# （1）

sns<-read.csv("snsdata.csv")

length(sns[is.na(sns$age)])

table(sns$gradyear)

n2006<-sns[sns$gradyear==2006,]

n2007<-sns[sns$gradyear==2007,]

n2008<-sns[sns$gradyear==2008,]

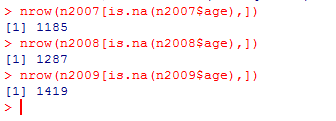
n2009<-sns[sns$gradyear==2009,]

nrow(n2006[is.na(n2006$age),])

nrow(n2007[is.na(n2007$age),])

nrow(n2008[is.na(n2008$age),])

nrow(n2009[is.na(n2009$age),])



# （2）

n1<-nrow(sns[sns$gender=="M",])

n2<-nrow(sns[sns$gender=="F",])

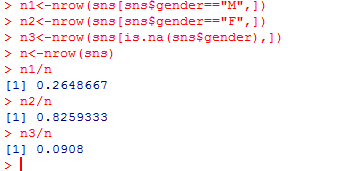
n3<-nrow(sns[is.na(sns$gender),])

n<-nrow(sns)

n1/n

n2/n

n3/n



# （3）

sns<-read.csv("snsdata.csv")

n2006<-sns[sns$gradyear==2006,]

n2007<-sns[sns$gradyear==2007,]

n2008<-sns[sns$gradyear==2008,]

n2009<-sns[sns$gradyear==2009,]

m2006<-mean(n2006[!is.na(n2006$age),]$age)

m2007<-mean(n2007[!is.na(n2007$age),]$age)

m2008<-mean(n2008[!is.na(n2008$age),]$age)

m2009<-mean(n2009[!is.na(n2009$age),]$age)

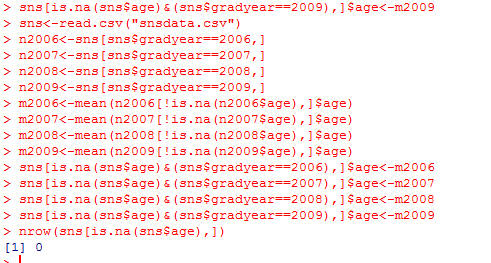
sns[is.na(sns$age)&(sns$gradyear==2006),]$age<-m2006

sns[is.na(sns$age)&(sns$gradyear==2007),]$age<-m2007

sns[is.na(sns$age)&(sns$gradyear==2008),]$age<-m2008

sns[is.na(sns$age)&(sns$gradyear==2009),]$age<-m2009

nrow(sns[is.na(sns$age),])



# （4）

#重新加载

sns<-read.csv("snsdata.csv")

f<-function(x){

x<-x[!is.na(x)]

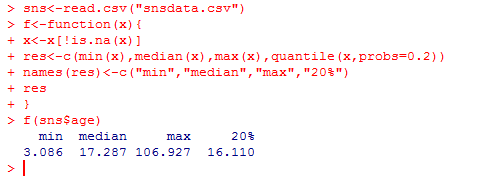
res<-c(min(x),median(x),max(x),quantile(x,probs=0.2))

names(res)<-c("min","median","max","20%")

res

}

f(sns$age)



# （5）

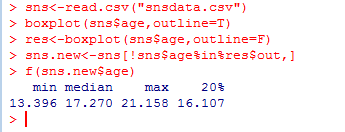
sns<-read.csv("snsdata.csv")

boxplot(sns$age,outline=T)

res<-boxplot(sns$age,outline=F)

sns.new<-sns[!sns$age%in%res$out,]

f(sns.new$age)



# （6）

#取数据

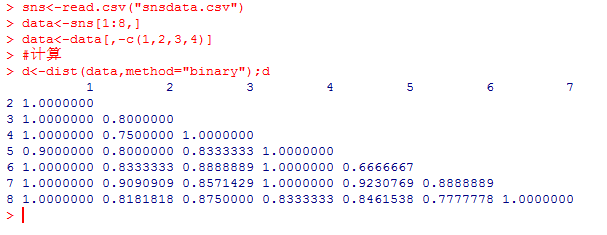
sns<-read.csv("snsdata.csv")

data<-sns[1:8,]

data<-data[,-c(1,2,3,4)]

#计算

d<-dist(data,method="binary");d



# （7）

#取数据

sns<-read.csv("snsdata.csv")

sns.new<-sns[!(is.na(sns$gender)|is.null(sns$gender)),]

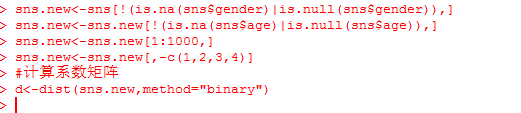
sns.new<-sns.new[!(is.na(sns$age)|is.null(sns$age)),]

sns.new<-sns.new[1:1000,]

sns.new<-sns.new[,-c(1,2,3,4)]

#计算系数矩阵

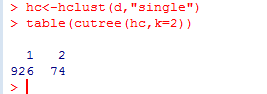
d<-dist(sns.new,method="binary")



# （8）

hc<-hclust(d,"single")

table(cutree(hc,k=2))



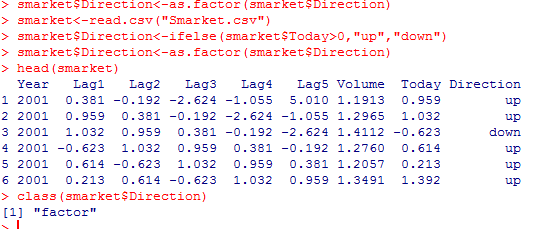
# （9）

smarket<-read.csv("Smarket.csv")

smarket$Direction<-ifelse(smarket$Today>0,"up","down")

smarket$Direction<-as.factor(smarket$Direction)

head(smarket)



# （10）

sub<-smarket[smarket$Year!=2001,]

table(sub$Year)

attach(sub)

sub.g<-glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,family=binomial)

summary(sub.g)

coef(sub.g)

> sub.g<-glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,family=binomial)

> summary(sub.g)

Call:

glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +

Volume, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.521 -1.210 1.034 1.132 1.440

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.04172 0.28203 -0.148 0.8824

Lag1 -0.11486 0.05869 -1.957 0.0503 .

Lag2 -0.05436 0.05853 -0.929 0.3530

Lag3 0.02180 0.05835 0.374 0.7087

Lag4 0.01212 0.05853 0.207 0.8360

Lag5 0.02661 0.05831 0.456 0.6482

Volume 0.09546 0.17866 0.534 0.5931

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

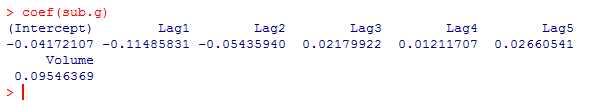
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1394.7 on 1007 degrees of freedom

Residual deviance: 1389.4 on 1001 degrees of freedom

AIC: 1403.4

Number of Fisher Scoring iterations: 4



# （11）

sub2<-smarket[smarket$Year==2001,]

pre<-predict(sub.g,newdata=sub2)

p<-exp(pre)/(1+exp(pre))

pred<-ifelse(p>0.5,"up","down")

pred<-as.factor(pred)

pred

> pred

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

up down down up up up down up up down up up up up down

16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

up up up up up up up up down up up down up up up

31 32 33 34 35 36 37 38 39 40 41 42 43 44 45

up up down down up up up down down down up up up up up

46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

up up down up up up down down down up up down up up up

61 62 63 64 65 66 67 68 69 70 71 72 73 74 75

down down up down up down up up down down up up up down down

76 77 78 79 80 81 82 83 84 85 86 87 88 89 90

down up up up up down up up up up up up up down down

91 92 93 94 95 96 97 98 99 100 101 102 103 104 105

up down up up up up up up up down down down up up up

106 107 108 109 110 111 112 113 114 115 116 117 118 119 120

up up up up up up down down down up up up up down up

121 122 123 124 125 126 127 128 129 130 131 132 133 134 135

down down up up up up up down down up up up up up up

136 137 138 139 140 141 142 143 144 145 146 147 148 149 150

up down down down up up up up up up up up up down down

151 152 153 154 155 156 157 158 159 160 161 162 163 164 165

up up up up up up up down down down up up up up down

166 167 168 169 170 171 172 173 174 175 176 177 178 179 180

up up up up up up up up up down down up up down up

181 182 183 184 185 186 187 188 189 190 191 192 193 194 195

up down up up up up down down up up up up up up down

196 197 198 199 200 201 202 203 204 205 206 207 208 209 210

up up down down up up up down down down down up up up up

211 212 213 214 215 216 217 218 219 220 221 222 223 224 225

down down up up up up up down down up up up up up down

226 227 228 229 230 231 232 233 234 235 236 237 238 239 240

down up up up up up up up down down down up up down up

241 242

down down

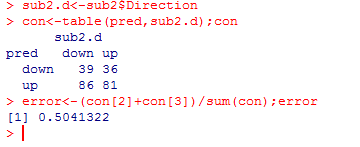
Levels: down up

# （12）

sub2.d<-sub2$Direction

con<-table(pred,sub2.d);con

error<-(con[2]+con[3])/sum(con);error



# 第三题判别分析

#取数据

se<-iris[iris$Species=="setosa",]

vir<-iris[iris$Species=="virginica",]

iris.new<-rbind(se,vir)

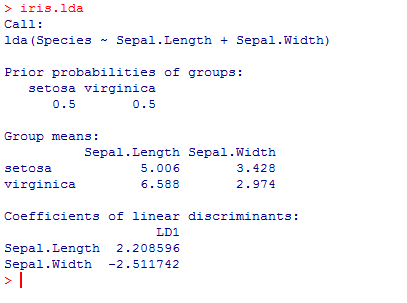
#判别

attach(iris.new)

library(MASS)

iris.lda<-lda(Species~Sepal.Length+Sepal.Width)

iris.lda



# （13）

#确定图的横纵轴区间

min(se$Sepal.Length)

min(vir$Sepal.Length)

max(se$Sepal.Length)

max(vir$Sepal.Length)

min(se$Sepal.Width)

min(vir$Sepal.Width)

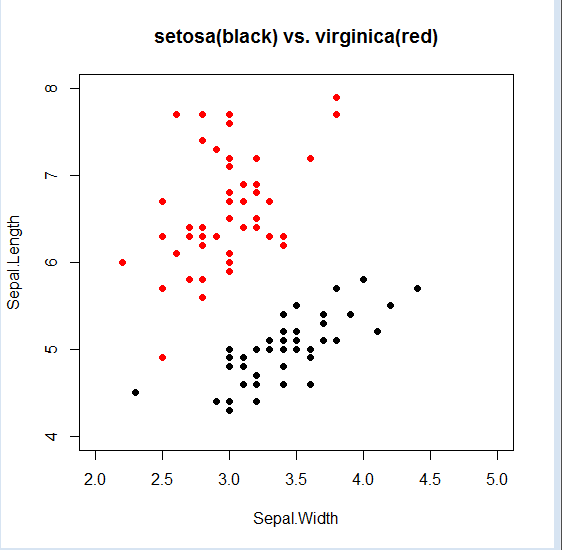
max(se$Sepal.Width)

max(vir$Sepal.Width)

#画图

plot(se$Sepal.Length~se$Sepal.Width,pch=16,xlab="Sepal.Width",ylab=c("Sepal.Length"),col="black",xlim=c(2,5),ylim=c(4,8),main=("setosa(black) vs. virginica(red)"))

points(vir$Sepal.Length~vir$Sepal.Width,pch=16,col="red")



# （14）

iris.c<-coefficients(iris.lda)

slope<--iris.c[1]/iris.c[2]



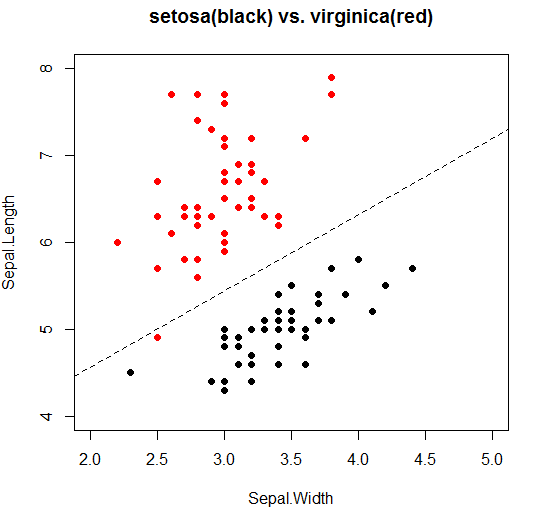
斜率为slope，截距可任取

不妨设：

Y=0.8793085\*X+2.8

# （15）

abline(2.8,slope,lty=2)



#为了美观将颜色设为绿色，线宽为2.不喜轻喷

abline(2.8,slope,col=3,lw=2,lty=2)

