第七讲 幂次律-简单线性模型的应用

2023.11.3

Benford定律

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

P2-7： 回顾（第六讲P15-27）

$e_i = y_i - \hat{a} - \hat{b}x_i$可看作是r.v. $\varepsilon_i = y_i - a - bx_i$的预测,

我们尝试基于$\{e_i，i = 1,2,\ldots,n\}$估计σ^2。

σ^2的“LS”估计取为： $\hat{\sigma}^2 = \frac{1}{n-2}RSS = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$

惯例：

- 虽然 $\hat{\sigma}^2$不是由最小二乘法直接得到的，但通常也称之为LS估计。
- 为什么除以$n-2$而不是$n-1$? 因为估计了两个参数a和b。

引理1.（1）$e_i = (\varepsilon_i - \bar{\varepsilon}) - (x_i - \bar{x})s_{x\varepsilon} / s_{xx}$.

（2）$RSS = s_{yy} - s_{xy}^2/s_{xx} = s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2 / s_{xx}$

$\boxed{\text{命题4. } \hat{\sigma}^2\text{是}\sigma^2\text{的无偏估计，即}E(\hat{\sigma}^2)=\sigma^2\text{。}}$

证明：利用引理1, $RSS = s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2 / s_{xx}$。

我们已知$\operatorname{var}(\hat{b}\,|\,\mathbf{x}) = \dfrac{\sigma^2}{s_{xx}}$，将其中的位置参数$\sigma^2$的估计代入其估计(plug - in)得：

$$\widehat{\operatorname{var}}(\hat{b}\,|\,\mathbf{x}) = \frac{\hat{\sigma}^2}{s_{xx}},$$

标准差：$se(\hat{b}) = \sqrt{\widehat{\operatorname{var}}(\hat{b}\,|\,\mathbf{x})} = \hat{\sigma} / \sqrt{s_{xx}}$,

截距项LS估计$\hat{a}$的方差估计类似得到（一般不关心）。

对于简单模型的假设$H_0 : b = b_0$（b_0已知），Wald检验统计量

$$W = \frac{\hat{b} - b_0}{se(\hat{b})} = \frac{\sqrt{s_{xx}}\,(\hat{b} - b_0)}{\hat{\sigma}}$$

为了对原假设进行检验或为了构造置信区间，需要建立检验统计量在原假设下的分布或枢轴量的分布，这里我们假设误差正态。

命题5. 假设模型$y_i = a + bx_i + \varepsilon_i, \varepsilon_1, ..., \varepsilon_n$ iid $\sim N(0, \sigma^2)$,则

(1) $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0,1)$

(2) $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, 且$\hat{\sigma}^2$与($\hat{a}$，$\hat{b}$)独立

(3) $W(b) = \dfrac{\sqrt{s_{xx}}(\hat{b} - b)}{\hat{\sigma}} \sim t_{n-2}$

证明：参见上一讲 P20 − 21.

注1： $W(b)$是枢轴量，可用于构建b的置信区间：

$$\left\{b: |W(b)| \le t_{n-2}\left(\frac{\alpha}{2}\right)\right\}$$

注2：在原假设$H_0: b = b_0$成立时，b_0已知，一般b_0=0。以b_0代替b, $W(b_0)$是Wald检验统计量：

$$t = W(b_0) = (\hat{b} - b_0)/se(\hat{b}) = \sqrt{s_{xx}}(\hat{b} - b_0)/\hat{\sigma},$$

由命题5知，$t \sim_{H_0} t_{n-2}$. $|t| \ge t_{n-2}(\alpha/2)$时拒绝原假设.

注3：（第六讲命题6）回归系数的显著性$H_0: b=0$的检验统计量

$$t=\sqrt{s_{xx}}\hat{b}/\hat{\sigma}=\sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}。$$

证：$\hat{b}=s_{xy}/s_{xx}$，$\hat{\sigma}^2=\frac{1}{n-2}RSS=\frac{1}{n-2}\left(s_{yy}-s_{xy}^2/s_{xx}\right)$，代入$t=\frac{\sqrt{s_{xx}}\hat{b}}{\hat{\sigma}}$。

注4：两样本t-检验是回归系数显著性检验的特殊情形

$$y_1,\ldots,y_{n_1}\ \text{iid}\sim N(\mu_1,\sigma^2) \qquad \leftarrow x_1,\ldots,x_{n_1}=1$$

$$y_{n_1},\ldots,y_{n_1+n_2}\ \text{iid}\sim N(\mu_2,\sigma^2) \qquad \leftarrow x_{n_1+1},\ldots,x_{n_1+n_2}=0$$

以线性模型表示两样本问题：

$$y_i=a+bx_i+\varepsilon_i \quad (a=\mu_2, b=\mu_1-\mu_2),\ \ \varepsilon_i\ \text{iid}\sim N(0,\sigma^2)$$

两样本$H_0:\mu_1=\mu_2 \Leftrightarrow$回归系数$H_0: b=0$。

容易验证$H_0: b=0$的检验统计量$t=\frac{\hat{b}}{\sqrt{\hat{\sigma}^2/s_{xx}}}=\frac{\bar{y}_1-\bar{y}_2}{\sqrt{(n_1^{-1}+n_2^{-1})s^2}}$

注5：其它参数得检验或置信区间类似构造。例如，均值函数/回归函数 $m(x_0)=E(y|x=x_0)=a+bx_0, x_0\in R$ 的置信区间

对于给定的x_0，$m(x_0)$的LS估计$\widehat{m}(x_0)=\hat{a}+\hat{b}x_0$，可以证明：

$$\widehat{m}(x_0)\sim N\left(m(x_0),\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)\right)$$

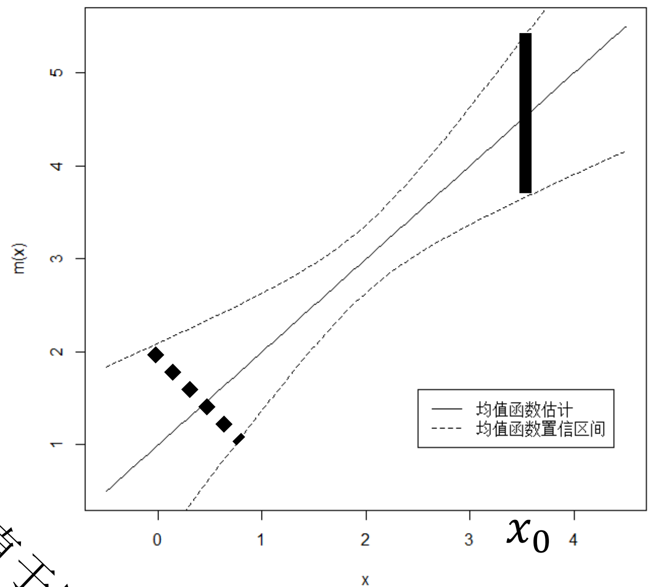
与$\hat{\sigma}^2$独立，类似于命题5可证

$$\frac{\widehat{m}(x_0)-m(x_0)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}}}\sim t_{n-2}$$

由此得 $m(x_0)$的置信水平$1-\alpha$的置信区间

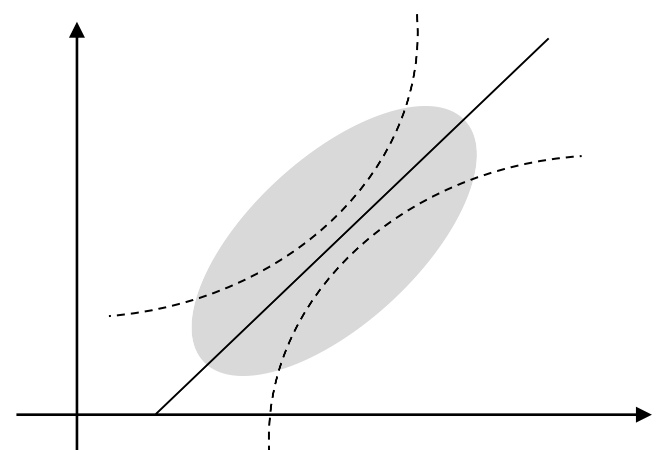
$$\left[\ \widehat{m}(x_0)\pm t_{n-2}(\alpha)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}}\ \right]$$

当 x_0 变化时，上述区间形成一个置信带



置信带在 y 轴方向
关于回归直线对称

在垂直于回归直线的方向不对称。（但我们一般会不自觉地以回归直线为中心，在垂直方向上看待对称性）



上图的回归线和置信带都不对

R软件： $\boldsymbol{lm(y\sim x)}$

例1. K.Pearson收集了1375对母女身高数据(单位：英寸)。前3行数据如右表所示。

	mheight	dheight
1	59.7	55.1
2	58.2	56.5
3	60.6	56.0

假设简单线性模型：

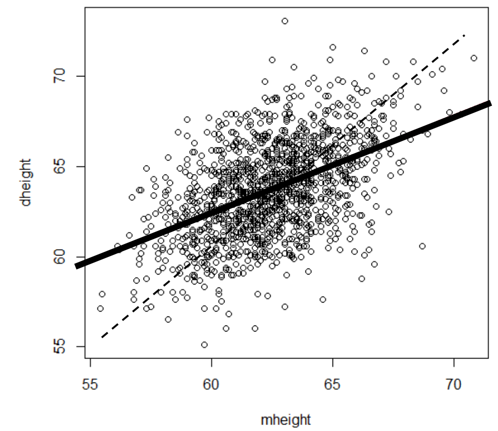
dheight = a+ b × mheigh + ε

```
> myfit <- lm(dheight ~ mheight, data=Heights)

Coefficients:
(Intercept)     mheight
   29.92         0.54

> plot(Heights)   #散点图
> abline(myfit)   #添加拟合回归直线
```

$\boxed{\hat{a}=29.92，\ \hat{b}=0.54}$



注意回归效应：若母亲身高80,则女儿身高的期望(预测)为 $\hat{a}+80\hat{b}$=73 < 80

结果汇总:
> summary(myfit)

t value=Estimate ÷ Std.Error
=第1列除以第2列

Coefficients:

	LS估计	**标准差**	**t检验**	**p值**
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.92	1.62	18.47	<2e-16 ***
mheight	0.54	0.026	20.77	<2e-16 ***

$$\hat{b} = 0.54,$$
$$se(\hat{b}) = 0.026$$
$$t = \frac{\hat{b}}{se(\hat{b})} = \frac{0.54}{0.026} = 20.77$$
p值 $< 2e-16$

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.266 on 1373 degrees of freedom
Multiple R-squared: 0.2408, Adjusted R-squared: 0.2402
F-statistic: 435.5 on 1 and 1373 DF, p-value: < 2.2e-16

$\hat{\sigma} = 2.266$
自由度：n-2=1373，
2个回归系数

R^2 =0.2408 (决定系数，R软件称之为"Multiple R−squared").

$$F = t^2 = 20.77^2$$

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:

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

无标度(scale-free)：若改变x的单位，则幂次律依然成立。

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、河流面积、排名次序、财富数量、朋友圈人数等。

例如，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 (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:

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

人口增加1倍, 加油站数量只多$2^{3/4}$ = 1.68 倍。规模越大，越有效。

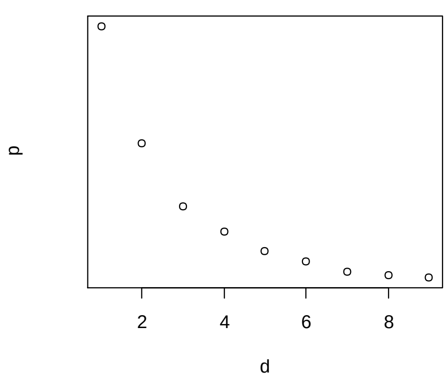
1. Benford定律 (首位数字)

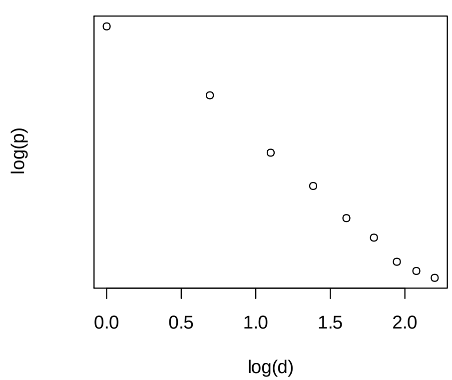
Newcomb(1881)， Benford (1938)发现“自然的”正数的首位非0数字并不是均匀分布，首位数字是1,2,⋯,9的概率依次下降，1的概率最大，即所谓Benford定律。“自然的数字”包括人口数、河流面积、财务报表、新闻中出现的数字等等。

320，0.032，3.2的首位非0数字都是3

例2.下表数据是2020年美国3144个县的人口数的首位数字(d) 的频率 (p)。

首位数字d	1	2	3	4	5	6	7	8	9
频数	957	581	381	298	237	204	172	161	153
频率p	0.304	0.185	0.121	0.095	0.075	0.065	0.055	0.051	0.049





(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 $\hat{a} = -1.153, \hat{b} = -0.874$

拟合得到的回归直线: $\log(p) = -1.153 - 0.874 \times \log(d)$

取指数函数得到拟合的幂次律：

$$p = p(d) = 0.316/d^{0.874}$$

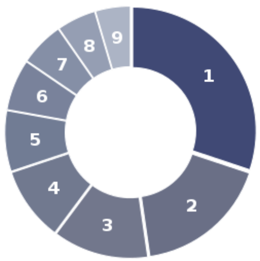
拟合值$p(1) = 0.316, p(2) = 0.316/2^{0.874} = 0.172, \ldots$

首位数字d	1	2	3	4	5	6	7	8	9
样本频率p	0.304	0.185	0.121	0.095	0.075	0.065	0.055	0.051	0.049
幂次律	0.316	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

Benford定律是一个严格的概率分布，不是数据分析得到的经验规律，而是在一定条件下经过严格的数学论证得到的分布。



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

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

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

该分布的假设是什么？有多种，我不确定，其中一个应该是：

若随机变量$x > 0, \log(x) \sim N(0,1)$, 则$x$的首位数字服从Benford定律。

Benford定律 ≈ 幂次律

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

应用: 通过Benford分布发现财务报表造假

2. 齐夫定律 Zipf’s law (体量与排名)

齐夫定律（Zipf‘s law）是由哈佛大学语言学家G.K.Zipf于1949年发现的单词使用频率与排名关系的经验幂次定律： 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：

$$p_k \propto 1/\, k^{\alpha}, k = rank，\alpha=1$$

一般的齐夫定律描述体量(size,可以是概率、体量等）与排名的幂次关系，刻画的是这样一种现象：排名越高，差距越大，比如第一名是第二名的2倍，第二名是第三名的1.25倍，…，排名靠后的差异很小。
例如size可以是财富榜的财富、金牌榜的金牌数目、河流面积等等。

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。

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

例3. 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，不一定等于1：

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

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

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

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

例4. 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

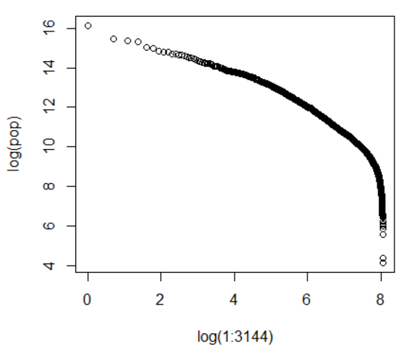
拟合线性模型

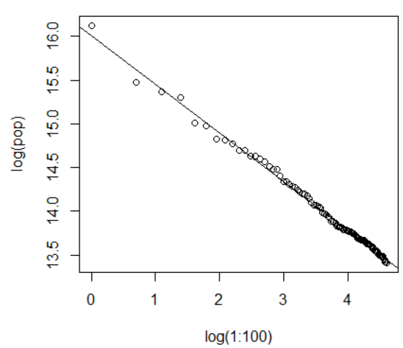
$\log(\text{Freq}) = 1.23 - 0.75 \times \log(\text{Rank})$，$\text{Freq} \propto \dfrac{1}{\text{Rank}^{0.75}}$，

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。

例2(续). 2020年美国3144个县的人口数与排名(rank)的关系，

前4个人口大县的人口数：10014042，5275522，4731129，4420574,...





拟合前100个县（右图）

$$\log(pop) = 16 - 0.56 \times \log(rank)$$

$$pop = c/rank^{0.56},\ c = \exp(16)$$

3. Kleiber定律与异速生长

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

Square - cube law（同速生长）：
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，比如眼睛生长较慢，而腿部生长较快，这称为异速生长(allometry)。

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。

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。

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。

Kleiber's law :

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

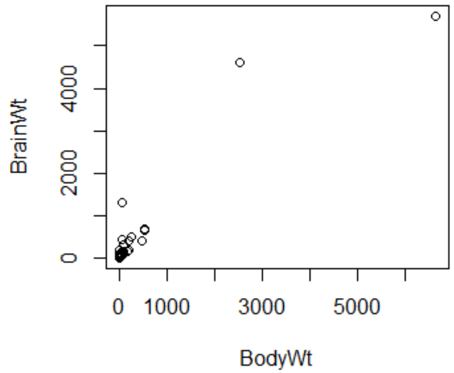
$$代谢速率 = 70 \times 体重^{3/4}, \quad 3/4 > 2/3$$

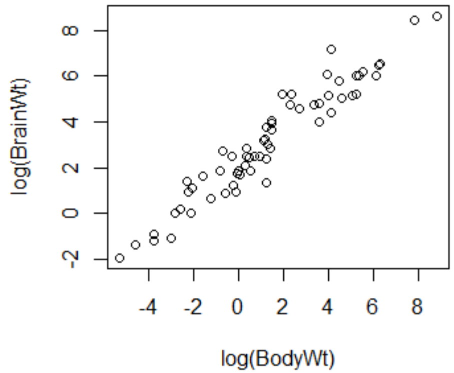
其中代谢速率单位为卡路里/秒,与表面积有关。

比如对于大体积动物，其表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，即要么增大表面积（大鸟的翅膀特别长），要么降低新陈代谢速度（大动物心跳慢、血液流动慢）。

再如， 小鸟的表面积足够大，所以可以快速扇动翅膀，并能把热量及时散发出去而保持正常体温。

例5.（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：

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

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

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

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

4. BMI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。**

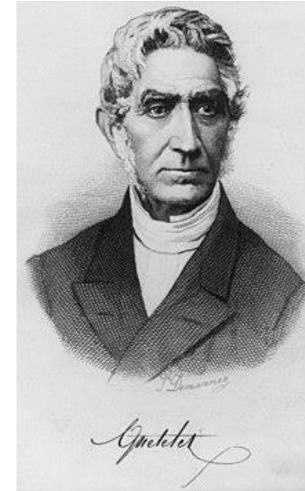
成年人体重指数BMI (Body Mass Index)定义为

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

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

from wiki

BMI是比利时/法国天文学家、数学家 L.A.J. Quetelet 在1830s提出的。法国人Quetelet和Laplace被认为是现代统计的先驱。



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），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。

为什么不是W除以H^3？文献中很难查到Quetelet的原始想法。

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（去相关化），且同分布(homogeneous)。

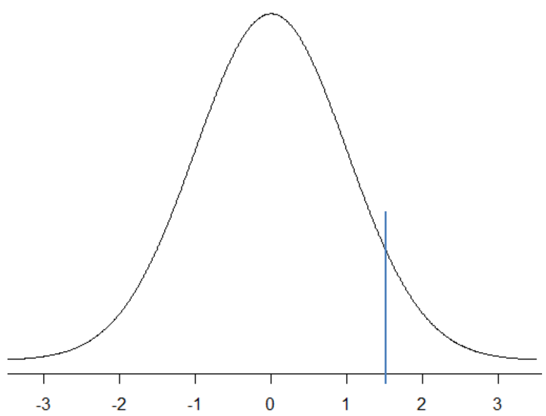
我们在对数尺度上建立线性模型：

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

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 $b \approx 2$.

例6. （身高与体重）

```
hw=read.table("http://staff.ustc.edu.cn/~ynyang/2023/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。

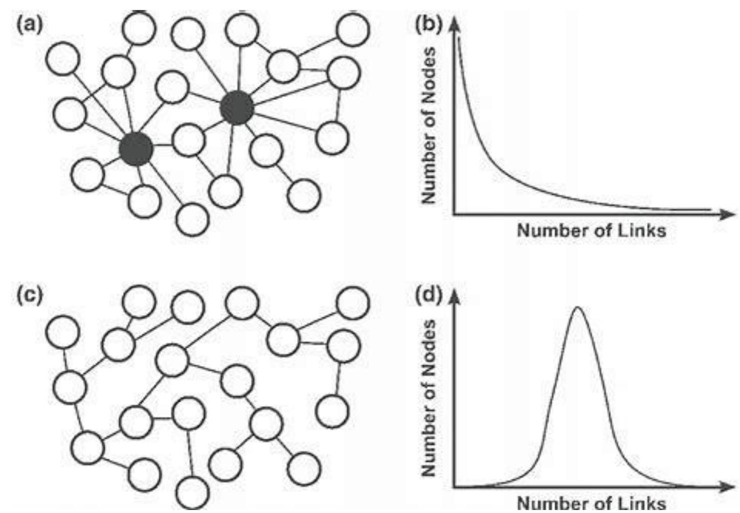
5. 无标度社交网络

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：

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

无标度网络，局部和整体看起来相像

随机网络，度数服从正态



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7 (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。