第三讲 多元线性回归模型

一、多元线性回归模型
二、多元线性回归模型的基本假定
三、多元线性回归模型的估计
四、多元线性回归模型的检验
五、多元线性回归模型的预测
六、案例分析

一、多元线性回归模型

多元线性回归模型:表现在线性回归模型中的解释变量有多个。

一般表现形式:

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

其中:k为解释变量的数目，β_j称为**回归参数**（regression coefficient）。

习惯上：把**常数项**看成为一**虚变量**的系数，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）

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i$$

也被称为总体回归函数的随机表达形式。它 的非随机表达式为:

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

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，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给出了X_j的单位变化对Y均值的“直接”或“净”（不含其他变量）影响。

总体回归模型n个随机方程的矩阵表达式为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$$

其中

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n\times(k+1)}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}_{n\times 1}$$

样本回归函数：用来估计总体回归函数

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki}$$

其随机表示式：

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki} + e_i$$

e_i称为残差或剩余项(residuals)，可看成是总体回归函数中随机扰动项μ_i的近似替代。

样本回归函数的矩阵表达：

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{或} \qquad \mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

其中：

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

二、多元线性回归模型的基本假定

假设1，解释变量是非随机的或固定的，且各X之间互不相关（无多重共线性）。

假设2，随机误差项具有零均值、同方差及不序列相关性

$$E(\mu_i) = 0$$

$$Var(\mu_i) = E(\mu_i^2) = \sigma^2 \qquad i \neq j \quad i, j = 1,2,\cdots,n$$

$$Cov(\mu_i, \mu_j) = E(\mu_i \mu_j) = 0$$

假设3，解释变量与随机项不相关

$$Cov(X_{ji}, \mu_i) = 0 \qquad j = 1,2\cdots,k$$

假设4，随机项满足正态分布

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

上述假设的矩阵符号表示 式：

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

假设2，

$$E(\boldsymbol{\mu})=E\begin{pmatrix}\mu_1\\ \vdots\\ \mu_n\end{pmatrix}=\begin{pmatrix}E(\mu_1)\\ \vdots\\ E(\mu_n)\end{pmatrix}=0$$

$$E(\boldsymbol{\mu}\boldsymbol{\mu}')=E\left(\begin{pmatrix}\mu_1\\ \vdots\\ \mu_n\end{pmatrix}\begin{pmatrix}\mu_1 & \cdots & \mu_n\end{pmatrix}\right)=E\begin{pmatrix}\mu_1^2 & \cdots & \mu_1\mu_n\\ \vdots & \ddots & \vdots\\ \mu_n\mu_1 & \cdots & \mu_n^2\end{pmatrix}$$

$$=\begin{pmatrix}\mathrm{var}(\mu_1) & \cdots & \mathrm{cov}(\mu_1,\mu_n)\\ \vdots & \ddots & \vdots\\ \mathrm{cov}(\mu_n,\mu_1) & \cdots & \mathrm{var}(\mu_n)\end{pmatrix}=\begin{pmatrix}\sigma^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma^2\end{pmatrix}=\sigma^2\mathbf{I}$$

假设3，$E(\mathbf{X}'\boldsymbol{\mu})=0$，即

$$E\begin{pmatrix}\sum\mu_i\\ \sum X_{1i}\mu_i\\ \vdots\\ \sum X_{Ki}\mu_i\end{pmatrix}=\begin{pmatrix}\sum E(\mu_i)\\ \sum X_{1i}E(\mu_i)\\ \vdots\\ \sum X_{Ki}E(\mu_i)\end{pmatrix}=0$$

假设4，向量$\boldsymbol{\mu}$有一多维正态分布，即

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

同一元回归一样，多元回归还具有如下两个重要假设：

*假设5，样本容量趋于无穷时，各解释变量的方差趋于有界常数，即$n \to \infty$时，

$$\frac{1}{n}\sum x_{ji}^2 = \frac{1}{n}\sum (X_{ji} - \bar{X}_j)^2 \to Q_j \qquad \text{或} \qquad \frac{1}{n}\mathbf{x}'\mathbf{x} \to \mathbf{Q}$$

其中：**Q**为一非奇异固定矩阵，矩阵**x**是由各解释变量的离差为元素组成的$n \times k$阶矩阵

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & \cdots & \vdots \\ x_{1n} & \cdots & x_{kn} \end{pmatrix}$$

假设6，回归模型的设定是正确的。

二、多元线性回归模型的估计

估计目标：结构参数$\hat{\beta}_j$及随机误差项的方差$\hat{\sigma}^2$

估计方法：OLS、ML或者MM

1、普通最小二乘估计

*2、最大或然估计

*3、矩估计

4、参数估计量的性质

5、样本容量问题

1、普通最小二乘估计

对于随机抽取的n组观测值 $(Y_i, X_{ji}), i = 1,2,\cdots,n,\ j = 0,1,2,\cdots k$

如果**样本函数**的参数估计值已经得到，则有：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{Ki} \qquad i=1,2\ldots n$$

根据**最小二乘原理**，参数估计值应该是下列方程组的解

$$\begin{cases} \dfrac{\partial}{\partial \hat{\beta}_0} Q = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_1} Q = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_2} Q = 0 \\ \vdots \\ \dfrac{\partial}{\partial \hat{\beta}_k} Q = 0 \end{cases}$$

其中 $Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$$= \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}))^2$$

于是得到关于待估参数估计值的**正规方程组**：

$$\begin{cases} \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki})=\Sigma Y_i \\ \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki})X_{1i}=\Sigma Y_iX_{1i} \\ \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_{2i}X_{2i}+\cdots+\hat{\beta}_kX_{ki})X_{2i}=\Sigma Y_iX_{2i} \\ \vdots \\ \Sigma(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki})X_{ki}=\Sigma Y_iX_{ki} \end{cases}$$

解该（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 $\hat{\beta}_j, j=0,1,2,\cdots,k$ 。

正规方程组的矩阵形式

$$\begin{pmatrix} n & \sum X_{1i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \cdots & \sum X_{1i}X_{ki} \\ \cdots & \cdots & \cdots & \cdots \\ \sum X_{ki} & \sum X_{ki}X_{1i} & \cdots & \sum X_{ki}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}$$

即

$$\mathbf{(X'X)\hat{\beta} = X'Y}$$

由于**X'X**满秩，故有

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

将上述过程用矩阵表示如下：

寻找一组参数估计值 $\hat{\boldsymbol{\beta}}$，使得残差平方和

$$Q=\sum_{i=1}^{n}e_i^2=\mathbf{e}'\mathbf{e}=(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

最小。

即求解方程组：$\dfrac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})=0$

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{Y}'\mathbf{Y}-\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}-\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})=0$$

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{Y}'\mathbf{Y}-2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})=0$$

$$-\mathbf{X}'\mathbf{Y}+\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=0$$

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

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

***例3.2.1：** 在例2.1.1的家庭收入-消费支出例中，

$$(\mathbf{X}'\mathbf{X})=\begin{pmatrix}1 & 1 & \cdots & 1\\ X_1 & X_2 & \cdots & X_n\end{pmatrix}\begin{pmatrix}1 & X_1\\ 1 & X_2\\ \cdots & \cdots\\ 1 & X_n\end{pmatrix}=\begin{pmatrix}n & \sum X_i\\ \sum X_i & \sum X_i^2\end{pmatrix}=\begin{pmatrix}10 & 21500\\ 21500 & 53650000\end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y}=\begin{pmatrix}1 & 1 & \cdots & 1\\ X_1 & X_2 & \cdots & X_n\end{pmatrix}\begin{pmatrix}Y_1\\ Y_2\\ \cdots\\ Y_n\end{pmatrix}=\begin{pmatrix}\sum Y_i\\ \sum X_iY_i\end{pmatrix}=\begin{pmatrix}15674\\ 39468400\end{pmatrix}$$

可求得

$$(\mathbf{X}'\mathbf{X})^{-1}=\begin{pmatrix}0.7226 & -0.0003\\ -0.0003 & 1.35E-07\end{pmatrix}$$

于是

$$\hat{\boldsymbol{\beta}}=\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\begin{pmatrix}0.7226 & -0.0003\\ -0.0003 & 1.35E-07\end{pmatrix}\begin{pmatrix}15674\\ 39648400\end{pmatrix}=\begin{pmatrix}-103.172\\ 0.7770\end{pmatrix}$$

正规方程组 的另一种写法

对于正规方程组

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

将 $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$ 代入得

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{e} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

于是

$$\mathbf{X}'\mathbf{e} = \mathbf{0} \qquad (*)$$

或

$$\begin{cases} \sum e_i = 0 \\ \sum_i X_{ji} e_i = 0 \end{cases} \qquad (**)$$

(*)或（**）是多元线性回归模型正规方程组的另一种写法

样本回归函数的离差形式

$$y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} + e_i \qquad \text{i=1,2…n}$$

其矩阵形式为

$$\mathbf{y} = \mathbf{x}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

其中：

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{x} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

在离差形式下，参数的最小二乘估计结果为

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \cdots - \hat{\beta}_k \overline{X}_k$$

随机误差项μ的方差σ的无偏估计

可以证明，随机误差项μ的方差的无偏估计量为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

*2、最大或然估计

对于多元线性回归模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i$$

易知 $$Y_i \sim N(\mathbf{X_i}\boldsymbol{\beta}, \sigma^2)$$

Y的随机抽取的n组样本观测值的联合概率

$$L(\hat{\boldsymbol{\beta}}, \sigma^2) = P(Y_1, Y_2, \cdots, Y_n)$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}\Sigma(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}))^2}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}$$

即为变量Y的**或然函数**

对数或然函数为

$$L^* = Ln(L)$$

$$= -nLn(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\,(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

对对数或然函数求极大值，也就是对

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

求极小值。

因此，参数的**最大或然估计**为

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

结果与参数的普通最小二乘估计相同

*3、矩估计（Moment Method, MM）

OLS估计是通过得到一个关于参数估计值的正规方程组

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

并对它进行求解而完成的。

该正规方程组 可以从另外一种思路来导：

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mu}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\boldsymbol{\mu}$$

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{X}'\boldsymbol{\mu}$$

求期望：

$$E(\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

$$E(\mathbf{X}'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})=\mathbf{0}$$

称为原总体回归方程的一组矩条件，表明了原总体回归方程所具有的内在特征。

如果随机抽出原总体的一个样本，估计出的样本回归方程

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

能够近似代表总体回归方程的话，则应成立：

$$\frac{1}{n}\mathbf{X}'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})=0$$

由此得到正规方程组

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

解此正规方程组即得参数的MM估计量。

易知MM估计量与OLS、ML估计量等价。

矩方法是工具变量方法(Instrumental Variables,IV)和广义矩估计方法(Generalized Moment Method, GMM)的基础

- 在矩方法中关键是利用了

$$E(X'\mu)=0$$

- 如果某个解释变量与随机项相关，只要能找到1个工具变量，仍然可以构成一组矩条件。这就是IV。

- 如果存在＞k+1个变量与随机项不相关，可以构成一组包含＞k+1方程的矩条件。这就是GMM。

4、参数估计量的性质

在满足基本假设的情况下，其结构参数β的普通最小二乘估计、**最大或然估计**及**矩估计**仍具有：

线性性、无偏性、有效性。

同时，随着样本容量增加，参数估计量具有：

渐近无偏性、渐近有效性、一致性。

（1）、线性性

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

其中, $\mathbf{C}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 为一仅与固定的$\mathbf{X}$有关的行向量

（2）、无偏性

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

这里利用了假设：$E(X'\mu)=0$

（3）、有效性（最小方差性）

参数估计量 $\hat{\beta}$ 的方差-协方差矩阵

$$
\begin{aligned}
Cov(\hat{\beta}) &= E(\hat{\beta}-E(\hat{\beta}))(\hat{\beta}-E(\hat{\beta}))' \\
&= E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'
\end{aligned}
$$

$$= E((XX')^{-1}X'\,\mu\,\mu'X(X'X)^{-1})$$
$$= (X'X)^{-1}X'E(\mu\,\mu')X(X'X)^{-1}$$
$$= E(\mu\,\mu')(X'X)^{-1}$$
$$= \sigma^2\mathbf{I}(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}$$

其中利用了

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

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

根据高斯—马尔可夫定理，$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$ 在所有无偏估计量的方差中是最小的。

5、样本容量问题

（1） 最小样本容量

所谓“**最小样本容量**”，即从最小二乘原理和最大或然原理出发，欲得到参数估计量，不管其质量如何，所要求的样本容量的下限。

样本最小容量必须不少于模型中解释变量的数目（包括常数项），即

$$n \geq k+1$$

因为，无多重共线性要求：秩$(\mathbf{X})=k+1$

（2）满足基本要求的样本容量

从统计检验的角度：

n>30 时，Z检验才能应用；

n-k≥8时，t分布较为稳定

一般经验认为：

当n≥30或者至少n≥3(k+1)时，才能说满足模型估计的基本要求。

模型的良好性质只有在大样本下才能得到理论上的证明

三、多元线性回归模型的统计检验

1、拟合优度检验

2、方程的显著性检验(F检验)

3、变量的显著性检验（t检验）

4、参数的置信区间

1、拟合优度检验

（1）可决系数与调整的可决系数

总离差平方和的分解

记 $TSS=\sum(Y_i-\bar{Y})^2$ 总离差平方和

$ESS=\sum(\hat{Y}_i-\bar{Y})^2$ 回归平方和

$RSS=\sum(Y_i-\hat{Y}_i)^2$ 剩余平方和

则

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

由于 $\sum (Y_i - \hat{Y})(\hat{Y}_i - \bar{Y}) = \sum e_i (\hat{Y}_i - \bar{Y})$

$$= \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum e_i X_{1i} + \cdots + \hat{\beta}_k \sum e_i X_{ki} + \bar{Y} \sum e_i$$

$$=0$$

所以有：

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

注意：一个有趣的现象

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

$$(Y_i - \bar{Y})^2 \neq (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2$$

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

可决系数

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

该统计量越接近于1，模型的拟合优度越高。

问题：

在应用过程中发现，如果在模型中增加一个解释变量，R^2往往增大（Why?）

这就给人一个错觉：**要使得模型拟合得好，只要增加解释变量即可**。

但是，现实情况往往是，由增加解释变量个数引起的R^2的增大与拟合好坏无关，**R^2需调整**。

调整的可决系数（adjusted coefficient of determination）

在样本容量一定的情况下，增加解释变量必定使得自由度减少，所以调整的思路是:**将残差平方和与总离差平方和分别除以各自的自由度，以剔除变量个数对拟合优度的影响:**

$$\overline{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

其中：n-k-1为残差平方和的自由度，n-1为总体平方和的自由度。

$\overline{R}^2$与R^2之间存在如下关系：

$$\overline{R}^2 = 1-(1-R^2)\frac{n-1}{n-k-1}$$

在中国居民消费支出的二元模型例中，$\overline{R}^2$=0.9954

在中国居民消费支出的一元模型例中，$\overline{R}^2$=0.9927

问题：$\overline{R}^2$多大才算通过拟合优度检验？

*（2）赤池信息准则和施瓦茨准则

为了比较所含解释变量个数不同的多元回归模型的拟合优度，常用的标准还有：

赤池信息准则（Akaike information criterion, AIC）

$$AIC = \ln \frac{\mathbf{e}'\mathbf{e}}{n} + \frac{2(k+1)}{n}$$

施瓦茨准则（Schwarz criterion，SC）

$$AC = \ln \frac{\mathbf{e}'\mathbf{e}}{n} + \frac{k}{n} \ln n$$

这两准则均要求仅当所增加的解释变量能够减少AIC值或AC值时才在原模型中增加该解释变量。

Eviews的估计结果显示：

中国居民消费一元例中：

AIC=6.68　　AC=6.83

中国居民消费二元例中：

AIC=7.09　　AC=7.19

从这点看，可以说前期人均居民消费CONSP(-1)应包括在模型中。

2、方程的显著性检验(F检验)

方程的显著性检验，旨在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作出推断。

（1）方程显著性的F检验

即检验模型

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\ \dots\ +\beta_kX_{ki}+\mu_i \qquad i=1,2,\ \dots,n$$

中的参数β_j是否显著不为0。

可提出如下原假设与备择假设：

H_0：　$\beta_0=\beta_1=\beta_2=\ \dots\ =\beta_k=0$

H_1：　β_j不全为0

F检验的思想来自于总离差平方和的分解式：

TSS=ESS+RSS

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

$$ESS / RSS = \sum \hat{y}_i^2 \Big/ \sum e_i^2$$

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。

根据数理统计学中的知识，在原假设H_0成立的条件下，统计量

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

服从自由度为$(k, n-k-1)$的F分布

给定显著性水平α，可得到临界值$F_\alpha(k,n-k-1)$，由样本求出统计量**F**的数值，通过

$$F > F_\alpha(k,n-k-1) \quad \text{或} \quad F \le F_\alpha(k,n-k-1)$$

来拒绝或接受原假设H_0，以判定原方程总体上的线性关系是否显著成立。

对于中国居民人均消费支出的例子：

一元模型：F=285.92

二元模型：F=2057.3

给定显著性水平α =0.05，查分布表，得到临界值：

一元例：$F_\alpha(1,21)=4.32$

二元例：$F_\alpha(2,19)=3.52$

显然有 $F > F_\alpha(k,n-k-1)$

即二个模型的线性关系在95%的水平下显著成立。

（2）关于拟合优度检验与方程显著性检验关系的讨论

由 $\bar{R}^2 = 1 - \dfrac{RSS/(n-k-1)}{TSS/(n-1)}$ 与 $F = \dfrac{ESS/k}{RSS/(n-k-1)}$

可推出： $\bar{R}^2 = 1 - \dfrac{n-1}{n-k-1+kF}$

或 $F = \dfrac{\bar{R}^2/k}{(1-\bar{R}^2)/(n-k-1)}$

F 与 $\bar{R}^2$ 同向变化：当 $\bar{R}^2 = 0$ 时，F = 0；

$\bar{R}^2$ 越大，F 值也越大；

当 $\bar{R}^2 = 1$ 时，F 为无穷大。

因此，F检验是所估计回归的总显著性的一个度量，也是$\bar{R}^2$的一个显著性检验。亦即

检验 $H_0: \beta_1 = 0, \beta_2 = 0, \cdots, \beta_k = 0$ 等价于检验 $\bar{R}^2 = 0$

回答前面的问题：$\bar{R}^2$多大才算通过拟合优度检验

- 在中国居民人均收入-消费一元模型中，

F>4.32 → $\bar{R}^2$>0.131 → 模型在95%的水平下显著成立

- 在中国居民人均收入-消费二元模型中，

F>3.52 → $\bar{R}^2$>0.194 → 模型在95%的水平下显著成立

3、变量的显著性检验（t检验）

方程的**总体线性**关系显著≠**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，以决定是否作为解释变量被保留在模型中。

这一检验是由对变量的 t 检验完成的。

（1）t统计量

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

以c_{ii}表示矩阵$(X'X)^{-1}$ 主对角线上的第i个元素，于是参数估计量的方差为：

$$Var(\hat{\beta}_i) = \sigma^2 c_{ii}$$

其中σ^2为随机误差项的方差，在实际计算时，用它的估计量代替：

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

易知 $\hat{\beta}$ 服从如下正态分布

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$$

因此，可构造如下t统计量

$$t = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} \frac{\hat{\beta}_i - \beta_i}{\sqrt{c_{ii} \dfrac{\mathbf{e}'\mathbf{e}}{n-k-1}}} \sim t(n-k-1)$$

（2）t检验

设计原假设与备择假设：

H_0：$\beta_i=0$　　（i=1,2…k）

H_1：$\beta_i\neq0$

给定显著性水平α，可得到临界值$t_{\alpha/2}(n-k-1)$，由样本求出统计量**t**的数值，通过

$|t|>t_{\alpha/2}(n-k-1)$　　或　　$|t|\leq t_{\alpha/2}(n-k-1)$

来拒绝或接受原假设H_0，从而**判定对应的解释变量是否应包括在模型中。**

注意：一元线性回归中，t检验与F检验一致

一方面，t检验与F检验都是对相同的原假设 H_0：$\beta_1=0$ 进行检验；

另一方面，两个统计量之间有如下关系：

$$F=\frac{\sum \hat{y}_i^2}{\sum e_i^2/(n-2)}=\frac{\hat{\beta}_1^2\sum x_i^2}{\sum e_i^2/(n-2)}=\frac{\hat{\beta}_1^2}{\sum e_i^2/(n-2)\sum x_i^2}$$

$$=\left(\frac{\hat{\beta}_1}{\sqrt{\sum e_i^2/(n-2)\sum x_i^2}}\right)^2=\left(\hat{\beta}_1\Big/\sqrt{\frac{\sum e_i^2}{n-2}\cdot\frac{1}{\sum x_i^2}}\right)^2=t^2$$

在中国居民人均收入-消费支出二元模型例中，由应用软件计算出参数的t值：

$$|t_0| = 3.306 \qquad |t_1| = 3.630 \qquad |t_2| = 2.651$$

给定显著性水平α=0.05，查得相应临界值：$t_{0.025}(19)$ =2.093。

可见，计算的所有t值都大于该临界值，所以拒绝原假设。即:

包括常数项在内的3个解释变量都在95%的水平下显著，都通过了变量显著性检验。

4、参数的置信区间

参数的置信区间用来考察：在一次抽样中所估计的参数值离参数的真实值有多“近”。

在变量的显著性检验中已经知道：

$$t=\frac{\hat{\beta}_i-\beta_i}{S_{\hat{\beta}_i}}\frac{\hat{\beta}_i-\beta_i}{\sqrt{c_{ii}\dfrac{\mathbf{e}'\mathbf{e}}{n-k-1}}}\sim t(n-k-1)$$

容易推出：在$(1-\alpha)$的置信水平下β_i的置信区间是

$$(\hat{\beta}_i-t_{\frac{\alpha}{2}}\times s_{\hat{\beta}_i},\hat{\beta}_i+t_{\frac{\alpha}{2}}\times s_{\hat{\beta}_i})$$

其中，$t_{\alpha/2}$为显著性水平为α 、自由度为n-k-1的临界值。

在中国居民人均收入-消费支出二元模型例中，

给定α=0.05，查表得临界值：$t_{0.025}(19)$=2.093

从回归计算中已得到：

$$\hat{\beta}_0 = 120.70 \qquad s_{\hat{\beta}_0} = 36.51$$

$$\hat{\beta}_1 = 0.2213 \qquad s_{\hat{\beta}_1} = 0.061$$

$$\hat{\beta}_2 = 0.4515 \qquad s_{\hat{\beta}_2} = 0.170$$

计算得参数的置信区间：

β_0：(44.284, 197.116)

β_1：(0.0937, 0.3489)

β_2：(0.0951, 0.8080)

如何才能缩小置信区间？

- 增大样本容量n，因为在同样的样本容量下，n越大，t分布表中的临界值越小，同时，增大样本容量，还可使样本参数估计量的标准差减小；

- 提高模型的拟合优度，因为样本参数估计量的标准差与残差平方和呈正比，模型优度越高，残差平方和应越小。

- 提高样本观测值的分散度，一般情况下，样本观测值越分散，$(X'X)^{-1}$的分母的$|X'X|$的值越大，致使区间缩小。

五、多元线性回归模型的预测

1、$E(Y_0)$的置信区间

2、Y_0的置信区间

对于模型

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

给定样本以外的解释变量的观测值$\mathbf{X}_0=(1,X_{10},X_{20},\ldots,X_{k0})$，可以得到被解释变量的预测值：

$$\hat{Y}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}}$$

它可以是总体均值$E(Y_0)$或个值Y_0的预测。

但严格地说，这只是被解释变量的预测值的估计值，而不是预测值。

为了进行科学预测，还需求出预测值的置信区间，包括$E(Y_0)$和Y_0的置信区间。

1、$E(Y_0)$的置信区间

易知

$$E(\hat{Y}_0) = E(\mathbf{X_0}\hat{\boldsymbol{\beta}}) = \mathbf{X_0}E(\hat{\boldsymbol{\beta}}) = \mathbf{X_0}\boldsymbol{\beta} = E(Y_0)$$

$$Var(\hat{Y}_0) = E(\mathbf{X_0}\hat{\boldsymbol{\beta}} - \mathbf{X_0}\boldsymbol{\beta})^2 = E(\mathbf{X_0}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathbf{X_0}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))$$

由于$\mathbf{X_0}(\hat{\beta} - \beta)$为标量，因此

$$\begin{aligned} Var(\hat{Y}_0) &= E(\mathbf{X_0}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}_0') \\ &= \mathbf{X_0}E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}_0' \\ &= \sigma^2\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0' \end{aligned}$$

容易证明

$$\hat{Y}_0 \sim N(\mathbf{X_0\beta}, \sigma^2 \mathbf{X_0(X'X)^{-1}X'_0})$$

取随机扰动项的样本估计量$\hat{\sigma}^2$，构造如下t统计量

$$\frac{\hat{Y}_0 - E(Y_0)}{\hat{\sigma}\sqrt{\mathbf{X_0(X'X)^{-1}X'_0}}} \sim t(n-k-1)$$

于是，得到$(1-\alpha)$的置信水平下$E(Y_0)$的**置信区间**：

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\mathbf{X}_0(\mathbf{X'X})^{-1}\mathbf{X'}_0} < E(Y_0) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\mathbf{X}_0(\mathbf{X'X})^{-1}\mathbf{X'}_0}$$

其中，$t_{\alpha/2}$为$(1-\alpha)$的置信水平下的**临界值**。

2、Y_0的置信区间

如果已经知道实际的预测值Y_0，那么预测误差为：

$$e_0 = Y_0 - \hat{Y}_0$$

容易证明

$$\begin{aligned} E(e_0) &= E(\mathbf{X}_0\boldsymbol{\beta} + \mu_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}) \\ &= E(\mu_0 - \mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &= E(\mu_0 - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu}) \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(e_0) &= E(e_0^2) \\ &= E(\mu_0 - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\mu})^2 \\ &= \sigma^2(1 + \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0') \end{aligned}$$

e_0服从正态分布，即

$$e_0 \sim N(0, \sigma^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'))$$

取随机扰动项的样本估计量$\hat{\sigma}^2$，可得e_0的方差的估计量

$$\hat{\sigma}_{e_0}^2 = \hat{\sigma}^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0')$$

构造t统计量

$$t = \frac{\hat{Y}_0 - Y_0}{\hat{\sigma}_{e_0}} \sim t(n-k-1)$$

可得给定$(1-\alpha)$的置信水平下Y_0的**置信区间**：

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'} < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$$

中国居民人均收入-消费支出二元模型例中：

2001年人均GDP：4033.1元，

于是人均居民消费的预测值为

$\hat{Y}_{2001}=120.7+0.2213\times 4033.1+0.4515\times 1690.8=1776.8$（元）

实测值（90年价）=1782.2元，**相对误差：**-0.31%

预测的置信区间：

在95%的置信度下，临界值$t_{\alpha/2}(19)=2.093$，$\hat{\sigma}^2=705.5$，

$$(\mathbf{X}'\mathbf{X})^{-1}=\begin{pmatrix} 1.88952 & 0.00285 & -0.00828 \\ 0.00285 & 0.00001 & -0.00001 \\ -0.00828 & -0.00001 & 0.00004 \end{pmatrix}$$

$$\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 = 0.3938$$

于是$E(\hat{Y}_{2001})$的95%的置信区间为:

$$1776.8 \pm 2.093 \times \sqrt{705.5} \times \sqrt{0.3938}$$

或　　（1741.8，1811.7）

同样，易得$\hat{Y}_{2001}$的95%的置信区间为

$$1776.8 \pm 2.093 \times \sqrt{705.5} \times \sqrt{1.3938}$$

或　　（1711.1, 1842.4）

六、 案例分析

研究的目的要求

为了研究影响中国税收收入增长的主要原因，分析中央和地方税收收入增长的数量规律，预测中国税收未来的增长趋势，需要建立计量经济模型。

研究范围： 1978年-2007年全国税收收入

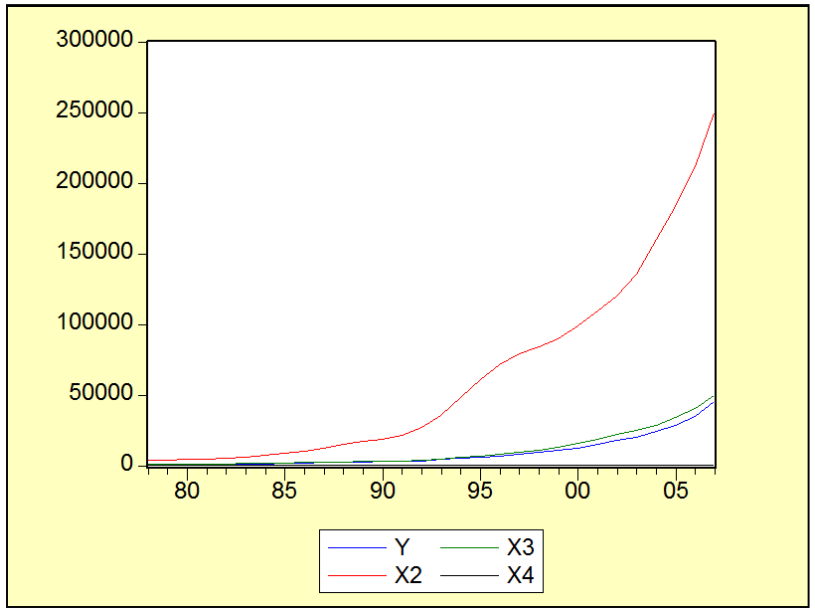
理论分析： 为了全面反映中国税收增长的全貌，选择包括中央和地方税收的“国家财政收入”中的“各项税收”（简称“税收收入”）作为被解释变量；选择国内生产总值（GDP）作为经济整体增长水平的代表；选择中央和地方“财政支出”作为公共财政需求的代表；选择“商品零售价格指数”作为物价水平的代表。

年份	税收收入（亿元）（Y）	国内生产总值（亿元）（X_2）	财政支出（亿元）（X_3）	商品零售价格指数（%）（X_4）
1978	519.28	3624.1	1122.09	100.7
1979	537.82	4038.2	1281.79	102.0
1980	571.70	4517.8	1228.83	106.0
1981	629.89	4862.4	1138.41	102.4
1982	700.02	5294.7	1229.98	101.9
1983	775.59	5934.5	1409.52	101.5
1984	947.35	7171.0	1701.02	102.8
1985	2040.79	8964.4	2004.25	108.8
1986	2090.73	10202.2	2204.91	106.0
1987	2140.36	11962.5	2262.18	107.3
1988	2390.47	14928.3	2491.21	118.5
1989	2727.40	16909.2	2823.78	117.8
1990	2821.86	18547.9	3083.59	102.1
1991	2990.17	21617.8	3386.62	102.9

1993	4255.30	34634.4	4642.30	113.2
1994	5126.88	46759.4	5792.62	121.7
1995	6038.04	58478.1	6823.72	114.8
1996	6909.82	67884.6	7937.55	106.1
1997	8234.04	74462.6	9233.56	100.8
1998	9262.80	78345.2	10798.18	97.4
1999	10682.58	82067.5	13187.67	97.0
2000	12581.51	89468.1	15886.50	98.5
2001	15301.38	97314.8	18902.58	99.2
2002	17636.45	104790.6	22053.15	98.7
2003	20017.31	135822.8	24649.95	99.9
2004	24165.68	159878.3	28486.89	102.8
2005	28778.54	183217.4	33930.28	100.8
2006	34804.35	211923.5	40422.73	101
2007	45621.97	249529.9	49781.35	103.8

序列Y、X2、X3、X4的线性图

可以看出Y、X2、X3都是逐年增长的，但增长速率有所变动，而且X4在多数年份呈现出水平波动。说明变量间不一定是线性关系，可探索将模型设定为以下对数模型：



$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_2 \ln X_{3t} + \beta_3 X_{4t} + u_t$$

注意这里的“商品零售价格指数”（X4）未取对数。

三、估计参数

Source	SS	df	MS
Model	52.7568562	3	17.5856187
Residual	.66290406	26	.02549631
Total	53.4197602	29	1.8420607

Number of obs	= 30
F(3, 26)	= 689.73
Prob > F	= 0.0000
R-squared	= 0.9876
Adj R-squared	= 0.9862
Root MSE	= .15968

lny	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnx2	.4512337	.1421285	3.17	0.004	.1590845	.7433829
lnx3	.6271328	.1615663	3.88	0.001	.2950285	.9592372
x4	.0101359	.0056449	1.80	0.084	-.0014675	.0217392
_cons	-2.755368	.6400803	-4.30	0.000	-4.071072	-1.439664

模型估计的结果为：

$$\ln \hat{Y}_i = -2.7554 + 0.4512 \ln X_2 + 0.6271 \ln X_3 + 0.0101 X_4$$

$$\begin{array}{lcccc} & (0.6401) & (0.1421) & (0.1616) & (0.0056) \\ t= & (-4.30) & (3.17) & (3.88) & (1.80) \end{array}$$

$$R^2 = 0.9876 \quad \bar{R}^2 = 0.9862 \quad F=679.73 \quad n=30$$

模型检验：

1、经济意义检验：

模型估计结果说明，在假定其它变量不变的情况下，当年GDP每增长1%，税收收入会增长0.4512%；当年财政支出每增长1%，平均说来税收收入会增长0.6271%；当年商品零售价格指数上涨一个百分点，平均说来税收收入会增长1.01%。这与理论分析和经验判断相一致。

2、统计检验：

拟合优度：$R^2 = 0.9876$ ，$\bar{R}^2 = 0.9862$ 表明样本回归方程较好地拟合了样本观测值。

F检验：对 $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ 已得到 F =689.73，给定 $\alpha = 0.05$ 查表得自由度k-1=3和n-k=26的临界值：$F_\alpha(3,26) = 2.98$ ，因为F=689.73> $F_\alpha(3,26) = 2.98$ 说明模型总体上显著，即“国内生产总值”、“财政支出”、“商品零售价格指数”等变量联合起来确实对“税收收入”有显著影响。

64

t 检验

分别针对 $H_0: \beta_j = 0 \ (j=1,2,3,4)$ 给定显著性水平 $\alpha=0.1$，

查t分布表得自由度为n-k=26的临界值 $t_{\alpha/2}(n-k)=$

由回归结果已知与 $\hat{\beta}_1$、$\hat{\beta}_2$、$\hat{\beta}_3$、$\hat{\beta}_4$ 对应的t值分别为：

-4.30、3.17、3.88、1.80，其绝对值均大于

$t_{\alpha/2}(n-k)=2.056$ ，这说明在显著性水平 $\alpha=0.1$ 下，分别都应当拒绝 $H_0: \beta_j = 0 \ (j=1,2,3,4)$

说明当在其它解释变量不变的情况下，解释变量“国内生产总值”、“财政支出”、“商品零售价格指数”分别对被解释变量“税收收入”Y都有显著的影响。

案例分析STATA命令语句

```
Gen lny=log(y)
Gen lnx1=log(x1)
Gen lnx2=log(x2)
reg lny lnx1 lnx2 x3
```

本章小结

1. 多元线性回归模型及其矩阵形式。
2. 多元线性回归模型中对随机扰动项u的假定，除了其他基本假定以外，还要求满足无多重共线性假定。
3. 多元线性回归模型参数的最小二乘估计量；在基本假定满足的条件下，多元线性回归模型最小二乘估计式是最佳线性无偏估计量。
4. 多元线性回归模型中参数区间估计的方法。

5. 多重可决系数的意义和计算方法，修正可决系数的作用和方法。

6. 对多元线性回归模型中所有解释变量联合显著性的F检验。

7. 多元回归分析中，对各个解释变量是否对被解释变量有显著影响的t检验。

8. 利用多元线性回归模型作被解释变量平均值预测与个别值预测的方法。

第三讲 结束了！

THANKS