课程主页：http://staff.ustc.edu.cn/~ynyang/2022

第九讲 简单线性模型的应用

2022.10.21

Benford定律
1----------2---------3-------4-----5----6---7--8--9

结果的解释: 因果还是关联?

$$y = a + bx + \varepsilon$$

- 随机化控制试验或天然试验：x是外生的，即x与ε独立，则LS估计$\hat{b}$是无偏的，结果可表述为因果关系：

  对同一个研究对象，x每增加一个单位，y的期望增加$\hat{b}$个单位。

- 观察研究：自变量一般是内生的，LS估计$\hat{b}$有偏，结果只能表述为关联关系：

  如果一个研究对象的x比另外一个研究对象大1个单位，则相应的y的期望大$\hat{b}$个单位。

例4 (Freedman book). 分析2001年人口抽样调查数据，得到妻子教育水平（上学的年数）与丈夫教育水平的回归方程如下：

$$\text{WifeEdLevel} = 5.60 + 0.57 \times \text{HusbandEdLevel} + \text{residual}$$

如果公司送王先生到大学在职培养一年，你是否预期王太太的教育水平会上升0.57年？若不是，0.57的含义是什么？

这是观察研究而非试验，没有证据表明误差与自变量独立，结果是关联而不是因果。 $b = 0.57$的含义是：如果该研究中某人比另外一个人多上一年学，那么他的妻子比另外一人的妻子期望多上0.57年学。

简单模型的应用：幂次律

幂次律

如果变量x, y满足

$$y = cx^{-k},$$

则称它们满足幂次律(power law)。 著名的幂次律包括牛顿万有引力定律、语言学中的Zipf定律、生物学中的Kleiber定律等等。幂次律的发现一般通过log尺度上的线性回归模型得到：

$$\log(y) = \log(c) - k\log(x)$$

幂次律最重要的特点是刻度不变性或无标度(scale invariance，scale-free):

假设幂次律$f(x) = cx^{-k}$，若改变x的刻度单位: $x \to sx$

$$f(sx) = c(sx)^{-k} = (s)^{-k}cx^{-k} \propto f(x)$$

自然界、社会经济中广泛存在幂次定律，比如许多涉及“规模(size)”的问题都可能存在幂次律。比如城市交通流量、森林大火面积、河流面积、排名次序、财富数量、朋友圈人数等。

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 (比如加油站数目) 、交通流量等 与城市人口规模(*size*)呈现一定规律, 服从幂次为 3/4 的幂次律:

$$\text{能量消耗} \propto (\text{人口数})^{3/4}$$

人口增加1倍, 加油站数量只多$2^{3/4} = 1.68$ 倍。这说明自然进化的生态系统规模越大，越有效。

幂次律的例子（https://en.wikipedia.org/wiki/Power_law）

Examples [edit]

More than a hundred power-law distributions have been identified in physics (e.g. sandpile avalanches), biology (e.g. species extinction and body mass), and the social sciences (e.g. city sizes and income).[15] Among them are:

Astronomy [edit]

- Kepler's third law
- The initial mass function of stars
- The differential energy spectrum of cosmic-ray nuclei
- The M–sigma relation

Physics [edit]

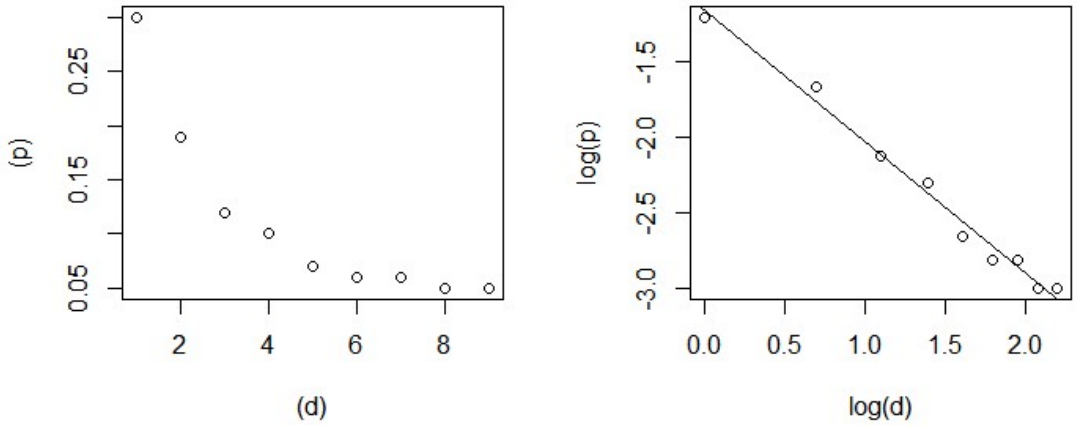
- The Angstrom exponent in aerosol optics
- The frequency-dependency of acoustic attenuation in complex media
- The Stefan–Boltzmann law
- The input-voltage–output-current curves of field-effect transistors and vacuum tubes approximate a square-law relationship, a factor in "tube sound".
- Square–cube law (ratio of surface area to volume)
- A 3/2-power law can be found in the plate characteristic curves of triodes.
- The inverse-square laws of Newtonian gravity and electrostatics, as evidenced by the gravitational potential and Electrostatic potential, respectively.
- Self-organized criticality with a critical point as an attractor
- Model of van der Waals force
- Force and potential in simple harmonic motion
- Gamma correction relating light intensity with voltage

1. Benford定律

例1. 下表数据是美国3142个县的人口数的首位数字(d) 的频率 (p)。首位数字并不均匀，首位数字为1,2,…,9的概率依次下降， 1出现的频率最大。

首位数字d	1	2	3	4	5	6	7	8	9
频数	956	593	380	301	225	203	177	159	148
频率p	0.30	0.19	0.12	0.10	0.07	0.06	0.06	0.05	0.05

(d, p) 散点图并非线性，但取对数之后基本线性。



拟合简单线性模型：$\log(p) = a + b\log(d) + error,$

LS估计 $\hat{a} = -1.16, \hat{b} = -0.87$

拟合得到的回归直线 : $\log(p) = -1.16 - 0.87 \times \log(d)$

取指数函数得到幂次律：

$$p = p(d) = 0.31 / d^{0.87}$$

拟合值与观察频率十分 接近：

首位数字d	1	2	3	4	5	6	7	8	9
样本频率p	0.30	0.19	0.12	0.10	0.07	0.06	0.06	0.05	0.05
幂次律	0.314	0.172	0.121	0.094	0.077	0.066	0.058	0.051	0.046
Benford律	**0.301**	**0.176**	**0.125**	**0.097**	**0.079**	**0.067**	**0.058**	**0.051**	**0.046**

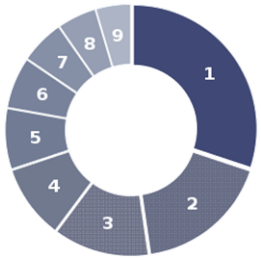
最后一行是Benford定律的理论概率值（下页）

Benford定律

对于“自然出现的”正数，比如人口数、河流面积、财务报表和新闻中出现的数据（都是正数），考虑首位非0数字，比如

0.032，3.2的首位非0数字是3

Newcomb(1881), Benford （1938）发现“自然出现的”正数的首位数字概率分布大多（近似）满足Benford定律：



Benford定律：在一定的假设下 ,自然数字的首位数字是 d的概率为

$$P(d)=\log_{10}(1+1/d), d=1,2,...,9$$

首位数d	1	2	3	4	5	6	7	8	9
概率P(d)	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Benford定律 ≈ 幂次律

$$p(d)=\log_{10}(1+1/d)=0.43\left(\frac{1}{d}-\frac{1}{2d^2}+\frac{1}{3d^3}-....\right)\approx 0.31/d^{0.86}$$

应用: 通过Benford分布发现财务报表造假

2. 齐夫定律 Zipf's law

Zipf's law

齐夫定律（Zipf's law）是由哈佛大学语言学家G.K.Zipf于1949年发表的实验定律。它可以表述为：

在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比：$p_k \propto 1/k$，k =rank

Zipf定律说明，频率最高的单词出现的频率大约是出现频率第二位的单词的2倍，而出现频率第二位的单词则是出现频率第四位的单词的2倍。下表是大众语言前三个高频词的统计频率：

单词	the	of	and	…
排序 k	1	2	3	…
概率 p_k	7%	3.5%	2.8%	…

排在前面的大多是冠词、代词等，高频动词、名词排名在30之后。

最频动词：say, go, make, see，look, come, think
高频名词：time, people, year, way, day, thing, man

联邦文献作者问题

不同类型、不同作者的词频分布可能不同，但一般服从如下一般的Zipf定律，其中指数α刻画了文本风格：

$$\text{Freq} \propto \frac{1}{\text{Rank}^{\alpha}}$$

例如美国Hamilton, Madison所作的联邦文献的前十个高频词分布如下，显然两人写作风格有差异。

两人数据分别拟合线性模型

$$\log(\text{Freq}) = \text{a} - \alpha \times \log(\text{Rank})$$

拟合效果良好，得到Hamilton的α=0.900，Madison $\alpha = 0.902$

	Hamilton		Madison	
Words	Freq	Rank	Freq	Rank
the	91.27	1	93.65	1
of	64.65	2	57.8	2
to	40.71	3	35.25	3
and	24.5	4	27.55	4
in	24.37	5	23.05	5
a	22.85	6	20.22	6
be	20.06	7	16.45	7
that	14.98	8	14.37	8
it	13.82	9	13.34	9
is	11.7	10	12.76	10

高频汉字

下表列出了中文报刊中前十个高频汉字（前42个高频词占25%）

单字	的	一	是	不	了	在	有	人	这	大
频率%	4.87	1.41	1.32	1.07	0.95	0.93	0.91	0.78	0.76	0.58

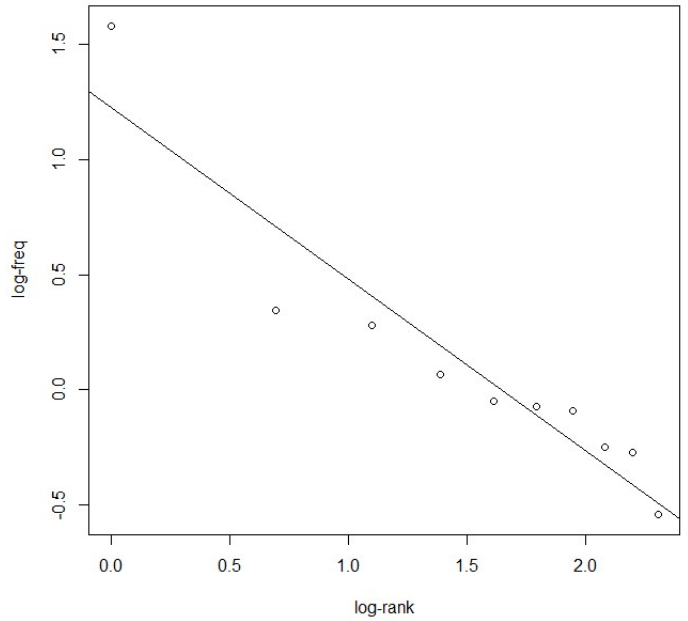
拟合线性模型

$$\log(\text{Freq}) = 1.23 - 0.75 \times \log(\text{Rank})$$

$$\alpha = 0.75,\ \text{Freq} \propto \frac{1}{\text{Rank}^{0.75}},$$

α远远偏离1，这可能是因为上述统计的是单字频率而不是词汇频率

（中文是否有词汇频率统计？）



3. 异速生长与Kleiber定律

Square-cube law

生物器官或物体的体积V（或重量）正比于长度的3次方，而表面积S正比于长度的2次方，所以面积正比于体积的$2/3$次方：

$$S \propto V^{2/3}$$

生物生长过程中，如果器官随着身体的增长而等比例地线性增长，满足square-cube law，称为同速生长（isometric scaling）。

但多数器官的生长不呈线性关系，比如眼睛显得生长较慢，而腿部生长较快，这称为异速生长(allometry)。

为什么大象的腿很粗？
为什么大飞机、大船难以制造？
为什么大鸟的翅膀很长且扇动很慢 (小鸟翅膀短且扇动快)？
为什么小动物心跳和呼吸快？
这都是异速生长的要求。

异速生长

体积或重量与表面积不服从同速生长规律就称为异速生长 (allometric scaling)。异速生长学是关于身体大小与形状 、解剖学、生理学及行为间关系的研究。 异速生长在形状分析及生物相对生长研究中是一个著名的研究论题。
参见维基百科wikipedia.org/wiki/Allometry。

例如，动物承重能力与腿的截面积成正比，同速生长情况下，体重/体积增加一倍，截面积只增加为原来的 $2^{2/3} = 1.59$倍，为了承重，腿的截面积需要生长的更快一些。这就是异速生长。

英国人JBS Haldane在著名科普文章 On being the right size (1923) 阐述了动物体积(size)变化时，形状(shape)的变化规律，特别是动物种群之间或内部存在不成比例的异速生长现象。最为著名的例子是Kleiber 幂次定律。

Kleiber定律

Kleiber's law :

动物代谢速率(Metabloic_Rate)与体重(Mass)存在如下幂次律关系

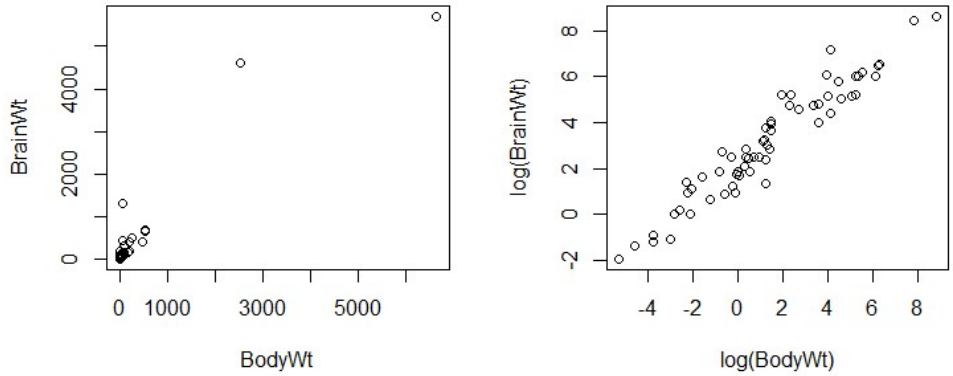
$$\text{代谢速率} = 70 \times \text{体重}^{3/4}, \quad 3/4 > 2/3$$

其中代谢速率单位为卡路里/秒,与表面积有关。

比如对于大体积动物，面积相对于体积显得太小，不能正常散热，所以需要异速生长：增大表面积（大鸟的翅膀特别长，一般不容易做到），或降低新陈代谢速度（大动物心跳慢、血液流动慢）。

再如，为了保持体温，小鸟散热快，所以可以快速扇动翅膀。大鸟体表面积相对较小散热较慢，故翅膀扇动较慢以减少热量的产生。

例2.（R Package alr3,数据集*brains*） *brains*数据给出了62种哺乳动物的脑重和体重数据. 散点图(左图)显示不出相关关系，但在对数尺度上呈现线性关系(右图)。



	BrainWt	BodyWt
Arctic fox	44.500	3.385
Owl monkey	15.499	0.480
Beaver	8.100	1.350
Cow	423.012	464.983
Gray wolf	119.498	36.328
…		

拟合简单线性模型：

$$\log(\text{BrainWt}) = a + b\log(\text{BodyWt}) + \varepsilon$$

$\hat{b} = 0.752 = 3/4, \hat{a} = 2.135,$

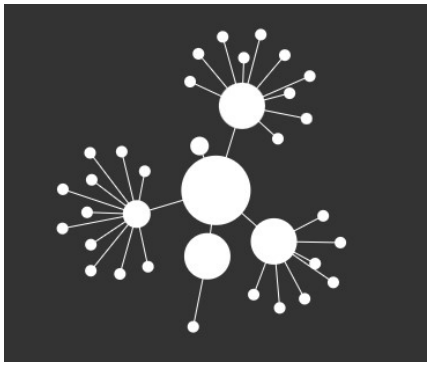
回归直线的估计为：$\log(\text{BrainWt}) = 2.135 + 0.752\log(\text{BodyWt})$

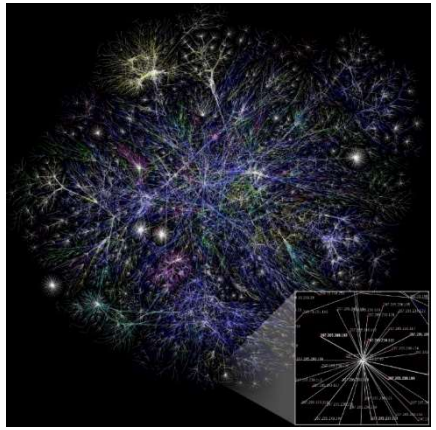
$\Leftrightarrow$ 幂次律：$\text{BrainWt} = 8.46 \times \text{BodyWt}^{3/4}$

4. 无标度社交网络

无标度社交网络（scale-free social network）中大多数成员有较少的连结，而少数点有较多的连结(称为hub)。每个节点的连结个数称为度数(degree)， 节点度数k服从Pareto分布/幂次律：

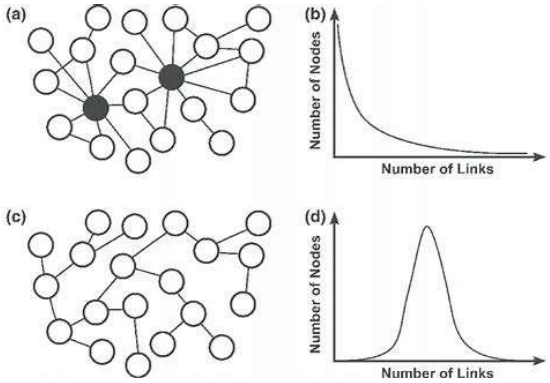
$$P(k) \propto k^{-r},$$





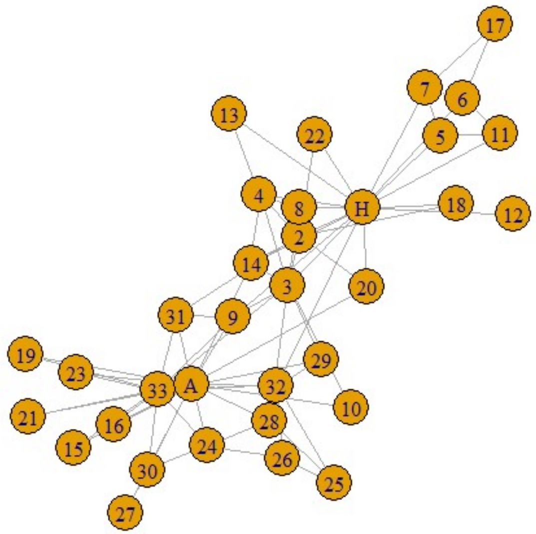
下图(a)是无标度网络，局部和整体看起来相像

(b)是一个随机网络，度数服从正态 。



例3 （karate club network 跆拳道俱乐部会员交往网络）共34人，其中有32个会员(Actor2-33)，A，H分别是经理和教练。 下面的邻接矩阵A表示34人之间的两两之间的关系，其中a_{ij} =1表示若i和j有密切交往，在社交网络图中i,j连线；否则$a_{ij} = 0$，不连线。

Mr Hi	.	1	1	1	1	1	1	1	1	.	1	1	1	1	.	.	.	1	.	1	.	1	.	.	.	.	.	.	.	.	.	1	.	.
Actor 2	1	.	1	1	.	.	.	1	.	.	.	.	.	1	.	.	.	1	.	1	.	1	.	.	.	.	.	.	.	.	1	.	.	.
Actor 3	1	1	.	1	.	.	.	1	1	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	1	.
Actor 4	1	1	1	.	.	.	.	1	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 5	1	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 6	1	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 7	1	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 8	1	1	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 9	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	1
Actor 10	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
Actor 11	1	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 12	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 13	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 14	1	1	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
Actor 15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1
Actor 16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1
Actor 17	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 18	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1
Actor 20	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
Actor 21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1
Actor 22	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Actor 23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1
Actor 24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	1	.	.	1	1
Actor 25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	1	.	.
Actor 26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	1	.	.
Actor 27	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1
Actor 28	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	1
Actor 29	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1
Actor 30	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	1	1
Actor 31	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1
Actor 32	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	1	.	.	.	1	1
Actor 33	.	.	1	.	.	.	.	.	1	.	.	.	.	.	1	1	.	.	1	.	1	.	1	1	.	.	.	.	.	1	1	1	.	1
John A	.	.	.	.	.	.	.	.	1	1	.	.	.	1	1	1	.	.	1	1	1	.	1	1	.	.	1	1	1	1	1	1	1	.

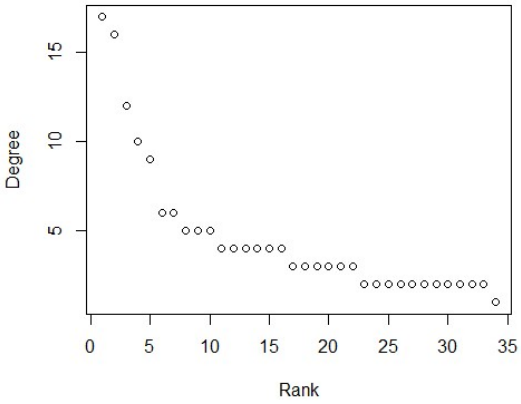


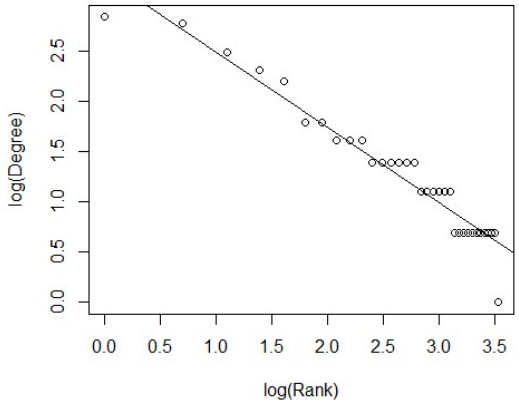
所有34人的度数(Degree)如右表（从大到小排列）

拟合简单线性模型得拟合直线

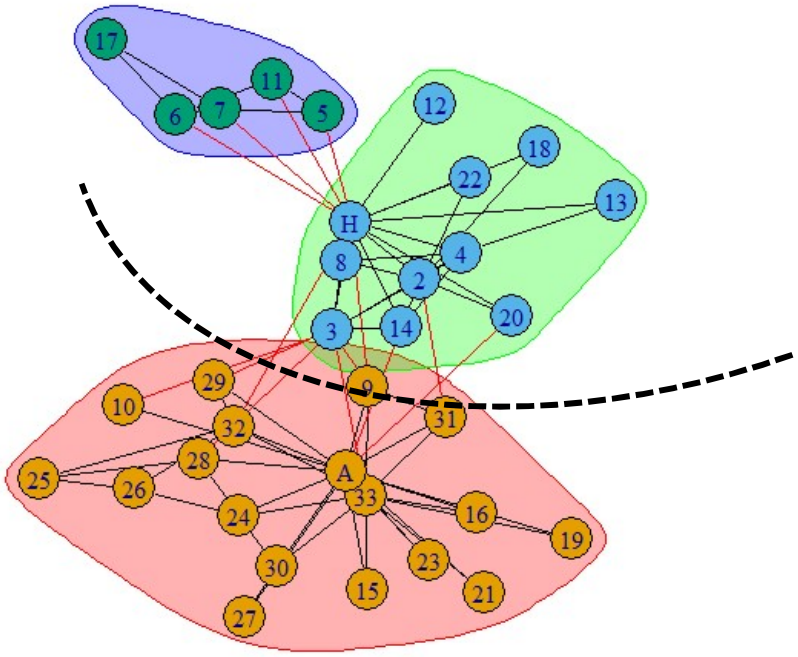
$$\mathrm{Log(Degree)} = 3.2 - 0.75 \times \log(\mathrm{Rank})$$

$$\mathrm{Degree} = 25/\mathrm{Rank}^{0.75}$$





会员	Degree
John A	17
Mr Hi	16
Actor 33	12
Actor 3	10
Actor 2	9
Actor 32	6
Actor 4	6
Actor 24	5
Actor 14	5
Actor 9	5
Actor 31	4
Actor 30	4
Actor 28	4
Actor 8	4
Actor 7	4
Actor 6	4
Actor 29	3
Actor 26	3
Actor 25	3
Actor 20	3
Actor 11	3
Actor 5	3
Actor 27	2
Actor 23	2
Actor 22	2
Actor 21	2
Actor 19	2
Actor 18	2
Actor 17	2
Actor 16	2
Actor 15	2
Actor 13	2
Actor 10	2
Actor 12	1



后期，俱乐部分裂为以H和A为中心的两家俱乐部。

聚类分析/社区检测

5. 帕累托分布

Pareto分布

帕累托（Pareto）分布是一种重尾、长尾、无标度分布(heavy-tailed, long-tailed, scale-free):

$$p(x) = k / x^r, \quad r > 1, x > c > 0.$$

重尾、长尾是相对于正态的指数阶尾概率而言的, 即能以较大概率取到很大的值。

Pareto法则(20-80法则，二八法则)： 80%的土地被20%的人所有。

简单模型的应用：体重指数

指数或指标（index）是反映复杂系统整体表现的度量，比如物价指数、消费指数、普尔指数。指数需要有普适性。

为了衡量人的体重是否超标，单纯用体重W作为指标不具有普适性，因为体重与身高（H）等因素有关（你可以对每个身高段定义体重标准，但这不够简洁）。**BMI被认为是适用于所有身高的成年人的一个体重（特别是脂肪）指数。**

BMI

成年人体重指数BMI (Body Mass Index)定义为

$$BMI = \frac{\text{体重}}{\text{身高}^2} \ (kg/m^2)$$

from wiki

Category	Underweight	Normal	Overweight	Obese I	Obese II	Obese III
BMI range	16.0-18.5	18.5-25	25-30	30-35	35-40	>40

BMI是比利时天文学家、数学家 L.A.J. Quetelet 在1830s提出的。人们普遍认为该指标适用于所有身高的人的体重度量，应用广泛。

显然，$BMI = W/H^2$ 在某种意义上校正了身高因素（消除了身高的影响），为什么不是W除以H^3？文献中很难查到Quetelet的原始想法，或许与去相关化或者回归有关。

线性模型与去相关化/标准化

假设$y_i, i = 1,\ldots,n$独立但不同分布（heterogeneous, 不一致),

假设y_i与变量x_i有关，满足线性模型：

$$y_i = a + bx_i + \varepsilon_i,\ \varepsilon_i\ iid \sim (0, \sigma^2),$$

其中误差

$$\varepsilon_i = y_i - a - bx_i$$

与y_i有关，但消除了x_i的影响，且同分布(iid, homogeneous),

因此可作为可替代y_i但又具有一致性(iid)。

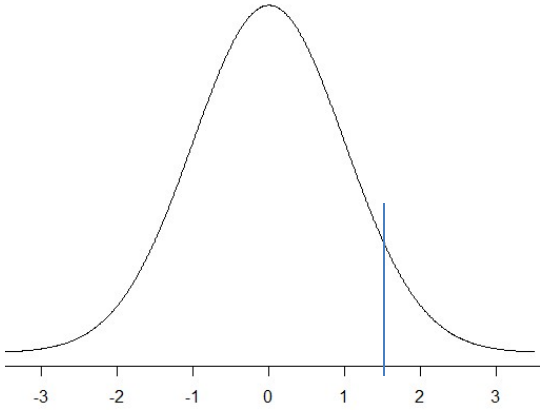
我们在对数尺度上建立线性模型：

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

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

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

z的分布$N(0,1)$与H无关，具有普适性。



若一个人的体重指标z超过95%的人，

即$z > 1.645$,则认为体重超标，

注意：$z > 1.645 \Leftrightarrow W/H^b > e^{a+1.645\sigma}$

结论：W/H^b 的分布与身高无关,可作为体重指数,其分布是与身高无关的对数正态分布，经验数据表明 $b \approx 2$.

```
hw=read.table("http://staff.ustc.edu.cn/~ynyang/2022/lab/height-weight.txt",head=T)
sex=hw[,1]
hw[sex==1,]->male #男性的身高-体重数据，单位：千克，米
lm(log(weight)~log(height) ,data=male) #拟合线性模型(对数尺度)

Coefficients:
(Intercept)  log(height)
      3.00         2.27
```

$a=3.00,\ b=2.27, \sigma=0.12$

$z>1.645 \Leftrightarrow W/H^b > e^{a+1.645\sigma} \Leftrightarrow W/H^{2.27} > e^{3+1.645\times 0.12}=24.5$

若某人的$W=70\text{kg}, \text{H}=1.8\text{m}$，$W/H^{2.27}=18.43$，小于95%阈值24.5。

指标18.43在群体处于什么水平？

$$z=\frac{\log(W)-(a+b\log(H))}{\sigma}=\frac{\log(70)-(3+2.27\log(1.8))}{0.12}=-0.715$$

$P(N(0,1)>-0.715)=76\%$，有76%的人的指标超过这个人。

简单模型的应用3：预 测

以自变量预测响应变量是统计学习的主要内容。有时，仅仅使用一个预测变量的简单线性回归模型可能比复杂的模型有更好的预测效果。

例3. 预测北京2008奥运会中国金牌数目。

主办国的表现与上一届的表现有关，

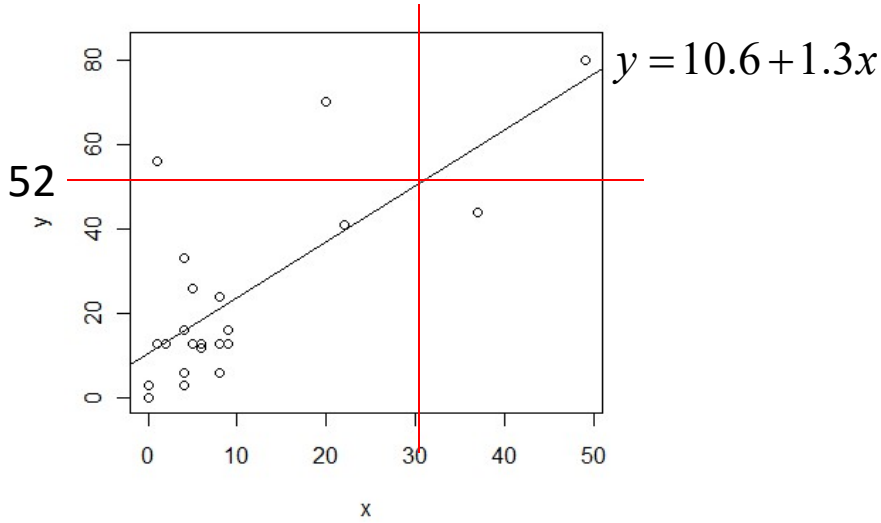
预测变量取为：上一届奥运会的金牌数(x)

历史数据：主办国金牌数(y)及其上届金牌数(x)，如左边表格所示。

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

假设模型$y = a + bx + \varepsilon$

$\hat{a} = 10.6, \hat{b} = 1.3, \hat{\sigma} = 15.2$



上一届(2004)中国金牌数$x = 32$,

预测08届金牌数：$10.6 + 1.3 \times 32 = 52$