第五讲 时间序列与线性回归

本节内容

一、时间序列

二、线性回归

一、时间序列

时间序列（或称动态数列）是指将同一统计指标的数值按其发生的时间先后顺序排列而成的数列。时间序列分析的主要目的是根据已有的历史数据对未来进行预测。

一、时间序列

- **时间序列根据所研究的依据不同，可有不同的分类。**

1．按所研究的对象的多少分，有一元时间序列和多元时间序列。

2．按时间的连续性可将时间序列分为离散时间序列和连续时间序列两种。

3．按序列的统计特性分，有平稳时间序列和非平稳时间序列。

4．按时间序列的分布规律来分，有高斯型时间序列和非高斯型时间序列。

一、时间序列

- 一个时间序列往往是以下几类变化形式的叠加或耦合。我们常认为一个时间序列可以分解为以下四个成分：长期趋势，季节变动，循环变动，不规则变动。

1）长期趋势（ T ）现象在较长时期内受某种根本性因素作用而形成的总的变动趋势。

2）季节变动（ S ）现象在一年内随着季节的变化而发生的有规律的周期性变动。

3）循环变动（ C ）现象以若干年为周期所呈现出的波浪起伏形态的有规律的变动。

4）不规则变动（I ）是一种无规律可循的变动，包括严格的随机变动和不规则的突发性影响很大的变动两种类型。

一、时间序列

➢ 三种时间序列模型

通常用T_t表示长期趋势项，S_t表示季节变动趋势项，C_t表示循环变动趋势项，R表示随机干扰项。常见的确定性时间序列模型有以下几种类型：

（1）加法模型

$$y_t = T_t + S_t + C_t + R_t$$

（2）乘法模型

$$y_t = T_t \cdot S_t \cdot C_t \cdot R_t$$

（3）混合模型

$$y_t = T_t \cdot S_t + R_t$$

$$y_t = S_t + T_t \cdot C_t \cdot R_t$$

其中y_t是观测目标的观测记录，$E(R_t) = 0$，$E(R_t^2) = \sigma^2$

一、时间序列

1.1 移动平均法

- 移动平均法可以作为一种数据平滑的方式。以每天的气温数据为例，今天的天气可能与过去的十天的气温有线性关系；或者有的人对食物有一种节俭的美德，他们做的饭菜能看出有些是上一顿的，当然也有一部分是今天的做的，再假设隔两顿的都被倒掉了，并且每天都是这样的，那么这碗饭菜可能就是一部分上一顿的再加上一部分今天现做的，这就是一个一阶的移动平均。
- 移动平均法是根据时间序列资料逐渐推移，依次计算包含一定项数的时序平均数，以反映长期趋势的方法。当时间序列的数值由于**受周期变动和不规则变动的影响**，起伏较大，不易显示出发展趋势时，可用**移动平均法，消除这些因素的影响**，分析、预测序列的长期趋势。移动平均法有简单移动平均法，加权移动平均法，趋势移动平均法等。

一、时间序列

（1）简单移动平均法

设观测序列为 $y_1,\cdots,y_T$，取移动平均的项数 $N<T$。一次简单移动平均值计算公式为：

$$M_t^{(1)}=\frac{1}{N}(y_t+y_{t-1}+\cdots+y_{t-N+1})$$

$$=\frac{1}{N}(y_{t-1}+\cdots+y_{t-N})+\frac{1}{N}(y_t-y_{t-N})=M_{t-1}^{(1)}+\frac{1}{N}(y_t-y_{t-N}) \quad (1)$$

当预测目标的基本趋势是在某一水平上下波动时，可用一次简单移动平均方法建立预测模型：

$$\hat{y}_{t+1}=M_t^{(1)}=\frac{1}{N}(\hat{y}_t+\cdots+\hat{y}_{t-N+1})\,,\ t=N,N+1,\cdots, \quad (2)$$

其预测标准误差为：

$$S=\sqrt{\frac{\sum_{t=N+1}^{T}(\hat{y}_t-y_t)^2}{T-N}}\,, \quad (3)$$

简单移动平均法只适合做近期预测，而且是预测目标的发展趋势变化不大的情况。如果目标的发展趋势存在其它的变化，采用简单移动平均法就会产生较大的预测偏差和滞后。

例1　某企业1月～11月份的销售收入时间序列如表1示。试用一次简单滑动平均法预测第 12 月份的销售收入。

表 1　企业销售收入

月份 t	1	2	3	4	5	6
销售收入 y_t	533.8	574.6	606.9	649.8	705.1	772.0
月份 t	7	8	9	10	11	
销售收入 y_t	816.4	892.7	963.9	1015.1	1102.7	

解：　分别取 $N=4, N=5$ 的预测公式

$$\hat{y}_{t+1}^{(1)}=\frac{y_t+y_{t-1}+y_{t-2}+y_{t-3}}{4},\quad t=4,5,\cdots,11$$

$$\hat{y}_{t+1}^{(2)}=\frac{y_t+y_{t-1}+y_{t-2}+y_{t-3}+y_{t-4}}{5},\quad t=5,\cdots,11$$

当$N=4$时，预测值$\hat{y}_{12}^{(1)}=993.6$，预测的标准误差为

$$S_2=\sqrt{\frac{\sum_{t=5}^{11}\left(\hat{y}_t^{(1)}-y_t\right)^2}{11-4}}=150.5$$

当$N=5$时，预测值$\hat{y}_{12}^{(2)}=958.2$，预测的标准误差为

$$S_2=\sqrt{\frac{\sum_{t=6}^{11}\left(\hat{y}_t^{(2)}-y_t\right)^2}{11-5}}=182.4$$

计算表明，$N=4$时，预测的标准误差较小，所以选取$N=4$。预测第12月份的销售金额为993. 6。

一、时间序列

（2）加权移动平均法

在简单移动平均公式中，每期数据在求平均时的作用是等同的。但是，每期数据所包含的信息量不一样，近期数据包含着更多关于未来情况的信息。因此，把各期数据等同看待是不合理的，应考虑各期数据的重要性，对近期数据给予较大的权重，这就是加权移动平均法的基本思想。

设时间序列为 $y_1, y_2, \cdots, y_t, \cdots$；加权移动平均公式为

$$M_{tw} = \frac{w_1 y_t + w_2 y_2 + \cdots + w_N y_{t-N+1}}{w_1 + w_2 + \cdots + w_N}, \quad t \geq N \tag{4}$$

式中 M_{tw} 为 t 期加权移动平均数；w_i 为 y_{t-i+1} 的权数，它体现了相应的 y_t 在加权平均数中的重要性。

利用加权移动平均数来做预测，其预测公式为

$$\hat{y}_{t+1} = M_{tw} \tag{5}$$

即以第 t 期加权移动平均数作为第 $t+1$ 期的预测值。

在加权移动平均法中，W_t 的选择，同样具有一定的经验性。一般的原则是：近期数据的权数大，远期数据的权数小。至于大到什么程度和小到什么程度，则需要按照预测者对序列的了解和分析来确定。

例2　我国1979～1988年原煤产量如表2所示，试用加权移动平均法预测1989年的产量

表 2　我国原煤产量统计数据及加权移动平均预测值表

年份	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
原煤产量 y_t	6.35	6.20	6.22	6.66	7.15	7.89	8.72	8.94	9.28	9.8
三年加权移动平均预测值				6.235	6.4367	6.8317	7.4383	8.1817	8.6917	9.0733
相对误差（%）				6.38	9.98	13.41	14.7	8.48	6.34	7.41

解　取 $w_1=3, w_2=2, w_3=1$，按预测公式

$$\hat{y}_{t+1}=\frac{3y_t+2y_{t-1}+y_{t-2}}{3+2+1}$$

计算三年加权移动平均预测值，其结果列于表 2 中。1989 年我国原煤产量的预测值为（亿吨）

$$\hat{y}_{1989}=\frac{3\times 9.8+2\times 9.28+8.94}{6}=9.48$$

这个预测值偏低，可以修正。其方法是：先计算各年预测值与实际值的相对误差，例如 1982 年为

$$\frac{6.66-6.235}{6.66}=6.38\%$$

将相对误差列于表 2 中，再计算总的平均相对误差。

$$\left(1-\frac{\sum \hat{y}_t}{\sum y_t}\right)\times 100\%=(1-\frac{52.89}{58.44})\times 100\%=9.5\%$$

由于总预测值的平均值比实际值低 9.5%，所以可将 1989 年的预测值修正为

$$\frac{9.48}{1-9.5\%}=10.4788$$

一、时间序列

（3）趋势移动平均法

简单移动平均法和加权移动平均法，在时间序列没有明显的趋势变动时，能够准确反映实际情况。但当时间序列出现直线增加或减少的变动趋势时，用简单移动平均法和加权移动平均法来预测就会出现滞后偏差。因此，需要进行修正，修正的方法是作**二次移动平均**，利用移动平均滞后偏差的规律来建立直线趋势的预测模型。这就是趋势移动平均法。一次移动的平均数为

$$M_t^{(1)} = \frac{1}{N}(y_t + y_{t-1} + \cdots + y_{t-N+1})$$

在一次移动平均的基础上再进行一次移动平均就是二次移动平均，其计算公式为

$$M_t^{(2)} = \frac{1}{N}(M_t^{(1)} + \cdots + M_{t-N+1}^{(1)}) = M_{t-1}^{(2)} + \frac{1}{N}(M_t^{(1)} - M_{t-N}^{(1)}) \qquad (6)$$

下面讨论如何利用移动平均的滞后偏差建立直线趋势预测模型。

设时间序列$\{y_t\}$从某时期开始具有直线趋势，且认为未来时期也按此直线趋势变化，则可设此直线趋势预测模型为

$$\hat{y}_{t+T} = a_t + b_t T\ ,\quad T = 1,2,\cdots \qquad (7)$$

其中t为当前时期数；T为由t至预测期的时期数；a_t为截距；b_t为斜率。两者又称为平滑系数。

现在，我们根据移动平均值来确定平滑系数。由模型（7）可知

$$
\begin{aligned}
a_t &= y_t \\
y_{t-1} &= y_t - b_t \\
y_{t-2} &= y_t - 2b_t \\
&\cdots \\
y_{t-N+1} &= y_t - (N-1)b_t
\end{aligned}
$$

所以

$$
\begin{aligned}
M_t^{(1)} &= \frac{y_t + y_{t-1} + \cdots + y_{t-N+1}}{N} = \frac{y_t + (y_t - b_t) + \cdots + [y_t - (N-1)b_t]}{N} \\
&= \frac{Ny_t - [1 + 2 + \cdots + (N-1)]b_t}{N} = y_t - \frac{N-1}{2} b_t
\end{aligned}
$$

因此

$$
y_t - M_t^{(1)} = \frac{N-1}{2} b_t \tag{8}
$$

由式（7），类似式（8）的推导，可得

$$
y_{t-1} - M_{t-1}^{(1)} = \frac{N-1}{2} b_t \tag{9}
$$

所以

$$
y_t - y_{t-1} = M_t^{(1)} - M_{t-1}^{(1)} = b_t \tag{10}
$$

类似式（8）的推导，可得

$$
M_t^{(1)} - M_t^{(2)} = \frac{N-1}{2} b_t \tag{11}
$$

于是，由式（8）和式（11）可得平滑系数的计算公式为

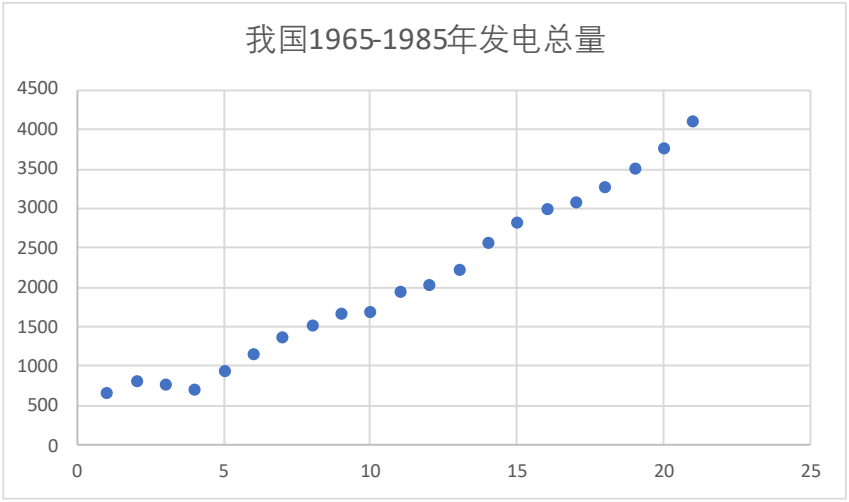
$$
\begin{cases}
a_t = 2M_t^{(1)} - M_t^{(2)} \\
b_t = \dfrac{2}{N-1}(M_t^{(1)} - M_t^{(2)})
\end{cases} \tag{12}
$$

例3　我国1965～1985年的发电总量如表3所示，试预测1986和1987年的发电总量。

表 3　我国发电量及一、二次移动平均值计算表

年份	t	发电总量 y_t	一次移动平均，$N=6$	二次移动平均，$N=6$
1965	1	676		
1966	2	825		
1967	3	774		
1968	4	716		
1969	5	940		
1970	6	1159	848.3	
1971	7	1384	966.3	
1972	8	1524	1082.8	
1973	9	1668	1231.8	
1974	10	1688	1393.8	
1975	11	1958	1563.5	1181.1
1976	12	2031	1708.8	1324.5
1977	13	2234	1850.5	1471.9
1978	14	2566	2024.2	1628.8
1979	15	2820	2216.2	1792.8
1980	16	3006	2435.8	1966.5
1981	17	3093	2625	2143.4
1982	18	3277	2832.7	2330.7
1983	19	3514	3046	2530
1984	20	3770	3246.7	2733.7
1985	21	4107	3461.2	2941.2

由散点图可以看出，发电总量基本呈直线上升趋势。可以用趋势移动平均法来预测。



取$N=6$，分别计算一次和二次移动平均值并列于表中

$$M_{21}^{(1)}=3461.2,\quad M_{21}^{(2)}=2941.2$$

再由公式（12），得：

$$a_{21}=2M_{21}^{(1)}-M_{21}^{(2)}=3981.1$$

$$b_{21}=\frac{2}{6-1}\left(M_{21}^{(1)}-M_{21}^{(2)}\right)=208$$

于是，得到$t=21$时直线趋势的预测模型为

$$\hat{y}_{21+T}=3981.1+208T$$

预测1986和1987年的发电总量为

$$\hat{y}_{1986}=\hat{y}_{22}=4189.1\ ,\ \hat{y}_{1987}=\hat{y}_{23}=4397.1$$

一、时间序列

1.2 指数平滑法

- 一次移动平均实际上认为近 N 期数据对未来值影响相同，都加权1/N；而 N 期以前的数据对未来值没有影响，加权为0。但是，二次及更高次移动平均数的权数却不是1/N，且次数越高，权数的结构越复杂，但永远保持对称的权数，即两端项权数小，中间项权数大，不符合一般系统的动态性。一般说来历史数据对未来值的影响是随时间间隔的增长而递减的。所以，更切合实际的方法应是对各期观测值依时间顺序进行加权平均作为预测值。指数平滑法可满足这一要求，而且具有简单的递推形式。
- 指数平滑法根据平滑次数的不同，又分为一次指数平滑法、二次指数平滑法和三次指数平滑法等。

（1）一次指数平滑法

设时间序列为 $y_1, y_2, \cdots, y_t, \cdots$，$\alpha$ 为加权系数，$0 < \alpha < 1$，一次指数平滑公式为：

$$S_t^{(1)} = \alpha y_t + (1-\alpha) S_{t-1}^{(1)} = S_{t-1}^{(1)} + \alpha (y_t - S_{t-1}^{(1)}) \tag{13}$$

式（13）是由移动平均公式改进而来的。由式（1）知，移动平均数的递推公式为

$$M_t^{(1)} = M_{t-1}^{(1)} + \frac{y_t - y_{t-N}}{N}$$

以 $M_{t-1}^{(1)}$ 作为 y_{t-N} 的最佳估计，则有

$$M_t^{(1)} = M_{t-1}^{(1)} + \frac{y_t - M_{t-1}^{(1)}}{N} = \frac{y_t}{N} + \left(1 - \frac{1}{N}\right) M_{t-1}^{(1)}$$

令 $\alpha = \frac{1}{N}$，以 S_t 代替 $M_t^{(1)}$，即得式（13）

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

为进一步理解指数平滑的实质，把式（13）依次展开，有

$$S_t^{(1)} = \alpha y_t + (1-\alpha)[\alpha y_{t-1} + (1-\alpha) S_{t-2}^{(1)}] = \cdots = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j y_{t-j} \tag{14}$$

（14）式表明 $S_t^{(1)}$ 是全部历史数据的加权平均，加权系数分别为 $\alpha, \alpha(1-\alpha), \alpha(1-\alpha)^2, \cdots$；显然有

$$\sum_{j=0}^{\infty} \alpha (1-\alpha)^j = \frac{\alpha}{1-(1-\alpha)} = 1$$

由于加权系数符合指数规律，又具有平滑数据的功能，故称为指数平滑。

以这种平滑值进行预测，就是一次指数平滑法。预测模型为

$$\hat{y}_{t+1} = S_t^{(1)}$$

即

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha) \hat{y}_t \tag{15}$$

也就是以第 t 期指数平滑值作为 $t+1$ 期预测值。

一、时间序列

- **加权系数的选择**

在进行指数平滑时，加权系数的选择是很重要的。由式（15）可以看出，α 的大小规定了在新预测值中新数据和原预测值所占的比重。α 值越大，新数据所占的比重就愈大，原预测值所占的比重就愈小，反之亦然。若把式（15）改写为

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t) \qquad (16)$$

则从上式可看出，新预测值是根据预测误差对原预测值进行修正而得到的。α 的大小则体现了修正的幅度，α 值愈大，修正幅度愈大；α 值愈小，修正幅度也愈小。

若选取 $\alpha = 0$，则 $\hat{y}_{t+1} = \hat{y}_t$，即下期预测值就等于本期预测值，在预测过程中不考虑任何新信息；若选取 $\alpha = 1$，则 $\hat{y}_{t+1} = y_t$，即下期预测值就等于本期观测值，完全不相信过去的信息。这两种极端情况很难做出正确的预测。因此，α 值应根据时间序列的具体性质在 0～1 之间选择。具体如何选择一般可遵循下列原则：①如果时间序列波动不大，比较平稳，则 α 应取小一点，如（0.1～0.5）。以减少修正幅度，使预测模型能包含较长时间序列的信息；②如果时间序列具有迅速且明显的变动倾向，则 α 应取大一点，如（0.6～0.8）。使预测模型灵敏度高一些，以便迅速跟上数据的变化。

在实用上，类似移动平均法，多取几个 α 值进行试算，看哪个预测误差小，就采用哪个。

一、时间序列

- **初始值的确定**

用一次指数平滑法进行预测，除了选择合适的 α 外，还要确定初始值 $s_0^{(1)}$。初始值是由预测者估计或指定的。当时间序列的数据较多，比如在 20 个以上时，初始值对以后的预测值影响很少，可选用第一期数据为初始值。如果时间序列的数据较少，在 20 个以下时，初始值对以后的预测值影响很大，这时，就必须认真研究如何正确确定初始值。一般以最初几期实际值的平均值作为初始值。

例4　某市1976～1987年某种电器销售额如表4所示。试预测1988年该电器销售额。

解　采用指数平滑法，并分别取 $\alpha = 0.2, 0.5$ 和 0.8 进行计算，初始值

$$S_0^{(1)} = \frac{y_1 + y_2}{2} = 51$$

即

$$\hat{y}_1 = S_0^{(1)} = 51$$

按预测模型

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t$$

计算各期预测值，列于表 4 中。

表 4　某种电器销售额及指数平滑预测值计算表　（单位：万元）

年份	t	实际销售额 y_t	预测值 $\hat{y}_t$ $\alpha = 0.2$	预测值 $\hat{y}_t$ $\alpha = 0.5$	预测值 $\hat{y}_t$ $\alpha = 0.8$
1976	1	50	51	51	51
1977	2	52	50.8	50.5	50.2
1978	3	47	51.04	51.25	51.64
1979	4	51	50.23	49.13	47.93
1980	5	49	50.39	50.06	50.39
1981	6	48	50.11	49.53	49.28
1982	7	51	49.69	48.77	48.26
1983	8	40	49.95	49.88	50.45
1984	9	48	47.96	44.94	42.09
1985	10	52	47.97	46.47	46.82
1986	11	51	48.77	49.24	50.96
1987	12	59	49.22	50.12	50.99

从表 4 可以看出，$\alpha = 0.2, 0.5$ 和 0.8 时，预测值是很不相同的。究竟 α 取何值为好，可通过计算它们的预测标准误差 S，选取使 S 较小的那个 α 值。预测的标准误差见表 5。

表 5　预测的标准误差

α	0.2	0.5	0.8
S	4.5029	4.5908	4.8426

计算结果表明：$\alpha = 0.2$ 时，S 较小，故选取 $\alpha = 0.2$，预测 1988 年该电器销售额为 $\hat{y}_{1988} = 51.1754$。

一、时间序列

（2）二次指数平滑法

一次指数平滑法虽然克服了移动平均法的缺点。但当时间序列的变动出现直线趋势时，用一次指数平滑法进行预测，仍存在明显的滞后偏差。因此，也必须加以修正。修正的方法与趋势移动平均法相同，即再作二次指数平滑，利用滞后偏差的规律建立直线趋势模型。这就是二次指数平滑法。其计算公式为

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

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \qquad (17)$$

式中 $S_t^{(1)}$ 为一次指数的平滑值； $S_t^{(2)}$ 为二次指数的平滑值。当时间序列 $\{y_t\}$，从某时期开始具有直线趋势时，类似趋势移动平均法，可用直线趋势模型

$$\hat{y}_{t+T} = a_t + b_t T \ , \ T = 1,2,\cdots \qquad (18)$$

$$\begin{cases} a_t = 2S_t^{(1)} - S_t^{(2)} \\ b_t = \dfrac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)}) \end{cases} \qquad (19)$$

进行预测。

例5 我国1965～1985年的发电总量如表3所示，试预测1986和1987年的发电总量。

年份	t	发电总量	一次平滑值	二次平滑值	yt+1的估计值(a=0.3)
1965	1	676	676	676	
1966	2	825	720.7	689.41	
1967	3	774	736.69	703.594	
1968	4	716	730.483	711.6607	
1969	5	940	793.3381	736.16392	
1970	6	1159	903.03667	786.225745	
1971	7	1384	1047.32567	864.5557222	
1972	8	1524	1190.32797	962.287396	
1973	9	1668	1333.62958	1073.690051	
1974	10	1688	1439.9407	1183.565247	
1975	11	1958	1595.35849	1307.103221	
1976	12	2031	1726.05095	1432.787538	
1977	13	2234	1878.43566	1566.481975	
1978	14	2566	2084.70496	1721.948872	
1979	15	2820	2305.29347	1896.952252	
1980	16	3006	2515.50543	2082.518206	
1981	17	3093	2688.7538	2264.388885	
1982	18	3277	2865.22766	2444.640518	
1983	19	3514	3059.85936	2629.206172	
1984	20	3770	3272.90155	2822.314786	
1985	21	4107	3523.13109	3032.559677	
1986					4223.94739
1987					4434.19228

取$\alpha = 0.3$，分别计算一次和二次指数平滑值

$$S_{21}^{(1)} = 3523.1, \quad S_{21}^{(2)} = 3032.6$$

再由公式（12），得：

$$a_{21} = 2S_{21}^{(1)} - S_{21}^{(2)} = 4013.7$$

$$b_{21} = \frac{0.3}{1-0.3}\left(S_{21}^{(1)} - S_{21}^{(2)}\right) = 210.2$$

于是，得到$t = 21$时直线趋势的预测模型为

$$\hat{y}_{21+T} = 4013.7 + 210.2T$$

预测1986和1987年的发电总量为

$$\hat{y}_{1986} = \hat{y}_{22} = 4223.9 \text{ ，} \hat{y}_{1987} = \hat{y}_{23} = 4434.2$$

一、时间序列

(3) 三次指数平滑法

当时间序列的变动表现为二次曲线趋势时，则需要用三次指数平滑法。三次指数平滑是在二次指数平滑的基础上，再进行一次平滑，其计算公式为

$$\begin{cases} S_t^{(1)} = \alpha y_t + (1-\alpha) S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \\ S_t^{(3)} = \alpha S_t^{(2)} + (1-\alpha) S_{t-1}^{(3)} \end{cases} \tag{21}$$

式中 $S_t^{(3)}$ 为三次指数平滑值。

三次指数平滑法的预测模型为

$$\hat{y}_{t+T} = a_t + b_t T + C_t T^2 \,,\ T = 1,2,\cdots \tag{22}$$

其中

$$\begin{cases} a_t = 3S_t^{(1)} - 3S_t^{(2)} + S_t^{(3)} \\ b_t = \dfrac{\alpha}{2(1-\alpha)^2}\left[(6-5\alpha)S_t^{(1)} - 2(5-4\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}\right] \\ c_t = \dfrac{\alpha^2}{2(1-\alpha)^2}\left[S_t^{(1)} - 2S_t^{(2)} + S_t^{(3)}\right] \end{cases} \tag{23}$$

一、时间序列

习题：我国 1974～1981 年布的产量如表 11 所示。

表 11　1974～1981 年布的产量

年份	1974	1975	1976	1977	1978	1979	1980	1981
产量(亿米)	80.8	94.0	88.4	101.5	110.3	121.5	134.7	142.7

（1）试用趋势移动平均法（取 $N=3$），建立布的年产量预测模型。

（2）分别取 $\alpha=0.3$，$\alpha=0.6$，$S_0^{(1)}=S_0^{(2)}=\dfrac{y_1+y_2+y_3}{3}=87.7$，建立布的直线指数平滑预测模型。

（3）计算模型拟合误差，比较 3 个模型的优劣。

（4）用最优的模型预测 1982 年和 1985 年布的产量。

（1）取预测公式

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + y_{t-2}}{3}, t = 3,4,\cdots,11,$$

其中 $y_t(t=9,10,11)$ 分别取为递推预测的预测值 $\hat{y}_t$。

预测的标准误差

$$S_1 = \sqrt{\frac{\sum_{t=4}^{8}(\hat{y}_t - y_t)^2}{8-3}} = 19.3542.$$

（2）二次指数平滑法的计算公式为

$$\begin{cases} S_t^{(1)} = \alpha y_t + (1-\alpha) S_{t-1}^{(1)}, \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)}. \end{cases}$$

式中：$S_t^{(1)}$ 为一次指数的平滑值；$S_t^{(2)}$ 为二次指数的平滑值。

当时间序列 $\{y_t\}$ 从某时期开始具有直线趋势时，可用直线趋势模型

$$\hat{y}_{t+m} = a_t + b_t m, m = 1,2,\cdots,$$

$$\begin{cases} a_t = 2S_t^{(1)} - S_t^{(2)}, \\ b_t = \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)}). \end{cases}$$

进行预测。

$\alpha=0.3$ 时，预测的标准误差

$$S_2 = \sqrt{\frac{\sum_{t=2}^{8}(\hat{y}_t - y_t)^2}{8-1}} = 11.7966.$$

当 $\alpha=0.6$，预测的标准误差 $S_3=7.0136$。

（3）从预测标准误差的角度考虑，选择 $\alpha=0.6$ 时的二次指数平滑模型。

（4）利用 $\alpha=0.6$ 时的二次指数平滑模型，得到 1982 年和 1985 年的产量预测值分别为 152.9452 亿米，182.8625 亿米。

一、时间序列

1.3 差分指数平滑法

在前面我们已经讲过，当时间序列的变动具有直线趋势时，用一次指数平滑法会出现滞后偏差，其原因在于数据不满足模型要求。因此，我们也可以从数据变换的角度来考虑改进措施，即在运用指数平滑法以前先对数据作一些技术上的处理，使之能适合于一次指数平滑模型，以后再对输出结果作技术上的返回处理，使之恢复为原变量的形态。差分方法是改变数据变动趋势的简易方法。

一、时间序列

(1) 一阶差分指数平滑法

当时间序列呈直线增加时，可运用一阶差分指数平滑模型来预测。其公式如下：

$$\nabla y_t = y_t - y_{t-1} \tag{25}$$

$$\nabla \hat{y}_{t+1} = \alpha \nabla y_t + (1-\alpha)\nabla \hat{y}_t \tag{26}$$

$$\hat{y}_{t+1} = \nabla \hat{y}_{t+1} + y_t \tag{27}$$

其中的 ∇ 为差分记号。式（25）表示对呈现直线增加的序列作一阶差分，构成一个平稳的新序列；式（26）表示把经过一阶差分后的新序列的指数平滑预测值与变量当前的实际值迭加，作为变量下一期的预测值。对于这个公式的数学意义可作如下的解释。

因为

$$y_{t+1} = y_{t+1} - y_t + y_t = \nabla y_{t+1} + y_t \tag{28}$$

当我们采用按式（26）计算的预测值去估计式（28）中的 y_{t+1}，从而式（28）等号左边的 y_{t+1} 也要改为预测值，亦即成为式（27）。

在前面我们已分析过，指数平滑值实际上是一种加权平均数。因此把序列中逐期增量的加权平均数（指数平滑值）加上当前值的实际数进行预测，比一次指数平滑法只用变量以往取值的加权平均数作为下一期的预测更合理。

(1) 一阶差分指数平滑法

例6　某工业企业 1977～1986 年锅炉燃料消耗量资料如表8所示，试预测1987年的燃料消耗量。

表 8　某企业锅炉燃料消耗量的差分指数平滑法计算表（$\alpha = 0.4$）（单位：百吨）

年份	t	燃料消耗量 y_t	差分	差分指数平滑值	预测值
1977	1	24			
1978	2	26	2		
1979	3	27	1	2	28
1980	4	30	3	1.6	28.6
1981	5	32	2	2.16	32.16
1982	6	33	1	2.10	34.10
1983	7	36	3	1.66	34.66
1984	8	40	4	2.19	38.19
1985	9	41	1	2.92	42.92
1986	10	44	3	2.15	43.15
1987	11			2.49	46.49

解：　由资料可以看出，燃料消耗量，除个别年份外，逐期增长量大体在 200 吨左右，即呈直线增长，因此可用一阶差分指数平滑模型来预测。我们取 α = 0.4，初始值为新序列首项值，计算结果列于表 8 中。预测 1987 年燃料消耗量为

$$\hat{y}_{1987} = 2.49 + 44 = 46.49 \text{（百吨）}$$

一、时间序列

(2) 二阶差分指数平滑法

当时间序列呈现二次曲线增长时，可用二阶差分指数平滑模型来预测，计算公式如下：

$$\nabla y_t = y_t - y_{t-1} \tag{29}$$

$$\nabla^2 y_t = \nabla y_t - \nabla y_{t-1} \tag{30}$$

$$\nabla^2 \hat{y}_{t+1} = \alpha \nabla^2 y_t + (1-\alpha)\nabla^2 \hat{y}_t \tag{31}$$

$$\hat{y}_{t+1} = \nabla^2 \hat{y}_{t+1} + \nabla y_t + y_t \tag{32}$$

其中 ∇^2 表示二阶差分。

因为

$$y_{t+1} = y_{t+1} - y_t + y_t = \nabla y_{t+1} + y_t = (\nabla y_{t+1} - \nabla y_t) + \nabla y_t + y_t = \nabla^2 y_{t+1} + \nabla y_t + y_t$$

同样，用 $\nabla^2 y_{t+1}$ 的估计值代替 $\nabla^2 y_{t+1}$ 得到式（32）。

差分方法和指数平滑法的联合运用，除了能克服一次指数平滑法的滞后偏差之外，对初始值的问题也有显著的改进。因为数据经过差分处理后，所产生的新序列基本上是平稳的。这时，初始值取新序列的第一期数据对于未来预测值不会有多大影响。其次，它拓展了指数平滑法的适用范围，使一些原来需要运用配合直线趋势模型处理的情况可用这种组合模型来取代。但是，指数平滑法存在的加权系数α的选择问题，以及只能逐期预测问题，差分指数平滑模型也没有改进。

一、时间序列

1.4 平稳时间序列模型

在时间序列分析中，我们关注平稳时间序列。平稳时间序列有几个关键的特点：

（1）时间序列的均值是一个常数，而不是随时间变化的一个值。下图左侧的时间序列是满足这个条件的，但是右侧的时间序列则不满足这个条件，因为它的均值随时间变化。

（2）时间序列的方差，不能是时间的一个函数。这个性质称为方差齐性(homoscedasticity)。下图中左侧的时间序列符合这个特点，右侧的时间序列则不符合这个特点。

一、时间序列

1.4 平稳时间序列模型

在时间序列分析中，我们关注平稳时间序列。平稳时间序列有几个关键的特点：

（3）第i数据点和第i+m数据点之间的协方差，不能是时间的一个函数。下图中左侧的时间序列符合这个特点，但是右侧的时间序列则随着时间的延伸，频率不断上升，意味着第i和第i+m数据点的协方差发生变化了，也就是它的协方差不是一个常数。

一、时间序列

1.4 平稳时间序列模型

白噪声模型：

高斯白噪声，是平稳的随机过程。它的均值是0，方差是σ^2。下图展示了一个白噪声时间序列。

一、时间序列

1.4 平稳时间序列模型

自相关（Autocorrelation）：

自相关，也称为序列相关。它所表达的，是一个信号和自身的一个延迟的拷贝之间的相关性。延迟作为变量，自相关性作为因变量，便是自相关函数（Auto Correlation Function，ACF）。

换句话来讲，它表达了观测值序列与其自身经过某些阶数滞后形成的序列之间，存在某种程度的相似性，该相似性是观察值的一定时间间隔（Time Lag）的函数。

表 1-2 自相关系数的计算

时间间隔	观察值					
原来的时间序列	X_{t-5}	X_{t-4}	X_{t-3}	X_{t-2}	X_{t-1}	X_t
0	X_{t-5}	X_{t-4}	X_{t-3}	X_{t-2}	X_{t-1}	X_t
1	…	X_{t-5}	X_{t-4}	X_{t-3}	X_{t-2}	X_{t-1}
2	…	…	X_{t-5}	X_{t-4}	X_{t-3}	X_{t-2}
…						

一、时间序列

1.4 平稳时间序列模型

自相关（Autocorrelation）：

相关系数计算公式：

$$\mathrm{correl}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{\mu}_1)(y_i-\bar{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{\mu}_1)^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{\mu}_2)^2}}$$

那么，自相关系数的计算公式为：

$$auto_{correl}(x,y)=\frac{\sum_{i=1}^{n-h}(x_i-\bar{\mu})(x_{i+h}-\bar{\mu})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{\mu})^2}\sqrt{\sum_{i=1}^{n}(x_i-\bar{\mu})^2}}$$

$$=\frac{\sum_{i=1}^{n-h}(x_i-\bar{\mu})(x_{i+h}-\bar{\mu})}{\sum_{i=1}^{n}(x_i-\bar{\mu})^2}$$

一、时间序列

1.4 平稳时间序列模型

偏自相关（Partial Autocorrelation）：

偏相关分析也称净相关分析，它在控制其它变量的线性影响的条件下，分析两变量间的线性相关性，所采用的工具是偏相关系数(净相关系数)。控制变量个数为一时，偏相关系数称为一阶偏相关系数；控制变量个数为二时，偏相关系数称为二阶偏相关系数；控制变量个数为为零时，偏相关系数称为零阶偏相关系数，也就是相关系数。时间序列和自身的偏相关系数，就是偏自相关系数。

形式化地定义，所谓滞后k的偏自相关系数，就是指在给定中间 k-1个随机变量 x_{t-1}，x_{t-2}，…，x_{t-k+1}的条件下，或者说剔除了中间 k-1个随机变量的干扰之后，x_{t-k}对 x_t的影响的相关度量，即

$$\rho_{x_t,x_{t-k}|x_{t-1},\ldots,x_{t-k+1}} = \frac{E[(x_t - \hat{E}x_t)[(x_{t-k} - \hat{E}x_{t-k})]]}{E[(x_t - \hat{E}x_t)^2]}$$

其中，$\hat{E}x_t = \mathrm{E}[x_t\,|x_{t-1},\ldots,x_{t-k+1}]$, $\hat{E}x_{t-k} = \mathrm{E}[x_{t-k}\,|x_{t-1},\ldots,x_{t-k+1}]$

一、时间序列

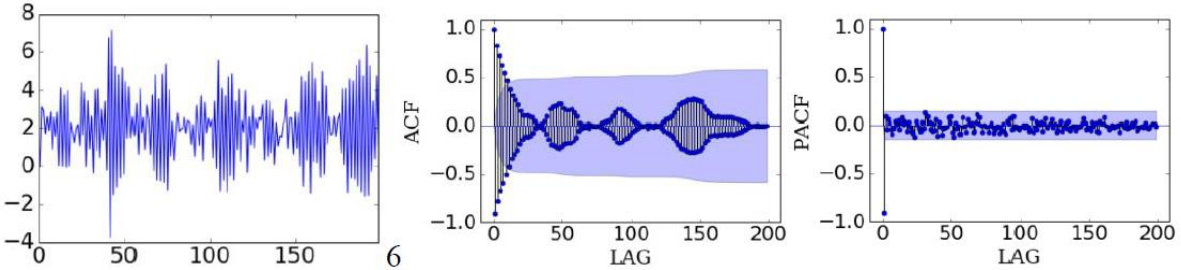
1.4 平稳时间序列模型

（1）自回归(Auto Regression)模型：

- AR模型是利用若干前期时刻的随机变量的线性组合来描述当前时刻随机变量的线性回归模型。描述时间序列 $\{x_t\}$某一个时刻t和前p个时刻序列的数据点的值之间关系的公式为：

$$x_t = \varepsilon_t + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p}$$

- 其中，随机序列$\{\varepsilon_t\}$是白噪声，和以前时刻的x_t (k < t)不相关，该模型称为p阶自回归模型，记为AR(p)。
- AR模型的特点为：(1)如果AR模型满足平稳性条件，则它的均值为0;(2) AR模型的自相关系数是呈复指数衰减，具有拖尾性;(3)AR模型的偏自相关系数有截尾性。第二和第三个特点，可以帮助我们做模型的识别。



(a) Time Series (b) ACF (c)PACF

一、时间序列

1.4 平稳时间序列模型

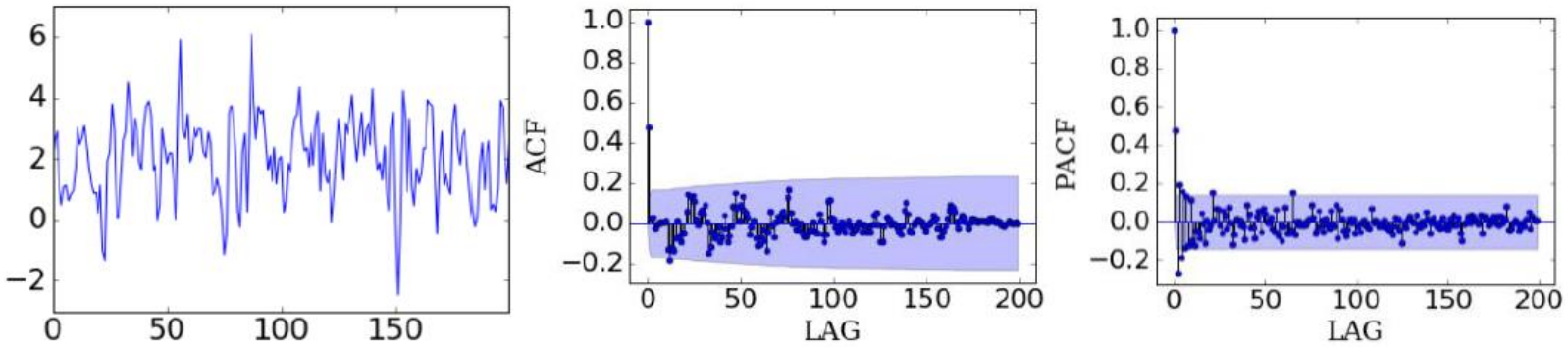
（2）移动平均（Moving Average)模型：

- 在序列$\{x_t\}$中，x_t表示为若干个白噪声的加权和，即：

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$$

- 其中随机序列$\{\varepsilon_t\}$是白噪声，该模型称为q阶移动平均模型，记为MA(q)。

- MA模型的特点为：(1)自相关系数q阶截尾；(2)偏自相关系数拖尾。也就是时间序列的自相关系数是截尾的，可用于对模型的识别。



(a) Time Series (b) ACF (c)PACF

一、时间序列

1.4 平稳时间序列模型

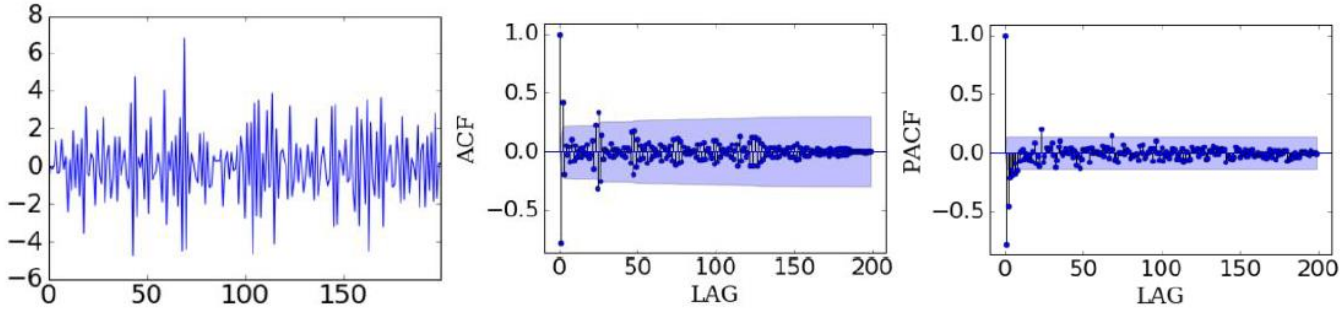
（3）自回归移动平均(ARMA)模型：

- ARMA模型是 AR模型和 MA模型的综合，它的形式为：

$$x_t = \varepsilon_t + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

- 记为ARMA(p，q)。AR模型和MA模型实际上是ARMA模型的特殊形式，即AR(p) = ARMA(p，0)，而MA(q)=ARMA(0，q)。

- ARMA模型兼具AR模型和MA模型的特点，它的自相关系数和偏自相关系数都具有拖尾性，这个特点也可以用于对模型进行识别。



(a) Time Series　　(b) ACF　　(c)PACF

一、时间序列

1.4 平稳时间序列模型

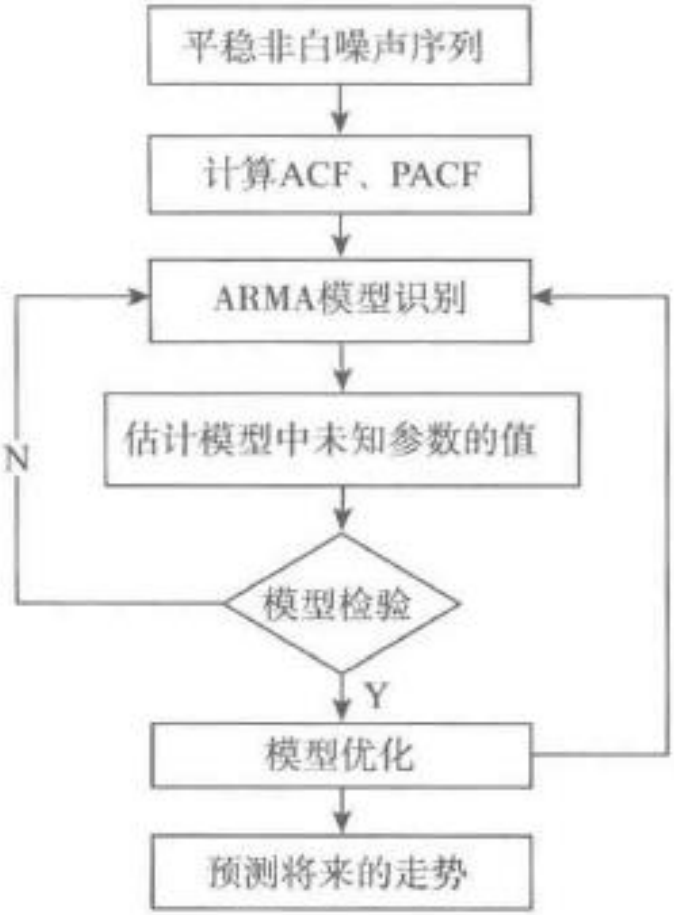
- 根据上述AR、MA、ARMA等模型的特点，我们可以通过ACF和PACF分析，来判断某个时间序列是纯自回归的、或纯移动平均类型的，或是这两者的一种混合体。对纯粹的AR模型或者MA模型可以定阶。

- 如果判别某个过程为ARMA过程，不能马上定阶，还需要进一步处理，需要赤池信息量准则（Akaike Information Criterion)或者贝叶斯信息量准则(Bayesian Information），权衡模型的复杂度和模型拟合数据的优良性，实现定阶。

自相关函数	偏自相关函数	模型定阶
拖尾	p 阶截尾	AR(p)模型
q 阶截尾	拖尾	MA(q)模型
拖尾	拖尾	ARMA(p, q)模型

一、时间序列

1.4 平稳时间序列建模步骤



一、时间序列

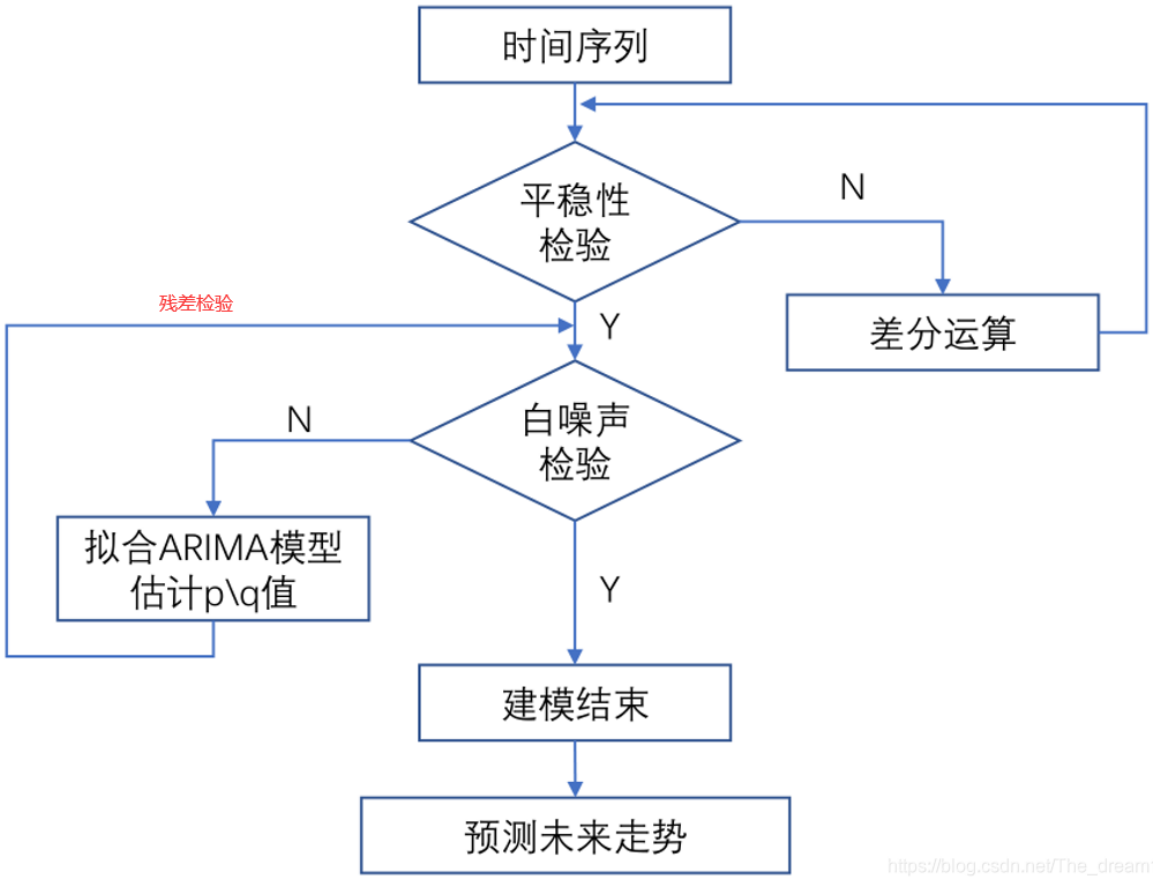
1.5 差分自回归移动平均（ARIMA)模型

- ARIMA模型试图通过数据的自相关性和差分的方式，提取出隐藏在数据背后的时间序列模式，然后用这些模式来预测未来的数据。其中：
  - AR部分用于处理时间序列的自回归部分，它考虑了过去若干时期的观测值对当前值的影响。
  - I部分用于使非平稳时间序列达到平稳，通过一阶或者二阶等差分处理，消除了时间序列中的趋势和季节性因素。
  - MA部分用于处理时间序列的移动平均部分，它考虑了过去的预测误差对当前值的影响。
- 差分自回归移动平均模型记为ARIMA（p、d、q），其中d是需要对数据进行差分的阶数。

一、时间序列

1.5 差分自回归移动平均（ARIMA)模型

- 生成 ARIMA 模型的基本步骤：



二、线性回归

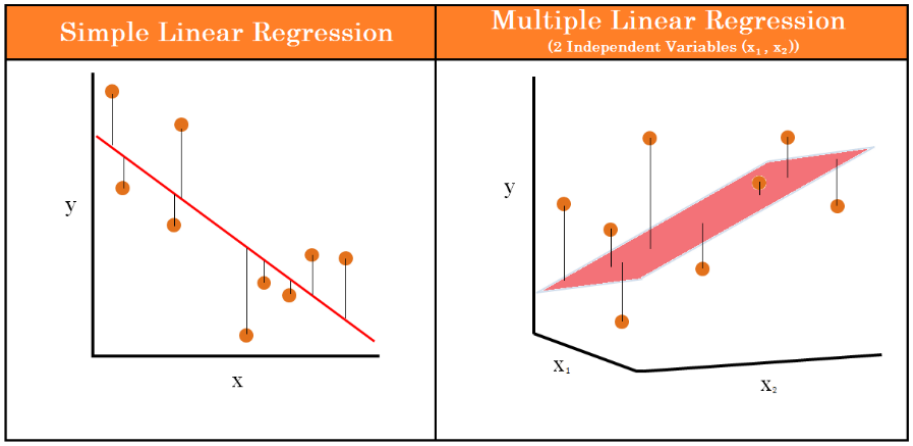
- **线性回归与多元线性回归**
- 回归分析是应用广泛的统计分析方法，用于分析事物之间的相关关系。

–一元线性回归(Linear Regression)模型，指的是只有一个解释变量的线性回归模型

–多元线性回归模型，则是包含多个解释变量的线性回归模型

–所谓解释变量就是自变量，而被解释变量则是因变量

- 回归模型就是描述因变量和自变量之间依存的数量关系的模型



二、线性回归

- **线性回归与多元线性回归**

- 一元线性回归模型具有$y = ax + b$的简单形式
  - a为自变量x的系数，b为截距
  - $y = ax + b$对应到图形，则为二维平面上的一条直线

- 多元线性回归模型具有$y = \sum_{i=1}^{n} a_i x_i + b$的形式
  - 方程包含有n个自变量$x_1, x_2, \ldots, x_n$，其系数分别是$a_1, a_2, \ldots a_n$
  - $y = a_1 x_1 + a_2 x_2 + b$对应到图形，则为一个平面
  - 多元（$n \geq 3$）线性回归模型对应到图形，则为一个超平面

二、线性回归

- **线性回归与多元线性回归**
- 现有12名大学一年级女生的体重、身高和肺活量数据。现在我们希望在这些数据上建立一个回归模型，这个模型的目的是解释这些数据，也就是肺活量和身高、体重有什么关系，模型的另外一个目的是进行预测，假设我们获得另外一个女生的身高、体重数据，我们希望用这个模型，预测出她的肺活量应该是多少。

表 5. 12 名大学一年级女生的身高、体重以及肺活量

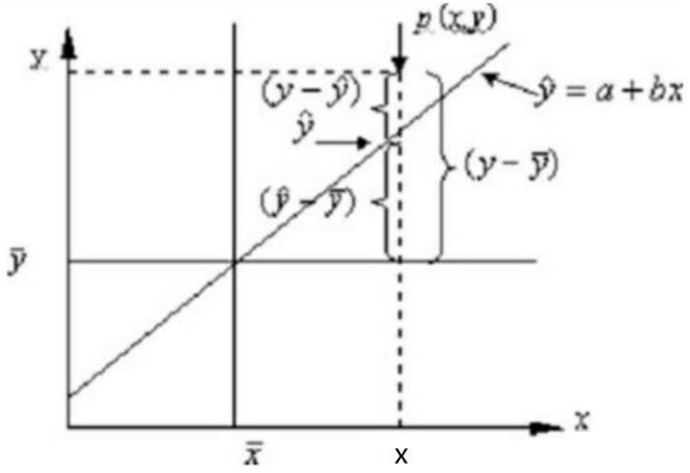
编号	身高(cm)	体重(kg)	肺活量(L)
1	161	42	2.55
2	162	42	2.2
3	165	46	2.75
4	162	46	2.4
5	166	46	2.8
6	167	50	2.81
7	165	50	3.41
8	166	50	3.1
9	168	52	3.46
10	165	52	2.85
11	170	58	3.5
12	168	58	3

二、线性回归

- **线性回归与多元线性回归**
- 我们把身高是作为第一个自变量x_1，体重作为第二个自变量x_2，而肺活量则作为因变量y。我们要建立的方程为$y = a_1x_1 + a_2x_2 + b$
- 经过计算(使用最小二乘法估计)，得出$b = -0.5657$，$a_1 = 0.005017$，$a_2 = 0.05406$。
  - 建立多元线性回归模型的目的，是解释数据以及进行预测。比如，现在遇到一个新的大学一年级女生，身高166cm，体重46公斤，把这两个数据带入上述方程，得到y=2.75，表示对于这样身高和体重的女生，估计的肺活量为2.75L。

二、线性回归

- **线性回归的评价**
- 我们把因变量的总变差(Sum of Squares for Total, SST)
  - -分解成自变量变动引起的变差(Sum of Squares for Regression, SSR)
  - -和其它因素造成的变差(Sum of Squares for Error, SSE)
  - -用数学语言来表达为
  - $-SST = \sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$=SSR+SSE
  - -式中，y表示因变量的实际值，$\bar{y}$表示样本均值，$\hat{y}$ 表示模型预测值



二、线性回归

- **线性回归的评价**

（1）**拟合优度检验**：回归方程的拟合优度，指的是回归方程对样本的各个数据点的拟合程度

- 拟合优度的度量一般使用判定系数R^2
  - 是在因变量的总变差中，由回归方程解释的变动(回归平方和)所占的比重
  - R^2越大，方程的拟合程度越高
  - R^2的计算公式为$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
    - 当一个多元线性回归模型的判定系数接近1.0时，说明其拟合优度较高

二、线性回归

- **线性回归的评价**

（2）回归方程显著性检验：回归方程的显著性检验目的是评价所有自变量和因变量的线性关系是否密切。

- 常用F检验统计量进行检验，F检验是对模型整体回归显著性的检验
  - F统计量的计算公式为$F=\frac{SSR/k}{SSE/(n-k-1)}$**式中**，n为样本容量，k为自变量个数
- F检验的原假设(H_0)为，自变量和因变量的线性关系不显著；备择假设(H_1)为，自变量和因变量的线性关系显著
  - 在给定的显著性水平(一般选0.05)下，查找自由度为(k，n-k-1)的F分布表，得到相应的临界值F_a
    - 如果上述公式计算得的$F>F_a$，那么拒绝原假设，回归方程具有显著意义，回归效果显著
    - 否则，$F<F_a$，那么接受原假设，回归方程不具有统计上的显著意义，回归效果不显著

二、线性回归

- **线性回归的评价**

（3）**回归系数的显著性检验**：使用t检验，分别检验回归模型中的各个回归系数是否具有显著性，以便使模型中只保留那些对因变量有显著影响的因素

- t检验是对单个解释变量回归系数的显著性检验
  - 回归系数i的t检验统计量为$t_i = \frac{\beta_i}{S_{\beta_i}}$，其中$S_{\beta_i}$表示回归系数$\beta_i$的标准误差
- t检验的原假设(H_0)为，a_i的值为0，即对应变量x_i的系数为0，该变量无需进入方程；备择假设(H_1)为，a_i的值不为0，即对应变量x_i的系数不为0，该变量需要进入方程
  - 给定显著性水平a(一般选0.05)，查找自由度为n-k-1的t分布表，得到临界值t_a，
    - 如果$t_i > t_a$，拒绝原假设，回归系数a_i与0有显著差异，对应的自变量x_i对因变量y有解释作用
    - 否则$t_i < t_a$，接受原假设，回归系数a_i与0没有显著差异，对应的自变量x_i对因变量y没有解释作用

二、线性回归

- **线性回归的评价**

（3）**回归系数的显著性检验**：使用t检验，分别检验回归模型中的各个回归系数是否具有显著性，以便使模型中只保留那些对因变量有显著影响的因素

- t检验的原假设、备择假设以及计算方法（假设有k个变量）

$$\begin{cases} H_0: \beta_j = 0, j = 1,2,\dots,k \\ H_1: \beta_j = 0, \end{cases}$$

$$t = \frac{\widehat{\beta_j} - \beta_j}{se(\widehat{\beta_j})} \sim t(n-k-1)$$

- $\hat{\beta}_j$为线性回归模型第j个系数估计值，
- β_j为原假设的值，也就是0，
- se($\hat{\beta}_j$)为回归系数的标准差

二、线性回归

- **共线性的检测**
- 所谓共线性，指的是自变量之间存在较强的线性关系，这种关系如果超越了因变量与自变量的线性关系，那么回归模型的就不准确了。在多元线性回归型中，共线性现象无法避免，只要不要太严重就可以了。
- 计算自变量之间的相关系数矩阵的特征值的条件数$k = \lambda_1/\lambda_p$（λ_1为最大的特征值，λ_p为最小的特征值）。如果$k \leq 100$，则自变量之间不存在严重的共线性。如果$100 \leq k \leq 1000$，则自变量之间存在较强的共线性。如果$k > 1000$，则自变量之间存在严重的共线性。降低共线性的办法，主要是转换自变量的取值，比如变绝对数为相对数或者平均数，或者更换其它的自变量。

二、线性回归

- **自变量筛选法**

- 在多元线性回归中，还存在一个自变量选择的问题，因为并不是所有的自变量都对因变量有解释作用。
- 比如，我们通过身高、体重(自变量)和肺活量(因变量)，建立回归模型
  - 这时候，我们引入一个血压数据，就可能和肺活量没有什么关系
  - 自变量间可能存在较强的线性关系，即共线性
  - 所以不能把所有的变量全部引入方程
- 变量选择的方法，包括**前向筛选法**、**后向筛选法**、和**逐步筛选法**三种

二、线性回归

- **自变量筛选法:**前向筛选法(Forward)

Step 1：因变量 y 对每个自变量 $x_1, x_2, \cdots, x_p$ 分别建立一元线性回归模型。分别计算这 p 个回归模型的 p 个回归系数的 F 值，记为 $F_1^{(1)}, F_2^{(1)}, \cdots, F_p^{(1)}$ 。选其最大者，记为

$$F_j^{(1)} = \max\left\{F_1^{(1)}, F_2^{(1)}, \cdots, F_p^{(1)}\right\} .$$

给定显著性水平 α ，若 $F_j^{(1)} > F_\alpha(1, n-2)$ ，则首先将 x_j 引入回归模型。不妨假设 x_j 就是 x_1 。

Step 2：因变量 y 对每组自变量 $(x_1, x_2), (x_1, x_3), \cdots, (x_1, x_p)$ 分别建立二元线性回归模型。分别计算这 $p-1$ 个回归模型中 $x_2, x_3, \cdots, x_p$ 的回归系数计算 F 值，记为 $F_2^{(2)}, F_3^{(2)} \cdots, F_p^{(2)}$ 。选其最大者，记为

$$F_j^{(2)} = \max\left\{F_2^{(2)}, F_3^{(2)}, \cdots, F_p^{(2)}\right\} .$$

给定显著性水平 α ，若 $F_j^{(2)} > F_\alpha(1, n-3)$ ，则首先将 x_j 引入回归模型。不妨假设 x_j 就是 x_2 。

Step 3：继续以上做法，假设已确定选入 q 个自变量 $x_1, x_2, \cdots, x_q$ ，在建立 $q+1$ 元线性回归模型时，若所有的 $x_{q+1}, \cdots, x_p$ 的 F 值均不大于 $F_\alpha(1, n-q-2)$ ，则变量选择结束。这时得到的 q 元线性回归模型就是向前法所确定的最优回归模型。

二、线性回归

- **自变量筛选法:**后向筛选法(Backward)

Step 1：因变量 y 对所有自变量 $x_1, x_2, \cdots, x_p$ 建立一个 p 元线性回归模型，分别计算 p 个回归系数的 F 值，记为 $F_1^{(p)}, F_2^{(p)}, \cdots, F_p^{(p)}$ 。选其最小者，记为

$$F_j^{(p)} = \min\left\{F_1^{(p)}, F_2^{(p)}, \cdots, F_p^{(p)}\right\} .$$

给定显著性水平 β，若 $F_j^{(p)} \leq F_\beta(1, n-p-1)$，则从回归模型中剔除 x_j 。不妨假设 x_j 就是 x_p 。

Step 2：因变量 y 对自变量 $x_1, x_2, \cdots, x_{p-1}$ 建立一个 $p-1$ 元线性回归模型。分别计算 $p-1$ 个回归系数的 F 值，记为 $F_1^{(p-1)}, F_2^{(p-1)}, \cdots, F_{p-1}^{(p-1)}$ 。选其最小者，记为

$$F_j^{(p-1)} = \min\left\{F_1^{(p-1)}, F_2^{(p-1)}, \cdots, F_{p-1}^{(p-1)}\right\} .$$

给定显著性水平 β，若 $F_j^{(p-1)} \leq F_\beta(1, n-p)$，则从回归模型中剔除 x_j 。不妨假设 x_j 就是 x_{p-1} 。

Step 3：继续以上做法，假设已确定剔除 $p-q$ 个自变量 $x_{q+1}, \cdots, x_p$，在 y 关于 $x_1, x_2, \cdots, x_q$ 建立 q 元线性回归模型时，若所有 $x_1, x_2, \cdots, x_q$ 的 F 值均大于 $F_\beta(1, n-q-1)$，则变量选择结束。这时得到的 q 元线性回归模型就是向后法所确定的最优回归模型。

二、线性回归

- **自变量筛选法:**前向筛选法与后向筛选法

- 若自变量$x_1,\ldots,x_m$是完全独立的,那么取相同的显著性水平时,前向法和后向法所建的回归方程是相同的。
- 然而这在实际中很难碰到,尤其在经济问题中。所研究的绝大部分问题，自变量间都有一定的相关性。
- 随着回归方程中自变量的增加和减少,某些自变量对回归方程的影响也会发生变化。因为自变量间的不同组合,由于它们相关的原因,对因变量的影响可能不太一样。
- 如果几个自变量的联合效应对y有重要作用,但是单个自变量对y的作用都不显著,那么前向筛选法就不能引入这几个自变量,而后向筛选法却可以保留这几个自变量.这是后向筛选法的一个优点。

二、线性回归

- **自变量筛选法:**逐步筛选法(Stepwise)

- 是“前向筛选法”和“后向筛选法”的结合
- 将自变量一个一个地引入回归模型，每引入一个自变量后，都要对已选入的自变量进行逐个检验，当先引入的自变量由于后引入的自变量而变得不再显著时，就要将其剔除。
- 将这个过程反复进行下去，直到既无显著的自变量引入回归模型，也无不显著的自变量从回归模型中剔除为止。
- 这样就避免了前向筛选法和后向筛选法各自的缺点，以保证最后得到的自变量子集是“最优”的自变量子集。
- 引入自变量和剔除自变量的显著性水平不同，引入自变量要比提出自变量的显著性水平要小。