常用时间序列统计预测方法

预测是一门预计未来事件的艺术，一门科学。它可以借助大量的信息资料和现代化计算手段，比较准确地揭示出客观事物运行中的本质联系及发展趋势。预测包含采集历史数据并应用某种模型来外推将来（定量预测）；也可以是对未来的主观或直觉的预期（定性预测）。而统计预测属于定量预测方法研究范畴。

统计预测是广泛的预测实践中，应用最广的预测方法，通过对大量的数据资料进行统计分析，以求得比较准确的预测结果的理论和方法。将统计预测的理论和方法应用于医学领域，即医学统计预测。

一、 统计预测分类

表 1 各类时间序列定量预测方法的特点及适用情况

方法	时间范围	适用情况
一元线性回归	短中期	自变量与因变量之间存在线性关系
多元线性回归	短中期	因变量与两个或两个以上的自变量之间存在线性关系
非线性方法	短中期	因变量与一个或多个自变量之间存在非线性关系
趋势外推法	中长期	被预测项目的有关变量用时间表示时，用非线性回归
分解分析法	短期	一次性或作为消除季节变动因素的方法
指数平滑方法	短期	具有或不具有季节变动的反复预测
灰色预测方法	短中期	时序的发展呈指数型趋势
ARMA 预测方法	短期	平稳时间序列预测
ARIMA 预测方法	短期	不带季节变动的反复预测
季节变动预测	短中期	一年内呈现一定周期规律的、每年（月、季度）重复出现的变动
景气预测方法	短中期	时序的延续及转折预测
状态空间模型	短中期	各类时序预测
马尔科夫分析法预测	短期	连续时间变化被划分为多个状态，计算系统状态转移率进行预测
数据挖掘方法	多种情况	高维时间序列或非时间序列的判断与预测

二、 指数平滑法

指数平滑法用序列过去值的加权均数来预测将来的值，并给序列中近期的数据以较大的权重，远期的数据给以较小的权重。该方法的目的是为了去除一些随机波动，从而找到其中显而易见的规律性，并对未来发展趋势进行合理的预测。

1. 基本思想

设时间序列为x1,x2,…xt,…,对新数据和前一时间点数据（老数据）是不同等对待的，α 是权重系数（1≥α ≥0），平滑值计算的思想是：平滑值= α *（新数据）+（1- α ）*（老数据）。指数平滑法的基本公式

$$S_t = \alpha x_t + (1-\alpha)S_{t-1}$$

其中S_t为时间 t 的平滑值；x_t为时间 t 的实际值；S_{t-1}为时间 t-1 的平滑值。

可见,实际观测值对预测值的影响随着时间距离的增大而呈指数级数衰减,这就是“指数”的由来。其衰减的速度由平滑系数 α 决定,如果 α =1,说明T+1时刻的预测值只由T时刻测值决定,而与T时刻之前的任何数值无关;当 α 接近1时,时间序列的衰减速度非常快,预测只受最近的几个观测值的影响,受远处的影响很少;当 α 接近0时,即使远处的观测值也对的预测有相当的影响力;如果 α =0,说明序列非常稳定,不受T时刻的观测值的影响,只由历史数据决定。

但是,指数平滑法在应用时也存在一些问题。

(1)指数平滑法只适合于影响随时间的消逝呈指数下降的数据。

(2)指数平滑法进行预测的关键是如何确定平滑参数α,从原理上讲它应该根据序列权重衰减的快慢来定,但这方面很难提供进一步可供操作的准则。一般来说,如果希望模型的灵敏度大些,即希望尽快跟上新的变化,让近期的影响更明显,α应该大些。反之，如果希望模型稳定些，不易受近期随机变动的影响,则把α放小些。

(3)指数平滑法的每次预测都是根据上一个数值得来的。那么第一个数,即初始值如何确定?一般来说,就用序列的第一个数作为初始值。如果数据点较多,那么经过指数衰减后,初始值的影响就不明显了。从经验上讲,当数据点多于40个,初始值的影响就不太明显了。但是如果数据点少,则初始值的影响会很大,甚至大于近期的数据点,这就会违背指数平滑影响呈指数衰减的假设，此时应该慎重考虑初始值的问题。

(4)指数平滑法适用于平均水平基本保持不变的序列。对于上升的数据,预测总偏低;下降的数据,预测总偏高。因此对于有上升或下降趋势的序列应当先通过差分使序列平稳化,对于有季节变化的数据可以用季节差分处理。

1. SPSS 软件实现

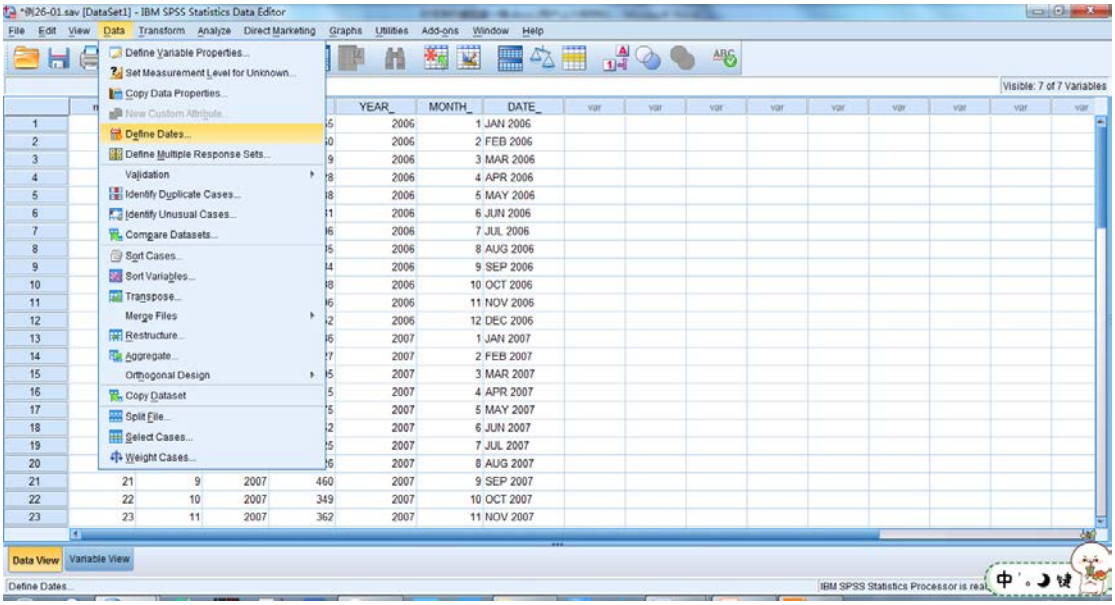
例 1　为了探索结核病发病规律，做好早期预防，某研究者整理了某地区结核病防控所登记的 2006-2009 年每月的结核病报告人数，资料见表 1，请对该地区结核病月报告人数进行预测分析。

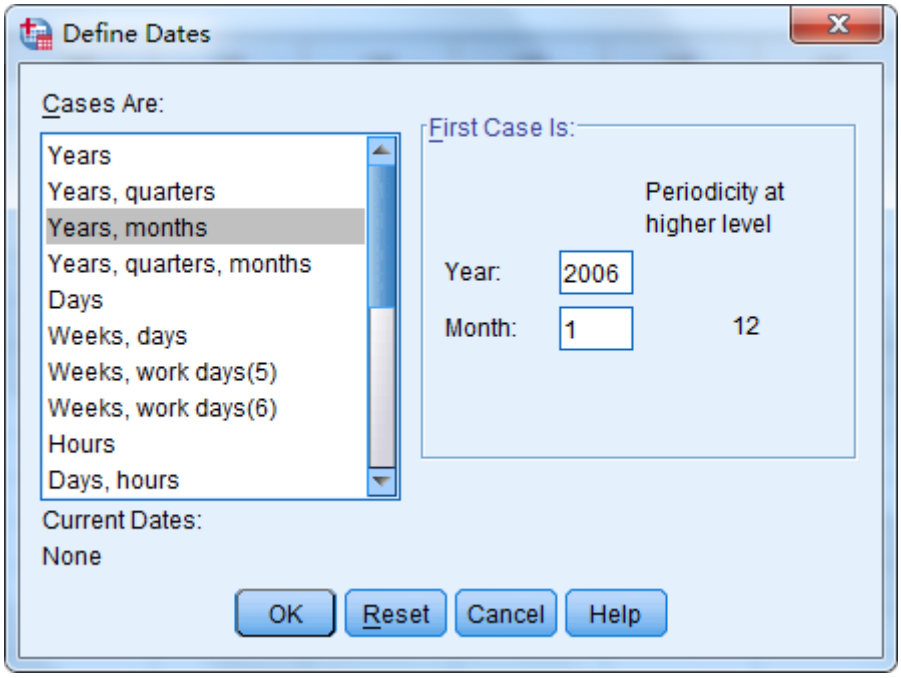
表 1　某地 2006-2009 年结核病的报告人数

序列号	月次	报告人数	序列号	月次	报告人数
1	1	255	9	9	434
2	2	260	10	10	338

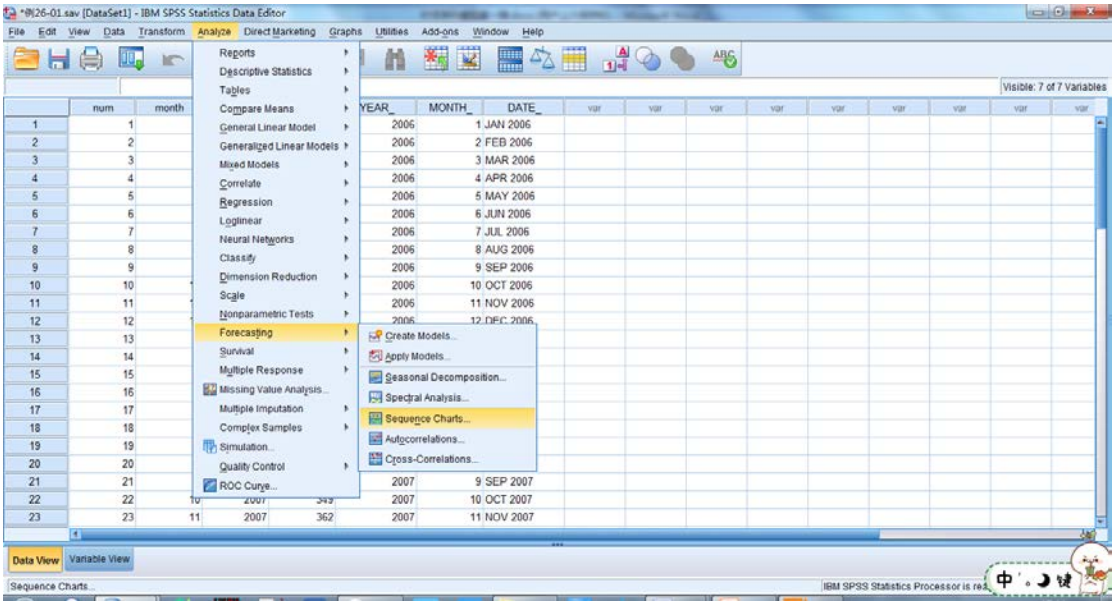
3	3	419	11	11	396
4	4	428	12	12	352
5	5	338	13	1	306
6	6	431	14	2	247
7	7	406	⋮	⋮	⋮
8	8	435	48	12	473

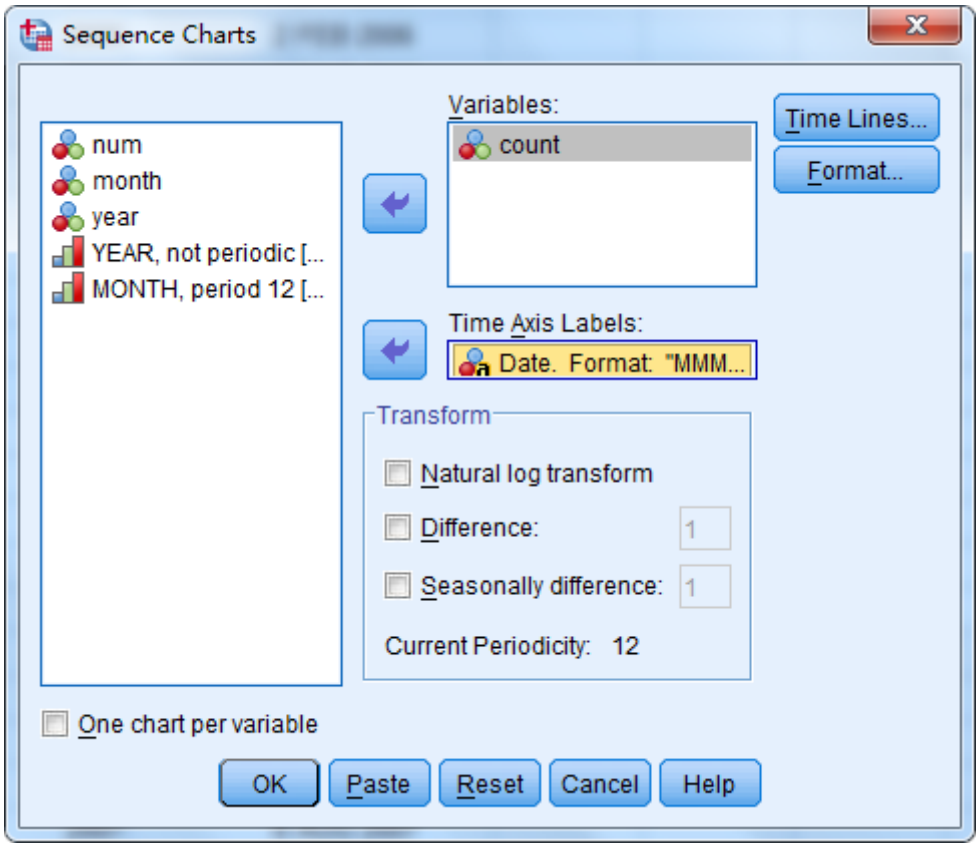
Step01：定义日期

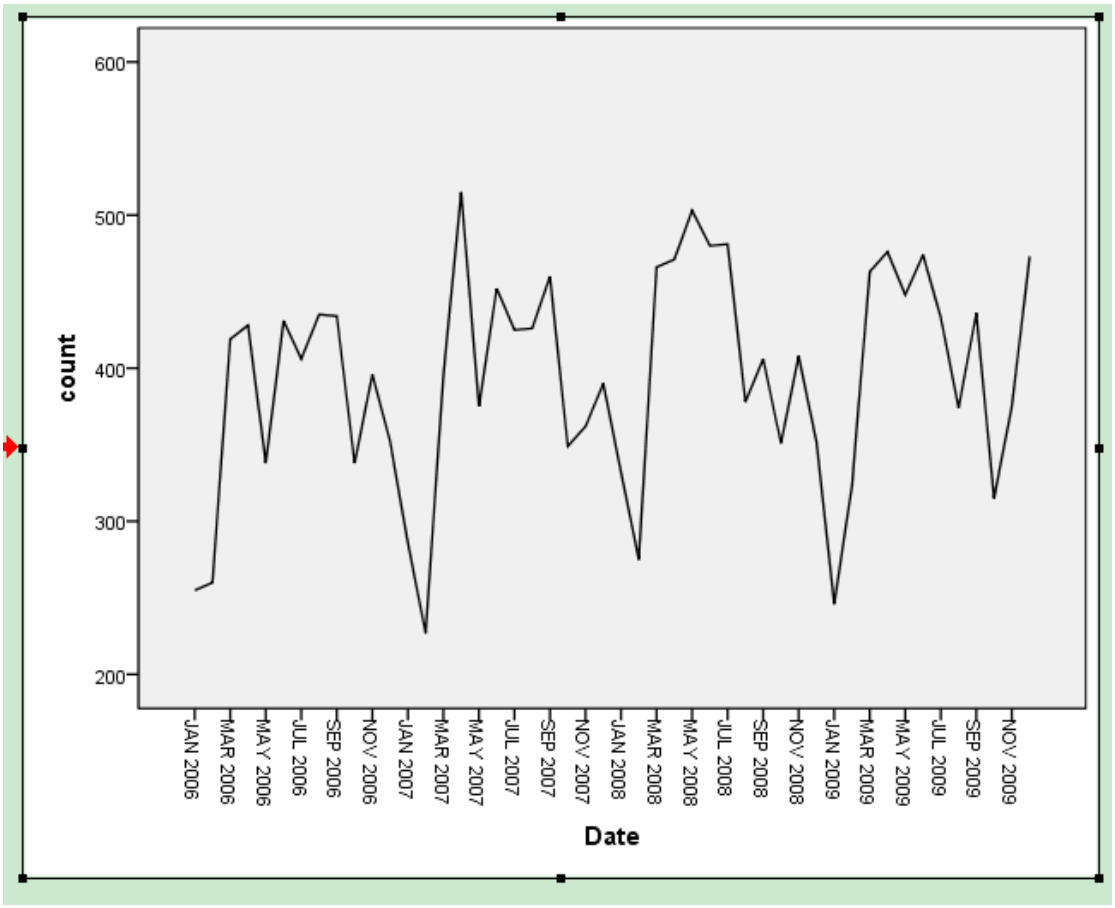




Step02：序列图

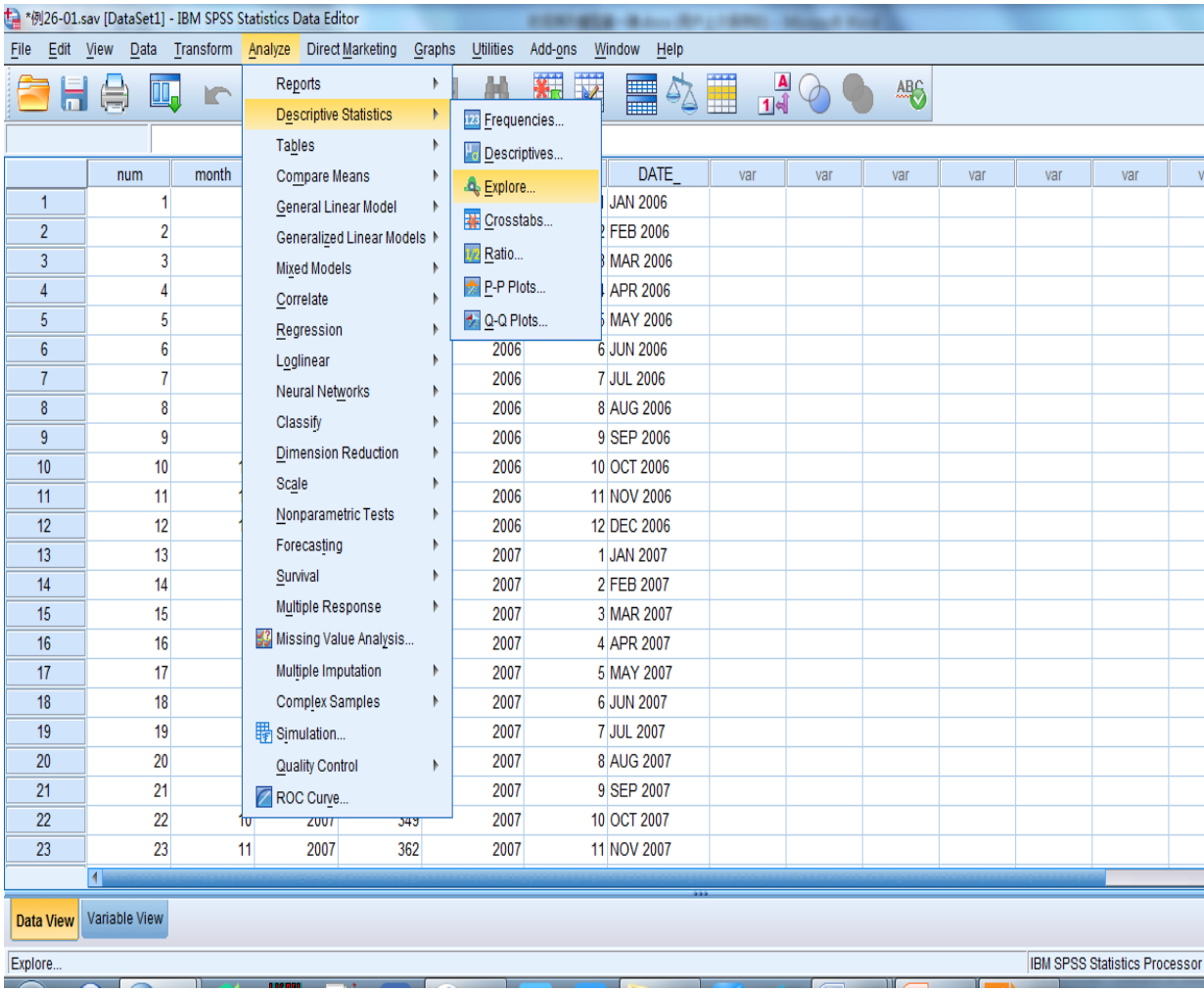

*例26-01.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Create Models...
Apply Models...
Seasonal Decomposition...
Spectral Analysis...
Sequence Charts...
Autocorrelations...
Cross-Correlations...
Visible: 7 of 7 Variables
num month YEAR_ MONTH_ DATE_
1 1 2006 1 JAN 2006
2 2 2006 2 FEB 2006
3 3 2006 3 MAR 2006
4 4 2006 4 APR 2006
5 5 2006 5 MAY 2006
6 6 2006 6 JUN 2006
7 7 2006 7 JUL 2006
8 8 2006 8 AUG 2006
9 9 2006 9 SEP 2006
10 10 2006 10 OCT 2006
11 11 2006 11 NOV 2006
12 12 2006 12 DEC 2006
21 21 2007 9 SEP 2007
22 22 2007 10 OCT 2007
23 23 11 2007 362 2007 11 NOV 2007
Data View
Variable View
Sequence Charts...
IBM SPSS Statistics Processor is re


Sequence Charts
num
month
year
YEAR, not periodic [...
MONTH, period 12 [...
Variables:
count
Time Lines...
Format...
Time Axis Labels:
Date. Format: "MMM...
Transform
Natural log transform
Difference: 1
Seasonally difference: 1
Current Periodicity: 12
One chart per variable
OK
Paste
Reset
Cancel
Help



从图中看出，对于该数据序列图没有明显的趋势，但有明显的季节性。而且没有离群值和缺失值。

Step03：特征分析


*例26-01.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
num month DATE_ var var var var var var
1 1 JAN 2006
2 2 FEB 2006
3 3 MAR 2006
4 4 APR 2006
5 5 MAY 2006
6 6 2006 6 JUN 2006
7 7 2006 7 JUL 2006
8 8 2006 8 AUG 2006
9 9 2006 9 SEP 2006
10 10 2006 10 OCT 2006
11 11 2006 11 NOV 2006
12 12 2006 12 DEC 2006
13 13 2007 1 JAN 2007
14 14 2007 2 FEB 2007
15 15 2007 3 MAR 2007
16 16 2007 4 APR 2007
17 17 2007 5 MAY 2007
18 18 2007 6 JUN 2007
19 19 2007 7 JUL 2007
20 20 2007 8 AUG 2007
21 21 2007 9 SEP 2007
22 22 2007 10 OCT 2007
23 23 11 2007 362 2007 11 NOV 2007
Data View
Variable View
Explore...
IBM SPSS Statistics Processor

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
count	48	100.0%	0	0.0%	48	100.0%

Descriptives

			Statistic	Std. Error
count	Mean		393.60	10.393
	95% Confidence Interval for Mean	Lower Bound	372.70	
		Upper Bound	414.51	
	5% Trimmed Mean		396.01	
	Median		406.00	
	Variance		5184.202	
	Std. Deviation		72.001	
	Minimum		227	
	Maximum		515	
	Range		288	
	Interquartile Range		102	
	Skewness		-.554	.343
	Kurtosis		-.394	.674

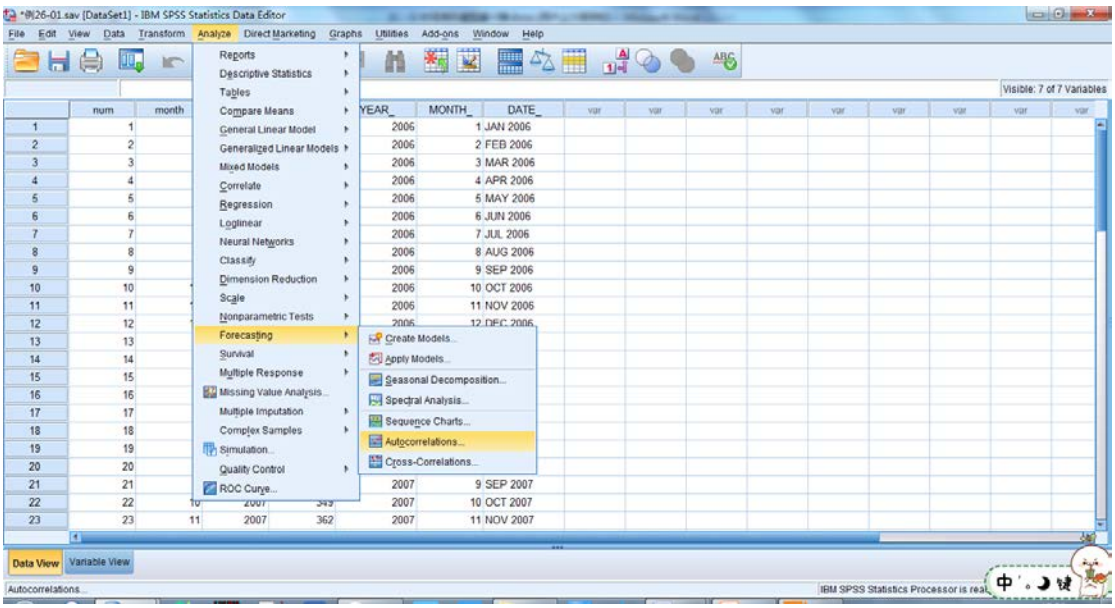
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
count	.106	48	.200*	.958	48	.086

*. This is a lower bound of the true significance.

该序列总共 48 例数据，服从正态分布，均数为 393.30，标准差为 72.001。

Step04：相关分析



A：样本自相关系数以及自相关图

在 SPSS 中给出了不同滞后期的样本自相关系数的值（自相关系数列），样本自相关系数的标准误差（标准误差列），以及 Box-Ljung 统计量的值、自由度和相伴概率。通过标准误差值以及 Box-Ljung 统计的相伴概率都可以说该时间序列不是白噪声，是具有自相关性的时间序列，可以建立指数平滑模型、ARIMA 等模型。

Autocorrelations

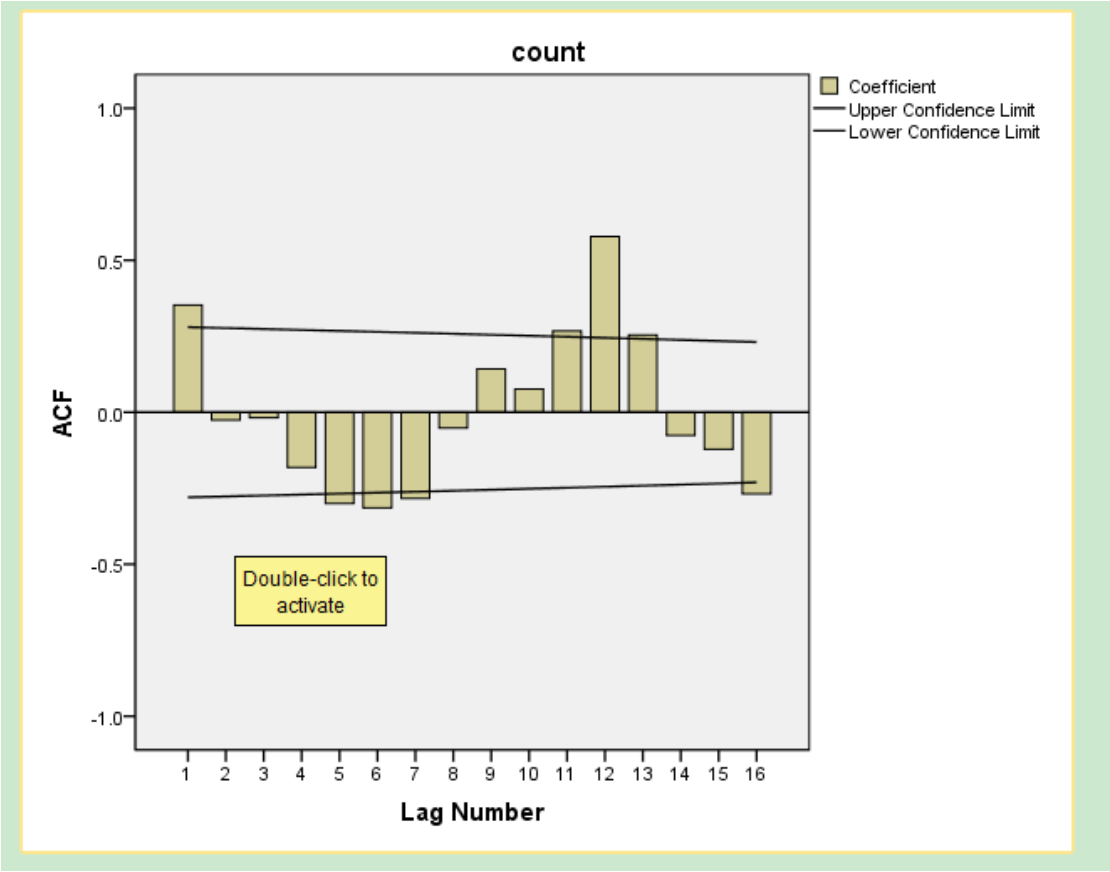
Series: count

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	.353	.140	6.347	1	.012
2	-.026	.138	6.382	2	.041
3	-.018	.137	6.399	3	.094
4	-.181	.135	8.193	4	.085
5	-.300	.134	13.214	5	.021
6	-.315	.132	18.875	6	.004
7	-.284	.131	23.588	7	.001
8	-.052	.129	23.750	8	.003
9	.142	.127	24.998	9	.003
10	.076	.126	25.363	10	.005
11	.267	.124	30.002	11	.002
12	.578	.122	52.307	12	.000
13	.254	.121	56.731	13	.000
14	-.076	.119	57.139	14	.000
15	-.121	.117	58.211	15	.000
16	-.268	.115	63.617	16	.000

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

在 SPSS 中画出了样本自相关系数图。图中的横轴为滞后期，纵轴为样本自相关系数。图中用条形形状来表示样本自相关系数，并画出了 95%的置信上下限的线条。从下图可以看出该时间序列的自相关系数并不呈负指数收敛到零，其衰减速度比较慢，不是平稳时间序列。



B：样本偏自相关系数以及偏自相关图

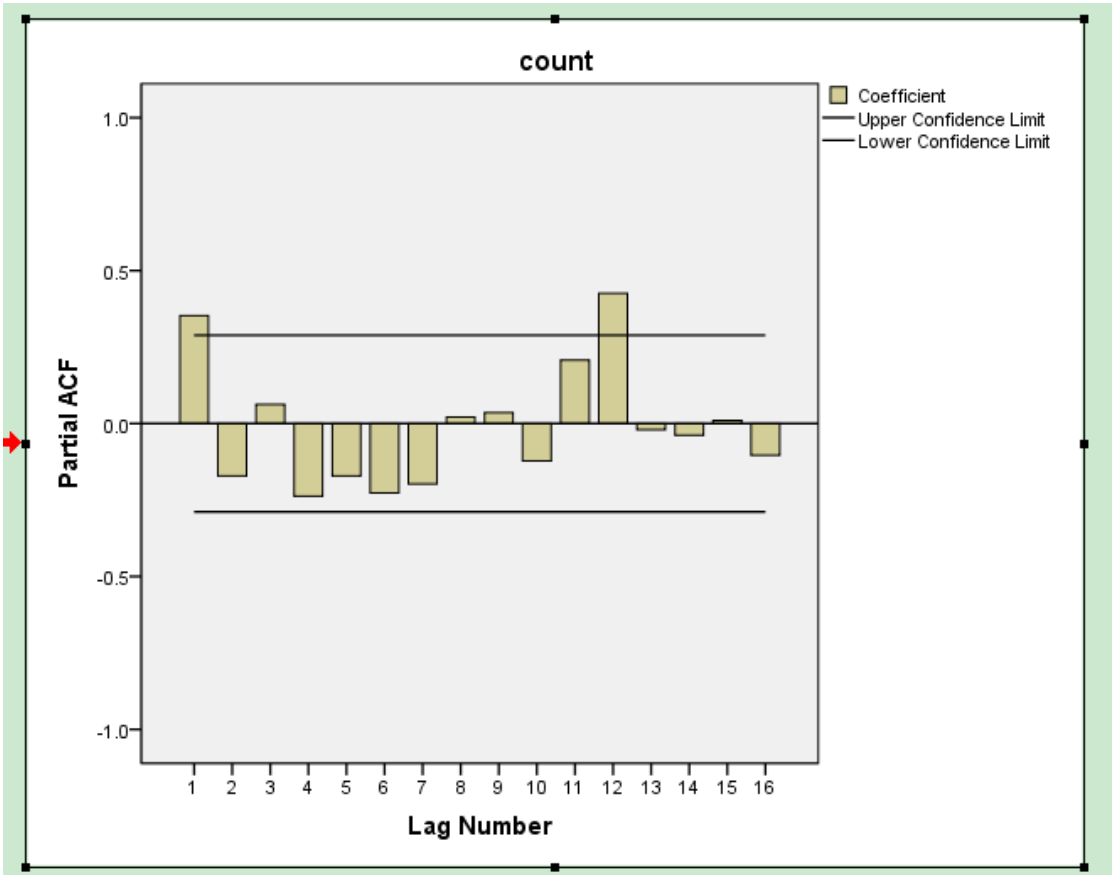
在 SPSS 中给出了不同滞后阶的样本偏相关系数的值，样本偏相关系数的标准误差（标准列）。从下表样本偏相关系数的数据表可以看出该时间序列不是白噪声。

Partial Autocorrelations

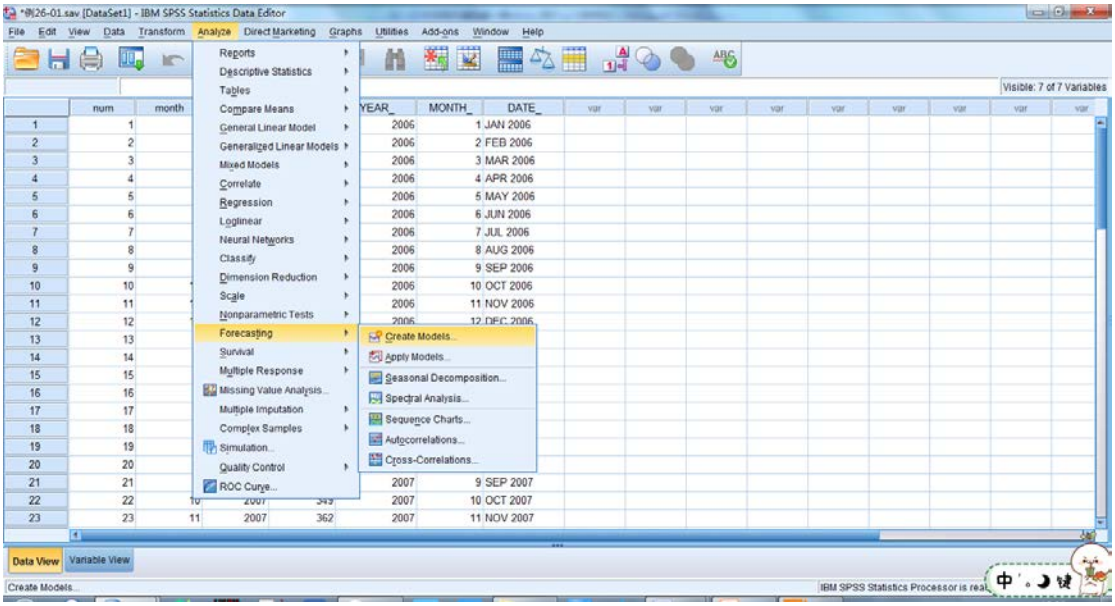
Series: count

Lag	Partial Autocorrelation	Std. Error
1	.353	.144
2	-.171	.144
3	.062	.144
4	-.237	.144
5	-.171	.144
6	-.227	.144
7	-.198	.144
8	.021	.144
9	.036	.144
10	-.122	.144
11	.207	.144
12	.426	.144
13	-.020	.144
14	-.038	.144
15	.010	.144
16	-.104	.144

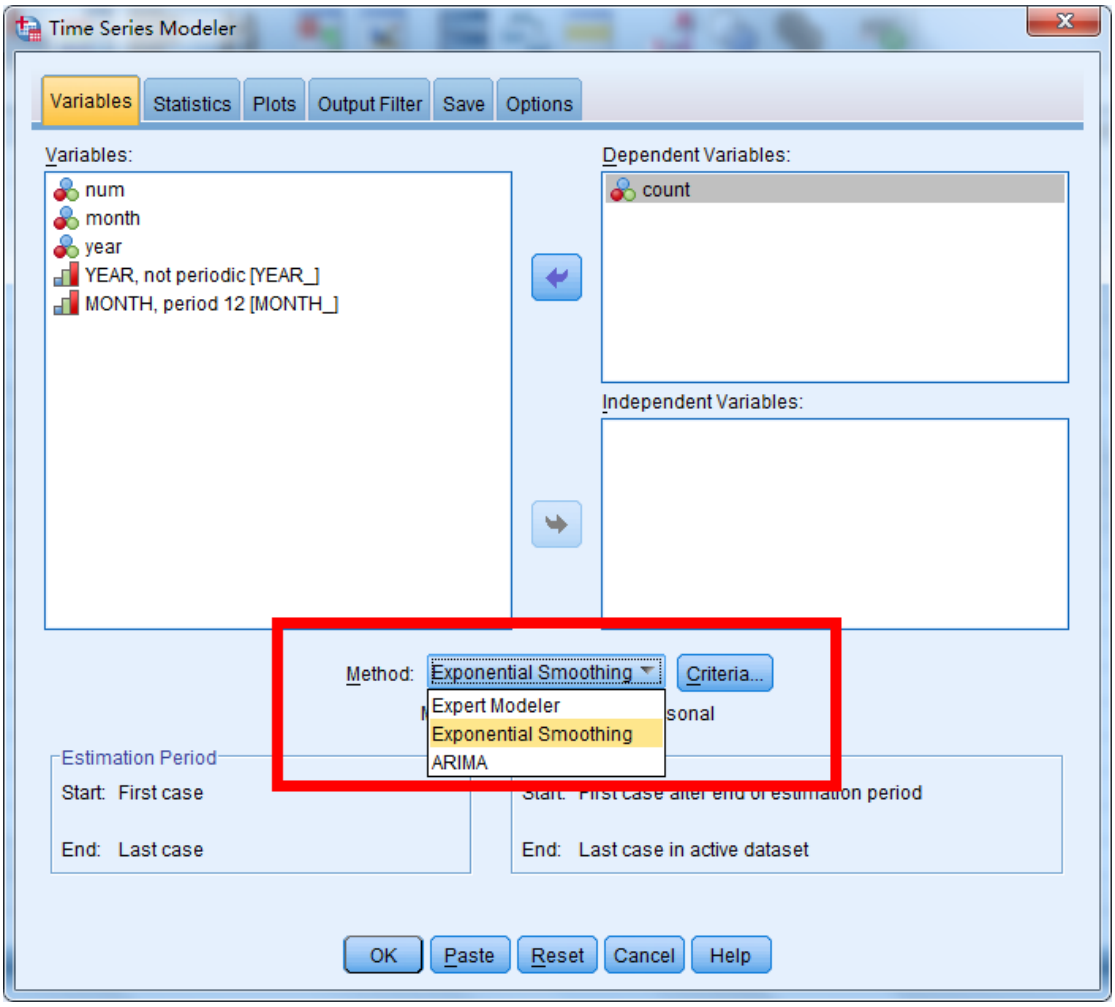
图中的横轴为滞后期，纵轴为样本偏相关系数（PACF）。图中用条形形状来表示样本偏相关系数，并画出了 95%的置信上下限的线条。从下图可以看出该时间序列的偏相关系数在一阶滞后期、12 阶滞后期比较大，说明该时间序列具有周期性，不是平稳时间序列。



Step05：创建模型



SPSS 提供了三大类预测方法：1：专家建模器，2：指数平滑法，3：ARIMA



本次笔记记录的为指数平滑法。

指数平滑法有助于预测存在趋势和/或季节的序列，此处数据同时体现上述两种特征。创建最适当的指数平滑模型包括确定模型类型（此模型是否需要包含趋势和/或季节），然后获取最适合选定模型的参数。

指数平滑法又包含几种模型，下面依次介绍。

指数平滑法模型 (Gardner, 1985) 分为季节性模型和非季节性模型。季节性模型只有在为活动数据集定义了周期时才可用（请参见下文的“当前周期性”）。

- **简单的.** 该模型适用于没有趋势或季节性的序列。其唯一的平滑参数是水平。简单指数平滑法与 ARIMA 模型极为相似，包含零阶自回归、一阶差分、一阶移动平均数，并且没有常数。

- **Holt 线性趋势.** 该模型适用于具有线性趋势并没有季节性的序列。其平滑参数是水平和趋势，不受相互之间的值的约束。Holt 模型比 Brown 模型更通用，但在计算大序列时要花的时间更长。Holt 指数平滑法与 ARIMA 模型极为相似，包含零阶自回归、二阶差分以及二阶移动平均数。

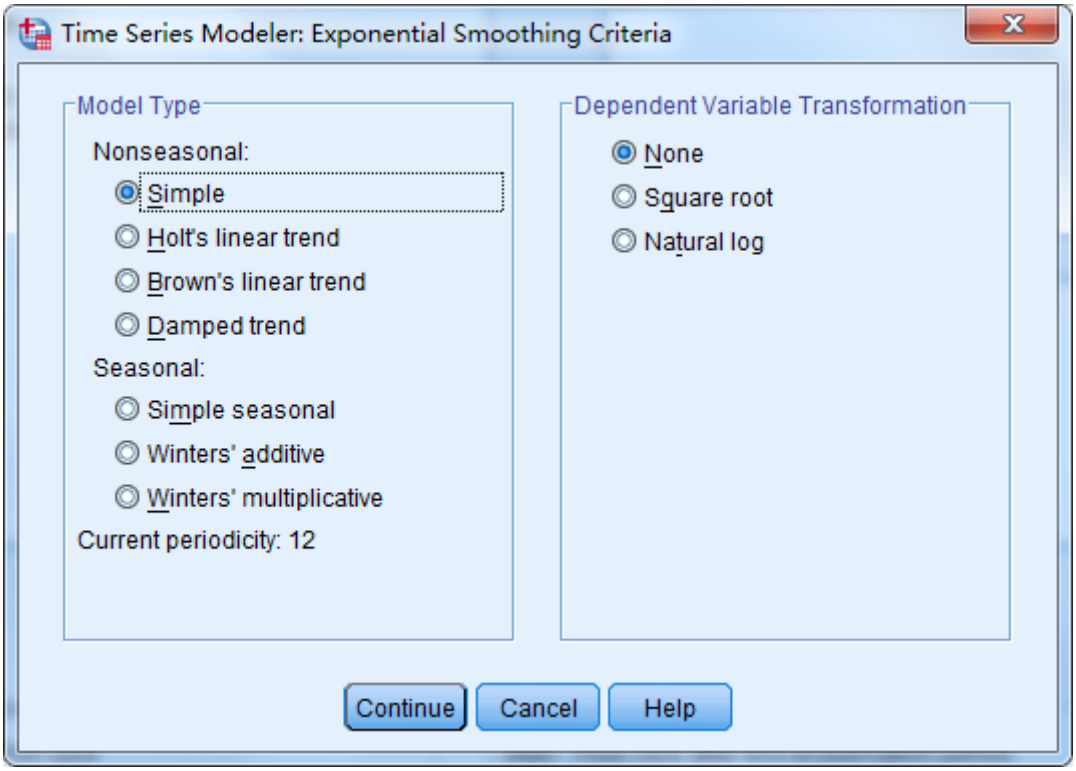
- **Brown 线性趋势.** 该模型适用于具有线性趋势并没有季节性的序列。其平滑参数是水平和趋势，并假定二者等同。因此，Brown 模型是 Holt 模型的特例。Brown 指数平滑法与具有零阶自回归、二阶差分和二阶移动平均的 ARIMA 模型极为相似，且移动平均第二阶的系数等于第一阶的系数二分之一的平方。

- **阻尼趋势.** 此模型适用于具有线性趋势的序列，且该线性趋势正逐渐消失并且没有季节性。其平滑参数是水平、趋势和阻尼趋势。阻尼指数平滑法与具有一阶自回归、一阶差分和二阶移动平均的 ARIMA 模型极为相似。

- **简单季节性.** 该模型适用于没有趋势并且季节性影响随时间变动保持恒定的序列。其平滑参数是水平和季节。简单季节性指数平滑法与 ARIMA 模型极为相似，包含零阶自回归、一阶差分、一阶季节性差分和一阶、p 阶和 p + 1 阶移动平均数，其中 p 是季节性区间中的周期数（对于月数据，p = 12）。

- **Winters 可加的.** 该模型适用于具有线性趋势和不依赖于序列水平的季节性效应的序列。其平滑参数是水平、趋势和季节。Winters 可加的指数平滑法与 ARIMA 模型极为相似，包含零阶自回归、一阶差分、一阶季节差分和 p +1 阶移动平均数，其中 p 是季节性区间中的周期数（对于月数据，p = 12）。

- **Winters 可乘的.** 该模型适用于具有线性趋势和依赖于序列水平的季节性效应的序列。其平滑参数是水平、趋势和季节。Winters 的可乘指数平滑法与任何 ARIMA 模型都不相似。

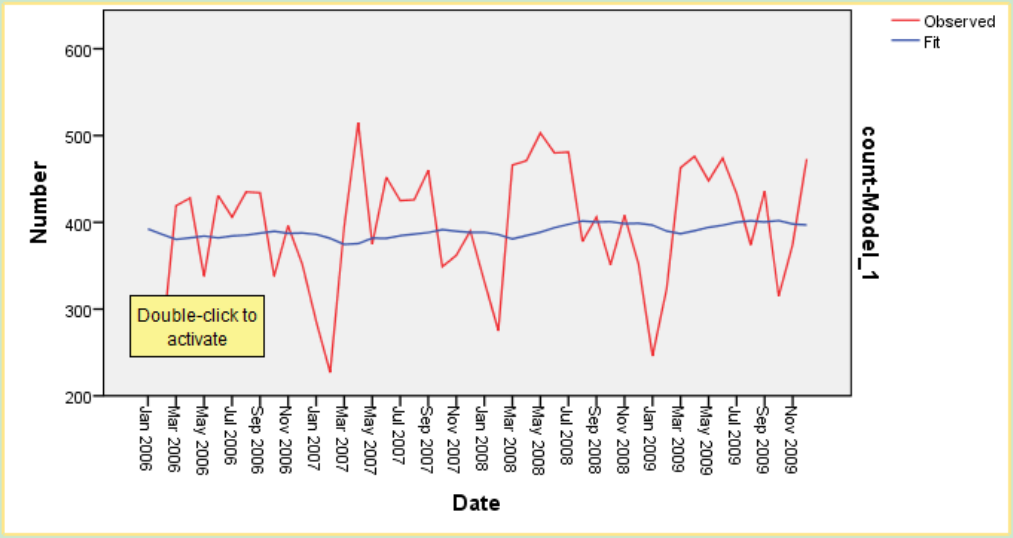


1．简单模型预测（即无趋势也无季节）

首先采用最为简单的建模方法，就是简单模型，这里不断尝试的目的是熟悉各种预测模型，了解模型在什么时候不适合数据，这是成功构建模型的基本技巧。

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	Ljung-Box Q(18): Statistics	Ljung-Box Q(18): DF	Ljung-Box Q(18): Sig.	Number of Outliers
count-Model_1	0	.197	76.232	17	.000	0



Ljung-Box Q 检验可以判断残差已违反白噪声假定，因此拒绝该模型。

从图中我们看到，虽然简单模型既没有考虑季节性变化，也没有周期性呈现，直观的讲基本上与线性预测没有差异。因此也拒绝此模型。

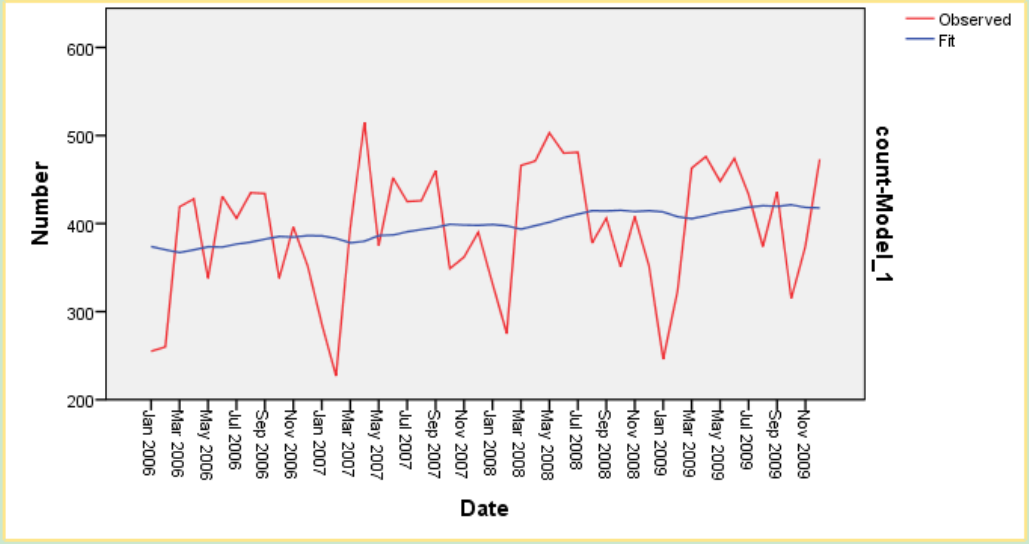
2. Holt 线性趋势预测

Holt 线性指数平滑法，一般选择：针对等级的平滑系数 lapha=0.1,针对趋势的平滑系数 gamma=0.2；

等级：	$L_t = \alpha y_t + (1-\alpha)F_t$
趋势：	$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$
初始值：	$L_2 = y_2$ and $T_2 = y_2 - y_1$

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
count-Model_1	0	.686	80.769	16	.000	0

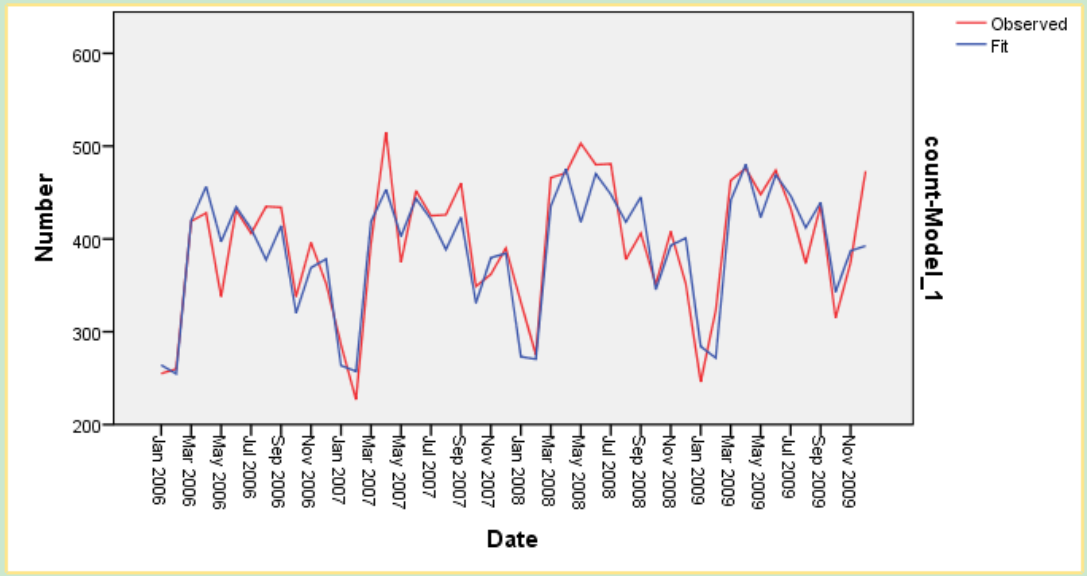


该模型依然不理想。

3．简单季节性模型

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
count-Model_1	0	.746	20.562	16	.196	0

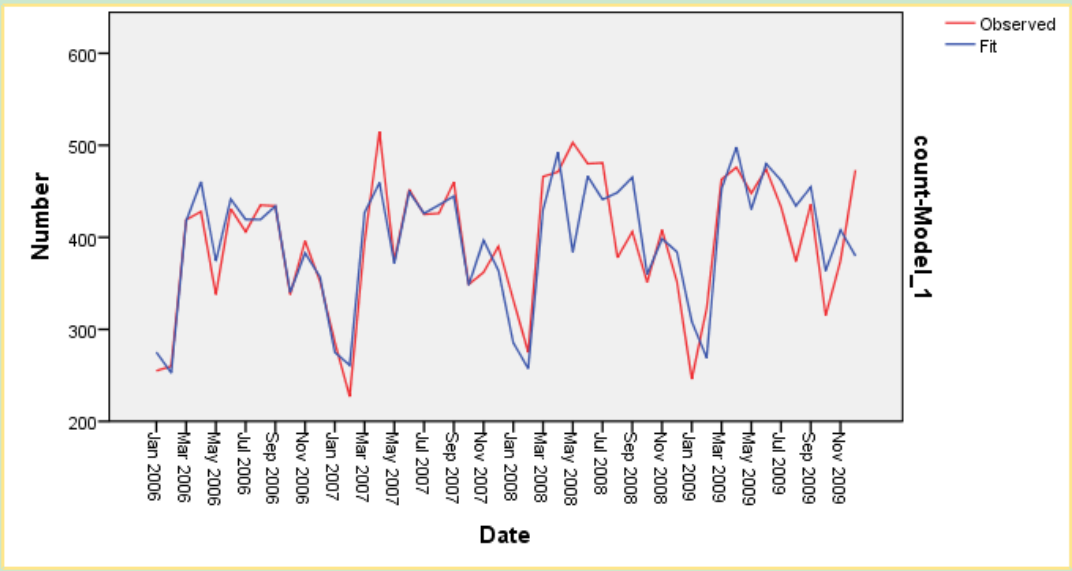


当考虑了季节性变化后，简单季节性预测模型基本上较好的拟合了数据的大趋势。

4. Winters 相乘法预测模型

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
count-Model_1	0	.629	20.022	15	.171	0



此时也说明，无论采用指数平滑的什么模型，只要考虑了季节因素，都可以得到较好结果，不同的季节性指数平滑方法只是细微差异。

不同模型之间优劣性的差异可以通过 AIC 准则、BIC 准则来判断。AIC 适用于自回归模型，而 BIC 准则是一个更通用的标准。

因此在这任意选择一个模型（简单季节模型）进行预测，预测 2010 结核病的人数。

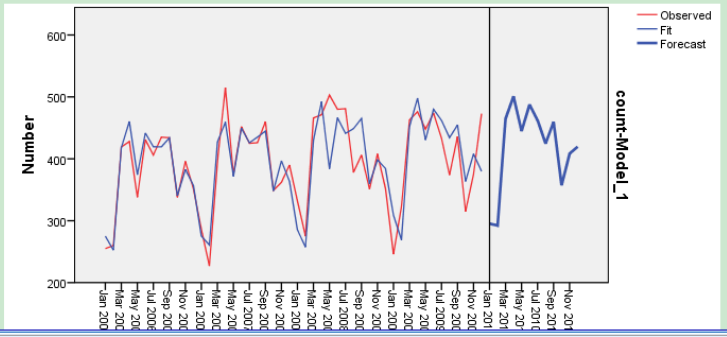
Model Statistics

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Normalized BIC	Statistics	DF	Sig.	
count-Model_1	0	.629	7.529	20.022	15	.171	0

Forecast

Model		Jan 2010	Feb 2010	Mar 2010	Apr 2010	May 2010	Jun 2010	Jul 2010	Aug 2010	Sep 2010	Oct 2010	Nov 2010	Dec 2010
count-Model_1	Forecast	295	292	465	501	445	488	462	425	460	357	408	420
	UCL	372	369	542	578	522	565	539	502	537	434	486	497
	LCL	218	215	388	424	368	410	385	348	382	280	331	342

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.



上图对该模型进行了评价，BIC 为 7.529，通过 Ljung-Box Q 检验可以判断残差未违反白噪声假定，因此接受该模型。

图中也给出了 2010 年全年结核病患病人数预测值，从图中看出，实际值和预测值比较吻合，该模型具有较高的预测精确度。

注：详见张文彤主编的《SPSS 统计分析高级教程(第 2 版)》以及 SPSS 官网帮助文件。