14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

# Correcting for preferential bias in NFL fourth-down decision making

Paper Track - Football
Paper ID - 987373

## 1. Introduction

Fourth-down decisions are some of the most important decisions NFL coaches make during a game. These decision involve teams and coaches deciding to either attempt to achieve a first-down (i.e. go-for-it), or kick the ball away (i.e. not go-for-it) by punting or attempting a field-goal. A fourth-down decision comes down to assessing the following probabilities: a successful fourth-down attempt, winning given a successful and unsuccessful fourth-down attempt, winning given a successful and unsuccessful field-goal attempt, and winning given a punt. Taking a weighted average of these win probabilities and choosing the maximum gives the decision that maximizes a team's win probability.

As discrete high-impact decisions, fourth downs present a unique opportunity for analytics to influence game outcomes. Given this opportunity, many papers have tried to analyze fourth down coaching decisions and suggest optimal choices depending on the game situation [1,2,3]. Public-facing models have been developed that automatically give recommendations for all fourth-down scenarios [4,5]. Several papers have shown that coaches act conservatively relative to what would be optimal to maximize their team's probability of winning [3, 6], though in recent years coaches have become more aggressive, adhering more closely to prescriptions suggested by these public-facing models [7].

While fourth-down play-by-play data is readily available for all fourth-down plays, we only observe fourth-down attempts conditional on teams being in a fourth-down situation and deciding to go-for-it. In general, we would expect teams that are more likely to succeed on a fourth-down attempt are more likely to go-for-it when given a choice, whereas teams less likely to succeed are more likely to punt or attempt a field goal. In certain situations where teams are forced to go-for-it given the game situation, for example when trailing by more than 3 points with less than 2 minutes left in the game, we expect teams less likely to succeed on a fourth down attempt to be in these situations more often. Teams with worse offenses, or who are playing worse that game, are more likely to be trailing and need to attempt fourth downs to get back into the game. Figure 1 shows empirical success probabilities for fourth down attempts separated by game situation. Fourth down attempts are successful more often when teams have a choice of whether to go-for-it, compared to when the game situation forces teams to go-for-it.

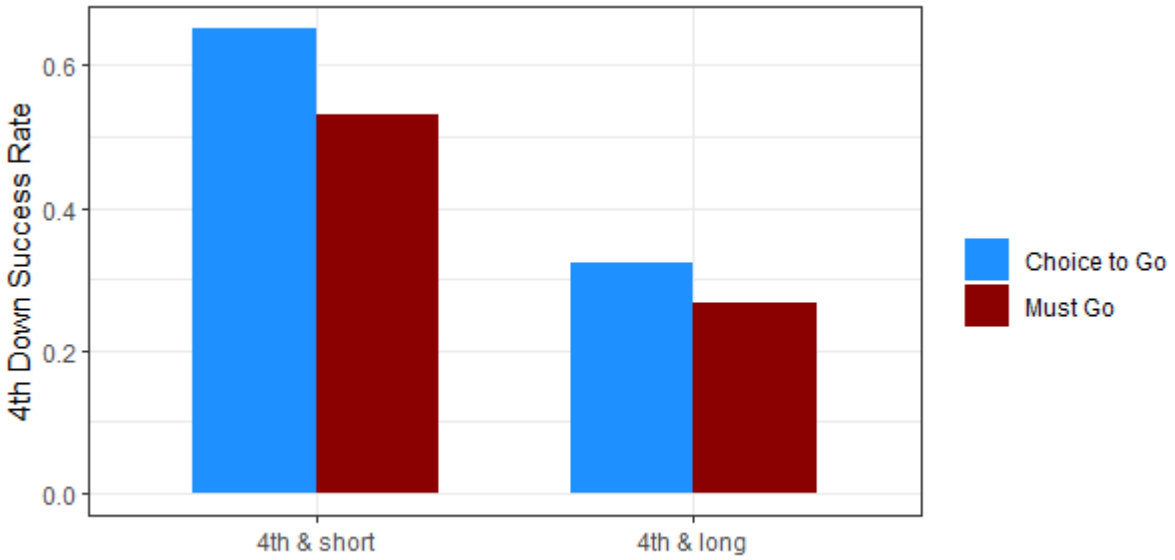


**Figure 1.** Average fourth-down success rates on attempts taken during the 2014-2021 regular seasons. Fourth and short is defined as 3 yards to-go or less, and fourth and long is defined as 8 yards to-go or more. The "must go" category are game situations where teams are forced to go-for-it, defined here as fourth downs with less than 2 minutes to go in the game, and either trailing by 1-16 points in their own half, or trailing by 4-8 or 12-16 points in the opponent's half. The "choice to go" category are game situations where teams have a choice whether to go-for-it or not, defined here as any fourth downs in the first or third quarters.

To produce unbiased fourth down probability estimates, models must include all variables that affect both the decision to go-for-it and the probability of success. If all variables are not included, comparisons made between the set of situations where teams go-for-it and situations where teams do not may be biased. This idea was explored in [8], who showed using tracking data that in the set of plays deemed "fourth and 1" in the play-by-play data, teams who went for it were on average closer to the first down line than teams who did not. This difference caused models that used play-by-play data and assumed all these plays had an equal distance to gain to be overly aggressive. In this paper we call this preferential bias, and our goal is to extend the ideas from [8] to correct for this bias over all fourth down success probability estimates. We frame this as a missing data problem, fitting a Heckman selection model to all fourth down play-by-play data from the 2014-2021 seasons, including situations where teams decided not to go-for-it. We compare these bias-corrected probabilities to those which ignore this bias to show current fourth down probabilities may be biased high and low in certain situations, leading to over- and under-aggressive decision recommendations. The remainder of our paper is outlined as follows. In Section 2 we introduce the missing data framework and the Heckman selection model. Section 3 describes the data and model fitting process. We describe our results in Section 4, and finish with a concluding discussion in Section 5.

# 2. Fourth downs as a missing data problem

In each fourth down scenario, we only observe one outcome of the two choices: either go-for-it or kick the ball away. The other outcome is missing, and if there is dependence between the decision-making process and the probability of fourth down success we will have bias in our fourth down estimates [9]. We can attempt to break the dependence by including all relevant covariates in our fourth down probability model. If, conditional on our included covariates, there is no dependence between the fourth down decision and outcome, then including these covariates would remove the bias in our estimated probabilities. This assumption is referred to as the missing at random (MAR) assumption [10]. However, it is unlikely that this assumption holds, especially when only using play-by-play data [8]. It is more likely that even conditional on included covariates, teams with higher probability of success are more likely to go-for-it when there are multiple viable choices. Additionally, teams that are forced to go-for-it based on the game situation likely have a lower probability of success given being in that situation usually means the team is trailing in the game. Data where there is dependence between missingness and the outcome is referred to as missing not at random (MNAR), and requires modelling of the missing data mechanism to avoid biased probability estimates [11].

## 2.1. Heckman selection model

In this paper we choose to use a Heckman selection model to account for the dependence between the missingness and outcome mechanisms [12, 13]. We assume a bivariate probit distribution for the probabilities of fourth down success and choosing to go-for-it. For fourth down play $i$ let $Y_i$ denote the binary outcome and $R_i$ denote the binary decision to go-for-it. We only have outcome data $Y_i$ for plays where $R_i = 1$, otherwise we consider the data missing. The Heckman model assumes the following models for $Y$ and $R$

$$\mathrm{P}(\mathrm{Y_i} = 1|X_i) = \Phi(X_i\beta) \quad (1)$$

$$\mathrm{P}(R_i = 1|Z_i) = \Phi(Z_i\gamma) \quad (2)$$

where $\Phi$ is the standard normal cumulative distribution, $\mathrm{Y_i} = 1$ if the fourth down play is successful and 0 otherwise, and $R_i$ is an indicator of missingness, equal to 0 if the team did not attempt to go-for-it on fourth down. The bivariate probit model assumes a latent bivariate normal distribution for $R_i^*$ and $Y_i^*$ of the form

$$\mathrm{Y}_\mathrm{i}^* = X_i\beta + \epsilon_i \tag{3}$$

$$R_\mathrm{i}^* = Z_i\gamma + \epsilon_i^r \tag{4}$$

$$\begin{pmatrix}\varepsilon\\ \varepsilon^r\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) \tag{5}$$

where $Y_i = 1$ if $Y_i^* > 0$ and $R_i = 1$ if $R_i^* > 0$. The correlation coefficient $\rho$ denotes the dependence between the probability of a successful fourth down and the probability of attempting the fourth down characteristic of our MNAR assumption. The MAR assumption is equivalent to assuming $\rho = 0$, in which case the selection model for $R$ would not affect our estimated probabilities in (1).

## 2.2 Generalized Heckman model

The Heckman model described above assumes a constant correlation parameter over all data. However, we do not expect this to be the case with our missing fourth down data. In scenarios where both going-for-it and kicking are viable options, we expect a positive correlation between Y and $R$ because teams more likely to succeed are more likely to choose to go-for-it. In situations where the game situation forces teams to go-for-it, we expect correlation to be in the opposite direction because teams in these situations are generally worse, or playing worse in that game, relative to teams not forced to go-for-it in these situations. While it is possible to condition our outcome model in (1) on the game situation, at best the correlation $\rho$ would be reduced to zero in these "must-go" situations, and we still suspect it to vary over the data (see Section 5 for further details). Instead, we decide to generalize the model to allow for the correlation $\rho$ to depend on game situation covariates [14, 15]. We assume

$$\arctan(\rho_i) = C_i\kappa \tag{6}$$

where $C_i$ are a set of game situation covariates, including time remaining and score differential. The hyperbolic tangent link function maps $C_i\kappa$ to the interval (-1, 1). Equations (1)-(6) make up our Heckman selection model. The joint modelling of outcome and missingness gives marginal probabilities of $Y$ estimated in (1) that account for the preferential bias in fourth down decision making.

# 3. Fitting the model

## 3.1. Data

We use play-by-play data from the 2014-2021 NFL regular seasons provided by the `nflfastR` package [16]. We include all fourth down plays where a rush, pass, punt, or field goal is attempted, excluding plays where a penalty resulted in a first down. This resulted in 29,239 fourth down plays, of which teams attempted to go-for-it 4,368 times. This data also includes Las Vegas closing spread and total lines taken from Pro-Football-Reference (PFR) which we use as proxies for team strength and offensive and defensive ratings. In addition, we use coaching data from PFR to determine which coach is involved in each fourth down decision, allowing us to encode coaching decision preferences into the model [17].

## 3.2. Model fitting

To fit the Heckman model described in Section 2, we estimate the parameters in (3)-(6) via Bayesian inference. We compute parameter posteriors using Hamiltonian MCMC methods with the `rstan` package V2.26.13 [18, 19]. Our covariates $X$ for the outcome model include yards to gain a first down, yards from the opponent's endzone, spread, and the total line, similar to the model given in [5]. In our selection model we include all covariates $X$ as well as game situation variables including time remaining, score differential, and timeouts remaining. We also include a coaching random effect variable to capture differences in preferences between coaches. See the Appendix for a full list of covariates.

Heckman models can suffer from collinearity issues when covariates in the outcome model (1) and selection model (2) are identical [20]. Typically, at least one instrumental variable is required; a covariate that is dependent with selection $R$, but independent of the outcome $Y$ given our covariates $X$. In our case we have a set of game situation variables that are included in (2) and not (1) that may fit this criterion. However, we also include these game situations covariates in our model for the correlation $\rho$ in (6). Since we expect coaching preferences by itself to be a weak instrument, we assist in the identifiability of parameters in (3) and (6) by treating a subset of fourth downs *a priori* as having a selection probability of 1. For these data the joint likelihood of (3), (4), and (5) marginalizes to the outcome model likelihood (3), allowing us to better estimate these parameters [21]. The criteria for a selection probability of 1 are given in the description of Figure 1. In total 502 out of the 4,368 fourth down attempts in our data satisfy these criteria, of which 497 (>99%) went for it. Code used in this project is publicly available on Github[1].

[1] https://github.com/danieldalygrafstein/nfl4th-heckman

# 4. Results

Fitting our model in Section 3 generates posterior estimates for all parameters $\beta, \gamma, \kappa$ and $\rho$ in equations (3)-(6). Results for the estimated posterior dependence $\rho$ between fourth down success and the fourth down decision process are given in Figure 2. We can see that fourth downs occurring during the start of the first or second half tend to have a positive correlation. This means we expect the observed probability of success of these fourth downs to be higher when teams decide to go-for-it, compared to the success we would have observed for teams that decided not to go-for-it. Similarly, near the end of the first and second halves, there tends to be a negative correlation between success probability and the decision to go-for-it. Teams in these situations are more likely to be forced into the decision based on the game situation, performing worse than teams that are not forced to go-for-it. Additionally, the correlation increases with the difference in score. Teams winning in the game are likely playing better, and thus more likely to succeed on an attempted fourth down relative to teams that are trailing.

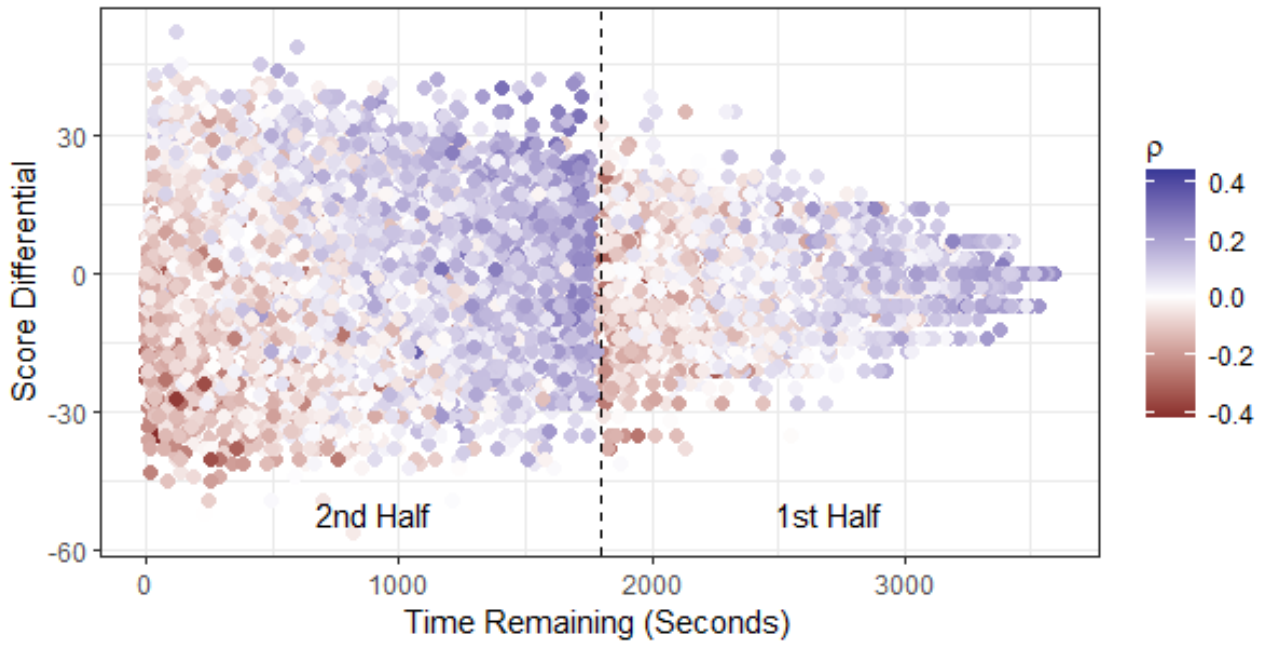


**Figure 2.** Posterior mean correlation $\rho$ for all 29,239 fourth down plays in our data. Score differentials are taken with respect to the team in possession, with a positive score differential indicating the team facing the fourth down is leading. While displayed over two covariate dimensions time and score differential, our model for correlation (6) includes additional covariates. See the Appendix for a full list of covariates.

### 4.1. Quantifying the preferential bias

Our model attempts to correct for the preferential bias captured in Figure 2. To quantify this bias we compare the posterior mean outcome probabilities in our model to a naïve model where just a probit regression on the observed fourth down outcomes is fit as in equations (1) and (3). In this naïve model we ignore all fourth down plays where teams decided to kick it away, implying a MAR for these fourth downs. Figure 3 compares the differences in the estimated probabilities between these two models over different yards-to-go and field positions. Overall, there is between a -0.15 and 0.09 difference between the mean posterior probability estimates of the naive and Heckman models. Positive differences (meaning the naïve model probabilities are higher) typically occur in short yardage situations in the opponent's half of the field. This is likely because better teams, or teams in better situations (in ways not captured by the success model (3)), will choose to go-for-it in these situations more often. Negative differences occur deep in a team's own half, or in fourth-and-long situations near the opponent's goal line. In these cases better teams are not going-for-it, because these fourth down attempts typically occur in desperate game situations. Overall, we find the model that does not correct for preferential decision-making bias is over-aggressive in fourth and short situations in the opponent's half, and under-aggressive in fourth and long situations near the opponent's endzone or in a team's own half.

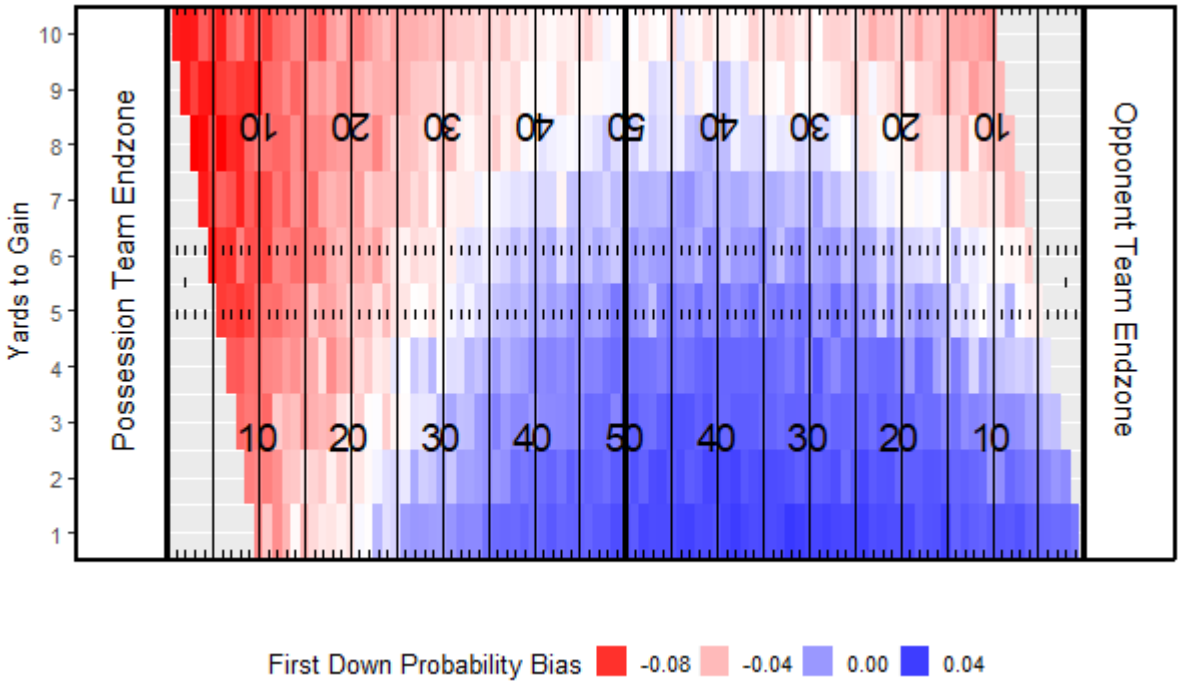


**Figure 3.** Difference in posterior mean probabilities between the Heckman model constructed in Section 2, and a naïve model probit regression using (3) over only fourth downs that were attempted. Positive bias indicates the naïve probability estimates are higher than the Heckman

model (i.e. a positive correlation between the decision and outcome) and negative bias indicates the opposite. Biases are averaged over all yards to gain, yard line combination.

## 4.2. Unbiased coaching preferences

In addition to game situation and team metrics, we include a coaching random effect in our selection model (4). We assume a $N{\sim}(0, \sigma^2_{coach})$ distribution for these parameters, with dummy variable encoding for each coach, equal to 1 if they were coaching when a fourth down play occurred. These random effects allow us to estimate coaching preferences independent of the game situation and team ability. Table 1 shows the top 5 and bottom 5 coaches in terms of fourth down aggressiveness. We find Doug Pederson to be the most aggressive coach, and Kyle Shanahan the least aggressive in our dataset.

| **Top 5 Aggressiveness** | | | **Bottom 5 Aggressiveness** | | |
|---|---|---|---|---|---|
| Coach | Random Effect | Go-for-it probability | Coach | Random Effect | Go-for-it probability |
| Doug Pederson | 0.278 | 0.671 | Kyle Shanahan | -0.168 | 0.499 |
| John Harbaugh | 0.203 | 0.643 | Jon Gruden | -0.167 | 0.499 |
| Raheem Morris | 0.172 | 0.631 | Mike McCoy | -0.164 | 0.500 |
| Brandon Staley | 0.162 | 0.628 | Bruce Arians | -0.154 | 0.504 |
| Mike McCarthy | 0.158 | 0.626 | Mike Mularkey | -0.152 | 0.505 |

**Table 1.** Mean posterior coaching preference random effects estimated as part of the selection model (4). The go-for-it probabilities are the estimated mean posterior probabilities that each coach will go-for-it on fourth down when in a fourth and 1 at midfield, in a tied game, at the start of the fourth quarter, during the 2021 season.

## 4.3. Examples where preferential bias influences fourth down recommendations

In this Section we give some examples where our Heckman model gives different fourth down recommendations than the naïve model conditioned only on fourth down plays where teams go-for-it. In general, our model gives slightly less aggressive recommendations in fourth-and-short situations, and slightly more aggressive recommendations in fourth-and-long.

For example, in a 2017 game between the Arizona Cardinals and Seattle Seahawks, the Seahawks were down 9 points with 13:08 left in the fourth quarter and faced with a fourth and 1 on the

Cardinals 31 yard line. In this situation the naïve model gives the Seahawks a 58% chance of succeeding on a fourth down attempt, while the Heckman model only gives them a 50% chance. Using win probabilities from the nfl4th package [5], the naïve model gives the Seahawks a 27.6% chance of winning if they go-for-it, while only a 26.2% chance of winning if they attempt a field goal. The Heckman model gives the Seahawks a 26.0% chance of winning if they go-for-it. The Seahawks kicked a field goal, which according to the naïve model is the wrong choice, but the correct one according to the Heckman model.

In another example, a 2018 week 16 game between the Pittsburgh Steelers and New Orleans Saints, New Orleans is trailing by 4 with 6:17 left in the fourth quarter. They are faced with a fourth and 11 from the Steelers 32 yard line. The naïve model gives the Saints a 33% chance of succeeding on a fourth down attempt, while the Heckman model gives the Saints a 41% chance of succeeding. Win probabilities give the Saints a 39.0% chance of winning if they go-for-it under the naïve model, a 41.8% chance of winning under the Heckman model, and a 40.5% chance of winning if they kick a field goal. The naïve model recommends a field-goal attempt, which is what the Saints did, while the Heckman model recommends they go-for-it.

# 5. Discussion

In this paper we developed a model to account for preferential decisions when modelling NFL fourth down success. We found there is a positive correlation between decisions and success when there are multiple viable choices for teams, and a negative correlation when teams are forced to go-for-it by the game situation. This causes modelled fourth down probabilities to be biased high in fourth-and-short scenarios, and biased low in fourth-and-long scenarios, which may result in over or under aggressive decision recommendations when not correcting for this bias. One alternative modelling approach we could have used would be to condition our fourth down model on the current game state by using win probability. This may remove the negative correlation found in fourth-down decisions in 'must-go' situations, however we would still expect the correlation between decision and outcome to vary over game state, with stronger positive correlation when both going-for-it and kicking are viable options (i.e. have similar win-probabilities).

We expect this preferential bias to decrease as the number of relevant covariates included in the model increases. The inclusion of tracking data, or team-specific covariates, would likely get us closer to the MAR assumption and reduce our estimated correlation coefficient $\rho$ [8]. On the other hand, the linearity assumptions made in our model equations (3)-(6) may have caused underfitting of the decision and selection models, causing an underestimation in the magnitude of $\rho$. Future work could explore extensions to the Heckman model that allow for nonparametric model equations [22, 23].

While in this paper we aimed to fit a model that would be applicable to teams league-wide, coaches almost certainly have fourth down decision models tailored to their specific teams. Conditioning on a given team should reduce preferential bias, as there is not variability between the skills of the different teams unaccounted for in the fourth down model. However, we would still expect some preferential bias, caused by week-to-week variability in team performance and opposition, dynamic in-game factors, play selection, or injuries. Thus even comparing within a team, we expect the same trend of positive and negative correlation seen in Figure 2. This means that conditional on a single

team, we expect the team is more likely to complete fourth downs they decide to go-for when given a choice, and less likely when forced into a decision by game situation. We recommend this preferential bias be accounted for when designing decision models, at least as part of a sensitivity analysis evaluating the MAR assumption of current models.

Almost all sports data we collect is observational, conditional on players, coaches, and teams deciding to perform an action. In many cases it is likely that models creating metrics and prescribing decisions based on these data do not satisfy the MAR assumption, and there is dependence between these decisions and the action's success. Examples from other sports could include choosing to take a shot in soccer or basketball, choosing to dump the puck in or attempt a controlled entry in hockey, or attempting to go for the green on a par 5 in golf. We should be aware of these potential biases, and correcting for them may lead to more accurate metrics and decision prescriptions.

# References


[1] Romer, D. (2006). Do firms maximize? Evidence from professional football. *Journal of Political Economy*, *114*(2), 340-365.

[2] Burke B., Carter S., Daniel J., Giratikanon T., & Quealy K. (2013). 4th Down: When to Go for and Why. https://www.nytimes.com/2014/09/05/upshot/4th-down-when-to-go-for-it-and-why.html?.

[3] Yam, D. R., & Lopez, M. J. (2019). What was lost? A causal estimate of fourth down behavior in the National Football League. *Journal of Sports Analytics*, *5*(3), 153-167.

[4] Burke, B., & Quely, K. (2013). How coaches and the NYT 4th down bot compare. http://www.nytimes.com/newsgraphics/2013/11/28/fourth-downs/post.html.

[5] Baldwin, B. NFL fourth-down decisions: The math behind the league's new aggressiveness. https://theathletic.com/2144214/2020/10/28/nfl-fourth-down-decisions-the-math-behind-the-leagues-new-aggressiveness.

[6] Sandholtz, N., Wu, L., Puterman, M., & Chan T. (2021) Risk measure inference in Markov decision processes: an application to fourth down decision making in American Football. New England Symposium on Statistics in Sports, 2021.

[7] Baldwin, B. (2021). 4th down research. https://www.nfl4th.com/articles/4th-down-research.html.

[8] Lopez, M. J. (2020). Bigger data, better questions, and a return to fourth down behavior: an introduction to a special issue on tracking data in the National football League. *Journal of Quantitative Analysis in Sports*, *16*(2), 73-79.

[9] Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley Sons.

[10] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

[11] Galimard, J. E., Chevret, S., Curis, E., & Resche-Rigon, M. (2018). Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC medical research methodology*, *18*(1), 1-13.

[12] Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4* (pp. 475-492). NBER.

[13] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.

[14] Bastos, F. D. S., Barreto-Souza, W., & Genton, M. G. (2020). A Generalized Heckman Model With Varying Sample Selection Bias and Dispersion Parameters. *arXiv preprint arXiv:2012.01807*.

[15] Saulo, H., Vila, R., Cordeiro, S. S., & Leiva, V. (2022). Bivariate symmetric Heckman models and their characterization. *Journal of Multivariate Analysis*, 105097.

[16] Carl, S., & Baldwin, B. (2022). *nflfastR: Functions to Efficiently Access NFL Play by Play Data*. https://www.nflfastr.com/.

[17] Sports Reference LLC. Pro-Football-Reference.com – Pro Football Statistics and History. https://www.pro-football-reference.com/. November 27, 2022.

[18] Stan Development Team. (2020). *RStan: the R interface to Stan*. http://mc-stan.org/.

[19] Young, J. C., & Minder, C. E. (1974). Algorithm AS 76: An integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics*, 455-457.

[20] Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of economic surveys*, *14*(1), 53-68.

[21] Campbell, H., de Valpine, P., Maxwell, L., de Jong, V. M., Debray, T. P., Jaenisch, T., & Gustafson, P. (2022). Bayesian adjustment for preferential testing in estimating infection fatality rates, as motivated by the COVID-19 pandemic. *The Annals of Applied Statistics*, *16*(1), 436-459.

[22] Ahn, H., & Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, *58*(1-2), 3-29.

[23] Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, *70*(1), 33-58.

# Appendix

## A.1. Covariates Included in the Generalized Heckman Model

| Covariate | Description | X | Z | C |
|---|---|---|---|---|
| ydstogo | Yards to go for a first down | ✓ | ✓ | ✓ |
| ydstogo_square | Yards to go for a first down squared | ✓ | ✓ | |
| yardline_100 | Yards from the opponent's goal | ✓ | ✓ | ✓ |
| yardline_100_square | Yards from the opponent's goal squared | ✓ | ✓ | |
| ydstogo_yardline_int | ydstogo*yardline_100 | ✓ | ✓ | ✓ |
| posteam_spread | Closing spread line of the team in possession | ✓ | ✓ | |
| total_line | Closing total line of the game | ✓ | ✓ | ✓ |
| year | Season the play occurred in (categorical) | ✓ | ✓ | |
| score_differential | Score difference | | ✓ | ✓ |
| half_seconds_remaining | Seconds remaining in the current half | | ✓ | ✓ |
| second_half | Second half binary variable | | ✓ | ✓ |
| posteam_timeouts_remaining | Timeouts remaining for the team in possession | | ✓ | ✓ |
| coach | Coach of the team in possession | | ✓ | |

**Table A1.** All covariates used in our Heckman model described in Section 2. The ticks in the final three columns indicate whether the covariate is included in model equations given by (3), (4), and (6), respectively. All numeric variables are standardized to have sample means of 0 and variances of 1. Second order and an interaction effect are included for the ydstogo and yardline_100 variables, while only main effects are included for other variables. Each team-year is assigned a single coach. If a team has multiple coaches in a season, we choose the one coaching for the most games. We encode coach parameters as a random effect, assuming coaching parameters follow a $N\sim(0,\sigma^2_{coach})$ distribution.