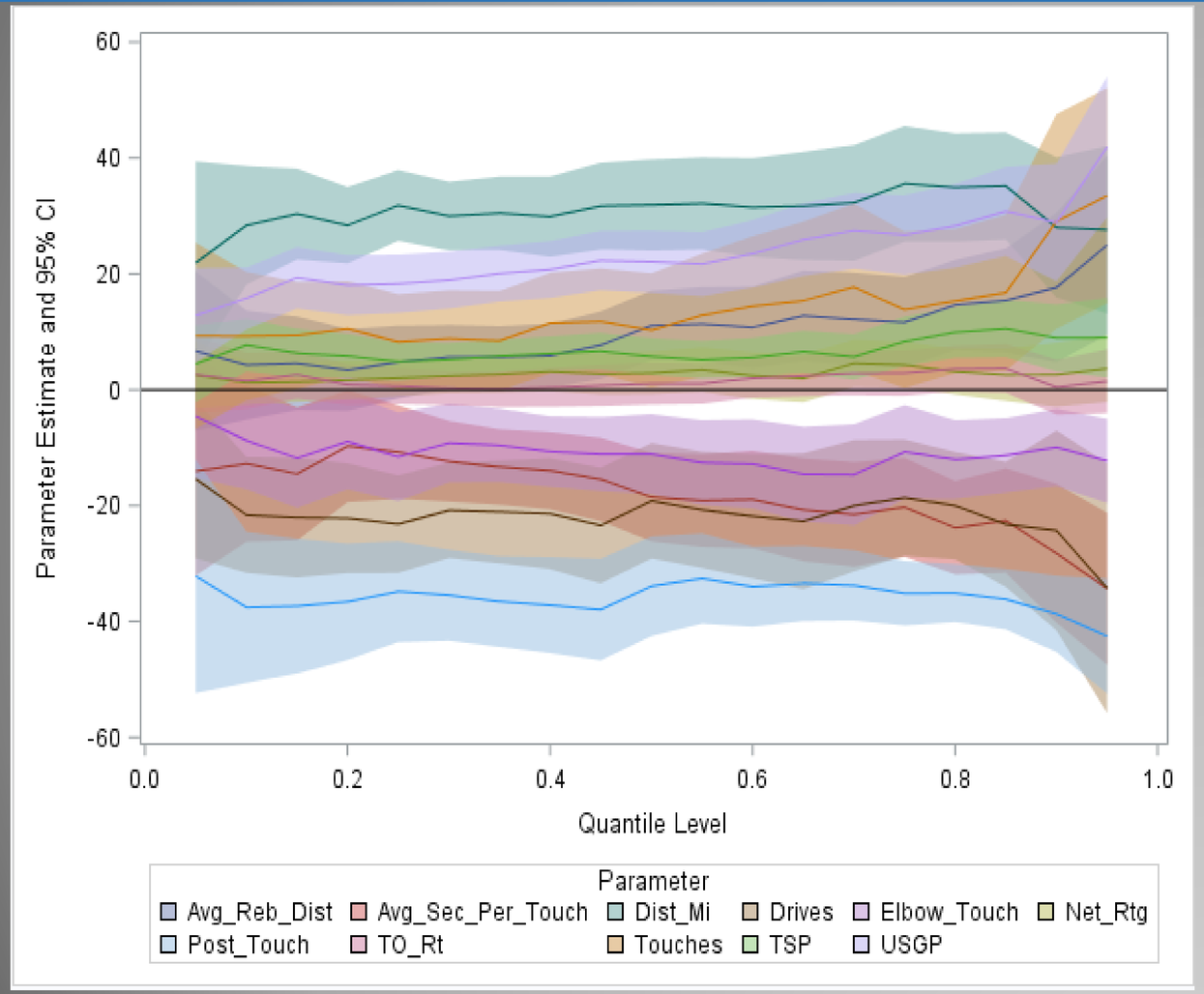

Figure 3: Evolution of coefficient paths across varying quantiles.

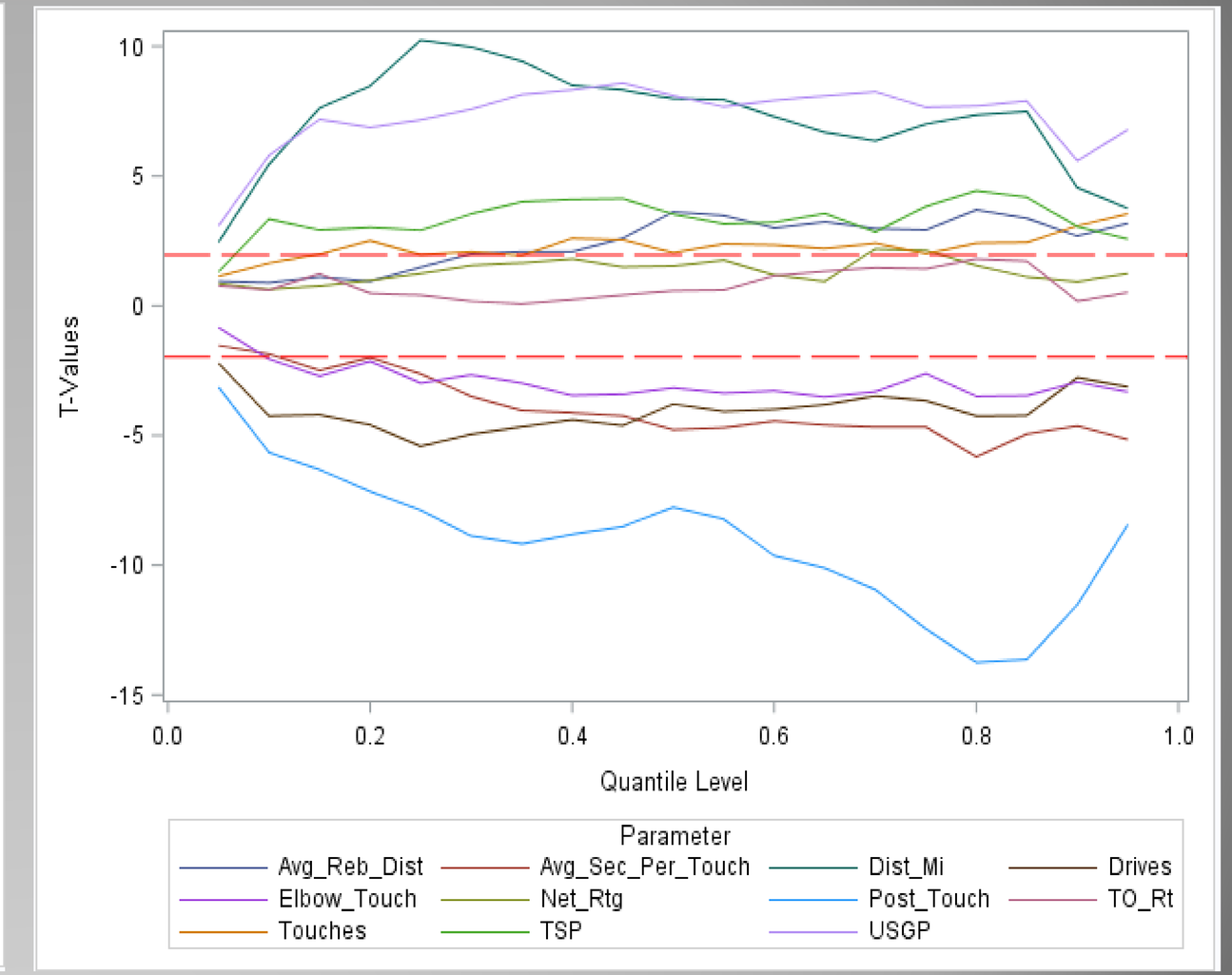

Figure 4: Evolution of t-values across varying quantiles. Red dashed lines denote 0.05 significance level.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

THE UNIVERSITY OF ALABAMA® | Culverhouse College of Commerce

What Does It All Mean?

Taylor K. Larkin and Denise J. McManus
The University of Alabama

Predictor Variable	Effect	Quantile P-value	Linear P-value	Disagreement
Avg_Reb_Dist	+	0.001	0.000	
Avg_Sec_Per_Touch	-	0.000	0.000	
Dist_Mi	+	0.000	0.000	
Drives	-	0.002	0.000	
Elbow_Touch	-	0.001	0.000	
Net_Rtg	+	0.213	0.011	X
Post_Touch	-	0.000	0.000	
TO_Rt	+	0.608	0.271	
Touches	+	0.000	0.001	
TSP	+	0.011	0.000	
USGP	+	0.000	0.000	

Table 2: Comparing p-values between quantile regression at the 95th percentile and linear regression.

Discussion

- *Avg_Reb_Dist*, *Touches,* and *USGP* all have positive relationships with *TM* while *Drives* and *Post_Touch* have a negative relationship. Each exhibit the largest effect at the 95th percentile. This makes sense given the fact that Curry as well as other top tier shooters typically play the guard position. These positions will naturally possess the ball more, attack the lane less if they are shooters, and be usually situated away from the post for touches and rebounds.
- *Avg_Sec_Per_Touch* has a significant, negative relationship and generally decreases as the quantiles increase closer to one. This indicates that great three point shooters do not hold the ball for relatively long periods of times in the presence of the other predictor variables. More than likely, this can be attributed to catch and shoot three pointers (example featured in the first few seconds of Animation 1). Therefore, defenses could focus on forcing Curry and other great shooters to take isolated three point shots (note Kevin Love's defense on Curry in the final seconds of game seven in the 2015-2016 NBA finals).
- If one used linear regression to draw inference, *Net_Rtg* would be declared as an informative predictor variable when in actuality, it is not the case for the best three point shooters (see Table 2).
- Depending on the response, different predictor variables may be selected (see Table 3).

Predictor Variable	Response = *TM*	Response = *PIE*
Avg_Reb_Dist	154	419
Avg_Sec_Per_Touch	711	49
Dist_Mi	905	971
Drives	383	223
Elbow_Touch	290	9
Net_Rtg	101	101
Post_Touch	934	20
TO_Rt	168	83
Touches	578	536
TSP	515	1000
USGP	923	1000

Table 3: Variable selection frequencies when modeling TM and *PIE.*

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

The Rise of Chef Curry: Studying Advanced Basketball Metrics with Quantile Regression in SAS®

Taylor K. Larkin and Denise J. McManus
The University of Alabama

Conclusions and Future Work

In this study, quantile regression is used to investigate the relevant player statistics associated with the ability to shoot at long ranges. Using PROC QUANTSELECT and a macro do loop, variable selection is performed 1,000 times in order to further reduce the manually chosen set of 26 advanced player statistics to 11. Then, implementing quantile process regression via the PROC QUANTREG procedure, the estimated coefficients, confidence intervals, and t-values are plotted for each predictor variable. Results show that quantile regression leads to a more comprehensive and accurate assessment for analyzing the best three point shooters as opposed to modeling the conditional mean. In addition, depending on the response chosen, different predictor variables can be selected via the PROC QUANTSELECT procedure.

While this study presents interesting results, these are based on only one season of data. Including more data may yield some differences. In addition, the p-values should be interpreted with care, as observational studies are more susceptible to systematic error compared to an experimental design (Schuemie, Ryan, DuMouchel, Suchard, & Madigan, 2014). Hence, it is important to supplement these with scientific reasoning (American Statistical Association, 2016), which is what this study has attempted to do in this section. Furthermore, although some of these findings may seem trivial, it can be the groundwork for analyzing other kinds of player information such as wearable technologies or even personality tests. Modeling the more intangible aspects of a basketball player in this way can unlock truly revolutionary discoveries.

References

1. American Statistical Association. (2016, March 7). *AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES* [Press release]. *ASA News*. Retrieved July 20, 2016, from https://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf

2. Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, *1*(8), 412-420.

3. Chen, C. (2005). An introduction to quantile regression and the QUANTREG procedure. In *Proceedings of the SUGI 30 Conference. Cary, NC: SAS Institute Inc.*

4. Kocherginsky, M., He, X., & Mu, Y. (2005). Practical Confidence Intervals for Regression Quantiles. *Journal of Computational and Graphical Statistics, 14*(1), 41-55. Retrieved from http://www.jstor.org/stable/27594096

5. Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33-50.

6. NBA. (n.d.). *NBA.com/Stats*. Retrieved April 25, 2016, from http://stats.nba.com/

7. [NBA]. (2016, Feb 27). *Steph Curry's Best Three-Pointers of 2015-16 Season*. [Video File]. Retrieved July 18, 2016, from https://www.youtube.com/watch?v=XbS9M4TO9z8.

8. Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., & Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*, *33*(2), 209-218.

9. SportVU Player Tracking | STATS SportVU Tracking Cameras. (n.d.). Retrieved July 18, 2016, from http://www.stats.com/sportvu/sportvu-basketball-media/

10. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.