Econometrics

计量经济学

第四章

多重共线性

Econometrics

引子：发展农业会减少财政收入吗？

为了分析各主要因素对财政收入的影响，建立财政收入模型：

$$CS_i = \beta_0 + \beta_1 NZ_i + \beta_2 GZ_i + \beta_3 JZZ_i + \beta_4 TPOP_i + \beta_5 CUM_i + \beta_6 SZM_i + u_i$$

其中：**CS**财政收入(亿元)；

NZ农业增加值(亿元)；　**GZ**工业增加值(亿元)；

JZZ建筑业增加值(亿元)；　**TPOP**总人口(万人)；

CUM最终消费(亿元)；　**SZM**受灾面积(万公顷)

数据样本时期**1978**年-**2007**年（资料来源：《中国统计年鉴**2008**》，中国统计出版社**2008**年版）

采用普通最小二乘法得到以下估计结果

Econometrics

财政收入模型的EViews估计结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
农业增加值	-1.907548	0.342045	-5.576888	0.0000
工业增加值建	0.045947	0.042746	1.074892	0.2936
筑业增加值	6.458374	0.765767	8.433867	0.0000
总人口	0.096022	0.091660	1.047591	0.3057
最终消费	0.003108	0.042807	0.072609	0.9427
受灾面积	-0.027627	0.048904	-0.564916	0.5776
截距	-5432.507	8607.753	-0.631118	0.5342

R-squared	0.989654	Mean dependent var	10049.04
Adjusted R-squared	0.986955	S.D. dependent var	12585.51
S.E. of regression	1437.448	Akaike info criterion	17.58009
Sum squared resid	47523916	Schwarz criterion	17.90704
Log likelihood	-256.7013	F-statistic	366.6801
Durbin-Watson stat	1.654140	Prob(F-statistic)	0.000000

Econometrics

模型估计与检验结果分析

- **可决系数为0.9897，校正的可决系数为0.9870，模型拟合很好。模型对财政收入的解释程度高达98.9%。**
- **F统计量为366.68，说明0.05水平下回归方程整体上显著。**
- **t 检验结果表明，除了农业增加值、建筑业增加值以外，其他因素对财政收入的影响均不显著。 ●农业增加值的回归系数是负数。**

农业的发展反而会使财政收入减少吗？！

这样的异常结果显然与理论分析和实践经验不相符。

若模型设定和数据真实性没问题，问题出在哪里呢？

Econometrics

第四章 多重共线性

本章讨论四个问题：

- 什么是多重共线性
- 多重共线性产生的后果
- 多重共线性的检验
- 多重共线性的补救措施

Econometrics

第一节 什么是多重共线性

本节基本内容:

- 多重共线性的含义
- 产生多重共线性的背景

Econometrics

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。在有截距项的模型中，截距项可以视为其对应的解释变量总是为**1**。对于解释变量 $1, X_2, X_3, \cdots, X_k$，如果存在不全为**0**的数 $\lambda_1, \lambda_2, \ldots \lambda_k$，使得

$$\lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \cdots + \lambda_k X_{ki} = 0 \qquad (i = 1, 2, \cdots, n)$$

则称解释变量 $1, X_2, X_3, \cdots, X_k$ 之间存在着完全的多重共线性。

或者说，当 $Rank(\boldsymbol{X}) < k$ 时，表明在数据矩阵 $\boldsymbol{X}$ 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。

Econometrics

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 $1, X_2, X_3, \cdots X_k$ ， 存在不全为**0**的数 $\lambda_1, \lambda_2, \cdots \lambda_k$ ， 使得

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

其中， u_i 为随机变量。这表明解释变量 $1, X_2, X_3, \cdots X_k$ 只是一种近似的线性关系。

在矩阵表示的线性回归模型

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

完全共线性指矩阵 $\boldsymbol{X}$的秩　$R(\boldsymbol{X}) < k$　即

$$|\boldsymbol{X'X}| = 0$$

近似共线性意味着

$$|\boldsymbol{X'X}| \approx 0$$

Econometrics

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β_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。

Econometrics

第二节 多重共线性产生的后果

本节基本内容:

- 完全多重共线性产生的后果
- 不完全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：$\mathrm{Var}(\hat{\beta}_2) = \infty$

- 如果解释变量存在完全共线性，则模型的参数β无法估计；

多元回归模型

$$Y = X\beta + \mu$$

的**OLS**估计量为

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

如果出现完全共线性，则 $(X'X)^{-1}$不存在，无法得到参数 β 的估计量。

二、不完全多重共线性产生的后果

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。

参数估计值的方差增大

如果解释变量之间存在近似共线性，则参数**OLS**估计量的方差随着多重共线程度的提高而增加；

在近似共线性下，虽然可以由式（**5-5**）得到参数**OLS**估计量，但

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

由于此时 $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$ ，引起 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 主对角线元素较大，且随着 $|\boldsymbol{X}'\boldsymbol{X}|$ 逼近于0而增大。这就使得参数估计量的方差增大，从而不能对总体参数做出准确推断。

例：

以二元回归模型 $Y=\beta_1+\beta_2X_2+\beta_3X_3+\mu$ 为例，$\hat{\beta}_2$ 的方差为

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

$$=\frac{\sigma^2\Big/\sum x_{2i}^2}{1-\dfrac{(\sum x_{2i}x_{3i})^2}{\sum x_{2i}^2\sum x_{3i}^2}}=\frac{\sigma^2}{\sum x_{2i}^2}\cdot\frac{1}{1-r_{23}^2}$$

其中 $r_{23}^2=(\sum x_{2i}x_{3i})^2\Big/\sum x_{2i}^2\sum x_{3i}^2$ 是X_1与X_2线性相关系数的平方，$r^2\leqslant 1$。

当X_2与X_3线性无关时，	当X_2与X_3 近似共线时，
$r_{23}^2=0,\quad Var(\hat{\beta}_2)=\dfrac{\sigma^2}{\sum x_{2i}^2}$	$0<r<1,\quad Var(\hat{\beta}_2)=\dfrac{\sigma^2}{\sum x_{2i}^2}\cdot\dfrac{1}{1-r_{23}^2}>\dfrac{\sigma^2}{\sum x_{2i}^2}$

可以看出，$|r|$ 越大，$Var(\hat{\beta}_2)$越大，多重共线性使得参数估计量方差增大，称 $\dfrac{1}{1-r^2}$ 为方差膨胀因子。其增大趋势如下表所示。

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

当完全共线性时，

$$r_{23}^2=1,\quad Var\left(\hat{\beta}_2\right)=\infty$$

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

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.可能造成参数估计量经济意义不合理

Econometrics

第三节 多重共线性的检验

本节基本内容：

- 简单相关系数检验法
- 方差膨胀因子法
- 直观判断法
- 逐步回归法
- 行列式检验法

Econometrics

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。

Econometrics

二、方差扩大（膨胀）因子法

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

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

其中的 VIF_j 是变量 X_j 的方差扩大因子

(Variance Inflation Factor)，即 $\mathrm{VIF}_j=\dfrac{1}{\left(1-R_j^2\right)}$

其中 R_j^2 是多个解释变量辅助回归的可决系数

经验规则

- 方差膨胀因子越大，表明解释变量之间的多重共性越严重。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**≥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。

Econometrics

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。

Econometrics

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。

Econometrics

五、行列式检验法

由于回归模型参数估计量的方差—协方差矩阵为

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

而

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{|\boldsymbol{X}'\boldsymbol{X}|}(\boldsymbol{X}'\boldsymbol{X})^{*}$$

所以

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 \frac{1}{|\boldsymbol{X}'\boldsymbol{X}|}(\boldsymbol{X}'\boldsymbol{X})^{*}$$

说明：
$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 \frac{1}{|\boldsymbol{X}'\boldsymbol{X}|}(\boldsymbol{X}'\boldsymbol{X})^*$$

(1) 当$|\boldsymbol{X}'\boldsymbol{X}|$较大时，$Var(\hat{\beta}_j)$ 较小

说明参数估计的精度较高，因而多重共线性不严重。

(2) 当$|\boldsymbol{X}'\boldsymbol{X}|$ 较小时，$Var(\hat{\beta}_j)$ 较大

说明参数估计的误差较大，因此表明模型的多重共线性严重。

(3) 当$|\boldsymbol{X}'\boldsymbol{X}|$ =0时，则 $Var(\hat{\beta}_j) \to \infty$

说明模型的解释变量之间完全相关，因而多重共线性最为严重，即存在完全多重共线性。

Econometrics

第四节 多重共线性的补救措施

本节基本内容：

- **修正多重共线性的经验方法**
- **逐步回归法**

岭回归法在本科教学中只是供选择使用的内容。

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。

注意：若剔除了重要变量，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样本数据可以改进模型参数的估计。

问题：增加样本数据在实际计量分析中常面临许多困难。

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能降低出现共线性的可能性，此时可直接估计差分方程。

问题：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关系，可以将这种关系作为约束条件，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。

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

Econometrics

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和 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 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，则认为该变量是多余变量。

若新变量的引入未能改进 R^2和 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。

Econometrics

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 并以铁路里程 X_5 作为相关基础设施的代表

理论模型的设定

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

其中： Y_t ——第 t 年全国国内旅游收入

Econometrics

数据的收集与处理

1994年—2011年中国旅游收入及相关数据

年份	国内旅游收入Y（亿元）	国内旅游人数X2（万人次）	城镇居民人均旅游花费X3（元）	农村居民人均旅游花费X4（元）	铁路里程X5（万公里）
1994	1023.5	52400	414.7	54.9	5.90
1995	1375.7	62900	464.0	61.5	5.97
1996	1638.4	63900	534.1	70.5	6.49
1997	2112.7	64400	599.8	145.7	6.60
1998	2391.2	69450	607.0	197.0	6.64
1999	2831.9	71900	614.8	249.5	6.74
2000	3175.5	74400	678.6	226.6	6.87
2001	3522.4	78400	708.3	212.7	7.01
2002	3878.4	87800	739.7	209.1	7.19
2003	3442.3	87000	684.9	200.0	7.30
2004	4710.7	110200	731.8	210.2	7.44
2005	5285.9	121200	737.1	227.6	7.54
2006	6229.74	139400	766.4	221.9	7.71
2007	7770.62	161000	906.9	222.5	7.80
2008	8749.3	171200	849.4	275.3	8.0

Econometrics

OLS估计的结果

Dependent Variable: Y
Method: Least Squares
Date: 08/24/13 Time: 11:01
Sample: 1994 2011
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	450.9799	3932.314	0.114686	0.9104
X2	0.073021	0.009533	7.659772	0.0000
X3	-6.655505	2.675543	-2.487534	0.0272
X4	14.15019	3.482846	4.062824	0.0013
X5	-230.9844	822.5258	-0.280823	0.7833

R-squared	0.985814	Mean dependent var	5567.064
Adjusted R-squared	0.981449	S.D. dependent var	4702.188
S.E. of regression	640.4485	Akaike info criterion	15.99235
Sum squared resid	5332266.	Schwarz criterion	16.23967
Log likelihood	-138.9311	F-statistic	225.8475
Durbin-Watson stat	1.297682	Prob(F-statistic)	0.000000

该模型 $R^2 = 0.9858$

$\bar{R}^2 = 0.9814$

可决系数较高，F检验值225.85,明显显著。

但是当 $\alpha = 0.05$ 时

$$t_{\alpha/2}(n-k) = t_{0.025}(18-5) = 2.16$$

不仅X5的系数不显著，而且X3、X5的符号与预期相反，这表明可能存在严重的多重共线性。

Econometrics

计算各解释变量的相关系数

	X2	X3	X4	X5
X2	1.000000	0.837135	0.846417	0.962193
X3	0.837135	1.000000	0.824165	0.902770
X4	0.846417	0.824165	1.000000	0.884279
X5	0.962193	0.902770	0.884279	1.000000

表明各解释变量间确实存在严重的线性关系

将每个解释变量分别作为被解释变量对其余的解释变量进行辅助回归，回归所得到的可决系数和方差扩大因子的数值见下表。

被解释变量	可决系数 R^2 的值	方差扩大因子 $VIF_j = \frac{1}{(1-R_j^2)}$
X2	0.9312	14.5349
X3	0.8310	5.9172
X4	0.7856	4.6642
X5	0.9618	26.1780

经验表明,方差扩大因子 **VIFj ≥ 10**时，通常说明该解释变量与其余解释变量之间有严重的多重共线性，这里**X2**、**X5**的方差扩大因子远大于**10**，表明存在严重多重共线性问题。

Econometrics

三、消除多重共线性

将各变量进行对数变换，再对以下模型进行估计

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{4t} + \beta_5 \ln X_{5t} + \varepsilon_t$$

将 Y_t、X_2、X_3、X_4、X_5 等数据取自然对数后，采用**OLS**方法估计模型参数，得到的回归结果

Dependent Variable: LNY
Method: Least Squares
Date: 08/24/13 Time: 11:30
Sample: 1994 2011
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8.440050	0.606171	-13.92355	0.0000
LNX2	0.916391	0.093951	9.753920	0.0000
LNX3	0.411645	0.139440	2.952138	0.0112
LNX4	0.289169	0.046084	6.274825	0.0000
LNX5	1.001875	0.422115	2.373463	0.0337

R-squared	0.997895	Mean dependent var	8.328844
Adjusted R-squared	0.997247	S.D. dependent var	0.792413
S.E. of regression	0.041574	Akaike info criterion	-3.292568
Sum squared resid	0.022469	Schwarz criterion	-3.045243
Log likelihood	34.63311	F-statistic	1540.781
Durbin-Watson stat	1.205065	Prob(F-statistic)	0.000000

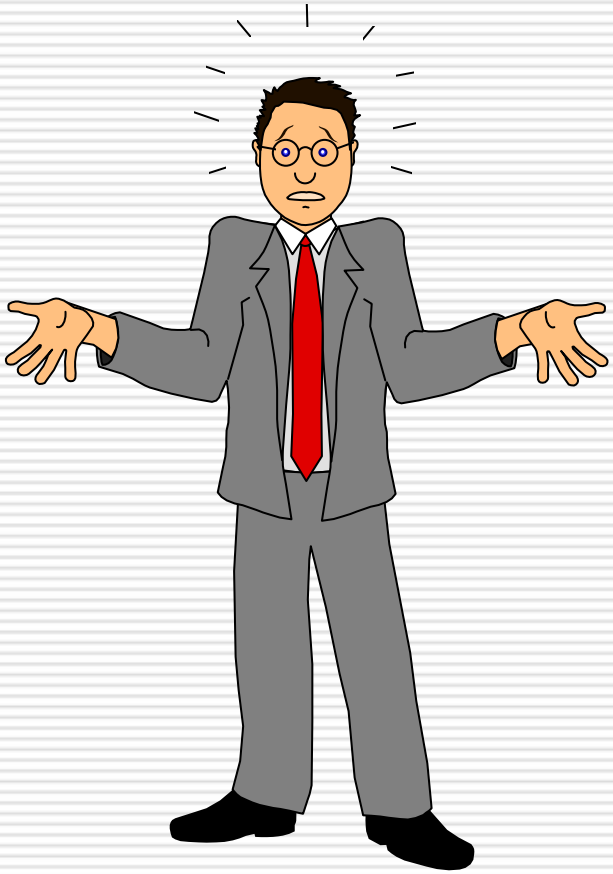
Econometrics

四、回归结果的解释与分析

最后消除多重共线性的结果

该模型 $R^2 = 0.9979, \bar{R}^2 = 0.9972$，可决系数很高，**F检验值1540.78**，明显显著。当 $\alpha = 0.05$时 $t_{\alpha/2}(n-k) = t_{0.025}(18-5) = 2.16$，所有系数估计值高度显著。

对系数估计值的解释：在其他变量保持不变的情况下，如果旅游人数每增加**1%**，则国内旅游收入平均增加**0.921%**；如果城镇居民旅游支出每增加**1%**，则国内旅游收入平均增加**0.41%**；如果农村居民旅游支出每增加**1%**，则国内旅游收入平均增加**0.29%**；如果铁路里程每增加**1%**，则国内旅游收入平均增加**1%**。



第四章结束了！

THANKS