课外资料

李 东

清华大学统计学研究中心

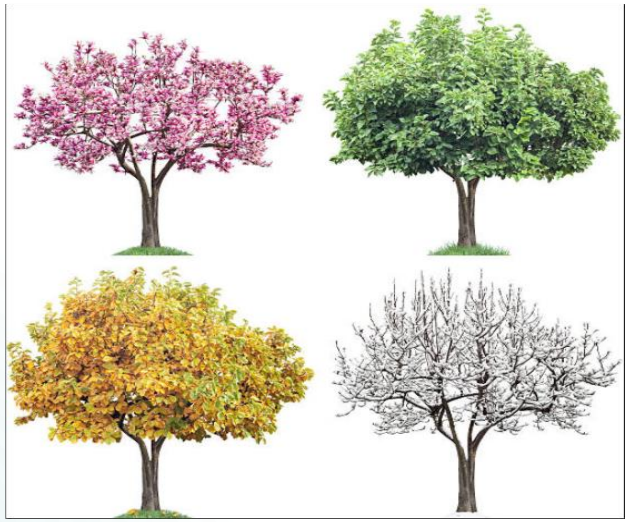
2019年3月

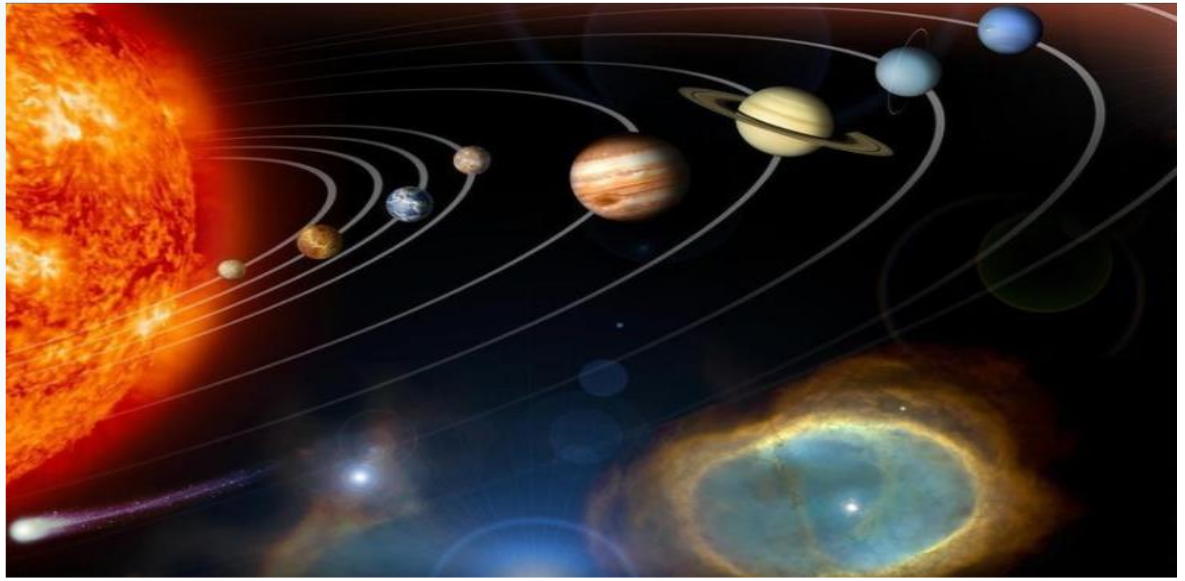
模型

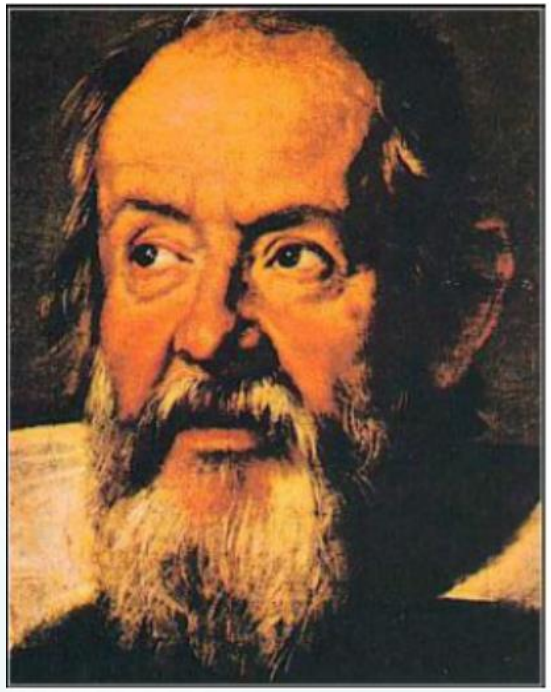
- 科学模型：科学研究中对事物的合理简化
  - 牛顿力学模型
  - 氢原子的玻尔模型
- 数学模型：对所描述的对象用数学语言所作出的描述和处理
- 统计模型：对所描述的对象用统计语言所作出的描述和处理
  - 线性回归模型
  - 时间序列模型
  - 空间统计模型
  - 时空统计模型
  - ……
- ……

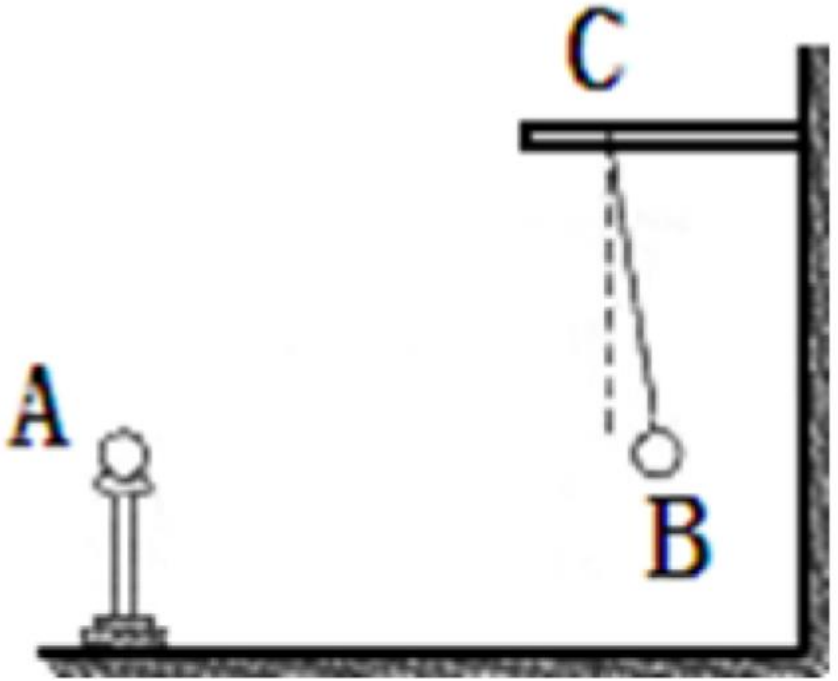
周期性：近代科学的起源


星期一
星期二
星期三
星期四
星期五
星期六
星期日
1M
2M
3M
4M
5M
6M
7M
8M
9M
10M
11M
12M
logo

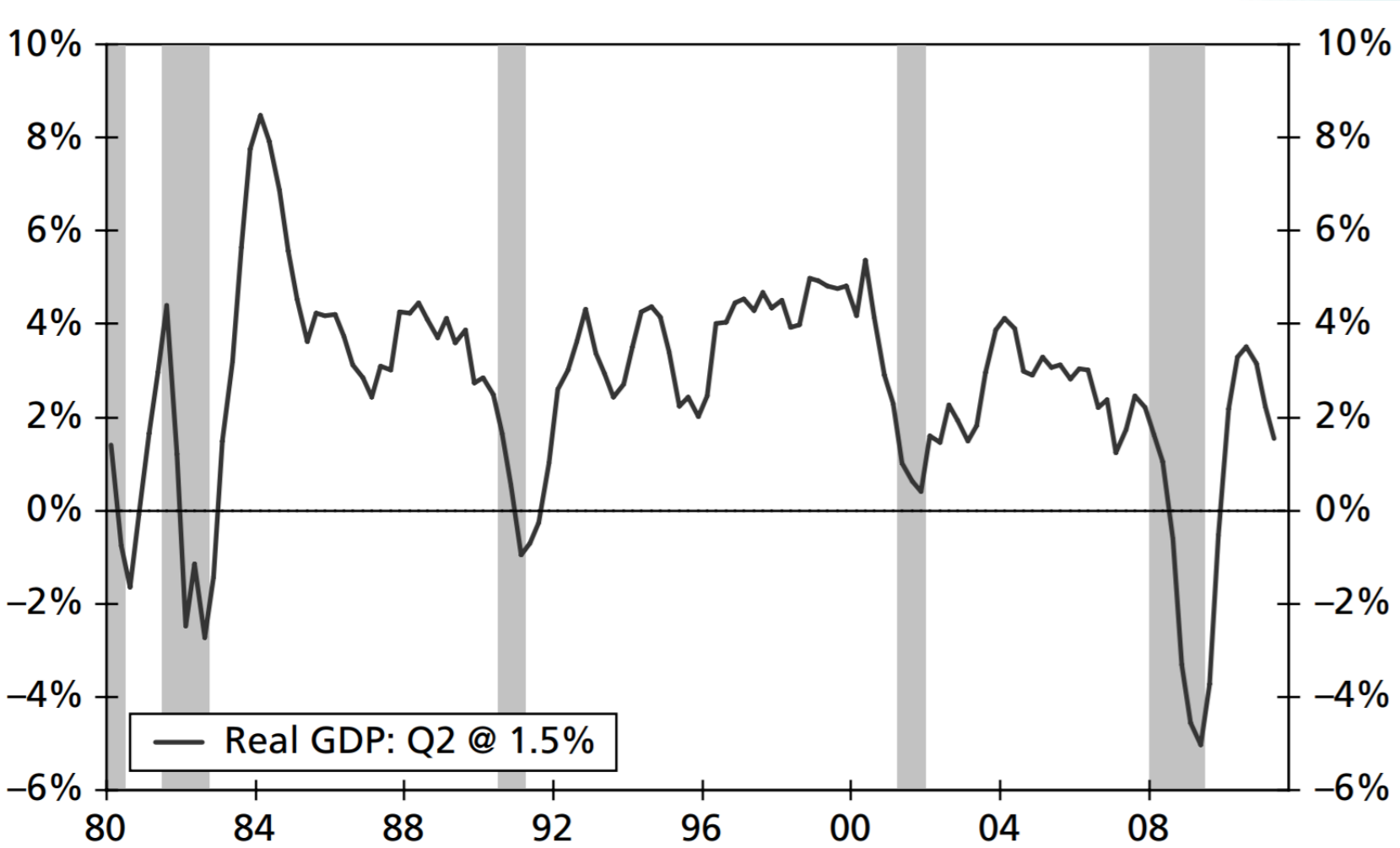








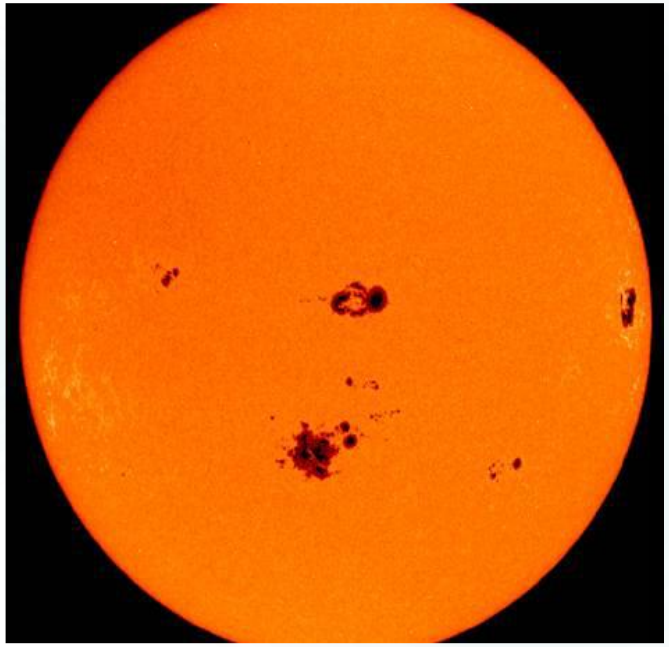
伽利略(1564/02/15-1642/01/08)是意大利的科学巨匠。在1583年，年仅19岁的他在教堂发现了吊灯的摆动现象.

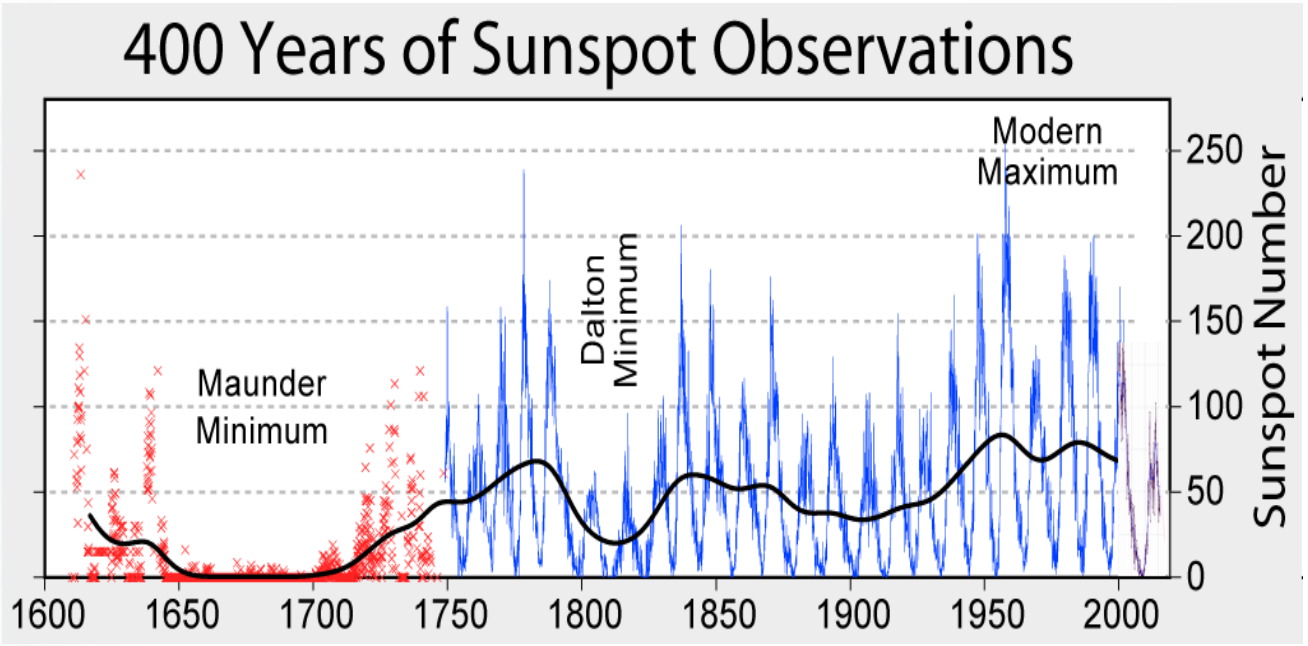


周期：
period
seasonal
cycle

数据实例

- 时间序列分析
  - 例1.【天文学：太阳黑子数据】

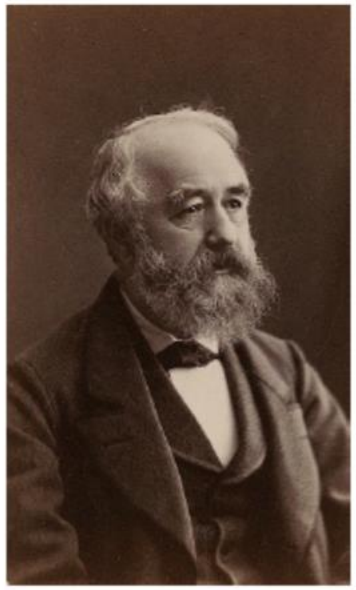
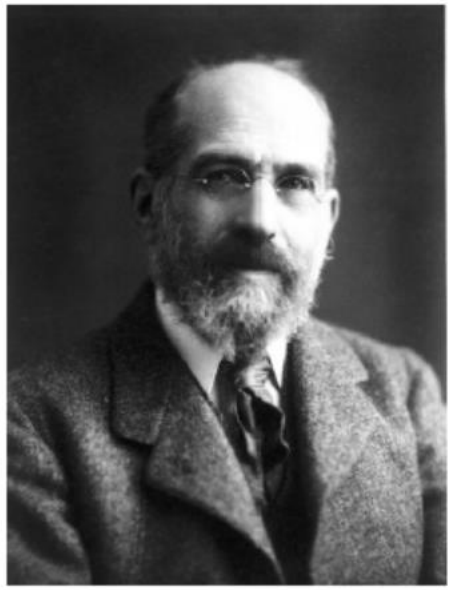




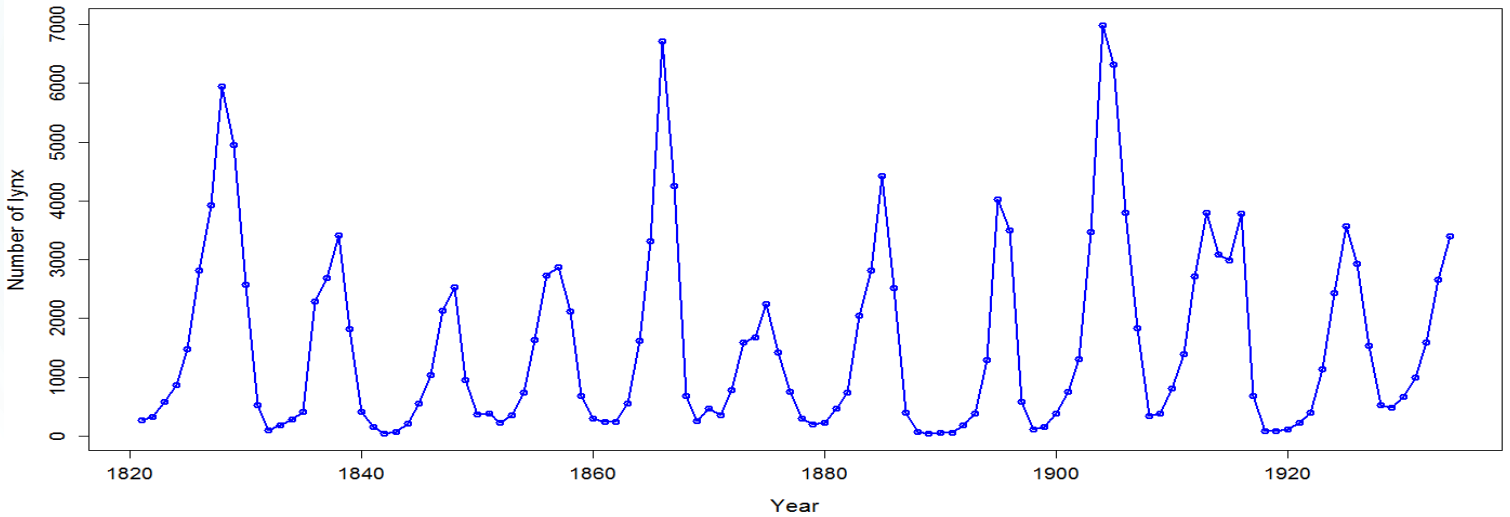
- 最早记录：《汉书.五行志》(公元前28年)：

  “三月乙未，日出黄，有黑气大如钱， 居日中央。”

- Samuel Heinrich Schwabe (太阳活动观测: 1826 -1843)
- Johann Rudolf Wolf (量化太阳黑子活动: 1848)
- Arthur Friedrich Schuster (周期图: 1898, 1906)
- George Udny Yule (AR(2) model, 1927)


- 例2.【生态学：加拿大山猫或猞猁数据】

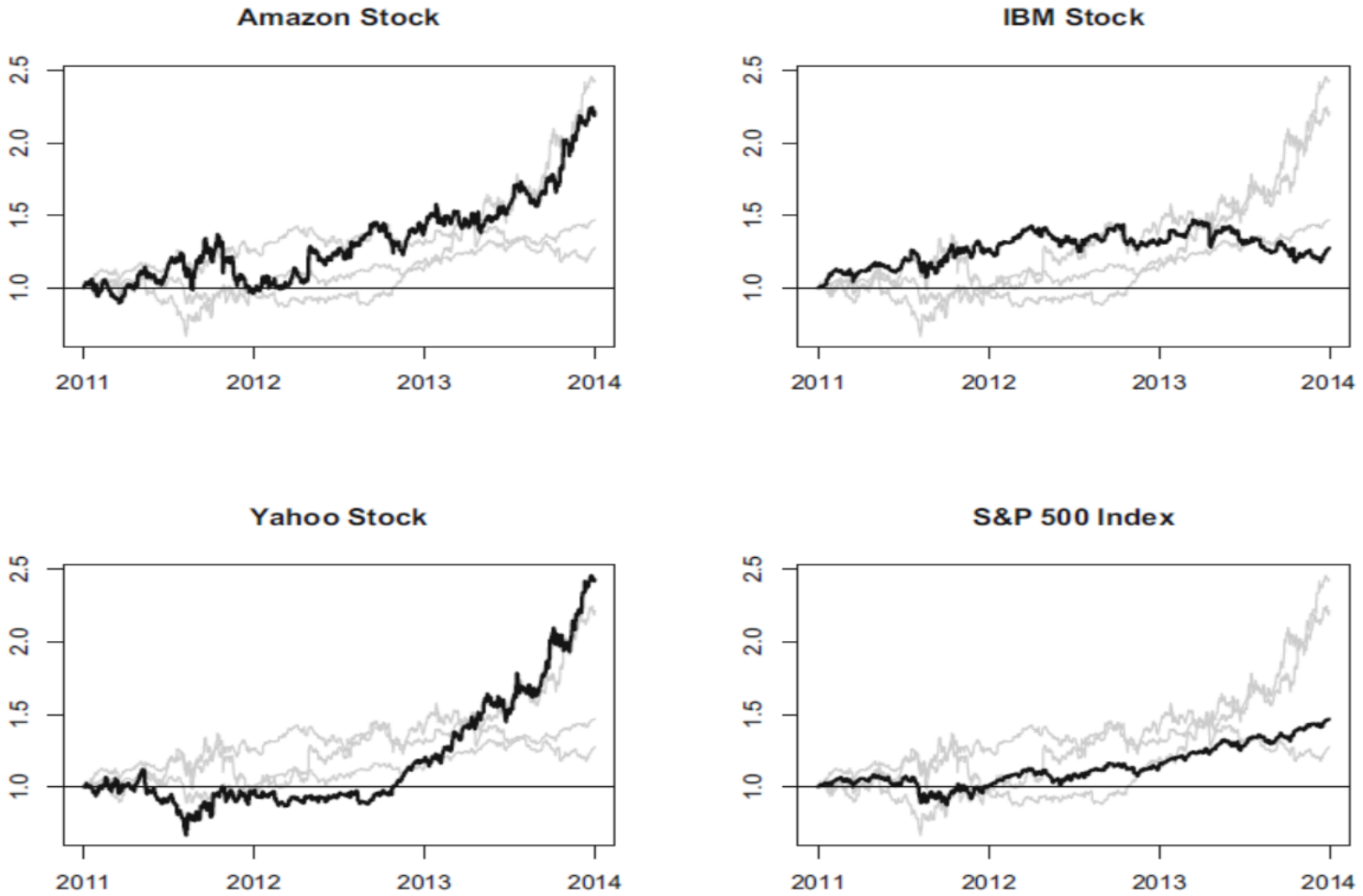


- 例3.【人口学：中国人口出生率】

- 例4.【金融：比特币+股票】


Bitcoin Price
Bitcoin.com
$18000.00
$16000.00
$14000.00
$12000.00
$9751.88
$8000.00
$6000.00
$4000.00
$2000.00
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2018
Feb
Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul Jan Jul
2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018

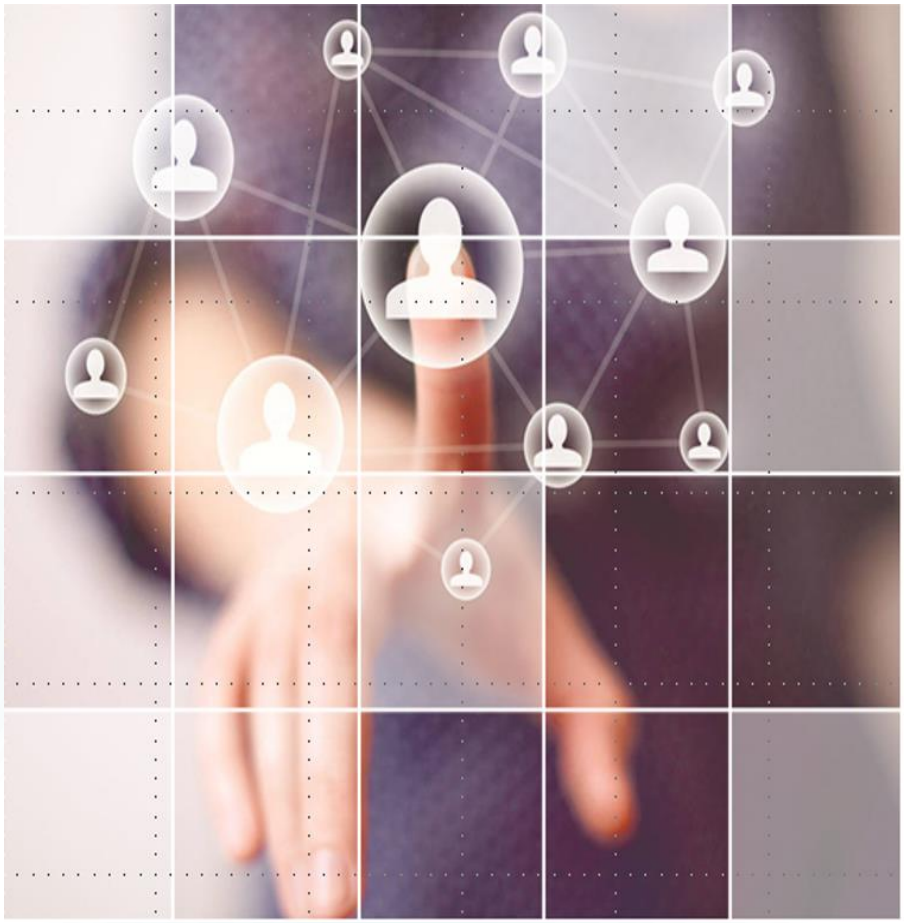
Value of $1 Invested in Amazon, IBM, Yahoo, and the Market
December 31, 2010 - December 31, 2013

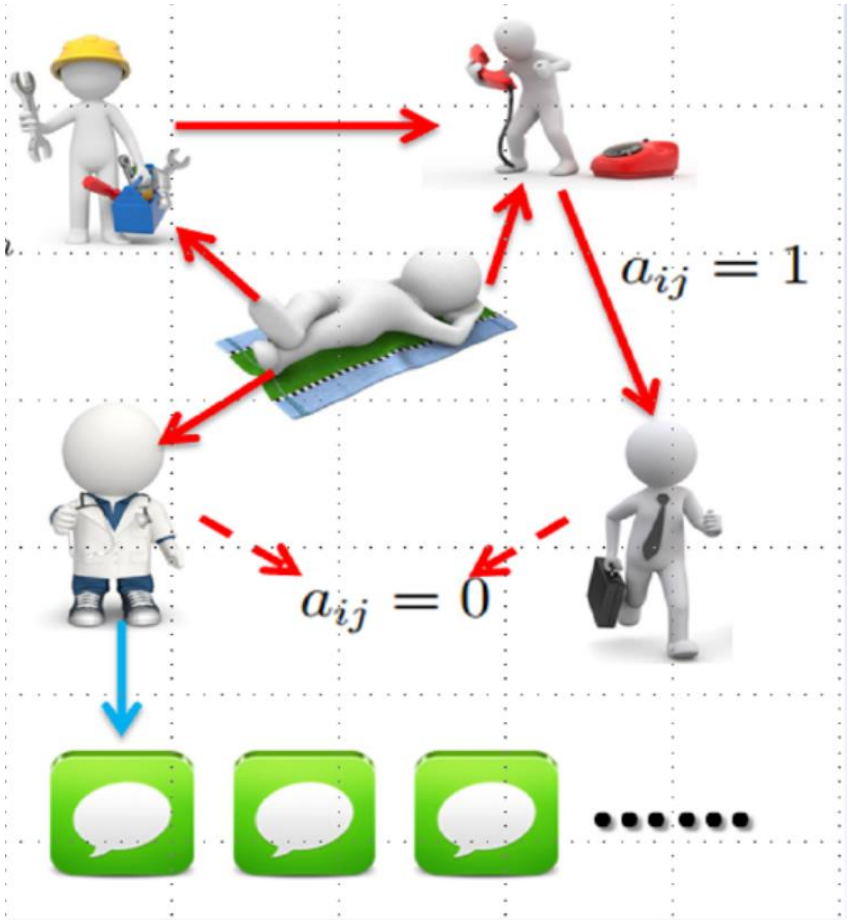


- 例5.【经济】


11.2 万亿美元（2016 年）
20兆
15兆
10兆
5兆
0
1970年
1980年
1990年
2000年
2010年
美国
18.57 万亿 USD
中华人民共和国
11.2 万亿 USD
印度
2.264 万亿 USD

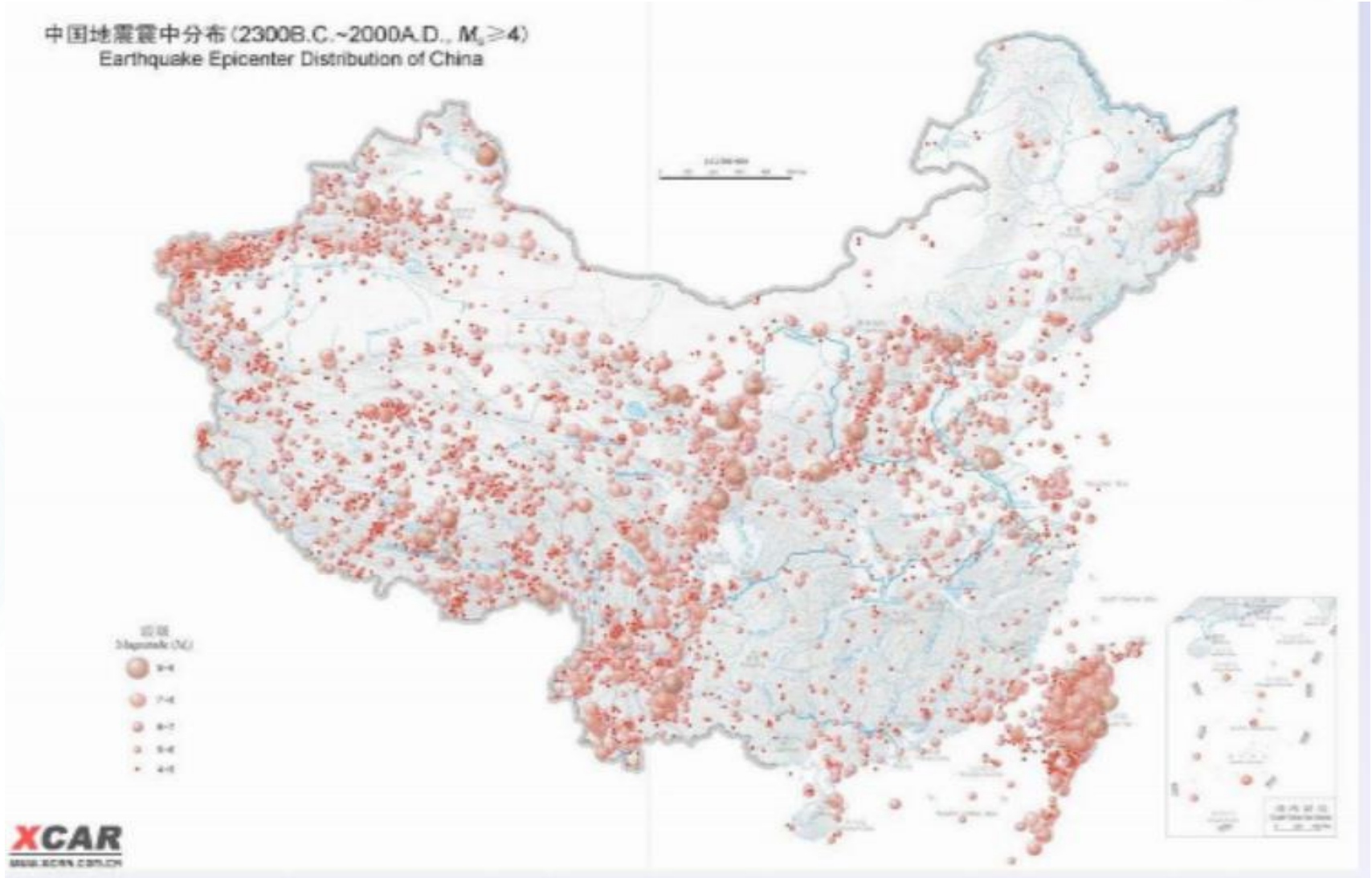
- 例6.【网络】





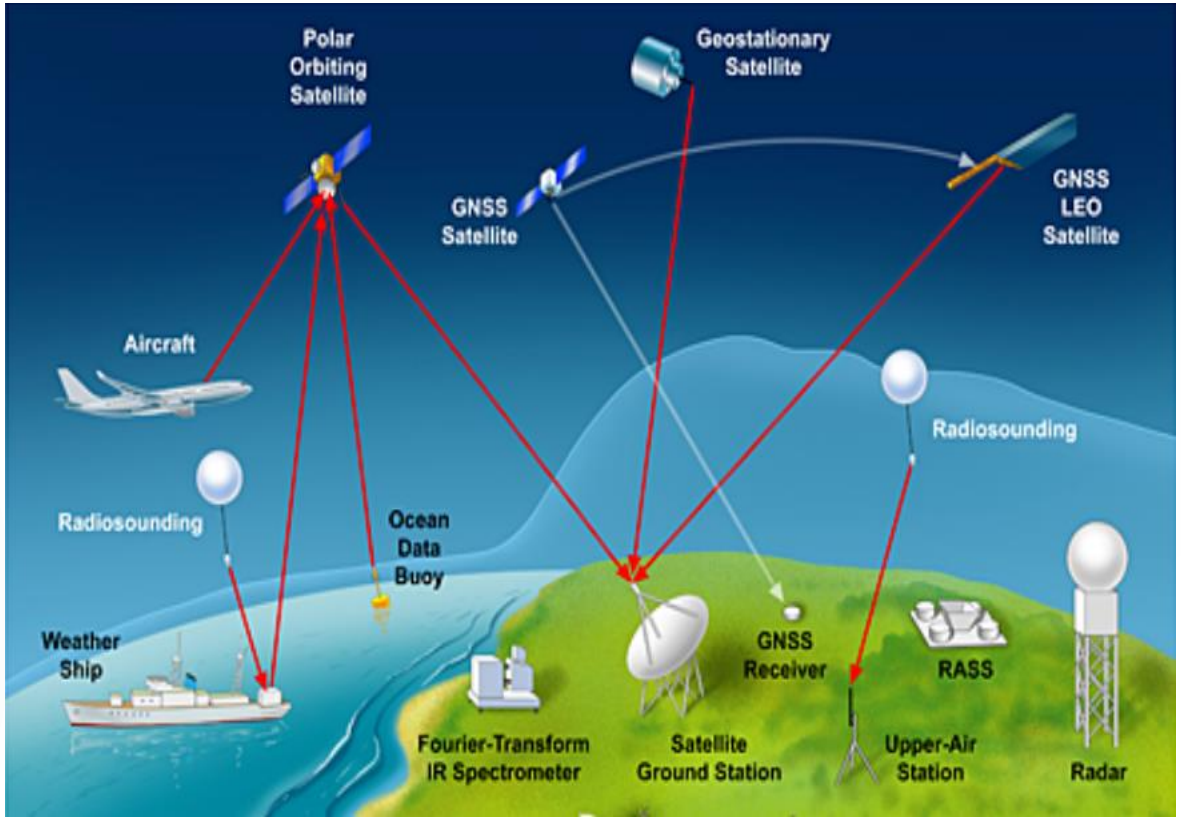
- 例7.【环境】

- 例8.【地震学】


中国地震震中分布(2300B.C.~2000A.D., $M_s \geq 4$)
Earthquake Epicenter Distribution of China
XCAR

- 例9.【空间经济学】

- 例10.【控制论+遥感技术】


Polar
Orbiting
Satellite
Geostationary
Satellite
GNSS
Satellite
GNSS
LEO
Satellite
Aircraft
Radiosounding
Radiosounding
Ocean
Data
Buoy
Weather
Ship
GNSS
Receiver
RASS
Fourier-Transform
IR Spectrometer
Satellite
Ground Station
Upper-Air
Station
Radar

时间序列分析的历史简介

- 单变量时间序列模型的历史
  - 线性时间序列分析
    - 开端
      - 线性自回归模型（autoregressive model, AR)

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t .$$

        - AR(2) 模型: George Udny Yule 在1927年提出
        - AR(p) 模型: Gilbert Thomas Walker 在1931年提出
      - 线性滑动平均模型（moving average model, MA)

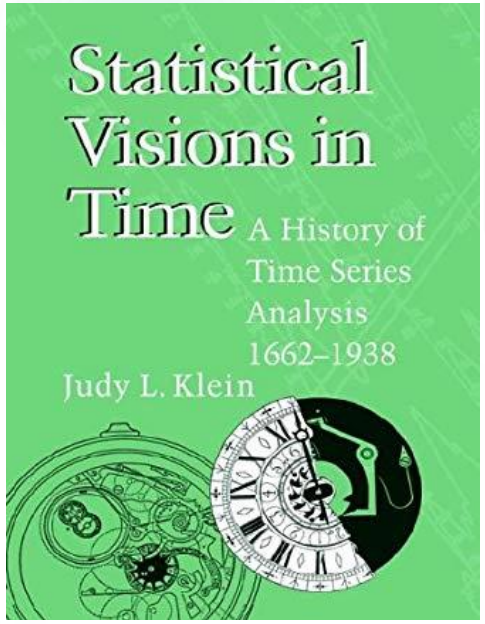
$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} .$$

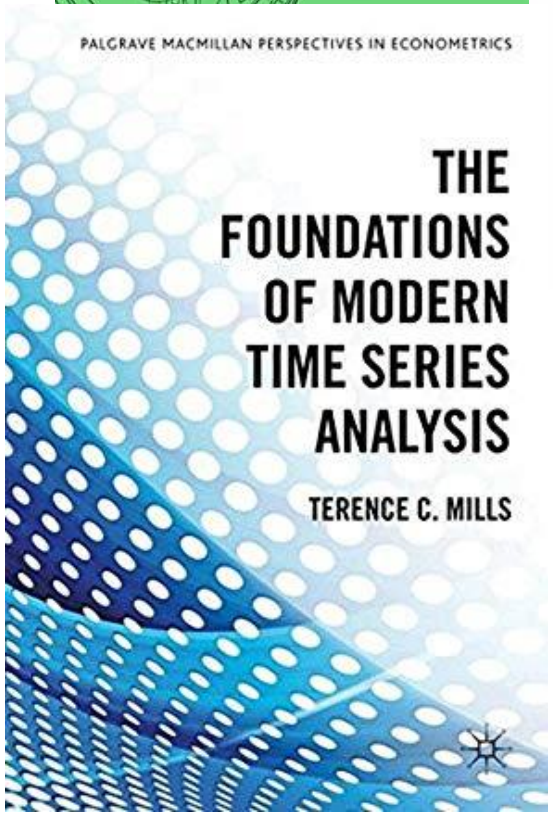
        - MA(q) 模型: Eugen Slutsky 在1927年提出

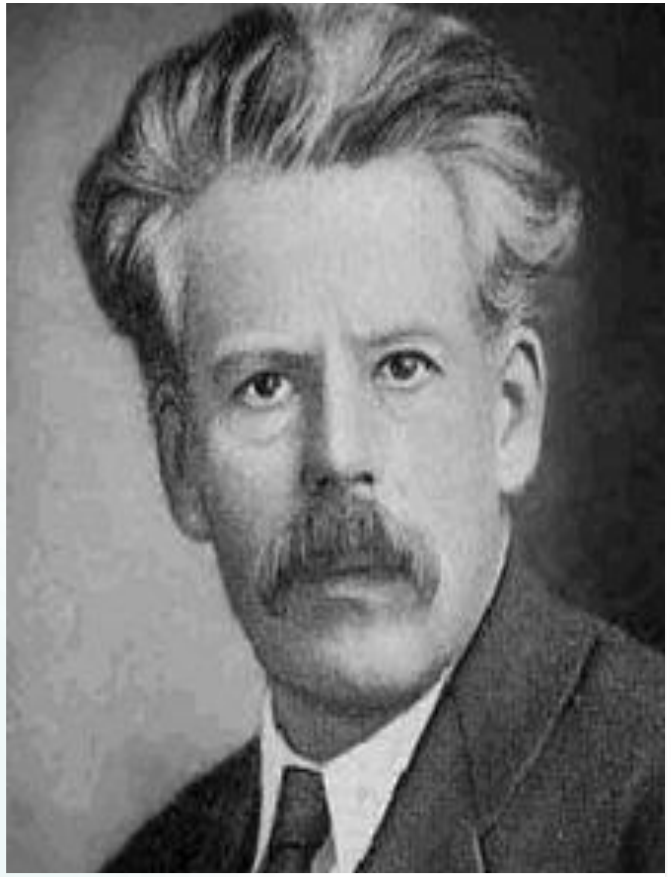
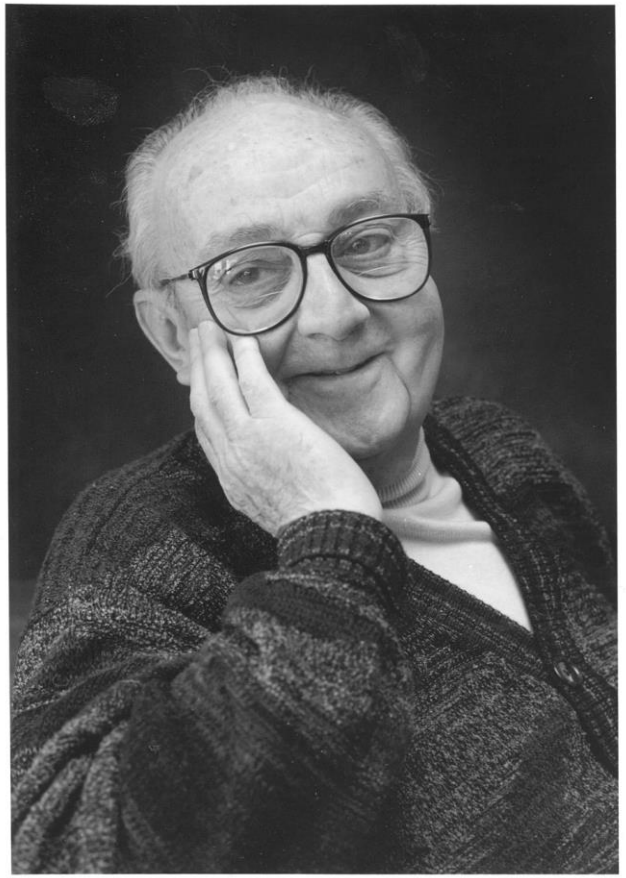
✓ Herman Ole Andreas Wold (25 Dec. 1908 - 16 Feb. 1992)

✓ Andrew Morris Walker (21 Dec. 1921 - 23 Dec. 2004) (ARMA model in 1950)

✓ Peter Whittle (27 Feb. 1927-) (ARMA model in 1951)





左: George Udny Yule (18 Feb. 1871-26 Jun. 1951), 英国统计学家；中: Eugen Slutsky (19 April 1880-10 Mar. 1948), 俄国/苏联数理统计学家、经济学家和政策经济学家；右: George Edward Pelham Box (18 Oct. 1919- 28 Mar. 2013), 英国统计学家(质量控制, 时间序列分析, 实验设计，贝叶斯推断). 被称之为 “one of the great statistical minds of the 20th century”.

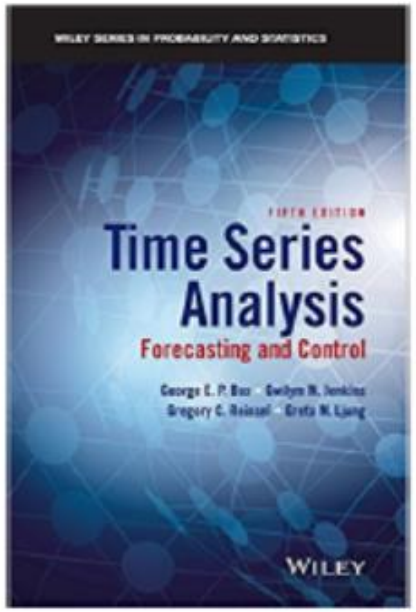
名言：“**Essentially, all models are wrong, but some are useful**.” —— George E. P. Box.

- 成熟期（1970年代中期）
  - ARIMA(p,d,q) 模型:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t\ .$$

  - Box-Jenkins multiplicative seasonal ARIMA$(p,\ d,\ q) \times (P,D,Q)_s$ 模型

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D(y_t-\mu) = \Theta_Q(B^s)\theta_q(B)\varepsilon_t.$$



- 非线性时间序列分析
  - 萌芽期（1953、1954）

Left: Patrick Alfred Pierce Moran (14 July 1917 - 19 Sept. 1988), commonly known as Pat Moran was an Australian statistician who made significant contributions to probability theory and its application to population and evolutionary genetics. Right: Peter Whittle (27 Feb. 1927 -, in Wellington, New Zealand) is a mathematician and statistician, working in the fields of stochastic nets, optimal control, time series analysis, stochastic optimization and stochastic dynamics.

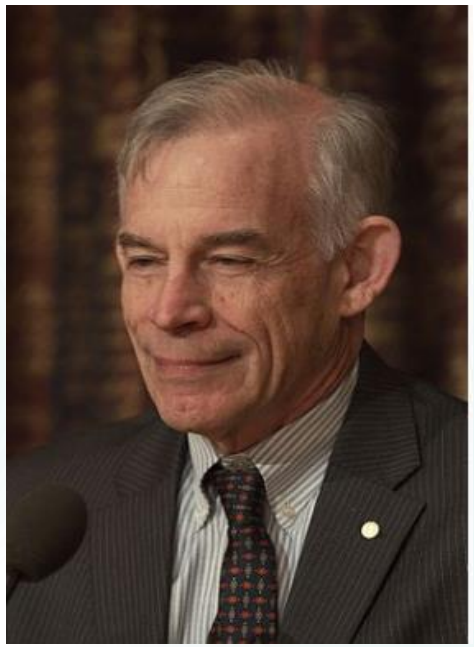
- 蓬勃期（Spring, 1980s）
  - 门限自回归模型（Threshold AR model）
  - 自回归条件异方差模型（ARCH）

Left: Howell Tong (1944-, in Hong Kong) is a pioneer and an acknowledged authority in the eld of nonlinear time series analysis, linking it with deterministic chaos. He is the father of the threshold time series models, which have extensive applications in ecology, economics, epidemiology and finance. Right: Robert Fry Engle (10 Nov. 1942 -) is an American economist and the winner of the 2003 Nobel Memorial Prize in Economic Sciences, sharing the award with Clive Granger, “for methods of analyzing economic time series with time-varying volatility (ARCH)”.

多变量时间序列模型的历史

- Vector AR model by Christopher A. Sims (1980).

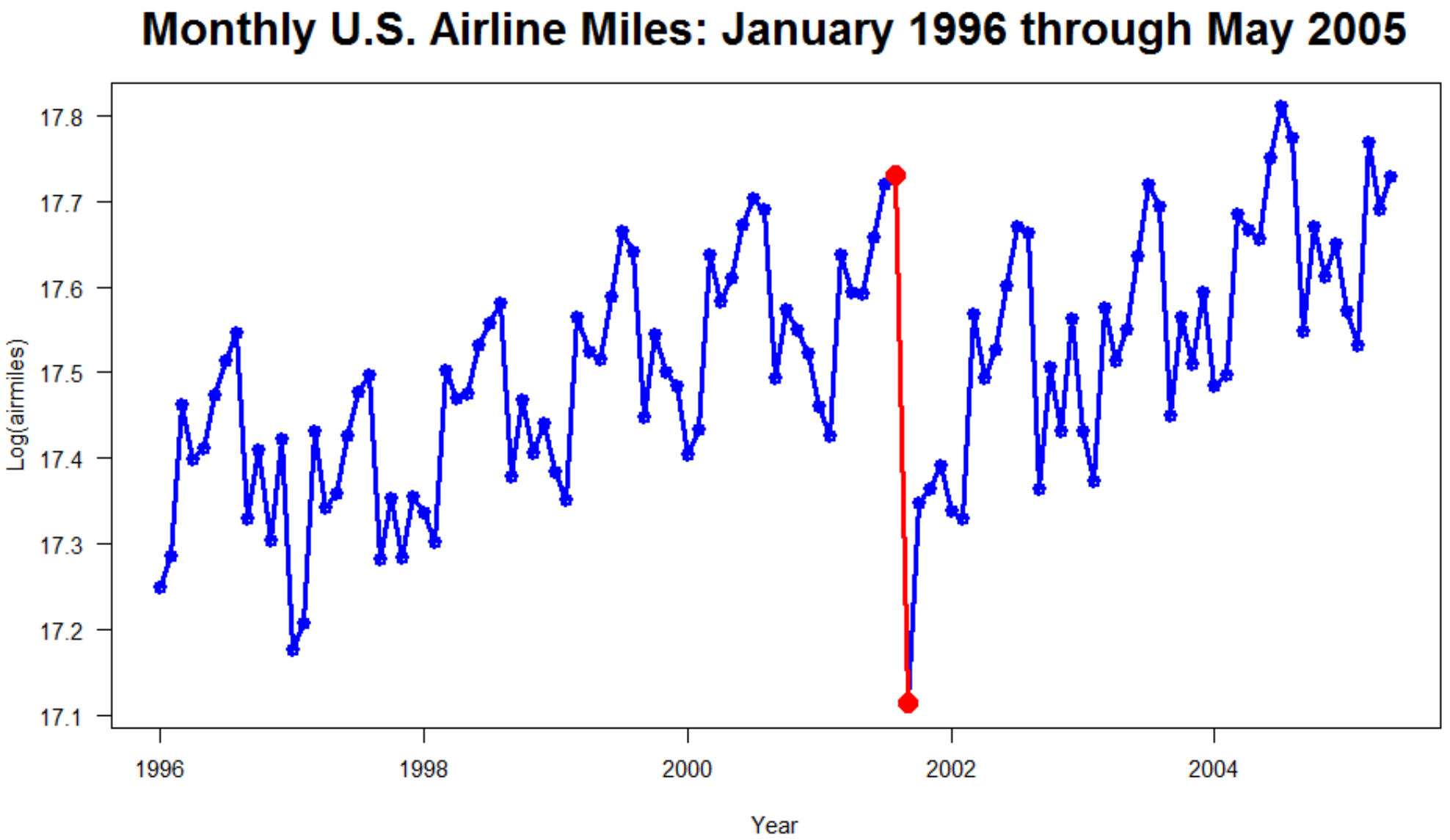
$$\mathbf{Y}_t = \Phi_1 \mathbf{Y}_{t-1} + \cdots + \Phi_p \mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t.$$



生于1942年10月21日，美国计量经济学家、宏观经济学家。因在“宏观经济中关于因果效应的经验研究”获得2011年的诺贝尔经济学奖。
Fellow of the Econometric Society (since 1974);
Member of the American Academy of Arts and Sciences (since 1988);
Member of the National Academy of Sciences (since 1989).

案例

- 案例1. “9/11事件”对美国航空业的影响


Monthly U.S. Airline Miles: January 1996 through May 2005
Log(airmiles)
17.8
17.7
17.6
17.5
17.4
17.3
17.2
17.1
1996
1998
2000
2002
2004
Year

- 问题：你如何评价或度量“9/11事件”对美国航空业的影响?
  - 想法：把想要分析的时间序列 $\{Y_t\}$分解为：$Y_t = N_t + m_t$， 其中m_t是均值函数的变化，N_t是某个ARIMA模型（可能包含季节性成分），通常表示没有被干预过的基本时间序列。干预性时间通常是已知的，记为：T. 在T之前，m_t通常假定为0. $\{Y_t : t < T\}$称之为“*干预前数据*”，常常用来识别过程N_t.
  - 如何刻画m_t? 主观的想法：用含有有限个参数的函数来表达.
    - 如何表达?

1. 阶梯函数（step function）：

$$S_t^{(T)} = \begin{cases} 1, & \text{如果 } t \geq T, \\ 0, & \text{如果 } t < T. \end{cases}$$

2. 脉冲函数 （pulse function）：$P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)} = \begin{cases} 1, & \text{如果 } t = T, \\ 0, & \text{如果 } t \neq T. \end{cases}$

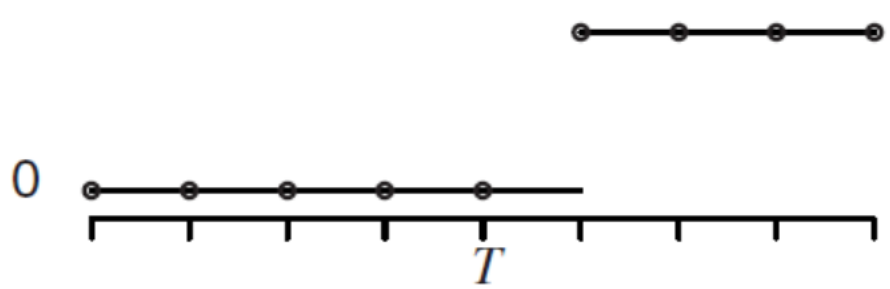
关系： $P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)} = (1 - B)S_t^{(T)}$.

- 通常，用ARMA类型的模型去识别m_t:

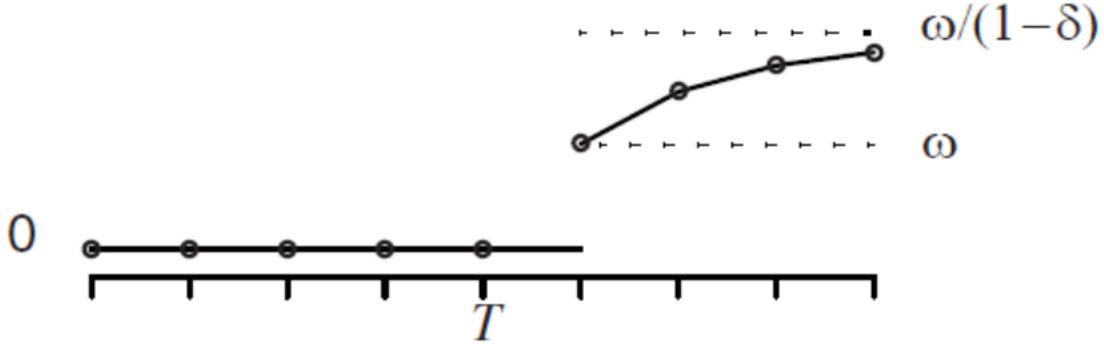
$m_t = \frac{\omega(B)}{\delta(B)} P_t^{(T)}$，其中$\omega(B)$ 和 $\delta(B)$是 B 的某个多项式.

阶梯型响应干预的常用类型：

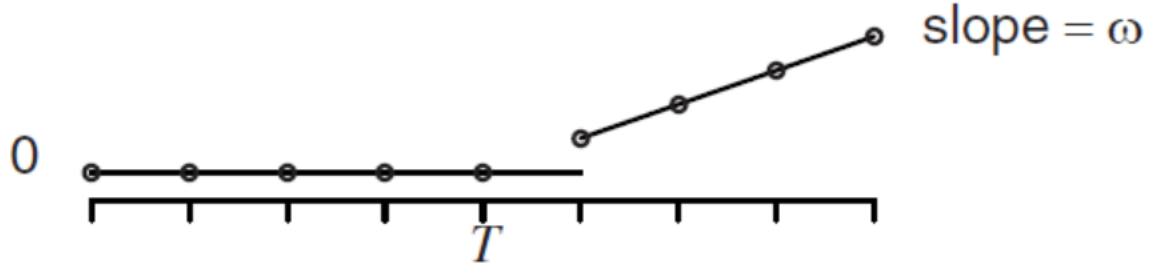
(a) $\omega B S_t^{(T)}$



(b) $\frac{\omega B}{1-\delta B} S_t^{(T)}$



(c) $\frac{\omega B}{1-B} S_t^{(T)}$

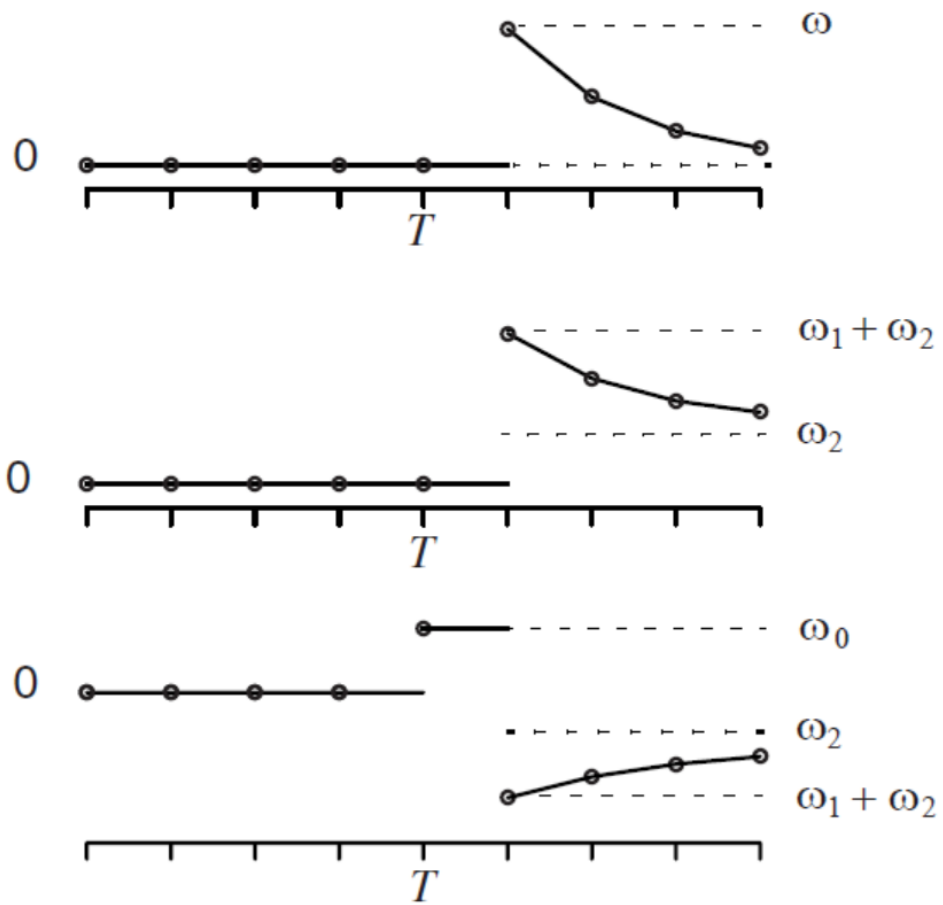


脉冲型响应干预的常用类型：

(a) $\dfrac{\omega B}{1-\delta B} P_t^{(T)}$

(b) $\left[\dfrac{\omega_1 B}{1-\delta B} + \dfrac{\omega_2 B}{1-B}\right] P_t^{(T)}$

(c) $\left[\omega_0 + \dfrac{\omega_1 B}{1-\delta B} + \dfrac{\omega_2 B}{1-B}\right] P_t^{(T)}$

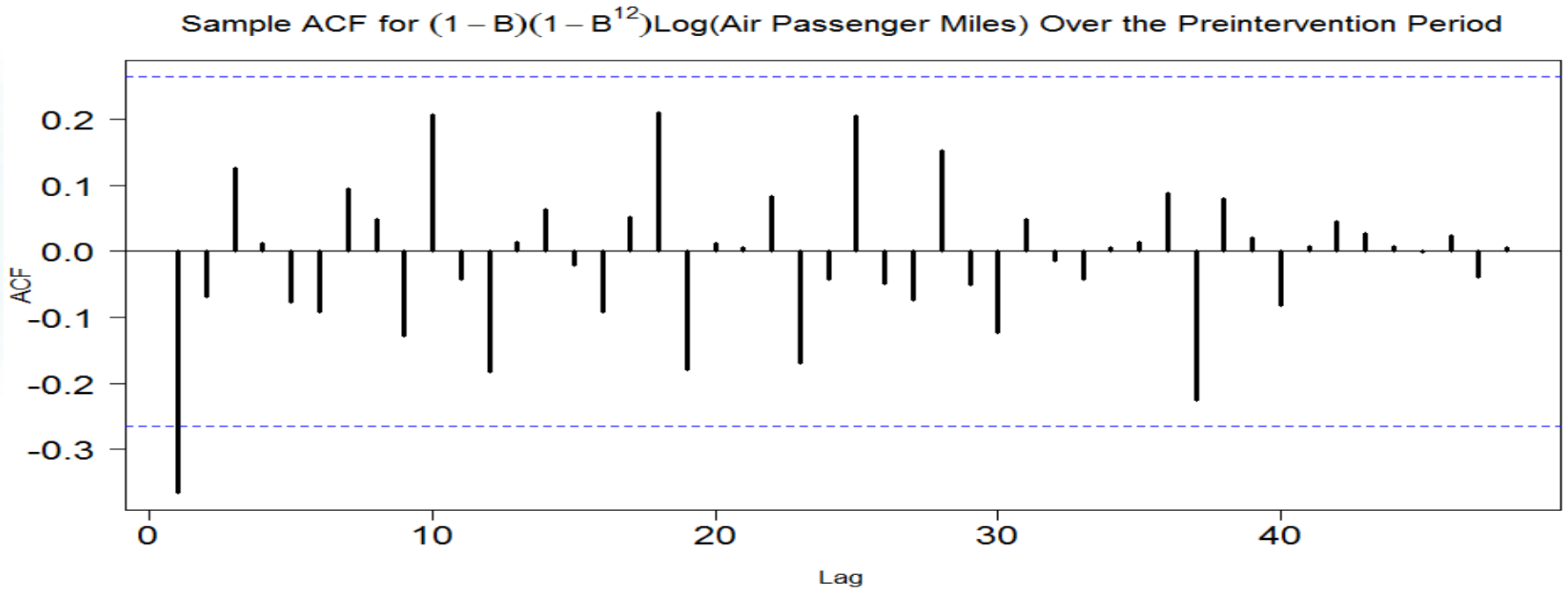


- 9/11效应定义为:

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B} P_t^{(T)},$$

其中T是2001年9月，$\omega_0 + \omega_1$ 表示9/11瞬时效应, $\omega_1 \omega_2^k$ 表示向前第k个月的9/11效应.

- 识别N_t: (利用干预前数据)　　$\text{ARIMA}(0,1,1) \times (0,1,0)_{12}$.

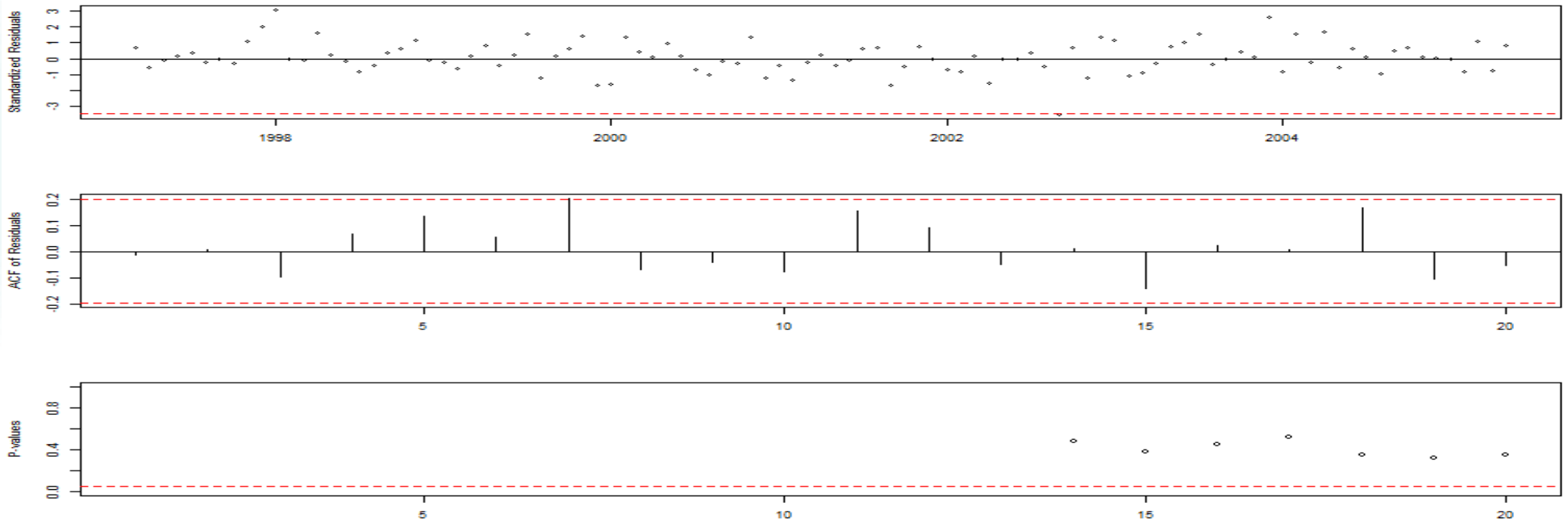


- 最终模型：ARIMA$(0,1,1)\times(0,1,1)_{12}$+9/11效应+3个可加离群值，其结果如下：

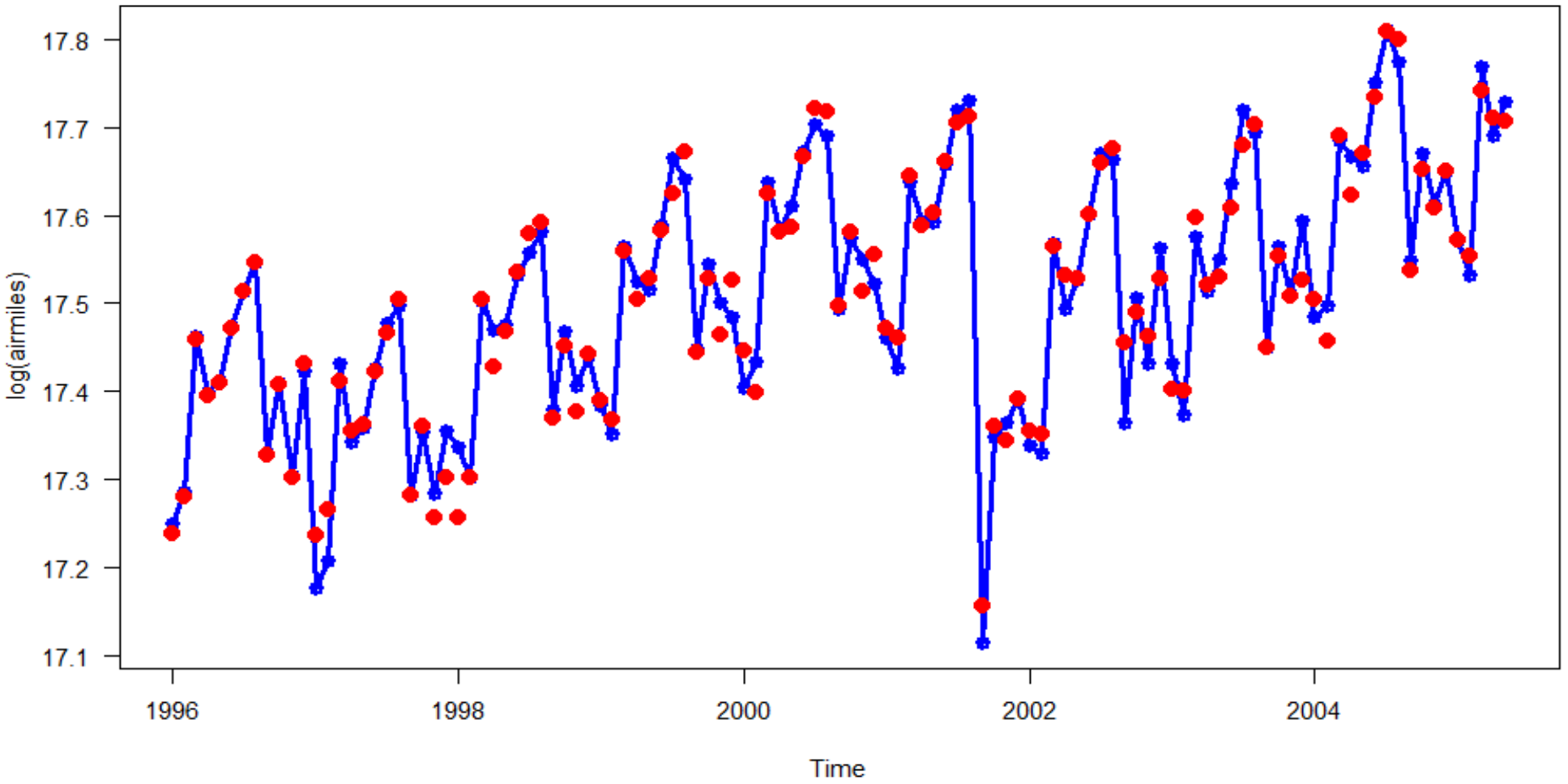
```
Coefficients:
          ma1      sma1    Dec96     Jan97    Dec02  I911-MA0  I911.1-AR1  I911.1-MA0
      -0.3825  -0.6499   0.0989  -0.0690   0.0810   -0.0949      0.8139     -0.2715
s.e.   0.0926   0.1189   0.0228   0.0218   0.0202    0.0462      0.0978      0.0439

sigma^2 estimated as 0.0006721:  log likelihood = 219.99,  aic = -423.98
```

- 模型诊断：



- 拟合数据与真实数据之比较



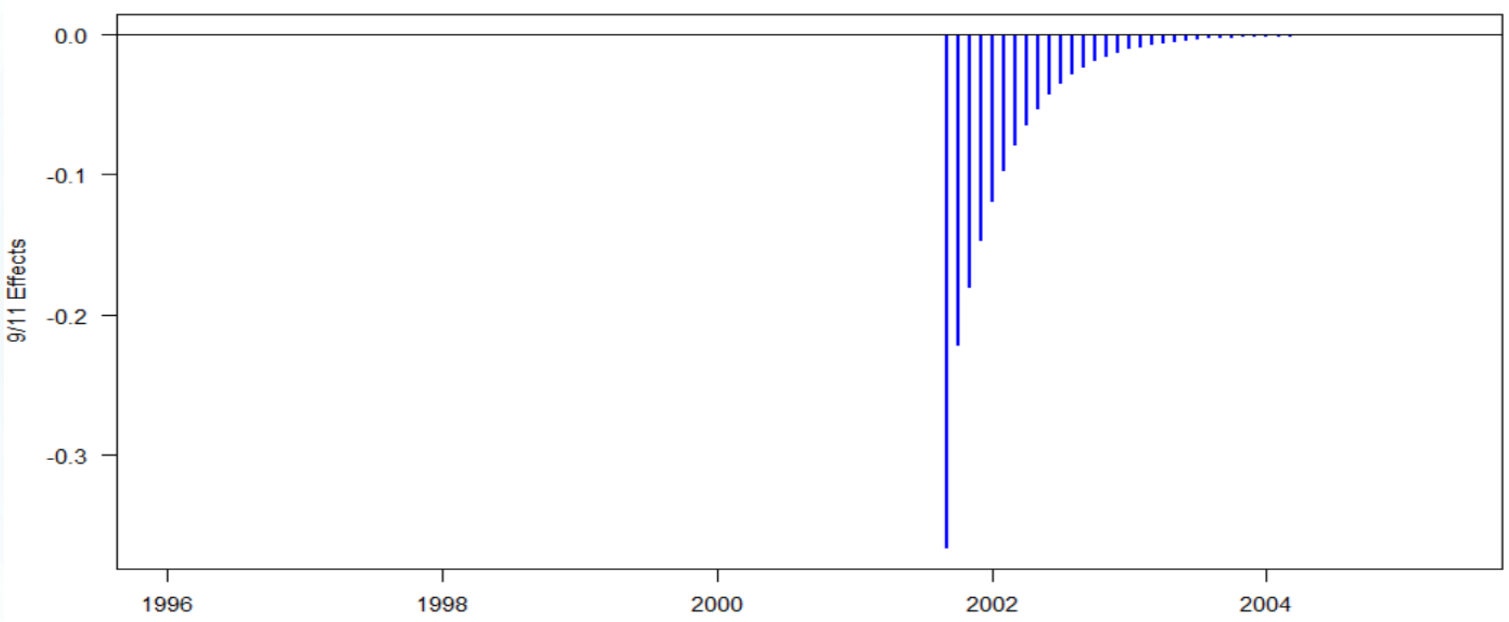
- 9/11效应的度量

$$\log y_t = m_t + \log N_t \longrightarrow \log\left(\frac{y_{T+k}}{N_{T+k}}\right) = m_{T+k} = (\omega_0 + \omega_1)I(k=0) + \omega_1\omega_2^k I(k>0)\ .$$

  - 瞬时9/11效应:

$$\frac{y_t - N_t}{N_t} \times 100\% = \{\exp(\omega_0 + \omega_1) - 1\} \times 100\% = -31\%. \quad \text{【增加量】}$$

  - 持续效应: $\{1 - \exp(\omega_1\omega_2^k)\} \times 100\%.$

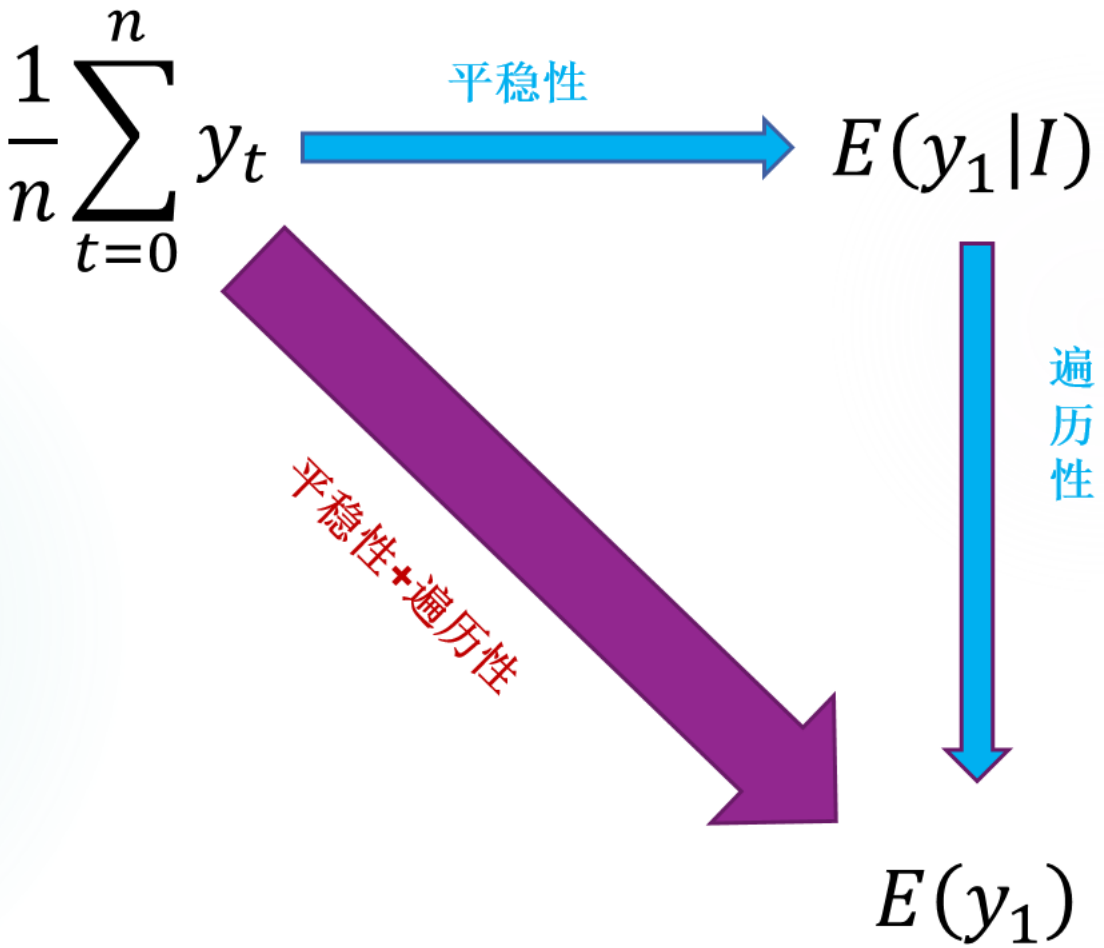


基础知识 I

- 数据的展示
- 数据的特征
  - 趋势项 【Hooke，消除趋势项技巧：平滑、差分等】
  - 季节项 【引入季节性因子】
  - 周期项
  - 平稳项 【ARMA模型】
- ARMA模型
  - AR模型
  - MA模型
  - ARMA模型
- 平稳性
  - 宽平稳、弱平稳、二阶平稳、协方差平稳
    (weak/second-order, covariance stationarity)
  - 严平稳 (strict stationarity)
  - （1） 严平稳+二阶矩有限 蕴含 宽平稳；（2）平稳高斯时间序列，宽平稳与严平稳等价！

- 为什么在时间序列分析中需要**平稳性**这个概念?

  设$\{y_1, y_2, \cdots, y_n\}$是来自某个时间序列的一组观测值.



- 可逆性（invertibility）
  - 保证信息$\sigma(y_j: j \le t) = \sigma(\varepsilon_j: j \le t)$, 对任意的 t 都成立.
  - 与reversibility区别开来.
- 因果性 （causality）
  - 问题：考虑AR(1)过程的平稳性：$y_t = \phi y_{t-1} + \varepsilon_t$, 其中$\{\varepsilon_t\}$是$i.i.d.$
- AR(∞) 与 MA(∞)表示
  - AR(∞)表示： $\varepsilon_t = y_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots$.
  - MA(∞)表示: $y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots$.
- 平稳时间序列的自协方差函数与自相关系数（ACF）

$$\gamma_k = \mathrm{Cov}(y_t, y_{t-k}), \qquad \rho_k = \frac{\gamma_k}{\gamma_0} = \mathrm{Corr}(y_t, y_{t-k}).$$

仅仅依赖于时间差，与起始点没关系！ $\{\gamma_k\}, \{\rho_k\}$ 作为k 的函数， 是非负定函数. 进而跟普分解、傅里叶分析联系在一起，产生时间序列的谱分析方法.

- 平稳时间序列的偏自相关系数 (Partial ACF, PACF): $\{\phi_{kk} : k \ge 1\}$, $\phi_{11} = \rho_1$.
- 样本ACF，PACF，EACF.

- 估计方法
  - Yule-Walker估计
    - ✓ 容易实现，但不是很有效；
    - ✓ 只适应于样本量很大的AR模型，对MA模型和ARMA模型则是非常复杂；
    - ✓ 无论对什么模型而言，此方法对于四舍五入所造成的误差敏感；
    - ✓ 可以为更有效的估计方法提供初始值；
    - ✓ 对接近非平稳或者接近非可逆的过程，不推荐使用此方法作为最终方法；
    - ✓ 虽然此方法不尽人意，对目前的高维AR模型来说，也没有其它更好的方法。
  - (条件)最小二乘估计
  - (条件)最小一乘估计
  - (条件)高斯伪极大似然估计
  - (条件)非高斯伪极大似然估计
  - ……

模型选择准则

- 想法：平衡参数与模型复杂度!
  - AIC（Akaike Information Criterion, 赤池信息准则）
  - BIC（Bayesian Information Criterion, 贝叶斯信息准则）
  - HQC（Hannan-Quinn Information Criterion）
  - ……
- 实际中，AIC和BIC相比较，到底选择哪个准则呢？答案：AIC.
  - ✓ 我们建模的目标是获得一个“好”的模型，而不是寻找一个“真实”的模型，记住：统计模型仅仅是对复杂系统的一种近似，估计真实的阶显然不是既定目标。真实模型或阶仅仅存在于随机模拟实验中。从模型是复杂现象的一种近似的观点来说，真实的阶可能无穷大.
  - ✓ 即使是真实的有限的阶存在，但是一个好的模型的阶未必就等于真实的阶.在观测值比较少的情况下，考虑到被估参数的不稳定性，AIC揭示了低阶模型可以获得较高预报精度的可能性.
  - ✓ 尽管信息准则使得自动模型选择称为可能，但是模型评价毕竟是相对的. 这意味着利用信息准则选择模型仅仅是从一个指定的模型类里面挑选一个适当的模型. 所以说，基于我们对目标的认识，我们的主要任务是提出更多适当合理的模型.
  - ✓ ……

- 模型诊断

  检验残差是不是白噪声或是否含有异方差

  - Ljung-Box 检验
  - McLeod-Li 检验

- 干预性分析

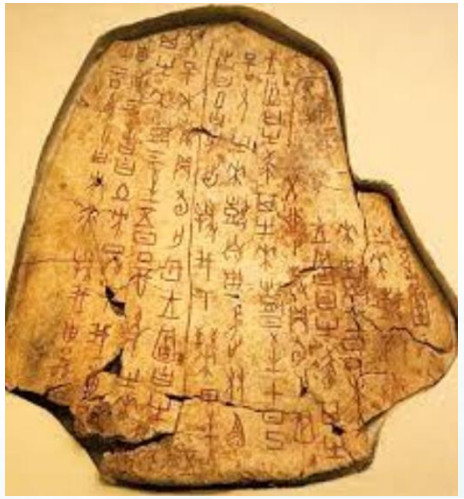
- 离群值探测
  - 可加离群值（Additive outlier, AO）
  - 新息离群值（Innovative outlier, IO）

- 转移函数方法 （transfer）

$$y_t = \nu(B)x_t + \varepsilon_t.$$

- 预报

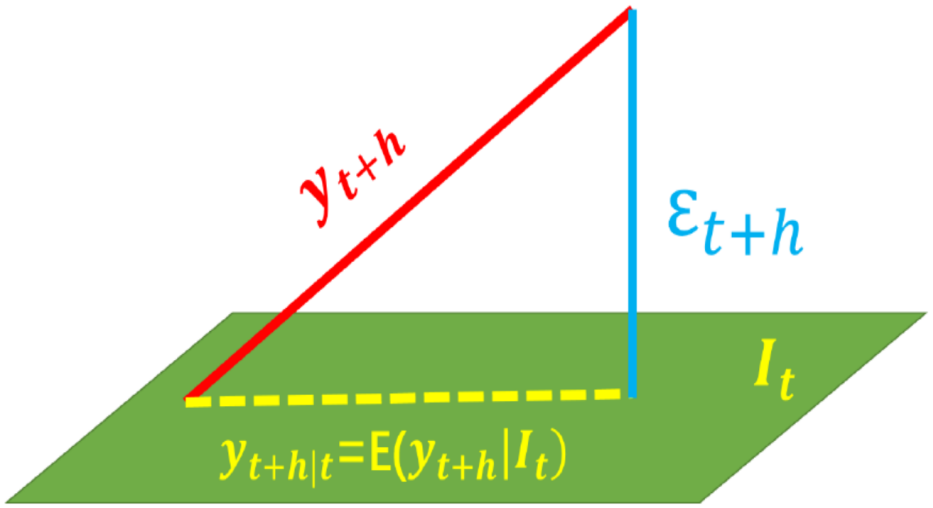
想法：利用现在及过去的信息，预测未来.

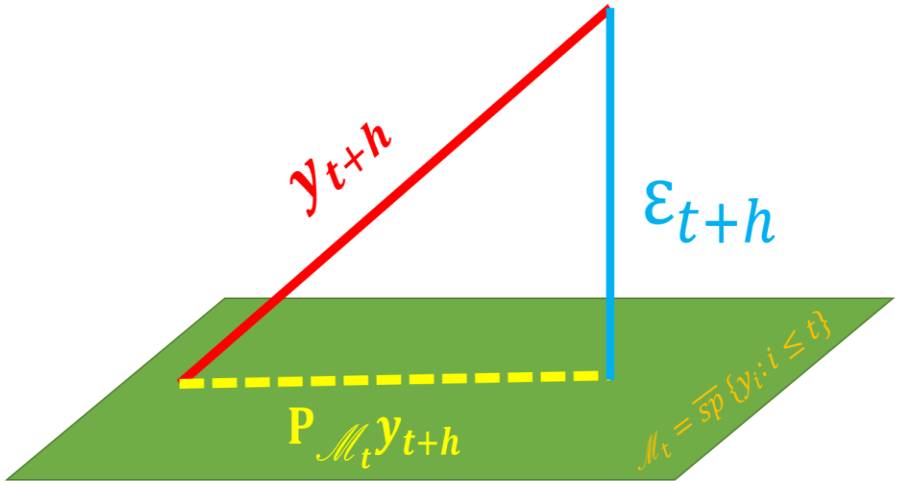


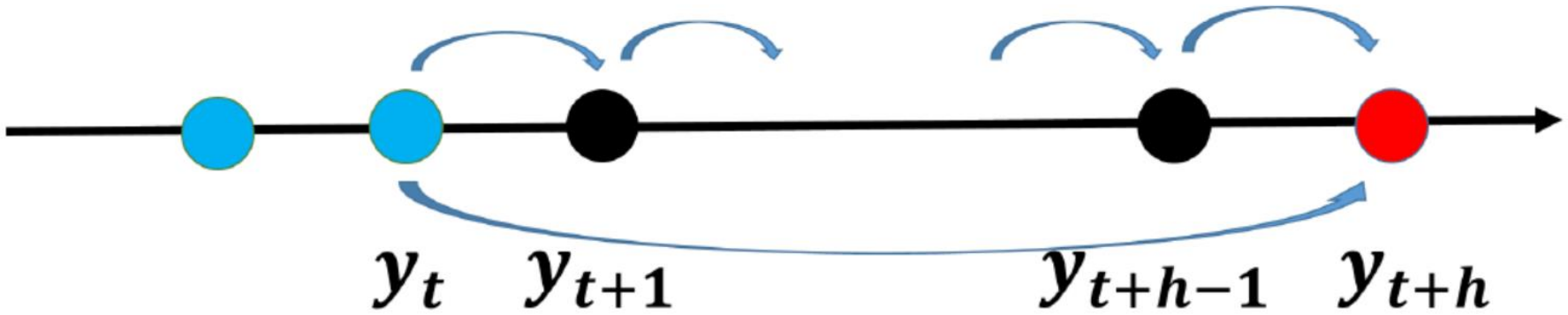
- 预报的准则

  均方误差最小： $\min_g E\left|y_{t+h} - g(y_j : j \le t)\right|^2$ ，对给定的视界水平h.

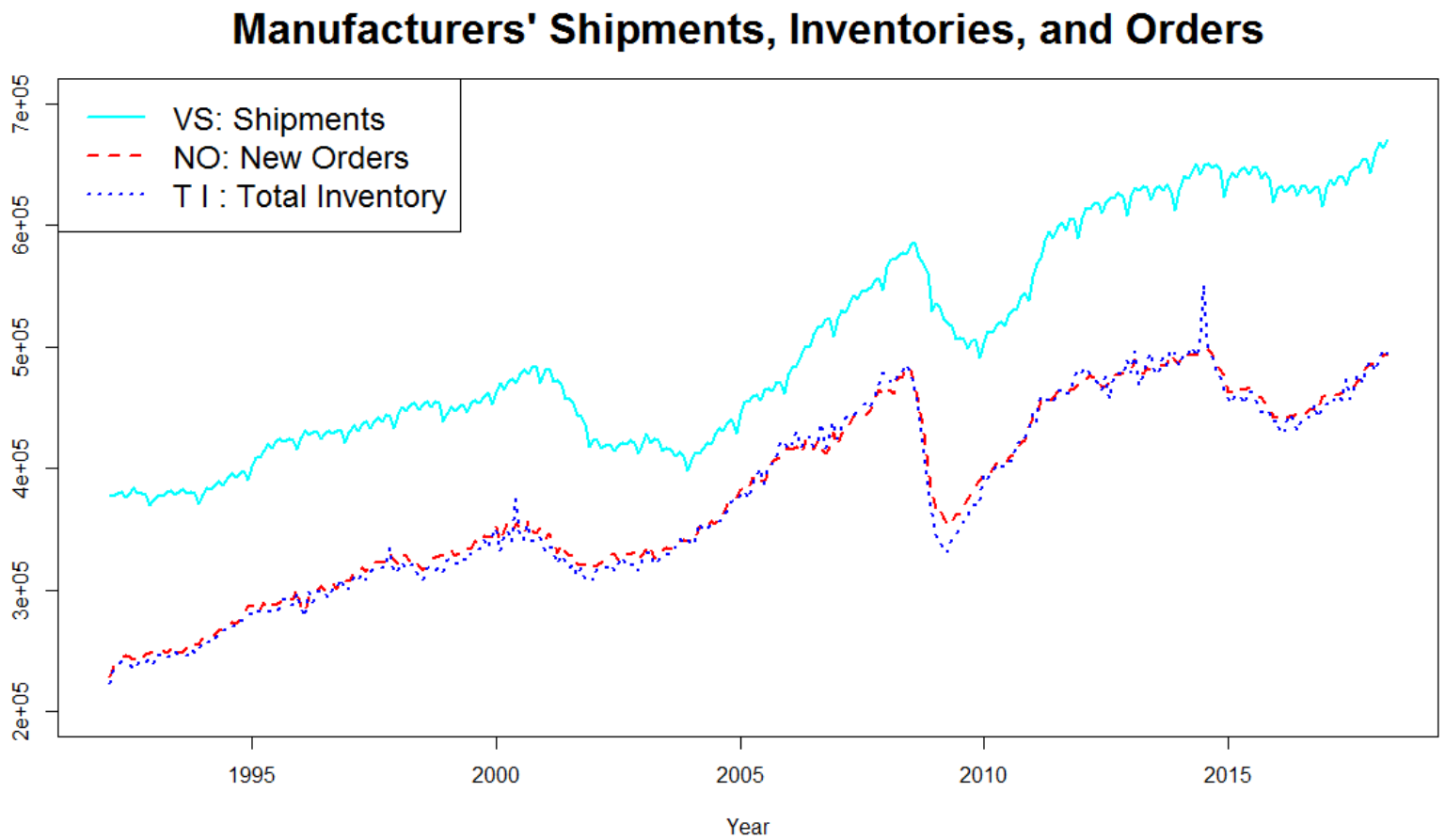
- 常用预报
  - 基于条件期望的预报
  - 基于线性投影的预报






y_t
y_{t+1}
y_{t+h-1}
y_{t+h}

- 案例2.（协整分析）
- 该商务数据是来自美国商务部网站公布的1992年2月到2013年9月的全部制造业的发货、新订单及总库存的月度数据，单位是百万美元：发货价值(VS)、新订单(NO)、总库存(TI).



基础知识 II

- 单位根：如果一个非平稳序列$\{y_t\}$在d次差分后 （即 $\Delta^d y_t\}$）成为平稳序列，则称其为d阶单整的 （integrated of order d），记作 $I(d)$. 比如：
  - 随机游动（random walk）：$y_t = \mu + y_{t-1} + w_t$, 其中$w_t \sim i.i.d.(0, \sigma^2)$.
  - 趋势平稳过程：$y_t = \mu + \beta t + w_t$.

  两种情况都可以写成：$\Delta y_t = \alpha + u_t$ 的形式，其中α为常数，u_t是平稳过程。
- 单位根检验
  - DF检验（Dickey-Fuller test）
  - ADF检验（Augmented DF test）
  - PP检验（Phillips and Perron test）
  - KPSS检验
  - 区别：1.前三种检验的原假设是序列有单位根，在数据量不够或缺乏足够证据时，往往无法拒绝原假设，这时，不能得出“序列有单位根”的结论，应该是“没有足够的证据说明没有单位根”的结论，而不是得到“有证据说不平稳”的结论; 第四种检验的原假设是序列平稳。2. DF检验只能处理一些简单的模型，往往用于1阶自相关模型的单位根检验，对于高阶的过程，需要用ADF检验。3. PP检验在于可以处理序列相关和误差项的异方差性。

- 协整（cointegration）：m-维随机向量$\mathbf{y}_t$称之为(d,b)阶协整的，如果（i）$\mathbf{y}_t$中的每一个分量都是$I(d)$的；（ii）存在向量$\boldsymbol{\beta} \neq 0$,使得$\boldsymbol{\beta}'\mathbf{y}_t \sim I(d-b)$, 其中$b>0$。记作：$\mathbf{y}_t \sim \mathrm{CI}(d,b)$.
  - $\boldsymbol{\beta}'\mathbf{y}_t$称之为“长期均衡关系”（long-run equilibrium relationship）.
- m-维向量自回归模型（vector autoregressive model, VAR, Sims (1980)）：

  $\mathbf{y}_t = \mu_t + \Phi_1\mathbf{y}_{t-1} + \cdots + \Phi_p\mathbf{y}_{t-p} + \varepsilon_t$，其中$\mu_t = \mu_0 + \mu_1 t$.
- 误差校正模型（error-correct model, ECM）：

  $\Delta\mathbf{y}_t = \mu_t + \Pi\mathbf{y}_{t-1} + \Phi_1^*\,\Delta\mathbf{y}_{t-1} + \cdots + \Phi_{p-1}^*\,\Delta\mathbf{y}_{t-p+1} + \varepsilon_t$, 其中

  $\Phi_j^* = -\sum_{i=j+1}^{p}\Phi_i,\ j = 1,2,\ldots,p-1.$

  $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}' = -\Phi(1) = \Phi_p + \cdots + \Phi_1 - I$.

  $\Pi\mathbf{y}_{t-1}$: 误差校正项 (error-corrected term)
  - 优点:
    - ✓ 综合考虑了水平与差分（Level & difference）；
    - ✓ 把数据按照长期、短期效应分类，给出了更多的解释；
    - ✓ 所有长期的信息均包含在矩阵Π中，矩阵Φ_j^*刻画了数据的短期动态.

- Π的秩：
  - Rank(Π)=0 (Difference stationary):

$$\Delta \mathbf{y}_t = \mu_t + \Phi_1^* \Delta \mathbf{y}_{t-1} + \cdots + \Phi_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \varepsilon_t,$$

即$\Delta \mathbf{y}_t$服从带有确定性项μ_t的VAR$(p-1)$.

  - Rank(Π)= m (Stationary): $\Delta \mathbf{y}_t \sim I(0)$.
  - 0<Rank(Π)= $r < m$ (Cointegration) :

$$\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}' = -\Phi(1) = \Phi_p + \cdots + \Phi_1 - I,$$

其中$\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 是列满秩的$m \times r$阶矩阵. 则

$$\Delta \mathbf{y}_t = \mu_t + \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{y}_{t-1} + \Phi_1^* \Delta \mathbf{y}_{t-1} + \cdots + \Phi_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \varepsilon_t.$$

这就意味着$\mathbf{y}_t$有r 个线性独立的协整向量，使得 $w_t = \boldsymbol{\beta}' \mathbf{y}_t \sim I(0)$, 有$m-r$个单位根，即$m-r$个公共随机趋势项. 获得此公共随机趋势项的简单方法：先把$\boldsymbol{\alpha}$ 正交扩充，即寻找$m \times (m-r)$阶矩阵$\boldsymbol{\alpha}_\perp$ 使得$\boldsymbol{\alpha}_\perp' \boldsymbol{\alpha} = 0$. 令$\mathbf{c}_t = \boldsymbol{\alpha}_\perp' \mathbf{y}_t$. 因此

$$\boldsymbol{\alpha}_\perp' \Delta \mathbf{y}_t = \boldsymbol{\alpha}_\perp' \mu_t + \boldsymbol{\alpha}_\perp' \Phi_1^* \Delta \mathbf{y}_{t-1} + \cdots + \boldsymbol{\alpha}_\perp' \Phi_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\alpha}_\perp' \varepsilon_t.$$

故$(m-r)$维序列$\mathbf{c}_t$有$m-r$个单位根.

- 极大似然估计

  给定样本或数据：$\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. 为了估计及其使用方便，通常假设$\mu_t = \mu[1, t]'$.

  假定Π的秩为r. ECM变为：

$$\Delta \mathbf{y}_t = \mu_t + \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{y}_{t-1} + \Phi_1^* \, \Delta \mathbf{y}_{t-1} + \cdots + \Phi_{p-1}^* \, \Delta \mathbf{y}_{t-p+1} + \varepsilon_t.$$

- VAR模型的最小二乘估计

  可以用软件包 vars 中的函数 VAR(). 对应模型的矩阵形式为：

$$\mathbf{X}_t = \mathbf{A}_1 \mathbf{X}_{t-1} + \cdots + \mathbf{A}_p \mathbf{X}_{t-p} + \boldsymbol{\Gamma} \, \mathbf{u}_t + \boldsymbol{\eta}_t.$$

- 高维AR模型
  - 带状高维AR模型
  - 误差校正因子模型

$$\Delta \mathbf{y}_t = \mathbf{C}\mathbf{y}_{t-1} + \mathbf{B}\mathbf{f}_t + \mathbf{e}_t,$$

$$\mathbf{f}_t = \sum_{i=1}^{m} \mathbf{E}_i \, \mathbf{f}_{t-i} + \varepsilon_t.$$

清華大學 Tsinghua University · 统计学研究中心 CENTER FOR STATISTICAL SCIENCE

欢迎关注“水木数据派”