



电子科技大学 经济与管理学院

School of Management and Economics of UESTC

计量经济学

Econometrics



电子科技大学 经济与管理学院

School of Management and Economics of UESTC

第六讲 多重共线性

(教材第8章)

第五讲 多重共线性



主要内容

- ❖ 回顾：经典线性回归模型的基本假设
- ❖ 多重共线性的定义
- ❖ 多重共线性产生的后果
- ❖ 多重共线性的来源和诊断
- ❖ 多重共线性的补救措施

第五讲 多重共线性



回顾：OLS的基本假设

假设1：回归模型是线性的，模型设定无误且含有误差项

假设2：误差项总体均值为零 $E(\varepsilon_i)=0$

假设3：所有解释变量与误差项都不相关 $Cov(X_i, \varepsilon_i)=0$

假设4：误差项观测值互不相关（无序列相关性） $Cov(\varepsilon_i, \varepsilon_j)=0$

假设5：误差项具有同方差（不存在异方差性） $Var(\varepsilon_i)=\sigma^2$

假设6：任何一个解释变量都不是其他解释变量的完全线性函数（不存在完全多重共线性）

第五讲 多重共线性



引例

❖ 回顾收入-消费问题

- 个人的消费支出不仅受个人收入的影响，也可能受财富、消费习惯的影响
- 建立消费对收入、财富的回归模型

$$Y_i = \beta_0 + \beta_1 X_{2i} + \beta_2 X_{3i} + \varepsilon_i$$

其中， Y_i 表示消费， X_{2i} 表示收入， X_{3i} 表示财富。

第五讲 多重共线性



引例

消费Y	收入X2	财富X3
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2052
140	220	2201
155	240	2435
150	260	2686

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24.77473	6.752500	3.668972	0.0080
X2	0.941537	0.822898	1.144172	0.2902
X3	-0.042435	0.080664	-0.526062	0.6151
R-squared	0.963504	Mean dependent var		111.0000
Adjusted R-squared	0.953077	S.D. dependent var		31.42893
S.E. of regression	6.808041	Akaike info criterion		6.917411
Sum squared resid	324.4459	Schwarz criterion		7.008186
Log likelihood	-31.58705	Hannan-Quinn criter.		6.817830
F-statistic	92.40196	Durbin-Watson stat		2.890614
Prob(F-statistic)	0.000009			

1. F 检验显著，单个系数的 t 检验不显著；
2. 财富变量X3的系数符号不符合预期。

将财富变量X3对收入变量X2做回归，结果如下：

$$X3 = \alpha + \beta X2 + \varepsilon$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.545455	29.47581	0.255988	0.8044
X2	10.19091	0.164262	62.04047	0.0000
R-squared	0.997926	Mean dependent var		1740.000
Adjusted R-squared	0.997667	S.D. dependent var		617.7312
S.E. of regression	29.83972	Akaike info criterion		9.806415
Sum squared resid	7123.273	Schwarz criterion		9.866932
Log likelihood	-47.03207	Hannan-Quinn criter.		9.740028
F-statistic	3849.020	Durbin-Watson stat		2.077534
Prob(F-statistic)	0.000000			

多重共线性

预示：财富与收入之间有显著的线性关系

第五讲 多重共线性

多重共线性的定义

❖ 多重共线性(multi-collinearity)的定义:

回归模型中的一些或全部解释变量之间存在一种完全或不完全的线性关系。

➤ 完全多重共线性

$$\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_k X_k = 0 \quad \text{等式成立,}$$

$\lambda_i \ (i = 1, \dots, k) \text{ 不全为零}$

➤ 不完全多重共线性

$$\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_k X_k + v_i = 0 \quad \text{等式成立,}$$

$\lambda_i \ (i = 1, \dots, k) \text{ 不全为零}$

v_i 为随机误差项

第五讲 多重共线性

多重共线性产生的后果

❖ 为什么要假设无多重共线性？

如果是**完全多重共线性**

若矩阵 $X'X$ 的逆不存在，下面的方程没有唯一解

$$X'X\beta = X'y$$

完全多重共线性只是一种极端的隐患，更常见的是出现不完全多重共线性。

第五讲 多重共线性

多重共线性产生的后果

如果是**不完全多重共线性**

矩阵 $X'X$ 的逆存在，下面的方程有唯一解

$$X'X\beta = X'y$$

且解为：

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

只要不是完全多重共线性，用OLS仍可得到参数的估计量及其标准误，并且仍是无偏；尽管无偏，但估计量的标准误非常大，即估计的精度很小。

第五讲 多重共线性

多重共线性产生的后果

❖ 思考题

- 为什么不完全多重共线性下估计量的标准误非常大？

$$\text{var}[\hat{\beta}] = \sigma^2 (X'X)^{-1}$$

- 多重共线性可能导致参数估计值的符号与预期符号不一致吗？
- 若总体中各解释变量X之间没有线性关系，样本中各解释变量X之间可能存在线性关系吗？

多重共线性本质上是一种样本现象

第五讲 多重共线性

➤ 以二元线性模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mu$ 为例:

$$\begin{aligned}\text{var}(\hat{\beta}_2) &= \frac{(\sum x_{3i}^2)}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} \sigma^2 \\ &= \frac{\sigma^2}{(\sum x_{2i}^2)(1 - r^2)} = \frac{\sigma^2}{(\sum x_{2i}^2)} * \boxed{\frac{1}{(1 - r^2)}} \\ \text{var}(\hat{\beta}_3) &= \frac{(\sum x_{2i}^2)}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} \sigma^2 \\ &= \frac{\sigma^2}{(\sum x_{3i}^2)(1 - r^2)} = \frac{\sigma^2}{(\sum x_{3i}^2)} * \boxed{\frac{1}{(1 - r^2)}}\end{aligned}$$

r 是变量 X_2 与 X_3 之间的相关系数。

➤ 由于 $r^2 \leq 1$, 故 $1/(1 - r^2) \geq 1$

➤ 当 $|r| \approx 1$ 时, $1/(1 - r^2)$ 很大 导致方差很大。

第五讲 多重共线性



➤例：对离差形式的二元回归模型10-24下次课

$$y = \beta_1 x_1 + \beta_2 x_2 + \mu$$

➤如果两个解释变量完全相关，如 $x_2 = \lambda x_1$ ，则

$$y = (\beta_1 + \lambda \beta_2) x_1 + \mu$$

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

$$\widehat{\beta_1 + \lambda \beta_2} = \sum x_{1i} y_i / \sum x_{1i}^2$$

第五讲 多重共线性

多重共线性产生的后果

❖ 不完全多重共线性的特征

- 偏回归系数的 t 值会降低，倾向于统计上不显著；
 - 估计量（偏回归系数）对模型设定的变化非常敏感，估计系数可能出现非预期的符号或难以置信的数值。
 - 虽然系数不显著，但总的
 - 可能出现每个偏回归系数
- 程的F值却很显著。

$$t_k = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)}$$

第五讲 多重共线性

多重共线性产生的后果

❖ 不完全多重共线性对预测的影响

- 如果回归分析的**唯一目的是预测**，并且如果**不完全共线性的结构在样本和未来都保持一致**，那么不完全多重共线性不是一个严重的问题。
- 如果不完全共线性的结构在未来发生变化，则预测是冒险的。

一个不完全多重共线性的例子

——消费对收入和财富的回归方程

Variable	Coefficient	Std. Error	t-Statistic	Prob.
	24.77473	6.752500	3.668972	0.0080
X2	0.941537	0.822898	1.144172	0.2902
	-0.042435	0.080664	-0.526062	0.6151
R-squared	0.963504	Mean dependent var		111.0000
Adjusted R-squared	0.953077	S.D. dependent var		31.42893
S.E. of regression	6.808041	Akaike info criterion		6.917411
Sum squared resid	324.4459	Schwarz criterion		7.008186
Log likelihood	-31.58705	Hannan-Quinn criter.		6.817830
F-statistic	92.40196	Durbin-Watson stat		2.890614
Prob(F-statistic)	0.000009			

符号

很大

不显著

高度显著

第五讲 多重共线性

多重共线性的来源

❖ 经济变量之间具有共同变化趋势

时间序列样本：经济繁荣时期，各基本经济变量（收入、消费、投资、价格）都趋于增长；衰退时期，又同时趋于下降。

❖ 模型中包含滞后变量

在经济计量模型中，往往需要引入滞后变量来反映真实的经济关系。

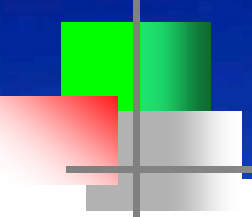
滞后变量：
前期收入

映真实的经

例如， $\text{消费} = f(\text{当期收入}, \text{前期收入})$

◆ 横截面数据之间存在相关性

生产函数中资本投入与劳动力投入往往出现高度相关情况，大企业二者都大，小企业都小



多重共线性的诊断

- ❖ 多重共线性是一个程度问题而不是有无的问题
- ❖ 以(经济)理论为基础，以经验(指标)为参考
 - 目前还没有一个被普遍接受的真正意义的检验多重共线性的统计量

第五讲 多重共线性

多重共线性的诊断

❖ 方法1：解释变量之间的相关系数

	X2	X3
X2	1.000000	0.998962
X3	0.998962	1.000000

- 相关系数多大才算大呢？
- 相关系数可以检验吗？
- 若解释变量多于两个，简单相关系数还合适吗？

第五讲 多重共线性

多重共线性的诊断

❖ 方法2：方差膨胀因子 (*VIF*)

- a) 把 X_i 对其他解释变量进行OLS回归，得到拟合系数 R_i^2
- b) 计算*VIF*: $VIF[\hat{\beta}_i] = (1 - R_i^2)^{-1}$
- c) 根据*VIF*判断，通常 $VIF > 5$ 被认为存在多重共线性

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
C	45.59625	9.837502	NA
X2	0.677162	4704.396	482.1275
X3	0.006507	4732.421	482.1275

第五讲 多重共线性

多重共线性的补救措施

❖ 剔除支配变量

- 支配变量(dominant variable):与被解释变量高度相关,以致于完全掩盖了方程中其他解释变量的影响(如销售量与销售额)

❖ 增加样本容量

- 样本越大,估计越精确

方差降低,
t检验的显著性增加

❖ 剔除多余的变量

- 潜在的理论假设作为剔除的主要依据

第五讲 多重共线性

多重共线性的补救措施

可支配收入

$$\hat{CO} = -367.83 + 0.51113Y_d + 0.0427LA$$

$$r_{Y_d, LA} = 0.986$$

(1.0307)

(0.0942)

学生的消费

$t = 0.496$

0.453

流动性资产

$\bar{R}^2 = 0.835$

$$\hat{CO} = -471.43 + 0.9714Y_d$$

(0.157)

$t = 6.187 \quad \bar{R}^2 = 0.861$

$$\hat{CO} = -199.44 + 0.088LA$$

(0.01443)

$t = 6.153 \quad \bar{R}^2 = 0.860$

多重共线性的补救措施

❖ 变换解释变量

- 一阶差分(时间序列分析)
- 两个变量相除
- 构造一个多重共线性的组合
 - ✓ 主成分分析法(principal components)
 - ✓ 因子分析法 (factor analysis)
-
- 参考：古扎拉蒂的计量经济学教材。

第五讲 多重共线性

差分法

针对时间序列数据、线性模型，

将原模型变换为**差分模型**：

$$\Delta Y_i = \beta_1 \Delta X_{1i} + \beta_2 \Delta X_{2i} + \dots + \beta_k \Delta X_{ki} + \Delta \mu_i$$

可以有效地消除原模型中的多重共线性。

一般讲，**增量**之间的线性关系远比总量之间的线性关系弱得多。

不好之处：可能出现误差项的自相关

比率变换法

针对时间序列数据、线性模型，

将原模型变换为比率模型：

$$\triangleright Y_i/x_{3i} = \beta_1 X_{1i}/x_{3i} + \beta_2 X_{2i}/x_{3i} + \beta_3 X_{3i}/x_{3i} + \mu_i/X_{3i}$$

不好之处：可能出现误差项的异方差性

第五讲 多重共线性

多重共线性的补救措施

❖ 无为而治: 什么也不做

- 剔除本应包含的解释变量会导致设定偏误；与遗漏变量造成的有偏估计相比较，较低的 t 统计值(显著性)似乎只是一个次要的问题
- 只有当后果很严重(参数估计值出现非预期的符号)，才应该采取其他补救措施。

第五讲 多重共线性

多重共线性的补救措施

饮料的平均价格

广告投入

$$\hat{S} = 3080 - 75000P_t + 4.23A_t - 1.04B_t$$

$$r_{A_t, B_t} = 0.974$$

$$t = -3.00 \quad 3.99 \quad -2.04$$

饮料的销售额

$$\bar{R}^2 = 0.825 \quad N = 28$$

竞争者的广告投入

$$\hat{S} = 2586 - 78000P_t + 0.52A_t$$

$$t = -3.25 \quad 0.12$$

$$\bar{R}^2 = 0.531 \quad N = 28$$

不用
处理

第五讲 多重共线性



本讲小结

- ❖ 多重共线性是违背了什么经典假设？
- ❖ 多重共线性的后果是什么？
- ❖ 怎样诊断多重共线性？
- ❖ 怎样补救多重共线性？
- ❖ 为什么“什么都不做”是对多重共线性最好的补救方法？

第五讲 多重共线性



作业

❖ 第8章作业：P148：习题2、6

第五讲 多重共线性

鱼肉/教皇案例

❖ 案例背景:

- 据推算，耶稣在星期五为人类受难钉死。为纪念耶稣苦难，天主教会规定，每星期五，凡年满十四岁之教友，该守小斋，即禁食热血动物的肉，但鱼、蛋及乳品不限(法典1251)。
- 1966年，教皇做出开戒决定，允许天主教徒在星期五斋戒日可吃猪肉。
- 教皇的决定会影响鱼的消费量吗？

第五讲 多重共线性

鱼肉/教皇案例

❖ 收集1966年前后的数据:

1946-1970

- 人均鱼肉消费量F
- 鱼肉价格(指数)PF
- 牛肉价格(指数)PB
- 人均可支配收入YD
- 虚拟变量P, 教皇公布
决定后取1, 其他取0
- 美国天主教人数N

表8-2 鱼肉/教皇案例的数据

年份	F	N	P	PB	PF	YD
1946	12.8	24 402	0	50.1	56	1 606
1947	12.3	25 268	0	71.3	64.3	1 513
1948	13.1	26 076	0	81	74.1	1 567
1949	12.9	26 718	0	76.2	74.5	1 547
1950	13.8	27 766	0	80.3	73.1	1 646
1951	13.2	28 635	0	91	83.4	1 657
1952	13.3	29 408	0	90.2	81.3	1 678
1953	13.6	30 425	0	84.2	78.2	1 726
1954	13.5	31 648	0	83.7	78.7	1 714
1955	12.9	32 576	0	77.1	77.1	1 795
1956	12.9	33 574	0	74.5	77	1 839
1957	12.8	34 564	0	82.8	78	1 844
1958	13.3	36 024	0	92.2	83.4	1 831
1959	13.7	39 505	0	88.8	84.9	1 881
1960	13.2	40 871	0	87.2	85	1 883
1961	13.7	42 105	0	88.3	86.9	1 909
1962	13.6	42 882	0	90.1	90.5	1 969
1963	13.7	43 847	0	88.7	90.3	2 015
1964	13.5	44 874	0	87.3	88.2	2 126
1965	13.9	45 640	0	93.9	90.8	2 239
1966	13.9	46 246	0	102.6	96.7	2 335
1967	13.6	46 864	1	100	100	2 403
1968	14	47 468	1	102.3	101.6	2 486
1969	14.2	47 873	1	111.4	107.2	2 534
1970	14.8	47 872	1	117.6	118	2 610

注: 表格数据文件名为FISH8。

资料来源: Historical Statistics of the U.S., Colonial Times to 1970(Washington, D.C.: U.S. Bureau of the Census, 1975).

第五讲 多重共线性

鱼肉/教皇案例

❖ 假定初步建立的方程为：

第t年鱼的人
均消费量

第t年鱼的价
格指数

第t年牛肉的
价格指数

第t年人均
可支配收入

$$F_t = \beta_0 + \beta_1 PF_t + \beta_2 PB_t + \beta_3 \ln Yd_t + \beta_4 P_t + \varepsilon_t$$

虚拟变量：1966年以后
取1，之前取0。

哪个是主要关心的解释变量？哪些是控制变量？

$$F_t = \beta_0 + \beta_1 PF_t + \beta_2 PB_t + \beta_3 \ln Yd_t + \beta_4 P_t + \varepsilon_t$$

Sample: 1946 1970

Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.961108	7.773354	1.024154	0.3180
PF	0.027993	0.028533	0.981075	0.3383
PB	0.004692	0.019336	0.242675	0.8107
LOG(YD)	0.360363	1.154974	0.312010	0.7583
P	-0.124462	0.257573	-0.483211	0.6342
R-squared	0.722869	Mean dependent var	13.44800	
Adjusted R-squared	<u>0.667443</u>	S.D. dependent var	0.533948	
S.E. of regression	0.307916	Akaike info criterion	0.658876	
Sum squared resid	1.896243	Schwarz criterion	0.902651	
Log likelihood	-3.235945	Hannan-Quinn criter.	0.726488	
F-statistic	13.04200	Durbin-Watson stat	2.236966	
Prob(F-statistic)	<u>0.000022</u>			

是否存在多重共线性呢？

各变量的相关系数表

	F	PF	PB	LOG(YD)	P
F	1.000000	0.847590	0.818532	0.780012	0.585630
PF	0.847590	1.000000	0.958096	0.915320	0.734643
PB	0.818532	<u>0.958096</u>	1.000000	0.814890	0.663162
LOG(YD)	0.780012	<u>0.915320</u>	<u>0.814890</u>	1.000000	0.744500
P	0.585630	0.734643	0.663162	0.744500	1.000000

可以剔除变量log(YD)吗？

可以剔除变量PF或PB吗？

变量转换：相对价格 $RP=PF/PB$

根据实证检验结果：教皇的决定影响鱼的消费了吗？

Sample: 1946 1970

Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.168676	4.832730	-1.069515	0.2970
PF/PB	-1.930897	1.430728	-1.349591	0.1915
LOG(YD)	2.711743	0.656781	4.128838	0.0005
P	0.005197	0.280080	0.018554	0.9854
R-squared	0.639721	Mean dependent var	13.44800	
Adjusted R-squared	0.588252	S.D. dependent var	0.533948	
S.E. of regression	0.342621	Akaike info criterion	0.841263	
Sum squared resid	2.465174	Schwarz criterion	1.036284	
Log likelihood	-6.515793	Hannan-Quinn criter.	0.895354	
F-statistic	12.42938	Durbin-Watson stat	1.597750	
Prob(F-statistic)	0.000069			

尽管仍然存在多重共线性，
但似乎不太严重了！

	F	N	P	PB	PF	YD
1946	12.80000	24402.00	0.000000	50.10000	56.00000	1606.000
1947	12.30000	25268.00	0.000000	71.30000	64.30000	1513.000
1948	13.10000	26076.00	0.000000	81.00000	74.10000	1567.000
1949	12.90000	26718.00	0.000000	76.20000	74.50000	1547.000
1950	13.80000	27766.00	0.000000	80.30000	73.10000	1646.000
1951	13.20000	28635.00	0.000000	91.00000	83.40000	1657.000
1952	13.30000	29408.00	0.000000	90.20000	81.30000	1678.000
1953	13.60000				78.20000	1726.000
1954	13.50000				78.70000	1714.000
1955	12.90000				77.10000	1795.000
1956	12.90000				77.00000	1839.000
1957	12.80000				78.00000	1844.000
1958	13.30000				83.40000	1831.000
1959	13.70000				84.90000	1881.000
1960	13.20000				85.00000	1883.000
1961	13.70000				86.90000	1909.000
1962	13.60000				90.50000	1969.000
1963	13.70000	43847.00	0.000000	88.70000	90.30000	2015.000
1964	13.50000	44874.00	0.000000	87.30000	88.20000	2126.000
1965	13.90000	45640.00	0.000000	93.90000	90.80000	2239.000
1966	13.90000	46246.00	0.000000	102.6000	96.70000	2335.000
1967	13.60000	46864.00	1.000000	100.0000	100.0000	2403.000
1968	14.00000	47468.00	1.000000	102.3000	101.6000	2486.000
1969	14.20000	47873.00	1.000000	111.4000	107.2000	2534.000
1970	14.80000	47872.00	1.000000	117.6000	118.0000	2610.000

数据收集
存在问题！