What was lost? A causal estimate of fourth down behavior in the National Football League [1]

Discussant: Derek Hansen

Sports Reading Group
University of Michigan Department of Statistics

September 25, 2020

[1] https://doi.org/10.3233/JSA-190294

Motivating Example

- https://youtu.be/fwWoBNELyeU?t=630

Background

- In American football, attempting to convert on fourth-down rather than punting or kicking a field-goal is a classic high-risk, high-reward decision.
- Successful conversion keeps possession and gives the team a chance to score a touchdown
- A failed conversion surrenders possession at the current spot on the field, which puts the opposing team in a good position to score.
- Did Bill Belichick make the right decision against the Colts in 2009?

Overview of Paper

- Previous studies in literature suggest NFL coaches are risk-averse and do not attempt to convert on fourth-down enough.
- The New York Times has a 4th Down Bot that gives real-time suggestions on whether a team should attempt to convert or not.
- This study focused on plays that the 4th Down Bot said to go for it
- This study used a matching algorithm based on propensity scores to estimate the difference in win probability if teams had gone for it instead.
- They find that the win probability increased in expectation, but also an increase in variance due to a bimodal distribution (confirming the high-risk, high-reward of going for it)

Past studies

- Carter and Machol (1978): Teams kick too many field goals (based on expected points framework)
- Romer (2006): Found that teams were not aggressive enough on fourth down (expected points)
- Burke and Quealy (2013): Expanded Romer (2006); authors helped create the NYT's "4th Down Bot"
- Causey et al (2015): Expanded Burke and Qualy to use more data and changed from expected points to win probability in cross-validated logistic regression

The Coachs' Opinions

- Despite empirical evidence mostly suggesting coaches are overly conservative, there is no known evidence that coaches' have modified their behavior since 2006.

> "There's so much more involved with the game than just sitting there, looking at the numbers and saying, 'OK, these are my percentages, then I'm going to do it this way,' because that one time it doesn't work could cost your team a football game, and that's the thing a head coach has to live with, not the professor." - Bill Cowher, former head coach of the Pittsburgh Steelers (Garber, 2002)

- "Analytics is not really my thing. I just try to evaluate what I see." - Bill Belichick [2]
- Belichick was also asked how much of a role analytics plays in his decision-making process, specifically in regards to two-point conversions and going for it on fourth down. Belichick responded, "less than zero."

[2] https://www.boston.com/sports/new-england-patriots/2019/09/27/bill-belichick-analytics-patriots-bills

Model

- Each observation is a play where the 4th Down Bot suggests to go for it.
- Y_i is the potential change in win probability [3] for the team on offense.
  - $Y_i(1)$ is the change in win probability if the team goes for it. (Treatment)
  - $Y_i(0)$ is the change in win probability if the team opts to kick or punt. (Control)

[3] As determined by a pre-trained model discussed in the Appendix. They tried two different models: random forest and GAM.

Data

- Data from ArmchairAnalysis.com from 2004 through 2016.
- $n = 13,172$ fourth downs identified where teams should have gone for it.
- $9,348$ teams (71.0%) did not go for it (i.e. the control group), and $3,824$ teams did go for it (i.e. the treatment group)
- Removed plays on which a penalty occured.
- https://github.com/statsbylopez/nfl-fourth-down for code (although data is behind paywall)

Model covariates used in win-probability model

Table 2

Descriptions of variables used in models of win probability (wp)

Variable	Description
Down	The current down (1st, 2nd, 3rd or 4th)
Score	Difference in offensive and defensive teams' score
Seconds	Number of seconds remaining in game
ScoreLeverage	$Score/\sqrt{Seconds+1}$
sprv	Las Vegas pre-game point spread
timo	Time outs remaining for the offensive team
timd	Time outs remaining for the defensive team
TotalPoints	Total points scored in the game
yfog	Yards from own goal
ytg	Yards to go for a first down

Model- Causal Inference Assumptions

- Stable unit treatment value assumption (SUTVA) is reasonable for this model
  - The potential outcome Y_i does not vary based on treatments assigned to other subjects.
  - There is no hidden variation of treatment ("decision to go for it is made prior to any play").
- W_i is the assignment of treatment.
- $Y_i^{obs} = Y_i(W_i)$ is the observed outcome, $Y_i^{mis} = Y_i(1 - W_i)$ is the missing outcome.
- X_i are covariates associated with play i (shown in a few slides)
- Assume $P(W_i = 0|X_i) > 0$ and $P(W_i = 1|X_i) > 0$.
- **Ignorability**: $P(W|X, Y_i(0), Y_i(1)) = P(W|X)$.

Average and Team-specific total treatment effects

- The Average Treatment effect on the Control (ATC) is defined when $W_i = 0$.

$$ATC_i = Y_i^{mis} - Y_i^{obs} = Y_i(1) - Y_i(0)$$

$$ATC = \frac{1}{N_c} \sum_{W_i=0} ATC_i$$

- The Total Treatment effect on the Control for team f (TTF_f) is:

$$TTC_f = \sum_{i:W_i=0} ATC_i 1[F_i = f]$$

where $f = 1, \ldots, 32$ are NFL teams and F_i is the team on offense for play i.

Propensity scores

- Very unlikely to find two plays where X_i is exactly the same
- Solution is to match plays with propensity scores, defined as:

$$e(X_i) = P(W_i = 1|X)$$

- These scores are estimated with a logistic regression model with spline terms for continuous variables, along with some interation terms. (see Table 3 in Appendix)
- Propensity scores are filtered out which are above maximum or below the minimum of either the treatment or control group.

Model covariates used in propensity scores

Covariates and Descriptions. All variables were obtained from Armchair Analysis unless otherwise indicated

Covariate	Description
yfog	Yards from own goal
ytg	Yards to go for a first down
pointdiff	Difference in offensive and defensive teams' scores: M4 (-17 or less), M3 (-16 to -9), M2 (-8 to -4), M1 (-3 to -1), T (0), P1 (1 to 3), P2 (4 to 8), P3 (9 to 16), P4 (17 or more)
time	Elapsed time in minutes
condcat	Weather category: Precipitation, Dry, or Dome
temp	Temperature at kickoff (in degrees Fahrenheit)
humd	Percent humidity
wspd	Windspeed at kickoff (in miles per hour)
sprv	Las Vegas point spread
ou	Las Vegas over-under (total points)
wp	Pre-snap win probability for the offensive team, averaged between two win probability models
Home	Factor variable for home or away
wk	Week of the season
OR.pass	Offensive team's pass offense rating (from Football Outsiders)
OR.rush	Offensive team's rush offense rating (from Football Outsiders)
DR.pass	Defensive team's pass defense rating (from Football Outsiders)
DR.rush	Defensive team's rush offense rating (from Football Outsiders)

Matching

- 1 : 1 nearest neightbor matching treatments to the control group using four play characteristics:
  1. $\text{logit}(e(X))$: propensity scores
  2. $\text{logit}(wp)$: pre-snap win probability
  3. *ytg* : yards to go
  4. *time* : number of minutes remaining in the game
- Calipers set to exclude observations that don't have a match which is within a certain threshold for each variable.

Matching Performance

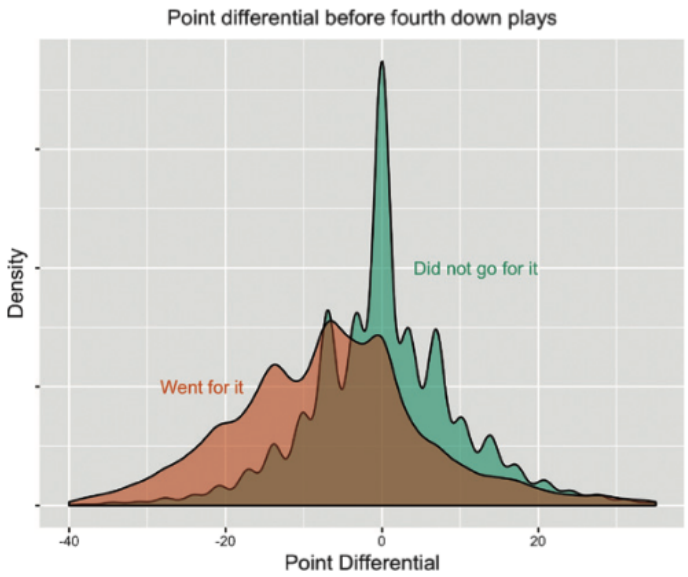


Fig. 1. Density curves showing the distributions of point differential (relative to the offensive team) among teams that went for it on fourth down and teams that did not. Shown are all fourth down plays from the 2004 through 2016 seasons (37,103 total plays) of the National Football League, using regular season games only.

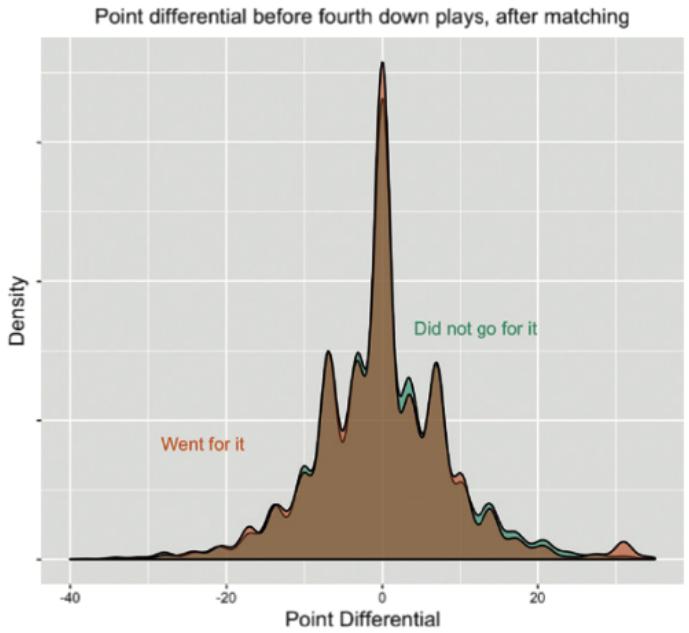
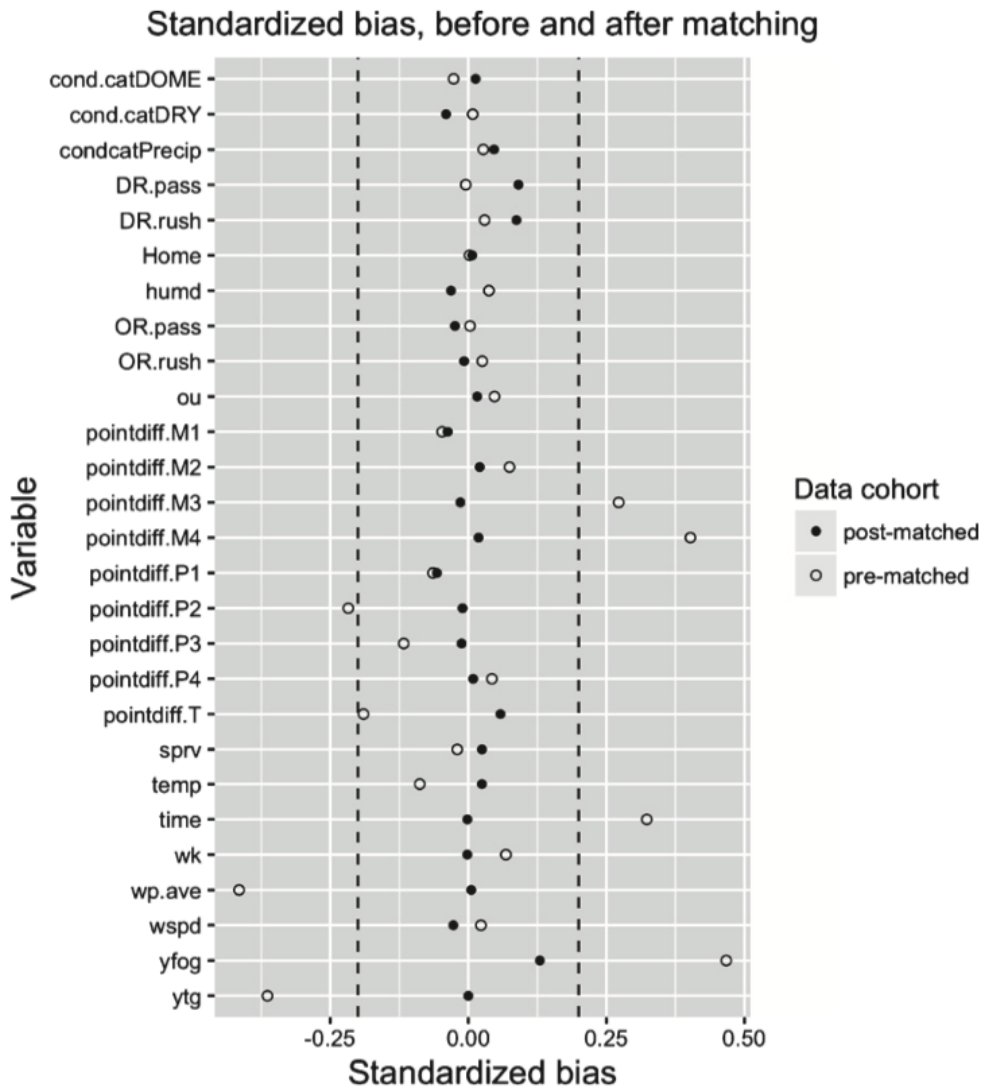


Fig. 3. Density curves showing the distributions of point differential among teams that went for it on fourth down and teams that did not. Shown are all 4th-down plays from the 2004 through 2016 seasons included in our matched analysis (7,698 pairs of plays).

Matching Performance



Findings

- Overall, teams that went for it have an average increase in win probability of 1.9%.
- The observed distribution of ATC_i was bimodal, corresponding to whether a team is successful or not.

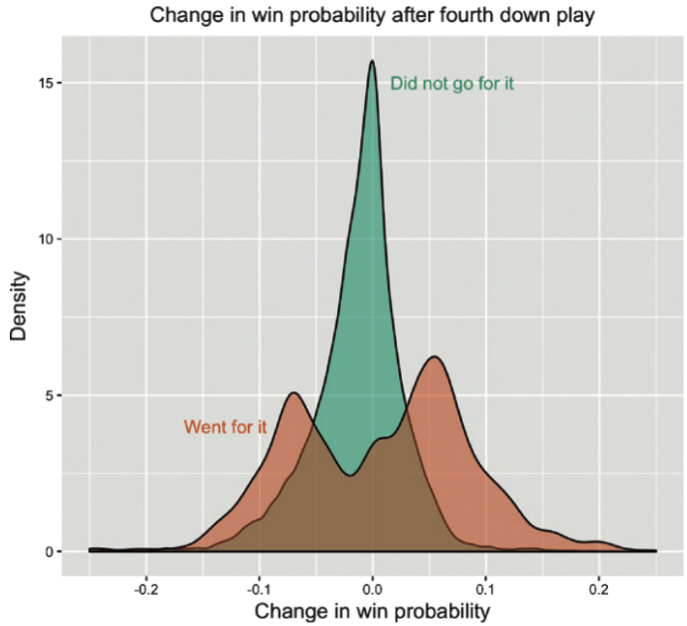


Fig. 5. Density curves for the changes in win probability on all matched fourth down plays. The average change in win probability for teams that went for it is about 1.9% higher than for teams that did not go for it (See Equation (2)). However, going for it also involves a larger variance in win probability changes, relative to not going for it.

Findings

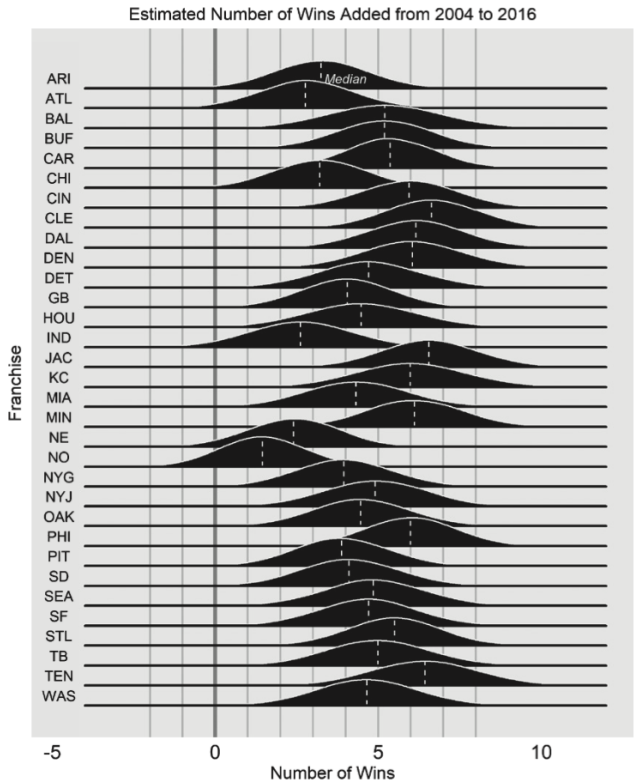


Fig. 6. Bootstrapped results for the estimated number of wins added per team (TTC_f) from 2004 to 2016, were each team to have adopted an aggressive fourth down strategy.

- Plot shows estimated change in wins from 2004-2016 if teams always went for it when 4th Down Bot said to (based on matched outcome).
- Team with highest number of expected increase in wins is Cleveland with 6.7.
- Only for New England, Indianapolis, and New Orleans is it feasible that the more aggresive strategy would not increase win total (matches perceptions that their coaches make better decisions on fourth down)

Discussion

- It would make more sense to group by Head Coaches than Teams to account for differences in philosophy
- It's hard to say that a coach isn't able to infer signal about $Y_i(1)$ and $Y_i(0)$ based on variables not in the covariates.

Discussion

- This paper maybe suggests that Bill Belichick was right to go for it on fourth down against the Colts.
  1. In 2005 against the Saints, New England faced "a 4th-and-2, 68 yards from its own goal, 10 minutes into a tied game, and successfully completed a four-yard pass, resulting in a $Y = \Delta wp = 8.5\%$".
  2. The increase in winning-percentage if New England had converted, $Y_i(1)|\{\text{converted}\}$, would likely have been higher. Indianapolis had one time-out and it would have been past the 2-minute warning.
  3. With two minutes on the clock, playing against a red-hot Peyton Manning and Reggie Wayne, the difference between $Y_i(1)|\{\text{not converted}\}$ and $Y_i(0)$ might not have been that high.