实用预测学

霍俊编著

第二册 因果分析 结构分析

中国发明创造者基金会
中国预测研究会

实 用 预 测 学

第 二 册

因果分析　结构分析

霍 俊 编 著

中国预测研究会
一九八四年六月

说　明

这本《实用预测学》第二册是作为两个大的方面进行编写的：因果预测分析（法）和结构预测分析（法）。事实上，结构分析是一种相互因果分析，前者是“因→果”，后者是“因⇄果”，即在各种因素之间是互为因果关系的。从数学方法上看，我们还可以把因果分析称为单方程模型分析。把结构分析称为多方程模型分析或联立方程模型分析。从预测应用上看，我们还可以把因果分析称为经济计量模型预测分析，把结构分析称为宏观经济计量模型预测分析。这本书，就是在我们早先编译的《经济计量模型与经济预测》、《宏观经济计量模型与预测》的基础上修改而成的，主要取材于《Econometric model and economic forecasts》R.S.Pindyck,D.L. Rubinfold, 1976年版一书。从多方面进行比较之后，我们认为这份材料很实用，故没有进行多大变动。但是，对初学者确实难处较大，必须在掌握了第一册中的常规预测方法和必要的数学准备之后才能看懂。如果只从应用上着眼，“照猫画虎”也是可行的。否则，只会用常规方法，就很难提高当前的预测水平了。

本册着重介绍建立和使用模型方面的科学和艺术。建模包括一组方法，其中绝大部份是定量方法。用这些方法来构造客观事物的数学表示式并对这些数学关系式进行检验。建模的艺术是难以用文字说明的，它主要是对建模过程进行直观判断。不幸的是，还没有明确的规则来对模型的建立作出判断。这就使得建模的艺术难以掌握。尽管如此，本册的目的之一是使读者尽可能了解建模艺术的性质。一方面通过实例和方法的讨论来达到这一目的，同时也鼓励读者通过建立它们自己的模型来掌握建模的艺术。

本册着重介绍企业、经济和社会科学中的模型，其中有综合模型、公司销售模型等。建立这些模型的直接目的是为预测和政策分析服务。但是读者应该记住这些内容的一般性质。

有人可能认为，还有许多通常用于预测和政策分析的模型，而本册并不包括全部模型类型和建模方法，只是集中讨论数据拟合模型及有关的检验理论。当然，也介绍了相当广泛的可供选择的模型。

1.为什么需要模型

许多人常常利用和进行这样或那样的预测，但是很少有人意识到，在对每一个社会或物质（实际）系统的预测和分析都内含着的某种逻辑结构和模型。例如，一个股票经纪人可能告诉你明年某种工业水平将上涨。他所以这样预测，因为他已经知道过去几年一直在上涨并且认为过去使之上涨的因素将继续存在并使之在将来也上涨。另一方面，他认为上涨是因为，他相信这个变量是通过一组复杂的关系与一组经济和政治变量相联系的结果。例如，在某种程度上与GNP国民生产总值和利率有关，由于他对GNP和利率未来的发展抱有最可能的信念，这就导致了他相信会上涨。

如果我们要用文字来说明股票经纪人进行预测的方法，我们可以说这是一种直观的。但是，股票经纪人肯定不会说，这是通过建立某种模型所进行的预测，的确，既没有写出数学方程式，也没有利用计算机，尽管如此，在每一种情况下，都内含着建立了某种隐式模型。

如果股票经纪人基于过去的增长，乐观地进行了有信心的预测。实际上他是构造了一个时间序列模型，并利用这个模型将过去的趋势外推到未来。如果预测是基于经济学知识，则是隐含着一个模型，这个模型包含由于过去的经验，股票经纪人思想上表达出来的不精确的各种关系。因此直观预测人员也构造了某种模式的模型，即使他本人并不意识到这一点。当然会有人提出，既然如此，为什么还需要显示模型来进行预测哩？回答是，利用显示模型具有某些好处。建立模型可以使人们清晰地思考和能够将全部重要的相关因素都能加以考虑。相信直观有时是危险的，因为可能忽略一些重要的关系或者没有正确地利用这样一些重要关系。另外，重要的是每一种关系都需要以这种或那种方式进行检验或确认。直观预测通常是做不到这一点的。但是建立模型的过程，不仅必须对整个模型进行检验或确认，还要对构成模型的每一个关系都要检验或确认。

当进行预测的时候，给预测使用者提供一个统计可信度，即预测精确度的某种程度，也是很重要的。单纯的利用直观方法通常不能对预测结果提供任何定量的可信量度。对构成模型的每一种关系和整个模型进行统计分析，就可以提供出模型预测的可信度。

一旦模型已经建立并对数据进行了拟合，就能利用敏感度分析来研究模型的许多性质。特别是，可以计算模型中各个变量的微小变化的影响。例如，用来描述和预测利率的模型，可以用来计算通货膨胀率的变化对利率的影响等。这种定量的敏感度研究不论是了解模型还是使用模型都是很重要的。这一点只有利用显式模型才能做到。

2.模型的种类

本册将讨论用于预测或政策分析的两类常用模型。每一类都包括模型的不同复杂程度和结构说明，每一类都设想是对现实世界过程的不同水平的理解。这二类模型是：

（1）单方程回归模型

这种情况是，研究的变量是由一些说明性变量组成的一个线性或非线性函数来说明的。这个方程常常是随时间变化的，也就是时间 t 将明显地出现在模型中，因此可以根据一个或多个说明性变量的变化来预测所研究的变量在所有时间的响应。

单方程回归模型的一个例子是关于具体利率（例如三个月国库券利率）与一组说明性变量（例如货币供应、通货膨胀率、和GNP的变化率等）关系的一个方程。回归模型通常被用来不仅仅对短期和长期利率的变化进行预测，也用来对其他经济和企业变量进行预测。

（2）多方程模拟（仿真）模型

研究的变量可以是一些说明性变量的函数，但是现在这些说明性变量以及研究的变量通过一方程组而彼此相关。仿真模型的建立是从说明一组单个关系开始的。其中的每一个关系式都要对可以利用的数据进行拟合，仿真是在某时间范围内联立求解这些方程的过程。

多方程仿真模型的一个例子是纺织工业模型。这个模型包括说明纺织品需求量、纺织品生产的产量、纺织工业中雇佣的生产工人数、纺织工业的投资以及纺织品价格等变量的方程式。这些变量是彼此相关的并且与其他一些变量（如国民收入、消费价格指数、利率等）有关。这些关系可以通过线性或非线性方程组来表示。假设已知国民收入、利率等变量的未来变化规律，就可以对未来进行仿真并获得每一个模型变量的预测值。这样的模型可以用来分析外部经济变量的变化对纺织工业的影响。

多方程仿真模型可以用来说明大量的具体过程的结构。它不仅仅能够解释单个的关系式，而且能够同时考虑所有相互关系之间的交叉影响。五个方程的仿真模型实际上比五个单个的

方程包含了更多的信息。五方程仿真模型不仅能够说明五个单独的关系，而且能够描述这些关系同时运行的动态结构。

选择模型的类型是比较困难的，要对时间、精力、费用、和所要求的预测精度之间进行权衡。建立一个多方程仿真模型可能要求花费大量的时间和金钱，这不仅是指实际工作，也包括利用计算机时间。从这种努力的结果所产生的收获可能包括对关系式和结构的更好的理解以及能够作出更好的预测。但是，在某些情况下，这样一些收获是不大的，以致花费是过重了。因为多方程模型需要对所研究的过程有很好的了解和具有这方面的大量的知识，所以建立这样的模型可能是相当困难的。

本册不仅要说明每一种模型如何建立和应用，而且要指出有关的费用和好处。但是这些都是一些困难的问题。例如，模型类型的选择常常是不清楚的。在任何情况下，都会有几种模型的讨论。

3．本册的内容

本册分二个部分，每一部分集中讨论一类模型。第一部分讨论最基本的一类模型，即单方程回归模型。用于建立单方程回归模型的经济计量方法，经过适当修改，可以用于建立多方程仿真模型和时间序列模型。

第一章是介绍回归分析的基本概念和双变量回归模型。接着第二章讨论多变量回归模型。同时也介绍了用于评价回归模型特性的统计试验和统计程序。

用于简单回归分析的估计方法要求对数据和模型作出某些假设。有时，这些假设被打破。第三章、四章将讨论不能满足假设时的一些情况。第三章是讨论异方差性和序列相关的问题，包括在这些情况下的统计检验以及对它们进行校正的估计方法。第四章是处理回归模型中说明性变量与误差项之间相关的问题，分析工具变量和二阶最小二乘估计方法。

第五章是讨论将单方程回归模型用于预测的目的。本章不仅叙述了进行预测的方法，而且也说明了预测值的可靠性的量度，例如置信区间和预测误差。

第一部分的最后一章即第六章，是回归模型的扩展。这章的内容具有更高级的性质，它将讨论规格误差、缺漏观察值、非线性估计、分布滞后模型以及抽样和时间序列数据共享等问题。第一部分介绍的经济计量学基础对于第二部分关于多方程模拟（仿真）模型的研究是基本的和必不可少的。本册第二部分，首先用了一章（即第七章）介绍联立方程模型的估计方法。这一章包括联立方程模型的识别和三阶最小二乘估计方法等。第八章、九章是介绍建立和利用多方程模型的方法。第八章是介绍仿真模型，包括仿真过程的讨论、仿真模型的评价方法、各种仿真模型的估计方法以及建立模型的一般方法。第九章的技术性更强，讨论了仿真模型动态特性的分析方法，其中包括模型的稳定性、动态响应与增益、以及仿真模型的调整和校正。这一章还讨论了敏感度分析和随机仿真。第十章详细介绍了建立和应用仿真模型的三个具体实例。第一个例子是建立一个比较小但是比较完整的经济模型，用于简单的政策分析。第二个例子是研究一个工业市场模型并说明了如何利用这个模型对生产和价格进行预测。最后一个例子是说明怎样将仿真技术用于一个公司的财务计划。

4．数学知识问题

数学推导和证明一般放在附录中或者是完全省去。第一部份关于矩阵形式的回归模型的讨论就放在附录中。因此，对于高年级大学生或大学毕业生来说是能够理解其中的绝大部份内容的。

要求读者应具有一些统计学的基本知识。虽然在附录中简要介绍了概率论和统计学知识，对于不具有一定统计学知识的人来说，在阅读某些部份时，可能会遇到困难。

5．一些特点

本册是试图从相当广泛的观点，来说明定量模型的研究和应用。大多数经济计量学教科书把单方程回归模型作为一个完全独立和孤立的部份。这样使读者常常就会这样认为，统计回归模型好像是与其他方面的模型是不同的和无关的。例如与模型的动态结构分析无关，也与时间序列分析用于预测无关。像任何实际工作者所知道的一样，情况并不是这样。例如，在研究多方程仿真模型的时候，不仅需要回归模型的知识，也需要知道由单个方程之间的相互影响产生的模型的动态行为。

本册讨论了建立两类模型的技术和方法。这样读者既可懂得单方程模型的单独应用，也可懂得将单方程作为研究多方程仿真模型的一个工具和作为研究随机时间序列预测模型的统计学基础。读者能够掌握到几类模型并能掌握到对于某一具体目的，而选用最适用的模型类型。

目 录

第一章　单方程一元回归模型

在第一章中将集中研究单方程一元线性回归模型。并通过对单方程一元线性回归模型的研究来介绍用经济计量方法进行因果分析。一开始先介绍线拟合的概念，接着提出线拟合的最小二乘准则并详细推导了最小二乘估计方法。为了对用最小二乘估计求得的单方程一元回归模型进行统计检验，介绍了经典回归模型的一些假设，分析了最小二乘参数估计值的统计性质。在某些假设条件下，最小二乘参数估计值是无偏、有效和具有一致性。利用估计参数的概率分布可以求置信区间和对模型进行统计检验。在这一章中还介绍了R^2的概念和回归模型的拟合优度的量度。

第一节　回归模型概述

一、线拟合

测定变量的数据，可能有许多来源和各种形式。描述变量在某一段时间的活动数据，称为时间序列数据，时间单位可能是日、周、月、季、年。描述变量在某一个时间点上的活动数据，称为抽样数据点（样本点）。抽样，是从能代表真实关系的基本母体中选择出来的。

我们从研究双变量模型入手，来讨论经济计量模型问题。双变量，是指一个自变量X与一个因变量Y；也称为一元，是指一个自变量X。为了用统计方式来描述这种关系，需要每一个变量进行一组观察，还要假设X和Y之间具有明确的数学形式。在此，假设X和Y之间的关系为线性，即X和Y之间的关系，可以用直线来描述。线性拟合，就是指用直线来描述数据点的分布规律。

我们要求的是，这种线拟合为最佳。例如，图1.1的数据点分布，可以用L_1、L_2或别的直线来表示。怎样来评定某直线的拟合为最佳呢？

一种方法是使数据点对称分布，线上偏差为正，线下偏差为负，各点偏差之和为零，如图1.2。这种方法，可以用来说明直线对数据点的拟合程度。但是，不能说明数据点对直线的离散程度。

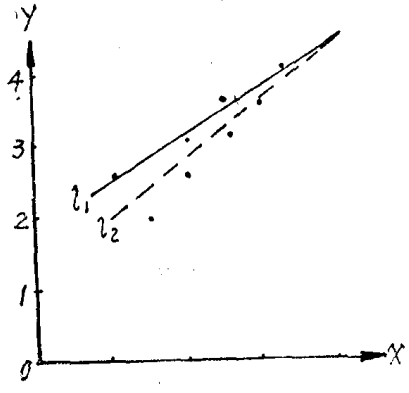

图 1.1

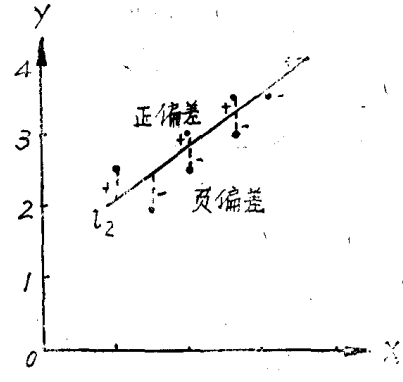

图 1.2

另一种方法是采用绝对值，使直线与数据点之间的偏差绝对值之和为最小，但是计算不方便。

最通用的方法是最小二乘法。最小二乘法，不仅能使数据点对称分布，还能使偏差的平方和最小，因而可以判断数据点对拟合直线的离散程度。

二、最小二乘法

建立统计关系的目的，通常是用来进行预测或说明一个或几个说明性变量（即自变量）的变化，对某一因变量所造成的影响。根据图 1.1 的数据点分布，可以写出如下线性方程

$$Y = a + bX$$

式中：Y——因变量；X——自变量。因为要根据X，来说明或预测Y的演变，要求拟合线与样本数据点之间垂直测定的偏差平方和为最小（见图 1.2），因此最小二乘法的准则是：

$$\mathrm{Min}\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2 \qquad (1.1)$$

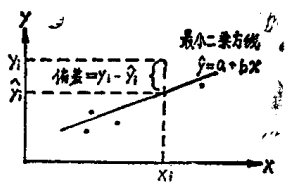

图 1.3

式中：$\hat{Y}_i = a + bX_i$ 是对应于具体观察X_i的Y的拟合值，N是观察次数，见图1.3。a和b的数值未知，需要根据最小二乘法准则（1.1）求系数a和b的数值。

求（1.1）式最小化的a和b的数值，利用普通微积分就可以做到。方法是将偏差平方和分别对a和b取偏导数，并令其等于零，然后解联立方程。

$$\frac{\partial}{\partial a}\sum(Y_i-a-bX_i)^2 = -2\sum(Y_i-a-bX_i) \qquad (1.2)$$

$$\frac{\partial}{\partial b}\sum(Y_i-a-bX_i)^2 = -2\sum X_i(Y_i-a-bX_i) \qquad (1.3)$$

使这些导数等于零，并除以－2，可得出：$\sum(Y_i-a-bX_i)=0$ (1.4)

$$\sum X_i(Y_i-a-bX_i)=0 \qquad (1.5)$$

改写（1.4）和（1.5）式，则可得出所谓正规方程的联立方程组：

$$\sum Y_i = aN + b\sum X_i; \qquad (1.6)$$

$$\sum X_iY_i = a\sum X_i + b\sum X_i^2 \qquad (1.7)$$

式中：Σ——求和符号；N——观察次数。现在就可以求出a和b，其方法是将（1.6）式乘以$\sum X_i$；（1.7）式乘以N：

$$\sum X_i\sum Y_i = aN\sum X_i + b(\sum X_i)^2 \qquad (1.8)$$

$$N\sum X_iY_i = aN\sum X_i + bN\sum X_i^2 \qquad (1.9)$$

由（1.9）式减去（1.8）式，得：

$$N\sum X_iY_i - \sum X_i\sum Y_i = b[N\sum X_i^2 - (\sum X_i)^2] \qquad (1.10)$$

由此；得：

$$b = \frac{N\sum X_iY_i - \sum X_i\sum Y_i}{N\sum X_i^2 - (\sum X_i)^2} \qquad (1.11)$$

已知b，则可由（1.6）式算出a：

$$a = \frac{\sum Y_i}{N} - b\frac{\sum X_i}{N} \qquad (1.12)$$

现在就可考虑在X和Y两者具有样本平均值为零时的特殊情况下，如何简化（1.11）和（1.12）公式，首先根据（1.12）式：$a = \bar{Y} - b\bar{X} = 0$ （1.13）式中，$\bar{Y}$和$\bar{X}$分别为Y和X的样本平均值。（1.13）式表明在X和Y的样本平均值为零时，拟合回归线的截距为零。为了在这种特殊情况下求出相应的斜率估计，我们将（1.11）式的分子和分母除以N^2：

$$b = \frac{\sum X_i Y_i/N - (\sum X_i/N)(\sum Y_i/N)}{\sum X_i^2/N - (\sum X_i/N)^2}$$

代入$\bar{X}$和$\bar{Y}$，

$$b = \frac{\sum X_i Y_i/N - \bar{X}\bar{Y}}{\sum X_i^2/N - \bar{X}^2}$$

因为$\bar{X} = \bar{Y} = 0$。所以

$$b = \frac{\sum X_i Y_i/N}{\sum X_i^2/N} \qquad (1.14)$$

（1.14）式比（1.11）式简单。这个事实说明，如果用样本平均值的偏差来表示变量，将使最小二乘法估计简化，而不论其平均值是否为0。为此，我们将数据都变换成偏差形式，方法是按各自平均值的偏差来表示对X和Y的每一次观察。我们将用小写字母x_i、y_i等来表示这种偏差形式，后面就不另做说明了。其定义如下：$x_i = X_i - \bar{X}$　$y_i = Y_i - \bar{Y}$　利用该定义，在一般情况下，可直接由（1.14）式得出斜率的最小二乘法估计，因为，x和y具有零平均值，因此，最小二乘法的斜率估计为：

$$b = \frac{\sum x_i y_i}{\sum x_i^2} \qquad (1.15)$$

变量转换成偏差形式见图1.4。在图1.4左图中，利用原始观察值绘出回归线，而图1.4右图则采用偏差形式。首先应注意两条回归线的估计斜率是相同的。由（1.15）式可以很明显看出，是用偏差形式的变量进行计算的。故图1.4右图的回归线截距等于零。这是由（1.13）式以及$\bar{X}$和$\bar{Y}$均等于零这一事实得出的。因此，假如我们决定使用偏差形式的数据，我们就变换回归线的原点，但是不改变斜率。最后，请注意图1.4右图中的线穿过原点。这就相当于图1.4左图中的线穿过平均值（$\bar{X}$，$\bar{Y}$）。

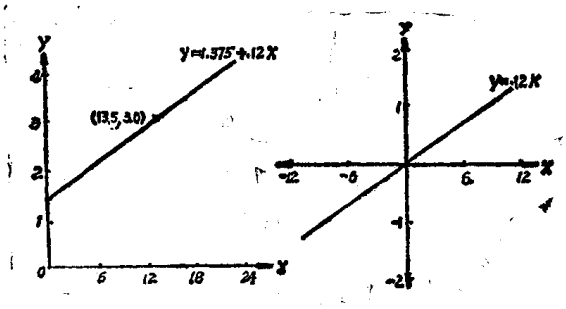

图1.4

在Y = a + bX模型中，斜率b是dY/dX的估计值，即Y的变化与X的变化之比。这使我们能够十分自然地解释回归斜率。另一方面，截距的解释则取决于那些接近X = 0的充分的观察次数，是否能用来产生具有统计意义的结果。如果是这样的话，那末，在X = 0时，我们就可以将截距解释为Y的一个估计值。但是，如果没有充分的观察次数，那末，截距就只是最小二乘法回归线的高度。

其次，假如我们估计了因果关系相反的回归线X = A + BY，则估计斜率将为：

$$B = \frac{\sum x_i y_i}{\sum y_i^2}$$

如果估计过程都相同的话，那末，相应的回归斜率必须等于原回归斜率的倒数，即：$B = 1/b$。但是，这种同一性只有在罕见的情况下，即：$(\Sigma x_i y_i)^2/(\Sigma x_i^2 \Sigma y_i^2) = 1$时才能成立。这种情况只有在全部样本点都位于拟合回归线上时才发生。

例：衣着费用支出实例

假定衣着费用支出与消费品总费用支出有关。更准确地说，希望通过对全部消费品支出的了解，能够对衣着费用支出进行预测。在法国曾经组织过这类研究，并获得112个居住在省城的低级官员的家庭抽样数据。所测定的两个变量为衣着费用支出Y和全部消费支出X，拟合直线为$Y = a + bX$。因为，当时相信每个人都是将衣着费用支出作为既定总开支的一个直接函数。这个估计方程式为：$\hat{Y} = 1.78 + 0.12X$

根据正的斜率系数，可得出这样的结论：该数据符合衣着费用支出与总支出之间具有直接关系的假设。

第二节　模型的随机性质

我们的目的在于了解回归模型的概率性质。为了做到这点，我们分析如下事实：当我们给定某一观察值X（自变量）时，就会观察到许多可能的数值Y（因变量）。举例来说，每年收入相同的人，每人的消费方式就不同。又由于每人所处的环境有变化，每个人每年的消费方式也可能有所改变。因此一般都假定：对应于观察值X（收入），Y值（如食品购买力）是不同的，是随机的。为了用公式来描述这种情况，我们给模型增加了一个随机误差分量，可写成：

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (1.16)$$

式中，每次观察时，Y是一随机变量；X是确定的(实验者知道)；ε是一随机误差项，其数值要根据概率分布。请注意，我们已改变符号，利用希腊字母来表示回归参数，因为，模型现在已含有一个随机误差项。

误差项是由于若干作用力的互相影响而出现的。第一，误差的出现是由于模型对现实的简化所造成的。例如，我们假定，价格是某一产品需求的唯一决定因素。事实上，有一些与需求有关的变量被省略了，如个人喜好、人口、收入以及气候，都可以归入误差项。假如这些被省略的影响很小，那末，假定误差项是随机的才算合理。出现误差的第二个原因，是与数据的收集和测定有关。经济数据和企业数据通常都是难以测定的。例如，某个商行可能不愿意放弃明确的成本数据，这样，他们就不会得到成本的无误差估计。必须将误差项ε和残差($\hat{\varepsilon}_i = Y_i - \hat{Y}_i$)，即和因变量与其拟合值的偏差加以区别。误差$\varepsilon$是与真实回归模型有关，而残差则是由估计过程产生的。每一个X值都有ε的概率分布，因此，就会有Y值的概率分布。可见图1.5双变量回归模型的概率分布。

现在我们就能够利用一些重要的假设来完全规定一元线性回归模型。

i，如（1.16）式所示，Y和X之间的关系为线性。

ii，X_i为非随机变量，其数值是确定的。

iii，（a）对所有观察来说，误差项具有零期望值和常数方差，即$E(\varepsilon_i) = 0$和$E(\varepsilon_i^2) = \sigma^2$

（b）就统计而言，随机变量ε_i是不相关的，即：对应于不同观察的误差具有零相

关。因此，i≠j时，$E(\varepsilon_i\varepsilon_j) = 0$。

遵循上述假设的模型称为经典的线性回归模型。

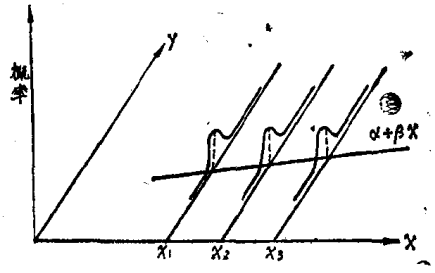

图 1.5

关于把X作为确定的非随机变量的假设，相当于自变量X完全可以由研究人员所控制，能够根据实验目的，而改变X的数值。这对于研究大多数的企业问题来说，是不现实的。因此，这种假设仅仅是为了说明的目的。最小二乘法回归分析的绝大多数结果，对于比这个假设更一般的假设也是有效的。但是，它的证明要困难得多。

为了方便起见，误差项具有零期望值的假设，可以部分地做出。为了理解这点，假定省略的误差的平均影响等于α'[即：$E(\varepsilon_i) = \alpha'$]。然后，将双变量模型改写为

$$Y_i = \alpha + \beta X_i + \varepsilon_i + (\alpha' - \alpha') = (\alpha + \alpha') + \beta X_i + (\varepsilon_i - \alpha') = \alpha^* + \beta X_i + \varepsilon_i^*$$

其中，$\alpha^* = (\alpha + \alpha'), \varepsilon_i^* = (\varepsilon_i - \alpha'), E(\varepsilon_i^*) = E(\varepsilon_i - \alpha') = E(\varepsilon_i) - \alpha' = \alpha' - \alpha' = 0$

如果误差项有一常数方差（根据假设），称为同方差性。但是，如果方差发生变化，则称为异方差性。异方差性可能会在研究某一工业企业的抽样时出现。图1.6是说明异方差性的两例。在图 1.6 右图中，误差项的方差随着X值的增加而减少，但在图1.6左图中，误差方差随X 增大而增加。

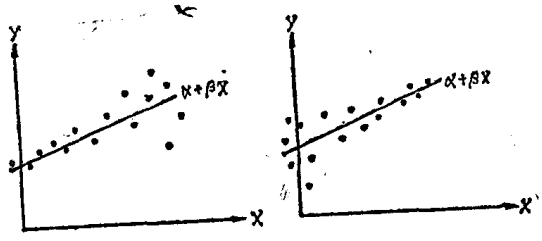

图 1.6

对应于不同观察的误差都不相关的这一假设，在时间序列和抽样研究中是很重要的。当不同观察的误差相关时，我们说，误差过程是序列相关（自相关）。图 1.7 说明了时间序列研究中的正、负序列相关的情况。负序列相关是指某一时间周期的负误差，伴随着下一时间周期的正误差，反之亦然。因此，在图 1.7 左图中，数据点在实际回归线上下呈某种规则性波动起伏。另一方面，当发生正序列相关时，某一周期的正（负）误差将趋于同下一周期的正（负）误差相联系。因此，在图 1.7 右图中，误差的图形是，先为负的（X的低数值），再为正的（X的高数值）。

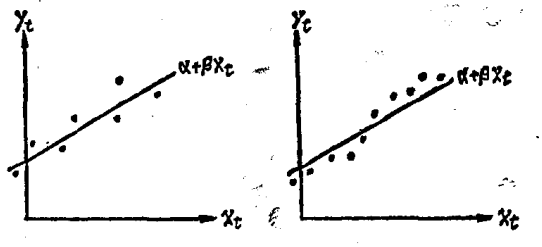

图 1.7

应该注意，作为假设ii与iii(a)的推论，我们可以毫无疑问地假定误差项是与X值不相关的。这是由于X值是非随机的这一假设推出的。因而：$E(X_i\varepsilon_i) = X_iE(\varepsilon_i) = 0$

如果将此分析推广到X值是随机的（也就是说，X值由概率分布任意得出的）变量时，也将需要这种假设。此外，假设iii(a)和iii(b)使我们得出的结论是，任何样本中的误差总和的期望值同样都为零，即：$E(\sum\varepsilon_i) = \sum E(\varepsilon_i) = 0$

不要把这个结果同iii(a)的说明混为谈。$E(\varepsilon_i)=0$的意思是，与某一个X_i相联系的Y的重复取样误差的期望值为0。为了得到此结果，先将某X值加以固定，然后从已知概率分布的母体中求出随机误差的样本。假设这些误差样本的期望值为0。

在我们研究随机模型的估计程序之前，有两种情况特别值得提出。第一，我们曾经说过，每个扰动即误差项都具有常数方差σ^2。当然，方差是一个未知参数，必须作为回归模型的一部分加以估计。因此，此处所叙述的随机回归模型具有三个未知参数，而在前面的线拟合模型中只具有两个未知参数。第二，根据曾经叙述过的与误差ε有关的模型假设，能够容易地写出与变量Y的概率分布有关的假设如下：

iii(a′) 随机变量Y的期望值为$\alpha+\beta X$和方差为σ^2：

$$E(Y_i)=E(\alpha+\beta X_i+\varepsilon i)=\alpha+\beta X_i+E(\varepsilon)=\alpha+\beta X_i$$

(b′) 随机变量Y_i是不相关的。

第三节 估计量的统计性质

为了对线性回归模型进行统计检验，我们需要确定误差项的概率分布。对于经典的正态线性回归模型，我们假定：

iii(c)误差项是正态分布的。

这一假设对模型的统计检验是很重要的。如果相信由于测定和省略所造成的误差都很小，而且相互独立，那末，正态性假设就是一种合理的假设。由误差项ε是正态分布的这一假设，可以得出，因变量Y也是正态分布的结论（因为Y_i是ε_i的线性组合，而每个ε_i都是正态分布的）。

现在，我们就能回到，选择随机线性回归模型的估计参数这一问题上来。我们的目标是拟合回归线$\hat{Y}=\hat{\alpha}+\hat{\beta}X$。这条回归线，在某种意义上讲，接近真实回归线。我们曾叙述过使样本点至回归线的偏差平方和最小化的方法。现在我们希望能检验该估计线与真实回归线的差别。为此，我们首先要描述估计参数的一些有用的统计性质。先进行介绍性的讨论，然后再详细研究最小二乘法估计量的统计性质。我们主要注意参数β估计量的性质，但是，这个讨论可以应用于所有的估计参数。

为了更清楚地了解这一点，请注意，我们分析的模型，其中自变量的N个值均假定为确定的。假若我们选择与自变量X值有关的Y为观察样本，那末，我们就能获得回归斜率的一个估计量，它是以已经出现的Y的具体观察值为基础的。假若我们以相同的X值重复该实验，我们就会得出Y的一组新的观察值（因为ε在新样本中是不同的），这样又能得出斜率参数的新估计值。如果我们能得到Y变量的充分样本，那末，我们就会得出参数β估计量的分布。由于每个Y的观察样本都产生一个β的具体估计量。当我们取Y值的样本时，就相当于从样本估计量分布中取β的一个估计值。

当然，这将会出现一个样本估计量的分布，它与从样本数据估计β的各种可能的方法有关。但是，我们现在的目的是研究最小二乘法估计程序所产生的分布（请注意，前面已求出$\hat{\beta}=\sum x_i y_i/\sum x_i^2$），以便评定最小二乘法是否是估计$\beta$的一种好的方法。重要的是，要认识到符号$\hat{\beta}$通常用于两种目的：一种是指由某一特定样本得出的斜率估计量；另一种是按照概率分布的估计量（与最小二乘法估计有关）。

一、无偏

只要我们能够对Y重复实验许多次，就能得到观察值的新的集合，我们就可以肯定，平均来说是正确的。如果$\hat{\beta}$的平均值或期望值等于真实值时，即$E(\hat{\beta})=\beta$，就说$\hat{\beta}$是一个无偏估计量。有偏与无偏估计量的区别可见图1.8，请注意，大小为N的样本，$\Sigma X_i/N$为母体真实平均值的一个无偏估计量，而$\Sigma(X_i-\bar{X})^2/(N-1)$则是母体真实方差的一个无偏估计量。为了说明起见，我们将估计参数的偏移定义如下：偏移$\text{Bias}=E(\hat{\beta})-\beta$

重要的是，应该认识到，一个估计量无偏是一种很理想的性质。因为，无偏意味着，估计量对真实参数的离中趋势没有意义。一般来说，人们希望估计量是无偏的，而且还希望估计量平均值周围的离中趋势很小。因此，我们就需要确立第二条准则，使我们能够在无偏估计量中间进行选择。

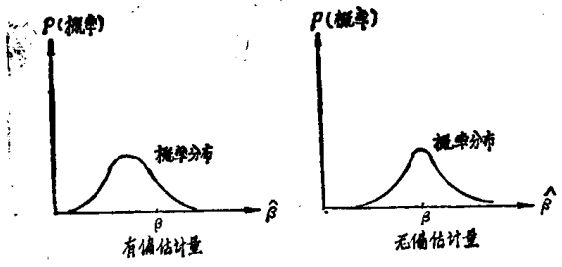

图 1.8

二、效率

假若一个无偏估计量$\hat{\beta}$的方差小于任何其他无偏估计量的方差，我们就可以认为$\hat{\beta}$是一个有效的无偏估计量。实际上，有时难以说明估计量是否有效，因此，按照它们有关的效率来描述估计量是很自然的。如果某一估计量的方差较小，那末，它就比其他估计量会更有效。比较有效和比较无效的估计量可见图1.9。效率是一个理想的目标，因为，估计过程越有效，人们对估计参数所作的统计说明就越有力。因此，在方差为零的（无偏）估计量这种极端情况下，我们可以肯定地指出真实回归参数的数值。

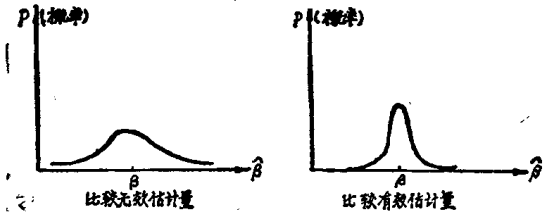

图 1.9

三、最小平均方差

有许多场合，人们不得不权衡估计量的方差和偏移。例如，当模型的目的是使预测的精度最高，那末，具有很小的方差但有某些偏移的估计量，可能比具有较大方差的无偏估计量要更理想。在这方面，有一条有用的准则，目标是使平均方差最小。其定义如下：平均方差$(\hat{\beta})=E(\hat{\beta}-\beta)^2$

可以证明平均方差$(\hat{\beta})=[\text{Bias}(\hat{\beta})]^2+\text{Var}(\hat{\beta})$

$$E(\hat{\beta}-\beta)^2=E[(\hat{\beta}-\bar{\hat{\beta}})+(\bar{\hat{\beta}}-\beta)]^2=E(\hat{\beta}-\bar{\hat{\beta}})^2+[E(\hat{\beta})-\beta]^2$$

$$+2(\bar{\hat{\beta}}-\beta)E(\hat{\beta}-\bar{\hat{\beta}})=\text{Var}(\hat{\beta})+[\text{Bias}(\hat{\beta})]^2\text{因为}E(\hat{\beta}-\bar{\hat{\beta}})=E(\hat{\beta})-\bar{\hat{\beta}}=0$$

其中$\bar{\hat{\beta}}$是估计回归系数β的期望值。根据定义$\bar{\hat{\beta}}=E(\hat{\beta})$。式中：Bias——偏移；Var——方差。

因此，平均方差最小化的准则，考虑了估计量偏移的平方及估计量的方差。如$\hat{\beta}$是无偏的，则$\hat{\beta}$的方差与平均方差相等。

四、一致性

当样本规模变大时，估计量就具有大样本的性质。我们希望估计量$\hat{\beta}$在某种意义上能在样本增大时接近真实β，这就是说，希望$\hat{\beta}$分布能收敛于β。更准确地说，我们希望，在样本无限变大时，$\hat{\beta}$不同于β的概率将非常小。为了将这种概率的概念应用于估计量选择，我们将$\hat{\beta}$的概率极限(Plim$\hat{\beta}$)定义如下：

假使N趋近无限大时，Plim$\hat{\beta}$等于β，即$|\beta-\hat{\beta}|$小于任意小正数的概率将接近于1。根据这个概念，很自然就能将一致性的准则规定如下：

如$\hat{\beta}$的概率极限为β，则$\hat{\beta}$为β的一致估计量。[严格说，对任意$\delta>0$，在概率极限$\hat{\beta}$收敛于β时$\lim_{N\to\infty} P(|\beta-\hat{\beta}|<\delta)=1$]

概言之，假若当样本任意变大时，估计量的概率分布压缩到一个点上（真实参数），那末这个估计量就是一致的。可见图1.10。

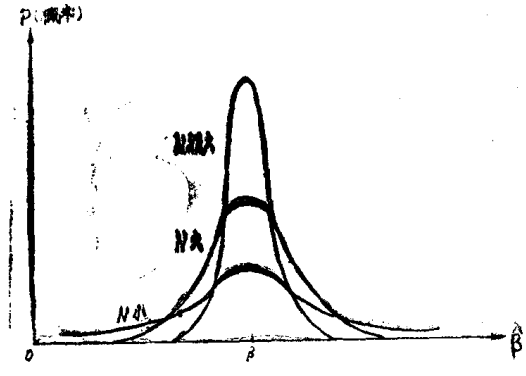

图 1.10

一般说来，经济计量学家倾向于将一致性而不是将无偏作为估计准则。一个有偏、然而一致的估计量，平均来说，可能不等于真实参数，但是，在样本数据变大时，这个估计量将近似真实参数。例如，规模为N的样本，$\Sigma(X_i-\bar{X})^2/N$就是母体方差的一个有偏而一致的估计量。从实际的观点来看，这要比得出一个无偏的、然而在样本变大时却继续大大地偏离真实参数的参数估计量更使人放心。

将估计量的平均方差，在样本增大时接近零这个目标当作一致性的选择准则，这是很自然的。平均方差准则的意思是，估计量是渐近无偏的，并且其方差在样本变得很大时接近于零。结果，平均方差接近零的估计量将是一个一致估计量。但是，逆推就不一定正确了。在多数实际应用中，一致估计量确实都具有接近零的平均方差，而且这两个准则是互相通用的。

第四节　最佳线性无偏估计

现在我们要回到最小二乘法估计量的问题。由于最小二乘法估计量与随机回归模型$Y_i=\alpha+\beta X_i+\varepsilon_i$有关，我们将这种估计量用$\hat{\alpha}$和$\hat{\beta}$表示。采用最小二乘法的最重要理由是由经典的高斯—马尔可夫定理给出的。根据这个定理，$\hat{\alpha}$和$\hat{\beta}$都是α和β的最佳（最有效）的线性无偏估计量。因为，它们与所有无偏估计量相比，具有最小的方差。重要的是要认识到高斯—马尔可夫定理不适用于非线性估计量。[如果$\hat{\beta}$能写成ΣC_iY_i，其中每个C_i都是一个常数，这样就能说$\hat{\beta}$与Y是线性关系。但是，如果$\beta'=\Sigma C_i\log Y_i$，那么$\beta'$就是Y的非线性估计量，非线性估计可能出现于最小二乘法以外的其他准则。]非线性估计量可以是无偏的，并且具有比最小二乘法线性估计量要小的方差。这一点是可能的。更为普遍的是，估计量可以有偏的，而其平均方差却比无偏估计量的要小。这就说明，会出现这样的情况，人们希望在选择估计程序时，利用的是客观的而不是“最佳、线性无偏的”估计程序。有偏但平均方差最小的非线性估计量，目前正在研究，看来很有用处。

在这方面，我们不想证明高斯—马尔可夫定理，但是，求出最小二乘法估计量的平均值和方差的表示式还是很有帮助的。为了把问题简化，我们将采用的数据变换成平均值的偏差形式，即采用$x_i = X_i - \bar{X}$和$y_i = Y - \bar{Y}$。请注意，真实回归线$E(y_i) = \beta x_i$以及估计斜率为：$\hat{\beta} = (\sum x_i y_i / \sum x_i^2)$ (1.17)

为简化以后的推导：令：$c_i = \dfrac{x_i}{\sum x_i^2}$

每一个C_i都是一个确定的常数（即非随机），因为样本中的X都是确定的。代入（1.17）式，可得出：$\hat{\beta} = \sum c_i y_i$

该式将估计斜率参数表示为因变量观察值的加权和。我们将用该表示式清楚地导出$\hat{\beta}$的平均值和方差。首先，$\hat{\beta} = \sum c_i y_i = \sum c_i(\beta x_i + \varepsilon_i) = \sum c_i \beta x_i + \sum c_i \varepsilon_i$ (1.18)

因而，$E(\hat{\beta}) = \sum c_i \beta x_i + \sum c_i E(\varepsilon_i)$，由于，$E(\varepsilon_i) = 0$，所以，$E(\hat{\beta}) = \sum c_i \beta x_i = \beta \sum c_i x_i = \beta$

$\sum c_i x_i = 1$这个事实是直接由c_i的定义而推出的。如：

$$\sum c_i x_i = \sum\left[\left(\frac{x_i}{\sum x_i^2}\right)x_i\right] = \sum\left[\frac{x_i^2}{\sum x_i^2}\right] = \frac{\sum x_i^2}{\sum x_i^2} = 1$$

因此，$\hat{\beta}$为β的无偏估计量。现在就可以考虑$\hat{\beta}$的方差：$Var(\hat{\beta}) = E(\hat{\beta} - \beta)^2$

但是，由（1.18）式，$(\hat{\beta} - \beta) = \sum c_i \beta x_i + \sum c_i \varepsilon_i - \beta = (\sum c_i x_i - 1)\beta + \sum c_i \varepsilon_i = \sum c_i \varepsilon_i$

所以，$(\hat{\beta} - \beta)^2 = (\sum c_i \varepsilon_i)^2$

以及$Var(\hat{\beta}) = E[\sum(c_i \varepsilon_i)]^2$

$$= E[(c_1\varepsilon_1)^2 + 2(c_1c_2\varepsilon_1\varepsilon_2) + (c_2\varepsilon_2)^2 + \cdots$$

但是，按照假设，ε_i是互不相关的；这就是说，$E(\varepsilon_i\varepsilon_j) = 0 (i \neq j$时$)$。

所以，$Var(\hat{\beta}) = E(c_1\varepsilon_1)^2 + E(c_2\varepsilon_2)^2 + \cdots = c_1^2E(\varepsilon_1^2) + c_2^2E(\varepsilon_2^2) + \cdots = \sum c_i^2 E(\varepsilon_i^2) = \sigma^2 \sum c_i^2$

由于，$\sum c_i^2 = \dfrac{\sum x_i^2}{(\sum x_i^2)^2} = \dfrac{1}{\sum x_i^2}$

所以，$Var(\hat{\beta}) = \sigma^2 / \sum x_i^2$ (1.19)

采用类似的推导，我们得出截距项估计量的平均值和方差为

$$E(\hat{\alpha}) = \alpha \tag{1.20}$$

$$Var(\hat{\alpha}) = \sigma^2\left[\frac{\sum X_i^2}{N\sum(X_i - \bar{X})^2}\right] \tag{1.21}$$

最后，还要考虑$\hat{\alpha}$和$\hat{\beta}$之间的协方差。再次采用类似的推导，我们可证明

$$Cov(\hat{\alpha}, \hat{\beta}) = \frac{-\bar{X}\sigma^2}{\sum x_i^2} \tag{1.22}$$

已知最小二乘法估计量的平均值和方差及其协方差的数据，我们就可以讨论线性模型的统计检验。首先，重要的是要记住：由于$\hat{\beta} = \sum c_i y_i$以及$y_i$都是正态分布的，因而估计量$\hat{\beta}$也都是正态分布的（正态分布的自变量的线性组合也是正态分布的）即使y_i不是正态分布的，根据统计学的中心极限定理（概略地说，中心极限定理说明，当样本规模无限增大时，一个独立分布变量的样本平均值将趋于正态分布。该定理可应用于$\hat{\beta}$，因为$\hat{\beta}$是y_i的线性组合）。$\hat{\beta}$的分布在合理条件下可表现为渐近于正态。因此，我们尽管没有假定正态性，但是获得的结果对大样本来说，将近似正态。在这方面总结一下，可写成：$\hat{\beta} \sim N\left[\beta, \dfrac{\sigma^2}{\sum x_i^2}\right]$ (1.23)