Multiple linear regression: A data driven approach to predicting midseason NBA point spreads

Mackenzie Qu

22 December 2020

Abstract

|Basketball has gained millions of passionate fans over the years, and its intensity often extends beyond the courts into various bets. For the purpose of predicting pointspread outcome for any midseason NBA games, a multiple linear regression model is developed using team's previous game stats from the same season. The model is then tested using randomly selected midseason games from 2015. Our model for pointspread is also applicable with regards to predicting the Money Line outcome, which may provide an overview of the upcoming season betting lines.

|**Keywords:** basketball; linear regression; NBA prediction; pointspread; statistics

Contents

1 Introduction

From the squeak of sneakers on the court to the cheers of fans from the bleachers, the excitement of basketball lies within its fast and unpredictable nature, which gratifies a massive crowd of gamblers. As the NBA remains to be the best basketball league in the world, its betting market also holds significant value. The NBA is composed of 30 teams; each team plays 82 games for a regular season which provides 1230 opportunities for gamblers to cash in. Each game is divided by 4 even quarters with the possibility of overtime if the result was a tie. Upon research, the online betting pool does not close until the fourth quarter, indicating that information gathered during the first three quarters of the game could be crucial.

In this paper, we are interested in predicting the outcome of two popular bets - the Point Spread and the Money Line. As the most popular wager, the Money Line simply allows one to bet on a desired team to win. A more intriguing bet on point spread is when one bets on a team to win by a specific number or not lose more than that number. In this paper, we have chosen to predict the Point Spread, which includes the Money Line outcome by experimenting with multiple linear regression models, hoping to predict a profitable betting line on the spreads of NBA games.

For the purpose of this paper, two datasets containing the actual NBA game statistics (Kelepouris [2018])(rLoper [2020]) are used to assist the model development. The data includes all games in a regular season, along with game stats such as field goal percentage and 3 point percentage generated by each team. We have chosen to use the cumulative average of field goal percentage, 3 point percentage, free throw percentage, total rebounds, assists, turnovers, as well as the total points until the third quarter as predictor variables. Our model for point spread has a residual standard error of 7.139, which is inadequate in accuracy in comparison to the betting lines. However, the accuracy for predicting the winning team appears to be significant through verification. The data wrangling and model for this paper is done using statistical language R Core Team [2020] in Rmarkdown(Allaire et al. [2020])(Xie et al. [2018])(Xie et al. [2020]). To reproduce the result, code can be accessed at: "https://github.com/mackenziequ/NBA-prediction".

The results are able to predict point-spread and winning margin in the upcoming NBA season for all purposes. However, the model is not sufficient enough to achieve the accuracy of the NBA Point Spread betting lines chosen by professional oddsmakers due to basketball's unpredictable nature. Our model could be significantly improved by replacing the cumulative averages with actual in game data by the third quarter. The rest of the report proceeds as follows. Section 2 describes the historical NBA data we have used to develop and verify the models, including some adjustments made. In section 3, we demonstrate the model development, aiming to provide a statistical reasoning behind the results. Section 4 shows the result of the models and verifies the model performance through randomly selected games and their betting lines respectively. In which the results will be discussed thoroughly in section 5, as well as some potential weaknesses and future work.

2 Data

With the intention of predicting the NBA game results, we have chosen two self-contained NBA datasets available on Kaggel. Both datasets are observational, in which it is gathered from historical NBA games in full measure. From all the historical data, we have sampled the 2014-2015 season for model training purposes. The 2014-2015 season has in total 1230 games, the datasets however, both contain 2460 variables as each game is recorded twice for both the home team and the away team. Though the complete historical data of each team may seem useful, we have chosen to not include any historical data from season 2013-2014 due to players' leave or trade, which will result in significant differences in a team's performance.

2.1 "NBA Team Game Stats from 2014 to 2018" - Data

"NBA Team Game Stats from 2014 to 2018" (Kelepouris [2018]) contains information of every NBA game played in 2014-2018. The dataset was mined from basketball reference(cit [2020]) and published in Kaggle. Its main focus are the points and game stats for both home and away teams in a game, and it contains the main predictor variables in the model. However, some variables, such as off-rebounds and assists are not used as they do not have a significant impact on the points made in a game. The dataset is beneficial such that it

is well organized, and majority of the variables contribute to the total points gained in a game. However, it is limited to the results of a game, in which some of the variables have to be altered for model fitting since we would not have any knowledge of the outcome prior to the game. Therefore, we have calculated and attached the cumulative averages of the Field Goal Percent, Three Point Percent, Free ThrowPercent, Total Rebounds, and Turnovers for each team before the game. This allows us to develop an actual prediction model of the game results without any information of the outcomes. We have also added point spread as the difference of two team's total points into the dataset for modeling purposes, which can be seen from the preview of the data(Table 1).

2.2 "1996-2019 NBA Stats Complete With Player Stats" - Data

"1996-2019 NBA Stats Complete With Player Stats"(rLoper [2020]) created by another Kaggle contributor contains NBA game statistics with a more detailed approach. In addition to the team stats for each game, player stats are also included which makes in total 142 variables. However, we have only chosen the most favorable variable for our model - points gained in each quarter. Though basketball games are fast-paced, and the result may change drastically in any given minute, this data gives us a better understanding of the in-game statistics and performance of each team; which is proven significant for the prediction.

For the purpose of our model, we simply added the points gained in the first three quarters for each game played by each team; the players stats and other variables for each game are not used since it will create too many categorical variables in the model with little significance. In addition, from the preview of the cleaned data(Table 2), it is shown that we have kept the date and team for merging purposes. Note that all variables from the preceding dataset is included, however, this data has presented more obstacles during the data cleaning process thus was not chosen.

Upon merging the two datasets by date and team, we have a complete dataset containing all the variables needed for visualization and modeling. Note that the dataset only contains games after the midseason, as we will provide the reasoning below.

Table 1: NBA Team Game Stats from 2014 to 2018 - 2014 data preview

Team	Opponent	Home	Game	Date	WINorLOSS	PointSpread	TeamPoints	OpponentPoints	FieldGoalPercent
ATL	TOR	Away	1	2014-10-29	0	-7	102	109	0.500
ATL	IND	Home	2	2014-11-01	1	10	102	92	0.507
ATL	SAS	Away	3	2014-11-05	0	-2	92	94	0.413
ATL	CHO	Away	4	2014-11-07	0	-3	119	122	0.462
ATL	NYK	Home	5	2014-11-08	1	7	103	96	0.407
ATL	NYK	Away	6	2014-11-10	1	6	91	85	0.380

OppFieldGoalPercent	ThreePointPercent	OppThreePointPercent	FreeThrowPercent	OppFreeThrowPercent	TotalRebounds	OppTotalRebound
0.411	0.591	0.308	0.529	0.818	42	48
0.383	0.350	0.375	0.758	0.857	37	44
0.449	0.320	0.294	0.727	0.711	37	50
0.495	0.394	0.286	0.769	0.741	38	51
0.476	0.409	0.381	0.778	0.727	41	44
0.434	0.370	0.231	0.964	0.583	38	40

Turnovers	OppTurnover	AvgTeamPoints	AvgOppPoints	AvgFieldGoalPercent	AvgOppFieldGoalPercent	AvgThreePointPercent
17	9	NaN	NaN	NaN	NaN	NaN
12	18	102.00000	109.00000	0.5000000	0.4110000	0.5910000
13	19	102.00000	100.50000	0.5035000	0.3970000	0.4705000
19	19	98.66667	98.33333	0.4733333	0.4143333	0.4203333
8	15	103.75000	104.25000	0.4705000	0.4345000	0.4137500
15	15	103.60000	102.60000	0.4578000	0.4428000	0.4128000

AvgOppThreePointPercent	AvgFreeThrowPercent	AvgOppFreeThrowPercent	AvgTotalRebounds	AvgOppTotalRebounds	AvgTurnovers
NaN	NaN	NaN	NaN	NaN	NaN
0.3080000	0.5290000	0.8180000	42.00000	48.00000	17.00
0.3415000	0.6435000	0.8375000	39.50000	46.00000	14.50
0.3256667	0.6713333	0.7953333	38.66667	47.33333	14.00
0.3157500	0.6957500	0.7817500	38.50000	48.25000	15.25
0.3288000	0.7122000	0.7708000	39.00000	47.40000	13.80

Table 2: 1996-2019 NBA Stats Complete With Player Stats - 2014 data preview

Date	Team	third_quarter_score
2014-10-28	ORL	64
2014-10-28	NOP	78
2014-10-28	DAL	73
2014-10-28	SAS	76
2014-10-28	HOU	85
2014-10-28	LAL	69

2.3 Data Discussion

The original dataset contains all games played in Season 2014; upon data wrangling, we first focus on some basic data among teams.

Figure 1 shows the distribution of total points scored by team. The graph shows the lowest point scored of 65 points by DEN(Denver Nuggets), and the highest point scored of 144. It is able to show the performance level of each team by their total points. In addition, we observe that the majority of the teams have a centered distribution. However, some teams such as CHI(Chicago Bulls) have less observable performance center, in which the point scored has a lower frequency and seems evenly distributed throughout. In this case, it becomes harder to predict the expected score in a game. Whereas MIL(Milwaukee Bucks) has shown to have a significantly high frequency scoring points in between 90-100, which leads to a slightly more accurate prediction.

Secondly, we are interested in observing the cumulative averages for total points, and whether it stabilizes or keeps fluctuating. Though the running average for each team is not one of the predictor variables with the reasoning in Section 3, it still provides us an overview of each team's performance throughout the season, which provides references to the other variables such as Field Goal Percent. Figure 2 demonstrates the cumulative average for each team through the whole season; it is indicated that the average stabilizes for the majority of the teams in the second half of the season with a couple exceptions. OKC(Oklahoma City Thunder) shows a constant increase in average throughout the season, and in the second half of the season it's average has gone up by more than 5 points. On the other hand, TOR(Toronto Raptors)'s average seems to fluctuate even in the second half.

From the overview of the data, we could conclude that for the majority of the teams, the total points is centered at its mean, and the cumulative averages stabilize after midseason. Therefore we filtered the data, starting from the 41st game. We have chosen to not present the point spread distribution since it is not sufficient to convey any information by itself. Though we can conclude some information about the teams, our data is limited to the cumulative average, which is not sufficient enough to predict the Point Spread without any in-game information. Therefore, the third quarter data from "1996-2019 NBA Stats Complete With Player Stats - Data" drastically improves the model. Still, we do not have any additional in-game information to verify the performance of each team during a game besides the total points in the third quarter, which as a result, is shown by our model.

Our data wrangling is completed in the statistical language R(R Core Team [2020]), using haven(Wickham and Miller [2020]), tidyverse(Wickham et al. [2019]), dplyr(Wickham et al. [2020]), ggplot2(Wickham [2016]), ggpubr(Kassambara [2020]), modelr(Wickham [2020a]), tydyr(Wickham [2020b]), knitr(Xie [2020])(Xie [2015])(Xie [2014]), kabelExtra(Zhu [2020]), broom(cite_broom), and AICcmodavg(Mazerolle [2020])

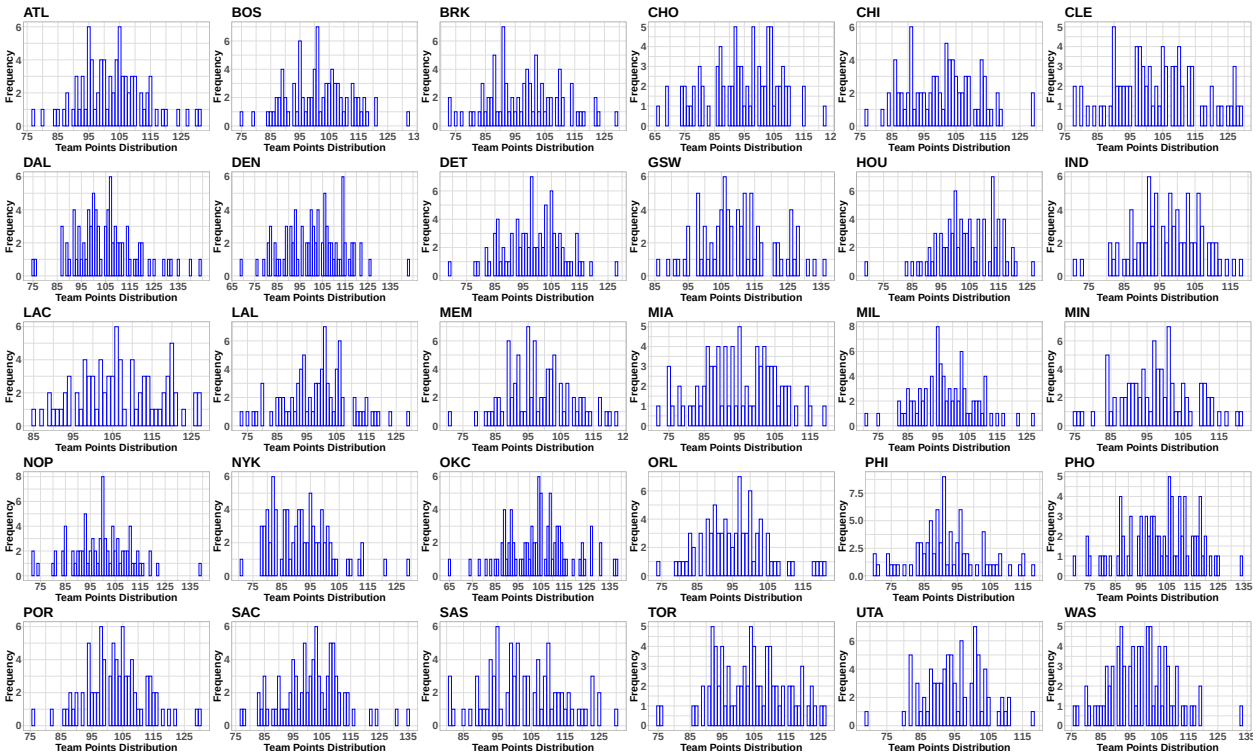

Figure 1: Total Points Scored in Season 2014 by Team

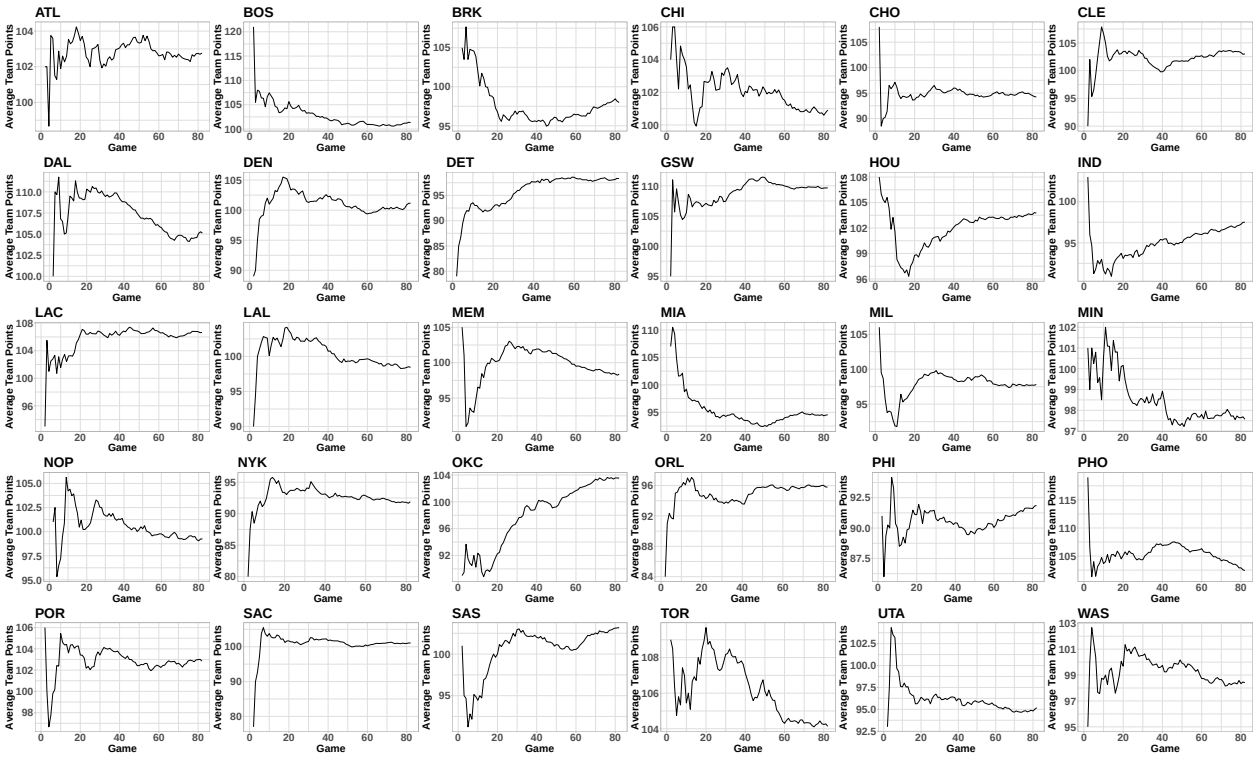

Figure 2: Cumulative Average for Total Points by Team

3 Model

3.1 Background

Prior to developing the model, some statistical backgrounds have to be acknowledged. We have chosen to use a multiple linear regression model to predict the dependent variable - Point Spread, with some independent, observable variables. Multiple linear regression focuses on suggesting a relationship between multiple independent variables and a dependent variable. General linear regression indicates that $y = X\beta + \epsilon$, where y is the $n \times 1$ column matrix, and each y_i is the response for the $i-th$ observation. X is an $n \times (k+1)$ matrix of observed constant, containing x_{ij} as the $j-th$ predictor for the $i-th$ observation. β is a $(k+1) \times 1$ matrix of unknown constant, including the regression intercept and slope; and ϵ is a $n \times 1$ matrix of unknown constant as an error term. The general model can be both denoted in matrix form or scalar form, and we are going to use the scalar form below to indicate each predictor variables. To develop a multiple linear regression model, some conditions have to be met, and failure to achieve all assumptions may result in inaccuracy. Most importantly, as a linear model, there has to be a linear relationship between each independent variable and the dependent variable, which is shown in Figure (3). Moreover, we also assumes the normality of residuals, no multicollinearlity, and its homoscedasticity, in which all assumptions will be verified in the diagnostics plot(Figure 4).

In this paper, the Point Spread of each NBA game in season 2014 is the independent variable, which is a 2460×1 vector containing all Point Spreads, and yi is the point spread for the i-th game. Team and Opponent are categorical variables, with team ATL(Atlanta Hawks) as reference. As a result, there are 58 dummy variables created by r, in which we will indicate as $\sum_{i=1}^{58} \beta_i team_i$ for convenience. All other variables are denoted by $x_1 - x_{11}$ respectively corresponding to Third Quarter Point Spread, Average Field Goal Percent, Average Opponent Field Goal Percent, Average Three Point Percent, Average Opponent Three Point Percent, Average Free Throw Percent, Average Opponent Free Throw Percent, Average Total Rebounds, Average Opponent Total Rebounds, Average Turnovers, and Average Opponent Turnovers.

As a result, the model is denoted as follows(Equation (1)):

$$y_i = \beta_0 + \sum_{i=1}^{58} \beta_i team_i + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6} + \beta_7 x_{i,7} + \beta_8 x_{i,8} + \beta_9 x_{i,9} + \beta_{10} x_{1,10} + \beta_{11} x_{i,11}$$

Equation: 1

3.2 Model Development

With the goal of predicting the Point Spread, we have chosen Team, Opponent, Third Quarter Point Spread, Average Field Goal Percent, Average Opponent Field Goal Percent, Average Three Point Percent, Average Opponent Three Point Percent, Average Free Throw Percent, Average Opponent Free Throw Percent, Average Total Rebounds, Average Opponent Total Rebounds, Average Turnovers, and Average Opponent Turnovers as our predictor variables. In the process of selecting the significant variables, we have tested different models with all variables and gradually remove each until drastic changes in model accuracy occur. Moreover, some variables in the raw data are contained in others, for example, offensive rebounds are included in total rebounds, and having it as a variable does not change the result. Therefore, we exclude all variables have either shown little relationship with the Point Spread, or are included in the existing variables. As a result, all the variables adopted in our model have shown significance, with the exception of Team and Opponent, which have shown to have the least relation with the Point Spread from the p-value. However, we decide to include them with the reasoning as follows. Consider the situation where two opponent teams have the exact performance history(ie. cumulative averages for every game being completely equal); though the probability is low, we would still wish to predict which team is more likely to be ahead for certain points. Therefore, having Team and Opponent as categorical variables will provide a prediction favorable to the winning Team.

That being said, we test out if the variables above can accurately predict the Point Spread, using just the variables given. The model is given in Equation(2) as follows:

$$y_i = \beta_0 + \sum_{i=1}^{58} \beta_i team_i + \beta_1 \text{ThirdQuarterPointSpread}_i + \beta_2 \text{FieldGoalPercent}_i + \beta_3 \text{OppnentFieldGoalPercent}_i$$

$$+\beta_4 \text{ThreePointPercent}_i + \beta_5 \text{OpponentFieldGoalPercent}_i + \beta_6 \text{FreeThrowPercent}_i + \beta_7 \text{OpponentFreeThrowPercent}_i$$

$$+\beta_8 \text{TotalRebounds}_i + \beta_9 \text{OpponentTotalRebounds}_i + \beta_{10} \text{Turnovers}_i + \beta_{11} \text{OpponentTurnovers}_i$$

Equation: 2

In Equation (2), all the predictor variables are from the result of each game, which is the information we would not know while predicting the Point Spread. It is shown in Table (4) that all variables besides the categorical variables have a p-value of less than 0.05, meaning they are significant. Though some of the categorical dummy variables have larger p-values, they still provide a general estimate of how the team performs. The summary of the model below(Table (4)) shows a residual standard error of 3.63, and a Multiple R-squared of 0.9304. The residual standard error error indicates that the actual point spread shows a 3.63 difference of the predicted point spread on average, and Multiple R-squared indicates that approximately 93% of the variance in Point Spread can be explained by the predictor variables. The model shown in equation (2) has been tested to be the most efficient reduction from all the variables in the data, meaning it has the least number of predictors while still producing reliable results. The results from this model indicates that if the cumulative sum of each predictor is close enough to the actual team performance in a game, then our final model in equation (1) would show similar result.

The next step is to use the cumulative averages of all the predictor variables, as shown in equation (1). If the cumulative average can provide an accurate estimation of a team's in game performance, then the model should have similar result as the prior model. Table (6) shows a summary of the final model. From Table (6), we can see that the adjusted R-squared changes to 0.71, which is significantly lower than the first model shown. This result is expected since the in game stats may differ from a team's average. 0.71 indicates that 71% of the variance can be explained using the model, though it is not as high as the previous model, it still gives reasonable results.

The model development is done in R(R Core Team [2020]), using corrr(Kuhn et al. [2020]) and pander(Daróczi and Tsegelskyi [2018]) for formatting tables. Plots for the final model are included in section 4, and we will further discuss the results and plots in section 5.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4174	4.284	0.09744	0.9224
TeamBOS	-0.6416	0.892	-0.7193	0.4722
TeamCHI	0.4082	0.9092	0.449	0.6536
TeamCLE	3.242	0.8946	3.624	0.0003059
TeamDAL	-0.6643	0.8855	-0.7502	0.4533
TeamDEN	-1.636	0.8864	-1.846	0.06521
TeamDET	0.907	0.8955	1.013	0.3114
TeamGSW	-0.1963	0.8937	-0.2196	0.8262
TeamHOU	2.011	0.9008	2.232	0.02584
TeamIND	-1.047	0.8948	-1.17	0.2423
TeamLAC	0.1276	0.9026	0.1414	0.8876
TeamLAL	-3.401	0.9018	-3.772	0.0001724
TeamMEM	-1.382	0.8744	-1.58	0.1144
TeamMIA	-1.944	0.8886	-2.188	0.02891
TeamMIL	-4.323	0.8822	-4.9	1.128e-06
TeamMIN	-1.699	0.89	-1.909	0.05656
TeamNOP	-1.135	0.8971	-1.265	0.2061
TeamNYK	-3.497	0.8921	-3.92	9.516e-05
TeamOKC	-0.599	0.9005	-0.6652	0.5061
TeamORL	-3.136	0.8859	-3.54	0.00042
TeamPHI	0.3093	0.8968	0.3449	0.7303
TeamPOR	1.604	0.9113	1.76	0.07868
TeamSAC	-2.473	0.9012	-2.744	0.006179
TeamSAS	1.625	0.8837	1.839	0.06621
TeamTOR	-0.4526	0.8935	-0.5066	0.6126
TeamUTA	-1.179	0.915	-1.288	0.198
TeamWAS	-4.021	0.9103	-4.417	1.12e-05
OpponentBOS	0.6269	0.8678	0.7223	0.4703
OpponentCHI	-0.3663	0.9092	-0.4029	0.6871
OpponentCLE	-3.141	0.8872	-3.541	0.0004191
OpponentDAL	0.8179	0.8922	0.9167	0.3595
OpponentDEN	1.804	0.8854	2.037	0.04189
OpponentDET	-0.8121	0.8882	-0.9143	0.3608
OpponentGSW	0.1464	0.8766	0.1671	0.8674
OpponentHOU	-1.728	0.8947	-1.932	0.05372
OpponentIND	1.141	0.8878	1.285	0.199
OpponentLAC	-0.1812	0.9025	-0.2008	0.8409
OpponentLAL	3.497	0.895	3.907	0.0001002
OpponentMEM	1.453	0.859	1.692	0.09102
OpponentMIA	1.992	0.8819	2.258	0.02415
OpponentMIL	4.374	0.8757	4.995	7.037e-07
OpponentMIN	1.817	0.8788	2.068	0.03895
OpponentNOP	1.261	0.8861	1.423	0.1549
OpponentNYK	3.622	0.8932	4.055	5.424e-05
OpponentOKC	0.7842	0.8886	0.8825	0.3777
OpponentORL	3.503	0.8982	3.9	0.000103
OpponentPHI	-0.1415	0.8862	-0.1597	0.8732
OpponentPOR	-1.293	0.9229	-1.401	0.1615
OpponentSAC	2.614	0.8963	2.916	0.003628
OpponentSAS	-1.482	0.8816	-1.681	0.09312
OpponentTOR	0.533	0.8865	0.6012	0.5479
OpponentUTA	1.259	0.9032	1.394	0.1637

	Estimate	Std. Error	t value	Pr(>|t|)
OpponentWAS	4.197	0.9098	4.613	4.524e-06
ThirdQuarterPointSpread	0.188	0.01556	12.08	2.679e-31
FieldGoalPercent	83.5	3.399	24.56	5.216e-103
OppFieldGoalPercent	-83.75	3.398	-24.65	1.482e-103
ThreePointPercent	16.9	1.394	12.13	1.63e-31
OppThreePointPercent	-17.08	1.391	-12.28	3.123e-32
FreeThrowPercent	9.908	1.191	8.321	3.1e-16
OppFreeThrowPercent	-10.08	1.191	-8.463	1.006e-16
TotalRebounds	0.5019	0.02724	18.43	7.609e-65
OppTotalRebound	-0.5081	0.02722	-18.67	2.891e-66
Turnovers	-1.002	0.03873	-25.86	1.848e-111
OppTurnover	1.006	0.03859	26.07	8.003e-113

Table 4: Test Ideal Model Fit Summary

Observations	Residual Std. Error	R^2	Adjusted R^2
991	3.672	0.929	0.9242

	Estimate	Std. Error	t value	Pr(>|t|)
TeamATL	-47.74	136.8	-0.349	0.7271
TeamBOS	-51.75	138.2	-0.3743	0.7082
TeamCHI	-52.25	136.9	-0.3816	0.7029
TeamCLE	-48.2	136.2	-0.3539	0.7235
TeamDAL	-51.75	138.5	-0.3737	0.7087
TeamDEN	-52.5	137.8	-0.3808	0.7034
TeamDET	-51.43	137.7	-0.3736	0.7088
TeamGSW	-44.9	140.1	-0.3204	0.7487
TeamHOU	-44.42	138	-0.3218	0.7476
TeamIND	-51.39	135.6	-0.379	0.7048
TeamLAC	-45.45	136.1	-0.3339	0.7385
TeamLAL	-57.09	137.5	-0.4152	0.6781
TeamMEM	-50.5	136.3	-0.3706	0.711
TeamMIA	-51.79	133.4	-0.3883	0.6979
TeamMIL	-47.29	137.1	-0.3448	0.7303
TeamMIN	-56.77	138.5	-0.4099	0.682
TeamNOP	-51.16	136.6	-0.3744	0.7082
TeamNYK	-59.03	134.9	-0.4376	0.6618
TeamOKC	-49.44	137.4	-0.3599	0.719
TeamORL	-50.72	137.5	-0.3689	0.7123
TeamPHI	-49.72	137.7	-0.3612	0.7181
TeamPOR	-53.4	138.6	-0.3851	0.7002
TeamSAC	-53.09	137.4	-0.3864	0.6993
TeamSAS	-50.36	137.5	-0.3663	0.7142
TeamTOR	-50.84	136	-0.3737	0.7087
TeamUTA	-52.12	134.5	-0.3875	0.6985
TeamWAS	-49.46	136.4	-0.3626	0.717
OpponentBOS	-0.5722	1.671	-0.3425	0.7321
OpponentCHI	-0.489	1.732	-0.2823	0.7778
OpponentCLE	-1.073	1.699	-0.6317	0.5277
OpponentDAL	0.6591	1.728	0.3814	0.703
OpponentDEN	1.938	1.702	1.138	0.2552
OpponentDET	0.4323	1.703	0.2538	0.7997
OpponentGSW	-0.1234	1.691	-0.07297	0.9418
OpponentHOU	1.093	1.718	0.6363	0.5248
OpponentIND	-0.4567	1.708	-0.2673	0.7893
OpponentLAC	-1.923	1.728	-1.113	0.2659
OpponentLAL	3.528	1.704	2.07	0.03872
OpponentMEM	-0.3579	1.658	-0.2159	0.8291
OpponentMIA	2.219	1.705	1.302	0.1933
OpponentMIL	2.579	1.689	1.527	0.1271
OpponentMIN	2.494	1.698	1.469	0.1422
OpponentNOP	0.6249	1.687	0.3704	0.7112
OpponentNYK	3.145	1.724	1.824	0.06849
OpponentOKC	0.6661	1.685	0.3954	0.6927
OpponentORL	0.7223	1.731	0.4173	0.6765
OpponentPHI	2.458	1.684	1.46	0.1448
OpponentPOR	1.192	1.739	0.6855	0.4932
OpponentSAC	3.562	1.716	2.077	0.03811
OpponentSAS	-1.058	1.702	-0.6219	0.5342
OpponentTOR	0.2737	1.72	0.1592	0.8736
OpponentUTA	0.7378	1.704	0.433	0.6651

	Estimate	Std. Error	t value	Pr(>|t|)
OpponentWAS	3.776	1.746	2.163	0.03083
ThirdQuarterPointSpread	0.7864	0.02008	39.17	9.475e-199
AvgFieldGoalPercent	-62.86	119.1	-0.5277	0.5978
AvgOppFieldGoalPercent	3.94	123.9	0.03179	0.9746
AvgThreePointPercent	-24.53	53.2	-0.461	0.6449
AvgOppThreePointPercent	81.23	55.33	1.468	0.1424
AvgFreeThrowPercent	51.34	43.9	1.169	0.2425
AvgOppFreeThrowPercent	37.62	50.82	0.7403	0.4593
AvgTotalRebounds	-0.2162	0.8383	-0.2579	0.7965
AvgOppTotalRebounds	0.4206	0.9274	0.4535	0.6503
AvgTurnovers	-0.4714	1.325	-0.3556	0.7222
AvgOppTurnovers	-0.8854	1.278	-0.6928	0.4886

Table 6: Real Model Fit Summary

Observations	Residual Std. Error	R^2	Adjusted R^2
991	7.139	0.7318	0.7133

4 Result

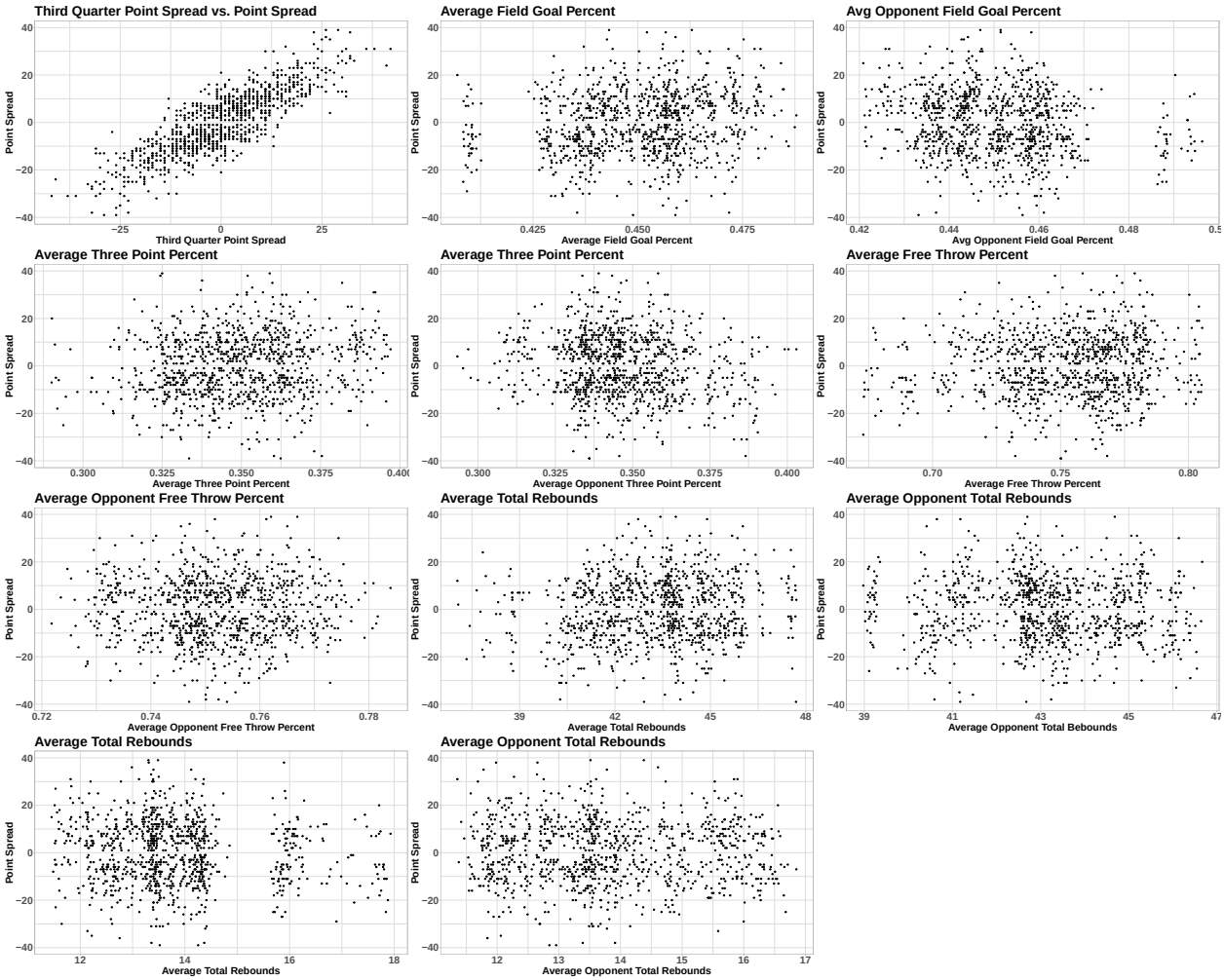

Figure 3: Check Linearality for each Predictor Variable

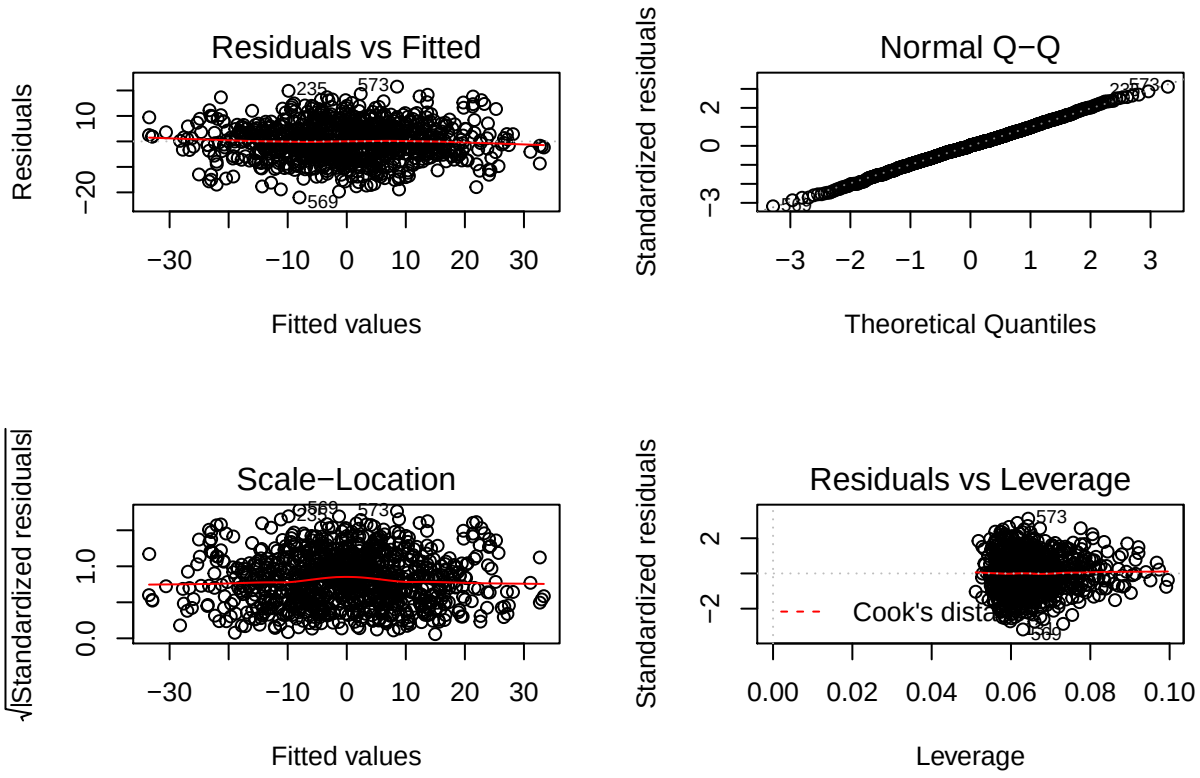

Figure 4: Diagnostic Plots for Regression Model

Table 7: Prediction of Point Spread with 10 randomly sampled games

Game	Team	Opponent	PointSpread	Predict.fit	Predict.lwr	Predict.upr
60	TOR	POR	2	14.399662	-1.626112	30.425437
68	SAS	POR	8	10.506143	-6.378235	27.390520
57	NYK	MIN	8	16.200200	-1.160206	33.560605
56	NOP	WAS	-20	-6.702546	-23.369003	9.963911
61	MIL	IND	-5	-6.063798	-22.214306	10.086711
59	CLE	IND	4	-0.323232	-15.104581	14.458117
66	ATL	MEM	12	12.724185	-3.243088	28.691457
60	MIL	HOU	7	3.605139	-12.550003	19.760282
58	WAS	CLE	14	21.029673	3.641135	38.418211
64	POR	DET	-20	-11.336417	-27.867806	5.194972

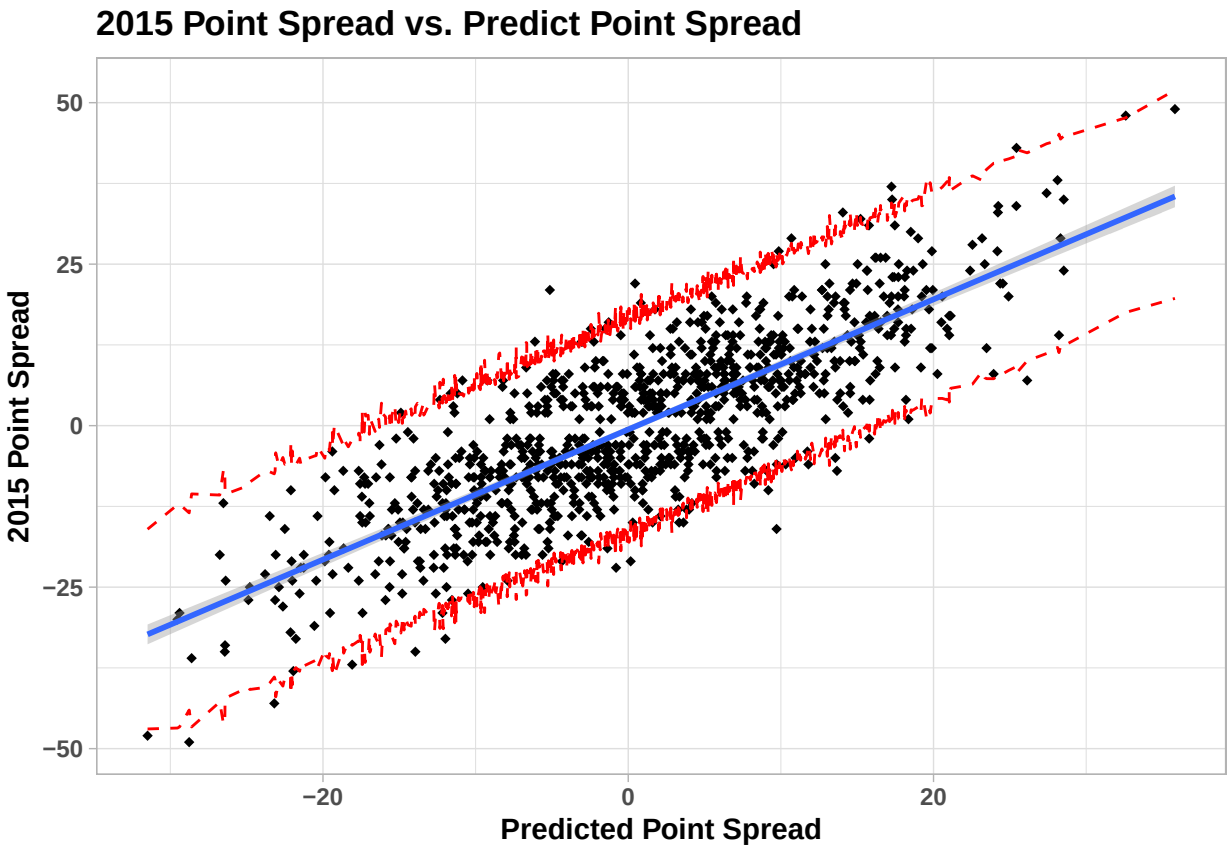

Figure 5: Season 2015 Predicted Point Spread compared to Actual Point Spread

5 Discussion

The section above demonstrates the results of our model, including diagnostics using season 2014 data, and some predictions using 2015 data. The figure indicates that our model failed to sufficiently predict the Point Spread, however, it could potentially be used to predict the Money Line result.

Figure 3 demonstrates the relationship of each independent variable and the Point Spread. From the graph, Third Quarter Point Spread appears to have the strongest linear relation with the final point spread with very few outliners. It explains the reason of Third Quarter Point Spread having the smallest p-value, rejecting the hypothesis that it is not significant. All other predictors tends to be linear as well, however, with higher variance. But we can use the method of least square to approximate a straight line through the points.

In addition, Figure 4 contains 4 diagnostic plots, in which we can use to analyze the model fit. First of all, the Residuals vs Fitted also shows the linear relationship between the independent variables and the dependent variables in a general sense, comparing to showing relationships of each variable separately in 3. It is demonstrated that the residuals follow a linear tendency as the red line appears to be horizontal; however, the variance spreads within a 40 points interval, which may the difference between the prediction and the actual point spread. Secondly, the Normal Q-Q plot from Figure4 successfully indicates that the residuals are normally distributed, which is what we are looking for in a model, satisfying the second assumption for applying linear regression. Scale-Location plot demonstrates that the residuals are evenly spread along the range of predictors, and the spread seems to be evenly distributed along the fitted value, which is also a good indicator. Moreover, the Residuals vs Leverage plot shows if there are influential cases such as outlines that will impact the accuracy of our model significantly. From the plot, all residuals seem to be clustered on the right, and it does not seem to have any influential cases that will impact the result drastically. Figure 4 suggests that our model is efficient to explain the relation of the independent and dependent variables.

For model verification, we have decided to use real NBA data from Season 2015, hoping that the model will be able to predict the Point Spread. We have plotted the predicted point spread against the actual point spread in figure 5. An ideal plot would have an approximate slope of 1, in which each predicted value perfectly corresponds to the actual Point Spread in 2015. However, the blue regression line shows an approximated slope of 1.4, meaning that we failed to accurately predict some of the Point Spread outcomes in 2015. Moreover, the dotted red lines are the upper and lower prediction interval, which reflects the uncertainly of each predicted value, whereas the gray strip indicates the confidence interval. It can be shown by the graph that our model is not accurate enough to predict the point spread in each game.

The results from figure 5 may seem a bit discouraging as it is not sufficient enough to predict the point spread, especially in betting when we are against the betting-line set by the sports book. However, the graph does indicate the majority of the positive predicted point spreads map to actual positive point spreads, negative point spreads map to negatives. Therefore, our model still holds its value for betting on the Money Line to predict the winning team. We took this idea further, and drew 10 random games from Season 2015. With the data we already have, Table 7 shows the prediction for each game's team spread, as well as the prediction interval. We only accurately predicted 1 of the point spread - game 66 of ATL vs. MEM. The prediction shows a 12.72 point Point Spread with ATL in favor, which is exactly the Point Spread of that game. Regardless, we can see that the model successfully predicted 9 out of 10 games for points being positively or negatively spread. A positive Point Spread indicates the Home Team winning, and from table 7, the only game we have wrongly predicted is the 59th game of CLE vs. IND. The prediction indicates that IND would win by approximately 0.3 points, however, CLE won by 4 percent. In conclusion, although our model is not sufficient enough to predict the Point Spread, it shows a high accuracy of predicting the winning team which could be used for betting the Money Line.

Our result have some potential weaknesses and ts limitations due to the unpredictable nature of basketball, which cannot be generalized by just the team performance. More specifically, each player has significant impact on the game, which are not accounted for in the model. For example, a team's best player being injured will not only drastically impact the game results, but also may impact the rest of the seasonal performance of that team. The players are not constant in each game, which should also be accounted for as it might cause the performance to differ from the average. Moreover, some of the predictor variables from our model are not independent, which results in potential multicollinearlity. A linear model assumes all

independent variables are independent from each other, in basketball, it is almost impossible to achieve as each aspects are closly related to each other. However, that being said, our model can be further improved in future studies. Figure 3 has shown that the most influencial perdictor is the point spread in the third quarter of that game, which is before bets close. Therefore, we could potentially find or create a dataset with in-game information of the third quarter, and use it to build a model, which may result in significantly higher accuracy.

6 Appendix

[1] github repo: "https://github.com/mackenziequ/NBA-prediction"

Reference

Basketball Statistics and History, 2020. URL https://www.basketball-reference.com/.

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2020. URL https://github.com/rstudio/rmarkdown. R package version 2.5.

Gergely Daróczi and Roman Tsegelskyi. *pander: An R 'Pandoc' Writer*, 2018. URL https://CRAN.R-project.org/package=pander. R package version 0.6.3.

Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2020. URL https://CRAN.R-project.org/package=ggpubr. R package version 0.4.0.

Ionas Kelepouris. *NBA Team Game Stats from 2014 to 2018*, 2018. URL https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018/metadata. Stats from every game during 2014 - 2018 nba period.

Max Kuhn, Simon Jackson, and Jorge Cimentada. *corrr: Correlations in R*, 2020. URL https://CRAN.R-project.org/package=corrr. R package version 0.4.3.

Marc J. Mazerolle. *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*, 2020. URL https://cran.r-project.org/package=AICcmodavg. R package version 2.3-1.

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

rLoper. *1996-2019 NBA Stats Complete With Player Stats*, 2020. URL https://www.kaggle.com/rloper/1996 2018-nba-stats-complete-with-player-stats/metadata. Includes every game since 1996/97 season, including the stats from top 5 players.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham. *modelr: Modelling Functions that Work with the Pipe*, 2020a. URL https://CRAN.R-project.org/package=modelr. R package version 0.1.8.

Hadley Wickham. *tidyr: Tidy Messy Data*, 2020b. https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.

Hadley Wickham and Evan Miller. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, 2020. http://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2020. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. URL http://www.crcpress.com/product/isbn/9781466561595. ISBN 978-1466561595.

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL https://yihui.org/knitr/. ISBN 978-1498716963.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2020. URL https://yihui.org/knitr/. R package version 1.30.

Yihui Xie, J.J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida, 2018. URL https://bookdown.org/yihui/rmarkdown. ISBN 9781138359338.

Yihui Xie, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook*. Chapman and Hall/CRC, Boca Raton, Florida, 2020. URL https://bookdown.org/yihui/rmarkdown-cookbook. ISBN 9780367563837.

Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2020. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.