第三章 多元线性回归

- 多元线性回归模型
- 多元线性回归模型的参数估计
- 多元线性回归模型的统计检验
- 多元线性回归模型的预测
- 回归模型的其他形式
- 回归模型的参数约束

§3.1 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

一、多元线性回归模型

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

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:

$$E(Y_i\mid X_{1i},X_{2i},\cdots X_{ki})=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}$$

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

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

其中

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

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

样本回归函数：用来估计总体回归函数

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

其**随机表示式**：

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

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**：

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

其中：

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

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**X**是非随机的，且**X**的秩$\rho=k+1$，即**X**满秩。

假设2，

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

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

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

假设3，E(**X'μ**)=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

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

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

假设6，回归模型的设定是正确的。

§3.2 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

一、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：

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

根据最小二乘原理，参数估计值应该是下列方程组的解

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

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

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

由于**X’X**满秩，故有

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

将上述过程用矩阵表示如下：

寻找一组参数估计值 $\hat{\beta}$，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：$\dfrac{\partial}{\partial \hat{\beta}}(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0$

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

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

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

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

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

正规方程组的另一种写法

对于正规方程组

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

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

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

于是

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

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

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

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

*二、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

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

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

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

即为变量Y的或然函数

对数或然函数为

$$
\begin{aligned}
L^* &= Ln(L) \\
&= -nLn(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\hat{\beta})\ (\mathbf{Y} - \mathbf{X}\hat{\beta})
\end{aligned}
$$

对对数或然函数求极大值，也就是对

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

求极小值。

因此，参数的最大或然估计为

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

结果与参数的普通最小二乘估计相同

*三、矩估计（Moment Method, MM）

OLS估计得到的正规方程组可以从另外一种思路来推导:

总体矩条件: $E(\mathbf{X}'\mu)=0$

样本矩条件: $\frac{1}{n}\mathbf{X}'(\mathbf{Y}-\mathbf{X}\hat{\beta})=0$

由此得到正规方程组 $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{Y}$

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

$$\mathrm{E}(\boldsymbol{X}'\boldsymbol{\mu})=\mathbf{0}$$

四、参数估计量的性质

在满足基本假设的情况下，其结构参数$\boldsymbol{\beta}$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

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

其中, $\mathbf{C=(X'X)^{-1} X'}$ 为一仅与固定的**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($\boldsymbol{X}'\boldsymbol{\mu}$)=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**

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

$$= E((\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}'\mu\mu'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mu\mu')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= E(\mu\mu')(\mathbf{X}'\mathbf{X})^{-1}$$

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

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

其中利用了

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

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

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

和 $E(\mu\mu') = \sigma^2\mathbf{I}$

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

五、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

§3.3 多元线性回归模型的统计检验

一、拟合优度检验

二、方程的显著性检验(F检验)

三、变量的显著性检验（t检验）

四、参数的置信区间

一、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

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

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

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

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

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。

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:**

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

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。

二、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： $\beta_1=\beta_2=\ \dots\ =\beta_k=0$

H_1： β_j不全为0（$j=1,2,\dots,k$）

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$(k, n\text{-}k\text{-}1)$的F分布

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

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

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。

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：

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

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。

三、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

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

以c_{ii}表示矩阵$(X'X)^{-1}$ 主对角线上的第i+1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：

$$H_0：\ \beta_i=0 \qquad （i=1,2\ldots k）$$

$$H_1：\ \beta_i\neq 0$$

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

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

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：

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

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

四、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−α)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。

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。

例 1. 某地区通过一个样本容量为 722 的调查数据得到劳动力受教育的一个回归方程为

$$edu = 10.36 - 0.094 sibs + 0.131 medu + 0.210 fedu$$

R^2=0.214

式中，edu 为劳动力受教育年数，sibs 为该劳动力家庭中兄弟姐妹的个数，medu 与 fedu 分别为母亲与父亲受到教育的年数。问

（1）sibs 是否具有预期的影响？为什么？若 medu 与 fedu 保持不变，为了使预测的受教育水平减少一年，需要 sibs 增加多少？

（2）请对 medu 的系数给予适当的解释。

（3）如果两个劳动力都没有兄弟姐妹，但其中一个的父母受教育的年数为 12 年，另一个的父母受教育的年数为 16 年，则两人受教育的年数预期相差多少？

例2 假设要求你建立一个计量经济模型来说明在学校跑道上慢跑一英里或一英里以上的人数，以便决定是否修建第二条跑道以满足所有的锻炼者。你通过整个学年收集数据，得到两个可能的解释性方程：

方程 A：$\hat{Y} = 125.0 - 15.0X_1 - 1.0X_2 + 1.5X_3$　　　$\bar{R}^2 = 0.75$

方程 B：$\hat{Y} = 123.0 - 14.0X_1 + 5.5X_2 - 3.7X_4$　　　$\bar{R}^2 = 0.73$

其中：Y——某天慢跑者的人数

X_1——该天降雨的英寸数

X_2——该天日照的小时数

X_3——该天的最高温度（按华氏温度）

X_4——第二天需交学期论文的班级数

请回答下列问题：（1）这两个方程你认为哪个更合理些，为什么？

（2）为什么用相同的数据去估计相同变量的系数得到不同的符号？

例3 下表给出三变量模型的回归结果：

方差来源	平方和（SS）	自由度（d.f.）	平方和的均值(MSS)
来自回归	65965	—	—
来自残差	—	—	—
总离差(TSS)	66042	14	

要求：（1）样本容量是多少？

（2）求 RSS？

（3）ESS 和 RSS 的自由度各是多少？

（4）求 R^2 和 $\overline{R}^2$？

（5）检验假设：X_2 和 X_3 对 Y 无影响。你用什么假设检验？为什么？

例4　已知线性回归模型 $\mathbf{Y}=\mathbf{X}B+\mathbf{U}$ 式中 $\mathbf{U}\sim$（0，$\sigma^2\mathbf{I}$），$n=13$ 且 $k=3$（n 为样本容量，k 为参数的个数），由二次型 $(\mathbf{Y}-\mathbf{X}B)'(\mathbf{Y}-\mathbf{X}B)$ 的最小化得到如下线性方程组：

$$\hat{\beta}_1+2\hat{\beta}_2+\hat{\beta}_3=3$$

$$2\hat{\beta}_1+5\hat{\beta}_2+\hat{\beta}_3=9$$

$$\hat{\beta}_1+\hat{\beta}_2+6\hat{\beta}_3=-8$$

要求：（1）把问题写成矩阵向量的形式；用求逆矩阵的方法求解之；

（2）如果 $\mathbf{Y}'\mathbf{Y}=53$，求 $\hat{\sigma}^2$；

（3）求出 $\hat{\beta}$ 的方差—协方差矩阵。

3-14．对模型 $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$ 应用 OLS 法，得到回归方程如下：

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

要求：证明残差 $\varepsilon_i = y_i - \hat{y}_i$ 与 $\hat{y}_i$ 不相关，即：$\sum \hat{y}_i \varepsilon_i = 0$。

3-19．假定以校园内食堂每天卖出的盒饭数量作为被解释变量，盒饭价格、气温、附近餐厅的盒饭价格、学校当日的学生数量（单位：千人）作为解释变量，进行回归分析；假设不管是否有假期，食堂都营业。不幸的是，食堂内的计算机被一次病毒侵犯，所有的存储丢失，无法恢复，你不能说出独立变量分别代表着哪一项！下面是回归结果（括号内为标准差）：

$$\hat{Y}_i = 10.6 + 28.4X_{1i} + 12.7X_{2i} + 0.61X_{3i} - 5.9X_{4i}$$

$$(2.6) \qquad (6.3) \qquad (0.61) \qquad (5.9) \qquad \overline{R}^2 = 0.63 \qquad n = 35$$

要求：

（1）试判定每项结果对应着哪一个变量？

（2）对你的判定结论做出说明。