第四章　统计预测及其常见错误

严格地说，统计推断与统计预测在内容上有相容的部分，但由于统计预测的方法具有数学上的相对独立性，在此单独列为一章，一是在原理上便于展开讨论，二是在方法上有利于读者对所讨论的内容进行分析。

第一节　相关与回归预测

在第一章，我们已经提到过相关这一概念，从字面上解释，相关就是客观事物或现象之间相互制约或相互影响的一种联系。从哲学的角度考虑，世界上任何事物都与周围的其他事物存在着联系，而且这种联系是客观的、普遍的。不过统计学上的相关指的是研究对象之间所具有的数量上的依存或制约的关系，根据影响因素和影响程度的大小，这种依存关系可以分为函数关系和相关关系两大类。

所谓函数关系，是指研究对象在数量上所表现出的严格的依赖关系，其精确的含义可以用数学中的函数的定义来描述。相关关系则表达研究对象在数量上反映出的非严格的、不确定的依存关系，这种依存关系有自身的特点：某一对象在数量上的变化会影响到另一对象，而且这种数量上的变化带有一定的随机性。具体地说就是当给定某一对象的一个数值时，另一对象会有若干数值与之

对应，并且这些数值随机地分布于某个确定的范围。研究对象在数量上之所以会发生“多值”及“随机”的对应，原因在于影响其数量变化的因素是不惟一的，有时也是不确定的。我们无法事先确知哪些条件和因素对研究对象构成何种程度的影响，只能根据研究对象表现出来的数量特征以一定的统计方法和数学手段对它们之间的因果关系进行描述。其中经常用到的方法是回归分析法。

回归分析是指对具有相互联系的现象，根据其关系形态，选择一个合适的数学模型，用以近似地表达变量的平均变化关系的一种统计分析方法[2]。根据变量的多少，常用到的回归主要是一元线性回归和多元线性回归，另外考虑到变量间的特殊关系，有时候也会用到非线性回归。在此需要强调一点，回归分析研究的变量要事先确定出自变量与因变量，并且自变量是普通变量，因变量是随机变量。相关分析与回归分析的步骤是：① 进行相关关系的定性分析；② 确定回归方程；③ 计算相关系数或相关指数，对回归方程变量之间的相关性进行显著性检验；④ 利用回归方程进行预测和推算；⑤ 对预测和推算作出区间估计。

一、一元线性回归分析预测

（一）一元线性回归模型的建立

根据回归分析方法得出的数学表达式称为回归方程。一元线性回归方程是根据某一变量与另一变量之间的关系用直线形式来表达的方程。如果自变量 X 与随机变量 Y 之间存在相关关系，由于 Y 是随机变量，对于 X 的各个确定的值，Y 有它的分布。对于两个变量的一组观察值 $(X_i, Y_i)(i = 1, 2, \cdots, n)$，可以在直角坐标系中画出它们的散点图，从图中如能观察出 X 与 Y 之间近似呈线性关系，就可以用一条直线 $\hat{Y} = a + bX$ 来描述它们之间的这种关系。当然，由于 Y 是随机变量，其值有一定的波动性，要使直线 $\hat{Y} = a + bX$ 通过所有的观察点是不现实的，但选择的用以描述 X 和 Y 之间关系

的直线要尽可能准确地体现两者之间变化的规律，就是说在给定的观察点上使 X 的观察值与理论值 $\hat{Y}$ 之间的总误差按某个标准最小。为了避免出现正、负误差相互抵消的情况，应使用两者之间误差的平方和最小这一标准，即使 $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - a - bX_i)^2$ 最小。由这个标准确定的直线还可以满足离差总和等于零，即 $\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$，因而是最优的。这种确定直线方程的方法称为最小二乘法。下面给出确定直线 $\hat{Y} = a + bX$ 的方法和步骤。

令 $Q = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - a - bX_i)^2$，由数学分析中求极值的原理，$Q$ 对 a，b 的一阶偏导数应该为零。即

$$\begin{cases} \dfrac{\partial Q}{\partial a} = \sum_{i=1}^{n} 2(Y_i - a - bX_i)(-1) = 0 \\ \dfrac{\partial Q}{\partial b} = \sum_{i=1}^{n} 2(Y_i - a - bX_i)(-X_i) = 0 \end{cases}$$

整理得

$$\begin{cases} na + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_iY_i \end{cases}$$

解得

$$a = \frac{\sum_{i=1}^{n} Y_i}{n} - b\frac{\sum_{i=1}^{n} X_i}{n} = \overline{Y} - b\overline{X}$$

$$b = \frac{n\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i\sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\sum_{i=1}^{n} X_i\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2} =$$

$$\frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

确定了参数 a,b 的值,回归直线就确定了。下面举一例说明。

例 1　在某种产品表面进行腐蚀性试验,得到腐蚀深度 Y 与腐蚀时间 t 之间对应的一组数据(见表 1),求腐蚀深度 Y 对腐蚀时间 t 的回归方程。

表 4.1　腐蚀深度 Y 与腐蚀时间 t 之间对应数据

t/s	5	10	15	20	30	40	50	60	70	90	120
Y/mm	6	10	10	13	16	17	19	23	25	29	46

解:在坐标系中描出点(t_i, Y_i),可以看出 Y 与 t 之间近似呈线性关系,故可建立 Y 与 t 的一元线性回归模型。

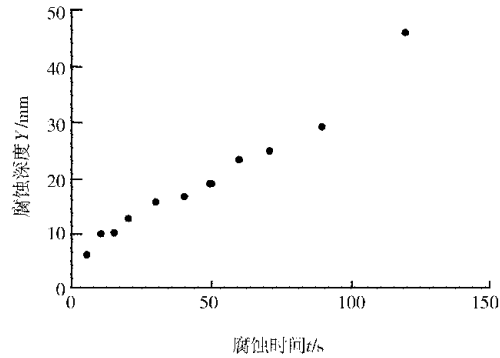

图 4.1　腐蚀深度 Y 对腐蚀时间 t 的散点图

为方便计,列出计算的相关数据,见表 4.2。

$$\bar{t} = \frac{510}{11} = 46.36, \bar{t}^2 = 2\,149.249\,6$$

$$\overline{Y} = \frac{214}{11} = 19.45, \overline{Y}^2 = 378.302\,5$$

表 4.2　腐蚀深度 Y 与腐蚀时间 t 的一元线性回归模型的计算表

i	Y_i^2	t_i^2	t_iY_i	Y_i	t_i
1	36	25	30	6	5
2	100	100	100	10	10
3	100	225	150	10	15
4	169	400	260	13	20
5	256	900	480	16	30
6	289	1600	680	17	40
7	361	2500	950	19	50
8	529	3600	1380	23	60
9	625	4900	1750	25	70
10	841	8100	2610	29	90
11	2116	14400	5520	46	120
$\sum$	5422	36750	13910	214	510

于是,有

$$b = \frac{\sum_{i=1}^{11} t_iY_i - 11\bar{t}\bar{Y}}{\sum_{i=1}^{11} t_i^2 - 11\bar{t}^2} = \frac{13910 - 11 \times 46.36 \times 19.45}{36750 - 11 \times 2149.2496} = 0.3$$

$$a = \bar{Y} - b\bar{t} = 19.45 - 0.3 \times 46.36 = 5.542$$

因此,所求的回归方程为

$$\hat{Y} = 5.542 + 0.3t$$

(二)一元线性回归模型的检验

用最小二乘法求出回归方程后,还需对变量 X 和 Y 之间的线性关系进行显著性检验,从而决定或判断能否利用该方程对因变量 Y 的取值进行控制或预测。

1. 回归方程的显著性检验(方差分析法)

回归方程的显著性检验就是要验证变量 X 和 Y 之间是否真正存在线性关系。显然,若回归系数 $b = 0$,则回归直线成为水平直

线，X 和 Y 之间无线性关系，$b \neq 0$ 时，X 和 Y 之间存在线性关系。检验步骤可按下面程序进行。

$$H_0: b = 0, H_1: b \neq 0$$

计算回归方程的 F 统计量，公式为

$$F = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})/1}{\sum_{i=1}^{n} (Y_i - \hat{Y})^2/(n-2)}$$

式中的1，$n-2$ 称为自由度。回归平方和（或称离差平方和）$\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$ 的自由度为1，因为回归方程只有一个自变量。误差平方和（或称剩余平方和）的自由度为 $n-2$，因为在 $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$ 公式中，参数 a，b 是由观察资料计算的，因而消失了两个自由度。

给定显著性水平 α 和两个自由度1，$n-2$，查 F 分布表，得临界值 F_α。若 $F \geqslant F_\alpha$，则拒绝 H_0，表明回归效果显著，否则接受 H_0，说明回归效果不显著。

2. 拟合程度的测定

把回归平方和与总平方和的比称为可决系数，即 $r^2 = \frac{\sum_{i=1}^{n} (\hat{Y} - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$，而 $r = \sqrt{\frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$

称为相关系数。r^2 和 r 可以反映各观察点在直线周围的紧密程度，称之为直线对样本数据的拟合程度。若全部观察点都落在直线上，则 $r^2 = 1$，这种情况称为完全相关；若自变量的变动对总变差毫无影响，则 $r = 0$，此种情况称为零相关。一般情况下，当 $r^2 \geqslant 0.64$ 时，

说明自变量 X 的变动对总变差的影响占一半以上，称为高度相关；$r^2 \leqslant 0.09$ 时，称为低度相关；$0.09 \leqslant r^2 \leqslant 0.64$ 时，称为中度相关。

通过计算 r^2 或 r 可以测定拟合的程度。

（三）一元线性回归模型的预测

如果一元线性回归模型经检验是显著的，就可以利用其进行预测，即给定自变量 X 的一个值 x_i，估计 Y 的取值的范围。由于各种因素的影响，对于给定的 X_i，实际观测值 Y_i 与由回归方程计算得到的理论值 $\hat{Y}_i$ 一般总存在一定的误差，即预测误差。测定预测误差变动范围，常采用估计标准误差来说明 Y 与 $\hat{Y}$ 的差异程度。计算公式为 $S_Y = \sqrt{\dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$，实际应用时，可用简捷的公式：$S_Y = \sqrt{\dfrac{\sum_{i=1}^{n} Y_i^2 - a\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i Y_i}{n-2}}$。若 Y 服从正态分布，当 n 较大，且 X_i 不远离 $\bar{X}$ 时，则可以确定预测区间为：

以 0.628 7 为概率保证的 Y_i 的预测区间为 $(\hat{Y}_i - S_Y, \hat{Y}_i + S_Y)$；

以 0.954 5 为概率保证的 Y_i 的预测区间为 $(\hat{Y}_i - 2S_Y, \hat{Y}_i + 2S_Y)$；

以 0.997 3 为概率保证的 Y_i 的预测区间为 $(\hat{Y}_i - 3S_Y, \hat{Y}_i + 3S_Y)$。

当 n 较小时，若给定置信概率 $1-\alpha$，则 Y_i 的预测区间为

$$\left(\hat{Y}_i - t_{\frac{\alpha}{2},(n-2)} S_Y \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}, \hat{Y}_i + \right.$$

$$t_{\frac{\alpha}{2},(n-2)} S_Y \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}}$$

其中，$t_{\frac{\alpha}{2},(n-2)}$ 可以通过查 t 分布表得到。

二、多元线性回归分析预测

多元回归分析是在研究多个自变量对一个因变量的相互关系的基础上，确定出它们之间的多元回归方程，进而根据各个自变量的变动情况来估计或预测因变量的变动程度。其实质是一元线性回归分析的扩充，计算原理及预测方法完全相同，只是计算较繁。

设因变量 Y 受 m 个自变量 $X_1, X_2, \cdots, X_m$ 的影响，其回归方程为

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m$$

由最小二乘法，要使 $\sum_{i=1}^{n} (Y_i - \hat{Y}_1)^2$ 最小，须满足

$$\begin{cases} \sum Y = na + b_1 \sum X_1 + b_2 \sum X_2 + \cdots + b_m \sum X_m \\ \sum X_1 Y = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 + \cdots + b_m \sum X_1 X_m \\ \sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 + \cdots + b_m \sum X_2 X_m \\ \sum X_3 Y = a \sum X_3 + b_1 \sum X_1 X_3 + b_2 \sum X_2 X_3 + \cdots + b_m \sum X_3 X_m \\ \quad \vdots \\ \sum X_m Y = a \sum X_m + b_1 \sum X_1 X_m + b_2 \sum X_2 X_m + \cdots + b_m \sum X_m^2 \end{cases}$$

从中解出 $a, b_1, b_2, \cdots, b_m$，就可以得到回归方程 $\hat{Y} = a + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m$。

多元线性回归模型的检验内容和方法与一元线性回归模型的检验相似，可分为回归系数的显著性检验、回归方程的显著性检验、拟合程度的检验、估计标准误差等。下面只介绍多元线性回归

方程的显著性检验。步骤如下：

（1）提出假设　$H_0: b_1 = b_2 = \cdots = b_m$；$H_1: b_i$ 不同时为0，$i = 1,2,\cdots,m$；

（2）计算 F 统计量 $F = \dfrac{\sum (\hat{Y} - \overline{Y})^2/m}{\sum (Y - \hat{Y})^2/(n - m - 1)}$；

（3）根据给定的显著性水平 α，自由度 $m, n - m - 1$，查 F 分布表，得到相应的临界值 F_α。若 $F > F_\alpha$，则拒绝 H_0，认为回归方程有显著意义；若 $F \leqslant F_\alpha$，则接受 H_0，认为回归方程回归效果不显著。

三、非线性回归分析预测

实际中，有时两个变量间的相关关系并非线性关系，而是存在某种曲线关系，这时需要应用适当形式的曲线回归模型来描述它们之间的关系。这种分析方法称为非线性回归分析。常见的一元非线性回归模型主要有下列几种：

（1）抛物线型：$\hat{Y} = a + bX + cX^2$；

（2）双曲线型：$\hat{Y} = a + \dfrac{b}{X}$；

（3）对数型：$\hat{Y} = a + b\ln X$；

（4）指数型：$\hat{Y} = a + \mathrm{e}^{bX}$。

对于抛物线型，可以直接应用最小二乘法求出 a, b, c 的值。具体做法是：令 $Q = \sum (Y - \hat{Y})^2 = \sum (Y - (a + bX + cX^2))^2$，欲使 Q 最小，则 Q 对 a, b, c 的偏导数为零。即

$$\begin{cases} \sum Y = na + b\sum X + c\sum X^2 \\ \sum XY = a\sum X + b\sum X^2 + c\sum X^3 \\ \sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 \end{cases}$$

从中解出 a,b,c 的值，就得到了回归方程。

对于双曲线型，令 $X' = \frac{1}{X}$，就可以化成一元线性回归模型：$\hat{Y} = a + bX'$；对于指数型，令 $\hat{Y}' = \ln\hat{Y}$，则可化为一元线性回归模型：$\hat{Y}' = \ln a + bX$；对于对数型，令 $X' = \ln X$ 即可化为一元线性回归模型：$\hat{Y} = a + bX'$。模型化为一元线性回归形式后，可以按照上面介绍的方法求解或分析，在此不再赘述。

第二节　时间序列预测

动态地分析研究对象的数量特征，是统计工作中研究随时间而连续变化的观察对象的客观需要。在及时地掌握分析有关时间序列资料的前提下，才能准确把握研究对象的发展变动规律。这里的时间序列指的是按固定时间距离将研究对象的某种指标观察数值顺序地排列而成的数列，也称为时间数列。时间数列按变量的表现形式可分为绝对时间数列、相对时间数列和平均时间数列。所谓绝对时间数列是指由反映某对象在时间上规模大小、数量多少的绝对数所构成的数列；相对时间数列是由反映研究对象之间数量对比关系的相对数所构成的数列；而平均时间数列则是由反映研究对象某一数量特征在不同时间上的一般水平的平均指标所构成的数列。其编制原则必须符合指标值的可比性原则。即各时期指标的计算方法、计算范围和计算内容要保持一致，而且时间的长短也要统一。在此基础上才可以展开对研究对象变化规律的分析和预测。

一、时间序列的分析指标

动态分析指标可分为两类：一类是研究对象的发展水平指标；

另一类是研究对象的发展速度指标。

（一）发展水平

发展水平是时间数列中原有的统计指标数值，它反映研究对象在不同时期的规模水平。通常用 $a_0, a_1, \cdots, a_n$ 表示各个时期的发展水平。其中 a_0 称为基期水平，其余 a_i 称为报告期水平或计算期水平。

（二）平均发展水平

平均发展水平是不同时期的发展水平的平均值，它从动态上反映研究对象在一段时间的一般发展水平，又称为“时序平均数”。若时间序列反映的是一个时期指标在各个时期的发展水平，则时序平均数的计算公式为：$\bar{a} = \dfrac{a_1 + a_2 + \cdots + a_n}{n} = \sum_{i=1}^{n} \dfrac{a_i}{n}$；若时间序列表现的是某一时点指标在不同时期的发展水平，当整个研究期的各个时点数据不齐备、不连续，但时点的间隔相等时，可采用下列公式计算时序平均数。

$$\bar{a} = \frac{\frac{1}{2}a_1 + a_2 + \cdots + a_{n-1} + \frac{1}{2}a_n}{n-1}$$

若时间数列反映某一时点指标在不同时期的发展水平，而掌握的时点数据资料的时间间隔不相等，则需以时间间隔长度 f 为权重，采用下式来计算时序平均数：

$$\bar{a} = \left[\left(\frac{a_1 + a_2}{2}\right)f_1 + \left(\frac{a_2 + a_3}{2}\right)f_2 + \cdots + \left(\frac{a_{n-1} + a_n}{2}\right)f_{n-1}\right] \Big/ \sum_{i=1}^{n-1} f_i$$

（三）增长量

增长量是指某一研究对象的某一统计指标在一定时期内增长或减少的数量。根据对比基期的不同，可分为逐期增长量和累积增长量。逐期增长量是以前期为基期计算的两相邻时期的增长量；累积增长量是所有逐期增长量的和。

（四）平均增长量

平均增长量是指一定时期内逐期增长量的算术平均值。即

$$\text{平均增长量} = \frac{(a_1 - a_0) + (a_2 - a_1) + \cdots + (a_n - a_{n-1})}{n} = \frac{a_n - a_0}{n}$$

（五）发展速度

发展速度是以相对数形式表示的两个不同时期发展水平的比率。根据计算采用的基期的不同，发展速度有定基和环比之分。定基发展速度是各报告期水平与某一固定时期的水平进行对比，其所形成的时间数列表明被研究现象在一定基础上较长时间内总的发展变化程度；环比发展速度是指各报告期水平与其前期的水平进行对比，其所形成的时间序列说明被研究现象逐期发展变化的情况。

（六）增长速度

增长速度是增长量与基期水平的比率，根据比较的基期不同，增长速度分为定基增长速度和环比增长速度。计算公式分别为

$$\text{定基增长速度：}\frac{a_1 - a_0}{a_0}, \frac{a_2 - a_0}{a_0}, \frac{a_3 - a_0}{a_0}, \cdots, \frac{a_n - a_0}{a_0}$$

$$\text{环比增长速度：}\frac{a_1 - a_0}{a_0}, \frac{a_2 - a_1}{a_1}, \frac{a_3 - a_2}{a_2}, \cdots, \frac{a_n - a_{n-1}}{a_{n-1}}$$

（七）平均发展速度

平均发展速度是各个时期环比发展速度的平均值。由于总速度不等于各个环比速度之和，而是等于相应时期内各环比发展速度的连乘积，故应采用环比速度的几何平均数来计算平均发展速度，即 $\overline{X}_G = \sqrt[n]{\frac{a_1}{a_0} \times \frac{a_2}{a_1} \times \cdots \times \frac{a_n}{a_{n-1}}} = \sqrt[n]{\frac{a_n}{a_0}}$。

（八）平均增长速度

平均增长速度是环比增长速度的平均值，根据它与平均发展速度的内在联系，可得平均增长速度的计算公式为：$\Delta\overline{X} = \overline{X}_G - 1$。

例2　已知我国钢产量1986－1990年的环比增长速度分别为15.6%，7.8%，5.6%，3.6%，7.2%。计算这期间我国钢产量的平均增长速度。

解：先求环比发展速度 $\overline{X}_G$

$$\overline{X}_G = \sqrt[n]{115.6\% \times 107.8\% \times 105.6\% \times 103.6\% \times 107.2\%} = 1.08$$

所以，$\Delta\overline{X} = \overline{X}_G - 1 = 1.08 - 1 = 0.08$，即平均增长速度为8%。

二、时间序列的分析与预测

一般情况下，对种群数量的变化和社会经济现象的分析预测常常使用时间序列预测法，它通过编制和分析时间序列，根据时间序列所反映出来的发展过程、方向和趋势，进行类推，借以预测下一时期或以后若干时期可能达到的水平。下面以社会经济现象为例展开讨论。

（一）趋势预测

趋势预测是根据较长时期的时间序列资料，在假定其过去的发展趋势及其变化规律性依然保持的前提下，探求其趋势并进行外推以测算其发展方向和变化程度的方法。常用的方法有时距扩大法、移动平均法、指数平滑法等，更有效的方法可以利用数学中的插值和拟合，即根据时间序列的数据资料，构造适合原始数据变动规律的直线或曲线，以此作为预测未来的根据。如最小二乘法就是预测中常用的典型的数据拟合方法。因为前面已作过介绍，在此不作重述。

（二）季节变动预测

季节变动预测是根据历年的时间数列资料，采用测定季节变

动的各种特有的方法，揭示客观事物季节变动的方向和程度，据以进行科学预测的方法。常用的有按月平均季节指数法和移动平均趋势剔除法。

1. 按月平均季节指数法

它是指对不含长期趋势，仅受季节变动因素影响而呈现周期性变动规律的时间数列进行预测的方法。一般步骤是：① 搜集历年各月（季）的资料；② 计算数年内同月（季）的平均数；③ 计算总的月（季）的平均数；④ 计算各月（季）的季节指数，即

$$\text{月的季节指数} = \frac{\text{月平均数}}{\text{总的月平均数}}$$

$$\text{季的季节指数} = \frac{\text{季平均数}}{\text{总的季平均数}}$$

⑤ 预测。根据季节指数和已知某年一个月或几个月的实际值就可以用比率法预测该年其他月或季的数值。

2. 移动平均趋势剔除法

该方法主要用于剔除季节指数计算时动态数列中的长期趋势、循环变动及不规则变动等影响因素，使预测结果更切合实际。其步骤是：① 计算 12 个月的移动平均数；② 将观察值除以对应的趋势值、循环变动值，得到季节变动和不规则变动的相对数；③ 将几年同月加总求月平均数，以消除随机的不规则变动因素；④ 将 12 个月的平均数加总算出总的月平均数，然后计算季节指数；⑤ 进行预测。这里需要说明的是，所谓按 12 个月计算的移动平均数，就是将时间数列的各月资料，逐月推移计算 12 个月的序时平均数。由于按时间顺序连续 12 个月的序时平均数，已经包含了一年四季的全部季节因素，因而“旺季”与“淡季”的平均会消除掉季节变动因素的影响，同时其他不规则波动的因素也会因为移动平均而受到削弱。

时间序列预测是统计预测中关于经济行为最常用的预测方

法。由于各种经济现象的特殊性，在具体应用时也要考虑经济现象变化的环境和条件，恰当选择适宜的预测工具和手段。并根据研究的需要，合理制定推断和预测的数学模型，使分析预测的结果能更准确、更有效地服务于社会经济建设。

第三节　统计预测中的注意事项及常见错误辨析

统计预测属于统计分析的范畴，作为统计工作的最终环节，它在很大程度上体现了统计工作的全部价值。统计的目的之一就是依据已掌握的资料或数据对未知的因素或参数进行推断和估计，以预测其发展或变化趋势，并将此作为制定计划、设立目标以及进行决策的理论依据。因此，恰当合理的预测分析方法对整个统计工作至关重要。下面针对统计预测的不同应用方法给出相应的注意事项，并结合具体的实例给出常见错误的辨识方法。

一、相关与回归分析中应注意的问题与常见错误

相关与回归是研究变量之间关系的直观方法，在应用分析时一般分为定性分析和定量分析两个阶段。定性分析是定量研究的基础和前提，也是选择回归模型的重要参照和依据，因此必须注意以下几个问题。

1. 分析判明变量之间确实存在的相关关系

在应用相关与回归分析时，首先要分析判明拟分析的变量之间确实存在一定的相关关系，否则，将风牛马不相及的变量强加联系进行分析就会得出荒谬的结论。例如，有人将某地区的降雨量与该地区的绿化面积联系在一起进行研究，得出正相关的结论；也有人将烟草的销售量与人的平均寿命结合在一起，也得出正相关的结论；更有甚者，有人将历年捕鲸的数量与股票价格指数凑合在一起，得出了负相关的结论。这种牵强的联系所得出的结论一旦用于

实际作为进行推断和预测的理论依据，势必会对整个工作造成损失。在医学研究中，这种错误十分普遍。如，将专业上毫无关联的两个变量放在一起作相关和回归分析；错误地将变量之间在统计学上的关系解释成在专业上的联系；用直线回归方程描述呈明显曲线变化趋势的实验资料；在利用回归方程预测时，随意将范围扩大，超出原观察值的范围等。以下几个问题在相关与回归分析中要引起注意。

（1）经样本求得的相关、回归系数要作假设检验

例如，某研究者欲探讨年龄（x）同肾病 LgA（y）的关系，经相关和回归分析，得方程 $y = -108.6 + 5.5x$，相关系数 $r = 0.340$。如果就此得出结论：LgA 随年龄增长而升高，则是不妥当的，必须经假设检验后，方能作出恰当的结论。用自由度 $\upsilon = n - 2 = 30$，可查得 $r_{0.05(30)} = 0.349$，$|r| < r_{0.05}$，$P > 0.05$。假设检验结果提示，没有足够的理由认为年龄和 LgA 有关。因此，回归方程 $y = -108.6 + 5.5x$ 并无实际意义。

（2）判断两变量线性关系的密切程度要看 r^2 的大小

例如，有人分析了年龄与淋巴细胞转化率的关系，调查了 252 人，求得回归方程为 $y = 76 - 0.4x$，$r = -0.20$，由于 $r_{0.05(250)} = 0.162$，$|r| > r_{0.01}$，$P < 0.01$，结论是淋巴细胞转化率与年龄呈负相关。有的研究人员在这种情况下可能得出的结论是“淋巴细胞转化率与年龄密切相关”，“淋巴细胞转化率与年龄明显相关”或“淋巴细胞转化率与年龄显著相关”。这些结论都不确切或可能导致误解。因为相关系数的统计学检验，不论 P 值多么小，只能提供两个变量是否相关的信息，却不提供相关是否密切的信息。相关的密切程度常以相关系数的数值大小为指标。相关系数的绝对值越接近 1，两变量的相关关系越密切；越接近 0，越不密切。如果 $r^2 = 0.6^2 = 0.36$，说明变量 y 的变异有 36% 与 x 有联系。对于前例，$r^2 = (-0.20)^2 = 0.04$，说明淋巴细胞转化率的变化中有4% 与年

龄有关。所以尽管在相关系数的假设检验中得$P < 0.01$,但两变量的线性相关程度还是很低的。

(3) 不要把相关、回归关系直接看作因果关系

两事物间有数量关系,可能是因果关系,也可能并不存在因果关系,而仅仅是伴随关系。例如,调查某地区的电视机拥有量和肺癌死亡率,这些年来,每百家庭的电视机拥有量在不断增加,该地区的肺癌死亡率也在不断上升,两者在数量上是相关的,但看电视不见得是肺癌发病的原因。要判断两事物间是否存在因果关系,必须作专门的研究。

2. 正确确定自变量和因变量

在回归分析中,一定要正确确定自变量和因变量。尤其是注意做回归分析的因变量是随机变量,颠倒二者之间的位置关系,必然造成回归模型的错误,从而得不出有效的结论。

3. 根据研究对象的具体数量特征,恰当地选用相应的回归表达式

由于现实中存在大量非线性相关的情况,而数学中所能应用的回归模型又极其丰富,这就给选择恰当的表达形式带来一定的难度。因此在作回归分析之前,必须认真分析变量之间的相关特性,通过在坐标系中描点并观察其散点图的分布趋势等手段从整体上把握变量之间所遵从的大致规律,尽可能地使回归模型的选择体现变量之间变化的真正规律。另外,在多元回归分析中,既要注意正确选用不与因变量关系密切的变量做自变量,又要在众多的自变量中进行筛选,留用与自变量关系不密切的变量,舍弃关系密切的变量,以防止多重共线性的发生。

4. 注意回归分析应用的范围和条件

由于用于建立回归模型的资料和数据具有应用范围和适用条件的限制,所以任何一个回归方程都只适用于该资料所能容许的有效范围,而不能无限制地加以外推。在具体应用中,要根据所建

立的模型的形式和适用条件进行必要的显著性检验，以确定模型所能预测的有效范围及可靠程度。

5. 模型的求解和检验要依据统计原理，不能仅凭主观作出判断或盲目套用固定的格式

下面举例说明回归分析中常见的且比较容易忽视的问题和错误。

例 1　假设儿子的身高(y) 与父亲的身高(x) 适合一元正态线性回归模型，表 4.3 是对英国 10 对父子身高的观察数据，① 建立 y 对 x 的回归方程；② 对线性回归方程作显著性检验(检验水平取为 0.05)。

表 4.3　10 对英国父子身高的对照表

x	60	62	64	65	66	67	68	70	72	74
y	63.6	65.2	66	65.5	66.9	67.1	67.4	63.3	70.1	70

一般解法如下：

(1) 按所给数据计算

$\sum_{i=1}^{10} x_i = 668$，$\bar{x} = 66.8$，$\sum_{i=1}^{10} x_i^2 = 44\,794$；$\sum_{i=1}^{10} y_i = 665$，$\bar{y} = 66.5$，$\sum_{i=1}^{10} y_i^2 = 44\,283.93$；

$\sum_{i=1}^{10} x_i y_i = 1\,149.24$，$S_{xx} = \sum_{i=1}^{10} x_i^2 - 10\bar{x}^2 = 171.6$，$S_{xy} = \sum_{i=1}^{10} x_i y_i - 10\,\overline{xy} = 63.72$；

故 $b = \dfrac{S_{xy}}{S_{xx}} = 0.3713$，因此所求回归方程为 $\hat{y} = 0.3713x$。

(2) 待检验的原假设为 $H_0: b = 0$ 的显著性假设检验问题，检验统计量是 $t = \dfrac{b}{\sigma}\sqrt{S_{xx}}$，水平为 α 的拒绝域为 $|t| = \dfrac{|b|}{\sigma}\sqrt{S_{xx}} \geqslant t_{\frac{\alpha}{2}}(n)$，再按所给数据计算

$$S_{yy}=\sum_{i=1}^{10}y_i^2-10\bar{y}^2=46.129;\sigma^2=\frac{1}{n-2}[S_{yy}-bS_{xy}]=3.0587$$

所以$|t|=\frac{0.3713}{\sqrt{3.0587}}\sqrt{171.6}=2.7811\geqslant 2.2281=t_{0.025}(10)$，于是拒绝原假设，认为回归方程显著。

分析：上述解法存在两个明显的错误：

(1) 求解回归方程时，漏估了常数项a，对于呈线性关系的两个变量，我们尤其注意不能认为回归直线一定会通过原点；

(2) 上一章我们已经提及，统计量t服从的分布是$t(n-2)$，而不是$t(n)$。由于该部分的计算较为复杂，这些细节和应用技巧常常会被忽略，在应用中应该引起充分的重视。

正确解法：在求出回归系数$b=\frac{S_{xy}}{S_{xx}}=0.3713$之后，再由公式求解$a=\bar{y}-b\bar{x}=41.7072$。因此，所求的回归方程为$\hat{y}=41.7072+0.3713x$。

在$H_0:b=0$的显著性假设检验中，$|t|=\frac{0.3713}{\sqrt{3.0587}}\sqrt{171.6}=2.7811\geqslant 2.3060=T_{0.025}(8)$，所以拒绝原假设，认为回归方程显著。

回归预测的另一类常见错误是在变量的观察值数量不足的情况下，盲目使用回归分析。事实上，在做统计预测时，用于回归的数据在理论上应该多于用于插值的数据。用插值法构造的曲线要经过所有的数据点，而用回归手段构造的预测曲线一般情况下不会经过所有的数据点，它只是在某种标准下使所有数据点与回归曲线上相同横坐标的点的总误差最小。因此，只有一定数量保证的观察点才能拟合出较高预测精度的回归曲线。如果数据点过少，我们根本无从判断变量间是否存在真正的相关关系，或者不能正确判断出变量间的大致变化趋势，更不用说准确选择回归曲线的类型

了。我们在科技期刊中可能经常会看到用回归作统计预测的实例，其实在变量的分布未知的情况下，有些数据是严重不足的，据此所得出的回归曲线根本不能用作统计预测，至多只能在非常有限的范围内对变量作粗糙的估计。

二、时间序列预测中的常见问题

时间序列预测在社会经济现象和生物种群的数量变化等统计问题上应用较广，由于涉及较多的概念，所以应用时常会发生误用、滥用统计指标的错误。在一些期刊上，有些统计术语张冠李戴的现象也屡见不鲜。而在预测手段上，也常会看到各种分析方法分辨不清的错误。下面对常见的典型错误分别进行描述。

（一）误用和滥用速度指标

速度是时间序列中用以描述研究对象的某一数量特征随时间而变化的一个统计指标。尽管速度的计算与分析相对简单，但以下两种情形尤其要引起重视：

（1）时间序列中的指标值为0或负数时，不宜直接计算速度。如，某企业连续五年的利润额分别为8，6，0，－2，1万元，对于这一数列，直接套用求速度公式是不恰当的，应该用绝对数进行分析；

（2）要注意速度指标与发展水平指标的结合。速度指标是相对数，其数值的大小取决于报告期和基期两个发展水平。速度快不一定增长量大，速度慢也未必增长量小。1949年，我国钢产量只有16万吨，增长1% 为1600吨，但到1998年，我国钢产量已达11559万吨，增长1% 就有115.59万吨，后者是前者的720多倍，只用前后两个增长速度显然不能准确评价两个时期的经济对比。因此，观察和研究问题不能只考察一个指标就盲目得出结论，要把上述两个指标结合起来。为避免统计中出现上述片面性，可以使用把增长速度和增长量结合起来的指标，即增长百分之一的绝对值，公式为：

$$\text{增长1\% 的绝对值} = \frac{\text{逐期增长量}}{\text{环比增长量}} = \frac{\text{前期水平}}{100}$$

这个公式不仅可用于比较同一事物不同时期增长速度的经济或生物学意义，还可以比较不同地域、不同单位之间同一研究对象增长速度所隐含的不同价值意义。从正确处理和评价速度与效益的关系角度来讲，这一指标有明显的优势。

（二）长期趋势预测中含有过多主观因素

长期趋势预测的真正目的在于通过对研究对象变化趋势的正确估计，来实现对未来工作计划的有效制定或对发展目标的合理决策。它是人们后续工作的导航和指引。错误的预测必然会给今后的工作带来巨大损失。而且这种错误产生的根源往往是人们对事物发展表面现象的主观惯性外推或仅凭臆断而作的宣传。《财经》杂志曾发表北京大学中国经济研究中心教授宋国青的文章，指出主观臆测对中国经济上造成的危害，以及错误的预测的来源。现摘录部分内容如下：

“从总量上看，现在整个经济的温度并不高，充其量只能说略为偏热（这是指绝对水平，并不是当前的变化情况），但是结构问题相当突出。除国内的电力和铁路运输紧张之外，一些进口大商品的价格仍在高位盘旋，国际海运费率依然过高。产生这些现象的一个基本原因是前些年的预期错误。1998 年初，中央政府提出了 8% 的经济增长目标。按那时绝大多数人的理解，控制经济过热比较难，而要让经济增长率和通货膨胀率高一点是很容易的。这个理解和现在流行的看法比较接近，理由是地方政府和国营企业具有投资冲动，中央政府手一松就会出现投资热。在 1998 年的大部分时间里，国内的绝大部分研究者相信政府的目标肯定能够实现。虽然后来报告的经济增长率距离政府目标只差一点点，但是考虑到目标留有的余地以及目标本身可能导致的统计数据问题（四季度的环比增长率异常偏高），当时的主流预测仍然是有相当大偏差的。

到了1999年,原来比较乐观的人有一部分调低了预测,但还有不少人相信在实施了积极财政政策等刺激措施之后,经济增长率能够恢复到8%以上。在1999年上半年的统计数据出来以后,绝大部分研究者都变成了悲观主义者,只有极少数人继续保持高调预测。在别人看来,这时的高调预测很可能是出于一定目的的宣传,并不是严肃的估计。研究者用文字和话语表达自己的预测,投资者用投资行动表达自己的预测。1999年的固定资产投资增长率掉到了5.1%,这还是积极财政政策拉起来的。在1996年到2000年的五年,工业尤其制造业的投资出现了一个巨大的低谷。预测错误的发生并不奇怪,古今中外都经常有,但是这一次的预测错误仍然有其独特的地方。首先是经济增长率下降和不景气持续的时间长,预测错误持续的时间也比较长,这使电力和铁路投资偏小的局面持续了比较长的时间。其次是中国经济与世界经济的互相影响尤其是中国对世界的影响比以往要大得多,这使很多商品的国际价格在更大程度上与中国经济相联系。最重要的是,由长时期的预测错误所导致的当前供求状况并不是长期的供求均衡状况。国内电力和铁路运输紧张是前几年需求预测偏低的结果,国际上金属材料价格和海运费率偏高也一样。如果在前些年都正确预测到了今天的中国需求,相关的投资就会多得多,今天的价格就不会是这样的。对于那些规模报酬不变或者下降的产业来说,均衡价格与预期到的需求数量没有关系甚至是相反的关系,无论需求来自中国还是中国之外。对大部分商品来说,不是需求的增加引起了价格上升,而是出乎意料的需求引起了价格的上升。从这个角度来看,有关国际价格的调整是必然的,就像国内的电力供给会增加到基本适应需求一样。”

由上面这段文字可以看出,长期预测中的错误并非人们的不智,而是没有以科学的统计方法为依据,是预测时过于主观的必然结果。由这类错误造成的损失巨大,所以应引起充分的重视。