第5回
入門機械学習　読書会

株価データを使って遊んでみる！

やじゅ＠静岡Developers勉強会

はじめに

- おわび m(_ _)m

8章の「PCA：株式市場指標の作成」を一通りやってみたのですが、説明する上では難しく、本書に沿って進行するのは、今回あきらめました。

テーマだけ頂いて、オリジナルで進行させて頂きます。
「相関」は5章でやっているけど、再度復習ってことで。

- 資料場所

https://github.com/yaju/ShizuDev_R

- 8章を実行する方への注意

ymd(Date)では、下記のエラーとなる。
'nzchar()' requires a character vector

Date = as.Date(Date, "%Y-%m-%d")　とする。

アジェンダ

- 株の基本知識
- 株価を取得してみよう
- 相関係数について
- 主成分分析(PCA)について
- その他

株の基本知識

- 日経平均株価とは

東証1部上場銘柄中から流動性や業種等のバランスを考慮して選んだ225銘柄の株価の単純平均。 代表例 トヨタ自動車、ソニー
日本経済新聞社が算出、公表しており、採用銘柄は毎年見直される他、臨時に入れ替えがなされることもある。

アメリカ合衆国では同じようにダウ平均株価があり、経済ニュース通信社であるダウ・ジョーンズ社(米)が算出する代表的な株価指数である。

銘柄コード	業種	銘柄コード	業種
1300番台	水産・農業	4000番台	化学・薬品
1500番台	鉱業	5000番台	資源・素材
1600番台	鉱業(石油/ガス開発)	6000番台	機械・電機
1700番台～1900番台	建設	7000番台	自動車・輸送機
2000番台	食品	8000番台	金融・商業・不動産
3000番台	繊維・紙	9000番台	運輸・通信・電気・ガス・サービス

- 株価データ

東証の場合には、平日の朝9時から11時30分まで（前場）と12時30分から15時まで（後場）

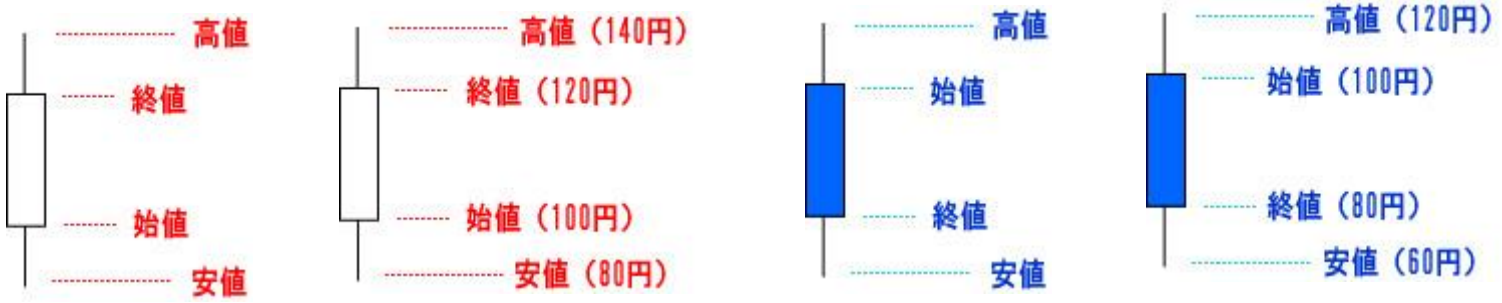
株価データには、始値、安値、高値、終値、出来高、調整後終値がある。

日付	始値	高値	安値	終値	出来高	調整後終値*
2013年8月23日	6,160	6,290	6,150	6,220	12,564,300	6,220
2013年8月22日	6,020	6,100	6,000	6,050	8,962,700	6,050
2013年8月21日	6,090	6,100	6,020	6,030	12,608,600	6,030
2013年8月20日	6,270	6,300	6,150	6,160	10,923,000	6,160
2013年8月19日	6,290	6,330	6,260	6,320	5,352,200	6,320

ローソク足

ローソク足は、江戸時代に出羽国の本間宗久が発案し、大阪・堂島の米取引で使われたという伝説が広く知られている。

始値が終値より高かったら陽線(白)、始値が終値より低かったら陰線(青)



リアルタイム株価

Yahoo!ファイナンスでは、2012年8月1日から、東証銘柄、札証銘柄、福証銘柄に限り、今まで20分遅れで表示されていた株価が、リアルタイムになりました。

※2013年7月16日より、東京証券取引所と大阪証券取引所の現物市場が統合され、大証上場銘柄の取引市場が東証に変更となりました。

その他

東証一部上場企業の最年少記録は、リブセンスの村上社長で当時25歳(2012年10月)。

その前は、グリーの田中良和社長(当時33歳)が最年少であった。

アルバイト情報を掲載するウェブサイト『ジョブセンス』を開設している。

■ビジネスモデル

ジョブセンスというアルバイト探しのウェブサービスは、リクルートが運営しているタウンワークや、フロム・エーみたいなサービスと思ってもらえれば問題ありません。

フロム・エーなどのサービスと違う点は大きく2つです。それは

・成功報酬型掲載モデル　・・・　求人が決まらなければ掲載費は不要

・祝い金モデル　・・・　仕事が決まった際はお金をプレゼント

という2つの要素を取り入れたビジネスモデルになっていることです。

株価を取得してみよう

- RFinanceYJ

ヤフーファイナンスから株価を取得することができる「RFinanceYJ」を使います。
作者は里 洋平(yokkuns)さん。マニュアルはここから。

Rstudioを使います。

```
install.packages("RFinanceYJ")
library(RFinanceYJ)

sony <- quoteStockTsData('6758.t', since='2013-01-01')
toyota <- quoteStockTsData('7203.t', since='2013-01-01')
nikkei_h <- quoteStockTsData('998407.O', since='2013-01-01')

head(sony)   #データの最初の15行くらいを出力することができる。
```

データの定義は下記
date:日付 open:始値 height:高値 low:低値 close:終値 volume:出来高
adj_close:調整後終値

plot()関数でグラフを作成

```
plot(sony[,1],sony[,5],type="l",xlab="月",ylab="終値")
```

相関係数について

- 相関係数の概念

2つの要因についてどれくらい関係が強いか？
というものを示す1つの数値データになります。

相関係数	相関関係
0.0～±0.2	ほとんど相関がない
±0.2～±0.4	やや相関がある
±0.4～±0.7	相関がある
±0.7～±0.9	強い相関がある
±0.9～±1.0	きわめて強い相関がある

トヨタ自動車とデンソーの株価

トヨタ自動車とトヨタグループであるデンソーの株価を見てみましょう。
すると、ほぼ同じ株価の動きをしていることが分かります。

Rstudioを使います。

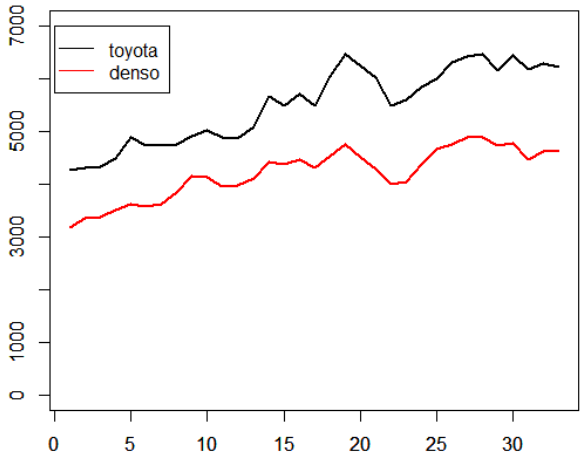
```
toyota_m <- quoteStockTsData('7203.t', since='2013-01-01',time.interval="weekly")
toyota_ts <- ts(data=toyota_m[,-1])
denso_m <- quoteStockTsData('6902.t', since='2013-01-01',time.interval="weekly")
denso_ts <- ts(data=denso_m[,-1])
```

```
plot(toyota_ts[,"close"],ann=FALSE,lwd=2,ylim=c(0,7000),col=1);par(new=T)
plot(denso_ts[,"close"],ann=FALSE,lwd=2,ylim=c(0,7000),col=2);par(new=T)
legend(0,7000,c("toyota","denso"), col = c(1:2),lwd=1, merge = TRUE)
```

実際に相関係数を求めてみます。

```
#トヨタとデンソーの相関係数
cor(toyota_ts[,"close"],denso_ts[,"close"])
```

相関係数　0.9506108



リスクヘッジ

リスクを分散させるために2つの銘柄を買うとして、それがトヨタ株とホンダ株では同じような値動きをして、リスクヘッジの意味を成しません。

このように自動車関連株の組み合わせは論外としても、トヨタ株とキャノン株のように輸出関連株同士の組み合わせも、リスク分散の観点で厳しいものがあります。

株式だけでのリスクヘッジを考えると真っ先に挙げられるのが、輸出関連株と内需関連株の組み合わせで、為替相場が円安に振れれば輸出関連株は利益の増大で値上がりし、内需関連株は輸入コストが上がり株価は下がる傾向にあります。

この「片方上がれば片方下がる」関係がリスクヘッジに最適な「逆相関関係」となります。

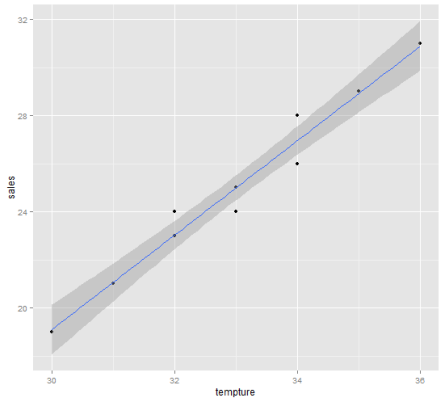
相関係数のペア抽出サイト
HTML5 kabu Charts (β)　～サヤ取り 必勝法～

相関係数のプラス値

気温とアイスクリームの売り上げのデータから見ると相関関係があります。
気温が高いほど、アイスクリームがよく売れます。

気温	30	32	33	35	36	34	32	31	33	34
売上	19	24	25	29	31	26	23	21	24	28

```
tempture <- c(30,32,33,35,36,34,32,31,33,34)
sales <- c(19,24,25,29,31,26,23,21,24,28)
ice  <- data.frame(tempture=tempture, sales=sales)
g = ggplot(ice, aes(tempture, sales))
g + geom_point()
g + geom_point() + stat_smooth(method = "lm")
```




```
#相関係数
cor(ice$sales,ice$tempture)
0.9833333
```

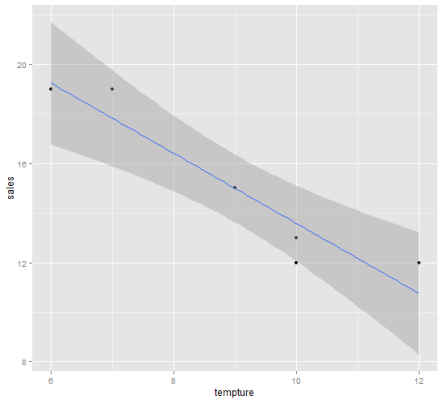
※Rのグラフィック作成パッケージ“ggplot2”について

相関係数のマイナス値

気温とおでんの売り上げのデータから見ると相関関係があります。
気温が低いほど、おでんがよく売れます。
気温とおでんの相関係数はマイナス値になります。

気温	10	7	6	12	9	10
売上	13	19	19	12	15	12

```
tempture <- c(10,7,6,12,9,10)
sales <- c(13,19,19,12,15,12)
oden  <- data.frame(tempture=tempture, sales=sales)
g = ggplot(oden, aes(tempture, sales))
g + geom_point()
g + geom_point() + stat_smooth(method = "lm")
```




```
#相関係数
cor(oden$sales,oden$tempture)
```

-0.9444444

- 相関係数と因果関係について

暑い日ほどアイスクリームがよく売れる。

⇒日中最高気温とアイスクリーム売上は正の相関関係

では、アイスクリーム売上を減らせば、日中最高気温を低くできるのか？

⇒No！

天気予報が雨ならば、実際に雨の降る日が多い。

⇒天気予報の降水確率と実際の降雨量に正の相関

では、天気予報士に「雨が降る」と言わせれば、雨で水不足を解消できるか？

⇒No！

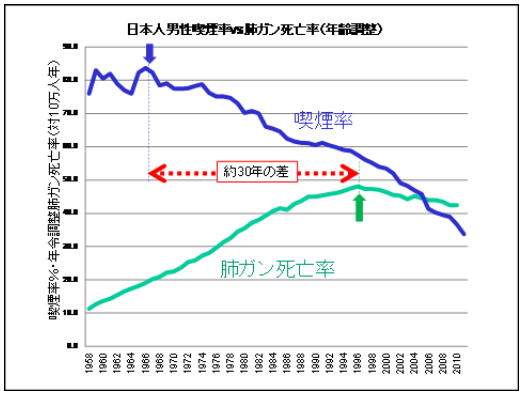
喫煙率と肺がんの因果関係

「喫煙率が下がると肺がん死が増える」のはなぜか？

http://d.hatena.ne.jp/NATROM/20120317

時系列研究を持ち出して喫煙と肺がんの関係を疑念を呈する

やり方は使い古されている



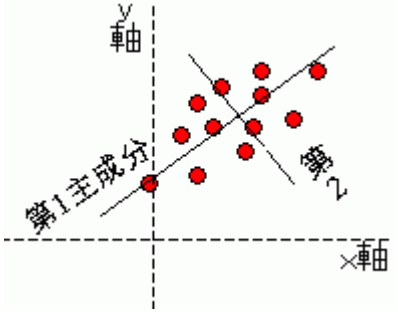
主成分分析(PCA)について

- 主成分分析(principal component analysis)

多くの変数により記述された量的データの変数間の相関を排除し、できるだけ少ない情報の損失で、少数個の無相関な合成変数に縮約して、分析を行う手法である。

・第1主成分
　X軸とY軸の散布図を書いて、点々の真中ほどに直線を引いたもの。

・第2主成分
　XとYの平均値（重心）を通って、第1主成分である直線に直角の線を引いたもの。



- 主分析分析の考え方

例）テストの合計得点の算出

・国語の平均が30点、数学の平均点が70点である時

－国語が得意なA君は国語が40点、数学が50点で、2教科の合計は90点。

－数学が得意なB君は国語が20点、数学が90点で、2教科の合計は110点。

単に足しあわせた合計得点には，数学の得点の影響がより大きく反映してしまうのではないか。

数学が得意な学生が上位を占め，国語が得意な学生の順位が低くなってしまうことになり，あまりフェアなやり方とはいえない。

主成分分析を用いると、各教科の点数に「重みづけをして」、合成得点を算出することができる。

- 大企業はどれか

参考資料:「はじめよう多変量解析～主成分分析編～」
http://www.slideshare.net/sanoche16/tokyor31-22291701

以下の8社の企業の規模順に並べたいとする。

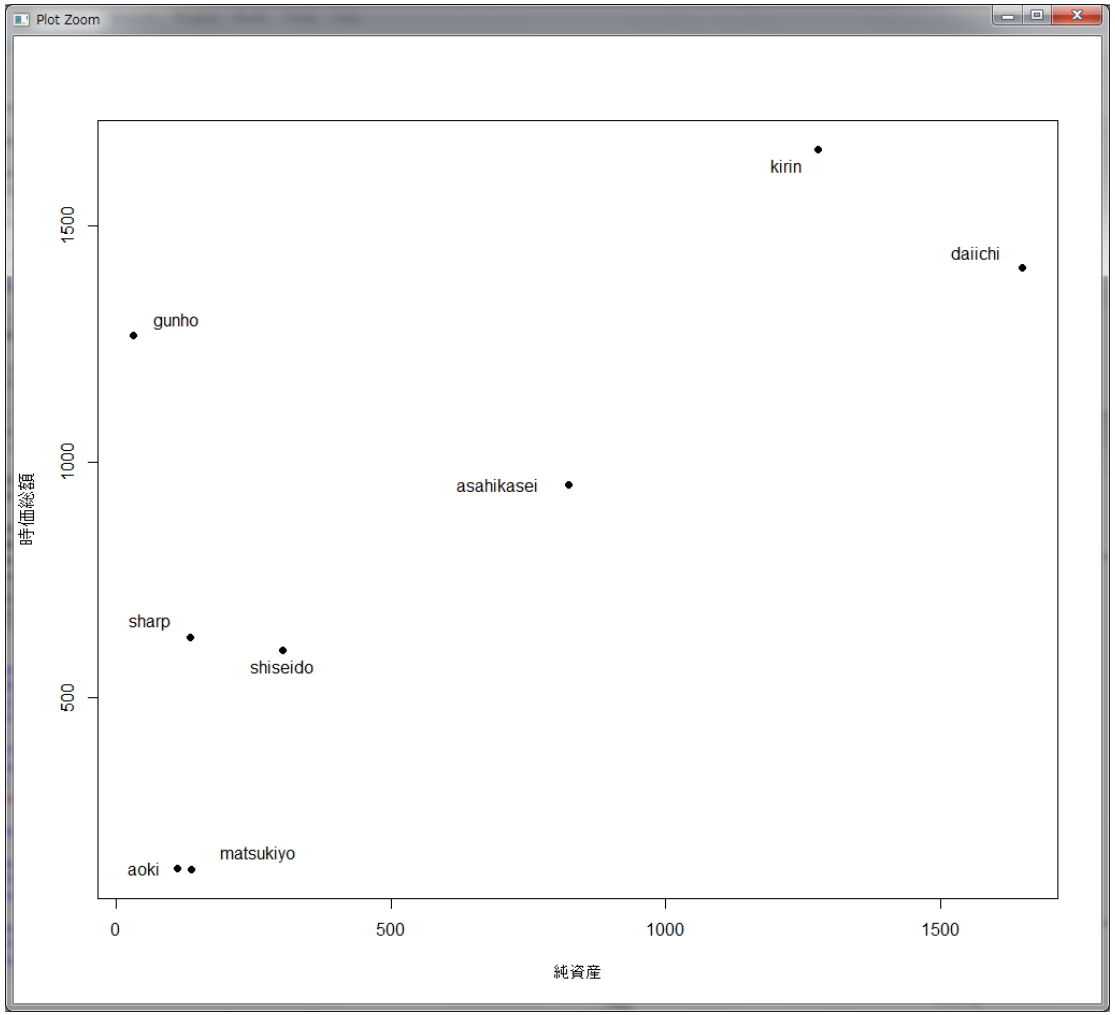
	時価総額(十億円)	純資産(十億円)
ガンホー	1,267	32
マツモトキヨシ	137	137
旭化成	952	824
キリン	1662	1278
アオキ	139	111
資生堂	601	304
第一生命	1412	1649
シャープ	629	135

散布図を書いてみる。

```
install.packages("maptools")
library(maptools)
# プロット
labs <-
c("gunho","matsukiyo","asahikasei","kirin","aoki","shiseido","daiichi","sharp")
market.price<- c(1267,137,952,1662,139,601,1412,629)
book.value <- c(32,137,824,1278,111,304,1649,135)
data <- data.frame(market.price, book.value, row.names=labs)

plot(data$book.value, data$market.price, pch=16, xlab="純資産", ylab="時価総額")
pointLabel(x=data$book.value, y=data$market.price, labels=rownames(data))
```

- 2次元だとどれが大企業がわかりにくい
  出来れば得点を付けて一列に並べたい。



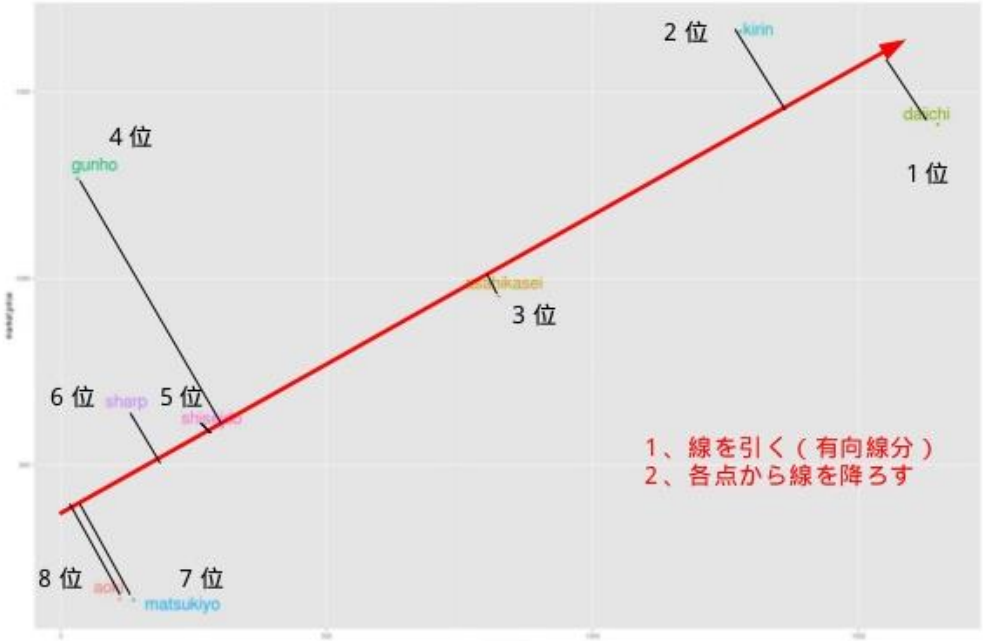
- 順位付け

得点を付ける方法を考える。

1. 有効線分

```
norns.lm <- lm(market.price~book.value , data=data)
abline(norns.lm , lwd=1 , col="red")
```

2. 各点から線を降ろす。(方法があるの？ 画像を加工しました)



主成分分析（2次元の場合）とは？

2次元データ（時価総額と純資産）を変換して1次元 （企業規模を表す得点）データに置き換えること

- 重み付けの係数

「企業の規模を時価総額と純資産の両方を考慮して評価したい」

重み付けして考える企業規模を z（主成分）とおく。

時価総額を x1 、純資産を x2 とおいて

一般式　$Z = a_1x_1 + a_2x_2$　　という式を作り上げればよい

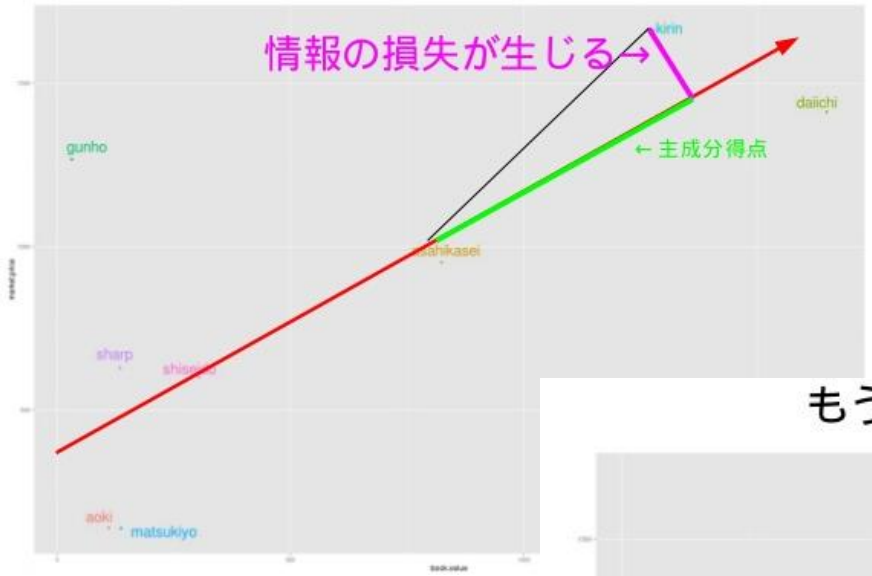
一般化

	時価総額 (x1)	純資産 (x2)	a1 × x1	a2 × x2	z
ガンホー	1,267	32	1267a1	32a2	1267a1 + 32 a2
マツモトキヨシ	137	137	137a1	137a2	137a1 + 137a2
旭化成	952	824	952a1	824a2	952a1 + 824a2
キリン	1662	1278	1662a1	1278a2	1662a1 + 1278a2
アオキ	139	111	139a1	111a2	139a1 + 111a2
資生堂	601	304	601a1	304a2	601a1 + 304a2
第一生命	1412	1649	1412a1	1649a2	1412a1 + 1649a2
シャープ	629	135	629a1	135a2	629a1 + 135a2

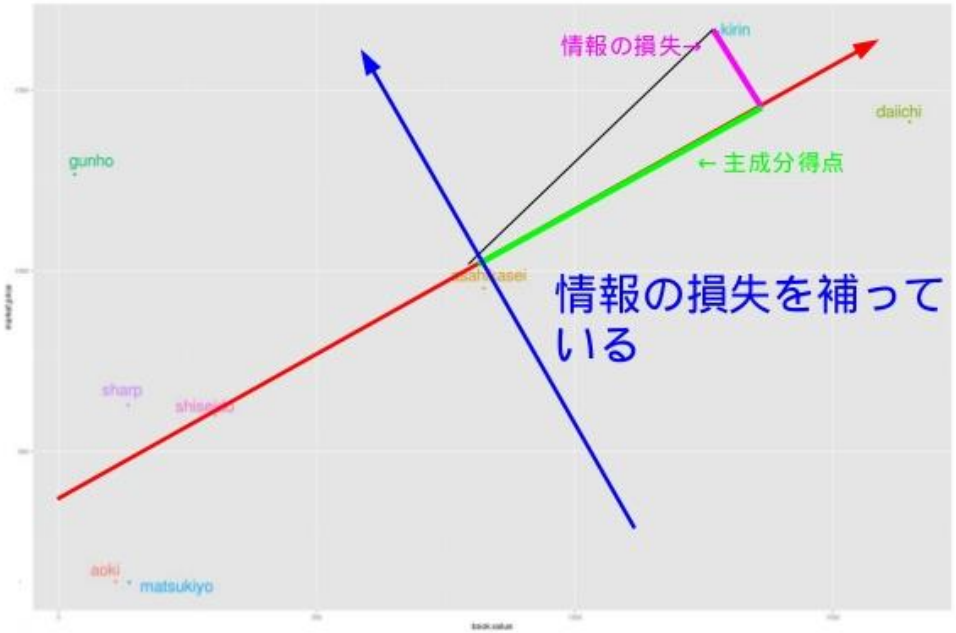
$z = a_1x_1 + a_2x_2$ とおく

- 情報の損失を補填するには

１本だけだと



もう１本引くことで



- 重み付け係数と主成分を求める

prcomp関数を使用する。

```
>pca = prcomp(data)
>pca
```

結果　PC1が第1主成分、PC2が第2主成分

```
Standard deviations:
[1] 781.1601 313.8103
Rotation:
                    PC1        PC2
market.price  -0.6672553 -0.7448291
book.value    -0.7448291  0.6672553
```

```
>summary(pca)
                          PC1      PC2
Standard deviation      781.160  313.810
Proportion of Variance    0.861    0.139  「寄与率」
Cumulative Proportion     0.861    1.000  「累積寄与率」
```

- 重み係数を当てはめる

	PC1	PC2	
market.price	-0.6672553	-0.7448291	時価総額
book.value	-0.7448291	0.6672553	純資産

一般式 $Z = a_1x_1 + a_2x_2$

Z=主成分が高いほど、企業規模が大きい。第一生命が第1位となる。

	時価総額 (x1)	純資産 (x2)	0.67 × x1	0.74 × x2	z
ガンホー	1,267	32	848.89	23.68	872.57
マツモトキヨシ	137	137	91.79	101.38	193.17
旭化成	952	824	637.84	609.76	1247.60
キリン	1662	1278	1113.54	945.72	2059.26
アオキ	139	111	93.13	82.14	175.27
資生堂	601	304	402.67	224.96	627.63
第一生命	1412	1649	946.04	1220.26	2166.30
シャープ	629	135	421.43	99.90	521.33

- 寄与率とは

>summary(pca)

	PC1	PC2	
Standard deviation	781.160	313.810	
Proportion of Variance	0.861	0.139	「寄与率」
Cumulative Proportion	0.861	1.000	「累積寄与率」

■寄与率

主成分1つだけで、どのくらいの割合の情報を説明しているかを表している。

第1主成分(PC1)が0.86あるため、第1主成分だけで、元データが持つ情報の約86%を説明していると読み取れる。

■累積寄与率

変数を縮約し、少ない数の主成分でデータを見ようとしたときに、元データの情報をほとんど含んでないまま分析を進めることを防ぐためと、すべての変数を分析に投入して、変数を縮約した意味がなくなってしまうことを防ぐための基準である。

- 主成分の解釈

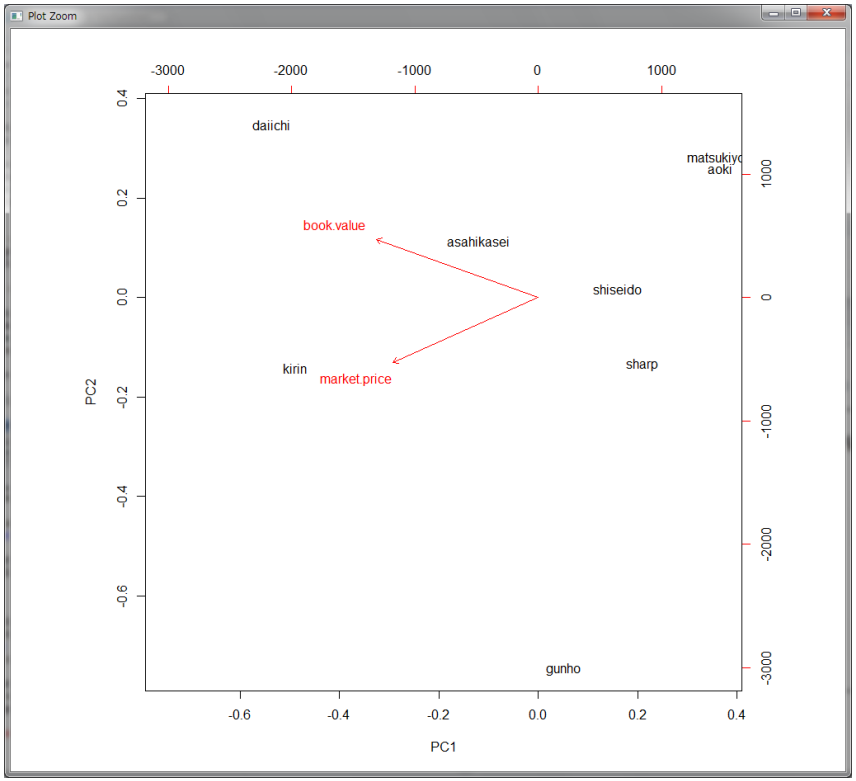
第1主成分 は、時価総額・純資産共に高ければ高いほど良い　=> 企業の規模を表す（はず）

第2主成分は時価総額が低いほどよく、純資産が高いほどよい　=> 企業への期待の少なさを表す（はず）

- 可視化

>biplot(pca)

主成分の2軸にそってデータをplotしてくれる。



- 2次元から多次元へ
  【主成分分析最終回】缶コーヒー総合力1位はどれだ？「コク」「香り」「酸味」の主成分得点を求め、散布図を描いて解釈する

http://markezine.jp/article/detail/17158

とても良い記事

Markezineの Excel ビジネス統計

	コク	香り	酸味
Sマルタ	-0.116248	1.2456822	1.5275252
モーニングS	-1.278724	-1.245682	0.0727393
BOSS	1.0462287	-0.415227	0.8001323
FIRE	1.0462287	0.4152274	-0.654654
サンタマルタ	1.0462287	1.2456822	1.5275252
BLACK無糖	0.4649906	-0.415227	-0.654654
UCCB	-1.278724	1.2456822	-1.382047
ジョージアB	-1.278724	-1.245682	-1.382047
ROOT	-0.697486	-1.245682	0.0727393
WANDA	1.0462287	0.4152274	0.0727393

- データ作成

```
KOKU<-c(-0.116248,-1.278724,1.0462287,1.0462287,1.0462287,0.4649906,-1.278724,-1.278724,-0.697486,1.0462287)
KAORI<-c(1.2456822,-1.245682,-0.415227,0.4152274,1.2456822,-0.415227,1.2456822,-1.245682,-1.245682,0.4152274)
SANMI<-c(1.5275252,0.0727393,0.8001323,-0.654654,1.5275252,-0.654654,-1.382047,-1.382047,0.0727393,0.0727393)
data<-data.frame(KOKU,KAORI,SANMI, row.names=c("Sマルタ","モーニングS","BOSS","FIRE","サンタマルタ","BLACK無糖","UCCB","ジョージアB","ROOT","WANDA"))
>data
```

- 主成分分析

```
> pca<-prcomp(data,scale.=TRUE)
> pca
Standard deviations: [1]
 1.3407225 0.8263832 0.7208008
Rotation:
                PC1          PC2          PC3
KOKU    0.6074840   -0.2324076   -0.7595722
KAORI   0.5376966    0.8241644    0.1778634
SANMI   0.5846756   -0.5164686    0.6256315
```

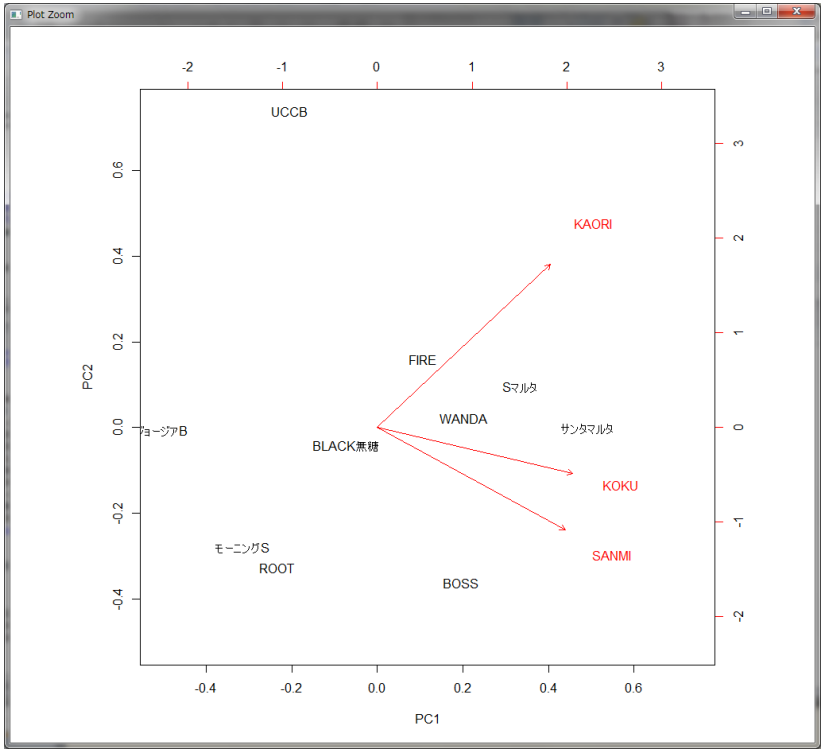
- 結果

> summary(pca)

Importance of components:

	PC1	PC2	PC3
Standard deviation	1.3407	0.8264	0.7208
Proportion of Variance	0.5992	0.2276	0.1732
Cumulative Proportion	0.5992	0.8268	1.0000

>biplot(pca)



その他

- 参考サイト

お正月だしRで株価をいじってみるよ！つまりRの紹介
http://chujo.hatenablog.com/entry/2013/01/02/235405

RFinanceYJ　マニュアル
http://cran.r-project.org/web/packages/RFinanceYJ/RFinanceYJ.pdf

はじめよう多変量解析～主成分分析編～
http://www.slideshare.net/sanoche16/tokyor31-22291701

コーヒーの「コク」「香り」「酸味」のデータから新たな評価軸を生み出すには？　Excelで相関係数行列、固有値と固有ベクトルを求める
http://markezine.jp/article/detail/17158

Chap5　主成分分析
http://www.msi.co.jp/splus/splusrescue/princomp.html

Rで学ぶデータマイニング〈2〉シミュレーション編
http://www.amazon.co.jp/dp/4274067475

ご清聴ありがとうございました！