法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

第五讲

时间序列数据分析

--梁斌

目录

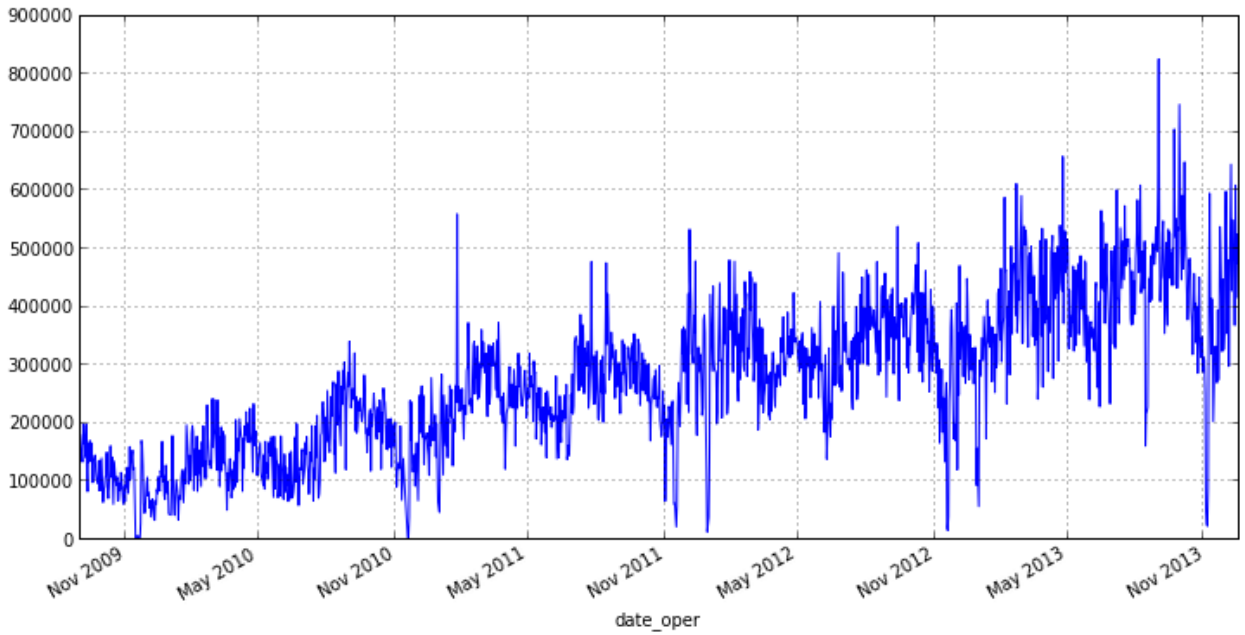
- Python的日期和时间处理及操作
- Pandas的时间序列数据处理及操作
- 时间数据重采样
- 时间序列数据统计—滑动窗口
- 时序模型：ARIMA
- 实战案例：股票数据分析

目录

Python的日期和时间处理

时间序列分类

- 时间戳（timestamp），特定的时刻
- 固定周期（period），某月或某年
- 时间间隔（interval），由起始时间戳和结束时间戳表示。



Python的日期和时间处理

datetime, time及calendar模块

- datetime, 以毫秒形式存储日期和时间
- datime.timedelta，表示两个datetime对象的时间差
- datetime模块中包含的数据类型

类型	说明
date	以公历形式存储日历日期（年、月、日）
time	将时间存储为时、分、秒、毫秒
datetime	存储日期和时间
timedelta	表示两个datetime值之间的差（日、秒、毫秒）

示例代码： 01_python_datetime.ipynb

Python的日期和时间处理

字符串和datetime转换

- datetime -> str,
    1. str(datetime_obj)
    2. datetime.strftime()
- str -> datetime
    1. datetime.strptime() 需要指定时间表示的形式
    2. dateutil.parser.parse() 可以解析大部分时间表示形式
    3. pd.to_datetime() 可以处理缺失值和空字符串

示例代码： 01_python_datetime.ipynb

Python的日期和时间处理

代码	说明
%Y	4位数的年
%y	2位数的年
%m	2位数的月[01, 12]
%d	2位数的日[01, 31]
%H	时（24小时制）[00, 23]
%I	时（12小时制）[01, 12]
%M	2位数的分[00, 59]
%S	秒[00, 61]（秒60和61用于闰秒）
%w	用整数表示的星期几[0（星期天）, 6]
%U	每年的第几周[00, 53]。星期天被认为是每周的第一天，每年第一个星期天之前的那几天被认为是“第0周”
%W	每年的第几周[00, 53]。星期一被认为是每周的第一天，每年第一个星期一之前的那几天被认为是“第0周”
%z	以+HHMM或−HHMM表示的UTC时区偏移量，如果时区为naive[译注3]，则返回空字符串
%F	%Y-%m-%d简写形式，例如2012-4-18[译注4]
%D	%m/%d/%y简写形式，例如04/18/12

字符串和datetime转换

- datetime常用格式定义

目录

Pandas的时间序列处理

- 基本类型，以时间戳为索引的Series -> DatetimeIndex
- 创建
  1. 指定index为datetime的list
  2. pd.date_range()
- 运算仍然符合按索引对齐，即按时间索引对齐运算
- 索引
  1. 索引位置
  2. 索引值
  3. 可以被解析的日期字符串
  4. 按“年份”、“月份”索引
  5. 切片操作
- 过滤 truncate

示例代码： 02_pandas_time.ipynb

Pandas的时间序列处理

- 生成日期范围 pd._date_range()
  1. 传入开始、结束日期，默认生成的该时间段的时间点是按天计算的 (频率是D)
  2. 只传入开始或结束日期，还需要传入时间段
  3. 规范化时间戳 normalize=True
- 频率Freq，由基础频率的倍数组成，基础频率包括：
  1. BM: business end of month，每个月最后一个工作日
  2. D: 天, M: 月 等
- 偏移量，每个基础频率对应一个偏移量
  1. 偏移量通过加法连接
- 移动数据 (shifting)，沿时间轴将数据前移或后移，保持索引不变

示例代码： 02_pandas_time.ipynb

Pandas的时间序列处理

- 基础频率Freq

别名	偏移量类型	说明
D	Day	每日历日
B	BusinessDay	每工作日
H	Hour	每小时
T或min	Minute	每分
S	Second	每秒
L或ms	Milli	每毫秒（即每千分之一秒）
U	Micro	每微秒（即每百万分之一秒）
M	MonthEnd	每月最后一个日历日
BM	BusinessMonthEnd	每月最后一个工作日
MS	MonthBegin	每月第一个日历日

示例代码： 02_pandas_time.ipynb

Pandas的时间序列处理

- 基础频率Freq (续)

BMS	BusinessMonthBegin	每月第一个工作日
W-MON、W-TUE…	Week	从指定的星期几（MON、TUE、WED、THU、FRI、SAT、SUN）开始算起，每周
WOM-1MON、WOM-2MON…	WeekOfMonth	产生每月第一、第二、第三或第四周的星期几。例如，WOM-3FRI表示每月第3个星期五
Q-JAN、Q-FEB…	QuarterEnd	对于以指定月份（JAN、FEB、MAR、APR、MAY、JUN、JUL、AUG、SEP、OCT、NOV、DEC）结束的年度，每季度最后一月的最后一个日历日
BQ-JAN、BQ-FEB…	BusinessQuarterEnd	对于以指定月份结束的年度，每季度最后一月的最后一个工作日

示例代码： 02_pandas_time.ipynb

Pandas的时间序列处理

时间周期计算

- Period类，通过字符串或整数及基础频率构造
- Period对象可进行数学运算，但要保证具有相同的基础频率
- period_range，创建指定规则的时间周期范围，生成PeriodIndex索引，可用于创建Series或DataFrame
- 时间周期的频率转换，asfreq
  - 如：年度周期->月度周期
- 按季度计算时间周期频率

示例代码： 02_pandas_time.ipynb

目录

时间数据重采样

重采样 (resampling)

- 将时间序列从一个频率转换到另一个频率的过程，需要聚合
- 高频率->低频率，downsampling，相反为upsampling
- pandas中的resample方法实现重采样
  - 产生Resampler对象
  - reample(freq).sum(), resampe(freq).mean(), ...

降采样 (downsampling)

- 将数据聚合到规整的低频率
- OHLC重采样, open, high, low, close
- 使用groupby降采样

示例代码： 03_resample.ipynb

时间数据重采样

升采样 (upsampling)

- 将数据从低频转到高频，需要插值，否则为NaN
- 常用的插值方法
  1. ffill(limit), 空值取前面的值填充，limit为填充个数
  2. bfill(limit)，空值取后面的值填充
  3. fillna（‘ffill’）或fillna（‘bfill’），
  4. interpolate，根据插值算法补全数据

     具体可以参考： http://pandas.pydata.org/pandas-docs/stable/generated/pandas.tseries.resample.Resampler.interpolate.html#pandas.tseries.resample.Resampler.interpolate

示例代码： 03_resample.ipynb

目录

滑动窗口

滑动窗口函数(moving window function)

- 在时间窗口上计算各种统计函数
- 窗口函数 (window functions)
  1. 滚动统计 (rolling)

     obj.rolling().func
  2. window

     窗口大小
  3. center

     窗口是否居中统计

Method	Description
count()	Number of non-null observations
sum()	Sum of values
mean()	Mean of values
median()	Arithmetic median of values
min()	Minimum
max()	Maximum
std()	Bessel-corrected sample standard deviation
var()	Unbiased variance
skew()	Sample skewness (3rd moment)
kurt()	Sample kurtosis (4th moment)
quantile()	Sample quantile (value at %)
apply()	Generic apply
cov()	Unbiased covariance (binary)
corr()	Correlation (binary)

示例代码： 04_mv_win_func.ipynb

目录

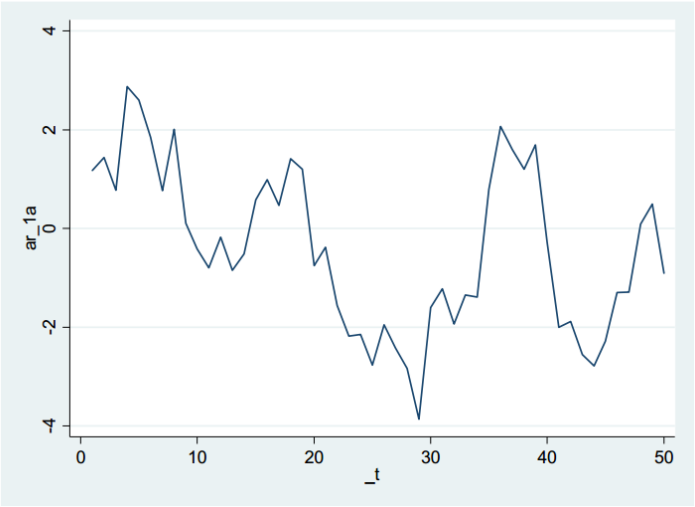
时序模型：ARIMA

AR（Autoregressive）模型

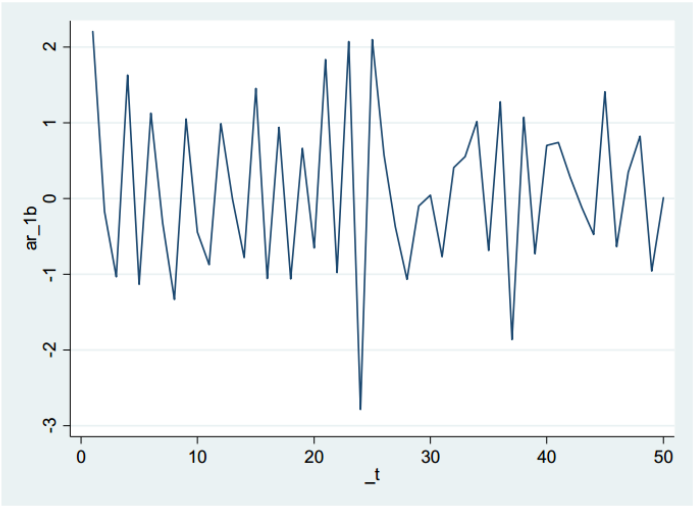
- 自回归模型描述的是当前值与历史值之间的关系
- 滞后p阶的AR模型AR(p)： $y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$

  μ 是常数，γ_p 是t-p时刻滞后变量的系数，ϵ_t 是误差
- AR(1): $y_t = \mu + \gamma y_{t-1} + \epsilon_t$

AR(1) with $\gamma = 0.8$



AR(1) with $\gamma = -0.8$

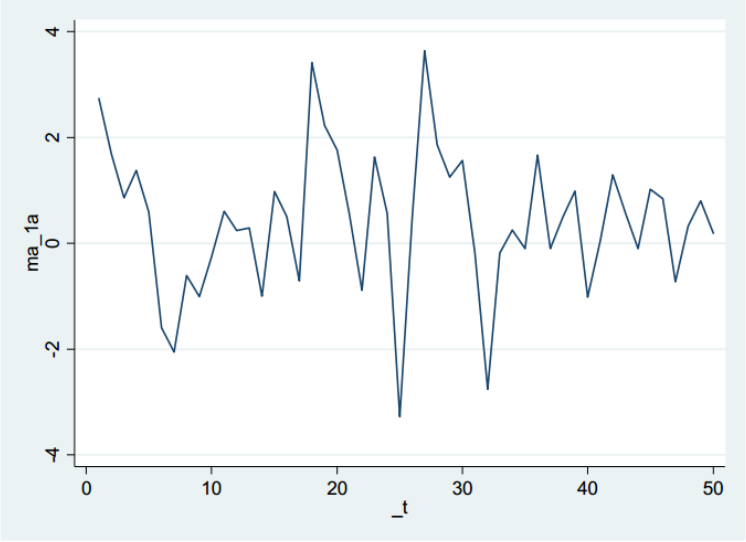


时序模型：ARIMA

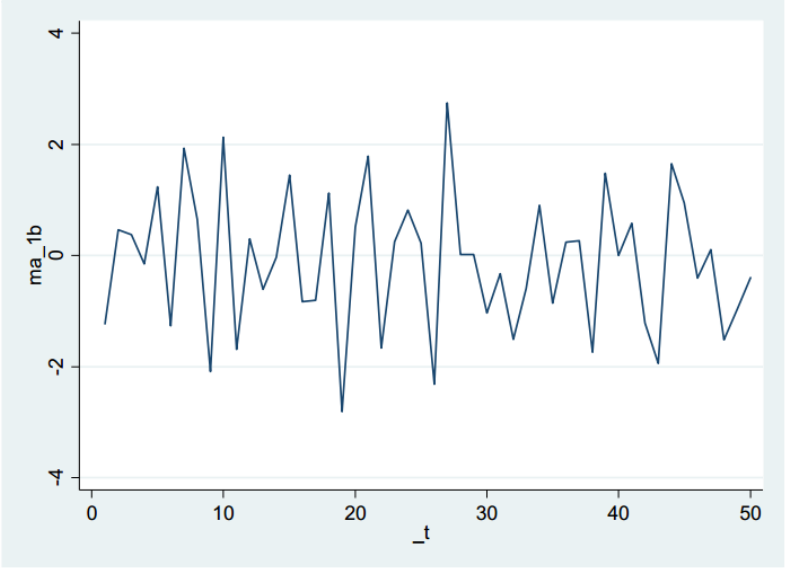
MA (Moving average) 模型

- 滑动平均模型描述的事自回归部分的误差累计
- MA(q): $y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$
- MA(1): $y_t = \mu + \epsilon_t + \theta \epsilon_{t-1}$

MA(1) with $\theta = 0.7$



MA(1) with $\theta = -0.7$



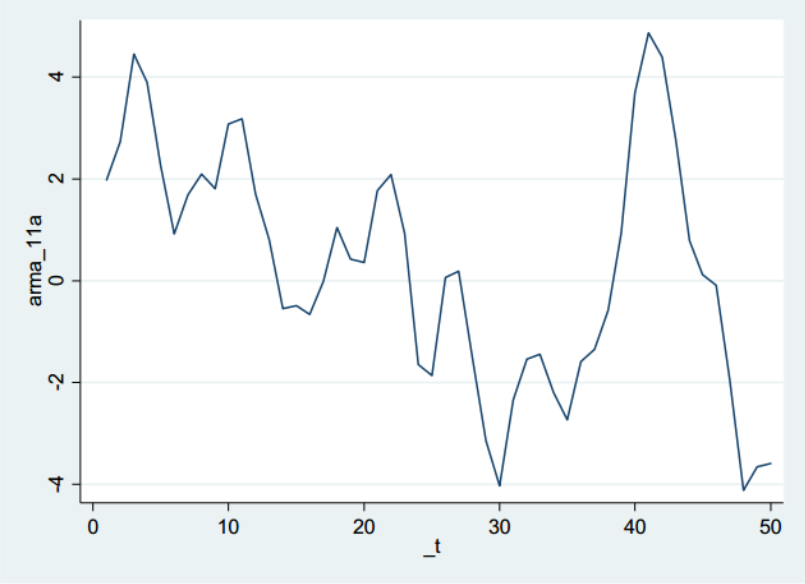
时序模型：ARIMA

ARMA (Autoregressive moving average) 模型

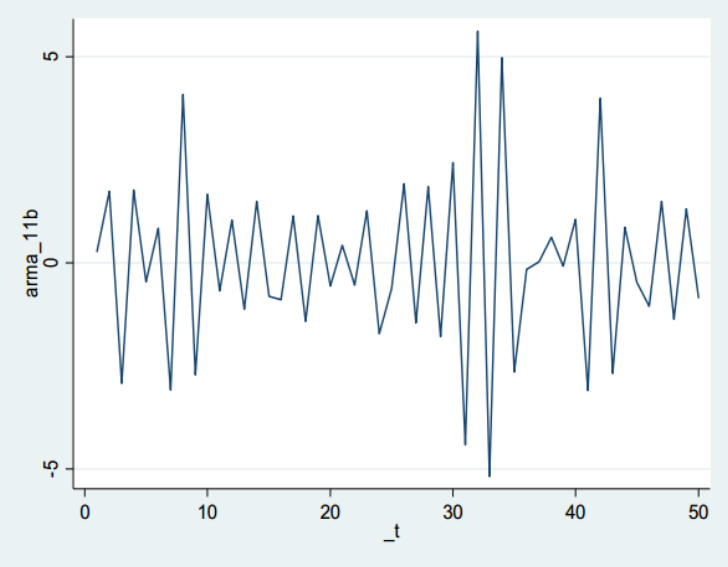
- AR与MA的结合 ARMA(p, q)

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

ARMA(1,1) with $\gamma = 0.8$ and $\theta = 0.7$



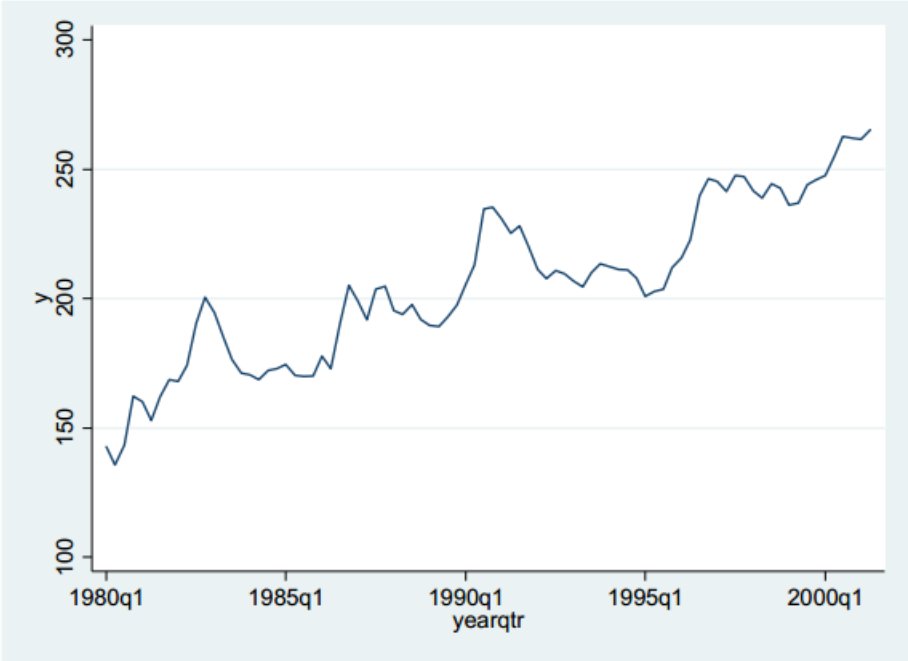
ARMA(1,1) with $\gamma = -0.8$ and $\theta = -0.7$



时序模型：ARIMA

平稳性

- ARMA模型要求时间序列是平稳的
- 一个时间序列，如果均值没有系统的变化（无趋势）、方差没有系统变化，且严格消除了周期性变化，就称为是平稳的



平稳？
非平稳？

示例代码： 05_arima.ipynb

时序模型：ARIMA

平稳性

- 严平稳
  - 如果对所有时刻t，任意整数k和任意k个正整数

    的联合分布与 $(t_1, t_2 \ldots\ldots t_k)$ 的联合分布相同，则序列$\{r_t\}$是

    严平 $(r_{t_1}, r_{t_2}, \ldots\ldots r_{t_k})$
- 弱平稳 $(r_{t_1+t}, r_{t_2+t}, \ldots\ldots r_{t_k+t})$
  - 若时间序列$\{r_t\}$满足两个条件：

    即序列的均值、r_t与r_{t-l}的协方差不随时间而改变，则序列是弱平稳的
  - 现实中所处理的时序通常是弱平 $E(r_t) = \mu$ $Cov(r_t, r_{t-l}) = \gamma_l$

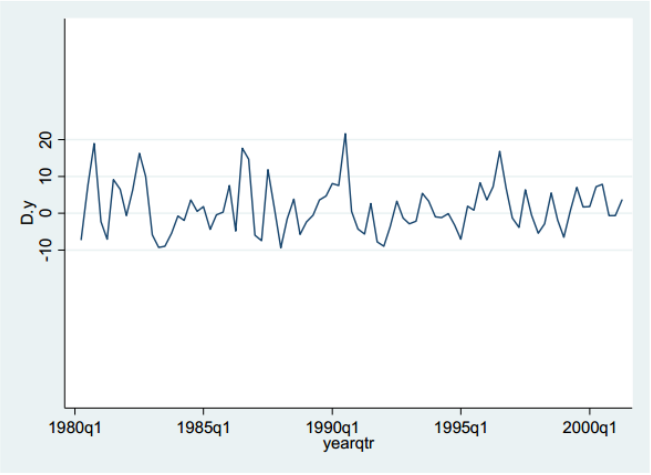
示例代码： 05_arima.ipynb

时序模型：ARIMA

差分

- 由时序 $\{r_t\}$ 在 t 时刻的值 r_t 与 t-1 时刻的值 r_{t-1} 的差 d_t 构造的新序列$\{d_t\}$为一阶差分。对一阶差分序列$\{d_t\}$进行相同的差分运算，可以得到二阶差分...
- 通常非平稳序列可以经过d阶差分得到弱平稳或近似弱平稳的时间序列
- d阶差分表示为I(d)
- ARIMA(p, d, q)模型：p阶自回归滞后项，q阶滑动平均滞后项，d阶差分

Differenced variable: $\Delta y_t = y_t - y_{t-1}$



示例代码： 05_arima.ipynb

时序模型：ARIMA

相关系数

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

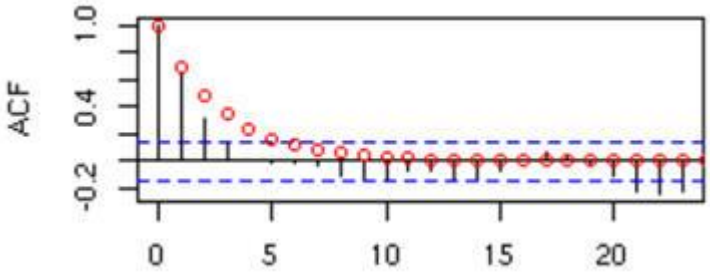
- 反映向量空间中两个向量之间相关关系密切程度的统计指标
  - 两个向量平行且同向，系数为1；平行且反向，系数为-1
  - 两个向量垂直，不相关，系数为0
  - 向量间夹角越小，相关系数越接近1，相关性越高

时序模型：ARIMA

自相关函数(Autocorrelation Function, ACF)

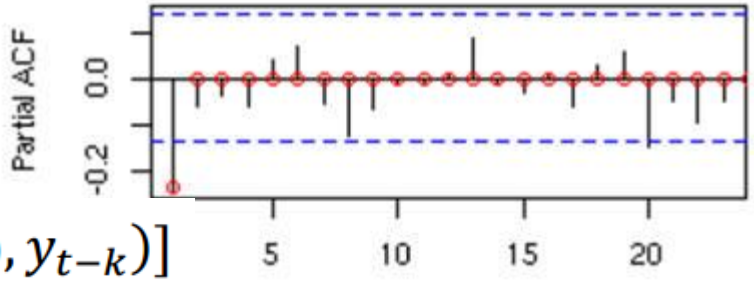
- 描述随机信号{r_t}在任意两个不同时刻t1，t2，的取值之间的相关程度

$$ACF(k) = \rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\mathrm{Var}(y_t)}$$



偏自相关函数(Partial Autocorrelation Function, PACF)

- 阶次为s的偏自相关：去除信号中所有滞后期小于s的信号影响后，当前信号与滞后s阶的信号之间的关系



$$\rho_k^* = \mathrm{Corr}[y_t - E^*(y_t|y_{t-1}, \cdots, y_{t-k+1}), y_{t-k})]$$

时序模型：ARIMA

ACF PACF 性质

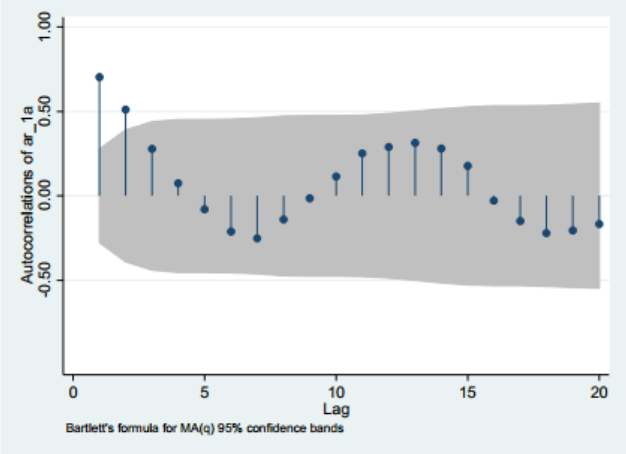
	AR(p)	MA(q)	ARMA(p, q)
ACF	趋势衰减	q阶后截尾	趋势衰减
PACF	p阶后截尾	趋势衰减	趋势衰减

示例代码： 05_arima.ipynb

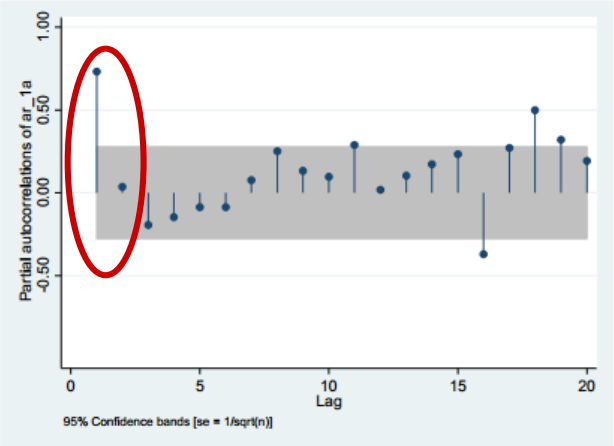
时序模型：ARIMA

AR(1)

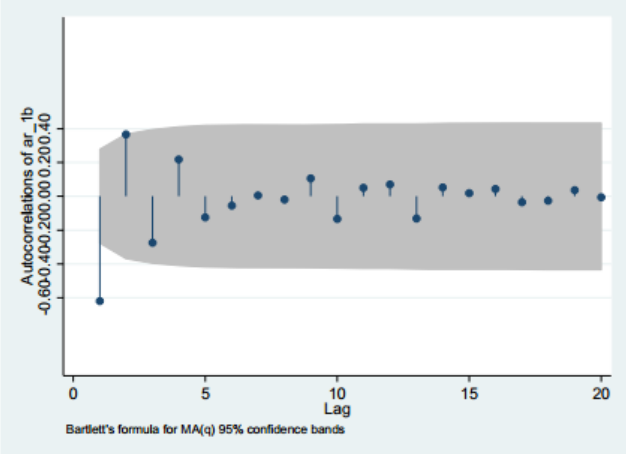
ACF of AR(1) with coefficient 0.8



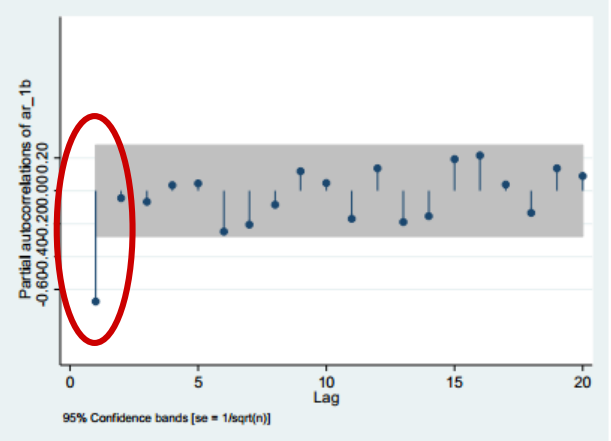
PACF of AR(1) with coefficient of 0.8



ACF of AR(1) with coefficient -0.8



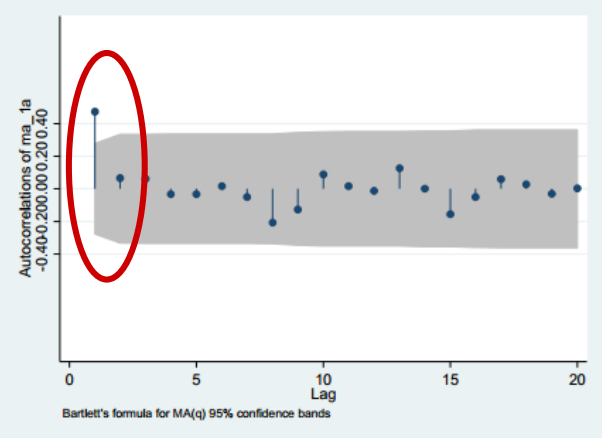
PACF of AR(1) with coefficient of -0.8



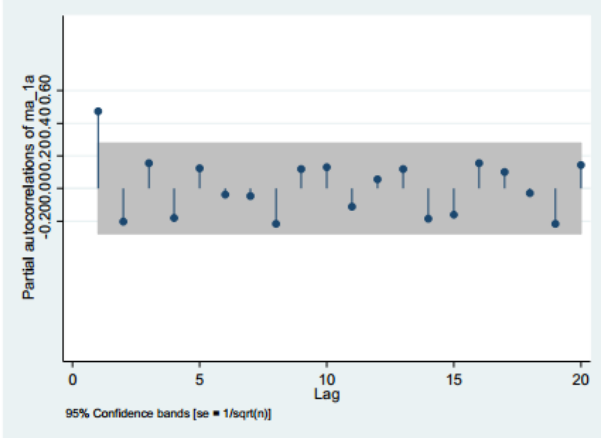
时序模型：ARIMA

MA(1)

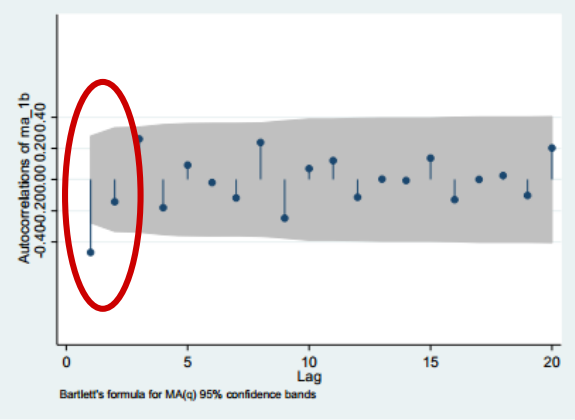
ACF of MA(1) with coefficient of 0.7



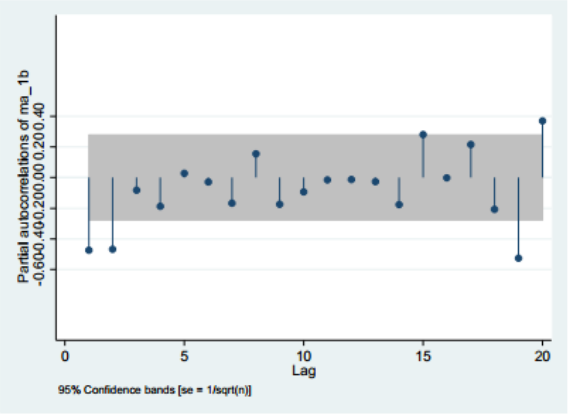
PACF of MA(1) with coefficient of 0.7



ACF of MA(1) with coefficient of -0.7



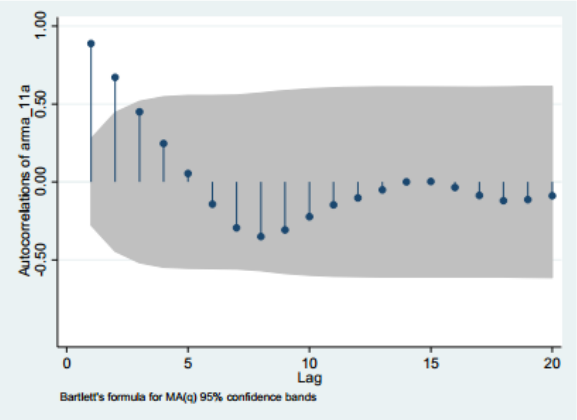
PACF of MA(1) with coefficient of -0.7



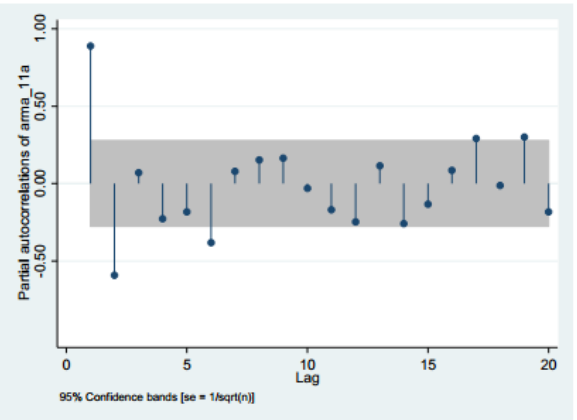
时序模型：ARIMA

ARMA(1,1)

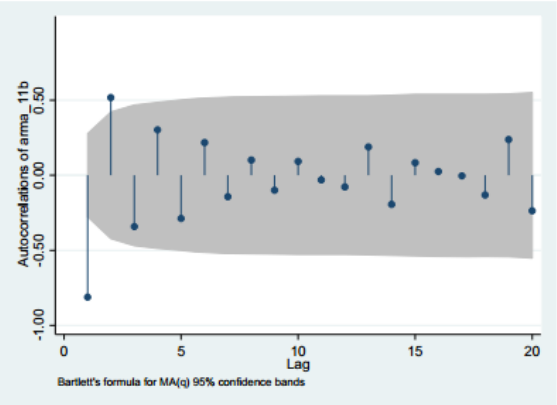
ACF of ARMA(1,1) with coeff 0.8 and 0.7



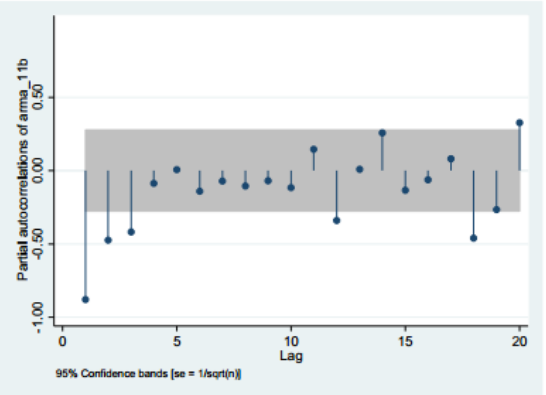
PACF of ARMA(1,1) with coeff 0.8 and 0.7



ACF of ARMA(1,1) with coeff −0.8 and −0.7



PACF of ARMA(1,1) with coeff −0.8 and −0.7



时序模型：ARIMA

ARIMA模型参数选择

1. 检查序列是否平稳
   - 若不平稳，使用差分平稳化序列，确定差分阶数d
2. ARMA定阶
   - 通过PACF确定AR的阶数p
   - 通过ACF确定MA的阶数q
3. 根据参数p, d, q建立模型ARIMA(p, d, q)

目录

实战案例：股票数据分析

- pandas_datareader

  安装 pip install pandas_datareader
- 通过pandas_datareader可以获取yahoo财经，Google财经，world bank等数据接口提供的股票数据

步骤

1. 准备数据
2. 可视化数据，审查数据
3. 处理数据（是否需要平稳处理）
4. 根据ACF, PACF定阶
5. 拟合ARIMA模型
6. 预测

示例代码： lect05_proj

参考

- Pandas数据重采样
  http://pandas.pydata.org/pandas-docs/stable/api.html#resampling
- Pandas滑动窗口函数
  http://pandas.pydata.org/pandas-docs/stable/computation.html#window-functions
- ARIMA模型详细讲解
  https://people.duke.edu/~rnau/411arim.htm
- ARIMA模型
  https://www.otexts.org/fpp/8
- pandas-reader模块
  https://pandas-datareader.readthedocs.io/en/latest/

参考

- Python时间序列预测案例

  http://it.sohu.com/20160320/n441240758.shtml

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

小象问答 @Robin_TY

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop