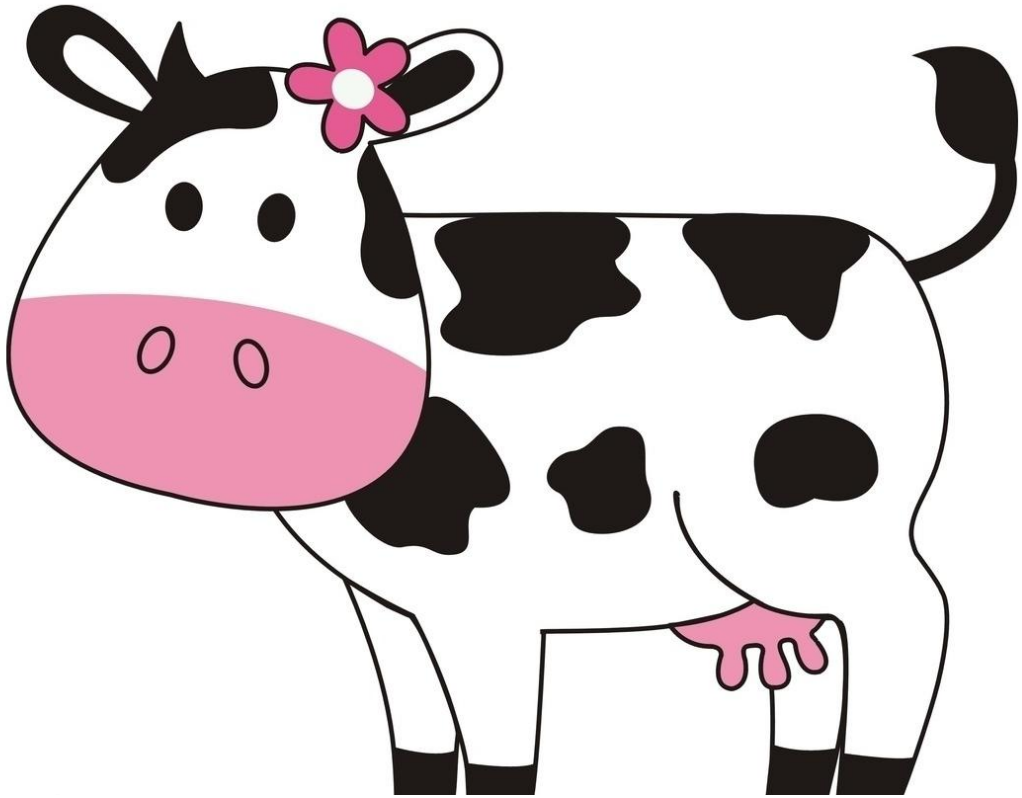
中山大學
SUN YAT-SEN UNIVERSITY



2012级《多元统计分析与数据挖掘》第3周

2015.3.17

中山大学数学与计算科学学院 黄志洪

假设检验的思路

- 作出原假设，替代假设
- 构造统计量，该统计量在假设成立的前提下，满足某种已知分布
- 根据样本计算统计量，看是否落在否定域
- 如果落在否定域内，则拒绝假设，接受替代假设。如果落在否定域外，则接受假设

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

- 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

内推插值与外推归纳

- 在身高与体重的例子中，我们注意到得到的回归方程中的截距项为-140.364 ，这表示身高为0的人的体重是负值，这明显是不可能的。所以这个回归模型对于儿童和身高特别矮的人不适用。
- 回归问题擅长于内推插值，而不擅长于外推归纳。在使用回归模型做预测时要注意x适用的取值范围
- 销售业绩预测适合使用回归吗？

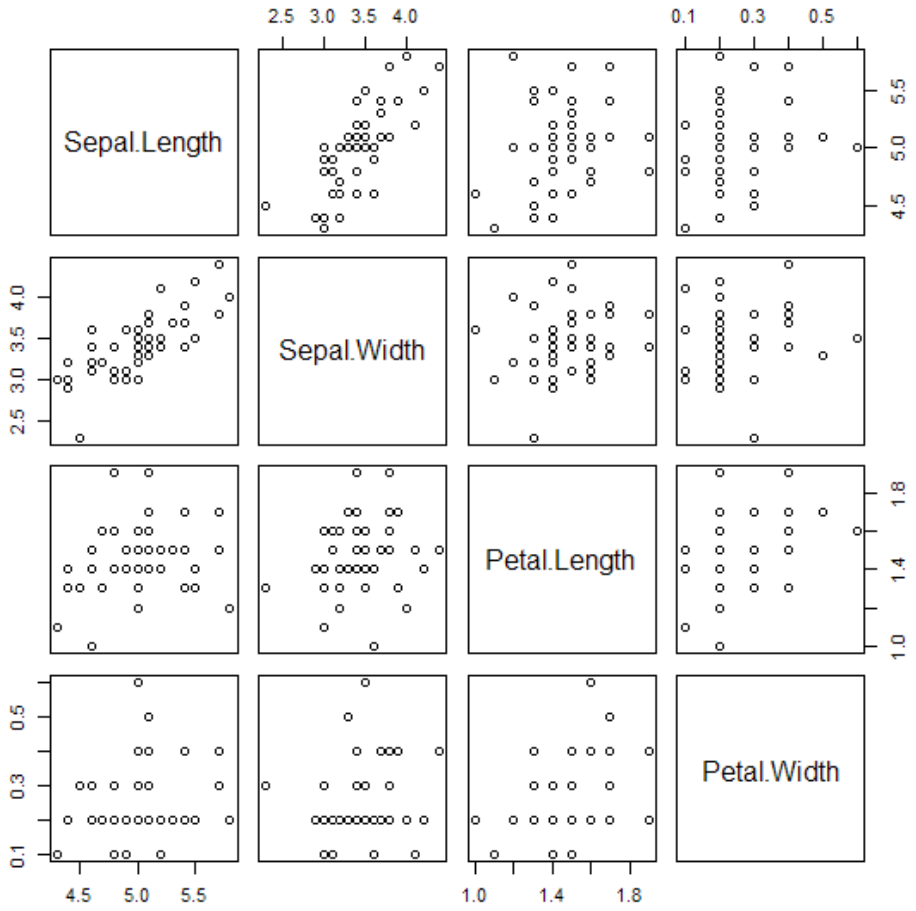
多元线性相关分析

- 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)



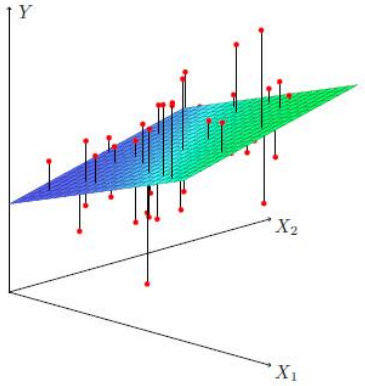
2015.3.17

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
> 
```

多元线性回归模型



- 当Y值的影响因素不唯一时，采用多元线性回归模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_m X_m + \varepsilon$$

- 例如商品的销售额可能与电视广告投入，收音机广告投入，报纸广告投入有关系，可以有

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_m \times \text{newspaper} + \varepsilon$$

参数估计

- 最小二乘法：
- 与一元回归方程的算法相似
- $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 是关于βi的函数。分别对βi求偏导并令偏导等于0，可以解出相应的βi的值

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
```

多元线性回归的几何意义

- 样本空间中的几何意义
- 如果把变量看成矢量（每个观测是矢量的一个分量），呈现的几何意义
- 矩阵形式表达的最小二乘解（薛毅书第268页）

与多元线性回归有关的假设检验

- 薛毅书纸介质第269页
- 回归系数的显著性检验
- 回归方程的显著性检验

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

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

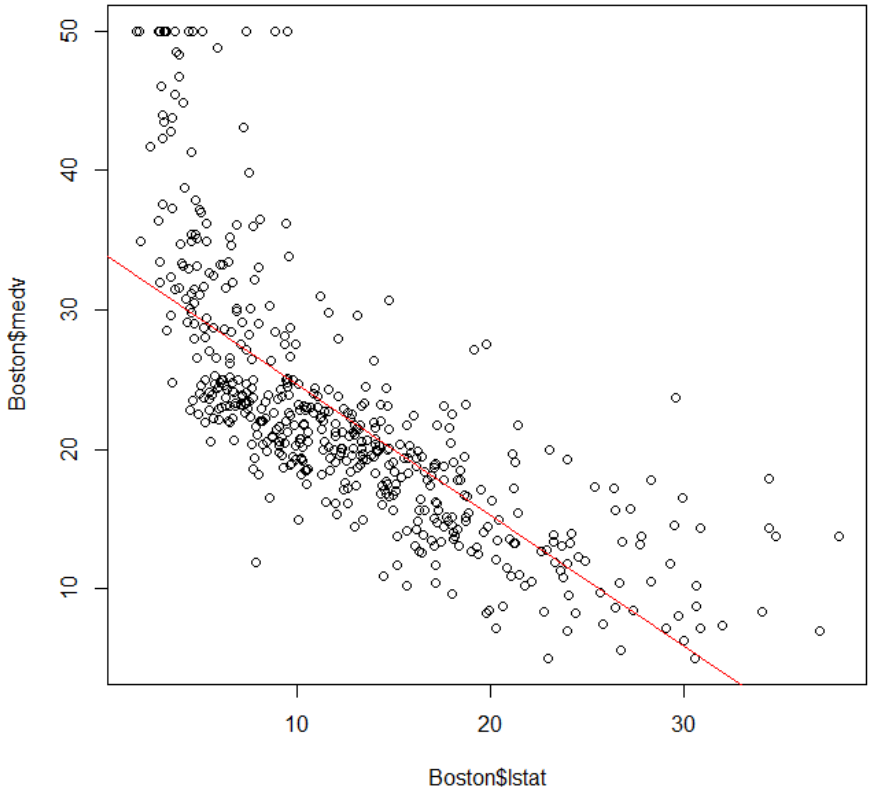
- Boston数据中，chas是一个虚拟变量，Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- 构建medv关于lstat与chas的回归模型
- Y=β0+β1*chas+β2*lstat= $\begin{cases} \beta0+\beta1+\beta2*lstat，chas=1 \\ \beta0+\beta2*lstat，chas=0 \end{cases}$
- 所以，虚拟变量影响的只是

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？

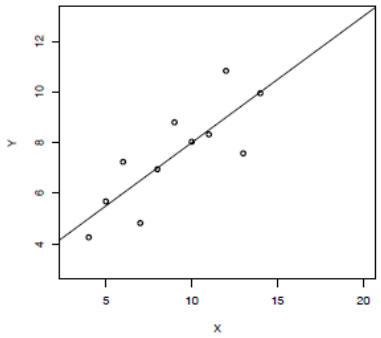
正态分布检验

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

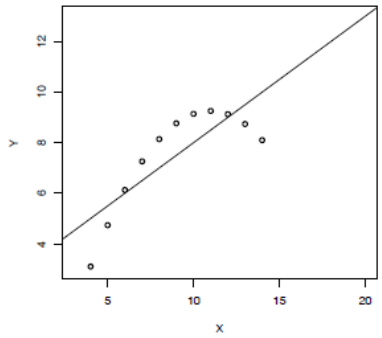
```
> shapiro.test(x$x1)

        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

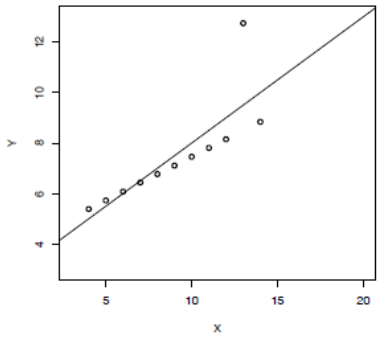
- 薛毅书纸介质p284，例6.11



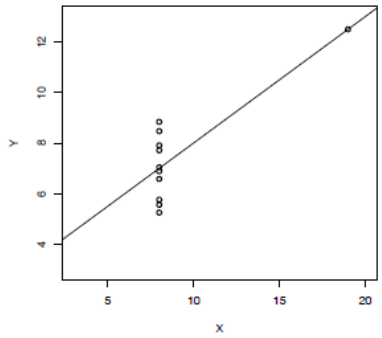
(a) 数据 1



(b) 数据 2



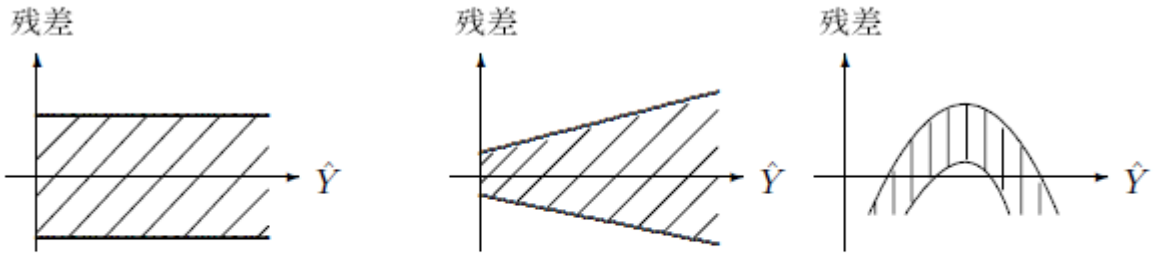
(c) 数据 3



(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

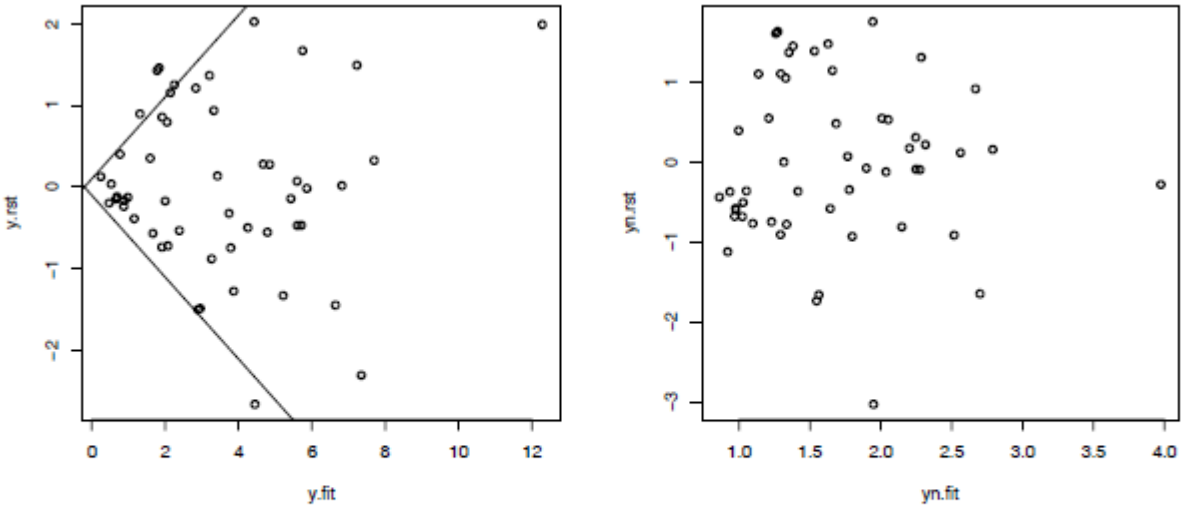


(a) 正常情况　(b) 异方差情况　(c) 非线性情况

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

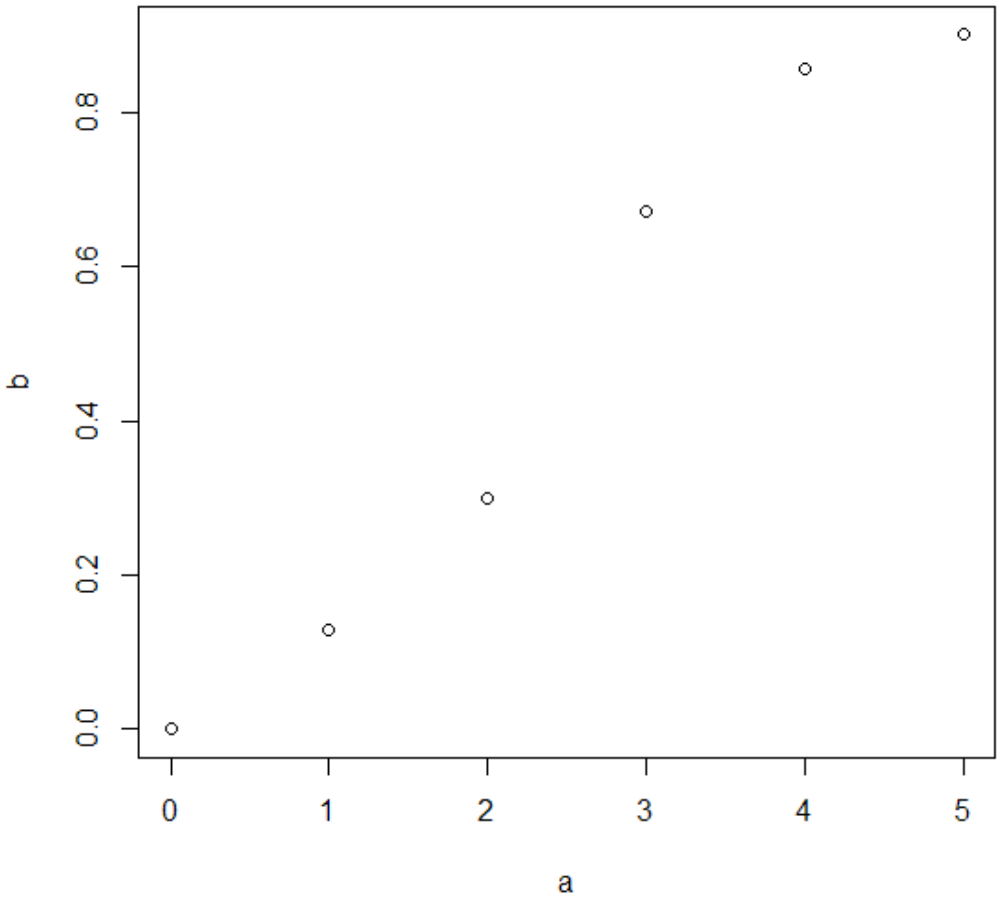
a=c(0:5)

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
```

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

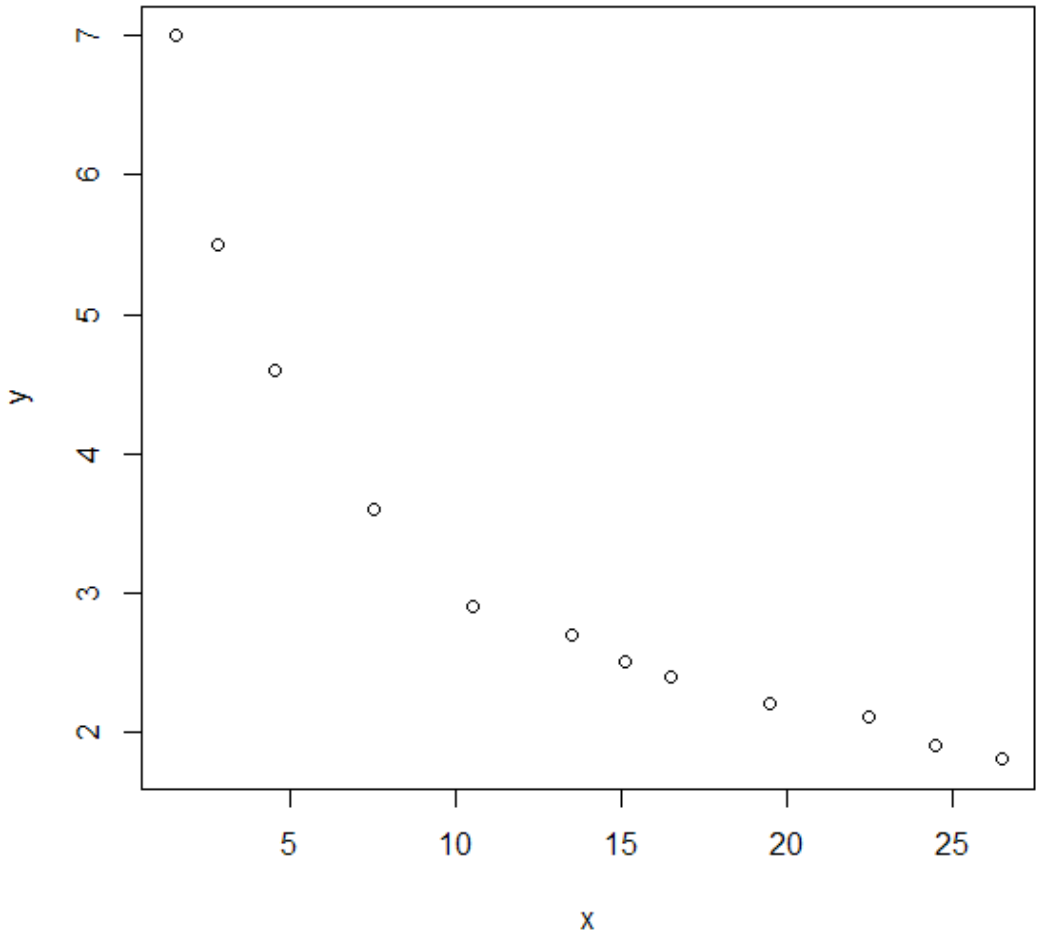
广义线性模型

- 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

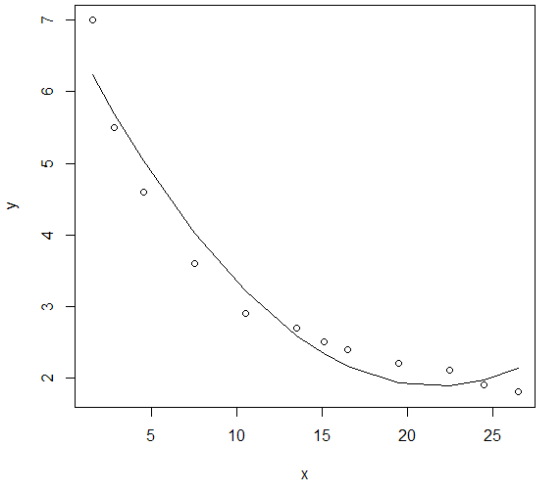
x2=x^2

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

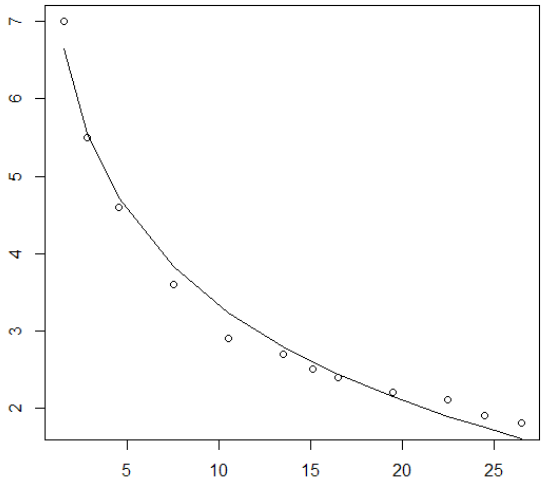
lm.log=lm(y~log(x))

Summar

plot(x,y)

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

.log)

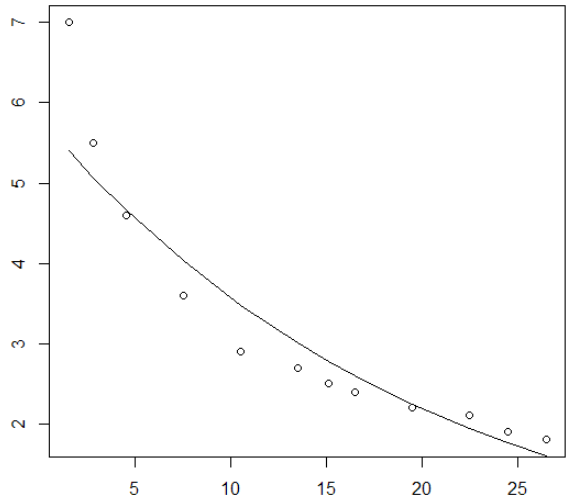



```
Call:
lm(formula = y ~ log(x))

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)))
```



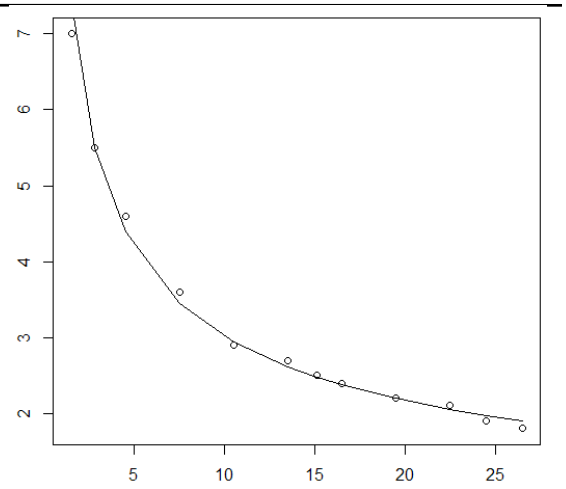

```
Call:
lm(formula = log(y) ~ x)

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

Thanks

FAQ时间