Players' Average Marginal Contribution in Basketball and Generalized Shapley Value

Rodolfo Metulini[1], **Giorgio Gnecco**[2]

July 14th, 2021

EURO 2021 - 31st European Conference on Operational Research

[1]Dept. of Economics and Statistics (DISES) - University of Salerno
[2]AXES Research Unit, IMT - School for Advanced Studies, Lucca

Context

- Data analytics is a common practice to help coaches and staff on the strategy to adopt in basketball (**Nikolaidis**; **2015**, **Sarlis**; **2020**) and it is increasingly used due to large amounts of different type of data.
- Literature has been devoted to both team (**Yang et al.**; **2014**, **Moreno and Lozano**; **2014**, **Hofler and Payne**; **2006**) and players performance (box-score **Cooper et al.**; **2009**; **Fearnhead and Taylor**; **2011**; **Page et al.**; **2013**, shooting variables **Piette et al.**; **2013**; **Metulini and LeCarre**; **2020**; **Sandri et al.**; **2020**, synthetic metrics **Terner and Franks**; **2020**), but, generally, separately.
- Team performance may be viewed as a network where each play represents a *pathway* through which ball and players cooperatively move to the goal. Players and team should be evaluated together, with an approach that take both of them into account.
- The line of research of this work is that of estimating players impact on winning the game (**Deshpande and Jensen**; **2016**).
- Aim: estimating the importance of players using average marginal contribution, borrowing the idea of the Shapley value (**Shapley**; **1953**) from Cooperative Game Theory (in the spirit of **Hernandez and Sanchez**; **2010** and **Hiller**; **2018**).

Motivation

- In basketball, five players in both teams rotate on the court. The five players of each single team on the court in that moment represent the lineup, while the ten players in the court represents the encounters.
- Ranking players and lineups has been addressed, for example, in **Barrientos et al.**; **2019**, via a Bayesian approach, for the analysis of encounters. **Kalman and Bosch**; **2020** detect more efficient lineups.
- The Shapley value has successfully been used in many political and economic games.
- The utilization of this value has not massively been percolated to team sports analysis (an exception being the works on Soccer by **Auer and Hiller**; **2015**; **Hiller**; **2015**).
- To the best of our knowledge, the Shapley value has never been used to evaluate players' performance in basketball, except for the proceeding article by **Yan et al.**; **2020**.

Methodological approach

- In the Shapley value, a player marginal utility is computed based on the difference between the values assumed by a characteristic function $v(.)$ - that measures the cohesion (performance) of each combination of players - calculated, respectively, with and without him in the court.
- A generalization of the Shapley value to the case of ordered coalitions, which was proposed by **Nowak and Radzik**; **1994** is more suitable than the Shapley value itself:
  - we need to take in account that only five players for team can play simultaneously (players that *virtually* enters a coalition after the fifth player has zero marginal utility, or worth)
  - this leads to multiple values for the coalition made of all players in the team (grand coalition): this issue has been only addressed by Nowak and Radzik.
- We model the generalized characteristic function in terms of the probabilities of winning the game and we estimate players' marginal contribution by proposing a three step approach.
- We employ our strategy to National Basketball Association (NBA) real data.

The three steps

1. We estimate logistic regression coefficients based on the full set of games and at single game level, where the dependent variable is dichotomous (win=1, defeat=0) and the *four Dean's factors* (**Oliver**; **2004**; **Kubatko et al.**; **2007**) are used as explanatories.
2. The coefficients estimated in step 1 are used to derive the probability to win associated to each lineup, by replacing the game level explanatories with those computed at lineup level.
3. With the probability of winning (generalized characteristic function) at hand for all the lineups, we compute, for each player, different versions of the generalized Shapley value.

With the generalized Shapley value for each player we put in relation:

- player's average marginal utility, and
- player's income

finding those players whose average marginal utility is higher than expected.

Generalized Shapley value - i

- In cooperative game theory, a "generalized" coalitional game (**Nowak and Radzik**; **1994**) is defined as a pair (N, v) where $N = \{1, 2, ..., n\}$ is the player set (cardinality $= n$).
- v is the generalized characteristic function, which assigns to every ordered coalition T extracted from the set N a certain worth $v(T)$ reflecting the "abilities" of such an ordered coalition.
  - This definition differs from the classical one of a coalitional game, whose characteristic function is defined on the set of unordered coalitions (**Maschler et al.**; **2013**, ch. 17).
- In the case of basketball, we have non-zero worth only on coalitions with cardinality $m = 5$, whereas $n > m$ is the total number of players rotating in the court in a game.

Generalized Shapley value - ii

- Let the elements of each ordered coalition $T \in \mathcal{T}$ denoted by $T_1, \ldots, T_{|T|}$, where the index refers to the order according to which a player enters that coalition, in a "virtual" process of its construction;
- for any ordered coalition T and any player i, let $T(i)$ denotes the ordered (sub)coalition formed by the players that precede i in T (this coincides with T if i is not present in T).
- The generalized Shapley (or Nowak-Radzik, NR) value of player i in a generalized coalition game is the average of the marginal contribution of that player when he enters an ordered subcoalition (of any cardinality) of an ordered coalition T with cardinality $|T| = n$:

$$\phi_i^{NR}(N, v) = \frac{1}{n!} \sum_{T \in \mathcal{T} \text{ with } |T|=n} (v((T(i), i)) - v(T(i))) \; . \qquad (1)$$

Generalized characteristic function - i

- We consider two alternatives for the generalized characteristic function $v(.)$, denoted respectively by $v_1(.)$ and $v_2(.)$:
  1. $v_1(.)$ is computed based on the probability of winning the game ($P(Win)$) for any specific lineup.
  2. $v_2(.)$ based on both $P(Win)$ and the probability of occurrence of that lineup on the court ($P(Occ)$).
- The two corresponding generalized Shapley values are, respectively, the "unweighted generalized Shapley value"($\phi_i^{NR}(N, v_1)$) and the "weighted generalized Shapley value" ($\phi_i^{NR}(N, v_2)$).

Generalized characteristic function - ii

- According to $v_1(.)$, we let (when $|(T(i), i)| = 5$)

$$v_1((T(i), i)) = P(Win)_{(T(i),i)} \tag{2}$$

be the probability of winning the game for the ordered (sub)coalition of players $(T(i), i)$, and we let

$$v_1(T) = P(Win)_{T(i)} \tag{3}$$

be the probability of winning the game for the ordered (sub)coalition of players $T(i)$, which does not contain player i ($|(T(i))| = 4$).

- According to $v_2(.)$ we use, respectively:

$$v_2((T(i), i)) = P(Occ)_{(T(i),i)} P(Win)_{(T(i),i)} \tag{4}$$

and

$$v_2(T) = P(Occ)_{T(i)} P(Win)_{T(i)}\,, \tag{5}$$

where $P(Occ)_{(T(i),i)}$ and $P(Occ)_{T(i)}$ are the probabilities of occurrence on the court of the ordered (sub)coalitions of players $(T(i), i)$ and $T(i)$, respectively.

Observed generalized Shapley value

- Let $\mathcal{L}_i$ be the set of *observed* lineups (coalitions with cardinality$=5$) in which player i appears, one gets the following estimate of his "empirical" generalized Shapley value:

$$\hat{\phi}_i^{NR}(N, v_k) = \frac{5}{n}\frac{1}{5|\mathcal{L}_i|}\sum_{L\in\mathcal{L}_i}(\hat{v}_k(L) - 0) = \frac{1}{n|\mathcal{L}_i|}\sum_{L\in\mathcal{L}_i}\hat{v}_k(L)\,, i = 1, 2. \tag{6}$$

  - $\hat{v}_k(T(i)) = 0$ because of zero worth when cardinality$=4$
  - The inclusion of factor $\frac{1}{5}$ is needed since, for any specific quintet, each player has the same probability of being the fifth to join all the other members of that quintet
  - The other factor $\frac{5}{n}$ expresses the probability that player i enters in one of the first 5 positions

Logistic model & P(Win) - i

- Studies using Statistics/machine learning techniques estimating the probability to win a game generally achieve an 70-80% accuracy:
  - Artificial Neural Networks (**Zhang**; **2000**);
  - Naive Bayes Classifier (**Langley et al.**; **1992**);
  - Support Vector Machines (**Cortes and Vapnik**; **1995**);
  - Logistic Regression (**Hosmer et al.**; **2013**), etc... .
- We adopt a logistic regression model strategy along with the four Dean's factors (**Oliver**; **2004**, **Kubatko et al.**; **2007**) which are well-known and agreed in the literature ($R^2 \sim 0.9$):
  1. effective field goal percentage (eFG%): $\frac{(FG+0.5*3P)}{FGA}$,
  2. turnover percentage (TOV%): $\frac{TOV}{(FGA+0.44*FTA+TOV)}$,
  3. offensive rebound percentage (ORB%): $\frac{ORB}{(ORB+OppDRB)}$,
  4. free throws percentage (FT%): $\frac{FT}{FGA}$.
- To account for both teams' features the four Dean's factors are actually eight (we use notation *off* when referring to the considered team, *def* when referring to the opponent team).

Logistic model & P(Win) - ii

Logistic regression model, to be estimated in step 1, reads, for game i, as:

$$log \frac{P(Y_i = 1 \mid \boldsymbol{X_i})}{P(Y_i = 0 \mid \boldsymbol{X_i})} = \boldsymbol{X}_i \beta \tag{7}$$

- The left part of the equation is the log-odds of $\boldsymbol{Y}$ conditional to $\boldsymbol{X}$.
- $\boldsymbol{Y}$: the response binary variable representing the outcome of the games, $\boldsymbol{Y}_i \in \{0, 1\}$, $i = 1, ..., g$, where g is the number of games.
- $\boldsymbol{X}_i$: the $i-$th row of the design matrix $\boldsymbol{X}$ with g rows and p columns (p=8, the eight Dean's factors used as explanatory variables, *eFG%_off*, *eFG%_def*, *TOV%_off*, *TOV%_def*, *ORB%_off*, *ORB%_def*, *FT%_off*, *FT%_def*, computed at game level).
- β: the vector containing the p regression parameters associated with the explanatory variables. These parameters have to be estimated from the data.

Logistic model & P(Win) - iii

- Single lineups do not play the full match, thus making not straightforward to determine variable Y for each quintet and thus to estimate logistic model at lineup level.
- So, in the second step, on dataset $\tilde{\boldsymbol{X}}$, where the eight Dean's factors are expressed at single lineup level[3] we predict the probability to win the game $P(Win)_{L_j}$ on each lineup L_j by using vector $\hat{\beta}$ estimated in step one:
  - let $\tilde{\boldsymbol{X}}_j$ be the j-th row of the matrix $\tilde{\boldsymbol{X}}$ with l rows (number of different considered lineups) and p=8 columns (the eight Dean's factors computed at lineup level), the probability to win the game for the lineup L_j is:

$$P(Win)_{L_j} = \frac{exp(\tilde{\boldsymbol{X}}_j \hat{\beta})}{1 + exp(\tilde{\boldsymbol{X}}_j \hat{\beta})}, \; j = 1, ..., l\,. \tag{8}$$

[3] An average of the Dean's factors is taken for the opponent lineups, considering that they are not fixed

Estimating players marginal utilities

- In the third step the generalized Shapley values are computed for each player, by using Equation 6 along with the winning probabilities provided by Equation 8 (step 2) computed for each lineup.

Data

- Play-by-play of all NBA games (both regular seasons and play-offs) for 14 seasons (2004–2005 – 2017–2018), available thanks to an agreement with BigDataBall (UK) (`www.bigdataball.com`).
- Start/end of the period, made/missed 2 points shots, made/missed 3 points shots, made/missed free throws, offensive/defensive rebounds, assists, steals, blocks and fouls, for each game and for both teams, associated with the event timestamp and with the lineup of both the two teams.
- $x-$axis and $y-$axis position, related, respectively, to court length and width, is also available for shots (made or missed).
- All features for both $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ have been computed from play-by-play dataset.

Logistic results

- A linear regression model with the Ordinary Least Squares (OLS) as done in **Kubatko et al.**; **2007**, using full set of games ($g = 18,109$) from all 14 seasons (play-offs included).
- When the dependent variable is binary, logistic model is preferable to linear regression to prevent us from having the estimated probability of success not included in the range [0,1] (**Wooldridge**; **2010**).
- According to the logistic results, we use the following values for vector $\hat{\beta}$ to be used to determine $P(Win)$ for each lineup L_j in the second step of our analysis:

$$[101.79,\ -101.30,\ -95.07,\ 93.99,\ 30.53,\ -31.10,\ 24.13,\ -23.99]'\ .$$

- ROC curve (**Krzanowski and Hand**; **2009**) confirms the high level of classification accuracy (AUC $= 0.993$). The hit-rate (**Bensic et al.**; **2005**) stands to 0.958.

Shapley values results

- We retrieve the winning probabilities for all relevant lineups of Houston Rockets in 2017/18 regular season (65-17, best record).
- We consider in the analysis only the 99 different lineups that were on the court more than 2 minutes in the 2017/18 regular season.
- With these lineups we cover about 85% of the total time of play.
- 13 different players rotate on the court considering the 99 lineups.
  - Just for curiosity, in most games, the starting lineup was: Chris Paul (*point guard*), James Harden (*shooting guard*), Trevor Ariza (*small forward*), Ryan Anderson (*power forward*) and Clint Capela (*center*).
- We compute the estimated values for winning probability for each lineup (according to equation 8), then we determine the (two versions of) generalized Shapley value for each player, as in Equation (6).
- According to results, best lineups can be employed during the game, in terms of role composition.

Marginal impact vs. income

- figures report the income[4] for the 13 Houston Rockets' players (season 2017/18) along with the weighted and the unweighted generalized Shapley values.
- We run a simple linear regression model with OLS on the n=13 sample, where the weighted (or unweighted[5]) generalized Shapley value is the dependent variable (Y), and income (in million of dollars) is the explanatory variable X_1.
  - The chart shed lights on which player presents an higher marginal utility than expected, according to his income.

[4] data found from the website `basketballinsiders.com`.

[5] We do not report results because the effect is not significant

Concluding remarks

- We provide a robust measure of player's utility on winning the game, using a method (Shapley value) which has never been used to this scope.
- With this measure, we can:
  - Rank players' utility within a team,
  - compare single player's average marginal utility with player's economic value, and
  - find those players whose performance is higher than expected.

- Further developments:
  - Methodology: to employ a version of the generalized Shapley value that excludes a-priori (it impose a zero worth) some coalitions, in such a way to account for impossible lineups.
  - It may be worth assessing the impact of shooting from specific locations of the court for estimating the probability to win the game.

Thanks

Thank you for listening!

A feedback is welcome

Slides will be soon available on my Website


Auer, B. R., & Hiller, T., On the evaluation of soccer players: a comparison of a new game-theoretical approach to classic performance measures. Applied Economics Letters, 22(14), 1100-1107 (2015)


Barrientos, A. F., Sen, D., Page, G. L., & Dunson, D. B., Bayesian inferences on uncertain ranks and orderings. arXiv preprint arXiv:1907.04842 (2019)


Bensic, M., Sarlija, N., & Zekic-Susac, M., Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intelligent Systems in Accounting, Finance & Management, 13(3), 133-150 (2005)


Cooper, W. W., Ruiz, J. L., & Sirvent, I., Selecting non-zero weights to evaluate effectiveness of basketball players with DEA, European Journal of Operational Research, 195, 563–574 (2009)


Langley, P., Iba, W., & Thompson, K., An analysis of Bayesian classifiers. In AAAI'92: Proceedings of the tenth national conference on Artificial intelligence, 223-228 (1992)


Deshpande, S. K., & Jensen, S. T., Estimating an NBA player's impact on his team's chances of winning. Journal of Quantitative Analysis in Sports, 12(2), 51-72 (2016)


Fearnhead, P., & Taylor, B. M., On estimating the ability of NBA players. Journal of Quantitative Analysis in Sports, 7(3), article number 11 (2011)


Hernández-Lamoneda, L., & Sánchez-Sánchez, F., Rankings and values for team games. International Journal of Game Theory, 39(3), 319-350 (2010)


Hiller, T., The importance of players in teams of the German Bundesliga in the season 2012/2013–a cooperative game theory approach. Applied Economics Letters, 22(4), 324-329 (2015)


Hiller, T., On the stability of couples. Games, 9(3), 48 (2018)


Hofler, R. A., & Payne, J. E., Efficiency in the National Basketball Association: a stochastic frontier approach with panel data. Managerial and Decision Economics, 27(4), 279-285 (2006)


Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X., Applied logistic regression. John Wiley & Sons (2013)


Kalman, S., & Bosch, J., NBA lineup analysis on clustered player tendencies: A new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates. 42 Analytics (2020).


Krzanowski, W. J., & Hand, D. J., ROC curves for continuous data. CRC Press (2009)


Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T., A starting point for analyzing basketball statistics. Journal of Quantitative Analysis in Sports, 3(3), article number 1 (2007)


Langley, P., Iba, W., & Thompson, K., An analysis of Bayesian classifiers. In AAAI'92: Proceedings of the tenth national conference on Artificial intelligence, 223-228 (1992)


Maschler, M., Solan, E., & Zamir, S., Game Theory. Cambridge University Press (2013)


Metulini, R., & Le Carre, M., Measuring sport performances under pressure by classification trees with application to basketball shooting. Journal of Applied Statistics, 47(12), 2120-2135 (2020)


Moreno, P., & Lozano, S., A network DEA assessment of team efficiency in the NBA. Annals of Operations Research, 214(1), 99-124 (2014)


Nikolaidis, Y., Building a basketball game strategy through statistical analysis of data. Annals of Operations Research, 227(1), 137-159 (2015)


Nowak, A., & Radzik, T., The Shapley Value for n-person games in generalized characteristic function form. Games and Economic Behavior 6(1), 150–161 (1994)


Oliver, D., Basketball on paper: rules and tools for performance analysis. Potomac Books, Inc. (2004)


Page, G. L., Barney, B. J., & McGuire, A. T., Effect of position, usage rate, and per game minutes played on NBA player production curves. Journal of Quantitative Analysis in Sports, 9(4), 337-345 (2013)


Piette, J., Anand, S., & Zhang, K., Scoring and shooting abilities of NBA players. Journal of Quantitative Analysis in Sports, 6(1), article number 1 (2013)


Sandri, M., Zuccolotto, P., & Manisera, M., Markov switching modelling of shooting performance variability and teammate interactions in basketball. Journal of the Royal Statistical Society: Series C (Applied Statistics), 69(5), 1337-1356 (2020)


Sarlis, V., & Tjortjis, C., Sports analytics–Evaluation of basketball players and team performance. Information Systems, 93, article number 101562 (2020)


Shapley, L. S., A value for n-person games, in Kuhn, H. W. and Tucker, A. W. (Eds.), Contributions to the theory of games, vol. 2, 307–17. Princeton University Press, Princeton, NJ (1953)


Terner, Z., & Franks, A., Modeling player and team performance in basketball. Annual Review of Statistics and Its Application, 8 (2021)


Wooldridge, J. M., Econometric analysis of cross section and panel data. MIT press (2010)


Yan, T., Kroer, C., & Peysakhovich, A., Evaluating and rewarding teamwork using cooperative game abstractions. Advances in Neural Information Processing Systems, 33 (2020)


Yang, C. H., Lin, H. Y., & Chen, C. P., Measuring the efficiency of NBA teams: additive efficiency decomposition in two-stage DEA. Annals of Operations Research, 217(1), 565-589 (2014)


Zhang, G. P., Neural networks for classification: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30(4), 451-462 (2000)

Regression results - I

	Dependent variable:	
	outcome (win: 1, defeat: 0)	
	(1)	(2)
eFG%_Off	4.085 (0.034) ***	3.938 (0.045) ***
eFG%_Def	-4.069 (0.034) ***	-3.935 (0.045) ***
TOV%_Off	-3.837 (0.064) ***	-3.576 (0.067) ***
TOV%_Def	3.841 (0.064) ***	3.610 (0.066) ***
ORB%_Off	1.190 (0.028) ***	1.135 (0.028) ***
ORB%_Def	-1.186 (0.028) ***	-1.136 (0.028) ***
FT%_Off	1.266 (0.026) ***	1.110 (0.038) ***
FT%_Def	-1.272 (0.026) ***	-1.074 (0.038) ***
AST_Off		0.004 (0.001) ***
AST_Def		-0.003 (0.001)***
BLK_Off		0.002 (0.001) **
BLK_Def		-0.001 (0.001)
FLS_Off		-0.007 (0.001) ***
FLS_Def		0.006 (0.001) ***
Constant	0.494 (0.029) ***	0.478 (0.032) ***
Observations	18,109	18,109
R^2	0.658	0.662
Adjusted R^2	0.658	0.661
Residual Std. Error	0.292 (df = 18100)	0.291 (df = 18094)
F Statistic	4,349.166*** (df = 8; 18,100)	2,526.909*** (df = 14; 18,094)
Note:		*p<0.1; **p<0.05; ***p<0.01

Go back to slide

Regression results - II

	Dependent variable:
	outcome (win: 1, defeat: 0)
eFG%_Off	101.788 (2.304) ***
eFG%_Def	-101.296 (2.300) ***
TOV%_Off	-95.067 (2.407) ***
TOV%_Def	93.993 (2.380) ***
ORB%_Off	30.530 (0.856) ***
ORB%_Def	-31.104 (0.875) ***
FT%_Off	24.125 (0.705) ***
FT%_Def	-23.994 (0.694) ***
Constant	0.031 (0.547)
Observations	18,109
Log Likelihood	-1,876.105
Akaike Inf. Crit.	3,770.211
McFadden pseudo R^2	0.850
Note:	*p<0.1; **p<0.05; ***p<0.01

Go back to slide

Regression results - III

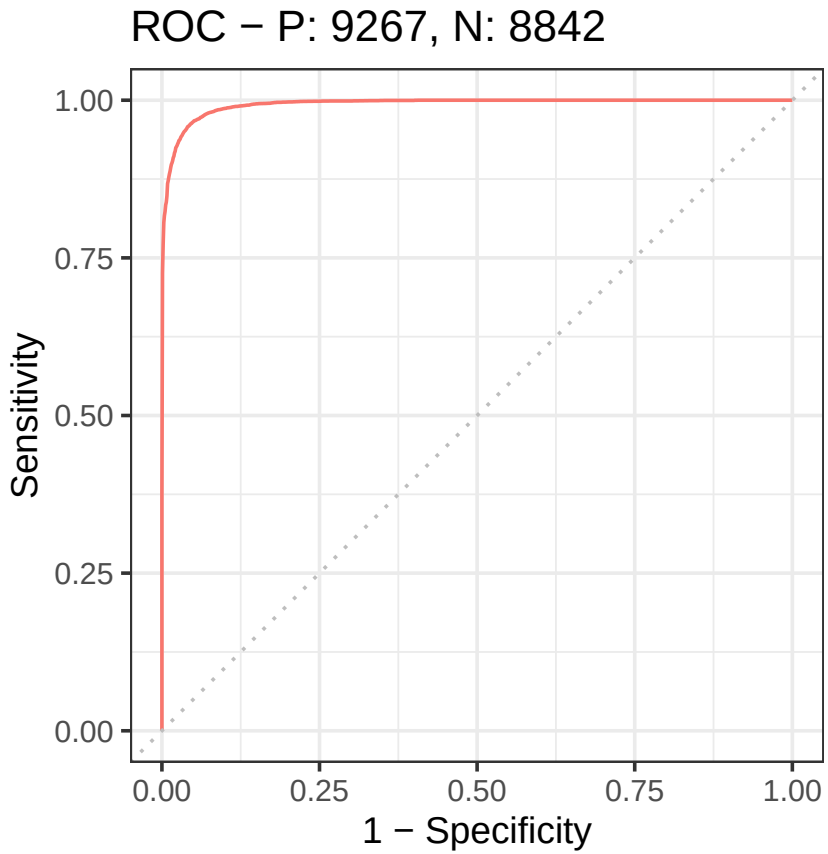

Figure: Receiving Operation Characteristic curve computed from the full sample of 18,109 games: 9,267 positives (true outcome = 1) and 8,842 negatives (true outcome = 0).

Go back to slide

Shapley values - I

Player (i)	n	%	unwgt_gen_Sh (rank)	wgt_gen_Sh (rank)
James Harden	68	75.2	0.127 (2)	0.137 (5)
Eric Gordon	67	56.4	0.121 (6)	0.094 (7)
PJ Tucker	65	52.1	0.123 (5)	0.096 (6)
Trevor Ariza	59	74.2	0.123 (4)	0.151 (3)
Clint Capela	57	73.9	0.132 (1)	0.169 (2)
Luc Mbah	41	32.6	0.127 (3)	0.084 (8)
Ryan Anderson	39	45.9	0.098 (13)	0.140 (4)
Chris Paul	35	48.2	0.115 (7)	0.179 (1)
Gerald Green	21	11.5	0.103 (9)	0.047 (12)
Tarik Black	18	9.9	0.100 (11)	0.048 (11)
Nene	16	10.7	0.113 (8)	0.054 (10)
Joe Johnson	5	3.9	0.101 (10)	0.079 (9)
Briante Weber	4	1.5	0.100 (12)	0.037 (13)

Table: *n*: number of different lineups where that player was in. %: the percentage of time that player was on the court, with respect to the time played by all the 99 considered lineups. *wgt_gen_Sh* values are multiplied by 100.

Go back to slide

Shapley values and Income - I

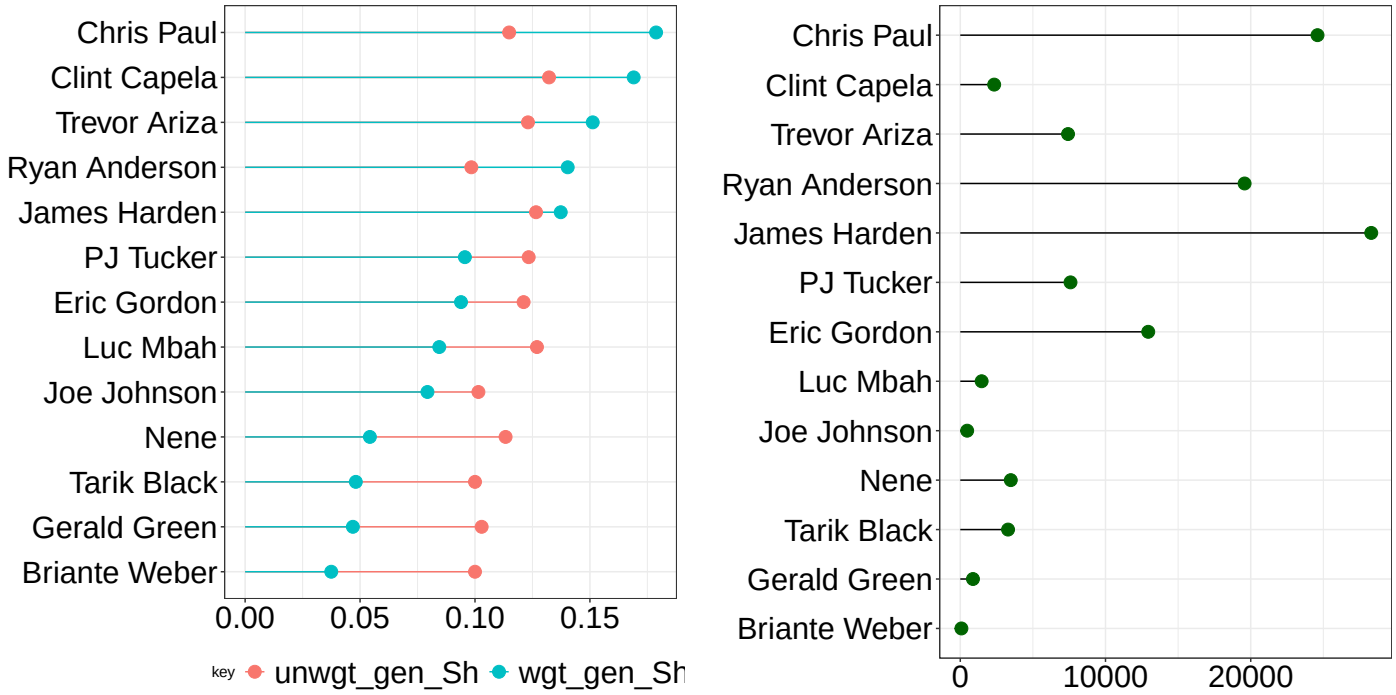

Figure: *wgt_gen_Sh* (cyan, in descending order), *unwgt_gen_Sh* (salmon) (left chart), and income (in thousands of dollars, middle chart) for the Houston Rockets' players.

Go back to slide

Shapley values and Income - II

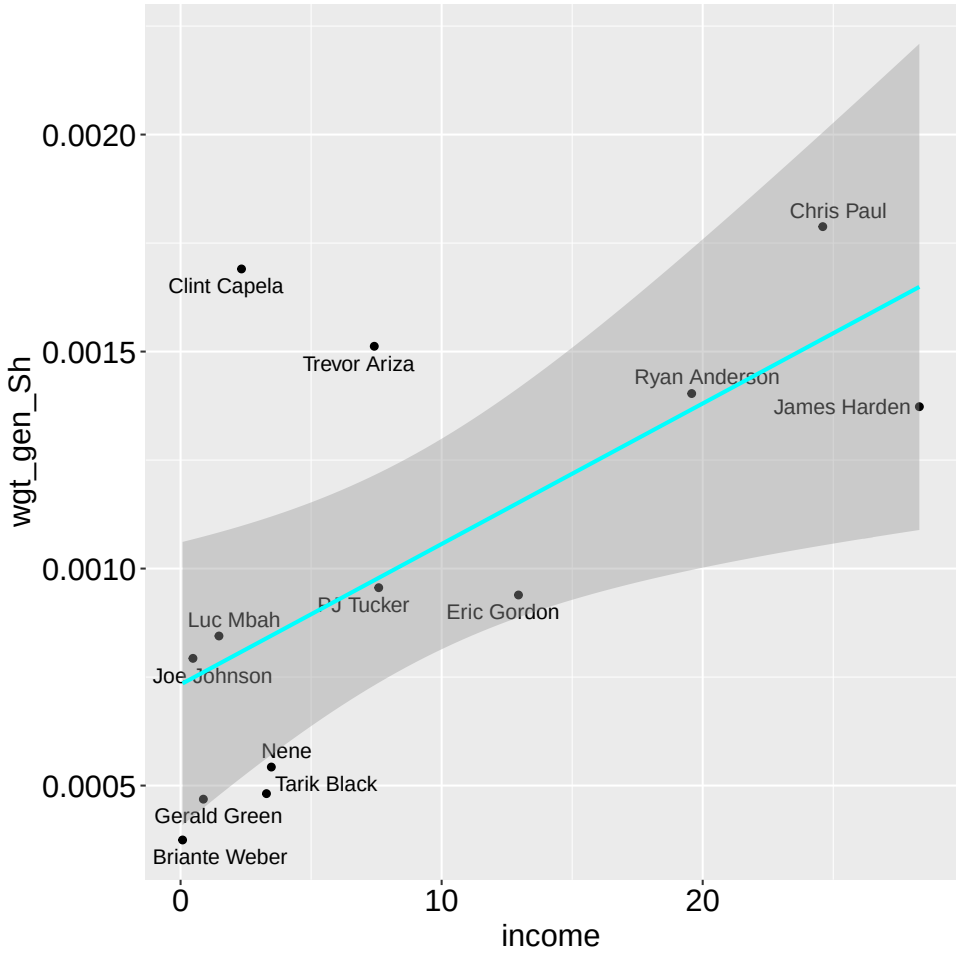

Figure: Scatterplot for the regression equation $\hat{y} = 0.073 + 0.003\ x_1$ (effect of income on *wgt_gen_Sh*). Confidence band (95%) reported in grey.

Go back to slide