A linear modelling exploration of goal scoring in the AFL

A Preprint

Trent Henderson
OLET5608

then6675@uni.sydney.edu.au

May 3, 2021

Abstract

Enter the text of your abstract here.

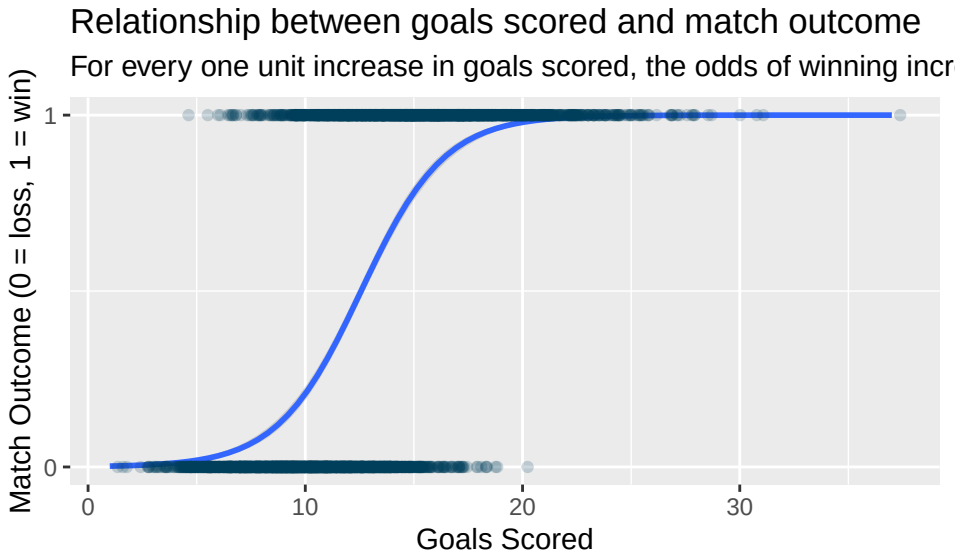
1 Introduction

AFL is a highly popular Australian sports league that began in 1896 and continues strongly today, with Grand Final match attendance (outside of the anomalous COID-19-impacted 2020 season) approximating a sold out 100,000 each year at the traditional host venue - the Melbourne Cricket Ground. An AFL match is won based on points, which can be accumulated by kicking either a goal (worth six points) or a behind (worth one point). Given the relative importance of goals-to-behinds and the rapid rise of advanced analytics in sport, many, if not all, AFL teams are interested in how they can score more goals.

Moreover, this pursuit of knowledge extends far past teams and players. Websites such as FiveThirtyEight and Advanced Sports Analytics have created a reliable source of insight and interactive analysis that only continue to rise in popularity and sophistication. However, this form of innovative and detailed sports analytics has yet to fully breach Australian sports. While the AFL has dedicated talk show analysis television programs such as AFL 360, The Front Bar, and Talking Footy, these programs focus mostly on qualitative breakdowns of high-level match statistics and not on statistical rigour. This report aims to bridge some of this gap by providing a preliminary statistical investigation of goal scoring in the AFL.

1.1 The importance of kicking goals

As described above, goals are worth six points in AFL, meaning most offensive activity is conducted so as to increase the likelihood and number of goals that are scored. Their contribution to the probability of a team winning a match is substantial, as highlighted by **FIGURE 1** below, where a match outcome of 1 = win and 0 = loss. The plot visualises the outputs of a logistic regression where the number of goals scored by a team for a given match was used as a predictor of binary match outcome (win versus loss). The number of goals scored was statistically significant, such that for every one unit increase in goals scored by a team, the odds of winning the match increase by 1.07 (p <.001, 95% confidence interval = 1.07-1.08). The case for a team focusing their efforts on scoring more goals is evident.



However, as is the case with most high-level sport, winning games is not as simple as *just* kicking more goals. There are sometimes long sequences of free-form and contested play that precede a goal being kicked that can be hard to directly influence through coaching interventions based off the notion of *as a team we need to kick more goals*. Instead, understanding the correlates and predictors of goal scoring using a data-driven approach may reveal subtle nuances in gameplay which could be used to tailor training and coaching approaches. At an even more granular level, differences in how these different gameplay attributes manifest for specific teams may lead to different recommendations.

Specifically, this report aims to explore the following research question: *Which gameplay attributes are predictors of the number of goals scored by teams in AFL matches?*

2 Data set

Historical AFL data has been made readily-accessible in an open-source setting through the R package `fitzRoy` (see Day, Nguyen, and Lane 2020). The package provides a simple API that accesses and integrates a range of data sources that collate AFL data. Examples of these sources include:

- AFL
- AFL Tables
- Squiggle
- FootyWire

The data itself is diverse, covering domains as broad as player and match statistics, Brownlow medal votes, betting odds, attendance numbers, and match times. This report focuses on player and match statistics by aggregating quantities of interest to team-per-match-level sums using data for the 2010-2019 seasons, inclusive. This time period is partially arbitrary, but was made on the basis of recency and potential homogeneity. The 2020 season is a strong counter example of this, where the season length was truncated and played almost entirely in Queensland due to the impacts of COVID-19 This means the standard set up of games - having a home and away team - was not normal in 2020 and thus data for the entire season may represent a heterogenous set.

A small subset of variables were retained from the broader dataset. The subset was developed based on the author's subject matter expertise of the sport of AFL, with additional consideration given to not wanting to specify an overly complex model. The variables retained were selected based on their likely relationship to a team's goal scoring and whether a team could change their gameplay to better target these predictors. For example, the variable *free kicks against* was not included, as the number of free kicks given away by a team is not a core contributor to the same team scoring goals.

The variables that were retained for the purposes of this analysis included team-match-level counts of goals, marks, handballs, hit outs, tackles, rebounds, inside 50s, clearances, clangers, free kicks for, contested possessions, contested marks, and marks inside 50.

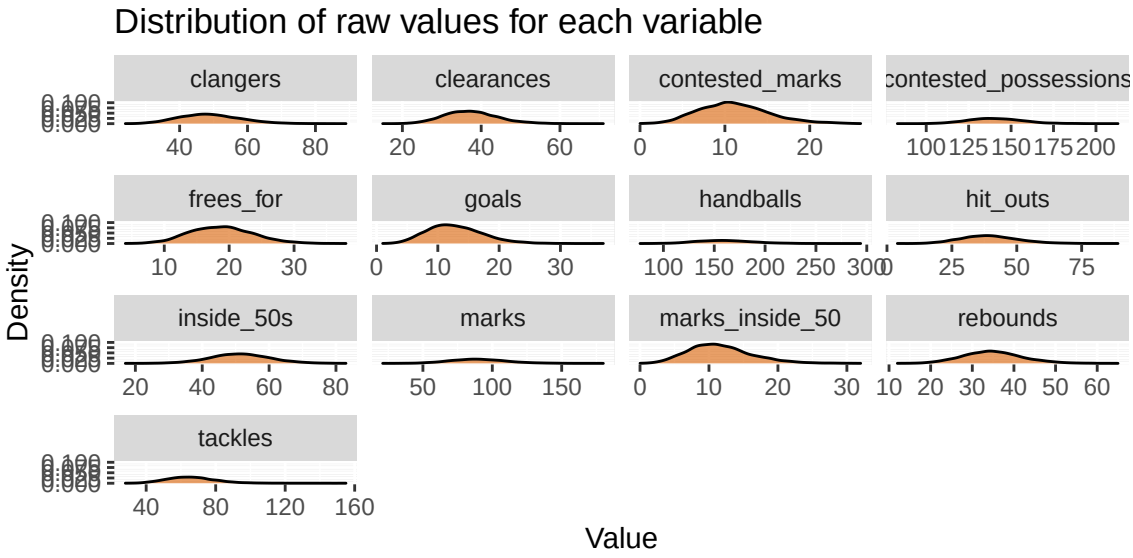
3 Analysis

A rigorous and detailed linear modelling pipeline was implemented. This involved the following steps:

1. Exploratory data analysis and preprocessing
2. Model fitting
3. Model assumption testing
4. Model re-specification (if required)
5. Model interpretation
6. Preliminary advanced model exploration

3.1 Exploratory data analysis and visualisation

Prior to modelling, the data were explored visually and numerically to understand the empirical structure. **FIGURE X** below shows the distributions of each quantitative variable.



The data were further explored using high-level summary statistics. These are presented below in **FIGURE XX**. Note the large difference in scales between the variables. To avoid issues with high-variance predictors influencing linear modelling or producing extremely low coefficients, all predictors were mean-centred and standardised (z-scored) prior to modelling.

% latex table generated in R 4.0.2 by xtable 1.8-4 package % Mon May 3 23:12:36 2021

	type
1	numeric

3.2 Model fitting

3.3 Model assumption testing

There are four core assumptions of linear regression model (see Faraway 2004). These include:

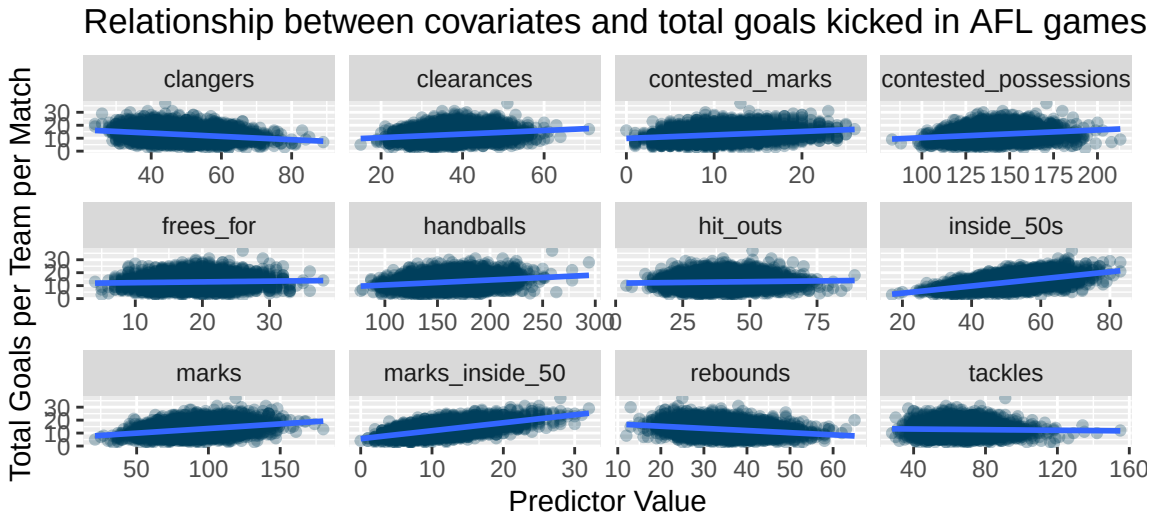
1. Linear relationship between `X` and `y`
2. Independent observations

3. Homogeneity of variance
4. Normality of residuals

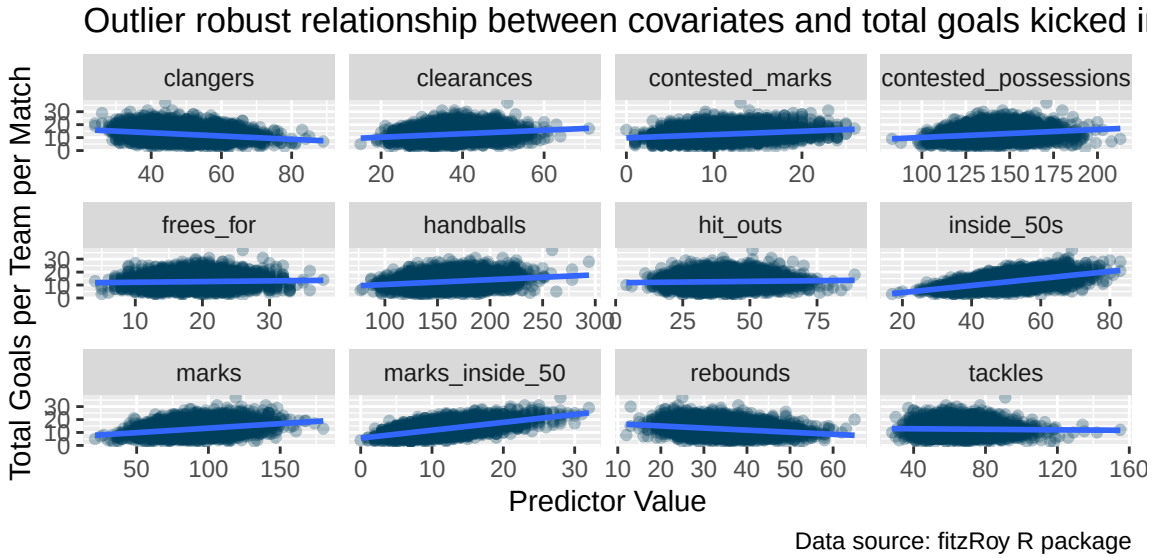
Since it is known that the data used for this report as independent observations, the following sections will focus on reporting the testing of the other assumptions.

3.3.1 Assumption 1: Linear relationship

The purpose of a linear model is to understand the relationship between some number of predictors and a quantitative response variable. As such, a linear model at its core assumes that all predictors are related linearly to the response variable.



A secondary visual test was conducted with a robust regression (using M-estimation) as the plots above appeared to contain some potential leverage points or outliers. The robust version of the plot is depicted below in **FIGURE X**. With almost no visual difference between the standard linear and the robust lienar approaches, the standard linear model was taken forward.

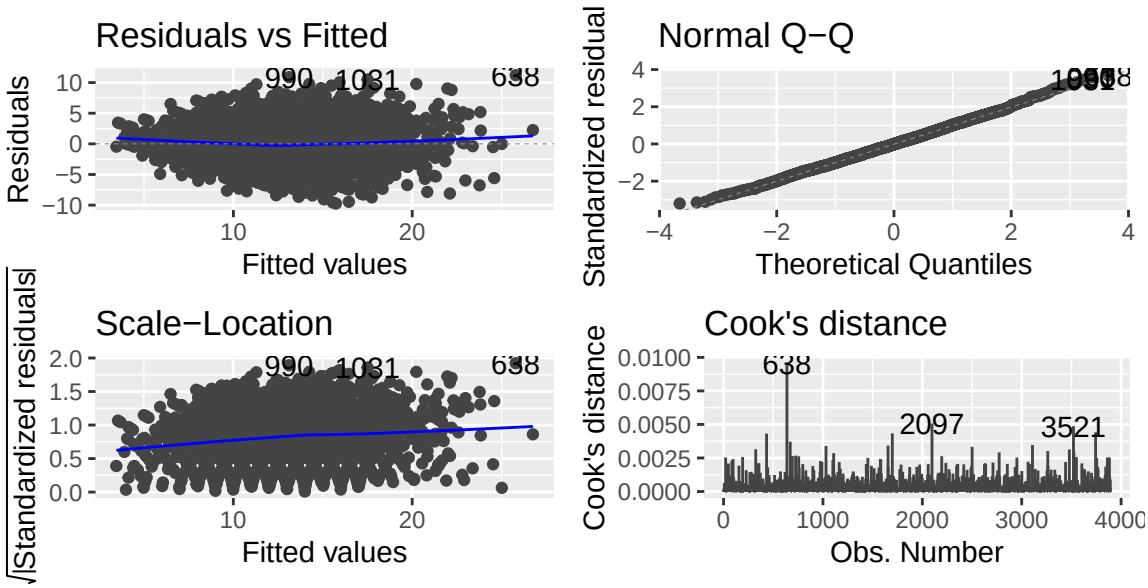


While all variables were still being tested for appropriateness, a variance inflation factor (VIF) test was undertaken to estimate potential multicollinearity between the predictors. Multicollinearity is an issue as it can drive imprecise estimates, change parameter value signs, and impact R^2. Different threshold values exist for the VIF, with cutoffs ranging from values less than four being acceptable (see Hair et al. 2010) to

values less than ten being acceptable (see Hair et al. 1995). Outputs from the VIF test are presented below. Evidently, no predictor violates even the lowest bound commonly cited in the literature, indicating no issue with multicollinearity.

```
##                 Variables Tolerance      VIF
## 1                   marks 0.5689153 1.757731
## 2               handballs 0.8070175 1.239130
## 3                hit_outs 0.7752018 1.289987
## 4                 tackles 0.7890447 1.267355
## 5                rebounds 0.7090882 1.410262
## 6              inside_50s 0.4915640 2.034323
## 7              clearances 0.5586642 1.789984
## 8                clangers 0.7929755 1.261073
## 9               frees_for 0.9170373 1.090468
## 10 contested_possessions 0.3844856 2.600878
## 11       contested_marks 0.7361132 1.358487
## 12       marks_inside_50 0.5647187 1.770793
```

3.3.2 Assumption 2: Homogeneity of variance



3.3.3 Assumption 3: Normality of residuals

3.4 Preliminary advanced model exploration

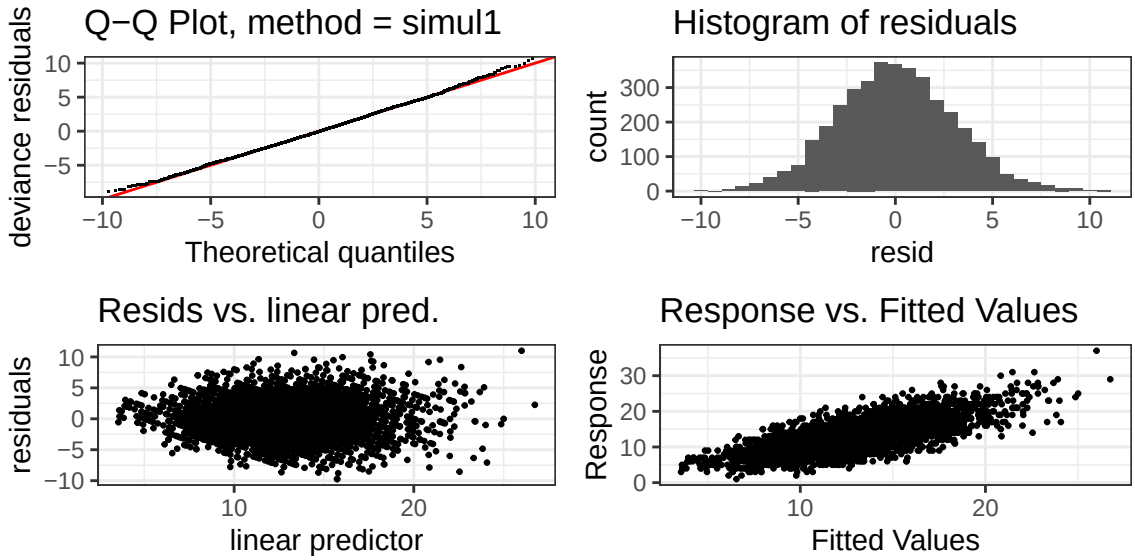
One other model was fit in addition to the standard linear model - a generalised additive model (GAM) (see Hastie and Tibshirani 1986). GAMs further generalise the commonly-used generalised linear model (GLM) to accommodate immensely powerful flexibility and potential to model non-linearities. GAMs achieve this through the use of splines and basis functions whose number is specified by a knot parameter, and who are connected by polynomials. GAMs essentially enable the fitting of wiggly functions over the data. The basic form of a GAM can be written as follows, where the predictors are still entered linearly, but they are instead modelled using some unknown smooth functions:

$$y_i = \beta_0 + f_1(x_i) + f_2(x_i)... + f_n(x_n) + \epsilon_i$$

Where ϵ is (in the standard linear modelling case) Gaussian noise $\mathcal{N}(\mu, \sigma^2)$, specified by its mean and standard deviation. Of course, similar to GLMs, this Gaussian noise assumption is generalised to other probability distributions, though these are not the considered in this report. The GAM for this report was fit in `R` using the `mgcv` package (see Wood 2011). It was fit using Restricted Maximum Likelihood for reduced-rank model parameter estimation, as per advice by Wood (see Wood, n.d.).

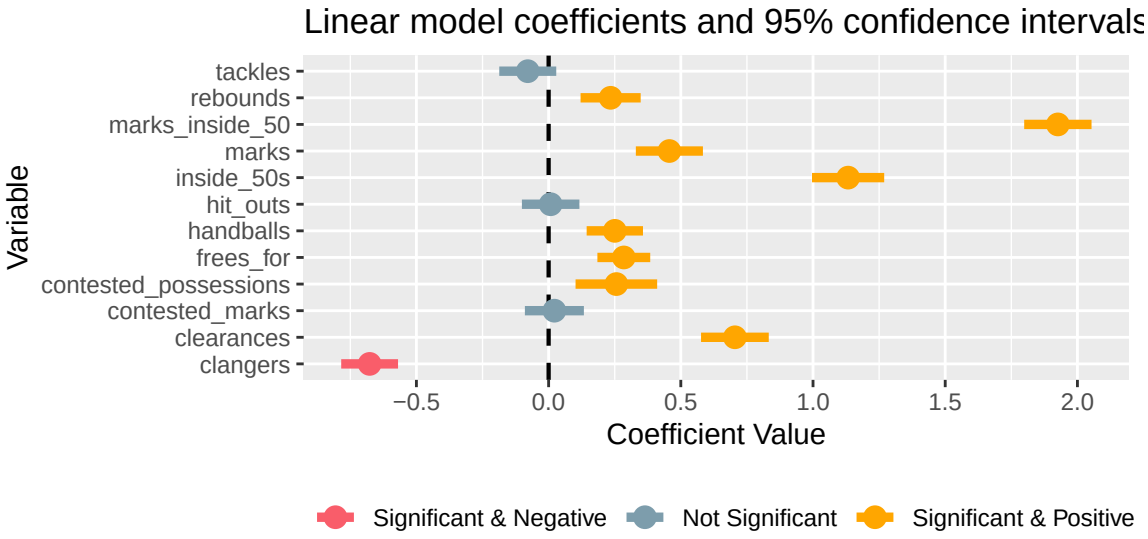
3.4.1 Model assumption testing

Similar to the standard linear model, core assumptions still need to hold for the GAM. These were also tested, with a summary output presented below in **FIGURE 1** generated by the `R` package `mgcViz` (see Fasiolo et al. 2018).



4 Results

4.1 Linear model



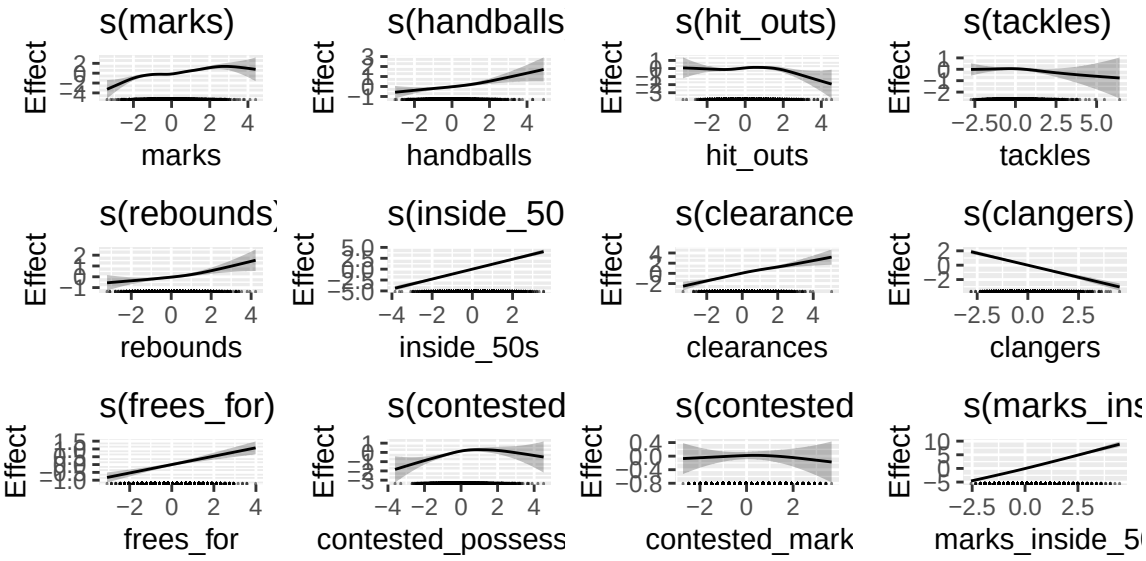
A numerical exploration of coefficients is presented below. This table also contains information regarding the overall model fit and F-statistic. The overall model is statistically significant, $F = 370.77$ ($df = 12; 3879$), and explains approximately 53.4% of the observed variance in goals scored.

4.2 GAM

XX

Table 1:

	Dependent variable:
	goals
marks	0.457***
	(0.065)
handballs	0.250***
	(0.054)
hit_outs	0.008
	(0.055)
tackles	−0.079
	(0.055)
rebounds	0.235***
	(0.058)
inside_50s	1.133***
	(0.070)
clearances	0.705***
	(0.065)
clangers	−0.677***
	(0.055)
frees_for	0.284***
	(0.051)
contested_possessions	0.256***
	(0.079)
contested_marks	0.022
	(0.057)
marks_inside_50	1.926***
	(0.065)
Constant	12.812***
	(0.049)
Observations	3,892
R^2	0.534
Adjusted R^2	0.533
Residual Std. Error	3.044 (df = 3879)
F Statistic	370.774*** (df = 12; 3879)
Note:	*p<0.1; **p<0.05; ***p<0.01



5 Discussion

References

Day, James, Robert Nguyen, and Oscar Lane. 2020. *FitzRoy: Easily Scrape and Process Afl Data*. https://CRAN.R-project.org/package=fitzRoy.

Faraway, Julian J. 2004. *Linear Models with R*. Chapman & Hall/CRC. http://www.maths.bath.ac.uk/%20jjf23/LMR/.

Fasiolo, Matteo, Raphael Nedellec, Yannig Goude, and Simon N. Wood. 2018. "Scalable Visualisation Methods for Modern Generalized Additive Models." *Arxiv Preprint*. https://arxiv.org/abs/1707.03307.

Hair, J. F., R. E. Anderson, R. L. Tatham, and W. C. Black. 1995. *Multivariate Data Analysis (3rd Ed.)*. Macmillan Publishing Company, New York.

Hair, J. F., W. C. Black, B. J. Babin, and R. E. Anderson. 2010. *Multivariate Data Analysis (7th Ed.)*. Upper saddle River, New Jersey: Pearson Education International.

Hastie, Trevor, and Robert Tibshirani. 1986. "Generalized Additive Models." *Statistical Science* 1 (3): 297–310. https://doi.org/10.1214/ss/1177013604.

Wood, S. N. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society (B)* 73 (1): 3–36.

———. n.d. "Frequently Asked Questions for Package Mgcv." http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/mgcv/html/mgcv-FAQ.html.