A linear modelling exploration of predictors of scores in the AFL

A Preprint

Trent Henderson
OLET5608

then6675@uni.sydney.edu.au

May 8, 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 sought to bridge this gap by providing a statistically robust exploration of predictors of scoring using data aggregated at the team and match level. A regression approach was adopted through the use of a hetereoscedastic-robust ordinary least squares model. Results found that hit outs contested possessions, and contested marks did not significantly predict scoring, tackles and unforced errors singificantly and negatively predicted scoring, and rebounds, marks inside 50, marks, inside 50s, handballs, free kicks for, and clearances all significantly and positively predicted scoring. Implications for coaching and gameplay strategy, as well as limitations are discussed.

***K*eywords** linear modelling, AFL, heteroscedasticity

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). 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, sport 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 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 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 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.

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 (listed above) all note potential caveats around their 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 scraped. 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.

2.2 Variable retention

A small subset of variables were retained from the much larger dataset. The subset was developed based on the author's subject matter expertise of AFL. 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 the 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, and it is likely near impossible to coach out of the 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
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 aggregated and explored visually and numerically to understand the empirical structure. The data was aggregated to match-level sums for each time by summing over individual

Figure 1: Distribution of raw values for each variable

player statistics. Figure @ref(fig:distplot) below shows the distributions of each aggregated quantitative variable.

The data were further explored using high-level summary statistics. These are presented below in Table @ref(tab:summarystats). 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. This also means the coefficients will have an intuitive interpretation compared to other rescaling methods.

% latex table generated in R 4.0.2 by xtable 1.8-4 package % Sat May 8 19:53:06 2021

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

A detailed linear model fitting process was undertaken, and included specification of initial models, assessment of diagnostics, and re-specification of the model. Details are presented in the following sections.

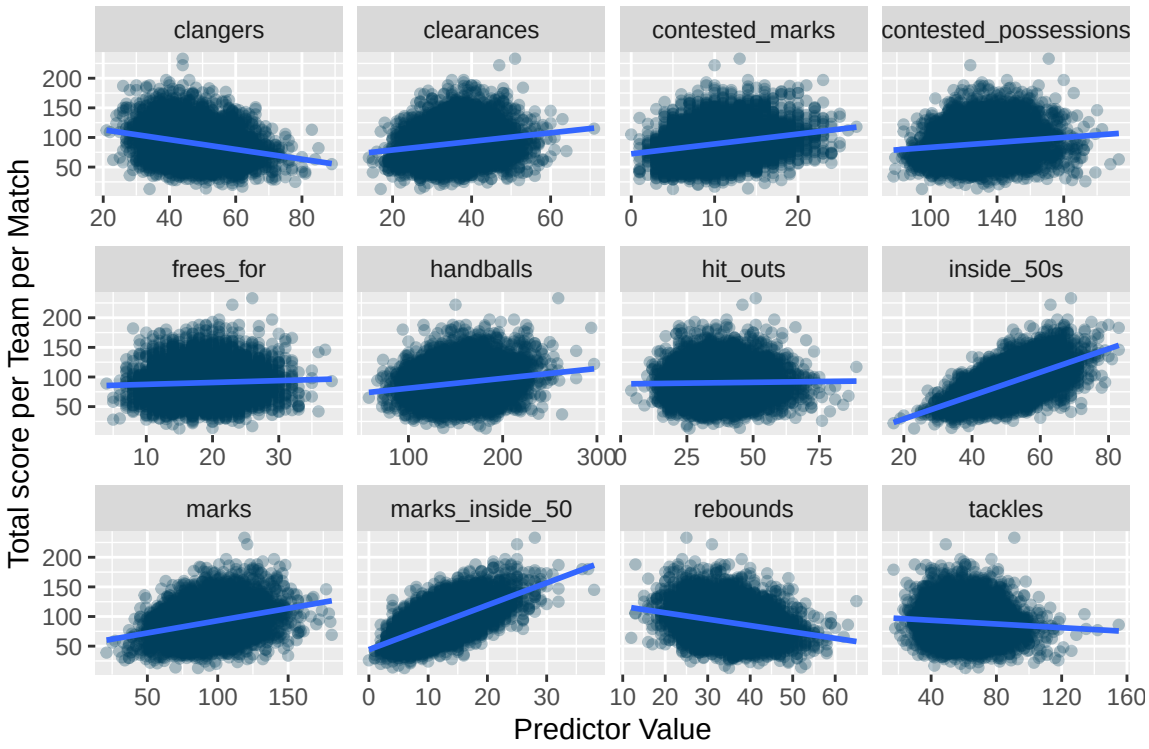

Figure 2: Linear relationships between covariates and total score in AFL games

3.3 Model 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`
3. Normality of residuals
4. Homogeneity of variance

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

3.3.1 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.

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 @ref(tab:vif). Evidently, no predictor violates even the lowest bound commonly cited in the literature, indicating no issue with multicollinearity.

3.3.2 Normality of residuals

Normality of residuals is typically assessed graphically using a Q-Q plot, as seen in the upper right graphic in the matrix presented below A model with normally-distributed residuals should lie directly on the diagonal

	Variables	Tolerance	VIF
1	marks	0.57	1.76
2	handballs	0.79	1.26
3	hit_outs	0.70	1.44
4	tackles	0.66	1.51
5	rebounds	0.76	1.32
6	inside_50s	0.54	1.86
7	clearances	0.50	2.02
8	clangers	0.79	1.26
9	frees_for	0.91	1.10
10	contested_possessions	0.30	3.33
11	contested_marks	0.74	1.35
12	marks_inside_50	0.57	1.76

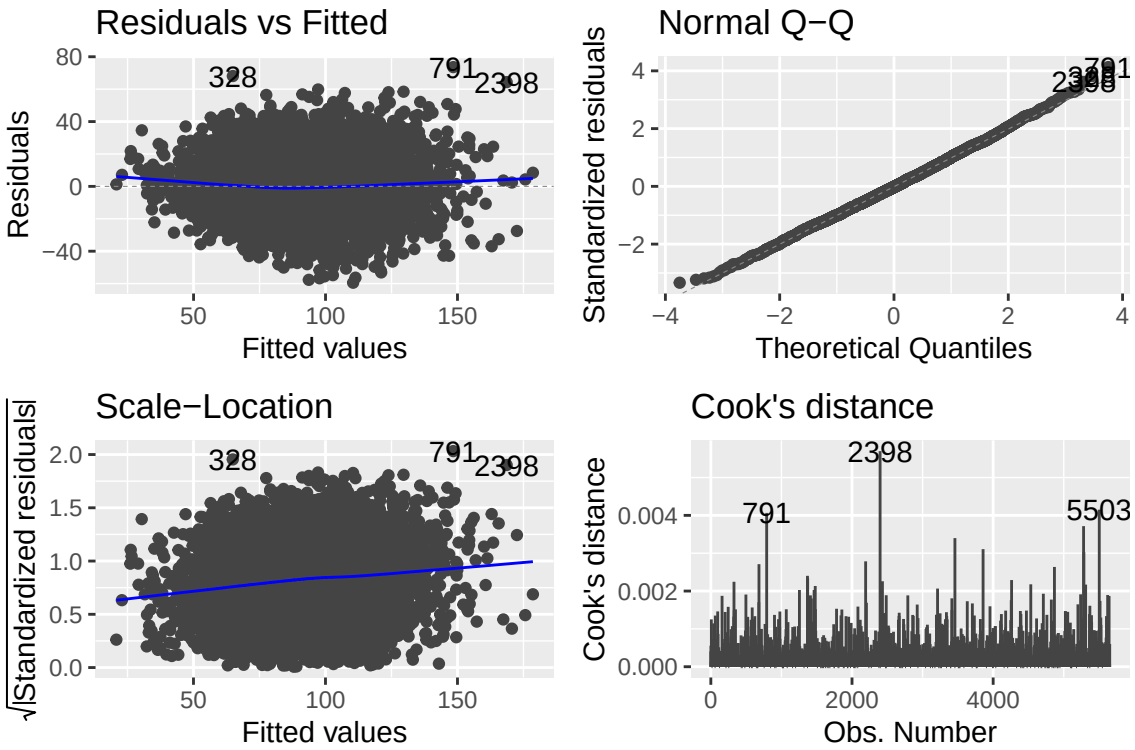

Figure 3: Linear model diagnostic plots

line. The residuals are almost entirely positioned on the line with very little variation at the ends indicating no issues with normality.

3.3.3 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 plot, where fitted values are plotted against model residuals. A model with homogeneity of variance should have no discernible pattern across the fitted values. This plot is depicted in the upper left in the graphics matrix above

Evidently, there is a quadratic-shaped sag in the line through this plot, indicating heteroscedasticity. 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. Since the maximum

residual value of 4.14 was less than the critical value of 4.45, it was declared that outliers were not a major issue.

Heteroscedasticity is a major issue for linear models. As a response to this initial violation, two follow-up models were specified: a weighted ordinary least squares (OLS) model, and an OLS with a square root transformed response variable. The weighted OLS model works by computing weights for each datapoint included, 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. Neither of these two models fixed the heteroscedasticity violation in the present study, and the square root transform actually introduced violations in other core assumptions, particularly normality of residuals. As a solution, a heteroscedastic-robust estimator was used, which produces robust estimations of standard errors, test statistics, and p-values. 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 acts on the diagonal of the variance-covariance matrix, and relaxes the assumption of homogeneity. 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.

4 Results

Coefficients and model outputs are presented below in Table @ref(tab:coeftable) where the first column of the dependent variable section is the standard OLS model and the second column is the heteroscedastic-robust correction. Intepretation 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 change in total score (response variable) for a one standard deviation change (increase or decrease, depending on sign of the coefficient) in the predictor.* The overall model is statistically significant, $F = 676.7$ ($df = 12$; 5639), and explains approximately 59% of the observed variance in scores.

Hit outs (t = -0.896, $p = 0.371$), contested possessions ($t = 1.94$, $p = 0.053$) and contested marks ($t = 1.1$, $p = 0.266$) were the only three statistically non-significant predictors. Of the remaining predictors, two were negative and statistically significant. These included tackles (t = -4.1, $p < .001$) and clangers (t = -15.4, $p < .001$), 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 4.1 in total score. On the positive predictors, the two with the strongest coefficients are mechanically related in terms of AFL gameplay: inside 50s ($t = 27.91$, $p < .001$) and marks inside 50 ($t = 37.7$, $p < .001$). 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 12.1 in total score.

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 accessed through the `R` package `fitzRoy`. While not 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 either or both of inside 50s and marks inside 50 is associated with a very substantial increase in total score. From a gameplay sense this is intuitive, as being closer to goal with possession of the ball would increase the likelihood of scoring, and a mark inside 50 means a guaranteed set shot (uninterrupted free shot) at goal, further increasingly 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

Table 2:

	Dependent variable:	
	score	
	OLS	*coefficient test*
	(1)	(2)
marks	2.233 (1.616, 2.851)	2.233 (1.615, 2.852)
	t = 7.089***	t = 7.079***
	p = 0.000	p = 0.000
handballs	1.238 (0.715, 1.761)	1.238 (0.710, 1.766)
	t = 4.637***	t = 4.592***
	p = 0.00001	p = 0.00001
hit_outs	−0.257 (−0.815, 0.300)	−0.257 (−0.821, 0.306)
	t = −0.905	t = −0.896
	p = 0.366	p = 0.371
tackles	−1.179 (−1.752, −0.607)	−1.179 (−1.746, −0.613)
	t = −4.035***	t = −4.081***
	p = 0.0001	p = 0.00005
rebounds	0.974 (0.439, 1.508)	0.974 (0.433, 1.514)
	t = 3.571***	t = 3.532***
	p = 0.0004	p = 0.0005
inside_50s	9.318 (8.683, 9.953)	9.318 (8.663, 9.972)
	t = 28.761***	t = 27.911***
	p = 0.000	p = 0.000
clearances	4.358 (3.697, 5.019)	4.358 (3.678, 5.038)
	t = 12.924***	t = 12.562***
	p = 0.000	p = 0.000
clangers	−4.148 (−4.671, −3.625)	−4.148 (−4.677, −3.619)
	t = −15.535***	t = −15.358***
	p = 0.000	p = 0.000
frees_for	2.020 (1.531, 2.508)	2.020 (1.529, 2.510)
	t = 8.101***	t = 8.074***
	p = 0.000	p = 0.000
contested_possessions	0.866 (0.017, 1.716)	0.866 (−0.009, 1.742)
	t = 1.999*	t = 1.939
	p = 0.046	p = 0.053
contested_marks	0.309 (−0.232, 0.851)	0.309 (−0.236, 0.855)
	t = 1.120	t = 1.113
	p = 0.263	p = 0.266
marks_inside_50	12.129 (11.511, 12.746)	12.129 (11.498, 12.759)
	t = 38.514***	t = 37.730***
	p = 0.000	p = 0.000
Constant	90.267 (89.802, 90.733)	90.267 (89.801, 90.733)
	t = 380.132***	t = 379.669***
	p = 0.000	p = 0.000
Observations	5,652	
R^2	0.590	
Adjusted R^2	0.589	
Residual Std. Error	17.852 (df = 5639)	
F Statistic	676.676*** (df = 12; 5639)	
Note:		*p<0.05; **p<0.01; ***p<0.001

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 uncovered in this report. The negative relationship between tackles and scores is a strong example of this. 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. However, it remains 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 subject matter expertise and prior knowledge of AFL on behalf of the author. 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 this broader set, 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 over each covariate - produced a better model fit at a lower Akaike information criterion (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 to solving the regression problem). Future research should aim to consider these modelling options, and potentially even perform a direct comparison.

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.

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.