A linear modelling exploration of predictors of scores in the AFL

A Preprint

Trent Henderson
OLET5608

then6675@uni.sydney.edu.au

May 27, 2021

Abstract

AFL is a highly-popular Australian sport that has garnered a lot of talk show attention, but suffers from a lack of statistical rigour. The present report seeks to bridge this gap by providing a statistically robust exploration of predictors of scores using data aggregated at the team and match level from the 2005-2019 seasons inclusive. An ordinary least squares regression model was used alongisde preliminary exploration of a more sophisticated generalised additive model approach. Robust variance-covariance matrix estimators were used due to the presence of mild heteroscedasticity. Results found that tackles and unforced errors significantly and negatively predicted match scores, and rebounds, marks inside 50, marks, inside 50s, handballs, free kicks for, and clearances all significantly and positively predicted match scores. Contested marks and hit outs did not significantly predict match scores. Implications for coaching and gameplay strategy as well as limitations are discussed.

***K*eywords** linear modelling, AFL, heteroscedasticity, robust estimation, sports analytics

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 COVID-19-impacted 2020 season) approximating a sold out 100,000 each year at the traditional host venue - the Melbourne Cricket Ground (Tables 2021). 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). Despite its popularity and complexity, AFL is a sport that has traditionally relied on subject matter expertise and the knowledge of past players to inform coaching strategies. Much like other Australian sports, a lack of empirical statistical sophistication is evident.

Globally, sports analytics has continued to generate increasing attention, with websites such as FiveThirtyEight and Advanced Sports Analytics creating stylish platforms that constitute a reliable source of insight and interactive analysis. However, this form of innovative and detailed analysis has yet to fully permeate Australian sports. While the AFL has many 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 descriptive statistics and not on statistical rigour. This report aims to bridge some of this gap by providing a preliminary statistical investigation of factors associated with scoring in the AFL. Specifically, this report aims to explore the following research question: *Which gameplay attributes are predictors of scores in AFL matches?*

2 Data set

Historical AFL data has been made readily-accessible in an open-source setting through the `R` package `fitzRoy` (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 2005-2019 seasons, inclusive. This time period is somewhat arbitrary, but was made on the basis of wanting a large sample size while balancing recency and 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.

2.1 Data limitations

Despite the availability of so much player-level data, the author of the `fitzRoy` package and the creators of the sources it pulls from all note potential caveats around the data. The main caveat is that the data is not official. Each source pulls from multiple others, and many individual people are involved in the continual updating of information. The accuracy of the data in `fitzRoy` is largely contingent on the accuracy of the sources underpinning the websites it scrapes. While this is cause for concern, there are a large number of industry-standard sources that comprise the majority of the data used in this report, including official statistics produced by the AFL, newspapers and magazines (such as The Herald Sun and Inside Football), and official books (such as Main and Holmesby (2018) and Rodgers (1996)). The open-source nature of many of the sources, especially AFL Tables, means continual improvement and accuracy is being achieved, further lending confidence to the available data, though some caution is still advised.

2.2 Variable retention

A small subset of variables were retained from the larger dataset. The subset was developed based on the author's AFL subject matter expertise. The variables retained were selected based on their likely relationship to a team's ability to score and whether a team could implement a training or coaching intervention off the back of this analysis to better target important 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 that team scoring, and it is likely near impossible to coach out of their game.

The variables that were retained for the purposes of this analysis included team-match-level counts of scores, marks, handballs, hit outs, tackles, rebounds, inside 50s, clearances, clangers (unforced errors), 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, each of which will be discussed in turn:

1. Exploratory data analysis and visualisation
2. Model fitting and assumption testing

3.1 Exploratory data analysis and visualisation

Prior to modelling, the data were aggregated and explored visually and numerically to understand the empirical structure. The data was aggregated to match-level sums for each team by summing over individual

Figure 1: Distribution of raw values for each variable

player statistics. Matches outside of regular season games (i.e. finals) were removed as finals likely represent a heterogenous set. Figure 1 shows the distributions of each aggregated quantitative variable.

The data were further explored using high-level summary statistics. These are presented in Table 1. Note the large difference in scales between the variables. To avoid issues with high-variance predictors (due to scale) influencing linear modelling or producing extremely low coefficients, all predictors were mean-centred and standardised (z-scored) prior to modelling. This also means the coefficients will have an intuitive interpretation compared to other rescaling methods.

Variable	**n**	**Min**	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	**Max**	**s**	**IQR**	**#NA**
score	5652	13	70	88	90.3	108	233	27.9	38	0
marks	5652	21	79	93	93.5	107	181	21.1	28	0
handballs	5652	58	134	155	155.5	175	297	30.6	41	0
hit_outs	5652	4	29	36	37.2	44	89	11.0	15	0
tackles	5652	17	51	61	61.6	71	155	15.1	20	0
rebounds	5652	12	30	35	34.9	39	65	6.9	9	0
inside_50s	5652	17	45	51	51.0	57	83	8.5	12	0
clearances	5652	14	31	35	35.8	40	71	6.9	9	0
clangers	5652	21	41	47	47.5	53	89	8.8	12	0
frees_for	5652	4	15	18	18.6	22	38	4.8	7	0
contested_possessions	5652	78	120	133	133.5	146	213	18.7	26	0
contested_marks	5652	0	8	10	10.7	13	27	4.0	5	0
marks_inside_50	5652	0	9	12	12.3	15	38	4.8	6	0

Table 1: Descriptive statistics for all quantitative variables

3.2 Model fitting and assumption testing

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

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

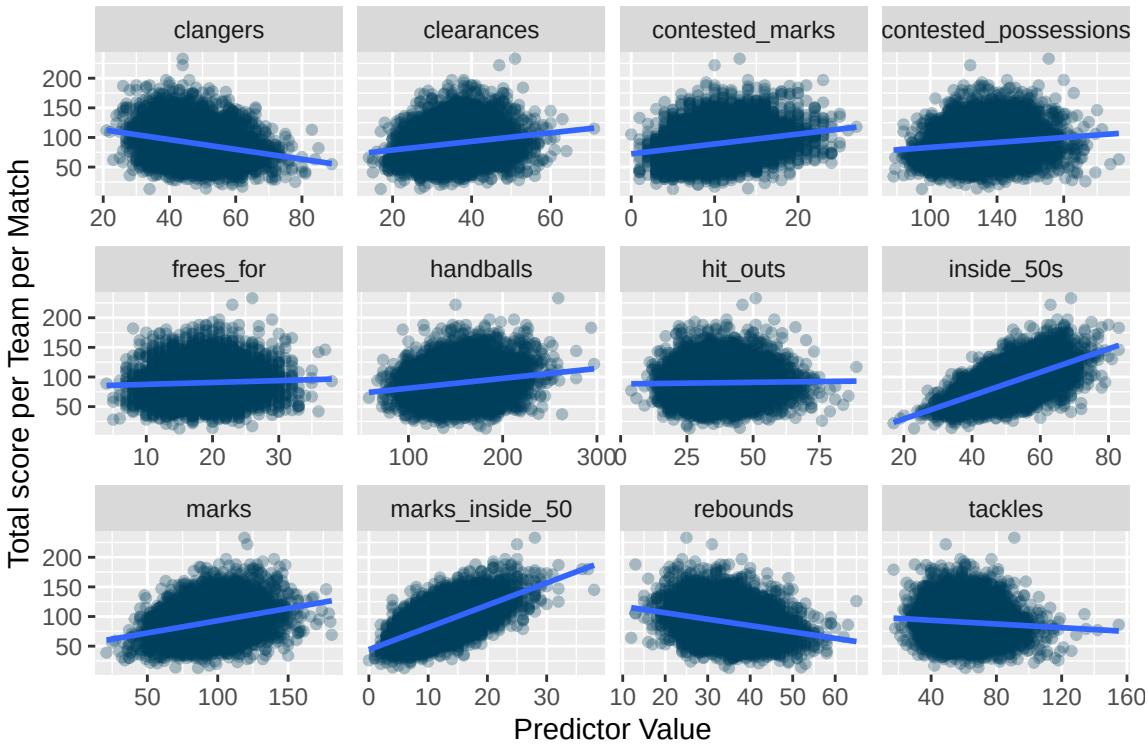

Figure 2: Relationships between covariates and total score in AFL games

3. Normality of residuals
4. Homogeneity of variance

Since the data is at the most independent level possible for the present analysis (acknowledging the potential that some relationship may exist between each team on the same match), the following sections will focus on reporting the testing of the other assumptions.

3.2.1 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. These bivariate relationships are presented in Figure 2. At this stage, a preliminary linear ordinary least squares (OLS) model was fit which confirmed the visual hypothesis that two variables - *contested marks* and *hit outs* - were not significantly associated with total scores. These variables were dropped for the remaining analysis. A follow-up assessment of linearity using a residuals versus fitted plot was conducted (see top left plot in Figure 3). In this plot, a slight quadratic shape is noted on the residuals versus fitted plot. Three new models were fit in response to this: OLS with second-degree polynomial terms on suspect predictors, OLS with square-root-transformed response, and OLS with log-transformed response. These models introduced new issues without addressing the underlying problem, and given that the quadratic shape was only slight with the data points themselves looking rather evenly-dispersed, the additional models were not retained.

While all variables were 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 (Hair et al. 2010). Different threshold values exist for VIF, with cutoffs ranging from values less than four being acceptable (Hair et al. 2010) to values less than ten being acceptable (Hair et al. 1995). Outputs from the VIF test are presented below in Table 2. Evidently, no predictor violates even the lowest bound commonly cited in the literature, indicating no issue with multicollinearity.

	Variables	Tolerance	VIF
1	marks	0.60	1.67
2	handballs	0.81	1.23
3	tackles	0.67	1.49
4	rebounds	0.76	1.31
5	inside_50s	0.55	1.83
6	clearances	0.57	1.75
7	clangers	0.80	1.25
8	frees_for	0.91	1.10
9	contested_possessions	0.36	2.77
10	marks_inside_50	0.59	1.68

Table 2: Variance inflation factor and tolerance estimates

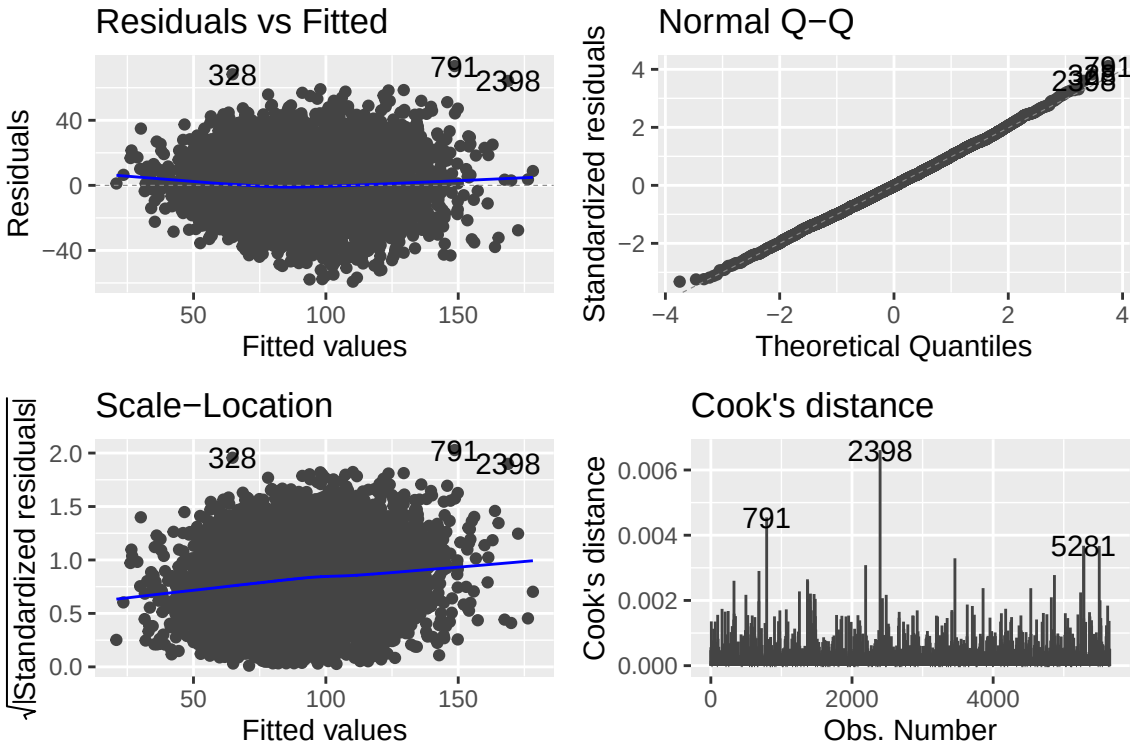

Figure 3: Linear model diagnostic plots

3.2.2 Normality of residuals

Normality of residuals is typically assessed graphically using a Q-Q plot, as seen in the upper right graphic in Figure 3. A model with normally-distributed residuals should lie directly on the diagonal line. The residuals are almost entirely positioned on the line with very little variation at the ends indicating no issues with normality.

3.2.3 Homogeneity of variance

Homoegeneity of variance - the lack of a systematic pattern or bias of residuals across model fitted or predictor values - is another core linear model assumption. This assumption is typically assessed graphically using a residuals and/or standardised plot. A model with homogeneity of variance should have no discernible pattern across the fitted values. This plot is depicted in the bottom left in Figure 3.

Evidently, there is a non-horizontal line through the plot, indicating potential heteroscedasticity. Visual inspection of the data points themselves suggests only mild heteroscedasticity as they look reasonably evenly-dispersed. It was first hypothesised that potential outliers might be influencing the results, despite

the lack of compelling visual evidence of leverage based on the Cook's distance plot. Following advice from (Faraway 2004), a test of the maximum studentised residual value against a Bonferroni-corrected critical value was conducted. Since the maximum residual value of 4.14 was less than the critical value of 4.45, it was declared that outliers were not an issue.

Heteroscedasticity is a major issue for linear models. As a response to this potential violation, a weighted OLS model was tested. The weighted OLS model works by computing weights for each data point, calculated by the inverse of squared fitted values from a linear regression of the absolute residuals of the original model as the response variable, and the fitted values of the original model as the predictor. The weight vector is then factored into the matrix decomposition to solve the linear regression problem. This method did not fix the heteroscedasticity issue.

As a solution, a heteroscedastic-robust estimator was used, which produces robust estimations of standard errors, test statistics, and p-values. This solution may seem like a highly conservative response to the relatively weak violation, however, erring on the side of caution could be considered a safe option in applied settings. Robust estimators are implemented in `R` using the `sandwich` package (Zeileis 2004; Zeileis, Köll, and Graham 2020). The estimators work by introducing a new term, Ω, that flexibly acts on the diagonal of the variance-covariance matrix, and relaxes the assumption of homogeneity by enabling differing variances along the matrix diagonal (see Equation (1) and (2)). The inclusion of heteroscedastic-robust estimators reduces the size of test statistics, drives significance values away from zero, and increases standard errors to reflect the variance structure of the data.

$$(X^{'}X)^{-1}X^{'}\Omega X(X^{'}X)^{-1} \tag{1}$$

$$\Omega = \sigma^2 I_n \tag{2}$$

Numerous Ω options are available in the `sandwich` package. Most of the options returned negligibly different values for the present analysis, so the default `HC3` parameter recommended by the authors was retained. It is defined according to Equation (3).

$$HC3 = \frac{\mu_i^2}{(1-h_i)^2} \tag{3}$$

4 Results

Coefficients and model outputs are presented in Table 3 where the first column of the dependent variable section is the standard OLS model and the second column is the heteroscedastic-robust corrected model. Interpretation will focus on the robust estimators, given the violation of homoscedasticity. For each predictor, coefficients, standard errors, t-statistics, and p-values are reported. Since all predictors were mean-centred and standardised (z-scored) prior to analysis, the interpretation is as follows: *the coefficient represents the expected change in total score (response variable) for a one standard deviation change in the predictor*. The overall model is statistically significant, $F = 811.8$ ($df = 10$; 5641), and explains approximately 59% of the observed variance in scores.

Two predictors were negative and statistically significant. These were tackles (t = -4.28, $p < .001$) and clangers (t = -15.54, $p < .001$), such that a one standard deviation increase in tackles is associated with mean reduction of 1.2 in total score, and a one standard deviation increase in clangers is associated with mean reduction of 4.2 in total score. For the positive predictors, the two with the strongest coefficients are mechanically related in terms of AFL gameplay: inside 50s ($t = 28.03$, $p < .001$) and marks inside 50 ($t = 38.6$, $p < .001$). The magnitude of both these predictors is noteworthy, as a one standard deviation increase in inside 50s is associated with a mean increase of 9.3 in total score, and a one standard deviation increase in marks inside 50 is associated with a mean increase of 12.2 in total score. The remaining positive predictors are reported in Table 3.

5 Discussion

The present analysis aimed to produce an innovative and statistically robust exploration of predictors of scoring in the AFL using team-per-match-level data for the 2005-2019 seasons inclusive accessed through the

Table 3: Model coefficients, confidence intervals, and signifiance tests

	Dependent variable:	
	score	
	OLS	*coefficient test*
	(1)	(2)
marks	2.321 (1.720, 2.922) t = 7.570***	2.321 (1.721, 2.921) t = 7.583***
	p = 0.000	p = 0.000
handballs	1.198 (0.681, 1.714) t = 4.541***	1.198 (0.676, 1.719) t = 4.502***
	p = 0.00001	p = 0.00001
tackles	−1.224 (−1.793, −0.656) t = −4.221***	−1.224 (−1.785, −0.663) t = −4.276***
	p = 0.00003	p = 0.00002
rebounds	0.980 (0.448, 1.513) t = 3.609***	0.980 (0.440, 1.521) t = 3.557***
	p = 0.0004	p = 0.0004
inside_50s	9.317 (8.687, 9.947) t = 28.974***	9.317 (8.666, 9.969) t = 28.029***
	p = 0.000	p = 0.000
clearances	4.198 (3.582, 4.815) t = 13.348***	4.198 (3.562, 4.835) t = 12.930***
	p = 0.000	p = 0.000
clangers	−4.163 (−4.684, −3.642) t = −15.667***	−4.163 (−4.688, −3.638) t = −15.542***
	p = 0.000	p = 0.000
frees_for	2.015 (1.528, 2.503) t = 8.102***	2.015 (1.526, 2.505) t = 8.071***
	p = 0.000	p = 0.000
contested_possessions	1.009 (0.233, 1.784) t = 2.550*	1.009 (0.208, 1.810) t = 2.469*
	p = 0.011	p = 0.014
marks_inside_50	12.202 (11.598, 12.806) t = 39.599***	12.202 (11.583, 12.822) t = 38.618***
	p = 0.000	p = 0.000
Constant	90.267 (89.802, 90.733) t = 380.134***	90.267 (89.801, 90.733) t = 379.740***
	p = 0.000	p = 0.000
Observations	5,652	
R^2	0.590	
Adjusted R^2	0.589	
Residual Std. Error	17.852 (df = 5641)	
F Statistic	811.824*** (df = 10; 5641)	
Note:		*p<0.05; **p<0.01; ***p<0.001

`R` package `fitzRoy`. While not necessarily causal, the analysis sought to quantify the type and magnitude of any relationships with end-of-match scores.

5.1 Implications for AFL teams

This report found some potentially informative relationships regarding scoring in the AFL that teams may seek to consider. First, teams should seek to deeply understand their potential to generate opportunities within the fifty-metre circle in front of goal. The analysis strongly supports this recommendation, as increases in inside 50s and marks inside 50 are both associated with a substantial increase in total score. This is intuitive from a gameplay sense, as being closer to goal with possession of the ball would increase the likelihood of scoring, and a mark inside 50 means a guaranteed uninterrupted set shot at goal, further increasing the likelihood of kicking a six-point goal.

Second, teams should also consider the importance of clearances. The strong positive association found between clearances and scores was surprising. This is because clearances involve a team kicking the ball away from their own goal area, which is a heavily defensive statistic. The positive relationship may suggest that the opposition team was unsuccessful in scoring on multiple occasions, and so the team could take advantage of converting a successful defense into attacking opportunities of their own.

Third, teams should be cautious not to interpret the causal direction of some of the relationships presented in this paper. The negative relationship between tackles and scores is one such example. It is not necessarily the case that tackling less directly results in higher scores at the end of a match. It is far more likely that teams who score more (therefore more likely winning more) are just more defensively efficient or spend more time attacking rather than defending. Both of these characteristics would manifest as noticeably lower tackle counts.

5.2 Limitations

Despite the potentially informative findings, there were some limitations with the analysis. The first, as described earlier, is that the data is not official, and therefore its accuracy is unknown. It is likely that the data quality is high, given that some of the underlying sources are official and published material and that the project is open-source with contributions for numerous high-profile researchers and analysts.

A second limitation is that of variable selection. The variables included in this analysis were selected based on the author's subject matter expertise and prior knowledge of AFL. However, these variables only explained roughly sixty per cent of the variance in match scores. It is highly likely that the addition of more variables included in the larger dataset of approximately sixty variables would help drive this number closer to a more respectable percentage, such as eighty or ninety per cent. Since factor variables are included in the broader dataset, their inclusion raises some interesting questions around interaction terms. For example, future research may seek to fit interaction terms by team, or by home versus away, to better understand the dynamics of AFL metrics on match scores. Of course, the inclusion of more covariates, especially large numbers of them, may raise serious issues around multicollinearity or other model assumptions. Researchers may seek to account for this by first applying variable selection procedures such as Lasso regression (Tibshirani 1996).

A third limitation is that of model selection. It remains unclear whether an ordinary least squares regression approach is the optimal modelling technique for this data. Preliminary follow-up analysis undertaken by the author revealed that a generalised additive model (Hastie and Tibshirani 1986; Wood, n.d., 2011) - a model that linearly adds estimated smooth functions using splines for each covariate - produced a better model fit at a lower Akaike information criterion value (Posada and Buckley 2004). Further, since the response variable is a non-zero count, it may be more appropriate to consider a generalised linear model with a link function appropriate to an integer response, such as a Poisson or negative binomial-distributed model. The added benefit of these models is that they correctly model the response as a discrete-valued probability mass function, instead of the probability density function assumed by a Gaussian linear model (if a maximum likelihood and not ordinary least squares approach is taken). This may be particularly pertinent if future endeavours focus on predictive applications. Future research should aim to consider these modelling options, and potentially even perform a direct comparison.

6 Available code

All code for this paper is available on GitHub.

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/.

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.

Main, Jim, and Russel Holmesby. 2018. *The Encyclopedia of Afl Footballers: Every Afl/Vfl Player Since 1897*. Bas Publishing.

Posada, David, and Thomas R. Buckley. 2004. "Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests." *Systematic Biology* 53 (5): 793–808. https://doi.org/10.1080/10635150490522304.

Rodgers, Stephen. 1996. *Every Game Ever Played: VFL/Afl Results, 1897-1995*. Viking.

Tables, AFL. 2021. "Grand Finals." https://afltables.com/afl/teams/allteams/gfgames.html.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88. https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

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.

Zeileis, Achim. 2004. "Econometric Computing with Hc and Hac Covariance Matrix Estimators." *Journal of Statistical Software, Articles* 11 (10): 1–17. https://doi.org/10.18637/jss.v011.i10.

Zeileis, Achim, Susanne Köll, and Nathaniel Graham. 2020. "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R." *Journal of Statistical Software, Articles* 95 (1): 1–36. https://doi.org/10.18637/jss.v095.i01.