2014.3.25

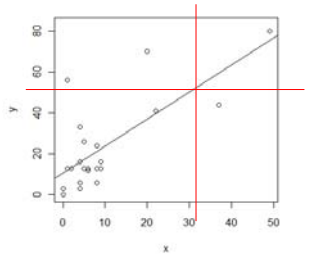
第十一讲 简单线性模型的应用

例2. (奥运主办国金牌预测). 为了预测北京奥运会中国金牌数目，首先需要确定预测变量.

主办国的表现与上一届密切相关

预测变量x: 上一届奥运会的金牌数(x)

历史数据：主办国金牌数(y)及其上届金牌数(x)。



拟合历史数据得回归方 程：$y = 10.6 + 1.3x$

上一届(2004)中国金牌数 $x = 32$，预测08届金牌数

$\hat{y} = 10.6 + 1.3 \times 32 = 52$（实际获得 51枚金牌）

主办国(y)	上届(x)
26	5
70	20
56	1
24	8
13	2
13	9
6	4
41	22
33	4
3	4
6	8
13	6
13	8
16	4
3	0
13	5
0	0
80	49
83	NA
12	6
13	1
44	37
16	9

例3. (标准化) 体重指数的定义 – 通过发现体重与身高的关系，消除身高的影响之后定义一个标准化的指数

如何计算一个体重指标，刻画体重是否正常或不合标准？

最简单的做法是将重量W作为指数：
假设W或者log(W) 服从正态分布N(μ , σ²), 重量处于群体95%置信区间定义为正常（比如：若log(W)> μ +1.645 σ，定义为偏胖（体重超过95%的人）

但显然，不同身高、性别、年龄的人不具可比性（不能认为所有人来自于同一总体或iid），即 μ 是若干因素的函数。

为了定义一个具有普适性的指数, 需要考虑到上式中 影响 μ 的其它因素, 比 如身高 H。假设成年男性的体重服从

假设身高H的成年男性的体重服从

$$\log(W) \sim N\left(a + b\log(H),\ \sigma^2\right)$$

$$\Leftrightarrow \log(W) = a + b\log(H) + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

标准化： $z = \dfrac{\log(W) - (a + b\log(H))}{\sigma} \sim N(0,1)$

例如，若 $z > 1.645 \Leftrightarrow \dfrac{W}{H^b} > e^{a + 1.645\sigma}$，定于为偏胖。

经验数据表明 $b \approx 2$

Body Mass Index : $BMI = \frac{W}{H^2}$ 单位kg/m^2

Category	BMI range – kg/m²	BMI Prime	Mass (weight) of a 1.8 metres (5 ft 11 in) person with this BMI.
Severely underweight	less than 16.0	less than 0.66	less than 51.8 kilograms (8.16 st; 114 lb)
Underweight	from 16.0 to 18.5	from 0.66 to 0.73	between 51.8 and 59.9 kilograms (8.16 and 9.43 st; 114 and 132 lb)
Normal	from 18.5 to 25	from 0.74 to 0.99	between 59.9 and 81.0 kilograms (9.43 and 12.76 st; 132 and 179 lb)
Overweight	from 25 to 30	from 1.0 to 1.19	between 81.0 and 97.2 kilograms (12.76 and 15.31 st; 179 and 214 lb)
Obese Class I	from 30 to 35	from 1.2 to 1.39	between 97.2 and 113.4 kilograms (15.31 and 17.86 st; 214 and 250 lb)
Obese Class II	from 35 to 40	from 1.4 to 1.59	between 113.4 and 129.6 kilograms (17.86 and 20.41 st; 250 and 286 lb)
Obese Class III	over 40	over 1.6	from 129.6 kilograms (20.41 st; 286 lb)

from wikP

例4. 幂次定律 (power law)

异速生长学(allometry)描研究植物的大小与其形状，结构，生理和行为的关系。**Kleiber's law**表明， 动物新陈代谢速率与体重的**3/4**次幂成正比

$$\text{Kleiber's law}: \quad \text{Metabloic Rate} = 70 \times \text{Mass}^{0.75}$$

这说明，体重/体积多1倍，代谢率只多$2^{3/4} = 1.68$ 倍.

再如, 纽约时报(2009)一篇题为"Math and the City"专栏文章中, 描述了城市能源消耗c (比如加油站数目) 、交通流量等与城市人口规模(s)呈现一定规律。服从幂次为 3/4 的幂次定律:

$$c \propto s^{3/4}$$

人口多1倍, 加油站数量只多$2^{3/4} = 1.68$ 倍。 很多自然进化的有机系统服从3/4 幂次律或其它幂次律: 自然进化的生态系统越大，越有效。

一般地，一个幂次律(power law) 指的是如下概率密度形式

$$p(x) \propto x^{-b}, x > x_{\min}, \; b > 0$$

也称作Pareto分布, 是一种重尾分布(heavy-tailed) 或长尾分布。

Pareto法则 (20-80法则, 二八法则): 意大利经济学家Pareto于1906年发现意大利80%的土地被20%的人所有。

现在，长尾经济用来描述诸如亚马逊和Netflix之类网站的商业和经济模式 ：关注尾部众多非热卖品

p(x) 代表概率，也可以是其他量, 比如规模(size):财富、人口、 森林大火面积、 河流面积…

若干幂次律的例子

- **Gutenberg-Richter**定律：***M*** 级以上的地震发生的频率***N***与能量成反比

  $N \propto \left(10^M\right)^{-b}$， b在0.5和1.5之间， 刻画地震活跃程度,一般为1

- 社交网络中成员的度(有连接的节点数目)的分布服从幂次律

  $P(k) \propto k^{-b}$， b在2和3之间(scale-free network)

  少数点(20%)有较多的连接(80%)，而多数点有较少的连接

- **Zipf**定律: 语言学家Zipf (1949) 发现大众语言用词频率满足幂次律：第 k 常用的单词的使用频率 f_k 与 k 成反比

单词	the	of	and	
排序 k	1	2	3	…
概率 f_k	7.5%	3.5%	2.8%	

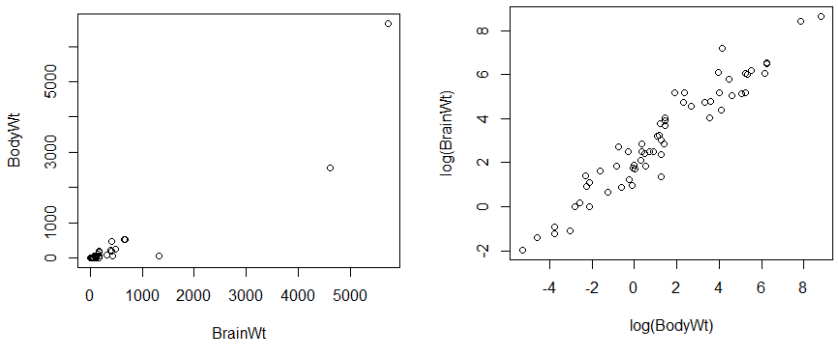
$\Rightarrow f_k \propto k^{-b}, \; b = 1$

不同的作者可能服从不同的幂次律(不同的b), 以此可判断某篇无名作品作者是否是莎士比亚，红楼梦作者问题等。

幂次率的发现通常基于对log变换数据的线性模型拟合

例4.1. 哺乳动物脑重与体重的关系.

alr3 数据集 brains 给出了62种哺乳动物的体重(kg)与脑重(g)数据



假设模型 : $\log(\text{BrainWt}) = a + b\log(\text{BodyWt}) + \varepsilon$

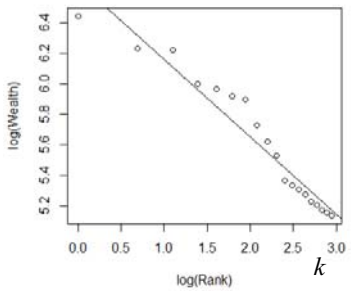

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
BodyWt       0.75169    0.02846   26.41   <2e-16 ***
---
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared: 0.9208,    Adjusted R-squared: 0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

$\hat{b} = 0.75$，显著关联 : 如果一种动物的体重是另外一种动物的2倍，则其平均脑重是其$2^{0.75} = 1.68$倍

拟合方程 : $\log(\text{BrainWt}) = 2.14 + 0.75 \times \log(\text{BodyWt})$

等价地, 有幂次律 : $\text{BrainWt} = e^{2.14+0.75\times\log(\text{BodyWt})} = 8.5 \times \text{BodyWt}^{0.75}$

例4.2. 2012福布斯中国富豪榜



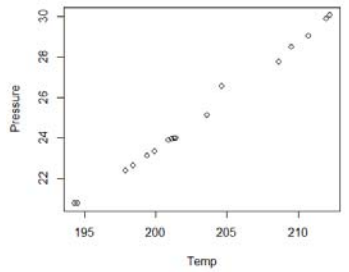
$\log(\text{Wealth}_k) = 6.67 - 0.5\log(k)$

$\text{Wealth}_k = 788/\sqrt{k}$

2012排名	2012财富(亿人民币)	姓名	公司
1	630	宗庆后	娃哈哈集团
2	510.3	李彦宏	百度
3	504	王健林	大连万达集团
4	403.2	马化腾	腾讯
5	390.6	吴亚军夫妇	龙湖地产
6	371.7	梁稳根	三一集团
7	365.4	刘永行	东方希望集团
8	308.7	许家印	恒大集团
9	277.2	杨惠妍	碧桂园
10	252	许荣茂	世茂集团
11	214.2	马云(微博)	阿里巴巴(微博)
12	207.9	何享健	美的集团
13	201.6	张近东	苏宁电器(微博)
14	195.3	孙广信	新疆广汇实业投资(集团)
15	185.9	丁磊	网易
16	182.7	魏建军	长城汽车(微博)
17	176.4	陈丽华家族	富华国际集团
18	173.3	刘永好家族	新希望集团
19	170.1	卢志强	泛海集团
19	170.1	周成建(微博)	美邦服饰

例5. 气压与沸点

1857年爱尔兰物理学家James D. Forbes发表了一篇文章研究气压与水的沸点之间的关系。 因为气压计容易破碎而不方便使用，所以Forbes 希望能得到沸点与气压的关系，从而把沸点作为气压的替代。 alr3 数据集 forbes 包含了他在山上不同高度测得的17组数值 (Temp:沸点,单位:F; Pressure: 气压, 单位: 水银柱英寸高度).

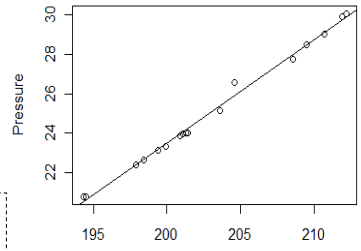


	Temp	Pressure
1	194.5	20.79
2	194.3	20.79
3	197.9	22.40
4	198.4	22.67
5	199.4	23.15
6	199.9	23.35
7	200.9	23.89
8	201.1	23.99
9	201.4	24.02
10	201.3	24.01
11	203.6	25.14
12	204.6	26.57
13	209.5	28.49
14	208.6	27.76
15	210.7	29.04
16	211.9	29.88
17	212.2	30.06

模型 1

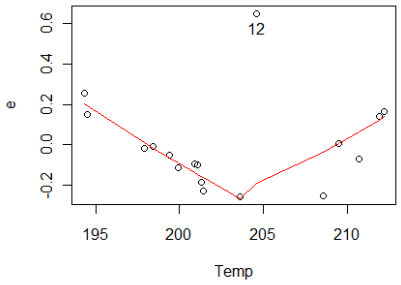
假设模型：

（1）$\text{Pressure} = a + b \times \text{Temp} + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2)$




```
> plot(Pressure~Temp, data=forbes)
> fit1=lm(Pressure~Temp, data=forbes)
> abline(fit1)
```

```
> plot(fit1,which=1)
```



残差分析:

(1) 残差是否是iid的?

(2) 残差是否有非线性的趋势?

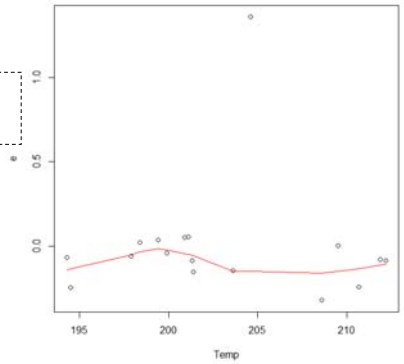
模型 2

*Forbes*根据物理知识假设模型：

（2）$\text{LPres} = a + b \times \text{Temp} + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2)$

其中$\text{LPres} = 100 \log_{10}(\text{Pressure})$

```
> fit2=lm(Lpres~Temp, data=forbes)
> plot(fit2,which=1)
```



模型 3

经典热力学中的Clausius–Clapeyron 公式(1850) 表明

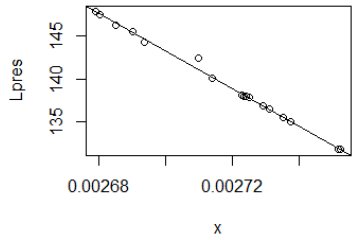
$$\text{LPres} = a + b \times \frac{1}{255.37 + \frac{5}{9}\text{Temp}} + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2)$$

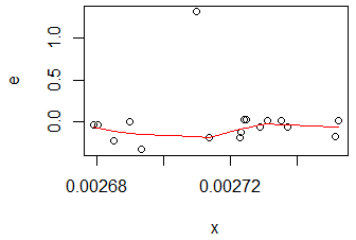
关于Temp不是一个线性模型，但 令$x = \dfrac{1}{255.37 + \frac{5}{9}\text{Temp}}$,

则为线性模型：

（3）$\text{LPres} = a + bx + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2)$


```
> x=1/(255.37+5/9*Temp)
> fit3=lm(Lpres~x, data=forbes)
> plot(x, Lpres)
> abline(fit3)
> plot(fit3, which=1)
```

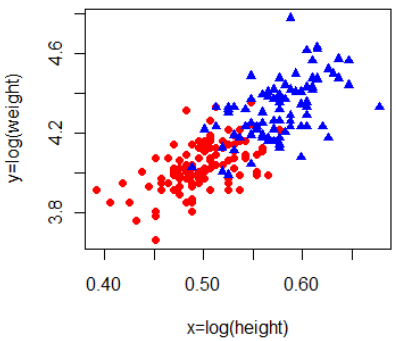




模型3的$R^2 = 0.9953$略大于模型2的$R^2 = 0.995$，所以模型3拟合效果略好。

注：模型2是模型3的一阶近似，两个模型都是同一个响 应变量*Lpres*对一个自变量的简单回归，所 以比较两者的R^2(等价地比较残差平方和)是有意义的。而模型1和3的响应变量不同，它们 的R^2没有可比性。

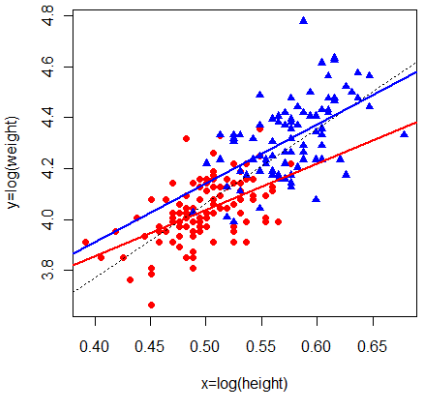
例6. 体重与身高
– 从简单回归到多重回归



	weight	height	sex
1	1.82	77	1
2	1.61	58	0
3	1.61	53	0
4	1.77	68	1
5	1.57	59	0
6	1.70	76	1
7	1.67	76	1
8	1.86	69	1
9	1.78	71	1
10	1.71	65	1
11	1.75	70	1
.....			

$$x = \log(weight), y = \log(height),$$

简单回归分析



不考虑性别：$y = 2.6 + 2.93x,$

男性：$y = 2.98 + 2.32x, \sigma = 0.128$

女性：$y = 3.12 + 1.83x, \sigma = 0.103$

除了截距不同，男女两组有共性: (1) 回归线斜率相同; (2) 误差方差相同.

这些共性使得我们将两种性别的数据置于同一模型下

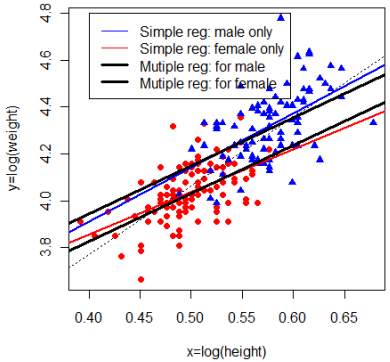
$$y = a + bx + cz + \varepsilon, \varepsilon \sim (0, \sigma^2), \varepsilon \perp x$$


```
> lm (log(weight)~log(height) + sex, data=weightheight)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  3.009     0.116       26.028   < 2e-16
log(height)  2.057     0.231       8.909    3.55e-16
sexM         0.124     0.0243      5.113    7.51e-07
---
Residual standard error: 0.1145 on 196 degrees of freedom
Multiple R-squared: 0.6601,    Adjusted R-squared: 0.6567
F-statistic: 190.4 on 2 and 196 DF,  p-value: < 2.2e-16
```

得到拟合方程：$y = 3 + 2x + 0.12z$

$z = 0$时 (女)：$y = 3 + 2x;$

$z = 1$时 (男)：$y = 3.12 + 2x;$



假设误差变量$\varepsilon \sim \mathrm{N}(0, \sigma^2)$，由拟合所得结果，$\hat{\sigma} = 0.115,$

$$\mathrm{u} = \frac{\log(W) - 3 - 2\log(L) - 0.12z}{0.115} \overset{\text{近似}}{\sim} N(0,1)$$

$$\mathrm{u} \geq \mathrm{C} \Leftrightarrow \frac{W}{H^2} > \exp(3 + 0.115\mathrm{C} + 0.12z) \triangleq C_z$$

若 $\mathrm{C} = 1.645$（标准正态分布的上95%分位点）

对于女性，$C_0 = \exp(3.19) = 24.3$

对于男性，$C_1 = \exp(3.19 + 0.12) = 27.4$