Average Points Scored by NBA Players: 2018-2019

General Linear Models I

Cesar Rene Pabon Bernal
City University of New York, Hunter College

Thursday, December 12th, 2019

Outline

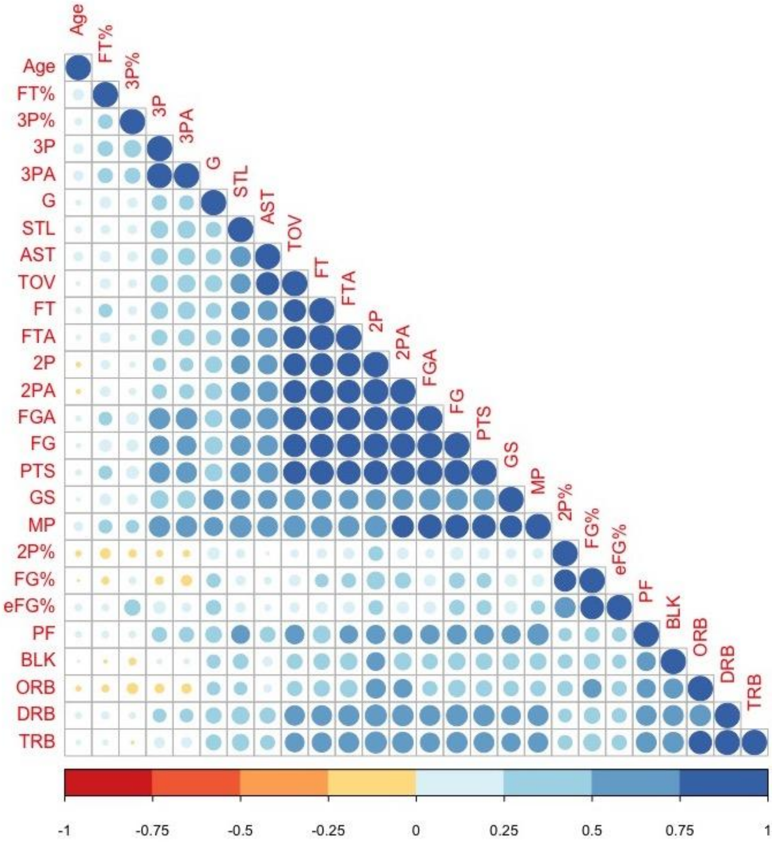
- Introduction
- Multiple Regression
  - Model Variable Selection
    - Cross Validation Using Subset Grouping: Stepwise forward & backward Regression
  - Final Model Diagnostics
- Simple Regression
  - Model Variable Selection
  - Relations Between Variables
  - Inferences in Regression and Correlation Analysis
  - Model Diagnostics and Possible Remedial Measures

Introduction

- Data Source: basketball-reference.com
  - 2018-19 NBA Player Stats: Average Per Game
  - 30 variables and 629 observations at each variable
- Using variable selection, we isolated the strongest predictor variables exploring both multiple and simple regression models to determine:
  - **Which variables best explain average number of points scored by an individual NBA player per game?**

Multiple Regression

Multiple Regression



$$Y_{14} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_{14} X_{14}$$ [1]

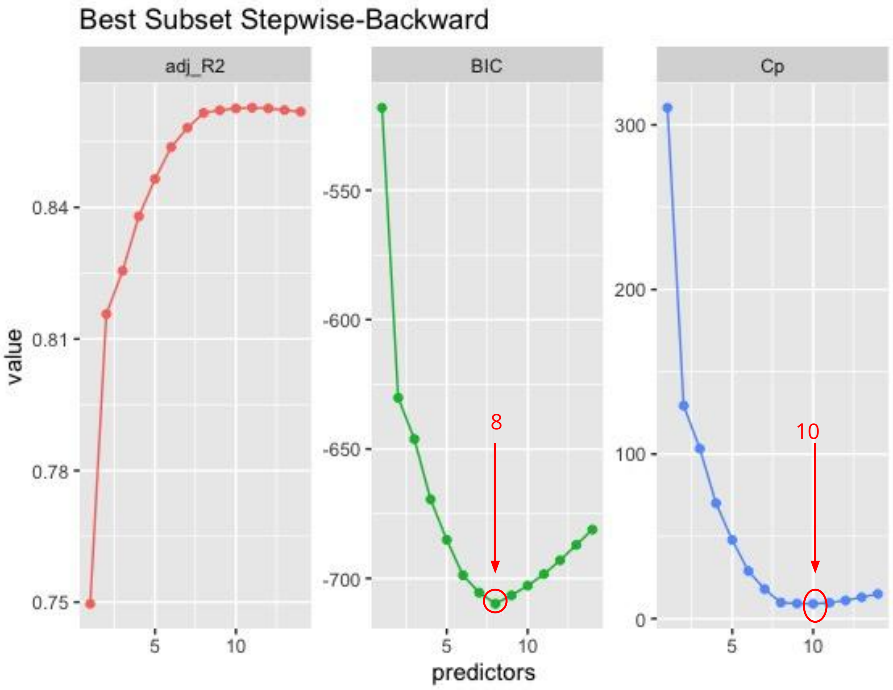
```
Call:
lm(formula = PTS ~ Age + G + GS + MP + ThreePoint_Perc + TwoPoint_Pe
+
    FT_Perc + ORB + DRB + AST + STL + BLK + TOV + PF, data = df3)

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
```

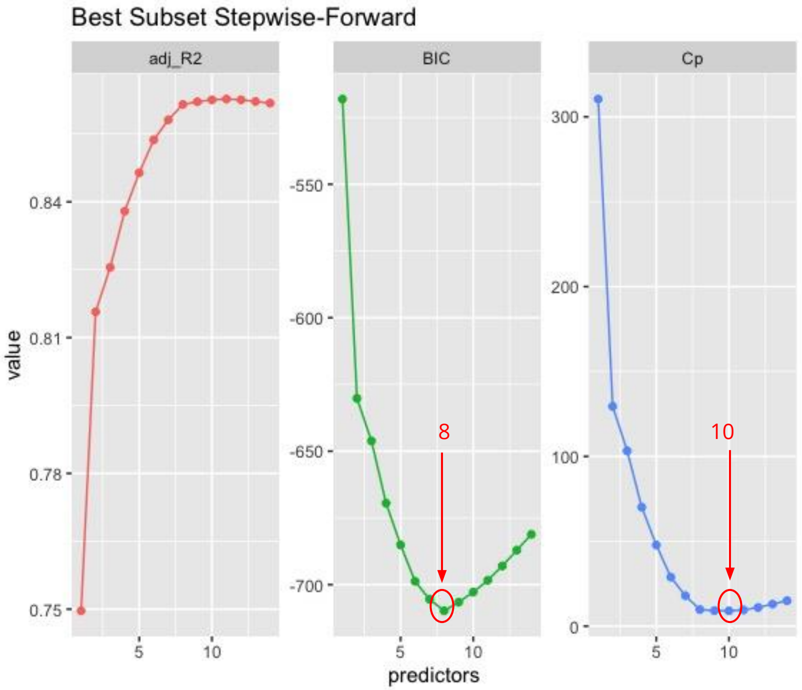
$$\frac{11}{14}$$

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.

Multiple Regression

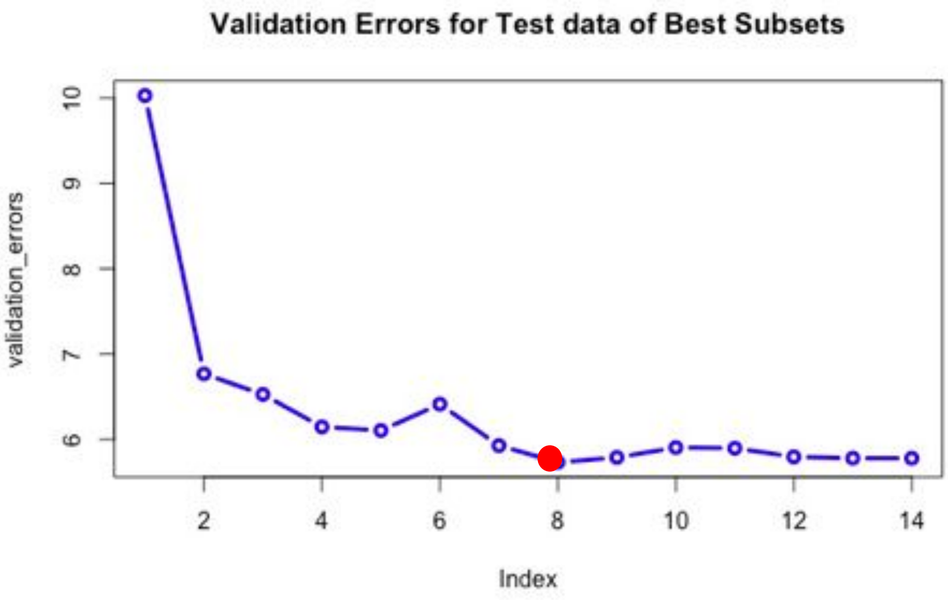
- Cross Validation Using Subset Grouping: Stepwise forward & backward Regression
  - 60% of the data was used for training

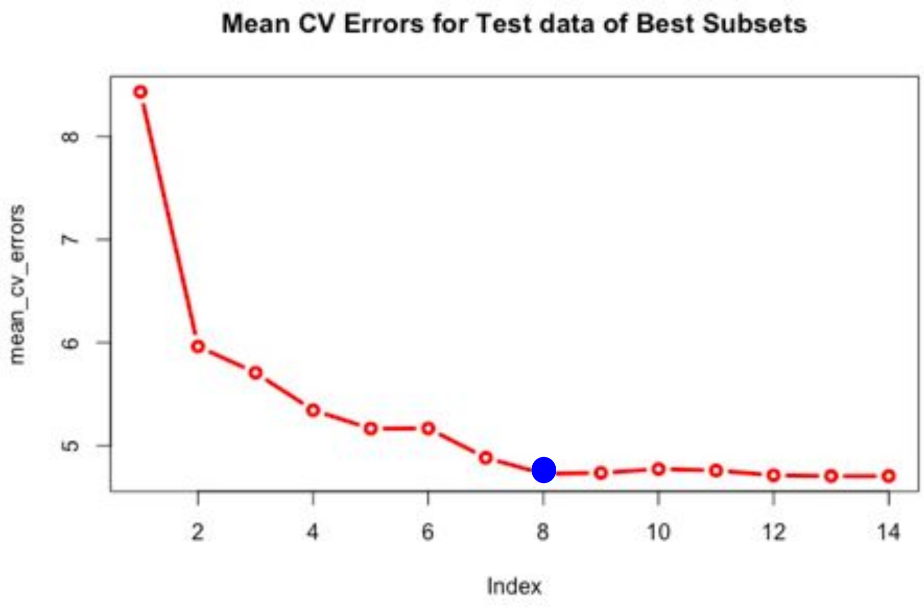




Multiple Regression

- Cross Validation Using Subset Grouping: Stepwise forward & backward
  - 40% of the data was used for testing





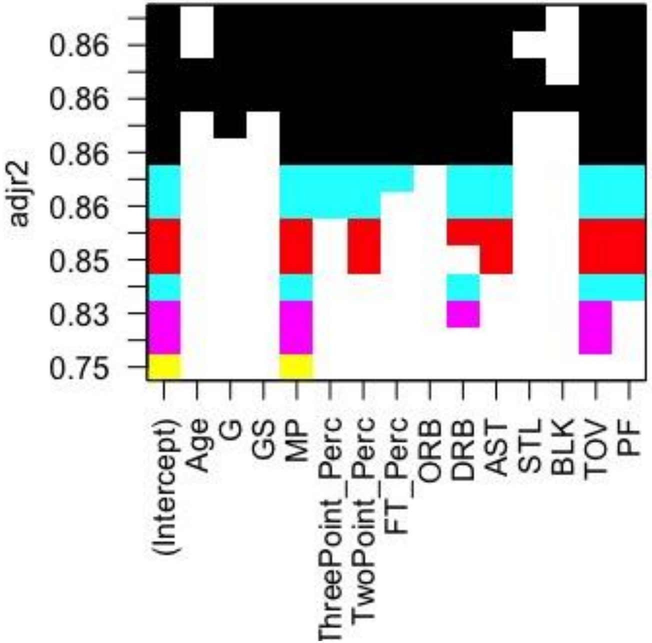
Multiple Regression

- 8 variable model (before testing for multicollinearity)

$$Y_8 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$

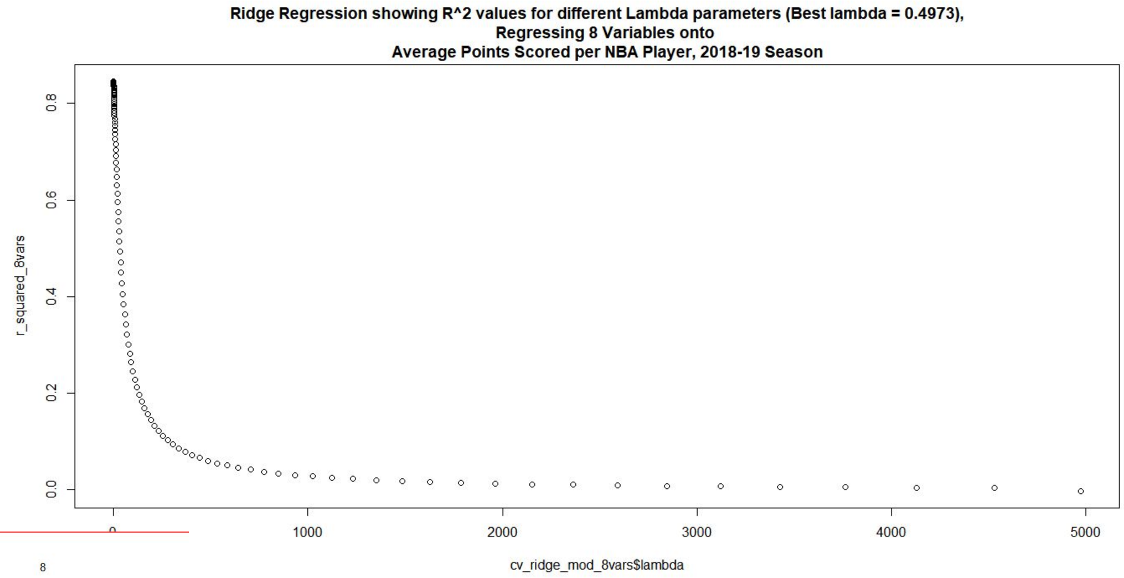
```
Call:
lm(formula = PTS ~ MP + ThreePoint_Perc + TwoPoint_Perc + FT_Perc +
    DRB + AST + TOV + PF, data = df3)

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
```

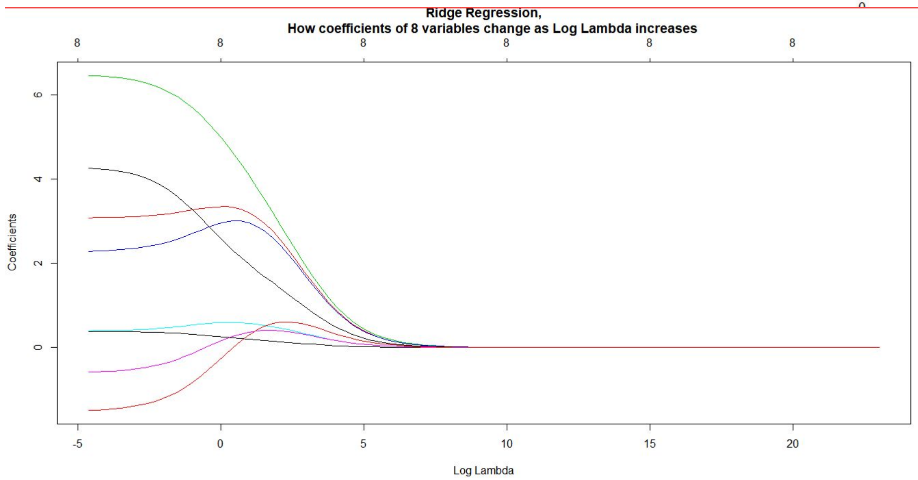
Adj. R^2 of 8 Variable Model vs Others



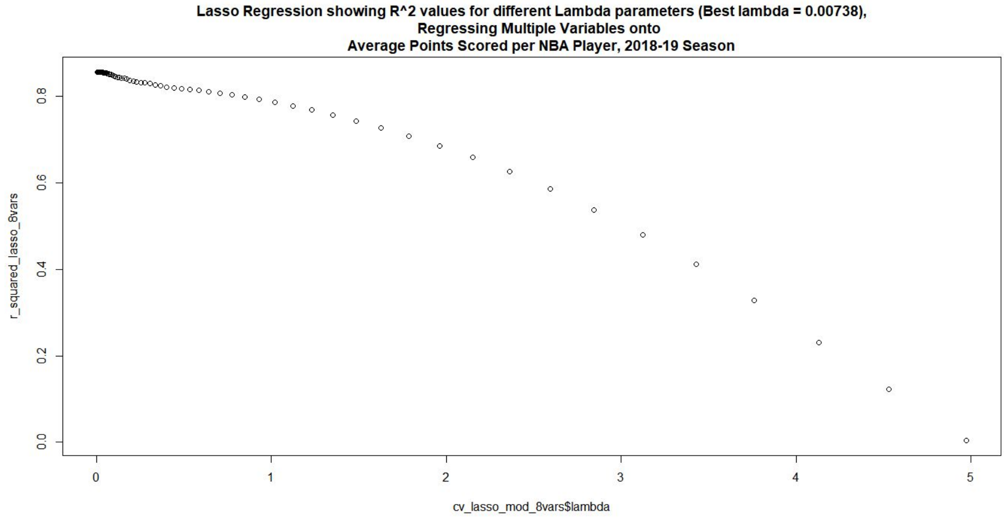
Multiple Regression

- <u>RIDGE (8 variable model)</u>
  - Reduce coefficients close to zero
  - Helps understand which variables are most important
    - Hurts interpretability
  - Looking for best lambda between 10^10 and 10^-2. Best is 0.4973
  - R^2 = 86.22%

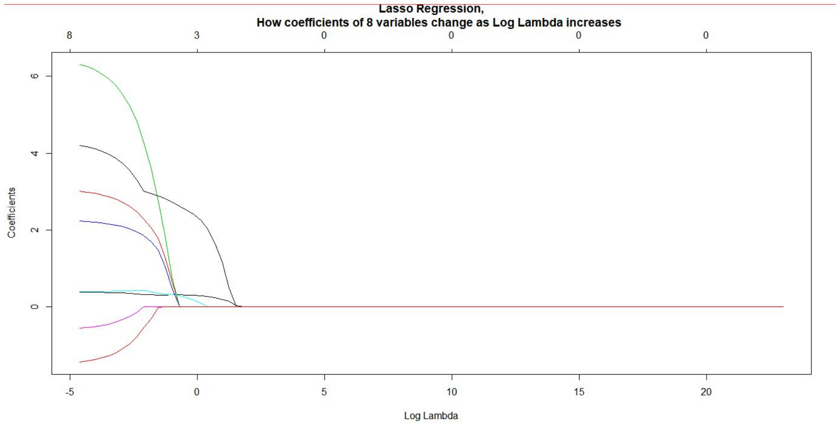




Multiple Regression



- <u>LASSO (8 variable model)</u>
  - Could reduce coefficients to exactly zero
  - Helps understand which variables are most important
    - Hurts interpretability
  - Looking for best lambda between 10^10 and 10^-2. Best is 0.00738
  - R^2 = 86.22%



Comparing Coefficients of Ridge, 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

- All variables must be standardized
  - Standardized variables are transformed to have a mean of 0 and a standard deviation of 1

1. 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.

1. Results are less interpretable since the coefficients are modified
   - The focus is on improving accuracy, e.g., R^2, as opposed to improving interpretability.

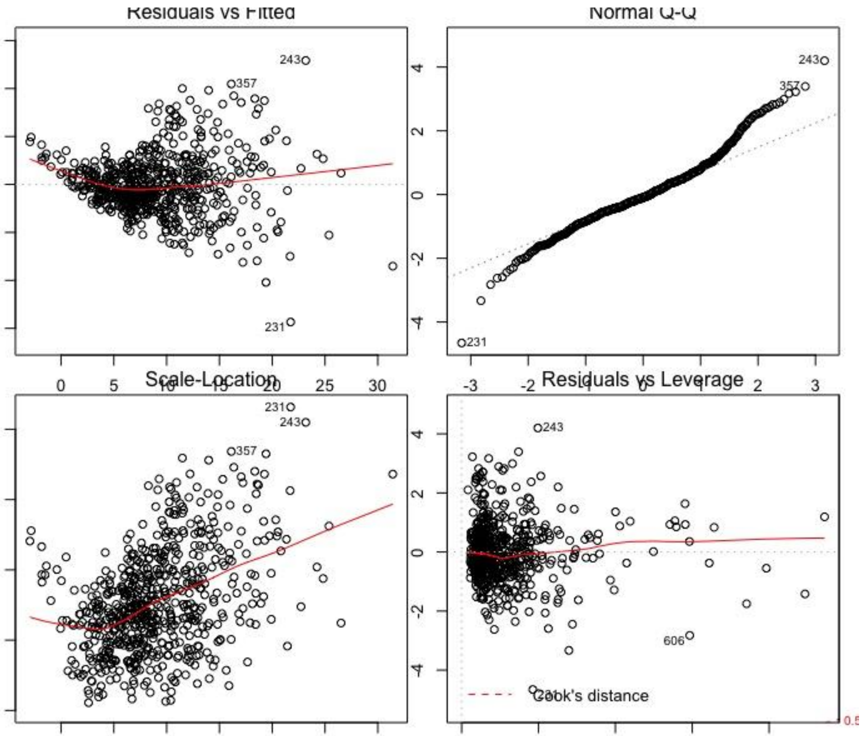
Ridge and lasso regression models were run for both 8 variable and 6 variable models. Cross validation was used to find the lambda (shrinkage value) term between 10^{10} and 10^{-2} that produces the highest R^2 value.

- 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.

Multiple Regression

DIAGNOSTICS FOR 6 VARIABLE MODEL

$$Y_8 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$



Although we've reduced the number of variables to 6, there are several assumptions to evaluate for linear regression to be suitable.

Top-left: Residuals vs Fitted plot shows our residuals do not have **constant error variance**, as noted by the rightward fan shape. Solution: Use a transformation on the data.

Top-right: Normal Q-Q plot looks at whether the residuals follow a **normal distribution** and mostly fall on the diagonal line. Towards the right side, there is clear violations of the assumption. Solution: Use a transformation on the data.

Bottom-left: The Scale-Location plot shows that as the fitted X values get larger, the standardized residuals are greatly affected. There is **non-independence of error terms**. Solution: Look to see which variables are related.

Bottom-right: The Residuals vs Leverage plot shows where **outliers, leverage points, and influential points** are located. Solution: Look at each of these observations to determine whether they should be kept, removed, or have their value(s) fixed.

Simple Regression

Simple Linear Regression: Variable Selection

- Do **individual** relationships exist between Points Scored per Game (PTS) and each of the six variables in our multiple regression model (MRM)?
- Looking at PTS vs. two_pt_perc , for example, we can see that there is not evidence of a linear relationship.
- Similarly, the other variables in our MRM had a an adj. $R^2 < 0.5$.
- From this, we determined while the predictors in the MRM explained much of the variation in Y, alone they were not significant.
- Instead, we looked at the remaining variables in the data set and ran simple regression on each to determine those with the highest adj. R^2 .

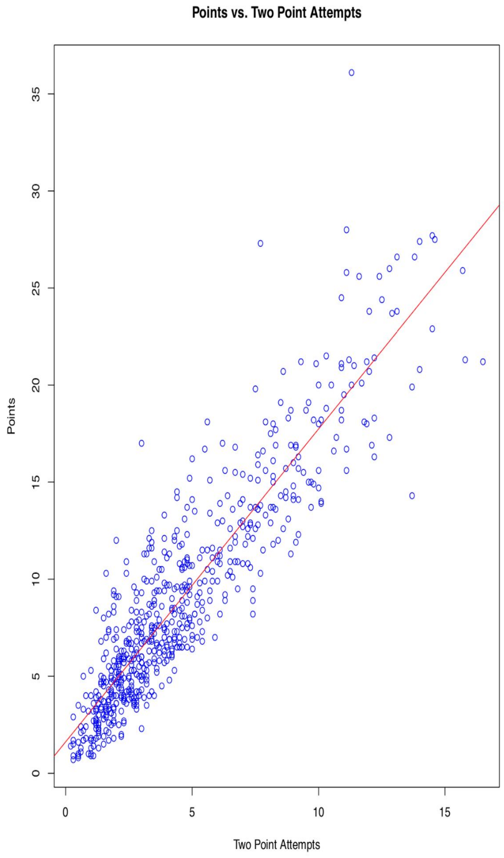
```
#Linear Regresson on two_pt_perc (adj. r^2= .03802)
df_linear6 = lm(PTS ~ two_pt_perc)
summary(df_linear6)
```

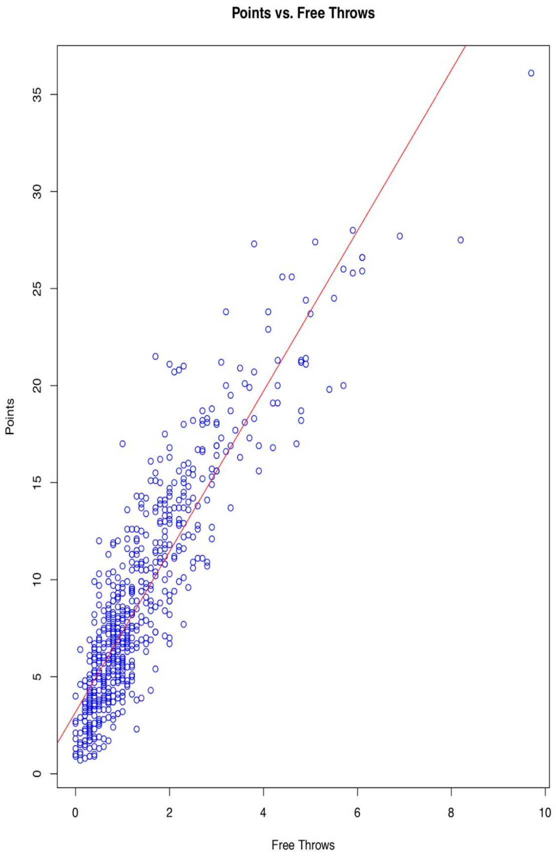
```
##
## Call:
## lm(formula = PTS ~ two_pt_perc)
##
## Residuals:
##      Min       1Q  Median      3Q     Max
## -11.477   -4.125  -1.378   2.989  26.707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.588      1.278   2.025   0.0433 *
## two_pt_perc   12.889      2.536   5.082 4.95e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.652 on 627 degrees of freedom
## Multiple R-squared:  0.03955,  Adjusted R-squared:  0.03802
## F-statistic: 25.82 on 1 and 627 DF,  p-value: 4.946e-07
```

Simple Linear Regression: Strongest Predictors

- The simple regression models that yielded the highest adj. R^2 were:
  - Average Two Point Field Goals Attempted per Game (2PA) Vs. PTS
    - Adjusted R^2 = .8173
  - Average Free Throws per Game (FT) Vs. PTS
    - Adjusted R^2 = .7965
- For the remainder of the presentation, these are the linear models we will be exploring.

Simple Linear Regression: Scatterplots





Simple Linear Regression

The least square estimates of the regression coefficients when the dependent variable is Average Points per Game (PTS) and the independent variable is:

- Average Two Point Field Goals Attempted per Game (2PA)

  $b_0 = 1.619534$ $b_1 = 1.612499$

  $\hat{Y} = 1.619534 + 1.612499X$

- Average Free Throws per Game (FT)

  $b_0 = 3.179097$ $b_1 = 4.133236$

  $\hat{Y} = 3.179097 + 4.133236X$

Simple Linear Regression: Confidence Intervals for β_0 and β_1

Separate 95% Confidence Intervals for β_0 and β_1 when the independent variable is:

- Average Two Point Field Goals Attempted per Game (2PA)

  $1.2856 \leq \beta_0 \leq 1.9535$
  $1.5528 \leq \beta_1 \leq 1.6722$

- Average Free Throws per Game (FT)

  $2.8721 \leq \beta_0 \leq 3.4861$
  $3.9685 \leq \beta_1 \leq 4.2969$

Confidence Band for the Reg. Lines and Prediction

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)

F-test for Lack of Fit

	H_0:	H_a:	F	F*	Conclude
2PA	$E(Y)= B_0+B_1X$	$E(Y)\neq B_0+B_1X$	1.25	1.503	H_a
FT	$E(Y)= B_0+B_1X$	$E(Y)\neq B_0+B_1X$	1.28	3.469	H_a

- For testing the appropriateness of a linear regression relation, we can use the F-test for lack of fit.
- Both tests conclude H_a
- Therefore, there is a linear association in our models.

Simultaneous Confidence Intervals: β_0 and β_1

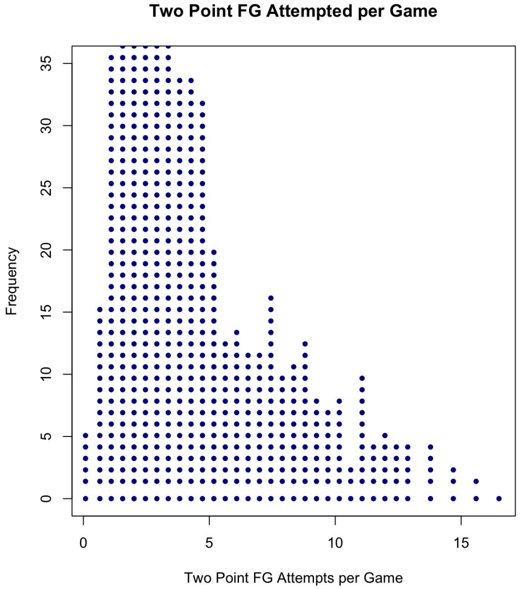
	Simultaneous Confidence Intervals for β0 and β1
2PA	$1.237 \leq \beta_0 \leq 2.002$
	$1.54 \leq \beta_1 \leq 1.681$
FT	$2.828 \leq \beta_1 \leq 3.530$
	$3.946 \leq \beta_1 \leq 4.320$

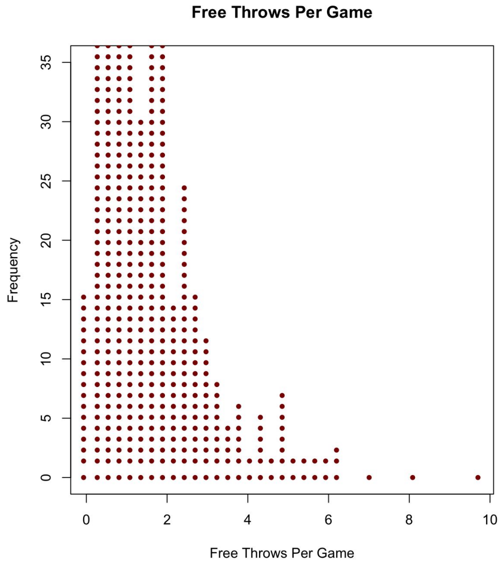
Simultaneous Confidence Intervals and Prediction

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)

Is the simple linear regression model appropriate?

- To test the validity of our simple regression model and to determine if transformations should be made, we must test the assumptions of a simple linear regression model.
- We first look at predictors 2PA and FT to determine if there are any **extreme** outliers.

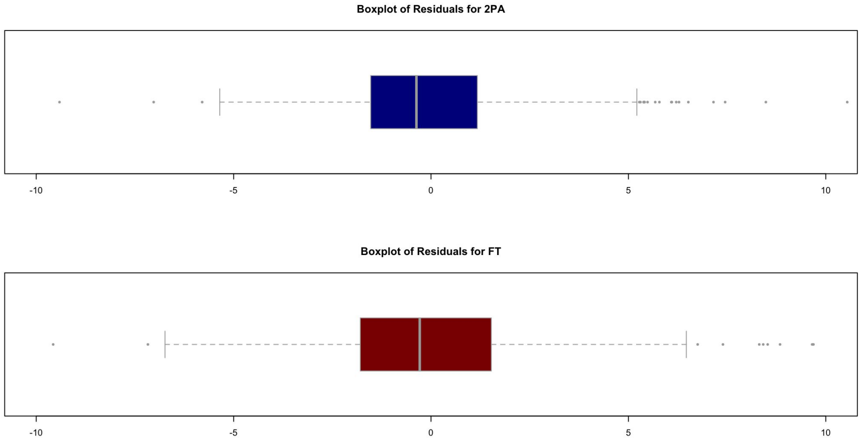


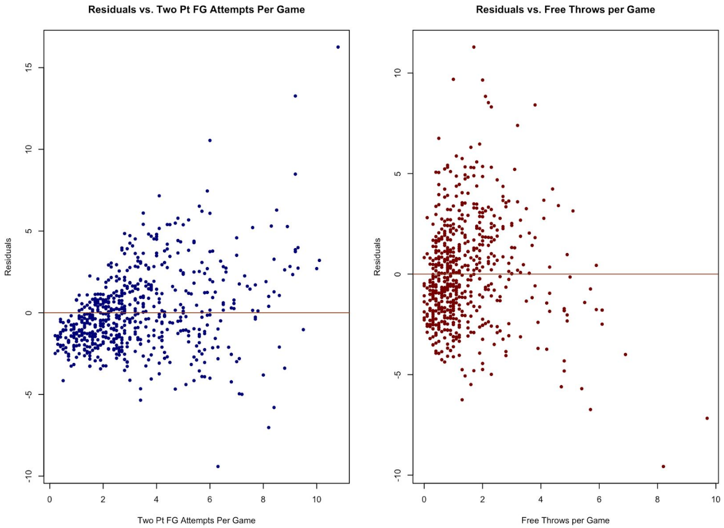


Model Diagnostics: Tests for Linear Association

- T-Test for Linear Association:
  - **H_0: $\beta_1 = 0$ H_a: $\beta_1 \neq 0$**
  - $T_{(.975,\ 629)} = 1.964$
  - $T^* = 53.0192$
  - Since $T^* > T_{(.975,\ 629)}$ we can reject H_0 and conclude with 95% confidence that there is a linear association between Two Point Field Goals Attempted and Average Points Scored per Game.
- Similar conclusions can be made FT.
- These same conclusions were also made by performing the F-test for linear association.
- **The regression functions for 2PA and FT are in fact linear.**

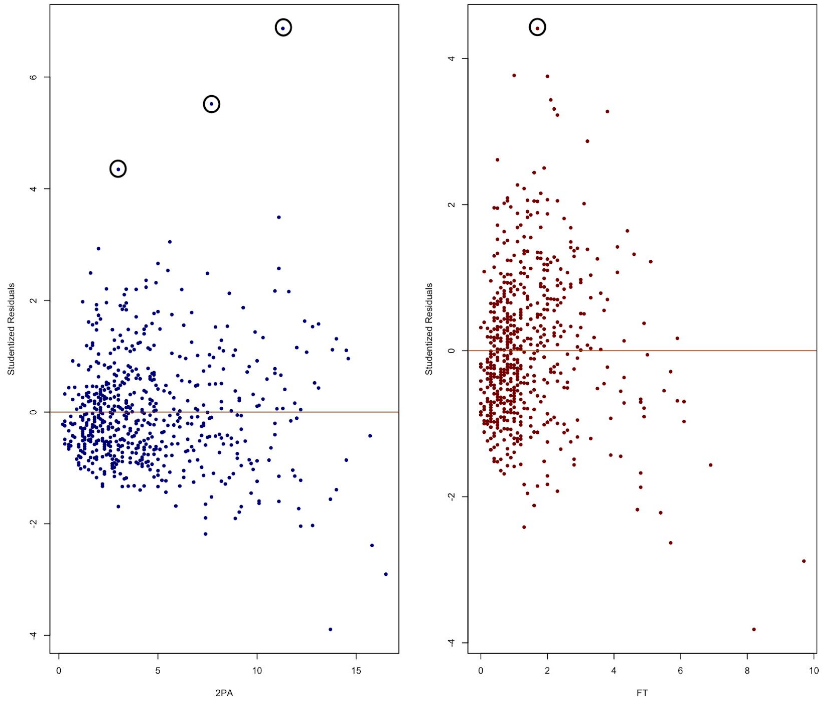
Model Diagnostics: Do Error Terms Have Constant Variance?



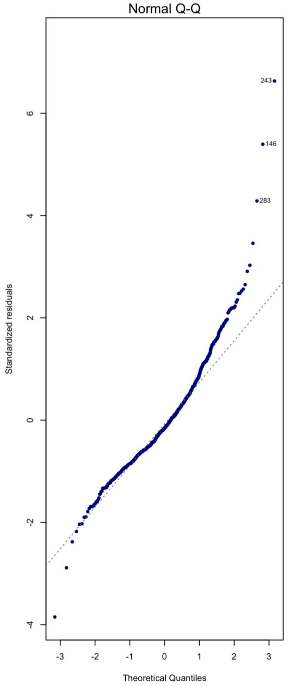


Model Diagnostics: Outliers?

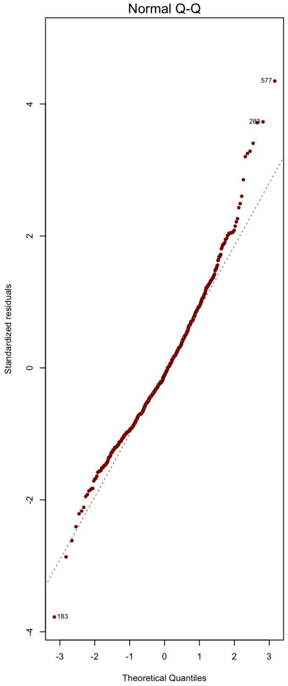
- Here we've used similar residual plots as we used for the detection of constant error variance however we've standardized the residuals to make it easier to see outliers.
- The circled points represent residual outliers (any residual value > |4|).
- Both models contain outliers, with FT containing two less than 2PA.



Model Diagnostics: Are Error Terms Normal?



2PA



FT

- To test the assumption of normality, we obtained the coefficient of correlation between the ordered residuals and their expected values under normality.

H_0: The residuals are normally distributed

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

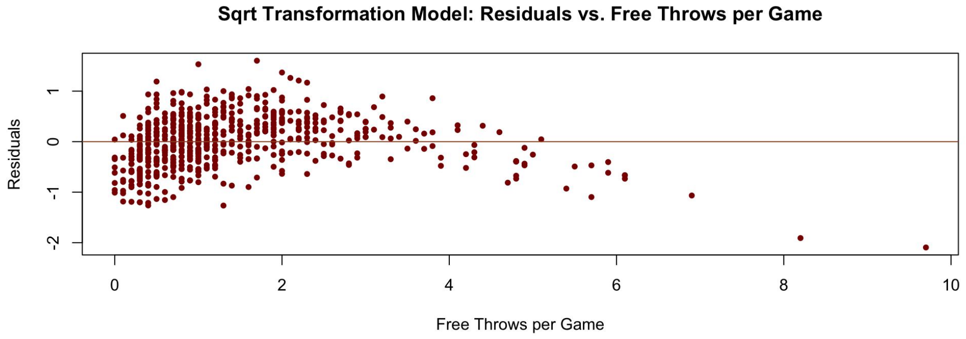
- After obtaining **$r = .965$ and $r_{crit}=.9976$** for 2PA, we can reject the null hypothesis since $r < r_{crit}$ and conclude that the residuals for the 2PA model are not normally distributed.
- This same conclusion was true for FT.

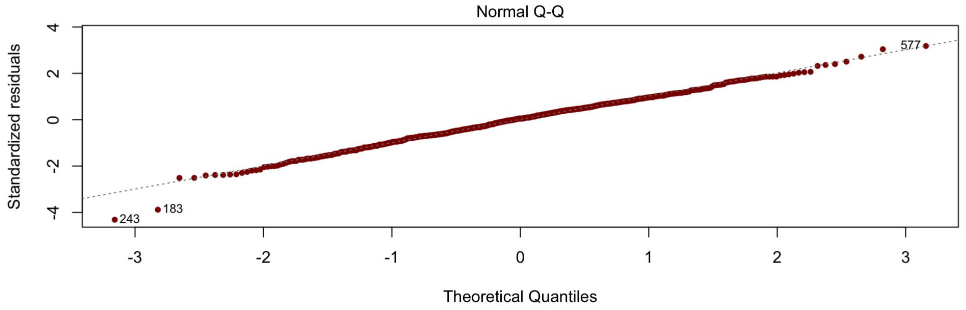
Remedial Measures

- Since all three of our models lack constance 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

$$Y' = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- While the square root transformation seems to correct for normality, the error terms still appear to be inconstant.
- Additionally the adj. R^2 decreased from .7965 to .7097 using the transformed model.





Conclusion

- For simple linear regression, some of the model assumptions were violated in both of the models:
    - Error variance was not constant
    - Residuals were not normally distributed
    - 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.