PREDICTION OF AVERAGE POINTS PER GAME: MULTIPLE LINEAR REGRESSION

Cesar Rene Pabon Bernal
Professor Chu Pan
Stat 706
City University of New York, Hunter College
Thursday, December 12th, 2019

TABLE OF CONTENTS

INTRODUCTION

For decades, many sports have been recording data about players and their performance. In the analytics age, even more fine-grain data are being recorded. With a focus on basketball, the purpose of this research is to see which metrics are related to a player's average number of points scored in a game.

DATA DESCRIPTION

Data from the 2018-2019 National Basketball Association (NBA) season and average player statistics per game was analyzed—it consisted of 30 variables and 629 observations.[1]

Upon evaluation, we propose:
Which variables best explain average number of points scored by an individual NBA player per game?

For the purpose of this class, two methodologies will be implemented in the journey of answering this question; multiple linear regression (MLR) and simple linear regression (SLR).

STATISTICAL METHODS

I. MULTIPLE LINEAR REGRESSION

A. Variable Selection

For the purpose of this report, the four qualitative variables were removed and the remaining 26 variables were explored. A correlation matrix of these variables is represented in **Figure 1.** Detailed exploration based on research[2] and expert analysis (from co-author and former professional basketball player Johnny Mathis), the MLR model was reduced to 14 predictors (labeled simple MLR model) and tabulated below. To determine which variables best explain the average number of points scored by an NBA player per game, a process of model selection was conducted. The Simple MLR model can be expressed as:

$$Y_{14} = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{13} X_{13} + \beta_{14} X_{14} + \epsilon_i \qquad \textbf{Eq. 1}$$

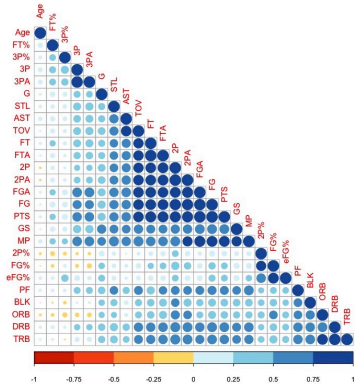


Figure 1 Correlation matrix of 26 variables

SIMPLE MLR MODEL	
Predictors	**Response (Outcome)**
1. Age 2. Game Played (G) 3. Games Started (GS) 4. Minutes Played (MP) 5. Three Point % (ThreePoint_Perc) 6. Two Point % (TwoPoint_Perc) 7. Free Throw % (FT_Perc) 8. Offensive Rebounds (ORB) 9. Defensive Rebounds (DRB) 10. Assists (AST) 11. Steals (STL 12. Blocks (BLK) 13. Turnovers (TOV) 14. Personal Fouls (PF)	1. Points (PTS)

Table 1 Simple MLR model 14 covariates

```
Call:
lm(formula = PTS ~ Age + G + GS + MP + ThreePoint_Perc + TwoPoint_Perc
+
    FT_Perc + ORB + DRB + AST + STL + BLK + TOV + PF, data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7644 -1.2571  0.0276  1.0525 11.9053

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6.368456   0.899722  -7.078 4.00e-12 ***
Age             -0.008388   0.021515  -0.390 0.696751
G               -0.014325   0.004590  -3.121 0.001889 **
GS               0.011802   0.005620   2.100 0.036123 *
MP               0.401962   0.025462  15.787  < 2e-16 ***
ThreePoint_Perc  3.492559   0.811340   4.305 1.95e-05 ***
TwoPoint_Perc    6.635965   1.068967   6.208 9.89e-10 ***
FT_Perc          2.483335   0.687293   3.613 0.000327 ***
ORB              0.438831   0.195323   2.247 0.025014 *
DRB              0.245958   0.099339   2.476 0.013557 *
AST             -0.541472   0.108652  -4.984 8.13e-07 ***
STL             -0.414707   0.330123  -1.256 0.209514
BLK              0.028226   0.339083   0.083 0.933687
TOV              4.243351   0.272700  15.560  < 2e-16 ***
PF              -1.573430   0.192243  -8.185 1.58e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.134 on 614 degrees of freedom
Multiple R-squared:  0.8659,      Adjusted R-squared:  0.8629
F-statistic: 283.3 on 14 and 614 DF,  p-value: < 2.2e-16
```

Figure 2 Regression output of simple MLR

[1] "2018-19 NBA Player Stats: Per Game." *Basketball*, https://www.basketball-reference.com/leagues/NBA_2019_per_game.html.

1 Sampaio, Jaime, et al. "Exploring Game Performance in the National Basketball Association Using Player Tracking Data." Plos One, vol. 10, no. 7, 2015, doi:10.1371/journal.pone.0132894.

where each β_i position is represented by a predictor. The regression output is represented in **Figure 2.**

Our simple MLR model can have three interpretations:

1. It can under fit our data which can lead to poor predictions (high bias, low variance).
2. It can overfit our data which can also lead to poor predictions (low bias, high variance).
3. It can appropriately fit our data (just right) leading to good predictions (balance between bias & variance)

We can see from **Figure 2**, that 3 variables (Age, Steal and Blocks) do not significantly contribute to our model due to their P-value results. In addition, a high $Adjusted\ R^2 = 0.8629$ value was found indicating that further analysis needs to be done.

The Simple MLR model data was split 60:40 for training:testing analysis. Cross validation using subset grouping (Stepwise forward & backward Regression) was implemented. The subset grouping stepwise forward selection algorithm starts with a null model and then gradually adds predictors. The backward selection starts with a full model (including all the 14 variables) and then you drop those that are not significant one at a time.

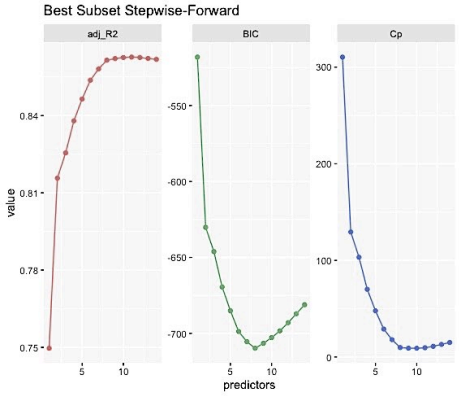


Figure 3 Adjusted R² , BIC, Cp for subset stepwise stepwise-forward

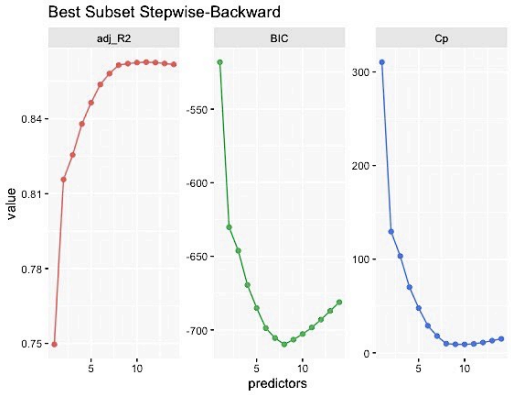


Figure 4 Adjusted R² , BIC, Cp for subset stepwise stepwise-backward

Figure 3, 4 both show commonly used metrics for model evaluation and selection using subset grouping stepwise forward and backward regression; the Adjusted R^2, BIC score and the Cp score.

1. The BIC is an estimate of a function of the posterior probability of a model being true (under a certain Bayesian setup)—it is used for selecting/interpreting the best predictors. Therefore, a lower BIC means that a model is considered to be more likely to be that of the true model.
    - For the stepwise forward regression, that model contained 8 variables
    - For the stepwise backward regression, that model also contained 8 variables
2. The Cp is essentially equivalent to the MSE + (some penalty): It is used to assess the fit of the multiple linear regression.
    - For the stepwise forward regression, that best fit contained 10 variables
    - For the stepwise backward regression, that best fit also contained 10 variables
3. The Adjusted R^2, values are selected by maximizing the R^2, which essentially leads to minimizing the MSE. As we can see from both figure 3 and 4, (forward and back), the Adjusted R^2 values started to plateau at models with variables 8.

Based on the high R^2 and relatively low error values from the best subset selection methods, a new 8 variable model is used. It can be expressed as:

where our 8 variables are:

$$Y_{14} = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \epsilon_i \qquad \textbf{Eq. 2}$$

8 MLR MODEL	
Predictors	**Response (Outcome)**
1. Minutes Played (MP) 2. Three Point % (ThreePoint_Perc) 3. Two Point % (TwoPoint_Perc) 4. Free Throw % (FT_Perc) 5. Defensive Rebounds (DRB) 6. Assists (AST) 7. Turnovers (TOV) 8. Personal Fouls (PF)	1. Points (PTS)

Table 2 MLR model 8 covariates

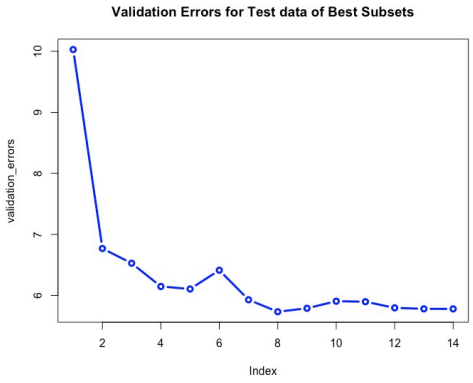


Figure 5 Validated Errors for Test Data of 8 MLR

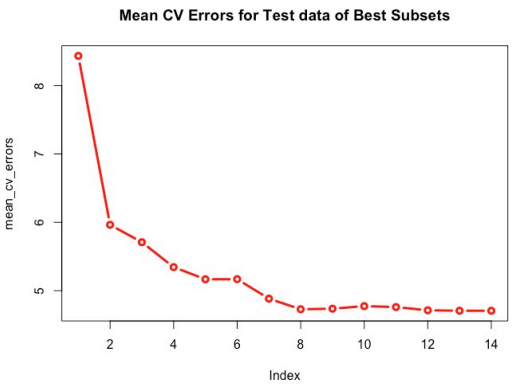


Figure 6 Mean CV Errors for Test Data of 8 MLR

Figure 5 and **Figure 6** confirm our analysis from the previous section. Using 10-fold cross validation, we can see that the 8 variable model is an appropriate model for predicting average points per player. We now assess our findings by evaluating statistics from the 8 variable MLR model.

Figure 7 is the output for the multiple linear regression model using the 8 predictors. Our Adjusted R^2 decreased by ~0.002 when comparing it to the initial MLR using 14 variables. With that said, a model with fewer variables is preferred as it may capture real-world effects better. All 8 predictors shown in **Figure 7** are statistically significant. **Figure 8** visualizes the R^2 values for each of the 14 variables.

```
Call:
lm(formula = PTS ~ MP + ThreePoint_Perc + TwoPoint_Perc + FT_Perc +
    DRB + AST + TOV + PF, data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0352 -1.2524 -0.0844  1.0614 12.5206

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.71370    0.74628  -8.996  < 2e-16 ***
MP              0.38812    0.02054  18.891  < 2e-16 ***
ThreePoint_Perc 3.07186    0.79329   3.872 0.000119 ***
TwoPoint_Perc   6.48897    1.04205   6.227 8.76e-10 ***
FT_Perc         2.26327    0.68027   3.327 0.000930 ***
DRB             0.39281    0.07740   5.075 5.12e-07 ***
AST            -0.60514    0.10164  -5.954 4.39e-09 ***
TOV             4.31338    0.26673  16.171  < 2e-16 ***
PF             -1.51889    0.18379  -8.264 8.54e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.153 on 620 degrees of freedom
Multiple R-squared:  0.8622,      Adjusted R-squared:  0.8604
F-statistic:   485 on 8 and 620 DF,  p-value: < 2.2e-16
```

Figure 7 Regression output of 8 MLR

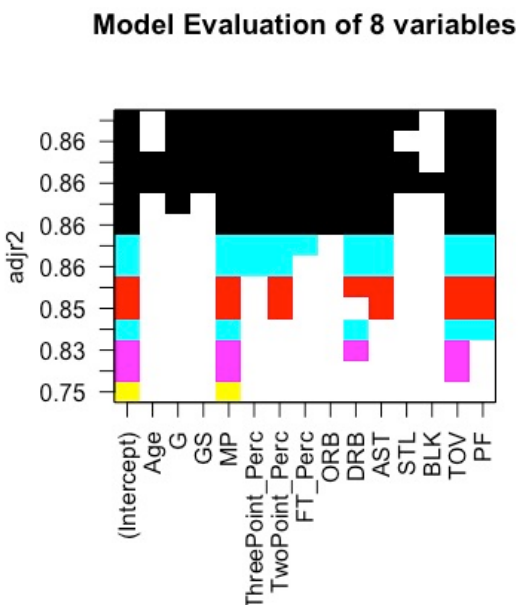


Figure 8 :Adjusted R^2 plot for multiple models

Notably, the six predictors with R^2 value < 0.86 are excluded from the best subset methods. After finding an 8 variable model that produces the highest R^2 and lowest error criterion (BIC, Cp), we decided to look at multicollinearity among the 8 predictors.

Variance inflation factor (VIF) values range from 1 to more than 10, where 1 represents zero collinearity and more than 10 represents high levels of collinearity with other predictors. A general rule of thumb is to consider the removal of predictors with VIF values that are > 5/10 as they may impact the results.[3] In R, the car::vif() function allowed us to find the level of multicollinearity among the predictors. Two

[3] James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2017

predictors had relatively moderate values VIF values: Average Turnovers Per Game had a VIF of 5.34 and Average Minutes Played Per Game had a VIF of 4.13. Other VIF values ranged from 1.15 to 2.53.

It's worth noting that in the 8 variable model, we had an R^2 value of 86.2%. All 8 variables are significant at the $\alpha = 0.05$ level (The chance that any of the 8 variables appears due to chance is < 5%) If the predictor with the highest VIF score was removed, the R^2 value would drop to 80.2%. If the two predictors with the highest VIF scores were removed, then the R^2 value would drop to 70.9%. While these two predictors have a moderate level of collinearity with other predictors, we decided to keep them in our model as their inclusion allows us to better explain the variance in PTS scored. More information on R^2 values for different models can be found in **Table A** in the Appendix.

In order to confirm our decision to go with the 8 variable model, we looked at how variables performed in other linear models, namely ridge and lasso regression, which implement shrinkage penalties. The purpose of the shrinkage penalties is to decrease the weight of less important coefficients: specifically, ridge regression brings values closer to zero while lasso regression sets them to exactly zero. These models focus more on model accuracy (i.e., R^2 or error values) as opposed to model interpretability (i.e., the relationship of each predictor with our outcome variable). By reducing coefficients, the model variance decreases and accuracy should improve.
Ridge and lasso regression models differ from ordinary least squares in a few ways:

1. All variables must be standardized
   - Standardized variables are transformed to have a mean of 0 and a standard deviation of 1
2. Coefficients of less important variables are brought closer to zero (ridge regression) or set to exactly zero (lasso regression)
   - As the shrinkage penalty lambda (λ) approaches infinity, the coefficients get closer to or are set to zero.
3. Results are less interpretable since the coefficients are modified
   - The focus is on improving accuracy, e.g., R^2, as opposed to improving interpretability.

Ridge and lasso regression models were run for both 8 variable and 6 variable models. Cross validation was used to find the lambda (shrinkage value) term between 10^{10} and 10^{-2} that produces the highest R^2 value.

- As shown in **Table 3** below, the R^2 values for the 8 variable models were 86.2% while they were ~71% for the 6 variable models.
- Lambda values were between 0 and 0.5.
- The coefficients for ordinary least squares and lasso regression were mostly the same while they are slightly different for ridge regression. It's worth noting that while Lasso regression can reduce coefficients down to exactly zero, this did not happen in the 6 or 8 variable model.

Predictor	OLS, 8 variables	Ridge, 8 variables	Lasso, 8 variables	OLS, 6 variables	Ridge, 6 variables	Lasso, 6 variables
Three Point %	3.1	3.3	3.0	5.4	5.3	5.3
Two Point %	6.5	5.5	6.4	4.7	4.6	4.5
Free Throw %	2.3	2.8	2.2	5.4	5.3	5.3
Defensive Rebounds (DRB)	0.4	0.6	0.4	1.4	1.3	1.4
Assists (AST)	-0.6	-0.1	-0.6	1.4	1.3	1.4
Personal Fouls (PF)	-1.5	-0.7	-1.5	0.9	1.1	0.9
Minutes Played (MP)	0.4	0.3	0.4	n/a	n/a	n/a
Turnovers (TOV)	4.3	3.1	4.2	n/a	n/a	n/a
Intercept	-6.7	-6.5	-6.6	-7.5	-7.1	-7.2
Lambda (shrinkage term)	n/a	0.507	0.007	n/a	0.407	0.018
R^2	86.22	83.89	84.71	71.17	69.72	69.85
Test Mean Squared Error (MSE)	4.57	4.46	4.15	9.56	9.19	9.13

Table 3 Values for 8 and 6 variable Ridge and Lasso models

B. Model Diagnostics

After deciding on an 8 variable model, we looked at model diagnostics to see whether the various assumptions for a linear regression model are met. Referencing the four plots in **Figure 9,** we see that:

- The top-left chart titled "Residuals vs Fitted" shows whether there is constant error variance for the multiple linear regression model. The chart informs us there is not constant error variance as there is a large variance among residuals as the fitted values increase, resulting in a rightward fan shape. A solution to would be to transform the predictors and/or the outcome variable.
- The top-right chart titled "Normal Q-Q plot" looks at whether the residuals follow a normal distribution which would have most points fall on the diagonal line. Since there are clear violations of the normality assumption (as the points on the right half deviated from the diagonal line), a transformation on the data would help. The same transformation(s) may be able to address the violations for both constant error variance and normality.
- The bottom-left plot titled "Scale-Location" shows that as fitted X values get larger, the standardized residuals are greatly affected. In other words, there is non-independence of error terms. To address this, further research would look into which variables are related and possibly apply transformations.
- The bottom-right plot titled "Residuals vs Leverage" shows where outliers, leverage points, and influential points are located. Outliers are extreme outcome values, high leverage points are extreme predictor values, and influential points have large residuals because the combination of predictor values is uncommon. Next steps would look into whether these observations should be kept, removed, or if their values should be adjusted.

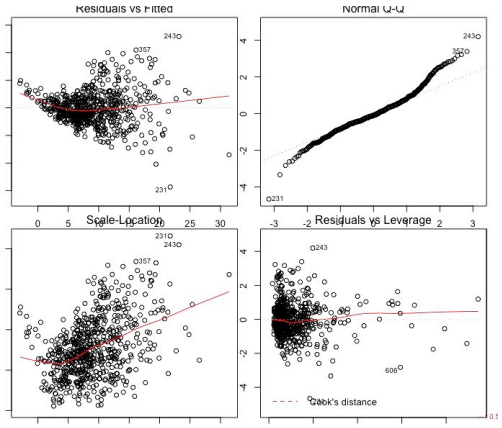


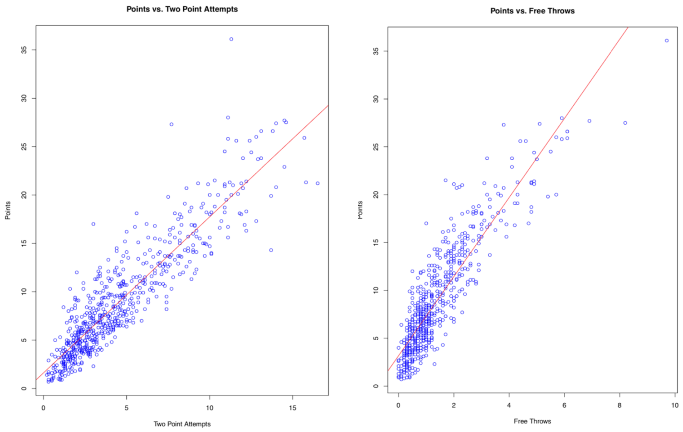
Figure 9 Diagnostic plots of 8 variables MLR model

II. Simple Linear Regression

A. Variable Selection

After looking at multiple linear regression, we decided to explore whether or not individual relationships exist between points scored per game (PTS) and each of the other 8 variables in our MLR model. None of these 8 variables, individually, had significant adjusted R^2 and therefore, we decided to explore the remaining variables in our dataset. Two predictors, two point attempts (2PA) and free throws made (FT) were evaluated for their high R^2 value with PTS. Scatterplots of the two models display this.

B. Relationships Between Variables



Figures 10 & 11 : Scatterplots of 2PA (left) and FT (right) Vs. PTS

The scatterplots in **Figures 10 and 11** show a strong positive linear relationship in each model. With adjusted R^2 values of .8173 for 2PA and .7965 for FT, a large amount of the variation in points can be described by these two variables separately. **Table 4** shows the corresponding regression functions.

Table 4 : Regression Coefficients and Equations for 2PA and FT

	b_0	b_1	regression equation
2PA	**1.619534**	**2.728632**	$\hat{Y}$ **= -.01874876+ 2.728632X**
FT	**3.179097**	**1.612499**	$\hat{Y}$**= 1.619534 + 1.612499X**

We estimate the mean number of points increases by approximately 1.612 for each two point attempt taken. When a player scores 0 points, it is estimated that the player will attempt 1.6195 two point shots. The same interpretations can be made for our free throw model.

C. Inferences in Regression and Correlation Analysis

We can perform s*eparate* confidence intervals on the regression coefficients. With 95% confidence, for two points attempted, we estimate the true value of β_1 to fall between 1.5528 and 1.6722 and similarly, for free throws made, we estimate the true value of β_1 to fall between 3.9695 and 4.2969. 95% confidence intervals were also conducted for β_0 for 2PA and FT determining $1.2856 \leq \beta_0 \leq 1.9535$ and $2.8721 \leq \beta_0 \leq 3.4861$, respectively.

Additionally, we performed 90% confidence intervals for Y_h, prediction intervals for Y_h, and confidence bands for both models, 2PA and FT vs. PTS. **Table 5** shows randomly chosen X_h levels 3 and 1.2. The Y_h's are our actual mean points scored at 6.7 for 2PA and 8.3 for FT and their respective $\hat{Y}h$ values at 6.5 and 8.1. None of our Y_h's fall within any of their respective intervals. However, they are very close.

Table 5: Other Interval Inferences

Variables	X_h	Y_h	$\hat{Y}_h$	90% CI for Y_h	90% Pred. Interval for Y_h	90% Confidence Band
2PA	**3**	**6.7**	**6.456**	**(6.4206, 6.4879)**	**(6.3811, 6.5308)**	**(6.4111, 6.5012)**
FT	**1.2**	**8.3**	**8.1386**	**(8.1278, 8.1493)**	**(8.1209, 8.1563)**	**(8.1233, 8.1539)**

Next we wanted to test our models that have been fitted to the dataset. We hypothesized that there was a linear association between average points made by a player and average 2PA and FT. We conducted the F-test for lack of fit. As you can see from **Table 6** the F-statistics are greater than their associated F critical values for all three instances. Therefore, we rejected our null hypothesis and concluded that there is a linear association in between each predictor and PTS.

Table 6: F-Test for Lack of Fit

	H*0*:	H*a*:	F	F*	Conclude
2PA	E(Y)= B*0*+B1X	E(Y)≠ B*0*+B1X	1.25	1.503	H*a*
FT	E(Y)= B*0*+B1X	E(Y)≠ B*0*+B1X	1.28	3.469	H*a*

We can also perform *joint* confidence intervals. The Bonferroni joint confidence intervals were conducted for simultaneous estimations of β_0 and β_1. The family confidence coefficient is at least .95 that the procedure leads to pairs of interval estimates. For 2PA, we conclude that the true value of β_0 is between 1.237 and 2.002 and the true value of β_1 is between 1.54 and 1.681, simultaneously. Similarly, for FT, the true value of β_0 is between 2.828 and 3.530 and the true value of β_1 is between 3.946 and 4.320, simultaneously.

Using the Work Hotelling procedure we also conducted simultaneous estimations of the mean responses at two levels. X_h values were chosen randomly from the dataset to obtain simultaneous confidence intervals at a 90% significance level for the mean responses at those X levels (**Table 7).**

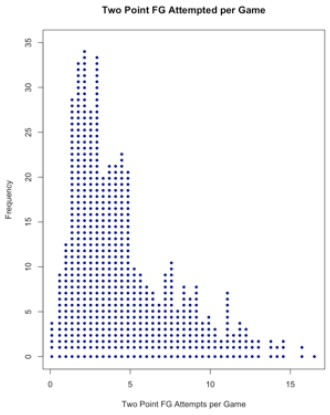
Table7: Family Confidence Intervals for Y_h

Variable	X_h	$\hat{Y}_h$	Y_h	Family Confidence Interval for Y_h
2PA	3.5	7.262	7.5	(7.231, 7.293)
	7.6	13.8712	13.6	(13.783, 13.959)
FT	1	7.312	7	(7.2818, 7.342)
	2	11.445	11.8	(11.400, 11.489)

None of our Y_h's fall within their respective family confidence Intervals. However our $\hat{Y}_h$ do. Again, the actual mean points scored are not far from their respective intervals. We are 90% confident that the mean responses for each chosen X level will fall between the intervals, simultaneously.

D. Model Diagnostics and Remedies

When a simple regression model is considered for application, we usually can't be certain in advance that the model is appropriate for that application. There are diagnostics for both the predictor variables and the residuals that must be done in order to determine whether the simple linear regression model is appropriate. We will now explore possible model violations and discuss possible remedial measures on our two simple regression models. Let us first look at the distribution of the predictor variables from our simple regression models, both 2PA and FT.



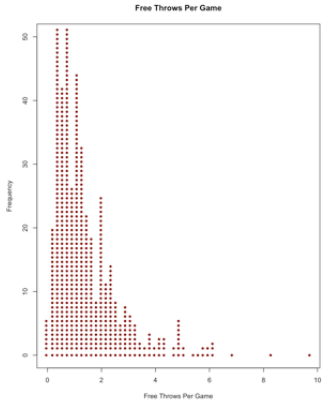


Figure 12: dot plot of predictor variable 2PA **Figure 13: dot plot of predictor variable FT**

We can see from **Figure 12**, that 2PA doesn't appear to have any extreme outliers. FT has three possible outliers but for now we leave them in because it is unclear whether or not they were in error. Next we must explore residual analysis to determine if departures from the model exist. Some important departures from the simple linear regression model with normal errors include:

1. The regression function is not linear.
2. The error terms do not have constant variance.
3. The model fits all but one or a few outlier observations.
4. The error terms are not normally distributed.

First, we can test whether or not our predictor variables 2PA and FT are linearly associated with our outcome variable, PTS. Using a T-test for linear association we tested the null hypothesis: $\beta_1 = 0$ on the variable 2PA. Since the P-value was less than an alpha level of .05, we were able to reject the null hypothesis and conclude that there is a linear association between 2PA and PTS. This same conclusion was made between FT and PTS. Additionally, we performed the F-test for linear association and concluded with 95% confidence that both 2PA and FT were linearly associated with PTS. With this knowledge, we can conclude that there was no model violation based on linearity.

Next, we must look at the error terms of both simple regression models to determine if they have constant variance. To do this we can examine a residual plot against our predictor variables 2PA and FT.

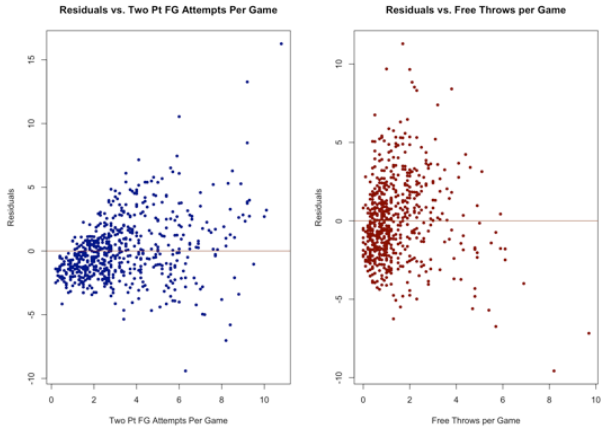


Figure 14: residuals plotted against predictor variables 2PA (left) and FT (right)

The residual plots in **Figure 14** both show what is often referred to as a megaphone configuration, where the residual variance increases as the value of the predictor variable increases. Both plots clearly indicate a non-constancy of error variance and a violation of the model assumptions.

Figure 15 although similar to **Figure 14**, indicates the presence of residual outliers. Rather than plotting residuals against the predictor variables, we have standardized the residuals in such a way that lends to easy comparison and indication of outliers. We consider a residual outlier to be any residual with a value greater than |4|. Both models contain residual outliers although FT has two less than 2PA. It is difficult to determine when we should remove them and when we should not. Further research should be done to determine if they were produced in error. Again, these outliers violate our model assumptions.

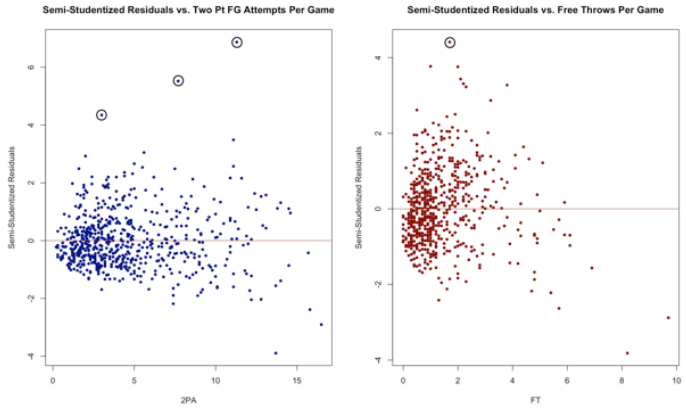


Figure 15: semi-studentized residuals plotted against predictor variables 2PA

Lastly, we must determine whether or not the error terms are normally distributed. QQ plots of residuals in **Figure 16** indicate heavy tails for both variables and lead us to believe the error terms are not normally distributed. This is confirmed by the test of normality where we obtained the coefficient of correlation between the ordered residuals and their expected values under normality (using **Eq. 3**) and compared it to the critical value for n=629 (the number of observations in our study).

$$\sqrt{MSE}[z(\frac{k-.375}{n+.25})] \quad \textbf{Eq. 3}$$

After obtaining r = .965 and r_{crit}=.9976 for 2PA, we can reject the null hypothesis since $r < r_{crit}$ and conclude that the residuals for the 2PA model are not normally distributed. The same conclusion was made for FT.

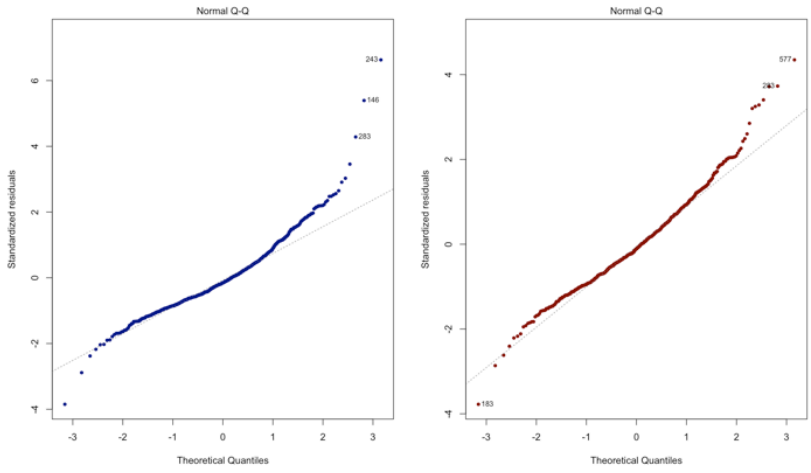


Figure 16: Normal Probability Plot of the residuals for 2PA (left) and FT

Since all three of our models lack constant error variance and normal error distributions we can perform a transformation on Y to try and remedy these departures from the simple regression model. Let's look at the predictor FT.

Here, we've let $Y' = \sqrt{Y}$

Where $Y' = \beta_0 + \beta_1 X_i + \epsilon_i$

Figure 17:Residuals plotted against FT and QQ plot of FT from square root transformation on Y

Figure 17 shows this transformation plotted with a residual plot against FT as well as a QQ plot of the transformed residuals. While the square root transformation seems to correct for normality, the error terms still appear to be inconstant. Additionally the adjusted R2 decreased from 0.7965 to 0.7097 using the transformed model. Similarly, the 2PA did not seem to have a suitable transformation to fix both departures from the simple regression model. All the transformations that we tried reduced the adjusted R2 and made the normality plot of residuals much worse.

III. Results and Conclusion

In conclusion, for simple linear regression, some of the model assumptions were violated in both of our simple regression models:

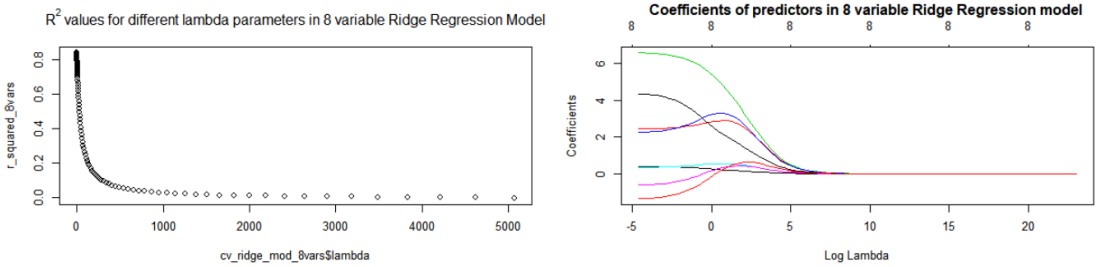
1. Error variance was not constant
2. Residuals were not normally distributed
3. Residual outliers existed

In order to use the simple linear regression model, some remedial measures must be taken. Perhaps we can remove some outliers or perform a better transformation on Y in order to reduce the inconstancy in error variance and normalize the distribution of the residuals. It may be possible that another model would work better for the data. Once appropriate adjustments are made, the inferences we made from our models should be retested. In summary, it does not appear to be effective to use simple regression to predict PTS. Multiple linear regression is a more appropriate model to use when considering the relationship between PTS and multiple other variables in this dataset.

APPENDIX:TABLE A Adjusted R^2 and variables removed in various-sized models

# Variables	14 variables	8 variables	7 variables	6 variables
Adjusted R^2	86.29%	86.22%	80.19%	70.89%
Variables removed	<u>Qualitative variables</u> 1. Player name 2. Team name 3. Player Position 4. Alpha ID <u>Quantitative variables</u> 5. Field Goals made 6. Field Goals attempted 7. Field Goals percentage 8. 3 pointers made 9. 3 pointers attempted 10. 2 pointers made 11. 2 pointers attempted 12. Effective Field Goal % 13. Free throws made 14. Total Rebounds attempted	1) Age 2) Games Played 3) Games Started 4) Offensive Rebounds 5) Steals 6) Blocks	1) Turnovers Vif=5.34	1. Turnovers, Vif=5.34 2. Minutes Played Vif=4.13
Methodology to get # Variables	Starting with 30 variable model, removed 4 qualitative variables and 11 quantitative variables based on research, professional experience, and variables that may "leak" into model performance.	Starting with 14 variable model, used Best Subset selection	Starting with 8 variable model, removed predictor with highest VIF value	Starting with 8 variable model, removed top two predictors with highest VIF values

APPENDIX:CHART 2: R^2, Coefficient values for 8 variable Ridge and Lasso Regression models



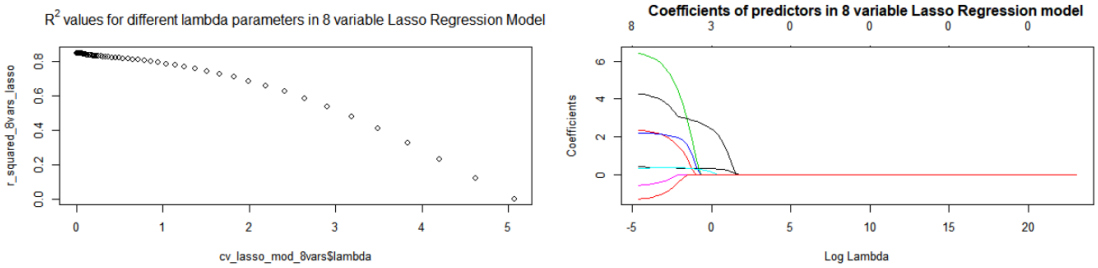


Chart Y: R^2, Coefficient values for 6 variable Ridge and Lasso Regression models