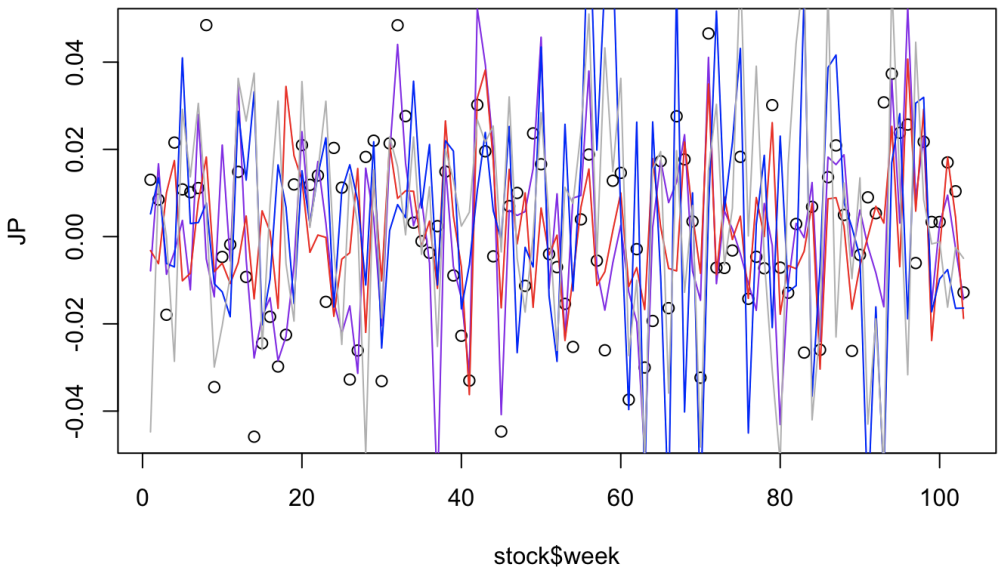
Conduct PCA.

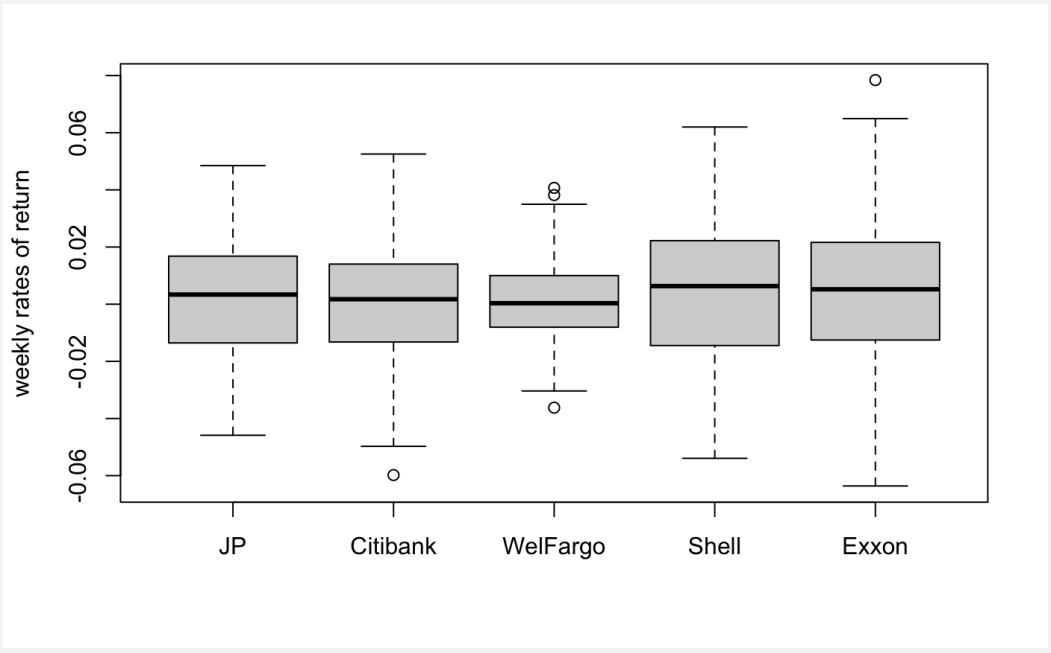
1. Introduction

Question:
How to find the correlation between the weekly rates of return for five stocks (JP Morgan (JP), Citibank, Wells Fargo, Royal Dutch Shell (Shell), and ExxonMobil (Exxon)) listed on the New York Stock Exchange were determined for the period January 2004 through December 2005. We see that there are 2 groups based on the fields: the first group is banks, the second group is gas and oil companies. However, when looking at graphs about standard deviation and boxplot, we can not see the correlation.
Answer:
PCA helps us see the correlation between 6 stocks.

2. Summary

Data includes weekly rates of return for five stocks. Through the boxplot, we can see the mean and standard deviation of the rate of return are nearly the same. So we need to use PCA to find another simple coordinate without much loss of information.
Line graph describes the rate of return of 5 stocks. We can guess they go together, but it's difficult to recognize.

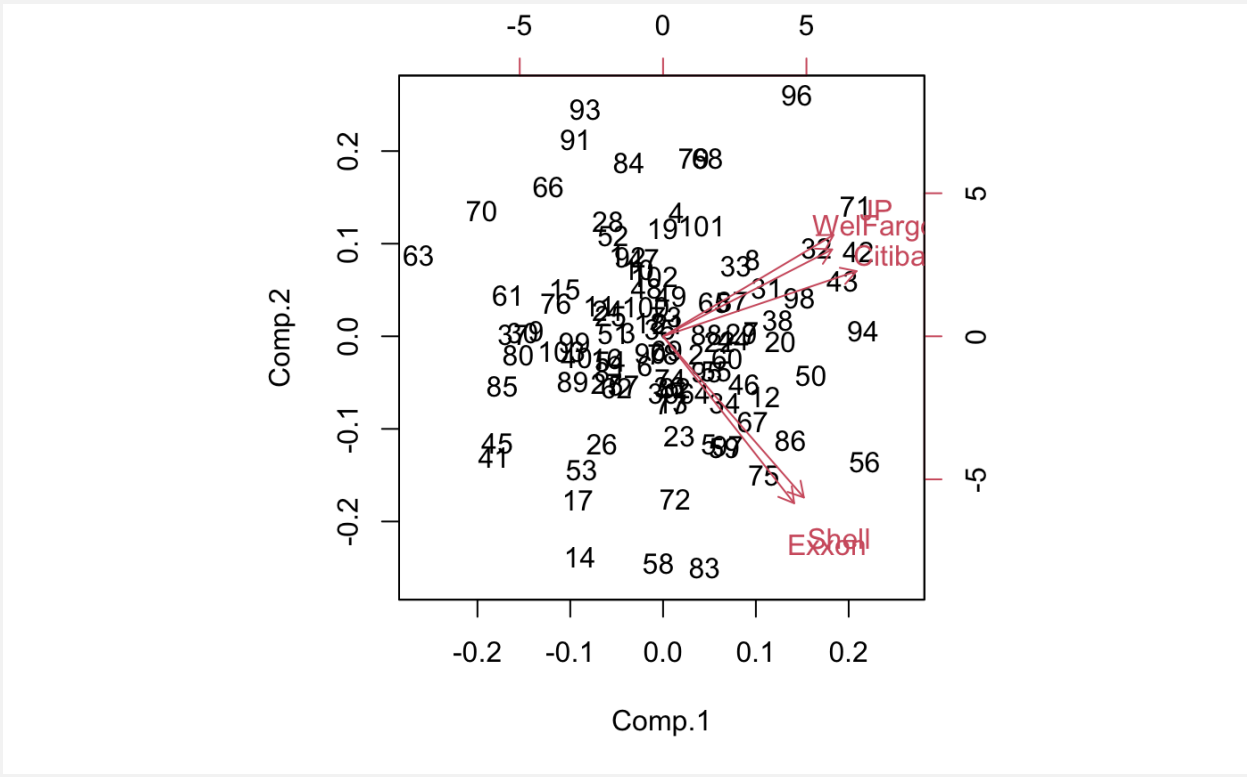




3. Analysis

Applying PCA, using the standardized variables, we obtain the first 2 sample principal components, which account for 76% of the total (standardized) sample variance, have interesting interpretations.

```
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Comp.5
Standard deviation     1.5611768 1.1861756 0.7074693 0.63248050
0.50514343
Proportion of Variance 0.4874546 0.2814025 0.1001025 0.08000632
0.05103398
Cumulative Proportion  0.4874546 0.7688572 0.8689597 0.94896602
1.00000000

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
JP        0.469  0.368  0.604  0.363  0.384
Citibank  0.532  0.236  0.136 -0.629 -0.496
WelFargo  0.465  0.315 -0.772  0.289
Shell     0.387 -0.585        -0.381  0.595
Exxon     0.361 -0.606  0.109  0.493 -0.498
```



We can see there are 2 group loading vectors: The first group is loading vectors of Shell and Exxon which have the same directions and are near together, the second is the group of 3 banks, so there is a contrast between the 2 groups. The second component represents a contrast between banking stocks and oil stocks.

4. Conclusion

PCA methods help us see the correlation between the stocks more clearly.

5. Code appendix

(in code.html file)

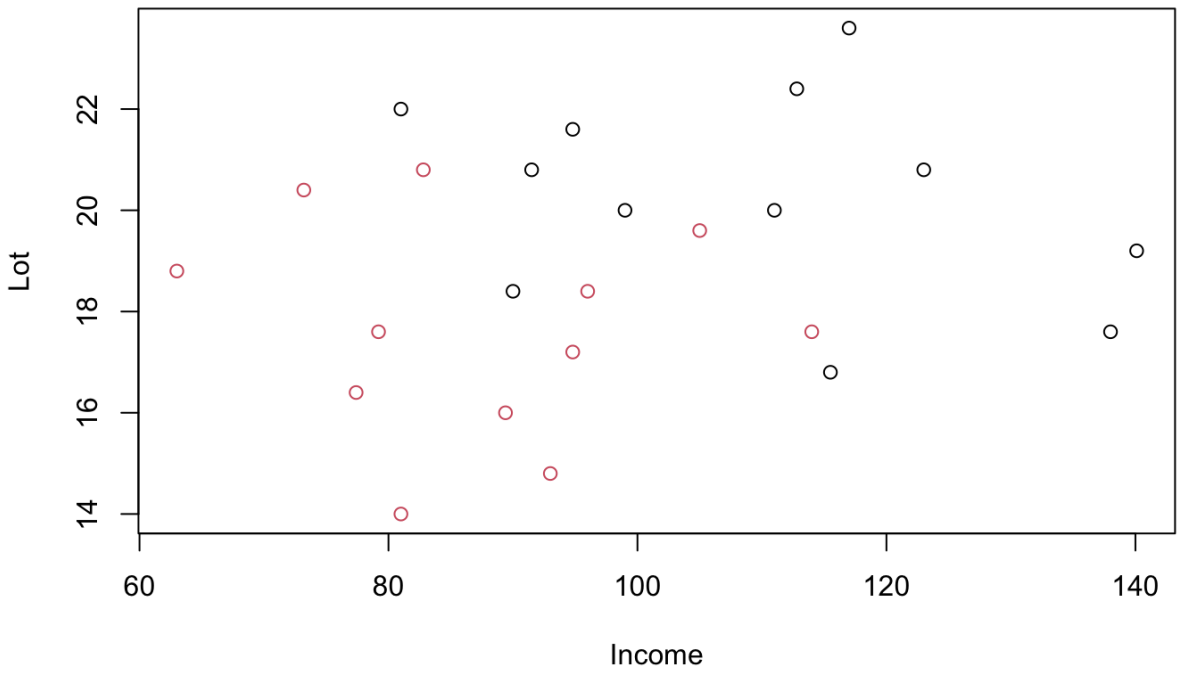
Conduct two-sample test and LDA

6. Introduction

I use the method of Linear Discriminant Analysis to classify new observations to discriminating owners from nonowners of riding mowers by income and lot size.

7. Summary

In table 11.1 there are 2 groups in a city: riding-mowers owners and those without riding mowers. In order to identify the best sales prospect for an intensive sales campaign, riding mower manufacturers are interested in classifying families as prospective owners or nonowners on the basis of income and lot size.
Dark spots represent owners, and red spots represent non-owners. We can see larger income and lot size are owners, but we can find the boundary of 2 groups. The best result of classification can reduce the cost for firms and help them understand their potential customers as long as customized the product by lot size and income.

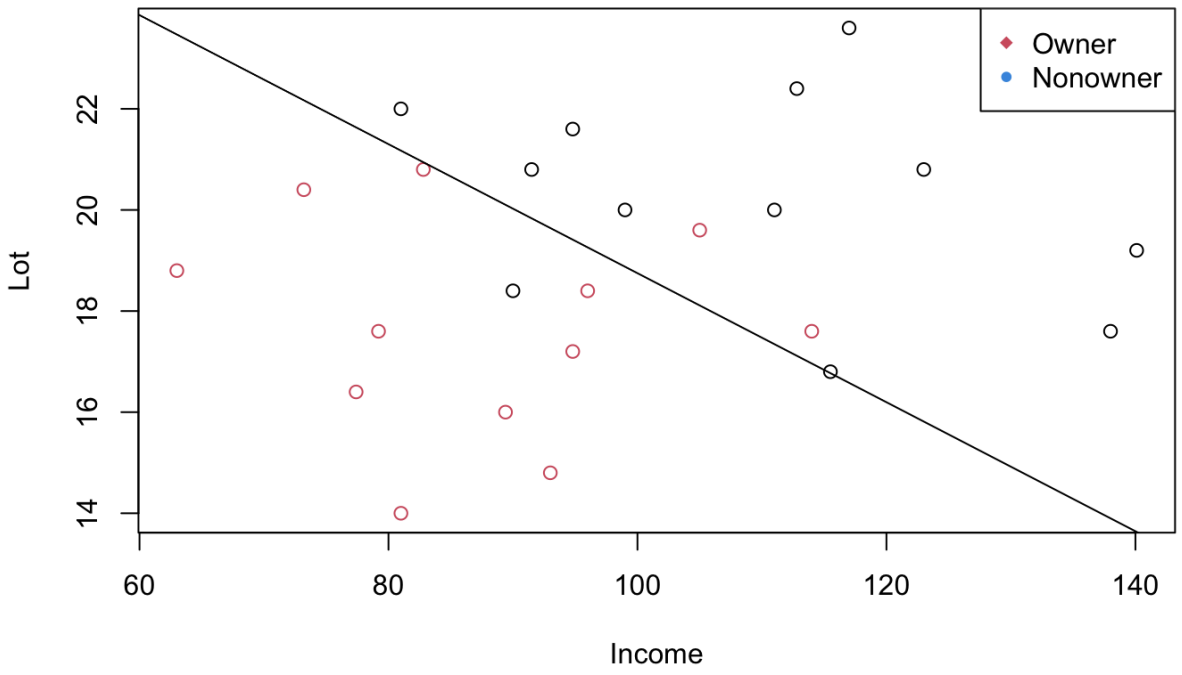


8. Analysis

```
Call:
lda(owner ~ Income + Lot, data = mover, prior = c(1, 1)/2)

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
   Income      Lot
1 109.475 20.26667
2  87.400 17.63333

Coefficients of linear discriminants:
              LD1
Income -0.0484468
Lot    -0.3795228
```

A good classification procedure should result in few misclassifications.
Using LDA we can draw the boundary of 2 groups, the coefficient of linear discriminants is Y = - 0.0484468 * income - 0.3795228 * lot.
We use the data in table 11.1 to test the result. The apparent error rate is 3/24.
The expected actual error rate by Lachenbruch's holdout is 5/25 = 20%.
If we have more data, the prediction is more accurate.

```
true_class  Income  Lot
Income         11  1
Lot             2 10
```



9. Conclusion

LDA methods help us predict who wants to buy riding mowers on income and lot size. The data about income and lot size is easy to collect, but firms can have huge benefits to recognizing their potential customers.

10. Code appendix

(in code.html file)

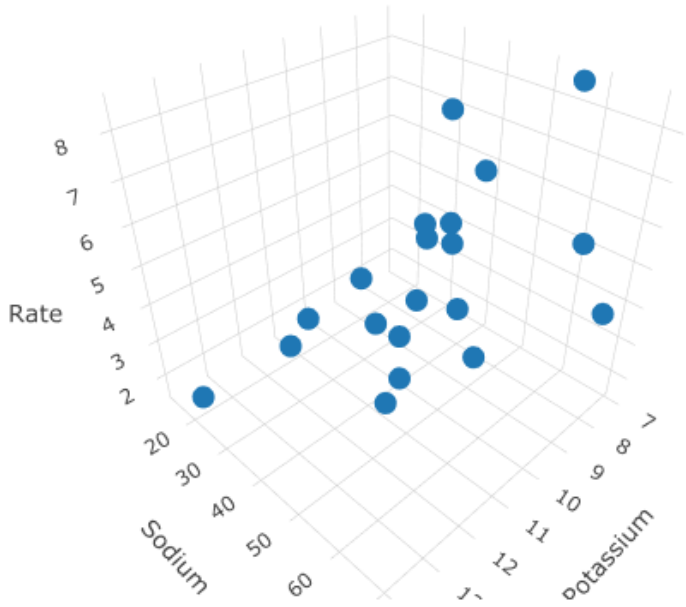
Conduct multiple linear regression

11. Introduction

I use the method of conducting multiple linear regression to find the trend of data and predict inference by testing a multivariate mean vector. Perspiration from 20 healthy females was analyzed. Three components are sweat rate, sodium content, and potassium content.

12. Summary

Looking into the plot we can see the trend of 3 components. Small rates go with small sodium and high potassium, so with the data, we can use linear regression to find the trend.



13. Analysis

In table 5.1, I use Sodium and Potassium components to find sweat rate:
Sweat rate ~ 0.03 * Sodium - 0.44 * Potassium

```
Call:
lm(formula = Y ~ Z)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.0372 -0.6842 -0.2219  0.8429  2.1846

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.31500    2.18870   3.342  0.00386 **
Z                NA         NA      NA       NA
ZSodium      0.03768    0.02292   1.644  0.11857
ZPotassium  -0.44010    0.17010  -2.587  0.01918 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.381 on 17 degrees of freedom
Multiple R-squared:  0.4075,     Adjusted R-squared:  0.3378
F-statistic: 5.846 on 2 and 17 DF,  p-value: 0.01169
```

The error rate of the model is 40%, because we only have 20 observations, and the sweat rate depends on more components. However, the model can help us little to predict the sweat rate in females.
The table below estimated covariance of hat{beta}. The linear between data is not clear, so I apply some hypothesis tests.

```
                            Sodium     Potassium
            4.79040135 -0.0319908667 -0.3254068133
Sodium     -0.03199087  0.0005253636  0.0008167945
Potassium -0.32540681  0.0008167945  0.0289337023
```

Confidence interval for beta_j from [-0.01067973 , 0.08603764]
Hypothesis: H0: beta_1 = beta_2 =0
Because F-test is larger than cval so we reject the null hypothesis. So we can’t conclude that there is no relationship between Potassium, Sodium with sweat rate.

14. Conclusion

After observing the relationship between Potassium, Sodium, and sweat rate we can conduct the model. The relationship between them is not clear so we use some hypothesis tests to test our hypothesis and we reject the null hypothesis there is no relationship between them.

.

15. Code appendix

(in code.html file)