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
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:

```r
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.
stats don't work as well in football as compared to other sports such as baseball. you can't statistically evaluate a running back without evaluating his offensive line. same thing with a QB. you can't 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 doesn't go crazy over stats--it's a team sport on every single play. the only stat that counts is the W.
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

Using `nflscrapR` we introduce `nflWAR` for offensive players:
- Reproducible framework for wins above replacement

Stats don't work as well in football as compared to other sports such as baseball. You can't statistically evaluate a running back without evaluating his offensive line. Same thing with a QB, you can't 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 doesn't go crazy over stats—it's a team sport on every single play. The only stat that counts is the W.
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

Type of Next Score for All Plays from 2009-16
(with respect to possession team)

- Touchdown (7)
- Field Goal (3)
- Safety (2)
- No Score (0)
- Safety (-2)
- Field Goal (-3)
- Touchdown (-7)

Number of Plays
What are the assumptions of linear regression?

\[ \epsilon_i \sim N(0, \sigma^2) \ (iid) \]
What are the assumptions of linear regression?

\[ \epsilon_i \sim \mathcal{N}(0, \sigma^2) \ (iid) \]
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) =** \( E(Y|X) = \sum_y P(Y = y|X) \times y \)
Expected Points Relationships

Expected Points Model Comparison: nflscrapR versus Carter and Hidden Game of Football

Yards from Opponent's End Zone

Expected Points Value

Model
- nflscrapR - 1st down
- nflscrapR - 2nd down
- nflscrapR - 3rd down
- nflscrapR - 4th down
- Carter
- Hidden Game of Football
Expected Points Relationships

Relationship between Next Score Probabilities and Field Position by Down

- 1st down
- 2nd down
- 3rd down
- 4th down

Yards from Opponent's End Zone

Predicted Probability

- Touchdown (7)
- Field Goal (3)
- Safety (2)
- No Score (0)
- Safety (-2)
- Field Goal (-3)
- Touchdown (-7)
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
Assign each play an appropriate value,

\[ \delta_{f,i} = V_f - V_i \]

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,

\[ \delta_{f,i} = V_f - V_i \]

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

\[
\delta_{f,i} \sim \text{Normal}(Q_q[i] + C_c[i] + X_i \cdot \beta, \sigma^2_\delta), \text{ for } i = 1, \ldots, n \text{ plays}
\]

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

\[
Q_q \sim \text{Normal}(\mu_Q, \sigma^2_Q), \text{ for } q = 1, \ldots, \# \text{ of QBs},
\]
\[
C_c \sim \text{Normal}(\mu_C, \sigma^2_C), \text{ for } c = 1, \ldots, \# \text{ of receivers}.
\]

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
nflWAR Modeling

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

Ron Yurko (@Stat_Ron)
nflWAR
Moneyball 2.0, 2018
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

(A)

(B)

Pass-catchers

Pure rushers

Short but accurate
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 $iPAA_{i,\text{total}} = iPAA_{i,\text{rush}} + iPAA_{i,\text{air}} + iPAA_{i,\text{yac}}$

Creates a replacement-level iPAA that “shadows” a player’s performance, denote as $iPAA_{i,\text{replacement}}$

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

$$iPAR_i = iPAA_{i,\text{total}} - iPAA^{\text{replacement}}_{i,\text{total}}$$
“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 \textbf{31 points per win}

and finally arrive at \textbf{wins above replacement (WAR)}:

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

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

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

Simulation Distributions of WAR Value by Type and QB

R. Wilson
C. Wentz
T. Brady
A. Smith
T. Taylor
P. Rivers
K. Cousins
C. Newton
D. Prescott
C. Keenum

Type of WAR:
- air WAR
- yac WAR
- rush WAR

Simulated WAR Values
Wilson vs Wentz 2017

WAR Simulations: Carson Wentz vs Russell Wilson

Wentz better 56.5%

Wilson better 43.5%
RB WAR in 2017

Simulation Distributions of WAR Value by Type and RB

Type of WAR
- air WAR
- yac WAR
- rush WAR

Simulated WAR Values
Kamara vs Hunt 2017

WAR Simulations: Kareem Hunt vs Alvin Kamara

Hunt better 28.1%

Kamara better 71.9%
Recap and Future of nflWAR

Properly evaluating every play 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 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

numberfire.


References IV

An analytical approach for fantasy football draft and lineup management.
*Journal of Quantitative Analysis in Sports, 12*(1).

Who controls the plate? isolating the pitcher/batter subgame.
*Journal of Quantitative Analysis in Sports, 2*(3).

Measuring productivity of nfl players.

Moving beyond wowy: A mixed approach to measuring catcher framing.
References V

Burke, B. (2009).  
Expected point values.

Advanced football analytics.

Burke, B. (2014b).  
Expected points and expected points added explained.

_The Hidden Game of Football_.

_The Hidden Game of Football: The Next Edition_.


*Data Analysis Using Regression and Multilevel/Hierarchical Models.*
Cambridge University Press, Cambridge, United Kingdom.

A markov model of football: Using stochastic processes to model a football drive.
*Journal of Quantitative Analysis in Sports, 8*(1).


James, B. (2017). *Judge and altuve*.


R package version 0.3.0.

A starting point for analyzing basketball statistics.
*Journal of Quantitative Analysis in Sports, 3*(3).

*Lahman’s Baseball Database.*

The anatomy of a 53-man roster in the nfl.

Lopez, M. (2017). All win probability models are wrong some are useful.


Paine, N. (2015a). Bryce Harper should have made $73 million more.


R Core Team (2017).  
*R: A Language and Environment for Statistical Computing.*  
R Foundation for Statistical Computing, Vienna, Austria.

The pff idp scoring system revisited.

Do firms maximize? evidence from professional football.  
*Journal of Political Economy, 114*(2).

Measuring how nba players help their teams win.

Methods to our madness.
Schoenfield, D. (2012). What we talk about when we talk about war.


