MATH220 - Final Project Report

Harry Nguyen - Khoi Van - Duc Bui

May 2025

1 Introduction

Bookmakers operate in highly competitive betting markets, and their profitability hinges on setting odds that include a built-in margin—commonly referred to as the overround. This margin ensures that the total implied probabilities across all possible outcomes exceed 100%, allowing the bookmaker to secure a profit regardless of the event's result (Hegarty and Whelan 2024). While this concept is fundamental to sports betting, the strategies behind how these margins are determined are less transparent. Understanding whether bookmakers apply their margins uniformly or strategically vary them based on match characteristics is a central question for both bettors and analysts.

This project aims to investigate how bookmaker margins are structured in European football matches over five seasons, from 2019–20 to 2023–24. Specifically, the analysis addresses several key questions:

- Do margins differ depending on whether a match ends in a home win, draw, or away win?
- Are margins higher when the home team is a favorite compared to when it is an underdog?
- Have margin levels changed over time, and if so, what might explain those shifts?
- How much of the variation in margins can be explained by odds sharpness (i.e., implied probability indices), favorite status, and season-level effects?

To answer these questions, we use a combination of multiple statistical methods including chi-square tests, two-sample t-tests, Welch's t-test, and multiple linear regression models. These methods allow us to test specific hypotheses about the relationship between margin levels and various predictor variables. The data includes odds and outcomes from thousands of matches across multiple seasons, ensuring a robust basis for inference.

The findings have both theoretical and practical significance. For bettors, recognizing when margins are lower can guide more strategic wagering—especially in seasons where competition or regulatory pressure pushes bookmakers to offer tighter odds. For bookmakers and analysts, the insights inform whether

margin-setting strategies should be adjusted based on match context, seasonality, or competitive conditions. More broadly, the study contributes to our understanding of market efficiency and pricing behavior in the sports betting industry.

2 Summary

This report explores how bookmakers set their profit margins in European football betting markets and whether those margins vary based on match characteristics or broader seasonal factors. A key concept in this analysis is the "overround"—a built-in margin in betting odds that ensures bookmakers make a profit regardless of the match outcome. This occurs because the total implied probabilities from the odds add up to more than 100%. While the use of the overround is widely understood, what determines the exact size of this margin remains less clear.

The analysis uses match data from five football seasons (2019–2020 to 2023–2024) to examine whether margin levels differ depending on the result of the match (home win, draw, or away win), whether the home team is the favorite, and how these margins vary across seasons. Statistical methods applied include chi-square tests to assess independence, Welch's t-tests to compare group means, and multiple linear regression to evaluate the influence of implied probabilities, team status, and seasonal effects on margin levels.

The findings suggest that bookmakers apply margins consistently across match outcomes and regardless of which team is favored, meaning bettors are unlikely to find better value by focusing on specific results. However, margin levels do shift significantly from season to season, likely due to changes in competition, regulation, or market behavior. These insights show that broader factors, rather than individual match details, play a more important role in shaping bookmaker strategy. Overall, the project demonstrates how statistical analysis can provide meaningful insights into pricing behavior in the sports betting industry.

3 Dataset

Our analysis utilizes match-level data from the English Premier League (EPL), sourced from the publicly available football-data archive at Football-Data.co.uk. This dataset compiles bookmaker odds and match outcomes across multiple seasons and is widely used in sports analytics and betting market research.

For the purposes of this project, we focused on five full EPL seasons, spanning from 2019–2020 through 2023–2024. Each season's data was provided in a separate CSV file and then combined into a single, clean dataset containing a total of 1,900+ observations. Each row corresponds to an individual match, with variables capturing both market expectations and outcomes.

Our primary variables of interest are the market average odds for each potential match result:

These odds represent aggregated figures across various bookmakers and serve as

Variable	Description
`AvgH`	Market average decimal odds for a home team win
`AvgD`	Market average decimal odds for a draw
`AvgA`	Market average decimal odds for an away team win

Table 1: Variables used in the EPL betting margin analysis

a proxy for market consensus regarding the probability of each outcome. Using these values, we derived implied probabilities and calculated the bookmaker's margin, which is the extent to which the sum of implied probabilities exceeds one—an indication of the bookmaker's expected profit.
We also constructed additional variables, such as the match result (coded as home win, draw, or away win), a margin category (high vs. low, based on the median), and seasonal indicators to allow for comparative analysis over time. Before analysis, we performed data cleaning steps including type conversions, missing value checks, and formatting adjustments to ensure analytical consistency.
This structured dataset provided a strong foundation for our statistical exploration into how bookmaker behavior varies across seasons and match conditions.

4 Statistical Tools

To investigate how bookmakers set profit margins in the English Premier League and whether those margins vary by outcome, season, or odds structure, we employed a suite of statistical methods including hypothesis testing, regression modeling, and exploratory data analysis.

4.1 Chi-Square Tests of Independence

We first conducted Chi-Square Tests of Independence to determine if categorical variables were associated with bookmaker margin behavior. Specifically, we tested whether there was a relationship between margin category (high vs. low) and match outcomes (home win, draw, away win). We also used a chi-square test to evaluate whether the distribution of high-margin games varied significantly across seasons. The second test revealed strong evidence of seasonal effects in margin setting ($\chi^2 = 333.12, p < 2.210^{-16}$).

4.2 t-tests and Welch's two-sample t-test

To compare average margins between different groups, we used t-tests. One test compared margins when the home team was favored versus when it was the underdog. Another compared average margins in the 2019–20 season to those

in 2023–24. In the latter, a Welch's two-sample t-test yielded a statistically significant difference ($p \approx 2.610^{-15}$), indicating that average margins decreased over time.

4.3 Linear Regression

We then developed a series of linear regression models to explain variability in bookmaker margins based on odds structure and match characteristics. Our first model regressed margin on the odds indices for each outcome (home, draw, away), revealing weak but significant predictive power. Subsequent models included additional covariates such as whether the home team was favored and the match season. These expanded models showed improved explanatory power, with the final model achieving an adjusted R of 0.226.

4.4 Residuals vs. Fitted Values Plot and QQ Plot

Finally, we validated our regression assumptions by plotting residuals against fitted values to assess linearity and homoscedasticity, and by examining Q-Q plots for normality. Although some heteroskedasticity was observed, the assumptions were largely met, supporting the reliability of our conclusions.
Together, these statistical tools allowed us to rigorously evaluate how bookmakers adjust margins over time and across different match conditions, providing insights into the underlying structure of the betting market.

5 Analysis

5.1 Association between margin category and match outcome

To begin the analysis, we examined whether bookmaker margin categories—classified as "high" or "low"—are associated with match outcomes, specifically home wins, draws, or away wins. The hypothesis was that if bookmakers systematically set higher margins for matches that are more predictable or one-sided, we would expect to see different distributions of margin categories across different match results.
A chi-square test of independence was performed to assess this relationship. The test yielded a chi-square statistic of $\chi^2 = 2.15$ with 2 degrees of freedom and a p-value of 0.3406. Since the p-value exceeds the standard significance threshold of 0.05, we fail to reject the null hypothesis. This result indicates that there is no statistically significant association between margin category and match outcome.
The bar chart in Figure 1 supports this conclusion. For each match outcome (home win, draw, away win), the distribution of high-margin and low-margin

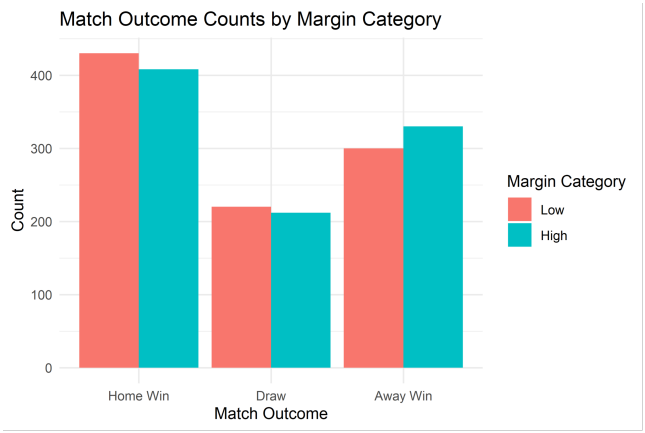


Figure 1: Match Outcome Counts by Margin Category

matches is fairly even. The visual similarity across categories suggests that bookmakers do not adjust margins based on how the match eventually turns out. Instead, this implies that margins are likely determined ahead of time based on general odds-setting models, rather than being tailored to the perceived difficulty or predictability of a particular result.

The results of this analysis highlight how bookmakers apply their margins evenly across different match outcomes. For bettors, this means there is no strategic advantage in focusing on matches expected to end a certain way, such as home wins or away wins, since margin levels are applied consistently across all results. In other words, the bookmaker's profit margin is not skewed to favor or penalize any specific outcome. For bookmakers, the result supports the effectiveness of a flat, standardized margin-setting approach. There appears to be no need to adjust margins based on the expected difficulty or predictability of match outcomes, as doing so does not significantly impact margin performance.

5.2 Season-to-season differences in high-margin matches

While margin-setting appears to be independent of individual match outcomes, it may still vary in broader patterns over time. To explore this possibility, we next examine how margin categories differ across football seasons.

A chi-square test was used to examine the relationship between margin category and season. The result, $\chi^2 = 333.12$, $df = 4$, $p < 2.2 \times 10^{-16}$, provides strong statistical evidence to reject the null hypothesis. This indicates a significant association between season and margin category, meaning that the distribution of high- and low-margin matches is not consistent over time.

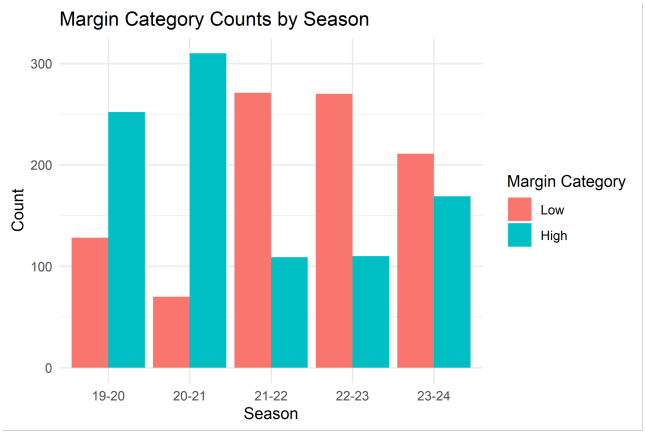


Figure 2: Margin Category Counts by Season

The bar chart in Figure 2 clearly illustrates this trend. For instance, the 2020–21 season saw a sharp increase in high-margin matches, whereas the 2021–22 and 2022–23 seasons show a noticeable skew toward low-margin games. This suggests that bookmakers' margin strategies shift over time, potentially in response to broader market factors such as regulation, competition, and bettor behavior.

Recent market analysis highlights that consolidation among major sportsbook operators, particularly in U.S. states like Colorado and Massachusetts, has reduced market competition and weakened promotional intensity (Birches Health 2024). As competition decreases, dominant sportsbooks may widen margins in less favorable odds and higher overrounds—without fear of losing market share. These structural shifts support our finding that margin-setting varies meaningfully across seasons.

The seasonal variation in margin levels has practical consequences for both bettors and bookmakers. For bettors, seasons like 2021–22 and 2022–23 offered more favorable conditions due to tighter margins, meaning better net odds. Tracking seasonal margin patterns may help identify when the betting environment is more advantageous. For bookmakers, the findings highlight the value of historical season-level analysis in guiding pricing strategies. In more volatile or competitive years, higher margins may be justified, while in more stable periods, tighter overrounds can help attract and retain market share.

5.3 Favorite vs. underdog average margin

To evaluate whether bookmakers systematically vary their margins depending on whether the home team is expected to win, we conducted a two-sample t-test comparing the average margin for matches where the home side was the favorite (AvgH $<$ AvgA) to those where it was the underdog.

The results showed no statistically significant difference in mean margins be-

tween the two groups. Specifically, the average margin when the home team was the favorite was approximately 0.04179, while it was 0.04178 when the home team was the underdog. The test yielded a test statistic of $t = 0.065$ with $p = 0.9483$, and the 95% confidence interval for the difference in means was $[-0.00022,\ 0.00023]$.
Overall, these findings suggest that bookmaker margin-setting is symmetric with respect to perceived team strength. Bookmakers do not inflate margins when the home team is favored, nor do they offer tighter pricing when the home team is the underdog. This uniform approach reinforces the idea that margins are not influenced by match balance but are instead applied uniformly across fixtures to maintain consistent profitability.

5.4 Change in average margin: 2019-20 vs. 2023-24

To determine whether bookmakers have adjusted their overall margin strategy over time, we performed a Welch's two-sample t-test comparing the average margins in the 2019–2020 and 2023–2024 Premier League seasons.
The analysis revealed a statistically significant decline in average margin. In 2019–2020, the mean margin was approximately 0.04272, whereas in 2023–2024, it had fallen to 0.04140. The test yielded a test statistic of $t = 8.08$ with a p-value of approximately 2.6×10^{-15}. The 95% confidence interval for the difference in means was $[0.0010,\ 0.0016]$, indicating that the decline was both statistically and practically significant.
Overall, bookmakers have gradually tightened their margins over the five-season span, reducing their typical profit "cut" by about 0.13 percentage points. This shift may reflect increasing market competition, regulatory pressure, or changing risk management strategies. For bettors, the trend signals a modest improvement in value over time. For bookmakers, it underscores the need to regularly reassess pricing strategies to stay competitive in an evolving market.

5.5 How much odd-indices explain margin

New Variable for Linear Regression: Odd Index Variables

To prepare our odds data for linear regression, we first convert each decimal odd x (AvgH, AvgD, AvgA) into its American odd equivalent (AvgH(Am), AvgD(Am), AvgA(Am)) :

$$\text{Am}(x) = \begin{cases} 100(x-1), x \geq 2 \\ -\frac{100}{x-1}, 1 < x < 2 \end{cases}$$

We then define an unbounded "odds index" (AvgHIndex, AvgDIndex, AvgAIndex):

$$\text{Index}(x) = \begin{cases} Am(x) - 100, Am(x) > 0 \\ Am(x) + 100, Am(x) < 0 \end{cases}$$

This transformation preserves monotonicity, which is higher decimal odds yield higher index and vice versa while mapping the set of only real numbers larger

than 1 to the set of all real numbers, making AvgHIndex, AvgDIndex and AvgAIndex more suitable linear predictors of betting margin.

Back to the linear regression itself:

To discover how odds explain margin, we will create a linear regression for the book margin based on odd indices AvgHINdex, AvgDIndex and AvgAIndex:

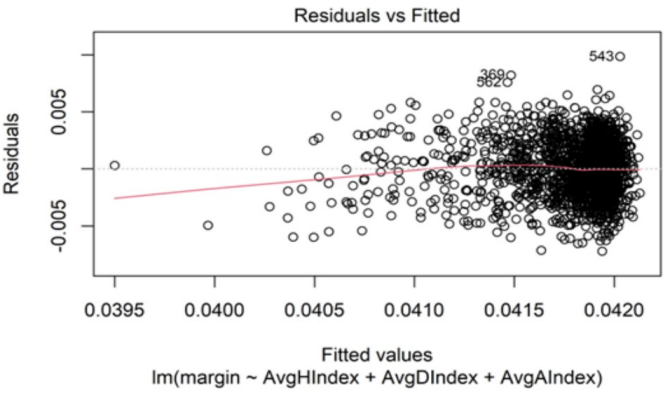
$$\text{margin} = \beta_0 + \beta_1 AvgHIndex + \beta_2 AvgDIndex + \beta_3 AvgAIndex.$$

Here is what we found:

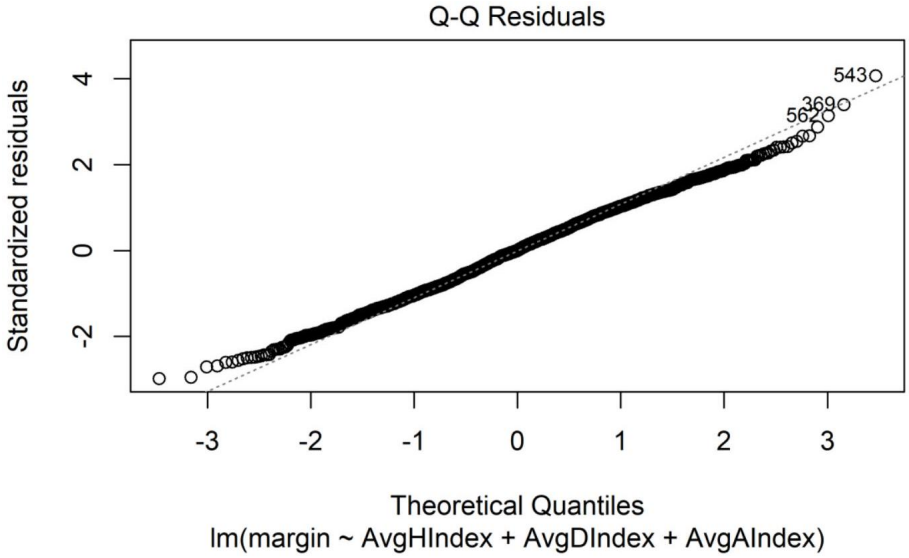
Predictor	Coefficient	p-value
β_0	$4.207 \cdot 10^{-2}$	< 0.001
β_1	$-1.991 \cdot 10^{-6}$	0.0106
β_2	$1.220 \cdot 10^{-6}$	0.3608
β_3	$-1.834 \cdot 10^{-6}$	0.0137

Based on this analysis, we can say that:
Across all matches, bookmakers have an approximately 4.2% margin for a match before accounting for any changes in odds.
Holding everything else constant, a one-unit increase in AvgHIndex is associated with a $1.991 \cdot 10^{-6}$ decrease in margin. The p-value in this case is less than 0.05, suggesting that AvgHIndex even though causes a tiny decrement in margin, it still has significant effect on it.
Holding everything else constant, a one-unit increase in AvgDIndex is associated with a $1.220 \cdot 10^{-6}$ increase in margin. The p-value is larger than 0.05 , suggesting that this variable has no significant effect on the margin.
Holding everything else constant, a one-unit increase in AvgAIndex is associated with a $1.834 \cdot 10^{-6}$ decrease in margin. Although this is a small decrement, the p-value in this case is less than 0.05, suggesting that AvgAIndex also has significant effect on the margin.
The R^2 for this model is 0.01403 , suggesting that this model explains a very marginal amount of variance in book margin.
Here are our diagnostic plots for this linear regression:

First, the residuals vs. fitted plot:



Based on this graph, we can say that the points in this plot do not follow a particular shape, which means that our linear regression has captured the linearity assumption between independent and dependent variables. Moreover, even though there are some outliers, the majority of the residuals have the same variance, which supports our homoscedastic assumption about the linear regression. Second, the $Q-Q$ plot:



Based on this plot, we can say that most of the points line on the 45-degree line, except for slight tail deviations, which suggest that the residuals are close to normal distribution, which satisfies our assumption about normality of residuals.

Key Takeaways:

R^2 is very small, which means that the odd indices all together account for a tiny amount of the match-to-match swings in the margin. Even though the model passes the assumptions for a linear regression and p-values shows that home and away odds cause significant effects to the margin, the changes in the real world are very small. This means that the vast majority of what makes one game's margin slightly higher or lower is either the same across all games or driven by other factors.
This implies that bookmakers set almost the same margin on every match. Tiny changes in the odds barely move the margin, which means that solely looking at the odds does not help to analyze the margin.

5.6 Controlling for odds, does favorite status matter?

To improve our model, we will add another variable to the linear regression in section 5.5, which is isHomeFav. This variable tells whether the home team is treated as a favorite or an underdog in a match:

$$\text{isHomeFav } = \begin{cases} 1 \text{ if } AvgH < AvgA \\ 0 \text{ if } AvgH \geq AvgA \end{cases}$$

Now our new model is:

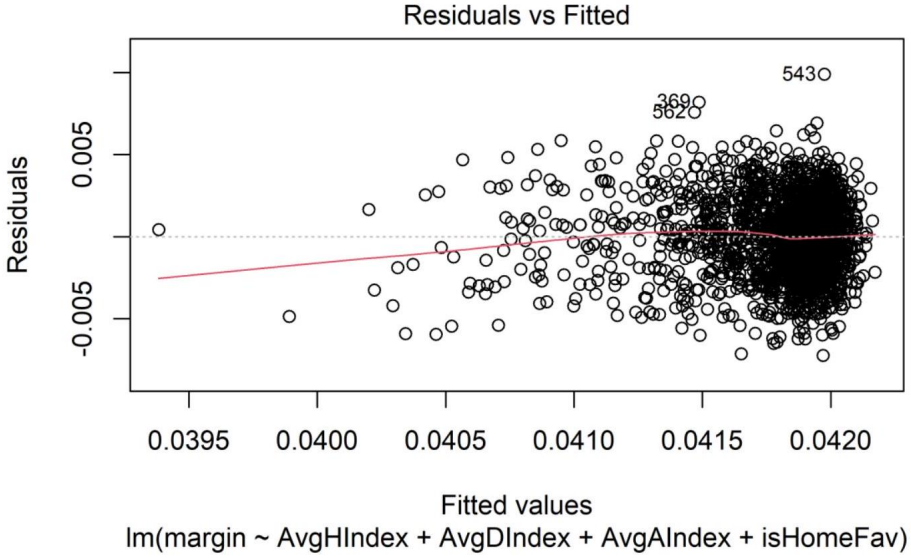
$$\text{margin} = \beta_0 + \beta_1 AvgHIndex + \beta_2 AvgDIndex + \beta_3 AvgAIndex + \beta_4 isHomeFav.$$

Here is what we found:

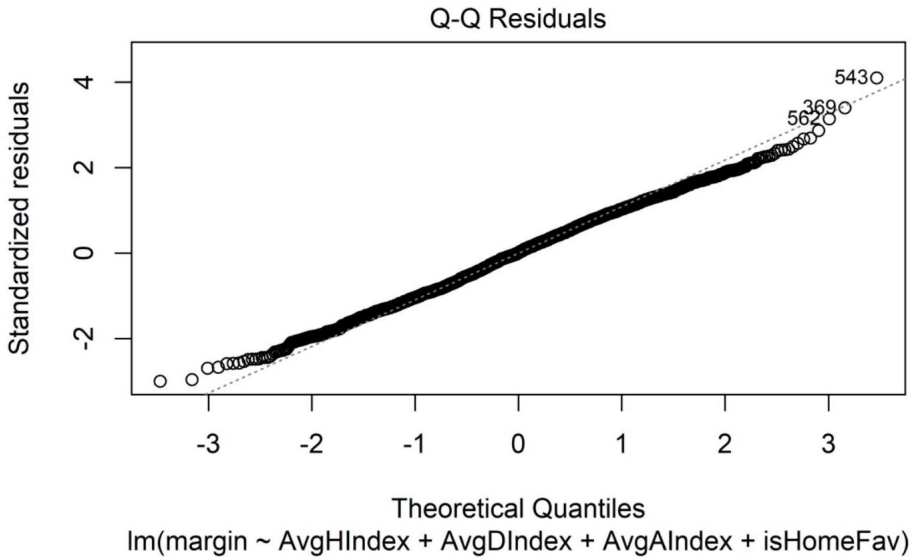
Predictor	Coefficient	p-value
β_0	$4.199 \cdot 10^{-2}$	< 0.001
β_1	$-1.932 \cdot 10^{-6}$	0.0138
β_2	$1.335 \cdot 10^{-6}$	0.3216
β_3	$-1.886 \cdot 10^{-6}$	0.0117
β_4	$9.917 \cdot 10^{-5}$	0.5075

Based on this analysis, we can say that:
Across all matches, bookmakers still have an approximately 4.2% margin for a match before accounting for any changes in odds. This is the same as the first model.
Keeping everything else constant, a one-unit increase in AvgHIndex is associated with a $1.932 \cdot 10^{-6}$ decrease in margin. The p-value in this case is less than 0.05 , suggesting that AvgHIndex even though causes a tiny decrement in margin, it still has significant effect on it.
Keeping everything else constant, a one-unit increase in AvgDIndex is associated with a $1.335 \cdot 10^{-6}$ increase in margin. The p-value is larger than 0.05 , suggesting that this variable has no significant effect on the margin.
Keeping everything else constant, a one-unit increase in AvgAIndex is associated with a $1.886 \cdot 10^{-6}$ decrease in the margin. Although this is a small decrease,

the p-value in this case is less than 0.05 , suggesting that AvgAIndex also has a significant effect on the margin.
Keeping everything else constant, the fact that the home team is considered the favorite is associated with a $9.917 \cdot 10^{-5}$ increase in the margin. The value of p is greater than 0.05, suggesting that this fact does not have a significant effect on the margin.
The R^2 for this model is 0.01426 , which is a very small change from the previous model, suggesting that after adding the isHomeFav, the new model still explains a very marginal amount of variance in the book margin. This is because the added independent variable (isHomeFav) has no significant effect on the dependent variable (the margin). Here are our diagnostic plots for this linear regression: First, the residuals vs. fitted plot:



Based on this graph, we can say that the points in this graph do not follow a particular shape, which means that our linear regression has captured the linearity assumption between independent and dependent variables. Moreover, even though there are some outliers, the majority of the residuals have the same variance, which supports our homoscedastic assumption about the linear regression. Second, the $Q - Q$ plot:



Based on this plot, we can say that most of the points line on the 45-degree line, except for slight tail deviations, which suggest that the residuals are close to normal distribution, which satisfies our assumption about normality of residuals.

Key Takeaways:

Since the new added variable has no significant effect on the margin, after added this variable, it does not improve the predictive power of it. A one-unit rise in AvgHIndex or AvgAIndex still reduces the margin statistically, but the effect is tiny. The high p-value of isHomeFav suggest that bookmakers do not raise the margin of a match just because one of the side is favored.
The R^2 is still very low (0.0143), which means that a huge majority of the variance in margin is still unexplained, and the odd indices and the fact that one team is favored only explain tiny tweaks in the margin.
In short, we can still say that the home or away odds contribute a significant yet very small amount to the margin, while the draw odds or favor status are irrelevant. This means that the margins set by bookmakers based on these factors are still stable. This means that trying to exploit home/away odd differences between matches or favorite/underdog status would not help creating a better margin for bookmakers.

5.7 Season effects on margin after odds

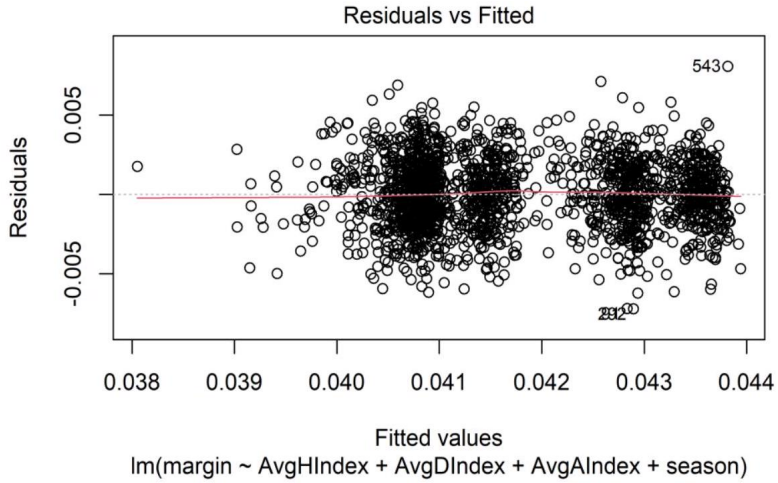
To improve our model, we will add another variable to the linear regression in section 5.5 , which is season.
Now our new model is:

$$\begin{aligned}\text{margin} = \beta_0 + \beta_1\ \text{AvgHIndex}\ + \beta_2\ \text{AvgDIndex}\ + \beta_3\ \text{AvgAIndex}\ + \beta_4\ \text{season2021} \\ + \beta_5\ \text{season2122}\ + \beta_6\ \text{season2223}\ + \beta_7\ \text{season2324}\ .\end{aligned}$$

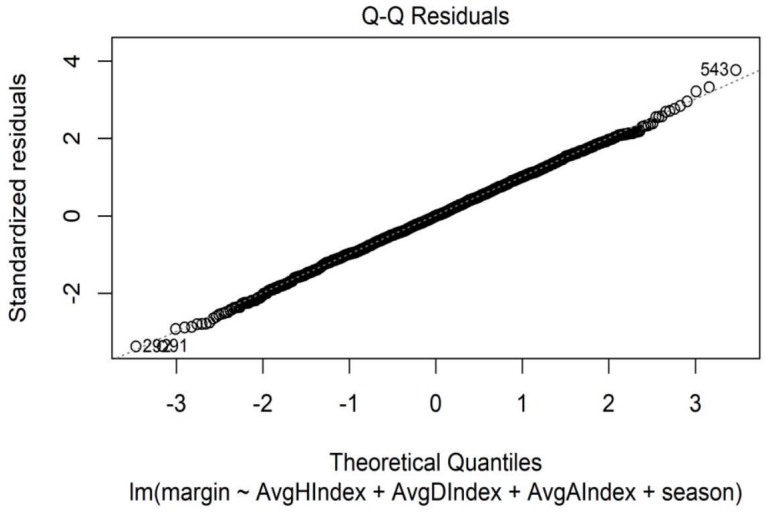
Here is what we found:

Predictor	Coefficient	p-value
β_0	$4.282 \cdot 10^{-2}$	< 0.001
β_1	$-3.470 \cdot 10^{-6}$	< 0.001
β_2	$4.123 \cdot 10^{-6}$	0.00157
β_3	$-3.288 \cdot 10^{-6}$	< 0.001
β_4	$7.436 \cdot 10^{-4}$	< 0.001
β_5	$-1.974 \cdot 10^{-3}$	< 0.001
β_6	$-2.100 \cdot 10^{-3}$	< 0.001
β_7	$-1.500 \cdot 10^{-3}$	< 0.001

Based on this analysis, we can say that:
All the seasons' coefficients have p-values less than 0.005 , and they all reflect big, significant changes to the betting margin. In particular, seasons $2020-21$, $2021-22$, $2022-23$ and $2023-24$ have a margin change of 0.074%, -0.197%, -0.210% and -0.150% compared to the season 2019-20. This means that the seasons matter, and they have a very big impact on the betting margin, which reflects in changes in coefficients of other variables.
Across all matches, now bookmakers have an approximately 4.28% margin for a match before accounting for any changes in odds. This is different from the number 4.2% in the previous models.
Holding everything else constant, a one-unit increase in AvgHIndex is now associated with a $3.470 \cdot 10^{-6}$ decrease in margin. The p-value in this case is less than 0.05 , suggesting that AvgHIndex has significant effect on margin. However, even though the coefficients are completely different from the previous two models, this coefficient still implies that this variable has a very minimal impact on the margin.
Holding everything else constant, a one-unit increase in AvgDIndex is now associated with a $4.123 \cdot 10^{-6}$ increase in margin. The p-value is now less than 0.05 , which means that in this model, unlike the previous models, the draw odd has a significant effect on the margin. However, this effect is still very tiny due to the very low coefficient.
Holding everything else constant, a one-unit increase in AvgAIndex is associated with a $3.288 \cdot 10^{-6}$ decrease in margin. The p-value in this case is less than 0.05 , suggesting that AvgHIndex has significant effect on margin. However, even though the coefficients are completely different from the previous two models, this coefficient still implies that this variable has a very minimal impact on the margin.
The R-squared for this model is 0.2288 , which is a huge leap from the R-squared value of the previous two models. This is because seasons matter significantly to the model, which also explains why it can help explain more variance in the margin.
Here are our diagnostic plots for this linear regression:
First, the residuals vs fitted plot:



Based on this graph, we can say that the points in this plot does not follow a particular shape, which means that our linear regression has captured the linearity assumption between independent and dependent variable. Moreover, even though there are some outliers, the majority of the residuals have the same variance, which supports our homoscedastic assumption about the linear regression. Second, the $Q-Q$ plot:



Based on this plot, we can say that most of the points line on the 45-degree line, except for slight tail deviations, which suggest that the residuals are close to normal distribution, which satisfies our assumption about normality of residuals.

Key Takeaways:

Nearly a quarter of the match-to-match margin variation is explained solely by which season it is, which is far more than by any odd-based measure. This means

that marketwide shifts are the primary drivers of margin. Without accounting for odd changes, according to this model, bookmakers now have a 4.28% margin across the matches, which is an increase reflecting changes in sportsbook making from $2019-20$ to $2023-24$. Compared to 2019-20, 2020-21 has a noticeably high interest in bet making for bookmakers when the margin raised by 0.074%. However, the next three seasons saw progressively tighter margins. Some possible reasons for this can be strategic or external forces, such as COVID. The bettor's edge comes from when they place bets and barely comes from the odds of the teams or whether the home team is a favorite or an underdog. Bookmakers should have season-level pricing strategies to maximize their profit. Since this model only accounts for less than a quarter of the variance in margin, in order to understand more about margin, we should look into other factors such as betting volume, game rule changes or competitor actions.

5.8 Best predictor set: indices + favorite status + season

Now, we will try to combine all of these factors into one linear regression to see how it works out and which ones best predict the margin.

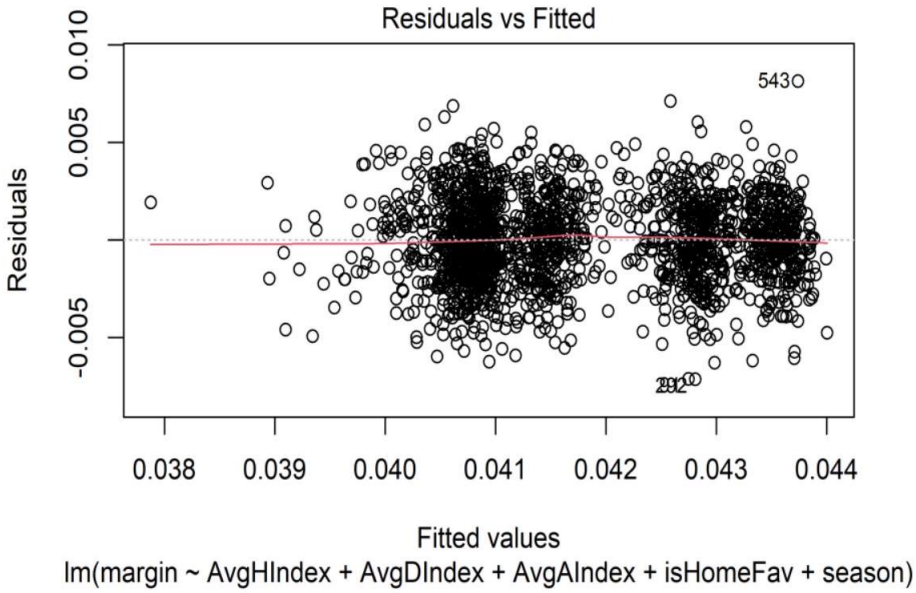
$$\begin{aligned}\text{margin} = \beta_0 &+ \beta_1\ \text{AvgHIndex}\ + \beta_2\ \text{AvgDIndex}\ + \beta_3\ \text{AvgAIndex}\ + \beta_4\ \text{isHomeFav} \\ &+ \beta_5\ \text{season2021}\ + \beta_6\ \text{season2122}\ + \beta_7\ \text{season2223}\ + \beta_8\ \text{season2324}\ .\end{aligned}$$

Here is what we found:

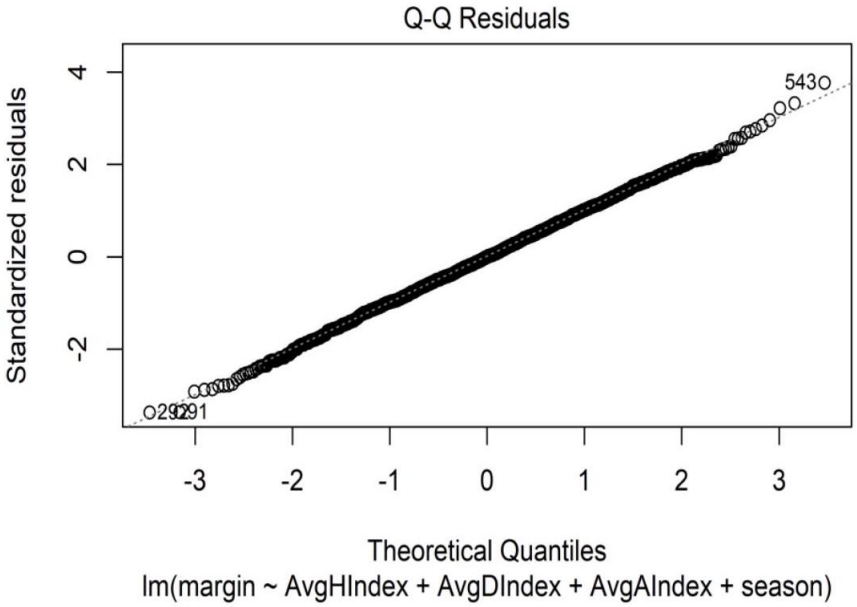
Predictor	Coefficient	p-value
β_0	$4.270 \cdot 10^{-2}$	< 0.001
β_1	$-3.388 \cdot 10^{-6}$	< 0.001
β_2	$4.302 \cdot 10^{-6}$	0.00106
β_3	$-3.369 \cdot 10^{-6}$	< 0.001
β_4	$1.471 \cdot 10^{-4}$	0.26723
β_5	$7.495 \cdot 10^{-4}$	< 0.001
β_6	$-1.971 \cdot 10^{-3}$	< 0.001
β_7	$-2.099 \cdot 10^{-3}$	< 0.001
β_8	$-1.499 \cdot 10^{-3}$	< 0.001

Based on these results, we can say that:
Before accounting for odds, favorite status, or season, bookmakers embed about a 4.27 % margin in each match ($p < 0.001$). A one-unit increase in AvgHIndex corresponds to a -3.388×10^{-6} change in margin ($p < 0.001$), confirming a tiny but statistically significant home-odds effect. AvgDIndex shows $a + 4.302 \times 10^{-6}$ change ($p = 0.00106$), now significant when all factors are combined, though the magnitude remains negligible. AvgAIndex yields -3.369×10^{-6} ($p < 0.001$), a similarly small yet significant away-odds effect. Matches where the home side is favorite have a $+1.471 \times 10^{-4}$ higher margin, but $p = 0.267$ indicates no statistical significance. The $2020-21$ season witnessed a 0.07495% rise in margin ($p < 0.001$), while the seasons $2021-22$, $2022-23$ and $2023-24$ witnessed statistically significant negative shifts of $-0.1971\%, -0.2099\%$ and

-0.1499% respectively (all $p < 0.001$). The magnitudes of these coefficients are much larger than those of the odd indices, meaning that seasonal effects dominate. The R^2 for this full model is 0.2289 , which is essentially the same as the previous model (0.2288). Adding favorite status does not materially boost explanatory power beyond what seasons already provide. Here are our diagnostic plots for this linear regression:
First, the residuals vs fitted plot:



Based on this graph, we can say that the points in this plot does not follow a particular shape, which means that our linear regression has captured the linearity assumption between independent and dependent variable. Moreover, even though there are some outliers, the majority of the residuals have the same variance, which supports our homoscedastic assumption about the linear regression.
Second, the Q-Q plot:



Based on this plot, we can say that most of the points line on the 45-degree line, except for slight tail deviations, which suggest that the residuals are close to normal distribution, which satisfies our assumption about normality of residuals.

Key Takeaways:

Seasonal shifts are by far the strongest predictors of margin, explaining nearly a quarter of its game-to-game variation. Odds indices, while statistically significant, have negligible real-world impact on the margin. Favorite status adds no predictive value even if seasons and odd indices are included.
In short, to forecast or interpret bookmaker margins, incorporate seasonal pricing strategies rather than focusing on home/away odds or favorite/underdog status. Oddsbased tweaks exist but are too small to exploit profitably.

6 Conclusion

Our analysis of English Premier League betting data from 2019–20 through 2023–24 reveals that bookmakers employ a remarkably uniform margin-setting strategy at the match level. We found no evidence that overrounds differ by match outcome (home win, draw, away win) or by whether the home side is favored (Sections 5.1 and 5.3). Although margins have tightened modestly over time—falling by about 0.13 percentage points between 2019–20 and 2023–24 (Section 5.4), every match continues to carry a baseline profit cushion of roughly 4.2 %. In contrast, season-to-season effects emerge as the dominant driver of margin variation. Chi-square tests showed highly significant shifts in the prevalence of "high-margin" matches across seasons (Section 5.2), and our regression

models demonstrated that seasonal indicators alone explain nearly 23% of game-to-game margin swings—far more than any odds-based measure (Sections 5.7 and 5.8). This suggests that broad market forces—such as regulatory changes, competitive dynamics, and overall betting volume—shape pricing behavior far more than individual match characteristics. For bettors, these findings imply that targeting specific outcomes or favorite/underdog scenarios is unlikely to yield a reliable edge; instead, value may be found by identifying seasons or periods when competition or regulation drives margins lower. For bookmakers, the results validate a standardized margin framework, while highlighting the importance of adjusting overrounds in response to season-level conditions rather than tailoring prices on a per-match basis. Limitations of our study include the exclusion of direct measures of betting volume, in-play pricing, and league-level heterogeneity. Future research could incorporate these factors, as well as extend the analysis to other sports or markets, to build a more complete picture of sportsbook pricing strategies. Overall, by combining hypothesis tests and regression analysis, this project illuminates both the stability and the seasonal dynamism of bookmaker margins, offering actionable insights for analysts, bettors, and industry participants alike.

7 References

Birches Health. 2024. *The consolidation of the sports betting industry.* Available from: https://bircheshealth.com/resources/consolidation-sports-betting-industry

Hegarty T, Whelan K. 2024. *Comparing two methods for testing the efficiency of sports betting markets.* UCD Centre for Economic Research Working Paper Series WP24/03. Available from: https://www.ucd.ie/economics/t4media/WP24_03.pdf