

nflWAR: A Reproducible Method for Offensive Player Evaluation in Football

Ron Yurko Sam Ventura Max Horowitz

Department of Statistics
Carnegie Mellon University

Moneyball 2.0: Winning in Sports with Data, Spring 2018

Recent work in football analytics is not easily reproducible:

- Reliance on proprietary and costly data sources
- Data quality relies on potentially biased human judgement

Reproducible Research with nflscrapR

Recent work in football analytics is not easily reproducible:

- Reliance on proprietary and costly data sources
- Data quality relies on potentially biased human judgement

nflscrapR:

- R package created by Maksim Horowitz to enable easy data access and promote reproducible NFL research
- Collects play-by-play data from NFL.com and formats into R data frames
- Data is available for all games starting in 2009

Available on Github, install with:

```
devtools::install_github(repo=maksimhorowitz/nflscrapR)
```

Pittsburgh Fans React

Pittsburgh Post-Gazette article by Liz Bloom covered recent `nflscrapR` research and status of statistics in football

Pittsburgh Fans React

Pittsburgh Post-Gazette article by Liz Bloom covered recent nflscrapR research and status of statistics in football

And the comments...

It's stat geeks like these that are ruining sports. They aren't athletic at all and need to find a way to make themselves relevant. Anyone can make up a stat and algorithms to fit their agenda. Both burkhead and gilislee were injured as wel (let's not forget the #1 most important stat here...IF a player played a whole season then their stats would be the best). Again, these stat geeks do not contemplate injuries and other, ya know, real life stuff.

Reply

▼ [3 replies](#)

-1  

Another Comment...

stats dont work as well in football as compared to other sports such as baseball. you cant statistically evaluate a running back without evaluating his offensive line. same thing with a QB. you cant evaluate a QB without evaluating his receivers (drop balls, wrong route etc). stats can only be helpful when an athlete is doing something completely on his own (pitching). that is why the nfl doesnt go crazy over stats--it's a team sport on every single play. the only stat that counts is the W.

Reply

0



Tremendous Insight!

Recognizes the key flaws of raw football statistics:

- Moving parts in every play
- Need to assign credit to each player involved in a play
- Ultimately evaluate players in terms of wins

stats dont work as well in football as compared to other sports such as baseball. you cant statistically evaluate a running back without evaluating his offensive line. same thing with a QB. you cant evaluate a QB without evaluating his receivers (drop balls, wrong route etc). stats can only be helpful when an athlete is doing something completely on his own (pitching). that is why the nfl doesnt go crazy over stats--it's a team sport on every single play. the only stat that counts is the W.

Reply

0



Using `nflscrapR` we introduce **nflWAR** for offensive players:

- Reproducible framework for **wins above replacement**

Goals of nflWAR

- Properly evaluate every play
- Assign individual player contribution on each play
- Evaluate relative to replacement level
- Convert to a wins scale
- Estimate the uncertainty in WAR

Apply this framework to each available season, 2009-2017

How to Value Plays?

Expected Points (EP): Value of play is in terms of $E(\text{points of next scoring play})$

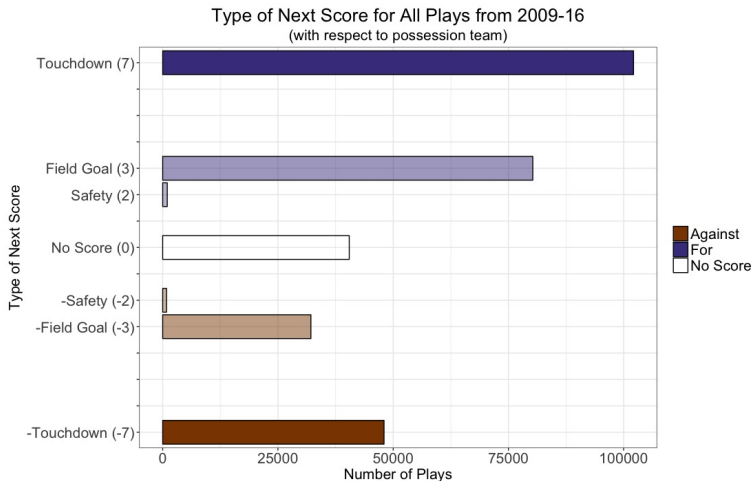
- How many points have teams scored when in similar situations?
- Several ways to model this

Win Probability (WP): Value of play is in terms of $P(\text{Win})$

- Have teams in similar situations won the game?
- Common approach is logistic regression

Can apply **nflWAR** framework to both

Distribution of Next Score



Linear Regression Approach...

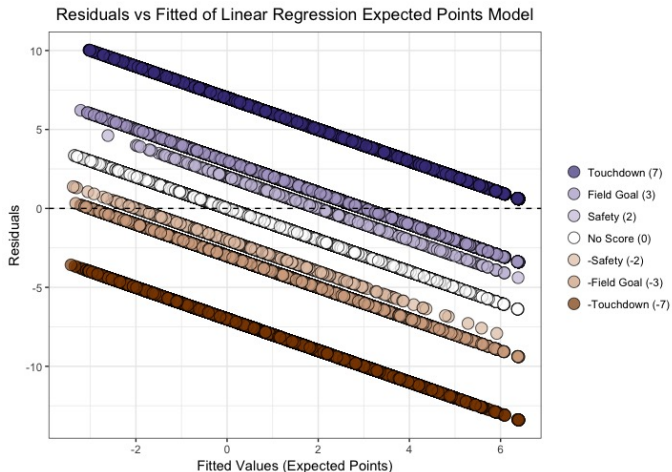
What are the assumptions of linear regression?

$$\epsilon_i \sim N(0, \sigma^2) \text{ (iid)}$$

Linear Regression Approach... IS A DISASTER!

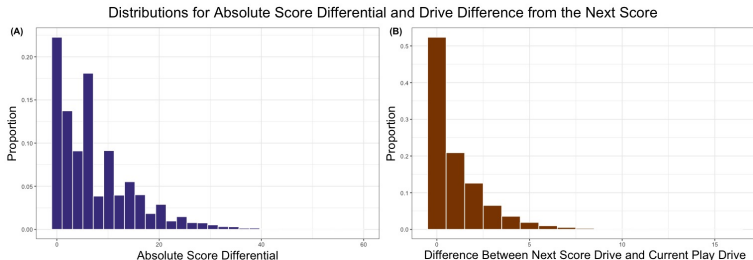
What are the assumptions of linear regression?

$$\epsilon_i \sim N(0, \sigma^2) \text{ (iid)}$$



Weighting Plays

Plays are weighted based on score differential and difference in drives from the next score



Multinomial Logistic Regression

Model is generating probabilities, agnostic of value associated with each next score type

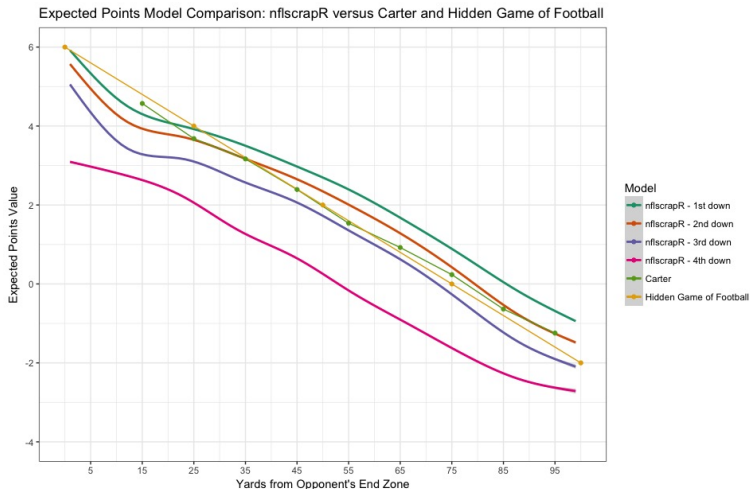
Next Score: $Y \in \{\text{Touchdown (7), Field Goal (3), Safety (2), No Score (0), -Safety (-2), -Field Goal (-3), -Touchdown (-7)}\}$

Situation: $X = \{\text{down, yards to go, yard line, ...}\}$

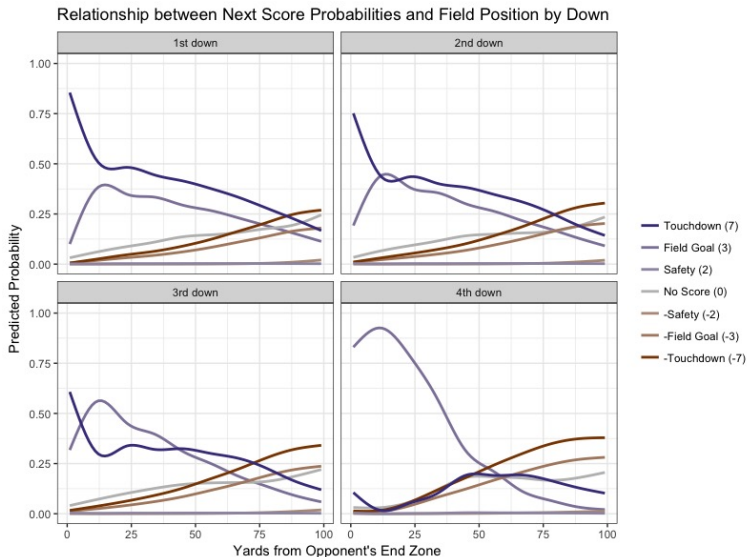
Outcome probabilities: $P(Y = y|X)$

Expected Points (EP) $= E(Y|X) = \sum_y P(Y = y|X) * y$

Expected Points Relationships



Expected Points Relationships



Win Probability Models

"All win probability models are wrong - some are useful"

-Michael Lopez

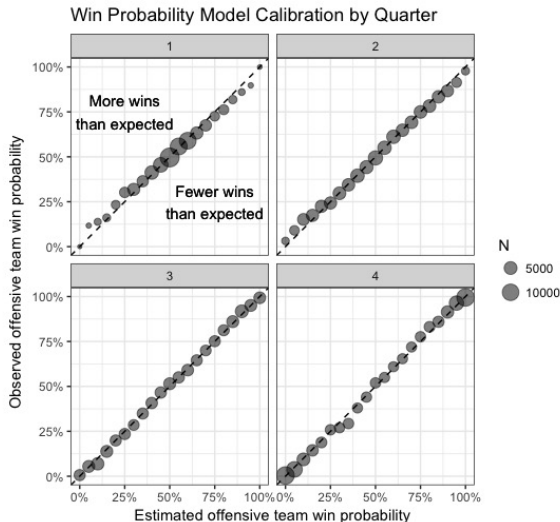
Hot topic in sports analytics community:

- StatsbyLopez blog post
- The Ringer: The Real Super Bowl Loser? Math

Lock and Nettleton (2014) use random forest model

nflscrapR model: Generalized Additive Model (GAM) using **expected score differential**, time remaining, and timeouts

Win Probability Calibration



Estimating the Value of a Play

Assign each play an appropriate value,

$$\delta_{f,i} = \mathbf{V}_f - \mathbf{V}_i$$

Two types of play valuations $\delta_{f,i}$:

- **Expected Points Added (EPA)**
- **Win Probability Added (WPA)**

Estimating the Value of a Play

Assign each play an appropriate value,

$$\delta_{f,i} = \mathbf{V}_f - \mathbf{V}_i$$

Two types of play valuations $\delta_{f,i}$:

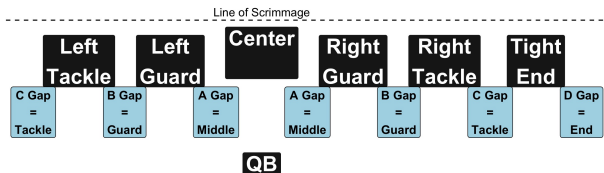
- **Expected Points Added (EPA)**
- **Win Probability Added (WPA)**

For passing plays can use **air yards** to separate the value added through the air and after the catch:

- **airEPA** and **yacEPA** (yards after catch EPA)
- **airWPA** and **yacWPA**

Publicly available data only includes those directly involved:

- Passing:
 - Players: passer, targeted receiver, tackler(s), and interceptor
 - Context: air yards, yards after catch, location (left, middle, right), and if the passer was hit on the play
- Rushing:
 - Players: rusher and tackler(s)
 - Context: run gap (end, tackle, guard, middle) and direction (left, middle, right)



Multilevel Modeling

Growing in popularity (and rightfully so):

- “Multilevel Regression as Default” - Richard McElreath
- Natural approach for data with **group structure**, and different levels of variation within each group
e.g. QBs have more pass attempts than receivers have targets
- Every play is a repeated measure of performance
- Baseball example: Deserved Run Average
(*Judge et al., 2015*)

Multilevel Modeling

Simple example of **varying-intercept** model:

$$\delta_{f,i} \sim \text{Normal}(Q_{q[i]} + C_{c[i]} + X_i \cdot \beta, \sigma_\delta^2), \text{ for } i = 1, \dots, n \text{ plays}$$

Key feature is the **groups are given a model** - treating the levels of groups as similar to one another with **partial pooling**

$$Q_q \sim \text{Normal}(\mu_Q, \sigma_Q^2), \text{ for } q = 1, \dots, \# \text{ of QBs,}$$
$$C_c \sim \text{Normal}(\mu_C, \sigma_C^2), \text{ for } c = 1, \dots, \# \text{ of receivers.}$$

Unlike linear regression, no longer assuming independence

Provides estimates for **average play effects** while providing necessary **shrinkage** towards the group averages

Use varying-intercepts for each of the grouped variables

With location and gap, create **Team-side-gap** as O-line proxy
e.g. PIT-left-end, PIT-left-tackle, PIT-left-guard, PIT-middle

Separate passing and rushing with different grouped variables

- Passing: Quarterback, receiver, defensive team
- Rushing: Team-side-gap, rusher, defensive team

Use varying-intercepts for each of the grouped variables

With location and gap, create **Team-side-gap** as O-line proxy
e.g. PIT-left-end, PIT-left-tackle, PIT-left-guard, PIT-middle

Separate passing and rushing with different grouped variables

- Passing: Quarterback, receiver, defensive team
- Rushing: Team-side-gap, rusher, defensive team

Each group intercept is an estimate for an individual or team effect,

- **individual points/probability added (iPA)**
- **team points/probability added (tPA)**

Multiply intercepts by attempts to get **points/probability above average (iPAA/tPAA)**

Rushing Breakdown

With EPA as the response, two separate models:

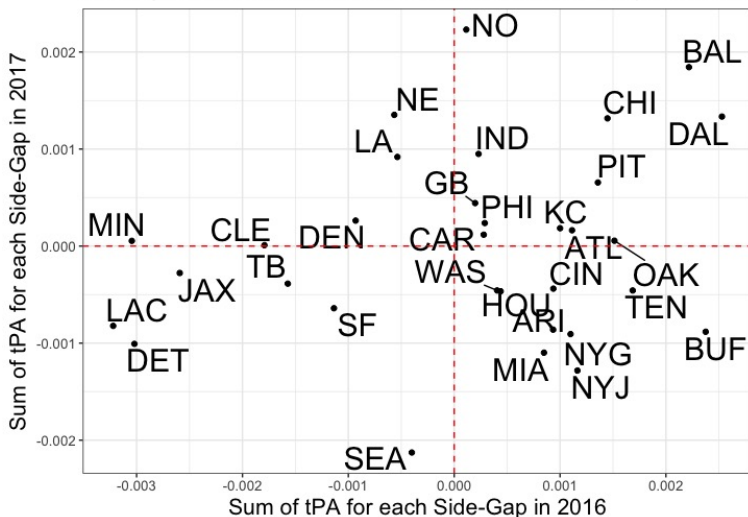
- RB/FB/WR/TE - designed rushing plays
 - Adjust for rusher position as non-grouped variable
- QB - designed runs, scrambles, and sacks
 - No longer use team-side-gap

Both models adjust for team passing strength using EPA per attempt

Provides iPA_{rush} and $tPA_{rush\ side-gap}$ estimates

Comparing Team Offensive Line Performance 2016-17

Summary of Each Team's Offensive Line Performance in 2017 Compared to 2016



Passing Breakdown

Could simply use raw EPA or WPA, or **take advantage of air yards**

Two separate models for air and yac value, where both models consider all pass attempts but the response depends on the model:

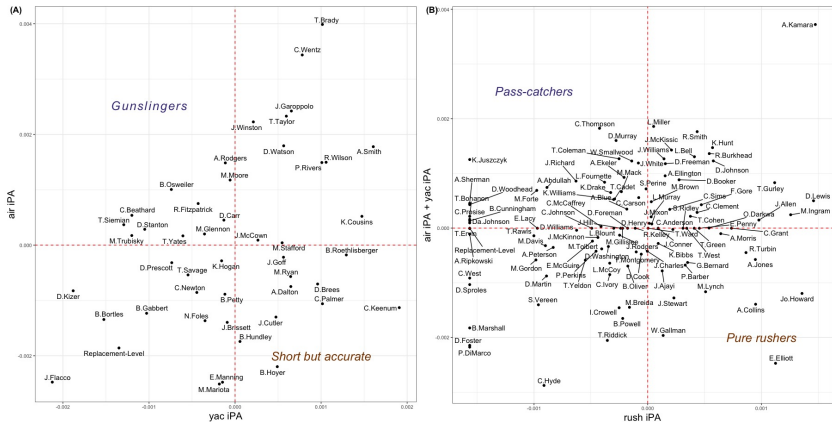
- Receptions assigned air and yac for respective models
- Incomplete passes use observed value
- **Emphasize importance of completions**
- Both adjust for QBs hit, receiver positions, and pass location
- yac model adjusts for air yards

Models adjust for team rushing strength using EPA per attempt

Provides iPA_{air} and iPA_{yac} estimates

QB and RB Efficiency in 2017

QB and RB Efficiency for Skill Separation



Relative to Replacement Level

Following an approach similar to **openWAR** (*Baumer et al., 2015*), defining replacement level based on roster

For each position sort by number of attempts, separate replacement level for rushing and receiving

Relative to Replacement Level

Following an approach similar to **openWAR** (*Baumer et al., 2015*), defining replacement level based on roster

For each position sort by number of attempts, separate replacement level for rushing and receiving

Player i 's $iPAA_{i,total} = iPAA_{i,rush} + iPAA_{i,air} + iPAA_{i,yac}$

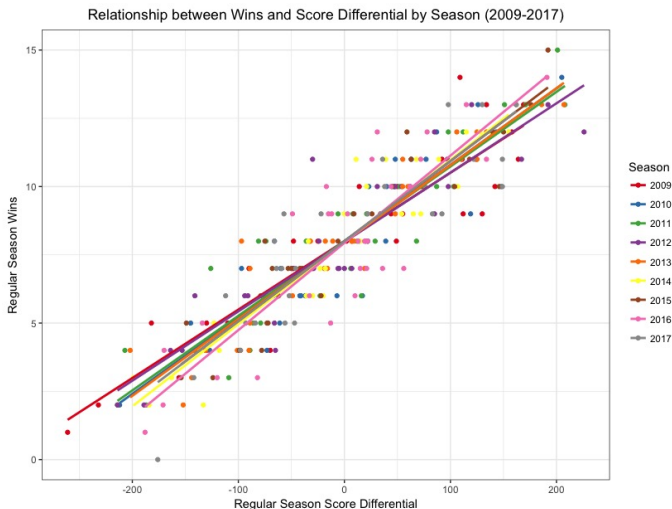
Creates a replacement-level iPAA that “shadows” a player's performance, denote as $iPAA_i^{replacement}$

Player i 's **individual points above replacement (iPAR)** as:

$$iPAR_i = iPAA_{i,total} - iPAA_{i,total}^{replacement}$$

Convert to Wins

“Wins & Point Differential in the NFL” - (*Zhou & Ventura, 2017*)
(CMU Statistics & Data Science **freshman** research project)



Fit a linear regression between wins and total score differential:

$$\text{Points per Win} = \frac{1}{\hat{\beta}_{\text{Score Diff}}}$$

e.g. In 2016 $\hat{\beta}_{\text{Score Diff}} = 0.0319$, roughly **31 points per win**

Fit a linear regression between wins and total score differential:

$$\text{Points per Win} = \frac{1}{\hat{\beta}_{\text{Score Diff}}}$$

e.g. In 2016 $\hat{\beta}_{\text{Score Diff}} = 0.0319$, roughly **31 points per win**

and finally arrive at **wins above replacement (WAR)**:

$$\text{EPA-based WAR} = \frac{iPAR}{\text{Points per Win}}$$

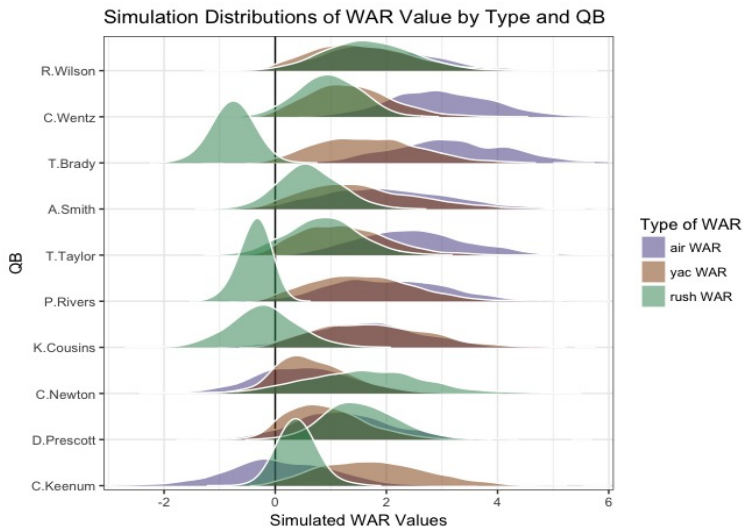
(WPA-based $\text{WAR} = iPAR$)

Similar to **openWAR** (again!) we use a resampling strategy to generate *WAR* distributions

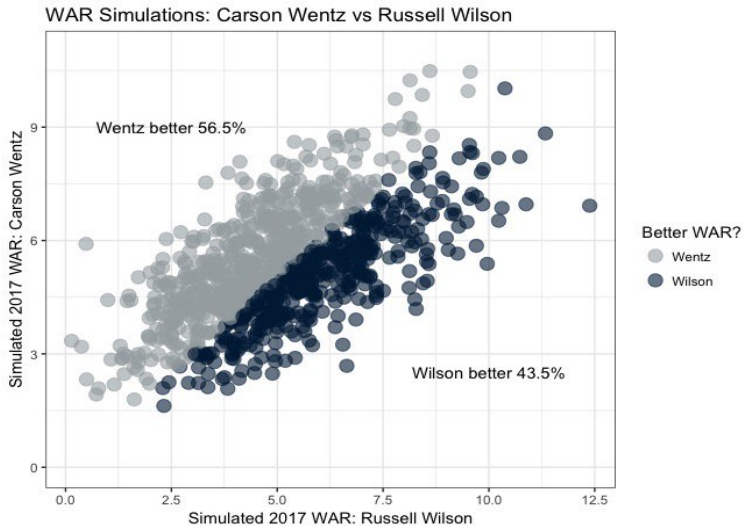
We resample **entire team drives** - *why does this make sense?*

Following estimates are based on 1000 simulations

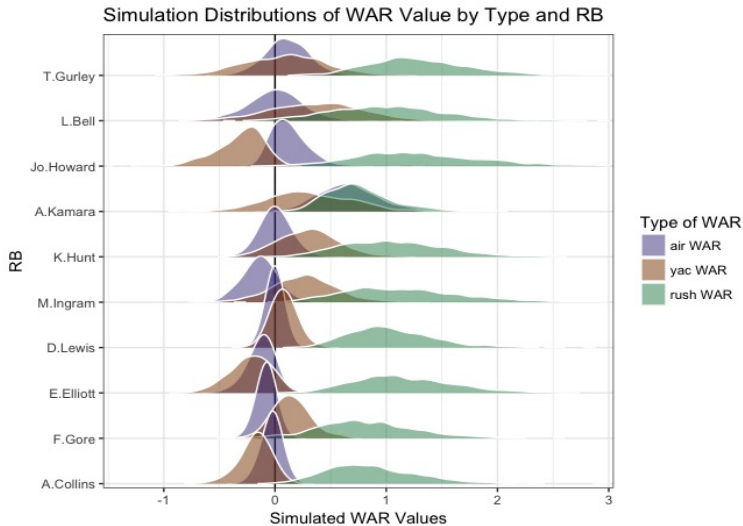
QB WAR in 2017



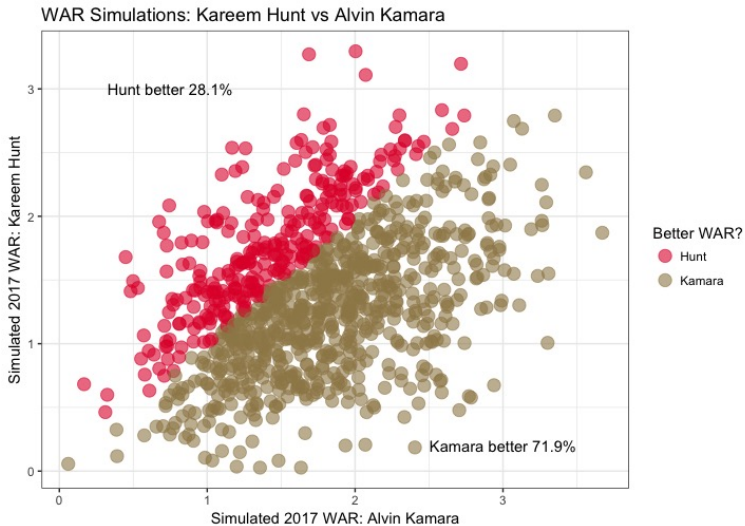
Wilson vs Wentz 2017



RB WAR in 2017



Kamara vs Hunt 2017



Recap and Future of nflWAR

Properly evaluating every play with with multinomial logistic regression model for EP and GAM for WP

Multilevel modeling provides an intuitive way for estimating player effects and **can be extended with data containing every player on the field for every play**

Estimate the uncertainty in the different types of iPA to generate intervals of WAR values

Recap and Future of nflWAR

Properly evaluating every play with with multinomial logistic regression model for EP and GAM for WP

Multilevel modeling provides an intuitive way for estimating player effects and **can be extended with data containing every player on the field for every play**

Estimate the uncertainty in the different types of iPA to generate intervals of WAR values

Naive to assume player has same effect for every play!

Refine the definition of replacement-level,
e.g. what about down specific players?

Acknowledgements





Max Horowitz for creating `nflscrapR`

Sam Ventura for advising every step in the process





Jonathan Judge for answering questions on multilevel modeling

Rebecca Nugent and **CMU Statistics and Data Science** for all of their instruction, motivation, and support!

References I

-  numberfire.
-  Alamar, B. (2010).
Measuring risk in nfl playcalling.
Journal of Quantitative Analysis in Sports, 6(2).
-  Alamar, B. and Goldner, K. (2011).
The blindside project: Measuring the impact of individual offensive linemen.
CHANCE, 24(4):25–29.
-  Alamar, B. and Weinstein-Gould, J. (2008).
Isolating the effect of individual linemen on the passing game in the national football league.
Journal of Quantitative Analysis in Sports, 4(2).





References II

-  Albert, J. (2006).
Pitching statistics, talent and luck, and the best strikeout seasons of all-time.
Journal of Quantitative Analysis in Sports, 2(1).
-  Albert, J. (2016).
Improved component predictions of batting measures.
Journal of Quantitative Analysis in Sports, 12(2).
-  Balreira, E. C., Miceli, B. K., and Tegtmeyer, T. (2014).
An oracle method to predict nfl games.
Journal of Quantitative Analysis in Sports, 10(2).
-  Baumer, B. (2017).
Color Palettes for Pro Sports Teams.
R package version 0.0.1.






References III

-  Baumer, B. and Badian-Pessot, P. (2017).
Evaluation of batters and base runners.
In Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H.,
editors, *Handbook of Statistical Methods and Analyses in Sports*,
pages 1–37. CRC Press, Boca Raton, Florida.
-  Baumer, B., Jensen, S., and Matthews, G. (2015).
openwar: An open source system for evaluating overall player
performance in major league baseball.
Journal of Quantitative Analysis in Sports, 11(2).
-  Baumer, B. and Zimbalist, A. (2014).
*The Sabermetric Revolution: Assessing the Growth of Analytics
in Baseball*.
University of Pennsylvania Press, Philadelphia, Pennsylvania.





References IV

-  Becker, A. and Sun, X. A. (2016).
An analytical approach for fantasy football draft and lineup management.
Journal of Quantitative Analysis in Sports, 12(1).
-  Benjamin, A., Jeff, M., M, D. G., and Lucas, R. (2006).
Who controls the plate? isolating the pitcher/batter subgame.
Journal of Quantitative Analysis in Sports, 2(3).
-  Berri, D. and Burke, B. (2017).
Measuring productivity of nfl players.
In Quinn, K., editor, *The Economics of the National Football League*, pages 137–158. Springer, New York, New York.
-  Brooks, D., Pavlidis, H., and Judge, J. (2015).
Moving beyond wowy: A mixed approach to measuring catcher framing.


References V

-  Burke, B. (2009).
Expected point values.
-  Burke, B. (2014a).
Advanced football analytics.
-  Burke, B. (2014b).
Expected points and expected points added explained.
-  Carroll, B., Palmer, P., Thorn, J., and Pietrusza, D. (1988).
The Hidden Game of Football.
Total Sports, Inc., New York, New York.
-  Carroll, B., Palmer, P., Thorn, J., and Pietrusza, D. (1998).
The Hidden Game of Football: The Next Edition.
Total Sports, Inc., New York, New York.

References VI

-  Carter, V. and Machol, R. (1971).
Operations research on football.
Operations Research, 19(2):541–544.
-  Causey, T. (2013).
Building a win probability model part 1.
-  Causey, T. (2015).
Expected points part 1: Building a model and estimating uncertainty.
-  Citrone, N. and Ventura, S. L. (2017).
A statistical analysis of the nfl draft: Valuing draft picks and predicting future player success.
Presented at the Joint Statistical Meetings.




References VII

-  Clark, T. K., Johnson, A. W., and Stimpson, A. J. (2013).
Going for three: Predicting the likelihood of field goal success
with logistic regression.
MIT Sloan Sports Analytics Conference.
-  Dasarathy, B. V. (1991).
*Nearest neighbor (NN) norms: NN pattern classification
techniques.*
IEEE Computer Society Press, Los Alamitos, CA.
-  Deshpande, S. K. and Jensen, S. T. (2016).
Estimating an nba player's impact on his team's chances of
winning.
Journal of Quantitative Analysis in Sports, 12(2).

References VIII

-  Douglas, B. and Eager, E. A. (2017).
Examining expected points and their interactions with pff grades.
Pro Football Focus Research and Development Journal, 1.
-  Drinen, D. (2013).
Approximate value: Methodolgy.
-  Eager, E. A., Chahrouri, G., and Palazzolo, S. (2017).
Using pff grades to cluster quarterback performance.
Pro Football Focus Research and Development Journal, 1.
-  Elmore, R. and DeWitt, P. (2017).
ballr: Access to Current and Historical Basketball Data.
R package version 0.1.1.
-  ESPN (2017).
Scoring formats.



References IX

-  Franks, A. M., D'Amour, A., Cervone, D., and Bornn, L. (2017).
Meta-analytics: tools for understanding the statistical properties
of sports metrics.
Journal of Quantitative Analysis in Sports, 4(4).
-  Gelman, A. and Hill, J. (2007).
*Data Analysis Using Regression and Multilevel/Hierarchical
Models*.
Cambridge University Press, Cambridge, United Kingdom.
-  Goldner, K. (2012).
A markov model of football: Using stochastic processes to model
a football drive.
Journal of Quantitative Analysis in Sports, 8(1).





References X

-  Goldner, K. (2017).
Situational success: Evaluating decision-making in football.
In Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H.,
editors, *Handbook of Statistical Methods and Analyses in Sports*,
pages 183–198. CRC Press, Boca Raton, Florida.
-  Gramacy, R. B., Taddy, M. A., and Jensen, S. T. (2012).
Estimating player contribution in hockey with regularized logistic
regression.
Journal of Quantitative Analysis in Sports, 9(1).
-  Grimshaw, S. D. and Burwell, S. J. (2014).
Choosing the most popular nfl games in a local tv market.
Journal of Quantitative Analysis in Sports, 10(3).





References XI

-  Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Springer, New York, New York.
-  Horowitz, M., Yurko, R., and Ventura, S. L. (2017).
nflscrapR: Compiling the NFL play-by-play API for easy use in R.
R package version 1.4.0.
-  James, B. (2017).
Judge and altuve.
-  James, B. and Henzler, J. (2002).
Win Shares.
STATS, Inc.





References XII

-  Jensen, J. A. and Turner, B. A. (2014).
What if statisticians ran college football? a re-conceptualization
of the football bowl subdivision.
Journal of Quantitative Analysis in Sports, 10(1).
-  Jensen, S., Shirley, K. E., and Wyner, A. (2009).
Bayesball: A bayesian hierarchical model for evaluating fielding in
major league baseball.
The Annals of Applied Statistics, 3(2).
-  Katz, S. and Burke, B. (2017).
How is total qbr calculated? we explain our quarterback rating.
-  Kelly, D. (2016).
The state of the two-point conversion.






References XIII

-  Knowles, J. E. and Frederick, C. (2016).
merTools: Tools for Analyzing Mixed Effect Regression Models.
R package version 0.3.0.
-  Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. T. (2007).
A starting point for analyzing basketball statistics.
Journal of Quantitative Analysis in Sports, 3(3).
-  Lahman, S. (1996 – 2017).
Lahman's Baseball Database.
-  Lillibridge, M. (2013).
The anatomy of a 53-man roster in the nfl.


References XIV

-  Lock, D. and Nettleton, D. (2014).
Using random forests to estimate win probability before each play of an nfl game.
Journal of Quantitative Analysis in Sports, 10(2).
-  Lopez, M. (2017).
All win probability models are wrong some are useful.
-  Macdonald, B. (2011).
A regression-based adjusted plus-minus statistic for nhl players.
Journal of Quantitative Analysis in Sports, 7(3).
-  Martin, R., Timmons, L., and Powell, M. (2017).
A markov chain analysis of nfl overtime rules.
Journal of Sports Analytics, Pre-print(Pre-print).

References XV

-  Meers, K. (2011).
How to value nfl draft picks.
-  Morris, B. (2015).
Kickers are forever.
-  Mulholland, J. and Jensen, S. T. (2014).
Predicting the draft and career success of tight ends in the
national football league.
Journal of Quantitative Analysis in Sports, 10(4).
-  Oliver, D. (2011).
Guide to the total quarterback rating.
-  Paine, N. (2015a).
Bryce harper should have made \$73 million more.






References XVI

-  Paine, N. (2015b).
How madden ratings are made: The secret process that turns nfl players into digital gods.
-  Pasteur, D. and Cunningham-Rhoads, K. (2014).
An expectation-based metric for nfl field goal kickers.
Journal of Quantitative Analysis in Sports, 10(1).
-  Pasteur, R. D. and David, J. A. (2017).
Evaluation of quarterbacks and kickers.
In Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H., editors, *Handbook of Statistical Methods and Analyses in Sports*, pages 165–182. CRC Press, Boca Raton, Florida.
-  Pattani, A. (2012).
Expected points and epa explained.





References XVII

-  Phillips, C. (2018).
Scoring and Scouting: How We Know What We Know about Baseball.
Forthcoming, Princeton University Press, Princeton, New Jersey.
-  Piette, J. and Jensen, S. (2012).
Estimating fielding ability in baseball players over time.
Journal of Quantitative Analysis in Sports, 8(3).
-  Pro-Football-Reference (2018).
Football glossary and football statistics glossary.
-  Quealy, K., Causey, T., and Burke, B. (2017).
4th down bot: Live analysis of every n.f.l. 4th down.

References XVIII

-  R Core Team (2017).
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria.
-  Ratcliffe, J. (2013).
The pff idp scoring system revisited.
-  Romer, D. (2006).
Do firms maximize? evidence from professional football.
Journal of Political Economy, 114(2).
-  Rosenbaum, D. T. (2004).
Measuring how nba players help their teams win.
-  Schatz, A. (2003).
Methods to our madness.

References XIX

-  Schoenfield, D. (2012).
What we talk about when we talk about war.
-  Sievert, C. (2015).
pitchRx: Tools for Harnessing 'MLBAM' 'Gameday' Data and Visualizing 'pitchfx'.
R package version 1.8.2.
-  Skiver, K. (2017).
Zebra technologies to bring live football tracking on nfl gamedays with rfid chips.
-  Smith, D., Siwoff, S., and Weiss, D. (1973).
Nfl's passer rating.

References XX



Snyder, K. and Lopez, M. (2015).

Consistency, accuracy, and fairness: a study of discretionary penalties in the nfl.

Journal of Quantitative Analysis in Sports, 11(4).



Sportradar (2017).

NFL Data Addendum.

Last updated 2017-08-11.



Sportradar (2018).

NFL Official API.





Version 2.



Tango, T. (2017).

War podcast.

References XXI

-  Tango, T., Lichtman, M., and Dolphin, A. (2007).
The Book: Playing the Percentages in Baseball.
Potomac Book, Inc., Washington, D.C.
-  Thomas, A. and Ventura, S. L. (2017).
nhlscrapr: Compiling the NHL Real Time Scoring System Database for easy use in R.
R package version 1.8.1.
-  Thomas, A. C. and Ventura, S. L. (2015).
The road to war.
-  Thomas, A. C., Ventura, S. L., Jensen, S. T., and Ma, S. (2013).
Competing process hazard function models for player ratings in ice hockey.
The Annals of Applied Statistics, 7(3).

References XXII

-  Turkenkopf, D., Pavlidis, H., and Judge, J. (2015).
Prospectus feature: Introducing deserved run average (dra) and all its friends.
-  Wakefield, K. and Rivers, A. (2012).
The effect of fan passion and official league sponsorship on brand metrics: A longitudinal study of official nfl sponsors and roo.
MIT Sloan Sports Analytics Conference.
-  Yam, D. R. and Lopez, M. J. (2018).
Quantifying the causal effects of conservative fourth down decision making in the national football league.
Under Review.
-  Yurko, R., Ventura, S., and Horowitz, M. (2017).
Nfl player evaluation using expected points added with nflscrapr.
Presented at the Great Lakes Sports Analytics Conference.



Zhou, E. and Ventura, S. (2017).
Wins and point differential in the nfl.