

相关与回归分析

相关分析与回归分析

1. 相关分析

是指研究一个变量与另一个变量或另一组变量之间相关方向和相关密切程度的统计分析方法。

2. 回归分析

是指根据相关关系的具体形态，选择一个合适的数学模型来近似地表达变量间平均变化关系的统计分析方法。

3. 相关分析与回归分析的联系

- (1) 相关分析与回归分析是研究现象之间相关关系的两种基本方法，两者有着密切的联系，它们不仅具有共同的研究对象，而且在具体应用时，常常必须互相补充。
- (2) 相关分析需要依靠回归分析来表明现象数量相关的具体形式，而回归分析则需要依靠相关分析来表明现象数量变化的相关程度。

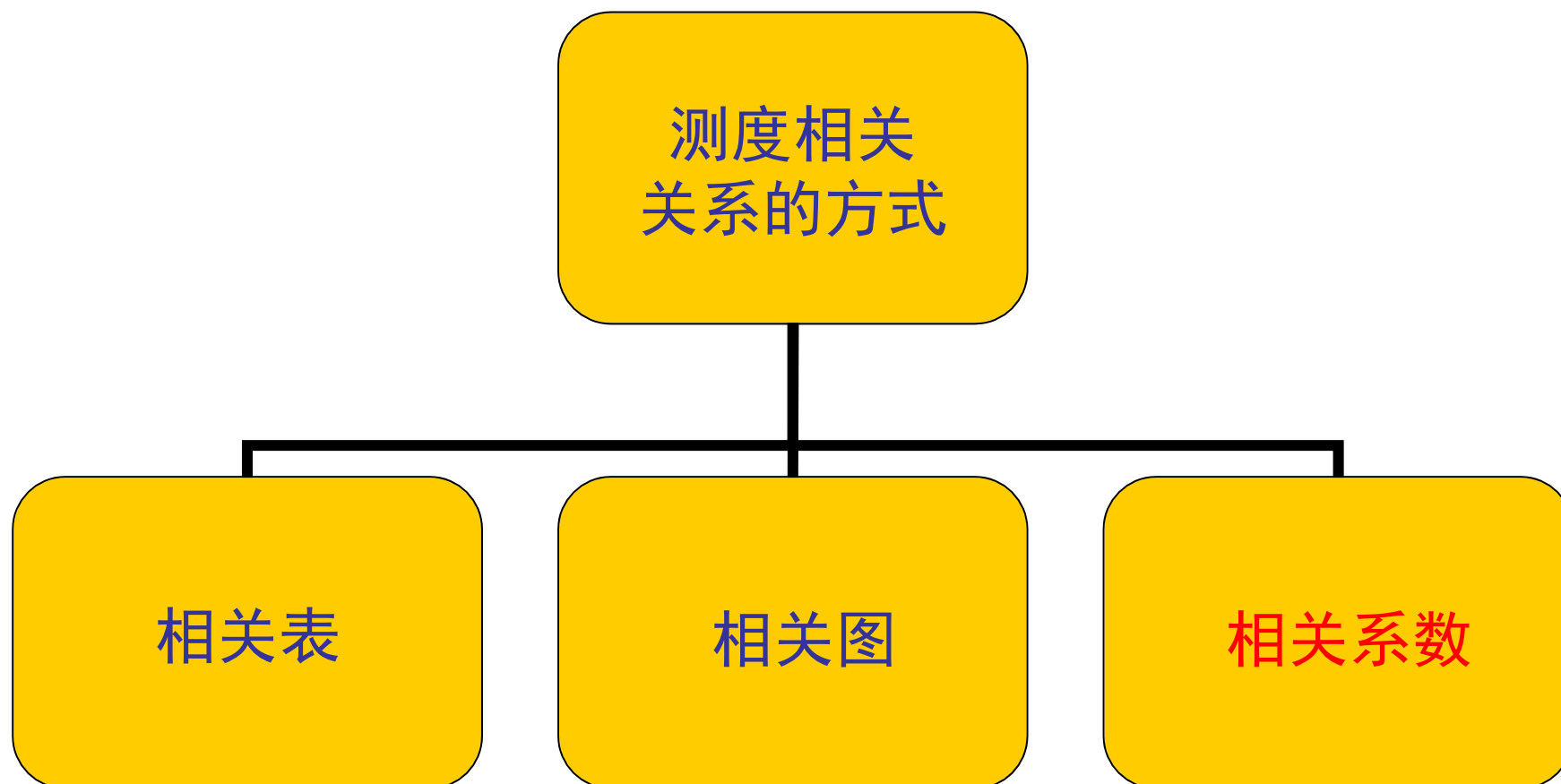
- (3) 只有当变量之间存在着高度相关时，进行回归分析寻求其相关的具体形式才有意义。
- (4) 由于上述原因，回归分析和相关分析在一些统计学的书籍中被合称为相关关系分析或广义的相关分析。

4. 相关分析与回归分析的区别

(1) 相关分析中，变量 x 与变量 y 处于平等地位，不需要区分自变量和因变量；回归分析中，变量 y 称为因变量，处在被解释的特殊地位。变量 x 称为自变量，可以通过 x 的变化来解释 y 的变化，故亦称为解释变量。

- (2)相关分析中所涉及的变量 y 与 x 全是随机变量。
而回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量。
- (3)相关分析的研究主要是刻画两类变量间线性相关的密切程度。而回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制。

相关关系的测度



相关系数

1. 相关系数概念

是反映变量之间线性相关密切程度的统计分析指标。
相关系数可依总体数据或样本数据计算，分别定义为总体相关系数 ρ 和样本相关系数 r 。

3. 简单相关系数的计算

设 $(x_i, y_i)(i = 1, 2, \dots, n)$ 是 (x, y) 的 n 组样本观察值，两个变量之间的简单线性相关系数计算公式如下：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$$

实际计算时也可以使用下列简捷公式：

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

r 的取值范围是 $[-1, 1]$ ， $|r|=1$ ，为完全相关， $r=1$ ，为完全正相关， $r=-1$ ，为完全负正相关， $r=0$ ，不存在线性相关关系， $-1 \leq r < 0$ ，为负相关， $0 < r \leq 1$ ，为正相关， $|r|$ 越趋于1表示关系越密切； $|r|$ 越趋于0表示关系越不密切

根据相关系数的取值判断相关程度的标准：

相关系数取值	相关程度
$ r \geq 0.8$	高度相关
$0.8 > r \geq 0.5$	中度相关
$0.5 > r \geq 0.3$	低度相关
$ r < 0.3$	不相关

必须注意，上述判断还只是针对样本而言的，样本范围内所存在的现象之间的相关程度是否在总体范围内也存在呢？就需要对总体相关程度进行假设检验。

4. 相关关系的显著性检验

r 是依据样本数据计算的，根据一个样本的相关系数能否说明总体的相关性呢？这需对样本相关系数的显著性进行检验。

样本相关系数的理论分布函数是很复杂的。 r 的抽样分布随总体相关系数和样本容量的大小而变化。

在进行这项检验时，通常假设 x 与 y 是正态变量，如果总体相关系数 $\rho=0$ ，则样本相关系数 r 服从 t 分布。

(二) 相关系数

4. 相关关系的显著性检验 检验的步骤为

(1) 提出假设: $H_0: \rho = 0$; $H_1: \rho \neq 0$

(2) 计算检验的统计量:

$$t = |r| \sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

(3) 确定显著性水平 α

(4) 作出决策

若 $|t| > t_{\alpha/2}$, 拒绝 H_0

若 $|t| < t_{\alpha/2}$, 不能拒绝 H_0

一元线性回归分析

一、一元线性回归模型

(一) 回归模型的基本形式

1. 总体回归模型

$$y = \beta_0 + \beta_1 x + \varepsilon$$

式中： y 为因变量（被解释变量）， x 为自变量(解释变量)， β_0 和 β_1 是未知参数，称为回归参数，称 β_1 为回归系数， ε 表示其他随机因素的影响，并假定 ε 是不可观测的随机误差，它是一个随机变量一般称之为变量 y 对 x 的一元线性理论回归模型，或称为总体回归模型。

(一) 回归模型的基本形式

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\beta_0 + \beta_1 x$$

$$\varepsilon$$

线性组合部分：确定部分

随机干扰部分：不确定部分

对于总体中的个体而言，有：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

(一) 回归模型的基本形式

2. 总体回归函数（方程）

对于总体回归模型中的 ε_i ，通常假设：

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

对总体回归模型两边取期望，得：

$$E(y_i) = \beta_0 + \beta_1 x_i$$

(一) 回归模型的基本形式

3. 样本回归模型

一般情况下，在研究某个实际问题时，对于获得的 n 组样本观测值来说，如果它们符合总体回归模型，则

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, 2, \dots, n$$

上式为样本回归模型，并假定 n 组数据是独立观测的，故 y_1, y_2, \dots, y_n 都是独立的随机变量， e_i 为残差，是对 ε_i 的估计， $\hat{\beta}_0, \hat{\beta}_1$ 是对 β_0, β_1 的估计。

(一) 回归模型的基本形式

4. 样本回归函数（方程）

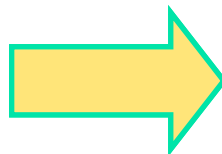
对于样本回归模型中的 e_i ，通常假设：

$$E(e_i) = 0$$

$$Var(e_i) = \hat{\sigma}^2$$

对总体回归模型两边取期望，得：

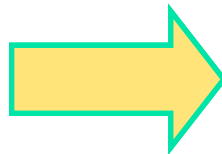
$$E(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$



样本回归函数（方程）



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$



估计的回归方程

(二) 回归模型的基本假设

假设1: 误差项的期望值为0, 即对所有的 i 有 $E(\varepsilon_i)=0$

假设2: 误差项的方差为常数, 即对所有的 i 有

$$\text{var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$$

假设3: 误差项之间不存在自相关关系, 其协方差为0, 即当

$$i \neq j \text{ 时, 有 } \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad ;$$

假设4: 自变量是给定的变量, 与随机误差项线性无关;

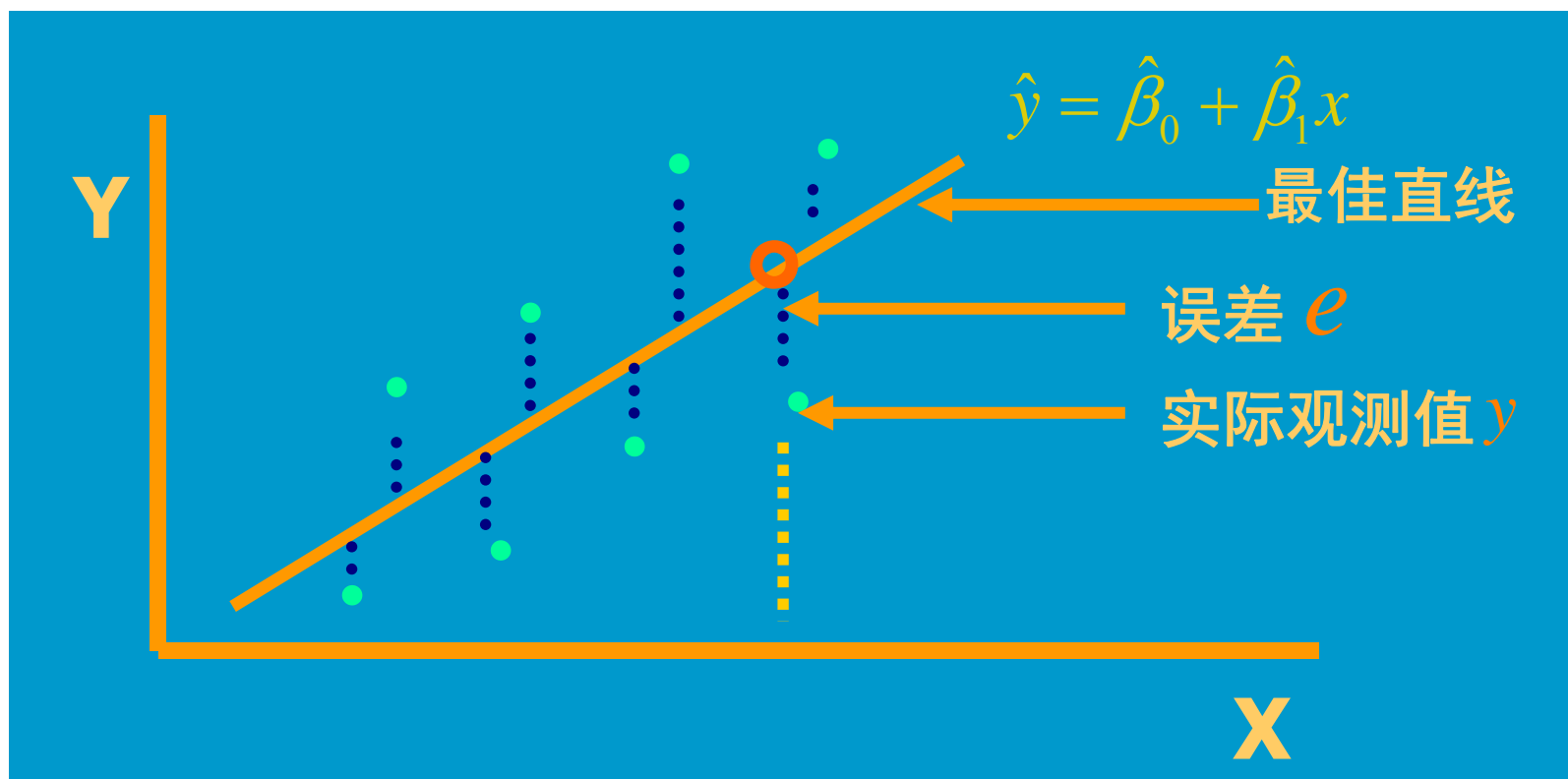
假设5: 随机误差项服从正态分布。

以上这些基本假设是德国数学家高斯最早提出的, 故也称为高斯假定或标准假定。

二、一元线性回归模型的估计

(一) 参数的最小二乘估计

参数的估计，就是寻求最佳直线来拟合变量间数量变化关系的过程。




(一) 参数的最小二乘估计

基本思想：使误差平方和最小



数学表达：

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1) &= \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \min \end{aligned}$$


解决办法：通过对Q求偏导数，确定使其最小的 $\hat{\beta}_0, \hat{\beta}_1$

(一) 参数的最小二乘估计

对Q求关于 $\hat{\beta}_0, \hat{\beta}_1$ 偏导数:

$$\begin{cases} \frac{\partial Q}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial Q}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$



$$\begin{cases} n\hat{\beta}_0 + (\sum x_i)\hat{\beta}_1 = \sum y_i \\ (\sum x_i)\hat{\beta}_0 + (\sum x_i^2)\hat{\beta}_1 = \sum x_i y_i \end{cases}$$

(一) 参数的最小二乘估计

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \end{cases}$$



$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{cases}$$

(三) 最小二乘估计量的性质---期望

最小二乘法是多种估计方法中的一种。按最小二乘法求得的总体回归系数的估计值被称为最小二乘估计量。最小二乘估计量的形式是不变的，但根据所选取的样本不同， $\hat{\beta}$ 的具体数值会随之变化，因此它是一种随机变量。可以证明，在基本假设能够得到满足的条件下，回归系数的最小二乘估计量的期望值等于真值，即有

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

(三) 最小二乘估计量的性质----方差

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差为：

$$\text{var}(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \sigma^2 = \left[\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}} \right] \sigma^2$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{L_{xx}}$$

$$L_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

(三) 最小二乘估计量的性质----方差

不难证明:

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}}\right)\sigma^2\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$$

还可以证明 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是 β_0 和 β_1 的最佳线性无偏估计, 也称为最小方差线性无偏估计, 也就是说, 在 β_0 和 β_1 的一切线性无偏估计中, 它们的方差最小.

(四) 回归系数的区间估计

回归分析中，有时需要知道回归系数的取值区间，此时就需要对回归系数进行区间估计。

对回归系数进行区间估计，就是在回归系数分布的基础上，以回归系数的估计值为中心，构造一个置信区间，使该区间以较大的概率包含总体回归系数的真值。

(四) 回归系数的区间估计

根据 $\beta_1 \sim N(\hat{\beta}_1, \frac{\sigma^2}{L_{xx}})$ 可得：

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} \\ &= \frac{\hat{\beta}_1 - \beta_1 \sqrt{L_{xx}}}{\hat{\sigma}} \end{aligned}$$

服从自由度为 $n - 2$ 的 t 分布，因而有

$$P \left[\left| \frac{\hat{\beta}_1 - \beta_1 \sqrt{L_{xx}}}{\hat{\sigma}} \right| < t_{\alpha/2}(n - 2) \right] = 1 - \alpha$$

$$P \left[\hat{\beta}_1 - t_{\alpha/2}(n - 2) \frac{\hat{\sigma}}{\sqrt{L_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}(n - 2) \frac{\hat{\sigma}}{\sqrt{L_{xx}}} \right] = 1 - \alpha$$

(四) 回归系数的区间估计

即得 β_1 的置信度 $1-\alpha$ 的置信区间为:

$$(\hat{\beta}_1 - t_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{L_{xx}}}, \hat{\beta}_1 + t_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{L_{xx}}})$$

同理,根据 $\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}})\sigma^2)$,可以推导出 β_0

的置信区间为

$$(\hat{\beta}_0 - t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}}}, \hat{\beta}_0 + t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}}})$$

(五) 总体方差的估计

除了参数 β_0, β_1 外，一元线性回归模型还有一个未知数，即总体随机误差项 ε 的方差 σ^2 。它是检验模型时必须利用的一个重要参数，用以反映理论模型误差的大小。由于随机误差项 ε 本身不能直接观测，因此需要用样本回归模型的残差 e 代替随机误差项来估计 σ^2 。数学上可以证明， σ^2 的无偏估计 $\hat{\sigma}^2$ 可由下式给出：

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

(五) 总体方差的估计

对 σ^2 开平方, 得
$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

$\hat{\sigma}$ 被称为回归标准差或估计标准差. 正如标准差可以说明平均数代表性大小一样, 估计标准差则可以说明回归线代表性的大小.

$\hat{\sigma}$ 越小表明实际观测值与所拟合的样本回归线的离差程度越小, 即回归具有较强的代表性. 反之, $\hat{\sigma}$ 越大表明实际观测值与所拟合的样本回归线的离差程度越大, 即回归线的代表性越差.

(五) 总体方差的估计

如果利用上述定义公式手工计算估计标准误差时需要求出每一项残差，计算工作较大。因此可以采用下列简捷公式计算：

$$\hat{\sigma} = \sqrt{\frac{\sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i}{n-2}}$$

上述简捷公式中所需的数据与计算相关系数和估计回归系数时所用数据相同，这样可以大减计算工作量，当然，如果是利用统计软件计算估计校准差，则无所谓简捷计算公式。

三、显著性检验

当我们建立了一个实际问题的经验回归方程之后，还不能马上用于分析和预测。因为得到的经验回归方程是否真正描述了变量之间的统计规律性，还需要运用统计方法对回归方程进行检验。

回归分析中的显著性包括两方面的内容：

一是对各回归系数的显著性检验（**t检验**）；

二是对整个回归方程的显著性检验（**F检验**）。

对于前者，通常采用t检验，而对于后者则是在方差分析的基础上采用F检验。在一元线性回归模型中，由于只有一个自变量，对 $\beta_1 = 0$ 的t检验与对整个方程的F检验是一致的。

(一) t检验

t检验是统计推断中常用的一种检验方法。在回归分析中主要用于检验回归系数的显著性。检验的原假设是

$$H_0: \beta_1 = 0, \text{ 备择假设是 } H_1: \beta_1 \neq 0。$$

回归系数的显著性检验就是检验因变量y对自变量x的影响程度是否显著。如果原假设 H_0 成立，则因变量y与自变量x之间并没有真正的线性关系，也就是说自变量x的变化对因变量y并没有影响。

由于 $\beta_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$ ，因而当原假设 $H_0: \beta_1 = 0$ 成立

时，有 $\beta_1 \sim N(0, \frac{\sigma^2}{L_{xx}})$

(一) t检验

此时, $\hat{\beta}_1$ 在零附近波动, 构造 t 统计量

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}}$$

$$\text{式中: } \hat{\sigma}^2 = \frac{1}{n-2} \sum e_i^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)$$

是 σ^2 的无偏估计, 称 $\hat{\sigma}$ 为回归标准差.

(一) t检验

当原假设 $H_0: \beta_1 = 0$ 成立时,由式(8.14)构造的 t 统计量服从自由度 $n-2$ 为的 t 分布.给定显著性水平 α ,双侧检验的临界值为 $t_{\alpha/2}$.当时 $|t| \geq t_{\alpha/2}$ 拒绝原假设 $H_0: \beta_1 = 0$,认为 β_1 显著不为零,因变量 y 对自变量 x 的一元线性回归成立;当 $|t| < t_{\alpha/2}$ 时,接受原假设 $H_0: \beta_1 = 0$,认为 β_1 为零,因变量 y 对自变量 x 的一元线性回归不成立.

(二) F检验

1.F检验的意义

t检验主要用来检验各个回归系数是否显著，F检验则主要用于检验整个回归方程是否有效。对于一元线性回归模型，由于只有一个回归系数，两种检验所得的结果是相同的。但对于多元线性回归模型则不同，t检验与F检验的结果可能相同也可能不相同，即会出现各个回归系数能通过检验而整个回归方程却不一定能够通过检验的情形，或者出现相反的情形。F检验的主要目的在于分析各个因变量值与其均值离差平方和中，由于自变量与因变量之间的回归关系所产生的影响情况。

(二) F检验

2.F检验的思想

F检验的目的，在于对回归方程的线性关系的显著性进行检验。不难理解，如果Y与X之间的线性关系显著，那么随着X朝某个方向变动，Y将会比较紧密地围绕一条直线朝某个方向变动，Y绕该条直线越紧密，则Y与X的线性关系越显著。此时，不论Y的波动变化如何，它总是以该条直线为中心变动的。这种波动既包括了X对Y的系统影响，也包括了随机因素对Y的影响。我们可以通过分析Y的总变动中，是X对Y的系统影响大，还是随机因素对Y的影响大，如果X对Y的系统影响大，就说明两者之间的线性关系越显著。所以，对回归方程的线性关系进行显著性检验是从分析Y的总波动原因入手的。

$$y = \beta_0 + \beta_1 x$$

线性关系：Y直线变动

$$+\varepsilon$$

随机干扰：Y绕直线震荡变化

(二) F检验

3.Y的波动原因分析——总离差平方和分解

根据方差分析原理，将 Y 的 n 个观察值之间的差异，用观察值 y_i 与其平均值 \bar{y} 的离差平方各来表示，并称之为总离差平方和，记为 **SST**。

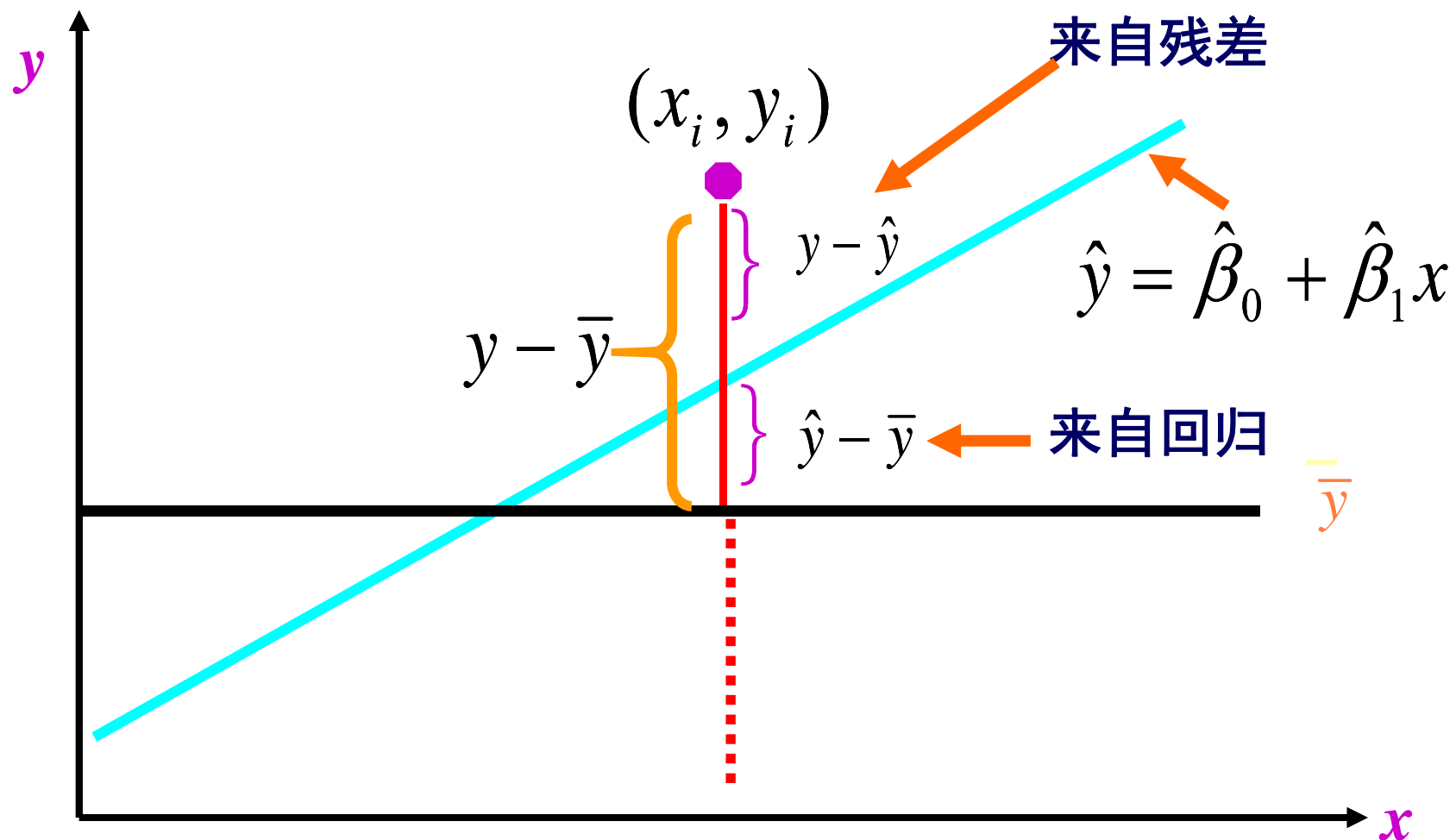
将**SST**分解如下：

$$\begin{aligned} SST &= \sum (y_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i)^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \end{aligned}$$

其中可以证明

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

(二) F检验



(二) F检验

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y})^2$$

总平方和
(SST)

回归平方和
(SSR)

残差平方和
(SSE)

$$SST = SSR + SSE$$

(二) F检验

4.三个平方和的意义

(1)总平方和(SST)

反映因变量的 n 个观察值与其均值的总离差

(2)回归平方和(SSR)

反映自变量 x 的变化对因变量 y 取值变化的影响，或者说，是由于 x 与 y 之间的线性关系引起的 y 的取值变化，也称为可解释的平方和

(3)残差平方和(SSE)

反映除 x 以外的其他因素对 y 取值的影响，也称为不可解释的平方和或剩余平方和

(二) F检验

4. 检验统计量的构造

直观地看，在SST中，如果SSR大于SSE，说明Y的变化主要受X的影响，两者之间线性关系明显。但由于SSR与SSE的大小与各自的自由度有关，故需要将其除以各自的自由度后才能直接比较。于是构造如下检验统计量：

$$F = \frac{SSR / p}{SSE / (n - p - 1)} \sim F(p, n - p - 1)$$

一元回归中：

$$F = \frac{SSR / 1}{SSE / (n - 2)} \sim F(1, n - 2)$$

(二) F检验

5.检验规则

若 $F \geq F_{\alpha}(1, n-2)$:

则拒绝 $H_0 : \beta_1 = 0$, 说明变量之间存在显著的线性关系;

若 $F < F_{\alpha}(1, n-2)$:

则接受 $H_0 : \beta_1 = 0$, 说明变量之间没有显著的线性关系。

(二) F检验

6. 方差分析表

表8.3 方差分析表

方差来源	平方和	自由度	均方误差	F统计量	显著性水平 (P值)
回归	SSR	1	SSR/1	$\frac{SSR/1}{SSE/n-2}$	F检验统计量大于临界值的概率
残差	SSE	n-2	SSE/ n-2		
总和	SST	n-1			

(三) 样本决定系数

根据回归平方和与残差平方和的意义，我们知道如果在总的离差平方和中回归平方和所占的比重越大，则线性回归效果就越好，这说明回归直线与样本观测值拟合优度就越好；如果残差平方和所占的比重越大，则回归直线与样本观测值拟合得就不理想。这里把回归平方和与总离差平方和之比定义为样本决定系数，记为 r^2 ，即

$$r^2 = \frac{SSR}{SST} = \frac{L_{xx}}{L_{xx} L_{yy}} = (r)^2$$

(四) 样本决定系数

由关系式: $\sum (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 \sum (\hat{x}_i - \bar{x})^2$, 可以证明上式的 r^2 正好是(8.1)式中相关系数 r 的平方:

$$r^2 = \frac{SSR}{SST} = \frac{L_{xx}}{L_{xx} L_{yy}} = (r)^2$$

四、回归模型的应用

(一) 单值预测

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

这就是变量新值： $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ 的单值预测.

(二) 区间预测

1. 因变量单个值区间估计

为了给出新值 y_0 的置信区间,需要求出其估计值 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 的分布,由于 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 都是 y_1, y_2, \dots, y_n 的线性组合,因而 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 也是 y_1, y_2, \dots, y_n 的线性组合,在正态假设定下 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 服从正态分布,其期望值为 $E(\hat{y}_0) = \beta_0 + \beta_1 x_0$.其方差计算如下:

1. 因变量单个值区间估计

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y}_0 - \hat{\beta}_1 \bar{x} + \hat{\beta}_0 x_0 = \sum \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_i - \bar{x})}{L_{xx}} \right] y_i$$

$$\text{var}(\hat{y}_0) = \sum \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_i - \bar{x})}{L_{xx}} \right]^2 \text{var}(y_i) = \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_i - \bar{x})}{L_{xx}} \right] \sigma^2$$

$$\text{可得到 } \hat{y}_0 \sim N \left\{ \beta_0 + \beta_1 x_0, \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_i - \bar{x})}{L_{xx}} \right] \sigma^2 \right\}$$

1. 因变量单个值区间估计

\hat{y}_0 是先前独立观测到的随机变量 y_1, y_2, \dots, y_n 的线性组合, 新值 y_0 与先前的观测值是独立的, 所以 y_0 与 \hat{y}_0 是独立的, 因而

$$\begin{aligned}\text{var}(y_0 - \hat{y}_0) &= \text{var}(y_0) + \text{var}(\hat{y}_0) \\ &= \sigma^2 + \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right] \sigma^2 = \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right) \sigma^2\end{aligned}$$

再由 $E(\hat{y}) = E(y)$ 有 $E(y_0 - \hat{y}_0) = 0$

于是有

$$y_0 - \hat{y}_0 \sim N\left(0, \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}\right) \sigma^2\right)$$

1. 因变量单个值区间估计

可知

$$t = \frac{y_0 - \hat{y}_0}{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}\right) \sigma^2} \sim t(n-2)$$

于是有

$$P \left\{ \left| \frac{y_0 - \hat{y}_0}{\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}} \hat{\sigma}} \right| \leq t_{\alpha/2}(n-2) \right\} = 1 - \alpha$$

1. 因变量单个值区间估计

由此我们可以求得 y_0 的置信概率 $1-\alpha$ 为的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2) \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}} \hat{\sigma} \dots \dots (8.15)$$

当样本容量 n 较大, $|x_0 - \bar{x}|$ 较小时, y_0 的置信度为95%的置信区间近似为

$$\hat{y}_0 \pm 2\hat{\sigma}$$

2. 因变量均值区间估计

(8.15)式给出的是因变量单个值的置信区间,我们关心的另外一种情况是因变量均值的置信区间.对于前面提出的人均消费性支出问题,如果有好几个地区的人均可支配收入同为 x_0 ,那么这些地区对应的人均消费性支出的平均数为多少?

这个问题就是要估计平均值 $E(y_0)$. $E(y_0)$ 的区间估计与因变量单个值 y_0 的置信区间有所不同,由于 $E(y_0) = \beta_0 + \beta_1 x_0$ 为常数,由(8.14)可知

2. 因变量均值区间估计

$$\hat{y}_0 - E(y_0) \sim N\left(0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}\right)\sigma^2\right)$$

进而得因变量均值 $E(y_0)$ 置信水平为 $1-\alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2) \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}} \hat{\sigma}$$

第三节 多元线性回归分析

一、多元线性回归模型

因为客观现象非常复杂，现象之间的联系方式和性质各不相同，影响因变量变化的自变量往往是多个而不仅仅是一个，其中既有主要因素也有次要因素。如果仅仅进行一元回归分析，不一定能得到满意的结果。因此，有必要将一个因变量与多个自变量联系起来进行分析。

在线性相关条件下，研究两个和两个以上自变量对一个因变量的数量变化关系，称为多元线性回归分析，表现这一数量关系的数学表达式则称为多元线性回归方程或多元线性回归模型。

一、多元线性回归模型

(一)多元线性回归模型的一般形式

设随机变量 y 与一组自变量 x_1, x_2, \dots, x_p 的线性回归模型为:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \dots \dots \dots (8.17)$$

式中, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 是个未知参数, β_0 称为回归常数,
 $\beta_1, \beta_2, \dots, \beta_p$ 称为回归系数. y 称为被解释变量(因变量),
而 x_1, x_2, \dots, x_p 是 p 个可以观测并可控制的般变量,称为解释变量(自变量).

(一)多元线性回归模型的一般形式

- 对于一个实际问题,如果我们获得了 n 组观测数据 $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i), i = 1, 2, \dots, n$, 则线性回归模型 (8.17) 式可表示为

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + \varepsilon_2 \\ \dots\dots\dots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

(一)多元线性回归模型的一般形式

写成矩阵形式: $y = X\beta + \varepsilon \dots \dots \dots (8.20)$

式中:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

矩阵 X 是一个 $n \times (p+1)$ 的矩阵,人们常称为回归设计矩阵或资料,在实验设计中, X 的元素是预先设定并可以控制的,人的主观因素可作用于其中,因而 X 称为设计矩阵.

(二) 多元线性回归模型的基本假定

1、 解释变量 x_1, x_2, \dots, x_p 是确定性变量，不是随机变量，且要求 $\text{rank}(X) = p + 1 < n$ ；

2、 随机误差项具有**0**均值和等方差，即

$$\begin{cases} E(\varepsilon_i) = 0, i = 1, 2, \dots, n \\ \text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases} (i, j = 1, 2, \dots, n) \end{cases}$$

3、 正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \dots, n \\ (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n) \text{ 是 } iid \end{cases}$$

(三) 多元线性回归方程的解释

- 在建立住房的预测模型时，用 y 来表示住房的销售量，用 x_1 表示住房的价格， x_2 表示消费者人均可支配收入。则可建立二元线性回归方程模型：
$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \end{cases}$$
- 在式（8.21）中，假如 x_2 保持不变，为一常数时，则有：
$$\frac{\partial E(y)}{\partial x_1} = \beta_1$$

(三) 多元线性回归方程的解释

- 即 β_1 可解释为在消费者人均收入 x_1 保持不变高时，住房价格 x_1 每变动一个单位，对住房销售量 y 的平均影响程度。一般来说，随着物价的提高，销售量是减少的，因此 β_1 将是负的。

- 在 (8.21) 式中，假如 x_1 保持不变，为一常数时，则有：

$$\frac{\partial E(y)}{\partial x_2} = \beta_2$$

- 即 β_2 可解释为在住房价格 x_1 保持不变时，消费者人均可支配收入 x_2 每变动一个单位，对住房销售量 y 的平均影响程度。一般来说，随着消费者人均可支配收入的增加，住房销售量是增加的，因此 β_2 将是正的。

二、多元线性回归模型的估计

(一)回归系数的估计

多元线性回归方程未知参数 $\beta_0, \beta_1, \dots, \beta_p$ 的估计与一元线性回归方程的参数估计原理一样,仍然可以采用最小二乘估计.对于(8.20)式矩阵形式表示的回归模型 $y = x\beta + \varepsilon$, 所谓最小二乘法,就是寻找参数 $\beta_0, \beta_1, \dots, \beta_p$ 的估计值 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 使离差平方和

$$Q(\beta_0, \beta_1, \dots, \beta_p) = \sum (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2$$

达到极小,即寻找 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 满足以下关系

(一) 回归系数的估计

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2 \\ &= \min_{\beta_0, \beta_1, \dots, \beta_p} \sum (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2 \dots \dots \dots (8.22) \end{aligned}$$

依照(8.22)式求出的 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 就称为回归参数 $\beta_0, \beta_1, \dots, \beta_p$ 的最小二乘估计.

从(8.22)式中求出 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 是一个求极值问题, 由于Q是关于 $\beta_0, \beta_1, \dots, \beta_p$ 的非负二次函数, 因而这的最小值总是存在的. 根据微积分中求极值的原理, $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 应满足下列方程组

(一) 回归系数的估计

$$\left\{ \begin{array}{l} \frac{\partial Q}{\partial \beta_0} \Big|_{\beta_0=\hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) = 0 \\ \frac{\partial Q}{\partial \beta_1} \Big|_{\beta_1=\hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{i1} = 0 \\ \dots\dots\dots \\ \frac{\partial Q}{\partial \beta_p} \Big|_{\beta_p=\hat{\beta}_p} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{ip} = 0 \end{array} \right.$$

以上方程组经整理后,得到用矩阵形式表示的正规方程组

$$X'(y - X\hat{\beta}) = 0$$

(一) 回归系数的估计

移项得:

$$X'X\hat{\beta} = X'y$$

当存在时,即得回归参数的最小二乘估计为:

$$\hat{\beta} = (X'X)^{-1}X'y$$

称: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \cdots + \hat{\beta}_px_p$

为经验回归方程.

(二) 最小二乘估计量的性质

- 数学上可以证明,在基本假定条件可以得到满足的情况下,多元回归模型中回归系数最小二乘估计量的期望值同样等于总体回归系数的真值,即 $\hat{\beta}$ 是 β 的无偏估计,可以表示为:

$$E(\hat{\beta}) = \beta$$

- 回归系数最小二乘估计量的方差,协方差矩阵为

$$D(\hat{\beta}) = \text{cov}(\hat{\beta}, \hat{\beta})$$

$$= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)$$

$$= \sigma^2 (X'X)^{-1}$$

(二) 最小二乘估计量的性质

该矩阵主对角元素是各回归系数估计量的方差 $E(\hat{\beta}_j - \beta_j)^2$, 其他元素是各回归系数估计量之间的协方差 $E(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j), (i \neq j)$. 在此基础上, 还可以进一步证明回归系数的最小二乘估计量是最优线性无偏估计量和一致估计量, 也就是说, 在基本假设条件得到满足的多元线性回归模型中, 高斯·马尔柯夫定理同样成立.

(三) 总体方差的估计

- 除了回归系数以外,多元线性回归模型中还包含了另一个未知参数,即随机误差项的方差 σ^2 ,与一元线性回归分析相似,多元线性回归模型中的 σ^2 也是利用残差平方和除以其自由度来估计.即有

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-p-1} SSE \\ &= \frac{1}{n-p-1} (e'e) = \frac{1}{n-p-1} \sum e_i^2\end{aligned}$$

- $\hat{\sigma}^2$ 是误差项方差 σ^2 的无偏估计量.

三、多元线性回归模型的检验

(一)回归系数的显著性检验(t检验)

多元线性回归模型中回归系数的检验同样采用t检验,其原理和基本步骤与一元回归模型中的t检验基本相同,就不在详细说明了.检验自变量 x_j 对因变量 y 的影响是否显著,等价于检验假设:

$$H_{0j} : \beta_j = 0, j = 1, 2, \dots, n$$

如果接受原假设 H_{0j} , 则 x_j 对 y 的影响不显著;如果拒绝原假设 H_{0j} , 则 x_j 对 y 的影响是显著的.

(一)回归系数的显著性检验(检验)

由上述讨论可知 $\hat{\beta}$ 服从均值为 β , 方差为 $\sigma^2(X'X)^{-1}$ 的正态分布, 即

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

记

$$(X'X)^{-1} = (c_{ij}), i, j = 0, 1, 2, \dots, p$$

(一)回归系数的显著性检验(t检验)

- 于是有 $E(\hat{\beta}_j) = \beta_j, \text{var}(\hat{\beta}_j) = c_{jj}\sigma^2$
$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2), j = 0, 1, 2, \dots, p$$
- 据此可以构造t统计量: $t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}} \dots (8.26)$
- 式中 c_{jj} 为矩阵主对角线 $(X'X)^{-1}$ 上第j个元素
$$\hat{\sigma} = \sqrt{\frac{1}{n-p-1} \sum e_i^2} = \sqrt{\frac{1}{n-p-1} \sum (y_i - \hat{y}_i)^2} \dots (8.27)$$
- 是回归标准差。

回归系数的显著性检验(t 检验)

当原假设 $H_{0j} : \beta_j = 0$ 成立时,(8.26)式构造的 t_j 统计量服从自由度为 $n - p - 1$ 的 t 分布.给定显著性水平 α ,查出双侧检验的临界值 $t_{\alpha/2}$.当 $|t_j| \geq t_{\alpha/2}$ 时拒绝原假设 $H_{0j} : \beta_j = 0$,认为 β_j 显著不为零,自变量 x_j 对因变量 y 的线性效果显著;当 $|t_j| < t_{\alpha/2}$ 时接受原假设 $H_{0j} : \beta_j = 0$,认为 β_j 为零,自变量 x_j 对因变量 y 的线性效果不显著.

(二) 回归方程显著性的F检验

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

- 如果 H_0 被接受，则表明随机变量 y 与 x_1, x_2, \cdots, x_p 之间的关系用线性回归模型表示不合适。类似一元线性回归检验，为了建立对 H_0 进行检验的F统计量，仍然利用总离差平方和的分解式，即；

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \cdots \cdots (8.28)$$

- 简写为 $SST=SSR+SSE$
- 构造F检验统计量如下：

$$F = \frac{SSR / p}{SSE / u - p - 1} \cdots \cdots (8.29)$$

(二) 回归方程显著性的F检验

在正态假设下，当原假设 $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 成立时，F服从自由度为 $(p, n-p-1)$ 的F分布。于是，可以利用F统计量对回归方程的总体显著性进行检验。对于给定的数据， $i = 1, 2, \cdots, n$ 计算出SSR和SSE，进而得到F的值，其计算过程一般列在表8.4的方差分析表中，再由给定的显著性水平 α 查F分布表，得临界值

$$F_{\alpha}(p, n-p-1)$$

(二)回归方程显著性的F检验

当 $F > F_{\alpha}(p, n-p-1)$ 时，拒绝原假设 H_0 ，认为在显著性水平 α 下， y 对 x_1, x_2, \dots, x_p 有显著的线性关系，也即认为回归方程是显著的。反之，当 $F \leq F_{\alpha}(p, n-p-1)$ 时，则认为回归方程不显著。

表8.4 方差分析表

方差来源	平方和	自由度	均方误差	F统计量	显著性水平 (P值)
回归	SSR	p	SSR/p	$\frac{SSR/p}{SSE/(n-p-1)}$	F检验统计量大于临界值的概率
残差	SSE	n-p-1	SSE/ n-p-1		
总和	SST	n-1			

(三) 回归系数的置信区间

由 $\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$, $j = 0, 1, 2, \dots, p$, 可知

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n - p - 1)$$

仿照一元线性回归系数区间估计的推导过程, 可得 β_j 的置信度为 $1 - \alpha$ 的置信区间为:

$$(\hat{\beta}_j - t_{\alpha/2} \sqrt{c_{jj}}\hat{\sigma}, \hat{\beta}_j + t_{\alpha/2} \sqrt{c_{jj}}\hat{\sigma})$$

四、拟合优度检验

- 在多元线性回归分析中，总离差平方和的分解公式依然成立。因此，也可以利用上一节所定义的样本决定系数作为评价模型拟合优度的一项指标。但为了避免混淆，多元回归的决定系数用 R^2 表示，并称为复决定系数，即定义样本复决定系数为

$$R^2 = SSR/SSE = 1 - (SSE/SST) \dots\dots\dots (8.30)$$

- 由样本复决定系数定义可知， R^2 的大小取决于残差平方和SSE在总离差平方和SST中所占的比重.在样本容量一定的条件下,总离差平方和与自变量的个数无关,而残差平方和则会随着模型中自变量个数的增加而不断减少,至少不会增加.

四、拟合优度检验

因此在多元回归分析中应该使用自由度调整后的决定系数,即利用各自的自由度对总离差平方和与残差平方和进行调整,然后再计算调整后的决定系数 R_{α}^2 。

$$\begin{aligned} R_{\alpha}^2 &= 1 - \frac{SSE / (n - p - 1)}{SST / (n - 1)} \\ &= 1 - \frac{n - 1}{n - p - 1} (1 - R^2) \end{aligned}$$