Finding the Next NBA Superstar: How to Predict a Player's Performance?

Hongyu Zhang

What factors determine an excellent basketball player? This is an interesting and significant question that concerns coaches, basketball managers, as well as many basketball fans. For instance, Michael Jordan, the well-known NBA superstar, had 6 NBA Championships and was selected to All Star 14 times[1]. In his career, he scored average 30.1 points per game[2] and was considered to be one of the greatest players in NBA history[3]. Similarly, Kobe Bryant, with 25.0 average points per game, won 5 NBA Championships and was selected to All Star 18 times[4]. Can we find any similarities between these two players that made them superstars in NBA history? Inspired by this intriguing problem, I want to find out whether there is a way to predict a player's performance using a variety of factors. The goal of this study is to construct a statistical model that uses players' statistics to explain his performance in game. Solving this particular problem will give us a clue about the question stated at the beginning: for basketball players, what factors influence their performance?

Introduction

The easiest way to evaluate a player's performance is to see how many points in average he scored per game. The higher his average points are, the better the player is. There are only 2 ways to score points in a game: free throw and field goal. Free throws result from a foul in opposing team. In a free throw, a player attempts to score points behind the free throw line without opposed interference[5]. A field goal is "a basket scored on any shot or tap other than a free throw"[6]. There are also many other moves in a basketball game, like rebounds, steals, blocks and so on. All of these data are purely numerical and objective, and they can be easily found from the official website of NBA or other sources. My plan is to construct a linear regression model to explain players' average points scored by using players' data like height, weight, field goal scored, free throw scored and so on. I plan to first construct a relatively simple model that fits using few regressors. Then I hope to find the best model from many variables using stepwise search.

Data Collection

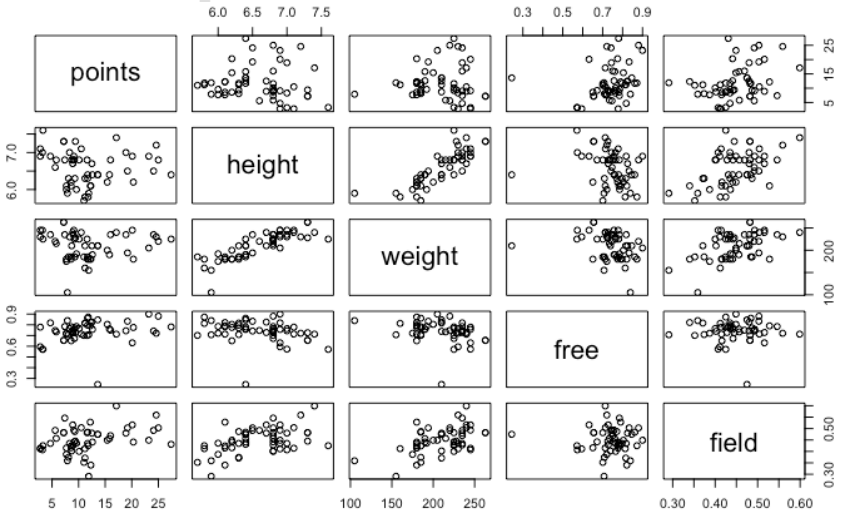
I was able to find this dataset online from Cengage[7]. This dataset referred *The official NBA basketball Encyclopedia*, Villard Books, which is an official book that records all NBA players' statistics. The dataset contains 54 players' data in 5 categories: column 1 is **height** (in feet), column 2 is **weight** (in pounds), column 3 is **field** (percent of successful field goals out of 100 attempted), column 4 is **free** (percent of successful free throws out of 100 attempted) and column 5 is **points** (average points scored per game). Here, the response variable is **points**, since it is the variable that I want to predict. I can read the data and extract some useful information about this dataset. Below are the first 6 rows of this dataset and basic summary statistics of these 5 variables.

	X1	X2	X3	X4	X5
1	6.8	225	0.442	0.672	9.2
2	6.3	180	0.435	0.797	11.7
3	6.4	190	0.456	0.761	15.8
4	6.2	180	0.416	0.651	8.6
5	6.9	205	0.449	0.900	23.2
6	6.4	225	0.431	0.780	27.4

```
> summary(points)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.80    8.15   10.75   11.79   13.60   27.40 
> summary(height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.700   6.225   6.650   6.587   6.900   7.600 
> summary(weight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  105.0   185.0   212.5   209.9   235.0   263.0 
> summary(field)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2910  0.4153  0.4435  0.4491  0.4835  0.5990 
> summary(free)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2440  0.7130  0.7535  0.7419  0.7953  0.9000 
```

To better understand this dataset, I can also visualize it by constructing scatter plot matrix of the average points scored and each of the regressors.

```
> # Scatter plot matrix of the response variable and each of the regressors
> pairs(points~height+weight+free+field,cex.labels=2)
```



scatter plot matrix

From this dataset, we can see that there might exist some relationship between average points scored and the other 4 regressors, although the variation is relatively high. Next, I am going to build linear regression models over this dataset and analyze this data.

Simple Approach

From the scatter plot, we can see that there probably exists some relationship between average points scored and percent of successful field goals and free throws. Based on our intuition, this also makes sense: if a player has high successful field goals rates and free throw rates, then he will probably have high average points per game because of his accuracy, and vice versa.

First, I need to make sure this model does not have multicollinearity issue. In the design matrix **X**, the first column is column vector 1, and the second column is the percent of successful field goals, and the third column is the percent of successful free throws. Using the eigen function in R, I calculate eigenvalues of $\mathbf{X}^T\mathbf{X}$:

94.7953878 0.3816589 0.1344273

The condition number of $\mathbf{X}^T\mathbf{X}$ is:

$$K = \frac{\lambda_{max}}{\lambda_{min}} = \frac{94.7953878}{0.1344273} = 705.1798$$

This condition number is not very large, which indicates that it is numerically stable to calculate the inverse of $\mathbf{X^TX}$. I also calculate the result of $\mathbf{X^TX} * (\mathbf{X^TX})^{-1}$:

$$\begin{bmatrix} 1.000000e^{+00} & 1.199041e^{-14} & 0 \\ 7.105427e^{-15} & 1.000000e^{+00} & 0 \\ 7.105427e^{-15} & -7.327472e^{-15} & 1 \end{bmatrix}$$

We can see that $\mathbf{X^TX} * (\mathbf{X^TX})^{-1}$ is very close to the identity matrix. These evidences show that our design matrix does not have multicollinearity issue. We can then start build the model.

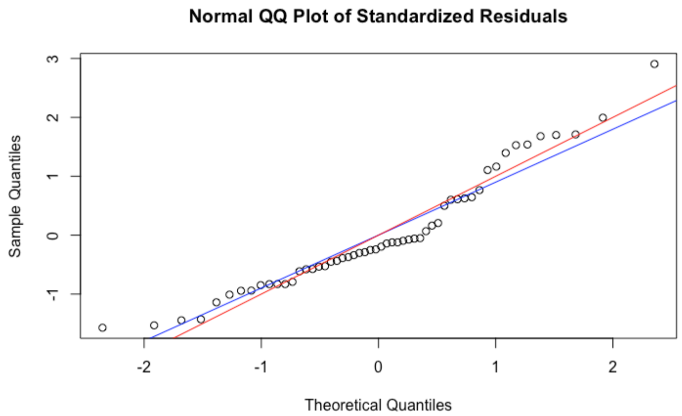
The model can be described as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where y is the average points scored, x_1 is the percent of successful field goals, and x_2 is the percent of successful free throws. In matrix notation, we have: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ has dimension 54×1, $\mathbf{X}$ has dimension 54×3, $\boldsymbol{\beta}$ has dimension 3×1, and $\boldsymbol{\varepsilon}$ has dimension 54×1. Using R, I obtain $\hat{\boldsymbol{\beta}}$:

$$\begin{bmatrix} -15.27738 \\ 35.82503 \\ 14.79905 \end{bmatrix}$$

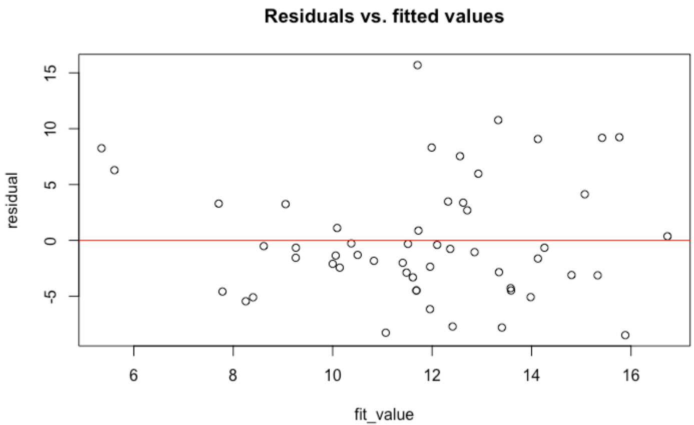
and SS_{res} is 1516.422 and $\widehat{\sigma^2}$ (also MS_{res}) is 29.16196. The fitting model becomes:

$$\hat{y} = -15.277 + 35.825x_1 + 14.799x_2.$$

Next, I need to investigate the normal assumption and homogeneous variance assumption. I plot the normal QQ plot of standardized residues:



From this graph, we observe that the blue line (straight line of standardized residuals) is close to the red line (reference line). This indicates that our normal assumption is good.
Also, I plot the residuals vs. fitted values graph:



In this graph, we observe roughly a random scatter of points around the horizontal axis. Therefore, our homogeneity assumption is good. Since the normal assumption and the homogeneity assumption both hold, our model is valid.

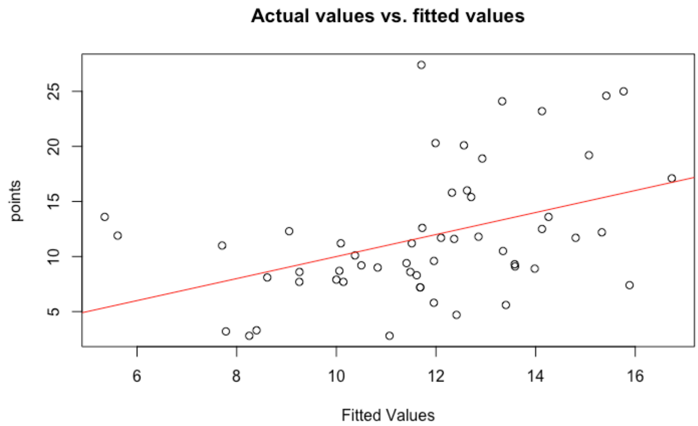
We have obtained the fitting model, and the next step is to conduct hypothesis tests about our estimated coefficients. Using R, the p-value for the hypothesis test of β_1 being 0 is 0.008606811, and the p-value for the hypothesis test of β_2 being 0 is 0.0509969. Because both p-values are relatively small, we do not have enough statistical evidence to support the null hypotheses. Therefore, we reject the claims that β_1 is 0 and β_2 is 0, and these two estimated coefficients are statistically significant.

I can verify my results by comparing them to the results obtained by lm function in R:

```
> m1 <- lm(points~field+free)
> summary(m1)

Call:
lm(formula = points ~ field + free)

Residuals:
   Min     1Q Median     3Q    Max 
-8.486 -3.267 -1.178  3.281 15.694 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -15.277      8.244  -1.853  0.06966 . 
field         35.825     13.247   2.704  0.00928 **
free          14.799      7.480   1.978  0.05330 . 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.453 on 51 degrees of freedom
Multiple R-squared:  0.1779,	Adjusted R-squared:  0.1456 
F-statistic: 5.516 on 2 and 51 DF,  p-value: 0.00678
```

My results and the results computed by lm function are almost exactly the same (considering rounding errors). Also, the overall F-test for this model is highly statistically significant because of the small p-value 0.00678.

Currently, my model is $\hat{y} = -15.277 + 35.825x_1 + 14.799x_2$. To find out whether this model fits the dataset well, we can plot the actual values vs. fitted values graph and add a straight fit line for reference:



In this plot, we observe that roughly most of the points are near around the straight fit line, which indicates that our model fits the dataset well in general. However, some points deviate from the straight fit line very much, and this provides evidence that this model is probably not the best model to fit our data. In the next step, I plan to use the stepwise approach to obtain the best model from these variables.

Stepwise Search to Find the Best Model

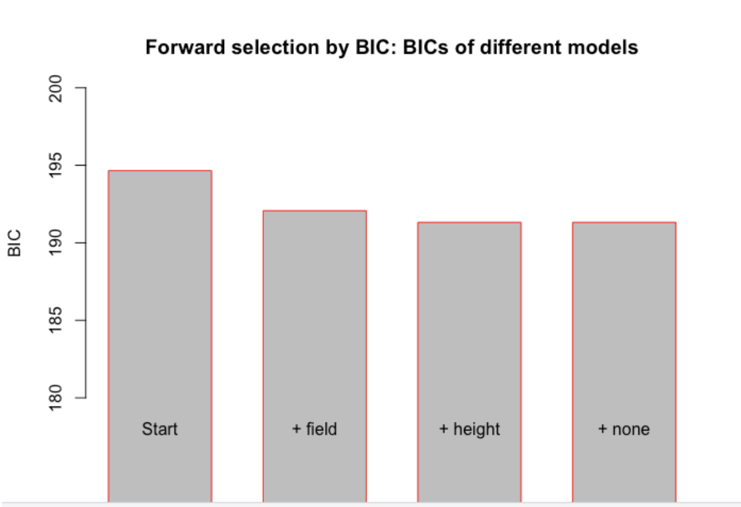
Given our dataset, there are 4 possible regressors: **height**, **weight**, **field** (percent of successful field goals) and **free** (percent of successful free throws). We can also add one interaction term between percent of successful field goals and percent of successful free throws to discover the association between these 2 variables. Our goal is to find the best fitting model built from these regressors. We need to be cautious about what variables we want to include in the model, because some of the variables may be highly correlated. For example, there might exist a positive correlation between height and weight, since in general taller people tend to also have higher weights. We can examine this correlation using R and the correlation between height and weight is 0.834324. This result verifies my conjecture: height and weight in this dataset are highly correlated. They carry similar information about the average points scored, and our model might not differentiate their separate effects on the response variable. Because of the high correlation, we cannot have both height and weight as regressors in our model, otherwise this will cause multicollinearity issue. In order to find the best set of variables, we can use the stepwise search approach.

To obtain the best set of variables, we can use forward selection based on Bayesian Information Criterion (BIC). The reason I use BIC is that I want to obtain the best fitting model using fewest variables. BIC is defined by:

$$\text{BIC} = -2\,log\text{L}(\hat{\theta}) + \text{K}\,log\text{n}.$$

Compared to Akaike's Information Criterion (AIC) which penalizes the complexity of model by a factor of 2, BIC penalizes the complexity of the model by a factor of *log*n, where n is the sample size. Using R, I do the forward selection process and graph the bar plot of the BICs of different models:

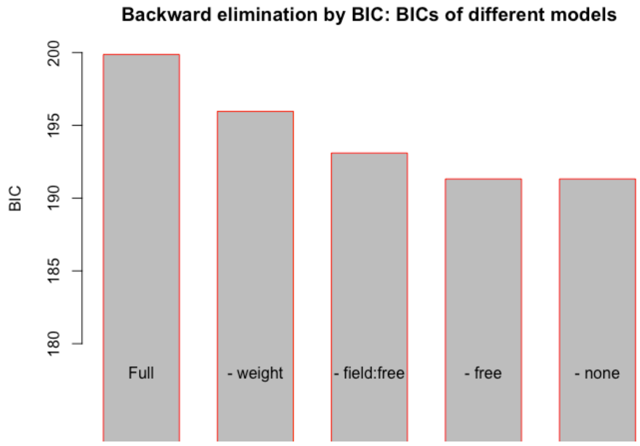
```
Start:  AIC=194.66
points ~ 1

         Df Sum of Sq    RSS    AIC
+ field   1   211.668 1632.8 192.07
<none>                1844.5 194.66
+ free    1   110.580 1733.9 195.31
+ height  1     8.758 1835.7 198.39
+ weight  1     0.179 1844.3 198.65

Step:  AIC=192.07
points ~ field

         Df Sum of Sq    RSS    AIC
+ height  1   137.066 1495.7 191.32
+ free    1   116.375 1516.4 192.06
<none>                1632.8 192.07
+ weight  1    85.725 1547.1 193.14

Step:  AIC=191.32
points ~ field + height

         Df Sum of Sq    RSS    AIC
<none>                1495.7 191.32
+ free    1    59.978 1435.8 193.10
+ weight  1     0.060 1495.7 195.31
```



From the forward selection result on the left and the bar plot on the right, we can see that first I start with regressor 1 and the model has BIC 194.66. Then I find out that adding the variable **field** can reduce the BIC to 192.07, so I add **field** to the model. Then I find out that adding the variable **height** can reduce the BIC to 191.32, so I add **height** to the model. In the end, I find out that there is no way to reduce the BIC further by adding more variables, so forward selection ends and the final model I obtain has 2 regressors: **height** and **field**.

Besides forward selection, I also use backward elimination based on BIC. I start with all potential regressors in the model, **height**, **weight**, **field**, **free**, and the interaction term between **field** and **free**. In each step, I remove the regressor that will make the new model have the lowest BIC. I continue this process until BIC increases or all variables have been removed. Using R, I obtain the following results:

```
Start:  AIC=199.87
points ~ height + weight + field + free + field:free

             Df Sum of Sq    RSS    AIC
- weight      1     1.964 1405.9 195.96
- field:free  1    30.564 1434.5 197.04
- height      1    41.906 1445.9 197.47
<none>                    1404.0 199.87

Step:  AIC=195.96
points ~ height + field + free + field:free

             Df Sum of Sq    RSS    AIC
- field:free  1    29.821 1435.8 193.10
- height      1    67.418 1473.3 194.50
<none>                    1405.9 195.96

Step:  AIC=193.1
points ~ height + field + free

         Df Sum of Sq    RSS    AIC
- free    1    59.978 1495.7 191.32
- height  1    80.668 1516.4 192.06
<none>                1435.8 193.10
- field   1   298.073 1733.8 199.30

Step:  AIC=191.32
points ~ height + field

         Df Sum of Sq    RSS    AIC
<none>                1495.7 191.32
- height  1    137.07 1632.8 192.07
- field   1    339.98 1835.7 198.39
```



I first start with the model with all potential regressors, which has BIC 199.87. I find out that removing **weight** will give me the model with the lowest BIC 195.96, so I remove **weight** from the model. Then I find out that removing the interaction term between **field** and **free** will give me the model with the lowest BIC 193.10, so I remove this interaction term. In the next step, removing **free** will give me the lowest BIC 191.32, so I remove **free**. In the end, there is no way to reduce the BIC further by removing regressors, so backward elimination ends, and I obtain the final model with 2 regressors: **height** and **field**, which is the same model I obtain in forward selection process.

In both forward selection and backward elimination, I get the same results. This provides further evidence that the model with 2 regressors **height** and **field** is our final model. Our final model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where y is **points**, x_1 is **height**, and x_2 is **field**.

Analysis of Final Model

In matrix notation, our final model is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ has dimension 54×1, $\mathbf{X}$ has dimension 54×3, $\boldsymbol{\beta}$ has dimension 3×1, and $\boldsymbol{\varepsilon}$ has dimension 54×1.

First, I check that $\mathbf{X}^T\mathbf{X}$ is not a singular matrix. The result of $\mathbf{X}^T\mathbf{X} * (\mathbf{X}^T\mathbf{X})^{-1}$:

$$\begin{bmatrix} 1.000000e^{+00} & 7.105427e^{-15} & -2.842171e^{-14} \\ -2.842171e^{-14} & 1.000000e^{+00} & 0.000000e^{+00} \\ -6.217249e^{-15} & -5.329071e^{-15} & 1.000000e^{+00} \end{bmatrix}$$

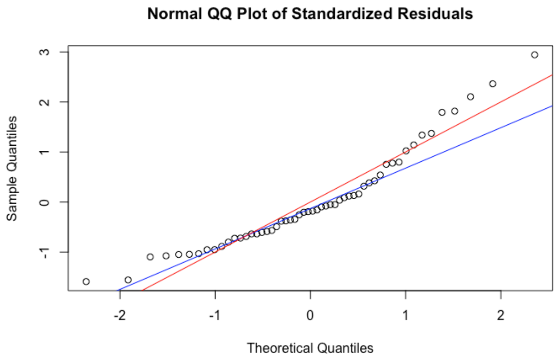
We can see that $\mathbf{X}^T\mathbf{X} * (\mathbf{X}^T\mathbf{X})^{-1}$ is very close to the identity matrix. This shows that our design matrix does not have multicollinearity issue. We can start to construct the model. Using R, I obtain $\hat{\boldsymbol{\beta}}$:

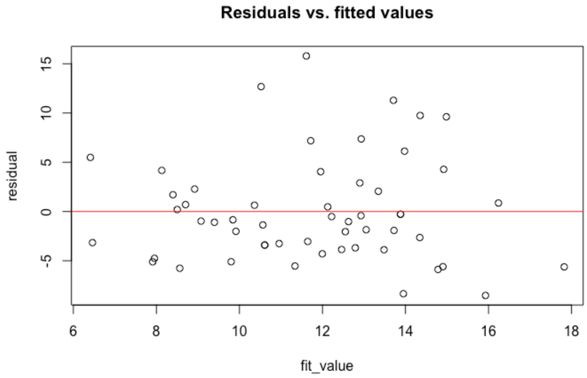
$$\begin{bmatrix} 15.209793 \\ -4.034628 \\ 51.562276 \end{bmatrix}$$

and SS_{res} is 1495.732 and $\widehat{\sigma^2}$ (also MS_{res}) is 28.76407. The fitting model becomes:

$$\hat{y} = 15.210 - 4.035x_1 + 51.562x_2.$$

In order to investigate the normal assumption and homogeneous variance assumption, I plot the normal QQ plot of standardized residues and the residuals vs. fitted values graph:





From the QQ plot, we observe that the blue line (straight line of standardized residuals) is relatively close to the red line (reference line). This indicates that our normal assumption is good. In the residuals vs. fitted values graph, we observe a random scatter of points around the horizontal axis. Therefore, our homogeneity assumption is good. Since the normal assumption and the homogeneity assumption both hold, our final model is valid.

After obtaining the final model, two questions arise: 1. Are there any differences in average points scored for players who have the same percent of successful field goals? 2. Are there any differences in average points scored for players who have the same height?

To answer these questions, we can conduct hypothesis tests about our estimated coefficients. Question 1 is equivalent to test whether β_1 is 0. Hypothesis testing: H_0: $\beta_1 = 0$, H_1: $\beta_1 \neq 0$. Using R, the p-value for the hypothesis test of β_1 being 0 is 0.03357903. Because the p-value is small (< 0.05), we reject the null hypothesis that $\beta_1 = 0$. Therefore, we have statistical evidence to support that there are differences in average points scored for players with the same successful field goal rates.

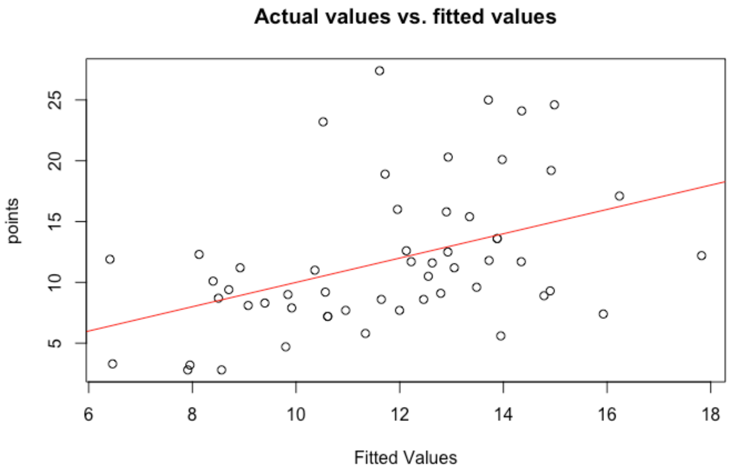
Question 2 is equivalent to test whether β_2 is 0. Hypothesis testing: H_0: $\beta_2 = 0$, H_1: $\beta_2 \neq 0$. Using R, the p-value for the hypothesis test of β_2 being 0 is 0.001161757. Because the p-value is very small, we reject the null hypothesis that $\beta_2 = 0$. Thus, we have statistical evidence for differences in average points scored for players who have the same height. Both estimated coefficients are statistically significant.

Using lm function, I can verify my results:

```
> m2 <- lm(points~height+field)
> summary(m2)

Call:
lm(formula = points ~ height + field)

Residuals:
   Min     1Q Median     3Q    Max 
-8.527 -3.621 -1.002  2.222 15.789 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   15.210     10.727   1.418   0.1623   
height        -4.035      1.866  -2.162   0.0353 * 
field         51.562     15.144   3.405   0.0013 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.416 on 51 degrees of freedom
Multiple R-squared:  0.1891,	Adjusted R-squared:  0.1573 
F-statistic: 5.945 on 2 and 51 DF,  p-value: 0.004776
```

My results and the results obtained by lm function are almost the same, considering rounding errors. In the summary, we see that the overall F-test for this model is highly statistically significant with p-value 0.004776.

In order to see if the final model fits the dataset well, we can plot the actual values vs. fitted values graph and add a straight fit reference line:



In this plot, we observe that roughly most of the points are near around the straight fit line. This provides further evidence that the final model fits the dataset well.

Summary & Discussion

In this project, I first construct a multiple linear regression model of **points** and **field** and **free**. Then, in order to find the best fitting model, I use stepwise search technique to find the final model, which has 2 regressors: **height** and **field**. The final model is: $\hat{y} = 15.210 - 4.035x_1 + 51.562x_2$. The mean square error of the final model is 28.76407, which is smaller than the previous model. And the sum of square residuals is 1495.732, which is also smaller than the previous model. These evidences show that the final model is better than the previous model. The forward selection and backward elimination based on BIC also show that this final model is the best fitting model. According to the final model, if player A is 6.3 feet tall and has 45% successful field goals, then his predicted average points scored is approximately 12.99466. If player B is 6.4 feet tall and has 46% successful field goals, then his predicted average points scored is approximately 13.10682. Basketball manager and team coach will probably choose player B instead of player A. This final model provides predictive power so that given a player's statistics, we can predict his average points scored per game.

Although this model fits the dataset well, future work can be done to improve it. First, I want to include more players' statistics in the current year, and include more variables in the dataset, like assists, rebounds, steals, blocks and so on. I think if I enlarge the sample size and include more variables in the dataset, I can improve the model to be more comprehensive and accurate. Also, currently this project is about NBA players, but in the future, I want to put NCAA (National Collegiate Athletic Association) players into consideration. I want to improve the model so that it has predictive power for both NBA players and NCAA basketball players. I can consider NBA player and NCAA player as a categorical variable and find out whether there are differences between these 2 types of basketball players.

References

1: Michael Jordan career statistics:
https://www.basketball-reference.com/players/j/jordami01.html
2: Michael Jordan career statistics:
https://www.basketball-reference.com/players/j/jordami01.html
3: Fox sports ranking: https://www.foxsports.com/nba/gallery/ranking-the-25-greatest-players-in-nba-history-100716
4: Kobe Bryant career statistics:
https://www.basketball-reference.com/players/b/bryanko01.html
5: Free throw definition: https://en.wikipedia.org/wiki/Free_throw
6: Field goal definition: https://en.wikipedia.org/wiki/Field_goal_(basketball)
7: Dataset source: http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr09.html