暨南大學
JINAN UNIVERSITY

多元统计分析及R语言建模

第4章 多元相关与回归分析及R使用

王斌会 教授

4 多元相关与回归分析及R使用

- ★内容与要求
  - 内容
  - 要求
- 4.1 变量间的关系分析
  - 变量间的关系
  - 简单线性相关
  - 简单线性回归
- 4.2 多元线性回归分析
  - 线性模型建立
  - 线性模型检验
- 4.3 多元线性相关分析
  - 多元线性相关
  - 复相关分析
    - 复相关系数
    - 决定系数
- 4.4 回归变量选择方法
  - 变量选择准则
  - 逐步回归分析
- 思考练习题
  - 思考题
  - 练习题

内容与要求

内容：

变量间的关系分析与回归分析。多元相关回归分析的目的和基本思想，回归变量选择及逐步回归分析方法。

要求：

在学生已具有的（一元）相关与回归分析的基础知识上，掌握和应用多元线性相关与回归分析。

本节内容

1 简单相关分析的R计算

2 一元线性回归分析的R计算

两变量线性相关系数

- 样本的线性相关系数：

$$r=\frac{s_{xy}}{\sqrt{s_x^2\cdot s_y^2}}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

- 离均差平方和与离均差积和：

$$\begin{cases} l_{xx}=\sum(x-\bar{x})^2=\sum x^2-\dfrac{\sum x^2}{n} \\ l_{yy}=\sum(y-\bar{y})^2=\sum y^2-\dfrac{\sum y^2}{n} \\ l_{xy}=\sum(x-\bar{x})(y-\bar{y})=\sum xy-\dfrac{(\sum x)(\sum y)}{n} \end{cases}$$

说明与举例

- 举例：

【例 4-1】（续例2-2）身高与体重的相关关系分析。下面以例2-2的身高与体重数据分析。

- 先建立一个离均差积和函数：

$$l_{xx} = 556.9, l_{yy} = 813, l_{xy} = 645.5$$

$$r = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}} = \frac{645.5}{\sqrt{559.6 * 813}} = 0.9593$$

数据输入：读取身高与体重的数据

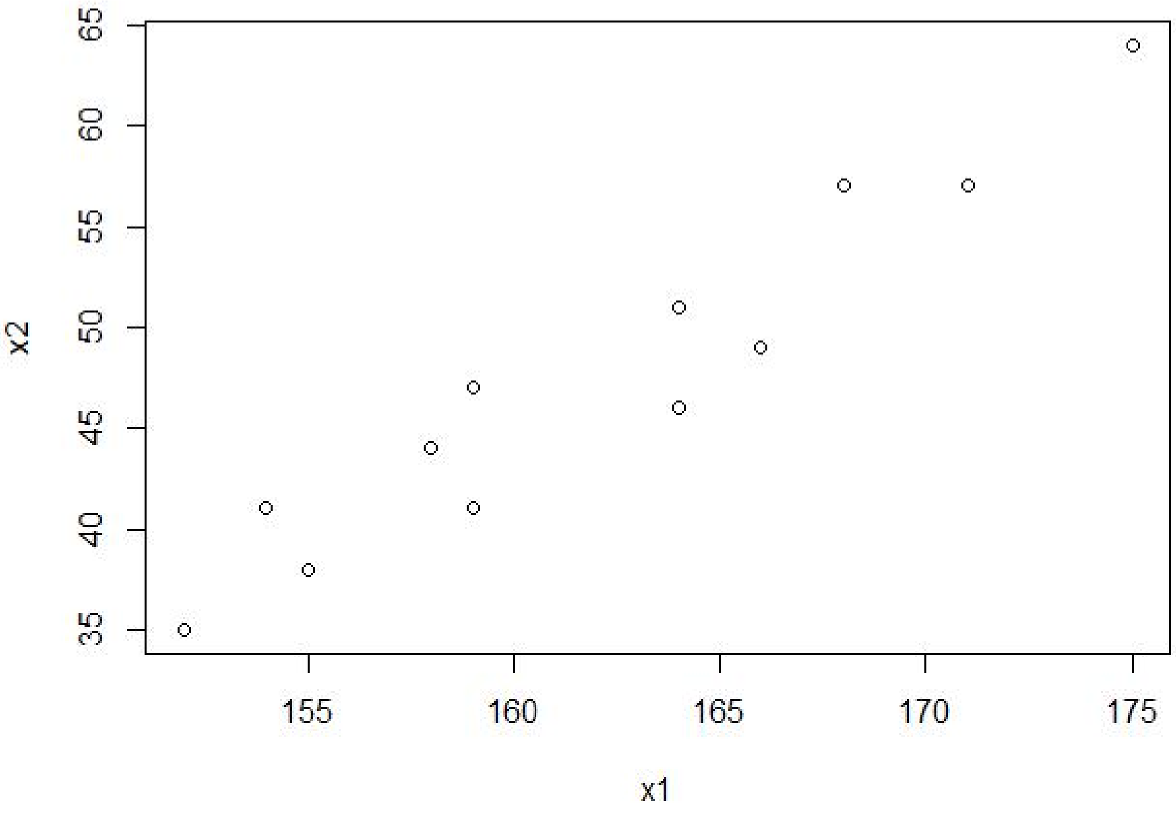
```
x1=c(171,175,159,155,152,158,154,164,168,166,159,164)
x2=c(57,64,41,38,35,44,41,51,57,49,47,46)
```

直观分析：图示法

通过散点图看身高与体重的关系

```
plot(x1,x2)
```

数据输出：



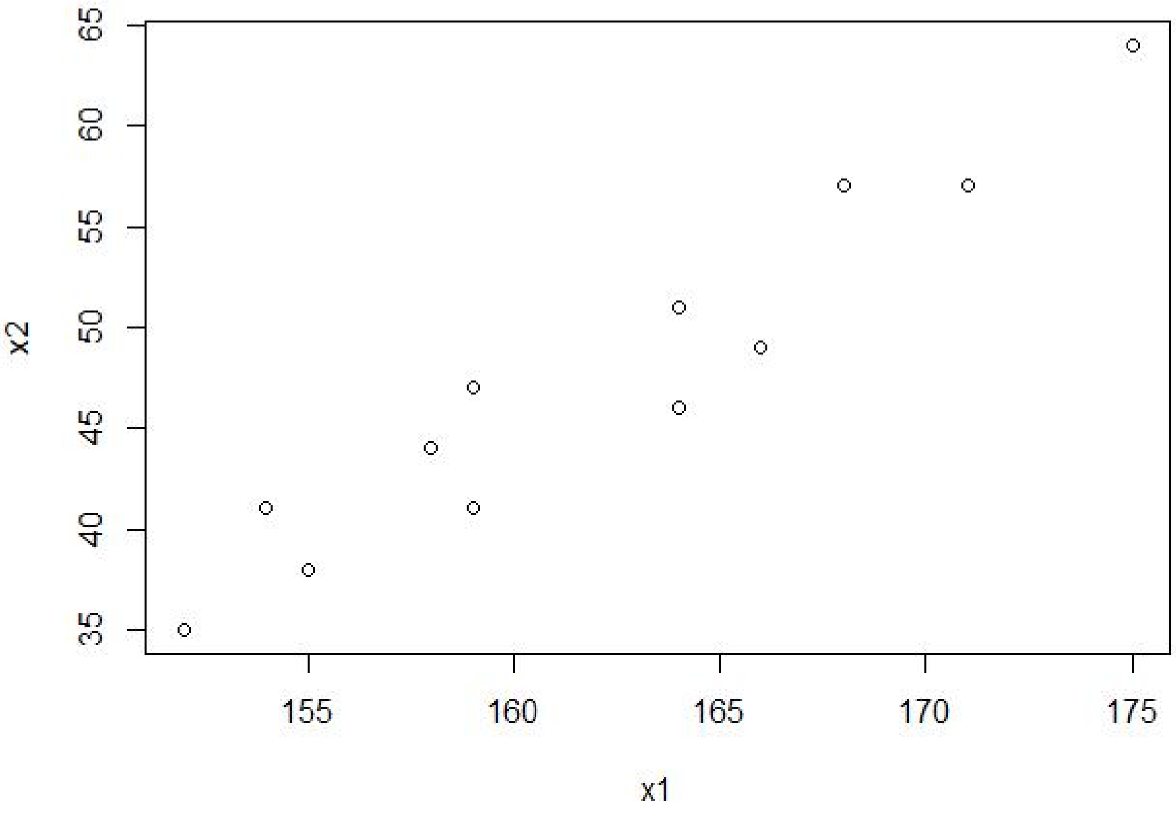
- 建立离均差乘积和函数：

```
lxy<-function(x,y)
sum(x*y)-sum(x)*sum(y)/length(x)
```

- 用离均差乘积和计算相关系数：

```
r=lxy(x1,x2)/sqrt(lxy(x1,x1)*lxy(x2,x2))
r
[1] 0.9593
```

- 数据输出：



相关系数的假设检验

- 建立检验假设：$H_0: \rho = 0, H_1: \rho \neq 0, \alpha = 0.05$
- 计算相关系数r的t值：$t_r = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.9593\sqrt{12-2}}{\sqrt{1-0.9593^2}} = 10.74$

```
n=length(x1)#向量的长度
tr=r/sqrt((1-r^2)/(n-2))#相关系数假设检验t统计量
tr
[1] 10.74
```

相关系数的假设检验

计算t值和P值，作结论：

```
cor.test(x1,x2)#相关系数假设检验
Pearson's product-moment correlation data:  x1 and x2
t = 10.743, df = 10, p-value = 8.21e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8574875 0.9888163
sample estimates:
cor
0.9593031
```

由于$p = 8.21e-07 < 0.05$，于是$\alpha = 0.05$在水准上拒绝H_0，接受H_1的，可认为该人群身高与体重呈现正的线性关系。

说明与举例

- 一元线性回归模型的参数估计：

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{l_{xy}}{l_{xx}} \qquad a=\bar{y}-b\bar{x}$$

- 举例：

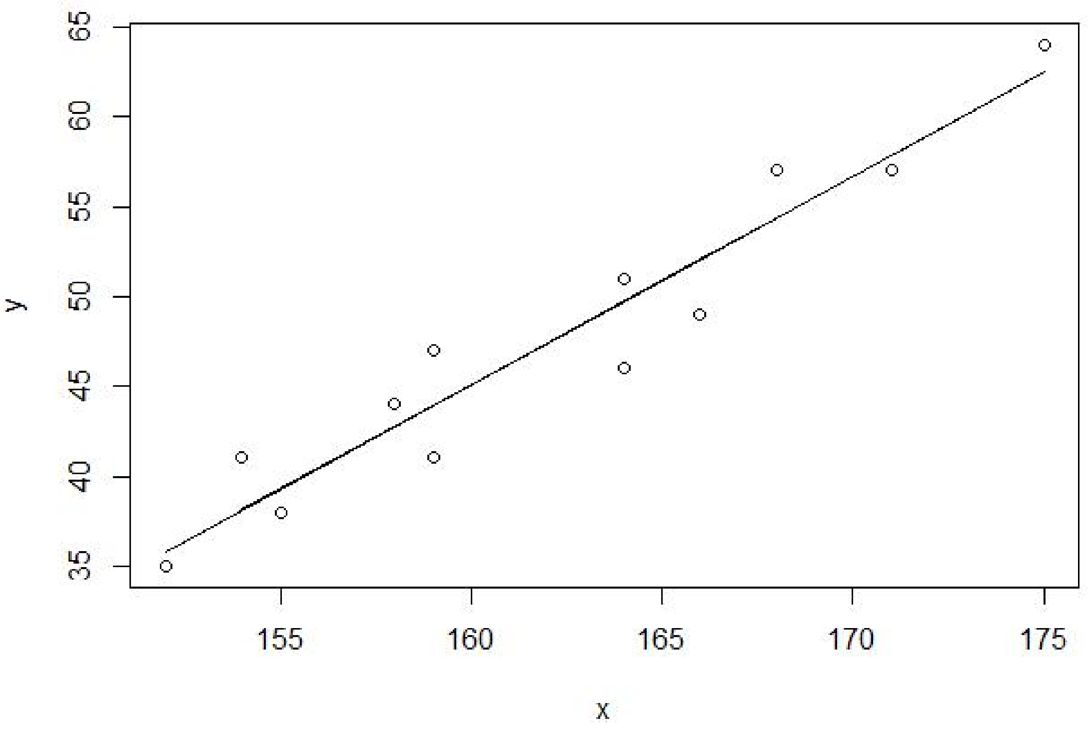
【例 4-2】下面仍以例2-2的数据来介绍建立直线回归方程的步骤。

建立直线回归方程：

```
x=x1#自变量,数据来自例2.2
y=x2#因变量,数据来自例2.2
b=lxy(x,y)/lxy(x,x)#线性回归方程斜率
a=mean(y)-b*mean(x)#线性回归方程截距
c(a=a,b=b)#显示线性回归方程估计值
 a                b
-140.36436    1.15906
```

散点图：

```
plot(x,y)#做散点图
lines(x,a+b*x)#添加估计方程线
```



4.1 变量间的关系分析 ⟹ 回归系数的假设检验

方差分析：

$$MS_R = \frac{SS_R}{df_R}, \quad MS_E = \frac{SS_E}{df_E}, \quad F = \frac{MS_R}{MS_E}$$

其中

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = b l_{xy}$$

$$SS_E = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (\widehat{y_i} - \bar{y})^2$$

t检验：

$$t = \frac{b - \beta}{s_b} \backsim t(n-2)$$

其中

$$s_b = \frac{s_{y,x}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{s_{y,x}}{\sqrt{l_{xx}}}$$

$$s_{y.x} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{MSE}$$

举例：

【例 4-3】以下收集了我国自1978年改革开放以来到2008年共31年的税收(x,百亿元)和财政收入(y,百亿元)数据，试分析税收与财政收入之间的依存关系。

mvstats4.xls

A1 fx

	A	B	C
1		y	x
2	1978	11.3262	5.1928
3	1979	11.4638	5.3782
4	1980	11.5993	5.717
5	1981	11.7579	6.2989
6	1982	12.1233	7.0002
7	1983	18.6695	7.5559
8	1984	16.4286	9.4735
9	1985	20.0482	20.4079
10	1986	21.2201	20.9073
11	1987	21.9935	21.4036
12	1988	23.5724	23.9047
13	1989	26.649	27.274
14	1990	29.371	28.2187
15	1991	31.4948	29.9017
16	1992	34.8337	32.9691
17	[illegible]	[illegible]	[illegible]

○ 数据输入：数据R语言读取

#在mvstats4.xls:d4.3中选取数据，拷贝

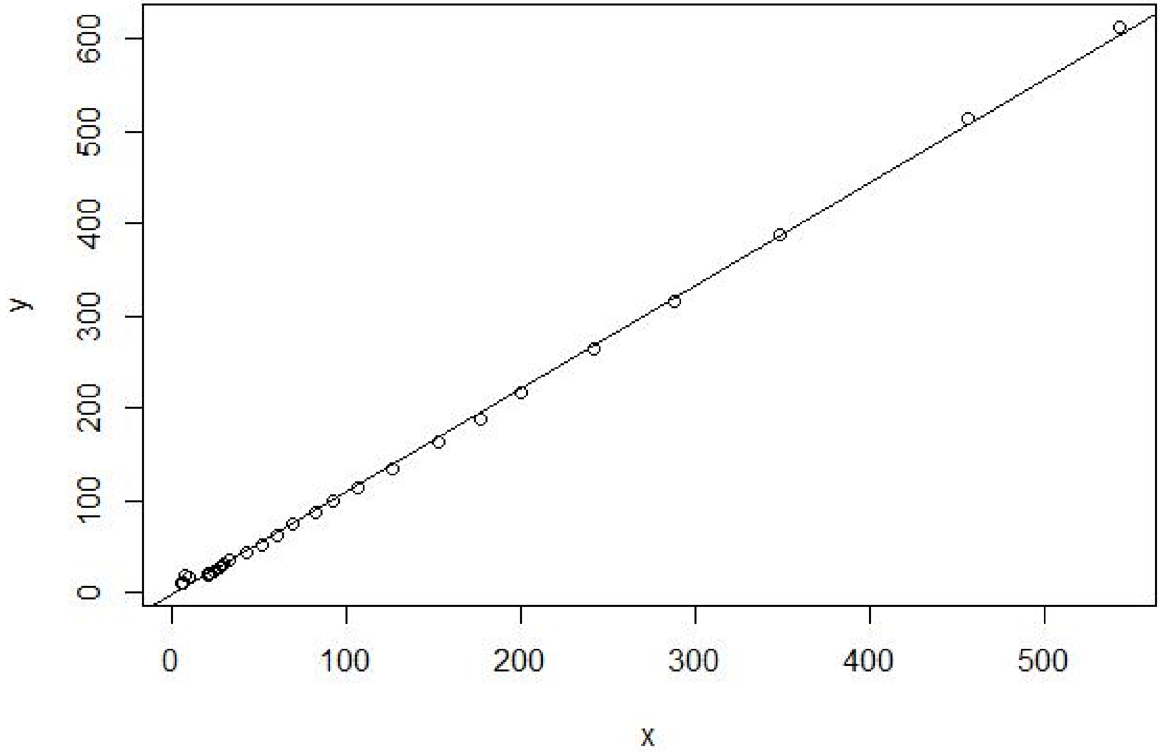
```
yX=read.table("clipboard",header=T)
```

○ 拟合模型

```
(fm=lm(y~x1+x2+x3+x4,data=yX))
Call:
lm(formula = y ~ x, data = yx)
Coefficients:
(Intercept)          x
    -1.197       1.116
```

● 作回归直线：

```
plot(y~x,data=yx)#做散点图
abline(fm)#添加回归线
```



模型的方差分析(ANOVA)

```
anova(fm)#模型方差分析
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 712077  712077   27427 < 2.2e-16 ***
Residuals 29    753      26
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
```

由于$P < 0.05$，于是在$\alpha = 0.05$水平处拒绝H_0，即本例回归系数有统计学意义，x与y间存在直线回归关系。

回归系数的t检验

```
summary(fm)#回归系数t检验
lm(formula = y ~ x, data = yx)
Residuals:
   Min     1Q Median     3Q    Max
-6.631 -3.692 -1.535  5.338 11.432
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.19660    1.16126   -1.03    0.311
x            1.11623    0.00674  165.61   <2e-16 ***
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
Residual standard error: 5.095 on 29 degrees of freedom
Multiple R-squared:  0.9989,     Adjusted R-squared:  0.9989
```

由于$P < 0.05$，于是在$\alpha = 0.05$水平处拒绝H_0，即本例回归系数有统计学意义，x与y间存在直线回归关系。

说明与举例

多元回归参数的最小二乘估计：

从多元线性模型的矩阵形式$y = X\beta + \epsilon$可知，若模型的参数β的估计量$\hat{\beta}$已获得，则$\hat{y} = X\hat{\beta}$，于是残差$e_i = y_i - \hat{y}_i$，根据最小二乘的原理，所选择的估计方法应是估计值$\hat{y}_i$与观察值y_i之间的残差e_i在所有样本点上达到最小，即使

$$Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = e'e = (y - X\hat{\beta})'(y - X\hat{\beta})$$

达到最小，根据微积分求极值的原理，Q对β求导且等于0，可求得使Q达到最小的$\hat{\beta}$，这就是所谓的最小二乘(LS)法。

举例：

【例 4-4】在例4-3中我们发现1978-2008年我国财政收入与税收之间存在线性回归关系，为进一步考察财政收入和其它变量之间的数量关系，需建立多元线性回归方程。

数据表如下：

表4.4 财政收入多因素分析数据

y	x1	x2	x3	x4
1978	11.3262	36.241	5.1928	3.550
1979	11.4638	40.382	5.3782	4.120
1980	11.5993	45.178	5.7170	5.700
…	…	…	…	…
2007	513.2178	2495.299	456.2197	1667.402
2008	613.3035	3006.7	542.1962	1778.8983

得到多元线性回归方程：

$$\hat{y} = 23.5321 - 0.003387x_1 + 1.1641x_2 + 0.000292x_3 - 0.04374x_4$$

建立多元线性回归方程：

```
yX=read.table("clipboard",header=T)
fm=lm(y~x1+x2+x3+x4,data=yX)
fm

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = yX)
Coefficients:
(Intercept)        x1         x2         x3         x4
 23.5321088   -0.0033866    1.1641150
0.0002919   -0.0437416
```

标准化偏回归系数：

$$\hat{\beta}^* = \hat{\beta}_i \frac{s_i}{s_y} \; (i = 1, 2, \ldots, p)$$

```
library(mvstats)
coef.sd(fm)#标准化偏回归系数结果

$coef.sd
      x1        x2        x3        x4
-0.01745   1.0423   0.00096  -0.037105
```

常用的统计软件都能给出标准化偏回归系数，但R语言中并不包含计算标准回归系数的函数，我们编写了$coef.sd$计算之。例4.4的R软件给出标准化偏回归系数如下：$\hat{\beta}_1^* = -0.01745, \hat{\beta}_2^* = 1.0424, \hat{\beta}_3^* = 0.00096, \hat{\beta}_4^* = -0.0371$，由标准化偏回归系数可见，税收对财政收入的线性影响最大。

多元回归方差分析：

$$F=\frac{MS_R}{MS_E}\sim F(p,n-p-1)$$

其中

$$MS_R=\frac{SS_R}{df_R}=\sum_{i=1}^{n}\frac{(\hat{y}_i-\bar{y})^2}{p}$$

$$MS_E=\frac{SS_E}{df_E}$$

方差分解为：

$$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=SS_R+SS_E$$

回归系数的t检验：

$$t_j=\frac{\hat{\beta}_j-\beta_j}{s_{\hat{\beta}_j}}\quad j=1,2,\ldots,p$$

其中

$$s_{\hat{\beta}_j}=\sqrt{c_{jj}}\,s_{y,x}$$

$$s_{y,x}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1}}=\sqrt{\frac{SS_E}{df_E}}=\sqrt{MS_E}$$

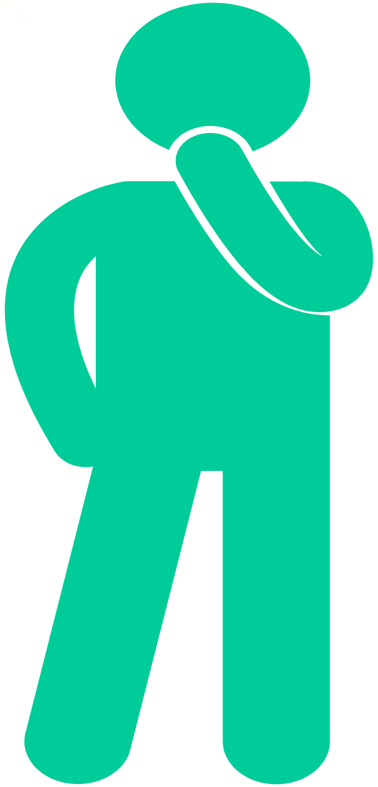
例4-4的t检验：

```
summary(fm)#多元线性回归系数t检验
lm(formula = y ~ x1 + x2 + x3 + x4, data = yX)
Residuals:
    Min      1Q  Median      3Q     Max
-5.0229 -2.1354  0.3297  1.2639  6.9690
```

```
Coefficients:
              Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 23.5321088  4.5990714   5.117   2.47e-05 ***
x1          -0.0033866  0.0080749  -0.419   0.678
x2           1.1641150  0.0404889  28.751   < 2e-16 ***
x3           0.0002919  0.0085527   0.034   0.973
x4          -0.0437416  0.0092638  -4.722   7.00e-05 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.79 on 26 degrees of freedom
Multiple R-squared:  0.9997,Adjusted R-squared:  0.9997
F-statistic: 2.289e+04 on 4 and 26 DF,  p-value: < 2.2e-16
```

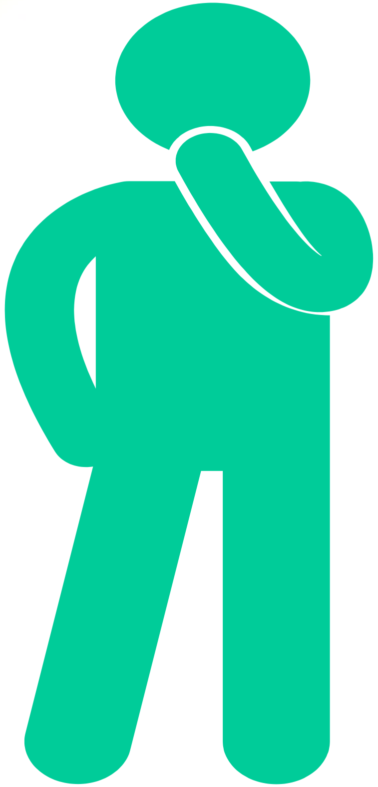
由t检验结果可见，偏回归系数b_2、b_4的P值都小于0.01，可认为解释变量税收x_2、经济活动人口x_4显著；b_1、b_3的P值大于0.50，不能否定$\beta_1=0, \beta_3=0$的假设，可认为国内生产总值x_1、进出口贸易总额x_3对财政收入y没有显著的影响。我们可以看到，国内生产总值、经济活动人口所对应的偏回归系数都为负，这与经济现实是不相符的。出现这种结果的可能原因在于，这些解释变量之间存在高度的共线性。

在相关分析中，研究较多的是两个变量之间的关系，称为简单相关；当涉及到的变量为三个或者三个以上时，称为偏相关或复相关。实际上，偏相关和复相关是对简单相关的一种推广。。。。



设样本矩阵为：

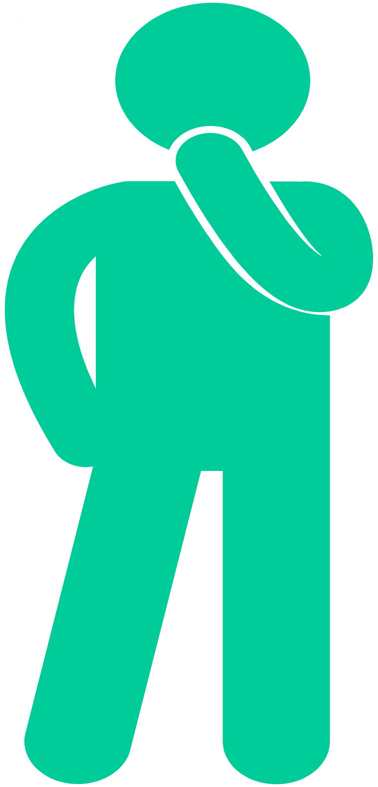
$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$



此时任意两个变量间相关系数构成的矩阵为：

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

$$=\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}=(r_{ij})_{p\times p}$$



其中r_{ij}为任意两变量之间的简单相关系数：

$$r_{ij}=\frac{\sum_{ij}(x_i-\bar{x})(y_j-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2\sum_j(y_j-\bar{y})^2}}$$

举例与说明

(续例4.4)财政收入与其他变量间的相关分析。

计算财政收入和国民生产总值及税收、进出口贸易总额、经济活动人口两两之间相关系数，表4.9给出了相关系数的假设检验统计量。

首先我们计算变量两两间的相关系数

R语言代码

```
#多元数据相关系数矩阵
cor(yX)
```

数据输出

	y	x1	x2	x3	x4
y	1.0000	0.9871	0.9995	0.9912	0.6957
x1	0.9871	1.0000	0.9907	0.9868	0.7818
x2	0.9995	0.9907	1.0000	0.9917	0.7154
x3	0.9912	0.9868	0.9917	1.0000	0.7074
x4	0.6957	0.7818	0.7154	0.7074	1.0000

函数说明

由于没有现成的进行相关系数矩阵的假设检验，

下面编写计算相关系数的值和值的函数corr.test()。

相关矩阵检验函数 **corr.test()**的用法
corr.test(X, ...)
X 数值矩阵或数据框

R语言代码

```
library(mvstats)
#多元数据相关系数检验
corr.test(yX)
```

数据输出

	y	x1	x2	x3	x4
y	0.000	0.000	0.000	0.000	0
x1	33.267	0.000	0.000	0.000	0
x2	165.614	39.214	0.000	0.000	0
x3	40.336	32.772	41.560	0.000	0
x4	5.215	6.752	5.514	5.389	0

左下角为 t 值，右上角为 p 值

从**结果**可以看出，财政收入和国民生产总值及税收、进出口贸易总额、经济活动人口之间的关系都非常密切，财政收入与税收之间的关系最为密切。

复相关分析

在实际分析中，一个变量的变化往往要受到多种变量的综合影响，这时就需要采用复相关分析方法。所谓复相关，就是研究多个变量同时与某个变量之间的相关关系，度量复相关程度的指标是复相关系数。

复相关系数

假定回归模型为：

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+e$$

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$$

复相关系数

复相关系数计算公式为：

$$R = corr(y, x_1, x_2, ..., x_p) = corr(y, \hat{y})$$

$$= \frac{cov(y, \hat{y})}{\sqrt{var(y)var(\hat{y})}} = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}}$$

决定系数

复相关系数：

$$R = \sqrt{\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}} = \sqrt{\frac{SS_R}{SS_T}}$$

决定系数：

$$R^2 = \frac{SS_R}{SS_T}$$

R语言代码

```
#显示多元线性回归模型决定系数
(R2=summary(fm)$r.sq)
```

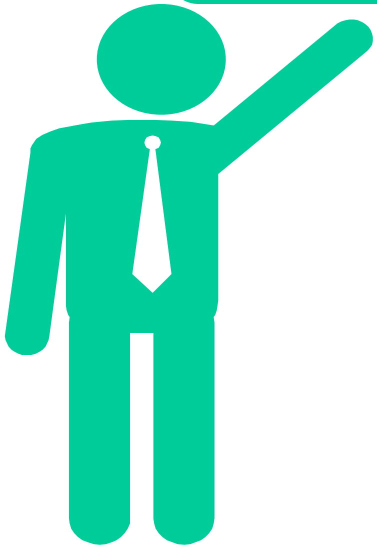
```
#显示多元数据复相关系数
(R=sqrt(R2))
```

数据输出

[1] 0.9997

[1] 0.9999

多元回归分析主要用途



- ✓ 用于描述解释现象，这时希望回归方程中所包含的自变量尽可能少一些
- ✓ 用于预测，这时希望预测的均方误差较小
- ✓ 用于控制,这时希望各回归系数具有较小的方差和均方误差

4.4 回归变量的选择方法

- ✓ 变量多增加了模型的复杂
- ✓ 计算量增大
- ✓ 估计和预测的精度下降
- ✓ 模型应用费用增加

4.4 回归变量的选择方法

- ✓ 全部子集法
- ✓ 向后删除法
- ✓ 向前引入法
- ✓ 逐步回归法

4.4 回归变量的选择方法

全局最优法

从理论上说，自变量选择最好的方法是所有可能回归法，即建立因变量和所有自变量全部子集组合的回归模型，也称全部子集法。

对于每个模型，在实用上，从数据与模型拟合优劣的直观考虑出发，基于残差（误差）平方和的变量选择准则使用的最多。

举例与说明

【例4.6】（续例4.4）在“财政收入”数据中，有4个自变量：x_1,x_2,x_3,x_4。所有可能的模型可分为5组子集：

子集$A: y = b_0 \Longrightarrow C_4^0 = 1$种可能模型。

子集$B: y = b_0 + b_i x_i, i = 1,2,3,4 \Longrightarrow C_4^1 = 4$种可能模型。

子集$C: y = b_0 + b_i x_i + b_j x_j, i \neq j, i,j = 1,2,3,4 \Longrightarrow C_4^2 = 6$种可能模型。

子集$D: y = b_0 + b_i x_i + b_j x_j + b_k x_k,, i \neq j \neq k, i,j,k = 1,2,3,4 \Longrightarrow C_4^3 = 4$种可能模型

子集$E: y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \Longrightarrow C_4^4 = 1$种可能模型。

$\Longrightarrow$总共有$C_4^0 + C_4^1 + C_4^2 + C_4^3 + C_4^4 = 2^4 = 16$个模型。

举例与说明

例4.4数据的RSS与R2准则回归子集:

子集	Models	RSS	R^2
子集 B	$y = b_0 + b_2x_2$	752.88	0.99894
子集 C	$y = b_0 + b_2x_2 + b_4x_4$	203.88	0.99971
子集 D	$y = b_0 + b_1x_1 + b_2x_2 + b_4x_4$	202.35	0.99972
子集 E	$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$	202.34	0.99972

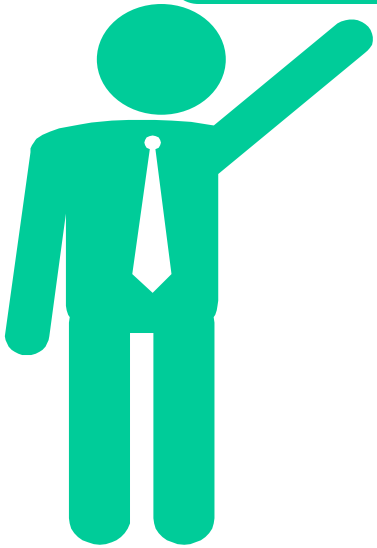
R语言代码

```
library(leaps) #加载leaps包
varsel=regsubsets(y~x1+x2+x3+x4,data
=yX)
result=summary(varsel)
data.frame(result＄outmat,RSS=result＄
rss,R2=result$rsq)
```

数据输出

	x1	x2	x3	x4	RSS	R2
1 (1)		*			752.88	0.99894
2 (1)		*		*	203.88	0.99971
3 (1)	*	*		*	202.35	0.99972
4 (1)	*	*	*	*	202.34	0.99972

R^2和RSS准则优缺点



具有较大R^2的对较少自变量的模型应该是好的选择，较大的意味着有好的拟合效果，而较少的变量个数可减轻信息的收集和控制。

对于有个自变量的回归模型来说，当自变量子集在扩大时，残差平方和随之减少。因此，如果按RSS“愈小愈好”和按R^2”愈大愈好”的原则来选择自变量子集，则毫无疑问应该选全部自变量

4.4 回归变量的选择方法

变量选择的常用准则

- ✓ **平均残差平方和**最小准则
- ✓ 误差均方根**MSE**最小准则
- ✓ 校正复相关系数平方（**Adjusted R^2**）准则
- ✓ C_p准则
- ✓ **AIC**准则**BIC**准则

举例与说明

表4.10例4.4数据的Cp与BIC准则回归子集

子集	Models	$AdjR^2$	Cp	BIC
子集B	$y = b_0 + b_2x_2$	0.9989	69.745	-205.6
子集C	$y = b_0 + b_2x_2 + b_4x_4$	0.9997	1.199	-242.6
子集 D	$y = b_0 + b_1x_1 + b_2x_2 + b_4x_4$	0.9997	3.001	-239.4
子集 E	$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$	0.9997	5.000	-236.0

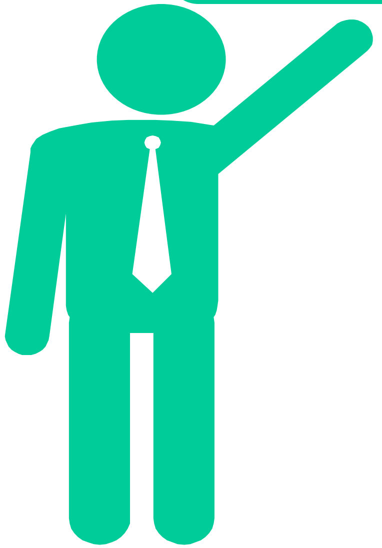
R语言代码

```
data.frame(result $ outmat,
adjR2=result $ adjr2,Cp=result $ cp,
BIC=result$bic)
```

数据输出

		x1	x2	x3	x4	adjR2	Cp	BIC
1	(1)		*			0.9989	69.745	-205.6
2	(1)		*		*	0.9997	1.199	-242.6
3	(1)	*	*		*	0.9997	3.001	-239.4
4	(1)	*	*	*	*	0.9997	5.000	-236.0

全局择优法的缺陷



如果自变量个数为4，则所有的回归有15个，当自变量个数为10时，所有可能的回归为1023个，…，当自变量数个数为50时，所有可能的回归为1015个，当p很大时，数字2^p大得惊人，有时计算是不可能的，于是就提出了所谓逐步回归的方法.

逐步回归分析

在作实际多元线性回归时常有这样情况，变量$x_1,x_2,\ldots x_p$相互之间常常是线性相关的，即在$x_1,x_2,\ldots x_p$中任何两个变量是完全线性相关的，即相关系数为1，则矩阵X^TX的秩小于p，X^TX就无解。当变量$x_1,x_2,\ldots x_p$中任有两个变量存在较大的相关性时，矩阵X^TX处于病态，会给模型带来很大误差。因此作回归时，应选变量$x_1,x_2,\ldots x_p$中的一部分作回归，剔除一些变量。逐步回归法就是寻找较优子空间的一种变量选择方法。

- ✓ 向前引入法
- ✓ 向后剔除法
- ✓ 逐步筛选法

R语言代码

```
fm=lm(y~x1+x2+x3+x4, data=yX)
fm.step=step(fm,direction="forward")
#向前引入法变量选择结果
```

数据输出

```
Start:  AIC=68.15
y ~ x1 + x2 + x3 + x4
```

4.4 回归变量的选择方法

R语言代码

```
fm.step=step(fm,direction="backward")
#向后剔除法变量选择结果
```

数据输出

```
Start:  AIC=68.15
y ~ x1 + x2 + x3 + x4
       Df Sum of Sq  RSS  AIC
- x3    1     0.009  202   66
- x1    1         1  204   66
<none>               202   68
- x4    1       174  376   85
- x2    1      6433 6635  174

Step:  AIC=66.16
y ~ x1 + x2 + x4
       Df Sum of Sq  RSS  AIC
- x1    1         2  204   64
<none>               202   66
- x4    1       197  400   85
- x2    1      7382 7585  176

Step:  AIC=64.39
y ~ x2 + x4
       Df Sum of Sq    RSS    AIC
<none>                 204     64
- x4    1       549    753    103
- x2    1    367655 367859    295
```

多元统计分析及R语言建模 → 4.4

R语言代码

```
fm.step=step(fm,direction="both")
 #逐步筛选法变量选择结果
```

数据输出

```
Start:  AIC=68.15
y ~ x1 + x2 + x3 + x4
       Df Sum of Sq  RSS  AIC
- x3    1     0.009  202   66
- x1    1         1  204   66
<none>               202   68
- x4    1       174  376   85
- x2    1      6433 6635  174

Step:  AIC=66.16
y ~ x1 + x2 + x4
       Df Sum of Sq  RSS  AIC
- x1    1         2  204   64
<none>               202   66
+ x3    1     0.009  202   68
- x4    1       197  400   85
- x2    1      7382 7585  176

Step:  AIC=64.39
y ~ x2 + x4
       Df Sum of Sq    RSS    AIC
<none>                 204     64
+ x1    1         2    202     66
+ x3    1      0.18    204     66
- x4    1       549    753    103
- x2    1    367655 367859    295
```