nflWAR:
A Reproducible Method for Offensive Player Evaluation in Football

Ron Yurko    Sam Ventura    Max Horowitz

Department of Statistics & Data Science
Carnegie Mellon University

RIT Sports Analytics Conference, Aug 11th 2018
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
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, game, roster data from NFL.com
- Data is available for all games starting in 2009 (soon 1998!)

Available on Github, install with:
devtools::install_github(repo=maksimhorowitz/nflscrapR)
Principles of nflWAR

- Publicly available data, code, and results; **reproducible**
- Interpretable in terms of game outcomes (e.g. points, wins)
- Account for uncertainty (football is a small sample game)
- Allow for objective decision-making by coaches/management
Our nflWAR Framework

- Create **open-source software** for the collection of NFL data
  see R package `nflscrapR` – Horowitz, Yurko, Ventura (2017)

- Properly model plays to determine **play value**

- Use play valuations to model **player value**

- Make player evaluation results **useful and interpretable:**
  - Evaluate relative to replacement level
  - Convert to a wins scale
  - Estimate the uncertainty in our evaluations of players
How to Value Plays?

Suppose it’s 4th down with 4 yards to go from the 40 yard line. You have three choices:

- **Punt:**
  You are sacrificing possession, but gaining (some) field position

- **Attempt a field goal:**
  You are sacrificing possession, but (possibly) gaining three points

- **“Go for it”:**
  You try to advance the ball four yards to maintain possession
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? (yard line, down, yards to go, etc.)
- Several ways to model this
- **Our approach:** multinomial logistic regression

**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
- **Our approach:** generalized additive model (GAM)

Can apply *nflWAR* framework to both (or any measure of play value)
Super Bowl LII Win Probability Chart

Data from nflscrapR
Win Probability Added (WPA) or Expected Points Added (EPA) Using air yards → airWPA/airEPA and yacWPA/yacEPA
How to Evaluate Players?

Comment from Pittsburgh Post-Gazette article on nflscrapR

Football is complex, need to divide credit, evaluate using wins!

Ron Yurko (@Stat_Ron)

nflWAR

RITSAC 2018
How to Evaluate Players?

Comment from Pittsburgh Post-Gazette article on nflscrapR

stats don't 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 doesn't go crazy over stats--it's a team sport on every single play. the only stat that counts is the W.
How to Evaluate Players?

Comment from Pittsburgh Post-Gazette article on nflscraper:

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.

Football is complex, need to divide credit, evaluate using wins!
Division of Credit

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)

---

![Diagram of the offensive line with labels for each position and gap]

QB
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

- Hockey example: WAR-on-ice (Thomas et al., 2013)

- Baseball example: Deserved Run Average (Judge et al., 2015)
Multilevel Modeling

Example of **varying-intercept** model:

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

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
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
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)**
Four different models to measure three skills:

- Two separate models for QB and non-QB **rushing**
- Two separate passing models for **air** and **yac**

Models adjust for team strength using opposite type of EPA per attempt (e.g. rushing models adjust for passing strength)

Every player has $iPA_{rush}$, $iPA_{air}$, and $iPA_{yac}$ estimates
QB Air vs Yac Efficiency in 2017

Gunslingers

Short but accurate
RB Receiving vs Rushing Efficiency in 2017

Pass-catchers

Pure rushers

air iPA + yAc iPA vs rush iPA

Replacement level

A.Kamara

C.Thompson

D.Sproles

C.Hyde

T.Gurley

M.Ingram

D.Lewis

Jo.Howard

E.Elliott

Ron Yurko (@Stat_Ron)
Comparing Team Offensive Line Performance 2016-17
Arriving at WAR

Evaluate relative to “shadow” replacement player based on rosters similar to openWAR (Baumer et al., 2015)

Results in individual points above replacement (iPAR)

Convert points to wins using regression approach

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

Two types of WAR:

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

\[
\text{WPA-based } WAR = iPAR
\]
Uncertainty is Mandatory

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

We resample entire team drives to preserve any play-sequencing tendencies that could affect our estimates

Following estimates are based on 1000 simulations
Top RBs by WAR in 2017

- A.Kamara
- T.Gurley
- K.Hunt
- L.Bell
- M.Ingram
- D.Lewis
- Jo.Howard
- L.Miller
- F.Gore
- D.Freeman

Simulated WAR Values

- air WAR
- yacc WAR
- rush WAR
Selection of QB WAR Distributions in 2017

![Graph showing simulated WAR values for multiple quarterbacks. Each quarterback is represented by a bell curve indicating their expected performance with shaded areas for different categories: air WAR, yac WAR, and rush WAR.](image-url)
Career WAR Leaders

WPA–Based, 2009 to 2018

A. Rodgers
D. Brees
T. Brady
P. Rivers
R. Wilson
B. Roethlisberger
P. Manning
C. Newton
M. Ryan
A. Smith
M. Stafford
C. Johnson
A. Brown
T. Taylor
J. Jones
T. Romo
B. Marshall

Total Career WAR

Data from NFL.com via nflscrapR; WAR from https://github.com/ryurko/nflscrapR–data
I’m Sorry Bills Fans

Good luck with Josh Allen!
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? QBs that rush more?

#GIVEMETRACKINGDATA
Future of Football Analytics

- **Brian Burke** (@bburkeESPN): father of modern football analytics - http://www.advancedfootballanalytics.com/

- **Josh Hermsmeyer** (@friscojosh): air yards, player stability, routes, etc - keeps work accessible, great visualizations

- **Zachary Binney** (@zbinney_NFLinj): NFL injury expert

- **Eric Eager** (@PFF_EricEager): Pro Football Focus collects everything - just don’t look at their barcharts...
Future of Football Analytics

- **Brian Burke** (@bburkeESPN): father of modern football analytics - http://www.advancedfootballanalytics.com/

- **Josh Hermsmeyer** (@friscojosh): air yards, player stability, routes, etc - keeps work accessible, great visualizations

- **Zachary Binney** (@zbinney_NFLinj): NFL injury expert

- **Eric Eager** (@PFF_EricEager): Pro Football Focus collects everything - just don’t look at their barcharts...

  ...there is another...
A New Hope

Michael Lopez (@StatsbyLopez) NFL Director of Data & Analytics
Clear your calendars for Oct 19-20th!
And visit https://cmusportsanalytics.com/conference2018.html for more information! #CMSAC18
Acknowledgements

Max Horowitz for creating nflscrapR

Sam Ventura for advising every step in the process

CMU Stats in Sports reading and research group

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

Max Horowitz for creating nflscrapR

Sam Ventura for advising every step in the process

CMU Stats in Sports reading and research group

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

Thanks for your attention!

No barcharts were harmed in the making of this presentation.
Expected Points Relationships

![Graph showing expected points relationships](image)

- **Yards from opponent's end zone**
- **Expected points value**

**Legend**:
- nflscrapR - 1st down
- nflscrapR - 3rd down
- nflscrapR - 2nd down
- nflscrapR - 4th down
- Carter
- Hidden Game of Football

**Predicted Probability**

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

- **Touchdown (7)**
- **Field Goal (3)**
- **Safety (2)**
- **No Score (0)**
- **-Safety (-2)**
- **-Field Goal (-3)**
- **-Touchdown (-7)**
Convert to Wins

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

$$iPAR_i = iPAA_{i,total} - iPAA_{i,\text{replacement}}^{i,\text{total}}$$
References I

numberfire.

Alamar, B. (2010).
Measuring risk in nfl playcalling.
*Journal of Quantitative Analysis in Sports, 6*(2).

The blindside project: Measuring the impact of individual offensive linemen.

Isolating the effect of individual linemen on the passing game in the national football league.
*Journal of Quantitative Analysis in Sports, 4*(2).


References V


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


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


References VII


References IX


References XII


References XIV


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


References XVI


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


References XVIII


Schoenfield, D. (2012). What we talk about when we talk about war.

*pitchRx: Tools for Harnessing 'MLBAM' 'Gameday' Data and Visualizing 'pitchfx'.*
R package version 1.8.2.


