

6.13 抽选 9 家自行车厂的定价和质量评定的名次如下：

工厂	1	2	3	4	5	6	7	8	9
名次	6	9	2	8	5	1	7	4	3
价格	480	395	575	550	510	545	400	465	420

用 $\alpha = 0.05$ 检验名次和价格之间的等级相关。

6.14 10 种不同品牌的矿泉水由消费者品尝排出名次及销售价格的高低名次如下：

价格名次	4	1	9	2	6	3	7	5	10	8
品尝名次	5	7	6	1	2	9	3	4	10	8

计算等级相关系数 r_{sp} , 用 $\alpha = 0.05$ 检验 $\rho = 0$ 。

第七章 相关与回归分析

相关与回归分析是处理变量与变量之间关系的一种统计方法。近年来,这种方法已被广泛应用于生物学、医学、心理学、教育学、社会学、经济学等诸多领域,并取得了一定成效。相关与回归分析从所处理的变量多少来看,如果研究的是两个变量之间的关系,称为简单相关与简单回归分析;如果研究的是两个以上变量之间的关系,称为多元相关与多元回归分析。从变量之间的关系形态上看,有线性相关与线性回归分析及非线性相关与非线性回归分析。从统计思想和方法上看,简单线性相关与简单线性回归是最基本的方法。因此本章重点讨论简单线性相关和简单线性回归的基本原理与方法。

第一节 简单线性相关

一、相关关系及其表现形态

(一)什么是相关关系

任何事物的变化都与之周围的其他事物相互联系和相互影响,用于描述事物数量特征的变量之间自然也存在一定的关系。统计分析的目的在于如何根据统计数据确定变量之间的关系形态及其关联的程度,并探索出其内在的数量规律性。人们在实践中发现,变量之间的关系形态可分为两种类型,即函数关系和统计关系。

函数关系是我们比较熟悉的。设有两个变量 x 和 y ,变量 y 随变量 x 一起变化,并完全依赖于 x ,当变量 x 取某个数值时, y 依确定的关系取相应的值,则称 y 是 x 的函数,记为 $y = f(x)$,其

中 x 称为自变量, y 称为因变量。例如, 某种商品的销售额与销售量之间的关系。设 y 为销售额, x 为销售量, p 为单价, 则 x 与 y 之间的关系可表示为 $y = px$ 。这就是说, 在销售价格不变的情况下, 对于该商品的某一销售量, 总有一个销售额与之对应, 即销售额完全由销售量所确定, 二者之间为线性函数关系。

函数关系是一一对应的确定关系。但在实际问题中, 变量之间的关系往往不那么简单。例如, 我们考察居民储蓄与居民家庭收入这两个变量, 它们之间就不存在完全确定的关系。也就是说, 收入水平相同的家庭, 它们的储蓄往往不同, 反之, 储蓄额相同的家庭, 它们的收入额也可能不同。可见家庭储蓄并不能完全由家庭收入所确定, 因为家庭收入尽管同家庭储蓄有密切的关系, 但它并不是影响储蓄的唯一因素, 还有银行利率、消费水平等其他因素的影响作用。正是由于影响一个变量的因素非常之多, 才造成了变量之间关系的不确定性。我们把变量之间存在的不确定的数量关系称为统计关系或相关关系。

相关关系的特点是, 一个变量的取值不能由另一个变量唯一确定。当变量 x 取某个值时, 变量 y 的取值可能有几个。对这种关系是确定的变量显然不能用函数关系进行描述, 但也不是无任何规律可寻。通过对大量数据的观察与研究, 我们就会发现许多变量之间确实存在着一定的客观规律性。例如, 平均来说, 收入水平高的家庭, 其家庭储蓄一般也较多。相关与回归分析正是描述与探索不确定变量之间关系及其规律的统计方法。

(二) 相关关系的表现形态

相关关系的表现形态大体上可分为线性相关、非线性相关、完全相关和不完全相关等几种。就两个变量而言, 如果变量之间的关系近似地表现为一条直线, 则称为线性相关; 如果变量之间的关系近似地表现为一条曲线, 则称为非线性相关或曲线相关; 如图 7.1(d); 如果一个变量的取值完全依赖与另一个变量, 各观察点落在一条线上, 称为完全相关, 如图 7.1(c) 和 (e), 这实际上就是函数关系; 如果两个变量的观察点很分散, 无任何规律, 则表示变量

之间没有相关关系,如图 7.1(f)。

在线性相关中,若两个变量的变动方向相同,一个变量的数值增加或减少,另一个变量也随之增加或减少,则称为正相关,如图 7.1(a);若两个变量的变动方向相反,一个变量数值的增大或减少,另一个变量随之减少或增大,则称为负相关,如图 7.1(b)。简单相关分析就是对两个变量之间线性关系的描述与度量。

二、相关关系的描述与测度

(一) 散点图

对于两个变量 x 和 y ,通过观察或试验我们可以得到若干组数据,记为 (x_i, y_i) ($i = 1, 2, \dots, n$)。相关分析所要解决的问题是,根据这些数据确定变量之间是否存在相关关系? 如果存在的话,如何描述出它们之间的关系并对其关系强度进行测度?

散点图是描述变量之间相关关系的一种直观方法。我们用横坐标代表自变量 x ,纵坐标代表因变量 y ,每组数据 (x_i, y_i) 在坐标系中用一个点表示, n 组数据在坐标系中形成的点称为散点,这样的图形称为散点图。散点图描述了两个变量之间的大致关系,从中可以直观地看出变量之间的关系形态及关系强度。图 7.1 就是不同形态的散点图。

【例 7.1】 家庭储蓄与家庭收入之间有一定关系。现从某城市家庭中随机抽取 12 个家庭,所得月收入与月储蓄的样本数据如表 7.1。试绘制散点图。

表 7.1 12 个家庭的月收入与储蓄额数据

家庭 编号	月收 入 (百元) x	月储 蓄 (百元) y	家庭 编号	月收 入 (百元) x	月储 蓄 (百元) y
1	9	3	7	22	8
2	13	5	8	20	7
3	15	4	9	23	10

续表

家庭 编号	月收入 (百元) x	月储蓄 (百元) y	家庭 编号	月收入 (百元) x	月储蓄 (百元) y
4	17	6	10	28	11
5	18	7	11	30	10
6	26	9	12	33	12

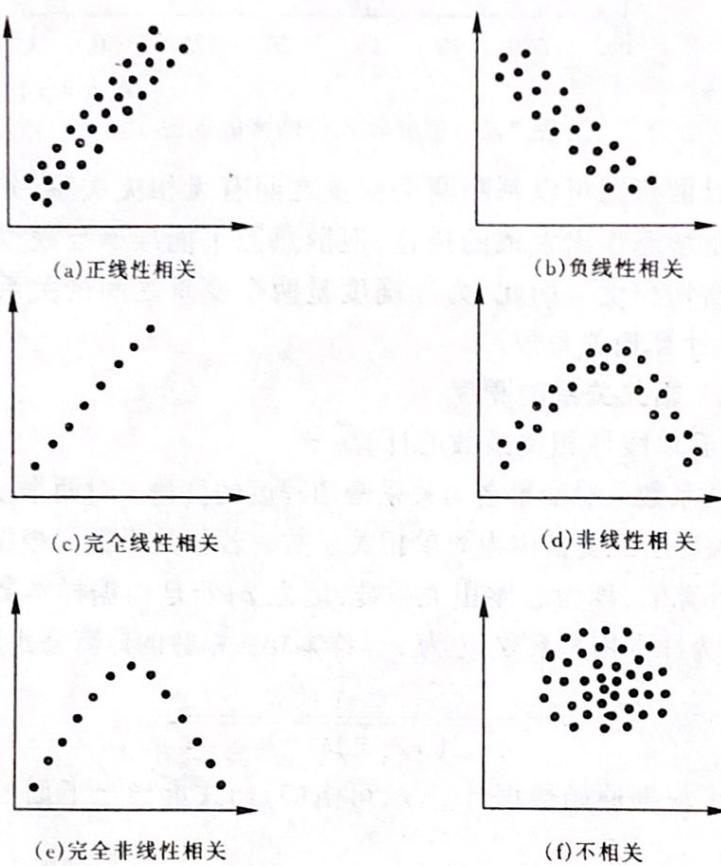


图 7.1 相关关系的表现形态(散点图)

根据表 7.1 中的数据绘制成散点图如图 7.2。从散点图可以看出,家庭储蓄随着收入的增加而增加,而且它们之间大致成一种线性相关关系。

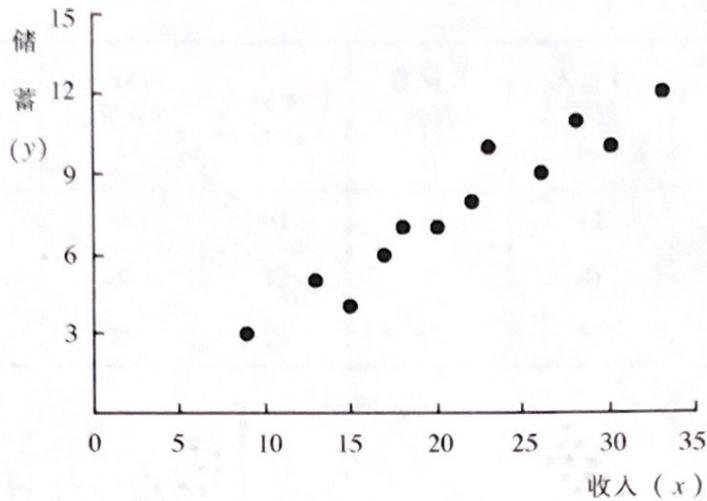


图 7.2 家庭收入与储蓄散点图

通过散点图可以判断两个变量之间有无相关关系，并对变量间的关系形态作出大致的描述，但散点图不能准确反映变量之间的关系密切程度。因此，为准确度量两个变量之间的关系密切程度，需要计算相关系数。

(二) 相关关系的测度

1. 简单线性相关系数的计算

相关系数是对变量之间关系密切程度的度量。对两个变量之间线性相关程度的度量称为简单相关系数。若相关系数是根据总体全部数据计算的，称为总体相关系数，记为 ρ ；若是根据样本数据计算的，则称为样本相关系数，记为 r 。样本相关系数的计算公式为：

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \cdot \sqrt{\sum (y - \bar{y})^2}} \quad (7.1)$$

为了根据原始数据计算 r ，可由(7.1)式推导出下面的简化计算公式：

$$r = \frac{n \sum xy - \sum x \cdot \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}} \quad (7.2)$$

【例 7.2】 根据表 7.1 中的数据，计算家庭月收入与家庭储蓄之间的相关系数。

解：计算过程见表 7.2。

表 7.2 家庭收入与储蓄相关系数计算表

家庭 编号	月收 入 (百元) <i>x</i>	月储 蓄 (百元) <i>y</i>	<i>x</i> ²	<i>y</i> ²	<i>xy</i>
1	9	3	81	9	27
2	13	5	169	25	65
3	15	4	225	16	60
4	17	6	289	36	102
5	18	7	324	49	126
6	26	9	676	81	234
7	22	8	484	64	176
8	20	7	400	49	140
9	23	10	529	100	230
10	28	11	784	121	308
11	30	10	900	100	300
12	33	12	1 089	144	396
合计	254	92	5 950	794	2 164

根据(7.2)式得：

$$r = \frac{12 \times 2 164 - 254 \times 92}{\sqrt{12 \times 5 950 - (254)^2} \times \sqrt{12 \times 794 - (92)^2}} = 0.9607$$

即家庭收入与储蓄之间的相关系数为 0.9607, 说明二者之间存在高度正线性相关关系。

2. 相关系数的意义

可以证明, 相关系数的取值范围在 +1 和 -1 之间, 即 $-1 \leq r \leq 1$ 。若 $0 < r \leq 1$, 表明 x 与 y 之间存在正相关关系, 如图 7.1(a); 若 $-1 \leq r < 0$, 表明 x 与 y 之间存在负相关关系, 如图 7.1(b)。若 $r = +1$, 表明 x 与 y 之间为完全正相关关系, 如图 7.1(c); 若 $r = -1$, 表明 x 与 y 之间为完全负相关关系。可见当 $|r| = 1$ 时, y 的取值完全依赖于 x , 二者之间即为函数关系; 当 $r = 0$ 时, 说明 y 的取值与 x 无关, 即二者之间不存在线性相关关系, 如图 7.1(f)。但需要注意的是, $r = 0$ 只表示两个变量之间不存在线性相关关系, 并不说明变量之间没有任何关系, 比如它们之间可能存在非线性相关关系。变量之间的非线性相关程度较大时, 就可能导致 r

$= 0$ 。因此,当 $r = 0$ 或很小时,不能轻易得出两个变量之间不存在相关关系的结论,而应结合散点图作出合理的解释。

根据实际数据计算出的 r ,其取值一般在 $-1 < r < 1$ 之间,在说明两个变量之间的线性关系的密切程度时,根据经验可将相关程度分为以下几种情况:当 $|r| \geq 0.8$ 时,视为高度相关; $0.5 \leq |r| < 0.8$ 时,视为中度相关; $0.3 \leq |r| < 0.5$ 时,视为低度相关; $|r| < 0.3$ 时,说明两个变量之间的相关程度极弱,可视为不相关。但这种说明必须建立在相关系数通过显著性检验的基础上。

第二节 一元线性回归

相关分析的目的在于测度变量之间的关系密切程度,它所使用的测度工具就是相关系数。而回归分析则侧重于考察变量之间的数量伴随关系,并通过一定的数学表达式将这种关系描述出来,进而确定一个或几个变量的变化对另一个特定变量的影响程度。具体来说,回归分析主要解决以下几个方面的问题:从一组样本数据出发,确定出变量之间的数学关系式;对这些关系式的可信程度进行各种统计检验,并从影响某一特定变量的诸多变量中找出哪些变量的影响是显著的,哪些是不显著的;利用所求的关系式,根据一个或几个变量的取值来估计或预测另一个特定变量的取值,并给出这种估计或预测的精确程度。

一、回归模型与回归方程

(一) 回归模型

在回归分析中,我们把被预测的变量称为因变量,用 y 表示;把用来预测因变量的一个或多个变量称为自变量,用 x 表示。例如,在分析家庭收入对储蓄的影响时,我们要预测一定的家庭收入下的储蓄额是多少。因此储蓄额应该作为因变量,而用来预测储蓄额的收入额应作为自变量。

当只涉及一个自变量时称为一元回归,若因变量 y 与自变量

x 之间为线性关系时称为一元线性回归。对于具有线性关系的两个变量,可以用一个线性方程来表示它们之间的关系。描述因变量 y 如何依赖于自变量 x 和误差项的方程称为回归模型,对于只涉及一个自变量的简单线性回归模型可表示为:

$$y = \beta_0 + \beta_1 x + \epsilon \quad (7.3)$$

在简单线性回归模型中, y 是 x 的线性函数 ($\beta_0 + \beta_1 x$ 部分) 加上误差项 ϵ 。 $\beta_0 + \beta_1 x$ 反映了由于 x 的变化而引起的 y 的线性变化; ϵ 是被称为误差项的随机变量,它反映了除 x 和 y 之间的线性关系之外的随机因素对 y 的影响,是不能由 x 和 y 之间的线性关系所解释的变异性;式中的 β_0 和 β_1 称为模型的参数。

(7.3)式被称为理论回归模型,对这一模型我们通常有三个假定:

(1) 误差项 ϵ 是一个期望值为 0 的随机变量,即 $E(\epsilon) = 0$ 。这意味着在(7.3)式中,由于 β_0 和 β_1 都是常数,所以有 $E(\beta_0) = \beta_0$, $E(\beta_1) = \beta_1$ 。因此对于一个给定的 x 值, y 的期望值为 $E(y) = \beta_0 + \beta_1 x$ 。

(2) 对于所有的 x 值, ϵ 的方差 σ^2 都相同。

(3) 误差项 ϵ 是一个服从正态分布的随机变量,且相互独立。即 $\epsilon \sim N(0, \sigma^2)$ 。独立性意味着对于一个特定的 x 值,它所对应的 ϵ 与其他 x 值所对应的 ϵ 不相关;因此,对于一个特定的 x 值,它所对应的 y 值与其他 x 所对应的 y 值也不相关。

(二) 回归方程

根据回归模型中的假定, ϵ 的平均值或期望值等于 0,因此 y 的平均值或期望值 $E(y) = \beta_0 + \beta_1 x$,也就是说, y 的平均值是 x 的线性函数。描述 y 的平均值或期望值如何依赖于 x 的方程称为回归方程。简单线性回归方程的形式如下:

$$E(y) = \beta_0 + \beta_1 x \quad (7.4)$$

简单线性回归方程的图示是一条直线,因此也称为直线回归方程。其中 β_0 是回归直线在 y 轴上的截距,是当 $x = 0$ 时 y 的期望值; β_1 是直线的斜率,它表示当 x 每变动一个单位时, y 的平均

变动值。

(三) 估计的回归方程

如果回归方程中的参数 β_0 和 β_1 已知, 对于一个给定的 x 的值, 利用(7.4)式就能计算出 y 的平均值。但总体回归参数 β_0 和 β_1 是未知的, 我们必需利用样本数据去估计它们。用样本统计量 b_0 和 b_1 代替回归方程中的未知参数 β_0 和 β_1 , 这时我们就得到了估计的回归方程, 简单线性回归中的估计的回归方程如下:

$$\hat{y}_i = b_0 + b_1 x_i$$

其中: b_0 是估计的回归直线在 y 轴上的截距, b_1 是直线的斜率, 它表示对于一个给定的 x 的值, \hat{y} 是 y 的估计值。 b_1 也表示 x 每变动一个单位时, y 的平均变动值的估计值。

二、最小二乘估计

对于第 i 个 x 值, 估计的回归方程可表示为:

$$\hat{y}_i = b_0 + b_1 x_i$$

对于 x 和 y 的 n 对观察值, 用于描述其关系的直线有很多条, 究竟用哪条直线来代表两个变量之间的关系, 需要有一个明确的原则。我们自然会想到距离各观察点最近的一条直线, 用它来代表 x 与 y 之间的关系与实际数据的误差比其他任何直线都小。根据这一思想确定直线中未知常数 b_0 和 b_1 的方法称为最小二乘法。换句话说, 最小二乘法是使因变量的观察值 y_i 与估计值 \hat{y}_i 之间的离差和达到最小来求得 b_0 和 b_1 的方法, 即

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2 = \text{最小}$$

令 $Q = \sum (y_i - \hat{y}_i)^2$, 在给定了样本数据后, Q 是 b_0 和 b_1 的函数, 且最小值总是存在。根据微积分的极值定理, 对 Q 求相应于 b_0 和 b_1 的偏导数, 并令其等于 0, 即可求出 b_0 和 b_1 , 即

$$\frac{\partial Q}{\partial b_0} = -2 \sum (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum x_i (y_i - b_0 - b_1 x_i) = 0$$

经化简得到求解 b_0 和 b_1 的标准方程组:

$$\begin{aligned} b_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \\ b_0 &= \frac{\sum y_i}{n} - b_1 \left(\frac{\sum x_i}{n} \right) = \bar{y} - b_1 \bar{x} \end{aligned} \quad (7.5)$$

【例 7.3】 根据表 7.1 中的数据, 拟合居民家庭储蓄与家庭月收入的回归直线。解: 根据表 7.2 中的计算结果, 由(7.5)式得:

$$b_1 = \frac{12 \times 2164 - 254 \times 92}{12 \times 5950 - (254)^2} = 0.3777$$

$$b_0 = \frac{92}{12} - 0.3777 \times \frac{254}{12} = -0.3280$$

家庭储蓄与家庭月收入的直线回归方程为:

$$\hat{y} = -0.3280 + 0.3777x$$

回归系数 $b_1 = 0.3777$ 表示, 家庭月收入每增加 100 元, 储蓄额平均增加 37.77 元。将 x 的各个取值代入上述回归方程, 可以得出家庭储蓄额的各估计值。由图 7.3 可以看出散点图与回归直线的关系。

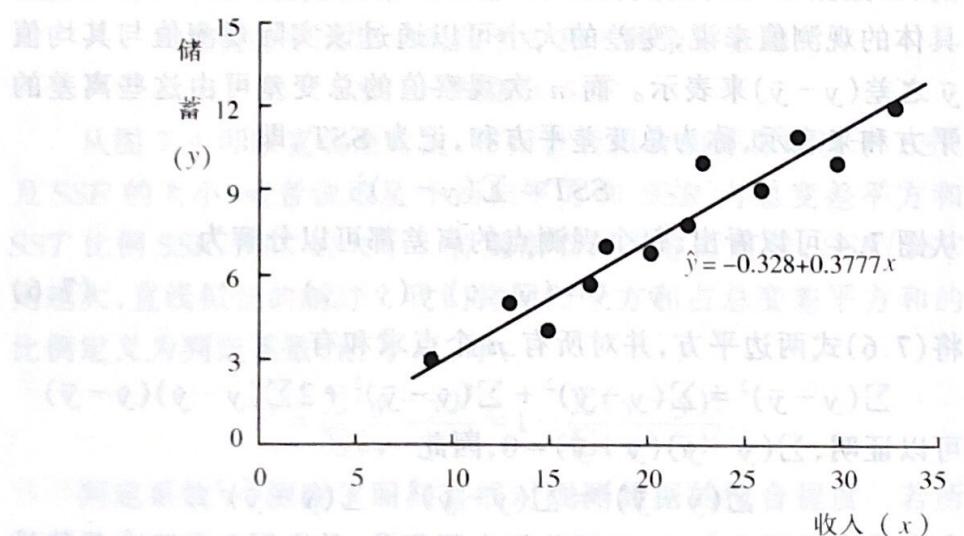


图 7.3 家庭储蓄对家庭月收入的回归直线

三、回归直线的拟合程度

回归直线 $\hat{y}_i = b_0 + b_1 x_i$ 在一定程度上描述了变量 x 与 y 之间的内在规律, 根据这一方程, 我们可由自变量 x 的取值来估计因变量 y 的取值。但估计的精度如何将取决于回归直线对观测数据的拟合程度。可以想象, 如果各观测数据的散点都落在这一直线上, 那么这条直线就是对数据的完全拟合, 直线充分代表了各个点, 此时用 x 来估计 y 是没有误差的。各观察点越是紧密围绕直线, 说明直线对观测数据的拟合程度越好, 反之则越差。我们把回归直线与各观察点的接近程度称为回归直线对数据的拟合程度。为说明直线的拟合程度, 我们需要研究因变量 y 取值的变化规律。

(一) 判定系数

判定系数是说明回归直线拟合程度的一个度量值。为说明它的含义, 我们需要对因变量 y 取值的变差进行研究。

因变量 y 的取值是不同的, y 取值的这种波动称为变差。变差的产生来自于两个方面: 一是由于自变量 x 的取值不同造成的; 二是除 x 以外的其他因素(包括测量误差等)的影响。对一个具体的观测值来说, 变差的大小可以通过该实际观测值与其均值 \bar{y} 之差 $(y - \bar{y})$ 来表示。而 n 次观察值的总变差可由这些离差的平方和来表示, 称为总变差平方和, 记为 SST , 即:

$$SST = \sum (y - \bar{y})^2$$

从图 7.4 可以看出, 每个观测点的离差都可以分解为

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y}) \quad (7.6)$$

将(7.6)式两边平方, 并对所有 n 个点求和有

$$\sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2 + 2 \sum (y - \hat{y})(\hat{y} - \bar{y})$$

可以证明, $\sum (y - \hat{y})(\hat{y} - \bar{y}) = 0$, 因此

$$\sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$$

即总的变差平方和 SST 可分解为两部分: 其中 $\sum (\hat{y} - \bar{y})^2$ 是估计值 \hat{y} 与均值 \bar{y} 的离差平方和, 根据回归方程, 估计值 $\hat{y} = b_0 +$

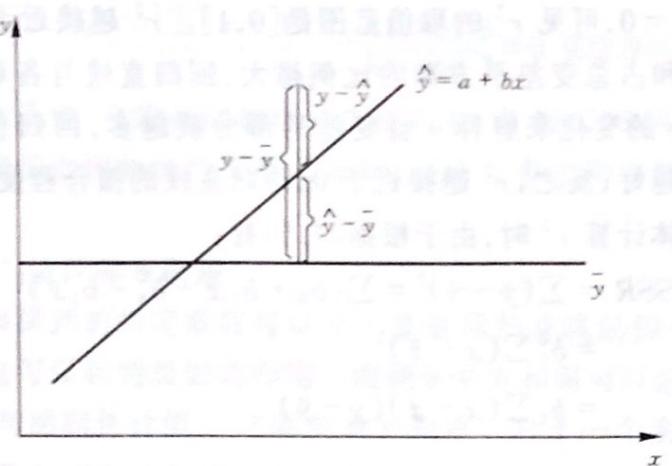


图 7.4 变差分解图

$b_1 x$, 因此可以把 $(\hat{y} - \bar{y})$ 看作是由于自变量 x 的变化引起的 y 的变化, 而其平方和 $\sum(\hat{y} - \bar{y})^2$ 则反映了 y 的总变差中由于 x 与 y 之间的线性关系引起的 y 的变化部分, 它可以由回归直线来解释, 因而称为可解释的变差或回归平方和, 记为 SSR 。另一部分 $\sum(y - \hat{y})^2$ 是各实际观测点与估计值的残差 $(y - \hat{y})$ 平方和, 它是除了 x 对 y 的线性影响之外的其他因素对 y 变差的作用, 是不能由回归直线来解释的, 因而称为不可解释的变差或剩余平方和, 记为 SSE 。三个平方和的关系为:

$$\text{总变差平方和} = \text{回归平方和} + \text{剩余平方和}$$

$$SST = SSR + SSE$$

从图 7.4 可以直观地看出, 回归直线拟合的好坏取决于 SSR 及 SSE 的大小, 或者说取决于回归平方和 SSR 占总变差平方和 SST 比例 SSR/SST 的大小。各观察点越是靠近直线, SSR/SST 则越大, 直线拟合的越好。我们将回归平方和占总变差平方和的比例定义为判定系数, 记为 r^2 , 即:

$$r^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

判定系数 r^2 测度了回归直线对观测数据的拟合程度。若所有观测值都落在直线上, 剩余平方和 $SSE = 0$, $r^2 = 1$, 拟合是完全的; 如果 x 的变化与 y 无关, x 完全无助于解释 y 的变差, 此时 \hat{y}

$= \bar{y}$, 则 $r^2 = 0$, 可见 r^2 的取值范围是 $[0, 1]$ 。 r^2 越接近于 1, 表明回归平方和占总变差平方和的比例越大, 回归直线与各观测点越接近, 用 x 的变化来解释 y 值变差的部分就越多, 回归直线的拟合程度就越好; 反之, r^2 越接近于 0, 回归直线的拟合程度就越差。

在具体计算 r^2 时, 由于根据(7.5)有:

$$\begin{aligned} SSR &= \sum (\hat{y} - \bar{y})^2 = \sum (b_0 + b_1 x - b_0 - b_1 \bar{x})^2 \\ &= b_1^2 \sum (x - \bar{x})^2 \\ &= b_1 \sum (x - \bar{x})(y - \bar{y}) \end{aligned}$$

所以

$$\begin{aligned} r^2 &= \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{b_1^2 \sum (x - \bar{x})^2}{\sum (y - \bar{y})^2} \\ &= \frac{b_1 \sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} \\ &= \left[\frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \cdot \sqrt{\sum (y - \bar{y})^2}} \right]^2 \end{aligned} \quad (7.7)$$

括号内的部分正是简单相关系数 r 。可见在一元线性回归中, 相关系数 r 实际上是判定系数的平方根。这一结论不仅可以使我们能由相关系数直接计算判定系数 r^2 , 也可以帮助我们进一步理解相关系数的意义。相关系数 r 与回归系数 b_1 的正负号是相同的, 实际上, 相关系数 r 也从另一个角度说明了回归直线的拟合程度。 $|r|$ 越接近 1, 表明回归直线对观测数据的拟合程度就越高。但用 r 说明回归直线的拟合程度需要慎重, 因为 r 的值总是大于 r^2 的值(除非 $r = 0$ 或 $|r| = 1$)。比如, 当 $r = 0.5$ 时, 表面上看似乎有一半的相关了, 但 $r^2 = 0.25$, 实际上我们只能解释因变量总变差的 25%。 $r = 0.7$ 才能解释近一半的变差, $|r| \leq 0.3$ 意味只有很少一部分变差可由回归直线来解释。

【例 7.4】 根据例 7.3 家庭储蓄对家庭月收入的回归计算判定系数。

解: 根据(7.7)式得:

$$r^2 = \frac{b_1^2 \sum (x - \bar{x})^2}{\sum (y - \bar{y})^2} = \frac{81.832^2}{88.666^2} = 0.9229$$

这就是说,在家庭储蓄额的总变差中,有 92.29% 可由家庭月收入与储蓄之间的线性关系来解释,说明二者之间有较强的线性关系。

(二) 估计标准误差

上面讲到的判定系数可以用于度量回归直线的拟合程度,相关系数也可以起到类似的作用。而剩余平方和则可以说明实际观测值 y_i 与回归估计值 \hat{y}_i 之间的差异程度。对于一个变量的诸多观测值,我们可以用标准差来测度各观测值在其平均数周围的分散程度。与之类似的一个量可以用来测度各实际观测点在直线周围的散布状况,这个量就是估计标准误差,其定义公式如下:

$$s_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} \quad (7.8)$$

估计标准误差 s_y 可以看作是在排除了 x 对 y 的线性影响后, y 随机波动大小的一个估计量。各观测点越靠近直线, s_y 越小, 回归直线对各观测点的代表性就越好;若各观测点全部落在直线上, 则 $s_y = 0$ 。可见 s_y 也从另一个角度说明了回归直线的拟合程度或两个变量之间的关系密切程度。

从(7.8)式容易看出,回归直线是对 n 个观测点拟合的所有直线中,估计标准误差最小的一条直线,因为回归直线是使 $\sum (y - \hat{y})^2$ 最小来确定的。为计算方便,根据回归方程可以得出估计标准误差的简化计算公式:

$$s_y = \sqrt{\frac{\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i}{n - 2}} \quad (7.9)$$

【例 7.5】 根据例 7.3 的有关结果,计算家庭储蓄对家庭月收入回归的估计标准误差。

解:根据(7.9)式得

$$s_y = \sqrt{\frac{794 - (-0.328) \times 92 - 0.3777 \times 2164}{12 - 2}} = 0.8266 (\text{百元})$$

四、回归分析中的显著性检验

上面介绍了如何根据样本数据拟合回归方程 $\hat{y}_i = b_0 + b_1 x_i$ ，并讨论了回归方程对各观察点的拟合程度。下面我们讨论回归分析中的假设检验问题。

(一) 为什么要进行显著性检验

对回归方程进行显著性检验基于以下两点理由：第一，在根据样本数据拟合回归方程时，我们首先是假设变量 x 与 y 之间存在着线性关系，因此可表示为 $y = \beta_0 + \beta_1 x + \epsilon$ ，根据样本数据得到的估计方程为 $\hat{y}_i = b_0 + b_1 x_i$ ，但这种假设是否成立，需要通过检验才能证实。第二，估计方程 $\hat{y}_i = b_0 + b_1 x_i$ 中的回归系数 b_1 反映的是变量 x 对 y 的影响程度，为分析 x 对 y 的影响是否显著，就需要对回归系数的显著性进行检验。

(二) 假设检验的内容

回归分析中的假设检验一般包括两个方面的内容：一是线性关系的检验；二是回归系数的检验。

1. 线性关系的检验。线性关系的检验是检验自变量和因变量之间的线性关系是否显著，或者说，它们之间能否用一个线性模型 $y = \beta_0 + \beta_1 x + \epsilon$ 来表示。具体方法是将回归平方和 (SSR) 同剩余平方和 (SSE) 加以比较，应用 F 检验来分析二者之间的差别是否显著。如果是显著的，说明两个变量之间存在着线性关系；如果不显著，说明两个变量之间不存在线性关系。检验的具体步骤如下：

第一步：提出假设

$$H_0: \text{线性关系不显著}$$

第二步：计算检验统计量 F

$$F = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\sum(\hat{y} - \bar{y})^2/1}{\sum(y - \hat{y})^2/(n-2)} \quad (7.10)$$

可以证明，在原假设成立的情况下， F 统计量服从 F 分布，第一自由度为 1，第二自由度为 $n-2$ ，即 $F \sim F(1, n-2)$ 。

第三步：确定显著性水平 α （通常取 $\alpha = 0.05$ ），并根据两个自

由自由度 $df_1 = 1$ 和 $df_2 = n - 2$ 查 F 分布表, 找到相应的临界值 F_a 。

第四步: 作出决策。若 $F \geq F_a$, 拒绝 H_0 , 说明两个变量之间的线性关系是显著的; 若 $F < F_a$, 不能拒绝 H_0 , 说明两个变量之间的线性关系不显著。

【例 7.6】 根据例 7.3 的回归方程, 对家庭储蓄与家庭月收入之间线性关系的显著性进行检验。

解: 假设二者之间的线性关系不显著 (H_0)。

根据(7.10)式得

$$F = \frac{\sum (\hat{y} - \bar{y})^2 / 1}{\sum (y - \hat{y})^2 / n - 2} = \frac{81.8323 / 1}{6.8374 / (12 - 2)} = 119.6834$$

取显著性水平 $\alpha = 0.05$, 根据自由度 $df_1 = 1$ 和 $df_2 = 10$, 查 F 分布表找到相应的临界值 $F_a = 4.96$ 。由于 $F = 119.6834 > F_{0.05} = 4.96$, 拒绝 H_0 , 说明家庭储蓄与家庭月收入之间的线性关系是显著的。

2. 回归系数的检验。在线性关系通过检验之后, 回归系数的显著性检验就是要检验自变量对因变量的影响是否显著的问题。在简单线性回归模型 $y = \beta_0 + \beta_1 x + \epsilon$ 中, 如果回归系数 $\beta_1 = 0$, 回归线是一条水平线, 表明自变量的变化对因变量没有影响, 即两个变量之间没有线性关系, 否则说明二者之间存在线性关系。因此, 回归系数的显著性检验就是检验回归系数 β_1 是否等于 0。检验的具体步骤如下:

第一步: 提出假设。假设样本是从一个没有线性关系的总体中选出的, 即

$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$

第二步: 计算检验的统计量 t 值:

$$t = \frac{b_1}{s_{b_1}}$$

在原假设成立的情况下, 统计量 t 服从自由度为 $n - 2$ 的 t 分布, 即 $t \sim t(n - 2)$ 。其中: s_{b_1} 是回归系数 b_1 的标准差, 可由下式求得:

$$s_{b_1} = \sqrt{\frac{s_y^2}{\sum(x_i - \bar{x})^2}}$$

第三步：确定显著性水平 α （通常取 $\alpha = 0.05$ ），并根据自由度 $df = n - 2$ 查 t 分布表，找到相应的临界值 $t_{\alpha/2}$ 。

第四步：作出决策。若 $|t| \geq t_{\alpha/2}$ ，拒绝 H_0 ，表明自变量 x 对因变量 y 的影响是显著的，换言之，两个变量之间确实存在着显著的线性相关关系；若 $|t| < t_{\alpha/2}$ ，则接受 H_0 ，表明 x 对 y 的影响是不显著的，二者之间不存在显著的线性相关关系。

【例 7.7】 根据例 7.3 的回归方程，对回归系数的显著性进行检验。

解：提出假设。假设家庭月收入对家庭储蓄的影响不显著，二者之间无线性关系，即：

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

计算统计量的 t 值：

$$t = \frac{b_1}{s_{b_1}} = \frac{0.3777}{\sqrt{0.8266^2 / 573.6667}} = 10.94$$

取显著性水平 $\alpha = 0.05$ ，根据自由度 $df = n - 2 = 10$ ，找到相应的临界值 $t_{\alpha/2} = t_{0.025} = 2.2281$ 。由于 $|t| = 10.94 > t_{\alpha/2} = 2.2281$ ，拒绝 H_0 ，表明样本回归系数是显著的，家庭储蓄与家庭月收入之间确实存在着显著的线性相关关系，家庭月收入是影响家庭储蓄的显著因素。

在进行显著性检验时，有以下两点需要注意：

第一，对回归系数进行的显著性检验也可能犯错误，如果我们拒绝了 $H_0: \beta_1 = 0$ ，仅表明在 x 的样本观察值范围内， x 和 y 是相关的，也就是说，由于样本的观察值不支持原假设 H_0 ，因此我们拒绝了 H_0 ，但并不等于 H_0 就一定不成立。

第二，在一元线性回归中，自变量只有一个，上面介绍的 F 检验和 t 检验是等价的，也就是说，如果 $H_0: \beta_1 = 0$ 被 t 检验拒绝（或接受），它也将被 F 检验所拒绝（或接受）。但在多元回归分析中，

这两种检验的意义是不同的, F 检验只是用来检验总体回归关系的显著性, 而 t 检验则是检验回归中各个系数的显著性。

五、利用回归方程进行估计和预测

对于一组样本数据, 我们利用最小二乘法得到了估计的回归方程, 如果回归方程通过了统计上的显著性检验, 我们就可以利用估计的回归方程, 根据样本观察值范围之内的 x 值对 y 进行估计和预测, 但一般情况下, 在自变量 x 的取值范围之外进行预测应十分谨慎。

(一) 点估计

利用估计的回归方程, 对于 x 的一个特定值 x_0 , 求出 y_0 的平均值 $E(y_0)$ 的一个估计值, 这就是平均值的点估计; 如果对于 x 的一个特定值 x_0 , 求出 y_0 的一个个别值的估计值 \hat{y}_0 , 则属于个别值的点估计。比如, 在例 7.3 中, 我们得到的估计的回归方程为 $\hat{y} = -0.328 + 0.3777x$, 假如我们要估计月收入为 2 000 元时所有家庭的平均月储蓄额, 就是平均值的点估计, 根据估计的回归方程得: $\hat{y}_0 = -0.328 + 0.3777 \times 20 = 722.6$ (元)。如果我们只是想知道月收入为 2 000 元的那个家庭的月储蓄额是多少, 则属于个别值的点估计。根据估计的回归方程得: $\hat{y}_0 = -0.328 + 0.3777 \times 20 = 722.6$ (元)。这就是说, 月收入为 2 000 元的那个家庭(这里是编号为 8 的那个家庭), 月储蓄额为 722.6 元。在点估计条件下, 平均值的点估计和个别值的点估计是一样的。但在区间估计中则是不同的。

(二) 区间估计

点估计不能给出估计的精度。实际上, 点估计值 \hat{y}_0 与 y_0 的实际值之间是有误差的, 因此需要进行区间估计。区间估计有两种类型: 一是置信区间估计, 它是对 x 的一个给定值, 求出 y 的平均值的估计区间, 这一估计区间称为置信区间; 二是预测区间估计, 它是对 x 的一个给定值, 求出 y 的一个个别值的估计区间, 这

一区间称为预测区间。

1. y 的平均值的置信区间估计。设 x_0 为自变量 x 的一个特定值或给定值; $E(y_0)$ 为对于给定的 x_0 , 因变量 y 的平均值或期望值。当 $x = x_0$ 时, $\hat{y}_0 = b_0 + b_1 x_0$ 为 $E(y_0)$ 估计值。

一般来说, 我们不能期望估计值 \hat{y}_0 精确地等于 $E(y_0)$ 。因此要想用 \hat{y}_0 推断 $E(y_0)$, 必需考虑根据估计的回归方程得到的 \hat{y}_0 的方差, 对于给定的 x_0 , 统计学家给出了估计 \hat{y}_0 方差的公式, 用 $s_{\hat{y}_0}^2$ 表示 \hat{y}_0 方差的估计量, 其计算公式为:

$$s_{\hat{y}_0}^2 = s_y^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad (7.11)$$

\hat{y}_0 标准差的估计量计算公式为:

$$s_{\hat{y}_0} = s_y \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

式中: s_y 为估计标准误差。

有了 \hat{y}_0 的标准差之后, 对于给定的 x_0 , $E(y_0)$ 在置信水平 $1 - \alpha$ 下的置信区间可表示为:

$$\hat{y}_0 \pm t_{\alpha/2} s_y \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (7.12)$$

【例 7.8】 根据例 7.3 的回归方程, 取 $x_0 = 20$, 建立家庭储蓄额 95% 的置信区间。

解: 根据前例计算结果, 已知 $n = 12$, $\hat{y} = -0.328 + 0.3777x$, $S_y = 0.8266$, $t_{\alpha/2} = t_{0.025} = 2.2281$ 。

当 $x_0 = 20$ 时, 根据回归方程得:

$$\hat{y}_0 = -0.328 + 0.3777 \times 20 = 7.226$$

根据(7.12)式得 $E(y_0)$ 的置信区间为:

$$7.226 \pm 2.2281 \times 0.8266 \sqrt{\frac{1}{12} + \frac{1.3611}{537.6667}}$$

即: $6.6863 \leq E(y_0) \leq 7.7657$, 也就是说, 我们可以 95% 的可靠性保证, 当家庭月收入为 2000 元时, 家庭平均月储蓄额在 668.63 元到 776.57 元之间。

当 $x_0 = \bar{x}$ 时, \hat{y}_0 的标准差的估计量最小, 此时有: $s_{\hat{y}_0} = s_y \sqrt{1/n}$ 。这就是说, 当 $x_0 = \bar{x}$ 时, 估计是最准确的。 x_0 偏离 \bar{x} 越远, y 的平均值的置信区间就变得越宽, 估计的效果也就越不好。

2. y 的个别值的预测区间估计。假如我们不是估计所有月收入为 2000 元的家庭的平均月储蓄额, 而只希望估计月收入为 2000 元的那个家庭的月储蓄额的区间是多少, 这个区间称为预测区间。

为求出预测区间, 我们首先必需确定当 $x = x_0$ 时与利用 \hat{y}_0 作为 y 的一个个别估计值的点估计相联系的方差, 这个方差由以下两部分组成:

(1) y 的个别值关于平均值的方差, 它的估计量由 s_y^2 (估计标准误差) 给出;

(2) 与利用 \hat{y}_0 估计 $E(y_0)$ 相联系的方差, 它的估计量由 $s_{\hat{y}_0}^2$ 给出。(见(7.11)式)。

统计学家已给出了 y 的一个个别估计 y_0 的方差的估计量, 我们用 s_{ind}^2 表示, 其计算公式为:

$$\begin{aligned} s_{ind}^2 &= s_y^2 + s_{\hat{y}_0}^2 \\ &= s_y^2 + s_y^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \\ &= s_y^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \end{aligned}$$

y 的一个个别估计值 y_0 的标准差的估计量为:

$$s_{ind} = s_y \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

因此,对于给定的 x_0 , y 的一个个别值 y_0 在置信水平 $1 - \alpha$ 下的预测区间可表示为:

$$\hat{y}_0 \pm t_{\alpha/2} s_y \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (7.13)$$

【例 7.9】根据例 7.3 的回归方程,建立家庭收入额为 2 000 元的那个家庭储蓄额 95% 的预测区间。

解:根据前例计算结果,已知 $n = 12$, $x_0 = 20$, $\hat{y} = -0.328 + 0.3777x$, $S_y = 0.8266$, $t_{\alpha/2} = t_{0.025} = 2.2281$ 。

当 $x_0 = 20$ 时,根据回归方程得:

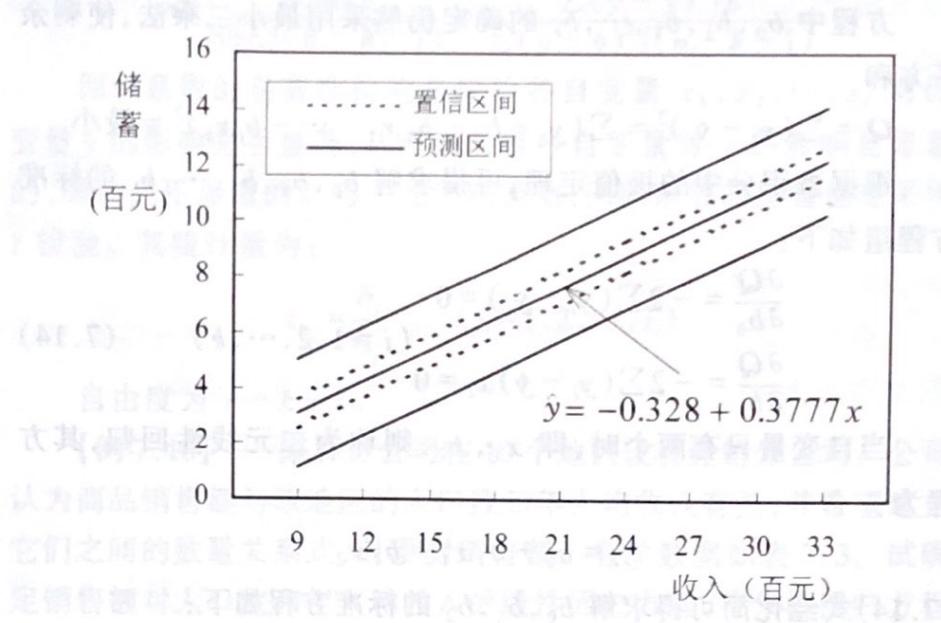
$$\hat{y}_0 = -0.328 + 0.3777 \times 20 = 7.226$$

根据(7.13)式得家庭收入额为 2 000 元的那个家庭储蓄额 95% 的预测区间为:

$$7.226 \pm 2.2281 \times 0.8266 \sqrt{1 + \frac{1}{12} + \frac{1.3611}{537.6667}}$$

即: $5.3068 \leq \hat{y}_0 \leq 9.1452$,也就是说,我们可以 95% 的可靠性保证,家庭月收入为 2 000 元的那个家庭的月储蓄额在 530.68 元到 914.52 元之间。

与月收入为 2 000 元时所有家庭的平均月储蓄额的置信区间 (6.6863, 7.7657) 相比, y 的个别值的预测区间是比较宽的。二者的差别表明,估计 y 的平均值比预测 y 的一个特定值或个别值更精确。只有当 $x_0 = \bar{x}$ 时,置信区间和预测区间都是最精确的。如图 7.5 所示。



(图 7.5) 图 7.5 家庭月收入与家庭储蓄的置信区间和预测区间

第三节 多元线性回归和非线性回归

一、多元线性回归

上面所讨论的回归问题只涉及一个自变量,但在实际问题中,影响因变量的因素往往有多个,这种一个因变量同多个自变量的回归问题称为多元回归,当因变量同各自变量之间为线性关系时,称为多元线性回归。多元线性回归分析的基本原理同一元线性关系相同,但计算上要复杂得多,一般需借助于计算机来完成。

多元线性回归的估计方程可以写为:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

其中: b_1, b_2, \dots, b_k 称为偏回归系数, b_1 表示在其他变量不变的条件下,自变量 x_1 变动一个单位引起的因变量 y 的平均变动数额; b_2, \dots, b_k 的意义类似。

方程中 $b_0, b_1, b_2, \dots, b_k$ 的确定仍然采用最小二乘法,使剩余平方和

$$Q = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_1 - \dots - b_k x_k)^2 = \text{最小}$$

根据微积分中的极值定理,可得求解 $b_0, b_1, b_2, \dots, b_k$ 的标准方程组如下:

$$\begin{aligned} \frac{\partial Q}{\partial b_0} &= -2 \sum (y_i - \hat{y}_i) = 0 \\ \frac{\partial Q}{\partial b_i} &= -2 \sum (y_i - \hat{y}) x_i = 0 \end{aligned} \quad (7.14)$$

当自变量只有两个时,即 x_1, x_2 ,则称为二元线性回归,其方程为:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2$$

(7.14)式经化简可得求解 b_0, b_1, b_2 的标准方程如下:

$$\begin{aligned} \sum y &= nb_0 + b_1 \sum x_1 + b_2 \sum x_2 \\ \sum x_1 y &= b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \\ \sum x_2 y &= b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 \end{aligned} \quad (7.15)$$

为计算简便,可令

$$x'_1 = x_1 - \bar{x}_1, x'_2 = x_2 - \bar{x}_2, y' = y - \bar{y}$$

(7.15)式可化简为:

$$\begin{aligned} b_1 \sum x'^1_1^2 + b_2 \sum x'_1 x'_2 &= \sum x'_1 y' \\ b_1 \sum x'_1 x'_2 + b_2 \sum x'^2_2 &= \sum x'_2 y' \\ b_0 &= \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \end{aligned} \quad (7.16)$$

对于多元线性回归,也需要测定回归方程的拟合程度、检验回归方程和回归系数的显著性。

测定多元线性回归方程的拟合程度,与一元回归中的判定系数类似,使用多重判定系数,也称为复判定系数,其定义为:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

R^2 的平方根称为多重相关系数,也称为复相关系数。 R^2 测度了样本回归方程的拟合程度。

在多元线性回归中,对回归方程线性关系的显著性检验也是采用 F 检验, F 统计量为: