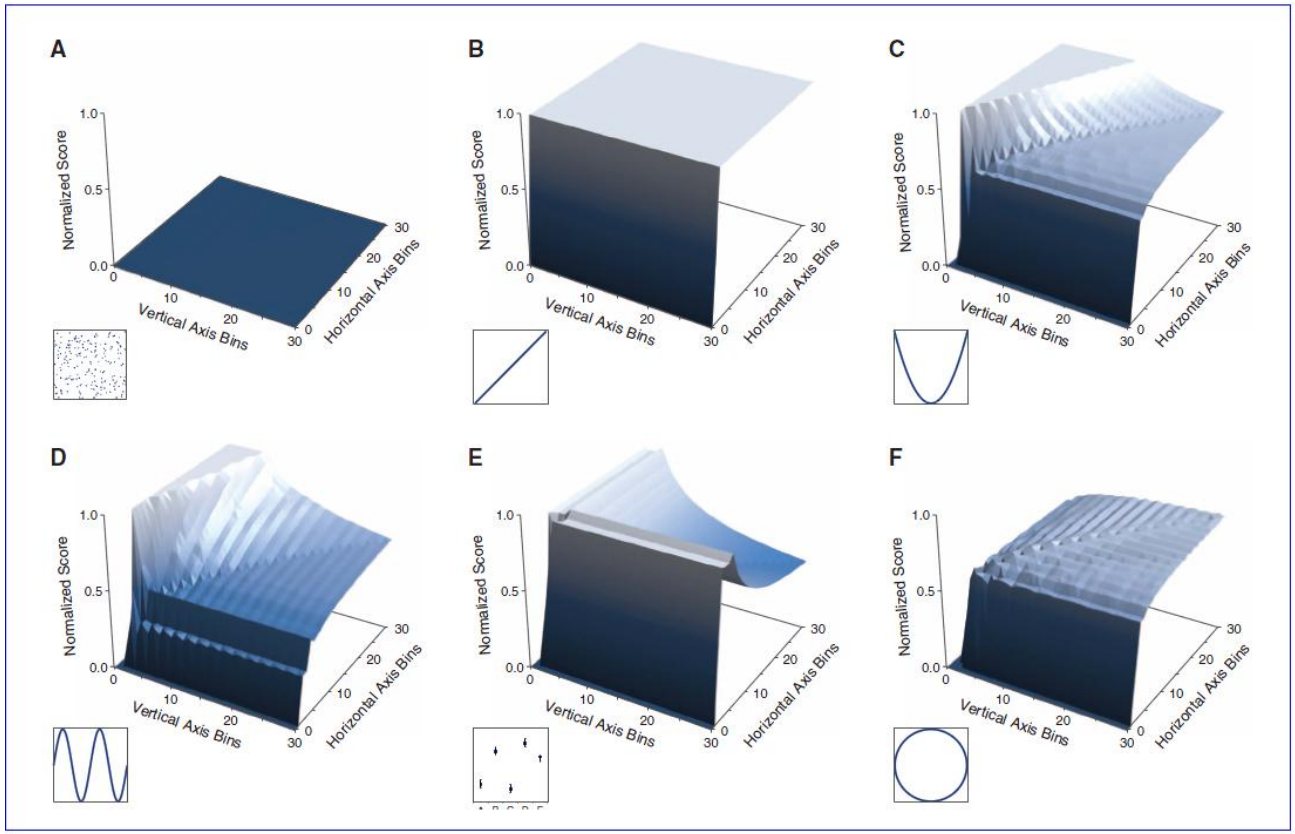
中山大學
SUN YAT-SEN UNIVERSITY



2012级《多元统计分析与数据挖掘》第3周

2015.3.19

中山大学数学与计算科学学院 黄志洪

虚拟变量

- 虚拟变量的定义
- 虚拟变量的作用
- 虚拟变量的设置

Boston数据集

中山大學 SUN YAT-SEN UNIVERSITY

- Boston数据集

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2
16	0.62739	0.0	8.14	0	0.5380	5.834	56.5	4.4986	4	307	21.0	395.62	8.47	19.9
17	1.05393	0.0	8.14	0	0.5380	5.935	29.3	4.4986	4	307	21.0	386.85	6.58	23.1

虚拟变量的使用

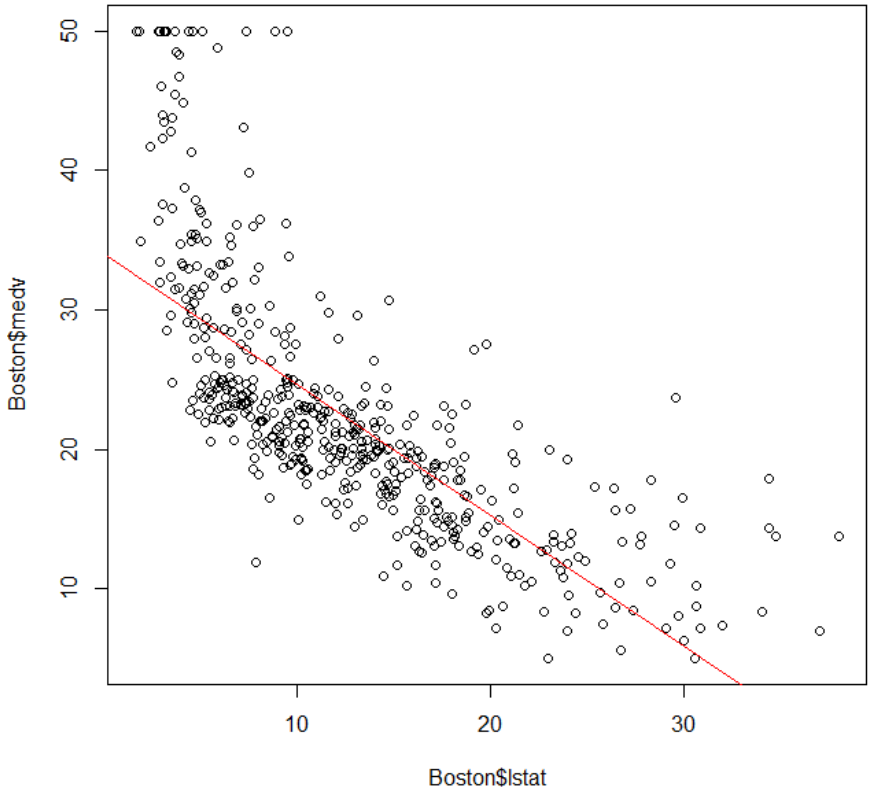
- Boston数据中，chas是一个虚拟变量，Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- 构建medv关于lstat与chas的回归模型
- Y=β0+β1*chas+β2*lstat= $\begin{cases} \beta0+\beta1+\beta2*lstat，chas=1 \\ \beta0+\beta2*lstat，chas=0 \end{cases}$
- 所以，虚拟变量影响的只是

截距项

```
> lm.fit=lm(medv~lstat+chas,data=Boston)
> summary(lm.fit)

Call:
lm(formula = medv ~ lstat + chas, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-14.782  -3.798  -1.286   1.769  24.870

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.09412    0.56067  60.809  < 2e-16 ***
lstat       -0.94061    0.03804 -24.729  < 2e-16 ***
chas         4.91998    1.06939   4.601 5.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.095 on 503 degrees of freedom
Multiple R-squared:  0.5626,    Adjusted R-squared:  0.5608
F-statistic: 323.4 on 2 and 503 DF,  p-value: < 2.2e-16
```

虚拟变量的使用

```
> plot(Boston$lstat,Boston$medv)
> abline(lm.fit,col="red")
```



回归诊断

- 样本是否符合正态分布假设？
- 是否存在离群值导致模型产生较大误差？
- 线性模型是否合理？
- 误差是否满足独立性、等方差、正态分布等假设条件？
- 是否存在多重共线性？

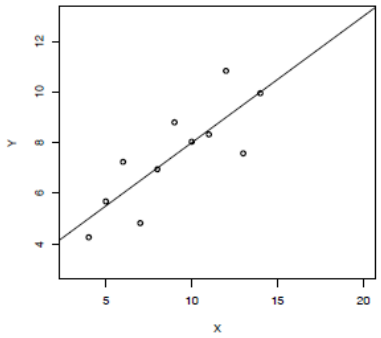
正态分布检验

- 正态性检验：函数shapiro.test( )
- P>0.05，正态性分布

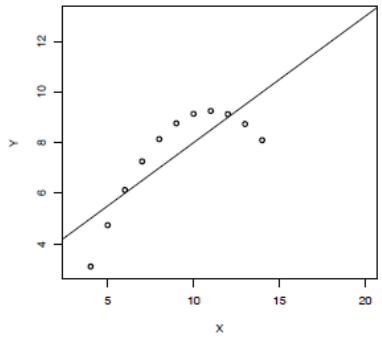
```
> shapiro.test(x$x1)

        Shapiro-Wilk normality test

data:  x$x1
W = 0.9937, p-value = 0.9259

> shapiro.test(x$x3)

        Shapiro-Wilk normality test

data:  x$x3
W = 0.9444, p-value = 0.0003618
```

散点图目测检验

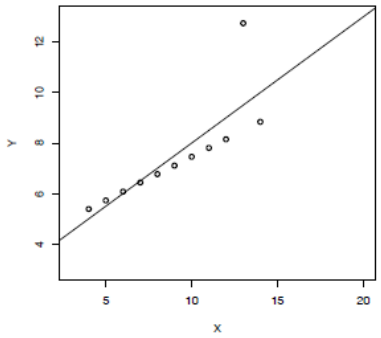
- 薛毅书纸介质p284，例6.11



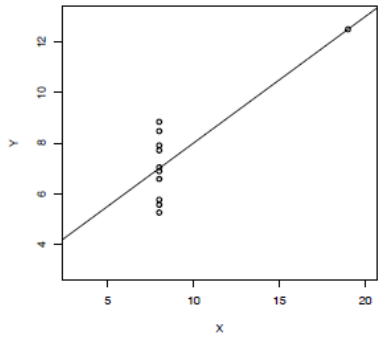
(a) 数据 1



(b) 数据 2



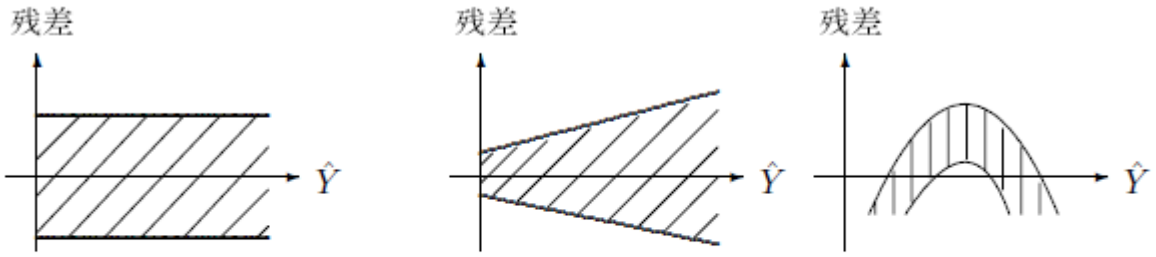
(c) 数据 3



(d) 数据 4

残差

- 残差计算函数residuals( )
- 对残差作正态性检验
- 残差图

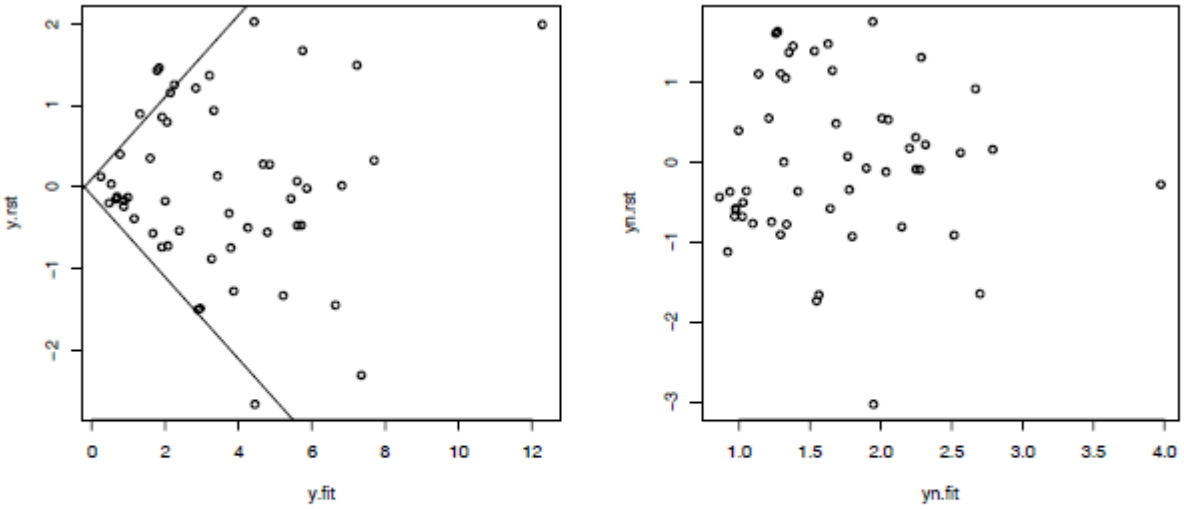


(a) 正常情况　(b) 异方差情况　(c) 非线性情况

图 6.7: 回归值 $\hat{Y}$ 与残差的散点图

例子

- 薛毅书p346例6.14



(a) 异方差情况　　(b) 变换后的情况

图 6.9: 例 6.6 的标准化残差图

多重共线性

- 什么是多重共线性
- 多重共线性对回归模型的影响
- 利用计算特征根发现多重共线性
- Kappa()函数

例 6.19 R. Norell 实验

为研究高压电线对牲畜的影响，*R. Norell* 研究小的电流对农场动物的影响. 他在实验中，选择了 *7* 头，*6* 种电击强度，*0,1,2,3,4,5* 毫安. 每头牛被电击 *30* 下，每种强度 *5* 下，按随机的次序进行. 然后重复整个实验，每头牛总共被电击 *60* 下. 对每次电击，响应变量 — 嘴巴运动，或者出现，或者未出现. 表 *6.13* 中的数据给出每种电击强度 *70* 次试验中响应的总次数. 试分析电击对牛的影响.

表 6.13: 7 头牛对 6 种不同强度的非常小的电击的响应

电流 (毫安)	试验次数	响应次数	响应的比例
0	70	0	0.000
1	70	9	0.129
2	70	21	0.300
3	70	47	0.671
4	70	60	0.857
5	70	63	0.900

广义线性模型

- 目标：求出电流强度与牛是否张嘴之间的关系
- 困难：牛是否张嘴，是0-1变量，不是变量，无法建立线性回归模型
- 矛盾转化：牛张嘴的概率是连续变量

广义线性模型

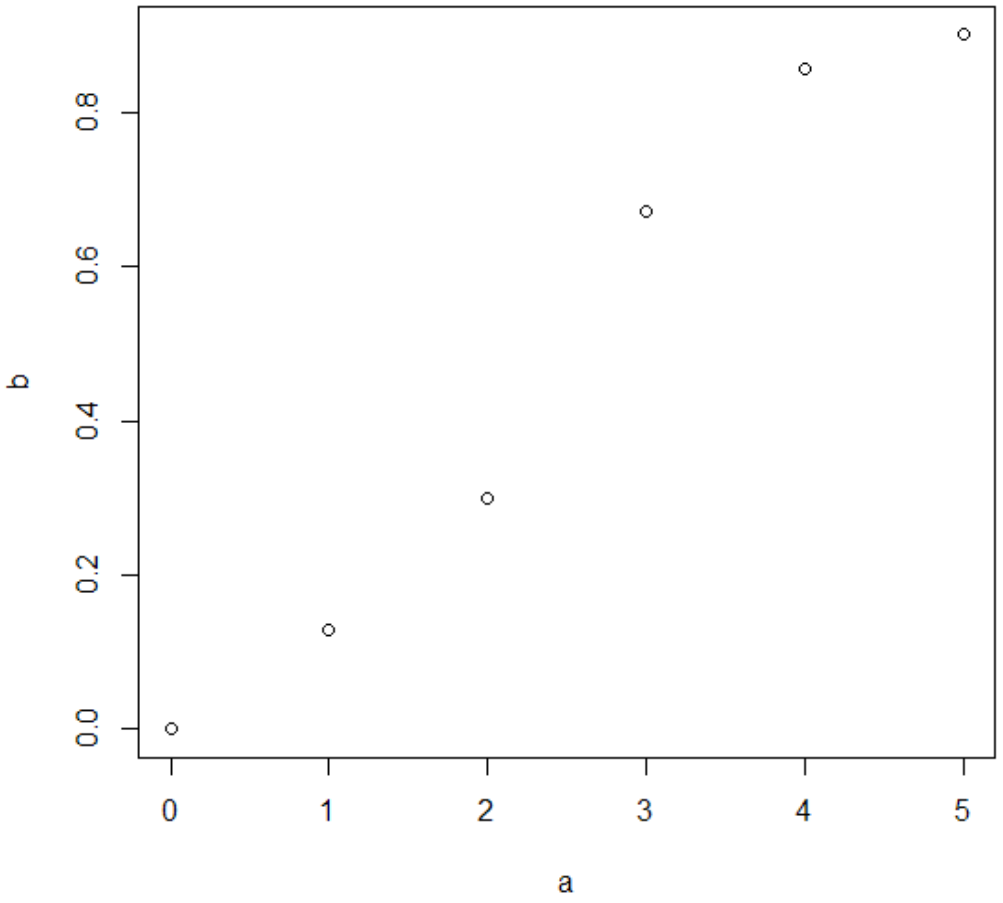
a=c(0:5)

b=c(0,0.129,0.3,0.671,0.857,0.9)

plot(a,b)

符合logistic回归模型的曲线特征

$$P = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 + \cdots + \beta_p X_p)}$$



广义线性模型

- Logit变换

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

- 常见连接函数与逆连接函数

表 6.11: 常见的连接函数和误差函数

	连接函数	逆连接函数 (回归模型)	典型误差函数
恒等	$x^T\beta = E(y)$	$E(y) = x^T\beta$	正态分布
对数	$x^T\beta = \ln E(y)$	$E(y) = \exp(x^T\beta)$	Poisson 分布
Logit	$x^T\beta = \text{Logit}E(y)$	$E(y) = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}$	二项分布
逆	$x^T\beta = \frac{1}{E(y)}$	$E(y) = \frac{1}{x^T\beta}$	Gamma 分布

广义线性模型

- 广义线性模型建模函数：glm( ) 。薛毅书p364

```
fitted.model <- glm(formula, family=family.generator,
                    data=data.frame)
```

```
fm <- glm(formula, family = binomial(link = logit),
          data=data.frame)
```

广义线性模型

```
norell<-data.frame(x=0:5,
    n=rep(70,6),
    success=c(0,9,21,47,60,63))
norell$Ymat<-
    cbind(norell$success,
    norell$n-norell$success)
glm.sol<-glm(Ymat~x,
    family=binomial,
    data=norell)
summary(glm.sol)
```

```
Call:
glm(formula = Ymat ~ x, family = binomial, data = norell)

Deviance Residuals:
      1       2       3       4       5       6
-2.2507  0.3892 -0.1466  1.1080  0.3234 -1.6679

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.3010     0.3238  -10.20   <2e-16 ***
x            1.2459     0.1119   11.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 250.4866  on 5  degrees of freedom
Residual deviance:   9.3526  on 4  degrees of freedom
AIC: 34.093

Number of Fisher Scoring iterations: 4
```

$$P=\frac{\exp(-3.3010+1.2459X)}{1+\exp(-3.3010+1.2459X)}$$

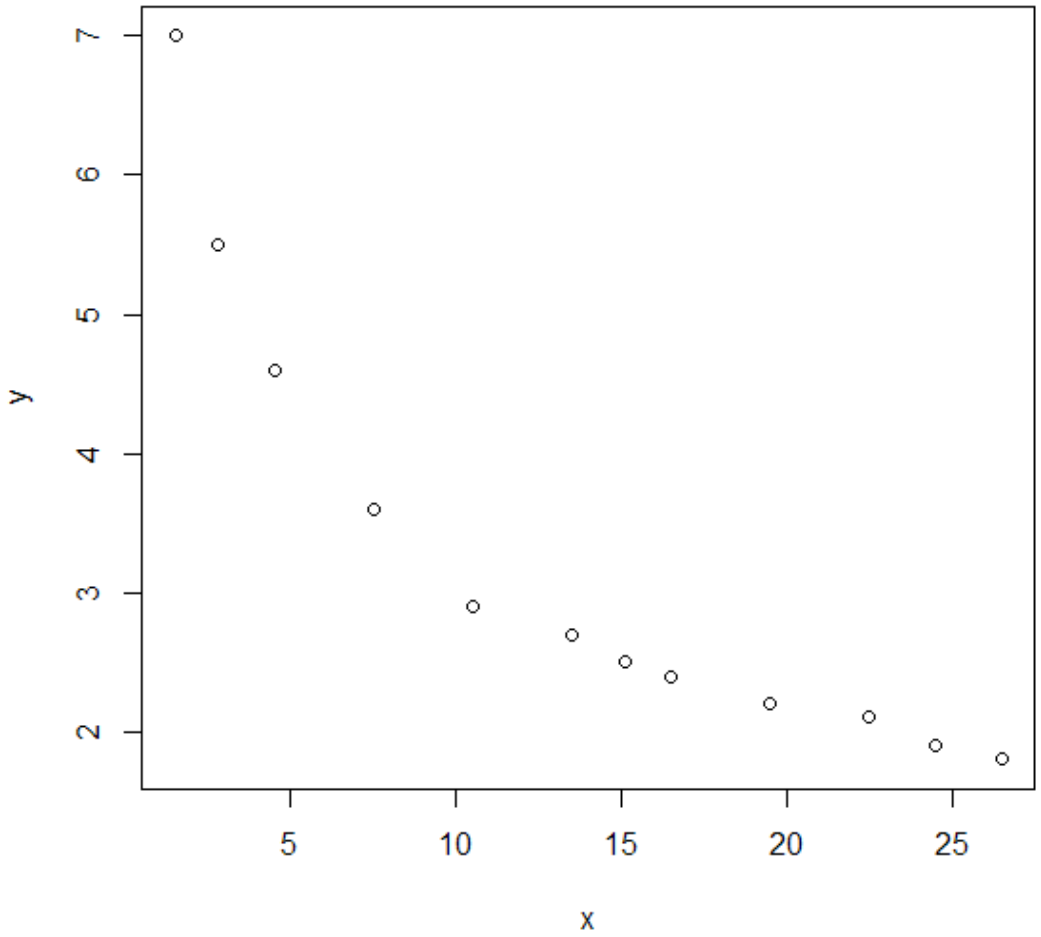
广义线性模型

- 多元的情形，逐步回归，step( )函数
- 例子，薛毅书P369
- 其它广义线性模型，薛毅书P374

非线性模型

- 例子：销售额x与流通费率y

```
x=c(1.5,2.8,4.5,7.5,10.5,13.5
    ,15.1,16.5,19.5,22.5,24.5
    ,26.5)
y=c(7.0,5.5,4.6,3.6,2.9,2.7,2.
    5,2.4,2.2,2.1,1.9,1.8)
plot(x,y)
```



非线性模型

- 直线回归（R^2值不理想）

lm.1=lm(y~x)

>summary(lm.1)

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9179 -0.5537 -0.1628  0.3953  1.6519

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.60316    0.43474  12.889 1.49e-07 ***
x           -0.17003    0.02719  -6.254 9.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7701 on 10 degrees of freedom
Multiple R-squared: 0.7964,     Adjusted R-squared: 0.776
F-statistic: 39.11 on 1 and 10 DF,  p-value: 9.456e-05
```

非线性模型

- 多项式回归，假设用二次多项式方程

  $y=a+bx+cx^2$

x1=x

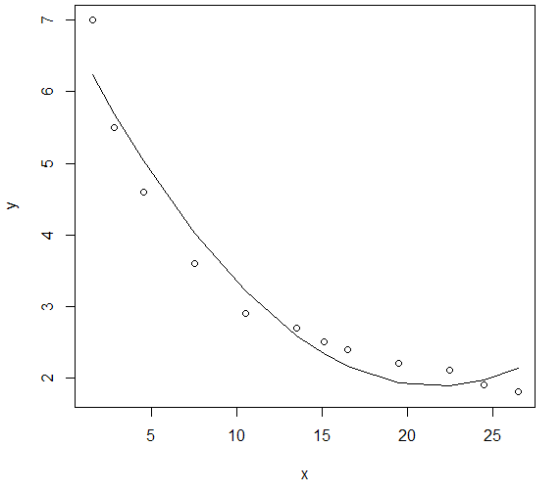
x2=x^2

lm.2=lm(y~x1+x2)

summary(lm.2)

plot(x,y)

lines(x,fitted(lm.2))




```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43718 -0.31604  0.02362  0.22211  0.75956

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.914687   0.331987  20.828 6.35e-09 ***
x1          -0.465631   0.056969  -8.173 1.86e-05 ***
x2           0.010757   0.002009   5.353  0.00046 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3969 on 9 degrees of freedom
Multiple R-squared: 0.9513,     Adjusted R-squared: 0.9405
F-statistic: 87.97 on 2 and 9 DF,  p-value: 1.237e-06
```

非线性模型

- 对数法，y=a+b logx

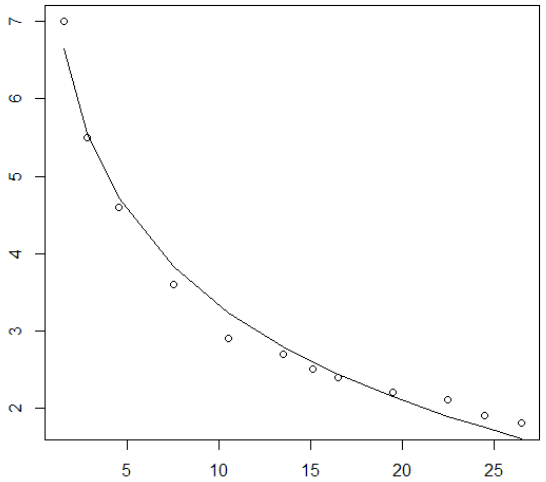
lm.log=lm(y~log(x))

Summar

plot(x,y)

lines(x,fitted(lm.log))y(lm

.log)

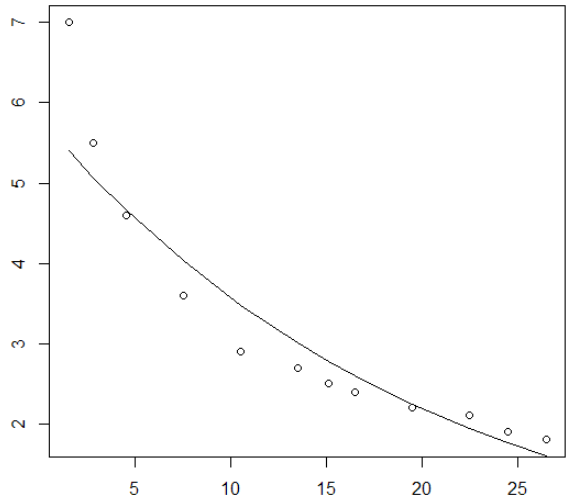



```
Call:
lm(formula = y ~ log(x))

Residuals:
     Min       1Q   Median       3Q      Max
-0.33291 -0.10133 -0.04693  0.16512  0.34844

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.3639     0.1688   43.64 9.60e-13 ***
log(x)       -1.7568     0.0677  -25.95 1.66e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2064 on 10 degrees of freedom
Multiple R-squared: 0.9854,     Adjusted R-squared: 0.9839
F-statistic: 673.5 on 1 and 10 DF,  p-value: 1.66e-10
```

非线性模型

- 指数法，$y=a\ e^{bx}$

```
lm.exp=lm(log(y)~x)
summary(lm.exp)
plot(x,y)
lines(x,exp(fitted(lm.
    exp)))
```



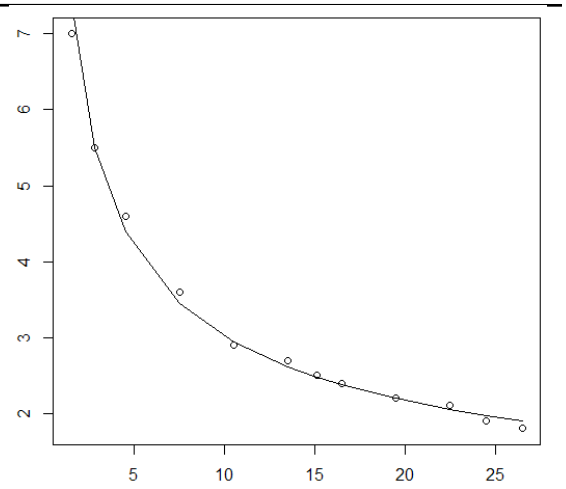

```
Call:
lm(formula = log(y) ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18246 -0.10664 -0.01670  0.08079  0.25946

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.759664   0.075101   23.43 4.54e-10 ***
x           -0.048809   0.004697  -10.39 1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.133 on 10 degrees of freedom
Multiple R-squared: 0.9153,     Adjusted R-squared: 0.9068
F-statistic:   108 on 1 and 10 DF,  p-value: 1.116e-06
```

非线性模型

- 幂函数法，$y=a\,x^b$

```
lm.pow=lm(log(y)~log(x))
summary(lm.pow)
plot(x,y)
lines(x,exp(fitted(lm.pow))
    )
```

对比以上各种拟合回归过程得出结论是幂函数法为最佳




```
Call:
lm(formula = log(y) ~ log(x))

Residuals:
      Min        1Q    Median        3Q       Max
-0.054727 -0.020805  0.004548  0.024617  0.045896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.19073    0.02951   74.23 4.81e-15 ***
log(x)      -0.47243    0.01184  -39.90 2.34e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0361 on 10 degrees of freedom
Multiple R-squared: 0.9938,     Adjusted R-squared: 0.9931
F-statistic:  1592 on 1 and 10 DF,  p-value: 2.337e-12
```

非线性模型

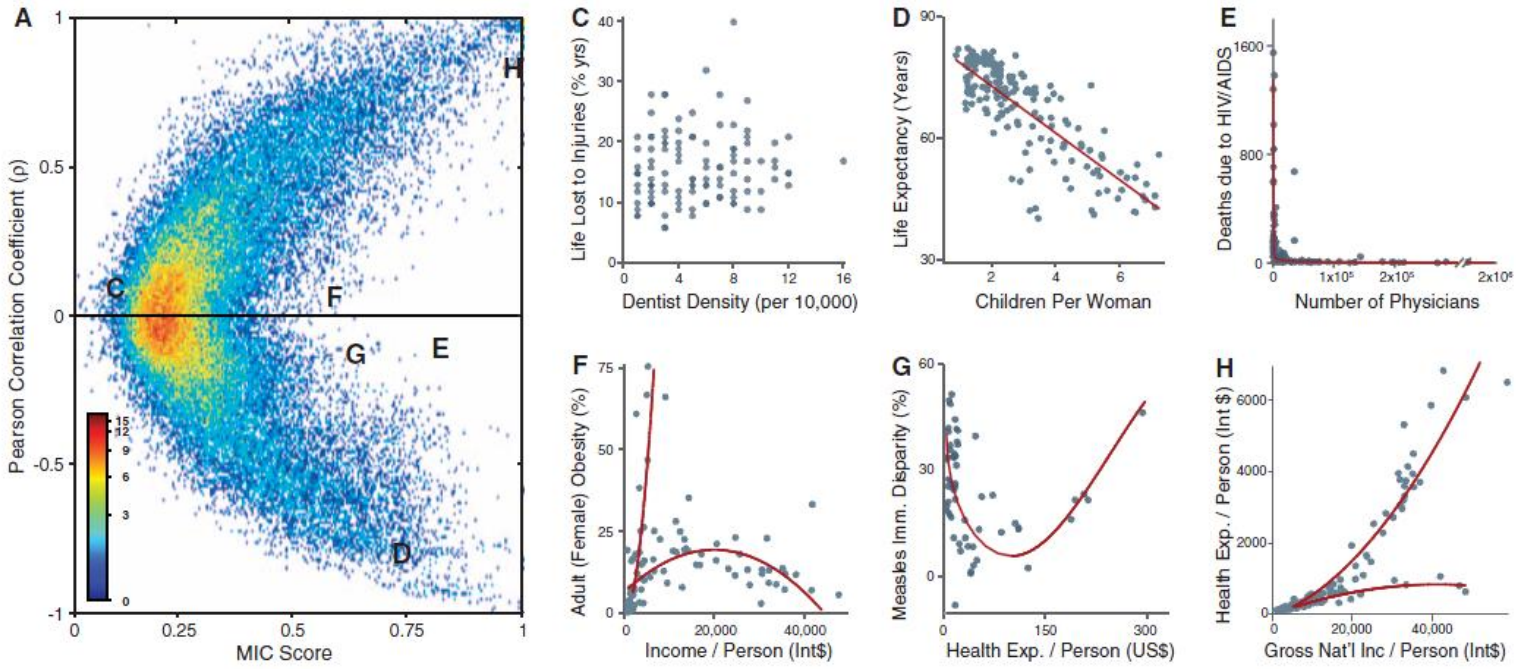
- 正交多项式回归
- 例子，薛毅书P378

非线性最小二乘问题

- nls( )函数
- 例子，薛毅书P384

传统回归模型的困难

- 为什么一定是线性？或某种非线性模型？
- 过分依赖于分析者的经验
- 对于非连续的离散数据难以处理



2015.3.19

网格方法

- 《Science》上的文章《Detecting Novel Associations in Large Data Sets》
- 方法概要：用网格判断数据的集中程度，集中程度意味着是否有关联关系
- 方法具有一般性，即无论数据是怎样分布的，不限于特定的关联函数类型，此判断方法都是有效
- 方法具有等效性，计算的熵值和噪音的程度有关，跟关联的类型无关
- MIC：the Maximal Information Coefficient
- MINE：Maximal Information-based Nonparametric Exploration

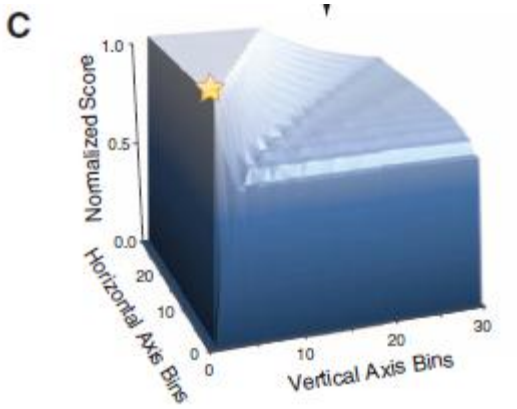
MIC值计算

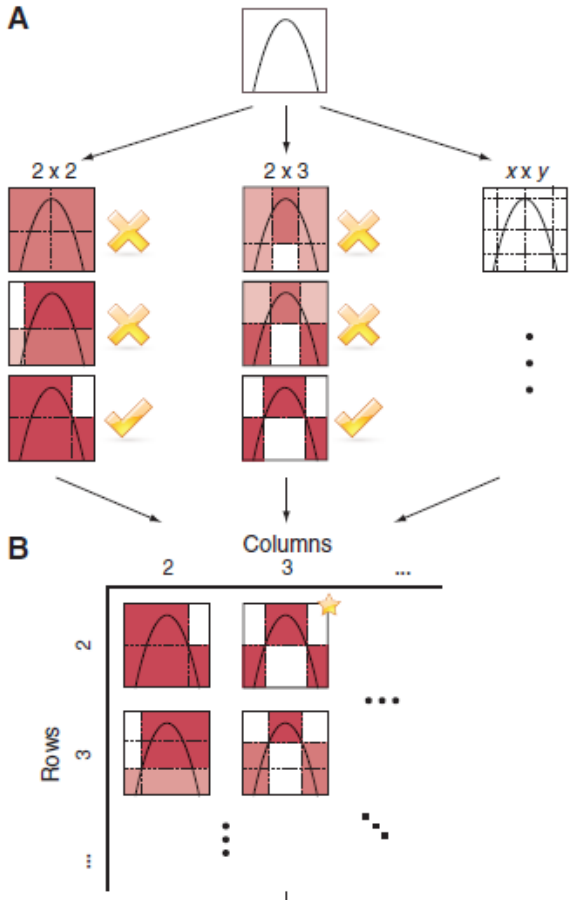
- 坐标平面被划分为(x,y)网格G（未必等宽），其中$xy<n^{0.6}$
- 在G上可以诱导出“自然概率密度函数”p(x,y)，任何一个方格（box）内的概率密度函数值为这个方格所 包含的样本点数量占全体样本点的比例
- 计算网格划分G下的 **mutual information值** I_G

$$I(X;Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) dx\, dy$$

MIC值计算

- 构造**特征矩阵**{m_{xy}}，矩阵的元素m_{xy}=max{I_G}/log min{x,y}。max取遍所有可能的(x,y)网格G
- MIC=max {m_{xy}}。Max取遍所有可能的(x,y)对





MIC值计算

- Mxy的计算是个难点，数据科学家构造了一个近似的逼近算法以提高效率

http://www.sciencemag.org/content/suppl/2011/12/14/334.6062.1518.DC1

在作者的网站上，可以下载MINE计算MIC的程序（Java和R）以及测试用数据集

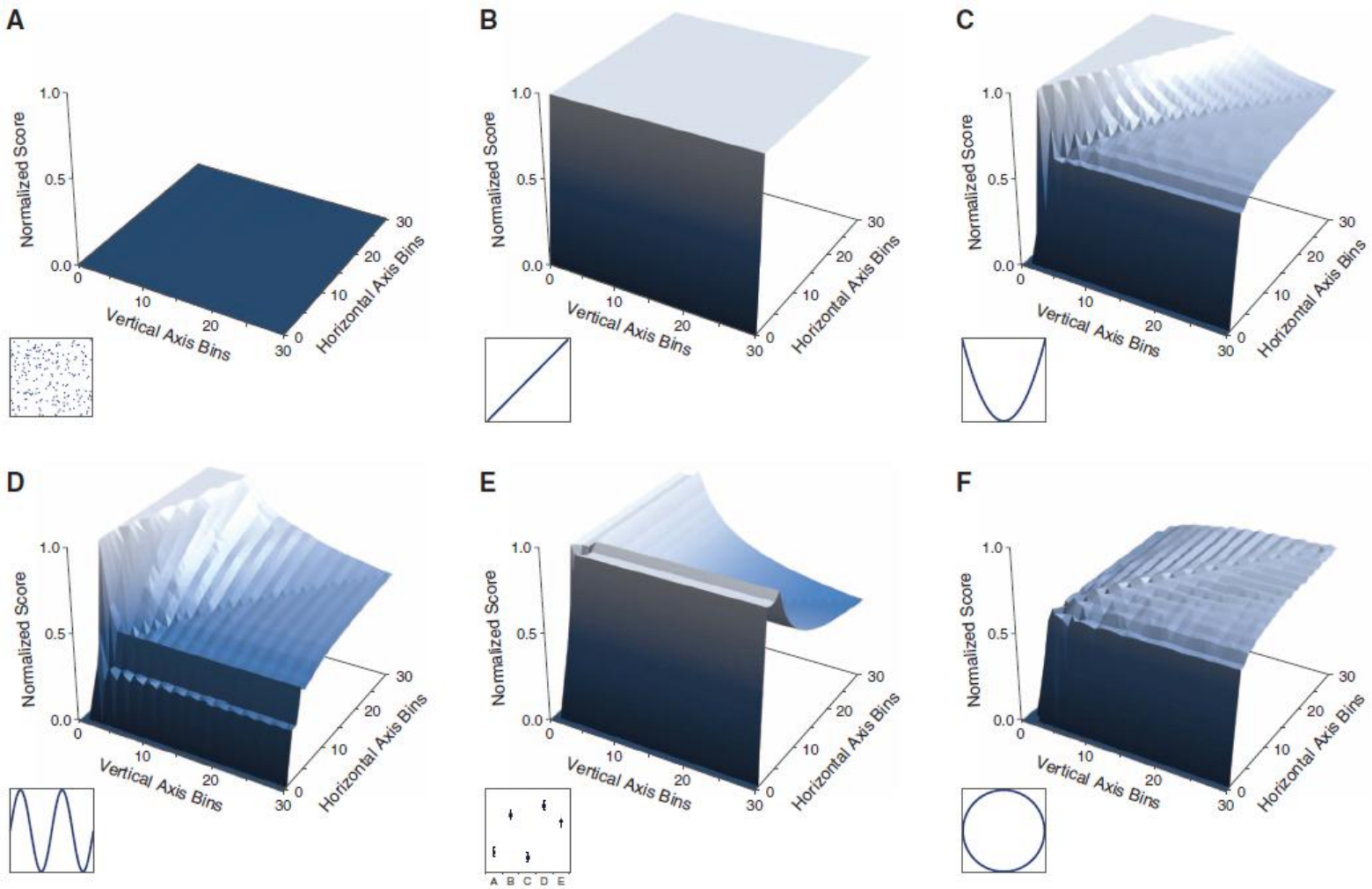
http://www.exploredata.net/Downloads

实验：WHO数据集，垒球数据集...

MIC的性质

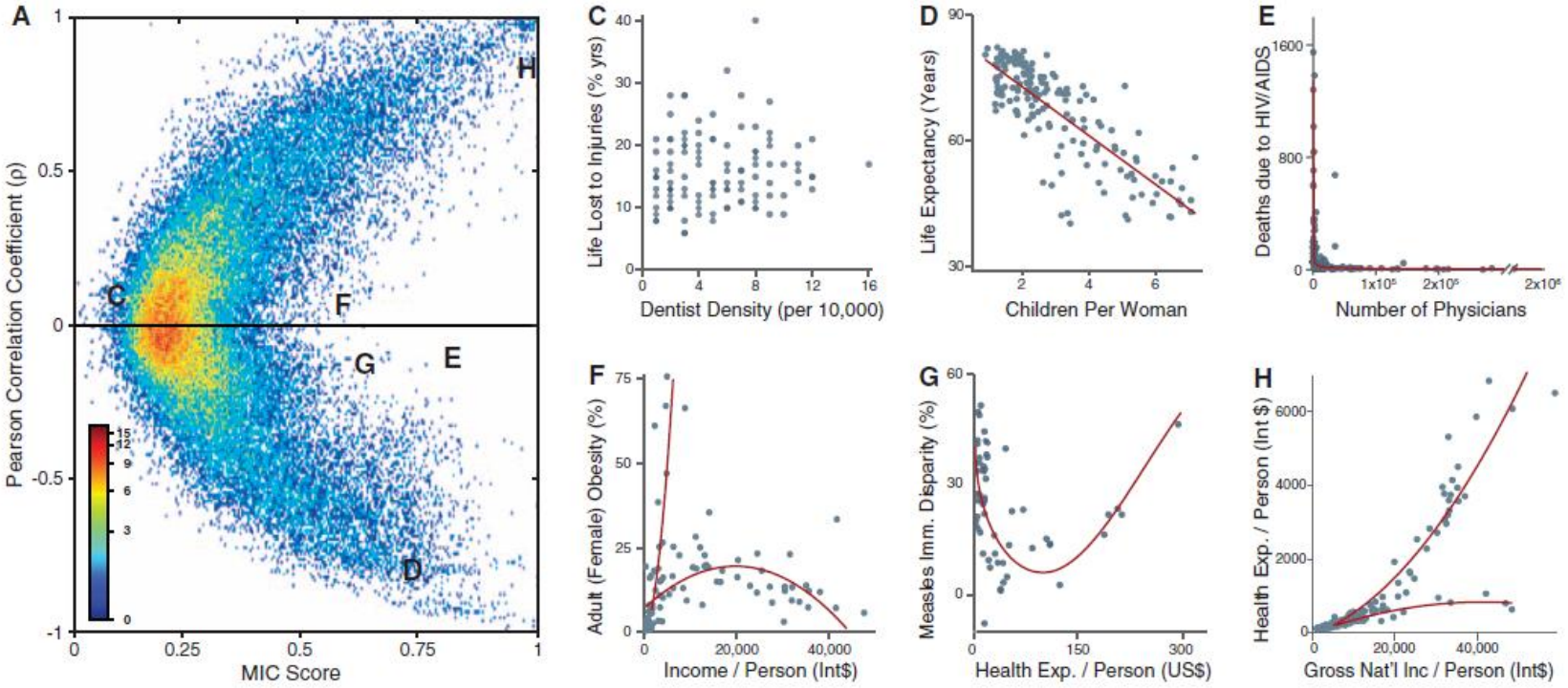
- 如果变量对x,y存在函数关系，则当样本数增加时，MIC必然趋向于1
- 如果变量对x,y可以由参数方程c(t)=[x(t),y(t)]所表达的曲线描画，则当样本数增加时，MIC必然趋于1
- 如果变量对x,y在统计意义下互相独立，则当样本数增加时，MIC趋于0

MIC观察



2015.3.19

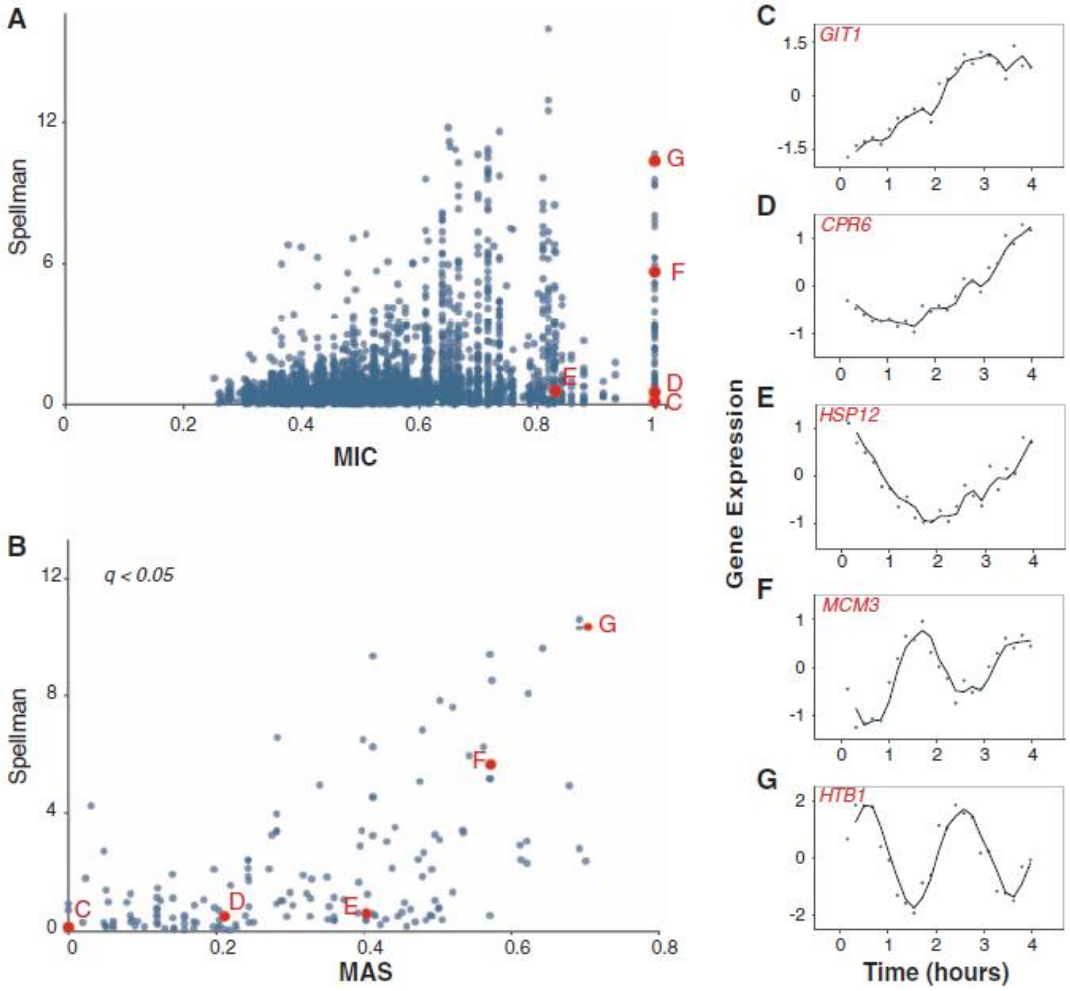
MIC与线性回归模型对比



2015.3.19

对基因数据集spellman的探索

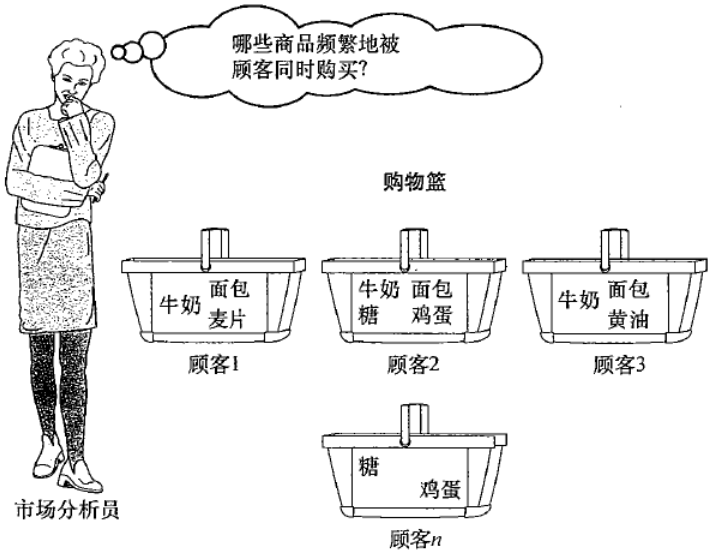
- 数据集包含6223组基因数据
- MINE对关联关系的辨认力明显强于以往的方法，例如双方都发现了HTB1，但MINE方法挖出了过去未被发现的HSP12



2015.3.19

数据挖掘：关联规则挖掘

- 例子：购物篮分析



购物篮分析的应用

- 超市里的货架摆设设计
- 电子商务网站的套餐推荐

英国史5

当当价 ¥60.70 （6.9折） 钻石VIP专享折上9.5折

定　价 ¥88.00

评　论 ★★★★★ 97.4%推荐 156条

配送至 广东广州市海珠区 有货 运费说明 本商品提供礼品包装服务

今天(3月16日)可送达，请在9小时24分钟内下单并选择 "普通快递送货上门"

作　者 [英]大卫·休谟 著，刘仲敬 译

出版社 吉林出版集团有限责任公司

出版时间 2013-7-1

ISBN 9787553405445

所属分类 图书 > 历史 > 世界史 > 欧洲史

我要买 1 件

加入购物车　一键购买　收藏商品

2015.3.19

购物篮分析的应用

- 推荐系统：网站或节目的阅读/收听推荐

2015.3.19

名词

- 挖掘数据集：购物篮数据
- 频繁模式：频繁地出现在数据集中的模式，例如项集，子结构，子序列等
- 挖掘目标：频繁模式，频繁项集，关联规则等
- 关联规则：牛奶=>鸡蛋【支持度=2%，置信度=60%】
- 支持度：分析中的全部事务的2%同时购买了牛奶和鸡蛋
- 置信度：购买了牛奶的筒子有60%也购买了鸡蛋
- 最小支持度阈值和最小置信度阈值：由挖掘者或领域专家设定

名词

- 项集：项（商品）的集合
- k-项集：k个项组成的项集
- 频繁项集：满足最小支持度的项集，频繁k-项集一般记为L_k
- 强关联规则：满足最小支持度阈值和最小置信度阈值的规则

关联规则挖掘： Apriori算法

- 两步过程：找出所有频繁项集；由频繁项集产生强关联规则
- 算法：Apriori
- 例子

表 6.1　AllElectronics 某分店的事务数据

TID	商品 *ID* 的列表	*TID*	商品 *ID* 的列表
T100	I1，I2，I5	T600	I2，I3
T200	I2，I4	T700	I1，I3
T300	I2，I3	T800	I1，I2，I3，I5
T400	I1，I2，I4	T900	I1，I2，I3
T500	I1，I3		

Apriori算法的工作过程

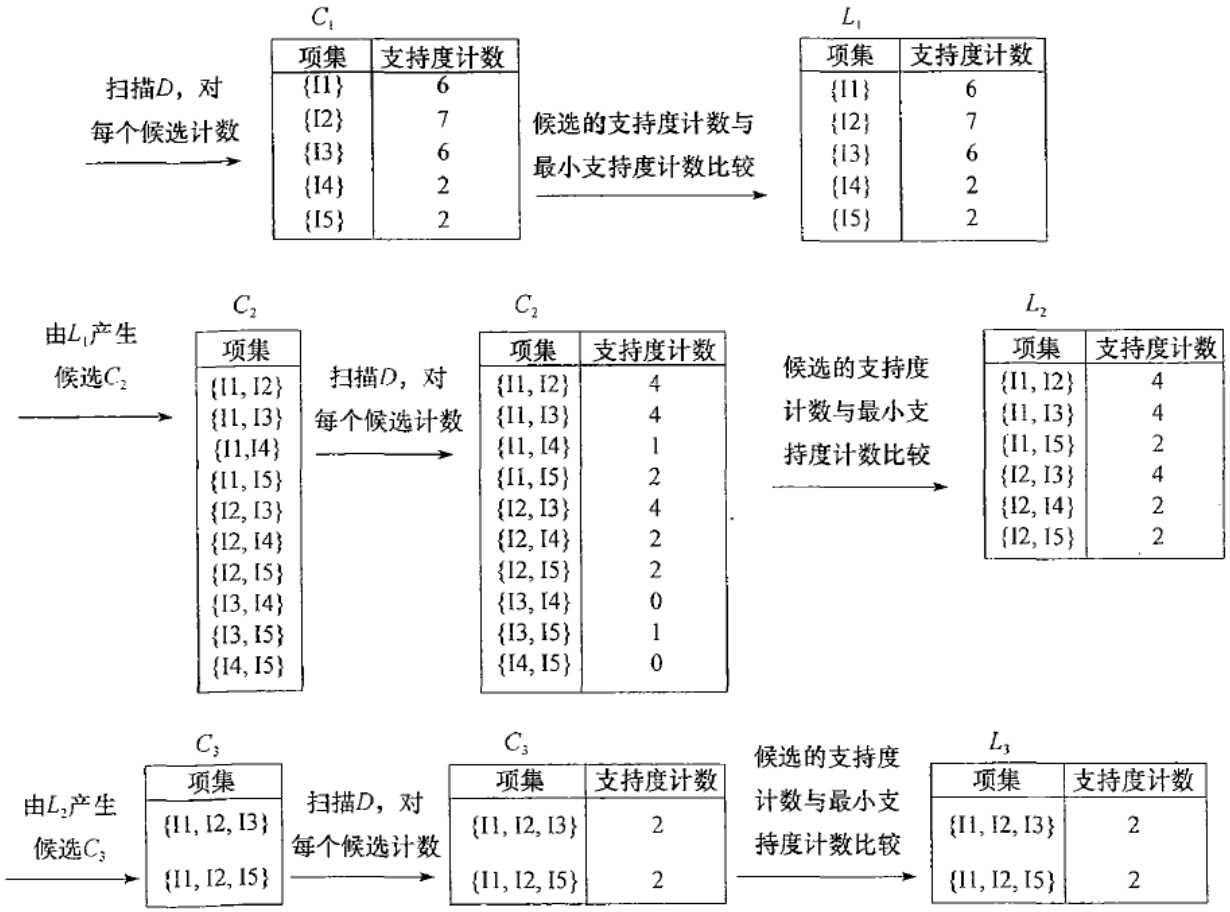


图 6.2　候选项集和频繁项集的产生，最小支持计数为 2

2015.3.19

步骤说明

- 扫描D，对每个候选项计数，生成候选1-项集C1
- 定义最小支持度阈值为2，从C1生成频繁1-项集L1
- 通过L1xL1生成候选2-项集C2
- 扫描D，对C2里每个项计数，生成频繁2-项集L2
- 计算L3xL3，利用apriori性质：频繁项集的子集必然是频繁的，我们可以删去一部分项，从而得到C3，由C3再经过支持度计数生成L3
- 可见Apriori算法可以分成 **连接，剪枝** 两个步骤不断循环重复

步骤说明

（a）连接：$C_3=L_2 \bowtie L_2=\{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$

$\bowtie\{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$

$=\{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\}, \{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}$

（b）使用先验性质剪枝：频繁项集的所有非空子集必须是频繁的。存在候选项集，其子集不是频繁的吗?

- {I1，I2，I3}的2项子集是{I1，I2}、{I1，I3}和{I2，I3}。{I1，I2，I3}的所有2项子集都是L_2的元素。因此，{I1，I2，I3}保留在C_3中。
- {I1，I2，I5}的2项子集是{I1，I2}、{I1，I5}和{I2，I5}。{I1，I2，I5}的所有2项子集都是L_2的元素。因此，{I1，I2，I5}保留在C_3中。
- {I1，I3，I5}的2项子集是{I1，I3}、{I1，I5}和{I3，I5}。{I3，I5}不是L_2的元素，因而不是频繁的。因此，从C_3中删除{I1，I3，I5}。
- {I2，I3，I4}的2项子集是{I2，I3}、{I2，I4}和{I3，I4}。{I3，I4}不是L_2的元素，因而不是频繁的。因此，从C_3中删除{I2，I3，I4}。
- {I2，I3，I5}的2项子集是{I2，I3}、{I2，I5}和{I3，I5}。{I3，I5}不是L_2的元素，因而不是频繁的。因此，从C_3中删除{I2，I3，I5}。
- {I2，I4，I5}的2项子集是{I2，I4}、{I2，I5}和{I4，I5}。{I4，I5}不是L_2的元素，因而不是频繁的。因此，从C_3中删除{I2，I3，I5}。

（c）因此，剪枝后$C_3=\{\{I1, I2, I3\}, \{I1, I2, I5\}\}$。

由频繁项集提取关联规则

- 例子：我们计算出频繁项集{I1,I2,I5}，能提取哪些规则？

I1^I2=>I5，由于{I1,I2,I5}出现了2次，{I1,I2}出现了4次，故置信度为2/4=50%

类似可以算出

$$\{I1, I2\} \Rightarrow I5, \quad confidence = 2/4 = 50\%$$
$$\{I1, I5\} \Rightarrow I2, \quad confidence = 2/2 = 100\%$$
$$\{I2, I5\} \Rightarrow I1, \quad confidence = 2/2 = 100\%$$
$$I1 \Rightarrow \{I2, I5\}, \quad confidence = 2/6 = 33\%$$
$$I2 \Rightarrow \{I1, I5\}, \quad confidence = 2/7 = 29\%$$
$$I5 \Rightarrow \{I1, I2\}, \quad confidence = 2/2 = 100\%$$

用 R 进行购物篮分析

- 安装arules包并加载
- 内置Groceries数据集

library(arules) #加载arules程序包

data(Groceries) #调用数据文件

Inspect(Groceries) #观看数据集里的数据

```
      specialty bar}
9823 {yogurt,
      long life bakery product}
9824 {pork,
      frozen vegetables,
      pastry}
9825 {ice cream,
      long life bakery product,
      specialty chocolate,
      specialty bar}
9826 {chicken,
      hamburger meat,
      citrus fruit,
```

用 R 进行购物篮分析

- 求频繁项集

frequentsets=eclat(Groceries,parameter=list(support=0.05,maxlen=10))

```
parameter specification:
 tidLists support minlen maxlen            target   ext
    FALSE    0.05      1     10 frequent itemsets FALSE

algorithmic control:
 sparse sort verbose
      7   -2    TRUE

eclat - find frequent item sets with the eclat algorithm
version 2.6 (2004.08.16)          (c) 2002-2004   Christian Borgelt
create itemset ...
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [28 item(s)] done [0.00s].
creating sparse bit matrix ... [28 row(s), 9835 column(s)] done [0.00s].
writing  ... [31 set(s)] done [0.02s].
Creating S4 object  ... done [0.00s].
```

用 R 进行购物篮分析

- 观看频繁项集

```
inspect(frequentsets[1:10])
inspect(sort(frequentsets,by="support")[1:10])    #根据支持度对求得的频繁项集排序并察看
```

```
   items                  support
1  {whole milk}           0.25551601
2  {other vegetables}     0.19349263
3  {rolls/buns}           0.18393493
4  {soda}                 0.17437722
5  {yogurt}               0.13950178
6  {bottled water}        0.11052364
7  {root vegetables}      0.10899847
8  {tropical fruit}       0.10493137
9  {shopping bags}        0.09852567
10 {sausage}              0.09395018
```

用 R 进行购物篮分析

- 利用apriori函数提取关联规则

rules=apriori(Groceries,parameter=list(support=0.01,confidence=0.5))

```
> rules=apriori(Groceries,parameter=list(support=0.01,confidence=0.5))

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE    0.01      1     10  rules FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [88 item(s)] done [0.00s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [15 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

用 R 进行购物篮分析

- 列出关联规则

summary(rules)　#察看求得的关联规则之摘要

inspect(rules)

```
> inspect(rules)
    lhs                          rhs                    support confidence     lift
1   {curd,
     yogurt}                  => {whole milk}       0.01006609  0.5823529 2.279125
2   {other vegetables,
     butter}                  => {whole milk}       0.01148958  0.5736041 2.244885
3   {other vegetables,
     domestic eggs}           => {whole milk}       0.01230300  0.5525114 2.162336
4   {yogurt,
     whipped/sour cream}      => {whole milk}       0.01087951  0.5245098 2.052747
5   {other vegetables,
     whipped/sour cream}      => {whole milk}       0.01464159  0.5070423 1.984385
6   {pip fruit,
     other vegetables}        => {whole milk}       0.01352313  0.5175097 2.025351
7   {citrus fruit,
     root vegetables}         => {other vegetables} 0.01037112  0.5862069 3.029608
8   {tropical fruit,
     root vegetables}         => {other vegetables} 0.01230300  0.5845411 3.020999
9   {tropical fruit,
```

用 R 进行购物篮分析

- 按需要筛选关联规则

```
x=subset(rules,subset=rhs%in%"whole milk"&lift>=1.2)    #求所需要的关联规则子集
inspect(sort(x,by="support")[1:5])    #根据支持度对求得的关联规则子集排序并察看
```

其中 lift=P(L,R)/(P(L)P(R)) 是一个类似相关系数的指标。lift=1时表示L和R独立。这个数越大，越表明L和R存在在一个购物篮中不是偶然现象。

Thanks

FAQ时间