第 6 章离散概率模型

韩建伟

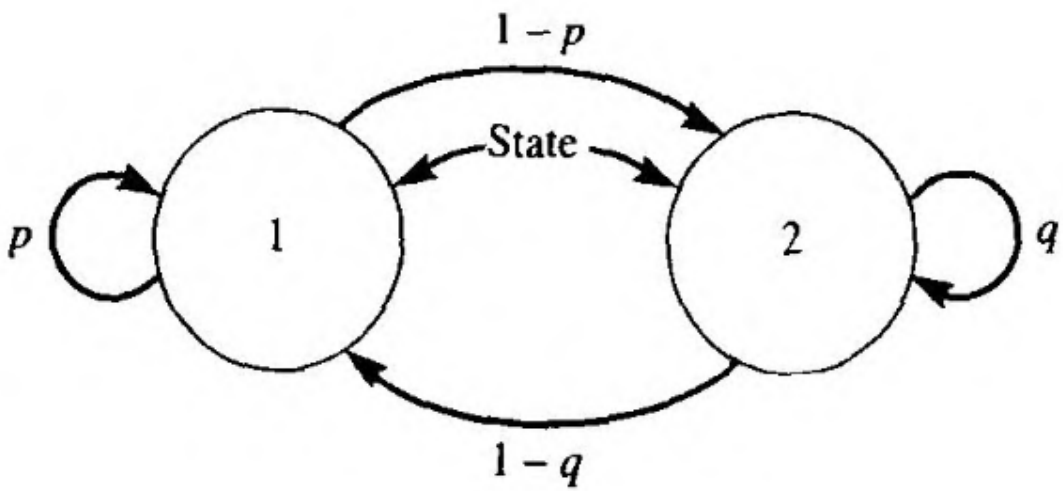
信息学院
mm@hanjianwei.com

2017/11/17

离散概率模型

马尔可夫链

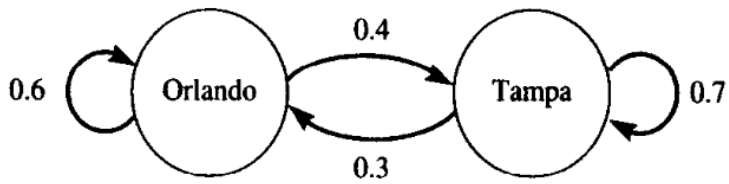
系统可以从一个状态转移到另外一个，每个时段转移一次，并且这种向每个可能结果的转移存在一定的概率。



再论汽车租赁公司

		Next state	
		Orlando	Tampa
Present state	Orlando	0.6	0.4
	Tampa	0.3	0.7

(a) 转移矩阵



(b) 二状态马尔可夫链

$$p_{n+1} = 0.6p_n + 0.3q_n$$

$$q_{n+1} = 0.4p_n + 0.7q_n$$

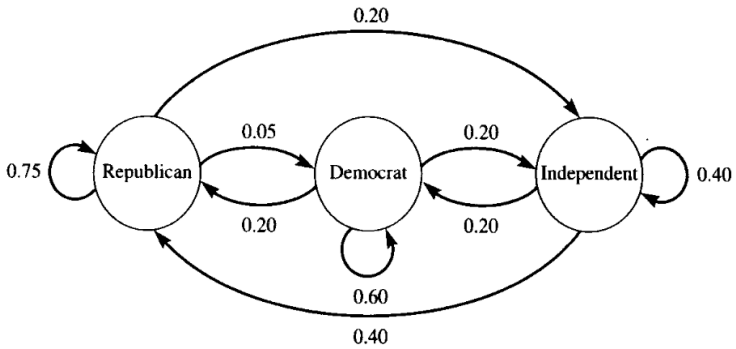
$$p_k \to 3/7 = 0.428571$$

$$q_k \to 4/7 = 0.571429$$

投票趋势

Present state \ Next state	Republicans	Democrats	Independents
Republicans	0.75	0.05	0.20
Democrats	0.20	0.60	0.20
Independents	0.40	0.20	0.40

(a) 转移矩阵



(b) 三状态马尔可夫链

$$R_{n+1} = 0.75R_n + 0.20D_n + 0.40I_n$$

$$D_{n+1} = 0.05R_n + 0.60D_n + 0.20I_n$$

$$I_{n+1} = 0.20R_n + 0.20D_n + 0.40I_n$$

马尔可夫链

定义

1. 一个事件有有限多个结果，称为状态，该过程总是这些状态中的一个
2. 在过程的每个阶段或者时段，一个特定的结果可以从它现在的状态转移到任何其它状态，或者保持原状态
3. 每个阶段从一个状态转移到其他状态的概率用一个转移矩阵表示，矩阵每行的各个元素在 0 到 1 之间，每行的和为 1, 这些概率只取决于当前状态，而与过去状态无关

部件和系统的可靠性建模

$f(t)$ 一个零件、部件或系统在时间 t 内的失效率 (概率分布)

$F(t)$ 相应于 $f(t)$ 的累计分布函数

$R(t)$ 一个零件、部件或系统的可靠性，$R(t) = 1 - F(t)$

串联系统

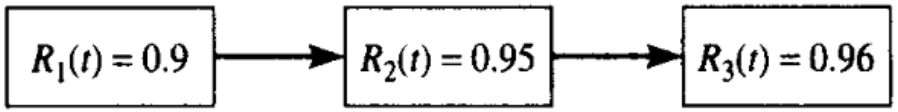


图: NASA 宇宙火箭推进系统

$$R_s(t) = R_1(t)R_2(t)R_3(t) = (0.90)(0.95)(0.96) = 0.8208$$

并联系统

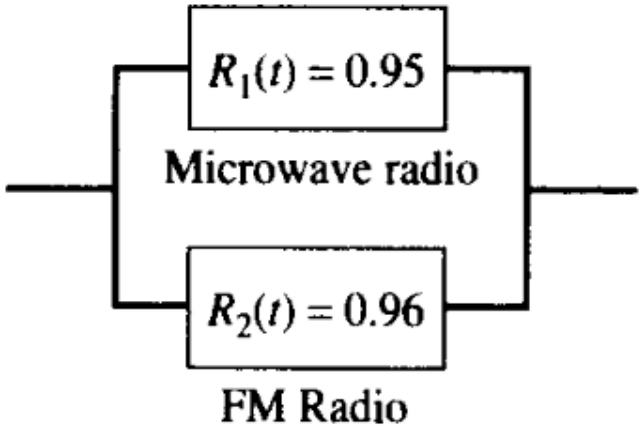


图: NASA 宇宙火箭通讯系统

$$R_s(t) = R_1(t) + R_2(t) - R_1(t)R_2(t) = 0.95 + 0.96 - (0.95)(0.96) = 0.998$$

串并联组合系统

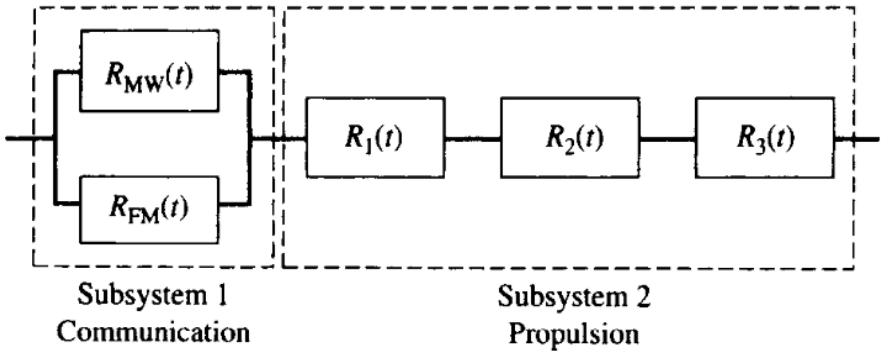


图: NASA 可控宇宙火箭推进点火系统

$$R_s(t) = R_{s_1}(t)R_{s_2}(t) = (0.998)(0.8208) = 0.8192$$

线性回归

线性回归 一种偏差平方和最小化的统计方法

1. 阐述基本的线性回归模型和它的假设
2. 定义并解释统计量 R^2
3. 利用检查和解释残差散点图对拟合线性回归模型做图形说明

统计量

基本的线性回归模型定义为 $y_i = ax_i + b$, y_i 的平均值记作 $\bar{y}$, 则:

误差平方和

$$SSE = \sum_{i=1}^{m} [y_i - (ax_i + b)]^2$$

总修正平方和

$$SST = \sum_{i=1}^{m} [y_i - \bar{y}]^2$$

回归平方和

$$SSR = \sum_{i=1}^{m} [\bar{y} - (ax_i + b)]^2 = SST - SSE \geq 0$$

决定系数

$$R^2 = 1 - \frac{SSE}{SST}$$

$0 \leq R^2 \leq 1$, 如果 $R^2 = 1$，那么数据精确地与回归直线吻合

1. R^2 的大小与两个自变量哪一个记作 x、哪一个记作 y 无关
2. R^2 的大小与 x, y 的单位无关

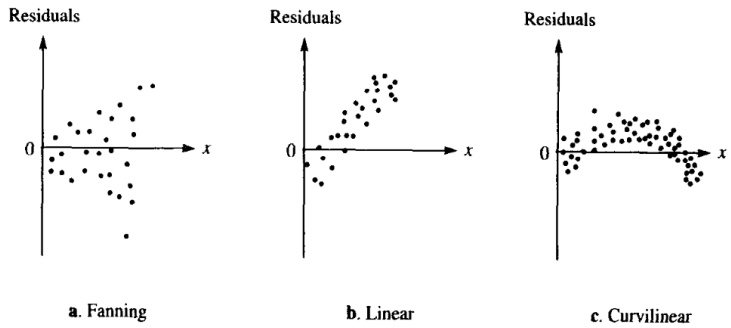
残差

残差是实际值和测量值之间的误差：

$$r_i = y_i - f(x_i) = y_i - (ax_i + b)$$

如果将残差对于自变量做图，会得到一些有价值的信息：

1. 残差应随机地分布在与数据精度同量级的、相当小的界限内
2. 遇到特别大的残差时，应对相应的数据点做进一步的研究，去发现其原因
3. 残差的模式或者趋势指出了可预测的影响因素仍有待建模，模式的性质常可提供使模型更精确的线索

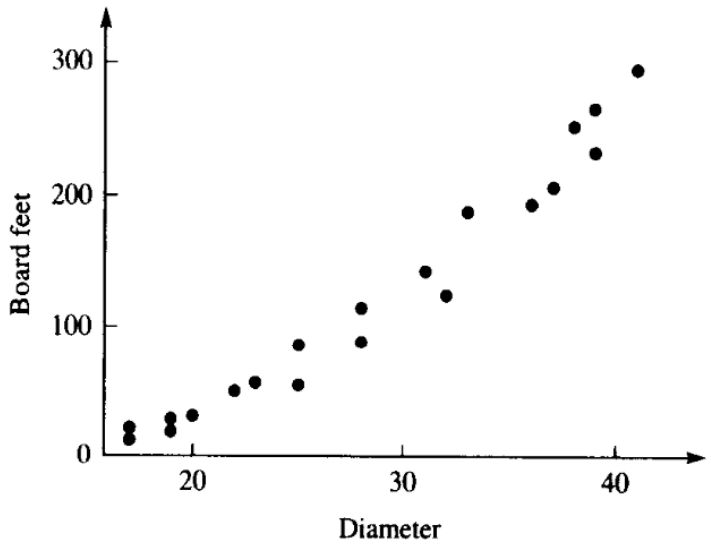


a. Fanning　　b. Linear　　c. Curvilinear

美国黄松

Diameter (in.)	Board feet
36	192
28	113
28	88
41	294
19	28
32	123
22	51
38	252
25	56
17	16
31	141
20	32
25	86
19	21
39	231
33	187
17	22
37	205
23	57
39	265

(a) 数据



(b) 散点图

建模

由几何相似性得到比例关系：

$$V \propto d^3$$

其中 d 是树的直径，如果假定高度相同，则：

$$V \propto d^2$$

假设树的根部的体积是常数那么上面两个模型进一步精细为：

$$V = ad^3 + b$$

$$V = \alpha d^2 + \beta$$

模型求解及分析

用计算机做 4 个模型的线性回归得到以下解:

$$V = 0.00431d^3$$

$$V = 0.00426d^3 + 2.08$$

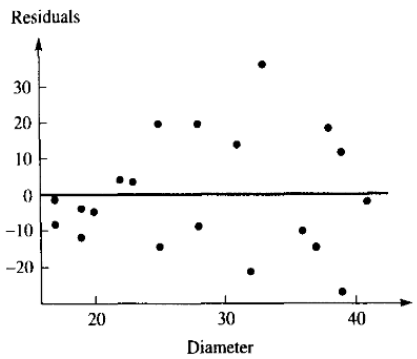
$$V = 0.152d^2$$

$$V = 0.194d^2 - 45.7$$

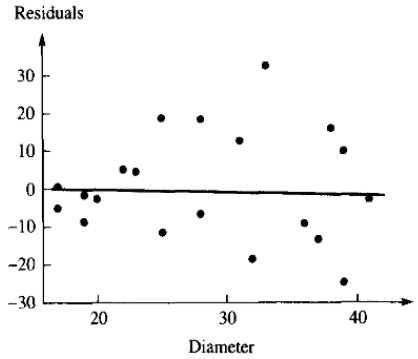
Model	SSE	SSR	SST	R^2
$V = 0.00431d^3$	3,742	458,536	462,278	0.9919
$V = 0.00426d^3 + 2.08$	3,712	155,986	159,698	0.977
$V = 0.152d^2$	12,895	449,383	462,278	0.9721
$V = 0.194d^2 - 45.7$	3,910	155,788	159,698	0.976

图: 四个回归模型的主要信息

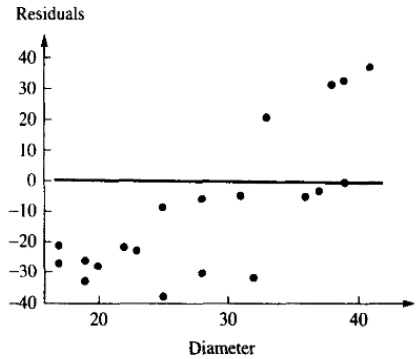
误差分析



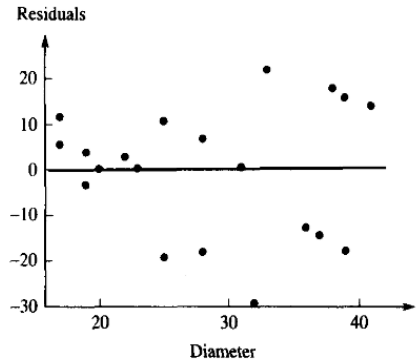
a. Residual plot for board feet = $0.00431d^3$; model appears adequate because no apparent trend appears in the residual plot



b. Residual plot for board feet = $0.00426d^3 + 2.08$; the model appears adequate because no trend appears in the plot



c. Residual plot for the model board feet = $0.152d^2$; model does not appear to be adequate because there is a linear trend in the residual plot



d. Residual plot for the model board feet = $0.194d^2 - 45.7$; the model appears to be adequate because there is no apparent trend in the residual plot

Matlab 中的线性回归

- [b, bint, r, rint, stats] = regress(y, x)是线性回归函数
- recplot(r, rint)做残差分析图