

数据分析与R语言 第4周

2012.5.28

第一步 提出原假设

$$H_0: a = a_0 \quad \text{或} \quad H_0: F(x) = F_0(x).$$

如例 8.1.1 中, $H_0: a = 2$. 例 8.1.2 中, $H_0: F(x) = F_0(x)$, 对于光通量 ξ 来说, $F_0(x)$ 是正态分布, 对于呼唤次数 η 来说, $F_0(x)$ 是泊松分布. 可见, 原假设 (又称作零假设) H_0 是我们所要进行检验的对象.

第二步 建立检验统计量

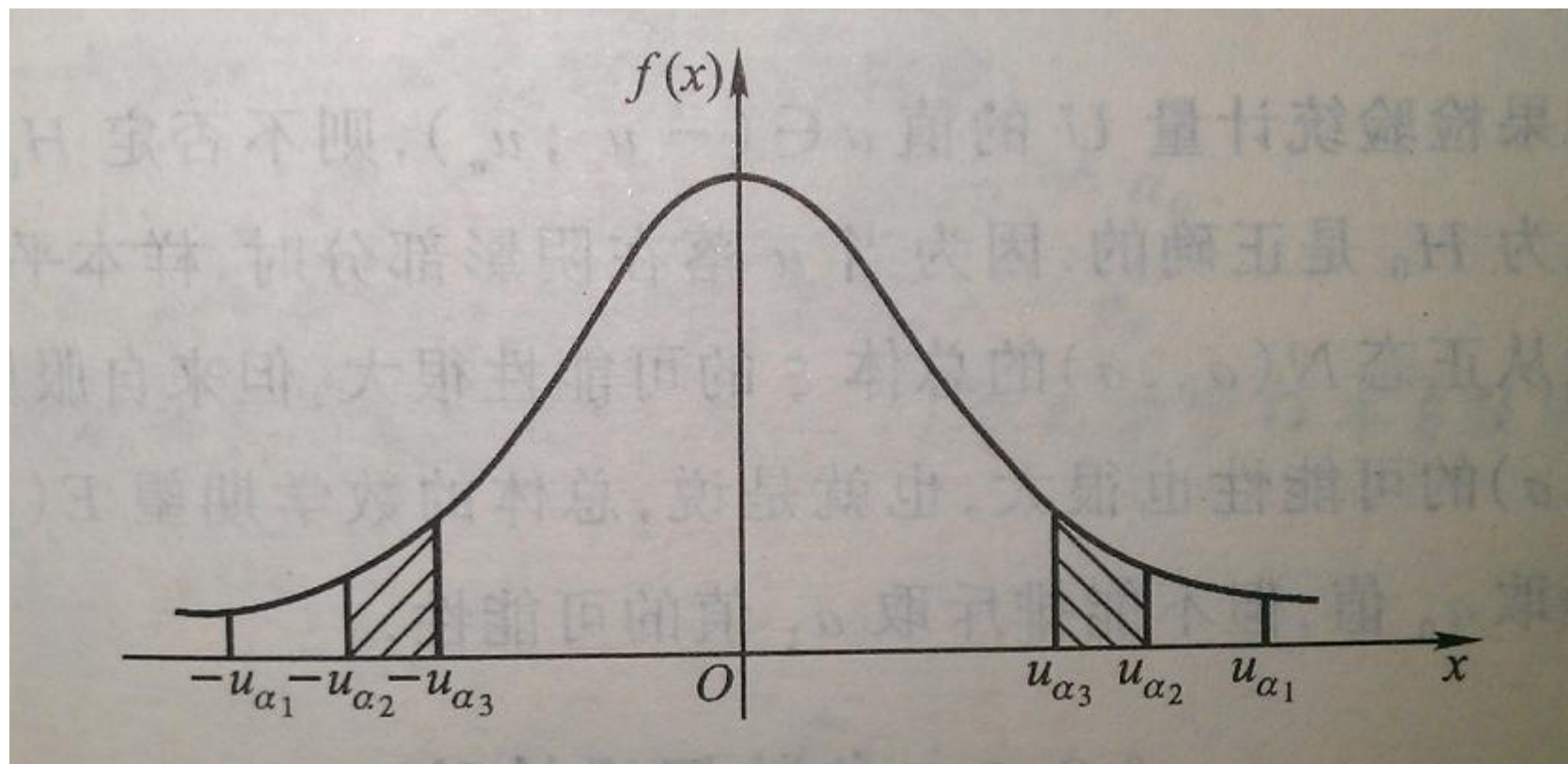
建立检验统计量是假设检验中重要的环节. 比如例 8.1.1 中, 在总体 ξ 服从正态 $N(a, \sigma_0)$ 的假定下, 当原假设 $H_0: a = a_0$ 成立时, 建立检验统计量 $U = \frac{\bar{\xi} - a_0}{\sigma_0 / \sqrt{n}}$, U 服从标准正态 $N(0, 1)$. 注意, 检验统计量是样本的函数, 要求不带有任何未知参数.

对于总体 ξ 的分布函数 $F(x; \theta_1, \theta_2)$ 中参数 θ_1, θ_2 的假设检验, 在 ξ 的分布函数为正态 $N(a, \sigma)$ 的基本假定下, 常用的检验统计量有 t -分布、 χ^2 -分布、 F -分布, 这些适合于小样问题. 如果总 ξ 不服从正态分布, 或总体 ξ 的分布函数未知, 这时检验统计量的精确分布难于求出或相当复杂, 如有可能求出其渐近分布, 则只适用于大样问题. 非参数性的检验问题, 一般都是大样问题, 如 § 8.3 中所讨论的检验问题.

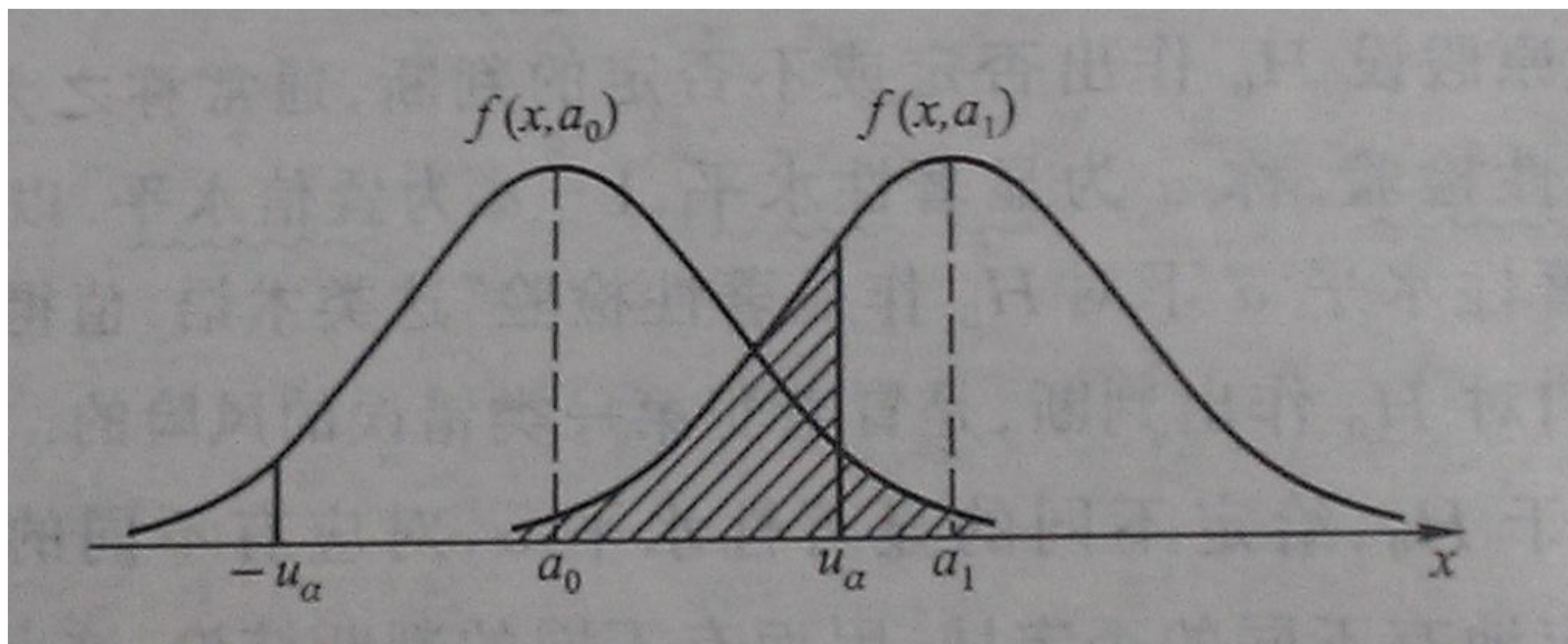
第三步 确定 H_0 的否定域

如例 8.1.1 中,当原假设 H_0 成立时,检验统计量 U 服从正态 $N(0,1)$,那么给定满足 $0 < \alpha < 1$ 的 α 值,在标准正态分布表中查得临界值 u_α ,使得

$$P\{|U| \geq u_\alpha\} = \alpha,$$



2012.5.28



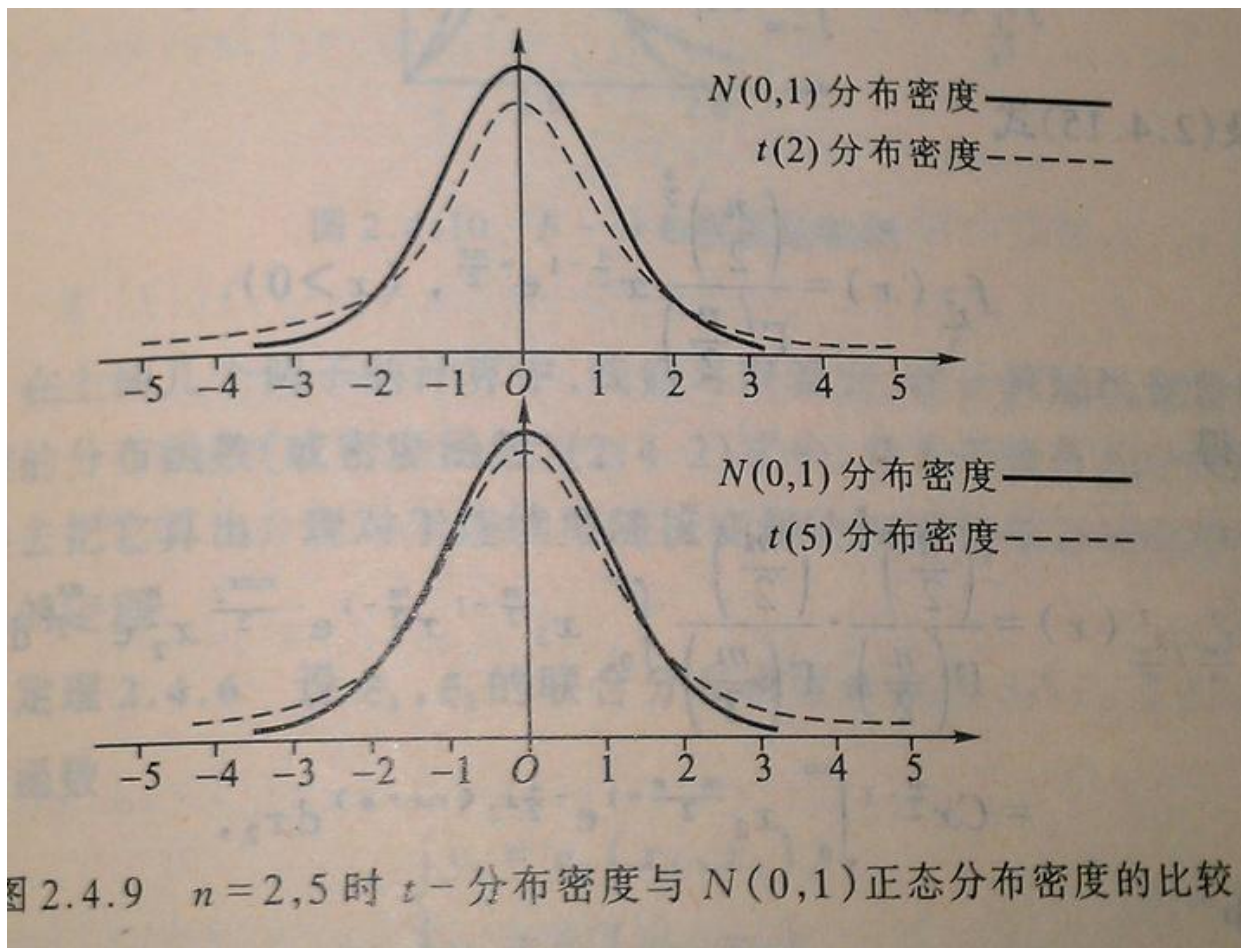
2012.5.28

定理 2.4.4(t -分布) 设 ξ, z 为相互独立随机变量, ξ 服从正态 $N(0, 1)$, z 服从自由度为 n 的 χ^2 -分布, 则 $t = \xi / \sqrt{\frac{z}{n}}$ 的密度函数为

$$f_t(x) = f_{\xi / \sqrt{z/n}}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad (2.4.17)$$

称 $f_t(x)$ 是自由度为 n 的 t -分布(或 Student 分布)的密度函数。

T分布密度函数



2012.5.28

例 8-2-1 药厂制剂车间用自动装瓶机封装药液，在装瓶机工作正常时，每瓶药液净重 500 克。某日随机抽取了 10 瓶成品，称重为：504, 498, 496, 487, 509, 476, 482, 510, 469, 472。问这时的装瓶机工作是否正常？

因此，当原假设 H_0 成立时，记

$$T = \sqrt{n-1} \frac{\bar{\xi} - a_0}{S} \sim t_{(n-1)}, \quad (8.2.2)$$

即统计量 T 服从自由度为 $n-1$ 的 t -分布，且不带有未知参数，它可作为判断 H_0 的检验统计量，这种检验法，称之为 t 检验法。

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 x_1 和 x_2 的多元回归模型 (隐含着截距项)。

- $y \sim 1 + x$ 或 $y \sim x$ 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： $y \sim x - 1$ 或 $y \sim x + 0$ 或 $y \sim 0 + x$

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

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
```

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
```

2012.5.28

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

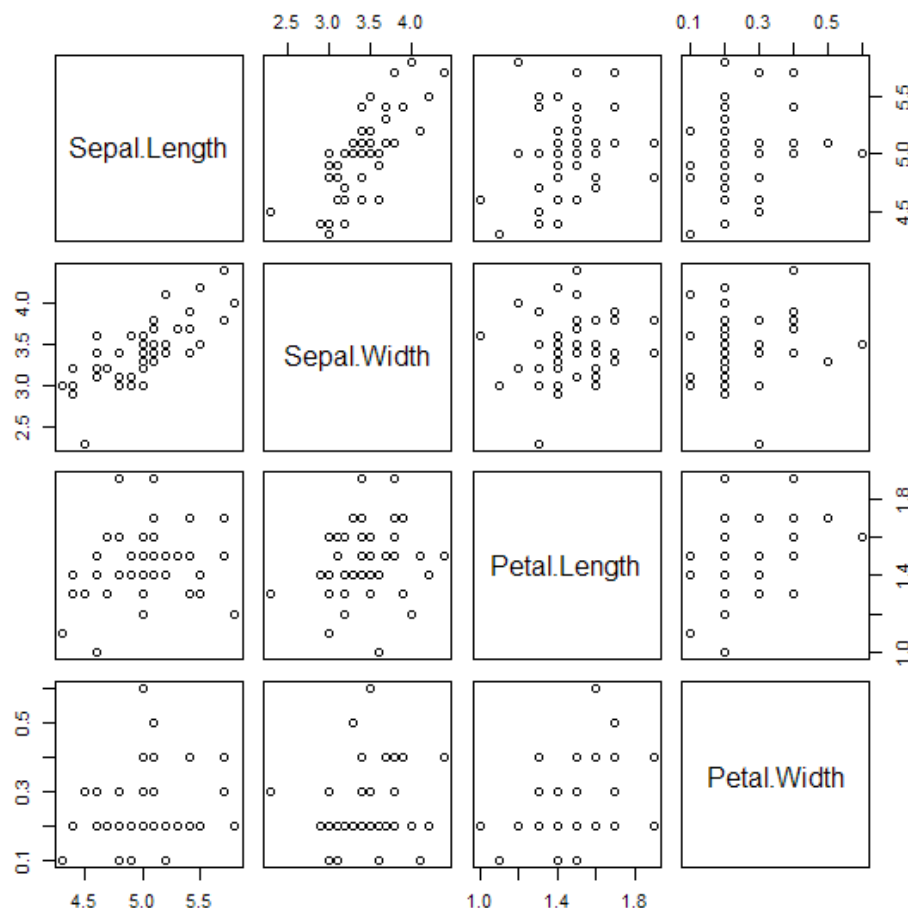
多元线性相关分析

- 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)



2012.5.28

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
> |
```

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

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelay	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

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		

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

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

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

■ 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

```
> |
```


- 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

- 薛毅书，p330例子



Thanks

FAQ时间