

第五章 经典单方程计量经济学模型： 专门问题

§ 5.1 虚拟变量模型**

§ 5.2 滞后变量模型

§ 5.3 模型设定误差*

§ 5.1 虚拟变量模型

Dummy Variables Regression Models

- 一、虚拟变量的基本含义
- 二、虚拟变量的引入**
- 三、虚拟变量的设置原则*

一、虚拟变量的基本含义**

1、虚拟变量（dummy variables）

- 许多经济变量，**可以定量度量**；但某些影响经济变量的因素，是**无法定量度量的**。
- 为了在模型中能够反映这些因素的影响，并提高模型的精度，需要将它们“量化”。
- 这种“量化”通常是通过引入“虚拟变量”来完成的：根据这些因素的属性类型，**构造只取“0”或“1”的人工变量**，通常称为**虚拟变量**，记为D。
- 虚拟变量，只作为解释变量。

- 一般地，在虚拟变量的设置中：
 - 关注类型、肯定类型取值为1；
 - 基础类型，否定类型取值为0。
- 例如，反映文程度的虚拟变量可取为：
 - $D=1$ ，本科学历
 - $D=0$ ，非本科学历
- 虚拟变量能否取1、0以外的数值？

2、虚拟变量模型

- 同时含有一般解释变量与虚拟变量的模型称为**虚拟变量模型或者方差分析（analysis-of variance）ANOVA模型**。
- 例如，一个以性别为虚拟变量考察企业职工薪金
的模型：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \mu_i$$

其中： Y_i 为企业职工的薪金； X_i 为工龄；

$D_i=1$ ，若是男性； $D_i=0$ ，若是女性。

二、虚拟变量的引入

1、加法方式

- 虚拟变量作为解释变量引入模型有两种基本方式：
加法方式和乘法方式。
- 上述企业职工薪金模型中性别虚拟变量的引入采取了加法方式。

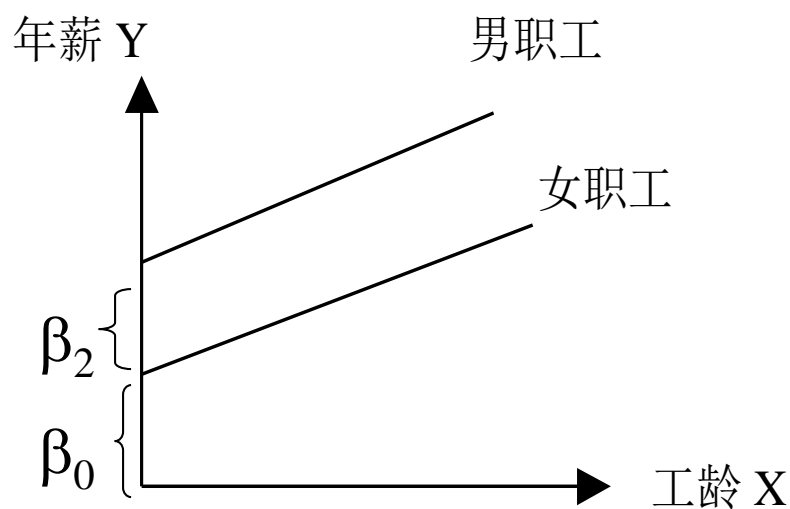
在该模型中，如果仍假定 $E(\mu_i)=0$ ，则企业男、女职工的平均薪金为：

$$E(Y_i | X_i, D_i = 1) = (\beta_0 + \beta_2) + \beta_1 X_i$$

$$E(Y_i | X_i, D_i = 0) = \beta_0 + \beta_1 X_i$$

假定 $\beta_2 > 0$ ，则两个函数有相同的斜率，但有不同的截距。意即，男女职工平均薪金对工龄的变化率是一样的，但两者的平均薪金水平相差 β_2 。

可以通过对 β_2 的统计显著性进行检验，以判断企业男女职工的平均薪金水平是否有显著差异。



- 将上例中的性别换成教育水平，教育水平考虑三个层次：高中以下、高中、大学及其以上。

$$D_1 = \begin{cases} 1 & \text{高中} \\ 0 & \text{其他} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{大学及其以上} \\ 0 & \text{其他} \end{cases}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_1 + \beta_3 D_2 + \mu_i$$

$$E(Y_i | X_i, D_1 = 0, D_2 = 0) = \beta_0 + \beta_1 X_i$$

高中以下

$$E(Y_i | X_i, D_1 = 1, D_2 = 0) = (\beta_0 + \beta_2) + \beta_1 X_i$$

高中

$$E(Y_i | X_i, D_1 = 0, D_2 = 1) = (\beta_0 + \beta_3) + \beta_1 X_i$$

大学及以上

- 在上例中 同时引入 性别和教育水平:

$$D_1 = \begin{cases} 1 & \text{男} \\ 0 & \text{女} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{大学及以上} \\ 0 & \text{大学以下} \end{cases}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_1 + \beta_3 D_2 + \mu_i$$

女职工本科以下学历的平均薪金：

$$E(Y_i | X_i, D_1 = 0, D_2 = 0) = \beta_0 + \beta_1 X_i$$

男职工本科以下学历的平均薪金：

$$E(Y_i | X_i, D_1 = 1, D_2 = 0) = (\beta_0 + \beta_2) + \beta_1 X_i$$

女职工本科以上学历的平均薪金：

$$E(Y_i | X_i, D_1 = 0, D_2 = 1) = (\beta_0 + \beta_3) + \beta_1 X_i$$

男职工本科以上学历的平均薪金：

$$E(Y_i | X_i, D_1 = 1, D_2 = 1) = (\beta_0 + \beta_2 + \beta_3) + \beta_1 X_i$$

2、乘法方式

- 加法方式引入虚拟变量，考察：截距的不同。
- 许多情况下，斜率发生变化，或斜率、截距同时发生变化。
- 斜率的变化，可通过以乘法的方式引入虚拟变量来测度。

- **例如，** 根据消费理论，收入决定消费。

但是，农村居民和城镇居民的边际消费倾向往往是不同的；这种消费倾向的不同，可通过在消费函数中引入虚拟变量来考察。

$$D_i = \begin{cases} 1 & \text{农村居民} \\ 0 & \text{城镇居民} \end{cases}$$

$$C_i = \beta_0 + \beta_1 X_i + \beta_2 D_i X_i + \mu_i$$

农村居民：

$$E(C_i | X_i, D_i = 1) = \beta_0 + (\beta_1 + \beta_2) X_i$$

城镇居民：

$$E(C_i | X_i, D_i = 0) = \beta_0 + \beta_1 X_i$$

- **例如，**根据消费理论，收入决定消费。

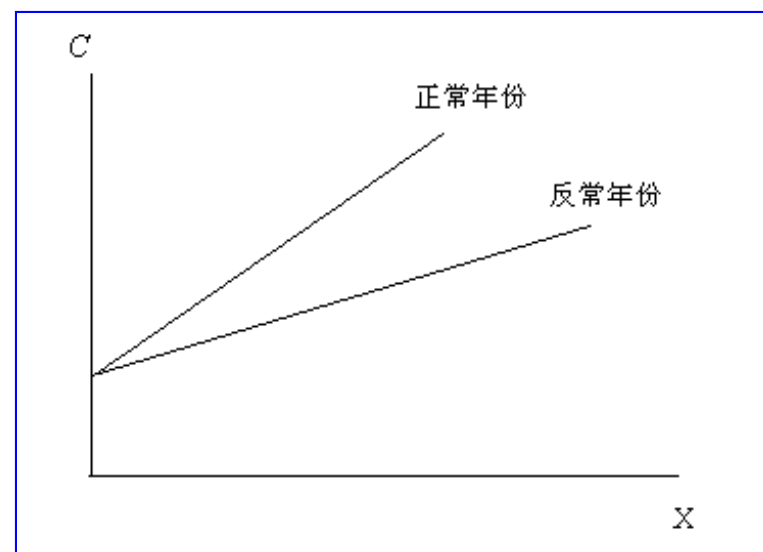
但是，在自然灾害、战争等反常年份，消费倾向往往发生变化；这种消费倾向的变化，可通过在消费函数中引入虚拟变量来考察。

$$D_t = \begin{cases} 1 & \text{正常年份} \\ 0 & \text{反常年份} \end{cases}$$

$$C_t = \beta_0 + \beta_1 X_t + \beta_2 D_t X_t + \mu_t$$

$$E(C_t | X_t, D_t = 1) = \beta_0 + (\beta_1 + \beta_2) X_t$$

$$E(C_t | X_t, D_t = 0) = \beta_0 + \beta_1 X_t$$



- **例如，**根据消费理论，收入决定消费。

但是，从某一个时点开始，消费倾向发生变化；这种消费倾向的变化，也可通过在消费函数中引入虚拟变量来考察。

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

$$C_t = \beta_0 + \beta_1 X_t + \beta_2 D_t X_t + \mu_t$$

$$E(C_t \mid X_t, D_t = 1) = \beta_0 + (\beta_1 + \beta_2) X_t$$

$$E(C_t \mid X_t, D_t = 0) = \beta_0 + \beta_1 X_t$$

3、同时引入加法与乘法形式的虚拟变量

- 当截距与斜率都发生变化时，则需要同时引入加法与乘法形式的虚拟变量。
- 对于一元模型，有两组样本，则有可能出现下述四种情况中的一种：
 - $\alpha_1 = \beta_1$ ，且 $\alpha_2 = \beta_2$ ，即两个回归相同，称为**重合回归**（Coincident Regressions）；
 - $\alpha_1 \neq \beta_1$ ，但 $\alpha_2 = \beta_2$ ，即两个回归的差异仅在其截距，称为**平行回归**（Parallel Regressions）；
 - $\alpha_1 = \beta_1$ ，但 $\alpha_2 \neq \beta_2$ ，即两个回归的差异仅在其斜率，称为**汇合回归**（Concurrent Regressions）；
 - $\alpha_1 \neq \beta_1$ ，且 $\alpha_2 \neq \beta_2$ ，即两个回归完全不同，称为**相异回归**（Dissimilar Regressions）。

- 例如，以1978-2009年的数据为样本，以GDP作为解释变量，建立居民消费函数。根据分析，1992年前后，自发消费和消费率都可能发生变化。

$$D_t = \begin{cases} 1 & \text{92年前} \\ 0 & \text{92年及以后} \end{cases}$$

$$C_t = \beta_0 + \beta_1 GDP_t + \beta_2 D_t + \beta_3 (D_t GDP_t) + \mu_t \\ t = 1978, \dots, 2009$$

通过统计检验，判断两个时期中消费函数的截距和斜率是否发生变化。

- **例5.1.1**以中国2007年各个地区城镇居民家庭人均可支配收入与人均生活消费支出，以及农村居民家庭人均纯收入与人均生活消费支出的相关数据，建立居民消费函数模型。
- 可以采用邹氏稳定性检验来考察农村居民与城镇居民边际消费倾向是否有差异。
- 也可以建立虚拟变量模型，考察农村居民与城镇居民边际消费倾向是否有差异。

$$Y_i = \beta_0 + \beta_1 X_i + \beta_3 D_i + \beta_4 (D_i X_i) + \mu_i$$

$$D_i = \begin{cases} 1 & \text{农村居民} \\ 0 & \text{城镇居民} \end{cases}$$

- 估计得到

$$\hat{Y}_i = 450.33 + 0.6920X_i - 271.14D_i + 0.0275D_iX_i$$

由变量显著性检验得到：2007年农村居民与城镇居民的边际消费倾向并无显著差异，他们有着共同的消费函数。

三、虚拟变量的设置原则

- “每一” 定性变量(qualitative variable)所需虚拟变量个数，要比该定性变量的状态类别数(categories)少1。
即若有m种状态，只在模型中引入m-1个虚拟变量。
- 例如，季节定性变量有春、夏、秋、冬4种状态，只需要设置3个虚变量：

$$D_1 = \begin{cases} 1 & \text{春季} \\ 0 & \text{其它} \end{cases} \quad D_2 = \begin{cases} 1 & \text{夏季} \\ 0 & \text{其它} \end{cases} \quad D_3 = \begin{cases} 1 & \text{秋季} \\ 0 & \text{其它} \end{cases}$$

若设置第4个虚变量，则出现“**虚拟变量陷阱**”
(Dummy Variable Trap)，为什么？

- 包含季节变量的正确模型：

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots \beta_k X_{kt} + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \mu_t$$

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots \beta_k X_{kt} + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + \mu_t$$

$$\mathbf{Y} = (\mathbf{X}, \mathbf{D}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} + \boldsymbol{\mu}$$

$$(\mathbf{X}, \mathbf{D}) = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} & 1 & 0 & 0 & 0 \\ 1 & X_{12} & \cdots & X_{k2} & 0 & 1 & 0 & 0 \\ 1 & X_{13} & \cdots & X_{k3} & 0 & 0 & 1 & 0 \\ 1 & X_{14} & \cdots & X_{k4} & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \cdots & & & & & \vdots \\ 1 & X_{16} & \cdots & X_{k6} & 1 & 0 & 0 & 0 \end{pmatrix}$$

4个D中每次必
有一个取1，解
释变量必然
完全共线性

- 如果在服装需求函数模型中必须包含3个定性变量：季节（4种状态）、性别（2种状态）、职业（5种状态），**应该设置多少虚变量？**
 - 模型含常数项
 - 模型不含常数项

讨论：定序定性变量可否按照状态赋值？

- 例如：表示居民对某种服务的满意程度，分5种状态：非常不满意、一般不满意、无所谓、一般满意、非常满意。在模型中按照状态分别赋值0、1、2、3、4或者-2、-1、0、1、2。
- 被经常采用，尤其在管理学、社会学研究领域。
- 正确的方法：
 - 设置多个虚拟变量，理论上正确，带来自由度损失。
 - 以定性变量为研究对象/因变量，构造多元排序离散选择模型，然后以模型结果、对定性变量的各种状态赋值；但这种办法，需要更多的信息支持。
- 赋值的方法、相当于对虚变量方法中的“各个虚变量的参数”施加了约束，但这种约束经常被检验为是不现实的。

讨论：虚变量与状态的不同对应关系对估计结果有无影响？

- 例3.2.2中，引入经济区位因素：东、中、西

$$\begin{array}{ll} D1 = \begin{cases} 1 & \text{西部} \\ 0 & \text{其它} \end{cases} & D2 = \begin{cases} 1 & \text{中部} \\ 0 & \text{其它} \end{cases} \\ DD1 = \begin{cases} 1 & \text{东部} \\ 0 & \text{其它} \end{cases} & DD2 = \begin{cases} 1 & \text{中部} \\ 0 & \text{其它} \end{cases} \end{array}$$

$$Y = -240.6137536 + 249.8125832 * D1 + 154.5909868 * D2 \\ + 0.6090284838 * X1 + 0.2032206892 * X2$$

$$Y = 9.198829575 - 249.8125832 * DD1 - 95.22159634 * DD2 \\ + 0.6090284838 * X1 + 0.2032206892 * X2$$

- 从上述2个得到：东部与中部自发性消费相差154.6，中部与西部相差95.2。
- 虚变量与状态的不同对应关系，对估计结果无实质影响。

§ 5.2 滞后变量模型

Lagged Variables Regression Models

- 一、滞后变量模型
- 二、分布滞后模型的参数估计
- 三、自回归模型的参数估计
- 四、格兰杰因果关系检验*

一、滞后变量模型

1、滞后变量

- 滞后被解释变量（Lagged explained variable）和滞后解释变量（Lagged explanatory variable）作为模型的解释变量。
- 一般出现在时间序列数据样本的模型中。
- 模型中出现滞后变量的原因：
 - 心理原因
 - 技术原因
 - 制度原因

2、滞后变量模型

- 以滞后变量作为解释变量，就得到**滞后变量模型**，**也称动态模型**。

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_q Y_{t-q} + \alpha_0 X_t + \alpha_1 X_{t-1} + \cdots + \alpha_s X_{t-s} + \mu_t$$

- **自回归分布滞后模型**（**Autoregressive Distributed Lag Model, ADL**）：既含有Y对自身滞后变量的回归，还包括着X分布在不同时期的滞后变量。

- **有限自回归分布滞后模型**：滞后期长度有限
- **无限自回归分布滞后模型**：滞后期无限

- **分布滞后模型 (distributed-lag model)** : 模型中不含滞后应变量, 仅有解释变量X的当期值及其若干期的滞后值。

$$Y_t = \alpha + \sum_{i=0}^s \beta_i X_{t-i} + \mu_t$$

β_0 : **短期(short-run)**或**即期乘数(impact multiplier)**, 表示本期X变化一单位对Y平均值的影响程度。

β_i ($i=1, 2, \dots, s$): **动态乘数**或**延迟系数**, 表示各滞后期X的变动对Y平均值影响的大小。

$\sum_{i=0}^s \beta_i$ 称为**长期（long-run）**或**均衡乘数（total distributed-lag multiplier）**，表示**X**变动一个单位，由于滞后效应而形成的对**Y**平均值总影响的大小。

如果各期的X值保持不变，则X与Y间的长期或均衡关系即为：

$$E(Y) = \alpha + \left(\sum_{i=0}^s \beta_i \right) X$$

- **自回归模型 (autoregressive model)** : 模型中的解释变量仅包含**X**的当期值与被解释变量**Y**的一个或多个滞后值 Y_{t-s} 。

$$Y_t = \alpha_0 + \alpha_1 X_t + \sum_{i=1}^q \beta_i Y_{t-i} + \mu_t$$

$$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + \mu_t$$

称为**一阶自回归模型 (first-order autoregressive model)**。

二、分布滞后DL模型的参数估计

1、分布滞后DL模型估计的困难

- 无限期的分布滞后模型，由于样本观测值的有限性，使得无法直接对其进行估计。
- 有限期的分布滞后模型，OLS会遇到如下问题：
 - 没有先验准则确定滞后期长度；
 - 如果滞后期较长，将缺乏足够的自由度进行估计和检验；
 - 同名变量滞后值之间可能存在高度线性相关，即模型存在高度的多重共线性。

2、分布滞后模型的修正估计方法

- 通过对各滞后变量加权，组成线性合成变量而有目的地减少滞后变量的数目，以缓解多重共线性、保证自由度。
- **经验加权法：**根据实际问题的特点和实际经验给各滞后变量指定权数，滞后变量按权数线性组合，构成新的变量。

权数据的类型有：**递减型、矩型、倒V型等。**

经验权数法的优点是：简单易行；缺点是：设置权数的随意性较大。

- 阿尔蒙（Almon）多项式法

主要思想：针对有限滞后期模型，通过阿尔蒙变换、定义新变量，以减少解释变量个数，然后用OLS法估计参数。

主要步骤为：

第一步，阿尔蒙变换

$$Y_t = \alpha + \sum_{i=0}^s \beta_i X_{t-i} + \mu_t$$

$$\beta_i = \sum_{k=0}^m \alpha_k (i)^k$$

$$i=0,1,\dots,s$$

例如取m=2

$$\beta_i = \sum_{k=0}^2 \alpha_k (i)^k = \alpha_0 + \alpha_1 (i) + \alpha_2 (i)^2$$

$$Y_t = \alpha + \sum_{i=0}^s \left(\sum_{k=0}^2 \alpha_k (i)^k \right) X_{t-i} + \mu_t$$

$$= \alpha + \alpha_0 \sum_{i=0}^s X_{t-i} + \alpha_1 \sum_{i=0}^s (i) X_{t-i} + \alpha_2 \sum_{i=0}^s (i)^2 X_{t-i} + \mu_t$$

$$Y_t = \alpha + \alpha_0 W_{0t} + \alpha_1 W_{1t} + \alpha_2 W_{2t} + \mu_t$$

第二步，模型的OLS估计

- 对变换后的模型进行OLS估计，得 α 的估计值；
- 计算滞后分布模型参数 β 的估计值。

在实际估计中，阿尔蒙多项式的阶数 m 一般取2或3，不超过4，否则达不到减少变量个数的目的。

由于 $m < s$ ，可以认为原模型存在的自由度不足和多重共线性问题已得到改善。

事实上，多项式分布滞后模型、比原分布滞后模型的多重共线性问题有可能增强了，而不是削弱了——引入 c 。

例5.2.2

- 发电量主要取决于电力部门固定资产，而固定资产是由历年的投资形成的，适合于建立分布滞后模型。
- 由于无法预知电力行业基本建设投资对发电量影响的时滞期，需取不同的滞后期试算。

经过试算发现，在2阶阿尔蒙多项式变换下，滞后期数取到第7期，估计结果的经济意义比较合理。

- 估计2阶阿尔蒙多项式模型：

$$\ln \hat{Y}_t = 6.732 + 0.025W_{t0} - 0.023W_{t1} + 0.006W_{t2}$$

- 计算分布滞后模型参数估计值，进而得到分布滞后模型估计式：

$$\ln \hat{Y}_t = 6.732 + 0.150 \ln X_t + 0.096 \ln X_{t-1} + 0.054 \ln X_{t-2} + 0.024 \ln X_{t-3}$$

$$+ 0.008 \ln X_{t-4} + 0.003 \ln X_{t-5} + 0.010 \ln X_{t-6} + 0.029 \ln X_{t-7}$$

- 直接对分布滞后模型进行OLS估计的结果：

$$\ln \hat{Y}_t = 6.74 + 0.124 \ln X_t + 0.167 \ln X_{t-1} - 0.010 \ln X_{t-2} + 0.049 \ln X_{t-3}$$

$$- 0.002 \ln X_{t-4} - 0.001 \ln X_{t-5} + 0.047 \ln X_{t-6} + 0.0003 \ln X_{t-7}$$

所有变量均未通过显著性检验，而且负值的出现也与实际经济意义不相符。

- 科伊克 (Koyck) 方法

科伊克方法是将无限分布滞后模型、转换为自回归模型，然后进行估计。

$$Y_t = \alpha + \sum_{i=0}^{\infty} \beta_i X_{t-i} + \mu_t$$

$$\beta_i = \beta_0 \lambda^i$$

$$Y_t = \alpha + \beta_0 \sum_{i=0}^{\infty} \lambda^i X_{t-i} + \mu_t$$

$$\lambda Y_{t-1} = \lambda \alpha + \beta_0 \sum_{i=1}^{\infty} \lambda^i X_{t-i} + \lambda \mu_{t-1}$$

$$Y_t - \lambda Y_{t-1} = (1 - \lambda) \alpha + \beta_0 X_t + \mu_t - \lambda \mu_{t-1}$$

$$Y_t = a + bX_t + cY_{t-1} + v_t$$

科伊克模型的特点：

- 以一个滞后因变量 Y_{t-1} 代替了大量的滞后解释变量 X_{t-i} ，最大限度地节省了自由度，解决了滞后期长度 s 难以确定的问题；
- 由于滞后一期的因变量 Y_{t-1} 与 X_t 的线性相关程度肯定小于 X 的各期滞后值之间的相关程度，从而缓解了多重共线性。

科伊克变换产生了两个新问题：

- 模型存在随机项 v_t 的一阶自相关性；
- 滞后被解释变量 Y_{t-1} 与随机项 v_t 不独立。

三、自回归模型的参数估计

1、自回归模型的构造

- 一个无限期分布滞后模型，可通过科伊克变换转化为自回归模型。
- 许多滞后变量模型都可以转化为自回归模型，自回归模型是经济生活中更常见的模型。
- 以适应预期模型以及局部调整模型为例进行说明。

- 自适应预期（Adaptive expectation, 调 X ）模型

$$Y_t = \beta_0 + \beta_1 X_t^e + \mu_t$$

$$X_t^e - X_{t-1}^e = r(X_t - X_{t-1}^e)$$

$$X_t^e = rX_t + (1-r)X_{t-1}^e$$

$$Y_t = \beta_0 + \beta_1[rX_t + (1-r)X_{t-1}^e] + \mu_t$$

$$(1-r)Y_{t-1} = \beta_0(1-r) + \beta_1(1-r)X_{t-1}^e + (1-r)\mu_{t-1}$$

$$Y_t = \beta_0 r + \beta_1 r X_t + (1-r)Y_{t-1} + v_t$$

$$v_t = \mu_t - (1-r)\mu_{t-1}$$

- 局部调整 (Partial Adjustment, 调Y) 模型

$$Y_t^e = \beta_0 + \beta_1 X_t + \mu_t$$

$$Y_t - Y_{t-1} = \delta(Y_t^e - Y_{t-1})$$

$$Y_t = \delta Y_t^e + (1 - \delta)Y_{t-1}$$

$$Y_t = \delta\beta_0 + \delta\beta_1 X_t + (1 - \delta)Y_{t-1} + \delta\mu_t$$

2、自回归模型的参数估计

- **自回归模型估计时的主要问题：**
 - 滞后被解释变量可能与随机扰动项相关；
 - 随机扰动项可能出现序列相关性。
- 视滞后被解释变量与随机扰动项之间的相关性选择估计方法。
- **工具变量法：**解释变量 \mathbf{Y}_{t-1} 与随机扰动项 μ_t 相关（例如科伊克模型、自适应预期模型）。
- **普通最小二乘法：**解释变量 \mathbf{Y}_{t-1} 与随机扰动项 μ_t 同期无关（例如局部调整模型）。

- 工具变量法只解决了解释变量与 μ_t 相关、对参数估计所造成的影响，但没有解决 μ_t 的自相关问题。
- 事实上，对于自回归模型， μ_t 项的自相关问题始终存在，对于此问题，至今没有完全有效的解决方法。唯一可做的，就是尽可能地建立“正确”的模型，以使序列相关性的程度减轻。
- 例5.2.3（自学）
 - 货币流通量局部调整模型的建立；
 - 货币流通量局部调整模型的估计。

四、格兰杰因果关系检验

Granger Test of Causality

1、原理

- 自回归分布滞后模型揭示：某变量变化，受其自身及其他变量过去行为的影响。
- 当两个变量在时间上有 先导—滞后关系 时，可从统计上考察这种关系是单向的还是双向。
 - 如果主要是一个变量过去的行为在影响另一个变量的当前行为，存在单向关系；
 - 如果双方的过去行为在相互影响着对方的当前行为，存在双向关系。
- 自回归/分布滞后ADL模型，可以用于变量间关系的检验。

2、格兰杰因果关系检验

$$Y_t = \sum_{i=1}^m \alpha_i X_{t-i} + \sum_{i=1}^m \beta_i Y_{t-i} + \mu_{1t}$$

$$X_t = \sum_{i=1}^m \lambda_i Y_{t-i} + \sum_{i=1}^m \delta_i X_{t-i} + \mu_{2t}$$

X对Y有单向影响： α 整体不为零，而 λ 整体为零；

Y对X有单向影响： λ 整体不为零，而 α 整体为零；

Y与X间存在双向影响： α 和 λ 整体不为零；

Y与X间不存在影响： α 和 λ 整体为零。

- 格兰杰检验，是通过受约束的F检验完成的。

如：

$$Y_t = \sum_{i=1}^m \alpha_i X_{t-i} + \sum_{i=1}^m \beta_i Y_{t-i} + \mu_{1t}$$

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

$$F = \frac{(RSS_R - RSS_U) / m}{RSS_U / (n - k)}$$

如果 $F > F_{\alpha}(m, n-k)$ ，则拒绝原假设。

能否说“X是Y的格兰杰原因”？为什么？

$$X_t = \sum_{i=1}^m \lambda_i Y_{t-i} + \sum_{i=1}^m \delta_i X_{t-i} + \mu_{2t}$$

$$H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_m = 0$$

$$F = \frac{(RSS_R - RSS_U) / m}{RSS_U / (n - k)}$$

如果 $F < F_\alpha(m, n-k)$ ， 则不拒绝原假设。

综合上述检验： **X是Y的格兰杰原因。**

- 格兰杰因果关系检验对于滞后期长度的选择有时很敏感；

不同滞后期，可能会得到完全不同检验结果。

- 一般首先以模型随机误差项、不存在序列相关为标准选取滞后期，然后进行因果关系检验。

3、例5.2.4

检验1978~2006年间中国当年价GDP (X) 与居民消费(Y)之间的因果关系。

数据

obs	X	Y
1	6678.8	3806.7
2	7551.6	4273.2
3	7944.2	4605.5
4	8438.0	5063.9
5	9235.2	5482.4
6	10074.6	5983.2
7	11565.0	6745.7
8	11601.7	7729.2
9	13036.5	8210.9
10	14627.7	8840.0
11	15794.0	9560.5
12	15035.5	9085.5
13	16525.9	9450.9
14	18939.6	10375.8
15	22056.5	11815.3
16	25897.3	13004.7
17	28783.4	13944.2
18	31175.4	15467.9
19	33853.7	17092.5
20	35956.2	18080.6
21	38140.9	19364.1
22	40277.0	20989.3
23	42964.6	22863.9
24	46385.4	24370.1
25	51274.0	26243.2
26	57408.1	28035.0
27	64623.1	30306.2
28	74580.4	33214.4
29	85623.1	36811.2

选择Granger检验

EViews - [Group: UNTITLED Workfile: E2.6\Untitled]

File Edit Object View Proc Quick Options Window Help

View Proc Object Print Name Freeze D

obs	X	
1	6678.8	3806.1
2	7551.6	4273.1
3	7944.2	4605.1
4	8438.0	5063.1
5	9235.2	5482.1
6	10074.6	5983.1
7	11565.0	6745.1
8	11601.7	7729.1
9	13036.5	8210.9
10	14627.7	8840.0
11	15794.0	9560.5
12	15035.5	9085.5

- Sample...
- Generate Series...
- Show ...
- Graph ▶
- Empty Group (Edit Series)
- Series Statistics ▶
- Group Statistics ▶**
 - Descriptive Statistics ▶
 - Covariances
 - Correlations
 - Cross Correlogram
 - Cointegration Test
 - Granger Causality Test**
- Estimate Equation...
- Estimate VAR...

选择检验的序列

View	Proc	Object	Print	Name	Freeze	Default	▼	Sort	Transpose	Edit+/-	Smpl+/-	InsDe
obs		X		Y								
1		6678.8		3806.7								
2		755										
3		794										
4		843										
5		923										
6		1007										
7		1156										
8		1160										
9		1303										
10		1462										
11		1579										
12		1503										
13		1652										
14		1893										
15		2205										
...		---		---								

Series List


List of series, groups, and/or series expressions

y x

OKCancel

确定滞后阶数（1阶）

View	Proc	Object	Print	Name	Freeze	Default	▼	Sort	Transpose	Edit+/-	Smpl-
obs		X		Y							
1		6678.8		3806.7							
2		7551.6		4273.2							
3		7944.2		4605.5							
4		8438.0		5063.9							
5		9235.2		5482.4							
6		10074.6		5983.2							
7		11565.0		6745.7							
8		11601.7		7729.2							
9		13036.5		8210.9							
10		14627.7		8840.0							
11		15794.0		9560.5							

Lag Specification 

Lags to

检验结果

View	Proc	Object	Print	Name	Freeze	Sample	Sheet	Stats	Spec
Pairwise Granger Causality Tests									
Date: 09/28/10 Time: 16:20									
Sample: 1 29									
Lags: 1									
Null Hypothesis:					Obs	F-Statistic	Probability		
X does not Granger Cause Y					28	15.1022	0.00066		
Y does not Granger Cause X						6.34368	0.01855		

由相伴概率知，在5%的显著性水平下，既拒绝“X不是Y的格兰杰原因”的假设，也拒绝“Y不是X的格兰杰原因”的假设。因此，从1阶滞后的情况看，可支配收入X的增长与居民消费支出Y增长互为格兰杰原因。

从检验模型随机干扰项1阶序列相关的LM检验看：以Y为被解释变量的模型的LM=0.897，对应的伴随概率P= 0.343，表明在5%的显著性水平下，该检验模型不存在序列相关性；

但是，以X为被解释变量的模型的LM=11.37，对应的伴随概率P= 0.001，表明在5%的显著性水平下，该检验模型存在严重的序列相关性。

检验结果

滞后长度	格兰杰因果性	F 检验的 P 值	LM(1)检验的 P 值	AIC 值	结论
1	$X \xrightarrow{*} Y$	0.001	0.343	14.36	拒绝
	$Y \xrightarrow{*} X$	0.019	0.001	16.87	拒绝
2	$X \xrightarrow{*} Y$	0.012	0.734	14.46	拒绝
	$Y \xrightarrow{*} X$	0.336	0.283	16.51	不拒绝
3	$X \xrightarrow{*} Y$	0.025	0.143	14.54	拒绝
	$Y \xrightarrow{*} X$	0.373	0.128	16.67	不拒绝
4	$X \xrightarrow{*} Y$	0.029	0.841	14.56	拒绝
	$Y \xrightarrow{*} X$	0.467	0.013	16.78	不拒绝

从2阶滞后期开始，检验模型都拒绝了“X不是Y的格兰杰原因”的假设，而不拒绝“Y不是X的原因”的假设。

滞后阶数为2或3时，两类检验模型都不存在序列相关性。

由赤池信息准则，发现滞后2阶检验模型拥有较小的AIC值。

可判断：**可支配收入X是居民消费支出Y的格兰杰原因，而不是相反，即国民收入的增加更大程度地影响着消费的增加。**

- **对于同阶单整的非平稳序列：**

- 理论上讲不能直接采用。
- 经过差分以后采用，经济意义发生变化。
- 模拟试验表明，当2个序列逐渐由平稳过程向非平稳过程过渡时，检验存在因果关系的概率出现一定程度的上升；但上升幅度，远小于2个序列之间因果关系的显著性增强时所引起的上升幅度。
- 同阶单整非平稳序列的Granger因果检验结果，具有一定的可靠性。

- **Granger因果检验是必要条件，不是充分条件。**

数据：农村居民消费 VS 城镇居民收入

obs	NCJMXF	CZJMSR
1995	1310.000	4283.000
1996	1572.000	4839.000
1997	1617.000	5160.000
1998	1590.000	5425.000
1999	1577.000	5854.000
2000	1670.000	6280.000
2001	1741.000	6860.000
2002	1834.000	7703.000
2003	1943.000	8472.000
2004	2185.000	9422.000
2005	2555.000	10493.00
2006	2829.000	11759.00

检验结果

Pairwise Granger Causality Tests

Date: 10/12/08 Time: 16:43

Sample: 1995 2006

Lags: 2

Null Hypothesis:	Obs	F-Statistic	Probability
<u>CZJMSR does not Granger Cause NCJMXF</u>	10	4.71731	0.07062
<u>NCJMXF does not Granger Cause CZJMSR</u>		0.51557	0.62579

- **结论：10%显著性水平下，城镇居民收入、是农村居民消费的单向原因，荒谬！**
- **统计检验、必须建立在经济关系分析的基础之上，结论才有意义。**

§ 5.3 模型设定偏误问题**

Model Specification Error(Bias)

- 一、模型设定偏误的类型
- 二、模型设定偏误的后果*
- 三、模型设定偏误的检验**

一、模型设定偏误的类型

Types of Specification errors(bias)

- Omission of a relevant variable(s)
- Inclusion of an unnecessary variable(s)
- Adopting the wrong functional form
- Errors of measurement
- Incorrect specification of the stochastic error term

To distinguish between model specification errors and model mis-specification errors

1、相关变量的遗漏（omitting relevant variables）

- 例如，如果“正确”的模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$$

而我们将模型设定为

$$Y = \alpha_0 + \alpha_1 X_1 + v$$

即设定模型时漏掉了一个相关的解释变量。
这类错误称为遗漏相关变量。

2、无关变量的误选 (including irrelevant variables)

- 例如，如果“真”的模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$$

但我们将模型设定为

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \mu$$

即设定模型时，多选了无关变量。

3、错误的函数形式 (wrong functional form)

- 例如，如果“真实”的回归函数为

$$Y = AX_1^{\beta_1} X_2^{\beta_2} e^{\mu}$$

但却将模型设定为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$$

二、模型设定偏误的后果*

1、遗漏相关变量偏误 (omitting relevant variable bias)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$$

$$Y = \alpha_0 + \alpha_1 X_1 + v$$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i - \bar{\mu}$$

$$\hat{\alpha}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}$$

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\sum x_{1i} y_i}{\sum x_{1i}^2} = \frac{\sum x_{1i} (\beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i - \bar{\mu})}{\sum x_{1i}^2} \\ &= \beta_1 + \beta_2 \frac{\sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i} (\mu_i - \bar{\mu})}{\sum x_{1i}^2}\end{aligned}$$

$$\hat{\alpha}_1 = \beta_1 + \beta_2 \frac{\sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i} (\mu_i - \bar{\mu})}{\sum x_{1i}^2}$$

- 若 X_2 与 X_1 相关， α_1 的估计量在小样本下有偏，在大样本下非一致。
- 若 X_2 与 X_1 不相关，则 α_1 的估计量满足无偏性与一致性；但这时 α_0 的估计却是有偏的。
- 随机扰动项的方差估计，也是有偏的。
- α_1 估计量的方差是有偏的。

$$Var(\hat{\alpha}_1) = \frac{\sigma^2}{\sum x_{1i}^2}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{1i}^2 (1 - r_{x_1 x_2}^2)}$$

2、包含无关变量偏误（including irrelevant variable bias）

$$Y = \alpha_0 + \alpha_1 X_1 + v$$



$$Var(\hat{\alpha}_1) = \frac{\sigma^2}{\sum x_{1i}^2}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$$



$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{1i}^2 (1 - r_{x_1 x_2}^2)}$$

- 对包含无关变量的模型进行估计，参数估计量是无偏的，但不具有最小方差性。

3、错误函数形式偏误（wrong functional form bias）

- 产生的偏误是全方位的。

三、模型设定偏误的检验**

1、检验是否含有无关变量

- **检验的基本思想**: 如果模型中误选了无关变量, 则其系数的真值应为零。

因此, 只须对无关变量系数的显著性进行检验。

- **t检验**: 检验某1个变量是否应包括在模型中;
- **F检验**: 检验若干个变量是否应同时包括在模型中。

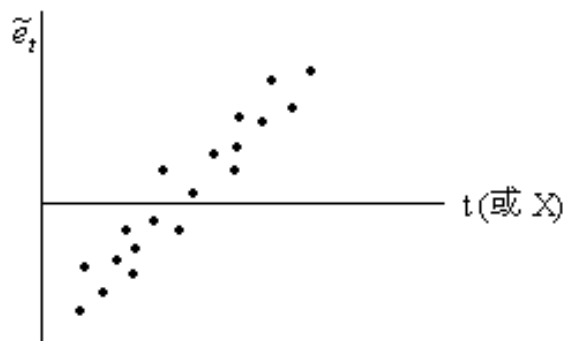
2、检验是否有相关变量的遗漏或函数形式设定偏误

- 残差图示法

对所设定的模型进行OLS回归，得到估计的残差序列 \tilde{e}_t ；

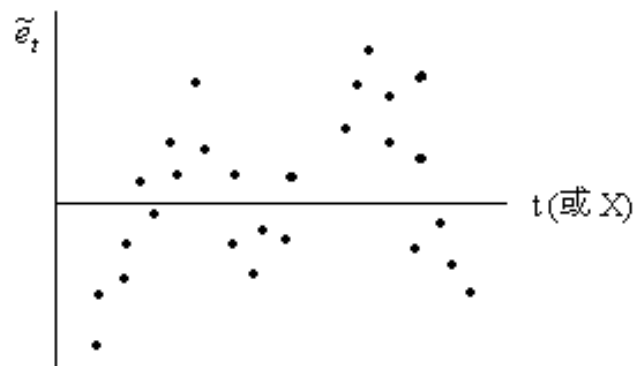
做出 \tilde{e}_t 与时间 t 或某解释变量 X 的散点图，考察 \tilde{e}_t 是否有规律地在变动，以判断是否遗漏了重要的解释变量或选取了错误的函数形式。

残差序列变化图



(a) 趋势变化：

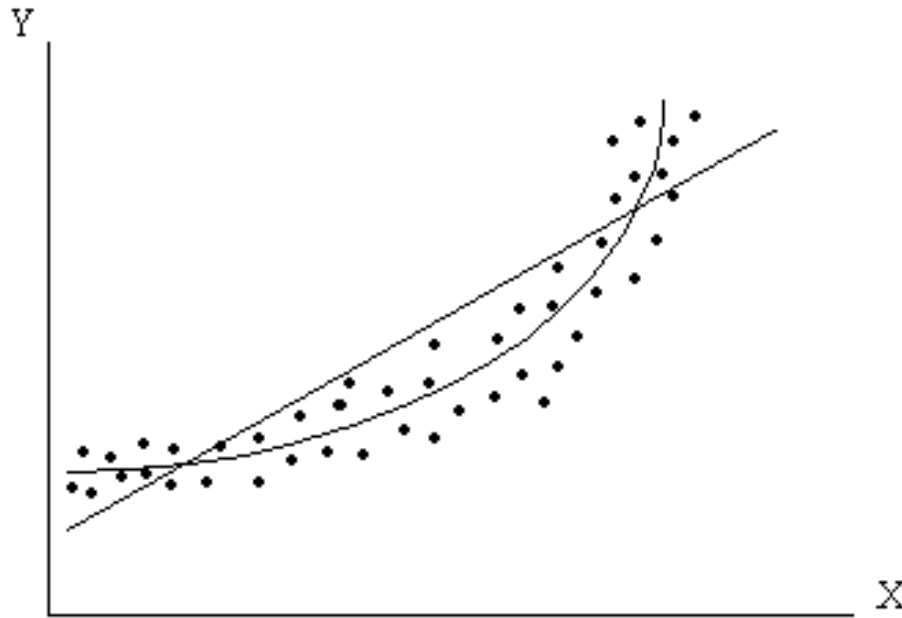
模型设定时可能遗漏了一随着时间的推移而持续上升的变量



(b) 循环变化：

模型设定时可能遗漏了一随着时间的推移而呈现循环变化的变量

模型函数形式设定偏误时，
残差序列呈现正负交替规律性变化



图示：一元回归模型中，真实模型呈幂函数形式，但却选取了线性函数进行回归。

- 一般性设定偏误检验

拉姆齐 (Ramsey) 于1969年提出的 RESET 检验
(regression error specification test) 。

RESET 检验基本思想：

- 如果事先知道遗漏了哪个变量，只需将此变量引入模型，估计并检验其参数是否显著不为零即可；
- 问题是不知道遗漏了哪个变量，需寻找一个替代变量Z，来进行上述检验。

RESET检验中，采用所设定模型中被解释变量Y的估计值 “ \hat{Y} 的若干次幂”、充当该“替代”变量。

RESET 检验步骤

- 估计原模型，得到残差和被解释变量的估计量；
- 根据它们的图形，判断应该引入 \hat{Y} 的若干次幂；
- 对增加变量的模型进行估计，并进行F检验或者t检验来判断是否增加这些“替代”变量。

RESET检验也可用来检验函数形式设定偏误的问题。

- 将非线性模型设定为线性，可近似认为遗漏了解释变量的2次、3次项；
- 引入模型，再进行检验。

RESET 检验例题

- 根据1978~2006年间中国当年价GDP (X) 与居民消费(Y)之间的因果关系检验结果：

以Y为被解释变量、X为解释变量，建立中国总量消费函数模型。

- 下面仅演示如何进行RESET检验，其它内容、见教科书例5.3.1。

原模型估计

View	Proc	Object	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: Y Method: Least Squares Date: 09/28/10 Time: 16:53 Sample: 1 29 Included observations: 29									
Variable		Coefficient	Std. Error	t-Statistic	Prob.				
C		2091.295	334.9869	6.242914	0.0000				
X		0.437527	0.009297	47.05950	0.0000				
R-squared		0.987955	Mean dependent var		14855.72				
Adjusted R-squared		0.987509	S.D. dependent var		9472.076				
S.E. of regression		1058.633	Akaike info criterion		16.83382				
Sum squared resid		30259014	Schwarz criterion		16.92811				
Log likelihood		-242.0903	F-statistic		2214.596				
Durbin-Watson stat		0.277155	Prob(F-statistic)		0.000000				

随机项具有强烈的1阶自相关性，
是否遗漏了重要的相关变量？

选择RESET检验

View	Proc	Object	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: Y Method: Least Squares Date: 09/28/10 Time: 16:53 Sample: 1 29 Included observations: 29						<div>Representations Estimation Output Actual, Fitted, Residual ▶ ARMA Structure... Gradients and Derivatives ▶ Covariance Matrix Coefficient Tests ▶ Residual Tests ▶ Stability Tests ▶ Label</div>			
Variable		Coefficient		rob.					
C		2091.29		0000					
X		0.43752							
R-squared		0.98795							
Adjusted R-squared		0.987509		S.D. dependent var		947			
S.E. of regression		1058.633		Akaike info criterion		16.			
Sum squared resid		30259014		Schwarz criterion		16.92811			
Log likelihood		-242.0903		F-statistic		2214.596			
Durbin-Watson stat		0.277155		Prob(F-statistic)		0.000000			

选择引入的变量数

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: Y
 Method: Least Squares
 Date: 09/28/10 Time: 16:53
 Sample: 1 29
 Included observations: 29

Variable	Coefficient	Prob.
C	2091.29	0.0000
X	0.43752	0.0000
R-squared	0.98796	4855.72
Adjusted R-squared	0.98750	472.076
S.E. of regression	1058.633	Akaike info criterion 16.83382
Sum squared resid	30259014	Schwarz criterion 16.92811
Log likelihood	-242.0903	F-statistic 2214.596
Durbin-Watson stat	0.277155	Prob(F-statistic) 0.000000

RESET Specifica...

Number of fitted

OK

Cancel

检验结果

View	Proc	Object	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Ramsey RESET Test:									
F-statistic		40.31008		Probability		0.000001			
Log likelihood ratio		27.15112		Probability		0.000000			
Test Equation:									
Dependent Variable: Y									
Method: Least Squares									
Date: 09/28/10 Time: 17:11									
Sample: 1 29									
Included observations: 29									
Variable		Coefficient		Std. Error		t-Statistic		Prob.	
C		421.2678		338.9401		1.242897		0.2250	
X		0.585373		0.024030		24.35981		0.0000	
FITTED^2		-8.63E-06		1.36E-06		-6.349022		0.0000	
R-squared		0.995277		Mean dependent var		14855.72			
Adjusted R-squared		0.994914		S.D. dependent var		9472.076			
S.E. of regression		675.5191		Akaike info criterion		15.96654			
Sum squared resid		11864476		Schwarz criterion		16.10798			
Log likelihood		-228.5148		F-statistic		2739.600			
Durbin-Watson stat		0.359172		Prob(F-statistic)		0.000000			

拒绝原模型与引入新变量的模型“可决系数无显著差异的 H_0 假设”，表明原模型确实存在遗漏相关变量的设定偏误。

- 线性模型与双对数线性模型的选择（仅供有兴趣的同学自学）