



数据分析、展现与R语言 第6周

2013.2.28

- 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)
```

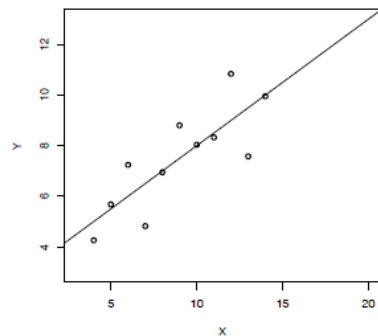
```
Shapiro-Wilk normality test
```

```
data: x$x3
```

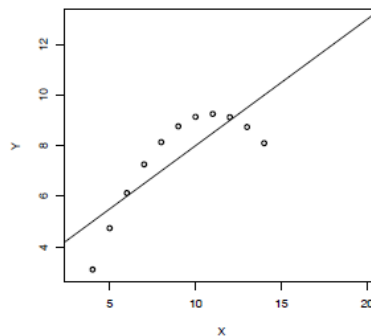
```
W = 0.9444, p-value = 0.0003618
```

散点图目测检验

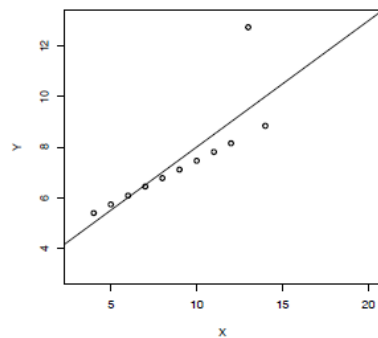
■ 薛毅书纸介质p284，例6.11



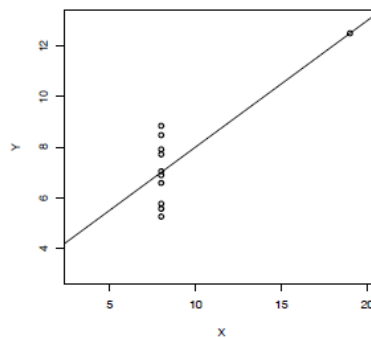
(a) 数据 1



(b) 数据 2



(c) 数据 3



(d) 数据 4

2013.2.28

- 残差计算函数residuals()
- 对残差作正态性检验
- 残差图

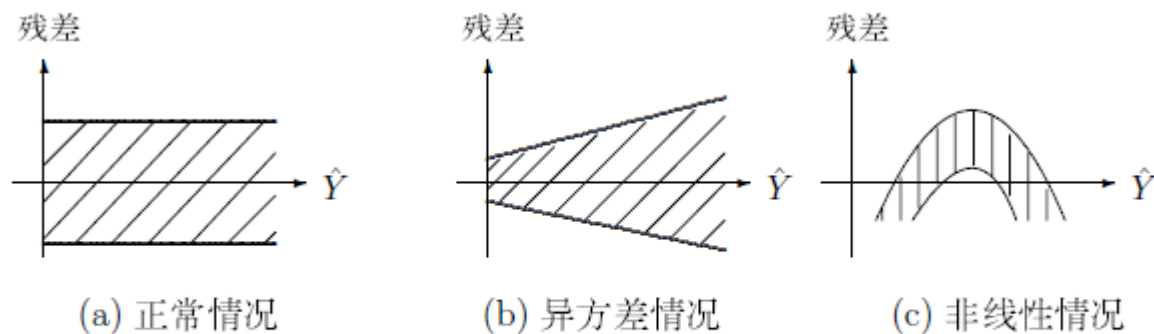
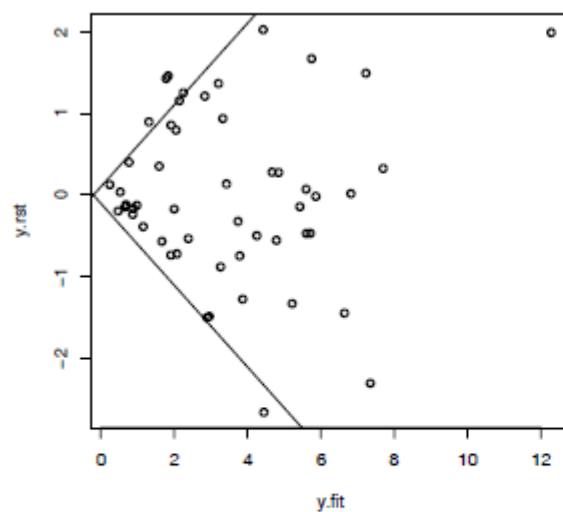
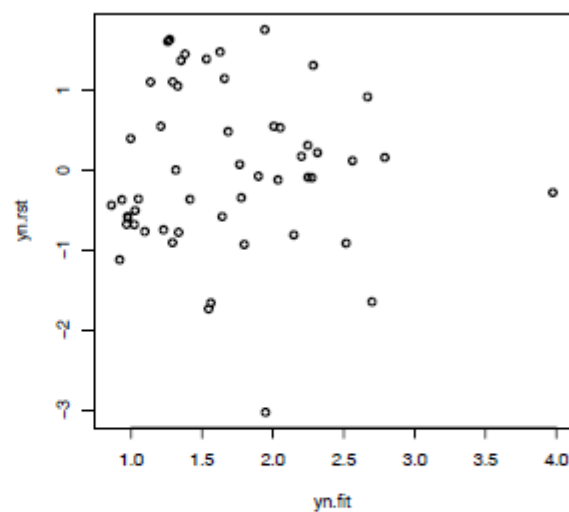


图 6.7: 回归值 \hat{Y} 与残差的散点图

■ 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

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

的影响.

- 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



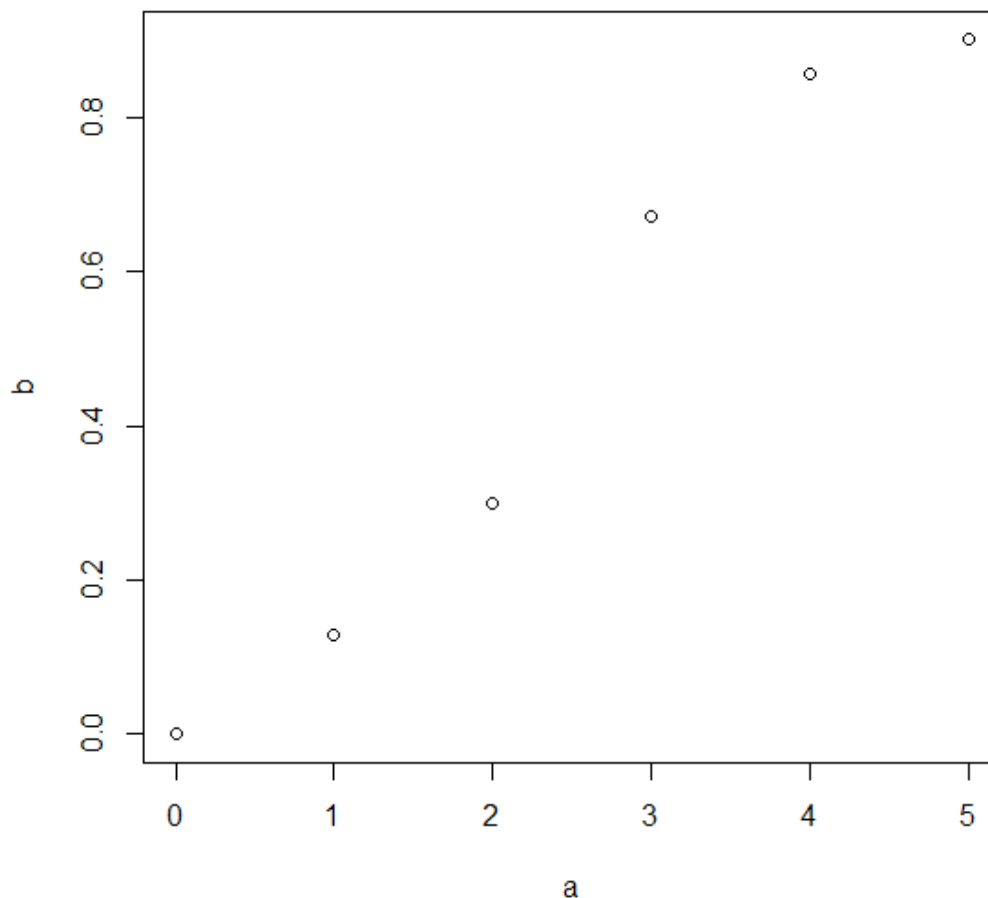
```
a=c(0:5)
```

```
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

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



■ 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)
```

```
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

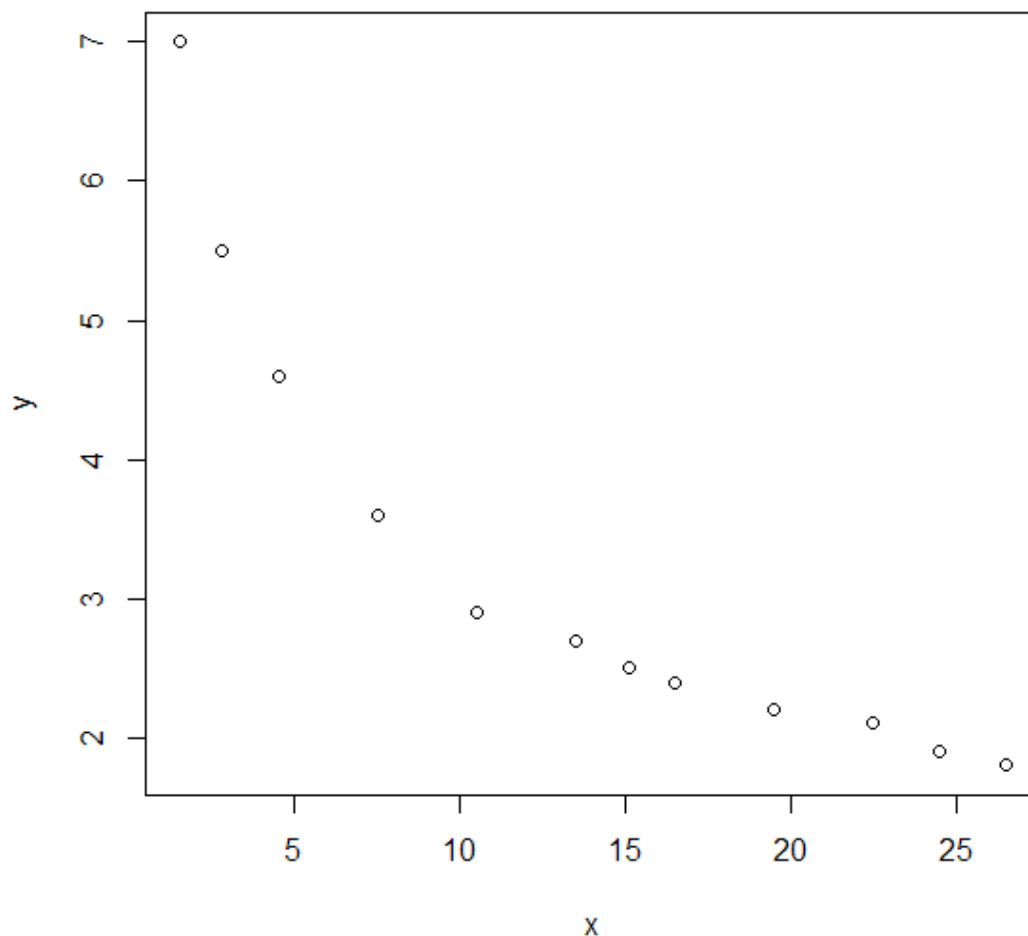
非线性模型

- 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)
```



2013.2.28

■ 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
```


- 多项式回归，假设
用二次多项式方程
 $y=a+bx+cx^2$

`x1=x`

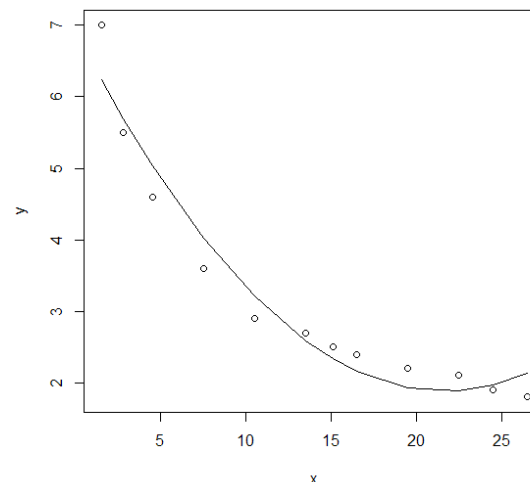
`x2=x^2`

`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
```

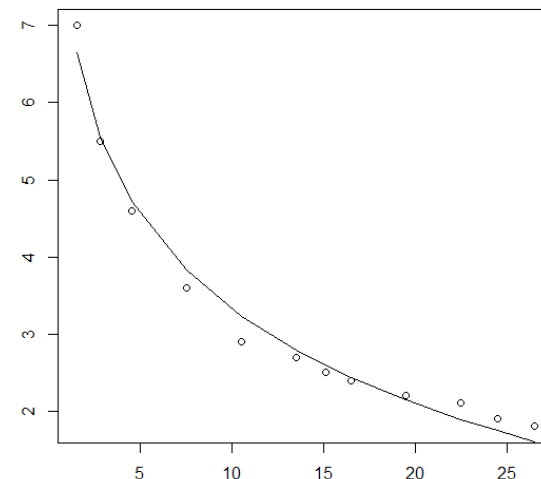
■ 对数法, $y = a + b \log x$

`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

■ 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

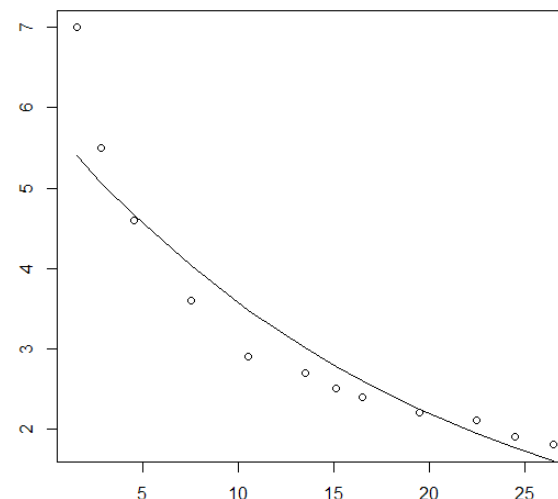
```
Coefficients:
```

	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, $y = a x^b$

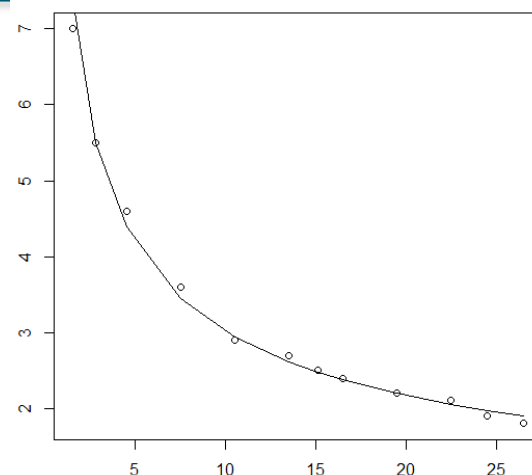
```
lm.pow = lm(log(y) ~ log(x))
```

```
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
```

- 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