

# 第四章

## 经典单方程计量经济学模型

### 放宽基本假定的模型

---

## 基本假设：

### 4.4 模型设定偏误问题

1.模型是正确设定的。

### 4.1 多重共线性

2. $X$ 是相互独立的，即 $X_j (j = 1, 2, \dots, k)$ 之间互不相关。

3.随机误差项具有给定 $X$ 条件下的零均值 $E(\mu_i|X) = 0$ ，推出 $Cov(X_i, \mu_i) = 0$ 。

### 4.3 内生解释变量问题

4.随机误差项具有同方差 $Var(\mu_i|X) = \sigma^2$ 、不序列相关 $Cov(\mu_i, \mu_j|X) = 0$ 。

### 4.2 异方差性

### 4.5 序列相关性

5.随机误差项服从正态分布。

**【例】**

$$Y = 0.472 + 0.32 \ln(X_1) + 0.05X_2$$

(1.37)      (0.22)      (0.046)      ———  $Se_{(\hat{\beta})}$

$$\bar{R}^2 = 0.76 \quad F = 34.8 \quad n = 32$$

其中： $Y$ —企业研发支出（R&D）占销售额的比重

$X_1$ —企业销售额

$X_2$ —企业利润占销售额的比重

问题：

- （1）解释 $\ln(X_1)$ 的系数。如果 $X_1$ 增加10%，估计 $Y$ 会变化多少个百分点？
- （2）对模型进行经济意义检验和统计检验（ $\alpha = 5\%$ ）。

表1 对数模型参数的经济含义

模型	因变量	自变量	对系数的解释
标准模型	$Y$	$X$	$\Delta Y = \beta \Delta X$
双对数模型	$\ln Y$	$\ln X$	$\% \Delta Y = \beta \% \Delta X$
半对数模型	$\ln Y$	$X$	$\% \Delta Y = (100\beta) \Delta X$
	$Y$	$\ln X$	$\Delta Y = \left( \frac{\beta}{100} \right) \% \Delta X$

## 第四章 放宽基本假定的模型

### 4.1 多重共线性

### 4.2 异方差性

### 4.3 内生解释变量问题

### 4.4 模型设定偏误问题

### 4.5 序列相关性

---

## 4.1 多重共线性

视频片段



概

为

## ● 多重共线性的类型

### 1. 如果存在

$$C_1X_{i1} + C_2X_{i2} + \cdots + C_kX_{ik} = 0$$

其中 $C_i$ 不全为0，则称为解释变量之间存在**完全共线性**。

( 不常见 )

### 2. 如果存在

$$\begin{cases} C_1X_{i1} + C_2X_{i2} + \cdots + C_kX_{ik} \approx 0 \\ C_1X_{i1} + C_2X_{i2} + \cdots + C_kX_{ik} + v_i = 0 \end{cases}, v_i \text{ 为随机干扰项}$$

其中 $C_i$ 不全为0，则称为解释变量之间存在**近似共线性**。

( 常见 )

## ● 多重共线性的矩阵表示

对于线性回归模型  $Y = X\beta + \mu$ ，完全共线性指  $R(X) < k + 1$ ，即矩阵

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}$$

中，至少有一个列向量可以由其他列向量线性表示出（不包括第一列）。

**【例】**  $X_2 = \lambda X_1$ ，这时， $X_1$ 和 $X_2$ 的相关系数为1， $X_2$ 的解释作用完全可以由 $X_1$ 代替。

**【注意】** 完全共线性并不多见，一般出现的是在一定程度上的共线性，即近似共线性。



## 二、多重共线性产生的原因

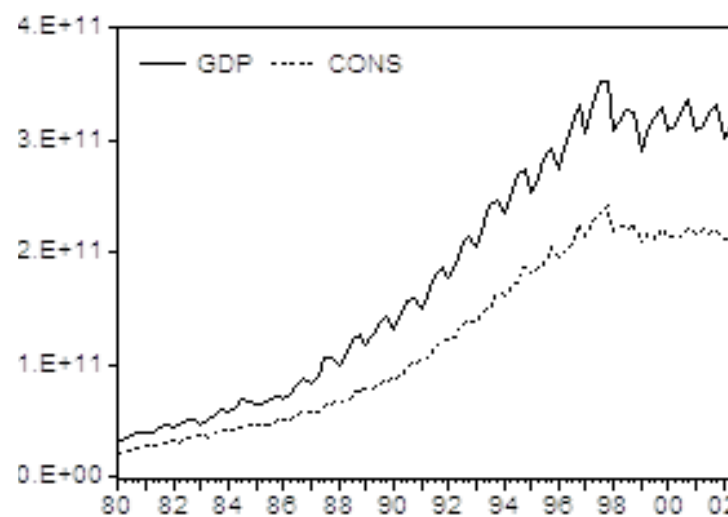
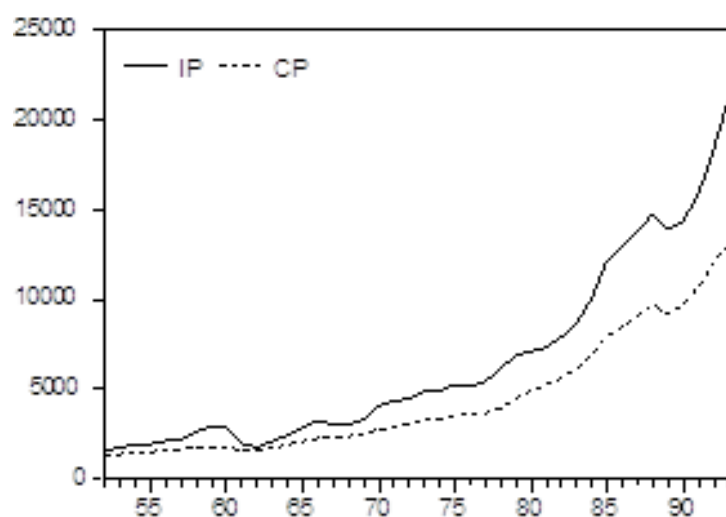
- 1.经济变量的共同趋势
- 2.模型设定不谨慎
- 3.样本资料的限制
- 4.滞后变量的引入

# 1.经济变量的共同趋势

- 时间序列数据

当经济繁荣时期，各基本经济变量（GDP、消费、投资、物价等）都趋于增长；

当经济衰退时期，各变量又同时趋于下降。



## ● 截面数据

生产函数中，资本投入与劳动投入往往呈现高度相关性，大企业资本投入与劳动投入往往都大，小企业资本投入与劳动投入往往都小。

## 2.模型设定不谨慎

【例】估计一个常性消费函数的扩展形式，将模型设定为

$$\ln C_i = \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln Y_i^2 + \mu_i$$

其中：  $C$  为家庭人均消费，  $Y$  为家庭人均收入。

【例】考察学校支出对学生平均成绩的影响

$$Y_i = \beta_0 + \beta_1 X_{i0} + \beta_2 X_{i1} + \beta_3 X_{i2} + \mu_i$$

其中：  $Y$  代表学校的平均成绩，  $X_0$  为学校总支出，  $X_1$   
对教职工的工资性支出，  $X_2$  为其他支出。

### 3.样本资料的限制

在可取得的数据范围内，样本恰好存在某种程度的多重共线性。

**【例】**对家庭孩子的考试分数进行回归，得到模型

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \mu_i$$

其中：Y孩子的考试分数， $X_1$ 为教育支出， $X_2$ 为家庭人均收入

## 4.滞后变量的引入

在计量经济模型中，往往需要引入滞后经济变量来反映真实的经济关系。

**【例】** 消费函数

$$C_t = \beta_0 + \beta_1 I_t + \beta_2 C_{t-1} + \mu_t$$

显然， $I_t$ （当期收入）与 $C_{t-1}$ （前期消费）之间有较强的线性相关性。

### 三、多重共线性产生的后果

#### (一) 完全共线性下，参数估计量不存在。

多元线性回归模型  $Y = X\beta + \mu$  的 OLS 估计量为：

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

如果存在完全共线性，则  $r(X) < k + 1$ ， $X$  为非满秩矩阵，则  $(X'X)^{-1}$  不存在，无法得到参数的估计量。

## 【以二元线性回归模型为例】

二元线性回归模型

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu_i$$

如果两个解释变量完全相关，即 $X_2 = \lambda X_1$ ，该二元线性模型退化为一元线性模型

$$Y_i = \beta_0 + (\beta_1 + \lambda\beta_2) X_1 + \mu_i$$

这时，只能确定综合参数 $\beta_1 + \lambda\beta_2$ 的估计值

$$\widehat{\beta_1 + \lambda\beta_2} = \frac{\sum X_{i1} y_i}{\sum X_{i1}^2}$$

无法确定 $\beta_1$ 和 $\beta_2$ 各自的估计值



## (二) 近似共线性下，OLS估计量无偏但是非有效。

近似共线性下，采用 $OLS$ ，可以得到参数的无偏估计量，估计量的方差表达式为：

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

由于 $|X'X| \approx 0$ ，导致矩阵 $(X'X)^{-1}$ 主对角线上元素较大，即各共线变量的参数 $OLS$ 估计值的方差很大，从而 $OLS$ 估计量非有效。

以二元模型为例，离差形式下 $\hat{\beta}_1$ 的方差为

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \frac{\sigma^2 \sum x_{i2}^2}{\sum x_{i1}^2 \sum x_{i2}^2 - (\sum x_{i1} x_{i2})^2} \\ &= \frac{\sigma^2}{\sum x_{i1}^2} \cdot \frac{1}{1 - r^2} \end{aligned}$$

其中， $\frac{(\sum x_{i1} x_{i2})^2}{\sum x_{i1}^2 \sum x_{i2}^2}$ 为 $X_1$ 与 $X_2$ 的线性相关系数的平方 $r^2$ ，而

$r^2 \leq 1$ ，因此， $\frac{1}{1-r^2} \geq 1$ 。

$$\text{对于 } Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{i1}^2} \cdot \frac{1}{1-r^2}$$

- 当完全不共线时,  $r^2 = 0$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{i1}^2}$$

- 当近似共线时,  $0 < r^2 < 1$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{i1}^2} \cdot \frac{1}{1-r^2} > \frac{\sigma^2}{\sum x_{i1}^2}$$

- 当完全共线性时,  $r^2 = 1$

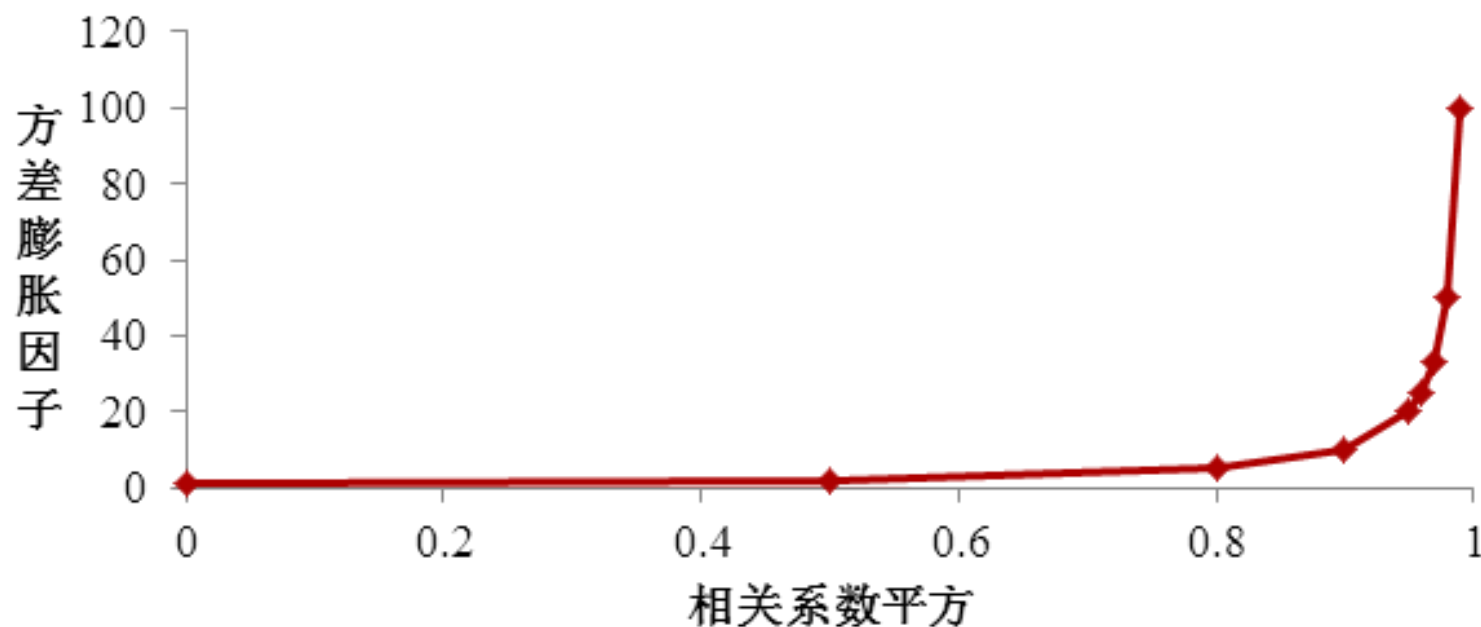
$$Var(\hat{\beta}_1) = +\infty$$

即多重共线性使参数估计量的方差增大, 其方差膨胀因子为

$$VIF(\hat{\beta}_1) = \frac{1}{1-r^2}$$

$$\text{方差膨胀因子} : VIF(\hat{\beta}_1) = \frac{1}{1-r^2}$$

相关系数平方	0	0.5	0.8	0.9	0.95	0.96	0.97	0.98	0.99	0.999
方差膨胀因子	1	2	5	10	20	25	33	50	100	1000



### （三）参数估计量的经济意义不合理。

如果模型中两个解释变量具有线性相关性，例如：  
 $X_2 = \lambda X_1$ ， $X_1$ 和 $X_2$ 前的系数 $\beta_1$ 、 $\beta_2$ 并不反映各自与被解释变量之间的**结构关系**，而是反映它们对被解释变量的**共同影响**。

参数 $\beta_1$ 、 $\beta_2$ 失去了应有的经济意义，经常表现出似乎反常的现象，例如： $\beta_1$ 本来应该是正数，估计结果却是负数。

#### （四）变量的显著性检验和模型的预测功能失去意义。

参数估计值的方差和标准差变大，从而 $t$ 值小于临界值，容易得到参数为0的判定。因而将重要的变量排除于模型之外。

$$t = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}}$$

## 四、多重共线性的检验

**检验任务：**1.检验多重共线性是否存在；

2.判断多重共线性存在的范围。

**检验方法：**主要为统计学方法，如判断系数检验，逐步

回归检验法等。

## (一) 检验多重共线性是否存在

### 1.二元模型中，采用简单相关系数法。

求出 $X_1$ 和 $X_2$ 的简单相关系数 $r$ ，若 $|r|$ 接近1，则说明二者存在较强的多重共线性。

$$r = \frac{\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{\sqrt{\sum(X_1 - \bar{X}_1)^2 \sum(X_2 - \bar{X}_2)^2}}$$



## 2.多元模型中，采用综合统计检验法。

在 $OLS$ 下，模型的 $R^2$ 与 $F$ 值较大，但参数 $t$ 值较小，说明各解释变量对 $Y$ 的联合线性作用显著，但是各解释变量间存在共线性而使它们对 $Y$ 的独立作用不能分辨，故 $t$ 检验不显著。

### 3.经济意义检验法。

回归参数估计值符号如不符合经济理论，模型有可能存在多重共线性。

### 4.回归参数检验法。

增加或减少解释变量个数时，回归参数估计值变化很大，说明模型有可能存在多重共线性。

---

## (二) 判断多重共线性存在范围

如果已经判定模型存在多重共线性，就需要进一步确定多重共线性究竟是由哪些变量引起的。

### 1.判定系数检验法

**第一步：**对模型中每个 $X_j$ 分别以其余解释变量为解释变量进行回归，并计算相应的拟合优度。

如果其中的某个回归

$$X_{ij} = \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_L X_{iL} + v_i$$

其判定系数较大，则说明 $X_j$ 与其他解释变量之间存在着共线性。

**第二步：**对判定系数较大的回归方程做 $F$ 检验，通过计算 $F$ 值与临界值进行比较判断

$$F = \frac{R_{j.}^2 / (k - 1)}{(1 - R_{j.}^2) / (n - k)} \sim F(k - 1, n - k)$$

$H_0$ ： $X_j$ 与其他解释变量间不存在显著线性相关关系，即

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_L = 0。$$

**该检验的等价检验：**在模型中排除某个解释变量 $X_j$ ，估计模型，如果拟合优度与包含 $X_j$ 时接近，就说明 $X_j$ 与其他解释变量之间存在共线性。

## 2.逐步回归法

**原理：**以 $Y$ 为被解释变量，逐个引入解释变量构成回归模型，估计模型，根据经济意义、 $t$ 值和 $R^2$ 的变化来决定新引入的解释变量是否独立。

- 如果判定系数变化显著，则说明新引入的解释变量是一个独立变量；
- 如果判定系数变化不显著，则说明新引入的解释变量与其它解释变量间存在共线性。

## 逐步回归法具体步骤：

### (1) 找到初始（基础）回归

初始回归：对被解释变量 $Y_i$ 分别做关于 $X_j$ 的回归，其中 $R^2$ 最大的为初始回归（因为对 $Y$ 的解释能力最强）

【例】三元模型中，对 $Y_i$ 做关于 $X_j$ 的回归

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_{i1} + \mu_i \\ Y_i = \beta_0 + \beta_2 X_{i2} + \mu_i \\ Y_i = \beta_0 + \beta_3 X_{i3} + \mu_i \end{cases}$$

运用OLS进行估计，并按步骤进行

$$\begin{cases} \text{①经济意义检验} \\ \text{②}t\text{检验} \\ \text{③按}R^2\text{进行排序} \end{cases},$$

排序重要性为①>②>③，①②不合理要排后，最后考虑 $R^2$ 。

## (2) 逐步回归

找到基础回归后，引入下一排序变量（如上例排序为132）

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i3} + \mu_i$$

此时要求：①经济意义合理；

② $t$ 必须显著；

③ $R^2$ 有显著提升。

对上例而言，若 $X_3$ 达到所有要求，则不再引入 $X_2$ 。

【115】

## 五、多重共线性的克服

如果模型被检验证明存在多重共线性，则需要发展新的

方法估计模型，最常用的方法有以下几类：

- 第一类方法：排除引起共线性的解释变量
- 第二类方法：差分法
- 第三类方法：减小参数估计量的方差



## （一）排除引起共线性的解释变量（最常用）

- 利用逐步回归法，找出引起多重共线性的解释变量，将其排除。
- 这是克服多重共线性最有效的方法。
- 需要注意，排除共线变量以后，剩余解释变量的参数的经济意义和数值都相应地发生变化。

## (二) 差分法

**原理：**一般来说，增量间的线性关系总是弱于总量间的线性关系。

所以，对于时间序列数据，通常将直接的线性模型转化为差分形式进行估计。

**作用机理：**削弱变量间的比例关系。

即

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i$$

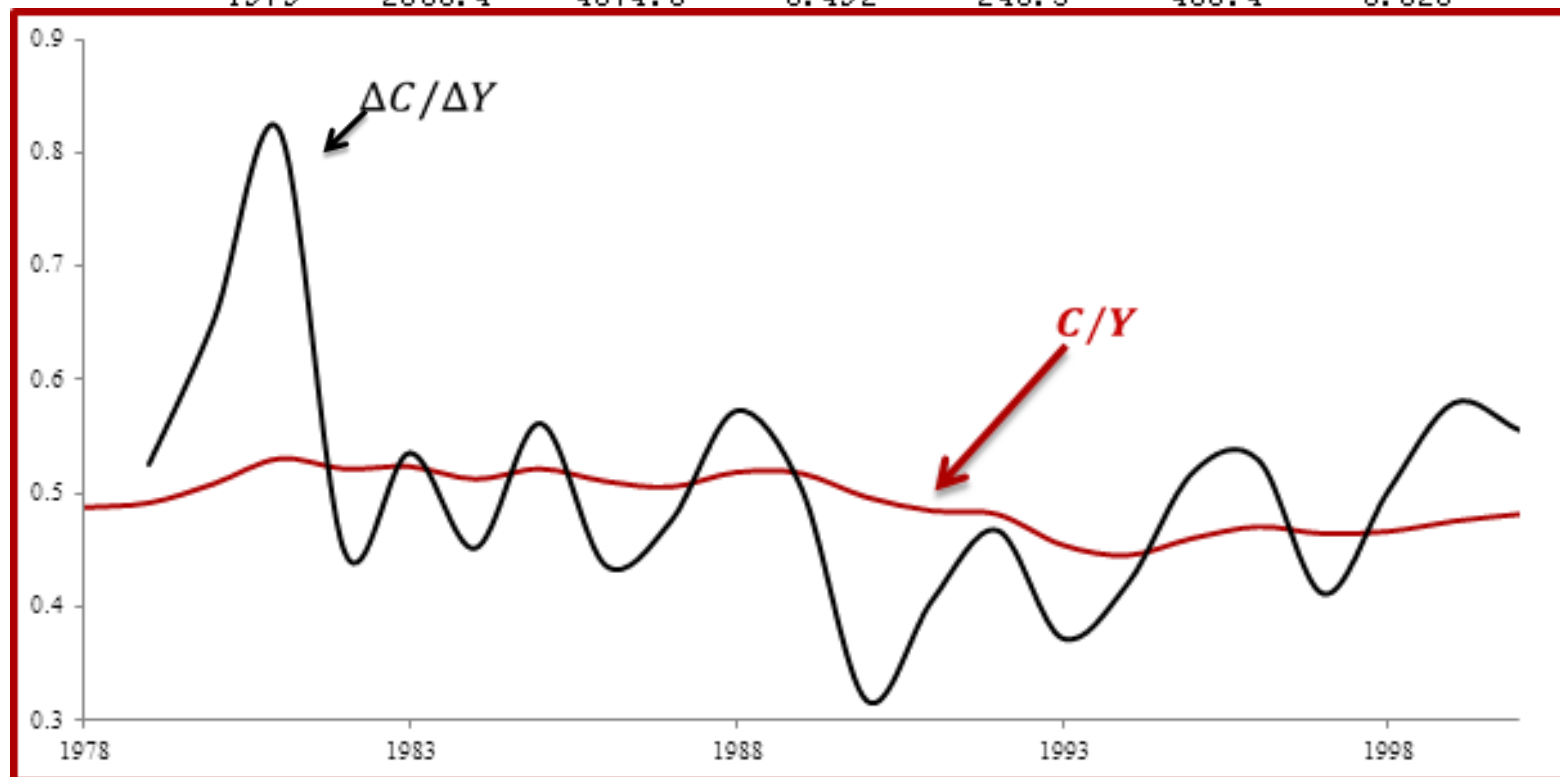
通过差分变换，转换为

$$\Delta Y_i = \beta_0 + \beta_1 \Delta X_{i1} + \beta_2 \Delta X_{i2} + \cdots + \beta_k \Delta X_{ik} + \Delta \mu_i$$

【例】

表 4.3.2 中国 GDP 与居民消费 C 的总量与增量数据 (亿元)

年份	C	Y	C/Y	$\Delta C$	$\Delta Y$	$\Delta C/\Delta Y$
1978	1759.1	3605.6	0.488			
1979	2005.4	4074.0	0.492	246.3	468.4	0.526



1997	34854.6	74894.2	0.465	2702.3	6563.8	0.412
1998	36921.1	79003.3	0.467	2066.5	4109.1	0.503
1999	39334.4	82673.1	0.476	2413.3	3669.8	0.658
2000	42911.9	89112.5	0.482	3577.5	6439.4	0.556

由表中的比值可以直观地看出，增量之间的线性关系弱于总量之间的线性关系。

进一步分析发现：

- $Y$ 与 $C$ 之间的相关系数为0.9845
- $\Delta Y$ 与 $\Delta C$ 之间的相关系数为0.9451
- 做 $Y$ 与 $C$ 的线性回归发现，其可决系数为0.9978
- 做 $\Delta Y$ 与 $\Delta C$ 的线性回归发现，其可决系数为0.8932

### (三) 减小参数估计量的方差

多重共线性的主要后果是参数估计量具有较大的方差，  
采取适当的方法减小参数估计量的方差，虽然不能消除  
模型的多重共线性，但能消除多重共线性造成的后果。

主要方法：①增加样本容量，减小参数估计值方差；  
②岭回归方法。

## ● 岭回归方法\*

1962年,A.E.Hoerl针对多重共线性的问题,提出了一种叫岭回归的回归估计方法,该方法以引入偏误为代价,减小参数估计量的方差。

### 1. 岭回归的基本原理

对线性模型

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i$$

引入矩阵 $D$ , 定义偏回归系数 $\beta$ 的岭估计为

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

$D$ 一般为主对角矩阵, 即

$$D = lI$$

其中 $l$ 为常数, 称为岭参数。

$\beta$ 的岭回归估计量为

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

- 当 $l = 0$ 时， $\hat{\beta}(l) = \hat{\beta}$ ，实际就是普通最小二乘估计。
- 当 $l > 0$ 时， $\hat{\beta}(l)$ 作为 $\beta$ 的估计量要比最小二乘估计量更为稳定，具有更小的方差。

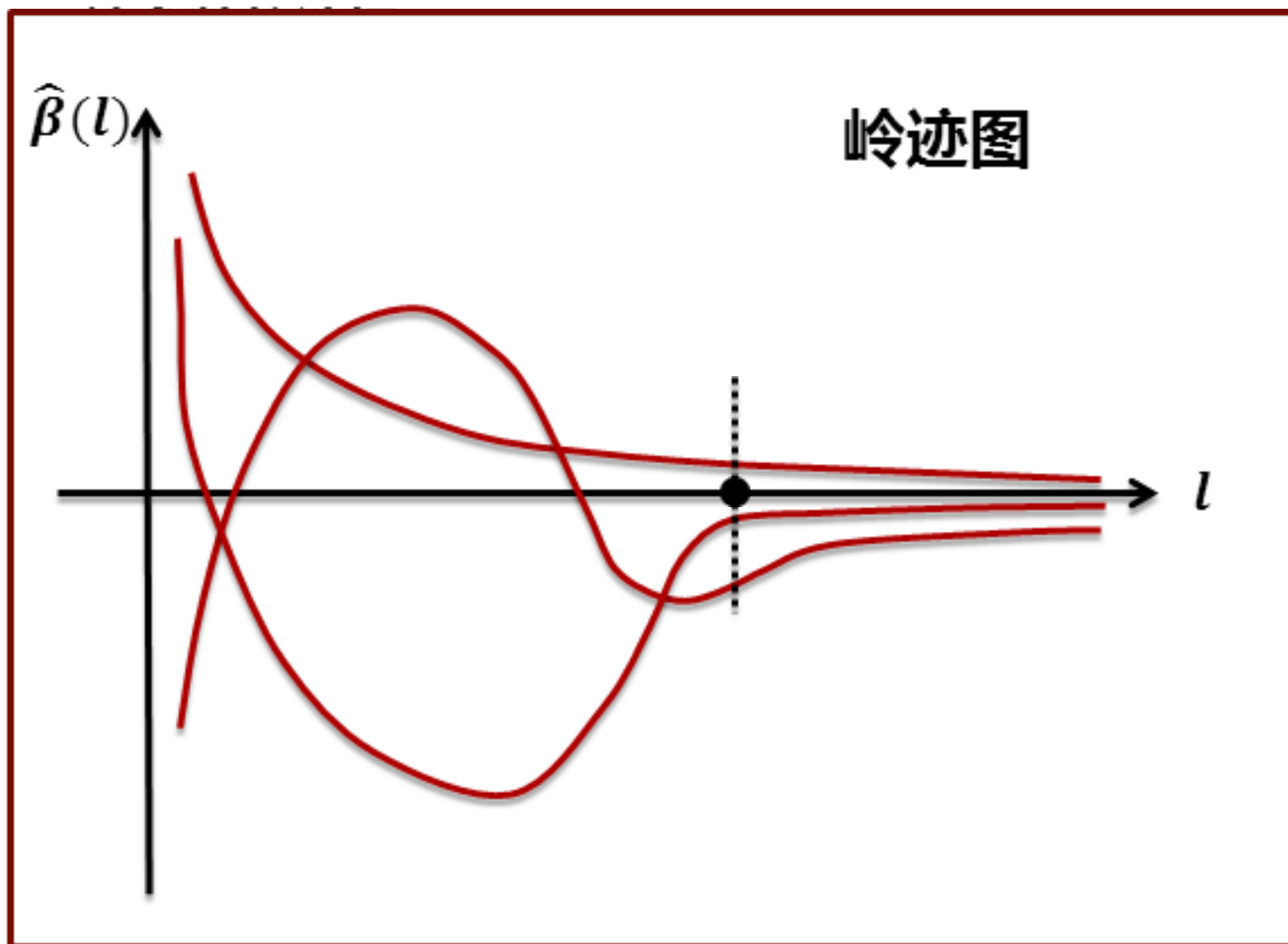
**【注意】**  $l$  一般只选择大于0的常数。

## 岭参数的特征：

- 当 $l$  较小时，回归系数很不稳定。
- 当 $l$  逐渐增大时，回归系数可能呈现稳定状态。

因此，要选择合适的 $l$  值，岭回归参数才会优于普通最小二乘参数。





## 多重共线性—案例分析

根据理论和经验分析，影响粮食产量( $Y$ )的主要因素有：

农业化肥施用量( $X_1$ )；粮食播种面积( $X_2$ )；受灾面积( $X_3$ )；农业机械总动力( $X_4$ )；农业劳动力( $X_5$ )。

已知我国粮食生产的相关数据，建立粮食生产函数：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \mu_i$$

表 4.3.3 中国粮食生产与相关投入资料

年份	粮食产量 $Y$ (万吨)	农业化肥施 用量 $X_1$ (万公斤)	粮食播种面 积 $X_2$ (千公顷)	受灾面积 $X_3$ (公顷)	农业机械总 动力 $X_4$ (万千瓦)	农业劳动 力 $X_5$ (万人)
1983	38728	1659.8	114047	16209.3	18022	31645.1
1984	40731	1739.8	112884	15264.0	19497	31685.0
1985	37911	1775.8	108845	22705.3	20913	30351.5
1986	39151	1930.6	110933	23656.0	22950	30467.0
1987	40208	1999.3	111268	20392.7	24836	30870.0
1988	39408	2141.5	110123	23944.7	26575	31455.7
1989	40755	2357.1	112205	24448.7	28067	32440.5
1990	44624	2590.3	113466	17819.3	28708	33330.4
1991	43529	2806.1	112314	27814.0	29389	34186.3
1992	44264	2930.2	110560	25894.7	30308	34037.0
1993	45649	3151.9	110509	23133.0	31817	33258.2
1994	44510	3317.9	109544	31383.0	33802	32690.3
1995	46662	3593.7	110060	22267.0	36118	32334.5
1996	50454	3827.9	112548	21233.0	38547	32260.4
1997	49417	3980.7	112912	30309.0	42016	32434.9
1998	51230	4083.7	113787	25181.0	45208	32626.4
1999	50839	4124.3	113161	26731.0	48996	32911.8
2000	46218	4146.4	108463	34374.0	52574	32797.5

## 1.估计模型

利用 $OLS$ 估计上述模型，得到

$$\hat{Y} = -12816.44 + 6.213X_1 + 0.421X_2 - 0.166X_3 - 0.098X_4 - 0.028X_5$$

(-0.91)      (8.39)      (3.32)      (-2.81)      (-1.45)      (-0.14)

$$R^2=0.9828 \quad \bar{R}^2 = 0.9756 \quad F = 137.11 \quad D.W. = 1.81$$

分析发现：

1. 调整的可决系数接近于1；
2. 给定 $\alpha = 5\%$ ，得 $F_{0.05}(5, 12) = 3.11$ ， $F = 137.11 > 3.11$ ，故认为上述粮食生产函数的总体线性关系显著成立。
3.  $X_4$ 、 $X_5$ 的参数未通过 $t$ 检验，且符号不正确，故解释变量间可能存在多重共线性。

## 2. 检验相关系数

列出 $X_1$ 、 $X_2$ 、 $X_3$ 、 $X_4$ 、 $X_5$ 的相关系数矩阵：

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
$X_1$	1.00	0.01	0.64	0.96	0.55
$X_2$	0.01	1.00	-0.45	-0.04	0.18
$X_3$	0.64	-0.45	1.00	0.69	0.36
$X_4$	0.96	-0.04	0.69	1.00	0.45
$X_5$	0.55	0.18	0.36	0.45	1.00

### 3.找到初始回归

分别作 $Y$ 关于 $X_1$ 、 $X_2$ 、 $X_3$ 、 $X_4$ 、 $X_5$ 的回归，根据估计结果的拟合优度选择初始的回归模型。

结果：选择 $Y$ 对 $X_1$ 的回归模型作为初始回归模型。

$$Y_i = \beta_0 + \beta_1 X_{i1} + \mu_i$$

## 4.逐步回归

将其他解释变量分别导入上述初始回归模型，寻找最佳回归方程。

	C	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$R^2$
$Y = f(X_1)$	30868	4.23					0.8852
t值	25.58	11.49					
$Y = f(X_1, X_2)$	-43871	4.65	0.67				0.9558
t值	-3.02	18.47	5.16				
$Y = f(X_1, X_2, X_3)$	-11978	5.26	0.41	-0.19			0.9752
t值	0.85	19.6	3.35	-3.57			
$Y = f(X_1, X_2, X_3, X_4)$	-13056	6.17	0.42	-0.17	-0.09		0.9775
t值	-0.97	9.61	3.57	-3.09	-1.55		
$Y = f(X_1, X_2, X_3, X_5)$	-12690	5.22	0.40	-0.20		0.07	0.9798
t值	-0.87	17.85	3.02	-3.47		0.37	

不显著，删除

## 5. 结论

视频片段



3