政治经济学前沿方法论与量化分析

第五讲 平稳时间序列的建模

上课地点： 善斋306C
上课时间：周二第六大节

龙治铭
善斋307C
zhiminglong@tsinghua.edu.cn

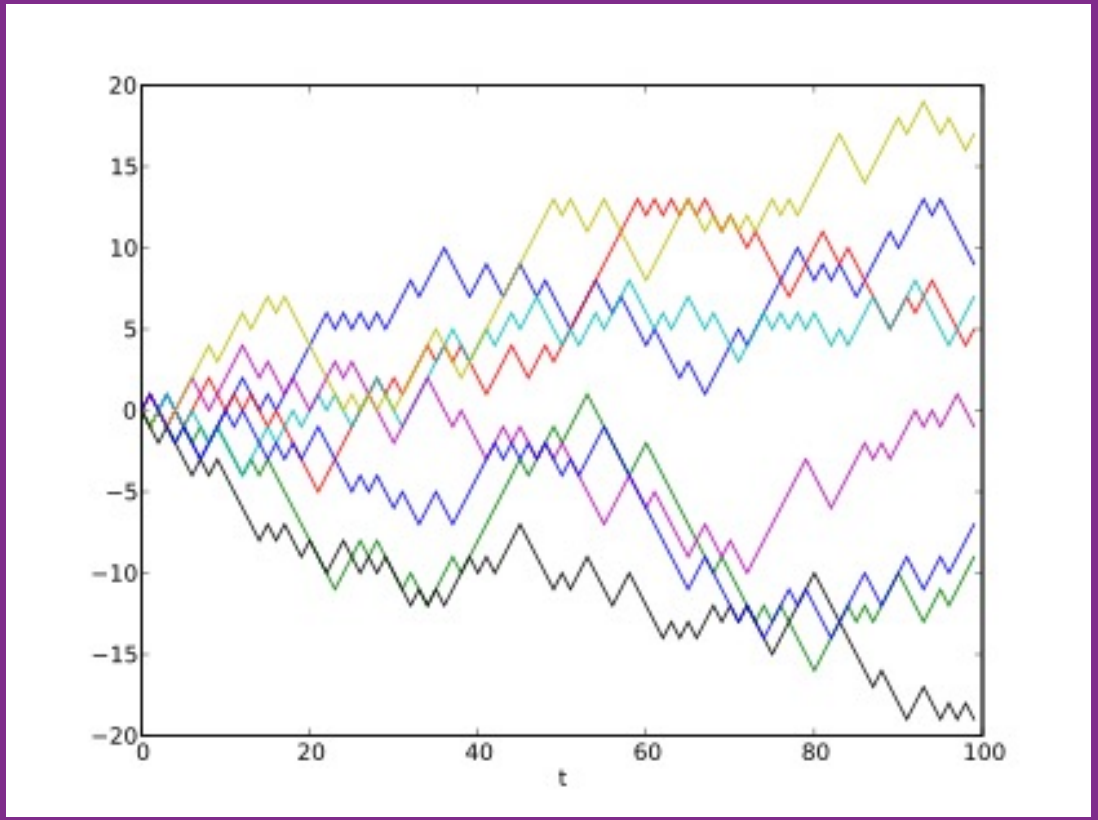
清華大學
Tsinghua University

目录

CONTENTS

1

伪回归问题和时间序列的平稳性



两个随机游走之间不可能存在任何因果关系，但OLS回归可能存在很高的R2

时间序列

※数据的三大类型：横截面数据（同一时间，同一空间，普通的OLS回归）、时间序列（time series）和面板数据（Panel data，两个维度：时间和空间）

※什么是时间序列？
时间序列变量：X_1、X_2、…X_T每一个都是随机变量，记$\{X_t\}_{t=1}^{T}$为时间序列
随机变量的取值：我们进行一次观察或实验，记录下一组观测值x_1、x_2、…x_T，$\{x_t\}_{t=1}^{T}$为时间序列的取值
和横截面数据的区别：横截面数据变量x是一个变量，x_1、x_2、…x_T是重复实验或抽样得到的T个观测值
例：每日股市收盘价是一个随机变量，从1991年至今是6000多个随机变量而不是一个变量！

※时间序列的特点：
有的事件可以在相同条件下重复进行，$\{x_t\}_{t=1}^{T}$有多组
有的事件不可能重复进行，尽管X_t是随机变量，但我们只可能观测到一个值，比如气温、股价、GDP等等。

※时间序列建模的基本假设：
1）遍历各态性（Ergodicity）:对于只可能有一个观测值的随机变量，我们如何得到它的数字特征、分布律等信息？假定时间足够长，样本空间里的每一个事件都会发生，因此可以用时间平均来代替集平均。
经济学含义：历史会重复自身，过去发生的事情未来必定会发生(后凯恩斯主义经济学的批评)。
物理学家：直接用，否则没办法研究问题
2）平稳性（Stationarity）:时间序列有某种规律其统计特性不随时间而变化，引入平稳性的概念，简单地说就是均值较为稳定。

伪回归（SPURIOUSREGRESSIONS）

※随机游走（Random walk）：$x_t = x_{t-1} + \varepsilon_t$, 其中$\varepsilon_t \sim WN$
随机游走的结果是没有规律的（不平稳），因为后来的取值取决于初始值和一个白噪音，而白噪音已经不不能从其他变量中得到多余的信息。

※ Granger and Newbold (1974) 用蒙特卡洛模拟的办法证明伪回归的存在
这个idea的绝妙之处：两个独立的随机游走之间不可能存在线性相关性，从理论上讲OLS估计量必然不显著且R2为0.

$$x_t = x_{t-1} + \varepsilon_t \quad \varepsilon_t \sim WN$$
$$y_t = y_{t-1} + v_t \quad v_t \sim WN$$

回归方程为：
$$y_t = \alpha + \beta 8\, x_t$$

Granger and Newbold重复了100次，本来应该为0的R2，经常很高（右侧表2，倒数第二列）

※理论证明：渐进一致理论（Consistent theory）,证明从略。
主要结论：
1）此时残差值不是随机扰动项的无偏估计，常用的基于残差值的分布的validation不成立（预测也不成立）
2）OLS估计量不再服从学生分布，显著性判断不再可靠（常用的F检验，卡方检验等也不再适用）

Table 2

Regressions of a series on m independent 'explanatory' series.

Series either all random walks or all A.R.I.M.A. (0, 1, 1) series, or changes in these. $Y_0 = 100$, $Y_t = Y_{t-1} + a_t$, $Y_t' = Y_t + kb_t$; $X_{j,0} = 100$, $X_{j,t} = X_{j,t-1} + a_{j,t}$ $X_{j,t}' = X_{j,t} + kb_{j,t}$; $a_{j,t}, a_{i,t}, b_{i,t}, b_{j,t}$ sets of independent $N(0, 1)$ white noises. $k = 0$ gives random walks, $k = 1$ gives A.R.I.M.A. (0, 1, 1) series. H_0 = no relationship, is true. Series length = 50, number of simulations = 100, R^2 = corrected R^2.

		Per cent times H_0 rejected*	Average Durbin–Watson d	Average R^2	Per cent $R^2 > 0.7$
		Random walks			
Levels	$m = 1$	76	0.32	0.26	5
	$m = 2$	78	0.46	0.34	8
	$m = 3$	93	0.55	0.46	25
	$m = 4$	95	0.74	0.55	34
	$m = 5$	96	0.88	0.59	37
Changes	$m = 1$	8	2.00	0.004	0
	$m = 2$	4	1.99	0.001	0
	$m = 3$	2	1.91	−0.007	0
	$m = 4$	10	2.01	0.006	0
	$m = 5$	6	1.99	0.012	0
		A.R.I.M.A. (0, 1, 1)			
Levels	$m = 1$	64	0.73	0.20	3
	$m = 2$	81	0.96	0.30	7
	$m = 3$	82	1.09	0.37	11
	$m = 4$	90	1.14	0.44	9
	$m = 5$	90	1.26	0.45	19
Changes	$m = 1$	8	2.58	0.003	0
	$m = 2$	12	2.57	0.01	0
	$m = 3$	7	2.53	0.005	0
	$m = 4$	9	2.53	0.025	0
	$m = 5$	13	2.54	0.027	0

*Test at 5% level, using an overall test on R^2.

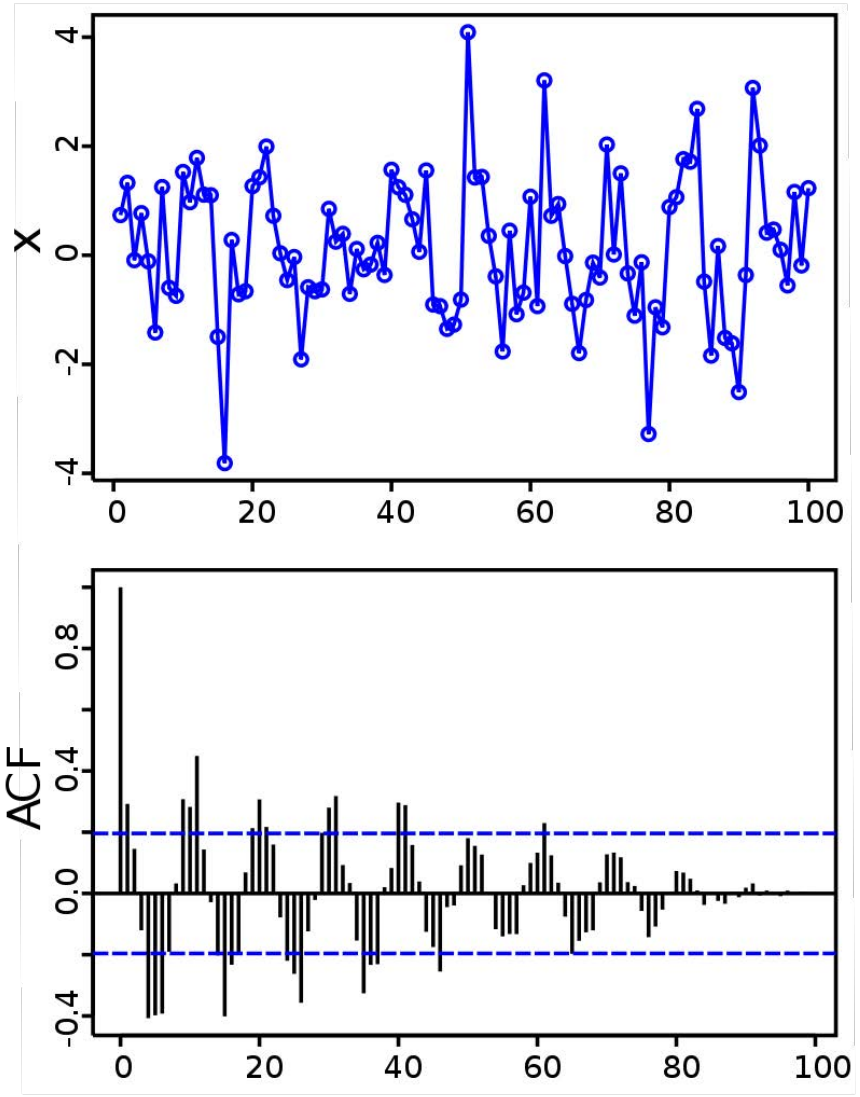
平稳性（Stationarity）

※时间序列有某种规律其统计特性不随时间而变化，表明它具有一定的平稳性。

强平稳 （严格平稳）：一个时间序列$\{X_t\}_{t=1}^{T}$的T个的变量的分布都是稳定的
弱平稳（二阶平稳）：均值稳定，方差稳定

如何判断时间序列是否平稳：除了看图还有：
※自相关函数（autocorrelation function）：相关性公式应用于时间间距为k的两个变量上，得到的相关系数即为自相关函数。
自相关函数给出了关于时间序列平稳性等重要信息（下降缓慢表示可能不平稳）。
偏自相关函数（Partial autocorrelation function）

※谱密度（Spectral density）
自相关函数的傅里叶变换就是谱密度，谱密度给出的信息和自相关函数是一样的（证明从略）。时域和频域是等价的，不过人们通常更习惯时域的表达方式，频谱分析在分析某些问题，如经济周期方面具有一定优势。（见我的博士论文第七章和第八章）



两种时间序列：TS和DS

※其他类型的伪回归：学堂路两侧的树每年长高，美国GDP每年增加，二者之间的回归，相关性和R2都可能显著和很高，很有可能仅仅是因为各自的趋势项（时间变量t）拟合得很好。
具有趋势的时间序列之间的高拟合度，很有可能仅仅是因为两个t而已。

※ Nelson (1980) 区别了两种非平稳的时间序列：
1）Trend stationary process（TS），从时间序列中移除趋势项后可以变为平稳
2）Difference stationary process (DS)，将时间序列差分后可以变为平稳
显然，随机游走是DS

※为什么要区别两种时间序列？Nelson and Kang (1981，Econometrica) 证明，在建模时区分这两种不同类型的时间序列，并找到准确的去趋势项方法是至关重要的。
1）将TS当作DS（即差分），随机扰动项人为引入短周期
2）将DS当作TS（即在回归中加入时间变量t），随机扰动项人为引入长周期

我的博士论文第二章进一步证明：将DS当作TS，趋势项的OLS 估计量趋近于0（去趋势的方法没有任何用处），而变量的OLS 估计量都会变得显著（样本足够大，任何变量都是显著的）

※观察自相关函数和谱密度，无法判断时间序列是否为DS(是否包含单位根)，需要用单位根检验。

第一代单位根检验Dickey-Fuller test

※随机游走模型为：$y_t = y_{t-1} + u_t$ $u_t \sim WN$ 我们称时间序列y_t包含单位根
假设时间序列的模型为：

$$y_t = \rho y_{t-1} + u_t$$

如果ρ=1，那么y_t包含单位根，因此检验单位根本质就是检验ρ=1，但前面指出，这时经典的Wald test 不再适用，需要构造新的检验和统计量。

※H0：注意！单位根检验里H0为：存在单位根（因为存在单位根而被当成不存在，危害害较大）

Model 1: $\Delta y_t = \delta y_{t-1} + u_t$

Model 2: $\Delta y_t = a_0 + \delta y_{t-1} + u_t$

Model 3: $\Delta y_t = a_0 + a_1 t + \delta y_{t-1} + u_t$

单位根的检验总是从模型3开始，即先检验模型是否包含趋势项，如果趋势项显著，就使用模型3进行临界值判断，如果趋势项不显著，则跳到模型2。再检查模型2中的常数项是否显著，如果常数项显著，就使用 模型2进行临界值判断，如果常数项不显著，就使用模型1进行临界值判断。

※判断：统计量（A）DF分布形式较为复杂，跟样本容量和模型形式都有关，一般的计量软件都会给出相应的临界值和p-value。
若ADF<临界值，拒绝原假设，不存在单位根，平稳 （注意！拒绝域跟通常的显著性检验是相反的）
若ADF>临界值，不能拒绝原假设，存在单位根，不平稳 （注：临界值都为负，ADF>0，不平稳）

ADF test和其他单位根检验

※ y_t不太可能是一个简单的AR(1)，更一般地：

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

检验方法相同，从具有趋势项的模型开始。
问题：涉及到滞后阶数p的选择。

Table 3.9 **Selection criteria for truncation setting of the number of delays (for T = 24)**

	Lardic and Mignon (2002)	Newey and West (1994)	Schwert (1989)
Formula(s)	$L = T^{1/4}$	$L = int\ [4\ (\frac{T}{100})^{2/9}]$	$l_4 = int\ [4\ (\frac{T}{100})^{1/4}]$, $l_{12} = int\ [12\ (\frac{T}{100})^{1/4}]$
Delay(s)	2.2	2	2 and/or 8

※截面参数p的选择方法：
方法1：利用信息准则自动选择最优值 AIC SIC HQ。缺点：不同信息准则给出的结果通常不一致。
方法2：根据样本容量大小选择固定的滞后阶数。缺点：检验结果可能跟预期效果不同。
常用的标准：Newey–West (1994), Schwert (1989), Mignon and Lardic (2002) 等
解决办法：见我的博士论文第五章及其附录：

Appendix 5.1 **Unit root tests applied to the different variables of the estimated models**

The correlograms suggested that all the series in the regressions are stationary. So, unit root tests were performed on all the variables used in our estimates. As is known, their results depend on the size of the sample, but also and above all on the choice of the truncation setting of the parameter number of lags of the autocorrelation function. According to the three formulas of the selected criteria (Schwert [1989], Newey and West [1994], Lardic and Mignon [2002]), the value obtained here is 3. So we fix the number of lags at 3 in a first stage. As the use of the Schwert (1989) criteria also suggests a different number of lags (10), we set in a second step – in the event that the first one would not be successful –

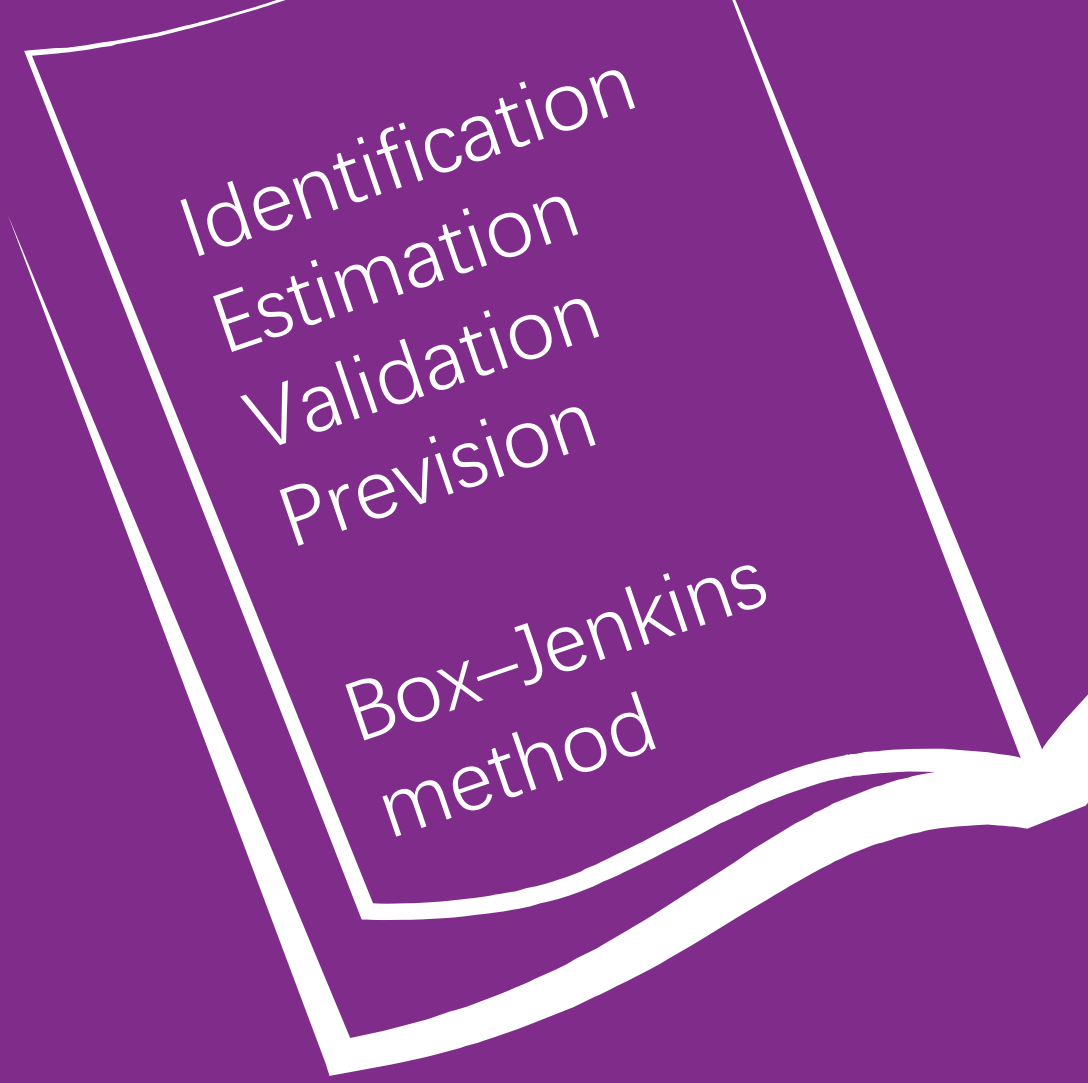
$$L_{maxi} = \max\left\{T^{\frac{1}{4}},\ int\left[4\left(\frac{T}{100}\right)^{\frac{1}{4}}\right],\ l12 = int\left[12\left(\frac{T}{100}\right)^{\frac{1}{4}}\right],\ l = int\left[4\left(\frac{T}{100}\right)^{\frac{2}{9}}\right]\right\} \quad (A5.1)$$

We use the maximum lag L_{maxi} = 10 in the unit root tests, then the information criteria (AIC SIC HQ and their modified forms) to determine the optimum lag L_{opti}. Critical values shown in the following

※其他单位根检验：
Phillips–Perron test
Zivot–Andrews test （断点）
ESR Test
KPSS test （注意H0：平稳）
单位根检验步骤大多相同，都要从包含趋势项的模型开始（**单位根检验争议很大，要多用几个检验**）
其他第二代单位根检验如贝叶斯方法等

2

ARMA模型的理论基础



ARMA模型

※ARMA是什么意思？Autoregressive–moving-average 自回归移动平均

p阶自回归模型AR（p):

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + u_t$$ 其中u_t ~N（0，σ^2）（白噪音即可）

基本思想：一个时间序列可以用过去解释未来

与横截面数据回归额区别：不同变量间的同期相关性vs同一变量不同时期的相关性

※q阶移动平均模型MA（q）：

$$X_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t$$ 其中v_t ~N（0，σ^2）（白噪音即可）

基本思想：一个时间序列的取值由过去外部冲击的影响累积造成

※ AR与MA的联系：任意一个平稳时间序列都可以写为MA(∞) (证明从略)，任意一个AR（p)模型都有一个等价的MA（∞），任意一个

※ ARMA模型

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

短期记忆：AR模型,ARMA模型

长期记忆：ARFIMA模型(Autoregressive fractionally integrated moving average)，分形

永久记忆：时间序列包含单位根，外部冲击的影响是永久的：ARIMA（Integrated差分至平稳）

ARMA与其他变量不同期相结合：结构模型、VAR

ARMA模型的稳定条件

※OLS要求解释变量严格外生且不共线，MA模型是OK的，AR模型中X_{t-1}、X_{t-2}、X_{t-p}相互联系，可能有多重共线性。φ_1不能过于接近于±1

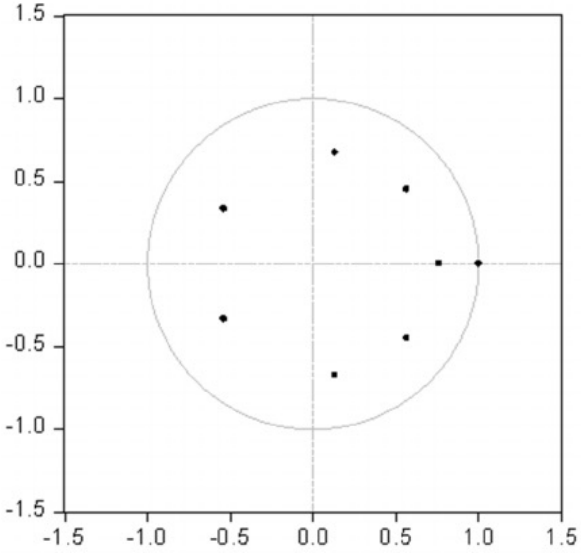
※ARMA模型的稳定条件：φ_1接近于1，时间序列接近随机游走，模型是不稳定的。可以证明（证明从略，见Sargent, 1987或Hamilton，1994第一章（不完整）），模型的稳定条件为：AR(p) 对应的特征方程的p个根在复数域内均位于单位圆内。

原因：AR模型本质是（向后）迭代方程，p阶滞后，意味着p次迭代，会得到p次方程。根据高斯代数基本定理和其他代数知识可以得到证明。

因此，如果得到的回归模型参数$|\varphi_1| < 1$, 是合理的，$|\varphi_1| = 1$或者接近于1，需要将原序列进行差分，$|\varphi_1| > 1$，需要反过来向前迭代。

※ Lag operator：$LX_t = X_{t-1}$，$L^pX_t = X_{t-p}$ 在计量经济学软件中，日线数据谨慎使用lag operator(周末问题)

※符合稳理想条件的ARMA模型的参数是一致估计量（证明从略）



例：有一个单位根位于单位圆上，模型是不稳定的

Identification

Box–Jenkins 建模思想：四个步骤：Identification，Estimation，Validation，Prevision

※ Identification（不是结构模型的识别问题）：

1）平稳性（stationarity）：

法一：自相关函数图（correlogram）：φ_1接近于1，自相关函数下降的非常缓慢（迭代使得φ_2接近于φ_1^2）

显然如果$|\varphi_1| < 1$,自相关函数应该下降得非常快（高阶自相关函数接近于φ_1 的高次方）

法二：单位根检验

2）季节性（seasonality）：某些时间序列呈现出非常显著的季节波动特征，如美国月度消费数据受圣诞节的影响，中国受春节的影响，A股市场逆回购受月末结算的影响等等。

方法一：观察自相关函数图，是否存在周期性规律

方法二：观察频谱密度图，是否在某个特定的频率存在峰值

3）p：AR（p）的偏自相关函数（PCF）在p+1阶为0，因此可以观察偏自相关函数图，显著的PCF阶数即为p。

4）Q：MA（q)的自相关函数（ACF）在q+1阶为0，因此可以观察自相关函数图，显著的ACF阶数即为q。

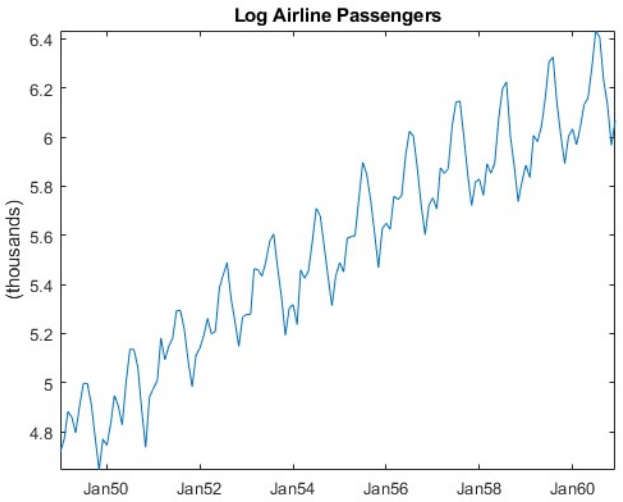
也可以用信息准则判断最优的p和q。

Identification:非理想情况的处理

※ 非平稳：弄清楚时间序列不平稳的原因是什么：
1）趋势项：去掉趋势项。线性趋势：移除时间趋势。非线性趋势：常用的是多项式趋势项
2）单位根：差分。差分几次才能得到平稳序列，就称原序列是几阶单整。 X_t~I（d）
大多数时间序列都是一阶单整，价格指数通常是二阶单整（LONG and Herrera, 2016, CER）,我从未见过三阶以上的单整经济时间序列。
使用ARIMA（p，d,q）,其中d表示经过几次差分后时间序列变为平稳。
3）非线性：nonlinear autoregressive–moving-average (NARMA) model
4） 断点：rupture是一种特殊的非线性模型，switch of regime models

※季节性：弄清楚时间序列呈现出季节性波动的原因是什么：
1）Seasonal ARIMA：
月度因素 $(1-L^{12})x_t = x_t - x_{t-12}$.
季度因素： $(1-L^{4})x_t = x_t - x_{t-12}$
节假日因素：X-N系列（中国节假日跟美国不同，NBS-X12）
中心移动平均回归（见我的硕士论文）

※异方差性：Autoregressive conditional heteroskedasticity (ARCH)
Granger（2003年诺奖） 的贡献：伪回归，ARCH, 因果检验，协整等等
Granger 的导师：清华大学刘大中教授



航空公司乘客数呈现出明显的季节性波动，建模时必须予以考虑

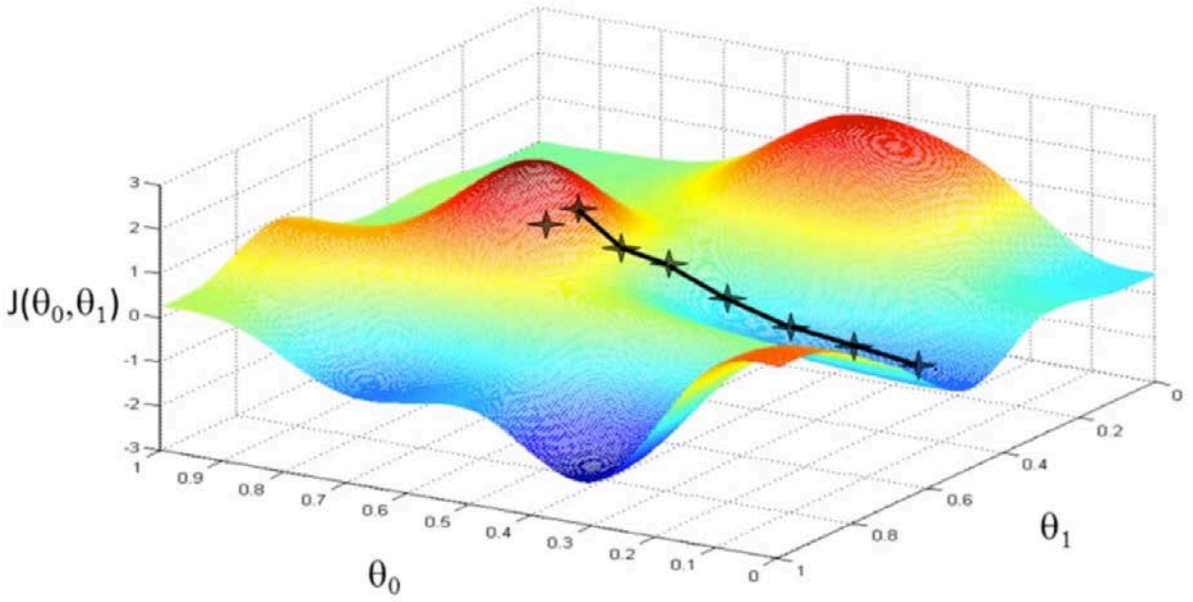
Estimation

※ 尽管现在计量经济学软件将估计过程都“黑箱化”，但是大多数研究人员只知使用，知其然而不知其所以然，容易出一些问题。
如果模型较为复杂，如p的阶数较高，黑箱化的估计过程可能导致我们得到的只是局部最优解。

※知道随机扰动项的分布，可以用MLE，不知道分布可用OLS/GLS等。
估计的过程一般是先对参数取一个初始值，再进行迭代，如stata中AR默认初始值是0.1,0.01,0.001··

这个过程如右图“登山”，选择斜率最高的方向进行迭代搜索，通常可以较快地登上峰顶，但由于我们通常受认知的限制（多峰函数），不知道全局最优解，很有可能只是得到局部的最优解。
结果的可复制危机

※解决办法：多试几个初始值，如果收敛结果很稳定，表明估计是可靠的。



在没有掌握全部信息的情况下，我们通常容易收敛于一个局部最优解而不自知。

Validation

※模型的诊断（diagnostics）：post-estimation tests 表明我们的模型时是正确的建模。

基本检验项跟OLS一样：残差值是否为白噪音等。

处理方法类似以及Identification中非理想情况的和estimation中参数不稳定的处理

3

ARMA模型在Stata中的实现

数据的处理

※使用第二讲的上证综指数据：

```
import delimited D:\new_zxzq_newfast\T0002\export\SH#999999.txt, varnames(2) encoding(GBK)
gen date=date(日期, "YMD" )
format date %td
tsset date, daily /* 时间序列的建模需要先定义时间变量*/
generate 涨幅 = 收盘/收盘[_n-1]-1   /* 日线数据谨慎使用lag operator*/
```

※ Box-Cox 变换：第三讲的描述性统计中我们得知，涨幅具有异方差性，此时需要使用Box-Cox 变换或者使用ARCH model (Autoregressive conditional heteroskedasticity).

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

可以简单地理解为取对数（如果对数存在）

$$x'_\lambda = \frac{e^{\lambda \log(x)} - 1}{\lambda} \approx \frac{\left(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \cdots\right) - 1}{\lambda} \to \log(x)$$

因此可以定义涨幅2：

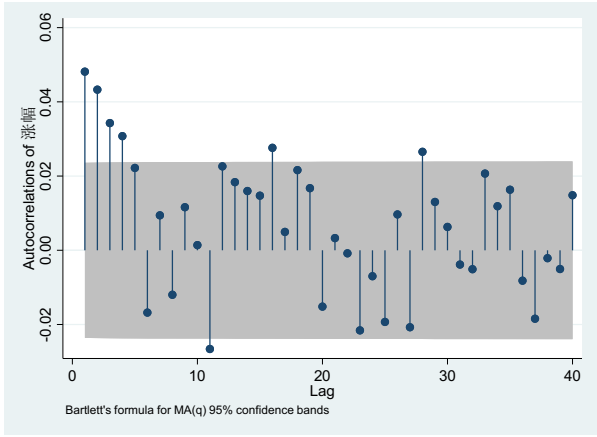
```
generate log收盘= log(收盘)
generate 涨幅2 = log收盘-log收盘[_n-1]
```

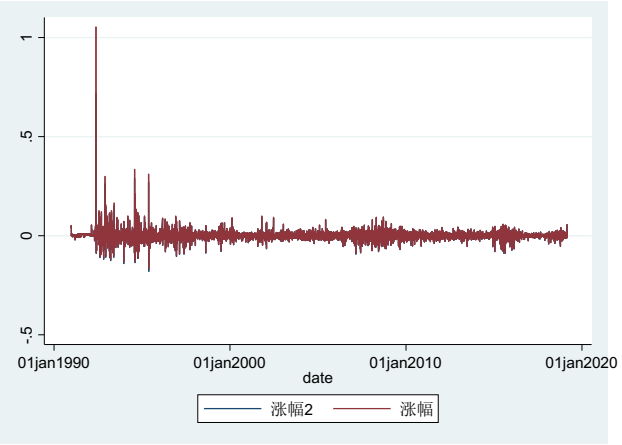
Identification: p & q

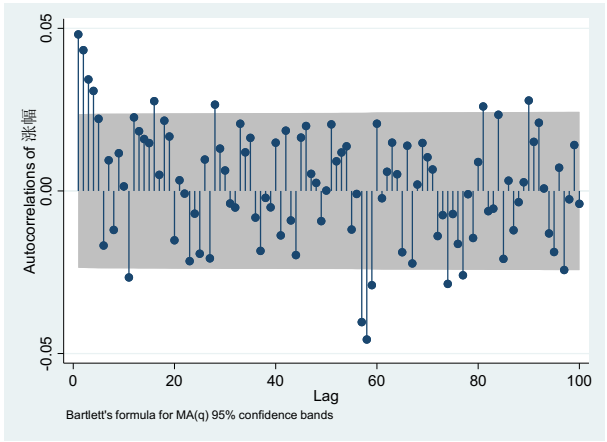
※使从图像上看，两种方式计算的涨幅基本上没有太大差别，
可以先忽略可能的异方差性，尝试直接对涨幅建模

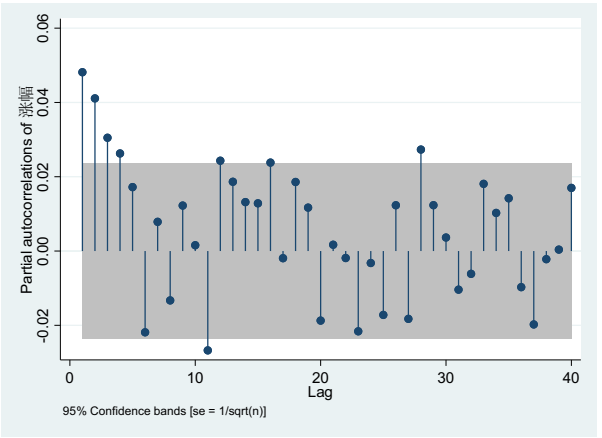
※统计性描述：见第三讲

※绘制correlogram识别p、q
ac 涨幅, lags(100)
pac 涨幅
偏自相关函数PAC 在lag=4时不为0，
在lag=5时为0，可能可以使用AR(4)
备选：AR(12), AR(16), AR(28)
自相关函数AC行为非常复杂，最大能
计算的lag只有100，可能的模型有：
MA(4）、MA(11)、MA(27),MA(60)等等。









这表明：
1）股市可能是一个长期记忆模型，此时应该使用长期记忆模型ARFIMA
2）从较长的correlogram来看，经过固定天数AC变得显著，这表明可能存在周期效应或者季节性因素
※越简单的模型越有效，复杂模型需要估计的参数过多，增加了估计难度和降低了统计可靠性

Identification: 平稳性

※使从图像上看，涨幅在0附近（略微为正），似乎没有趋势项（不是TS），均值平稳而方差不平稳

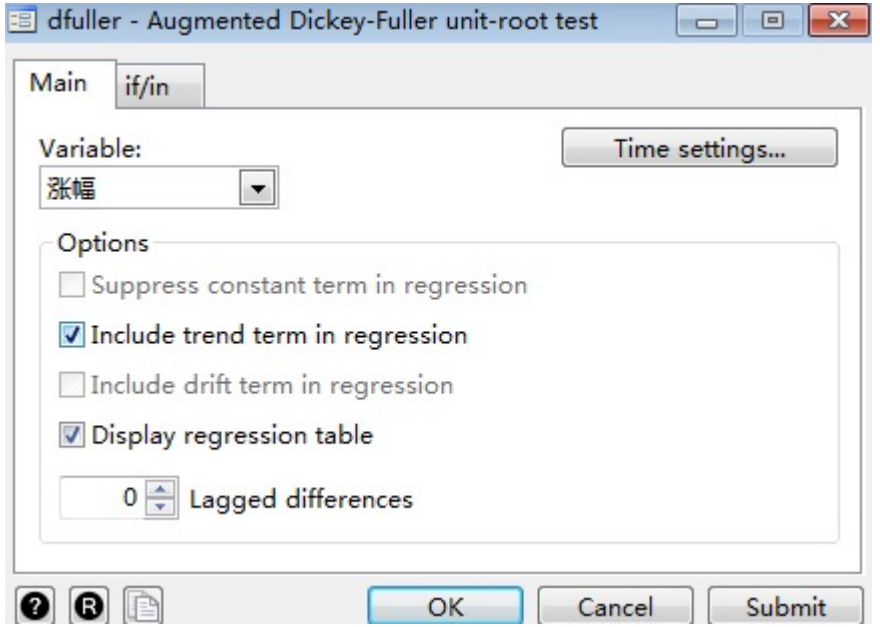
※单位根检验，以ADF检验为例
即使我们已经有了预判，涨幅的模型可能是模型2，仍然需要从模型3开始检验。

※ Lag的选择：根据样本容量使用固定截面参数：6896开四次=9.1
但由于我们定义了date作为时间变量，节假日不开市，没有数据，系统会默认为缺失值，因此lag不可能超过3.
同类问题：(note: time series has 1440 gaps)，以及lag operator

※解决办法：重新定义一个新的时间变量（注意不要选daily）

```
generate obs = 1
generate trend = sum(obs)
tsset trend
dfuller 涨幅, trend regress lags(9)
dfuller 涨幅, drift regress lags(9)
```

P-value很小，我们拒绝原假设，涨幅是平稳的。



_trend	-2.94e-07	1.51e-07	-1.94	0.053	-5.91e-
_cons	.0016487	.000605	2.73	0.006	.00046

_cons	.0006313	.0003011	2.10	0.036	.0000

趋势项不显著，常数项显著因此我们应当使用模型2

	Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
		Z(t) has t-distribution		
Z(t)	-25.306	-2.327	-1.645	-1.282

p-value for Z(t) = 0.0000

Estimation：AR(4)

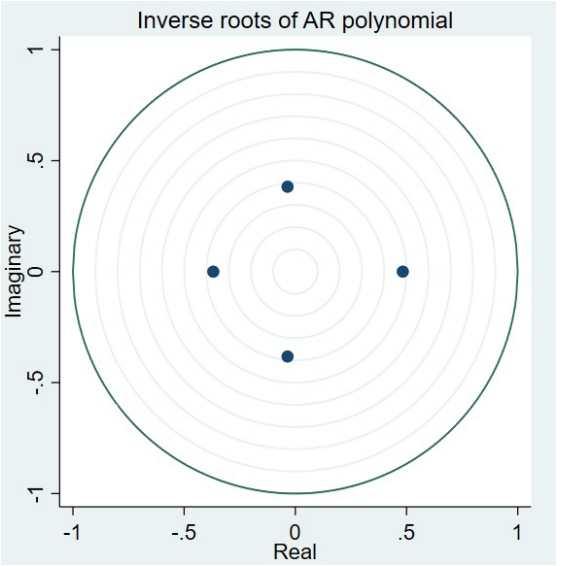
※我们分别估计以下模型: AR(4), MA（4）， 以及它们的组合ARMA(4，4)等三个模型

AR(4): arima 涨幅, arima(4,0,0)
※估计过程：在默认设置下，经过15次迭代得到收敛的结果（改变默认初始值可以测试收敛的稳定性）
注意: ARMA使用MLE迭代，没有R2
(可以手动计算一个“伪”R2，R2=ESS/TS
此时可以通过信息准则来判断拟合程度
estat ic

Sample: 2 - 6895 Number of obs = 6894
Wald chi2(4) = 267.68
Log likelihood = 15676.92 Prob > chi2 = 0.0000

涨幅	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
涨幅						
_cons	.0007728	.0004195	1.84	0.065	-.0000494	.001595
ARMA						
ar						
L1.	.0441605	.0074108	5.96	0.000	.0296357	.0586854
L2.	.0387157	.0053914	7.18	0.000	.0281487	.0492827
L3.	.0293772	.0038761	7.58	0.000	.0217802	.0369743
L4.	.0263035	.0041411	6.35	0.000	.018187	.03442
/sigma	.0248988	.0000181	1375.89	0.000	.0248634	.0249343

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.



Wald (卡方)检验：模型具有整体显著性
/sigma 表示残差值的标准差
OPG表示Robust Regression (相比于White 和NW correction, OPG在时间序列回归中更稳定，Bollerslev, Engle, and Nelson (1994))

※模型的稳定性： estat aroots
特征根均在单位圆内，模型是稳定的

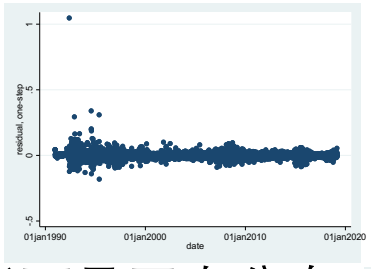
Akaike's information criterion and Bayesian information criterion

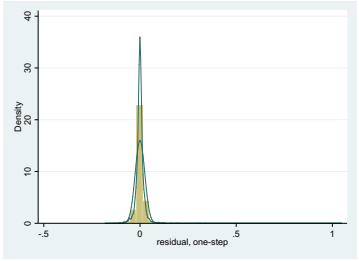
Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	6,894	.	15676.92	6	-31341.84	-31300.81

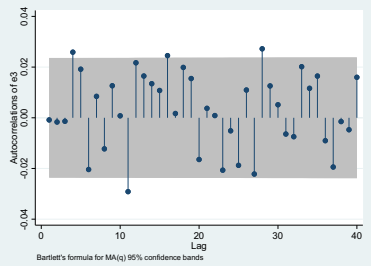
Note: N=Obs used in calculating BIC; see [R] BIC note.

Validation：AR(4)

我们使用了MLE，因此暗含“随机扰动项是高斯白噪音”的假设，需要检验残差值是否服从正态分布。







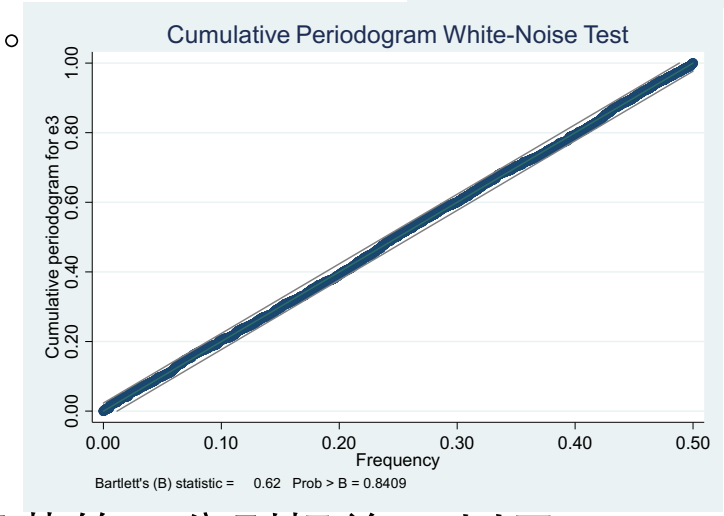
※储存残差值：predict e3, residual
绘制残差值的图像： twoway (scatter e3 date)
绘制频率直方图/概率密度：histogram e3, normal kdensity
从图像中我们可以查看出，残差值的峰度非常高，很有可能不是正态分布。

※ K-S检验等正态性检验证明了我们的猜想：残差值不是正态分布
summarize e3
ksmirnov e3 = normal((e3-r(mean))/r(sd))

```
. ksmirnov e3 = normal((e3-r(mean))/r(sd))

One-sample Kolmogorov-Smirnov test against
          normal((e3-r(mean))/r(sd))

Smaller group       D          P-value
-------------------------------------------
e3:                  0.1454     0.000
Cumulative:         -0.1370     0.000
Combined K-S:        0.1454     0.000
```



※退而求其次，检验是否为白噪音：
绘制残差值的correlogram: ac e3
残差值呈现平稳时间序列的特征，但似乎存在周期性（lag=4,11,16,28是显著的，分别相差7,5以及7+5=12，似乎对应一周7天、一周5个交易日和二者混的的周期）
corrgram e3
白噪音检验： wntestb e3 /* Bartlett's periodogram-based test for white noise*/

```
Portmanteau test for white noise
-------------------------------------------
Portmanteau (Q) statistic =      8.0490
Prob > chi2(9)            =      0.5292
```

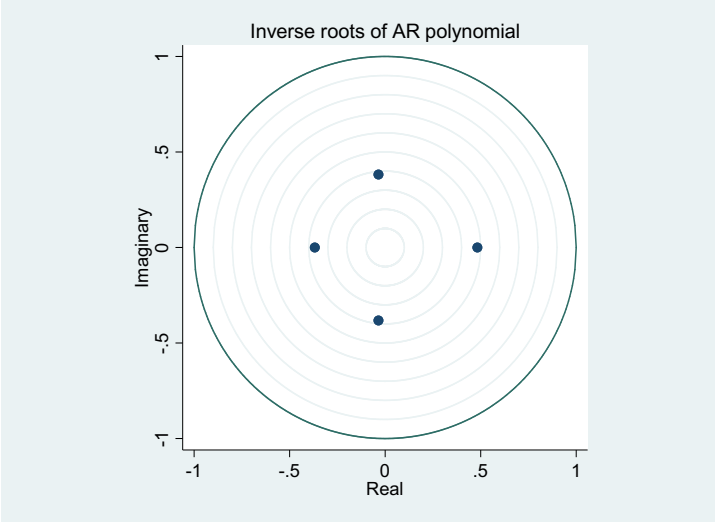
检验结果表明，尽管存在一定的异方差性和自相关性（周期），但残差值通过了白噪音检验（其他白噪音检验：wntestq e3, lags(9)，直到lag=9,仍然是白噪音）。
※结论：模型没有成功地Validation（残差值不是正态分布）

Re-Estimation：AR(4)

AR(3)的残差值通过了白噪音检验但没有通过正态性检验，但是我们使用了MLE，模型的可靠性值得审视，这种情况称为mis-specified

※Conditional MLE和quasi-maximum likelihood (QML)：可以证明（linear exponential family证明从略），在某些情况下，即使是mis-specified模型， Conditional MLE和QML仍然是一致估计量。因此我们就可以将估计方法改为Conditional MLE：
arima 涨幅, arima(3,0,0) condition

通过与Unconditional MLE比较，我们发现回归结果相当稳定，这表明尽管是一个mis-specified model，结果在一定程度上是可以接受的。



Sample: 2 - 6895 — Number of obs = 6,894
Distribution: Gaussian — Wald chi2(4) = 265.84
Log likelihood = 15676.87 — Prob > chi2 = 0.0000

涨幅	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
涨幅						
_cons	.0007731	.0004193	1.84	0.065	-.0000487	.0015949
ARMA						
ar						
L1.	.0441214	.0074105	5.95	0.000	.0295971	.0586456
L2.	.0386342	.0053907	7.17	0.000	.0280686	.0491999
L3.	.0292639	.0038785	7.55	0.000	.0216621	.0368657
L4.	.0261691	.0041443	6.31	0.000	.0180464	.0342918
/SIGMA2	.00062	9.01e-07	687.89	0.000	.0006182	.0006217

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	6,894	.	15676.87	6	-31341.75	-31300.72

Note: N=Obs used in calculating BIC; see [R] BIC note.

Re-Validation：AR(4)

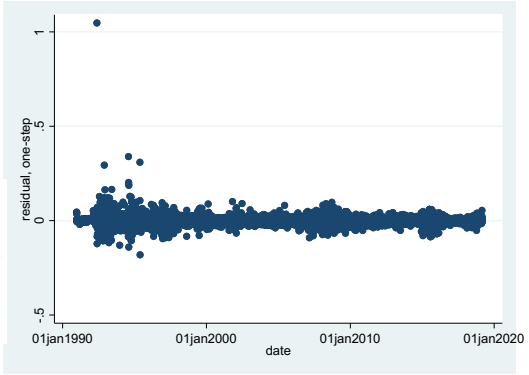
※储存新模型的残差值：predict e3new, residual
新模型的残差值仍然是可以通过白噪音检验，但无法通过正态性检验

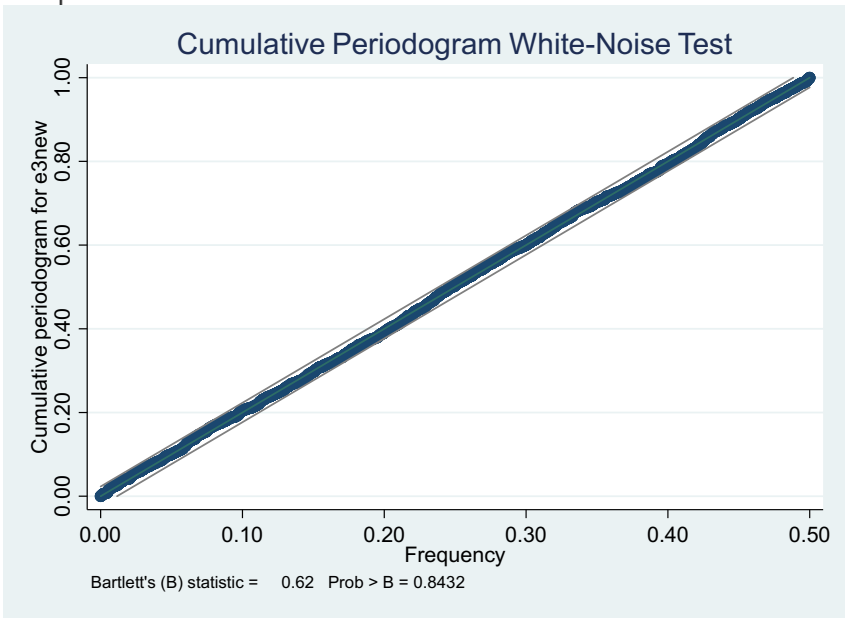
LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	-0.0005	-0.0005	.00146	0.9695		
2	0.0000	0.0000	.00148	0.9993		
3	-0.0002	-0.0002	.00188	1.0000		
4	0.0002	0.0002	.00214	1.0000		
5	0.0178	0.0178	2.1888	0.8225		
6	-0.0214	-0.0214	5.3507	0.4997		
7	0.0084	0.0083	5.8334	0.5593		
8	-0.0137	-0.0137	7.1255	0.5232		
9	0.0116	0.0115	8.0482	0.5293		

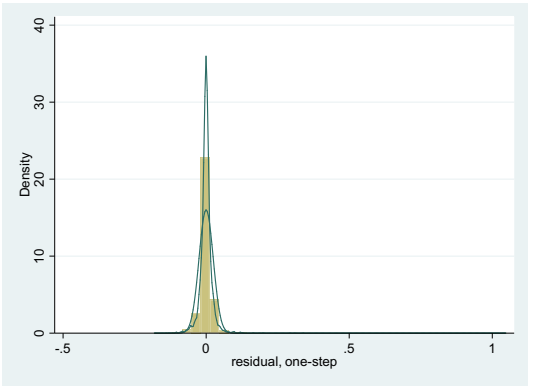
Smaller group	D	P-value
e3new:	0.1454	0.000
Cumulative:	-0.1370	0.000
Combined K-S:	0.1454	0.000

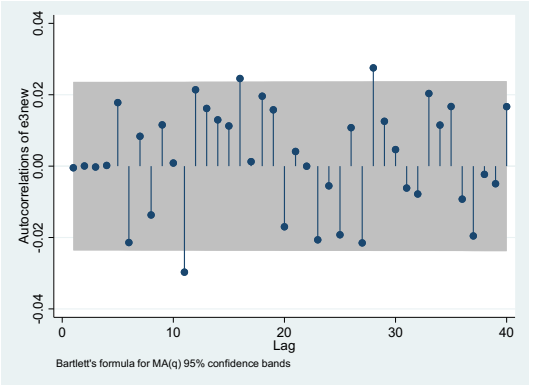
Portmanteau test for white noise

Portmanteau (Q) statistic =		8.0482
Prob > chi2(9)	=	0.5293









※ LM检验表明：AC和PAC一直到lag=10都不存在自相关性：corrgram e3new
注释：某些异方差性检验不适用于自回归模型（如果要用，需要先用OLS对滞后阶进行回归，在进行检验）
如：estat bgodfrey
estat archlm

※结论：尽管模型存在一定缺陷（残差值的周期性、异方差性（OPG）和非正态性（CMLE）），在一定程度上可以接受。

Prevision：AR(3)

※横截面数据无法预测未来，除非有样本外的数据，但自回归模型可以，这是自回归模型的优势

※样本内预测：根据估计的模型对已有的样本进行预测（本质上使用了未来的数据）

记录样本内预测值：predict predicted, xb

判断预测涨跌是否正确：

gen correct=1 if 涨幅*predicted>0

补全缺失值：

replace correct=0 if missing(correct)

乱猜有50%的正确性，因此我们检验正确性平均值是否显著大于0.5

结论：样本内预测效果非常好，但是没有太大意义。

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
correct	6,896	.5269722	.0060127	.4993082	.5151854	.5387589

mean = mean(correct) t = 4.4859

Ho: mean = 0.5 degrees of freedom = 6895

Ha: mean < 0.5 Pr(T < t) = 1.0000

Ha: mean != 0.5 Pr(|T| > |t|) = 0.0000

Ha: mean > 0.5 Pr(T > t) = 0.0000

	日期	predicted	预测	previsiondy
6895	2019/03/04	.0012415	.0012415	.0012415
6896	数据来源:通达信	.0018399	.0018399	.0013998
6897		.	.0015946	.0011898
6898		.	.0016079	.0012812
6899		.	.0011461	.0008422

※样本外预测：滞后阶数为3，最多可以预测未来3天

第一步：将样本外推3天: tsappend,add(3)

第二步：静态预测predict predicted, xb或predict 预测, y

或者动态预测：predict previsiondy, dynamic(6895) y

样本外日期	样本内预测	静态预测	动态预测	实际
3月5日	涨0.18%	涨0.18%	涨0.13%	0.88% 正确
3月6日		涨0.15%	涨0.11%	1.57% 正确
3月7日		涨0.16%	涨0.12%	0.14% 正确

样本外预测，静态预测one-step和动态预测等价,表中最后一列预测不等是因为包含了一期的样本内预测

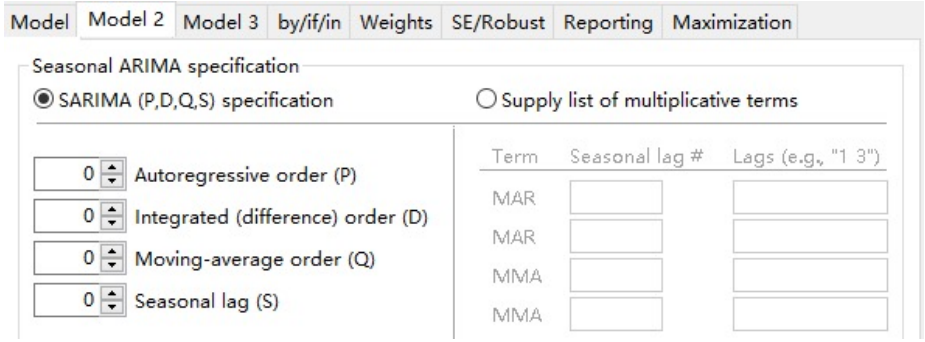
模型的改进：季节性因素和节假日效应

※周期效应和节假日效应:
A股市场是T+1交易制度，而政策存在不稳定性，例如央行喜欢在星期天晚上突然加息，证监会会在周五晚上发行新股等。因此部分持股者会为了规避周末消息的风险，会在周四或者周五清仓，而在周一重新开仓，造成周四周五经常性下跌而周一经常性上涨。（国外市场的无效性与此相反，国外市场机构投资者较多，周五需要买入证券平掉融券的空头头寸，周一重新卖空开仓）。此外，春节前因流动性紧张等因素，两会等，也会存在较强的季节性因素。

※识别某一天是星期几： gen week=dow(date)
可以用logit model (第七讲介绍)
也可以用SARIMA

※使用简单的统计性描述验证该直觉：
mean 涨幅 in 6000/6896, over(week)

日期	涨幅	trend	week
1990/12/19	.	1	3
1990/12/20	.0441088	2	4
1990/12/21	.0454066	3	5
1990/12/24	.0496656	4	1
1990/12/25	.0497599	5	2
1990/12/26	.0417463	6	3
1990/12/27	.0000798	7	4




```
Mean estimation                   Number of obs   =        896

                1: week = 1
                2: week = 2
                3: week = 3
                4: week = 4
                5: week = 5
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
涨幅				
1	.0000284	.0013462	-.0026137	.0026706
2	.0009056	.0010772	-.0012086	.0030198
3	-.0011084	.0009302	-.0029339	.0007171
4	-.0011204	.0010683	-.003217	.0009762
5	.0002379	.0010203	-.0017646	.0022403

作业

※按照Box–Jenkins估计MA(4), ARMA(4,4),
越简单的模型越有效，不提倡使用ARMA(28,60)一类过于高阶的模型

※加分作业：预测能力更好的ARMA(p,q) 或者直接对 “收盘价” 建模（Box-Cox变换）
ARIMA(p,d,q)

4

参考文献


站在巨人的肩膀上
"If I have seen further it is by standing on the shoulders of Giants. “
by Isaac Newton in 1675

参考文献&数学基础知识&样本数据

见网络学堂附件

下节课见

马克思主义学院
龙治铭

清華大學
Tsinghua University