回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第九讲 简单线性模型的应用

2020.3.18

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

例子

- 异速增长
  - Kleiber's law
- Pareto分布：概率幂次律
  - 无标度社交网络
  - Zipf定律
  - Gutenberg-Richter定律
  - Benford定律
- 标准化/不相关化
  - BMI
- 预测

异速增长

异速生长（ **Allometry** ）是关于身体大小与形状 、解剖学、生理学及至行为 间关系的研究。 异速生长在统计形状分析及生物各部分相对生长率研究中是一个著名的研究论题。参见维基百科**wikipedia.org/wiki/Allometry**

等速增长(isometric scaling):

等速增长情形下，生物或物体(立方体、球等)的体积V正比于长度的3次方，表面积S正比于长度的3次方：

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

这称为square - cube law.

等速增长会使得表面积相对于体积显得太小，异速生长指的是面积与体积（或有关的量）的不成比例的生长关系。

异速增长现象(allometric scaling)

英国人J.B.S. Haldane发现，动物种群之间或种群内部存在异速增长现象, 比如大象不是老鼠的比例扩大版 : 大象的骨骼面积和强度的增加比例远大于体重的增加比例。Haldane的论文"On Being the Right Size", 研究了动物体积(size)的变化时，形状(shape)的变化规律。

例如：

- 动物承重能力与腿的截面积成正比，如果是同速生长,则体积/重量增加1倍时动物承重能力(腿的截面积)只增加为原来的 $2^{2/3}=1.59<2$倍。所以为了生存大动物的腿相对于小动物必须是异速增长的.

- 如果鸟的翅膀与体重同速生长，为了飞行，需要增加翅膀扇动频率（但这导致心脏和呼吸负担加重），或者需增大翅膀面积（异速生长）。

- 质量为M(假设$\mathrm{M} \propto V$)的飞行物以加速度a飞行，压力$F = Ma$,
  如果等速增长(表面积$A \propto V^{2/3} \propto M^{2/3}$),表面单位承受压力

$$T = F/A = Ma/A \propto M^{1/3}a$$

  当质量M增加1倍时，表面单位承受压力为原来的$2^{1/3} = 1.26 > 1$倍。
  这说明了为什么大飞机难以制造(减小材料密度,增大材料抗压能力，异速增大表面积）。

异速增长与幂次律

异速增长现象通常可以用幂次律表达：

$$y = kx^r \quad (\text{幂次} r \neq 2/3 \text{代表异速增长})$$

其中, x是体积或与体积有关的度量，y是与面积有关的量.

最著名的例子是Kleiber's law：

$$\text{Kleiber's law}: \quad \text{Metabloic\_Rate} = 70 \times \text{Mass}^r, \quad r = 3/4$$

其中，Metabloci_Rate为动物新陈代谢速率(卡路里/秒), M为体重.

$r > 2/3$ 说明动物代谢率是异速增长的，比同速增长情形代谢率要高;
但 $r < 1$，动物的散热量与体表面积成正比，故为了维持一定的体温,
大动物心跳慢, 血液流速也慢.

社会经济中也广泛存在类似于Kleiber ‘s law 的幂次定律，比如，城市交通流量、森林大火面积、河流面积等

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 (*c*) 与城市人口规模(*s*， *size*)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。

例1.（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，在对数尺度上拟合简单线性模型得到斜率估计为$0.7517 = 3/4$,截距项2.1348，回归直线的估计为

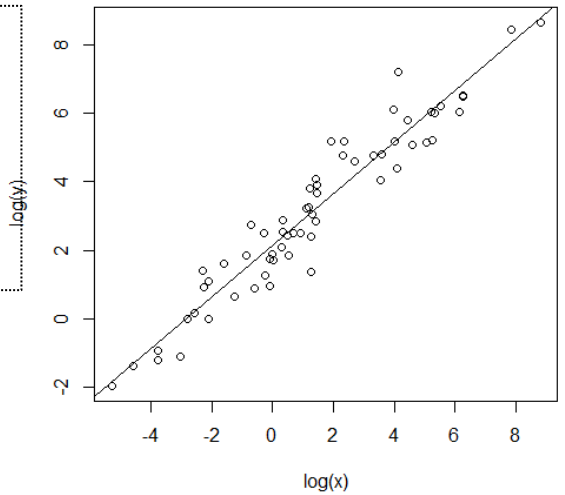
$$\log(\text{BrainWt}) = 2.1348 + 0.7517\log(\text{BodyWt})$$

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

R 命令：

```
library(alr3)
x=brains[,2]
y=brains[,1]
plot(x,y)
plot(log(x), log(y) )

myfit=lm(log(y)~log(x))
abline(myfit)
```

```
> myfit

Coefficients:
(Intercept)      log(x)
     2.1348      0.7517
```



```
> summary(myfit)

Call:
lm(formula = log(y) ~ log(x))

Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43598  1.94833

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
log(x)       0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

$\hat{b} = 0.75169, se(\hat{b}) = 0.02846, t = \hat{b} / se(\hat{b}) = 26.41$

截距a

斜率b

$\hat{\sigma} = 0.6943$

$R^2 = 0.9208$

Pareto 分布：概率幂次律

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

Pareto 分布相对于正态的指数阶尾概率, 是一种重尾、长尾、无标度分布(heavy - tailed, long - tailed, scale - free).

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。

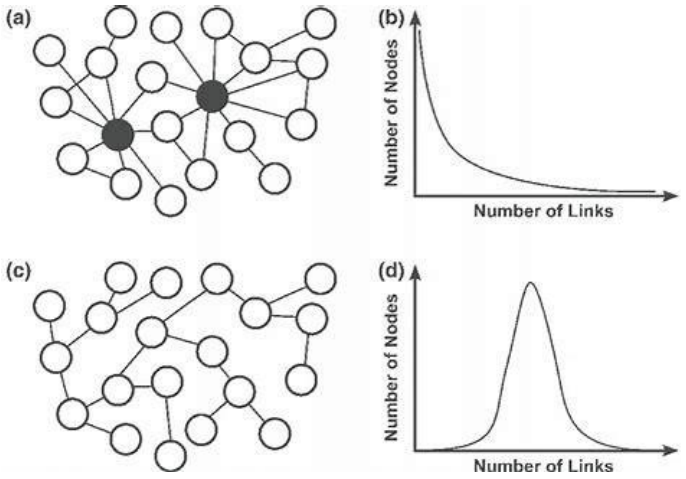
幂次律

1. 无标度社交网络（scale-free social network）

社交网络中大多数成员有较少的连接，而少数点有较多的连接。成员的度(连接的节点数目k)的分布服从幂次律/Pareto分布：

$$P(k) \propto k^{-r}，\ r在2和3之间$$

图(a)是无标度网络，两个黑点具有较多连接的检点。



2. 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 p_k 与 k 成反比

$$p_k \propto k^{-r}, \quad r=1$$

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

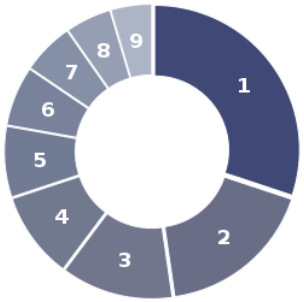
作者判定：不同的作者可能服从不同的幂次律 (不同的 r)。

3. Gutenberg-Richter（里氏）定律

M 级地震发生的频率N与能量10^M成反比

$N \propto \left(10^M\right)^{-r}$，$r$在0.5和1.5之间，刻画地震活跃程度,正常地区为1

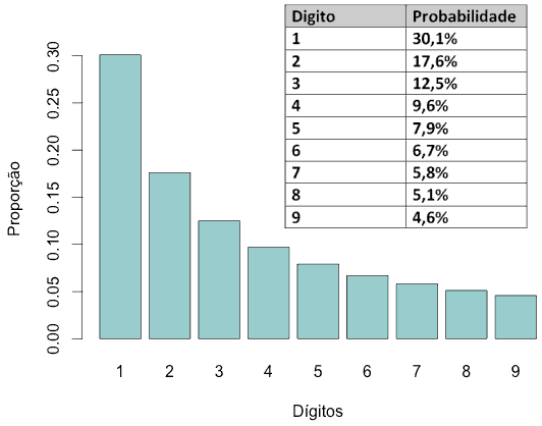
4. Benford定律



Newcom(1881), Benford(1938）发现“自然出现的”数字，比如新闻中的数字，人口数字，河流面积等首位数字的分布并非均匀，首位数字是1,2,…,9的概率依次下降。在一定理论假设下，有如下分布：

Benford定律

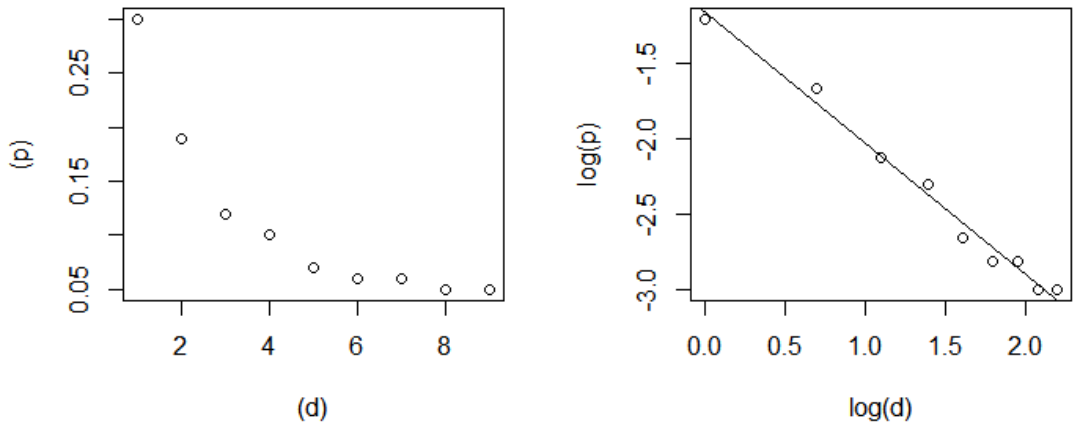
首位数字是d的概率为：$p(d)=\log_{10}(1+1/d)$



例2. Benford定律给出了一定理论假设下首位数的规律表格，并不是幂次律。对于右边表格数据，美国3142个县的人口数的首位数字(d) 的频率 (p)，我们拟合幂次律

d	个数	p
1	956	30%
2	593	19%
3	380	12%
4	301	10%
5	225	7%
6	203	6%
7	177	6%
8	159	5%
9	148	5%

(d,p)散点图并非线性，但取对数之后基本线性：




```
d=1:9；  p=c(0.30, 0.19, 0.12, 0.10, 0.07, 0.06, 0.06, 0.05, 0.05)
plot(d,p);   plot(log(d),log(p));
```

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

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

拟合得到的回归直线：

$$\log(p) = -1.16 - 0.87 \times \log(d) \Rightarrow p = p(d) = 0.31/d^{0.87}$$

```
Myfit = lm(log(p)~log(d))
abline(Myfit)
```

下表第3行给出了由上述幂次律得到的拟合值$p(d) = 0.31/d^{0.87}$，
第4行是Benford公式$p(d) = \log_{10}(1+1/d)$。

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

标准化（不相关化）

一个指标如果在群体中不是一致的（不是同分布的），即与其它变量有关，那么该指标仅仅适用于一致的子群体，为了适用于全部群体，需要校正或消除其它变量的影响（标准化），回归模型中的误差即是一个标准化的量。

例3.（体重指数）如何判断体重是否超标？需要定义一个体重指数/度量 B，假设服从正态，一个人的指数B如果超过95%的人，则认为体重超标。显然B取为体重是不恰当的，因为身高以及其它因素包括性别、年龄等与体重有关。

为了定义一个具有普适性的指数, 需要消除其它因素(比如身高H，性别S)的影响。下面我们仅考虑成年男性

假设身高*H*的成年男性的体重*W在对数尺度上*满足模型：

$$\log(W) \sim N\left(\mu,\ \sigma^2\right),\quad \mu = \mu(\mathrm{H})\text{与H有关，}$$

我们假设线性模型：$\mu = a + b\log(H)$

此即线性模型：$\log(W) = a + b\log(H) + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$

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

log(W)的分布不是一个一致的(homogeneous)总体（均值不为常数的正态分布），但z的分布是一个单一总体,因而具有普适性.

若$z > C$ (比如C = 1.65，标准正态的上95%分位数)，可定义为偏胖

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

W / H^b 即可作为校正了身高影响之后的体重指数,经验数据表明$b \approx 2$

故定义体重指数(Body Mass Index)：$BMI = \frac{W}{H^2}$，单位kg/m^2

```
hw=read.table("http://staff.ustc.edu.cn/~ynyang/lm2020/lab/height-weight.txt",head=T)

sex=hw[,1]
hw[sex==1,]->male
lm(log(weight)~log(height) ,data=male)->myfit
```

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

类别由标准正态分布分位数决定,而阈值由a,σ及$\mathrm{N}(0,1)$分位数决定

from wiki

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。

首先需要确定预测变量.

主办国的表现与上一届的表现有关

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)。

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

R 命令：

> lm (y ~ x, data = data.name)

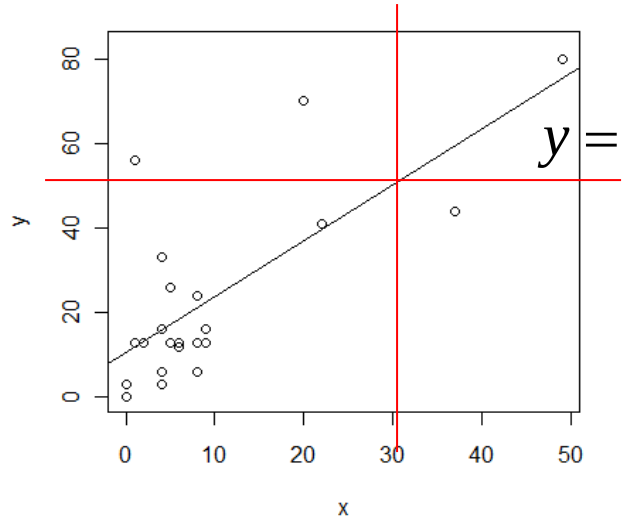
 # lm: linear model

```
> myfit =  lm(y~x, data=gold)
> summary(myfit)
Residuals:
   Min     1Q  Median     3Q    Max
-15.540  -9.192  -4.895   2.422  44.054

Coefficients:   β̂          sd(β̂)      t = β̂ / sd(β̂)      p值
              Estimate   Std. Error    t value          Pr(>|t|)
(Intercept)  10.6236     4.1617        2.553            0.019 *
x             1.3221     0.2707        4.884            8.98e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.21 on 20 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5439,    Adjusted R-squared:  0.5211
F-statistic: 23.85 on 1 and 20 DF,  p-value: 8.979e-05
```

$\hat{\sigma} = 15.21,$

$df = n - 2 = 20$

复相关系数平方 : $R^2 = 0.5439$



$y = 10.6 + 1.3x$

上一届(2004)中国金牌数$x = 32$,

预测08届金牌数：

$$10.6 + 1.3 \times 32 = 52$$