APPLICATION NOTE

Partially constrained group variable selection to adjust for complementary unit performance in American college football

A. Skripnikov
New College of Florida, Sarasota, FL, USA

ABSTRACT
Given the importance of accurate team rankings in American college football (CFB) – due to heavy title and playoff implications – strides have been made to improve metrics for team performance evaluation, going from basic averages (e.g. points scored per game) to metrics that adjust for a team’s strength of schedule, but one aspect that’s yet to be accounted for is the ability of team’s offense and defense to complement one another, termed ‘complementary football’. American football is unique because the same team’s offensive and defensive units typically consist of separate player sets that don’t share the field simultaneously, which tempts one to evaluate them independently. Yet, some aspects of your team’s defensive (offensive) performance may directly impact the complementary unit, e.g. turnovers forced by your defense could lead to easier scoring chances for your offense. Our main goal is to identify the most consistently influential features of complementary football in a data-driven way, subsequently adjusting each team’s offensive (defensive) performance for that of their complementary unit. To achieve that, for the 2009–2019 CFB seasons*, we incorporate natural splines with group penalty approaches, conducting partially constrained optimization to guarantee the full adjustment for the strength of schedule and home-field factor.

ARTICLE HISTORY
Received 17 November 2021
Accepted 4 January 2023

KEYWORDS
Group penalty; LASSO; natural splines; regularized estimation; reverse causality; sports statistics

1. Introduction

American football, while requiring immense physical ability, is an example of an extremely strategic sport. Requiring full mental engagement on every play from all 11 players your team puts on the field, underlines the importance of tactical preparation and accurate evaluation of the opposing team in advance of the match-up. Besides that, in American college football specifically, the ability to evaluate team’s performance objectively is pivotal for rankings, with the latter fully dictating the college teams that get into the championship playoffs and high-profile bowl games, including all the financial benefits that come with it.

CONTACT A. Skripnikov askripnikov@ncf.edu New College of Florida, Sarasota, FL 34243, USA

*Data and source code files are made available at https://github.com/UsDAnDreS/SUBMISSIONS_OffenseST_DefenseST_GLM_LASSO_adjusted_rankings

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02664763.2023.2166905.
Historically, classical averages have been used to evaluate a team’s performance in a certain facet of the game [1]. For example, the American football team’s scoring ability is ubiquitously described by an average amount of points scored per game. In contrast, its tendency to turn the ball over to the opposing team is typically captured in their average turnover count per contest. Such basic averages, albeit useful, are extremely flawed whenever trying to objectively evaluate and compare teams within the setup of Football Bowl Subdivision (FBS) of American college football, as they completely disregard contextual components such as quality of opposition and home field advantage, to name a few.

Football Bowl Subdivision (FBS), as of 2019, consists of 131 participating college football programs, split into 11 conferences. Each team is typically scheduled to play only 12 opponents during the regular season, with most of the games taking place against teams from the same conference, hence leaving the ranking picture incomplete in terms of relative strengths for teams across the entire FBS. Certain conferences could end up being especially strong or weak during a particular season, leading to the uneven quality of opposition which can’t be reflected in calculating each team’s basic per-game averages alone. That gave rise to what’s now known as ‘strength of schedule’ adjustment, originating in Harville [15] via an offense-defense model where points scored by a team are assumed to be a function of that team’s offensive strength and opposing team’s defensive strength, therefore adjusting for quality of the opponent. Somewhat analogous philosophy was implemented in Govan et al. [12] to calculate offense-defense ratings of teams in college basketball and American football, both college and professional. Lastly, given the importance of home-field advantage [9,24], the model in Harville [15] also accounted for whether a team played most of its games at home or away.

In this work, besides the strength of schedule and home field adjustments, we wanted to also incorporate the complementary nature between offense and defense in American football. Unlike most other sports, in American football, offense and defense are typically played by non-overlapping sets of players that don’t take the field at the same time. It frequently leads one to assume that performance of a team’s offense could be treated independently of that same team’s defense. Nonetheless, historically it’s been shown that the two sides are likely to complement each other, e.g. turnovers forced by your defense could lead to easier scoring chances for your offense. In contrast, your offense’s ability to control the clock may help your defense. That concept has been termed ‘complementary football,’ and the focus of this work was determining the most critical features of collegiate American football in affecting the complementary side’s performance. Specifically, we utilized a variety of regularized estimation techniques that impose a penalty for variable selection, implementing partially constrained optimization to guarantee strength of schedule and home-field adjustment. Somewhat related ideas were studied in Karlis and Ntzoufras [18], Boshnakov et al. [3] but in application to modeling the dependence between goals scored and allowed by a team in a soccer match, which doesn’t translate as well to our example due to soccer having a much more continuous flow between offensive and defensive play, with the same players being involved in both facets while always sharing the field at the same time. The only other sport that bears some resemblance to American football’s offense-defense dynamics is baseball, with each inning split into halves based on which team plays offense, so there’s no continuous flow between offense and defense for the same team. Pitchers play almost exclusively on defense, although it’s not unheard of to have a pitcher take some at-bats. With that being said, the vast majority of players on a baseball
team participate on both offense and defense, and, more importantly, no one ever uses the term ‘complementary baseball’ when it comes to the same team’s offense-defense dynamic. For example, regardless of what your offense does at the top of the inning, your pitcher would still be faced with the task of giving up the least amount of runs against the same opposing batters at the bottom of that inning. The only instances we could find of using the word ‘complementary’ when it comes to sports like baseball or soccer, in both popular and scientific literature, were when it came to players complementing each other’s skill sets as they share the field at the same time [11,20]. This indeed shows the uniqueness of American football from that standpoint and, to the best of our knowledge, there hasn’t been any involved research done on complementary impacts an offense can have on its defense, and vice versa, after having adjusted for strength of schedule and home field factor in college football.

The remainder of the paper is structured as follows. Section 2 lays out the details of data collection and cleaning, subsequently formulating all regularized estimation approaches that were used to find the most critical complementary football features affecting certain aspects of team performance. The results of feature selection, ensuing ranking adjustments, and their impact on predicting game outcomes are presented in Section 3. Moreover, the issue of endogeneity is brought up as a byproduct of reverse causal within-game dynamics of a football game, proposing use of solely efficiency-based statistics as a partial solution. Other potential solutions, along with general concluding remarks and discussions of future work, are covered in Section 4.

2. Materials and methods
2.1. Data collection and cleaning

Game-by-game data for 2009-2019 FBS seasons were obtained via web scraping from two primary sources: www.sports-reference.com/cfb/ and www.cfbstats.com/. Scraping was performed by leveraging rvest [26] and RSelenium packages [14], both being a part of R Statistical Software [22] – our primary tool for data handling and statistical analysis in this work. Certain defensive play statistics (e.g. tackles, sacks, forced fumbles) were lacking for ‘non-major’ teams, in other words, colleges not in the FBS. In these cases, which constituted 7% of our data, we imputed [10] the missing values using linear regression on other statistics that were available. For example, if values for a statistic such as sacks were missing, they were imputed by regressing on values of statistics that were available such as yards, touchdowns, and punts, using available sack values of FBS teams as a response.

Game by game data were converted into a format whereby each observation described both the offensive side and defensive side’s performance for an entire game. For instance, an offense’s statistics included yards gained, touchdowns scored, passes completed, and field goals scored, while a defense’s statistics included total tackles, tackles for loss, sacks, and quarterback hurries. To clarify, the ‘offensive side’ also incorporates offensive special teams, which includes field goal kicking and kick/punt returning, while the ‘defensive side’ includes such defensive special team plays as field goal blocking and kick/punt coverage. This data format resulted in two observations per game, with each observation representing a pairing of one team’s offense with its opponent’s defense. This scheme
implied a potential for within-game dependence of said observations. However, analysis of variance and Intraclass Correlation Coefficients [2] calculations for the residuals across all 66 models (spanning 11 CFB seasons and six response variables), concluded that game-by-game grouping effects were statistically insignificant for virtually all of those models, with intraclass correlation coefficients being consistently close to 0, providing evidence against within-game dependence.

2.2. Variable selection methodology outline

Below, we introduce the modeling notation, and main variable selection approaches to help adjust the team’s offensive (or defensive) quality for the strength of schedule, home-field factor, and complementary nature of American football.

2.2.1. Adjustment for the strength of schedule and home-field factor

For a total of \( n \) teams in the league, and a given response variable \( y \) (e.g. points per game), let \( y_{ij} \) denote the performance of team \( i \) against team \( j \) with regard to \( y \), \( i, j = 1, \ldots, n \).

For example, if \( y = \) (total points scored per game), then \( y_{ij} \) would represent points team \( i \) scored against team \( j \) and, by symmetry, points team \( j \) allowed against team \( i \). In case teams \( i \) and \( j \) faced each other \( L_{ij} \) times during the season, we add another index \( l \). Then, \( y_{ijl} \) corresponds to performance of team \( i \) with regard to \( y \) during its \( l^{th} \) meeting against team \( j \), \( i, j = 1, \ldots, n, \ l = 1, \ldots, L_{ij} \).

When adjusting teams’ performances in regard to response variable \( y \) for strength of schedule and home-field factor, we leverage the following modeling formula

\[
y_{ijl} = \mu + \alpha_i + \beta_j + \gamma h_{ijl} + \epsilon_{ijl}, \epsilon_{ijl} \sim N(0, \sigma^2), \quad i, j = 1, \ldots, n, \ l = 1, \ldots, L_{ij},
\]

where \( \mu \) simultaneously represents both the offensive and the defensive worth of the ‘league-average opponent’, which is possible due to the aforementioned symmetry between offensive and defensive outputs for two opponents (e.g. if \( y \) is points, then \( \mu \) denotes the average points per game, both scored and allowed, by all teams across all games that could have been played against one another throughout the course of the season); \( \alpha_i \) represents offensive margin of team \( i \); \( \beta_j \) denotes defensive margin of team \( j \); \( h_{ijl} \) denotes home-field indicator for \( l^{th} \) game between teams \( i \) and \( j \), taking on value 1 if team \( i \) is at home, 0 if the game site is neutral, \(-1\) if team \( i \) is away (such intuitive encoding was also confirmed by running a dummy-variable regression that showed this exact hierarchy and equidistant differences in team performances depending on the home-field value); \( \epsilon_{ijl} \) is the error term.

Offensive margin \( \alpha \) (defensive margin \( \beta \)) describes how much a team would outperform the defensive (offensive) worth \( \mu \) of the league-average opponent. The main assumption when adjusting for the strength of schedule is that performance of team \( i \) against team \( j \) in regard to response variable \( y \) is attributable to both the offensive margin \( \alpha_i \) of team \( i \) and the defensive margin \( \beta_j \) of team \( j \). Model (1) achieves two goals at once: it accounts for the quality of opposition by projecting each team’s performance onto the same average-league opponent (hence making it a more fair comparison as opposed to basic averages that disregard the strength of the opposition), and for the possibility of a team having an imbalanced schedule between home and away games, projecting everyone’s performances onto a neutral site.
2.2.2. Adjustment for complementary unit performance

To model potentially non-linear effects of complementary football features, as a well-known method, we utilized natural cubic splines [16], where one uses a mixture of piece-wise cubic and linear polynomials, smoothly connected at a set of $K$ knots placed across the range of the explanatory variable. It results in each complementary statistic $\alpha$ being represented by a set of basis functions $N_1(x_c), N_2(x_c), \ldots, N_{K-1}(x_c)$. For more detail, see [16]. We chose to use $K = 5$ knots placed at 0.00-, 0.25-, 0.50-, 0.75- and 1.00-quantiles, providing just enough flexibility to capture any clear non-linearity yet keep it interpretable and avoid overfitting.

Natural splines approach leads to each complementary football feature being represented by a group of several parameters, where the inclusion of a feature in the model implies including the entire said parameter group. At the same time, exclusion indicates potentially aiding with estimation efficiency and model selection. In our case, it leads to penalty application of penalties of different strengths to different groups of coefficients, potentially aiding with estimation efficiency and model selection. In our case, it leads to penalty application of penalties of different strengths to different groups of coefficients, potentially aiding with estimation efficiency and model selection.

For more detail, see [17, 25]. As we also want to guarantee adjustment for home-field effect and implying setting of the entire said parameter group to zero. Such setting warrants group penalization approaches [17, 25, 27]. As we also want to guarantee adjustment for home-field effect and strength of schedule adjustment effects, certain parameters will be left unpenalized, resulting in the following partially constrained group penalization criteria:

$$
\min_{\mu, \delta, |\alpha_i|^2, |\beta_j|^2} \sum_{i, j=1}^{n} \sum_{l=1}^{L_{ij}} (y_{ij} - (\mu + \delta h_{ij} + \alpha_i + \beta_j + \sum_{k=1}^{K-1} \gamma_{k} N_k(x_{ij})))^2 + \lambda \sum_{c=1}^{C} P_c(||\gamma_c||_2),
$$

(2)

where $\gamma_c = (\gamma_{c,1}, \ldots, \gamma_{c,K-1})$, capturing all the natural cubic spline coefficients for $c^{th}$ complementary statistic; $||\gamma_c||_2 = \sqrt{\sum_{k=1}^{K-1} \gamma_{c,k}^2}$ – the Euclidian ($L_2$) norm; $P_c(\cdot)$ indicates a specific group penalty function to be utilized for the coefficients of $c^{th}$ complementary statistic; $\lambda$ – tuning parameter responsible for the strength of penalty (larger $\lambda$ implies larger penalty for high values of $||\gamma_c||_2$). This partially constrained optimization criteria guarantees the inclusion of home-field ($\delta$) and strength of schedule adjustment effects ($\alpha_i, \beta_j, i, j = 1, \ldots, n$), while selecting the most critical complementary football statistics via group penalty imposed solely on coefficients $\gamma_{1, \ldots, C}$.

To identify the most consistently selected features regardless of a particular penalty choice, we considered an array of various group penalties. First, we utilized a classic group least absolute shrinkage and selection operator (group LASSO, [27]), $P_c(||\gamma_c||_2) = ||\gamma_c||_2$, which selects groups of variables. Adaptive group LASSO [25] is an extension that allows application of penalties of different strengths to different groups of coefficients, potentially aiding with estimation efficiency and model selection. In our case, it leads to penalty function $P_c(||\gamma_c||_2) = w_c ||\gamma_c||_2$, $w_c = 1/||\hat{y}_c^{LS}||_2$, where $\hat{y}_c^{LS}$ denotes the least squares estimate for the effect of $c^{th}$ complementary statistic resulting from optimization criteria (2) with the penalty term excluded. As an alternative to the two-stage approach that is adaptive group LASSO, we have considered single-stage approaches involving non-convex penalties – smoothly clipped absolute deviations (SCAD) and minimax concave penalty (MCP) – that might achieve analogous bias reduction for large regression coefficients. For details on the exact functional form $P_c(\cdot)$ for those penalties in the case of group LASSO,
To implement all of the aforementioned group LASSO penalties for optimization task (2) we utilized `grpreg` package \cite{6}, along with 20-fold cross-validation (CV) \cite{7} and BIC to select several candidate values for the tuning parameter \(\lambda\). For CV, we opted to use a \(\lambda\) value that achieved CV test error within one standard deviation of the minimal CV test error while providing a sparser model (‘sparse CV’). Moreover, to reduce the randomness effect of the 20-fold CV procedure, we ran it five times for each model, only counting features selected at least four out of five times. We used BIC and sparse CV because they imposed the strongest penalties on overly complicated models, virtually guaranteeing that if a complementary football feature gets selected by these criteria, it would have also been selected by any other ubiquitous criteria.

Each complementary statistic was evaluated on selection consistency by the four main variable selection methods outlined above (classic and adaptive group LASSO, group SCAD, and MCP penalties). Each method was given an equal weight of \(\frac{1}{4}\) in the final consistency metric, while within any individual method, an equal weight of \(\frac{1}{2}\) was granted to BIC and sparse CV as approaches to select the tuning parameter. Lastly, to calculate the final adjustment for each season, natural cubic splines were fitted solely with the consistently selected complementary football features while centering those features around their respective means. That way, the team \(i\)'s offensive worth \(\mu + \alpha_i\) and defensive worth \(\mu + \beta_i\) gets projected not only onto the league-average opponent and neutral site but also onto the league-average complementary side.

\subsection{2.2.3. Modeling assumptions given response variable type}

We considered six statistics to be adjusted, breaking it down by the type of a statistic (points, touchdowns, yards) and whether ‘counter-plays’ are accounted for or not. ‘Counter-plays’ are plays that happen directly after a turnover, before (if at all) the complementary side takes the field. For example, we could solely count touchdowns that your offense scored, or we could also take the margin between touchdowns they scored and touchdowns they allowed the opposing defense to score (via a returned interception or fumble). Same for points and yards. These margin statistics are useful as they account for the totality of the impact whenever a particular side takes the field, rather than reflecting their positive impact only. Lastly, while points and yards are well-approximated by continuous distributions, touchdowns exhibit more of a discrete and low-count dynamic, which led us to consider count-response approaches such as Poisson and negative binomial generalized linear models. Alas, those didn’t improve on the classic linear regression fit; hence we stuck with the latter moving forward.

\section{3. Results}

This section is divided into two main parts based on the complementary football feature sets being considered. The first part works with the entire set of statistics originally obtained from \texttt{www.sports-reference.com/cfb/} and \texttt{www.cfbstats.com/}. The second part discusses using an efficiency-based feature set of solely per-play and per-possession
statistics as a way to partially address the issue of reverse causality for the first set. Both sets can be found in the supplement.

3.1. Using the entire feature set of complementary football statistics

3.1.1. Collinearity
Prioritizing the interpretation and consistency of variable selection results, we disposed of perfect and near-perfect multicollinearity among our features. For perfect multicollinearity cases of one variable representing a sum of its ‘sub-categories’, we chose to retain only the sub-categories. For example, passing attempts equals completions plus incompletions, so we dropped passing attempts, retaining only completions and incompletions. The only exceptions were the cases where breakdown into sub-categories was not useful in terms of its effects on the complementary side. For example, as long as it is a turnover, distinguishing between fumbles and interceptions is unnecessary, as the impact on the complementary side won’t change based solely on that aspect. On the other hand, there is a considerable difference between pass completions and incompletions, where the latter stops the game clock while the former doesn’t. For near-perfect multi-collinearity cases, we picked the more ubiquitously used statistic to represent the highly correlated group. For example, points per game are strongly correlated with touchdowns per game (≥ 0.95), so we pick points per game as a more all-encompassing and recognizable statistic. For the resulting set of roughly 50 features, see the supplement.

3.1.2. Selected complementary statistics, issue of reverse causality
Among variables that were consistently selected as affecting the team’s complementary side regarding points scored/allowed in a game, we got rushing attempts, pass incompletions, non-scoring turnovers, number of attempted special teams returns of the ball, and yards gained on those returns. Some of these are perfectly reasonable, e.g. non-scoring turnovers provide the complementary side with the ball and a potentially easy scoring opportunity. Meanwhile, for a variable such as pass incompletions, while intuitively having a positive impact on the points scored on the complementary side (due to stopping the clock and leaving more game time to the complementary side), it rarely features in discussions about factors driving complementary scoring. Lastly, some variables simply don’t make as much sense, e.g. the number of attempted returns or rushing plays, regardless of gained yardage.

Why do we observe such unintuitive strong effects for some of the statistics? Reverse causality is the most likely answer. For example, pass incompletions can be a partial function of play selection by the team trailing in the score. Such a team would be forced to throw the ball more often in order to gain yards quicker while also stopping the clock in case of an unsuccessful play (unlike an ineffective rushing attempt), trying to score while taking as little of the game clock as possible. For rushing attempts, on the other hand, reverse logic applies: the leading team is likely to call more rushing plays to keep the game clock running, while the trailing team is less likely to do so. Lastly, the special teams return data (attempts, yards) suffers from the same issue but has an even clearer time-sequential component: a return attempt tends to follow directly after a score on the complementary side. Despite a lack of access to sequential play-by-play data, our data allowed us to calculate that 68% of attempted returns followed after a kickoff, which in its turn is most likely to come after a score on the previous possession by the complementary side (60% chance it was a
touchdown, 18% chance – a field goal). The aforementioned examples point to reverse causality permeating the estimation of complementary feature effects.

How can we potentially alleviate these reverse-causal effects? Even though there are several potential solutions (see Section 4 for an extended discussion), as the most feasible one, we decided to pursue an intuition-driven approach of focusing solely on efficiency statistics. While such cumulative metrics as pass incompletions or rushing attempts can be a strong function of strategic play selection within the game context, efficiency-based statistics are more reflective of a team’s effectiveness once a certain play is called, rather than heavily depending on the overall game context. For example, suppose a trailing team pursues a pass-heavy strategy. In that case, their cumulative passing statistics (attempts, completions, incompletions, yards) would inevitably increase, while the same isn’t necessarily true for their efficiency-based alternatives (completion percentage, yards per attempt). Hence, such efficiency-based statistics aren’t as affected by the game context, thereby reducing the extent of reverse causality.

3.2. Using only efficiency-based complementary football statistics

3.2.1. Efficiency-based features

Having disposed of collinearity issues via analogous reasoning to that of Section 3.1.1, in the row names of Figure 1 one can witness the efficiency-based complementary football statistics that ended up being utilized for variable selection. For each statistic, we had to decide between using a per-play or per-possession efficiency. We treated plays that terminate a possession (turnover, points scored, punt, etc.) on a per-possession basis, making them into possession percentages. For example, \textit{Off\_ST\_NonScoring\_TO\_PossPct} denotes the percentage of possessions ending in the team’s offensive side turning the ball over to their defensive side (as opposed to an instant score by the opposing defensive side, which is captured via \textit{Def\_ST\_TD\_PossPct}). For offensive scoring, we simply went with points scored per possession (\textit{Off\_ST\_Points\_PerPos}), as any scoring play resulted in the same direct impact on the complementary side (a kickoff). Lastly, punts and safeties were combined into a single variable because, despite safeties resulting in two points scored (to punt’s zero), more importantly both plays result in change of possession along via the ball getting kicked across the field.

In the meantime, virtually all other metrics were treated on a per-play basis, including yards or first downs gained, overall tackles, tackles for loss, and forced fumbles. For yards gained, there were several alternatives to looking at average yards per play by the offense: breaking it down by play type (pass or rush) and whether to include special teams returns as plays. The former idea was tested via cross-validation, having shown that total yards per play outperformed the pass/rush yardage breakdown approach. The latter idea resulted in special teams’ return yardage overly affecting the per-play efficiency values, leading to this statistic getting selected every single time, mostly due to the aforementioned reverse causality issue. One may notice that we didn’t include statistics such as field goal percentage, average yards per punt or per attempted return, red zone conversion percentage, etc. That is due to the issue of a low denominator, where efficiency is calculated over a really small sample size within a game (e.g. only over two field goal attempts), and hence is unreliable, and at times inaccessible (e.g. zero attempts).
Lastly, such statistics as ‘sacks to pass attempts ratio’ were defined in order to adjust for the effects of pass-heavy play calling being conducive to such defensive plays as sacks or quarterback hurries. Given that sacks don’t count as a passing attempt, and that we didn’t have data on quarterback dropbacks, this approach was a reasonable alternative. Similarly, the percentage of forced fumbles was counted for rush attempts and pass completions – the only plays where a fumble was possible.

### 3.2.2. Selected complementary statistics

Figure 1 demonstrates the efficiency-based features of complementary football (rows) that affected each respective response variable to be adjusted (columns). As described in Section 2.2.3, ‘margin’ accounts for counter-plays with regard to the respective statistic. One can witness non-scoring turnover percentage forced by your defensive side or committed by the opponent’s offensive side (hence the ‘Off_ST’ prefix) as the sole statistic that is consistently picked (≥ 74% of the time) in terms of its impact on points and touchdowns scored by your complementary offensive side. Meanwhile, for both total yardage response variables, the only complementary football statistic stably selected is the total yards per play (≥ 94% of the time). To evaluate the nature of these consistently selected complementary statistics, we fit the respective natural cubic spline regressions, with Figure 2 depicting the effects of your defense’s forced non-scoring turnovers on your offensive scoring (left), and of per-play yardage allowed by your defense, or gained by the opponent’s offense, on total yardage gained by your offense (right). Each non-scoring turnover forced by your defense mostly leads to higher offensive scoring, with the majority of years exhibiting a roughly linear relationship. This is a reasonable finding due to each such turnover instantly providing your offense with a ball, likely in a good scoring position. On the other hand, yards-per-play allowed by your defense shows a distinct non-linear impact on your offense’s ability to gain yards, which actually aligns well with the intuition. Defenses that allow either extremely few
Figure 2. Non-linear effects of complementary statistics on respective response variables, controlling for strength of schedule and homefield factor. Left: Effect of non-scoring turnovers forced by your defense on points scored by your offense. Right: Effect of per-play yardage allowed by your defense on yards gained by your offense.

(< 2.5) or many (> 5) yards per play tend to take less of the game clock themselves, albeit for polar opposite reasons. Good defenses (low per-play yardage allowed) would quickly stop the opposing offense and force them to turn the ball over to the complementary side, putting the latter in a position to gain yards of their own. Bad defenses (high per-play yardage allowed) would let the opposing offense score quickly, which would still lead to the complementary offense getting the ball. Meanwhile, if your defense allows a moderate 2.5–5 yards per play, that can make for longer offensive possessions by the opponent, taking away game time from your offensive unit that could be used to gain more yards.

Note that, given the symmetric two-sided nature of game sports, the reasoning above could also be observed from the opposing team’s viewpoint of their offensive side affecting their complementary defensive side. Each gain by one team’s offense could be considered a loss by the opponent’s defense and vice versa.

3.2.3. Ranking shifts in points scored per game

To illustrate the mechanism behind adjustments for complementary football features, we showcase its impact on a team’s offensive scoring ranking if one accounts for turnovers forced by that team’s defense. Table 1 demonstrates the ranking shifts in points per game scored by the offense in the 2012 season when, on top of adjusting for strength of schedule and home-field factor, the defense’s ability to turn the ball over to the offense is accounted for. Oregon’s scoring, while remaining the best in FBS, dipped by a whole point per game once projected onto an average complementary unit, vastly due to their defense being elite in forcing the non-scoring turnovers (#4 in the nation). A similar example happened to the Washington offense, complemented by a Top 5 defense. On the other hand, Texas A&M’s offense climbed into the Top 3, as its complementary defense ranked outside the Top 100 in takeaways, not providing the offense with as many easy scoring opportunities. The biggest increases in projected points per game happened to Texas Tech and South Florida offenses,
both working aside some of the worst defenses when it came to forcing turnovers – ranked 124th and 123rd, respectively, out of 125 teams total. The intuition behind ranking shifts in other years was identical.

We additionally provide Table 2 to demonstrate the defensive ranking shifts for the same season. Symmetrically, the strongest positive (negative) impacts happened to the defensive units accompanied by offensive sides unable (able) to consistently take care of the ball and avoid turning it over to the opponent’s defense. For example, teams like Illinois, Florida State, and South Carolina, all sporting below-average offensive units in ball security, had

### Table 1. Final 2012 season rankings (and ranking shifts) in points scored per game by team’s offense, adjusted for defense’s ability to turn the ball over to team’s offense.

| Team              | Points Scored (Per Game) | Non-Scoring Turnovers Forced (Per Possession) |
|-------------------|--------------------------|----------------------------------------------|
|                   | Value (Shift) Rank (Shift) | Value Rank                                   |
| Oregon            | 48.43 (↓ 1.02) 1 (0) | 0.19 Rank 4                                 |
| Louisiana Tech    | 47.12 (↑ 0.19) 2 (0) | 0.13 Rank 52                               |
| Texas A&M         | 46.67 (↑ 0.75) 3 (↑ 1) | 0.08 Rank 114                            |
| Oklahoma State    | 46.15 (↑ 0.11) 4 (↓ 1) | 0.12 Rank 75                               |
| Baylor            | 45.25 (↑ 0.01) 5 (0) | 0.12 Rank 61                               |
| Clemson           | 41.02 (↑ 0.08) 6 (0) | 0.12 Rank 67                               |
| Oklahoma          | 40.93 (↑ 0.26) 7 (↑ 1) | 0.09 Rank 111                            |
| Alabama           | 40.57 (↑ 0.17) 8 (↓ 1) | 0.17 Rank 13                               |
| West Virginia     | 39.74 (↑ 0.06) 9 (0) | 0.10 Rank 97                               |
| Tennessee         | 39.69 (↑ 0.58) 10 (↑ 2) | 0.09 Rank 108                           |
| Texas Tech        | 39.44 (↑ 1.22) 11 (↑ 4) | 0.06 Rank 124                           |
| Georgia           | 38.98 (↑ 0.19) 12 (↑ 1) | 0.16 Rank 23                              |
| South Florida     | 21.99 (↑ 1.16) 87 (↑ 6) | 0.06 Rank 123                           |
| Washington        | 23.72 (↓ 1.02) 80 (↓ 5) | 0.19 Rank 5                                |

### Table 2. Final 2012 season rankings (and ranking shifts) in points allowed per game by team’s defense, adjusted for offense’s ability to avoid turning the ball over to opposing offense.

| Team              | Points Allowed (Per Game) | Non-Scoring Turnovers Committed (Per Possession) |
|-------------------|---------------------------|--------------------------------------------------|
|                   | Value (Shift) Rank (Shift) | Value Rank                                     |
| Alabama           | 8.70 (↑ 0.42) 1 (0) | 0.09 Rank 21                                 |
| Notre Dame        | 11.85 (↑ 0.42) 2 (↑ 1) | 0.10 Rank 32                                 |
| Florida           | 11.88 (↑ 0.80) 3 (↓ 1) | 0.09 Rank 19                                 |
| Michigan State    | 14.03 (↑ 0.46) 4 (↑ 1) | 0.09 Rank 14                                 |
| Stanford          | 14.05 (↑ 0.88) 5 (↓ 1) | 0.10 Rank 29                                 |
| Brigham Young     | 14.26 (↑ 0.10) 6 (0) | 0.13 Rank 70                                 |
| South Carolina    | 15.74 (↑ 0.60) 7 (↑ 1) | 0.13 Rank 77                                 |
| Florida State     | 15.97 (↑ 0.44) 8 (↑ 1) | 0.15 Rank 92                                 |
| Rutgers           | 16.56 (↑ 0.31) 9 (↑ 2) | 0.13 Rank 69                                 |
| Kansas State      | 16.84 (↑ 0.88) 10 (↑ 3) | 0.08 Rank 9                                  |
| Louisiana State   | 17.06 (↑ 0.25) 11 (↓ 1) | 0.10 Rank 33                                 |
| Texas Christian   | 17.31 (↑ 0.60) 12 (↑ 2) | 0.16 Rank 104                               |
| Pittsburgh        | 25.07 (↑ 1.26) 51 (↓ 8) | 0.06 Rank 2                                  |
| Illinois          | 27.41 (↑ 1.27) 64 (↑ 5) | 0.20 Rank 124                               |
Table 3. Predicting binary game outcomes (won/lost) via logistic regression, and numerical scoring differentials via linear regression, across 2009–2019 seasons by using variously adjusted points per game margin stats.

| Adjustment                | Binary game outcome (won/lost, AUROC) | Score differential (MAE) |
|---------------------------|---------------------------------------|--------------------------|
|                           | Training     | Test       | Training | Test       |
| None                      | 0.790 (0.017) | 0.704 (0.028) | 14.76 (0.57) | 17.39 (1.20) |
| SOS+HF                    | 0.810 (0.017) | 0.715 (0.026) | 13.78 (0.58) | 17.12 (1.18) |
| SOS+HF & Efficiency Cmpl. | 0.808 (0.016) | 0.717 (0.033) | 13.83 (0.57) | 17.13 (1.18) |
| SOS+HF & Any Cmpl.        | 0.790 (0.019) | 0.707 (0.026) | 14.58 (0.58) | 17.53 (0.93) |

Note: Training metric obtained from fitting the model to entire season, while test metric is obtained from using only first 9 weeks to predict the rest of the season.

3.2.4. Using adjusted statistics as features for game outcome prediction

In addition to the exploration of the most impactful complementary football features, we compared predictive performances for statistics that were adjusted in various ways. Specifically, we pitched the following four methods against one another: no adjustment (classical averages), adjusting solely for the strength of schedule and home-field factor (SOS+HF), adjusting also for efficiency-based complementary statistics (our main method) or any complementary statistic. The last approach is included for sanity check purposes to demonstrate how reverse causality can hurt predictive performance (see discussion in Sections 3.1.2 and 4 for more details).

Table 3 shows the results of logistic regression for binary game outcome prediction of whether a team won or lost and classic linear regression for modeling the score differential, respectively. As our five predictors, we used offensive and defensive points-per-game margins for both teams, along with the home-field indicator. As prediction quality metrics, we used the area under the curve (AUROC) [13] for the binary game outcome and mean absolute error (MAE) for score differential. To produce training errors, we utilized the full season data for each year in the 2009–2019 span. For test errors, we used the first nine weeks to train the model and subsequently yield predictions for the rest of the season (using other cut-offs, e.g. 8, 10, and 11 weeks, led to similar results).

Assuming prediction error normality, performances are mostly similar across the methods, with all the metrics being within less than two standard deviations of one another. Nonetheless, SOS+HF and our method consistently show better average performances, with higher AUROC and lower MAE values. When adjusting for any complementary football features, rather than only efficiency-based ones, performance was as subpar as for the classical averages, confirming the dangers of reverse causality in the case of using the game totals.

4. Discussion and future work

We have applied several group penalization techniques in combination with natural cubic splines to detect the most consistent features of American college football in terms of the
impact on the complementary side of your team (e.g. your offense affecting your defense, or vice versa). Partially constrained optimization was implemented to also guarantee the adjustment for the strength of the opponent and home-field factor.

We instantly stumbled upon the issue of reverse causality regarding within-game dynamics and how play-calling could be affected by the scoreline. We have decided to go with an intuitive approach of using solely the efficiency-based statistics (e.g. yards per play, points per possession, etc.), which are less subject to being impacted by strategic play choices compared to game totals. Among the findings, we unsurprisingly showed the non-scoring turnovers forced by your defensive unit as the most consistent feature in positively impacting the scoring by your offensive unit. This is reasonable due to such turnovers giving the offensive side an extra possession and the potential for a good starting field position. Hence, the ranking shifts in points or touchdowns scored per game that resulted from this adjustment tended to penalize offensive units complemented by a defensive side that forced many non-scoring turnovers. Conversely, in points or touchdowns allowed, most penalized were defensive units complemented by offensive sides that committed fewer turnovers.

On the other hand, the team’s defensive statistic that affected the total yardage gained by the offensive unit the most was the yards allowed per play. This effect was non-linear, with defenses allowing moderate 3–5 yards per play shown to be less conducive to their complementary offensive units gaining yards, while defenses on the more extreme ends of the spectrum – allowing either many (≥ 6) or few (≤ 2.5) yards per play – tend to provide more opportunities for their offense to gain yards. As thoroughly discussed in Section 3.2.2, this finding is intuitive due to moderate yardage allowance by your defense being indicative of the opposing offense controlling the game clock, preventing your offense from gaining yards by keeping it on the sidelines.

While utilizing solely efficiency-based statistics led to reasonable results, its lack of improvement in game outcome predictions over basic strength-of-schedule and home-field factor adjustment (SOS+HF) – as shown in Section 3.2.4 – was indicative of reverse causality issues still lingering to an extent, hence leaving one in search of alternative approaches. One such approach could be adjusting for endogeneity, which broadly refers to cases of an explanatory variable being correlated with an error term, hence violating one of the least squares modeling assumptions and subsequently biasing the estimates. Reverse causality is a special case of endogeneity, with explanatory variables suffering from it being called ‘endogenous’ due to being determined by factors ‘within the system’ (e.g. within the game situation), as opposed to ‘exogenous,’ which are fully defined by external factors (e.g. weather). The issue of endogeneity has been discussed across several application domains [8,19,23], with instrumental variables (IV) [5] proposed as one of the most ubiquitous remedial approaches. A strong IV is an exogenous variable that strongly correlates with the endogenous explanatory variable while being uncorrelated with the error term. Such variables have been notoriously difficult to find, resulting in cases where weak IVs only exacerbated estimation bias [4]. Moreover, in our specific case, virtually every complementary football feature can be considered endogenous due to being a partial byproduct of within-game dynamics, making it that much harder to find good exogenous instrumental variables, although that could still constitute an avenue for future research.

The more obvious method for addressing reverse causality would be to obtain play-by-play data across all games. With careful consideration and proper domain understanding, knowledge of play sequencing could provide us with further insight into the causal flow
of within-game dynamics. At the very least, it could make clear that return attempts by one’s offense typically follow directly after the score by the opposing offense, avoiding the aforementioned reverse-causal trap. Although accessing and processing such fine-grained play-by-play data is not a trivial task, we fully intend on carrying it out as part of future research, making sure to incorporate it into the analysis for the purpose of alleviating the reverse causality issue, resulting in a more accurate evaluation of complementary football feature effects.

Acknowledgments

The author is grateful to www.sports-reference.com/cfb/ and www.cfbstats.com/ for providing open access comprehensive game-by-game college football data, and Denise Bradford, PhD and Communications Officer for the ASA ‘Statistics in Sports’ section, for proof-reading the article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] J. Albert, J. Bennett, and J.J. Cochran, *Anthology of Statistics in Sports*, SIAM, Philadelphia, 2005.
[2] J.J. Bartko, *The intraclass correlation coefficient as a measure of reliability*, Psychol. Rep. 19 (1966), pp. 3–11.
[3] G. Boshnakov, T. Kharrat, and I.G. McHale, *A bivariate Weibull count model for forecasting association football scores*, Int. J. Forecast. 33 (2017), pp. 458–466.
[4] J. Bound, D.A. Jaeger, and R.M. Baker, *Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak*, J. Amer. Statist. Assoc. 90 (1995), pp. 443–450.
[5] R.J. Bowden and D.A. Turkington, *Instrumental Variables*, Cambridge University Press, Cambridge, 1990. No. 8.
[6] P. Breheny and M.P. Breheny, Package ‘grpreg’, 2021.
[7] M.W. Browne, *Cross-validation methods*, J. Math. Psychol. 44 (2000), pp. 108–132.
[8] G.J. Duncan, K.A. Magnuson, and J. Ludwig, *The endogeneity problem in developmental studies*, Res. Hum. Dev. 1 (2004), pp. 59–80.
[9] J. Edwards and D. Archambault, *The home field advantage*, Sports Games Play Soc. Psychol. Viewpoints (1979), pp. 409–438.
[10] B. Efron, *Missing data, imputation, and the bootstrap*, J. Amer. Statist. Assoc. 89 (1994), pp. 463–475.
[11] L.O. Gavião, A.P. Sant’Anna, G.B. Alves Lima, and P.A. de Almada Garcia, *Evaluation of soccer players under the moneyball concept*, J. Sports. Sci. 38 (2020), pp. 1221–1247.
[12] A.Y. Govan, A.N. Langville, and C.D. Meyer, *Offense-defense approach to ranking team sports*, J. Quant. Anal. Sports. 5 (2009).
[13] J.A. Hanley and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (roc) curve*, Radiology 143 (1982), pp. 29–36.
[14] J. Harrison, *RSEelenium: R Bindings for ‘Selenium WebDriver’*, 2020. R package version 1.7.7.
[15] D. Harville, *The use of linear-model methodology to rate high school or college football teams*, J. Amer. Statist. Assoc. 72 (1977), pp. 278–289.
[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2 ed., Springer, New York, NY, 2009.
[17] J. Huang, P. Breheny, and S. Ma, *A selective review of group selection in high-dimensional models*, Stat. Sci. Rev. J. Inst. Math. Stat. 27 (2012), pp. 481–499.
[18] D. Karlis and I. Ntzoufras, *Analysis of sports data by using bivariate poisson models*, J. R. Stat. Soc. Ser. D. Stat. 52 (2003), pp. 381–393.

[19] J. Kmenta, *Elements of Econometrics*, 2d ed., The Macmillan Company, New York City, 1986.

[20] M. Lewis, *Moneyball: the art of winning an unfair game*, WW Norton & Company, New York, 2004.

[21] A.A. Neath and J.E. Cavanaugh, *The Bayesian information criterion: Background, derivation, and applications*, Wiley Interdiscip. Rev. Comput. Stat. 4 (2012), pp. 199–203.

[22] R Core Team, *R, A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Version 4.0.3, 2020.

[23] O.J. Rutz and G.F. Watson, *Endogeneity and marketing strategy research: an overview*, J. Acad. Mark. Sci. 47 (2019), pp. 479–498.

[24] R.C. Vergin and J.J. Sosik, *No place like home: An examination of the home field advantage in gambling strategies in nfl football*, J. Econ. Bus. 51 (1999), pp. 21–31.

[25] H. Wang and C. Leng, *A note on adaptive group lasso*, Comput. Stat. Data. Anal. 52 (2008), pp. 5277–5286.

[26] H. Wickham, *rvest: Easily Harvest (Scrape) Web Pages*, 2019. R package version 0.3.5.

[27] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 68 (2006), pp. 49–67.