Exploring NFL Rushing Plays Using Linear Models

...

Matt Bundas - AST550 - 12/8/2020

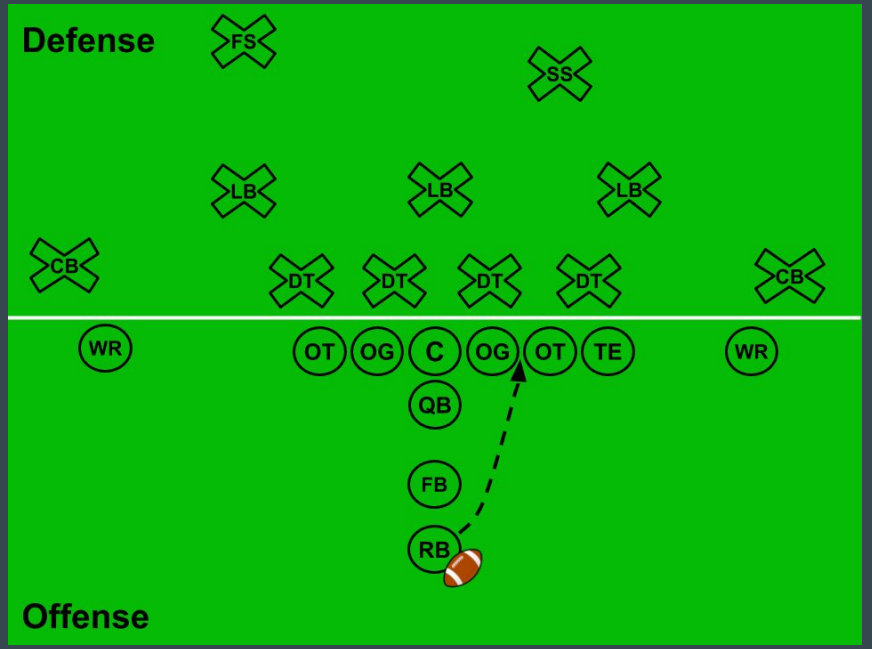
Overview Of Project

- Data of NFL rushing plays from 2017-2019, NFL Big Data Bowl Kaggle competition https://www.kaggle.com/c/nfl-big-data-bowl-2020
- ~30,000 rushing plays, used ~10,000 for computation sake
- Used MCMC LM in JAGS to explore what features can be used to predict play outcome
- Also did some simple statistical analysis on questions I had

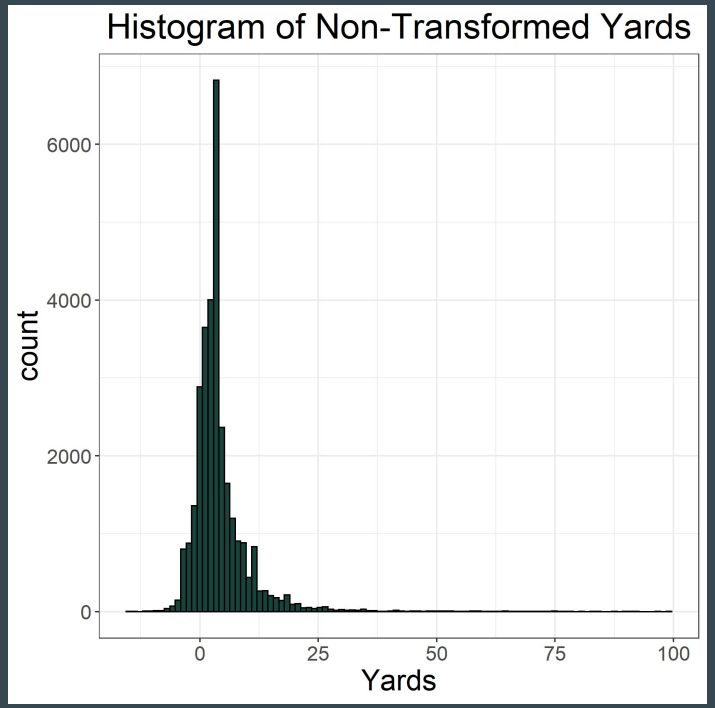
[1]

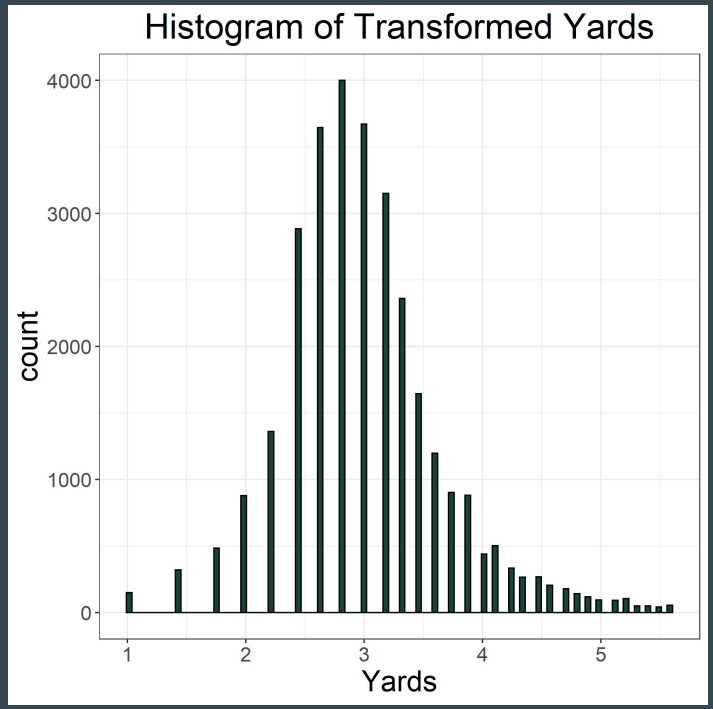
American Football Background

- Team with the ball tries to move it down the field with passing plays or rushing plays, they are the offense
- Team without the ball tries to stop them from doing that, they are the defense
- 11 players on offense, 11 on defense
- Rushing play - One player gets handed the ball, rest block for him.



Yards





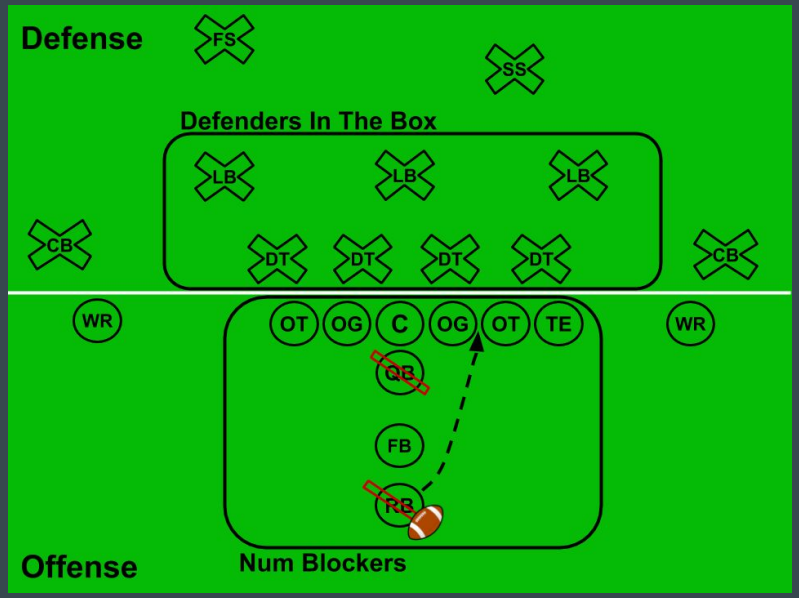
Average Play ~ 2-5 Yards

Bad Play ~ -5-0 Yards

Great Play ~ 10+ Yards

Features I Used in the Model

- Original data contains 50 features, rows for all 22 players on the field.
- Defenders In the Box - how many defenders in a good position to stop a run play.
- Num Blockers - how many players on offense in a good position to block for ball carrier.
- Winning - 0 if team is not winning, 1 if team is winning
- Distance - Yards offense needs for first down
- Weight - Weight of the ball carrier.



MCMC Linear Model 1

Yards ~ □0 + □1*DefendersInTheBox + □2*Winning + □3*Distance + □4*PlayerWeight + □5*NumBlockers

Features all scaled

Betas ~ dnorm(0, 0.01)

number of chains = 6

nu

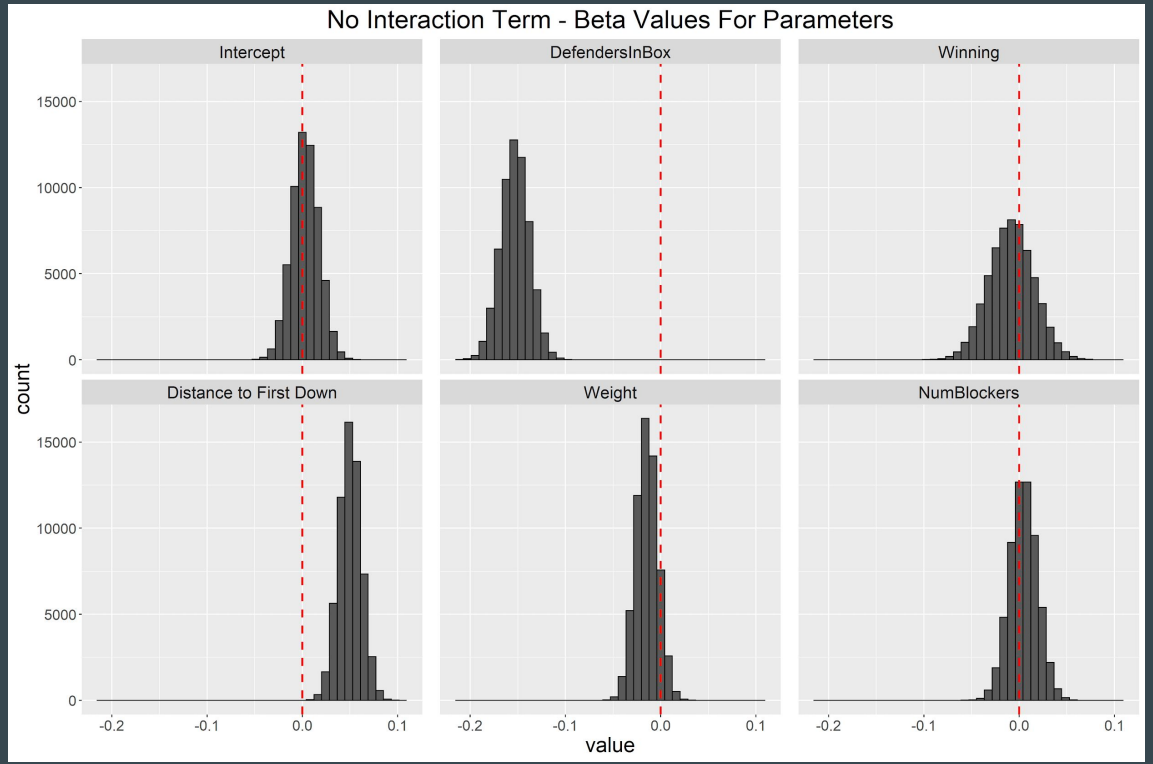
```
Potential scale reduction factors:

      Point est. Upper C.I.
beta0          1          1
beta1          1          1
beta2          1          1
beta3          1          1
beta4          1          1
beta5          1          1
sigma          1          1
tau            1          1
```

	beta0	beta1	beta2	beta3	beta4	beta5	sigma	tau
Lag 0	1.000000000	1.0000000000	1.000000000	1.0000000000	1.000000000	1.000000000	1.000000000	1.000000000
Lag 1	0.357196335	0.4259503938	0.367817012	0.0698940172	0.021282945	0.399131419	-0.002889549	-0.002919629
Lag 5	0.007613026	0.0138595295	0.004310701	0.0043936536	-0.001161046	0.014808253	-0.002506047	-0.002454289
Lag 10	-0.002091319	0.0008169144	-0.004531788	-0.0006546427	-0.001937407	-0.002111731	-0.004046905	-0.004118993
Lag 50	-0.007941936	0.0016641063	-0.004264180	0.0027232855	0.004338264	-0.002290092	0.003217879	0.003308968

beta0	28135.574143488
beta1	24315.3473057277
beta2	27637.8335542998
beta3	50984.1379327309
beta4	58180.0183492472
beta5	25094.2610080176
sigma	59692.010020745
tau	59687.5348642128

Linear Model 1 Results



95% Credible Intervals

Intercept	lower	-0.0241908926286246
	upper	0.0312609463913783
DefendersInBox	lower	-0.181477259435315
	upper	-0.12284402847121
Winning	lower	-0.0546886836310527
	upper	0.0377740612249682
Distance To FD	lower	0.0268157211023344
	upper	0.0728687093293225
Weight	lower	-0.0370436781581218
	upper	0.00770721887330902
NumBlockers	lower	-0.02367675707506
	upper	0.0337221230256186

MCMC Linear Model 2

Yards ~ □0 + □1*(NumBlockers - DefendersInTheBox) + □2*Winning + □3*Distance + □4*PlayerWeight

Features all scaled

Betas ~ dnorm(0, 0.01)

number of chains = 6

num[illegible] of iterations = 10,000

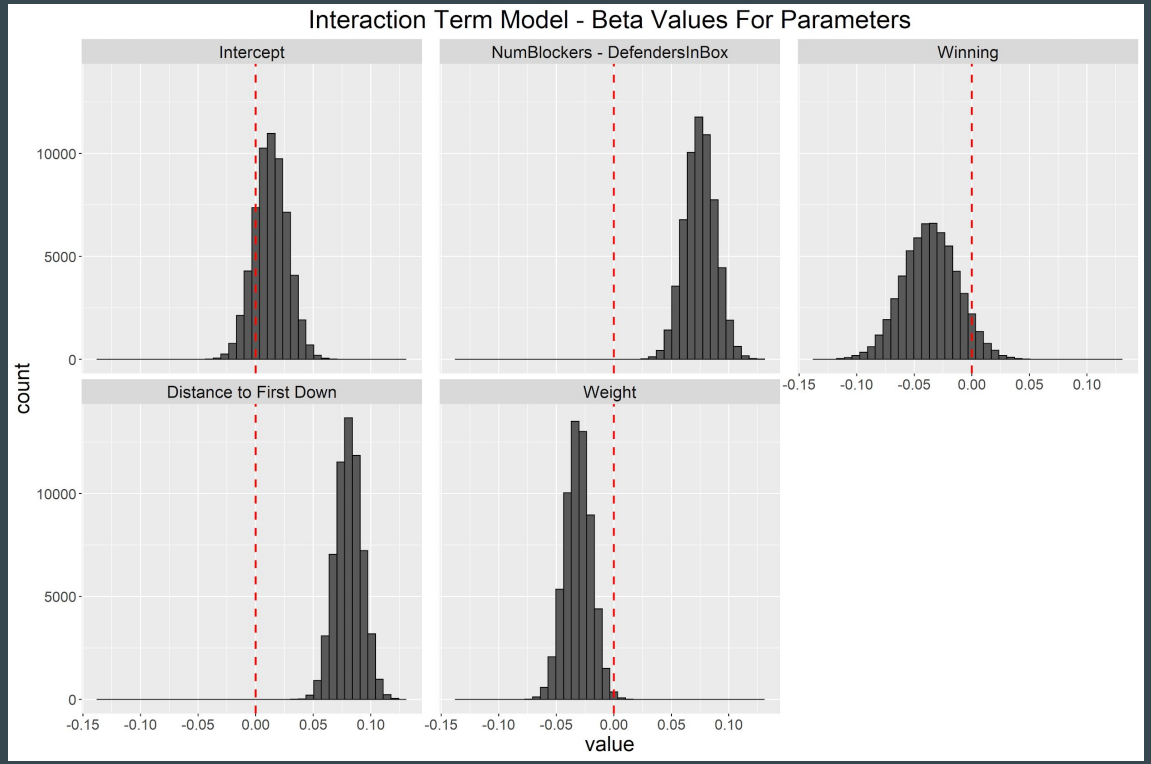
Potential scale reduction factors:

	Point est.	Upper C.I.
beta0	1	1
beta1	1	1
beta2	1	1
beta3	1	1
beta4	1	1
sigma	1	1
tau	1	1

	beta0	beta1	beta2	beta3	beta4	sigma	tau
Lag 0	1.0000000000	1.000000000	1.000000e+00	1.000000000	1.0000000000	1.0000000000	1.0000000000
Lag 1	0.3666966913	0.023988932	3.690151e-01	0.015854954	0.0006754666	-0.0017666119	-0.0018051772
Lag 5	0.0014076362	-0.003136446	3.178972e-03	-0.002066348	-0.0027017682	0.0038148966	0.0037493279
Lag 10	0.0011170330	0.003907541	6.909161e-05	0.005110798	-0.0027713085	0.0023868678	0.0024127676
Lag 50	0.0001604007	-0.001901811	6.744263e-04	0.001372938	0.0039702705	-0.0001393633	-0.0001492794

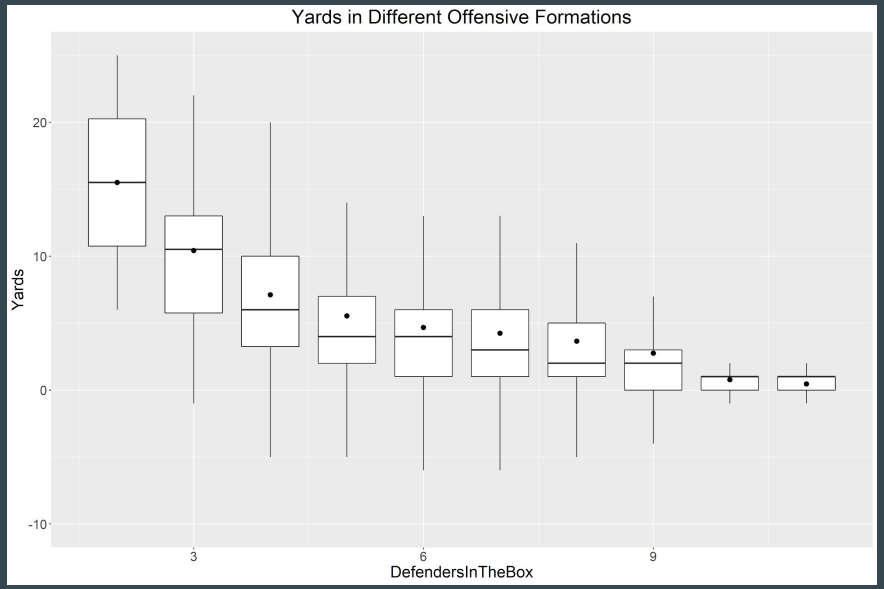
beta0	26417.2906399021
beta1	57160.5305739561
beta2	27621.2655852248
beta3	58634.0811290727
beta4	60138.374232259
sigma	60501.0064900432
tau	60525.2254105587

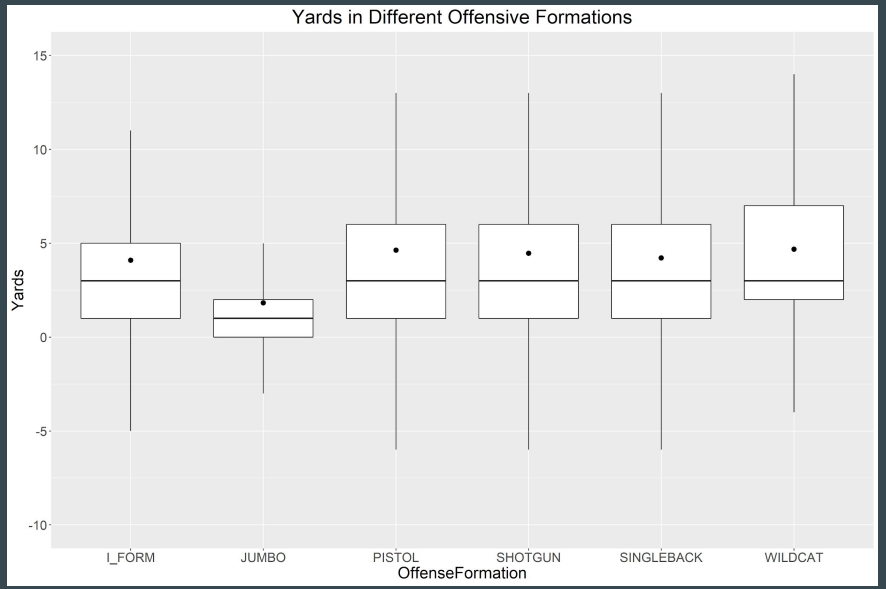
Linear Model 2 Results



95% Credible Intervals

Intercept	lower	-0.0146585471218115
	upper	0.041221054202261
NumBlockers - DefendersInBox	lower	0.0487394787746154
	upper	0.100918899205894
Winning	lower	-0.0820955099140881
	upper	0.0106675388990485
Distance To FD	lower	0.0587494404608197
	upper	0.103677314932593
Weight	lower	-0.0531853674299943
	upper	-0.00840223845067171





T-Tests

First Down vs Third Down

```
        Welch Two Sample t-test

data:  Yards by Down
t = 2.276, df = 3088.9, p-value = 0.02291
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04491775 0.60355279
sample estimates:
mean in group 1 mean in group 3
       4.307293        3.983058
```

RB vs WR

```
        Welch Two Sample t-test

data:  Yards by Position
t = -6.4212, df = 854.06, p-value = 2.241e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.496344 -1.327524
sample estimates:
mean in group RB mean in group WR
       4.193139        6.105072
```

Left vs Right

```
        Welch Two Sample t-test

data:  Yards by PlayDirection
t = 1.1589, df = 30990, p-value = 0.2465
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05869002  0.22848865
sample estimates:
 mean in group left mean in group right
          4.268371           4.183471
```

References

[1] NFL Big Data Bowl. (2020). Retrieved December 09, 2020, from https://www.kaggle.com/c/nfl-big-data-bowl-2020

[2] Trainor, P. (2020). *GLMs in JAGS*. Lecture.

[3] Trainor, P. (2020). *Lecture10NB* [Ipynb].

[4] Trainor, P. (2020). *Markov Chain Monte Carlo*. Lecture.