A Regression-based Comparison of NBA Players' Height and Gameplay Performance Between the 1996 and 2021 Seasons.

Lucas Lyons

December 2022

Introduction

Gameplay tendencies in the NBA have changed over the past several decades. In more recent years many teams have adopted a “small-ball” style of play characterized by increased three-point attempts and positional fluidity (Mandić, 2019). Existing studies characterize broad trends in NBA style of play (da Silva, 2021), but to the author’s knowledge, none have recently examined the relationship between NBA players’ heights and their performance on the court over time. This study aims to do exactly that, using linear regression to predict player height as a function of player statistics across five seasons of NBA data.

Methods

Data Set

To answer the research question, a data set titled “NBA Players” was located on Kaggle containing 11981 points, one for each player who has played in the NBA between the 1996-97 season and the 2020-2021 season. For each player, the data set contained biographical information such as age and height, and season-long averages of statistics such as points per game and rebounds per game. The data set was chosen because of the relevance of its contents to the research question. The variables average assist percentage, average defensive rebound percentage, average offensive rebound percentage, average usage percentage, and average points per game were chosen as predictors of player height (cm) to reflect the various roles players are capable of performing in NBA games.

Preparation

First, the chosen predictor variables were plotted with histograms and then plotted against the predictor variable to verify the data set’s suitability for regression analysis. Then, players with less than 10 games played in a season were removed from the data set to exclude potential cases of players with extreme statistics due to low sample size. The complete data set containing all 25 seasons from 1996-2021 was subsetted to contain only the 1996-97, 2002-03, 2008-09, 2014-15, and 2020-2021 seasons to allow for analysis on multiple time periods but reduce the number of models requiring validation. The data was then further subsetted into training (70%) and test (30%) data for each season.

Model Assessment Framework

This section will outline the framework used in this study to assess regression models. First, fitted values for the models were plotted against the response variable to assess condition 1 for linear regression. Then pairwise predictor plots were generated and inspected for nonlinear relations to assess condition 2. VIF values were calculated to diagnose multicollinearity. Residual QQ plots were generated and inspected for violations of normality, and fitted vs. residual plots were inspected for violations of homeoskedasticity, linearity, and correlation of residuals. Cutoffs were generated to identify leverage points, outliers, and influential points (Cook’s, DFFITS, DFBETAS). Leverage/outlier/influential points were inspected and the impact they were having on the model was recorded.

Regression Analysis

Regression models were fit to each training data set with the specified predictor variables and player height as the response. Each model was assessed following the model validation framework outlined in the preceding section of this report. Right skew was observed to affect these models so new models with square root transformed models were fit and assessed. The transformed and untransformed models were compared by several metrics including AIC, maximum VIF value, Rˆ2 value, and number of influential points, and the transformed models were seen to perform slightly better. Transformed usage percentage and points were not significant predictors for every season, so a partial ANOVA F-test (p = 0.05) was conducted to determine if these variables could be dropped, but the null hypothesis was rejected so they were kept. The final models were selected and fit to the test data and assessed according to the framework. The test and training models were then directly compared by several metrics, including a confidence interval test, to validate the final model. Lastly, the final models were evaluated by their Rˆ2 values, predictor coefficients, and predictor significance levels over time. These values were put into context to formulate an answer for the research question.

Results

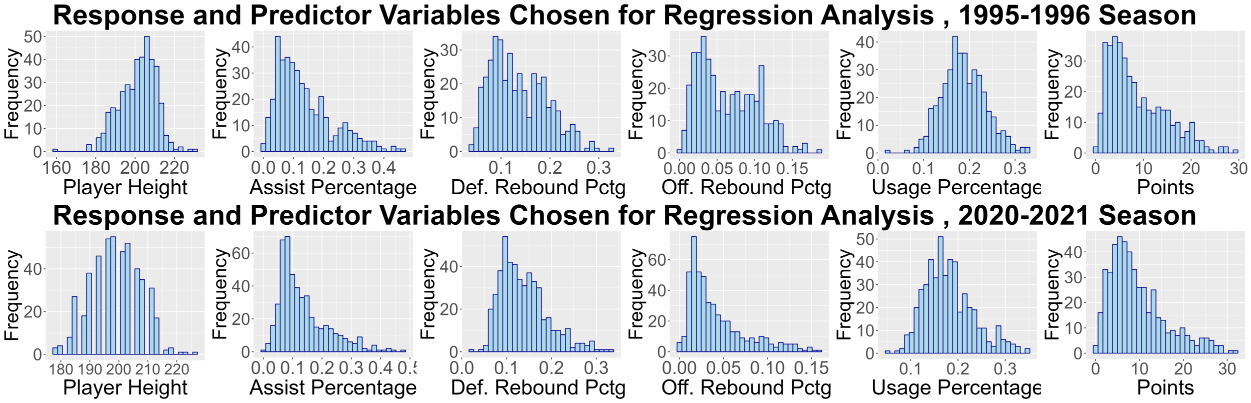


Figure 1: Frequency histograms of the analysis variables for the 1995-96 and 2020-21 seasons (complete table in Appendix A).

Analysis Process and Model Selection

Preliminary investigation of the variables chosen for analysis showed that for each season, the distributions for player height and usage percentage somewhat resembled the normal distribution while the distributions for assist percentage, defensive and offensive rebound percentage, and points had pronounced right skew (see figure 1). When the models were fit to the data, the results varied by season but assists and offensive/defensive rebound percentage were highly significant predictors ($p < 0.001$) for almost every season, but the significance of points and usage percentage varied by season. No obvious violations of condition 1 for linear regression were detected in the models, however "L-shape" patterns were observed in pairwise predictor plots, indicating condition 2 was possibly not met. When the models were inspected for evidence of violations of normality, constant variance, and non-correlation of residuals, no violations were found except for slight curvature in the fitted against residuals plots indicating minor violations of linearity. No outliers or influential points as measured by Cook's distance were identified for any model, however all models had points identified as influential by DFFITS and DFBETAS. None of the influential points were found to be entry errors. The models with the transformed predictors were fit and found to reduce violations of condition 2 compared to the original models (see appendix B).

The transformed models were assessed to perform as well or better than the untransformed models based on inspection of QQ plots and fitted vs residual plots and were selected for subsequent analyses after a comparison of multiple metrics which is displayed partially in Figure 3.

The usage and points variables were kept in the models following ANOVA partial F-tests on each model, hence the full transformed models were selected as the final models. They were fit to the test data set. Different predictors were found to be significant in the training and test models, and 10 out of 24 of the estimated coefficients for the test models were outside the standard errors for the estimated training model slopes. Additionally, the corresponding Rˆ2 values for certain training and test models differed substantially. Based on this information, the models failed to validate. Appendix C demonstrates a comparison of the final models fit to the training and test data.

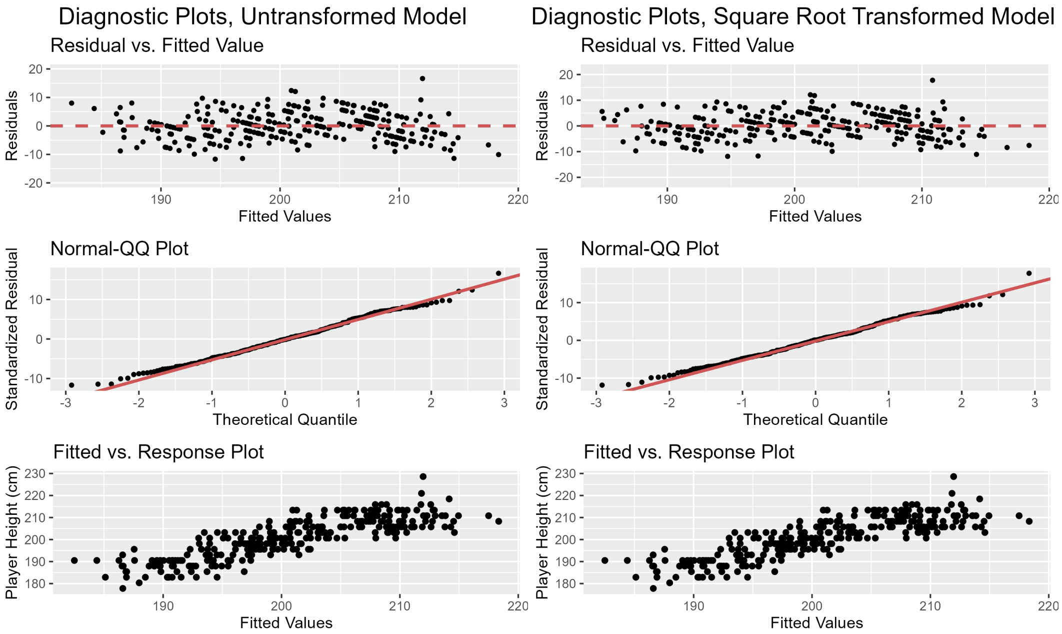
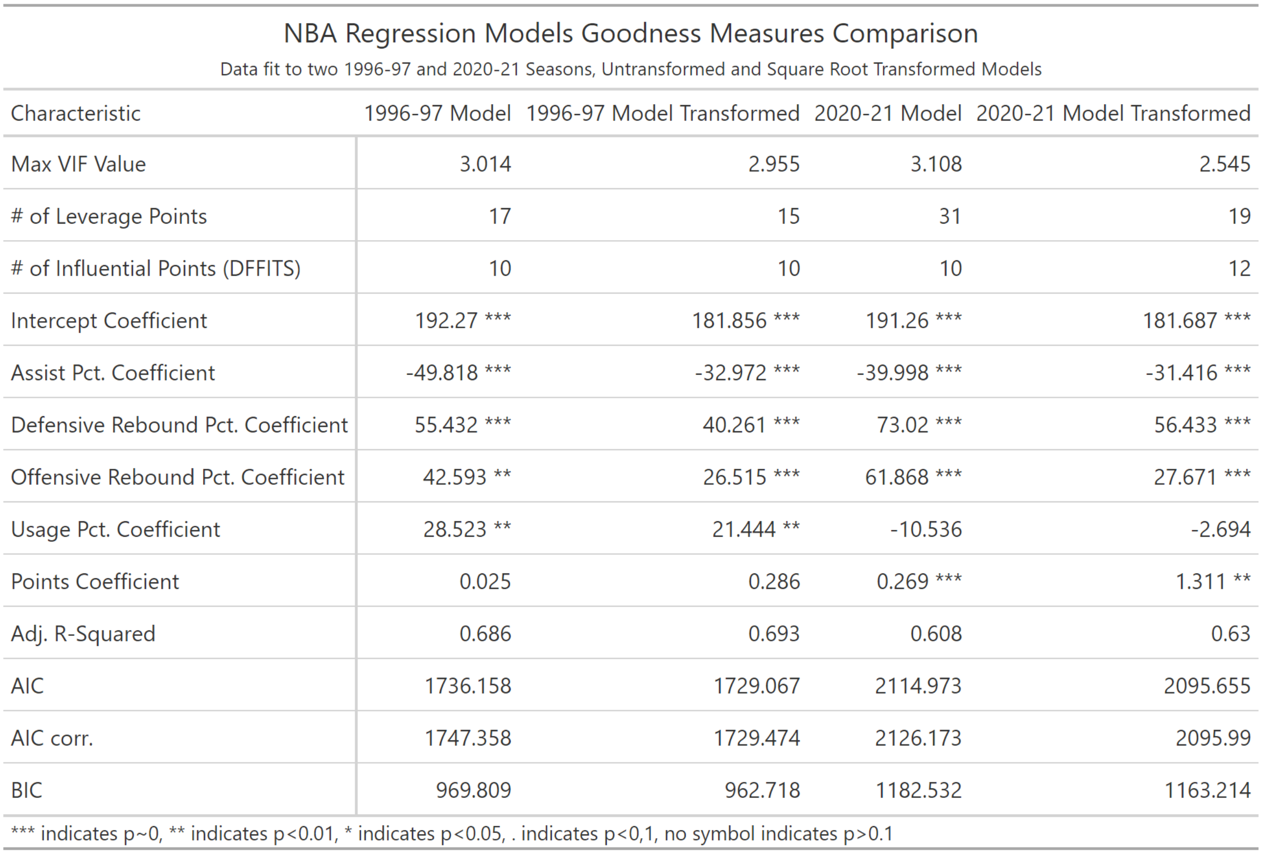


Figure 2: Comparison of Diagnostic Plots for Transformed and Untransformed Models Fit to 2008-09 NBA Season Data.



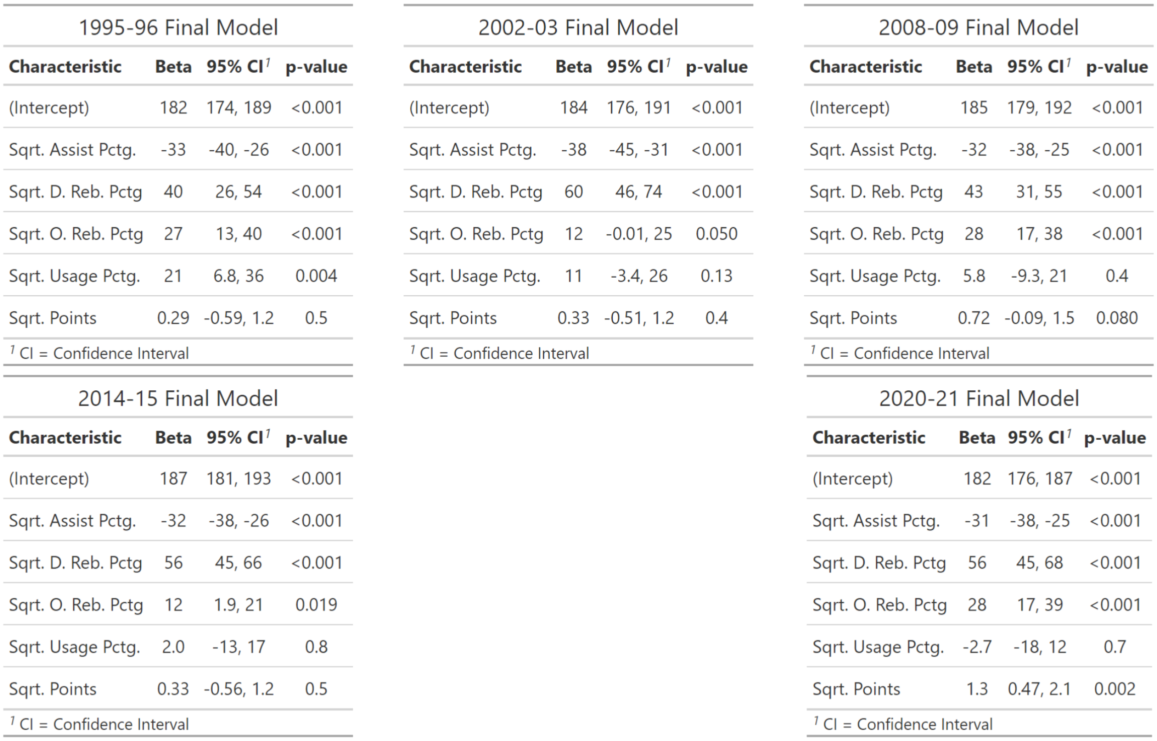
NBA Regression Models Goodness Measures Comparison

Data fit to two 1996-97 and 2020-21 Seasons, Untransformed and Square Root Transformed Models

Characteristic	1996-97 Model	1996-97 Model Transformed	2020-21 Model	2020-21 Model Transformed
Max VIF Value	3.014	2.955	3.108	2.545
# of Leverage Points	17	15	31	19
# of Influential Points (DFFITS)	10	10	10	12
Intercept Coefficient	192.27 ***	181.856 ***	191.26 ***	181.687 ***
Assist Pct. Coefficient	-49.818 ***	-32.972 ***	-39.998 ***	-31.416 ***
Defensive Rebound Pct. Coefficient	55.432 ***	40.261 ***	73.02 ***	56.433 ***
Offensive Rebound Pct. Coefficient	42.593 **	26.515 ***	61.868 ***	27.671 ***
Usage Pct. Coefficient	28.523 **	21.444 **	-10.536	-2.694
Points Coefficient	0.025	0.286	0.269 ***	1.311 **
Adj. R-Squared	0.686	0.693	0.608	0.63
AIC	1736.158	1729.067	2114.973	2095.655
AIC corr.	1747.358	1729.474	2126.173	2095.99
BIC	969.809	962.718	1182.532	1163.214

*** indicates p~0, ** indicates p<0.01, * indicates p<0.05, . indicates p<0,1, no symbol indicates p>0.1

Figure 3: Partial summary of model goodness measures for transformed and square root transformed models fit to the 1996-7 and 2020-2021 NBA seasons.



1995-96 Final Model

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	182	174, 189	<0.001
Sqrt. Assist Pctg.	-33	-40, -26	<0.001
Sqrt. D. Reb. Pctg	40	26, 54	<0.001
Sqrt. O. Reb. Pctg	27	13, 40	<0.001
Sqrt. Usage Pctg.	21	6.8, 36	0.004
Sqrt. Points	0.29	-0.59, 1.2	0.5

[1] CI = Confidence Interval

2002-03 Final Model

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	184	176, 191	<0.001
Sqrt. Assist Pctg.	-38	-45, -31	<0.001
Sqrt. D. Reb. Pctg	60	46, 74	<0.001
Sqrt. O. Reb. Pctg	12	-0.01, 25	0.050
Sqrt. Usage Pctg.	11	-3.4, 26	0.13
Sqrt. Points	0.33	-0.51, 1.2	0.4

[1] CI = Confidence Interval

2008-09 Final Model

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	185	179, 192	<0.001
Sqrt. Assist Pctg.	-32	-38, -25	<0.001
Sqrt. D. Reb. Pctg	43	31, 55	<0.001
Sqrt. O. Reb. Pctg	28	17, 38	<0.001
Sqrt. Usage Pctg.	5.8	-9.3, 21	0.4
Sqrt. Points	0.72	-0.09, 1.5	0.080

[1] CI = Confidence Interval

2014-15 Final Model

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	187	181, 193	<0.001
Sqrt. Assist Pctg.	-32	-38, -26	<0.001
Sqrt. D. Reb. Pctg	56	45, 66	<0.001
Sqrt. O. Reb. Pctg	12	1.9, 21	0.019
Sqrt. Usage Pctg.	2.0	-13, 17	0.8
Sqrt. Points	0.33	-0.56, 1.2	0.5

[1] CI = Confidence Interval

2020-21 Final Model

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	182	176, 187	<0.001
Sqrt. Assist Pctg.	-31	-38, -25	<0.001
Sqrt. D. Reb. Pctg	56	45, 68	<0.001
Sqrt. O. Reb. Pctg	28	17, 39	<0.001
Sqrt. Usage Pctg.	-2.7	-18, 12	0.7
Sqrt. Points	1.3	0.47, 2.1	0.002

[1] CI = Confidence Interval

Figure 4: Summary of final models chosen for analysis.

Key Findings

Transformed assist percentage had a highly significant negative ($p \sim 0$) relationship with player height for all five seasons. Transformed offensive and defensive rebound percentages had a significant ($p < 0.1$) positive relationship with player height. Transformed points and usage percentage were not always significant predictors of player height, but for the 2020-2021 season points was significant ($p < 0.05$). Additionally, the adjusted Rˆ2 values for the models remained consistent with the exception of the model for the 2020-2021 season.

Discussion

Trends in Player Roles

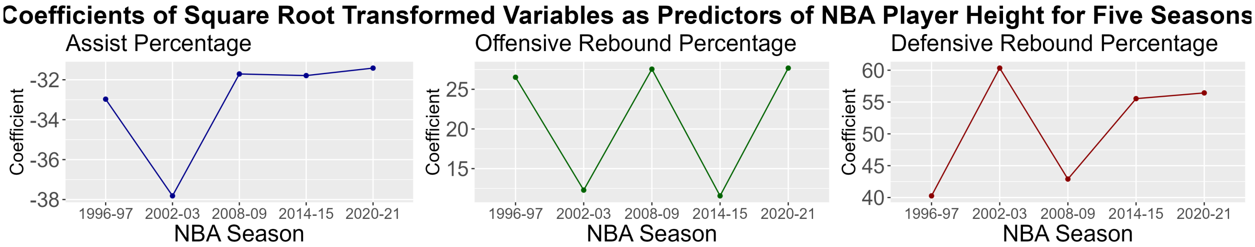


Figure 5: Estimated coefficient values for three predictor variables from the final regression models.

The most robust finding from this analysis is that a strong and negative linear relationship was observed between player height and the square root of assist percentage across all seasons. The estimated coefficient of this relationship was observed to be near -30 for each season except 2002-03. Thus, holding fixed the other predictors, a one percent increase of square rooted assist percentage predicted roughly a 30 centimeter decrease in a player's height for most seasons. That the R^2 values for the models remained relatively constant over time indicates that these predictors remained effective at predicting variation in player height across seasons. All of these findings are weakened by inconsistencies between the training and test models. The existence of negative relationships between height and assist percentage is not surprising; point guards pass the most and are generally the shortest players by position. However, the expectation was that this relationship would become less negative over time as teams moved to small-ball. This data suggests that in the NBA, taller players are still passing the ball less and still performing the vital role of rebounding more. A possible reason for this could be that that not all teams currently play a small-ball style of offense, so players may still be assigned to more traditional height-based roles. Additionally, this data set does not contain information on three-point shooting. A distinct possibility is that a relationship exists between player height and three-point shooting which is not captured in this study.

A Brief Discourse on Influential Points

The impact of influential points on the quality of the study appeared limited at first, however they may have played a role in the failure of the final model to validate. In particular certain estimated coefficients of the test models differed largely from the corresponding training model coefficients, and influential points impacting coefficient values were identified in both data sets. Even so, influential points were a feature of the data reflecting the diversity of players' ability in the NBA. Influential points generally fell in two categories of players; specialists who were very strong in certain areas (or just tall) but weak in others, or superstars who were well above average in many metrics.

Limitations

The failure to validate the models indicates that the results of the study need to be interpreted with caution. It is possible the models were overfitted. The usage and points predictors were kept due to the results of the ANOVA partial F-tests conducted, but they were precisely the predictors whose significance varied from the training to test models. As previously mentioned, lack of data on three-point shooting lowered the utility of this study. Finally, that only five seasons were analyzed means there are gaps in the study. A more complete study would analyze every season available to better clarify how trends have changed over time.

Citations

Kaggle Dataset: https://www.kaggle.com/datasets/justinas/nba-players-data/code

Mandić R, Jakovljević S, Erčulj F, Štrumbelj E (2019) Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017. PLoS ONE 14(10): e0223524. https://doi.org/10.1371/journal.pone.0223524

da Silva, R., Vítor, J., & Rodrigues, Canas, P. (2021). The Three Eras of the NBA Regular Seasons: Historical Trend and Success Factors. Journal of Sports Analytics, 7(4), 263-275. https://content.iospress.com/articles/journal-of-sports-analytics/jsa200525

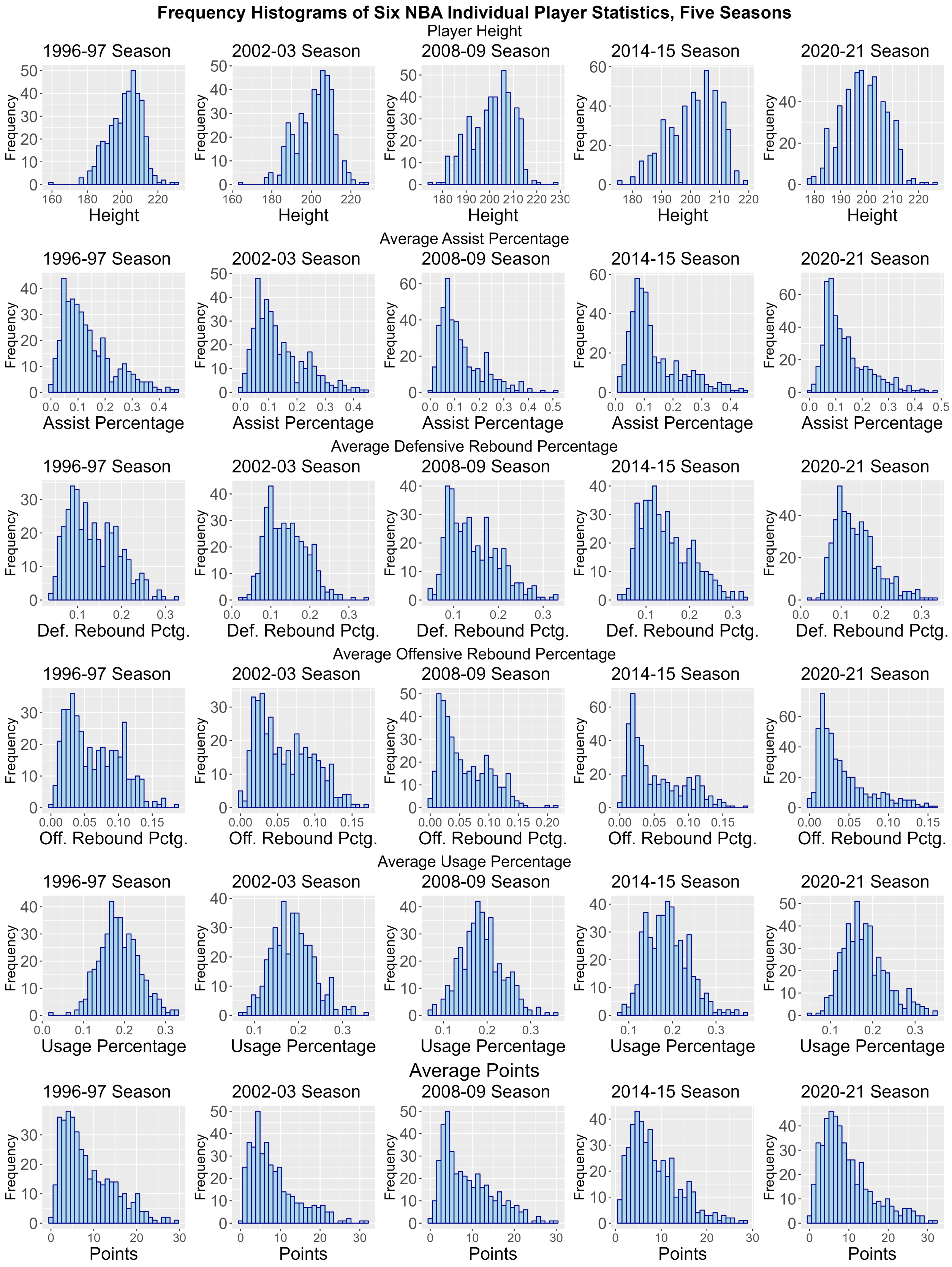


Figure 6: Appendix A: Complete graphical summary of all predictor variables used in regression analysis, for all seasons used in regression analysis.

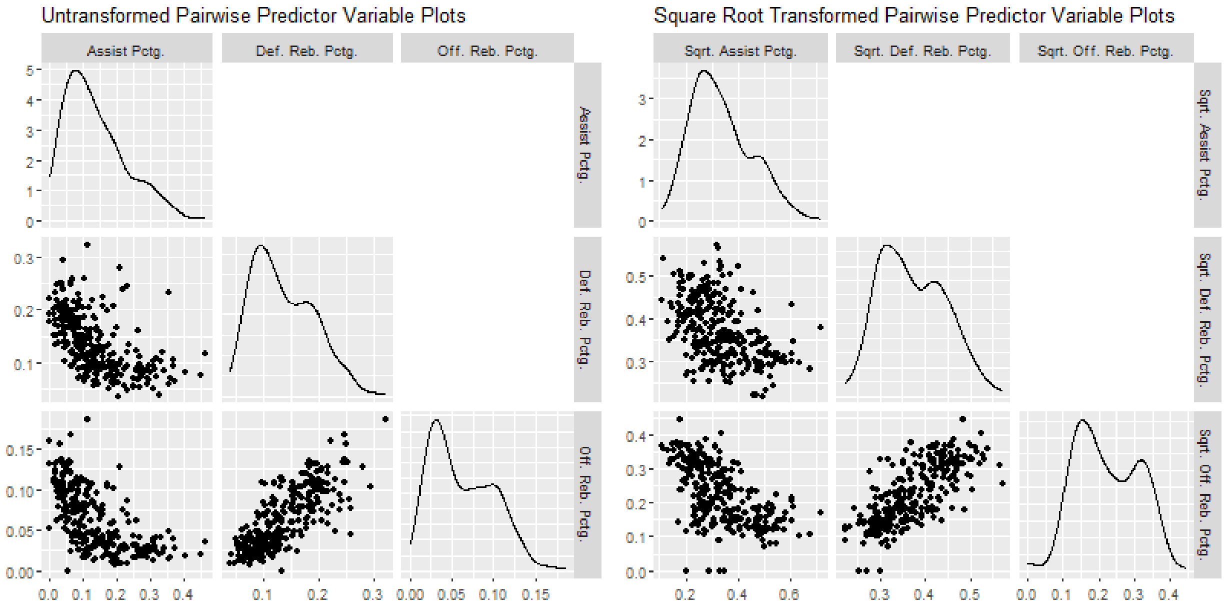
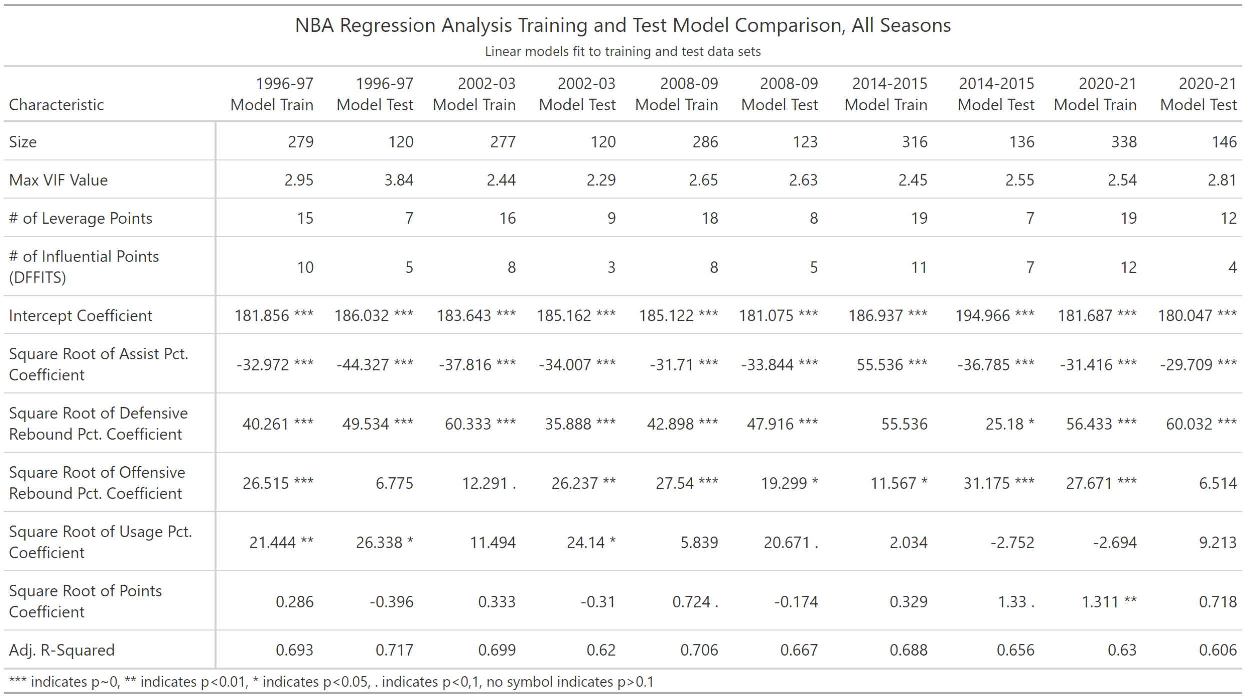


Figure 7: Appendix B: Pairwise predictor plots for untransformed and transformed models, 2008-09 Season.



NBA Regression Analysis Training and Test Model Comparison, All Seasons

Linear models fit to training and test data sets

Characteristic	1996-97 Model Train	1996-97 Model Test	2002-03 Model Train	2002-03 Model Test	2008-09 Model Train	2008-09 Model Test	2014-2015 Model Train	2014-2015 Model Test	2020-21 Model Train	2020-21 Model Test
Size	279	120	277	120	286	123	316	136	338	146
Max VIF Value	2.95	3.84	2.44	2.29	2.65	2.63	2.45	2.55	2.54	2.81
# of Leverage Points	15	7	16	9	18	8	19	7	19	12
# of Influential Points (DFFITS)	10	5	8	3	8	5	11	7	12	4
Intercept Coefficient	181.856 ***	186.032 ***	183.643 ***	185.162 ***	185.122 ***	181.075 ***	186.937 ***	194.966 ***	181.687 ***	180.047 ***
Square Root of Assist Pct. Coefficient	-32.972 ***	-44.327 ***	-37.816 ***	-34.007 ***	-31.71 ***	-33.844 ***	55.536 ***	-36.785 ***	-31.416 ***	-29.709 ***
Square Root of Defensive Rebound Pct. Coefficient	40.261 ***	49.534 ***	60.333 ***	35.888 ***	42.898 ***	47.916 ***	55.536	25.18 *	56.433 ***	60.032 ***
Square Root of Offensive Rebound Pct. Coefficient	26.515 ***	6.775	12.291 .	26.237 **	27.54 ***	19.299 *	11.567 *	31.175 ***	27.671 ***	6.514
Square Root of Usage Pct. Coefficient	21.444 **	26.338 *	11.494	24.14 *	5.839	20.671 .	2.034	-2.752	-2.694	9.213
Square Root of Points Coefficient	0.286	-0.396	0.333	-0.31	0.724 .	-0.174	0.329	1.33 .	1.311 **	0.718
Adj. R-Squared	0.693	0.717	0.699	0.62	0.706	0.667	0.688	0.656	0.63	0.606

*** indicates p~0, ** indicates p<0.01, * indicates p<0.05, . indicates p<0,1, no symbol indicates p>0.1

Figure 8: Appendix C: Complete validation comparison table of final model fit to training data and test data.