§2.8多重共线性
Multi-Collinearity

一、多重共线性的概念

二、多重共线性的后果

三、多重共线性的检验

四、克服多重共线性的方法

五、案例

六、分部回归与多重共线性

一、多重共线性的概念

非多重共线性假定

$$\mathrm{rk}(X'X) = \mathrm{rk}(X) = k$$

解释变量不是完全线性相关的或接近完全线性相关的。

$|r_{xi\,xj}| \neq 1$，$|r_{xi\,xj}|$ 不近似等于1。

就模型中解释变量的关系而言，有三种可能。

（1）$r_{xi\,xj} = 0$，解释变量间相关系数等于0。（少见）

（2）$|r_{xi\,xj}| = 1$，解释变量间完全相关。（少见）

（3）$0 < |r_{xi\,xj}| < 1$，解释变量间存在一定程度的线性相关。（常见）

因此我们关心的不是有无多重共线性，而是多重共线性的程度。随着共线性程度的加强，对参数估计值的准确性、稳定性带来影响。

1、多重共线性

对于模型

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\ldots+\beta_kX_{ki}+\mu_i$$

$$i=1,2,\ldots,n \qquad (2.6.1)$$

其基本假设之一是解释变量是互相独立的。

如果某两个或多个解释变量之间出现了相关性，则称为多重共线性。

如果存在

$$c_1X_{1i}+c_2X_{2i}+\ldots+c_kX_{ki}=0$$

$$i=1,2,\ldots,n \qquad (2.6.2)$$

其中：c_i不全为0，即某一个解释变量可以用其它解释变量的线性组合表示，则称为解释变量间存在**完全共线性**。

如果存在

$$c_1X_{1i}+c_2X_{2i}+\ldots+c_kX_{ki}+v_i=0$$

$$i=1,2,\ldots,n \qquad (2.6.3)$$

其中ci不全为0，为随机误差项，则称为**一般共线性(近似共线性)或交互相关(intercorrelated)**。

在矩阵表示的线性回归模型

$$\mathbf{Y=XB+N}$$

中，**完全共线性指：秩(X)<k+1**，即矩阵

$$X=\begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix}$$

中，至少有一列向量可由其他列向量（不包括第一列）线性表出。

例如，$X2=\lambda X1$，这时X1与X2的相关系数为1，解释变量X2对因变量的作用完全可由X1代替。

注意：

完全共线性的情况并不多见，一般出现的是在一定程度上的共线性，即近似共线性。

2、实际经济问题中的多重共线性现象

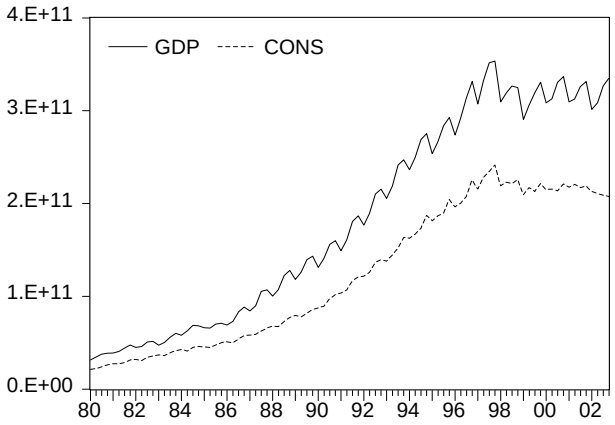
- 经济变量的共同变化趋势

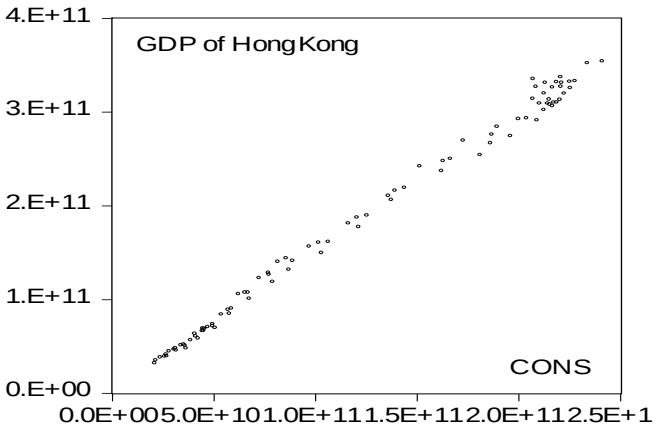
时间序列样本：经济繁荣时期，各基本经济变量（收入、消费、投资、价格）都趋于增长；衰退时期，又同时趋于下降。

横截面数据：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。

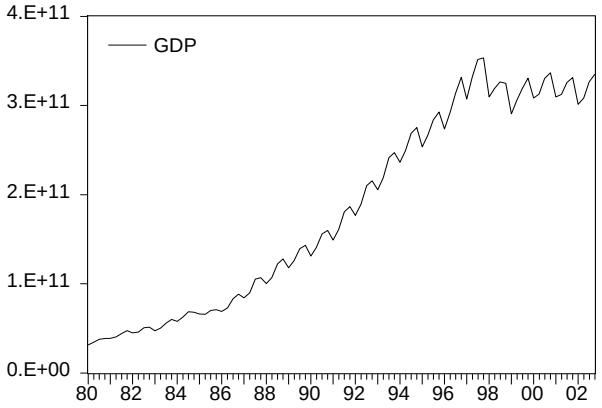
多重共线性的经济解释

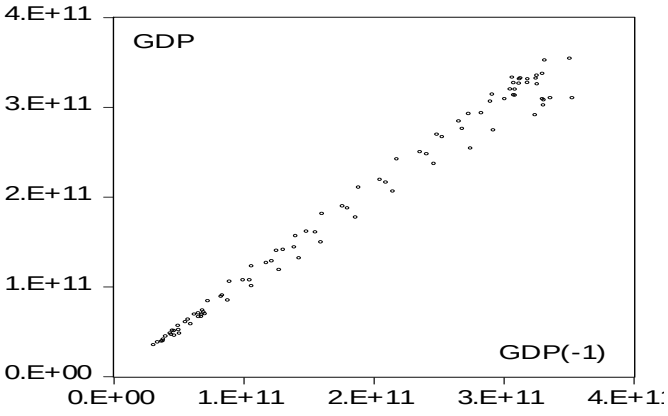
（1）经济变量在时间上有共同变化的趋势。如在经济上升时期，收入、消费、就业率等都增长，当经济收缩期，收入、消费、就业率等又都下降。当这些变量同时进入模型后就会带来多重共线性问题。





（2）解释变量与其滞后变量同作解释变量。





- **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。

例如，消费=f(当期收入, 前期收入）

显然，两期收入间有较强的线性相关性。

- 一般经验

对于采用时间序列数据作样本、以简单线性形式建立的计量经济学模型，往往存在多重共线性。

以截面数据作样本时，问题不那么严重，但多重共线性仍然是存在的。

二、多重共线性的后果

1、完全共线性下参数估计量不存在

多元线性模型

$$\mathbf{Y} = \mathbf{X}\mathrm{B} + \mathrm{N}$$

的普通最小二乘参数估计量为：

$$\hat{\mathrm{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \qquad (2.6.4)$$

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

例如：对一个离差形式的二元回归模型

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

如果两个解释变量完全相关，如 $x_2 = \lambda x_1$，则有

$$X'X = \begin{pmatrix} \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i}x_{1i} & \sum x_{2i}^2 \end{pmatrix} = \begin{pmatrix} \sum x_{1i}^2 & \lambda\sum x_{1i}^2 \\ \lambda\sum x_{1i}^2 & \lambda^2\sum x_{1i}^2 \end{pmatrix} = \sum x_{1i}^2 \begin{pmatrix} 1 & \lambda \\ \lambda & \lambda^2 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{pmatrix} = \sum x_{1i}y_i \begin{pmatrix} 1 \\ \lambda \end{pmatrix}$$

该回归模型的正规方程为

$$(\mathbf{X}'\mathbf{X})\hat{\mathbf{B}} = \mathbf{X}'\mathbf{Y}$$

或

$$\begin{cases} \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i}x_{2i} = \sum x_{1i}y_i \\ \hat{\beta}_1 \sum x_{2i}x_{1i} + \hat{\beta}_2 \sum x_{2i}^2 = \sum x_{2i}y_i \end{cases}$$

解该线性方程组得：

$$\hat{\beta}_1 = \frac{\begin{vmatrix} \sum x_{1i}y_i & \sum x_{1i}x_{2i} \\ \sum x_{2i}y_i & \sum x_{2i}^2 \end{vmatrix}}{\begin{vmatrix} \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i}x_{1i} & \sum x_{2i}^2 \end{vmatrix}} = \frac{\begin{vmatrix} \sum x_{1i}y_i & \lambda\sum x_{1i}^2 \\ \lambda\sum x_{1i}y_i & \lambda^2\sum x_{1i}^2 \end{vmatrix}}{\begin{vmatrix} \sum x_{1i}^2 & \lambda\sum x_{1i}^2 \\ \lambda\sum x_{1i}^2 & \lambda^2\sum x_{1i}^2 \end{vmatrix}} = \frac{0}{0}$$

$\hat{\beta}_1$为不定式；

同理，$\hat{\beta}_2$也为不定式，其值无法确定。

事实上，当 $x_2 = \lambda x_1$ 时，原二元回归模型退化为一元回归模型：

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

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

$$\hat{\beta}_1 + \lambda\hat{\beta}_2 = \sum x_{1i} y_i \Big/ \sum x_{1i}^2$$

2、近似共线性下普通最小二乘法参数估计量非有效

在一般共线性（或称近似共线性）下，虽然可以得到OLS法参数估计量，但是由参数估计量方差的表达式为

$$Cov(\hat{\mathrm{B}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

可见，由于此时|X’X|≈0，引起(X’X) $^{-1}$主对角线元素较大，从而使参数估计值的方差增大，OLS参数估计量非有效。

仍以一元模型中 $\hat{\beta}_1$ 为例，$\hat{\beta}_1$ 的方差为

$$\mathrm{var}(\hat{\beta}_1) = \sigma^2 (X'X)^{-1}_{11} = \frac{\sigma^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

$$= \frac{\sigma^2 / \sum x_{1i}^2}{1 - (\sum x_{1i} x_{2i})^2 / \sum x_{1i}^2 \sum x_{2i}^2}$$

$\frac{(\sum x_{1i} x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}$ 恰为 x_1 与 x_2 的线性相关系数的平方 r^2，由于 $r^2 \leq 1$，故 $\frac{1}{1-r^2} \geq 1$。

当完全不共线时，$r^2=0$，$\text{var}(\hat{\beta}_1)=\sigma^2/\sum x_{1i}^2$

当不完全共线（近似共线）时，$0<r^2<1$，

$$\text{var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum x_{1i}^2}\bullet\frac{1}{1-r^2}>\frac{\sigma^2}{\sum x_{1i}^2}$$

即：**多重共线性使参数估计值的方差增大，方差扩大因子(Variance Inflation Factor)为$1/(1-r^2)$，**其增大趋势见下表：

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
方差扩大因子	1	2	5	10	20	25	33	50	100	1000

当完全共线时，$r^2=1$，$\text{var}(\hat{\beta}_1)=\infty$

3、参数估计量经济含义不合理

如果模型中两个解释变量具有线性相关性，例如X_1和X_2，那么它们中的一个变量可以由另一个变量表征。

这时，X_1和X_2前的参数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。

所以各自的参数已经失去了应有的经济含义，于是经常表现出似乎反常的现象，例如本来应该是正的，结果恰是负的。

4、变量的显著性检验失去意义

存在多重共线性时

↓

参数估计值的方差与标准差变大

↓

使t统计量的拒绝域变小（临界值增大）

↓

容易使通过样本计算的t值小于临界值，
误导作出参数为0的推断

↓

可能将重要的解释变量排除在模型之外

5、模型的预测功能失效

- 变大的方差容易使区间预测的“区间”变大，使预测失去意义。

- 能否说：如果存在完全共线性，预测值的置信区间为（－∞，+∞）？

三、多重共线性的检验

- 由于多重共线性表现为解释变量之间具有相关关系，所以用于多重共线性的检验方法主要是统计方法：如判定系数检验法、逐步回归检验法等。

- **多重共线性检验的任务**是：

（1）检验多重共线性是否存在；

（2）估计多重共线性的范围，即判断哪些变量之间存在共线性。

多重共线性的检验

（1）初步观察。当模型的拟合优度（R^2）很高，F值很高，而每个回归参数估计值的方差$\mathrm{Var}(\beta_j)$又非常大（即t值很低）时，说明解释变量间可能存在多重共线性。

（2）**Klein**判别法。计算多重可决系数R^2及解释变量间的简单相关系数$r_{xi\,xj}$。若有某个$|r_{xi\,xj}| > R^2$，则x_i，x_j间的多重共线性是有害的。

（3）回归参数估计值的符号不符合经济理论。

（4）增加或减少解释变量个数时，回归参数估计值变化很大。

1、检验多重共线性是否存在

(1)对两个解释变量的模型，采用简单相关系数法

求出X1与X2的简单相关系数r，若|r|接近1，则说明两变量存在较强的多重共线性。

(2)对多个解释变量的模型，采用综合统计检验法

若 在OLS法下，模型的R^2与F值较大，但各参数估计值的t检验值较小，说明各解释变量对Y的联合线性作用显著，但各解释变量间存在共线性而使得它们对Y的独立作用不能分辨，故t检验不显著。

2、判明存在多重共线性的范围

(1) 判定系数检验法

- 使模型中每一个解释变量分别以其余解释变量为解释变量进行回归计算，并计算相应的拟合优度，也称为判定系数。如果在某一种形式

$$X_{ji}=\alpha_1 X_{1i}+\alpha_2 X_{2i}+\ldots\alpha_L X_{Li}$$

中判定系数较大，则说明在该形式中作为被解释变量的Xj可以用其他X的线性组合代替，即Xj与其他X之间存在共线性。

- 等价的检验是对上述回归方程作F检验

构造如下 F 统计量：

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

式中：$R_{j.}^2$为第j个解释变量对其他解释变量的回归方程的决定系数，

若存在较强的共线性，则$R_{j.}^2$较大且接近于1，这时（1- $R_{j.}^2$ ）较小，从而F_j的值较大。因此，可以在给定的显著性水平α下，通过计算F值的方法进行检验。

- 另一等价的检验:

在模型中排除某一个解释变量X_j，估计模型，如果拟合优度与包含X_j时十分接近，则说明X_j与其它解释变量之间存在共线性。

(2) 逐步回归法

- 以**Y**为被解释变量，逐个引入解释变量，构成回归模型，进行模型估计。
- 根据拟合优度的变化决定新引入的变量是否可以用其它变量的线性组合代替，而不作为独立的解释变量。
- 如果拟合优度变化显著，则说明新引入的变量是一个独立解释变量；
- 如果拟合优度变化很不显著，则说明新引入的变量不是一个独立解释变量，它可以用其它变量的线性组合代替，也就是说它与其它变量之间存在共线性关系。

四、克服多重共线性的方法

多重共线性的克服方法

1 直接合并解释变量

当模型中存在多重共线性时，在不失去实际意义的前提下，可以把有关的解释变量直接合并，从而降低或消除多重共线性。

如果研究的目的是预测全国货运量，那么可以把重工业总产值和轻工业总产值合并为工业总产值，甚至还可以与农业总产值合并，变为工农业总产值。解释变量变成了一个，自然消除了多重共线性。

2 利用已知信息合并解释变量

通过经济理论及对实际问题的深刻理解，对发生多重共线性的解释变量引入附加条件从而减弱或消除多重共线性。

比如有二元回归模型 $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t$

x_1与x_2间存在多重共线性。如果能给出β_1与β_2的某种关系，$\beta_2 = \lambda\beta_1$其中 λ 为常数。

$$y_t = \beta_0 + \beta_1 x_{t1} + \lambda\beta_1 x_{t2} + u_t = \beta_0 + \beta_1 (x_{t1} + \lambda x_{t2}) + u_t$$

令 $x_t = x_{t1} + \lambda x_{t2}$ 得$y_t = \beta_0 + \beta_1 x_t + u_t$

模型是一元线性回归模型，所以不再有多重共线性问题。

多重共线性的克服方法

3 增加样本容量或重新抽取样本

这种方法主要适用于那些由测量误差而引起的多重共线性。当重新抽取样本时，克服了测量误差，自然也消除了多重共线性。有时，增加样本容量也可以减弱多重共线性的程度。

4 利用解释变量之间的关系

如果解释变量之间存在多重共线性，那么可以利用它们之间的关系，引入附加方程，从而将单方程模型转化为联立方程模型，克服多重共线性。

5 变换模型形式

通过变换模型形式克服多重共线性。例如某产品销量**Y**取决于其出厂价格**X1**，市场价格**X2**，和市场供应量**X3**。模型为

$$LnY = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + u_t$$

通常，*X*1与*X*2是高度相关的，如果研究的目的是预测销售量**Y**，则可以用相对价格*X*1/ *X*2代替*X*1与*X*2对销售量**Y**的影响，

$$LnY = \beta_0 + \beta_1 (X1/X2) + \beta_3 X3 + u_t$$

从而克服了*X*1与*X*2的多重共线性。

多重共线性的克服方法

5.6 把数据中心化

把数据中心化有时也是克服多重共线性的有效方法。

例如多项式回归模型

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + u_t$$

中，变量之间常存在多重共线性。

可以把解释变量先中心化（各自减自己的均值），

然后建立多元回归模型

$$y_t = \beta_0 + \beta_1 (x_{t1} - \bar{x}_1) + \beta_2 (x_t^2 - \bar{x}_2) + \beta_3 (x_t^3 - \bar{x}_3) + u_t$$

多重共线性的克服方法

5.7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。按可决系数大小给解释变量重要性排序。

（2）以可决系数最大的回归方程为基础，按解释变量重要性大小为顺序逐个引入其余的解释变量。这个过程会出现3种情形。

①若新变量的引入改进了R^2，且回归参数的t检验在统计上也是显著的，则该变量在模型中予以保留。

②若新变量的引入未能改进R^2，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的，应该舍弃。

③若新变量的引入未能改进R^2，且显著地影响了其他回归参数估计值的符号与数值，同时本身的回归参数也通不过t检验，这说明出现了严重的多重共线性。舍弃该变量。

1、第一类方法：排除引起共线性的变量

• 找出引起多重共线性的解释变量，将它排除出去，是最为有效的克服多重共线性问题的方法。以逐步回归法得到最广泛的应用。

• 注意：

剩余解释变量参数的经济含义和数值都发生了变化。

2、第二类方法：差分法

- 对于以时间序列数据为样本、以直接线性关系为模型关系形式的计量经济学模型，将原模型变换为差分模型

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

可以有效地消除存在于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。

例如：在中国消费模型中的2个变量：

收入(Y：GDP)与消费 C 的总量与增量数据

	Y	*C(-1)*	*C(-1)/Y*	*△Y*	*△C(-1)*	*△C(-1)/△Y*
1981	4901	2976	0.6072			
1982	5489	3309	0.6028	588	333	0.5663
1983	6076	3638	0.5996	587	329	0.5605
1984	7164	4021	0.5613	1088	383	0.3520
1985	8792	4694	0.5339	1628	673	0.4134
1986	10133	5773	0.5697	1441	1079	0.7488
1987	11784	6542	0.5552	1651	769	0.4658
1988	14704	7451	0.5067	2920	909	0.3113
1989	16466	9360	0.5684	1762	1909	1.083
1990	18320	10556	0.5762	1854	1196	0.6451
1991	21280	11362	0.5339	2960	806	0.2723
1992	25864	13146	0.5083	4584	1784	0.3892
1993	34501	15952	0.4624	8637	2806	0.3249
1994	47111	20182	0.4284	12610	4230	0.3354
1995	59405	27216	0.4581	12294	7034	0.5721
1996	68498	34529	0.5041	9093	7313	0.8042

- 由表中的比值可以直观地看到，两变量增量的线性关系弱于总量之间的线性关系。

- 进一步分析：

Y与C(-1)之间的判定系数为0.9845，

△Y与△C(-1)之间的判定系数为0.7456。

一般认为：两个变量之间的判定系数大于0.8时，二者之间存在线性关系。

所以，原模型经检验地被认为具有多重共线性，而差分模型则可认为不具有多重共线性。

3、第三类方法：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，所以采取适当方法减小参数估计量的方差，虽然没有消除模型中的多重共线性，但确能消除多重共线性造成的后果。

- 例如，增加样本容量，可使参数估计量的方差减小。

- 再如：岭回归法（**Ridge Regression**）

70年代发展的岭回归法，以引入偏误为代价减小参数估计量的方差，受到人们的重视。

具体方法是：引入矩阵**D**，使参数估计量为

$$\hat{B} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D})^{-1}\boldsymbol{X}'\boldsymbol{Y} \qquad (2.6.5)$$

其中矩阵**D**一般选择为主对角阵，即

$$D = aI \qquad (2.6.6)$$

a为大于0的常数。

显然，与未含**D**的参数**B**的估计量相比，**(2.6.5)**的估计量有较小的方差。

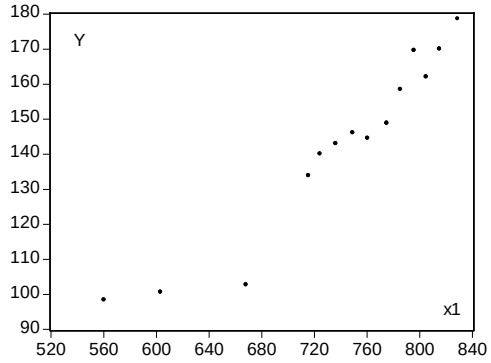
五、 案例分析（3例）

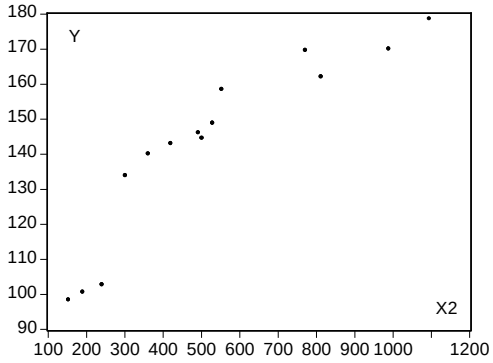
例8.1：天津市粮食需求模型（1974-1987）（file: li-7-1）

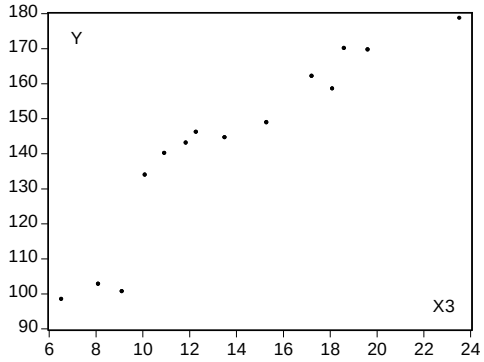
y：粮食销售量（万吨 / 年），x_1：市常住人口数（万人），

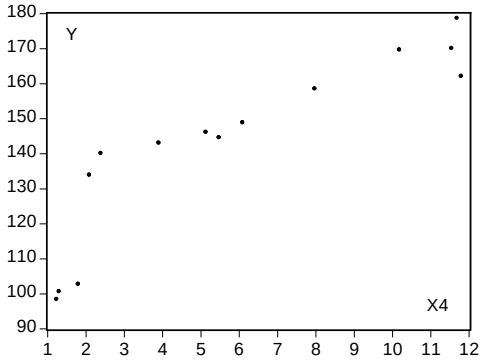
x_2：人均收入（元 / 年），x_3：肉销售量（万吨 / 年），

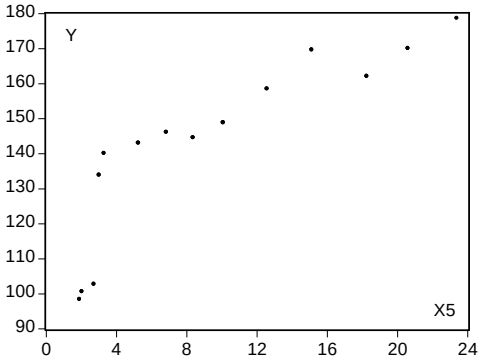
x_4：蛋销售量（万吨 / 年），x_5：鱼虾销售量（万吨 / 年）。











案例分析（例8.1）

Dependent Variable: Y
Method: Least Squares
Date: 11/02/00 Time: 12:17
Sample: 1974 1987
Included observations: 14

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.496563	30.00659	-0.116526	0.9101
X1	0.125330	0.059139	2.119245	0.0669
X2	0.073667	0.037877	1.944897	0.0877
X3	2.677589	1.257293	2.129646	0.0658
X4	3.453448	2.450850	1.409082	0.1965
X5	-4.491117	2.214862	-2.027719	0.0771

R-squared	0.970442	Mean dependent var	142.7129
Adjusted R-squared	0.951968	S.D. dependent var	26.09805
S.E. of regression	5.719686	Akaike info criterion	3.785355
Sum squared resid	261.7185	Schwarz criterion	4.059237
Log likelihood	-40.36262	F-statistic	52.53086
Durbin-Watson stat	1.972755	Prob(F-statistic)	0.000007

$$y = -3.497 + 0.125\,x_1 + 0.074\,x_2 + 2.678\,x_3 + 3.453\,x_4 - 4.491\,x_5$$

(-0.1) (2.1) (1.9) (2.1) (1.4) (-2.0)

$R^2 = 0.97$, $F = 52.59$, DW = 1.97, $t_{0.05}(8) = 2.31$, $T = 14$, (1974-1987)

$R^2 = 0.97$，而每个回归参数的t检验在统计上都不显著，这说明模型中存在严重的多重共线性。

案例分析（3例）

把解释变量换成对数形式建模还是存在多重共线性。

$$y = -134.248 + 0.013x_1 + 33.611Lnx_2 + 34.363Lnx_3 + 27.280Lnx_4 - 34.906Lnx_5$$

(-2.0)　(0.1)　(1.7)　(1.8)　(1.3)　(-1.6)

$R^2 = 0.97$, $F = 50.2$, DW = 1.96, $T = 14$, $t_{0.05}(8) = 2.31$, (1974-1987)

用Klein判别法进行分析。首先给出解释变量间的简单相关系数矩阵。

Correlation Matrix

	X1	X2	X3	X4	X5
X1	1.000000	0.866552	0.882293	0.852449	0.821305
X2	0.866552	1.000000	0.945896	0.964773	0.982532
X3	0.882293	0.945896	1.000000	0.940506	0.948361
X4	0.852449	0.964773	0.940506	1.000000	0.981979
X5	0.821305	0.982532	0.948361	0.981979	1.000000

因为其中有两个简单相关系数大于$R^2 = 0.97$，

所以根据Klein判别法，模型中存在严重的多重共线性。

案例分析（3例）

用逐步回归法筛选解释变量。

（1）用每个解释变量分别对被解释变量做简单回归，

以可决系数为标准确定解释变量的重要程度，为解释变量排序。

$y = -90.921 + 0.317\ x_1$

(-4.7)　(12.2)　$R^2 = 0.92$, $F = 147.6$, $T = 14$, (1974-1987)

$y = 99.613 + 0.082\ x_2$

(15.4)　(7.6)　$R^2 = 0.83$, $F = 57.6$, $T = 14$, (1974-1987)

$y = 74.648 + 4.893\ x_3$

(9.0)　(8.7)　$R^2 = 0.86$, $F = 75.4$, $T = 14$, (1974-1987)

$y = 108.865 + 5.740\ x_4$

(18.3)　(6.8)　$R^2 = 0.80$, $F = 46.8$, $T = 14$, (1974-1987)

$y = 113.375 + 3.081\ x_5$

(18.7)　(6.0)　$R^2 = 0.75$, $F = 36.1$, $T = 14$, (1974-1987)

解释变量的重要程度依次为x_1, x_3, x_2, x_4, x_5。

案例分析（3例）

（2）以第一个回归方程$y = -90.921 + 0.317\ x_1$为基础，

依次引入x_3, x_2, x_4, x_5。首先把x_3引入模型，

$$y = -39.795 + 0.212\ x_1 + 1.909\ x_3$$

(-1.6)　(4.7)　　(2.6)　　$R^2 = 0.95, F = 114, T = 14$, (1974-1987)

因为R^2从0.92增至0.95，且x_3的系数通过显著性检验，所以在模型中保留x_3。

再把x_2引入模型，

$$y = -34.777 + 0.207\ x_1 + 0.009\ x_2 + 1.456\ x_3$$

(-1.3)　　(4.3)　　(0.5)　　(1.2)　$R^2 = 0.96, F = 70.8, T = 14$, (1974-1987)

因为x_2的引入没有使R^2得到改善，同时还使各回归系数的t值下降，所以应剔除x_2。

把x_4引入模型，

$$y = -37.999 + 0.210\ x_1 + 1.746\ x_3 + 0.235\ x_4$$

(-1.4)　　(4.4)　　(1.5)　　(0.2)　　$R^2 = 0.95, F = 69, T = 14$, (1974-1987)

同理，应剔除x_4。

案例分析（3例）

把x_5引入模型，

$$y = -40.823 + 0.211\,x_1 + 2.145\,x_3 - 0.157\,x_5$$

(-1.5)　　(4.4)　　(1.6)　　(-0.2)

$R^2 = 0.95$, $F = 69$, $T = 14$, (1974-1987)

同理，应剔除x_5。最后确定的模型是

$$y = 0.141\,x_1 + 2.80\,x_3$$

(14.6)　　(5.8)　　$R^2 = 0.94$, $F = 119.8$, $T = 14$, (1974-1987)

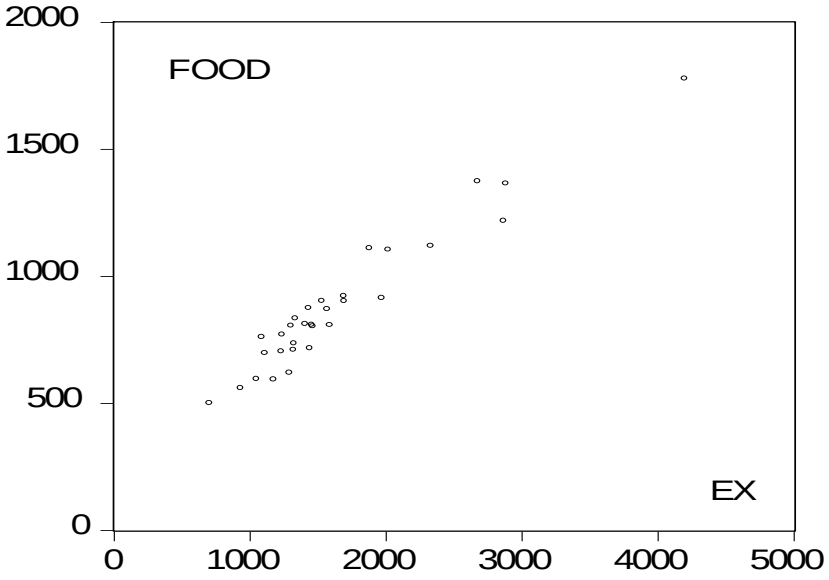
或者以y与x_1, x_3, x_2, x_4, x_5的相关系数大小排序。按如上的方法逐步回归，得最终结果：

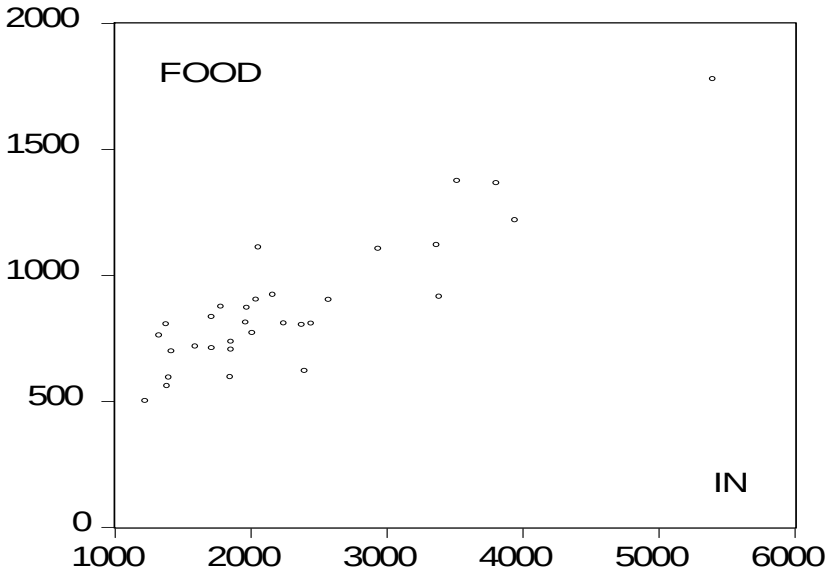
$$y = 0.164\,x_1 + 0.042\,x_2$$

(26.3)　　(5.4)　　$R^2 = 0.94$, $F = 182.0$, $T = 14$, (1974-1987)

补充案例1：1998年农村居民食品支出（file:B1E4）

1998年31省市自治区农村居民人均年食品支出（food，元）、人均年总支出（EX，元）和人均年可支配收入（IN，元）。见散点图，food与EX 和IN都是正相关的，





补充案例1：1998年农村居民食品支出（file:B1E4）

建立2元回归模型：

Dependent Variable: FOOD
Method: Least Squares
Date: 01/01/02 Time: 05:49
Sample: 1 31
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	278.6733	30.84747	9.033914	0.0000
EX	0.507690	0.055109	9.212529	0.0000
IN	-0.105945	0.041317	-2.564165	0.0160

[illegible]ared	0.948247	Mean dependent var	877.2419
Adjusted R-squared	0.944550	S.D. dependent var	273.1305
S.E. of regression	64.31622	Akaike info criterion	8.419389
Sum squared resid	115824.1	Schwarz criterion	8.558162
Log likelihood	-171.4876	F-statistic	256.5145
Durbin-Watson stat	0.977692	Prob(F-statistic)	0.000000

估计结果IN回归系数是负的。显然与事实不符、与经济理论不符。原因是EX和IN之间的多重共线性（高度相关）。

r(EX, IN) = 0.9537大于可决系数**0.9482**。按**Klein**判别准则模型存在严重的多重共线性。

Correlation Matrix

	EX	IN
EX	1.000000	0.953720
IN	0.953720	1.000000

补充案例1：1998年农村居民食品支出（file:B1E4）

另外，如果用**food**只对**IN**回归，回归系数是正的。这也说明上述二元回归结果中存在多重共线性。

$$Food_t = 285.5945 + 0.2571\ In_t$$

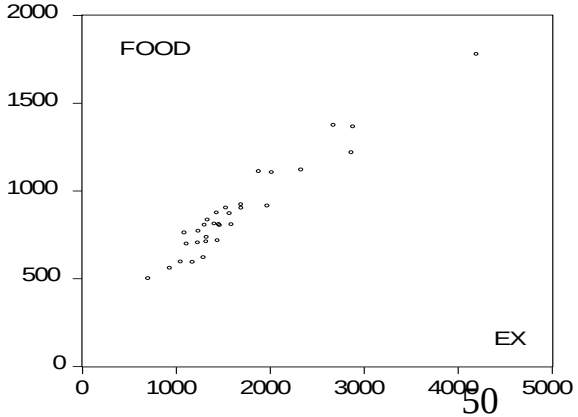
$$(4.7) \qquad (10.5) \qquad R^2 = 0.79,\ F = 110,\ T = 31$$

处理方法是用**food**只对**EX**回归。效果很好。

Dependent Variable: FOOD
Method: Least Squares
Date: 01/01/02 Time: 06:19
Sample: 1 31
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	258.4643	32.56437	7.937028	0.0000
EX	0.372922	0.018094	20.61049	0.0000

R-squared	0.936094	Mean dependent var	877.2419
Adjusted R-squared	0.933890	S.D. dependent var	273.1305
S.E. of regression	70.22670	Akaike info criterion	8.565798
Sum squared resid	143021.9	Schwarz criterion	8.658313
Log likelihood	-174.7570	F-statistic	424.7922
Durbin-Watson stat	0.729924	Prob(F-statistic)	0.000000



补充案例2：中国私人轿车拥有量决定因素分析（file：nonli14）

1985-2002年中国私人轿车拥有量以年增长率**23%**，年均增长**55**万辆的速度飞速增长。

考虑到目前农村家庭购买私人轿车的现象还很少，在建立中国私人轿车拥有量模型时，主要考虑如下因素：（**1**）城镇居民家庭人均可支配收入；（**2**）城镇总人口；（**3**）轿车产量；（**4**）公路交通完善程度；（**5**）轿车价格。

定义变量名如下：

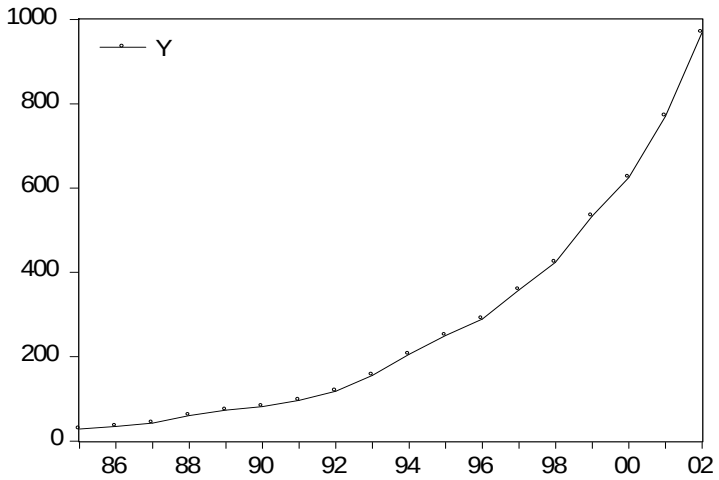
Y：中国私人轿车拥有量（万辆）

X1：城镇居民家庭人均可支配收入（元）

X2：全国城镇人口（亿人）

X3：全国汽车产量（万辆）

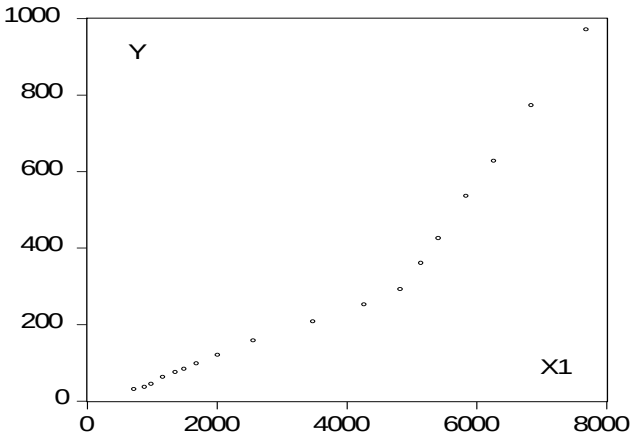
X4；全国公路长度（万公里）

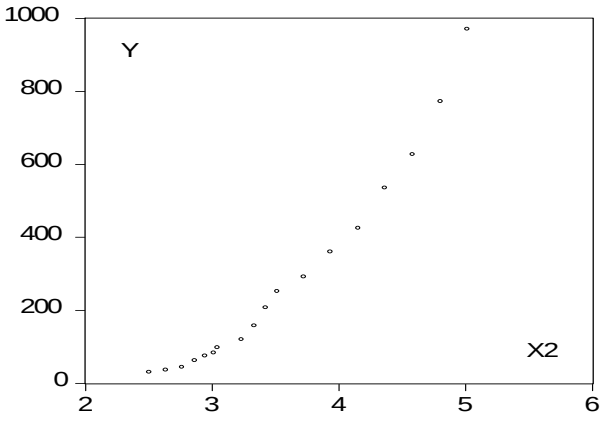


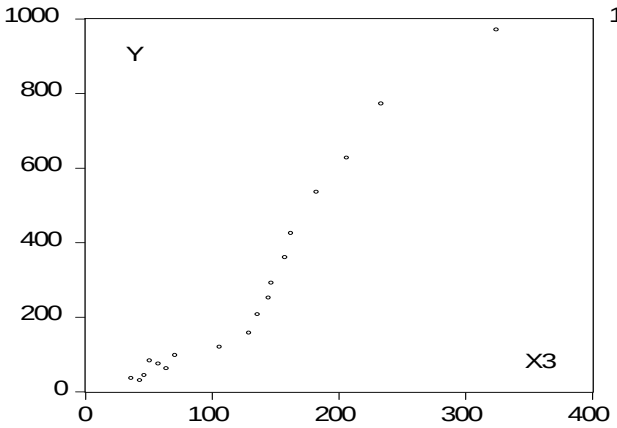
案例分析

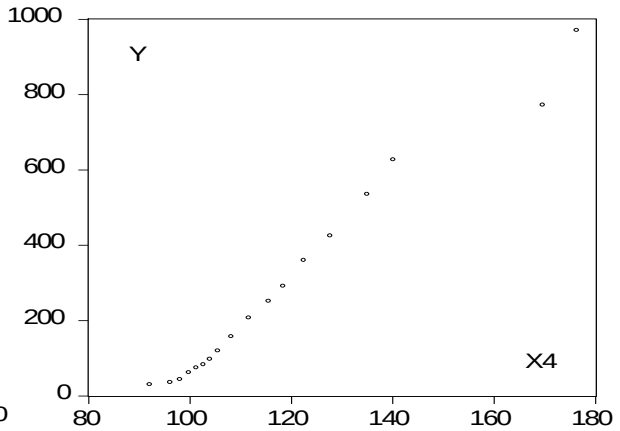
补充案例2

中国私人轿车拥有量决定因素分析









Correlation Matrix				
	X1	X2	X3	X4
X1	1.000000	0.983022	0.958465	0.929555
X2	0.983022	1.000000	0.962856	0.958785
X3	0.958465	0.962856	1.000000	0.955281
X4	0.929555	0.958785	0.955281	1.000000

案例分析

补充案例2

中国私人轿车拥有量决定因素分析

Dependent Variable: Y
Method: Least Squares
Date: 02/02/05 Time: 20:10
Sample: 1985 2002
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-925.6637	163.8137	-5.650711	0.0001
X1	0.005702	0.023424	0.243424	0.8115
X2	62.94281	84.36504	0.746077	0.4689
X3	0.411564	0.507786	0.810506	0.4322
X4	7.729285	1.560120	4.954290	0.0003

R-squared	0.986721	Mean dependent var	284.2606
Adjusted R-squared	0.982635	S.D. dependent var	278.4439
S.E. of regression	36.69198	Akaike info criterion	10.27313
Sum squared resid	17501.92	Schwarz criterion	10.52045
Log likelihood	-87.45814	F-statistic	241.4995
Durbin-Watson stat	1.400855	Prob(F-statistic)	0.000000

看相关系数阵，**Y**与**X1**，**X2**，**X3**，**X4**的相关系数都在**0.9**以上，但输出结果中，解释变量**X1**，**X2**，**X3**的回归系数却通不过显著性检验。这预示着解释变量之间一定存在多重共线性。

案例分析

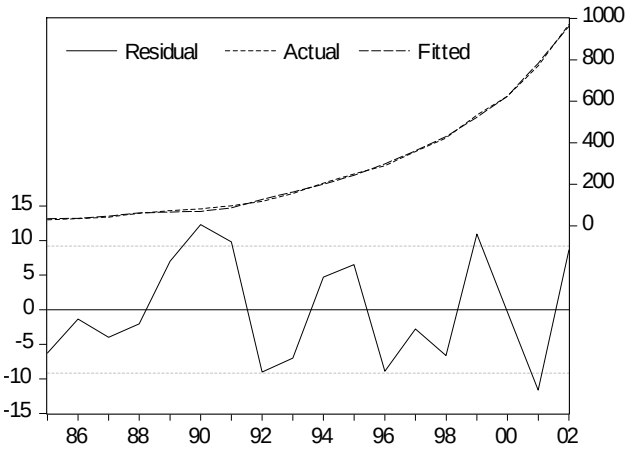
补充案例2
中国私人轿车拥有量决定因素分析

看散点图，把**Y**与**X3**，**X4**处理成线性关系，把**Y**与**X1**，**X2**处理成幂函数（抛物线）关系，得结果如下，

Dependent Variable: Y
Method: Least Squares
Date: 02/02/05 Time: 20:10
Sample: 1985 2002
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	440.7471	112.4087	3.920935	0.0020
X1^2	3.31E-06	8.74E-07	3.786107	0.0026
X2	-429.4148	47.49970	-9.040367	0.0000
X2^2	80.78993	8.597263	9.397168	0.0000
X3	0.558851	0.127728	4.375321	0.0009
X4	1.480844	0.588140	2.517841	0.0270

R-squared	0.999232	Mean dependent var	284.2606
Adjusted R-squared	0.998912	S.D. dependent var	278.4439
S.E. of regression	9.184921	Akaike info criterion	7.534205
Sum squared resid	1012.353	Schwarz criterion	7.830995
Log likelihood	-61.80784	F-statistic	3122.264
Durbin-Watson stat	1.887328	Prob(F-statistic)	0.000000



五、案例：服装市场需求函数□

1、建立模型

• 根据理论和经验分析，影响居民服装类支出的主要因素有：可支配收入、居民流动资产拥有量、服装价格指数、物价总指数。

• 已知某地区的有关资料，根据散点图判断，建立线性服装消费支出模型：

$$Y=\beta_0+\beta_1X+\beta_2K+\beta_3P_1+\beta_4P_0+\mu$$

2、样本数据

单位：万元，指数：1993=100

年份	服装支出 Y	可支配收入 X	流动资产 K	服装类价格指数 P1	总物价指数 P0
1989	8.4	82.9	17.1	92	94
1990	9.6	88	21.3	93	96
1991	10.4	99.9	25.1	96	97
1992	11.4	105.3	29	94	97
1993	12.2	117.7	34	100	100
1994	14.2	131	40	101	101
1995	15.8	148	44	105	104
1996	17.9	161.8	49	112	109
1997	19.3	174.2	51	112	111
1998	20.8	184.7	53	112	111

3、估计模型

(1)用 OLS 法估计上述模型：

$$\hat{Y} = -13.20 + 0.10X + 0.001K - 0.197P_1 + 0.334P_0$$

(-1.76)　(3.71)　(0.30)　(-2.20)　(2.24)

$r^2=0.9980$　　$R^2=0.9965$　　F=638.4

由于R^2较大且接近于1，而且 F=638.4，大于临界值：$F_{0.05}(4,5)=15.19$，故认为服装支出与上述解释变量间总体线性关系显著。

但由于参数K的估计值的t检验值较小（未能通过检验），故解释变量间存在多重共线性。

（2）检验简单相关系数

列出 X，K，P1，P0 的相关系数矩阵：

	X	K	P1	P0
X	1	0.9883	0.9804	0.9878
K	0.9883	1	0.9700	0.9695
P1	0.9804	0.9700	1	0.9918
P0	0.9878	0.9695	0.9918	1

- 各解释变量间存在高度相关性，其中尤其以**P1，P0**间的相关系数为最高。

（3）找出最简单的回归形式

分别作 Y 与 X，K，P1，P0 间的回归：

①$\hat{Y} = -1.24 + 0.118X$

(-3.36)　(42.48)

R^2=0.9950　　F=1805.1

②$\hat{Y} = 2.118 + 0.327K$

(2.58)　(15.31)

R^2=0.9629　F=234.4

③$\hat{Y} = -38.5 + 0.516P_1$

(-9.16)　　(12.53)

R^2=0.9455　　F=157.1

④$\hat{Y} = -53.7 + 0.663P_0$

(-14.77)　(18.66)

R^2=0.9747　F=348.1

- 可见，应选①为初始的回归模型。

（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。

Y	C	X	K	P1	P0	R^2	F
=f(X)	-1.25	0.12				0.9950	1805.1
t 值	-3.36	42.49					
=f(X,P1)	1.53	0.13		-0.04		0.9958	826.9
t	0.31	8.57		-0.57			
=f(X,P1,K)	1.06	0.14	-0.04	-0.04		0.9941	509.0
t	0.21	5.70	-0.68	-0.53			
=f(X,P1,P0)	-12.45	0.10		-0.19	0.31	0.9970	1003.6
t	-1.92	7.55		-2.47	2.59		
=f(X,P1,P0,K)	-13.20	0.10	0.01	-0.20	0.33	0.9965	638.4
	-1.79	3.71	0.30	-2.20	2.24		

4、讨论:

①在初始模型中引入P_1，模型拟合优度提高，且参数符号合理，但P_1的t检验未通过；

②再引入K，拟合优度虽有提高，但K与P_1的t检验未能通过，且X与P_1的t检验值及F检验值有所下降，表明引入K并未对回归模型带来明显的“好处”，K可能是多余的；

③去掉K，加入P_0，拟合优度有所提高，且各解释变量的t检验全部通过，F值也增大了。

④将4个解释变量全部包括进模型，拟合优度未有明显改观，K的t检验未能通过，K显然是多余的。

5、结论

回归方程以Y=f(X,P1,P0)为最优：

$$Y=-12.45+0.10X-0.19P1+0.31P0$$

五、案例二：中国消费函数模型

1、OLS估计结果

Dependent Variable: CONS
Method: Least Squares
Date: 03/01/03 Time: 00:46
Sample: 1981 1996
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	540.5286	84.30153	6.411848	0.0000
GDP	0.480948	0.021861	22.00035	0.0000
CONS1	0.198545	0.047409	4.187969	0.0011

R-squared	0.999773	Mean dependent var	13618.94
Adjusted R-squared	0.999739	S.D. dependent var	11360.47
S.E. of regression	183.6831	Akaike info criterion	13.43166
Sum squared resid	438613.2	Schwarz criterion	13.57652
Log likelihood	-104.4533	F-statistic	28682.51
Durbin-Watson stat	1.450101	Prob(F-statistic)	0.000000

2、差分法估计结果

Dependent Variable: DCONS

Method: Least Squares

Date: 03/18/03 Time: 23:18

Sample(adjusted): 1982 1996

Included observations: 15 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DGDP	0.496723	0.026879	18.48006	0.0000
DCONS1	0.158504	0.051678	3.067122	0.0090

R-squared	0.992686	Mean dependent var	2457.533
Adjusted R-squared	0.992123	S.D. dependent var	2422.687
S.E. of regression	215.0169	Akaike info criterion	13.70288
Sum squared resid	601019.5	Schwarz criterion	13.79728
Log likelihood	-100.7716	Durbin-Watson stat	2.612102

3、比较

β1：0.48095→0.49672

β2：0.19854→0.15850

在消除了共线性后，**GDP对CONS的影响增大，CONS1对CONS的影响减少。**

六、分部回归与多重共线性

1、分部回归法(Partitioned Regression)

对于模型

$$\mathbf{Y} = \mathbf{X}\mathrm{B} + \mathrm{N}$$

将解释变量分为两部分，对应的参数也分为两部分：

$$\mathbf{Y} = \mathbf{X}_1\mathrm{B}_1 + \mathbf{X}_2\mathrm{B}_2 + \mathrm{N}$$

在满足解释变量与随机误差项不相关的情况下，可以写出关于参数估计量的方程组：

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 \\ \mathbf{x}_2'\mathbf{x}_1 & \mathbf{x}_2'\mathbf{x}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathrm{B}}_1 \\ \hat{\mathrm{B}}_2 \end{bmatrix}$$

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\mathbf{B}}_2$$

$$= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{Y} - \mathbf{X}_2\hat{\mathbf{B}}_2)$$

如果存在 $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$

则有 $\hat{\mathbf{B}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$

这就是仅以X_1作为解释变量时的参数估计量。

同样有 $\hat{\mathbf{B}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}$

这就是仅以X_2作为解释变量时的参数估计量。

2、由分部回归法导出

- 如果一个多元线性模型的解释变量之间完全正交，可以将该多元模型分为多个一元模型、二元模型、...进行估计，参数估计结果不变；
- 实际模型由于存在或轻或重的共线性，如果将它们分为多个一元模型、二元模型、...进行估计，参数估计结果将发生变化；

- 当模型存在共线性，将某个共线性变量去掉，剩余变量的参数估计结果将发生变化，而且经济含义有发生变化；

- 严格地说，实际模型由于总存在一定程度的共线性，所以每个参数估计量并不 真正反映对应变量与被解释变量之间的结构关系。