计量经济学

第 三 章

多元线性回归模型

引子：中国汽车的保有量会超过一亿辆吗？

影响中国汽车行业发展的因素是多方面的:经济增长、消费趋势、市场行情、业界心态，内外环境，都会使中国汽车行业面临机遇和挑战。

应当具体分析这样一些问题：

中国汽车市场发展的状况如何？（用销售量观测）

影响中国汽车销量的主要因素是什么？（如收入、价格、费用、道路状况、政策环境等）

各种因素对汽车销量影响的性质怎样？（正、负）

各种因素影响汽车销量的具体数量关系是什么？

所得到的数量结论是否可靠？

中国汽车行业今后的发展前景怎样？应当如何制定汽车的产业政策？

很明显，还需要寻求有多个解释变量的回归分析方法。

第一节 多元线性回归模型及古典假定

一、多元线性回归模型的意义

例如:电力供应模型 $Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u_i$

其中：Y_i 为各地区电力消费量；X_2 为各地区国内生产总值（GDP）；X_3 为各地区电力价格变动

一般形式：对于有K个解释变量的线性回归模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

模型中 β_j（j=1,2,---k）是偏回归系数

偏回归系数：

控制其它解释量不变的条件下，第j个解释变量的单位变动对应变量平均值的影响

多元线性回归：指对各个回归系数而言是“线性”的，对变量则可是线性的，也可是非线性的

例如：生产函数

$$Y = AL^{\alpha} K^{\beta} u$$

取对数

$$\ln Y = \ln A + \alpha \ln L + \beta \ln K + \ln u$$

多元总体回归函数与多元样本回归函数

多元总体回归函数： Y 的总体条件均值表示为多个解释变量的函数

$$E(Y_i \mid X_{2i}, X_{3i}, \cdots X_{ki}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}$$

注意：这时Y总体条件均值的轨迹是K维空间的一条线

或表示为 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$

多元样本回归函数： Y 的样本条件均值表示为多个解释变量的函数

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}$$

或

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + e_i$$

回归剩余（残差）：

$$e_i = Y_i - \hat{Y}_i$$

二、多元线性回归模型的矩阵表示

K个解释变量的多元线性回归模型的 n个观测样本，可表示为

$$Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1$$
$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2$$
$$\cdots\cdots\cdots\cdots$$
$$Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n$$

用矩阵表示

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \cdots & \cdots & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\mathbf{Y} \quad \mathbf{X} \quad \boldsymbol{\beta} \quad \mathbf{u}$$

$$n\times 1 \quad n\times k \quad k\times 1 \quad n\times 1$$

用矩阵表示

总体回归函数 $E(\mathbf{Y})=\mathbf{X}\boldsymbol{\beta}$ 或 $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$

样本回归函数 $\hat{\mathbf{Y}}=\mathbf{X}\hat{\boldsymbol{\beta}}$ 或 $\mathbf{Y}=\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}$

其中：$\mathbf{Y},\hat{\mathbf{Y}},\mathbf{u},\mathbf{e}$ 都是有n个元素的列向量

$\boldsymbol{\beta},\hat{\boldsymbol{\beta}}$ 是有k 个 元素的列向量

$\mathbf{X}$ 是第一列为1的n×k阶解释变量数据矩阵 (截距项可视为解释变量取值为1)

三、多元线性回归中的基本假定

假定1：零均值假定

$$E(u_i)=0 \qquad (\text{i=1,2,---n}) \quad \text{或} \quad \text{E（u）=0}$$

假定2和假定3：同方差和无自相关假定：

$$Cov(u_i,u_j)=E[(u_i-Eu_i)(u_j-Eu_j)]=E(u_iu_j)=\begin{cases}\sigma^2 & (\text{i=j})\\ 0 & (\text{i}\neq\text{j})\end{cases}$$

假定4：随机扰动项与解释变量不相关

$$\text{Cov}(X_{ki},u_i)=0 \qquad \text{k=2,3,----k}$$

假定5:无多重共线性假定（多元中）

假定各解释变量之间不存在线性关系，或各个解释变量观测值之间线性无关。或解释变量观测值矩阵X列满秩(K列)。

$$\text{Ran(X)= k} \longrightarrow \text{Rak(X`X)=K} \longrightarrow \text{即 (X`X) 可逆}$$

假定6：正态性假定 $u_i \sim N(0,\sigma^2)$

第二节　多元线性回归模型的估计

一、普通最小二乘法（OLS）

原则：剩余平方和最小　$\min: \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$

$$\min: \sum e_i^2 = \sum [Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki})]^2$$

求偏导,令其为0　$\dfrac{\partial(\sum e_i^2)}{\partial \hat{\beta}_j} = 0$

即

$$-2\sum \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_{ki} X_{ki}) \right] = 0 \rightarrow \sum e_i = 0$$

$$-2\sum X_{2i} \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_{ki} X_{ki}) \right] = 0 \rightarrow \sum X_{2i} e_i = 0$$

$$\vdots$$

$$-2\sum X_{ki} \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_{ki} X_{ki}) \right] = 0 \rightarrow \sum X_{ki} e_i = 0$$

注意到　$\left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}) \right] = e_i$

用矩阵表示

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

$\mathbf{X'}$ $\mathbf{e}$

因为样本回归函数为 $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$

两边乘 $\mathbf{X'}$ $\mathbf{X'Y} = \mathbf{X'X}\hat{\boldsymbol{\beta}} + \mathbf{X'e}$

因为 $\mathbf{X'e} = \mathbf{0}$

则正规方程为 $\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'Y}$

◆ **OLS估计式：**

由正规方程 $\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'Y}$ $(\mathbf{X'X})_{k\times k}$是满秩矩阵,其逆存在

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

二元回归中： $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3$

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

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

注意：x 和 y 为**X**、**Y**的离差

二、OLS估计式的性质

OLS估计式

1、 线性特征 $\hat{\beta} = (X'X)^{-1}X'Y$

$\hat{\beta}$ 是Y是线性函数，因 $(X'X)^{-1}X'$ 是非随机或取固定值的矩阵

2、 无偏特性 $E(\hat{\beta}_K) = \beta_K$

3、 最小方差特性

在 β_k 所有的线性无偏估计中，**OLS**估计 $\hat{\beta}_K$ 具有最小方差

结论：在古典假定下，多元线性回归的OLS估计式是最佳线性无偏估计式（BLUE）

三、OLS估计的分布性质

基本思想：

- $\hat{\beta}$ 是随机变量，必须确定其分布性质才可能进行区间估计和假设检验
- u_i 是服从正态分布的随机变量，决定了Y也是服从正态分布的随机变量
- $\hat{\beta}$ 是Y的线性函数，决定了 $\hat{\beta}$ 也是服从正态分布的随机变量

- $\hat{\beta}$ 的期望　　$E(\hat{\beta}) = \beta$　　（由无偏性）

- $\hat{\beta}$ 的方差和标准误差：

可以证明 $\hat{\beta}$ 的方差—协方差矩阵为

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

$$Var(\hat{\beta}_j) = \sigma^2 c_{jj}$$　　这里的 $(X`X)^{-1} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{bmatrix}$

$$SE(\hat{\beta}_j) = \sigma \sqrt{c_{jj}}$$

（其中 c_{jj} 是矩阵$(X`X)^{-1}$ 中第j行第j列的元素）

所以　$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$　　（j=1,2,---k）

四、随机扰动项方差 σ^2 的估计：

多元回归中 σ^2 的无偏估计为：

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} \quad \text{或表示为} \quad \hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k}$$

将 $\hat{\boldsymbol{\beta}}$ 作标准化变换：

$$z_k = \frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} = \frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{c_{jj}}} \sim N(0,1)$$

因 σ^2 是未知的，可用 $\hat{\sigma}^2$ 代替 σ^2 去估计参数 $\hat{\boldsymbol{\beta}}$ 的标准误差：

- 当为大样本时，用估计的参数标准误差对 $\hat{\boldsymbol{\beta}}$ 作标准化变换，所得 Z 统计量仍可视为服从正态分布
- 当为小样本时，用估计的参数标准误差对 作标准化变换，所得的 t 统计量服从 t 分布：

$$t = \frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} \sim t(n-k)$$

五、回归系数的区间估计

由于 $$t^* = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-k)$$

给定 α ，查t分布表的自由度为n-k的临界值 $t_{\alpha/2}(n-k)$

$$P[-t_{\alpha/2}(n-k) \le t^* = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \le t_{\alpha/2}(n-k)] = 1-\alpha \qquad (j=1\cdots k)$$

或

$$P[\hat{\beta}_j - t_{\alpha/2} SE(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2} SE(\hat{\beta}_j)] = 1-\alpha$$

或表示为

$$P[\hat{\beta}_j - t_{\alpha/2}\hat{\sigma}\sqrt{c_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2}\hat{\sigma}\sqrt{c_{jj}}] = 1-\alpha$$

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

第三节　多元线性回归模型的检验

一、多元回归的拟合优度检验

多重可决系数：在多元回归模型中，由各个解释变量联合解释了的Y的变差，在Y的总变差中占的比重，用 R^2 表示

与简单线性回归中可决系数 r^2 的区别只是 $\hat{Y}_i$ 不同，多元回归中

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}$$

多重可决系数也可表示为

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{TSS - RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

多重可决系数的矩阵表示：

$$TSS = \mathbf{Y}'\mathbf{Y} - N\bar{Y}^2$$

$$ESS = \hat{\beta}\,\mathbf{X}'\mathbf{Y} - N\bar{Y}^2$$

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}\,\mathbf{X}'\mathbf{Y} - N\bar{Y}^2}{\mathbf{Y}'\mathbf{Y} - N\bar{Y}^2}$$

特点：

可以证明 $R^2 = \dfrac{\hat{\beta}_2 \sum x_{2i} y_i + \hat{\beta}_3 \sum x_{3i} y_i + \cdots + \hat{\beta}_k \sum x_{ki} y_i}{\sum y_i^2}$

多重可决系数是模型中解释变量个数的不减函数，这给对比不同模型的多重可决系数带来缺陷，所以需要修正

修正的可决系数

思想：可决系数只涉及变差，没有考虑自由度。如果用自由度去校正所计算的变差，可纠正解释变量个数不同引起的对比困难。

自由度：统计量的自由度指可自由变化的样本观测值个数，它等于所用样本观测值的个数减去对观测值的约束个数。

可决系数的修正方法：

总变差TSS $= \sum (Y_i - \bar{Y})^2 = \sum y_i^2$ 自由度为n—1

解释了的变差ESS $= \sum (\hat{Y}_i - \bar{Y})^2$ 自由度为k—1

剩余平方和RSS $= \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$ 自由度为n—k

修正的可决系数为

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

修正的可决系数$\overline{R}^2$与可决系数R^2的关系

关系：

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

特点：

可决系数R^2必定非负，但修正的可决系数$\overline{R}^2$可能为负值，这时规定$\overline{R}^2=0$

二、回归方程的显著性检验（F检验）

基本思想：

在多元回归中有多个解释变量，需要说明所有解释变联合起来对应变量影响的总显著性,或整个方程总的联合显著性。对方程总显著性检验需要在方差分析的基础上进行F检验。

1、方差分析

在讨论可决系数时已经分析了总变差TSS的分解及自由度：

$$TSS=ESS+RSS$$

Y的样本方差为：总变差/自由度　即

$$\hat{\sigma}_{Y_i}^2 = \frac{TSS}{n-1} = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

显然，Y的方差也可分解为两部分，可用方差分析表分解：

方差分析表

总变差 **TSS=** $\sum (Y_i - \bar{Y})^2$ 自由度 **n－1**

模型解释了的变差**ESS=** $\sum (\hat{Y}_i - \bar{Y})^2$ 自由度 **k－1**

剩余变差**RSS=** $\sum (Y_i - \hat{Y}_i)^2$ 自由度 **n－k**

变差来源	平方和	自由度	方差
归于回归模型	**ESS=** $\sum (\hat{Y}_i - \bar{Y})^2$	**k－1**	**ESS/(k－1)**
归于剩余	**RSS=** $\sum (Y_i - \hat{Y}_i)^2$	**n－k**	**RSS/(n－k)**
总变差	**TSS=** $\sum (Y_i - \bar{Y})^2$	**n－1**	**TSS/(n－1)**

2、F检验

原假设　$H_0:\beta_2=\beta_3=\cdots=\beta_k=0$

备择假设　$H_1:\beta_j(j=1,2,\cdots k)$　不全为**0**

建立统计量(可以证明):

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

给定显著性水平α，查**F**分布表中自由度为 **k-1** 和 **n-k** 的临界值 $F_\alpha(k-1,n-k)$，并通过样本观测值计算**F**值

▼如果计算的**F**值大于**F**临界值 $F_\alpha(k-1,n-k)$ (小概率)

则拒绝 $H_0:\beta_2=\beta_3=\cdots=\beta_k=0$，说明回归模型有显著意义，即所有解释变量联合起来对**Y**有显著影响。

▼如果计算的**F**值小于临界值) $F_\alpha(k-1,n-k)$ (大概率)

则接受 $H_0:\beta_2=\beta_3=\cdots=\beta_k=0$，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。

3、可决系数的显著性检验

由方差分析可以看出，F检验与可决系数有密切联系，二者都建立在对应变量变差分解的基础上。F统计量也可通过可决系数计算：

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

可看出：当 R^2 =0时，F=0

当 R^2 越大时，F值也越大

当 R^2 =1时，F→∞

结论：对方程联合显著性检验的F检验，实际上也是对R^2的显著性检验。

三、各回归系数的假设检验（t检验）

目的：

在多元回归中，分别检验当其他解释变量保持不变时，各个解释变量X对应变量Y是否有显著影响。

方法：

原假设　$H_0: \beta_j = 0$　（j=1,2,……k）

备择假设　$H_1: \beta_j \neq 0$

统计量t为：

$$t^* = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-k)$$

t 检验的方法

给定显著性水平 α，查自由度为n－k时t分布表的临界值为 $t_{\alpha/2}(n-k)$

如果 $-t_{\alpha/2}(n-k) \le t^* \le t_{\alpha/2}(n-k)$

就接受 $H_1: \beta_j = 0$ $H_1: \beta_j \neq 0$ 而拒绝

即认为 β_j 所对应的解释变量 对应变量Y的影响不显著。

$t^* < -t_{\alpha/2}(n-k)$ 或 $t^* > t_{\alpha/2}(n-k)$

如果 H_0 $H_1: \beta_j \neq 0$

就拒绝 而接受 X_j

即认为 所对应的解释变量 对应变量Y的影响是显著的。

（注意：这里是双尾检验）

$F = t^2$

在多元回归中，可分别对每个回归系数逐个地进行 t 检验。

注意:在一元回归中F检验与 t 检验等价.且

第四节　多元线性回归模型的预测

一、应变量平均值预测

1、Y平均值的点预测

将解释变量预测值代入估计的方程：

多元回归时：

$$\hat{Y}_F = \hat{\beta}_1 + \hat{\beta}_2 X_{F2} + \hat{\beta}_3 X_{F3} + \cdots + \hat{\beta}_K X_{Fk}$$

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

注意:预测期的 X_F是第一个元素为1的行向量,不是矩阵,也不是列向量

2、Y平均值的区间预测

基本思想：

- 由于存在抽样波动，预测的平均值 $\hat{Y}_F$ 不一定等于真实平均值 $E(Y_F|X_F)$，还需要对 $E(Y_F|X_F)$ 作区间估计
- 为对Y作区间预测，必须确定平均值预测值 $\hat{Y}_F$ 的抽样分布
- 必须找出与 $\hat{Y}_F$ 和 $E(Y_F|X_F)$ 都有关的统计量

具体作法 (回顾一元回归)

一元中已知 $\hat{E(Y_F)} = E(Y_F | X_F) = \beta_1 + \beta_2 X_F$

$$\hat{SE(Y_F)} = \sigma \sqrt{\frac{1}{n} + \frac{(X_F - \overline{X})^2}{\sum x_i^2}}$$

$$\hat{Var(Y_F)} = \sigma^2 [\frac{1}{n} + \frac{(X_F - X)^2}{\sum x_i^2}]$$

当 σ^2 未知时，只得用 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 代替，这时

$$\hat{Var(Y_F)} = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_F - \overline{X})^2}{\sum x_i^2} \right]$$

多元回归时，与 $\hat{Y}_F$ 和 $E(Y_F|X_F)$ 都有关的是偏差 w_F

$$w_F = \hat{Y}_F - E(Y_F|X_F)$$

w_F 服从正态分布，可证明

$$E(w_F)=0 \qquad Var(w_F)=\sigma^2\mathbf{X_F}(\mathbf{X'X})^{-1}\mathbf{X_F}$$

用 $\hat{\sigma}^2=\sum e_i^2/(n-k)$ 代替 σ^2，可构造 t 统计量

$$t^*=\frac{w_F-E(w_F)}{\hat{SE}(w_F)}=\frac{\hat{Y}_F-E(Y_F|X_F)}{\hat{\sigma}\sqrt{\mathbf{X_F}(\mathbf{X'X})^{-1}\mathbf{X'_F}}}\sim t(n-k)$$

则给定显著性水平α，查t分布表，得自由度 n-k 的临界值 $t_{\alpha/2}(n-k)$，则

$$P\{[(\hat{Y}_F - t_{\alpha/2}\hat{SE}(\hat{Y}_F)] \le E(Y_F) \le [(\hat{Y}_F + t_{\alpha/2}\hat{SE}(\hat{Y}_F)]\}$$

$$= 1-\alpha$$

或

$$P\{[\hat{Y}_F - t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{X_F}(\mathbf{X`X})^{-1}\mathbf{X'_F}}] \le E(Y_F) \le [\hat{Y}_F + t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{X_F}(\mathbf{X`X})^{-1}\mathbf{X'_F}}]\}$$

$$= 1-\alpha$$

三、应变量个别值预测

基本思想：

- $\hat{Y}_F$ 既是对Y平均值的点预测，也是对Y个别值的点预测。
- 由于存在随机扰动 u_i 的影响，Y的平均值并不等于Y的个别值
- 为了对Y的个别值 Y_F 作区间预测，需要寻找与预测值 $\hat{Y}_F$ 和个别值 Y_F 有关的统计量，并要明确其概率分布

具体作法：

已知剩余项 e_F 是与预测值 $\hat{Y}_F$ 和个别值 Y_F 都有关的变量

并且已知 e_F 服从正态分布，且可证明

$$E(e_F)=0$$

$$Var(e_F)=\sigma^2[1+\mathbf{X_F}(\mathbf{X'X})^{-1}\mathbf{X_F}]$$

当用 $\hat{\sigma}^2=\sum e_i^2/(n-k)$ 代替 σ^2 时，对 e_F 标准化的变量 t 为：

$$t=\frac{e_F-E(e_F)}{\hat{SE}(e_F)}=\frac{Y_F-\hat{Y}_F}{\hat{\sigma}\sqrt{1+\mathbf{X}_F(\mathbf{X'X})^{-1}\mathbf{X'_F}}}\sim t(n-k)$$

给定显著性水平 α ，查t分布表得自由度为 n-k 的临界值 $t_{\alpha/2}(n-k)$ 则

$$P(\{[\hat{Y}_F - t_{\alpha/2}\hat{SE}(e_F)] \le Y_F \le [\hat{Y}_F + t_{\alpha/2}\hat{SE}(e_F)]\} = 1-\alpha$$

因此，多元回归时Y的个别值的置信度1－α的预测区间的上下限为

$$Y_F = \hat{Y}_F \mp t_{\alpha/2}\hat{\sigma}\sqrt{1+\mathbf{X_F}(\mathbf{X'X})^{-1}\mathbf{X'_F}}$$

第五节 案例分析

案例一：中国税收增长的分析

提出问题： 改革开放以来，随着经济体制改革的深化和经济的快速增长，中国的财政收支状况发生很大变化，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的增长规律，预测中国税收未来的增长趋势，需要建立计量经济模型。

理论分析： 影响中国税收收入增长的主要因素可能有：

（1）从宏观经济看，经济整体增长是税收增长的基本源泉。

（2）社会经济的发展和社会保障等都对公共财政提出要求，公共财政的需求对当年的税收收入可能会有一定的影响。

（3）物价水平。中国的税制结构以流转税为主，以现行价格计算的GDP和经营者的收入水平都与物价水平有关。

（4）税收政策因素。

建立模型：

分析:以各项税收收入作为被解释变量

以GDP表示经济整体增长水平

以财政支出表示公共财政的需求

以商品零售价格指数表示物价水平

税收政策因素较难用数量表示,暂时不予考虑。

模型设定为：

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

其中：Y —各项税收收入（亿元）

X2 —国内生产总值（亿元）

X3 —财政支出（亿元）

X4 —商品零售价格指数（%）

数据收集：

obs	Y	X2	X3	X4
1978	519.2800	3624.100	1122.090	100.7000
1979	537.8200	4038.200	1281.790	102.0000
1980	571.7000	4517.800	1228.830	106.0000
1981	629.8900	4862.400	1138.410	102.4000
1982	700.0200	5294.700	1229.980	101.9000
1983	775.5900	5934.500	1409.520	101.5000
1984	947.3500	7171.000	1701.020	102.8000
1985	2040.790	8964.400	2004.250	108.8000
1986	2090.730	10202.20	2204.910	106.0000
1987	2140.360	11962.50	2262.180	107.3000
1988	2390.470	14928.30	2491.210	118.5000
1989	2727.400	16909.20	2823.780	117.8000
1990	2821.860	18547.90	3083.590	102.1000
1991	2990.170	21617.80	3386.620	102.9000
1992	3296.910	26638.10	3742.200	105.4000
1993	4255.300	34634.40	4642.300	113.2000
1994	5126.880	46759.40	5792.620	121.7000
1995	6038.040	58478.10	6823.720	114.8000
1996	6909.820	67884.60	7937.550	106.1000
1997	8234.040	74462.60	9233.560	100.8000
1998	9262.800	78345.20	10798.18	97.40000
1999	10682.58	82067.50	13187.67	97.00000
2000	12581.51	89468.10	15886.50	98.50000
2001	15301.38	97314.80	18902.58	99.20000
2002	17636.45	104790.6	22053.15	98.70000

数据来源：

《中国统计年鉴》

其中：

Y——各项税收收入（亿元）

X2——国内生产总值（亿元）

X3——财政支出（亿元）

X4——商品零售价格指数（%）

参数估计：

假定模型中随机项满足基本假定，可用OLS法估计其参数。

具体操作：用Eviews软件包，估计结果为

Dependent Variable: Y
Method: Least Squares
Date: 07/05/05 Time: 16:54
Sample: 1978 2002
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2582.791	940.6128	-2.745860	0.0121
X2	0.022067	0.005577	3.956605	0.0007
X3	0.702104	0.033236	21.12466	0.0000
X4	23.98541	8.738302	2.744859	0.0121

R-squared	0.997430	Mean dependent var	4848.366
Adjusted R-squared	0.997063	S.D. dependent var	4870.971
S.E. of regression	263.9599	Akaike info criterion	14.13512
Sum squared resid	1463172.	Schwarz criterion	14.33014
Log likelihood	-172.6890	F-statistic	2717.238
Durbin-Watson stat	0.948542	Prob(F-statistic)	0.000000

模型估计的结果可表示为：

$$\hat{Y}_i = -2582.791 + 0.022067X_2 + 0.702104X_3 + 23.98541X_4$$

(940.6128) (0.0056) (0.0332) (8.7363)

t= (-2.7459) (3.9566) (21.1247) (2.7449)

$R^2 = 0.9974$ $\bar{R}^2 = 0.9971$ F=2717.238 df=21

模型检验：

- 拟合优度：可决系数 $R^2 = 0.9974$ 较高，修正的可决系数 $\bar{R}^2 = 0.9971$ 也较高，表明模型拟合较好。

●显著性检验：

F检验：针对 $H_0:\beta_2=\beta_3=\beta_4=0$ $\alpha=0.05$ ，取 ，查出自由度为k-1=3和n-k=21的临界值 。由于F=2717.238＞3.075，应拒绝 $H_0:\beta_2=\beta_3=\beta_4=0$ ，说明回归方程显著，即

$F_\alpha(3,21)=3.075$

“国内生产总值”、“财政支出”、“商品零售物价指数”等变量联合起来确实对“税收收入”有显著影响。

t 检验：给定 $\alpha=0.05$ ，查 t 分布表，在自由度为n-3=18-3=15 时临界值为 $t_{0.025}(21)=2.080$ ，因为X2、X3、X4参数对应的 t

● 经济意义检验：

本模型 $\hat{\beta}_2 = 0.022067, \hat{\beta}_3 = 0.702104, \hat{\beta}_4 = 23.98541$ 中

，所估计的参数的符号与经济理论分析一致，说明在其他因素不变的情况下，国内生产总值每增加1亿元，平均说来财政收入将增加220·67万元；财政支出每增加1亿元，平均说来财政收入将增加7021·04万元;商品零售物价指数每增加1%,平均说来财政收入将增加23.98541亿元。

第三章小结

1、多元线性回归模型是将总体回归函数描述为一个被解释变量与多个解释变量之间线性关系的模型。

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + u_i$$

通常多元线性回归模型可以用矩阵形式表示。

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

2、多元线性回归模型中对随机扰动项u的假定:零均值假定、方差假定、无自相关假定、随机扰动与解释变量不相关假定、正态性假定、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式;

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

参数估计式的分布性质及期望、方差和标准误差;

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \qquad \hat{Var}(\hat{\beta}_j) = \hat{\sigma}^2 C_{jj} = \left(\frac{\sum e_i^2}{n-k}\right) C_{jj} \qquad \hat{SE}(\beta_j) = \sigma\sqrt{C_{jj}}$$

4、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式。

5、多元线性回归模型中参数区间估计的方法。

$$P[\hat{\beta}_j - t_{\alpha/2}\sigma\sqrt{c_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2}\sigma\sqrt{c_{jj}}] = 1-\alpha$$

6、多重可决系数的意义和计算方法：

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

修正可决系数的作用和方法：

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(n-k)}{\sum (Y_i - \bar{Y})^2/(n-1)} = 1 - \frac{n-1}{n-k}\frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

7、F检验是对多元线性回归模型中所有解释变量联合显著性的检验，F检验是在方差分析基础上进行的。

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

8、多元回归分析中，为了分别检验当其它解释变量不变时，各个解释变量是否对被解释变量有显著影响，需要分别对所估计的各个回归系数作t检验。

$$t^* = \frac{\hat{\beta}_j - \beta_j}{\hat{SE}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-k)$$

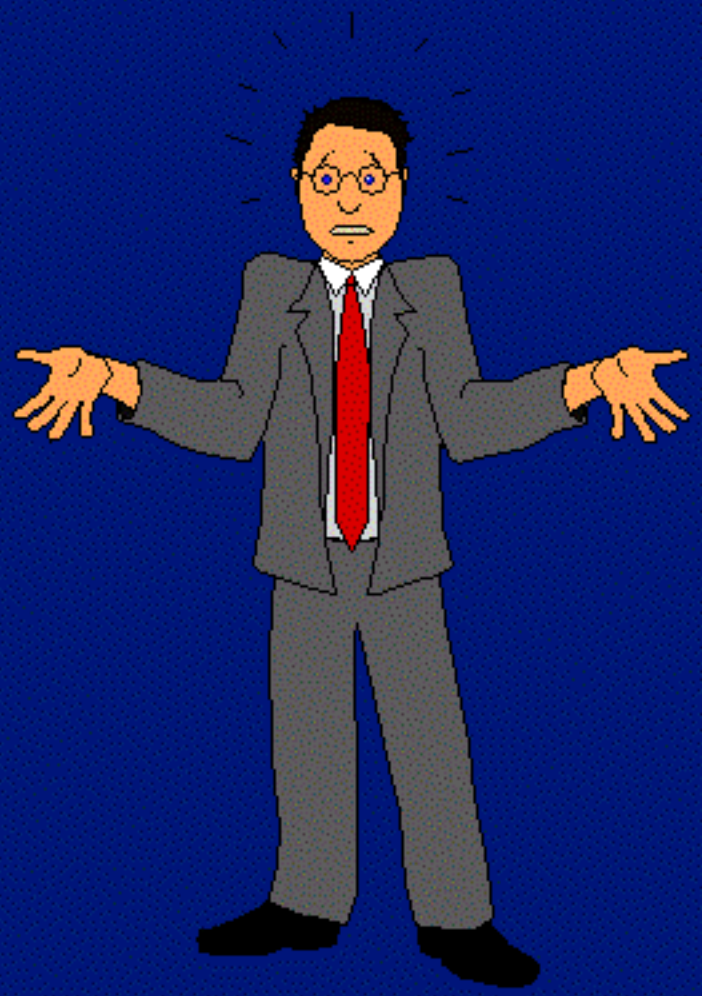
9、利用多元线性回归模型作被解释变量平均值预测与个别值预测的方法。

点预测：$\hat{Y}_f = \mathbf{X}_f\hat{\boldsymbol{\beta}}$

平均值：$\hat{Y}_f - t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f'} \le E(Y_f) \le \hat{Y}_f + t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f'}$

个别值：$\hat{Y}_f - t_{\alpha/2}\hat{\sigma}\sqrt{1+\mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f'} \le Y_f \le \hat{Y}_f + t_{\alpha/2}\hat{\sigma}\sqrt{1+\mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f'}$

第三章结束了！



THANKS