多元时间序列分析

08

本章内容

ARIMAX模型

- 1976年，Box和Jenkins采用带输入变量的ARIMA模型为平稳多元序列建模。他们建立的这个模型简记为ARIMAX模型。因为该模型引入了自回归系数多项式和移动平均多项式结构，所以也称为传递函数模型。
- ARIMAX模型结构

$$y_t = \beta_0 + \sum_{i=1}^{k} \frac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} x_{it} + \frac{\Theta(B)}{\Phi(B)} a_t$$

式中，$\Phi_i(B)$ 第 i 个自变量 $\{x_{it}\}$ 的 p_i 阶自回归系数多项式

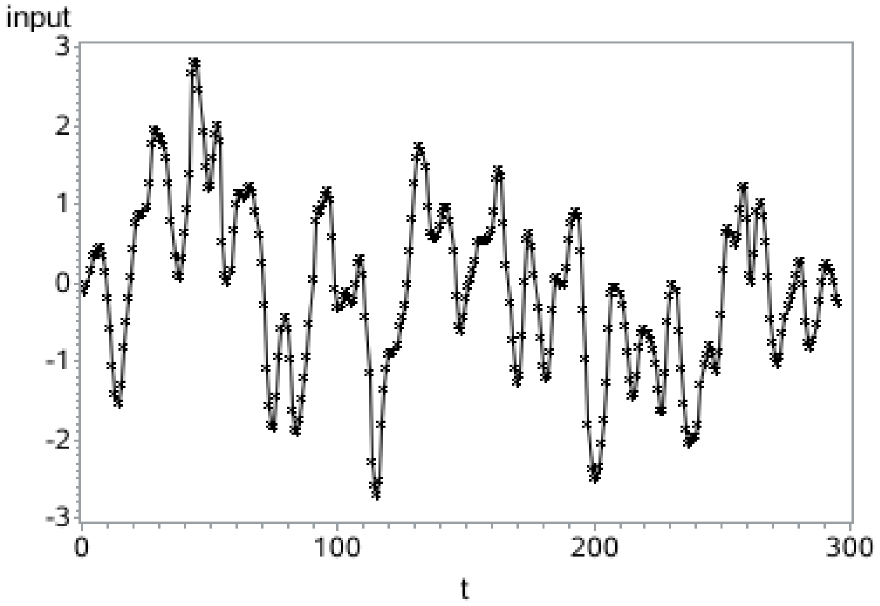
$\Theta_i(B)$ 第 i 个自变量 $\{x_{it}\}$ 的 q_i 阶移动平均系数多项式

$\Phi(B)$ 为残差序列自回归系数多项式

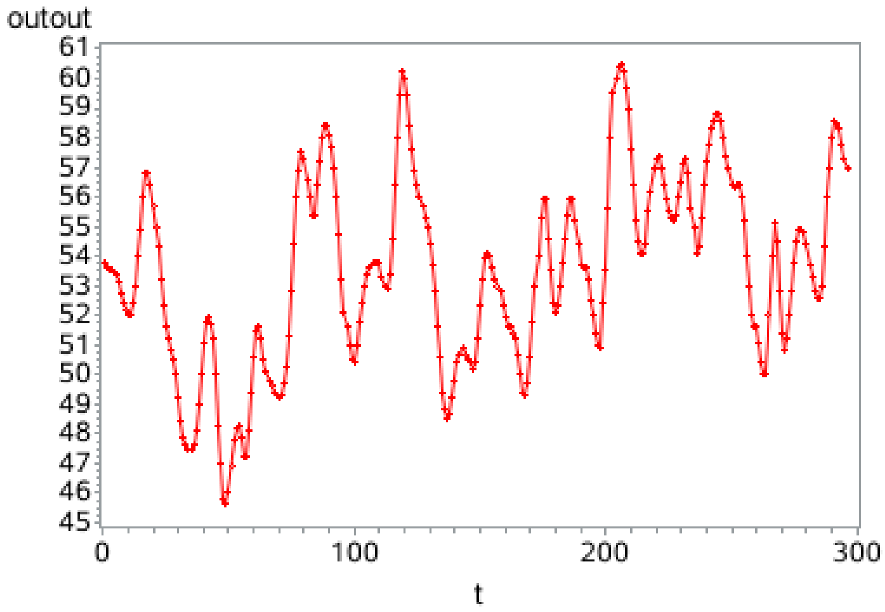
$\Theta(B)$ 为残差序列移动平均序数多项式

例8-1

- 在天然气炉中，输入的是天然气，输出的是 CO2，CO2的输出浓度与天然气的输入速率有关。现在以中心化后的天然气输入速率为输入序列，建立CO2的输出百分浓度模型。



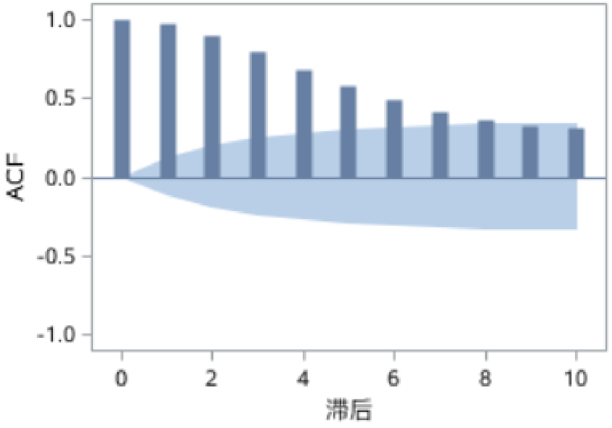
天然气输入速率时序图

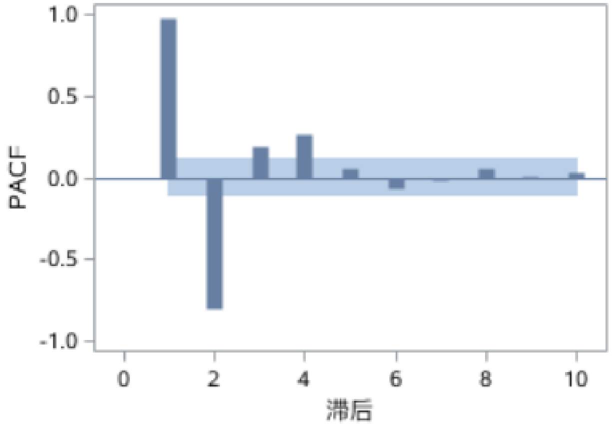


CO2输出浓度时序图

对输出序列建立单变量ARIMA模型

- 如果不考虑输入序列和输出序列之间的相关性，将它们作为两个独立的时间序列看待。对输出序列建立单变量ARIMA模型





- 自相关图呈现拖尾属性，偏自相关图4阶截尾。所以对输出序列拟合AR(4)模型。
- 根据系数显著性检验结果，最后确定的拟合模型为AR(1,2,4)疏系数模型。

$$y_t = 53.90176 + \frac{a_t}{1 - 2.10703B + 1.34005B^2 - 0.21274B^4}$$

- 输出序列模型的 AIC = 1 9 6，SBC = 2 1 1

互相关系数

- 考虑到输入天然气速率与输出CO2的浓度之间有逻辑上的因果关系，将输入天然气速率作为输入变量考虑进输出序列的模型中。通过互相关函数或互相关系数的特征，考察回归模型的结构。
- 延迟k阶互相关函数（cross covariance）的定义

$$Cov_k = Cov(y_t, x_{t-k}) = E\left[\left(y_t - E(y_t)\right)\left(x_{t-k} - E(x_{t-k})\right)\right]$$

- 延迟k阶互相关系数（cross correlation coefficient）的定义

$$C\rho_k = \frac{Cov(y_t, x_{t-k})}{\sqrt{Var(y_t)}\sqrt{Var(x_{t-k})}}$$

  - 如果k>0，计算的是y序列滞后于x序列k期的相关系数。
  - 如果k<0，计算的是x序列滞后于y序列k期的相关系数。

互相关系数的分布特征

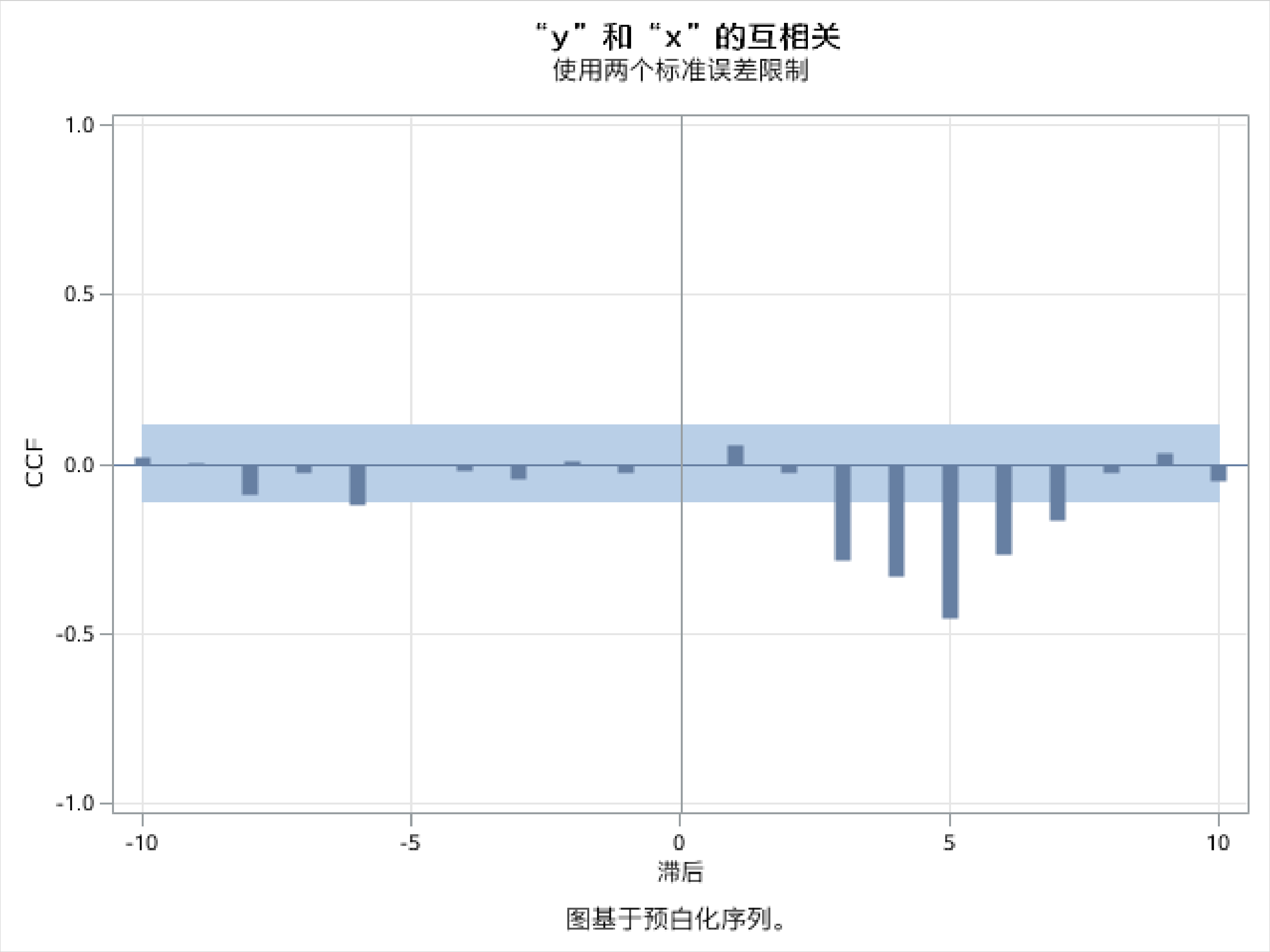
- 和自相关系数、偏自相关系数一样，根据Bartlett定理，互相关系数近似服从零均值正态分布

$$C\rho_k \dot{\sim} N\left(0, \frac{1}{n-|k|}\right)$$

- 超过2倍标准差的互相关系数可以认为显著非零，即相应序列和自变量序列之间具有显著相关性

$$C\rho_k > \frac{2}{\sqrt{n-|k|}}$$

互相关系数图



从左图中可以看出：

- 延迟3阶到延迟7阶，互相关系数都显著大于2倍标准差。
- 这说明输出序列和输入序列之间有3期滞后效应
- 5期（延迟3-7阶）显著互相关

构建ARIMAX模型

- 传统线性回归模型
  - 互相关系数图显示输出序列和输入序列之间有3期滞后效应，5期显著互相关，如果构建传统线性回归模型，模型结构可以表达为

$$y_t = \beta_0 + \beta_1 x_{t-3} + \beta_2 x_{t-4} + \beta_3 x_{t-5} + \beta_4 x_{t-6} + \beta_5 x_{t-7} + \varepsilon_t$$

  - 该回归模型的问题：（1）自变量太多；（2）自变量之间有显著的相关性，容易出现多元共线性问题
- ARIMAX模型
  - Box和Jenkins建议当自变量延迟阶数比较多时，可以考虑采用传递函数模型结构，以减少待估参数的个数。本例回归模型不妨采用ARMA(1,2)结构替代

$$y_t = \beta_0 + \frac{\theta_0 - \theta_1 B - \theta_2 B^2}{1 - \phi_1 B} B^3 x_t + \varepsilon_t$$

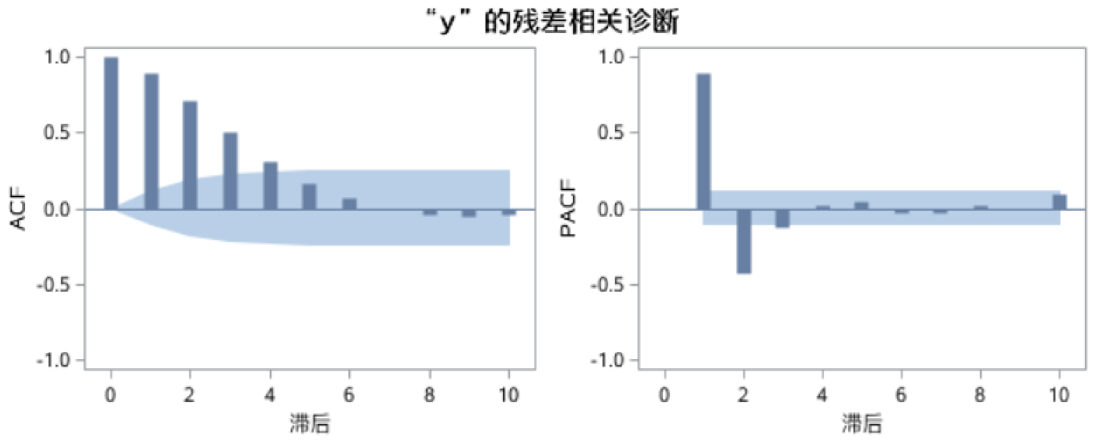
ARIMAX模型拟合

- 第一步：建立响应变量与自变量的传递函数模型

$$y_t = 53.322 + \frac{-0.565 - 0.426B - 0.299B^2}{1 - 0.601B} B^3 x_t + \varepsilon_t$$

- 第二步：建立残差序列的ARMA模型

残差序列自相关拖尾，偏自相关2阶截尾，所以对残差序列拟合AR(2)模型



- 第三步：得到ARIMAX模型

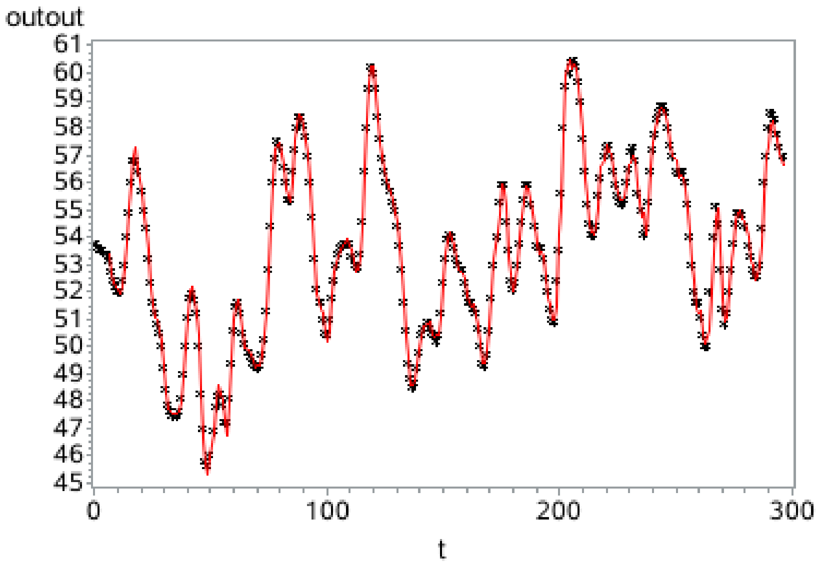
$$y_t = 53.322 + \frac{-0.565 - 0.426B - 0.299B^2}{1 - 0.601B} B^3 x_t + \frac{1}{1 - 1.53B + 0.64B^2} a_t$$

模型比较

- 输出序列ARIMAX模型的 AIC = 8， SBC = 3 4。显然，这个ARIMAX模型比不考虑输入序列的单纯的AR(1,2,4)疏系数模型优化多了。

拟合模型	AIC	BIC
AR(1,2,4)	196	211
ARIMAX	8	34

- ARIMAX模型拟合效果图



本章内容

干预分析

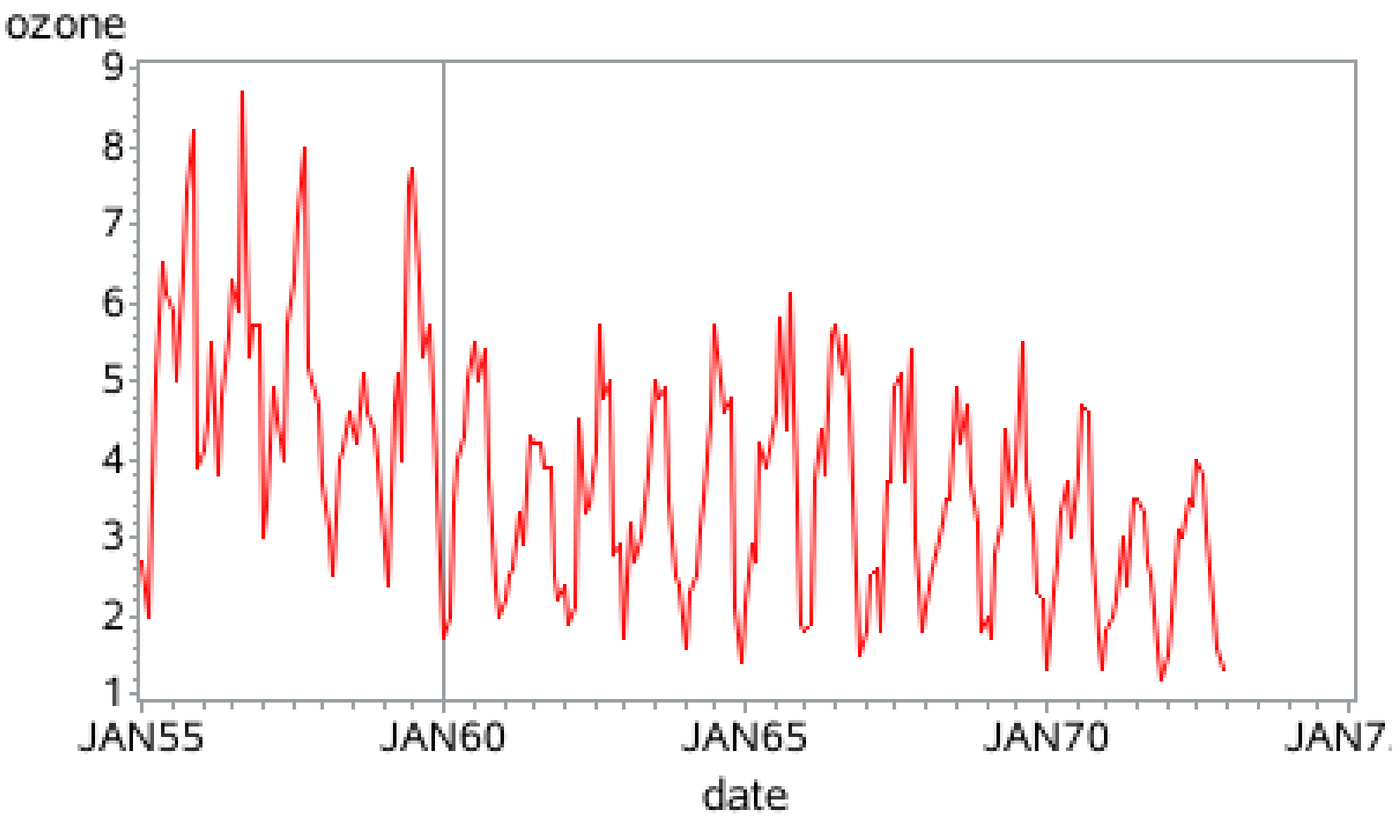
- 干预分析的定义
  - 时间序列常常受到某些外部事件的影响，诸如：假期，罢工，促销，或者政策的改变等。我们称这些外部事件为“干预”。评估外部事件对序列产生的影响的分析，称为干预分析（intervention analysis）。
- 干预分析的产生背景
  - 最早的干预分析是1975年Box和刁锦寰（Tiao）对加州63号法令是否有效抑制了加州空气污染问题的研究。他们首次将干预事件以虚拟变量的方式进行标注，然后把虚拟变量作为输入变量引入序列分析，构建ARIMAX模型。
- 干预分析的实质
  - 所谓干预模型实际上是带虚拟变量回归的ARIMAX模型，所以干预模型实质上就是ARIMAX模型的一种特例。

例8-2

- 二战之后加利福尼亚州经济高速发展，蓬勃发展的经济也带来了严重的空气污染。由于工厂排放的废气、汽车排放的尾气、家庭使用的燃气排放物中都含有大量的氮氧化物和活性碳氢化物。
- 废气在阳光的作用下产生化学反应，这些化学反应物形成了严重的雾霾，造成大量人群流眼泪，咳嗽、肺部受损等身体伤害。经测量光化学污染程度的标志是臭氧的含量。
- 为了解决污染问题，加州政府在1959年颁布了63号法令。该法令要求从1960年1月起，在当地销售的汽油中减少碳氢化物的容许比。
- Box和Tiao在1975年，根据他们收集的1955年1月—1972年12月的月度臭氧浓度序列，分析63号法令的颁布执行，对控制加州的光学污染有没有起到作用？如果起了作用，起了多大的作用。

干预分析步骤一

- 考察序列的时序图和互相关图，研究干预变量对序列的干预机制。



臭氧序列时序图

干预变量设定

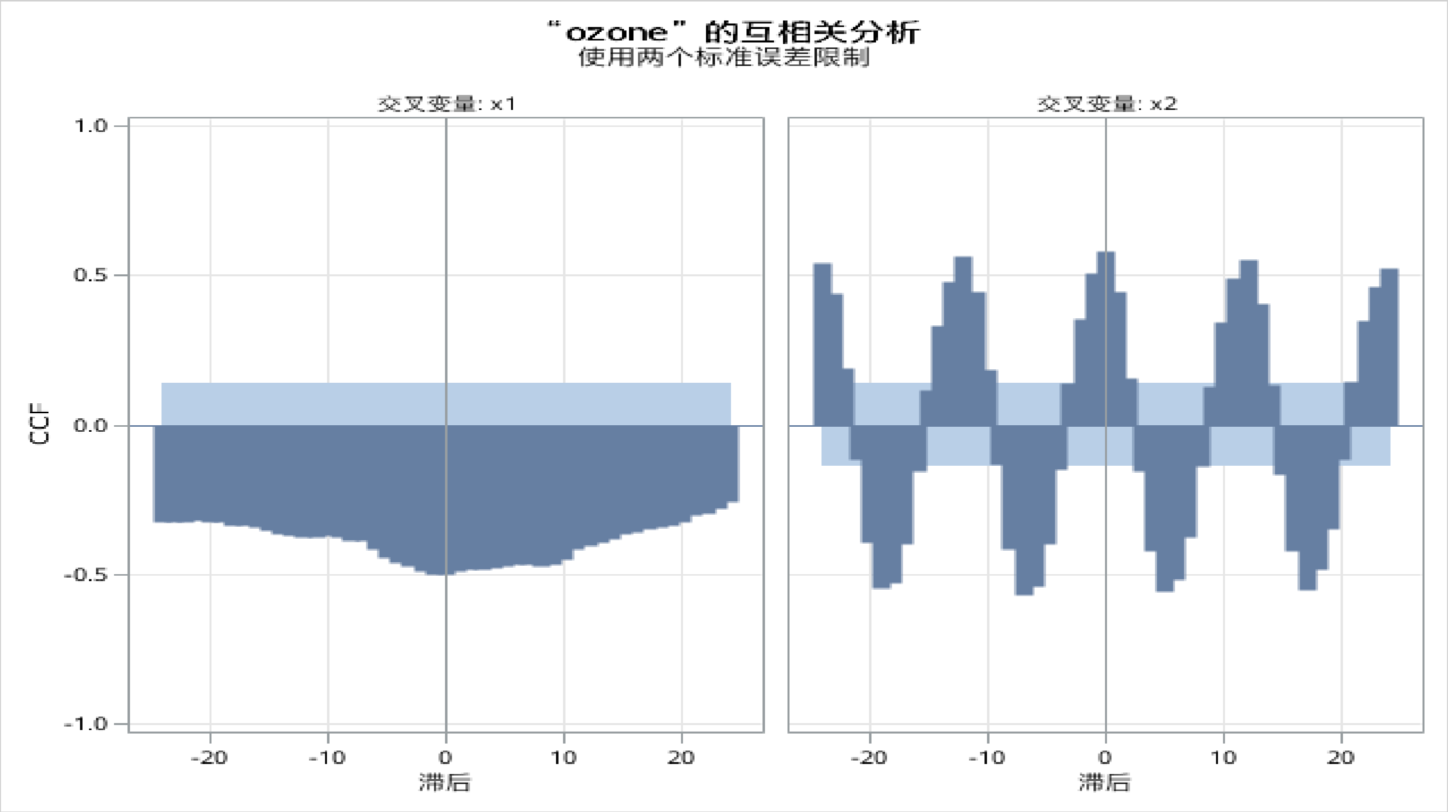
- 在这个研究中，干预变量是63号法令的颁布和执行。这是一个定性变量，它没有数值，只有两个属性：（1）1960年之前没有执行，（2）1960年之后执行了。基于这种情况，Box和Tiao把干预变量以虚拟变量的方式进行处理

$$x_{1t}=\begin{cases}0 & ,t<1960\text{年}1\text{月}\\ 1 & ,t\geq 1960\text{年}1\text{月}\end{cases}$$

- 在研究中，Box和Tiao发现，除了政策法规这个干预变量之外，影响臭氧浓度的还有一个定性变量，那就是季节。因为冬季有供暖需求，废气排放比夏天多。其次冬季的温度低，污染物的扩散慢，所以冬季和夏季对臭氧浓度可能有不同的干预力度。所以他们又构造了两个虚拟变量，用以描述季节对臭氧序列的影响（这两个变量选其一即可）

$$x_{2t}=\begin{cases}1 & ,t\text{处于}6-10\text{月}\\ 0 & ,t\text{处于}11-5\text{月}\end{cases}\qquad x_{3t}=\begin{cases}1 & ,t\text{处于}11-5\text{月}\\ 0 & ,t\text{处于}6-10\text{月}\end{cases}$$

互相关图



“ozone”的互相关分析
使用两个标准误差限制
交叉变量: x1
交叉变量: x2
CCF
1.0
0.5
0.0
-0.5
-1.0
-20
-10
0
10
20
滞后

干预机制

- 时序图显示，序列有明显的季节效应。６３号法令执行之后（参照线前后），序列的周期波动特征没有明显改变，但是序列的波动水平比以前明显降低。所以季节效应和63号法令作为两个干预变量引入臭氧序列拟合。
- 互相关图显示，两个干预变量都是0阶滞后互相关系数最大。所以假定干预变量对序列的干预只是水平影响，且无延迟。确定干预模型结构如下

$$ozone_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \frac{\Theta(B)}{\Phi(B)} a_t$$

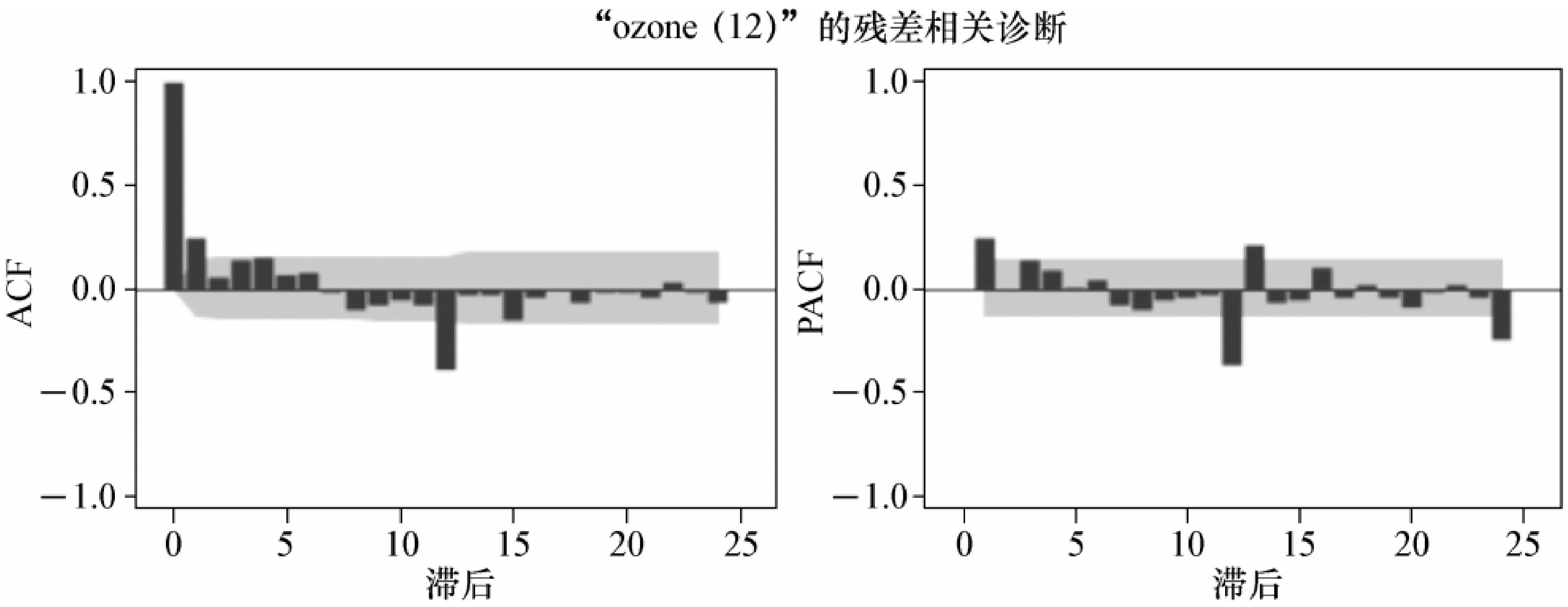
干预分析步骤二

- 对臭氧浓度序列进行１２步差分，实现差分平稳

类型	延迟阶数	τ统计量的值	$Pr<\tau$
类型一	0	−6.89	<0.000 1
	1	−6.25	<0.000 1
	2	−4.35	<0.000 1
类型二	0	−6.97	<0.000 1
	1	−6.41	<0.000 1
	2	−4.53	0.000 4
类型三	0	−6.99	<0.000 1
	1	−6.5	<0.000 1
	2	−4.67	0.001 3

干预分析步骤三

- 考察残差序列 $\{\varepsilon_t\}$ 的自相关图和偏自相关图，为残差序列指定模型结构为 $ARIMA(1,0,0)\times(0,1,1)_{12}$



- 确定干预模型结构为

$$\nabla_{12} ozone_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \frac{1-\theta_{12}B^{12}}{1-\phi_1 B} a_t$$

干预分析步骤四

- 参数估计

参数	估计	标准误差	t 值	近似 $Pr>\|t\|$	滞后	变量	位移
θ_{12}	0.721 24	0.061 42	11.74	<0.000 1	12	ozone	0
ϕ_1	0.314 34	0.064 87	4.85	<0.000 1	1	ozone	0
β_1	−1.224 64	0.251 84	−4.86	<0.000 1	0	x1	0
β_2	−0.067 52	0.040 54	−1.67	0.095 8	0	x2	0

- 干预模型拟合结果

$$ozone_t = ozone_{t-12} - 1.2246x_{1t} - 0.0675x_{2t} + \frac{1-0.7212B^{12}}{1-0.3143B}a_t$$

$$a_t \sim N(0, 0.6674)$$

干预分析步骤五

- 拟合模型显著性检验
  - 对干预模型残差序列进行白噪声检验，LB检验结果如下表所示。检验结果显示残差序列为白噪声序列，所以拟合模型显著成立

延迟阶数	纯随机性检验	
	LB 检验统计量的值	P 值
6	8.04	0.090 2
12	10.58	0.391 2
18	15.64	0.478 0
24	21.17	0.510 1
30	28.65	0.430 5

干预分析步骤六

- 干预效果解读

  - 根据 $\beta_1 = -1.225$ ，而且该系数t检验显著非零的特征，可以认为63号法令的颁布和实施有效降低了加州臭氧浓度。这说明这个法令的颁布和实施对治理加州的空气污染是显著有效的。又因为 $mean(ozone_t | t < 1960\text{年}) = 4.177$ ，即在1960年之前，臭氧序列的平均浓度等于4.177，而因为63号法令的执行，使得臭氧浓度平均降低了1.225，所以63号法令的执行，使得加州臭氧浓度比法令执行之前下降了30%左右。
  - 根据 $\beta_2 = -0.0675$ ，说明夏季比冬季的臭氧浓度低，但我们的季节性划分太粗糙，这个系数并不显著非零。
  - 消除政策因素和季节因素的干预影响，臭氧浓度序列自身的波动服从季节乘法模型。干预因素会影响臭氧序列的浓度水平，但不会改变臭氧序列的波动规律。

干预分析步骤七

- 序列预测：根据拟合的干预模型，事先确定未来各期干预变量的取值，还可以对序列进行短期预测。

预测时期	预测值	标准差	95%置信下限	95%置信上限
$T+1$	1.602 3	0.816 9	0.001 2	3.203 5
$T+2$	2.098 3	0.856 4	0.419 9	3.776 7
$T+3$	2.749 4	0.860 1	1.063 5	4.435 2
$T+4$	3.130 4	0.860 5	1.443 8	4.817
$T+5$	3.427 9	0.860 6	1.741 2	5.114 6
$T+6$	3.314 7	0.860 6	1.628	5.001 4
$T+7$	3.894 1	0.860 6	2.207 4	5.580 7
$T+8$	4.054 5	0.860 6	2.367 9	5.741 2
$T+9$	3.480 8	0.860 6	1.794 1	5.167 5
$T+10$	2.850 9	0.860 6	1.164 2	4.537 5
$T+11$	2.056 1	0.860 6	0.369 4	3.742 8
$T+12$	1.529 8	0.860 6	−0.156 9	3.216 4

干预机制的选择

- 干预模型是进行政策效果评估或分析特殊事件影响的有用模型。而干预模型的关键是将干预事件以虚拟变量的形式引入响应序列分析。干预事件根据作用机制可以分为三种类型：

阶梯干预	$x_t=\begin{cases}0\ ,\text{干预事件发生之前}(x<T)\\1\ ,\ \text{干预事件发生之后}(x\geq T)\end{cases}$
脉冲干预	$x_t=\begin{cases}1\ ,\ \text{干预事件发生时}(x=T)\\0\ ,\ \text{其他时刻}(x\neq T)\end{cases}$
其他干预	用阶梯干预和脉冲干预的转换函数或组合来生成

本章内容

多元时序回归分析

- 计量经济学家很早就开始构建多元时序的回归模型

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + + \beta_k x_{kt} + \varepsilon_t$$

- 20世纪60-70年代，Granger等统计学家，对多元时序回归模型提出了伪回归的概念。
- 他们提醒计量经济学家，在使用时间序列进行线性回归分析时，回归模型很容易通过方程显著性检验。很多时候不是因为这些序列之间真的具有因果关系，而是时间的相关性，造成非平稳序列之间的“伪”回归。

伪回归定义

- 对回归模型进行方程显著性检验

$$y_t = \beta_0 + \beta_1 x_{1t} + \varepsilon_t$$

- 假设条件

$$H_0 : \beta_1 = 0 \leftrightarrow H_1 : \beta_1 \neq 0$$

- 检验统计量

$$t = \frac{\beta_1}{\sigma_\beta}$$

- 伪回归：真实拒绝原假设的概率大于理论上的拒真概率 α

$$\Pr\{|t| \leq t_{\alpha/2}(n) \,|\, \text{非平稳序列}\} \geq \alpha$$

伪回归随机模拟试验

- 1974年，Granger和Newbold进行了非平稳序列伪回归的随机模拟试验。检验结果说明在非平稳的场合，参数显著性检验犯第一类错误的概率远远大于显著性水平，伪回归显著成立。这导致多元非平稳序列的分析埋有隐患。
- 试验设计思想：分别拟合两个随机游走序列：

$$(1)\ y_t = y_{t\text{-}1} + \omega_t \qquad (2)\ x_t = x_{t\text{-}1} + \upsilon_t$$

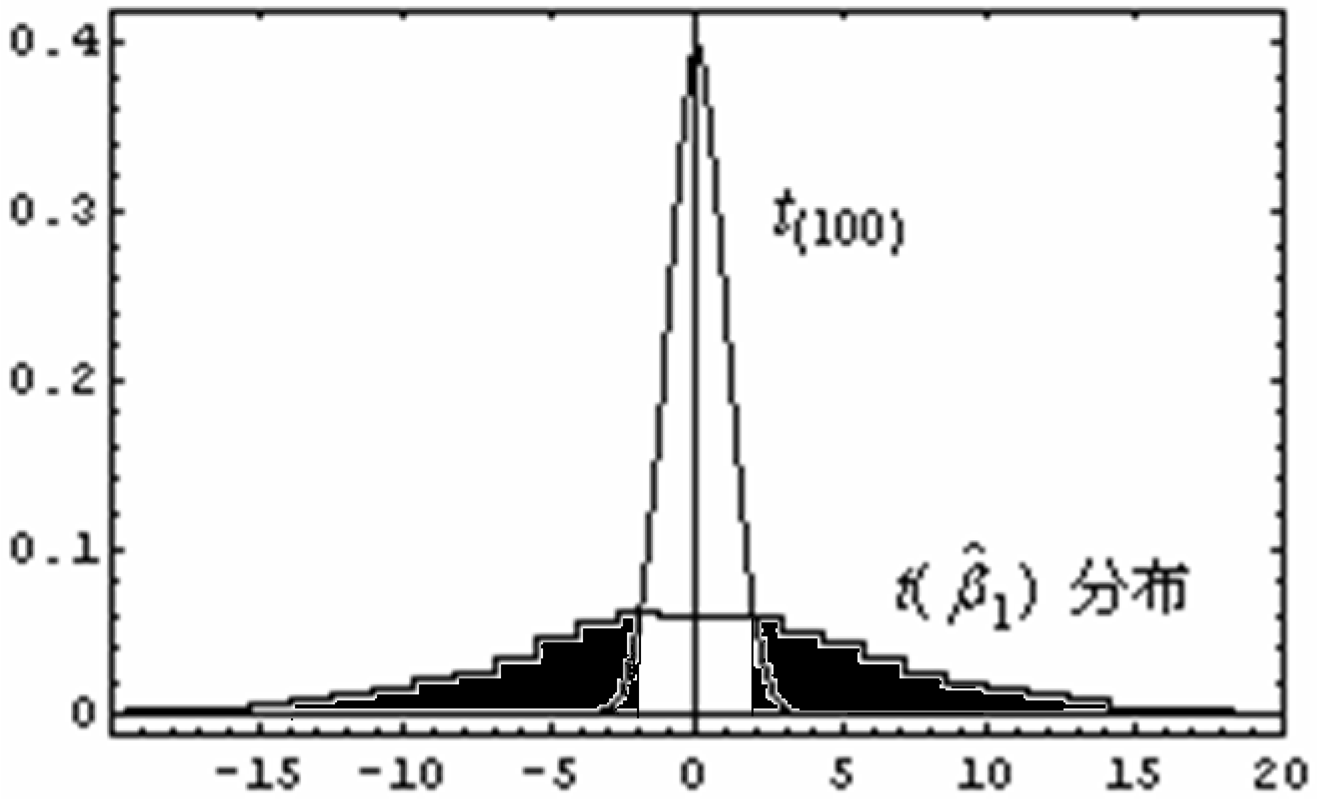
其中：$\omega_t \overset{iid}{\sim} N(0,\sigma_\omega^2), \upsilon_t \overset{iid}{\sim} N(0,\sigma_\upsilon^2)$，且$Cov(\omega_t,\upsilon_s)=0, \quad \forall t,s \in T$

构建回归模型：$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$，并进行参数显著性检验。

试验结果

- 由于这是两个独立的随机游走模型，所以理论上它们应该没有任何相关性，即模型检验应该显著支持 $\beta_1=0$ 的假设。如果模拟结果显示拒绝原假设的概率远远大于拒真概率 α ，即认为伪回归显著成立。
- 大量随机拟合的结果显示，每100次回归拟合中，平均有75次拒绝 $\beta_1=0$ 的假设，拒真概率高达75%。这说明在非平稳的场合，参数显著性检验犯拒真错误的概率远远大于 α ，伪回归显著成立。

伪回归产生原因



产生伪回归的原因是在非平稳场合，参数的 t 检验统计量不再服从 t 分布。统计量真实的抽样分布 $t(\hat{\beta}_1)$ 尾部肥，方差大，比 t 分布要扁平很多。如果继续使用t分布的临界值做方程显著性判断，则会导致很大的犯第一类错误的概率（如图阴影部分所示）。

本章内容

单整

- 单整的概念
  - 如果序列平稳，说明序列不存在单位根，这时称序列为零阶单整序列，简记为

$$x_t \sim I(0)$$

  - 假如原序列一阶差分后平稳，说明序列存在一个单位根，这时称序列为一阶单整序列，简记为

$$x_t \sim I(1)$$

  - 假如原序列至少需要进行d阶差分才能实现平稳，说明原序列存在d个单位根，这时称原序列为d阶单整序列，简记为

$$x_t \sim I(d)$$

单整的性质

- 若 $x_t \sim I(0)$ ，对任意非零实数a，b，有

$$a + bx_t \sim I(0)$$

- 若 $x_t \sim I(d)$ ，对任意非零实数a，b，有

$$a + bx_t \sim I(d)$$

- 若 $x_t \sim I(0)$ ，$y_t \sim I(0)$ 对任意非零实数a，b，有

$$z_t = ax_t + by_t \sim I(0)$$

- 若 $x_t \sim I(d)$ ，$y_t \sim I(c)$ 对任意非零实数a，b，有 $k \le \max[d, c]$

$$z_t = ax_t + by_t \sim I(k)$$

协整的概念

- 假定自变量序列为 $\{x_1\},\{x_2\}\cdots,\{x_k\}$，响应变量序列为 $\{y_t\}$，构造回归模型

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{it} + \varepsilon_t$$

  假如回归残差序列 $\{\varepsilon_t\}$ 平稳，我们称响应序列与自变量序列之间具有协整关系（Cointegration）。

- 变量之间如果具有协整关系，意味着它们具有长期稳定的均衡变化关系。

协整检验

- 假设条件
  - 原假设：多元非平稳序列之间不存在协整关系

$$H_0 : \varepsilon_t \sim I(k), k \geq 1$$

  - 备择假设：多元非平稳序列之间存在协整关系

$$H_1 : \varepsilon_t \sim I(0)$$

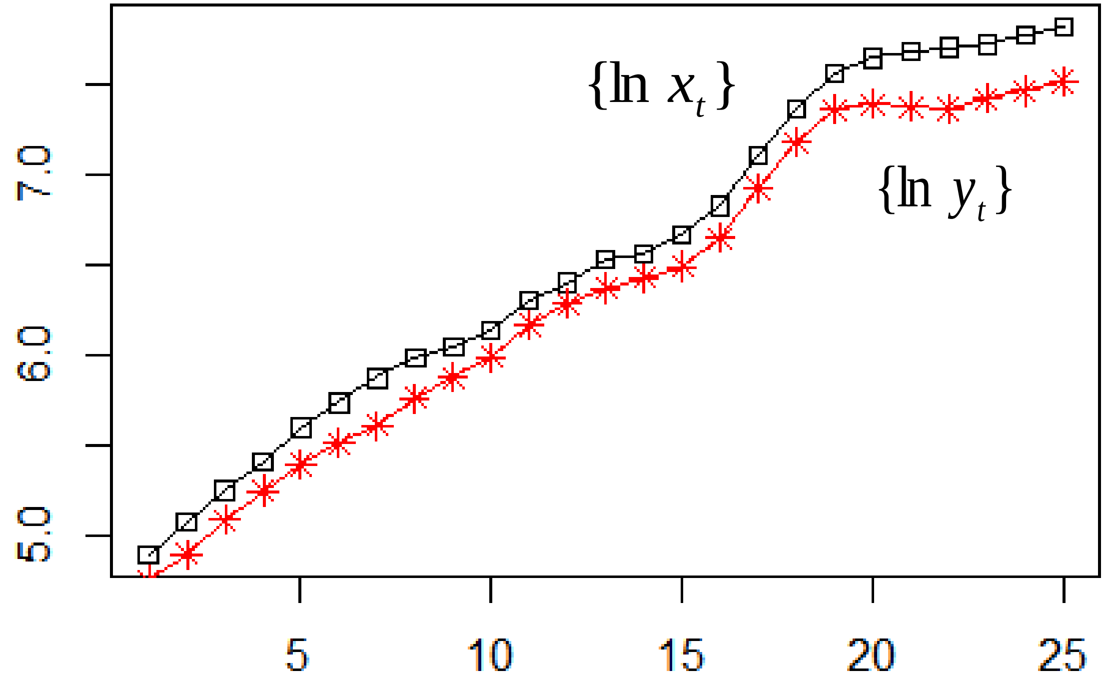
- 检验方法（EG两步法）

  第一步：建立响应序列与输入序列之间的回归模型

  第二步：对回归残差序列进行平稳性检验

例8-3

- 对1978-2002年中国农村居民家庭人均纯收入对数序列 $\{\ln x_t\}$ 和生活消费支出对数序列 $\{\ln y_t\}$ 进行协整检验



构造协整模型

- 拟合一元线性回归模型

$$\ln y_t = 0.9683 \ln x_t + \varepsilon_t$$

- 残差序列平稳性检验

类型	延迟阶数	τ 检验统计量的值	$Pr<\tau$
类型 1	0	−1.33	0.162 9
	1	−1.69	0.084 5
	2	−1.93	0.052 8
类型 2	0	−1.28	0.622 2
	1	−1.64	0.445 5
	2	−1.85	0.348 4

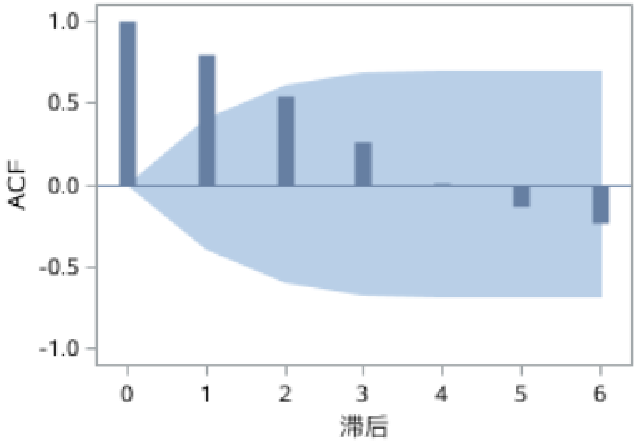
- 残差序列平稳，说明收入序列和支出序列之间协整关系成立。

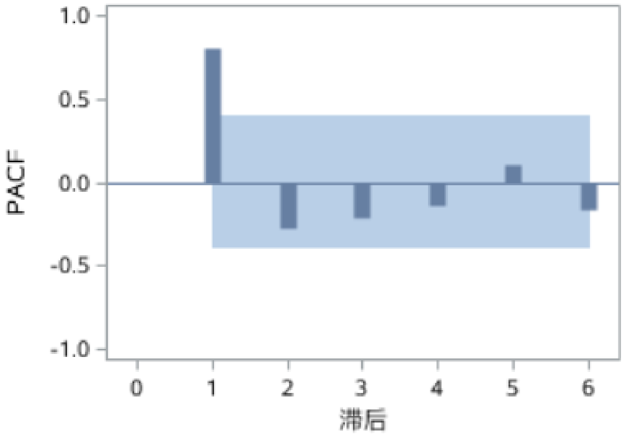
残差序列分析

- 残差白噪声检验

延迟	LB统计量	*P* 值
6	31.29	<0.001

- EG检验和白噪声检验结果显示回归残差序列为平稳非白噪声序列。我们还需要进一步提取残差序列中的相关信息。根据残差序列的自相关图和偏自相关图属性，对残差序列拟合AR(1)模型



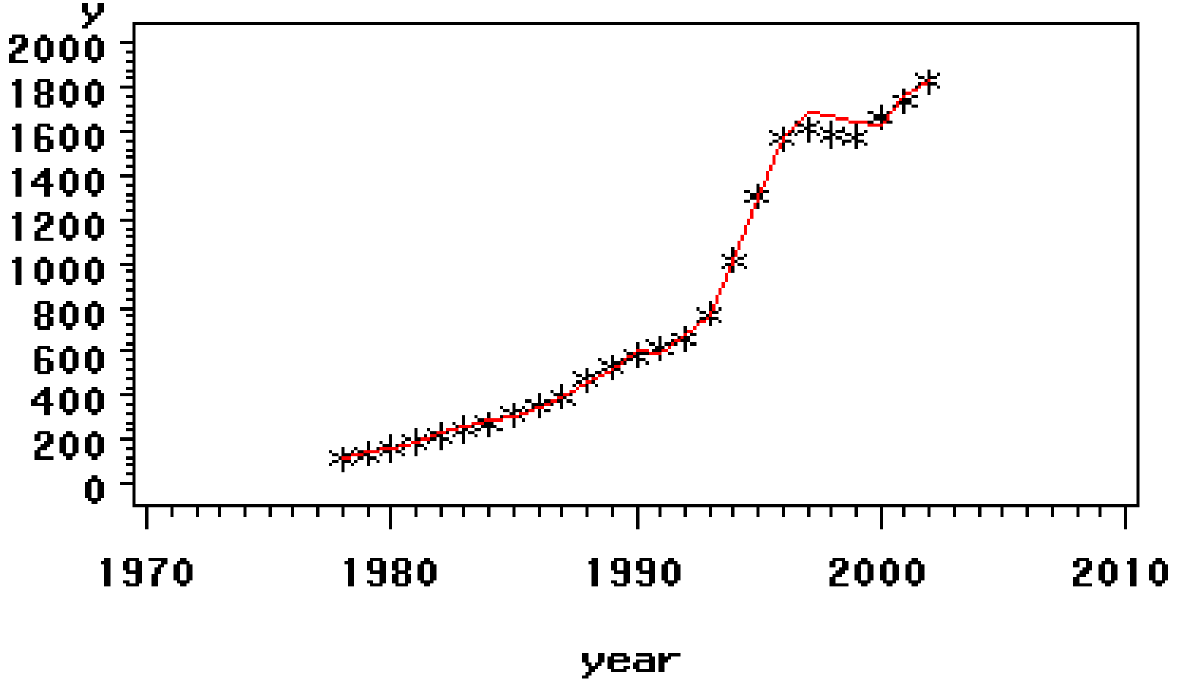


最终拟合模型

- 最终拟合模型

$$\ln y_t = 0.9682 \ln x_t + \frac{\varepsilon_t}{1-0.8372B} \quad \varepsilon_t \sim N(0, 0.0009)$$

- 模型检验
  - 参数显著非零
  - 模型显著成立



协整模型拟合效果图

序列预测

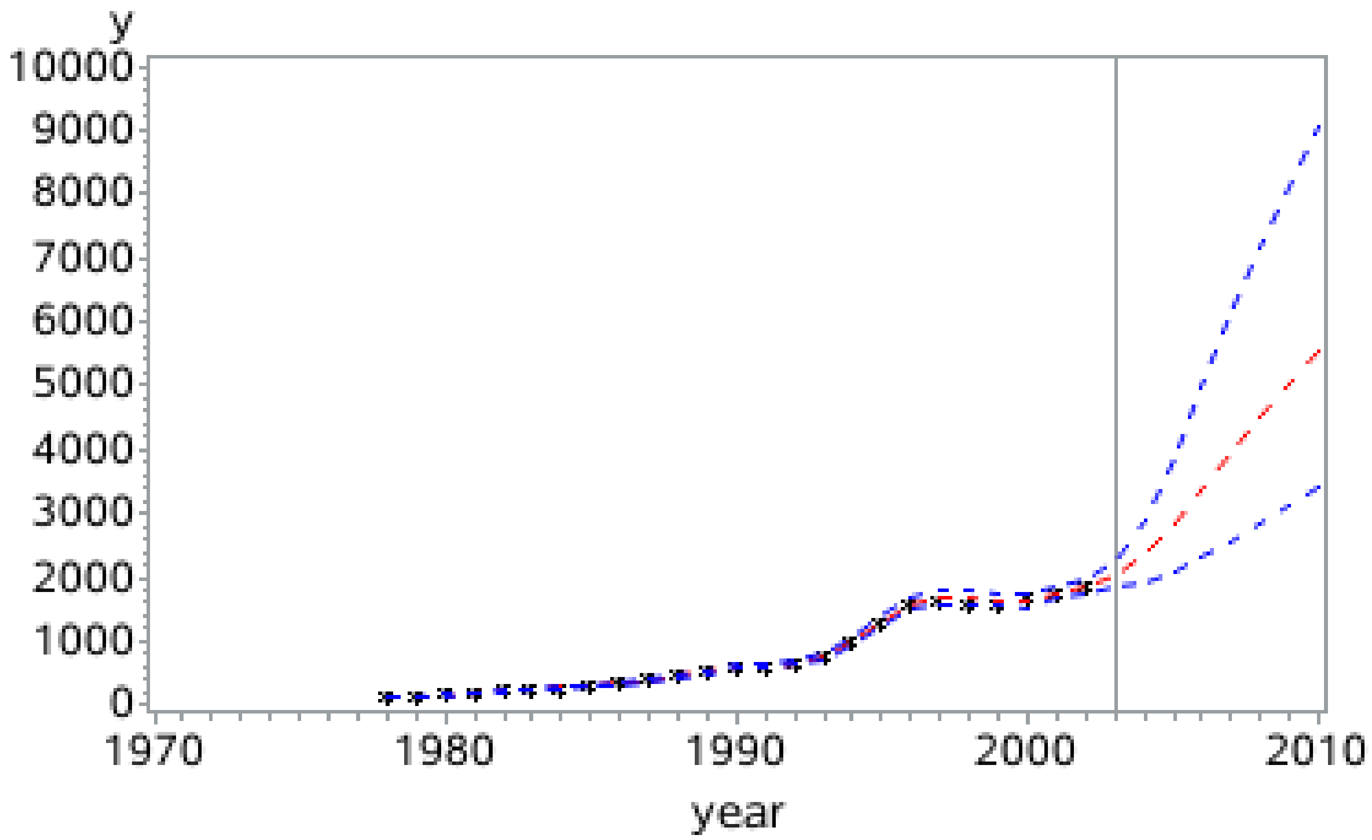
- 首先通过单变量拟合的方法获得人均收入序列的预测值

$$\nabla \ln x_t = 0.0802 + \frac{\varepsilon_t}{1 - 0.7909B + 0.4206B^3}$$

- 将人均收入对数序列的预测值代入协整动态模型，可以得到人均消费支出对数序列的预测值。然后对对数序列进行指数运算，就得到人均消费支出序列预测值

预测时期	预测值	95%置信下限	95%置信上限
2003	2 053.19	1 840.77	2 290.13
2004	2 379.33	1 940.39	2 917.56
2005	2 837.64	2 092.15	3 848.79
2006	3 374.67	2 300.60	4 950.19
2007	3 944.17	2 552.51	6 094.59
2008	4 489.90	2 822.22	7 143.02
2009	5 009.49	3 098.08	8 100.20
2010	5 539.48	3 387.72	9 057.98

协整模型的拟合与预测效果图



误差修正模型

- 误差修正模型（Error Correction Model）简称为ECM，最初由Hendry和Anderson于1977年提出，它常常作为协整回归模型的补充模型出现。
- 协整模型度量序列之间的长期均衡关系，而ECM模型则解释序列的短期波动关系。
- 误差修正模型构造思想：对协整模型进行等价差分运算，就得到短期波动关系

$$
\begin{aligned}
& y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \\
\Rightarrow\ & y_t - y_{t-1} = \beta_0 + \beta_1 x_t + \varepsilon_t - \left(\beta_0 + \beta_1 x_{t-1} + \varepsilon_{t\text{-}1}\right) \\
\Rightarrow\ & \nabla y_t = \beta_1 \nabla x_t - \varepsilon_{t-1} + \varepsilon_t
\end{aligned}
$$

- $\varepsilon_{t-1}=y_{t-1}-\beta_1 x_{t-1}$ 代表的是上一期的误差，特别记作 ECM_{t-1} ，则上式可以表达为

$$\nabla y_t = \beta_1 \nabla x_t - ECM_{t-1} + \varepsilon_t$$

短期影响因素分析

- 响应序列的当期波动 ∇y_t 主要会受到三方面短期波动的影响
  - 输入序列的当期波动 ∇x_t
  - 上一期的误差 ECM_{t-1}
  - 纯随机波动 ε_t
- 为了定量地测定这三方面影响的大小，尤其是上期误差 ECM_{t-1} 对当期波动 ∇y_t 的影响，可以构建ECM模型，模型结构如下：

$$\nabla y_t = \beta_0 \nabla x_t + \beta_1 ECM_{t-1} + \varepsilon_t$$

误差修正模型的负反馈机制

- 模型结构

$$\nabla y_t = \beta_0 \nabla x_t + \beta_1 ECM_{t-1} + \varepsilon_t$$

- 负反馈机制

$$\beta_1 < 0$$

- 以例8-3的应用背景对误差修正模型的负反馈机制进行直观解释
  - 当 $ECM_{t-1} > 0$时，即上期真实支出比估计支出大，这种误差反馈回来，会导致下期支出适当压缩
  - 当 $ECM_{t-1} < 0$ 时，即上期真实支出比估计支出小，这种误差反馈回来，会导致下期支出适当增加

例8-3续

- 对1978年 - 2002年中国农村居民家庭人均纯收入对数序列 $\ln\{x_t\}$和生活消费支出对数序列 $\ln\{y_t\}$构造ECM模型。
- 根据协整方程，得到误差序列为

$$\mathrm{ECM}_{t-1}=\ln y_{t-1}-0.9683\ln x_{t-1}$$

- 拟合误差修正模型

$$\nabla\ln y_t=0.9579\,\nabla\ln x_t-0.1537\mathrm{ECM}_{t-1}+\varepsilon_t$$

- 对误差修正模型的解读：
  - 收入的当期波动对生活消费支出当期波动的调整幅度很大，每增加1单位的对数收入，会增加0.9579单位的对数生活消费支出
  - 上期误差对生活消费支出当期波动的调整幅度不大，单位调整比例为 - 0.1537。

本章内容

因果关系的识别

- 对于多元时间序列而言，如果能找到对响应变量有显著影响的输入序列，并且能验证它们之间具有协整关系，这对准确预测响应变量的波动，或者通过控制输入序列的取值，间接控制响应变量的发展都是非常有用的。但前提是输入序列和响应序列之间具有真正的因果关系，而且一定是输入序列为因，响应序列为果。
- 这种因果关系的认定，在某些情况下是清晰明确的。比如例8-3，对于中国农村家庭而言，一定是量入为出，收入的多少影响了支出的多少，所以一定是收入为因，支出为果。这个例子中输入变量和因变量很好识别。
- 但是在有些领域，变量之间的关系可能比较复杂，因果关系的识别并不是一目了然的。尤其是在经济、金融领域，多个变量都来自相同领域，甚至是同一个系统，彼此之间多半有相关关系，但彼此之间的因果关系却并不明确。
- 我们在建立协整模型时，需要先确定变量之间的因果关系。

例8-4

- 1983年D.A.Nicols想研究对白领阶层薪水调整有决定性影响的宏观经济因素。他收集了四个相关变量：

  （1）白领阶层的平均年薪 W；

  （2）当年的通货膨胀率CPI；

  （3）当年的失业率 U；

  （4）当年的最低工资标准 MW。

- 他想研究的响应变量是白领阶层的平均年薪 W，那么剩下的三个变量是不是导致年薪变化的因变量呢？
- 如果单纯从逻辑上分析，上述四个变量之间的因果关系，我们很难直接识别。

序列因果关系的定义

- 因果关系，一定是原因导致了结果。所以从时间上说，应该是原因发生在前，结果产生在后。就影响效果而言， X 事件发生在前，而且对 Y 事件的发展结果有影响， X 事件才能称为 Y 事件的因。基于对这种因果关系的理解，1969年Granger给出了序列间因果关系的定义，我们称之为Granger因果关系定义。
- Granger因果关系定义：假设X和Y是宽平稳序列，记

（1）I_t 为 t 时刻所有有用信息的集合

$$I_t = \{x_t, x_{t-1}, x_{t-2}, \cdots, y_t, y_{t-1}, y_{t-2}, \cdots\}$$

（2）X_t 为 t 时刻所有 x 序列信息的集合

$$X_t = \{x_t, x_{t-1}, x_{t-2}, \cdots\}$$

（3）$\sigma^2(\cdot)$ 为方差函数。

Granger因果定义

- 则序列 x 是序列 y 的Granger原因，当且仅当 y 的最优线性预测函数使得下式成立：

$$\sigma^2(y_{t+1} \mid I_t) < \sigma^2(y_{t+1} \mid I_t - X_t)$$

式中，$\sigma^2(y_{t+1} \mid I_t)$ 是使用所有可获得的历史信息（其中也包含 x 序列的历史信息）得到的 y 序列一期预测值的方差；$\sigma^2(y_{t+1} \mid I_t - X_t)$ 是从所有信息中刻意扣除 x 序列的历史信息得到的 y 序列一期预测值的方差。

如果 $\sigma^2(y_{t+1} \mid I_t) < \sigma^2(y_{t+1} \mid I_t - X_t)$，则说明 x 序列历史信息的加入能提高 y 序列的预测精度。进而反推出序列 x 是因，序列 y 是果，简记为 $x \rightarrow y$。

两变量之间的4种因果关系

- 根据Granger因果关系定义，在两个序列之间存在 4 种不同的因果关系 (不考虑两变量之间当期影响)

（1）x 和 y 相互独立，简记为（x，y）；

（2）x 是 y 的 Granger 原因，简记为（$x \rightarrow y$）；

（3）y 是 x 的 Granger 原因，简记为（$x \leftarrow y$）；

（4）x 和 y 互为因果，这种情况称为 x 和 y 之间存在反馈（feedback）关系，简记为（$x \leftrightarrow y$）。

Granger因果检验

- 假设条件
  - Granger因果检验认为绝大多数时间序列的生成过程是相互独立的，所以原假设是序列 x 不是序列 y 的Granger原因，备择假设是序列 x 是序列 y 的Granger原因。
  - 构造序列 y 的最优线性预测函数，不妨记作：

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{q} \alpha_k x_{t-k} + \sum_{k=1}^{l} \gamma_k z_{t-k} + \varepsilon_t$$

此时，Granger因果检验的假设条件可以等价表达为

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_q = 0 \leftrightarrow H_1: \alpha_1, \alpha_2, \cdots, \alpha_q \text{ 不全为 } 0$$

Granger因果检验

- 检验统计量：有多种方法构建Granger因果检验的统计量，在此介绍 F 检验统计量的构造原理。在该检验方法下，需要拟合两个回归模型
  - 在原假设成立的情况下，拟合序列 y 的有约束预测模型

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{l} \gamma_k z_{t-k} + \varepsilon_{1t} \qquad SSE_r = \sum_{t=1}^{n} \varepsilon_{1t}^2 = SST - SSR_{yz}$$

  - 在备择假设成立的情况下，拟合序列 y 的无约束预测模型

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{q} \alpha_k x_{t-k} + \sum_{k=1}^{l} \gamma_k z_{t-k} + \varepsilon_t \qquad SSE_u = \sum_{t=1}^{n} \varepsilon_{2t}^2 = SST - SSR_x - SSR_{yz}$$

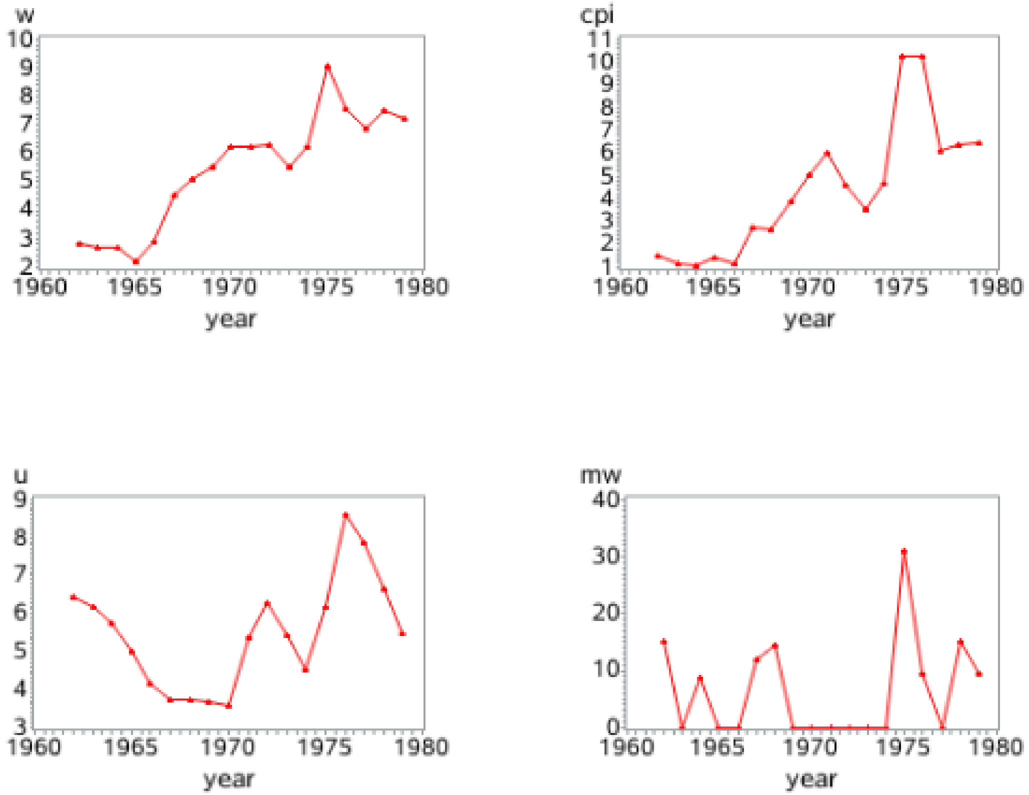
  - 检验统计量

$$F = \frac{(SSE_r - SSE_u)/q}{SSE_u/n-q-p-1} \sim F(q, n-q-p-1)$$

例8-4

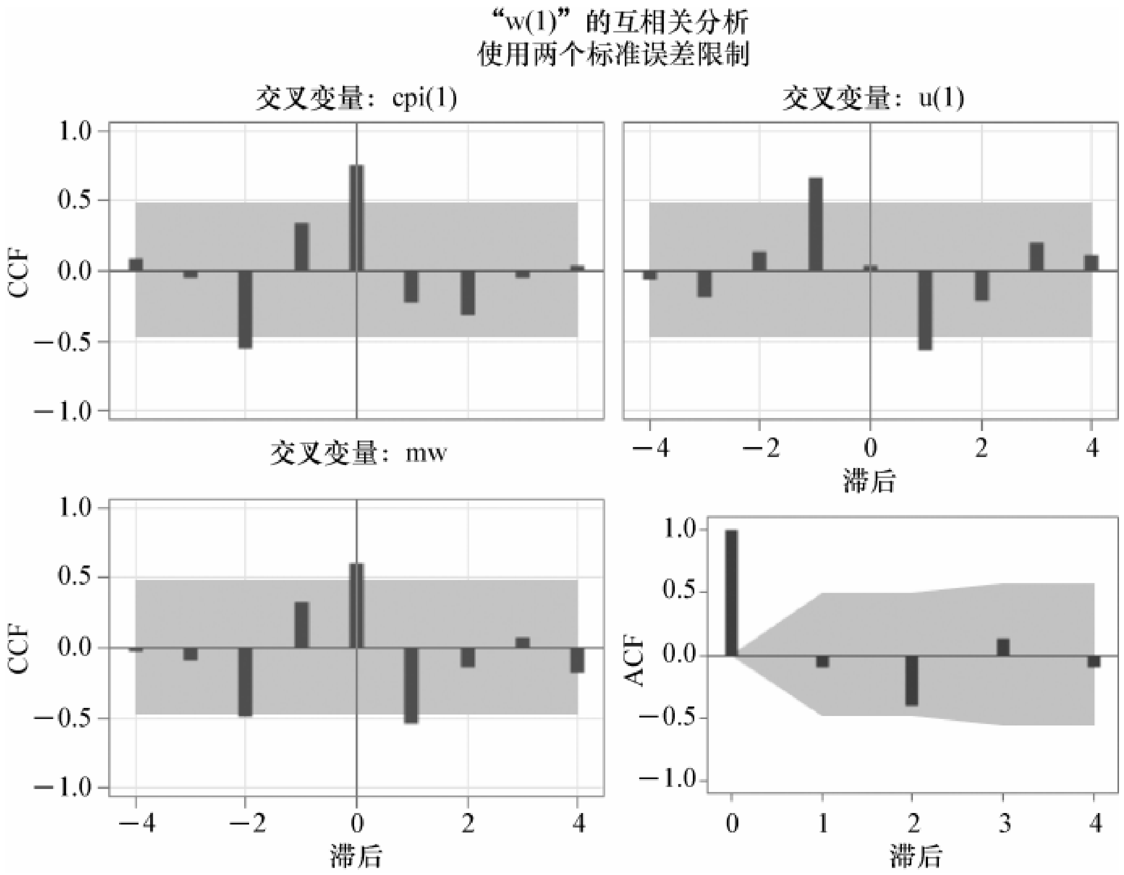
- 对1962-1979年美国白领阶层平均年薪和可能对它有显著影响的宏观经济因素进行Granger因果检验。

- 第一步：检验序列平稳性
  - 前三个序列 (W,CPI,U) 1 阶差分后平稳
  - 第四个序列(MW)原序列平稳



例8-4

- 第二步：考察年薪变量的自相关图和三个宏观经济变量与年薪变量的互相关系数图，确定输入变量的延迟阶数



例8-4

- 第三步：检验各宏观经济变量是不是年薪变量的 Granger原因

检验变量	SSE_r	SSE_u	F 统计量	P 值
CPI	3.170 9 (df=12)	2.581 7 (df=11)	2.510 4	0.139 1
U	3.207 4 (df=12)		2.666 0	0.128 5
MW	6.586 9 (df=13)		8.532 6	0.005 8

- 检验结果显示：这三个宏观经济变量只有最低工资(MW)可以认为是白领年薪波动的 Granger原因，通货膨胀率(CPI)和失业率(U)不能视为年薪变量的Granger原因。

进行Granger因果检验应该注意的问题

- 检验结果只说明样本数据特征。例8-4的Granger因果检验得出结论：白领年薪的波动受最低工资的影响，但不受通货膨胀率和失业率的显著影响。这个因果结论是基于这批样本数据得出的。如果换一批数据，或增加样本数据量，得出的因果判别可能会完全不一样。这也就是说，Granger因果检验的结果会受到样本随机性的影响。样本容量越小，样本随机性的影响就越大。所以最好在样本容量比较大时进行Granger因果检验，以保证检验结果相对稳健。
- Granger因果检验即使显著拒绝原假设，也不能说明两个序列间具有真正的因果关系。Granger因果检验的构造思想是：使响应变量预测精度有显著提高的自变量可以视作响应变量的因。这里面存在一个逻辑漏洞：如果变量 x 是变量 y 的因，那么知道 x 的信息对预测 y 是有帮助的，这个结论是对的。但反过来认为，有助于预测精度提高的变量都是响应变量的因，就不一定正确了。这就意味着，在进行 Granger因果检验时，即使得出因果关系显著成立的结论，也仅仅是预测精度提高的统计显著性判断，并不意味着两个变量之间一定存在真正的因果关系。

08

THANKS