The Success of College Basketball Teams can be predicted using several advanced team statistics*

A Linear Model Approach

Thomas D'Onofrio

25 April 2022

Abstract

The March Madness college basketball tournament is one of the most exciting sports events in the world due to the intense games, unpredictable results, and the creation of brackets by millions. In this paper, data displaying each college basketball schools' advanced team statistics was used to create a linear model with team winning percentage as the response variable, where the explanatory variables are the team statistics that play the largest factor in leading to team success. It was found that [insert statistics here once model has been created] are the most important team statistics that affect the amount of wins a team gets the most, meaning fans can use these statistics to fill out their brackets more successfully. Unfortunately, the data is slightly biased due to inconsistent scorekeeping across teams and arenas, and the racial biases of referees when calling fouls, causing some implications to the findings.

1 Introduction

March is a very exciting time for sports fans in the US and many other countries, as it is when the NCAA Men's Division 1 Basketball Tournament, more commonly known as March Madness, is played. 68 college basketball teams go head to head, playing single elimination games until there is only one team remaining, making for some of the most intense and unpredictable sports games in the world. As a result of this, March Madness draws in a lot of viewers. 16.9 million people tuned in to the 2021 March Madness finals, making it the 2nd most viewed North American sports event only behind the Superbowl (Baker 2021). A big reason for the massive following is that many people create brackets for the tournament, trying to predict the results. With 9.2 quintillion different bracket possibilities, all the insane upsets that nobody sees coming, and the overall chaotic nature of the tournament, creating a bracket and seeing how well you do is very appealing to both sports fans and regular people, causing all the hype around March Madness (Geiling 2014). People put a lot of work into their brackets, comparing teams to one another and trying to figure out who has the best odds of winning. With this in mind, is there a way fans can more successfully predict which teams will have more success than others in the tournament?

In this paper, we attempt to answer this intriguing question by analyzing a variety datasets from the college basketball section of Sports Reference (Sports Reference 2022), using R (R Core Team 2020). Datasets containing college basketball statistics from 2017 to 2020 were used, where each dataset contains many different advanced team statistics for all NCAA Division 1 schools with a Men's basketball team in a specific season. The data was first graphically analyzed to determine which team statistics have a strong or noticeable relationship with how successful a team is in their games. These statistics were then used to create a linear model that attempts to predict a teams Simple Rating System, the response variable that is being used to measure team success, where only the best and most necessary team statistics were included. As a result of this, it was determined that the following statistics [Insert statistics here once model is completed], play the most important factor in leading to team success. Thus, when filling out their brackets, fans can compare

*Code and data are available at: https://github.com/TDonofrio62/Paper-On-NCAAB-Stats.

these statistics across teams to determine which teams have a good chance of going far or even winning the tournament, as opposed to just looking at wins or their seeding like many people do.

Unfortunately, there are some implications with these findings. There appears to be several aspects of bias in the datasets being used. For one, there is a lot of human error in the tracking of statistics by scorekeepers, causing certain ambiguously defined statistics in the data to be incorrectly tracked and thus skewed (Williams 2017). This means that data is tracked differently for each team in the league in their respective arenas, made evident by Madison Square Garden's scorekeepers overstating the quality of shots in the NHL (McCurdy 2020). Thus, the data for some teams may be biased, making them look either better or worse relative to the other teams in the league. There is also unfortunate racial biases in basketball, as referees tend to show racism by calling more fouls on players with different races than them, leading to these teams scoring less points in games where this racial injustice occurs (Wolfers and Price 2012). With more white referees in the league, teams with more white players may have team statistics skewed to make them seem better than they truly are, where teams with more players of colour will look worse because of the bias. We must be aware that our findings have been affected by these biases and as a result may not be perfectly correct. Nevertheless, a lot can be learned from the datasets which are still quite informative.

The rest of this paper will be as follows: Section 2 goes over the datasets being used in the paper, and analyzes the data with the use of tables and graphs to see which team statistics have a relationship with winning percentage. Section 3 includes the creation of a linear regression model that uses only the most important team statistics as explanatory variables to predict winning percentage, the response variable that determines team success. Section 4 interprets the coefficients and provides a statistical analysis of the model in order to understand what is being done. Finally, Section 5 discusses what exactly the model and its outcome is telling us, what can be learned from it, and what implications come along with the findings.

2 Data

The data being used to answer our research question in this paper is from Sports Reference, and specifically the college basketball section of the site (Sports Reference 2022). Sports Reference is a fantastic site that contains hundreds of thousands statistics for all of the most popular American sports, making it a site that many rely on constantly for their sports information. Their college basketball section includes data dating all the way back to the late 1800's, and includes the results from every single game since 2010, giving us a plethora of information to potentially use. This data is very reliable as it is taken directly from the box scores of each game, which has been tracked by both NCAA certified referees and scorekeepers (Sports Reference 2022). Each section of data on their site also gives you an option to export their data, stating that use of their datasets is allowed as long as they are giving proper citation, meaning that using their information for research is ethical.

Of the websites abundance of data, we will only be working with team statistics from the 2017-18 season to the most recent 2021-22 season, a 5 season period. The reason for this is that basketball is a constantly changing sport where teams adapt to new trends and game plans that optimize their success. The game is being played differently today than it was even a decade ago, so our results will be able to more accurately predict future results if only more recent data is used. Each of the five datasets contain several advanced team regular season statistics for every NCAA Division 1 school in a given season. These statistics include and average a teams results from each game in the season. With statistics from the 5 datasets being used, which were all combined into one large dataset for use, there are 1762 observations of schools and their results that we can work with. Since only 5 seasons are used, this dataset is still only a sample of all college basketball statistics of all time, as opposed to being true population data. Also, as a result of COVID-19, many games were canceled from 2020 to 2022 due to outbreaks and team sicknesses, so the data still would not be considered population data for the past 5 seasons. Although we are working with sample data, between the amount of seasons used, the number of schools taking part in each season, and the amount of games each team ends up playing, our dataset contains a lot of valuable information. It should give us a large enough sample size to work with that will allow us to accurately make predictions based off of it in our model.

While the data given by Sports Reference is very accurate and comes from a reliable source, there are some

issues with the tracking of this data that must be considered as they may cause future implications. Sports data in general is often inaccurate as it is all recorded by scorekeepers who are prone to both human error and having biases. This is especially true for basketball, as there is a lot of ambiguity in the definitions of certain statistics such as assists, blocks, and steals, meaning different scorekeepers interpret them in different ways (Williams 2017). This causes some team score keepers to either under count or over count certain statistics, and skew the results away from their true values, causing bias in the tracked data. A particularly notable example of scorekeepers tracking data incorrectly due to human error is seen in the National Hockey League. Popular arena Madison Square Garden (MSG), home to the New York Rangers, and its scorekeepers for are known to track shot locations far too close towards the net. This causes them to extremely overstate the quality of the shots in games played at MSG, measured with the statistic "Expected Goals" predicting the chance of a shot resulting in a goal, relative to the other arenas in the league (McCurdy 2020). As a result of this, the Rangers team statistics are often extremely biased and skewed to show them having stronger offensive and goaltending performances, and weaker defensive performances (McCurdy 2020). It is likely that situations like this also occur in college basketball, especially considering some NCAA basketball games are played at MSG as well and would cause our data to be biased. Finally, there is known racial bias in basketball and data may not be tracked ethically. Studies have shown that the referees in the National Basketball Association call less fouls against players who share the same race as them, and more fouls against those who are a different race, ultimately causing players of the same race to score more points in the game being played (Wolfers and Price 2012). This likely goes for NCAA basketball as well and may be even worse since it is a lower and non-professional level of play. Considering there typically are more white referees than other races in basketball, due to the makeup of the US population, this is problematic and unfair towards players of colour. This racial inequality could mean that teams with a larger white makeup will have team statistics skewed positively since they will receive more calls and thus score more points, and those with more players of colour will have negatively skewed team statistics.

Even with these biases, the data should not be affected too substantially, and information can still be learned by analyzing it. This can be done using packages such as Tidyverse (**???**), Dplyr (**???**), modelsummary (**???**), and more in R (R Core Team 2020), where the data will be looked at graphically and modeled to answer the research question and tell a story about which team statistics play the biggest factor in team success. Before getting to the modeling, we ust first take a deeper look into the data that we will be using to do so.

Table 1: 10 Observations from dataset of NCAA Basketball team statistics

Year	School	Over .500	SRS	Pace	O-rating	3pa Rate	TS %	TRB %	AST %	STL %	BLK %	TOV %	ORB %
2017-18	Austin Peay	TRUE	-3.28	70.1	104.7	0.294	53.2	52.2	49.5	9.4	8.9	16.8	35.7
2017-18	Green Bay	FALSE	-8.51	74.1	102.2	0.401	53.7	48.7	53.4	9.2	7.8	16.2	26.3
2018-19	South Carolina Upstate	FALSE	-15.30	69.4	96.9	0.436	51.5	46.7	56.0	9.2	8.1	17.3	26.2
2018-19	Xavier	TRUE	9.61	66.5	107.0	0.374	55.3	53.4	56.3	8.1	10.6	16.5	32.2
2019-20	Duke	TRUE	22.55	72.9	111.4	0.315	55.9	53.8	52.1	11.2	13.6	15.2	34.8
2019-20	Eastern Michigan	TRUE	-3.12	68.1	95.4	0.378	50.9	48.6	42.9	14.4	11.6	18.3	27.8
2020-21	Central Connecticut State	FALSE	-15.12	72.0	95.2	0.376	51.6	44.7	51.7	9.6	7.1	17.6	23.4
2020-21	Marshall	TRUE	5.70	73.1	108.7	0.410	57.1	48.7	53.2	10.6	13.2	14.6	23.8
2021-22	Eastern Kentucky	FALSE	-6.75	73.0	106.5	0.484	52.5	47.5	54.7	13.7	10.1	13.3	29.7
2021-22	Lipscomb	FALSE	-8.42	71.2	104.0	0.417	56.3	49.8	59.1	6.2	7.0	16.4	23.1

Table 1 allows us to look at the actual raw data as it contains an extract of 10 of the 1762 observations from the dataset being used. Two observations from each of the five seasons were randomly selected to be used. With that being said, each observation shown contains the team statistics for a given school in a given season or year. For example, row one shows the team statistics for the school team of Austin Peay in the 2017-18 season. Of the team statistics, two important variables to notice are SRS and Over .500, as they will be the team statistics used to identify a teams success. SRS, Simple Rating System, is a rating given to a team that takes into account both how much the team has been winning or losing their games, and the quality of opponents their games have been against (Kubatko 2008). This will be our response variable that is used to quantify team success, as it is a better indicator of a teams performance than something like straight up wins or winning percentage that would typically be used. For example, if Team A beats the first place team by 10, and Team B beats the last place team by 1, although they both have 1 win, Team A's win is clearly more

impressive and successful, and Simple Rating System takes this into account. Over .500 is a created boolean variable that is True if a team won more than half their games, or had a winning percentage over .500, and False otherwise. While once again winning percentage is not as good at showing team success as SRS and thus will not be used in the model, it is still a good indicator that will be used in visualizations to show the difference between the better teams that win more and the losing teams.

All the variables to the right of SRS in Table 1 are the numeric team statistics that can potentially be explanatory variable in the model. The team statistics included in the model will only be the most important ones that clearly play a role in making a team better or more successful. There were additional statistics in the original dataset, but they were removed as they did not reasonably seem like they had an impact on how well a team did, so they were removed. The remaining statistics in order are as follows; pace, offensive rating, 3-point attempt rate, true shooting percentage, total rebound percentage, assist percentage, steal percentage, block percentage, turnover percentage, and offensive rebound percentage. Once only the main and important statistics from this list are chosen they will be defined and looked at further, but for now we will focus on analyzing our response variable and its relationship with these team statistics.

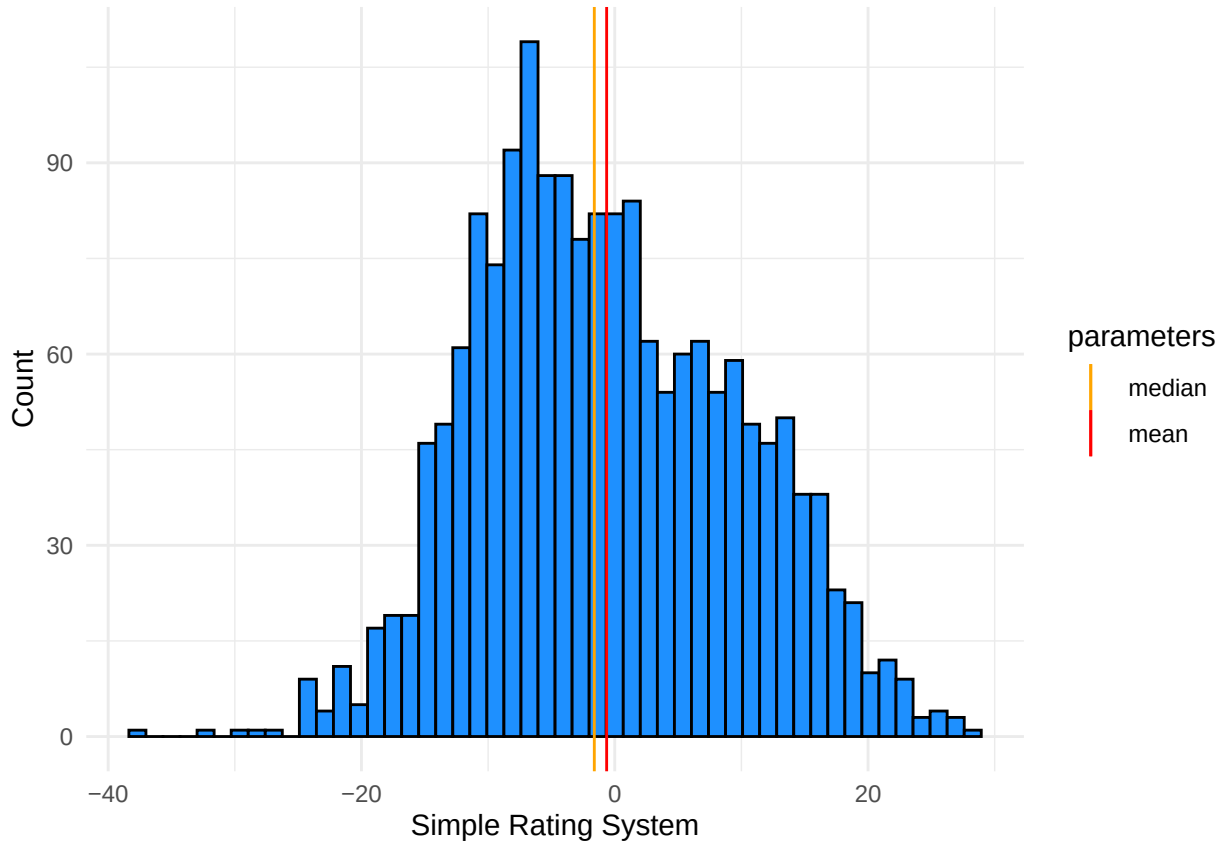


Figure 1: Distribution of Simple Rating System for NCAA Basketball teams

Figure 1 shows the distribution of our response variable simple rating system. It can be seen that the variable is close to resembling a normal distribution based on its shape, and even a standard normal distribution since it is almost centered around 0. The distribution is slightly skewed to the right, evident by the larger mean (red line) than median (orange line), and the mode being lower than both. There are also some noticeable outliers on the negative side of the distribution, which may be dragging the mean down lower than it would be without them. Even with the skew, it seems that there are close to similar amounts of teams with postive and negative SRS's, which makes sense as the better half of the league will often beat the worse half of the league, making the better teams have positive ratings and the poorer teams have negative ratings.Overall,

the fact that SRS is almost a normal distribution if not for the slight skew is a good sign that our model will be able to predict the variable successfully. One of the most important model assumptions the Normality assumption stating the models errors must be normally distributed. This is more likely to occur when the response is normally distributed, and thus it is likely we will not have to make any transformation on our variable to get more accurate and trustworthy results.

Now, we must compare simple rating system to each of our potential predictor variable team statistics, and Figure 2 shows just this. The figure contains 10 different scatter plots, with each one showing the relationship between simple rating system and another team statistics. This will allow us to see which team statistics play a role in whether a team is successful and has a higher simple rating system value or not. If SRS increases as a certain team statistic increases, they have a positive relationship and the variable will be considered for the model. The same goes for variables that have a negative relationship with simple rating system, meaning SRS decreases as the variable increases. The plots where the points are randomly scattered all throughout suggest that there is no relationship between simple rating system and a given team statistic. Additionally, the points are coloured in turquoise if the team the point is representing had a winning percentage over .500 in the given season, and a light shade of red if the team had a winning percentage under .500. This gives us an additional piece on insight on whether a variable leads to team success or not. If the blue points representing teams that win more often are more towards a certain side of the plot, it means that a given team statistic causes teams to win more when it is higher or lower. For example, if all the blue dots are on the right of the plot showing the relationship between SRS and statistic x, and all the red dots are on the left, it means more successful teams have higher values of statistic x and worse teams have lower values. Again, this is not as informative as just looking at simple rating system, but it adds a visual that can easily be seen and interpreted which should for the most part back up what the relationship between SRS and another variable is telling us. So, overall, we are looking for variables to include in the model that have a noticeable upwards or downwards trend with SRS, and also for plots where the colours are more separate towards different sides.

Looking at the plots, there are four variables that immediately stand out and have a noticeable trend with simple rating system, and should be considered for use in the model. These variables are offensive rating, true shooting percentage, total rebound percentage percentage, and turnover percentage. Offensive rating is simply defined as the amount of points scored per 100 team possessions, so the positive relationship with SRS makes sense as the more points a team scores the more likely they will win games and win them by a larger margin. Next, true shooting percentage is essentially the percentage of all the shots a team takes that they actually make. This was chosen for use over normal shooting percentage as unlike the normal statistic it includes free throws, an important part of the game, and also is adjusted based on where the shot was taken from, whether it was a free throw, 2-pointer, or 3-pointer, through the use of different weightings in the calculation formula (Jacobs 2017). This means the stat truly captures just how strong or weak a team is at overall shooting. Once again, the positive relationship with SRS makes sense as better shooting teams will score more shots and be able to keep up with other good scoring teams, leading them to more success. Futhermore, total rebound percentage, which shows the percent of possible rebounds a team grabs, also seems to have a positive upwards trend with simple rating system. Although on the surface this is less obvious that what was determined for the other variables, it means that teams that grab more rebounds on both ends of the court will have more success. Finally, turnover percentage is defined as the amount of team possesstions that end in a turnover. The lower this stat is the better, and thus unlike the other variables it has a negative relationship with SRS, where teams that turn the ball over more often have lower ratings. This is also reasonable as less turnovers should lead to both more scoring opportunities for the team while simultaneously leading to less chances for the opposition.

Each of these relationships are further evident by looking at the difference between where the blue dots representing teams above .500 are and where the red points representing teams below .500 are. In the plots for offensive rating and total rebound percentage, the two colours for the most part are almost completely separated, where blue points are more towards the right and red dots are on the left. This means that teams with higher values for these two stats win more games than those with lower values, which agrees with the variables positive trends with simple rating system. Although the teams with better records are a little more mixed in with the teams with worse records in the plots for true shooting percentage and turnover percentage, they still differ greatly towards the far ends of the plots. The right side of true shooting percentages plot

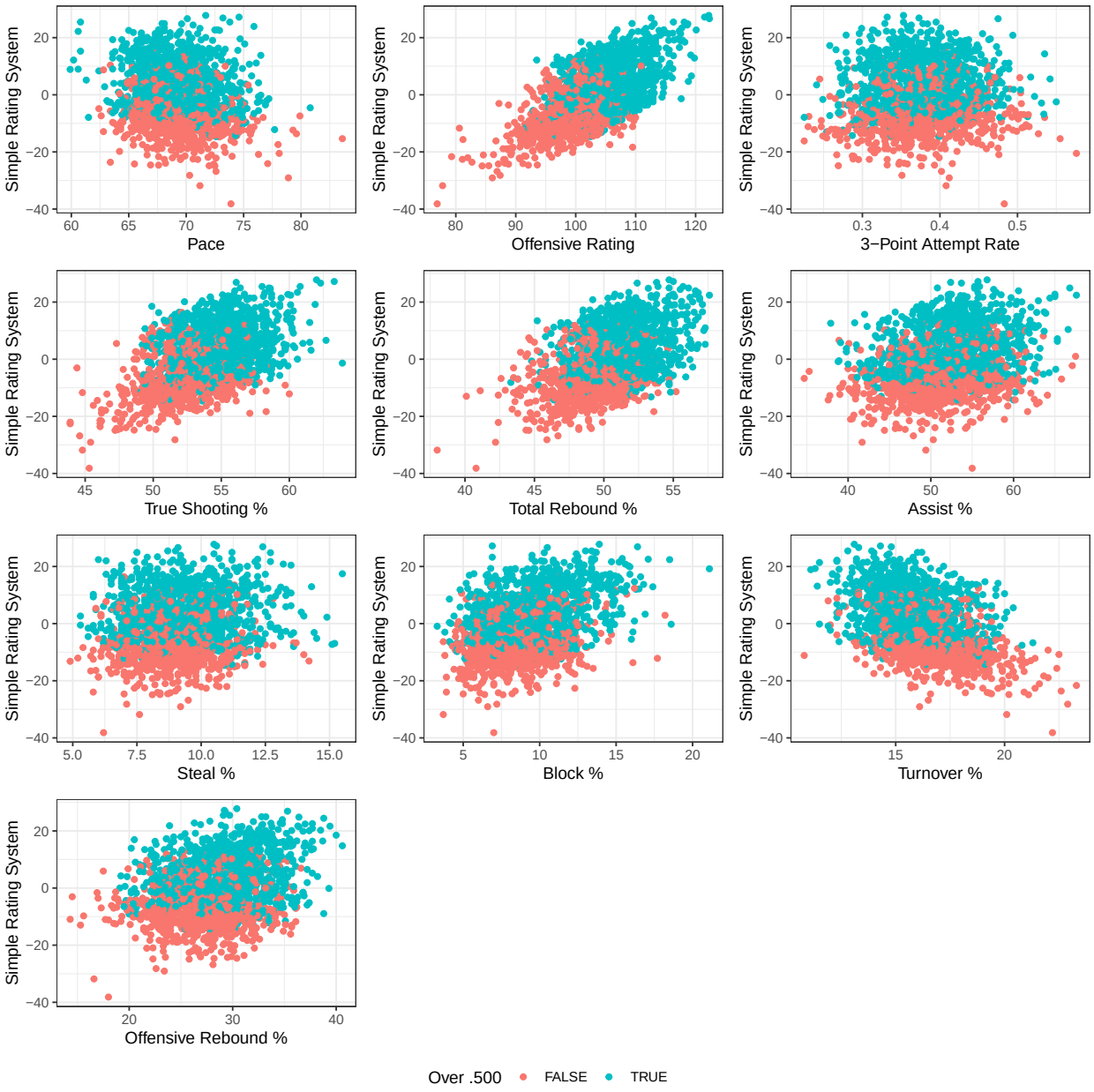


Figure 2: Relationship between Simple Rating System and several adcanced team statistics, coloured by whether a schools winning % was above .500 or not

with SRS is mostly blue points showing more winning teams, and the left side has more red points showing losing teams, backing up the positive trend of the plot. The same goes for the plot with turnover percentage, although the colours and their sides are flipped since it has a negative relationship with SRS.

The remaining 6 variables from the dataset that were plotted, pace, 3-point attempt rate, assist percentage, steal percentage, block percentage, and offensive rebound percentage, do not seem to have a noticeable trend in their relationships with simple rating system. The points for each plot are randomly scattered throughtout, and the structure of the plot does not change or differ as the variables increase or decrease, meaning simple rating system is not affected by any of these team statistics. The two colours showing teams above and below .500 also do not differ greatly in the plots, and are mostly just stacked on top of each other, with the blue dots at the top, which is of course caused by simple rating system as opposed to the variables on the x-axis. This further proves none of these varibales have any real and consistent affect on how successful a college basketball team is. The one variable that you can make a case for is offensive rebound percentage as there seems to be a slight positive trend with SRS in its plot. Despite this, the teams with better records and worse records differ minimally in the plot, telling us that the stat does not affect team success too greatly. Also, offensive rebounding percentage is likely to have a lot of multicollinearity with total rebounding percentage since the have the same definition except offensive rebound percentage only includes offensive rebounds. Due to this, it would be best if we do not include both and just stick with total rebound percentage which had the stronger relationship.

So, none of these 6 variables will be considered for the model, and only offensive rating, true shooting percentage, total rebound percentage, and turnover percentage will potentially be used to predict simple rating system, our response variable, and help us determine what leads to team success. Each of these variables are once again numeric, with offensiove rating being a rate per 100 possessions, and the remaining three other variables all being in the form of percentages,

Table 2: Summary statistics of important variables

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
srs	1418	0	−0.6	10.3	−38.2	−1.6	27.8
o_rtg	295	0	102.9	6.1	76.9	103.0	122.3
ts_percent	154	0	53.8	2.8	43.9	53.7	63.9
trb_percent	146	0	50.3	2.6	38.0	50.3	57.6
tov_percent	108	0	16.2	1.8	10.8	16.2	23.3

Table 2 shows several summary statistics of each of these statistics that will be considered for the model, including the response. Unique shows the number of different unique values for each team statistic. Missing shows the percent of values missing in the data for each variable, and each one is 0 since all NA values were removed in the initial cleaning process of the dataset. The noteworthy statistics included in the table are mean, standard deviation, minimum, median, and maximum. We once again see how simple rating system has a mean and median quite close together and is centered very close to one, showing how the distribution is only slightly skewed and is close to being a normal distribution. With a standard deviation of just over 10, there is a decent amount of spread in the distribution but nothing too substantial. The larger max and min may cause some implications in the future but only minimally.

All the potential explanatory team statistics seem to have little to no skew at all as the mean and median of each variable are extremely similar. They are actually the exact same for total rebound percentage, with a mean and median of 50.3, and for turnover percentage, with a mean and median of 16.2. Although they are not identical, the mean and medians for offensive rating and true shooting percentage are nearly the same as well, as they only differ by a tenth for both, as they are 102.9 and 103 respectively for offensive rating, and 53.8 and 53.7 for true shooting percentage. Each variable also seems to have fairly reasonably low standard deviations, meaning no variable has a spread too high that will affect our model. Based on the maximum's and minimum's, there does not seem to be any crazy high or low values that are serious outliers and would skew our results too drastically either, so overall the data looks quite good.

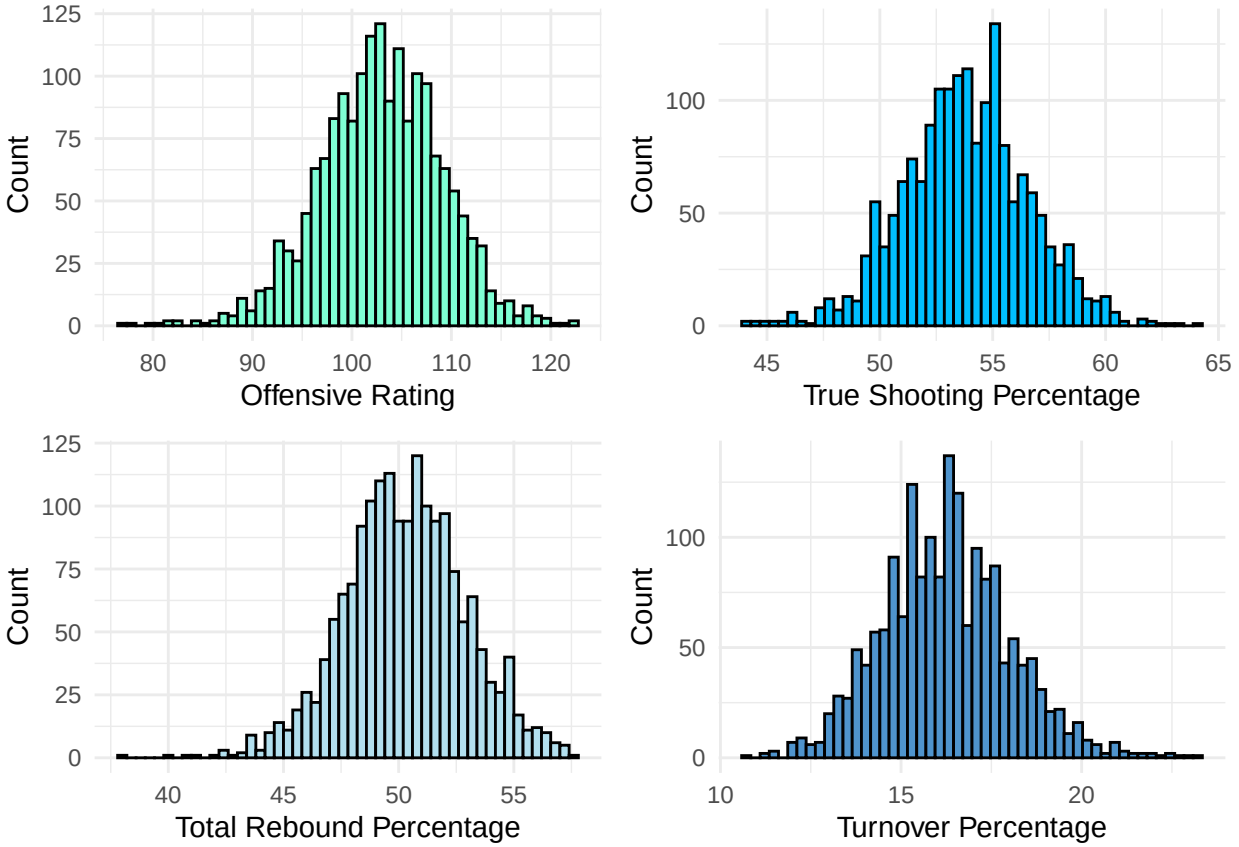


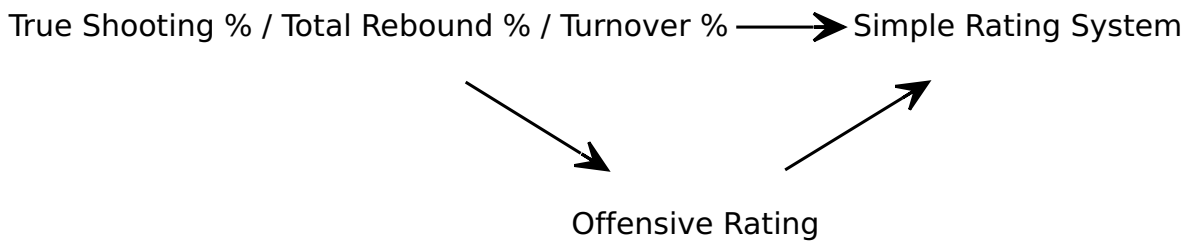
Figure **??** shows the distributions for each of these important team statistics that may be used as predictor variables, which allows us to more easily and visually see what was learned in the summary table. It can be seen that each variable seems to very closely resemble normal distributions, where they are centered around their mean and are fairly symmetric on both sides of it. Of course they are not perfectly symmetric as that would be near impossible, but the overall shape of the distributions are still very similar to that of a normal distribution. Offensive rating and total rebound percentage have a slight tail to the left of the distributions skewing them slightly to the left, and true shooting percentage and turnover percentage are the opposite where they are slightly skewed to the right due to their tails. However, they tails are quite small and not too noticeable, so the distributions on not skewed very much, which is evident by the means and medians of each variable being the exact same or nearly the same. Additionally, although there are some points at the end of the tails that may alter the distribution slightly, we can now more clearly see that there does not seem to be any substantially large outliers for any of the variables that would skew results too drastically. All in all, these team statistics should make for great explanatory variables and should not cause any issues in our model.

Now that our dataset and the variables within it have been looked at and analyzed through the use of numeric tables and visualizations, we understand what exactly we are working with and can being to create our model that will help answer our research question.

3 Model

While we do have four strong candidate explanatory variables that may be used in the model, we still must decide what combination of these variables will make for the best possible model. First, it is a good idea to think logically if each of these variables should be included in the model, and the variable in particular we should analyze is offensive rating. Based on the visualizations analyzed in Section 2, it is clear that there is a relationship between offensive rating and a teams simple rating system, but that does not necessarily mean there is direct causality. Of course if a team scores more points and has a higher offensive rating, they

will win more games and have a higher simple rating system value, but a high offensive rating is not what is causing a team to have success. It is instead underlying statistics that cause a team to score more points and thus have a higher simple rating system value. These underlying statistics, such as true shooting percentage, total rebounding percentage, and turnover percentage, affect both offensive rating and simple rating system and thus make it seem like offensive rating and SRS have a causal relationship. For example, teams with a higher true shooting percentages will hit more of their shots causing them to score more points, leading to a higher offensive rating, and thus causing them to win more games by larger amounts, and have a higher SRS value. So, these statistics are the true variables affecting simple rating system, and offensive rating is actually just acting as a mediator for the relationship between them. This relationship can be shown through the use of directed acrylic graphs (DAGs) created with DiagrammeR (**???**), seen in Figure **??**.



In the DAG, it can be seen how true shooting percentage, total rebound percentage, and turnover percentage all affect and have causality with simple rating system. It can also be seen that offensive rating affects simple rating system too, but it is also affected by the three variables listed above and is being used as a mediator between these variables and simple rating system. As this mediator, offensive rating is explaining exactly why these three variables affect simple rating system (Bhandari 2021). The three variables affect team success because the three variables cause teams to either score more or less points on offense, giving them either a better or worse offensive rating, and thus causing them to either win or lose more games, which is essentially what the DAG is explaining. As a result of this, offensive rating should not be used in the model as the mediator will affect the strength of the relationship between the true predictor variables and simple rating system.

With the final three predictor variables, there are 7 possible combinations that can be used as predictors for a linear regression model. We can then test these models against one another to determine which model is the best at predicting simple rating system, and the variables used in the model will be the main team statistics that affect how much success a college basketball team has. Before creating the models and testing them, we must first split our data into two groups, a training group and a testing group. This is done to ensure we do not overfit our model, which would make it difficult for our model to be applied to other datasets and accurately make predictions on different data (Alexander 2022). The training data will be used to actually build and fit the model, and thus is also used to get our parameter estimates. As a result of this, the data should be split in a way that the training dataset is larger, allowing it could make more accurate estimates, and thus an 80:20 split has been chosen. The testing data will be used later on to validate our model and see if it can be used to make predictions on different datasets than the one used to build it.

Using our training dataset, the seven different models were created, one with all three final variables included, three with two of the variables included, and three where one variable is used as a single predictor. We can then obtain the values of several different goodness of fit and maximum-likelihood measures that tell us how well each model fits the data. Table 3 displays the estimated coefficients of each potential final model, and the several different goodness of fit measures that were aforementioned, where each column represents a different model. Based on the measurements in the table, it seems as if the largest model that uses true

Table 3: Comparing goodness of fit measures of several potential models predicting Simple Rating System. Column names represent which variables are included in each model.

	mod_ts_trb_tov	mod_ts_trb	mod_ts_tov	mod_trb_tov	mod_ts	mod_trb	mod_tov
(Intercept)	−98.62	−152.76	−44.00	−58.93	−98.57	−107.96	44.20
	se = 5.42	se = 4.96	se = 5.43	se = 4.53	se = 4.60	se = 4.41	se = 2.26
ts_percent	0.92	1.27	1.44		1.82		
	se = 0.08	se = 0.08	se = 0.08		se = 0.09		
trb_percent	1.63	1.66		1.94		2.13	
	se = 0.08	se = 0.09		se = 0.08		se = 0.09	
tov_percent	−2.04		−2.10	−2.41			−2.77
	se = 0.11		se = 0.13	se = 0.12			se = 0.14
R2	0.512	0.402	0.361	0.461	0.244	0.297	0.221
R2 Adj.	0.511	0.401	0.360	0.461	0.244	0.296	0.220
AIC	9602.5	9886.8	9980.8	9739.6	10 215.0	10 113.4	10 257.6
BIC	9628.8	9907.8	10 001.8	9760.6	10 230.7	10 129.1	10 273.4
RMSE	7.29	8.07	8.34	7.66	9.07	8.75	9.20

shooting percentage, total rebound percentage, and turnover percentage is the best possible model.

The estimated coefficients are not of too much use when comparing the models, and the coefficients of the final model will be analyzed in 4. The standard errors of the estimates are of use to us though, as a smaller standard error means the model has estimated the parameters more precisely. The model with all three variables used has the lowest standard error for each of the slope coefficients, showing it is doing the best job at estimating these parameters. It's intercept standard error is quite high on the contrary, which is expected since it is the largest model with several predictors in use. Next, the R^2 and R^2_{adj} measures also point to the largest model being the best. The measures describe how much of the variance in our response is explained by the model, with R^2_{adj} penalizing larger models since more predictors causes R^2 to increase. Even with the penalty, the model with all three variables still has the highest value at 0.511, which is 0.05 points higher than the next best model, meaning it fits the data much better than the rest. Ideally that number would be closer to 1, but explaining over half of the varation still makes the model fairly successful. Additionally, the Aikake's Information Criterion (AIC) and Bayesian Information Criteria (BIC) log-likelihood based measures also point towards the most complex model being the best. These measures tell us how well the model fits the data, while also penalizing models with more parameters, with BIC having a stronger penalty than AIC (S. 2010). Even as the largest model and being penalized the most, the model will all three variable once again has the best results, as both its AIC and BIC are much lower than each of the other models values, and it is not particularly close. While it already fairly evident the full model is the best, looking at the Root Mean Squared Error (RMSE) only backs up this claim further. The RMSE is the variance of all the models residuals, and shows us how close the models predicted fitted values are to the actual points in our data (Grace-Martin 2013). A lower RMSE means a model has less error and fits the data much better. The largest model once again has the lowest value at 7.29, which is significantly better than the next best model, and further shows that this model fits the data the best and should be used as the final model. The only model that comes relatively close to the full model is the one with total rebound percentage and turnover percentage, as it proved to be the second best model for each of the measurements, with the second highest R^2 and R^2_{adj}, and second lowest AIC, BIC, and RMSE values. While its measurements were fairly good and it was considered as a possibility for the final model, the model with all three variables still was better in each measurement and by a significantly large amount as well, making it the clear and obvious choice to be used as our final model.

So, now that we know which variables will be included in the final and best model, the form of our final model can be seen below.

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \epsilon$$

where:

- $\hat{y_i}$ *represents the predicted value of SRS given specific values of the predictor variables*
- $\hat{\beta}_0$ *is the estimated intercept coefficient*
- x_1 *represents the predictor variable of True Shooting Percentage*
- $\hat{\beta}_1$ *is the estimated slope coefficient for True Shooting Percentage*
- x_2 *represents the predictor variable of Total Rebound Percentage*
- $\hat{\beta}_2$ *is the estimated slope coefficient for Total Rebound Percentage*
- x_3 *represents the predictor variable of Turnover Percentage*
- $\hat{\beta}_3$ *is the estimated slope coefficient for Turnover Percentage*
- *and* ϵ *is an error term which shoud be 0 assuming our model is unbiased*

Essentially, the model works by plugging in values of our predictor variables for the x's, and then based these given values and the intercept and slope coeffcients, the model will predict what a teams Simple Rating System Value will be. In Section 4, we will analyze the results of the model and its parameters, plug the coefficient parameters into the model, and interpret what exactly the model means.

4 Results

Table 4: Summary of parameter values and statistics for linear model predicting Simple Rating System using various team statistics

	Estimated Coefficient	95% Confidence Interval	P-Value
Intercept	-98.62	(-109.25, -87.98)	1.38e-66
True Shooting %	0.92	(0.77, 1.07)	5.39e-32
Total Rebound %	1.63	(1.47, 1.78)	2.01e-84
Turnover %	-2.04	(-2.27, -1.82)	5.13e-64

Table 4 displays the results of our models estimated parameter statistics. It displays each of the estimated coefficient values, the 95% confidence interval for each estimated coefficient, and a p-value stating the significance of each variable. The parameters can be interpreted in order for us to better understand the relationship between the three predictor variables used in the model and Simple Rating System, or a college basketballs team success, shown in the model, and this is especially true for the estimated coefficients. The intercept coefficient of -98.62 tells us what a teams SRS will be if all predicted variable values are equal to zero. So, if in a given season a team hits 0% of their shots, grabs 0% of all possible rebounds, and turns the ball over in 0% of their positions, they will have a Simple Rating System value of -98.62. This is a reasonable intercept value as if a team hits no shots they will score no points, and by grabbing no rebounds they will give the other team many chances to score, so the team will lose many games by a large amount, resulting in the predicted very low SRS value. Of course it is near impossible for a team to have a season where they do not make a shot, do not grab a rebound, and never turn the ball over, but this is just a baseline value that helps us understand how the model works.

Next, the slope coefficient for the true shooting percentage variable is 0.92, meaning that when all other predictor variables are held at a fixed value, if true shooting percentage increases by 1% then their simple rating system will increase by 0.92. Total rebound percentage has a slope coefficient even larger at 1.63. This coefficient value of 1.63 can be interpreted as how much a teams simple rating system increases by when their total rebound percentage goes up 1% and all other variables are fixed. Both of these coefficients make sense as simple rating system will logically increase if a team shoots more efficiently and scores more points or if they grab more rebounds and get more attempts to score that their opposition, which was already determined by looking at the relation graphically earlier. The estimated slope coefficient for turnover percentage is different from the other as it is negative, with a value of -2.04. This lines up with what was seen in the graphical

analysis as simple rating system and turnover percentage have a negative relationship, since turning the ball over more will allow the other team to score more and lead to a worse SRS. This negative value tells us that when a teams turnover percentage increases by 1% and all other variables are held fixed, the teams simple rating system value will decrease by 2.04 units.

We can plug all these coefficients into the $\beta's$ in the formula explained in Section 3 to get the completed formula for our final model, $\hat{y_i} = -98.62 + 0.92x_1 + 1.63x_2 - 2.04x_3$. Doing so makes our model much easier to interpret. In order to estimate what a teams simple rating system value would be, you plug in their true shooting percentage for x_1, their total rebound percentage for x_2, and their turnover percentage for x_3, and the output of $\hat{y_i}$ would be their rating. For example, if a team has a true shooting percentage of 50%, a total rebounding percentage of 40%, and a turnover percentage of 20%, the model estimates that their simple rating system would be $\hat{y_i} = -98.62 + 0.92(50) + 1.63(40) - 2.04(20) = -28.22$. This shows how the model can successfully predict a teams success with the use of advanced team statistics such as our predictor variables that were used, answering the research question.

Table 4 also contains 95% confidence intervals. These intervals give us a range of plausible values for each estimate coefficient, and we are 95% certain that the true population parameter value is within this interval. For example, the true shooting percentage slope coefficient can reasonably be any value between 0.77 and 1.07, and our estimate coeffcient value of 0.92 for true shooting percentage falls right in the middle of this range. There is a chance this interval does not capture the true value but it is unlikely, as if we took 100 samples of college basketball data and created models, only 95% of the models and their confidence intervals for the estimated coefficients would capture the true population values. Fortunately, each of the estimated slope coefficients confidence intervals are quite narrow, as the lower and upper bounds only differ by 0.3, 0.31, and 0.45 respectively. The narrowness of the intervals is good as it means that our estimated coefficients are most likely quite accurate, and any coefficients and their confidence intervals that are estimated from other samples would give values very similar to the ones obtained from our sample that was used (Bassett 1999). The intercept coefficient also has a fairly narrow confidence interval as well, from -109.25 to -87.98, even if it may not seem like it relative to the other coefficients. This is because it is not dealing with percentage variables like the slope coefficient and instead has to do with simple rating system which has a wider range of values.

The final value in the summary table is the p-value of each estimated coefficient, which tell us how significant each estimated coefficient is. A low p-value, and one below 0.05 in particular, for a variables coefficient means that there is evidence against a null hypothesis stating the coefficient should be zero as the variable does not affect the response in the presence of the other variables in the model. Looking at the p-values for our coefficients, they are all extremely small and are very close to 0, meaning the null hypothesis is rejected and thus each of the coefficients are significant and belong in the model. However, this information must taken with a grain of salt, as p-values are difficult to work with since they assume all model assumptions are satisfied and the data was properly collected and cleaned (Alexander 2022). As a result of this, no decisions were made based on these values, but they do provide us with a piece of information that allow us to further see that the decision we did make are likely good ones due to their low values.

The relationship between simple rating system and our three explanatory variables can also be visualized despite being a 4D relationship, thanks to the package plot3D (**???**).

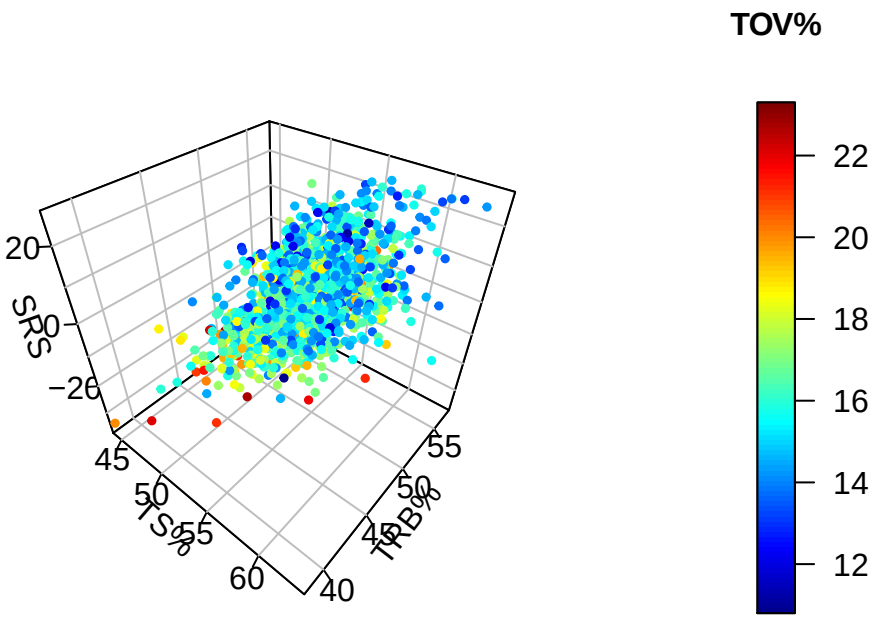


Figure **??** shows this relationship through the use of a 3D plot, which is technically a 4D plot based on the use of a colour gradient to show a 4th variable. The vertical z-axis represents our response variable SRS, the x-axis closest to the viewer represents true shooting percentage, and the remaining y-axis on the right represents total rebound percentage. Also, the colour gradient is used to represent turnover percentage where each point is coloured in differently based on its turnover percentage value, giving the plot its 4th dimension.

Although it is difficult, it can be seen that the overall trend of the points moves from the bottom left of the plot towards the top right of the plot. So, as true shooting percentage increases along the x-axis and the points move right, simple rating system increases since the points are also moving upwards. Additionally, as total rebound percentage increases up the y-axis, simple rating system increases as well since the points are moving from the bottom of the z-axis to the top. Finally, the points towards the bottom of the plot where SRS is low are mostly red, orange and yellow. Then as SRS increases, the points gradually become more green and then more blue. This final dimension shows that as turnover percentage decreases and the colours go from red to blue, simple rating system increases. Each of these relationships shown by the 4D plot line up with what has been seen in both the graphical analysis and the analysis of the models parameter statistics. It should be noted that all three of these relationships are in the presence of the other predictor variables, and when you combine them all together you get the upwards trend and colour change from red to blue seen in the plot, which shows us the overall relationship estimated by the model.

Before going over what we have learned from our model, there are two things that must be done. First, we need to check to see if all of the assumptions that where made with respect to the model were all satisfied, or else there will be some serious limitations to our findings. Then, we need to use our testing dataset to validate our model and see if these results can be used to make predictions on other datasets.

```
ggarrange(assump_plot1,assump_plot5,assump_plot2,assump_plot3, assump_plot4, ncol=2, nrow=3)
```

The first assumption is the linearity assumption where all explanatory variables must have a linear relationship with the response, and this was confirmed to be true in Section 2 through the use of scatter plots. The

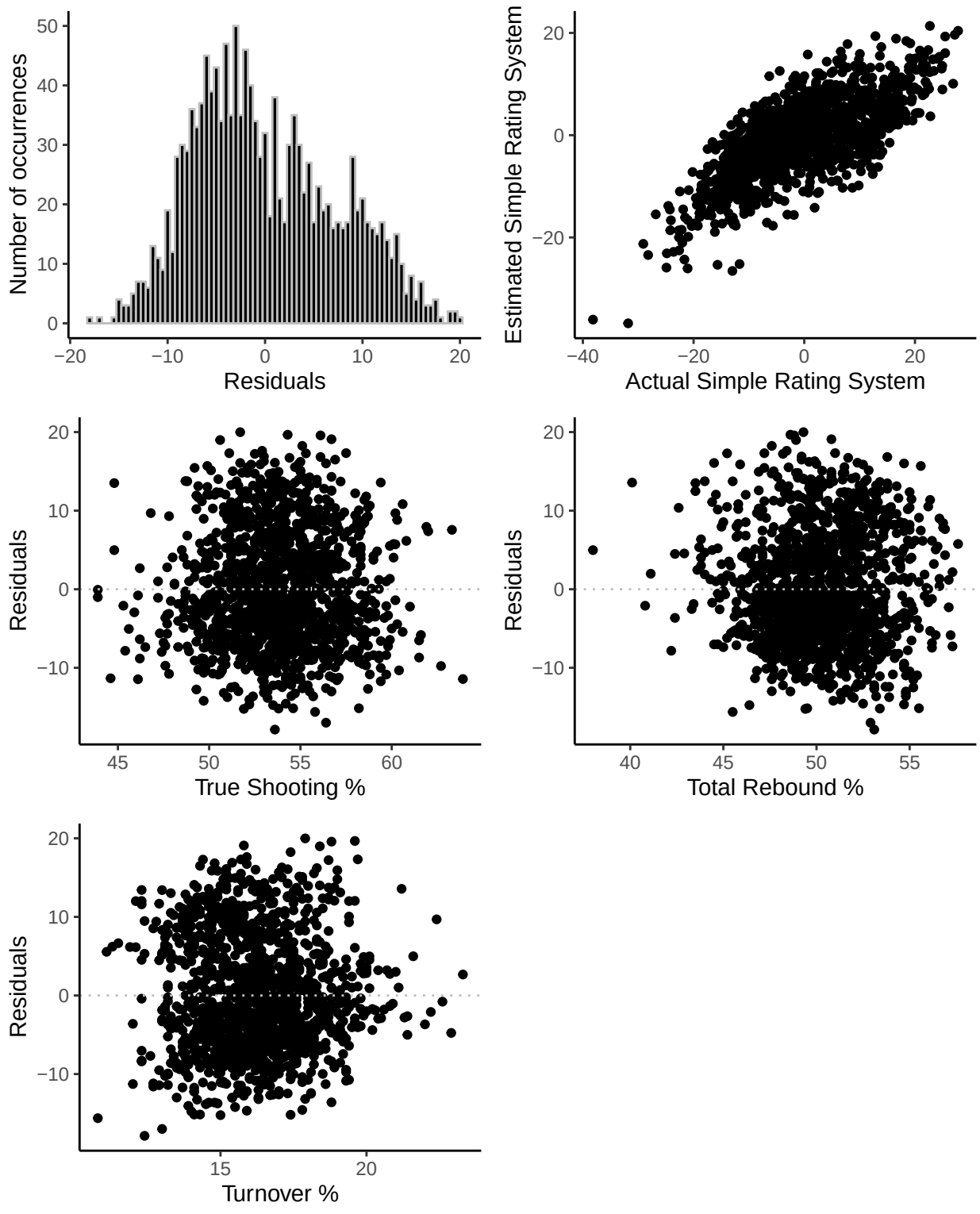


Figure 3: Residuals and Fitted Value plots for linear regression model predicting Simple Rating System using a training dataset on college basketball statistics

explanatory variables also should not have a strong linear relationship with each other, and this was also already considered and dealt with earlier through the use of DAGs in Section 3. The next three assumptions will be checked using the plots in Figure 3. The next assumptions is the independence errors assumptions, as errors should be independent of one another and show no correlation. The bottom three plots in Figure 3, which show the relationship between each predictor variable and the models residuals, prove that this assumption is satisfied. This is because the points are all randomly scattered around a horizontal line at 0 on the y-axis for each one of the three plots, and there are no separated groups of points or noticeable trends that would suggest any correlation in the residuals. The third assumption is the homoscedasticity of errors assumption which simply means the models residual errors have constant variance. The plot in the top right of the figure shows the residuals have constant variance since the points have an upwards trend along an invisible diagonal line, meaning that variance is mostly similar and constant across all fitted values. Finally, the last assumption states that the errors should be normally distributed. The histogram in the top left of Figure 3 is used to see if this assumption is satisfied, and it should show a normal distribution if it is. The histogram has a distribution very similar to the distribution of our response variable simple rating system, which we previously determined that despite being slightly right skewed, the distribution is very close to being a normal distribution centered around its mean. This seems to be the case here as well, and while it is not perfect, it is still very close to being so, meaning it should not cause many issues with our model, if any at all, and the assumption can be considered satisfied.

Finally, our model must be validated using the testing dataset. One way to do this is by making predictions on the test dataset, using the values of the explanatory variables in the data. We then calculate the RMSE using the errors between the predicted values and actual values of the testing data, and compare it to the RMSE of the original model built with the training data (Arnholt 2021). After using the model to make predictions on the testing data, it was found that the predicted values had a RMSE of 7.74. This is very similar to the original models RMSE based on the training data, which was 7.29. Since these values are very similar to one another, it means that our model did a good job at predicting values using the testing dataset, and thus it should also successfully be able to make predictions on other datasets, like for example future college basketball seasons. This means that the model has successfully been validated using the testing dataset, and we can now analyze what exactly was learned from our model.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional details

References

Alexander, Rohan. 2022. *Telling Stories with Data*. https://www.tellingstorieswithdata.com/index.html.

Arnholt, Alan T. 2021. *Machine Learning with Caret in R*. https://stat-ata-asu.github.io/MachineLearningToolbox/.

Baker, Alison. 2021. *The Most Watched Sporting Events in the World*. Road Trips: The Ultimate in Sports Travel. https://www.roadtrips.com/blog/the-most-watched-sporting-events-in-the-world/.

Bassett, Mary T. 1999. *Confidence Intervals - Statistics Teaching Tools*. New York State Department of Health. https://www.health.ny.gov/diseases/chronic/confint.htm#:~:text=What%20does%20a%20confidence%20interval,if%20the%20survey%20were%20repeated.

Bhandari, Pritha. 2021. *Mediator Vs Moderator Variables | Differences & Examples*. Scribbr. https://www.scribbr.com/methodology/mediator-vs-moderator/.

Geiling, Natasha. 2014. *When Did Filling Out a March Madness Bracket Become Popular?* Smithsonian Magazine. https://www.smithsonianmag.com/history/when-did-filling-out-march-madness-bracket-become-popular-180950162/.

Grace-Martin, Karen. 2013. *Assessing the Fit of Regression Models*. The Analysis Factor. https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/.

Jacobs, Justin. 2017. *Relationship Between Ts*. Squared 2020. https://squared2020.com/2017/10/10/relationship-between-ts-and-efg/.

Kubatko, Justin. 2008. *The Simple Rating System*. Basketball Reference. https://www.basketball-reference.com/blog/indexba52.html?p=39.

McCurdy, Micah Blake. 2020. *Scorer Bias Adjustments*. Hockey Viz. https://hockeyviz.com/txt/scorerBias.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

S., Prabhat. 2010. *Difference Between Aic and Bic*. Difference Between Similar Terms; Objects. http://www.differencebetween.net/miscellaneous/difference-between-aic-and-bic/#:~:text=AIC%20and%20BIC%20are%20widely,two%20approaches%20of%20model%20selection.

Sports Reference. 2022. *NCAA Seasons Index*. https://www.sports-reference.com/cbb/seasons/.

Williams, Rob. 2017. *SFU Study Reveals There's Scorekeeper Bias in the Nba*. Daily Hive. https://dailyhive.com/vancouver/sfu-study-nba-scorekeeper-bias.

Wolfers, Justin, and Joseph Price. 2012. *Racial Discrimination Among Nba Referees*. University of Pennsylvania. https://users.nber.org/~jwolfers/papers/NBARace(QJE).pdf.