Assessing the accuracy of betting odds in European football

Joseph Edward Pym
8404110

Supervisor
Maha Moustafa

A dissertation presented in the School of Computing, Electronics & Mathematics, Coventry University, for the degree of Bachelor of Science in Mathematics & Statistics.

January First Check Submission
January 25th 2021

0.1 Acknowledgements

Insert my thanks at the end here.

0.2 Abstract

The last thing to do: This is a page (limited) summarising the dissertation.

Contents

Disclaimer

Involvement in gambling can often lead to highly dangerous and damaging consequences, both for the bettor and those around them: problem gambling is characterised as "*persistent and recurrent problematic gambling behaviour leading to clinically significant impairment or distress*" (**Dsm5**).

This dissertation will not seek to solve the social, economic or political issues surrounding gambling, but will instead focus wholly on the mathematics and statistics upon which sports gambling is based.

1

Introduction

1.1 Background Information

1.1.1 Association football in Europe

Association football, also known as ‘soccer’ or simply ‘football’, is the most popular sports in the world (**Giulianotti12**), with—according to a survey by the world’s governing body, FIFA, in 2001—240 million players worldwide (**FifaSurvey01**).

Professional football in Europe is governed by UEFA, with each nation having their own governing body (such as the Deutscher Fußball-bund, DFB, in Germany) administering football in that country (**UefaMembers**), including the league systems (pyramids). Generally, the systems follow a similar structure, with a number of promotion and relegation places contested for throughout the course of a year-long season, ensuring each division has a similar ability (**PremierLeagueHandbook**). The top division in each country is allocated a number of qualification places to the two Europe-wide club competitions, the Champions’ League and Europa League.[1] This allows the best teams from each country to compete against each other.

The number of allocated places is chosen via the UEFA country coefficient, which ranks the countries by the performance of their collective clubs in these competitions: we choose the top six[2] of these to be considered the *elite* European leagues. These are given in Table 1.1 (**UefaCoeffs**).

Later in this dissertation, we will consider the English and Scottish football pyramids; the order of their leagues used in the analysis are given in Table 1.2. We choose these two leagues as `football-data.co.uk` have data on five English leagues and four Scottish leagues. No other league has more than two; this will allow us to compare between different *tiers* in the pyramid.

[1] A third competition, the Europa Conference League, is planned for the 2021/22 season (**UefaEcl20**).

[2] As of 5th January, 2021.

Table 1.1: The UEFA Country Coefficient for the top six European leagues.

Country	*Top Division*	*Coefficient*
Spain	La Liga	92.283
England	Premier League	90.712
Italy	Serie A	72.295
Germany	Bundesliga	71.856
France	Ligue Une	54.915
Portugal	Premiera Liga	47.349

Table 1.2: The English and Scottish football league pyramids.

Tier in Pyramid	*English Pyramid*	*Scottish Pyramid*
1	Premier League (EPL)	Premier League (SPL)
2	Championship	Division 1
3	League One	Division 2
4	League Two	Division 3
5	Conference/National League	—

1.1.2 Probabilities, odds and gambling

Every event, say i, has a PROBABILITY of occuring, denoted as $\mathbb{P}(\text{Event } i) = p_i$, between 0 (almost never occurs) and 1 (almost certain to occur), with the sum of all possible outcomes being equal to 1 (**Grinstead12**). For example, when rolling a fair, six-sided die, numbered one to six, we can draw a table with the probabilities of rolling the die and number X being rolled:

x	1	2	3	4	5	6
$\mathbb{P}(X = x) = p_i$	1/6	1/6	1/6	1/6	1/6	1/6

In gambling, the ODDS of an event are used, rather than the probabilities (**Strumbelj14**). For a completely fair, non-profit casino or bookmaker, the odds offered for an outcome would be equal to the inverse of the probability of the event: for our fair die scenario above, the odds O_i (written in different styles[3]) would be:

	x	1	2	3	4	5	6
	British Style/Fractional	6/1	6/1	6/1	6/1	6/1	6/1
O_x	European Style/Decimal	7.0	7.0	7.0	7.0	7.0	7.0
	American Style/Moneyline	600	600	600	600	600	600

If one was to place a bet worth £1 (the STAKE) on rolling a three, and this happened, the casino or bookmaker would—for British and European odds—multiply the stake, S, by the odds O to find one's profit: the total pay-out, P would include the original stake: $P = SO + S$. A positive American odd shows the amount of money you would win for a \$100 bet (or whichever currency), whereas a negative odd is the amount you need to bet in order to win \$100 (**AmericanOdds19**).

[3] Throughout this report, we will use the European/decimal odds system: this is also favoured by `football-data.co.uk`.

1.1.3 Gambling markets and betting in football

There are a vast range of *markets* used in football betting (**Bet365Markets**); the ones we are most focused on in this report, are:

- 1X2 (or full-time result)—betting on the final outcome of the match being a home win, a draw or an away win.
- Goal Markets—this is a bet on the amount of total goals or number goals for either side, usually given as a half (e.g., 0.5, 1.5, 2.5, etc.), with odds offered for under/over the given amount.
- Asian Handicap (AH)—this style of betting allows a seemingly one-sided fixture to become competitive. For AH bets, a handicap given can be a whole number or half-number, used as a headstart (if positive) or a detriment (negative) to a team; for example, if, in a match between Chelsea and West Brom, Chelsea's handicap was -2.5, they would need to win by 3 goals in order for a bet on them to win: West Brom would just need to avoid a loss by 3 goals for a bet on them to win. In addition, Asian handicaps can also be quarter-numbers (say $\frac{3}{4}$): the bet is then split into two: half the stake on $\frac{1}{2}$ and half on 1, in this case.

Gambling is a huge part of football culture, with 27 of the 44 teams (61.3%) in the English Premier League and Championship having a gambling company as their main shirt sponsor (**Davey20**); those that don't likely have a betting company as a 'Club Partner', for example Arsenal, who have Fly Emirates on their shirts, are partnered with SportsBet.io; Manchester City—Etihad Airways on their shirts—are partnered with Marathon Bet (**PremierLeagueHandbook**). The combined income of betting partnerships in the Premier League is around £70 million. 8 of the 20 Spanish La Liga sides have a gambling company on their shirts: BetWay alone sponsor three (**LaLigaSponsors**). This leads one to wonder: how accurate are the betting odds by these companies, so invested in football, on the matches?

1.2 Literature review

The Literature Review will be an assessment of previous works in this field, analysing what they did, how it differs from my work and why I will use their methods, or change them.

1.3 Rationale

Motivations about why I am doing a study into this — it will likely link to the literature review

1.4 Aims and objectives

Throughout this project, we will aim to assess the accuracy of betting odds offered by a range of bookmakers in European football. We will look into different markets, such as the Asian Handicap and goal markets, and try to determine factors that may or may not influence the accuracy of the odds, such as the level of the match and country (league) of the match. We will also investigate whether or not there is an existence of the "FAVOURITE-LONGSHOT BIAS", where, on average, bettors tend to overvalue the underdogs and undervalue the favourites.

In order to answer this, we will utilise data from `football-data.co.uk`—a website set up by Joseph Buchdahl, a betting analyst and author of multiple published works about betting (**BuchdahlAbout**)—which provides historical results and odds in easy-to-access comma-seperated-value (.csv) files (**FootballDataAbout**). With this data, we will explore different angles assessing the accuracy of the odds, including exploratory analysis (**Hoaglin77**) and calculating the predictive power of the odds (**Owen09**). In addition, we will look at a range of visual aids such as tile plots, histograms and density plots. We will create models to predict the actual observed probability of an event from the bookmaker consensus probability, obtained by taking the inverse of the consensus odds and normalising.

Our objectives are as follows: first, we will look into the accuracy of odds offered in the 1X2 market in *elite* European leagues: the Spanish La Liga, English Premier League, Italian Serie A, German Bundesliga, French Ligue Une, and Portuguese Premiera Liga. We will combine this analysis with a review into whether or not COMPETITIVE BALANCE impacts the accuracy of the odds offered.

Secondly, we will look into the accuracy of odds offered in different levels of the English and Scottish league systems. *Combining this analysis with an investigation into the Favourite-Longshot bias, goal markets and Asian Handicap markets. Finish this paragraph after analysis is done.*

1.5 Methods

All of the analysis will be conducted via **R**, using the `car`, `mass`, `ggplot2` and `gridExtra` packages: the full code scripts for each chapter can be found in the Appendices. Not included in these scripts are the installation of packages; these are done by the following:

```
install.packages("PackageName")
```

1.5.1 The Data

Sourcing the Data

For this report, I have used `www.football-data.co.uk` (F-D) to collect my data. F-D use the following bookmakers for their odds (as of 02/10/20): Bet365, Blue Square Bet, Bet & Win, Gamebookers, Interwetten, Ladbrokes, Pinnacle, Sporting Odds, Stan James, VC Bet, William Hill. For game statistics, such as home/away corners, free kicks, shots (on/off target), offsides and cards, F-D uses BBC Sport, ESPN Soccer, Gazzetta.it and Football.fr, with betting odds taken from the individual bookmakers. The odds for matches during the weekend are collected on Friday afternoons; odds for midweek matches are collected on Tuesday afternoons. Statistics for 2000-01 and 2001-02 for the English, Scottish and German leagues were provided by Sports.com (which is under now ownership and now unavailable) (**FootballDataNotes**).

All the data is stored in comma-separated value (.csv) files, with each fixture having its own row. The columns from these .csv files that are required for our analyses are given in Table 1.3.

Table 1.3: Columns from `Football-Data`'s Datasets Used

Name	*Meaning*
`div`	Division
`date`	Date of the match
`HomeTeam`, `AwayTeam`	The home/away side in the match
`FTHG`, `FTAG`	Full-time home/away goals
`FTR`	Full-time result: Equal to 'H' for a home win, 'A' for an away win, and 'D' for a draw.
`BbAvH`, `BbAvA`, `BbAvD`	The bookmaker consensus odds for a home win, away win and draw, respectively. From the 2019/20 seasons onwards, these were renamed `AvgH`, `AvgA` and `AvgD`.

Data Storage

With R, we can access the .csv files without downloading them, using the web URL in place of a file name with the `read.csv("")` command. Further, we can use a `for` loop to download and store all the datasets in R; this is shown in Section 2.1.

The country codes given to each country are:

Spain	`es`
England	`en`
Italy	`it`
Germany	`de`
France	`fr`
Portugal	`po`
Scotland	`sc`

1.6 Structure

This dissertation will be split into two main chapters: Chapters 2 and *3 (English/Scot)*, with an introductory chapter (Chapter 1) and a conclusion (Chapter *4*). In both Chapters 2 and *3*, we will present a question, before using previous literature, data and figures generated through `ggplot2` to answer the question. Each of the chapters will begin with exploratory data analysis, before further exploring themes we find, and ending with a brief conclusion. In the appendices, we include a list of the statistical theorems, distributions and calculations we use (Appendix *0*), a list of definitions (Appendix A)[4] and the

[4] The words in SMALL CAPS in the dissertation are defined here.

full R code used (Appendices B and *3*), as well as the ethics approval cerificate *and any other stuff I need to include, done at the end.*

2

Assessing the accuracy of betting odds in European elite leagues between 2005 and 2020

This chapter will aim to assess the accuracy of betting odds offered by bookmakers in *elite* European leagues, as defined prior (Section 1.1.1, Table 1.1). To begin our analysis—after reading in and cleaning our data—we will use the analytical techniques of initial and exploratory data analysis, with the former looking at a smaller sample of data and ensuring it behaves as expected the latter looking at four "major themes" (**Hoaglin77**) to answer questions about our dataset and, using the `ggplot2` package, we will create visual aids to further explore and analyse the dataset.

Then—using correlation analysis—we will create models to predict the observed probability of a result from the bookmaker's offered odds: from these models we will be able to find statistics to test the FIT of the data. To aid this, we will find the P_1 and P_2 statistics for predictive performance (**Owen09**) before finally looking at the impact the COMPETITIVE BALANCE of a league has on the accuracy of the bookmaker and presenting our conclusions.

2.1 Initial data analysis

Initial data analysis (IDA) is used to answer four questions about our data:

- What is the quality of our data?;
- What is the quality of the measurements?;
- Did the implementation of the study fulfil the intentions of the research design?;
- What are the characteristics of the data sample? (**Ader08**)

To answer these questions, we will conduct simple tasks—such as calculating the mean consensus probabilities and comparing them with the observed probabilities, and plotting a histogram to assess the distribution of the consensus probabilities—on a sample data set: we choose one league for one season. Chosen at random, we will look into the **French Ligue Une** over the **2016/17** season.

Our first step is reading in the data from `football-data.co.uk`, and removing the columns we do not need. We use the code below to do this.

```
fr_l1_1617 <-
    read.csv("https://www.football-data.co.uk/mmz4281/1617/F1.csv")
fr_l1_1617 <- fr_l1_1617[,c("Div", "Date", "HomeTeam", "AwayTeam",
    "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD", "BbAvA")]
fr_l1_1617 <- na.omit(fr_l1_1617)
```

As mentioned in Section 1.5, we will be considering probabilities throughout the analyses, rather than odds: our first step is to calculate the bookmaker consensus probabilities, before *normalising* them—that is, ensuring the sum equals one. The equation for this is given in Equation 2.1, where j is the outcome being normalised; i is all possible outcomes. The code to perform this, rounded to 4 decimal places, is below.

$$\frac{\mathbb{P}_{\text{cons}}(\text{Outcome}_j)}{\sum_{i=1}^{3}\mathbb{P}_{\text{cons}}(\text{Outcome}_i)} \tag{2.1}$$

```
fr_l1_1617$AvgHProbPN <- with(fr_l1_1617, round(1/BbAvH, 4))
fr_l1_1617$AvgDProbPN <- with(fr_l1_1617, round(1/BbAvD, 4))
fr_l1_1617$AvgAProbPN <- with(fr_l1_1617, round(1/BbAvA, 4))
fr_l1_1617$Overround <- with(fr_l1_1617, (AvgHProbPN + AvgDProbPN +
    AvgAProbPN))
fr_l1_1617$AvgHProb <- with(fr_l1_1617, round(AvgHProbPN/Overround,4))
fr_l1_1617$AvgDProb <- with(fr_l1_1617, round(AvgDProbPN/Overround,4))
fr_l1_1617$AvgAProb <- with(fr_l1_1617, round(AvgAProbPN/Overround,4))
```

We normalise the probabilities to counter the bookmaker commission, or *overround*, although this will differ between games and between outcomes (**Henery99**). Doing this allows us to treat the consensus probabilities as statistical probabilities.

To begin our analysis, we will compute the consensus mean probabilities μ_i and consensus standard deviation σ_i for each outcome i, in order to compare with the observed probabilities, and to compare variation in each outcome. The results, to four decimal places, are shown in Table 2.1.

Table 2.1: IDA Calculations

	Home Win	*Draw*	*Away Win*
Mean Consensus Probability μ_i	0.4414	0.2701	0.2886
Observed Probability	0.4895	0.2474	0.2632
Consensus Standard Deviation σ_i	0.1540	0.0469	0.1366

From this table, we can see the mean consensus probabilities are very close to the observed probabilities: less than 0.05 off for each outcome. Interestingly, the standard deviation of a draw is much lower than the standard deviation for both home and away wins, indicating that the variation in the odds offered for draws are much lower than those for a clear winner[1].

We will now produce histograms to plot the distribution of the consensus probabilities for each outcome. As mentioned, we will be using the `ggplot2` package for our graphics; to produce the histogram for the distribution of the consensus probabilities of a home win, $\mathbb{P}_{\text{cons}}(\text{Home Win})$, we use the code below. Using the `gridExtra` package, we can add the histograms for each outcome into one figure, shown in Figure 2.1.

```
ggplot(fr_l1_1617, aes(AvgHProb)) + geom_histogram(binwidth=0.05,
    fill="blue") + coord_cartesian(xlim=c(0,1)) + theme_light() +
    labs(title="Home Win", x=NULL, y=NULL)
```

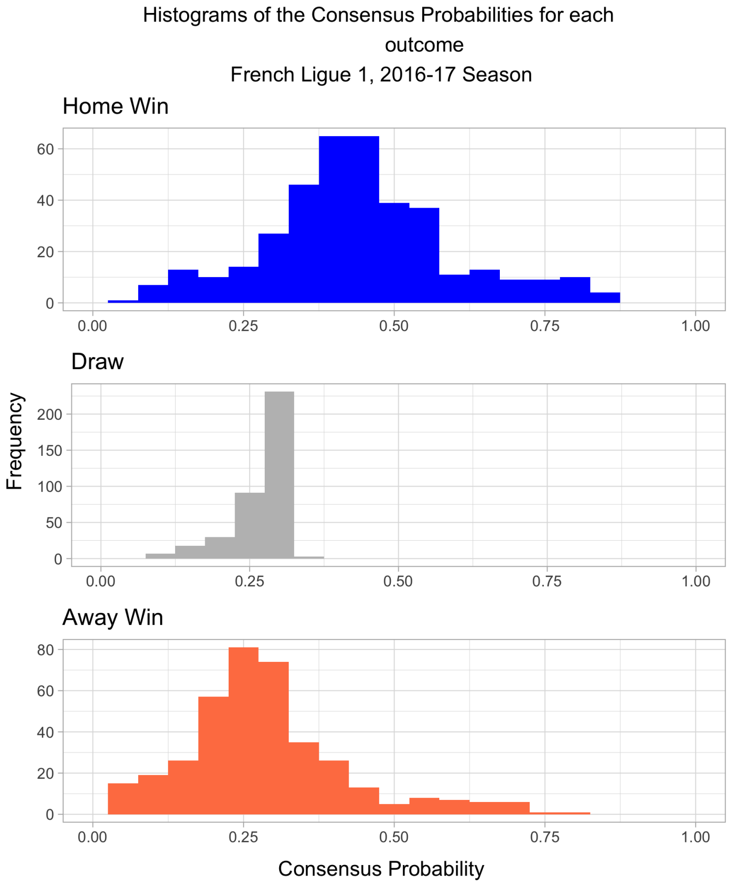

Figure 2.1: N.B. the change of scale on the y axes.

This figure shows that the consensus probabilities of a home win are sym-

[1] I.E. a home or away win.

metrically distributed around the mean ($\mu_{\mathrm{H}} = 0.44$, with a distribution similar to the Normal bell curve. For the consensus probabilities of an away win, the probabilites are positively skewed[2] suggesting a greater proportion of measurements lie to the right/are greater than the peak value (**Mendenhall13**). This means our data will include a few unusually large measurements. Contextually, these could be the league leaders playing away against a relegation-battling side.

The $\mathbb{P}_{\text{cons}}$(Draw) graph display assists our finding (based on σ_{Draw}) that the variation of bookmaker consensus probabilities of a draw is very low: Over 225 ($N = 380$ matches) of the games lie in the same bin[3] with no values at all recorded bove 0.4.

The rationale behind conducting IDA was to answer the four questions set out above (**Ader08**), about the quality of our data and measurements, whether the implementation of the study fulfilled the intention of the research design, and the characteristics of our sample. The quality of the data and measurements are both acceptable for our research: no measurements are missing, our values were as expected ($\mathbb{P}_{\text{cons}} \approx \mathbb{P}_{\text{obs}}$), and we were able to conduct computations and produce plots without any issues. The question of the study, about assessing the accuracy of betting odds, will be able to be answered: we have access to the information we need (consensus odds and actual results)—in fact, we have much more information we can read in if needed—and our final question, in reference to the characteristics of our sample, can be answered by saying that our data suggests the mean consensus probabilities are roughly equal to the observed probability for each outcome; bookmakers offer less variation on their odds for a draw than for a clear winner; and the consensus probability of a home win is symmetic, whereas the consensus probability of an away win is positively skewed.

2.2 Exploratory data analysis

Exploratory data analysis, EDA, is how we will initially explore the data. There are four "major themes which motivate many of the techniques" we will apply (**Hoaglin77**). These are:

- **Displays**: reveal major features, outliers, non-linearities, discontinunities, skewness, ETC. that calculations such as means, standard deviations and least square regressions cannot show;
- **Residuals**: defined as the observed data minus the fitted data, a clear pattern in the plot of the residuals v. fitted values indicates improvement is possible;
- **Resistance**: dealing with outliers. If we find them, we can run parallel tests (one with outliers; one without) and comparing, testing the resistance of our data to the outliers (not dissimilar to the idea of statistical leverage);
- **Transformations**: does adding a transformation, such as taking the n-th root, logarithms, logits/probits, ETC. allow us to make sense of the data?

[2] Or right-skewed.

[3] The width of all bins across all three histograms is 0.05.

In this section, we will also create visual analyses of the bookmaker's odds and performance; namely, boxplots, density plots and a tile plot. To begin, however, we must first read in our data. As we will be considering six leagues and 15 seasons (90 datasets), it is efficient to use a `for` loop to do this, and to clean our data. The code to do so—and to normalise our underlying probabilities—is below.

```r
countries <- c("de", "en", "es", "fr", "it", "po")
countries.web <- c("D1", "E0", "SP1", "F1", "I1", "P1")
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112",
    "1213", "1314", "1415", "1516", "1617", "1718", "1819", "1920")
eliteTemp <- NULL
elite <- NULL

for (i in seasons){
   for (j in 1:6){
      eliteTemp <-
          read.csv(paste0("https://www.football-data.co.uk/mmz4281/", i,
          "/", countries.web[j], ".csv"))
      eliteTemp$Country <- with(eliteTemp, countries[j])
      eliteTemp$Season <- with(eliteTemp, i)
      if (i=="1920"){
         eliteTemp$BbAvH<-eliteTemp$AvgH
         eliteTemp$BbAvA<-eliteTemp$AvgA
         eliteTemp$BbAvD<-eliteTemp$AvgD
      }
      else{}
      eliteTemp <- eliteTemp[ ,c("Div", "Date", "HomeTeam", "AwayTeam",
          "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD", "BbAvA", "Country",
          "Season")]
      elite <- rbind(elite, eliteTemp)
   }
}
elite <- na.omit(elite)

elite$AvgHProbPN <- with(elite, round(1/BbAvH, 4))
elite$AvgDProbPN <- with(elite, round(1/BbAvD, 4))
elite$AvgAProbPN <- with(elite, round(1/BbAvA, 4))

elite$overround <- with(elite, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
elite$AvgHProb <- with(elite, round(AvgHProbPN/overround, 4))
elite$AvgDProb <- with(elite, round(AvgDProbPN/overround, 4))
elite$AvgAProb <- with(elite, round(AvgAProbPN/overround, 4))
```

The number of matches in our `elite` dataset is 31,346. As this is a large set, we can apply the central limit theorem, and assume the mean of the random variables (contextually, matches) follows the Normal distribution (see **Section on Statistical Theorems**). For our later analysis, we also will need to find the *correct* probability (and the natural logarithm of it)—the bookmaker consensus probability of the event that was observed; the two *incorrect* probabilities; and we will need to create smaller datasets, one for each league.

Our first step after cleaning our data is to compute, as in the IDA, the bookmaker consensus mean probabilities and compare them to the observed probabilities. These are shown in Table 2.2.

Table 2.2

	Home Win	*Draw*	*Away Win*
Mean Consensus Probability μ_i	0.4472	0.2620	0.2908
Observed Probability	0.4589	0.2566	0.2845
Consensus Standard Deviation σ_i	0.1714	0.0478	0.1536

As in Table 2.1, the mean consensus probabilities are extremely close to the observed probabilities: the magnitude difference is than 0.02 for all three outcomes. Similarly, the standard deviations indicate that the consensus probabilities offered for a draw vary significantly less than for those with a clear winner.

Our first graphical step with the elite dataset is to create a boxplot for each outcome. This will allow us to see whether or not there is a significant difference in the observed outcome for the bookmaker probabilities of each event, as well as possibly seeing a trend in the bookmaker's consensus probabilities. This is shown in Figure 2.2.

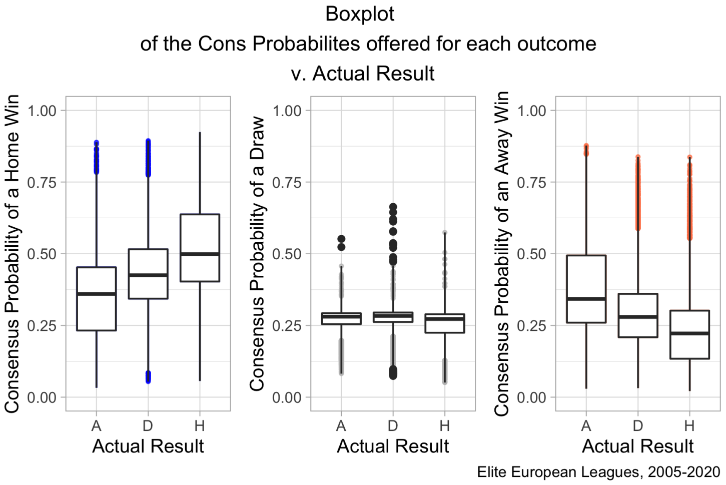

Figure 2.2

The figure implies there is no significant difference; whilst it appears there is a correlation—for games that end in home wins, the consensus probability of a home win is higher, and vice versa—this plot shows no strong significance. The plot does, however, strongly reiterate our previous point about variation, or lack thereof, of the consensus probabilities of a draw (the box is smaller). Interestingly, however, the draw plot has a large number of outliers. We will consider these later.

Instead of creating a binned histogram, as in Figure 2.1, we will use a density plot: using KERNAL DENSITY ESTIMATION, we will smooth out a histogram using a number of equally spaced points at which the *density* of the distribution is estimated. This number of spaces is a power of two: we choose $2^9 = 512$ (and a Normal (Gaussian) kernal, due to the CLT). To do this using `ggplot2`, we can use the code below (this shows our density plot for home and away wins—we include draws seperately due to a different y axis scale needed. This code, along with that for the draws, results in Figure 2.3

```
ggplot(elite, aes(x=AvgHProb)) + geom_density(color="blue") +
    geom_density(data=elite, mapping=aes(x=AvgAProb), color="coral",
    show.legend = TRUE) + coord_cartesian(xlim=c(0,1)) + labs(title =
    "Home and Away Wins", caption = "Elite Leagues, 2005-2020", x =
    "Consensus Probability", y = "density") + theme_light()
```

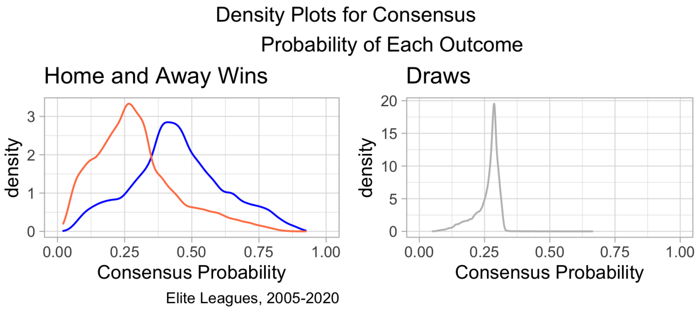

Figure 2.3

Interpreting this figure, we can immediately see that the trend shown from Figure 2.1—$\mathbb{P}_{\text{cons}}$(Home Win) is symmetrically distributed around the peak (mean $\mu_{HW} = 0.45$); $\mathbb{P}_{\text{cons}}$(Away Win) has a positive skew; and $\mathbb{P}_{\text{cons}}$(Draw) has much less variation, with few matches above the peak—still holds. However, we observe a number of matches with a large consensus probability of a draw: in fact, some have $\mathbb{P}_{\text{cons}}$(Draw) around 0.65. Using R to find these matches, we can find the extremes: Table 2.3 shows matches with a consensus probability of a draw greater than 0.6.

Table 2.3: Matches with $\mathbb{P}_{\text{cons}}(\text{Draw}) > 0.6$

League	*Date*	*Home Team*	*Away Team*	*FTR*	$\mathbb{P}_{\text{cons}}$(Draw)
Serie A	09/05/10	Bologna	Catania	1-1	0.6634
Serie A	08/05/11	Bologna	Parma	0-0	0.6445
Serie A	20/05/07	Torino	Livorno	0-0	0.6208
Serie A	03/04/11	Chievo	Sampdoria	0-0	0.6121

Immediately, one will notice all of these matches are both a) in the Italian Serie A; and b) in the late stages of the football season[4], leading us to two

[4] The season normally starts in August and ends in May (**PremierLeagueHandbook**).

possible reasons behind this:

- The Italian Serie A has a history of match fixing in recent times: the CALCIOPOLI which occured during the 2004/05 and 2005/06 season involved Juventus, A.C. Milan and Lazio, among others—three of Italy's largest clubs; in 2015, Catania's president was one of several arrested for match-fixing in Serie B matches (**Calciopoli**).
- Due to these games occuring in the late stages, it is possible for a scenario where both teams would benefit from a certain result.[5]

Whilst fixed matches would naturally impact our results, due to the small number of games impacted, it is unneccesary to exclude them from our analysis: this is explained in Section 2.3 with an investigation into the leverage.

As these games are all in the Serie A (in fact, 54 of the 55 matches with the highest probability of a draw are in the Italian top league[6]), it makes sense to split this density plot into the different leagues. The Home Win (and to a lesser extent Away Win) plots in Figure 2.4 imply leagues can be split into two categories: those with a *unimodal* density (one peak), and those with a *trimodal* density (three peaks). The latter group (England, Portugal, Spain) have a peak between 0 and 0.25 (very low probability of a home/away win, depending on the plot) and a peak between 0.75 and 1 (very high probability of a home/away win). One reason could be COMPETITIVE BALANCE: we investigate this later in Section 2.6.

The final visual aid in this section we will use is a tile plot. Similar to a heat map, this will allow us to create a three dimensional representation of data in the 2D plane. For our dataset, we will plot the match result on the x (home goals) and y (away goals) axes, allowing us to see both the full time result—tiles on the diagonal ($x = y$) are draws; upper triangle are away wins ($y > x$); lower triangle are home wins ($x > y$)—and the *magnitude* of the result—or how convincing the result is: a match further from the diagonal has a greater disparity in goals, and so can be considered a more convincing win. Representing the z axis, each square will be shaded in with the correct bookmaker probability: with a low consensus probability, the square will be lighter. We would expect these to be closer to the diagonal. Due to a low number of *extreme* results (more than 5 goals scored for a team), we will group these into a 5+ tile[7]. Before we can analyse the plot, it is important to know the bin sizes for each tile. These are given in Table 2.4

Inspecting the figure, we notice that the results on the diagonal all have a similar consensus probability, around 0.3, as one would expect. Considering the lower triagle (home wins) first, we notice the pattern expected holds: that is, tiles furthest from the diagonal are darker. Interestingly, the darkest tile is for a 5 (plus) - 2 home win; the bin size for this is low, though, with $N = 106$ matches. Whilst there are only 12 games in the 5 (plus)-4 tile, it is striking that it has one

[5] The DISGRACE OF GIJÓN in the 1982 World Cup is a particularly famous example of this (See Appendix A.2, Definition 5).

[6] The fiftieth highest was a game between Granada and Atletico Madrid on 23rd May 2015: the $\mathbb{P}_{\text{cons}}$(Draw) was 0.3940; it finished 0-0.

[7] One may notice here that a game finishing in a non-draw but with both sides scoring 5 goals or more will be plotted as a draw; using `View(elite[elite$FTR=="D",])` to view all the draws, we can see the highest-scoring draws in our dataset are two 5-5 matches: Lyon v. Marseille (2009), West Bromwich Albion v. Manchester United (2013).

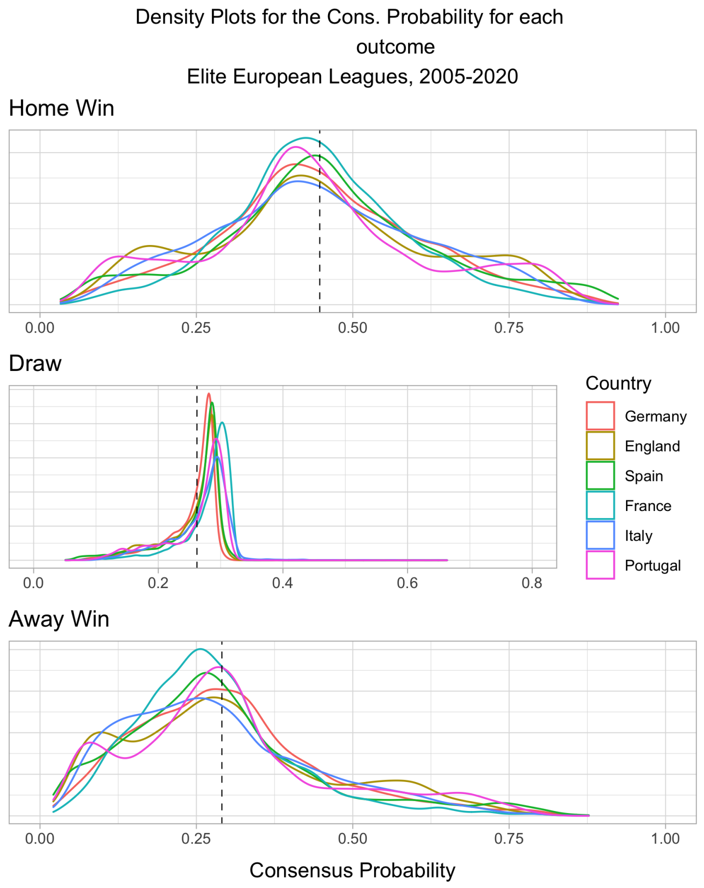

Figure 2.4

Table 2.4: The Bin Size for each tile of Figure 2.5

	Home Goals					
Away Goals	0	1	2	3	4	5+
5+	119	105	67	26	7	2
4	257	294	174	63	36	12
3	636	869	547	299	109	45
2	1451	2003	1508	712	265	106
1	2299	3688	2756	1375	528	281
0	2510	3342	2575	1371	594	315

of the lowest correct consensus probabilities of all the tiles. Finally, considering the upper triangle (away wins), our pattern is not as consistent as with home wins, but there is still evidence of it, with the highest probabilities being further from the diagonal. As a whole, the figure is implying what we would expect: games with greater disparity in full time goals (i.e. more convincing wins) have a higher bookmaker consensus probability of the correct result than closer games

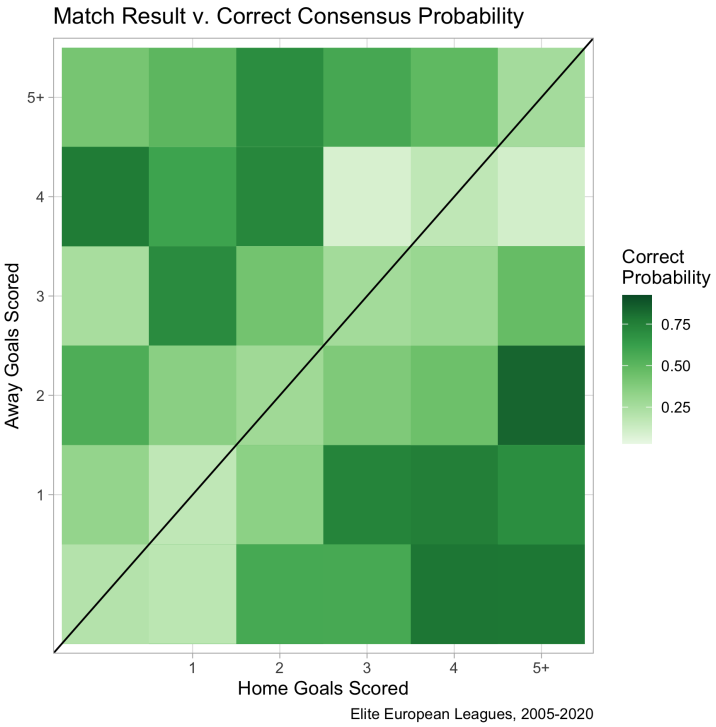

Figure 2.5: Tile Plot of the Correct Consensus Probability for each possible result

and draws.

We conducted EDA to check four themes: displays, residuals, resistance and transformations.

- Displays: Whilst there was no discontinuity, we saw—as in our IDA in Section 2.1—the consensus probabilities of away wins are positively/right-skewed, and of draws have a very low variance. We saw, and assessed, the outlying matches with a large $\mathbb{P}_{\text{cons}}(\text{Draw})$.
- Residuals & Resistance: As we have not created a model (we do this in Section 2.3), we cannot plot the residuals.
- Transformations: Our data, so far, makes sense—and is as expected. There is nothing in the plots to suggest we need to, or to justify, taking any transformations.

2.3 Correlation analysis

In this section, we will determine the ccoefficient of determination R^2 and the root mean square error RMSE. We can do this by creating a linear model to predict the observed probability, from a given consensus probability. For high levels of accuracy, we will find a high R^2 and low RMSE (**Mendenhall13**). Alternatives to RMSE include mean square error (MSE), mean absolute error and median absolute error: RMSE is more sensitive to outliers than mean and median absolute error; we choose to use RMSE due to a relatively low amount of outliers, and its greater theoretical relevance (**Hyndman06**). R^2 describes the percentage of the variation in a variable—in our case, observed probability—due to a predictor—bookmaker consensus probability (**Draper98**).

Once we have determined R^2 and RMSE, both overall and for each outcome, we will determine them for each league, which will be used in Section 2.6. To perform this in R, we will games with similar consensus probabilities into bins, and see the actual v. consensus probabilty for each. For home wins, we use the code below. N.B., we use `tapply` to find the mean of the bin (R's default is midpoint); in the line defining the observed probabilites, we use `elite.observed.probabilites.TabH[3,]` to select the third row (home wins, alphabetically). For away wins and draws, we use `[2,]` and `[1,]` respectively. Finally, we use 124 bins, meaning that each bin has over 250 matches. Doing this for each outcome and overall, we find our R^2 and RMSE values, shown in Table 2.5.

```
elite$AvgHProb.cut <- cut(elite$AvgHProb, 124, inclues.lowest=TRUE)
levels(elite$AvgHProb.cut) <- tapply(elite$AvgHProb, elite$AvgHProb.cut,
    mean)
elite.observed.probabilites.TabH <- prop.table(table(elite$FTR,
    elite$AvgHProb.cut), 2)[c(1, 2, 3),]
elite.observed.probabilites.H <- elite.observed.probabilites.TabH[3,]
elite.bookmaker.probabilites.H <-
    as.numeric(names(elite.observed.probabilites.H))

elite.lm.home <- lm(elite.observed.probabilites.H ~
    elite.bookmaker.probabilites.H)
round(summary(elite.lm.home)$r.squared, 5)
round(sqrt(mean(elite.lm.home$residuals^2)), 5)
```

Table 2.5: R^2 and RMSE values for the elite leagues, 2005-20

	Home Win	*Draw*	*Away Win*	*Overall*
R^2	0.98665	0.52008	0.96411	0.83832
RMSE	0.03241	0.20767	0.05166	0.11768

These values show that, for the home and away win models, a very large (> 95%) amount of the variation in the model generated is explained by the consensus probabilities: this is extremely strong[8]. In addition, the RMSE values for the two are very low, around 0.05 for away wins, and 0.03 for home wins. For draws, however, the R^2 is 0.52 and the RMSE is 0.21 indicating that the bookmaker's odds are not as accurate for draws than for clear results. This also explains why our overall R^2 and RMSE values are lower than home and away wins, at 83.8% and 0.12 respectively.

[8] For a perfect correlation, R^2=1.

To futher demonstrate this point, in Figure 2.6, we have a scatter plot of the observed probability v. bookmaker consensus probability using the bins created (N.B. points on the $y = 1.0$ or $= 0.0$ lines are where all or none of the games in the bin resulted in the event), along with the linear models (with a 95% CONFIDENCE INTERVAL around them in light gray) and the $x = y$ line for reference (dashed black).

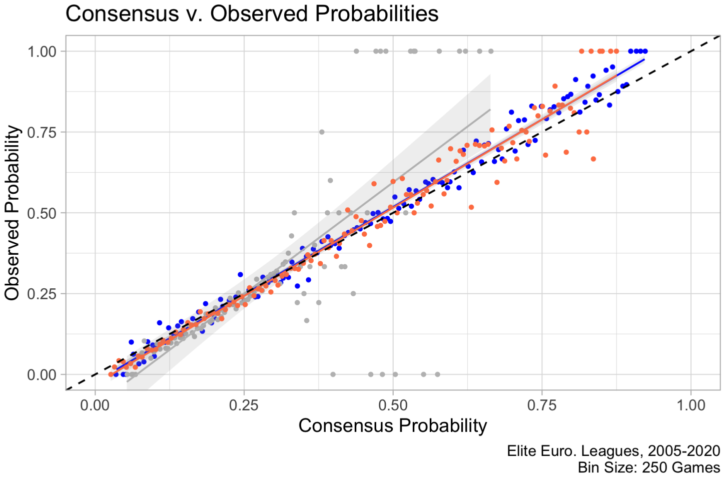

Figure 2.6: Scatter plot of the linear models created

The figure affirms our observations from the values in Table 2.5, and suggests that the bookmakers accuracy for home and away wins is extremely high. Both the 95% confidence intervals are extremely small, and there are no major outliers. Noticable, however, is a group of away win bins with a high consensus probability, but a low observed probability: contextually, this could be down to HOME ADVANTAGE, with a teams performance being poor away from home. This is an area for future research. **Is this a good thing to add or no?** As aforementioned, the bookmaker performance is poor for draws: this is reiterated in the plot, with a large 95% confidence interval. It is worth noting that the accuracy is higher from $\mathbb{P}_{\text{cons}}(\text{Draw}) \in (0, 0.3)$, after which is varies more. The values on the $y = 1$ line—where all games in the bin were draws—are likely to be due to the aforementioned match fixing.

Another feature of the plot is the distinctly large 95% confidence interval—calculated by Equation 2.2 (**Mendenhall13**) —for the draw model. In the equation, $\bar{x}$ is the point estimate, μ is the true value and $Z_{\alpha=0.025}$ is the t statistic as N approaches infinity (following the central limit theorem, we can assume this with a large sample size). This indicates that the STANDARD ERROR ($SE = \frac{s}{\sqrt{N}}$, where s is the point estimate of standard deviation) is large: in fact, the standard error of the bookmaker consensus probability coefficient for each model—overall, home wins, away wins and draws—(found by using the R code `summary(model)` are shown in Table 2.6.

$$\mathbb{P}\bigg(\mu \in \big(\bar{x} \pm Z_{\alpha=0.025} \cdot SE\big)\bigg) = 0.95 \tag{2.2}$$

Table 2.6: Standard error for the model for each outcome.

	Model			
	Home Wins	*Draws*	*Away Wins*	*Overall*
SE	0.011404	0.13835	0.019151	0.02651

We can use the idea of statistical LEVERAGE to investigate whether or not any of our points should be removed. **Here onwards:-** Using R, we can review the influence of certain bins on our model. In R, to do this, we find the HAT values (the statistical theory behind these is explained in Section **Sec**) (**HoaglinWelsch78**). In R, we can use the code below, giving us our Hat values with leverage greater than the critical, given by Equation 2.3, where h_i is the ith diagonal element of the Hat matrix, H; k is the number of independent variables in the equation; and n is the number of values (in our case, bins) **Can I cite a lecture? If so, how?**.

$$h_i > \frac{3 \times (k+1)}{n} = \frac{3 \times 2}{94} = 0.0638 \tag{2.3}$$

```
hats <- as.data.frame(hatvalues(elite.lm.draw))
leverageCrit <- (2*3)/length(orderedhats)
hats[hats$‘hatvalues(elite.lm.draw)‘>leverageCrit, ]
```

The code gives us one value where the leverage is greater than the critical value: `[1] 0.06880155`. Knowing there is one point, we can single it out in a plot, given in Figure 2.7. From this figure, we can see that it is a bin where no draws were seen.

2.4 Predictive performance

We have two measures for calculating short-term predictive performance, P_1 and P_2, defined in Equations 2.4 and 2.5 respectively (**Owen09**), with N being the number of matches in the sample; k being the match number; $\mathbb{P}(O_k)$ being the correct/observed bookmaker probability for match k; and $\mathbb{P}(N_{1k})$, $\mathbb{P}(N_{2k})$ being the two incorrect/not observed bookmaker probabilities for match k: for example, if match m finished in a home win, $\mathbb{P}(O_m) = \mathbb{P}_{\text{cons}}(\text{Home Win})$.

$$P_1 = \exp\left\{\frac{1}{N}\sum_{k=1}^{N}\log_e\big[\mathbb{P}(O_k)\big]\right\}, \quad k \in [1,\ N] \tag{2.4}$$

$$P_2 = \frac{1}{N}\sum_{k=1}^{N}\left\{\big[1-\mathbb{P}(O_k)\big]^2 + \mathbb{P}(N_{1k})^2 + \mathbb{P}(N_{2k})^2\right\}, \quad k \in [1,\ N] \tag{2.5}$$

For better predictive performance, we seek higher P_1 and lower P_2 values, though on their own will not tell us much. We will instead use the values to

Figure 2.7: Leverage Plot for the Draw Linear Model

compare between leagues. To find these values in R, we use a `for` loop to find the correct ($\mathbb{P}(O_k)$) (and log thereof) and the incorrect ($\mathbb{P}(N_{ik})$) probabilities. Then, we can use the code below to compute P_1 and P_2, shown in Table 2.7.

```
P1<-exp( (1/N)*sum(elite$logCorrect) )
P2<-(1/N)*sum( (1-elite$Correct)**2 + (elite$Incorr1)**2 +
    (elite$Incorr2)**2 )
```

Table 2.7: P_1 and P_2 value for all elite leagues.

P_1	0.3788607
P_2	0.5776072

2.5 Comparing leagues

In this section, we will compare the six leagues of our elite group, in turn allowing us to investigate a possible reason for any disparity in bookmaker accuracy between leagues—competitive balance—in Section 2.6. In order to compare the six elite leagues, we will conduct a near-identical correlation analysis as in Section 2.3 to find each league's R^2 and RMSE value and we will find the corresponding P_1 and P_2 value. The only difference in the correlation analysis is that we choose to add a weight to our bins: we bin the more varied home and away wins into 20 bins (around 200 games per bin) and the less varied draws into 5 bins (800). To do this in R, we will use a `for` loop to create a temporary linear model, extract the R^2 and RMSE values, and create a vector with each leagues values. The results (and their respective ranking—R^2 and P_1 decending; RMSE and P_2 ascending—with higher ranking indicating better performance) shown in Table 2.8.

Table 2.8: R^2, RMSE, P_1, P_2 for all elite leagues.

	R^2	RMSE	P_1	P_2	Average Rank
Germany	0.95489 6th	0.05472 6th	0.36958 5th	0.59429 5th	5.50
England	0.96773 5th	0.04783 5th	0.38454 2nd	0.56661 2nd	3.50
Spain	0.96950 4th	0.04711 4th	0.38267 4th	0.57029 3rd	3.75
France	0.97938 3rd	0.03845 3rd	0.36591 6th	0.60248 6th	4.50
Italy	0.98599 1st	0.03297 1st	0.38287 3rd	0.57085 4th	2.25
Portugal	0.98216 2nd	0.03642 2nd	0.38894 1st	0.55963 1st	1.50

Looking at the average ranking, we can see that the bookmakers performed most accurately in the Portuguese and Italian leagues, followed by the English and Spanish leagues, and least accurately in the French and German leagues. We investigate whether this is due to competitive balance in Section 2.6.

2.6 The effect of competitive balance on bookmaker accuracy

2.6.1 What is competitive balance?

Competitive balance is a concept that weighs heavily on economics. It is defined, by the Cambridge Dictionary, as "the situation in which no one business of a group of competing businesses has an unfair advantage over the others," with a monopoly being a situation with no competitive balance (**CompetitiveBalance**).

The nature of competitive balance in baseball that "competitors must be of approximately equal 'size' if any are to be successful; this seems to be a unique attribute of professional competitive sports" (**Rottenberg56**). However, if we look into elite European football, this doesn't hold: in the 2015/16 season, for the English Premier League, Manchester United—who finished fifth—had a turnover of £515m; paid £232m in wages, competed against Watford—finished thirteenth—who had a turnover of £94m; wage bill of £58m (**EplFinance17**). In 2020, Statista published a ranking of the German Bundesliga teams by market (transfer) value. FC Bayern Munich had a value of €875m; DSC Arminia Bielefeld had a value of €47m (**BundesligaFinance20**). This shows that football may not, in terms of competitive balance, follow the same guides that American sport does.

2.6.2 Quantifying competitive balance

We introduce the NAMSI (NAtional Measure of Seasonal Imbalance), shown in equation 2.6, which is, for a league of n teams, the ratio between two standard deviations: σ_{Season} the observed standard deviation of winning percentage for team i, W_i; and $\sigma_{\text{Certainty}}$ the theoretical standard deviation of a certain season, where the team in 1st place wins every game, the team in 2nd wins every game except those against the team in first, ETC, and the team in ith place loses to all teams above them in the table, and defeats all those below them. The NAMSI statistic lies in the range between 0 and 1, with a lower value corresponding to lower seasonal imbalance (**Goossens05**).

$$\text{NAMSI} = \frac{\sigma_{\text{Season}}}{\sigma_{\text{Certainty}}} = \frac{\sqrt{\dfrac{\sum_{i=1}^{n}(W_i - 0.5)^2}{n}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(C_i - 0.5)^2}{n}}} \tag{2.6}$$

NAMSI is a "static measure since it only looks at one season independently of other seasons." In football leagues, a poorly competitively balanced league system would have the same few teams competing for the same places each season. A fluid measure is the Top K ranking—we let $K = 3$.[9] Our statistic is the number of teams entering the top three in three consecutive years, with data from 1963/64 to 2004/05, with a value of 3 showing perfect imbalance, and 9 showing perfect balance, we call this value κ (**Goossens05**).

A further method of quantifying imbalance is the Gini coefficient, G, which is often used to compare wealth inequality between different nations (**GiniWorldCoeffs18**). It measures the ratio of the area between the Lorenz Curve of the country and the $y = x$ (perfect equality) line. The Lorenz Curve is a graphical measure which shows the overall income distribution: on the x axis is the cumulated percent of the population, from poorest to richest; on the y axis is the percent of the total wealth of the country held by this x% (**Lorenz05**).

In Table 2.9, we have the NAMSI, κ and Gini coefficients (**Goossens05**) with the leagues in our analysis ranked (NAMSI, Gini descending; κ ascending: higher ranking indicates less balance in the league), and other top-tier European leagues included for reference.

Our values in Table 2.8 showed that bookmakers performed most accurately in the Portuguese and Italian leagues; Table 2.9 implies the Portuguese is the least competitively balance by all measures. The Italian league has a high NAMSI and low κ value[10], but—contrastingly—a respectively low Gini coefficient. The German and French leagues were where the bookmakers are least accurate: by the NAMSI and κ values, the French Ligue Une was the most competitive league, and by the Gini coefficient, the German Bundesliga was. These

[9] "Because in most European countries it are two or three teams that in general are considered to be dominant. Taking up more teams underrates the dominance since the top 4 and 5 often change," (**Goossens05**).

[10] Indicating low competitive balance

Table 2.9: Data from 1963/64-2004/05 Seasons (**Goossens05**).

	Values					
	NAMSI		κ		Gini	
Germany	0.374	3rd	5.71	4th	0.723	6th
England	0.372	4th	5.79	5th	0.826	3rd
Spain	0.364	5th	5.07	2nd	0.861	2nd
France	0.342	6th	6.00	6th	0.784	4th
Italy	0.418	2nd	5.36	3rd	0.737	5th
Portugal	0.505	1st	4.07	1st	0.898	1st
Greece	0.488	—	4.14	—	0.870	—
The Netherlands	0.494	—	4.36	—	0.888	—
Sweden	0.410	—	6.07	—	0.692	—
Belgium	0.452	—	5.07	—	0.801	—
Denmark	0.412	—	6.43	—	0.581	—

findings indicate that there is a possible link between bookmaker accuracy and competitive balance.

Further, if we consider Figure 2.4—which implies the Spanish, Portuguese and English leagues follow a trimodal distribution; the French, German and Italian leagues follow a unimodal distibution—we can infer a link between competitive balance and the distribution. Portugal's unbalanced Primiera Liga has a trimodal distribution, and higher levels of bookmaker accuracy; both Spain's La Liga and England's Premier League also follow the trimodal distribution, and had the median levels of balance and accuracy. The well-balanced German Bundesliga and French Ligue Une both have a unimodal distribution and lower levels of bookmaker accuracy. The only unexpected result here is the Italian Serie A, with low balance and high accuracy but a unimodal distribution.

A study into the Italian and Spanish football, analysing styles of play (offensive or defensive) showed that whilst the Italian league requires strong defensive efficiency to achieve a high ranking; the contrary is true in Spain, where the "best-rewarded strategy consists in improving offensive efficiency" (**bosca09**). Perhaps, the two distributions are due to the styles of play in those countries. **Here, I could find the average shots per game in each league?**

2.7 Conclusion

To conclude, we have in this chapter shown that the levels of bookmaker accuracy are high: in each of our elite European leagues, the coefficient of determination R^2 was above 95%, with the RMSE below 0.055. The variation of linear model where one uses the bookmaker's odds to predict the actual probability of a result is 95% determined by the bookmaker's odds given, with low errors, on average. From our scatter plot (Figure 2.6), we saw the linear models from home and away wins have a much lower standard error (and therefore a narrower confidence interval) than the model for draws. This, along with the R^2 and RMSE for each result (Table 2.5), indicates whilst bookmakers enjoy high accuracy with clear results, their predictions for draws have considerable room for improvement.

In addition, we have shown that the accuracy is impacted by the competitive balance in each league: bookmakers perform better in countries in unbalanced leagues, such as Portugal's Primiera Liga and Italy's Serie A, than in balanced leagues, Germany's Bundesliga and France's Ligue Une.

Finally, we have located room for future research: there may be a relationship between home advantage and bookmaker's accuracy of away games (perhaps this differs between leagues, too?), and there may also be a relationship between the distribution of bookmaker's odds, and the style of play, in certain leagues, such as the defensive Serie A displaying a unimodal distribution and the offensive La Liga displaying a trimodal distribution.

Appendices

Appendix A

Definitions

Throughout the report, several words have been written in SMALL CAPS: these will be defined in this appendix.

A.1 Mathematical and Statistical

Unless stated, these definitions have been taken from the Fifth Edition of the Oxford Concise Dictionary of Mathematics, published in 2014 (**OxfordMathsDict**).

(i) CENTRAL LIMIT THEOREM — "[T]he distribution of the mean of a sequence of random variables tends to a normal distribution as the number in the sequence increases."

(ii) DISTRIBUTIONS — "[This is] concerned with the way in which the probability of its taking a certain value, or a value within a certain interval, varies. It may be given by the cumulative distribution function[,] its probability mass function [or] its probability density function."

(iii) FIT — "[T]he degree of of correspondence between the observations and the model's predictions."

(iv) KERNEL DENSITY ESTIMATION — ?

(v) ODDS — "...expressed in the form $r : s$ corresponding in theory to a probability of $\frac{r}{r+s}$ of winning."

(vi) PROBABILITY — "...is a measure of the possibility of the event occuring as the result of an experiment." Note that for an event A, the probability of its complement, $\mathbb{P}(A') = 1 - \mathbb{P}(A)$

(vii) SKEWNESS — "The amount of asymmetry of a distribution... If the distribution has a long tail to the left... it is said to be skewed to the left and to have negative skewness."

(viii) STANDARD ERROR — "The standard deviation of an estimator of a population parameter."

A.2 Gambling Terms

(i) Bettor — Someone who places a bet. Also *punter*, *gambler*.

(ii) Bookmaker — Organisation that accepts and pays off bets. Also *House*, *Bookie*.

(iii) Calciopoli — A match fixing scandal which occurred in the 2004/05 and 2005/06 seasons, where Italian Serie A teams (A.C. Milan, Fiorentina, Juventus, Lazio and Reggina) "systematically influenced referees." (**Calciopoli**).

(iv) Competitive Balance — "The situation in which no one business of a group of competing businesses has an unfair advantage over the others." (**CompetitiveBalance**).

(v) Disgrace of Gijón — A football match between West Germany and Austria at the 1982 F.I.F.A. World Cup where a "mutually suitable scoreline" was played out, ensuring both sides progressed to the knockout rounds, leading to the final pair of matches of World Cup group stages being played simultaneously (**DisgraceofGijon**).

(vi) Favourite-Longshot Bias — ?

(vii) Stake — The amount of money placed onto a bet by the bettor (the amount of money *at stake*); *the ante*.

A.3 Acronyms

	Football Associations
F.I.F.A.	Fédération Internationale de Football Association (Global)
U.E.F.A.	the Union of European Football Associations (Europe)
F.A.	the Football Association (England)
D.F.B.	Deutscher Fußball-bund (Germany)
F.F.F.	Fédération Française de Football (France)
F.I.G.C.	Federazione Italiana Giuoco Calcio (Italy)
F.P.F.	Federação Portuguesa de Futebol (Portugal)
R.F.E.F.	Real Federación Española de Fútbol (Spain)
S.F.A.	the Scottish Football Association (Scotland)
	Football Leagues
E.P.L.	English Premier League
S.P.L.	Scottish Premier League

Appendix B

Chapter 2 Code

```
### ANALYSING THE ACCURACY OF BETTING ON ELITE EUROPEAN FOOTBALL
    LEAGUES, 2005-2020
##Set directory, libraries, etc.----
#Setting our directory and clearing the environment
setwd("~/Desktop/University/University Year 3/331MP/Data")
rm(list=ls())
#On all plots, we will set the colours for:
  #Home Win = Blue; Draw = Gray; Away Win = Coral
##LIBRARIES WE NEED
library(car)
library(MASS)
library(ggplot2)
library(gridExtra)

##IDA; Analysis of one season, one league----
#Reading the data:
fr_l1_1617 <-
    read.csv("https://www.football-data.co.uk/mmz4281/1617/F1.csv")
fr_l1_1617 <- fr_l1_1617[,c("Div", "Date", "HomeTeam", "AwayTeam",
    "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD", "BbAvA")]
fr_l1_1617 <- na.omit(fr_l1_1617)

#Finding consensus probabilities (we need to Normalise the inverse odds)
fr_l1_1617$AvgHProbPN <- with(fr_l1_1617, round(1/BbAvH, 4))
fr_l1_1617$AvgDProbPN <- with(fr_l1_1617, round(1/BbAvD, 4))
fr_l1_1617$AvgAProbPN <- with(fr_l1_1617, round(1/BbAvA, 4))
fr_l1_1617$Overround <- with(fr_l1_1617,
    (AvgHProbPN+AvgDProbPN+AvgAProbPN))
fr_l1_1617$AvgHProb <- with(fr_l1_1617, round(AvgHProbPN/Overround,4))
fr_l1_1617$AvgDProb <- with(fr_l1_1617, round(AvgDProbPN/Overround,4))
fr_l1_1617$AvgAProb <- with(fr_l1_1617, round(AvgAProbPN/Overround,4))

##IDA: Simple Calculations -- Means, Std. Devs, Observed Probabilities
cat("FRENCH LIGUE UNE 2016/17:-\nMEANS\nMean Consensus Probability of a
    HOME WIN: "
   ,mean(fr_l1_1617$AvgHProb),
    "\nMean Consensus Probability of an AWAY WIN: ",
        mean(fr_l1_1617$AvgAProb),
    "\nMean Consensus Probability of a DRAW: ",
        mean(fr_l1_1617$AvgDProb))

cat("STANDARD DEVIATIONS\n
    Std Dev for Consensus Probability of a HOME WIN: ",
        sd(fr_l1_1617$AvgHProb),
    "\nStd Dev for Consensus Probability of an AWAY WIN: ",
        sd(fr_l1_1617$AvgAProb),
    "\nStd Dev for Consensus Probability of a DRAW: ",
        sd(fr_l1_1617$AvgDProb))
round(prop.table(table(fr_l1_1617$FTR)), 4)

##Histograms
idaHomeHIST <- ggplot(fr_l1_1617, aes(AvgHProb)) +
    geom_histogram(binwidth=0.05, fill="blue") +
    coord_cartesian(xlim=c(0,1)) + theme_light() + labs(title="Home
    Win", x=NULL, y=NULL)
idaAwayHIST <- ggplot(fr_l1_1617, aes(AvgAProb)) +
    geom_histogram(binwidth=0.05, fill="coral") +
    coord_cartesian(xlim=c(0,1)) + theme_light() + labs(title="Away
    Win", x=NULL, y=NULL)
idaDrawHIST <- ggplot(fr_l1_1617, aes(AvgDProb)) +
    geom_histogram(binwidth=0.05, fill="gray") +
    coord_cartesian(xlim=c(0,1)) + theme_light() + labs(title="Draw",
    x=NULL, y=NULL)
ida.histogram <- grid.arrange(idaHomeHIST, idaDrawHIST, idaAwayHIST,
    nrow=3, ncol=1, left="Frequency", bottom="Consensus Probability",
    top="Histograms of the Consensus Probabilities for each
    outcome\nFrench Ligue 1, 2016-17 Season")
ggsave(path="./writeup/images", filename="elite_01_idahist.png",
    plot=ida.histogram, unit="cm", width=15, height=18) #Saves the
    graph as a file

##EDA and Visual Analysis----
#Reading the data using a For Loop: We first need to define which
    leagues and seasons we are using:
countries <- c("de", "en", "es", "fr", "it", "po")
countries.web <- c("D1", "E0", "SP1", "F1", "I1", "P1") #N.b. The
    Premier League's code is 0; other countries are 1.
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112",
    "1213", "1314", "1415", "1516", "1617", "1718", "1819", "1920")
eliteTemp <- NULL
elite <- NULL

for (i in seasons){
  for (j in 1:6){
    eliteTemp <-
        read.csv(paste0("https://www.football-data.co.uk/mmz4281/", i,
        "/", countries.web[j], ".csv"))
    eliteTemp$Country <- with(eliteTemp, countries[j])
    eliteTemp$Season <- with(eliteTemp, i)
    if (i=="1920"){#Football-data changed the column heading for the
         average odds from the 19/20 season from BbAv_ to Avg_. This
         accounts for that change
      eliteTemp$BbAvH<-eliteTemp$AvgH
      eliteTemp$BbAvA<-eliteTemp$AvgA
      eliteTemp$BbAvD<-eliteTemp$AvgD
    }
    else{}
    eliteTemp <- eliteTemp[ ,c("Div", "Date", "HomeTeam", "AwayTeam",
         "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD", "BbAvA", "Country",
         "Season")]
    elite <- rbind(elite, eliteTemp)
  }
}
elite <- na.omit(elite)

#Pre-Normalised Probabilities
elite$AvgHProbPN <- with(elite, round(1/BbAvH, 4))
elite$AvgDProbPN <- with(elite, round(1/BbAvD, 4))
elite$AvgAProbPN <- with(elite, round(1/BbAvA, 4))
#To normalise them:
elite$overround<-with(elite, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
     #For perfect probabilities -- this should equal 1
elite$AvgHProb <-with(elite, round(AvgHProbPN/overround, 4))
elite$AvgDProb <-with(elite, round(AvgDProbPN/overround, 4))
elite$AvgAProb <-with(elite, round(AvgAProbPN/overround, 4))

#For later analysis, we need these: The 'Correct'/Incorrect Probabilities
N<-nrow(elite)
elite$Correct<-with(elite, rep(0, N))
elite$Incorr1<-with(elite, rep(0, N))
elite$Incorr2<-with(elite, rep(0, N))

for (i in 1:N){
  if ((elite$FTR[i])=="A"){
    elite$Correct[i]<-(elite$Correct[i] + elite$AvgAProb[i])
    elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgDProb[i])
    elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgHProb[i])}
  else if ((elite$FTR[i])=="H"){
    elite$Correct[i]<-(elite$Correct[i] + elite$AvgHProb[i])
    elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgDProb[i])
    elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgAProb[i])}
  else {
    elite$Correct[i]<-(elite$Correct[i] + elite$AvgDProb[i])
    elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgAProb[i])
    elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgHProb[i])}
}

elite$logCorrect<-with(elite, rep(0,N))
for (j in 1:N){
  elite$logCorrect[j]<-log(elite$Correct[j], base=exp(1))
}

#Also, we will need smaller datasets for each country:
```

```
elite.de<-elite[elite$Country=="de",]
elite.en<-elite[elite$Country=="en",]
elite.es<-elite[elite$Country=="es",]
elite.fr<-elite[elite$Country=="fr",]
elite.it<-elite[elite$Country=="it",]
elite.po<-elite[elite$Country=="po",]
##CALCULATIONS
cat("ELITE LEAGUES, 05-20:-\nMEANS\nMean Consensus Probability of a HOME
    WIN: "
    ,mean(elite$AvgHProb),
    "\nMean Consensus Probability of an AWAY WIN: ",
        mean(elite$AvgAProb),
    "\nMean Consensus Probability of a DRAW: ", mean(elite$AvgDProb))

cat("STANDARD DEVIATIONS\n
    Std Dev for Consensus Probability of a HOME WIN: ",
        sd(elite$AvgHProb),
    "\nStd Dev for Consensus Probability of an AWAY WIN: ",
        sd(elite$AvgAProb),
    "\nStd Dev for Consensus Probability of a DRAW: ",
        sd(elite$AvgDProb))
round(prop.table(table(elite$FTR)), 4)
##PLOTS
##Boxplots
#This will see simply if there is a significant different between
    outcomes and the consensus probabilities (it doesn't)
  bp.home<-ggplot(elite, aes(x=FTR, y=AvgHProb)) +
      geom_boxplot(outlier.size=0.75,outlier.alpha=0.7, color="blue") +
      theme_light() + stat_boxplot(coef=5) + labs(x="Actual Result",
      y="Consensus Probability of a Home Win", caption="") +
      coord_cartesian(ylim=c(0,1))
  bp.draw<-ggplot(elite, aes(x=FTR, y=AvgDProb)) +
      geom_boxplot(outlier.size=0.75, outlier.alpha=0.7, color="gray")
      + theme_light() + stat_boxplot(coef=5) + labs(x="Actual Result",
      y="Consensus Probability of a Draw", caption="") +
      coord_cartesian(ylim=c(0,1))
  bp.away<-ggplot(elite, aes(x=FTR, y=AvgAProb)) +
      geom_boxplot(outlier.size=0.75,outlier.alpha=0.7, color="coral")
      + theme_light() + stat_boxplot(coef=5) + labs(x="Actual Result",
      y="Consensus Probability of an Away Win", caption="Elite European
      Leagues, 2005-2020") + coord_cartesian(ylim=c(0,1))
  eda.bp.all<-grid.arrange(bp.home, bp.draw, bp.away, nrow=1, ncol=3,
      top="Boxplot of the Consensus Probabilites offered for each
      outcome\nagainst Actual (Observed) Result")
  ggsave(path = "./writeup/images", filename = "elite_02_boxplot.png",
      plot = eda.bp.all, unit="cm", width=15, height=10)

##Density Plots
#Instead of a binned histogram, we will instead use a density Plot,
    which is essentially a smoothed histogram:-
  #All matches; not split by league
  eda.wins.dens.all<-ggplot(elite, aes(x=AvgHProb)) +
    geom_density(color="blue") +
   geom_density(data=elite, mapping=aes(x=AvgAProb), color="coral",
        show.legend = TRUE) +
   coord_cartesian(xlim=c(0,1)) +
   labs(title = "Home and Away Wins",
        caption = "Elite Leagues, 2005-2020",
        x = "Consensus Probability", y = "density") +
   theme_light()
  eda.draw.dens.all<-ggplot(elite, aes(x=AvgDProb)) +
   geom_density(color="gray") +
   coord_cartesian(xlim=c(0,1)) +
   labs(title = "Draws",
        caption="",
        x = "Consensus Probability", y = "density") +
   theme_light()
  eda.density.all<-grid.arrange(eda.wins.dens.all, eda.draw.dens.all,
      nrow=1, ncol=2, left = "", bottom = "", top = "Density Plots for
      Consensus Probability of Each Outcome")
  ggsave(path = "./writeup/images", filename =
      "elite_02_edadensall.png", plot=eda.density.all, unit="cm",
      width=15, height=7)

#Splitting these by league:
  #Home Wins
  eda.home.dens<-ggplot(elite, aes(x=AvgHProb, color=Country)) +
      geom_density() + coord_cartesian(xlim=c(0,1)) +
   labs(title="Home Win", x=NULL, y=NULL) +
        geom_vline(aes(xintercept=mean(AvgHProb)), linetype="dashed",
        size=0.3) + guides(y="none") + theme(legend.position="none")
  #Away Wins
  eda.away.dens<-ggplot(elite, aes(x=AvgAProb, color=Country)) +
      geom_density() + coord_cartesian(xlim=c(0,1))+
   labs(title="Away Win", x=NULL, y=NULL) +
        geom_vline(aes(xintercept=mean(AvgAProb)), linetype="dashed",
        size=0.3) + guides(y="none") + theme(legend.position="none")
  #Draws
  eda.draw.dens<-ggplot(elite, aes(x=AvgDProb, color=Country)) +
      geom_density() + coord_cartesian(xlim=c(0,0.8))+
   labs(title="Draw", x=NULL, y=NULL) +
        geom_vline(aes(xintercept=mean(AvgDProb)), linetype="dashed",
        size=0.3) + guides(y="none") + scale_colour_discrete(labels =
        c("Germany","England","Spain","France","Italy","Portugal"))
  eda.density<-grid.arrange(eda.home.dens, eda.draw.dens, eda.away.dens,
      nrow=3, ncol=1, left="", bottom="Consensus Probability",
      top="Density Plots for the Consensus Probability for each
      outcome\nElite European Leagues, 2005-2020")
  ggsave(path = "./writeup/images", filename = "elite_02_edadens.png",
      plot=eda.density, unit="cm", width=15, height=18)


#We are going to create a Tile Plot to see when the bookmakers are most
   accurate.
#We will be using the Correct column from S4, but neither of the
    Incorrect ones.
```

```
#For ease, we will encode that any match where team A scores 5+ goals
    will be counted as 5 for the tile plot
elite$FTHG.Tile<-with(elite,rep(0,N))
elite$FTAG.Tile<-with(elite,rep(0,N))
for (k in 1:N){
  if ((elite$FTHG[k])>=5){elite$FTHG.Tile[k]<-5}
  else{elite$FTHG.Tile[k]<-elite$FTHG[k]}}
for (k in 1:N){
  if ((elite$FTAG[k])>=5){elite$FTAG.Tile[k]<-5}
  else{elite$FTAG.Tile[k]<-elite$FTAG[k]}}

elite.tile<-ggplot(elite, aes(y=FTAG.Tile, x=FTHG.Tile)) +
  geom_tile(aes(fill = Correct)) +
  scale_fill_distiller(palette = "Greens", direction = 1,
      name="Correct\nProbability") +
  theme_light() +
  labs(title="Match Result v. Correct Consensus Probability", x="Home
      Goals Scored", y="Away Goals Scored", caption="Elite European
      Leagues, 2005-2020") +
  scale_y_discrete(limits=factor(c(1:4, "5+"))) +
  scale_x_discrete(limits=factor(c(1:4, "5+"))) +
  geom_abline(intercept=0, slope=1) +
  coord_cartesian(xlim=c(0,5), ylim=c(0,5))
ggsave(path="./writeup/images", filename="elite_05_tile.png",
    plot=elite.tile, unit="cm", width=15, height=15)

#Bin Sizes--Whilst we can’t overlay this onto the graph, it is important
    to calculate
  binSizeTemp<-0
  binSize<-c(NULL)
  for (i in 0:5){
    for (j in 0:5){
      binSizeTemp<-nrow(elite[elite$FTHG.Tile==i & elite$FTAG.Tile==j,])
      binSize<-c(binSize, binSizeTemp)
    }
  }
  binSize<-matrix(binSize, nrow=6)
  rownames(binSize)<-c("AG, 0", 1:4, "5+") #Rows = AWAY GOALS
  colnames(binSize)<-c("HG, 0", 1:4, "5+") #Cols = HOME GOALS
  binSize<-as.table(binSize)
  binSize

##Correlation Analysis----
#For each outcome, we will group a number of games with that outcome,
    and see the actual probability, v. the mean consensus probability
    in that bin:
#This is the data that will be plotted for HOME WINS
elite$AvgHProb.cut <- cut(elite$AvgHProb, 124, inclues.lowest=TRUE)
#First, we cut the data into 100 ’bins’
levels(elite$AvgHProb.cut) <- tapply(elite$AvgHProb, elite$AvgHProb.cut,
    mean)
#Tapply finds the mean of the bin, rather than taking the midpoint
elite.observed.probabilites.TabH <- prop.table(table(elite$FTR,
    elite$AvgHProb.cut), 2)[c(1, 2, 3),] #The latter will remove an
    extra row--sometimes a blank row stays by mistake
elite.observed.probabilites.H <- elite.observed.probabilites.TabH[3,] #1
    = Away; 2 = Draw; 3 = Home (alphabetic)
elite.bookmaker.probabilites.H <-
    as.numeric(names(elite.observed.probabilites.H))
#AWAY WINS
elite$AvgAProb.cut <- cut(elite$AvgAProb, 124, inclues.lowest=TRUE)
levels(elite$AvgAProb.cut) <- tapply(elite$AvgAProb, elite$AvgAProb.cut,
    mean)
elite.observed.probabilites.TabA <- prop.table(table(elite$FTR,
    elite$AvgAProb.cut), 2)[c(1, 2, 3), ]
elite.observed.probabilites.A <- elite.observed.probabilites.TabA[1, ]
elite.bookmaker.probabilites.A <-
    as.numeric(names(elite.observed.probabilites.A))
#DRAWS
elite$AvgDProb.cut <- cut(elite$AvgDProb, 124, inclues.lowest=TRUE)
levels(elite$AvgDProb.cut) <- tapply(elite$AvgDProb, elite$AvgDProb.cut,
    mean)
elite.observed.probabilites.TabD <- prop.table(table(elite$FTR,
    elite$AvgDProb.cut), 2)[c(1, 2, 3), ]
elite.observed.probabilites.D <- elite.observed.probabilites.TabD[2, ]
elite.bookmaker.probabilites.D <-
    as.numeric(names(elite.observed.probabilites.D))

#Calculating R squared and RMSE:
#First, we will concatenate the three outcomes’ bookmaker v. observed
    probabilities to find an overall R^2 and RMSE
elite.bookmaker.probabilities <- c(elite.bookmaker.probabilites.H,
    elite.bookmaker.probabilites.D, elite.bookmaker.probabilites.A)
elite.observed.probabilities <- c(elite.observed.probabilites.H,
    elite.observed.probabilites.D, elite.observed.probabilites.A)

#Home Wins
elite.lm.home <- lm(elite.observed.probabilites.H ~
    elite.bookmaker.probabilites.H) #Creates the linear model
round(summary(elite.lm.home)$r.squared, 5) #R Squared
round(sqrt(mean(elite.lm.home$residuals^2)), 5) #RMSE

#Away Wins
elite.lm.away <- lm(elite.observed.probabilites.A ~
    elite.bookmaker.probabilites.A)
round(summary(elite.lm.away)$r.squared, 5) #R Squared
round(sqrt(mean(elite.lm.away$residuals^2)), 5) #RMSE

#Draws
elite.lm.draw<-lm(elite.observed.probabilites.D ~
    elite.bookmaker.probabilites.D)
round(summary(elite.lm.draw)$r.squared, 5) #R Squared
round(sqrt(mean(elite.lm.draw$residuals^2)),5) #RMSE

#OVERALL (General trend)
elite.linear.model<-lm(elite.observed.probabilities ~
    elite.bookmaker.probabilities)
elite.rsqu <- round(summary(elite.linear.model)$r.squared, 5) #R Squared
elite.rmse <- round(sqrt(mean(elite.linear.model$residuals^2)), 5) #RMSE

#Our final plot:-
elite.scatter<-ggplot(data=NULL,aes()) + geom_smooth() +
  geom_jitter(aes(x=elite.bookmaker.probabilites.H,
      y=elite.observed.probabilites.H), col="blue", size=0.75,
      show.legend=TRUE) +
  geom_smooth(aes(x=elite.bookmaker.probabilites.H,
      y=elite.observed.probabilites.H), col="blue", method=lm,
      alpha=.15, size=0.5) +
  geom_jitter(aes(x=elite.bookmaker.probabilites.D,
      y=elite.observed.probabilites.D), col="gray", size=0.75,
      show.legend=TRUE) +
  geom_smooth(aes(x=elite.bookmaker.probabilites.D,
      y=elite.observed.probabilites.D), col="gray", method=lm,
      alpha=.15, size=0.5) +
  geom_jitter(aes(x=elite.bookmaker.probabilites.A,
      y=elite.observed.probabilites.A), col="coral", size=0.75,
      show.legend = TRUE) +
  geom_smooth(aes(x=elite.bookmaker.probabilites.A,
      y=elite.observed.probabilites.A), col="coral", method=lm,
      alpha=.15, size=0.5) +
  geom_abline(intercept = 0, slope = 1, linetype="dashed") +
  labs(title="Consensus v. Observed Probabilities", x="Consensus
      Probability", y="Observed Probability", caption="Elite European
      Leagues, 2005-2020\nBin Size: 250 Games") +
  coord_cartesian(xlim=c(0, 1), ylim=c(0, 1)) + theme_light()
ggsave(path = "./writeup/images", filename = "elite_03_scatter.png",
    plot=elite.scatter, unit="cm", width=15, height=10)

##Hat values (Statistical Leverage)----
hats <- as.data.frame(hatvalues(elite.lm.draw))
leverageCrit <- (2*3)/nrow(hats)
hats[hats$‘hatvalues(elite.lm.draw)‘>leverageCrit,]
drawLevPlot<-leveragePlot(elite.lm.draw,elite.bookmaker.probabilites.D,col="gray",
                       id =
                           list(method=list(abs(residuals(elite.lm.draw,
                           type="pearson"))), n=1, cex=1, col="red",
                           location="lr"),
                       xlab = "Consensus Probability (Draw) | Others",
                           ylab = "Observed Probability (Draw) |
                           Others")
#png(path = "./writeup/images", filename = "elite_03_drawLeverage.png",
    plot=drawLevPlot, units="cm", width=20, height=10, res=450)

##Overall P1 and P2 values----
#Calculating Owen (2009)'s P1 and P2 values:
#Overall:
#P1 is defined as: = exp{ 1/N * sum^N_k=1 ( log_e [ P(Ok) ] ) } where k
    is the match; Ok is the correct probability
P1<-exp( (1/N)*sum(elite$logCorrect) )
#P2 is defined as: = 1/N * sum^N_k=1 { (1-P(Ok))^2 + P(NO1)^2 + P(NO2)^2
    } where NO1 and NO2 are the incorrect probabilities
```

```
P2<-(1/N)*sum( (1-elite$Correct)**2 + (elite$Incorr1)**2 +
    (elite$Incorr2)**2 )


##Linear Models for each country----
RSqu.H <- NULL
RMSE.H <- NULL
RSqu.D <- NULL
RMSE.D <- NULL
RSqu.A <- NULL
RMSE.A <- NULL
RSqu.O <- NULL
RMSE.O <- NULL
P1.split <- NULL
P2.split <- NULL
for(i in countries){
  ##Pre Model
  modelTempData <- elite[elite$Country==i, ] #Splits elite dataset into
      leagues
  #Cuts the average probabilities into 10 groups (around 400 games)
  modelTempData$AvgHProb.cut <- cut(modelTempData$AvgHProb, 20,
      include.lowest=TRUE)
  modelTempData$AvgDProb.cut <- cut(modelTempData$AvgDProb, 5,
      include.lowest=TRUE)
  modelTempData$AvgAProb.cut <- cut(modelTempData$AvgAProb, 20,
      include.lowest=TRUE)
  #Finds the mean of each group (cut)
  levels(modelTempData$AvgHProb.cut) <- tapply(modelTempData$AvgHProb,
      modelTempData$AvgHProb.cut, mean)
  levels(modelTempData$AvgDProb.cut) <- tapply(modelTempData$AvgDProb,
      modelTempData$AvgDProb.cut, mean)
  levels(modelTempData$AvgAProb.cut) <- tapply(modelTempData$AvgAProb,
      modelTempData$AvgAProb.cut, mean)
  #Finds the observed probability for each cut
  modelTemp.obs.prob.tabH <- prop.table(table(modelTempData$FTR,
      modelTempData$AvgHProb.cut), 2)[c(1, 2, 3), ]
  modelTemp.obs.prob.H <- modelTemp.obs.prob.tabH[3, ]
  modelTemp.obs.prob.tabD <- prop.table(table(modelTempData$FTR,
      modelTempData$AvgDProb.cut), 2)[c(1, 2, 3), ]
  modelTemp.obs.prob.D <- modelTemp.obs.prob.tabD[2, ]
  modelTemp.obs.prob.tabA <- prop.table(table(modelTempData$FTR,
      modelTempData$AvgAProb.cut), 2)[c(1, 2, 3), ]
  modelTemp.obs.prob.A <- modelTemp.obs.prob.tabA[1, ]
  #Finds the bookmaker probabilities for each group and creates vectors
  modelTemp.boo.prob.H <- as.numeric(names(modelTemp.obs.prob.H))
  modelTemp.boo.prob.D <- as.numeric(names(modelTemp.obs.prob.D))
  modelTemp.boo.prob.A <- as.numeric(names(modelTemp.obs.prob.A))
  modelTemp.bookmakers <- c(modelTemp.boo.prob.H, modelTemp.boo.prob.D,
      modelTemp.boo.prob.A)
  modelTemp.observed <- c(modelTemp.obs.prob.H, modelTemp.obs.prob.D,
      modelTemp.obs.prob.A)
  ##Making the model
  modelTempH <- lm(modelTemp.obs.prob.H˜modelTemp.boo.prob.H)
  modelTempD <- lm(modelTemp.obs.prob.D˜modelTemp.boo.prob.D)
  modelTempA <- lm(modelTemp.obs.prob.A~modelTemp.boo.prob.A)
  modelTempO <- lm(modelTemp.observed~modelTemp.bookmakers)
  #Finding the R2 and RMSE values
  RSqu.H <- c(RSqu.H, round(summary(modelTempH)$r.squared, 5))
  RMSE.H <- c(RMSE.H, round(sqrt(mean(modelTempH$residuals^2)), 5))
  RSqu.D <- c(RSqu.D, round(summary(modelTempD)$r.squared, 5))
  RMSE.D <- c(RMSE.D, round(sqrt(mean(modelTempD$residuals^2)), 5))
  RSqu.A <- c(RSqu.A, round(summary(modelTempA)$r.squared, 5))
  RMSE.A <- c(RMSE.A, round(sqrt(mean(modelTempA$residuals^2)), 5))
  RSqu.O <- c(RSqu.O, round(summary(modelTempO)$r.squared, 5))
  RMSE.O <- c(RMSE.O, round(sqrt(mean(modelTempO$residuals^2)), 5))
  #Finding P1 and P2
  P1.split <- c(P1.split, round(exp(
      (1/(nrow(modelTempData)))*sum(modelTempData$logCorrect) ),5) )
  P2.split <- c(P2.split, round((1/(nrow(modelTempData)))*sum(
      (1-modelTempData$Correct)**2 + (modelTempData$Incorr1)**2 +
      (modelTempData$Incorr2)**2 ),5) )
}
#Putting this into an easy-to-see Table:
league.values <- matrix(c(RSqu.H, RMSE.H, RSqu.D, RMSE.D, RSqu.A,
    RMSE.A, RSqu.O, RMSE.O, P1.split, P2.split), ncol=6, byrow=TRUE)
rownames(league.values) <- c("RSqu.H", "RMSE.H", "RSqu.D", "RMSE.D",
    "RSqu.A", "RMSE.A", "RSqu.O", "RMSE.O", "P1.split", "P2.split")
colnames(league.values) <- countries
league.values <- as.table(league.values)
league.values

##Competitive Balance----
#We first define the statistics (Gini, NAMSI and K) from Goossens (05):
namsi <- c(0.374, 0.372, 0.364, 0.342, 0.418, 0.505)
kappa <- c(5.71, 5.79, 5.07, 6.00, 5.36, 4.07)
gini  <- c(0.723, 0.826, 0.861, 0.784, 0.737, 0.898)
comp.bal.stats <- matrix(c(namsi, kappa, gini), ncol=3, byrow=FALSE)
colnames(comp.bal.stats) <- c("namsi", "kappa", "gini")
rownames(comp.bal.stats) <- countries
comp.bal.stats <- as.table(comp.bal.stats)
#Defining our accuracy stats we will use:
accuracy.stats <- matrix(c(RSqu.O, RMSE.O, P1.split, P2.split), ncol=4,
    byrow=FALSE)
colnames(accuracy.stats) <- c("rsq", "rmse", "p1", "p2")
rownames(accuracy.stats) <- countries
accuracy.stats <- as.table(accuracy.stats)
#Creating a new data frame:
compbal <- cbind(accuracy.stats, comp.bal.stats)
compbal <- data.frame(compbal)
```