Exploring Variables that Predict Success in the Australian National Basketball League: a Pilot Study

David Lillis and Jacob Kempt

New Zealand Institute of Sport

This report emerged from a research study undertaken by Jacob Kempt, in partial fulfilment of a Bachelor of Applied Science degree from the Open Polytechnic of New Zealand. Jacob was supervised by David Lillis of the New Zealand Institute of Sport. Jacob played basketball for several years with the Middleton Grange Gators in Christchurch, but now plays socially in the Christchurch Social Basketball league.

Abstract

The study described in this report explores variables that predict success in the Australian National Basketball League (NBL) and proposes statistical models that relate those variables to success. The variables considered here are technical variables that are specific to basketball, such as blocks, assists and turnovers. For the purposes of this study success is measured through three variables - win ratio, points differential and a team's final ranking in the league. Multiple regression models and generalised linear models were developed in order to identify the key variables that predict success collectively.

The main findings of the study are as follows:

1. The variables that collectively predict both win ratio and final position in the NBL are rebounds for, rebounds against, turnovers for, turnovers against, three point shots for, and field goals against.
2. The variables that collectively predict points differential are those that predict win ratio and final position, but including three pointers against and fieldgoals for.
3. Turnovers for and turnovers against are very important predictors of success.

1. Introduction

1.1 The Australian Basketball League

The Australian National Basketball League (NBL) is an Australian and New Zealand basketball league, which has been running since 1978, when games were played to a few hundred people in small suburban stadia. Today, basketball has one of the highest participation rates of all sports in Australia. The NBL now attracts more than 750,000 spectators each season, and prime-time television audiences for games that are broadcast across Australia (The National Basketball League, 2017).

We have sourced information on ten teams that have played in the NBL between 2011 and 2017, but, currently, eight teams play in the league. Eight were or are based in Australia (Adelaide, Brisbane, Cairns, Gold Coast, Illawarra, Melbourne, Perth, Sydney and Townsville), and one is from New Zealand – the Breakers.

For the purposes of this study, technical variables are basketball-specific variables that include blocks, assists, turnovers, field goals, three-pointers, rebounds, free throws and steals. All of these variables are scored both for a team or against a team. A comprehensive list of these technical variables is given in Appendix 1 of this report. The variables points for, points against and points differential are not considered technical variables, but nevertheless these variables are included within the first of the modelling procedures of this exploratory study.

1.2 Basketball–specific Variables

The NBL compiles statistics on the technical variables specific to basketball, points for and against, and each team's final position at the end of the season. The variables considered here are technical variables that are specific to basketball, such as blocks, assists, turnovers, rebounds, assists, steals, blocks, fieldgoals and three point shots. However, little research has been conducted on the relationships between the technical variables and success in the NBL, and this study provides an initial exploration of those relationships.

2. Research Questions and Methods

2.1 Research Questions

The research question of this study is: which variables collectively predict the success of teams in the NBL? This question has been delineated into three sub-questions, as follows:

1. Which variables collectively predict a team's win ratio (ratio of wins to the number of games played)?
2. Which variables collectively predict a team's points differential (mean difference between points scored for and points scored against the team across all games in a given season)?
3. Which variables collectively predict a team's final ranking on the league table (position)?

2.2 Data Collection and Analysis

NBL game statistics were sourced online through the basketball statistics website: basketball.realgm.com, and time series data on all ten teams were compiled into a single data set. This data set consists of 49 observations across 25 variables, including team name, year, points for and against, and all 16 technical variables. The data set was incomplete in that six year time series of results were available for some teams, whereas, for other teams time series of only one, two or three years were available, and only one observation was available for two particular teams. However, the data were judged to be sufficient for an exploratory study at the level of an undergraduate research project. Finally, autocorrelation (correlation between data from one period of time and data from later periods of time) was ignored for this study.

The open source software application, R Studio, was used to explore the data and investigate relationships between the technical variables and success (outcome) variables. Multiple regression models were developed in order to identify variables that predict success collectively. Multiple regression provides for one continuous outcome variable to be modelled by several continuous predictors (here the technical basketball-specific variables) and assumes approximately linear relationships between the outcome variable and the predictors. After creating initial versions of such models, stepwise elimination was used to remove any non-significant predictors (those with the highest p-value) and produce a Minimum Adequate Model – a model that explains the variation in the outcome variable while embodying as few predictors as possible. In addition, several generalised linear models were developed to model final position in the league, where position is treated as a count variable.

3. Analysis and Discussion

3.1 Our Success (Outcome) Variables

Clearly, teams with high win ratios and points differentials tend to rank high on the league table and may proceed to win the championship - the objective of every team in the NBL and indeed many sporting competitions. Clearly, our outcome variables are related. The following table gives the calculated correlations between the three outcome (success) variables used in this study (win ratio, points differential and position):

	WR	pd	Pos
WR	1	0.90	-0.91
pd		1	-0.81
Pos			1

We see that win ratio and points differential are correlated extremely strongly (r = 0.9) and we have very strong negative correlations between those two variables and position in the league. The observed negative correlations arise because high values of win ratio and points differential lead to low numeric values of Pos (i.e. low values reflect better placings). These three variables are very similar and possibly a single model is sufficient to explore the technical variables that predict success. However, for the purposes of this study, each of them was modelled separately.

3.2 Multiple Regression Models

In multiple regression we aim to create a linear model which includes two or more continuous predictors (independent variables) that together predict a single dependent continuous variable. Often we are interested in the predictive power of a particular variable above and beyond the impacts of other variables (which we call control variables).

In an Ordinary Least Squares regression with multiple predictors, we fit a model of the form:

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{1i} + \boldsymbol{\beta}_2 X_{2i} \ldots + \boldsymbol{\beta}_k X_{ki} \ldots + \mathbf{e}_i$$

. . . where $\boldsymbol{\beta}_0$ is the intercept, $\boldsymbol{\beta}_k$ are the coefficients, and $\mathbf{e}_i$ are the error terms.

In creating multiple regression models it is highly important to check for normality of errors and homoscedasticity (approximate constancy of variance about the regression line across the range of values of the predictor), and that the predictors are not too highly correlated (we have the problem of multi-collinearity when correlations between predictors exceeds approximately 0.8). In addition, we want to know about any interactions between the predictors and the extent of any non-linearity (curvature) in the relationships between predictors and the outcome (dependent) variable. Interactions may arise when the influence of two variables on a third variable is not additive. When two independent variables interact, the relationship between either one of them and the third (dependent) variable depends on the value of the other independent variable.

Homoscedasticity means that the variance of the outcome (dependent) variable is approximately the same across the range of data. Homoscedasticity is very desirable because most linear models are predicated on the assumption of approximately constant variance. Heteroscedasticity arises when the variance of the dependent variable varies across the data and may invalidate our analysis unless we can find ways of controlling it. In creating the four models discussed in this report, we took the appropriate steps to ensure that our data meet the requirements of multiple regression.

First, it was established that multicollinearity between the 16 technical variables was not problematic. The mean correlation across all variables was 0.14 and the maximum correlation was just under 0.8 - that between turnovers against (ta) and steals for (sf). Appendix 2 gives the complete correlation table for all technical variables of this study. Strong correlations (> 0.5) are evident between af and sf (0.59), ta and sf (just below 0.8), ftf and rf (-0.56), a3 and aa (0.53), fga and a3 (0.67).

Second, we investigated interactions between selected pairs of variables, but found few significant interactions. In any case, fitting multiple regression models for all technical variables and their interactions was not possible because of the limited data set. Thus, interaction effects were ignored in all of the modelling procedures of this study.

Third, we investigated constancy of variance (homoscedasticity) for each model and we discuss this issue in section 3.10.

3.3 Win Ratio as the Outcome Variable

We define the win ratio (WR) as the ratio of wins to the total number of games played in a given season. This variable provides a convenient continuous outcome variable for our multiple regression models. Figure 1 gives a boxplot of the win ratio for the participating teams across all years within our data set.

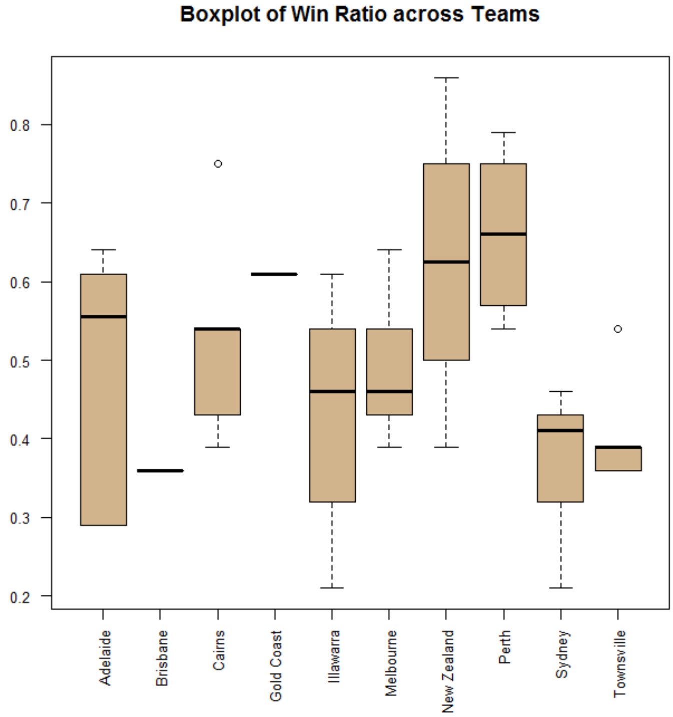

Figure 1: Box plot of win ratio for each team

Figure 1 shows that the New Zealand team (the Breakers) has performed strongly by comparison with the Australian teams. We have data for this team for the period from 2011 to 2017. Across those years, New Zealand had the second highest median win ratio and second highest mean win ratio (0.63) of all teams for which we sourced data. Perth had the highest median and mean win ratio (0.66), while Townsville had the lowest mean (0.41). Note that we have only one observation for each of Brisbane and the Gold Coast.

It is self-evident that any team that scores many points (pf) and concedes few points (pa) will tend to have a high win ratio, except for specific situations in which, for example, a team loses most of its games but plays a small number of games with high winning scores. Figure 2 gives a graph of win ratio against points differential, along with a fitted quadratic curve, created in R through an Ordinary Least Squares linear model in which both points differential and its square were included as predictors (independent variables).

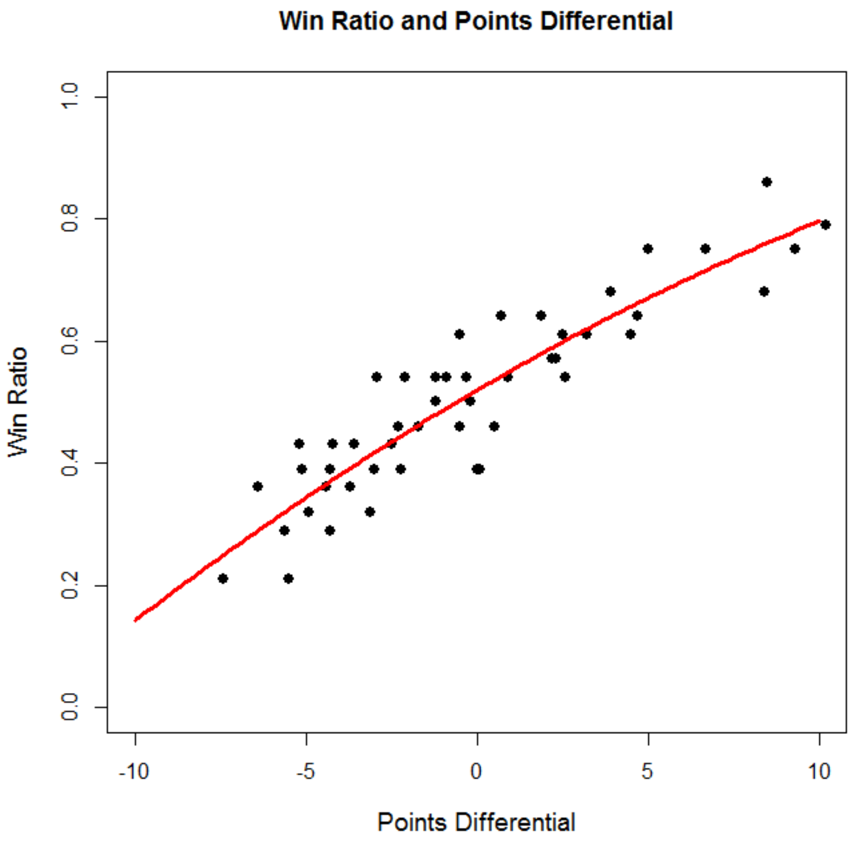

Figure 2: Win ratio plotted against points differential, with a fitted quadratic curve

Figure 2 confirms that a high win ratio is associated very strongly with high points differential. The correlation between these two variables is very high (0.90), so that the two variable share 82% of variance. A small degree of curvature is apparent in this plot and the curvature (the quadratic term) is highly significant (p-value = 5.0 e-4).

3.4 A Simple Model for Win Ratio

The process of identifying the variables that predict win ratio involved creating a multiple regression model that includes win ratio as the outcome variable (Model A). Every available technical variable, along with points for and points against, was placed initially into Model A.

Of course, fitting a model with 49 observations and 18 predictors (our initial model) gives a small number of observations for each predictor, and our model may be over-fitted. The justification for proceeding was that the present study, an undergraduate project, essentially constitutes a pilot study on which future research may be based.

At each iteration step, non-significant predictors were eliminated until a model was created in which all predictors exhibited p-values below 0.05. The variables with the greatest predictive power (points for and points against) remain in Model A, while all technical variables have been eliminated. This model yielded an Adjusted R-squared of 0.81 and a highly significant p-value (< 2.2e-16). The model explains over 81% of the variance in win ratio and the overall model p-value is highly significant. Here is the R Studio output for Model A:

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  0.610391  0.175452    3.479    0.00111 **
pf           0.030855  0.002431   12.693    < 2e-16 ***
pa          -0.032077  0.002427  -13.215    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06502 on 46 degrees of freedom
Multiple R-squared:  0.8189,     Adjusted R-squared:  0.8111
F-statistic:   104 on 2 and 46 DF,  p-value: < 2.2e-16
```

Thus, the Model A regression equation is: WR = 0.61 + 0.03*pf - 0.03*pa

We see that only points for (pf) and points against (pa) have remained in the model and none of the technical variables have been retained. This is because these two variables predict win ratio much more strongly than the technical variables, which become non-significant when we control for points for and points against. The similarity of the two coefficients is reassuring, because each variable is measured on the same scale (points per game).

Of course, in a single game, the difference in points scored for and points scored against determine the result, so that points differential (and therefore pf and pa) must of necessity be an extremely strong statistical predictor.

Mikołajec, Maszczyk & Zając (2013) discuss the importance of factors that influence success in basketball, including points scored for and against. They find that the main variables that influence results in the NBA (the major US basketball competition) relate more to offense than defence. We tested this finding by calculating standardized coefficients for Model A and all subsequent multiple models. Standardized coefficients measure the number of standard deviations by which the outcome variable changes for every standard deviation increase in the predictor(s). Standardized coefficients help to identify the predictors that influence the outcome variable most strongly, given that each predictor may be measured in completely different units (clearly not the case for pf and pa, which are both measured in points per game). The standardised regression coefficients for Model A are as follows:

pf	pa
1.01	-1.05

These standardised coefficients arising from Model A suggest that points for and points against are of roughly equal importance in predicting the win ratio. Naturally, pf is associated positively with win ratio, while pa is associated negatively with win ratio. Of course, in a complete data set of all games within a season, the total points for all teams and points against all teams must be equal.

One of the primary uses of any model is as a predictive tool. For example, inserting the mean values of pa (82.18) and pf (82.41) into Model A predicts a win ratio of 0.60. Increasing pf to 85.0, while holding pa at its mean value, predicts a win ratio of 0.69. Holding pf at its mean value, while reducing pa to 80.0, predicts a win ratio of 0.68.

3.5 A Model for Win Ratio that includes only Technical Variables

Our next model (Model B) also embodies win ratio as the outcome variable, but omits pf and pa, because we wish to identify the purely technical variables that predict success. Omitting points for and against obviates the issue of non-significance of the purely technical variables when controlling for pa and pf.

However, before embarking on our first multiple regression, we investigated relationships between variables using regression tree models, available in R through the tree package. Regression trees indicate the expected values of the dependent variable for particular ranges of values of the predictors and they help to identify relationships and interactions between variables before we undertake more complex modelling (see Crawley, pages 197 and 204 for a helpful explanation of tree models).

Figure 3 gives a tree plot of the technical variables as they predict win ratio.

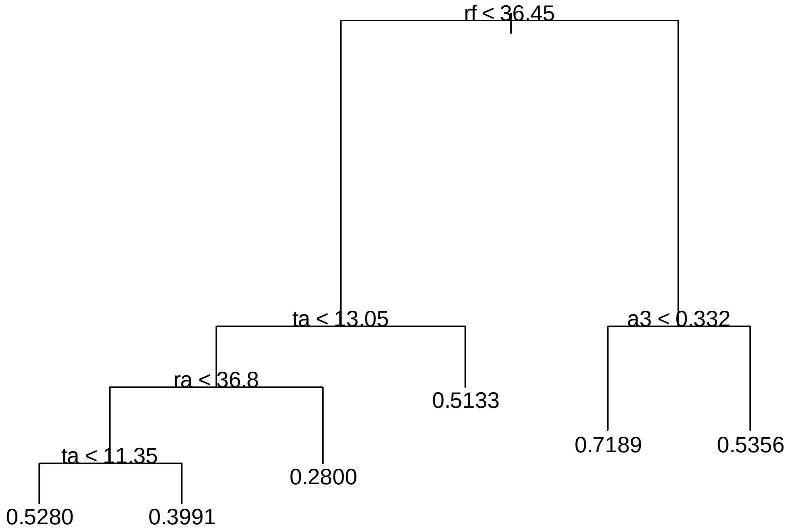

Figure 3: Tree plot of the technical variables of Model B

A tree plot indicates the relative strength of each predictor. In a tree plot, the longer the branches, the greater the deviance explained (a statistical term meaning that the more strongly the independent variable predicts the outcome variable). The tree model has identified differences in relationships between variables for values of rf less than and above 36.45, and other differences in relationships lower down the tree at the indicated values above each pair of branches.

The figures at the ends of the branches give the mean win ratio for the relevant values of the predictors (technical variables). We see that rebounds for (rf) is the strongest predictor. Turnovers against (ta) is important at low values of rf and three-pointers (a3) against are important at higher values of rf. We see that high win ratio is associated with lower mean a3 (i.e. the mean win ratio at the end of the left a3 branch is 0.7189, greater than the mean value of 0.5356 on the right hand branch).

Every technical variable was placed into the initial version of Model B, excluding pf, pa and points differential (pd is, of course, an exact linear combination of pf and pa). Eleven iteration steps were required to produce a model in which all predictors were significant. Model B has an Adjusted R-squared of 0.746 (i.e. it predicts nearly 75% of the variability in win ratio) and a p-value of 3.0e-12 (i.e. the model is highly significant). Here is the R Studio output for Model B:

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  0.805487   0.574129    1.403   0.167974
rf           0.018064   0.005443    3.319   0.001875 **
ra          -0.016959   0.004281   -3.962   0.000283 ***
tf          -0.043077   0.008183   -5.264   4.50e-06 ***
ta           0.047167   0.007367    6.403   1.05e-07 ***
f3           2.257193   0.477643    4.726   2.58e-05 ***
fga         -2.657666   0.727751   -3.652   0.000716 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07539 on 42 degrees of freedom
Multiple R-squared:  0.7777,    Adjusted R-squared:  0.746
F-statistic: 24.49 on 6 and 42 DF,  p-value: 3.023e-12
```

Thus, the Model B regression equation is:

$$WR = 0.81 + 0.02*rf - 0.02*ra - 0.04*tf + 0.05*ta + 2.26*f3 - 2.66*fga$$

Model B gives the technical basketball-specific variables that combine to predict win ratio. These variables are: rf (rebounds for), ra (rebounds against), tf (turnovers for), ta (turnovers against), f3 (three point shots for), and fga (field goals against). The variable three-pointers against (a3) was evident in the tree model (where it was important only for high values of rf) but has been eliminated from our Model B.

The standardised regression coefficients for Model B are as follows:

rf	ra	tf	ta	f3	fga
0.33	-0.31	-0.43	0.51	0.37	-0.35

Turnovers against and turnovers for appear to be the strongest predictors of win ratio, but the other variables of the model are also strong predictors.

As with Model A, we can experiment with this model by setting desired values for the predictors and evaluating the impact on win ratio. Inserting the mean values for each of the technical variables into Model B predicts a win ratio of 0.60. Increasing rf from its mean value of 35.0 to 38.0 and reducing ra from its mean value of 34.9 to 34.0, while holding all other variables at their mean values, predicts a win ratio of 0.62. Decreasing tf from its mean value of 12.87 to 12.00 and increasing ta from its mean value of 12.81 to 13.00, while holding all other variables at their mean values, predicts a win ratio of 0.59.

3.6 Points Differential as the Success (Outcome) Variable

Points differential (pd) was the second outcome variable considered in this study. Points differential is simply the mean difference between points scored for and points scored against a given team across a given season. Figure 4 gives a boxplot of the points differential for the participating teams across the years within our data set.

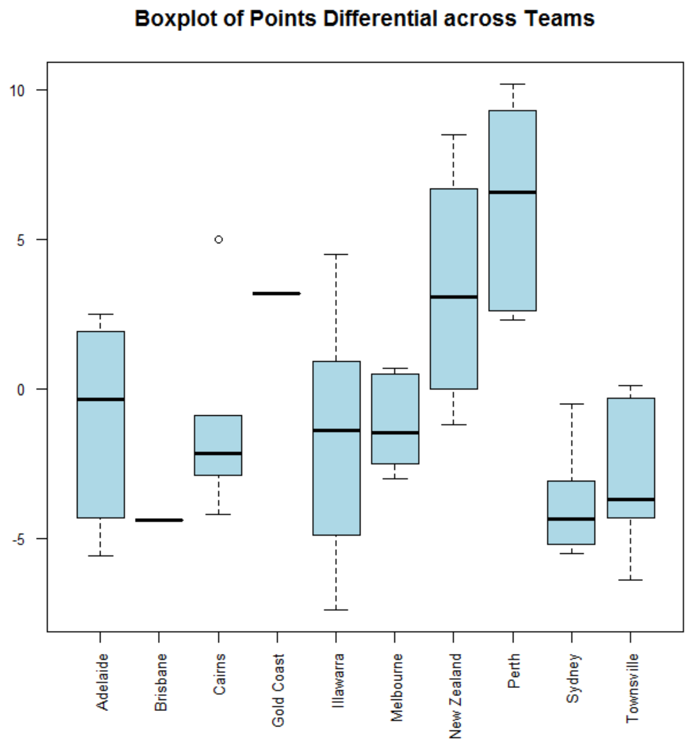

Figure 4: Box plot of points differential across teams

Figure 4 shows that the New Zealand team has performed strongly in points differential by comparison to the Australian teams. Again, we have only one observation for Brisbane and the Gold Coast. Of course, Figure 4 looks very similar to Figure 1 because the correlation between win ratio and points differential is very high (0.90).

3.7 A Model with Points Differential as the Outcome Variable

A third multiple regression model (Model C) was created in order to identify the technical variables with collective predictive power for points differential as the outcome variable. Every technical variable was included within the initial version of Model C (i.e. excluding pf and pa). Figure 5 gives a tree plot of the variables of Model C.

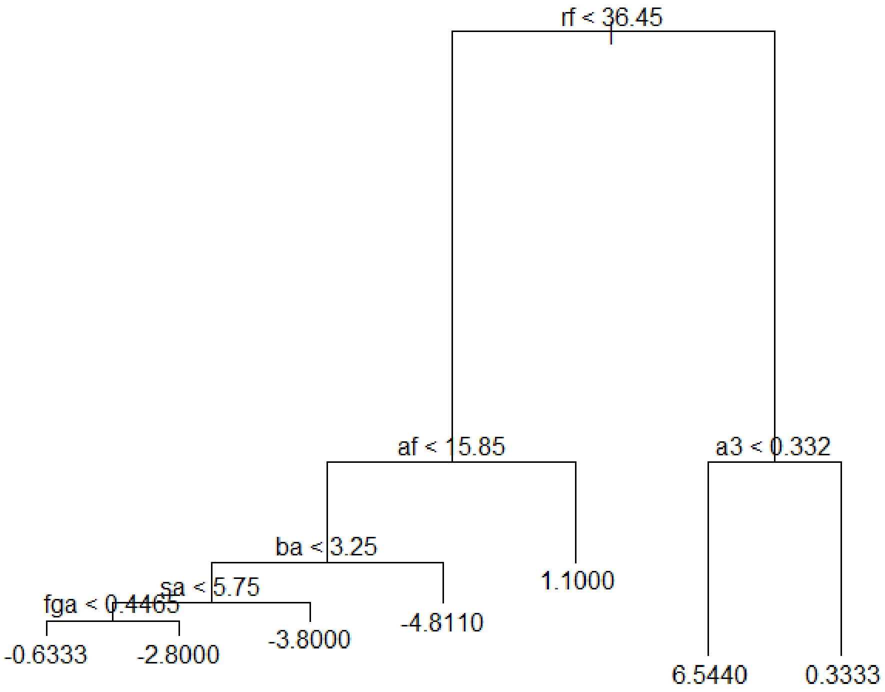

Figure 5: Tree plot of the technical variables of Model C

The tree plot of figure 5 is similar to that for win ratio as the outcome variable. We see that rebounds for (rf) is the strongest predictor, as it was for win ratio as the outcome variable. Assists for (af) is important at low values of rf (whereas in Figure 3 turnovers against was the important variable here) and three-pointers (a3) against are important at higher values of rf. Again, we see that high points differential is associated with lower mean a3 (i.e. the mean win ratio at the end of the left a3 branch is 6.5440, greater than the mean value of 0.3333 on the right hand branch). Of course, low values of fieldgoals against (fga) is associated with higher mean points differential.

Eight elimination steps were required to create Model C, and an Adjusted R-Squared of over 0.9 and a highly significant model p-value were obtained. Model C showed great promise and was considered worthy of further development.

Figure 6 presents diagnostic plots of the residuals of Model C, and indicates the strong influence of three outliers, but of one outlier in particular (point 14).

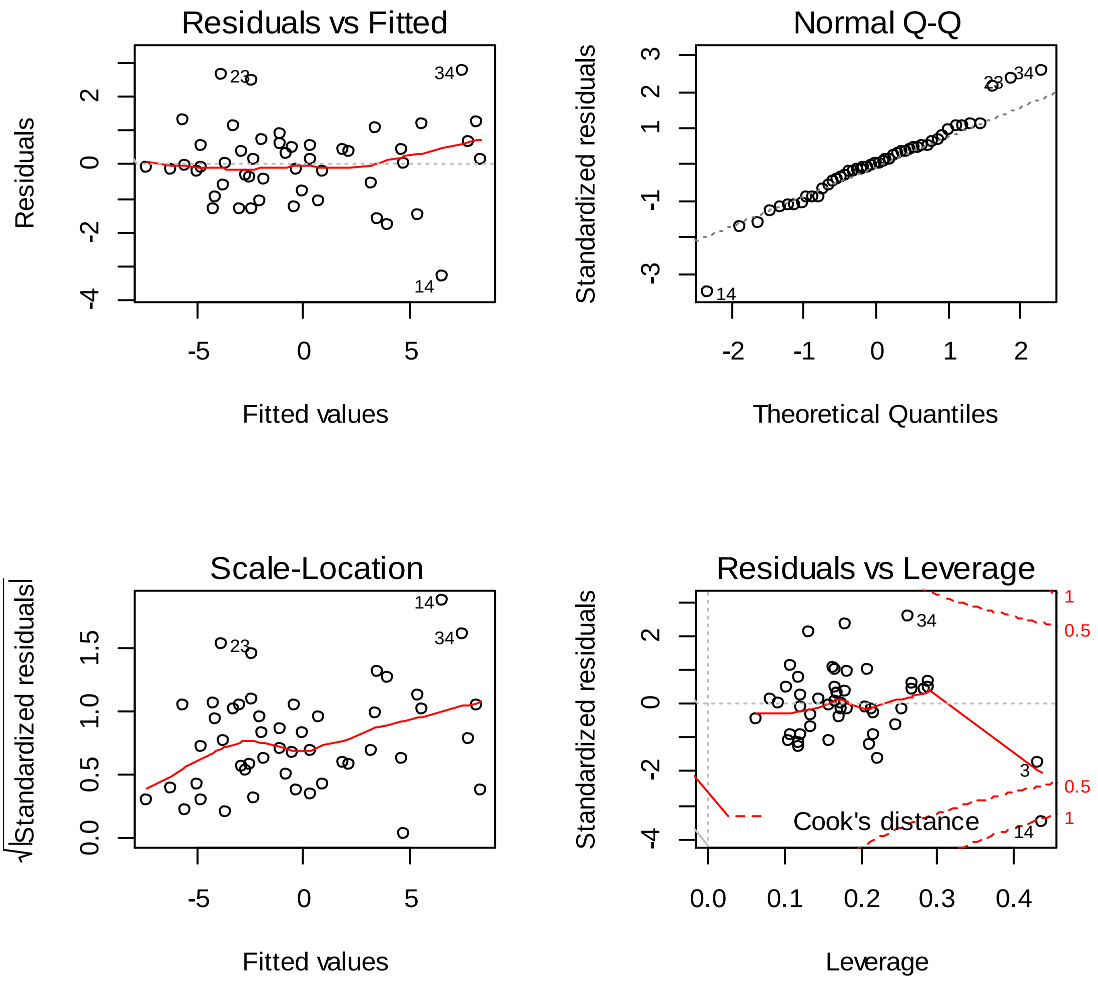

Figure 6: Diagnostic Residual plots for Model C

Within Figure 6 the first plot (the plot of residuals versus fitted values) is roughly random. Apart from three outliers, we have approximate constancy of variance and the residuals display little inherent structure. Any visible structure in this plot suggests that other models (possibly non-linear models) may be more appropriate than Model C.

The normal QQ plot (the second plot) suggests a roughly normal distribution, apart from points 14, 23 and 34. The third plot (the scale-location plot of the square root of the standardized residuals) provides more or less the same information as in the first plot, but is scaled differently.

Finally, the last plot (residuals versus leverage) gives the Cook's Distance of each point in the set of predictors, a measure of the influence that any observation exerts on the model; that is, the influence of each observation on the fitted outcome variable. A commonly accepted criterion (adopted for our study) is that an observation with a Cook's distance of 1.0 or more can be considered an outlier. According to this criterion, point 14 is highly influential (more influential than points 23 and 34, for example). In fact, this point represents the only datum we have for one particular team (the Gold Coast). We considered this point to be anomalous and we removed it so as to create an improved version of Model C. After removal of this point, the plot of residuals versus fitted values is approximately random and the normal QQ plot suggests a roughly normal distribution. Elimination of point 14 resulted in a clear improvement of the model.

Here is the output from R studio for Model C, where we have omitted point 14 and implemented stepwise elimination:

```
Coefficients:
            Estimate Std.  Error     t value   Pr(>|t|)
(Intercept) -12.42277   8.31047  -1.495    0.143006
rf            0.61783   0.07953   7.769    1.93e-09 ***
ra           -0.54666   0.07617  -7.177    1.23e-08 ***
tf           -1.28646   0.11488 -11.198    9.44e-14 ***
ta            1.47231   0.10901  13.507    2.74e-16 ***
f3           59.36158   8.01799   7.404    6.02e-09 ***
a3          -67.92495  12.07846  -5.624    1.73e-06 ***
fgf          53.77610  13.86480   3.879    0.000393 ***
fga         -30.00672  13.24110  -2.266    0.029067 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.051 on 39 degrees of freedom
Multiple R-squared:  0.9508,      Adjusted R-squared:  0.9407
F-statistic: 94.13 on 8 and 39 DF,  p-value: < 2.2e-16
```

Model C has an adjusted R-squared of over 0.94 (i.e. it predicts approximately 94% of the variability in points differential) and has a p-value of < 2.2e-16. (i.e. the model is highly significant).

The Model C regression equation is:

$$PD = -12.42 + 0.62*rf - 0.55*ra - 1.29*tf + 1.47*ta + 59.36*f3 - 67.92*a3 + 53.78* fgf - 30.00*fga$$

Clearly, the following variables have a strong collective predictive power for points differential: rf (rebounds for), ra (rebounds against), tf (turnovers for), ta (turnovers against), f3 (three point shots for), a3 (three point shots against), fgf (field goals for) and fga (field goals against). Two variables that were retained in Model C were eliminated in Model B (a3 and fgf).

The following table gives the standardised regression coefficients for Model C.

rf	ra	tf	ta	f3	a3	fgf	fga
0.40	-0.27	-0.44	0.55	0.34	-0.31	0.19	-0.14

As for Model B, these standardised regression coefficients suggest that the two strongest predictors of points differential are ta and tf, so that turnovers have a considerable influence on points differential. Another strong predictor is rf, which relates closely to turnovers because catching a rebound means either retaining or gaining possession of the ball, and turnovers occur when a team loses possession of the ball.

Trninic, Disdar & Luksic (2002) attempted to identify factors that differentiated between winning and losing teams in the final tournaments of the European club championships from 1992 to 2000. They found that the highest discriminative power for separating winning and losing teams was demonstrated by defensive rebounds (ra), and by field goals (fgf and fga). All three variables appear in Model C.

The two variables f3 and a3 are also strong predictors, suggesting that three point shooting for and against a team influences points differential considerably. The standardised regression coefficients suggest that the technical variables fgf and fga of Model C determine points differential somewhat less strongly.

Again, we can experiment with particular values of the predictors and evaluate their impact on points differential. Inserting the mean values for each of the technical variables into Model C predicts a points differential of 0.42. Increasing rf from its mean value of 35.0 to 38.0, reducing ra from its mean value of 34.9 to 34.0, decreasing tf from its mean value of 12.87 to 12.00 and increasing ta from its mean value of 12.81 to 13.00, while holding all other variables at their mean values, predicts a points differential of 3.47.

3.8 Turnovers and Rebounds

Model B and Model C share the following variables: rf (rebounds for), ra (rebounds against), tf (turnovers for), ta (turnovers against), f3 (three point shots for) and a3 (three point shots against).

Ruano et al (2006) analyzed game-related statistics that discriminate between winning and losing teams in female basketball. They identified defensive rebounds (ra in our study) and assists as key variables that discriminated between winning and losing teams. Gómez, Lorenzo & Sampaio (2008) also noted the importance of defensive rebounds. They found that winning balanced games (where the two teams score similar numbers of points) is often the result of greater success in defensive rebounds and that defensive rebounds are very important in balanced games. They also found that success in defensive rebounds contributed strongly to winning unbalanced games.

Lorenzo, Gómez, Ortega, Ibáñez & Sampaio (2010) analysed statistics that discriminate between winning and losing in male under-16 basketball games. They also identified turnovers and assists as the critical variables that discriminated between successful and unsuccessful teams. Lorenzo et al. emphasise the importance of reducing the impact of turnovers, especially in close games, because in close games winning teams tend to have better turnover statistics. A controlled style of play reduces risk because it reduces the frequency of turnovers. Our models include turnovers for and against (tf and ta) and, indeed, these variables appear to be the strongest predictors.

3.9 Fieldgoals and Three Point Shots

Model B shows that the technical variable f3 (three point shots for) has strong predictive power for win ratio and our Model C shows that both a3 and f3 are important for points differential in the NBA. Lorenzo et al. found that in both balanced games (score differences between 10 and 29 points) and unbalanced games (score differences above 30 points), two-point fieldgoals discriminated strongly between successful teams. Ruano et al also found that successful free-throws and three point field-goals were important predictors in female basketball.

Watson (2016) underscores the importance of three point shots in influencing the outcome of the game. He notes that the average number of three-point attempts per game is increasing and will continue to increase in the future. He suggests that focusing on defence is the best way of preventing three pointers. Watson discusses the Spurs - the best defensive team in the NBA (the major US basketball competition), as follows:

The Spurs are really locking in on the elite shooters of the NBA. The top 10 shooters in the league are on average shooting 40% less three-point attempts against the Spurs than they usually do. This is a huge advantage.

Watson's research is specific to the NBA, but the variables a3 and f3 emerge from our study as important to success in the NBA. Watson finds that defending the three point shot is very advantageous and that three point shots scored against a team constitute a major disadvantage. This finding is confirmed in Model C in which a3 has a strong negative association with points differential.

3.10 Checking the Assumptions of Multiple Regression Models

Multiple regression models are developed on the assumption that the predictors and outcome variables are related linearly (or approximately so). Non-linearity invalidates multiple regression. Thus, bivariate analysis was undertaken at the very beginning of this study to assess the extent of linearity between variables. Because the data set involves 16 technical variables and three outcome variables, a complete set of all bivariate plots across all pairs of variables is beyond the scope of this report. However, Figure 7 gives bivariate plots with fitted curves (using the default Loess procedure within R and R Studio), but including only four of the key variables that remained in Model C following stepwise elimination and removal of point 14.

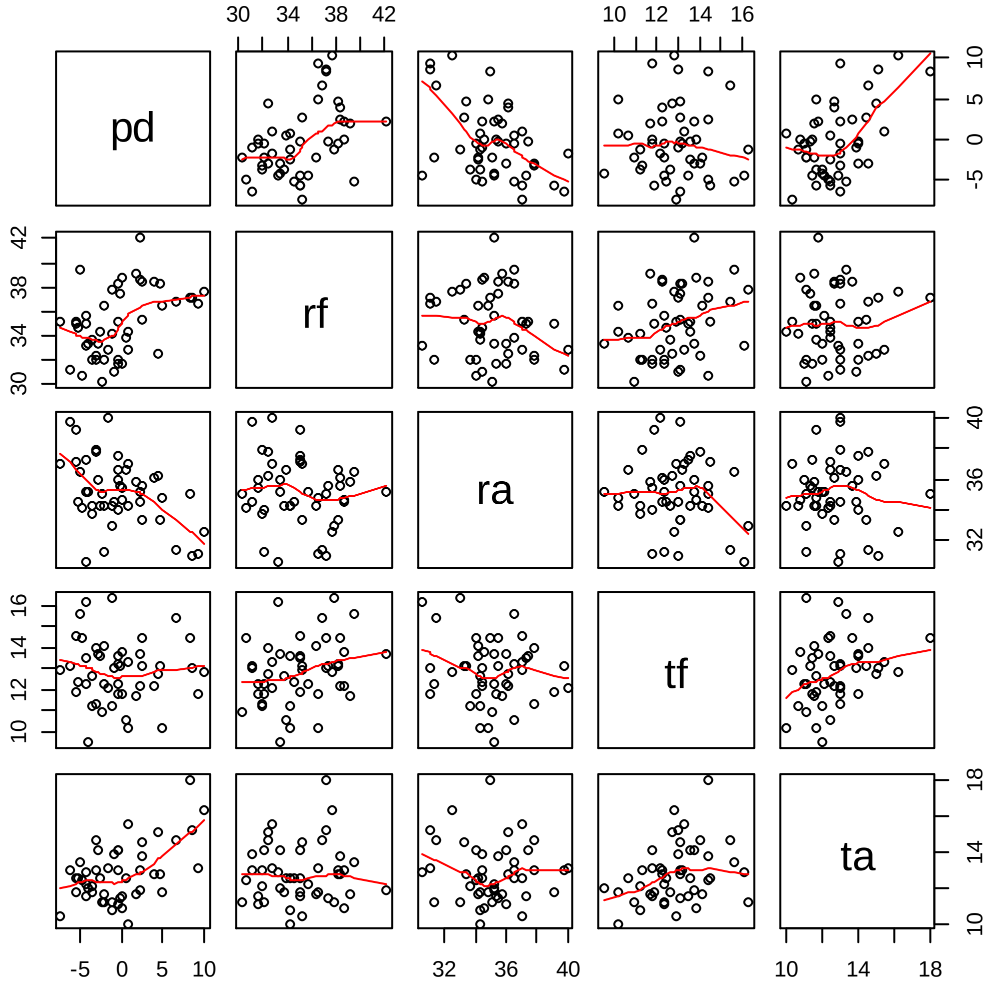

Figure 7: Bivariate plots of the key variables of Model C

The bivariate plots show considerable scatter but, for the purposes of an undergraduate research study, the variables relate linearly enough to justify the development of multiple regression models.

4. Final Position in the League as the Outcome Variable

4.1 Final Position and Points Differential

Figure 8 gives a graph of position against points differential, along with a fitted curve, developed through a generalised linear model using a Poisson error structure (see section 4.3).

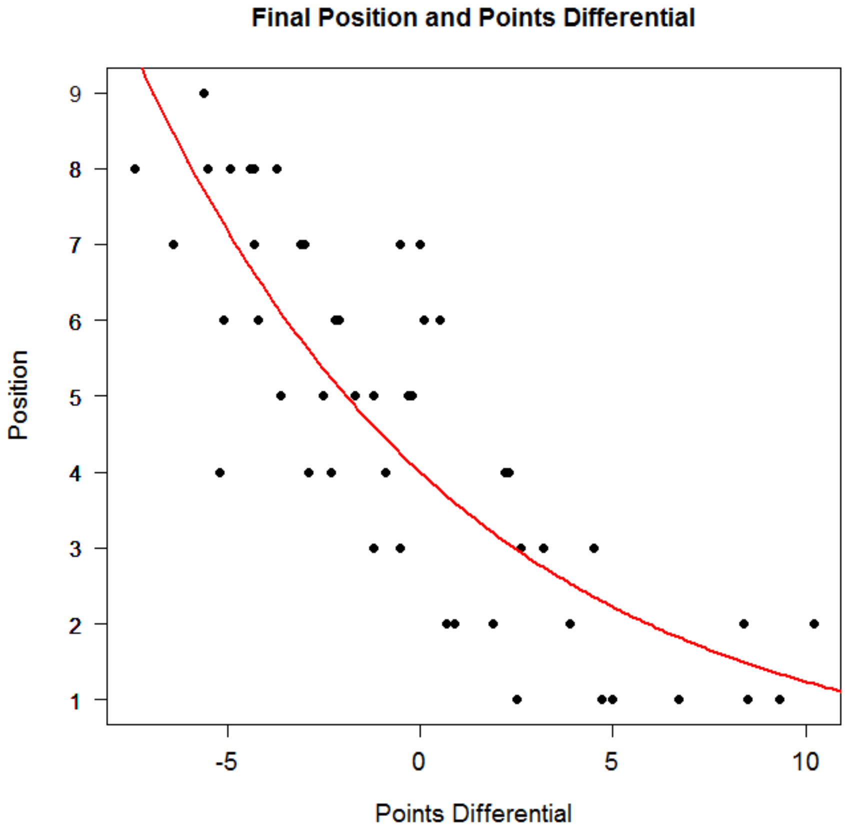

Figure 8: Position in the league against points differential with a fitted curve

Figure 8 confirms visually that a strong finishing position in the league is associated with high points differential. A degree of curvature is apparent in the fitted curve. For our data set, at finishing positions at or close to the top position (i.e. first place), somewhat greater points differentials are required in order to gain an additional placing than at lower finishing positions. It would be interesting to explore this finding with more comprehensive data sets than that of this study.

4.2 Correlations between Final Position and the Technical Variables

We found moderate to strong correlations (> 0.3) between the teams' finishing positions and the following variables (Spearman's correlations are given in parentheses):

1. Points difference (-0.81)
2. Three pointers against (0.50) but not three pointers for (-0.21)
3. Fieldgoals against (0.50) but not fieldgoals for (-0.16)
4. Assists for (-0.43)
5. Rebounds for (-0.42) and rebounds against (0.32)
6. Rebounds for (-0.42) and rebounds against (0.32)
7. Steals for (-0.40) but not steals against (0.03)
8. Points for (-0.38) and against (0.34)
9. Turnovers against (-0.34) but not turnovers against (0.01)

On the basis of these moderately strong correlations, it was decided to develop generalised linear models.

4.3 Generalised Linear Models for modelling Position

We developed several generalised linear models (GLM) using final position in the league table as the outcome variable, using both Poisson error structures and Quasi-Poisson error structures. We also developed Negative Binomial models. In these models we consider position as a count variable which must always be greater than or equal to one. In count data the errors may be distributed non-normally and the variance may increase with the mean values, so that multiple regression models based on Ordinary Least Squares are no longer valid.

Our Poisson model uses the natural logarithm as the link function (the default link function for the Poisson error distribution). The Poisson error distribution assumes that the variance is approximately equal to the mean, so that specifying a Poisson error distribution accounts well for integer data meeting that criterion. For the variable position, the mean (5.0) and variance (7.5) are similar enough to justify a model with a Poisson error structure. Specifying a logarithm as the link function forces all of the predicted values to be positive. Clearly, this must be the case if we are modelling position in the league table, where the smallest possible value is unity.

Position was taken as the outcome variable, and all technical variables were included in the initial versions of our models. Eleven iteration steps were required to create Model D, using Poisson errors. No advantage was found either by adopting a Quasi-Poisson error structure or a Negative Binomial model. Here is the output from R Studio for Model D, in which we use the Poisson error structure:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.42437    3.62080  -0.393  0.69403
rf          -0.07146    0.03275  -2.182  0.02911 *
ra           0.07067    0.03273   2.159  0.03083 *
tf           0.17040    0.05186   3.286  0.00102 **
ta          -0.16522    0.05269  -3.136  0.00171 **
f3          -8.36050    3.03927  -2.751  0.00594 **
fga         12.85640    4.66098   2.758  0.00581 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 64.366  on 47  degrees of freedom
Residual deviance: 20.695  on 41  degrees of freedom
AIC: 190.29

Number of Fisher Scoring iterations: 4
```

In generalized linear models, deviance is a measure of goodness of fit. R (and R Studio) report two forms of deviance – the null deviance and the residual deviance. The null deviance shows how well the outcome variable is predicted by a model that includes only the intercept (grand mean). We use the residual deviance to test the goodness of fit. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model, where the predicted values are identical to the observed. If we have two or more similar models (where the predictors of one model are also predictors of the other models), the rule of thumb is to select the model with the lowest residual deviance.

We note that Model D is under-dispersed (i.e. the residual deviance is less than the degrees of freedom), indicating that the predicted variability in the outcome variable is greater than the observed variability. Thus, our model provides a good, but not perfect, description of the variability in Position. To obtain the final version of Model D we take the exponential of the model provided by the initial Model D coefficients because those estimates are scaled in natural logs. The final Model D regression equation is as follows:

$$\log_e(\text{Pos}) = -1.42 - 0.07*rf + 0.07ra + 0.17*tf - 0.17*ta - 8.36*f3 + 12.86*fga$$

The final set of technical variables is identical to that of Model B. Again, we can experiment with our model. Taking the mean values of each of the technical variables in Model D predicts a finishing position of 3.99 (i.e. marginally better than fourth place).

Holding all other variables at their mean values, but increasing field goals against from its mean value of 0.44 to 0.50, results in a predicted finishing position of 8.8 (i.e. barely better than ninth place). Holding all other variables at their mean values, but decreasing field goals against from its mean value of 0.44 to 0.35, results in a predicted finishing position of 1.27 (i.e. nearly first place).

Holding all other variables at their mean values, but increasing three pointers for from its mean value of 0.44 to 0.40, results in a predicted finishing position of 2.45 (i.e. better than third place). Holding all other variables at their mean values, but decreasing three pointers for from its mean value of 0.44 to 0.30, results in a predicted finishing position of 5.66 (i.e. worse than fifth place).

Holding all other variables at their mean values, but increasing rebounds for (rf) from its mean value of 35.03 to 40.0, increasing tf from its mean of 12.85 to 14.0 and decreasing ta from its mean of 12.78 to 12.0, results in a predicted finishing position of 2.73 (i.e. better than third place).

4.4 Extending the Analysis for Future Work

A principal components analysis (PCA) was performed on the technical variables in order to reduce the number of predictors to s smaller subset of linear combinations of the original set of predictors. Figure 9 gives a scree plot of the technical variables.

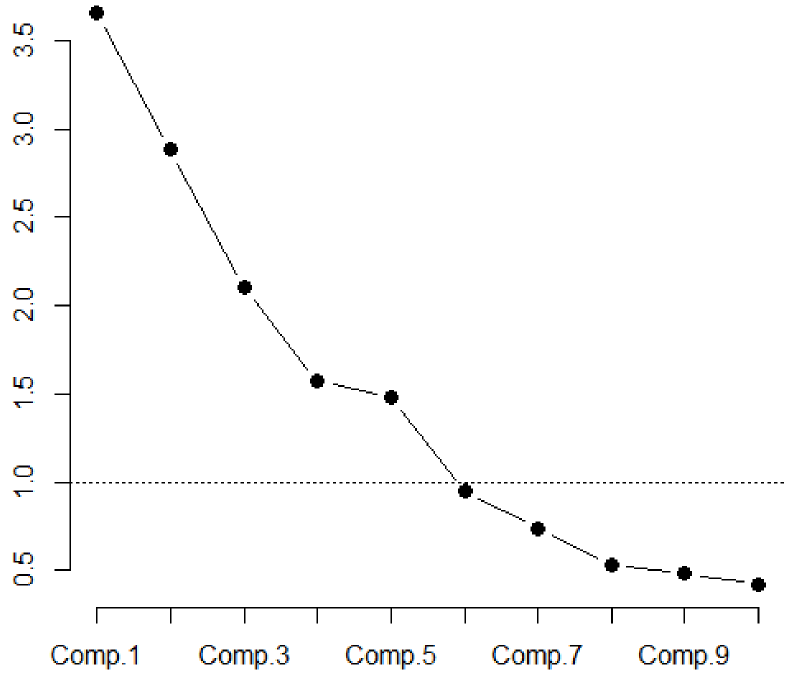

Figure 9: PCA Analysis of the Technical Variables

Five principal components with eigenvalues above 1.0 are present but few of the variables load strongly on any one component. However, the variables rf (0.389), ftf (-0.310), a3 (-0.368) and fga (-0.413) have the strongest loadings on the first component. PCA analysis of larger datasets similar to our may provide further insight into relationships between the technical variables.

5. Summary and Recommendations

The findings of this study, obtained mainly through multiple regression and generalised linear models, are of potential importance for coaching and management of basketball. We found that particular technical, game-related variables predict strongly the outcome variables win ratio, points differential and position in the league table, while other technical, game-related variables are not strong predictors. Our main findings are as follows:

1. Apart from points for and points against, it is clear that the strongest predictors of a team's win ratio and final position in the league table are: rf (rebounds for), ra (rebounds against), tf (turnovers for), ta (turnovers against), f3 (three point shots for) and fga (field goals against)

2. The variables that predict a team's points differential are those that predict win ratio and position, but also including a3 and fgf.

3. Turnovers for and against appear to be very strong predictors of success.

Further research will be undertaken on the basis of more comprehensive statistics from the NBL and other competitions in other countries, such as the US competition - the NBA. If a much bigger sample of data can be compiled, then more sophisticated models can be developed that include interaction effects and that take account of possible non-linear relationships between technical variables and selected outcome variables.

APPENDIX 1

Definitions of the Technical Variables

Points for (pf): the points scored for a team

Minimum: 72.30; Maximum: 93.00; Mean: 82.15; Median: 82.15

Points against (pa): the points scored against a team

Minimum: 68.10; Maximum: 91.80; Mean: 82.45; Median: 83.40

Blocks for (bf): when a player stops the opposition's shot at the basketball hoop mid-air with their hand

Minimum: 1.90; Maximum: 5.40; Mean: 3.22; Median: 3.05

Blocks against (ba): when an opposition player stops the shot mid-air with their hand while defending their basketball hoop

Minimum: 2.20; Maximum: 4.40; Mean: 3.21; Median: 3.20

Steals for (sf): taking the ball off the opponent while they are dribbling the ball

Minimum: 3.30; Maximum: 8.40; Mean: 5.50; Median: 5.20

Steals against (sa): when the ball is taken by an opposition player while dribbling the ball

Minimum: 4.30; Maximum: 7.50; Mean: 5.58; Median: 5.50

Assists for (af): passing the ball to a team member 1-3 seconds before they score

Minimum: 12.60; Maximum: 19.30; Mean: 15.38; Median: 15.10

Assists against (aa): when the opposition passes the ball to a team member 1-3 seconds before they score

Minimum: 12.40; Maximum: 18.80; Mean: 15.26; Median: 15.20

Rebounds for (rf): catching the ball before it hits the ground after a missed shot attempt from either team
Minimum: 30.30; Maximum: 42.10; Mean: 35.03; Median: 35.00

Rebounds against (ra): when the opposition catches the ball before it hits the ground after a missed shot attempt from either team
Minimum: 30.60; Maximum: 40.00; Mean: 35.13; Median: 35.15

Turnovers for (tf): giving away possession of the ball
Minimum: 9.50; Maximum: 16.30; Mean: 12.85; Median: 12.95

Turnovers against (ta): when the opposition gives away possession of the ball
Minimum: 10.00; Maximum: 18.00; Mean: 12.79; Median: 12.50

Free throws for (ftf): successfully converting a free shot at the basket from the free-throw line after a player has been fouled
Minimum: 0.62; Maximum: 0.83; Mean: 0.73; Median: 0.74

Free throws against (fta): when an opposition player successfully converts a free shot at the basket from the free-throw line after they have been fouled
Minimum: 0.67; Maximum: 0.78; Mean: 0.73; Median: 0.74

Three point shots for (f3): successfully shooting the basketball from beyond the three point line
Minimum: 0.29; Maximum: 0.39; Mean: 0.34; Median: 0.34

Three point shots against (a3): when the opposition shoots the basketball successfully from beyond the three point line
Minimum: 0.30; Maximum: 0.38; Mean: 0.34; Median: 0.34

Field goals for (fgf): successfully shooting the basketball within the three point line

Minimum: 0.40;   Maximum: 0.47; Mean: 0.44; Median: 0.44

Field goals against (fga): when the opposition shoots the basketball successfully within the three point line
Minimum: 0.40;   Maximum: 0.48; Mean: 0.44; Median: 0.44

APPENDIX 2

Correlation Table for all Technical Variables

	bf	ba	sf	sa	af	aa	rf	ra	tf	ta	ftf	fta	f3	a3	fgf	fga
bf	1	0.38	0.29	-0.01	0.25	0.07	0.4	0.17	0.08	0.11	-0.37	0.14	-0.17	-0.08	0.12	-0.23
ba	0.38	1	-0.08	0	0.09	0.29	0.29	0.13	0.11	-0.1	-0.25	0	-0.22	0.21	0.15	0.04
sf	0.29	-0.08	1	0.16	0.59	0.12	0.16	-0.13	0.19	0.8	-0.01	0.12	0.05	-0.34	0.02	-0.2
sa	-0.01	0	0.16	1	0.38	0.2	0.19	-0.01	0.71	0.18	-0.11	0.03	0.03	0	0.09	-0.09
af	0.25	0.09	0.59	0.38	1	0.26	0.39	-0.21	0.3	0.49	-0.06	0.16	0.18	-0.06	0.37	-0.1
aa	0.07	0.29	0.12	0.2	0.26	1	-0.11	0.3	0.04	0.11	0.2	0.2	0.01	0.53	0.39	0.41
rf	0.4	0.29	0.16	0.19	0.39	-0.11	1	-0.19	0.3	0.06	-0.56	-0.32	-0.23	-0.33	-0.04	-0.62
ra	0.17	0.13	-0.13	-0.01	-0.21	0.3	-0.19	1	-0.17	-0.15	-0.03	-0.15	-0.15	0.1	-0.15	0.01
tf	0.08	0.11	0.19	0.71	0.3	0.04	0.3	-0.17	1	0.32	-0.2	-0.17	-0.05	-0.16	0.08	-0.17
ta	0.11	-0.1	0.8	0.18	0.49	0.11	0.06	-0.15	0.32	1	0.12	-0.08	-0.14	-0.37	-0.12	-0.16
ftf	-0.37	-0.25	-0.01	-0.11	-0.06	0.2	-0.56	-0.03	-0.2	0.12	1	0.24	0.21	0.31	0.07	0.54
fta	0.14	0	0.12	0.03	0.16	0.2	-0.32	-0.15	-0.17	-0.08	0.24	1	0.48	0.21	0.36	0.29
f3	-0.17	-0.22	0.05	0.03	0.18	0.01	-0.23	-0.15	-0.05	-0.14	0.21	0.48	1	0.16	0.54	0.23
a3	-0.08	0.21	-0.34	0	-0.06	0.53	-0.33	0.1	-0.16	-0.37	0.31	0.21	0.16	1	0.44	0.67
fgf	0.12	0.15	0.02	0.09	0.37	0.39	-0.04	-0.15	0.08	-0.12	0.07	0.36	0.54	0.44	1	0.35
fga	-0.23	0.04	-0.2	-0.09	-0.1	0.41	-0.62	0.01	-0.17	-0.16	0.54	0.29	0.23	0.67	0.35	1

References

2016-2017 Australian NBL Standings. (2017). Retrieved from http://basketball.realgm.com/international/league/5/Australian-NBL/standings/427/2017

Crawley, M,. J. *Statistics: an Introduction using R*. Wiley.

Gómez, M., Lorenzo, A., & Sampio, J. (2008). *Game-Related Statistics that Discriminated Winning and Losing Teams from the Spanish Men's Professional Basketball Teams*. Coll Antropol. 32(2), pages 451-456.

The National Basketball League, (2017). Retrieved 1 March 2017, from http://www.nbl.com.au/history/

Lorenzo, A., Ángel Gómez, M., Ortega, E., José Ibáñez, S., & Sampaio, J. (2010). *Game related statistics which discriminate between winning and losing under-16 male basketball games.* Journal of Sports Science and Medicine. Vol. 9, pages 664 - 668
http://www.jssm.org/vol9/n4/17/v9n4-17text.php

Mikołajec, K., Maszczyk, A., & Zając, T. (2013). *Game Indicators Determining Sports Performance in the NBA.* Journal of Human Kinetics, Vol. *37*(1).
http://dx.doi.org/10.2478/hukin-2013-0035

Ruano, M., Calvo, A., Sampaio, J., & Ibáñez, S. (2006). *Differences in game-related statistics between winning and losing teams in women's basketball.* Journal of Human Movement Studies, Vol. 51(5): pages 357-369
https://www.researchgate.net/publication/255172750_Differences_in_game-related_statistics_between_winning_and_losing_teams_in_women%27s_basketball

Trninic, S., Disdar, D., & Luksic, E. (2002). *Differences between Winning and Defeated Top Quality Basketball Teams in Final Tournaments of European Club Championship*. Coll Antropol. Vol. 26(2), pages 521-531.

Watson, M. (2016). *How elite NBA teams defend the 3 Point Shot*. NBA Basketball Analytics. Retrieved 16 March 2017, from http://www.zigzaganalytics.com/home/-how-elite-teams-defend-the-3-point-shot-in-the-modern-nba