Spring 4-29-2024

An Investigation into the Causes of Home Field Advantage in Professional Soccer

Paige E. Tomer
Macalester College, petomer@icloud.com

An Investigation into the Causes of Home Field Advantage in Professional Soccer

Paige Tomer

Vittorio Addona, Advisor

Leslie Myint, Reader

Kelsey Grinde, Reader

Department of Mathematics, Statistics, and Computer Science

May, 2024

Abstract

Home-field advantage is the sporting phenomenon in which the home team outperforms the away team. Despite its widespread occurrence across sports, the underlying reasons for home-field advantage remain uncertain. In this paper, we employ a range of statistical methods to explore the causal relationships of potential determinants of home-field advantage. We measure home-field advantage using match outcomes and differential metrics (e.g., differences in yellow cards received). In an attempt to narrow the research disparity between men's and women's sports, we utilize data from the National Women's Soccer League (NWSL) and the English Premier League (EPL) to investigate potential causes of home-field advantage.

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor, Victor Addona, for his help and guidance throughout this project. His assistance and expertise were not only invaluable during this experience but have also been foundational to my entire academic journey at Macalester. Additionally, I want to thank Leslie Myint and Kelsey Grinde for serving as my readers and lending their advice and support to my paper. Moreover, I want to thank all of my math and statistics professors and friends at Macalester, who make the department wonderful.

And, of course, thank you to my friends, roommates, and family — who have supported me from the start and who are likely fed up with hearing about my project, but would never say it to me. Thank you to my friends, who have brought me so much joy and given me so much encouragement throughout this process and over the last four years. To my brother who read my drafts; to my mom, who is always proud of me; and to my grandma, who listened to me talk about this paper and remembered the little details. And, lastly, thank you to my dad, who took days off work to bring me to women's soccer games when I was little and who sat with me while I cried over math homework in high school. I could not have completed this project without them and their support over the past years – thank you!

Contents

1. Introduction

1.1 Background

The concept of home-field advantage describes the sporting phenomenon in which the home team outperforms the away team, and is is visible across a variety of sports. In the National Basketball Association (NBA), for example, the home team wins about 60% of the time and averages 3.5 more points per game (Jones, 2007). In the National Football League (NFL), during the 2008-2009 season, 73% of teams had a higher win percentage when playing at home (Wang et al., 2011). In the Turkish Premier League (soccer), 61.5% of points gained are won by the home team (Seçkin & Pollard, 2008), while in the Australian League, the home teams score about 58% of total recorded points (Goumas, 2014).

Due to its consistency across various sports, extensive research has attempted to isolate the causes of home advantage, yet the underlying factors contributing to this phenomenon remain uncertain. However, numerous explanations have been proposed by scholars and general fan speculation. The most common reasons cited are referee bias, attendance, distance traveled, rule factors, familiarity, and psychological factors. Do referees, for instance, make calls in favor of the home team? Or does a large home crowd urge the home team to perform better or discourage the away team who then performs worse? Additionally, visiting teams may be fatigued and consequently under-perform after traveling long distances and staying in a hotel. Familiarity with the field may prompt a home team to play better: in the NBA, for example, minuscule changes in the ball's trajectory can affect whether a shot is made, and home players may feel more comfortable (familiar) shooting on their home court into hoops and against visual backdrops on which they practice with daily. Rule factors are also hypothesized to contribute to home advantage in sports that have rules benefiting the home team. For example, in the National Hockey League (NHL), the home team gets to choose their lineup after the away team has declared theirs (the so-called "last change privilege").

Despite the numerous possible explanations, this paper concentrates solely on the initial three factors. We investigate whether:

1. referees exhibit bias in favor of the home team,

2. a vocal and supportive home crowd influences team performance,
3. distance traveled by the away team is associated with under-performance.

Though home advantage exists in most (if not all) sports, the remainder of this paper will focus only on professional soccer. This is due to a personal connection to the sport, along with an interest in addressing a data analysis gap between men's and women's soccer. In recent years, we have seen a significant increase in support for women's leagues, with funding and viewership on the rise. For example, the final two teams in the 2023 Women's World Cup (England and Spain) had increased investment in their women's youth leagues in the decade leading up to the tournament. This funding has dramatically improved the quality and quantity of female soccer players from these countries, as evidenced by their World Cup performances (Smith, 2023).

This surge in funding for women's leagues is not limited to specific regions, but is rather an international trend, attracting global attention which is evidenced by escalating viewership trends. Articles from noteworthy publications, such as Forbes' "The Women's World Cup was TV's Most-Watched Show Amid Record-Breaking Viewership," underscore the heightened interest in women's soccer, citing unprecedented viewership figures (Roeloffs, 2023). While women's soccer continues to gain traction with its fan base, financial backing, and overall quality, data analysis has yet to catch up with these upward trends. Despite its growth, little to no data analysis considers women's leagues either on their own or in conjunction with men's leagues. This neglect of women's sports in data analysis is a sexist tradition that utterly fails to recognize the contributions made by women's soccer both domestically and internationally. This paper intends to build a blueprint for examining home advantage in the National Women's Soccer League (NWSL) and to contribute to filling the data analysis gap.

However, due to sample size issues, we had to broaden the scope of our analysis. We first used NWSL data and attempted to isolate the causes of home advantage. Then, we proceeded to do a similar analysis on a men's soccer league, or the English Premier League (EPL) - for which a plethora of data is available extending back many more years - to compare and contrast conclusions.

The paper begins in Section 1.2, where we summarize previous literature surrounding home advantage in soccer. This literature informed our methods and outcome measures, which are discussed in Section 2.1, along with background on the data we use. Section 1.2 and Section 2.2 inform and set the foundation for our models and analysis on the aforementioned three factors (referee bias, attendance, and distance traveled). The results are presented in Chapter 3. We discuss our findings in Chapter 4, before providing some concluding remarks, limitations, and potential future directions in Chapter 5.

1.2 Literature Review

Any successful research project begins by learning from and building upon previous explorations on the given topic. We first considered several overview papers to select the potential causes of home advantage we were interested in. These papers provided a comprehensive overview of the relevant literature and highlighted the key causes of home advantage that may exist in soccer, including familiarity, crowd size/density, travel fatigue, and referee bias (Nevill & Holder, 1999; Pollard & Pollard, 2005).

For each factor we chose to analyze, we reviewed numerous papers examining its contribution to home advantage in soccer. For instance, Benz and Lopez (2020) employed COVID-19 as a natural instrumental variable to investigate the effect of crowd size on home advantage across European leagues. Their bivariate Poisson models showed that goal difference was significant pre-COVID, but post-COVID, goal differences were not significant to the same extent, suggesting that crowd size may contribute to the existence of home advantage.

Benz and Lopez (2020) are also part of a broader collection of papers that consider team quality as a potential confounding variable in their modeling. They simulated plausible ranges for team strength estimates. Then, based on a bivariate normal distribution, Benz and Lopez simulate estimates across seasons and leagues to derive a range of plausible correlation values between team strength and its effect on home advantage. Finally, the authors include this estimated correlation in their model of home advantage.

Boyko et al. (2007) are also among the papers that quantified and controlled team ability. These authors did so by calculating four measurements. For each match in their dataset, they calculated the expected home goals for, expected home goals against, expected away goals for, and expected away goals against. These measurements were calculated by finding the mean of the goals for/against the team in question (home/away) after excluding the current match outcome from consideration and filtering by season. Our research follows this method for controlling for team ability.

Boyko et al.'s primary objective was to examine referee bias in the English Premier League by focusing on individual referees. The authors used outcomes such as goal differential, yellow cards, and foul differential. Boyko et al. saw significant effects of crowd size, referee, and team ability on the outcome in their goal difference models. However, to ensure this significance was not due to outliers (i.e. matches with very high goal differentials), the authors reran their model on a 'truncated' dataset, or a dataset that only included matches with goal differentials ± 3 or less. The results in this truncated dataset were extremely similar to the full dataset – with significance remaining for all variables. Boyko et al. also found referees gave significantly different numbers of yellow cards, red cards, and penalties per game – with the home team receiving fewer cards and penalties – suggesting that referees may contribute to home advantage.

Distance traveled is an additional factor often considered in the literature (Nevill & Holder, 1999; Pollard & Pollard, 2005). Pollard and Da Silva (2008) investigated the impact of distance traveled in the Brazilian football league over the 2003 to 2007 seasons. These authors accounted for team quality by grouping teams from various regions, assuming the quality of these regions would be fairly similar. Then, they conducted a multiple linear regression with goal difference as the outcome metric, which revealed significant results. They found that the home teams are expected to score .115 more goals for every 1,000km the away team travels. Also, by using an ordinal logistic regression – a common method employed in most literature we read– Pollard and Da Silva found the effect of distance traveled was significant in terms of match outcome, with the home team's odds of an advantageous outcome increasing as distance traveled by the away team increases.

Broadly, ordinal logistic regressions were used to model match result, as an outcome while multiple linear regressions modeled score, card, and foul differentials throughout the literature (Pollard & Da Silva, 2008; Hattum, 2017). Therefore, this paper employs similar modeling techniques.

The papers previously mentioned exhibited considerable variation in their focal points, yet they shared a consistent theme: an exclusive emphasis on men's football. Few papers examined multiple factors simultaneously in their scope. Therefore, this paper intends to fill the data analysis gap in women's soccer, while considering multiple potential causal factors in our analysis.

2. Data and Methods

2.1 Data

We use two datasets to help us answer our research questions. Both were retrieved from *FBRef*, a popular repository for worldwide football data. The National Women's Soccer League (NWSL) dataset contains information from 1,112 matches, ranging from 2013 to 2022. The other dataset has information on 10,874 matches from the English Premier League (EPL), covering the 1993 season through the 2022 season. Key variables in both datasets include goals, season, referee, and yellow cards. Attendance data was also included but was missing for 8,000 of the EPL matches.

Along with the variables included in the dataset, we manufactured a few additional measurements of our own. Among them was our measure of team ability. For each case/row in our data frame, we calculated four measurements that encompassed team ability: home team expected goals for (HExpGF), home team expected goals against (HExpGA), away team expected goals for (AExpGF), and away team expected goals against (AExpGA). These variables are the average of goals scored/conceded for the team in question in that season, but exclude that match from the calculation.

Figure 2.1 depicts a directed acyclic graph (DAG) summarizing what we believe to be plausible causal connections in the sphere of our analysis. Our DAG helped inform the variables we should control for in our models. For instance, there was a possible non-causal pathway from our predictor of interest through attendance. Controlling for attendance would block this causal pathway from an unmeasured variable to an outcome, to ensure that we isolated only the effect of the predictor of interest on the outcome. By controlling for attendance, we blocked this pathway as a precautionary measure, as we were unsure whether there was an unmeasured variable causing referee and attendance. For instance, we hypothesized that match location could potentially impact attendance and referee, yet we were uncertain whether this was true, especially in the EPL, where the matches are not extremely dispersed and budgets are higher.

Additionally, the arrow between team ability and attendance (where higher quality teams likely lead to higher attendance) makes attendance a collider variable. Therefore, by only controlling for attendance in our models, we would be creating a superficial relationship between team ability,

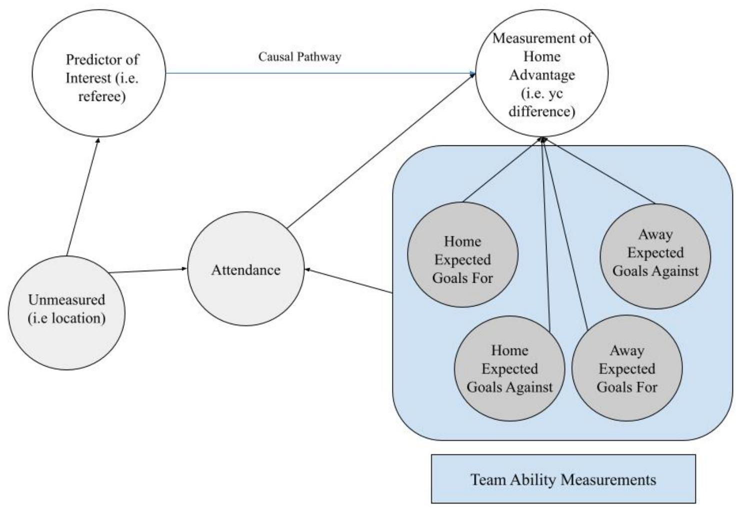

Figure 2.1 Suspected Relationship Between Variables

any unmeasured confounder, and, thus, our predictor of interest. Including team ability in our models mitigated the false inverse relationship that this collider would introduce.

However, in the EPL, we did not have enough data on attendance to include it in our model. And, since referee is a categorical variable, we were unable to conduct a sensitivity analysis. Therefore, in the EPL analysis, we relied on the belief that there is no unmeasured variable causing both referee and attendance and that attendance is not a confounder variable. This is one of the limitations of our paper and will be discussed further in Section 5.

Furthermore, our DAG concludes that team ability is not a direct confounding variable due to the lack of a causal path to our predictor of interest. For instance, we do not believe that the ability of the teams impacts which referee is chosen for a match and how far the away team travels. Yet, we maintained that, in these models, team ability should be included as a precision variable to isolate the effect of referee or distance traveled on the outcome in question. However, team ability is a confounding variable when attendance is considered the predictor of interest. This effect was mitigated by including ability in the attendance models. Therefore, the measures of team ability were included in all of our statistical models.

In addition to team ability, other necessary variables had to be added to our dataset. Firstly, two measures of match result were created. One measure included results with numerical values associated with each possible outcome (0 as loss, 1 as draw, 2 as win). This ordered measure of result was essential to employ ordinal logistic regression models. We also created a binary version of the match result, defined as "win" or "not win" for the home team. Note that, while modeling, we ran into separation issues whenever match attendance was included, either as a control or the primary predictor of interest. In such instances, we used a Firth Bias Reduced Logistic Regression model. Finally, we added several differential variables to our datasets (for example, goal differential and yellow card differential) so that these could serve as outcome measures.

2.2 Methods: Ordinal Logistic Regression and Firth Bias-Reduced Regression

An ordinal logistic regression (also called a proportional odds logistic regression) can be parameterized as:

$$\log \frac{P(Y \leq j)}{P(Y > j)} \;=\; \beta_{j0} - n_1 x_1 - \ldots - n_p x_p \tag{2.1}$$

In this parametrization, Y is the ordered outcome, with J levels. Note that:

$$\frac{P(Y \leq j)}{P(Y > j)}$$

are the odds of Y being at most a category j, so that equation (2.1) is analogous to a traditional "log(odds)" model for a binary outcome. This transformation of odds is a common practice in most statistical software so that the coefficients are more easily interpreted.

Also note that the parametrization in equation (2.1), used by the R Statistical programming language, makes an important stipulation where it multiplies every coefficient, n_p, by -1 instead of using a more familiar parametrization, or:

$$\log \frac{P(Y \leq j)}{P(Y > j)} \;=\; \beta_{j0} + \beta_1 x_1 + \ldots + \beta_p x_p \tag{2.2}$$

This multiplication interprets the model output returned by R more intuitive, as it implies that a *positive* coefficient, n_i, means that larger values of x_i are associated with a *lower* chance of Y being *small* (or, stated alternatively, a *higher* chance of Y being *large*). This allows, for example, for positive model coefficients reported by R to imply a positive association between x_i and Y. We urge the reader to keep this in mind as they examine our model later in this paper.

Fitting an ordinal logistic regression model relies on an assumption known as the *proportional odds assumption* (hence the analogous "proportional odds logistic regression" name). This assumption is that the odds and odds ratio between each category is equivalent, making interpretations of the coefficients possible. Therefore, the intercept is the only variable parameter between categories in the output, while the coefficients remain constant.

However, models, including ordinal logistic regression, can encounter an issue known as *separation*. This occurs when the outcome variable separates a predictor variable perfectly by its values. Separation can make coefficient estimation impossible or lead to unreliable estimation of coefficients, as separation violates the assumption of a unique solution within maximum likelihood estimation. Furthermore, separation can also lead to infinite or N/A odds ratios. This is because, when a predictor perfectly predicts the outcome, there is no variability in certain levels of the outcome, making it so that we are dividing by zero in an odds ratio calculation. Separation can be addressed by introducing a penalization or regularization term that penalizes extreme coefficient estimates (Shen & Gao, 2008; Nusrat & Rahman, 2021).

In our models, attendance (both as a categorical variable and a numerical variable) was often responsible for the separation problem. To fix this issue, we employed a Firth Bias Reduced Logistic Model, which is a regularizing model. A Firth model handles separation by introducing a penalization for probabilities with very high or very low likelihoods within the likelihood function. After introducing a penalization term for extreme coefficients, a parameter estimate is obtained. As with any regularization method, this fix comes with the introduction of some bias, but we were willing to accept such a trade-off to arrive at more realistic model coefficient estimates (Puhr et al., 2016).

In our analysis, ordinal logistic regression, Firth bias reduced logistic regression, binary logistic regression, and multiple linear regression are all used to help identify associations. The associated outcome measures and control variables for each of these methods are presented in Table 2.1, where attendance is a control variable in models unless it is the predictor of interest. As mentioned previously, the EPL data had many missing attendance values in its dataset, which led to us excluding it as a control variable in our EPL models.

Method	***Outcome Measure***	***Control Variables***
Multiple Linear Regression	Goal Difference; Yellow Card Difference	Attendance (NWSL); Team Ability
Ordinal Logistic Regression	Ordered Match Result (loss, tie, win)	Attendance (NWSL); Team Ability
Firth Bias Reduced Logistic Regression	Binary Match Result (win, not-win)	Attendance (NWSL); Team Ability

Table 2.1 Methods and Outcome Measures in Analysis

3. Results

Before exploring the causes of home-field advantage, we first demonstrated its presence using our available data. To do so in the NWSL, we considered game outcomes in terms of the home team (see Figure 6.1). Broadly, we can see that the home team wins far more than they lose or draw. For a more specific and numerical approach, we investigated a goal differential model with no explanatory variables. The model revealed that, on average, the home team scores 0.29 more goals per game than the away team. This was a statistically significant finding with a p-value of approximately 0 (see Table 6.1). Similarly, we confirmed the existence of home advantage in yellow card decisions. A simple linear regression concluded that the away team is expected to obtain .16 *more* yellow cards than the home team ($p \approx 0$). We assert that these findings are evidence of home advantage in the NWSL, which prompts our analysis of the causes behind home advantage.

However, non-rejections of the null hypothesis regularly occurred throughout our NWSL analysis. Of course, this may be because the null hypotheses we were testing are true. Still, we were concerned that it might be because we had insufficient data to yield statistically significant results (in other words, we may have had low statistical power for the magnitude of effects that we were attempting to identify). Because of this, we supported our analyses with data from the English Premier League (EPL), arguably the most famous football league in the world. This league predates the NWSL and possesses more comprehensive match data, offering potentially clearer insights into the causes of home-field advantage.

Because we attempted to explain the causes of home advantage in the EPL, we must establish its existence in this data. Firstly, we see that the home team wins more than they draw or lose (see Figure 6.2). Moreover, a simple linear model of goal differential (see Table 6.2) supports the phenomenon's existence. This EPL model indicates that the home team is expected to score 0.38 more goals than the away team ($p \approx 0$). Additionally, we found (using a similar measure) that the away teams are expected to receive 0.41 more yellow cards than the home team ($p \approx 0$). Therefore, we can conclude that there is a home-field advantage in the EPL and attempt to explain its causes as well.

3.1 Referee Bias

3.1.1 NWSL

Let us first look at referee bias in the NWSL and its effect on home advantage. To reduce noise, the following analysis includes only referees who have officiated ten or more games in the NWSL for our data. After removing officials based on this threshold, 33 referees and 625 matches remained.

Outcome measures we use include result (loss-draw-win [3-level ordinal] and win-not win [binary]), goal difference, and yellow card difference.

Firstly, looking at match results, we used the ordinal measure (ordered loss, draw, win) with an ordinal logistic regression (a.k.a. a proportional odds logistic regression). At the 5% level, eight referees were statistically significant compared to the reference referee. The 95% confidence intervals of the referee model coefficients (see Figure 6.3) show that all significant referee coefficients are in the positive direction – meaning that the home team has higher odds of an advantageous outcome when these referees are in charge of a match, relative to the reference referee. Despite the significance of a few referees, an analysis-of-variance (ANOVA) test yielded a p-value of 0.89 with 33 degrees of freedom, concluding that the referee variable as a whole is not statistically significant.

We used the binary measure of the match result and the Firth Bias Reduced Logistic Regression model to combat separation introduced by the attendance variable. From this model, ten referees were statistically significant in the positive direction, indicating that the home team had higher odds of a positive outcome compared to the reference referee. With a p-value of 0.72 and 33 degrees of freedom, however, an ANOVA comparing this model to its referee-excluding counterpart echoed the previous conclusion: despite the significance of some referees, the overall referee variable is insignificant, which does not justify its inclusion in our model.

In our differential outcome measures, we considered our models in two ways. To begin, we used a measure of referee bias where significance is determined relative to a reference referee. This method is deemed a "conservative" approach because a lack of statistical significance in such models would not rule out a referee bias as a whole. That is, if all referees were equally biased in favor of the home team, such models would not identify any statistically significant results. But, if the referee variable in such a "conservative" approach was significant, this would be strong evidence that referees are distributing yellow cards differently (possibly due to bias).

This is because, for a confounder to influence our results, there would have to be a significant difference between individual referees. For the remainder of this paper, we frequently use the term 'conservative measure/model' to refer to this method of testing differences compared to a reference referee. The counters to this conservative measure are discussed further in this section.

In a goal difference (defined as home goals scored minus away goals scored) model, we used a multiple linear regression while controlling attendance and team ability. Table 6.3 presents a few versions of the model, where the model (3) displays estimated coefficients for all explanatory variables (excluding the referee variable). Figure 6.4 shows the model with the referee variable included, revealing that only two referees were statistically different from the reference referee. We were concerned, however, that this lack of significance was due to low statistical power. Therefore, we were not confident in this non-rejection.

A similar conclusion is reached for our last outcome measure of yellow card difference (defined home yellow cards minus away yellow cards). In this multiple linear model, only one referee was statistically different from the reference referee (see Figure 6.5). Yet, this significant referee does not warrant the whole coefficient as a worthwhile predictor of yellow card difference, with an ANOVA table revealing a p-value of 0.815.

As mentioned earlier, comparing referees to a reference referee is a conservative approach – by only testing discrepancies between referees, we may not capture referee bias that manifests equally across referees. We thus considered two alternative methods. First, we ran a mixed effects model allowing for a random effect of referees. This model still checks whether referees vary in their association with the outcome measure. Yet, compared to our fixed effects linear model, a mixed effects model lets referees enter the model as a random effect so that each referee will *not* have their own coefficient. Rather, we assume that the referee effects are all drawn from a single distribution, with a mean equal to 0 and some variance (σ^2). By evaluating the magnitude of σ^2, we assess the variability between referees without estimating many model coefficients. We hoped to circumvent some of our statistical power concerns by only calculating a single variance. Unfortunately, this approach also left us unable to conclude that referees varied in their association with the outcome measure.

Secondly, to consider an approach that may be considered less conservative, we reran the linear models without a reference referee (therefore, no intercept). Instead of testing differences from each other, this approach tests whether individual referees are associated with the outcome measures (that is, whether each referee has a coefficient different from 0). Although it makes it easier to establish an association between referees and differential measures, this procedure may leave doubt as to whether these associations (if we can establish them) are causal. In the context of this paper, we will first implement this approach and then subsequently discuss the issue of correlation versus causation.

The coefficients of the goal difference model with no reference referee are seen in Figure 6.6. This "non-conservative" approach found no referee coefficient statistically significantly different from 0. The analogous model for yellow card difference found only one referee with a coefficient that was statistically significantly different from 0.

In both models without a reference referee, we do not have evidence to reject our null hypothesis – that referees do not impact yellow card difference. However, by not testing the differences between referees, the possibility of an omitted confounding variable influencing the results arises. With attendance and team ability accounted for, we assert that team psychology remains the most important uncontrolled variable. Consequently, the question arises: How strong would team psychology have to be to impact our results?

Though outside of the realm of this paper, this question is discussed in further detail in Chapter 5, calling for a more standardized measure of team psychology to be able to consider it as a potential confounder.

3.1.2 EPL

Because of consistent non-rejections with potentially low power in the NWSL data, we ran analogous models using the English Premier League (EPL) dataset. Since the EPL contains far more data, we filtered out officials with less than 50 games and were still left with a robust dataset of 9,874 matches, encompassing 56 referees. However, the data only contains attendance information on approximately 2,600 matches. Therefore, including attendance as a confounder in our analyses would remove many data points. Because of this, we chose to exclude attendance in our models and assume that the missingness of the attendance data is "at random".

We followed the procedures outlined in Section 3.1.1 to analyze the EPL data. Starting with the ordinal logistic regression, using the ordered result as the outcome, three referees significantly differed from the reference referee (see Figure 6.8). Yet, an ANOVA test yields a p-value of 0.49 with 55 degrees of freedom, despite the significance of a few referees.

Using the win-not win measure of result, we employed a binary logistic model (since attendance is not included, separation was not an issue and therefore does not have to be combated by a Firth model). We found 3 statistically significant referees, all of which had a positive coefficient compared to the reference referee. When these 3 referees are in charge of the match, the home team has higher odds of winning than the reference referee. However, an ANOVA test again returns an insignificant result (p = 0.36 with 55 degrees of freedom). This reiterates the conclusion that the significance of a few individual referees does not warrant the inclusion of the referee variable as a whole.

We proceeded to use multiple linear regression on the differential measure outcomes. First, our conservative measure found statistically significant results in the goal differential model. Four referees are statistically significant (see Figure 6.9) and the ANOVA yields a p-value of 0.028 with 55 degrees of freedom, which implies that the referee variable is significant. This result indicates that referees contribute to home advantage through goal differential, and is an important conclusion, especially considering this is our conservative measure.

A more dramatic conclusion arises from the yellow card differential model. 95% confidence intervals of the referee coefficients are depicted in Figure 6.10, where sixteen referees are statistically different from the reference referee. Most of the significant referees are in the negative direction, indicating these referees are expected to give comparatively *more* yellow cards to the away team compared to the reference referee. An ANOVA test gives a p-value significant at the 5% level, with 55 degrees of freedom ($p \approx 0$). Therefore, the referee variable is a worthwhile and significant inclusion in the yellow card differential model.

Following these significant findings using a reference referee, we were interested in how this changed using no reference. In a goal difference model, seven referees were statistically significant in the positive direction (see Figure 6.11) – implying that we expect the home team to score more goals when officiated by these seven referees. An ANOVA supports this conclusion, with a p-value of approximately $.023$. Therefore, the referee variable is a significant predictor in explaining goal differential, providing evidence of a bias in favor of the home team.

In the yellow card differential model that uses no reference referees, the majority of referees are statistically significant (refer to Figure 6.12). Moreover, all but two of the significant referees are significant in the negative direction – meaning these referees give significantly more yellow cards to the away team. An ANOVA supports this notion ($p \approx 0$), concluding that the referee variable is a statistically significant predictor.

Since this is a less conservative procedure, an unmeasured variable (e.g., attendance and team psychology) may be impacting our findings. This possibility introduces a necessary discussion surrounding the attendance variable. We argue that attendance cannot be a confounding variable. The definition of a confounder indicates that it has a direct causal link to the outcome (e.g., yellow cards) and an association with the predictor of interest (in this case, the referee assigned to the match). Though attendance could certainly have some link to yellow cards, we argue that there is no plausible link between attendance and the referee who is assigned to officiate the match. One possible dissenting view would posit that attendance *may* be linked to the referee officiating the match, *if* an underlying cause of <u>both</u> is, say, a high-intensity match (i.e., a match between two very strong clubs or two bitter rivals). Under this scenario, attendance may be higher *and* the league may carefully choose a certain type of referee (e.g., a very experienced one). We still believe that this scenario seems rather unlikely. Outside of this possible case, we argue that there is no reason to believe in a link between attendance and match referee. Attendance may be an effect modifier, altering the magnitude of the effect for different referees, but this question is difficult to answer given our less-than-ideal EPL attendance data. This issue is discussed further in Chapter 5.

Therefore, since we believe attendance is not a confounder, we do not expect it to impact the results of our non-conservative models. This still leaves team psychology as a possible confounding factor. This introduces another realization: that team psychology– if it were a confounder – should only impact the non-conservative measurement. For example, if away teams consistently play more aggressively and receive *more* yellow cards, we would expect to see referees significantly differ from 0 in the negative direction – which is what we see in our non-conservative measure. *However*, we would also assume that referees would be impacted by this more aggressive playing style *similarly*. In other words, we would not expect referees to statistically differ from each other. This contradicts what we see in our conservative measurement.

Despite its drawbacks, significance concerning a reference referee builds confidence that referees contribute to home advantage. This is because the confounder would have to affect referees differently to find significance in the conservative measure due to this variable. In other words, the effect of the confounder would have to impact referees so differently that we find overwhelming statistical significance. We argue that such a variable does not exist. Therefore, we assert that referees within the EPL contribute to home advantage through yellow card difference.

3.2 Attendance

3.2.1 NWSL

Attendance was an additional factor we analyzed as a cause of home advantage. To begin, attendance was categorized into quartiles. This action was taken to mitigate separation concerns and reflects our belief that the impact of increasing attendance is not consistently predictable. Moreover, we categorized attendance to capture the overall effect of crowds on teams; in other words, we did not want to test attendance by every additional person in the crowd (what is the difference of *one* more person in a massive crowd?) but instead tested attendance as broad categories.

After deciding to categorize attendance, the variable was sorted into four groups based on the quartiles of the data. For example, the 0 category (reference) includes matches with 0 to 2,910 people, which constitutes approximately 25% of the NWSL matches in our dataset.

To analyze this categorical measure of attendance, we utilized the familiar techniques from previous sections. Firstly, using the ordinal result as the outcome, we found that attendance was insignificant ($p \approx .69$) while team ability (the only control variable in our model) remained significant. Though insignificant, the coefficients of attendance were positive, which is the direction we would expect (as a positive value indicates greater odds of the home team *not* losing). Additionally, the test statistics increased in magnitude as the attendance category increased, which was an interesting finding and could prompt future research with a broader range of data. We reached similar conclusions in the Firth Bias Reduced model with the binary result measure.

Considering goal differential as the outcome measure, we began with the hypothesis that a larger home crowd might impact the psychological effects of the home team – encouraging goals to be scored. Our model reflects this, reporting positive coefficients increasing in magnitude as attendance increases. For instance, in matches with the highest category of attendance (over 6,354 people), we expect the home team to score .21 more goals compared to matches with under 2,910 people (note that this result is insignificant however, with a $p \approx .18$). In this model, the attendance categories are not significant (see Table 6.6). However, we see that as attendance increases, the p-value decreases. This is an interesting note and may hint towards possible significance once more data is available.

A last outcome measure for attendance is yellow card difference, which could capture the effect of a large crowd on team psychology (with more home fans possibly leading to more aggression) or on referees. The two largest attendance categories are statistically significant ($p \approx .0195, p \approx .0023$ respectively) (see Table 6.7). Furthermore, both coefficients are significant in the negative direction, suggesting that larger crowds are associated with more yellow cards for the away team. For example, in games with over 6,354 people (compared to matches with 2,910 people), we expect the away team to receive .356 more yellow cards on average.

3.2.2 EPL

We analyzed attendance in the EPL using the same methods as our NWSL analysis. However, it is important to note that only approximately 2,600 matches in the EPL dataset provided attendance information, leaving roughly 8,000 matches out of our models. For efficiency, we assume that the missingness of this data is random.

Attendance was again categorized into four groups, ranging from low to high attendance. As stated in Section 3.2.1, categorization addresses separation and represents larger trends in the data (with categories representing a general crowd size rather than individual people). In the EPL data, match attendance ranges from 0 to 90,000 people which is much broader than the NWSL. This increase in the range offers unique insights because of the more dramatic differences between low and high attendance.

Attendance measures were subjected to the same analytical approaches used throughout our study. Starting with the usual ordinal result measure as the outcome, this logistic regression model for match results showed significance for the highest attendance category. The positive coefficient suggests that the home team has higher odds of achieving a favorable outcome when the attendance surpasses 51,792 people compared to games with under 24,968 people. An ANOVA table supports this finding with a p-value of 0.006 on 4 degrees of freedom, indicating a significant association between attendance and the home team's result. The Firth Logistic Regression yielded similar findings.

The goal differential model had a parallel conclusion and highlighted the significance of the highest attendance category. The model output is presented in Table 6.8, which shows that the home team is expected to score .266 more goals in matches with over $\approx$ 51,000 people compared to matches with under $\approx$ 25,000 people. An ANOVA table reveals that attendance is a statistically significant predictor of goal difference (with a p-value of .03 and 4 degrees of freedom).

Similarly, the highest level of attendance yielded significance in our yellow card differential model (see Table 6.9). In games with over 51,792 people compared to games with under 24,968 people, we expect the home team to receive .214 *more* yellow cards. Perhaps, a larger crowd may influence and encourage the home team to adopt a more aggressive playing style. An ANOVA reveals a p-value of 0.01 on 4 degrees of freedom.

3.3 NWSL: Distance Traveled

Distance traveled was a final factor considered in our analysis; due to the lack of long-distance traveling in the English Premier League, we exclusively considered this predictor in the NWSL.

Ordinal and binary logistic models were applied using the respective match result outcome measures. A model for goal differential was also considered. However, we did not consider yellow cards as an outcome measure because we did not believe that distance traveled impacts home advantage through yellow card difference.

The ordinal logistic regression model showed no significant conclusions regarding the impact of distance traveled (in miles) on match outcomes. The results of our model revealed a coefficient of -.00001, meaning that every additional mile the away team travels is associated with the log odds of an advantageous outcome for the home team being lower by .00001. This result is not the "expected" (or intuitive) direction, but, with a test statistic of .13, this was not a significant result.

The binary result model reached a similar insignificant conclusion using the Firth Bias-Reduced Logistic Regression model. This model reveals a distance traveled coefficient of -.000037; therefore, when all other variables are held constant, every additional mile the away team travels is associated with the log odds of the home team winning decreasing by .000037. However, since 0 is included in our confidence interval here [(-0.00022, 0.00014)], we do not have evidence to reject the null hypothesis (or determine that distance traveled does not impact the binary match result).

In the goal difference model, the distance traveled coefficient also does not attain statistical significance, rendering inconclusive results on this relationship (see Table 6.5). This model reports a coefficient of -.0000241, meaning that for every additional mile the away team travels, we expect the away team to score .000024 *more* goals. Despite again being in the unexpected direction, the coefficient has an insignificant p-value ($p \approx .734$) which does not warrant any surprising conclusions.

It is interesting to note that in all the relationships we observed, despite insignificance, the coefficients were not in the expected direction. Yet, by increasing the data range, this result could change. The issue of low power is a large obstacle in our analysis and future work with more data is necessary to draw stronger conclusions about non-rejections in the NWSL.

4. Discussion

Our paper considers multiple causes of home-field advantage in both the National Women's Soccer League (NWSL) and the English Premier League (EPL). Though our original intention was to pursue analysis in only women's soccer, we ran into challenges of low power. With frequent non-rejections and small sample sizes, we were unable to be confident in our conclusions. Therefore, we introduced a more extensive dataset from the EPL which allowed us to draw stronger conclusions and build a blueprint for future analysis as data on women's professional soccer becomes more abundant.

Despite the general inconclusiveness in the NWSL, we did fail to reject the null hypotheses in our analysis of distance traveled on home advantage, where our models of goal difference and match outcome report no significant effects. This finding varies from previous literature with papers such as *Home advantage in football in Brazil: differences between teams and the effects of distance traveled* (Pollard and Da Silva, 2008) finding statistical significance in favor of home advantage. This paper compares regions of the Brazilian national league and discovered that the home team is expected to score 0.115 more goals for every additional 1,000 kilometers the away team travels. Our conclusions in the NWSL differed greatly from Pollard and Da Silva (2008), who did not conclude that distance traveled impacts measures of home advantage. Further research with more data could provide insight as to why the differences in the distance traveled effect exist.

For attendance, we found some evidence of an association with yellow card difference in the NWSL. The two highest attendance categories predict *more* yellow cards for the away team at a 5% significance level. This could be because a large home crowd motivates the away team to be more aggressive in response to the crowd's large opposition to them. Or, a large home crowd could cause the referee to make calls in favor of the home team, causing more yellow cards for the away team. This possible relationship could be analyzed with an interaction term and is discussed in our conclusion. Additionally, we could introduce a measure of team psychology (how offensive/defensive a team is playing or how the players feel prior to a match) to root out the reason behind the link between attendance and yellow cards. This would be an interesting line of research to consider in future analyses.

In the Premier League, we found evidence of an effect of attendance on home advantage in goal difference. In the EPL, the highest level of attendance (over $\approx$ 51,000 people) is associated with the home team scoring 0.26 more goals than the away team, compared to a match with less than approximately 25,000 people in the stands. This could be for a multitude of reasons; for instance, the encouragement from the large home crowd could spur the home team to score, or somehow discourage the away team. This result is also consistent with previous literature; consider Benz and Lopez (2020) who used COVID-19 as an instrumental variable to study the effect of a crowd on multiple soccer leagues. They found that home advantage (through goal difference) was smaller in 11 out of the 17 leagues they considered post-COVID, which is analogous to our conclusion in the EPL with low attendance games.

In addition to attendance, we also analyzed the relationship between referees and home advantage. In this analysis, a central challenge revolved around precisely defining the concept of 'referee bias.' In our models employing a reference referee, a straightforward interpretation of individual referees exists, which reflects the causal question at hand: on average, how different do we expect our outcome variable to be if a specific referee officiated the match instead of the reference referee? Essentially, we are attempting to determine whether variation exists between the referees observed in our dataset. Comparatively, in our models without a reference referee, the causal question is less direct. In these models, it is helpful to consider what question we are truly answering when determining whether individual referee coefficients are different from 0. After considering this issue, we arrive at the following causal formulation: how different, on average, do we expect our outcome variable to be if a specific referee officiated the match instead of a hypothetical referee who has *no* impact on the outcome variable of interest (i.e., a hypothetical referee whose true coefficient value is 0)? While this approach may seem conceptually challenging because it relies on a fictional referee who isn't present in our data, it helps quantify referee bias by testing against a theoretically 'unbiased' referee.

This leads to our most significant finding regarding referee bias in the EPL. We conclude that referees, both as a whole and individually, significantly contribute to home advantage through yellow card decisions. The conservative and non-conservative models both discovered that the home team is expected to receive fewer yellow cards. This significant effect of referees on yellow cards could be due to officials – consciously or unconsciously – making calls that favor the home team. Additionally, a loud home crowd team could pressure an official to make calls to please the overwhelming crowd. As mentioned previously, a possible interaction between attendance and referee decisions should be analyzed in future work.

5. Conclusion

We were prompted to incorporate the English Premier League in our research because we could not make confident conclusions about the National Women's Soccer League. With over 10,000 more matches, we strengthened our conclusions and found significance within some home advantage contributors. For instance, we found that referees impacted home advantage through yellow card distribution and goal differences, both as a whole coefficient and as individual referees. Adding in data on the EPL also allowed us to observe conclusions that could potentially be evident in the NWSL, should we have more data available. For example, with more data, we would be able to see if NWSL referees mirror the same trend as the EPL or whether they stay insignificant. If the latter were the case, we would then be able to ask: why are referees in the EPL biased, whereas NWSL referees seem not to be? Yet, we can only do so when more women's sports data is available.

This is a part of the biggest limitation in our analysis: the issue of low power in both the NWSL data and attendance data in the EPL (as only 2,000 matches contained attendance information). Despite this challenge, we were able to make certain conclusions about attendance in both leagues. Though we were able to find some significance, collecting more data would potentially allow us to find more renowned conclusions.

More data would also allow us to examine a possible interactive model between referees and attendance. As hinted at in Chapter 4, we believe there could be a possible interaction between attendance and referee decisions (i.e. does a larger home crowd mean more yellow cards for the away team?). This refers to arrow B in Figure 5.1, where attendance causally impacts a referee's contribution to a home advantage measure. This relationship could be accounted for with an interaction term, yet only more data on attendance would allow us to make an interaction model with high power.

Furthermore, extended data on attendance would allow us to examine a possible causal relationship between attendance and the referee chosen for the match (arrow A in Figure 5.1), which would make attendance a confounder. In modeling, we considered (and believed) attendance and referee not to have a direct causal relationship, but this may not be the case. For example, with high-attendance games or contentious matches, we would likely see more experienced (and possibly less biased?) referees chosen for these matches.

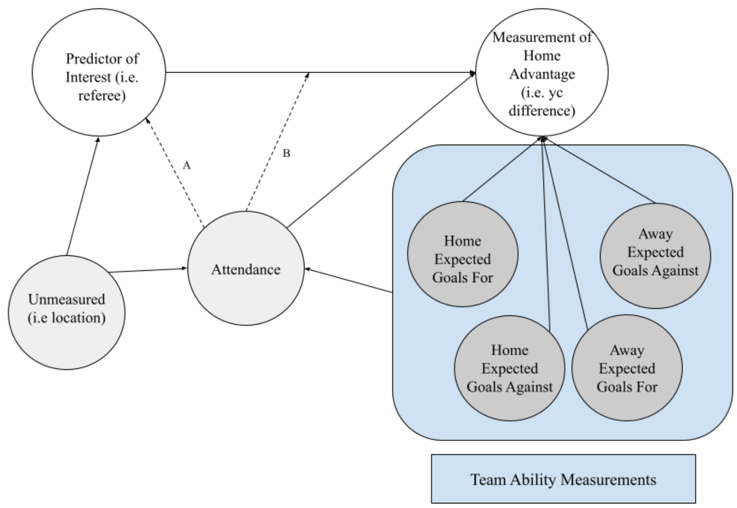

Figure 5.1 Possible Relationships Between Variables

Similarly to collecting data on attendance, collecting data on crowd density is another avenue for future research. This collection would allow us to see if attendance and crowd density differ in their effects on the teams. For instance, does a full, smaller stadium have the same impact as a larger stadium with the same number of people that feels emptier due to its size?

Outside of attendance, creating an accurate measurement for team psychology would be another recommendation for future research. When we tried to analyze team psychology, an issue appeared: there is, at this point, no consistent, standardized way to define team psychology. A measurement that computes how defensive/offensive a team plays in a match is not included in the data or even regularly computed in soccer; further research on developing an adequate measure would allow us to control for this variable in the future and provide the opportunity for interesting explorations. Team psychology is one of the potential unmeasured confounding variables that we considered. Though we do not believe it would impact our conservative approaches, it would be intriguing to see the results of the differential models if, say, away teams tended to play more aggressively. We would also be able to analyze how this measure could affect our non-conservative approach in the referee bias analysis. That is, would incorporating team psychology metrics eliminate the significance of the referee variable?

In summary, the limitations of our paper were the lack of a standardized team psychology measure along with smaller datasets concerning attendance and women's soccer. These limitations then informed our current recommendations for future research. Overall, this paper accomplished its goal: to conduct an investigation of potential causes of home advantage with a spotlight on women's soccer data. We aspired to lay the groundwork for future analyses in women's soccer, whilst also demonstrating the feasibility of data analysis in women's sports, which are worthy of inclusion in comprehensive studies.

6. Tables and Figures

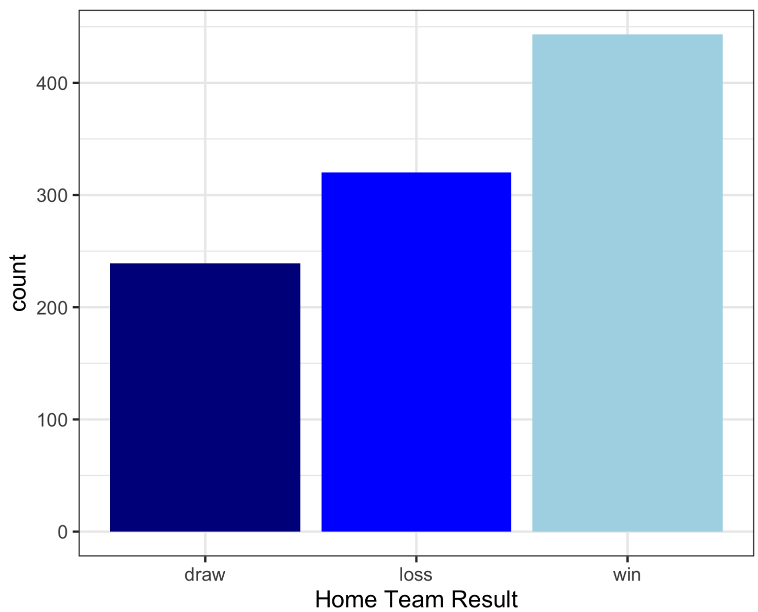

Figure 6.1 NWSL Home Team Results

Table 6.1 **NWSL** Goal Difference: Existence of Home Advantage

	Dependent variable:
	goal_diff
Constant	0.307***
	(0.069)
Observations	625
R^2	0.000
Adjusted R^2	0.000
Residual Std. Error	1.728 (df = 624)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.2 **EPL** Goal Difference: Existence of Home Advantage

	Dependent variable:
	goal_diff
Constant	0.378***
	(0.017)
Observations	10,874
R^2	0.000
Adjusted R^2	0.000
Residual Std. Error	1.777 (df = 10873)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.3 **NWSL** Goal Difference Models

	Dependent variable:		
	goal_diff		
	(1)	(2)	(3)
HExpGF		0.755***	0.729***
		(0.144)	(0.146)
HExpGA		−0.714***	−0.680***
		(0.148)	(0.155)
AExpGF		−0.633***	−0.604***
		(0.144)	(0.147)
AExpGA		0.614***	0.629***
		(0.148)	(0.152)
Attendance: 2,911 - 4,078			0.076
			(0.149)
Attendance: 4,079 - 6,354			0.092
			(0.151)
Attendance: 6,355 - 33,000			0.206
			(0.153)
Constant	0.292***	0.250	0.092
	(0.054)	(0.468)	(0.502)
Observations	1,002	936	906
R^2	0.000	0.135	0.134
Adjusted R^2	0.000	0.131	0.127
Residual Std. Error	1.694 (df = 1001)	1.585 (df = 931)	1.587 (df = 898)
F Statistic		36.341*** (df = 4; 931)	19.859*** (df = 7; 898)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6.4 **NWSL** YC Difference Models

	Dependent variable:		
	yc_diff		
	(1)	(2)	(3)
HExpGF		−0.143	−0.147
		(0.109)	(0.110)
HExpGA		0.252**	0.178
		(0.113)	(0.117)
AExpGF		0.034	0.043
		(0.110)	(0.111)
AExpGA		−0.192*	−0.209*
		(0.113)	(0.115)
Attendance: 2,911 - 4,078			−0.171
			(0.113)
Attendance: 4,079 - 6,354			−0.267**
			(0.114)
Attendance: 6,355 - 33,000			−0.356***
			(0.116)
Constant	−0.163***	−0.111	0.215
	(0.035)	(0.355)	(0.380)
Observations	1,112	936	906
R^2	0.000	0.016	0.028
Adjusted R^2	0.000	0.012	0.021
Residual Std. Error	1.171 (df = 1111)	1.204 (df = 931)	1.201 (df = 898)
F Statistic		3.758*** (df = 4; 931)	3.735*** (df = 7; 898)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6.5 **NWSL** Goal Difference by Distance Traveled

	Dependent variable:
	goal_diff
Distance (miles)	−0.00002
	(0.0001)
HExpGF	0.514***
	(0.112)
HExpGA	−0.560***
	(0.124)
AExpGF	−0.484***
	(0.126)
AExpGA	0.476***
	(0.117)
Attendance: 2,911 - 4,078	0.081
	(0.148)
Attendance: 4,079 - 6,354	0.119
	(0.150)
Attendance: 6,355 - 33,000	0.204
	(0.154)
Constant	−0.004
	(0.374)
Observations	959
R^2	0.093
Adjusted R^2	0.085
Residual Std. Error	1.615 (df = 950)
F Statistic	12.120*** (df = 8; 950)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.6 **NWSL** Goal Difference by Attendance

	Dependent variable:
	goal_diff
HExpGF	0.729***
	(0.146)
HExpGA	−0.680***
	(0.155)
AExpGF	−0.604***
	(0.147)
AExpGA	0.629***
	(0.152)
Attendance: 2,911 - 4,078	0.076
	(0.149)
Attendance: 4,079 - 6,354	0.092
	(0.151)
Attendance: 6,355 - 33,000	0.206
	(0.153)
Constant	0.092
	(0.502)
Observations	906
R^2	0.134
Adjusted R^2	0.127
Residual Std. Error	1.587 (df = 898)
F Statistic	19.859*** (df = 7; 898)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.7 **NWSL** Yellow Card Differential by Attendance

	Dependent variable:
	yc_diff
HExpGF	−0.147
	(0.110)
HExpGA	0.178
	(0.117)
AExpGF	0.043
	(0.111)
AExpGA	-0.209^{*}
	(0.115)
Attendance: 2,911 - 4,078	−0.171
	(0.113)
Attendance: 4,079 - 6,354	-0.267^{**}
	(0.114)
Attendance: 6,355 - 33,000	-0.356^{***}
	(0.116)
Constant	0.215
	(0.380)
Observations	906
R^2	0.028
Adjusted R^2	0.021
Residual Std. Error	1.201 (df = 898)
F Statistic	3.735^{***} (df = 7; 898)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 6.8 **EPL** Goal Difference by Attendance

	Dependent variable:
	goal_diff
HExpGF	0.905***
	(0.111)
HExpGA	−0.660***
	(0.129)
AExpGF	−0.810***
	(0.098)
AExpGA	0.791***
	(0.128)
Attendance: 24,969 - 32,059	−0.028
	(0.091)
Attendance: 32,060 - 51,792	0.040
	(0.096)
Attendance: 51,793 - 84,000	0.266**
	(0.111)
Constant	−0.054
	(0.411)
Observations	2,600
R^2	0.236
Adjusted R^2	0.234
Residual Std. Error	1.627 (df = 2592)
F Statistic	114.488*** (df = 7; 2592)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6.9 **EPL** Yellow Card Difference By Attendance

	Dependent variable:
	yc_diff
HExpGF	-0.280^{**}
	(0.117)
HExpGA	0.449^{***}
	(0.136)
AExpGF	0.281^{***}
	(0.103)
AExpGA	0.012
	(0.135)
Attendance: 24,969 - 32,059	-0.023
	(0.096)
Attendance: 32,060 - 51,792	-0.106
	(0.101)
Attendance: 51,793 - 84,000	0.214^{*}
	(0.116)
Constant	-0.894^{**}
	(0.432)
Observations	2,600
R^2	0.026
Adjusted R^2	0.023
Residual Std. Error	1.708 (df = 2592)
F Statistic	9.722^{***} (df = 7; 2592)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

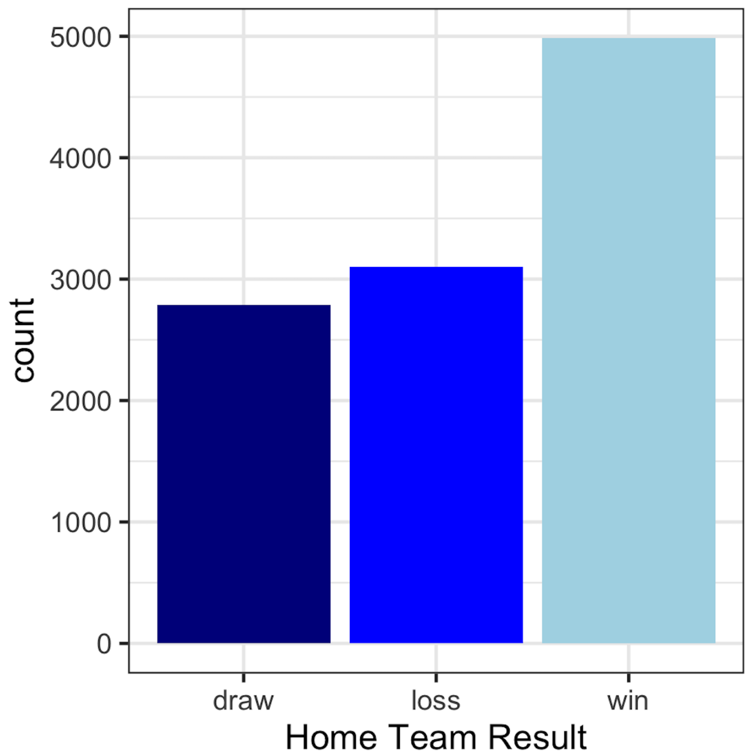

Figure 6.2 EPL Home Team Results

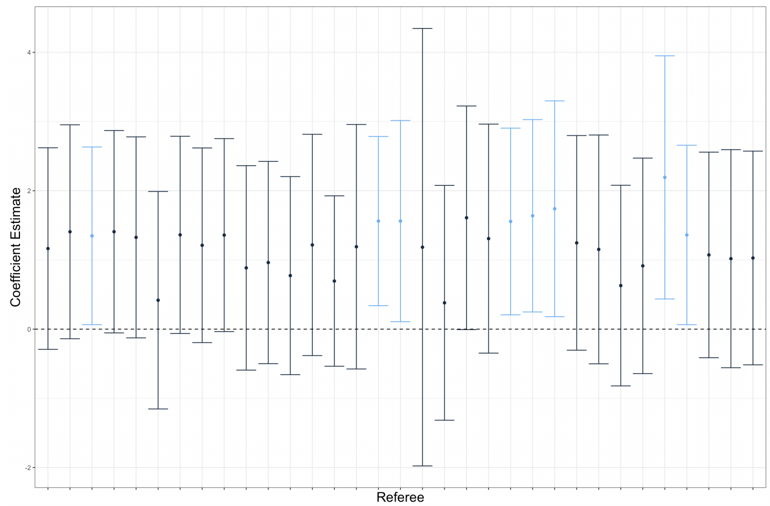

Figure 6.3 NWSL Referee Significance in Ordinal Result Model

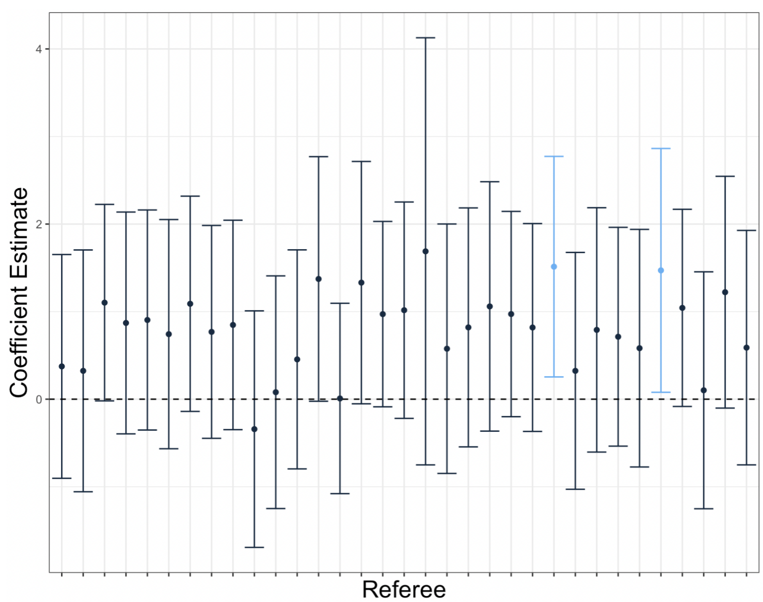

Figure 6.4 NWSL Referee Significance in Goal Difference Multiple Linear Model

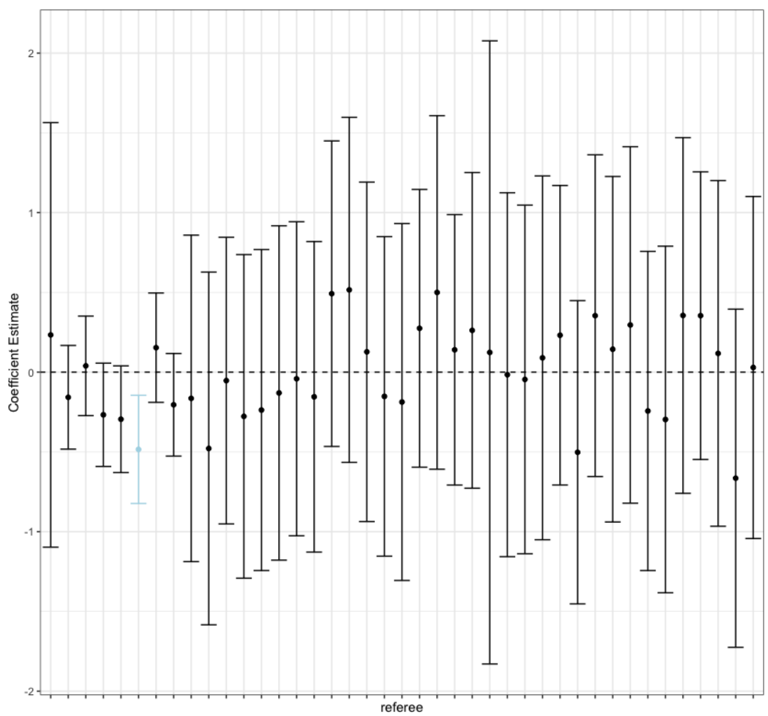

Figure 6.5 NWSL Referee Significance in Yellow Card Linear Model

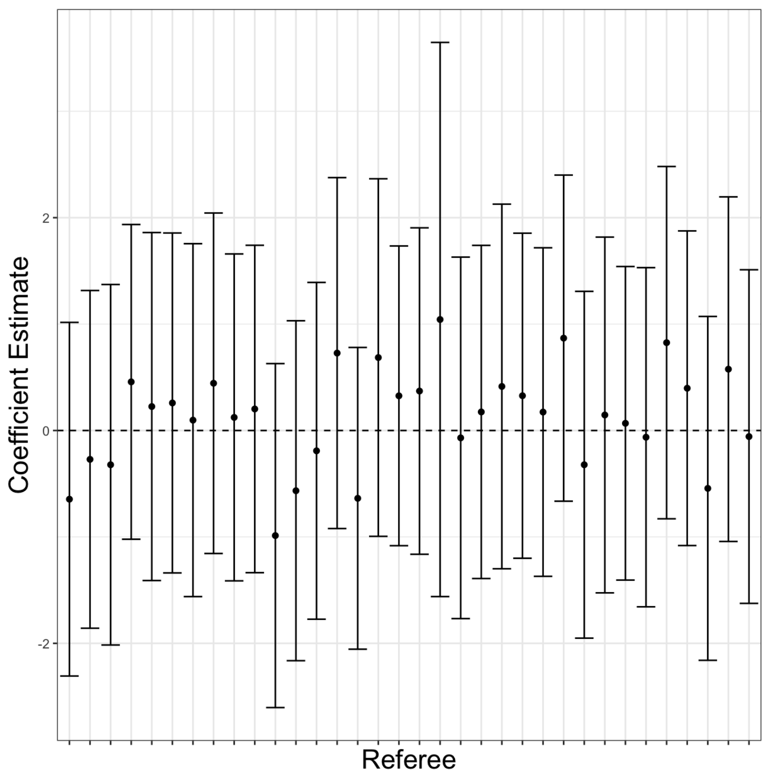

Figure 6.6 NWSL Referee Significance with no Reference Referee in Goal Difference Model

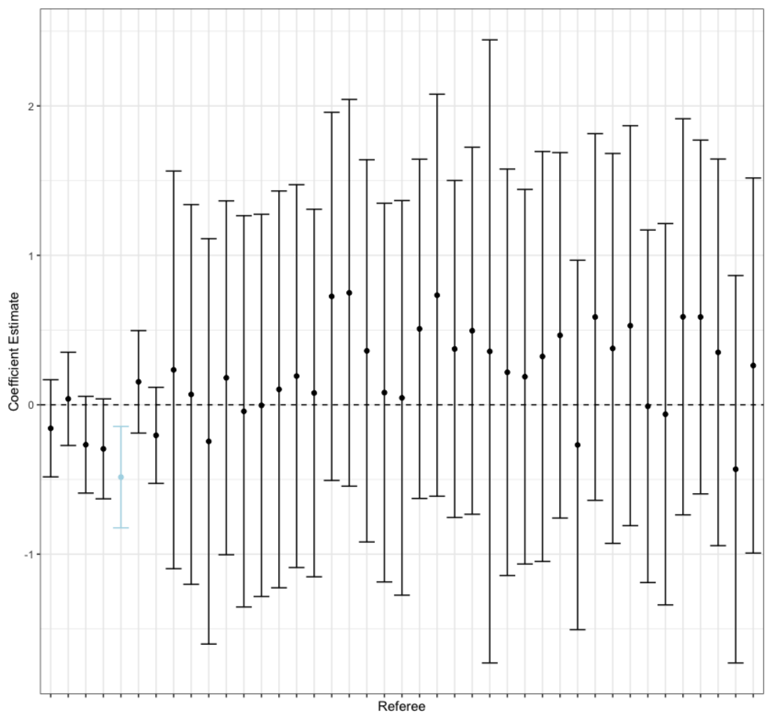

Figure 6.7 NWSL Referee Significance with no Reference Referee in Goal Difference Model

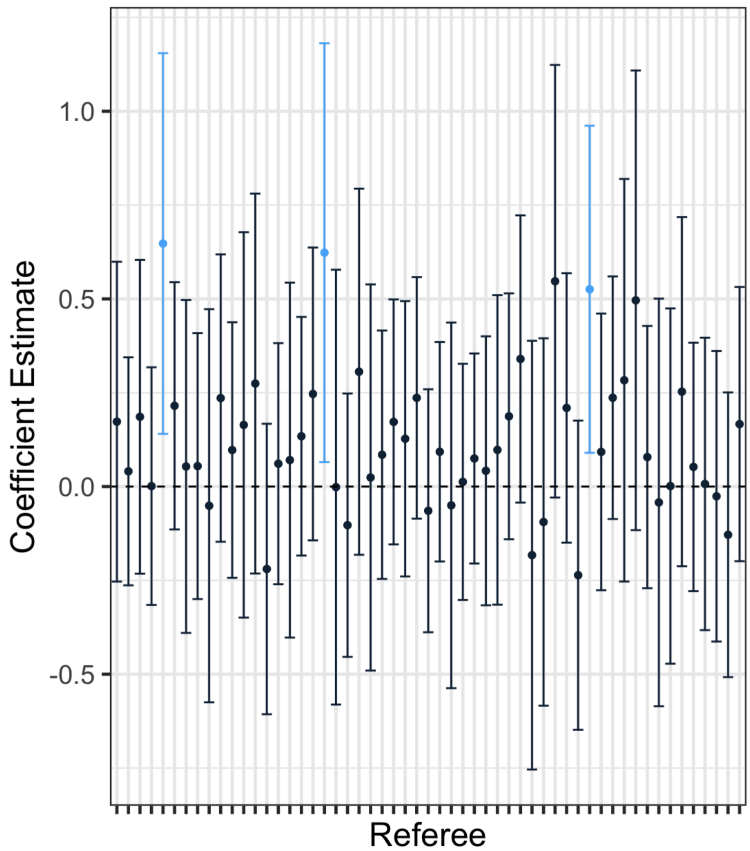

Figure 6.8 EPL Referee Significance in Ordinal Result Model

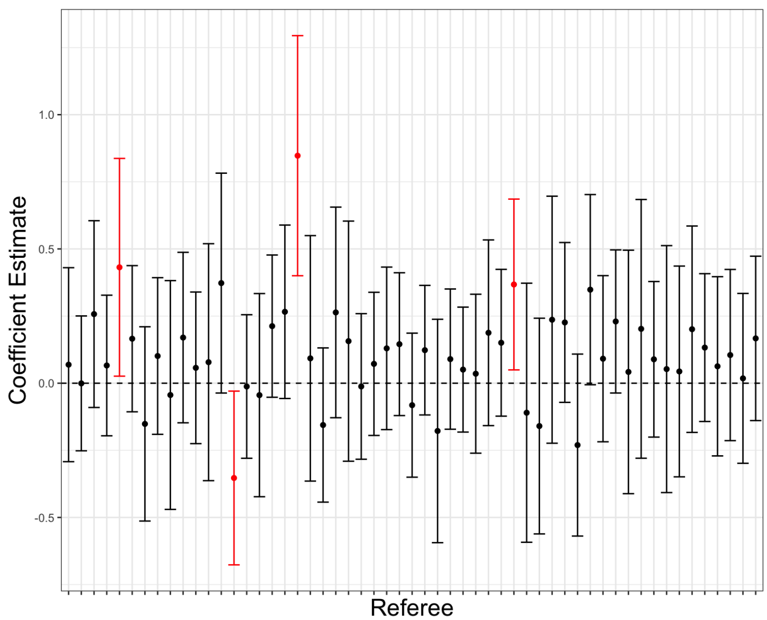

Figure 6.9 EPL Referee Significance in Goal Difference Model

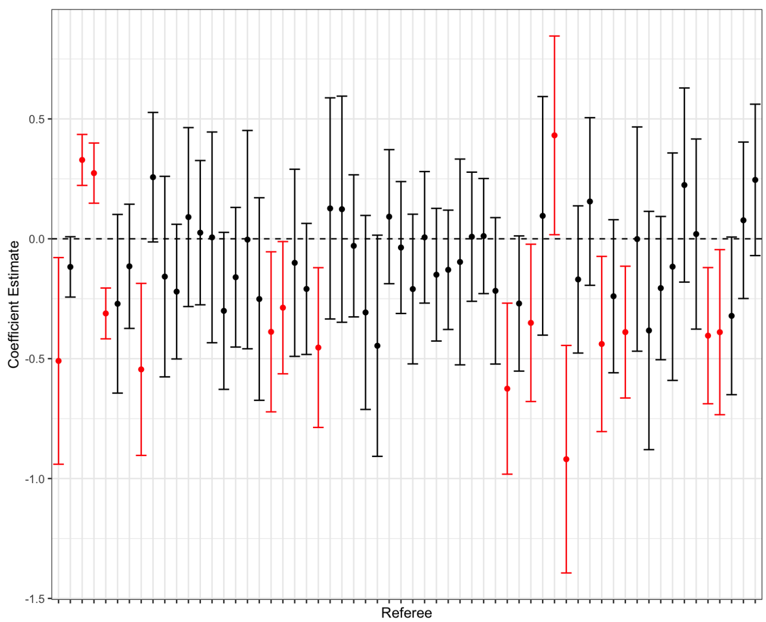

Figure 6.10 EPL Referee Significance in Yellow Card Difference Model

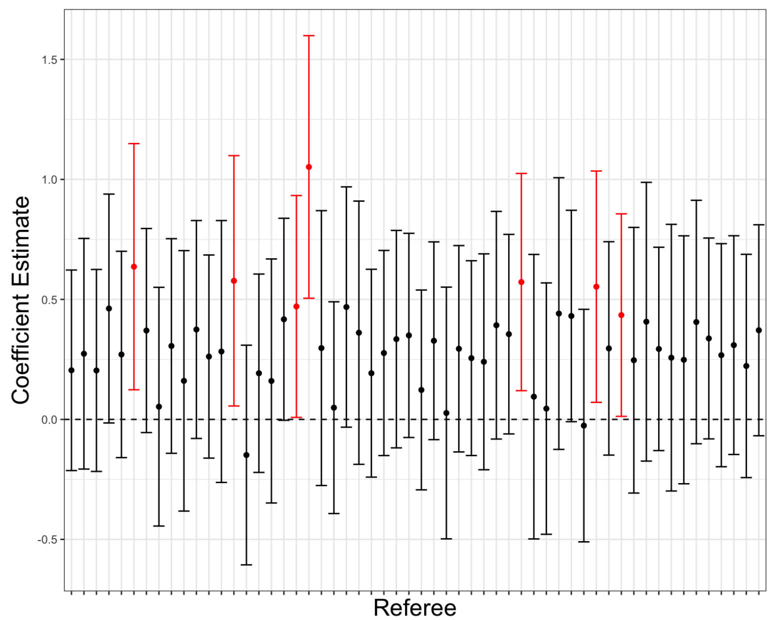

Figure 6.11 EPL Referee Significance in Goal Difference Model, With no Reference Referee

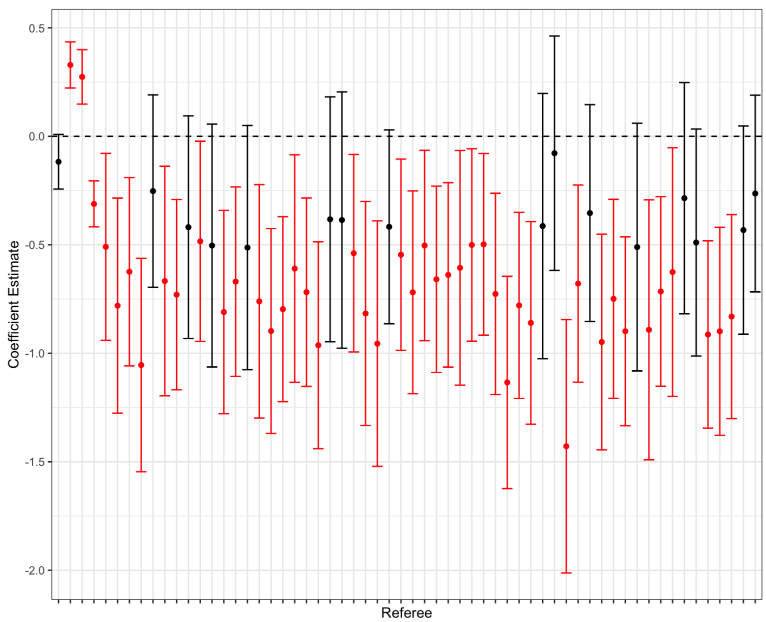

Figure 6.12 EPL Referee Significance in Yellow Card Difference Model, With no Reference Referee

Bibliography

[1] Benz, L. & Lopez, M. (2020). *Estimating the change in soccer's home advantage during the Covid-19 pandemic using bivariate Poisson regression.* AStA Advances in Statistical Analysis, 107(2).

[2] Boyko, A., Boyko, M., & Boyko, R. (2007). *Referee bias contributes to home advantage in English Premiership football.* Journal of Sports Sciences, 25(11), 1185-1194.

[3] Cook, S., Niehaus, J., & Zuhlke, S. (June 2018). *A warning on separation in multinomial logistic models.* Research and Politics, 5(2).

[4] Goumas, C. (2014). *Home advantage in Australian soccer.* Journal of Science and Medicine in Sport, 17(1).

[5] Hattum, W. (2017). *Home Advantage in Football: Does distance play a role?* Erasumus University Thesis Repository.

[6] Henningsen, A., & Toomet, O. (2023). *logistf: Firth's Bias-Reduced Logistic Regression.* R package version 1.26.

[7] Hlavac, M. (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 6.0.5.

[8] Jones, M. (February 2007). *Home advantage in the NBA as a game-long process.* Journal of Quantitative Analysis in Sports, 3(4).

[9] Király, F. (2020). *worldfootballR: FIFA World Football Data R Package.* R package version 0.1.0.

[10] Nevill, A. & Holder, R. (1999). *Home advantage in sport: an overview of studies on the advantage of playing at home.* Sports Medicine, 28, 221-236.

[11] Nusrat, N. & Rahman, M. (2021). *Dealing with separation or near-to-separation in the model for multinomial response with application to childhood health seeking behavior data from a complex survey.* Sports Medicine, 28, 221-236.

[12] Pollard, G. & Pollard, R. (2005) *Home advantage in soccer: a review of its existence and causes.* International Journal of Soccer and Science Journal, 3(1).

[13] Puhr, R., Heinze, G., Nold, M., Lard, L., & Geroldinger, A. (May 2016). *Firth's logistic regression with rare events: accurate effect estimates AND predictions?* Journal of Quantitative Analysis in Sports, 1(1).

[14] Roeloffs, M. (August 2023). *The Women's World Cup was TV's most-watched show amid record-breaking viewership.* Forbes.

[15] Seçkin, A. & Pollard, R. (September 2008). *Home advantage in Turkish professional soccer.* Perceptual and Motor Skills, 107(1).

[16] Shen, J. & Gao, S. (October 2008). *A solution to separation and multicollinearity in multiple logistic regression.* Journal of Data Science, 6(4).

[17] Smith, R. (August 2023). *At the World Cup, the field thins, and the contenders expand.* The New York Times.

[18] *Ordinal logistic regression | R data analysis examples.* University of California Lose Angeles, Statistical Methods and Data Analysis.

[19] Wang, W., Johnston, R., & Jones, K. (September 2011). *Home advantage in American college football games: A multilevel modelling approach.* Journal of Quantitative Analysis in Sports, 7(3).