统计预测

——方法与应用

易丹辉 编著

中国统计出版社
China Statistics Press

统计预测
——方法与应用

易丹辉　编著

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

统计预测:方法与应用/易丹辉编著.
-北京:中国统计出版社,2001.4
ISBN 7-5037-3453-1

Ⅰ.统…
Ⅱ.易…
Ⅲ.统计预测-分析方法
Ⅳ.C8

中国版本图书馆 CIP 数据核字(2001)第 02620 号

责任编辑/吕　军
责任校对/刘开颜
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 75 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电　　话/(010)63459084、63266600-22500(发行部)
印　　刷/科伦克三莱印务(北京)有限公司
经　　销/新华书店
开　　本/787×1092mm　1/18
字　　数/300 千字
印　　张/17.75
印　　数/1-3000 册
版　　别/2001 年 4 月第 1 版
版　　次/2001 年 4 月北京第 1 次印刷
书　　号/ISBN 7-5037-3453-1/C·1849
定　　价/28.00 元

再版序言

自1988年2月《统计预测—方法与应用》完稿以后，我一直在使用这本书为本科生、研究生讲授相关课程。十几年来，我国发生了很大变化，科学技术飞速发展，计算机技术广泛应用，网络在经济社会中起着越来越大的作用，人们对现象之间的数量关系、对经济社会未来的发展趋势越来越关注，应运而生的预测方法层出不穷。虽然我一直在注意跟踪方法的发展，并力求将其运用于我国实际问题的研究，但总感欠缺太多。近些年来，教学内容作了许多补充，也几次想修订该书，但终因事务性工作太多，难得静下来思考、修改。眼看进入21世纪，再不修订，实在是愧对学生。我终于下决心进行修订。在出版社的热情鼓励下，现在总算完稿，但有些方法的阐述仍不尽人意。将这样一本不甚满意的书奉献给读者，我是深感歉意。好在网络的发展，为和学生以及读者架起了一座沟通的桥梁，我会在中国人民大学统计学系的网站上公布习题并通过网络与读者进行交流，探讨预测方法及其应用。由于篇幅的限制，本书未对Eviews软件的使用作说明。有愿使用的读者可以参看将要出版的《数据分析与Eviews应用》。期望有更多的人运用统计方法分析研究我国的实际问题。

易丹辉

2001年1月

前　言

在社会经济活动中，无论从宏观的角度还是从微观的角度，都存在着许多未知的因素，影响着各级的管理决策。为了克服未知因素可能带来的消极后果，必须进行有科学根据的预测。所谓预测，是人们在观察和分析客观事物发展过程的历史及现状的基础上，通过对客观事物发展规律的认识，进而推断其未来状况的过程。为了收到预期的预测效果，对于预测对象最好提出几种不同的预测方案，在各种方案中，充分衡量预测对象变化的条件以及可能变化的幅度，相应地采取有关措施，以便保持最佳的管理过程。换句话说，预测是在制定切实可行的计划时，为了避免可能产生的缺点和失误，而对事物的未来发展预先进行的多种方案的设计和研究。

预测可以按不同的标准进行分类。预测方法基本上分为两大类，即定性分析法和定量分析法。本书比较详尽地介绍了用于预测的定量分析方法：因果回归分析法和时间序列分析法。为了将每种具体方法与我国的社会经济实际相结合，在每一方法介绍之后，都配有实例说明其应用，书中所有计算均应用电子计算机完成。为帮助读者掌握和运用各种方法，特别是无法进行手工计算的方法，书后附有 TSP 软件的使用说明，它适用于 IBM－PC 机以及与它兼容的微型机，如长城 0520。介绍方法时，涉及到的比较复杂的数学公式推导和证明，均列入各章附录中，供读者参考。

本书编写的过程中，得到中国人民大学计划统计学院计算机室刘延军、陈虹同志，计划经济学系成晓梅同志以及校信息中心的

同志们的帮助与支持。书中采用的某些实例，是我系袁卫同志在硕士研究生学习期间收集的资料，他为编写此书提出了不少建议。在此一并表示衷心的感谢。

本书试图将各种预测方法与我国的实际结合运用，由于水平有限，编写时间又较仓促，一定存在不少缺点，殷切期望读者们随时给予批评指教。

1988 年 2 月

目　录

第一章　简单回归分析法

客观事物之间常存在着某种因果关系，如工业产品成本的降低常导致利润的上升；某种消费品价格的提高往往造成销售量的下降，等等。这种因果关系往往无法用精确的数学表达式描述，只有通过对大量观察数据的统计处理，才能找到它们之间的关系和规律。回归分析就是通过对观察数据的统计分析和处理，研究与确定事物间相关关系和联系形式的方法。运用回归分析法寻找预测对象与影响因素之间的因果关系，建立回归模型进行预测的方法，称为因果回归分析法。其特点是，将影响预测对象的因素分解，在考察各个因素的变动中，估计预测对象未来的数量状况。它建立的是预测对象与影响因素之间的单一方程，因此也被称为单方程模型分析。按方程中影响预测对象因素的多少，分为简单回归分析法和多重回归分析法。

回归分析法在预测中主要用以解决下面的问题：

(1)分析所获得的统计数据，确定几个特定变量之间的数学关系形式，即建立回归模型。

(2)对回归模型的参数进行估计和统计检验，分析影响因素对预测对象的影响程度，确定预测模型。

(3)利用确定的回归模型和自变量的未来可能值，估计预测对象的未来可能值，并分析研究预测结果的误差范围及精度。

第一节　模型和参数估计

如果影响预测对象的主要因素只有一个，并且它们之间呈线性关系，那么可以采用简单回归分析法预测。由于这种方法只涉及一个自变量，也称为一元线性回归分析法。

1.1.1　理论回归模型

将预测对象作为因变量 Y，主要影响因素为自变量 X，它们之间的线性关系，从理论上说，能够表述为下面的形式

$$Y = \alpha + \beta X + \varepsilon \tag{1.1}$$

式中：α 和 β 是固定的但未知的参数，它们反映了变量 X 与 Y 之间应该有的一种线性关系；α 是常数项；β 是理论回归系数；ε 是那些除 X 以外，被忽略和(或)无法考虑到的因素，被称为随机项。对于每一组可以观察到的因变量、自变量数值、Y_i、X_i，(1.1)式可以写成

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{1.2}$$

式中：ε_i 满足

$$E(\varepsilon_i) = 0$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases} \tag{1.3}$$

1.1.2 实际回归模型

要得到(1.1)式中参数 α 和 β 的精确值几乎不可能，因为通常只有有限的样本数据和情报。利用有限的资料，只能得到参数 α 和 β 的估计值 a 和 b。实际上，因变量 Y 和自变量 X 之间的简单线性关系能够表述为

$$Y = a + bX + + e \tag{1.4}$$

这里，a、b 不是象 α、β 那样固定的数值，而是能够取多个数值的统计估计值；e 是残差项，也被称为回归余项，它是由于用 $a + bX$ 估计因变量 Y 的数值所造成的，是估计值与实际数值之间的离差。

相对于(1.2)式，实际回归模型也可以写成

$$Y_i = a + bX_i + e_i \tag{1.5}$$

这里，e_i 是 $a + bX_i$ 的估计值 $\hat{Y}_i$ 与实际观察值 Y_i 的离差，即

$$e_i = Y_i - \hat{Y}_i$$

1.1.3 预测模型

实际预测时，残差项 e_i 是无法预测的，我们的目的是借助 $a + bX$ 得到预测对象 Y 的估计值，所以预测模型为

$$\hat{Y} = a + bX \tag{1.6}$$

式中：a 为回归常数，是回归直线的截距。其实际含义为，若在某一刻不考虑自变量时，因变量所能达到的数值。b 为回归系数，是回归直线的斜率。其实际含义为，当自变量 X 每变动一个单位时，因变量 Y 的平均变动量。

可以看出，(1.6)式 $\hat{Y} = \alpha + bX$ 实际上是(1.4)式 $Y = \alpha + bX + e$ 的主体部分。相对于(1.5)式，(1.6)式也可以写成

$$\hat{Y}_i = a + bX_i \tag{1.7}$$

1.1.4 参数估计

对(1.1)式中 α、β 进行估计，依照不同的准则，采用不同的统计方法，可以得到不同的数值，因而(1.4)式中的 a、b 不是唯一确定的。预测中，通常采用最小平方法[Least Squres]，亦称最小二乘法。其准则是，选择的参数 a、b 要使因变量 Y 的观察值 Y_i 与估计值 $\hat{Y}_i$ 之间的离差平方和最小，即 $\sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2 = \min$[①]。

采用最小二乘法得到 a、b 的计算公式为

$$\begin{cases} b = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ a = \bar{y} - b\bar{x} \end{cases} \tag{1.8}$$

式中：x_i 为自变量 X 的第 i 个观察值；y_i 为因变量 Y 的第 i 个观察值；n 为观察值的个数亦称样本数据个数；$\bar{x}$ 为 n 个自变量观察值的平均数；$\bar{y}$ 为 n 个因变量观察值的平均数。

(1.8) 式还可以写成

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ a = \bar{y} - b\bar{x} \end{cases} \tag{1.9}$$

例1.1 根据表 1.1 的数据，分析可能建立的预测国民收入中消费额的模型。

分析 设国民收入总额与国民收入中消费额分别作为自变量 X 和因变量 Y。将表1.1中有关数据绘制成图(见图1.1)。从图中可以看出，国民收入与消费之间大致是线性关系，计算它们之间的简单相关系数 $r_{xy} = 0.9954$，这说明，X 与 Y 之间，可以建立线性回归模型

$$\hat{Y} = a + bX$$

---

① 有关参数 a、b 的确定参见本章附录 1 - A。

表 1.1　　**国民收入总额与消费额**　　单位:亿元

年　份	国民收入总额(X)	国民收入消费额(Y)	年份	国民收入总额(X)	国民收入消费额(Y)
1952	589	477	1969	1617	180
1953	709	559	1970	1926	1 258
1954	748	570	1971	2077	1324
1955	788	622	1972	2136	1404
1956	882	671	1973	2318	1511
1957	908	702	1974	2348	1550
1958	1118	738	1975	2503	1621
1959	1222	716	1976	2427	1676
1960	1220	763	1977	2644	1741
1961	996	818	1978	3010	1888
1962	924	849	1979	3350	2195
1963	1000	864	1980	3688	2531
1964	1166	921	1981	3940	2799
1965	1387	982	1982	4261	3054
1966	1586	1065	1983	473	3358
1967	1487	1124	1984	5630	3895
1968	1415	1111	1985	6822	4820

资料来源:《中国统计年鉴(1986)》,中国统计出版社 1986 年版。

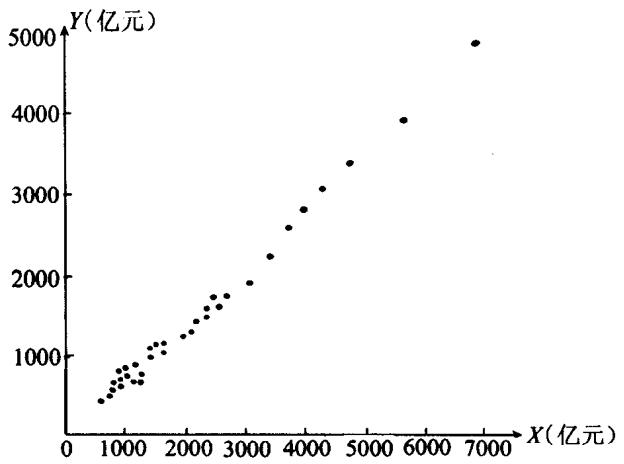

图 1.1　国民收入总额与消费额

根据表 1.1 中的数据，采用(1.9) 式，得到

$$b = 0.6835, \quad a = 31.5327$$

我国国民收入中消费额的预测模型可以是

$$\hat{Y} = 31.5327 + 0.6835X$$

这个模型表明，在 1952 ～ 1985 年 30 多年间，我国每增加 1 元的国民收入，平均就有 0.68 元用于消费。这就是回归系数 b 在这里所提供的经济意义。

模型中的参数，对不同的预测对象有不同的含义。参数估计值的符号和大小，要符合它的实际意义。例 1.1 中当国民收入增加时，消费额一般也有增加，因此 b 必须大于 0，若得到的估计值 b 小于 0，则模型应否定。b 的变动范围是否适当，主要根据预测人员的经验确定。参数估计值的符号和大小不符合其实际含义，其主要原因可能是：所选用的模型不能代表变量之间的关系；统计数据不足或口径不一致；违反了最小二乘法的某些假定。

预测模型中的回归系数 b，反映了因变量 Y 和自变量 X 之间的一种变动结构关系。这种变动结构对未来是否合适，决定着模型能否用于预测。这一点是预测时应该予以考虑的。

第二节　模型的检验

数理统计理论证明，采用最小二乘法得到的估计值 a、b 是 α、β 的最小方差无偏估计[①]，它们是较为理想和实用的估计值。在这一过程中，实际上我们是承认了几点假设：

(1) 变量 X 与 Y 之间为线性关系；

(2) 回归余项线性独立，即 $E(e_i e_j) = 0 (i \neq j)$；

(3) 回归余项服从正态分布，即 $e_i \sim N(0, \sigma^2)$。

当利用变量的样本数据(实际观察值) 建立起预测模型后，需要判断所做的各种假设的合理性以及模型的优劣。模型检验就是利用各种统计检验来判别模型的适用性。

1.2.1　回归系数的显著性检验

对于预测模型 $\hat{Y} = a + bX$，变量 X、Y 之间的线性假设是否合理，可以通过回归系数的显著性检验得到判别。回归系数的显著性检验由于要用参数的

① 有关证明参看参考书目[10]。

t 值，因而也称为参数的 t 检验。

检验假设

$$H_o : b = 0^{①}$$

计算参数 b 的 t 值

$$t_b = \frac{b}{S_b} \tag{1.10}$$

式中：S_b 是参数 b 的标准差，$S_b = S_y / \sqrt{\sum (x - \bar{x})^2}$②，这样，(1.10) 式可以写成

$$t_b = (b \cdot \sqrt{\sum (x - \bar{x})^2}) / S_y \tag{1.11}$$

式中：S_y 为回归标准差，$S_y^2 = \sum (Y - \hat{Y})^2 / n - 2$，$n$ 是样本数据个数，2 是参数个数。

$t_b = b / S_b$ 服从 t 分布，即 $t_b \sim t(n-2)$，因此，可以通过 t 分布表查得显著性水平为 α③，自由度为 $n-2$ 的数值 t_c。将 t_b 与 t_c 比较，可决定是接受还是否定 H_0 假设。

若 $| t_b | > t_c$，可以拒绝 H_0。它表明回归系数显著不为 0，参数的 t 检验通过。回归系数显著，说明变量 X 与 Y 之间的线性假设合理，这意味着，所选择的自变量能比较有效的解释预测对象的变化。

若 $| t_b | \leqslant t_c$，则不能拒绝 H_0，它表明回归系数为 0 的可能性较大，参数的 t 检验未通过。回归系数不显著，说明对于变量 X 与 Y 之间的线性假设不合理，意味着模型中的自变量无法较好地解释预测对象的变化，应重新考虑。

例 1.2　国民收入中消费额的预测模型的 t 检验(续例 1.1)。

分析　例 1.1 中建立的回归模型为

$$\hat{Y} = 31.5327 + 0.6835X$$

对这个模型的回归系数进行显著性检验。

$$t_b = 59.0858$$

查 t 分布表，$a = 0.05$，自由度 $df = 34 - 2 = 32$，得

---

① 检验回归常数 a 是否为 0 的意义不大，故通常只检验参数 b。

② 为方便，本书中凡有求和符号 $\sum$ 若无注明均表示 $\sum_{i=1}^{n}$。

③ 预测时，只需检验 b 是否为 0，故为双侧假设检验。查 t 分布表时，则应以显著水平为 $\frac{\alpha}{2}$ 查找。有时 t_c 也写成 $t_\alpha(n-2)$，或 $t_{\alpha/2}(n-2)$。

$$t_c = 2.0369$$

显然 $$t_b = 59.0858 > t_c = 2.0369$$

因此，参数的 t 检验通过。说明国民收入总额对消费额有显著影响。这一检验也可以采用 t_b 为 0 的概率大小作出判定。一般来说，$|t_b|$ 为 0 的概率小于 5%，则拒绝 H_0。

1.2.2 回归方程的显著性检验

参数的 t 检验，考察的是自变量 X 与因变量 Y 之间线性假设的合理性。但预测模型 $\hat{Y} = a + bX$ 作为一个整体，在一定程度上也反映了变量 X 与 Y 之间的统计线性关系，其是否适用于预测，仍需检验。回归方程的显著性检验，是利用方差分析所提供的 F 统计量，检验预测模型的总体线性关系的显著性，也被称为方程的 F 检验。

检验假设

$$H_0 : b = 0$$

计算回归方程的 F 值[①]

$$F = \frac{\sum(\hat{Y} - \bar{Y})^2/1}{\sum(Y - \hat{Y})^2/n - 2} \tag{1.12}$$

统计量 F 服从 F 分布，即 $F \sim F(1, n-2)$。在 F 分布表中，查找显著性水平为 α，自由度 $n_1 = 1, n_2 = n - 2$ 的 F 值 $F_\alpha(1, n-2)$。将 F 与 $F_\alpha(1, n-2)$ 比较，能够判定接受 H_0 还是否定 H_0。

若 $$F(1, n-2) > F_\alpha(1, n-2)$$

拒绝 H_0，回归方程较好地反映了变量 X 与 Y 之间的线性关系，回归效果显著，方程的 F 检验通过。这意味着，预测模型从整体上看适用。若

$$F(1, n-2) \leqslant F_\alpha(1, n-2)$$

不能拒绝 H_0，回归方程不能很好地反映变量 X 与 Y 之间的关系，回归效果不显著，方程的 F 检验未通过。这意味着，预测模型不能被采用。

例 1.3 国民收入中消费额的预测模型的 F 检验(续例 1.1)。

分析 对预测模型进行 F 检验，计算方程的 F 值

$$F(1, 34-2) = F(1, 32) = 3491.127$$

以显著性水平 $\alpha = 0.05$，自由度 $n_1 = 1, n_2 = 32$，查 F 分布表，得到

---

① 有关 F 检验的详细内容参见本章附录 1 - B。

$$F_{0.05}(1,32) = 4.17$$

显然，$F(1,32) = 3491.127 > F_{0.05}(1,32) = 4.17$，方程的 F 检验通过。预测模型 $\hat{Y} = 31.5327 + 0.6835X$ 在置信度为 95% 的情况下，回归效果显著，因而以此模型对消费额进行的预测，可靠性较高。

回归效果不显著或说回归方程的 F 检验通不过的原因可能是：某些对预测对象有影响的重要因素被忽略，未包括在方程中；变量 Y 与 X 之间不是线性关系；变量 X 与 Y 无关。究竟是何原因，需结合其他检验及具体情况而定。无论何种原因，所建的预测模型都不能用于预测，而应另行建立。

在简单回归分析中，t 检验与 F 检验的作用基本相同，这一点，从检验假设可以看出。但在多重回归分析中却有很大差异，在实际运用中应予以注意。

1.2.3 D.W 检验

上面的两个检验，能够对线性假设的合理性加以识别。对回归余项正态分布的假设，要求并不十分严格。因为虽然迄今尚无有效的统计检验，但由于它来自大量的、对因变量 Y 影响很小的不重要因素，因而当样本数据个数较多，其他假设得到满足时，可以认为回归余项服从正态分布。回归余项线性独立的假设是严格的。如果回归余项存在自相关，即回归余项之间不是相互独立，而是有较明显的相关关系，那么，应用最小二乘法就会产生很多问题，如回归系数的估计值虽然是无偏的，但不具有最小方差性质，t 检验、F 检验不再能严格应用等。为保证预测模型适用，需要对回归余项的线性独立假设进行检验。通常采用 D.W 检验，即序列的自相关检验。

设回归余项序列的自相关系数为 ρ。则检验假设

$$H_0 : \rho = 0$$

计算回归余项的 D.W 值即统计量 d

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{1.13}$$

根据给定的显著性水平 α，样本数据的个数 n 和自变量的个数 k 查找由 Durbin-Watson 建立的 D.W 表，得到下限值 d_L 和上限值 d_U。将计算的 d 与 d_L、d_U 比较，可以判定接受或否定 H_0。其判定标准如图 1.2 所示。

若 $d_U < d < 4 - d_U$，不能拒绝 H_0，回归余项无序列相关。

若 $0 < d < d_L$，拒绝 H_0，回归余项有正序列相关。