



金融计量学

(Financial Econometrics)

主讲教师：任 飞

商学院金融系

办公地点：商学院1601

email: olrenfei@163.com



第四章 多重共线性和虚拟变量的应用

第一节 多重共线性的概念和后果

- 一、多重共线性的概念和产生
- 二、多重共线性的后果

第二节 多重共线性的检验

- 一、检验多重共线性问题是否严重
- 二、判断多重共线性的存在范围
- 三、检验多重共线性的表现形式

第三节 多重共线性的修正

- 一、删除不必要的变量
- 二、改变解释变量的形式
- 三、补充新数据
- 四、利用先验信息法

第四节 金融数据的多重共线性处理

第五节 虚拟变量模型

- 一、虚拟变量的性质和设置原则
- 二、虚拟变量模型的运用

第六节 回归模型的结构稳定性检验——邹氏检验

- 一、邹氏检验的过程

第七节 回归模型的结构稳定性检验——虚拟变量法

第八节 实例——虚拟变量在金融数据处理中的作用

- 一、简单理论回顾
- 二、实证检验



一、多重共线性的概念和产生

（一）概念

多重共线性（multicollinearity）：最早由挪威经济学家弗瑞希（Ragnar Frisch）于1934年提出，是指回归模型中的一些或全部解释变量中存在的一种完全（perfect）或准确（exact）的**线性关系**，也包括近似多重共线性（near multicollinearity）。

以一组解释变量 X_1, X_2, \dots, X_n 为例：

- **完全多重共线性**：存在一组不完全为零的常数 $\lambda_1, \lambda_2, \dots, \lambda_n$ ，满足 $\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_n X_n = 0$
- **近似多重共线性**： $\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_n X_n + \mu = 0$ ，其中 μ 表示随机误差项



一、多重共线性的概念和产生

(二) 多重共线性产生的原因

(1) 数据收集及计算方法

- 对数据进行处理或者交叉计算。

(2) 模型或从中取样的总体受到限制

- 金融系统中的金融变量相互依存、相互制约
- 某些金融变量可能受某同一决定因素的影响

例：深交所股指和上交所股指共同受我国各种经济、政治因素的影响

(3) 模型设定偏误

- 回归模型中加入多项式项，且自变量的变动范围较窄

此外，观测值个数较少，或同时使用解释变量的当期值和滞后值，也容易产生多重共线性。



二、多重共线性的后果

以某一离差形式（即 $x_i = X_i - \bar{X}$ ）表示的二元线性回归模型 $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + v_i$ 为例：

（1）若存在完全多重共线性，假设 $x_{1i} = \lambda x_{2i}$ ， $\lambda \neq 0$ ，则

$$\hat{\beta}_1 = \frac{\sum y_i x_{1i} \sum x_{2i}^2 - \sum x_{1i} x_{2i} \sum y_i x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2} = \frac{\lambda \sum y_i x_{2i} \sum x_{2i}^2 - \lambda \sum y_i x_{2i} \sum x_{2i}^2}{\lambda^2 (\sum x_{2i}^2)^2 - \lambda^2 (\sum x_{2i}^2)^2} = \frac{0}{0} \quad (4.1)$$

同理， $\hat{\beta}_2$ 也无法确定，即不能求得参数估计值。

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_v^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2} = \frac{\sigma_v^2 \sum x_{2i}^2}{\lambda^2 (\sum x_{2i}^2)^2 - \lambda^2 (\sum x_{2i}^2)^2} = \infty \quad (4.2)$$

同理， $\hat{\beta}_2$ 的方差也无限大。因此，存在多重共线性时，参数估计值的方差无限大。



二、多重共线性的后果

以某一离差形式（即 $x_i = X_i - \bar{X}$ ）表示的二元线性回归模型
 $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + v_i$ 为例：

（2）若存在近似多重共线性，假设 $x_{1i} = \lambda x_{2i} + \varepsilon_i$ ，并且回归残差 v_i 满足 $\sum x_{2i} \varepsilon_i = 0$ ， $\lambda \neq 0$ ，则

$$\hat{\beta}_1 = \frac{\sum y_i x_{1i} \sum x_{2i}^2 - \sum x_{1i} x_{2i} \sum y_i x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2} = \frac{\sum x_{2i}^2 [\sum (\lambda x_{2i} + \varepsilon_i) y_i] - \sum y_i x_{2i} [\sum (\lambda x_{2i} + \varepsilon_i) x_{2i}]}{\sum x_{2i}^2 \sum (\lambda x_{2i} + \varepsilon_i)^2 - [\sum (\lambda x_{2i} + \varepsilon_i) x_{2i}]^2} = \frac{\sum y_i \varepsilon_i}{\sum \varepsilon_i^2} \quad (4.3)$$

$\hat{\beta}_1$ 取决于 ε_i ，数值很不稳定。

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_v^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2} = \frac{\sigma_v^2 / \sum x_{1i}^2}{1 - \gamma_{12}^2} \quad (4.4)$$

其中， γ_{12} 为 x_1 ， x_2 之间的相关系数。 $|\gamma_{12}|$ 的值越大，共线性程度越高，方差越大。



二、多重共线性的后果

在实际金融数据中，完全多重共线性只是一种极端情况，通常所说多重共线性造成的后果是指近似多重共线性造成的后果，具体而言，它将造成如下的后果：

- （1）回归方程参数估计值将变得不精确。
- （2）由于参数估计值的标准差变大， t 值将缩小，使得 t 检验有可能得出错误的结论。
- （3）将无法区分单个变量对被解释变量的影响作用。



第四章 多重共线性和虚拟变量的应用

第二节 多重共线性的检验

- 一、检验多重共线性问题是否严重
- 二、判断多重共线性的存在范围
- 三、检验多重共线性的表现形式



一、检验多重共线性问题是否严重

检验多重共线性问题的经验法则

(1) 回归模型的 R^2 值比较高（如 $R^2 > 0.8$ ），或 F 检验显著，但单个解释变量系数估计值不显著；或从金融理论知某个解释变量对因变量有重要影响，但其估计值却不显著，则可以认为存在严重的多重共线性问题。

多重共线性效应系数：度量多重共线性的严重程度

$$R^2 - \sum_{h=1}^k (R^2 - R_h^2) \quad (4.5)$$

其中， k 表示解释变量的个数， R_h^2 为因变量去掉解释变量 X_h 后对其余 $k-1$ 个解释变量进行回归的 R^2 值。

若多重共线性效应系数接近于0，可以认为多重共线性问题不严重；若接近于1，多重共线性问题比较严重。



一、检验多重共线性问题是否严重

检验多重共线性问题的经验法则

(2) 若两个解释变量之间的相关系数比较高（如大于0.8），则可认为存在严重的多重共线性。

对解释变量两两之间的相关系数组成的矩阵 R 的行列式建立假设：

$$H_0: \det(R) = 1, \quad H_1: \det(R) \neq 1$$

上述假设的检验统计量： $FG = -\left[n - 1 - \frac{1}{6}(2k + 5)\right] \ln[\det(R)]$

服从自由度为 $\frac{1}{2}k(k - 1)$ 的 χ^2 分布，其中 n 为样本容量， k 为解释变量的个数。

若拒绝原假设，即拒绝所有变量均独立的假设，认为模型中存在多重共线性。



二、判断多重共线性的存在范围

可以采用**辅助回归法**（auxiliary regression method）确定多重共线性是由哪些主要变量引起的。

辅助回归：某一解释变量对其余解释变量的回归

$$X_{1n} = \lambda_0 + \lambda_1 X_{2n} + \cdots \lambda_{k-1} X_{kn} + \mu_n$$

辅助回归法构造的检验统计量 F_i ：

$$F_i = \frac{R_i^2 / (k - 1)}{(1 - R_i^2) / (n - k)} \sim F(k - 1, n - k)$$

其中， R_i^2 为第 i 个解释变量 X_i 关于其余解释变量的辅助回归的拟合优度， k 为解释变量的个数， n 为样本容量。

若 F_i 大于临界 F 值，可以认为 X_i 与其他解释变量之间存在共线性；否则，说明 X_i 与其他解释变量之间不存在共线性。



三、检验多重共线性的表现形式

当确定多重共线性是由主要变量 X_i 引起后，可采用偏相关系数法找出与 X_i 有共线性的解释变量。

为判断 X_i 与 X_j 是否存在共线性，构造检验统计量 t_i ：

$$t_i = \sqrt{n - k - 1} \frac{\rho_{ij}}{\sqrt{1 - \rho_{ij}^2}} \sim t(n - k - 1)$$

其中， n 为样本容量， k 为解释变量的个数， ρ_{ij} 为 X_i 与 X_j 的偏相关系数，即在其他解释变量固定的情况下， X_i 与 X_j 之间的相关系数。若 t_i 的绝对值大于临界值，则认为 X_i 、 X_j 是引起多重共线性的原因，否则不是。



第四章 多重共线性和虚拟变量的应用

第三节 多重共线性的修正

- 一、删除不必要的变量
- 二、改变解释变量的形式
- 三、补充新数据



一、删除不必要的变量

若多重共线性对重要因素的系数估计值有严重影响，则必须补救。如果在产生多重共线性的因素中有相对不重要的变量，则可试着将其删除，但删除变量也可能会导致新问题的产生：

（1）被删除变量对因变量的影响将被其它解释变量和随机误差项所吸收，这可能**增强另一部分变量的多重共线性问题**，还可能使**随机误差项的自相关程度增强**。

（2）删除某个变量可能会导致**模型设定误差**（specification error）。

↑
在建立回归模型的过程中，因为错误设定模型结构而产生的误差



二、改变解释变量的形式

（一）差分法

对于时间序列数据，可以考虑对变量取差分形式，如对于模型

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \mu_t$$

若 X_{1t} 和 X_{2t} 有强相关性，可把变量变换为差分形式：

$$\Delta Y_t = \alpha_0 + \alpha_1 \Delta X_{1t} + \alpha_2 \Delta X_{2t} + \Delta \mu_t$$

其中， $\Delta Y_t = Y_t - Y_{t-1}$ ， $\Delta X_{it} = X_{it} - X_{i,t-1}$ ($i = 1, 2$)，

$$\Delta \mu_t = \mu_t - \mu_{t-1}$$

差分后的数据相关性一般会明显降低。



二、改变解释变量的形式

（二）指数增长率法

在模型中存在指数变量时，指数增长率法可能会有较好效果。如研究三种指数 p_1 ， p_2 ， p_3 关系时，采用模型

$$p_{1t} = \beta_1 + \beta_2 p_{2t} + \beta_3 p_{3t} + \mu_t$$

对模型进行如下变换：

$$\dot{p}_{1t} = \beta_1 + \beta_2 \dot{p}_{2t} + \beta_3 \dot{p}_{3t} + v_t$$

其中， $\dot{p}_{it} = \frac{p_{it} - p_{i,t-1}}{p_{i,t-1}} = \ln\left(\frac{p_{it}}{p_{i,t-1}}\right) = \ln(p_{it}) - \ln(p_{i,t-1})$ ，表示 p_{it} 的增长率。



二、改变解释变量的形式

（三）以比率代替高度相关的变量

对于模型

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \mu_t$$

若 X_{2t} 和 X_{3t} 高度相关，则可令 $R_t = \frac{X_{2t}}{X_{3t}}$ ，并将其代替 X_{2t} ， X_{3t} ：

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 R_t + v_t$$



三、补充新数据

多重共线性是一样本特征，故有可能在关于同样变量的另一样本中共线性没有第一个样本严重。**增加样本容量**，就可以避免或减轻多重共线性的危害。



四、利用先验信息法

若发生多重共线性的那些解释变量之间的关系可由先验信息得到，则在所研究的模型中利用这种关系，便可以减轻多重共线性的程度。如对于存在严重多重共线性的模型

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \mu_t$$

若由先验信息得到 Y 对 X_1 的变化率是 Y 对 X_2 变化率的两倍，即 $\beta_1 = 2\beta_2$ ，则原模型可以变换为：

$$Y_t = \beta_0 + 2\beta_2 X_{1t} + \beta_2 X_{2t} + \mu_t = \alpha_0 + \alpha_1 X_t + v_t$$

其中， $X_t = 2X_{1t} + X_{2t}$ ，则原模型的多重共线性可以消除。



第四章 多重共线性和虚拟变量的应用

第四节 金融数据的多重共线性处理



例4.1 对影响股票价格指数宏观经济因素的实证分析

我们选择上证综指(以 Y 表示)作为股票价格指数的代表。对于影响股票价格指数的宏观经济因素,初步选定如下的八个宏观变量:居民消费价格指数(X_1)、商品零售价格指数(X_2)、企业商品价格指数(X_3)、汇率(X_4)、货币供应量M1 (X_5)、进出口总值(X_6)、股市成交金额(X_7)、外汇市场交易量(X_8)。

采用的数据是从2016年1月—2021年12月的月度数据,对于价格指数变量以及汇率,以原变量形式进入模型,而对于其它变量,我们取其对数形式进入模型。建立如下模型:

$$\ln Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{4t} + \alpha_5 \ln X_{5t} + \alpha_6 \ln X_{6t} + \alpha_7 \ln X_{7t} + \alpha_8 \ln X_{8t} + \mu_t \quad (4.6)$$

例4.1 对影响股票价格指数宏观经济因素的实证分析

“Quick” → “Estimate Equation”

Equation Estimation

Specification Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$.

$\log(y)$ c x1 x2 x3 x4 $\log(x5)$ $\log(x6)$ $\log(x7)$ $\log(x8)$

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 2016m01 2021m12

确定 取消

例4.1 对影响股票价格指数宏观经济因素的实证分析

表4-1 OLS回归结果

Dependent Variable: LOG(Y)					
Method: Least Squares					
Date: 02/17/22 Time: 20:13					
Sample: 2016M01 2021M12					
Included observations: 72					
Variable	Coefficient	Std. Error	t-Statistic	Prob.	
C	3.376744	0.851231	3.966896	0.0002	
X1	0.056286	0.011927	4.719142	0.0000	
X2	-0.076491	0.012109	-6.317066	0.0000	
X3	0.016989	0.001626	10.44735	0.0000	
X4	-0.125972	0.022827	-5.518595	0.0000	
LOG(X5)	0.274360	0.064252	4.270054	0.0001	
LOG(X6)	0.123392	0.045617	2.704934	0.0088	
LOG(X7)	0.106823	0.010680	10.00204	0.0000	
LOG(X8)	-0.150598	0.035471	-4.245669	0.0001	
R-squared	0.894815	Mean dependent var		8.041658	
Adjusted R-squared	0.881458	S.D. dependent var		0.091619	
S.E. of regression	0.031544	Akaike info criterion		-3.958371	
Sum squared resid	0.062688	Schwarz criterion		-3.673788	
Log likelihood	151.5014	Hannan-Quinn criter.		-3.845078	
F-statistic	66.99296	Durbin-Watson stat		1.222117	
Prob(F-statistic)	0.000000				

例4.1 对影响股票价格指数宏观经济因素的实证分析

利用普通最小二乘法回归方程，得到如下的结果：

$$\begin{aligned} \ln Y_t = & 3.38 + 0.06X_{1t} - 0.08X_{2t} + 0.02X_{3t} - 0.13X_{4t} + 0.27\ln X_{5t} \\ & (3.97^*) \quad (4.72^*) \quad (-6.32^*) \quad (10.45^*) \quad (-5.52^*) \quad (4.27^*) \\ & + 0.12\ln X_{6t} + 0.11\ln X_{7t} - 0.15\ln X_{8t} \\ & (2.70^*) \quad (10.00^*) \quad (-4.25^*) \end{aligned}$$

其中，括号内表示的是 t 值，*表示在1%的显著性水平上显著。

模型的 $R^2 = 0.8948$ ， $\bar{R}^2 = 0.8814$

➤ 检验多重共线性是否严重

1. 计算多重共线性效应系数 $R^2 - \sum_{h=1}^k (R^2 - R_h^2) = 0.3183$ ，接近于0，初步判断模型中可能并不存在严重的多重共线性。

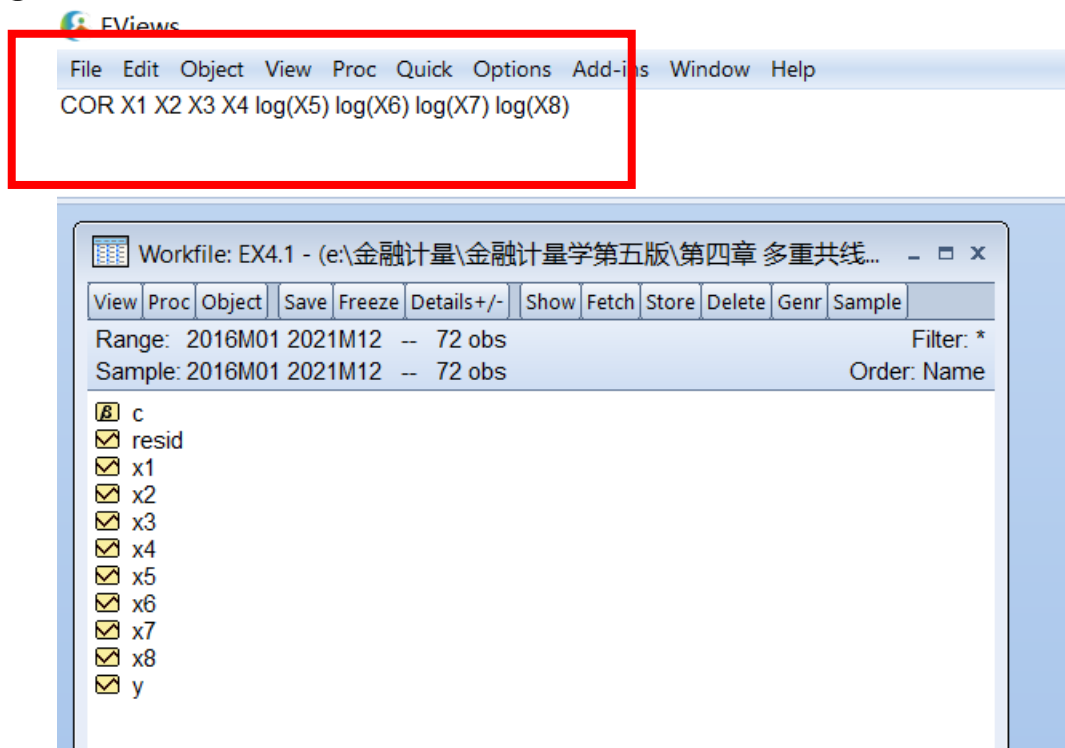
R^2	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2	R_6^2	R_7^2	R_8^2
0.8948	0.8576	0.8282	0.7126	0.8440	0.8644	0.8826	0.7278	0.8647

其中， R_h^2 为因变量在去掉解释变量 X_h 后，对其余 $k - 1$ 个解释变量进行回归的 R^2 值（ $h = 1, \dots, 8$ ）。

例4.1 对影响股票价格指数宏观经济因素的实证分析

2. 计算解释变量之间的相关系数

在Eviews软件的命令窗口键入：**COR X1 X2 X3 X4 log(X5) log(X6) log(X7) log(X8)**，回车，得到包含相关系数矩阵的窗口。



例4.1 对影响股票价格指数宏观经济因素的实证分析

G Group: UNTITLED Workfile: EX4.1::Untitled\								
View	Proc	Object	Print	Name	Freeze	Sample	Sheet	Stats Spec
Correlation								
	X1	X2	X3	X4	LOG(X5)	LOG(X6)	LOG(X7)	LOG(X8)
X1	1.000000	0.750608	-0.333174	0.549989	-0.137201	-0.294856	-0.215097	-0.216322
X2	0.750608	1.000000	0.236935	0.267539	0.264996	0.169089	-0.009487	0.205483
X3	-0.333174	0.236935	1.000000	-0.302768	0.345931	0.422301	0.121258	0.331744
X4	0.549989	0.267539	-0.302768	1.000000	-0.008425	-0.244570	-0.233450	-0.104595
LOG(X5)	-0.137201	0.264996	0.345931	-0.008425	1.000000	0.801600	0.517066	0.831219
LOG(X6)	-0.294856	0.169089	0.422301	-0.244570	0.801600	1.000000	0.468651	0.880093
LOG(X7)	-0.215097	-0.009487	0.121258	-0.233450	0.517066	0.468651	1.000000	0.500828
LOG(X8)	-0.216322	0.205483	0.331744	-0.104595	0.831219	0.880093	0.500828	1.000000

图4-1 相关系数矩阵

可以看到，变量 $\log(X_5)$ 、 $\log(X_6)$ 、 $\log(X_8)$ 两两之间的相关系数很高，变量 X_1 和 X_2 之间的相关系数也较高，因此可以认为他们之间分别存在共线性。



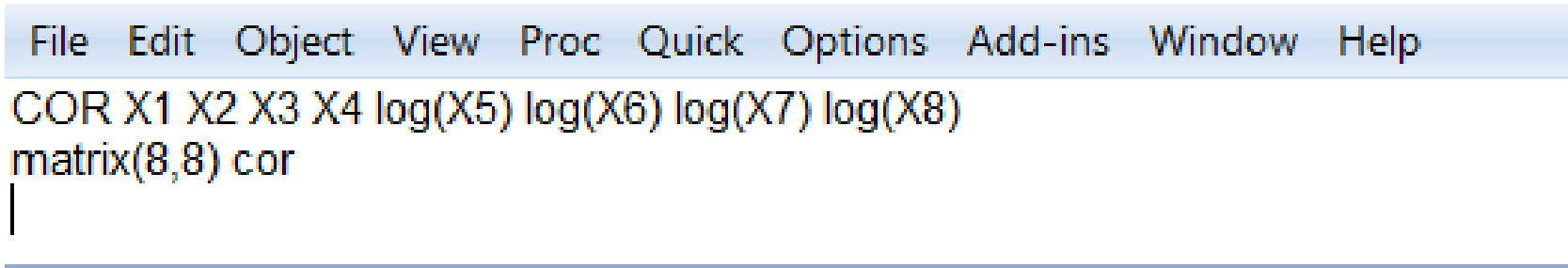
例4.1 对影响股票价格指数宏观经济因素的实证分析

为综合探测所有解释变量间的多重共线性，可对解释变量两两之间的相关系数组成的矩阵 R 的行列式建立假设：

$$H_0: \det(R) = 1, \quad H_1: \det(R) \neq 1$$

在Eviews计算矩阵行列式的步骤如下：

首先在命令窗口键入：matrix (8,8) cor



例4.1 对影响股票价格指数宏观经济因素的实证分析

生成8行8列的空矩阵cor:

Matrix: COR Workfile: EX4.1::Untitled\								
View	Proc	Object	Print	Name	Freeze	Edit+/-	Label+/-	Sheet
COR								
	C1	C2	C3	C4	C5	C6	C7	C8
Last updated: 02/22/22 - 21:11								
R1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
R2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
R3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
R4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
R5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
R6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
R7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
R8	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

进一步将上一步得到的相关系数矩阵复制到cor的对应元素中:

Matrix: COR Workfile: EX4.1::Untitled\								
View	Proc	Object	Print	Name	Freeze	Edit+/-	Label+/-	Sheet
COR								
Last updated: 02/22/22 - 21:15								
R1	1.000000	0.750608	-0.333174	0.549989	-0.137201	-0.294856	-0.215097	-0.216322
R2	0.750608	1.000000	0.236935	0.267539	0.264996	0.169089	-0.009487	0.205483
R3	-0.333174	0.236935	1.000000	-0.302768	0.345931	0.422301	0.121258	0.331744
R4	0.549989	0.267539	-0.302768	1.000000	-0.008425	-0.244570	-0.233450	-0.104595
R5	-0.137201	0.264996	0.345931	-0.008425	1.000000	0.801600	0.517066	0.831219
R6	-0.294856	0.169089	0.422301	-0.244570	0.801600	1.000000	0.468651	0.880093
R7	-0.215097	-0.009487	0.121258	-0.233450	0.517066	0.468651	1.000000	0.500828
R8	-0.216322	0.205483	0.331744	-0.104595	0.831219	0.880093	0.500828	1.000000

例4.1 对影响股票价格指数宏观经济因素的实证分析

在命令窗口键入： `scalar cordet=@det(cor)`



File Edit Object View Proc Quick Options Add-ins Window Help

COR X1 X2 X3 X4 log(X5) log(X6) log(X7) log(X8)

matrix(8,8) cor

scalar cordet=@det(cor)|

得到变量 *cordet*，其值即为相关系数矩阵的行列式，为0.001662。

# Scalar: CORDET Workfile: EX4.1::Untitled\				
View	Proc	Object	Print	Name Freeze Edit+/-
0.001661565809220868				
		Value		
CORDET		0.001662		



例4.1 对影响股票价格指数宏观经济因素的实证分析

计算检验统计量： $FG = -\left[n - 1 - \frac{1}{6}(2k + 5)\right] \ln[\det(R)] =$
 $-\left[72 - 1 - \frac{1}{6}(2 * 8 + 5)\right] \ln(0.001662) = 432.0 > \chi_{0.05}^2(28) = 41.34$
因此拒绝原假设，认为解释变量之间存在多重共线性。

例4.1 对影响股票价格指数宏观经济因素的实证分析

➤ 采用辅助回归法判断多重共线性的存在范围

以变量 X_1 为例，构建辅助回归模型：

$$X_{1t} = \lambda_0 + \lambda_1 X_{2t} + \lambda_2 X_{3t} + \lambda_3 X_{4t} + \lambda_4 \log(X_{5t}) + \lambda_5 \log(X_{6t}) \\ + \lambda_6 \log(X_{7t}) + \lambda_7 \log(X_{8t}) + \mu_t$$

Equation Estimation

Specification Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$.

x1 c x2 x3 x4 log(x5) log(x6) log(x7) log(x8)

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 2016M01 2021M12

确定 取消

例4.1 对影响股票价格指数宏观经济因素的实证分析

采用辅助回归法判断多重共线性的存在范围

F统计值对应的p值为0.00，因此可认为 X_1 与其他解释变量之间存在多重共线性；同时采用克里安的经验法则，辅助回归的 R^2 （0.9212）大于主回归的 R^2 （0.8948），可以认为 X_1 与其他解释变量之间存在多重共线性。

同理对其他解释变量进行检验，最终确定多重共线性主要由变量 X_1 与 X_2 引起。

表4-2 X_1 对其余各解释变量回归的OLS估计结果

Dependent Variable: X1					
Method: Least Squares					
Date: 02/22/22 Time: 21:34					
Sample: 2016M01 2021M12					
Included observations: 72					
Variable	Coefficient	Std. Error	t-Statistic	Prob.	
C	30.82422	8.046138	3.830934	0.0003	
X2	0.951858	0.044125	21.57174	0.0000	
X3	-0.102855	0.011187	-9.194112	0.0000	
X4	0.880716	0.212394	4.146611	0.0001	
LOG(X5)	-0.690044	0.667831	-1.033262	0.3054	
LOG(X6)	-0.298695	0.476621	-0.626692	0.5331	
LOG(X7)	0.045243	0.111787	0.404726	0.6870	
LOG(X8)	-0.629406	0.363323	-1.732358	0.0880	
R-squared	0.921222	Mean dependent var		102.0014	
Adjusted R-squared	0.912606	S.D. dependent var		1.118285	
S.E. of regression	0.330593	Akaike info criterion		0.728584	
Sum squared resid	6.994685	Schwarz criterion		0.981547	
Log likelihood	18.22901	Hannan-Quinn criter.		0.829289	
F-statistic	106.9157	Durbin-Watson stat		0.705862	
Prob(F-statistic)	0.000000				

➤ 采用偏相关系数法检验多重共线性的表现形式

为判断 X_1 与 X_2 是否存在共线性，构造检验统计量 t_1 ：

$$t_1 = \sqrt{n - k - 1} \frac{\rho_{12}}{\sqrt{1 - \rho_{12}^2}} \sim t(n - k - 1)$$

其中， ρ_{12} 为解释变量 X_1 与 X_2 的偏相关系数。在Eviews中计算偏相关系数的步骤如下：

- 估计 X_1 关于其余解释变量的辅助回归，得到模型残差 e_1 （“Proc” → “Make Residual Series”）；
- 估计 X_2 关于其余解释变量的辅助回归，得到模型残差 e_2 ；
- 计算 e_1 与 e_2 之间的相关系数，即为解释变量 X_1 与 X_2 的偏相关系数 ρ_{12} 。

按照上述步骤，计算得到 X_1 与其他各解释变量之间的偏相关系数与对应的t值，如下表所示。 $t_{0.025}(63) = 1.998$ ，因此认为 X_2 、 X_3 、 X_4 与 X_1 有共线性。

表4-3 X_1 与其余各解释变量之间的偏相关系数与t值

	X_2	X_3	X_4	$\log(X_5)$	$\log(X_6)$	$\log(X_7)$	$\log(X_8)$
偏相关系数	-0.94	0.75	-0.46	0.13	0.08	-0.05	0.21
t值	-21.40	9.12	-4.11	1.03	0.62	-0.40	1.72

对变量 X_2 的检验与 X_1 类似，发现 X_1 、 X_3 、 X_4 与 X_2 有共线性。

例4.1 对影响股票价格指数宏观经济因素的实证分析

➤ 多重共线性的处理

1. 删除变量法

分别删除 X_1 、 X_2 、 X_3 、 X_4 ，再进行回归得到的结果如下：

表4-4 删除 X_1 后的回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.111721	0.886186	5.768226	0.0000
X2	-0.022914	0.004860	-4.714988	0.0000
X3	0.011200	0.001232	9.089835	0.0000
X4	-0.076400	0.023393	-3.265957	0.0018
LOG(X5)	0.235520	0.073554	3.202020	0.0021
LOG(X6)	0.106580	0.052494	2.030315	0.0465
LOG(X7)	0.109369	0.012312	8.883139	0.0000
LOG(X8)	-0.186025	0.040016	-4.648790	0.0000

R-squared	0.857632	Mean dependent var	8.041658
Adjusted R-squared	0.842061	S.D. dependent var	0.091619
S.E. of regression	0.036411	Akaike info criterion	-3.683458
Sum squared resid	0.084848	Schwarz criterion	-3.430495
Log likelihood	140.6045	Hannan-Quinn criter.	-3.582752
F-statistic	55.07712	Durbin-Watson stat	0.874544
Prob(F-statistic)	0.000000		

表4-5 删除 X_2 后的回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.932419	1.033227	4.773800	0.0000
X1	-0.014357	0.005259	-2.730122	0.0082
X3	0.009115	0.001324	6.883015	0.0000
X4	-0.077191	0.027238	-2.833912	0.0061
LOG(X5)	0.210770	0.080467	2.619319	0.0110
LOG(X6)	0.104030	0.057713	1.802534	0.0762
LOG(X7)	0.111666	0.013508	8.266787	0.0000
LOG(X8)	-0.197246	0.043993	-4.483622	0.0000

R-squared	0.828189	Mean dependent var	8.041658
Adjusted R-squared	0.809397	S.D. dependent var	0.091619
S.E. of regression	0.039999	Akaike info criterion	-3.495474
Sum squared resid	0.102396	Schwarz criterion	-3.242511
Log likelihood	133.8371	Hannan-Quinn criter.	-3.394769
F-statistic	44.07165	Durbin-Watson stat	0.742448
Prob(F-statistic)	0.000000		

例4.1 对影响股票价格指数宏观经济因素的实证分析

➤ 多重共线性的处理

1. 删除变量法

分别删除 X_1 、 X_2 、 X_3 、 X_4 ，再进行回归得到的结果如下：

表4-6 删除 X_3 后的回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.153560	1.367921	3.767439	0.0004
X1	-0.037718	0.012840	-2.937410	0.0046
X2	0.020477	0.012754	1.605546	0.1133
X4	-0.088589	0.036974	-2.395962	0.0195
LOG(X5)	0.264837	0.105367	2.513481	0.0145
LOG(X6)	0.144194	0.074744	1.929171	0.0581
LOG(X7)	0.100823	0.017491	5.764383	0.0000
LOG(X8)	-0.225892	0.056961	-3.965734	0.0002

R-squared	0.712582	Mean dependent var	8.041658
Adjusted R-squared	0.681146	S.D. dependent var	0.091619
S.E. of regression	0.051735	Akaike info criterion	-2.980934
Sum squared resid	0.171295	Schwarz criterion	-2.727971
Log likelihood	115.3136	Hannan-Quinn criter.	-2.880229
F-statistic	22.66747	Durbin-Watson stat	0.517469
Prob(F-statistic)	0.000000		

表4-7 删除 X_4 后的回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.943839	1.021105	3.862324	0.0003
X1	0.025996	0.012796	2.031583	0.0464
X2	-0.053886	0.013769	-3.913444	0.0002
X3	0.015582	0.001941	8.028988	0.0000
LOG(X5)	0.165894	0.073920	2.244222	0.0283
LOG(X6)	0.182262	0.053596	3.400695	0.0012
LOG(X7)	0.121687	0.012489	9.743751	0.0000
LOG(X8)	-0.188174	0.042066	-4.473317	0.0000

R-squared	0.843967	Mean dependent var	8.041658
Adjusted R-squared	0.826901	S.D. dependent var	0.091619
S.E. of regression	0.038118	Akaike info criterion	-3.591805
Sum squared resid	0.092992	Schwarz criterion	-3.338842
Log likelihood	137.3050	Hannan-Quinn criter.	-3.491100
F-statistic	49.45286	Durbin-Watson stat	0.923436
Prob(F-statistic)	0.000000		

可以看到，删除 X_1 对回归方程拟合优度造成的下降最小，因此，先删除 X_1 。再考虑分别删除 X_2 、 X_3 、 X_4 ，再进行回归得到的结果如下，同理，选择进一步删除变量 X_4 。

表4-8 删除 X_1 、 X_2 的回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.570943	0.948774	3.763747	0.0004
X3	0.009693	0.001371	7.072147	0.0000
X4	-0.109692	0.025687	-4.270337	0.0001
LOG(X5)	0.198743	0.084241	2.359219	0.0213
LOG(X6)	0.110889	0.060453	1.834288	0.0712
LOG(X7)	0.113449	0.014146	8.019966	0.0000
LOG(X8)	-0.191439	0.046071	-4.155307	0.0001
R-squared	0.808179	Mean dependent var	8.041658	
Adjusted R-squared	0.790473	S.D. dependent var	0.091619	
S.E. of regression	0.041938	Akaike info criterion	-3.413087	
Sum squared resid	0.114321	Schwarz criterion	-3.191745	
Log likelihood	129.8711	Hannan-Quinn criter.	-3.324970	
F-statistic	45.64298	Durbin-Watson stat	0.650612	
Prob(F-statistic)	0.000000			

表4-9 删除 X_1 、 X_3 后的回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.396988	1.300477	2.612109	0.0112
X2	-0.011456	0.007049	-1.625168	0.1090
X4	-0.145872	0.033206	-4.392964	0.0000
LOG(X5)	0.320194	0.109582	2.921951	0.0048
LOG(X6)	0.181364	0.077868	2.329130	0.0230
LOG(X7)	0.093683	0.018309	5.116727	0.0000
LOG(X8)	-0.210700	0.059962	-3.513883	0.0008
R-squared	0.673833	Mean dependent var	8.041658	
Adjusted R-squared	0.643725	S.D. dependent var	0.091619	
S.E. of regression	0.054686	Akaike info criterion	-2.882239	
Sum squared resid	0.194389	Schwarz criterion	-2.660897	
Log likelihood	110.7606	Hannan-Quinn criter.	-2.794122	
F-statistic	22.38073	Durbin-Watson stat	0.432498	
Prob(F-statistic)	0.000000			

表4-10 删除 X_1 、 X_4 后的回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.829677	0.945277	5.109272	0.0000
X2	-0.027705	0.004966	-5.579217	0.0000
X3	0.012514	0.001248	10.02692	0.0000
LOG(X5)	0.168146	0.075669	2.222115	0.0298
LOG(X6)	0.158837	0.053585	2.964211	0.0042
LOG(X7)	0.119752	0.012748	9.393492	0.0000
LOG(X8)	-0.200268	0.042633	-4.697532	0.0000
R-squared	0.833905	Mean dependent var	8.041658	
Adjusted R-squared	0.818573	S.D. dependent var	0.091619	
S.E. of regression	0.039025	Akaike info criterion	-3.557087	
Sum squared resid	0.098989	Schwarz criterion	-3.335745	
Log likelihood	135.0551	Hannan-Quinn criter.	-3.468970	
F-statistic	54.39025	Durbin-Watson stat	0.832020	
Prob(F-statistic)	0.000000			



例4.1 对影响股票价格指数宏观经济因素的实证分析

删除 X_1 与 X_4 的模型是：

$$\ln Y_t = 4.83 - 0.03X_{2t} + 0.01X_{3t} + 0.17\ln X_{5t} + 0.16\ln X_{6t} + 0.12\ln X_{7t} - 0.20\ln X_{8t}$$

此时，采用辅助回归法并应用克里安的经验法则，可以认为所有变量之间不存在共线性。

例4.2 我国进口支出与国民生产总值和消费者价格指数的关系

选择我国1995年至2020年宏观经济变量：国民生产总值（ GNP ）、进口总额（ IM ）、居民消费价格指数（ CPI ，1978=100）。构建多元线性回归模型，检验我国进口需求与国民生产总值和消费者价格指数的关系。多元回归模型为： $IM = \beta_0 + \beta_1 GNP + \beta_2 CPI + \mu$

Equation Estimation

Specification Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$.

im c cpi gnp

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 1995 2020

确定 取消

例4.2 我国进口支出与国民生产总值和消费者价格指数的关系

对原始模型的OLS回归结果如表4-13所示。模型的 R^2 值比较高，为0.935，但国民生产总值（*GNP*）的系数为负且不显著，与经济理论不符，说明模型中可能存在严重的多重共线性问题。

表4-13 OLS回归结果

Dependent Variable: IM
Method: Least Squares
Date: 02/23/22 Time: 16:32
Sample: 1995 2020
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-281290.9	81154.98	-3.466095	0.0021
CPI	721.8054	197.0078	3.663842	0.0013
GNP	-0.060825	0.056313	-1.080122	0.2913
R-squared	0.935263			
Adjusted R-squared	0.929634			
S.E. of regression	12658.13			
Sum squared resid	3.69E+09			
Log likelihood	-280.8960			
F-statistic	166.1431			
Prob(F-statistic)	0.000000			
Mean dependent var		71844.13		
S.D. dependent var		47718.73		
Akaike info criterion		21.83815		
Schwarz criterion		21.98332		
Hannan-Quinn criter.		21.87996		
Durbin-Watson stat		0.654848		

例4.2 我国进口支出与国民生产总值和消费者价格指数的关系

➤ 多重共线性的检验

计算解释变量之间的相关系数

与例4.1的步骤类似，在Eviews软件的命令窗口键入：**COR CPI GNP**。

计算结果显示，国民生产总值（*GNP*）与居民消费价格指数（*CPI*）的相关系数为0.9898，表明模型中确实存在严重的多重共线性。

G Group: UNTITLED Workfile: EX4.2::Untitled\			
View	Proc	Object	Print Name Freeze Sample Sheet Stats S
Correlation			
	GNP	CPI	
GNP	1.000000	0.989781	
CPI	0.989781	1.000000	

图4-2 相关系数矩阵

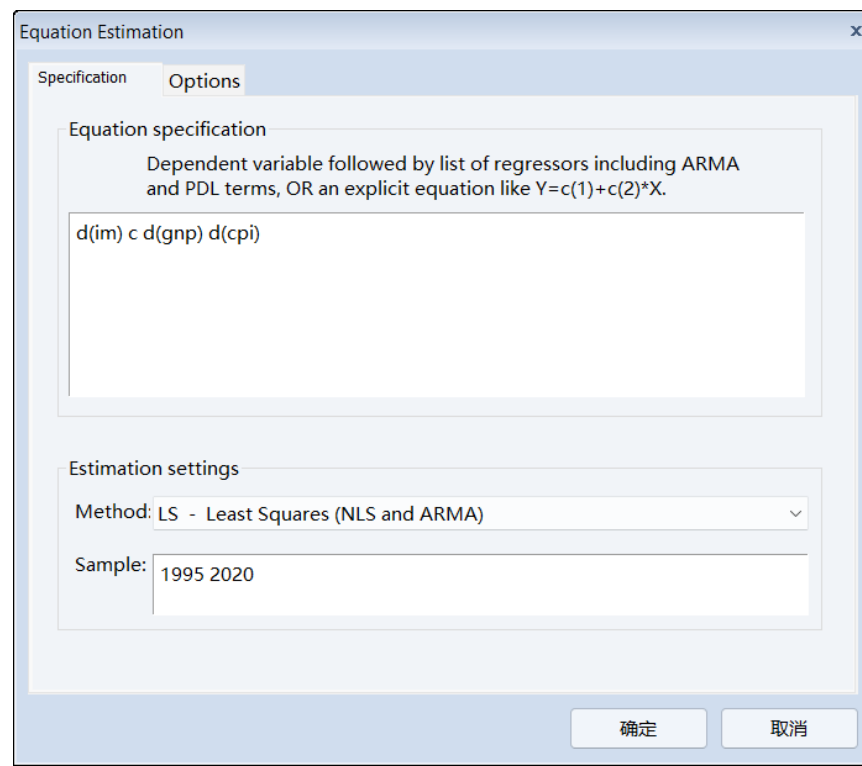


例4.2 我国进口支出与国民生产总值和消费者价格指数的关系

➤ 多重共线性的修正

1. 差分法

对所有变量取差分形式后进行OLS估计：



Equation Estimation

Specification Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$.

d(im) c d(gnp) d(cpi)

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 1995 2020

确定 取消

例4.2 我国进口支出与国民生产总值和消费者价格指数的关系

对变量进行差分后， $D(GNP)$ 与 $D(CPI)$ 的相关系数为0.471，在一定程度上降低了多重共线性的程度。差分模型的OLS回归结果如表4-14所示，此时国民生产总值（ GNP ）的系数在10%置信水平下显著为正。

表4-14 差分模型的OLS回归结果

Dependent Variable: D(IM)				
Method: Least Squares				
Date: 02/23/22 Time: 16:44				
Sample (adjusted): 1996 2020				
Included observations: 25 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1325.888	3098.050	-0.427975	0.6728
D(GNP)	0.139351	0.074255	1.876667	0.0739
D(CPI)	114.6693	184.9305	0.620067	0.5416
R-squared	0.226088	Mean dependent var	5275.531	
Adjusted R-squared	0.155733	S.D. dependent var	9076.269	
S.E. of regression	8339.641	Akaike info criterion	21.00759	
Sum squared resid	1.53E+09	Schwarz criterion	21.15386	
Log likelihood	-259.5949	Hannan-Quinn criter.	21.04816	
F-statistic	3.213506	Durbin-Watson stat	1.429661	
Prob(F-statistic)	0.059649			

例4.2 我国进口支出与国民生产总值和消费者价格指数的关系

➤ 多重共线性的修正

2. 变量变换

将国民生产总值（ GNP ）和进口总额（ IM ）分别除以居民消费价格指数（ CPI ），得到实际国民生产总值（ $RGNP$ ）和实际进口总额（ RIM ），构建新的回归模型： $RIM = \alpha_0 + \alpha_1 RGNP + \varepsilon$

Equation Estimation

Specification Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$.

im/cpi*100 c gnp/cpi*100

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 1995 2020

确定 取消

例4.2 我国进口支出与国民生产总值和消费者价格指数的关系

➤ 多重共线性的修正

2. 变量变换

将名义变量转化为实际变量后，原模型的多重共线性消除。模型的OLS回归结果如表4-15所示，此时实际国民生产总值（*RGNP*）的系数在1%置信水平下显著为正。

表4-15 差分模型的OLS回归结果

Dependent Variable: IM/CPI*100 ← 实际进口总额（*RIM*）
Method: Least Squares
Date: 02/23/22 Time: 17:02
Sample: 1995 2020
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2992.468	1038.911	2.880389	0.0082
GNP/CPI*100	0.143511	0.012853	11.16586	0.0000
R-squared				
0.838576		Mean dependent var		12682.18
Adjusted R-squared		S.D. dependent var		7102.674
0.831850		Akaike info criterion		18.86524
S.E. of regression		Schwarz criterion		18.96201
2912.532		Hannan-Quinn criter.		18.89310
Sum squared resid		Durbin-Watson stat		0.243271
2.04E+08				
Log likelihood				
-243.2481				
F-statistic				
124.6765				
Prob(F-statistic)				
0.000000				

实际国民生产总值 →
（*RGNP*）



第四章 多重共线性和虚拟变量的应用

第五节 虚拟变量模型

- 一、虚拟变量的性质和设置原则
- 二、虚拟变量模型的运用

一、虚拟变量的性质和设置原则

(一) 虚拟变量的性质

定量变量：可以直接度量数量

定性变量：某一性质、属性是否出现

表示
虚拟变量 D

1, 表示某一性质出现 (不出现)

0, 表示某一性质不出现 (出现)

(dummy variables)

例: $D_1 \begin{cases} 1, \text{男性} \\ 0, \text{女性} \end{cases}$

$D_2 \begin{cases} 1, \text{吸烟} \\ 0, \text{不吸烟} \end{cases}$

$D_3 \begin{cases} 1, \text{战争时期} \\ 0, \text{非战争时期} \end{cases}$

虚拟变量和定量变量在回归模型中的应用相同。

方差分析模型(analysis of variance model): 解释变量全部为虚拟变量

协方差分析模型(analysis of covariance model): 解释变量同时包含定量变量与虚拟变量



一、虚拟变量的性质和设置原则

(一) 虚拟变量的性质

例4.3 目前在我国上市公司中，主要是国家或法人作为公司的第一大股东。总体而言，国家作为第一大股东的上市公司的绩效要低于法人作为第一大股东的上市公司的绩效。为验证该结论，建立模型：

$$R_i = \alpha + \beta D_i + \mu_i$$

其中， R_i 为每股收益，代表公司绩效。

$$D_i = \begin{cases} 1, & \text{国家是公司} i \text{的第一大股东} \\ 0, & \text{法人是公司} i \text{的第一大股东} \end{cases}$$

国家为第一大股东公司的平均每股收益 $E(R_i | D_i = 1) = \alpha + \beta$

法人为第一大股东公司的平均每股收益 $E(R_i | D_i = 0) = \alpha$

β 代表了国家、法人分别为公司第一大股东时平均每股收益的差距。



一、虚拟变量的性质和设置原则

（二）虚拟变量的设置原则

虚拟变量的引入方式包括：

（1）加法方式。 $Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 D_t + \varepsilon_t$

（2）乘法方式。 $Y_t = \alpha_0 + (\alpha_1 + \alpha_{11} D_t) X_t + \varepsilon_t$

（3）同时以加法方式和乘法方式引入。

$$Y_t = \alpha_0 + (\alpha_1 + \alpha_{11} D_{1t}) X_t + \alpha_2 D_{2t} + \varepsilon_t$$

若一个定性变量有 m 个类别，则仅要引入 $m - 1$ 个虚拟变量。若引入 m 个虚拟变量，则解释变量之间存在完全多重共线性。



二、虚拟变量模型的运用

（一）虚拟变量模型在调整季节波动中的应用

例4.4 利用虚拟变量进行季节调整（seasonal adjustment）

设可用如下模型对以季度数据表示的通货膨胀率进行回归：

$$R_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \mu_t \quad (4.8)$$

其中， R_t 表示通货膨胀率， X_{1t}, \dots, X_{pt} 表示影响通货膨胀率的因素。

（1）为研究四个季度对 R_t 的影响，引入三个虚拟变量，若以第一季度为基准季节， $D_i = \begin{cases} 1, & \text{对第 } i+1 \text{ 个季度} \\ 0, & \text{对其他季度} \end{cases}, i = 1, 2, 3$

则模型变为：

$$R_t = \beta_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \mu_t$$

其中， $\alpha_1, \alpha_2, \alpha_3$ 分别体现了第二、三、四季度相对于第一季度对 R_t 的影响。



二、虚拟变量模型的运用

(一) 虚拟变量模型在调整季节波动中的应用

(2) 利用最小二乘法回归模型，得到 $\hat{\alpha}_1$ ， $\hat{\alpha}_2$ ， $\hat{\alpha}_3$ ，若 $\hat{\alpha}_1$ ， $\hat{\alpha}_2$ ， $\hat{\alpha}_3$ 显著，则表明季节因素对 R_t 有显著影响，需要进行季节调整。

(3) 当需要进行季节调整时，将第二、三、四季度的数据分别减去 $\hat{\alpha}_1$ ， $\hat{\alpha}_2$ ， $\hat{\alpha}_3$ ，即为季节调整后的数据。

二、虚拟变量模型的运用

（二）虚拟变量模型在分段线性回归中的应用

在金融理论中，常常会出现一种情况：当某影响因素越过某一临界值，或时间过了某一临界点之后，因变量对影响因素的变化率将发生变化，在图形中就表现为斜率不同的两段连续折线。

分段线性回归：对构成折线的数据的回归。

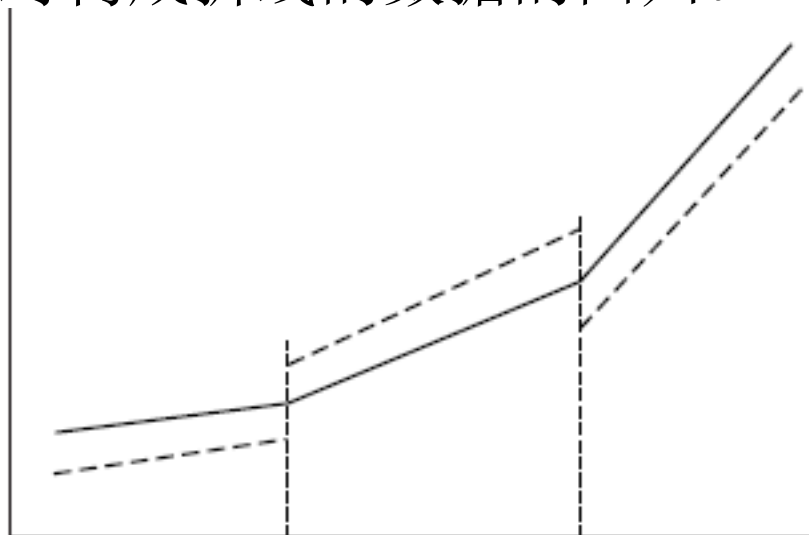


图4-3 有两个转折点的联系折线



二、虚拟变量模型的运用

（二）虚拟变量模型在分段线性回归中的应用

利用虚拟变量模型进行分段线性回归。假设因变量 Y_t 仅受唯一解释变量 X_t 的影响。存在 t^* ，当 $t \geq t^*$ 时， Y_t 对 X_t 的变化率将发生变换。建立如下回归模型：

$$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 (X_t - X_t^*) D_t + \varepsilon_t$$

其中，
$$D_t = \begin{cases} 1, & t \geq t^* \\ 0, & t < t^* \end{cases}$$

结构发生转折时间前 D_t 为0， $E(Y_t) = \alpha_0 + \alpha_1 X_t$ ；

结构发生转折时间后 D_t 为1， $E(Y_t) = (\alpha_0 - \alpha_2 X_t^*) + (\alpha_1 + \alpha_2) X_t$ ；

在时刻 t^* ，无论 D_t 为0还是1， $E(Y_t^*) = \alpha_0 + \alpha_1 X_t$



二、虚拟变量模型的运用

(三) 利用虚拟变量模型对平行数据进行混合回归

平行数据：是时间数据和横截面数据的有机结合，即多个个体同样变量的时间序列数据按照一定顺序排列得到的数据集。

例4.5 假定要研究某一类型上市公司资本结构与影响因素之间的关系，以总负债率(Y)代表资本结构。

影响因素：股权结构(X_1)，公司治理结构(X_2)，成长性(X_3)

假设样本公司只有两家，若对每家公司分别应用普通最小二乘回归，两方程的随机误差项可能同期相关。

可以利用虚拟变量建立如下模型：

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \beta_0 D_t + \beta_1 D_t X_{1t} + \beta_2 D_t X_{2t} + \beta_3 D_t X_{3t} + \mu_t \quad (4.9)$$

其中， $D_t = \begin{cases} 1, & \text{公司乙} \\ 0, & \text{公司甲} \end{cases}$



二、虚拟变量模型的运用

(三) 利用虚拟变量模型对平行数据进行混合回归

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \beta_0 D_t + \beta_1 D_t X_{1t} + \beta_2 D_t X_{2t} + \beta_3 D_t X_{3t} + \mu_t \quad (4.9)$$

其中, $D_t = \begin{cases} 1, & \text{公司乙} \\ 0, & \text{公司甲} \end{cases}$

对上述模型应用普通最小二乘法进行回归。

对于 F 检验, $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$,

若不能拒绝原假设, 则说明甲、乙公司的回归模型没有显著差别, 可以合并;

若拒绝上述假设, 同时不能拒绝假设 $\beta_i = 0$ ($i = 1, 2, 3$), 则说明甲、乙两公司 Y_t 对于 X_i 的变化率没有显著差别, 可以去掉 $\beta_i D_t X_{it}$ 项。
通过应用虚拟变量模型, 相当于把不同属性类型的样本合并。



第四章 多重共线性和虚拟变量的应用

第六节 回归模型的结构稳定性检验——邹氏检验

一、邹氏检验的过程



第六节 回归模型的结构稳定性检验——邹氏检验

在金融模型中，各变量相互之间的关系有时会因某些外部冲击或自身性质的改变而发生**结构性的变化**。

若仍然同时利用变化前后的数据进行回归分析，模型的精确度将大大降低。因此，有必要对回归模型进行结构稳定性检验。

邹氏检验（Chow's test）是检验模型结构稳定性的一种较为简单的方法。



一、邹氏检验的过程

理论前提：在可能发生的结构变化前后，随机误差项具有相同的方差；随机误差项满足独立正态分布。

检验步骤：

(1) 将数据以可能发生结构变化的点为界分为两部分。分别利用全部数据、两分样本对模型进行回归，并获得三次回归的残差平方和。

(2) 对整段数据回归得到的残差平方和大于对两分样本进行回归得到的残差平方和之和，可建立如下的 F 检验：

$$F = \frac{[RSS - (RSS_1 + RSS_2)]/k}{(RSS_1 + RSS_2)/(T - 2k)} \sim F(k, T - 2k)$$

其中， RSS 为对全部数据回归的残差平方和， RSS_1 、 RSS_2 表示对两分样本回归得到的残差平方和， T 为样本数， k 为所估计参数的个数。

(3) 查表求得在一定显著性水平下的 F 临界值。如果第二步计算出的 F 值大于临界 F 值，则拒绝模型结构稳定的假设；如果小于临界 F 值，则不能拒绝模型结构稳定性假设。



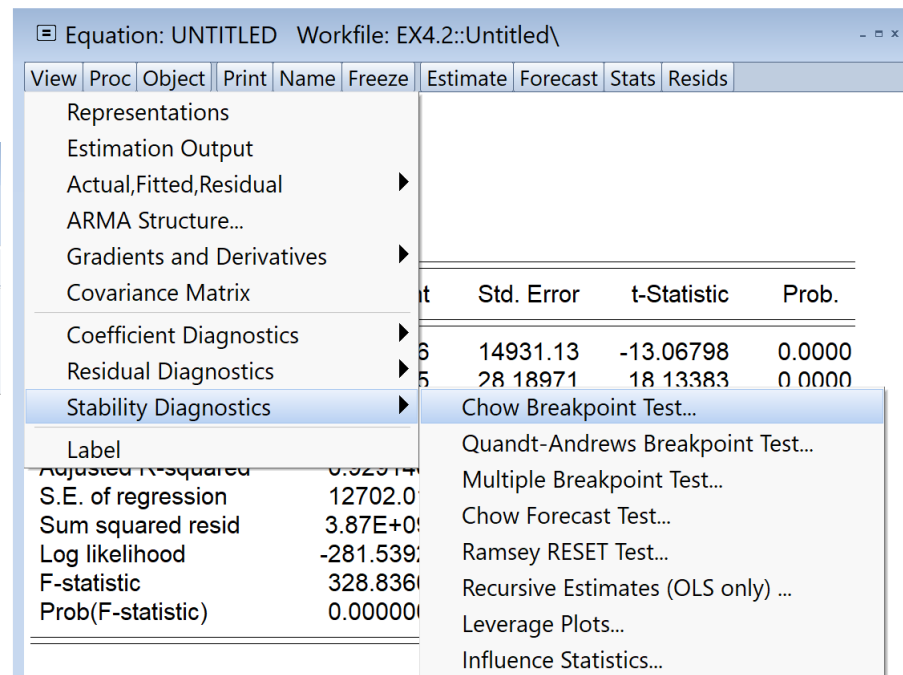
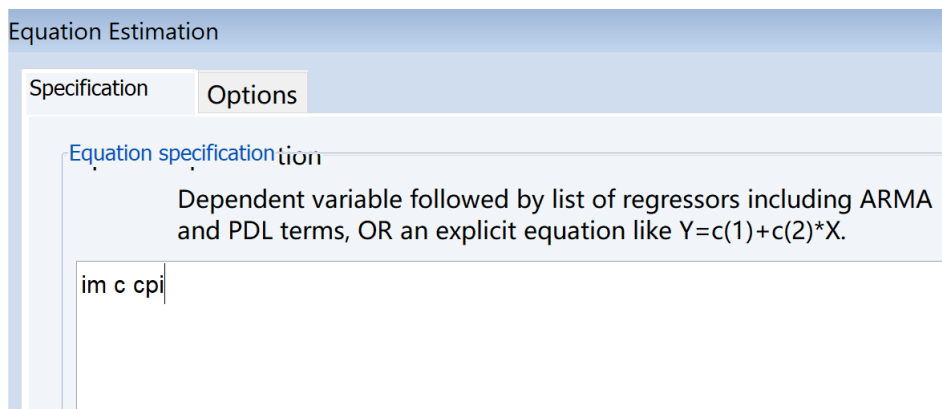
一、邹氏检验的过程

应用邹氏检验的过程中应注意以下几点：

- （1）必须满足前提假设条件。
- （2）邹氏检验仅仅告诉我们模型结构是否稳定，而不能告诉我们如果结构不稳定，到底是截距还是斜率抑或两者都发生了变化。
- （3）邹氏检验需要知道结构可能发生的时间点。

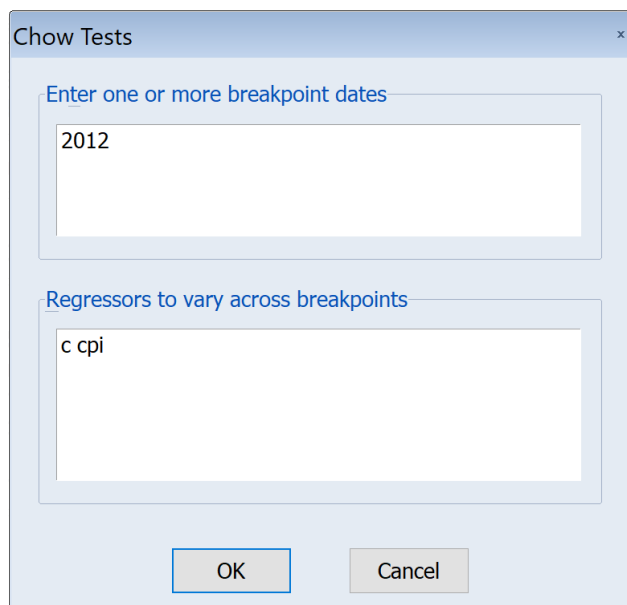
二、在Eviews软件中使用邹氏检验

回归结果视图里面点击“View” —> “Stability tests” —> “Chow Breakpoint Test”，例4.2（breakpoint 2012）



二、在Eviews软件中使用邹氏检验

回归结果视图里面点击“View” —> “Stability tests” —> “Chow Breakpoint Test”，例4.2（breakpoint 2012）



The image shows the 'Chow Tests' dialog box in EViews. It has two main sections. The first section, 'Enter one or more breakpoint dates', contains a text box with '2012' entered. The second section, 'Regressors to vary across breakpoints', contains a text box with 'c cpi' entered. At the bottom, there are 'OK' and 'Cancel' buttons.

Chow Tests

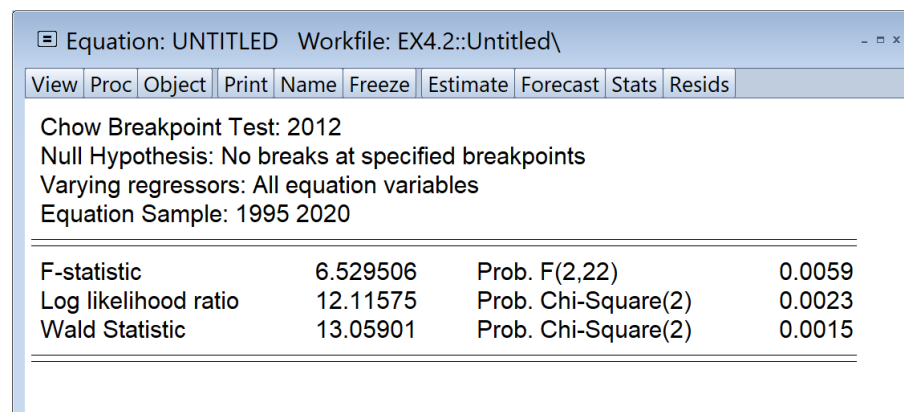
Enter one or more breakpoint dates

2012

Regressors to vary across breakpoints

c cpi

OK Cancel



The image shows the 'Equation: UNTITLED' window in EViews. It displays the results of a Chow Breakpoint Test for the year 2012. The window includes a menu bar with 'View', 'Proc', 'Object', 'Print', 'Name', 'Freeze', 'Estimate', 'Forecast', 'Stats', and 'Resids'. The main text area shows the test name, null hypothesis, varying regressors, and equation sample. Below this is a table with test statistics and p-values.

Equation: UNTITLED Workfile: EX4.2::Untitled\

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Chow Breakpoint Test: 2012
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1995 2020

F-statistic	6.529506	Prob. F(2,22)	0.0059
Log likelihood ratio	12.11575	Prob. Chi-Square(2)	0.0023
Wald Statistic	13.05901	Prob. Chi-Square(2)	0.0015



第四章 多重共线性和虚拟变量的应用

第七节 回归模型的结构稳定性检验——虚拟变量法



第七节 回归模型的结构稳定性检验——虚拟变量法

对于一元线性模型 $Y_t = \alpha_0 + \alpha_1 X_t + \mu_t$ ，假定在时刻 t^* ，由于外部事件冲击，模型结构可能发生了变化。为验证该观点，建立如下虚拟变量模型：

$$Y_t = \alpha_0 + \beta_0 D_t + \alpha_1 X_t + \beta_1 X_t D_t + \mu_t \quad (4.10)$$

其中， $D_t = \begin{cases} 1, t \geq t^* \\ 0, t < t^* \end{cases}$ 。

发生结构变化前： $E(Y_t|D_t = 0, X_t) = \alpha_0 + \alpha_1 X_t$

可能的结构变化发生后： $E(Y_t|D_t = 1, X_t) = \alpha_0 + \beta_0 + (\alpha_1 + \beta_1)X_t$

β_0 ：极差截距(differential intercept)

β_1 ：极差斜率系数(differential slope coefficient)

利用全部数据对上述模型进行最小二乘回归，对参数估计值进行显著性检验：

	$\hat{\beta}_0$ 显著	$\hat{\beta}_0$ 不显著
$\hat{\beta}_1$ 显著	截距、斜率均发生变化	截距未变化，斜率发生变化
$\hat{\beta}_1$ 不显著	截距发生变化，斜率未变化	模型结构稳定



第四章 多重共线性和虚拟变量的应用

第八节 实例——虚拟变量在金融数据处理中的作用

一、简单理论回顾

二、实证检验



例4.6 虚拟变量在金融数据处理中的作用

一、简单理论回顾

有效市场理论(Efficient Market Hypothesis, EMH):

在有效市场中, 由于股票价格能够及时地反映所有的信息, 因此股价会呈现出随机波动的特征。

各国学者利用各国的证券市场数据对有效市场理论的实证检验, 结果发现许多与该理论相背离的现象, **周内效应**就是其中之一。

在股票市场中, 一周内的某些交易日的收益率显著高于或低于其他交易日的收益率。



例4.6 虚拟变量在金融数据处理中的作用

二、实证检验

下面我们将利用虚拟变量模型对这一现象进行实证检验：

数据描述：我们利用的是上海股票市场上证指数2016年1月4日到2021年12月31日的日收盘价数据，共1461个观测值。收益率的计算采用的是连续收益率法，计算公式如下：

$$R_t = \ln P_t - \ln P_{t-1}$$

其中， P_t 为 t 时期的收盘价， P_{t-1} 为 $t-1$ 时期的收盘价。

建立如下的虚拟变量模型：

$$R_t = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{4t} + \alpha_5 D_{5t} + \mu_t \quad (4.11)$$

其中，虚拟变量 $D_{it} = \begin{cases} 1, t \text{ 为每周的第 } i \text{ 天} \\ 0, \text{其他} \end{cases} (i = 1, 2, 4, 5)。$

α_0 表示周三的平均收益， α_1 、 α_2 、 α_4 、 α_5 分别表示周一、周二、周四、周五与周三的平均收益的差异。

例4.6 虚拟变量在金融数据处理中的作用

导入股票日数据： 在eviews中点击File---Open---Foreign Data as workfile在弹出的对话框中选中数据文件ex4.6.xls,下面就是打开的对话框页面。单击完成。

Excel 97-2003 Read - Step 1 of 3

Cell Range

☒ Predefined range

Sheet: Sheet1

Start cell: \$A\$1

☐ Custom range

Sheet1!\$A\$1:\$G\$1462

End cell: \$G\$1462

date	r	weekday	D1	D2	D4	D5
2016-01-04	-0.0711077413149788	1	1	0	0	0
2016-01-05	-0.00259630772901431	2	0	1	0	0
2016-01-06	0.0222968651803601	3	0	0	0	0
2016-01-07	-0.0730535199459794	4	0	0	1	0
2016-01-08	0.0194605956955698	5	0	0	0	1
2016-01-11	-0.0547306742259757	1	1	0	0	0
2016-01-12	0.00203888926865581	2	0	1	0	0
2016-01-13	-0.0245351830761917	3	0	0	0	0
2016-01-14	0.0194901595989174	4	0	0	1	0
2016-01-15	-0.036113546624974	5	0	0	0	1
2016-01-18	0.00442550257073826	1	1	0	0	0
2016-01-19	0.0317180670537498	2	0	1	0	0
2016-01-20	-0.0103755119283511	3	0	0	0	0
2016-01-21	-0.0328557191302989	4	0	0	1	0
2016-01-22	0.0124481601009894	5	0	0	0	1
2016-01-25	0.00749865518098236	1	1	0	0	0

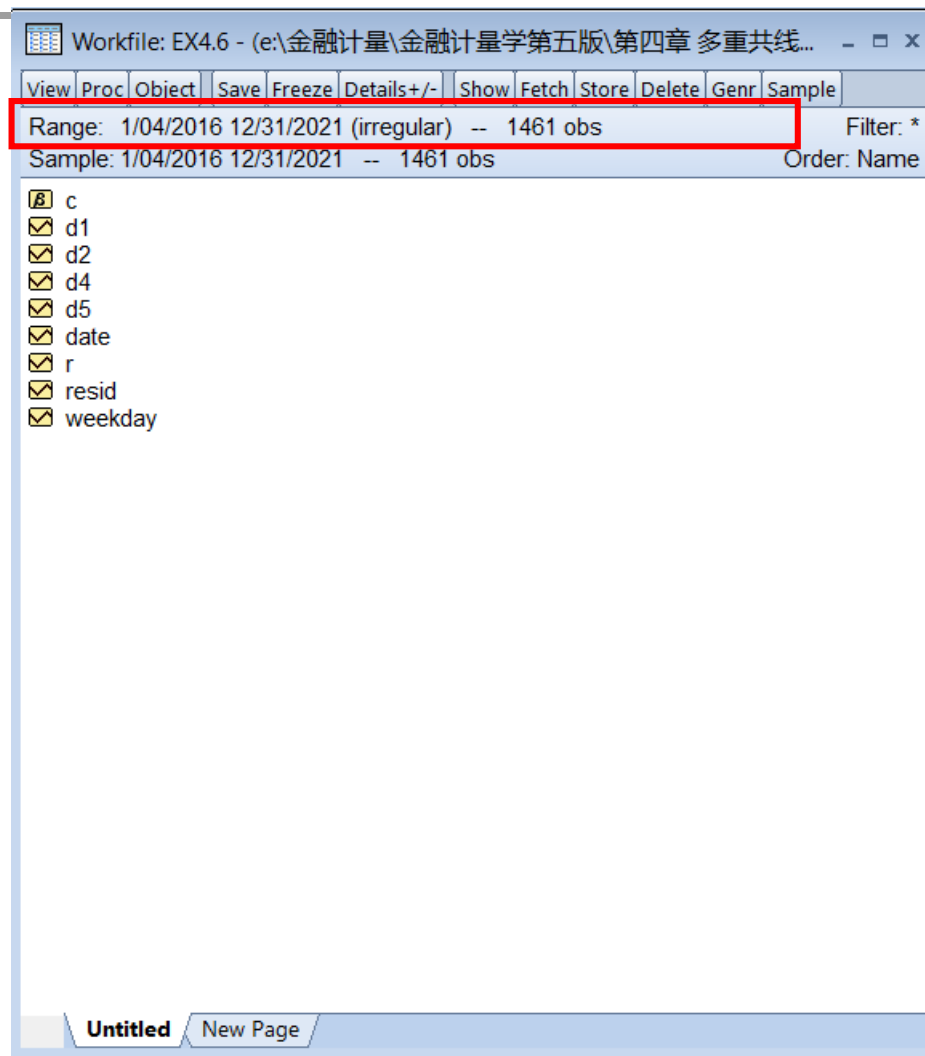
☐ Read series by row (transpose incoming data)

取消 < 上一步(B) 下一页(N) > 完成

例4.6 虚拟变量在金融数据处理中的作用

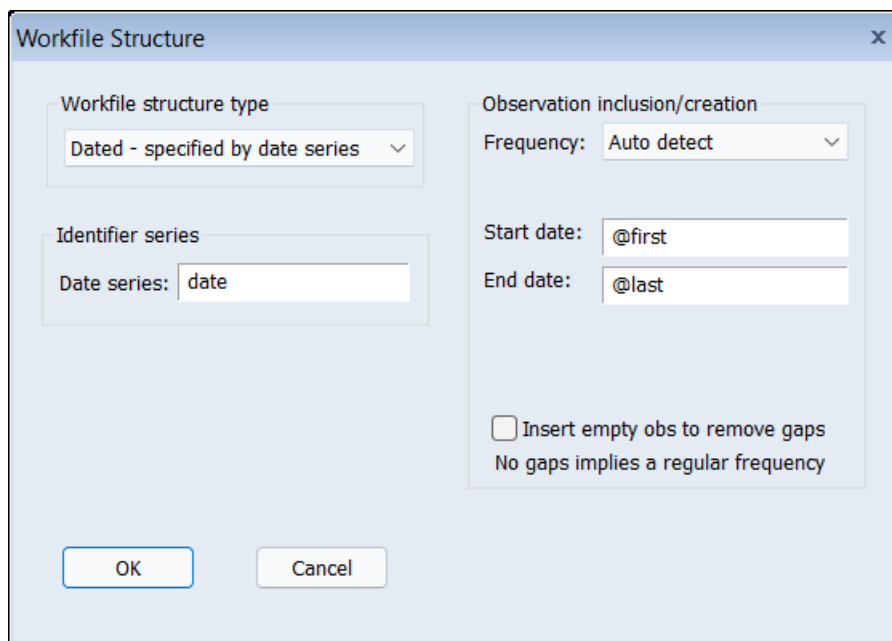
导入股票日数据:

双击workfile对话框的Range



例4.6 虚拟变量在金融数据处理中的作用

导入股票日数据：在弹出的对话框中选择Dated-specified by date series,这是eviews提供的处理非规范日期数据的工具。在Date Series 中输入你导入的数据中属于日期的列名。本例的日期列名为`date`，点击OK，至此完成股票数据的导入。



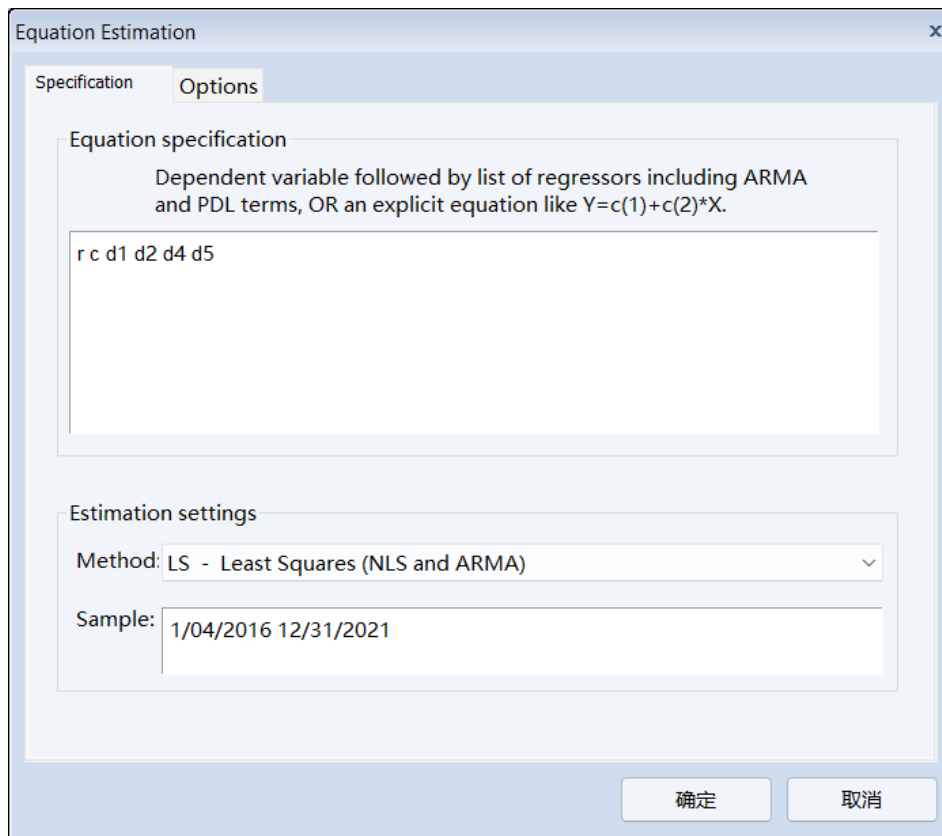
The screenshot shows the 'Workfile Structure' dialog box with the following settings:

- Workfile structure type:** Dated - specified by date series (selected in the dropdown)
- Identifier series:** Date series: date (entered in the text field)
- Observation inclusion/creation:**
 - Frequency: Auto detect (selected in the dropdown)
 - Start date: @first (entered in the text field)
 - End date: @last (entered in the text field)
 - ☐ Insert empty obs to remove gaps
No gaps implies a regular frequency
- Buttons:** OK and Cancel



例4.6 虚拟变量在金融数据处理中的作用

“Quick” → “Estimate Equation”



Equation Estimation

Specification Options

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$.

r c d1 d2 d4 d5

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 1/04/2016 12/31/2021

确定 取消

例4.6 虚拟变量在金融数据处理中的作用

表4-16 OLS估计结果

Dependent Variable: R
Method: Least Squares
Date: 02/17/22 Time: 20:38
Sample: 1/04/2016 12/31/2021
Included observations: 1461

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.000575	0.000660	-0.870664	0.3841
D1	0.001179	0.000945	1.247946	0.2123
D2	0.002089	0.000937	2.230421	0.0259
D4	-0.000601	0.000936	-0.641937	0.5210
D5	0.000330	0.000940	0.351425	0.7253
R-squared	0.006834	Mean dependent var	1.92E-05	
Adjusted R-squared	0.004106	S.D. dependent var	0.011419	
S.E. of regression	0.011395	Akaike info criterion	-6.107790	
Sum squared resid	0.189071	Schwarz criterion	-6.089696	
Log likelihood	4466.740	Hannan-Quinn criter.	-6.101040	
F-statistic	2.504829	Durbin-Watson stat	2.057879	
Prob(F-statistic)	0.040565			



例4.6 虚拟变量在金融数据处理中的作用

二、实证检验

使用Eviews软件对上述模型进行OLS回归，得到如下结果（括号内为相应的 t 值）：

$$R_t = -0.0006 + 0.0012D_{1t} + 0.0021D_{2t} - 0.0006D_{4t} + 0.0003D_{5t}$$

(-0.8707)	(1.2479)	(2.2304*)	(-0.6419)	(0.3514)
-----------	----------	-----------	-----------	----------

对模型各系数估计值进行联合 F 检验，看各系数值是否同时为零，结果的到 F 值为2.5048，其概率值为0.0406，因此在5%的显著性水平下拒绝各系数值同时为零的假设，则可以得出结论，上海股票市场存在周内效应。

D_{2t} 的系数为0.0023，其对应的 t 值为2.2304，概率值为0.0259，因此在5%的显著性水平下拒绝系数 α_2 为零的假设，表明星期二与星期三的平均收益率有显著差异。



本章小结

本章主要分为两部分，在第一部分中我们主要讨论了金融数据中存在的多重共线性现象。第二部分我们主要介绍了虚拟变量的应用。在接下来的部分中，我们主要介绍了如何进行模型的结构稳定性检验的两种方法：邹氏检验和虚拟变量法。



本章小结

- 多重共线性的含义
- 多重共线性产生的原因
- 多重共线性的后果
- 判断多重共线性的方法及其修正方法
- 虚拟变量的设置原则
- 虚拟变量模型的应用
- 邹氏检验的做法及缺陷
- 虚拟变量法检验结构稳定性的优点