The regression analysis of basketball players in videogame

Author: Jian Sun

GWID: G43474152

Date: 11/28/2015

Abstract

Regression analysis is a very useful knowledge, in this paper, we plan to use multiple regression analysis and principle component analysis to study the data we gathered from the NBA 2K16. Then we fit a model to reflect player's overall value and predict his overall interval. Then we use principle component analysis to show player's characteristic. We use some example to check our result and find the satisfying conclusion. It may be helpful for coach to choose player. For the future work, we plan to predict the player's overall value in each game.

Keywords: regression analysis, multiple linear regression, confidence interval, principle component analysis

Content

1 Introduction

Recently, we learned simple linear regression and multiple linear regression and regression diagnostics in the course. So useful they are that could solve many practical problems, such as evaluating the gender discrimination in salary of a company, judging whether the publisher should print more books and the relationship between mother's height and daughter's height. And I find that the more I learn, the more problem I can solve by this models. In this paper, we are going to study an interesting thing, how to evaluate a basketball player in videogame.

When we play basketball videogame, we would like to choose players who have high overall value. But few people focus on how to calculate the overall value by mathematical formula. Whether we can get this formula via linear model. If we could part the players to different position in our new model. These kind of questions are our target in this paper.

In this paper, we are going to fit a linear model via the basketball players' capability data. And judging whether we could drop some predict variables. We would predict the confidence interval of a player's overall value. Finally, we plan to predict the players' style and their position. About the outlines, we will discuss the method section in the part 2, and fit our model and do relative testing in part 3. In part 4, we will try to get principle component analysis. Then we need to do data analyzing in part 5, finally we will draw a conclusion in part 6.

2 Method section

In this paper, we try to evaluate videogame's basketball players' capability. We collect data from the NBA 2K16. And we will mainly discuss four problems, fitting a model, doing hypothesis testing, calculating a player's confidence interval and doing principle component regression to estimate the given players' feature.

For the fitting a model, we directly use R language to fit the data and check if there are collinear data. If so, we will use hypothesis testing to judge if we could drop those collinear predict variables. Given a player's capability data, we will use the confidence interval formula to predict his overall value interval.

But the fitting model cannot reflect players' characteristic, to solve this problem we add more predict variables and do principle component analysis.

Then according to the calculating conclusion we get, we will analysis them and tell the practical meaning behind them. Finally, we will draw our conclusion about our researching.

3 Basketball player's overall model

3.1 Fitting data

we choose the overall value as response variable, named Y;

meaning	variable
Speed	Z1
Moving Shot Close	Z2
Moving Shot Mid-Range	Z3
Draw Foul	Z4
Offensive Rebound^2+Defensive Rebound^2/ Offensive	Z5
Steal	Z6
Block	Z7
Vertical	Z8
Strength	Z9

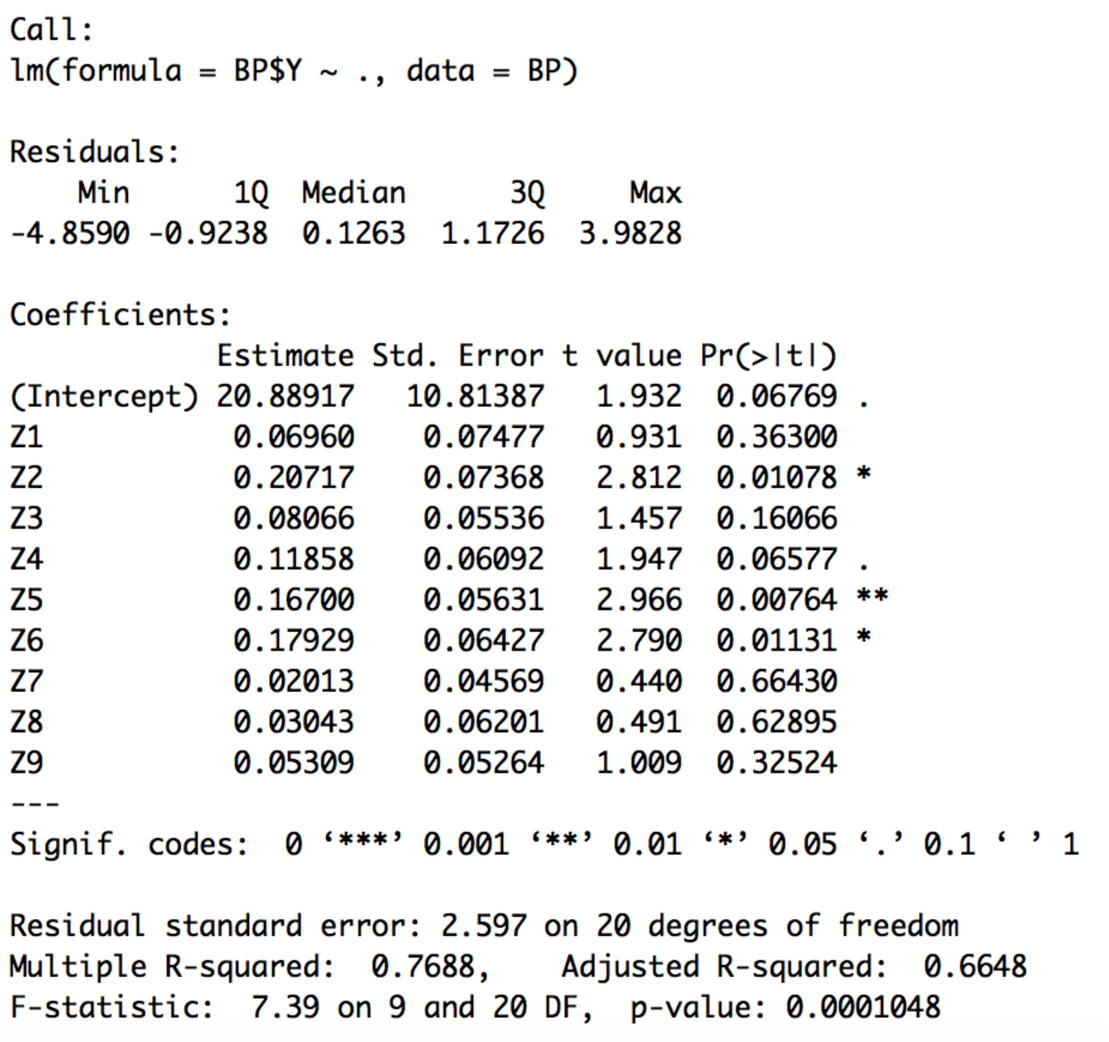
Then we use the least square method to get a multiple regression equation in R.

$$\hat{Y} = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_9 Z_9$$

The result is

Yhat=20.89+0.07Z1+0.21Z2+0.08Z3+0.12Z4+0.17Z5+0.18Z6+0.02Z7+0.03Z8+0.05Z9

Then we check the model's coefficients table



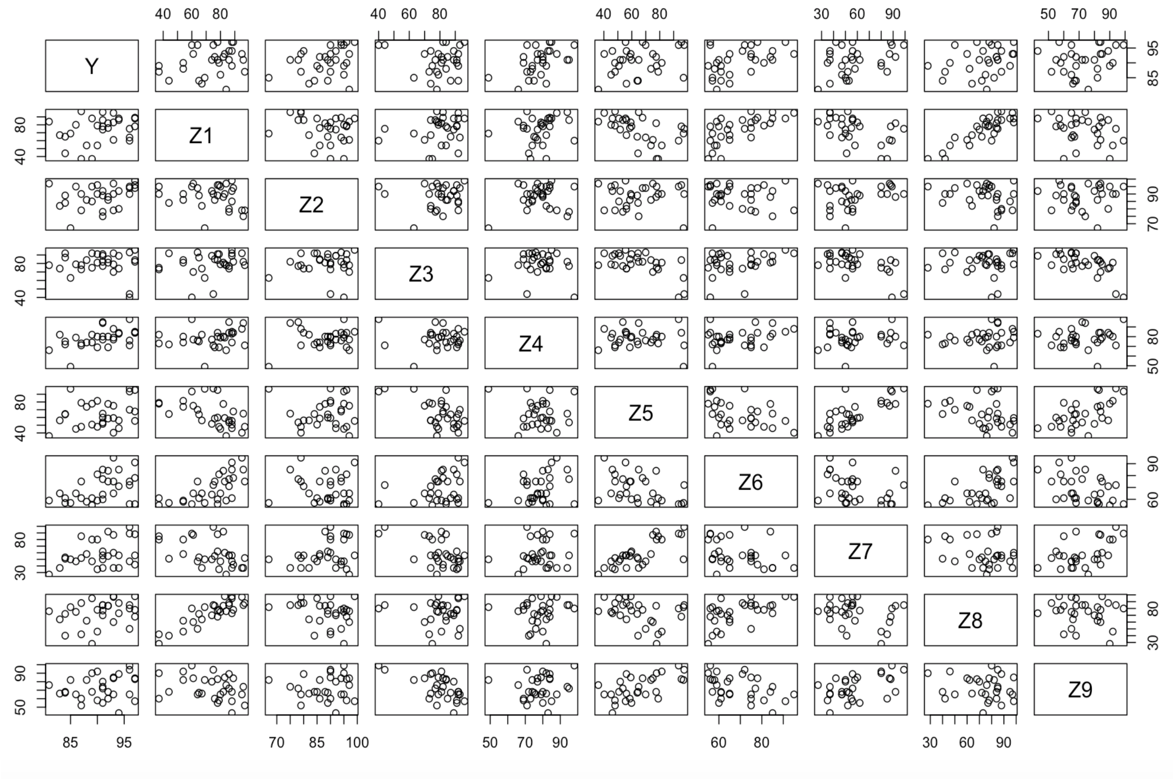
```
Call:
lm(formula = BP$Y ~ ., data = BP)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8590 -0.9238  0.1263  1.1726  3.9828 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 20.88917   10.81387   1.932  0.06769 . 
Z1           0.06960    0.07477   0.931  0.36300   
Z2           0.20717    0.07368   2.812  0.01078 * 
Z3           0.08066    0.05536   1.457  0.16066   
Z4           0.11858    0.06092   1.947  0.06577 . 
Z5           0.16700    0.05631   2.966  0.00764 **
Z6           0.17929    0.06427   2.790  0.01131 * 
Z7           0.02013    0.04569   0.440  0.66430   
Z8           0.03043    0.06201   0.491  0.62895   
Z9           0.05309    0.05264   1.009  0.32524   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.597 on 20 degrees of freedom
Multiple R-squared:  0.7688,	Adjusted R-squared:  0.6648 
F-statistic:  7.39 on 9 and 20 DF,  p-value: 0.0001048
```

we cannot deny that only three predict variables are significant. But taking the realistic meaning into consideration, we cannot easily drop them, so we continue to check the full model's scatter plot.



As we can see, there is a strong linear relationship between Z1 and Z8, Z5 and Z7, Z5 and Z9.

Then we let each of Z8, Z7 and Z9 be response variable and fit a new model with other predict variables. Using R, we get VIF (Variance Inflation Factor, $VIF = 1/1 - R_8^2$).

VIF(Z8) = 5.34<10,

VIF(Z7) = 3.52<10,

VIF(Z9) = 2.29<10.

Although the result does not show any collinear relationship between Z1 and Z8, Z5 and Z7, Z5 and Z9, we need to check the parameters as well.

3.2 Hypothesis testing

let check if all the parameters could be zero. Let a=5%

H0: $Y=\beta_0+ \qquad\qquad\qquad\qquad\qquad +\varepsilon$

H1: $Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+...+\beta_9 Z_9+\varepsilon$

$$\begin{aligned} F &= \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/(p+1-k)}{\text{SSE(FM)}/(n-p-1)} \\ &= \frac{[\text{SST} - \text{SSE}]/p}{\text{SSE}/(n-p-1)}. \end{aligned}$$

We know SST=583.369, SSE=134.867, p=9,n=30, so F=7.39,

F(9,20,0.05)=2.39.

F> F(9,20,0.05), so we can reject the null hypothesis that all parameters are zero.

Then we are going to test if the parameter of Z7=that of Z8= that of Z9=0,

$$H0: Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+...+\beta_6 Z_6+\varepsilon$$

$$H1: Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+...+\beta_8 Z_8+\beta_9 Z_9+\varepsilon$$

F=[SSE(RM)-SSE(FM)]/(p+1-k)/SSE(FM)/(n-p-1)

= (145.324-134.867)/3/134.867/20

= 0.517

F(3,20,0.05) = 3.098

F< F(3,20,0.05), so we cannot reject the null hypothesis that B7=B8=B9=0.

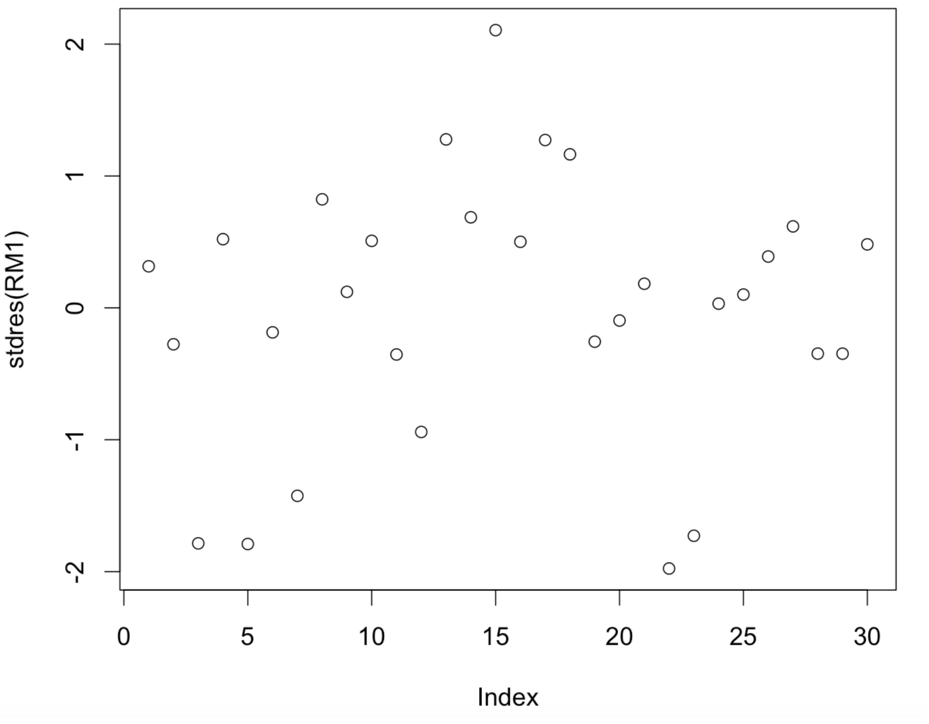
So the new model is

Y=24.95+0.102Z1+0.22Z2+0.05Z3+0.137Z4+0.198Z5+0.162Z6+e

Yhat=24.95+0.102Z1+0.22Z2+0.05Z3+0.137Z4+0.198Z5+0.162Z6

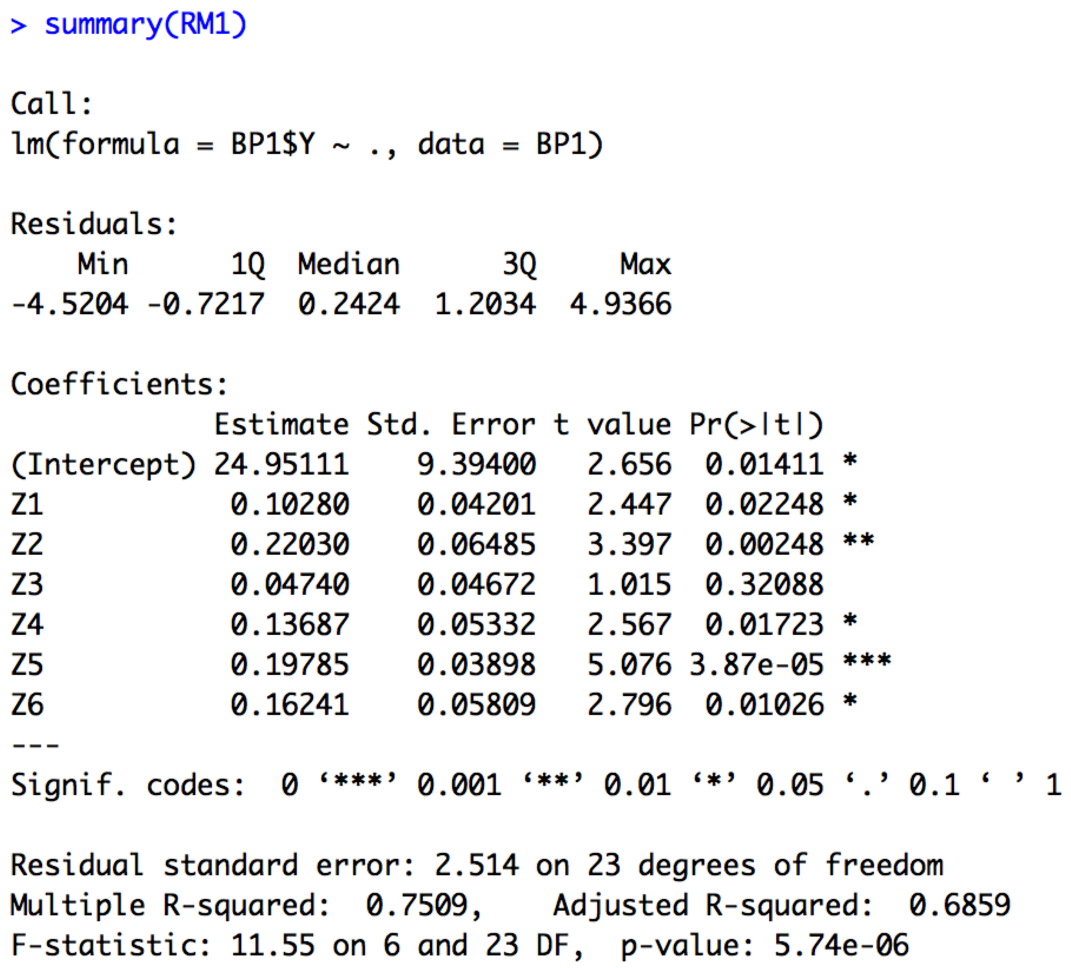
.

And here is the standard residual plot



From the plot, we can see the residuals are random distribution and can prove the new reduced model is reasonable.

Then we still need to check the model's coefficients table



```
> summary(RM1)

Call:
lm(formula = BP1$Y ~ ., data = BP1)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5204 -0.7217  0.2424  1.2034  4.9366 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.95111    9.39400   2.656  0.01411 *  
Z1           0.10280    0.04201   2.447  0.02248 *  
Z2           0.22030    0.06485   3.397  0.00248 ** 
Z3           0.04740    0.04672   1.015  0.32088    
Z4           0.13687    0.05332   2.567  0.01723 *  
Z5           0.19785    0.03898   5.076 3.87e-05 ***
Z6           0.16241    0.05809   2.796  0.01026 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.514 on 23 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.6859 
F-statistic: 11.55 on 6 and 23 DF,  p-value: 5.74e-06
```

we find that the parameter of Z3 is much bigger than 0.05 and extremely insignificant, we decide to drop it.

Now, we need to refit the model by Y, Z1, Z2, Z4, Z5, Z6. The result is

Y=29.933+0.095Z1+0.226Z2+0.13Z4+0.176Z5+0.175Z6+e.

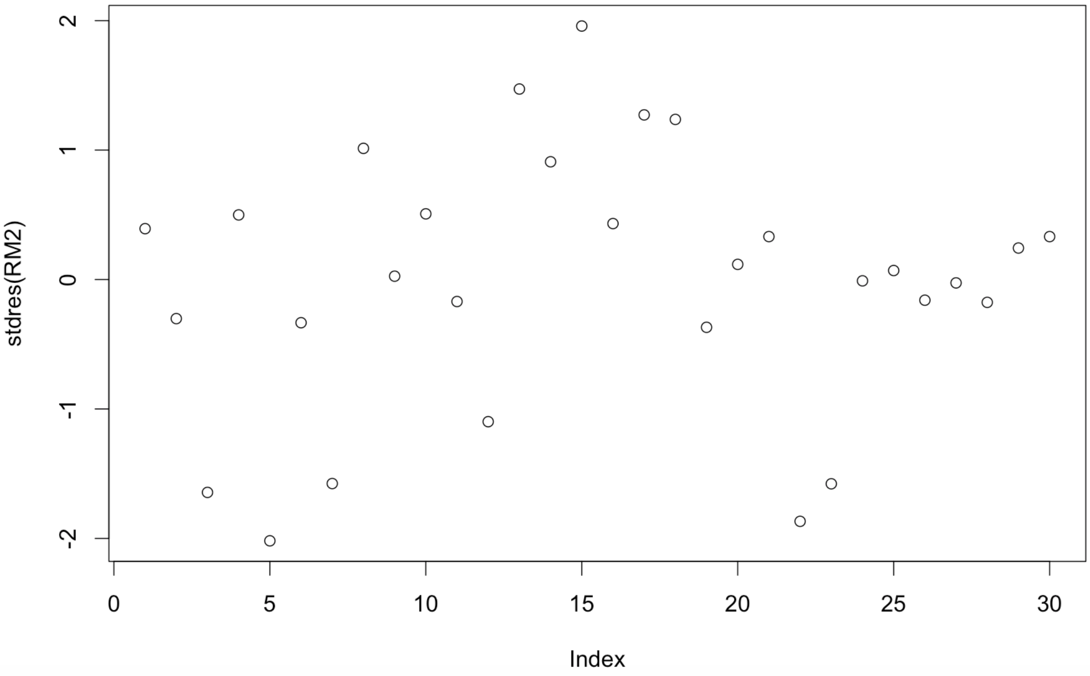
```
> summary(RM2)

Call:
lm(formula = BP2$Y ~ ., data = BP2)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2955 -0.7415  0.1048  1.0737  4.6315 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.93347    8.01302   3.736  0.00102 ** 
Z1           0.09481    0.04129   2.296  0.03072 *  
Z2           0.22583    0.06466   3.492  0.00188 ** 
Z4           0.12981    0.05289   2.454  0.02176 *  
Z5           0.17587    0.03242   5.424 1.42e-05 ***
Z6           0.17511    0.05676   3.085  0.00506 ** 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.515 on 24 degrees of freedom
Multiple R-squared:  0.7397,    Adjusted R-squared:  0.6855 
F-statistic: 13.64 on 5 and 24 DF,  p-value: 2.32e-06
```



we find that in this model, all the parameters are significant. We also find that the correlation coefficient is 73.97%. Hence, the new reduced model could explain 73.97% response variable.

Checking it's standardizing residual plot, we find the residuals obey random distribution.

3.3 Predicting a player's overall interval

In this part, we plan to use the new regression model and player's data to predict the interval of his overall value.

The player we choose is Blake Griffin, his capability value is

Y=90, Z1=74, Z2=90, Z3=87, Z4=87,

Z5 = (62^2+78^2)/(62+78)=70.91, Z6=61.

So

Y0hat=29.933+0.095*74+0.226*90+0.13*87+0.176*70.91+0.175*61

=91.768

$$\text{s.e.}(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p-1}$$

sigma^2=6.326,

s.e.y0hat=3.44,

t(23,0.025)=2.0639

$$\hat{y}_0 \pm t_{(n-p-1,\alpha/2)}\ \text{s.e.}(\hat{y}_0).$$

The result is 84.67 and 98.87, so the predict interval of Griffin's overall value is from 84.67 to 98.87. At a meanwhile, the overall value of Blake Griffin we collected is 90, which falls into the confidence interval [84.67, 98.87]

4 Players' feature analysis

In the above researching, we find that the function of the multiple regression model is limited, it is difficult for us to understand the players' characteristic. Therefore, we plan to do principle analysis to get more information. And we add more predict variables to 20. Initially, we set the cumulative proportion to be 85%. By using function prcomp(), we notice that we need five principle components to meet the 85%.

```
PCTL<-cbind(PD1)
PCFM<-prcomp(PD1,scale=TRUE,scores=TRUE)
summary(PCFM)

> summary(PCFM)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     2.9269 1.8164 1.4825 1.19290 1.04720
Proportion of Variance 0.4284 0.1650 0.1099 0.07115 0.05483
Cumulative Proportion  0.4284 0.5933 0.7032 0.77434 0.82917
                           PC6
Standard deviation     0.98070
Proportion of Variance 0.04809
Cumulative Proportion  0.87726
```

And then, we use R language to get correlation coefficient matrix and it's eigenvectors. We keep the first 5 eigenvectors which are

```
             [,1]        [,2]         [,3]        [,4]
 [1,]  0.10552236  0.21055342  0.476682347  0.33179663
 [2,] -0.19121037  0.35592455  0.104234811 -0.27706697
 [3,] -0.29472886  0.15606305 -0.098814332 -0.09294418
 [4,]  0.04017204  0.21635440  0.514711443  0.34780898
 [5,] -0.23787746  0.28488679  0.050982721 -0.26834844
 [6,] -0.29533331  0.04602938 -0.086355937 -0.02562325
 [7,] -0.18584617  0.31150867  0.193775497 -0.21183992
 [8,] -0.27846897 -0.17738183  0.150260986  0.10169843
 [9,] -0.04551325 -0.12001782  0.405327318 -0.07571083
[10,] -0.30581039 -0.12411937  0.123812423  0.13422107
[11,] -0.25384692 -0.17994893  0.122103353  0.06714324
[12,]  0.29317377 -0.13941564  0.073670971 -0.14202946
[13,]  0.29898923 -0.02750869  0.029842566 -0.26752138
[14,] -0.21099822 -0.22960362  0.237851363 -0.26855172
[15,]  0.28283286 -0.05860206  0.217585040 -0.08459219
[16,] -0.23529531 -0.33573297  0.055607349  0.09758731
[17,] -0.15804662 -0.42118682 -0.008335855  0.11599932
[18,]  0.22598021 -0.15047553  0.059136317  0.10582268
[19,]  0.12129184 -0.17700735  0.287083147 -0.50981664
[20,] -0.04384444 -0.25181833  0.152295826 -0.24957342

              [,5]
 [1,] -0.182888835
 [2,] -0.010417188
 [3,] -0.232330072
 [4,] -0.138183809
 [5,] -0.032403493
 [6,] -0.336367222
 [7,]  0.063834676
 [8,] -0.002272508
 [9,]  0.327578237
[10,]  0.060417518
[11,]  0.054548704
[12,] -0.081887985
[13,] -0.082052990
[14,]  0.165624341
[15,]  0.121980239
```

```
[16,]   0.036482688
[17,]   0.005791361
[18,] -0.326221385
[19,]   0.059231009
[20,] -0.702938616
```

let the Xi, i=1,2,3…,20 be the standardizing predict variables

Xi=(Zi-Zibar)/sqrt(sum(Z-Zbar)^2/(n-1))

According to the Page 192 in the textbook, we can draw the principle component equation,

PC1=0.106X1-0.191X2-0.295X3+0.04X4-0.238X5-0.295X6-0.18§X7-0.278X8-0.046X9-0.306X10-0.254X11+0.293X12+0.299X13-0.211X14+0.283X15-0.235X16-0.158X17+0.226X18+0.121X19-0.044X20

PC2=0.211X1+0.356X2+0.156X3+0.216X4+0.285X5+0.046X6+0.312X7-0.177X8-0.12X9-0.124X10-0.18X11-0.139X12-0.028X13-0.23X14-0.059X15-0.336X16-0.421X17-0.15X18-0.177X19-0.252X20

PC3=0.477X1+0.104X2-0.099X3+0.515X4+0.051X5-0.086X6+0.194X7+0.15X8+0.405X9+0.124X10+0.122X11+0.074X12+0.03X13+0.238X14+0.218X15+0.056X16-0.008X17+0.059X18+0.287X19+0.152X20

PC4=0.332X1-0.277X2-0.093X3+0.348X4-0.268X5-0.026X6-0.212X7+0.102X8-0.076X9+0.134X10+0.067X11-0.142X12-0.268X13-0.269X14-0.085X15+0.098X16+0.116X17+0.106X18-0.51X19-0.25X20

PC5=-0.183X1-0.010X2-0.232X3-0.138X4-0.032X5-0.336X6+0.064X7-0.002X8+0.328X9+0.060X10+0.055X11-0.082X12-0.082X13+0.166X14+0.122X15+0.036X16+0.006X17-0.326X18+0.059X19-0.703X20

In PC 1, the parameter of Standing Shot Close, Moving Shot Close, Offensive Rebound, Defensive Rebound, Block, Strength, Reaction Time is positive, so we let PC1 check if the player is good at defense.

In PC 2, the parameter of Standing Shot Close, Standing Shot Mid-Range, Standing Shot Three, Moving Shot Close, Moving Shot Mid-Range, Moving Shot Three, Free Throw is positive, so we let PC 2 check if the player is good at attacking and score capability. In PC 3, except the parameter of Standing Shot Three, Moving Shot Three, Vertical, the rest parameters are positive, so we use PC 3 to check if the player has a comprehensive ability.

In PC 4, the parameter of Standing Shot Close, Moving Shot Close, Driving Layup, Ball Control, Passing Accuracy, Speed, Vertical, Strength is positive, so we let PC 4 check if the player is good at breaking through.

In PC 5, the parameter of Free Throw, Draw Foul, Ball Control, Passing Accuracy, Steal, Block, Speed, Vertical , Reaction Time is positive, so we let PC 5 check if the player is good at basic skills.

Firstly, let us take LeBron James into example, and the conclusion is PC1=-1.0275
PC2=-2.3707
PC3=-0.6348

PC4=-0.4051
PC5=-0.7888

The highest one is PC4, so LeBron James is good at breakthrough.

Then we begin to research Tim Duncan into example, and the conclusion is PC1=4.4460
PC2=2.0203
PC3=-0.5134
PC4=-0.1616
PC5=1.8579

The highest one is PC1, and the PC2 is also high, so Tim Duncan is excellent at defense and good at attacking.

Next, we decide to study Michael Jordan, and the conclusion is
PC1=-2.7218
PC2=-0.4809
PC3=3.0197
PC4=-0.9712
PC5=-0.7105

The highest one is PC3, so Michael Jordan is a comprehensive player.

Finally, we take Larry Bird into example, and the conclusion is
PC1=-1.1484
PC2=1.9187
PC3=1.3100
PC4=-1.4902
PC5=-1.3467

The highest one is PC2, and PC3 is relative high as well, so Larry Bird is great at attacking and is a comprehensive player.

5 Data analysis

The above section 3 and section 4 show us the result of four problems. As we can see, in the full model, the parameter of Block, Vertical, Strength can be 0, since the rest predict variables can explain the response variable (overall). That's also mean the more predict variables the model has, the more possible it has collinearity. And the scatter plot of standard residuals is random distribution, which proves the rationality of the new reduced model. But we find there is a insignificant variable, so we drop it and do a regression analysis in R again. In this way, the second model is fitted. What is more, given the correlation coefficient of new reduced model is 73.97%, so the variable Z1 to Z6 except Z3 are account for 73.97% of the model. The relationship between response variable and predict variable in this model is tight. Secondly, the predict value of Blake Griffin's overall is 91.77, the predicting confidence interval of his overall value is from 84.67 to 98.87. It is true that the predict interval has more error. But we use a predict interval because we did not put Blake Griffin's data into dataset. So we chose the predict formula. From this interval we can find that the Blake Griffin has potential to increase his overall value higher than 92.

Thirdly, in section 3, we do fit a linear model, but it could solve limited problems. We want to get a more satisfying result via the gathered data. Therefore, we decide to do principle component analysis in section 4. We find that LeBron James is good at breakthrough, Tim Duncan is excellent at defense and good at attacking, Michael Jordan is a comprehensive player and Larry Bird is great at attacking and is a comprehensive player. Considering the fact, we do a correct judging by our model. At a meanwhile, Both Larry Bird and LeBron James are SF in the field and excellent players in whole league. It seems like hard for coach to make decision in starting line-up. However, according to the PC1 to PC5, MR Bird has a high score in attacking and comprehensive ability than MR James. So if I am the coach I would let MR Bird be the starting line-up. Additionally, by PCA, coach will know if the player A is good at different positions. So when there is a team member getting injured, the coach can judge if he would put this player A to alter the injured one. Or the coach would judge if the player A is valuable for team to pay salary.

6 Conclusion

In the end, we can draw our conclusion.

Firstly, the multiple linear model of NBA player's overall value is Y=29.933+0.095Z1+0.226Z2+0.13Z4+0.176Z5+0.175Z6+e. This is an adjusted result after finishing F-test. The predict variables in this model are able to account for 73.97% of response variable (overall value). By this model, under the 95% confidence level, we predict the confidence interval of a NBA star -Blake Griffin- is from 84.67 to 98.87. Given the limited information getting from this model, we turn to principle component analysis to get more results. By using prcomp(), we lower the dimension to 5 principle components, which PC1, PC2, PC3, PC4, PC5.

And PC1 reflects player's defense ability;
PC2 reflects the player's attacking and score capability;
PC3 reflects the player's comprehensive ability.
PC4 reflects the player's breakthrough capability.
PC5 reflects the player's basic skills.

According to these 5 elements, we analysis 4 players, LeBron James, Tim Duncan, Michael Jordan and Larry Bird. The results we get are that LeBron James is expert at breakthrough, Tim Duncan is a defense guru who can get score well in basketball court, Michael Jordan is master of omnipotence and Larry Bird is a good attacker and a comprehensive player as well. These results are accord with the realistic situation. What's more, we think our results do a favor for coach to put better players into starting line-up. For example, between Larry Bird and LeBron James, we will choose Larry Bird to be the starting line-up based on our principle component analysis. Even the coach can use this model to check whether the player could do more job. If there are more players in our dataset, we could evaluate other players too. For future researching, we plan to modifying our dataset. The existing dataset is adjusted one, it provides error from the beginning of the researching. We plan to use original players' data to fit model and do PCA, even we want to predict the player's overall value in each game.

Reference

1. S. Chatterjee and A. S. Hadi, *regression analysis by example (fifth edition)*, China Machine Press, Beijing, China, 2013. ISBN 978-7-111-43156-5

Appendix

```
PD=read.table("/Users/LoveChina/Desktop/DATA.txt", header = TRUE)
PD1=read.table("/Users/LoveChina/Desktop/partdat1.txt",header = TRUE)
PD2=read.table("/Users/LoveChina/Desktop/partata1.txt",header = TRUE)
#section 3
#fit full model and check colinearity
Z1=PD$S.1
Z2=PD$MSC
Z3=PD$MSM
Z4=PD$DF
Z5=(PD$OR^2+PD$DR^2)/(PD$OR+PD$DR)
Z6=PD$S
Z7=PD$B
Z8=PD$V
Z9=PD$S.2
BP<-data.frame(Y,Z1,Z2,Z3,Z4,Z5,Z6,Z7,Z8,Z9)
plot(BP)
FM1<-lm(BP$Y~.,data=BP)
summary(FM1)
anova(FM1)
plot(stdres(FM1))
(583.369-134.867)/9/(134.867/20)
qf(0.05,9,20,lower.tail=FALSE)
qt(p=0.025,df=20,lower.tail=FALSE)
colinear1<-lm(Z8~Z2+Z3+Z4+Z5+Z6+Z7+Z1+Z9)
summary(colinear1)
1/(1- 0.8128)
colinear2<-lm(Z7~Z2+Z3+Z4+Z5+Z6+Z8+Z1+Z9)
summary(colinear2)
1/(1-0.7162)
colinear3<-lm(Z9~Z2+Z3+Z4+Z5+Z6+Z7+Z1+Z8)
summary(colinear3)
1/(1-0.5627)

> summary(FM1)

Call:
lm(formula = BP$Y ~ ., data = BP)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-4.8590 -0.9238   0.1263   1.1726   3.9828

Coefficients:
            Estimate Std. Error t value Pr(>ltl)
(Intercept) 20.88917   10.81387   1.932  0.06769 .
Z1           0.06960    0.07477   0.931  0.36300
Z2           0.20717    0.07368   2.812  0.01078 *
Z3           0.08066    0.05536   1.457  0.16066
Z4           0.11858    0.06092   1.947  0.06577 .
Z5           0.16700    0.05631   2.966  0.00764 **
Z6           0.17929    0.06427   2.790  0.01131 *
Z7           0.02013    0.04569   0.440  0.66430
Z8           0.03043    0.06201   0.491  0.62895
Z9           0.05309    0.05264   1.009  0.32524
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.597 on 20 degrees of freedom
Multiple R-squared:  0.7688,  Adjusted R-squared:  0.6648
F-statistic:  7.39 on 9 and 20 DF,  p-value: 0.0001048

> anova(FM1)
Analysis of Variance Table

Response: BP$Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
Z1         1  46.532  46.532  6.9004  0.016159 *
Z2         1  82.058  82.058 12.1687  0.002317 **
Z3         1  21.266  21.266  3.1536  0.090979 .
Z4         1  74.731  74.731 11.0822  0.003346 **
Z5         1 164.062 164.062 24.3295 8.028e-05 ***
Z6         1  49.395  49.395  7.3251  0.013584 *
Z7         1   1.307   1.307  0.1938  0.664466
Z8         1   2.292   2.292  0.3398  0.566449
Z9         1   6.859   6.859  1.0172  0.325241
Residuals 20 134.867   6.743
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> (583.369-134.867)/9/(134.867/20)
[1] 7.39003
> qf(0.05,9,20,lower.tail=FALSE)
```

```
[1] 2.392814
> qt(p=0.025,df=20,lower.tail=FALSE)
[1] 2.085963
> 1/(1- 0.8128)
[1] 5.34188
> 1/(1-0.7162)
[1] 3.523608
> 1/(1-0.5627)
[1] 2.28676

#fit reduced model
BP1<-data.frame(Y,Z1,Z2,Z3,Z4,Z5,Z6)
plot(BP1)
RM1<-lm(BP1$Y~.,data=BP1)
summary(RM1)
anova(RM1)
((145.324-134.867)/3)/(134.867/20)
qf(0.05,3,20,lower.tail = FALSE)
plot(stdres(RM1))

> summary(RM1)

Call:
lm(formula = BP1$Y ~ ., data = BP1)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5204 -0.7217  0.2424  1.2034  4.9366 

Coefficients:
            Estimate Std. Error t value Pr(>ltl)    
(Intercept) 24.95111    9.39400   2.656  0.01411 *  
Z1           0.10280    0.04201   2.447  0.02248 *  
Z2           0.22030    0.06485   3.397  0.00248 ** 
Z3           0.04740    0.04672   1.015  0.32088    
Z4           0.13687    0.05332   2.567  0.01723 *  
Z5           0.19785    0.03898   5.076 3.87e-05 ***
Z6           0.16241    0.05809   2.796  0.01026 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.514 on 23 degrees of freedom
```

```
Multiple R-squared:  0.7509,  Adjusted R-squared:  0.6859
F-statistic: 11.55 on 6 and 23 DF,  p-value: 5.74e-06

> anova(RM1)
Analysis of Variance Table

Response: BP1$Y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Z1         1  46.532  46.532  7.3644  0.012386 *  
Z2         1  82.058  82.058 12.9870  0.001496 ** 
Z3         1  21.266  21.266  3.3656  0.079539 .  
Z4         1  74.731  74.731 11.8274  0.002236 ** 
Z5         1 164.062 164.062 25.9655 3.681e-05 ***
Z6         1  49.395  49.395  7.8176  0.010264 *  
Residuals 23 145.324   6.318                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ((145.324-134.867)/3)/(134.867/20)
[1] 0.5169043
> qf(0.05,3,20,lower.tail = FALSE)
[1] 3.098391

#fitting reduced model 2
BP2<-data.frame(Y,Z1,Z2,Z4,Z5,Z6)
plot(BP2)
RM2<-lm(BP2$Y~.,data=BP2)
summary(RM2)
anova(RM2)
plot(stdres(RM2))

> summary(RM2)

Call:
lm(formula = BP2$Y ~ ., data = BP2)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2955 -0.7415  0.1048  1.0737  4.6315 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 29.93347    8.01302    3.736   0.00102 **
Z1           0.09481    0.04129    2.296   0.03072 *
Z2           0.22583    0.06466    3.492   0.00188 **
Z4           0.12981    0.05289    2.454   0.02176 *
Z5           0.17587    0.03242    5.424 1.42e-05 ***
Z6           0.17511    0.05676    3.085   0.00506 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.515 on 24 degrees of freedom
Multiple R-squared:   0.7397,   Adjusted R-squared:   0.6855
F-statistic: 13.64 on 5 and 24 DF,   p-value: 2.32e-06

> anova(RM2)
Analysis of Variance Table

Response: BP2$Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
Z1         1  46.532  46.532  7.3555  0.012164 *
Z2         1  82.058  82.058 12.9712  0.001432 **
Z4         1  81.738  81.738 12.9207  0.001457 **
Z5         1 160.990 160.990 25.4483 3.713e-05 ***
Z6         1  60.221  60.221  9.5194  0.005062 **
Residuals 24 151.828   6.326
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#predict overall
PDM=cbind(Z1,Z2,Z4,Z5,Z6)
BG=c(74,90, 87,70.91,61)
sey0hat<-sqrt(6.326*(1+t(BG)%*%(t(PDM)%*%PDM)^-1%*%BG))
qt(0.025,24,lower.tail = FALSE)
91.768-3.44*2.0639
91.768+3.44*2.0639

> qt(0.025,24,lower.tail = FALSE)
[1] 2.063899
> 91.768-3.44*2.0639
[1] 84.66818
> 91.768+3.44*2.0639
[1] 98.86782
```

```
#section 4
#decide principle component and the eigenvector of correlation matrix

PCTL<-cbind(PD1)
PCFM<-prcomp(PD1,scale=TRUE,scores=TRUE)
summary(PCFM)
CRTL<-cor(PD1)
eigen(CRTL)

> summary(PCFM)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     2.9269 1.8164 1.4825 1.19290 1.04720
Proportion of Variance 0.4284 0.1650 0.1099 0.07115 0.05483
Cumulative Proportion  0.4284 0.5933 0.7032 0.77434 0.82917
                           PC6
Standard deviation     0.98070
Proportion of Variance 0.04809
Cumulative Proportion  0.87726

> eigen(CRTL)

$vectors
             [,1]        [,2]         [,3]        [,4]
 [1,]  0.10552236  0.21055342  0.476682347  0.33179663
 [2,] -0.19121037  0.35592455  0.104234811 -0.27706697
 [3,] -0.29472886  0.15606305 -0.098814332 -0.09294418
 [4,]  0.04017204  0.21635440  0.514711443  0.34780898
 [5,] -0.23787746  0.28488679  0.050982721 -0.26834844
 [6,] -0.29533331  0.04602938 -0.086355937 -0.02562325
 [7,] -0.18584617  0.31150867  0.193775497 -0.21183992
 [8,] -0.27846897 -0.17738183  0.150260986  0.10169843
 [9,] -0.04551325 -0.12001782  0.405327318 -0.07571083
[10,] -0.30581039 -0.12411937  0.123812423  0.13422107
[11,] -0.25384692 -0.17994893  0.122103353  0.06714324
[12,]  0.29317377 -0.13941564  0.073670971 -0.14202946
[13,]  0.29898923 -0.02750869  0.029842566 -0.26752138
[14,] -0.21099822 -0.22960362  0.237851363 -0.26855172
[15,]  0.28283286 -0.05860206  0.217585040 -0.08459219
[16,] -0.23529531 -0.33573297  0.055607349  0.09758731
[17,] -0.15804662 -0.42118682 -0.008335855  0.11599932
[18,]  0.22598021 -0.15047553  0.059136317  0.10582268
```

```
[19,]  0.12129184 -0.17700735  0.287083147 -0.50981664
[20,] -0.04384444 -0.25181833  0.152295826 -0.24957342

               [,5]
 [1,] -0.182888835
 [2,] -0.010417188
 [3,] -0.232330072
 [4,] -0.138183809
 [5,] -0.032403493
 [6,] -0.336367222
 [7,]  0.063834676
 [8,] -0.002272508
 [9,]  0.327578237
[10,]  0.060417518
[11,]  0.054548704
[12,] -0.081887985
[13,] -0.082052990
[14,]  0.165624341
[15,]  0.121980239
[16,]  0.036482688
[17,]  0.005791361
[18,] -0.326221385
[19,]  0.059231009
[20,] -0.702938616
```

# standardize the variable

M1=mean(PD1$SSC);	M11=mean(PD1$PA);
M2=mean(PD1$SSM);	M12=mean(PD1$OR);
M3=mean(PD1$SST);	M13=mean(PD1$DR);
M4=mean(PD1$MSC);	M14=mean(PD1$S);
M5=mean(PD1$MSM);	M15=mean(PD1$B);
M6=mean(PD1$MST);	M16=mean(PD1$Sp);
M7=mean(PD1$FT);	M17=mean(PD1$V);
M8=mean(PD1$DL);	M18=mean(PD1$St);
M9=mean(PD1$DF);	M19=mean(PD1$RT);
M10=mean(PD1$BC);	M20=mean(PD1$OD);

```
M=matrix(c(M1,M2,M3,M4,M5,M6,M7,M8,M9,M10,M11,M12,M13,M14,M15,M16,M17,M18,M19,M20),nrow = 20)
```

PV1=PD1$SSC;	PV6= PD1$MST;	PV11= PD1$PA;	PV16= PD1$Sp;
PV2= PD1$SSM;	PV7= PD1$FT;	PV12= PD1$OR;	PV17= PD1$V;
PV3= PD1$SST;	PV8= PD1$DL;	PV13= PD1$DR;	PV18= PD1$St;
PV4= PD1$MSC;	PV9= PD1$DF;	PV14= PD1$S;	PV19= PD1$RT;
PV5= PD1$MSM;	PV10= PD1$BC;	PV15= PD1$B;	PV20= PD1$OD;

```
PV=matrix(c(PV1,PV2,PV3,PV4,PV5,PV6,PV7,PV8,PV9,PV10,PV11,PV12,PV13,PV14,
PV15,PV16,PV17,PV18,PV19,PV20),nrow = 30)
```

S1=sqrt(sum((PV1-M1)^2)/29)	S11=sqrt(sum((PV11-M11)^2)/29)
S2=sqrt(sum((PV2-M2)^2)/29)	S12=sqrt(sum((PV12-M12)^2)/29)
S3=sqrt(sum((PV3-M3)^2)/29)	S13=sqrt(sum((PV13-M13)^2)/29)
S4=sqrt(sum((PV4-M4)^2)/29)	S14=sqrt(sum((PV14-M14)^2)/29)
S5=sqrt(sum((PV5-M5)^2)/29)	S15=sqrt(sum((PV15-M15)^2)/29)
S6=sqrt(sum((PV6-M6)^2)/29)	S16=sqrt(sum((PV16-M16)^2)/29)
S7=sqrt(sum((PV7-M7)^2)/29)	S17=sqrt(sum((PV17-M17)^2)/29)
S8=sqrt(sum((PV8-M8)^2)/29)	S18=sqrt(sum((PV18-M18)^2)/29)
S9=sqrt(sum((PV9-M9)^2)/29)	S19=sqrt(sum((PV19-M19)^2)/29)
S10=sqrt(sum((PV10-M10)^2)/29)	S20=sqrt(sum((PV20-M20)^2)/29)

```
S=matrix(c(S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11,S12,S13,S14,S15,S16,S17,S18,S19,
S20),nrow = 20)
lj=c(79,75,75,80,74,75,71,96,84,81,81,36,67,71,56,86,88,89,90,95)
LJ=matrix(lj,nrow = 20)
for (i in 1:20)
{
  LBJ[i]=(LJ[i,1]-M[i,1])/S[i,1]

}

td=c(92,75,35,90,73,25,73,63,73,40,56,67,88,61,85,37,42,60,88,74)
TD=matrix(td,nrow = 20)
for (i in 1:20)
{
  TD[i]=(TD[i,1]-M[i,1])/S[i,1]

}
```

```
mj=c(98,97,75,99,96,74,85,96,85,88,85,47,62,91,56,88,98,57,98,98)
MJ=matrix(mj,nrow=20)
for (i in 1:20)
{
   MJ[i]=(MJ[i,1]-M[i,1])/S[i,1]

}
lb=c(97,95,94,94,93,90,89,72,76,79,94,61,77,75,47,64,50,65,97,91)
LB=matrix(lb,nrow=20)
for (i in 1:20)
{
   LB[i]=(LB[i,1]-M[i,1])/S[i,1]

}
```

```
        LBJ           TD           MJ           LB
  -1.323539    0.3469472     1.117941    0.9894421
 -0.5231858   -0.5231858     1.353459    1.1828549
  0.2982821   -1.6902653    0.2982821    1.2428421
  -1.089746     0.223201    1.4048536    0.7483799
 -0.4521834   -0.5297006    1.2531941    1.0206426
  0.4525602   -2.0249593    0.4030098    1.1958161
 -0.6331559   -0.4513887    0.6392148    1.0027493
  0.9172888     -1.54373    0.9172888    -0.872543
  0.5958749   -0.5608235    0.7010293   -0.2453603
  0.4175822   -1.7048399     0.779947    0.3140495
  0.3294579   -1.2443857    0.5812729    1.1478567
 -0.8817733    0.5904873   -0.3593582    0.3055336
 -0.2691229    1.0350881   -0.5796494      0.35193
  0.1508305   -0.7193456    1.8911828     0.498901
 -0.1211373    1.3426056   -0.1211373   -0.5754024
  0.7366212   -2.2219393    0.8573787   -0.5917121
  0.8010579    -1.757877     1.357348   -1.3128448
   1.171817   -0.9215728   -1.1381303   -0.5606435
  0.3768984    0.2176456    1.0139099    0.9342835
   1.354328   -1.5676816    1.7717585    0.7977551
```

```
>> LBJ'*eignv

ans =
   -1.0275   -2.3707   -0.6348   -0.4051   -0.7888

>> TD'*eignv

ans =

    4.4460    2.0203   -0.5134   -0.1616    1.8579

>> MJ'*eignv

ans =

   -2.7218   -0.4809    3.0197   -0.9712   -0.7105

>> LB'*eignv

ans =

   -1.1484    1.9187    1.3100   -1.4902   -1.3467
```