Determining variables which best explain average points scored by NBA players, 2018-2019

General Linear Models I - Fall 2019

Yamar Ba
Samantha Benedict
Cesar Rene Pabon Bernal
Johnny C. Mathis
Fouad Yared

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.

This study will rely on basketball-reference.com's 2018-2019 Average NBA Player Statistics per Game1 which can be accessed here:

https://www.basketball-reference.com/leagues/NBA_2019_per_game.html

Multiple linear regression

The original data set consisted of 30 variables. For the purpose of this report, the four qualitative variables were removed and the remaining 26 variables were explored. Detailed exploration based on research[1] and expert analysis (from co-author and former professional basketball player Johnny Mathis), the multiple linear regression model (MLR) was reduced to 14 predictors, which are calculated as averages per game:

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

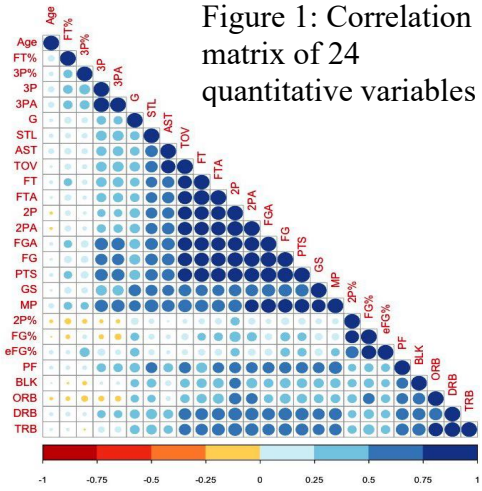


Figure 1: Correlation matrix of 24 quantitative variables

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. Our initial 14 variable MLR model can be expressed as:

$$Y_{14} = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \epsilon_i + \beta_{13} X_{13} + \beta_{14} X_{14} + \varepsilon_i$$

where each β_i position is represented by a predictor.

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.

Figure 2: Statistics for 14 variable multiple linear regression

```
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
```

Our simple MLR model can have three interpretations:

- It can underfit our data which can lead to poor predictions (high bias, low variance).
- It can overfit our data which can also lead to poor predictions (low bias, high variance).
- It can appropriately fit our data leading to good predictions (balance between bias & variance)

We can see from Figure 2, that 3 variables (Age, Steals and Blocks) do not significantly contribute to our model due to their P-value results. Additionally, we wanted to see whether we could remove some variables while maintaining much of the explanatory power of the 14 variable model which had an adjusted R^2 of 86.29%.

To evaluate the model against unseen observations, 60% of the data was set aside to be the training data set while 40% of the data was set aside to be the testing data set.

Subset grouping methods, namely stepwise forward and stepwise backward methods, were implemented. The purpose of subset grouping is to find the number of variables that produces the highest R^2 and/or lowest error criterion.

- For forward selection, the algorithm starts with a null model and then gradually adds predictors that contribute the most and stops adding variables when they are no longer statistically significant.

Backward selection starts with all variables in the model and removes variables, one-at-a-time, that contribute the least. The resulting model retains only variables that are statistically significant.

Figure 3, 4: Adjusted R^2 , BIC, Cp values for stepwise forward and stepwise backward methods

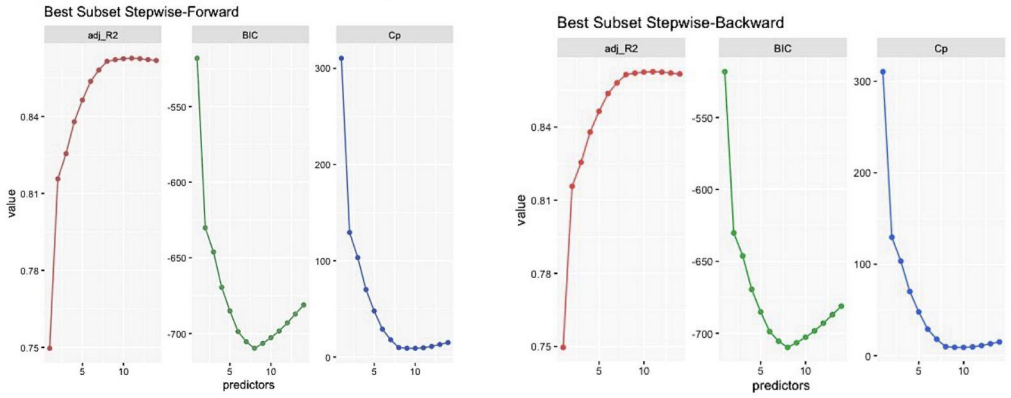


Figure 3, 4 show commonly used metrics for model evaluation and variable selection. Using subset grouping stepwise forward and stepwise backward methods, the Adjusted R^2, BIC score and the Cp score were evaluated.

1. For Adjusted R^2, values are selected that maximize the R^2, which essentially leads to minimizing the MSE. As we can see from both figure 3 and 4, the Adjusted R^2 values started to plateau for models with 8 variables.
2. 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 more likely closer to the true model.
   - The nadir, or lowest BIC value, is associated with 8 variables in both the stepwise forward and backward methods.
3. The Cp is essentially equivalent to the MSE + (some penalty): It is used to assess the fit of the multiple linear regression model.
   - For both stepwise forward and backward methods, the lowest Cp is associated with 10 variables. It is worth noting that the values for Cp plateaued at 8 variables.

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:

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

where our 8 variables are:

List of 8 Predictors		
1) Minutes Played (MP)	4) Free Throw Percentage (FT_Perc)	7) Turnovers (TOV)
2) Three Point Percentage (ThreePoint_Perc)	5) Defensive Rebounds (DRB)	8) Personal Fouls (PF)
3) Two Point Percentage (TwoPoint_Perc)	6) Assists (AST)	
Outcome variable		
A) Points (PTS)		

Figure 5, 6 : 10-fold Cross validation of the 8 variable model

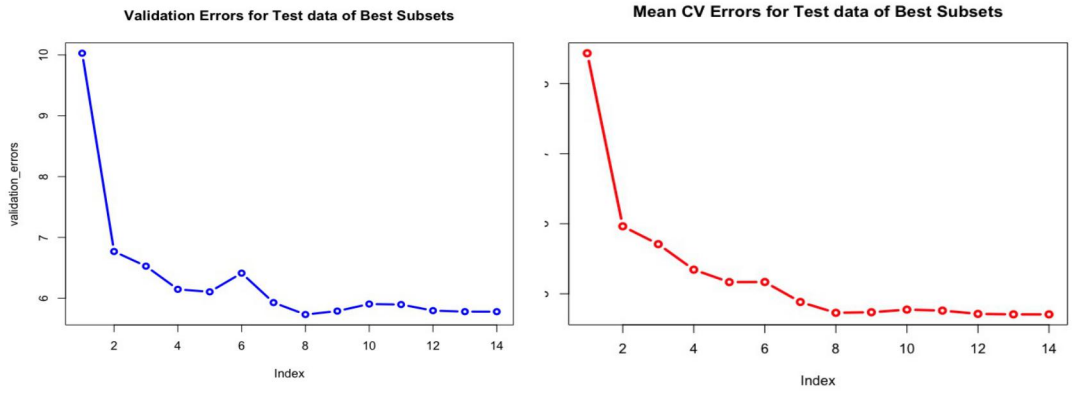


Figure 5 and **Figure 6** confirm our analysis from the previous section. Using 10-fold cross validation, we can see that the 8 variable model is an appropriate model for predicting average points per player. We now assess our findings by evaluating statistics from the 8 variable MLR model.

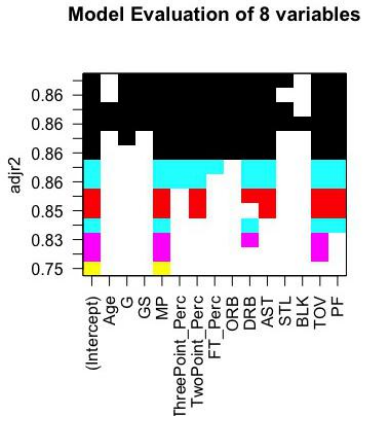
Figure 7: Statistics for 8 variable multiple linear regression

```
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 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. Notably, the six predictors with R^2 values less than 0.86 are excluded from the best subset methods.

Figure 8: Adjusted R^2 plot for multiple models



After finding an 8 variable model that produces the highest R^2 and lowest error criterion (BIC, C_p), 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 greater than 5 or 10 as they may impact the results. (Introduction to Statistical Learning, p 101-102).

In R, the car::vif() function allowed us to find the level of multicollinearity among the predictors. Two 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 less than 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 1 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
    a) 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)
    a) 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
    a) 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 1 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.

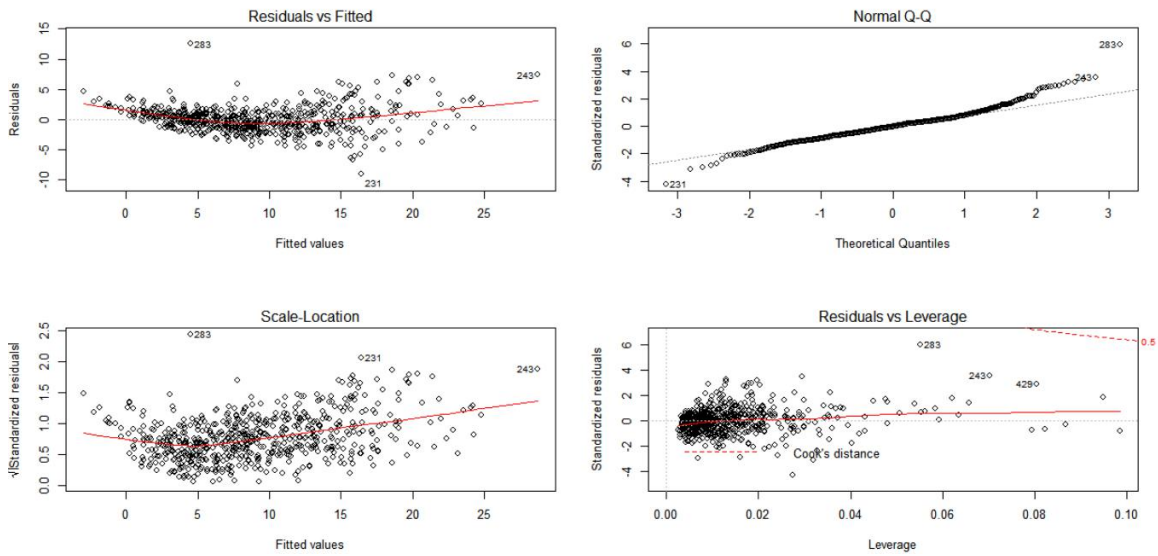
Table 1: Values for 8 and 6 variable Ridge and Lasso models

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

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 variable multiple linear regression model



Simple linear regression

After looking at multiple linear regression, we decided to look into how individual variables relate to average points scored for each player. Two predictors, two_PA (two pointers attempted) and free throws made were evaluated for their high R^2 value with PTS. Making a scatterplot and calculating the coefficient of determination showed that these 2 predictor variables would be a good fit for our simple linear regression model.

Figure 10
>cor(bbdata_1_$PTS,bbdata_1_$two_PA)
[1] 0.9042281

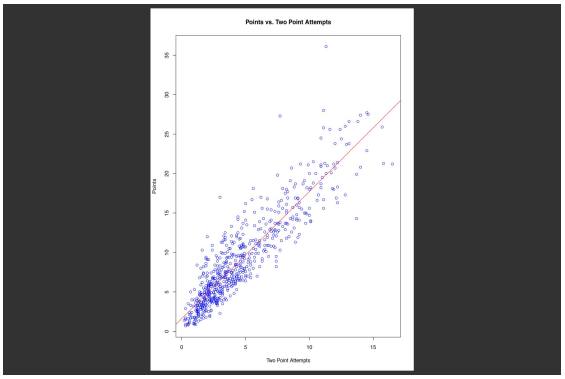
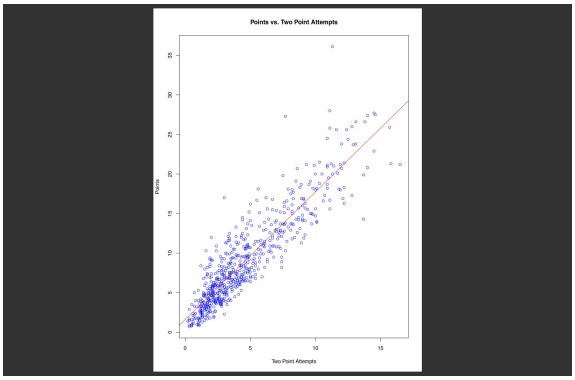


Figure 11
cor(bbdata_1_$PTS,bbdata_1_$FT)
[1] 0.8926232



The scatterplots in Figure 1 show a strong positive linear relationship in each model. The interpretation of the information in figure 2 is approximately 90.42% of the variation in points can be described by the variation in two point attempts and approximately 89.26% of the variation in points can be described by the variation in free throws.

To obtain the estimated regression function, we needed to find b0 and b1 for each regression model.

Figure 3.

```
coef(lm(bbdata_1_$PTS~bbdata_1_$two_PA))
    (Intercept) bbdata_1_$two_PA
       1.619534         1.612499
> coef(lm(bbdata_1_$PTS~bbdata_1_$FT))
 (Intercept) bbdata_1_$FT
   3.179097     4.133236
```

The interpretation of b1 In Figure 3 is we estimate the mean number of points increases by approximately 1.612 for each two point attempt taken and we estimate the mean number of points increases by approximately 4.1332 for each free throw taken. The interpretation of b0 is when a player scores 0 points, it is estimated that the player will make approximately 3.1791 free throws and attempt 1.6195 two point shots. Therefore, our estimated regression function is $\hat{Y}$= 1.619534 + 1.612499X and $\hat{Y}$ = 3.179097 + 4.133236X respectively.

In R, 95% confidence intervals of 0 and 1 were obtained.

Figure 4

```
> confint(lm(bbdata_1_$PTS~bbdata_1_$two_PA))
                   2.5 %   97.5 %
(Intercept)      1.285568 1.953500
bbdata_1_$two_PA 1.552774 1.672224

> confint(lm(bbdata_1_$PTS~bbdata_1_$FT))
                2.5 %   97.5 %
(Intercept)  2.872099 3.486095
bbdata_1_$FT 3.969531 4.296942
```

Figure 4 can be interpreted that with 95% confidence, we estimate that the mean number of points increases approximately between 1.5528 to 1.6722 for each two points attempted and the mean number of points increases approximately between 3.9695 to 4.2969 for each free throw made. Also, with 95% confidence the range of two point attempts and free throws are approximately 1.2856 to 1.9535 and 2.8721 to 3.4861, respectively when a player scores 0 points.

Confidence Intervals

We Obtained 90% confidence intervals for mean points scored on two variables:

- Mean two point field goal attempts (2PA)
- Mean Free-throws (FT).

In figure 1.1, the first column of the table are our variables. The second and third columns are the variables respective b_0's and b_1's. We randomly chose the Xh 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.

Figure 1.1

Variables	b_0	b_1	X_h	Y_h (actual value)	$\hat{Y}_h$	90% CI for Y_h	90% Pred. Interval for Y_h	90% Confidence Band
2PA	1.62	1.612	3	6.7	6.456	(6.4206, 6.4879)	(6.3811, 6.5308)	(6.4111, 6.5012)
FT	3.179	4.133	1.2	8.3	8.1386	(8.1278, 8.1493)	(8.1209, 8.1563)	(8.1233, 8.1539)

We conducted three interval tests. We have 90% confidence intervals, 90% Predicted intervals for Y_h and 90% confidence bands. None of our Y_h's fall within any of their respective intervals. However, they are very close.

F-Test for Lack of Fit

We wanted to test our models that have been fitted to our 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 figure 1.2, the F-tests or F-stars are greater their associated F-criticals for all 3 instances.

Figure 1.2

	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*

Therefore, we rejected our hull hypothesis and concluded that there is a linear association in our models.

Bonferroni Procedure

The Bonferroni joint confidence intervals procedure was also conducted. We wanted to estimate where the betas for each variable would fall simultaneously given any Xh.

The joint confidence intervals for our variables and their associated beta's are listed in table figure 1.3. With 2PA, we conclude that its b_0 is between 1.237 and 2.002 and its b_1 is between 1.54 and 1.681. Similarly, for FT as indicated in the table.

Figure 1.3

	Simultaneous Confidence Intervals for 0 and 1
2PA	1.237 ≤ $_0$ ≤ 2.002
	1.54 ≤ $_1$ ≤ 1.681

FT	2.828 ≤ $_1$ ≤ 3.530
	3.946 ≤ $_1$ ≤ 4.320

The family confidence coefficient is at least .95 that the procedure leads to pairs of interval estimates.

Working Hotelling Procedure

We conducted simultaneous estimations of the mean responses at two levels. Xh values were chosen from the dataset and the Working-Hotelling procedure was used to obtain simultaneous confidence intervalsI at 90% for mean responses at those X levels (figure 1.4).

Figure 1.4

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. As before, in figure 1.4, the actual mean points scored are not far from their respective intervals. For instance, 2PA at Xh 3.5 has a difference of 2.6% from the higher bound value of its interval of 7.293 and a difference of 1.3% for Xh 7.6 from its lower bound interval value of 13.783. Similarly, for FT as listed in the table. Therefore, we are 90% confident that the mean responses for each chosen X level will fall between the intervals, simultaneously.

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.

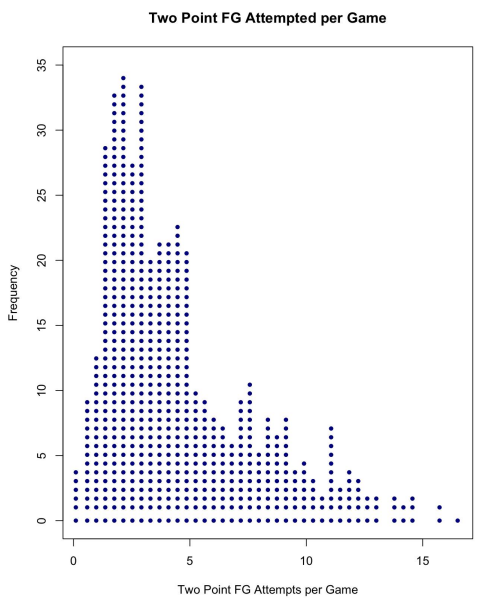


Figure 1: dot plot of predictor variable 2PA

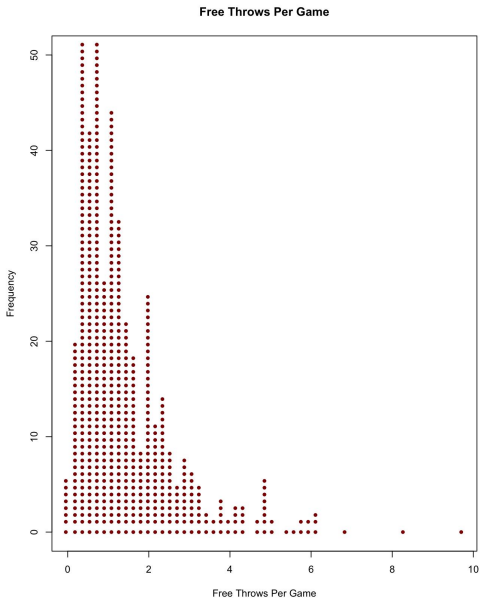


Figure 2: dot plot of predictor variable FT

We can see from **Figure 1,** 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: $\boldsymbol{\beta}_1 = 0$ 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.

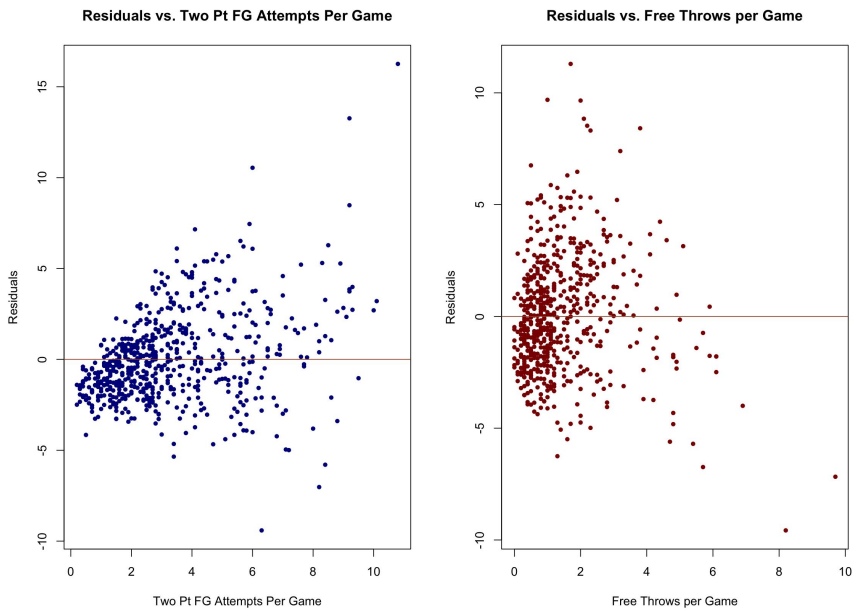


Figure 3: residuals plotted against predictor variables 2PA (left) and FT (right)

The residual plots in **Figure 3** 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.

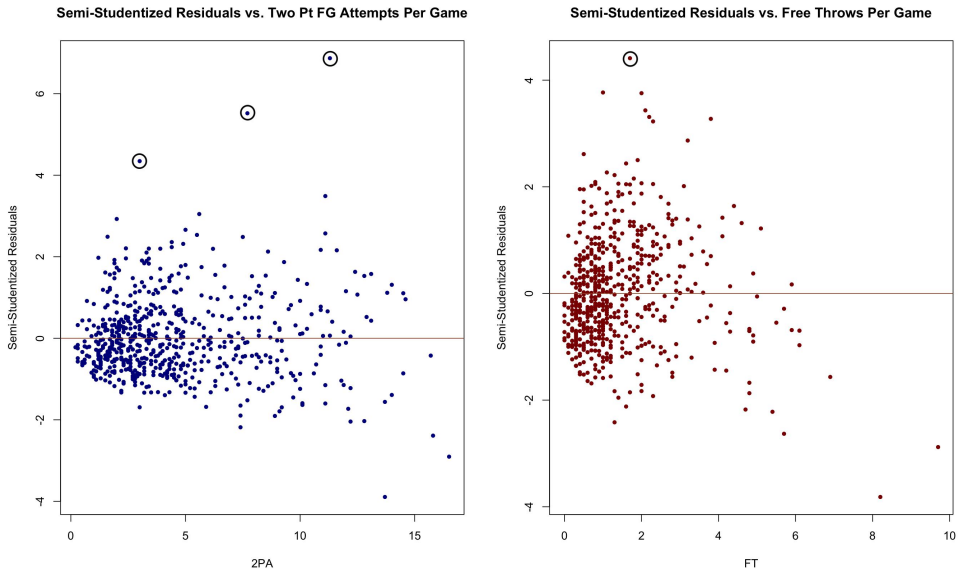


Figure 4: semi-studentized residuals plotted against predictor variables 2PA (left) and FT (right)

Figure 4, although similar to **Figure 3**, 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.

Lastly, we must determine whether or not the error terms are normally distributed. QQ plots of residuals in **Figure 5** 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. 1**) and compared it to the critical value for n=629 (the number of observations in our study).

$$\sqrt{MSE}\left[z\left(\frac{k-.375}{n+.25}\right)\right]$$

Eq. 1

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.

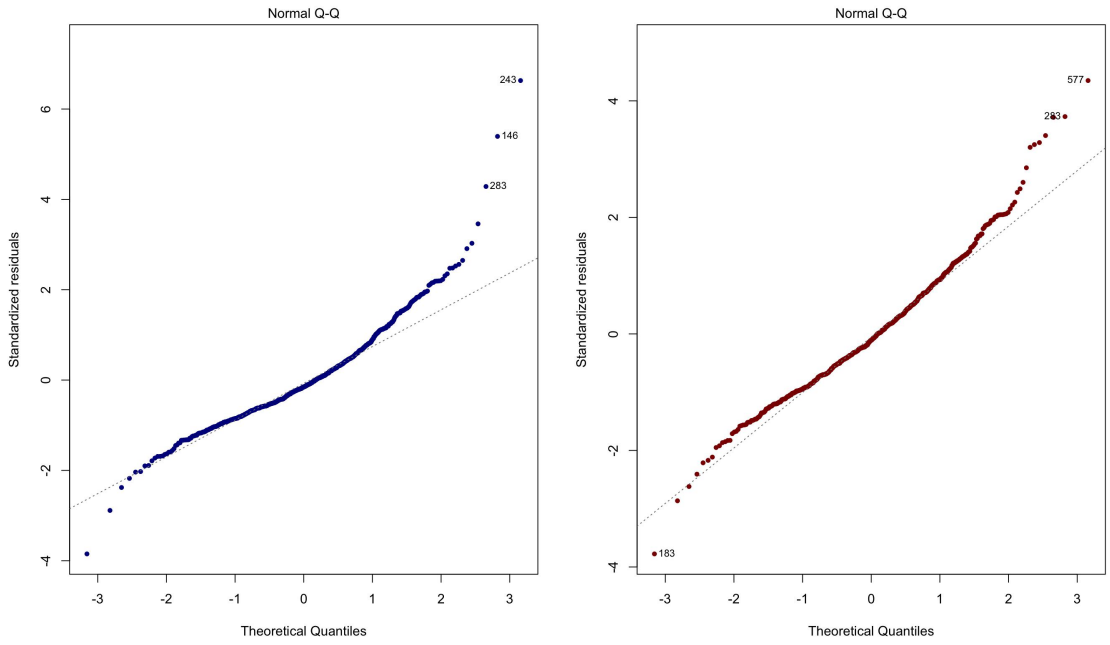


Figure 5: Normal Probability Plot of the residuals for 2PA (left) and FT (right)

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 $\mathbf{Y'} = \sqrt{Y}$

Where $\boldsymbol{Y'} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_i} + \boldsymbol{\varepsilon_i}$

Figure 6: Residuals plotted against FT and QQ plot of FT from square root transformation on Y

Figure 6 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 R^2 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 R^2 and made the normality plot of residuals much worse.

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 1: 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	Qualitative variables 1. Player name 2. Team name 3. Player Position and 4. Alpha ID Quantitative variables 1) Field Goals made 2) Field Goals attempted 3) Field Goals percentage 4) 3 pointers made 5) 3 pointers attempted 6) 2 pointers made 7) 2 pointers attempted 8) Effective Field Goal % 9) Free throws made 10) Free throws attempted 11) 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

Chart X: R^2, Coefficient values for 8 variable Ridge and Lasso Regression models

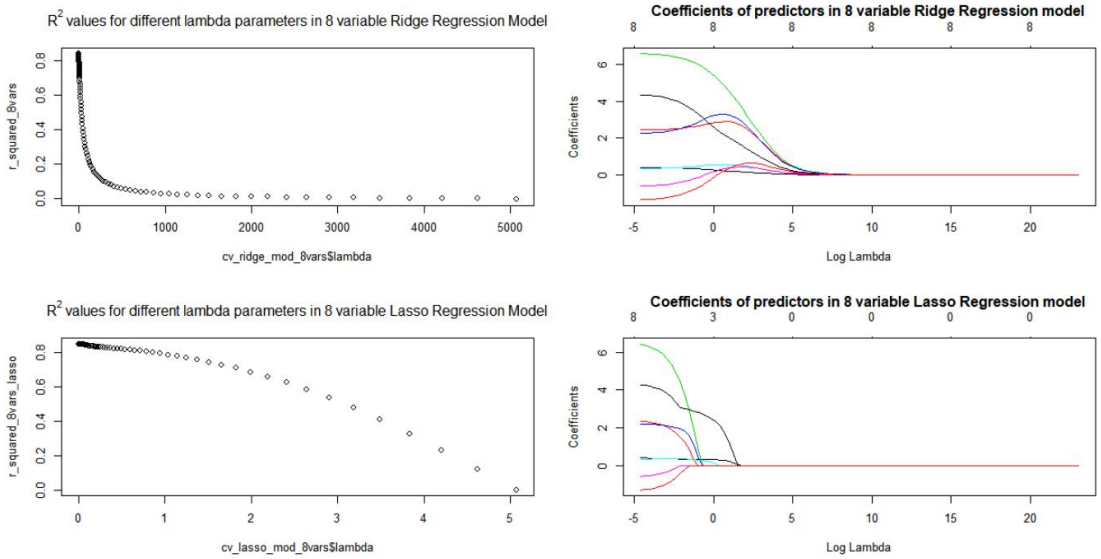


Chart Y: R^2, Coefficient values for 6 variable Ridge and Lasso Regression models