Stockholms universitet

A COMPARISON BETWEEN STEP-WISE REGRESSION MODELS TO PREDICT FOOTBALL GAMES

Natanael Blomberg

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2022:6
Matematisk statistik
Juni 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Mathematical Statistics
Stockholm University
Bachelor Thesis **2022:6**
http://www.math.su.se

A COMPARISON BETWEEN STEPWISE REGRESSION MODELS TO PREDICT FOOTBALL GAMES

Natanael Blomberg*

June 2022

Abstract

This study analyzes and compares which variables have the most significant impact on football games. Four models have been created using linear regression and different stepwise regression procedures. The methods used for comparison are backward elimination, forward selection, AIC backward elimination, and AIC forward selection. Additionally, various transformations have been applied, such as a Yeo-Johnson transformation and others, to see if the model could be improved to be more accurate. The model was then used to predict the outcome and analyze which stepwise regression models gave the best prediction. It was discovered that most of the response variables already described the explanatory variable so well that no transformation was needed. Through various prediction tests and measurements, it was concluded that AIC backward elimination and AIC forward selection provided the best models for predicting the outcome of football matches.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: natanaelblomberg@gmail.com. Supervisor: Ola Hössjer & Nils Engler.

Preface

I want to thank my supervisors, Ola Hössjer and Nils Engler, for their guidance and support during my bachelor thesis. I also want to thank my family, but above all, my grandfather and father, who showed great interest in my studies and helped me through tough times.

Contents

1 Introduction

Today, there are about 240 million active football players in 200 different countries. So it is safe to say that football has, at some point, been a part of everyone's life. Either you have a relative that has played, or you have played or watched a football game yourself.

This thesis is about predicting the outcome of football games. This is done using regression analysis. The most common parameters are investigated to understand how they impact the final result. Stepwise regression was chosen as the statistical technique used to predict the outcome of the response variable. The thesis starts by verifying if there is any multicollinearity or correlation between the explanatory variables by applying the variance inflation factor (VIF) method. Then, the status of the correlation between the explanatory variables is analyzed using a correlation plot. A correlation plot lets you view and understand the pairwise relationship between the variables in the dataset. The purpose is to find the best model for predicting the outcome of football games. Different types of transformations and stepwise regression techniques will be used. The stepwise regression techniques implemented are backward elimination, forward selection, Akaike's information criterion (AIC) with backward elimination, and AIC with forward selection. When the best model is found for the different stepwise regression models, this will be used to predict all games during an entire premier league season. When all games have been investigated, points are handed out for the teams based on the output, and a results table containing all the teams' total points is created. A comparison between the predicted outcome table to the actual outcome table for that particular season was made. Based on this information, a conclusion is made as to which stepwise regression technique and model gives the best prediction.

There are a lot of different parameters that may impact the final result, such as shots on target, ball possession, corners, and much more. Hence, it would be precious for players, coaches, and teams to have this information. Many teams may think they have a game plan, but applying regression analysis can determine which aspects are the most important. This information will not just be highly relevant to active footballers but also for people betting and commentating on football since this gives a better understanding of the game.

The thesis is divided into five main sections. The *Introduction* summarizes this thesis' methods and main results. The *Theory* section consists of the mathematical models and tools utilized. The *Data* section deals with the dataset and eventual transformations that are used. The *Analysis and Results* section contains calculations and implementations of the theories.

The *Discussion* section discusses the results and explains what has been accomplished in the thesis.

2 Theory

2.1 Linear regression

When applying the simplest form of linear regression, certain assumptions must be met to keep processing the data. These assumptions include:

- The observations must be independent
- The residuals in the model must follow a normal distribution
- The variance in the residuals must be constant.

2.1.1 Simple linear regression

Simple linear regression is defined as follows (Alm & Britton, 2008, Ch. 9.2, page 424):

Let $x_1, ..., x_n$ be given (non-random) quantities. Assume further that $Y_1, ...Y_n$ are independent random variables with common variance σ^2 and that $\mu_i = E(Y_i)$ is given by $\mu_i = \alpha + \beta x_i$. Then a simple linear regression model has been obtained. The quantity x is called the regressor or explanatory variable and the random variable Y (or its observed value y) is the response variable. The "expected value line" $y = \alpha + \beta x$ is called the theoretical regression line, α is called the intercept and β is the slope coefficient.

An alternative way of defining the model above are as follows:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

where $\epsilon_i, i = 1, ..., n$, is independent with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$.

2.1.2 Multiple linear regression

If it were to be a case that several explanatory variables could affect the response variable, then a multiple linear regression model can be applied. This model is defined as follows (Alm & Britton, 2008, Ch. 9.3, page 442):

Let $x = (x_1, ..., x_k)$ be given (non-random) vectors of dimension k ($x_i = (x_{1i}, ..., x_{ki})$). Assume further that $Y_1, ..., Y_n$ are independent random variables with common variance σ^2 and that $\mu_i = E(Y_i)$ is given by $\mu_i = \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki}$. Then a multiple linear regression model was obtained.

An alternative way of defining the model above is as follows:

$$Y_i = \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \epsilon_i, \quad i = 1, ..., n, \tag{2}$$

where $\epsilon_i, i = 1, ..., n$, are independent with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$.

2.1.3 Estimates

The unknown parameters α and β can be estimated by applying the least-squares method to data. This means minimizing the sum of the squares of the residuals (Andersson et al., 2019, Ch 2):

$$\sum_{i=1}^{n} (Y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_k x_{ki})^2 \tag{3}$$

Put $\bar{Y} = \sum_i Y_i/n$ and $\bar{x}_{j\cdot} = \sum_i x_{ji}/n$ for sample averages of the observations and the j:th predictor variable respectively. Let $\tilde{\alpha} = \alpha + \bar{x}_{j\cdot}\beta$ be the centered intercept, $X = (x_1^T, \ldots, x_n^T)^T$ the design matrix, $S = X^T X$, and $Y = (Y_1, \ldots, Y_n)^T$ the observational vector. According to Sundberg (2021, Ch 3.2, page 80), the following estimates for the multiple regression model have been acquired:

$$\hat{\beta} = S^{-1} X^T Y, \quad \hat{\beta} \sim N(\beta, \sigma^2 S^{-1}) \tag{4}$$

$$\hat{\tilde{\alpha}} = \bar{Y}, \quad \hat{\tilde{\alpha}} \sim N(\tilde{\alpha}, \sigma^2/N) \tag{5}$$

$$\hat{\alpha} = \hat{\tilde{\alpha}} - \sum_j \hat{\beta}_j \bar{x}_{j\cdot}, \tag{6}$$

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (Y_i - \bar{Y} - \sum_{j=1}^{k} \hat{\beta}_j (x_{ij} - \bar{x}_{j\cdot}))^2, \tag{7}$$

$$(n-k-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-k-1), \tag{8}$$

where $N(\mu, \sigma^2)$ refers to a normal distribution with mean μ and variance σ^2, whereas $\chi^2(f)$ is a chisquare distribution with f degrees of freedom.

2.2 Hypothesis testing

It is important to know how much the explanatory variables $x_{1i}, ..., x_{ki}$ affect the response variable Y_i. This means that a hypothesis test must be formulated, or a so-called t-test for β_j where $j = 1, ..., k$ and k is the number of regression parameters. For instance, suppose the following null hypothesis is to be tested (Alm & Britton, 2008, Ch 7.4):

$$H_0 : \theta = \theta_0 \tag{9}$$

against the alternative hypothesis

$$H_1 : \theta \neq \theta_0 \tag{10}$$

If the test confirms a systematic departure from H_0, then H_0 is rejected in favor of H_1. It is then usually said that θ is significantly different from θ_0.

In Ninorta Malki's bachelor thesis (Malki, 2022) a t-test is described as follows: Suppose $Z \sim N(0, \sigma^2 I_k)$, where I_k is an identity matrix of order k. Then $\hat{\beta} = \beta + (X^T X)^{-1} X^T Z$. The conclusion is then drawn that $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$. This further leads to the test statistic

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{V}(\hat{\beta}_i)}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{c_{ii}}}, \tag{11}$$

where c_{ii} is the j:th diagonal element of $(X^T X)^{-1}$. Finally we have have the test statistic

$$T = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{((X^T X)^{-1})_{ii}}} \sim t(n-k), \tag{12}$$

where $t(f)$ is a t-distribution with f degrees of freedom. Since σ^2 is rarely known, the estimated value $\hat{\sigma}^2$ of equation (7) is used. Due to the null hypothesis, this is a two-sided test. There will be a rejection of the null hypothesis when $|T| \geq t_{\alpha/2}(n-k)$ at the significance level α, with $t_\alpha(f)$ the $1-\alpha$ quantile of a t distribution with f degrees of freedom. This means that the p-value $P(|T| \geq |t||H_0)$ is the probability that the observed value t of T is at least as extreme of an outcome as the sample obtained when the null hypothesis is true.

2.3 Adjusted measures

2.3.1 Coefficient of determination

The most common model adaptation measures in connection with linear models in general and multiple regression models, in particular, is the so-called degree of explanation R^2. The degree of explanation can be defined as the proportion of the total variation that the model "explains." If the total variation, as usual, is indicated by the total sum of squares (TSS), the (total squared variation around the total mean), then the proportion of explained variations is the sum of squares ratio

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}, \tag{13}$$

where RSS is the residual sum of squares of equation (7). When adding an x-variable to a model, R^2 always increases. To see if it pays off to add a

variable, one may look at whether $\hat{\sigma}^2$ decreases. If this is the case, it can then be interpreted as if there is less unexplained variation left in the model. One such measure is what is called the adjusted R^2, from now on referred to as R^2_{adj}. R^2_{adj} measures precisely how much variance reduction is achieved in the current model. An explicit formula is

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2_0}, \tag{14}$$

where $\hat{\sigma}^2_0 = TSS/(n-1)$ is the σ^2-estimate when one does not have any x-variable in the model (Sundberg, 2021, Ch. 3.2, page 89).

2.4 Multicollinearity

2.4.1 Variance inflation factor

The variance inflation factor, also called VIF, is the variance inflation which means that collinear variables can be strongly significant together in one model but individually non-significant in the same model. This measure is first described by looking at $Var(\hat{\beta}_j)$ for a single regression coefficient β_j. If S is invertible with inverse S^{-1}, then $Var(\hat{\beta}_j)$ is expressed as follows:

$$Var(\hat{\beta}_j) = \sigma^2 (S^{-1})_{jj} = \frac{\sigma^2}{s_{jj}} \text{VIF}, \tag{15}$$

where s_{jj} is the j:th diagonal element of S. Here the first of the two factors is the variance of $\hat{\beta}_j$ one would have in a simple linear regression of y on x_j, or if x_j had been orthogonal to all the other $x-$variables and the second terms is the VIF factor. The VIF factor can be calculated as

$$\text{VIF} = \frac{1}{1 - R^2_j}, \tag{16}$$

where R^2_j stands for the degree of explanation that indicates how much the other $x-$variables explain the variation in x_j. The VIF factor thus expresses according to (16)-(17) how much larger the variance of the least-squares estimate $\hat{\beta}_j$ is in the presence of the other regression terms than it would have been if the variable x_j had been the only predictor variable or at least would be orthogonal to the others (Sundberg, 2021, Ch 3.2, page 94).

2.4.2 Correlation

Assume that two explanatory variables of a system have the measured values x_i and y_i, $i = 1, 2, ..., n$. The correlation coefficient is defined as

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}, \tag{17}$$

where $\bar{x}$ and $\bar{y}$ are the means of x_i and y_i, respectively.

According to Cauchy-Schwarz' inequality one has $-1 \leq r_{xy} \leq 1$ and there is a linear relationship between the variables if and only if $r_{xy} = \pm 1$ (Tambour, 2022).

2.5 Variable selection methods

Regression analysis often has a large set of potential explanatory variables. On the other hand, problems often arise, such as how many and which of them you should include in your regression model. To solve this problem, several stepwise procedures are in use. This means that one successively either increase or decrease the number of variables, one at a time, until a stop criterion is met (Sundberg, 2021, Ch. 3.2, page 91).

The different types of stepwise variable selection models that will be used in this thesis will now be explained.

2.5.1 Backward elimination

The backward elimination procedure is based on the model that includes all k variables. One variable at a time is excluded until the procedure stops. In each step, the hypothesis $\beta_j = 0$ is tested by means of a t-test (11) for all the remaining x-variables x_j. For this thesis, the most classical stopping criterion will be used. All remaining parameters β_j are individually different from zero when tested at a pre-selected significance level. If one or more variables are not significant, the variable with the highest p-value is eliminated. The same applies for the variable whose exclusion gives the highest coefficient of determination R^2 (Sundberg, 2021, Ch 3.2, page 92).

2.5.2 Forward selection

This class of procedures starts with the model completely without x-variables, ie $\mu_i = \alpha$ for $i = 1, ..., n$. The model is extended by adding one variable at a time. In each step, the current model consists of those explanatory variables that have been included so far. The method then includes, by means of a t-test, the "most significant" variables outside the model as long as any such variable is significant on a pre-specified significance level (Sundberg, 2021, Ch 3.2, page 93).

2.5.3 Akaike information criterion

Akaike's Information Criterion (AIC) measures how well a model fits the data set without adding too many explanatory variables. The model with the lowest AIC score is the most suitable model to use. For example, let

ψ denote the entire parameter vector. Then AIC is defined by (Sundberg, 2021, Ch 6.8, page 275):

$$AIC = -2\log L(\hat{\psi}) + 2\dim(\psi), \tag{18}$$

where L is the likelihood function. In this thesis, AIC is combined with backward elimination and forward selection, where variables will be added or removed precisely as described above in sections 2.5.1-2.5.2.

2.6 Transformations

In section 2.1 on linear regression, some assumptions have been described that must be met to keep processing the data. However, problems can arise if these assumptions are not met. One example is that the functional relation between explanatory and response variable is not linear. In some cases, these issues can be solved by transforming either the response variable or the explanatory variables. Still, in some cases, a transformation of both the response variable and the explanatory variables may be needed.

The transformations used in the thesis, as well as their application will now be described.

2.6.1 Log transformation

One of the most common methods is log transformation. Assume the multiplicative model

$$Y = \alpha e^{\beta x}\epsilon, \tag{19}$$

where the error term ϵ enters in a multiplicative fashion. Through a log transformation, the model in (19) is converted to an additive model

$$\log Y = \log\alpha + \beta x + \log\epsilon, \quad \log\epsilon \sim N(0, \sigma^2). \tag{20}$$

Due to the logarithms, the multiplicative model is transformed into a simple linear regression model (Sundberg, 2021, Ch 3.3, page 108).

2.6.2 Box Cox transformation

The one-parameter family of Box-Cox transformations is defined as

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log y_i & \text{if } \lambda = 0, \end{cases}$$

and the two-parameter family of Box-Cox transformations as

$$y_i^{((\boldsymbol{\lambda_1},\boldsymbol{\lambda_2}))} = \begin{cases} \dfrac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0, \\ \log(y_i + \lambda_2) & \text{if } \lambda_1 = 0. \end{cases}$$

Moreover, the first family of transformations are defined for $y_i > 0$, and the second one for $y_i > -\lambda_2$. The parameter λ is estimated using the profile likelihood function and goodness-of-fit tests (Box and Cox, 1964, "An analysis of transformations").

2.6.3 Yeo-Johnson power transformation

The Yeo-Johnson power transformation is an extended version of the Box-Cox transformation that handles zero and negative values of y. λ can be any real number, where $\lambda = 1$ gives rise to the identity transformation. The transformation law reads as follows (Yeo and Johnson, 2000):

$$y_i^{(\lambda)} = \begin{cases} ((y_i + 1)^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0, \\ \log(y_i + 1) & \text{if } \lambda = 0, y \geq 0, \\ -((-y_i + 1)^{(2-\lambda)} - 1)/(2 - \lambda) & \text{if } \lambda \neq 2, y < 0, \\ -\log(-y_i + 1) & \text{if } \lambda = 2, y < 0. \end{cases} \tag{21}$$

2.7 Predictions

2.7.1 Mean squared error of prediction

For the chosen model, the prediction error should be as low as possible. One way of expressing this is to use the Mean Squared Prediction Error (MSEP), the average value of the squared difference between the response variables and their predictions. The observation i is temporarily removed from the data. The regression is then done without observation i, with the predicted value $\hat{y}_{-i}$ where the notation $-i$ is used to indicate that observation i has been omitted. This is done to predict the value y_i. The prediction error is calculated as $y_i - \hat{y}_{-i}$. The prediction measure is then given by (Sundberg, 2021, Ch 3.2, page 90):

$$MSEP = \frac{1}{n} \sum_{1}^{n} (y_i - \hat{y}_{-i})^2. \tag{22}$$

It should also be mentioned that the lower the value of MSEP is, the more adapted the dataset is to the regression model.

3 Data

The data used for this thesis is obtained from the website FootyStats. FootyStats is the premier football stats and analysis site, with data coverage of 1000+ football leagues worldwide, including the UK, Europe, and South America (FootyStats, 2022). The data obtained consists of all the teams' matches for the premier league season 2018-2019. As there are 20

playing teams in the premier league, and all teams play 38 games over a season, this generates 380 rows of the xlsx file, one for each game. The data contains a lot of variables from football matches. In addition, it also includes several different betting sites' odds before the games, the number of spectators, the name of the referee that judged, and a lot more that adds up to a total of 64 columns. However, these will not be used since the focus is on which variables determine the outcome of the match on the field and not other external effects.

3.1 Description of used variables

The variables that will be used for the project are the following:

Table 1: The variables of each game used in this thesis and their description.

Variable name	Variable description	Variable type
home_team_name	Name of the home team	Character
away_team_name	Name of the away team	Character
home_ppg	Points Per Game for Home Team - Current	Float
away_ppg	Points Per Game for Away Team - Current	Float
home_team_goal_count	Number of goals scored by the home team	Integer
away_team_goal_count	Number of goals scored by the away team	Integer
home_team_goal_count_half_time	Number of goals scored by the home team by half-time	Integer
away_team_goal_count_half_time	Number of goals scored by the away team by half-time	Integer
home_team_corner_count	Home Team corner count	Integer
away_team_corner_count	Away Team corner count	Integer
home_team_yellow_cards	Number of yellow cards for home team	Integer
away_team_yellow_cards	Number of yellow cards for away team	Integer
home_team_red_cards	Number of red cards for home team	Integer
away_team_red_cards	Number of red cards for away team	Integer
home_team_shots	Total number of shots for home team	Integer
away_team_shots	Total number of shots for away team	Integer
home_team_shots_on_target	Total number of shots on target for home team	Integer
away_team_shots_on_target	Total number of shots on target for away team	Integer
home_team_fouls	Total number of fouls for home team	Integer
away_team_fouls	Total number of fouls for away team	Integer
home_team_possession	Total number of shots for home team	Integer
away_team_possession	Total number of shots for away team	Integer

3.2 Variable transformation & Final dataset

To analyze each match and draw conclusions, a transformation of data is needed. This is done by calculating the difference between each pair of parameters of Table 1 (except for the names of the teams) for each match. Thus, differences of all variables are formed, subtracting the away team from the home team, reducing the number of columns from 22 to 10. After all these differences have been formed, the final dataset contains the following variable names, with the following meaning:

Table 2: Final set of variables of each game, used in this thesis.

Variable name	Variable transformation	Variable description
$scorediff$	$home\_team\_goal\_count - away\_team\_goal\_count$	Response variable that indicates the goal difference
sd	$home\_team\_shots - away\_team\_shots$	Discrete variable indicating shot difference
td	$home\_team\_shots\_on\_target - away\_team\_shots\_on\_target$	Discrete variable indicating the shots on target difference
fd	$home\_team\_fouls - away\_team\_fouls$	Discrete variable indicating foul difference
cd	$home\_team\_corner\_count - away\_team\_corner\_count$	Discrete variable indicating corner difference
yd	$home\_team\_yellow\_cards - away\_team\_yellow\_cards$	Discrete variable indicating yellow card difference
rd	$home\_team\_red\_cards - away\_team\_red\_cards$	Discrete variable indicating the difference in the number of red cards
hd	$home\_team\_goal\_count\_half\_time - away\_team\_goal\_count\_half\_time$	Discrete variable indicating the half time result difference
$ppgd$	$home\_ppg - away\_ppg$	Continuous variable indicating the average point per game difference
pd	$home\_team\_possession - away\_team\_possession$	Variable indicating the possession difference in percentage. It is discrete, as it can only attain values between 0 and 100

4 Analysis and Results

The models created through different stepwise regression methods will be presented and used to predict upcoming football games. In this way, a conclusion can be drawn about which model and method gives the best result in comparison with the actual result.

4.1 Simple linear regression

Simple linear regression is applied to explore whether a linear relationship between the individual explanatory variables is present. These values are presented below in the form of a table where the most important aspects are included.

Table 3: Coefficient of determination (R^2) for all explanatory variables of Table 2, against the response variable, when simple linear regression is used.

Explanatory variable	R^2	p-value
sd	0.1887	$<$ 2e-16
td	0.3876	$<$ 2e-16
fd	0.002089	0.3743
cd	0.02467	0.002135
yd	0.01061	0.0448
rd	0.01404	0.002089
hd	0.5151	$<$ 2.2e-16
ppgd	0.3724	$<$ 2.2e-16
pd	0.08916	2.892e-09

From Table 3, the only variable that is not significant is fd (foul difference) when applying simple linear regression. On the other hand, by looking at the R^2 values, it can be seen that variables such as td (shots on target difference), hd (halftime score difference) and ppgd (points per game difference) have an enormous influence on the response variable scorediff.

4.2 Multicollinearity

A strong relationship between the explanatory variables and the response variable, but not between the explanatory variables among themselves, is preferable. Therefore, in order to check correlations between all pairs of variables, correlation coefficients are provided in Table 4.

Table 4: Correlation coefficients between all pairs of variables.

	scorediff	sd	td	fd	cd	yd	rd	hd	ppgd	pd
scorediff	1.00	0.43	0.62	-0.05	0.16	-0.10	-0.12	0.72	0.61	0.30
sd	0.43	1.00	0.81	-0.15	0.65	-0.19	-0.15	0.26	0.54	0.60
td	0.62	0.81	1.00	-0.08	0.51	-0.16	-0.09	0.41	0.57	0.60
fd	-0.05	-0.15	-0.08	1.00	-0.08	0.30	0.05	-0.08	-0.16	-0.25
cd	0.16	0.65	0.51	-0.08	1.00	-0.18	-0.08	0.01	0.43	0.58
yd	-0.10	-0.19	-0.16	0.30	-0.18	1.00	0.04	-0.12	-0.11	-0.16
rd	-0.12	-0.15	-0.09	0.05	-0.08	0.04	1.00	-0.04	-0.01	-0.10
hd	0.72	0.26	0.41	-0.08	0.01	-0.12	-0.04	1.00	0.47	0.13
ppgd	0.61	0.54	0.57	-0.16	0.43	-0.11	-0.01	0.47	1.00	0.67
pd	0.30	0.60	0.50	-0.25	0.58	-0.16	-0.10	0.13	0.67	1.00

From Table 4, it can be seen that there is a very high correlation between sd and td and a reasonably high correlation between, for instance, ppgd and pd and cd and sd.

The degree of multicollinearity that arises depends entirely on which question is asked. For example, how many observations there are, the size of the estimated effects, and how much variable these estimates are. A sign of a small degree of multicollinearity is a low standard error for the estimated effect parameters. One can also measure the degree of multicollinearity in other ways by, for example, calculating VIF values. VIF provides a value for multicollinearity between the independent variables in multiple regression. A VIF value of 1 means no correlation between an explanatory variable and the other explanatory variables in the model. A VIF value in the range $1-5$ indicates an amount of correlation between explanatory variables but it is often insufficient to implement a change in the model (i.e. removal of some explanatory variables). A VIF value higher than 5 means a high correlation between the explanatory variables (Sundberg, 2021, page 95). First of all a calculation of VIF for the independent variable sd is done by fitting a multiple regression model

$$sd \sim td + fd + cd + yd + rd + hd + ppgd + pd, \tag{23}$$

and then applying (16). Then the same is done for the other independent variables. This gives the following values:

Table 5: Values of the Variance Inflation Factor (VIF) for all explanatory variables.

sd	td	fd	cd	yd	rd	hd	ppgd	pd
4.03	3.42	1.18	2.07	1.15	1.04	1.62	2.66	2.61

One can see from Table 5 that each explanatory variable exhibits a relatively low value of VIF, with no numbers above 5. In line with the previous correlation table (Table 2), the variables *sd* and *td* have the highest VIF.

4.3 Modeling

Since all explanatory variables' VIF values were lower than 5, all explanatory variables are kept. Thus, the whole model, including outliers, will be as follows:

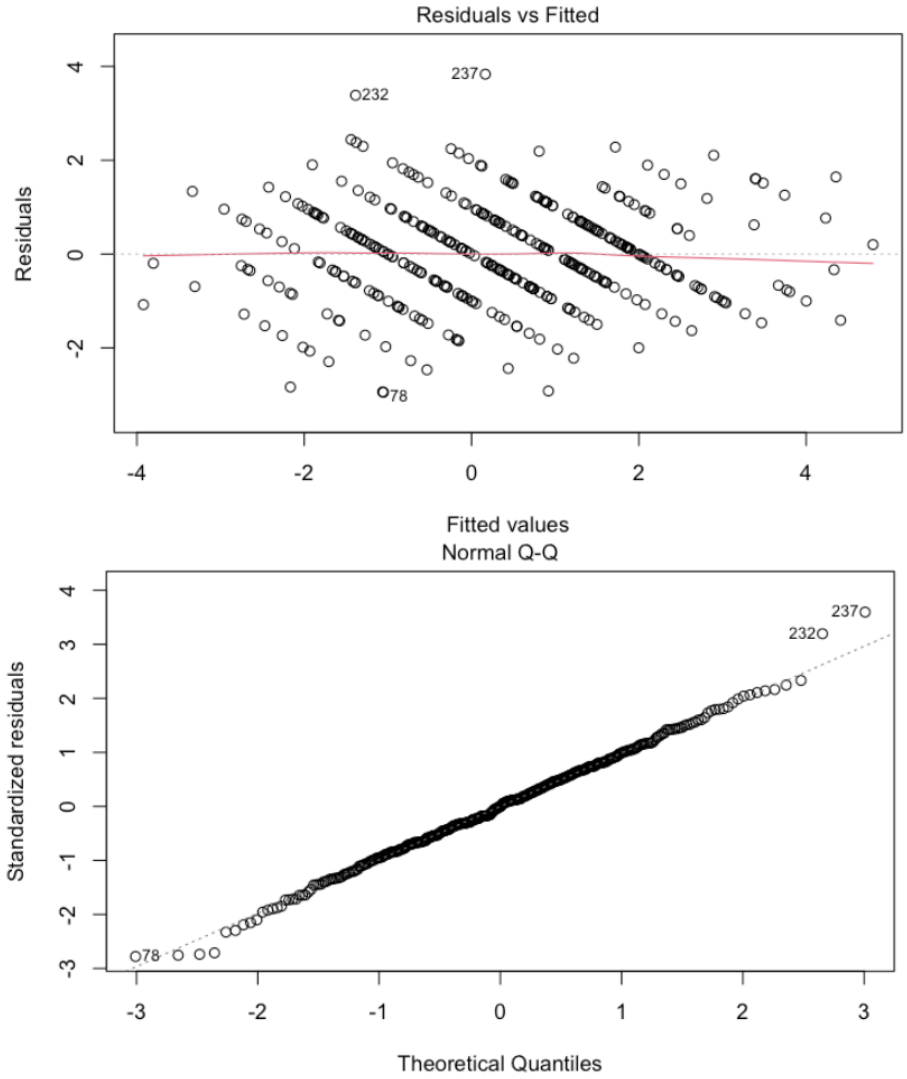

Figure 1: Residual and QQ-plots for the fit of model 1 of equation (24).

Model 1 is given by:

$$\begin{aligned} \text{scorediff}_i &= \alpha + \beta_1 sd + \beta_2 td + \beta_3 fd + \beta_4 cd + \beta_5 yd \\ &+ \beta_6 rd + \beta_7 hd + \beta_8 ppgd + \beta_9 pd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{24}$$

From Figure 1 it seems that the residuals are normally distributed with a constant variance. A log-transformation of the response variable was also tried, with no further success. Even though it can be seen below that there is constant variance and normally distributed data, the adjusted coefficient of determination R^2_{adj} is 0.5244 for the log-transformation, which indicates that there are many less significant explanatory variables not included in the model.

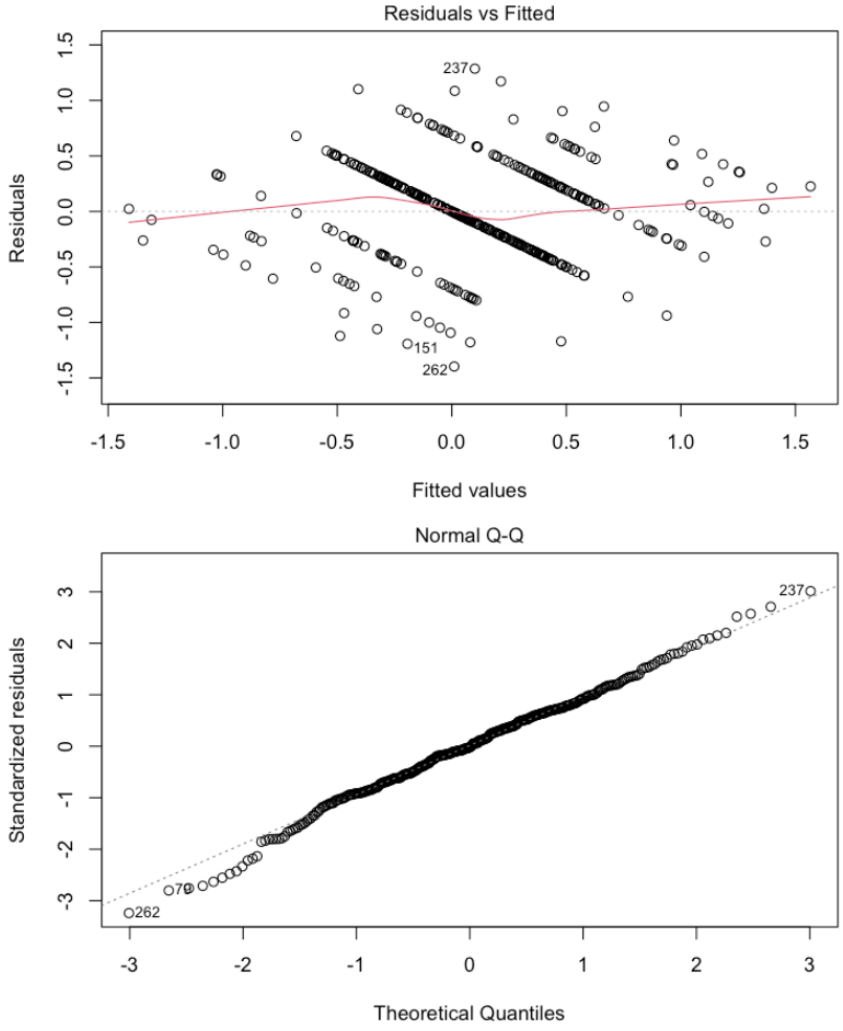

Figure 2: Residual and QQ-plots for the fit of a model where the response variablescorediff_i of equation (24) is log-transformed

On the other hand, from Figure 3 it can be seen that a Yeo-Johnson transformation of the response variable gave much better results than the log transformation did.

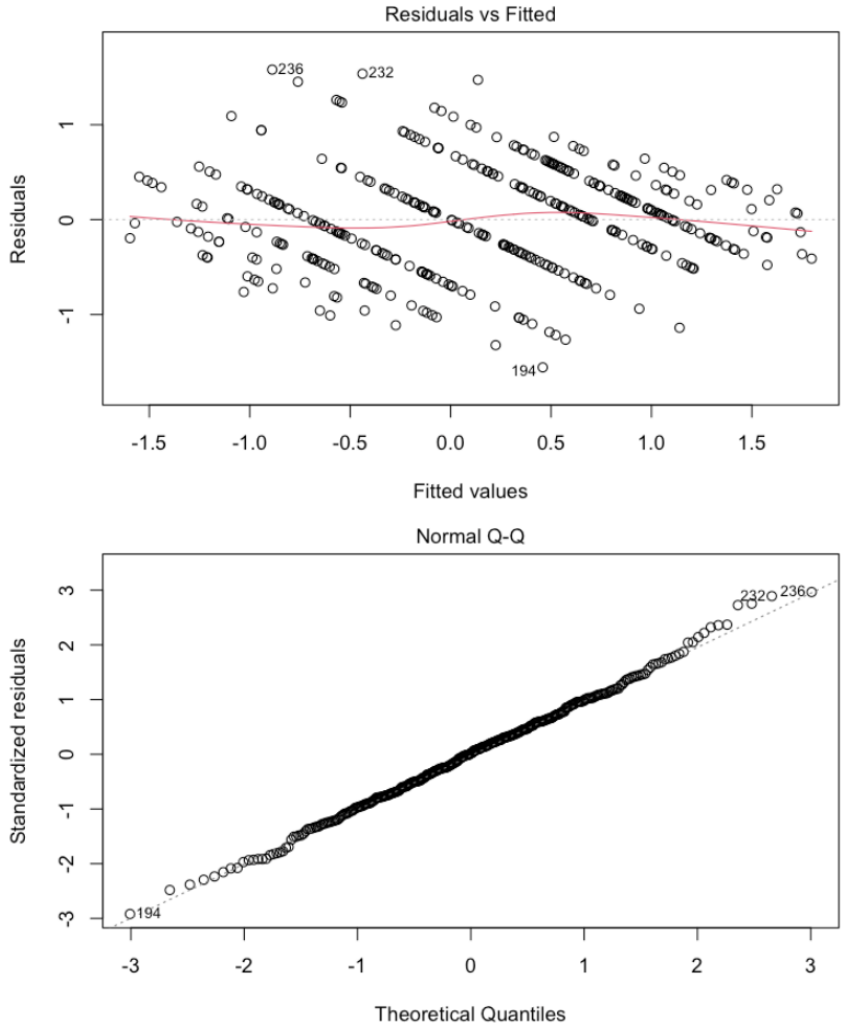

Figure 3: Residual and QQ-plots for the fit of a model where a Yeo-Johnson transformation (21) is applied to the response variable scorediff_i of equation (24).

However, R^2_{adj} is marginally higher for the original, untransformed data. This is not so unreasonable as the explanatory variables seem to explain the response variable in a linear way. For this reason, in the sequel we will only consider untransformed response variables.

Variable selection aims to find the best model for predicting football results, which is now known to be based on the initial model (24) without any transformations of variables. The goal is to find the explanatory variables that correspond to the highest value of the adjusted R^2. Therefore, finding the best models for each stepwise regression method can begin.

The AIC has a lower threshold for including additional explanatory variables compared to the backward elimination and forward selection with a p-value threshold of 0.1. Therefore different results can be obtained to see which types of thresholds work the best.

Assume the initial and maximal model (24) is used. Then, after a fit of this multiple regression model to data, the following table is obtained.

Table 6: p-values for the estimated effect parameters of all explanatory variables of the linear multiple regression model (24).

Explanatory variable	p-value
sd	0.16950
td	$9.67e-14$
fd	0.17050
cd	0.01491
yd	0.90791
rd	0.00276
hd	$< 2e-16$
ppgd	4.05e-10
pd	0.21407

4.3.1 Forward selection model

When forward selection is performed, the procedure begins by selecting the most significant variable (hd). This process is ongoing until the most significant explanatory variable no longer has a p-value under 0.1 (or that the adjusted R^2 decreases). This is illustrated below:

Model 1:

$$\text{scorediff}_i = \alpha + \beta_7 hd + \epsilon_i, i = 1, ...380 \tag{25}$$

Model 2:

$$\text{scorediff}_i = \alpha + \beta_2 td + \beta_7 hd + \epsilon_i, i = 1, ...380 \tag{26}$$

Model 3:

$$\text{scorediff}_i = \alpha + \beta_2 td + \beta_7 hd + \beta_8 ppgd + \epsilon_i, i = 1, ...380 \tag{27}$$

Model 4:

$$\text{scorediff}_i = \alpha + \beta_2 td + \beta_4 cd + \beta_7 hd + \beta_8 ppgd + \epsilon_i, i = 1, ...380 \tag{28}$$

Model 5:

$$\begin{aligned} scorediff_i &= \alpha + \beta_2 td + \beta_4 cd + \beta_6 rd \\ &+ \beta_7 hd + \beta_8 ppgd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{29}$$

Model 6:

$$\begin{aligned} scorediff_i &= \alpha + \beta_2 td + \beta_4 cd + \beta_6 rd \\ &+ \beta_7 hd + \beta_8 ppgd + \beta_9 pd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{30}$$

Where the following plot can show the progress:

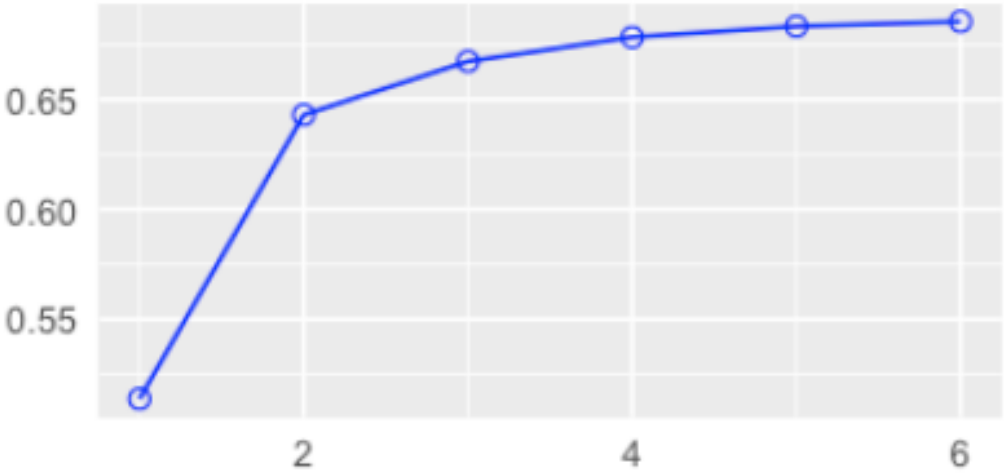

Figure 4: Forward selection: Adjusted R^2 as a function of the number of included explanatory variables, showing an improvement for each step.

4.3.2 Backward elimination model

Backward elimination is based on the initial model (24) and then the explanatory variables are reduced one by one until all remaining explanatory variables are significant. To get a stopping criterion and see for how long the model fit can be improved, a limit is set so that all included variables in the final model must have a p-value below 0.1, when testing whether or not they should be removed. The steps are illustrated below:

Model 1 is given by equation (24).

Model 2:

$$\begin{aligned} \text{scorediff}_i &= \alpha + \beta_1 sd + \beta_2 td + \beta_3 fd + \beta_4 cd + \beta_6 rd \\ &+ \beta_7 hd + \beta_8 ppgd + \beta_9 pd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{31}$$

Model 3:

$$\begin{aligned} \text{scorediff}_i &= \alpha + \beta_1 sd + \beta_2 td + \beta_3 fd + \beta_4 cd + \beta_6 rd \\ &+ \beta_7 hd + \beta_8 ppgd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{32}$$

Model 4:

$$\begin{aligned} \text{scorediff}_i &= \alpha + \beta_2 td + \beta_3 fd + \beta_4 cd + \beta_6 rd \\ &+ \beta_7 hd + \beta_8 ppgd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{33}$$

The following plot shows the progress of the backward elimination method:

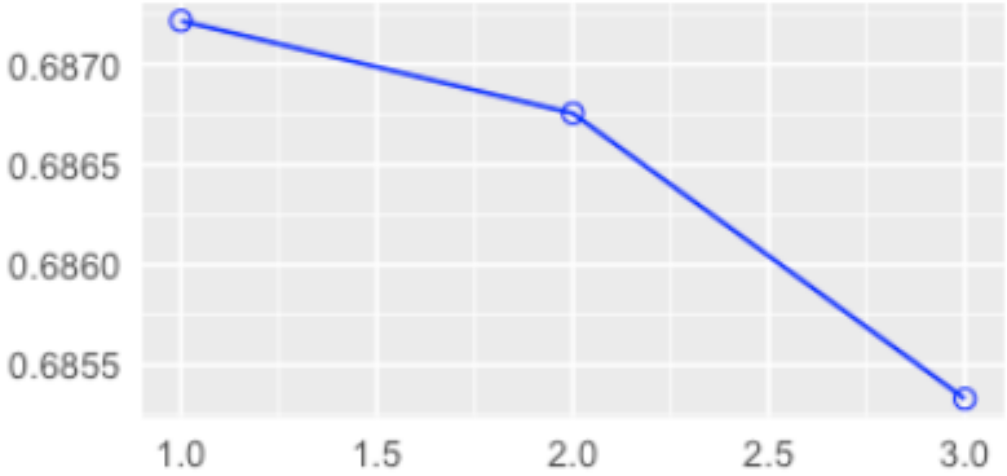

Figure 5: Backward elimination: Adjusted R^2 as a function of the number of removed explanatory variables, showing a slight decrease for each step.

Even though the adjusted R^2 is decreasing in each step, the decrease is so marginal that no action is taken, i.e. the model with six remaining explanatory variables is regarded as the best one.

4.3.3 AIC forward selection model

Unlike traditional forward selection, AIC forward selection is based on adding the explanatory variable that lowers the AIC the most. This process continues until no explanatory variable is left that can lower the AIC-value. The following sequence of models is obtained:

Model 1:

$$scorediff_i = \alpha + \beta_7 hd + \epsilon_i, i = 1, ...380 \tag{34}$$

Model 2:

$$scorediff_i = \alpha + \beta_2 td + \beta_7 hd + \epsilon_i, i = 1, ...380 \tag{35}$$

Model 3:

$$scorediff_i = \alpha + \beta_2 td + \beta_7 hd + \beta_8 ppgd + \epsilon_i, i = 1, ...380 \tag{36}$$

Model 4:

$$scorediff_i = \alpha + \beta_2 td + \beta_4 cd + \beta_7 hd + \beta_8 ppgd + \epsilon_i, i = 1, ...380 \tag{37}$$

Model 5:

$$\begin{aligned} scorediff_i &= \alpha + \beta_2 td + \beta_4 cd + \beta_6 rd \\ &+ \beta_7 hd + \beta_8 ppgd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{38}$$

Model 6:

$$\begin{aligned} scorediff_i &= \alpha + \beta_2 td + \beta_4 cd + \beta_6 rd \\ &+ \beta_7 hd + \beta_8 ppgd + \beta_9 pd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{39}$$

Model 7:

$$\begin{aligned} scorediff_i &= \alpha + \beta_2 td + \beta_3 fd + \beta_4 cd \\ &+ \beta_6 rd + \beta_7 hd + \beta_8 ppgd + \beta_9 pd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{40}$$

These models' successively reduced AIC-values are shown below:

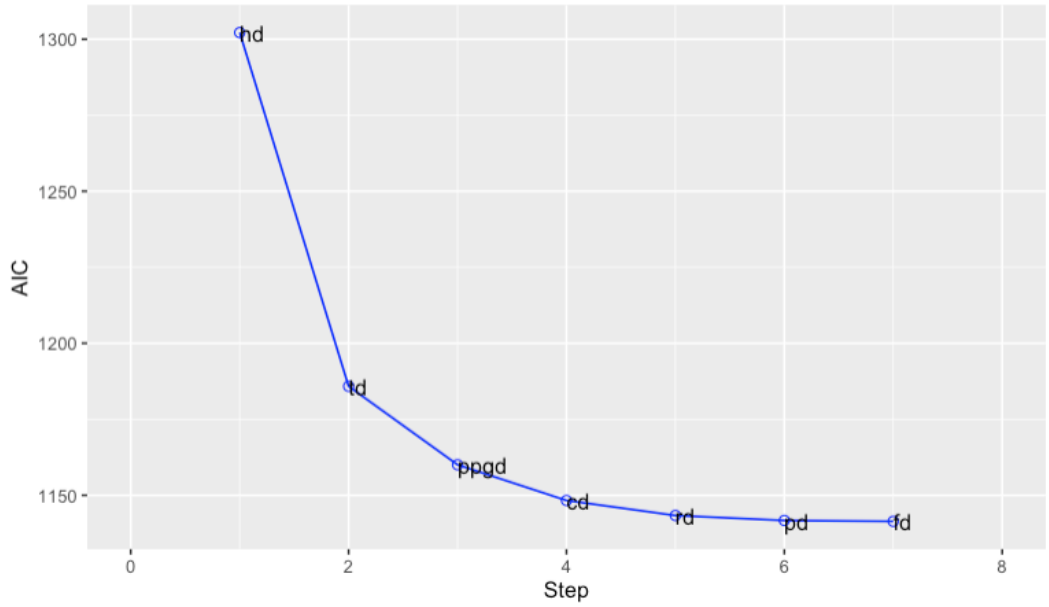

Figure 6: AIC-values based on forward selection, as a function of the number of included explanatory variables.

4.3.4 AIC backward elimination model

Like the previous model, AIC backward elimination aims to obtain the lowest possible AIC value. On the other hand, it starts with the full model (24), just as in the case of backward elimination, and then removes explanatory variables that lower the AIC the most, one by one, until no further improvement is attained. Referring to (24) as Model 1, the following results are obtained:

Model 1 is given by equation (24).

Model 2:

$$\begin{aligned} scorediff_i &= \alpha + \beta_1 sd + \beta_2 td + \beta_3 fd + \beta_4 cd \\ &+ \beta_6 rd + \beta_7 hd + \beta_8 ppgd + \beta_9 pd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{41}$$

Model 3:

$$\begin{aligned} scorediff_i &= \alpha + \beta_1 sd + \beta_2 td + \beta_3 fd + \beta_4 cd \\ &+ \beta_6 rd + \beta_7 hd + \beta_8 ppgd + \epsilon_i, \quad i = 1, \ldots, 380. \end{aligned} \tag{42}$$

The following plot illustrates the progress of AIC backward elimination:

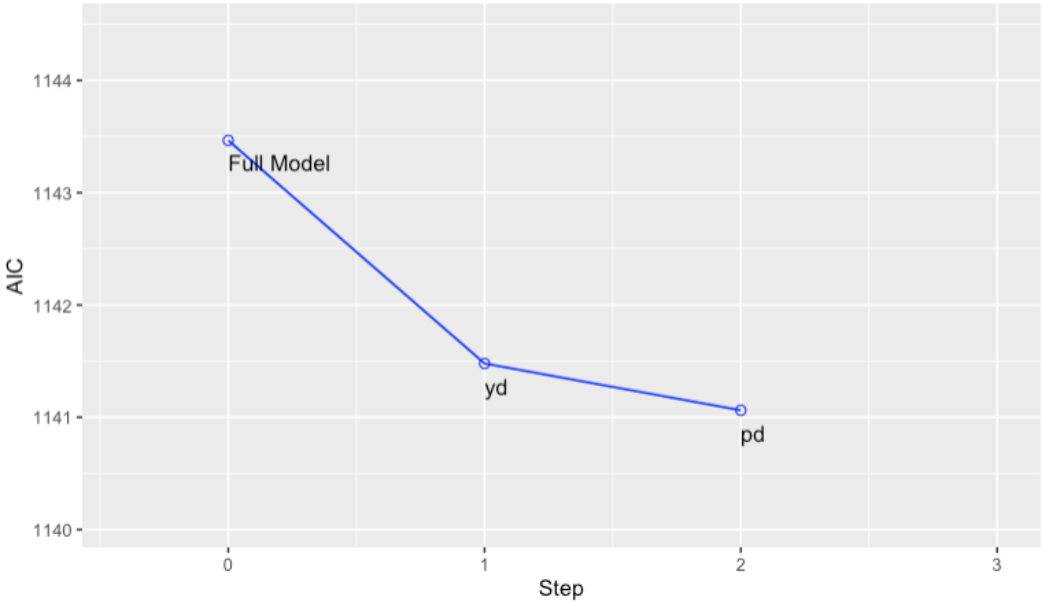

Figure 7: AIC-values based on backward elimination, as a function of the number of removed explanatory variables.

4.4 Prediction

Training data for the premier league season 2018-2019 has been used to fit the best multiple linear regression model (2) for predicting upcoming games. Therefore, data for the next season, 2019-2020, has been chosen for validation. The predicted score difference $\hat{E}(Y_i) = \hat{\mu}_i = \hat{\alpha} + \hat{\beta}x_i$ as a function of the explanatory variables x_i of each game i of the validation set, was computed for each model. This will generate a predicted score difference $\hat{\mu}_i$ for each game i. If this score difference has a value over 0.5, then 3 points are given to the home team. If the score difference is below -0.5, 3 points are provided for the away team. Anything in between will generate 1 point for both teams. When this is done for all teams over the whole season, each team will have a cumulative predicted sum, which is compared to the actual sum for every team for that particular season 2019-2020. Thus, a conclusion can be made about which method gives the best prediction.

When subtracting each team's actual number of points from the predicted number of points during the whole season, the following results are obtained:

Table 7: Sum of the differences between actual and predicted number of points during the whole season 2019-2020. The table displays the sum of the differences for all 20 teams.

Method	Total sum diff
Backward elimination	8
Forward selection	4
AIC Backward elimination	8
AIC Forward selection	1

The table above does not show how accurate each model's prediction is since one team might have too few points predicted in some games, and too many points predicted in others, and still these prediction errors cancel out during a whole season. However, these total score differences of Table 7 give a glimpse of which model might be the best. Furthermore, since the total amount of points can vary from season to season depending on how many games result in a draw and two points are handed out, and how may games result in a win and 3 points are given, this shows that a total score difference of 1 for the AIC forward selection is very good and indeed, Table 7 indicates that all four methods make very good predictions.

An additional prediction measure was made to see how well all the different methods predicted each game. For this, the actual result for each game was compared with the predicted outcome. This could be seen in how many of all 380 games were predicted correctly. The result can be seen in Table 8.

Table 8: The number and fraction of all 380 games during 2019-2020 that were predicted correctly.

Method	Correct predictions	Percentage
Backward elimination	267	70.26%
Forward selection	268	70.53%
AIC Backward elimination	268	70.53%
AIC Forward selection	270	71.05%

Thus, the AIC forward selection model has a slight advantage over the models obtained from the other methods, but the difference is so minimal that it is almost impossible to draw any firm conclusions.

4.4.1 Forward selection prediction

The estimated effect parameters $\hat{\beta}_j$ for all the explanatory variables x_j are as follows:

Table 9: The estimated effect parameters for all explanatory variables of the forward selection model

Explanatory variable	Estimate
td	0.179
cd	−0.047
rd	−0.461
hd	0.714
ppgd	0.716
pd	−0.006

Based on Table 9 and the estimated intercept, the difference between the actual total score and predicted total score of each team will look like the following:

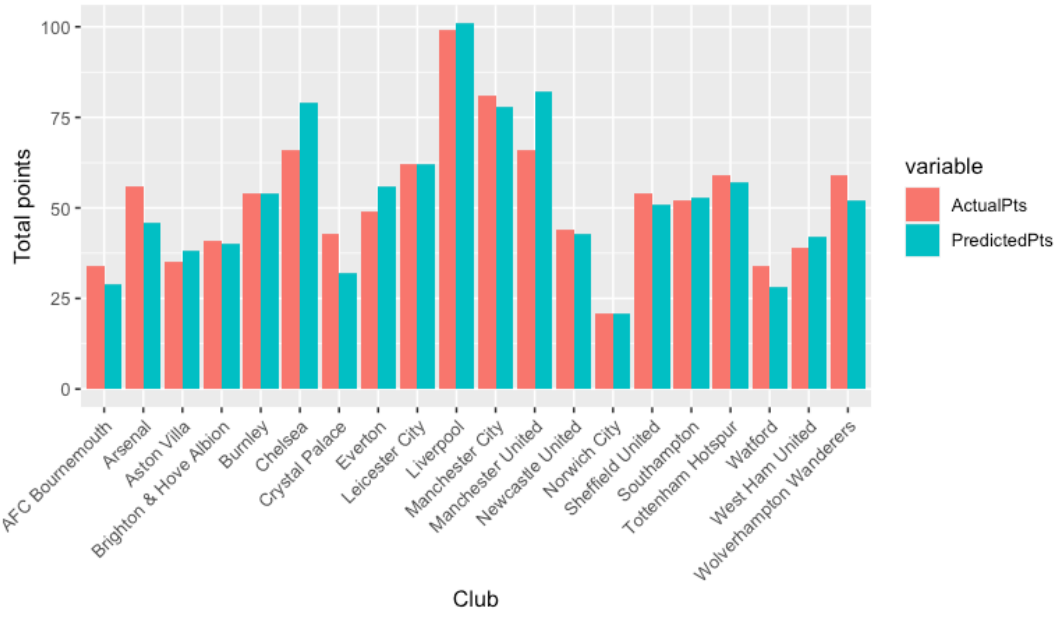

Figure 8: Actual and predicted total scores for forward selection.

4.4.2 Backward elimination prediction

When applying the backward elimination the estimated effect parameters are the following:

Table 10: The estimated effect parameters for all explanatory variables of the backward elimination model

Explanatory variable	Estimate
td	0.174
cd	−0.056
fd	0.023
rd	−0.442
hd	0.741
ppgd	0.620

Based on Table 10 and the estimated intercept, the difference between the actual and predicted total scores of all teams will look like the following:

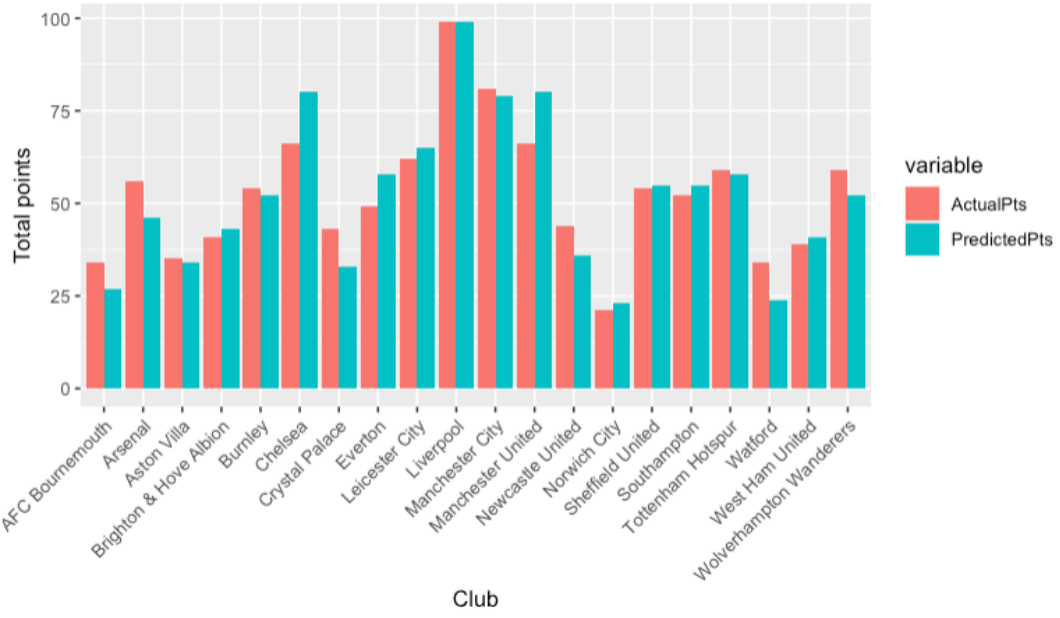

Figure 9: Actual and predicted total scores of all teams for backward elimination.

4.4.3 AIC forward selection prediction

When applying the AIC forward selection method we get the following result:

Table 11: The estimated effect parameters for all explanatory variables of the chosen AIC forward selection model

Explanatory variable	Estimate
td	0.177
cd	−0.048
fd	0.019
rd	−0.468
hd	0.719
ppgd	0.713
pd	−0.005

Using the estimated intercept and the estimated effect parameters of table 11, a comparison between the actual and predicted total scores of each team can be seen below

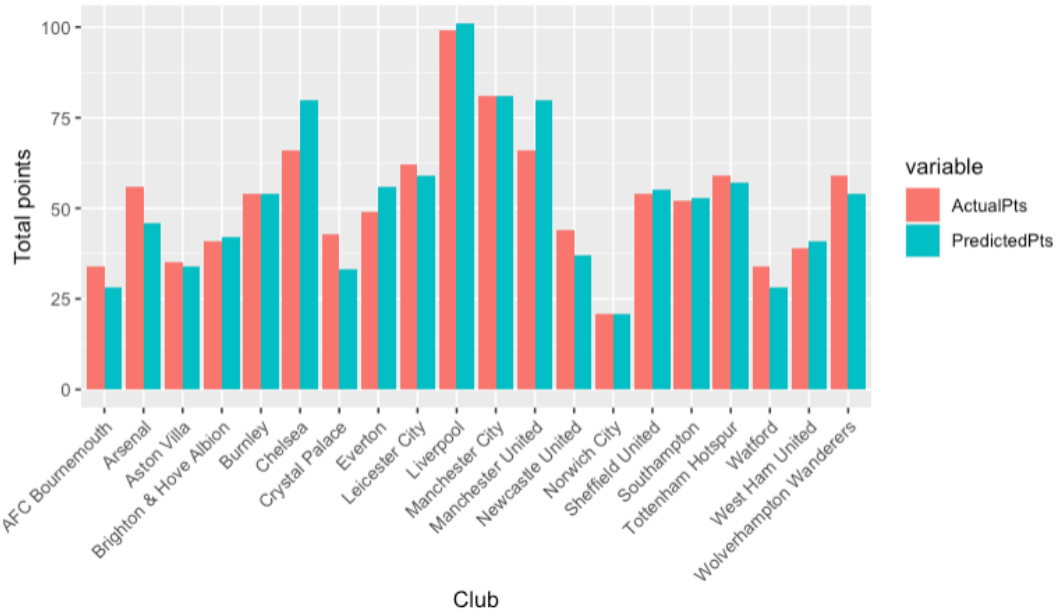

Figure 10: Difference between the actual and predicted total score of each team, based on AIC forward selection.

4.4.4 AIC backward elimination prediction

The estimated effect parameters for the explanatory variables when using AIC backward elimination are the following:

Table 12: The estimated effect parameters for all explanatory variables of the AIC backward elimination model

Explanatory variable	Estimate
sd	0.028
td	0.203
cd	−0.045
fd	0.021
rd	−0.478
hd	0.734
ppgd	0.634

The estimated intercept and the estimated effect parameters of table 12 have then been used to generate the following figure:

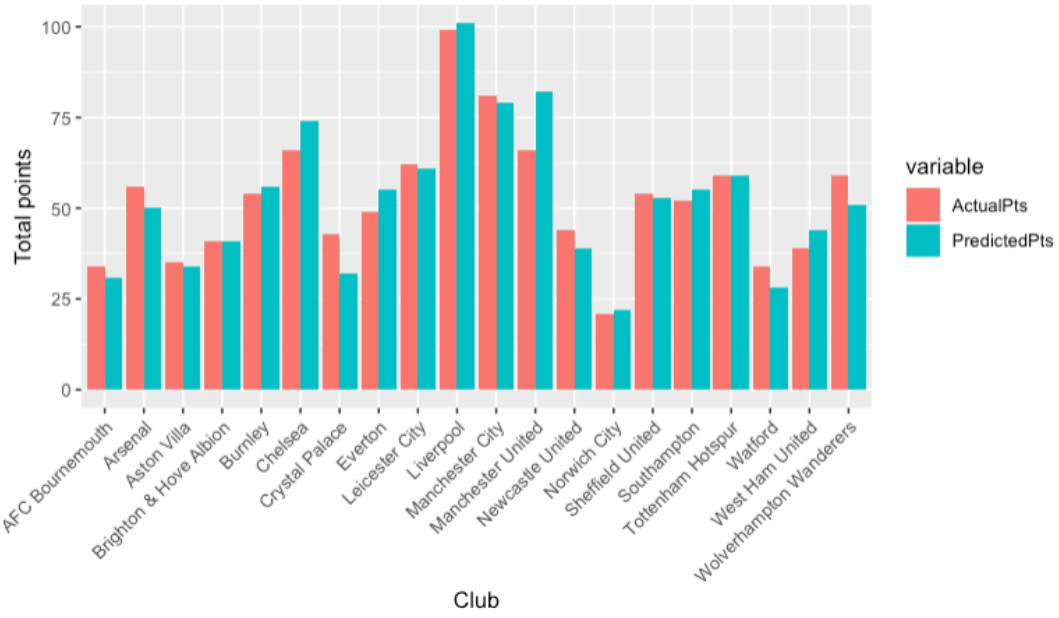

Figure 11: actual and predicted total scores of all teams when using AIC backward elimination.

4.5 Msep & Rmsep

For a good model, one prefers as low a prediction error as possible. To measure this error, the mean squared prediction error (22) of the different models can be computed, which is the average value of the square difference between the response variables and their predictions. The RMSEP is described as the square root of the MSEP. The values for these will be shown below for the full model (24), the four different stepwise regression models and the simple linear regression models with only one explanatory variable.

Table 13: MSEP and RMSEP values for a number of simple and multiple linear regression models.

Model	Msep	Rmsep
Scorediff $\sim sd$	2.998918	1.731738
Scorediff $\sim td$	2.263850	1.50461
Scorediff $\sim cd$	3.602648	1.898064
Scorediff $\sim fd$	3.689317	1.92076
Scorediff $\sim yd$	3.655647	1.911975
Scorediff $\sim rd$	3.644002	1.908927
Scorediff $\sim hd$	1.792453	1.338825
Scorediff $\sim ppgd$	2.318299	1.522596
Scorediff $\sim pd$	3.367239	1.835004
Backward elimination	1.173330	1.083204
Forward selection	1.172982	1.083043
AIC forward selection	1.172489	1.082816
AIC backward elimination	1.170806	1.082038
The full model (24)	1.180037	1.086295

5 Discussion

When evaluating all the different plots and comparing the actual results to the predicted, it could be seen that the AIC backward elimination and AIC forward selection methods gave the models with best predictive performance. This can be expected since the lowest MSEP and RMSEP values were obtained with the AIC backward elimination and AIC forward selection methods.

When evaluating the results one can see that all methods have similar performance. One of the reasons why the AIC generated better results could be that the AIC selects a larger model than hypothesis tests since the penalty term of AIC corresponds to a smaller threshold for including new explanatory variables than the hypothesis tests' cutoff at a p-value of 0.1. Therefore more variables can be included in the AIC-based models, which could be the reason for the lower MSEP values. However, it is hard to conclude whether backward or forward selection works best. One of the reasons why the backward elimination and forward selection methods have similar performance is the fact that when the explanatory variables are independent, the two methods will provide the same result.

Something that can be said about the four different model selection methods is that they all give a better prediction than the original full model (24), containing all explanatory variables. However, it can be concluded that the full model (24) still seems to provide decent predictions. As discussed earlier, hd, td, and ppgd seem to be the best explanatory variables in predicting and evaluating the most critical aspects of a game. This makes very much sense since halftime results tend to have a massive impact on the final result. The same is for the shots on target difference. The more shots a team has on target, the greater are the chances for that team to win. Furthermore, the average number of points per game difference indicates how big a favorite a team is before the game. And that also turned out to be an important factor for predicting how the game will end, in agreement with the fact that favorites very often tend to win.

5.1 Possible future enhancements

It is generally and inherently hard to predict the outcomes of football games since there is a tiny margin between a win and a draw, and that may be one of the most prominent reasons why some teams had worse predictions than others. In football, you can win a game by only taking one shot, and this may affect the accuracy of the predicted outcomes. The same applies for more passive teams. Some teams are satisfied with a one-goal win, and others strive to win by a lot. This can as well lead to misleading predictions. Thus,

it is hard to predict the actual results with good accuracy since so little can make the difference between a win and a loss. This can be seen in Figures 8-11. The predictions for the worse teams that probably had the lowest amount of shots, ball possession, etc., had much lower predicted total scores than their actual total scores. But the two AIC models resulted in better predictions than the other two models, and this may reflect that the larger AIC-models (obtained with lower thresholds to include more explanatory variables) might capture these tiny margins.

Another thing that can be said is that it was challenging to find good data. Many websites have some good statistics, but either one has to pay for it, or it is not possible to download. But most websites barely contain anything of value at all. Hence it would have been interesting to have more than 10 observations to see if this could positively affect the prediction. This thesis has only treated the aspects that the players can affect themselves but no external features. Hence one might want to look at how well the teams perform on rainy days versus sunny days and are some teams better when the games are played on normal grass or artificial grass.

We also tried to have data sets with a larger number of games and more explanatory variables, in order to see if this increases prediction accuracy. If scorediff has a value over 0.5, then 3 points are given to the home team. If it is below -0.5, 3 points are given to the away team and everything in between results in 1 point each. However, one exciting aspect would be to change this value of 0.5 to see if better predictions were obtained for the outcomes of games. But it turned out that this was barely not the case. One thought that changing the value in favor of the away team might provide an even better result since these teams have a disadvantage of being the away team and tend to be less dominant in the aspects that have been taken into account. But the best predictions were obtained when we assigned a draw for score differences between 0.5 and -0.5.

6 Appendix

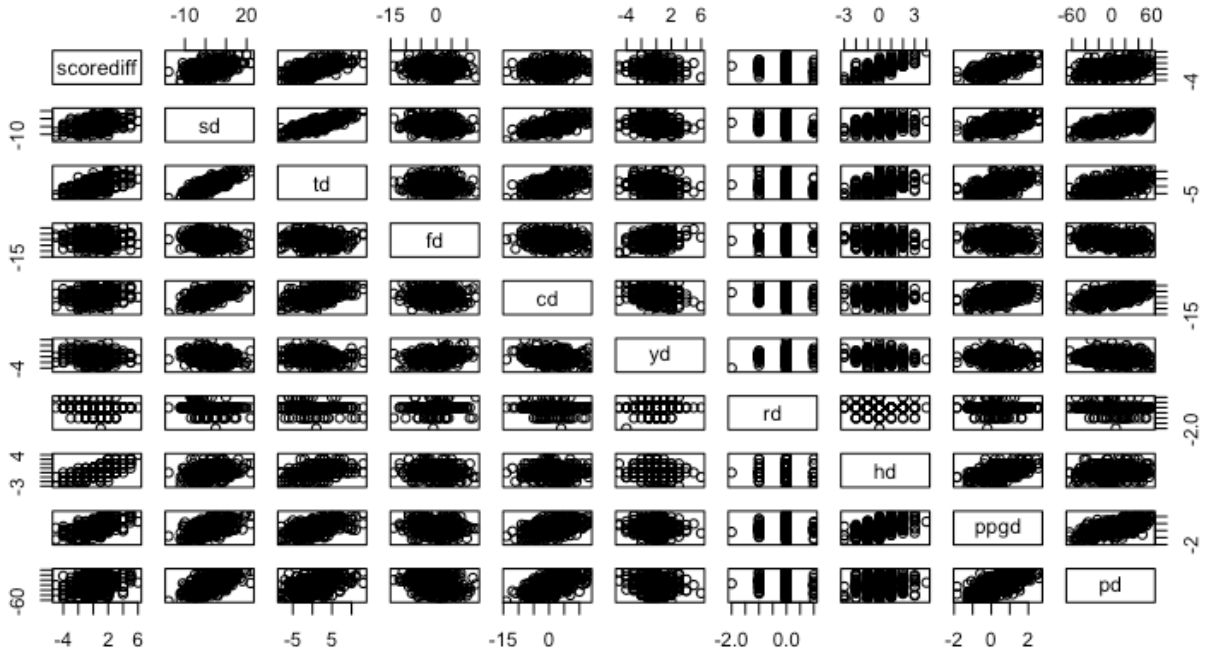

Figure 12: Pairs plot for the full model (24).

The pairs plot above shows the pairwise dependence, in terms of scatter plots, between the response and explanatory variables of the full model (24).

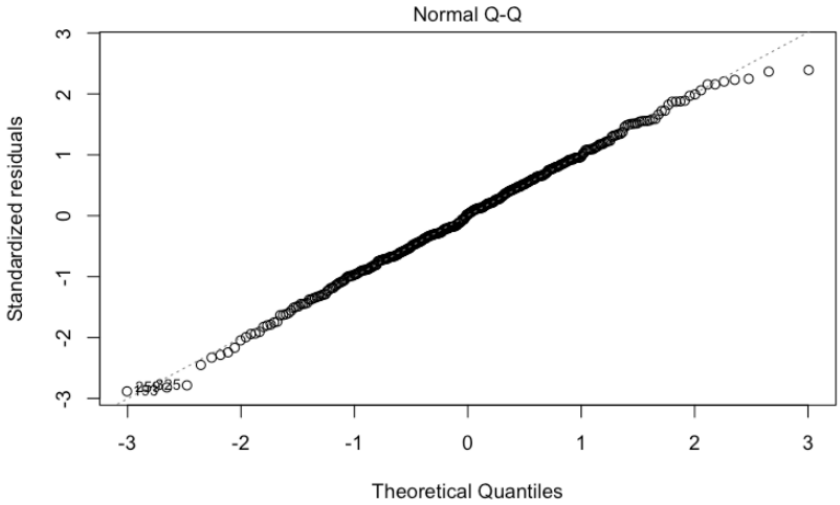

Figure 13: QQ-plots for the residuals of the full model (24), excluding outliers.

It can be seen on Figure 14 that when removing outliers, the fit of a normal distribution to the residuals gets worse.

Table 14: Sum of the differences between actual and predicted total scores, for all teams during the 2018-2019 season.

Method	Total sum diff
Backward elimination	14
Forward selection	15
AIC Backward elimination	18
AIC Forward selection	14

Showing the difference for each method when subtracting the predicted points with the actual points for each team over a whole season:

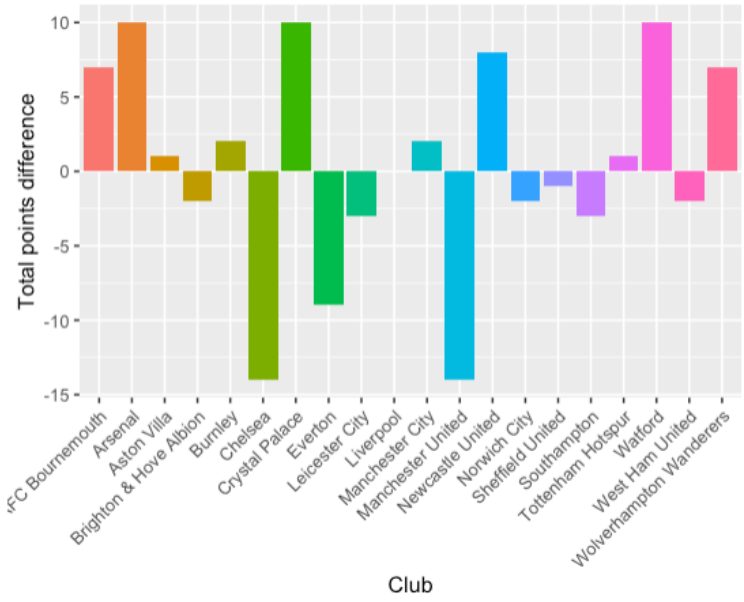

Figure 14: Difference between the actual and predicted total scores for each team over the 2019-2020 season using backward elimination

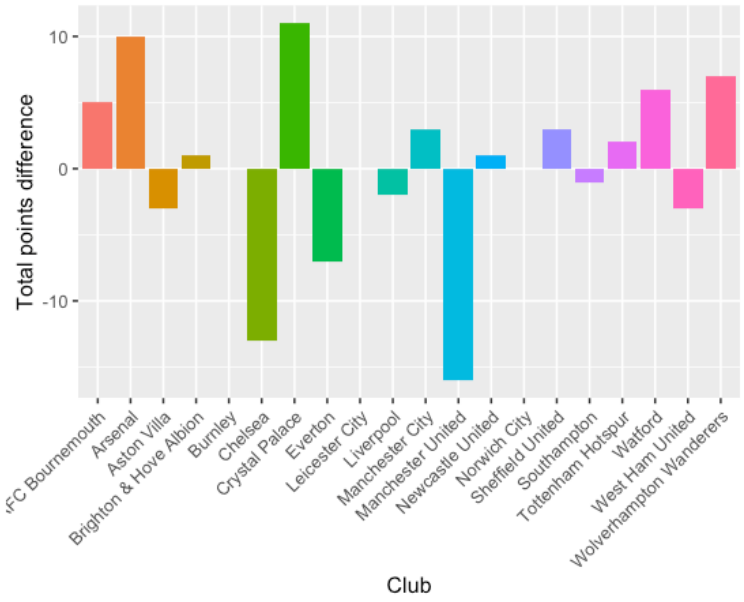

Figure 15: Difference between the actual and predicted total scores for each team over the 2019-2020 season using forward selection

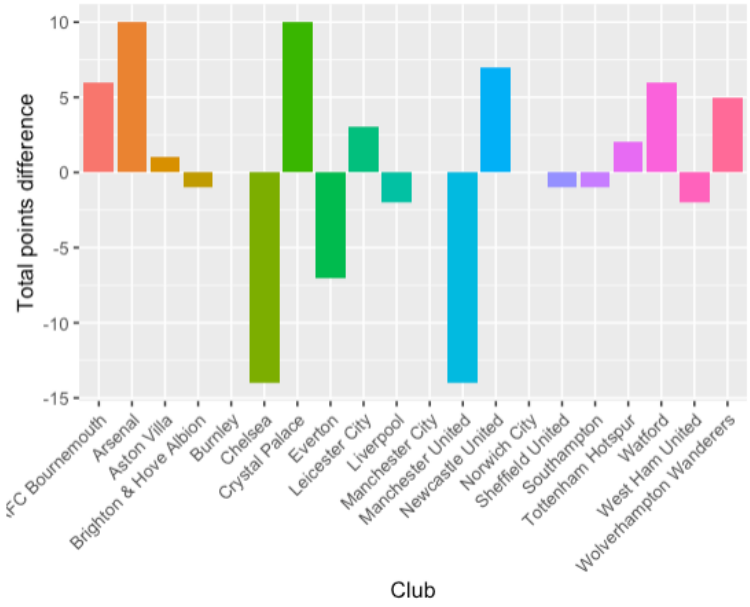

Figure 16: Difference between the actual and predicted total scores for each team over the 2019-2020 season using AIC forward selection

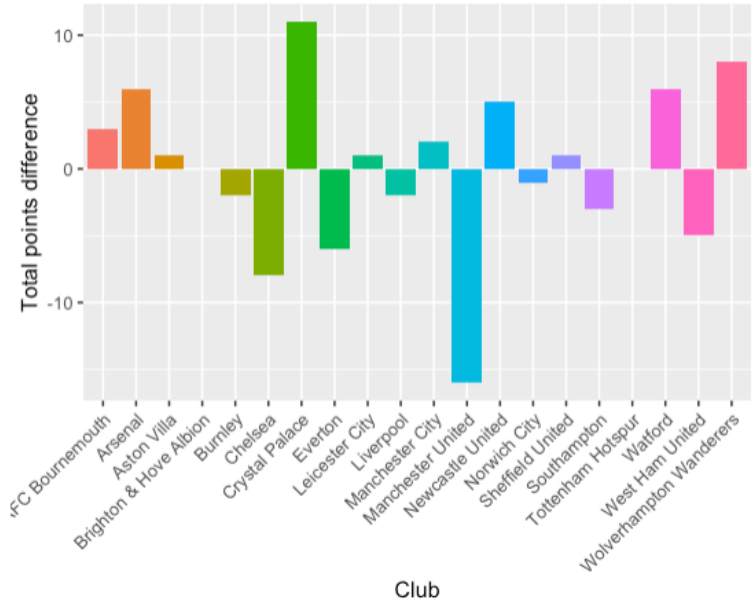

Figure 17: Difference between the actual and predicted total scores for each team over the 2019-2020 season using AIC backward elimination

The difference between the actual and predicted total scores of all teams

during the 2018-2019 season is now illustrated:

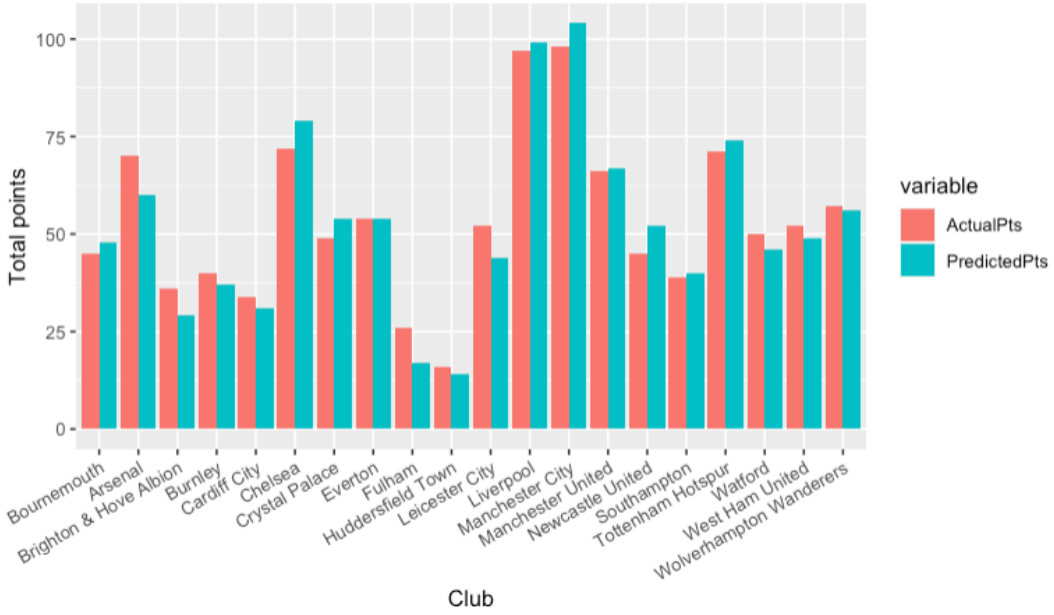

Figure 18: Difference between the actual and predicted total scores of all teams during the 2018-2019 season, based on forward selection.

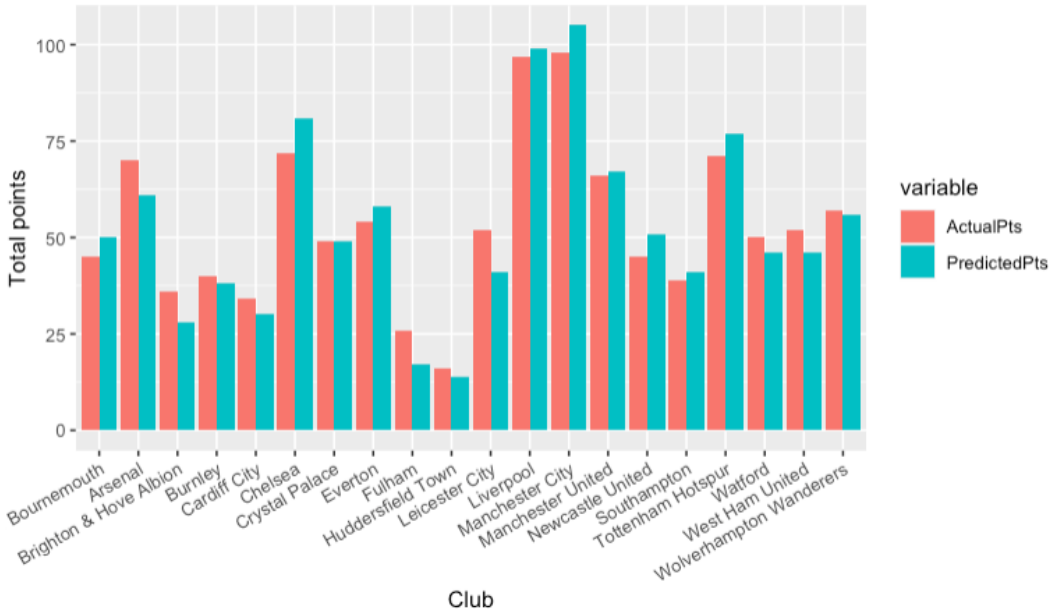

Figure 19: Difference between the actual and predicted total scores of all teams during the 2018-2019 season, based on backward elimination.

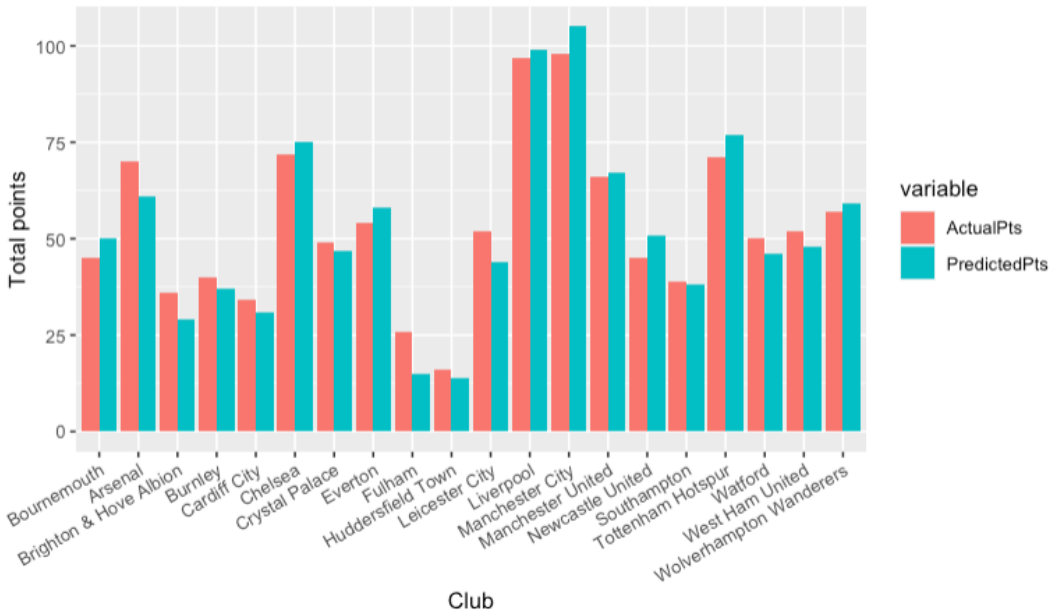

Figure 20: Difference between the actual and predicted total scores of all teams during the 2018-2019 season, based on AIC forward selection.

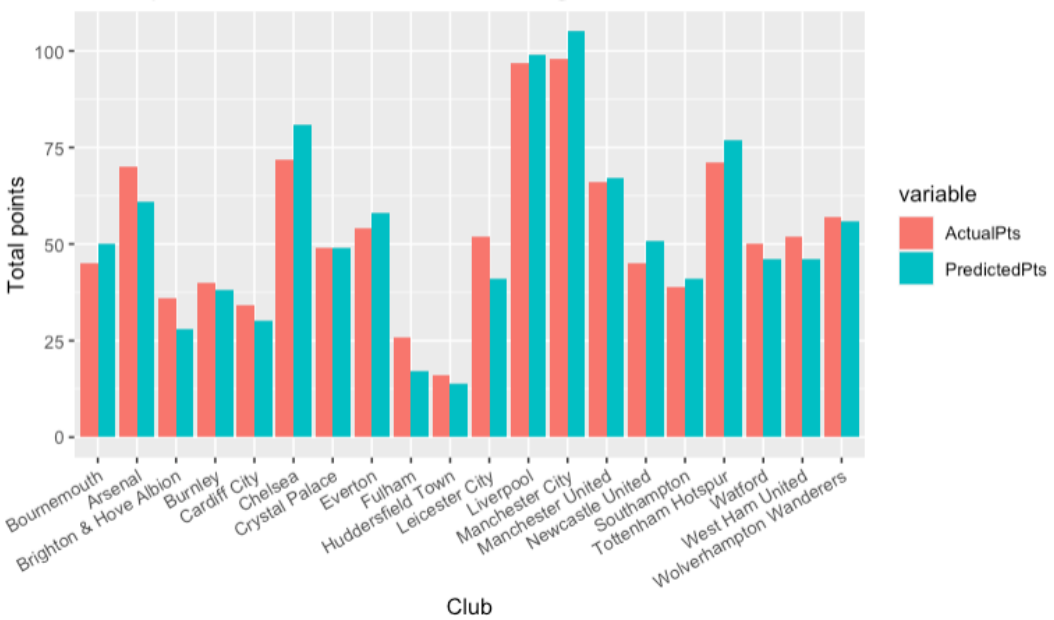

Figure 21: Difference between the actual and predicted total scores of all teams during the 2018-2019 season, based on AIC backward elimination.

7 Sources

- Sundberg, R. (2021). Lineära statistiska modeller. Department of Mathematics. Stockholm University.

- Britton, T. Alm, S.E. (2008). Stokastik. Sannolikhetsteori och statistikteori med tillämpningar. Liber AB, Stockholm.

- Held, L. Sabanés Bové, D. (2020). Likelihood and Bayesian Inference. Stockholm's University.

- Data. (2022). `https://footystats.org/download-stats-csv`.

- Yeo. Johnson. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. Biometrika 87(4), 954-959.

- Box. Cox. (1964). An analysis of transformations. Journal of the Royal Statistical Society Series B, 26(2), 211-252.

- Tambour, T. (2022). Department of Mathematics, Stockholm University. Matematik, vetenskap och samhälle. Lecture notes day 11.

- Malki, N (2022). Regressionsanalys av nyproducerade bostadsrätts priser i Solna-Sundbybergs kommun. Department of Mathematics, Stockholm University (Bachelor thesis in mathematical statistics). Section 2.1.4.