ECONOMETRICS

计量经济学

第三讲

多元线性回归模型

ECONOMETRICS 引子:中国已成为世界汽车产销第一大国

2009年，为应对国际金融危机、确保经济平稳较快增长，国家出台了一系列促进汽车消费的政策，有效刺激了汽车消费市场，汽车产销呈高增长态势，首次成为世界汽车产销第一大国。2009年，汽车产销分别为1379.1万辆和1364.5万辆，同比增长48.3%和46.15%。

是什么因素导致中国汽车数量的增长?

影响中国汽车行业发展的因素并不是单一的，经济增长、消费趋势、市场行情、业界心态、能源价格、道路发展、内外环境，都会使中国汽车行业面临机遇和挑战。

ECONOMETRICS

怎样分析多种因素的影响？

分析中国汽车行业未来的趋势,应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、能源、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，只用一个解释变量已很难分析汽车产业的发展，还需要寻求有更多个解释变量情况的回归分析方法。

本章主要讨论：

- 多元线性回归模型及古典假定
- 多元线性回归模型的估计
- 多元线性回归模型的检验
- 多元线性回归模型的预测

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

一般形式：对于有**K-1**个解释变量的线性回归模型

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

$$(i = 1, 2, \cdots n)$$

注意：模型中的β_j （j=1,2,---k）是偏回归系数

样本容量为n

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被解释变量平均值的影响，即对Y平均值“直接”或“净”的影响。

5

多元线性回归中的“线性”

指对各个回归系数而言是“线性”的，对变量则可以是线性的，也可以是非线性的

例如：生产函数

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

取对数

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

这也是多元线性回归模型，只是这时变量为ln***Y***、ln***L***、ln***K***

ECONOMETRICS

多元总体回归函数

条件期望表现形式：

将Y的总体条件期望表示为多个解释变量的函数，如：

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

$$(i = 1, 2, \cdots n)$$

注意：这时Y总体条件期望的轨迹是K维空间的一条线

个别值表现形式：

引入随机扰动项 $u_i = Y_i - E(Y_i \mid X_{2i}, X_{3i} \cdots X_{ki})$

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

$$(i = 1, 2, \cdots n)$$

多元样本回归函数

Y 的样本条件均值可表示为多个解释变量的函数

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

或回归剩余（残差）：　$e_i = Y_i - \hat{Y}_i$

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

其中　$i = 1, 2, \cdots n$

三、多元线性回归模型的矩阵表示

多个解释变量的多元线性回归模型的n组样本观测值，可表示为

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

用矩阵表示

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

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

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

9

矩阵表示方式

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

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

其中： $Y, \hat{Y}, u, e$ 都是有n个元素的列向量

$\beta, \hat{\beta}$ 是有k 个 元素的列向量

（ k = 解释变量个数 + 1 ）

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总是取值为1)

三、多元线性回归中的基本假定

假定1：零均值假定

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

假定2和假定3：同方差和无自相关假定：

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

或用方差-协方差矩阵表示为：

$$Cov(u_i,u_j)=E\{[u_i-E(u_i)][u_j-E(u_j)]\}=E(\boldsymbol{uu}')$$

$$(i=1,2,\cdots n \qquad j=1,2,\cdots n)$$

$$=\begin{bmatrix}E(u_1u_1) & E(u_1u_2) & \cdots & E(u_1u_n)\\ E(u_2u_1) & E(u_2u_2) & \cdots & E(u_2u_n)\\ \vdots & \vdots & \vdots & \vdots\\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_nu_n)\end{bmatrix}=\sigma^2\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}=\sigma^2\boldsymbol{I}$$

假定4：随机扰动项与解释变量不相关

$$Cov(X_{ji}, u_i) = 0 \qquad (j = 2, 3, \cdots, k)$$

假定5：无多重共线性假定（多元中增加的）

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***的秩为K(注意***X***为n行K列)。

Ran(***X***)= k ⟷ Rak(***X′X***)=k

即 (***X′X***) 可逆

假定6：正态性假定

$$u_i \sim N(0, \sigma^2) \qquad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

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：寻求剩余平方和最小的参数估计式 $\min: \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$

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

即 $$\min: \sum e_i^2 = \min: \mathbf{e}'\mathbf{e} = \min: (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

求偏导，并令其为0 $\partial(\sum e_i^2)/\partial\hat{\beta}_j = 0$ 其中 $(i = 1, 2, \cdots n)$ $(j = 1, 2, \cdots n)$

即

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

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

$$\vdots$$

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

用矩阵表示的正规方程

偏导数

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

X' e 0

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

两边左乘 X' $X'Y = X'X\hat{\beta} + X'e$

根据最小二乘原则 $X'e = 0$

则正规方程为 $X'X\hat{\beta} = X'Y$

ECONOMETRICS

OLS估计式

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

多元回归的OLS估计量为 $\hat{\beta} = (X'X)^{-1}X'Y$

当只有两个解释变量时为：

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$

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

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

对比

简单线性回归中

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

注意： x、y 为X、Y的离差

OLS回归线的数学性质 (与简单线性回归相同)

- 回归线通过样本均值 $\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_3 \overline{X}_3 + \cdots + \hat{\beta}_k \overline{X}_k$
- 估计值 $\hat{Y}_i$ 的均值等于实际观测值 Y_i 的均值 $\sum \hat{Y}_i / n = \overline{Y}$
- 剩余项 e_i 的均值为零 $\overline{e_i} = \sum e_i / n = 0$
- 被解释变量估计值 $\hat{Y}_i$ 与剩余项 e_i 不相关

$$Cov(\hat{Y}_i, e_i) = 0 \qquad \text{或} \qquad \sum (e_i \hat{y}_i) = 0$$

- 解释变量 X_i 与剩余项 e_i 不相关

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

16

ECONOMETRICS

三、OLS估计式的统计性质

1、 **线性特征** $\hat{\beta} = (X'X)^{-1}X'Y$

$\hat{\beta}$ 是**Y**的线性函数，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，$Y = X\beta + u$

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

$\hat{\beta}$ 的期望与方差

- $\hat{\beta}$ 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（见下页）

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

$$Var(\hat{\beta}_j) = \sigma^2 c_{jj}$$

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

这里的 $(X'X)^{-1} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{bmatrix}$

（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）

$\hat{\beta}$ 的方差-协方差

$$COV(\hat{\boldsymbol{\beta}}) = E\{[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})]'\}$$

注意 $\hat{\boldsymbol{\beta}}$是向量 $(i = 1, 2, \cdots n)$ $(j = 1, 2, \cdots n)$

$$= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']$$

(由无偏性)

$$= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}]$$

(由OLS估计式)

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}\boldsymbol{u}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

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

(由同方差性)

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

其中：

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

$$E(\boldsymbol{u}\boldsymbol{u}') = \sigma^2\boldsymbol{I}$$

20

四、 随机扰动项方差 σ^2 的估计

σ^2 一般未知，可证明多元回归中 σ^2 的无偏估计为：

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$$

或表示为 $\hat{\sigma}^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}$

对比： 一元回归中 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$

将 $\hat{\boldsymbol{\beta}}$ 作标准化变换：

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

21

σ^2 未知时 $\hat{\beta}$ 的标准化变换

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

● 当为大样本时，用估计的参数标准误差对 $\hat{\beta}$ 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 $\hat{\beta}$ 作标准化变换，所得的 t 统计量服从 t 分布：

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

22

五、回归系数的区间估计

由于

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

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

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

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

或

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

或表示为

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

23

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起来解释了的**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

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

（注意：红色字体是与一元回归不同的部分）

24

多重可决系数的矩阵表示

$$TSS=\sum(Y_i-\bar{Y})^2=\boldsymbol{Y'Y}-n\bar{Y}^2 \qquad ESS=\sum(\hat{Y}_i-\bar{Y})^2=\hat{\boldsymbol{\beta}}'\boldsymbol{X'Y}-n\bar{Y}^2$$

$$R^2=\frac{ESS}{TSS}=\frac{\hat{\boldsymbol{\beta}}'\boldsymbol{X'Y}-n\bar{Y}^2}{\boldsymbol{Y'Y}-n\bar{Y}^2}$$

可用代数式表达为

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

特点: 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。

ECONOMETRICS

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。

回顾：

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

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

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

修正的可决系数为

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

修正的可决系数$\overline{R}^2$与可决系数R^2的关系

已经导出：

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

注意：

可决系数 R^2 必定非负，但所计算的修正可决系数 $\overline{R}^2$ 有可能为负值

解决办法： 若计算的 $\overline{R}^2<0$ ，规定 $\overline{R}^2$ 取值为0

二、回归方程的显著性检验（F检验）

基本思想：

在多元回归中包含多个解释变量，它们与被解释变量是否有显著关系呢?

当然可以分别检验各个解释变量对被解释变量影响的显著性。

但是我们首先关注的是所有解释变量联合起来对被解释变量影响的显著性,或整个方程总的联合显著性，需要对方程的总显著性在方差分析的基础上进行F检验。

ECONOMETRICS

1.方差分析

在讨论可决系数时已经分析了被解释变量总变差

TSS的分解及自由度：

TSS=ESS+RSS

注意：　　**Y的样本方差= 总变差/自由度**

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

显然，**Y**的样本方差也可分解为两部分，可用方差分析表分解

30
30

ECONOMETRICS

方差分析表

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

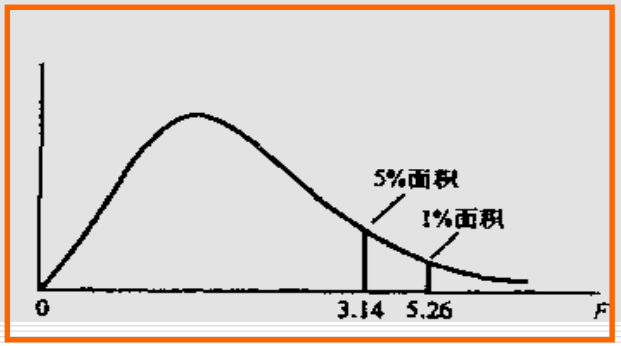
模型解释了的变差 ESS= $\sum(\hat{Y}_i-\bar{Y})^2$ 自由度 K−1

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	$\sum(\hat{Y}_i-\bar{Y})^2/(k-1)$
归于剩余	RSS= $\sum(Y_i-\hat{Y}_i)^2$	n-k	$\sum(Y_i-\hat{Y}_i)^2/(n-k)$
总变差	TSS= $\sum(Y_i-\bar{Y})^2$	n-1	$\sum(Y_i-\bar{Y})^2/(n-1)$

基本思想: 如果多个解释变量联合起来对被解释变量的影响不显著，“归于回归的方差“ 比 “归于剩余的方差” 显著地小应是大概率事件。

2. F检验



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

（所有解释变量联合起来对被解释变量的影响不显著）

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)：

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/(k-1)}{\sum(Y_i - \hat{Y}_i)^2/(n-k)} \sim F(k-1, n-k)$$

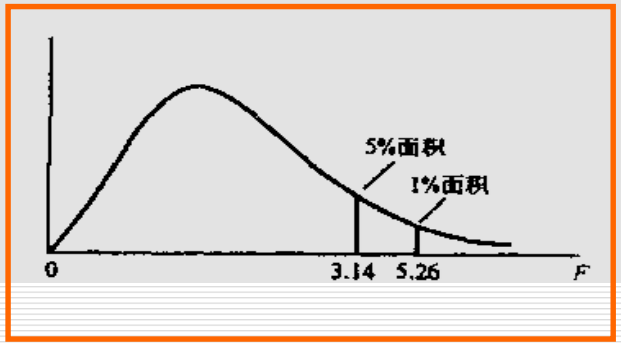
给定显著性水平 α，查**F**分布表中自由度为 **k-1** 和 **n-k** 的临界值 $F_\alpha(k-1, n-k)$，并通过样本观测值计算**F**值

32

F检验方式

▼如果计算的F值大于临界值 $F_\alpha(k-1,n-k)$， 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确有显著影响。

▼如果计算的F值小于临界值 $F_\alpha(k-1,n-k)$，则不拒绝 $H_0:\beta_2=\beta_3=\cdots=\beta_k=0$ ，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。



三、各回归系数的假设检验

注意：在一元回归中F检验与t检验等价，且 $F = t^2$

但在多元回归中，**F**检验显著，不一定每个解释变量都对**Y**有显著影响。还需要分别检验当其他解释变量保持不变时，各个解释变量**X**对被解释变量**Y**是否有显著影响。

方法：

原假设 $H_0 : \beta_j = 0$

备择假设 $H_1 : \beta_j \neq 0$ （j=1,2,……k）

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

对各回归系数假设检验的作法

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

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

就不拒绝 $H_0: \beta_j = 0$ ，而拒绝 $H_1: \beta_j \neq 0$

即认为 β_j 所对应的解释变量 X_j 对被解释变量Y的影响不显著。

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

就拒绝 $H_0: \beta_j = 0$ 而不拒绝 $H_1: \beta_j \neq 0$

即认为 β_j 所对应的解释变量 X_j 对被解释变量Y的影响是显著的。

讨论：在多元回归中，可以作F检验，也可以分别对每个回归系数逐个地进行 t 检验。F 检验与t检验的关系是什么？

第四节 多元线性回归模型的预测

一、被解释变量平均值预测

1.Y平均值的点预测

方法：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

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

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

$$\boldsymbol{X}_F = (1 \quad X_{F2} \quad X_{F3} \quad \cdots X_{Fk} \quad)$$

ECONOMETRICS

2. Y平均值的区间预测

基本思想： （与简单线性回归时相同）

- 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。

- 为了对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，并要明确其概率分布性质。

37

区间预测的具体作法 (回顾简单线性回归)

简单线性回归中 $E(\hat{Y}_F) = E(Y_F \mid X_F) = \beta_1 + \beta_2 X_F$

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

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

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

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

38

区间预测的具体作法（多元时）

多元回归时，与预测的平均值 $\hat{Y}_F$ 和真实平均值 $E(Y_F \mid \boldsymbol{X}_F)$ 都有关的是二者的偏差 w_F：

$$w_F = \hat{Y}_F - E(Y_F \mid \boldsymbol{X}_F)$$

服从正态分布，可证明

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

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

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

或者

$\hat{Y}_F$ 服从正态分布，可证明

$$E(\hat{Y}_F) = E(Y_F \mid \boldsymbol{X_F})$$

$$Var(\hat{Y}_F) = \sigma^2 \boldsymbol{X_F}(\boldsymbol{X'X})^{-1}\boldsymbol{X_F'}$$

即 $$\hat{Y}_F \sim N\{E(Y_F \mid \boldsymbol{X_F}), \sigma^2 \boldsymbol{X_F}(\boldsymbol{X'X})^{-1}\boldsymbol{X_F'}\}$$

标准化 $$t^* = \frac{\hat{Y}_F - E(\hat{Y}_F)}{SE(\hat{Y}_F)} = \frac{\hat{Y}_F - E(Y_F \mid \boldsymbol{X_F})}{\sigma\sqrt{\boldsymbol{X_F}(\boldsymbol{X'X})^{-1}\boldsymbol{X_F'}}} \sim N(0,1)$$

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

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

40

区间预测的具体作法

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

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

或 $=1-\alpha$

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

$$=1-\alpha$$

二、被解释变量个别值预测

基本思想： （与简单线性回归时相同）

- 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性质。

个别值区间预测具体作法

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 $e_F = Y_F - \hat{Y}_F$

并且已知 e_F 服从正态分布，且多元回归时可证明

$$E(e_F) = 0$$

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

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：

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

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

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

ECONOMETRICS

第五节 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

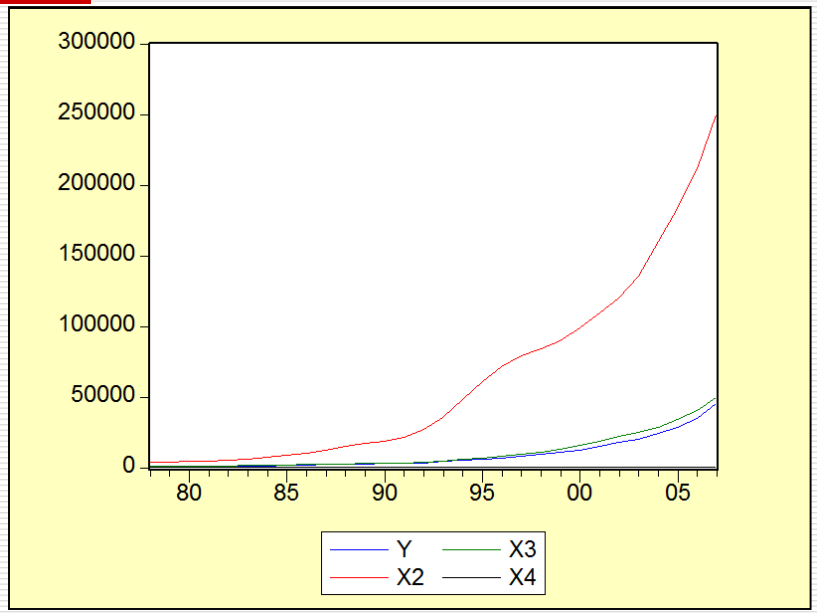
理论分析： 为了全面反映中国税收增长的全貌，选择包括中央和地方税收的“国家财政收入”中的“各项税收”（简称“税收收入”）作为被解释变量；选择国内生产总值（GDP）作为经济整体增长水平的代表；选择中央和地方“财政支出”作为公共财政需求的代表；选择“商品零售价格指数”作为物价水平的代表。

ECONOMETRICS

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
1992	3296.91	26638.1	3742.20	105.4

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	Number of obs =	30
				F(3, 26) =	689.73
Model	52.7568562	3	17.5856187	Prob > F =	0.0000
Residual	.66290406	26	.02549631	R-squared =	0.9876
				Adj R-squared =	0.9862
Total	53.4197602	29	1.8420607	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：

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

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

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

ECONOMETRICS

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。

50

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。

ECONOMETRICS

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
```

ECONOMETRICS

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。

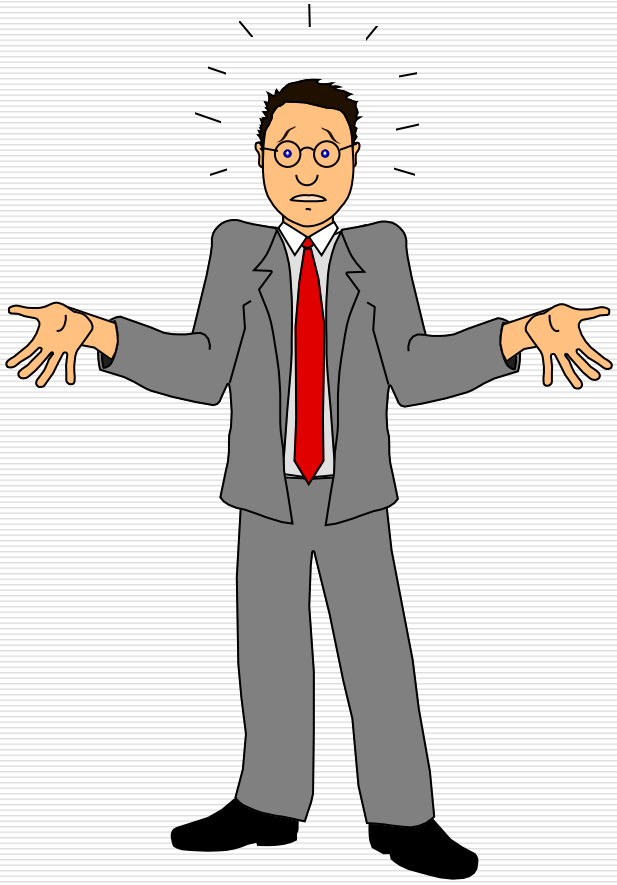
4. 多元线性回归模型中参数区间估计的方法。

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