



高等学校本科生公共课教材

时间序列分析与SAS应用

■ 肖枝洪 郭明月 编著



WUHAN UNIVERSITY PRESS

武汉大学出版社



高等学校本科生公共课教材

时间序列分析与SAS应用

■ 肖枝洪 郭明月 编著



WUHAN UNIVERSITY PRESS

武汉大学出版社

图书在版编目(CIP)数据

时间序列分析与 SAS 应用/肖枝洪,郭明月编著. —武汉:武汉大学出版社,2009.1

高等学校本科生公共课教材

ISBN 978-7-307-06780-6

I. 时… II. ①肖… ②郭… III. 时间序列分析—应用软件, SAS
IV. O211.61

中国版本图书馆 CIP 数据核字(2009)第 001336 号

责任编辑:杨 华

责任校对:黄添生

版式设计:马 佳

出版发行:武汉大学出版社 (430072 武昌 珞珈山)

(电子邮件: cbs22@whu.edu.cn 网址: www.wdp.com.cn)

印刷:荆州市鸿盛印务有限公司

开本: 720 × 1000 1/16 印张: 13 字数: 228 千字 插页: 1

版次: 2009 年 1 月第 1 版 2009 年 1 月第 1 次印刷

ISBN 978-7-307-06780-6/O · 400 定价: 20.00 元

版权所有,不得翻印;凡购我社的图书,如有缺页、倒页、脱页等质量问题,请与当地图书销售部门联系调换。

内 容 提 要

时间序列分析是数理统计的一个分支。它是一种利用具有“时间特性”的观测数据,根据研究对象的特征发掘内在规律性建立动态模型,并对之进行模式识别、参数估计,然后以此为依据对未来的行为进行科学的预测和控制的统计方法,在工程技术、经济管理、气象学、地球物理学等方面有着广泛的应用。

SAS 软件是国际上流行的统计分析的标准软件,本教材只介绍与时间序列有关的程序编写和结果分析。本教材主要介绍时间序列的概念、奇异点的诊断、自相关分析、偏自相关分析、时序模型的识别、时序模型参数的估计、预测以及多元时间序列分析。本书既可作为数学与信息专业、统计专业、经济管理专业以及工程方面的本科生教材,也可以作为科技工作者的参考书。

前 言

在自然现象和经济现象中,人们为了对某些事物或系统的运行规律探索其究竟,需要观测所要研究的某种现象,从而得到一定顺序的数据资料。通过分析这些数据资料,对事物或系统的未来发展进行预测或控制的方法,称为时间序列分析。

在实际问题中,当数据很多时,如果没有计算程序,人们很难完成工作。因此,我们在介绍时间序列分析理论的同时,对涉及的具体计算,将给出相应的 SAS 程序以及比较详细的结果分析。这样做,一方面使读者有一种能够解决问题的工具在手的感觉,以增强读者的自信心;另一方面,可以使读者进一步加深对理论分析的理解。我们认为这样处理也符合现代教学理念,因为我们现在课堂教学基本上使用多媒体,在讲解理论的同时,也可以演示计算机运行的结果。

为了方便读者阅读,在编写过程中,我们将数据与问题同时陈述出来,将程序与解答过程同时陈述出来。这样做可以帮助读者不断积累识别时间序列模型的经验,熟悉程序运行结果的解释。

在本书中,肖枝洪编写第 1 章、第 4 章和第 5 章的内容,并负责统稿;郭明月负责编写第 2 章和第 3 章的内容,并负责收集资料。本书是湖北省省级教改项目立项研究的一项成果。本书的出版得到了武汉大学出版社的大力支持,在此我们表示诚挚的谢意!

由于编者水平有限,不当之处,恳请读者批评指正!

编 者

2008 年 8 月于狮子山

目 录

1 时间序列的基本知识	1
1.1 时间序列概念	1
1.2 SAS 介绍	5
1.2.1 SAS 的显示管理系统	5
1.2.2 SAS 的程式结构	6
1.2.3 SAS 程式的输入及运行	6
1.2.4 DATA 语句	8
1.2.5 CARDS 语句	8
1.2.6 INPUT 语句	9
1.2.7 PROC 语句	9
1.2.8 PRINT 过程	10
1.3 时间序列的平稳性	13
1.3.1 统计特征	13
1.3.2 时间序列的平稳性	14
1.3.3 严平稳与宽平稳的关系	15
1.3.4 样本均值、方差、自协方差与自相关函数	16
1.3.5 平稳时间序列的意义	18
1.4 异常点检验与缺省值的补足	18
1.4.1 时间序列数据的采集	18
1.4.2 异常点的检验与处理	19
1.4.3 缺省值的补足	20
1.5 平稳性检验	21
1.6 纯随机性检验	28
1.7 方差的同质性检验	33
1.7.1 方差的同质性检验	33
1.7.2 方差的稳定性转换	35
1.8 差分运算与后移算子	38
1.8.1 差分运算	38

1.8.2 后移算子	41
习题 1	42
2 平稳时间序列	44
2.1 AR(p)模型	44
2.1.1 p 阶自回归模型	44
2.1.2 p 阶自回归模型的统计特性	49
2.2 MA 模型	58
2.2.1 q 阶移动平均模型	58
2.2.2 移动平均模型的统计特性	59
2.3 ARMA 模型(Auto Regression Moving Average Model)	65
2.3.1 ARMA(p, q)模型	65
2.3.2 ARMA(p, q)模型的统计特性	66
2.4 ARMA 模型的识别与参数估计	77
2.4.1 模型的初步识别	78
2.4.2 模型定阶	81
2.4.3 模型参数估计	88
2.4.4 模型的适应性检验和参数的显著性检验	95
2.5 平稳时间序列的预测	98
2.6 实例分析(I)	104
习题 2	109
3 非平稳时间序列的确定性分析	112
3.1 时间序列的分解	112
3.1.1 Gramer 分解定理	112
3.1.2 确定性因素分解	113
3.2 长期趋势分析及预报	114
3.2.1 平滑法	115
3.2.2 趋势拟合法	119
3.3 季节变动分析及预报	125
3.3.1 季节变动及其测定目的	125
3.3.2 季节变动分析及预测的原理与方法	125
3.4 X-11 方法简介	134
3.4.1 X-11 方法的基本思想	135
3.4.2 X-11 方法	135

习题 3	142
4 ARIMA 模型	145
4.1 平稳化方法	145
4.1.1 差分运算的实质	145
4.1.2 平稳化方法	146
4.1.3 过差分	152
4.2 ARIMA(p, d, q)模型	153
4.2.1 ARIMA(p, d, q)模型	153
4.2.2 ARIMA(p, d, q)模型参数统计与预报	155
4.3 实例分析(II)	163
习题 4	168
5 传递函数模型	172
5.1 传递函数模型	172
5.2 传递函数模型的识别	173
5.3 干预模型	189
习题 5	194
附表	196
参考文献	198

1

时间序列的基本知识

1.1 时间序列概念

在人们的社会活动和科学试验中,经常会碰到按照一定的顺序观察得到的数据,如股票市场的每日波动,气象变化,某工厂装船货物数量的月度序列,公路事故数量的周度序列,某化工生产过程按小时观测的产量,等等,且这些数据之间具有相依性.例如,平台式惯性导航系统的陀螺漂移数据(表 1.1.1),中国 1950—2006 年财政收入总额(表 1.1.2).而人们要根据这些数据进行观察、研究,寻找它的变化发展规律来拟合出某种最优的数学模型,用以预测未来的发展趋势.概率统计的一个分支——时间序列分析就是专门处理此类问题的.

表 1.1.1

某陀螺逐日漂移测试数据

99.12.21	99.12.22	99.12.23	99.12.24	99.12.25	99.12.26
-0.927	-0.919	-0.968	-1.013	-1.02	-1.028
99.12.27	99.12.28	99.12.29	99.12.30	00.01.01	00.01.02
-1.035	-1.03	-1.025	-1.019	-1.01	-1.005
00.01.03	00.01.04	00.01.05	00.01.06	00.01.07	00.01.08
-0.996	-0.99	-0.985	-1.003	-1.007	-1.006

(数据见《战术导弹控制技术》,2006 年第 2 期)

表 1.1.2

中国 1950—2006 年财政收入总额表

单位:亿元

年份	1950	1951	1952	1953	1954	1955
财政收入	62.17	124.96	173.96	213.24	245.17	249.27
年份	1956	1957	1958	1959	1960	1961
财政收入	280.19	303.20	379.62	487.12	572.29	356.66
年份	1962	1963	1964	1965	1966	1967
财政收入	313.55	342.25	399.54	473.32	558.71	419.36
年份	1968	1969	1970	1971	1972	1973
财政收入	361.25	526.76	662.90	744.73	766.56	809.61
年份	1974	1975	1976	1977	1978	1979
财政收入	783.14	815.61	776.58	874.46	1132.26	1146.26
年份	1980	1981	1982	1983	1984	1985
财政收入	1159.93	1175.79	1212.33	1366.95	1642.86	2004.82
年份	1986	1987	1988	1989	1990	1991
财政收入	2122.01	2199.35	2357.24	2664.90	2937.10	3149.48
年份	1992	1993	1994	1995	1996	1997
财政收入	3483.37	4348.95	5218.10	6242.20	7407.89	8651.14
年份	1998	1999	2000	2001	2002	2003
财政收入	9875.95	11444.08	13395.20	16386.04	18903.64	21715.25
年份	2004	2005	2006			
财政收入	26355.88	31649.29	35423.38			

(数据来源: 中国统计局网站 www.stats.gov.cn(2008-03-28))

1. 时间序列的概念

从统计意义上来看, 所谓时间序列就是将某一个指标在不同时间上的不同数值, 按照时间的先后次序排列而成的数列. 这种数列由于受到各种偶然

因素的影响,往往表现出某种随机性,彼此之间存在着统计上的依赖关系。例如,北京市区 76—7 号孔 1983—1991 年实测水位月度序列(表 1.1.3)。

表 1.1.3 北京市区 76—7 号孔 1983—1991 年实测水位数据表

年份 月	实测水位								
	1983	1984	1985	1986	1987	1988	1989	1990	1991
1	42.50	42.56	42.14	42.23	42.50	42.35	42.73	42.61	42.36
2	42.52	42.60	42.11	42.29	42.48	42.36	42.61	42.53	42.26
3	42.39	42.62	42.21	42.44	42.48	42.35	42.63	42.59	42.41
4	42.08	42.45	42.16	42.49	42.58	42.24	42.72	42.61	42.59
5	42.28	42.37	42.15	42.63	42.57	42.50	42.61	42.60	42.68
6	42.18	42.31	42.39	42.49	42.75	42.35	42.82	42.70	42.82
7	42.15	42.34	42.83	43.01	42.53	42.81	42.91	42.82	42.95
8	42.42	42.73	43.10	43.00	42.84	43.07	42.65	42.67	42.67
9	42.33	42.44	42.78	43.11	42.71	42.83	42.82	42.77	42.83
10	42.23	42.16	42.37	42.79	42.60	42.75	42.82	42.59	42.48
11	42.24	42.08	42.19	42.75	42.40	42.44	42.58	42.56	42.26
12	42.51	42.15	42.22	42.62	42.34	42.56	42.59	42.51	42.19

(数据来源:《工程勘察》,1997 年第 1 期)

从系统意义来看,时间序列就是某一系统在不同时间(地点、条件等)的响应。

这个定义从系统运行的观点出发,指出时间序列是按一定顺序排列而成的,此处“一定顺序”既可以是时间顺序,也可以是具有不同意义的物理量,如温度、速度、长度或其他单调递增取值的物理量。例如,气压随着高度的增加而变化,这里的“高度”就可以看做“时间”。可见,时间序列只强调顺序的重要性,而并非强调必须以时间顺序排列。另外,时间序列也是所研究系统的历史行为的客观记录,因而它包含了系统结构特征及运行规律。

所以,我们可以通过对时间序列的研究来认识所研究系统的结构特征(如周期波动的周期、振幅、趋势的种类),揭示其运行规律,进而用以预测,控制其未来行为,修正和重新设计系统(如改变其周期、参数),使之按照新的结构运行。

从数学意义来讲,对某一过程中的某一个变量或一组变量 $X(t)$ 进行观察测量,在一系列时刻 $t_1, t_2, \dots, t_N$ ($t_1 < t_2 < \dots < t_N$) 得到的离散有序数集合 $X(t_1), X(t_2), \dots, X(t_N)$, 称为离散时间序列。

由以上可知,时间序列有以下特点:

- ① 顺序性;
- ② 随机性;
- ③ 前后时刻(不一定相邻)的数据具有相依性,这种关系就是系统的动态规律性;

④ 从整体上看,时间序列呈现某种趋势性或周期性.

2. 时间序列分类

时间序列根据所研究的依据不同,有如下分类:

- ① 按所研究对象的多少分,有一元时间序列和多元时间序列.
- ② 按时间的连续性可分为离散时间序列和连续时间序列. 本书主要研究离散时间序列.

③ 按序列的统计特性分,有平稳时间序列分析和非平稳时间序列分析.

④ 按序列的分布规律分,有 Gauss 时间序列和 Non-Gauss 时间序列分析.

本书所介绍的模型多数是假设时间序列模型为 Gauss 时序模型.

3. 时间序列分析的主要方法

时间序列分析的主要方法有两种:

- ① 平稳时间序列分析;
- ② 非平稳时间序列分析.

而非平稳时间序列分析又分为两种:

- ① 非平稳时间序列确定性分析;
- ② 非平稳时间序列随机分析.

本书将分别在第 2,3,4 章中对之进行详细介绍.

4. 时间序列分析的方法

时间序列分析的方法若按采用的手段不同则可分为数据图法、指标法和模型法.

数据图法是将时间序列在平面坐标系中绘出坐标图.

根据图形直接观察序列的总趋势和周期变化以及异常点、升降转折点等.

优点:简单直观,易懂易用.

缺点:获取的信息少且肤浅,需要有相当丰富的经验,分析结构的主观性较大.

指标法是指通过计算一系列核心指标来反映所研究系统的动态特征.

优点：客观。

缺点：获取的信息少、肤浅且有限。

模型法是对给定的时间序列，根据统计理论和方法，建立描述该序列的适应或最优的统计模型，并据此进行预测或控制。本书将着重介绍这部分内容。

- 思考题**
1. 时间序列中的“时间”只是特定的时间吗？
 2. 时间序列在不同时间上的表现是否要相关？

1.2 SAS 介绍

Statistical Analysis System 简称 SAS，可用来分析数据和编写报告。它是美国 SAS 研究所的产品，在国际上被誉为标准通用软件，在我国深受医学、农林、财经、社会科学、行政管理等众多领域的专业工作者的好评。

有关 SAS 的最新信息，可以查看 <http://www.sas.com.cn>。

SAS 采用积木式模块结构，其中的 SAS/STAT 模块是目前功能最强的多元统计分析程序集，可以作回归分析、聚类分析、判别分析、主成分分析、因子分析、典型相关分析、各种试验设计的方差分析、协方差分析以及时间序列分析。

1.2.1 SAS 的显示管理系统

启动计算机，点击 SAS 图标后，即可进入 SAS 的显示管理系统。

在 View 中有 4 个主要的窗口(其他的先不考虑)：

- ① 编辑窗口(Program Editor)——编辑程式和数据文件；
- ② 日志窗口(Log)——记录运行情况，显示 ERROR 信息；
- ③ 输出窗口(Output)——输出运行的结果；
- ④ 图形窗口(Graph)——输出图形。

点击 View 菜单中的 Program Editor, Log, Output, Graph 命令可以分别进入编辑、日志、输出及图形窗口。

按功能键 F5, F6, F7 也可以分别进入编辑、日志及输出窗口。

退出 SAS 有两种方法：

- ① 点击 File 菜单中的 Exit 命令；
- ② 点击窗口右上角的×。

1.2.2 SAS 的程式结构

在 SAS 中, 对数据的处理可分为两大步骤:

- ① 将数据读入 SAS 建立的 SAS 数据集, 称为数据步;
- ② 调用 SAS 的模块处理和分析数据集中的数据, 称为过程步。

每一数据步都是以 DATA 语句开始, 以 RUN 语句结束。而每一过程步则都是以 PROC 语句开始, 以 RUN 语句结束。当有多个数据步或过程步时, 由于后一个 DATA 或 PROC 语句可以起到前一步的 RUN 语句的作用, 两步中间的 RUN 语句也就可以省略。但是最后一步的后面必须有 RUN 语句, 否则不能运行。

SAS 还规定, 每个语句的后面都要用符号“;”作为这个语句结束的标志。

在编辑 SAS 程式时, 一个语句可以写成多行, 多个语句也可以写成一行, 可以从一行的开头写起, 也可以从一行的任一位置写起。每一行输入完成后, 用 ENTER 键可以使光标移到下一行的开头处。

1.2.3 SAS 程式的输入及运行

SAS 程式的输入及运行步骤如下:

- ① 进入 SAS 的显示管理系统;
- ② 进入并扩大编辑窗口;
- ③ 调出、编辑或修改 SAS 程式或数据文件;
- ④ 将编辑窗口的 SAS 程式或数据文件存盘;
- ⑤ 按功能键 F8 或点击“”图标运行 SAS 程式并注意观察日志窗口中的信息, 如有 ERROR 出现, 则将光标移到日志窗口, 找到错误的所在;
- ⑥ 将光标移到编辑窗口, 按功能键 F4 或点击 Locals 菜单中的 Recall text 命令调出已经运行的 SAS 程式, 改正错误后转入步骤④, 直到日志窗口中的信息没有 ERROR 出现为止;
- ⑦ 将光标移到输出窗口, 用 PgUp 和 PgDn 两键翻页阅读输出的结果。

【例 1.2.1】 进入 SAS 的显示管理系统, 在编辑窗口输入程式:

```
data example1_2_1;
  a=1; b=2; c=3;  x=2*a+b;  y=a-b/2;  z=b+c*3;
proc print; run;
```

在程式中, data 后面的 example1_2_1 是给数据集所取的名字, 在这个数据集内将储存变量 a, b, c, x, y, z 及它们所取的值, proc 后面的 print 是要打印 6 个变量所取的值。

(注:标点符号一定要是在英文状态下的形式.)

将程式提交运行后,注意日志窗口中的信息:

NOTE: The data set WORK.EXAMPLE1_2_1 has 1 observation and 6 variables.

NOTE: The DATA statement used 0.71 seconds.

若将例 1.2.1 的程序最后一行改为

```
proc; print; run;
```

将程式提交运行后,注意日志窗口中的信息:

```
3 proc;
```

ERROR 10-205: Expecting the name of the procedure to be executed.

```
3 print;
```

ERROR 180-322: Statement is not valid or it is used out of proper order.

```
3 run;
```

ERROR 180-322: Statement is not valid or it is used out of proper order.

在日志中记录了一个错误,那就是在 proc 与 print 之间不应该有“;”这个符号. 将它去掉以后,请再注意日志窗口中的信息并阅读输出窗口中输出的结果.

【例 1.2.2】 进入 SAS 的显示管理系统,在编辑窗口输入程式:

```
data example1_2_2; input no$ sex$ age h w@@;
cards;
10. f 47 156.3 47.1 24 m 38 172.4 61.5 53 m 41 169.2
64.5 46 f 52 158.2 53.6 38 f 39 160.1 48
;
proc sort; by sex;
proc means; by sex; var h w;
run;
```

在程式中, input 定义与数据相对应的变量名、顺序及类型,其中 no\$ 与 sex\$ 是非数字型的字符串变量, cards 是数据行开始的标志,数据行下面的“;”独占一行是数据行结束的标志, sort 是将 SAS 数据集中的观测值按一个或多个变量进行排序,后面的 by sex 是要根据 sex 的值对上述观测值进行分组处理,而 means 则是要在分组处理后按 var 的要求给出变量 h 和 w 的简单的描述性统计分析结果,如图 1.2.1 所示.

在程式中, cards; 与“;”之间是数据,第一列是变量 no\$ 的数据,表示学生的号码;第二列是变量 sex\$ 的数据,表示性别;第三列是变量 age 的数

-----SEA=f-----						
N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
3	H	3	156.3	160.1	158.2	1.90
	W	3	47.1	53.6	49.6	3.52
-----SEA=m-----						
N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
2	H	2	169.2	172.4	170.8	2.26
	W	2	61.5	64.5	63.0	2.12

图 1.2.1

据，表示年龄；第四列是变量 h 的数据，表示身高；第五列是变量 w 的数据，表示体重。

而输出的结果为：女生组有 3 人，身高在 156.3 至 160.1 之间，均值为 158.2，标准差为 1.90；体重在 47.1 至 53.6 之间，均值为 49.6，标准差为 3.52；男生组有 2 人，身高在 169.2 至 172.4 之间，均值为 170.8，标准差为 2.26；体重在 61.5 至 64.5 之间，均值为 63，标准差为 2.12。

1.2.4 DATA 语句

DATA 语句的作用是表明数据步的开始并给出数据集的名称。

DATA 语句的格式为

DATA 数据集的名称；

数据集的名称必须以英文字母开始，最长不超过 8 个字符。

例如建立一个名为 EX 的数据集的语句为“DATA EX;”，这时 SAS 系统会自动地把 EX 作为数据集名添加到 WORK 中，因此在日志窗口显示的信息中记该数据集为 WORK.EX。但是这个数据集是临时的，它仅仅在程式运行期间有效，关闭 SAS 系统后它就被覆盖，不能重新调用。

1.2.5 CARDS 语句

CARDS 语句的作用是与“;”呼应，标志数据行的开始与结束。

CARDS 语句的格式为

CARDS;

数据行

;

如果使用 CARDS 语句，在 CARDS 的后面必须紧跟数据行，并且在一个数据步中最多只能有一个 CARDS 语句。

1.2.6 INPUT 语句

INPUT 语句的作用是描述输入记录中的数据,并把输入值赋给相应的变量。

INPUT 语句的格式为

INPUT 数据的变量名、顺序及类型;

用 INPUT 语句是为了读外部文件的数据或跟在 CARDS 语句后面的数据。除非在 INPUT 语句中的变量名后有串符号或用字符的输入格式表示、或该变量事先已被定义为字符型,否则 SAS 认定用 INPUT 语句读入的是数值型变量的值。

用 INPUT 语句时,外部文件中的数据和 CARDS 语句后面的数据都采取列表输入的方法,各个变量的值由它们之间的空格来分隔。为从一行读入多个观测值,应使用行保持符 @@ 限制读数指针,使其保持在这一行上读数,直到数据读完为止。

【例 1.2.3】 输入程式:

```
input x y@@;
cards;
3.16 2.9 5.8 3 4.9 4.17
;
```

1.2.7 PROC 语句

PROC 语句的作用是指定需要调用的程序(或过程)以及该程序的若干选择项。

PROC 语句的格式为

PROC SAS 的程序名;

例如,调用 PRINT 过程,打印数据集 EX 的内容:

```
PROC PRINT DATA=EX;
```

这里的 DATA=数据集名,用来指定本过程所要处理的数据集名,如缺省则处理最新建立的数据集。

在 PROC 步中,还必须确认一些最基本的信息,包括:

① 处理的数据集名,格式为

DATA=数据集名;

② 所涉及的变量名,格式为

VAR 变量名;