

第 3 章 多元线性回归

在实际应用中,我们常常需要考虑多个因素对目标变量的影响,而此时简单的一元线性回归模型已无法满足我们的需求。为了更准确地描述变量之间的复杂关系,我们引入多元线性回归模型。多元线性回归模型是一种用于分析多个解释变量(也称为自变量)与一个因变量之间线性关系的统计模型。本章将深入探讨多元线性回归模型的基本概念和原理,以及其在经济学和其他领域中的应用。通过学习本章,读者将全面了解多元线性回归模型的建模过程和推导方法。

3.1 多元线性回归模型

3.1.1 多元线性回归模型

令 $\{(X_{i1}, X_{i2}, \dots, X_{ik}, Y_i) (i=1, 2, \dots, n)\}$ 表示从总体中抽取的容量为 n 的随机样本,多元线性回归模型的一般形式为

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \mu_i \quad (3-1)$$

其中 k 为解释变量的数目, $\beta_j (j=0, 1, \dots, k)$ 称为回归参数(regression coefficient)。习惯上把常数项看成一个虚变量的参数,在参数估计过程中该虚变量的样本观测值始终取 1。这样,模型中解释变量的数目为 $(k+1)$ 。同一元回归分析一样,式(3-1)也被称为总体回归模型(population regression model)。它的非随机表达式称为总体回归函数(population regression function),表达式为

$$E(Y_i | X_{1i}, X_{2i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} \quad (3-2)$$

可见,多元回归分析是以多个解释变量的固定值为条件的回归分析,方程表示各变量 X 值固定时 Y 的平均响应。 β_j 也被称为偏回归系数(partial regression coefficient),表示在其他解释变量保持不变的情况下, X_j 每变化 1 个单位时, Y 的均值 $E(Y)$ 的变化,或者说 β_j 给出 X_j 的单位变化对 Y 均值的“直接”或“净”(不含其他变量)影响。

由式(3-1)表示的 n 个随机方程的矩阵表达式为

$$Y = X\beta + \mu \quad (3-3)$$

其中

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n \times (k+1)}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1) \times 1}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}_{n \times 1}$$

同一元回归分析相仿,在给出总体中的一个样本时,我们估计样本回归函数(sample regression function),并让它近似代表未知的总体回归函数。

样本回归函数可表示为

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki} \quad (3-4)$$

样本回归模型为

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki} + e_i \quad (3-5)$$

其中 e_i 称为残差或剩余项(residuals),可看成总体回归函数中随机误差项 μ_i 的近似替代。

同样地,式(3-4)与式(3-5)中样本回归函数的矩阵表达式分别为

$$\hat{Y} = X\hat{\beta} \quad (3-6)$$

$$Y = X\hat{\beta} + e \quad (3-7)$$

其中

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

3.1.2 多元线性回归模型的基本假定

为了使参数估计量具有良好的统计性质,对多元线性回归模型式(3-1)或式(3-3)的随机误差项可作出与一元回归分析类似的若干基本假定。

假定 1,解释变量 $X_1, X_2, \dots, X_k$ 是非随机的或固定的,且各 X 之间互不相关,即各 X 间无多重共线性(no multicollinearity)。

假定 2,随机误差项具有零均值。

$$E(\mu_i) = 0 \quad (i = 1, 2, \dots, n)$$

假定 3,随机误差项具有同方差。

$$\text{Var}(\mu_i) = E(\mu_i^2) = \sigma^2 \quad (i = 1, 2, \dots, n)$$

假定 4,随机误差项不序列相关。

$$\text{Cov}(\mu_i, \mu_j) = E(\mu_i \mu_j) = 0 \quad (i \neq j; i, j = 1, 2, \dots, n)$$

假定 5,解释变量与随机误差项不相关。

$$\text{Cov}(X_{ji}, \mu_i) = 0 \quad (j = 1, 2, \dots, k; i = 1, 2, \dots, n)$$

假定 6,随机误差项满足正态分布。

$$\mu_i \sim N(0, \sigma^2)$$

假定 1~假定 4 以及假定 6 可用矩阵表示如下。

假定 1, $n \times (k+1)$ 矩阵 $\mathbf{X}$ 是非随机的, 且 $\mathbf{X}$ 的秩 $\rho(\mathbf{X}) = k+1$, 即 $\mathbf{X}$ 列满秩。

$$\text{假定 2, } E(\boldsymbol{\mu}) = E \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E(\mu_1) \\ \vdots \\ E(\mu_n) \end{pmatrix} = \mathbf{0}$$

假定 3 和假设 4 可以合并为

$$\begin{aligned} E(\boldsymbol{\mu}\boldsymbol{\mu}') &= E \begin{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} & (\mu_1 \quad \cdots \quad \mu_n) \end{pmatrix} = E \begin{pmatrix} \mu_1^2 & \cdots & \mu_1\mu_n \\ \vdots & & \vdots \\ \mu_n\mu_1 & \cdots & \mu_n^2 \end{pmatrix} \\ &= \begin{pmatrix} \text{Var}(\mu_1) & \cdots & \text{Cov}(\mu_1, \mu_n) \\ \vdots & & \vdots \\ \text{Cov}(\mu_n, \mu_1) & \cdots & \text{Var}(\mu_n) \end{pmatrix} = \begin{pmatrix} \sigma^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I} \end{aligned}$$

假定 6, 向量 $\boldsymbol{\mu}$ 服从多维正态分布, 即

$$\boldsymbol{\mu} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

3.2 多元线性回归模型的参数估计

同一元回归模型的估计一样, 多元回归模型参数估计的目标仍有两项: 一是求得变量之间数量关系的结构参数的估计量 $\hat{\beta}_j$ ($j=1, 2, \dots, k$); 二是求得随机误差项的方差估计 $\hat{\sigma}^2$ 。式(3-1)或式(3-2)在满足 3.1 节所列的基本假定的情况下, 可以采用普通最小二乘法进行参数估计。

3.2.1 普通最小二乘估计

随机抽取被解释变量和解释变量的 n 组样本观测值

$$(Y_i, X_{ji}) \quad (i=1, 2, \dots, n; j=0, 1, 2, \dots, k)$$

如果样本函数的参数估计值已经得到, 则有

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki} \quad (i=1, 2, \dots, n) \quad (3-8)$$

那么, 根据普通最小二乘原理, 参数估计值应该是下列方程组的解

$$\begin{cases} \frac{\partial}{\partial \hat{\beta}_0} Q = 0 \\ \frac{\partial}{\partial \hat{\beta}_1} Q = 0 \\ \frac{\partial}{\partial \hat{\beta}_2} Q = 0 \\ \vdots \\ \frac{\partial}{\partial \hat{\beta}_k} Q = 0 \end{cases} \quad (3-9)$$

其中

$$\begin{aligned}
 Q &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\
 &= \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki})]^2
 \end{aligned} \quad (3-10)$$

于是得到关于待估参数估计值的正规方程组

$$\begin{cases}
 \sum Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} + \cdots - \hat{\beta}_k X_{ki} = 0 \\
 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} + \cdots - \hat{\beta}_k X_{ki}) X_{1i} = 0 \\
 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} + \cdots - \hat{\beta}_k X_{ki}) X_{2i} = 0 \\
 \vdots \\
 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} + \cdots - \hat{\beta}_k X_{ki}) X_{ki} = 0
 \end{cases} \quad (3-11)$$

式(3-11)也可以表示为

$$\begin{cases}
 \sum e_i = 0 \\
 \sum e X_{1i} = 0 \\
 \sum e X_{2i} = 0 \\
 \vdots \\
 \sum e X_{ki} = 0
 \end{cases} \quad (3-12)$$

或用矩阵表示为

$$\begin{bmatrix} \sum e_i \\ \sum X_{2i} e_i \\ \vdots \\ \sum X_{ki} e_i \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{X}' \mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

求解式(3-11) $(k+1)$ 个方程组成的线性代数方程组,即可得到 $(k+1)$ 个待估参数的估计值 $\hat{\beta}_j$ ($j=0,1,2,\cdots,k$)。

因为样本回归模型为 $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$,两边左乘 $\mathbf{X}'$ 可以得到

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{e} \quad (3-13)$$

由于 $\mathbf{X}'\mathbf{e} = \mathbf{0}$

因此式(3-13)可变为

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

结合基本假定 1, $\mathbf{X}$ 列满秩,因此有 $\mathbf{X}'\mathbf{X}$ 满秩,从而可以推出 $\mathbf{X}'\mathbf{X}$ 可逆。

因此参数 $\boldsymbol{\beta}$ 的普通最小二乘估计结果可以表示为

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3-14)$$

当 $k=2$ 时,可以得到二元线性回归模型参数的普通最小二乘估计结果为

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \quad (3-15)$$

$$\hat{\beta}_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \quad (3-16)$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \quad (3-17)$$

在第2章得到一元线性回归参数的普通最小二乘估计结果为

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (3-18)$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (3-19)$$

可以把二元线性回归参数的估计结果与一元线性回归参数的估计结果进行对比。通过对比可以发现,当增加一个解释变量时,普通最小二乘估计结果的表达式复杂了很多。因此,对于多元线性回归,采用矩阵的形式表示其估计结果能带来很大的方便。

最后,可以证明(证明过程略),随机误差项 μ 的方差的无偏估计量为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1} = \frac{e'e}{n-k-1} \quad (3-20)$$

3.2.2 参数估计量的性质

在多元线性回归模型满足基本假定的情况下,其参数的普通最小二乘估计仍具有线性、无偏性和有效性。

1. 线性

由于

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{Y}$$

其中, $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ 为仅与给定的 $\mathbf{X}$ 有关的矩阵。可见,参数估计量是被解释变量 $\mathbf{Y}$ 的线性组合。

2. 无偏性

参数估计量 $\hat{\beta}$ 的无偏性证明如下。

$$\begin{aligned} E(\hat{\beta}) &= E((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}) \\ &= E((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\beta + \mu)) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'E(\mu) \\ &= \beta \end{aligned} \quad (3-21)$$

这里利用了随机误差项均值为零的假设,即 $E(\mu) = 0$ 。

3. 有效性

首先给出参数估计量 $\hat{\beta}$ 的方差-协方差矩阵

$$\begin{aligned}
\text{Cov}(\hat{\beta}) &= E(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))' \\
&= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\
&= E((\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}'\mu\mu'\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mu\mu')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= E(\mu\mu')(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned} \tag{3-22}$$

其中利用了

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mu) \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mu
\end{aligned}$$

和

$$E(\mu\mu') = \sigma^2\mathbf{I}$$

$\mathbf{I}$ 为单位矩阵。

根据高斯-马尔可夫定理,式(3-22)表示的方差在所有无偏估计量的方差中是最小的。所以该参数估计量具有有效性。

3.3 多元线性回归模型的统计检验

在多元线性回归模型的参数估计出来后,即求出样本回归函数后,还需进一步进行统计检验,以判断该样本回归函数估计的可靠程度。统计检验包括对模型拟合优度的检验、回归方程的显著性检验、参数的显著性检验以及参数的置信区间等方面。

3.3.1 拟合优度

在一元线性回归模型中,可以使用可决系数 R^2 来衡量样本回归线对样本观测值的拟合程度。在多元线性回归模型中,也可以采用可决系数 R^2 来判断样本回归线对样本观测值的拟合程度。

记 $\text{TSS} = \sum (Y_i - \bar{Y})^2$ 为总离差平方和, $\text{ESS} = \sum (\hat{Y}_i - \bar{Y})^2$ 为回归平方和, $\text{RSS} = \sum (Y_i - \hat{Y}_i)^2$ 为剩余平方和,则有

$$\begin{aligned}
\text{TSS} &= \sum (Y_i - \bar{Y})^2 \\
&= \sum ((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}))^2 \\
&= \sum (Y_i - \hat{Y}_i)^2 + 2 \sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum (\hat{Y}_i - \bar{Y})^2
\end{aligned}$$

由于

$$\begin{aligned}
 \sum (Y_i - \hat{Y})(\hat{Y}_i - \bar{Y}) &= \sum e_i (\hat{Y}_i - \bar{Y}) \\
 &= \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum e_i X_{1i} + \cdots + \hat{\beta}_k \sum e_i X_{ki} + \bar{Y} \sum e_i \\
 &= 0
 \end{aligned}$$

所以有

$$TSS = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 = RSS + ESS \quad (3-23)$$

即总离差平方和可分解为剩余平方和与回归平方和两部分。回归平方和反映了总离差平方和中可由样本回归线解释的部分,它越大,剩余平方和越小,表明样本回归线与样本观测值的拟合程度越高。因此,可用回归平方和占总离差平方和的比重来衡量样本回归线对样本观测值的拟合程度,即

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad (3-24)$$

该统计量越接近 1,模型的拟合优度越高。

剩余平方和有一个显著特点,即当在模型里增加一个解释变量,然后用修改后的模型重新估计时,剩余平方和的值会减小,或者严格地说,至少不会增大。在实际应用中,我们会发现,如果在模型里增加一个解释变量,那么拟合优度 R^2 往往会随之增大。这就导致一种情况:只要不断增加模型中解释变量的数量,就能增大拟合优度 R^2 的值。因此容易给人造成一种错误认知,即若想让模型拟合效果好,单纯增加解释变量的个数就可以了。但实际情况并非如此,增加解释变量个数导致的拟合优度 R^2 增大,与模型真正的拟合优劣并没有实质关联。所以,在对多元回归模型的可靠程度进行比较时,拟合优度 R^2 并非一个恰当的衡量指标,必须对其进行相应的调整,才能更准确地评估模型的可靠性。

在样本容量固定的前提下,每增加一个解释变量,模型的自由度就会相应减少。鉴于此,为了更准确地衡量模型的拟合效果,避免因变量个数的变化对拟合优度评估产生干扰,我们需要对拟合优度的计算方式进行调整。具体的调整思路是,将剩余平方和除以它对应的自由度,同时把总离差平方和也除以其对应的自由度,通过这种方式来剔除变量个数对拟合优度的影响,从而得到更具参考价值的拟合优度指标,以便更科学地评估模型的拟合效果。记 $\bar{R}^2$ 为调整的可决系数(adjusted coefficient of determination),则有

$$\bar{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)} \quad (3-25)$$

式(3-25)中 $(n-k-1)$ 为剩余平方和的自由度, $(n-1)$ 为总离差平方和的自由度。可以看到,当增加一个解释变量时,如果新增加的解释变量并没有解释能力,对剩余平方和(RSS)的减小就不能提供贡献,而增加待估参数的个数,将使得 $\bar{R}^2$ 有较大幅度的下降。

调整的可决系数与未经调整的可决系数之间的关系为

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k-1} \quad (3-26)$$

需要注意的是,可决系数 R^2 必定非负,但按照式(3-26)所计算的调整的可决系数 $\bar{R}^2$ 有可能为负值。解决办法:若计算的 $\bar{R}^2 < 0$,规定 $\bar{R}^2$ 取值为 0。

在实际运用中,判断一个模型拟合程度 $\bar{R}^2$ 是否良好,并没有一个绝对标准。模型的拟合优度数值达到多少才算是拟合程度较好,往往因具体的应用场景、研究目的和数据特征等

因素而异。另外,拟合优度虽然是衡量模型表现的一个重要指标,但它绝非判断模型质量的唯一依据。在某些特定情况下,尤其是当模型的经济意义与拟合优度发生冲突时,为确保模型能够准确反映经济理论和实际经济关系,我们可能会适当降低对拟合优度的要求,牺牲一定的拟合优度。

3.3.2 回归方程的显著性检验(F 检验)

方程的显著性检验,其核心目的是推断模型里被解释变量和解释变量之间的线性关系是否显著成立。

通过拟合优度知识可知,对于一个回归函数而言,拟合优度越高,表明解释变量对被解释变量的解释力度越强。基于此,可以初步推测模型的总体线性关系是成立的;反之,若拟合优度较低,那么模型总体线性关系成立的可能性就较小。

然而,在判断拟合优度时,并没有一个确切的标准值可供参考,这种判断更多的是基于经验的模糊推测,无法从统计学意义上给出严谨且确凿的结论。为弥补这一不足,需要进行方程的显著性检验。方程的显著性检验采用的依旧是数理统计学领域中的假设检验方法,以此来对模型中被解释变量与解释变量之间的线性关系在总体上是否显著成立进行科学判断。

1. 方程显著性的 F 检验

方程显著性的 F 检验是要检验模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i \quad (i=1,2,\cdots,n)$$

中参数是否显著不为零。按照假设检验的原理与程序,原假设与备择假设分别为

$$H_0: \beta_1 = 0, \beta_2 = 0, \cdots, \beta_k = 0$$

$$H_1: \beta_i (i=1,2,\cdots,k) \text{ 不全为零}$$

F 检验的思想来自总离差平方和的分解式

$$TSS = ESS + RSS \quad (3-27)$$

由于回归平方和 $ESS = \sum \hat{y}_i^2$ 是解释变量 X 的联合体对被解释变量 Y 的线性作用的结果,考虑比值

$$ESS/RSS = \sum \hat{y}_i^2 / \sum e_i^2 \quad (3-28)$$

如果这个比值较大,则 X 的联合体对 Y 的解释程度高,可认为总体存在线性关系,反之总体上可能不存在线性关系。因此,可通过该比值的大小对总体线性关系进行推断。

根据数理统计学知识,在原假设 H_0 成立的条件下,统计量

$$F = \frac{ESS/k}{RSS/(n-k-1)} \quad (3-29)$$

服从自由度为 $(k, n-k-1)$ 的 F 分布。

给定一个显著性水平 α ,可得到一个临界值 $F_\alpha(k, n-k-1)$,根据样本在求出 F 统计量的数值后,如果 $F > F_\alpha(k, n-k-1)$,则拒绝原假设 $H_0: \beta_1 = 0, \beta_2 = 0, \cdots, \beta_k = 0$,说明回归模型有显著意义,即所有解释变量联合起来对 Y 确有显著影响;如果 $F \leq F_\alpha(k, n-k-1)$,则不能拒绝原假设 $H_0: \beta_1 = 0, \beta_2 = 0, \cdots, \beta_k = 0$,说明回归模型没有显著意义,即所有解释变量联合起来对 Y 没有显著影响。

2. 拟合优度检验与方程显著性检验关系

目前,我们已经掌握了对模型拟合优度的检验和方程显著性检验相关知识。这两类检验基于不同原理,前者以估计出的模型为起点,检验模型对样本观测值的拟合程度;而后者从样本观测值出发,判断模型总体线性关系是否显著。

虽然它们的检验角度和方法有所不同,但二者存在内在联系。在通常情况下,如果模型对样本观测值的拟合程度较高,那么模型总体线性关系的显著性也会更强。在实际应用中,我们能够探寻出这两个检验统计量之间的数量关系,通过相互验证,进一步提升模型回归结果的可靠性和准确性,这对于提升模型回归的质量和效果具有重要意义。

根据式(3-29),分别用式(3-30)和式(3-31)表示的两个统计量之间存在下列关系:

$$R^2 = \frac{kF}{n-k-1+kF} \quad (3-30)$$

或

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \quad (3-31)$$

由式(3-31)可知, F 与 R^2 同向变化:当 $R^2=0$ 时, $F=0$; R^2 越大, F 值也越大;当 $R^2=1$ 时, F 为无穷大。因此, F 检验是所估计回归的方程总体显著性的度量,也是 R^2 的显著性检验。也就是说,检验原假设 $H_0: \beta_1=0, \beta_2=0, \dots, \beta_k=0$,等价于检验 $R^2=0$ 这一虚拟假设。 F 检验实际上也是判定系数的显著性检验。如果模型对样本有较高的拟合优度, F 检验一般都能通过。

在应用中不必苛求 R^2 ,重要的是需考察模型的经济关系是否合理。例如 $n=20, k=2, n-k-1=17$,取显著性水平 $\alpha=0.01$ 时,由 F 分布查得临界值 $F_{0.01}(2, 17)=6.11$;因此只要 F 统计量的值大于6.11,就能以99%的置信度认为模型的线性关系是显著的。由 F 统计量与 R^2 的关系可以计算出

$$R^2 = \frac{kF}{n-k-1+kF} = \frac{2 \times 6.11}{17 + 2 \times 6.11} = 0.418$$

也就是说,只要 $R^2 > 0.418$,模型线性关系显著的概率就达到99%。

3. 一般线性约束的检验

对于多元线性回归模型,可以设置更一般的线性约束。例如, $\beta_1 + \beta_2 = 1$ 。把这个约束条件代入式(3-1),可以得到

$$Y = \beta_0 + \beta_1 X_1 + (1 - \beta_1) X_2 + \dots + \beta_{k-1} X_k + \mu \quad (3-32)$$

$$Y^* = \beta_0 + \beta_1 X_1^* + \beta_3 X_3 + \dots + \beta_k X_k + \mu^* \quad (3-33)$$

其中, $Y^* = Y - X_2, X_1^* = X_1 - X_2$ 。

对模型(3-33)进行估计,可以得到受约束样本回归模型的剩余平方和 RSS_R ,对式(3-1)所作的回归可以得到无约束回归模型的剩余平方和 RSS_U 。可以证明,当式(3-1)参数受到约束时,模型所得的剩余平方和变大,因此 $RSS_R \geq RSS_U$ 。

但是,如果约束条件为真,则受约束回归模型与无约束回归模型具有相同的解释能力,从而使得 RSS_U 与 RSS_R 的差异变小。于是,可用 $RSS_R - RSS_U$ 的大小来检验约束条件的真实性。

根据数理统计学的知识, $RSS/\sigma^2 \sim \chi^2(n-k-1)$,其中 k 为回归模型中解释变量的个

数, σ^2 为回归模型随机误差项的方差。于是有

$$\text{RSS}_U/\sigma^2 \sim \chi^2(n - k_U - 1) \quad (3-34)$$

$$\text{RSS}_R/\sigma^2 \sim \chi^2(n - k_R - 1) \quad (3-35)$$

$$(\text{RSS}_R - \text{RSS}_U)/\sigma^2 \sim \chi^2(k_U - k_R) \quad (3-36)$$

其中, k_U 、 k_R 分别为无约束与受约束回归模型的解释变量的个数。

由式(3-34)与式(3-36), 可进一步得到 F 统计量

$$F = \frac{(\text{RSS}_R - \text{RSS}_U)/(k_U - k_R)}{\text{RSS}_U/(n - k_U - 1)} \sim F(k_U - k_R, n - k_U - 1) \quad (3-37)$$

根据式(3-37), 如果约束条件无效, 则 RSS_R 与 RSS_U 的差异较大, 计算的 F 值也较大。于是, 可用计算的 F 统计量的值与所给定的显著性水平下的临界值作比较, 来对约束条件的真实性进行检验。需注意的是, $k_U - k_R$ 恰为约束条件的个数。

例 3-1 柯布和道格拉斯于 1928 年使用美国 1899—1922 年制造业产出(Y)、资本(K)与劳动力(L)的数据, 估计生产函数为 $Y = AK^{\beta_1}L^{\beta_2}e^{\mu}$, 两边取对数得到

$$\ln(Y) = \ln(A) + \beta_1 \ln(K) + \beta_2 \ln(L) + \mu.$$

利用柯布和道格拉斯提供的原始数据进行回归, 可以得到回归结果式(3-38)。

$$\begin{aligned} \widehat{\ln(Y)} &= -0.1773 + 0.2331\ln(K) + 0.8073\ln(L) \\ \text{se} &= (0.4343) \quad (0.0635) \quad (0.1451) \\ \text{RSS} &= 0.0710 \end{aligned} \quad (3-38)$$

如果限制条件 $\beta_1 + \beta_2 = 1$ 成立, 则可以得到估计结果式(3-39)。

$$\begin{aligned} \widehat{\ln(Y/L)} &= 0.0145 + 0.2541\ln(K/L) \\ \text{se} &= (0.0120) \quad (0.1412) \\ \text{RSS} &= 0.0716 \end{aligned} \quad (3-39)$$

式(3-38)和式(3-39)括号中的数字表示各参数估计的标准差。

根据式(3-38)与式(3-39)估计结果可以得到 $F = \frac{(0.0716 - 0.0710)/1}{0.0710/20} = 0.1690$ 。

根据上面计算得到的 F 值可以看出, 约束条件成立。

3.3.3 参数的显著性检验(t 检验)

对于多元线性回归模型来说, 通过回归方程的显著性检验, 可以检测出方程的总体线性关系是否显著, 但实际上, 方程的总体线性关系是显著的并不能代表每个解释变量对被解释变量的影响都是显著的, 因此必须对每个解释变量进行显著性检验。通过检验, 如果发现某个变量对被解释变量的影响并不显著, 那么就应该将它剔除, 以建立更为简单和精确的模型。在对变量进行显著性检验中应用最为普遍的是 t 检验, 在目前常用的计量经济学软件包中, 也都包含关于 t 统计量的计算结果。

1. t 统计量

在 3.2 节中, 已经导出了参数估计量的方差为

$$\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$