大数据的统计学基础——第12周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关注炼数成金企业微信

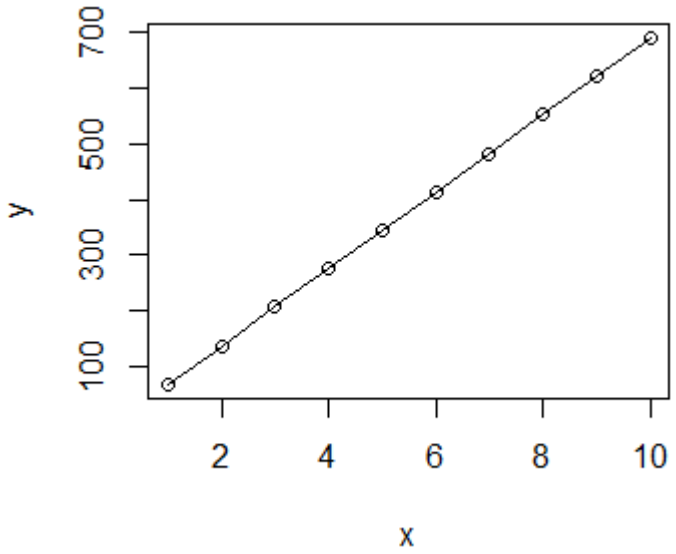
■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、 企业信息化、 数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

本周内容

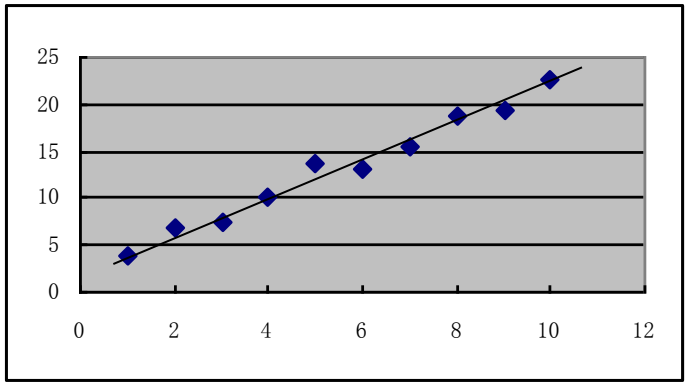
- 相关系数
- ——相关系数检验
- 一元线性回归
- ——回归系数检验
- ——残差图分析
- 多元线性回归
- 虚拟变量

关系

- 函数关系：确定性关系，y=3+10*x
- 相关关系：非确定性关系



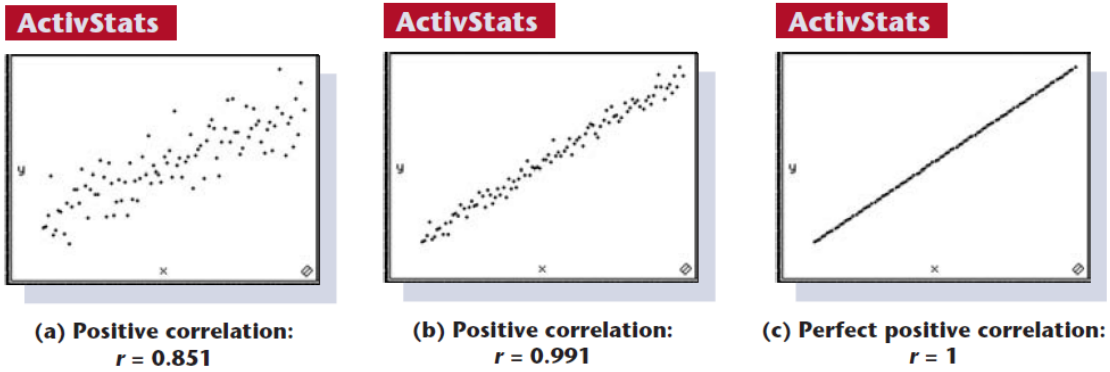
函数关系



相关系数

- 我们使用相关系数这一指标去衡量两个变量之间的线性相关程度。

正相关

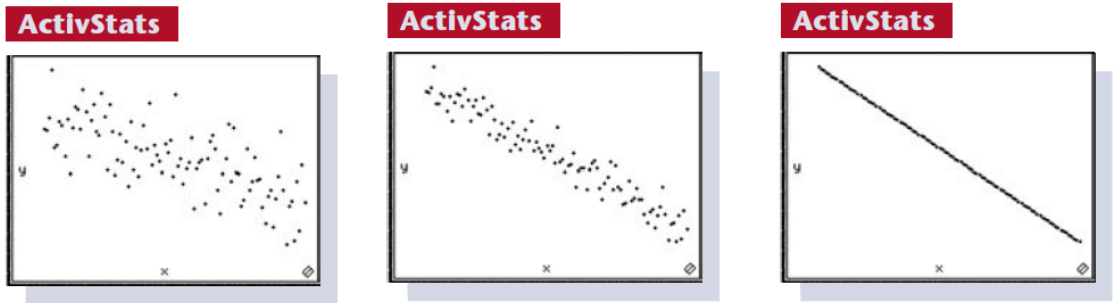


(a) Positive correlation: $r = 0.851$

(b) Positive correlation: $r = 0.991$

(c) Perfect positive correlation: $r = 1$

负相关



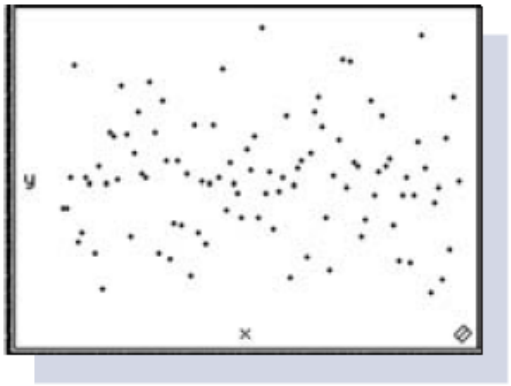
(d) Negative correlation: $r = -0.702$

(e) Negative correlation: $r = -0.965$

(f) Perfect negative correlation: $r = -1$

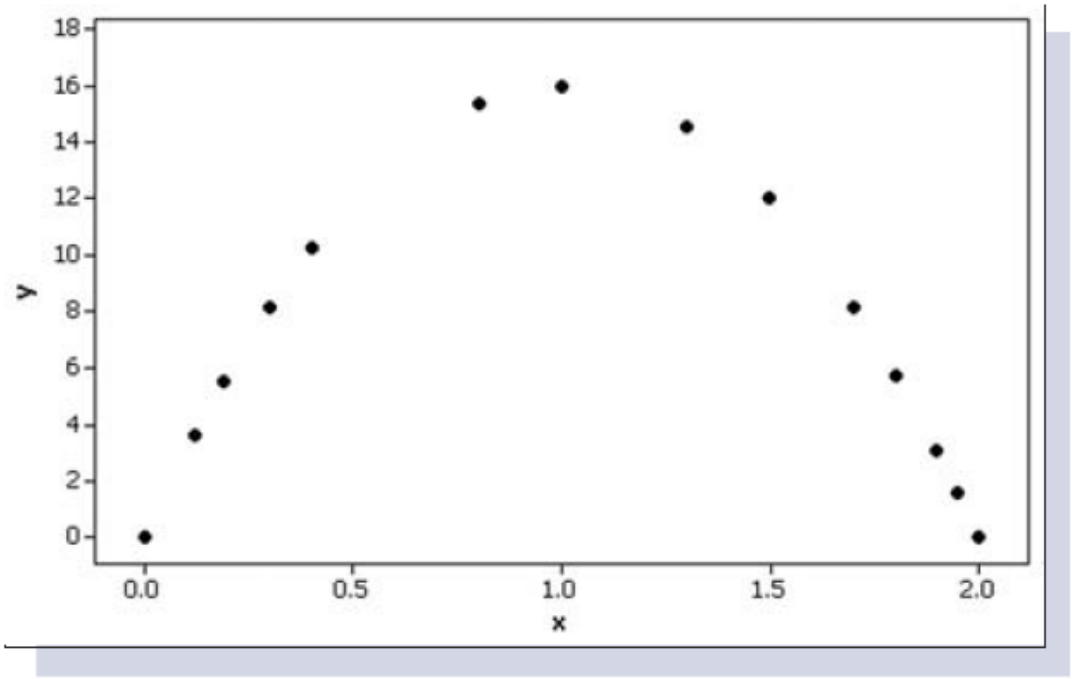
相关系数

ActivStats



(g) No correlation: $r = 0$

Minitab



(h) Nonlinear relationship: $r = -0.087$

相关系数

- 要求：
- 1. 成对数据(x,y)组成的样本是一个随机样本
- 2. 数据对(x,y)的散点图要呈现出近似线性相关性
- 3. 要把离群值排除

- 相关系数计算公式

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} = \frac{n(\sum XY)-(\sum X)(\sum Y)}{\sqrt{n(\sum X^2)-(\sum X)^2}\sqrt{n(\sum Y^2)-(\sum Y)^2}}$$

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

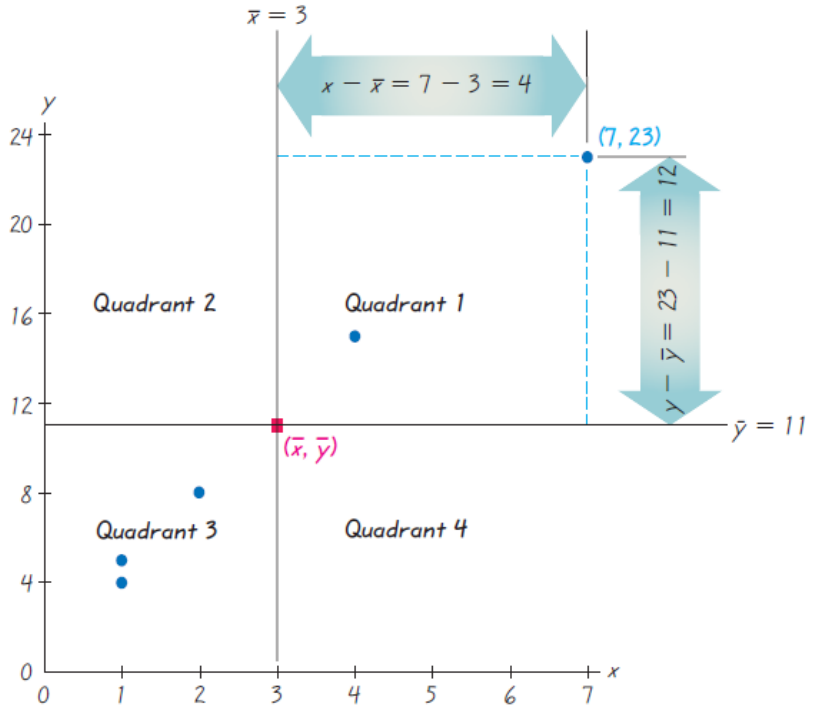
相关系数的等价表示

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$r = \frac{\Sigma\left[\frac{(x - \bar{x})}{s_x}\frac{(y - \bar{y})}{s_y}\right]}{n - 1}$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$



相关系数r的性质

- r的性质
- 1.r的范围是-1到1
- 2.如果某个变量的所有值都转化为一个不同的度量单位，r值不变。
- 3. r值不受x、y的选择影响。交换所有的x值与y值，r不变
- 4. r是线性相关性的强度的度量，不适用于非线性相关的关系
- 5. r非常容易受到离群值的影响，当有离群值存在的时候，r可能变得非常不一样

可解释变异r^2

- Y变异的来源
- 1. x变异造成的——可解释变异r^2
- 2. 除x外的因素变异造成的，如随机抽样引起的误差

- r^2放映了y变异中由x变异引起的变异所占总变异的比例，也就是r^2的值是由x和y之间的线性相关性所解释的y的变异比例。

常见错误

- 1. 误将相关关系认为是因果关系
- 2. 局部求平均数后再用于计算会使变异减少，相关性增大
- 3. 不存在线性相关关系，不意味着两个变量没有关系，可能会存在其他非线性关系

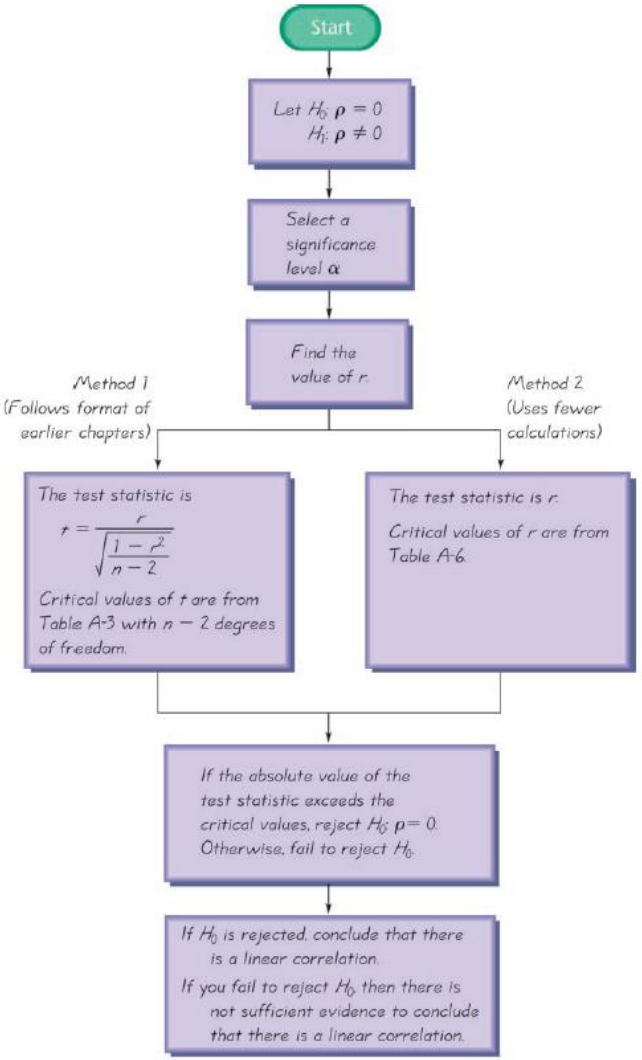
相关系数检验

- 要求：
- 1. 数据对(x,y)来自一个随机样本
- 2. (X,Y)服从二元正态分布
- 3. X与Y的散点图中的散点近似分布在一条直线附件

$$H_0: \rho = 0 \quad v.s \qquad H_1: \rho \neq 0$$

- 检验统计量：$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t(n-2)$

相关系数检验


Start
Let H_0: $\rho = 0$
H_1: $\rho \neq 0$
Select a significance level α
Find the value of r.
Method 1
(Follows format of earlier chapters)
Method 2
(Uses fewer calculations)
The test statistic is $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$
Critical values of t are from Table A-3 with $n - 2$ degrees of freedom.
The test statistic is r.
Critical values of r are from Table A-6.
If the absolute value of the test statistic exceeds the critical values, reject H_0: $\rho = 0$. Otherwise, fail to reject H_0.
If H_0 is rejected, conclude that there is a linear correlation.
If you fail to reject H_0, then there is not sufficient evidence to conclude that there is a linear correlation.

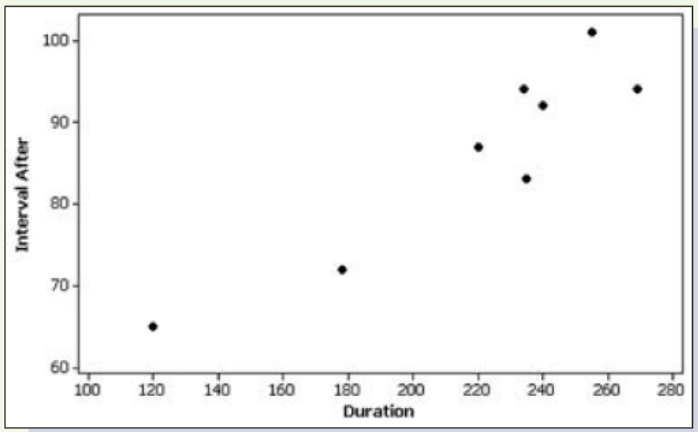
例子——old faithful

- 检验duration和interval after这两个变量的相关系数是否为0

Table 10-1 Eruptions of the Old Faithful Geyser

Duration	240	120	178	234	235	269	255	220
Interval Before	98	90	92	98	93	105	81	108
Interval After	92	65	72	94	83	94	101	87
Height	140	110	125	120	140	120	125	150

Minitab



(a)

例子——old faithful

- $H_0: \rho = 0 \quad v.s \quad H_1: \rho \neq 0$
- 设α=0.05
- $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.926}{\sqrt{\frac{1-0.926*0.926}{8-2}}} = 6.008$
- 方法1：查表可得，临界值为 $t = \pm 2.447$，故检验统计量落在拒绝域中，从而拒绝零假设
- 方法2：P(|t|>6.008)=2(1-P(t<=6.008)<0.05

```
> 2*(1-pt(6.008,6))
[1] 0.0009578141
```

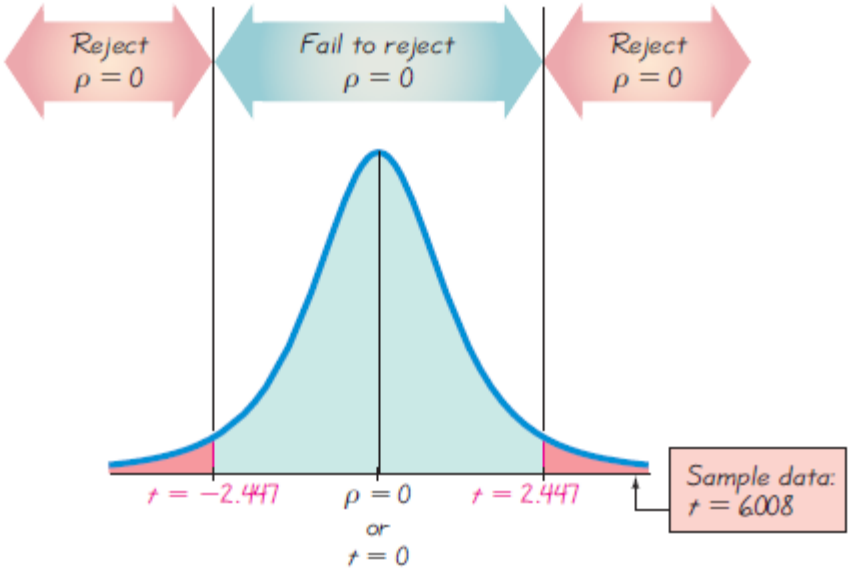


Figure 10-4 Testing H_0: $\rho = 0$ with Method 1

例子——old faithful

- 方法3：可以直接根据r的值查相应的临界值，0.707
- 0.927>0.707,故拒绝零假设

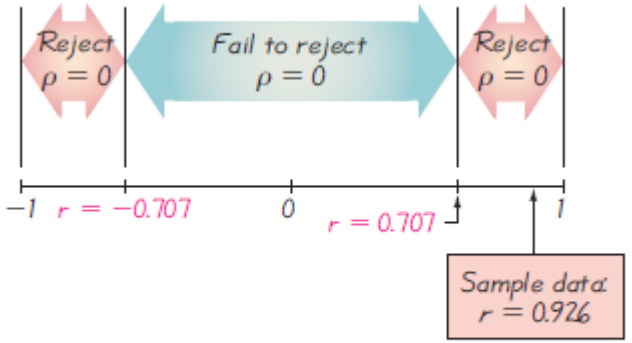
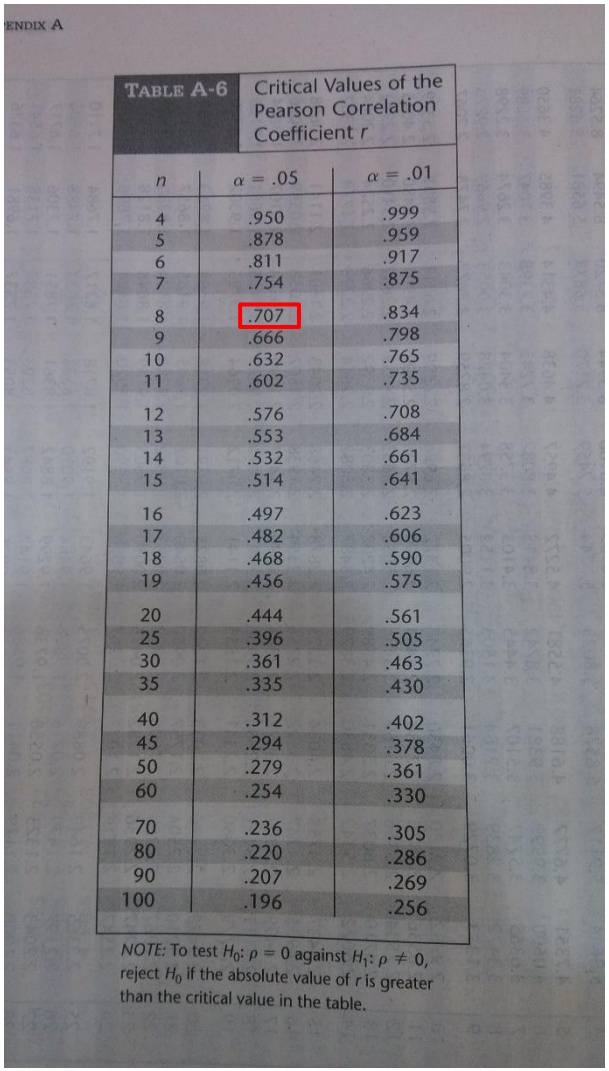


Figure 10-5 Testing H_0: $\rho = 0$ with Method 2



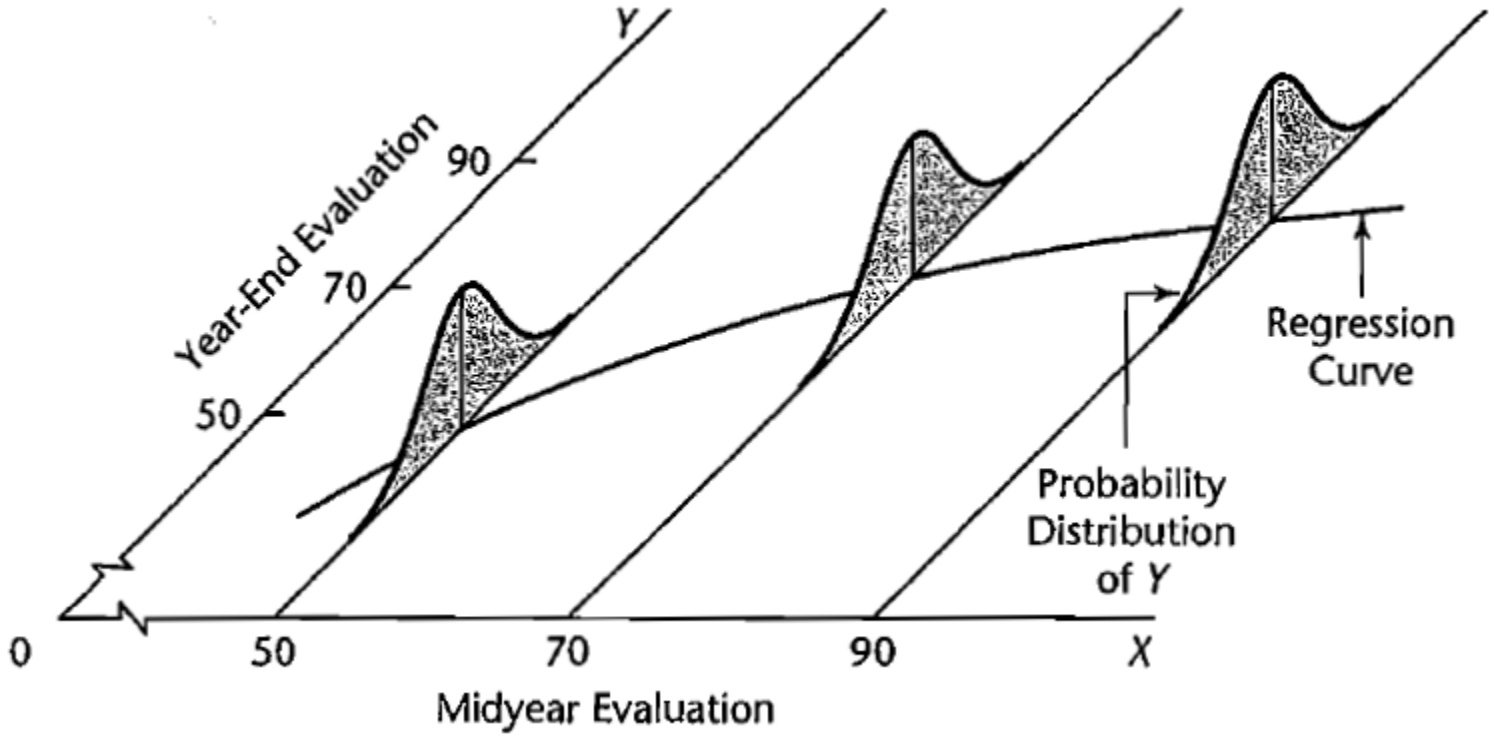
ENDIX A

TABLE A-6 Critical Values of the Pearson Correlation Coefficient *r*

n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

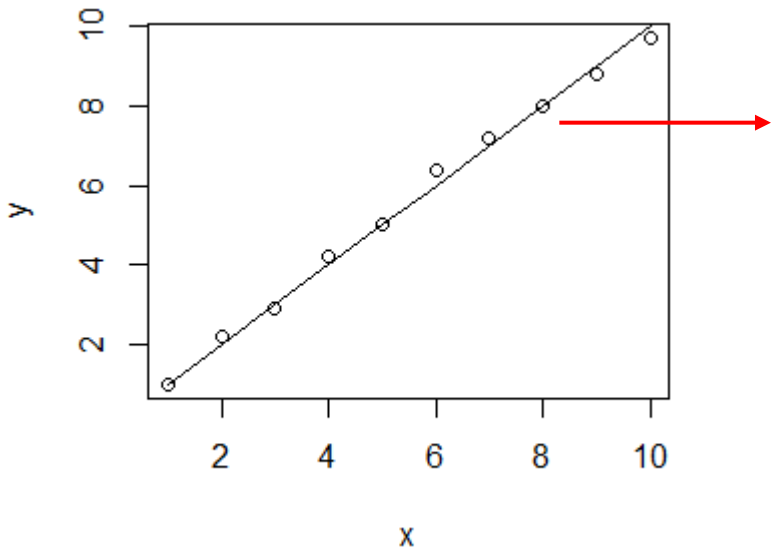
NOTE: To test H_0: $\rho = 0$ against H_1: $\rho \neq 0$, reject H_0 if the absolute value of *r* is greater than the critical value in the table.

回归模型


Y
Year-End Evaluation
90
70
50
0
50
70
90
X
Midyear Evaluation
Regression Curve
Probability Distribution of Y

一元线性回归模型

- 若X与Y之间存在着较强的相关关系，则我们有Y≈α+βX
- 若α与β的值已知，则给出相应的X值，我们可以根据Y≈α+βX得到相应的Y的预测值



回归直线方程：y=x

一元线性回归模型

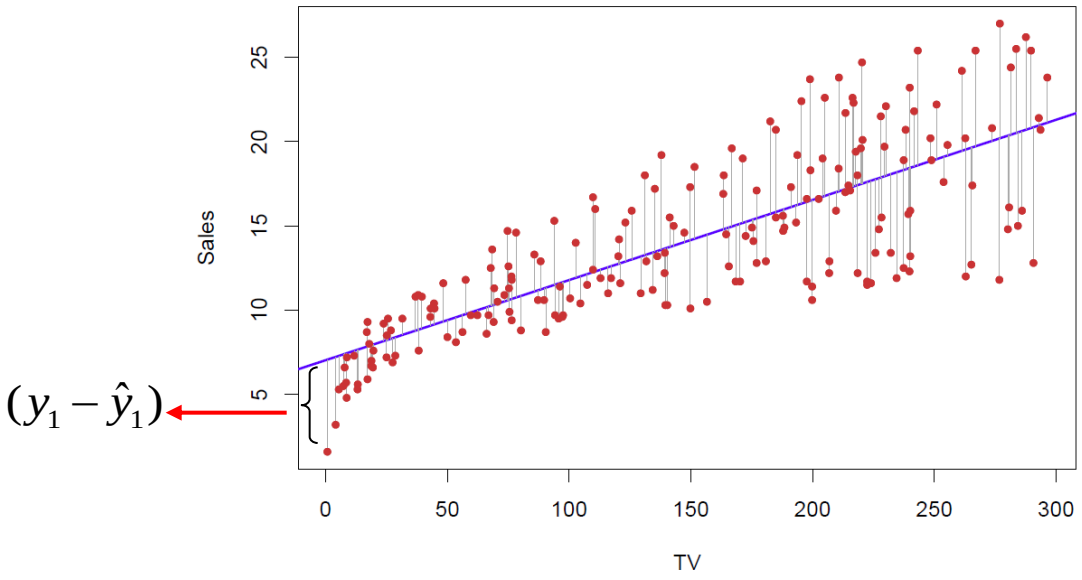
- 要求：
- 1. 数据对(x,y)来自一个随机样本
- 2. 只对线性相关的数据进行考察
- 3. 排除离群值

- 对于两个随机变量X与Y，定义两者的回归方程，其中ε表示的是有随机性引起的误差，

$$Y = \alpha + \beta X + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

如何确定参数

- 使用平方误差和衡量预测值与真实值的差距
- 平方误差真实值y，预测值 $\hat{y}$，则误差平方就是 $(y - \hat{y})^2$
- 寻找合适的参数，使得误差平方和 $\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ 最小。



如何确定参数

- 最小二乘法：

$$SSR = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} [y_i - (\alpha + \beta x_i)]^2$$

- SSR其实是关于α与β的函数，分别对α与β求偏导并令偏导等于0，就可以得出α与β的值

$$\beta = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$$

$$\alpha = \bar{Y} - \beta \bar{X}$$

如何确定参数

- 由于总体未知，采用样本值估计：

$$b = \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \hat{\alpha} = \bar{y} - b\bar{x}$$

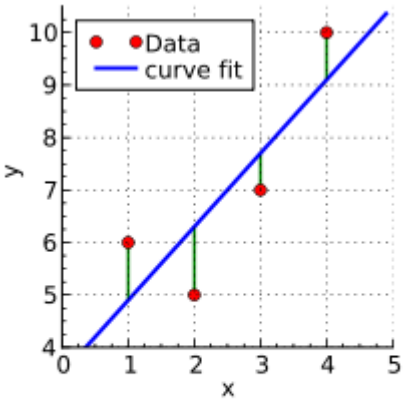
- 从而，对于每个xi，我们可以通过 $\hat{y}_i = a + bx_i$ 预测相应的y值

例子

- x=c(1,2,3,4)，y=c(6,5,7,10)。构建y关于x的回归方程y=α+βx
- 使用最小二乘法求解参数：

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=1.4$$

$$a=\bar{y}-b\bar{x}=3.5$$



数据点（红色）、使用最小二乘法求得的最佳解（蓝色）、误差（绿色）。

- 得到y=3.5+1.4x
- 如果有新的点x=2.5，则我们预测相应的y值为3.5+1.4*2.5=7

回归系数显著性检验

- 回归方程：$Y_i = \alpha + \beta X_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

$$b = \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- b的性质：
- $E(b) = \beta$
- $V(b) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

回归系数显著性检验

- 假设检验：$H_0: \beta = 0 \ \ v.s \ \ \ H_1: \beta \neq 0$
- $\frac{b-\beta}{\sqrt{\sigma^2/\sum(X_2-\bar{X})^2}} \sim N(0,1)$
- b的样本方差：$s^2(b) = \frac{MSE}{\sum(X_2-\bar{X})^2}$
- 其中MSE为总体方差的无偏估计量，$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{\sum(Y_i-\hat{Y}_i)^2}{n-2}$
- 故检验统计量：$\frac{b}{\sqrt{\mathrm{MSE}/\sum(X_2-\bar{X})^2}} \sim t(n-2)$

回归系数显著性检验

- 关于截距项： $a = \hat{\alpha} = \bar{y} - b\bar{x}$

$$E(a) = E(\bar{Y} - \beta\bar{X}) = \alpha$$

$$V(a) = V(\bar{Y} - \beta\bar{X}) = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2})$$

$$s^2(a) = MSE(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2})$$

$$s(a) = \sqrt{MSE(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2})}$$

$$\frac{a - \alpha}{\sqrt{V(a)}} \sim N(0,1), \frac{a - \alpha}{s(a)} \sim t(n - 2)$$

例子

构建身高~体重线性回归模型

x=c(171,175,159,155,152,158,
154,164,168,166,159,164)

y=c(57,64,41,38,35,44,41,51,5
7,49,47,46)

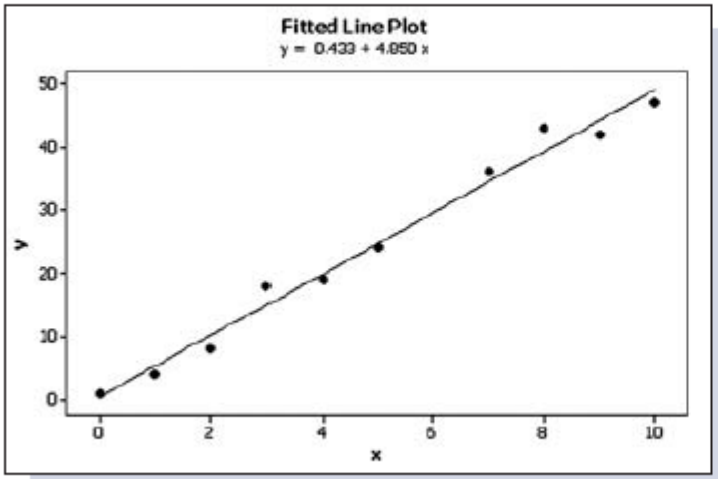
summary(lm(y~x))

```
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

   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
```

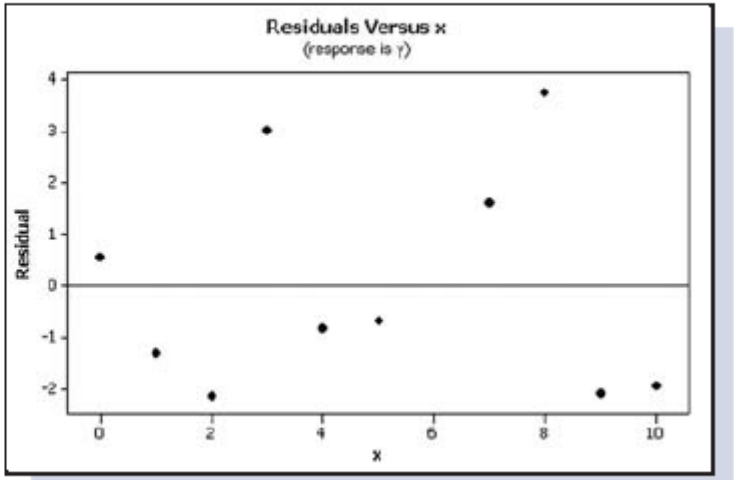
回归诊断——残差图

- 残差e：观察值-预测值=$y - \hat{y}$
- 残差图：x与残差e的散点图

Minitab



Minitab

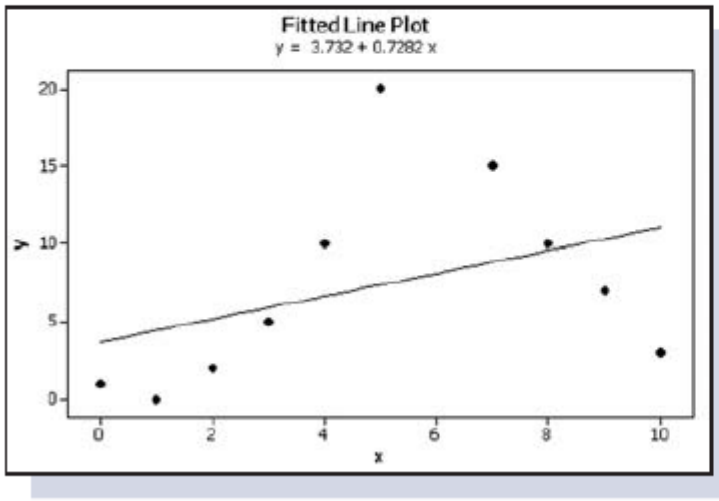


回归诊断——残差图

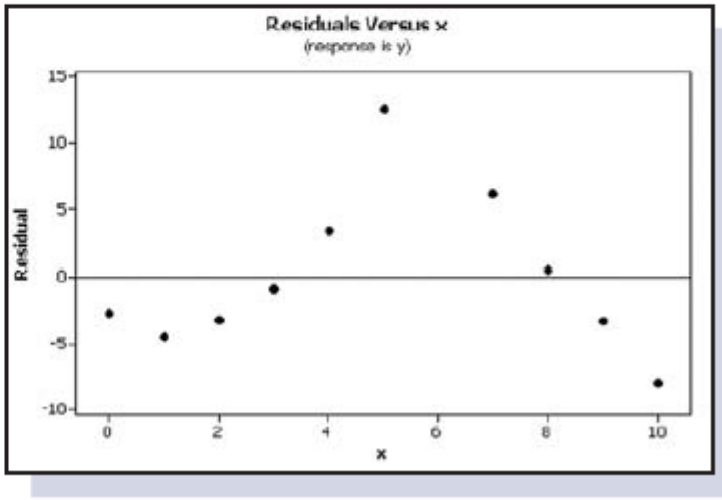
The scatterplot shows that the association is not linear.

The residual plot reveals a distinct pattern.

Minitab



Minitab

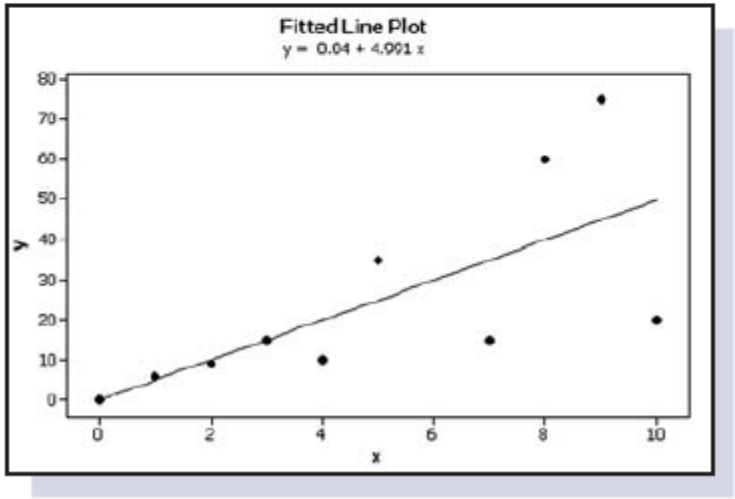


回归诊断——残差图

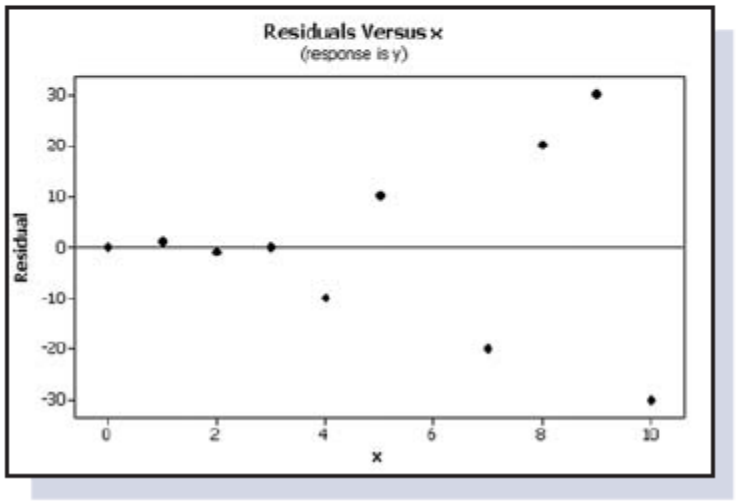
The scatterplot shows increasing variation of points away from the regression line

The residual plot reveals this pattern: Going from left to right, the points show more spread. (This is contrary to the requirement that for the different values of x, the distributions of y values have the same variance.)

Minitab



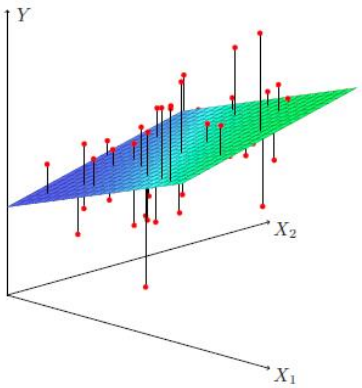
Minitab



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
- $\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ 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
- 使用矩阵方法表示更简练

R^2

- R^2：复判定系数（**multiple coefficient of determination**），衡量多元线性回归方程对数据的拟合程度。越接近1，拟合效果越好，相反，越接近0，拟合效果越差
- 缺点：增加自变量， R^2会增大
- 调整的复判定系数（ **adjusted coefficient of determination** ）

$$\text{adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

其中，n为样本容量，k为自变量个数

例子

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

	row.names	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
1	Courtelary	80.2	17.0	15	12	9.96	22.2
2	Delemont	83.1	45.1	6	9	84.84	22.2
3	Franches-Mnt	92.5	39.7	5	5	93.40	20.2
4	Moutier	85.8	36.5	12	7	33.77	20.3
5	Neuveville	76.9	43.5	17	15	5.16	20.6
6	Porrentruy	76.1	35.3	9	7	90.57	26.6
7	Broye	83.8	70.2	16	7	92.85	23.6
8	Glane	92.4	67.8	14	8	97.16	24.9
9	Gruyere	82.4	53.3	12	7	97.67	21.0
10	Sarine	82.9	45.2	16	13	91.38	24.4
11	Veveyse	87.1	64.5	14	6	98.61	24.5
12	Aigle	64.1	62.0	21	12	8.52	16.5
13	Aubonne	66.9	67.5	14	7	2.27	19.1
14	Avenches	68.9	60.7	19	12	4.43	22.7

例子

```
> swiss.lm=lm(Fertility~.,data=swiss)
> summary(swiss.lm)

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
```

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

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
- $Y=\beta0+\beta1*chas+\beta2*lstat=\begin{cases}\beta0+\beta1+\beta2*lstat，chas=1\\ \beta0+\beta2*lstat，chas=0\end{cases}$

```
lm1=lm(medv~lstat+chas,data=Boston)
lm2=lm(medv~lstat,data=Boston)
```

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间