

nflWAR: A Reproducible Method for Offensive Player Evaluation in Football

Ron Yurko Sam Ventura Max Horowitz

Department of Statistics
Carnegie Mellon University

NESSIS, 2017

Recent work in football analytics is not easily reproducible:

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

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 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, 2009-2016

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, but will focus on **EP** today

How to Calculate EP?

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

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

“Nearest Neighbors”:

- Identify similar plays in historical data based on down, yards to go, yard line, etc. and take the average

How to Calculate EP?

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

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

“Nearest Neighbors”:

- Identify similar plays in historical data based on down, yards to go, yard line, etc. and take the average

But what defines a similar play?

How to Calculate EP?

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

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

“Nearest Neighbors”:

- Identify similar plays in historical data based on down, yards to go, yard line, etc. and take the average

But what defines a similar play?

Linear Regression:

$$E(Y|X) = \beta_0 + \beta_1 X_{\text{down}} + \beta_2 X_{\text{yards to go}} + \dots$$

How to Calculate EP?

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

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

“Nearest Neighbors”:

- Identify similar plays in historical data based on down, yards to go, yard line, etc. and take the average

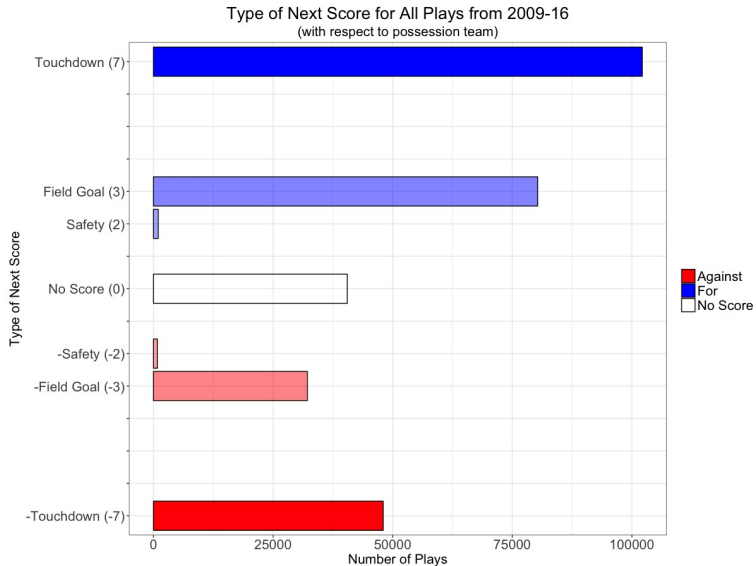
But what defines a similar play?

Linear Regression:

$$E(Y|X) = \beta_0 + \beta_1 X_{\text{down}} + \beta_2 X_{\text{yards to go}} + \dots$$

But is treating the next score as continuous appropriate?

Distribution of Next Score



Linear Regression Approach...

What are the assumptions of linear regression?

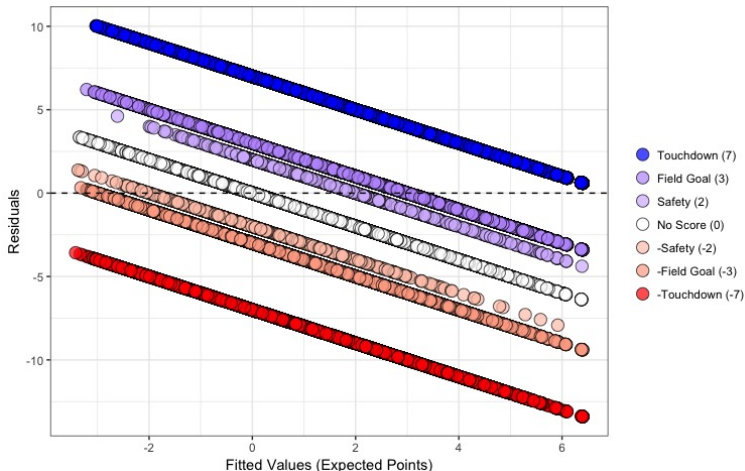
$$\epsilon_i \sim N(0, \sigma^2) \text{ (iid)}$$

Linear Regression Approach... IS A DISASTER!

What are the assumptions of linear regression?

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

Residuals vs Fitted of Linear Regression Expected Points Model



Multinomial Logistic Regression

Logistic regression to **model the probabilities** of
 $Y \in \{\text{Touchdown (7)}, \text{Field Goal (3)}, \text{Safety (2)},$
-Touchdown (-7), -Field Goal (-3), -Safety (-2),
No Score (0)}

Specified with 6 logit transformations relative to No Score:

$$\log\left(\frac{P(Y = \text{Touchdown}|X = x)}{P(Y = \text{No Score}|X = x)}\right) = \beta_{10} + \beta_1^T x$$
$$\log\left(\frac{P(Y = \text{Field Goal}|X = x)}{P(Y = \text{No Score}|X = x)}\right) = \beta_{20} + \beta_2^T x$$
$$\vdots$$
$$\log\left(\frac{P(Y = -\text{Touchdown}|X = x)}{P(Y = \text{No Score}|X = x)}\right) = \beta_{60} + \beta_6^T x$$

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 Added

Expected Points Added (EPA) estimates a play's value based on the change in situation, providing a point value

- $EPA_{play_i} = EP_{play_{i+1}} - EP_{play_i}$

Expected Points Added

Expected Points Added (EPA) estimates a play's value based on the change in situation, providing a point value

- $EPA_{play_i} = EP_{play_{i+1}} - EP_{play_i}$

For passing plays can use **air yards** to calculate **airEPA** and **yacEPA** (yards after catch EPA):

- $airEPA_{play_i} = EP_{in\ air\ play_i} - EP_{start\ play_i}$

- $yacEPA_{play_i} = EP_{play_{i+1}} - EP_{in\ air\ play_i}$

Expected Points Added

Expected Points Added (EPA) estimates a play's value based on the change in situation, providing a point value

- $EPA_{play_i} = EP_{play_{i+1}} - EP_{play_i}$

For passing plays can use **air yards** to calculate **airEPA** and **yacEPA** (yards after catch EPA):

- $airEPA_{play_i} = EP_{in\ air\ play_i} - EP_{start\ play_i}$

- $yacEPA_{play_i} = EP_{play_{i+1}} - EP_{in\ air\ play_i}$

But how much credit does each player deserve?

e.g. On a pass play, how much credit does a QB get vs the receiver?

One player is not solely responsible for a play's EPA

How to Allocate EPA?

Using proprietary, manually collected data, **Total QBR** (*Oliver et al., 2011*) divides credit between those involved in passing plays

How to Allocate EPA?

Using proprietary, manually collected data, **Total QBR** (*Oliver et al., 2011*) divides credit between those involved in passing plays

Publicly available data only includes those directly involved:

- Passing:
 - Individuals: passer, target receiver, tackler(s), interceptor
 - Context: air yards, yards after catch, location, and if the passer was hit on the play
- Rushing:
 - Individuals: rusher and tackler(s)
 - Context: run gap and location

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

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

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

$$EPA_i \sim N(QB_{j[i]} + REC_{k[i]} + \beta x_i, \sigma_{EPA}^2), \text{ for } i = 1, \dots, \# \text{ of plays,}$$

$$QB_j \sim \text{Normal}(\mu_{QB}, \sigma_{QB}^2), \text{ for } j = 1, \dots, \# \text{ of QBs,}$$

$$REC_k \sim \text{Normal}(\mu_{REC}, \sigma_{REC}^2), \text{ for } k = 1, \dots, \# \text{ of Receivers}$$

Multilevel Modeling

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

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

$$\begin{aligned}EPA_i &\sim N(QB_{j[i]} + REC_{k[i]} + \beta x_i, \sigma_{EPA}^2), \text{ for } i = 1, \dots, \# \text{ of plays,} \\QB_j &\sim \text{Normal}(\mu_{QB}, \sigma_{QB}^2), \text{ for } j = 1, \dots, \# \text{ of QBs,} \\REC_k &\sim \text{Normal}(\mu_{REC}, \sigma_{REC}^2), \text{ for } k = 1, \dots, \# \text{ of Receivers}\end{aligned}$$

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-guard, PIT-middle

Separate passing and rushing with different grouped variables

- Passing: Offensive team, QB, 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-guard, PIT-middle

Separate passing and rushing with different grouped variables

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

Each individual intercept for player groups is an estimate for a player's effect, **individual points added (iPA)**

Intercepts for team groups are **team points added (tPA)**

Multiply iPA/tPA by attempts to get
individual/team points above average (iPAA/tPAA)

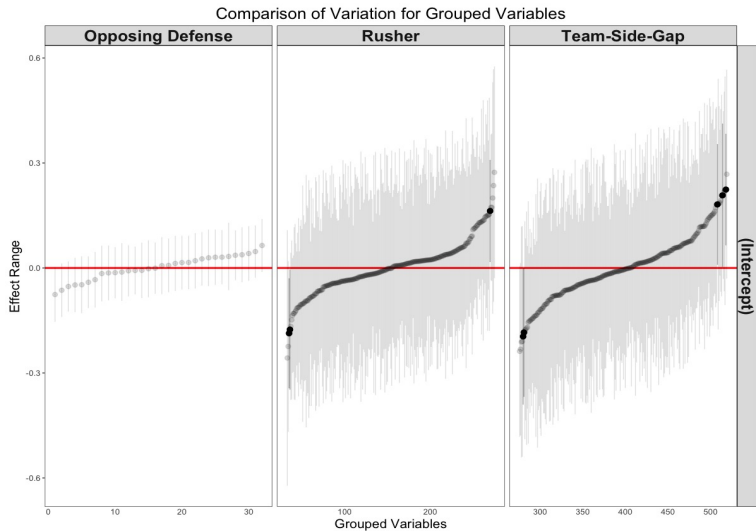
Rushing Breakdown

With EPA as the response, two separate models:

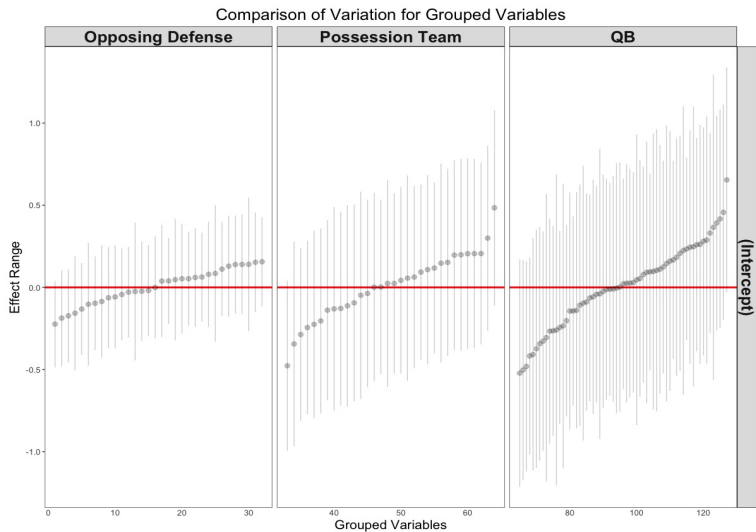
- RB/FB/WR/TE - designed rushing plays
 - Adjust for rusher position as non-grouped variable
- QB - designed runs, scrambles, and sacks
 - Replace Team-side-gap with offensive team

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

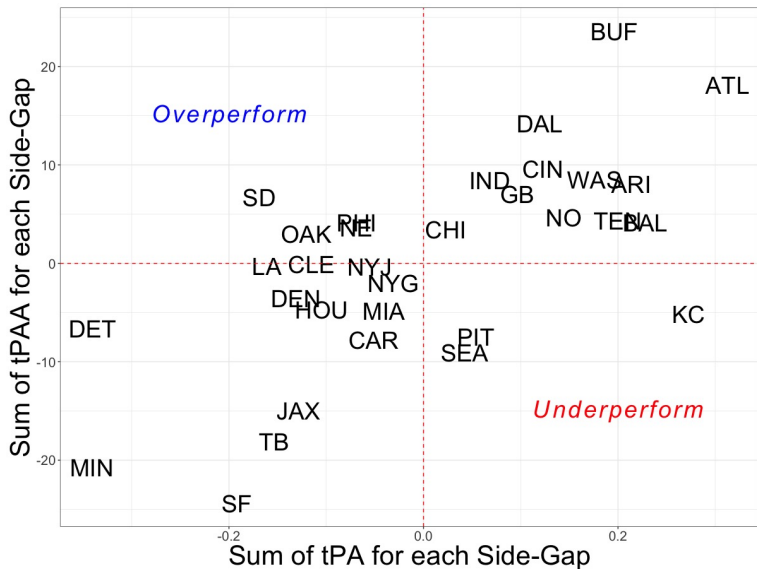
Group Variation for RB/FB/WR/TE Rushing Model



Group Variation for QB Rushing Model



Which Teams Ran Efficiently in 2016?



Passing Breakdown

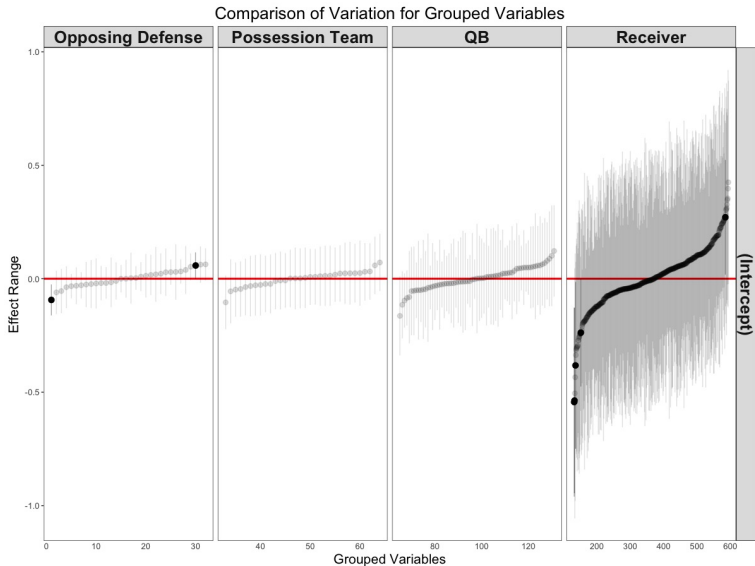
Could simply use EPA, or **take advantage of air yards**

Two separate models for airEPA and yacEPA, where both models consider all pass attempts but the response depends on the model:

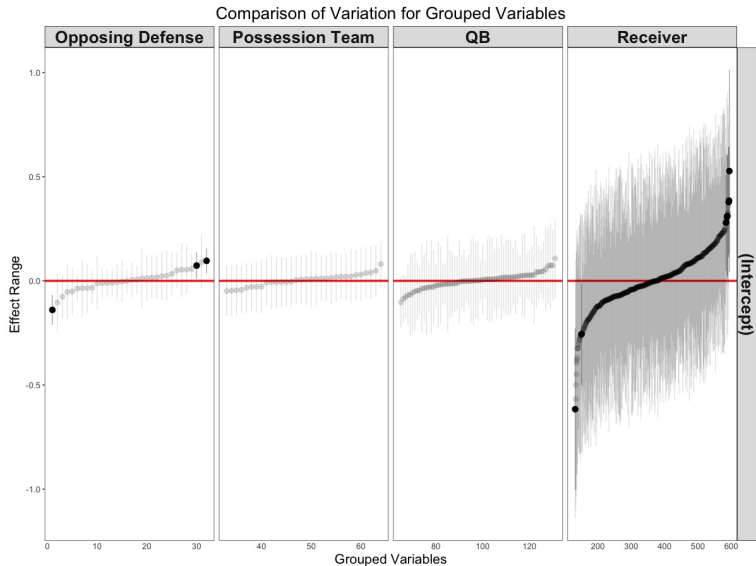
- Receptions assigned airEPA and yacEPA for respective models
- Incomplete passes use observed EPA
- **Emphasize importance of completions**
- Both adjust for QBs hit, receiver positions, and pass location
- yacEPA model adjusts for air yards

Provides iPA_{air} and iPA_{yac} estimates

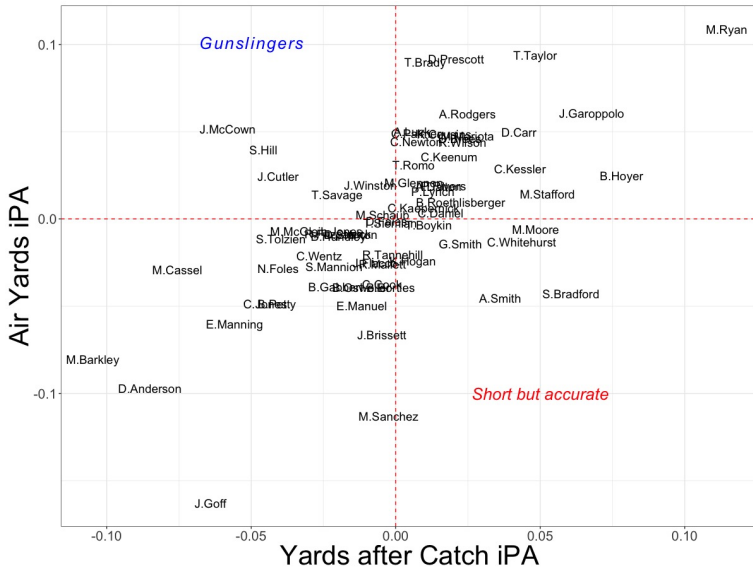
Variation of Passing Intercepts (airEPA)



Variation of Passing Intercepts (yacEPA)



Passing Efficiency in 2016



Relative to Replacement Level

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

For each team and position sort by number of attempts (separate RB/FB 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 team and position sort by number of attempts (separate RB/FB replacement level for rushing and receiving)

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

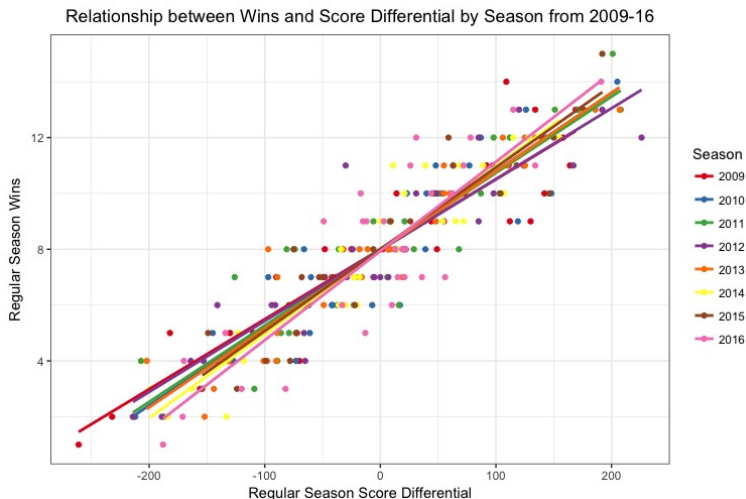
Creates a replacement-level iPAA that “shadows” a player's performance, denote as $iPAA_i^{replacement}$

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:

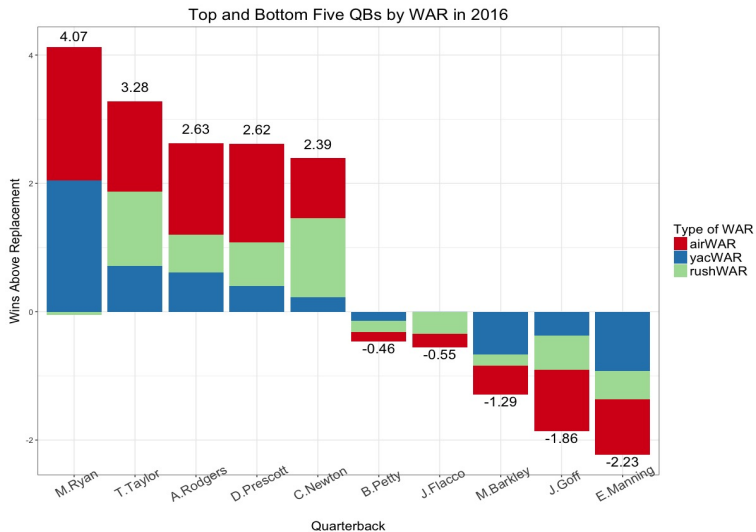
$$\text{Points per Win} = \frac{1}{\hat{\beta}_{\text{Score Diff}}}$$

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

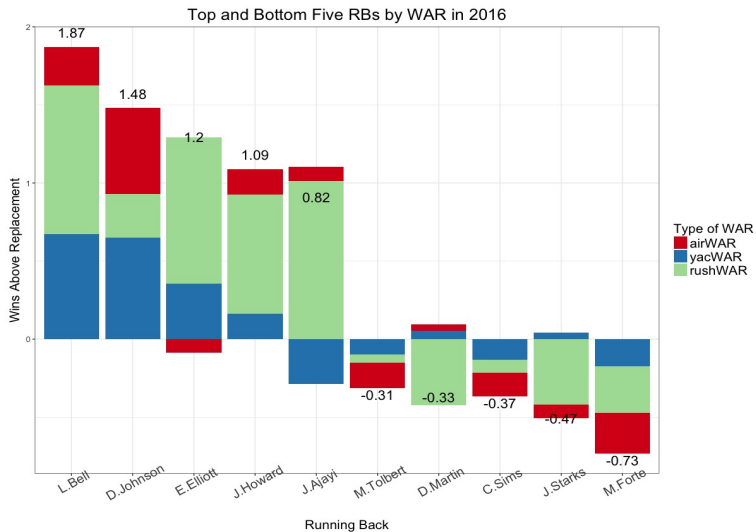
and finally arrive at **wins above replacement (WAR)**:

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

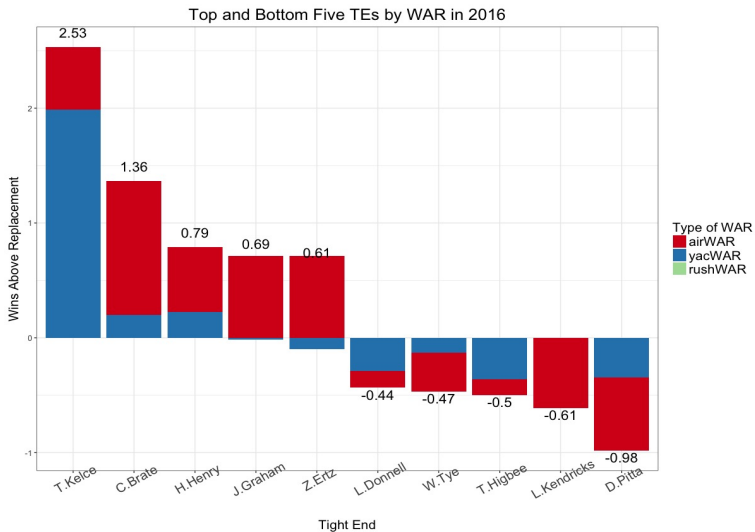
QB WAR in 2016



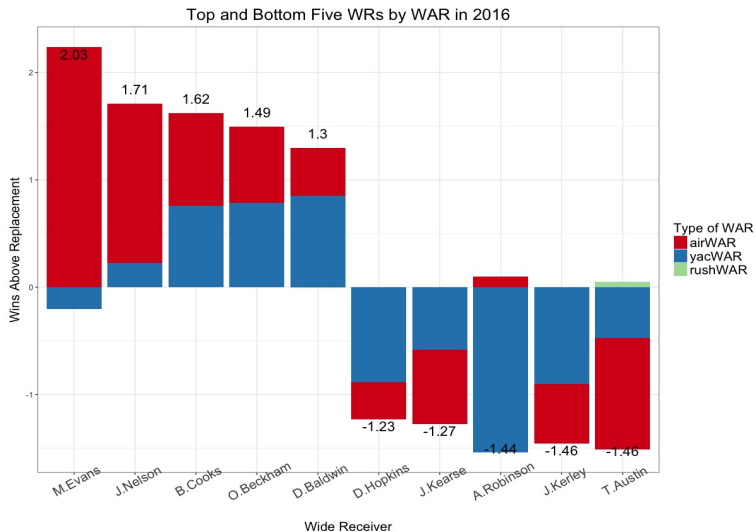
RB WAR in 2016



TE WAR in 2016



WR WAR in 2016



Recap and Future of nflWAR

Properly evaluating every play with EPA generated with multinomial logistic regression model

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**

Recap and Future of nflWAR

Properly evaluating every play with EPA generated with multinomial logistic regression model

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**

Naive to assume player has same effect for every play!

Need to **estimate the uncertainty** in the different types of iPA to generate intervals of WAR values

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

CARNEGIE MELLON SPORTS ANALYTICS

Clear your calendars for Oct 28th!

And visit www.cmusportsanalytics.com/conference

for more information! #CMSAC

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!

[Gelman and Hill, 2007]. [Baumer et al., 2015]. [Oliver,].
[Hastie et al., 2009]. [Carroll et al., 1998].
[Pasteur and David, 2017]. [Goldner, 2017]. [Berri and Burke, 2017].
[Carter and Machol, 1971]. [Burke,]. [out,]. [num,].



Football outsiders.



numberfire.







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).

References II

-  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.
-  Burke, B.
Advanced football analytics.
-  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.
-  Carter, V. and Machol, R. (1971).
Operations research on football.
Operations Research, 19(2):541–544.

References III

-  Gelman, A. and Hill, J. (2007).
Data Analysis Using Regression and Multilevel/Hierarchical Models.
Cambridge University Press, Cambridge, United Kingdom.
-  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.
-  Hastie, T., Tibshirani, R., and Friedman, J. (2009).
*The Elements of Statistical Learning: Data Mining, Inference,
and Prediction*.
Springer, New York, New York.

-  Oliver, D.
Guide to the total quarterback rating.
-  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.