大数据的统计学基础——第14周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

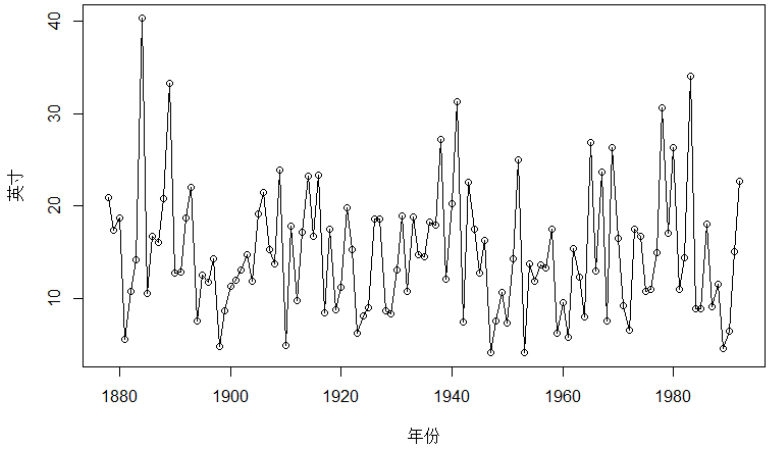
http://edu.dataguru.cn

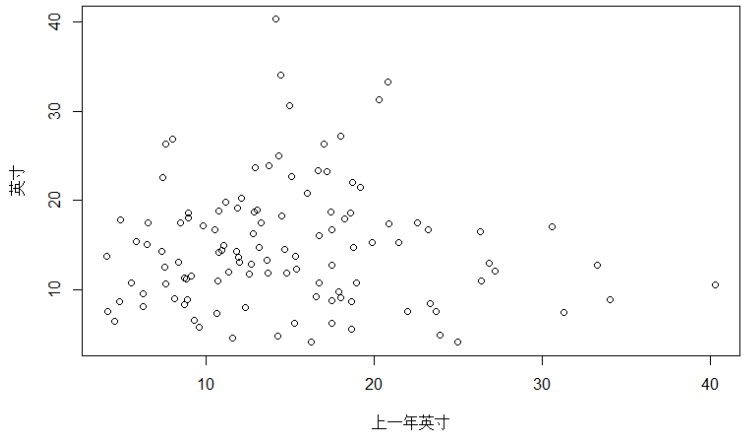
关注炼数成金企业微信

■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

时间序列举例

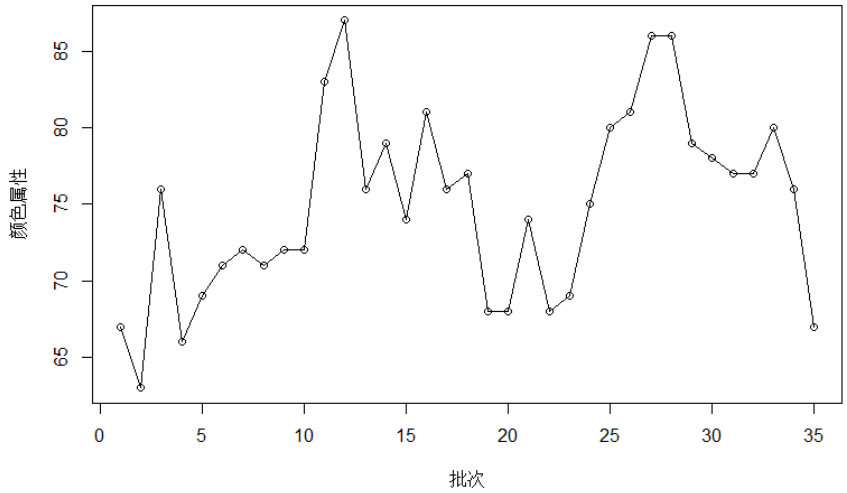
- 洛杉矶年降水量

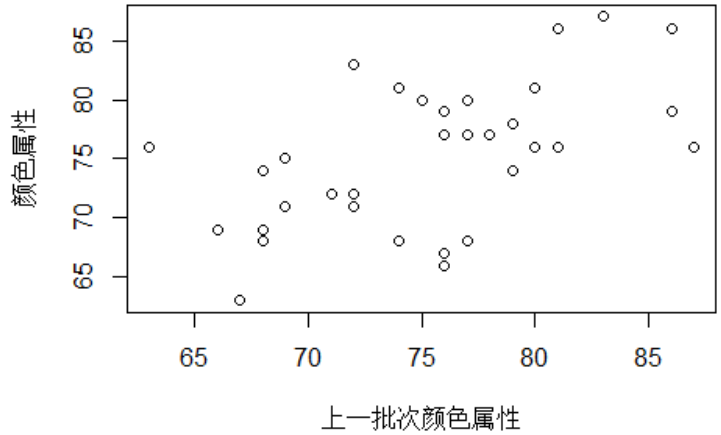




时间序列举例

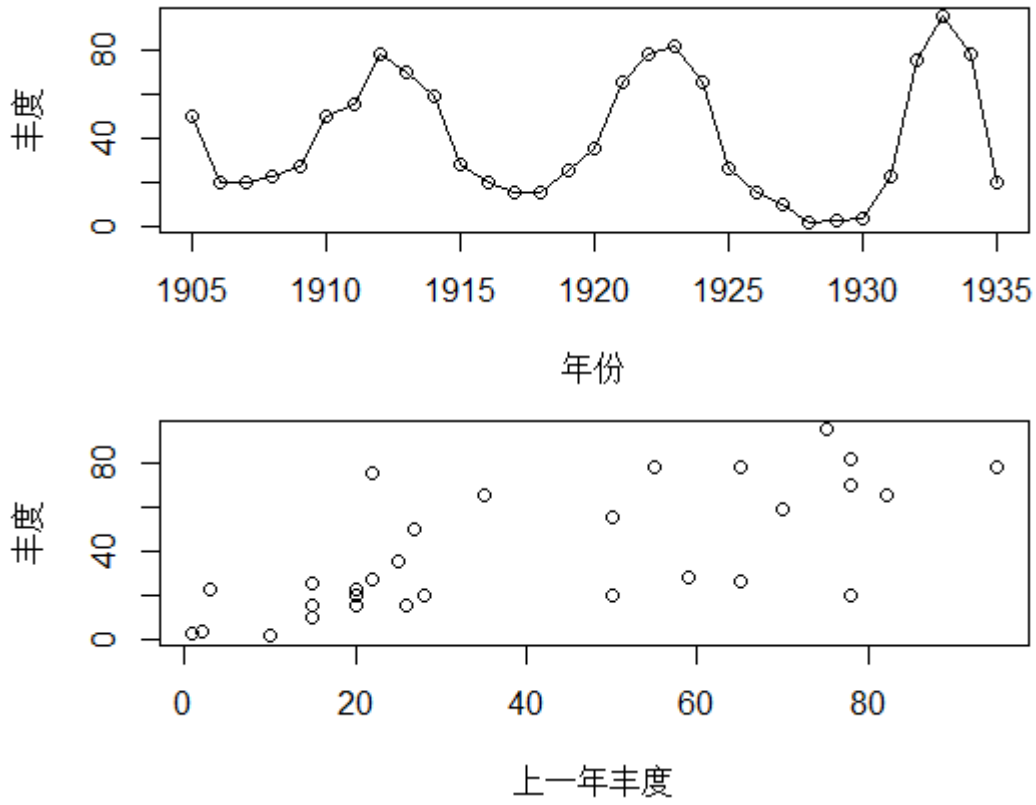
- 化工过程


颜色属性
65
70
75
80
85
0
5
10
15
20
25
30
35
批次


颜色属性
65
70
75
80
85
65
70
75
80
85
上一批次颜色属性

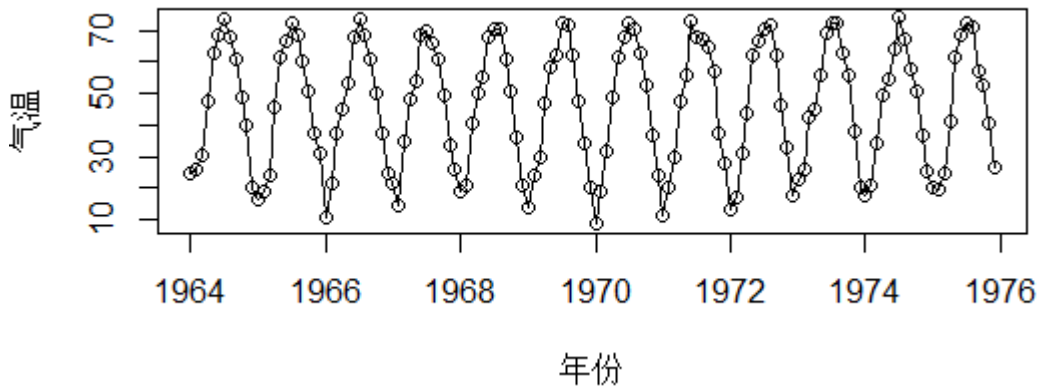
时间序列举例

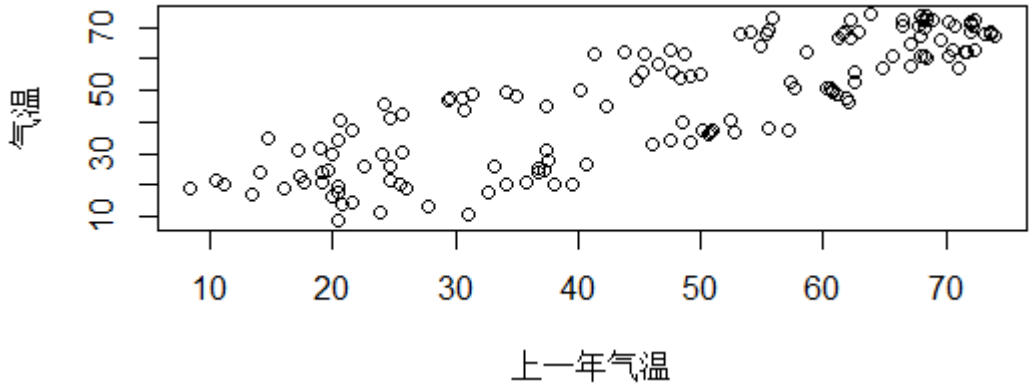
- 加拿大野兔年丰度



时间序列举例

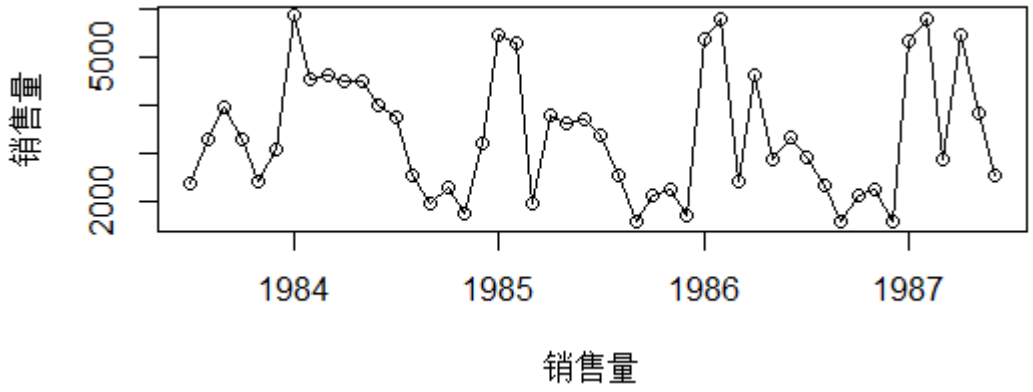
- 某市月平均气温

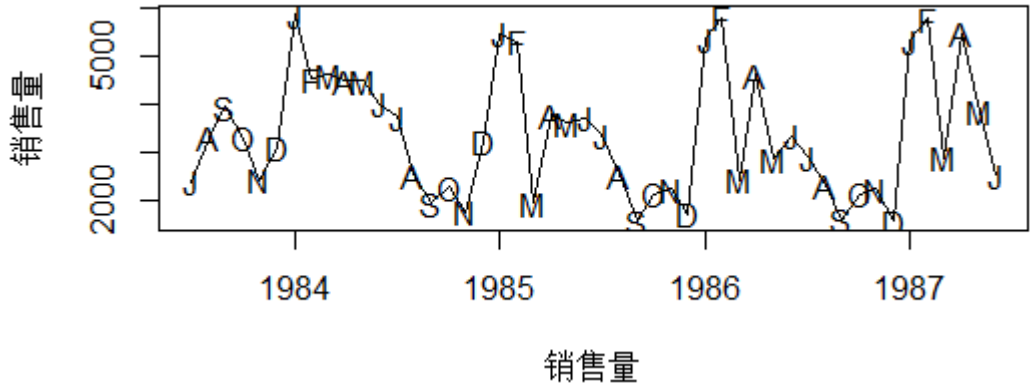




时间序列举例

- 滤油器月销售量





时间序列分类

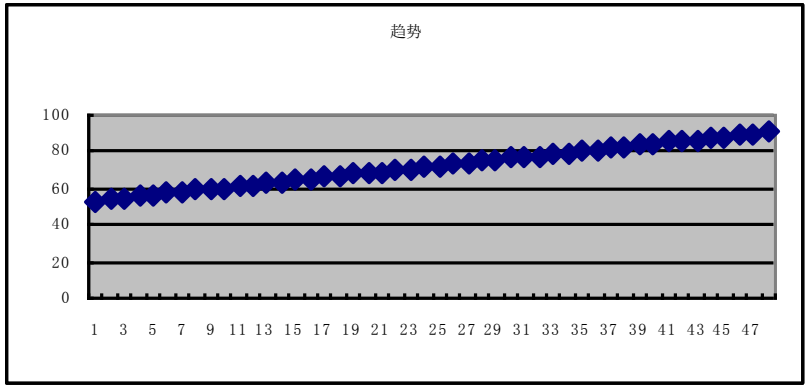
- 所研究的对象的多少
  - 一元
  - 多元
- 时间的连续性
  - 连续
  - 离散
- 序列的统计特性
  - 平稳时间序列
  - 非平稳时间序列

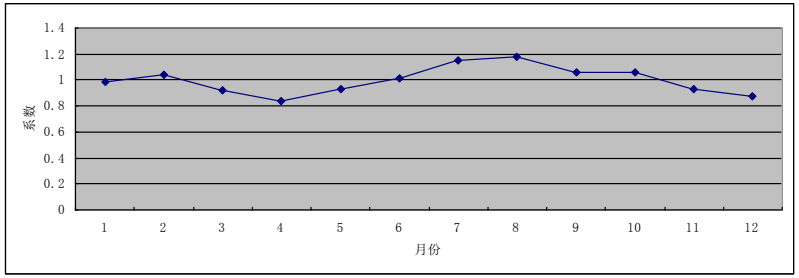
时间序列分析方法

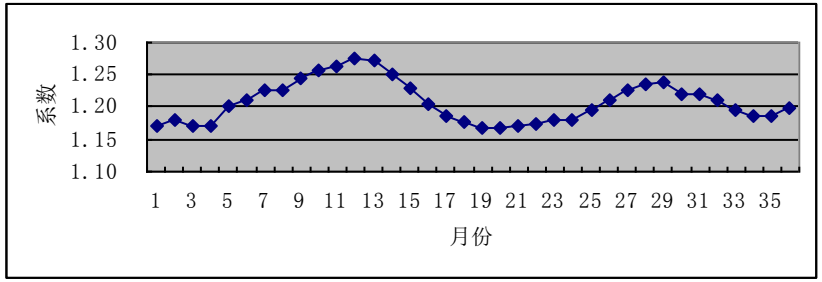
- 随机性时间序列分析
  - 自回归模型(AR)
  - 滑动平均模型(MA)
  - 自回归滑动平均模型(ARMA)
  - 差分自回归滑动平均模型(ARIMA)
- 确定性时间序列分析
  - 趋势预测法
  - 平滑预测法
  - 分解分析法

时间序列影响因素

- 时间序列影响因素：
- 1. 长期趋势Trend
- 2. 循环变动\周期性Cyclic
- 3. 季节性变化Seasonal variation
- 4. 不规则变化Irregular movement







影响因素的叠加

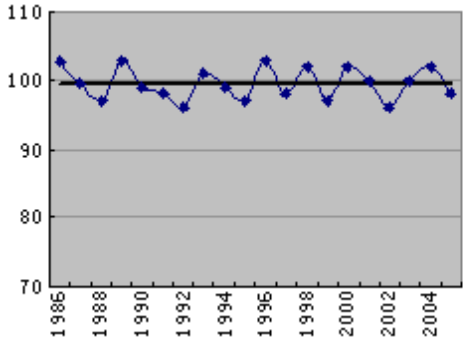


图1.1　平稳序列

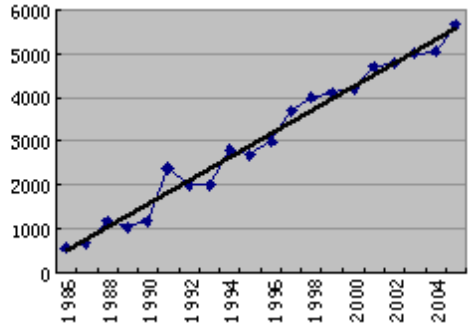


图1.2　趋势序列

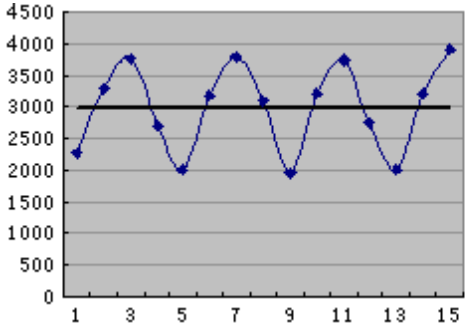


图1.3　季节型序列

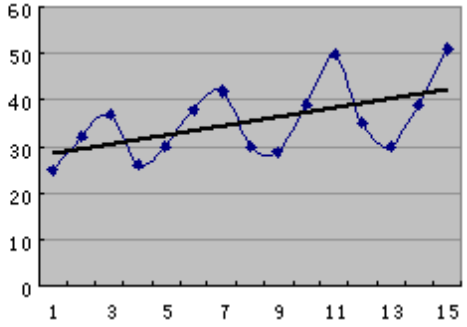


图1.4　含有季节与趋势因素的序列

时间序列分析模型

- 乘法模型：Y=T*S*C*I
- 加法模型：Y=T+S+C+I

时间序列

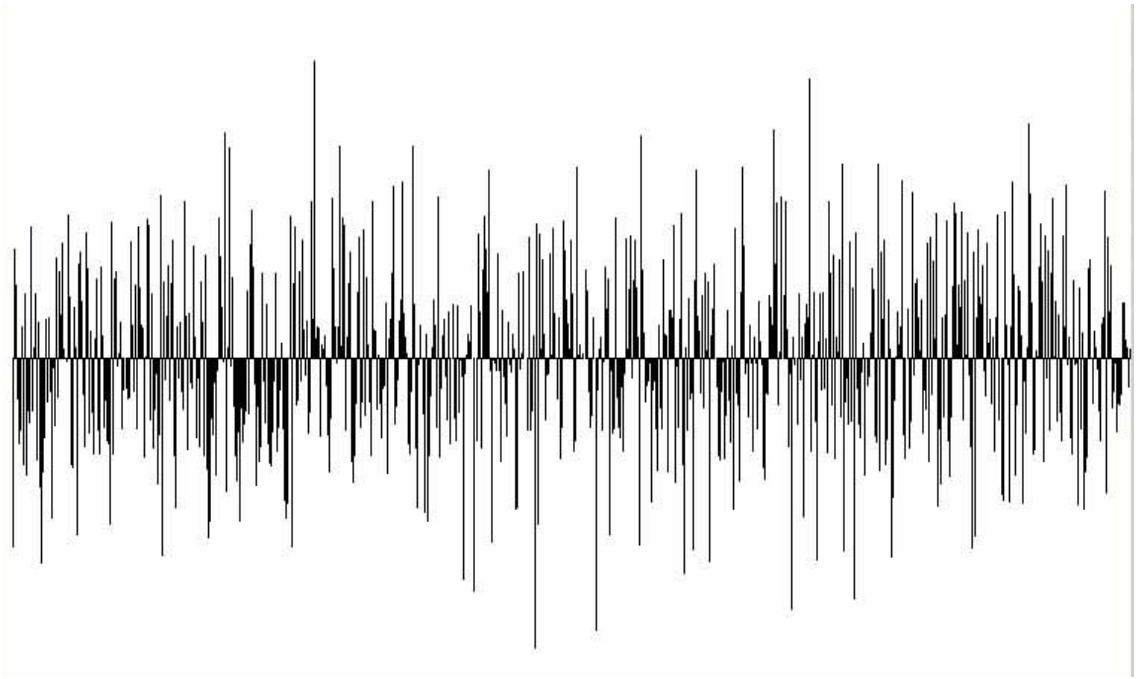
- 随机变量序列$\{Y_t: t = 0, 1, 2, \ldots\ldots\}$称为一个时间序列模型。
- 均值函数：$\mu_t = E(Y_t), t = 0, 1, 2, \ldots\ldots$
- 自协方差函数：$\gamma_{t,s} = Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s, t, s = 0, 1, 2, \ldots\ldots$
- 自相关函数：$\rho_{s,t} = Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t) Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$

平稳性

- 假定时间序列$\{Y_t: t = 0, \pm 1, \pm 2, \ldots\ldots\}$的每一个数值都是从一个概率分布中随机得到，如果满足下列条件：
  - 1）均值$E(Y_t) = \mu$与时间t无关的常数；
  - 2）方差$Var(Y_t) = \gamma$与时间t 无关的常数；
  - 3）协方差$Cov(Y_t, Y_{t+k}) = \gamma_{0,k}$只与时期间隔k 有关，与时间t 无关的常数。
- 则称该随机时间序列是平稳的（stationary)

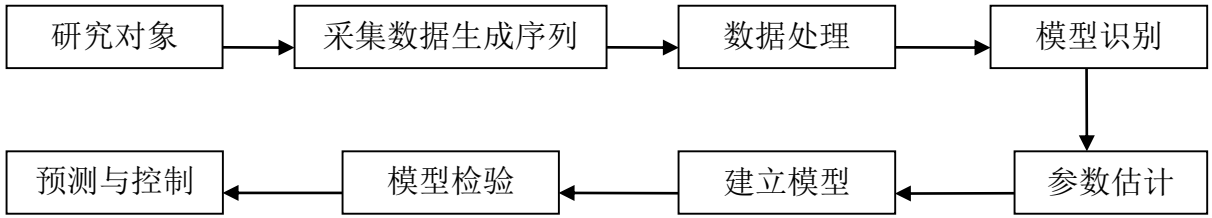
白噪声(White Noise)

- 纯随机过程
- 白噪声



平稳时间序列分析

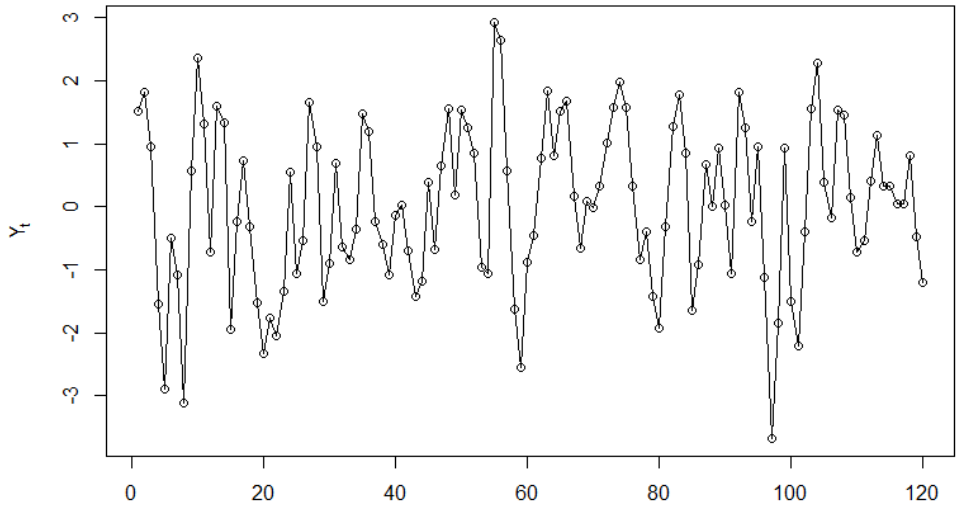
- 自回归模型(AR)
- 滑动平均模型(MA)
- 自回归滑动平均模型(ARMA)

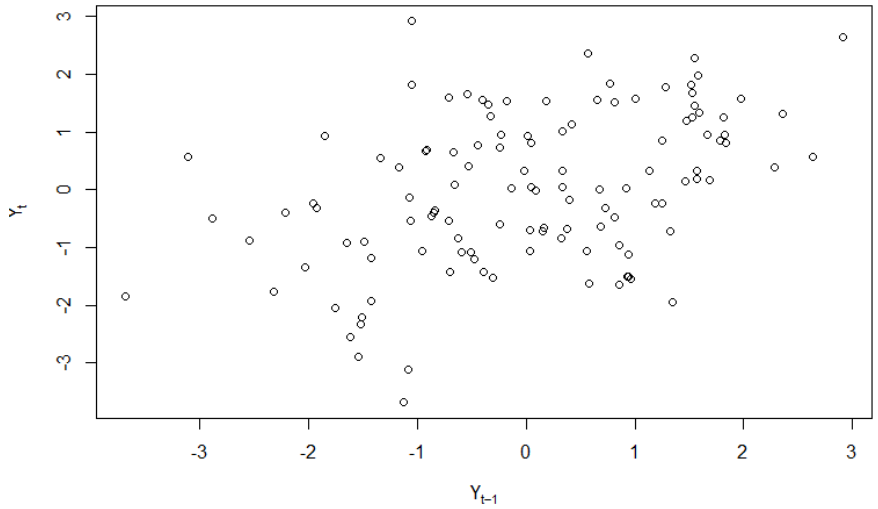


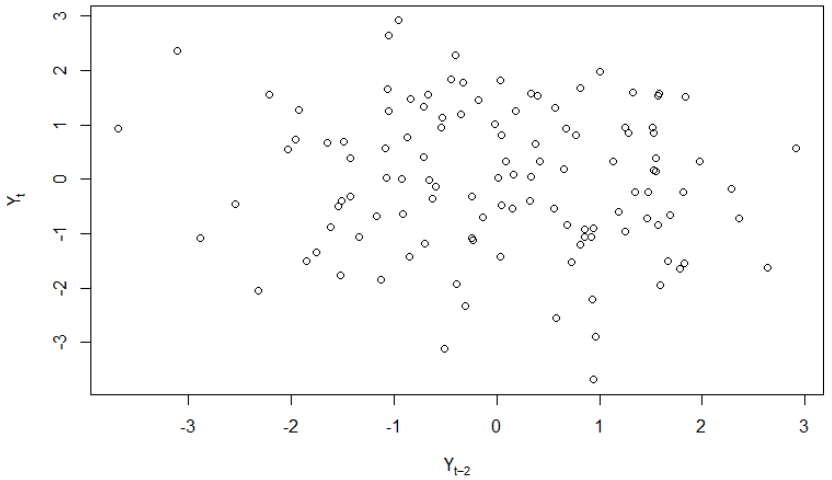
自回归模型(AR)

- $$\begin{cases} Y_t = \varphi_1 Y_{t-1} + + \cdots \ldots + \varphi_p Y_{t-p} + \varepsilon_t \\ \varphi_p \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(Y_s \varepsilon_t) = 0, \forall s < t \end{cases}$$
- 称上述模型为p阶自回归模型——AR(p)
- 中心化模型

例子







滑动平均模型(MA)

- $$\begin{cases} Y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots \ldots - \theta_q\varepsilon_{t-q} \\ \theta_p \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, s \neq t \end{cases}$$
- 称上述模型为q阶滑动平均模型——MA(q)
- 中心化模型

自回归滑动平均模型(ARMA)

- $$\begin{cases} Y_t = \varphi_1 Y_{t-1} + + \cdots \ldots + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots \ldots - \theta_q \varepsilon_{t-q} \\ \varphi_p \neq 0, \theta_p \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(Y_s \varepsilon_t) = 0, \forall s < t \end{cases}$$
- 称上述模型为自回归滑动平均模型——ARMA(p,q)

非平稳时间序列

- 随机性时间序列分析
  - 差分自回归滑动平均模型(ARIMA)
- 确定性时间序列分析
  - 平滑预测法
  - 趋势预测法
  - 分解分析法

ARIMA模型

- ARIMA(p,d,q)
- p——自回归阶数
- d——差分阶数
- q——移动平均阶数
- 通过差分运算讲非平稳时间序列模型转化为平稳时间序列模型

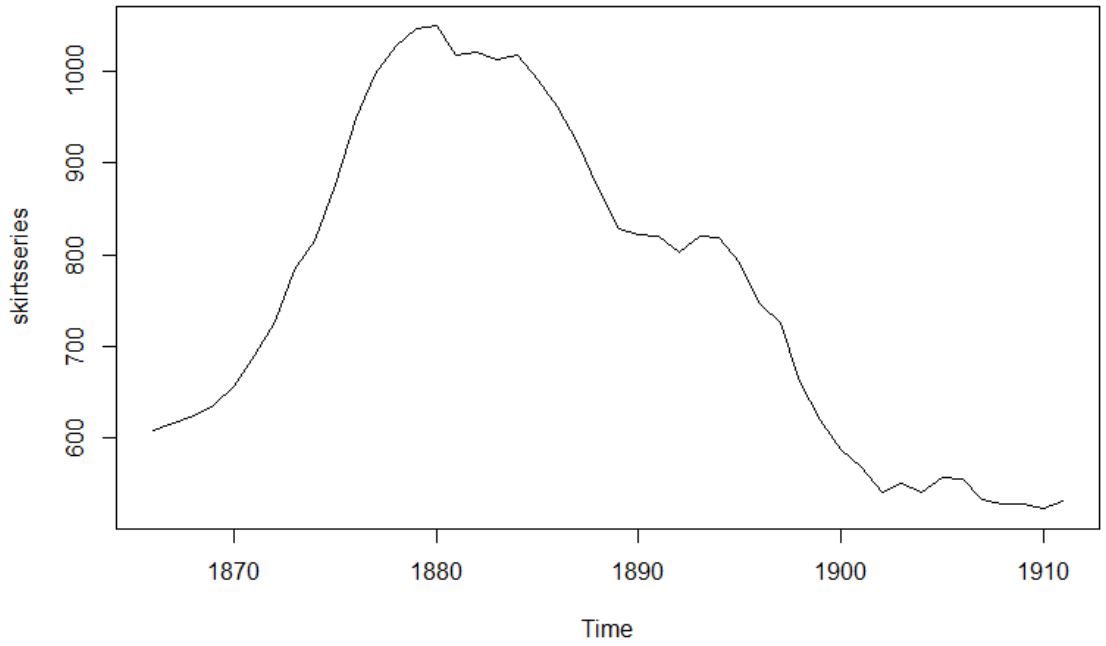
差分

- 对于时间序列$\{Y_t: t = 0, 1, 2, \ldots\ldots\}$
- 1阶差分运算：$\nabla X_t = X_t - X_{t-1}$
- 2阶差分运算：$\nabla^2 X_t = \nabla X_t - \nabla X_{t-1}$
- ……
- p阶差分运算：$\nabla^p X_t = \nabla^{p-1} X_t - \nabla^{p-1} X_{t-1}$

- k步差分：$\nabla_k X_t = X_t - X_{t-k}$

例子

- 1866年到1911年每年女人们裙子的直径——非平稳



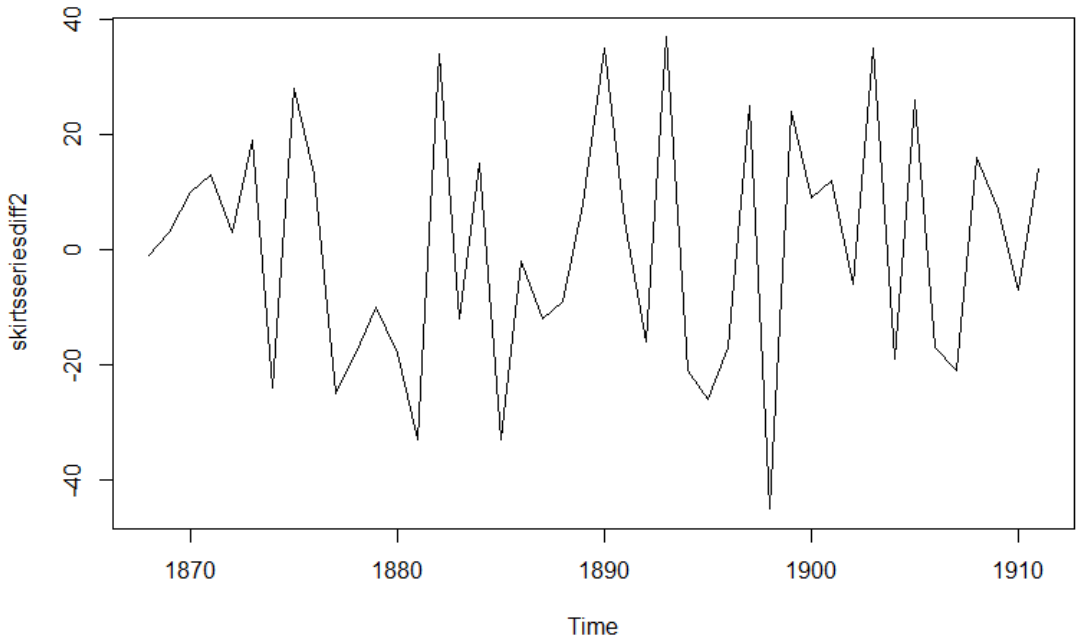
例子

- 1阶差分


skirtsseriesdiff1
60
40
20
0
-20
-40
-60
1870
1880
1890
1900
1910
Time

例子

- 2阶差分


skirtsseriesdiff2
40
20
0
-20
-40
1870
1880
1890
1900
1910
Time

平滑法

- 简单移动平均法

$$\hat{y}_t(1) = \hat{y}_{t+1} = (y_t + y_{t-1} + \ldots + y_1)/t$$

- 加权移动平均法

对不同时间点的数据不同的权重：

$$\alpha_i > 0; \sum \alpha_i = 1$$

$$\hat{y}_t(1) = \hat{y}_{t+1} = \sum_1^t \alpha_i y_i$$

- k期移动平均法

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + \ldots\ldots + y_{t-k+1}}{k}$$

平滑法

- 指数平滑法
  - 一次指数平滑法

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

  - 二次指数平滑法
- 平滑系数α的选择

例子

- 某百货公司一柜台2003年下半年各月的销售额分别为18、17、19、20、17、19万元，试用简单移动平均法预测2004年1月份该柜台的销售额。
- 设2003年7-12月的权数分别为0.5、1.0、2.5、3.5、5.0，使用加权移动平均法预测2004年1月份该柜台的销售额

例子

- 下表是我国1980-1981年平板玻璃月产量，试选用$N=3$和$N=5$用一次移动平均法进行预测。计算结果列入表中。

时间	序号	实际观测值	三个月移动平均值	五个月移动平均值
1980.1	1	203.8		
1980.2	2	214.1		
1980.3	3	229.9		
1980.4	4	223.7		
1980.5	5	220.7		
1980.6	6	198.4		
1980.7	7	207.8		
1980.8	8	228.5		
1980.9	9	206.5		
1980.10	10	226.8		
1980.11	11	247.8		
1980.12	12	259.5		

例子

时间	序号	实际观测值	三个月移动平均值	五个月移动平均值
1980.1	1	203.8	-	-
1980.2	2	214.1	-	-
1980.3	3	229.9	-	-
1980.4	4	223.7	215.9	-
1980.5	5	220.7	222.6	-
1980.6	6	198.4	224.8	218.4
1980.7	7	207.8	214.6	217.4
1980.8	8	228.5	209.0	216.1
1980.9	9	206.5	211.6	215.8
1980.10	10	226.8	214.3	212.4
1980.11	11	247.8	220.6	213.6
1980.12	12	259.5	227.0	223.5

例子

- 利用下表数据运用一次指数平滑法对1981年1月我国平板玻璃月产量进行预测（取α=0.3，0.5 ，0.7）。并计算均方误差选择使其最小的α进行预测。

时间	序号	实际观测值	指数平滑法		
			α=0.3	α=0.5	α=0.7
1980.01	1	203.8	—	—	—
1980.02	2	214.1	203.8	203.8	203.8
1980.03	3	229.9	206.9	209.0	211.0
1980.04	4	223.7	213.8	230.0	224.2
1980.05	5	220.7	216.8	226.9	223.9
1980.06	6	198.4	218.0	223.8	221.7
1980.07	7	207.8	212.1	211.1	205.4
1980.08	8	228.5	210.8	209.5	207.1
1980.09	9	206.5	216.1	219.0	222.1
1980.10	10	226.8	213.2	212.8	211.2
1980.11	11	247.8	217.3	219.8	222.1
1980.12	12	259.5	226.5	233.8	240.1
1981.01					

趋势拟合法

- 线性趋势预测模型

$$\mu_t = \alpha + \beta t$$

$$Y_t = \alpha + \beta t + \varepsilon_t$$

$$E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2$$

- 利用最小二乘法估计参数

$$\begin{cases} b = (n\sum ty - \sum t\sum y)\Big/[n\sum t^2 - \left(\sum t\right)^2] \\ a = \bar{y} - b\bar{t} \end{cases}$$

例子

- 某单位十年的商品销售额
- 试采用线性趋势预测2011年的销售额

$y_c = a + bt$，根据表中资料，可算出：

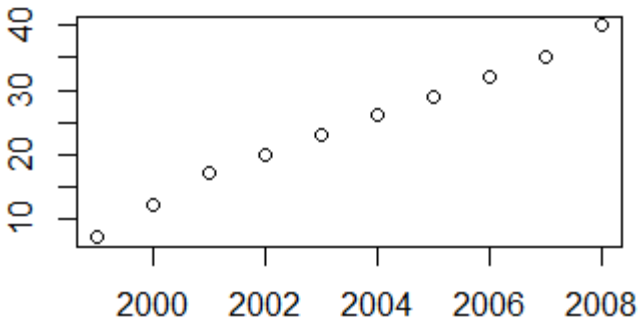
$$\begin{cases} b = \dfrac{n\sum ty - \sum t \sum y}{n\sum t^2 - \left(\sum t\right)^2} = \dfrac{10 \times 1607 - 55 \times 241}{10 \times 385 - 55^2} = 3.4 \\ a = \bar{y} - b\bar{t} = \dfrac{241}{10} - 3.4 \times \dfrac{55}{10} = 5.4 \end{cases}$$

$$y_c = 5.4 + 3.4t$$

将代表各年度的t值代入方程即可计算出各年的预测值。如预测2011年的商品销售额，相对应的

$$t = 13 \quad y_c = 5.4 + 3.4 \times 13 = 49.6$$

年份	商品销售额
1999	7
2000	12
2001	17
2002	20
2003	23
2004	26
2005	29
2006	32
2007	35
2008	40
合计	241



趋势拟合法

- 非线性趋势预测模型
  - 二次曲线

$$\mu_t = a + bt + ct^2$$

$$Y_t = a + bt + ct^2 + \varepsilon_t$$

$$E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2$$

  - 指数曲线

$$\mu_t = ab^t$$

$$Y_t = ab^t + \varepsilon_t$$

$$E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2$$

- 转化为线性模型求解系数

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间