**Probability and Statistics**

Project 2



|  |  |
| --- | --- |
| 학번 | 2016025469 |
| 이름 | 서건식 |
| 수업 | 확률통계론 |
| 교수님 | 배완덕 교수님 |
| 제출일 | 2017.06.13 |

1. Question studied

 The population, explanatory variables, response variables, and the question

1) The population : The basketball player of NBA

2) Explanatory variables :

X1 = height in feet

X2 = weight in pounds

X3 = percent of successful field goals (out of 100 attempted)

X4 = percent of successful free throws (out of 100 attempted)

3) response variable

Y = average points scored per game

4) The question

1. Is there a relationship between average points scored per game and explanatory variables?

2. How strong is that relationship?

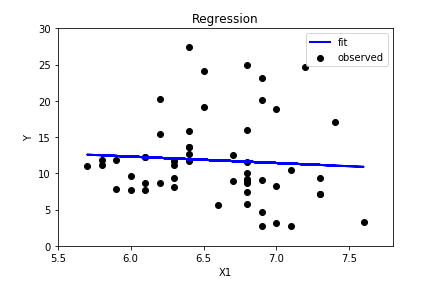
3. which goals(free throw or field goal) more contribute to average points scored per game?

4. In height in feet or weight in pounds, which one more contribute to average points scored per game?

1. Correlation analysis on all pairs of the explanatory variables and the response variable.

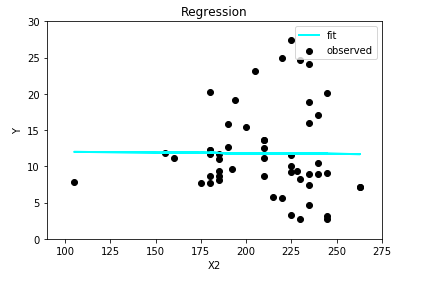
First, I’ll show the Estimating Model (using scatterplot)

1. Y with X1: height in feet



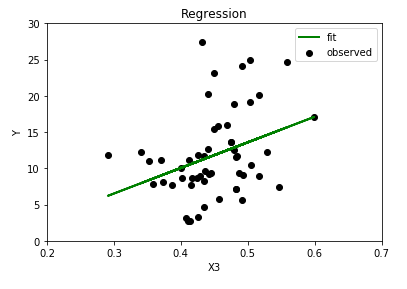
Correlation coefficient r = 0.33876

1. Y with X2: weight in pounds



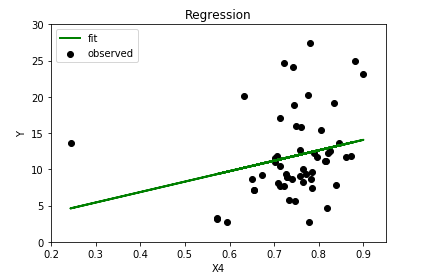
Correlation coefficient r = -0.00984

1. Y with X3: percent of successful field goals (out of 100 attempted)



Correlation coefficient r = 0.33876

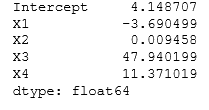
1. Y with X4: percent of successful free throws (out of 100 attempted)



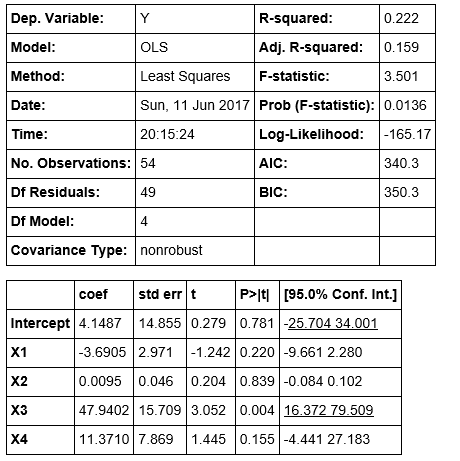
Correlation coefficient r =0.244852

1. A regression analysis.

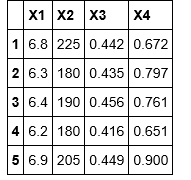
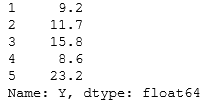
First , I’ll show each coefficient. (Multiple Linear regression)



Second, I’ll show the summary of values.



Step 1: Preparing X and y



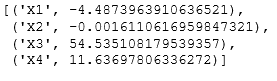
Step 2: Splitting X and Y into training and testing sets.

each, X-train , Y-train, X-test, Y-test.

Step 3: Develop a linear regression model.



Interpreting model coefficients



Y = 4.148707 -3.6905X1 + 0.009458X2 + 47.9402X3 + 11.37102X4

Step 4: Making predictions.

from sklearn import metrics

y\_pred = linreg.predict(X\_test)

Using this code, make a predictions.

Step 5: Computing the RMSE for the predictions



Step 6: Feature selection consideration

When I remove X1: 4.064799636

When I remove X2: 4.09287593868

When I remove X3: 4.68224795025

When I remove X4: 4.13965456846

The RMSE **decreased when** we removed X1 from the model. (the highest decreasing)

Conclusion:

The coefficient of X1 is -3.690499. This means that the larger the height, the lower the average or score.

The coefficient of X2 is 0.009458, which means that the weight of the athlete and the number of goals are almost irrelevant.

The coefficient of X3 is 47.940199, which means that the score in the field is very much related.

The coefficient of X4 is 11.371019, which means that the free throw success rate also affects, but not as much as, the score in the field.

Let’s answer the question.

1. Is there a relationship between average points scored per game and explanatory variables?

X2 has nothing to do with average points scored per game. It can be seen that the higher the player's height through X1, the lower the score. And through X3 and X4, we can see that the scoring rate in the field or free throw situation has a big influence on the average points

1. How strong is that relationship?

The coefficient value of X3 is 47.940199. This means that there is a strong relationship with the response variable.

1. which goals(free throw or field goal) more contribute to average points scored per game?

Field goal. It can be seen by comparing the values ​​of the coefficients.

1. In height in feet or weight in pounds, which one more contribute to average points scored per game?

Height. Weight was found to have little effect on average points scored per game.