# 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。

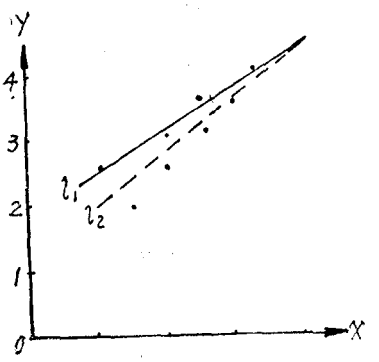


图 1.1

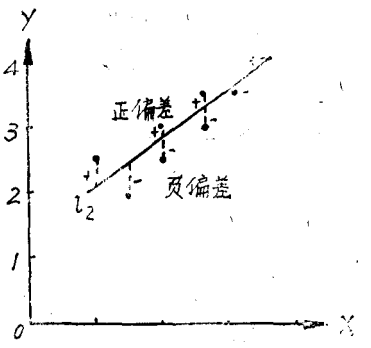


图 1.2

需要全本请访问 www.ertongbook.com

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：

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

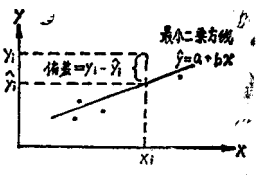


图 1.3

式中：$Y_i = a + bX_i$ 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：

$$\sum Y_i = aN + b\sum X_i; \qquad (1.6) \quad \sum X_iY_i = a\sum X_i + b\sum X_i^2 \qquad (1.7)$$

式中：$\sum$——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：$N\sum X_iY_i - \sum X_i\sum Y_i = b[N\sum X_i^2 - (\sum X_i)^2]$ (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：$$= \frac{\sum Y_i}{N} - b\frac{\sum X_i}{N} \qquad (1.12)$$

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：

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

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（$\bar{X}$，$\bar{Y}$）。

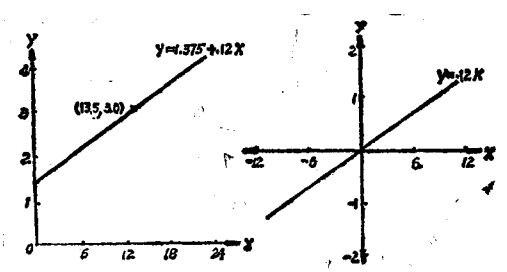


图 1.4

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：$(\sum x_i y_i)^2/(\sum x_i^2 \sum y_i^2) = 1$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：$Y_i = \alpha + \beta X_i + \varepsilon_i$　　(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)；$\varepsilon$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 $\varepsilon$ 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$\varepsilon$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$E(\varepsilon_i^2) \approx \sigma^2$

（b）就统计而言，随机变量$\varepsilon_i$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。

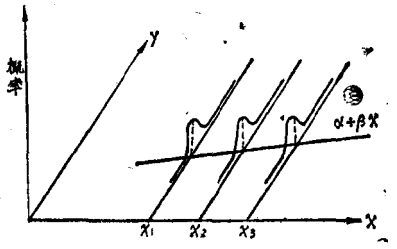


图 1.5

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$\alpha'$[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。

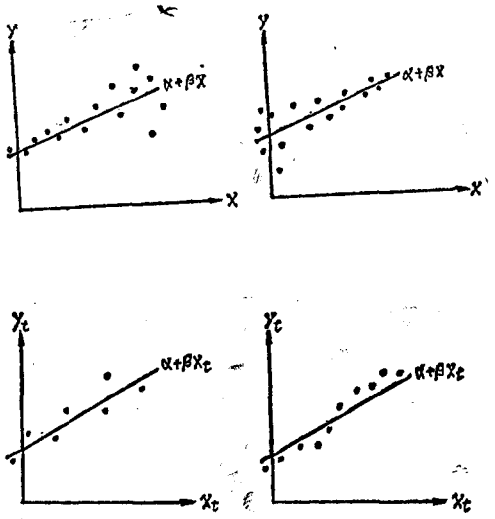


图 1.7

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）。

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：$E(\sum\varepsilon_i) = \sum E(\varepsilon_i) = 0$

需要全本请访问 www.ertongbook.com

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$\sigma^2$。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$\varepsilon$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$\sigma^2$：

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

(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$\varepsilon$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$\varepsilon_i$的线性组合，而每个$\varepsilon_i$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$\beta$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$\varepsilon$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$\beta$估计量的分布。由于每个Y的观察样本都产生一个$\beta$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$\beta$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$\beta$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。

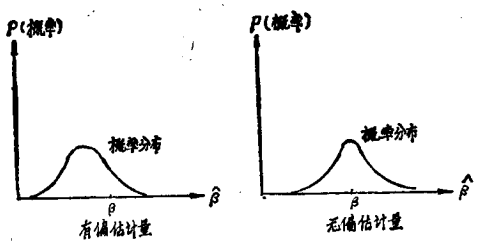


图 1.8

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。

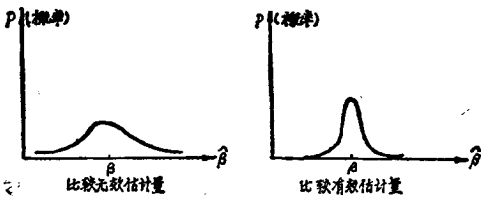


图 1.9

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$(\hat{\beta}) = [\text{Bias}(\hat{\beta})]^2 + \text{Var}(\hat{\beta})$

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

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

其中$\bar{\hat{\beta}}$是估计回归系数$\beta$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$\beta$，这就是说，希望$\hat{\beta}$分布能收敛于$\beta$。更准确地说，我们希望，在样本无限变大时，$\beta$不同于$\beta$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$(Plim\beta)$定义如下：

假使N趋近无限大时，$Plim\hat{\beta}$等于$\beta$，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$\beta$，则$\hat{\beta}$为$\beta$的一致估计量。[严格说，对任意$\delta>0$，在概率极限$\hat{\beta}$收敛于$\beta$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。

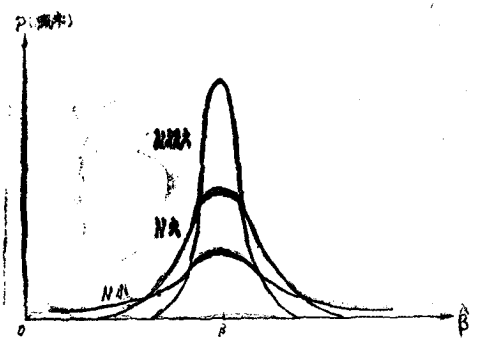

图 1.10

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，$\sum(X_i-\bar{X})^2/N$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$\alpha$和$\beta$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$\sum C_iY_i$，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$\beta'=\sum C_i\log Y_i$，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，我们不想证明高斯一马尔可夫定理，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$x_i = X_i - X$和$y_i = Y - Y$。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：$c_i = \frac{x_i}{\sum x_i^2}$

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：$\beta = \sum c_i y_i$

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$\beta$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，$\varepsilon_i$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，$\sum c_i^2 = \frac{\sum x_i^2}{(\sum x_i^2)^2} = \frac{1}{\sum x_i^2}$

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

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

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

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

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

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：$\hat{\beta} \sim N\left[\beta, \frac{\sigma^2}{\sum x_i^2}\right]$ (1.23)

$$\hat{\alpha} \sim N\left[\alpha, \sigma^2\left(\frac{\sum X_i^2}{N\sum x_i^2}\right)\right] \tag{1.24}$$

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X}\sigma^2}{\sum x_i^2} \tag{1.25}$$

请注意，$\hat{\beta}$的方差直接随$\varepsilon$的方差而变化。因此，在其他情况相同时，我们就能在误差项的方差较小时获得更为精确的斜率参数估计量。另一方面，$\hat{\beta}$方差随$\sum x_i^2$成反比变化。因此，$X_i$方差越大，对$\beta$的估计就可能更好一些。简言之，当X的全部数据限于较小的范围时，就难于准确地确定回归线的斜率。

尽管$\hat{\alpha}$的方差表示式看起来十分复性，但是我们要注意：当X平均值等于零时，表示式简化为$\sigma^2/N$，在这种情况下，$\hat{\alpha}$方差为最小值。这一结果在应用中意义不大，因为，截距参数通常不很重要。还要注意：$\alpha$和$\hat{\beta}$协方差的符号与$\bar{X}$的符号相反。假若X的平均值为正，则$\hat{\alpha}$的过高估计可能与$\hat{\beta}$的过低估计有关。

我们的分析还没有结束，因为我们需要利用样本数据来获得真实母体方差$\sigma^2$的估计量。这种估计量的实际推导十分复杂，此处就不再证明。我们将用以下的真实方差$\sigma^2$的样本估计量：

$$S^2 = \hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{N-2} = \frac{\sum (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2}{N-2} \tag{1.26}$$

式中，$\hat{\varepsilon}_i = Y_i - \hat{Y}$为回归残差，$S^2$为残差方差，可作为真实方差$\sigma^2$的无偏和一致的估计量。（S称为回归的标准误差，SER、SE和SEE有时也都用来表示S，这个字母是估计标准误差的缩写）读者可能会问：为什么残差平方和$\sum \hat{\varepsilon}_i^2$要被$N-2$来除，得出真实方差的无偏估计量。回答是在于以下事实：在估计过程中，具有N个数据点，但是，斜率和截距的估计为数据提出了两个约束条件。这样就剩下$N-2$个无约束的观察数据，用其来估计残差方差。故$N-2$的除数被称为自由度数。

利用$\sigma^2$的估计量$S^2$，我们可以通过（1.23）式至（1.25）式获得估计参数$\hat{\alpha}$和$\hat{\beta}$的方差样本估计量以及估计参数$\hat{\alpha}$、$\hat{\beta}$之间的协方差估计量。这三个估计量可分别表述如下：

$$S^2_{\hat{\beta}} = \frac{S^2}{\sum x_i^2} \tag{1.27}$$

$$S^2_{\hat{\alpha}} = S^2\left(\frac{\sum X^2}{N\sum x_i^2}\right) \tag{1.28}$$

$$\mathrm{Cov}\ (\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X}S^2}{\sum x_i^2} \tag{1.29}$$

$S_{\hat{\beta}}$和$S_{\hat{\alpha}}$是估计参数$\hat{\beta}$和$\hat{\alpha}$的标准误差，它们分别提供估计量的离散量度。

## 第五节假设检验与置信区间

已知$\hat{\alpha}$和$\hat{\beta}$的分布，就可以确定置信区间以及检验关于估计参数的假设。置信区间即区间估计量，能提供可能包含有真实回归参数的一个范围。对每一个置信区间，我们都使其带有一个统计的显著水平。已知显著水平，就可能确定置信区间，以便使该区间包括真实回归参数的概率等于1减去显著水平。置信区间对于检验关于估计回归参数的统计假设特别有用。最常用的检验是虚假设。一般来说，虚假设是一种说明或假设，它假定不会出现某种影响。由于研究人员希望能接受模型，通常虚假设的构成都是为了使假设本身能被否定。举例来说，

如果要想说明按照总的个人收入DI研究总消费支出C，则合理的模型规格将为$E(C)=\alpha+\beta(DI)$。我们期望消费与个人收入互相有正的关系，这样，$\beta$就必须为正。但是，为了检验模型的有效性，我们提出$\beta$等于0的虚假设。我们希望否定虚假设，所用的方法是得出一个$\hat{\beta}$值，这个$\hat{\beta}$值大大超过0，这样就能对$\beta$等于0的虚假设产生显著怀疑。例如，假定$\hat{\beta}$被估计为0.9，如果我们选定10%的显著水平，则$\beta$的90%置信区间将为：$0.6<\beta<1.2$
这就是说，$\beta$在范围（0.6，1.2）以内的概率为0.9。此外，它还意味着：我们可以否定$\beta$等于0的虚假设，并且具有大于90%的置信度。

在经济计量学中，我们首要的目标是，用一种使我们能够检验和评定自己模型的方式来分析数据。实际上，一个研究人员可能有几个似乎可行的模型加以选择，而且尽管这种选择，部分取决于模型的用途，但是它还取决于经验信息（即数据）。同时比较许多模型，在方法上说，是困难的。通常我们按顺序研究模型，企图证实所研究的每一个模型。这就是说，确定每一个模型时必须具有一种能够产生在经验上可以检验的假设形式。如果数据与模型不一致，则模型就要被否定，可以考虑其他方案。如果数据与模型相一致，则模型无疑被接受，一直到产生新的可以检验的假设或新的数据时为止。因此，假设检验适合于简单的模型，而且其结果不是模型被否定就是被接受。

在进行假设检验时，关于接受和否定的规则必须在研究数据以前就加以选出。有一条常用的规则是具有5%的显著水平，它确立了这样一种准则，即：在事实上，出现虚假设的机会少于5%时，否定虚假设。显著水平的选择是任意的，而实际上取决于由模型得出的结论类型。但是，重要的是要看到，经典的经济计量学的假设检验几乎只是讨论错误地否定一个真实假设的问题。由于规定的可检验假设的性质，供选择的假设经常得不到明确的定义，以致不能判断虚假设在事实上为人们错误接受的次数。为此，在应用时，我们要注意区别虚假设的否定及其接受。所以，我们常常要说明虚假设已在5%显著水平时遭到否定，同时无疑将接受供选择的假设。

例如，对消费模型，检验$\beta$等于0的虚假设。按照所估计的斜率参数，否定了虚假设而同意斜率为非零的供选择的假设。虽然，我们可以采用更具体的供选择的假设，即：斜率参数为正，来开始分析的。这样就会稍稍改变有关的统计检验，并且在虚假设被否定时增加一些补充的信息。

在经济计量应用中，标准的做法是细心研究检验统计量以及系数的标准误差。当虚假设的否定成立时，通常模型就被接受，至少在得到进一步相反的信息以前是如此。接受模型所需的显著水平在研究人员之间以及在所研究的模型之间都有很大的不同，例如，根据大量观察次数而估计的模型，可以使人否定许多说明性变量的零系数的虚假设。我们可以不采用严格的5%显著检验，却只要按排检验统计量的次序，并将变量从那个与最低检验统计量有关的模型中消去。

否定与回归系数有关的虚假设所需的统计检验通常是以t分布为基础的。t分布所以恰当是因为在统计检验时，我们需要利用误差方差的样本估计而不是它的真实值。为了使用t分布来确定估计参数的95%的置信区间，我们首先将估计的回归参数（例如$\hat{\beta}$）加以标准化，其方法是减去其真实值，并除以标准误差的估计量［见（1.27)式］：

$$t_{N-2}=\frac{\hat{\beta}-\beta}{S_{\hat{\beta}}}=\frac{\hat{\beta}-\beta}{S/\sqrt{\sum x_i^2}} \qquad (1.30)$$

需要全本请访问 www.ertongbook.com

标准化变量$t_{N-2}$将按照t分布形式，并具有$N-2$自由度。现在我们可以选一适当数值$t_C$，称为临界值，该值确保t分布的2.5%在其两个尾部以内，即

$$P(-t_c < t_{N-2} < t_c) = 0.95 \qquad (1.31)$$

式中P代表概率。分布的临界值随自由度数而变化，但是，在观察数较大时，它将等于1.96。这样，代入（1.30）式，则得出下式：

$$P\left(-t_c < \frac{\hat{\beta}-\beta}{S/\sqrt{\sum x_i^2}} < t_c\right) = 0.95 \qquad (1.32)$$

（1.30）式中的t统计量临界值将对要不要否定关于真实斜率参数$\beta$的虚假设起作用。但是，更有直接重要意义的是，可以将（1.32）式稍微加以改动使其得出：

$$P\left(\hat{\beta} - t_c\frac{S}{\sqrt{\sum x_i^2}} < \beta < \hat{\beta} + t_c\frac{S}{\sqrt{\sum x_i^2}}\right) = 0.95 \qquad (1.33)$$

由此就可得出$\beta$的95%置信区间，等于：$\beta \pm t_c\frac{S}{\sqrt{\sum x_i^2}} = \hat{\beta} \pm t_c S_{\hat{\beta}}$ （1.34）

采用类似程序，我们得出a的95%置信区间：$\hat{\alpha} \pm t_c S_{\hat{\alpha}} = \hat{\alpha} \pm t_c S\frac{\sqrt{\sum X_i^2}}{\sqrt{N\sum x_i^2}}$ （1.35）

当然，只要t分布的临界值选得正确，就有可能确定各种显著水平的置信区间。未知参数的置信区间，可使我们获得关于那些可能包含真实参数的数值范围的统计说明。所以，（1.34）式告诉我们，估计斜率参数两侧的$t_c$标准误差区间包含真实参数的概率为0.95。

在双变量回归模型中最常见的假设检验是：变量X和Y之间不存在关系。这种检验是按以下假设考虑的：虚假设：$\beta = 0$

选择假设：$\beta \neq 0$

按照虚假设，$\beta$等于0，因而合适的检验统计量变为$\hat{\beta}/S_{\hat{\beta}}$。假若估计参数与其标准误差之比大于或等于$t_c$的绝对值，那么，我们就否定虚假设。对于大样本和5%显著水平的情况下$t_c = 1.96$。我们可以经常采用这样的经验规则，即：假若检验统计量的绝对值大于2时，应否定虚假设。关于这个问题的另一种方法是根据$\hat{\beta}$的置信区间。假若0在置信区间以内，我们就不能否定虚假设，但是，假若0在该区间以外，就可以否定它。重要的是，要再次强调，如果不否定虚假设，并不一定是说它就是真实的。我们推定假设和置信区间的方式是，假若虚假设不是真实的，那末，我们否定它，其不可靠程度只占5%的机会。检验程序很少告诉我们这种情况，即：我们接受虚假设时，事实上，这个虚假设是错误的。我们在接受选择假设时，必须记住这一点。

消费支出实例

为了说明起见，我们提出一个回归式，式中，消费是因变量，而任意时间序列X则为自变量。数据按季度来自1958年第一季度至1973年第四季度。估计的回归式如下：

$$\hat{C} = 404.32 + 7.57X$$

截距项的标准误差为19.17，因而t统计量为404.32/19.17 = 21.09。由于回归式包括64次观察与两个估计参数，我们可以找到在5%显著水平时的t分布临界值为2。因此，我们可以否定截距为0的这一虚假设。但是，X系数的标准误差为36.85，因而t统计量为0.20。所以，我们不能否定5%（或10%）水平时的零斜率的虚假设。因此，结论是，消费者的开支不是任意决定的，故应否定该模型，并寻找更好的可供选择的模型。

说明性变量更合理的选择是总的个人收入DI。如消费按个人收入列成回归式，可得出以下结果：

$$\hat{C} = -12.73 + 0.79DI$$

在这种情况下，截距在5%水平时是显著的（t统计量为-3.24），但是，更重要的是，与个人收入变量系数有关的t统计量为108.98。因此，我们可以否定零斜率的虚假设，而赞成斜率为非零的选择假设。否定虚假设使我们至少暂时接受双变量回归模型。当然，如进一步研究还会使我们可能得到比方才描述的更为适合的总消费模型。

## 第六节　方差与相关分析

### 一、拟合优度

回归的残差有助于衡量估计回归线拟合样本数据点的程度。概略地说，良好的回归方程能够说明或解释Y的方差的一大部分。大的残差表明拟合不好，而小的残差则说明拟合良好。将残差（或误差项方差的估计量）作为拟合优度的量度所带来的问题是，残差值是有单位的；它取决于因变量的量度单位。我们想要得出的拟合优度量度是没有单位的，这对于某些统计检验是很重要的。从直观来看，用Y的方差除以残差变量能够提供所谋求的一种拟合优度量度，这一点似乎是很合理的。

我们首先将Y对其平均值的变差定义如下：变差（Y）$= \sum(Y_i - \bar{Y})^2$

我们的目的是，将Y的变差分成两个不同部分，第一部分由回归方程加以说明，第二部分则与模型的未说明部分（误差项）有关。首先假定线性回归模型的斜率已知为0，我们拟合的回归线只需要估计一个截距。那末，与X任一值有关的$Y_i$的最佳预测就可以由Y的样本平均值得出：$\hat{Y}_i = \hat{\alpha} + 0 \cdot X_i = \hat{\alpha} = \bar{Y}$

在这种特殊情况下，结论是，Y的变差可测定观察值$Y_i$和预测值Y之间差的平方。更一般的是，我们可以通过考虑这样的事实来改进预测，即：$Y_i$的拟合或预测值都取决于对$X_i$的观察，也就是：$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$

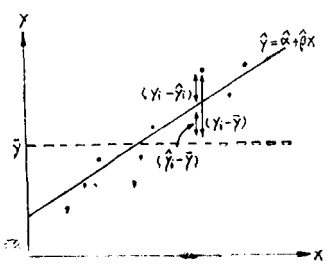


图 1.11

补充的信息可以使我们改进预测并减少Y的变差的未说明部分。为此，可考虑下列恒等式，该式适用于所有样本观察：

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (Y_i - Y) \qquad (1.36)$$

等式左边的一项代表样本点与Y的平均值之间的差，等式右边的第一项为残差，第二项为Y的预测值与Y的平均值之间的差。其图示可见图1.11。

上面的分析是针对一个样本点。为了将分析扩展到所有样本点并测定其变差，我们使（1.36）式两侧平方，然后将全部观察i=1，2，…，N相加。

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 + 2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \qquad (1.37)$$

该表示式的最后一项可证明等于0。（因为$\sum\hat{\varepsilon}_i = 0$和$\sum\hat{\varepsilon}_i X_i = 0$）。因此，

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \qquad (1.38)$$

Y的总变差　　Y的残余变差　　Y的说明变差

（或总平方和）　　（或误差平方和）　　（或回归平方和）

$$TSS = ESS + RSS$$

为了将Y的变差分类进行标准化，将方程（1.38）的两侧除以变差的总平方和，得出：

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

我们将回归方程的R平方（$R^2$）定义如下：

$$R^2 = 1 - \frac{ESS}{TSS} \quad \frac{RSS}{TSS} \qquad (1.39)$$

$R^2$是Y的说明变差与总变差之比。由于误差平方和的数值范围在0与总平方和之间，所以很容易发现$R^2$的数值范围在0与1之间。如果线性回归模型无助于说明Y的变差，这时$R^2$就出现0。例如，当全部样本点在水平线上Y = Y，或样本点在圆周上（图1.12b)时，就会发生这种情况。只有在全部样本点都在估计回归线上（如图1.12a 所示的完全拟合情况）时，$R^2$才会出现1。

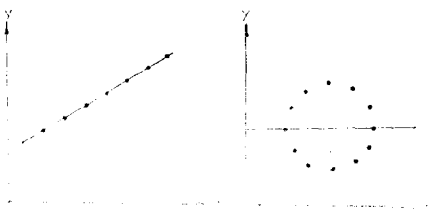

图 1.12a



图 1.12b

参数相联系，这将是很有用的。为此，我们仍将测定的变量写成它们平均值的偏差形式，即$y_i = Y_i - \bar{Y}$，$x_i = X_i - \bar{X}$。$y_i$的预测值为：

$$\hat{y}_i = \hat{\beta} x_i$$

每个因变量观察值可以表示为：$y_i = \hat{y}_i + \hat{\varepsilon}$

式中$\hat{\varepsilon}_i$为回归残差。从而$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{\varepsilon}_i^2 = \beta^2 \sum x_i^2 + \sum \varepsilon_i^2$

因为，$(\sum \hat{y}_i \hat{\varepsilon}_i = \hat{\beta} \sum x_i \hat{\varepsilon}_i = 0)$

由此可得：$R^2 = \frac{RSS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \hat{\beta}^2 \frac{\sum x_i^2}{\sum y_i^2}$

或者

$$R^2 = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum y_i^2} \qquad (1.40)$$

(1.40）式系计算$R^2$的一个简单公式。

请注意，在我们的讨论中，$R^2$只是一个说明性统计量。概略地说，我们将$R^2$的较高值同回归线的良好拟合相联系，将$R^2$的较低值同较差的拟合相联系。但是，必须看到，$R^2$的较低值的发生可能是由于一个或几个相互联系的原因。在有些情况下，X变量并不是一个良好的说明性变量。尽管有理由相信，X确实对预测Y有帮助，但是，甚至变量X在方程中已经出现以后，还可能有很大的Y的未说明变差。换句话说，X的说明能力同残差和总变差相比可能是很小的。例如，在时间序列研究中，我们常常获得高的$R^2$值，这只是因为随时间增长的变量，可能会很好地说明任何其它随时间增长的变量。另一方面，在抽样研究中，尽管模型很理想，但仍会出现低$R^2$值，这是因为在数据中固有的个别观察具有较大的变差所造成的。这说明，$R^2$单独应用不一定是模型满意程度的合适量度。一种较好的全面量度，可能是能说明模型对新数据的预测能力的一种统计量。

有时根据方差分析将Y的变差的分类加以概括是很有用的。在这种情况下，Y的全部未说明和已说明的变差都转换成方差，其方法是，除以适当的自由度数。因此，Y的方差是总变差除以N－1（因为计算样本平均值损失一个自由度）；已说明方差等于已说明变差（因为回归式除去用干估计Y平均值的自由度以外，只有一个自由度）；残余方差为残余变差除

以N－2（因为计算样本的平均值和斜率参数损失了二个自由度）。需要说明一下，计算一个统计量有关的自由度数目，等于变量的观察次数（即数据点数），减去由于统计计算程序加到数据上的约束或限制条件数目（即参数个数）。

二、相关

$R^2$通常对于分析一个具有因变量Y与自变量X之间因果关系的回归模型是很有价值的。因此，$R^2$被认为不只是若干变量中间相关的一个量度。回归与相关的区别在于，相关是没有内含的因果关系假设，而回归则具有内含的这种假设，我们已经知道，对回归模型的因变量和自变量的选择是一个关键。因变量是需要说明和预测的变量，而自变量则是推动力量或原因变量。最小二乘法回归技术只有在研究数据以前，能够确定出模型的因果关系时才适用。最小二乘法回归技术的目的是证实早先导出的模型。如果模型规定为$Y=\alpha+\beta X$，我们就可以将回归斜率参数的显著t统计量作为能证实模型的根据。另一方面，不显著的统计量将不能证实模型。

作为没有因果关系的相关的一个例子，可以考虑一系列随时间的观察，如从十九世纪对非洲医学的研究中得出的观察数据。我们能在一个地区的医生数目与同一地区的流行疾病之间找出较高的相关。但是，如果推论认为有医生实际上是疾病流行的原因，那就错了。

因此，高相关并不是为了推论因果关系。我们必须事先规定基于以前的信息，某一地区的医生数目随疾病流行而变化，并且从统计上检验在正确使用回归技术时这种关系是否成立。

请注意，回归技术只能假定而不能证明因果关系。相关技术通常为研究人员用来提出假设或证实原先的怀疑。只要研究人员小心谨慎不直接从数据分析中推论出因果关系，这些程序才能被接受。在经济、企业以及其他领域，都有两个变量高度相关，且它们二者又由第三个基本变量确定的许多例子。如果情况确实如此的话，那末，第三个基本变量在回归模型中应作为自变量而出现。

当作出不正确的因果关系式时，回归斜率参数将会出现什么情况。让我们回到有关回归模型的讨论，并且比较与以下回归模型有关的斜率参数：

i, $Y=a+bX+e$; ii, $X=A+BY+e'$

B和b的最小二乘法估计量为：$\hat{b}=\dfrac{\sum x_i y_i}{\sum x_i^2}$　　$\hat{B}=\dfrac{\sum x_i y_i}{\sum y_i^2}$

只要$\hat{b}=1/\hat{B}$或者$R^2=1$，这两个回归斜率才会对X与Y之间的关系产生相同的结论。因此，回归模型规格的选择将影响我们的参数估计量和预测。

三、回归方程的检验

将Y的变差分成两部分的方法可以对Y与X之间存在线性关系进行统计检验。请考虑以下比值：$F_{1,N-2}=\dfrac{\text{已说明方差}}{\text{未说明方差}}=\dfrac{RSS/1}{ESS/(N-2)}=\dfrac{\hat{\beta}^2\sum x_1^2}{S^2}$

在其他条件都相同的情况下，可以予计X与Y之间的强（线性）统计关系将导致已说明方差与未说明方差的比值大。这种检验可以直接应用，因为，$F_{1,N-2}$遵循一已知分布（F分布），并具有1和N－2自由度。F的下标分别表示分子和分母的自由度数目。只有在回归式中的已说明方差为0时，F统计量的数值才是0。人们总是把F的低数值同X与Y之间的弱线性关系相联系。把高数值同强线性关系相联系。幸而F统计量的准确数字分布是已知的。例如，我们要否定在5%显著水平时Y与X之间无关系的虚假设，其方法是，查找具有1和N－2自由度的F分布

（5％显著性）的相应临界值。如果由回归式算出的$F_{1,N-2}$的数值大于临界数值，那末，我们就可以在5％显著水平时否定虚假设，如果$F_{1,N-2}$数值低于临界值，我们就不能否定虚假设。

上面叙述的F检验同那个与$\beta=0$的虚假设有联系的t检验两者具有密切关系。事实上，我们可以一般地说明如下，即：在相同的显著水平时，$F_{1,N-2}=t^2_{N-2}$。我们所以要明确地介绍F检验是因为它对于包括多元回归方程显著性检验在内的联合假设检验是很有用的。为此，我们将在后面使用F检验。

汽车零售实例

美国对汽车零售额（因变量）与经济中的工资、薪金总额（自变量）之间的关系进行了研究。可以预计工资和薪金的水平高一些，可以使汽车销售额增加。以下系根据月度时间序列数据的零售额对工资和薪金回归的概括。所估计的方程式为

$$S=\alpha+\beta W+\varepsilon$$

式中，S＝1963年1月至1970年4月的汽车月度零售额，以百万美元为单位

W＝1963年1月至1970年4月的月度工资，以十亿美元为单位

拟合回归线如下。我们将t统计量写入估计系数的下面括号内。我们还在因变量符号上面加了一顶“帽子”，说明该方程式用于计算因变量的估计数值。

$$\hat{S}=1,767.61+7.48W \qquad R^2=0.80 \qquad F_{1,86}=344.0$$
$$(7.4) \quad (18.5)$$

正的常数截距项的含意是，根据假设，如果在某月份不考虑工资，个别人仍会购置汽车。工资变量系数可解释为：工资和薪金每增加十亿美元将使汽车销售额增加748万美元。该模型可用来预测以未来工资为条件的未来的汽车销售水平。请注意，斜率系数一般作为测定与自变量较小变化相关的因变量变化（事实上，在线性模型中，$\hat{\beta}=\partial s/\partial W$适用于W的所有变化）。估计系数并不是没有单位的。其数值直接与因变量S（百万美元单位）与自变量W（十亿美元单位）的单位有关。在本例中，我们决定在括号内写出t统计量而不写出估计标准误差。我们利用t统计量就可以否定以下虚假设，即可以否定在1％以及5％显著水平时，截距和斜率都为0的虚假设。0.80的$R^2$表示，回归式能说明因变量中变差的80％。$F_{1,86}$的数值为344使我们否定以下虚假设，即否定在汽车销售额同工资和薪金之间（在1％水平时）没有关系的虚假设。

如果人们有理由坚信汽车销售额与工资薪金的关系线穿过原点，而且不管0截距的虚假设已经被否定，那末，运算回归式时自然就会没有截距常数项。同一样本的结果如下式：

$$S=11.71W \qquad R^2=0.54$$
$$(10.2)$$

尽管t检验使我们能够否定虚假设，但显然将常数项删减后已经降低了方程式的说明能力，因此，我们需要估计有截距的回归模型。只有在具有充分理由使方程式穿过原点时，才可以使截距等于0。所以在没有常数项时使用$R^2$，必须谨慎地加以说明；在省略常数项时，必须修改检验的推导。不然，方程的说明能力要减弱。

实际上，对与没有常数项方程有关的$R^2$必须非常谨慎地加以说明，当常数项被省略时，检验的推导必须修改。$R^2$在0和1范围之内，这点是特别重要的，在这种情况下，比较拟合值很清楚地表明，方程的说明能力是减弱了。

## 第七节　附录一：基础知识介绍

### 一、随机变量

随机变量是根据偶然机会选择数值的一种变量。更准确地说，随机变量的性质是，它包含了不同的数值，每个数值出现的概率小于 1。我们还可以通过研究产生这些数值的机理来描述随机变量。这种机理叫做“概率分布”，它能说明所有可能的结果以及它们各自发生的概率。我们可以将随机变量定义为一个函数，这个函数规定实验的每一个结果为一实数。例

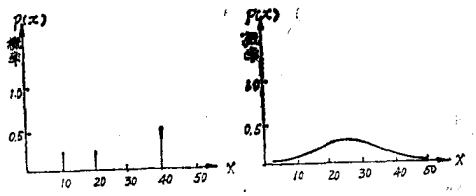


图 1.13

1，反面定为数值 0 （如果用一个完好的钱币，正面概率为1/2）我们可以将上抛钱币正反面的数值作为随机变量。在这种具体情况下，产生随机变量的机理是二项式概率分布。

区别离散和连续的随机变量是有用的。连续随机变量可以是某一实数区间的任一数值，而离散随机变量则可能只具有实数的特定值。图1.13说明了离散（左图）和连续（右图）的概率函数P(x)。

### 二、期望值

随机变量的平均值或期望值是与变量有关的可能结果的加权平均数，在这种场合，概率只作为适当的加权。在离散条件下，如果与随机变量X有关的N个可能的结果为$X_1$，$X_2$，…，$X_N$，那末，X的平均值将等于：

$$P_1X_1 + P_2X_2 + \cdots + P_NX_N = \sum_{i=1}^{N} P_iX_i = E(X) = \mu x \qquad (1.41)$$

式中，$\sum P_i = 1$，$P_i = X_i$ 发生的概率（即 $X = X_i$的概率），E （ ） =“期望算子”

我们将常常利用这个期望算子，E(X)有如下性质：

（1） $E(aX + b) = aE(X) + b$　　（2） $E(aX)^2 = a^2E(X^2)$

其中a和b是常数，在连续随机变量情况下，可写成下式： $\mu_X = E(X) = \int_{-\infty}^{+\infty} XP(X)dX$

式中，X为一连续随机变量，其概率分布由P(X 表示。假若概率分布是对称的， 那末，平均值就位于中间。因此，平均值通常作为随机变量X的平均数或中心倾向的量度。

### 三、方差

方差是X变量围绕平均值的分散或离中趋势的量度，并定义如下：

$$\sigma_x^2 = \sum_{i=1}^{N} P_i[X_i - E(X)]^2 \qquad \text{离散情况} \qquad (1.42)$$

$$\sigma_x^2 = \int_{-\infty}^{+\infty} P(X)[X - E(X)]^2 dX \qquad \text{连续情况} \qquad (1.43)$$

方差的正平方根称为“标准偏差”。根据定义，方差和标准偏差都不是负的。

还有两个定义是很有用的。如果我们对随机变量X和Y之间的关系感兴趣，我们通过研

需要全本请访问 www.ertongbook.com