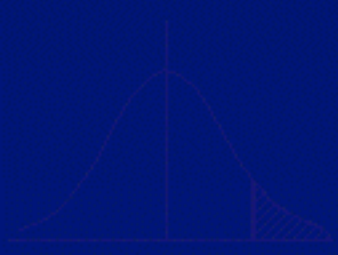
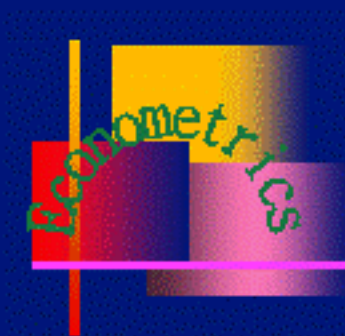


# 计量经济学

## 第四章

### 多重共线性





# 多重共线性

- 第一节 什么是多重共线性
- 第二节 多重共线性产生的后果
- 第三节 多重共线性的检验
- 第四节 多重共线性的补救措施
- 第五节 案例分析





# 第一节 什么是多重共线性

## 一、多重共线性的含义

在计量经济学中所谓的多重共线性(Multi-Collinearity), 不仅包括完全的多重共线性, 还包括不完全的多重共线性。

对于解释变量  $X_2, X_3, \dots, X_k$  如果存在不全为0的数  $\lambda_1, \lambda_2, \dots, \lambda_k$ , 使得

$$\lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_k X_{ki} = 0 \quad i = 1, 2, \dots, n$$

则称解释变量  $X_2, X_3, \dots, X_k$  之间存在着完全的多重共线性。

当  $\text{Rank}(X) < k$  时, 表明在数据矩阵  $X$  中, 至少有一个列向量可以用其余的列向量线性表示, 则说明存在完全的多重共线性。



# 不完全的多重共线性

实际中，常见的情形是解释变量之间存在不完全的多重共线性。

对于解释变量  $X_2, X_3, \dots, X_k$ ，存在不全为0的数

$\lambda_1, \lambda_2, \dots, \lambda_k$ ，使得

$$\lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_k X_{ki} + u_i = 0 \quad i = 1, 2, \dots, n$$

其中， $u_i$  为随机变量。这表明解释变量

$X_2, X_3, \dots, X_k$  只是一种近似的线性关系。



# 回归模型中解释变量的关系

可能表现为三种情形：

(1)  $r_{x_i x_j} = 0$ ，解释变量间毫无线性关系，变量间相互正交。这时已不需要作多元回归，每个参数 $\beta_j$ 都可以通过 $Y$ 对 $X_j$ 的一元回归来估计。

(2)  $r_{x_i x_j} = 1$ ，解释变量间完全共线性。此时模型参数将无法确定。

(3)  $0 < r_{x_i x_j} < 1$ ，解释变量间存在一定程度的线性关系。实际中常遇到的情形。





## 二、产生多重共线性的背景

多重共线性产生的经济背景主要有几种情形：

- 1、经济变量之间具有共同变化趋势。
- 2、模型中包含滞后变量。
- 3、利用截面数据建立模型也可能出现多重共线性。
- 4、样本数据自身的原因。



## 第二节 多重共线性产生的后果

### 一、完全多重共线性产生的后果

#### 1、参数的估计值不确定

当解释变量完全线性相关时——OLS估计式不确定

▲ 从偏回归系数意义看：在  $X_2$  和  $X_3$  完全共线性时，无法保持  $X_3$  不变，去单独考虑  $X_2$  对  $Y$  的影响（ $X_2$  和  $X_3$  的影响不可区分）

▲ 从OLS估计式看：可以证明此时 
$$\hat{\beta}_2 = \frac{0}{0}$$

#### 2、参数估计值的方差无限大

OLS估计式的方差成为无穷大

$$\text{Var}(\hat{\beta}_2) = \infty$$



## 二、不完全多重共线性下产生的后果

如果模型中存在不完全的多重共线性，可以得到参数的估计值，但是对计量经济分析可能会产生一系列的影响。

### 1、参数估计值的方差增大

$$\text{Var}(\hat{\beta}_2) = \sigma^2 \frac{1}{\sum x_{2i}^2 (1 - r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2} \frac{1}{(1 - r_{23}^2)}$$

当  $r_{23}$  增大时  $\text{Var}(\hat{\beta}_2)$  也增大

- 2、对参数区间估计时，置信区间趋于变大
- 3、假设检验容易作出错误的判断
- 4、可能造成可决系数较高，但对各个参数单独的 t 检验却可能不显著，甚至可能使估计的回归系数符号相反，得出完全错误的结论。



## 第三节 多重共线性的检验

### 一、简单相关系数检验法

**含义：**简单相关系数检验法是利用解释变量之间的线性相关程度去判断是否存在严重多重共线性的一种简便方法。

**判断规则：**一般而言，如果每两个解释变量的简单相关系数(零阶相关系数)比较高，例如大于0.8，则可认为存在着较严重的多重共线性。

**注意事项：**较高的简单相关系数只是多重共线性存在的充分条件，而不是必要条件。特别是在多于两个解释变量的回归模型中，有时较低的简单相关系数也可能存在多重共线性。因此并不能简单地依据相关系数进行多重共线性的准确判断。



## 二、方差扩大（膨胀）因子法

含义：

统计上可以证明，解释变量  $X_j$  的参数估计值  $\hat{\beta}_j$  的方差可表示为

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \cdot \frac{1}{1-R_j^2} = \frac{\sigma^2}{\sum x_j^2} \cdot VIF_j$$

其中的  $VIF_j$  是变量  $X_j$  的方差扩大因子 (Variance Inflation Factor)，即  $VIF_j = \frac{1}{(1-R_j^2)}$

其中  $R_j^2$  是多个解释变量辅助回归的可决系数



方差膨胀因子越大，表明解释变量之间的多重共线性越严重。反过来，方差膨胀因子越接近于1，多重共线性越弱。

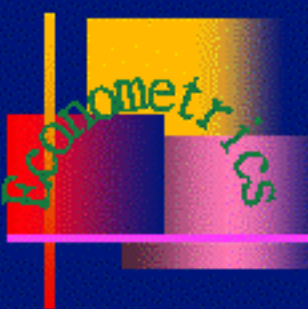
经验表明，方差膨胀因子 $\geq 10$ 时，说明解释变量与其余解释变量之间有严重的多重共线性，且这种多重共线性可能会过度地影响最小二乘估计。



### 三、直观判断法

- (1)当增加或剔除一个解释变量，或者改变一个观测值时，回归参数的估计值发生较大变化，回归方程可能存在严重的多重共线性。
- (2)从定性分析认为，一些重要的解释变量的回归系数的标准误差较大，在回归方程中没有通过显著性检验时，可初步判断可能存在严重的多重共线性。
- (3)有些解释变量的回归系数所带正负号与定性分析结果违背时，很可能存在多重共线性。
- (4)解释变量的相关矩阵中，自变量之间的相关系数较大时，可能会存在多重共线性问题。





## 四、逐步回归检测法

### 逐步回归的基本思想

将变量逐个的引入模型，每引入一个解释变量后，都要进行F检验，并对已经选入的解释变量逐个进行t检验，当原来引入的解释变量由于后面解释变量的引入而变得不再显著时，则将其剔除。以确保每次引入新的变量之前回归方程中只包含显著的变量。

在逐步回归中，高度相关的解释变量，在引入时会被剔除。因而也是一种检测多重共线性的有效方法。



## 第四节 多重共线性的补救措施

### 一、修正多重共线性的经验方法

#### 1、剔除变量法

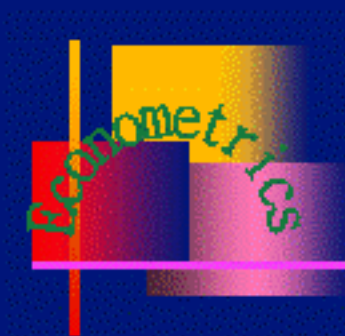
把方差扩大因子最大者所对应的自变量首先剔除再重新建立回归方程，直至回归方程中不再存在严重的多重共线性。

可能引起模型的设定误差。

#### 2、增大样本容量

如果样本容量增加，会减小回归参数的方差，标准误差也同样会减小。因此尽可能地收集足够多的数据可以改进模型参数的估计。





### 3、变换模型形式

一般而言，差分后变量之间的相关性要比差分前弱得多，所以差分后的模型可以有效地降低出现共线性的可能性，此时可直接估计差分方程。

差分会丢失一些信息，差分模型的误差项可能存在序列相关，可能会违背经典线性回归模型的相关假设，在具体运用时要慎重。



## 4、利用非样本先验信息

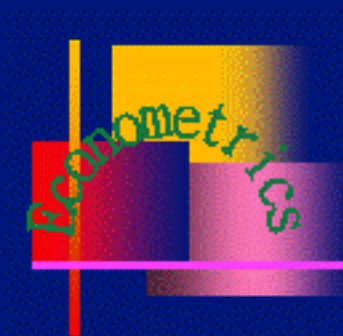
通过经济理论分析能够得到某些参数之间的线性关系，可以将这种线性关系作为约束条件，将此约束条件和样本信息结合起来进行约束最小二乘估计。

## 5、横截面数据与时序数据并用

首先利用横截面数据估计出部分参数，再利用时序数据估计出另外的部分参数，最后得到整个方程参数的估计。

包含假设：参数的横截面估计和从纯粹时间序列分析中得到的估计是一样的。





## 6、变量变换

变量变换的主要方法：

- (1)计算相对指标
- (2)将名义数据转换为实际数据
- (3)将小类指标合并成大类指标

变量数据的变换只是有时可得到较好的结果，但谁也无法保证一定可以得到很好的结果。



## 二、逐步回归法

- (1) 用被解释变量对每一个所考虑的解释变量做简单回归。
- (2) 以对被解释变量贡献最大的解释变量所对应的回归方程为基础，按对被解释变量贡献大小的顺序逐个引入其余的解释变量。

若新变量的引入改进了 $R^2$ 和F检验，且回归参数的 $t$ 检验在统计上也是显著的，则在模型中保留该变量。

若新变量的引入未能改进 $R^2$ 和F检验，且对其他回归参数估计值的 $t$ 检验也未带来什么影响，则认为该变量应该舍弃。

若新变量的引入未能改进 $R^2$ 和F检验，且显著地影响了其他回归参数估计值的数值或符号，同时本身的回归参数也通不过 $t$ 检验，说明出现了严重的多重共线性，应剔除该变量。



## 第五节 案例分析

### 一、研究的目的要求

提出研究的问题——为了规划中国未来国内旅游产业的发展，需要定量地分析影响中国国内旅游市场发展的主要因素。

### 二、模型设定及其估计

影响因素分析与确定——影响因素主要有国内旅游人数  $X_2$ ，城镇居民人均旅游支出  $X_3$ ，农村居民人均旅游支出  $X_4$ ，并以公路里程  $X_5$  和铁路里程  $X_6$  作为相关基础设施的代表

#### 理论模型的设定

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \beta_6 X_{6t} + u_t$$

其中： $Y_t$ ——第t年全国国内旅游收入



# 数据的收集与处理

年份	国内旅游收入Y (亿元)	国内旅游人数X2 (万人次)	城镇居民人均旅游支出X3 (元)	农村居民人均旅游支出X4 (元)	公路里程 X5 (万公里)	铁路里程X6 (万公里)
1994	1023.5	52400	414.7	54.9	111.78	5.90
1995	1375.7	62900	464.0	61.5	115.70	5.97
1996	1638.4	63900	534.1	70.5	118.58	6.49
1997	2112.7	64400	599.8	145.7	122.64	6.60
1998	2391.2	69450	607.0	197.0	127.85	6.64
1999	2831.9	71900	614.8	249.5	135.17	6.74
2000	3175.5	74400	678.6	226.6	140.27	6.87
2001	3522.4	78400	708.3	212.7	169.80	7.01
2002	3878.4	87800	739.7	209.1	176.52	7.19
2003	3442.3	87000	684.9	200.0	180.98	7.30

数据来源:《中国统计年鉴2004》



对模型进行OLS回归，结果如表4.3  
表4.3

Dependent Variable: Y  
Method: Least Squares  
Date: 07/18/05 Time: 18:16  
Sample: 1994 2003  
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-274.3773	1316.690	-0.208384	0.8451
X2	0.013088	0.012692	1.031172	0.3607
X3	5.438193	1.380395	3.939591	0.0170
X4	3.271773	0.944215	3.465073	0.0257
X5	12.98624	4.177929	3.108296	0.0359
X6	-563.1077	321.2830	-1.752685	0.1545

R-squared	0.995406	Mean dependent var	2539.200
Adjusted R-squared	0.989664	S.D. dependent var	985.0327
S.E. of regression	100.1433	Akaike info criterion	12.33479
Sum squared resid	40114.74	Schwarz criterion	12.51634
Log likelihood	-55.67396	F-statistic	173.3525
Durbin-Watson stat	2.311565	Prob(F-statistic)	0.000092

由此可见，该模型

$$R^2 = 0.9954$$

$$\bar{R}^2 = 0.9897$$

可决系数很高，F检验值  
173.3525，明显显著。

但是当  $\alpha = 0.05$  时

$$t_{\alpha/2}(n-k) = t_{0.025}(10-6) = 2.776$$

，不仅  $X_2$   $X_6$  系数的  
t检验不显著，而且

$X_6$  系数的符号与预期的相反，这表明很可能存在严重的多重共线性。



# 计算各解释变量的相关系数

	X2	X3	X4	X5	X6
X2	1.000000	0.918851	0.751960	0.947977	0.941681
X3	0.918851	1.000000	0.865145	0.859191	0.963313
X4	0.751960	0.865145	1.000000	0.664946	0.818137
X5	0.947977	0.859191	0.664946	1.000000	0.897708
X6	0.941681	0.963313	0.818137	0.897708	1.000000



### 三、消除多重共线性

采用逐步回归法检验和解决多重共线性问题。  
分别作Y对X2、X3、X4、X5、X6的一元回归

变量	X2	X3	X4	X5	X6
参数估计值	0.0842	9.0523	11.6673	34.3324	2014.146
t 统计量	8.6659	13.1598	5.1967	6.4675	8.7487
$R^2$	0.9037	0.9558	0.7715	0.8394	0.9054

$R^2$  的大小排序为：X3、X6、X2、X5、X4。

以X3为基础，顺次加入其他变量逐步回归



## 四、回归结果的解释与分析

最后消除多重共线性的结果

$$\hat{Y}_t = -2441.161 + 4.215884X_3 + 13.62909X_5 + 3.221965X_4$$

t = (3.944983)                      (4.692961)                      (3.06767)

$$R^2 = 0.991445 \quad \bar{R}^2 = 0.987186 \quad F=231.7935 \quad DW=1.952587$$

这说明，在其他因素不变的情况下，当城镇居民人均旅游支出  $X_3$  和农村居民人均旅游支出  $X_4$  分别增长1元时，国内旅游收入  $Y_t$  将分别增长4.21亿元和3.22 亿元。

在其他因素不变的情况下，作为旅游设施的代表，公路里程  $X_5$  每增加1万公里时，国内旅游收入  $Y_t$  将增长13.63亿元。



## 第四章小结

1、多重共线性是指各个解释变量 $X$ 之间有准确或近似准确的线性关系。

2、多重共线性的后果：

如果各个解释变量 $X$ 之间有完全的共线性，则它们的回归系数是不确定的，并且它们的方差会无穷大。

如果共线性是高度的但不完全的，回归系数可估计，但有较大的标准误差。回归系数不能准确地估计。

3、诊断共线性的经验方法：

(1) 表现为可决系数异常高而回归系数的 $t$ 检验不显著。

(2) 变量之间的零阶或简单相关系数。多个解释变量时，较低的零阶相关也可能出现多重共线性，需要检查偏相关系数。

(4) 如果 $R^2$ 高而偏相关系数低，则多重共线性是可能的。

(5) 用解释变量间辅助回归的可决系数判断。



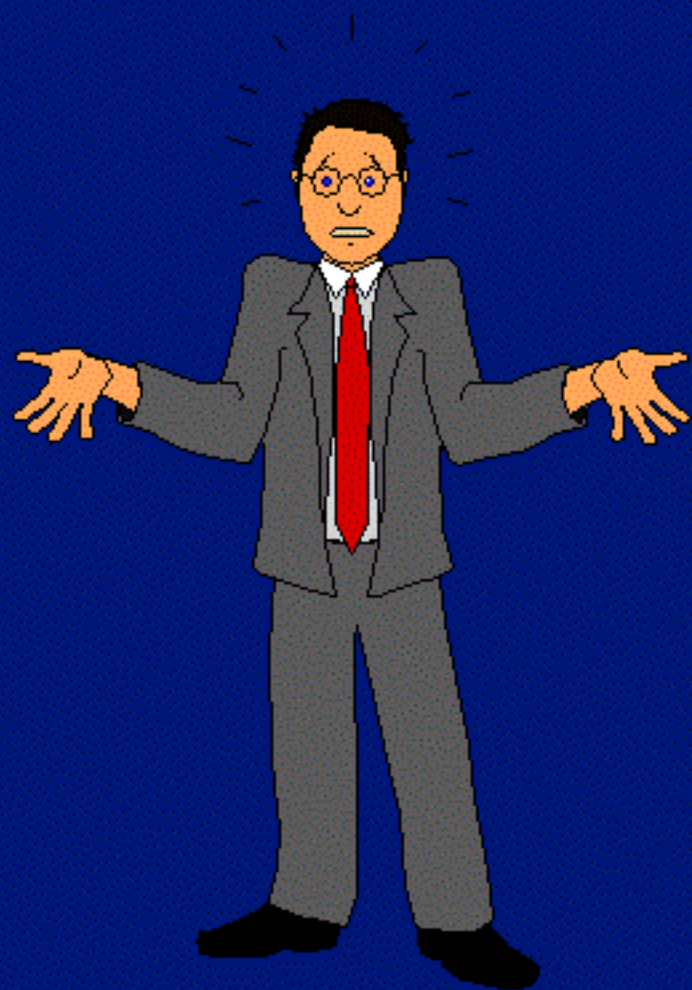
#### 4、降低多重共线性的经验方法有：

- (1)利用外部或先验信息；
- (2)横截面与时间序列数据并用；
- (3)剔除高度共线性的变量(如逐步回归)；
- (4)数据转换；
- (5)获取补充数据或新数据；
- (6)选择有偏估计量（如岭回归）。

经验方法的效果取决于数据的性质和共线性的严重程度。



# 第四章 结束了!



THANKS