第四章 多重共线性

基本假定的回顾与分析

- **零均值假定**（主要影响截距项的估计）
- **同方差假定**
- **无自相关假定**
- **解释变量非随机，或虽随机但与 u 不相关**

在单一方程模型中，从重复抽样的角度一般是合理的。

在某些单一方程模型中和联立方程模型中可能会违反。

- **无多重共线性假定**
- **正态性假定**（不影响OLS估计是BLUE）
- 根据中心极限定理，样本容量无限增大时，OLS趋于正态分布
- **结论：**需要专门讨论无多重共线性、同方差、 无自相关

引子：发展农业和建筑业会减少财政收入吗？

为了分析各主要因素对国家财政收入的影响，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-2003年（资料来源：《中国统计年鉴2004》，中国统计出版社2004年版）

采用普通最小二乘法得到以下估计结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
农业增加值NZ	-1.535090	0.129778	-11.82861	0.0000
工业增加值GZ	0.898788	0.245466	3.661558	0.0017
建筑业增加值JZZ	-1.527089	1.206242	-1.265989	0.2208
总人口TPOP	0.151160	0.033759	4.477646	0.0003
最终消费CUM	0.101514	0.105329	0.963783	0.3473
受灾面积SZM	-0.036836	0.018460	-1.995382	0.0605
截距项	-11793.34	3191.096	-3.695704	0.0015

R-squared	**0.995015**	Mean dependent var	5897.824
Adjusted R-squared	**0.993441**	S.D. dependent var	5945.854
S.E. of regression	481.5380	Akaike info criterion	15.41665
Sum squared resid	4405699.	Schwarz criterion	15.75537
Log likelihood	-193.4165	**F-statistic**	**632.0999**
Durbin-Watson stat	1.873809	Prob(F-statistic)	0.000000

模型估计检验结果分析：

- 可决系数为0.995，校正的可决系数为0.993，模型拟合很好。模型对财政收入的解释程度高达99.5%。

F统计量为632.10，说明0.05水平下回归方程整体上显著

- t 检验结果表明，工业、农业增加值和总人口对财政收入影响显著，其他因素对财政收入的影响均不显著。
- 农业增加值和建筑业增加值的回归系数为负数，

农业和建筑业的发展反而会使财政收入减少吗？！

这样的结果显然与理论分析和实践经验不相符。

为什么会出现这样的异常结果呢？

如果模型设定和数据真实性没有问题，问题出在哪里呢？

从实例谈起

- **实例：消费支出与收入和财富的关系**
  - 根据经济理论，研究消费问题，假定消费与收入和财富有线性关系，建立计量经济模型。

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

Y_i：消费

X_{2i}：收入

X_{3i}：财富

从实例谈起

消费Y_i	收入X_2	财富X_3
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

Equation: UNTITLED Workfile: BOOK1\Book1

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 11/17/07 Time: 11:52
Sample: 1 10
Included observations: 10

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	F-statistic	92.40196
Durbin-Watson stat	2.890614	Prob(F-statistic)	0.000009

1. F检验显著，单个系数*t*检验不显著；

2. 财富变量*X*3的系数符号错误。

财富X3对收入X2做回归

Equation: UNTITLED Workfile: BOOK1\Book1

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: X3
Method: Least Squares
Date: 11/17/07 Time: 16:31
Sample: 1 10
Included observations: 10

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	F-statistic	3849.020
Durbin-Watson stat	2.077534	Prob(F-statistic)	0.000000

多重共线性

- 财富与收入之间有显著的线性关系

本章讨论四个问题：

- 多重共线性的实质与产生原因
- 多重共线性的后果
- 多重共线性的检测（判断）方法
- 多重共线性的补救方法

第一节　什么是多重共线性

一、多重共线性的概念

在多元回归模型中，各个解释变量之间可能存在一定的线性相关关系

完全线性关系

能找到不全为0的数 $\lambda_1,\lambda_2\cdots\lambda_k$，使得

$$\lambda_1+\lambda_2X_2+\lambda_3X_3+\cdots+\lambda_kX_K=0$$

不完全线性关系

$$\lambda_1+\lambda_2X_2+\lambda_3X_3+\cdots+\lambda_kX_K+v_i=0$$

完全无线性关系

（正交变量）

多重共线性——指解释变量间的线性关系，既包括完全的线性关系，又包括不完全的线性关系

注意：▲ 多重共线性不是有无问题，而是程度高低问题

▲ 多重共线性关注的是解释变量间的线性关系，而非相互之间的非线性关系

二、多重共线性产生的原因

（1）时间序列数据在时间上常有共同变动的趋势

如工业产值、商品零售额、固定资产投资、财政收入常有共同趋势

（2）模型中放入同一变量的滞后变量或不同形式

如：$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + u_t$　　$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + u$

（3）经济变量间本身具有内在联系

如截面数据中某行业企业的资本量、劳动投入等都与企业规模相关

（4）样本数据自身的原因（抽样有限）

本质上多重共线性就是一个样本问题

第二节　多重共线性产生的后果

从参数估计看，在完全无多重共线性时，各解释变量都独立地影响被解释变量，多元回归是否还有必要呢？

例如，对于 $Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+u_i$

X_2 与 X_3 完全不相关时，相关系数 $\gamma_{X_2X_3}=\dfrac{\sum x_{2i}x_{3i}}{\sqrt{\sum x_{2i}^2\sum x_{3i}^2}}=0$

即 $\sum x_{2i}x_{3i}=0$

此时

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

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

对比一元回归时

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

一、存在多重共线性时

——OLS估计式变得不确定或不精确

1. 解释变量完全线性相关时 **——OLS** 估计式不确定

- 从**OLS**估计式看：此时 $X_{3i}=\lambda X_{2i}$ 可以证明（见96页）

$$\hat{\beta}_2=\frac{(\sum y_i x_{2i})[\sum(\lambda x_{2i})^2]-[\sum y_i(\lambda x_{2i})][\sum x_{2i}(\lambda x_{2i})]}{(\sum x_{2i}^2)[\sum(\lambda x_{2i})^2]-[\sum x_{2i}(\lambda x_{2i})]^2}=\frac{0}{0}$$

同理 $\hat{\beta}_3=\frac{0}{0}$

- 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 的作用不可区分）

2. 解释变量不完全线性相关，但存在高度多重共线性时——回归系数虽可以确定，但方差会变得很大，**OLS**估计式不精确（下面讲）

多重共线性的后果

- **出现高度但不完全多重共线性时的估计：**
  - 完全多重共线性只是一种极端的隐患，在时间序列数据中经常会出现欠完全的线性关系

$$x_{3i} = \lambda x_{2i} + v_i, \lambda \neq 0, 并且 \sum x_{2i} v_i = 0$$

  - 这种情况下偏回归系数的估计可以实现

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

二、OLS估计式方差变得很大，标准误差增大

1. 当X_2和X_3完全线性相关时——OLS估计式的方差成为无穷大

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

（证明见P97）

2. 当X_2和X_3不完全线性相关时——OLS估计式的方差会增大，例如在有两个解释变量时，可证明(见P98)

$$Var(\hat{\beta}_2) = \sigma^2 \frac{1}{\sum x_{2i}^2 (1-r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2} \frac{1}{(1-r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2} \cdot VIF_2$$

当r_{23}增大时，VIF_2增大，$Var(\hat{\beta}_2)$也会增大，

思考：当$r_{23} \to 0$时$Var(\hat{\beta}_2) \to \sigma^2 / \sum x_{2i}^2$ (与一元回归比较)

多重共线性的后果

OLS估计量的大方差：

$$\mathrm{var}(\hat{\beta}_2)=\frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)} \qquad \mathrm{var}(\hat{\beta}_3)=\frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)}$$

- ✓ 方差增大的速度可由方差膨胀因子（Variance Inflation Factor）给出：

$$VIF=\frac{1}{1-r_{23}^2}$$

- ✓ 无共线性时VIF达到最小值1，完全共线性时VIF达到最大值无穷大

三、区间估计和假设检验会出现错误

1. 多重共线性严重时，对总体参数的置信区间趋于增大。因为

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

（共线性越严重，$Var(\hat{\beta}_2)$ 和 $SE(\hat{\beta}_2)$越大，置信区间也增大）

2. 严重多重共线时，假设检验作出错误判断的概率增大

因为 $t = \hat{\beta}_2 / \sqrt{Var(\hat{\beta}_2)}$，当方差变大时会使 t 绝对值减小，导致使本应否定的“参数为 0 ”的原假设被接受

可能造成参数的联合显著性很高（通过F检验），但各个参数单独的 t 检验却不显著

四、当多重共线性严重时，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

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

例如当 $X_{4t} = kX_{2t}$ 时，引入任意不为0的数 $\beta^* \neq 0$

模型变换 $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} - \beta^* X_{2t} + \beta^* X_{2t} + u_t$

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} - \beta^* X_{2t} + \beta^* \frac{X_{4t}}{k} + u_t$$

$$Y_t = \beta_1 + (\beta_2 - \beta^*) X_{2t} + \beta_3 X_{3t} + (\beta_4 + \frac{1}{k}\beta^*) X_{4t} + u_t$$

估计结果 $\hat{Y}_t = \hat{\beta}_1 + (\hat{\beta}_2 - \beta^*) X_{2t} + \hat{\beta}_3 X_{3t} + (\hat{\beta}_4 + \frac{1}{k}\beta^*) X_{4t}$

当 $\beta^* > \hat{\beta}_2$ 时，所估计的 X_2 的参数与真实 β_2 的符号可能相反。

分析多重共线性后果时应注意：

- 存在多重共线性时,OLS估计式还是最佳线性无偏估计式（BLUE）

理解：　无偏性是重复抽样的特性；

"最小方差"是相对于其他估计方法而言：

（相对于其他方法方差最小，并不是说相对于估计量的值就小）

“方差变大”是相对于无多重共线性而言

- 多重共线性的影响程度与解释变量在方程中的相对“地位”有关

- 如果研究目的仅在于预测Y，而解释变量X之间的多重共线性关系的性质在未来将继续保持（前提条件），这时多重共线性可能并不是严重问题，而应着重于可决系数高，F检验显著。

（理解：出现高度共线性时，虽然无法精确估计个别回归系数，但可精确估计这些系数的某些线性组合。）

多重共线性的后果

- 从另一个角度，由 $X_{3i}=\lambda X_{2i},\lambda\neq 0$有

$$
\begin{aligned}
y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3(\lambda x_{2i}) + \hat{\mu}_i \\
&= (\hat{\beta}_2 + \lambda\hat{\beta}_3)x_{2i} + \hat{\mu}_i \\
&= \hat{\alpha} x_{2i} + \hat{\mu}_i
\end{aligned}
$$

$$
\hat{\alpha} = \hat{\beta}_2 + \lambda\hat{\beta}_3 = \frac{\sum x_{2i}y_i}{\sum x_{2i}^2}
$$

- 虽然两偏回归系数的线性组合$\hat{\beta}_2 + \lambda\hat{\beta}_3$可以唯一地估计，但无法得到每个回归系数的唯一解

第三节　多重共线性的检验（判断是否严重）

一、利用解释变量之间的相关系数去判断

1. 只有两个解释变量时：用二者相关系数 r_{23} 判断

2. 两个以上解释变量时：可用两两变量的相关系数 r_{ij} 判断（K个变量可用相关系数矩阵）

例如

Correlation Matrix

	NZ	GZ	JZZ	TPOP	CUM	SZM
NZ	1.000000	0.969053	0.972087	0.966446	0.971170	0.237891
GZ	0.969053	1.000000	0.998744	0.989328	0.997229	0.284421
JZZ	0.972087	0.998744	1.000000	0.989777	0.995402	0.277182
TPOP	0.966446	0.989328	0.989777	1.000000	0.994551	0.326340
CUM	0.971170	0.997229	0.995402	0.994551	1.000000	0.276963
SZM	0.237891	0.284421	0.277182	0.326340	0.276963	1.000000

注意：简单相关系数只是多重共线性的充分条件，不是必要条件。在有多个解释变量时，较低的相关系数也可能存在较严重多重共线性

二、 直观判断法 （经验方法）

以下情况的出现提示可能存在较严重多重共线性：

(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

(2) 从定性分析认为一些是重要的解释变量，但其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

(3) 有些解释变量的回归系数的正负号与定性分析结果违背

(4) 可决系数较高，F检验显著，但偏回归系数的 t 检验不显著

三、利用解释变量之间的辅助回归及检验判断

辅助回归：逐次将每一个解释变量作为被解释 变量对其它解释变量进行回归

分别估计其参数、计算可决系数、作F检验

- 若辅助回归的F检验显著，认为该变量与其它变量可能存在较严重的多重共线性
- 若F检验不显著，认为该变量与其它变

量不存在严重的多重共线性

四、 方差扩大因子法（容许度）

多元线性回归模型中，可分别以每个解释变量为被解释变量，作与其他解释变量的辅助回归。以x_j为被解释变量作对其他解释变量的辅助线性回归的可决系数用R_j^2表示。

原回归方程中解释变量X_j的参数估计值$\hat{\beta}_j$的方差可表示为

（证明从略）

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

其中 $VIF_j=\frac{1}{(1-R_j^2)}$

其中的VIF_j是变量X_j所对应参数估计量的方差扩大因子，也称容许度。

对比

在只有两个解释变量时(如前面的讨论)

$$Var(\hat{\beta}_2)=\sigma^2\frac{1}{\sum x_{2i}^2(1-r_{23}^2)}=\frac{\sigma^2}{\sum x_{2i}^2}\frac{1}{(1-r_{23}^2)}=\frac{\sigma^2}{\sum x_{2i}^2}\cdot VIF_2$$

当有多个解释变量时，作 X_j 对其他解释变量的辅助回归，并计算可决系数 R_j^2 ，

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

注意： R_j^2 是多个解释变量辅助回归的多重可决系数，而相关系数 r_{23}^2 只是说明两个变量的线性关系 。

（一元回归中可决系数的数值等于相关系数的平方）

方差扩大因子的作用

由 $VIF_J = 1/(1-R_j^2)$

R_j^2 越大 ⟹ 多重共线性越严重 ⟹ VIF_j越大

VIF_j的大小可以反映解释变量之间存在多重共线性的严重程度。

优点：可从数量上判断多重共线性的程度

（给出了一种经验规则）

经验表明： $VIF_j \geq 10$时，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，且这种多重共线性可能会过度地影响最小二乘估计。

五、 逐步回归检测法

基本思想：将变量逐个的引入模型，每引入一个解释变量后，都要观察可决系数的变化，进行F检验，并对已经选入的解释变量逐个进行t检验。

（1）当引入新变量后可决系数显著改善，原来的解释变量的显著性不变化，说明新变量是独立解释变量

（2）当引入新变量后可决系数变化不显著，或使得原来的解释变量变得不再显著（或符号违背直观）时，说明新变量不是独立解释变量，则提示很可能引起了多重共线性。

当出现多个解释变量之间高度相关的时候，逐步回归方法是一种检测多重共线性的方法。

第四节 多重共线性的补救

一、增加样本容量

多重共线性的后果主要是方差变大，在有两个解释变量时

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

式中σ^2为常数，r_{23}^2 确定后，当样本容量越大时，$\sum x_{2i}^2$ 越大，可使 $Var(\hat{\beta}_2)$ 变小，从而减轻多重共线性的影响

注意：

- 增大样本容量只能减轻多重共线性的影响，不能根本解决它，当 $r_{23}^2\to 1$ 时，仍有 $Var(\hat{\beta}_2)\to\infty$
- 增大样本容量有时十分困难，受到数据来源的限制

二、利用先验信息

先验信息：在此之前的研究所提供的信息。

利用某些先验信息可把有共线性的变量组成新的变量，从而消除多重共线性

（举例：生产函数,利用规模报酬不变 $\alpha+\beta=1$ 的先验信息，把有共线性的变量组成新的变量,可避免共线性）

信息：$\alpha+\beta=1$

$$Q_t = AL_t^{\alpha}K_t^{\beta}u \longrightarrow Q_t = AL_t^{\alpha}K_t^{1-\alpha}u \rightarrow \frac{Q_t}{K_t} = A(\frac{L_t}{K_t})^{\alpha}u$$

$$\ln Q_t = \ln A + \alpha \ln L_t + \beta \ln K_t + \ln u$$

（$\ln L_t$与 $\ln K_t$ 有多重共线性）

$$\ln \frac{Q_t}{K_t} = \ln A + \alpha \ln \frac{L_t}{K_t} + \ln u$$

三、截面数据与时间序列数据的结合

有时在时间序列数据中多重共线性严重的变量，在截面数据中不一定有严重的共线性

假定前提：截面数据估计出的参数在时间序列中变化不大

方法：先用截面数据估计出一个变量的参数，再代入原模型

用时间序列数据估计另一个变量的参数

如 $$Y_t = \beta_1 + \beta_2 P_t + \beta_3 I_t + u_t$$

（Y商品销售量，P价格，I收入）

先用截面数据估计 $\hat{\beta}_3$（若各截面价格视为相同，即“保持价格不变”），

即 $$Y_i = \beta_1^* + \beta_3 I_i + v_i$$

再用时序数据估计 $\hat{\beta}_2$ $$Y_t^* = Y_t - \hat{\beta}_3 I_t = \beta_1 + \beta_2 P_t + u_{2t}$$

四、变换模型的形式

对存在多重共线性的变量，进行**对数变换、一阶差分变换、比率变换**等，有时可消除或减轻多重共线性的影响。

如一阶差分：　$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

$$Y_t - Y_{t-1} = \beta_1 + \beta_2 (X_{2t} - X_{2t-1}) + \beta_3 (X_{3t} - X_{3t-1}) + (u_t - u_{t-1})$$

注意：一阶差分可能带来新的问题：

- 虽然u_t和u_{t-1}都是序列无关的，但　$v_t = (u_t - u_{t-1})$ 常常是序列相关的，可能会违反无自相关假定.
- 一阶差分中减少了一个自由度
- 一阶差分不适于截面数据，因截面数据没有先后顺序

五、逐步回归法

基本思想：

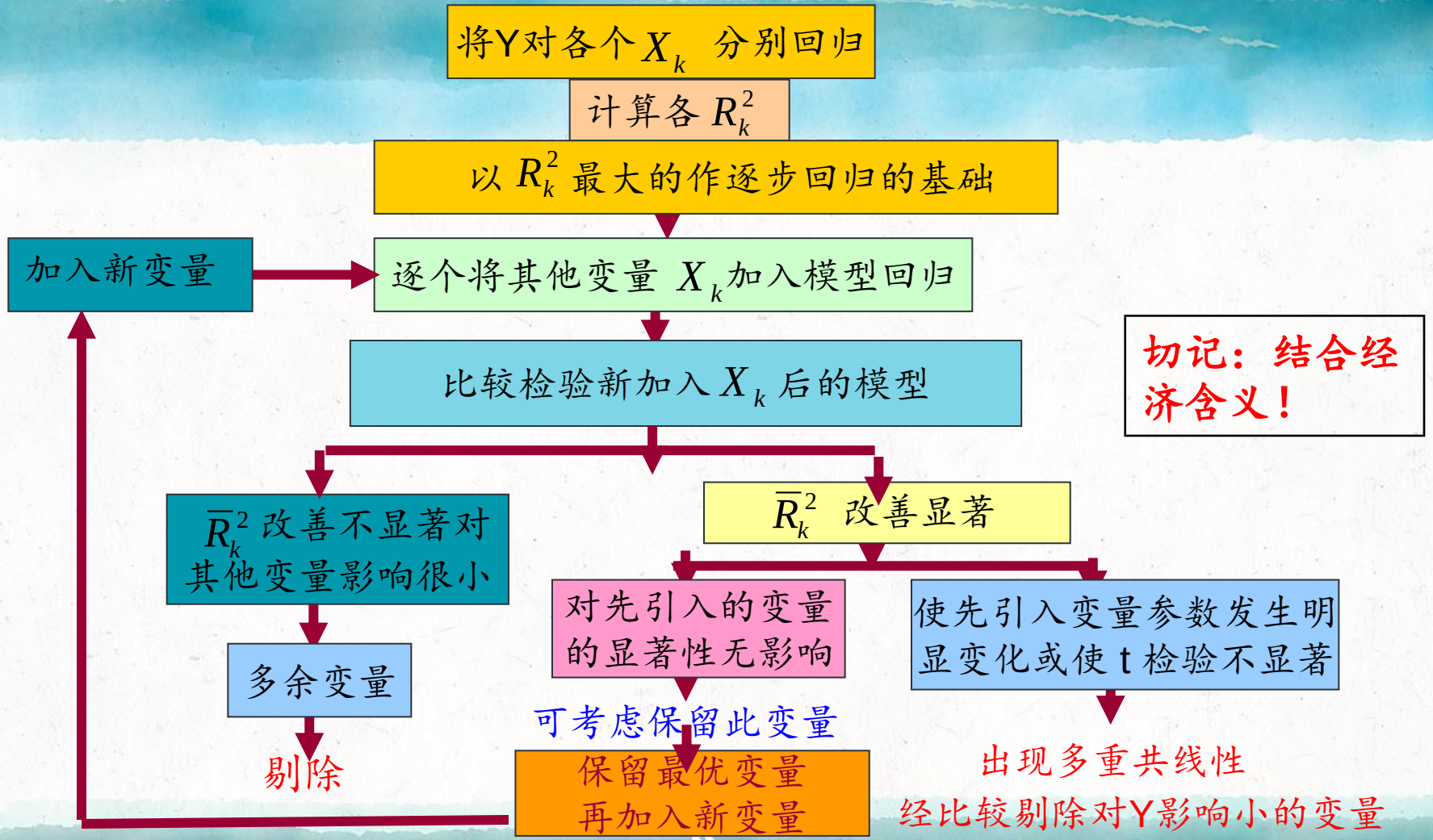
用逐步引入变量回归的方法，发现产生共线性的解释变量，并按一定原则将其剔除，从而减少多重共线性影响。

方法：

这既是判断是否存在多重共线性的方法，又是解决多重共线性的方法：基本思路的框图为：(见下页)

注意:逐步回归剔除变量时应非常谨慎，若剔除了重要变量，可能导致设定误差，而带来更严重的后果。

使用逐步回归剔除变量时要格外小心！


将Y对各个X_k 分别回归
计算各 R_k^2
以 R_k^2 最大的作逐步回归的基础
加入新变量
逐个将其他变量 X_k 加入模型回归
比较检验新加入 X_k 后的模型
切记：结合经济含义！
$\overline{R}_k^2$ 改善不显著对其他变量影响很小
$\overline{R}_k^2$ 改善显著
多余变量
对先引入的变量的显著性无影响
使先引入变量参数发生明显变化或使 t 检验不显著
剔除
可考虑保留此变量
保留最优变量
再加入新变量
出现多重共线性
经比较剔除对Y影响小的变量

第五节　案例分析

新案例：中国国内旅游收入的分析

研究目的：

近年来，中国旅游业一直保持高速发展，旅游业作为国民经济新的增长点，在整个社会经济发展中的作用日益显现。中国的旅游业分为国内旅游和入境旅游两大市场，入境旅游外汇收入年均增长22.6%，与此同时国内旅游也迅速增长。改革开放30多年来，特别是进入90年代后，中国的国内旅游收入年均增长14.4%，远高于同期GDP 9.76%的增长率。为了规划中国未来旅游产业的发展，需要定量地分析影响中国旅游市场发展的主要因素。

模型设定：

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

其中：

Y_t ——第 t年全国国内旅游总花费（亿元）

X_{2t} ——国内旅游人数（万人）

X_{3t} ——城镇居民人均旅游支出 （元）

X_{4t} ——农村居民人均旅游支出 （元）

X_{5t} ——铁路里程（万公里）

OLS回归结果

Dependent Variable: Y
Method: Least Squares
Date: 09/09/19 Time: 14:09
Sample: 1994 2016
Included observations: 23

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7460.354	4663.388	-1.599771	0.1271
X2	7.196343	1.319938	5.452033	0.0000
X3	-16.13588	3.116747	-5.177154	0.0001
X4	11.89220	4.882899	2.435479	0.0255
X5	1846.233	905.8380	2.038149	0.0565

R-squared	0.993759	Mean dependent var	11003.76
Adjusted R-squared	0.992372	S.D. dependent var	11666.83
S.E. of regression	1018.994	Akaike info criterion	16.88068
Sum squared resid	18690294	Schwarz criterion	17.12753
Log likelihood	-189.1278	Hannan-Quinn criter.	16.94276
F-statistic	716.4832	Durbin-Watson stat	1.145433
Prob(F-statistic)	0.000000		

相关系数法检验

在EVews中选择 X2、X3 、X4 、X5数据，点右键“open as group”，然后点 “view/Covariance Analysis”，在对话框中记得将默认的“covariance”改为 “correlation”，然后点“ok”，即得相关系数矩阵:

Correlation				
	X2	X3	X4	X5
X2	1.000000	0.874954	0.952018	0.989868
X3	0.874954	1.000000	0.894003	0.894863
X4	0.952018	0.894003	1.000000	0.954808
X5	0.989868	0.894863	0.954808	1.000000

由相关系数矩阵可以看出，各解释变量相互之间的相关系数较高，证实确实多重共线性较为严重。

方差膨胀因子法检验

在Eviews中，也可以直接计算解释变量的方差扩大因子，在“Equation”回归结果中点“View/Coefficient Diagnostics/ Variance Inflation Factors”即可，其中以“Centered VIF”即为方差扩大因子VIF:

Variance Inflation Factors
Date: 09/09/19 Time: 14:47
Sample: 1994 2016
Included observations: 23

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
C	21747191	481.7119	NA
X2	1.742235	165.2576	53.88798
X3	9.714109	129.4083	5.827392
X4	23.84270	55.78583	13.10557
X5	820542.5	1274.998	61.28948

经验表明,如果方差扩大因子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、X4、X5的方差扩大因子大于10，表明存在严重多重共线性问题。

对多重共线性的处理

为避免删除重要解释变量引起设定误差，不随意删除解释变量。考虑将各变量进行对数变换，再对以下模型进行估计：

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

Dependent Variable: LNY
Method: Least Squares
Date: 09/09/19 Time: 15:00
Sample: 1994 2016
Included observations: 23

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.026304	0.554904	-7.255852	0.0000
LNX2	1.032261	0.068831	14.99705	0.0000
LNX3	0.389739	0.124204	3.137902	0.0057
LNX4	0.328204	0.041489	7.910593	0.0000
LNX5	0.469701	0.213969	2.195183	0.0415

R-squared	0.998843	Mean dependent var	8.759408
Adjusted R-squared	0.998586	S.D. dependent var	1.091519
S.E. of regression	0.041039	Akaike info criterion	-3.358915
Sum squared resid	0.030316	Schwarz criterion	-3.112069
Log likelihood	43.62753	Hannan-Quinn criter.	-3.296834
F-statistic	3886.192	Durbin-Watson stat	0.903516
Prob(F-statistic)	0.000000		

在Eviews的命令窗分别输入以下命令就可以实现对数变换：

```
genr lny=log(y)
genr lnx2=log(x2)
genr lnx3=log(x3)
genr lnx4=log(x4)
genr lnx5=log(x5)
```

对多重共线性的处理

也可以在Eviews的命令窗直接输入“LS log(y) c log(x2) log(x3) log(x4) log(x5)”，敲回车即可：

Dependent Variable: LOG(Y)
Method: Least Squares
Date: 09/09/19 Time: 15:06
Sample: 1994 2016
Included observations: 23

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.026304	0.554904	-7.255852	0.0000
LOG(X2)	1.032261	0.068831	14.99705	0.0000
LOG(X3)	0.389739	0.124204	3.137902	0.0057
LOG(X4)	0.328204	0.041489	7.910593	0.0000
LOG(X5)	0.469701	0.213969	2.195183	0.0415

R-squared	0.998843	Mean dependent var	8.759408
Adjusted R-squared	0.998586	S.D. dependent var	1.091519
S.E. of regression	0.041039	Akaike info criterion	-3.358915
Sum squared resid	0.030316	Schwarz criterion	-3.112069
Log likelihood	43.62753	Hannan-Quinn criter.	-3.296834
F-statistic	3886.192	Durbin-Watson stat	0.903516
Prob(F-statistic)	0.000000		

在这个案例中，经过数据变换后，尽管解释变量之间高度相关，但相关的统计检验指标，如回归方程检验的F统计量、各回归系数的 t 统计量都高度显著，且所有系数都具有正确的符号，这表明所有这些变量一起对国内旅游收入具有显著的影响。

第五节　案例分析

老案例：中国国内旅游收入的分析

研究目的： 中国国内旅游市场发展迅速，需要定量地研究影响中国国内旅游市场发展的主要原因。经分析，可以旅游收入表示旅游市场发展，除了国内旅游人数和旅游支出外，还可能与旅游基础设施有关。

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

其中：

Y_t ——第 t年全国旅游收入

X_{2t} ——国内旅游人数（万人）

X_{3t} ——城镇居民人均旅游支出 （元）

X_{4t} ——农村居民人均旅游支出 （元）

X_{5t} ——公路里程（万公里）

X_{6t} ——铁路里程（万公里）

1994—2003年的统计数据(教材数据)

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

OLS回归结果

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

结果分析

该模型 $R^2 = 0.9954$，$\overline{R}^2 = 0.9897$ 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。

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

各解释变量相互之间的相关系数较高，证实确实存在严重多重共线性。

用方差扩大因子法检验

例如作X3对X2、X4、X5、X6的辅助回归得

$$R_{X3}^2 = 0.948332$$

方差扩大因子为：

$$VIF_{X3} = \frac{1}{(1-R_{X3}^2)} = \frac{1}{(1-0.948332)} = 19.3543$$

由于 $VIF_{X3} = 19.3543 \geq 10$ ，根据经验，说明X3与其他解释变量间有严重多重共线性。

其他变量间的多重共线性可用类似方式检验。

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）	公路里程X5（万km）	铁路里程X6（万km）
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
2004	4710.7	110200	731.8	210.2	187.07	7.44
2005	5285.9	121200	737.1	227.6	193.05	7.54
2006	6229.74	139400	766.4	221.9	345.70	7.71

Dependent Variable: Y
Method: Least Squares
Date: 10/19/09 Time: 09:18
Sample: 1994 2007
Included observations: 14

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1471.956	1137.046	-1.294544	0.2316
X2	0.042510	0.004613	9.216082	0.0000
X3	4.432478	1.063341	4.168445	0.0031
X4	2.922273	1.093665	2.672001	0.0283
X5	1.426786	1.417555	1.006512	0.3436
X6	-354.9821	244.8486	-1.449802	0.1852

R-squared	0.997311	Mean dependent var	3527.783
Adjusted R-squared	0.995630	S.D. dependent var	1927.495
S.E. of regression	127.4135	Akaike info criterion	12.83028
Sum squared resid	129873.5	Schwarz criterion	13.10416
Log likelihood	-83.81195	F-statistic	593.4168
Durbin-Watson stat	1.558415	Prob(F-statistic)	0.000000

结果：

可决系数F统计量有改善

X2变得显著了，

但X5变得不显著.

X6参数的符号仍然为负

说明：

多重共线性问题

还没有解决！

修正多重共线性 —模型变换

Dependent Variable: LNY
Method: Least Squares
Date: 10/19/09 Time: 09:28
Sample: 1994 2007
Included observations: 14

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8.923615	0.998396	-8.937951	0.0000
LNX2	0.833541	0.139286	5.984390	0.0003
LNX3	0.841457	0.222852	3.775853	0.0054
LNX4	0.240369	0.048235	4.983246	0.0011
LNX5	0.027838	0.093990	0.296179	0.7746
LNX6	0.354642	0.594876	0.596161	0.5675

R-squared	0.997477	Mean dependent var	8.021852
Adjusted R-squared	0.995900	S.D. dependent var	0.580626
S.E. of regression	0.037177	Akaike info criterion	-3.448708
Sum squared resid	0.011057	Schwarz criterion	-3.174826
Log likelihood	30.14095	F-statistic	632.5752
Durbin-Watson stat	1.355879	Prob(F-statistic)	0.000000

结果：

可决系数改变不大，

F统计量有改善

X2、X3、X4都显著，

但X5、X6不显著.

X6参数的符号变为正，与经验符合

说明：

多重共线性问题有改善，但需分析X5、X6的影响和多重共线性的作用.

修正多重共线性—逐步回归

采用逐步回归的办法，去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。

一元回归结果：

变量	X2	X3	X4	X5	X6
参数估计值	0.0588	14.0225	19.6103	22.5957	3025.062
t 统计量	18.2488	9.3090	3.2710	8.7084	9.1392
R^2	0.9652	0.8784	0.4714	0.8634	0.8744
$\overline{R}^2$	0.9623	0.8682	0.4273	0.8520	0.8639

加入X2的方程 $\overline{R}^2$ 最大，以X2为基础，顺次加入其他变量逐步回归

加入新变量回归结果（一）

$[t_{0.025}(n-k)=2.201]$

$[t_{0.05}(n-k)=1.796]$

	X2	X3	X4	X5	X6	$\overline{R}^2$
X2、X3	0.0410 （15.2635）	5.1427 （7.6657）				0.9935 ↑
X2、X4	0.0523 （5.3186）		5.4830 （5.3186）			0.9885 ↑
X2、X5	0.0587 （5.6753）			0.0536 （0.0128）		0.9589 ↓
X2、X6	0.0434 （8.2145）				935.0066 （3.2754）	0.9792 ↑

新加入X3的方程 $\overline{R}^2=0.9935$，改进最大，且t检验显著。保留X3，再加入其他新变量逐步回归

$[t_{0.025}(n-k)=2.228]$

$[t_{0.05}(n-k)=1.812]$

加入新变量的回归结果（二）

	X2	X3	X4	X5	X6	$\overline{R}^2$
X2、X3、X4	0.0435 (16.0418)	3.6660 (3.8314)	2.1786 (1.9744)			0.9949 ↑
X2、X3、X5	0.0379 (7.5541)	5.1881 (7.5308)		1.2342 (0.7205)		0.9932 ↓
X2、X3、X6	0.0418 (13.7021)	5.7560 (4.8365)			-178.7471 (-0.6325)	0.9931 ↓

在X2、X3基础上加入X4后的方程$\overline{R}^2$明显增大，而且各个参数t检验都显著。加入X5后不仅$\overline{R}^2$下降，而且X5参数的t检验不显著；加入X6后不仅$\overline{R}^2$下降，X6参数的t检验不显著，甚至X6的符号也变得不合理。保留X4，再加入其他新变量逐步回归

加入新变量的回归结果（三）

$[t_{0.025}(n-k)=2.262]$

$[t_{0.05}(n-k)=1.833]$

	X2	X3	X4	X5	X6	$\overline{R}^2$
X2、X3、X4、X5	0.0394 (9.1108)	3.5794 (3.8145)	2.4034 (2.1951)	1.7859 (1.2078)		0.9951 ↑
X2、X3、X4、X6	0.0461 (15.6295)	4.6031 (4.3817)	2.8112 (2.5817)		-398.0537 (-1.6499)	0.9956 ↑

加入X5后 $\overline{R}^2$ 有改进，但X5参数的t检验不显著。加入X6后 $\overline{R}^2$ 有改进，但X6参数的t检验不显著，并且参数为负值不合理。从相关系数也可看出，**X5**、**X6** 与其他变量高度相关，这说明主要是X5、X6引起严重多重共线性，应予剔除。

修正严重多重共线性影响后的回归结果

$$\hat{Y} = -3136.713 + 0.0435X_2 + 3.6660X_3 + 2.1786X_4$$

t= （-10.5998）(16.0418) （3.8314） (1.9744)

$R^2 = 0.9961$ $\bar{R}^2 = 0.9949$

F=841.4324 DW=1.1763

存在的问题：

1.样本容量过小，自由度太小（n-k)=14-4=10，其可靠性受到一定影响。

2.剔除的X5、X6有可能是重要变量，容易引起设定误差。

作业

本科教材练习题4.4

电子版，附详细过程及必要截图

（建议再自己练习一下4.5）