

向量内积(点乘/数量积)

向量是由n个实数组成的一个n行1列 ($n \times 1$) 或一个1行n列 ($1 \times n$) 的有序数组

向量的内积 (点乘/数量积): 对两个向量执行点乘运算, 就是对这两个向量对应位一一相乘之后求和的操作, 点乘的结果是一个标量。

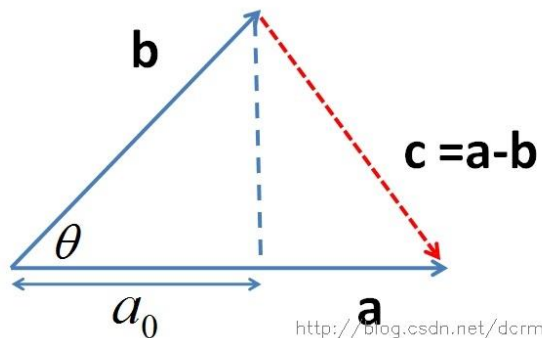
$$a = [a_1, a_2, \dots, a_n] \quad b = [b_1, b_2, \dots, b_n] \longrightarrow a \bullet b = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

要求一维向量a和向量b的行列数相同。

点乘几何意义

用来表征或计算两个向量之间的夹角, 以及在b向量在a向量方向上的投影, 有公式:

$$a \bullet b = |a| |b| \cos \theta$$



向量外积(叉乘/向量积/叉积)

向量外积的运算结果是一个向量而不是一个标量。并且两个向量的叉积与这两个向量组成的坐标平面垂直。

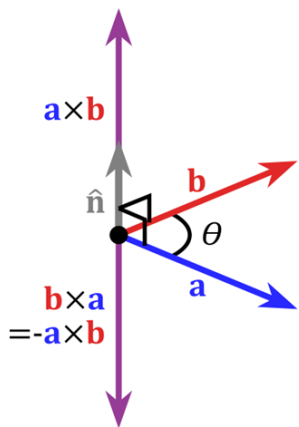
$$\begin{aligned} a &= (x_1, y_1, z_1) \\ b &= (x_2, y_2, z_2) \end{aligned} \longrightarrow a \times b = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix} = (y_1 z_2 - y_2 z_1) \mathbf{i} - (x_1 z_2 - x_2 z_1) \mathbf{j} + (x_1 y_2 - x_2 y_1) \mathbf{k}$$

$\mathbf{i} = (1, 0, 0) \quad \mathbf{j} = (0, 1, 0) \quad \mathbf{k} = (0, 0, 1)$

$$a \times b = (y_1 z_2 - y_2 z_1, -(x_1 z_2 - x_2 z_1), x_1 y_2 - x_2 y_1)$$


叉乘几何意义

在三维几何中，向量a和向量b的叉乘结果是一个向量(法向量)，该向量垂直于a和b向量构成的平面。在3D图像学中，叉乘的概念非常有用，可以通过两个向量的叉乘，生成第三个垂直于a, b的法向量，从而构建X、Y、Z坐标系。如下图所示：



在二维空间中，叉乘还有另外一个几何意义就是： $\mathbf{a} \times \mathbf{b}$ 等于由向量a和向量b构成的平行四边形的面积。

协方差(covariance)

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$


向量内积

标准差和方差一般是用来描述一维数据的

协方差只能处理二维问题

维数多了就需要计算多个协方差，会使用矩阵来组织这些数据。也就是协方差矩阵

假设数据集有三个维度，则协方差矩阵为：

$$C = \begin{pmatrix} \text{cov}(x, x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{cov}(y, y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{cov}(z, z) \end{pmatrix}$$

可见，协方差矩阵是一个对称的矩阵，而且对角线是各个维度上的方差。

统计相关系数

相关系数：考察两个事物（在数据里我们称之为变量）之间的相关程度。

如果有两个变量：X、Y，最终计算出的相关系数的含义可以有如下理解：

- (1)、当相关系数为0时，X和Y两变量无关系。
- (2)、当X的值增大（减小），Y值增大（减小），两个变量为正相关，相关系数在0.00与1.00之间。
- (3)、当X的值增大（减小），Y值减小（增大），两个变量为负相关，相关系数在-1.00与0.00之间。

相关系数的绝对值越大，相关性越强，相关系数越接近于1或-1，相关度越强，相关系数越接近于0，相关度越弱。

通常情况下通过以下相关系数取值范围判断变量的相关强度：

- 0.8-1.0 极强相关
- 0.6-0.8 强相关
- 0.4-0.6 中等程度相关
- 0.2-0.4 弱相关
- 0.0-0.2 极弱相关或无相关

Pearson (皮尔逊) 相关系数

- 皮尔逊相关也称为积差相关（或积矩相关），是英国统计学家皮尔逊于20世纪提出的一种计算直线相关的方法。
- 假设有两个变量X、Y，那么两变量间的皮尔逊相关系数可通过以下公式计算：

- 公式一：
$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$
； $\xrightarrow{\text{协方差/标准差的积}}$ $\xrightarrow{\text{向量夹角的余弦值}(\cos\theta)}$

- 公式二：
$$\rho_{x,y} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

- 公式三：
$$\rho_{x,y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

- 公式四：
$$\rho_{x,y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

当两个变量的标准差都不为零时，相关系数才有定义，皮尔逊相关系数适用于：

- (1)、两个变量之间是线性关系，都是连续数据。
- (2)、两个变量的总体是正态分布，或接近正态的单峰分布。
- (3)、两个变量的观测值是成对的，每对观测值之间相互独立。

Spearman Rank 相关系数

Kendall Rank 相关系数

拟合，一般选择直线或者次数比较低得曲线

- 回归分析就是利用样本（已知数据），产生拟合方程，从而（对未知数据）进行预测

- 用途：预测，判别合理性

- 例子：利用身高预测体重；利用广告费用预测商品销售额；等等.

一个自变量 自变量多个，一次方程，是个曲面，高维空间中的超平面

- 线性回归分析：一元线性；多元线性；广义线性

经过转换后为线性模型

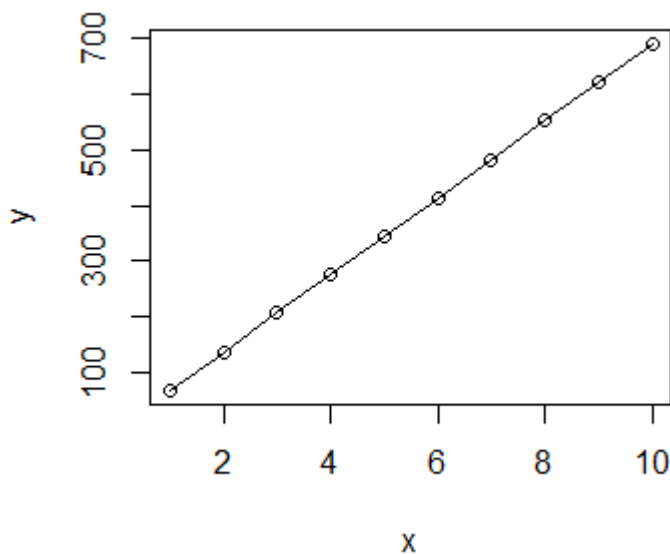
- 非线性回归分析

- 困难：选定变量（多元），避免多重共线性，观察拟合方程，避免过度拟合，检验模

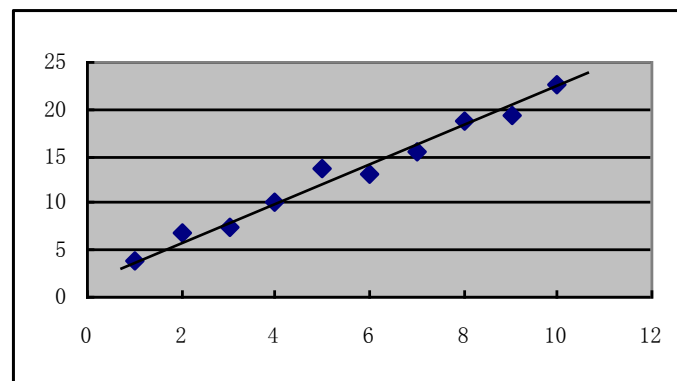
型是否合理 困难：选定变量(多元)，降维是回归模型中的难点。世界规律都是用很简单的东西
多重共线性：有些变量是打酱油的，怎么判断，怎么去掉
怎样检验模型是否合理，需要一些检验手段。

自变量和因变量的关系有两种：

- **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



- 我们使用相关系数去衡量线性相关性的强弱。

r取值范围： $-1 < r < 1$

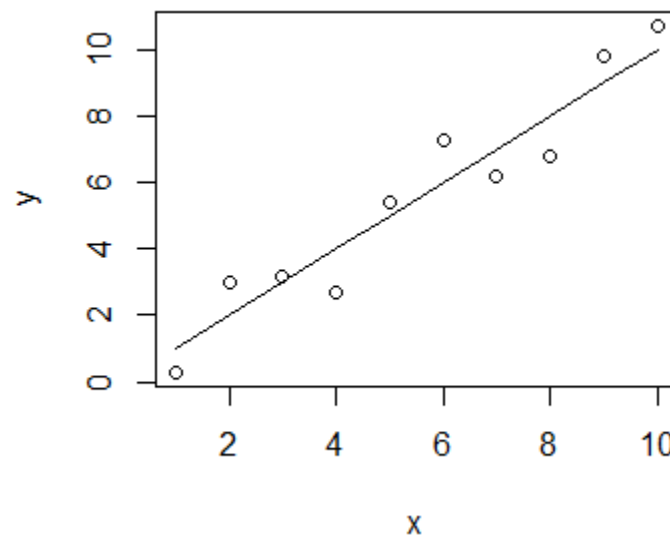
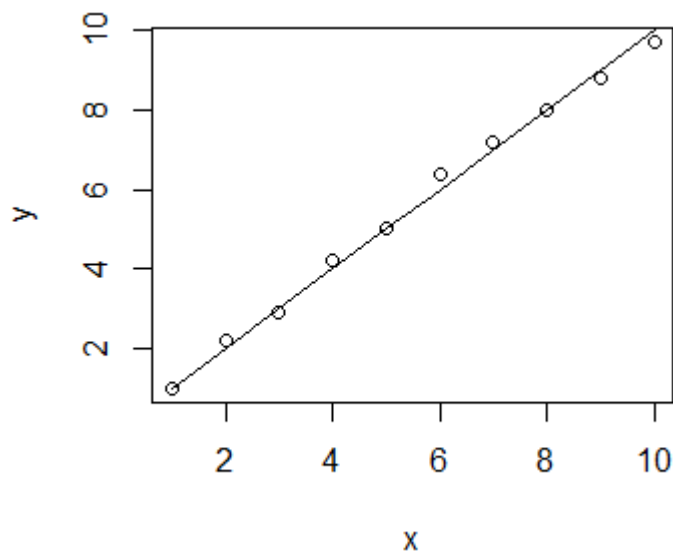
r越接近于1或-1，表明相关性越强。越适合用线性回归模型

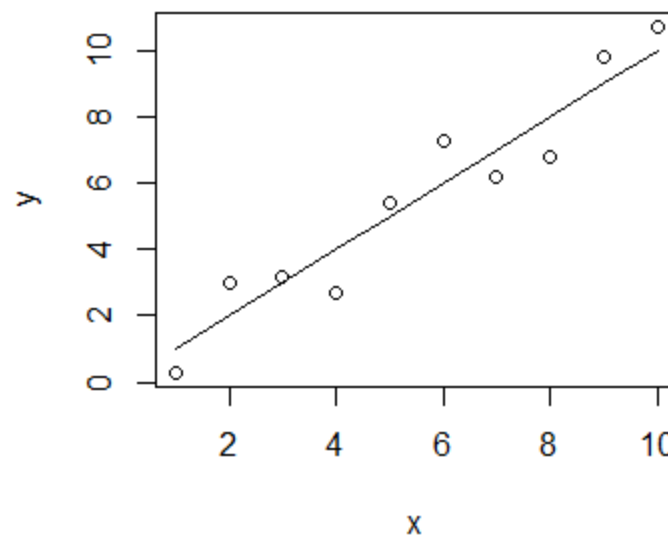
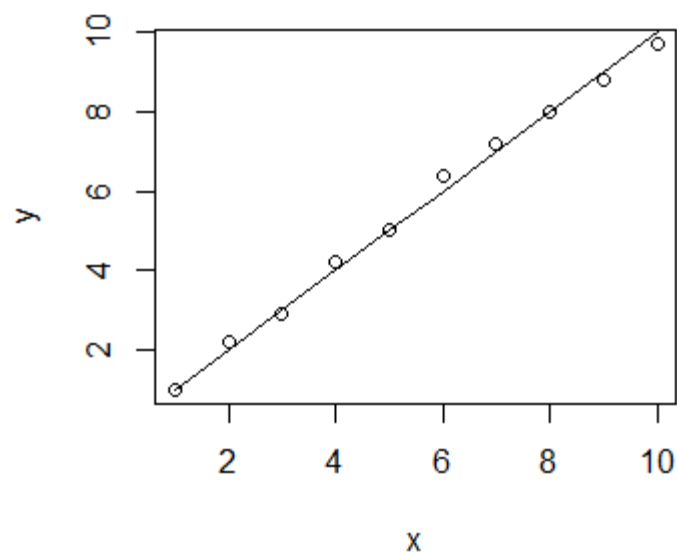
$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

X_i, Y_i : 第i个X, 第i个Y

\bar{X}, \bar{Y} : X, Y的平均值

Σ : 求和

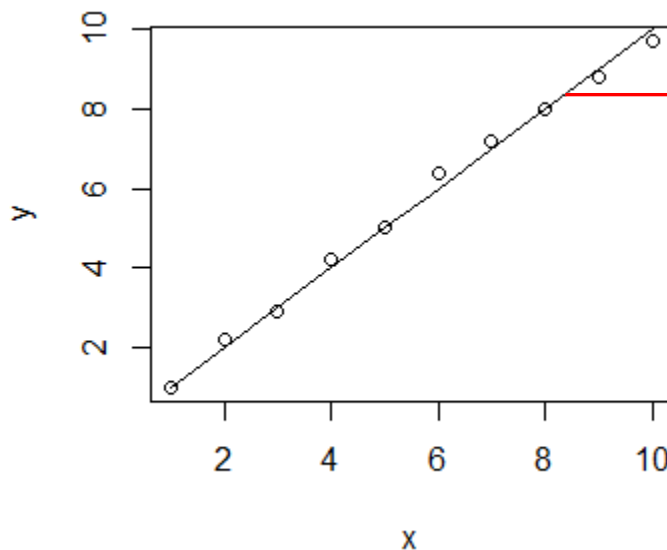




- 通过计算，左图中的相关系数为0.9930858，右图的相关系数为0.9573288

- 若X与Y之间存在着较强的相关关系，则我们有 $Y \approx \alpha + \beta X$
- 若 α 与 β 的值已知，则给出相应的X值，我们可以根据 $Y \approx \alpha + \beta X$ 得到相应的Y的预测值
- $\hat{y}_i = \alpha + \beta x_i$

也是Y的回归值，即
根据回归方程得出来的值

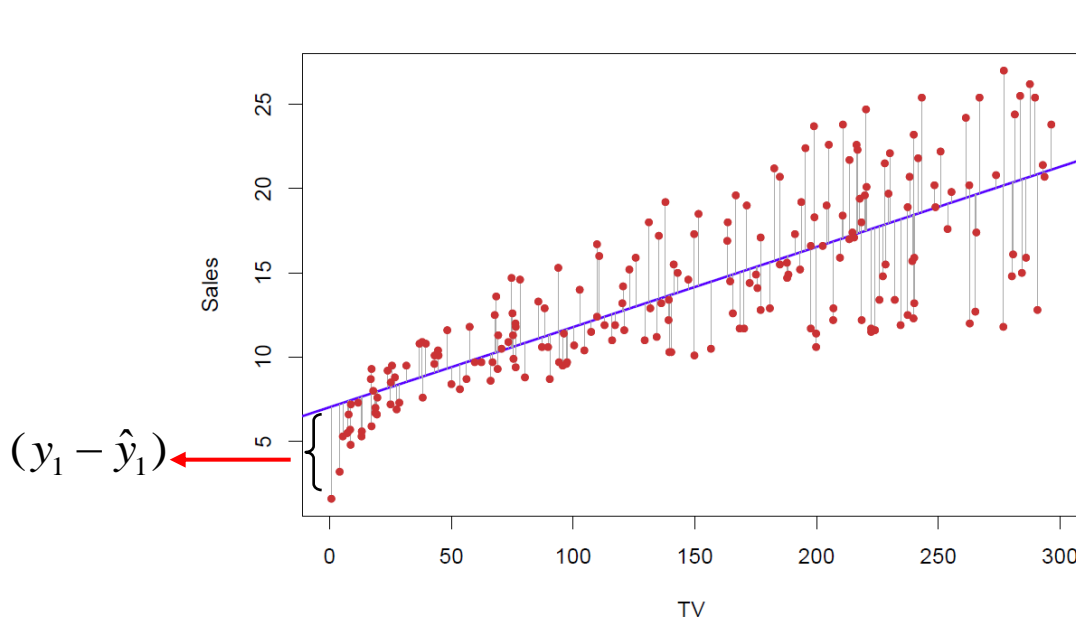


回归直线方程： $y=x$

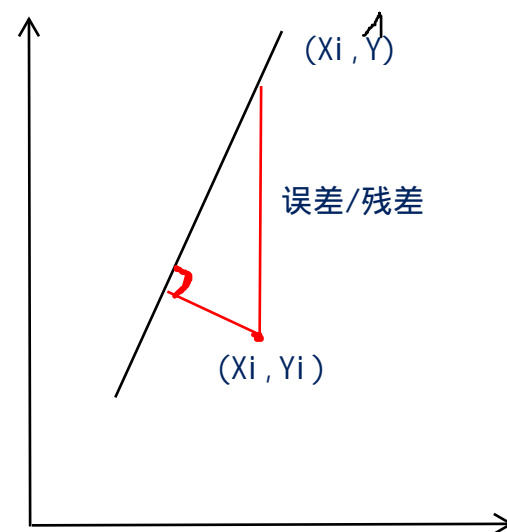
- $Y = \alpha + \beta X + \varepsilon$
- 截距项 α
- 斜率 β
- 误差项 ε
- 例子：商品销量s关于电视广告费用t的回归方程： $s=10+3.4*t$ （单位：万元）

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} ，则平方误差就是 $(y - \hat{y})^2$
- 寻找合适的参数，使得平方误差和 $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 最小。



一元线性回归产生的是一个平面



$$\begin{aligned} \text{误差/残差} &= (\hat{Y} - Y_i) \\ &= (bX_i + a - Y_i) \end{aligned}$$

- **最小二乘法**：哪个回归线效果最好呢：比较直观的做法，点到直线的距离，使得所有点的距离之和最小。问题是，距离涉及到开方，很难转换为极值。就改为垂直线，即平行于y轴，称为残差绝对值在数学里不好求极值，所以改为求平方

Residual Sum of Squares
残差/误差平方和
衡量预测值和真实值的差距

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

- RSS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

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

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

- 由于总体未知，采用样本值估计：

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

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

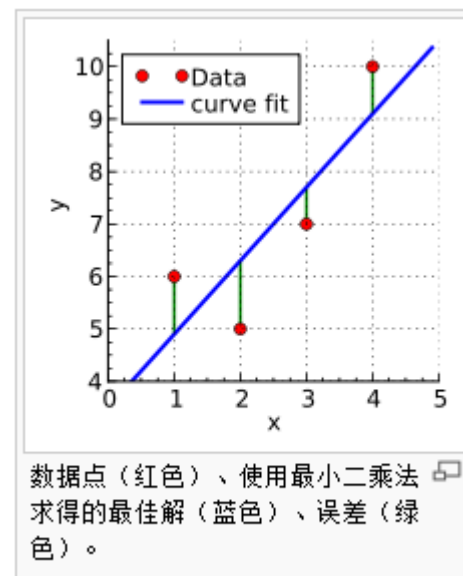
- 从而，对于每个 x_i ，我们可以通过 $\hat{y}_i = a + bx_i$ 预测相应的 y 值

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

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

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

- 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 $3.5+1.4*2.5=7$



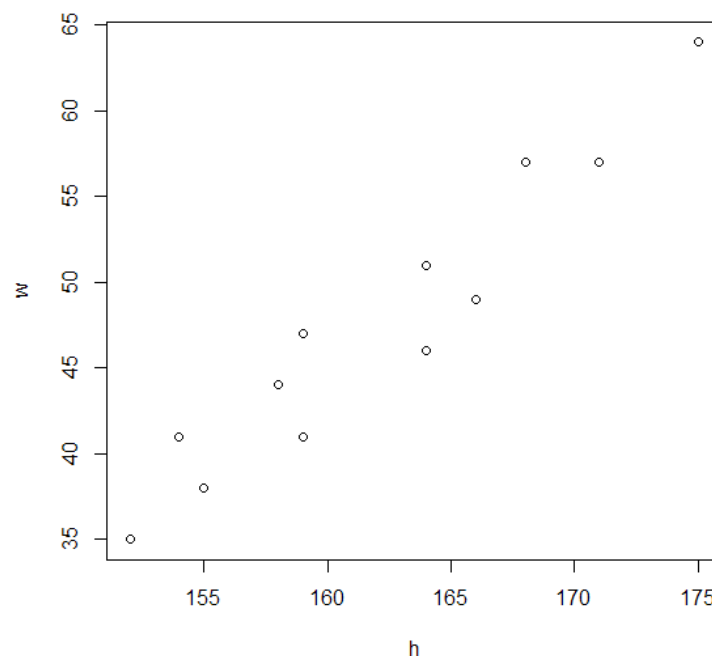
- 原理，最小二乘法
- 步骤：建立回归模型，求解回归模型中的参数，对回归模型进行检验
- 例子

数据：身高-体重

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

$w = c(57, 64, 41, 38, 35, 44, 41, 51, 57, 49, 47, 46)$

$\text{plot}(w \sim h + 1)$



自定义函数 lxy <-

```
function(x,y){n=length(x);sum(x*  
y)-sum(x)*sum(y)/n}
```

假设 $w = a + bh$

则有

```
> b=lxy(h,w)/lxy(h,h)
```

```
> a=mean(w)-b*mean(h)
```

```
> a
```

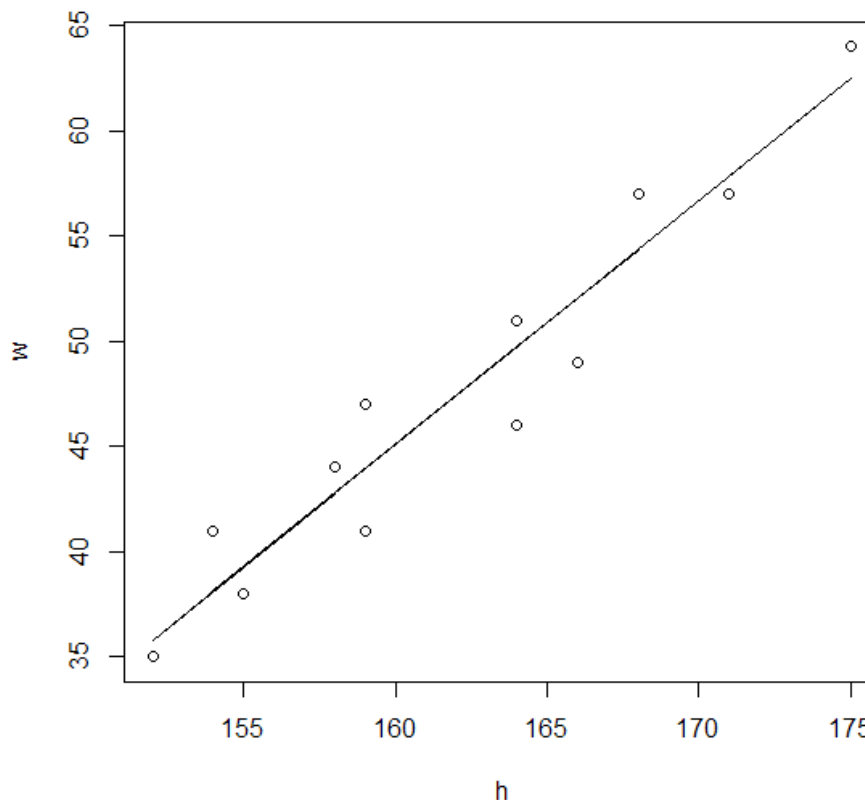
```
[1] -140.3644
```

```
> b
```

```
[1] 1.15906
```

作回归直线

```
lines(h,a+b*h)
```



- 回归系数的假设检验
- 建立线性模型

```
> a=lm(w~1+h)  lm 假定 y=ax+b, 后面的+1可以不写  
> a
```

```
Call:  
lm(formula = w ~ 1 + h)
```

```
Coefficients:  
(Intercept)              h  
    -140.364             1.159
```

■ 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:
```

假设检验的统计量t值

	Estimate	Std. Error	t value	Pr(> t)	Pr(> t) t以外的面积有多大，这个值越小越好。
截距 (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
```

残差的标准差误差

```
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
```

Multiple R-squared: 相关系数平方，越高表示相关性越好

Adjusted R-squared: 调整后的拟合优度，作用有限

p-value 整体的假设检验。不能说我有错。假设不对，回归模型是无效的

- 汇总数据的解释
- Residuals : ~~参差~~分析数据 残差
- 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

■ 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
```

■ 预测：一个身高185的人，体重大约是多少？

```
> a+b*185
```

```
[1] 74.0618
```

```
>
```


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)`

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
```

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		

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
```

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

- 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？

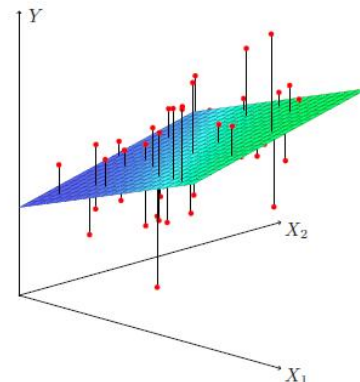
多元线性回归模型

自变量多个，一次方程，是个曲面，高维空间中的超平面

多元线性回归产生的是超平面

- 当Y值的影响因素不唯一时，采用多元线性回归模型

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



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

- 最小二乘法：

- 与一元回归方程的算法相似

求各个点到超平面的距离(残差/误差)的平方和RSS

- $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 是关于 β_i 的函数。分别对 β_i 求偏导并令偏导等于0，可以解出相应的 β_i 的值

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

	row.names	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
1	Courtellary	88.2	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

```
> 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

- 虚拟变量的定义
- 虚拟变量的作用
- 虚拟变量的设置

哑变量/虚拟变量(dummy variable)，如sex这个分类变量用两个哑变量表示：`isman`, `iswoman`

加法模型，哑变量用来调整截距：，如 $w=a+bh+c*isman$

乘法模型，哑变量用来调整斜率，如 $w=a+bh+c*isman*h$

混合模型，即影响截距和斜率上， $w=a+bh+c*isman+d*isWoman+e*isman*h+f*isWoman*h+g$

■ 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

- 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

- $$Y = \beta_0 + \beta_1 \text{chas} + \beta_2 \text{lstat} = \begin{cases} \beta_0 + \beta_1 + \beta_2 \text{lstat}, & \text{chas} = 1 \\ \beta_0 + \beta_2 \text{lstat}, & \text{chas} = 0 \end{cases}$$

- 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)

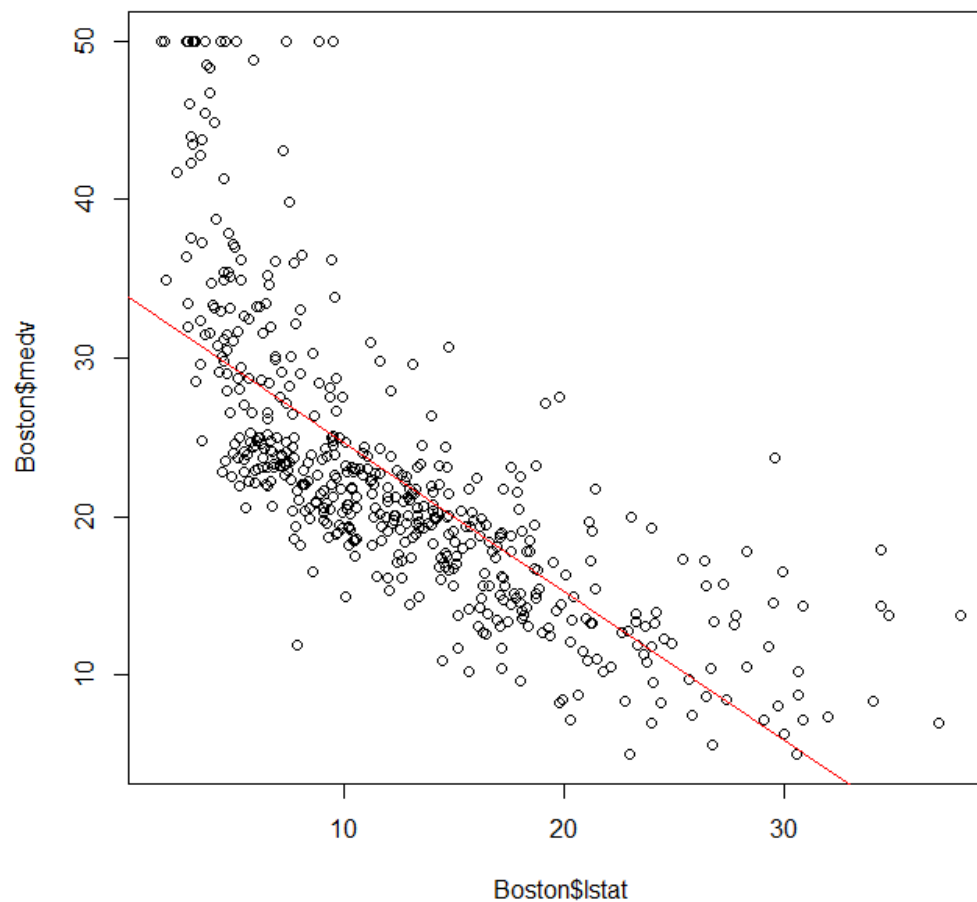
Residuals:
    Min       1Q   Median       3Q      Max
-14.782  -3.798  -1.286   1.769  24.870

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")
```




```
> lm.fit =lm(medv~., data= Boston ); summary (lm.fit )

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-15.595  -2.730  -0.518   1.777   26.199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn          4.642e-02  1.373e-02   3.382 0.000778 ***
indus       2.056e-02  6.150e-02   0.334 0.738288
chas       2.687e+00  8.616e-01   3.118 0.001925 **
nox        -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm         3.810e+00  4.179e-01   9.116 < 2e-16 ***
age         6.922e-04  1.321e-02   0.052 0.958229
dis        -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad         3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax        -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

- 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**

模型修正, 参见R-Modeling 324页

```
lm.new<-update(lm.sol, .~.+I(X2^2)) #I(X2^2) 表示X2的平方项
```

```
lm2.new<-update(lm.new, .~-X2) #去掉X2的一次项
```

```
lm3.new<-update(lm.new, .~.+X1*X2) #增加考虑X1和X2的一次项
```

说明：这个修正都是靠着分析师的经验和肉眼观察，碰出来的。统计的方法还需要用以下方法。

- 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

```
s=lm(Fertility~., data=swiss)
```

```
向前引入法：s1=step(s, direction="forward")
```

```
向后删除法：s1=step(s, direction="backward")
```

```
逐步筛选法：s1=step(s, direction="both")
```

■ step()函数

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？
手工回归, R-Modeling 334页
add1()
drop1()
- 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
```

- 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？

模型检验三要素：

正态分布

等方差

非多重共线性

- 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)
```

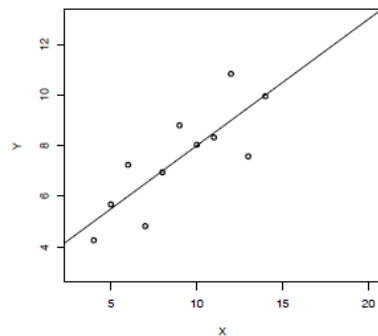
```
Shapiro-Wilk normality test
```

```
data:  x$x3
```

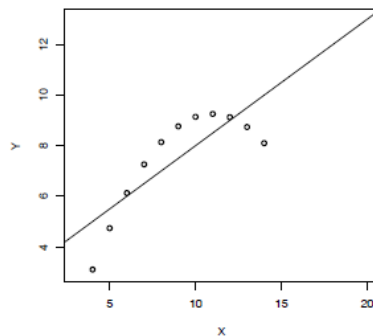
```
W = 0.9444, p-value = 0.0003618
```

散点图目测检验

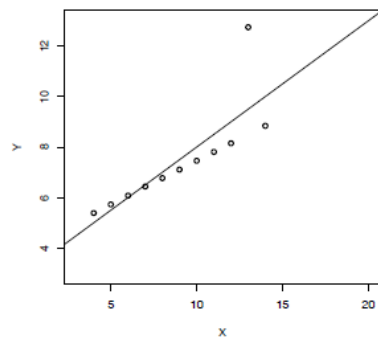
■ 薛毅书纸介质p284，例6.11



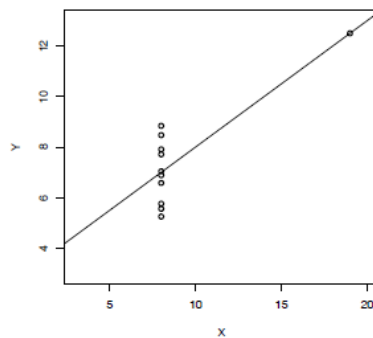
(a) 数据 1



(b) 数据 2



(c) 数据 3



(d) 数据 4

误差是否满足独立性、等方差（误差与y大小没有关系）
如果样本是正态分布的，残差residuals()也是正态分布的

- 残差计算函数residuals()
- 对残差作正态性检验
- 残差图

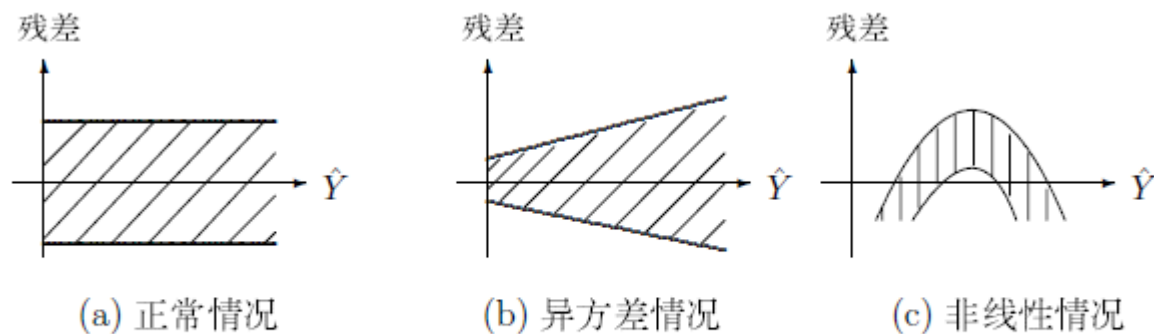
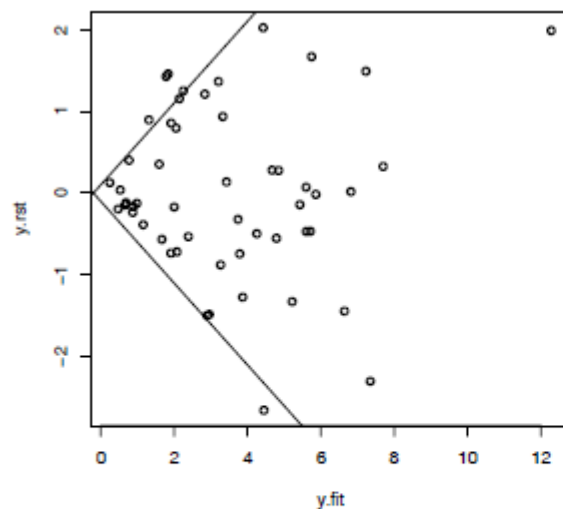
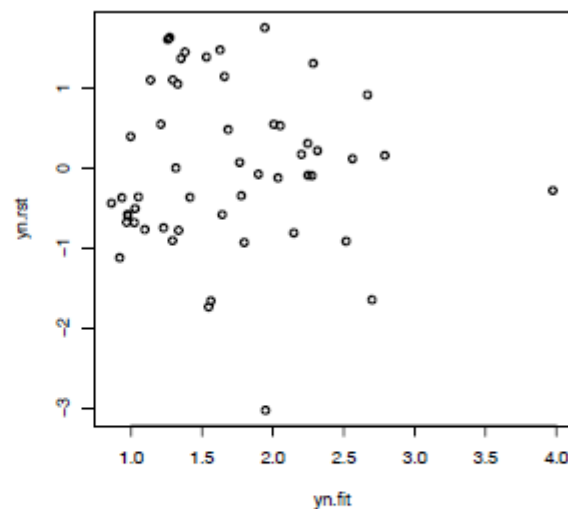


图 6.7: 回归值 \hat{Y} 与残差的散点图
回归值：经过回归方程算出来的值

■ 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

- 什么是多重共线性 自变量不是独立的
- 多重共线性对回归模型的影响 多重共线性存在，会导致求逆矩阵的结果非常不确定
- 利用计算特征根发现多重共线性
- Kappa()函数

Kappa值，希腊字母，把样本的数据乘以它的矩阵的转置，在求特征根，最大值除以最小值
 $k < 100$, 多重共线性程度小；
 $100 < k < 1000$, 中等程度或较强的多重共线性；
 $k > 1000$, 严重的多重共线性

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

此为非线性回归模型

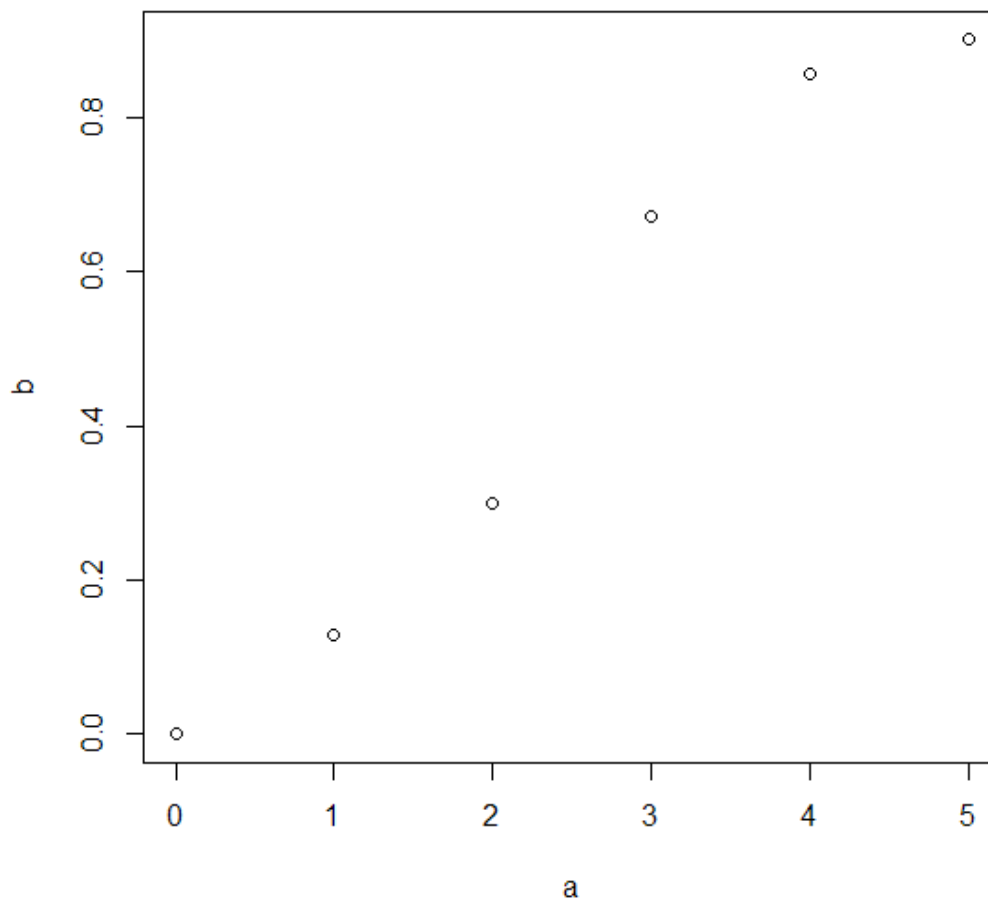
$$P = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 + \cdots + \beta_p X_p)}$$

β_0 为常数项或截距

$\beta_1, \beta_2, \dots, \beta_p$ 为logistic模型回归系数

$\exp(n)$: e的n次方。以e为底数的指数函数

S型曲线，统计学非常有名，叫logistic曲线



■ Logit变换 变换为线性回归模型

Logit模型
$$\text{logit}(P) = \ln \left(\frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

■ 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

Logit 模型和Logistic模型是一回事。

Logit模型和Logistic模型的区别：

- 1)Logit模型的左侧是Odds的对数，而Logistic模型的左侧是概率。
- 2)Logit模型的右侧是一个线性结构，而Logistic模型的右侧是非线性的。
- 3)二者可以相互转化。

- 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)
```

连接函数为二项分布


```
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
```

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

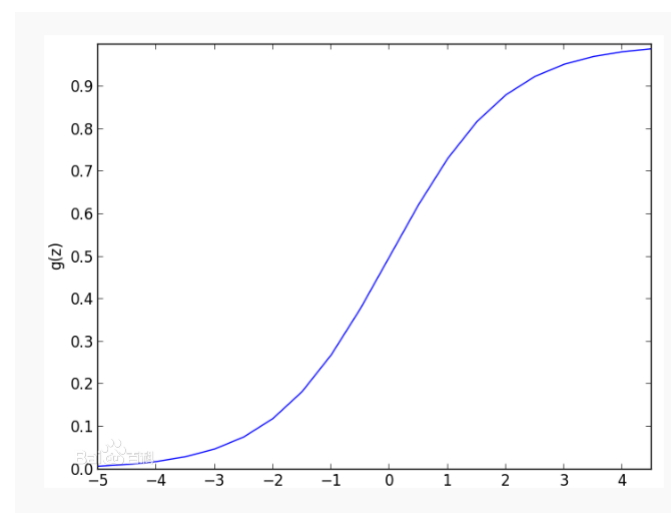
- 当要预测的y值不是连续的实数，而是定性变量，例如某个客户是否购买某件商品，这时线性回归模型不能直接应用。

- 为了让模型适用，我们对p做logistic变换，得到

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

- 其中p表示Y=1的概率

S型曲线，统计学非常有名，叫logistic曲线

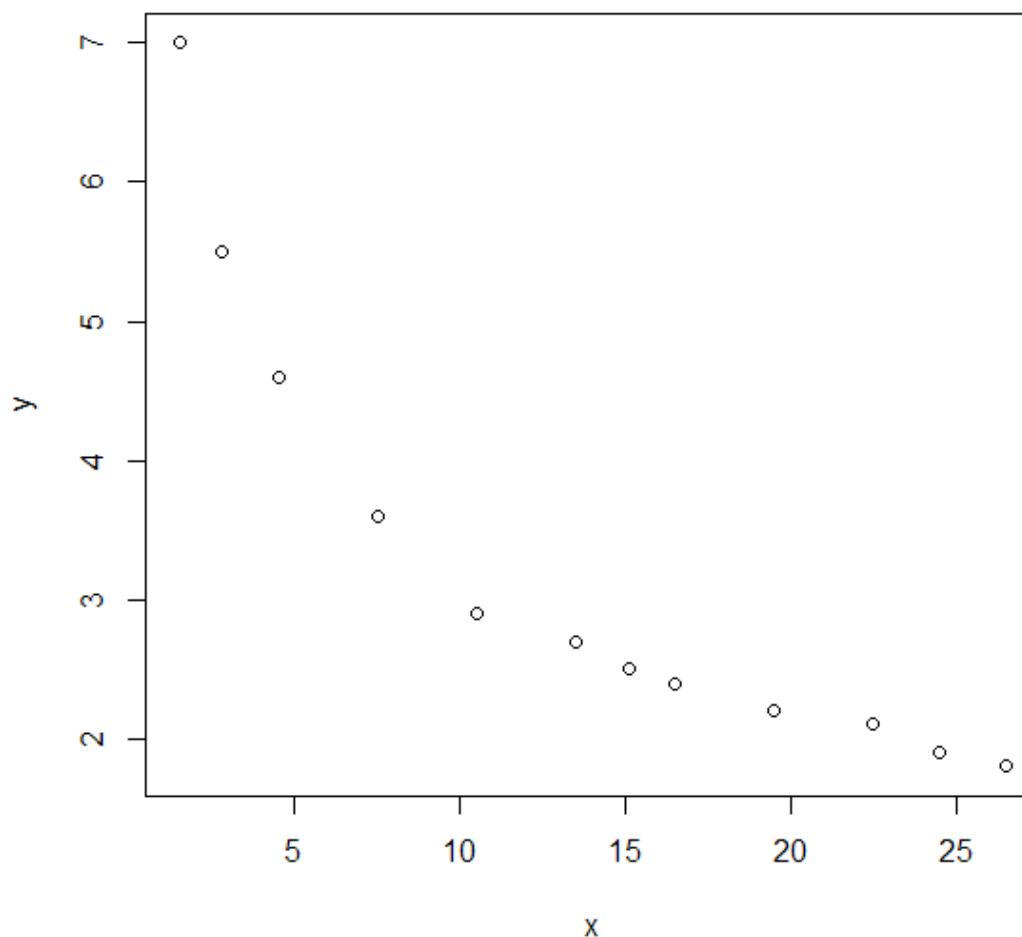


■ 例子：销售额x与流通费率y

```
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)
```



■ 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

$$y=a+bx+cx^2$$

$x_1=x$

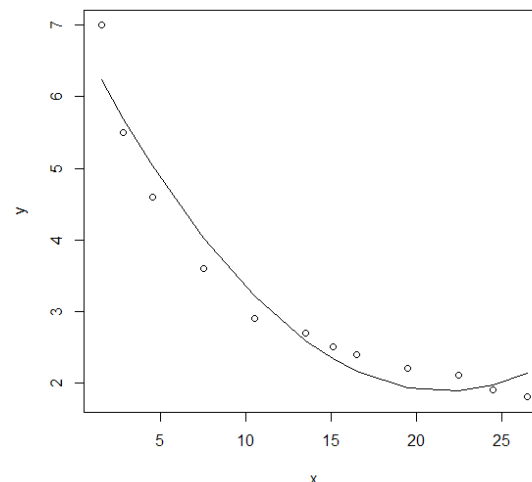
$x_2=x^2$

$\text{lm.2}=\text{lm}(y\sim x_1+x_2)$

$\text{summary}(\text{lm.2})$

$\text{plot}(x,y)$

$\text{lines}(x,\text{fitted}(\text{lm.2}))$



```
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
```

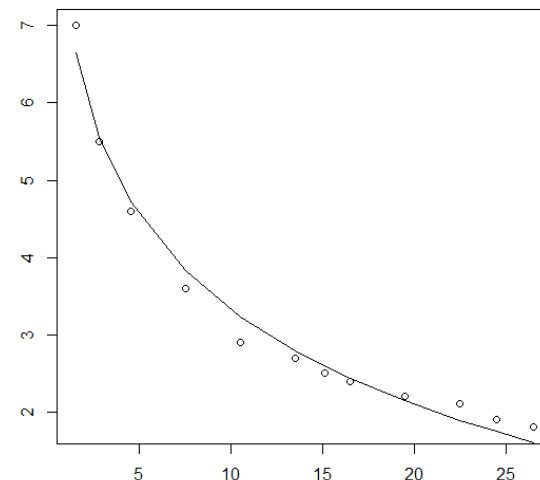
■ 对数法 , $y=a+b \log x$

`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:

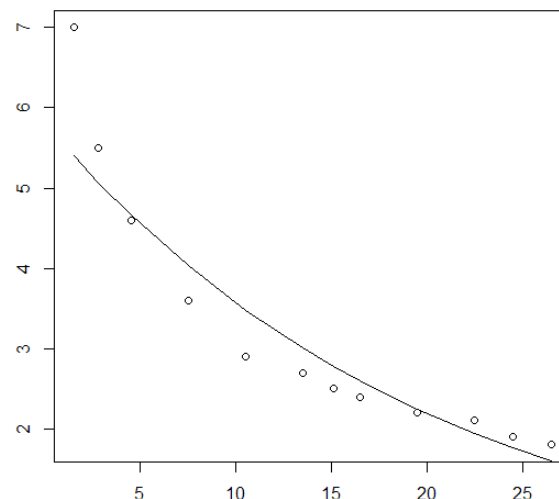
Min	1Q	Median	3Q	Max
-0.18246	-0.10664	-0.01670	0.08079	0.25946

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



■ 幂函数法 , $y=a x^b$

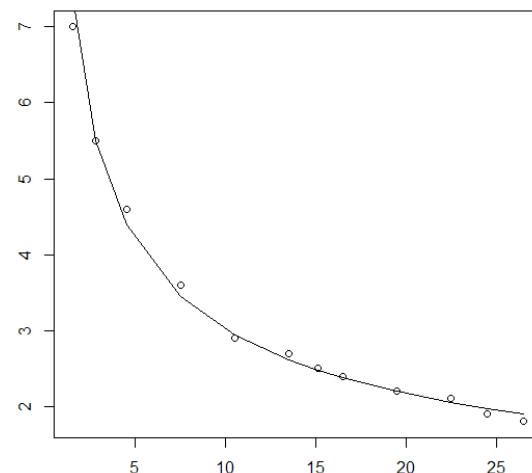
```
lm.pow=lm(log(y)~log(x))
```

```
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
```


案例分析——预测网页流量

- 使用互联网排名前1000的网站的数据
- Rank : 排名
- PageViews : 网站访问量
- UniqueVisitor : 访问用户数目
- HasAdvertising : 是否有广告
- IsEnglish : 主要使用的语言是否为英语

Rank	Site	Category	UniqueVisitors	Reach	PageViews	HasAd- vertising	InEnglish	TLD
1	facebook.com	Social Net- works	880000000	47.2	9.1e+11	Yes	Yes	com
2	youtube.com	Online Video	800000000	42.7	1.0e+11	Yes	Yes	com
3	yahoo.com	Web Portals	660000000	35.3	7.7e+10	Yes	Yes	com
4	live.com	Search En- gines	550000000	29.3	3.6e+10	Yes	Yes	com
5	wikipedia.org	Dictionaries & Encyclo- pedias	490000000	26.2	7.0e+09	No	Yes	org

案例分析——预测网页流量

```
top.1000.sites <- read.csv('data/top_1000_sites.tsv',  
                           sep = '\t',  
                           stringsAsFactors = FALSE)
```

```
ggplot(top.1000.sites, aes(x = PageViews, y = UniqueVisits))  
  geom_point()
```

- 可以看到数据分布很不均匀，集中在左下角坐标原点附近。数据间差异太大时，可以考虑对数据进行log变换

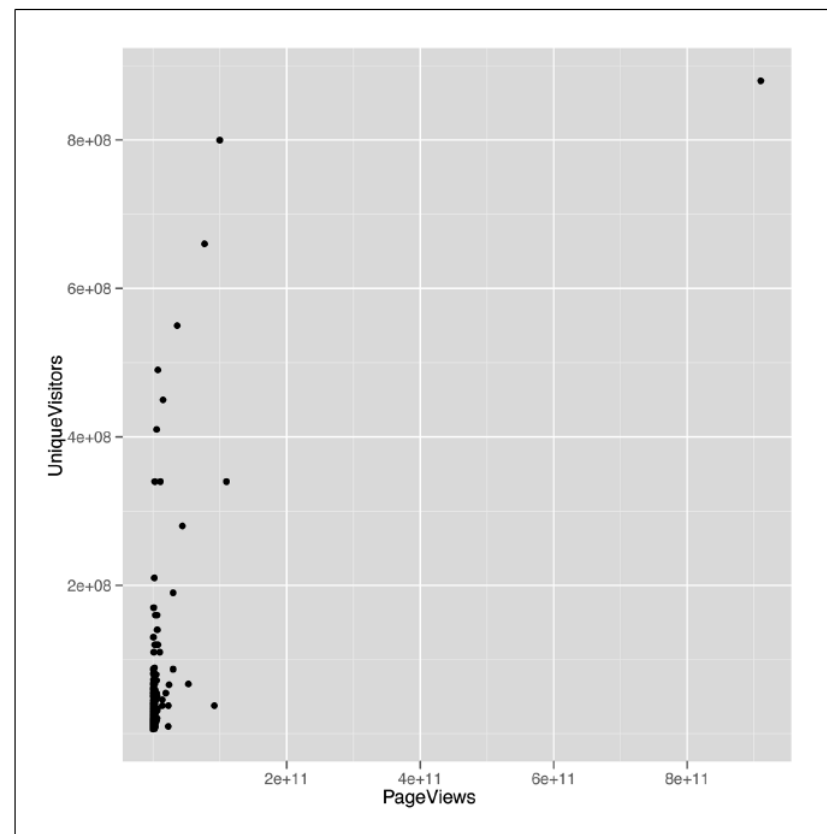


Figure 5-6. Scatterplot of UniqueVisitors versus PageViews

案例分析——预测网页流量

```
ggplot(top.1000.sites, aes(x = log(PageViews), y = log(UniqueVisitors))) +  
  geom_point()
```

- 可以看到经过对数变换后问题得到了改善

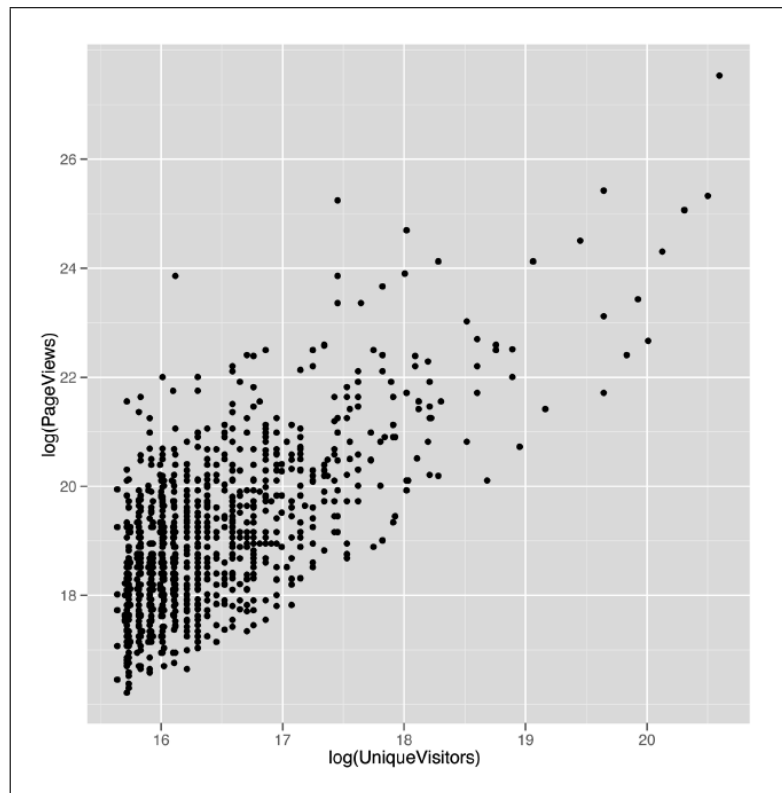


Figure 5-9. Log-scale scatterplot of UniqueVisitors versus PageViews

■ 先对单一一个变量做一元线性回归分析

```
lm.fit <- lm(log(PageViews) ~ log(UniqueVisitors),
             data = top.1000.sites)

summary(lm.fit)

#Call:
#lm(formula = log(PageViews) ~ log(UniqueVisitors), data = top.1000.sites)
#
#Residuals:
#   Min       1Q   Median       3Q      Max
#-2.1825 -0.7986 -0.0741  0.6467  5.1549
#
#Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
#(Intercept)    -2.83441    0.75201  -3.769 0.000173 ***
#log(UniqueVisitors) 1.33628    0.04568  29.251 < 2e-16 ***
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#Residual standard error: 1.084 on 998 degrees of freedom
#Multiple R-squared:  0.4616,    Adjusted R-squared:  0.4611
#F-statistic: 855.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

■ 对多个变量做多元线性回归分析

```
lm.fit <- lm(log(PageViews) ~ HasAdvertising + log(UniqueVisitors) + InEnglish,
             data = top.1000.sites)
summary(lm.fit)

#Call:
#lm(formula = log(PageViews) ~ HasAdvertising + log(UniqueVisitors) +
#   InEnglish, data = top.1000.sites)
#
#Residuals:
#   Min       1Q   Median       3Q      Max
#-2.4283 -0.7685 -0.0632  0.6298  5.4133
#
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)    -1.94502     1.14777   -1.695  0.09046 .
#HasAdvertising  0.30595     0.09170    3.336  0.00088 ***
#log(UniqueVisitors) 1.26507     0.07053   17.936 < 2e-16 ***
#InEnglishNo      0.83468     0.20860    4.001 6.77e-05 ***
#InEnglishYes     -0.16913     0.20424   -0.828  0.40780
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#Residual standard error: 1.067 on 995 degrees of freedom
#Multiple R-squared:  0.4798,    Adjusted R-squared:  0.4777
#F-statistic: 229.4 on 4 and 995 DF,  p-value: < 2.2e-16
```