Credit: https://wallpapersafari.com/w/ynWIzP

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

STAT 410 Project

Hongyu Zhang

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Slide-Master.jpg

What factors determine an excellent basketball player?

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Hall of Fame

14x All Star	10x Scoring Champ
3x STL Champ	6x NBA Champ
11x All-NBA	9x All-Defensive
1984-85 All-Rookie	3x AS MVP
1987-88 Def. POY	6x Finals MVP
5x MVP	1984-85 ROY

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg
https://cdn.revistagq.com/uploads/images/thumbs/201340/michael_jordan_3508_645x485.jpg
https://www.basketball-reference.com/players/j/jordami01.html

Goal

Construct a statistical model that can use players' statistics to explain players' performance in game.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

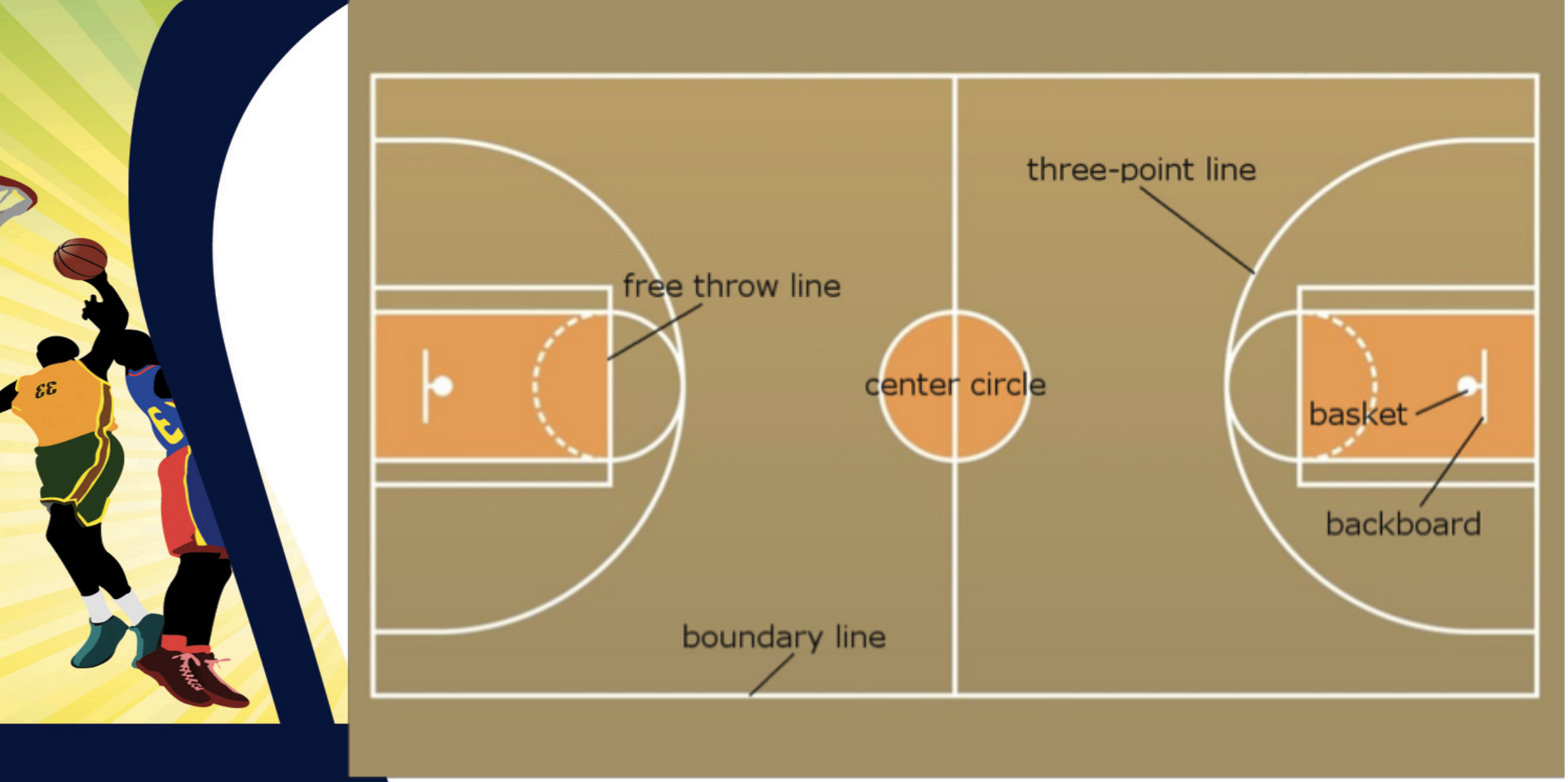
Introduction

1. How to evaluate performance?

Average points scored in game.

2. 2 ways to score points in game: field goals and free throws.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg



Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg
http://www.danna.it/Risorse/DAN/Public/O_D9046/D9046/Materiali_disponibili/images/basket1.jpg

Data Collection

I obtained the dataset online from Cengage. This dataset referred *The official NBA basketball Encyclopedia*, Villard Books.

This dataset contains 54 players' data in 5 categories:

height (in feet), **weight** (in pounds), **field** (percent of successful field goals out of 100 attempted), **free** (percent of successful free throws out of 100 attempted) and **points** (average points scored per game).

Here, the response variable is **points**, since it is the variable that I want to predict.

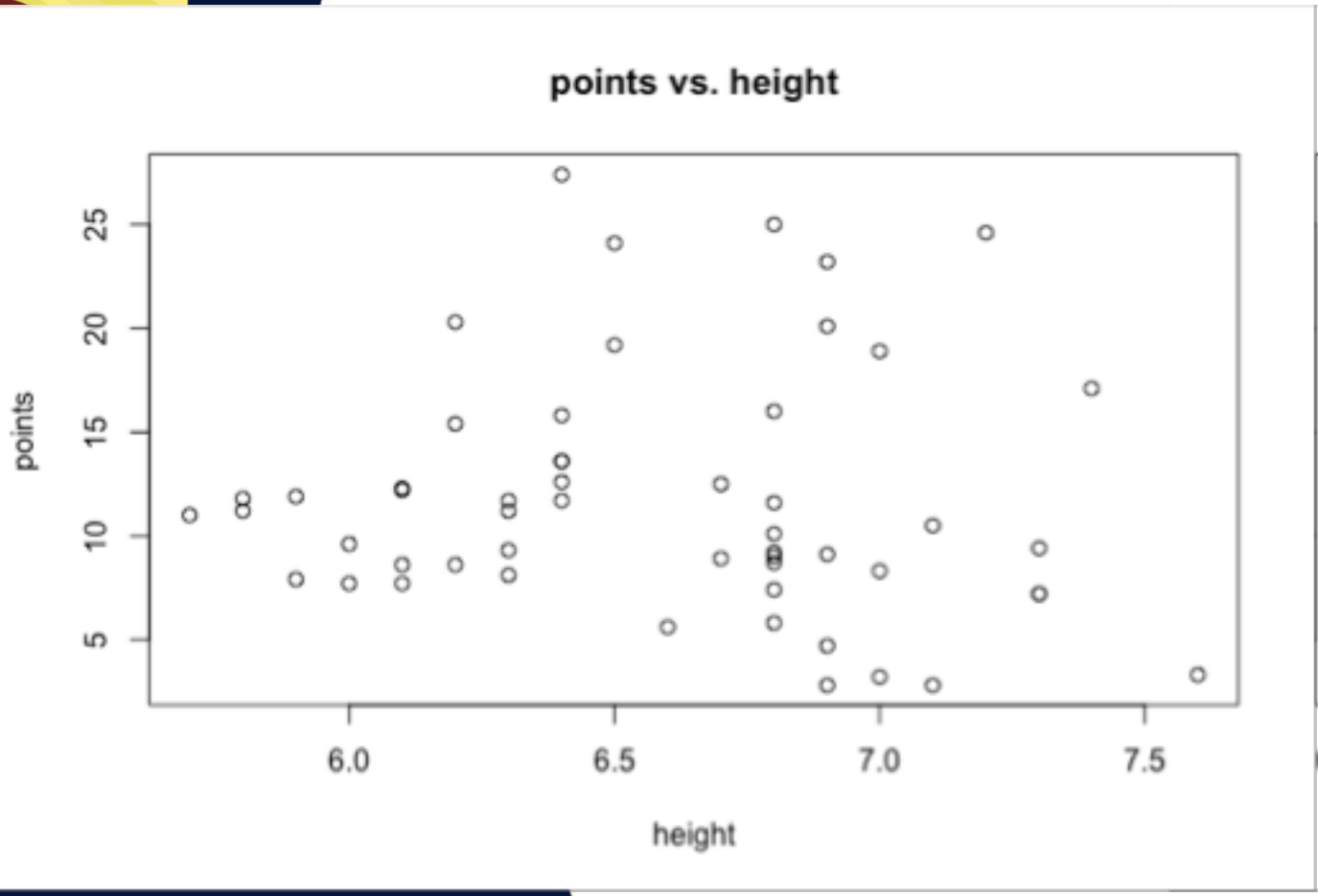
Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg
http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr09.html

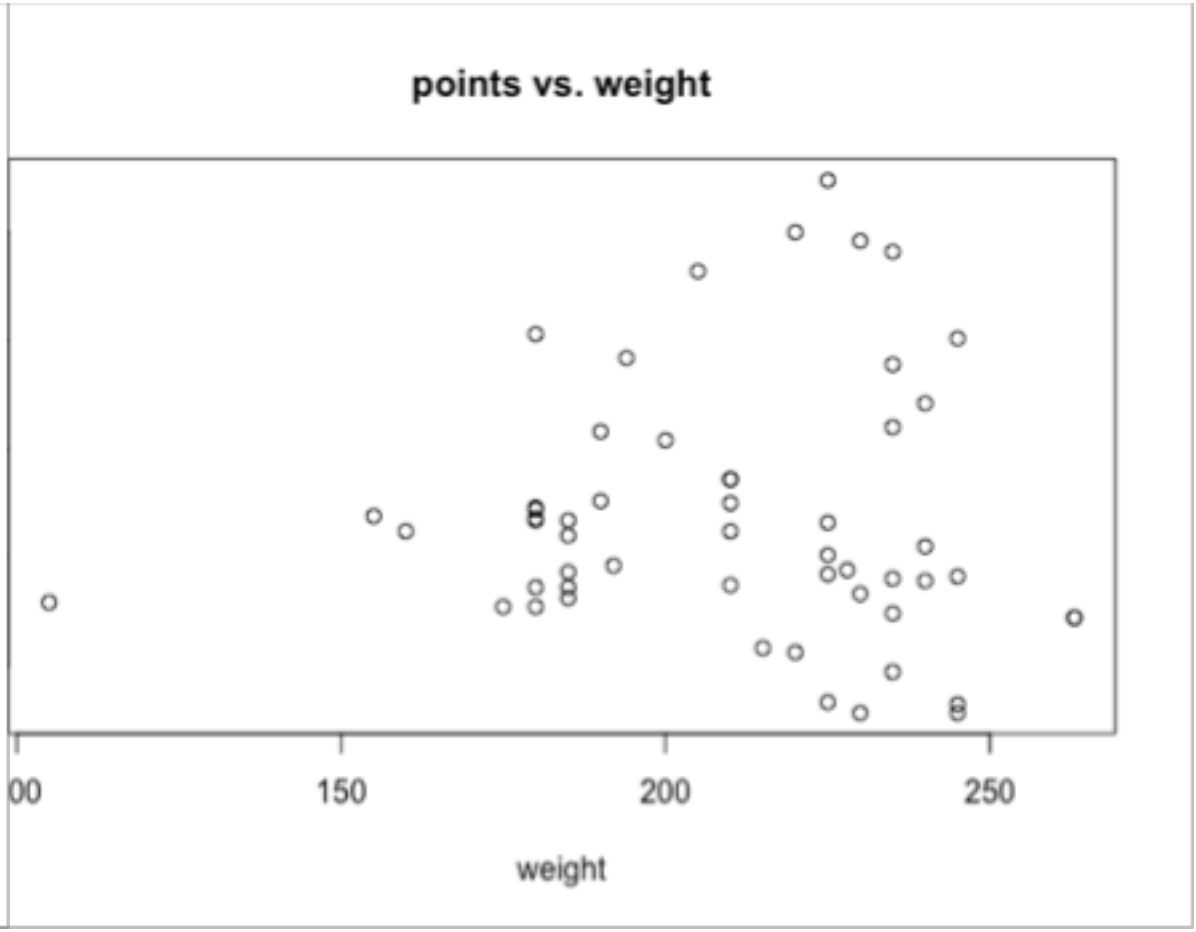
Visualization

	height	weight	percent_field_goals	percent_free_throws	avg_points
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

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

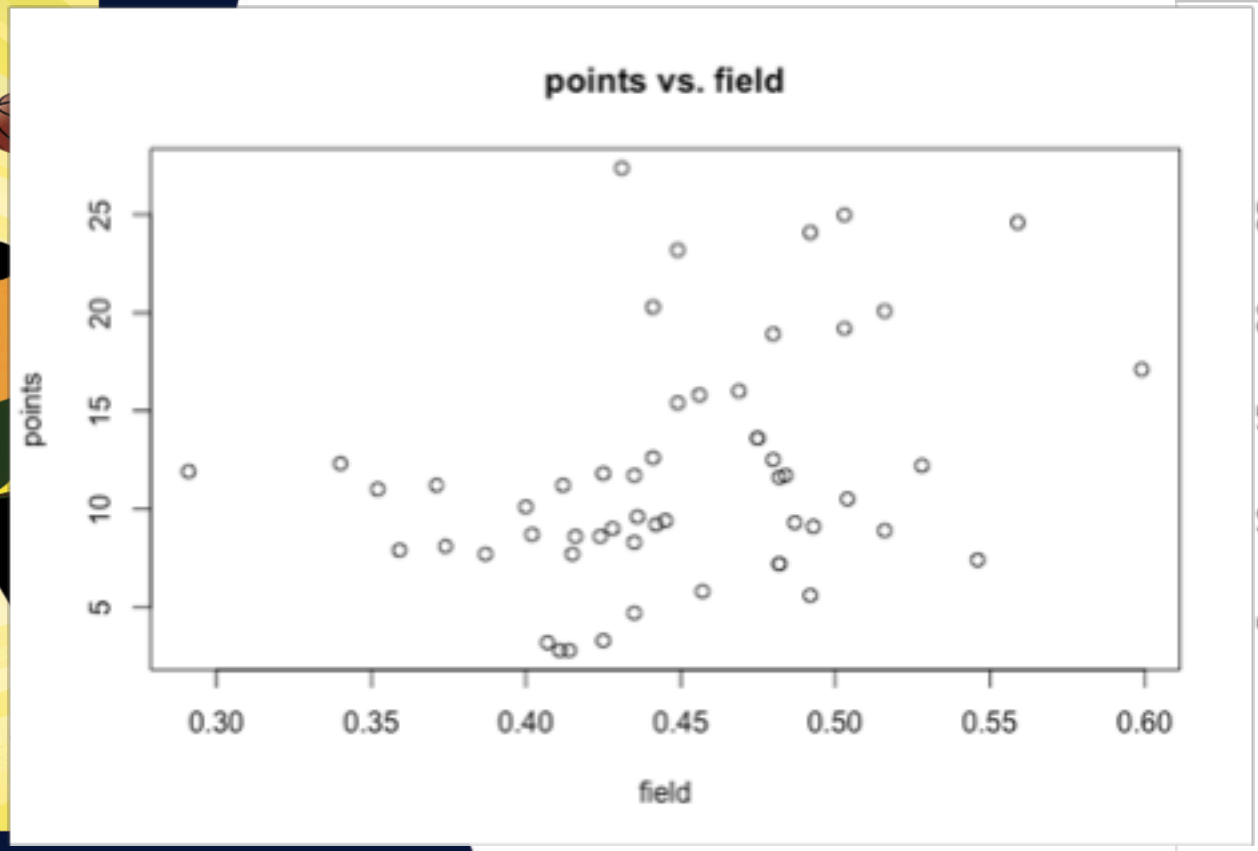
Visualization

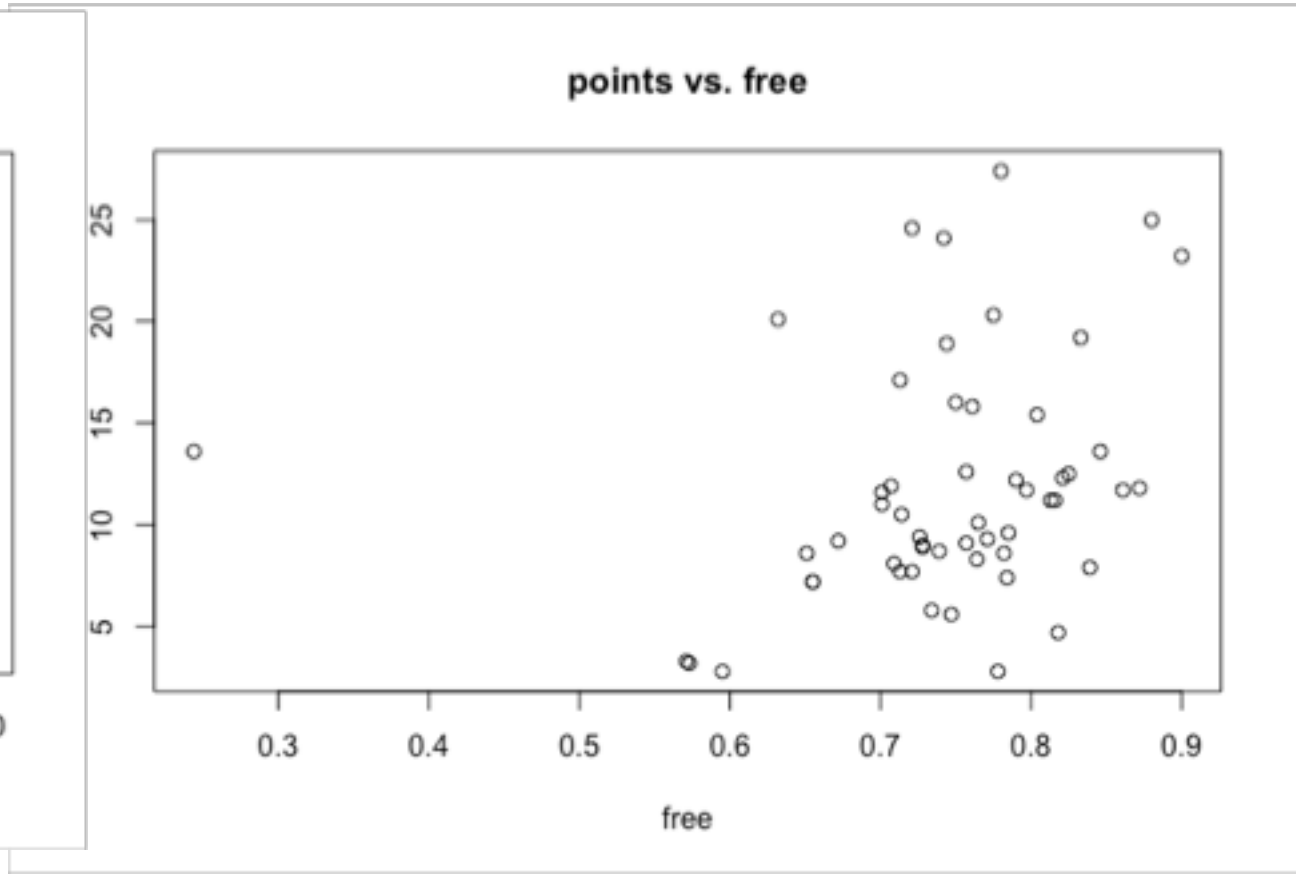




Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Visualization





Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Simple Approach

Relationship between average points scored and percent of successful field goals and free throws.

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.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Simple Approach

Here I try to build a multiple linear regression model of average points scored and 2 regressors: **field** (successful field goal percentage) and **free** (successful free throw percentage).

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Simple Approach

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

In matrix form, $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

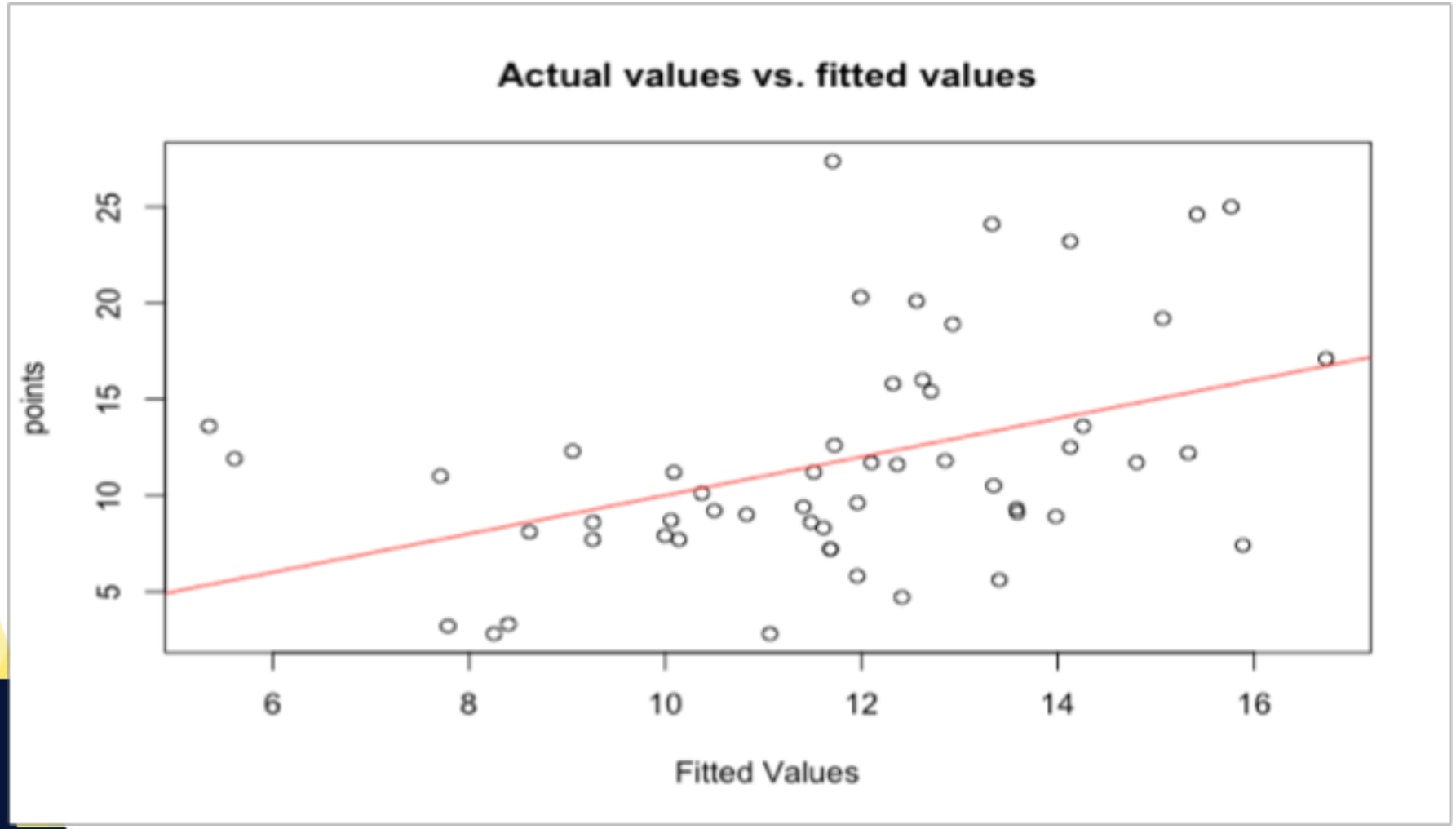
Using R, $\hat{\beta} = \begin{bmatrix} -15.27738 \\ 35.82503 \\ 14.79905 \end{bmatrix}$

$SS_{res} = 1516.422$

$\widehat{\sigma^2}$ (MS_{res}) = 29.16196

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

actual values vs. fitted values



Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Stepwise Search to Find the Best Model

4 possible regressors: **height**, **weight**, **field** and **free**.

There might be a 5th possible regressor: interaction term between **field** and **free**. I want to find out whether there exists association between percent of successful field goals and percent of successful free throws.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Stepwise Search to Find the Best Model

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.

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

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.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

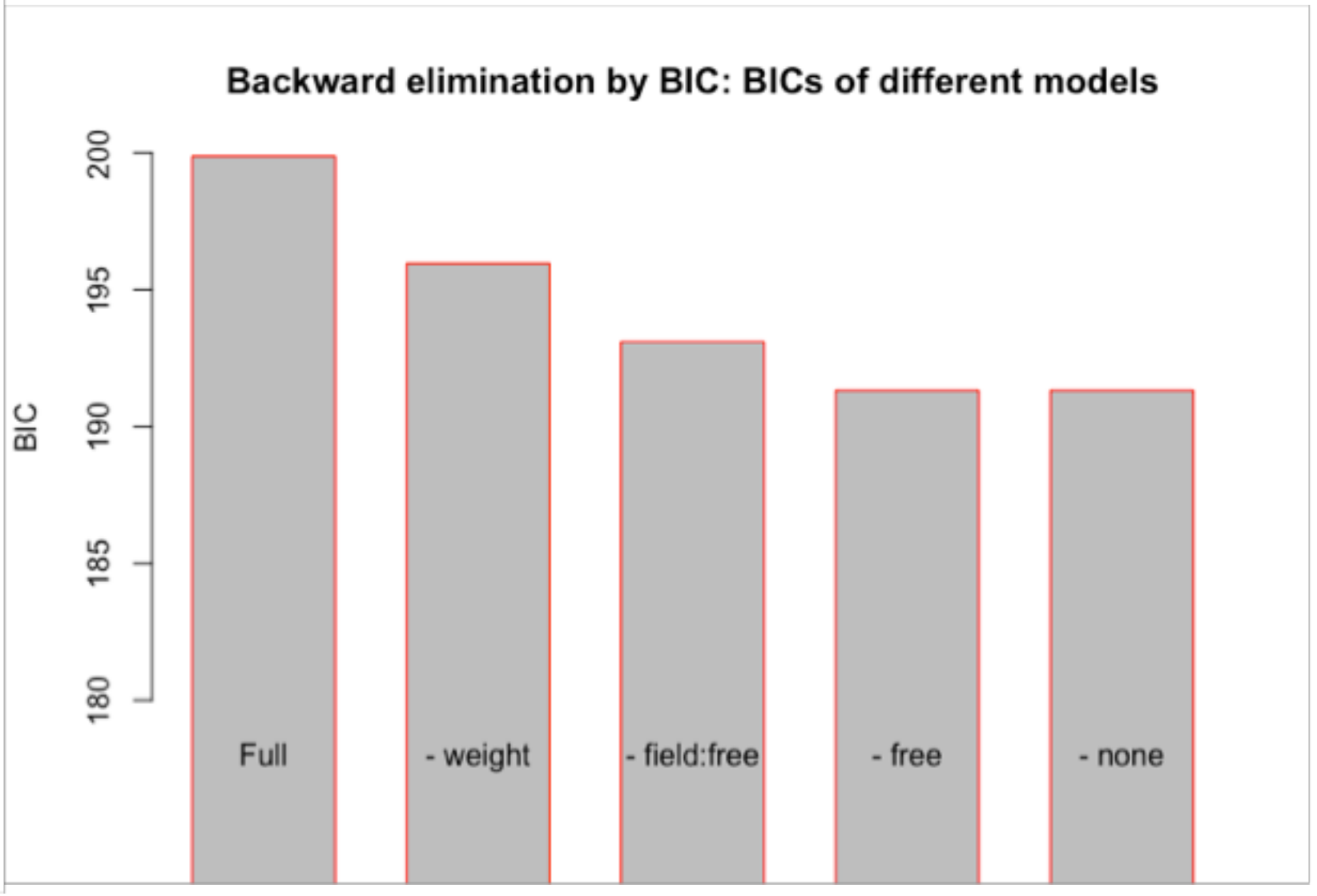
Stepwise Search to Find the Best Model

```
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
```

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Stepwise Search to Find the Best Model

```
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
```



Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Stepwise Search to Find the Best Model

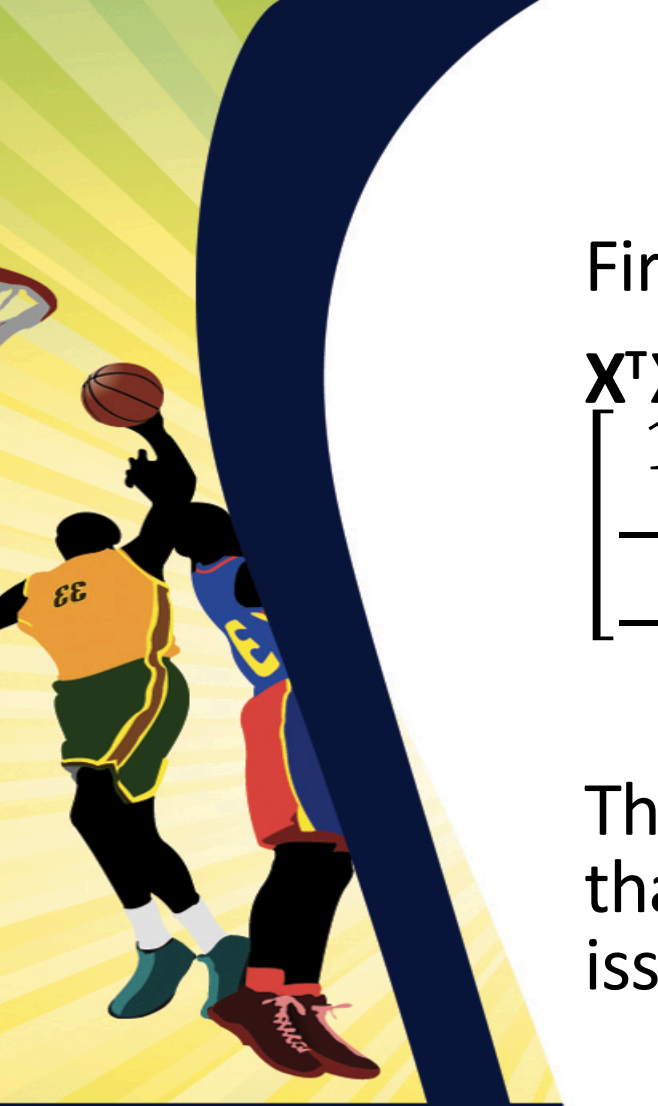
Final model with 2 regressors **height** and **field**.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Analysis of 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**.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg



Analysis of Final Model

First check that $\mathbf{X}^T\mathbf{X}$ is not a singular matrix.

$\mathbf{X}^T\mathbf{X} * (\mathbf{X}^T\mathbf{X})^{-1} =$

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

The result is very close to the identity matrix. This shows that our design matrix does not have multicollinearity issue.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Analysis of Final Model

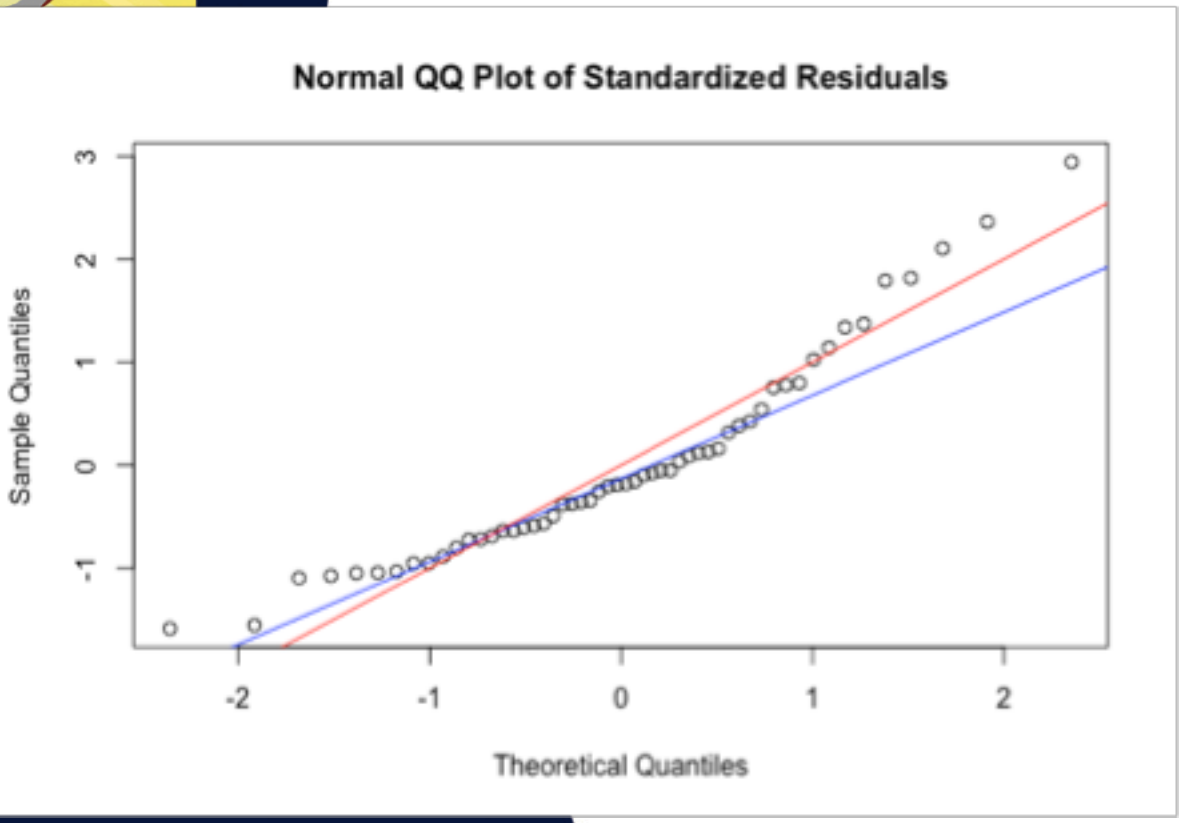
- Using R, I obtain $\hat{\beta} = \begin{bmatrix} 15.209793 \\ -4.034628 \\ 51.562276 \end{bmatrix}$.
- SS_{res} = 1495.732
- $\widehat{\sigma^2}$ (MS_{res}) = 28.76407
- The fitting model becomes:

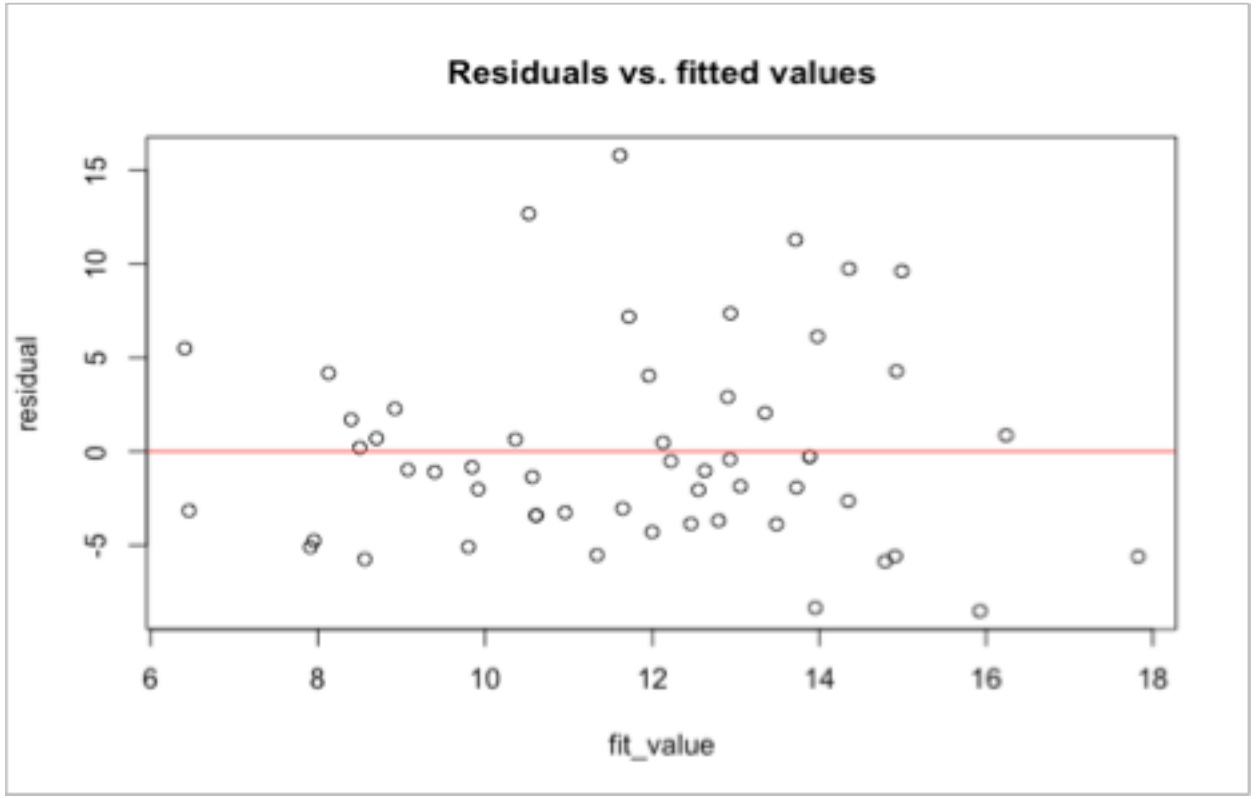
$$\hat{y} = 15.210 - 4.035x_1 + 51.562x_2.$$

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Analysis of Final Model

Normal assumption and homogeneity assumption.





Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Analysis of Final Model

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

Hypothesis testing 1:

Are there differences in average points scored for players who have the same percent of successful field goals?

$H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$

p-value 0.03357903

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Analysis of Final Model

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

Hypothesis testing 2:

Are there differences in average points scored for players who have the same height?

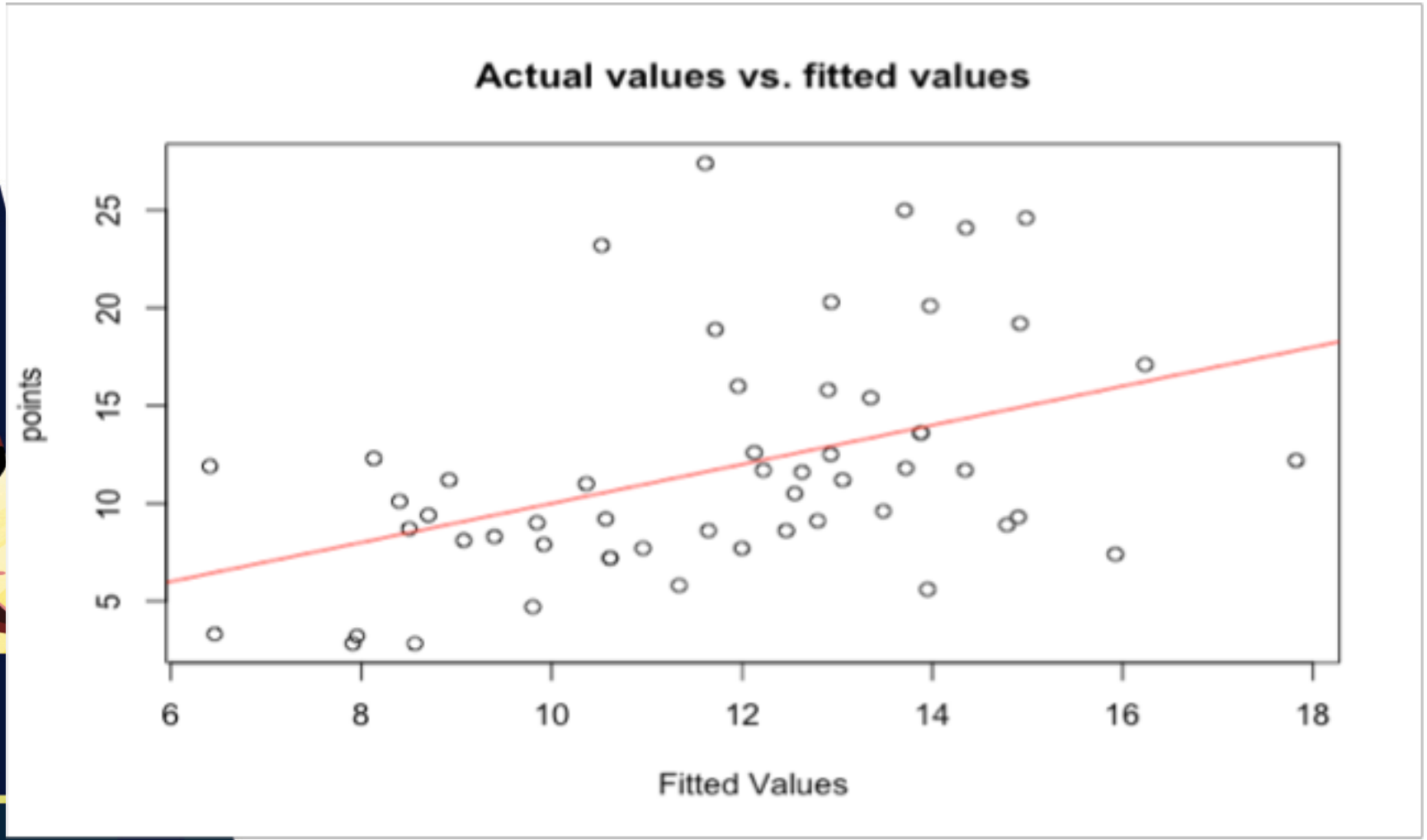
$H_0: \beta_2 = 0$, $H_1: \beta_2 \neq 0$

p-value 0.001161757

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Analysis of Final Model

actual values vs. fitted values



Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Discussion

Final model: $\hat{y} = 15.210 - 4.035x_1 + 51.562x_2$

If player A is 6.3 feet tall and has 45% successful field goals, then his predicted average points scored is approximately:

15.210 – 4.035 * 6.3 + 51.562 * 0.45 = 12.9924.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Discussion

If player B is 6.4 feet tall and has 46% successful field goals, then his predicted average points scored is approximately:

$$15.210 - 4.035 * 6.4 + 51.562 * 0.46 = 13.10452.$$

Basketball manager and team coach will probably choose player B instead of player A.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Discussion

Final model: $\hat{y} = 15.210 - 4.035x_1 + 51.562x_2$

Performance determined by 2 factors: height and percent of successful field goals.

Percent of successful field goals is the more influential factor in this model.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Future Work

1. Include more up-to-date players' statistics in the current year, and include more variables in the dataset, like assists, rebounds, steals, blocks and so on.

2. Compare NCAA (National Collegiate Athletic Association) players with NBA players.

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg

Thank you!

Credit: http://www.freepptbackgrounds.net/wp-content/uploads/2013/11/Sports-Basketball-Print-Master.jpg