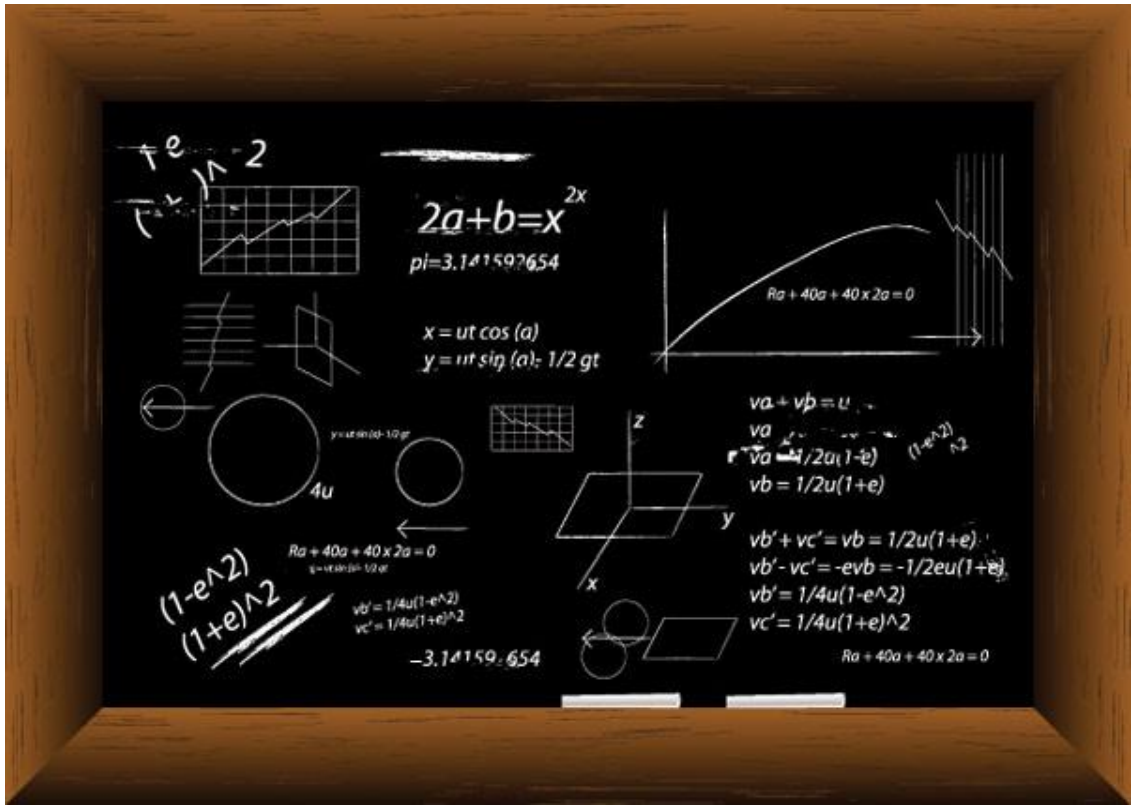
中山大學
SUN YAT-SEN UNIVERSITY



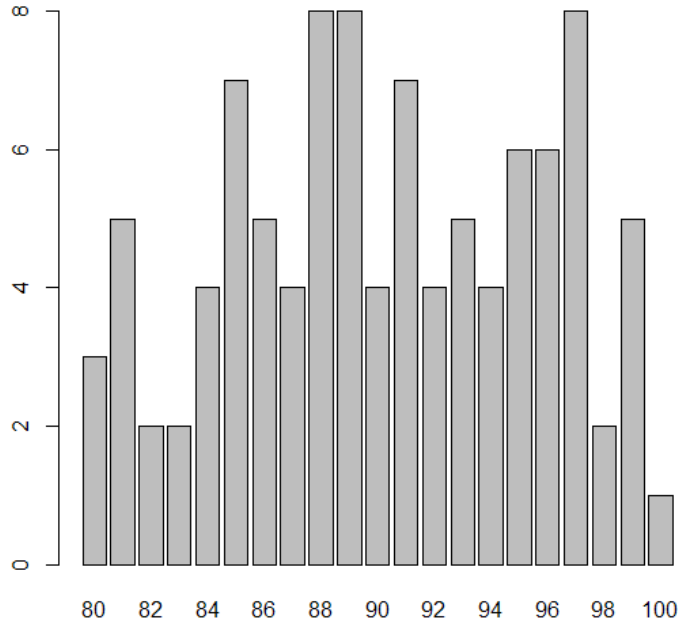
2012级《多元统计分析与数据挖掘》第2周

2015.3.12

中山大学数学与计算科学学院 黄志洪

列联表分析

- 列联函数table( )，柱状图绘制函数barplot( )



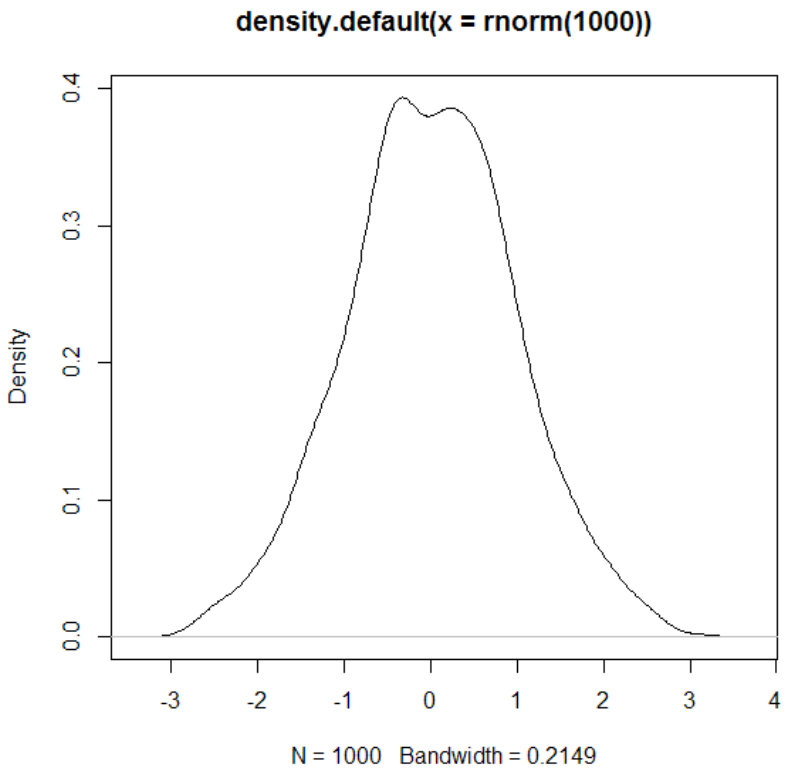

```
> table(x$x1)

 80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98
  3   5   2   2   4   7   5   4   8   8   4   7   4   5   4   6   6   8   2
 99 100
  5   1
> barplot(table(x$x1))
```

密度图

- 函数density( )

plot(density(rnorm(1000)))



R内置数据集

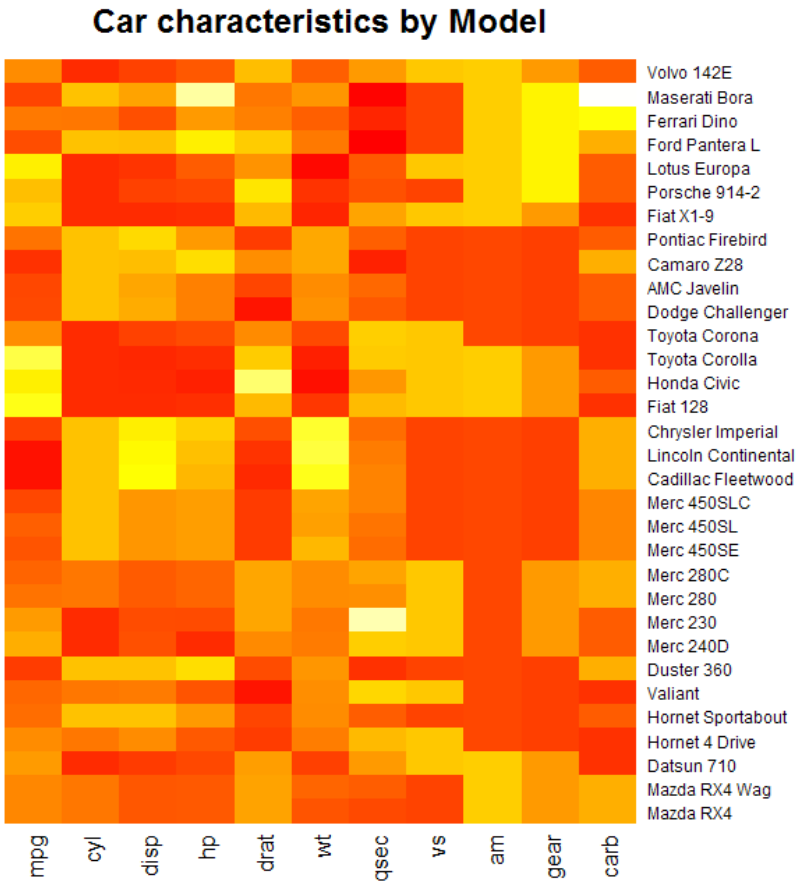
- 函数data( )列出内置数据

```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
```

热力图

- 利用内置的mtcars数据集绘制

```
heatmap(as.matrix(mtcars),
Rowv=NA,
Colv=NA,
col = heat.colors(256),
scale="column",
margins=c(2,8),
main = "Car characteristics by
    Model")
```



Iris（鸢尾花）数据集

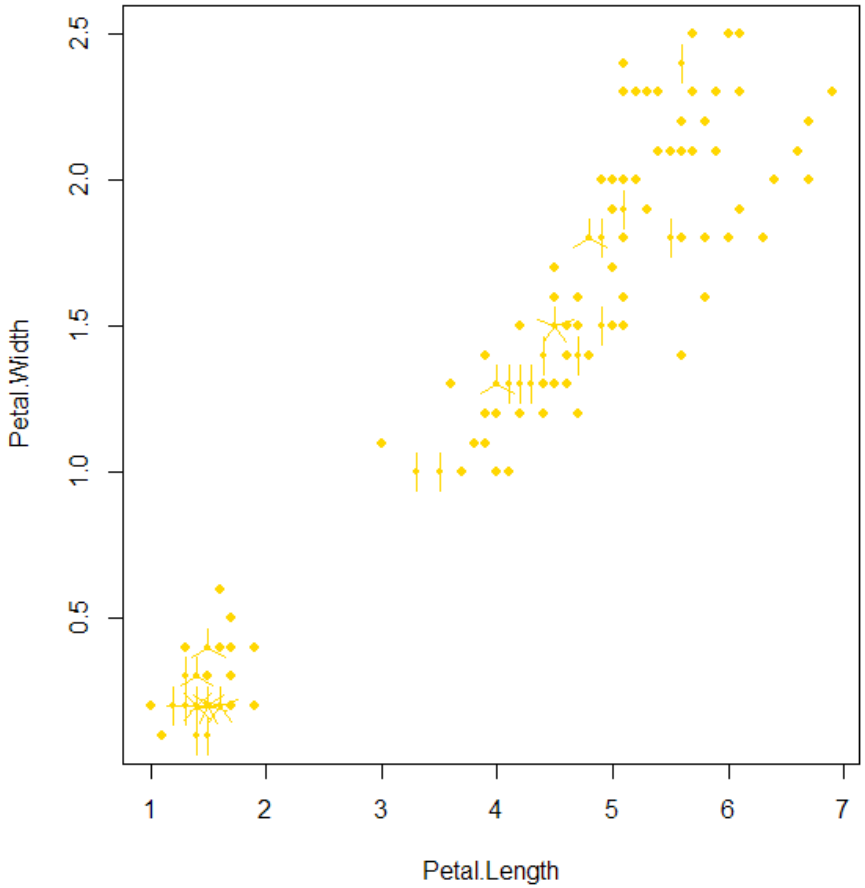
- Sepal 花萼
- Petal 花瓣
- Species 种属

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
```

向日葵散点图

- 用来克服散点图中数据点重叠问题
- 在有重叠的地方用一朵“向日葵花”的花瓣数目来表示重叠数据的个数

```
sunflowerplot(iris[, 3:4], col =
    "gold", seg.col = "gold")
```

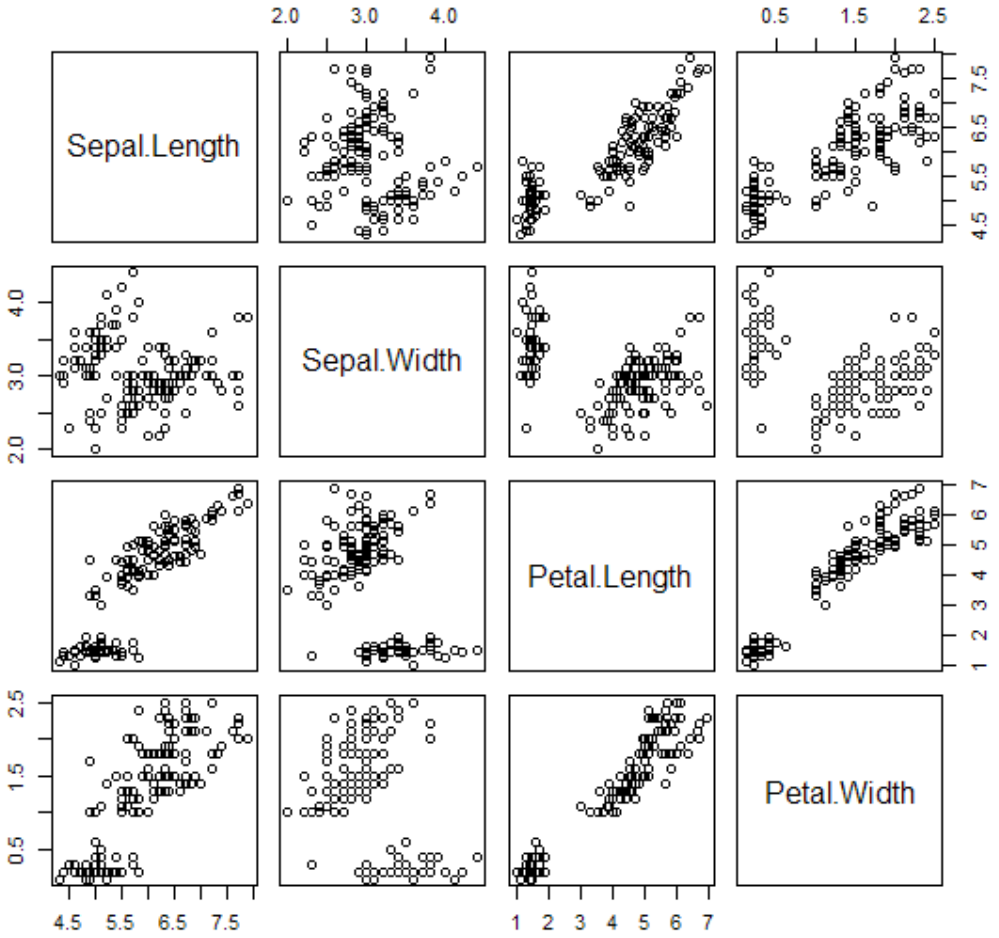


2015.3.12

散点图集

- 遍历样本中全部的变量配对画出二元图
- 直观地了解所有变量之间的关系

pairs(iris[,1:4])

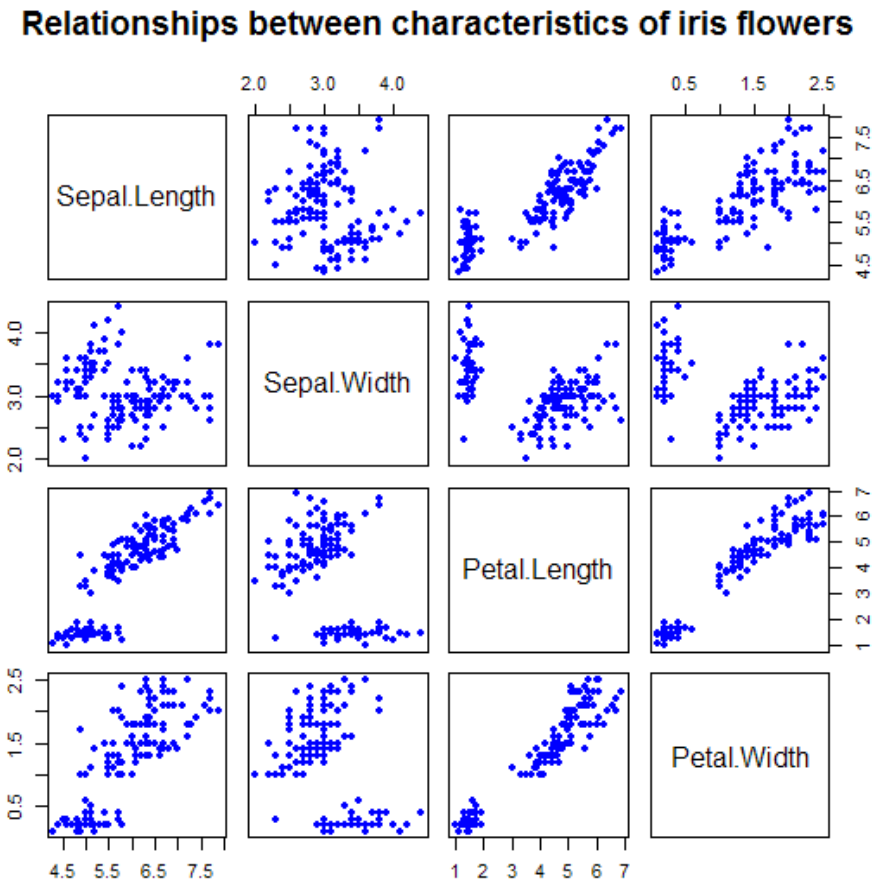


2015.3.12

散点图集

- 用plot也可以实现同样的效果

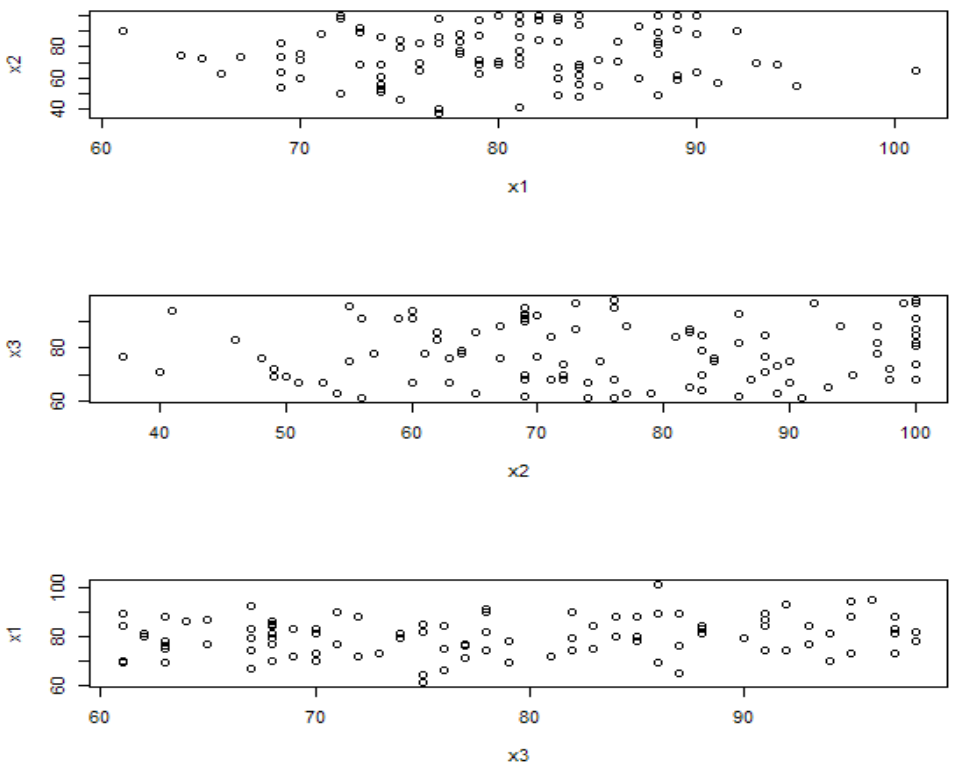
```
plot(iris[,1:4],
main="Relationships between
      characteristics of iris flowers",
pch=19,
col="blue",
cex=0.9)
```



散点图集

- 利用par( )在同一个device输出多个散点图
- Par命令博大精深，用于设置绘图参数，help(par)

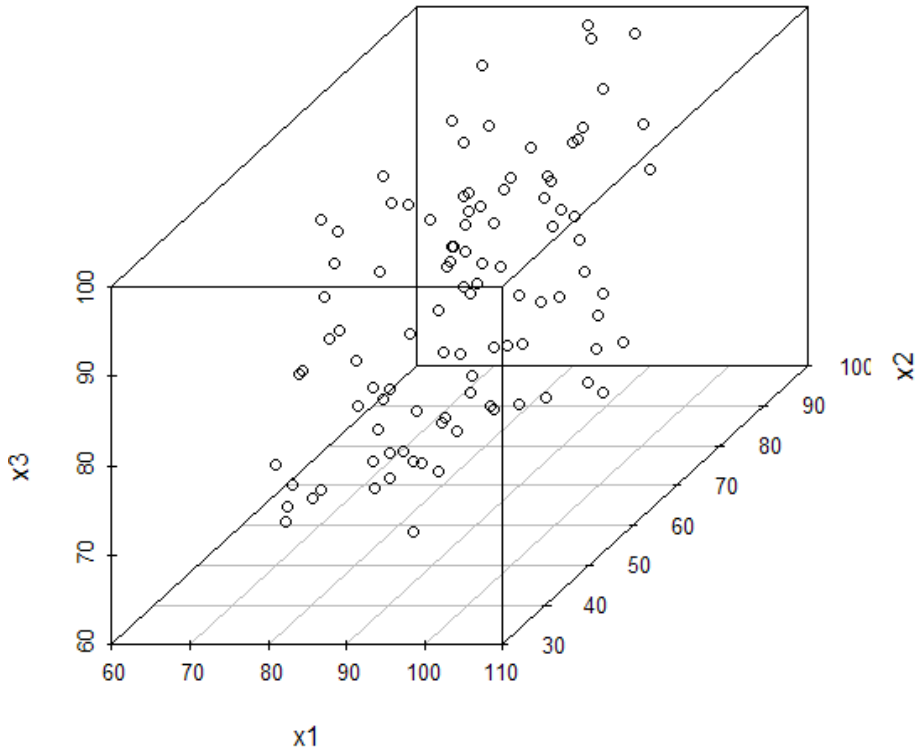
```
par(mfrow=c(3,1))
plot(x1,x2);plot(x2,x3);plot(x3,x1)
```



三维散点图

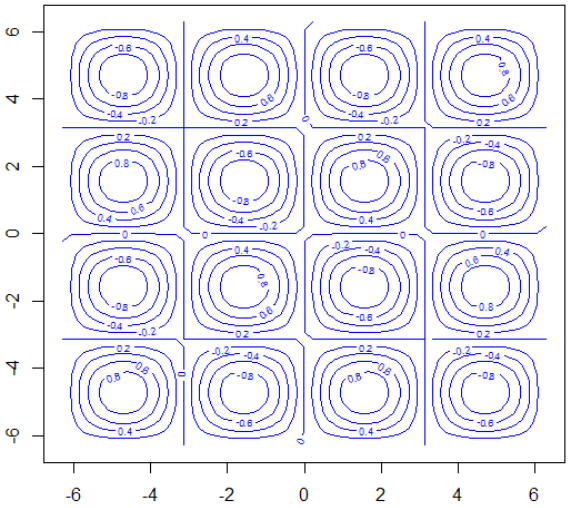
- 安装scatterplot3d 包

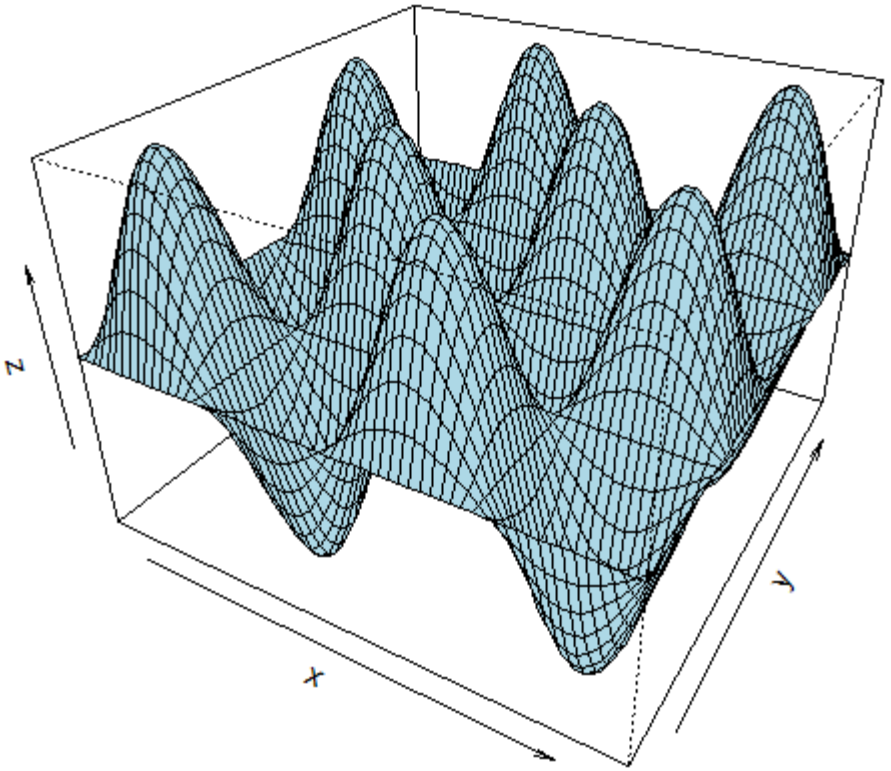
scatterplot3d(x[2:4])



三维作图

```
x<-y<-seq(-2*pi, 2*pi, pi/15)
f<-function(x,y) sin(x)*sin(y)
z<-outer(x, y, f)
contour(x,y,z,col="blue")
persp(x,y,z,theta=30, phi=30,
      expand=0.7,col="lightblue")
```





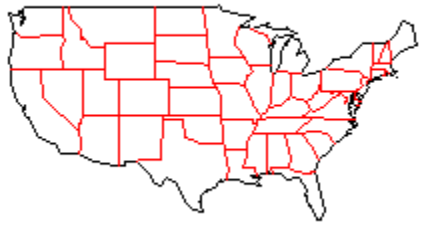
2015.3.12

地图

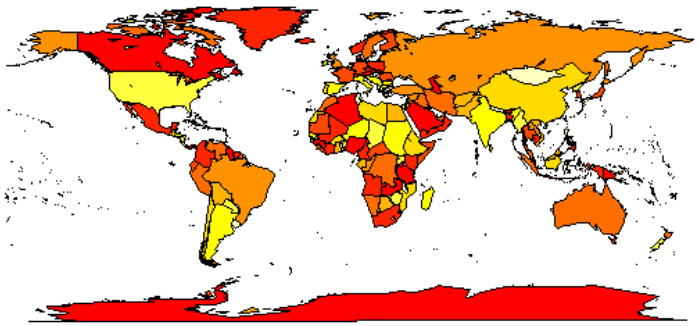
- 安装maps包

```
map("state", interior = FALSE)
```

```
map("state", boundary = FALSE, col="red",
    add = TRUE)
```



```
map('world', fill = TRUE,col=heat.colors(10))
```



调和曲线图

调和曲线图是 Andrews (安德鲁斯) 在 1972 年提出来的三角表示法，其思想是将多维空间中的一个点对应于二维平面的一条曲线，对于 p 维数据，假设 X_r 是第 r 观测值，即

$$X_r^T = (x_{r1}, x_{r2}, \cdots, x_{rp}),$$

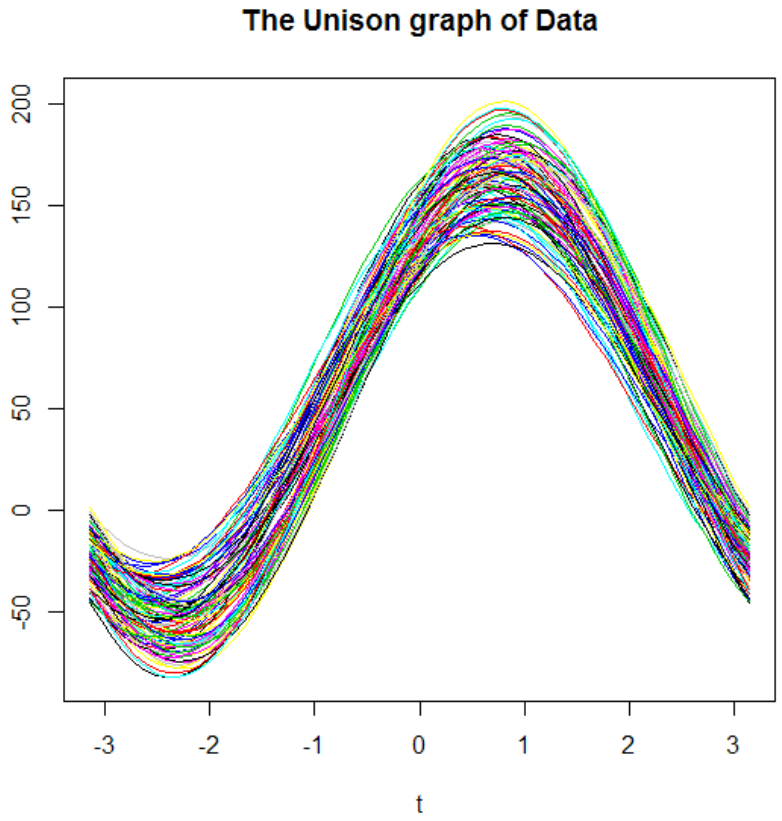
则对应的调和曲线是

$$\begin{aligned} f_r(t) &= \frac{x_{r1}}{\sqrt{2}} + x_{r2} \cdot \sin(t) + x_{r3} \cdot \cos(t) + x_{r4} \cdot \sin(2t) + x_{r5} \cdot \cos(2t) + \\ &\quad + \cdots +, \qquad -\pi \le t \le \pi. \end{aligned} \tag{3.29}$$

调和曲线图

- unison.r的代码
- 自定义函数
- 调和曲线用于聚类判断非常方便

```
> source("d:\\unison.R")
> unison(x[2:4])
>
```



R实验：社交数据可视化

- 先下载安装maps包和geosphere包并加载

```
library(maps)
library(geosphere)
```

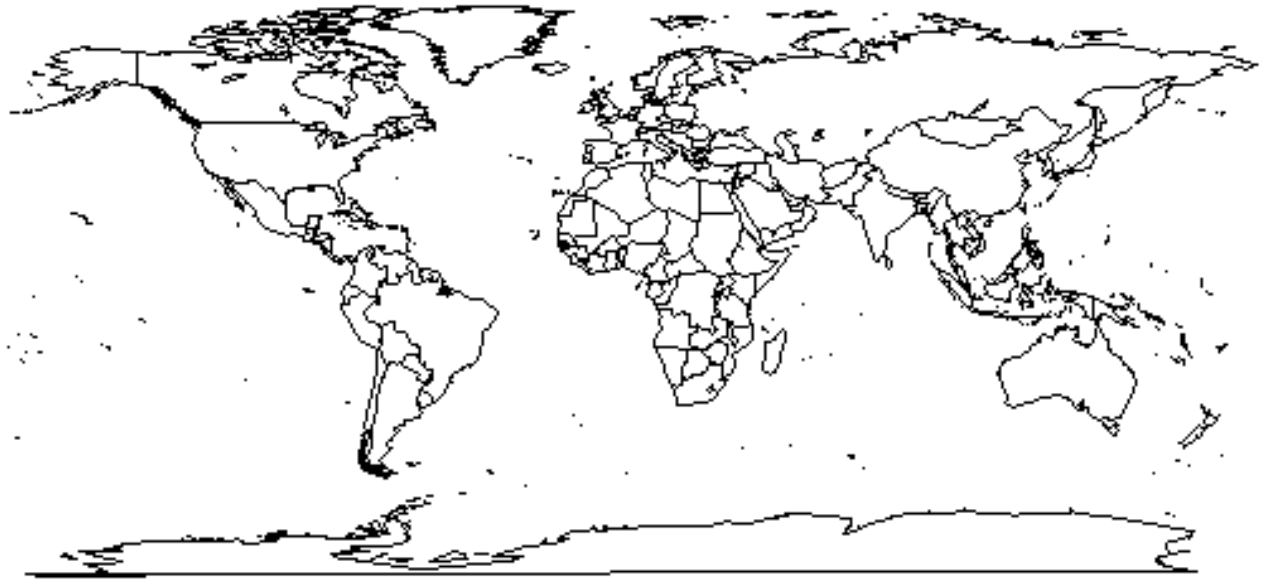
- 画出美国地图

```
map("state")
```

R实验：社交数据可视化

- 画世界地图

map("world")



R实验：社交数据可视化

- 通过设置坐标范围使焦点集中在美国周边，并且设置一些有关颜色

```
xlim <- c(-171.738281, -56.601563)
ylim <- c(12.039321, 71.856229)
map("world", col="#f2f2f2", fill=TRUE, bg="white", lwd=0.05, xlim=xlim, ylim=ylim)
```

R实验：社交数据可视化

- 画一条弧线连线，表示社交关系

```
lat_ca <- 39.164141
lon_ca <- -121.64062
5lat_me <- 45.21300
4lon_me <- -68.906250
inter <-
    gcIntermediate(c(lon_c
    a, lat_ca), c(lon_me,
    lat_me), n=50,
    addStartEnd=TRUE)
lines(inter)
```

R实验：社交数据可视化

- 继续画弧线

```
lat_tx <- 29.954935
lon_tx <- -98.701172
inter2 <-
    gcIntermediate(c(lon_ca
    , lat_ca), c(lon_tx, lat_tx),
    n=50,
    addStartEnd=TRUE)
lines(inter2, col="red")
```

R实验：社交数据可视化

- 装载数据

```
airports <- read.csv("http://datasets.flowingdata.com/tuts/maparcs/airports.csv",
    header=TRUE)
flights <- read.csv("http://datasets.flowingdata.com/tuts/maparcs/flights.csv",
    header=TRUE, as.is=TRUE)
```

R实验：社交数据可视化

- 画出多重联系

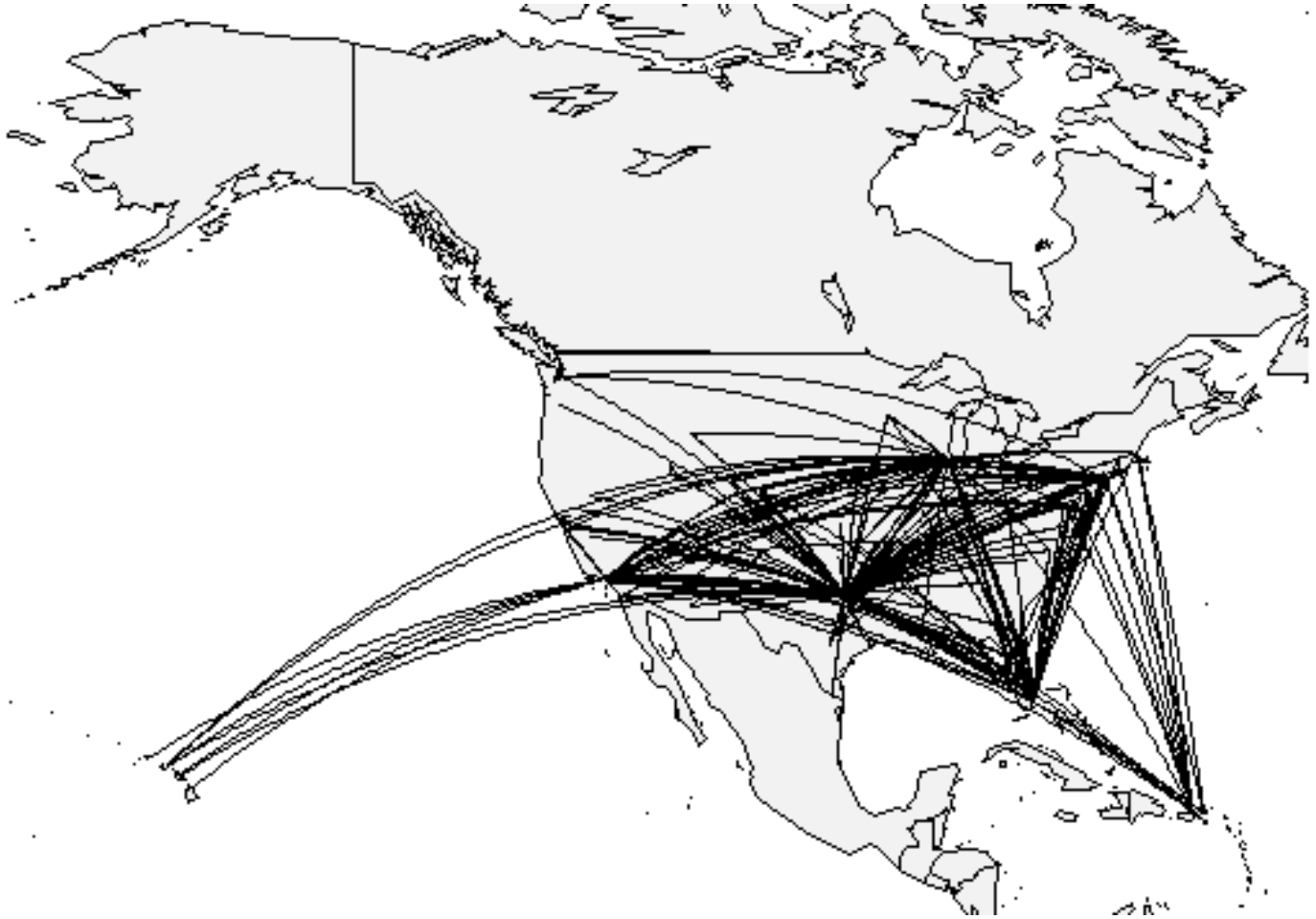
```
map("world", col="#f2f2f2", fill=TRUE, bg="white", lwd=0.05, xlim=xlim, ylim=ylim)

fsub <- flights[flights$airline == "AA",]
for (j in 1:length(fsub$airline)) {
    air1 <- airports[airports$iata == fsub[j,]$airport1,]
    air2 <- airports[airports$iata == fsub[j,]$airport2,]

    inter <- gcIntermediate(c(air1[1,]$long, air1[1,]$lat), c(air2[1,]$long, air2[1,]$lat), n=100,
    addStartEnd=TRUE)

    lines(inter, col="black", lwd=0.8)
}
```

R实验：社交数据可视化



2015.3.12

R实验：社交数据可视化

http://flowingdata.com/2011/05/11/how-to-map-connections-with-great-circles/

知识补漏：关于逻辑运算符

- 与 &
- 或 |
- 否 !
- 举例

```
> a=0
> b=1
> if (a==0 & b==1) print(1);
[1] 1
> if (a==0 | b==2) print(1);
[1] 1
> if (!b==2) print(1);
[1] 1
> if (!b==1) print(1);
>
```

知识补漏：seq()的along参数

- 生成一个和指定向量长度一样的等差数列
- 经常用在for循环里产生循环变量的变化范围

```
>
> a=c(1,2,4,3,2,5,6,2,1,3,5,6,7,4,3,7,8,2,3,5,9)
> length(a)
[1] 21
> seq(along=a)
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
> for (i in seq(along=a)) {if (a[i]>5) a[i]=5}
> a
 [1] 1 2 4 3 2 5 5 2 1 3 5 5 5 4 3 5 5 2 3 5 5
>
```

知识补漏：集合运算

- 求并函数union(x,y)
- 求交函数intersect(x,y)
- 求差函数setdiff(x,y)
- 判断属于关系is.elment(元素,集合)，相当于%in%的作用

```
> x=c(1,3,5,4,3,5,6,7)
> y=c(9,6,4,5,3,2,8)
> union(x,y)
[1] 1 3 5 4 6 7 9 2 8
> intersect(x,y)
[1] 3 5 4 6
> setdiff(x,y)
[1] 1 7
```

```
> is.element(9,y)
[1] TRUE
> is.element(10,y)
[1] FALSE
> is.element(c(2,3),y)
[1] TRUE TRUE
> 5 %in% y
[1] TRUE
```

知识补漏：is和as

- is往往用来做某种判断，返回逻辑值
- as通常用于把某种类型的数据，转换为另外一种类型

```
> x=c(1:100)
> y=c(1:100)
> is.array(x)
[1] FALSE
> is.vector(x)
[1] TRUE
> z=data.frame(x,y)
> is.dataframe(z)
错误: 没有"is.dataframe"这个函数
> is.data.frame(z)
[1] TRUE
```

```
> x=3
> as.complex(x)
[1] 3+0i
> z=as.complex(x)
> Re(x)
[1] 3
> Im(x)
[1] 0
> y=4+5i
> x*y
[1] 12+15i
```

知识补漏：因子与factor()，聚组

- 什么是因子？
- 什么是聚组？

```
> x=c("ABC","DEF","MNL","ABC","MNL","MNL","DEF")
> factor(x)
[1] ABC DEF MNL ABC MNL MNL DEF
Levels: ABC DEF MNL
```

因为离散变量有各种不同表示方法，在 R 软件中，为了统一起见，使用因子 (factor) 来表示这种类型的变量. 例如，知道 5 位学生的性别，用因子变量表示

```
> sex <- c("M","F","M","M", "F")
> sexf <- factor(sex); sexf
[1] M F M M F
Levels:  F M
```

函数 factor() 用来把一个向量编码成为一个因子. 其一般形式为:

```
factor(x, levels = sort(unique(x), na.last = TRUE),
       labels, exclude = NA, ordered = FALSE)
```

知识补漏：关于绘图参数

- help(par)
- 有哪些颜色？ colors( )

```
> colors()
  [1] "white"           "aliceblue"       "antiquewhite"
  [4] "antiquewhite1"   "antiquewhite2"   "antiquewhite3"
  [7] "antiquewhite4"   "aquamarine"      "aquamarine1"
 [10] "aquamarine2"     "aquamarine3"     "aquamarine4"
 [13] "azure"           "azure1"          "azure2"
 [16] "azure3"          "azure4"          "beige"
 [19] "bisque"          "bisque1"         "bisque2"
 [22] "bisque3"         "bisque4"         "black"
 [25] "blanchedalmond"  "blue"            "blue1"
 [28] "blue2"           "blue3"           "blue4"
 [31] "blueviolet"      "brown"           "brown1"
 [34] "brown2"          "brown3"          "brown4"
 [37] "burlywood"       "burlywood1"      "burlywood2"
 [40] "burlywood3"      "burlywood4"      "cadetblue"
 [43] "cadetblue1"      "cadetblue2"      "cadetblue3"
 [46] "cadetblue4"      "chartreuse"      "chartreuse1"
 [49] "chartreuse2"     "chartreuse3"     "chartreuse4"
```

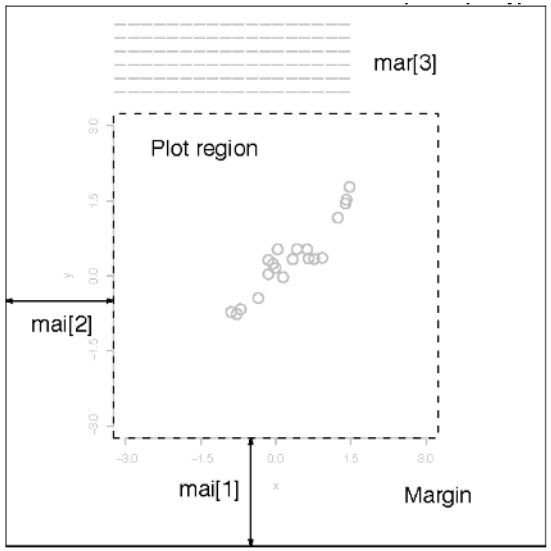
知识补漏：关于绘图参数

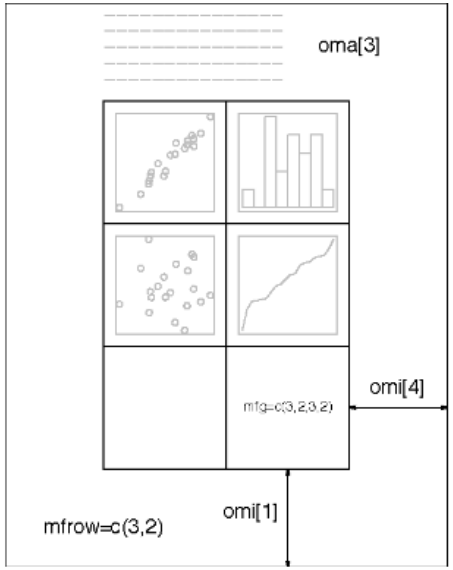
- 绘图设备

```
dev.cur()
dev.list()
dev.next(which = dev.cur())
dev.prev(which = dev.cur())
dev.off(which = dev.cur())
dev.set(which = dev.next())
dev.new(...)
graphics.off()
```

知识补漏：关于绘图参数

- 位置控制参数
- mai参数：A numerical vector of the form c(bottom, left, top, right) which gives the margin size specified in inches.
- oma参数：A vector of the form c(bottom, left, top, right) giving the size of the outer margins in lines of text.





2015.3.12

多元数据的数据特征

- 方差与协方差、相关系数

$$
\begin{aligned}
s_{xx} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2,\\
s_{yy} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2,\\
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}),
\end{aligned}
$$

则称 s_{xx} 为变量 X 的观测样本的方差，称 s_{yy} 为变量 Y 的观测样本的方差，称 s_{xy} 为变量 X,Y 的观测样本的协方差．称

$$
S=\begin{bmatrix} s_{xx} & s_{xy}\\ s_{xy} & s_{yy}\end{bmatrix}
$$

为观测样本的协方差矩阵．称

$$
r=\frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}
$$

为观测样本的相关系数．

协方差与相关系数计算

- 函数cov( )和cor( )

```
> cov(x$x1,x$x2)
[1] 4.928283
> cor(x$x1,x$x2)
[1] 0.03982364

> cov(x[2:4])
           x1          x2         x3
x1 57.626263   4.928283  16.15152
x2  4.928283 265.759495  10.61010
x3 16.151515  10.610101 125.03030
> cor(x[2:4])
            x1          x2          x3
x1 1.00000000 0.03982364 0.19028099
x2 0.03982364 1.00000000 0.05820596
x3 0.19028099 0.05820596 1.00000000
>
```

相关性检验

```
> cor.test(x$x1,x$x2)

        Pearson's product-moment correlation

data:  x$x1 and x$x2
t = 0.3945, df = 98, p-value = 0.694
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1578290  0.2344082
sample estimates:
       cor
0.03982364
```

相关分析与回归分析

- 变量之间的关系

函数关系：有精确的数学表达式

相关关系：非确定性关系

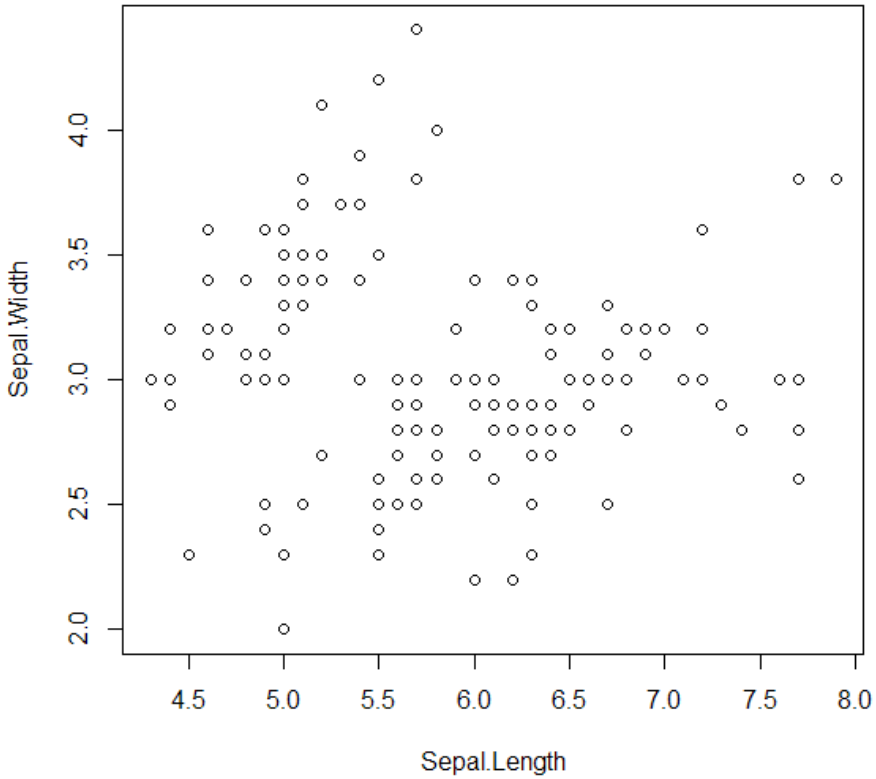
平行关系：相关分析（一元，多元）

依存关系：回归分析（一元，多元）

相关分析的例子

- Iris数据集
- 目测相关性

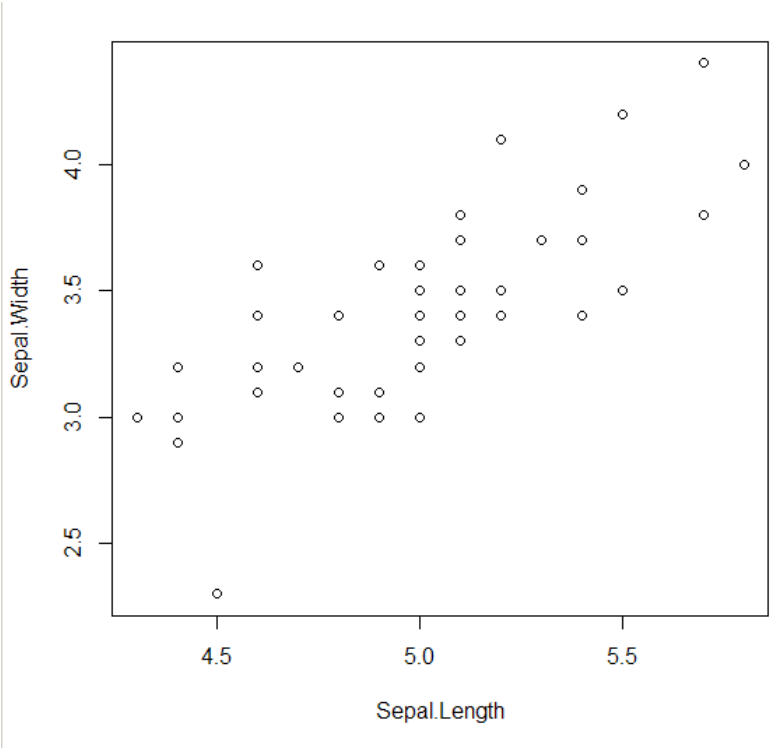
plot(iris[1,2])



相关分析的例子

- 分离种属

```
i1=iris[which(iris$Species=="setosa"),1:2]
plot(i1)
```



相关分析的例子

- 求相关系数
- 相关系数是否显著，不能只根据值的大小还需要进行假设检验

```
> cor(i1[1],i1[2])
              Sepal.Width
Sepal.Length   0.7425467
```

相关分析的例子

- 相关系数显著性的假设检验
- 假设r0为总体相关系数，r0=0则说明没有相关关系，建立假设H0:r0=0，H1:r0<>0(alpha=0.05)
- 计算相关系数r的t值和P-值

```
> cor.test(i1$Sepal.Length,i1$Sepal.Width)

        Pearson's product-moment correlation

data:  i1$Sepal.Length and i1$Sepal.Width
t = 7.6807, df = 48, p-value = 6.71e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5851391 0.8460314
sample estimates:
      cor
0.7425467
```

一元线性回归分析

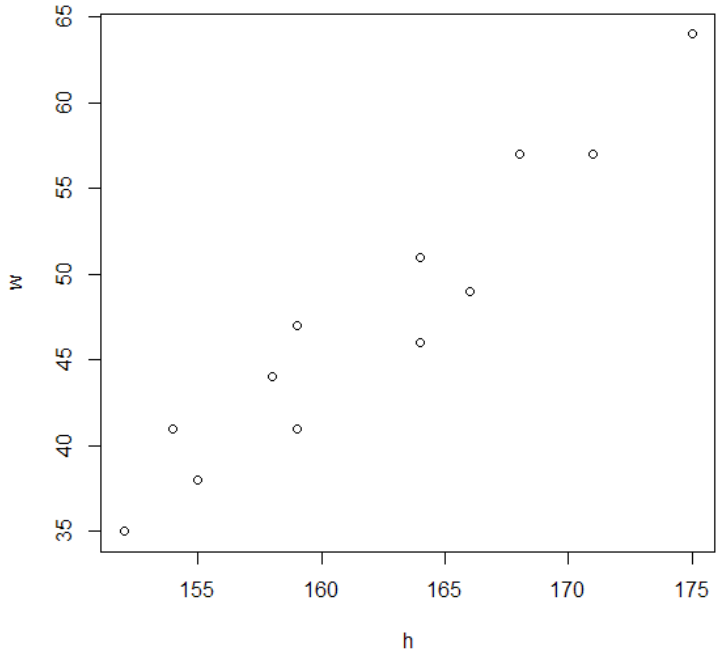
- 原理，最小二乘法
- 步骤：建立回归模型，求解回归模型中的参数，对回归模型进行检验
- 例子

数据：身高-体重

h=c(171,175,159,155,152,158,154,164,168,166,159,164)

w=c(57,64,41,38,35,44,41,51,57,49,47,46)

plot(w~h+1)



一元线性回归分析

自定义函数 lxy<-

```
    function(x,y){n=length(x);sum(x*
    y)-sum(x)*sum(y)/n}
```

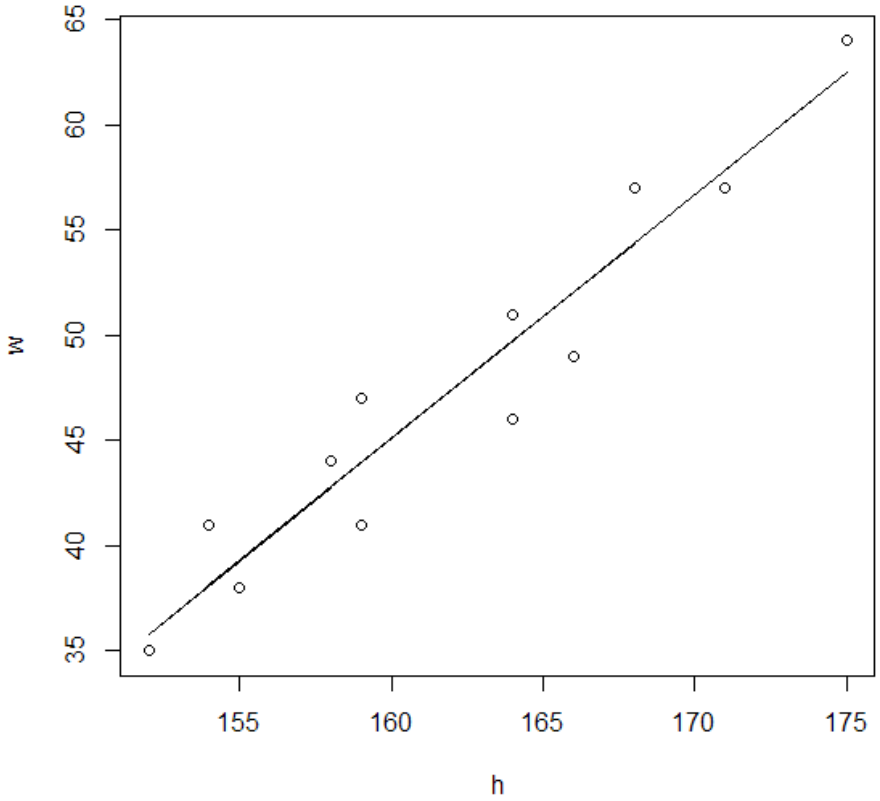
假设w=a+bh

则有

```
> b=lxy(h,w)/lxy(h,h)
> a=mean(w)-b*mean(h)
> a
[1] -140.3644
> b
[1] 1.15906
```

作回归直线

```
lines(h,a+b*h)
```



一元线性回归分析

- 回归系数的假设检验
- 建立线性模型

```
> a=lm(w~1+h)
> a

Call:
lm(formula = w ~ 1 + h)

Coefficients:
(Intercept)            h
   -140.364        1.159
```

一元线性回归分析

- 线性模型的汇总数据，t检验，summary()函数

```
> summary(a)

Call:
lm(formula = w ~ 1 + h)

Residuals:
   Min     1Q Median     3Q    Max
-3.721 -1.699  0.210  1.807  3.074

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -140.3644    17.5026   -8.02 1.15e-05 ***
h              1.1591     0.1079   10.74 8.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.546 on 10 degrees of freedom
Multiple R-squared: 0.9203,     Adjusted R-squared: 0.9123
F-statistic: 115.4 on 1 and 10 DF,  p-value: 8.21e-07
```

一元线性回归分析

- 汇总数据的解释
- Residuals：参差分析数据
- Coefficients：回归方程的系数，以及推算的系数的标准差，t值，P-值
- F-statistic：F检验值
- Signif：显著性标记，***极度显著，**高度显著，*显著，圆点不太显著，没有记号不显著

一元线性回归分析

- 方差分析，函数anova()

```
> anova(a)
Analysis of Variance Table

Response: w
          Df Sum Sq Mean Sq F value    Pr(>F)
h          1 748.17  748.17  115.41 8.21e-07 ***
Residuals 10  64.83    6.48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

一元线性回归分析

- 预测：一个身高185的人，体重大约是多少？

```
> a+b*185
[1] 74.0618
>
```

lm()线性模型函数

适应于多元线性模型的基本函数是 lm(), 其调用形式是

```
fitted.model <- lm(formula, data = data.frame)
```

其中 formula 为模型公式. data.frame 为数据框. 返回值为线性模型结果的对象存放在 fitted.model 中. 例如

```
fm2 <- lm(y ~ x1 + x2, data = production)
```

适应于 y 关于 $x1$ 和 $x2$ 的多元回归模型（隐含着截距项）。

- y~1+x或y~x均表示y=a+bx有截距形式的线性模型
- 通过原点的线性模型可以表达为：y ~ x - 1 或y ~ x + 0 或 y ~ 0 + x

参见help(formula)

与线性模型有关的函数

建立数据：身高-体重

```
x=c(171,175,159,155,152,158,154,164,168,166,159,164)
y=c(57,64,41,38,35,44,41,51,57,49,47,46)
```

建立线性模型

```
a=lm(y~x)
```

求模型系数

```
> coef(a)
(Intercept)           x
 -140.36436     1.15906
```

提取模型公式

```
> formula(a)
y ~ x
```

与线性模型有关的函数

计算残差平方和（什么是残差平方和）

```
> deviance(a)
[1] 64.82657
```

绘画模型诊断图（很强大，显示残差、拟合值和一些诊断情况）

```
> plot(a)
```

计算残差

```
> residuals(a)
        1         2         3         4         5         6         7
-0.8349544  1.5288044 -2.9262307 -1.2899895 -0.8128086  1.2328296  2.8690708
        8         9        10        11        12
 1.2784678  2.6422265 -3.0396529  3.0737693 -3.7215322
```

与线性模型有关的函数

打印模型信息

```
> print(a)


Call:
lm(formula = y ~ x)


Coefficients:
(Intercept)            x
   -140.364        1.159
```

与线性模型有关的函数

计算方差分析表

,

```
> anova(a)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 748.17  748.17  115.41 8.21e-07 ***
Residuals 10  64.83    6.48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

与线性模型有关的函数

中山大學 SUN YAT-SEN UNIVERSITY

提取模型汇总资料

```
> summary(a)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-3.721 -1.699  0.210  1.807  3.074

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -140.3644    17.5026   -8.02 1.15e-05 ***
x              1.1591     0.1079   10.74 8.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.546 on 10 degrees of freedom
Multiple R-squared: 0.9203,     Adjusted R-squared: 0.9123
F-statistic: 115.4 on 1 and 10 DF,  p-value: 8.21e-07
```

与线性模型有关的函数

作出预测

```
> z=data.frame(x=185)
> predict(a,z)
       1
74.0618
> predict(a,z,interval="prediction", level=0.95)
       fit     lwr      upr
1 74.0618 65.9862 82.13739
```

课后阅读：薛毅书，p308，计算实例

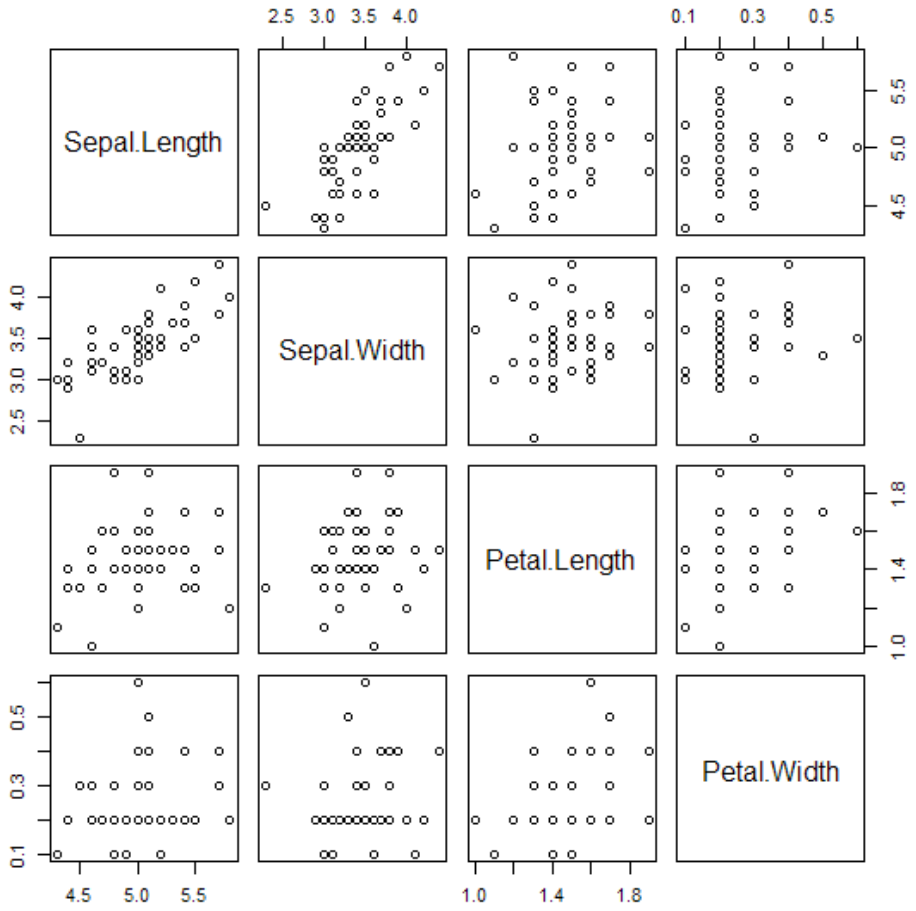
多元线性相关分析

- 研究多个变量之间的关系
- 例子：iris数据集，研究花瓣和花萼的长度、宽度之间的联系

准备数据：

```
x=iris[which(iris$Species
==="setosa"),1:4]
```

画出散点图集：plot(x)



2015.3.12

多元线性相关分析

- 计算相关系数矩阵，cor()函数
- 暂时没有发现可以在多元情况下进行相关性检验的函数，只能对变量两两进行检验

```
> cor(x)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
> 
```

多元线性回归

- Swiss数据集： Swiss Fertility and Socioeconomic Indicators (1888) Data

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2
Grandson	71.7	34.0	17	8	3.30	20.0
Lausanne	55.7	19.4	26	28	12.11	20.2
La Vallee	54.3	15.2	31	20	2.15	10.8
Lavaux	65.1	73.0	19	9	2.84	20.0
Morges	65.5	59.8	22	10	5.23	18.0

多元线性回归

建立多元线性模型

```
> s=lm(Fertility ~ ., data = swiss)
> print(s)

Call:
lm(formula = Fertility ~ ., data = swiss)

Coefficients:
     (Intercept)       Agriculture       Examination         Education
         66.9152           -0.1721           -0.2580           -0.8709
        Catholic  Infant.Mortality
          0.1041            1.0770
```

多元线性回归

模型汇总信息

```
> summary(s)

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       66.91518   10.70604   6.250 1.91e-07 ***
Agriculture       -0.17211    0.07030  -2.448  0.01873 *
Examination       -0.25801    0.25388  -1.016  0.31546
Education         -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic           0.10412    0.03526   2.953  0.00519 **
Infant.Mortality   1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared: 0.7067,     Adjusted R-squared: 0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

多元线性回归

- 多元线性回归的核心问题：**应该选择哪些变量？**
- 一个非典型例子（薛毅书p325)
- RSS（残差平方和）与R^2（相关系数平方）选择法：遍历所有可能的组合，选出使RSS最小，R^2最大的模型
- AIC（Akaike information criterion）准则与BIC （Bayesian information criterion ）准则

$$AIC=n\ \ln\left(RSS_p/n\right)+2p$$

n为变量总个数，p为选出的变量个数，**AIC越小越好**

多元线性回归

- 逐步回归
- 向前引入法：从一元回归开始，逐步增加变量，使指标值达到最优为止
- 向后剔除法：从全变量回归方程开始，逐步删去某个变量，使指标值达到最优为止
- 逐步筛选法：综合上述两种方法

多元线性回归

- step( )函数

```
> s1=step(s,direction="forward")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
```

```
> s1=step(s,direction="backward")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                         2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                         2158.1 189.86
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43
```

```
> s1=step(s,direction="both")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                         2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                         2158.1 189.86
+ Examination       1     53.03 2105.0 190.69
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43
>
```

多元线性回归

- 是否还有优化余地？
- 使用drop1作删除试探，使用add1函数作增加试探

```
> drop1(s1)
Single term deletions

Model:
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
                  Df Sum of Sq    RSS    AIC
<none>                          2158.1 189.86
Agriculture        1    264.18 2422.2 193.29
Education          1   2249.97 4408.0 221.43
Catholic           1    956.57 3114.6 205.10
Infant.Mortality   1    409.81 2567.9 196.03
```

多元线性回归

- 薛毅书，p330例子

Thanks

FAQ时间